Stat 5101 Lecture Notes

Charles J. Geyer

May 7, 2001

Contents

Stat 5101 Lecture Notes

Charles J. Geyer

January 16, 2001

Contents

Chapter 1

Random Variables and Change of Variables

1.1 Random Variables

1.1.1 Variables

Before we tackle *random* variables, it is best to be sure we are clear about the notion of a mathematical variable. A *variable* is a symbol that stands for an unspecified mathematical object, like x in the expression $x^2 + 2x + 1$.

Often, it is clear from the context what kind of object the variable stands for. In this example, x can be any real number. But not all variables are numerical. We will also use vector variables and variables taking values in arbitrary sets.

Thus, when being fussy, we specify the kind of mathematical objects a variable can symbolize. We do this by specifying the set of objects which are possible *values* of the variable. For example, we write

$$x^2 + 2x + 1 = (x + 1)^2, \qquad x \in \mathbb{R},$$

to show that the equality holds for any real number x, the symbol $\mathbb{R}$ indicating the set of all real numbers.

1.1.2 Functions

In elementary mathematics, through first year calculus, textbooks, teachers, and students are often a bit vague about the notion of a function, not distinguishing between a function, the value of a function, the graph of a function, or an expression defining a function. In higher mathematics, we are sometimes just as vague when it is clear from the context what is meant, but when clarity is needed, especially in formal definitions, we are careful to distinguish between these concepts.

A *function* is a rule f that assigns to each element x of a set called the *domain* of the function an object $f(x)$ called the *value* of the function at x. Note the distinction between the function f and the value $f(x)$. There is also a distinction between a function and an expression defining the function. We say, let f be the function defined by

$$f(x) = x^2, \qquad x \in \mathbb{R}. \tag{1.1}$$

Strictly speaking, (1.1) isn't a function, it's an expression defining the function f. Neither is x^2 the function, it's the *value* of the function at the point x. The function f is the rule that assigns to each x in the domain, which from (1.1) is the set $\mathbb{R}$ of all real numbers, the value $f(x) = x^2$.

As we already said, most of the time we do not need to be so fussy, but some of the time we do. Informality makes it difficult to discuss some functions, in particular, the two kinds described next. These functions are important for other reasons besides being examples where care is required. They will be used often throughout the course.

Constant Functions

By a constant function, we mean a function that has the same value at all points, for example, the function f defined by

$$f(x) = 3, \qquad x \in \mathbb{R}. \tag{1.2}$$

We see here the difficulty with vagueness about the function concept. If we are in the habit of saying that x^2 is a function of x, what do we say here? The analogous thing to say here is that 3 is a function of x. But that looks and sounds really weird. The careful statement, that f is a function defined by (1.2), is wordy, but not weird.

Identity Functions

The *identity function* on an arbitrary set S is the function f defined by

$$f(x) = x, \qquad x \in S. \tag{1.3}$$

Here too, the vague concept seems a bit weird. If we say that x^2 is a function, do we also say x is a function (the identity function)? If so, how do we distinguish between the variable x and the function x? Again, the careful statement, that f is a function defined by (1.3), is wordy, but not weird.

Range and Codomain

If f is a function with domain A, the *range* of f is the set

$$\operatorname{range} f = \{\, f(x) : x \in S \,\}$$

of all values $f(x)$ for all x in the domain.

Sometimes it is useful to consider f as a map from its domain A into a set B. We write $f : A \to B$ or

$$A \xrightarrow{f} B$$

to indicate this. The set B is called the *codomain* of f.

Since all the values $f(x)$ of f are in the codomain B, the codomain necessarily includes the range, but may be larger. For example, consider the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^2$. The *codomain* is $\mathbb{R}$, just because that's the way we defined f, but the *range* is the interval $[0, \infty)$ of nonnegative real numbers, because squares are nonnegative.

1.1.3 Random Variables: Informal Intuition

Informally, a *random variable* is a variable that is *random*, meaning that its value is unknown, uncertain, not observed yet, or something of the sort. The probabilities with which a random variable takes its various possible values are described by a probability model.

In order to distinguish random variables from ordinary, nonrandom variables, we adopt a widely used convention of denoting random variables by capital letters, usually letters near the end of the alphabet, like X, Y, and Z.

There is a close connection between random variables and certain ordinary variables. If X is a random variable, we often use the corresponding small letter x as the ordinary variable that takes the same values.

Whether a variable corresponding to a real-world phenomenon is considered random may depend on context. In applications, we often say a variable is random *before it is observed* and nonrandom *after it is observed* and its actual value is known. Thus the same real-world phenomenon may be symbolized by X before its value is observed and by x after its value is observed.

1.1.4 Random Variables: Formal Definition

The formal definition of a random variable is rather different from the informal intuition. Formally, a random variable isn't a *variable*, it's a *function*.

Definition 1.1.1 (Random Variable).
A ***random variable*** *in a probability model is a function on the sample space of a probability model.*

The capital letter convention for random variables is used here too. We usually denote random variables by capital letters like X. When considered formally a random variable X a function on the sample space S, and we can write

$$S \xrightarrow{X} T$$

if we like to show that X is a map from its domain S (always the sample space) to its codomain T. Since X is a function, its values are denoted using the usual notation for function values $X(s)$.

An Abuse of Notation

A widely used shorthand that saves quite a bit of writing is to allow a relation specifying an event rather than an event itself as the apparent argument of a probability measure, that is, we write something like

$$P(X \in A) \tag{1.4}$$

or

$$P(X \le x). \tag{1.5}$$

Strictly speaking, (1.4) and (1.5) are nonsense. The argument of a probability measure is an event (a subset of the sample space). Relations are not sets. So (1.4) and (1.5) have the wrong kind of arguments.

But it is obvious what is meant. The events in question are the sets defined by the relations. To be formally correct, in place of (1.4) we should write $P(B)$, where

$$B = \{\, s \in S : X(s) \in A \,\}, \tag{1.6}$$

and in place of (1.5) we should write $P(C)$, where

$$C = \{\, s \in S : X(s) \le x \,\}. \tag{1.7}$$

Of course we could always plug (1.6) into $P(B)$ getting the very messy

$$P(\{\, s \in S : X(s) \in A \,\}) \tag{1.8}$$

It is clear that (1.4) is much simpler and cleaner than (1.8).

Note in (1.5) the role played by the two exes. The "big X" is a random variable. The "little x" is an ordinary (nonrandom) variable. The expression (1.5) stands for any statement like

$$P(X \le 2)$$

or

$$P(X \le -4.76)$$

Why not use *different* letters so as to make the distinction between the two variables clearer? Because we want to make an association between the random variable "big X" and the ordinary variable "little x" that stands for a *possible value* of the random variable X. Anyway this convention is very widely used, in all probability and statistics books, not just in this course, so you might as well get used to it.

The Incredible Disappearing Identity Random Variable

By "identity random variable" we mean the random variable X on the sample space S defined by

$$X(s) = s, \qquad s \in S,$$

that is, X is the identity function on S.

As we mentioned in our previous discussion of identity functions, when you're sloppy in terminology and notation the identity function disappears. If you don't distinguish between functions, their values, and their defining expressions x is both a variable and a function. Here, sloppiness causes the disappearance of the distinction between the random variable "big X" and the ordinary variable "little s." If you don't distinguish between the function X and its values $X(s)$, then X is s.

When we plug in $X(s) = s$ into the expression (1.6), we get

$$B = \{ \, s \in S : s \in A \, \} = A.$$

Thus when X is the identity random variable $P(X \in A)$ is just another notation for $P(A)$. Caution: when X is *not* the identity random variable, this isn't true.

Another Useful Notation

For probability models (distributions) having a standard abbreviation, like $\text{Exp}(\lambda)$ for the exponential distribution with parameter λ we use the notation

$$X \sim \text{Exp}(\lambda)$$

as shorthand for the statement that X is a random variable with this probability distribution. Strictly speaking, X is the identity random variable for the $\text{Exp}(\lambda)$ probability model.

Examples

Example 1.1.1 (Exponential Random Variable).
Suppose

$$X \sim \text{Exp}(\lambda).$$

What is

$$P(X > x),$$

for $x > 0$?

The definition of the probability measure associated with a continuous probability model says

$$P(A) = \int_A f(x) \, dx.$$

We only have to figure what event A we want and what density function f.

> *To calculate the probability of an event A. Integrate the density over A for a continuous probability model (sum over A for a discrete model).*

The event A is

$$A = \{ \, s \in \mathbb{R} : s > x \, \} = (x, \infty),$$

and the density of the Exp(λ) distribution is from the handout

$$f(x) = \lambda e^{-\lambda x}, \qquad x > 0.$$

We only have to plug these into the definition and evaluate the integral.

But when we do so, we have to be careful. We cannot just put in the limits of integration x and ∞ giving

$$P(A) = \int_x^\infty f(x)\,dx, \tag{1.9}$$

because the x in the limit of integration isn't the same as the x that is the variable of integration (in $f(x)\,dx$. In fact, this formula is obviously wrong because it violates a basic sanity check of calculus

> *The "dummy" variable of integration* ***never*** *appears in the limits of integration or in the expression that is the value of the integral.*

Thus we need to use some other variable, say s, as the dummy variable of integration (it's called a "dummy" variable, because the value of the integral doesn't contain this variable, so it doesn't matter what variable we use.) This gives

$$\begin{aligned} P(A) &= \int_x^\infty f(s)\,ds \\ &= \int_x^\infty \lambda e^{-\lambda s}\,ds \\ &= -e^{-\lambda s}\Big|_x^\infty \\ &= e^{-\lambda x} \end{aligned}$$

Note that in the second line

$$f(s) = \lambda e^{-\lambda s}.$$

When we replace $f(x)$ by $f(s)$, we replace x by s everywhere x appears in the definition of $f(x)$.

Example 1.1.2 (A More Complicated Event).
Suppose, as before,

$$X \sim \text{Exp}(\lambda).$$

But know we want to know

$$P\big((X-\mu)^2 < a^2\big), \tag{1.10}$$

where μ and a are positive real numbers.

We follow the same strategy as before. We need to evaluate (1.9), where A is the event implicitly defined in (1.10), which is

$$\begin{aligned} A &= \{\, x > 0 : x < \mu - a \,\text{or}\, x > \mu + a \,\} \\ &= (0, \mu - a) \cup (\mu + a, \infty) \end{aligned}$$

the union of two disjoint intervals unless $\mu - a < 0$, in which case the lower interval is empty.

This mean that (1.9) becomes the sum of integrals over these two disjoint sets

$$\begin{aligned} P(A) &= \int_0^{\mu - a} f(x) \, dx + \int_{\mu + a}^\infty f(x) \, dx \\ &= -e^{-\lambda x} \Big|_0^{\mu - a} - e^{-\lambda x} \Big|_{\mu + a}^\infty \\ &= (1 - e^{-\lambda(\mu - a)}) + e^{-\lambda(\mu + a)} \end{aligned}$$

unless $\mu - a < 0$, in which case it is

$$\begin{aligned} P(A) &= \int_{\mu + a}^\infty f(x) \, dx \\ &= e^{-\lambda(\mu + a)} \end{aligned}$$

1.1.5 Functions of Random Variables

One immediate consequence of the formal definition of random variables is that any function of random variables is another random variable. Suppose X and Y are real valued random variables and we define $Z = X^2 Y$. Then Z is also a function on the sample space S defined by

$$Z(s) = X(s)^2 Y(s), \qquad s \in S,$$

and similarly for any other function of random variables.

1.2 Change of Variables

1.2.1 General Definition

Consider a random variable X and another random variable Y defined by $Y = g(X)$, where g is an arbitrary function. Every function of random variables is a random variable!

Note that

$$P(Y \in A) = P\bigl(g(X) \in A\bigr). \tag{1.11}$$

In one sense (1.11) is trivial. The two sides are equal because $Y = g(X)$.

In another sense (1.11) is very deep. It contains the heart of the most general change of variable formula. It tells how to calculate probabilities for Y in terms of probabilities for X. To be precise, let P_X denote the probability measure for the model in which X is the identity random variable, and similarly P_Y for the analogous measure for Y. Then the left hand side of (1.11) is trivial is $P_Y(A)$ and the right hand side is $P_X(B)$, where

$$B = \{\, s \in S : g(s) \in A \,\} \tag{1.12}$$

where S is the sample space of the probability model describing X. We could have written $g\bigl(X(s)\bigr)$ in place of $g(s)$ in (1.12), but since X is the identity random variable for the P_X model, these are the same. Putting this all together, we get the following theorem.

Theorem 1.1. *If $X \sim P_X$ and $Y = g(X)$, then $Y \sim P_Y$ where*

$$P_Y(A) = P_X(B),$$

the relation between A and B being given by (1.12)*.*

This theorem is too abstract for everyday use. In practice, we will use at lot of other theorems that handle special cases more easily. But it should not be forgotten that this theorem exists and allows, at least in theory, the calculation of the distribution of *any* random variable.

Example 1.2.1 (Constant Random Variable).
Although the theorem is hard to apply to complicated random variables, it is not too hard for simple ones. The simplest random variable is a constant one. Say the function g in the theorem is the constant function defined by $g(s) = c$ for all $s \in S$.

To apply the theorem, we have to find, for any set A in the sample of Y, which is the codomain of the function g, the set B defined by (1.12). This sounds complicated, and in general it is, but here is it fairly easy. There are actually only two cases.

Case I: Suppose $c \in A$. Then

$$B = \{\, s \in S : g(s) \in A \,\} = S$$

because $g(s) = c \in A$ for all s in S.

Case II: Conversely, suppose $c \notin A$. Then

$$B = \{\, s \in S : g(s) \in A \,\} = \varnothing$$

because $g(s) = c \notin A$ for all s in S, that is there is no s such that the condition holds, so the set of s satisfying the condition is empty.

Combining the Cases: Now for any probability distribution the empty set has probability zero and the sample space has probability one, so $P_X(\varnothing) = 0$ and $P_X(S) = 1$. Thus the theorem says

$$P_Y(A) = \begin{cases} 1, & c \in A \\ 0, & c \notin A \end{cases}$$

Thus even constant random variables have probability distributions. They are rather trivial, all the probabilities being either zero or one, but they are probability models that satisfy the axioms.

Thus in probability theory we treat nonrandomness as a special case of randomness. There is nothing uncertain or indeterminate about a constant random variable. When Y is defined as in the example, we always know $Y = g(X) = c$, regardless of what happens to X. Whether one regards this as mathematical pedantry or a philosophically interesting issue is a matter of taste.

1.2.2 Discrete Random Variables

For discrete random variables, probability measures are defined by sums

$$P(A) = \sum_{x \in A} f(x) \tag{1.13}$$

where f is the density for the model (Lindgren would say p. f.)

Note also that for discrete probability models, not only is there (1.13) giving the measure in terms of the density, but also

$$f(x) = P(\{x\}). \tag{1.14}$$

giving the density in terms of the measure, derived by taking the case $A = \{x\}$ in (1.13). This looks a little odd because x is a point in the sample space, and a point is not a set, hence not an event, the analogous event is the set $\{x\}$ containing the single point x.

Thus our job in applying the change of variable theorem to discrete probability models is much simpler than the general case. We only need to consider sets A in the statement of the theorem that are one-point sets. This gives the following theorem.

Theorem 1.2. *If X is a discrete random variable with density f_X and sample space S, and $Y = g(X)$, then Y is a discrete random variable with density f_Y defined by*

$$f_Y(y) = P_X(B) = \sum_{x \in B} f_X(x),$$

where

$$B = \{\, x \in S : y = g(x) \,\}.$$

Those who don't mind complicated notation plug the definition of B into the definition of f_Y obtaining

$$f_Y(y) = \sum_{\substack{x \in S \\ y = g(x)}} f_X(x).$$

In words, this says that to obtain the density of a *discrete* random variable Y, one sums the probabilities of all the points x such that $y = g(x)$ for each y.

Even with the simplification, this theorem is still a bit too abstract and complicated for general use. Let's consider some special cases.

One-To-One Transformations

A transformation (change of variable)

$$S \xrightarrow{g} T$$

is *one-to-one* if g maps each point x to a different value $g(x)$ from all other points, that is,

$$g(x_1) \neq g(x_2), \qquad \text{whenever } x_1 \neq x_2.$$

A way to say this with fewer symbols is to consider the equation

$$y = g(x).$$

If for each fixed y, considered as an equation to solve for x, there is a *unique* solution, then g is one-to-one. If for any y there are multiple solutions, it isn't.

Whether a function is one-to-one or not may depend on the domain. So if you are sloppy and don't distinguish between a function and an expression giving the value of the function, you can't tell whether it is one-to-one or not.

Example 1.2.2 (x^2).
The function $g : \mathbb{R} \to \mathbb{R}$ defined by $g(x) = x^2$ is *not* one-to-one because

$$g(x) = g(-x), \qquad x \in \mathbb{R}.$$

So it is in fact two-to-one, except at zero.

But the function $g : (0, \infty) \to \mathbb{R}$ defined by the *very same formula* $g(x) = x^2$ *is* one-to-one, because there do not exist distinct *positive* real numbers x_1 and x_2 such that $x_1^2 = x_2^2$. (Every positive real number has a unique *positive* square root.)

This example seems simple, and it is, but every year some students get confused about this issue on tests. If you don't know whether you are dealing with a one-to-one transformation or not, you'll be in trouble. And you can't tell without considering the domain of the transformation as well as the expression giving its values.

Inverse Transformations

A function is *invertible* if it is *one-to-one* and *onto*, the latter meaning that its codomain is the same as its range.

Neither of the functions considered in Example 1.2.2 are invertible. The second is one-to-one, but it is not onto, because the g defined in the example maps positive real numbers to positive real numbers. To obtain a function that is invertible, we need to restrict the codomain to be the same as the range, defining the function

$$g : (0, \infty) \to (0, \infty)$$

by

$$g(x) = x^2.$$

Every invertible function

$$S \xrightarrow{g} T$$

has an *inverse* function

$$T \xrightarrow{g^{-1}} S$$

(note g^{-1} goes in the direction opposite to g) satisfying

$$g\left(g^{-1}(y)\right) = y, \qquad y = inT$$

and

$$g^{-1}\bigl(g(x)\bigr) = x, \qquad x = inS.$$

A way to say this that is a bit more helpful in doing actual calculations is

$$y = g(x) \quad \text{whenever} x = g^{-1}(y).$$

The inverse function is *discovered* by trying to solve

$$y = g(x)$$

for x. For example, if

$$y = g(x) = x^2$$

then

$$x = \sqrt{y} = g^{-1}(y).$$

If for any y there is no solution or multiple solutions, the inverse does not exist (if no solutions the function is not onto, if multiple solutions it is not one-to-one).

Change of Variable for Invertible Transformations

For invertible transformations Theorem 1.2 simplifies considerably. The set B in the theorem is always a singleton: there is a unique x such that $y = g(x)$, namely $g^{-1}(y)$. So

$$B = \{\, g^{-1}(y) \,\},$$

and the theorem can be stated as follows.

Theorem 1.3. *If X is a discrete random variable with density f_X and sample space S, if $g : S \to T$ is an invertible transformation, and $Y = g(X)$, then Y is a discrete random variable with density f_Y defined by*

$$f_Y(y) = f_X\bigl(g^{-1}(y)\bigr), \qquad y \in T. \tag{1.15}$$

Example 1.2.3 (The "Other" Geometric Distribution).
Suppose $X \sim \text{Geo}(p)$, meaning that X has the density

$$f_X(x) = (1-p)p^x, \qquad x = 0, 1, 2, \ldots \tag{1.16}$$

Some people like to start counting at one rather than zero (Lindgren among them) and prefer to call the distribution of the random variable $Y = X + 1$ the "geometric distribution" (there is no standard, some people like one definition, some people like the other).

The transformation in question is quite simple

$$y = g(x) = x + 1$$

has inverse

$$x = g^{-1}(y) = y - 1$$

if (big if) we get the domains right. The domain of X is the set of nonnegative integers $\{0, 1, \ldots\}$. The transformation g maps this to the set of *positive* integers $\{1, 2, \ldots\}$. So that is the range of g and the domain of g^{-1} and hence the sample space of the distribution of Y. If we don't get the domains right, we don't know the sample space for Y and so can't completely specify the distribution.

Now we just apply the theorem. The density f_X in the theorem is defined by (1.16). The expression $f_X\bigl(g^{-1}(y)\bigr)$ in the theorem means that everywhere we see an x in the definition of $f_X(x)$, we plug in $g^{-1}(y) = y - 1$. This gives

$$f_Y(y) = (1-p)p^{y-1}, \qquad y - 1 = 0, 1, 2, \ldots.$$

The condition on the right giving the possible values of y is not in the usual form. If we clean it up, we get

$$f_Y(y) = (1-p)p^{y-1}, \qquad y = 1, 2, 3, \ldots \tag{1.17}$$

Note that this does indeed say that Y has the domain (sample space) we figured out previously.

Example 1.2.4 (A Useless Example).
Again consider the geometric distribution with density (1.16), but now consider the transformation $g(x) = x^2$. Since the domain is the nonnegative integers, g is one-to-one. In order to make it onto, we must make the codomain equal to the range, which is the set $\{0, 1, 4, 9, 16, \ldots\}$ of perfect squares. The inverse transformation is $x = \sqrt{y}$, and applying the theorem gives

$$f_Y(y) = (1-p)p^{\sqrt{y}}, \qquad y = 0, 1, 4, 9, 16, \ldots$$

for the density of $Y = g(X)$.

The reason this is called a "useless example" is that the formula is fairly messy, so people avoid it. In general one never *has to* do a change of variable unless a test question or homework problem makes you. One can always do the calculation using f_X rather than f_Y. The question is which is easier.

1.2.3 Continuous Random Variables

For continuous random variables, probability measures are defined by integrals

$$P(A) = \int_A f(x)\,dx \tag{1.18}$$

where f is the density for the model (Lindgren would say p. d. f.)

So far (one sentence) this section looks much like the section on discrete random variables. The only difference is that (1.18) has an integral where (1.13) has a sum. But the next equation (1.14) in the section on discrete random variables has no useful analog for continuous random variables. In fact

$$P(\{x\}) = 0, \qquad \text{for all } x$$

(p. 32 in Lindgren). Because of this there is no simple analog of Theorem 1.2 for continuous random variables.

There is, however, an analog of Theorem 1.3.

Theorem 1.4. *If X is a continuous random variable with density f_X and sample space S, if $g : S \to T$ is an invertible transformation with differentiable inverse $h = g^{-1}$, and $Y = g(X)$, then Y is a continuous random variable with density f_Y defined by*

$$f_Y(y) = f_X\bigl(h(y)\bigr) \cdot |h'(y)|, \qquad y \in T. \tag{1.19}$$

The first term on the right hand side in (1.19) is the same as the right hand side in (1.15), the only difference is that we have written h for g^{-1}. The second term has no analog in the discrete case. Here summation and integration, and hence discrete and continuous random variables, are not analogous.

We won't bother to prove this particular version of the theorem, since it is a special case of a more general theorem we will prove later (the multivariable continuous change of variable theorem).

Example 1.2.5.
Suppose

$$X \sim \operatorname{Exp}(\lambda).$$

What is the distribution of $Y = X^2$?

This is just like Example 1.2.4 except now we use the continuous change of variable theorem.

The transformation in question is $g : (0, \infty) \to (0, \infty)$ defined by

$$g(x) = x^2, \qquad x > 0.$$

The inverse transformation is, of course,

$$h(y) = g^{-1}(y) = y^{1/2}, \qquad y > 0,$$

and it also maps from $(0, \infty)$ to $(0, \infty)$. Its derivative is

$$h'(y) = \tfrac{1}{2} y^{-1/2}, \qquad y > 0.$$

The density of X is

$$f_X(x) = \lambda e^{-\lambda x}, \qquad x > 0.$$

Plugging in $h(y) = \sqrt{y}$ everywhere for x gives

$$f_X\bigl(h(y)\bigr) = \lambda e^{-\lambda \sqrt{y}}$$

And multiplying by the derivative term gives the density of Y.

$$\begin{aligned} f_Y(y) &= f_X\bigl(h(y)\bigr) \cdot |h'(y)| \\ &= \lambda e^{-\lambda \sqrt{y}} \cdot \tfrac{1}{2} y^{-1/2} \\ &= \frac{\lambda e^{-\lambda \sqrt{y}}}{2 \sqrt{y}}, \qquad y > 0. \end{aligned}$$

Note that we tack the range of y values on at the end. The definition of f_Y isn't complete without it.

1.3 Random Vectors

A *vector* is a mathematical object consisting of a *sequence* or *tuple* of real numbers. We usually write vectors using boldface type

$$\mathbf{x} = (x_1, \ldots, x_n)$$

The separate numbers x_1, …, x_n are called the *components* or *coordinates* of the vector. We can also think of a vector as a point in n-dimensional Euclidean space, denoted $\mathbb{R}^n$.

A *random vector* is simply a vector-valued random variable. Using the "big X" and "little x" convention, we denote random vectors by capital letters and their possible values by lower case letters. So a random vector

$$\mathbf{X} = (X_1, \ldots, X_n)$$

is a vector whose components are real-valued random variables X_1, …, X_n. For contrast with *vectors*, real numbers are sometimes called *scalars.* Thus most of the random variables we have studied up to now can be called *random scalars* or scalar-valued random variables.

Strictly speaking, there is a difference between a function f of a vector variable having values $f(\mathbf{x})$ and a function f of several scalar variables having values $f(x_1, \ldots, x_n)$. One function has one argument, the other n arguments. But in practice we are sloppy about the distinction, so we don't have to write $f\bigl((x_1, \ldots, x_n)\bigr)$ when we want to consider f a function of a vector variable and explicitly show the components of the vector. The sloppiness, which consists in merely omitting a second set of parentheses, does no harm.

That having been said, there is nothing special about random vectors. They follow the same rules as random scalars, though we may need to use some boldface letters to follow our convention.

1.3.1 Discrete Random Vectors

A real-valued function f on a countable subset S of $\mathbb{R}^n$ is the *probability density* (Lindgren would say p. f.) of a discrete random vector if it satisfies the following two properties

$$f(\mathbf{x}) \geq 0, \qquad \text{for all } \mathbf{x} \in S \tag{1.20a}$$

$$\sum_{\mathbf{x} \in S} f(\mathbf{x}) = 1 \tag{1.20b}$$

The corresponding *probability measure* ("big P") is defined by

$$P(A) = \sum_{\mathbf{x} \in A} f(\mathbf{x}) \tag{1.20c}$$

for all events A (events being, as usual, subsets of the sample space S).

Except for the boldface type, these are exactly the same properties that characterize probability densities and probability measures of a discrete random scalar. The only difference is that $\mathbf{x}$ is really an n-tuple, so f is "really" a function of several variables, and what looks simple in this notation, may be complicated in practice. We won't give an example here, but will wait and make the point in the context of continuous random vectors.

1.3.2 Continuous Random Vectors

Similarly, a real-valued function f on a subset S of $\mathbb{R}^n$ is the *probability density* (Lindgren would say p. d. f.) of a continuous random vector if it satisfies the following two properties

$$f(\mathbf{x}) \geq 0, \qquad \text{for all } \mathbf{x} \in S \tag{1.21a}$$

$$\int_S f(\mathbf{x})\, d\mathbf{x} = 1 \tag{1.21b}$$

The corresponding *probability measure* is defined by

$$P(A) = \int_A f(\mathbf{x})\, d\mathbf{x} \tag{1.21c}$$

for all events A (events being, as usual, subsets of the sample space S).

Again, except for the boldface type, these are exactly the same properties that characterize probability densities and probability measures of a continuous random scalar. Also note that the similarity between the discrete and continuous cases, the only difference being summation in one and integration in the other.

To pick up our point about the notation hiding rather tricky issues, we go back to the fact that f is "really" a function of several random variables, so the integrals in (1.21b) and (1.21c) are "really" multiple (or iterated) integrals. Thus (1.21c) could perhaps be written more clearly as

$$P(A) = \iint_A \cdots \int f(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n$$

Whether you prefer this to (1.21c) is a matter of taste. It does make some of the difficulty more explicit.

Example 1.3.1.
Suppose that f is the probability density on the unit square in $\mathbb{R}^2$ defined by

$$f(x, y) = x + y, \qquad 0 < x < 1 \text{ and } 0 < y < 1. \tag{1.22}$$

Suppose we wish to calculate $P(X + Y > 1)$, or written out more explicitly, the probability of the event

$$A = \{\, (x, y) : 0 < x < 1 \text{ and } 0 < y < 1 \text{ and } x + y > 1 \,\}$$

We have to integrate over the set A. How do we write that as an iterated integral?

Suppose we decide to integrate over y first and x second. In the first integral we keep x fixed, and consider y the variable. What are the limits of integration for y? Well, y must satisfy the inequalities $0 < y < 1$ and $1 < x + y$. Rewrite the latter as $1 - x < y$. Since $1 - x$ is always greater than zero, the inequality $0 < y$ plays no role, and we see that the interval over which we integrate y is $1 - x < y < 1$.

Now we need to find the limits of integration of x. The question is whether the interval over which we integrate is $0 < x < 1$ or whether there is some other restriction limiting us to a subinterval. What decides the question is whether it is always possible to satisfy $1 - x < y < 1$, that is, whether we always have $1 - x < 1$. Since we do, we see that $0 < x < 1$ is correct and

$$P(A) = \int_0^1 \int_{1-x}^1 f(x, y) \, dy \, dx$$

The inner integral is

$$\int_{1-x}^1 (x + y) \, dy = xy + \frac{y^2}{2} \bigg|_{1-x}^1 = \left(x + \frac{1}{2} \right) - \left(x(1 - x) + \frac{(1 - x)^2}{2} \right) = x + \frac{x^2}{2}$$

So the outer integral is

$$\int_0^1 \left(x + \frac{x^2}{2} \right) dx = \frac{x^2}{2} + \frac{x^3}{6} \bigg|_0^1 = \frac{2}{3}$$

In more complicated situations, finding the limits of integration can be much trickier. Fortunately, there is not much use for this kind of trickery in probability and statistics. In principle arbitrarily obnoxious problems of this sort can arise, in practice they don't.

Note that we get an exactly analogous sort of problem calculating probabilities of arbitrary events for discrete random vectors. The iterated integrals become iterated sums and the limits of integration are replaced by limits of summation. But the same principles apply. We don't do an example because the sums are harder to do in practice than integrals.

1.4 The Support of a Random Variable

The *support* of a random variable is the set of points where its density is positive. This is a very simple concept, but there are a few issues about supports that are worthwhile stating explicitly.

If a random variable X has support A, then $P(X \in A) = 1$, because if S is the sample space for the distribution of X

$$\begin{aligned}
1 &= \int_S f_X(x) \, dx \\
&= \int_A f_X(x) \, dx + \int_{A^c} f_X(x) \, dx \\
&= \int_A f_X(x) \, dx \\
&= P(X \in A)
\end{aligned}$$

because f_X is zero on A^c and the integral of zero is zero.

Thus, as long as the only random variables under consideration are X and functions of X it makes no difference whether we consider the sample space to be S (the original sample space) or A (the support of X). We can use this observation in two ways.

- If the support of a random variable is not the whole sample space, we can throw the points where the density is zero out of the sample space without changing any probabilities.
- Conversely, we can always consider a random variable to live in a larger sample space by defining the density to be zero outside of the original sample space.

Simple examples show the idea.

Example 1.4.1.
Consider the $\mathcal{U}(a, b)$ distribution. We can consider the sample space to be the interval (a, b), in which case we write the density

$$f(x) = \frac{1}{b - a}, \qquad a < x < b. \tag{1.23a}$$

On the other hand, we may want to consider the sample space to be the whole real line, in which case we can write the density in two different ways, one using case splitting

$$f(x) = \begin{cases} 0, & x \le a \\ \frac{1}{b-a}, & a < x < b \\ 0, & b \le x \end{cases} \tag{1.23b}$$

and the other using indicator functions

$$f(x) = \frac{1}{b - a} I_{(a,b)}(x), \qquad x \in \mathbb{R}. \tag{1.23c}$$

In most situations you can use whichever form you prefer. Why would anyone every use the more complicated (1.23b) and (1.23c)? There are several reasons. One good reason is that there may be many different random variables, all with different supports, under consideration. If one wants them all to live on the *same* sample space, which may simplify other parts of the problem, then one needs something like (1.23b) or (1.23c). Another reason not so good is mere habit or convention. For example, convention requires that the domain of a c. d. f. be the whole real line. Thus one commonly requires the domain of the matching density to also be the whole real line necessitating something like (1.23b) or (1.23c) if the support is not the whole real line.

1.5 Joint and Marginal Distributions

Strictly speaking, the words "joint" and "marginal" in describing probability distributions are unnecessary. They don't describe kinds of probability distributions. They are just probability distributions. Moreover, the same probability distribution can be either "joint" or "marginal" depending on context. Each is the probability distribution of a set of random variables. When two different sets are under discussion, one a subset of the other, we use "joint" to indicate the superset and "marginal" to indicate the subset. For example, if we are interested in the distribution of the random variables X, Y, and Z and simultaneously interested in the distribution of X and Y, then we call the distribution of the three variables with density $f_{X,Y,Z}$ the "joint" distribution and density, whereas we call the distribution of the two variables X and Y with density $f_{X,Y}$ the "marginal" distribution and density. In a different context, we might also be interested in the distribution of X alone with density f_X. In that context we would call $f_{X,Y}$ the joint density and f_X the marginal density. So whether $f_{X,Y}$ is "joint" or "marginal" depends entirely on context.

What is the relationship between joint and marginal densities? Given $f_{X,Y}$, how do we obtain f_X? (If we can see that, other questions about joint and marginal densities will be obvious by analogy.)

First, note that this is a question about change of variables. Given the "original" random vector (X, Y) what is the distribution of the random variable defined by the transformation

$$X = g(X, Y)?$$

This is not the sort of transformation covered by any of the special-case change of variable theorems (it is certainly not one-to-one, since any two points with the same x value but different y values map to the same point x). However, the general change of variable theorem, Theorem 1.1, does apply (it applies to *any* change of variables).

Theorem 1.1 applied to this case says that

$$P_X(A) = P_{X,Y}(B), \tag{1.24}$$

where

$$\begin{aligned} B &= \{\, (x, y) \in \mathbb{R}^2 : g(x, y) \in A \,\} \\ &= \{\, (x, y) \in \mathbb{R}^2 : x \in A \,\} \\ &= A \times \mathbb{R}. \end{aligned}$$

because $g(x, y) = x$, the notation $A \times \mathbb{R}$ indicating the Cartesian product of A and $\mathbb{R}$, the set of all points (x, y) with $x \in A$ and $y \in \mathbb{R}$.

Now the definition of the density of a continuous (scalar) random variable applied to the left hand side of (1.24) gives us

$$P_X(A) = \int_A f_X(x)\,dx,$$

whereas the definition of the density of a continuous (bivariate) random vector applied to the right hand side of (1.24) gives us

$$\begin{aligned} P_{X,Y}(B) &= \iint_B f_{X,Y}(x, y)\,dx\,dy \\ &= \iint_{A\times\mathbb{R}} f_{X,Y}(x, y)\,dx\,dy \\ &= \int_A \int_{-\infty}^{+\infty} f_{X,Y}(x, y)\,dy\,dx \end{aligned}$$

Thus we can calculate $P(X \in A)$ in two different ways, which must be equal

$$\int_A f_X(x)\,dx = \int_A \int_{-\infty}^{+\infty} f_{X,Y}(x, y)\,dy\,dx$$

Equality of the two expressions for arbitrary events A requires that $f_X(x)$ be the result of the y integral, that is,

$$f_X(x) = \int f_{X,Y}(x, y)\,dy. \tag{1.25}$$

In words we can state this result as follows

> *To go from joint to marginal you integrate (or sum) out the variables you don't want.*

Those readers who are highlighting with a marker, should change colors here and use fire engine red glitter sparkle for this one, something that will *really* stand out. This point is *very* important, and *frequently* botched by students. If you don't remember the slogan above, you will only know that to produce the marginal of X you integrate with respect to x or y. Not knowing which, you will guess wrong half the time. Of course, if you have good calculus awareness you know that

$$\int f_{X,Y}(x, y)\,dx$$

like *any* integral

- *cannot* be a function of the *dummy variable of integration* x, and
- *is* a function of the *free variable* y.

Thus

$$\int f_{X,Y}(x,y)\,dx = \text{some function of } y \text{ only}$$

and hence can only be $f_Y(y)$ and *cannot* be $f_X(x)$. Thus making the mistake of integrating with respect to the wrong variable (or variables) in attempting to produce a marginal is really dumb on two counts: first, you were warned but didn't get it, and, second, it's not only a mistake in probability theory but also a calculus mistake. I do know there are other reasons people can make this mistake, being rushed, failure to read the question, or whatever. I know *someone* will make this mistake, and I apologize in advance for insulting you by calling this a "dumb mistake" if that someone turns out to be you. I'm only trying to give this lecture now, when it may do some good, rather than later, written in red ink all over someone's test paper. (I will, of course, be shocked but very happy if *no one* makes the mistake on the tests.)

Of course, we sum out discrete variables and integrate out continuous ones. So how do we go from $f_{W,X,Y,Z}$ to $f_{X,Z}$? We integrate out the variables we don't want. We are getting rid of W and Y, so

$$f_{X,Z}(x,z) = \iint f_{W,X,Y,Z}(w,x,y,z)\,dw\,dy.$$

If the variables are discrete, the integrals are replaced by sums

$$f_{X,Z}(x,z) = \sum_w \sum_y f_{W,X,Y,Z}(w,x,y,z).$$

In principle, it couldn't be easier. In practice, it may be easy or tricky, depending on how tricky the problem is. Generally, it is easy if there are no worries about domains of integration (and tricky if there are such worries).

Example 1.5.1.
Consider the distribution of Example 1.3.1 with joint density of X and Y given by (1.22). What is the marginal distribution of Y? We find it by integrating out X

$$f_Y(y) = \int f(x,y)\,dx = \int_0^1 (x+y)\,dx = \frac{x^2}{2} + xy\bigg|_0^1 = \left(\frac{1}{2} + y\right)$$

Couldn't be simpler, so long as you don't get confused about which variable you integrate out.

That having been said, it is with some misgivings that I even mention the following examples. If you are having trouble with joint and marginal distributions, don't look at them yet! They are tricky examples that very rarely arise. If you never understand the following examples, you haven't missed much. If you never understand the preceding example, you are in big trouble.

Example 1.5.2 (Uniform Distribution on a Triangle).
Consider the uniform distribution on the triangle with corners $(0, 0)$, $(1, 0)$, and $(0, 1)$ with density

$$f(x, y) = 2, \qquad 0 < x \text{ and } 0 < y \text{ and } x + y < 1$$

What is the marginal distribution of X? To get that we integrate out Y. But the fact that the support of the distribution is not rectangular with sides parallel to the axes means we must take care about limits of integration.

When integrating out y we consider x fixed at one of its possible values. What are the possible values? Clearly $x > 0$ is required. Also we must have $x < 1 - y$. This inequality is least restrictive when we take $y = 0$. So the range of the random variable X is $0 < x < 1$.

For x fixed at a value in this range, what is the allowed range of y? By symmetry, the analysis is the same as we did for x. We must have $0 < y < 1-x$, but now we are considering x fixed. So we stop here. Those are the limits. Thus

$$f_X(x) = \int_0^{1-x} f(x, y)\, dy = \int_0^{1-x} 2\, dy = 2y\Big|_0^{1-x} = 2(1 - x), \qquad 0 < x < 1.$$

Note that the marginal is *not* uniform, although the joint *is* uniform!

Example 1.5.3 (The Discrete Analog of Example 1.5.2).
We get very similar behavior in the discrete analog of Example 1.5.2. Consider the uniform distribution on the set

$$S_n = \{\, (x, y) \in \mathbb{Z}^2 : 1 \le x \le y \le n \,\}$$

for some positive integer n (the symbol $\mathbb{Z}$ denotes the set of integers, so $\mathbb{Z}^2$ is the set of points in $\mathbb{R}^2$ with integer coordinates).

Of course the density of the uniform distribution is constant

$$f(x, y) = \frac{1}{\operatorname{card}(S_n)}, \qquad (x, y) \in S_n.$$

We only have to count the points in S_n to figure out what it is.

We do the count in two bits. There are n points of the form (i, i) for $i = 1$, $\ldots$, n, and there are $\binom{n}{2}$ points of the form (i, j) with $1 \le i < j \le n$. Hence

$$\operatorname{card}(S_n) = n + \binom{n}{2} = n + \frac{n(n-1)}{2} = \frac{n(n+1)}{2}$$

Now in order to have a problem we need a question, which we take to be the same as in the preceding example: what is the marginal of X? To find that we sum out y

$$f_X(x) = \sum_{y=x}^{n} f(x, y) = \frac{2}{n(n+1)} \sum_{y=x}^{n} 1 = \frac{2(n - x + 1)}{n(n+1)}$$

because there are $n - x + 1$ integers between x and n (including both ends).

1.6 Multivariable Change of Variables

1.6.1 The General and Discrete Cases

This section is very short. There is nothing in the general change of variable theorem (Theorem 1.1 about dimension. It applies to all problems, scalar, vector, or whatever.

Similarly, there is nothing in the specializations of the general theorem to the discrete case (Theorems 1.2 and 1.3) about dimension. These too apply to all problems, scalar, vector, or whatever.

1.6.2 Continuous Random Vectors

Derivatives of Vector Functions

But Theorem 1.4 obviously doesn't apply to the vector case, at least not unless it is made clear what the notation $|h'(y)|$ in (1.19) might mean when h is a vector-valued function of a vector variable. For future reference (to be used next semester) we develop the general case in which the dimensions of the domain and codomain are allowed to be different, although we only want the case where they are the same right now.

Let $\mathbf{g}$ be a function that maps n-dimensional vectors to m-dimensional vectors (maps $\mathbb{R}^n$ to $\mathbb{R}^m$). If we write $\mathbf{y} = \mathbf{g}(\mathbf{x})$, this means $\mathbf{y}$ is m-dimensional and $\mathbf{x}$ is n-dimensional. If you prefer to think in terms of many scalar variables instead of vectors, there are really m functions, one for each component of $\mathbf{y}$

$$y_i = g_i(x_1, \ldots, x_n), \qquad i = 1, \ldots, m.$$

So $\mathbf{g}(\mathbf{x})$ really denotes a vector of functions

$$\mathbf{g}(\mathbf{x}) = \begin{pmatrix} g_1(\mathbf{x}) \\ \vdots \\ g_m(\mathbf{x}) \end{pmatrix}$$

which, if you want to write the functions as having n scalar arguments rather than just one vector argument, can also be written

$$\mathbf{g}(\mathbf{x}) = \begin{pmatrix} g_1(x_1, \ldots, x_n) \\ \vdots \\ g_m(x_1, \ldots, x_n) \end{pmatrix}$$

Vector notation is very compact! A few symbols say a lot.

The derivative of the function $\mathbf{g}$ at the point $\mathbf{x}$ (assuming it exists) is the matrix of partial derivatives. It is written $\nabla \mathbf{g}(\mathbf{x})$ and pronounced "del g of x." Throughout this section we will also write it as the single letter $\mathbf{G}$. So

$$\mathbf{G} = \nabla \mathbf{g}(\mathbf{x})$$

is the matrix with elements

$$g_{ij} = \frac{\partial g_i(\mathbf{x})}{\partial x_j}$$

Note that if $\mathbf{g}$ maps n-dimensional vectors to m-dimensional vectors, then it is an $m \times n$ matrix (rather than the $n \times m$). The reason for this choice will become apparent eventually, but not right now.

Example 1.6.1.
Suppose we are interested in the map from 3-dimensional space to 2-dimensional space defined by

$$\begin{aligned} u &= \frac{x}{\sqrt{x^2 + y^2 + z^2}} \\ v &= \frac{y}{\sqrt{x^2 + y^2 + z^2}} \end{aligned}$$

where the 3-dimensional vectors are (x, y, z) and the 2-dimensional vectors (u, v). We can write the derivative matrix as

$$\mathbf{G} = \begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} & \frac{\partial u}{\partial z} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} & \frac{\partial v}{\partial z} \end{pmatrix}$$

This is sometimes written in calculus books as

$$\mathbf{G} = \frac{\partial(u, v)}{\partial(x, y, z)}$$

a notation Lindgren uses in Section 12.1 in his discussion of Jacobians. This notation has never appealed to me. I find it confusing and will avoid it.

Calculating these partial derivatives, we get

$$\begin{aligned} \frac{\partial u}{\partial x} &= (x^2 + y^2 + z^2)^{-1/2} - \frac{1}{2} x (x^2 + y^2 + z^2)^{-3/2} 2x \\ &= \frac{y^2 + z^2}{r^3} \end{aligned}$$

(where we have introduced the notation $r = \sqrt{x^2 + y^2 + z^2}$),

$$\begin{aligned} \frac{\partial u}{\partial y} &= -\frac{1}{2} x (x^2 + y^2 + z^2)^{-3/2} 2y \\ &= -\frac{xy}{r^3} \end{aligned}$$

and so forth (all the other partial derivatives have the same form with different letters), so

$$\nabla \mathbf{g}(x, y, z) = \frac{1}{r^3} \begin{pmatrix} y^2 + z^2 & -xy & -xz \\ -xy & x^2 + z^2 & -yz \end{pmatrix} \tag{1.26}$$

To be careful, we should point out that the function $\mathbf{g}$ is undefined when its argument is zero, but it exists and is differentiable with derivative (1.26) everywhere else.

Note that the derivative matrix is 2×3 as required in mapping 3-dimensional vectors to 2-dimensional vectors.

Invertible Transformations

A multivariate change of variables $\mathbf{h}$ cannot be invertible unless it maps between spaces of the same dimension, that is, from $\mathbb{R}^n$ to $\mathbb{R}^n$ for some n. The determinant of its derivative matrix is called the *Jacobian* of the mapping, denoted

$$J(\mathbf{x}) = \det\bigl(\nabla \mathbf{h}(\mathbf{x})\bigr).$$

(In an alternative terminology, some people call the derivative matrix $\nabla \mathbf{h}(\mathbf{x})$ the *Jacobian matrix* and its determinant the *Jacobian determinant*, but "Jacobian" used as a noun rather than an adjective usually means the determinant.)

The Jacobian appears in the change of variable theorem for multiple integrals.

Theorem 1.5 (Change of Variables in Integration). *Suppose that* $\mathbf{h}$ *is an invertible, continuously differentiable mapping with nonzero Jacobian defined on an open subset of* $\mathbb{R}^n$*, and suppose that* A *is a region contained in the domain of* $\mathbf{h}$ *and that* f *is an integrable function defined on* $\mathbf{h}(A)$*, then*

$$\int_{\mathbf{h}(A)} f(\mathbf{x})\, d\mathbf{x} = \int_A f[\mathbf{h}(\mathbf{y})] \cdot |J(\mathbf{y})|\, d\mathbf{y},$$

where J *is the Jacobian of* $\mathbf{h}$*.*

The notation $\mathbf{h}(A)$ means the image of the region A under the mapping $\mathbf{h}$, that is

$$\mathbf{h}(A) = \{\, \mathbf{h}(\mathbf{x}) : \mathbf{x} \in A \,\}.$$

Corollary 1.6 (Change of Variables for Densities). *Suppose that* $\mathbf{g}$ *is an invertible mapping defined on an open subset of* $\mathbb{R}^n$ *containing the support of a continuous random vector* $\mathbf{X}$ *having probability density* $f_{\mathbf{X}}$*, and suppose that* $\mathbf{h} = \mathbf{g}^{-1}$ *is continuously differentiable with nonzero Jacobian* J*. Then the random vector* $\mathbf{Y} = \mathbf{g}(\mathbf{X})$ *has probability density*

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}[\mathbf{h}(\mathbf{y})] \cdot |J(\mathbf{y})| \tag{1.27}$$

If we plug the definition of the Jacobian into (1.27) we get

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}[\mathbf{h}(\mathbf{y})] \cdot \bigl|\det\bigl(\nabla \mathbf{h}(\mathbf{y})\bigr)\bigr|.$$

Note that the univariate change-of-variable formula

$$f_Y(y) = f_X[h(y)] \cdot |h'(y)|.$$

is a special case.

Proof. The general change of variable theorem (Theorem 1.1) says

$$P_{\mathbf{Y}}(A) = P_{\mathbf{X}}(B) \tag{1.28}$$

where

$$B = \{ \mathbf{x} \in S : \mathbf{g}(\mathbf{x}) \in A \}$$

where S is the sample space of the random vector $\mathbf{X}$, which we may take to be the open subset of $\mathbb{R}^n$ on which $\mathbf{g}$ is defined. Because $\mathbf{g}$ is invertible, we have the relationship between A and B

$$\begin{aligned} B &= \mathbf{h}(A) \\ A &= \mathbf{g}(B) \end{aligned}$$

Rewriting (1.28) using the definition of measures in terms of densities gives

$$\int_A f_{\mathbf{Y}}(\mathbf{y})\, d\mathbf{y} = \int_B f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} = \int_{\mathbf{h}(A)} f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} \tag{1.29}$$

Now applying Theorem 1.5 to the right hand side gives

$$\int_A f_{\mathbf{Y}}(\mathbf{y})\, d\mathbf{y} = \int_A f_{\mathbf{X}}[\mathbf{h}(\mathbf{y})] \cdot |J(\mathbf{y})|\, d\mathbf{y}.$$

This can be true for all sets A only if the integrands are equal, which is the assertion of the theorem. □

Calculating determinants is difficult if n is large. However, we will usually only need the bivariate case

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

Example 1.6.2.
Suppose f is the density on $\mathbb{R}^2$ defined by

$$f(x, y) = \frac{1}{2\pi} \exp\left(-\frac{x^2}{2} - \frac{y^2}{2} \right), \qquad (x, y) \in \mathbb{R}^2.$$

Find the joint density of the variables

$$\begin{aligned} U &= X \\ V &= Y/X \end{aligned}$$

(This transformation is undefined when $X = 0$, but that event occurs with probability zero and may be ignored. We can redefine the sample space to exclude the y-axis without changing any probabilities).

The inverse transformation is

$$\begin{aligned} X &= U \\ Y &= UV \end{aligned}$$

This transformation has derivative

$$\begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ v & u \end{pmatrix}$$

and Jacobian $1 \cdot u - v \cdot 0 = u$.

Thus the joint density of U and V is

$$\begin{aligned} g(u, v) &= \frac{1}{2\pi} \exp\left(-\frac{u^2}{2} - \frac{(uv)^2}{2}\right) \cdot |u| \\ &= \frac{|u|}{2\pi} \exp\left(-\frac{u^2(1 + v^2)}{2}\right) \end{aligned}$$

As another example of the multivariate change-of-variable formula we give a correct proof of the *convolution formula* (Theorem 23 of Chapter 4 in Lindgren)[1]

Theorem 1.7 (Convolution). *If X and Y are independent continuous real-valued random variables with densities f_X and f_Y, then $X + Y$ has density*

$$f_{X+Y}(z) = \int f_X(z - y) f_Y(y) \, dy. \tag{1.30}$$

This is called the *convolution formula*, and the function f_{X+Y} is called the *convolution* of the functions f_X and f_Y.

Proof. Consider the change of variables

$$\begin{aligned} u &= x + y \\ v &= y \end{aligned}$$

(this is the mapping $\mathbf{g}$ in the corollary, which gives the new variables in terms of the old) having inverse mapping

$$\begin{aligned} x &= u - v \\ y &= v \end{aligned}$$

(this is the mapping $\mathbf{h}$ in the corollary, which gives the old variables in terms of the new). The Jacobian is

$$J(u, v) = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 1 & -1 \\ 0 & 1 \end{vmatrix} = 1$$

[1] What's wrong with Lindgren's proof is that he differentiates under the integral sign without any justification. Every time Lindgren uses this differentiation under the integral sign trick, the same problem arises. The right way to prove all such theorems is to use the multivariate change of variable formula.

The joint density of X and Y is $f_X(x)f_Y(y)$ by independence. By the change-of-variable formula, the joint density of U and V is

$$\begin{aligned} f_{U,V}(u,v) &= f_{X,Y}(u-v,v)|J(u,v)| \\ &= f_X(u-v)f_Y(v) \end{aligned}$$

We find the marginal of U by integrating out V

$$f_U(u) = \int f_X(u-v)f_Y(v)\,dv$$

which is the convolution formula. □

Noninvertible Transformations

When a change of variable $\mathbf{Y} = \mathbf{g}(\mathbf{X})$ is not invertible, things are much more complicated, except in one special case, which is covered in this section. Of course, the general change of variable theorem (Theorem 1.1) always applies, but is hard to use.

The special case we are interested in is exemplified by the univariate change of variables

$$\mathbb{R} \xrightarrow{g} [0,\infty)$$

defined by

$$g(x) = x^2, \qquad x \in \mathbb{R}^2. \tag{1.31}$$

This function is not invertible, because it is not one-to-one, but it has two "sort of" inverses, defined by

$$h_+(y) = \sqrt{y}, \qquad y \geq 0. \tag{1.32a}$$

and

$$h_-(y) = -\sqrt{y}, \qquad y \geq 0. \tag{1.32b}$$

Our first task is to make this notion of a "sort of" inverse mathematically precise, and the second is to use it to get a change of variable theorem. In aid of this, let us take a closer look at the notion of inverse functions. Two functions g and h are inverses if, first, they map between the same two sets but in opposite directions

$$\begin{gathered} S \xrightarrow{g} T \\ S \xleftarrow{h} T \end{gathered}$$

and, second, if they "undo" each other's actions, that is,

$$h[g(x)] = x, \qquad x \in S \tag{1.33a}$$

and

$$g[h(y)] = y, \qquad y \in T. \tag{1.33b}$$

Now we want to separate these two properties. We say h is a *left inverse* of g if (1.33a) holds and a *right inverse* of g if (1.33b) holds. Another name for *right inverse* is *section.* It turns out that the important property for change of variable theorems is the right inverse property (1.33b), for example, the function g defined by (1.31) has two right inverses defined by (1.32a) and (1.32b).

The next concept we need to learn in order to state the theorem in this section is "partition." A *partition* of a set S is a family of sets $\{A_i : i \in I\}$ that are disjoint and cover S, that is,

$$A_i \cap A_j = \varnothing, \qquad i \in I,\ j \in I, \text{ and } i \neq j$$

and

$$\bigcup_{i \in I} A_i = S.$$

The last concept we need to learn, or more precisely relearn, is the notion of the support of a random variable. This should have been, perhaps, run into Section 1.4, but too late now. A more general notion of the support of a random variable is the following. An event A is a (not the) *support* of a random variable X if $P(X \in A) = 1$. The support defined Section 1.4 is a support under the new definition, but not the only one. For example, if X is a continuous random variable, we can throw out any single point, any finite set of points, even a countable set of points, because any such set has probability zero. We will see why this more general definition is important in the examples.

These three new concepts taken care of, we are now ready to state the theorem.

Theorem 1.8. *Suppose* $\mathbf{g} : U \to V$ *is a mapping, where* U *and* V *are open subsets of* $\mathbb{R}^n$*, and* U *is a support of a continuous random variable* $\mathbf{X}$ *having probability density* $f_{\mathbf{X}}$*. Suppose that* $\mathbf{h}_i$*,* $i \in I$ *are continuously differentiable sections (right inverses) of* $\mathbf{g}$ *with nonzero Jacobians* $J_i = \det(\nabla \mathbf{h}_i)$*, and suppose the sets* $\mathbf{h}_i(V)$*,* $i \in I$ *form a partition of* U*. Then the random vector* $\mathbf{Y} = \mathbf{g}(\mathbf{X})$ *has probability density*

$$f_{\mathbf{Y}}(\mathbf{y}) = \sum_{i \in I} f_{\mathbf{X}}[\mathbf{h}_i(\mathbf{y})] \cdot |J_i(\mathbf{y})| \tag{1.34}$$

Proof. The proof starts just like the proof of Theorem 1.6, in particular, we still have

$$P_{\mathbf{Y}}(A) = P_{\mathbf{X}}(B)$$

where

$$B = \{\, \mathbf{x} \in U : \mathbf{g}(\mathbf{x}) \in A \,\}$$

Now $\mathbf{g}$ is not invertible, but the sets $\mathbf{h}_i(A)$ form a partition of B. Hence we have

$$\begin{aligned}
\int_A f_{\mathbf{Y}}(\mathbf{y})\,d\mathbf{y} &= \int_B f_{\mathbf{X}}(\mathbf{x})\,d\mathbf{x} \\
&= \sum_{i\in I} \int_{\mathbf{h}_i(A)} f_{\mathbf{X}}(\mathbf{x})\,d\mathbf{x} \\
&= \sum_{i\in I} \int_A f_{\mathbf{X}}[\mathbf{h}_i(\mathbf{y})] \cdot |J_i(\mathbf{y})|\,d\mathbf{y}.
\end{aligned}$$

This can be true for all sets A only if the integrands are equal, which is the assertion of the theorem. □

Example 1.6.3.
Suppose X is a random variable with density

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad x \in \mathbb{R}$$

(that this is a probability density will be proved in Chapter 6 in Lindgren), and suppose $Y = X^2$. What is the density of Y?

In order to apply the theorem, we need to delete the point zero from the sample space of X, then the transformation

$$(-\infty, 0) \cup (0, +\infty) \xrightarrow{g} (0, +\infty)$$

defined by $g(x) = x^2$ has the two sections (right inverses)

$$(-\infty, 0) \xleftarrow{h_-} (0, +\infty)$$

and

$$(0, +\infty) \xleftarrow{h_+} (0, +\infty)$$

defined by $h_-(y) = -\sqrt{y}$ and $h_+(y) = +\sqrt{y}$. And the ranges of the sections do indeed form a partition of the domain of g.

The sections have derivatives

$$\begin{aligned}
h'_-(y) &= -\frac{1}{2} y^{-1/2} \\
h'_+(y) &= +\frac{1}{2} y^{-1/2}
\end{aligned}$$

and applying the theorem gives

$$\begin{aligned}
f_Y(y) &= f_X(\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} + f_X(-\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} \\
&= \frac{1}{\sqrt{y}} f_X(\sqrt{y}) \\
&= \frac{1}{\sqrt{2\pi y}} e^{-y/2}, \qquad y > 0.
\end{aligned}$$

because f_X happens to be a symmetric about zero, that is, $f_X(x) = f_X(-x)$.

Note that it is just as well we deleted the point zero at the beginning, because the resulting density is undefined at zero anyway.

It is worthwhile stating a couple of intermediate results of the preceding example in a corollary.

Corollary 1.9. *Suppose X is a continuous random scalar with density f_X, then $Y = X^2$ has density*

$$f_Y(y) = \frac{1}{2\sqrt{y}}\bigl[f_X(\sqrt{y}) + f_X(-\sqrt{y})\bigr], \qquad y > 0.$$

Moreover, if f_X is symmetric about zero, then

$$f_Y(y) = \frac{1}{\sqrt{y}} f_X(\sqrt{y}), \qquad y > 0.$$

Chapter 2

Expectation

2.1 Introduction

Expectation and *probability* are equally important concepts. An important educational objective of this course is that students become "ambidextrous" in reasoning with these two concepts, able to reason equally well with either.

Thus we don't want to think of expectation as a derived concept—something that is calculated from probabilities. We want the expectation concept to stand on its own. Thus it should have the same sort of treatment we gave probability. In particular, we need to have the connection between expectation and the law of large numbers (the analog of Section 2.2 in Lindgren) and axioms for expectation (the analog of Section 2.4 in Lindgren).

Suppose you are asked to pick a single number to stand in for a random variable. Of course, the random variable, when eventually observed, will probably differ from whatever number you pick (if the random variable is continuous it will match whatever number you pick with probability zero). But you still have to pick a number. Which number is best?

The *expectation* (also called *expected value*) of a real-valued random variable, if it exists, is one answer to this problem. It is the single number that a rational person "should" expect as the value of the random variable when it is observed. Expectation is most easily understood in economic contexts. If the random variable in question is the value of an investment or other uncertain quantity, the expectation is the "fair price" of the investment, the maximum amount a rational person is willing to pay to pay for the investment.

The notion of expectation of a non-monetary random variable is less clear, but can be forced into the monetary context by an imaginary device. Suppose the random variable in question is the weight of a student drawn at random from a list of all students at the university. Imagine you will be paid a dollar per pound of that student's weight. How much would you be willing to pay to "invest" in this opportunity? That amount is (or should be) the expected value of the student's weight.

2.2 The Law of Large Numbers

What Lindgren describes in his Section 2.2 is not the general form of the law of large numbers. It wasn't possible to explain the general form then, because the general form involves the concept of expectation.

Suppose X_1, X_2, $\ldots$ is an independent and identically distributed sequence of random variables. This means these variables are the same function X (a random variable is a function on the sample space) applied to independent repetitions of the same random process. The average of the first n variables is denoted

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i. \tag{2.1}$$

The general form of the law of large numbers says the average converges to the expectation $E(X) = E(X_i)$, for all i. In symbols

$$\overline{X}_n \to E(X), \qquad \text{as } n \to \infty. \tag{2.2}$$

It is not clear at this point, just what the arrow on the left in (2.2) is supposed to mean. Vaguely it means something like convergence to a limit, but $\overline{X}_n$ is a random variable (any function of random variables is a random variable) and $E(X)$ is a constant (all expectations are numbers, that is, constants), and we have no mathematical definition of what it means for a sequence of random variables to converge to a number. For now we will make do with the sloppy interpretation that (2.2) says that $\overline{X}_n$ gets closer and closer to $E(X)$ as n goes to infinity, in some sense that will be made clearer later (Chapter 5 in Lindgren and Chapter 4 of these notes).

2.3 Basic Properties

2.3.1 Axioms for Expectation (Part I)

In this section, we begin our discussion of the formal mathematical properties of expectation. As in many other areas of mathematics, we start with fundamental properties that are not proved. These unproved (just assumed) properties are traditionally called "axioms." The axioms for expectation are the mathematical definition of the expectation concept. Anything that satisfies the axioms is an instance of mathematical expectation. Anything that doesn't satisfy the axioms isn't. Every other property of expectation can be derived from these axioms (although we will not give a completely rigorous derivation of all the properties we will mention, some derivations being too complicated for this course).

The reason for the "Part I" in the section heading is that we will not cover all the axioms here. Two more esoteric axioms will be discussed later (Section 2.5.4 of these notes).

Expectation is in some respects much a much simpler concept that probability and in other respects a bit more complicated. The issue that makes expectation more complicated is that not all real-valued random variables have expectations. The set of real valued random variables that have expectation is denoted L^1 or sometimes $L^1(P)$ where P is the probability measure associated with the expectation, the letter "L" here being chosen in honor of the French mathematician Henri Lebesgue (1875–1941), who invented the general definition of integration used in advanced probability theory (p. 67 of these notes), the digit "1" being chosen for a reason to be explained later. The connection between integration and expectation will also be explained later.

An *expectation operator* is a function that assigns to each random variable $X \in L^1$ a real number $E(X)$ called the *expectation* or *expected value* of X. Every expectation operator satisfies the following axioms.

Axiom E1 (Additivity). *If X and Y are in L^1, then $X + Y$ is also in L^1, and*

$$E(X + Y) = E(X) + E(Y).$$

Axiom E2 (Homogeneity). *If X is in L^1 and a is a real number, then aX is also in L^1, and*

$$E(aX) = aE(X).$$

These properties agree with either of the informal intuitions about expectations. Prices are additive and homogeneous. The price of a gallon of milk and a box of cereal is the sum of the prices of the two items separately. Also the price of three boxes of cereal is three times the price of one box. (The notion of expectation as fair price doesn't allow for volume discounts.)

Axiom E3 (Positivity). *If X is in L^1, then*

$$X \geq 0 \textit{ implies } E(X) \geq 0.$$

The expression $X \geq 0$, written out in more detail, means

$$X(s) \geq 0, \qquad s \in S,$$

where S is the sample space. That is, X is always nonnegative.

This axiom corresponds to intuition about prices, since goods always have nonnegative value and prices are also nonnegative.

Axiom E4 (Norm). *The constant random variable I that always has the value one is in L^1, and*

$$E(I) = 1. \tag{2.3}$$

Equation (2.3) is more commonly written

$$E(1) = 1, \tag{2.4}$$

and we will henceforth write it this way. This is something of an abuse of notation. The symbol "1" on the right hand side is the number one, but the symbol "1" on the left hand side must be a random variable (because the argument of an expectation operator is a random variable), hence a function on the sample space. So in order to understand (2.4) we must agree to interpret a number in a context that requires a random variable as the constant random variable always equal to that number.

2.3.2 Derived Basic Properties

Theorem 2.1 (Linearity). *If X and Y are in L^1, and a and b are real numbers then $aX + bY$ is also in L^1, and*

$$E(aX + bY) = aE(X) + bE(Y). \tag{2.5}$$

Proof of Theorem 2.1. The existence part of Axiom E2 implies $aX \in L^1$ and $bY \in L^1$. Then the existence part of Axiom E1 implies $aX + bY \in L^1$.

Then Axiom E1 implies

$$E(aX + bY) = E(aX) + E(bY)$$

and Axiom E2 applied to each term on the right hand side implies (2.5). □

Corollary 2.2 (Linear Functions). *If X is in L^1, and $Y = a + bX$, where a and b are real numbers, then Y is also in L^1, and*

$$E(Y) = a + bE(X). \tag{2.6}$$

Proof. If we let X in Theorem 2.1 be the constant random variable 1, then (2.5) becomes

$$E(a \cdot 1 + bY) = aE(1) + bE(Y),$$

and applying Axiom E4 to the $E(1)$ on the right hand side gives

$$E(a + bY) = E(a \cdot 1 + bY) = a \cdot 1 + bE(Y) = a + bE(Y),$$

and reading from end to end gives

$$E(a + bY) = a + bE(Y), \tag{2.7}$$

which except for notational differences is what was to be proved. □

If the last sentence of the proof leaves you unsatisfied, you need to think a bit more about "mathematics is invariant under changes of notation" (Problem 2-1).

Example 2.3.1 (Fahrenheit to Centigrade).
Corollary 2.2 arises whenever there is a change of units of measurement. All changes of units are linear functions. Most are purely multiplicative, 2.54 centimeters to the inch and so forth, but a few are the more general kind of linear transformation described in the corollary. An example is the change of temperature units from Fahrenheit to centigrade degrees. If X is a random variable having units of degrees Fahrenheit and Y is the a random variable that is the same measurement as X but in units of degrees centigrade, the relation between the two is

$$Y = \frac{5}{9}(X - 32).$$

The corollary then implies

$$E(Y) = \frac{5}{9}[E(X) - 32],$$

that is, the expectations transform the same way as the variables under a change of units. Thus, if the expected daily high temperature in January in Minneapolis is 23 °F, then this expected value is also −5 °C. Expectations behave sensibly under changes of units of measurement.

Theorem 2.3 (Linearity). *If X_1, …, X_n are in L^1, and a_1, …, a_n are real numbers then $a_1 X_1 + \cdots a_n X_n$ is also in L^1, and*

$$E(a_1 X_1 + \cdots + a_n X_n) = a_1 E(X_1) + \cdots + a_n E(X_n).$$

Theorem 2.1 is the case $n = 2$ of Theorem 2.3, so the latter is a generalization of the former. That's why both have the same name. (If this isn't obvious, you need to think more about "mathematics is invariant under changes of notation." The two theorems use different notation, a_1 and a_2 instead of a and b and X_1 and X_2 instead of X and Y, but they assert the same property of expectation.)

Proof of Theorem 2.3. The proof is by mathematical induction. The theorem is true for the case $n = 2$ (Theorem 2.1). Thus we only need to show that the truth of the theorem for the case $n = k$ implies the truth of the theorem for the case $n = k + 1$. Apply Axiom E1 to the case $n = k + 1$ giving

$$E(a_1 X_1 + \cdots + a_{k+1} X_{k+1}) = E(a_1 X_1 + \cdots + a_k X_k) + E(a_{k+1} X_{k+1}).$$

Then apply Axiom E2 to the second term on the right hand side giving

$$E(a_1 X_1 + \cdots + a_{k+1} X_{k+1}) = E(a_1 X_1 + \cdots + a_k X_k) + a_{k+1} E(X_{k+1}).$$

Now the $n = k$ case of the theorem applied to the first term on the right hand side gives the $n = k + 1$ case of the theorem. □

Corollary 2.4 (Additivity). *If X_1, …, X_n are in L^1, then $X_1 + \cdots X_n$ is also in L^1, and*

$$E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n).$$

This theorem is used so often that it seems worth restating in words to help you remember.

> *The expectation of a sum is the sum of the expectations.*

Note that Axiom E1 is the case $n = 2$, so the property asserted by this theorem is a generalization. It can be derived from Axiom E1 by mathematical induction or from Theorem 2.3 (Problem 2-2).

Corollary 2.5 (Subtraction). *If X and Y are in L^1, then $X - Y$ is also in L^1, and*

$$E(X - Y) = E(X) - E(Y).$$

Corollary 2.6 (Minus Signs). *If X is in L^1, then $-X$ is also in L^1, and*

$$E(-X) = -E(X).$$

These two properties are obvious consequences of linearity (Problems 2-3 and 2-4).

Corollary 2.7 (Constants). *Every constant random variable is in L^1, and*

$$E(a) = a.$$

This uses the convention we introduced in connection with (2.4). The symbol "a" on the right hand side represents a real number, but the symbol "a" on the left hand side represents the constant random variable always equal to that number. The proof is left as an exercise (Problem 2-6).

Note that a special case of Corollary 2.7 is $E(0) = 0$.

Theorem 2.8 (Monotonicity). *If X and Y are in L^1, then*

$$X \le Y \text{ implies } E(X) \le E(Y).$$

The expression $X \le Y$, written out in full, means

$$X(s) \le Y(s), \qquad s \in S,$$

where S is the sample space. That is, X is always less than or equal to Y.

Note that the positivity axiom (E3) is the special case $X = 0$ of this theorem. Thus this theorem is a generalization of that axiom.

This theorem is fairly easily derived from the positivity axiom (E3) and the Theorem 2.5 (Problem 2-7).

All of the theorems in this section and the axioms in the preceding section are exceedingly important and will be used continually throughout the course. You should have them all at your fingertips. Failure to recall the appropriate axiom or theorem when required will mean failure to do many problems. It is not necessary to memorize all the axioms and theorems. You can look them up when needed. But you do need to have *some* idea what each axiom and theorem is about so you will know that there is something to look up. After all, you can't browse the entire course notes each time you use something.

Axiom E3 and Theorem 2.8 are important in what I call "sanity checks." Suppose you are given a description of a random variable X and are told to calculate its expectation. One of the properties given is $X \ge 3$, but your answer is $E(X) = 2$. This is obviously wrong. It violates Theorem 2.8. You must have made a mistake somewhere! Sanity checks like this can save you from many mistakes if you only remember to make them. A problem isn't done when you obtain an answer. You should also take a few seconds to check that your answer isn't obviously ridiculous.

2.3.3 Important Non-Properties

What's a non-property? It's a property that students often use but isn't true. Students are mislead by analogy or guessing. Thus we stress that the following are not true in general (although they are sometimes true in some special cases).

The Multiplicativity Non-Property

One might suppose that there is a property analogous to the additivity property, except with multiplication instead of addition

$$E(XY) = E(X)E(Y), \qquad \text{Uncorrelated } X \text{ and } Y \text{ only!} \tag{2.8}$$

As the editorial comment says, this property does *not* hold in general. We will later see that when (2.8) does hold we have a special name for this situation: we say the variables X and Y are *uncorrelated*.

Taking a Function Outside an Expectation

Suppose g is a linear function defined by

$$g(x) = a + bx, \qquad x \in \mathbb{R}, \tag{2.9}$$

where a and b are real numbers. Then

$$E\{g(X)\} = g(E\{X\}), \qquad \text{Linear } g \text{ only!} \tag{2.10}$$

is just Theorem 2.2 stated in different notation. The reason for the editorial comment is that (2.10) does *not* hold for general functions g, only for *linear* functions. Sometime you will be tempted to use (2.10) for a nonlinear function g. Don't! Remember that it is a "non-property."

For example, you may be asked to calculate $E(1/X)$ for some random variable X. The "non-property," if it were true, would allow to take the function outside the expectation and the answer would be $1/E(X)$, but it isn't true, and, in general

$$E\left(\frac{1}{X}\right) \neq \frac{1}{E(X)}$$

There may be a way to do the problem, but the "non-property" isn't it.

2.4 Moments

If k is a positive integer, then the real number

$$\alpha_k = E(X^k) \tag{2.11}$$

is called the *k-th moment* of the random variable X.

If p is a positive real number, then the real number

$$\beta_p = E(|X|^p) \tag{2.12}$$

is called the *p-th absolute moment* of the random variable X.

If k is a positive integer and $\mu = E(X)$, then the real number

$$\mu_k = E\{(X - \mu)^k\} \tag{2.13}$$

is called the *k-th central moment* of the random variable X. (The symbols α, β, and μ are Greek letters. See Appendix A).

Sometimes, to emphasize we are talking about (2.11) rather than one of the other two, we will refer to it as the *ordinary moment*, although, strictly speaking, the "ordinary" is redundant.

That's not the whole story on moments. We can define lots more, but all moments are special cases of one of the two following concepts.

If k is a positive real number and a is any real number, then the real number $E\{(X-a)^k\}$ is called the *k-th moment about the point a* of the random variable X. We introduce no special symbol for this concept. Note that the k-th ordinary moment is the special case $a = 0$ and the k-th central moment is the case $a = \mu$.

If p is a positive real number and a is any real number, then the real number $E\{|X-a|^p\}$ is called the *p-th absolute moment about the point a* of the random variable X. We introduce no special symbol for this concept. Note that the p-th absolute moment is the special case $a = 0$.

2.4.1 First Moments and Means

The preceding section had a lot of notation and definitions, but nothing else. There was nothing there you could use to calculate anything. It seems like a lot to remember. Fortunately, only a few special cases are important. For the most part, we are only interested in p-th moments when p is an integer, and usually a small integer. By far the most important cases are $p = 1$, which is covered in this section, and $p = 2$, which is covered in the following section. We say p-th moments (of any type) with $p = 1$ are *first moments*, with $p = 2$ are *second moments*, and so forth (third, fourth, fifth, ...).

First ordinary moment is just a fancy name for *expectation*. This moment is so important that it has yet another name. The first ordinary moment of a random variable X is also called the *mean* of X. It is commonly denoted by the Greek letter μ, as we did in (2.13). Note that α_1, μ, and $E(X)$ are different notations for the same thing. We will use them all throughout the course.

When there are several random variables under discussion, we denote the mean of each using the same Greek letter μ, but add the variable as a subscript to distinguish them: $\mu_X = E(X)$, $\mu_Y = E(Y)$, and so forth.

Theorem 2.9. *For any random variable in L^1, the first central moment is zero.*

The proof is left as an exercise (Problem 2-9).

This theorem is the first one that allows us to actually calculate a moment of a nonconstant random variable, not a very interesting moment, but it's a start.

Symmetric Random Variables

We say two random variables X and Y *have the same distribution* if

$$E\{g(X)\} = E\{g(Y)\}$$

holds for all real-valued functions g such that the expectations exist and if both expectations exist or neither. In this case we will say that X and Y are *equal in distribution* and use the notation

$$X \overset{\mathcal{D}}{=} Y.$$

This notation is a bit misleading, since it actually says nothing about X and Y themselves, but only about their distributions. What is does imply is any of the following

$$\begin{aligned} P_X &= P_Y \\ F_X &= F_Y \\ f_X &= f_Y \end{aligned}$$

that is, X and Y have the same *probability measure*, the same *distribution function*, or the same *probability density.* What it does *not* imply is anything about the values of X and Y themselves, which like all random variables are functions on the sample space. It may be that $X(\omega)$ is not equal to $Y(\omega)$ for any ω. Nevertheless, the notation is useful.

We say a real-valued random variable X is *symmetric about zero* if X and $-X$ have the same distribution, that is, if

$$X \overset{\mathcal{D}}{=} -X.$$

Note that this is an example of the variables themselves not being equal. Clearly, $X(\omega) \neq -X(\omega)$ unless $X(\omega) = 0$, which may occur with probability zero (will occur with probability zero whenever X is a continuous random variable).

We say a real-valued random variable X is *symmetric about a point a* if $X - a$ is symmetric about zero, that is, if

$$X - a \overset{\mathcal{D}}{=} a - X.$$

The point a is called the *center of symmetry* of X. (Note: Lindgren, definition on p. 94, gives what is at first glance a completely unrelated definition of this concept. The two definitions, his and ours, do in fact define the same concept. See Problem 2-11.)

Some of the most interesting probability models we will meet later involve symmetric random variables, hence the following theorem is very useful.

Theorem 2.10. *Suppose a real-valued random variable X is symmetric about the point a. If the mean of X exists, it is equal to a. Every higher odd integer central moment of X that exists is zero.*

In notation, the two assertions of the theorem are

$$E(X) = \mu = a$$

and

$$\mu_{2k-1} = E\{(X - \mu)^{2k-1}\} = 0, \qquad \text{for any positive integer } k.$$

The proof is left as an exercise (Problem 2-10).

2.4.2 Second Moments and Variances

The preceding section says all that can be said in general about first moments. As we shall now see, second moments are much more complicated.

The most important second moment is the second central moment, which also has a special name. It is called the *variance* and is often denoted σ^2. (The symbol σ is a Greek letter. See Appendix A). We will see the reason for the square presently. We also use the notation $\operatorname{var}(X)$ for the variance of X. So

$$\sigma^2 = \mu_2 = \operatorname{var}(X) = E\{(X - \mu)^2\}.$$

As we did with means, when there are several random variables under discussion, we denote the variance of each using the same Greek letter σ, but add the variable as a subscript to distinguish them: $\sigma_X^2 = \operatorname{var}(X)$, $\sigma_Y^2 = \operatorname{var}(Y)$, and so forth.

Note that variance is just an expectation like any other, the expectation of the random variable $(X - \mu)^2$.

All second moments are related.

Theorem 2.11 (Parallel Axis Theorem). *If X is a random variable with mean μ and variance σ^2, then*

$$E\{(X - a)^2\} = \sigma^2 + (\mu - a)^2$$

Proof. Using the fact

$$(b + c)^2 = b^2 + 2bc + c^2 \tag{2.14}$$

from algebra

$$\begin{aligned} (X - a)^2 &= (X - \mu + \mu - a)^2 \\ &= (X - \mu)^2 + 2(X - \mu)(\mu - a) + (\mu - a)^2 \end{aligned}$$

Taking expectations of both sides and applying linearity of expectation (everything not containing X is nonrandom and so can be pulled out of expectations) gives

$$\begin{aligned} E\{(X - a)^2\} &= E\{(X - \mu)^2\} + 2(\mu - a)E(X - \mu) + (\mu - a)^2 E(1) \\ &= \sigma^2 + 2(\mu - a)\mu_1 + (\mu - a)^2 \end{aligned}$$

By Theorem 2.9, the middle term on the right hand side is zero, and that completes the proof. □

The name of this theorem is rather strange. It is taken from an analogous theorem in physics about moments of inertia. So the name has nothing to do with probability in general and moments (as understood in probability theory rather than physics) in particular, and the theorem is not commonly called by that name. We will use it because Lindgren does, and perhaps because the theorem doesn't have any other widely used name. In fact, since it is so simple, it is often not called a theorem but just a calculation formula or method. Sometimes it is called "completing the square" after the method of that name from high-school algebra, although that name isn't very appropriate either. It is a very simple theorem, just the algebraic identity (2.14), which is related to "completing the square" plus linearity of expectation, which isn't. Whatever it is called, the theorem is exceedingly important, and many important facts are derived from it. I sometimes call it "the most important formula in statistics."

Corollary 2.12. *If X is a random variable having first and second moments, then*

$$\operatorname{var}(X) = E(X^2) - E(X)^2.$$

The proof is left as an exercise (Problem 2-13).

This corollary is an important special case of the parallel axis theorem. It also is frequently used, but not quite as frequently as students want to use it. It should not be used in every problem that involves a variance (maybe in half of them, but not all). We will give a more specific warning against overusing this corollary later.

There are various ways of restating the corollary in symbols, for example

$$\sigma_X^2 = E(X^2) - \mu_X^2,$$

and

$$\mu_2 = \alpha_2 - \alpha_1^2.$$

As always, mathematics is invariant under changes of notation. The important thing is the concepts symbolized rather than the symbols themselves.

The next theorem extends Theorem 2.2 from means to variances.

Theorem 2.13. *Suppose X is a random variable having first and second moments and a and b are real numbers, then*

$$\operatorname{var}(a + bX) = b^2 \operatorname{var}(X). \tag{2.15}$$

Note that the right hand side of (2.15) does not involve the constant part a of the linear transformation $a + bX$. Also note that the b comes out squared. The proof is left as an exercise (Problem 2-15).

Before leaving this section, we want to emphasize an obvious property of variances.

Sanity Check: *Variances are nonnegative.*

This holds by the positivity axiom (E3) because the variance of X is the expectation of the random variable $(X - \mu)^2$, which is nonnegative because squares are nonnegative. We could state this as a theorem, but won't because its main use is as a "sanity check." If you are calculating a variance and don't make any mistakes, then your result must be nonnegative. The only way to get a negative variance is to mess up somewhere. If you are using Corollary 2.12, for example, you can get a negative number as a result of the subtraction, if you have calculated one of the quantities being subtracted incorrectly.

So whenever you finish calculating a variance, check that it is nonnegative. If you get a negative variance, and have time, go back over the problem to try to find your mistake. There's never any question such an answer is wrong.

A more subtile sanity check is that a variance should rarely be zero. We will get to that later.

2.4.3 Standard Deviations and Standardization

Standard Deviations

The nonnegative square root of the variance is called the *standard deviation*. Conversely, the variance is the square of the standard deviation. The symbol commonly used for the standard deviation is σ. That's why the variance is usually denoted σ^2.

As with the mean and variance, we use subscripts to distinguish variables σ_X, σ_Y, and so forth. We also use the notation $\operatorname{sd}(X)$, $\operatorname{sd}(Y)$, and so forth. Note that we always have the relations

$$\begin{aligned} \operatorname{sd}(X) &= \sqrt{\operatorname{var}(X)} \\ \operatorname{var}(X) &= \operatorname{sd}(X)^2 \end{aligned}$$

So whenever you have a variance you get the corresponding standard deviation by taking the square root, and whenever you have a standard deviation you get the corresponding variance by squaring. Note that the square root always is possible because variances are always nonnegative. The σ and σ^2 notations make this obvious: σ^2 is the square of σ (duh!) and σ is the square root of σ^2. The notations $\operatorname{sd}(X)$ and $\operatorname{var}(X)$ don't make their relationship obvious, nor do the names "standard deviation" and "variance" so the relationship must be kept in mind.

Taking the square root of both sides of (2.15) gives the analogous theorem for standard deviations.

Corollary 2.14. *Suppose X is a random variable having first and second moments and a and b are real numbers, then*

$$\operatorname{sd}(a + bX) = |b| \operatorname{sd}(X). \tag{2.16}$$

It might have just occurred to you to ask why anyone would want two such closely related concepts. Won't one do? In fact more than one introductory (freshman level) statistics textbook does just that, speaking only of standard deviations, never of variances. But for theoretical probability and statistics, this will not do. Standard deviations are almost useless for theoretical purposes. The square root introduces nasty complications into simple situations. So for theoretical purposes *variance* is the preferred concept.

In contrast, for all practical purposes *standard deviation* is the preferred concept, as evidenced by the fact that introductory statistics textbooks that choose to use only one of the two concepts invariably choose standard deviation.

The reason has to do with units of measurement and measurement scales. Suppose we have a random variable X whose units of measurement are inches, for example, the height of a student in the class. What are the units of $E(X)$, $\operatorname{var}(X)$, and $\operatorname{sd}(X)$, assuming these quantities exist?

The units of an expectation are the same as the units of the random variable, so the units of $E(X)$ are also inches. Now $\operatorname{var}(X)$ is also just an expectation, the expectation of the random variable $(X - \mu)^2$, so its units are the units of $(X - \mu)^2$, which are obviously inches squared (or square inches, if you prefer). Then obviously, the units of $\operatorname{sd}(X)$ are again inches. Thus X, $E(X)$, and $\operatorname{sd}(X)$ are comparable quantities, all in the same units, whereas $\operatorname{var}(X)$ is *not*. You can't understand what $\operatorname{var}(X)$ tells you about X without taking the square root. It's isn't even in the right units of measurement.

The theoretical emphasis of this course means that we will be primarily interested in variances rather than standard deviations, although we will be interested in standard deviations too. You have to keep in mind which is which.

Standardization

Given a random variable X, there is always a linear transformation $Z = a + bX$, which can be thought of as a change of units of measurement as in Example 2.3.1, that makes the transformed variable Z have mean zero and standard deviation one. This process is called *standardization*.

Theorem 2.15. *If X is a random variable having mean μ and standard deviation σ and $\sigma > 0$, then the random variable*

$$Z = \frac{X - \mu}{\sigma} \tag{2.17}$$

has mean zero and standard deviation one.

Conversely, if Z is a random variable having mean zero and standard deviation one, μ and σ are real numbers, and $\sigma \geq 0$, then the random variable

$$X = \mu + \sigma Z \tag{2.18}$$

has mean μ and standard deviation σ.

The proof is left as an exercise (Problem 2-17).

Standardization (2.17) and its inverse (2.18) are useful in a variety of contexts. We will use them throughout the course.

2.4.4 Mixed Moments and Covariances

When several random variables are involved in the discussion, there are several moments of each type, as we have already discussed. If we have two random variables X and Y, then we also have two (ordinary) first moments μ_X and μ_Y and two second central moments σ^2_X and σ^2_Y, but that is not the whole story. To see why, it is helpful to make a brief digression into the terminology of polynomials.

Polynomials and Monomials

Forget random variables for a second and consider polynomials in two (ordinary) variables x and y. A general polynomial of degree zero is a constant function f defined by

$$f(x, y) = a, \qquad x, y \in \mathbb{R},$$

where a is a constant. A general polynomial of degree one is a linear function f defined by

$$f(x, y) = a + bx + cy, \qquad x, y \in \mathbb{R},$$

where a, b, and c are constants. A general polynomial of degree two is a quadratic function f defined by

$$f(x, y) = a + bx + cy + dx^2 + exy + ky^2, \qquad x, y \in \mathbb{R},$$

where a, b, c, d, e, and k are constants. The point is that we have a new kind of term, the term exy that contains both variables in the polynomial of degree two. In general, we say the degree of a term is the sum of the exponents of all the variables in the term, so x^2 and $xy = x^1 y^1$ are both terms of degree two.

One term of a polynomial is called a *monomial.* The convention that the degree of a monomial is the sum of the exponents of the variables is arbitrary, but it is a useful convention for the following reason. It seems sensible to consider $(x + y)^2$ a quadratic polynomial because it is the square of a linear polynomial, but the identity

$$(x + y)^2 = x^2 + 2xy + y^2$$

shows us that this sort of quadratic polynomial involves the "mixed" monomial xy. The reason why this monomial is said to have degree two rather than degree one will become clearer as we go along.

Mixed Moments

We apply the same sort of thinking to moments. We say $E(XY)$ is a "mixed" second moment if X and Y are two random variables and in general that an expectation of the form

$$E\left(\prod_{i=1}^n X_i^{k_i}\right), \tag{2.19}$$

where X_1, $\ldots$, X_n are n random variables, is a "mixed" K-th moment, where

$$K = \sum_{i=1}^n k_i \tag{2.20}$$

is the sum of the exponents. If you are not familiar with the product notation in (2.19), it is analogous to the summation notation in (2.20). The expression (2.19) can also be written

$$E\bigl(X_1^{k_1} X_2^{k_2} \cdots X_n^{k_n}\bigr)$$

just as (2.20) can be written

$$K = k_1 + k_2 + \cdots + k_n.$$

The general formula (2.19) allows for the possibility that some of the k_i may be zero if we adopt the convention that ($a^0 = 1$ for all real a so, for example $x^0 y^2 z^1 = y^2 z$).

Even more general than (2.19) we allow, just as in the non-mixed case, moments about arbitrary points, so we also say

$$E\left\{ \prod_{i=1}^n (X_i - a_i)^{k_i} \right\}$$

is a K-th moment, where K is again the sum of the exponents (2.20) and a_1, a_2, $\ldots$, a_n are arbitrary real numbers. We say this sort of mixed moment is a *central* moment if it is a moment about the means, that is,

$$E\left\{ \prod_{i=1}^n (X_i - \mu_i)^{k_i} \right\}$$

where

$$\mu_i = E(X_i), \qquad i = 1, \ldots, n.$$

(The convention that we use the random variable as a subscript would require μ_{X_i} here rather than μ_i, but the simplicity of avoiding the extra level of subscripts makes the simpler form preferable.)

Covariance

All of that is a lot of abstract notation and complicated definitions. As in the case of non-mixed moments, by far the most important case, the one we will be concerned with more than all the higher-order moments together, is the second central mixed moment, which has a special name. The *covariance* of two random variables X and Y, written $\operatorname{cov}(X, Y)$, is the second central mixed moment

$$\operatorname{cov}(X, Y) = E\bigl\{(X - \mu_X)(Y - \mu_Y)\bigr\},$$

where, as usual, $\mu_X = E(X)$ and $\mu_Y = E(Y)$.

Note a fact that follows trivially from the definition: a covariance is a symmetric function of its arguments, that is, $\operatorname{cov}(X, Y) = \operatorname{cov}(Y, X)$ for any two random variables X and Y.

Note that *variance* is a special case of *covariance*. When X and Y are the same random variable, we get

$$\operatorname{cov}(X, X) = E\bigl\{(X - \mu_X)^2\bigr\} = \operatorname{var}(X).$$

The covariance of a random variable with itself is its variance. This is one reason why covariance is considered a (mixed) second moment (rather than some sort of first moment). A more important reason arises in the following section.

For some unknown reason, there is no standard Greek-letter notation for covariance. We can always write σ_X^2 instead of $\operatorname{var}(X)$ if we like, but there is no standard analogous notation for covariance. (Lindgren uses the notation $\sigma_{X,Y}$ for $\operatorname{cov}(X, Y)$, but this notation is nonstandard. For one thing, the special case $\sigma_{X,X} = \sigma_X^2$ looks weird. For another, no one who has not had a course using Lindgren as the textbook will recognize $\sigma_{X,Y}$. Hence it is better not to get in the habit of using the notation.)

Variance of a Linear Combination

A very important application of the covariance concept is the second-order analog of the linearity property given in Theorem 2.3. Expressions like the $a_1 X_1 + \cdots + a_n X_n$ occurring in Theorem 2.3 arise so frequently that it is worth having a general term for them. An expression $a_1 x_1 + \cdots a_n x_n$, where the a_i are constants and the x_i are variables is called a *linear combination* of these variables. The same terminology is used when the variables are random. With this terminology defined, the question of interest in this section can be stated: what can we say about variances and covariances of linear combinations?

Theorem 2.16. *If X_1, ..., X_m and Y_1, ..., Y_n are random variables having first and second moments and a_1, ..., a_m and b_1, ..., b_n are constants, then*

$$\operatorname{cov}\left(\sum_{i=1}^m a_i X_i, \sum_{j=1}^n b_j Y_j\right) = \sum_{i=1}^m \sum_{j=1}^n a_i b_j \operatorname{cov}(X_i, Y_j). \tag{2.21}$$

Before we prove this important theorem we will look at some corollaries that are even more important than the theorem itself.

Corollary 2.17. *If X_1, ..., X_n are random variables having first and second moments and a_1, ..., a_n are constants, then*

$$\operatorname{var}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n \sum_{j=1}^n a_i a_j \operatorname{cov}(X_i, X_j). \tag{2.22}$$

Proof. Just take $m = n$, $a_i = b_i$, and $X_i = Y_i$ in the theorem. □

Corollary 2.18. *If X_1, ..., X_m and Y_1, ..., Y_n are random variables having first and second moments, then*

$$\operatorname{cov}\left(\sum_{i=1}^m X_i, \sum_{j=1}^n Y_j\right) = \sum_{i=1}^m \sum_{j=1}^n \operatorname{cov}(X_i, Y_j). \tag{2.23}$$

Proof. Just take $a_i = b_j = 1$ in the theorem. □

Corollary 2.19. *If X_1, …, X_n are random variables having first and second moments, then*

$$\operatorname{var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \sum_{j=1}^n \operatorname{cov}(X_i, X_j). \tag{2.24}$$

Proof. Just take $a_i = 1$ in Corollary 2.17. □

The two corollaries about variances can be rewritten in several ways using the symmetry property of covariances, $\operatorname{cov}(X_i, X_j) = \operatorname{cov}(X_j, X_i)$, and the fact that variance is a special case of covariance, $\operatorname{cov}(X_i, X_i) = \operatorname{var}(X_i)$. Thus

$$\begin{aligned}
\operatorname{var}\left(\sum_{i=1}^n a_i X_i\right) &= \sum_{i=1}^n \sum_{j=1}^n a_i a_j \operatorname{cov}(X_i, X_j) \\
&= \sum_{i=1}^n a_i^2 \operatorname{var}(X_i) + \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n a_i a_j \operatorname{cov}(X_i, X_j) \\
&= \sum_{i=1}^n a_i^2 \operatorname{var}(X_i) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n a_i a_j \operatorname{cov}(X_i, X_j) \\
&= \sum_{i=1}^n a_i^2 \operatorname{var}(X_i) + 2 \sum_{i=2}^{n} \sum_{j=1}^{i-1} a_i a_j \operatorname{cov}(X_i, X_j)
\end{aligned}$$

Any of the more complicated re-expressions make it clear that some of the terms on the right hand side in (2.22) are "really" variances and each covariance "really" occurs twice, once in the form $\operatorname{cov}(X_i, X_j)$ and once in the form $\operatorname{cov}(X_j, X_i)$. Taking $a_i = 1$ for all i gives

$$\begin{aligned}
\operatorname{var}\left(\sum_{i=1}^n X_i\right) &= \sum_{i=1}^n \sum_{j=1}^n \operatorname{cov}(X_i, X_j) \\
&= \sum_{i=1}^n \operatorname{var}(X_i) + \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \operatorname{cov}(X_i, X_j) \\
&= \sum_{i=1}^n \operatorname{var}(X_i) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n \operatorname{cov}(X_i, X_j) \\
&= \sum_{i=1}^n \operatorname{var}(X_i) + 2 \sum_{i=2}^{n} \sum_{j=1}^{i-1} \operatorname{cov}(X_i, X_j)
\end{aligned} \tag{2.25}$$

We also write out for future reference the special case $m = n = 2$.

Corollary 2.20. *If W, X, Y, and Z are random variables having first and second moments and a, b, c, and d are constants, then*

$$\begin{aligned} \operatorname{cov}(aW + bX, cY + dZ) &= ac \operatorname{cov}(W, Y) + ad \operatorname{cov}(W, Z) \\ &\quad + bc \operatorname{cov}(X, Y) + bd \operatorname{cov}(X, Z) \qquad (2.26) \\ \operatorname{var}(aX + bY) &= a^2 \operatorname{var}(X) + 2ab \operatorname{cov}(X, Y) + b^2 \operatorname{var}(Y) \qquad (2.27) \\ \operatorname{cov}(W + X, Y + Z) &= \operatorname{cov}(W, Y) + \operatorname{cov}(W, Z) \\ &\quad + \operatorname{cov}(X, Y) + \operatorname{cov}(X, Z) \qquad (2.28) \\ \operatorname{var}(X + Y) &= \operatorname{var}(X) + 2 \operatorname{cov}(X, Y) + \operatorname{var}(Y) \qquad (2.29) \end{aligned}$$

No proof is necessary, since all of these equations are special cases of those in Theorem 2.16 and its corollaries.

This section contains a tremendous amount of "equation smearing." It is the sort of thing for which the acronym MEGO (my eyes glaze over) was invented. To help you remember the main point, let us put Corollary 2.19 in words.

> *The variance of a sum is the sum of the variances* plus *the sum of twice the covariances.*

Contrast this with the much simpler slogan about expectations on p. 35.

The extra complexity of the of the variance of a sum contrasted to the expectation of a sum is rather annoying. We would like it to be simpler. Unfortunately it isn't. However, as elsewhere in mathematics, what cannot be achieved by proof can be achieved by definition. We just make a definition that describes the nice case.

Definition 2.4.1.
Random variables X and Y are uncorrelated *if* $\operatorname{cov}(X, Y) = 0$.

We also say a set X_1, $\ldots$, X_n of random variables are uncorrelated if each pair is uncorrelated. The reason for the name "uncorrelated" will become clear when we define correlation.

When a set of random variables are uncorrelated, then there are no covariance terms in the formula for the variance of their sum; all are zero by definition.

Corollary 2.21. *If the random variables X_1, $\ldots$, X_n are uncorrelated, then*

$$\operatorname{var}(X_1 + \ldots + X_n) = \operatorname{var}(X_1) + \ldots + \operatorname{var}(X_n).$$

In words,

> *The variance of a sum is the sum of the variances* if *(big if) the variables are uncorrelated.*

Don't make the mistake of using this corollary or the following slogan when its condition doesn't hold. When the variables are correlated (have nonzero covariances), the corollary is false and you must use the more general formula of Corollary 2.19 or its various rephrasings.

What happens to Corollary 2.17 when the variables are uncorrelated is left as an exercise (Problem 2-16).

At this point the reader may have forgotten that nothing in this section has yet been proved, because we deferred the proof of Theorem 2.16, from which everything else in the section was derived. It is now time to return to that proof.

Proof of Theorem 2.16. First define

$$\begin{aligned} U &= \sum_{i=1}^{m} a_i X_i \\ V &= \sum_{j=1}^{n} b_j Y_j \end{aligned}$$

Then note that by linearity of expectation

$$\begin{aligned} \mu_U &= \sum_{i=1}^{m} a_i \mu_{X_i} \\ \mu_V &= \sum_{j=1}^{n} b_j \mu_{Y_j} \end{aligned}$$

Then

$$\begin{aligned} \operatorname{cov}(U, V) &= E\{(U - \mu_U)(V - \mu_V)\} \\ &= E\left\{ \left(\sum_{i=1}^{m} a_i X_i - \sum_{i=1}^{m} a_i \mu_{X_i} \right) \left(\sum_{j=1}^{n} b_j Y_j - \sum_{j=1}^{n} b_j \mu_{Y_j} \right) \right\} \\ &= E\left\{ \sum_{i=1}^{m} (a_i X_i - a_i \mu_{X_i}) \sum_{j=1}^{n} \bigl(b_j Y_j - b_j \mu_{Y_j} \bigr) \right\} \\ &= E\left\{ \sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_j (X_i - \mu_{X_i})(Y_j - \mu_{Y_j}) \right\} \\ &= \sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_j E\bigl\{ (X_i - \mu_{X_i})(Y_j - \mu_{Y_j}) \bigr\} \\ &= \sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_j \operatorname{cov}(X_i, Y_j), \end{aligned}$$

the last equality being the definition of covariance, the next to last linearity of expectation, and the rest being just algebra. And this proves the theorem because $\operatorname{cov}(U, V)$ is the left hand side of (2.21) in different notation. □

2.4.5 Exchangeable Random Variables

We say random variables X_1, …, X_n are exchangeable if

$$(X_1, \ldots, X_n) \stackrel{\mathcal{D}}{=} (X_{i_1}, \ldots, X_{i_n})$$

for any of the $n!$ permutations i_1, …, i_n of the integers 1, …, n. (This is equivalent to the definition in Section 3.8 in Lindgren.) In particular, if we look at marginal distributions, this implies

$$X_1 \stackrel{\mathcal{D}}{=} X_i, \qquad i = 1, \ldots, n,$$

that is, all of the X_i have the same distribution,

$$(X_1, X_2) \stackrel{\mathcal{D}}{=} (X_i, X_j), \qquad i = 1, \ldots, n,\ j = 1, \ldots, n,\ i \neq j,$$

and analogous statements for triples, quadruples, and so forth. In turn, these imply

$$\begin{aligned} E(X_1) &= E(X_i), \\ \operatorname{var}(X_1) &= \operatorname{var}(X_i), \end{aligned}$$

and analogous statements for all moments of X_1 and X_i, for all i,

$$\operatorname{cov}(X_1, X_2) = \operatorname{cov}(X_i, X_j),$$

and analogous statements for all mixed moments of X_1 and X_2 and X_i and X_j, for all i and j, and so forth for moments involving three or more variables.

Theorem 2.22. *If X_1, …, X_n are exchangeable random variables, then*

$$\operatorname{var}(X_1 + \cdots + X_n) = n \operatorname{var}(X_1) + n(n-1) \operatorname{cov}(X_1, X_2). \tag{2.30}$$

Proof. Apply (2.25). All n terms $\operatorname{var}(X_i)$ are equal to $\operatorname{var}(X_1)$, which accounts for the first term on the right hand side of (2.30). All the $\operatorname{cov}(X_i, X_j)$ terms for $i \neq j$ are equal to $\operatorname{cov}(X_1, X_2)$, and there are

$$2 \binom{n}{2} = n(n-1)$$

of these, which accounts for the second term on the right hand side of (2.30). □

2.4.6 Correlation

The Cauchy-Schwarz Inequality

Theorem 2.23 (Cauchy-Schwarz Inequality). *For any random variables X and Y having first and second moments*

$$E(|XY|) \leq \sqrt{E(X^2) E(Y^2)}. \tag{2.31}$$

This inequality is also called the *Schwarz inequality* or the *Cauchy-Schwarz-Buniakowski inequality* Statisticians generally prefer two-name eponyms, so that's what we've used.

Proof. By the positivity property of expectation for any $a \in \mathbb{R}$

$$0 \leq E\{(X + aY)^2\} = E(X^2) + 2aE(XY) + a^2 E(Y^2).$$

There are only two ways the right hand side can be nonnegative for all a.

Case I. $E(Y^2) = 0$, in which case we must also have $E(XY) = 0$, so the right hand side is equal to $E(X^2)$ regardless of the value of a.

Case II. $E(Y^2) > 0$, in which case the right hand side is a quadratic function of a that goes to infinity as a goes to plus or minus infinity and achieves its minimum where its derivative

$$2E(XY) + 2aE(Y^2)$$

is equal to zero, that is, at

$$a = -E(XY)/E(Y^2),$$

the minimum being

$$E(X^2) - 2\frac{E(XY)}{E(Y^2)}E(XY) + \left(-\frac{E(XY)}{E(Y^2)}\right)^2 E(Y^2) \;=\; E(X^2) - \frac{E(XY)^2}{E(Y^2)}$$

And this is nonnegative if and only if

$$E(XY)^2 \leq E(X^2)E(Y^2).$$

Taking the square root of both sides gives almost what we want

$$|E(XY)| \leq \sqrt{E(X^2)E(Y^2)}. \tag{2.32}$$

Plugging $|X|$ in for X and $|Y|$ in for Y in (2.32) gives (2.31). □

Note that the proof establishes (2.32) as well as (2.31). Both of these inequalities are useful and we can regard one as a minor variant of the other. The proof shows that (2.32) implies (2.31). We will eventually see (Theorem 2.28) that the implication also goes the other way, that (2.31) implies (2.32). For now, we will just consider them to be two inequalities, both of which have been proved.

Correlation

The *correlation* of real-valued random variables X and Y having strictly positive variances is

$$\begin{aligned} \operatorname{cor}(X, Y) &= \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X) \operatorname{var}(Y)}} \\ &= \frac{\operatorname{cov}(X, Y)}{\operatorname{sd}(X) \operatorname{sd}(Y)} \end{aligned}$$

If $\operatorname{var}(X)$ or $\operatorname{var}(Y)$ is zero, the correlation is undefined.

Again we might ask why two such closely related concepts as correlation and covariance. Won't just one do? (Recall that we asked the same question about variance and standard deviation.) Here too we have the same answer. The covariance is simpler to handle theoretically. The correlation is easier to understand and hence more useful in applications. Correlation has three important properties.

First, it is a dimensionless quantity, a pure number. We don't think much about units, but if we do, as we noted before the units X and $\operatorname{sd}(X)$ are the same and a little thought shows that the units of $\operatorname{cov}(X, Y)$ are the product of the units of X and Y. Thus in the formula for the correlation all units cancel.

Second, correlation is unaltered by changes of units of measurement, that is,

$$\operatorname{cor}(a + bX, c + dY) = \operatorname{sign}(bd) \operatorname{cor}(X, Y), \tag{2.33}$$

where $\operatorname{sign}(bd)$ denotes the sign (plus or minus) of bd. The proof is left as an exercise (Problem 2-25).

Third, we have the correlation inequality.

Theorem 2.24 (Correlation Inequality). *For any random variables X and Y for which correlation is defined*

$$-1 \le \operatorname{cor}(X, Y) \le 1. \tag{2.34}$$

Proof. This is an immediate consequence of Cauchy-Schwarz. Plug in $X - \mu_X$ for X and $Y - \mu_Y$ for Y in (2.32), which is implied by Cauchy-Schwarz by the comment following the proof of the inequality, giving

$$|\operatorname{cov}(X, Y)| \le \sqrt{\operatorname{var}(X) \operatorname{var}(Y)}.$$

Dividing through by the right hand side gives the correlation inequality. □

The correlation has a widely used Greek letter symbol ρ (lower case rho). As usual, if correlations of several pairs of random variables are under consideration, we distinguish them by decorating the ρ with subscripts indicating the random variables, for example, $\rho_{X,Y} = \operatorname{cor}(X, Y)$. Note that by definition of correlation

$$\begin{aligned} \operatorname{cov}(X, Y) &= \operatorname{cor}(X, Y) \operatorname{sd}(X) \operatorname{sd}(Y) \\ &= \rho_{X,Y} \sigma_X \sigma_Y \end{aligned}$$

This is perhaps one reason why covariance doesn't have a widely used Greek-letter symbol (recall that we said the symbol $\sigma_{X,Y}$ used by Lindgren is nonstandard and not understood by anyone who has not had a course using Lindgren as the textbook).

Problems

2-1. Fill in the details at the end of the proof of Corollary 2.2. Specifically, answer the following questions.

(a) Why does (2.7) assert the same thing as (2.6) in different notation?

(b) What happened to the existence assertion of the corollary? Why it is clear from the use made of Theorem 2.1 in the proof that $a + bX$ has expectation whenever X does?

2-2. Prove Corollary 2.4. As the text says, this may be done either using Axiom E1 and mathematical induction, the proof being similar to that of Theorem 2.3 but simpler, or you can use Theorem 2.3 without repeating the induction argument (the latter is simpler).

In all of the following problems the rules are as follows. You may assume in the proof of a particular theorem that all of the preceding theorems have been proved, whether the proof has been given in the course or left as an exercise. But you may not use any later theorems. That is, you may use without proof any theorem or corollary with a lower number, but you may not use any with a higher number. (The point of the rule is to avoid circular so-called proofs, which aren't really proofs because of the circular argument.)

2-3. Prove Corollary 2.5.

2-4. Prove Corollary 2.6.

2-5. If X_1, X_2, $\ldots$ is a sequence of random variables all having the same expectation μ, show that

$$E(\overline{X}_n) = \mu,$$

where, as usual, $\overline{X}_n$ is defined by (2.1).

2-6. Prove Corollary 2.7.

2-7. Prove Theorem 2.8 from Axiom E3 and Theorem 2.5.

2-8. A gambler makes 100 one-dollar bets on red at roulette. The probability of winning a single bet is 18/38. The bets pay even odds, so the gambler gains \$1 when he wins and loses \$1 when he loses.

What is the mean and the standard deviation of the gambler's net gain (amount won minus amount lost) on the 100 bets?

2-9. Prove Theorem 2.9.

2-10. Prove Theorem 2.10.

2-11. Lindgren (Definition on p. 94) defines a continuous random variable to be *symmetric about a point a* if it has a density f that satisfies

$$f(a + x) = f(a - x), \qquad \text{for all } x.$$

We, on the other hand, gave a different definition (p. 39 in these notes) gave a different definition (that $X - a$ and $a - X$ have the same distribution), which is more useful for problems involving expectations and is also more general (applying to arbitrary random variables, not just continuous ones). Show that for continuous random variables, the two definitions are equivalent, that is, suppose X is a continuous random variable with density f_X, and

(a) Find the density of $Y = X - a$.

(b) Find the density of $Z = a - X$.

(c) Show that these two densities are the same function if and only if

$$f_X(a + x) = f_X(a - x), \qquad \text{for all } x.$$

2-12. For the densities in Problem 4-8 in Lindgren, find the medians of the distributions.

2-13. Prove Corollary 2.12.

2-14. Suppose X is a zero-one-valued random variable, that is, $X(s)$ is either zero or one for all s. Suppose X has mean μ.

(a) Show that $\alpha_k = \mu$ for all positive integers k.

(b) Show that $0 \le \mu \le 1$.

(c) Show that $\operatorname{var}(X) = \mu(1 - \mu)$.

2-15. Prove Theorem 2.13. **Hint:** It helps to define $Y = a + bX$ and to use Property 2.2. Since there are now two random variables under discussion, the means must be denoted μ_X and μ_Y (what does Property 2.2 say about μ_Y) and similarly for the variances (what is to be shown is that $\sigma_Y^2 = b^2 \sigma_X^2$).

2-16. Give the general formula for the variance of a linear combination of uncorrelated random variables.

2-17. Prove Theorem 2.15.

2-18. Suppose X is a random variable having mean μ_X and standard deviation σ_X and $\sigma_X > 0$. Find a linear transformation $Y = a + bX$ so that Y has mean μ_Y and σ_Y, where μ_Y is any real number and σ_Y is any nonnegative real number.

2-19. If X_1, X_2, $\ldots$ is a sequence of uncorrelated random variables all having the same expectation μ and variance σ^2, show that

$$\operatorname{sd}(\overline{X}_n) = \frac{\sigma}{\sqrt{n}},$$

where, as usual, $\overline{X}_n$ is defined by (2.1).

2-20. State the result analogous to Theorem 2.22 giving $\operatorname{var}(\overline{X}_n)$. You need not prove your theorem (the proof is an obvious variation of the proof of Theorem 2.22).

2-21. Suppose X_1, X_2, $\ldots$, X_n are exchangeable with nonzero variance and

$$X_1 + X_2 + \cdots + X_n = 0.$$

What is $\operatorname{cor}(X_i, X_j)$ for $i \neq j$.

2-22. Suppose X_1, $\ldots$, X_n are exchangeable random variables. Show that

$$-\frac{1}{n-1} \le \operatorname{cor}(X_i, X_j).$$

Hint: Consider $\operatorname{var}(X_1 + \cdots + X_n)$. Compare with the preceding problem.

2-23. An infinite sequence of random variables X_1, X_2, $\ldots$ is said to be *exchangeable* if the finite sequence X_1, $\ldots$, X_n is exchangeable for each n.

(a) Show that correlations $\operatorname{cor}(X_i, X_j)$ for an exchangeable infinite sequence must be nonnegative. **Hint:** Consider Problem 2-22.

(b) Show that the following construction gives an exchangeable infinite sequence X_1, X_2, $\ldots$ of random variables having any correlation in the range $0 \le \rho \le 1$. Let Y_1, Y_2, $\ldots$ be an i. i. d. sequence of random variables with variance σ^2, let Z be a random variable independent of all the Y_i with variance τ^2, and define $X_i = Y_i + Z$.

2-24. Consider an infinite sequence of random variables X_1, X_2, $\ldots$ having covariances

$$\operatorname{cov}(X_i, X_j) = \rho^{|i-j|}\sigma^2$$

where $-1 < \rho < 1$ and $\sigma > 0$. Find $\operatorname{var}(\overline{X}_n)$ where, as usual, $\overline{X}_n$ is defined by (2.1). Try to simplify your formula so that it does not have an explicit sum. **Hint:** The geometric series

$$\sum_{k=0}^{n-1} a^k = \frac{1-a^n}{1-a}, \qquad -1 < a < 1$$

helps.

2-25. Prove (2.33).

2-26. Show that for any linear function, that is, a function T satisfying (2.35), $T(0) = 0$.

2.5 Probability Theory as Linear Algebra

This section has two objectives.

The minor objective is to explain something that might be bothering the astute reader. What is the connection between the linearity property of expectation (Property 2.1) and the linearity property that defines linear transformations in linear algebra. They look similar. What's the connection?

The major objective is to provide some mathematical models for expectation. Everything we have done so far, important as it is, mostly tells us how some expectations relate to other expectations. Linearity of expectation, for example tells us that if we know $E(X)$ and $E(Y)$, then we can calculate $E(aX + bY)$. It doesn't tell us where $E(X)$ and $E(Y)$ come from in the first place.

2.5.1 The Vector Space L^1

Although we haven't gotten to it yet, we will be using linear algebra in this course. The linearity property of linear transformations between vector spaces will be important. If these two linearity properties (the one from linear algebra and the one from probability theory) are different, what is the difference and how can you keep from confusing them?

Fortunately, there is nothing to confuse. The two properties are the same, or, more precisely, expectation is a linear transformation.

Theorem 2.25. *L^1 is a real vector space, and E is a linear functional on L^1.*

The proof is trivial (we will give it below). The hard part is understanding the terminology, especially if your linear algebra is a bit rusty. So our main effort will be reviewing enough linear algebra to understand what the theorem means.

Vector Spaces

Every linear algebra book starts with a definition of a *vector space* that consists of a long list of formal properties. We won't repeat them. If you are interested, look in a linear algebra book. We'll only review the facts we need here.

First a vector space is a set of objects called *vectors.* They are often denoted by boldface type. It is associated with another set of objects called *scalars.* In probability theory, the scalars are always the real numbers. In linear algebra, the scalars are often the complex numbers. More can be proved about complex vector spaces (with complex scalars) than about real vector spaces (with real scalars), so complex vector spaces are more interesting to linear algebraists. But they have no application in probability theory. So to us "scalar" is just a synonym for "real number."

There are two things you can do with vectors.

- You can add them (vector addition). If $\mathbf{x}$ and $\mathbf{y}$ are vectors, then there exists another vector $\mathbf{x} + \mathbf{y}$.
- You can multiply them by scalars (scalar multiplication). If $\mathbf{x}$ is a vector and a is a scalar, then there exists another vector $a\mathbf{x}$.

If you got the impression from your previous exposure to linear algebra (or from Chapter 1 of these notes) that the typical vector is an n-tuple

$$\mathbf{x} = (x_1, \ldots, x_n)$$

or perhaps a "column vector" ($n \times 1$ matrix)

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

you may be wondering what the connection between random variables and vectors could possibly be. Random variables are functions (on the sample space) and functions aren't n-tuples or matrices.

But n-tuples are functions. You just have to change notation to see it. Write $x(i)$ instead of x_i, and it's clear that n-tuples are a special case of the function concept. An n-tuple is a function that maps the index i to the value x_i.

So the problem here is an insufficiently general notion of vectors. You should think of functions (rather than n-tuples or matrices) as the most general notion of vectors. Functions can be added. If f and g are functions on the same domain, then $h = f + g$ means

$$h(s) = f(s) + g(s), \qquad \text{for all } s \text{ in the domain.}$$

Functions can be multiplied by scalars. If f is a function and a is a scalar (real number), then $h = af$ means

$$h(s) = af(s), \qquad \text{for all } s \text{ in the domain.}$$

Thus the set of scalar-valued functions on a common domain form a vector space. In particular, the scalar-valued random variables of a probability model (all real-valued functions on the sample space) form a vector space. Theorem 2.25 asserts that L^1 is a subspace of this vector space.

Linear Transformations and Linear Functionals

If U and V are vector spaces and T is a function from U to V, then we say that T is *linear* if

$$T(a\mathbf{x} + b\mathbf{y}) = aT(\mathbf{x}) + bT(\mathbf{y}), \qquad \text{for all vectors } \mathbf{x} \text{ and } \mathbf{y} \text{ and scalars } a \text{ and } b. \tag{2.35}$$

Such a T is sometimes called a *linear transformation* or a *linear mapping* rather than a *linear function.*

The set of scalars (the real numbers) can also be thought of as a (one-dimensional) vector space, because scalars can be added and multiplied by scalars. Thus we can also talk about scalar-valued (real-valued) linear functions on a vector space. Such a function satisfies the same property (2.35). The only difference is that it is scalar-valued rather than vector-valued. In linear algebra, a scalar-valued linear function is given the special name *linear functional.*

Theorem 2.25 asserts that the mapping from random variables X to their expectations $E(X)$ is a linear functional on L^1. To understand this you have to think of E as a function, a rule that assigns values $E(X)$ to elements X of L^1.

Proof of Theorem 2.25. The existence assertions of Properties E1 and E2 assert that random variables in L^1 can be added and multiplied by scalars yielding a result in L^1. Thus L^1 is a vector space. Property 2.1 now says the same thing as (2.35) in different notation. The map E, being scalar-valued, is thus a linear functional on L^1. □

2.5.2 Two Notions of Linear Functions

The preceding section showed that there was no difference between the notion of linearity used in linear algebra and linearity of expectation in probability theory.

There is, however, another notion of linearity. In fact, we already used it in (2.9) and silently skipped over the conflict with (2.35). To be more precise, we should say that (2.9) defines a function that is linear in the sense of high-school algebra or first-year calculus (or in the sense used in statistics and various other kinds of applied mathematics), and (2.35) defines a function that is linear in the sense of linear algebra (and other higher mathematics).

To simplify terminology and indicate the two notions with single words, mathematicians call the first class of functions *affine* and the second class *linear*. Note that affine functions are what everyone but pure mathematicians calls linear functions.

The two notions are closely related, but slightly different. An affine function is a linear function plus a constant. If T is a linear function from a vector space U to a vector space V, that is, a function satisfying (2.35), and $\mathbf{a}$ is any vector in V, then the map A defined by

$$A(\mathbf{x}) = \mathbf{a} + T(\mathbf{x}), \qquad \mathbf{x} \in V \tag{2.36}$$

is an affine function.

If we were mathematical purists, we would always call functions of the form (2.36) "affine," but if we taught you to do that, no one would understand what you were talking about except for pure mathematicians. So we won't. We will call functions of the form (2.36) "linear," like everyone but pure mathematicians. Only when we think confusion is likely will we call them "linear in the ordinary sense" or "affine."

Confusion between the two is fairly easy to clear up. Linear functions (in the strict sense) are a special case of affine functions. They are the ones satisfying $T(0) = 0$ (Problem 2-26). So just check whether this holds. If so, linear is meant in the strict sense, if not, linear is meant in the ordinary sense.

So that explains the difference between affine and linear. The only question remaining is why (2.9) defines an affine function. What does (2.9) have to do with (2.36)? First (2.9) defines a scalar-valued affine function of a scalar variable. This makes both the constant and the function values in (2.36) scalar, so we can rewrite it as

$$g(x) = a + h(x), \qquad x \in \mathbb{R},$$

where a is a scalar and h is a scalar-valued linear function on $\mathbb{R}$. To get this in the form (2.9) we only need to show that the most general scalar-valued linear function on $\mathbb{R}$ has the form

$$h(x) = bx, \qquad x \in \mathbb{R},$$

where b is a real number. The homogeneity property applied to h says

$$h(x) = h(x \cdot 1) = xh(1), \qquad x \in \mathbb{R}.$$

So we are done, the identification $b = h(1)$ makes the two equations the same.

2.5.3 Expectation on Finite Sample Spaces

Consider a finite set S and define L^1 to be the set of all real-valued functions on S. This makes L^1 a finite-dimensional vector space. The elements of L^1 differ from n-tuples only in notation. A random variable $X \in L^1$ is determined by its values

$$X(s), \qquad s \in S,$$

and since S is finite, this means X is determined by a finite list of real numbers. If S is indexed

$$S = \{s_1, \ldots, s_n\}$$

then we could even, if we wanted, collect these values into an n-tuple

$$(x_1, \ldots, x_n)$$

where

$$x_i = X(s_i), \qquad i = 1, \ldots, n,$$

which shows explicitly the correspondence between n-tuples and functions on a set of cardinality n.

However, we don't want to make too much of this correspondence. In fact the only use we want to make of it is the following fact: every linear functional T on an n-dimensional vector space has the form

$$T(\mathbf{x}) = \sum_{i=1}^{n} a_i x_i \tag{2.37}$$

where, as usual, $\mathbf{x} = (x_1, \ldots, x_n)$. This is sometimes written

$$T(\mathbf{x}) = \mathbf{a}'\mathbf{x}$$

where $\mathbf{a} = (a_1, \ldots, a_n)$ the prime indicating transpose and $\mathbf{a}$ and $\mathbf{x}$ being considered as column vectors. Other people write

$$T(\mathbf{x}) = \mathbf{a} \cdot \mathbf{x}$$

the operation indicated by the dot being called the *scalar product* or *dot product* of the vectors $\mathbf{a}$ and $\mathbf{x}$.

We now want to change back to our original notation, writing vectors as functions on a finite set S rather than n-tuples, in which case (2.37) becomes

$$T(\mathbf{x}) = \sum_{s \in S} a(s)x(s)$$

Now we want to make another change of notation. If we want to talk about vectors that are elements of L^1 (and we do), we should use the usual notation, denoting those elements (which are random variables) by X rather than $\mathbf{x}$ and their components by $X(s)$ giving

$$T(X) = \sum_{s \in S} a(s) X(s). \tag{2.38}$$

To summarize the argument of this section so far

Theorem 2.26. *For probability models on a finite sample space S, every linear functional on L^1 has the form* (2.38)*.*

But not every linear functional is an expectation operator. Every linear functional satisfies two of the probability axioms (homogeneity and additivity). But a linear functional need not satisfy the other two (positivity and norm).

In order that (2.38) be positive whenever $X \geq 0$, that is, when $X(s) \geq 0$, for all s, it is required that

$$a(s) \geq 0, \qquad s \in S. \tag{2.39a}$$

In order that (2.38) satisfy the norm property (2.4) it is required that

$$\sum_{s \in S} a(s) = 1, \tag{2.39b}$$

because $X = 1$ means $X(s) = 1$, for all s. We have met functions like this before: a function a satisfying (2.39a) and (2.39b) we call a *probability density.* Lindgren calls them probability functions (p. f.'s).

Theorem 2.27. *For probability models on a finite sample space S, every expectation operator on L^1 has the form*

$$E(X) = \sum_{s \in S} p(s) X(s) \tag{2.40}$$

for some function $p : S \to \mathbb{R}$ satisfying

$$p(s) \geq 0, \qquad s \in S, \tag{2.41a}$$

and

$$\sum_{s \in S} p(s) = 1. \tag{2.41b}$$

A function p as defined in the theorem is called a *probability density* or just a *density.*

Remark. Theorem 2.27 is also true if the word "finite" in the first sentence is replaced by "countable" (see Theorem 2.30).

A little section about *mathematics is invariant under changes of notation.* We often write (2.40) in different notation. If X is a random variable with density f_X having domain S (the range of possible values of X), then

$$E\{g(X)\} = \sum_{x \in S} g(x) f_X(x). \tag{2.42}$$

Note that (2.42) is exactly the same as (2.40) except for purely notational differences. The special case where g is the identity function

$$E(X) = \sum_{x \in S} x f_X(x) \tag{2.43}$$

is of some interest. Lindgren takes (2.43) as the *definition* of expectation. For us it is a trivial special case of the more general formula (2.42), which in turn is not a definition but a theorem (Theorem 2.27). For us the *definition* of expectation is "an operator satisfying the axioms."

Example 2.5.1 (The Binomial Distribution).
Recall the binomial distribution (Section B.1.2 of Appendix B) having density

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, \ldots, n.$$

We want to calculate $E(X)$. By the formulas in the preceding discussion

$$\begin{aligned}
E(X) &= \sum_{x=0}^{n} x f(x) \\
&= \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} \\
&= \sum_{k=0}^{n} k \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\
&= \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} p^k (1-p)^{n-k} \\
&= np \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} (1-p)^{n-k} \\
&= np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} \\
&= np \sum_{m=0}^{n-1} \binom{n-1}{m} p^m (1-p)^{n-1-m}
\end{aligned}$$

- Going from line 1 to line 2 we just plugged in the definition of $f(x)$ and changed the dummy variable of summation from x to k.

- Going from line 2 to line 3 we just plugged in the definition of the binomial coefficient.
- Going from line 3 to line 4 we just observed that the $k = 0$ term is zero and then canceled the k in the numerator with the k in the $k!$ in the denominator.
- Going from line 4 to line 5 we pulled an n out of the $n!$ and a p out of the p^k.
- Going from line 5 to line 6 we just used the definition of the binomial coefficient again.
- Going from line 6 to line 7 we changed the dummy variable of summation to $m = k - 1$.

Now the binomial theorem says the sum in the last line is equal to one. Alternatively, the sum in the last line is equal to one because the summand is the $\text{Bin}(n - 1, p)$ density, and *every* probability density sums to one. Hence

$$E(X) = np.$$

2.5.4 Axioms for Expectation (Part II)

Absolute Values

Axiom E5 (Absolute Values). *If X is in L^1, then so is $|X|$.*

Note that this axiom trivially applies to the probability models on a finite sample space discussed in the preceding section, because *every* real-valued function is in L^1. This axiom is only interesting when the sample space is infinite.

With this axiom, we can prove another basic property of expectation that is mostly used in theoretical arguments.

Theorem 2.28 (Absolute Values). *If X is in L^1, then*

$$|E(X)| \le E(|X|).$$

The name of this theorem is "taking an absolute value inside an expectation can only increase it." That's a long-winded name, but there is no widely used short name for the theorem.

Derivation of Property 2.28. First note that $X \le |X|$. Applying Property 2.8 to these two random variables gives

$$E(X) \le E(|X|),$$

which is what was to be proved in the case that $E(X)$ is nonnegative.

To prove the other case, we start with the fact that $-X \le |X|$. Another application of Property 2.8 along with Property 2.6 gives

$$-E(X) = E(-X) \le E(|X|).$$

But when $E(X)$ is negative $-E(X) = |E(X)|$, so that proves the other case. □

Note that there is no explicit mention of Axiom E5 in the proof. The *implicit* mention is that only the axiom allows us to talk about $E(|X|)$. None of the other axioms guarantee that $|X|$ has expectation.

Monotone Convergence

The last axiom for expectation analogous to the countable additivity axiom for probability (called Axiom 3a on p. 30 in Lindgren). This is the monotone convergence axiom. To understand it we need a preliminary definition. For a sequence of numbers $\{x_n\}$, the notation $x_n \uparrow x$ means $x_1 \le x_2 \le \ldots$ and $x_n \to x$. For a sequence of random variables $\{X_n\}$ on a sample space S, the notation $X_n \uparrow X$ means $X_n(s) \uparrow X(s)$ for all $s \in S$.

Axiom E6 (Monotone Convergence). *Suppose X_1, X_2, ... is a sequence of random variables in L^1 such that $X_n \uparrow X$. If*

$$\lim_{n \to \infty} E(X_n) < \infty,$$

then $X \in L^1$ and

$$E(X_n) \uparrow E(X).$$

Conversely, if

$$\lim_{n \to \infty} E(X_n) = \infty,$$

then $X \notin L^1$.

The monotone convergence axiom is a fairly difficult subject, so difficult that Lindgren omits it entirely from his book, although this makes no sense because the countable additivity axiom for probability is equally difficult and is included. So this is really more treating expectation is a second class concept, subsidiary to probability. Our insistence on including it is part and parcel of our notion that probability and expectation are equally important and deserve equal treatment.

That having been said, this axiom can be considered the dividing line between material at the level of this course and material over our heads. If a proof involves monotone convergence, it is too hard for us. We will state some results that can only be proved using the monotone convergence axiom, but we will leave the proofs for more advanced courses.

There is a "down arrow" concept defined in obvious analogy to the "up arrow" concept (the sequence converges down rather than up), and there is an analogous form of monotone convergence

Corollary 2.29 (Monotone Convergence). *Suppose X_1, X_2, ... is a sequence of random variables in L^1 such that $X_n \downarrow X$. If*

$$\lim_{n \to \infty} E(X_n) > -\infty,$$

then $X \in L^1$ and

$$E(X_n) \downarrow E(X).$$

Conversely, if

$$\lim_{n \to \infty} E(X_n) = -\infty,$$

then $X \notin L^1$*.*

2.5.5 General Discrete Probability Models

If the sample space S of a probability model is countably infinite, we would like to use the same formulas (2.40), (2.41a) and (2.41b), that we used for finite sample spaces, but we run into problems related to infinite series. The sum may not exist (the series may not converge), and if it does exist, its value may depend on the particular enumeration of the sample space that is used. Specifically, there are many ways to enumerate the sample space, writing it as a sequence $S = \{s_1, s_2, \ldots\}$, and when we write out the infinite sum explicitly as

$$E(X) = \sum_{i=1}^{\infty} X(s_i)p(s_i) = \lim_{n \to \infty} \sum_{i=1}^{n} X(s_i)p(s_i)$$

the limit may depend on the particular enumeration chosen. The axioms of expectation, however, solve both of these problems.

First, not all random variables have expectation, only those in L^1 so the fact that expectation may not be defined for some random variables should not bother us. For discrete probability models on a sample space S defined by a probability density p, we define L^1 to be the set of all functions $X : S \to \mathbb{R}$ satisfying

$$\sum_{s \in S} |X(s)|p(s) < \infty. \tag{2.44}$$

This definition trivially satisfies Axiom E5 and also satisfies the existence parts of Axioms E1, E2, and E4.

For $X \in L^1$ we define expectation by the same formula (2.40) as in the finite sample space case. Note that then the sum in (2.44) is $E(|X|)$. Thus our definition says that X has expectation if and only if $|X|$ also has expectation. Another way to say the same thing is that (2.40) defines an expectation if and only if the series is *absolutely summable*, which means the sum of the absolute values of the terms of the series exists.

Because of the rearrangement of series theorem from calculus, which says that if a series is absolutely summable then the sum of the series does not depend on the order in which the terms are summed, we can rearrange the terms in the sum as we please without changing the result. That is why we can write (2.40) as an *unordered* sum using notation that does not specify any particular ordering.

Theorem 2.30. *All probability models on a countable sample space S are defined by a function function $p : S \to \mathbb{R}$ satisfying*

$$p(s) \geq 0, \qquad s \in S, \tag{2.45a}$$

and

$$\sum_{s \in S} p(s) = 1. \tag{2.45b}$$

The corresponding expectation operator is $E : L^1 \to \mathbb{R}$, where L^1 is the set of functions $X : S \to \mathbb{R}$ such that

$$\sum_{s \in S} p(s) |X(s)| < \infty,$$

and

$$E(X) = \sum_{s \in S} p(s) X(s) \tag{2.46}$$

Following our policy that any proof that involves dominated convergence is beyond the scope of this course, we won't try to prove the theorem.

Note that the remarks about *mathematics is invariant under changes of notation* in the preceding section apply here too. In particular, (2.42) and (2.43) apply just as well in the case that S is countably infinite (so long as the expectation in question exists).

Example 2.5.2 (The Poisson Distribution).
The the Poisson distribution is the discrete distribution having density

$$f(x) = \frac{\mu^x}{x!} e^{-\mu}, \qquad x = 0, 1, \ldots.$$

(Section B.1.4 of Appendix B). If $X \sim \text{Poi}(\mu)$, then

$$\begin{aligned} E(X) &= \sum_{x=0}^{\infty} x f(x) \\ &= \sum_{k=0}^{\infty} k \frac{\mu^k}{k!} e^{-\mu} \\ &= \mu \sum_{k=1}^{\infty} \frac{\mu^(k-1)}{(k-1)!} e^{-\mu} \\ &= \mu \sum_{m=0}^{\infty} \frac{\mu^m}{m!} e^{-\mu} \end{aligned}$$

- Going from line 1 to line 2 we just plugged in the definition of $f(x)$ and changed the dummy variable of summation from x to k.

- Going from line 2 to line 3 we just observed that the $k = 0$ term is zero, then canceled the k in the numerator with the k in the $k!$ in the denominator, and then pulled a μ out of the μ^k.

- Going from line 3 to line 4 we changed the dummy variable of summation to $m = k - 1$.

The sum in the last line is equal to one because the summand is the Poi(μ) density, and *every* probability density sums to one. Hence

$$E(X) = \mu.$$

2.5.6 Continuous Probability Models

When the sample space is uncountable, like $\mathbb{R}$ or $\mathbb{R}^d$ we cannot use the formulas of Theorem 2.27 to define expectation. There is no notion of sums with an uncountably infinite number of terms.

There is, however, another concept that behaves much like summation, which is integration. We just replace the sums by integrals.

Theorem 2.31. *Probability models on having a subset S of $\mathbb{R}$ or $\mathbb{R}^d$ can be defined by a function function $f : S \to \mathbb{R}$ satisfying*

$$f(x) \ge 0, \qquad x \in S, \tag{2.47a}$$

and

$$\int_S f(x)\, dx = 1. \tag{2.47b}$$

The space L^1 of random variables having expectations is the set of real-valued functions $g : S \to \mathbb{R}$ such that

$$\int_S |g(x)| f(x)\, dx < \infty.$$

The corresponding expectation operator is $E : L^1 \to \mathbb{R}$ is defined by

$$E\{g(X)\} = \int_S g(x) f(x)\, dx. \tag{2.48}$$

As in the discrete case, we define expectation so that Y has expectation only if $|Y|$ also has expectation. Since we are using integrals rather than sums, we are now interested in absolute integrability rather than absolute summability, but there is a complete analogy between the two cases.

Similar formulas hold when the sample space is $\mathbb{R}^d$ or a subset S of $\mathbb{R}^d$. The general formula, written in vector notation and ordinary multiple-integral notation is

$$\begin{aligned} E\{g(\mathbf{X})\} &= \int_S g(\mathbf{x}) f(\mathbf{x})\, d\mathbf{x} \\ &= \underset{S}{\iint \cdots \int} g(x_1, x_2, \ldots, x_n) f(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n \end{aligned} \tag{2.49}$$

Now we take a time out for a comment that is "beyond the scope of this course." We just lied to you, sort of. Theorem 2.31 is not true if the integral signs indicate the kind of integral (the so-called *Riemann integral*) described in elementary calculus courses. All the axioms except monotone convergence are satisfied, and monotone convergence

$$\lim_{n \to \infty} \int g_n(x) f(x) \, dx = \int g(x) f(x) \, dx, \qquad \text{if } g_n \uparrow g. \tag{2.50}$$

holds sometimes but not always.

The problem is that the limit of a sequence of Riemann integrable functions is not necessarily Riemann integrable. So even though (2.50) is true whenever all the functions involved are Riemann integrable, that isn't enough to satisfy the monotone convergence axiom. The way around this problem is a *tour de force* of higher mathematics. One just makes (2.50) hold *by definition*. First one shows that for two sequences $g_n \uparrow g$ and $h_n \uparrow g$ increasing to the same limit

$$\lim_{n \to \infty} \int g_n(x) f(x) \, dx = \lim_{n \to \infty} \int h_n(x) f(x) \, dx \tag{2.51}$$

Therefore if we just *define* the right hand side of (2.50) to be the left hand side, the equation is then true by definition. This definition is unambiguous because the value of the limit does not depend on the sequence chosen (2.51). This "extension by monotone convergence" of the definition of the integral is called the *Lebesgue integral* .

Note that the Riemann integral always agrees with the Lebesgue integral whenever both are defined, so this is not a totally new concept. Every function you already know how to integrate has the same integral in both senses. The only point of Lebesgue integration is that it allows the integration of some *really weird* functions, too weird to have Riemann integrals. Since no really weird functions are of any practical interest, the only point of the whole exercise is to prove theorems using the monotone convergence axiom. And since that is beyond the scope of this course, we won't worry about it.

Example 2.5.3 (The Gamma Distribution).
The the Gamma distribution is the continuous distribution having density

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha - 1} e^{-\lambda x}, \qquad x > 0$$

(Section B.2.3 of Appendix B). If $X \sim \text{Gam}(\alpha, \lambda)$, then

$$\begin{aligned} E(X) &= \int_0^\infty x f(x) \, dx \\ &= \int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} x^\alpha e^{-\lambda x} \, dx \\ &= \frac{\Gamma(\alpha + 1)}{\lambda \Gamma(\alpha)} \int_0^\infty \frac{\lambda^{\alpha+1}}{\Gamma(\alpha + 1)} x^\alpha e^{-\lambda x} \, dx \end{aligned}$$

- Going from line 1 to line 2 we just plugged in the definition of $f(x)$ and collected the x and $x^{\alpha-1}$ terms together.

- Going from line 2 to line 3 we just pulled some constants outside of the integral.

The integral in the last line is equal to one because the integrand is the density of the $\text{Gam}(\alpha + 1, \lambda)$ distribution, and *every* probability density integrates to one. Hence

$$E(X) = \frac{\Gamma(\alpha + 1)}{\lambda \Gamma(\alpha)} = \frac{\alpha}{\lambda}$$

the second equality being the recurrence relation for the gamma function (B.3) in Section B.3.1 of Appendix B.

2.5.7 The Trick of Recognizing a Probability Density

The astute reader may have recognized a pattern to Examples 2.5.1, 2.5.2, and 2.5.3. In each case the sum or integral was done by recognizing that by moving certain constants (terms not containing the variable of summation or integration) outside of the sum or integral leaving only the sum or integral of a *known probability density*, which is equal to one by definition.

Of course, you don't have to use the trick. There is more than one way to do it. In fact, we even mentioned that you could instead say that we used the binomial theorem to do the sum in Example 2.5.1. Similarly, you could say we use the Maclaurin series for the exponential function

$$e^x = 1 + x + \frac{x^2}{2} + \cdots + \frac{x^k}{k!} + \cdots$$

to do the sum in Example 2.5.2, and you could say we use the definition of the gamma function, (B.2) in Appendix B plus the change-of-variable formula to do the integral in Example 2.5.3. In fact, the argument we gave using the fact that densities sum or integrate to one as the case may be does use these *indirectly*, because those are the reasons why these densities sum or integrate to one.

The point we are making here is that in every problem involving an expectation in which you are doing a sum or integral, you already have a know sum or integral to work with. This is expecially important when there is a whole *parametric family* of densities to work with. In calculating the mean of a $\Gamma(\alpha, \lambda)$ distribution, we used the fact that a $\Gamma(\alpha + 1, \lambda)$ density, like all densities, integrates to one. This is a very common trick. One former student said that if you can't do an integral using this trick, then you can't do it at all, which is not quite true, but close. Most integrals and sums you will do to calculate expectations can be done using this trick.

2.5.8 Probability Zero

Events of probability zero are rather a nuisance, but they cannot be avoided in continuous probability models. First note that every outcome is an event of probability zero in a continuous probability model, because by definition

$$P(X = a) = \int_a^a f(x) \, dx,$$

and a definite integral over an interval of length zero is zero.

Often when we want to assert a fact, it turns out that the best we can get from probability is an assertion "with probability one" or "except for an event of probability zero." The most important of these is the following theorem, which is essentially the same as Theorem 5 of Chapter 4 in Lindgren.

Theorem 2.32. *If $Y = 0$ with probability one, then $E(Y) = 0$. Conversely, if $Y \geq 0$ and $E(Y) = 0$, then $Y = 0$ with probability one.*

The phrase "$Y = 0$ with probability one" means $P(Y = 0) = 1$. The proof of the theorem involves dominated convergence and is beyond the scope of this course.

Applying linearity of expectation to the first half of the theorem, we get an obvious corollary.

Corollary 2.33. *If $X = Y$ with probability one, then $E(X) = E(Y)$.*

If $X = Y$ with probability one, then the set

$$A = \{ s : X(s) \neq Y(s) \}$$

has probability zero. Thus a colloquial way to rephrase the corollary is "what happens on a set of probability zero doesn't matter." Another rephrasing is "a random variable can be arbitrarily redefined on a set of probability zero without changing any expectations."

There are two more corollaries of this theorem that are important in statistics.

Corollary 2.34. $\operatorname{var}(X) = 0$ *if and only if X is constant with probability one.*

Proof. First, suppose $X = a$ with probability one. Then $E(X) = a = \mu$, and $(X - \mu)^2$ equals zero with probability one, hence by Theorem 2.32 its expectation, which is $\operatorname{var}(X)$, is zero.

Conversely, by the second part of Theorem 2.32, $\operatorname{var}(X) = E\{(X-\mu)^2\} = 0$ implies $(X-\mu)^2 = 0$ with probability one because $(X-\mu)^2$ is a random variable that is nonnegative and integrates to zero. Since $(X - \mu)^2$ is zero only when $X = \mu$, this implies $X = \mu$ with probability one. □

Corollary 2.35. $|\operatorname{cor}(X, Y)| = 1$ *if and only if there exist constants α and β such that $Y = \alpha + \beta X$ with probability one.*

Proof. First suppose $Y = \alpha + \beta X$ with probability one. Then by (2.33)

$$\operatorname{cor}(\alpha + \beta X, X) = \operatorname{sign}(\beta) \operatorname{cor}(X, X) = \pm 1.$$

That proves one direction of the "if and only if."

To prove the other direction, we assume $\rho_{X,Y} = \pm 1$ and have to prove that $Y = \alpha + \beta X$ with probability one, where α and β are constants we may choose. I claim that the appropriate choices are

$$\beta = \rho_{X,Y} \frac{\sigma_Y}{\sigma_X}$$
$$\alpha = \mu_Y - \beta \mu_X$$

(these are just pulled out of the air here, the choice will make sense after we have done best linear prediction).

We want to prove that $Y = \alpha + \beta X$ with probability one. We can do this by showing that $(Y - \alpha - \beta X)^2$ is zero with probability one, and this will follow from Theorem 2.32 if we can show that $(Y - \alpha - \beta X)^2$ has expectation zero. Hence let us calculate

$$\begin{aligned}
E\left\{(Y - \alpha - \beta X)^2\right\} &= E\left\{\left(Y - \mu_Y - \rho_{X,Y}\frac{\sigma_Y}{\sigma_X}(X - \mu_X)\right)^2\right\} \\
&= E\left\{(Y - \mu_Y)^2\right\} \\
&\qquad - 2E\left\{(Y - \mu_Y)\left(\rho_{X,Y}\frac{\sigma_Y}{\sigma_X}(X - \mu_X)\right)\right\} \\
&\qquad + E\left\{\left(\rho_{X,Y}\frac{\sigma_Y}{\sigma_X}(X - \mu_X)\right)^2\right\} \\
&= \operatorname{var}(Y) - 2\rho_{X,Y}\frac{\sigma_Y}{\sigma_X}\operatorname{cov}(X, Y) + \rho_{X,Y}^2\frac{\sigma_Y^2}{\sigma_X^2}\operatorname{var}(X) \\
&= \sigma_Y^2 - 2\rho_{X,Y}^2\sigma_Y^2 + \rho_{X,Y}^2\sigma_Y^2 \\
&= \sigma_Y^2(1 - \rho_{X,Y}^2)
\end{aligned}$$

which equals zero because of the assumption $|\rho_{X,Y}| = 1$. □

2.5.9 How to Tell When Expectations Exist

We say a random variable Y *dominates* a random variable X if $|X| \le |Y|$.

Theorem 2.36. *If Y dominates X and Y has expectation, then X also has expectation. Conversely if Y dominates X and the expectation of X does not exist, then the expectation of Y does not exist either.*

The proof involves monotone convergence and is hence beyond the scope of this this course.[1]

We say a random variable X is *bounded* if $|X| \le a$ for some constant a.

[1] Actually this theorem is way, way beyond the scope of this course, the one subject we will touch on that is really, really, really weird. Whether this theorem is true or false is a matter of taste. Its truth depends on an axiom of set theory (the so-called axiom of choice), which can be assumed or not without affecting anything of practical importance. If the theorem is false, that means there exists a random variable X dominated by another random variable Y such that Y is in L^1 and X isn't. However, the usual assumptions of advanced probability theory imply that every Riemann integrable random variable dominated by Y is in L^1, hence X cannot be written as the limit of a sequence $X_n \uparrow X$ for a sequence of Riemann integrable random variables X_n. This means that X is weird indeed. Any conceivable description of X (which like any random variable is a function on the sample space) would have not only infinite length but uncountably infinite length. That's weird! What is not widely known, even among experts, is that there is no need to assume such weird functions actually exist. The entirety of advanced probability theory can be carried through under the assumption that Theorem 2.36 is true (R. M. Solovay, "A Model of Set-Theory in Which Every Set of Reals is Lebesgue Measurable," *Annals of Mathematics*, 92:1-56, 1970).

Corollary 2.37. *Every bounded random variable is in L^1.*

Corollary 2.38. *In a probability model with a finite sample space, every random variable is in L^1.*

The corollaries take care of the trivial cases. Thus the question of existence or non-existence of expectations only applies to unbounded random variables in probability models on infinite sample spaces. Then Theorem 2.36 is used to determine whether expectations exist. An expectation is an infinite sum in the discrete case or an integral in the continuous case. The question is whether the integral or sum converges absolutely. That is, if we are interested in the expectation of the random variable $Y = g(X)$ where X has density f, we need to test the integral

$$E(|Y|) = \int |g(x)| f(x) \, dx$$

for finiteness in the continuous case, and we need to test the corresponding sum

$$E(|Y|) = \sum_{x \in S} |g(x)| f(x)$$

for finiteness in the discrete case. The fact that the integrand or summand has the particular product form $|g(x)|f(x)$ is irrelevant. What we need to know here are the rules for determining when an integral or infinite sum is finite.

We will cover the rules for integrals first. The rules for sums are very analogous. Since we are only interested in nonnegative integrands, we can always treat the integral as representing "area under the curve" where the curve in question is the graph of the integrand. Any part of the region under the curve that fits in a finite rectangle is, of course, finite. So the only way the area under the curve can be infinite is if part of the region does not fit in a finite rectangle: either the integrand has a singularity (a point where it goes to infinity), or the domain of integration is an unbounded interval. It helps if we focus on each problem separately: we test whether integrals over neighborhoods of singularities are finite, and we test whether integrals over unbounded intervals are finite. Integrals over bounded intervals not containing singularities do not need to be checked at all.

For example, suppose we want to test whether

$$\int_0^\infty h(x) \, dx$$

is finite, and suppose that the only singularity of h is at zero. For any numbers a and b such that $0 < a < b < \infty$ we can divide up this integral as

$$\int_0^\infty h(x) \, dx = \int_0^a h(x) \, dx + \int_a^b h(x) \, dx + \int_b^\infty h(x) \, dx$$

The first integral on the right hand side may be infinite because of the singularity. The third integral on the right hand side may be infinite because of the unbounded domain of integration. The second integral on the right hand side must be finite: the integral of a bounded function over a bounded domain is always finite, we do not need to check.

It is rare that we can exactly evaluate the integrals. Usually we have to use Theorem 2.36 to settle the existence question by comparing with a simpler integral. The following lemmas give the most useful integrals for such comparisons. While we are at it, we give the analogous useful infinite sums. The proofs are all elementary calculus.

Lemma 2.39. *For any positive real number a or any positive integer m*

$$\int_a^\infty x^b \, dx \qquad \text{and} \qquad \sum_{n=m}^\infty n^b$$

exist if and only if $b < -1$.

Lemma 2.40. *For any positive real number a*

$$\int_0^a x^b \, dx$$

exists if and only if $b > -1$.

Lemma 2.41. *For any positive real number a or any positive integer m and any positive real number c and any real number b (positive or negative)*

$$\int_a^\infty x^b e^{-cx} \, dx \qquad \text{and} \qquad \sum_{n=m}^\infty n^b e^{-cn}$$

exist.

The following two lemmas give us more help using the domination theorem.

Lemma 2.42. *Suppose g and h are bounded, strictly positive functions on an interval $[a, \infty)$ and*

$$\lim_{x \to \infty} \frac{g(x)}{h(x)} = k, \tag{2.52}$$

where k is a strictly positive constant, then either both of the integrals

$$\int_a^\infty g(x) \, dx \qquad \text{and} \qquad \int_a^\infty h(x) \, dx \tag{2.53}$$

are finite, or neither is. Similarly, either both of the sums

$$\sum_{k=m}^\infty g(k) \qquad \text{and} \qquad \sum_{k=m}^\infty h(k) \tag{2.54}$$

are finite, or neither is, where m is any integer greater than a.

Example 2.5.4 (Exponentially Decreasing Tails).
The following densities

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad -\infty < x < \infty$$

and

$$f(x) = \frac{1}{2} e^{-|x|}, \qquad -\infty < x < \infty$$

have moments of all orders, that is, $E(|X|^p)$ exists for all $p > 0$.

Why? Because the densities are bounded (no singularities) and have exponentially decreasing tails, so Lemma 2.41 assures us that all moments exist.

Example 2.5.5 (Polynomially Decreasing Tails).
The following densities

$$f(x) = \frac{1}{\pi(1 + x^2)}, \qquad -\infty < x < \infty$$

and

$$f(x) = \frac{6}{\pi^2 x^2}, \qquad x = 1, 2, \ldots$$

do not have moments of all orders. In fact, for both $E(|X|^p)$ exists for $p > 0$ if and only if $p < 1$. Thus for these two distributions, neither the mean, nor the variance, nor any higher moment exists.

Why? In both cases, if we look at the integrand or summand $|x|^p f(x)$ in the integral or sum we need to check, we see that it behaves like $|x|^{p-2}$ at infinity. (More formally, the limit of the integrand or summand divided by $|x|^{p-2}$ converges to a constant as x goes to plus or minus infinity. Hence by Lemma 2.42, the expectation exists if and only if the integral or sum of $|x|^{p-2}$ exists.) By Lemma 2.39 the integral or sum exists if and only if $p - 2 < -1$, that is, $p < 1$.

To do problems involving singularities, we need another lemma analogous to Lemma 2.42. This lemma involves only integrals not sums because sequences cannot go to infinity except at infinity (all the terms are actually finite).

Lemma 2.43. *Suppose g and h are strictly positive functions on an interval (a, b) and both have singularities at a but are bounded elsewhere, and suppose*

$$\lim_{x \to a} \frac{g(x)}{h(x)} = k,$$

where k is a strictly positive constant, then either both of the integrals

$$\int_a^b g(x)\,dx \qquad \textit{and} \qquad \int_a^b h(x)\,dx$$

are finite, or neither is.

Example 2.5.6 (The Gamma Distribution Again).
The the Gamma distribution is the continuous distribution having density

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \qquad x > 0$$

(Section B.2.3 of Appendix B). For $X \sim \text{Gam}(\alpha, \lambda)$, we consider here when X^p is in L^1 for any real number p, positive or negative. The integral that defines the expectation is

$$E(X^p) = \int_0^\infty x^p \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \, dx = \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha+p-1} e^{-\lambda x} \, dx$$

if the integral exists (which is the question we are examining).

From Lemma 2.41, the integral over (a, ∞) exists for for any $a > 0$ and any p positive or negative. The only issue is the possible singularity of the integrand at the origin. There is a singularity if $\alpha + p - 1 < 0$. Otherwise the integrand is bounded and the expectation exists.

Since $e^0 = 1$, the integrand behaves like $x^{\alpha+p-1}$ at zero and according to Lemma 2.43 this is integrable over a neighborhood of zero if and only if $\alpha + p - 1 > -1$, that is, if and only if $p > -\alpha$.

2.5.10 L^p Spaces

We start with another consequence of the domination theorem and the methods for telling when expectations exist developed in the preceding section.

Theorem 2.44. *If X is a real-valued random variable and $|X - a|^p$ is in L^1 for some constant a and some $p \geq 1$, then*

$$|X - b|^q \in L^1,$$

for any constants b and any q such that $1 \leq q \leq p$.

Proof. First the case $q = p$. The ratio of the integrands defining the expectations of $|X - a|^p$ and $|X - b|^p$ converges, that is

$$\frac{|x - b|^p f(x)}{|x - a|^p f(x)} = \left| \frac{x - b}{x - a} \right|^p$$

goes to 1 as x goes to plus or minus infinity. Thus both integrals exist, and $|X - b|^p \in L^1$.

In the case $q < p$, the ratio of integrands

$$\frac{|x - b|^q f(x)}{|x - a|^p f(x)} = \frac{|x - b|^q}{|x - a|^p}$$

converges to zero as x goes to plus or minus infinity. Again this implies both integrals exist and $|X - b|^p \in L^1$. □

Definition 2.5.1 (L^p Spaces).
For any $p \geq 1$, the set of random variables X such that $|X|^p \in L^1$ is called L^p.

With this definition, we can rephrase the theorem. The condition of the theorem can now be stated concisely as $X \in L^p$, because if $|X - a|^p \in L^1$, then the theorem implies $|X|^p \in L^1$ too, which is the same as $X \in L^p$. The conclusion of the theorem can also be restated as $X \in L^q$. Hence $L^1 \supset L^q \supset L^p$ when $1 \leq q \leq p$.

The reason for the name "L^p space" is the following theorem, which we will not prove.

Theorem 2.45. *Each L^p is a vector space.*

What the theorem says is that L^p is closed under addition and scalar multiplication, that is,

$$X \in L^p \text{ and } Y \in L^p \text{ implies } X + Y \in L^p$$

and

$$X \in L^p \text{ and } a \in \mathbb{R} \text{ implies } aX \in L^p.$$

All of this having been said, I have to admit that the main use of the L^p concept at this level is purely as a shorthand. L^2 is the set of random variables having variances. By Theorem 2.44 and the following comment $L^1 \supset L^2$ so these random variables also have means. Thus we could have stated the condition "X is a random variable having first and second moments" in Corollary 2.12 and succeeding theorems about second moments much more concisely as "$X \in L^2$." Whether you like the shorthand or not is a matter of taste. One thing, though, that we did learn in this section is that the words "first and" could have been deleted from the condition of Corollary 2.12 and theorems with similar conditions. If second moments exist, then so do first moments by Theorem 2.44.

2.6 Probability is a Special Case of Expectation

A special kind of random variable is the *indicator function* (or indicator random variable) of an event A (a random variable is a function on the sample space, so an indicator function is a random variable). This is denoted I_A and defined by

$$I_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & \omega \notin A \end{cases}$$

The indicator function characterizes the set A. It is the set of points ω such that $I_A(\omega) = 1$. More importantly from our point of view, indicator functions connect probability and expectation. The relation

$$P(A) = E(I_A) \tag{2.55}$$

holds for all events A. Probability is just expectation of indicator functions. Thus probability is a dispensable concept. It is just a special case of expectation.

The proof of (2.55) for discrete probability models is trivial.

$$\begin{aligned} E(I_A) &= \sum_{\omega \in \Omega} I_A(\omega) p(\omega) \\ &= \sum_{\omega \in A} p(\omega) \\ &= P(A) \end{aligned}$$

The first equality is the definition (2.40), the third is the definition of probability (p. 30 in Lindgren), and the middle equality just uses the definition of indicator functions: terms for $\omega \in A$ have $I_A(\omega) = 1$ and terms for $\omega \notin A$ have $I_A(\omega) = 0$ and can be dropped from the sum. The proof of (2.55) for continuous probability models is the same except that we replace sums by integrals.

All of the probability axioms can be derived from the expectation axioms by just taking the special case when the random variables are indicator functions. Since indicator functions are nonnegative, Axiom E1 implies

$$E(I_A) = P(A) \geq 0$$

which is the first probability axiom. Axiom E2 implies

$$E(1) = E(I_\Omega) = P(\Omega) = 1$$

which is the second probability axiom. The sum of indicator functions is not necessarily an indicator function, in fact

$$I_A + I_B = I_{A \cup B} + I_{A \cap B}. \tag{2.56}$$

This is easily verified by checking the four possible cases, ω in or not in A and in or not in B. Applying Axiom E4 to both sides of (2.56) gives

$$\begin{aligned} P(A) + P(B) &= E(I_A) + E(I_B) \\ &= E(I_{A \cup B}) + E(I_{A \cap B}) \\ &= P(A \cup B) + P(A \cap B) \end{aligned}$$

which is the general addition rule for probabilities and implies the third probability axiom, which is the special case $A \cap B = \varnothing$.

The countable additivity axiom is applied by the monotone convergence. A nondecreasing sequence of indicator functions corresponds to a nondecreasing sequence of sets. Hence Axiom E5 implies

$$P(A_n) \uparrow P(A), \qquad \text{whenever } A_n \uparrow A$$

This statement, continuity of probability, implies countable additivity (just run the proof on p. 29 in Lindgren backwards).

2.7 Independence

2.7.1 Two Definitions

Lindgren (p. 79, equation (3)) gives the following as a definition of independent random variables.

Definition 2.7.1 (Independent Random Variables).
Random variables X and Y are independent *if*

$$P(X \in A \, \text{and} \, Y \in B) = P(X \in A)P(Y \in B). \tag{2.57}$$

for every event A in the range of X and B in the range of Y.

We take a quite different statement as the definition.

Definition 2.7.2 (Independent Random Variables).
Random variables X and Y are independent *if*

$$E\{g(X)h(Y)\} = E\{g(X)\}E\{h(Y)\} \tag{2.58}$$

for all real-valued functions g and h such that these expectations exist.

These two definitions are equivalent—meaning they define the same concept. That means that we could take either statement as the definition and prove the other. Lindgren takes (2.57) as the definition and "proves" (2.58). This is Theorem 11 of Chapter 4 in Lindgren. But the "proof" contains a lot of hand waving. A correct proof is beyond the scope of this course.

That's one reason why we take Definition 2.7.2 as the definition of the concept. Then Definition 2.7.1 describes the trivial special case of Definition 2.7.2 in which the functions in question are indicator functions, that is, (2.57) says exactly the same thing as

$$E\{I_A(X)I_B(Y)\} = E\{I_A(X)\}E\{I_B(Y)\}. \tag{2.59}$$

only in different notation. Thus if we take Definition 2.7.2 as the definition, we easily (trivially) prove (2.57). But the other way around, the proof is beyond the scope of this course.

2.7.2 The Factorization Criterion

Theorem 2.46 (Factorization Criterion). *A finite set of real-valued random variables is independent if and only if their joint distribution is the product of the marginals.*

What this says is that X_1, $\ldots$, X_n are independent if and only if

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i) \tag{2.60}$$

One direction of the theorem is easy to establish. If (2.60) holds

$$\begin{aligned} E\left\{\prod_{i=1}^n g_i(X_i)\right\} &= \int \cdots \int \left(\prod_{i=1}^n g_i(x_i) f_{X_i}(x_i)\right) dx_1 \cdots dx_n \\ &= \prod_{i=1}^n \int g_i(x_i) f_{X_i}(x_i) dx_i \\ &= \prod_{i=1}^n E\left\{g_i(X_i)\right\} \end{aligned}$$

So the X_i are independent. The proof of the other direction of the theorem is beyond the scope of this course.

The simple statement of Theorem 2.46 assumes the marginal densities are defined on the whole real line If necessary, they are extended by zero off the supports of the variables.

> *It is not enough to look only at the formulas defining the densities. You must also look at the domains of definition.*

The following example shows why.

Example 2.7.1 (A Cautionary Example).
The random variables X and Y having joint density

$$f(x, y) = 4xy, \qquad 0 < x < 1 \text{ and } 0 < y < 1 \tag{2.61}$$

are independent, but the random variables X and Y having joint density

$$f(x, y) = 8xy, \qquad 0 < x < y < 1 \tag{2.62}$$

are not! For more on this, see Problem 2-35.

The difference is easy to miss. The formulas defining the densities are very similar, both factor as a function of x times a function of y. The difference is in the domains of definition. The one for which the factorization criterion holds is a rectangle with sides parallel to the axes. The other isn't.

2.7.3 Independence and Correlation

Theorem 2.47. *Independent random variables are uncorrelated.*

The converse is false!

Example 2.7.2.
Suppose X is a nonconstant random variable having a distribution symmetric about zero, and suppose $Y = X^2$ is also nonconstant. For example, we could take $X \sim \mathcal{U}(-1, 1)$, but the details of the distribution do not matter, only that it is symmetric about zero and nonconstant and that X^2 also has a nonconstant distribution.

Then X and Y are uncorrelated (Problem 2-37) but not independent. Independence would require that

$$E\{g(X)h(Y)\} = E\{g(X)\}E\{h(Y)\}$$

hold for *all* functions g and h. But it obviously does not hold when, to pick just one example, g is the squaring function and h is the identity function so $g(X) = Y$ and $h(Y) = Y$, because no nonconstant random variable is independent of itself.[2]

Problems

2-27. Suppose $X \sim \text{Bin}(n, p)$.

(a) Show that

$$E\{X(X-1)\} = n(n-1)p^2$$

Hint: Follow the pattern of Example 2.5.1.

(b) Show that

$$\text{var}(X) = np(1-p).$$

Hint: Use part (a).

2-28. Suppose $X \sim \text{Poi}(\mu)$.

(a) Show that

$$E\{X(X-1)\} = \mu^2$$

Hint: Follow the pattern of Example 2.5.2.

(b) Show that

$$\text{var}(X) = \mu.$$

Hint: Use part (a).

2-29. Verify the moments of the $\mathcal{DU}(1, n)$ distribution given in Section B.1.1 of Appendix B.
Hint: First establish

$$\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$$

by mathematical induction.

[2] Bizarrely, constant random variables are independent of all random variables, including themselves. This is just the homogeneity axiom and the "expectation of a constant is the constant" property:

$$E\{g(a)h(X)\} = g(a)E\{h(X)\} = E\{g(a)\}E\{h(X)\}$$

for any constant a and random variable X.

2-30. Verify the moments of the $\mathcal{U}(a, b)$ distribution given in Section B.2.1 of Appendix B.

2-31. The proof of Corollary 2.35 used $\operatorname{cor}(X, X) = 1$ without comment. Prove this.

2-32. Suppose $X \sim \operatorname{Gam}(\alpha, \lambda)$.

(a) For any real number $p > -\alpha$, the p-th ordinary moment

$$\alpha_p = E(X^p)$$

exists. Calculate it.

Hint: Follow the pattern of Example 2.5.3. Your answer will involve gamma functions that cannot be simplified using the recurrence relation if p is not an integer (which we didn't say it was).

(b) Show that

$$\operatorname{var}(X) = \frac{\alpha}{\lambda^2}$$

Hint: Use part (a) and the recurrence relation for gamma functions, (B.3) in Appendix B.

2-33. Suppose X has probability density

$$f(x) = \frac{3}{x^4}, \qquad x > 1$$

(note the domain).

(a) For what positive integers k is X^k in L^1?

(b) Calculate $E(X^k)$ for the positive integers k such that the expectation exists.

2-34. Suppose X has probability density

$$f(x) = \frac{1}{2\sqrt{x}}, \qquad 0 < x < 1$$

(note the domain).

(a) For what positive integers k is X^k in L^1?

(b) Calculate $E(X^k)$ for the positive integers k such that the expectation exists.

2-35. Calculate the marginal distributions for

(a) the density (2.61) and

(b) the density (2.62).

Show that the factorization criterion

(c) holds for the density (2.61) and

(d) fails for the density (2.62).

2-36. Prove Theorem 2.47.

2-37. This fills in some details left unsaid in Example 2.7.2.

(a) Prove that X and Y defined in Example 2.7.2 are uncorrelated.

Hint: Use Theorem 2.10.

(b) Prove that no nonconstant random variable is independent of itself.

Hint: If all we know is that X is nonconstant, then all we know is that there exists an event A such that $0 < P(X \in A) < 1$. Now use Definition 2.7.1.

2-38. Prove the following identities. For any $n \geq 1$

$$\mu_n = \sum_{k=0}^{n} \binom{n}{k} (-1)^k \alpha_1^k \alpha_{n-k}$$

and

$$\alpha_n = \sum_{k=0}^{n} \binom{n}{k} \alpha_1^k \mu_{n-k}$$

where, as defined in Section 2.4, μ_k is the k-th central moment and α_k is the k-th ordinary moment.
Hint: Use the binomial theorem (Problem 1-14 on p. 7 of Lindgren).

Chapter 3

Conditional Probability and Expectation

3.1 Parametric Families of Distributions

Scalar Variable and Parameter

Sometimes, like in the "brand name distributions" in Appendix B of these notes, we consider probability models having an adjustable constant in the formula for the density. Generically, we refer to such a constant as a *parameter* of the distribution. Usually, though not always, we use Greek letters for parameters to distinguish them from random variables (large Roman letters) and possible values of random variables (small Roman letters). A lot of different Greek letters are used for parameters (check out Appendix B), the Greek letter used for a "generic" parameter (when we are talking generally, not about any particular distribution) is θ (lower case theta, see Appendix A).

When we want to emphasize the dependence of the density on the parameter, we write f_θ or $f(\,\cdot \mid \theta)$ rather than just f for the density function and $f_\theta(x)$ or $f(x \mid \theta)$ for the value of the density function at the point x. Why two notations? The former is simpler and a good deal less clumsy in certain situations, but the latter shows explicitly the close connection between conditional probability and parametric families, which is the subject of this section and the following section.

Thus we say: let X be a random variable having density f_θ on a sample space S. This means that for each particular value of the parameter θ the function f_θ is a density, that is,

$$f_\theta(x) \ge 0, \qquad x \in S \tag{3.1a}$$

and

$$\int f_\theta(x)\,dx = 1 \tag{3.1b}$$

(with, as usual, the integral replaced by a sum in the discrete case). Note that this is exactly the usual condition for a function to be a probability density, just like (2.47a) and (2.47b). The *only* novelty is writing f_θ in place of f. If you prefer the other notation, this condition would become

$$f(x \mid \theta) \geq 0, \qquad x \in S \tag{3.2a}$$

and

$$\int f(x \mid \theta) \, dx = 1 \tag{3.2b}$$

Again, there is no novelty here except for the purely notational novelty of writing $f(x \mid \theta)$ instead of $f_\theta(x)$ or $f(x)$.

Example 3.1.1 (The Exponential Distribution).
We want to write the exponential distribution (Section B.2.2 in Appendix B) in the notation of parametric families. The parameter is λ. We write the density as

$$f_\lambda(x) = \lambda e^{-\lambda x}, \qquad x > 0$$

or as

$$f(x \mid \lambda) = \lambda e^{-\lambda x}, \qquad x > 0$$

the only difference between either of these or the definition in Section B.2.2 being the notation on the left hand side: $f(x)$ or $f_\lambda(x)$ or $f(x \mid \lambda)$.

Each different value of the parameter θ gives a different probability distribution. As θ ranges over its possible values, which we call the *parameter space*, often denoted Θ when the parameter is denoted θ, we get a *parametric family* of densities

$$\{ \, f_\theta : \theta \in \Theta \, \}$$

although we won't see this notation much until we get to statistics next semester.

Vector Variable or Parameter

Vector Variable

Another purely notational variant involves random vectors. We typically indicate vector variables with boldface type, as discussed in Section 1.3 of these notes, that is, we would write $f(\mathbf{x})$ or $f_\theta(\mathbf{x})$ or $f(\mathbf{x} \mid \theta)$. As usual we are sloppy about whether these are functions of a single vector variable $\mathbf{x} = (x_1, \ldots, x_n)$ or of many scalar variables x_1, …, x_n. When we are thinking in the latter mode, we write $f(x_1, \ldots, x_n)$ or $f_\theta(x_1, \ldots, x_n)$ or $f(x_1, \ldots, x_n \mid \theta)$.

Example 3.1.2 (The Exponential Distribution).
Suppose X_1, …, X_n are independent and identically distributed $\text{Exp}(\lambda)$ random

variables. We write the density of the random vector $\mathbf{X} = (X_1, \ldots, X_n)$ as

$$\begin{aligned} f_\lambda(\mathbf{x}) &= \prod_{i=1}^n \lambda e^{-\lambda x_i} \\ &= \lambda^n \exp\left(-\lambda \sum_{i=1}^n x_i\right), \qquad x_i > 0,\ i = 1, \ldots, n. \end{aligned}$$

or, according to taste, we might write the left hand side as $f_\lambda(x_1, \ldots, x_n)$ or $f(\mathbf{x} \mid \lambda)$ or $f(x_1, \ldots, x_n \mid \lambda)$.

Vector Parameter

Similarly, when we have a vector parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$, we write the density as $f_{\boldsymbol{\theta}}(x)$ or $f(x \mid \boldsymbol{\theta})$. And, as usual, we are sloppy about whether there is really one vector parameter or several scalar parameters θ_1, …, θ_m. When we are thinking in the latter mode, we write $f_{\theta_1,\ldots,\theta_m}(x)$ or $f(x \mid \theta_1, \ldots, \theta_m)$.

Example 3.1.3 (The Gamma Distribution).
We want to write the gamma distribution (Section B.2.3 in Appendix B) in the notation of parametric families. The parameter is $\boldsymbol{\theta} = (\alpha, \lambda)$. We write the density as

$$f_{\boldsymbol{\theta}}(x) = f_{\alpha,\lambda}(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \qquad x > 0$$

or if we prefer the other notation we write the left hand side as $f(x \mid \boldsymbol{\theta})$ or $f(x \mid \alpha, \lambda)$.

The parameter space of this probability model is

$$\Theta = \{\, (\alpha, \lambda) \in \mathbb{R}^2 : \alpha > 0,\ \lambda > 0 \,\}$$

that is, the first quadrant with boundary points excluded.

Vector Variable and Vector Parameter

And, of course, the two preceeding cases can be combined. If we have a vector random variable $\mathbf{X} = (X_1, \ldots, X_n)$ and a vector parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$, we can write write the density as any of

$$\begin{gathered} f_{\boldsymbol{\theta}}(\mathbf{x}) \\ f(\mathbf{x} \mid \boldsymbol{\theta}) \\ f_{\theta_1,\ldots,\theta_m}(x_1, \ldots, x_n) \\ f(x_1, \ldots, x_n \mid \theta_1, \ldots, \theta_m) \end{gathered}$$

according to taste.

3.2 Conditional Probability Distributions

Scalar Variables

The conditional probability distribution of one random variable Y given another X is the probability model you are supposed to use in the situation when you have seen X and know its value but have not yet seen Y and don't know its value. The point is that X is no longer random. Once you know its value x, it's a constant not a random variable.

We write the density of this probability model, the *conditional distribution of Y given X* as $f(y \mid x)$. We write expectations with respect to this model as $E(Y \mid x)$, and we write probabilities as

$$P(Y \in A \mid x) = E\{I_A(Y) \mid x\}$$

(couldn't resist an opportunity to reiterate the lesson of Section 2.6 that probability is a special case of expectation).

We calculate probabilities or expectations from the density in the usual way with integrals in the continuous case

$$E\{g(Y) \mid x\} = \int g(y) f(y \mid x) \, dy \tag{3.3}$$

$$P\{Y \in A \mid x\} = \int_A f(y \mid x) \, dy \tag{3.4}$$

and with the integrals replaced by sums in the discrete case.

Note that

> *A conditional probability density is just an ordinary probability density, when considered as a function of the variable(s) in front of the bar alone with the variable(s) behind the bar considered fixed.*

This means that in calculating a conditional probability or expectation from a conditional density

> *always integrate with respect to the variable(s) in front of the bar*

(with, of course, "integrate" replaced by "sum" in the discrete case).

Example 3.2.1 (Exponential Distribution).
Of course, one doesn't always have to do an integral or sum, expecially when a "brand name" distribution is involved. Suppose the conditional distribution of Y given X is $\text{Exp}(X)$, denoted

$$Y \mid X \sim \text{Exp}(X)$$

for short. This means, of course, that the conditional density is

$$f(y \mid x) = x e^{-xy}, \qquad y > 0$$

(just plug in x for λ in the formula in Section B.2.2 in Appendix B), but we don't need to use the density to calculate the conditional expectation, because we know that the mean of the Exp(λ) distribution is $1/\lambda$, hence (again just plugging in x for λ

$$E(Y \mid x) = \frac{1}{x}$$

or

$$E(Y \mid X) = \frac{1}{X}$$

depending on whether we are thinking of the variable behind the bar as random (big X) or fixed (little x) As we shall see, both viewpoints are useful and we shall use both in different situations.

If the known formulas for a "brand name" distribution don't answer the question, then we do need an integral

$$\begin{aligned} P(a < Y < b \mid x) &= \int_a^b f(y \mid x) \, dy \\ &= \int_a^b x e^{-xy} \, dy \\ &= -e^{-xy} \Big|_a^b \\ &= e^{-xa} - e^{-xb} \end{aligned}$$

and, of course, if we are thinking of X as being random too, we would write

$$P(a < Y < b \mid X) = e^{-aX} - e^{-bX}$$

just the same except for big X instead of little x.

The astute reader will by now have understood from the hint given by the notation why this chapter started with a section on the seemingly unrelated topic of parametric families of distributions.

> *Conditional probability distributions are no different from parametric families of distributions.*

For each fixed value of x, the conditional density $f(y \mid x)$, considered as a function of y alone, is just an ordinary probability density. Hence it satisfies the two properties

$$f(y \mid x) \geq 0, \qquad \text{for all } y \tag{3.5a}$$

and

$$\int f(y \mid x) \, dy = 1 \tag{3.5b}$$

(with the integral replaced by a sum in the discrete case). Notice that there is no difference, except a purely notational one, between the pair of conditions (3.5a) and (3.5b) and the pair of conditions (3.2a) and (3.2b). Here we have a

Roman letter behind the bar; there we had a Greek letter behind the bar, but (mathematics is invariant under changes of notation) that makes no *conceptual* difference whatsoever.

The fact that conditional probability is a special case of ordinary probability (when we consider the variable or variables behind the bar fixed) means that we already know a lot about conditional probability. Every fact we have learned so far in the course about *ordinary* probability and expectation applies to its special case *conditional* probability and expectation. **Caution:** What we just said applies *only* when the variable(s) behind the bar are considered fixed. As we shall see, things become more complicated when both are treated as random variables.

Vector Variables

Of course, either of the variables involved in a conditional probability distribution can be vectors. Then we write either of

$$f(\mathbf{y} \mid \mathbf{x})$$
$$f(y_1, \ldots y_n \mid x_1, \ldots x_m)$$

according to taste, and similarly either of

$$E(\mathbf{Y} \mid \mathbf{x})$$
$$E(Y_1, \ldots Y_n \mid x_1, \ldots x_m)$$

Since we've already made this point in the context of parametric families of distributions, and conditional probability distributions are no different, we will leave it at that.

3.3 Axioms for Conditional Expectation

The conditional expectation $E(Y \mid x)$ is just another expectation operator, obeying all the axioms for expectation. This follows from the view explained in the preceeding section that conditional expectation is a special case of ordinary unconditional expectation (at least when we are considering the variable or variables behind the bar fixed). If we just replace unconditional expectations with conditional expectations everywhere in the axioms for unconditional expectation, they are still true.

There are, however, a couple of additional axioms for conditional expectation. Axiom E2 can be strengthened (as described in the next section), and an entirely new axiom (described in the two sections following the next) can be added to the set of axioms.

3.3.1 Functions of Conditioning Variables

Any function of the variable or variables behind the bar (the *conditioning* variables) behaves like a constant in conditional expectations.

Axiom CE1. *If Y is in L^1 and a is any function, then*

$$E\{a(X)Y \mid X\} = a(X)E(Y \mid X).$$

We don't have to verify that conditional expectation obeys the axioms of ordinary unconditional expectation, because conditional expectation is a special case of unconditional expectation (when thought about the right way), but this axiom isn't a property of unconditional expectation, so we do need to verify that it holds for conditional expectation as we have already defined it. But the verification is easy.

$$\begin{aligned} E\{a(X)Y \mid X\} &= \int a(X) y f(y \mid X) \, dy \\ &= a(X) \int y f(y \mid X) \, dy \\ &= a(X)E(Y \mid X) \end{aligned}$$

because any term that is not a function of the variable of integration can be pulled outside the integral (or sum in the discrete case).

Two comments:

- We could replace big X by little x if we want

  $$E\{a(x)Y \mid x\} = a(x)E(Y \mid x)$$

  though, of course, this now follows from Axiom E2 of ordinary expectation because $a(x)$ is a constant when x is a constant.

- We could replace big Y by any random variable, for example, $g(Y)$ for any function g, obtaining

  $$E\{a(X)g(Y) \mid X\} = a(X)E\{g(Y) \mid X\}.$$

3.3.2 The Regression Function

It is now time to confront squarely an issue we have been tiptoeing around with comments about writing $E(Y \mid x)$ or $E(Y \mid X)$ "according to taste." In order to clearly see the contrast with unconditional expectation, let first review something about ordinary unconditional expectation.

$E(X)$ is not a function of X. It's a constant, not a random variable.

This doesn't conflict with the fact that an expectation operator is a function $E : L^1 \to \mathbb{R}$ when considered abstractly. This is the usual distinction between a function and it's values: E is indeed a function (from L^1 to $\mathbb{R}$), but $E(X)$ isn't a function, it's the value that the expectation operator assigns to the random variable X, and that value is a real number, a constant, not a random variable (not a function on the sample space).

So $E(X)$ is very different from $g(X)$, where g is an ordinary function. The latter is a random variable (any function of a random variable is a random variable).

So what's the corresponding fact about conditional expectation?

> $E(Y \mid X)$ *is not a function of* Y*, but it is a function of* X*, hence a random variable.*

We saw this in Example 3.2.1

$$Y \mid X \sim \operatorname{Exp}(X)$$

implies

$$E(Y \mid X) = \frac{1}{X}$$

which is, apparently, a function of X and not a function of Y.

In a way, there is nothing surprising here. If we consider the conditioning variable fixed, then $E(Y \mid x)$ is just a special case of ordinary expectation. Hence $E(Y \mid x)$ is not a function of Y any more than $E(Y)$ is. Furthermore, $E(Y \mid x)$ is not a random variable because x isn't a random variable (little x).

In another way, this is surprising. If we consider the conditioning variable to be random, then it no longer looks like conditional expectation is a special case of ordinary expectation, because the former is a random variable and the latter isn't! What happens is that which is a special case of which gets turned around.

> *Unconditional expectation is the special case of conditional expectation obtained by conditioning on an empty set of variables.*

This accords with the naive view that a conditional probability model for Y given X is what you use when you have seen X but not yet seen Y. Clearly, what you use when you have seen (nothing) but not yet seen Y is the the ordinary unconditional models we have been using all along. It says that $E(Y)$ can be thought of as $E(Y \mid \quad)$ with nothing behind the bar. Applying our other slogan to this special case we see that

> $E(Y) = E(Y \mid \quad)$ *is not a function of* Y*, but it is a function of (nothing), hence a constant random variable.*

Thus when we think of unconditional expectation as a special case of conditional expectation $E(Y)$ isn't a constant but a constant random variable, which is almost the same thing—only a mathematician and a rather pedantic one could care about the difference.

So we have two somewhat conflicting views of conditional probability and expectation.

- When we consider the conditioning variables (the variables behind the bar) fixed, conditional expectation is just a special case of ordinary unconditional expectation. The conditioning variables behave like parameters of the probability model.

- When we consider the conditioning variables (the variables behind the bar) random, unconditional expectation is just a special case of conditional expectation, what happens when we condition on an empty set of variables.

What's to blame for the confusion is partly just the notation, it's not clear from the notation that $E(Y \mid X)$ is a function of X but not a function of Y, and partly the real conflict between seeing the conditioning variable sometimes as random and sometimes as constant. There's nothing to be done about the second problem except to be very careful to always understand which situation you are in. For the first, we can change terminology and notation.

If $E(Y \mid X)$ is a function of X, we can write it as a function of X, say $g(X)$. In Example 3.2.1 we had

$$E(Y \mid X) = g(X) = \frac{1}{X}$$

which means that g is the function defined by

$$g(x) = \frac{1}{x}, \qquad x > 0$$

just an ordinary function of an ordinary variable, that is, g is an ordinary function, and $g(x)$ is an ordinary number, but, of course, $g(X)$ is a random variable (because of the big X).

Another name for this function g is the *regression function of Y on X*. When it's clear from the context which is the conditioning variable and which is the other variable, we can say just *regression function*. But when any confusion might arise, the longer form is essential. The regression function of Y on X, that is, $E(Y \mid X)$ is quite different from the regression function of X on Y, that is, $E(X \mid Y)$. For one thing, the former is a function of X and the latter is a function of Y. But not only that, they are in general quite different and unrelated functions.

3.3.3 Iterated Expectations

We saw in the preceding section that $E(Y \mid X)$ is a random variable, a function of X, say $g(X)$. This means we can take its expectation

$$E\{g(X)\} = E\{E(Y \mid X)\}.$$

The left hand side is nothing unusual, just an expectation like any other. The right hand side looks like something new. We call it an "iterated expectation" (an unconditional expectation of a conditional expectation). Iterated expectation has a very important property which is the last axiom for conditional probability.

Axiom CE2. *If $Y \in L^1$, then*

$$E\{E(Y \mid X)\} = E(Y). \tag{3.6}$$

A proof that the notion of conditional expectation we have so far developed satisfies this axiom will have to wait until the next section. First we give some examples and consequences.

Example 3.3.1 (Random Sum of Random Variables).
Suppose X_0, X_1, $\ldots$ is an infinite sequence of identically distributed random variables, having mean $E(X_i) = \mu_X$, and suppose N is a nonnegative integer-valued random variable independent of the X_i and having mean $E(N) = \mu_N$. It is getting a bit ahead of ourselves, but we shall see in the next section that this implies

$$E(X_i \mid N) = E(X_i) = \mu_X. \tag{3.7}$$

Question: What is the expectation of

$$S_N = X_1 + \cdots + X_N$$

(a sum with a random number N of terms and each term X_i a random variable) where the sum with zero terms when $N = 0$ is defined to be zero?

Linearity of expectation, which applies to conditional as well as unconditional probability, implies

$$\begin{aligned} E(S_N \mid N) &= E(X_1 + \cdots X_n \mid N) \\ &= E(X_1 \mid N) + \cdots + E(X_n \mid N) \\ &= E(X_1) + \cdots + E(X_N) \\ &= N\mu_X \end{aligned}$$

the next to last equality being (3.7). Hence by the iterated expectation axiom

$$E(S_N) = E\{E(S_N \mid N)\} = E(N\mu_X) = E(N)\mu_X = \mu_N\mu_X.$$

Note that this example is impossible to do any other way than using the iterated expectation formula. Since no formulas were given for any of the densities, you can't use any formula involving explicit integrals.

If we combine the two conditional probability axioms, we get the following.

Theorem 3.1. *If X and Y are random variables and g and h are functions such that $g(X)$ and $h(Y)$ are in L^1, then*

$$E\{g(X)E[h(Y) \mid X]\} = E\{g(X)h(Y)\}. \tag{3.8}$$

Proof. Replace Y by $g(X)h(Y)$ in Axiom CE2 obtaining

$$E\{E[g(X)h(Y) \mid X]\} = E\{g(X)h(Y)\}.$$

then apply Axiom CE1 to pull $g(X)$ out of the inner conditional expectation obtaining (3.8). □

The reader should be advised that our treatment of conditional expectation is a bit unusual. Rather than state two axioms for conditional expectation, standard treatments in advanced probability textbooks give just one, which is essentially the statement of this theorem. As we have just seen, our two axioms imply this one, and conversely our two axioms are special cases of this one: taking $g = a$ and h the identity function in (3.8) gives our Axiom CE1, and taking $g = 1$ and h the identity function in (3.8) gives our Axiom CE2. Thus our treatment characterizes the same notion of conditional probability as standard treatments.

Another aspect of advanced treatments of conditional probability is that standard treatments usually take the statement Theorem 3.1 as a *definition* rather than an *axiom*. The subtle difference is the following uniqueness assertion.

Theorem 3.2. *If X and Y are random variables and h is a function such that $h(Y) \in L^1$, then then there exists a function f such that $f(X) \in L^1$ and*

$$E\{g(X)f(X)\} = E\{g(X)h(Y)\} \tag{3.9}$$

for every function g such that $g(X)h(Y) \in L^1$. The function f is unique up to redefinition on sets of probability zero.

The proof of this theorem is far beyond the scope of this course. Having proved this theorem, advanced treatments take it as a definition of conditional expectation. The unique function f whose existence is guaranteed by the theorem is defined to be the conditional expectation, that is,

$$E\{h(Y) \mid X\} = f(X).$$

The theorem makes it clear that (as everywhere else in probability theory) redefinition on a set (event) of probability zero makes no difference.

Although we cannot prove Theorem 3.2, we can use it to prove a fancy version of the iterated expectation formula.

Theorem 3.3. *If $Y \in L^1$, then*

$$E\big\{E(Z \mid X, Y) \bigm| X\big\} = E(Z \mid X). \tag{3.10}$$

Of course, the theorem also holds when the conditioning variables are vectors, that is, if $m < n$

$$E\big\{E(Z \mid X_1, \ldots, X_n) \bigm| X_1, \ldots X_m\big\} = E(Z \mid X_1, \ldots, X_m).$$

In words, an iterated conditional expectation (a conditional expectation inside another conditional expectation) is just the conditional expectation conditioning on the set of variables of the outer conditional expectation, if the set of conditioning variables in the outer expectation is a *subset* of the conditioning variables in the inner expectation. That's a mouthful. The formula (3.10) is simpler.

Proof of Theorem 3.3. By Theorem 3.2 and the following comment,

- $E(Z \mid X, Y)$ is the unique (up to redefinition on sets of probability zero) function $f_1(X, Y)$ such that

  $$E\{g_1(X, Y)f_1(X, Y)\} = E\{g_1(X, Y)Z\} \tag{3.11a}$$

  for all functions g_1 such that $g_1(X, Y)Z \in L^1$.

- The iterated expectation on the left hand side of (3.10) is the unique (up to redefinition on sets of probability zero) function $f_2(X)$ such that

  $$E\{g_2(X)f_2(X)\} = E\{g_2(X)f_1(X, Y)\} \tag{3.11b}$$

  for all functions g_2 such that $g_2(X)f_1(X, Y) \in L^1$.

- $E(Z \mid X)$ is the unique (up to redefinition on sets of probability zero) function $f_3(X)$ such that

  $$E\{g_3(X)f_3(X)\} = E\{g_3(X)Z\} \tag{3.11c}$$

  for all functions g_3 such that $g_3(X)Z \in L^1$.

Since (3.11a) holds for any function g_1, it holds when $g_1(X, Y) = g_3(X)$, from which, combining (3.11a) and (3.11c), we get

$$E\{g_3(X)f_3(X)\} = E\{g_3(X)Z\} = E\{g_3(X)f_1(X, Y)\} \tag{3.11d}$$

Reading (3.11d) from end to end, we see it is the same as (3.11b), because (3.11d) must hold for any function g_3 and (3.11b) must hold for any function g_2. Thus by the uniqueness assertion of Theorem 3.2 we must have $f_2(X) = f_3(X)$, except perhaps on a set of probability zero (which does not matter). Since $f_2(X)$ is the left hand side of (3.10) and $f_3(X)$ is the right hand side, that is what was to be proved. □

Theorem 3.2 can also be used to prove a very important fact about independence and conditioning.

Theorem 3.4. *If X and Y are independent random variables and h is a function such that $h(Y) \in L^1$, then*

$$E\{h(Y) \mid X\} = E\{h(Y)\}.$$

In short, conditioning on an *independent* variable or variables is the same as conditioning on *no* variables, making conditional expectation the same as unconditional expectation.

Proof. If X and Y are independent, the right hand side of (3.9) becomes $E\{g(X)\}E\{h(Y)\}$ by Definition 2.7.2. Hence, in this special case, Theorem 3.2 asserts that $E\{h(Y) \mid X\}$ is the unique function $f(X)$ such that

$$E\{g(X)f(X)\} = E\{g(X)\}E\{h(Y)\}$$

whenever $g(X) \in L^1$. Certainly the constant $f(X) = a$, where $a = E\{h(Y)\}$ is one such function, because

$$E\{g(X)a\} = E\{g(X)\}a = E\{g(X)\}E\{h(Y)\}$$

so by the uniqueness part of Theorem 3.2 this is the conditional expectation, as was to be proved. □

3.4 Joint, Conditional, and Marginal

As was the case with unconditional expectation, our “axioms first” treatment of conditional expectation has been a bit abstract. When the problem is solved by pulling a function of the conditioning variables outside of a conditional expectation or by the iterated expectation formula, either the special case in Axiom CE2 with the outside expectation an unconditional one or the general case in Theorem 3.3 in which both expectations are conditional, then the axioms are just what you need. But for other problems you need to be able to calculate conditional probability densities and expectations by doing sums and integrals, and that is the subject to which we now turn.

3.4.1 Joint Equals Conditional Times Marginal

Note that the iterated expectation axiom (Axiom CE2), when we write out the expectations as integrals, equates

$$\begin{aligned} E\{E(Y \mid X)\} &= \int \left(\int y f(y \mid x) \, dy \right) f_X(x) \, dx \\ &= \iint y f(y \mid x) f_X(x) \, dx \, dy \end{aligned} \tag{3.12a}$$

and

$$E(Y) = \iint y f(x, y) \, dx \, dy. \tag{3.12b}$$

Equation (3.12b) is correct, because of the general definition of expectation of a function of two variables:

$$E\{g(X, Y)\} = \iint g(x, y) f(x, y) \, dx \, dy$$

whenever the expectation exists. Now just take $g(x, y) = y$.

One way that the right hand sides of (3.12a) and (3.12b) can be equal is if

$$f(x, y) = f(y \mid x) f_X(x) \tag{3.13}$$

or in words,

$$\text{joint} = \text{conditional} \times \text{marginal}$$

In fact, by the uniqueness theorem (Theorem 3.2), this is the only way the iterated expectation axiom can hold, except, as usual, for possible redefinition on sets of probability zero.

This gives a formula for calculating a conditional probability density from the joint

$$f(y \mid x) = \frac{f(x, y)}{f_X(x)} \tag{3.14}$$

or in words,

$$\text{conditional} = \frac{\text{joint}}{\text{marginal}}$$

Of course, there is a slight problem with (3.14) when the denominator is zero, but since the set of x such that $f_X(x) = 0$ is a set of probability zero, this does not matter, and $f(y \mid x)$ can be defined arbitrarily for all such x.

Example 3.4.1 (Uniform Distribution on a Triangle).
This continues Example 1.5.2. Recall from that example that if X and Y have joint density

$$f(x, y) = 2, \qquad 0 < x \text{ and } 0 < y \text{ and } x + y < 1$$

that the marginal of X is

$$f_X(x) = 2(1 - x), \qquad 0 < x < 1.$$

Thus the conditional is

$$f(y \mid x) = \frac{2}{2(1 - x)} = \frac{1}{1 - x}$$

Or we should say this is the conditional for *some* values of x and y. As usual, we have to be careful about domains of definition or we get nonsense. First, the marginal only has the formula we used when $0 < x < 1$, so that is one requirement. Then for x in that range, the joint is only defined by the formula we used when $0 < y$ and $x + y < 1$, that is, when $0 < y < 1 - x$. Thus to be precise, we must say

$$f(y \mid x) = \frac{1}{1 - x}, \qquad 0 < y < 1 - x \text{ and } 0 < x < 1. \tag{3.15}$$

What about other values of x and y? What if we want the definition for all real x and y? First, for $f(y \mid x)$ to be a probability density (considered as a function of y for fixed x) it must integrate to 1 (integrating with respect to y). Since our formula already does integrate to one over its domain of definition $0 < y < 1 - x$, it must be zero elsewhere. Thus when $0 < x < 1$

$$f(y \mid x) = \begin{cases} \frac{1}{1-x}, & 0 < y < 1 - x \\ 0, & \text{elsewhere} \end{cases}$$

or, if you prefer a definition using an indicator function,

$$f(y \mid x) = \frac{1}{1 - x} I_{(0, 1-x)}(y), \qquad y \in \mathbb{R}.$$

What about x outside $(0, 1)$? Those are x such that the marginal is zero, so the formula “joint over marginal” is undefined. As we have already said, the definition is then arbitrary, so we may say

$$f(y \mid x) = 42$$

or whatever we please when $x \leq 0$ or $1 \leq x$. (It doesn’t even matter that this function doesn’t integrate to one!) Mostly we will ignore such nonsense and only define conditional densities where the values are not arbitrary and actually matter. The only reason we mention this issue at all is so that you won't think $f(y \mid x)$ has to have a sensible definition for all possible x.

So how about conditional expectations? Given the formula (3.15) for the conditional density, we just plug and chug

$$E(Y \mid x) = \int y f(y \mid x) \, dy = \frac{1 - x}{2} \tag{3.16a}$$

$$E(Y^2 \mid x) = \int y^2 f(y \mid x) \, dy = \frac{(1 - x)^2}{3} \tag{3.16b}$$

$$\operatorname{var}(Y \mid x) = E(Y^2 \mid x) - E(Y \mid x)^2 = \frac{(1 - x)^2}{12} \tag{3.16c}$$

and so forth, (3.16c) holding because of Corollary 2.12, which like every other fact about unconditional expectation, also holds for conditional expectation so long as we are considering the conditioning variables fixed.

We could end Section 3.4 right here. Formulas (3.13) and (3.14) tell us how to calculate conditionals from joints and joints from conditionals and marginals. And the fact that "conditional expectation is a special case of ordinary expectation" (so long as we are considering the conditioning variables fixed) tells how to compute expectations. So what else is there to know? Well, nothing, but a lot more can be said on the subject. The rest of Section 3.4 should give you a much better feel for the subject and allow you to calculate conditional densities and expectations more easily.

3.4.2 Normalization

A standard homework problem for courses like this specifies some nonnegative function $h(x)$ and then asks for what real number k is $f(x) = kh(x)$ a probability density.

Clearly we must have $k > 0$, because $k < 0$ would entail negative probabilities and $k = 0$ would make the density integrate (or sum in the discrete case) to zero. Either violates the defining properties for a probability density, which are (1.20a) and (1.20b) in the discrete case and (1.21a) and (1.21b) in the continuous case.

For reasons that will soon become apparent, we prefer to use $c = 1/k$. This is allowed because $k \neq 0$. Thus the problem becomes: for what real number c is

$$f(x) = \frac{1}{c} h(x)$$

a density function? The process of determining c is called *normalization* and c is called the *normalizing constant* for the *unnormalized density* $h(x)$.

To determine c we use the second defining property for a probability density (1.20b) or (1.21b) as the case may be, which implies

$$c = \int h(x) \, dx \tag{3.17}$$

(with integration replaced by summation if the probability model is discrete). In order for c to be a positive number, the integral (or sum in the discrete case) must exist and be nonzero. This gives us two conditions on unnormalized densities. A real-valued function $h(x)$ is an *unnormalized density* provided the following two conditions hold.

- It is nonnegative: $h(x) \geq 0$, for all x.
- It is integrable in the continuous case or summable in the discrete case and the integral or sum is nonzero.

Then

$$f(x) = \frac{1}{c} h(x)$$

is a normalized probability density, where c is given by (3.17) in the continuous case and by (3.17) with the integral replaced by a sum in the discrete case.

Example 3.4.2.
Consider the function

$$h(x) = x^{\alpha - 1} e^{-x}, \qquad x > 0,$$

where $\alpha > 0$. How do we normalize it to make a probability density?

The normalizing constant is

$$c = \int_0^\infty x^{\alpha - 1} e^{-x} \, dx = \Gamma(\alpha)$$

by (B.2) in Appendix B. Thus we obtain a gamma distribution density

$$f(x) = \frac{1}{\Gamma(\alpha)} x^{\alpha - 1} e^{-x}.$$

So what's the big deal? We already knew that! Is "normalization" just a fancy name for something trivial? Well, yes and no. You can form your own opinion, but not until the end of Section 3.4.

3.4.3 Renormalization

We start with a slogan

Conditional probability is renormalization.

What this means will become apparent presently.

First, $f(y \mid x)$ is just an ordinary probability density when considered as a function of y for fixed x. We maintain this view, y is the variable and x is fixed, throughout this subsection.

Second, since x is fixed, the denominator in

$$f(y \mid x) = \frac{f(x, y)}{f_X(x)} \tag{3.18}$$

is constant (not a function of y). Thus we can also write

$$f(y \mid x) \propto f(x, y) \tag{3.19}$$

the symbol $\propto$ meaning "proportional to" (still thinking of y as the only variable, the proportionality does not hold if we vary x). This says the joint is just like the conditional, at least proportional to it, the only thing wrong is that it doesn't integrate to one (still thinking of y as the only variable, the joint does, of course, integrate to one if we integrate with respect to x and y). Formula (3.19) says that if we graph the conditional and the joint (as functions of y!) we get the same picture, they are the same shape, the only difference is the scale on the vertical axis (the constant of proportionality). So if we put in the constant of proportionality, we get

$$f(y \mid x) = \frac{1}{c(x)} f(x, y). \tag{3.20}$$

We have written the "constant" as $c(x)$ because it is a function of x, in fact, comparing with (3.18) we see that

$$c(x) = f_X(x).$$

We call it a "constant" because we are considering x fixed.

All of this can be summarized in the following slogan.

> *A joint density is an unnormalized conditional density. Its normalizing constant is a marginal density.*

Spelled out in more detail, the joint density $f(x, y)$ considered as a function of y alone is an unnormalized probability density, in fact, is it proportional to the conditional density (3.19). In order to calculate the conditional density, we need to calculate the normalizing constant, which just happens to turn out to be the marginal $f_X(x)$, and divide by it (3.18).

If we take this argument a bit further and plug the definition of the marginal into (3.18), we get

$$f(y \mid x) = \frac{f(x, y)}{\int f(x, y) \, dy} \tag{3.21}$$

This shows more explicitly how "conditional probability is renormalization." You find a conditional probability density by dividing the joint density by what it integrates to. How do we remember which variable is the variable of integration here? That's easy. In this whole subsection y is the only variable; x is fixed. In general, a conditional density is an ordinary density (integrates to one, etc.) when considered a function of the variable "in front of the bar" with the conditioning variable, the variable "behind the bar" *fixed.* That's what we are doing here. Hence we divide by the integral of the joint density with respect to the variable "in front of the bar."

It is occasionally useful that (3.21) holds whether or not the joint density is normalized. Suppose we are given an unnormalized joint density $h(x, y)$ so that

$$f(x, y) = \frac{1}{c} h(x, y)$$

for some normalizing constant c. Plugging this into (3.21) gives

$$f(y \mid x) = \frac{h(x, y)}{\int h(x, y)\, dy} \tag{3.22}$$

The c's cancel in the numerator and denominator.

Our slogan about conditional probability and renormalization helps us remember which marginal is meant in

$$\text{conditional} = \frac{\text{joint}}{\text{marginal}}$$

- If the conditional in question is $f(y \mid x)$, then we are considering y the variable (x is fixed).
- Thus the marginal in question is the one obtained by integrating with respect to y (that's what we are considering variable).
- The marginal obtained by integrating out y is the marginal of the *other* variable (slogan on p. 19 in these notes). Hence the marginal is $f_X(x)$.

But even if you are confused about how to calculate marginals or which marginal you need to divide by, you should still be able to calculate conditionals using (3.21) and (3.22), which *contain no marginals* and are in fact derivable on the spot. Both are obvious consequences of the facts that

- Conditional densities are proportional to joint densities considered as functions of the variable(s) in front of the bar.
- Conditional densities integrate to one considered as functions of the variable(s) in front of the bar.

Example 3.4.3.
Consider the function

$$h(x, y) = (x + y^2)e^{-x-y}, \qquad x > 0,\ y > 0.$$

If we take this to be an unnormalized joint density, what are the two conditional densities $f(x \mid y)$ and $f(y \mid x)$?

Integrating with respect to x gives

$$\begin{aligned} \int_0^\infty h(x, y)\, dx &= e^{-y} \int_0^\infty x e^{-x}\, dx + y^2 e^{-y} \int_0^\infty e^{-x}\, dx \\ &= (1 + y^2)e^{-y} \end{aligned}$$

We used the formula

$$\int_0^\infty x^n e^{-x}\, dx = \Gamma(n + 1) = n! \tag{3.23}$$

to evaluate the integrals. Hence

$$f(x \mid y) = \frac{f(x,y)}{\int f(x,y)\,dx} = \frac{x+y^2}{1+y^2}e^{-x}$$

Similarly

$$\begin{aligned}\int_0^\infty h(x,y)\,dy &= xe^{-x}\int_0^\infty e^{-y}\,dy + e^{-x}\int_0^\infty y^2 e^{-y}\,dy \\ &= (x+2)e^{-x}\end{aligned}$$

Again, we used (3.23) to evaluate the integrals. So

$$f(y \mid x) = \frac{f(x,y)}{\int f(x,y)\,dy} = \frac{x+y^2}{x+2}e^{-y}$$

Things become considerably more complicated when the support of the joint density is not a rectangle with sides parallel to the axes. Then the domains of integration depend on the values of the conditioning variable.

Example 3.4.4 (A Density with Weird Support).
Consider the function

$$h(x,y) = \begin{cases} x+y^2, & x>0,\ y>0,\ x+y<1 \\ 0, & \text{otherwise} \end{cases}$$

If we take this to be an unnormalized joint density, what is the conditional density $f(x \mid y)$?

Integrating with respect to x gives

$$\int_{-\infty}^{\infty} h(x,y)\,dx = \int_0^{1-y} (x+y^2)\,dx = \frac{x^2}{2} + xy^2 \Big|_0^{1-y} = \tfrac{1}{2}(1-y)(1-y+2y^2)$$

What is tricky is that the formula $x+y^2$ for $h(x,y)$ is valid only when $x>0$ and $y>0$ and $x+y<1$. This means $0<x<1-y$. For other values of x, the integrand is zero. Hence the domain of integration in the second integral must be $0<x<1-y$. If you miss this point about the domain of integration, you make a complete mess of the problem. If you get this point, the rest is easy

$$f(x \mid y) = \frac{f(x,y)}{\int f(x,y)\,dx} = \frac{2(x+y^2)}{(1-y)(1-y+2y^2)}$$

3.4.4 Renormalization, Part II

This subsection drops the other shoe in regard to “conditional probability is renormalization.” So is conditional expectation. Plugging the definition (3.21) of conditional densities into (3.3) gives

$$E\{g(Y) \mid x\} = \frac{\int g(y)f(x,y)\,dy}{\int f(x,y)\,dy} \tag{3.24}$$

(and, of course, the discrete case is analogous with the integrals replaced by sums). It is a useful mnemonic device to write (3.24) lining up the analogous bits in the numerator and denominator

$$E\{g(Y) \mid x\} = \frac{\int g(y) f(x, y) \, dy}{\int f(x, y) \, dy}.$$

This looks a little funny, but it reminds us that the density in the numerator and denominator is the same, and the variable of integration is the same. The only difference between the numerator and denominator is the function $g(y)$ appearing in the numerator.

If we plug in (3.22) instead of (3.21) for $f(y \mid x)$ we get

$$E\{g(Y) \mid x\} = \frac{\int g(y) h(x, y) \, dy}{\int h(x, y) \, dy} \tag{3.25}$$

where $h(x, y)$ is an *unnormalized* joint density.

These formulas make it clear that we are choosing the denominator so that $E(1 \mid x) = 1$, which is the form the norm axiom takes when applied to conditional probability. That is, when we take the special case in which the function $g(y)$ is equal to one for all y, the numerator and denominator are the same.

Example 3.4.5.
Suppose X and Y have the unnormalized joint density

$$h(x, y) = (x + y) e^{-x-y}, \qquad x > 0, \ y > 0,$$

what is $E(X \mid y)$?

Using (3.25) with the roles of X and Y interchanged and g the identity function we get

$$\begin{aligned} E(X \mid y) &= \frac{\int x h(x, y) \, dx}{\int h(x, y) \, dx} \\ &= \frac{\int x (x + y) e^{-x-y} \, dx}{\int (x + y) e^{-x-y} \, dx} \end{aligned}$$

Using (3.23) the denominator is

$$\begin{aligned} \int_0^\infty (x + y) e^{-x-y} \, dx &= e^{-y} \int_0^\infty x e^{-x} \, dx + y e^{-y} \int_0^\infty e^{-x} \, dx \\ &= (1 + y) e^{-y} \end{aligned}$$

and the numerator is

$$\begin{aligned} \int_0^\infty x (x + y) e^{-x-y} \, dx &= e^{-y} \int_0^\infty x^2 e^{-x} \, dx + y e^{-y} \int_0^\infty x e^{-x} \, dx \\ &= (2 + y) e^{-y} \end{aligned}$$

Hence

$$E(X \mid y) = \frac{2 + y}{1 + y}, \qquad y > 0.$$

Recall from p. 90 in these notes

> **Sanity Check:** $E(X \mid Y)$ *is a function of* Y *and is not a function of* X.

Good. We did get a function of y. If you get confused about which variable to integrate with respect to, this sanity check will straighten you out. If you through some mistake get a function of both variables, this sanity check will at least tell you that you messed up somewhere.

3.4.5 Bayes Rule

Now we want to study the consequences of

$$\text{joint} = \text{conditional} \times \text{marginal} \tag{3.26}$$

Again we have the problem of remembering which marginal. If we recall our analysis of

$$\text{conditional} = \frac{\text{joint}}{\text{marginal}}$$

on p. 100 in these notes, we recall that it is the marginal of the variable "behind the bar."

Because "mathematics is invariant under changes of notation" (3.26) is also true when we interchange the roles of the variables Hence we can "factor" a joint density into marginal and conditional two different ways

$$f(x, y) = f(x \mid y) f_Y(y) \tag{3.27}$$

$$f(x, y) = f(y \mid x) f_X(x) \tag{3.28}$$

Plugging (3.27) into (3.21) gives

$$f(y \mid x) = \frac{f(x \mid y) f_Y(y)}{\int f(x \mid y) f_Y(y) \, dy} \tag{3.29}$$

This equation is called *Bayes rule*. It allows us to "turn around" conditional probabilities. That is, it is useful for problems that say: given $f(x \mid y)$, find $f(y \mid x)$. Or vice versa. Of course, because "mathematics is invariant under changes of notation" (3.29) is also true with all the x's and y's interchanged.

Example 3.4.6.
Suppose that X and Y are positive real-valued random variables and

$$f(x \mid y) = \tfrac{1}{2} x^2 y^3 e^{-xy}$$

$$f_Y(y) = e^{-y}$$

what is $f(y \mid x)$?

Note that this is slightly tricky in that the conditional wanted is not the one given by the Bayes rule formula (3.29). You need to interchange x's and y's in (3.29) to get the formula needed to do this problem

$$f(y \mid x) = \frac{f(x \mid y) f_Y(y)}{\int f(x \mid y) f_Y(y) \, dy}$$

The denominator is

$$\int_0^\infty x^2 y^3 e^{-xy-y} \, dy = x^2 \int_0^\infty y^3 e^{-(1+x)y} \, dy$$

The change of variable $y = u/(1+x)$ makes the right hand side

$$\frac{x^2}{(1+x)^4} \int_0^\infty u^3 e^{-u} \, du = \frac{6x^2}{(1+x)^4}$$

Thus

$$f(y \mid x) = \tfrac{1}{6}(1+x)^4 y^3 e^{-(1+x)y}, \qquad y > 0$$

Example 3.4.7 (Bayes and Brand Name Distributions).
Suppose

$$\begin{aligned} X &\sim \text{Exp}(\lambda) \\ Y \mid X &\sim \text{Exp}(X) \end{aligned}$$

meaning the marginal distribution of X is $\text{Exp}(\lambda)$ and the conditional distribution of Y given X is $\text{Exp}(X)$, that is,

$$f(y \mid x) = x e^{-xy}, \qquad y > 0. \tag{3.30}$$

This is a bit tricky, so let's go through it slowly. The formula for the density of the exponential distribution given in Section B.2.2 in Appendix B is

$$f(x \mid \lambda) = \lambda e^{-\lambda x}, \qquad x > 0. \tag{3.31}$$

We want to change x to y and λ to x. Note that it matters which order we do the substitution. If we change λ to x first, we get

$$f(x \mid x) = \lambda e^{-x^2}, \qquad x > 0.$$

but that's nonsense. First, the right hand side isn't a density. Second, the left hand side is the density of X given X, but this distribution is concentrated at X (if we know X, then we know X) and so isn't even continuous. So change x in (3.31) to y obtaining

$$f(y \mid \lambda) = \lambda e^{-\lambda y}, \qquad y > 0.$$

and then change λ to x obtaining (3.30).

Of course, the joint is conditional times marginal

$$f(x, y) = f(y \mid x) f_X(x) = x e^{-xy} \cdot \lambda e^{-\lambda x} = \lambda x e^{-(\lambda + y) x} \tag{3.32}$$

Question: What is the other marginal (of Y) and the other conditional (of X given Y)? Note that these two problems are related. If we answer one, the answer to the other is easy, just a division

$$f(x \mid y) = \frac{f(x, y)}{f_Y(y)}$$

or

$$f_Y(y) = \frac{f(x, y)}{f(x \mid y)}$$

I find it a bit easier to get the conditional first. Note that the joint (3.32) is an unnormalized conditional when thought of as a function of x alone. Checking our inventory of "brand name" distributions, we see that the only one like (3.32) in having both a power and an exponential of the variable is the gamma distribution with density

$$f(x \mid \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha - 1} e^{-\lambda x}, \qquad x > 0. \tag{3.33}$$

Comparing the analogous parts of (3.32) and (3.33), we see that we must match up x with $x^{\alpha-1}$, which tells us we need $\alpha = 2$, and we must match up $e^{-(\lambda+y)x}$ with $e^{-\lambda x}$ which tells us we need $\lambda + y$ in (3.32) to be the λ in (3.33), which is the second parameter of the gamma distribution. Thus (3.32) must be an unnormalized $\Gamma(2, \lambda + y)$ density, and the properly normalized density is

$$f(x \mid y) = (\lambda + y)^2 x e^{-(\lambda + y) x}, \qquad x > 0 \tag{3.34}$$

Again this is a bit tricky, so let's go through it slowly. We want to change α to 2 and λ to $\lambda + y$ in (3.33). That gives

$$f(x \mid y) = \frac{(\lambda + y)^2}{\Gamma(2)} x^{2-1} e^{-(\lambda + y) x}, \qquad x > 0.$$

and this cleans up to give (3.34).

3.5 Conditional Expectation and Prediction

The parallel axis theorem (Theorem 2.11 in these notes)

$$E[(X - a)^2] = \operatorname{var}(X) + [a - E(X)]^2$$

has an analog for conditional expectation. Just replace expectations by conditional expectations (and variances by conditional variances) and, because functions of the conditioning variable behave like constants, replace the constant by a function of the conditioning variable.

Theorem 3.5 (Conditional Parallel Axis Theorem). *If $Y \in L^1$*

$$E\{[Y - a(X)]^2 \mid X\} = \operatorname{var}(Y \mid X) + [a(X) - E(Y \mid X)]^2 \tag{3.35}$$

The argument is exactly the same as that given for the unconditional version, except for the need to use Axiom CE1 instead of Axiom E2 to pull a function of the conditioning variable out of the conditional expectation. Otherwise, only the notation changes.

If we take the unconditional expectation of both sides of (3.35), we get

$$E\bigl(E\{[Y - a(X)]^2 \mid X\}\bigr) = E\{\operatorname{var}(Y \mid X)\} + E\{[a(X) - E(Y \mid X)]^2\}$$

and by the iterated expectation axiom, the left hand side is the the unconditional expectation, that is,

$$E\{[Y - a(X)]^2\} = E\{\operatorname{var}(Y \mid X)\} + E\{[a(X) - E(Y \mid X)]^2\} \tag{3.36}$$

This relation has no special name, but it has two very important special cases. The first is the prediction theorem.

Theorem 3.6. *For predicting a random variable Y given the value of another random variable X, the predictor function $a(X)$ that minimizes the expected squared prediction error*

$$E\{[Y - a(X)]^2\}$$

is the conditional expectation $a(X) = E(Y \mid X)$.

The proof is extremely simple. The expected squared prediction error is the left hand side of (3.36). On the right hand side of (3.36), the first term does not contain $a(X)$. The second term is the expectation of the square of $a(X) - E(Y \mid X)$. Since a square is nonnegative and the expectation of a nonnegative random variable is nonnegative (Axiom E1), the second term is always nonnegative and hence is minimized when it is zero. By Theorem 2.32, that happens if and only if $a(X) = E(Y \mid X)$ with probability one. (Yet another place where redefinition on a set of probability zero changes nothing of importance).

Example 3.5.1 (Best Prediction).
Suppose X and Y have the unnormalized joint density

$$h(x, y) = (x + y)e^{-x-y}, \qquad x > 0,\ y > 0,$$

what function of Y is the best predictor of X in the sense of minimizing expected squared prediction error?

The predictor that minimizes expected squared prediction error is the regression function

$$a(Y) = E(X \mid Y) = \frac{2 + Y}{1 + Y}$$

found in Example 3.4.5.

The other important consequence of (3.36) is obtained by taking $a(X) = E(Y) = \mu_Y$ (that is, a is the constant function equal to μ_Y). This gives

$$E\{[Y - \mu_Y]^2\} = E\{\operatorname{var}(Y \mid X)\} + E\{[\mu_Y - E(Y \mid X)]^2\} \tag{3.37}$$

The left hand side of (3.37) is, by definition $\operatorname{var}(Y)$. By the iterated expectation axiom, $E\{E(Y \mid X)\} = E(Y) = \mu_Y$, so the second term on the right hand side is the expected squared deviation of $E(Y \mid X)$ from its expectation, which is, by definition, its variance. Thus we have obtained the following theorem.

Theorem 3.7 (Iterated Variance Formula). *If $Y \in L^2$,*

$$\operatorname{var}(Y) = E\{\operatorname{var}(Y \mid X)\} + \operatorname{var}\{E(Y \mid X)\}.$$

Example 3.5.2 (Example 3.3.1 Continued).
Suppose X_0, X_1, $\ldots$ is an infinite sequence of identically distributed random variables, having mean $E(X_i) = \mu_X$ and variance $\operatorname{var}(X_i) = \sigma_X^2$, and suppose N is a nonnegative integer-valued random variable independent of the X_i having mean $E(N) = \mu_N$ and variance $\operatorname{var}(N) = \sigma_N^2$. Note that we have now tied up the loose end in Example 3.3.1. We now know from Theorem 3.4 that independence of the X_i and N implies

$$E(X_i \mid N) = E(X_i) = \mu_X.$$

and similarly

$$\operatorname{var}(X_i \mid N) = \operatorname{var}(X_i) = \sigma_X^2.$$

Question: What is the variance of

$$S_N = X_1 + \cdots + X_N$$

expressed in terms of the means and variances of the X_i and N?

This is easy using the iterated variance formula. First, as we found in Example 3.3.1,

$$E(S_N \mid N) = N E(X_i \mid N) = N\mu_X.$$

A similar calculation gives

$$\operatorname{var}(S_N \mid N) = N \operatorname{var}(X_i \mid N) = N\sigma_X^2$$

(because of the assumed independence of the X_i and N). Hence

$$\begin{aligned}
\operatorname{var}(S_N) &= E\{\operatorname{var}(S_N \mid N)\} + \operatorname{var}\{E(S_N \mid N)\} \\
&= E(N\sigma_X^2) + \operatorname{var}(N\mu_X) \\
&= \sigma_X^2 E(N) + \mu_X^2 \operatorname{var}(N) \\
&= \sigma_X^2 \mu_N + \mu_X^2 \sigma_N^2
\end{aligned}$$

Again notice that it is impossible to do this problem any other way. There is not enough information given to use any other approach.

Also notice that the answer is not exactly obvious. You might just guess, using your intuition, the answer to Example 3.3.1. But you wouldn't guess this. You need the theory.

Problems

3-1. In class we found the moment generating function of the geometric distribution (Section B.1.3 in Appendix B) is defined by

$$\psi(t) = \frac{1 - p}{1 - pe^t}$$

on some neighborhood of zero. Find the variance of this random variable.

3-2. Verify the details in (3.16a), (3.16b), and (3.16c).

3-3. Suppose X is a positive random variable and the density of Y given X is

$$f(y \mid x) = \frac{2y}{x^2}, \qquad 0 < y < x.$$

(a) Find $E(Y \mid X)$.

(b) Find $\operatorname{var}(Y \mid X)$.

3-4. For what real values of θ is

$$f_\theta(x) = \frac{1}{c(\theta)} x^\theta, \qquad 0 < x < 1$$

a probability density, and what is the function $c(\theta)$?

3-5. Suppose X, Y, and Z are random variables such that

$$E(X \mid Y, Z) = Y \quad \text{and} \quad \operatorname{var}(X \mid Y, Z) = Z.$$

Find the (unconditional) mean and variance of X in terms of the means, variances, and covariance of Y and Z.

3-6. Suppose the random vector (X, Y) is uniformly distributed on the disk

$$S = \{\, (x, y) \in \mathbb{R}^2 : x^2 + y^2 < 4 \,\}$$

that is, (X, Y) has the $\mathcal{U}(S)$ distribution in the notation of Section B.2.1 of Appendix B.

(a) Find the conditional distributions of X given Y and of Y given X.

(b) Find the marginal distributions of X and Y.

(c) Find $E(Y \mid x)$.

(d) Find $P(|Y| < 1 \mid x)$.

3-7. Suppose the conditional distribution of Y given X is $\mathcal{N}(0, 1/X)$ and the marginal distribution of X is $\operatorname{Gam}(\alpha, \lambda)$.

(a) What is the conditional density of X given Y?

(b) What is the marginal density of Y?

3-8. Suppose X and Z are independent random variables and $E(Z) = 0$. Define $Y = X + X^2 + Z$.

(a) Find $E(Y \mid X)$.

(b) Find $\operatorname{var}(Y \mid X)$.

(c) What function of X is the best predictor of Y in the sense of minimizing expected squared prediction error?

(d) What is the expected squared prediction error of this predictor?

Note: Any of the answers may involve moments of X and Z.

Chapter 4

Parametric Families of Distributions

The first thing the reader should do before reading the rest of this chapter is go back and review Section 3.1, since that establishes the basic notation for parametric families of distributions.

4.1 Location-Scale Families

Consider a probability density f of a real-valued random variable X. By the theorem on linear changes of variables (Theorem 7 of Chapter 3 in Lindgren), for any real number μ and any positive real number σ, the random variable $Y = \mu + \sigma X$ has the density

$$f_{\mu,\sigma}(y) = \frac{1}{\sigma} f\left(\frac{y - \mu}{\sigma}\right).$$

This generates a two-parameter family of densities called the *location-scale* family generated by the *reference density* f. The parameter μ is called the *location* parameter, and the parameter σ is called the *scale* parameter.

We could choose any distribution in the family as the reference distribution with density f. This gives a different *parameterization* of the family, but the *same* family. Suppose we choose $f_{\alpha,\beta}$ as the reference density. The family it generates has densities

$$\begin{aligned}
f_{\mu,\sigma}(y) &= \frac{1}{\sigma} f_{\alpha,\beta}\left(\frac{y - \mu}{\sigma}\right). \\
&= \frac{1}{\sigma\beta} f\left(\frac{1}{\beta}\left[\frac{y - \mu}{\sigma} - \alpha\right]\right) \\
&= \frac{1}{\sigma\beta} f\left(\frac{y - \mu - \sigma\alpha}{\sigma\beta}\right)
\end{aligned}$$

It is clear that as μ and σ run over all possible values we get the same *family* of distributions as before. The parameter values that go with each particular distribution have changed, but each density that appears in one family also appears in the other. The correspondence between the parameters in the two parameterizations is

$$\begin{aligned} \mu &\longleftrightarrow \mu + \sigma\alpha \\ \sigma &\longleftrightarrow \sigma\beta \end{aligned}$$

If the reference random variable X has a variance, then every distribution in the family has a variance (by Theorem 2.44 in these notes), and the distributions of the family have every possible mean and variance. Since we are free to choose the reference distribution as any distribution in the family, we may as well choose so that $E(X) = 0$ and $\operatorname{var}(X) = 1$, then μ is the mean and σ the standard deviation of the variable Y with density $f_{\mu,\sigma}$.

But the distributions of the family do not have to have either means or variances. In that case we cannot call μ the mean or σ the standard deviation. That is the reason why in general we call μ and σ the location and scale parameters.

Example 4.1.1 (Uniform Distributions).
The $\mathcal{U}(a, b)$ family of distribution defined in Section B.2.1 of Appendix B has densities

$$f(x \mid a, b) = \frac{1}{b - a}, \qquad a < x < b \tag{4.1}$$

and moments

$$\begin{aligned} E(X \mid a, b) &= \frac{a + b}{2} \\ \operatorname{var}(X \mid a, b) &= \frac{(b - a)^2}{12} \end{aligned}$$

Therefore the parameters a and b of the distribution having mean zero and standard deviation one is found by solving

$$\frac{a + b}{2} = 0$$

(from which we see that $b = -a$) and

$$\frac{(b - a)^2}{12} = 1$$

which becomes, plugging in $b = -a$,

$$\frac{(2 \cdot b)^2}{12} = 1$$

Hence $b = \sqrt{3}$. Giving the density

$$f(x) = \frac{1}{2\sqrt{3}}, \qquad -\sqrt{3} < x < +\sqrt{3}$$

Then use the formula for a general location-scale family, obtaining

$$f(y \mid \mu, \sigma) = \frac{1}{\sigma} f\left(\frac{y - \mu}{\sigma}\right) = \frac{1}{2\sigma\sqrt{3}}$$

on the domain of definition, whatever that is. The change of variable is $y = \mu + \sigma x$, so $x = \pm\sqrt{3}$ maps to $\mu \pm \sigma\sqrt{3}$, and those are the endpoints of the domain of definition. So

$$f(y \mid \mu, \sigma) = \frac{1}{2\sigma\sqrt{3}}, \qquad \mu - \sigma\sqrt{3} < y < \mu + \sigma\sqrt{3} \tag{4.2}$$

The reader may have lost track in all the formula smearing of how simple this all is. We have another description of the *same* family of densities. The correspondence between the two parameterizations is

$$\begin{aligned} a &\longleftrightarrow \mu - \sigma\sqrt{3} \\ b &\longleftrightarrow \mu + \sigma\sqrt{3} \end{aligned}$$

It should be clear that (4.2) defines a density that is constant on an interval, just like (4.1) does. Furthermore, it should also be clear that as μ and σ range over all possible values we get distributions on all possible intervals. This is not so obvious from the range specification in (4.2), but is clear from the definition of μ and σ in terms of a and b

$$\begin{aligned} \mu &= \frac{a + b}{2} \\ \sigma &= \sqrt{\frac{(b - a)^2}{12}} \end{aligned}$$

The only virtue of the new parameterization (4.2) over the old one (4.1) is that it explicitly describes the density in terms of the mean and standard deviation (μ is the mean and σ is the standard deviation, as explained in the comments immediately preceding the example). But for most people that is not a good enough reason to use the more complicated parameterization. Hence (4.1) is much more widely used.

Example 4.1.2 (Cauchy Distributions).
The function

$$f(x) = \frac{1}{\pi(1 + x^2)}, \qquad -\infty < x < +\infty$$

is a probability density, because

$$\int_{-\infty}^{\infty} \frac{1}{1 + x^2}\,dx = \tan^{-1} x \Big|_{-\infty}^{\infty} = \pi$$

This density is called the standard Cauchy density (Section 6.12 in Lindgren). This distribution has no mean or variance. If we try to calculate

$$E(|X|) = \int_{-\infty}^{\infty} \frac{|x|}{1 + x^2}\,dx = 2\int_{0}^{\infty} \frac{x}{1 + x^2}\,dx$$

we see that, because the integrand is bounded, only the behavior of the integrand near infinity is important. And for large x

$$\frac{x}{1+x^2} \approx \frac{1}{x}$$

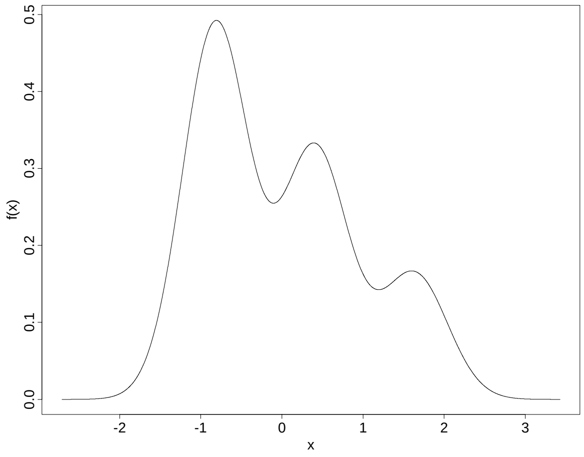
and so by Lemma 2.39 the integral does not exist. Hence by Theorem 2.44 neither does any moment of first or higher order. That is, no moments exist.

The Cauchy location-scale family has densities

$$f_{\mu,\sigma}(x) = \frac{\sigma}{\pi(\sigma^2 + [x-\mu]^2)}, \qquad -\infty < x < +\infty \tag{4.3}$$

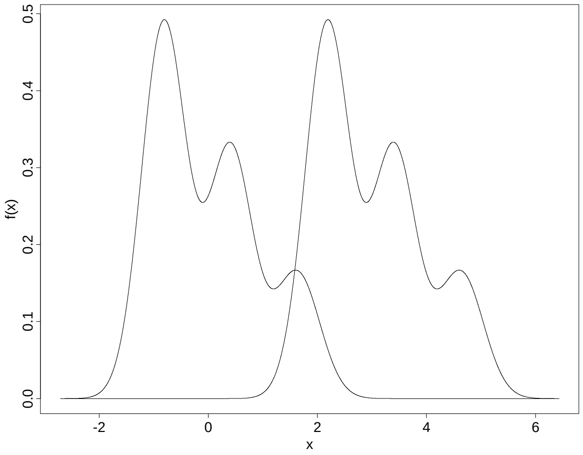
Here μ is not the mean, because Cauchy distributions do not have means. It is, however, the median because this distribution is symmetric with center of symmetry μ. Neither is σ the standard deviation, because Cauchy distributions do not have variances.

Example 4.1.3 (Blurfle Distributions).
All of the distributions in a location-scale family have the same *shape*. In fact we could use the same curve as the graph of every density in the family. Changing μ and σ only changes the scales on the axes, not the shape of the curve. Consider the distribution with the density shown below, which is of no particular interest, just an arbitrary p. d. f. Call it the "blurfle" distribution. It has been chosen so to have mean zero and variance one, so we can refer to it as the *standard* blurfle distribution.

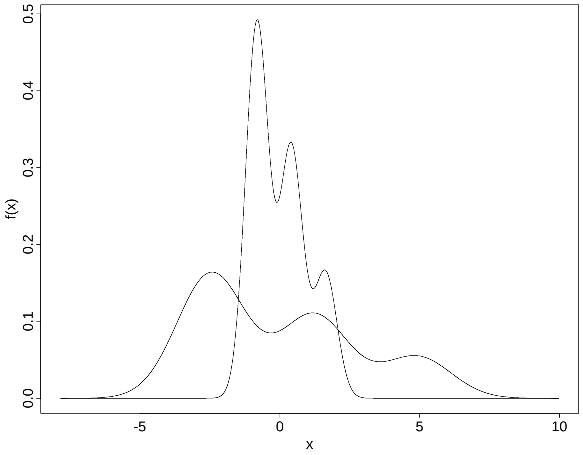


Like any other distribution, it generates a location-scale family, which we can call the blurfle family. Different blurfle distributions have the same shape, just different location and scale parameters. Changing the location parameter, but leaving the scale parameter unchanged just shifts the curve to the right or left along the number line.

Shown below are two different blurfle densities with same scale parameter but different location parameters.



And shown below are two different blurfle densities with same location parameter but different scale parameters.



4.2 The Gamma Distribution

The gamma function is defined for all real $\alpha > 0$ by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} \, dx. \tag{4.4}$$

Theorem 4.1 (Gamma Function Recursion Relation).

$$\Gamma(\alpha + 1) = \alpha \Gamma(\alpha) \tag{4.5}$$

holds for all $\alpha > 0$.

Proof. This can be proved using the integration by parts formula: $\int u\,dv = uv - \int v\,du$. Let $u = x^\alpha$ and $dv = e^{-x}\,dx$, so $du = \alpha x^{\alpha-1}\,du$ and $v = -e^{-x}$, and

$$\begin{aligned} \Gamma(\alpha+1) &= \int_0^\infty x^\alpha e^{-x}\,dx \\ &= -x^\alpha e^{-x}\Big|_0^\infty - \int_0^\infty \alpha x^{\alpha-1} e^{-x}\,dx \\ &= \alpha\Gamma(\alpha) \end{aligned}$$

The uv term in the integration by parts is zero, because $x^\alpha e^{-x}$ goes to zero as x goes to either zero or infinity. □

Since

$$\Gamma(1) = \int_0^\infty e^{-x}\,dx = -e^{-x}\Big|_0^\infty = 1,$$

the gamma function interpolates the factorials

$$\begin{aligned} \Gamma(2) &= 1 \cdot \Gamma(1) = 1! \\ \Gamma(3) &= 2 \cdot \Gamma(2) = 2! \\ &\vdots \\ \Gamma(n+1) &= n \cdot \Gamma(n) = n! \end{aligned}$$

In a later section, we will find out that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, which can be used with the recursion relation (4.5) to find $\Gamma(\frac{n}{2})$ for odd positive integers n.

The integrand in the integral defining the gamma function (4.4) is nonnegative and integrates to a finite, nonzero constant. Hence, as we saw in Example 3.4.2, dividing it by what it integrates to makes a probability density

$$f(x \mid \alpha) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}, \qquad x > 0. \tag{4.6}$$

The parameter α of the family is neither a location nor a scale parameter. Each of these densities has a different shape. Hence we call it a *shape parameter.*

It is useful to enlarge the family of densities by adding a scale parameter. If X has the density (4.6), then σX has the density

$$f(x \mid \alpha, \sigma) = \frac{1}{\sigma} f\Big(\frac{x}{\sigma} \;\Big|\; \alpha\Big) = \frac{1}{\sigma^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\sigma}. \tag{4.7}$$

For reasons that will become apparent later Lindgren prefers to use the reciprocal scale parameter $\lambda = 1/\sigma$. If the units of X are feet, then so are the units of σ. The units of λ are reciprocal feet (ft^{-1}). In this parameterization the densities are

$$f(x \mid \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}. \tag{4.8}$$

You should be warned that there is no generally accepted parameterization of the gamma family of densities. Some books prefer one, some the other. In this course we will always use (4.8), and following Lindgren we will use the notation $\text{Gam}(\alpha, \lambda)$ to denote the distribution with density (4.8). We will call λ the *inverse scale parameter* or, for reasons to be explained later (Section 4.4.3), the *rate parameter*. The fact that (4.8) must integrate to one tells us

$$\int_0^\infty x^{\alpha-1} e^{-\lambda x} \, dx = \frac{\Gamma(\alpha)}{\lambda^\alpha}.$$

We can find the mean and variance of the gamma using the trick of recognizing a probability density (Section 2.5.7).

$$\begin{aligned} E(X) &= \int_0^\infty x f(x \mid \alpha, \lambda) \, dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^\alpha e^{-\lambda x} \, dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+1)}{\lambda^{\alpha+1}} \\ &= \frac{\alpha}{\lambda} \end{aligned}$$

(we used the recursion (4.5) to simplify the ratio of gamma functions). Similarly

$$\begin{aligned} E(X^2) &= \int_0^\infty x^2 f(x \mid \alpha, \lambda) \, dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha+1} e^{-\lambda x} \, dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+2)}{\lambda^{\alpha+2}} \\ &= \frac{(\alpha+1)\alpha}{\lambda^2} \end{aligned}$$

(we used the recursion (4.5) twice). Hence

$$\operatorname{var}(X) = E(X^2) - E(X)^2 = \frac{(\alpha+1)\alpha}{\lambda^2} - \left(\frac{\alpha}{\lambda}\right)^2 = \frac{\alpha}{\lambda^2}$$

The sum of independent gamma random variables with the same scale parameter is also gamma. If X_1, $\ldots$, X_k are independent with $X_i \sim \text{Gam}(\alpha_i, \lambda)$, then

$$X_1 + \cdots + X_k \sim \text{Gam}(\alpha_1 + \cdots + \alpha_k, \lambda).$$

This will be proved in the following section (Theorem 4.2).

4.3 The Beta Distribution

For any real numbers $s > 0$ and $t > 0$, the function

$$h(x) = x^{s-1}(1-x)^{t-1}, \qquad 0 < x < 1$$

is an unnormalized probability density. This is clear when $s \geq 1$ and $t \geq 1$, because then it is bounded. When $s < 1$, it is unbounded near zero. When $t < 1$, it is unbounded near one. But even when unbounded it is integrable. For x near zero

$$h(x) \approx x^{s-1}$$

Hence h is integrable on $(0, \epsilon)$ for any $\epsilon > 0$ by Lemmas 2.40 and 2.43 because the exponent $s - 1$ is greater than -1. The same argument (or just changing the variable from x to $1 - x$) shows that the unnormalized density h is integrable near one.

The normalizing constant for h depends on s and t and is called the beta function

$$B(s, t) = \int_0^1 x^{s-1}(1 - x)^{t-1} \, dx.$$

Dividing by the normalizing constant gives normalized densities

$$f(x \mid s, t) = \frac{1}{B(s, t)} x^{s-1}(1 - x)^{t-1}, \qquad 0 < x < 1.$$

The probability distributions having these densities are called *beta distributions* and are denoted $\text{Beta}(s, t)$.

The next theorem gives the "addition rule" for gamma distributions mentioned in the preceding section and a connection between the gamma and beta distributions.

Theorem 4.2. *If X and Y are independent random variables*

$$\begin{aligned} X &\sim \text{Gam}(s, \lambda) \\ Y &\sim \text{Gam}(t, \lambda) \end{aligned}$$

Then

$$\begin{aligned} U &= X + Y \\ V &= \frac{X}{X + Y} \end{aligned}$$

are also independent random variables, and

$$\begin{aligned} U &\sim \text{Gam}(s + t, \lambda) \\ V &\sim \text{Beta}(s, t) \end{aligned}$$

Proof. To use the multivariate change of variable formula, we first solve for the old variables x and y in terms of the new

$$\begin{aligned} x &= uv \\ y &= u(1 - v) \end{aligned}$$

Hence the Jacobian is

$$J(u,v) = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} v & u \\ 1-v & -u \end{vmatrix} = -u$$

The joint density of X and Y is $f_X(x)f_Y(y)$ by independence. By the change of variable formula, the joint density of U and V is

$$\begin{aligned} f_{U,V}(u,v) &= f_{X,Y}[uv, u(1-v)]|J(u,v)| \\ &= f_X(uv)f_Y[u(1-v)]u \\ &= \frac{\lambda^s}{\Gamma(s)}(uv)^{s-1}e^{-\lambda uv}\frac{\lambda^t}{\Gamma(t)}[u(1-v)]^{t-1}e^{-\lambda u(1-v)}u \\ &= \frac{\lambda^{s+t}}{\Gamma(s)\Gamma(t)}u^{s+t-1}e^{-\lambda u}v^{s-1}(1-v)^{t-1} \end{aligned}$$

Since the joint density factors into a function of u times a function of v, the variables U and V are independent. Since these functions are proportional to the gamma and beta densities asserted by the theorem, U and V must actually have these distributions. □

Corollary 4.3.

$$B(s,t) = \frac{\Gamma(s)\Gamma(t)}{\Gamma(s+t)}$$

Proof. The constant in the joint density found in the proof of the theorem must be the product of the constants for the beta and gamma densities. Hence

$$\frac{\lambda^{s+t}}{\Gamma(s)\Gamma(t)} = \frac{\lambda^{s+t}}{\Gamma(s+t)}\frac{1}{B(s,t)}$$

Solving for $B(s,t)$ gives the corollary. □

For moments of the beta distribution, see Lindgren pp. 176–177.

4.4 The Poisson Process

4.4.1 Spatial Point Processes

A *spatial point process* is a random pattern of points in a region of space. The space can be any dimension.

A point process is *simple* if it never has points on top of each other so that each point of the process is at a different location in space. A point process is *boundedly finite* if with probability one it has only a finite number of points in any bounded set.

Let N_A denote the number of points in a region A. Since the point pattern is random, N_A is a random variable. Since it counts points, N_A is a discrete random variable taking values 0, 1, 2, ... If A is a bounded set and the point process is boundedly finite, then the event $N_A = \infty$ has probability zero.

A point x is a *fixed atom* if $P(N_{\{x\}} > 0) > 0$, that is, if there is positive probability of seeing a point at the particular location x in every random pattern. We are interested in point processes in which the locations of the points are continuous random variables, in which case the probability of seeing a point at any particular location is zero, so there are no fixed atoms.

For a general spatial point process, the joint distribution of the variables N_A for various sets A is very complicated. There is one process for which it is not complicated. This is the Poisson process, which is a model for a "completely random" pattern of points. One example of this process is given in Figure 4.1.

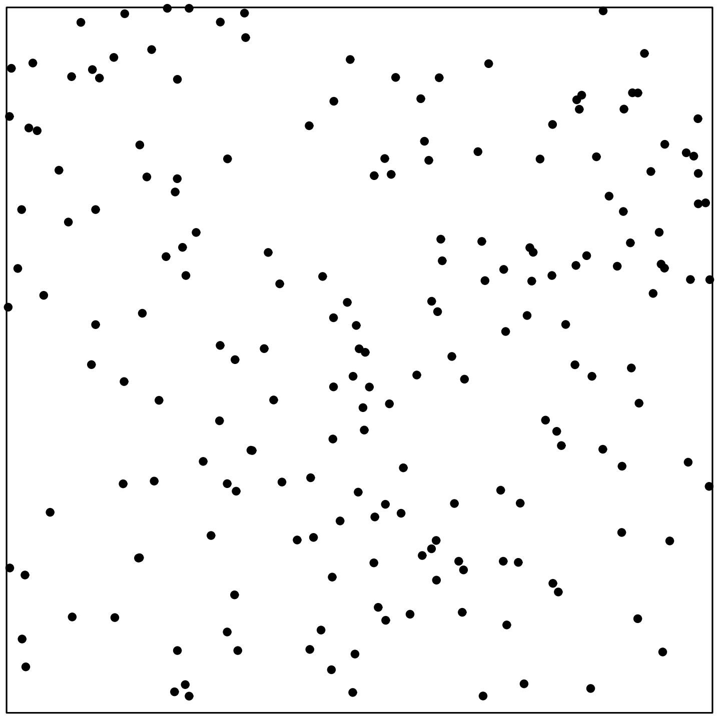


Figure 4.1: A single realization of a homogeneous Poisson process.

4.4.2 The Poisson Process

A Poisson process is a spatial point process characterized by a simple independence property.

Definition 4.4.1.
A ***Poisson process*** *is a simple, boundedly finite spatial point process with no fixed atoms having the property that N_{A_1}, N_{A_2}, …, N_{A_k} are independent random variables, whenever A_1, A_2, …, A_k are disjoint bounded sets.*

In short, counts of points in disjoint regions are independent random variables. It is a remarkable fact that the independence property alone determines the distribution of the counts.

Theorem 4.4. *For a Poisson process, N_A has a Poisson distribution for every bounded set A. Conversely, a simple point process with no fixed atoms such that N_A has a Poisson distribution for every bounded set A is a Poisson process.*

Write $\Lambda(A) = E(N_A)$. Since the parameter of the Poisson distribution is the mean, the theorem says N_A has the Poisson distribution with parameter $\Lambda(A)$. The function $\Lambda(A)$ is called the *intensity measure* of the process.

An important special case of the Poisson process occurs when the intensity measure is proportional to ordinary measure (length in one dimension, area in two, volume in three, and so forth): if we denote the ordinary measure of a region A by $m(A)$, then

$$\Lambda(A) = \lambda m(A) \tag{4.9}$$

for some $\lambda > 0$. The parameter λ is called the *rate parameter* of the process. A Poisson process for which (4.9) holds, the process is said to be a *homogeneous Poisson process.* Otherwise it is *inhomogeneous.*

The space could be the three-dimensional space of our ordinary experience. For example, the points could be the locations of raisins in a carrot cake. If the process is homogeneous, that models the situation where regions of equal volume have an equal number of raisins on average, as would happen if the batter was stirred well and the raisins didn't settle to the bottom of the cake pan before baking. If the process is inhomogeneous, that models the situation where some regions get more raisins per unit volume than others on average. Either the batter wasn't stirred well or the raisins settled or something of the sort.

There are two important corollaries of the characterization theorem.

Corollary 4.5. *The sum of independent Poisson random variables is a Poisson random variable.*

If $X_i \sim \text{Poi}(\mu_i)$ then the X_i could be the counts N_{A_i} in disjoint regions A_i having measures $m(A_i) = \mu_i$ in a homogeneous Poisson process with unit rate parameter. The sum is the count in the combined region

$$X_1 + \cdots + X_n = N_{A_1 \cup \cdots \cup A_n}$$

which has a Poisson distribution with mean

$$m(A_1 \cup \cdots \cup A_n) = m(A_1) + \cdots m(A_n)$$

because the measure of the union of disjoint regions is the sum of the measures. This is also obvious from linearity of expectation. We must have

$$E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n).$$

Corollary 4.6. *The conditional distribution of a Poisson process in a region A^c given the process in A is the same as the unconditional distribution of the process in A^c.*

In other words, finding the point pattern in A tells you nothing whatsoever about the pattern in A^c. The pattern in A^c has the same distribution conditionally or unconditionally.

Proof. By Definition 4.4.1 and Theorem 4.4 N_B is independent of N_C when $B \subset A^c$ and $C \subset A$. Since this is true for all such C, the random variable N_B is independent of the whole pattern in A, and its conditional distribution given the pattern in A is the same as its unconditional distribution. Theorem 4.4 says Poisson distributions of the N_B for all subsets B of A^c imply that the process in A^c is a Poisson process. □

4.4.3 One-Dimensional Poisson Processes

In this section we consider Poisson processes in one-dimensional space, that is, on the real line. So a realization of the process is a pattern of points on the line. For specificity, we will call the dimension along the line "time" because for many applications it is time. For example, the calls arriving at a telephone exchange are often modeled by a Poisson process. So are the arrivals of customers at a bank teller's window, or at a toll plaza on an toll road. But you should remember that there is nothing in the theory specific to time. The theory is the same for all one-dimensional Poisson processes.

Continuing the time metaphor, the points of the process will always in the rest of this section be called *arrivals*. The time from a fixed point to the next arrival is called the *waiting time* until the arrival.

The special case of the gamma distribution with shape parameter one is called the exponential distribution, denoted $\text{Exp}(\lambda)$. Its density is

$$f(x) = \lambda e^{-\lambda x}, \qquad x > 0. \tag{4.10}$$

Theorem 4.7. *The distribution of the waiting time in a homogeneous Poisson process with rate parameter λ is* $\text{Exp}(\lambda)$. *The distribution is the same unconditionally, or conditional on the past history up to and including the time we start waiting.*

Call the waiting time X and the point where we start waiting a. Fix an $x > 0$, let $A = (a, a + x)$, and let $Y = N_{(a,a+x)}$ be the number of arrivals in the interval A. Then Y has a Poisson distribution with mean $\lambda m(A) = \lambda x$, since measure in one dimension is length. Then the c. d. f. of X is given by

$$\begin{aligned} F(x) &= P(X \le x) \\ &= P(\text{there is at least one arrival in } (a, a + x)) \\ &= P(Y \ge 1) \\ &= 1 - P(Y = 0) \\ &= 1 - e^{-\lambda x} \end{aligned}$$

Differentiating gives the density (4.10).

The assertion about the conditional and unconditional distributions being the same is just the fact that the process on $(-\infty, a]$ is independent of the process on $(a, +\infty)$. Hence the waiting time distribution is the same whether or not we condition on the point pattern in $(-\infty, a]$.

The length of time between two consecutive arrivals is called the *interarrival time*. Theorem 4.7 also gives the distribution of the interarrival times, because it says the distribution is the same whether or not we condition on there being an arrival at the time we start waiting. Finally, the theorem says an interarrival time is independent of any past interarrival times. Since independence is a symmetric property (X is independent of Y if and only if Y is independent of X), this means all interarrival times are independent.

This means we can think of a one-dimensional Poisson process two different ways.

- The number of arrivals in disjoint intervals are independent Poisson random variables. The number of arrivals in an interval of length t is $\text{Poi}(\lambda t)$.

- Starting at an arbitrary point (say time zero), the waiting time to the first arrival is $\text{Exp}(\lambda)$. Then all the successive interarrival times are also $\text{Exp}(\lambda)$. And all the interarrival times are independent of each other and the waiting time to the first arrival.

  Thus if X_1, X_2, ... are i. i. d. $\text{Exp}(\lambda)$ random variables, the times T_1, T_2, ... defined by

  $$T_n = \sum_{i=1}^{n} X_i \tag{4.11}$$

  form a Poisson process on $(0, \infty)$.

Note that by the addition rule for the gamma distribution, the time of the nth arrival is the sum of n i. i. d. $\text{Gam}(1, \lambda)$ random variables and hence has a $\text{Gam}(n, \lambda)$ distribution.

These two ways of thinking give us a c. d. f. for the $\text{Gam}(n, \lambda)$ distribution of T_n.

$$\begin{aligned} F(x) &= P(T_n \le x) \\ &= P(\text{there are at least n arrivals in } (0, x)) \\ &= 1 - P(\text{there are no more than } n-1 \text{ arrivals in } (0, x)) \\ &= 1 - \sum_{k=0}^{n-1} \frac{(\lambda x)^k}{k!} e^{-\lambda x} \end{aligned}$$

Unfortunately, this trick does not work for gamma distributions with noninteger shape parameters. There is no closed form expression for the c. d. f. of a general gamma distribution.

In problems, it is best to use the way of thinking that makes the problem easiest.

Example 4.4.1.
Assume the service times for a bank teller form a homogeneous Poisson process with rate parameter λ. I arrive at the window, and am fifth in line with four people in front of me. What is the expected time until I leave?

There are four interarrival times and the waiting time until the first person in line is finished. All five times are i. i. d. $\text{Exp}(\lambda)$ by the Poisson process assumption. The times have mean $1/\lambda$. The expectation of the sum is the sum of the expectations $5/\lambda$.

Alternatively, the distribution of the time I leave is the sum of the five interarrival and waiting times, which is $\text{Gam}(5, \lambda)$, which has mean $5/\lambda$.

Example 4.4.2.
With the same assumptions in the preceding example, suppose $\lambda = 10$ per hour. What is the probability that I get out in less than a half hour.

This is the probability that there are at least five points of the Poisson process in the interval $(0, 0.5)$, measuring time in hours (the time I leave is the fifth point in the process). The number of points Y has a $\text{Poi}(\lambda t)$ distribution with $t = 0.5$, hence $\lambda t = 5$. From Table II in the back of Lindgren $P(Y \ge 5) = 1 - P(Y \le 4) = 1 - .44 = .56$.

Problems

4-1. Prove Corollary 4.5 for the case of two Poisson random variables directly using the convolution formula Theorem 1.7 from Chapter 1 of these notes. Note that the two Poisson variables are allowed to have different means.
Hint: Use the binomial theorem (Problem 1-14 on p. 7 of Lindgren).

4-2. Suppose X_1, X_2, ... are i. i. d. random variables with mean μ and variance σ^2, and N is a $\text{Geo}(p)$ random variable independent of the X_i. What is the mean and variance of

$$Y = X_1 + X_2 + \cdots + X_N$$

(note N is random).

4-3. A brand of raisin bran averages 84.2 raisins per box. The boxes are filled from large bins of well mixed raisin bran. What is the standard deviation of the number of raisins per box.

4-4. Let X be the number of winners of a lottery. If we assume that players pick their lottery numbers at random, then their choices are i. i. d. random variables and X is binomially distributed. Since the mean number of winners is small, the Poisson approximation is very good. Hence we may assume that $X \sim \operatorname{Poi}(\mu)$ where μ is a constant that depends on the rules of the lottery and the number of tickets sold.

Because of our independence assumption, what other players do is independent of what you do. Hence the conditional distribution of the number of other winners given that you win is also $\operatorname{Poi}(\mu)$. If you are lucky enough to win, you must split the prize with X other winners. You win $A/(X+1)$ where A is the total prize money. Thus

$$E\left(\frac{A}{X+1}\right)$$

is your expected winnings given that you win. Calculate this expectation.

4-5. Suppose X and Y are independent, but not necessarily identically distributed Poisson random variables, and define $N = X + Y$.

(a) Show that

$$X \mid N \sim \operatorname{Bin}(N, p),$$

where p is some function of the parameters of the distributions of X, Y. Specify the function.

(b) Assume

$$Z \mid N \sim \operatorname{Bin}(N, q),$$

where $0 < q < 1$. Show that

$$Z \sim \operatorname{Poi}(\mu),$$

where μ is some function of q and the parameters of the distribution of X, Y. Specify the function.

4-6. Suppose $X \sim \operatorname{Gam}(\alpha, \lambda)$. Let $Y = 1/X$.

(a) For which values of α and λ does $E(Y)$ exist?

(b) What is $E(Y)$ when it exists?

4-7. Suppose that X, Y, and Z are independent $\mathcal{N}(2, 2)$ random variables. What is $P(X > Y + Z)$? **Hint:** What is the distribution of $X - Y - Z$?

Chapter 5

Multivariate Distribution Theory

5.1 Random Vectors

5.1.1 Vectors, Scalars, and Matrices

It is common in linear algebra to refer to single numbers as *scalars* (in contrast to vectors and matrices). So in this chapter a real variable x or a real-valued random variable X will also be referred to as a *scalar variable* or a *scalar random variable*, respectively.

A *matrix* (plural *matrices*) is a rectangular array of scalars, called called the *elements* or *components* of the matrix, considered as a single mathematical object. We use the convention that matrices are denoted by boldface capital letters. The elements of a matrix are indicated by double subscripts, for example the elements of a matrix $\mathbf{A}$ may be denoted a_{ij}. Conventionally, the array is displayed as follows

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ & \vdots & & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{pmatrix} \tag{5.1}$$

The first index indicates the element's row, and the second index indicates the column. The matrix (5.1) has *row dimension* m and *column dimension* n, which is indicated by saying it is an $m \times n$ matrix.

The *transpose* of a matrix $\mathbf{A}$ with elements a_{ij} is the matrix $\mathbf{A}'$ with elements a_{ji}, that is, $\mathbf{A}'$ is obtained from $\mathbf{A}$ by making the rows columns and vice versa.

There are several ways to think of vectors. In the preceeding chapters of these notes we wrote vectors as tuples $\mathbf{x} = (x_1, \ldots, x_n)$. Now we will also consider vectors as special cases of matrices. A *column vector* is an $n \times 1$ matrix

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \tag{5.2}$$

and a *row vector* is a $1 \times n$ matrix

$$\mathbf{x}' = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix} \tag{5.3}$$

Note that (5.2) is indeed the transpose of (5.3) as the notation $\mathbf{x}$ and $\mathbf{x}'$ indicates. Note that even when we consider vectors as special matrices we still use boldface lower case letters for nonrandom vectors, as we always have, rather than the boldface capital letters we use for matrices.

5.1.2 Random Vectors

A *random vector* is just a vector whose components are random scalars. We have always denoted random vectors using boldface capital letters $\mathbf{X} = (X_1, \ldots, X_n)$, which conflicts with the new convention that matrices are boldface capital letters. So when you see a boldface capital letter, you must decide whether this indicates a random vector or a constant (nonrandom) matrix. One hint is that we usually use letters like $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ for random vectors, and we will usually use letters earlier in the alphabet for matrices. If you are not sure what is meant by this notation (or any notation), look at the context, it should be defined nearby.

The *expectation* or *mean* of a random vector $\mathbf{X} = (X_1, \ldots, X_n)$ is defined componentwise. The mean of $\mathbf{X}$ is the vector

$$\boldsymbol{\mu}_{\mathbf{X}} = E(\mathbf{X}) = \bigl(E(X_1), \ldots, E(X_n)\bigr)$$

having components that are the expectations of the corresponding components of $\mathbf{X}$.

5.1.3 Random Matrices

Similarly, we define *random matrix* to be a matrix whose components are random scalars. Let $\mathbf{X}$ denote a random matrix with elements X_{ij}. We can see that the boldface and capital letter conventions have now pooped out. There is no "double bold" or "double capital" type face to indicate the difference between a random vector and a random matrix.[1] The reader will just have to remember in this section $\mathbf{X}$ is a matrix not a vector.

[1] This is one reason to avoid the "vectors are bold" and "random objects are capitals" conventions. They violate "mathematics is invariant under changes of notation." The type face conventions work in simple situations, but in complicated situations they are part of the problem rather than part of the solution. That's why modern advanced mathematics doesn't use the "vectors are bold" convention. It's nineteenth century notation still surviving in statistics.

Again like random vectors, the *expectation* or *mean* of a random matrix is a nonrandom matrix. If $\mathbf{X}$ is a random $m \times n$ matrix with elements X_{ij}, then the mean of $\mathbf{X}$ is the matrix $\mathbf{M}$ with elements

$$\mu_{ij} = E(X_{ij}), \tag{5.4}$$

and we also write $E(\mathbf{X}) = \mathbf{M}$ to indicate all of the mn equations (5.4) with one matrix equation.

5.1.4 Variance Matrices

In the preceding two sections we defined random vectors and random matrices and their expectations. The next topic is variances. One might think that the variance of a random vector should be similar to the mean, a vector having components that are the variances of the corresponding components of $\mathbf{X}$, but it turns out that this notion is not useful. The reason is that variances and covariances are inextricably entangled. We see this in the fact that the variance of a sum involves both variances and covariances (Corollary 2.19 of these notes and the following comments). Thus the following definition.

The *variance matrix* of an n-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_n)$ is the nonrandom $n \times n$ matrix $\mathbf{M}$ having elements

$$m_{ij} = \operatorname{cov}(X_i, X_j). \tag{5.5}$$

As with variances of random scalars, we also use the notation $\operatorname{var}(\mathbf{X})$ for the variance matrix. Note that the diagonal elements of $\mathbf{M}$ are variances because the covariance of a random scalar with itself is the variance, that is,

$$m_{ii} = \operatorname{cov}(X_i, X_i) = \operatorname{var}(X_i).$$

This concept is well established, but the name is not. Lindgren calls $\mathbf{M}$ the *covariance matrix* of $\mathbf{X}$, presumably because its elements are covariances. Other authors call it the *variance-covariance* matrix, because some of its elements are variances too. Some authors, to avoid the confusion about variance, covariance, or variance-covariance, call it the *dispersion* matrix. In my humble opinion, “variance matrix” is the right name because it is the generalization of the variance of a scalar random variable. But you’re entitled to call it what you like. There is no standard terminology.

Example 5.1.1.
What are the mean vector and variance matrix of the random vector (X, X^2), where X is some random scalar? Let

$$\alpha_k = E(X^k)$$

denote the ordinary moments of X. Then, of course, the mean and variance of X are $\mu = \alpha_1$ and

$$\sigma^2 = E(X^2) - E(X)^2 = \alpha_2 - \alpha_1^2,$$

but it will be simpler if we stick to the notation using the α's. The mean vector is

$$\boldsymbol{\mu} = \begin{pmatrix} E(X) \\ E(X^2) \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \tag{5.6}$$

The moment matrix is the 2 × 2 matrix $\mathbf{M}$ with elements

$$\begin{aligned} m_{11} &= \operatorname{var}(X) \\ &= \alpha_2 - \alpha_1^2 \\ m_{22} &= \operatorname{var}(X^2) \\ &= E(X^4) - E(X^2)^2 \\ &= \alpha_4 - \alpha_2^2 \\ m_{12} &= \operatorname{cov}(X, X^2) \\ &= E(X^3) - E(X)E(X^2) \\ &= \alpha_3 - \alpha_1 \alpha_2 \end{aligned}$$

Putting this all together we get

$$\mathbf{M} = \begin{pmatrix} \alpha_2 - \alpha_1^2 & \alpha_3 - \alpha_1 \alpha_2 \\ \alpha_3 - \alpha_1 \alpha_2 & \alpha_4 - \alpha_2^2 \end{pmatrix} \tag{5.7}$$

5.1.5 What is the Variance of a Random Matrix?

By analogy with random vectors, the variance of $\mathbf{X}$ should be a mathematical object with four indexes, the elements being

$$v_{ijkl} = \operatorname{cov}(X_{ij}, X_{kl}).$$

Even naming such an object takes outside the realm of linear algebra. One terminology for objects with more than two indices is *tensors*. So we can say that the variance of a random matrix is a nonrandom tensor. But this doesn't get us anywhere because we don't know anything about operations that apply to tensors.

Thus we see that random matrices present no problem so long as we only are interested in their means, but their variances are problematical. Fortunately, we can avoid random matrices except when we are interested only in their means, not their variances.[2]

[2] A solution to the problem of defining the variance of a random matrix that avoids tensors is to change notation and consider the random matrix a random vector. For example, a random $m \times n$ matrix $\mathbf{X}$ can be written as a vector

$$\mathbf{Y} = (X_{11}, X_{12}, \dots X_{1n}, X_{21}, X_{22}, \dots, X_{2n}, \dots X_{m1}, X_{m2}, \dots X_{mn})$$

So $Y_1 = X_{11}$, $Y_2 = X_{12}$, …, $Y_{n+1} = X_{2n}$, and so forth. Now there is no problem defining the variance matrix of $\mathbf{Y}$, but this is unnatural and clumsy notation that will in most problems make things exceedingly messy.

5.1.6 Covariance Matrices

The *covariance matrix* of an m-dimensional random vector $\mathbf{X}$ and an n-dimensional random vector $\mathbf{Y}$ is the nonrandom matrix $\mathbf{C}$ with elements

$$c_{ij} = \operatorname{cov}(X_i, Y_j), \tag{5.8}$$

(where, as usual X_i is an element of $\mathbf{X}$ and Y_j an element of $\mathbf{Y}$). Note that if $\mathbf{X}$ is an m-dimensional vector and $\mathbf{Y}$ is an n-dimensional vector, then $\mathbf{C} = \operatorname{cov}(\mathbf{X}, \mathbf{Y})$ is an $m \times n$ matrix. Swapping the roles of $\mathbf{X}$ and $\mathbf{Y}$ we see that $\operatorname{cov}(\mathbf{Y}, \mathbf{X})$ is an $n \times m$ matrix. Thus it is obvious that the property $\operatorname{cov}(X, Y) = \operatorname{cov}(Y, X)$ that holds for covariances of scalar random variables, does not hold for covariances of random vectors. In fact, if we write

$$\begin{aligned} \mathbf{C} &= \operatorname{cov}(\mathbf{X}, \mathbf{Y}) \\ \mathbf{D} &= \operatorname{cov}(\mathbf{Y}, \mathbf{X}), \end{aligned}$$

then the elements of $\mathbf{C}$ are given by (5.8) and the elements of $\mathbf{D}$ are

$$d_{ij} = \operatorname{cov}(Y_i, X_j) = c_{ji}$$

Thus the two matrices are transposes of each other: $\mathbf{D} = \mathbf{C}'$.

With these definitions, we can easily generalize most of the formulas about variances and covariances of scalar random variables to vector random variables. We won't bother to go through all of them. The most important one is the formula for the variance of a sum of random vectors.

$$\operatorname{var}\left(\sum_{i=1}^{n} \mathbf{X}_i\right) = \sum_{i=1}^{n} \sum_{j=1}^{n} \operatorname{cov}(\mathbf{X}_i, \mathbf{X}_j) \tag{5.9}$$

which is the same as Corollary 2.19, except that it applies to vector random variables in place of scalar ones. The special case in which $\mathbf{X}_1$, …, $\mathbf{X}_n$ are *uncorrelated* random vectors, meaning $\operatorname{cov}(\mathbf{X}_i, \mathbf{X}_j) = 0$ when $i \neq j$, gives

$$\operatorname{var}\left(\sum_{i=1}^{n} \mathbf{X}_i\right) = \sum_{i=1}^{n} \operatorname{var}(\mathbf{X}_i) \tag{5.10}$$

that is, the variance of the sum is the sum of the variances, which is the same as Corollary 2.21, except that it applies to vector random variables in place of scalar ones.

As with random scalars, independence implies lack of correlation, because $\mathbf{C} = \operatorname{cov}(\mathbf{X}, \mathbf{Y})$ has elements $c_{ij} = \operatorname{cov}(X_i, Y_j)$ which are all zero by this property for random scalars (Theorem 2.47). Hence (5.10) also holds when $\mathbf{X}_1$, …, $\mathbf{X}_n$ are *independent* random vectors. This is by far the most important application of (5.10). As in the scalar case, you should remember

> *Independent implies uncorrelated, but uncorrelated does **not** imply independent.*

Thus independence is a sufficient but not necessary condition for (5.10) to hold. It is enough that the variables be uncorrelated.

In statistics, our main interest is not in sums per se but rather in averages

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i. \tag{5.11a}$$

The analogous formula for random vectors is just the same formula with boldface

$$\overline{\mathbf{X}}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i. \tag{5.11b}$$

Warning: the subscripts on the right hand side in (5.11b) do not indicate components of a vector, rather $\mathbf{X}_1$, $\mathbf{X}_2$, $\ldots$ is simply a sequence of random vectors just as in (5.11a) X_1, X_2, $\ldots$ is a sequence of random scalars. The formulas for the mean and variance of a sum also give us the mean and variance of an average.

Theorem 5.1. *If $\mathbf{X}_1$, $\mathbf{X}_2$, $\ldots$ are random vectors having the same mean vector $\boldsymbol{\mu}$, then*

$$E(\overline{\mathbf{X}}_n) = \boldsymbol{\mu}. \tag{5.12a}$$

If $\mathbf{X}_1$, $\mathbf{X}_2$, $\ldots$ also have the same variance matrix $\mathbf{M}$ and are uncorrelated, then

$$\operatorname{var}(\overline{\mathbf{X}}_n) = \frac{1}{n}\mathbf{M}. \tag{5.12b}$$

This is exactly analogous to the scalar case

$$E(\overline{X}_n) = \mu \tag{5.13a}$$

and

$$\operatorname{var}(\overline{X}_n) = \frac{\sigma^2}{n} \tag{5.13b}$$

Theorem 5.2 (Alternate Variance and Covariance Formulas). *If $\mathbf{X}$ and $\mathbf{Y}$ are random vectors with means $\boldsymbol{\mu}_{\mathbf{X}}$ and $\boldsymbol{\mu}_{\mathbf{Y}}$, then*

$$\operatorname{cov}(\mathbf{X}, \mathbf{Y}) = E\{(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}})'\} \tag{5.14a}$$

$$\operatorname{var}(\mathbf{X}) = E\{(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}})'\} \tag{5.14b}$$

This hardly deserves the name "theorem" since it is obvious once one interprets the matrix notation. If $\mathbf{X}$ is m-dimensional and $\mathbf{Y}$ is n-dimensional, then when we consider the vectors as matrices ("column vectors") we see that the dimensions are

$$\begin{array}{cc} (\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}) & (\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}})' \\ m \times 1 & 1 \times n \end{array}$$

so the "sum" implicit in the matrix multiplication has only one term. Thus (5.14a) is the $m \times n$ matrix with i, j element

$$E\{(X_i - \mu_{X_i})(Y_j - \mu_{Y_j})\} = \operatorname{cov}(X_i, Y_j)$$

and hence is the covariance matrix $\operatorname{cov}(\mathbf{X}, \mathbf{Y})$. Then we see that (5.14b) is just the special case where $\mathbf{Y} = \mathbf{X}$.

5.1.7 Linear Transformations

In this section, we derive the analogs of the formulas

$$E(a + bX) = a + bE(X) \tag{5.15a}$$

$$\operatorname{var}(a + bX) = b^2 \operatorname{var}(X) \tag{5.15b}$$

(Corollary 2.2 and Theorem 2.13 in Chapter 2 of these notes) that describe the moments of a linear transformation of a random variable. A general linear transformation has the form

$$\mathbf{y} = \mathbf{a} + \mathbf{B}\mathbf{x}$$

where $\mathbf{y}$ and $\mathbf{a}$ are m-dimensional vectors, $\mathbf{B}$ is an $m \times n$ matrix, and $\mathbf{x}$ is an n-dimensional vector. The dimensions of each object, considering the vectors as column vectors (that is, as matrices with just a single column), are

$$\begin{array}{ccccccc} \mathbf{y} & = & \mathbf{a} & + & \mathbf{B} & \mathbf{a} \\ m \times 1 & & m \times 1 & & m \times n & n \times 1 \end{array} \tag{5.16}$$

Note that the column dimension of $\mathbf{B}$ and the row dimension of $\mathbf{x}$ must agree, as in any matrix multiplication. Also note that the dimensions of $\mathbf{x}$ and $\mathbf{y}$ are not the same. We are mapping n-dimensional vectors to m-dimensional vectors.

Theorem 5.3. *If $\mathbf{Y} = \mathbf{a} + \mathbf{B}\mathbf{X}$, where $\mathbf{a}$ is a constant vector, $\mathbf{B}$ is a constant matrix, and $\mathbf{X}$ is a random vector, then*

$$E(\mathbf{Y}) = \mathbf{a} + \mathbf{B}E(\mathbf{X}) \tag{5.17a}$$

$$\operatorname{var}(\mathbf{Y}) = \mathbf{B} \operatorname{var}(\mathbf{X})\mathbf{B}' \tag{5.17b}$$

If we write $\boldsymbol{\mu}_{\mathbf{X}}$ and $\mathbf{M}_{\mathbf{X}}$ for the mean and variance of $\mathbf{X}$ and similarly for $\mathbf{Y}$, then (5.17a) and (5.17b) become

$$\boldsymbol{\mu}_{\mathbf{Y}} = \mathbf{a} + \mathbf{B}\boldsymbol{\mu}_{\mathbf{X}} \tag{5.18a}$$

$$\mathbf{M}_{\mathbf{Y}} = \mathbf{B}\mathbf{M}_{\mathbf{X}}\mathbf{B}' \tag{5.18b}$$

If we were to add dimension information to (5.18a), it would look much like (5.16). If we add such information to (5.18b) it becomes

$$\begin{array}{ccccc} \mathbf{M}_{\mathbf{Y}} & = & \mathbf{B} & \mathbf{M}_{\mathbf{X}} & \mathbf{B}' \\ m \times m & & m \times n & n \times n & n \times m \end{array}$$

Note again that, as in any matrix multiplication, the column dimension of the left hand factor agrees with row dimension of the right hand factor. In particular, the column dimension of $\mathbf{B}$ is the row dimension of $\mathbf{M}_{\mathbf{X}}$, and the column dimension of $\mathbf{M}_{\mathbf{X}}$ is the row dimension of $\mathbf{B}'$. Indeed, this is the only way these matrices can be multiplied together to get a result of the appropriate dimension. So merely getting the dimensions right tells you what the formula has to be.

Proof of Theorem 5.3. Since our only definition of the mean of a random vector involves components, we will have to prove this componentwise. The component equations of $\mathbf{Y} = \mathbf{a} + \mathbf{B}\mathbf{X}$ are

$$Y_i = a_i + \sum_{j=1}^{n} b_{ij} X_j$$

(where, as usual, the a_i are the components of $\mathbf{a}$, the b_{ij} are the components of $\mathbf{B}$, and so forth). Applying linearity of expectation for scalars gives

$$E(Y_i) = a_i + \sum_{j=1}^{n} b_{ij} E(X_j),$$

which are the component equations of (5.18a).

Now we can be a bit slicker about the second half of the proof using the alternate variance formula (5.14b).

$$\begin{aligned}
\operatorname{var}(\mathbf{a} + \mathbf{B}\mathbf{X}) &= E\{(\mathbf{a} + \mathbf{B}\mathbf{X} - \boldsymbol{\mu}_{\mathbf{a}+\mathbf{B}\mathbf{X}})(\mathbf{a} + \mathbf{B}\mathbf{X} - \boldsymbol{\mu}_{\mathbf{a}+\mathbf{B}\mathbf{X}})'\} \\
&= E\{(\mathbf{B}\mathbf{X} - \mathbf{B}\boldsymbol{\mu}_{\mathbf{X}})(\mathbf{B}\mathbf{X} - \mathbf{B}\boldsymbol{\mu}_{\mathbf{X}})'\} \\
&= E\{\mathbf{B}(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})'\mathbf{B}'\} \\
&= \mathbf{B}E\{(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})'\}\mathbf{B}'
\end{aligned}$$

Going from the first line to the second is just (5.18a). Going from the second line to the third uses the fact that the transpose of a matrix product is the product of the transposes in reverse order, that is, $(\mathbf{B}\mathbf{C})' = \mathbf{C}'\mathbf{B}$. And going from the third line to the forth uses (5.18a) again to pull the constant matrices outside the expectation. □

Of particular interest is the special case in which the linear transformation is scalar-valued, that is, $m = 1$ in (5.16). Then the matrix $\mathbf{B}$ must be $1 \times n$, hence a row vector. We usually write row vectors as transposes, say $\mathbf{c}'$, because convention requires unadorned vectors like $\mathbf{c}$ to be column vectors. Thus we write $\mathbf{B} = \mathbf{c}'$ and obtain

Corollary 5.4. *If $Y = a + \mathbf{c}'\mathbf{X}$, where a is a constant scalar, $\mathbf{c}$ is a constant vector, and $\mathbf{X}$ is a random vector, then*

$$E(Y) = a + \mathbf{c}'E(\mathbf{X}) \tag{5.19a}$$

$$\operatorname{var}(Y) = \mathbf{c}' \operatorname{var}(\mathbf{X})\mathbf{c} \tag{5.19b}$$

Or, if you prefer the other notation

$$\mu_Y = a + \mathbf{c}'\boldsymbol{\mu}_{\mathbf{X}} \tag{5.20a}$$

$$\sigma_Y^2 = \mathbf{c}'\mathbf{M}_{\mathbf{X}}\mathbf{c} \tag{5.20b}$$

Note that, since $m = 1$, both Y and a are scalars (1×1 matrices), so we have written them in normal (not boldface) type and used the usual notation σ_Y^2 for the variance of a scalar. Also note that because $\mathbf{B} = \mathbf{c}'$ the transposes have switched sides in going from (5.18b) to (5.20b).

Example 5.1.2.
(This continues Example 5.1.1.) What are the mean and variance of $X + X^2$, where X is some random scalar? We don't have to use an multivariate theory to answer this question. We could just use the formulas for the mean and variance of a sum of random variables from Chapter 2 of these notes. But here we want to use the multivariate theory to illustrate how it works.

Write $Y = X + X^2$ and let

$$\mathbf{Z} = \begin{pmatrix} X \\ X^2 \end{pmatrix}$$

be the random vector whose mean vector and variance matrix were found in Example 5.1.1. Then $Y = \mathbf{u}'\mathbf{Z}$, where

$$\mathbf{u} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Thus by (5.20a) and (5.6)

$$E(Y) = \mathbf{u}'\boldsymbol{\mu}_\mathbf{Z} = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \alpha_1 + \alpha_2$$

And by (5.20b) and (5.7)

$$\begin{aligned} \operatorname{var}(Y) &= \mathbf{u}'\mathbf{M}_\mathbf{Z}\mathbf{u} \\ &= \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} \alpha_2 - \alpha_1^2 & \alpha_3 - \alpha_1\alpha_2 \\ \alpha_3 - \alpha_1\alpha_2 & \alpha_4 - \alpha_2^2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ &= \alpha_2 - \alpha_1^2 + 2(\alpha_3 - \alpha_1\alpha_2) + \alpha_4 - \alpha_2^2 \end{aligned}$$

Alternate Solution We could also do this problem the "old fashioned way" (without matrices)

$$\begin{aligned} \operatorname{var}(X + X^2) &= \operatorname{var}(X) + 2\operatorname{cov}(X, X^2) + \operatorname{var}(X^2) \\ &= (\alpha_2 - \alpha_1^2) + 2(\alpha_3 - \alpha_1\alpha_2) + (\alpha_4 - \alpha_2^2) \end{aligned}$$

Of course, both ways must give the same answer. We're just using matrices here to illustrate the use of matrices.

5.1.8 Characterization of Variance Matrices

A matrix $\mathbf{A}$ is said to be *positive semi-definite* if

$$\mathbf{c}'\mathbf{A}\mathbf{c} \geq 0, \qquad \text{for every vector } \mathbf{c} \tag{5.21}$$

and *positive definite* if

$$\mathbf{c}'\mathbf{A}\mathbf{c} > 0, \qquad \text{for every nonzero vector } \mathbf{c}.$$

Corollary 5.5. *The variance matrix of any random vector is symmetric and positive semi-definite.*

Proof. Symmetry follows from the symmetry property of covariances of random scalars: $\operatorname{cov}(X_i, X_j) = \operatorname{cov}(X_j, X_i)$.

The random scalar Y in Corollary 5.4 must have nonnegative variance. Thus (5.19b) implies $\mathbf{c}' \operatorname{var}(\mathbf{X})\mathbf{c} \ge 0$. Since $\mathbf{c}$ was an arbitrary vector, this proves $\operatorname{var}(\mathbf{X})$ is positive semi-definite. □

The corollary says that a *necessary condition* for a matrix $\mathbf{M}$ to be the variance matrix of some random vector $\mathbf{X}$ is that $\mathbf{M}$ be symmetric and positive semi-definite. This raises the obvious question: is this a *sufficient condition*, that is, for any symmetric and positive semi-definite matrix $\mathbf{M}$ does there exist a random vector $\mathbf{X}$ such that $\mathbf{M} = \operatorname{var}(\mathbf{X})$? We can't address this question now, because we don't have enough examples of random vectors for which we know the distributions. It will turn out that the answer to the sufficiency question is "yes." When we come to the multivariate normal distribution (Section 5.2) we will see that for any symmetric and positive semi-definite matrix $\mathbf{M}$ there is a multivariate normal random vector $\mathbf{X}$ such that $\mathbf{M} = \operatorname{var}(\mathbf{X})$.

A *hyperplane* in n-dimensional space $\mathbb{R}^n$ is a set of the form

$$H = \{\, \mathbf{x} \in \mathbb{R}^n : \mathbf{c}'\mathbf{x} = a \,\} \tag{5.22}$$

for some nonzero vector $\mathbf{c}$ and some scalar a. We say a random vector $\mathbf{X}$ is *concentrated* on the hyperplane H if $P(\mathbf{X} \in H) = 1$. Another way of describing the same phenomenon is to say that that H is a *support* of $\mathbf{X}$.

Corollary 5.6. *The variance matrix of a random vector* $\mathbf{X}$ *is positive definite if and only if* $\mathbf{X}$ *is not concentrated on any hyperplane.*

Proof. We will prove the equivalent statement that the variance matrix is *not* positive definite if and only if is *is* concentrated on some hyperplane.

First, suppose that $\mathbf{M} = \operatorname{var}(\mathbf{X})$ is not positive definite. Then there is some nonzero vector $\mathbf{c}$ such that $\mathbf{c}'\mathbf{M}\mathbf{c} = 0$. Consider the random scalar $Y = \mathbf{c}'\mathbf{X}$. By Corollary 5.4 $\operatorname{var}(Y) = \mathbf{c}'\mathbf{M}\mathbf{c} = 0$. Now by Corollary 2.34 of these notes $Y = \mu_Y$ with probability one. Since $E(Y) = \mathbf{c}'\mu_{\mathbf{X}}$ by (5.19a), this says that $\mathbf{X}$ is concentrated on the hyperplane (5.22) where $a = \mathbf{c}'\mu_{\mathbf{X}}$.

Conversely, suppose that $\mathbf{X}$ is concentrated on the hyperplane (5.22). Then the random scalar $Y = \mathbf{c}'\mathbf{x}$ is concentrated at the point a, and hence has variance zero, which is $\mathbf{c}'\mathbf{M}\mathbf{c}$ by Corollary 5.4. Thus $\mathbf{M}$ is not positive definite. □

5.1.9 Degenerate Random Vectors

Random vectors are sometimes called *degenerate* by those who believe in the kindergarten principle of calling things we don't like bad names. And why wouldn't we like a random vector concentrated on a hyperplane? Because it doesn't have a density. A hyperplane is a set of measure zero, hence any integral over the hyperplane is zero and cannot be used to define probabilities and expectations.

Example 5.1.3 (A Degenerate Random Vector).
Suppose U, V, and W are independent and identically distributed random variables having a distribution not concentrated at one point, so $\sigma^2 = \operatorname{var}(U) = \operatorname{var}(V) = \operatorname{var}(W)$ is strictly positive. Consider the random vector

$$\mathbf{X} = \begin{pmatrix} U - V \\ V - W \\ W - U \end{pmatrix} \tag{5.23}$$

Because of the assumed independence of U, V, and W, the diagonal elements of $\operatorname{var}(\mathbf{X})$ are all equal to

$$\operatorname{var}(U - V) = \operatorname{var}(U) + \operatorname{var}(V) = 2\sigma^2$$

and the off-diagonal elements are all equal to

$$\operatorname{cov}(U - V, V - W) = \operatorname{cov}(U, V) - \operatorname{cov}(U, W) - \operatorname{var}(V) + \operatorname{cov}(V, W) = -\sigma^2$$

Thus

$$\operatorname{var}(\mathbf{X}) = \sigma^2 \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}$$

Question Is $\mathbf{X}$ degenerate or non-degenerate? If degenerate, what hyperplane or hyperplanes is it concentrated on?

Answer We give two different ways of finding this out. The first uses some mathematical cleverness, the second brute force and ignorance (also called plug and chug).

The first way starts with the observation that each of the variables U, V, and W occurs twice in the components of $\mathbf{X}$, once with each sign, so the sum of the components of $\mathbf{X}$ is zero, that is $X_1 + X_2 + X_3 = 0$ with probability one. But if we introduce the vector

$$\mathbf{u} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

we see that $X_1 + X_2 + X_3 = \mathbf{u}'\mathbf{X}$. Hence $\mathbf{X}$ is concentrated on the hyperplane defined by

$$H = \{\, \mathbf{x} \in \mathbb{R}^3 : \mathbf{u}'\mathbf{x} = 0 \,\}$$

or if you prefer

$$H = \{\, (x_1, x_2, x_3) \in \mathbb{R}^3 : x_1 + x_2 + x_3 = 0 \,\}.$$

Thus we see that $\mathbf{X}$ is indeed degenerate (concentrated on H). Is is concentrated on any other hyperplanes? The answer is no, but our cleverness has run out. It's hard so show that there are no more except by the brute force approach.

The brute force approach is to find the eigenvalues and eigenvectors of the variance matrix. The random vector in question is concentrated on hyperplanes defined by eigenvectors corresponding to zero eigenvalues (Lemma 5.7 below). Eigenvalues and eigenvectors can be found by many numerical math packages. Here we will just demonstrate doing it in R.

```
> M <- matrix(c(2, -1, -1, -1, 2, -1, -1, -1, 2), nrow=3)
> M
     [,1] [,2] [,3]
[1,]    2   -1   -1
[2,]   -1    2   -1
[3,]   -1   -1    2
> eigen(M)
$values
[1]  3.000000e+00  3.000000e+00 -8.881784e-16

$vectors
           [,1]       [,2]      [,3]
[1,]  0.8156595  0.0369637 0.5773503
[2,] -0.3758182 -0.7248637 0.5773503
[3,] -0.4398412  0.6879000 0.5773503
```

Each eigenvector corresponding to a zero eigenvalue is a vector $\mathbf{c}$ defining a hyperplane by (5.22) on which the random vector is concentrated. There is just one zero eigenvalue. The corresponding eigenvector is

$$\mathbf{c} = \begin{pmatrix} 0.5773503 \\ 0.5773503 \\ 0.5773503 \end{pmatrix}$$

(the eigenvectors are the columns of the `$vectors` matrix returned by the `eigen` function). Since $\mathbf{c}$ is a multiple of $\mathbf{u}$ in the first answer, they define the same hyperplane. Since there is only one zero eigenvalue, there is only one hyperplane supporting the random vector.

Lemma 5.7. *A random vector* $\mathbf{X}$ *is concentrated on a hyperplane* (5.22) *if and only if the vector* $\mathbf{c}$ *in* (5.22) *is an eigenvector of* $\operatorname{var}(\mathbf{X})$ *corresponding to a zero eigenvalue.*

Proof. First suppose $\mathbf{c}$ is an eigenvector of $\mathbf{M} = \operatorname{var}(\mathbf{X})$ corresponding to a zero eigenvalue. This means $\mathbf{M}\mathbf{c} = 0$, which implies $\mathbf{c}'\mathbf{M}\mathbf{c} = 0$, which, as in the proof of Corollary 5.6, implies that $\mathbf{X}$ is concentrated on the hyperplane defined by (5.22).

Conversely, suppose $\mathbf{X}$ is concentrated on the hyperplane defined by (5.22), which, as in the proof of Corollary 5.6, implies $\mathbf{c}'\mathbf{M}\mathbf{c} = 0$. Write, using the spectral decomposition (Theorem E.4 in Appendix E) $\mathbf{M} = \mathbf{O}\mathbf{D}\mathbf{O}'$, where $\mathbf{D}$ is diagonal and $\mathbf{O}$ is orthogonal. Then

$$0 = \mathbf{c}'\mathbf{M}\mathbf{c} = \mathbf{c}'\mathbf{O}\mathbf{D}\mathbf{O}'\mathbf{c} = \mathbf{w}'\mathbf{D}\mathbf{w}$$

where we have written $\mathbf{w} = \mathbf{O}'\mathbf{c}$. Writing out the matrix multiplications with subscripts

$$\mathbf{w}'\mathbf{D}\mathbf{w} = \sum_i d_{ii} w_i^2 = 0$$

which implies, since $d_{ii} \geq 0$ for all i that

$$d_{ii} = 0 \quad \text{or} \quad w_i = 0, \qquad \text{for all } i$$

and this implies that actually $\mathbf{D}\mathbf{w} = 0$. Hence, plugging back in the definition of $\mathbf{w}$, that $\mathbf{D}\mathbf{O}'\mathbf{c} = 0$, and, multiplying on the left by $\mathbf{O}$, that

$$\mathbf{M}\mathbf{c} = \mathbf{O}\mathbf{D}\mathbf{O}'\mathbf{c} = 0$$

which says that $\mathbf{c}$ is an eigenvector of $\mathbf{M}$ corresponding to a zero eigenvalue, which is what we were proving. □

Degeneracy is not solely a phenomenon of concentration on hyperplanes. We say a random vector is degenerate if it is concentrated on any set of measure zero.

Example 5.1.4.
In Example 2.7.2 we considered the random vector $\mathbf{Z} = (X, Y)$, where $Y = X^2$ and X was any nonconstant random variable having a distribution symmetric about zero. It served there as an example of random variables X and Y that were uncorrelated but not independent.

Here we merely point out that the random vector $\mathbf{Z}$ is degenerate, because it is clearly concentrated on the parabola

$$S = \{\, (x, y) \in \mathbb{R}^2 : y = x^2 \,\}$$

which is, being a one-dimensional curve in $\mathbb{R}^2$, a set of measure zero.

So how does one handle degenerate random vectors? If they don't have densities, and most of the methods we know involve densities, what do we do? First let me remind you that we do know some useful methods that don't involve densities.

- The first part of Chapter 2 of these notes, through Section 2.4 never mentions densities. The same goes for Sections 3.3 and 3.5 in Chapter 3.

- In order to calculate $E(\mathbf{Y})$ where $\mathbf{Y} = \mathbf{g}(\mathbf{X})$, you don't need the density of $\mathbf{Y}$. You can use

  $$E(\mathbf{Y}) = \int \mathbf{g}(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}$$

  instead. Thus even if $\mathbf{Y}$ is degenerate, but is a known function of some non-degenerate random vector $\mathbf{X}$, we are still in business.

When a random vector $\mathbf{X}$ is degenerate, it is always possible in theory (not necessarily in practice) to eliminate one of the variables. For example, if $\mathbf{X}$ is concentrated on the hyperplane H defined by (5.22), then, since $\mathbf{c}$ is nonzero, it has at least one nonzero component, say c_j. Then rewriting $\mathbf{c}'\mathbf{x} = a$ with an explicit sum we get

$$\sum_{i=1}^{n} c_i X_i = a,$$

which can be solved for X_j

$$X_j = \frac{1}{c_j} \left(a - \sum_{\substack{i=1 \\ i \neq j}}^{n} c_i X_i \right)$$

Thus we can eliminate X_j and work with the remaining variables. If the random vector

$$\mathbf{X}' = (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n)$$

of the remaining variables is non-degenerate, then it has a density. If $\mathbf{X}'$ is still degenerate, then there is another variable we can eliminate. Eventually, unless $\mathbf{X}$ is a constant random vector, we get to some subset of variables that have a non-degenerate joint distribution and hence a density. Since the rest of the variables are a function of this subset, that indirectly describes all the variables.

Example 5.1.5 (Example 5.1.3 Continued).
In Example 5.1.3 we considered the random vector

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} U - V \\ V - W \\ W - U \end{pmatrix}$$

where U, V, and W are independent and identically distributed random variables. Now suppose they are independent standard normal.

In Example 5.1.3 we saw that $\mathbf{X}$ was degenerate because $X_1 + X_2 + X_3 = 0$ with probability one. We can eliminate X_3, since

$$X_3 = -(X_1 + X_2)$$

and consider the distribution of the vector (X_1, X_2), which we will see (in Section 5.2 below) has a non-degenerate multivariate normal distribution.

5.1.10 Correlation Matrices

If $\mathbf{X} = (X_1, \ldots, X_n)$ is a random vector having no constant components, that is, $\operatorname{var}(X_i) > 0$ for all i, the *correlation matrix* of $\mathbf{X}$ is the $n \times n$ matrix $\mathbf{C}$ with elements

$$c_{ij} = \frac{\operatorname{cov}(X_i, X_j)}{\sqrt{\operatorname{var}(X_i) \operatorname{var}(X_i)}} = \operatorname{cor}(X_i, X_j)$$

If $\mathbf{M} = \operatorname{var}(\mathbf{X})$ has elements m_{ij}, then

$$c_{ij} = \frac{m_{ij}}{\sqrt{m_{ii} m_{jj}}}$$

Note that the diagonal elements c_{ii} of a correlation matrix are all equal to one, because the correlation of any random variable with itself is one.

Theorem 5.8. *Every correlation matrix is positive semi-definite. The correlation matrix of a random vector* $\mathbf{X}$ *is positive definite if and only the variance matrix of* $\mathbf{X}$ *is positive definite.*

Proof. This follows from the analogous facts about variance matrices. □

It is important to understand that the requirement that a variance matrix (or correlation matrix) be positive semi-definite is a much stronger requirement than the correlation inequality (correlations must be between -1 and $+1$). The two requirements are related: positive semi-definiteness implies the correlation inequality, but not vice versa. That positive semi-definiteness implies the correlation inequality is left as an exercise (Problem 5-4). That the two conditions are not equivalent is shown by the following example.

Example 5.1.6 (All Correlations the Same).
Suppose $\mathbf{X} = (X_1, \ldots, X_n)$ is a random vector and all the components have the same correlation, as would be the case if the components are *exchangeable* random variables, that is, $\operatorname{cor}(X_i, X_j) = \rho$ for all i and j with $i \neq j$. Then the correlation matrix of $\mathbf{X}$ has one for all diagonal elements and ρ for all off-diagonal elements. In Problem 2-22 it is shown that positive definiteness of the correlation matrix requires

$$-\frac{1}{n-1} \leq \rho.$$

This is an additional inequality not implied by the correlation inequality.

The example says there is a limit to how negatively correlated a sequence of exchangeable random variables can be. But even more important than this specific discovery, is the general message that there is more to know about correlations than that they are always between -1 and $+1$. The requirement that a correlation matrix (or a variance matrix) be positive semi-definite is much stronger. It implies a lot of other inequalities. In fact it implies an infinite family of inequalities: $\mathbf{M}$ is positive semi-definite only if $\mathbf{c}'\mathbf{M}\mathbf{c} \geq 0$ for every vector $\mathbf{c}$. That's a different inequality for every vector $\mathbf{c}$ and there are infinitely many such vectors.

5.2 The Multivariate Normal Distribution

The *standard multivariate normal distribution* is the distribution of the random vector $\mathbf{Z} = (Z_1, \ldots, Z_n)$ having independent and identically $\mathcal{N}(0, 1)$ distributed components. Its density is, of course,

$$f_{\mathbf{Z}}(\mathbf{z}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-z_i^2/2} = \frac{1}{(2\pi)^{n/2}} e^{-\mathbf{z}'\mathbf{z}/2}, \qquad \mathbf{z} \in \mathbb{R}^n$$

Note for future reference that

$$\begin{aligned} E(\mathbf{Z}) &= 0 \\ \operatorname{var}(\mathbf{Z}) &= \mathbf{I} \end{aligned}$$

where $\mathbf{I}$ denotes an identity matrix. These are obvious from the definition of $\mathbf{Z}$. Its components are independent and standard normal, hence have mean zero, variance one, and covariances zero. Thus the variance matrix has ones on the diagonal and zeroes off the diagonal, which makes it an identity matrix.

As in the univariate case, we call a linear transformation of a standard normal random vector a (general) normal random vector. If we define $\mathbf{X} = \mathbf{a} + \mathbf{B}\mathbf{Z}$, then by Theorem 5.3

$$\begin{aligned} E(\mathbf{X}) &= \mathbf{a} + \mathbf{B}E(\mathbf{Z}) \\ &= \mathbf{a} \\ \operatorname{var}(\mathbf{X}) &= \mathbf{B} \operatorname{var}(\mathbf{Z})\mathbf{B}' \\ &= \mathbf{B}\mathbf{B}' \end{aligned}$$

We say that $\mathbf{X}$ has the multivariate normal distribution with mean (vector) $\mathbf{a}$ and variance (matrix) $\mathbf{M} = \mathbf{B}\mathbf{B}'$, and abbreviate it as $\mathcal{N}_n(\mathbf{a}, \mathbf{M})$ if we want to emphasize the dimension n of the random vector, or just as $\mathcal{N}(\mathbf{a}, \mathbf{M})$ if we don't want to explicitly note the dimension. No confusion should arise between the univariate and multivariate case, because the parameters are scalars in the univariate case and a vector and a matrix in the multivariate case and are clearly distinguishable by capitalization and type face.

Lemma 5.9. *If* $\mathbf{M}$ *is a positive semi-definite matrix, then there exists a normal random vector* $\mathbf{X}$ *such that* $E(\mathbf{X}) = \mu$ *and* $\operatorname{var}(\mathbf{X}) = \mathbf{M}$.

Proof. In Corollary E.7 in Appendix E the symmetric square root $\mathbf{M}^{1/2}$ of $\mathbf{M}$ is defined. Now define $\mathbf{X} = \boldsymbol{\mu} + \mathbf{M}^{1/2}\mathbf{Z}$, where $\mathbf{Z}$ is multivariate standard normal. Then by Theorem 5.3

$$E(\mathbf{X}) = \boldsymbol{\mu} + \mathbf{M}^{1/2} E(\mathbf{Z}) = \boldsymbol{\mu}$$

and

$$\operatorname{var}(\mathbf{X}) = \mathbf{M}^{1/2} \operatorname{var}(\mathbf{Z}) \mathbf{M}^{1/2} = \mathbf{M}^{1/2}\mathbf{I}\mathbf{M}^{1/2} = \mathbf{M}^{1/2}\mathbf{M}^{1/2} = \mathbf{M}$$

□

5.2.1 The Density of a Non-Degenerate Normal Random Vector

How about the density of the multivariate normal distribution? First we have to say that it may not have a density. If the variance parameter $\mathbf{M}$ is not positive definite, then the random vector will be concentrated on a hyperplane (will be degenerate) by Theorem 5.6, in which case it won't have a density. Otherwise, it will.

Another approach to the same issue is to consider that $\mathbf{X}$ will have support on the whole of $\mathbb{R}^n$ only if the transformation

$$\mathbf{g}(\mathbf{z}) = \mathbf{a} + \mathbf{B}\mathbf{z}$$

is invertible, in which case its inverse is

$$\mathbf{h}(\mathbf{x}) = \mathbf{B}^{-1}(\mathbf{x} - \mathbf{a})$$

and has derivative matrix

$$\nabla \mathbf{h}(\mathbf{x}) = \mathbf{B}^{-1}$$

Thus we find the density of $\mathbf{X}$ by the multivariate change of variable theorem (Corollary 1.6 of Chapter 1 of these notes)

$$\begin{aligned}
f_{\mathbf{X}}(\mathbf{x}) &= f_{\mathbf{Z}}[\mathbf{h}(\mathbf{x})] \cdot \bigl|\det\bigl(\nabla \mathbf{h}(\mathbf{x})\bigr)\bigr| . \\
&= f_{\mathbf{Z}}\bigl(\mathbf{B}^{-1}(\mathbf{x} - \mathbf{a})\bigr) \cdot \bigl|\det\bigl(\mathbf{B}^{-1}\bigr)\bigr| . \\
&= \frac{\bigl|\det\bigl(\mathbf{B}^{-1}\bigr)\bigr|}{(2\pi)^{n/2}} \exp\bigl(-\tfrac{1}{2}[\mathbf{B}^{-1}(\mathbf{x} - \mathbf{a})]'\mathbf{B}^{-1}(\mathbf{x} - \mathbf{a})\bigr) \\
&= \frac{\bigl|\det\bigl(\mathbf{B}^{-1}\bigr)\bigr|}{(2\pi)^{n/2}} \exp\bigl(-\tfrac{1}{2}(\mathbf{x} - \mathbf{a})'(\mathbf{B}^{-1})'\mathbf{B}^{-1}(\mathbf{x} - \mathbf{a})\bigr)
\end{aligned}$$

Now we need several facts about matrices and determinants to clean this up. First, $(\mathbf{B}^{-1})'\mathbf{B}^{-1} = \mathbf{M}^{-1}$, where, as above, $\mathbf{M} = \operatorname{var}(\mathbf{X})$ because of two facts about inverses, transposes, and products.

- The inverse of a transpose is the transpose of the inverse.

  Hence $(\mathbf{B}^{-1})' = (\mathbf{B}')^{-1}$

- The inverse of a product is the product of the inverses in reverse order, that is, $(\mathbf{CD})^{-1} = \mathbf{D}^{-1}\mathbf{C}^{-1}$ for any invertible matrices $\mathbf{C}$ and $\mathbf{D}$.

  Hence $(\mathbf{B}')^{-1}\mathbf{B}^{-1} = (\mathbf{B}\mathbf{B}')^{-1} = \mathbf{M}^{-1}$.

Second, $|\det\bigl(\mathbf{B}^{-1}\bigr)| = \det(\mathbf{M})^{-1/2}$ because of two facts about determinants, inverses, and products.

- The determinant of an inverse is the multiplicative inverse (reciprocal) of the determinant.

  Hence $\det\bigl(\mathbf{B}^{-1}\bigr) = \det(\mathbf{B})^{-1}$.

- The determinant of a matrix and its transpose are the same.

  Hence $\det(\mathbf{B}) = \det(\mathbf{B}')$.

- The determinant of a product is the product of the determinants, that is, $\det(\mathbf{CD}) = \det(\mathbf{C}) \det(\mathbf{D})$ for any matrices $\mathbf{C}$ and $\mathbf{D}$.

  Hence $\det(\mathbf{M}) = \det(\mathbf{BB}') = \det(\mathbf{B})^2$.

Putting this all together, we get

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} \det(\mathbf{M})^{1/2}} \exp\left(-\tfrac{1}{2}(\mathbf{x}-\mathbf{a})'\mathbf{M}^{-1}(\mathbf{x}-\mathbf{a})\right), \qquad \mathbf{x} \in \mathbb{R}^n \quad (5.24)$$

Note that the formula does not involve $\mathbf{B}$. The distribution does indeed only depend on the parameters $\mathbf{a}$ and $\mathbf{M}$ as the notation $\mathcal{N}_n(\mathbf{a}, \mathbf{M})$ implies.

Recall from Lemma 5.9 that there exists a $\mathcal{N}(\mathbf{a}, \mathbf{M})$ for every vector $\mathbf{a}$ and every symmetric positive semi-definite matrix $\mathbf{M}$. If $\mathbf{M}$ is not positive definite, then the distribution is degenerate and has no density. Otherwise, it has the density (5.24).

While we are on the subject, we want to point out that every density that looks like even vaguely like (5.24) is multivariate normal. Of course, we will have to be a bit more precise than "even vaguely like" to get a theorem. A general *quadratic form* is a function $q : \mathbb{R}^n \to \mathbb{R}$ defined by

$$q(\mathbf{x}) = \tfrac{1}{2}\mathbf{x}'\mathbf{A}\mathbf{x} + \mathbf{b}'\mathbf{x} + c \qquad (5.25)$$

where $\mathbf{A}$ is an $n \times n$ matrix, $\mathbf{b}$ is an n vector, and c is a scalar. There is no loss of generality in assuming $\mathbf{A}$ is symmetric, because

$$\tfrac{1}{2}\mathbf{x}'\mathbf{A}\mathbf{x} = \tfrac{1}{2}\mathbf{x}'\mathbf{A}'\mathbf{x} = \mathbf{x}'(\mathbf{A} + \mathbf{A}')\mathbf{x},$$

the first equality following from the rule for the transpose of a product, and the second equality coming from averaging the two sides of the first equality. The matrix in the middle of the expression on the right hand side is symmetric. If we replaced $\mathbf{A}$ in the definition of q by the symmetric matrix $\frac{1}{2}(\mathbf{A} + \mathbf{A}')$, we would still be defining the same function. Thus we assume from here on that the matrix in the definition of any quadratic form is symmetric.

Theorem 5.10. *If q is a quadratic form defined by* (5.25) *and*

$$f(\mathbf{x}) = e^{-q(\mathbf{x})}, \qquad \mathbf{x} \in \mathbb{R}^n$$

is the probability density of a random variable $\mathbf{X}$, then

(a) $\mathbf{A}$ *is positive definite,*

(b) $\mathbf{X}$ *has a non-degenerate multivariate normal distribution,*

(c) $\operatorname{var}(\mathbf{X}) = \mathbf{A}^{-1}$*, and*

(d) $E(\mathbf{X}) = -\mathbf{A}^{-1}\mathbf{b}$.

Proof. The proof that $\mathbf{A}$ is positive definite has to do with the existence of the integral $\int f(\mathbf{x})\,d\mathbf{x} = 1$. We claim that unless $\mathbf{A}$ is positive definite the integral does not exist and cannot define a probability density.

First note that, since the density is continuous, it is bounded on bounded sets. We only need to worry about the behavior of the integrand near infinity. Second, since

$$\frac{f(\mathbf{x})}{e^{-\mathbf{x}'\mathbf{A}\mathbf{x}/2}} \to 1, \qquad \text{as } \mathbf{x} \to \infty,$$

we may in determining when the integral exists consider only the quadratic part in the definition of q. Let $\mathbf{A} = \mathbf{ODO}'$ be the spectral decomposition (Theorem E.4 in Appendix E) of $\mathbf{A}$, and consider the change of variables $\mathbf{y} = \mathbf{O}'\mathbf{x}$, which has inverse transformation $\mathbf{x} = \mathbf{O}\mathbf{y}$ and Jacobian one. Using this change of variables we see

$$\begin{aligned}
\int e^{-\mathbf{x}'\mathbf{A}\mathbf{x}/2}\,d\mathbf{x} &= \int e^{-\mathbf{y}'\mathbf{D}\mathbf{y}/2}\,d\mathbf{y} \\
&= \iint \cdots \int \exp\left(-\frac{1}{2}\sum_{i=1}^{n} d_{ii} y_i^2\right)\,dy_1\,dy_2 \cdots dy_n \\
&= \prod_{i=1}^{n}\left(\int_{-\infty}^{\infty} e^{-d_{ii}y_i^2/2}\,dy_i\right)
\end{aligned}$$

It is obvious that all the integrals in the last line exist if and only if each d_{ii} is strictly positive, which happens if and only if $\mathbf{A}$ is positive definite. That proves (a).

Now we just "complete the square." We want to put $q(\mathbf{x})$ in the same form as the quadratic form

$$\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\mathbf{M}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \tag{5.26}$$

in the exponent of the usual expression for the normal distribution. Expand (5.26)

$$\begin{aligned}
\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\mathbf{M}^{-1}(\mathbf{x} - \boldsymbol{\mu}) &= \tfrac{1}{2}\mathbf{x}'\mathbf{M}^{-1}\mathbf{x} - \tfrac{1}{2}\mathbf{x}'\mathbf{M}^{-1}\boldsymbol{\mu} - \tfrac{1}{2}\boldsymbol{\mu}'\mathbf{M}^{-1}\mathbf{x} + \tfrac{1}{2}\boldsymbol{\mu}'\mathbf{M}^{-1}\boldsymbol{\mu} \\
&= \tfrac{1}{2}\mathbf{x}'\mathbf{M}^{-1}\mathbf{x} - \boldsymbol{\mu}'\mathbf{M}^{-1}\mathbf{x} + \tfrac{1}{2}\boldsymbol{\mu}'\mathbf{M}^{-1}\boldsymbol{\mu}
\end{aligned}$$

(the second equality holding because of the rule for the transpose of a product). Now the only way $q(\mathbf{x})$ can match up with this is if the constants in the quadratic and linear terms both match, that is,

$$\mathbf{A} = \mathbf{M}^{-1}$$

and

$$\mathbf{b}' = -\boldsymbol{\mu}'\mathbf{M}^{-1},$$

and these in turn imply

$$\boldsymbol{\mu} = -\mathbf{A}^{-1}\mathbf{b} \tag{5.27}$$

$$\mathbf{M} = \mathbf{A}^{-1} \tag{5.28}$$

which in turn are (c) and (d) if (b) is true. So all that remains is to prove (b).

We have now shown that the quadratic and linear terms of $q(\mathbf{x})$ and (5.26) match when we define $\boldsymbol{\mu}$ and $\mathbf{M}$ by (5.27) and (5.28). Hence

$$q(\mathbf{x}) = \tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\mathbf{M}^{-1}(\mathbf{x} - \boldsymbol{\mu}) + c - \tfrac{1}{2}\boldsymbol{\mu}'\mathbf{M}^{-1}\boldsymbol{\mu}$$

and

$$f(\mathbf{x}) = \exp\left(-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\mathbf{M}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)\exp\left(c - \tfrac{1}{2}\boldsymbol{\mu}'\mathbf{M}^{-1}\boldsymbol{\mu}\right)$$

Since the first term on the right hand side is an unnormalized density of the $\mathcal{N}(\boldsymbol{\mu}, \mathbf{M})$ distribution, the second term must be the reciprocal of the normalizing constant so that $f(\mathbf{x})$ integrates to one. That proves (b), and we are done. □

I call this the "e to a quadratic" theorem. If the density is the exponential of a quadratic form, then the distribution must be non-degenerate multivariate normal, and the mean and variance can be read off the density.

5.2.2 Marginals

Lemma 5.11. *Every linear transformation of a multivariate normal random vector is (multivariate or univariate) normal.*

This obvious because a linear transformation of a linear transformation is linear. If $\mathbf{X}$ is multivariate normal, then, by definition, it has the form $\mathbf{X} = \mathbf{a} + \mathbf{B}\mathbf{Z}$, where $\mathbf{Z}$ is standard normal, $\mathbf{a}$ is a constant vector, and $\mathbf{B}$ is a constant matrix. So if $\mathbf{Y} = \mathbf{c} + \mathbf{D}\mathbf{X}$, where $\mathbf{c}$ is a constant vector and $\mathbf{D}$ is a constant matrix, then

$$\begin{aligned} \mathbf{Y} &= \mathbf{c} + \mathbf{D}\mathbf{X} \\ &= \mathbf{c} + \mathbf{D}(\mathbf{a} + \mathbf{B}\mathbf{Z}) \\ &= (\mathbf{c} + \mathbf{D}\mathbf{a}) + (\mathbf{D}\mathbf{B})\mathbf{Z}, \end{aligned}$$

which is clearly a linear transformation of $\mathbf{Z}$, hence normal.

Corollary 5.12. *Every marginal distribution of a multivariate normal distribution is (multivariate or univariate) normal.*

This is an obvious consequence of the lemma, because the operation of finding a marginal defines a linear transformation, simply because of the definitions of vector addition and scalar multiplication, that is, because the i-th component of $a\mathbf{X} + b\mathbf{Y}$ is $aX_i + bY_i$.

5.2.3 Partitioned Matrices

This section has no probability theory, just an odd bit of matrix algebra. The notation

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix} \tag{5.29}$$

indicates a *partitioned matrix*. Here each of the $\mathbf{B}_{ij}$ is itself a matrix. $\mathbf{B}$ is just the matrix having the elements of $\mathbf{B}_{11}$ in its upper left corner, with the elements of $\mathbf{B}_{12}$ to their right, and so forth. Of course the dimensions of the $\mathbf{B}_{ij}$ must fit together the right way.

One thing about partitioned matrices that makes them very useful is that matrix multiplication looks "just like" matrix multiplication of non-partitioned matrices. You just treat the matrices like scalar elements of an ordinary array

$$\begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{B}_{11}\mathbf{C}_{11} + \mathbf{B}_{12}\mathbf{C}_{21} & \mathbf{B}_{11}\mathbf{C}_{12} + \mathbf{B}_{12}\mathbf{C}_{22} \\ \mathbf{B}_{21}\mathbf{C}_{11} + \mathbf{B}_{22}\mathbf{C}_{21} & \mathbf{B}_{21}\mathbf{C}_{12} + \mathbf{B}_{22}\mathbf{C}_{22} \end{pmatrix}$$

If one of the matrixes is a partitioned column vector, it looks like the multiplication of a vector by a matrix

$$\begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{B}_{11}\mathbf{x}_1 + \mathbf{B}_{12}\mathbf{x}_2 \\ \mathbf{B}_{21}\mathbf{x}_1 + \mathbf{B}_{22}\mathbf{x}_2 \end{pmatrix}$$

and similarly for

$$\begin{aligned} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}' \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} &= \begin{pmatrix} \mathbf{x}_1' & \mathbf{x}_2' \end{pmatrix} \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{x}_1' & \mathbf{x}_2' \end{pmatrix} \begin{pmatrix} \mathbf{B}_{11}\mathbf{x}_1 + \mathbf{B}_{12}\mathbf{x}_2 \\ \mathbf{B}_{21}\mathbf{x}_1 + \mathbf{B}_{22}\mathbf{x}_2 \end{pmatrix} \\ &= \mathbf{x}_1'\mathbf{B}_{11}\mathbf{x}_1 + \mathbf{x}_1'\mathbf{B}_{12}\mathbf{x}_2 + \mathbf{x}_2'\mathbf{B}_{21}\mathbf{x}_1 + \mathbf{x}_2'\mathbf{B}_{22}\mathbf{x}_2 \end{aligned}$$

Of course, in all of these, the dimensions have be such that the matrix multiplications make sense.

If $\mathbf{X}$ is a partitioned random vector

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}, \tag{5.30a}$$

then its mean mean vector is

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \tag{5.30b}$$

where

$$\boldsymbol{\mu}_i = E(\mathbf{X}_i),$$

and its variance matrix is

$$\mathbf{M} = \begin{pmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{pmatrix}, \tag{5.30c}$$

where

$$\mathbf{M}_{ij} = \operatorname{cov}(\mathbf{X}_i, \mathbf{X}_j).$$

Again, every thing looks very analogous to the situation with scalar rather than vector or matrix components.

A partitioned matrix is called *block diagonal* if the "off-diagonal" matrices are all zero. The partitioned matrix (5.29) is block diagonal if $\mathbf{B}_{12} = 0$ and $\mathbf{B}_{21} = 0$. The partitioned matrix (5.30c) is block diagonal if $\mathbf{X}_1$ and $\mathbf{X}_2$ are uncorrelated, that is, $\operatorname{cov}(\mathbf{X}_1, \mathbf{X}_2) = 0$.

A block diagonal matrix with square blocks on the diagonal, is easy to invert, just invert each block. For example, if (5.30c) is block diagonal, then

$$\mathbf{M}^{-1} = \begin{pmatrix} \mathbf{M}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{M}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{22}^{-1} \end{pmatrix} \tag{5.31}$$

5.2.4 Conditionals and Independence

In this section we consider a normal random vector $\mathbf{X}$ partitioned as in (5.30a) with variance matrix $\mathbf{M}$, which must be partitioned as in (5.30c). We will need a notation for the inverse variance matrix: we adopt $\mathbf{W} = \mathbf{M}^{-1}$. Of course, it can be partitioned in the same way

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{21} & \mathbf{W}_{22} \end{pmatrix} \tag{5.32}$$

Note from (5.31) that if $\mathbf{M}$ is block diagonal and invertible, then so is $\mathbf{W}$ and $\mathbf{W}_{ii} = \mathbf{M}_{ii}^{-1}$. When $\mathbf{M}$ is not block diagonal, then neither is $\mathbf{W}$ and the relation between the two is complicated.

Theorem 5.13. *Random vectors that are jointly multivariate normal and uncorrelated are independent.*

In notation, what the theorem says is that if $\mathbf{X}$ is multivariate normal and partitioned as in (5.30a) with variance matrix (5.30c), then

$$\mathbf{M}_{12} = \operatorname{cov}(\mathbf{X}_1, \mathbf{X}_2) = 0$$

implies that $\mathbf{X}_1$ and $\mathbf{X}_2$ are actually independent random vectors.

Please note the contrast with the general case.

> *In general independent implies uncorrelated, but uncorrelated does* ***not*** *imply independent.*
>
> *Only when the random variables are* ***jointly multivariate normal*** *does uncorrelated imply independent.*

Proof. Without loss of generality, we may assume the means are zero, because $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent if and only if $\mathbf{X}_1 - \boldsymbol{\mu}_1$ and $\mathbf{X}_2 - \boldsymbol{\mu}_2$ are independent.

We first prove the special case in which $\mathbf{X}$ has a non-degenerate distribution. Then the unnormalized density (ignoring constants) is

$$\exp\left(-\tfrac{1}{2}\mathbf{x}'\mathbf{W}\mathbf{x}\right) = \exp\left(-\tfrac{1}{2}\mathbf{x}_1'\mathbf{W}_{11}\mathbf{x}_1\right)\exp\left(-\tfrac{1}{2}\mathbf{x}_2'\mathbf{W}_{22}\mathbf{x}_2\right)$$

In general, there is also a $\mathbf{x}_1'\mathbf{W}_{12}\mathbf{x}_2$ term in the exponent, but it vanishes here because $\mathbf{W}$ is block diagonal because of (5.31). Since the density factors, the random vectors are independent.

We now prove the general case by expressing some variables in terms of the others. If $\mathbf{X}$ is concentrated on a hyperplane, then we can express one variable as a linear combination of the remaining $n - 1$ variables. If these are still concentrated on a hyperplane, then we can express another variable as a linear combination of the remaining $n - 2$ and so forth. We stop when we have expressed some variables as linear combinations of a set of k variables which have a non-degenerate multivariate normal distribution. We can now partition $\mathbf{X}$ as

$$\mathbf{X} = \begin{pmatrix} \mathbf{U}_1 \\ \mathbf{V}_1 \\ \mathbf{U}_2 \\ \mathbf{V}_2 \end{pmatrix}$$

where $(\mathbf{U}_1, \mathbf{U}_2)$ has a non-degenerate multivariate normal distribution and

$$\begin{aligned} \mathbf{V}_1 &= \mathbf{B}_{11}\mathbf{U}_1 + \mathbf{B}_{12}\mathbf{U}_2 \\ \mathbf{V}_2 &= \mathbf{B}_{21}\mathbf{U}_1 + \mathbf{B}_{22}\mathbf{U}_2 \end{aligned}$$

for some matrix $\mathbf{B}$ partitioned as in (5.29), and $\mathbf{X}_i = (\mathbf{U}_i, \mathbf{V}_i)$. Note that the assumption that $\mathbf{X}_1$ and $\mathbf{X}_2$ are uncorrelated implies that $\mathbf{U}_1$ and $\mathbf{U}_2$ are also uncorrelated and hence, by what has already been proved independent (since they are jointly non-degenerate multivariate normal).

Then, using the additional notation

$$\begin{aligned} \operatorname{var}(\mathbf{U}_1) &= \mathbf{S}_{11} \\ \operatorname{var}(\mathbf{U}_2) &= \mathbf{S}_{22} \end{aligned}$$

we calculate that $\operatorname{var}(\mathbf{X})$ is

$$\begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{11}\mathbf{B}'_{11} & \mathbf{0} & \mathbf{S}_{11}\mathbf{B}'_{21} \\ \mathbf{B}_{11}\mathbf{S}_{11} & \mathbf{B}_{11}\mathbf{S}_{11}\mathbf{B}'_{11} + \mathbf{B}_{12}\mathbf{S}_{22}\mathbf{B}'_{12} & \mathbf{B}_{12}\mathbf{S}_{22} & \mathbf{B}_{11}\mathbf{S}_{11}\mathbf{B}'_{21} + \mathbf{B}_{12}\mathbf{S}_{22}\mathbf{B}'_{22} \\ \mathbf{0} & \mathbf{S}_{22}\mathbf{B}'_{12} & \mathbf{S}_{22} & \mathbf{S}_{22}\mathbf{B}'_{22} \\ \mathbf{B}_{21}\mathbf{S}_{11} & \mathbf{B}_{21}\mathbf{S}_{11}\mathbf{B}'_{11} + \mathbf{B}_{22}\mathbf{S}_{22}\mathbf{B}'_{12} & \mathbf{B}_{22}\mathbf{S}_{22} & \mathbf{B}_{21}\mathbf{S}_{11}\mathbf{B}'_{21} + \mathbf{B}_{22}\mathbf{S}_{22}\mathbf{B}'_{22} \end{pmatrix}$$

Now the assumption of the theorem is that this matrix is block diagonal, with the blocks now 2×2. Since $\mathbf{U}_1$ and $\mathbf{U}_2$ are nondegenerate, their variance matrices are invertible, thus the only way we can have $\mathbf{B}_{21}\mathbf{S}_{11} = 0$ and $\mathbf{B}_{12}\mathbf{S}_{22} = 0$ is if $\mathbf{B}_{21} = 0$ and $\mathbf{B}_{12} = 0$. But this implies

$$\mathbf{X}_i = \begin{pmatrix} \mathbf{U}_i \\ \mathbf{B}_{ii}\mathbf{U}_i \end{pmatrix}$$

for $i = 1$, 2, and since these are functions of the independent random vectors $\mathbf{U}_1$ and $\mathbf{U}_2$, they are independent. □

Every conditional of a normal random vector is normal too, but it is hard for us to give an explicit expression for the degenerate case. This is not surprising, because all our methods for finding conditional distributions involve densities and degenerate normal distributions don't have densities.

First a lemma.

Lemma 5.14. *Suppose* $\mathbf{X}$ *is partitioned as in* (5.30a) *and has variance matrix* (5.30c)*, and suppose that* $\mathbf{M}_{22}$ *is positive definite. Then*

$$\mathbf{X}_1 - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{X}_2 \quad \textit{and} \quad \mathbf{X}_2$$

are uncorrelated.

And, we should note, by Theorem 5.13, if $\mathbf{X}$ is multivariate normal, then $\mathbf{X}_1 - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{X}_2$ is independent of $\mathbf{X}_2$.

Proof. Obvious, just calculate the covariance

$$\begin{aligned} \operatorname{cov}(\mathbf{X}_1 - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{X}_2, \mathbf{X}_2) &= \operatorname{cov}(\mathbf{X}_1, \mathbf{X}_2) - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\operatorname{cov}(\mathbf{X}_2, \mathbf{X}_2) \\ &= \mathbf{M}_{12} - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{22} \\ &= 0 \end{aligned}$$

□

Every conditional of a normal random vector is also normal, but it is hard for us to give an explicit expression for the degenerate case. This is not surprising, because all our methods for finding conditional densities and degenerate normal distributions don't have densities. So here we will be satisfied with describing the non-degenerate case.

Theorem 5.15. *Every condition distribution of a non-degenerate multivariate normal distribution is non-degenerate (multivariate or univariate) normal.*

In particular, if $\mathbf{X}$ *is partitioned as in* (5.30a)*, has the multivariate normal distribution with mean vector* (5.30b) *and variance matrix* (5.30c)*, then*

$$\mathbf{X}_1 \mid \mathbf{X}_2 \sim \mathcal{N}(\boldsymbol{\mu}_1 + \mathbf{M}_{12}\mathbf{M}_{22}^{-1}[\mathbf{X}_2 - \boldsymbol{\mu}_2], \mathbf{M}_{11} - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{21}). \tag{5.33}$$

Proof. First note that the conditional distribution is multivariate normal by Lemma 5.10, because the joint density is the exponential of a quadratic, hence so is the conditional, which is just the joint density considered as a function of x_1 with x_2 fixed renormalized.

So all that remains to be done is figuring out the conditional mean and variance. For the conditional mean, we use Lemma 5.14 and the comment following it. Because of the independence of $\mathbf{X}_1 - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{X}_2$ and $\mathbf{X}_2$,

$$E(\mathbf{X}_1 - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{X}_2 \mid \mathbf{X}_2) = E(\mathbf{X}_1 - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{X}_2)$$

but

$$E(\mathbf{X}_1 - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{X}_2 \mid \mathbf{X}_2) = E(\mathbf{X}_1 \mid \mathbf{X}_2) - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{X}_2$$

by linearity of expectations and functions of the conditioning variable behaving like constants, and

$$E(\mathbf{X}_1 - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{X}_2) = \boldsymbol{\mu}_1 - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\boldsymbol{\mu}_2.$$

Thus

$$E(\mathbf{X}_1 \mid \mathbf{X}_2) - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{X}_2 = \boldsymbol{\mu}_1 - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\boldsymbol{\mu}_2,$$

which establishes the conditional expectation given in (5.33).

To calculate the variance, we first observe that

$$\operatorname{var}(\mathbf{X}_1 \mid \mathbf{X}_2) = W_{11}^{-1} \tag{5.34}$$

where $\mathbf{W} = \mathbf{M}^{-1}$ is partitioned as in (5.32), because the quadratic form in the exponent of the density has quadratic term $\mathbf{x}_1 W_{11} \mathbf{x}_1$ and Theorem 5.10 says that is the inverse variance matrix of the vector in question, which in this case is x_1 given x_2. We don't know what the form of W_{11} or it's inverse it, but we do know it is a constant matrix, which is all we need. The rest of the job can be done by the vector version of the iterated variance formula (Theorem 3.7)

$$\operatorname{var}(\mathbf{X}_1) = \operatorname{var}\{E(\mathbf{X}_1 \mid \mathbf{X}_2)\} + E\{\operatorname{var}(\mathbf{X}_1 \mid \mathbf{X}_2)\} \tag{5.35}$$

(which we haven't actually proved but is proved in exactly the same way as the scalar formula). We know

$$\operatorname{var}(\mathbf{X}_1) = \mathbf{M}_{11}$$

but

$$\begin{aligned}
\operatorname{var}\{E(\mathbf{X}_1 \mid \mathbf{X}_2)\} &+ E\{\operatorname{var}(\mathbf{X}_1 \mid \mathbf{X}_2)\} \\
&= \operatorname{var}\{\boldsymbol{\mu}_1 + \mathbf{M}_{12}\mathbf{M}_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2)\} + E\{\mathbf{W}_{11}^{-1}\} \\
&= \operatorname{var}(\mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{X}_2) + \mathbf{W}_{11}^{-1} \\
&= \mathbf{M}_{12}\mathbf{M}_{22}^{-1} \operatorname{var}(\mathbf{X}_2)\mathbf{M}_{22}^{-1}\mathbf{M}_{12}' + \mathbf{W}_{11}^{-1} \\
&= \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{22}\mathbf{M}_{22}^{-1}\mathbf{M}_{21} + \mathbf{W}_{11}^{-1} \\
&= \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{21} + \mathbf{W}_{11}^{-1}
\end{aligned}$$

Equating the two gives

$$\mathbf{M}_{11} = \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{21} + \mathbf{W}_{11}^{-1}$$

which along with (5.34) establishes the conditional variance given in (5.33). □

5.3 Bernoulli Random Vectors

To start we generalize the notion of a Bernoulli random variables. One might think that should be a vector with i. i. d. Bernoulli components, but something quite different is in order. A (univariate) Bernoulli random variable is really an indicator function. *All* zero-or-one valued random variables are indicator functions: they indicate the set on which they are one. How do we generalize the notion of an indicator function to the multivariate case? We consider a vector of indicator functions.

We give three closely related definitions.

Definition 5.3.1 (Bernoulli Random Vector).
A random vector $\mathbf{X} = (X_1, \ldots, X_k)$ is ***Bernoulli*** *if the X_i are the indicators of a partition of the sample space, that is,*

$$X_i = I_{A_i}$$

where

$$A_i \cap A_j = \varnothing, \qquad i \neq j$$

and

$$\bigcup_{i=1}^{k} A_i$$

is the whole sample space.

Definition 5.3.2 (Bernoulli Random Vector).
A random vector $\mathbf{X} = (X_1, \ldots, X_k)$ is ***Bernoulli*** *if the X_i are zero-or-one-valued random variables and*

$$X_1 + \cdots + X_k = 1.$$

with probability one.

Definition 5.3.3 (Bernoulli Random Vector).
A random vector $\mathbf{X} = (X_1, \ldots, X_k)$ is ***Bernoulli*** *if the X_i are zero-or-one-valued random variables and with probability one exactly one of X_1, $\ldots$, X_k is one and the rest are zero.*

The equivalence of Definitions 5.3.2 and 5.3.3 is obvious. The only way a bunch of zeros and ones can add to one is if there is exactly one one.

The equivalence of Definitions 5.3.1 and 5.3.3 is also obvious. If the A_i form a partition, then exactly one of the

$$X_i(\omega) = I_{A_i}(\omega)$$

is equal to one for any outcome ω, the one for which $\omega \in A_i$. There is, of course, exactly one i such that $\omega \in A_i$ just by definition of "partition."

5.3.1 Categorical Random Variables

Bernoulli random vectors are closely related to *categorical random variables* taking values in an arbitrary finite set. You may have gotten the impression up to know that probability theorists have a heavy preference for numerical random variables. That's so. Our only "brand name" distribution that is not necessarily numerical valued is the discrete uniform distribution. In principle, though a random variable can take values in *any* set. So although we haven't done much with such variables so far, we haven't ruled them out either. Of course, if Y is a random variable taking values in the set

$$S = \{\text{strongly agree}, \text{agree}, \text{neutral}, \text{disagree}, \text{strongly disagree}\} \tag{5.36}$$

you can't talk about expectations or moments, $E(Y)$ is defined only for numerical (or numerical vector) random variables, not for categorical random variables.

However, if we number the categories

$$S = \{s_1, s_2, \ldots, s_5\}$$

with $s_1 =$ strongly agree, and so forth, then we can identify the categorical random variable Y with a Bernoulli random vector $\mathbf{X}$

$$X_i = I_{\{s_i\}}(Y)$$

that is

$$X_i = 1 \quad \text{if and only if} \quad Y = s_i.$$

Thus Bernoulli random variables are an artifice. They are introduced to inject some numbers into categorical problems. We can't talk about $E(Y)$, but we can talk about $E(\mathbf{X})$. A thorough analysis of the properties of the distribution of the random vector $\mathbf{X}$ will also tell us everything we want to know about the categorical random variable Y, and it will do so allowing us to use the tools (moments, etc.) that we already know.

5.3.2 Moments

Each of the X_i is, of course, univariate Bernoulli, write

$$X_i \sim \text{Ber}(p_i)$$

and collect these parameters into a vector

$$\mathbf{p} = (\mathbf{p}_1, \ldots, \mathbf{p}_k)$$

Then we abbreviate the distribution of $\mathbf{X}$ as

$$\mathbf{X} \sim \text{Ber}_k(\mathbf{p})$$

if we want to indicate the dimension k or just as $\mathbf{X} \sim \text{Ber}(\mathbf{p})$ if the dimension is clear from the context (the boldface type indicating a vector parameter makes it clear this is not the univariate Bernoulli).

Since each X_i is univariate Bernoulli,

$$\begin{aligned} E(X_i) &= p_i \\ \text{var}(X_i) &= p_i(1 - p_i) \end{aligned}$$

That tells us

$$E(\mathbf{X}) = \mathbf{p}.$$

To find the variance matrix we need to calculate covariances. For $i \neq j$,

$$\text{cov}(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j) = -p_i p_j,$$

because $X_i X_j = 0$ with probability one.

Hence $\operatorname{var}(\mathbf{X}) = \mathbf{M}$ has components

$$m_{ij} = \begin{cases} p_i(1 - p_i), & i = j \\ -p_i p_j & i \neq j \end{cases} \tag{5.37}$$

We can also write this using more matrixy notation by introducing the diagonal matrix $\mathbf{P}$ having diagonal elements p_i and noting that the "outer product" $\mathbf{p}\mathbf{p}'$ has elements $p_i p_j$, hence

$$\operatorname{var}(\mathbf{X}) = \mathbf{P} - \mathbf{p}\mathbf{p}'$$

Question: Is $\operatorname{var}(\mathbf{X})$ positive definite? This is of course related to the question of whether $\mathbf{X}$ is degenerate. We haven't said anything explicit about either, but the information needed to answer these questions is in the text above. It should be obvious if you know what to look for (a good exercise testing your understanding of degenerate random vectors).

5.4 The Multinomial Distribution

The multinomial distribution is the multivariate analog of the binomial distribution. It is sort of, but not quite, the multivariate generalization, that is, the binomial distribution is sort of, but not precisely, a special case of the multinomial distribution. Thus is unlike the normal, where the univariate normal distribution is precisely the one-dimensional case of the multivariate normal.

Suppose $\mathbf{X}_1$, $\mathbf{X}_2$ are an i. i. d. sequence of $\text{Ber}_k(\mathbf{p})$ random vectors (caution: the subscripts on the $\mathbf{X}_i$ indicate elements of an infinite sequence of i. i. d. random vectors, not components of one vector). Then

$$\mathbf{Y} = \mathbf{X}_1 + \cdots + \mathbf{X}_n$$

has the *multinomial distribution* with *sample size* n and dimension k, abbreviated

$$\mathbf{Y} \sim \text{Multi}_k(n, \mathbf{p})$$

if we want to indicate the dimension in the notation or just $\mathbf{Y} \sim \text{Multi}(n, \mathbf{p})$ if the dimension is clear from the context.

Note the dimension is k, not n, that is, both $\mathbf{Y}$ and $\mathbf{p}$ are vectors of dimension k.

5.4.1 Categorical Random Variables

Recall that a multinomial random vector is the sum of i. i. d. Bernoullis

$$\mathbf{Y} = \mathbf{X}_1 + \cdot + \mathbf{X}_n$$

and that each Bernoulli is related to a categorical random variable: $X_{i,j} = 1$ if and only if the i-th observation fell in the j-th category. Thus $Y_j = \sum_i X_{i,j}$ is the number of individuals that fell in the j-th category.

This gives us another distribution of multinomial random vectors. A random vector $\mathbf{Y} = \text{Multi}(n, p)$ arises by observing a sequence of n independent random variables (taking values in any set) and letting Y_j be the number of observations that fall in the j-th category. The parameter p_j is the probability of any one individual observation falling in the j-th category.

5.4.2 Moments

Obvious, just n times the moments of $\text{Ber}(\mathbf{p})$

$$\begin{aligned} E(\mathbf{X}) &= n\mathbf{p} \\ \text{var}(\mathbf{X}) &= n(\mathbf{P} - \mathbf{p}\mathbf{p}') \end{aligned}$$

5.4.3 Degeneracy

Since the components of a $\text{Ber}(\mathbf{p})$ random vector sum to one, the components of a $\text{Multi}(n, \mathbf{p})$ random vector sum to n. That is, if $\mathbf{Y} \sim \text{Multi}(n, \mathbf{p})$, then

$$Y_1 + \cdots Y_k = n$$

with probability one. This can be written $\mathbf{u}'\mathbf{Y} = n$ with probability one, where $\mathbf{u} = (1, 1, \ldots, 1)$. Thus $\mathbf{Y}$ is concentrated on the hyperplane

$$H = \{\, \mathbf{y} \in \mathbb{R}^k : \mathbf{u}'\mathbf{y} = n \,\}$$

Is $\mathbf{Y}$ concentrated on any other hyperplanes? Since the $\text{Ber}_k(\mathbf{p})$ distribution and the $\text{Multi}_k(n, \mathbf{p})$ distribution have the same variance matrices except for a constant of proportionality ($\mathbf{M}$ and $n\mathbf{M}$, respectively), they both are supported on the same hyperplanes. We might as well drop the n and ask the question about the Bernoulli.

Let $\mathbf{c} = (c_1, \ldots, c_k)$ be an arbitrary vector. Such a vector is associated with a hyperplane supporting the distribution if

$$\begin{aligned} \mathbf{c}'\mathbf{M}\mathbf{c} &= \sum_{i=1}^{k} \sum_{j=1}^{k} m_{ij} c_i c_j \\ &= \sum_{i=1}^{k} p_i c_i^2 - \sum_{i=1}^{k} \sum_{j=1}^{k} p_i p_j c_i c_j \\ &= \sum_{i=1}^{k} p_i c_i^2 - \left(\sum_{j=1}^{k} p_j c_j \right)^2 \end{aligned}$$

is zero. Thinking of this as a function of $\mathbf{c}$ for fixed $\mathbf{p}$, write it as $q(\mathbf{c})$. Being a variance, it is nonnegative, hence it is zero only where it is achieving its minimum value, and where, since it is a smooth function, its derivative must be zero, that is,

$$\frac{\partial q(\mathbf{c})}{\partial c_i} = 2p_i c_i - 2p_i \left(\sum_{j=1}^k p_j c_j \right) = 0$$

Now we do not know what the quantity in parentheses is, but it does not depend on i or j, so we can write it as a single letter d with no subscripts. Thus we have to solve

$$2p_i c_i - 2dp_i = 0 \tag{5.38}$$

for c_i. This splits into two cases.

Case I. If none of the p_i are zero, the only solution is $c_i = d$. Thus the only null eigenvectors are proportional to the vector $\mathbf{u} = (1, 1, \ldots, 1)$. And all such vectors determine the same hyperplane.

Case II. If any of the p_i are zero, we get more solutions. Equation (5.38) becomes $0 = 0$ when $p_i = 0$, and since this is the only equation containing c_i, the equations say nothing about c_i, thus the solution is

$$\begin{aligned} c_i &= d, & p_i &> 0 \\ c_i &= \text{arbitrary}, & p_i &= 0 \end{aligned}$$

In hindsight, case II was rather obvious too. If $p_i = 0$ then $X_i = 0$ with probability one, and that is another degeneracy. But our real interest is in case I. If none of the success probabilities are zero, then the only degeneracy is $Y_1 + \cdots + Y_k = n$ with probability one.

5.4.4 Density

Density? Don't degenerate distribution have no densities? In the continuous case, yes. Degenerate continuous random vectors have no densities. But discrete random vectors always have densities, as always, $f(\mathbf{x}) = P(\mathbf{X} = \mathbf{x})$.

The derivation of the density is exactly like the derivation of the binomial density. First we look at one particular outcome, then collect the outcomes that lead to the same $\mathbf{Y}$ values. Write $X_{i,j}$ for the components of $\mathbf{X}_i$, and note that if we know $X_{i,m} = 1$, then we also know $X_{i,j} = 0$ for $j \neq m$, so it is enough to determine the probability of an outcome if we simply record the X_{ij} that are equal to one. Then by the multiplication rule

$$\begin{aligned} P(X_{1,j_1} = 1 \text{ and} \cdots \text{and } X_{n,j_n} = 1) &= \prod_{i=1}^n P(X_{i,j_i} = 1) \\ &= \prod_{i=1}^n p_{j_i} \\ &= \prod_{j=1}^k p_j^{y_j} \end{aligned}$$

The last equality records the same kind of simplification we saw in deriving the binomial density. The product from 1 to n in the next to last line may repeat some of the p's. How often are they repeated? There is one p_j for each X_{ij} that is equal to one, and there are $Y_j = \sum_i X_{ij}$ of them.

We are not done, however, because more than one outcome can lead to the same right hand side here. How many ways are there to get exactly y_j of the X_{ij} equal to one? This is the same as asking how many ways there are to assign the numbers $i = 1, \ldots, n$ to one of k categories, so that there are y_i in the i-th category, and the answer is the multinomial coefficient

$$\binom{n}{y_1, \ldots, y_k} = \frac{n!}{y_1! \cdots y_k!}$$

Thus the density is

$$f(\mathbf{y}) = \binom{n}{y_1, \ldots, y_k} \prod_{j=1}^{k} p_j^{y_j}, \qquad \mathbf{y} \in S$$

where the sample space S is defined by

$$S = \{\, \mathbf{y} \in \mathbb{N}^k : y_1 + \cdots y_k = n \,\}$$

where $\mathbb{N}$ denotes the "natural numbers" 0, 1, 2, ... In other words, the sample space S consists of vectors $\mathbf{y}$ having nonnegative integer coordinates that sum to n.

5.4.5 Marginals and "Sort Of" Marginals

The univariate marginals are obvious. Since the univariate marginals of $\text{Ber}(\mathbf{p})$ are $\text{Ber}(p_i)$, the univariate marginals of $\text{Multi}(n, \mathbf{p})$ are $\text{Bin}(n, p_i)$.

Strictly speaking, the multivariate marginals do not have a brand name distribution. Lindgren (Theorem 8 of Chapter 6) says the marginals of a multinomial are multinomial, but this is, strictly speaking, complete rubbish, given the way he (and we) defined "marginal" and "multinomial." It is obviously wrong. If $\mathbf{X} = (X_1, \ldots, X_k)$ is multinomial, then it is degenerate. But $(X_1, \ldots, X_{k-1})$ is not degenerate, hence not multinomial (all multinomial distributions are degenerate). The same goes for any other subvector, (X_2, X_5, X_{10}), for example.

Of course, Lindgren knows this as well as I do. He is just being sloppy about terminology. What he means is clear from his discussion leading up to the "theorem" (really a non-theorem). Here's the correct statement.

Theorem 5.16. *Suppose $\mathbf{Y} = \text{Multi}_k(n, \mathbf{p})$ and $\mathbf{Z}$ is a random vector formed by collapsing some of the categories for $\mathbf{Y}$, that is, each component of $\mathbf{Z}$ has the form*

$$Z_j = Y_{i_1} + \cdots + Y_{i_{m_j}}$$

where each Y_i contributes to exactly one Z_j so that

$$Z_1 + \cdots + Z_l = Y_1 + \cdots + Y_k = n,$$

then

$$\mathbf{Z} \sim \text{Multi}_l(n, \mathbf{q})$$

where the parameter vector $\mathbf{q}$ *has components*

$$q_j = p_{i_1} + \cdots + p_{i_{m_j}}$$

is formed by collapsing the categories in the same way as in forming $\mathbf{Z}$ *from* $\mathbf{Y}$*.*

No wonder Lindgren felt the urge to sloppiness here. The correct statement is a really obnoxious mess of notation. But the idea is simple and obvious. If we collapse some categories, that gives a different (coarser) partition of the sample space and a multinomial distribution with fewer categories.

Example 5.4.1.
Consider the multinomial random vector $\mathbf{Y}$ associated with i. i. d. sampling of a categorical random variable taking values in the set (5.36). Let $\mathbf{Z}$ be the multinomial random vector associated with the categorical random variable obtained by collapsing the categories on the ends, that is, we collapse the categories "strongly agree" and "agree" and we collapse the categories "strongly disagree" and "disagree." Thus

$$\begin{aligned} \mathbf{Y} &\sim \text{Multi}_5(n, \mathbf{p}) \\ \mathbf{Z} &\sim \text{Multi}_3(n, \mathbf{q}) \end{aligned}$$

where

$$\begin{aligned} Z_1 &= Y_1 + Y_2 \\ Z_2 &= Y_3 \\ Z_3 &= Y_4 + Y_5 \end{aligned}$$

and

$$\begin{aligned} q_1 &= p_1 + p_2 \\ q_2 &= p_3 \\ q_3 &= p_4 + p_5 \end{aligned}$$

The notation is simpler than in the theorem, but still messy, obscuring the simple idea of collapsing categories. Maybe Lindgren has the right idea. Slop is good here. The marginals of a multinomial are sort of, but not precisely, multinomial. Or should that be the sort-of-but-not-precisely marginals of a multinomial are multinomial?

Recall that we started this section with the observation that one-dimensional marginal distributions of a multinomial are binomial (with no "sort of"). But two-dimensional multinomial distributions must also be somehow related to the binomial distribution. The $k = 2$ multinomial coefficients *are* binomial coefficients, that is,

$$\binom{n}{y_1, y_2} = \frac{n!}{y_1! y_2!} = \binom{n}{y_1} = \binom{n}{y_2}$$

because the multinomial coefficient is only defined when the numbers in the second row add up to number in the first row, that is, here $y_1 + y_2 = n$.

And the relation between distributions is obvious too, just because the marginals are binomial. If

$$\mathbf{Y} = \text{Multi}_2(n, \mathbf{p}),$$

then

$$Y_i = \text{Bin}(n, p_i)$$

and

$$Y_2 = n - Y_1.$$

Conversely, if

$$X \sim \text{Bin}(n, p),$$

then

$$(X, n - X) \sim \text{Multi}_2\bigl(n, (p, 1 - p)\bigr)$$

So the two-dimensional multinomial is the distribution of $(X, n - X)$ when X is binomial. Recall the conventional terminology that X is the number of "successes" in n Bernoulli "trials" and $n - X$ is the number of "failures." Either of the successes or the failures considered by themselves are binomial. When we paste them together in a two-dimensional vector, the vector is degenerate because the successes and failures sum to the number of trials, and that degenerate random vector is the two-dimensional multinomial.

5.4.6 Conditionals

Theorem 5.17. *Every conditional of a multinomial is multinomial. Suppose* $\mathbf{Y} \sim \text{Multi}_k(n, \mathbf{p})$*, then*

$$(Y_1, \ldots, Y_j) \mid (Y_{j+1}, \ldots, Y_k) \sim \text{Multi}_j(n - Y_{j+1} - \cdots - Y_k, \mathbf{q}), \tag{5.39a}$$

where

$$q_i = \frac{p_i}{p_1 + \cdots + p_j}, \qquad i = 1, \ldots, j. \tag{5.39b}$$

In words, the variables that are still random (the ones "in front of the bar") are multinomial. The number of categories is the number (here j) of such variables. The sample size is the number of observations still random, which is the original sample size minus the observations in the variables now known (the ones "behind the bar"). And the parameter vector $\mathbf{q}$ is the part of the original parameter vector corresponding to the variables in front of the bar renormalized.

Renormalized? Why are we renormalizing parameters? The parameter vector for a multinomial distribution can be thought of as a probability density (it's numbers that are nonnegative and sum to one). When we take a subvector, we need to renormalize to get another multinomial parameter vector (do what it takes to make the numbers sum to one). That's what's going on in (5.39b).

Proof of Theorem 5.17. Just calculate. The relevant marginal is the distribution of $(Y_{j+1}, \ldots, Y_k)$ but that isn't a brand name distribution. Almost as good is the marginal of

$$\mathbf{Z} = (Y_1 + \cdots + Y_j, Y_{j+1}, \ldots, Y_k) = (n - Y_{j+1} - \cdots - Y_k, Y_{j+1}, \ldots, Y_k) \tag{5.40}$$

which is $\text{Multi}_{k-j+1}(n, \mathbf{q})$ with

$$\mathbf{q} = (p_1 + \cdots + p_j, p_{j+1}, \ldots, p_k) = (n - p_{j+1} - \cdots - p_k, p_{j+1}, \ldots, p_k)$$

It's almost the same thing really, because the right hand side of (5.40) is a function of Y_{j+1}, …, Y_k alone, hence

$$\begin{aligned} P(Y_i = y_i,\ i = j+1, \ldots, k) & \\ &= \binom{n}{n - y_{j+1} - \cdots - y_k, y_{j+1}, \ldots, y_k} \\ &\qquad \times (1 - p_{j+1} - \cdots - p_k)^{n - y_{j+1} - \cdots - y_k} p_{j+1}^{y_{j+1}} \cdots p_k^{y_k} \end{aligned}$$

And, of course, conditional equals joint over marginal

$$\begin{aligned} &\frac{\binom{n}{y_1, \ldots, y_k} p_1^{y_1} \cdots p_k^{y_k}}{\binom{n}{n - y_{j+1} - \cdots - y_k, y_{j+1}, \ldots, y_k} (1 - p_{j+1} - \cdots - p_k)^{n - y_{j+1} - \cdots - y_k} p_{j+1}^{y_{j+1}} \cdots p_k^{y_k}} \\ &\qquad = \frac{n!}{y_1! \cdots y_k!} \cdot \frac{(n - y_{j+1} - \cdots - y_k)! y_{j+1}! \cdots y_k!}{n!} \\ &\qquad\qquad \times \frac{p_1^{y_1} \cdots p_j^{y_j}}{(1 - p_{j+1} - \cdots - p_k)^{n - y_{j+1} - \cdots y_k}} \\ &\qquad = \frac{(n - y_{j+1} - \cdots - y_k)!}{y_1! \cdots y_j!} \prod_{i=1}^{j} \left(\frac{p_i}{1 - p_{j+1} - \cdots p_k} \right)^{y_j} \\ &\qquad = \binom{n - y_{j+1} - \cdots - y_k}{y_1, \ldots, y_j} \prod_{i=1}^{j} \left(\frac{p_i}{p_1 + \cdots + p_j} \right)^{y_j} \end{aligned}$$

and that's the conditional density asserted by the theorem. □

Problems

5-1. Is

$$\begin{pmatrix} 3 & 2 & -1 \\ 2 & 3 & 2 \\ -1 & 2 & 3 \end{pmatrix}$$

a covariance matrix? If not, why not?

5-2. Is

$$\begin{pmatrix} 3 & 2 & -1/3 \\ 2 & 3 & 2 \\ -1/3 & 2 & 3 \end{pmatrix}$$

a covariance matrix? If not, why not? If it is a covariance matrix, is a random vector having this covariance matrix degenerate or non-degenerate?

5-3. Consider the degenerate random vector (X, Y) in $\mathbb{R}^2$ defined by

$$\begin{aligned} X &= \sin(U) \\ Y &= \cos(U) \end{aligned}$$

where $U \sim \mathcal{U}(0, 2\pi)$. We say that (X, Y) has the uniform distribution on the unit circle. Find the mean vector and covariance matrix of (X, Y).

5-4. Let $\mathbf{M}$ be any symmetric positive semi-definite matrix, and denote its elements m_{ij}. Show that for any i and j

$$-1 \le \frac{m_{ij}}{\sqrt{m_{ii} m_{jj}}} \le 1 \tag{5.41}$$

Hint: Consider $\mathbf{w}'\mathbf{M}\mathbf{w}$ for vectors $\mathbf{w}$ having all elements zero except the i-th and j-th.

The point of the problem (this isn't part of the problem, just the explanation of why it is interesting) is that if $\mathbf{M}$ is a variance, then the fraction in (5.41) is $\operatorname{cor}(X_i, X_j)$. Thus positive semi-definiteness is a stronger requirement than the correlation inequality, as claimed in Section 5.1.4.

5-5. Show that the usual formula for the univariate normal distribution is the one-dimensional case of the formula for the multivariate normal distribution.

5-6. Show that a constant random vector (a random vector having a distribution concentrated at one point) is a (degenerate) special case of the multivariate normal distribution.

5-7. Suppose $\mathbf{X} = (X_1, \ldots, X_k)$ has the multinomial distribution with sample size n and parameter vector $\mathbf{p} = (p_1, \ldots, p_k)$, show that for $i \neq j$

$$\frac{\operatorname{var}(X_i - X_j)}{n} = p_i + p_j - (p_i - p_j)^2$$

5-8. If $\mathbf{X} \sim \mathcal{N}(0, \mathbf{M})$ is a non-degenerate normal random vector, what is the distribution of $\mathbf{Y} = \mathbf{M}^{-1}\mathbf{X}$?

5-9. Prove (5.35).
Hint: Write

$$\mathbf{X}_1 - \boldsymbol{\mu}_1 = [\mathbf{X}_1 - E(\mathbf{X}_1 \mid \mathbf{X}_2)] + [E(\mathbf{X}_1 \mid \mathbf{X}_2) - \boldsymbol{\mu}_1]$$

then use the alternate variance and covariance expressions in Theorem 5.2 and linearity of expectation.

5-10. Specialize the formula (5.24) for the non-degenerate multivariate normal density to the two-dimensional case, obtaining

$$
\begin{aligned}
f(x, y) = {} & \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \times \\
& \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right)
\end{aligned}
$$

Hint: To do this you need to know how to invert a 2 × 2 matrix and calculate its determinant. If

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

then

$$\det(\mathbf{A}) = a_{11}a_{22} - a_{12}a_{21}$$

and

$$\mathbf{A}^{-1} = \frac{\begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}}{\det(\mathbf{A})}$$

(This is a special case of Cramer's rule. It can also be verified by just doing the matrix multiplication. Verification of the formulas in the hint is *not* part of the problem.)

5-11. Specialize the conditional mean and variance in Theorem 5.15 to the two-dimensional case, obtaining

$$
\begin{aligned}
E(X \mid Y) &= \mu_X + \rho\frac{\sigma_X}{\sigma_Y}(Y - \mu_Y) \\
\operatorname{var}(X \mid Y) &= \sigma_X^2(1 - \rho^2)
\end{aligned}
$$

5-12 (Ellipsoids of Concentration). Suppose $\mathbf{X}$ is a non-degenerate normal random variable with density (5.24), which we rewrite as

$$f(\mathbf{x}) = \frac{e^{-q(\mathbf{x})/2}}{(2\pi)^{n/2}\det(\mathbf{M})^{1/2}}$$

A *level set* of the density, also called a *highest density region* is a set of the form

$$S = \{\, \mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) > c \,\}$$

for some constant c. Show that this can also be written

$$S = \{\, \mathbf{x} \in \mathbb{R}^n : q(\mathbf{x}) < d \,\}$$

for some other constant d. (A set like this, a level set of a positive definite quadratic form, is called an ellipsoid.) Give a formula for $P(\mathbf{X} \in S)$ as a function of the constant d in terms of the probabilities for a univariate brand name distribution. (**Hint:** Use Problem 12-32 in Lindgren.)

5-13. For the random vector $\mathbf{X}$ defined by (5.23) in Example 5.1.3 suppose U, V, and W are i. i. d. standard normal random variables.

(a) What is the joint distribution of the two-dimensional random vector whose components are the first two components of $\mathbf{X}$?

(b) What is the conditional distribution of the first component of $\mathbf{X}$ given the second?

5-14. Suppose Z_1, Z_2, $\ldots$ are i. i. d. $\mathcal{N}(0, \tau^2)$ random variables and X_1, X_2, $\ldots$ are defined recursively as follows.

- X_1 is a $\mathcal{N}(0, \sigma^2)$ random variable that is independent of all the Z_i.
- for $i > 1$
$$X_{i+1} = \rho X_i + Z_i.$$

There are three unknown parameters, ρ, σ^2, and τ^2, in this model. Because they are variances, we must have $\sigma^2 > 0$ and $\tau^2 > 0$. The model is called an autoregressive time series of order one or AR(1) for short. The model is said to be *stationary* if X_i has the same marginal distribution for all i.

(a) Show that the joint distribution of X_1, X_2, $\ldots$, X_n is multivariate normal.

(b) Show that $E(X_i) = 0$ for all i.

(c) Show that the model is stationary only if $\rho^2 < 1$ and
$$\sigma^2 = \frac{\tau^2}{1 - \rho^2}$$

Hint: Consider $\operatorname{var}(X_i)$.

(d) Show that
$$\operatorname{cov}(X_i, X_{i+k}) = \rho^k \sigma^2, \qquad k \geq 0$$
in the stationary model.

Chapter 6

Convergence Concepts

6.1 Univariate Theory

Chapter 5 in Lindgren is a jumble of convergence theory. Here we will follow one thread through the jumble, ignoring many of the convergence concepts discussed by Lindgren. The only ones widely used in statistics are *convergence in distribution* and its special case *convergence in probability to a constant*. We will concentrate on them.

6.1.1 Convergence in Distribution

Definition 6.1.1 (Convergence in Distribution).
A sequence of random variables X_1, X_2, ... with X_n having distribution function F_n converges in distribution *to a random variable X with distribution function F if*

$$F_n(x) \to F(x), \qquad \text{as } n \to \infty$$

for every real number x that is a continuity point of F. We indicate this by writing

$$X_n \xrightarrow{\mathcal{D}} X, \qquad \text{as } n \to \infty.$$

"Continuity point" means a point x such that F is continuous at x (a point where F does not jump). If the limiting random variable X is continuous, then every point is a continuity point. If X is discrete or of mixed type, then $F_n(x) \to F(x)$ must hold at points x where F does not jump but it does not have to hold at the jumps.

From the definition it is clear that convergence in distribution is a statement about distributions not variables. Though we write $X_n \xrightarrow{\mathcal{D}} X$, what this means is that the *distribution of* X_n converges to the *distribution of* X. We could dispense with the notion of convergence in distribution and always write $F_{X_n}(x) \to F_X(x)$ for all continuity points x of F_X in place of $X_n \xrightarrow{\mathcal{D}} X$, but that would be terribly cumbersome.

There is a much more general notion of convergence in distribution (also called *convergence in law* or *weak convergence*) that is equivalent to the concept defined in Definition 6.1.1.

Theorem 6.1 (Helly-Bray). *A sequence of random variables X_1, X_2, ... converges in distribution to a random variable X if and only if*

$$E\{g(X_n)\} \to E\{g(X)\}$$

for every bounded continuous function $g : \mathbb{R} \to \mathbb{R}$.

For comparison, Definition 6.1.1 says, when rewritten in analogous notation

$$E\{I_{(-\infty,x]}(X_n)\} \to E\{I_{(-\infty,x]}(X)\}, \qquad \text{whenever } P(X = x) = 0. \tag{6.1}$$

Theorem 6.1 doesn't explicitly mention continuity points, but the continuity issue is there implicitly. Note that

$$E\{I_A(X_n)\} = P(X_n \in A)$$

may fail to converge to

$$E\{I_A(X)\} = P(X \in A)$$

because indicator functions, though bounded, are not continuous. And (6.1) says that expectations of some indicator functions converge and others don't (at least not necessarily).

Also note that $E(X_n)$ may fail to converge to $E(X)$ because the identity function, though continuous, is unbounded. Nevertheless, the Theorem 6.1 does imply convergence of expectations of many interesting functions.

How does one establish that a sequence of random variables converges in distribution? By writing down the distribution functions and showing that they converge? No. In the common applications of convergence in distribution in statistics, convergence in distribution is a consequence of the central limit theorem or the law of large numbers.

6.1.2 The Central Limit Theorem

Theorem 6.2 (The Central Limit Theorem (CLT)). *If X_1, X_2, ... is a sequence of independent, identically distributed random variables having mean μ and variance σ^2 and*

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{6.2}$$

is the sample mean for sample size n, then

$$\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow{\mathcal{D}} Y, \qquad \text{as } n \to \infty, \tag{6.3}$$

where $Y \sim \mathcal{N}(0, \sigma^2)$.

It simplifies notation if we are allowed to write a distribution on the right hand side of a statement about convergence in distribution, simplifying (6.3) and the rest of the sentence following it to

$$\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2), \qquad \text{as } n \to \infty. \tag{6.4}$$

There's nothing wrong with this mixed notation because (to repeat something said earlier) convergence in distribution is a statement about distributions of random variables, not about the random variables themselves. So when we replace a random variable with its distribution, the meaning is still clear.

The only requirement for the CLT to hold is that the variance σ^2 exist (this implies that the mean μ also exists by Theorem 2.44 of Chapter 2 of these notes. No other property of the distribution of the X_i matters.

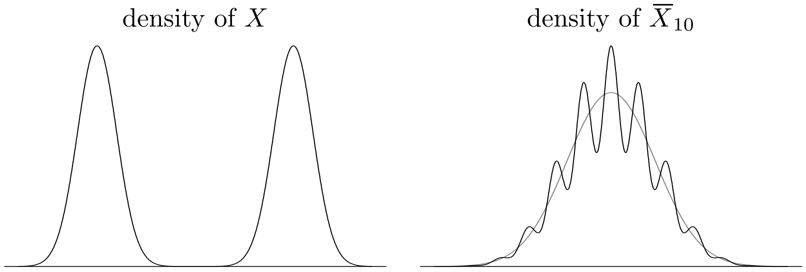
The left hand side of (6.3) always has mean zero and variance σ^2 for all n, regardless of the distribution of the X_i so long as the variance exists. Thus the central limit theorem doesn't say anything about means and variances, rather it says that the *shape* of the distribution of $\overline{X}_n$ approaches the bell-shaped curve of the normal distribution as $n \to \infty$.

A sloppy way of rephrasing (6.3) is

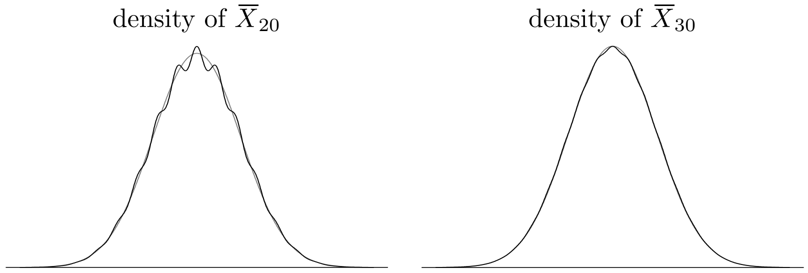
$$\overline{X}_n \approx \mathcal{N}\left(\mu, \tfrac{\sigma^2}{n}\right)$$

for "large n." Most of the time the sloppiness causes no harm and no one is confused. The mean and variance of $\overline{X}_n$ are indeed μ and σ^2/n and the shape of the distribution is approximately normal if n is large. What one cannot do is say $\overline{X}_n$ converges in distribution to Z where Z is $\mathcal{N}(\mu, \sigma^2/n)$. Having an n in the supposed limit of a sequence is mathematical nonsense.

Example 6.1.1 (A Symmetric Bimodal Distribution).
Let us take a look at how the CLT works in practice. How large does n have to be before the distribution of $\overline{X}_n$ is approximately normal?



On the left is a severely bimodal probability density function. On the right is the density of (6.2), where $n = 10$ and the X_i are i. i. d. with the density on the left. The wiggly curve is the density of $\overline{X}_{10}$ and the smooth curve is the normal density with the same mean and variance. The two densities on the right are not very close. The CLT doesn't provide a good approximation at $n = 10$.



density of $\overline{X}_{20}$

density of $\overline{X}_{30}$

At $n = 20$ and $n = 30$ we have much better results. The density of $\overline{X}_{30}$ is almost indistinguishable from the normal density with the same mean and variance. There is a bit of wiggle at the top of the curve, but everywhere else the fit is terrific. It is this kind of behavior that leads to the rule of thumb propounded in elementary statistics texts that $n > 30$ is "large sample" territory, thirty is practically infinity.

The symmetric bimodal density we started with in this example is of no practical importance. Its only virtue giving rise to a density for $\overline{X}_n$ that is easy to calculate. If you are not interested in the details of this example, skip to the next example. If you wish to play around with this example, varying different aspects to see what happens, go to the web page

`http://www.stat.umn.edu/geyer/5101/clt.html#bi`

The symmetric bimodal density here is the density of $X = Y + Z$, where $Y \sim \text{Ber}(p)$ and $Z \sim \mathcal{N}(0, \sigma^2)$, where $p = \frac{1}{2}$ and $\sigma = 0.1$. If Y_i and Z_i are i. i. d. sequences, then, of course

$$\sum_{i=1}^{n} Y_i \sim \text{Bin}\,(n, p)$$

$$\sum_{i=1}^{n} Z_i \sim \mathcal{N}\left(0, n\sigma^2\right)$$

So by the convolution theorem the density of their sum is

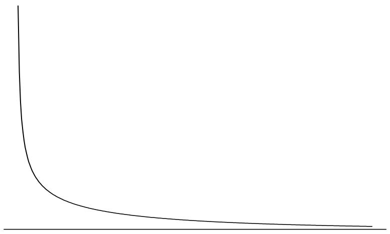
$$f_{X_1+\cdots+X_n}(s) = \sum_{k=0}^{n} f(k \mid n, p)\phi(s - k \mid 0, n\sigma^2)$$

where $f(k \mid n, p)$ is the the $\text{Bin}(n, p)$ density and $\phi(z \mid \mu, \sigma^2)$ is the $\mathcal{N}(\mu, \sigma^2)$ density. The the distribution of $\overline{X}_n$ is given by

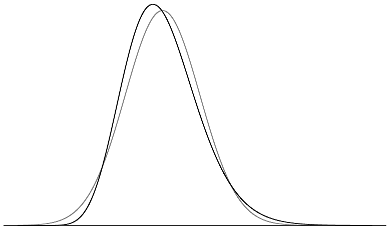
$$f_{\overline{X}_n}(w) = n f_{X_1+\cdots+X_n}(nw) = n \sum_{k=0}^{n} f(k \mid n, p)\phi(nw - k \mid 0, n\sigma^2) \tag{6.5}$$

Example 6.1.2 (A Skewed Distribution).
The $30 = \infty$ "rule" promulgated in introductory statistics texts does not hold for skewed distributions. Consider X having the chi-square distribution with one degree of freedom.

density of X

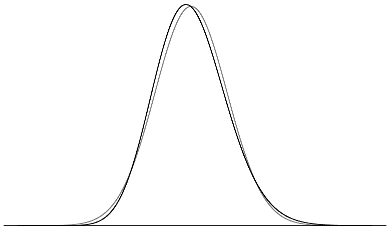


density of $\overline{X}_{30}$

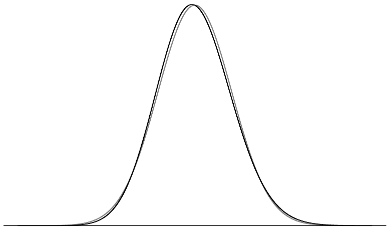


The density of X is shown on the left. It is extremely skewed going to infinity at zero. On the right is the density of $\overline{X}_{30}$ and the normal density with the same mean and variance. The fit is not good. The density of $\overline{X}_{30}$, a rescaled chi$^2(30)$ density, is still rather skewed and so cannot be close to a normal density, which of course is symmetric.

density of $\overline{X}_{100}$



density of $\overline{X}_{300}$



The fit is better at $n = 100$ and $n = 300$, but still not as good as our bimodal example at $n = 30$. The moral of the story is that skewness slows convergence in the central limit theorem.

If you wish to play around with this example, varying different aspects to see what happens, go to the web page

```
http://www.stat.umn.edu/geyer/5101/clt.html#expo
```

6.1.3 Convergence in Probability

A special case of convergence in distribution is convergence in distribution to a degenerate random variable concentrated at one point, $X_n \xrightarrow{\mathcal{D}} a$ where a is a constant. Theorem 2 of Chapter 5 in Lindgren says that this is equivalent to the following notion.

Definition 6.1.2 (Convergence in Probability to a Constant).
A sequence of random variables X_1, X_2, $\ldots$ converges in probability *to a constant a if for every $\epsilon > 0$*

$$P(|X_n - a| > \epsilon) \to 0, \qquad \text{as } n \to \infty.$$

We indicate X_n converging in probability to a by writing

$$X_n \xrightarrow{P} a, \qquad \text{as } n \to \infty.$$

Convergence in probability to a constant and convergence in distribution to a constant are the same thing, so we could write $X_n \xrightarrow{\mathcal{D}} a$ instead of $X_n \xrightarrow{P} a$, but the latter is traditional. There is also a more general notion of convergence in probability to a *random variable*, but it has no application in statistics and we shall ignore it.

6.1.4 The Law of Large Numbers

One place convergence in probability appears is in the law of large numbers.

Theorem 6.3 (Law of Large Numbers (LLN)). *If X_1, X_2, ... is a sequence of independent, identically distributed random variables having mean μ, and*

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

is the sample mean for sample size n, then

$$\overline{X}_n \xrightarrow{P} \mu, \qquad \text{as } n \to \infty. \tag{6.6}$$

The only requirement is that the mean μ exist. No other property of the distribution of the X_i matters.

6.1.5 The Continuous Mapping Theorem

Theorem 6.4 (Continuous Mapping). *If g is a function continuous at all points of a set A, if $X_n \xrightarrow{\mathcal{D}} X$, and if $P(X \in A) = 1$, then $g(X_n) \xrightarrow{\mathcal{D}} g(X)$.*

The main point of the theorem is the following two corollaries.

Corollary 6.5. *If g is an everywhere continuous function and $X_n \xrightarrow{\mathcal{D}} X$, then $g(X_n) \xrightarrow{\mathcal{D}} g(X)$.*

Here the set A in the theorem is the whole real line. Hence the condition $P(X \in A) = 1$ is trivial.

Corollary 6.6. *If g is a function continuous at the point a and $X_n \xrightarrow{P} a$, then $g(X_n) \xrightarrow{P} g(a)$.*

Here the set A in the theorem is just the singleton set $\{a\}$, but the limit variable in question is the constant random variable satisfying $P(X = a) = 1$.

These theorems say that convergence in distribution and convergence in probability to a constant behave well under a continuous change of variable.

Rewriting the CLT

The CLT can be written in a variety of slightly different forms. To start, let us rewrite (6.3) as

$$\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow{\mathcal{D}} \sigma Z, \qquad \text{as } n \to \infty,$$

where now Z is a standard normal random variable. If $\sigma > 0$, then we can divide both sides by σ. This is a simple application of the continuous mapping theorem, the function defined by $g(x) = x/\sigma$ being continuous. It gives

$$\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \xrightarrow{\mathcal{D}} Z$$

Moving the $\sqrt{n}$ from the numerator to the denominator of the denominator gives

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{D}} Z \tag{6.7}$$

6.1.6 Slutsky's Theorem

Theorem 6.7 (Slutsky). *If $g(x, y)$ is a function jointly continuous at every point of the form (x, a) for some fixed a, and if $X_n \xrightarrow{\mathcal{D}} X$ and $Y_n \xrightarrow{P} a$, then*

$$g(X_n, Y_n) \xrightarrow{\mathcal{D}} g(X, a).$$

Corollary 6.8. *If $X_n \xrightarrow{\mathcal{D}} X$ and $Y_n \xrightarrow{P} a$, then*

$$X_n + Y_n \xrightarrow{\mathcal{D}} X + a,$$
$$Y_n X_n \xrightarrow{\mathcal{D}} aX,$$

and if $a \neq 0$

$$X_n / Y_n \xrightarrow{\mathcal{D}} X/a.$$

In other words, we have all the nice properties we expect of limits, the limit of a sum is the sum of the limits, and so forth. The point of the theorem is this is *not true* unless one of the limits is a constant. If we only had $X_n \xrightarrow{\mathcal{D}} X$ and $Y_n \xrightarrow{\mathcal{D}} Y$, we couldn't say anything about the limit of $X_n + Y_n$ without knowing about the *joint* distribution of X_n and Y_n. When Y_n converges to a constant, Slutsky's theorem tells us that we don't need to know anything about joint distributions.

A special case of Slutsky's theorem involves two sequences converging in probability. If $X_n \xrightarrow{P} a$ and $Y_n \xrightarrow{P} b$, then $X_n + Y_n \xrightarrow{P} a + b$, and so forth. This is a special case of Slutsky's theorem because convergence in probability to a constant is the same as convergence in distribution to a constant.

6.1.7 Comparison of the LLN and the CLT

When X_1, X_2, $\ldots$ is an i. i. d. sequence of random variables having a variance, both the law of large numbers and the central limit theorem apply, but the CLT tells us much more than the LLN.

It could not tell us less, because the CLT implies the LLN. By Slutsky's theorem, the CLT (6.3) implies

$$\overline{X}_n - \mu = \frac{1}{\sqrt{n}} \cdot \sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow{\mathcal{D}} 0 \cdot Y = 0$$

where $Y \sim \mathcal{N}(0, \sigma^2)$. Because convergence in distribution to a constant and convergence in probability to a constant are the same thing, this implies the LLN.

But the CLT gives much more information than the LLN. It says that the size of the estimation error $\overline{X}_n - \mu$ is about $\sigma / \sqrt{n}$ and also gives us the *shape* of the error distribution (i. e., normal).

So why do we even care about the law of large numbers? Is it because there are lots of important probability models having a mean but no variance (so the LLN holds but the CLT does not)? No, not any used for real data. The point is that sometimes we don't care about the information obtained from the central limit theorem. When the only fact we want to use is $\overline{X}_n \xrightarrow{P} \mu$, we refer to the law of large numbers as our authority. Its statement is simpler, and there is no point in dragging an unnecessary assumption about variance in where it's not needed.

6.1.8 Applying the CLT to Addition Rules

The central limit theorem says that the sum of i. i. d. random variables with a variance is approximately normally distributed if the number of variables in the sum is "large." Applying this to the addition rules above gives several normal approximations.

Binomial The $\text{Bin}(n, p)$ distribution is approximately normal with mean np and variance $np(1 - p)$ if n is large.

Negative Binomial The $\text{NegBin}(n, p)$ distribution is approximately normal with mean n/p and variance $n(1 - p)/p^2$ if n is large.

Poisson The $\text{Poi}(\mu)$ distribution is approximately normal with mean μ and variance μ if μ is large.

Gamma The $\text{Gam}(\alpha, \lambda)$ distribution is approximately normal with mean α/λ and variance α/λ^2 if α is large.

Chi-Square The $\text{chi}^2(n)$ distribution is approximately normal with mean n and variance $2n$ if n is large.

Comment The rules containing n are obvious combinations of the relevant addition rule and the CLT. The rules for the Poisson and gamma distributions are a bit weird in that there is no n. To understand them we need the notion of an infinitely divisible distribution.

Definition 6.1.3.
*A probability distribution P is **infinitely divisible** if for every positive integer n there exist independent and identically distributed random variables X_1, ..., X_n such that $X_1 + \cdots + X_n$ has the distribution P.*

Example 6.1.3 (Infinite Divisibility of the Poisson).
By the addition rule for Poisson random variables, $X_1 + \cdots + X_n \sim \text{Poi}(\mu)$ when the X_i are i. i. d. $\text{Poi}(\mu/n)$. Thus the $\text{Poi}(\mu)$ distribution is infinitely divisible for any $\mu > 0$.

Example 6.1.4 (Infinite Divisibility of the Gamma).
By the addition rule for gamma random variables, $X_1 + \cdots + X_n \sim \text{Gam}(\alpha, \lambda)$ when the X_i are i. i. d. $\text{Gam}(\alpha/n, \lambda)$. Thus the $\text{Gam}(\alpha, \lambda)$ distribution is infinitely divisible for any $\alpha > 0$ and $\lambda > 0$.

The infinite divisibility of the Poisson and gamma distributions explains the applicability of the CLT. But we have to be careful. Things are not quite as simple as they look.

A Bogus Proof that Poisson is Normal Every Poisson random variable is the sum of n i. i. d. random variables and n can be chosen as large as we please. Thus by the CLT the Poisson distribution is arbitrarily close to normal. Therefore it is normal.

Critique of the Bogus Proof For one thing, it is obviously wrong. The Poisson discrete is discrete. The Normal distribution is continuous. They can't be equal. But what's wrong with the proof?

The problem is in sloppy application of the CLT. It is often taken to say what the bogus proof uses, and the sloppy notation (6.4) encourages this sloppy use, which usually does no harm, but is the problem here.

A more careful statement of the CLT says that for any fixed μ and large enough n the $\text{Poi}(n\mu)$ distribution is approximately normal. The n that is required to get close to normal depends on μ. This does tell us that for sufficient large values of the parameter, the Poisson distribution is approximately normal. It does *not* tell us the Poisson distribution is approximately normal for *any* value of the parameter, which the sloppy version seems to imply.

The argument for the gamma distribution is exactly analogous to the argument for the Poisson. For large enough values of the parameter α involved in the infinite divisibility argument, the distribution is approximately normal. The statement about the chi-square distribution is a special case of the statement for the gamma distribution.

6.1.9 The Cauchy Distribution

The Cauchy location-scale family, abbreviated Cauchy(μ, σ) is described in Section B.2.7 of Appendix B an addition rule given by (C.11) in Appendix C, which we repeat here

$$X_1 + \cdots + X_n \sim \text{Cauchy}(n\mu, n\sigma) \tag{6.8}$$

from which we can derive the distribution of the sample mean

$$\overline{X}_n \sim \text{Cauchy}(\mu, \sigma) \tag{6.9}$$

(Problem 6-1).

The Cauchy family is not a useful model for real data, but it is theoretically important as a source of counterexamples. A Cauchy(μ, σ) distribution has center of symmetry μ. Hence μ is the median, but μ is not the mean because the mean does not exist.

The rule for the mean (6.9) can be trivially restated as a convergence in distribution result

$$\overline{X}_n \xrightarrow{\mathcal{D}} \text{Cauchy}(\mu, \sigma), \qquad \text{as } n \to \infty \tag{6.10}$$

a "trivial" result because $\overline{X}_n$ actually has exactly the Cauchy(μ, σ) distribution for all n, so the assertion that is gets close to that distribution for large n is trivial (exactly correct is indeed a special case of "close").

The reason for stating (6.10) is for contrast with the law of large numbers (LLN), which can be stated as follows: if X_1, X_2, $\ldots$ are i. i. d. from a distribution with mean μ, then

$$\overline{X}_n \xrightarrow{P} \mu \qquad \text{as } n \to \infty \tag{6.11}$$

The condition for the LLN, that the mean exist, does not hold for the Cauchy. Furthermore, since μ does not exist, $\overline{X}_n$ cannot converge to it. But it is conceivable that

$$\overline{X}_n \xrightarrow{P} c \qquad \text{as } n \to \infty \tag{6.12}$$

for some constant c, even though this does not follow from the LLN. The result (6.10) for the Cauchy rules this out. Convergence in probability to a constant is the same as convergence in distribution to a constant (Theorem 2 of Chapter 5 in Lindgren). Thus (6.12) and (6.10) are contradictory. Since (6.10) is correct, (6.12) must be wrong. For the Cauchy distribution $\overline{X}_n$ does not converge in probability to anything.

Of course, the CLT also fails for the Cauchy distribution. The CLT implies the LLN. Hence if the CLT held, the LLN would also hold. Since the LLN doesn't hold for the Cauchy, the CLT can't hold either.

Problems

6-1. Derive (6.9) from (6.8) using the change of variable theorem.

6-2. Suppose that S_1, S_2, ... is any sequence of random variables such that $S_n \xrightarrow{P} \sigma$, and X_1, X_2, ... are independent and identically distributed with mean μ and variance σ^2 and $\sigma > 0$. Show that

$$\frac{\overline{X}_n - \mu}{S_n / \sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1), \qquad \text{as } n \to \infty,$$

where, as usual,

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

6-3. Suppose X_1, X_2, ... are i. i. d. with common probability measure P, and define $Y_n = I_A(X_n)$ for some event A, that is,

$$Y_n = \begin{cases} 1, & X_n \in A \\ 0, & X_n \notin A \end{cases}$$

Show that $\overline{Y}_n \xrightarrow{P} P(A)$.

6-4. Suppose the sequences X_1, X_2, ... and Y_1, Y_2, ... are defined as in Problem 6-3, and write $P(A) = p$. Show that

$$\sqrt{n}(\overline{Y}_n - p) \xrightarrow{\mathcal{D}} \mathcal{N}\bigl(0, p(1-p)\bigr)$$

and also show that

$$\frac{\overline{Y}_n - p}{\sqrt{\overline{Y}_n(1 - \overline{Y}_n)/n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$$

Hint: What is the distribution of $\sum_i Y_i$? Also use Problem 6-2.

Chapter 7

Sampling Theory

7.1 Empirical Distributions

In statistics, we often deal with complicated data, but for learning it is best to start simple. The simplest sort of data is just a set of numbers that are measurements of one variable on a set of individuals. In the next section we will see that it is important that these individuals be a random sample from some population of interest. For now we will just treat the data as a set of numbers.

Example 7.1.1 (A Data Set).
The numbers below were generated by computer and are a random sample from an Exp(1) distribution rounded to one significant figure. Because of the rounding, there are duplicate values. If not rounded the values would all be different, as would be the case for any sample from any continuous distribution.

0.12 3.15 0.77 1.02 0.08 0.35 0.29 1.05 0.49 0.81

A vector

$$\mathbf{x} = (x_1, \ldots, x_n) \tag{7.1}$$

can be thought of as a function of the index variable i. To indicate this we can write the components as $x(i)$ instead of x_i. Then x is a function on the index set $\{1, \ldots, n\}$. Sometimes we don't even bother to change the notation but still think of the vector as being the function $i \mapsto x_i$.

This idea is useful in probability theory because of the dogma "a random variable is a function on the sample space." So let us think of the index set $S = \{1, \ldots, n\}$ as the sample space, and X as a random variable having values $X(i)$, also written x_i. When we consider a uniform distribution on the sample space, which means each point gets probability $1/n$ since there are n points, then the distribution of X is called the *empirical distribution* associated with the vector (7.1).

By definition, the probability function of X is

$$f(x) = P(X = x) = \sum_{\substack{i \in S \\ x_i = x}} \frac{1}{n} = \frac{\operatorname{card}(\{\, i \in S : x_i = x \,\})}{n}$$

where, as usual, $\operatorname{card}(A)$ denotes the *cardinality* of the set A. If all of the x_i are distinct, then the distribution of X is also uniform. Otherwise, it is not. If the point x occurs m times among the x_i, then $f(x) = m/n$. This makes the definition of the empirical distribution in terms of its probability function rather messy. So we won't use it.

The description in terms of expectation is much simpler.

Definition 7.1.1 (Empirical Expectation).
The empirical expectation operator associated with the vector $(x_1, \ldots, x_n)$ is denoted E_n and defined by

$$E_n\{g(X)\} = \frac{1}{n} \sum_{i=1}^{n} g(x_i). \tag{7.2}$$

Example 7.1.2.
For the data in Example 7.1.1 we have for the function $g(x) = x$

$$E_n(X) = \frac{1}{n} \sum_{i=1}^{n} x_i = 0.813$$

and for the function $g(x) = x^2$

$$E_n(X^2) = \frac{1}{n} \sum_{i=1}^{n} x_i^2 = 1.37819$$

The corresponding probability measure P_n is found by using "probability is just expectation of indicator functions."

Definition 7.1.2 (Empirical Probability Measure).
The empirical probability measure associated with the vector $(x_1, \ldots, x_n)$ is denoted P_n and defined by

$$P_n(A) = \frac{1}{n} \sum_{i=1}^{n} I_A(x_i). \tag{7.3}$$

Example 7.1.3.
For the data in Example 7.1.1 we have for the event $X > 2$

$$P_n(X > 2) = \frac{1}{n} \sum_{i=1}^{n} I_{(2,\infty)}(x_i) = \frac{\text{number of } x_i \text{ greater than 2}}{n} = 0.1$$

and for the event $1 < X < 2$

$$P_n(1 < X < 2) = \frac{1}{n} \sum_{i=1}^{n} I_{(1,2)}(x_i) = \frac{\text{number of } x_i \text{ between 1 and 2}}{n} = 0.2$$

7.1.1 The Mean of the Empirical Distribution

For the rest of this section we consider the special case in which all of the x_i are real numbers.

The *mean* of the empirical distribution is conventionally denoted by $\bar{x}_n$ and is obtained by taking the case $g(x) = x$ in (7.2)

$$\bar{x}_n = E_n(X) = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

7.1.2 The Variance of the Empirical Distribution

The variance of the empirical distribution has no conventional notation, but we will use both $\operatorname{var}_n(X)$ and v_n. Just like any other variance, it is the expected squared deviation from the mean. The mean is $\bar{x}_n$, so

$$v_n = \operatorname{var}_n(X) = E_n\{(X - \bar{x}_n)^2\} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \tag{7.4}$$

It is important that you think of the empirical distribution as a probability distribution just like any other. This gives us many facts about empirical distributions, that are derived from general facts about probability and expectation. For example, the parallel axis theorem holds, just as it does for any probability distribution. For ease of comparison, we repeat the general parallel axis theorem (Theorem 2.11 of Chapter 2.27 of these notes).

If X is a real-valued random variable having finite variance and a is any real number, then

$$E\{(X - a)^2\} = \operatorname{var}(X) + [a - E(X)]^2 \tag{7.5}$$

Corollary 7.1 (Empirical Parallel Axis Theorem).

$$E_n\{(X - a)^2\} = \operatorname{var}_n(X) + [a - E_n(X)]^2$$

or, in other notation,

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - a)^2 = v_n + (a - \bar{x}_n)^2 \tag{7.6}$$

In particular, the case $a = 0$ gives the empirical version of

$$\operatorname{var}(X) = E(X^2) - E(X)^2$$

which is

$$\operatorname{var}_n(X) = E_n(X^2) - E_n(X)^2$$

or, in other notation,

$$v_n = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}_n^2. \tag{7.7}$$

Example 7.1.4.
In Example 7.1.2 we found for the data in Example 7.1.1

$$\bar{x}_n = E_n(X) = 0.813$$

and

$$E_n(X^2) = 1.37819.$$

Although we could use the definition (7.4) directly, we can also use the empirical parallel axis theorem in the form (7.7)

$$v_n = 1.37819 - 0.813^2 = 0.717221.$$

7.1.3 Characterization of the Mean

Considering a as a variable in (7.5) or (7.6) gives the following pair of theorems. The first one is just the corollary to the parallel axis theorem in Lindgren (p. 107) in different language. It is also the special case of the characterization of conditional expectation as best prediction (Theorem 3.6 in Chapter 3 of these notes) when the conditional expectation is actually unconditional.

Corollary 7.2 (Characterization of the Mean). *The mean of a real-valued random variable X having finite variance is the value of a that minimizes the function*

$$g(a) = E\{(X - a)^2\}$$

which is the expected squared deviation from a.

Corollary 7.3 (Characterization of the Empirical Mean). *The mean of the empirical distribution is the value of a that minimizes the function*

$$g(a) = E_n\{(X - a)^2\} = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2$$

which is the average squared deviation from a.

The point of these two corollaries is that they describe the sense in which the mean is the "center" of a distribution. It is the point to which all other points are closest on average, when "close" is defined in terms of squared differences. The mean is the point from which the average squared deviation is the smallest. We will contrast this characterization with an analogous characterization of the median in Section 7.1.7.

7.1.4 Review of Quantiles

Recall from Section 3.2 in Lindgren the definition of a quantile of a probability distribution.

Definition 7.1.3 (Quantile).
For $0 < p < 1$, a point x is a p-th quantile *of the distribution of a real-valued random variable X if*

$$P(X \leq x) \geq p \quad \textit{and} \quad P(X \geq x) \geq 1 - p$$

If the c. d. f. of X is invertible, then there is a much simpler characterization of quantiles. For $0 < p < 1$, the p-th quantile is the unique solution x of the equation

$$F(x) = p, \tag{7.8a}$$

or in other notation

$$x = F^{-1}(p). \tag{7.8b}$$

The following lemma tells us we are usually in this situation when dealing with continuous random variables

Lemma 7.4. *A continuous random variable having a strictly positive p. d. f. has an invertible c. d. f.*

Proof. There exists a solution to (7.8a), by the intermediate value theorem from calculus, because F is continuous and goes from zero to one as x goes from $-\infty$ to $+\infty$. The solution is unique because

$$F(x + h) = F(x) + \int_x^{x+h} f(x)\,dx$$

and the integral is not zero unless $h = 0$, because the integral of a strictly positive function cannot be zero. □

In general, the p-th quantile need not be unique and it need not be a point satisfying $F(x) = p$ (see Figure 3.3 in Lindgren for examples of each of these phenomena). Hence the technical fussiness of Definition 7.1.3. That definition can be rephrased in terms of c. d. f.'s as follows. A point x is a p-th quantile of a random variable with c. d. f. F if

$$F(x) \geq p \qquad \text{and} \qquad F(y) \leq p, \quad \text{for all } y < x$$

Here the asymmetry of the definition of c. d. f.'s (right continuous but not necessarily left continuous) makes the two conditions asymmetric. Definition 7.1.3 makes the symmetry between left and right clear. If x is a p-th quantile of X, then $-x$ is also a q-th quantile of $-X$, where $q = 1 - p$.

7.1.5 Quantiles of the Empirical Distribution

Now we want to look at the quantiles of the empirical distribution associated with a vector $\mathbf{x}$. In order to discuss this, it helps to establish the following notation. We denote the sorted values of the components of $\mathbf{x}$ by

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}.$$

That is, when we put parentheses around the subscripts, that means we have put the values in ascending order. For any real number x, the notation $\lceil x \rceil$ denotes the smallest integer greater than or equal to x, which is called the *ceiling* of x, and the notation $\lfloor x \rfloor$ denotes the largest integer less than or equal to x, which is called the *floor* of x,

Theorem 7.5. *If np is not an integer, then the p-th quantile of the empirical distribution associated with the vector $\mathbf{x}$ is unique and is equal to $x_{(\lceil np \rceil)}$.*

When np is an integer, then any point x such that

$$x_{(np)} \le x \le x_{(np+1)} \tag{7.9}$$

is a p-th quantile.

Proof. The p-th quantile must be a point x such that there are at least np of the x_i at or below x and at least $n(1-p)$ at or above x.

In the case that np is not an integer, let $k = \lceil np \rceil$. Since np is not an integer, and $\lceil np \rceil$ is the least integer greater than k, we have $k > np > k - 1$. What we must show is that $x_{(k)}$ is the unique p-th quantile.

There are at least $k > np$ data points

$$x_{(1)} \le \cdots \le x_{(k)}$$

at or below $x_{(k)}$. Furthermore, if $i < k$, then $i \le k - 1 < np$ so there are fewer than np data points at or below $x_{(i)}$ unless $x_{(i)}$ happens to be equal to $x_{(k)}$.

Similarly, there are at least $n - k + 1 > n(1-p)$ data points

$$x_{(k)} \le \cdots \le x_{(n)}$$

at or above $x_{(k)}$. Furthermore, if $i > k$, then $n - i + 1 \le n - k < n(1-p)$ so there are fewer than $n(1-p)$ data points at or above $x_{(i)}$ unless $x_{(i)}$ happens to be equal to $x_{(k)}$.

In the case $np = k$, let x be any point satisfying (7.9). Then there are at least $k = np$ data points

$$x_{(1)} \le \cdots \le x_{(k)} \le x$$

at or below x, and there are at least $n - k = n(1-p)$ data points

$$x \le x_{(k+1)} \le \cdots \le x_{(n)}$$

at or above x. Hence x is a p-th quantile. □

Example 7.1.5.
The data in Example 7.1.1 have 10 data points. Thus by the theorem, the empirical quantiles are uniquely defined when np is not an integer, that is, when p is not a multiple of one-tenth.

The first step in figuring out empirical quantiles is always to *sort the data.* Don't forget this step. The sorted data are

0.08	0.12	0.29	0.35	0.49	0.77	0.81	1.02	1.05	3.15

To find the 0.25 quantile, also called the 25-th percentile, the theorem says we find $\lceil np \rceil$, which is the integer above $np = 2.5$, which is 3, and then the empirical quantile is the corresponding order statistic, that is $x_{(3)} = 0.29$.

We remark in passing that if the 25-th percentile is 3 in from the left end of the data in *sorted order*, then the 75-th percentile is 3 in from the right end, so the definition behaves as we expect. Let's check this. First $np = 7.5$. Rounding up gives 8. And $x_{(8)} = 1.02$ is indeed the third from the right.

The definition gets tricky is when np is an integer. If we want the 40-th percentile, $np = 4$. Then the theorem says that any point x between $x_{(4)} = 0.35$ and $x_{(5)} = 0.49$ is a 40-th percentile (0.4 quantile) of these data. For example, 0.35, 0.39, 0.43, and 0.49 are all 40-th percentiles. A bit weird, but that's how the definition works.

7.1.6 The Empirical Median

The median of the empirical distribution we denote by $\tilde{x}_n$. It is the p-th quantile for $p = 1/2$. By the theorem, the median is unique when np is not an integer, which happens whenever n is an odd number. When n is an even number, the empirical median is not unique. It is any point x satisfying (7.9), where $k = n/2$. This nonuniqueness is unsettling to ordinary users of statistics, so a convention has grown up of taking the empirical median to be the midpoint of the interval given by (7.9).

Definition 7.1.4 (Empirical Median).
The median *of the values $x_1, \ldots, x_n$ is the middle value in sorted order when n is odd*

$$\tilde{x}_n = x_{(\lceil n/2 \rceil)}$$

and the average of the two middle values when n is even

$$\tilde{x}_n = \frac{x_{(n/2)} + x_{(n/2+1)}}{2}$$

Example 7.1.6.
The data in Example 7.1.1 have 10 data points. So we are in the "n even" case, and the empirical median is the average of the two middle values of the data *in sorted order*, that is,

$$\tilde{x}_n = \frac{x_{(5)} + x_{(6)}}{2} = \frac{0.49 + 0.77}{2} = 0.63$$

7.1.7 Characterization of the Median

Corollary 7.6 (Characterization of the Median). *If X is a real-valued random variable having finite expectation, then a median of X is any value of a that minimizes the function*

$$g(a) = E\{|X - a|\}$$

which is the expected absolute deviation from a.

Proof. What we need to show is that if m is a median, that is, if

$$P(X \le m) \ge \tfrac{1}{2} \qquad \text{and} \qquad P(X \ge m) \ge \tfrac{1}{2}$$

and a is any real number, then

$$E(|X - a|) \ge E(|X - m|).$$

Without loss of generality, we may suppose $a > m$. (The case $a = m$ is trivial. The case $a < m$ follows from the other case by considering the distribution of $-X$.)

Define

$$g(x) = |x - a| - |x - m|$$

so, by linearity of expectation,

$$E(|X - a|) - E(|X - m|) = E(|X - a| - |X - m|) = E\{g(X)\}$$

So what must be shown is that $E\{g(X)\} \ge 0$.

When $x \le m < a$,

$$g(x) = (a - x) - (m - x) = a - m.$$

Similarly, when $m < a \le x$,

$$g(x) = -(a - m).$$

When $m < x < a$,

$$g(x) = (x - a) - (m - x) = 2(x - m) - (a - m) \ge -(a - m).$$

Thus $g(x) \ge h(x)$ for all x, where

$$h(x) = \begin{cases} a - m, & x \le m \\ -(a - m), & x > m \end{cases}$$

The point is that h can be written in terms of indicator functions

$$h(x) = (a - m)\bigl[I_{(-\infty, m]}(x) - I_{(m, +\infty)}(x)\bigr]$$

so by monotonicity of expectation, linearity of expectation, and "probability is expectation of indicator functions"

$$E\{g(X)\} \ge E\{h(X)\} = (a - m)\bigl[P(X \le m) - P(X > m)\bigr]$$

Because m is a median, the quantity in the square brackets is nonnegative. □

Corollary 7.7 (Characterization of the Empirical Median). *A median of the empirical distribution is a value of a that minimizes the function*

$$g(a) = E_n\{|X - a|\} = \frac{1}{n} \sum_{i=1}^{n} |x_i - a| \tag{7.10}$$

which is the average absolute deviation from a.

There is no end to this game. Every notion that is defined for general probability models, we can specialize to empirical distributions. We can define empirical moments and central moments of all orders, and so forth and so on. But we won't do that in gory detail. What we've done so far is enough for now.

7.2 Samples and Populations

7.2.1 Finite Population Sampling

It is common to apply statistics to a *sample* from a *population.* The population can be any finite set of individuals. Examples are the population of Minnesota today, the set of registered voters in Minneapolis on election day, the set of wolves in Minnesota. A sample is any subset of the population. Examples are the set of voters called by an opinion poll and asked how they intend to vote, the set of wolves fitted with radio collars for a biological experiment. By convention we denote the population size by N and the sample size by n. Typically n is much less than N. For an opinion poll, n is typically about a thousand, and N is in the millions.

Random Sampling

A *random sample* is one drawn so that every individual in the population is equally likely to be in the sample. There are two kinds.

Sampling without Replacement The model for sampling without replacement is dealing from a well-shuffled deck of cards. If we deal n cards from a deck of N cards, there are $(N)_n$ possible outcomes, all equally likely (here we are considering that the order in which the cards are dealt matters). Similarly there are $(N)_n$ possible samples without replacement of size n from a population of size N. If the samples are drawn in such a way that all are equally likely we say we have a *random sample without replacement* from the population.

Sampling with Replacement The model for sampling with replacement is spinning a roulette wheel. If we do n spins and the wheel has N pockets, there are N^n possible outcomes, all equally likely. Similarly there are N^n possible samples with replacement of size n from a population of size N. If the samples are drawn in such a way that all are equally likely we say we have a *random sample with replacement* from the population.

Lindgren calls this a *simple random sample*, although there is no standard meaning of the word "simple" here. Many statisticians would apply "simple" to sampling either with or without replacement using it to mean that all samples are equally likely in contrast to more complicated sampling schemes in which the samples are not all equally likely.

Random Variables

Suppose we are interested in a particular variable, which in principle could be measured for each individual in the population. Write the vector of population values

$$\mathbf{x} = (x_1, \ldots, x_N).$$

Sometimes when x is the only variable of interest we think of this collection of x values as being the population (as opposed to the population being the collection of individuals on whom these measurements could be made).

The vector of population values is *not* a random vector.[1] The population is what it is, and the value x_i for the i-th individual of the population is what it is. Because $\mathbf{x}$ is not random, we use a lower case letter, following the "big X" for random and "little x" for nonrandom convention.

When we take a random sample of size n from the population we obtain a sequence X_1, ..., X_n of values of the variable. Each sample value X_i is one of the population values x_j, but which one is random. That is why we use capital letters for the sample values. When we think of the sample as one thing rather than n things, it is a vector

$$\mathbf{X} = (X_1, \ldots, X_n).$$

Thus we can talk about the probability distributions of each X_i and the joint distribution of all the X_i, which is the same thing as the distribution of the random vector $\mathbf{X}$.

Theorem 7.8 (Sampling Distributions). *If X_1, ..., X_n are a random sample from a population of size n, then the marginal distribution of each X_i is the empirical distribution associated with the population values x_1, ..., x_N.*

If the sampling is with replacement, then the X_i are independent and identically distributed. If the sampling is without replacement, then the X_i are exchangeable but not independent.

Proof. The X_i are exchangeable by definition: every permutation of the sample is equally likely. Hence they are identically distributed, and the marginal distribution of the X_i is the marginal distribution of X_1. Since every individual is equally likely to be the first one drawn, X_1 has the empirical distribution.

Under sampling with replacement, every sample has probability $1/N^n$, which is the product of the marginals. Hence the X_i are independent random variables. Under sampling without replacement, every sample has probability $1/(N)_n$, which is not the product of the marginals. Hence the X_i are dependent random variables. □

Thus, when we have sampling with replacement, we can use formulas that require independence, the most important of these being

[1] When we get to the chapter on Bayesian inference we will see that this sentence carries unexamined philosophical baggage. A Bayesian would say the population values are random too. But we won't worry about that for now.

- the variance of a sum is the sum of the variances

$$\operatorname{var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{var}(X_i) = n\sigma^2 \tag{7.11}$$

  where we have written σ^2 for the variance of all of the X_i (they must have the same variance because they are identically distributed), and

- the joint density is the product of the marginals

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} f_{X_i}(x_i) = \prod_{i=1}^{n} f(x_i) \tag{7.12}$$

  where we have written f for the marginal density of all of the X_i (they must have the same density because they are identically distributed).

When we have sampling without replacement *neither* (7.11) nor (7.12) holds. The analog of (7.11) is derived as follows.

Theorem 7.9 (Finite Population Correction). *If X_1, X_2, …, X_n are a random sample without replacement from a finite population of size N, then all the X_i have the same variance σ^2 and*

$$\operatorname{var}\left(\sum_{i=1}^{n} X_i\right) = n\sigma^2 \cdot \frac{N-n}{N-1} \tag{7.13}$$

The factor $(N-n)/(N-1)$ by which (7.13) differs from (7.11) is called the *finite population correction.*

Proof. Since the X_i are exchangeable, each X_i has the same variance σ^2 and each pair X_i and X_j has the same correlation ρ. Thus

$$\begin{aligned}\operatorname{var}\left(\sum_{i=1}^{n} X_i\right) &= \sum_{i=1}^{n}\sum_{j=1}^{n} \operatorname{cov}(X_i, X_j) \\ &= n\sigma^2 + n(n-1)\sigma^2\rho \\ &= n\sigma^2\left[1 + (n-1)\rho\right]\end{aligned} \tag{7.14}$$

The correlation ρ does not depend on the sample size, because by exchangeability it is the correlation of X_1 and X_2, and the marginal distribution of these two individuals does not depend on what happens after they are drawn. Therefore (a trick!) we can determine ρ by looking at the special case when $N = n$, when the sample is the whole population and

$$\sum_{i=1}^{n} X_i = \sum_{i=1}^{N} x_i$$

is not random (as is clear from the "little x" notation on the right hand side). Hence when $N = n$ the variance is zero, and we must have

$$1 + (N - 1)\rho = 0$$

which, solving for ρ, implies

$$\rho = -\frac{1}{N - 1}$$

Plugging this into (7.14) gives (7.13). □

7.2.2 Repeated Experiments

If X_1, ..., X_n are the outcomes of a series of random experiments which are absolutely identical and have nothing to do with each other, then they are independent and identically distributed, a phrase so widely used in statistics that its abbreviation i. i. d. is universally recognized.

This situation is analogous to sampling with replacement in that the variables of interest are i. i. d. and all the consequences of the i. i. d. property, such as (7.11) and (7.12), hold. The situation is so analogous that many people use the language of random sampling to describe this situation too. Saying that X_1, ..., X_n are a random sample from a hypothetical infinite population. There is nothing wrong with this so long as everyone understands it is only an analogy. There is no sense in which i. i. d. random variables actually are a random sample from some population.

We will use the same language. It lends color to otherwise dry and dusty discussions if you imagine we are sampling a population to answer some interesting question. That may lead us into some language a pedant would call sloppy, such as, "suppose we have a sample of size n from a population with finite variance." If the population is finite, then it automatically has a finite variance. If the population is infinite, then the variance is not really defined, since infinite populations don't exist except as a vague analogy. What is meant, of course, is "suppose X_1, ..., X_n are i. i. d. and have finite variance." That's well defined.

7.3 Sampling Distributions of Sample Moments

7.3.1 Sample Moments

If X_1, ..., X_n are a random sample, the *sample moments* are the moments of the empirical distribution associated with the vector $\mathbf{X} = (X_1, \ldots, X_n)$. The first moment is the *sample mean*

$$\overline{X}_n = E_n(X) = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{7.15}$$

The k-th moment is

$$A_{k,n} = E_n(X^k) = \frac{1}{n} \sum_{i=1}^n X_i^k.$$

The central moments of this empirical distribution are

$$M_{k,n} = E_n\bigl\{[X - E_n(X)]^k\bigr\} = E_n\bigl\{(X - \overline{X}_n)^k\bigr\} = \frac{1}{n} \sum_{i=1}^n (X_i - \overline{X}_n)^k$$

As with any distribution, the first central moment is zero, and the second is

$$V_n = \operatorname{var}_n(X) = E_n\bigl\{(X - \overline{X}_n)^2\bigr\} = \frac{1}{n} \sum_{i=1}^n (X_i - \overline{X}_n)^2. \tag{7.16}$$

If there were any logic to statistics V_n would be called the "sample variance," but Lindgren, agreeing with most other textbooks, uses that term for something slightly different

$$S_n^2 = \frac{n}{n-1} V_n = \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X}_n)^2. \tag{7.17}$$

The $n-1$ rather than n in the definition makes all of the formulas involving S_n^2 ugly, and makes S_n^2 not satisfy any of the usual rules involving variances. So be warned, and be careful! For example, V_n obeys the parallel axis theorem, hence

$$V_n = E_n(X^2) - E_n(X)^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \overline{X}_n^2.$$

Clearly S_n^2 cannot satisfy the same rule or it would be V_n. The only way to figure out the analogous rule for S_n^2 is to remember the rule for V_n (which makes sense) and derive the one for S_n^2.

$$\begin{aligned}
S_n^2 &= \frac{n}{n-1} V_n \\
&= \frac{n}{n-1} \left[\frac{1}{n} \sum_{i=1}^n X_i^2 - \overline{X}_n^2 \right] \\
&= \frac{1}{n-1} \sum_{i=1}^n X_i^2 - \frac{n}{n-1} \overline{X}_n^2
\end{aligned}$$

No matter how you try to write it, it involves both n and $n-1$, and makes no sense.

Since S_n^2 is so ugly, why does anyone use it? The answer, as with so many other things, is circular. Almost everyone uses it because it's the standard, and it's the standard because almost everyone uses it. And "almost everyone" includes a lot of people, because S_n^2 is a topic in most introductory statistics courses.

Our position is that it simply does not matter whether you use V_n or S_n^2. Since one is a constant times the other, any place you could use one, you could use the other, so long as you make the appropriate changes in formulas. So the only reason for using S_n^2 is to avoid fighting tradition. Sometimes it's easier to follow the herd.

7.3.2 Sampling Distributions

Since a sample moment is a random variable, it has a probability distribution. We may not be able to give a formula for the density or distribution function, but it does have a distribution. So we can talk about its distribution and investigate its properties.

In a few specific cases we know the distribution of $\overline{X}_n$. It is given implicitly by what we call "addition rules" and which are summarized in Appendix C of these notes. They give the distribution of $Y = \sum_i X_i$ when the X_i are i. i. d.

- Binomial (including Bernoulli)
- Negative Binomial (including Geometric)
- Poisson
- Gamma (including Exponential and Chi-Square)
- Normal
- Cauchy

Given the distribution of Y, the distribution of $\overline{X}_n$ is found by a simple change of scale. If the X_i are continuous random variables, then

$$f_{\overline{X}_n}(z) = n f_Y(nz). \tag{7.18}$$

Example 7.3.1 (I. I. D. Exponential).
Let X_1, $\ldots$, X_n be i. i. d. $\text{Exp}(\lambda)$. Then the distribution of $Y = X_1 + \cdots + X_n$ is $\text{Gam}(n, \lambda)$ by the addition rule for Gamma distributions (Appendix C) and the fact that the $\text{Exp}(\lambda)$ is $\text{Gam}(1, \lambda)$. Hence by Problem 7-10

$$\overline{X}_n \sim \text{Gam}(n, n\lambda).$$

Many statistics textbooks, including Lindgren, have no tables of the gamma distribution. Thus we have to use the fact that gamma random variables having integer and half-integer values of their shape parameters are proportional to chi-square random variables, because $\text{chi}^2(n) = \text{Gam}(\frac{n}{2}, \frac{1}{2})$ and the second parameter of the gamma distribution is a shape parameter (Problem 7-10).

Lemma 7.10. *Suppose*

$$X \sim \text{Gam}(n, \lambda)$$

where n is an integer, then

$$2\lambda X \sim \text{chi}^2(2n).$$

The proof is Exercise 7-2.

Example 7.3.2 (Table Look-Up).
(Continues Example 7.3.1). Using the lemma, we can calculate probabilities for the sampling distribution of the sample mean of i. i. d. Exp(λ) data. Suppose $\lambda = 6.25$ so $\mu = 1/\lambda = 0.16$, and $n = 9$. What is $P(\overline{X}_n > 0.24)$.

In Example 7.3.1 we figured out that

$$\overline{X}_n \sim \text{Gam}(n, n\lambda)$$

so in this case

$$\overline{X}_n \sim \text{Gam}(9, 56.25) \tag{7.19}$$

($n\lambda = 9 \times 6.25 = 56.25$).

But to use the tables in Lindgren, we must use the lemma, which says

$$2n\lambda\overline{X}_n \sim \text{chi}^2(2n).$$

(there is an n on the left hand side, because the scale parameter of the gamma distribution is $n\lambda$ here rather than λ).

If $\overline{X}_n = 0.24$, then $2n\lambda\overline{X}_n = 2 \cdot 9 \cdot 6.25 \cdot 0.24 = 27.0$, and the answer to our problem is $P(Y > 27.0)$, where $Y \sim \text{chi}^2(18)$. Looking this up in Table Va in Lindgren, we get 0.079 for the answer.

Example 7.3.3 (Table Look-Up using Computers).
(Continues Example 7.3.1). The tables in Lindgren, or in other statistics books are not adequate for many problems. For many problems you need either a huge book of tables, commonly found in the reference section of a math library, or a computer.

Many mathematics and statistics computer software packages do calculations about probability distributions. In this course, we will only describe two of them: the statistical computing language R and the symbolic mathematics language Mathematica.

R In R the lookup is very simple. It uses the function `pgamma` which evaluates the gamma c. d. f.

```
> 1 - pgamma(0.24, 9, 1 / 56.25)
[1] 0.07899549
```

This statement evaluates $P(X \le x)$ when $X \sim \text{Gam}(9, 56.25)$ and $x = 0.24$, as (7.19) requires. We don't have to use the property that this gamma distribution is also a chi-square distribution. One caution: both R and Mathematica use a different parameterization of the gamma distribution than Lindgren. The shape parameter is the same, but the scale parameter is the reciprocal of Lindgren's scale parameter (See Problem 7-10). That's why the third argument of the `pgamma` function is 1/56.25 rather than 56.25.

Mathematica Mathematica makes things a bit more complicated. First you have to load a special package for probability distributions (always available, but not loaded by default), then you have to tell Mathematica which distribution you want, then you do the calculation

```
In[1]:= <<Statistics‘ContinuousDistributions‘

In[2]:= dist = GammaDistribution[9, 1 / 56.25]

Out[2]= GammaDistribution[9, 0.0177778]

In[3]:= F[x_] = CDF[dist, x]

Out[3]= GammaRegularized[9, 0, 56.25 x]

In[4]:= 1 - F[0.24]

Out[4]= 0.0789955
```

of course, the last three statements can be combined into one but just plugging in definitions

```
In[5]:= 1 - CDF[GammaDistribution[9, 1 / 56.25], 0.24]

Out[5]= 0.0789955
```

but that's a cluttered and obscure. For more on computing in general see the course computing web page

```
http://www.stat.umn.edu/geyer/5101/compute
```

and the pages on *Probability Distributions* in *R* and *Mathematica* in particular (follow the links from the main computing page).

Example 7.3.4 (I. I. D. Bernoulli).
If X_1, ..., X_n are i. i. d. Ber(p) random variables, then $Y = \sum_i X_i$ is a Bin(n, p) random variable, and since $\overline{X}_n = Y/n$, we also have

$$n\overline{X}_n \sim \text{Bin}(n, p).$$

Example 7.3.5 (Another Computer Table Look-Up).
(Continues Example 7.3.4). Suppose $\overline{X}_n$ is the sample mean of 10 i. i. d. Ber(0.2) random variables. What is the probability $P(\overline{X}_n \le 0.1)$?

By the preceding example, $n\overline{X}_n \sim \text{Bin}(10, 0.2)$ and here $n\overline{X}_n = 10 \cdot 0.1 = 1$. So we need to look up $P(Y \le 1)$ when $Y \sim \text{Bin}(10, 0.2)$. In R this is

```
> pbinom(1, 10, 0.2)
[1] 0.3758096
```

In Mathematica it is

```
In[1]:= <<Statistics‘DiscreteDistributions‘

In[2]:= dist = BinomialDistribution[10, 0.2]

Out[2]= BinomialDistribution[10, 0.2]

In[3]:= F[x_] = CDF[dist, x]

Out[3]= BetaRegularized[0.8, 10 - Floor[x], 1 + Floor[x]]

In[4]:= F[1]

Out[4]= 0.37581
```

Our textbook has no tables of the binomial distribution, so there is no way to do this problem with pencil and paper except by evaluating the terms

$$\binom{n}{0} p^0 q^n + \binom{n}{1} p^1 q^{n-1}$$

(not so hard here, but very messy if there are many terms). You can't use the normal approximation because n is not large enough. Anyway, why use an approximation when the computer gives you the exact answer?

We can calculate the density using the convolution theorem. Mathematical induction applied to the convolution formula (Theorem 23 of Chapter 4 in Lindgren) gives the following result.

Theorem 7.11. *If X_1, ..., X_n are i. i. d. continuous random variables with common marginal density f_X, then $Y = X_1 + \cdots + X_n$ has density*

$$f_Y(y) = \int \cdots \iint f_X(y - x_2 - \cdots - x_n) f_X(x_2) \cdots f_X(x_n) \, dx_2 \cdots dx_n \quad (7.20)$$

Then (7.18) gives the density of $\overline{X}_n$. But this is no help if we can't do the integrals, which we usually can't, with the notable exceptions of the "brand name" distributions with "addition rules" (Appendix C).

Higher Moments So far we haven't considered any sample moment except $\overline{X}_n$. For other sample moments, the situation is even more complicated.

It is a sad fact is that the methods discussed in this section don't always work. In fact they usually don't work. Usually, nothing works, and you just can't find a closed form expression for the sampling distribution of a particular sample moment.

What is important to understand, though, and understand clearly, is that every sample moment does *have* a sampling distribution. Hence we can talk about properties of that distribution. The properties exist in principle, so we can talk about them whether or not we can calculate them.

7.3.3 Moments

In this section we calculate moments of sample moments. At first this sounds confusing, even bizarre, but sample moments are random variables and like any random variables they have moments.

Theorem 7.12. *If X_1, …, X_n are identically distributed random variables with mean μ and variance σ^2, then*

$$E(\overline{X}_n) = \mu. \tag{7.21a}$$

If in addition, they are uncorrelated, then

$$\operatorname{var}(\overline{X}_n) = \frac{\sigma^2}{n}. \tag{7.21b}$$

If instead they are samples without replacement from a population of size N, then

$$\operatorname{var}(\overline{X}_n) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}. \tag{7.21c}$$

Note in particular, that because independence implies lack of correlation, (7.21a) and (7.21b) hold in the i. i. d. case.

Proof. By the usual rules for linear transformations, $E(a + bX) = a + bE(X)$ and $\operatorname{var}(a + bX) = b^2 \operatorname{var}(X)$

$$E(\overline{X}_n) = \frac{1}{n} E\left(\sum_{i=1}^n X_i\right)$$

and

$$\operatorname{var}(\overline{X}_n) = \frac{1}{n^2} \operatorname{var}\left(\sum_{i=1}^n X_i\right)$$

Now apply Corollary 1 of Theorem 9 of Chapter 4 in Lindgren and (7.11) and (7.13). □

Theorem 7.13. *If X_1, …, X_n are uncorrelated, identically distributed random variables with variance σ^2, then*

$$E(V_n) = \frac{n-1}{n}\sigma^2, \tag{7.22a}$$

and

$$E(S_n^2) = \sigma^2. \tag{7.22b}$$

Proof. The reason why (7.22a) doesn't work out simply is that V_n involves deviations from the sample mean $\overline{X}_n$ and σ^2 involves deviations from the population mean μ. So use the empirical parallel axis theorem to rewrite V_n in terms of deviations from μ

$$E_n\{(X - \mu)^2\} = V_n + (\overline{X}_n - \mu)^2. \tag{7.23}$$

The left hand side is just $\overline{Y}_n$, where $Y_i = (X_i - \mu)^2$. Taking expectations of both sides of (7.23) gives

$$E(\overline{Y}_n) = E(V_n) + E\{(\overline{X}_n - \mu)^2\}$$

On the left hand side we have

$$E(\overline{Y}_n) = E(Y_i) = \operatorname{var}(X_i) = \sigma^2$$

And the second term on the right hand side is

$$\operatorname{var}(\overline{X}_n) = \frac{\sigma^2}{n}.$$

Collecting terms gives (7.22a). Then linearity of expectation gives (7.22b). □

The assertions (7.22a) and (7.22b) of this theorem are one place where S_n^2 seems simpler than V_n. It's why S_n^2 was invented, to make (7.22b) simple.

The sample moment formulas (7.21a), (7.21b), and (7.22b) are the ones most commonly used in everyday statistics. Moments of other sample moments exist but are mostly of theoretical interest.

Theorem 7.14. *If X_1, ..., X_n are i. i. d. random variables having moments of order k, then all sample moments of order k have expectation. If the X_i have moments of order $2k$, then sample moments of order k have finite variance. In particular,*

$$E(A_{k,n}) = \alpha_k$$

and

$$\operatorname{var}(A_{k,n}) = \frac{\alpha_{2k} - \alpha_k^2}{n},$$

where α_k is the k-th population moment.

We do not give formulas for the central moments because they are a mess. Even the formula for the variance of the sample variance given (though not proved) in Theorem 7 of Chapter 7 in Lindgren is already a mess. The formulas for higher moments are worse. They are, however, a straightforward mess. The proof below shows how the calculation would start. Continuing the calculation without making any mistakes would produce an explicit formula (a symbolic mathematics computer package like Maple or Mathematica would help a lot).

Proof. The k-th sample moment $A_{k,n}$ is the sample average of the random variables $Y_i = X_i^k$. Since

$$E(Y_i) = E(X_i^k) = \alpha_k \tag{7.24a}$$

and

$$\begin{aligned} \operatorname{var}(Y_i) &= E(Y_i^2) - E(Y_i)^2 \\ &= E(X_i^{2k}) - E(X_i^k)^2 \\ &= \alpha_{2k} - \alpha_k^2 \end{aligned} \tag{7.24b}$$

the formulas in the theorem follow by the usual rules for the moments of a sample mean.

The k-th central sample moment

$$\begin{aligned} M_{k,n} &= \frac{1}{n} \sum_{i=1}^n \bigl(X_i - \overline{X}_n\bigr)^k \\ &= \frac{1}{n} \sum_{i=1}^n \left(\frac{n-1}{n} X_i - \sum_{j \neq i} \frac{1}{n} X_j \right)^k \end{aligned}$$

is a k-th degree polynomial in the X_i. A single term of such a polynomial has the form

$$a \prod_{i=1}^n X_i^{m_i}$$

where the m_i are nonnegative integers such that $m_1 + \cdots + m_n = k$, and a is some constant (a different constant for each term of the polynomial, although the notation doesn't indicate that). By independence

$$E\left(a \prod_{i=1}^n X_i^{m_i} \right) = a \prod_{i=1}^n E(X_i^{m_i}) = a \prod_{i=1}^n \alpha_{m_i}. \tag{7.25}$$

If k-th moments exist, then all of the moments α_{m_i} in (7.25) exist because $m_i \le k$.

Similarly, $M_{k,n}^2$ is a polynomial of degree $2k$ in the X_i and hence has expectation if population moments of order $2k$ exist. Then $\operatorname{var}(M_{k,n}) = E(M_{k,n}^2) - E(M_{k,n})^2$ also exists. □

7.3.4 Asymptotic Distributions

Often we cannot calculate the exact sampling distribution of a sample moment, but we can always get large sample properties of the distribution from law of large numbers, the central limit theorem, and Slutsky's theorem.

Theorem 7.15. *Under i. i. d. sampling every sample moment converges in probability to the corresponding population moment provided the population moment exists.*

Proof. For ordinary moments, this was done as a homework problem (Problem 5-3 in Lindgren). If we let α_k be the k-th ordinary population moment and $A_{k,n}$ be the corresponding ordinary sample moment for sample size n, then

$$A_{k,n} = \frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{P} E(X_1^k) = \alpha_k.$$

Let μ_k be the k-th population central moment and $M_{k,n}$ be the corresponding sample central moment, then

$$\begin{aligned} M_{k,n} &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^k &\quad (7.26a)\\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{k}\binom{k}{j}(-1)^j(\overline{X}_n - \mu)^j(X_i - \mu)^{k-j} \\ &= \sum_{j=0}^{k}\binom{k}{j}(-1)^j(\overline{X}_n - \mu)^j\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^{k-j} \\ &= \sum_{j=0}^{k}\binom{k}{j}(-1)^j(\overline{X}_n - \mu)^j M'_{k-j,n} &\quad (7.26b) \end{aligned}$$

where we have introduced the notation

$$M'_{k,n} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^k.$$

This is almost the same as (7.26a), the only difference being the replacement of $\overline{X}_n$ by μ. The asymptotics of $M'_{k,n}$ are much simpler than those for $M_{k,n}$ because $M'_{k,n}$ is the sum of i. i. d. terms so the LLN and CLT apply directly to it. In particular

$$M'_{k,n} \xrightarrow{P} E\{(X_i - \mu)^k\} = \mu_k \qquad (7.27)$$

also

$$\overline{X}_n - \mu \xrightarrow{P} 0 \qquad (7.28)$$

by the LLN and the continuous mapping theorem. Then (7.28) and Slutsky's theorem imply that every term of (7.26b) converges in probability to zero except the $j = 0$ term, which is $M'_{k,n}$. Thus (7.27) establishes

$$M_{k,n} \xrightarrow{P} \mu_k \qquad (7.29)$$

which is what was to be proved. □

Theorem 7.16. *Under i. i. d. sampling every sample k-th moment is asymptotically normal if population moments of order $2k$ exist. In particular,*

$$\sqrt{n}(A_{k,n} - \alpha_k) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \alpha_{2k} - \alpha_k^2) \qquad (7.30)$$

and

$$\sqrt{n}(M_{k,n} - \mu_k) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mu_{2k} - \mu_k^2 - 2k\mu_{k-1}\mu_{k+1} + k^2\mu_2\mu_{k-1}^2) \tag{7.31}$$

For ordinary moments, this is a homework problem (Problem 7-17 in Lindgren). For central moments, the proof will have to wait until we have developed multivariate convergence in distribution in the following chapter.

The special case $k = 2$ is worth noting.

Corollary 7.17. *Suppose X_1, X_2, ... are i. i. d. and have fourth moments. Then*

$$\sqrt{n}(V_n - \sigma^2) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mu_4 - \mu_2^2)$$

where V_n is defined by (7.16)*.*

This is the case $V_n = M_{2,n}$ of the theorem. The third and forth terms of the asymptotic variance formula are zero because $\mu_1 = 0$ (Theorem 2.9 in Chapter 2 of these notes).

Example 7.3.6 (I. I. D. Normal).
Suppose X_1, ..., X_n are i. i. d. $\mathcal{N}(\mu, \sigma^2)$. What is the asymptotic distribution of $\overline{X}_n$, of V_n, of $M_{3,n}$?

The CLT, of course, tells us the asymptotic distribution of $\overline{X}_n$. Here we just want to check that the $k = 1$ case of (7.30) agrees with the CLT. Note that $A_{1,n} = \overline{X}_n$ and $\alpha_1 = \mu$, so the left hand side of (7.30) is the same as the left hand side of the CLT (6.7). Also $\alpha_2 - \alpha_1^2 = \sigma^2$ because this is just $\text{var}(X) = E(X^2) - E(X)^2$ in different notation. So the $k = 1$ case of (7.30) does agree with the CLT.

The asymptotic distribution of $V_n = M_{2,n}$ is given by the $k = 2$ case of (7.31) or by Theorem 7.17. All we need to do is calculate the asymptotic variance $\mu_4 - \mu_2^2$. The fourth central moment of the standard normal distribution is given by the $k = 2$ case of equation (5) on p. 178 in Lindgren to be $\mu_4 = 3$. A general normal random variable has the form $X = \mu + \sigma Z$, where Z is standard normal, and this has fourth central moment $3\sigma^4$ by Problem 7-11. Thus $\mu_4 - \mu_2^2 = 3\sigma^4 - \sigma^4 = 2\sigma^4$, and finally we get

$$V_n \approx \mathcal{N}\left(\sigma^2, \frac{2\sigma^4}{n}\right)$$

Note this formula holds for i. i. d. normal data *only*. Other statistical models can have rather different distributions (Problem 7-12).

The asymptotic distribution of $M_{3,n}$ is given by the $k = 3$ case of (7.31)

$$\begin{aligned}\mu_6 - \mu_3^2 - 2\cdot 3\mu_2\mu_4 + 3^2\mu_2\cdot\mu_2^2 &= \mu_6 - \mu_3^2 - 6\mu_2\mu_4 + 9\mu_2^3 \\ &= \mu_6 - 6\mu_2\mu_4 + 9\mu_2^3\end{aligned}$$

because odd central moments are zero (Theorem 2.10 of Chapter 2 of these notes). We already know $\mu_2 = \sigma^2$ and $\mu_4 = 3\sigma^2$. Now we need to use the

$k = 3$ case of equation (5) on p. 178 in Lindgren and Problem 7-11 to get to be $\mu_6 = 15\sigma^2$. Hence the asymptotic variance is

$$\mu_6 - 6\mu_2\mu_4 + 9\mu_2^3 = (15 - 6 \cdot 1 \cdot 3 + 9)\sigma^6 = 6\sigma^6$$

and

$$M_{3,n} \approx \mathcal{N}\left(0, \frac{6\sigma^6}{n}\right)$$

(the asymptotic mean is $\mu_3 = 0$).

7.3.5 The t Distribution

We now derive two other "brand name" distributions that arise as exact sampling distributions of statistics derived from sampling normal populations. The distributions are called the t and F distributions (whoever thought up those names must have had a real imagination!)

Before we get to them, we want to generalize the notion of degrees of freedom to noninteger values. This will be useful when we come to Bayesian inference.

Definition 7.3.1 (Chi-Square Distribution).
The chi-square with noninteger degrees of freedom $\nu > 0$ is the $\text{Gam}(\frac{\nu}{2}, \frac{1}{2})$ *distribution.*

This agrees with our previous definition when ν is an integer.

Definition 7.3.2 (Student's t Distribution).
If Z and Y are independent random variables, Z is standard normal and Y is $\text{chi}^2(\nu)$*, then the random variable*

$$T = \frac{Z}{\sqrt{Y/\nu}}$$

is said to have a t-distribution with ν degrees of freedom, *abbreviated $t(\nu)$. The parameter ν can be any strictly positive real number.*

The reason for the "Student" sometimes attached to the name of the distribution is that the distribution was discovered and published by W. S. Gosset, the chief statistician for the Guiness brewery in Ireland. The brewery had a company policy that employees were not allowed to publish under their own names, so Gosset used the pseudonym "Student" and this pseudonym is still attached to the distribution by those who like eponyms.

Theorem 7.18. *The p. d. f. of the $t(\nu)$ distribution is*

$$f_\nu(x) = \frac{1}{\sqrt{\nu\pi}} \cdot \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \cdot \frac{1}{\left(1 + \frac{x^2}{\nu}\right)^{(\nu+1)/2}}, \qquad -\infty < x < +\infty \tag{7.32}$$

The normalizing constant can also be written using a beta function because $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. Thus

$$\frac{1}{\sqrt{\nu\pi}} \cdot \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} = \frac{1}{\sqrt{\nu}} \cdot \frac{1}{B(\frac{\nu}{2}, \frac{1}{2})}$$

The connection with the beta distribution is obscure but will be clear after we finish this section and do Problem 7-3.

Proof. The joint distribution of Z and Y in the definition is

$$f(z, y) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \frac{\left(\frac{1}{2}\right)^{\nu/2}}{\Gamma(\nu/2)} y^{\nu/2-1} e^{-y/2}$$

Make the change of variables $t = z/\sqrt{y/\nu}$ and $u = y$, which has inverse transformation

$$\begin{aligned} z &= t\sqrt{u/\nu} \\ y &= u \end{aligned}$$

and Jacobian

$$\begin{vmatrix} \sqrt{u/\nu} & t/2\sqrt{u\nu} \\ 0 & 1 \end{vmatrix} = \sqrt{u/\nu}$$

Thus the joint distribution of T and U given by the multivariate change of variable formula is

$$\begin{aligned} f(t, u) &= \frac{1}{\sqrt{2\pi}} e^{-(t\sqrt{u/\nu})^2/2} \frac{\left(\frac{1}{2}\right)^{\nu/2}}{\Gamma(\nu/2)} u^{\nu/2-1} e^{-u/2} \cdot \sqrt{u/\nu} \\ &= \frac{1}{\sqrt{2\pi}} \frac{\left(\frac{1}{2}\right)^{\nu/2}}{\Gamma(\nu/2)} \frac{1}{\sqrt{\nu}} u^{\nu/2-1/2} \exp\left\{-\left(1 + \frac{t^2}{\nu}\right)\frac{u}{2}\right\} \end{aligned}$$

Thought of as a function of u for fixed t, this is proportional to a gamma density with shape parameter $(\nu + 1)/2$ and inverse scale parameter $\frac{1}{2}(1 + \frac{t^2}{\nu})$. Hence we can use the "recognize the unnormalized density trick" (Section 2.5.7 in Chapter 2 of these notes) to integrate out u getting the marginal of t

$$f(t) = \frac{1}{\sqrt{2\pi}} \cdot \frac{\left(\frac{1}{2}\right)^{\nu/2}}{\Gamma(\nu/2)} \cdot \frac{1}{\sqrt{\nu}} \cdot \frac{\Gamma(\frac{\nu+1}{2})}{[\frac{1}{2}(1 + \frac{t^2}{\nu})]^{(\nu+1)/2}}$$

which, after changing t to x, simplifies to (7.32). □

The formula for the density of the t distribution shows that it is symmetric about zero. Hence the median is zero, and the mean is also zero when it exists. In fact, all odd central moments are zero when they exist, because this is true of any symmetric random variable (Theorem 2.10 of Chapter 2 of these notes).

The question of when moments exist is settled by the following theorem.

Theorem 7.19. *If X has a Student t distribution with ν degrees of freedom, then moments of order k exist if and only if $k < \nu$.*

Proof. The density (7.32) is clearly bounded. Hence we only need to check whether $|x|^k f(x)$ is integrable near infinity. Since the density is symmetric, we only need to check one tail. For x near $+\infty$

$$|x|^k f(x) \approx k x^{k-(\nu+1)}$$

for some constant k. From Lemma 2.39 of Chapter 2 of these notes the integral is finite if and only if $k - (\nu + 1) < -1$, which the same as $\nu > k$. □

We also want to know the variance of the t distribution.

Theorem 7.20. *If $\nu > 2$ and $X \sim t(\nu)$, then*

$$\operatorname{var}(X) = \frac{\nu}{\nu - 2}.$$

The proof is a homework problem (7-5).

Another important property of the t distribution is given in the following theorem, which we state without proof since it involves the Stirling approximation for the gamma function, which we have not developed, although we will prove a weaker form of the second statement of the theorem in the next chapter after we have developed some more tools.

Theorem 7.21. *For every $x \in \mathbb{R}$*

$$f_\nu(x) \to \phi(x), \qquad \text{as } \nu \to \infty,$$

where ϕ is the standard normal density, and

$$t(\nu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1), \qquad \text{as } \nu \to \infty.$$

Comparison of the $t(1)$ density to the standard Cauchy density given by equation (1) on p. 191 in Lindgren shows they are the same (it is obvious that the part depending on x is the same, hence the normalizing constants must be the same if both integrate to one, but in fact we already know that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ also shows the normalizing constants are equal). Thus $t(1)$ is another name for the standard Cauchy distribution. The theorem above says we can think of $t(\infty)$ as another name for the standard normal distribution. Tables of the t distribution, including Tables IIIa and IIIb in the Appendix of Lindgren include the normal distribution labeled as ∞ degrees of freedom. Thus the t family of distributions provides lots of examples between the best behaved distribution of those we've studied, which is the normal, and the worst behaved, which is the Cauchy. In particular, the $t(2)$ distribution has a mean but no variance, hence the sample mean of i. i. d. $t(2)$ random variables obeys the LLN but not the CLT. For $\nu > 2$, The $t(\nu)$ distribution has both mean and variance, hence the sample mean of i. i. d. $t(\nu)$ random variables obeys both LLN and CLT, but the $t(\nu)$ distribution is much more heavy-tailed than other distributions we have previously considered.

7.3.6 The F Distribution

The letter F for the random variable having the "F distribution" was chosen by Snedecor in honor of R. A. Fisher who more or less invented the F distribution. Actually, he proposed a monotone transformation of this variable $Z = \frac{1}{2} \log F$, which has a better normal approximation.

Definition 7.3.3 (The F Distribution).
If Y_1 and Y_2 are independent random variables, and $Y_i \sim \text{chi}^2(\nu_i)$, then the random variable

$$U = \frac{Y_1/\nu_1}{Y_2/\nu_2}$$

has an F distribution with ν_1 numerator degrees of freedom and ν_2 denominator degrees of freedom, abbreviated $F(\nu_1, \nu_2)$.

Theorem 7.22. *If Y_1 and Y_2 are independent random variables, and $Y_i \sim \text{chi}^2(\nu_i)$, then the random variable*

$$W = \frac{Y_1}{Y_1 + Y_2}$$

has a $\text{Beta}(\frac{\nu_1}{2}, \frac{\nu_2}{2})$ *distribution.*

Proof. Since we know that the chi-square distribution is a special case of the gamma distribution $\text{chi}^2(k) = \text{Gam}(\frac{k}{2}, \frac{1}{2})$, this is one of the conclusions of Theorem 4.2 of Chapter 4 of these notes. □

Corollary 7.23. *If $U \sim F(\nu_1, \nu_2)$, then*

$$W = \frac{\frac{\nu_1}{\nu_2} U}{1 + \frac{\nu_1}{\nu_2} U}$$

has a $\text{Beta}(\frac{\nu_1}{2}, \frac{\nu_2}{2})$ *distribution.*

Hence the F distribution is not really new, it is just a transformed beta distribution. The only reason for defining the F distribution is convention. Tables of the F distribution are common. There is one in the appendix of Lindgren. Tables of the beta distribution are rare. So we mostly use F tables rather than beta tables. When using a computer, the distinction doesn't matter. Mathematica and R have functions that evaluate either F or beta probabilities.

7.3.7 Sampling Distributions Related to the Normal

When the data are i. i. d. normal, the exact (not asymptotic) sampling distributions are known for many quantities of interest.

Theorem 7.24. *If X_1, …, X_n are i. i. d. $\mathcal{N}(\mu, \sigma^2)$, then $\overline{X}_n$ and S_n^2 given by* (7.15) *and* (7.17) *are independent random variables and*

$$\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \tag{7.33a}$$

$$(n-1)S_n^2/\sigma^2 \sim \text{chi}^2(n-1) \tag{7.33b}$$

This is a combination of Theorems 9, 10, and 11 and the Corollary to Theorem 10 in Section 7.5 of Lindgren.

Note that the theorem implicitly gives the distribution of S_n^2, since $\text{chi}^2(n-1)$ is just another name for $\text{Gam}(\frac{n-1}{2}, \frac{1}{2})$ and the second parameter of the gamma is an upside down scale parameter, which implies

$$S_n^2 \sim \text{Gam}\left(\frac{n-1}{2}, \frac{n-1}{2\sigma^2}\right) \tag{7.34}$$

The theorem is stated the way it is because chi-square tables are widely available (including in the Appendix of Lindgren) and gamma tables are not. Hence (7.33b) is a more useful description of the sampling distribution of S_n^2 than is (7.34) when you are using tables (if you are using a computer, either works).

The main importance of the t distribution in statistics comes from the following corollary.

Corollary 7.25. *If X_1, ..., X_n are i. i. d. $\mathcal{N}(\mu, \sigma^2)$, then*

$$T = \frac{\overline{X}_n - \mu}{S_n / \sqrt{n}}$$

has a $t(n-1)$ distribution.

Proof.

$$Z = \frac{\overline{X}_n - \mu}{\sigma / \sqrt{n}}$$

is standard normal, and independent of $Y = (n-1)S_n^2/\sigma^2$ which is $\text{chi}^2(n-1)$ by Theorem 7.24. Then $Z/\sqrt{Y/(n-1)}$ is T. □

One use of the F distribution in statistics (not the most important) comes from the following corollary.

Corollary 7.26. *If X_1, ..., X_m are i. i. d. $\mathcal{N}(\mu_X, \sigma_X^2)$ and Y_1, ..., Y_n are i. i. d. $\mathcal{N}(\mu_Y, \sigma_Y^2)$, and all of the X_i are independent of all of the Y_j, then*

$$F = \frac{S_{m,X}^2}{S_{n,Y}^2} \cdot \frac{\sigma_Y^2}{\sigma_X^2}$$

has an $F(m-1, n-1)$ distribution, where $S_{m,X}^2$ is the sample variance of the X_i and $S_{n,Y}^2$ is the sample variance of the Y_i.

The proof is obvious from Theorem 7.24 and the definition of the F distribution.

Example 7.3.7 (T Distribution).
Suppose X_1, ..., X_{20} are i. i. d. standard normal. Compare $P(\overline{X}_n > \sigma/\sqrt{n})$

and $P(\overline{X}_n > S_n/\sqrt{n})$. We know that

$$\frac{\overline{X}_n}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$
$$\frac{\overline{X}_n}{S_n/\sqrt{n}} \sim t(19)$$

So we need to compare $P(Z > 1)$ where Z is standard normal and $P(T > 1)$ where $T \sim t(19)$.

From Tables I and IIIa in Lindgren, these probabilities are .1587 and .165, respectively. The following R commands do the same lookup

```
> 1 - pnorm(1)
[1] 0.1586553
> 1 - pt(1, 19)
[1] 0.1649384
```

Example 7.3.8 (F Distribution).
Suppose S_1^2 and S_2^2 are sample variances of two independent samples from two normal populations with equal variances, and the sample sizes are $n_1 = 10$ and $n_2 = 20$, respectively. What is $P(S_1^2 > 2S_2^2)$? We know that

$$\frac{S_1^2}{S_2^2} \sim F(9, 19)$$

So the answer is $P(Y > 2)$ where $Y \sim F(9, 19)$. Tables IVa and IVb in Lindgren (his only tables of the F distribution) are useless for this problem. We must use the computer. In R it's simple

```
> 1 - pf(2, 9, 19)
[1] 0.0974132
```

For this example, we also show how to do it in Mathematica

```
In[1]:= <<Statistics`ContinuousDistributions`

In[2]:= dist = FRatioDistribution[9, 19]

Out[2]= FRatioDistribution[9, 19]

In[3]:= F[x_] = CDF[dist, x]

                             19         19  9
Out[3]= BetaRegularized[--------, 1, --, -]
                         19 + 9 x       2   2

In[4]:= 1 - F[2]
```

```
                               19      19  9
Out[4]= 1 - BetaRegularized[--, 1, --, -]
                               37      2   2
```

```
In[5]:= N[%]

Out[5]= 0.0974132
```

(The last command tells Mathematica to evaluate the immediately preceding expression giving a numerical result). This can be done more concisely if less intelligibly as

```
In[6]:= N[1 - CDF[FRatioDistribution[9, 19], 2]]

Out[6]= 0.0974132
```

7.4 Sampling Distributions of Sample Quantiles

The *sample quantiles* are the quantiles of the empirical distribution associated with the data vector $\mathbf{X} = (X_1, \ldots, X_n)$. They are mostly of interest only for continuous population distributions. A sample quantile can always be taken to be an order statistic by Theorem 7.5. Hence the exact sampling distributions of the empirical quantiles are given by the exact sampling distributions for order statistics, which are given by equation (5) on p. 217 of Lindgren

$$f_{X_{(k)}}(y) = \frac{n!}{(k-1)!(n-k)!} F(y)^{k-1} [1 - F(y)]^{n-k} f(y) \tag{7.35}$$

when the population distribution is continuous, (where, as usual, F is the c. d. f. of the X_i and f is their p. d. f.). Although this is a nice formula, it is fairly useless. We can't calculate any moments or other useful quantities, except in the special case where the X_i have a $\mathcal{U}(0,1)$ distribution, so $F(y) = y$ and $f(y) = 1$ for all y and we recognize

$$f_{X_{(k)}}(y) = \frac{n!}{(k-1)!(n-k)!} y^{k-1} (1 - y)^{n-k} \tag{7.36}$$

as a $\text{Beta}(k, n-k+1)$ distribution.

Much more useful is the asymptotic distribution of the sample quantiles given by the following. We will delay the proof of the theorem until the following chapter, where we will develop the tools of multivariate convergence in distribution used in the proof.

Theorem 7.27. *Suppose X_1, X_2, … are continuous random variables that are independent and identically distributed with density f that is nonzero at the p-th quantile x_p, and suppose*

$$\sqrt{n} \left(\frac{k_n}{n} - p \right) \to 0, \qquad \textit{as } n \to \infty, \tag{7.37}$$

then

$$\sqrt{n}\bigl(X_{(k_n)} - x_p\bigr) \xrightarrow{\mathcal{D}} \mathcal{N}\biggl(0, \frac{p(1-p)}{f(x_p)^2}\biggr), \qquad \text{as } n \to \infty. \tag{7.38}$$

Or the sloppy version

$$X_{(k_n)} \approx \mathcal{N}\biggl(x_p, \frac{p(1-p)}{nf(x_p)^2}\biggr).$$

In particular, if we define $k_n = \lceil np \rceil$, then $X_{(k_n)}$ is a sample p-th quantile by Theorem 7.5. The reason for the extra generality, is that the theorem makes it clear that $X_{(k_n+1)}$ also has the same asymptotic distribution. Since $X_{(k_n)} \le X_{(k_n+1)}$ always holds by definition of order statistics, this can only happen if

$$\sqrt{n}\bigl(X_{(k_n+1)} - X_{(k_n)}\bigr) \xrightarrow{P} 0.$$

Hence the average

$$\widetilde{X}_n = \frac{X_{(k_n)} + X_{(k_n+1)}}{2}$$

which is the conventional definition of the sample median, has the same asymptotic normal distribution as either $X_{(k_n)}$ or $X_{(k_n+1)}$.

Corollary 7.28. *Suppose X_1, X_2, … are continuous random variables that are independent and identically distributed with density f that is nonzero the population median m, then*

$$\sqrt{n}\bigl(\widetilde{X}_n - m\bigr) \xrightarrow{\mathcal{D}} \mathcal{N}\biggl(0, \frac{1}{4f(x_p)^2}\biggr), \qquad \text{as } n \to \infty.$$

This is just the theorem with $x_p = m$ and $p = 1/2$. The sloppy version is

$$\widetilde{X}_n \approx \mathcal{N}\biggl(m, \frac{1}{4nf(m)^2}\biggr).$$

Example 7.4.1 (Median, Normal Population).
If $X_1, X_2, \ldots$ are i. i. d. $\mathcal{N}(\mu, \sigma^2)$, then the population median is μ by symmetry and the p. d. f. at the median is

$$f(\mu) = \frac{1}{\sigma\sqrt{2\pi}}$$

Hence

$$\widetilde{X}_n \approx \mathcal{N}\biggl(\mu, \frac{\pi\sigma^2}{2n}\biggr).$$

or, more precisely,

$$\sqrt{n}(\widetilde{X}_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}\biggl(0, \frac{\pi\sigma^2}{2}\biggr)$$

Problems

7-1. The *median absolute deviation from the median* (MAD) of a random variable X with unique median m is the median of the random variable $Y = |X - m|$. The MAD of the values $x_1, \ldots, x_n$ is the median of the values $x_i - \tilde{x}_n$, where $\tilde{x}_n$ is the empirical median defined in Definition 7.1.4. This is much more widely used than the "other MAD," mean absolute deviation from the mean, discussed in Lindgren.

(a) Show that for a symmetric continuous random variable with strictly positive p. d. f. the MAD is half the interquartile range. (The point of requiring a strictly positive p. d. f. is that this makes all the quantiles unique and distinct. The phenomena illustrated in the middle and right panels of Figure 3-3 in Lindgren cannot occur.)

(b) Calculate the MAD for the standard normal distribution.

(c) Calculate the MAD for the data in Problem 7-4 in Lindgren.

7-2. Prove Lemma 7.10.

7-3. Show that if $T \sim t(\nu)$, then $T^2 \sim F(1, \nu)$.

7-4. Show that if $X \sim F(\mu, \nu)$ and $\nu > 2$, then

$$E(X) = \frac{\nu}{\nu - 2}$$

7-5. Prove Theorem 7.20.

7-6. Find the asymptotic distribution of the sample median of an i. i. d. sample from the following distributions:

(a) Cauchy(μ, σ) with density $f_{\mu,\sigma}$ given by

$$f_{\mu,\sigma}(x) = \frac{\sigma}{\pi(\sigma^2 + [x - \mu]^2)}, \qquad -\infty < x < +\infty$$

(b) The double exponential distribution (also called Laplace distribution) having density

$$f_{\mu,\sigma}(x) = \frac{1}{2\sigma} e^{-|x-\mu|/\sigma}, \qquad -\infty < x < +\infty$$

7-7. Suppose X_1, X_2, $\ldots$ are i. i. d. $\mathcal{U}(0, \theta)$. As usual $X_{(n)}$ denotes the n-th order statistic, which is the maximum of the X_i.

(a) Show that

$$X_{(n)} \xrightarrow{P} \theta, \qquad \text{as } n \to \infty.$$

(b) Show that

$$n\bigl(\theta - X_{(n)}\bigr) \xrightarrow{\mathcal{D}} \text{Exp}(1/\theta), \qquad \text{as } n \to \infty.$$

Hints This is a rare problem (the only one of the kind we will meet in this course) when we can't use the LLN or the CLT to get convergence in probability and convergence in distribution results (obvious because the problem is not about $\overline{X}_n$ and the asymptotic distribution we seek isn't normal). Thus we need to derive convergence in distribution directly from the definition (Definition 6.1.1 in these notes or the definition on p. 135 in Lindgren).
Hint for Part (a): Show that the c. d. f. of $X_{(n)}$ converges to the c. d. f. of the constant random variable θ. (Why does this do the job?)
Hint for Part (b): Define

$$Y_n = n\bigl(\theta - X_{(n)}\bigr)$$

(the random variable we're trying to get an asymptotic distribution for). Derive its c. d. f. $F_{Y_n}(y)$. What you need to show is that

$$F_{Y_n}(y) \to F(y), \qquad \text{for all } y$$

where F is the c. d. f. of the Exp$(1/\theta)$ distribution. The fact from calculus

$$\lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n = e^x$$

is useful in this.

You can derive the c. d. f. of Y_n from the c. d. f. of $X_{(n)}$, which is given in the first displayed equation (unnumbered) of Section 7.6 in Lindgren.

7-8. Suppose X_1, ..., X_n are i. i. d. $\mathcal{N}(\mu, \sigma^2)$. What is the probability that $|\overline{X}_n - \mu| > 2S_n/\sqrt{n}$ if $n = 10$?

7-9. Suppose X_1, ..., X_n are i. i. d. $\mathcal{N}(\mu, \sigma^2)$. What is the probability that $S_n^2 > 2\sigma^2$ if $n = 10$?

7-10. R and Mathematica and many textbooks use a different parameterization of the gamma distribution. They write

$$f(x \mid \alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} \tag{7.39}$$

rather than

$$f(x \mid \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \tag{7.40}$$

Clearly the two parameterizations have the same first parameter α, as the notation suggests, and second parameters related by $\lambda = 1/\beta$.

(a) Show that β is the usual kind of scale parameter, that if X has p. d. f. (7.39), then σX has p. d. f. $f(x \mid \alpha, \sigma\beta)$, where again the p. d. f. is defined by (7.39).

(b) Show that λ is an "upside down" scale parameter, that if X has p. d. f. (7.40), then σX has p. d. f. $f(x \mid \alpha, \lambda/\sigma)$, where now the p. d. f. is defined by (7.40).

7-11. Show if X has k-th central moment

$$\mu_k = E\{(X - \mu)^k\}$$

where, as usual, $\mu = E(X)$, then $Y = a + bX$ has k-th central moment $b^k \mu_k$.

7-12. What is the asymptotic distribution of the variance V_n of the empirical distribution for an i. i. d. $\text{Exp}(\lambda)$ sample?

7-13. Suppose X is standard normal (so $\mu_X = 0$ and $\sigma_X = 1$).

(a) What is $P(|X| > 2\sigma_X)$?

In contrast, suppose X has a $t(3)$ distribution (so $\mu_X = 0$ and the variance σ_X^2 is given by Problem 7-5)

(b) Now what is $P(|X| > 2\sigma_X)$?

7-14. With all the same assumptions as in Example 7.3.8, what are

(a) $P(S_2^2 > S_1^2)$?

(b) $P(S_2^2 > 2S_1^2)$?

7-15. Suppose X_1, X_2, X_3, $\ldots$ is an i. i. d. sequence of random variables with mean μ and variance σ^2, and $\overline{X}_n$ is the sample mean. Show that

$$\sqrt{n}\left(\overline{X}_n - \mu\right)^k \xrightarrow{P} 0$$

for any integer $k > 1$. (**Hint:** Use the CLT, the continuous mapping theorem for convergence in distribution, and Slutsky's theorem.)

Appendix A

Greek Letters

Table A.1: Table of Greek Letters (Continued on following page.)

name	capital letter	small letter	pronunciation	sound
alpha	A	α	AL-fah	short a
beta	B	β	BAY-tah	b
gamma	Γ	γ	GAM-ah	g
delta	Δ	δ	DEL-tah	d
epsilon	E	ϵ	EP-si-lon	e
zeta	Z	ζ	ZAY-tah	z
eta	H	η	AY-tah	long a
theta	Θ	θ or ϑ	THAY-thah	soft th (as in thin)
iota	I	ι	EYE-oh-tah	i
kappa	K	κ	KAP-ah	k
lambda	Λ	λ	LAM-dah	l
mu	M	μ	MYOO	m
nu	N	ν	NOO	n
xi	Ξ	ξ	KSEE	x (as in box)
omicron	O	o	OH-mi-kron	o
pi	Π	π	PIE	p
rho	R	ρ	RHOH	rh[1]
sigma	Σ	σ	SIG-mah	s
tau	T	τ	TAOW	t
upsilon	Υ	υ	UP-si-lon	u

[1] The sound of the Greek letter ρ is not used in English. English words, like *rhetoric* and *rhinoceros* that are descended from Greek words beginning with ρ have English pronunciations beginning with an “r” sound rather than “rh” (though the spelling reminds us of the Greek origin).

Table A.2: Table of Greek Letters (Continued.)

name	capital letter	small letter	pronunciation	sound
phi	Φ	ϕ or φ	FIE	f
chi	X	χ	KIE	guttural ch[2]
psi	Ψ	ψ	PSY	ps (as in stops)[3]
omega	Ω	ω	oh-MEG-ah	o

[2] The sound of the Greek letter χ is not used in English. It is heard in the German *Buch* or Scottish *loch*. English words, like *chemistry* and *chorus* that are descended from Greek words beginning with χ have English pronunciations beginning with a "k" sound rather than "guttural ch" (though the spelling reminds us of the Greek origin).

[3] English words, like *pseudonym* and *psychology* that are descended from Greek words beginning with ψ have English pronunciations beginning with an "s" sound rather than "ps" (though the spelling reminds us of the Greek origin).

Appendix B

Summary of Brand-Name Distributions

B.1 Discrete Distributions

B.1.1 The Discrete Uniform Distribution

The Abbreviation $\mathcal{DU}(S)$.

The Sample Space Any finite set S.

The Density

$$f(x) = \frac{1}{n}, \qquad x \in S,$$

where $n = \operatorname{card}(S)$.

Specialization The case in which the sample space consists of consecutive integers $S = \{m, m+1, \ldots, n\}$ is denoted $\mathcal{DU}(m, n)$.

Moments If $X \sim \mathcal{DU}(1, n)$, then

$$\begin{aligned} E(X) &= \frac{n+1}{2} \\ \operatorname{var}(X) &= \frac{n^2 - 1}{12} \end{aligned}$$

B.1.2 The Binomial Distribution

The Abbreviation $\text{Bin}(n, p)$

The Sample Space The integers 0, …, n.

The Parameter p such that $0 < p < 1$.

The Density

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, \ldots, n.$$

Moments

$$\begin{aligned} E(X) &= np \\ \operatorname{var}(X) &= np(1-p) \end{aligned}$$

Specialization

$$\operatorname{Ber}(p) = \operatorname{Bin}(1, p)$$

B.1.3 The Geometric Distribution, Type II

Note This section has changed. The roles of p and $1 - p$ have been reversed, and the abbreviation Geo(p) is no longer used to refer to this distribution but the distribution defined in Section B.1.8. All of the changes are to match up with Chapter 6 in Lindgren.

The Abbreviation No abbreviation to avoid confusion with the other type defined in Section B.1.8.

Relation Between the Types If $X \sim \operatorname{Geo}(p)$, then $Y = X - 1$ has the distribution defined in this section.

X is the number of *trials* before the first success in an i. i. d. sequence of Ber(p) random variables. Y is the number of *failures* before the first success.

The Sample Space The integers 0, 1, ...

The Parameter p such that $0 < p < 1$.

The Density

$$f(x) = p(1-p)^x, \qquad x = 0, 1, \ldots.$$

Moments

$$\begin{aligned} E(X) &= \frac{1}{p} - 1 = \frac{1-p}{p} \\ \operatorname{var}(X) &= \frac{1-p}{p^2} \end{aligned}$$

B.1.4 The Poisson Distribution

The Abbreviation Poi(μ)

The Sample Space The integers 0, 1, ….

The Parameter μ such that $\mu > 0$.

The Density

$$f(x) = \frac{\mu^x}{x!} e^{-\mu}, \qquad x = 0, 1, \ldots.$$

Moments

$$\begin{aligned} E(X) &= \mu \\ \operatorname{var}(X) &= \mu \end{aligned}$$

B.1.5 The Bernoulli Distribution

The Abbreviation Ber(p)

The Sample Space The integers 0 and 1.

The Parameter p such that $0 < p < 1$.

The Density

$$f(x) = \begin{cases} p, & x = 1 \\ 1 - p & x = 0 \end{cases}$$

Moments

$$\begin{aligned} E(X) &= p \\ \operatorname{var}(X) &= p(1 - p) \end{aligned}$$

Generalization

$$\operatorname{Ber}(p) = \operatorname{Bin}(1, p)$$

B.1.6 The Negative Binomial Distribution, Type I

The Abbreviation NegBin(k, p)

The Sample Space The integers k, $k + 1$, ….

The Parameter p such that $0 < p < 1$.

The Density

$$f(x) = \binom{x-1}{k-1} p^k (1-p)^{x-k}, \qquad x = k, k+1, \ldots.$$

Moments

$$\begin{aligned} E(X) &= \frac{k}{p} \\ \operatorname{var}(X) &= \frac{k(1-p)}{p^2} \end{aligned}$$

Specialization

$$\text{Geo}(p) = \text{NegBin}(1, p)$$

B.1.7 The Negative Binomial Distribution, Type II

The Abbreviation No abbreviation to avoid confusion with the other type defined in Section B.1.6.

Relation Between the Types If $X \sim \text{NegBin}(k, p)$, then $Y = X - k$ has the distribution defined in this section.

X is the number of *trials* before the k-th success in an i. i. d. sequence of $\text{Ber}(p)$ random variables. Y is the number of *failures* before the k-th success.

The Sample Space The integers 0, 1, ...

The Parameter p such that $0 < p < 1$.

The Density

$$f(x) = \binom{x-1}{k-1} p^k (1-p)^{x}, \qquad x = 0, 1, \ldots.$$

Moments

$$\begin{aligned} E(X) &= \frac{k}{p} - k = \frac{k(1-p)}{p} \\ \operatorname{var}(X) &= \frac{k(1-p)}{p^2} \end{aligned}$$

B.1.8 The Geometric Distribution, Type I

The Abbreviation $\text{Geo}(p)$

The Sample Space The integers 1, 2, ...

The Parameter p such that $0 < p < 1$.

The Density

$$f(x) = p(1-p)^{x-1}, \qquad x = 1, 2, \ldots.$$

Moments

$$\begin{aligned} E(X) &= \frac{1}{p} \\ \operatorname{var}(X) &= \frac{1-p}{p^2} \end{aligned}$$

Generalization

$$\operatorname{Geo}(p) = \operatorname{NegBin}(1, p)$$

B.2 Continuous Distributions

B.2.1 The Uniform Distribution

The Abbreviation $\mathcal{U}(S)$.

The Sample Space Any subset S of $\mathbb{R}^d$.

The Density

$$f(x) = \frac{1}{c}, \qquad x \in S,$$

where

$$c = m(S) = \int_S dx$$

is the measure of S (length in $\mathbb{R}^1$, area in $\mathbb{R}^2$, volume in $\mathbb{R}^3$, and so forth).

Specialization The case having $S = (a, b)$ in $\mathbb{R}^1$ and density

$$f(x) = \frac{1}{b-a}, \qquad a < x < b$$

is denoted $\mathcal{U}(a, b)$.

Moments If $X \sim \mathcal{U}(a, b)$, then

$$\begin{aligned} E(X) &= \frac{a+b}{2} \\ \operatorname{var}(X) &= \frac{(b-a)^2}{12} \end{aligned}$$

B.2.2 The Exponential Distribution

The Abbreviation Exp(λ).

The Sample Space The interval $(0, \infty)$ of the real numbers.

The Parameter λ such that $\lambda > 0$.

The Density

$$f(x) = \lambda e^{-\lambda x}, \qquad x > 0.$$

Moments

$$\begin{aligned} E(X) &= \frac{1}{\lambda} \\ \operatorname{var}(X) &= \frac{1}{\lambda^2} \end{aligned}$$

Generalization

$$\operatorname{Exp}(\lambda) = \operatorname{Gam}(1, \lambda)$$

B.2.3 The Gamma Distribution

The Abbreviation Gam(α, λ).

The Sample Space The interval $(0, \infty)$ of the real numbers.

The Parameters α and λ such that $\alpha > 0$ and $\lambda > 0$.

The Density

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha - 1} e^{-\lambda x}, \qquad x > 0.$$

where $\Gamma(\alpha)$ is the gamma function (Section B.3.1 below).

Moments

$$\begin{aligned} E(X) &= \frac{\alpha}{\lambda} \\ \operatorname{var}(X) &= \frac{\alpha}{\lambda^2} \end{aligned}$$

Specialization

$$\begin{aligned} \operatorname{Exp}(\lambda) &= \operatorname{Gam}(1, \lambda) \\ \operatorname{chi}^2(k) &= \operatorname{Gam}\left(\tfrac{k}{2}, \tfrac{1}{2}\right) \end{aligned}$$

B.2.4 The Beta Distribution

The Abbreviation Beta(s, t).

The Sample Space The interval $(0, 1)$ of the real numbers.

The Parameters s and t such that $s > 0$ and $t > 0$.

The Density

$$f(x) = \frac{1}{B(s,t)} x^{s-1}(1-x)^{t-1} \qquad 0 < x < 1.$$

where $B(s, t)$ is the *beta function* defined by

$$B(s,t) = \frac{\Gamma(s)\Gamma(t)}{\Gamma(s+t)} \tag{B.1}$$

Moments

$$\begin{aligned} E(X) &= \frac{s}{s+t} \\ \operatorname{var}(X) &= \frac{st}{(s+t)^2(s+t+1)} \end{aligned}$$

B.2.5 The Normal Distribution

The Abbreviation $\mathcal{N}(\mu, \sigma^2)$.

The Sample Space The real line $\mathbb{R}$.

The Parameters μ and σ^2 such that $\sigma^2 > 0$.

The Density

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad x \in \mathbb{R}.$$

Moments

$$\begin{aligned} E(X) &= \mu \\ \operatorname{var}(X) &= \sigma^2 \\ \mu_4 &= 3\sigma^4 \end{aligned}$$

B.2.6 The Chi-Square Distribution

The Abbreviation chi$^2(k)$.

The Sample Space The interval $(0, \infty)$ of the real numbers.

The Parameter A positive integer k.

The Density

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}, \qquad x > 0.$$

Moments

$$\begin{aligned} E(X) &= k \\ \operatorname{var}(X) &= 2k \end{aligned}$$

Generalization

$$\operatorname{chi}^2(k) = \operatorname{Gam}\left(\tfrac{k}{2}, \tfrac{1}{2}\right)$$

B.2.7 The Cauchy Distribution

The Abbreviation Cauchy(μ, σ).

The Sample Space The real line $\mathbb{R}$.

The Parameters μ and σ such that $\sigma > 0$.

The Density

$$f(x) = \frac{1}{\pi} \cdot \frac{\sigma}{\sigma^2 + (x - \mu)^2}, \qquad x \in \mathbb{R}.$$

Moments None: $E(|X|) = \infty$.

B.3 Special Functions

B.3.1 The Gamma Function

The Definition

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} \, dx, \qquad \alpha > 0 \tag{B.2}$$

The Recursion Relation

$$\Gamma(\alpha + 1) = \alpha \Gamma(\alpha) \tag{B.3}$$

Known Values

$$\Gamma(1) = 1$$

and hence using the recursion relation

$$\Gamma(n+1) = n!$$

for any nonnegative integer n.

Also

$$\Gamma(\tfrac{1}{2}) = \sqrt{\pi}$$

and hence using the recursion relation

$$\begin{aligned} \Gamma(\tfrac{3}{2}) &= \tfrac{1}{2}\sqrt{\pi} \\ \Gamma(\tfrac{5}{2}) &= \tfrac{3}{2} \cdot \tfrac{1}{2}\sqrt{\pi} \\ \Gamma(\tfrac{7}{2}) &= \tfrac{5}{2} \cdot \tfrac{3}{2} \cdot \tfrac{1}{2}\sqrt{\pi} \end{aligned}$$

and so forth.

B.3.2 The Beta Function

The function $B(s, t)$ defined by (B.1).

B.4 Discrete Multivariate Distributions

B.4.1 The Multinomial Distribution

The Abbreviation $\text{Multi}_k(n, \mathbf{p})$ or $\text{Multi}(n, \mathbf{p})$ if the dimension k is clear from context.

The Sample Space

$$S = \{\, \mathbf{y} \in \mathbb{N}^k : y_1 + \cdots y_k = n \,\}$$

where $\mathbb{N}$ denotes the "natural numbers" 0, 1, 2, ...

The Parameter $\mathbf{p} = (p_1, \ldots, p_k)$ such that $p_i \geq 0$ for all i and $\sum_i p_i = 1$.

The Density

$$f(\mathbf{y}) = \binom{n}{y_1, \ldots, y_k} \prod_{j=1}^{k} p_j^{y_j}, \qquad \mathbf{y} \in S$$

Moments

$$\begin{aligned} E(\mathbf{Y}) &= n\mathbf{p} \\ \operatorname{var}(\mathbf{Y}) &= \mathbf{M} \end{aligned}$$

where $\mathbf{M}$ is the $k \times k$ matrix with elements

$$m_{ij} = \begin{cases} np_i(1 - p_i), & i = j \\ -np_i p_j & i \neq j \end{cases}$$

Specialization The special case $n = 1$ is called the multivariate Bernoulli distribution

$$\mathrm{Ber}_k(\mathbf{p}) = \mathrm{Bin}_k(1, \mathbf{p})$$

but for once we will not spell out the details with a special section for the multivariate Bernoulli. Just take $n = 1$ in this section.

Marginal Distributions Distributions obtained by collapsing categories are again multinomial (Section 5.4.5 in these notes).

In particular, if $\mathbf{Y} \sim \mathrm{Multi}_k(n, \mathbf{p})$, then

$$(Y_1, \ldots, Y_j, Y_{j+1} + \cdots + Y_k) \sim \mathrm{Multi}_{j+1}(n, \mathbf{q}) \tag{B.4}$$

where

$$\begin{aligned} q_i &= p_i, & i \le j \\ q_{j+1} &= p_{j+1} + \cdots p_k \end{aligned}$$

Because the random vector in (B.4) is degenerate, this equation also gives implicitly the marginal distribution of Y_1, …, Y_j

$$\begin{aligned} & f(y_1, \ldots, y_j) \\ & \quad = \binom{n}{y_1, \ldots, y_j, n - y_1 - \cdots - y_j} p_1^{y_1} \cdots p_j^{y_j} (1 - p_1 - \cdots - p_j)^{n - y_1 - \cdots - y_j} \end{aligned}$$

Univariate Marginal Distributions If $\mathbf{Y} \sim \mathrm{Multi}(n, \mathbf{p})$, then

$$Y_i \sim \mathrm{Bin}(n, p_i).$$

Conditional Distributions If $\mathbf{Y} \sim \mathrm{Multi}_k(n, \mathbf{p})$, then

$$(Y_1, \ldots, Y_j) \mid (Y_{j+1}, \ldots, Y_k) \sim \mathrm{Multi}_j(n - Y_{j+1} - \cdots - Y_k, \mathbf{q}),$$

where

$$q_i = \frac{p_i}{p_1 + \cdots + p_j}, \qquad i = 1, \ldots, j.$$

B.5 Continuous Multivariate Distributions

B.5.1 The Uniform Distribution

The uniform distribution defined in Section B.2.1 actually made no mention of dimension. If the set S on which the distribution is defined lies in $\mathbb{R}^n$, then this is a multivariate distribution.

Conditional Distributions Every conditional distribution of a multivariate uniform distribution is uniform.

Marginal Distributions No regularity. Depends on the particular distribution. Marginals of the uniform distribution on a rectangle with sides parallel to the coordinate axes are uniform. Marginals of the uniform distribution on a disk or triangle are not uniform.

B.5.2 The Standard Normal Distribution

The distribution of a random vector $\mathbf{Z} = (Z_1, \ldots, Z_k)$ with the Z_i i. i. d. standard normal.

Moments

$$\begin{aligned} E(\mathbf{Z}) &= 0 \\ \operatorname{var}(\mathbf{Z}) &= \mathbf{I}, \end{aligned}$$

where $\mathbf{I}$ denotes the $k \times k$ identity matrix.

B.5.3 The Multivariate Normal Distribution

The distribution of a random vector $\mathbf{X} = \mathbf{a} + \mathbf{B}\mathbf{Z}$, where $\mathbf{Z}$ is multivariate standard normal.

Moments

$$\begin{aligned} E(\mathbf{X}) &= \boldsymbol{\mu} = \mathbf{a} \\ \operatorname{var}(\mathbf{X}) &= \mathbf{M} = \mathbf{B}\mathbf{B}' \end{aligned}$$

The Abbreviation $\mathcal{N}_k(\boldsymbol{\mu}, \mathbf{M})$ or $\mathcal{N}(\boldsymbol{\mu}, \mathbf{M})$ if the dimension k is clear from context.

The Sample Space If $\mathbf{M}$ is positive definite, the sample space is $\mathbb{R}^k$.

Otherwise, X is concentrated on the intersection of hyperplanes determined by null eigenvectors of $\mathbf{M}$

$$S = \{\, \mathbf{x} \in \mathbb{R}^k : \mathbf{z}'\mathbf{x} = \mathbf{z}'\boldsymbol{\mu} \text{ whenever } \mathbf{M}\mathbf{z} = 0 \,\}$$

The Parameters The mean vector $\boldsymbol{\mu}$ and variance matrix $\mathbf{M}$.

The Density Only exists if the distribution is nondegenerate ($\mathbf{M}$ is positive definite). Then

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} \det(\mathbf{M})^{1/2}} \exp\left(-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu})' \mathbf{M}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right), \qquad \mathbf{x} \in \mathbb{R}^k$$

Marginal Distributions All are normal. If

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$$

is a partitioned random vector with (partitioned) mean vector

$$E(\mathbf{X}) = \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}$$

and (partitioned) variance matrix

$$\operatorname{var}(\mathbf{X}) = \mathbf{M} = \begin{pmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{pmatrix}$$

and $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{M})$, then

$$\mathbf{X}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{M}_{11}).$$

Conditional Distributions All are normal. If $\mathbf{X}$ is as in the preceding section and $\mathbf{X}_2$ is nondegenerate, then the conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_2$ is normal with

$$\begin{aligned} E(\mathbf{X}_1 \mid \mathbf{X}_2) &= \boldsymbol{\mu}_1 + \mathbf{M}_{12}\mathbf{M}_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2) \\ \operatorname{var}(\mathbf{X}_1 \mid \mathbf{X}_2) &= \mathbf{M}_{11} - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{21} \end{aligned}$$

If $\mathbf{X}_2$ is degenerate so $\mathbf{M}_{22}$ is not invertible, then the conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_2$ is still normal and the same formulas work if $\mathbf{M}_{22}^{-1}$ is replaced by a generalized inverse.

B.5.4 The Bivariate Normal Distribution

The special case $k = 2$ of the preceeding section.

The Density

$$\begin{aligned} f(x, y) = {} & \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \times \\ & \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right) \end{aligned}$$

Marginal Distributions

$$Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$$

Conditional Distributions The conditional distribution of X given Y is normal with

$$\begin{aligned} E(X \mid Y) &= \mu_X + \rho \frac{\sigma_X}{\sigma_Y}(Y - \mu_Y) \\ \operatorname{var}(X \mid Y) &= \sigma_X^2(1 - \rho^2) \end{aligned}$$

where $\rho = \operatorname{cor}(X, Y)$.

Appendix C

Addition Rules for Distributions

"Addition rules" for distributions are rules of the form: if X_1, …, X_k are independent with some specified distributions, then $X_1 + \cdots + X_k$ has some other specified distribution.

Bernoulli If X_1, …, X_k are i. i. d. Ber(p), then

$$X_1 + \cdots + X_k \sim \text{Bin}(k, p). \tag{C.1}$$

- All the Bernoulli distributions must have the *same* success probability p.

Binomial If X_1, …, X_k are independent with $X_i \sim \text{Bin}(n_i, p)$, then

$$X_1 + \cdots + X_k \sim \text{Bin}(n_1 + \cdots + n_k, p). \tag{C.2}$$

- All the binomial distributions must have the *same* success probability p.
- (C.1) is the special case of (C.2) obtained by setting $n_1 = \cdots = n_k = 1$.

Geometric If X_1, …, X_k are i. i. d. Geo(p), then

$$X_1 + \cdots + X_k \sim \text{NegBin}(k, p). \tag{C.3}$$

- All the geometric distributions must have the *same* success probability p.

Negative Binomial If X_1, …, X_k are independent with $X_i \sim \text{NegBin}(n_i, p)$, then

$$X_1 + \cdots + X_k \sim \text{NegBin}(n_1 + \cdots + n_k, p). \tag{C.4}$$

- All the negative binomial distributions must have the *same* success probability p.
- (C.3) is the special case of (C.4) obtained by setting $n_1 = \cdots = n_k = 1$.

Poisson If X_1, …, X_k are independent with $X_i \sim \text{Poi}(\mu_i)$, then

$$X_1 + \cdots + X_k \sim \text{Poi}(\mu_1 + \cdots + \mu_k). \tag{C.5}$$

Exponential If X_1, …, X_k are i. i. d. $\text{Exp}(\lambda)$, then

$$X_1 + \cdots + X_k \sim \text{Gam}(n, \lambda). \tag{C.6}$$

- All the exponential distributions must have the *same* rate parameter λ.

Gamma If X_1, …, X_k are independent with $X_i \sim \text{Gam}(\alpha_i, \lambda)$, then

$$X_1 + \cdots + X_k \sim \text{Gam}(\alpha_1 + \cdots + \alpha_k, \lambda). \tag{C.7}$$

- All the gamma distributions must have the *same* rate parameter λ.
- (C.6) is the special case of (C.7) obtained by setting $\alpha_1 = \cdots = \alpha_k = 1$.

Chi-Square If X_1, …, X_k are independent with $X_i \sim \text{chi}^2(n_i)$, then

$$X_1 + \cdots + X_k \sim \text{chi}^2(n_1 + \cdots + n_k). \tag{C.8}$$

- (C.8) is the special case of (C.7) obtained by setting

$$\alpha_i = n_i/2 \quad \text{and} \quad \lambda_i = 1/2, \qquad i = 1, \ldots, k.$$

Normal If X_1, …, X_k are independent with $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, then

$$X_1 + \cdots + X_k \sim \mathcal{N}(\mu_1 + \cdots + \mu_k, \sigma_1^2 + \cdots + \sigma_k^2). \tag{C.9}$$

Linear Combination of Normals If X_1, …, X_k are independent with $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ and a_1, …, a_k are constants, then

$$\sum_{i=1}^k a_i X_i \sim \mathcal{N}\left(\sum_{i=1}^k a_i \mu_i, \sum_{i=1}^k a_i^2 \sigma_i^2\right). \tag{C.10}$$

- (C.9) is the special case of (C.10) obtained by setting $a_1 = \cdots = a_k = 1$.

Cauchy If X_1, …, X_k are independent with $X_i \sim \text{Cauchy}(\mu, \sigma)$, then

$$X_1 + \cdots + X_k \sim \text{Cauchy}(n\mu, n\sigma). \tag{C.11}$$

Appendix D

Relations Among Brand Name Distributions

D.1 Special Cases

First there are the special cases, which were also noted in Appendix B.

$$\begin{aligned} \text{Ber}(p) &= \text{Bin}(1, p) \\ \text{Geo}(p) &= \text{NegBin}(1, p) \\ \text{Exp}(\lambda) &= \text{Gam}(1, \lambda) \\ \text{chi}^2(k) &= \text{Gam}\left(\tfrac{k}{2}, \tfrac{1}{2}\right) \end{aligned}$$

The main point of this appendix are the relationships that involve more theoretical issues.

D.2 Relations Involving Bernoulli Sequences

Suppose X_1, X_2, $\ldots$ are i. i. d. Ber(p) random variables.

If n is a positive integer and

$$Y = X_1 + \cdots + X_n$$

is the number of "successes" in the n Bernoulli trials, then

$$Y \sim \text{Bin}(n, p).$$

On the other hand, if y is positive integer and N is the trial at which the y-th success occurs, that is the random number N such that

$$\begin{aligned} X_1 + \cdots + X_N &= y \\ X_1 + \cdots + X_k &< y, \qquad k < N, \end{aligned}$$

then

$$N \sim \text{NegBin}(y, p).$$

D.3 Relations Involving Poisson Processes

In a one-dimensional homogeneous Poisson process with rate parameter λ, the counts are Poisson and the waiting and interarrival times are exponential. Specifically, the number of points (arrivals) in an interval of length t has the $\text{Poi}(\lambda t)$ distribution, and the waiting times and interarrival times are independent and indentically $\text{Exp}(\lambda)$ distributed.

Even more specifically, let X_1, X_2, $\ldots$ be i. i. d. $\text{Exp}(\lambda)$ random variables. Take these to be the waiting and interarrival times of a Poisson process. This means the arrival times themselves are

$$T_k = \sum_{i=1}^{k} X_i$$

Note that

$$0 < T_1 < T_2 < \cdots$$

and

$$X_i = T_i - T_{i-1}, \qquad i > 1$$

so these are the interarrival times and $X_1 = T_1$ is the waiting time until the first arrival.

The characteristic property of the Poisson process, that counts have the Poisson distribution, says the number of points in the interval $(0, t)$, that is, the number of T_i such that $T_i < t$, has the $\text{Poi}(\lambda t)$ distribution.

D.4 Normal and Chi-Square

If Z_1, Z_2, $\ldots$ are i. i. d. $\mathcal{N}(0, 1)$, then

$$Z_1^2 + \ldots Z_n^2 \sim \text{chi}^2(n).$$

Appendix E

Eigenvalues and Eigenvectors

E.1 Orthogonal and Orthonormal Vectors

If $\mathbf{x}$ and $\mathbf{y}$ are vectors of the same dimension, we say they are *orthogonal* if $\mathbf{x}'\mathbf{y} = 0$. Since the transpose of a matrix product is the product of the transposes in reverse order, an equivalent condition is $\mathbf{y}'\mathbf{x} = 0$. Orthogonality is the n-dimensional generalization of perpendicularity. In a sense, it says that two vectors make a right angle.

The *length* or *norm* of a vector $\mathbf{x} = (x_1, \ldots, x_n)$ is defined to be

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}'\mathbf{x}} = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

Squaring both sides gives

$$\|\mathbf{x}\|^2 = \sum_{i=1}^{n} x_i^2,$$

which is one version of the Pythagorean theorem, as it appears in analytic geometry.

Orthogonal vectors give another generalization of the Pythagorean theorem. We say a set of vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ is *orthogonal* if

$$\mathbf{x}_i'\mathbf{x}_j = 0, \qquad i \neq j. \tag{E.1}$$

Then

$$\begin{aligned}\|\mathbf{x}_1 + \cdots + \mathbf{x}_k\|^2 &= (\mathbf{x}_1 + \cdots + \mathbf{x}_k)'(\mathbf{x}_1 + \cdots + \mathbf{x}_k) \\ &= \sum_{i=1}^k \sum_{j=1}^k \mathbf{x}_i'\mathbf{x}_j \\ &= \sum_{i=1}^k \mathbf{x}_i'\mathbf{x}_i \\ &= \sum_{i=1}^k \|\mathbf{x}_i\|^2\end{aligned}$$

because, by definition of orthogonality, all terms in the second line with $i \neq j$ are zero.

We say an orthogonal set of vectors is *orthonormal* if

$$\mathbf{x}_i'\mathbf{x}_i = 1. \tag{E.2}$$

That is, a set of vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ is orthonormal if it satisfies both (E.1) and (E.2).

An orthonormal set is automatically linearly independent because if

$$\sum_{i=1}^k c_i\mathbf{x}_i = 0,$$

then

$$0 = \mathbf{x}_j'\left(\sum_{i=1}^k c_i\mathbf{x}_i\right) = c_j\mathbf{x}_j'\mathbf{x}_j = c_j$$

holds for all j. Hence the only linear combination that is zero is the one with all coefficients zero, which is the definition of linear independence.

Being linearly independent, an orthonormal set is always a *basis* for whatever subspace it spans. If we are working in n-dimensional space, and there are n vectors in the orthonormal set, then they make up a basis for the whole space. If there are $k < n$ vectors in the set, then they make up a basis for some proper subspace.

It is always possible to choose an orthogonal basis for any vector space or subspace. One way to do this is the Gram-Schmidt orthogonalization procedure, which converts an arbitrary basis $\mathbf{y}_1$, $\ldots$, $\mathbf{y}_n$ to an orthonormal basis $\mathbf{x}_1$, $\ldots$, $\mathbf{x}_n$ as follows. First let

$$\mathbf{x}_1 = \frac{\mathbf{y}_1}{\|\mathbf{y}_1\|}.$$

Then define the $\mathbf{x}_i$ in order. After $\mathbf{x}_1$, $\ldots$, $\mathbf{x}_{k-1}$ have been defined, let

$$\mathbf{z}_k = \mathbf{y}_k - \sum_{i=1}^{k-1} \mathbf{x}_i\mathbf{x}_i'\mathbf{y}$$

and

$$\mathbf{x}_k = \frac{\mathbf{z}_k}{\|\mathbf{z}_k\|}.$$

It is easily verified that this does produce an orthonormal set, and it is only slightly harder to prove that none of the $\mathbf{x}_i$ are zero because that would imply linear dependence of the $\mathbf{y}_i$.

E.2 Eigenvalues and Eigenvectors

If $\mathbf{A}$ is any matrix, we say that λ is a *right eigenvalue* corresponding to a *right eigenvector* $\mathbf{x}$ if

$$\mathbf{A}\mathbf{x} = \lambda \mathbf{x}$$

Left eigenvalues and eigenvectors are defined analogously with "left multiplication" $\mathbf{x}'\mathbf{A} = \lambda \mathbf{x}'$, which is equivalent to $\mathbf{A}'\mathbf{x} = \lambda \mathbf{x}$. So the right eigenvalues and eigenvectors of $\mathbf{A}'$ are the left eigenvalues and eigenvectors of $\mathbf{A}$. When $\mathbf{A}$ is symmetric ($\mathbf{A}' = \mathbf{A}$), the "left" and "right" concepts are the same and the adjectives "left" and "right" are unnecessary. Fortunately, this is the most interesting case, and the only one in which we will be interested. From now on we discuss only eigenvalues and eigenvectors of *symmetric* matrices.

There are three important facts about eigenvalues and eigenvectors. Two elementary and one very deep. Here's the first (one of the elementary facts).

Lemma E.1. *Eigenvectors corresponding to distinct eigenvalues are orthogonal.*

This means that if

$$\mathbf{A}\mathbf{x}_i = \lambda_i \mathbf{x}_i \tag{E.3}$$

then

$$\lambda_i \neq \lambda_j \qquad \text{implies} \qquad \mathbf{x}_i'\mathbf{x}_j = 0.$$

Proof. Suppose $\lambda_i \neq \lambda_j$, then at least one of the two is not zero, say λ_j. Then

$$\mathbf{x}_i'\mathbf{x}_j = \frac{\mathbf{x}_i'\mathbf{A}\mathbf{x}_j}{\lambda_j} = \frac{(\mathbf{A}\mathbf{x}_i)'\mathbf{x}_j}{\lambda_j} = \frac{\lambda_i \mathbf{x}_i'\mathbf{x}_j}{\lambda_j} = \frac{\lambda_i}{\lambda_j} \cdot \mathbf{x}_i'\mathbf{x}_j$$

and since $\lambda_i \neq \lambda_j$ the only way this can happen is if $\mathbf{x}_i'\mathbf{x}_j = 0$. □

Here's the second important fact (also elementary).

Lemma E.2. *Every linear combination of eigenvectors corresponding to the same eigenvalue is another eigenvector corresponding to that eigenvalue.*

This means that if

$$\mathbf{A}\mathbf{x}_i = \lambda \mathbf{x}_i$$

then

$$\mathbf{A}\left(\sum_{i=1}^{k} c_i \mathbf{x}_i\right) = \lambda \left(\sum_{i=1}^{k} c_i \mathbf{x}_i\right)$$

Proof. This is just linearity of matrix multiplication. □

The second property means that all the eigenvectors corresponding to one eigenvalue constitute a subspace. If the dimension of that subspace is k, then it is possible to choose an orthonormal basis of k vectors that span the subspace. Since the first property of eigenvalues and eigenvectors says that (E.1) is also satisfied by eigenvectors corresponding to different eigenvalues, all of the eigenvectors chosen this way form an orthonormal set.

Thus our orthonormal set of eigenvectors spans a subspace of dimension m which contains all eigenvectors of the matrix in question. The question then arises whether this set is *complete*, that is, whether it is a basis for the whole space, or in symbols whether $m = n$, where n is the dimension of the whole space ($\mathbf{A}$ is an $n \times n$ matrix and the $\mathbf{x}_i$ are vectors of dimension n). It turns out that the set *is* always complete, and this is the third important fact about eigenvalues and eigenvectors.

Lemma E.3. *Every real symmetric matrix has an orthonormal set of eigenvectors that form a basis for the space.*

In contrast to the first two facts, this is deep, and we shall not say anything about its proof, other than that about half of the typical linear algebra book is given over to building up to the proof of this one fact.

The "third important fact" says that *any* vector can be written as a linear combination of eigenvectors

$$\mathbf{y} = \sum_{i=1}^{n} c_i \mathbf{x}_i$$

and this allows a very simple description of the action of the linear operator described by the matrix

$$\mathbf{A}\mathbf{y} = \sum_{i=1}^{n} c_i \mathbf{A}\mathbf{x}_i = \sum_{i=1}^{n} c_i \lambda_i \mathbf{x}_i \tag{E.4}$$

So this says that *when we use an orthonormal eigenvector basis*, if $\mathbf{y}$ has the representation $(c_1, \ldots, c_n)$, then $\mathbf{A}y$ has the representation $(c_1\lambda_1, \ldots, c_n\lambda_n)$. Let $\mathbf{D}$ be the representation in the orthonormal eigenvector basis of the linear operator represented by $\mathbf{A}$ in the standard basis. Then our analysis above says the i-the element of $\mathbf{D}\mathbf{c}$ is $c_i\lambda_i$, that is,

$$\sum_{j=1}^{n} d_{ij} c_j = \lambda_i c_i.$$

In order for this to hold for all real numbers c_i, it must be that $\mathbf{D}$ is diagonal

$$\begin{aligned} d_{ii} &= \lambda_i \\ d_{ij} &= 0, \qquad i \neq j \end{aligned}$$

In short, using the orthonormal eigenvector basis *diagonalizes* the linear operator represented by the matrix in question.

There is another way to describe this same fact without mentioning bases. Many people find it a simpler description, though its relation to eigenvalues and eigenvectors is hidden in the notation, no longer immediately apparent. Let $\mathbf{O}$ denote the matrix whose columns are the orthonormal eigenvector basis ($\mathbf{x}_1$, …, $\mathbf{x}_n$), that is, if o_{ij} are the elements of $\mathbf{O}$, then

$$\mathbf{x}_i = (o_{1i}, \ldots, o_{ni}).$$

Now (E.1) and (E.2) can be combined as one matrix equation

$$\mathbf{O}'\mathbf{O} = \mathbf{I} \tag{E.5}$$

(where, as usual, $\mathbf{I}$ is the $n \times n$ identity matrix). A matrix $\mathbf{O}$ satisfying this property is said to be *orthogonal.* Another way to read (E.5) is that it says $\mathbf{O}' = \mathbf{O}^{-1}$ (an orthogonal matrix is one whose inverse is its transpose). The fact that inverses are two-sided ($\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ for any invertible matrix $\mathbf{A}$) implies that $\mathbf{O}\mathbf{O}' = \mathbf{I}$ as well.

Furthermore, the eigenvalue-eigenvector equation (E.3) can be written out with explicit subscripts and summations as

$$\sum_{j=1}^{n} a_{ij} o_{jk} = \lambda_k o_{ik} = o_{ik} d_{kk} = \sum_{j=1}^{n} o_{ij} d_{jk}$$

(where $\mathbf{D}$ is the the diagonal matrix with eigenvalues on the diagonal defined above). Going back to matrix notation gives

$$\mathbf{A}\mathbf{O} = \mathbf{O}\mathbf{D} \tag{E.6}$$

The two equations (E.3) and (E.6) may not look much alike, but as we have just seen, they say exactly the same thing in different notation. Using the orthogonality property ($\mathbf{O}' = \mathbf{O}^{-1}$) we can rewrite (E.6) in two different ways.

Theorem E.4 (Spectral Decomposition). *Any real symmetric matrix* $\mathbf{A}$ *can be written*

$$\mathbf{A} = \mathbf{O}\mathbf{D}\mathbf{O}' \tag{E.7}$$

where $\mathbf{D}$ *is diagonal and* $\mathbf{O}$ *is orthogonal.*

Conversely, for any real symmetric matrix $\mathbf{A}$ *there exists an orthogonal matrix* $\mathbf{O}$ *such that*

$$\mathbf{D} = \mathbf{O}'\mathbf{A}\mathbf{O}$$

is diagonal.

(The reason for the name of the theorem is that the set of eigenvalues is sometimes called the *spectrum* of $\mathbf{A}$). The spectral decomposition theorem says nothing about eigenvalues and eigenvectors, but we know from the discussion above that the diagonal elements of $\mathbf{D}$ are the eigenvalues of $\mathbf{A}$, and the columns of $\mathbf{O}$ are the corresponding eigenvectors.

E.3 Positive Definite Matrices

Using the spectral theorem, we can prove several interesting things about positive definite matrices.

Corollary E.5. *A real symmetric matrix* $\mathbf{A}$ *is positive semi-definite if and only if its spectrum is nonnegative. A real symmetric matrix* $\mathbf{A}$ *is positive definite if and only if its spectrum is strictly positive.*

Proof. First suppose that $\mathbf{A}$ is positive semi-definite with spectral decomposition (E.7). Let $\mathbf{e}_i$ denote the vector having elements that are all zero except the i-th, which is one, and define $\mathbf{w} = \mathbf{O}\mathbf{e}_i$, so

$$0 \le \mathbf{w}'\mathbf{A}\mathbf{w} = \mathbf{e}_i'\mathbf{O}'\mathbf{O}\mathbf{D}\mathbf{O}'\mathbf{O}\mathbf{e}_i = \mathbf{e}_i'\mathbf{D}\mathbf{e}_i = d_{ii} \tag{E.8}$$

using $\mathbf{O}'\mathbf{O} = I$. Hence the spectrum is nonnegative.

Conversely, suppose the d_{ii} are nonnegative. Then for any vector $\mathbf{w}$ define $\mathbf{z} = \mathbf{O}'\mathbf{w}$, so

$$\mathbf{w}'\mathbf{A}\mathbf{w} = \mathbf{w}'\mathbf{O}\mathbf{D}\mathbf{O}'\mathbf{w} = \mathbf{z}'\mathbf{D}\mathbf{z} = \sum_i d_{ii} z_i^2 \ge 0$$

Hence $\mathbf{A}$ is positive semi-definite.

The assertions about positive definiteness are proved in almost the same way. Suppose that $\mathbf{A}$ is positive definite. Since $\mathbf{e}_i$ is nonzero, $\mathbf{w}$ in (E.8) is also nonzero because $\mathbf{e}_i = \mathbf{O}'\mathbf{w}$ would be zero (and it isn't) if $\mathbf{w}$ were zero. Thus the inequality in (E.8) is actually strict. Hence the spectrum of is strictly positive.

Conversely, suppose the d_{ii} are strictly positive. Then for any nonzero vector $\mathbf{w}$ define $\mathbf{z} = \mathbf{O}'\mathbf{w}$ as before, and again note that $\mathbf{z}$ is nonzero because $\mathbf{w} = \mathbf{O}\mathbf{z}$ and $\mathbf{w}$ is nonzero. Thus $\mathbf{w}'\mathbf{A}\mathbf{w} = \mathbf{z}'\mathbf{D}\mathbf{z} > 0$, and hence $\mathbf{A}$ is positive definite. □

Corollary E.6. *A positive semi-definite matrix is invertible if and only if it is positive definite.*

Proof. It is easily verified that the product of diagonal matrices is diagonal and the diagonal elements of the product are the products of the diagonal elements of the multiplicands. Thus a diagonal matrix $\mathbf{D}$ is invertible if and only if all its diagonal elements d_{ii} are nonzero, in which case $\mathbf{D}^{-1}$ is diagonal with diagonal elements $1/d_{ii}$.

Since $\mathbf{O}$ and $\mathbf{O}'$ in the spectral decomposition (E.7) are invertible, $\mathbf{A}$ is invertible if and only if $\mathbf{D}$ is, hence if and only if its spectrum is nonzero, in which case

$$\mathbf{A}^{-1} = \mathbf{O}\mathbf{D}^{-1}\mathbf{O}'.$$

By the preceding corollary the spectrum of a positive semi-definite matrix is nonnegative, hence nonzero if and only if strictly positive, which (again by the preceding corollary) occurs if and only if the matrix is positive definite. □

Corollary E.7. *Every real symmetric positive semi-definite matrix* $\mathbf{A}$ *has a symmetric square root*

$$\mathbf{A}^{1/2} = \mathbf{O}\mathbf{D}^{1/2}\mathbf{O}' \tag{E.9}$$

where (E.7) *is the spectral decomposition of* $\mathbf{A}$ *and where* $\mathbf{D}^{1/2}$ *is defined to be the diagonal matrix whose diagonal elements are* $\sqrt{d_{ii}}$*, where* d_{ii} *are the diagonal elements of* $\mathbf{D}$.

Moreover, $\mathbf{A}^{1/2}$ *is positive definite if and only if* $\mathbf{A}$ *is positive definite.*

Note that by Corollary E.5 all of the diagonal elements of $\mathbf{D}$ are nonnegative and hence have real square roots.

Proof.

$$\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{O}\mathbf{D}^{1/2}\mathbf{O}'\mathbf{O}\mathbf{D}^{1/2}\mathbf{O}' = \mathbf{O}\mathbf{D}^{1/2}\mathbf{D}^{1/2}\mathbf{O}' = \mathbf{O}\mathbf{D}\mathbf{O}' = \mathbf{A}$$

because $\mathbf{O}'\mathbf{O} = \mathbf{I}$ and $\mathbf{D}^{1/2}\mathbf{D}^{1/2} = \mathbf{D}$.

From Corollary E.5 we know that $\mathbf{A}$ is positive definite if and only if all the d_{ii} are strictly positive. Since (E.9) is the spectral decomposition of $\mathbf{A}^{1/2}$, we see that $\mathbf{A}^{1/2}$ is positive definite if and only if all the $\sqrt{d_{ii}}$ are strictly positive. Clearly $d_{ii} > 0$ if and only if $\sqrt{d_{ii}} > 0$. □

Appendix F

Normal Approximations for Distributions

F.1 Binomial Distribution

The $\mathrm{Bin}(n, p)$ distribution is approximately normal with mean np and variance $np(1-p)$ if n is large.

F.2 Negative Binomial Distribution

The $\mathrm{NegBin}(n, p)$ distribution is approximately normal with mean n/p and variance $n(1-p)/p^2$ if n is large.

F.3 Poisson Distribution

The $\mathrm{Poi}(\mu)$ distribution is approximately normal with mean μ and variance μ if μ is large.

F.4 Gamma Distribution

The $\mathrm{Gam}(\alpha, \lambda)$ distribution is approximately normal with mean α/λ and variance α/λ^2 if α is large.

F.5 Chi-Square Distribution

The $\mathrm{chi}^2(n)$ distribution is approximately normal with mean n and variance $2n$ if n is large.

Chapter 8

Convergence Concepts Continued

8.1 Multivariate Convergence Concepts

When we covered convergence concepts before (Chapter 6 of these notes), we only did the scalar case because of the semester system. Logically, this chapter goes with Chapter 6, but if we had done it then, students transferring into this section this semester would be lost because the material isn't in Lindgren. Then we only covered convergence in probability and in distribution of *scalar* random variables. Now we want to cover the same ground but this time for random *vectors*. It will also be a good review.

8.1.1 Convergence in Probability to a Constant

Recall that *convergence in probability to a constant* has a definition (Definition 6.1.2 in Chapter 6 of these notes), but we never used the definition. Instead we obtained all of our convergence in probability results, either directly or indirectly from the law of large numbers (LLN).

Now we want to discuss convergence of random vectors, which we can also call *multivariate convergence in probability to a constant.* It turns out, that the multivariate concept is a trivial generalization of the univariate concept.

Definition 8.1.1 (Convergence in Probability to a Constant).
A sequence of random vectors

$$\mathbf{X}_n = (X_{n1}, \ldots, X_{nm}), \qquad n = 1, 2, \ldots$$

converges in probability to a constant vector

$$\mathbf{a} = (a_1, \ldots, a_m)$$

written

$$\mathbf{X}_n \xrightarrow{P} \mathbf{a}, \qquad \textit{as } n \to \infty$$

if the corresponding components converge in probability, that is, if

$$X_{ni} \xrightarrow{P} a_i, \qquad \text{as } n \to \infty$$

for each i.

The reader should be warned that this isn't the usual definition, but it is equivalent to the usual definition (we have defined the usual concept but not in the usual way).

8.1.2 The Law of Large Numbers

The componentwise nature of convergence in probability to a constant makes the multivariate law of large numbers a trivial extension of the univariate law (Theorem 6.3 in Chapter 6 of these notes).

Theorem 8.1 (Multivariate Law of Large Numbers). *If $\mathbf{X}_1$, $\mathbf{X}_2$, ... is a sequence of independent, identically distributed random vectors having mean vector $\boldsymbol{\mu}$, and*

$$\overline{\mathbf{X}}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i$$

is the sample mean for sample size n, then

$$\overline{\mathbf{X}}_n \xrightarrow{P} \boldsymbol{\mu}, \qquad \text{as } n \to \infty. \tag{8.1}$$

The only requirement is that the mean $\boldsymbol{\mu}$ exist. No other property of the distribution of the $\mathbf{X}_i$ matters.

We will use the abbreviation LLN for either theorem. The multivariate theorem is only interesting for giving us a notational shorthand that allows us to write the law of large numbers for all the components at once. It has no mathematical content over and above the univariate LLN's for each component. Convergence in probability (to a constant) of random vectors says no more than the statement that each component converges. In the case of the LLN, each statement about a component is just the univariate LLN.

8.1.3 Convergence in Distribution

Convergence in distribution is different. Example 8.1.1 below will show that, unlike convergence in probability to a constant, convergence in distribution for random vectors is not just convergence in distribution of each component.

Univariate convergence in distribution has a definition (Theorem 6.1.1 of these notes), but the definition was not used except in Problem 7-7 in Chapter 7 of these notes, which is an odd counterexample to the usual behavior of statistical estimators.

Instead we obtained all of our convergence in distribution results, either directly or indirectly, from the central limit theorem (CLT), which is Theorem 6.2 of Chapter 6 of these notes. Multivariate convergence in distribution has a definition is much the same, we will obtain almost all such results directly or indirectly from the (multivariate) CLT. Hence here we define multivariate convergence in distribution in terms of univariate convergence in distribution.

Definition 8.1.2 (Convergence in Distribution).
A sequence of random vectors $\mathbf{X}_1$*,* $\mathbf{X}_2$*, ...* converges in distribution *to a random vector* $\mathbf{X}$*, written*

$$\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{X}, \qquad \textit{as } n \to \infty.$$

if

$$\mathbf{t}'\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{t}'\mathbf{X}, \qquad \textit{for all constant vectors } \mathbf{t}. \tag{8.2}$$

Again, the reader should be warned that this isn't the usual definition, but it is equivalent to the usual definition (we have defined the usual concept but not in the usual way, the equivalence of our definition and the usual definition is called the Cramér-Wold Theorem).

This shows us in what sense the notion of multivariate convergence in distribution is determined by the univariate notion. The multivariate convergence in distribution $\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{X}$ happens if and only if the univariate convergence in distribution $\mathbf{t}'\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{t}'\mathbf{X}$ happens for every constant vector $\mathbf{t}$. The following example shows that convergence in distribution of each component of a random vector is not enough to imply convergence of the vector itself.

Example 8.1.1.
Let $\mathbf{X}_n = (U_n, V_n)$ be defined as follows. Define U_n to be standard normal for all n. Then trivially $U_n \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$. Define $V_n = (-1)^n U_n$ for all n. Then V_n is also standard normal for all n, and hence trivially $V_n \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$. Thus both components of $\mathbf{X}_n$ converge in distribution. But if $\mathbf{t}' = \begin{pmatrix} 1 & 1 \end{pmatrix}$, then

$$\mathbf{t}'\mathbf{X}_n = U_n + V_n = \begin{cases} 2U_n, & n \text{ even} \\ 0, & n \text{ odd} \end{cases}$$

This clearly does not converge in distribution, since the even terms all have one distribution, $\mathcal{N}(0,4)$ (and hence trivially converge to that distribution), and the odd terms all have another, the distribution concentrated at zero (and hence trivially converge to that distribution).

Thus, unlike convergence in probability to a constant, multivariate convergence in distribution entails more than univariate convergence of each component. Another way to say the same thing is that *marginal* convergence in distribution does not imply *joint* convergence in distribution. Of course, the converse does hold: joint convergence in distribution does imply marginal convergence in distribution (just take a vector $\mathbf{t}$ in the definition having all but one component equal to zero).

8.1.4 The Central Limit Theorem

We can now derive the multivariate central limit theorem from the univariate theorem (Theorem 6.2 of Chapter 6 of these notes).

Theorem 8.2 (Multivariate Central Limit Theorem). *If $\mathbf{X}_1$, $\mathbf{X}_2$, ... is a sequence of independent, identically distributed random vectors having mean vector $\boldsymbol{\mu}$ and variance matrix $\mathbf{M}$ and*

$$\overline{\mathbf{X}}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i$$

is the sample mean for sample size n, then

$$\sqrt{n}\left(\overline{\mathbf{X}}_n - \boldsymbol{\mu}\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{M}). \tag{8.3}$$

The only requirement is that second moments (the elements of $\mathbf{M}$) exist (this implies first moments also exist by Theorem 2.44 of Chapter 2 of these notes). No other property of the distribution of the $\mathbf{X}_i$ matters.

We often write the univariate CLT as

$$\overline{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \tag{8.4}$$

and the multivariate CLT as

$$\overline{\mathbf{X}}_n \approx \mathcal{N}\left(\boldsymbol{\mu}, \frac{\mathbf{M}}{n}\right) \tag{8.5}$$

These are simpler to interpret (though less precise and harder to use theoretically). We often say that the right hand side of one of these equations is the *asymptotic distribution* of the left hand side.

Derivation of the multivariate CLT from the univariate. For each constant vector $\mathbf{t}$, the scalar random variables $\mathbf{t}'(\mathbf{X}_n - \boldsymbol{\mu})$ have mean 0 and variance $\mathbf{t}'\mathbf{M}\mathbf{t}$ and hence obey the univariate CLT

$$\sqrt{n}\mathbf{t}'\left(\overline{\mathbf{X}}_n - \boldsymbol{\mu}\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{t}'\mathbf{M}\mathbf{t}).$$

The right hand side is the distribution of $\mathbf{t}'\mathbf{Z}$ where $\mathbf{Z} \sim \mathcal{N}(0, \mathbf{M})$, hence (8.3) follows by our definition of multivariate convergence in distribution. □

Example 8.1.2.
(This continues Example 5.1.1.) Let X_1, X_2, ... be a sequence of i. i. d. random variables, and define random vectors

$$\mathbf{Z}_i = \begin{pmatrix} X_i \\ X_i^2 \end{pmatrix}$$

Then $\mathbf{Z}_1$, $\mathbf{Z}_2$, ... is a sequence of i. i. d. random vectors having mean vector $\boldsymbol{\mu}$ given by (5.6) and variance matrix $\mathbf{M}$ given by (5.7), and the CLT applies

$$\sqrt{n}\left(\overline{\mathbf{Z}}_n - \boldsymbol{\mu}\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{M}).$$

8.1.5 Slutsky and Related Theorems

As with univariate convergence in distribution, we are forced to state a number of theorems about multivariate convergence in distribution without proof. The proofs are just too hard for this course.

Theorem 8.3. *If* $\mathbf{a}$ *is a constant, then* $\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{a}$ *if and only if* $\mathbf{X}_n \xrightarrow{P} \mathbf{a}$.

Thus, as was true in the univariate case, convergence in probability to a constant and convergence in distribution to a constant are equivalent concepts. We could dispense with one, but tradition and usage do not allow it. We must be able to recognize both in order to read the literature.

Theorem 8.4 (Slutsky). *If* $g(\mathbf{x}, \mathbf{y})$ *is a function jointly continuous at every point of the form* $(\mathbf{x}, \mathbf{a})$ *for some fixed* $\mathbf{a}$*, and if* $\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{X}$ *and* $\mathbf{Y}_n \xrightarrow{P} \mathbf{a}$*, then*

$$g(\mathbf{X}_n, \mathbf{Y}_n) \xrightarrow{\mathcal{D}} g(\mathbf{X}, \mathbf{a}).$$

The function g here can be either scalar or vector valued. The continuity hypothesis means that $g(\mathbf{x}_n, \mathbf{y}_n) \to g(\mathbf{x}, \mathbf{a})$ for all nonrandom sequences $\mathbf{x}_n \to \mathbf{x}$ and $\mathbf{y}_n \to \mathbf{a}$.

Sometimes Slutsky's theorem is used in a rather trivial way with the sequence "converging in probability" being nonrandom. This uses the following lemma.

Lemma 8.5. *If* $\mathbf{a}_n \to \mathbf{a}$ *considered as a nonrandom sequence, then* $\mathbf{a}_n \xrightarrow{P} \mathbf{a}$ *considered as a sequence of constant random vectors.*

This is an obvious consequence of the definition of convergence in probability (Definition 6.1.2 in Chapter 6 of these notes).

Example 8.1.3.
The so-called sample variance S_n^2 defined on p. 204 in Lindgren or in (7.17) in Chapter 7 of these notes is asymptotically equivalent to the variance of the empirical distribution V_n also defined on p. 204 in Lindgren or in (7.4) in Chapter 7 of these notes. The two estimators are related by

$$S_n^2 = \frac{n}{n-1} V_n.$$

We know the asymptotics of V_n because it is a sample moment. By Theorem 7.15 of Chapter 7 of these notes

$$V_n \xrightarrow{P} \sigma^2 \tag{8.6}$$

and by Theorem 7.16 of Chapter 7 of these notes

$$\sqrt{n}(V_n - \sigma^2) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mu_4 - \mu_2^2) \tag{8.7}$$

(strictly speaking, we don't actually know this last fact yet, because we haven't proved Theorem 7.16 yet, but we will).

One application of Slutsky's theorem shows that

$$S_n^2 \xrightarrow{P} \sigma^2 \tag{8.8}$$

because

$$\frac{n}{n-1} V_n \xrightarrow{P} 1 \cdot \sigma^2$$

because of (8.6) and

$$\frac{n}{n-1} \to 1$$

and Lemma 8.5.

To get the next level of asymptotics we write

$$\begin{aligned} \sqrt{n}(S_n^2 - \sigma^2) &= \sqrt{n}\left(\frac{n}{n-1} V_n - \sigma^2\right) \\ &= \frac{n}{n-1}\left[\sqrt{n}\left(V_n - \sigma^2\right) + \frac{\sqrt{n}}{n-1}\sigma^2\right] \end{aligned} \tag{8.9}$$

Then two applications of Slutsky's theorem give us what we want. Let W be a random variable having the distribution of the right hand side of (8.7) so that equation can be rewritten

$$\sqrt{n}(V_n - \sigma^2) \xrightarrow{\mathcal{D}} W.$$

Then one application of Slutsky's theorem (and the corollary following it in Chapter 6 of these notes) shows that the term in square brackets in (8.9) also converges to W

$$\sqrt{n}\left(V_n - \sigma^2\right) + \frac{\sqrt{n}}{n-1}\sigma^2 \xrightarrow{\mathcal{D}} W + 0$$

because of

$$\frac{\sqrt{n}}{n-1}\sigma^2 \to 0$$

and Lemma 8.5. Then another application of Slutsky's theorem shows what we want

$$\frac{n}{n-1} \times \text{term in square brackets} \xrightarrow{\mathcal{D}} 1 \cdot W.$$

The special cases of Slutsky's theorem which we only have only one sequence of random variables converging in distribution or in probability are known as "continuous mapping theorems."

Theorem 8.6 (Continuous Mapping, Convergence in Distribution). *If g is an everywhere continuous function and $\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{X}$, then $g(\mathbf{X}_n) \xrightarrow{\mathcal{D}} g(\mathbf{X})$.*

The function g here can be either scalar or vector valued. The only requirement is that it be continuous, that is, $g(\mathbf{x}_n) \to g(\mathbf{x})$ for any point $\mathbf{x}$ and any sequence $\mathbf{x}_n \to \mathbf{x}$.

Theorem 8.7 (Continuous Mapping, Convergence in Probability). *If g is a function continuous at the point $\mathbf{a}$ and $\mathbf{X}_n \xrightarrow{P} \mathbf{a}$, then $g(\mathbf{X}_n) \xrightarrow{P} g(\mathbf{a})$.*

The function g here can be either scalar or vector valued. The only requirement is that it be continuous at $\mathbf{a}$, that is, $g(\mathbf{x}_n) \to g(\mathbf{a})$ for any sequence $\mathbf{x}_n \to \mathbf{a}$.

Example 8.1.4.
Suppose (8.8) holds for some sequence of random variables S_n^2 and $\sigma > 0$, then the continuous mapping theorem for convergence in probability immediately gives many other convergence in probability results, for example,

$$\begin{aligned} S_n &\xrightarrow{P} \sigma \qquad (8.10) \\ \frac{1}{S_n} &\xrightarrow{P} \frac{1}{\sigma} \\ \log(S_n) &\xrightarrow{P} \log(\sigma) \end{aligned}$$

All of these applications are fairly obvious. These conclusions seem so natural that it is hard to remember that we need the continuous mapping theorem to tell us that they hold.

We will use the continuous mapping theorem for convergence in probability many times in our study of statistics. In contrast, our uses of the continuous mapping theorem for convergence in distribution will all be rather trivial. We will only use it to see that we can divide both sides of a convergence in distribution statement by the same constant, or add the same constant to both sides, and so forth.

Example 8.1.5.
This was assigned for homework (Problem 6-2 of Chapter 6 of these notes) last semester. We will see many other examples later, but all will be similar to this one, which is by far the most important in statistics. Suppose X_1, X_2, $\ldots$ are i. i. d. random variables with mean μ and variance σ^2, suppose that S_n is any sequence of random variables satisfying (8.10), and suppose $\sigma > 0$, then

$$\frac{\overline{X}_n - \mu}{S_n / \sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1). \qquad (8.11)$$

The most obvious choice of a sequence S_n satisfying (8.10) is the sample standard deviation. That's what Examples 8.1.3 and 8.1.4 showed. But the exact way S_n is defined isn't important for this example. In fact there are many sequences of random variables having this property. The only thing that is important is that such sequences exist.

How do we show (8.11) using the CLT and Slutsky's theorem? First the CLT says

$$\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow{\mathcal{D}} Y$$

where $Y \sim \mathcal{N}(0, \sigma^2)$. Define the function

$$g(u, v) = \frac{u}{v}.$$

This is continuous everywhere except where $v = 0$, where it is undefined. Now define

$$U_n = \sqrt{n}\left(\overline{X}_n - \mu\right)$$

and apply Slutsky's theorem to $g(U_n, S_n)$. The first argument converges in distribution and the second argument converges to a constant, so Slutsky's theorem does hold and says

$$\sqrt{n}\frac{\overline{X}_n - \mu}{S_n} = g(U_n, S_n) \xrightarrow{\mathcal{D}} g(Y, \sigma) = \frac{Y}{\sigma}$$

and the right hand side does have a standard normal distribution, as asserted, by the rule giving the variance of a linear transformation (5.15b).

8.2 The Delta Method

8.2.1 The Univariate Delta Method

Suppose T_n is any sequence of random variables converging in probability to a constant θ. Many such examples arise in statistics. Particular cases are $\overline{X}_n \xrightarrow{P} \mu$ (which is the LLN) and $S_n \xrightarrow{P} \sigma$ (which we showed in Examples 8.1.3 and 8.1.4). It is conventional in statistics to use T_n as a default notation for all such sequences and θ as a default notation for all the constants.

The continuous mapping theorem for convergence in probability tells us that $g(T_n) \xrightarrow{P} g(\theta)$ for any function g that is continuous at the point θ. Many different functions g arise in applications. The continuity requirement is not very restrictive. Almost any function will do.

What this says is that $g(T_n)$ gets closer and closer to $g(\theta)$ as n gets large. The obvious next question is "How close?" We want a statement analogous to the CLT that tells us the distribution of the "error" $g(T_n) - g(\theta)$. The continuous mapping theorem for convergence in distribution doesn't do this. It would tell us, for example, that

$$g\Big(\sqrt{n}\left(\overline{X}_n - \mu\right)\Big) \xrightarrow{\mathcal{D}} g(Y) \tag{8.12a}$$

where $Y \sim \mathcal{N}(0, \sigma^2)$, but that's not what we want. We want a convergence in distribution result for

$$\sqrt{n}\big(g(T_n) - g(\theta)\big). \tag{8.12b}$$

If g is a linear function, then (8.12b) and the left hand side of (8.12a) are equal. If g is not linear, then they aren't, and the continuous mapping theorem is of no use. The delta method does do what we want: a convergence in distribution result for (8.12b).

Theorem 8.8 (Univariate Delta Method). *Suppose*

$$\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{D}} Y \tag{8.13}$$

and g is any function differentiable at θ, then

$$\sqrt{n}\bigl(g(T_n) - g(\theta)\bigr) \xrightarrow{\mathcal{D}} g'(\theta)Y. \tag{8.14}$$

By far the most important applications of the delta method have Y normally distributed with mean zero, say $Y \sim \mathcal{N}(0, \sigma_Y^2)$. In that case, we can put (8.14) in "sloppy form" with "double squiggle" notation like (8.4) or (8.5). It becomes

$$g(T_n) \approx \mathcal{N}\left(g(\theta), \frac{g'(\theta)^2\sigma_Y^2}{n}\right)$$

and we say that the right hand side is the *asymptotic distribution* of $g(T_n)$.

It is called the "delta method" because of the important role played by the derivative. The "delta" is supposed to remind you of

$$\frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}$$

the triangles being capital Greek letter deltas, and the fraction on the right being pronounced "delta y over delta x." The earlier term for this concept, used throughout the nineteenth century and still used by some people, was "propagation of errors."

It is important to understand that the delta method does not produce a convergence in distribution result out of thin air. It turns one convergence in distribution statement (8.13) into another (8.14). In order to use the delta method we must already have one convergence in distribution result. Usually that comes either from the CLT or from a previous application of the delta method.

Example 8.2.1.
Suppose X_1, X_2, ... are i. i. d. Exp(λ). Then

$$\begin{aligned} E(X_i) &= \frac{1}{\lambda} \\ \operatorname{var}(X_i) &= \frac{1}{\lambda^2} \end{aligned}$$

the LLN says

$$\overline{X}_n \xrightarrow{P} \frac{1}{\lambda}$$

and the CLT says

$$\sqrt{n}\left(\overline{X}_n - \frac{1}{\lambda}\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{1}{\lambda^2}\right) \tag{8.15}$$

That's all well and good, but it seems more natural to look at the sequence of random variables

$$W_n = \frac{1}{\overline{X}_n} \tag{8.16}$$

because then the continuous mapping theorem for convergence in probability gives

$$W_n \xrightarrow{P} \lambda.$$

So what is the asymptotic distribution of W_n?

We want to apply the delta method. To do that we already need one convergence in distribution result. What we have is the CLT (8.15). This tells we want to use the delta method with $T_n = \overline{X}_n$ and $\theta = 1/\lambda$. Then, since we want $g(T_n) = W_n$, we must have

$$g(t) = \frac{1}{t}$$

and hence

$$g'(t) = -\frac{1}{t^2}$$

So

$$g(\theta) = \lambda$$

and

$$g'(\theta) = -\lambda^2.$$

And the delta method tells us that W_n is asymptotically normally distributed with mean λ and variance

$$\frac{g'(\theta)^2\sigma^2}{n} = \frac{\left(-\lambda^2\right)^2}{n\lambda^2} = \frac{\lambda^2}{n}$$

The argument is a bit involved, but in the end we arrive at the fairly simple statement

$$W_n \approx \mathcal{N}\left(\lambda, \frac{\lambda^2}{n}\right).$$

Proof of the Univariate Delta Method. By definition, the derivative is

$$g'(\theta) = \lim_{t \to \theta} \frac{g(t) - g(\theta)}{t - \theta}$$

To be useful in our proof we need to rewrite this slightly. For $t \neq \theta$ define the function

$$w(t) = \frac{g(t) - g(\theta)}{t - \theta} - g'(\theta) \tag{8.17}$$

then the definition of differentiation says that $w(t) \to 0$ as $t \to \theta$, which is the same thing as saying that w is continuous at the point θ if we define $w(\theta) = 0$. (The reason for phrasing things in terms of continuity rather than limits is because Slutsky's theorem uses continuity.) Then (8.17) can be rewritten as

$$g(t) - g(\theta) = g'(\theta)(t - \theta) + w(t)(t - \theta). \tag{8.18}$$

Now plug in T_n for t and multiply by $\sqrt{n}$ giving

$$\sqrt{n}\bigl(g(T_n) - g(\theta)\bigr) = g'(\theta)\sqrt{n}(T_n - \theta) + w(T_n)\sqrt{n}(T_n - \theta).$$

By the continuous mapping theorem, the first term on the right hand side converges in distribution to $g'(\theta)Y$. By Slutsky's theorem, the second term on the right hand side converges in distribution to $w(\theta)Y = 0$. Hence by another application of Slutsky's theorem, the right hand side converges to $g'(\theta)Y$, which is the assertion of the delta method. □

8.2.2 The Multivariate Delta Method

The multivariate delta method is a straightforward extension of the univariate delta method, obvious if you know about derivatives of general vector-valued functions. You already know this material because it was used in the change of variable theorem for random vectors (Section 1.6.2 in Chapter 1 of these notes). You may need to go back and review that.

In brief, that section introduced the derivative of a vector-valued function of a vector variable $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^m$, which can also be thought of a a vector of scalar-valued functions

$$\mathbf{g}(\mathbf{x}) = \begin{pmatrix} g_1(\mathbf{x}) \\ \vdots \\ g_m(\mathbf{x}) \end{pmatrix}$$

The derivative of the function $\mathbf{g}$ at the point $\mathbf{x}$ (assuming it exists) is the matrix of partial derivatives. It is written $\nabla\mathbf{g}(\mathbf{x})$ and pronounced "del g of x." Throughout this section we will also write it as the single letter $\mathbf{G}$. So

$$\mathbf{G} = \nabla\mathbf{g}(\mathbf{x})$$

is the matrix with elements

$$g_{ij} = \frac{\partial g_i(\mathbf{x})}{\partial x_j}$$

Note that if $\mathbf{g}$ maps n-dimensional vectors to m-dimensional vectors, then it is an $m \times n$ matrix (rather than the other way around). A concrete example that may help you visualize the idea is Example 1.6.1 in Chapter 1 of these notes.

Theorem 8.9 (Multivariate Delta Method). *Suppose*

$$\sqrt{n}(\mathbf{T}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} \mathbf{Y} \tag{8.19}$$

and $\mathbf{g}$ *is any function differentiable*[1] *at* $\boldsymbol{\theta}$*, then*

$$\sqrt{n}\bigl(\mathbf{g}(\mathbf{T}_n) - \mathbf{g}(\boldsymbol{\theta})\bigr) \xrightarrow{\mathcal{D}} \nabla\mathbf{g}(\boldsymbol{\theta})\mathbf{Y}. \tag{8.21}$$

[1] The notion of multivariate differentiability of is actually a bit complicated. We presented only a simplified version of the facts, which is not completely correct. Here are the facts. Most readers will only want to know the first item below, maybe the second. The third is the pedantically correct mathematical definition of multivariate differentiability, which is of theoretical interest only. It won't help you do any problems. You are free to ignore it.

By far the most important applications of the delta method have $\mathbf{Y}$ normally distributed with mean zero, say $\mathbf{Y} \sim \mathcal{N}(0, \mathbf{M}_\mathbf{Y})$. In that case, we can put (8.21) in "sloppy form" with "double squiggle" notation like (8.4) or (8.5). It becomes

$$\mathbf{g}(\mathbf{T}_n) \approx \mathcal{N}\left(\mathbf{g}(\boldsymbol{\theta}), \frac{\mathbf{G}\mathbf{M}_\mathbf{Y}\mathbf{G}'}{n}\right),$$

where, as we said we would, we are now defining $\mathbf{G} = \nabla \mathbf{g}(\boldsymbol{\theta})$ to simplify notation. We say that the right hand side is the *asymptotic distribution* of $\mathbf{g}(\mathbf{T}_n)$.

Example 8.2.2.
Suppose

$$\mathbf{Z}_i = \begin{pmatrix} X_i \\ Y_i \end{pmatrix}, \qquad i = 1, 2, \dots$$

are an i. i. d. sequence of random vectors with mean vector $\boldsymbol{\mu}$ and variance matrix $\mathbf{M}$. Suppose we are interested in the parameter

$$\omega = \log\left(\frac{\mu_1}{\mu_2}\right) = \log(\mu_1) - \log(\mu_2)$$

The continuous mapping theorem applied to the LLN gives

$$W_n = \log(\overline{X}_n) - \log(\overline{Y}_n) \xrightarrow{P} \omega$$

and we want to use the delta method to find the asymptotic distribution of the difference $W_n - \omega$. The "g" involved is

$$g(x, y) = \log(x) - \log(y)$$

which has partial derivatives

$$\frac{\partial g(x,y)}{\partial x} = \frac{1}{x}$$
$$\frac{\partial g(x,y)}{\partial y} = -\frac{1}{y}$$

1. If a function is differentiable, then the derivative is the matrix of partial derivatives.
2. If the partial derivatives exist and are continuous, then the function is differentiable.
3. A function can be differentiable without the partial derivatives being continuous. The exact condition required is the multivariate analog of (8.18) in the proof of the univariate delta method

$$\mathbf{g}(\mathbf{t}) - \mathbf{g}(\boldsymbol{\theta}) = \mathbf{G}(\mathbf{t} - \boldsymbol{\theta}) + \|\mathbf{t} - \boldsymbol{\theta}\|\mathbf{w}(\mathbf{t}) \tag{8.20}$$

where the double vertical bars indicate the norm of a vector

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}'\mathbf{x}} = \sqrt{\sum_{i=1}^{n} x_i^2}$$

A function $\mathbf{g}$ is differentiable at $\boldsymbol{\theta}$ if there exists a matrix $\mathbf{G}$ and a function $\mathbf{w}$ that is continuous at $\boldsymbol{\theta}$ with $\mathbf{w}(\boldsymbol{\theta}) = 0$ such that (8.20) holds, in which case $\mathbf{G}$ is the derivative matrix $\nabla \mathbf{g}(\boldsymbol{\theta})$.

The nice thing about this definition, pedantic though it may be, is that it makes the proof of the multivariate delta method just like the proof of the univariate proof. Start from (8.20) and proceed just like the univariate proof, changing notation as necessary.

Thus the derivative matrix is

$$\nabla g(x, y) = \begin{pmatrix} \frac{1}{x} & -\frac{1}{y} \end{pmatrix}$$

Evaluating at $\boldsymbol{\mu}$, we get

$$\mathbf{G} = \begin{pmatrix} \frac{1}{\mu_1} & -\frac{1}{\mu_2} \end{pmatrix}$$

As always, the asymptotic distribution produced by the delta method has mean $g(\boldsymbol{\mu}) = \omega$ and variance $\mathbf{G}\mathbf{M}\mathbf{G}'/n$. We just have to work out the latter

$$\begin{pmatrix} \frac{1}{\mu_1} & -\frac{1}{\mu_2} \end{pmatrix} \begin{pmatrix} m_{11} & m_{12} \\ m_{12} & m_{22} \end{pmatrix} \begin{pmatrix} \frac{1}{\mu_1} \\ -\frac{1}{\mu_2} \end{pmatrix} = \frac{m_{11}}{\mu_1^2} - \frac{2m_{12}}{\mu_1 \mu_2} + \frac{m_{22}}{\mu_2^2}$$

If you prefer to phrase everything in terms of the usual notation for the moments of the components X and Y, this becomes

$$\sigma_W^2 = \frac{\sigma_X^2}{\mu_X^2} - \frac{2\rho_{X,Y}\sigma_X\sigma_Y}{\mu_X\mu_Y} + \frac{\sigma_Y^2}{\mu_Y^2}$$

Thus the result of applying the delta method is

$$\sqrt{n}(W_n - \omega) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_W^2),$$

where the asymptotic variance σ_W^2 is defined above.

8.2.3 Asymptotics for Sample Moments

This section supplies the proof of Theorem 7.16 which we stated in Chapter 7 but could not prove because it requires the multivariate delta method.

Proof of Theorem 7.16. For ordinary moments, this is a homework problem (Problem 7-17 in Lindgren).

For $M_{k,n}$ we proceed as in the proof of Theorem 7.15, using (7.26b), which implies

$$\sqrt{n}(M_{k,n} - \mu_k) = \sqrt{n}(M'_{k,n} - \mu_k) + \sum_{j=1}^{k} \binom{k}{j} (-1)^j \sqrt{n}(\overline{X}_n - \mu)^j M'_{k-j,n}$$

the first term arising from the $j = 0$ term in (7.26b). Now the CLT says

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\mathcal{D}} Z$$

where $Z \sim \mathcal{N}(0, \mu_2)$, because $\mu_2 = \sigma^2$. Then Slutsky's theorem implies

$$\sqrt{n}(\overline{X}_n - \mu)^j \xrightarrow{\mathcal{D}} 0$$

for any $j > 1$ (Problem 7-15). Thus all the terms for $j > 1$ make no contribution to the asymptotics, and we only need to figure out the asymptotics of the sum of the first two ($j = 0$ and $j = 1$) terms

$$\sqrt{n}(M'_{k,n} - \mu_k) - k\sqrt{n}(\overline{X}_n - \mu)M'_{k-1,n}.$$

By Slutsky's theorem and (7.29) this converges to

$$W - k\mu_{k-1} Z \tag{8.22}$$

where W and Z are defined by the multivariate CLT

$$\sqrt{n} \begin{pmatrix} \overline{X}_n - \mu \\ M'_{k,n} - \mu_k \end{pmatrix} \xrightarrow{\mathcal{D}} \begin{pmatrix} Z \\ W \end{pmatrix} \sim \mathcal{N}(0, \mathbf{M})$$

where

$$\mathbf{M} = \operatorname{var} \begin{pmatrix} X_i - \mu \\ (X_i - \mu)^k \end{pmatrix} = \begin{pmatrix} \mu_2 & \mu_{k+1} \\ \mu_{k+1} & \mu_{2k} - \mu_k^2 \end{pmatrix}$$

Now calculating the variance of (8.22) using the usual formulas for the variance of a sum gives the asserted asymptotic variance in (7.31). □

8.2.4 Asymptotics of Independent Sequences

In several places throughout the course we will need the following result. In particular, we will use it in the section immediately following this one.

Theorem 8.10. *Suppose*

$$\begin{aligned} X_n &\xrightarrow{\mathcal{D}} X \\ Y_n &\xrightarrow{\mathcal{D}} Y \end{aligned} \tag{8.23}$$

and all of the X_i are independent of all of the Y_i, and suppose

$$\begin{aligned} k_n &\to \infty \\ m_n &\to \infty \end{aligned}$$

Then

$$\begin{pmatrix} X_{k_n} \\ Y_{m_n} \end{pmatrix} \xrightarrow{\mathcal{D}} \begin{pmatrix} X \\ Y \end{pmatrix}$$

where the right hand side denotes the random vector having independent components having the same marginal distributions as the variables in (8.23)*.*

As with many of the theorems in this section, we omit the proof.[2] The theorem seems very obvious. In fact, the marginal laws must be as stated in the theorem by the continuous mapping theorem (the map that takes a vector to one of its components being continuous). So the only nontrivial assertion of the theorem is that the joint distribution of the limiting random variable has

[2] It It can be proved fairly easily from the relationship between characteristic functions and convergence in distribution, slightly misstated as Theorem 26 of Chapter 4 in Lindgren and the characteristic function uniqueness theorem, Theorem 25 of Chapter 4 in Lindgren, or more precisely from the multivariate versions of these theorems, but since we gave no proof of those theorems and didn't even state their multivariate versions, there seems no point in proofs using them.

independent components. That seems obvious. What else could happen? The only point of stating the theorem is to point out that this actually needs a proof, which is given in texts on advanced probability theory.

The conclusion of the theorem is sometimes stated a bit less precisely as

$$\begin{pmatrix} X_k \\ Y_m \end{pmatrix} \xrightarrow{\mathcal{D}} \begin{pmatrix} X \\ Y \end{pmatrix} \qquad \text{as } k \to \infty \text{ and } m \to \infty$$

(you have to imagine the sequences k_n and m_n if you want a more precise statement), the point being that whenever k and m are both large the distribution of the left hand side is close to the distribution of the right hand side (so the latter can be used as an approximation of the former).

Corollary 8.11. *Under the hypotheses of the theorem*

$$X_k + Y_m \xrightarrow{\mathcal{D}} X + Y, \qquad \textit{as } k \to \infty \textit{ and } m \to \infty,$$

where X and Y are independent random variables having the same marginal distributions as the variables in (8.23)*.*

This follows directly from the theorem by the continuous mapping theorem for multivariate convergence in distribution (addition of components of a vector being a continuous operation).

8.2.5 Asymptotics of Sample Quantiles

In this section we give a proof of Theorem 7.27, which we were also unable to give in Chapter 7 because it too requires the multivariate delta method. We give a proof not because it represents a useful technique. The proof is a rather specialized trick that works only for this particular theorem. The reason we give the proof is to show how asymptotic normality arises even when there are no obvious averages anywhere in sight. After all, sample quantiles have nothing to do with any averages. Still, asymptotic normality arises anyway. This is typical. Most statistics that arise in practice are asymptotically normal whether or not there is any obvious connection with the CLT. There are exceptions (Problem 7-7), but they arise rarely in practice.

Before we begin the proof, we take a closer look at the relationship between the general case and the $\mathcal{U}(0, 1)$ case. It turns out that the latter can be derived from the former using the so-called quantile transformation.

Lemma 8.12. *Suppose X is a continuous random variable having an invertible c. d. f. F, then $F(X)$ has the $\mathcal{U}(0, 1)$ distribution. Conversely if $U \sim \mathcal{U}(0, 1)$, then $F^{-1}(U)$ has the same distribution as X.*

The first assertion is Theorem 9 of Chapter 3 in Lindgren. The second assertion is a special case of Problem 3-35 in Lindgren. The transformation $X = F^{-1}(U)$ is called the *quantile transformation* because it maps p to the p-th quantile x_p, and the transformation $U = F(X)$ is called the *inverse quantile* *transformation.* These transformations are a bit odd at first sight because they use F two different ways, both as a c. d. f. and as a change-of-variable function. From calculus, we know that these two transformations have derivatives that are inverses of each other, that is, if $u = F(x)$, so $x = F^{-1}(u)$, then

$$f(x) = F'(x) = \frac{d}{dx} F(x)$$

and

$$\frac{d}{du} F^{-1}(u) = \frac{1}{f(x)}. \tag{8.24}$$

Because we want to use the quantile transformation, we need to add an additional condition to the theorem, that the variables have an invertible c. d. f., which will be the case when $f(x) > 0$ for all x by Lemma 7.4 (the theorem is true without the additional condition, the proof is just a bit messier).

Proof of Theorem 7.27 assuming an invertible c. d. f. First we show how to derive the general case from the $\mathcal{U}(0, 1)$ case

$$\sqrt{n}\bigl(U_{(k_n)} - p\bigr) \xrightarrow{\mathcal{D}} \mathcal{N}\bigl(0, p(1-p)\bigr), \tag{8.25}$$

where the $U_{(k)}$ are the order statistics of a sample from the $\mathcal{U}(0, 1)$ distribution. Apply the quantile transformation so that $F^{-1}\bigl(U_{(k_n)}\bigr)$ has the same distribution as $X_{(k_n)}$ and apply the delta method with the derivative of the transformation given by (8.24). The result is assertion of the theorem (7.38). Thus it is enough to prove (8.25).

Now we use some facts about the relationships between various "brand name" distributions. The distribution of $U_{(k)}$ is $\text{Beta}(k, n-k+1)$. By Theorem 4.2, this distribution is the same as the distribution of $V_k/(V_k + W_k)$, where V_k and W_k are independent and

$$\begin{aligned} V_k &\sim \text{Gam}(k, \lambda) \\ W_k &\sim \text{Gam}(n-k+1, \lambda) \end{aligned}$$

where λ can have any value, for simplicity chose $\lambda = 1$. Then we use the normal approximation for the gamma distribution (Appendix C of these notes) which arises from the addition rule for the gamma distribution and the CLT

$$\begin{aligned} V_k &\approx \mathcal{N}(k, k) \\ W_k &\approx \mathcal{N}(n-k+1, n-k+1) \end{aligned}$$

So

$$\sqrt{n}\left(\frac{V_{k_n}}{n} - \frac{k_n}{n}\right) \approx \mathcal{N}\left(0, \frac{k_n}{n}\right)$$

and because of the assumption (7.37) and Slutsky's theorem we can replace k_n/n on both sides by p giving

$$\sqrt{n}\left(\frac{V_{k_n}}{n} - p\right) \approx \mathcal{N}\left(0, p\right),$$

and, similarly,

$$\sqrt{n}\left(\frac{W_{k_n}}{n} - (1-p)\right) \approx \mathcal{N}\left(0, 1-p\right).$$

Because of the assumed independence of V_k and W_k we can use Theorem 8.10 to get a multivariate CLT for the joint distribution of these two random vectors

$$\sqrt{n}\begin{pmatrix} \frac{V_{k_n}}{n} - p \\ \frac{W_{k_n}}{n} - (1-p) \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{M})$$

where

$$\mathbf{M} = \begin{pmatrix} p & 0 \\ 0 & 1-p \end{pmatrix}$$

Note that we can write this as

$$\sqrt{n}(\mathbf{T}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{M})$$

where

$$\mathbf{T}_n = \frac{1}{n}\begin{pmatrix} V_{k_n} \\ W_{k_n} \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\theta} = \begin{pmatrix} p \\ 1-p \end{pmatrix}$$

So we apply the multivariate delta method to this convergence in distribution result. Note that $g(V_{k_n}/n, W_{k_n}/n)$ has the same distribution as $U_{(k)}$. Hence we want to use the transformation

$$g(v, w) = \frac{v}{v+w},$$

which has partial derivatives

$$\begin{aligned} \frac{\partial g(v, w)}{\partial v} &= \frac{w}{(v+w)^2} \\ \frac{\partial g(v, w)}{\partial w} &= -\frac{v}{(v+w)^2} \end{aligned}$$

and derivative matrix

$$\mathbf{G} = \nabla g(\boldsymbol{\theta}) = \begin{pmatrix} 1-p & -p \end{pmatrix}$$

Thus, finally, we see that $U_{(k)}$ is asymptotically normal with mean

$$g(\boldsymbol{\theta}) = p$$

and variance

$$\mathbf{G}\mathbf{M}\mathbf{G}' = \begin{pmatrix} 1-p & -p \end{pmatrix}\begin{pmatrix} p & 0 \\ 0 & 1-p \end{pmatrix}\begin{pmatrix} 1-p \\ -p \end{pmatrix} = p(1-p)$$

and we are done. □

Problems

8-1. Suppose X is a random scalar with ordinary moments $\alpha_k = E(X^k)$.

(a) What are the mean vector and variance matrix of the random vector

$$\mathbf{Z} = \begin{pmatrix} X \\ X^2 \\ X^3 \end{pmatrix}$$

(b) Suppose $\mathbf{Z}_1$, $\mathbf{Z}_2$, $\ldots$ is an i. i. d. sequence of random vectors having the same distribution as $\mathbf{Z}$. What are the mean vector and variance matrix of $\overline{\mathbf{Z}}_n$?

8-2. Suppose Y is a random scalar having mean μ and variance σ^2 and $\mathbf{Z}$ is a random vector with i. i. d. components Z_i having mean zero and variance τ^2, and suppose also that Y is independent of $\mathbf{Z}$. Define $\mathbf{X} = Y + \mathbf{Z}$ (that is, $\mathbf{X}$ has components $X_i = Y + Z_i$).

(a) What are the mean vector and variance matrix of $\mathbf{X}$?

(b) Suppose $\mathbf{X}_1$, $\mathbf{X}_2$, $\ldots$ is an i. i. d. sequence of random vectors having the same distribution as $\mathbf{X}$. What is the asymptotic distribution of $\overline{\mathbf{X}}_n$?

8-3. Suppose

$$\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{D}} Y.$$

Use the delta method to find convergence in distribution results for

(a) $\log(T_n)$

(b) $\sqrt{T_n}$

(c) $\dfrac{\exp(T_n)}{1 + \exp(T_n)}$

Note: In (a) and (b) we need to assume $\theta > 0$.

8-4. Suppose X_1, X_2, X_3, $\ldots$ is an i. i. d. sequence of $\text{Ber}(p)$ random variables.

(a) State the LLN for $\overline{X}_n$, expressing all constants in terms of the parameter p (that is, don't use μ and σ, express them as functions of p).

(b) State the CLT for $\overline{X}_n$, expressing all constants in terms of the parameter p.

(c) To what does $\overline{X}_n(1 - \overline{X}_n)$ converge in probability? What theorem allows you to conclude this?

(d) Use the delta method to determine the asymptotic distribution of the random variable $\overline{X}_n(1 - \overline{X}_n)$. (Note: there is something funny about the case $p = 1/2$. Theorem 8.8 applies but its conclusion doesn't satisfactorily describe the "asymptotic distribution".)

8-5. Suppose $X_n \xrightarrow{\mathcal{D}} X$ where $X \sim \mathcal{N}(0, 1)$. To what does X_n^2 converge in distribution? What is the *name* of the limiting distribution (it is some "brand name" distribution). What theorem allows you to conclude this?

8-6. Suppose X_1, X_2, X_3, $\ldots$ is an i. i. d. sequence of random variables with mean μ and variance σ^2, and $\overline{X}_n$ is the sample mean. To what does

$$\sqrt{n} \left(\overline{X}_n - \mu \right)^2$$

converge in probability? (**Hint:** Use the CLT, the continuous mapping theorem for convergence in distribution, Slutsky's theorem, and Lemma 8.5.)

8-7. Suppose X_1, X_2, X_3, $\ldots$ is an i. i. d. sequence of random variables with mean μ and variance σ^2, and $\overline{X}_n$ is the sample mean. Define

$$Y_i = a + bX_i,$$

where a and b are constants, and

$$\overline{Y}_n = a + b\overline{X}_n.$$

Derive the asymptotic distribution of $\overline{Y}_n$ in two different ways.

(a) Use the delta method with $g(u) = a + bu$.

(b) Use the CLT applied to the sequence Y_1, Y_2, $\ldots$.

8-8. Suppose

$$\mathbf{Z}_i = \begin{pmatrix} X_i \\ Y_i \end{pmatrix}, \qquad i = 1, 2, \ldots$$

are an i. i. d. sequence of random vectors with mean vector $\boldsymbol{\mu}$ and variance matrix $\mathbf{M}$. What is the asymptotic distribution of

$$W_n = \frac{\overline{X}_n}{\overline{Y}_n}$$

assuming $\mu_2 \neq 0$.

8-9. Suppose

$$\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{D}} Y$$

Show that

$$T_n \xrightarrow{P} \theta.$$

(**Hint:** Lemma 8.5 and the continuous mapping theorem.)

8-10. In Example 8.1.3 we showed that S_n^2 and V_n have the same asymptotic distribution, which is given by Corollary 7.17 in Chapter 7 of these notes. Find the asymptotic distribution of S_n.

8-11. (This problem has nothing to do with convergence concepts. It is a lemma for the following problem.)

Consider two arbitrary events A and B, and, as usual, let I_A and I_B denote their indicator functions. Show that

$$\operatorname{cov}(I_A, I_B) = P(A \cap B) - P(A)P(B).$$

Hint: $I_{A \cap B} = I_A I_B$.

8-12. (This is the bivariate analog of Problem 6-4 of Chapter 6 of these notes.)

Suppose X_1, X_2, $\ldots$ are i. i. d. with common probability measure P, and define

$$\begin{aligned} Y_n &= I_A(X_n) \\ Z_n &= I_B(X_n) \end{aligned}$$

for some events A and B. Find the asymptotic distribution of the vector $(\overline{Y}_n, \overline{Z}_n)$.

Chapter 9

Frequentist Statistical Inference

9.1 Introduction

9.1.1 Inference

Statistics is Probability done backwards.

Probability theory allows us to do calculations given a probability model. If we assume a random variable X has a particular distribution, then we can calculate, at least in principle, $P(|X| \geq c)$ or $E(X)$ or $\operatorname{var}(X)$. Roughly speaking, given the distribution of X we can say some things about X. Statistics tries to solve the inverse problem: given an observation of X, say some things about the distribution of X. This is called *statistical inference.*

Needless to say, what statistics can say about the distribution of random data is quite limited. If we ask too much, the problem is impossible. In a typical situation, any value of the observed data is possible under any of the distributions being considered as possible models. Thus an observation of the data does not completely rule out any model. However, the observed data will be more probable under some models and less probable under others. So we ought to be able to say the data favor some distributions more than others or that some distributions seem very unlikely (although not impossible).

9.1.2 The Sample and the Population

Usually the data for a statistical problem can be considered a random vector $\mathbf{X} = (X_1, \ldots, X_n)$ representing a sample from a population. The statistical problem is to infer the population distribution (which determines the probability distribution of $\mathbf{X}$). The simple-minded approach to this problem, which seems natural to those who have not been exposed to formal statistical thinking, is to just treat the sample as if it were the population. This is clearly a mistake, obvious as soon as it is described in these words. It violates the fundamental issue that statistics must deal with.

> *The sample is not the population.*

In introductory courses I try to repeat this slogan so often that it is drilled into students by sheer force of repetition. Despite its simplicity, it is very important to remember. I often see it violated by scientists who are trained in formal statistics and think they are correctly applying their statistical training. It's easy to do. Just a little bit of confusion obscures the issues enough to make one rely on intuition, which is always wrong.

> *Intuition is always wrong about statistics.*
>
> *Statistics courses don't develop good intuition about statistics. They teach you how to calculate instead of using intuition.*
>
> *Expert intuition is no better than anyone else's. Experts are just better at calculating and knowing what to calculate.*

Intuition treats the sample as the population. Whenever you ignore "the sample is not the population," you will say stupid things and do stupid things.

9.1.3 Frequentist versus Bayesian Inference

The word "frequentist" in the chapter title refers to a great philosophical divide in statistical inference. All statistical inference is divided into two parts. One part is generally called "Bayesian" because of the prominent role played by Bayes' rule. It will be covered in a later chapter. The other part has no generally accepted name. We could call it non-Bayesian, but that sounds negative and is also inaccurate because it also sometimes uses Bayes' rule. We could call it "statistical inference based on sampling distributions," which would be accurate but too long for everyday use.

Devotees of Bayesian inference generally call the other camp "frequentist." This is intended to be pejorative, saddling the enemy with the philosophical baggage of the frequentist philosophy of probability, which says that it only makes sense to talk about probabilities in an infinite sequence of identical random experiments. The idea is that it only makes sense to apply "frequentist" statistics to actual infinite sequences of identical random experiments. Since that doesn't describe any real data, one should never use "frequentist" statistics.

These days, however, no one is really a "frequentist" about probability. Probably no one, except a few philosophers, ever was. Everyone is a formalist, holding the view that anything that satisfies the probability axioms is probability (if it waddles like a duck and quacks like a duck, then it is a duck). This takes all the sting out of the "frequentist" label. No one minds the label, because everyone knows it isn't accurate. As we will see, the only thing required for so-called "frequentist" inference is probability models for data. It doesn't matter what you think probability really is.

9.2 Models, Parameters, and Statistics

A *statistical model* is a family of probability distributions. This differs from a *probability model*, which is only a single distribution.

9.2.1 Parametric Models

Often we consider probability models having an adjustable constant in the formula for the density.[1] Generically, we refer to such a constant as a *parameter* of the distribution. This notion was introduced back in Section 3.1 of these notes, and all of the "brand name distributions" described in Chapter 4 of these notes and Chapter 6 of Lindgren and summarized in Appendix B of these notes are examples of parametric models.

Usually, though not always, we use Greek letters for parameters to distinguish them from random variables (large Roman letters) and possible values of random variables (small Roman letters). Among the brand-name distributions (Appendix B of these notes), the only parameters we do not use Greek letters for are the success probability p occurring in the Bernoulli, binomial, geometric, and negative binomial distributions and the analogous vector parameter $\mathbf{p}$ of the multinomial distribution, the parameters s and t of the beta distribution, and the parameters a and b of the uniform distribution on the interval (a, b). All the rest are Greek letters.

When we say let X be a random variable having density f_θ, this means that for each fixed value of the parameter θ the function f_θ is a probability density, which means it satisfies (3.1a) and (3.1b) of Chapter 3 of these notes. Of course, we didn't use θ for the parameter of any brand-name distribution. The idea is that θ can stand for any parameter (for μ of the Poisson, for λ of the exponential, and so forth).

Each different value of the parameter θ gives a different probability distribution. As θ ranges over its possible values, which we call the *parameter space*, often denoted Θ when the parameter is denoted θ, we get a *parametric family* of densities

$$\{\, f_\theta : \theta \in \Theta \,\}$$

Even the notation for parametric families is controversial. How can that be? Mere notation generate controversy? You can see it in the conflict between the notation $f_\theta(x)$ used in these notes and the notation $f(x \mid \theta)$ used in Lindgren.

Lindgren uses the same notation that one uses for conditional probability densities. The reason he uses that notation is because he belongs to the Bayesian

[1] In these notes and in the lecture we use the term *density* to refer to either of what Lindgren calls the *probability function* (p. f.) of a discrete distribution of the *probability density function* (p. d. f.) of a continuous distribution. We have two reasons for what seems at first sight a somewhat eccentric notion (failing to draw a terminological distinction between these two rather different concepts). First, these two concepts are special cases of a more general concept, also called *density*, explained in more advanced probability courses. Second, and even more important, these two concepts are used the same way in statistics, and it is a great convenience to say "density" rather than "p. f. or p. d. f." over and over.

camp, and as a matter of philosophical principle is committed to the notion that all parameters are random variables. Bayesians consider $f(x \mid \theta)$ the conditional distribution of the random variable X given the random variable θ. We can't use the "big X" and "little x" distinction for Greek letters because we often use the corresponding capital letters for something else. In particular, as explained above, we use Θ for the parameter space, not for the parameter considered as a random variable. But regardless of whether the "big X" and "little x" convention can be applied, the important point is that Bayesians do consider $f(x \mid \theta)$ a conditional density. The rest of the statistical profession, call them "non-Bayesians" is at least willing to think of parameters as not being random. They typically use the notation $f_\theta(x)$ to show that θ is an adjustable constant and is not being treated like a random variable.

It is also important to realize that in a statistical model, probabilities and expectations depend on the actual value of the parameter. Thus it is ambiguous to write $P(A)$ or $E(X)$. Sometimes we need to explicitly denote the dependence on the parameter by writing $P_\theta(A)$ and $E_\theta(X)$, just as we write $f_\theta(x)$ for densities.

Location-Scale Families

Location-scale families were introduced in Section 4.1 of Chapter 4 of these notes. The only brand name location-scale families are $\mathcal{U}(a, b)$, $\mathcal{N}(\mu, \sigma^2)$, and Cauchy$(\mu, \sigma)$. But, as shown in Example 4.1.3 in Chapter 4 of these notes, there are many more "non-brand-name" location scale families. In fact, every probability distribution of a real-valued random variable generates a location-scale family.

The only reason for bring up location-scale families here is to make the point that a statistical model (family of probability distributions) can have many different parameterizations. Example 4.1.1 of Chapter 4 of these works out the relation between the usual parameters a and b of the $\mathcal{U}(a, b)$ distribution and the mean μ and variance σ, which can also be used to parameterize this family of distributions. The relation between the two parameterizations is given by unnumbered displayed equations in that example, which we repeat here

$$\begin{aligned} \mu &= \frac{a + b}{2} \\ \sigma &= \sqrt{\frac{(b - a)^2}{12}} \end{aligned}$$

This is an invertible change of parameters, the inverse transformation being

$$\begin{aligned} a &= \mu - \sigma\sqrt{3} \\ b &= \mu + \sigma\sqrt{3} \end{aligned}$$

This illustrates a very important principle.

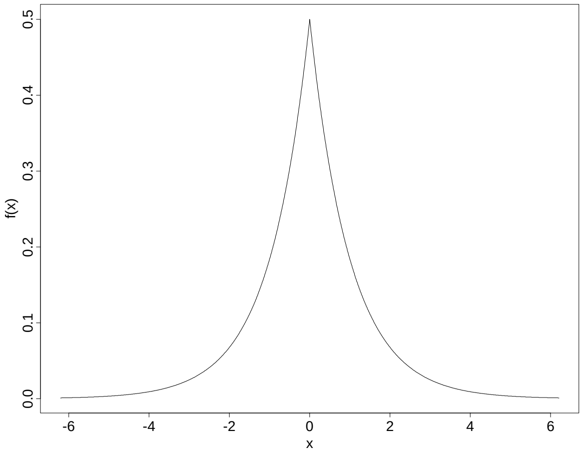
A single statistical model can have many different parameterizations.

We often change parameters, using the parameterization that seems simplest in a particular problem.

Example 9.2.1 (Laplace Distributions).
The density

$$f(x) = \frac{1}{2} e^{-|x|}, \qquad -\infty < x < +\infty \tag{9.1}$$

is called a *Laplace* or *double exponential* density. It is two Exp(1) densities back to back. The density is graphed below.



The Laplace location-scale family thus has densities

$$f_{\mu,\sigma}(x) = \frac{1}{2\sigma} e^{-|x-\mu|/\sigma}, \qquad -\infty < x < +\infty \tag{9.2}$$

Note (σ^2 is *not* the variance, see Problem 9-1).

Models and Submodels

Any set of probability distributions is a statistical model. A statistical model need not include *all* distributions of a certain type. It might have only a subset of them. We then say we have a *submodel* of the larger family. In parametric families, we specify submodels by specifying their parameter spaces.

Example 9.2.2 (All Normal Distributions).
The family of $\mathcal{N}(\mu, \sigma^2)$ distributions for $-\infty < \mu < +\infty$ and $0 < \sigma < \infty$ is a statistical model. (As mentioned in the preceding section, it is a location-scale family, μ is the location parameter and σ is the scale parameter.) Because the model has two parameters, the parameter space is a subset of $\mathbb{R}^2$

$$\Theta = \{ (\mu, \sigma) \in \mathbb{R}^2 : \sigma > 0 \}.$$

Example 9.2.3 (Normal Distributions, Unit Variance).
The family of $\mathcal{N}(\mu, 1)$ distributions for $-\infty < \mu < +\infty$ is a also statistical model. It is a submodel of the family of all normal distributions in the preceding example. The model has one parameter μ, so the parameter space is a subset of $\mathbb{R}$. In fact, since μ is unrestricted, the parameter space is the whole real line: $\Theta = \mathbb{R}$.

It is important to realize that the examples describe two different models. It is not enough to say we are talking about a normal model. Different parameter spaces make different models.

Example 9.2.4 (Translation Families).
If we take a location-scale family and fix the scale parameter, then we have a one-parameter family. Example 9.2.3 is an example of this. Such a family is called a *location family* or a *translation family*, the latter name arising because the different random variables in the family are related by *translations*, which are changes of variables of the form

$$Y = \mu + X.$$

The distributions in the family differ only in location. They all have the same shape and scale.

Example 9.2.5 (Scale Families).
Conversely, if we take a location-scale family and fix the location parameter, then we also have a one-parameter family. But now the varying parameter is the scale parameter, so the family is called a *scale family.* The distributions in the family differ only in scale. They all have the same shape. They may also differ in location, because a scale transformation of the form

$$Y = \sigma X$$

changes both location and scale. For example, if X has a variance, then

$$\begin{aligned} E(Y) &= \sigma E(X) \\ \operatorname{var}(Y) &= \sigma^2 \operatorname{var}(X) \end{aligned}$$

9.2.2 Nonparametric Models

Some families of distributions are too big to specify in parametric form. No finite set of real parameters can serve to describe the family.

Example 9.2.6 (All Distributions with Finite Variance).
The family of all distributions with finite variance is a statistical model.

At the level of mathematics used in this course, it is hard to see that this model cannot be parameterized, but that does not really matter. The important point is that this is a statistical model even though we do not specify it using a parametric family of densities.

It is important to realize that probabilities and expectations depend on the actual probability distribution of the data, in nonparametric models just as in parametric models. Thus it is still ambiguous to write $P(A)$ or $E(X)$. The probabilities and expectations depend on the actual distribution of X. Since the model is not parametric, we cannot write $P_\theta(A)$ or $E_\theta(X)$ to remind us of the dependence. But it is still there and must be kept in mind.

9.2.3 Semiparametric Models

Sometimes, rather confusingly, we speak of parameters of nonparametric distributions. In this usage a *parameter* is any quantity that is determined by a distribution. In Example 9.2.6 we can still speak of the mean μ as being a parameter of the family. Every distribution in the family has a mean (because it has a variance and this implies existence of moments of lower order). Many different distributions in the family have the same mean, so the mean doesn't determine the distribution, and hence we don't have a parametric family with the mean as its parameter. But we still speak of the mean as being a parameter (rather than *the* parameter) of the family.

Models of this sort are sometimes called *semiparametric*, meaning they have a parametric part of the specification and a nonparametric part. In the example, the parametric part of the specification of the distribution is the mean μ and the nonparametric part is the rest of the description of the distribution (whatever that may be).

9.2.4 Interest and Nuisance Parameters

In multiparameter models, we divide parameters into two categories: *parameters of interest* (also called *interest parameters* though that is not idiomatic English) and *nuisance parameters*. The parameter or parameters of interest are the ones we want to know something about, the nuisance parameters are just complications. We have to deal with the nuisance parameters, but they are not interesting in themselves (in the particular application at hand).

In semiparametric models, the parametric part is typically the parameter of interest, the nonparametric part is the "nuisance" part of the model specification, although we can no longer call it the "nuisance parameter" when it is nonparametric.

Example 9.2.7 (All Distributions with Finite Variance).
The family of all distributions with finite variance is a semiparametric statistical model when we consider the mean μ the parameter of interest.

9.2.5 Statistics

Also somewhat confusingly, the term "statistic" is used as a technical term in this subject. Please do not confuse it with the name of the subject, "statistics." A *statistic* is a function of the data of a random experiment. It cannot involve any parameters or otherwise depend on the true distribution of the data. Since the data make up a random vector and a function of a random variable is a random variable, statistics are random variables. But a random variable can depend on a parameter, and a statistic cannot.

> *All statistics are random variables, but some random variables are not statistics.*

For each fixed μ, the function $X - \mu$ is (if X is the data) a random variable. But if μ is a parameter of the statistical model under consideration, $X - \mu$ is not a statistic. The reason for the distinction is that assuming a statistical model doesn't completely specify the distribution. It only says that X has some distribution in the family, but doesn't say which one. Hence we don't know what μ is. So we can't calculate $X - \mu$ having observed X. But we can calculate any statistic, because a statistic is a function of the observed data only.

9.3 Point Estimation

One form of statistical inference is called *point estimation.* Given data X_1, ..., X_n that are a random sample from some population or that are i. i. d. having a distribution in some statistical model, the problem is to say something about a parameter θ of the population or model, as the case may be. A *point estimate* (also called *point estimator*) of the parameter is a function of the data

$$\hat{\theta}_n(\mathbf{X}) = \hat{\theta}_n(X_1, \ldots, X_n) \tag{9.3}$$

that we use as an estimate of the true unknown parameter value.

This is our first example of "hat" notation. The symbol $\hat{\theta}$ is read "theta hat," the symbol on top of the letter being universally called a "hat" in mathematical contexts (outside math it is called a "caret" or "circumflex accent"). It changes the convention that parameters are Greek letters (like θ) and random variables are Roman letters (like X). Now we are adding Greek letters with hats to the random variables. Since $\hat{\theta}_n$ given by (9.3) is a function of the data, it is a random variable (and not just a random variable, more precisely, it is a *statistic*). The reason we do this is to make the connection between the two clear: $\hat{\theta}_n$ is a point estimator of θ. We often denote a point estimator of a parameter by putting a hat on the parameter. Remember that this puts $\hat{\theta}_n$ in a different conceptual category from θ. The point estimate $\hat{\theta}_n$ is a random variable. The parameter θ is a nonrandom constant.

Example 9.3.1 (Estimating the Mean).
Given i. i. d. data X_1, ..., X_n from a distribution having a mean μ, one point estimator of μ is the sample mean $\overline{X}_n$ defined by (7.15). Another point estimator of μ is the sample median of the empirical distribution of the data $\widetilde{X}_n$ defined in Definition 7.1.4. Yet another point estimator of μ is the constant estimator $\hat{\mu}_n \equiv 42$ that completely ignores the data, producing the estimate 42 for any data.

It is important to understand that any function whatsoever of the data is a point estimate of θ as long as it is a statistic (a function of data values only, not parameters). Even really dumb functions, such as the constant function $\hat{\theta}_n \equiv 42$ that completely ignores the data, are point estimates. Thus calling a function of the data a "point estimate" doesn't say anything at all about the properties of the function except that it is a statistic. The only point of calling a statistic a point estimate of θ is to establish a context for subsequent discussion. It only says we *intend to use* the statistic as a point estimate.

As we just said, every statistic whatsoever is a point estimate of θ, but that doesn't end the story. Some will be better than others. The ones of actual interest, the ones that get used in statistical practice, are the good ones. Much of the theory of point estimation is about which estimators are good. In order to characterize good estimators, we need some criterion of goodness. In fact, there are several different criteria in common use, and different criteria judge estimators differently. An estimator might be good under some criteria and bad under others. But at least if we use an estimator that is good according to some particular criterion, that says something.

The most obvious criterion, how often an estimator is correct, is unfortunately worthless.

> *With continuous data, every continuous estimator is wrong with probability one.*

The true parameter value θ is just a point in the parameter space. We don't know which point, but it is some point. If $\hat{\theta}(\mathbf{X})$ is a continuous random variable, then $P_\theta\{\hat{\theta}(\mathbf{X}) = \theta\}$ is zero, because the probability of every point is zero, as is true for any continuous random variable.

9.3.1 Bias

An estimator T of a parameter θ is *unbiased* if $E_\theta(T) = \theta$, that is, if θ is the mean of the sampling distribution of T when θ is the true parameter value. An estimator that is not unbiased is said to be *biased*, and the difference

$$b(\theta) = E_\theta(T) - \theta$$

is called the *bias* of the estimator.

Example 9.3.2 (Estimating σ^2).
By equations 7.22a and 7.22b, S_n^2 is an unbiased estimator of σ^2 and V_n is a biased estimator of σ^2.

The bias of S_n^2 as an estimator of σ^2 is zero (zero bias is the same as unbiased).

The bias of V_n as an estimator of σ^2 is

$$E(V_n) - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}$$

Note that the bias is negative, meaning that the estimator V_n is below the parameter σ^2 on average.

In a multiparameter problem with a single parameter of interest φ and a nuisance parameter ψ, an estimator T of φ is *unbiased* if $E_\theta(T) = \varphi$, where $\theta = (\varphi, \psi)$ is the parameter vector, and the difference

$$b(\theta) = b(\varphi, \psi) = E_\theta(T) - \varphi$$

is called the *bias* of the estimator. Note that the bias generally depends on both interest and nuisance parameters.

It is generally a bad idea to give mathematical concepts emotionally charged names, and this concept is particularly badly named. Naturally we all want to be unbiased, so we should avoid biased estimators right? Wrong! It is important to remember that this mathematical concept has nothing whatsoever to do with what we call bias in everyday life.

In fact, it can mean the exact opposite! Psychologists call an achievement test unbiased if it satisfies the statistical definition, if it has the correct expectation. The tests are supposed to predict grades in school, and the test is unbiased if it is wrong high and wrong low about equally often so that it is right on average. But grades in school are themselves biased in the everyday sense (unless teachers all turned into saints when I wasn't looking). So in order to be unbiased in the statistical sense the tests must accurately track whatever bias in the everyday sense there is in the grades.

Note that this argument has nothing whatsoever to do with whether the questions on the tests appear to be biased, which is what the argument about "culturally biased" tests usually revolves around. Whatever the appearances, enough cultural bias (or other kinds of bias) must be somehow built into the tests, perhaps without any effort on the part of the people constructing the tests, perhaps even despite efforts to avoid it, to exactly match whatever cultural bias (or whatever) is in grades.

There are also technical arguments against unbiased estimation. I once had a friend who claimed he was ambidextrous because he did equally poorly with both hands. That's the idea behind unbiased estimation, doing equally poorly high and low.

Example 9.3.3 (Constrained Estimation).
Suppose the parameter satisfies a constraint, for example, θ might be a variance, in which case $\theta \ge 0$. Any sensible estimator should take values in the parameter space. Hence we should have $T(\mathbf{X}) \ge 0$ for all values of the data $\mathbf{X}$. Suppose also that our statistical model consists of probability distributions with the same support, hence the same events of probability zero. Then $P_\theta\{T(\mathbf{X}) > 0\}$ is either zero for all θ or nonzero for all θ, and by Theorem 5 of Chapter 4 in Lindgren, this implies

$$E_\theta(T) = 0, \qquad \theta \in \Theta \tag{9.4a}$$

or

$$E_\theta(T) > 0, \qquad \theta \in \Theta \tag{9.4b}$$

If (9.4a) holds, then the estimator is biased because $E_\theta(T) = \theta$ does not hold when $\theta \ne 0$. If (9.4b) holds, then the estimator is also biased, because $E_\theta(T) = \theta$ does not hold when $\theta = 0$.

Hence either way we get a biased estimator. The only way to get an unbiased estimator is to sometimes estimate ridiculous (that is, negative) values of the parameter.

So why specifically do I call the principle of unbiasedness the principle of "doing equally poorly with both hands" in this situation? The constraint allows you to do much better on the low side than the high side. Take any estimator $T(\mathbf{X})$, perhaps an unbiased one that sometimes estimates negative parameter values. The estimator is clearly improved by setting all the negative estimates to zero, that is,

$$T_{\text{improved}}(\mathbf{X}) = \begin{cases} T(\mathbf{X}), & T(\mathbf{X}) > 0 \\ 0, & \text{otherwise} \end{cases}$$

is a better estimator, because when $T(\mathbf{X})$ is not ridiculous (i. e., not negative) $T_{\text{improved}}(\mathbf{X})$ has the same value and when when $T(\mathbf{X})$ is ridiculous (negative) $T_{\text{improved}}(\mathbf{X})$ is not (is zero). But $T_{\text{improved}}(\mathbf{X})$ is *biased* whereas $T(\mathbf{X})$ may be *unbiased*. Adopting the principle of unbiasedness here means accepting that one should, as a matter of principle, increase the errors of estimation on the low side to make them as large as the inevitable errors on the high side. Stated that way, it is a mystery why anyone thinks unbiasedness is a good thing. (The solution to the mystery is that people who think unbiasedness is a good thing have never seen this example or other examples where unbiasedness is clearly a bad thing.)

Lest you think the example contrived, let me assure you that it does arise in practice, and I have actually seen real scientists using ridiculous estimators in order to achieve unbiasedness. They must have had a bad statistics course that gave them the idea that unbiasedness is a good thing.

Even though it is not a particularly good thing, unbiasedness is an important theoretical concept. We will meet several situations in which we can prove something about unbiased estimators, but can't do anything with estimators in general. For example, there are theorems that say under certain circumstances that a particular estimator is uniformly minimum variance unbiased (UMVU). It is easy to misinterpret the theorem to say that the best estimator is unbiased, but it doesn't say that at all. In fact, it implicitly says the opposite. It says the particular estimator is better than any other *unbiased* estimator. It says nothing about *biased* estimators, presumably some of them are better still, otherwise we could prove a stronger theorem.

Another issue about bias is that nonlinear functions of unbiased estimators are usually not unbiased. For example, suppose T is an unbiased estimator of θ. Is T^2 also an unbiased estimator of θ^2? No! By the parallel axis theorem

$$E_\theta(T^2) = \operatorname{var}_\theta(T) + E_\theta(T)^2.$$

Unless the distribution of T is concentrated at one point, $\operatorname{var}_\theta(T)$ is strictly greater than zero, and T^2 is biased high, that is, $E_\theta(T^2) > \theta^2$, when T is unbiased for θ. Conversely, if T^2 is unbiased for θ^2, then T is biased low for θ, that is $E_\theta(T) < \theta$.

Example 9.3.4 (Estimating σ).
Since S_n^2 is an unbiased estimator of σ^2, it follows that S_n itself is a biased estimator of σ, in fact $E(S_n) < \sigma$ always holds.

9.3.2 Mean Squared Error

The *mean squared error* (m. s. e.) of an estimator T of a parameter θ is

$$\operatorname{mse}_\theta(T) = E_\theta\bigl\{(T - \theta)^2\bigr\}$$

By the parallel axis theorem

$$\operatorname{mse}_\theta(T) = \operatorname{var}_\theta(T) + b(\theta)^2$$

So mean squared error is variance plus bias squared, and for unbiased estimators m. s. e. is just variance.

Mean squared error provides one sensible criterion for goodness of point estimators. If T_1 and T_2 are estimators of the same parameter θ, then we can say that T_1 is better than T_2 if $\operatorname{mse}(T_1) < \operatorname{mse}(T_2)$. It goes without saying that if we choose a different criterion, the order could come out differently.

An example of another criterion is mean absolute error $E_\theta\{|T-\theta|\}$, but that one doesn't work so well theoretically, because the parallel axis theorem doesn't apply, so there is less we can say about this criterion than about m. s. e.

Example 9.3.5.
Consider the class of estimators of σ^2 of the form kV_n, where $k > 0$ is some constant. The choice $k = 1$ gives V_n itself. The choice $k = n/(n-1)$ gives S_n^2. It turns out that neither of these estimators is the best in this class when we use mean squared error as the criterion. The best in the class is given by the choice $k = n/(n+1)$. No proof is given here. It is a homework problem (Problem 8-7 in Lindgren).

Note that the optimal estimator is *biased.* This gives us yet another example showing that unbiasedness is not necessarily a good thing. The same sort of calculation that shows the choice $k = n/(n+1)$ is optimal, also shows that $\operatorname{mse}(V_n) < \operatorname{mse}(S_n^2)$. So among the two more familiar estimators, the unbiased one is worse (when mean square error is the criterion).

9.3.3 Consistency

Definition 9.3.1 (Consistency).
A sequence of point estimators $\{T_n\}$ of a parameter θ is consistent *if*

$$T_n \xrightarrow{P} \theta, \qquad \text{as } n \to \infty.$$

Generally we aren't so pedantic as to emphasize that consistency is really a property of a sequence. We usually just say T_n is a consistent estimator of θ.

Consistency is not a very strong property, since it doesn't say anything about how fast the errors go to zero nor does it say anything about the distribution of the errors. So we generally aren't interested in estimators that are merely consistent unless for some reason consistency is all we want. We will see the most important such reason in the following section. For now we just list a few consistent estimators.

By the law of large numbers, if X_1, X_2, $\ldots$ are i. i. d. from a distribution with mean μ, then the sample mean $\overline{X}_n$ is a consistent estimator of μ. The only requirement is that the expectation defining μ exist.

Similarly, by Theorem 7.15 every sample moment (ordinary or central) is a consistent estimator of the corresponding population moment. The only requirement is that the population moment exist. Here we have fallen into the sloppy terminology of referring to i. i. d. random variables as a "sample" from a hypothetical infinite "population." What is meant, of course, is that if X_1, X_2, $\ldots$ are i. i. d. from a distribution having ordinary moment α_k or central moment μ_k and if the corresponding sample moments are $A_{k,n}$ and $M_{k,n}$ in the notation of Section 7.3.1, then

$$A_{k,n} \xrightarrow{P} \alpha_k$$
$$M_{k,n} \xrightarrow{P} \mu_k$$

provided only that the moments α_k and μ_k exist.

That doesn't give us a lot of consistent estimators, but we can get a lot more with the following.

Theorem 9.1. *Any continuous function of consistent estimators is consistent. Specifically, if $T_{i,n} \xrightarrow{P} \theta_i$, as $n \to \infty$ for $i = 1, \ldots, m$, then*

$$g(T_{1,n}, \ldots, T_{m,n}) \xrightarrow{P} g(\theta_1, \ldots, \theta_m), \qquad \text{as } n \to \infty$$

if g is jointly continuous at the point $(\theta_1, \ldots, \theta_m)$.

This is just the multivariate version of the continuous mapping theorem for convergence in probability (Theorem 8.7).

9.3.4 Asymptotic Normality

As we said in the preceding section, mere consistency is a fairly uninteresting property, unless it just happens to be all we want. A much more important property is asymptotic normality. Another way to restate the definition of consistency is

$$T_n - \theta \xrightarrow{P} 0.$$

The estimator T_n is supposed to estimate the parameter θ, so $T_n - \theta$ is the error of estimation. Consistency says the error goes to zero. We would like to know more than that. We would like to know how about big the error is, more specifically we would like an approximation of its sampling distribution.

It turns out that almost all estimators of practical interest are not just consistent but also *asymptotically normal*, that is,

$$\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2) \tag{9.5}$$

holds for some constant σ^2, which may depend on the true distribution of the data. We say an estimator T_n that satisfies (9.5) is *consistent and asymptotically normal* (that is, asymptotically normal when centered at θ). It may be very hard, even impossible, to work out theoretically what the constant σ^2 actually is, although these days one can often use computer simulations to calculate it when pencil and paper analysis fails. Examples of consistent and asymptotically normal estimators are ordinary and central sample moments (Theorems 7.15 and 7.16).

The property (9.5) is not much help by itself, because if σ^2 actually depends on the true distribution of the data (that is, σ^2 is actually a function of the parameter θ, although the notation doesn't indicate this), then we don't know what it actually is because we don't know the true distribution (or the true value of θ). Then the following theorem is useful.

Theorem 9.2 (Plug-In). *Suppose* (9.5) *holds and S_n is any consistent estimator of σ, then*

$$\frac{T_n - \theta}{S_n / \sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$$

This was proved as a homework problem last semester, and we repeated the proof in Example 8.1.5. It is just (9.5) and Slutsky's theorem. We call this the "plug-in" theorem, because it says asymptotics still works when you plug in S_n for σ.

9.3.5 Method of Moments Estimators

A function of sample moments is called a *method of moments estimator* of a parameter θ if the function evaluated at the corresponding population moments is equal to θ.

Example 9.3.6.
Trivially, sample moments are "method of moments estimators" of the corresponding population moments.

Example 9.3.7 (The Two-Parameter Gamma Model).
The mean and variance of the $\text{Gam}(\alpha, \lambda)$ distribution are

$$\begin{aligned} \mu &= \frac{\alpha}{\lambda} \\ \sigma^2 &= \frac{\alpha}{\lambda^2} \end{aligned}$$

Solving for α and λ gives

$$\alpha = \frac{\mu^2}{\sigma^2}$$
$$\lambda = \frac{\mu}{\sigma^2}$$

Plugging in the corresponding sample moments gives

$$\hat{\alpha}_n = \frac{\overline{X}_n^2}{V_n} \tag{9.6a}$$
$$\hat{\lambda}_n = \frac{\overline{X}_n}{V_n} \tag{9.6b}$$

These are method of moments estimators because they are functions of sample moments, for example

$$\hat{\alpha}_n = g(\overline{X}_n, V_n),$$

where

$$g(u, v) = \frac{u^2}{v}, \tag{9.7}$$

and the function evaluated at the population moments is the parameter to be estimated, for example

$$g(\mu, \sigma^2) = \frac{\mu^2}{\sigma^2} = \alpha.$$

Method of moments estimators are always consistent and asymptotically normal if enough population moments exist and they are nice functions of the sample moments.

Theorem 9.3. *A method of moments estimator involving sample moments of order k or less is consistent provided population moments of order k exist and provided it is a continuous function of the sample moments.*

This is just Theorem 7.15 combined with Theorem 9.1.

Theorem 9.4. *A method of moments estimator involving sample moments of order k or less is asymptotically normal provided population moments of order $2k$ exist and provided it is a differentiable function of the sample moments.*

The proof of this theorem is just the multivariate delta method (Theorem 8.9) applied to the multivariate convergence in distribution of sample moments, for which we have not stated a completely general theorem. What is needed is the multivariate analog of Theorem 7.16 which would give the asymptotic *joint* distribution of several sample moments, rather than the asymptotic *marginal* of just one. Rather than state such a general theorem, we will be content by giving the specific case for the first two moments.

Theorem 9.5. *If X_1, X_2, $\ldots$ are i. i. d. from a distribution having fourth moments, then*

$$\sqrt{n} \begin{pmatrix} \overline{X}_n - \mu \\ A_{2,n} - \alpha_2 \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{M}_1), \tag{9.8}$$

where $A_{2,n}$ is the sample ordinary second moment and

$$\mathbf{M}_1 = \begin{pmatrix} \alpha_2 - \alpha_1^2 & \alpha_3 - \alpha_1 \alpha_2 \\ \alpha_3 - \alpha_1 \alpha_2 & \alpha_4 - \alpha_2^2 \end{pmatrix} \tag{9.9}$$

and

$$\sqrt{n} \begin{pmatrix} \overline{X}_n - \mu \\ V_n - \sigma^2 \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{M}_2), \tag{9.10}$$

where

$$\mathbf{M}_2 = \begin{pmatrix} \mu_2 & \mu_3 \\ \mu_3 & \mu_4 - \mu_2^2 \end{pmatrix} \tag{9.11}$$

Proof. The first assertion of the theorem was proved in Examples 5.1.1 and 8.1.2. The second assertion was almost, but not quite, proved while we were proving Theorem 7.16. In that theorem, we obtained the asymptotic marginal distribution of V_n, but not the asymptotic joint distribution of $\overline{X}_n$ and V_n. However in the unlabeled displayed equation just below (8.22) we determined

$$\sqrt{n} \begin{pmatrix} \overline{X}_n - \mu \\ M'_{2,n} - \mu_2 \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{M}_2)$$

where by the empirical central axis theorem

$$M'_{2,n} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2 = V_n + (\overline{X}_n - \mu)^2$$

Hence

$$\sqrt{n} \begin{pmatrix} \overline{X}_n - \mu \\ M'_{2,n} - \mu_2 \end{pmatrix} = \sqrt{n} \begin{pmatrix} \overline{X}_n - \mu \\ V_n - \mu_2 \end{pmatrix} + \sqrt{n} \begin{pmatrix} 0 \\ (\overline{X}_n - \mu)^2 \end{pmatrix}$$

and by Problem 8-6 the second term converges in probability to zero, hence Slutsky's theorem gives the asserted result. □

Example 9.3.8.

In Example 8.2.1 we essentially did a method of moments estimator problem. We just didn't know at the time that "method of moments" is what statisticians call that sort of problem. There we looked at the asymptotic behavior of $1/\overline{X}_n$ for an i. i. d. sample from an Exp(λ) distribution. From the fact that

$$E(X) = \frac{1}{\lambda}$$

we see that

$$\hat{\lambda}_n = \frac{1}{\overline{X}_n}$$

is the obvious method of moments estimator of λ. In Example 8.2.1 we calculated its asymptotic distribution

$$\hat{\lambda}_n \approx \mathcal{N}\left(\lambda, \frac{\lambda^2}{n}\right)$$

Example 9.3.9.
For the method of moments estimators $\hat{\alpha}_n$ and $\hat{\lambda}_n$ defined in Example 9.3.7, these theorems imply that these estimators are consistent and asymptotically normal, because the gamma distribution has moments of all orders and both estimators are differentiable functions of the sample moments they involve.

Getting the actual asymptotic distribution of the estimators is more work. We have to apply the multivariate delta method to the result of Theorem 9.5. We'll just do $\hat{\alpha}_n$ As was pointed out in Example 9.3.7 $\hat{\alpha}_n = g(\overline{X}_n, V_n)$, where the function g is given by (9.7), which has derivative

$$\mathbf{G} = \nabla g(\mu, \sigma^2) = \begin{pmatrix} \frac{2\mu}{\sigma^2} & -\frac{\mu^2}{\sigma^4} \end{pmatrix} \tag{9.12}$$

The specific form of the asymptotic variance matrix of $\overline{X}_n$ and V_n is given by (9.11) with the specific moments of the gamma distribution plugged in. Of course, we already know

$$\mu = \frac{\alpha}{\lambda}$$

and

$$\sigma^2 = \mu_2 = \frac{\alpha}{\lambda^2}$$

Plugging these into (9.12) gives

$$\mathbf{G} = \begin{pmatrix} 2\lambda & -\lambda^2 \end{pmatrix} \tag{9.13}$$

To calculate $\mathbf{M}_2$, we need to also calculate μ_3 and μ_4.

$$\begin{aligned}
\mu_3 &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty \left(x - \frac{\alpha}{\lambda}\right)^3 x^{\alpha-1} e^{-\lambda x} \, dx \\
&= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty \left(x^3 - \frac{3\alpha x^2}{\lambda} + \frac{3\alpha^2 x}{\lambda^2} - \frac{\alpha^3}{\lambda^3}\right) x^{\alpha-1} e^{-\lambda x} \, dx \\
&= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha+3-1} e^{-\lambda x} \, dx - \frac{3\alpha\lambda^{\alpha-1}}{\Gamma(\alpha)} \int_0^\infty x^{\alpha+2-1} e^{-\lambda x} \, dx \\
&\qquad + \frac{3\alpha^2\lambda^{\alpha-2}}{\Gamma(\alpha)} \int_0^\infty x^{\alpha+1-1} e^{-\lambda x} \, dx - \frac{\alpha^3\lambda^{\alpha-3}}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-\lambda x} \, dx \\
&= \frac{\lambda^\alpha \Gamma(\alpha+3)}{\lambda^{\alpha+3}\Gamma(\alpha)} - \frac{3\alpha\lambda^{\alpha-1}\Gamma(\alpha+2)}{\lambda^{\alpha+2}\Gamma(\alpha)} + \frac{3\alpha^2\lambda^{\alpha-2}\Gamma(\alpha+1)}{\lambda^{\alpha+1}\Gamma(\alpha)} - \frac{\alpha^3\lambda^{\alpha-3}\Gamma(\alpha)}{\lambda^\alpha\Gamma(\alpha)} \\
&= \frac{(\alpha+2)(\alpha+1)\alpha}{\lambda^3} - \frac{3\alpha(\alpha+1)\alpha}{\lambda^3} + \frac{3\alpha^2\alpha}{\lambda^3} - \frac{\alpha^3}{\lambda^3} \\
&= \frac{(\alpha+2)(\alpha+1)\alpha}{\lambda^3} - \frac{3\alpha(\alpha+1)\alpha}{\lambda^3} + \frac{3\alpha^2\alpha}{\lambda^3} - \frac{\alpha^3}{\lambda^3} \\
&= \frac{2\alpha}{\lambda^3}
\end{aligned}$$

A fairly horrible calculation, but we can check it in Mathematica. In fact we did this problem in the lab as the last question on the quiz. Let's do μ_4 in Mathematica.

```
In[1]:= <<Statistics‘ContinuousDistributions‘

In[2]:= dist = GammaDistribution[alpha, 1 / lambda]

                                   1
Out[2]= GammaDistribution[alpha, ------]
                                 lambda

In[3]:= f[x_] = PDF[dist, x]

                     -1 + alpha
                    x
Out[3]= ------------------------------------
         lambda x    1    alpha
        E         (------)      Gamma[alpha]
                   lambda

In[4]:= mu = Integrate[ x f[x], {x, 0, Infinity},
        Assumptions -> {alpha > 0 && lambda > 0} ]

              -1 - alpha
        lambda           Gamma[1 + alpha]
Out[4]= ---------------------------------
              1    alpha
           (------)      Gamma[alpha]
            lambda

In[5]:= mu = FullSimplify[mu]

                    -1 - alpha
        alpha lambda
Out[5]= ----------------------
               1    alpha
            (------)
             lambda

In[6]:= mu = PowerExpand[mu]

        alpha
Out[6]= ------
        lambda
```

```
In[7]:= mu4 = Integrate[ (x - mu)^4 f[x], {x, 0, Infinity},
        Assumptions -> {alpha > 0 && lambda > 0} ]

               -4 - alpha           4
Out[7]= (lambda           (-3 alpha  Gamma[alpha] +

                2
>        6 alpha  Gamma[2 + alpha] - 4 alpha Gamma[3 + alpha] +

                                  1    alpha
>        Gamma[4 + alpha])) / ((------)      Gamma[alpha])
                                lambda

In[8]:= mu4 = PowerExpand[FullSimplify[mu4]]

        3 alpha (2 + alpha)
Out[8]= -------------------
                    4
              lambda
```

Thus we finally obtain

$$\mu_4 - \mu_2^2 = \frac{2\alpha(3+\alpha)}{\lambda^4}$$

and

$$\mathbf{M}_2 = \begin{pmatrix} \frac{\alpha}{\lambda^2} & \frac{2\alpha}{\lambda^3} \\ \frac{2\alpha}{\lambda^3} & \frac{2\alpha(3+\alpha)}{\lambda^4} \end{pmatrix} \tag{9.14}$$

So the asymptotic variance of $\hat{\alpha}_n$ is

$$\begin{aligned} \mathbf{G}\mathbf{M}_2\mathbf{G}' &= \begin{pmatrix} 2\lambda & -\lambda^2 \end{pmatrix} \begin{pmatrix} \frac{\alpha}{\lambda^2} & \frac{2\alpha}{\lambda^3} \\ \frac{2\alpha}{\lambda^3} & \frac{2\alpha(3+\alpha)}{\lambda^4} \end{pmatrix} \begin{pmatrix} 2\lambda \\ -\lambda^2 \end{pmatrix} \\ &= 2\alpha(1+\alpha) \end{aligned}$$

and

$$\hat{\alpha}_n \approx \mathcal{N}\left(\alpha, \frac{2\alpha(1+\alpha)}{n}\right) \tag{9.15}$$

9.3.6 Relative Efficiency

Definition 9.3.2 (Relative Efficiency).
The relative efficiency *of two estimators of the same parameter is the ratio of their mean squared errors.*

Lindgren (p. 260) adds an additional proviso that the ratio must not depend on the parameter, but this needlessly restricts the concept. Of course, if the relative efficiency does depend on the parameter then in actual practice you don't know exactly what it is because you don't know the true parameter value. However, you can still make useful statements comparing the estimators. It might be, for example, that the relative efficiency is large for all likely values of the parameter.

Unfortunately, this criterion is almost useless except in toy problems because it is often impossible to calculate mean squared errors of complicated estimators. A much more useful criterion is given in the following section.

9.3.7 Asymptotic Relative Efficiency (ARE)

Definition 9.3.3 (Asymptotic Relative Efficiency).
The asymptotic relative efficiency *of two consistent and asymptotically normal estimators of the same parameter is the ratio of their asymptotic variances.*

Expressed in symbols, if S_n and T_n are two estimators of θ and

$$\sqrt{n}(S_n - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$$
$$\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \tau^2)$$

then the asymptotic relative efficiency is the ratio of σ^2 to τ^2.

It is unimportant whether you say the ARE is σ^2/τ^2 or τ^2/σ^2. No one can remember which way is up anyway. It is much clearer if you say something like S_n is better than T_n and the ARE is 0.95. It is then clear that σ^2 is smaller than τ^2, because "better" means smaller asymptotic variance. Hence it is clear that the ARE is in this case σ^2/τ^2. Typically the ARE depends on the true distribution of the data.

Example 9.3.10 (Mean versus Median).
Whenever all the distributions in a statistical model are symmetric and have means, the center of symmetry is both the mean and the median. Hence both $\overline{X}_n$ and $\widetilde{X}_n$ are sensible estimators. Which is better?

Generally, it depends on the shape of the population distribution. For a concrete example, we will do the normal distribution. Then the asymptotic variance of $\widetilde{X}_n$ is given in Example 7.4.1, and the asymptotic variance of $\overline{X}_n$ is of course σ^2/n. Hence the sample mean is the better estimator and the ARE is $2/\pi = 0.6366$. Thus the sample median is only about 64% as efficient as the sample mean for normal populations.

For other population distributions the conclusion can be reversed and the sample median may be much better than the mean (Problem 9-12).

Why is ARE interesting? Why ratio of variances? Why not ratio of standard deviations for example? The reason is that ARE has a direct relation to actual costs. To get the same accuracy, we need the same variance. The asymptotic variances are σ^2/m and τ^2/n if we choose sample sizes m and n for the two estimators. So in order to have the same accuracy, we must have

$$m = \frac{\sigma^2}{\tau^2} n = \text{ARE} \times n$$

A large part of the costs of any random experiment will be proportional to the sample size. Hence ARE is the right scale, the one proportional to costs.

9.4 Interval Estimation

Since point estimators are never right, at least when the statistical model is continuous, it makes sense to introduce some procedure that is right some of the time. An *interval estimate* of a real-valued parameter θ is a random interval having endpoints that are statistics, call them $\hat{\theta}_L(\mathbf{X})$ and $\hat{\theta}_R(\mathbf{X})$, the L and R being for "left" and "right." The idea is that the interval estimate says the true parameter value is somewhere between these endpoints. The event $\hat{\theta}_L(\mathbf{X}) < \theta < \hat{\theta}_R(\mathbf{X})$ that the true parameter value is actually in the interval is described as saying the interval *covers* the parameter, and the probability of this event

$$P_\theta\big\{\hat{\theta}_L(\mathbf{X}) < \theta < \hat{\theta}_R(\mathbf{X})\big\} \tag{9.16}$$

is called the *coverage probability* of the interval estimator. This terminology, "interval estimate," "covers," and "coverage probability," is not widely used, appearing only in fairly technical statistical literature, but the same concepts are widely known under different names. The interval $\big(\hat{\theta}_L(\mathbf{X}), \hat{\theta}_R(\mathbf{X})\big)$ is called a *confidence interval* and the probability (9.16) is called the *confidence level*, conventionally expressed as a percentage. If the coverage probability is 0.95, then the interval is said to be a "95 percent confidence interval."

The careful reader may have noticed that an important issue was passed over silently in defining the coverage probability (9.16). As the notation indicates, the coverage probability depends on θ, but we don't know what the value of θ is. The whole point of the exercise is to estimate θ. If we knew what θ was, we wouldn't care about a confidence interval.

There are three solutions to this problem.

- Sometimes, by everything working out nicely, the coverage probability (9.16) does not actually depend on θ, so the issue goes away.

- Often, we can't calculate (9.16) exactly anyway and are using the central limit theorem or other asymptotic approximation to approximate the coverage probability. If the asymptotic approximation doesn't depend on θ, the issue goes away.

- Rarely, we can get a lower bound on the coverage probability, say

  $$P_\theta\big\{\hat{\theta}_L(\mathbf{X}) < \theta < \hat{\theta}_R(\mathbf{X})\big\} \ge p, \qquad \theta \in \Theta$$

  then we are entitled to call p expressed as a percentage the confidence level of the interval. This procedure is conservative but honest. It understates the actual coverage, but guarantees a certain minimum coverage.

9.4.1 Exact Confidence Intervals for Means

Theorem 9.6. *If X_1, …, X_n are i. i. d. $\mathcal{N}(\mu, \sigma^2)$ where σ is known, then*

$$\overline{X}_n \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \tag{9.17}$$

is a $100(1-\alpha)\%$ confidence interval for μ, where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

The notation in (9.17) means that the confidence interval is

$$\overline{X}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \overline{X}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \tag{9.18}$$

Typically, confidence levels of 90%, 95%, or 99% are used. The following little table gives the corresponding $z_{\alpha/2}$ values.

confidence level	z critical value
90%	1.645
95%	1.960
99%	2.576

As the heading of the second column says, the $z_{\alpha/2}$ values are often called "z critical values" for a reason that will become apparent when we get to tests of significance.

The proof of the theorem is trivial. Equation (9.18) holds if and only if

$$\left| \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \right| < z_{\alpha/2}$$

The fraction in the absolute value signs is a standard normal random variable by Theorem 7.12. Hence the confidence level is

$$P(|Z| < z_{\alpha/2}) = 1 - 2P(Z > z_{\alpha/2}) = 1 - \alpha$$

by the symmetry of the normal distribution and the definition of $z_{\alpha/2}$.

This theorem is not much use in practical problems because σ is almost never known. Hence the following.

Theorem 9.7. *If X_1, …, X_n are i. i. d. $\mathcal{N}(\mu, \sigma^2)$, then*

$$\overline{X}_n \pm t_{\alpha/2} \frac{S_n}{\sqrt{n}} \tag{9.19}$$

is a $100(1-\alpha)\%$ confidence interval for μ, where $t_{\alpha/2}$ is the $1-\alpha/2$ quantile of the $t(n-1)$ distribution.

The proof is again trivial, and follows exactly the same pattern as the previous theorem. Now, however, we can't use our little table of z critical values. We must use a big table of t critical values. The columns labeled 95, 97.5, and 99.5 in Table IIIb in the Appendix of Lindgren give the critical values $t_{\alpha/2}$ for 90%, 95%, and 99% confidence intervals respectively. (Why? Figure it out.) The bottom row of the table, labeled ∞ degrees of freedom gives the corresponding z critical values, so if you forget which column you need to use but can remember what the z critical value would be, that will tell you the right column. Also keep in mind that the degrees of freedom are $n-1$, not n.

9.4.2 Pivotal Quantities

The random variables

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1) \tag{9.20a}$$

used in the proof of Theorem 9.6 and

$$\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim t(n-1) \tag{9.20b}$$

used in the proof of Theorem 9.7 are called "pivotal quantities."

More generally, a random variable is called a *pivotal quantity* or *pivot* if its distribution does not depend on the true distribution of the data so long as that distribution remains in the statistical model under consideration. For a parametric model, this means the distribution of the the pivot does not depend on the parameters of the model. For a nonparametric model, we need the more general definition.

Any pivotal quantity that involves only one parameter can always be used to make confidence intervals for that parameter if its sampling distribution is known. If $g(\mathbf{X}, \theta)$ is a pivotal quantity with known sampling distribution, then we can find numbers a and b such that

$$P\{a < g(\mathbf{X}, \theta) < b\}$$

is our desired confidence level. Then

$$\{\, \theta \in \Theta : a < g(\mathbf{X}, \theta) < b \,\} \tag{9.21}$$

is the desired confidence interval, or perhaps to be precise we should say "confidence set" because the set (9.21) is not necessarily an interval, though it is in cases of practical interest.

Both theorems of the preceding section are examples of intervals derived by the method of pivotal quantities. Another interesting example is Example 8.8b in Lindgren which gives a confidence interval for the parameter of the $\text{Exp}(1/\theta)$ distribution (that is, θ is the mean) using an i. i. d. sample. The pivotal quantity is

$$\frac{2}{\theta} \sum_{i=1}^{n} X_i = \frac{2n\overline{X}_n}{\theta} \sim \text{chi}^2(2n)$$

That this random variable has the asserted distribution comes from Lemma 7.10. The exact assertion we need is given in the unnumbered displayed equation below (7.19) in the example following the lemma (recalling that $1/\theta = \lambda$).

Generally it is not clear how to choose a and b in (9.21) unless the sampling distribution of the pivot is symmetric, as it is for the confidence intervals in the preceding section. Lindgren in Example 8.8b chooses a so-called equal-tailed interval with a and b satisfying

$$P\{g(\mathbf{X}, \theta) < a\} = P\{b < g(\mathbf{X}, \theta)\}$$

but the only reason for doing this is the limitations of the chi-square table used to find a and b. With better tables or a computer, we find that if instead of the 5th and 95th percentiles of the $\text{chi}^2(20)$ used by Lindgren, we use the 0.086 and 0.986 quantiles we get the interval

$$\frac{2 \sum_i X_i}{36.35} < \theta < \frac{2 \sum_i X_i}{12.06}$$

which is about 8% shorter than the equal-tailed interval. In fact, this is the shortest possible 90% confidence interval based on this pivot.

9.4.3 Approximate Confidence Intervals for Means

If no pivotal quantity is known, there still may be an asymptotically pivotal quantity, a function $g_n(X_1, \ldots, X_n, \theta)$ satisfying

$$g_n(X_1, \ldots, X_n, \theta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1) \tag{9.22}$$

regardless of the true distribution of the data so long as that distribution remains in the statistical model under consideration. Then

$$\{\, \theta \in \Theta : |g_n(X_1, \ldots, X_n, \theta)| < z_{\alpha/2} \,\} \tag{9.23}$$

is an asymptotic $100(1 - \alpha)\%$ confidence interval for θ, meaning the coverage probability converges to $1 - \alpha$ as $n \to \infty$, and where, as before, $z_{\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution.

Theorem 9.8. *If X_1, ..., X_n are i. i. d. from a distribution having mean μ and finite variance σ^2 and S_n is any consistent estimator of σ, then*

$$\overline{X}_n \pm z_{\alpha/2} \frac{S_n}{\sqrt{n}} \tag{9.24}$$

is an asymptotic $100(1-\alpha)\%$ confidence interval for μ, where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

This is Theorem 9.2 restated in confidence interval jargon.

Example 9.4.1 (I. I. D. Exponential).
In his Example 8.8b Lindgren gave the following exact equal-tailed 90% confidence interval for the mean θ of an exponential model

$$\frac{2 \sum_i X_i}{31.4} < \theta < \frac{2 \sum_i X_i}{10.9} \tag{9.25}$$

Here we compare this with the asymptotic confidence interval. Since the variance of the exponential is $\sigma^2 = \frac{1}{\lambda^2} = \theta^2$, $\overline{X}_n$ is a consistent estimator of both θ and σ. Hence by the theorem

$$\overline{X}_n \pm 1.645 \frac{\overline{X}_n}{\sqrt{n}}$$

is an asymptotic 90% confidence interval for θ. In Lindgren's example $n = 10$ so the asymptotic interval works out to be

$$0.48 \overline{X}_n < \theta < 1.52 \overline{X}_n \tag{9.26}$$

For comparison with the exact interval (9.25), we rewrite this as

$$\frac{2 \sum_i X_i}{30.4} < \theta < \frac{2 \sum_i X_i}{9.60}$$

Since $2 \sum_i X_i / \theta$ has a $\text{chi}^2(20)$ distribution (see Example 8.8b in Lindgren), the exact confidence level of this interval is

$$P(9.60 < \chi^2_{20} < 30.4) = 0.911$$

Not perfect, but not too shabby, especially since $n = 10$ is not a very large sample size. We don't usually expect agreement this good.

A useful bit of terminology for discussing asymptotic confidence intervals is the following. In the example, the approximate confidence interval (9.26) has a *nominal* confidence level of 90%, meaning only that we are calling it a 90% interval ("nominal" meaning having a certain name). The *actual* confidence level turns out to be 91.1%. In most applications we have no idea what the actual confidence level of an asymptotic interval really is. The CLT assures us that the actual level is close to the nominal if the sample size is large enough. But we rarely know how large is large enough.

There is in general no reason to expect that the actual level will be greater than the nominal level. It just happened to turn out that way in this example. In another application, actual might be less than nominal.

Most people would find the performance of the asymptotic interval satisfactory in this example and would not bother with figuring out the exact interval. In fact, with the single exception of intervals based on the t distribution, very few exact intervals are widely known or used. None (except t) are mentioned in most introductory statistics courses.

9.4.4 Paired Comparisons

In this section and in the next several sections we cover what is probably the single most useful application of statistics: comparison of the means of two populations. These can be divided into two kinds: paired comparisons and comparisons using independent samples.

In this section we deal with paired comparisons, leaving the other to following sections. The message of this section is that paired comparison problems naturally transform to the one-sample problems already studied. Hence they involve no new theory, just a new application of old theory.

In a paired comparisons problem we observe i. i. d. bivariate data (X_i, Y_i), $i = 1, \ldots, n$. The parameter of interest is

$$\mu_X - \mu_Y = E(X_i) - E(Y_i) \tag{9.27}$$

The standard solution to this problem is to reduce the data to the random variables

$$Z_i = X_i - Y_i, \qquad i = 1, \ldots, n,$$

which are i. i. d. and have mean

$$\mu_Z = \mu_X - \mu_Y,$$

which is the parameter of interest. Hence standard one-sample procedures applied to the Z_i provide point estimates and confidence intervals in this case.

It is an important point that X_i and Y_i do not have to be independent. In fact it is sometimes better, in the sense of getting more accurate estimates of the parameter of interest, if they are dependent. The typical paired comparison situation has X_i and Y_i being different measurements on the same individual, say arm strength of left and right arms or MCAT scores before and after taking a cram course. When X_i and Y_i are measurements on the same individual, they are usually correlated.

The procedure recommended here that reduces the original data to the differences Z_i and then uses one-sample procedures is the only widely used methodology for analyzing paired comparisons. We will study other procedures for paired comparisons when we come to nonparametrics (Chapter 13 in Lindgren), but those procedures also use the same trick of reducing the data to the differences Z_i and then applying one-sample procedures to the Z_i. The only difference between the nonparametric procedures and those described here is that the nonparametric one-sample procedures are not based on the normal or t distributions and do not require normality of the population distribution.

9.4.5 Independent Samples

A more complicated situation where the paired difference trick is not appropriate arises when we have X_1, …, X_m i. i. d. from one population and Y_1, …, Y_n i. i. d. from another population. We assume the samples are independent, that is, $\mathbf{X} = (X_1, \ldots, X_m)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$ are independent random vectors. The parameter of interest is still (9.27).

Now there is no obvious way to do pairing if $m \neq n$. Even if $m = n$, the pairing is arbitrary and unnatural when X_i and Y_i are measurements on independent randomly chosen individuals.

Asymptotic Confidence Intervals

The obvious estimate of $\mu_X - \mu_Y$ is $\overline{X}_m - \overline{Y}_n$, which has variance

$$\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n} \tag{9.28}$$

An obvious estimator of (9.28) is

$$\frac{S_{X,m}^2}{m} + \frac{S_{Y,n}^2}{n}$$

where $S_{X,m}$ and $S_{Y,n}$ are any consistent estimators of σ_X and σ_Y, such as the usual sample standard deviations. We can use this to construct asymptotic confidence intervals for the parameter of interest as follows.

Theorem 9.9. *Suppose X_1, X_2, ... are i. i. d. with mean μ_X and variance σ_X^2 and Y_1, Y_2, ... are i. i. d. with mean μ_Y and variance σ_Y^2 and $(X_1, \ldots, X_m)$ and $(Y_1, \ldots, Y_n)$ are independent random vectors for each m and n and*

$$S_{X,m} \xrightarrow{P} \sigma_X, \qquad \text{as } m \to \infty \tag{9.29a}$$

$$S_{Y,n} \xrightarrow{P} \sigma_Y, \qquad \text{as } n \to \infty \tag{9.29b}$$

Then

$$\frac{(\overline{X}_m - \overline{Y}_n) - (\mu_X - \mu_Y)}{\sqrt{\frac{S_{X,m}^2}{m} + \frac{S_{Y,n}^2}{n}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1), \qquad \text{as } m \to \infty \text{ and } n \to \infty.$$

Partial Proof. We will prove the assertion of the theorem under the additional condition that m and n go to infinity in a certain special way, that they are given by sequences m_k and n_k such that

$$\frac{\frac{\sigma_X^2}{m_k}}{\frac{\sigma_X^2}{m_k} + \frac{\sigma_Y^2}{n_k}} \to \alpha \tag{9.30}$$

where α is some constant, necessarily satisfying $0 \le \alpha \le 1$, since the left hand side of (9.30) is always between zero and one.

Then the CLT says

$$\frac{\overline{X}_{m_k} - \mu_X}{\sigma_X / \sqrt{m_k}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

$$\frac{\overline{Y}_{n_k} - \mu_Y}{\sigma_Y / \sqrt{n_k}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

If we let α_k denote the left hand side of (9.30), then by Corollary 8.11 and Slutsky's theorem

$$\frac{(\overline{X}_{m_k} - \overline{Y}_{n_k}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{m_k} + \frac{\sigma_Y^2}{n_k}}} = \sqrt{\alpha_k}\frac{\overline{X}_{m_k} - \mu_X}{\sigma_X / \sqrt{m_k}} + \sqrt{1 - \alpha_k}\frac{\overline{Y}_{n_k} - \mu_Y}{\sigma_Y / \sqrt{n_k}}$$
$$\xrightarrow{\mathcal{D}} \sqrt{\alpha} Z_1 + \sqrt{1 - \alpha} Z_2,$$

where Z_1 and Z_2 are independent standard normal random variables. The limit is a linear combination of independent normal random variables, hence is normal. It has mean zero by linearity of expectation and variance $\alpha + (1 - \alpha) = 1$. Hence it is standard normal. Thus we have established

$$\frac{(\overline{X}_{m_k} - \overline{Y}_{n_k}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{m_k} + \frac{\sigma_Y^2}{n_k}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1) \tag{9.31}$$

Similarly

$$\frac{\frac{S_{X,m_k}^2}{m_k} + \frac{S_{Y,n_k}^2}{n_k}}{\frac{\sigma_X^2}{m_k} + \frac{\sigma_Y^2}{n_k}} = \alpha_k \frac{S_{X,m_k}^2}{\sigma_X^2} + (1 - \alpha_k)\frac{S_{Y,n_k}^2}{\sigma_Y^2} \xrightarrow{P} 1 \tag{9.32}$$

Combining (9.31) and (9.32) and using Slutsky's theorem gives the assertion of the theorem in the presence of our additional assumption (9.30).

The fact that the limit does not depend on α actually implies the theorem as stated (without the additional assumption) but this involves a fair amount of advanced calculus (no more probability) that is beyond the prerequisites for this course, so we will punt on the rest of the proof. □

Corollary 9.10.

$$\overline{X}_m - \overline{Y}_n \pm z_{\alpha/2} \sqrt{\frac{S_{X,m}^2}{m} + \frac{S_{Y,n}^2}{n}}$$

is an asymptotic $100(1 - \alpha)\%$ *confidence interval for* $\mu_X - \mu_Y$*, where* $z_{\alpha/2}$ *is the* $1 - \alpha/2$ *quantile of the standard normal distribution.*

Exact Confidence Intervals

Exact confidence intervals are more problematic. If we assume both populations are normal, then

$$\frac{(m-1)S_{X,m}^2}{\sigma_X^2} \sim \text{chi}^2(m-1) \quad \text{and} \quad \frac{(n-1)S_{Y,n}^2}{\sigma_Y^2} \sim \text{chi}^2(n-1) \tag{9.33}$$

and are independent. Hence the sum

$$\frac{(m-1)S_{X,m}^2}{\sigma_X^2} + \frac{(n-1)S_{Y,n}^2}{\sigma_Y^2} \tag{9.34}$$

is $\text{chi}^2(m+n-2)$. But this doesn't help much. Since it involves the population variances, which are unknown parameters, we can't use (9.34) to make a t distributed pivotal quantity that contains only the parameter of interest. Hence we can't use it to make an exact confidence interval.

In order to make progress, we need to add an additional assumption $\sigma_X = \sigma_Y = \sigma$. Then the variance of the point estimator $\overline{X}_m - \overline{Y}_n$ (9.28) becomes

$$\sigma^2 \left(\frac{1}{m} + \frac{1}{n} \right) \tag{9.35}$$

and (9.34) becomes

$$\frac{(m-1)S^2_{X,m} + (n-1)S^2_{Y,n}}{\sigma^2} \tag{9.36}$$

This gives us a useful pivot. Dividing the standard normal random variable

$$\frac{(\overline{X}_m - \overline{Y}_n) - (\mu_X - \mu_Y)}{\sigma \sqrt{\frac{1}{m} + \frac{1}{n}}}$$

by the square root of (9.36) divided by its degrees of freedom gives

$$\frac{(\overline{X}_m - \overline{Y}_n) - (\mu_X - \mu_Y)}{S_{p,m,n} \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t(m+n-2) \tag{9.37}$$

where

$$S^2_{p,m,n} = \frac{(m-1)S^2_{X,m} + (n-1)S^2_{Y,n}}{m+n-2}$$

It is clear from the fact that (9.36) is $\text{chi}^2(m+n-2)$ that $S^2_{p,m,n}$ is an unbiased estimator of σ^2, but that is not the reason we use it. Rather we use it because of the way the t distribution is defined. $S^2_{p,m,n}$ is called the "pooled" estimator of variance (hence the subscript p).

Thus under the assumptions that

- both samples are i. i. d.
- the samples are independent of each other
- both populations are exactly normal
- both populations have exactly the same variance

an exact $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$ is given by

$$\overline{X}_m - \overline{Y}_n \pm t_{\alpha/2} S_{p,m,n} \sqrt{\frac{1}{m} + \frac{1}{n}}$$

where $t_{\alpha/2}$ is the $1-\alpha/2$ quantile of the $t(m+n-2)$ distribution.

The sad truth about this procedure is that, although it is taught in many introductory statistics textbooks, it has (or should have) little practical application. The assumption $\sigma_X = \sigma_Y$ is the archetype of an assumption made for "reasons of mathematical convenience" rather than practical or scientific reasons. If we make the assumption, then we get an exact confidence interval. If we do not make the assumption, then we don't. But the assumption is almost never justified. If you don't know the true population means, how are you to know the population variances are the same?

Welch's Approximation

A better procedure was proposed by Welch in 1937. Unlike the the procedure of the preceding section, there is no set of assumptions that make it "exact." But it is correct for large m and n under any assumptions (like the asymptotic interval) and is a good approximation for small m and n. Welch's procedure uses the same asymptotically pivotal quantity

$$T = \frac{(\overline{X}_m - \overline{Y}_n) - (\mu_X - \mu_Y)}{\sqrt{\frac{S^2_{X,m}}{m} + \frac{S^2_{Y,n}}{n}}} \tag{9.38}$$

as the one used to make the asymptotic confidence interval. It just uses a better approximation to its sampling distribution than the $\mathcal{N}(0, 1)$ approximation appropriate for large m and n.

The key idea goes as follows. The numerator of (9.38) is normal. When standardized, it becomes

$$Z = \frac{(\overline{X}_m - \overline{Y}_n) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma^2_X}{m} + \frac{\sigma^2_Y}{n}}} \tag{9.39}$$

Recall that a random variable has a t distribution if it is a standard normal divided by the square root of a chi-square divided by its degrees of freedom. The quantity (9.38) can be rewritten $T = Z/\sqrt{W}$, where Z is given by (9.39) and

$$W = \frac{\frac{S^2_{X,m}}{m} + \frac{S^2_{Y,n}}{n}}{\frac{\sigma^2_X}{m} + \frac{\sigma^2_Y}{n}} \tag{9.40}$$

Unfortunately, W is not a chi-square divided by its degrees of freedom, so (9.38) is not exactly t distributed. Welch's idea is that although W does not have exactly the desired distribution it will typically have approximately this distribution, so if we figure out the degrees of freedom ν for which a $\text{chi}^2(\nu)$ random variable divided by ν best approximates W, then the distribution of T will be approximately $t(\nu)$.

There could be several definitions of "best approximates." Welch's choice was to match moments. Rewrite W as

$$W = \lambda \frac{U}{m-1} + (1 - \lambda) \frac{V}{n-1} \tag{9.41}$$

where

$$\begin{aligned} U &= \frac{(m-1)S_{X,m}^2}{\sigma_X^2} \\ V &= \frac{(n-1)S_{Y,n}^2}{\sigma_Y^2} \end{aligned}$$

and

$$\lambda = \frac{\frac{\sigma_X^2}{m}}{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}$$

Since $U \sim \text{chi}^2(m-1)$ and $V \sim \text{chi}^2(n-1)$ and λ is a constant, we can easily calculate moments.

$$E(W) = \lambda \frac{m-1}{m-1} + (1-\lambda)\frac{n-1}{n-1} = 1$$

which is the right expectation for a chi-square divided by its degrees of freedom, and

$$\begin{aligned} \operatorname{var}(W) &= \left(\frac{\lambda}{m-1}\right)^2 2(m-1) + \left(\frac{1-\lambda}{n-1}\right)^2 2(n-1) \\ &= 2\left[\frac{\lambda^2}{m-1} + \frac{(1-\lambda)^2}{n-1}\right] \end{aligned}$$

Since if $Y \sim \text{chi}^2(\nu)$, then

$$\operatorname{var}\left(\frac{Y}{\nu}\right) = \frac{1}{\nu^2}\operatorname{var}(Y) = \frac{2}{\nu}$$

the Y/ν that gives the best approximation to W in the sense of having the right mean and variance is the one with

$$\frac{1}{\nu} = \frac{\lambda^2}{m-1} + \frac{(1-\lambda)^2}{n-1} \tag{9.42}$$

Thus we arrive at Welch's approximation. The distribution of (9.38) is approximated by a $t(\nu)$ distribution where ν is defined by (9.42).

There are two problems with this approximation. First, we have no t tables for noninteger degrees of freedom and must use computers to look up probabilities. Second, we don't know know ν and must estimate it, using

$$\hat{\nu} = \frac{\left(\frac{S_{X,m}^2}{m} + \frac{S_{Y,n}^2}{n}\right)^2}{\frac{1}{m-1}\left(\frac{S_{X,m}^2}{m}\right)^2 + \frac{1}{n-1}\left(\frac{S_{Y,n}^2}{n}\right)^2} \tag{9.43}$$

Thus (finally) we arrive at the approximate confidence interval based on Welch's approximation. An approximate $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$ is given by

$$\overline{X}_m - \overline{Y}_n \pm t_{\alpha/2} \sqrt{\frac{S_{X,m}^2}{m} + \frac{S_{Y,n}^2}{n}}$$

where $t_{\alpha/2}$ is the $1-\alpha/2$ quantile of the $t(\hat{\nu})$ distribution, $\hat{\nu}$ being given by (9.43).

Example 9.4.2.
We make a confidence interval for the difference of means using the data in Example 9.7b in Lindgren. The statistics are (p. 327 in Lindgren)

	n	$\overline{X}$	S^2
Soil A	6	24	18.000
Soil B	8	29	23.714

Our estimate of the degrees of freedom is

$$\hat{\nu} = \frac{\left(\frac{18.000}{6} + \frac{23.714}{8}\right)^2}{\frac{1}{5}\left(\frac{18.000}{6}\right)^2 + \frac{1}{7}\left(\frac{23.714}{8}\right)^2} = 11.643$$

To get the critical value for this noninteger degrees of freedom we must interpolate in Table IIIb in the Appendix of Lindgren. The critical values for a 95% confidence interval are 2.20 for 11 d. f. and 2.18 for 12 d. f. Interpolating gives 2.19 for 11.6 d. f. R gives

```
> qt(0.975, 11.643)
[1] 2.186245
```

but 2.19 is good enough for all practical purposes.

The 95% confidence interval is thus

$$24 - 29 \pm 2.1862 \sqrt{\frac{18.000}{6} + \frac{23.714}{8}}$$

which is $(-10.34, 0.34)$.

For comparison, the procedure of the preceding section gives an interval

$$24 - 29 \pm 2.18 \sqrt{\frac{5 \cdot 18.000 + 7 \cdot 23.714}{12} \left(\frac{1}{6} + \frac{1}{8}\right)}$$

which is $(-10.435, 0.435)$.

The two confidence intervals are very similar. Why bother with the more complicated procedure? Because the "exact" procedure makes an assumption which is almost certainly false and is hence indefensible. If we had only done the exact procedure we would have no idea how wrong it was. It is only after we have also used Welch's procedure that we see that *in this particular case* the simpler procedure worked fairly well. In other cases, when the variances or sample sizes are more uneven, there will be an unacceptably large difference between the two answers. For this reason, many statistical computing packages now use Welch's approximation as the primary method of analysis of data like this or at least provide it as an option. Several introductory statistics texts (including the one I use for Statistics 3011) now explain Welch's approximation and recommend its use, although this is still a minority view in textbooks. Textbooks are slow to catch on, and it's only been 60 years.

Example 9.4.3 (Worse Examples).
This analyzes two artificial examples where the standard deviations and sample sizes vary by a factor of 3. First consider

n	S
10	1
30	3

Then

	standard	
	error	d. f.
pooled	0.9733	38
Welch	0.6325	37.96

here "standard error" is the estimated standard deviation of the point estimate, the thing you multiply by the critical value to get the "plus or minus" of the confidence interval. The degrees of freedom, hence the critical values are almost the same, but the standard error using the "pooled" estimator of variance is way too big. Thus the interval is way too wide, needlessly wide, because the only reason it is so wide is that is based on an assumption $\sigma_X = \sigma_Y$ that is obviously false.

Now consider

n	S
10	3
30	1

Then

	standard	
	error	d. f.
pooled	0.6213	38
Welch	0.9661	9.67

Here the "exact" procedure is more dangerous. It gives confidence intervals that are too narrow, not only wrong but wrong in the wrong direction, having far less than their nominal coverage probability.

For example, consider a difference of sample means of 2.0. Then the "exact" procedure gives a 95% confidence interval $2 \pm 2.0244 \cdot 0.6213$ which is

$(0.742, 3.258)$, whereas Welch's procedure gives a a 95% confidence interval $2 \pm 2.2383 \cdot 0.96609$ which is $(-0.162, 4.162)$. Using Welch's approximation to calculate probabilities, the coverage probability of the "exact" interval is about 77.7%. Of course, its coverage probability would be the nominal level 95% if the assumption of equal population variances were true, but here it is obviously false. Welch's approximation isn't exact, so we don't know what the true coverage probability actually is, but it is surely far below nominal.

9.4.6 Confidence Intervals for Variances

Sometimes confidence intervals for variances are wanted. As usual, these come in two kinds, asymptotic and exact.

Asymptotic Intervals

An asymptotic interval for σ^2 can be derived from asymptotic distribution for V_n given by Theorem 7.16 and the "plug-in" theorem (Theorem 9.2).

Theorem 9.11. *If $X_1, \ldots, X_n$ are i. i. d. from a distribution having finite fourth moments, V_n is given by* (7.16) *and $M_{4,n}$ is any consistent estimator of μ_4, for example, the fourth sample central moment*

$$M_{4,n} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^4$$

then

$$V_n \pm z_{\alpha/2} \sqrt{\frac{M_{4,n} - V_n^2}{n}} \tag{9.44}$$

is an asymptotic $100(1-\alpha)\%$ *confidence interval for σ^2, where $z_{\alpha/2}$ is the* $1-\alpha/2$ *quantile of the standard normal distribution.*

Example 9.4.4.
Suppose we have i. i. d. data, the sample size is $n = 200$, the sample second central moment is $V_n = 1.7314$, and the sample fourth central moment is $M_{4,n} = 14.2728$.

Plugging this into (9.44), we get

$$1.96 \sqrt{\frac{14.2728 - 1.7314^2}{200}} = 0.46537$$

for the half-width of the asymptotic 95% confidence interval, that is, the interval is 1.73 ± 0.47 or $(1.27, 2.20)$.

Exact Intervals

If the data are assumed to be exactly normally distributed, then by Theorem 7.24

$$\frac{nV_n}{\sigma^2} = \frac{(n-1)S_n^2}{\sigma^2} \sim \text{chi}^2(n-1)$$

and this is a pivotal quantity that can be used to make an exact confidence interval for σ^2 or σ. The calculations are almost exactly like those for the mean of the exponential distribution discussed in Section 9.4.2 that also involved a chi-square distributed pivotal quantity.

Theorem 9.12. *If $X_1, \ldots, X_n$ are i. i. d. $\mathcal{N}(\mu, \sigma^2)$ random variables, V_n is given by* (7.16), *and $0 < \beta < \alpha < 1$, then*

$$\frac{nV_n}{\chi^2_{1-\alpha+\beta}} < \sigma^2 < \frac{nV_n}{\chi^2_\beta} \tag{9.45}$$

is an exact $100(1-\alpha)\%$ confidence interval for σ^2, where χ^2_β is the β-th quantile of the chi-square distribution with $n-1$ degrees of freedom.

Of course, one can replace nV_n by $(n-1)S_n^2$ both places it appears in (9.45), if one pleases. Usually, one uses $\beta = \alpha/2$ in (9.45), giving a so-called "equal-tailed" interval (equal chance of missing high or low), but other choices of β also give valid confidence intervals, and such intervals may be shorter than the equal-tailed interval.

Example 9.4.5.
Consider the data in Example 8.9a in Lindgren for which $n = 14$ and $S_n^2 = 85.912$. Suppose we want a 95% confidence interval for σ^2. To get an equal-tailed interval, we look up the 0.025 and 0.975 quantiles of the chi^2(13) distribution in Table Vb of Lindgren. They are 5.01 and 24.7. Hence an exact 95% confidence interval is given by

$$\frac{(n-1)S_n^2}{24.7} < \sigma^2 < \frac{(n-1)S_n^2}{5.01}$$

which after plugging the values of n and S_n^2 becomes

$$45.15 < \sigma^2 < 222.98.$$

Taking square roots gives us a 95% confidence interval for σ

$$6.72 < \sigma < 14.93.$$

As always, we can find a shorter interval if we give up on the equal-tailed idea and use a computer search to find the β that gives the shortest interval for the desired confidence level. The β will depend on whether we are getting an interval for σ^2 or for σ. For σ^2, the optimal β is 0.0465 and the corresponding 95% confidence interval

$$36.12 < \sigma^2 < 192.90.$$

For σ, the optimal β is 0.0414 and the corresponding 95% confidence interval is

$$6.30 < \sigma < 14.08.$$

The Ratio of Variances (Two Samples)

We now go back to the situation of two independent samples from two populations studied in Section 9.4.5. The samples are $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ and the sample variances are denoted $S^2_{X,m}$ and $S^2_{Y,n}$, respectively. Then (9.33) gives the sampling distributions of these sample variances. If we divide the chi-square random variables by their degrees of freedom and form the ratio, we get an F random variable

$$\frac{S^2_{X,m}}{S^2_{Y,n}} \cdot \frac{\sigma^2_Y}{\sigma^2_X} \sim F(m-1, n-1).$$

Hence if a and b are numbers such that $P(a < X < b) = 1 - \alpha$ when $X \sim F(m-1, n-1)$, then

$$a\frac{S^2_{Y,n}}{S^2_{X,m}} < \frac{\sigma^2_Y}{\sigma^2_X} < b\frac{S^2_{Y,n}}{S^2_{X,m}}$$

is a $100(1-\alpha)\%$ confidence interval for the ratio of variances. Taking square roots gives a confidence interval for the ratio of standard deviations.

Of course, there are asymptotic confidence intervals for ratios of variances (or differences of variances) that do not require normal data (Problem 9-19).

9.4.7 The Role of Asymptotics

Why do asymptotics as the sample size goes to infinity matter? Real data have a sample size that is not going anywhere. It just is what it is. Why should anyone draw comfort from the fact that if the sample size were very large, perhaps billions of times larger than the actual sample size, the asymptotics would give a good approximation to the correct sampling distribution of the estimator?

The answer is, of course, that no one does draw comfort from that. What they draw comfort from is that asymptotics actually seem to work, to provide good approximations, at relatively small sample sizes, at least in simple well-behaved situations. Hence the rules of thumb promulgated in introductory statistics books that $n > 30$ is enough to apply "large sample theory" in i. i. d. sampling, except for the binomial distribution and contingency tables,[2] where the rule is the expected value in each cell of the table should be at least five. These rules are known to be simplistic. For skewed distributions n must be larger than 30 for good approximation, much larger if the distribution is highly skewed. Similarly, there are cases where the contingency table rule holds but the distribution of the chi-square statistic is not well approximated by the chi-square distribution. But the rules of thumb are good enough so that textbook authors do not feel negligent in teaching them.

[2] The chi-square test for contingency tables gets us ahead of ourselves. This is the subject of much of Chapter 10 in Lindgren, which we will get to eventually. But as long as we are discussing rules of thumb, we might as well mention all of them, and this is all there are.

If one is worried about the validity of asymptotics, the standard cure is to look at computer simulations. Sometimes simulations show that asymptotic approximations are bad, but asymptotics look good in simulations often enough that people keep on using asymptotics.

So people don't use asymptotics because of the theorems. They use asymptotics because most of the time they actually work in practice. That they work is not something explained by asymptotic theory, because theory only says they work for sufficiently large n. There is no guarantee for the actual n in an actual application.

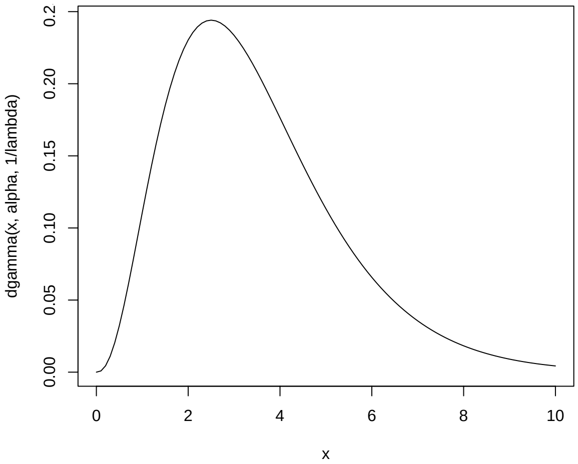
> *Asymptotic theory is only a heuristic. It is a device for producing approximations that may or may not be any good.*

Whether they are any good in an actual application, is something on which the theory is silent.

> *If you are worried the validity of asymptotics, you do simulations. Theory is no help.*

Example 9.4.6 (A Simulation Study).
In Example 9.3.9 and Problem 9-11 we derived the asymptotic distributions of the method of moments estimators of the two parameters of the gamma distribution. What if we are curious whether the asymptotics are good for sample size $n = 25$? Whether the asymptotics are valid also will depend on the shape of the distribution. Skewed distributions require larger sample sizes. Hence it will depend on the shape parameter α, but not on the scale parameter λ. Let's check the case $\alpha = 3.5$. Which a plot shows is moderately skewed.

```
> alpha <- 3.5
> lambda <- 1       # irrelevant, choose something
> curve(dgamma(x, alpha, 1 / lambda), from=0, to=10)
```



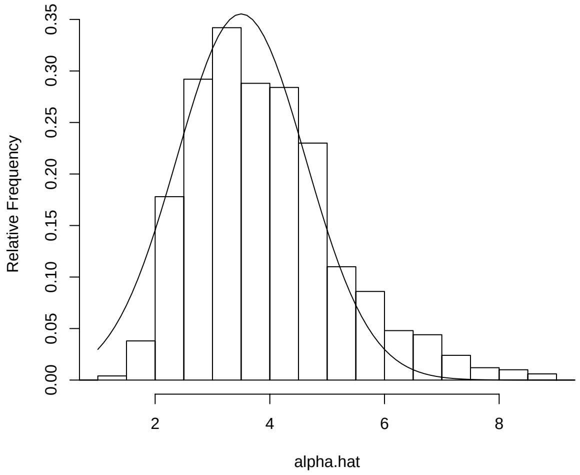
Now what we want to do is simulate lots of random samples of size n and see what the distribution of one of these statistics is. The following code does the job.

```
> alpha <- 3.5
> lambda <- 1       # irrelevant, choose something
> n <- 25           # sample size
> nsim <- 1000      # number of simulations to do
> alpha.hat <- double(nsim)
> for (i in 1:nsim) {
+     x <- rgamma(n, alpha, 1 / lambda)
+     xbar <- mean(x)
+     v <- var(x)
+     alpha.hat[i] <- xbar^2 / v
+ }
```

This only takes a few seconds of computer time. Now we look at a histogram of the data and compare it to the normal density of the asymptotic distribution. Actually, a histogram of all the `alpha.hat` values (not shown) when I did this simulation (of course, you would get different results because the results are random) was clearly nonnormal because it contained two "outliers" very far from the rest of the data. Below is a histogram of all but those two outliers produced by the following R statements

```
> hist(alpha.hat, probability=TRUE, xlim=c(1,9),
+     breaks=seq(0,20,.5))
> curve(dnorm(x, alpha, sqrt(2 * alpha * (1 + alpha) / n)),
+     from=1, to=9, add=TRUE)
```

where in the last line the standard deviations comes from (9.15).



As can be clearly seen from the histogram, the middle of the distribution of $\hat{\alpha}_n$ is clearly skewed and so not very normal (because the normal distribution is symmetric, not skewed). The outliers in the distribution of $\hat{\alpha}_n$ are not a big problem. We are not usually interested in the distribution far out in the tails. The skewness is a big problem. Because of it we should have asymmetric confidence intervals (not $\hat{\alpha}_n$ plus or minus the same thing), and asymptotics doesn't do that. (At least the kind of asymptotics we've studied doesn't do that. So-called "higher order" asymptotics, does correct for skewness, but that subject is beyond the scope of this course.)

Hence the simulation study shows that asymptotics doesn't work, at least for $n = 25$ and $\alpha = 3.5$. For large enough n, it will work, regardless of the value of α. For larger α it will work better for the same n, because the distribution of the X_i will be less skewed. For example, another simulation study (not shown), which I did to check my algebra in deriving the asymptotics, showed that the asymptotics worked fine for $n = 100$ and $\alpha = 2.3$ (there was only a little skewness visible in the histogram, the fit was pretty good).

9.4.8 Robustness

"Robustness" is a word with a lot of meanings in statistics. All of the meanings have something to do with a procedure being insensitive to violation of its assumptions. The differences have to do with what kind of violations are envisaged and what effects are considered important.

Asymptotic Robustness

A confidence interval is *asymptotically robust* or *asymptotically distribution free* for a specified statistical model if it is based on an asymptotically pivotal quantity (the property of being asymptotically pivotal depending, of course, on the model).

Example 9.4.7 (One-Sample t Intervals).
The "exact" confidence interval for the population mean given by Theorem 9.7, which uses the t distribution and assumes normal data, is asymptotically robust (more precisely asymptotically distribution free within the class of all distributions with finite variance) because it is based on the same pivotal quantity as the "large sample" interval given by Theorem 9.8, which needs no assumption about the data except finite variance. Thus we say these one-sample t confidence intervals are robust against departures from the normality assumptions.

Of course asymptotic robustness is inherently a "large sample" result. It may not say much about small sample sizes. Hence the analysis above does not really justify use of the t distribution for small sample sizes when we are worried that the population may be nonnormal. However, one can make the following argument. While it is true that the t confidence intervals are no longer exact if we do not assume exact normality, it is clear that we should make *some* adjustment of the critical value for small sample sizes. Just using the asymptotic interval based on the z critical value is obviously wrong. While the t critical value may not be exactly right, at least it will be a lot closer to the right thing than using the z critical value. If we are really worried we could do some simulations, which would show us that so long as the population distribution is symmetric or close to symmetric the distribution of $\sqrt{n}(\overline{X}_n - \mu)/S_n$ is very well approximated by the $t(n-1)$ distribution even when the population distribution has much heavier tails than normal. When the population is highly skewed, then the distribution of $\sqrt{n}(\overline{X}_n - \mu)/S_n$ is also skewed when n is small and hence cannot be well approximated by a t distribution, which, of course, is symmetric.

Example 9.4.8 (Two-Sample t Intervals).
The "exact" confidence interval for the difference of population means using the pooled estimator of variance is not asymptotically robust within the class of all distributions with finite variances, because the asymptotic distribution of the pivotal quantity (9.37) depends on the ratio of variances σ_X^2/σ_Y^2. Welch's approximate confidence interval is asymptotically robust within this class because it uses the same pivotal quantity as the asymptotic interval.

Both intervals are robust against departures from normality, but the "exact" interval is not robust against departures from its extra assumption $\sigma_X^2 = \sigma_Y^2$.

If we wanted to be pedantic we could say that the "exact" two-sample interval is asymptotically robust within the class of all distributions with finite variances and satisfying the additional assumption $\sigma_X^2 = \sigma_Y^2$, but this would only satisfy a pedant. It only emphasizes that the critical assumption of equality of population variances cannot be violated without destroying any nice properties the procedure is supposed to have. Calling that "robust" is perverse, although it does satisfy the technical condition. Robustness is defined relative to a statistical model. You can always make up a statistical model that makes a procedure robust with respect to that model. The question is whether that model is interesting.

Example 9.4.9 (Variances).
The "exact" confidence interval for the variance using the chi-square distribution is not asymptotically robust with the class of all distributions with finite fourth moments. This is clear because it is not based on the same pivotal quantity as the asymptotic confidence interval given by Theorem 9.11. Hence the "exact" interval is not robust against departures for normality. It critically depends on the property $\mu_4 = 3\sigma^4$ of the normal distribution.

Consider the data from Example 9.4.4 which were computer simulated from a Laplace (double exponential) distribution. The reason for using this distribution as an example is that it has heavier tails than the normal distribution, but not too heavy (the distribution still has moments of all orders, for example). The so-called exact 95% confidence interval for the variance using Theorem 9.12 is

$$1.44 < \sigma^2 < 2.14$$

but here this interval is inappropriate because the normality assumption is incorrect.

In Example 9.4.4 we calculated the correct asymptotic interval using the sample fourth central moment to be

$$1.27 < \sigma^2 < 2.20.$$

Comparing the two, we see that the correct interval is longer than the so-called exact interval based on what is in this case an incorrect assumption (of normality of the population distribution). Hence the so-called exact but in fact incorrect interval will have insufficient coverage. How bad this interval will be, whether it will have 90% coverage or 80% coverage or what instead of its nominal 95% coverage only a simulation study can tell. But we are sure it won't have its nominal coverage, because its assumptions do not hold.

Although we haven't worked out the correct asymptotic interval for the ratio of variances, we can easily believe the "exact" interval for the ratio of variances is also not robust and depends critically on the normality assumption.

These robustness considerations are important. You can find in various textbooks strong recommendations that the "exact" procedures that we have just found to be nonrobust should never be used. Simulations show that they are so critically dependent on their assumptions that even small violations lead to large errors. The robustness analysis shows us why.

Breakdown Point

This section takes up a quite different notion of robustness. The *breakdown point* of a point estimator is the limiting fraction of the data that can be dragged off to infinity without taking the estimator to infinity. More precisely if for each n we have an estimator $T_n(x_1, \ldots, x_n)$ which is a function of the n data values and k_n is the largest integer such that k_n of the x_i can be taken to infinity (with the other $n - k_n$ remaining fixed) and $T_n(x_1, \ldots, x_n)$ remain bounded, then $\lim_{n\to\infty} k_n/n$ is the breakdown point of the sequence of estimators T_n.[3]

The idea behind this technical concept is how resistant the estimator is to junk data. Roughly speaking, the breakdown point is the fraction of junk the estimator can tolerate. Here "junk" is generally considered to consist of gross errors, copying mistakes and the like, where recorded data has nothing whatsoever to do with the actual properties supposedly measured. It can also model rare disturbances of the measurement process or individuals that wind up in a sample though they weren't supposed to be and similar situations.

Example 9.4.10 (The Mean).
The sample mean has breakdown point zero, because

$$\frac{x_1 + \cdots + x_n}{n} \to \infty, \qquad \text{as } x_i \to \infty$$

[3] Some authorities would call this the *asymptotic* breakdown point, since they only use "breakdown point" to describe finite sample properties, that is, they say k_n/n is the breakdown point of T_n. But that needlessly complicates discussions of estimators since k_n is typically a complicated function of n, but the limit is simple.

with x_j, $i \neq j$ fixed. Hence $k_n = 0$ for all n.

Thus the sample mean tolerates zero junk and should only be used with perfect data.

Example 9.4.11 (The Median).
The sample median has breakdown point one-half. If n is odd, say $n = 2m + 1$, then we can drag off to infinity m data points and the sample median will remain bounded (in fact it will be one of the other $m + 1$ data points left fixed). Thus $k_n = \lfloor n/2 \rfloor$ when n is odd. When n is even, say $n = 2m$, then we can drag off to infinity $m - 1$ data points and the sample median will remain bounded, since we are leaving fixed $m + 1$ points and two of them will be the two points we average to calculate the sample median. Thus $k_n = n/2 - 1$ when n is even. In either case k_n is nearly $n/2$ and clearly $k_n/n \to 1/2$ as $n \to \infty$.

This example shows why the finite-sample notion of breakdown points is not so interesting.

Thus we see that the while the mean tolerates only perfect data, the median happily accepts any old junk and still gets decent answers. This is not to say that junk doesn't affect the median at all, only that any amount of junk up to 50% doesn't make the median completely useless. That wouldn't seem like a very strong recommendation until we remember that the mean is completely useless when there is any junk at all no matter how little.

9.5 Tests of Significance

In one sense we are now done with this chapter. This section is just a rehash of what has gone before, looking at the same stuff from a different angle. In another sense we are only half done. Tests of significance (also called hypothesis tests) are as important as confidence intervals, if not more important. So we have to redo everything, this time learning how it all relates to tests of significance. Fortunately, the redo won't take as much time and effort as the first time through, because all of the sampling theory is the same.

The simple story on tests of significance is that they are essentially the same thing as confidence intervals looked from a slightly different angle.

Example 9.5.1 (Difference of Population Proportions).
Suppose two public opinion polls are taken, one four weeks before an election and the other two weeks before. Both polls have sample size 1000. The results in percents were

	1st poll	2nd poll
Jones	36.1	39.9
Smith	30.1	33.0
Miller	22.9	17.4
Undecided	10.9	9.7

The typical reporter looks at something like this and says something like "Jones and Smith both gained ground since the last poll two weeks ago, Jones picking up 4 percent and Smith 3 percent, while Miller lost ground, losing 5 percentage points." This is followed by "news analysis" which reports that Jones is picking up support among college educated voters or whatever. Somewhere down toward the end of the article there may be some mention of sampling variability, a statement like "the polls have a margin of error of 3 percentage points," but it's not clear what anyone is supposed to make of this. Certainly the reporter ignored it in his analysis.

A skeptic might ask the question: has *anything* really changed in two weeks? We know the poll results are random. They are not the true population proportions but only estimates of them (the sample is not the population). Maybe the population proportions haven't changed at all and the apparent change is just chance variation. We don't yet know how to analyze the question of whether anything at all has changed in two weeks—we will get to this in Section 10.5 in Lindgren—but we do know how to analyze whether anything has changed in regard to one candidate, say Jones. The number of people in the samples who expressed a preference for Jones, 361 two weeks ago and 399 now, are binomial random variables with success probabilities p and q (the population proportions). These are estimated by the sample proportions $\hat{p} = 0.361$ and $\hat{q} = 0.399$. An asymptotic confidence interval for $q - p$ is (9.66). Plugging in the numbers gives the 95% confidence interval

$$0.399 - 0.361 \pm 1.960\sqrt{\frac{0.361 \times 0.639}{1000} + \frac{0.399 \times 0.601}{1000}}$$

or 0.038 ± 0.0425, which is $(-0.0045, 0.0805)$. Multiplying by 100 gives the interval expressed in percent $(-0.45, 8.05)$.

Since the confidence interval contains zero, it is not clear that there has been any increase in voter preference for Jones. No change in preference corresponds to $q - p = 0$, which is a point in the confidence interval, hence is a parameter value included in the interval estimate. It is true that the confidence interval includes a much wider range of positive values than negative values, so the confidence interval includes big increases but only small decreases, but decreases or no change are not ruled out.

Thus it is a bit premature for reporters to be bleating about an increase in the support for Jones. Maybe there wasn't any and the apparent increase is just chance variation in poll results.

The argument carried out in the example is called a test of significance or a statistical hypothesis test. The hypothesis being tested is that there is no real change in the population proportions, in symbols $p = q$. Alternatively, we could say we are testing the complementary hypothesis $p \neq q$, because if we decide that $p = q$ is true this is equivalent to deciding that $p \neq q$ is false and vice versa. We need general names for these two hypotheses, and the rather colorless names that are generally used in the statistical literature are the *null hypothesis* and the *alternative hypothesis*. Lindgren denotes them H_0 and H_A, respectively. These are always two complementary hypotheses, each the negation of the other, so we could do with just one, but we usually wind up mentioning both in discussions, hence the names are handy.

Summarizing what was just said, there are two possible decisions a test of significance can make. It can decide in favor of the null or the alternative. Since exactly one of the two hypotheses is true, deciding that one is true is tantamount to deciding that the other is false. Hence one possible decision is that the null hypothesis is true and the alternative hypothesis is false, which is described as *accepting the null hypothesis* or *rejecting the alternative hypothesis*. The other possible decision, that the alternative hypothesis is true and the null hypothesis is false, is described as described as *accepting the alternative hypothesis* or *rejecting the null hypothesis*.

In the opinion poll example, the null hypothesis is $p = q$ and the alternative hypothesis is $p \neq q$. We accept the null hypothesis if the confidence interval covers the parameter value $q - p = 0$ hypothesized by the null hypothesis. Otherwise we reject the null and accept the alternative. Since the confidence interval $(-0.45, 8.05)$ covers the hypothesized value, we accept the null. We conclude that the hypothesis may well be true, there being no strong evidence against it. In less technical language we conclude that the apparent change in the poll results may be just chance variation. Hence there is no real evidence that Jones is gaining, and hence there is no point in any news analysis of the reasons the gain has occurred.

Of course, the result of the test depends on which confidence interval we use, in particular, on the confidence level. A 90% confidence interval for $q - p$ expressed in percent is $(0.23, 7.37)$. Since this interval does not contain zero, we now reject the null hypothesis and accept the alternative. Thus we come to the opposite conclusion, Jones really is gaining support.

Thus it isn't enough to simply state whether the null hypothesis is accepted or rejected. We must also give the confidence level. For reasons of tradition we actually give something a little bit different. If the test involves a $100(1 - \alpha)\%$ confidence interval, we say we did a test with *significance level* α. People who think it is cool to talk in jargon rather than plain words often call the significance level the "α level," making α a frozen letter in this context and thus violating the principle of "mathematics is invariant under changes of notation." "Significance level" is a much better name.

The significance level is the probability that the confidence interval fails to cover. When the null hypothesis is true and the confidence interval fails to cover, we reject the null hypothesis erroneously. Thus another way of describing the significance level that does not mention confidence intervals is that it is *the probability of erroneously rejecting the null hypothesis*.

You may now be wondering what tests of significance are worth if one can always make a test come out either way by simply choosing to use a higher or lower significance level. The answer is they are worthless if you only pay attention to the decisions ("accept" or "reject") and ignore the significance levels, but when the decision and the significance level are considered together a test does provide useful information. A test using the 0.05 level of significance will erroneously reject the null hypothesis 5% of the time. A test using the 0.10

level of significance will erroneously reject the null hypothesis 10% of the time. That is a weaker evidentiary standard. The 0.10 level test rejects the null, and it may be right in doing so, but it may also be wrong, and the probability of its being wrong is twice that of the 0.05 level test.

That is the basic story on tests of significance. We now begin a systematic development of the theory.

9.5.1 Interest and Nuisance Parameters Revisited

Recall that in Section 9.2.4 we divided parameters into *parameters of interest* and *nuisance parameters*. The parameter or parameters of interest are the ones we want to know something about, the parameter the confidence interval is for, or the parameters involved in the null and alternative hypotheses of a test of significance.

In Example 9.5.1 there are two parameters p and q. Neither is the parameter of interest. The parameter of interest is $q - p$. Thus we see that sometimes we have to reparameterize the model in order to make the parameter of interest one of the parameters of the model. For example, we could choose new parameters α and δ defined by

$$\begin{aligned} \alpha &= p + q \\ \delta &= q - p \end{aligned}$$

Then δ is the parameter of interest, and α is a nuisance parameter.

In general, we write $\theta = (\varphi, \psi)$, where φ is the parameter of interest and ψ is the nuisance parameter. Either or both can be vectors, so

$$\theta = (\theta_1, \ldots, \theta_{k+m}) = (\varphi_1, \ldots, \varphi_k, \psi_1, \ldots, \psi_m)$$

if there are k parameters of interest and m nuisance parameters.

In dealing with confidence intervals there is always exactly one parameter of interest. The confidence interval is an interval estimate of that parameter. There may be many nuisance parameters. When we are estimating a difference of means from independent samples (Section 9.4.5) the parameter of interest is $\mu_X - \mu_Y$. Everything else is a nuisance parameter. As in Example 9.5.1, the parameter of interest is not one of the original parameters. The reparameterization

$$\begin{aligned} \alpha &= \mu_X + \mu_Y \\ \delta &= \mu_X - \mu_Y \end{aligned}$$

makes δ the parameter of interest and α, σ^2_X and σ^2_Y the nuisance parameters.

9.5.2 Statistical Hypotheses

In general, a statistical hypothesis can be any statement at all about the parameters of interest. In actual practice, almost all tests involve two kinds of hypotheses.

- The null hypothesis specifies a value of the parameter of interest. Thus it can be written $\varphi = \varphi_0$, where φ_0 is a fixed known value.
- There is a single parameter of interest φ and the null hypothesis is of the form $\varphi \le \varphi_0$ or $\varphi \ge \varphi_0$, where φ_0 is a fixed known value.

When there is a single parameter of interest these have widely used names. A test of

$$\begin{aligned} H_0 &: \varphi \le \varphi_0 \\ H_A &: \varphi > \varphi_0 \end{aligned} \tag{9.46a}$$

is called a *one-tailed test* and H_A is called a *one-sided alternative* (and the same names are used if both inequalities are reversed). A test of

$$\begin{aligned} H_0 &: \varphi = \varphi_0 \\ H_A &: \varphi \ne \varphi_0 \end{aligned} \tag{9.46b}$$

is called a *two-tailed test* and H_A is called a *two-sided alternative*.

When there are several parameters of interest only (9.46b) makes sense so there is usually no need of distinguishing terminology, but in order to discuss the cases of one or several parameters of interest together we will call null hypotheses form in (9.46b) *equality-constrained null hypotheses*.

We also will need notation for the sets of parameter values corresponding to the hypotheses

$$\begin{aligned} \Theta_0 &= \{\, (\varphi, \psi) \in \Theta : \varphi \le \varphi_0 \,\} \\ \Theta_A &= \{\, (\varphi, \psi) \in \Theta : \varphi > \varphi_0 \,\} \end{aligned} \tag{9.47a}$$

in the case of a one-sided alternative and

$$\begin{aligned} \Theta_0 &= \{\, (\varphi, \psi) \in \Theta : \varphi = \varphi_0 \,\} \\ \Theta_A &= \{\, (\varphi, \psi) \in \Theta : \varphi \ne \varphi_0 \,\} \end{aligned} \tag{9.47b}$$

in the case of an equality-constrained null.

As we said above, one can in principle test any hypothesis, but hypotheses other than the two types just described lead to complexities far beyond the scope of this course (in fact beyond the scope of PhD level theoretical statistics courses). So these two kinds of tests are all we will cover. For now we will concentrate on tests of equality-constrained null hypotheses and leave one-tailed tests for a later section (they require only minor changes in the theory).

9.5.3 Tests of Equality-Constrained Null Hypotheses

Most tests of significance (all tests we will consider) are determined by a *test statistic* $T(\mathbf{X})$. The null hypothesis is rejected for large values of the test statistic and accepted for small values. More precisely, there is a number c called the *critical value* for the test such that the decision rule for the test is the following

$$\begin{array}{ll} T(\mathbf{X}) \ge c & \text{reject } H_0 \\ T(\mathbf{X}) < c & \text{accept } H_0 \end{array}$$

Exact Tests

The *significance level* of the test is the probability of rejecting H_0 when H_0 is in fact true, when this probability does not depend on the parameter so long as the parameter remains in Θ_0, that is,

$$\alpha = P_\theta(\text{reject } H_0) = P_\theta\bigl(T(\mathbf{X}) \ge c\bigr), \qquad \theta \in \Theta_0. \tag{9.48}$$

Note that, since the null hypothesis fixes the value of the parameter of interest (for tests we are considering in this section), this means

$$P_\theta(\text{reject } H_0) = P_{(\varphi_0, \psi)}(\text{reject } H_0)$$

does not depend on the value of the nuisance parameter ψ.

How can we arrange for this probability to not depend on the nuisance parameter? We use a pivotal quantity $g(\mathbf{X}, \varphi)$ that only contains the parameter of interest. By definition, its distribution does not depend on the parameter. More precisely, the c. d. f. of the pivotal quantity

$$F(x) = P_{(\varphi, \psi)}\bigl(g(\mathbf{X}, \varphi) \le x\bigr), \qquad (\varphi, \psi) \in \Theta. \tag{9.49}$$

does not depend on the parameter when the parameter value that is the argument of the pivotal quantity and the true parameter value (i. e., both φ's on the right hand side of the equation) are the same. Of course, $g(\mathbf{X}, \varphi)$ is not a statistic, since it depends on a parameter, but we are only interested right now in parameter values in Θ_0, which means $\varphi = \varphi_0$. Plugging in the hypothesised value φ_0 for φ does give us a statistic $g(\mathbf{X}, \varphi_0)$, and from (9.49) we see that its distribution does not depend on θ for $\theta \in \Theta_0$, which is what is required. Since any function of a statistic is a statistic, any function of $g(\mathbf{X}, \varphi_0)$ can be used as a test statistic.

Example 9.5.2 (t Tests).
Suppose X_1, ..., X_n are i. i. d. $\mathcal{N}(\mu, \sigma^2)$ and we wish to test

$$\begin{aligned} H_0 &: \mu = 0 \\ H_A &: \mu \ne 0 \end{aligned}$$

(μ is the parameter of interest and σ^2 is a nuisance parameter.) We know that

$$\frac{\overline{X}_n - \mu}{S_n / \sqrt{n}} \sim t(n-1)$$

is a pivotal quantity. Plugging in μ_0 for μ gives a statistic

$$T = \frac{\overline{X}_n - \mu_0}{S_n / \sqrt{n}}$$

the distribution of which is $t(n-1)$ when the null hypothesis it true (and μ_0 is the true value of the parameter of interest). Which function of T do we use as our test statistic? The analogy with confidence intervals and the symmetry of the t distribution suggest the absolute value $|T|$. Thus we determine the critical value by solving

$$P\bigl(|T| > c\bigr) = \alpha$$

or what is equivalent

$$P\bigl(T > c\bigr) = \alpha/2. \tag{9.50}$$

We denote the solution of (9.50) $c = t_{\alpha/2}$. It is, as we defined it throughout Section 9.4, the $1 - \alpha/2$ quantile of the $t(n-1)$ distribution.

Why would we ever want to test whether μ is zero? Remember paired comparisons, where the test for no difference of population means reduces to a one sample test of $\mu = 0$ after the data are reduced to the differences of the pair values.

Asymptotic Tests

Asymptotic tests work much the same way as exact tests. We just substitute an asymptotically pivotal quantity for an exactly pivotal quantity and substitute asymptotic approximations for exact probabilities.

Sometimes there is a choice of pivotal quantity, as the following examples show.

Example 9.5.3 (Binomial, One Sample).
Suppose X is $\text{Bin}(n, p)$ and we wish to test

$$\begin{aligned} H_0 &: p = p_0 \\ H_A &: p \neq p_0 \end{aligned}$$

where p_0 is a particular number between zero and one. There are two asymptotically pivotal quantities we used to make confidence intervals in this situation

$$g_{1,n}(X, p) = \frac{\hat{p}_n - p}{\sqrt{p(1-p)/n}}$$

and

$$g_{2,n}(X, p) = \frac{\hat{p}_n - p}{\sqrt{\hat{p}_n(1-\hat{p}_n)/n}}$$

where, as usual, $\hat{p}_n = X/n$. Both are asymptotically standard normal. Both can be used to make confidence intervals, although the latter is much easier to use for confidence intervals. When it comes to tests, both are easy to use. Plugging in the hypothesized value of the parameter gives test statistics

$$Z_1 = \frac{\hat{p}_n - p_0}{\sqrt{p_0(1-p_0)/n}} \tag{9.51a}$$

and

$$Z_2 = \frac{\hat{p}_n - p_0}{\sqrt{\hat{p}_n(1 - \hat{p}_n)/n}} \tag{9.51b}$$

The two-tailed test with significance level α rejects H_0 when $|Z_i| \ge z_{\alpha/2}$ where, as usual, $z_{\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution.

For large n the two test statistics will be very close to each other and the two tests agree with very high probability. Asymptotics gives no reason to choose one over the other. The vast majority of statistics textbooks, however, recommend the test using Z_1. There is a sense in which Z_1 is closer to the standard normal than Z_2. The variance of Z_1 is exactly one, whereas Z_2 does not even have a variance because the denominator is zero when $X = 0$ or $X = n$. Still, neither test is exact, and when n is large enough so that the asymptotics are working well both tests give similar answers. So there is no real reason other than convention for using one or the other. But convention is a good enough reason. Why get into fights with people whose introductory statistics course insisted that the test using Z_1 was the only right way to do it?

Example 9.5.4 (Binomial, Two Sample).
Suppose $X \sim \text{Bin}(m, p)$ and $Y \sim \text{Bin}(n, q)$ and we wish to test

$$\begin{aligned} H_0 &: p = q \\ H_A &: p \ne q \end{aligned}$$

The asymptotically pivotal quantity we used to make confidence intervals in this situation was

$$g_{1,n}(X, Y, p - q) = \frac{(\hat{p}_m - \hat{q}_n) - (p - q)}{\sqrt{\frac{\hat{p}_m(1-\hat{p}_m)}{m} + \frac{\hat{q}_n(1-\hat{q}_n)}{n}}}$$

where, as usual, $\hat{p}_m = X/m$ and $\hat{q}_n = Y/n$. Plugging in the hypothesized value under the null hypothesis ($p - q = 0$) gives the test statistic

$$Z_1 = \frac{\hat{p}_m - \hat{q}_n}{\sqrt{\frac{\hat{p}_m(1-\hat{p}_m)}{m} + \frac{\hat{q}_n(1-\hat{q}_n)}{n}}} \tag{9.52}$$

A different quantity that is pivotal under the null hypothesis uses a "pooled" estimator of p similar to the pooled estimator of variance used in the two-sample t confidence interval based on the assumption of equality of variances. Here there is nothing controversial or nonrobust about the assumption $p = q$. The theory of tests of significance requires a probability calculation assuming H_0, so that's what we do. Under the null hypothesis $X + Y \sim \text{Bin}(m + n, p)$, hence

$$\hat{r}_{m,n} = \frac{X + Y}{m + n} = \frac{m\hat{p}_m + n\hat{q}_n}{m + n}$$

is the sensible estimator of p (and q). Also, still assuming $p = q$, the variance of the numerator of Z_q is

$$\operatorname{var}(\hat{p}_m - \hat{q}_n) = p(1-p)\left(\frac{1}{m} + \frac{1}{n}\right)$$

Hence under H_0

$$Z_2 = \frac{\hat{p}_m - \hat{q}_n}{\sqrt{\hat{r}_{m,n}(1 - \hat{r}_{m,n})\left(\frac{1}{m} + \frac{1}{n}\right)}} \tag{9.53}$$

is also asymptotically standard normal. Thus we again have two test statistics. Either is asymptotically correct. Both will agree with very high probability when m and n are large. Neither is exact. Convention (meaning the vast majority of introductory statistics courses) recommends the test based on Z_2.

9.5.4 P-values

Tests of significance are often done when they do not decide a course of action. This is almost always the case when the issue is a scientific inference. Data are collected, a test of significance is done, a paper is written, and readers are left to judge what it all means. Even the readers make no decisions in the statistical sense, accepting or rejecting some hypothesis solely on the basis of the data reported in the paper.

In such situations it is absurd to simply report that the test of significance rejected H_0 at a particular level of significance chosen by the authors (or accepted H_0 if that is what happened). What if a reader wants a different level of significance?

When the test is based on a test statistic, there is a much more sensible procedure. Suppose our test statistic is $T(\mathbf{X})$ and the value for the actual data being analyzed is $T(\mathbf{x})$. This is the usual "big $\mathbf{X}$" for random vectors and "little $\mathbf{x}$" for possible values (in this case the actual observed value). The significance level of the test corresponding to the critical value c is

$$\alpha(c) = P_\theta\bigl(T(\mathbf{X}) \ge c\bigr),$$

which we assume is the same for all $\theta \in \Theta_0$. Note that as c increases the event $\{\, \mathbf{X} : T(\mathbf{X}) \ge c \,\}$ decreases, and hence $\alpha(c)$ is a decreasing function of c by monotonicity of probability. Since the test rejects H_0 if $T(\mathbf{x}) \ge c$ and otherwise accepts H_0, the null hypothesis is rejected for all critical values c such that $c \le T(\mathbf{x})$ and hence for all significance levels α greater than or equal to

$$\alpha\bigl(T(\mathbf{x})\bigr) = P_\theta\bigl\{T(\mathbf{X}) \ge T(\mathbf{x})\bigr\}.$$

Definition 9.5.1 (P-value).
*The P-*value *of a test based on a test statistic $T(\mathbf{X})$ is*

$$P_\theta\bigl\{T(\mathbf{X}) \ge T(\mathbf{x})\bigr\}$$

provided this does not depend on θ for $\theta \in \Theta_0$. (This definition will later be generalized in Definition 9.5.3).

Summarizing the argument preceding the definition, the relationship between P-values (P), significance levels (α), and decisions is

$$\begin{array}{ll} P \le \alpha & \text{reject } H_0 \\ P > \alpha & \text{accept } H_0 \end{array} \tag{9.54}$$

The P-value (according to textbooks also called "observed level of significance," but I have never seen that outside of textbooks) is the answer to the quandary about different readers wanting different levels of significance. If the scientists report the P-value, then every reader can choose his or her own individual α and apply the rule (9.54) to determine the appropriate decision.

Example 9.5.5 (Example 9.5.1 Continued).
The hypothesis test in Example 9.5.1 is a two-tailed test based on the test statistic $|Z_1|$ where Z_1 is given by (9.52), which has the observed value

$$\frac{0.399 - 0.361}{\sqrt{\frac{0.361 \times 0.639}{1000} + \frac{0.399 \times 0.601}{1000}}} = 1.752$$

Under the null hypothesis, Z_1 is asymptotically standard normal. Hence the P-value is

$$P(|Z_1| > z) \approx 1 - 2\Phi(z) = 2\Phi(-z)$$

where $z = 1.752$ is the observed value of the test statistic, and Φ is, as usual, the standard normal c. d. f. Thus the P-value is $2 \times 0.0399 = 0.0798$.

Interpretation of P-values

P-values have two different interpretations, one trivial and the other controversial. Both interpretations support the following slogan.

> *The lower the P-value, the stronger the evidence against H_0.*

What is controversial is what "strength of evidence" is supposed to mean.

The trivial sense in which the slogan is true is that, if we consider a group of readers with a range of individual significance levels, a lower P-value will convince more readers. So there is no question that a lower P-value is more evidence against H_0. What is controversial is the question: "How much more?"

The controversial interpretation of P-values is the following. When we reject H_0 there are two possibilities

- H_0 actually is false.
- H_0 actually is true, in which case the P-value measures the probability of an actual event $T(\mathbf{X}) \ge T(\mathbf{x})$, which can be stated in words as the probability of seeing data at least as extreme as the data actually observed, where "extreme" is defined by largeness of $T(\mathbf{X})$.

Thus smallness of the P-value is a measure of the improbability of the second possibility.

This "controversial" argument is perfectly correct and strongly disliked by many statisticians, who claim that most people cannot or will not understand it and will confuse the P-value with a quite different notion: the probability of H_0. So let us be very clear about this. Since in frequentist statistics the parameter is not considered a random quantity, H_0 is not an event and $P(H_0)$ is a meaningless expression. Thus not only is the P-value *not* the "probability of the null hypothesis" neither is anything else.

Bayesians do consider parameters to be random quantities and hence do consider $P(H_0)$ to be meaningful. Dogmatic Bayesians consider all non-Bayesian statistics to be bogus, hence they particularly dislike P-values (though they like tests as decision procedures). They write papers[4] with titles like "the irreconcilability of P-values and evidence" by which they mean irreconcilability with *Bayesian notions of evidence.* This paper shows that a P-value always overstates the evidence against H_0 as compared to standard Bayesian notions of evidence.

Bayesian versus frequentist arguments aside, there is nothing controversial about P-values, as the "trivial" argument above makes clear. Lindgren in Section 9.3 gives a long discussion of tests as evidence raising a number of troubling issues, but only the very last paragraph, which mentions the Bayesian view involves P-values *per se.* The rest are troubling issues involving all tests of significance, whether or not P-values are used. We will return to our own discussion of interpretation of tests of significance after we discuss one-tailed tests, themselves one of the most controversial issues.

9.5.5 One-Tailed Tests

Theory

One-tailed tests require changes of the definitions of significance level and P-value. For one-tailed tests, we can no longer arrange for (9.48) to hold. Since the null hypothesis no longer fixes the value of the parameter of interest (only asserts an inequality), the probability we previously used to define the significance level will now depend on the parameter. This leads to the following definition

Definition 9.5.2 (Significance Level).
The significance level *of a test of significance based on a test statistic $T(\mathbf{X})$ is*

$$\alpha = \sup_{\theta \in \Theta_0} P_\theta(\textit{reject } H_0) = \sup_{\theta \in \Theta_0} P_\theta\bigl(T(\mathbf{X}) \geq c\bigr). \tag{9.55}$$

In words, the significance level is *the maximum probability of erroneously rejecting the null hypothesis.* Note that (9.55) reduces to our former definition (9.48) in the special case where $P_\theta(\text{reject } H_0)$ is actually the same for all $\theta \in \Theta_0$.

[4] Berger and Sellke, "Testing a point null hypothesis: The irreconcilability of P values and evidence" (with discussion), *Journal of the American Statistical Association*, 82:112-122, 1987

When we carry this new definition through the argument about P-values we obtain the promised generalization of Definition 9.5.1

Definition 9.5.3 (P-value).
The P-value of a test based on a test statistic $T(\mathbf{X})$ is

$$\sup_{\theta \in \Theta_0} P_\theta\bigl\{T(\mathbf{X}) \ge T(\mathbf{x})\bigr\}$$

Suppose the two-tailed test of the hypotheses (9.46b) is based on a pivotal quantity

$$g(\mathbf{X}, \varphi) = \frac{a(\mathbf{X}) - \varphi}{b(\mathbf{X})} \tag{9.56}$$

having a symmetric distribution that does not depend on the true parameter value. The primary examples are the one-sample z and t tests based on the pivotal quantities (9.20a) and (9.20b) and the two-sample test with the pooled estimate of variance based on the pivotal quantity (9.37). Other examples are the sign test and the two Wilcoxon tests that we will meet when we get to nonparametrics (Chapter 13 in Lindgren).

If we follow what we did with two-tailed tests and plug in φ_0 for φ in (9.56), we obtain

$$T(\mathbf{X}) = \frac{a(\mathbf{X}) - \varphi_0}{b(\mathbf{X})} \tag{9.57}$$

The idea of a one-tailed test is to use $T(\mathbf{X})$ itself as the test statistic, rather than $|T(\mathbf{X})|$, which is what we used for two-tailed z and t tests.

Lemma 9.13. *For the test with test statistic* (9.57) *based on the pivotal quantity* (9.56)*, the significance level corresponding to the critical value c is*

$$\alpha = P_{\varphi_0}\bigl\{T(\mathbf{X}) \ge c\bigr\},$$

and the P-value is

$$P_{\varphi_0}\bigl\{T(\mathbf{X}) \ge T(\mathbf{x})\bigr\}.$$

Proof. What we must show is that

$$P_{\varphi_0}\bigl\{T(\mathbf{X}) \ge c\bigr\} = \sup_{\theta \in \Theta_0} P_\theta\bigl\{T(\mathbf{X}) \ge c\bigr\}. \tag{9.58}$$

The assumption that (9.56) is a pivotal quantity means

$$\begin{aligned} P_\theta\bigl\{g(\mathbf{X}, \varphi) \ge c\bigr\} &= P_\theta\left\{\frac{a(\mathbf{X}) - \varphi}{b(\mathbf{X})} \ge c\right\} \\ &= P_{(\varphi,\psi)}\bigl\{a(\mathbf{X}) - cb(\mathbf{X}) \ge \varphi\bigr\} \end{aligned} \tag{9.59a}$$

does not depend on φ and ψ. Now

$$P_\theta\bigl\{T(\mathbf{X}) \ge c\bigr\} = P_{(\varphi,\psi)}\bigl\{a(\mathbf{X}) - cb(\mathbf{X}) \ge \varphi_0\bigr\} \tag{9.59b}$$

does depend on φ and ψ, but for parameter values in the null hypothesis $\varphi \le \varphi_0$ monotonicity of probability implies that (9.59b) is less than or equal to (9.59a), that is,

$$\begin{aligned} P_{\varphi_0}\bigl\{T(\mathbf{X}) \ge c\bigr\} &= P_{(\varphi_0,\psi)}\bigl\{a(\mathbf{X}) - cb(\mathbf{X}) \ge \varphi_0\bigr\} \\ &= P_{(\varphi,\psi)}\bigl\{a(\mathbf{X}) - cb(\mathbf{X}) \ge \varphi\bigr\} \\ &\ge P_{(\varphi,\psi)}\bigl\{a(\mathbf{X}) - cb(\mathbf{X}) \ge \varphi_0\bigr\} \\ &= P_{\varphi}\bigl\{T(\mathbf{X}) \ge c\bigr\} \end{aligned}$$

whenever $\varphi \le \varphi_0$. And this implies (9.58). □

There is an entirely analogous lemma for asymptotic tests, which we won't bother to prove, since the proof is so similar.

Lemma 9.14. *Suppose*

$$g_n(\mathbf{X}, \varphi) = \frac{a_n(\mathbf{X}) - \varphi}{b_n(\mathbf{X})/\sqrt{n}}$$

is an asymptotically pivotal quantity converging to a standard normal distribution as $n \to \infty$, and let

$$T_n(\mathbf{X}) = g_n(\mathbf{X}, \varphi_0).$$

Then a one-tailed test of (9.46a) *rejects H_0 when $T_n \ge z_\alpha$, where z_α is the $1-\alpha$ quantile of the standard normal distribution. The P-value is*

$$P\bigl\{Z > T_n(\mathbf{x})\bigr\},$$

where Z is standard normal and $T_n(\mathbf{x})$ is the observed value of the test statistic.

Practice

The theory in the preceding section is complicated but in practice one-tailed tests are simple as falling off a log. We will just give an example.

Example 9.5.6 (t Tests).
The following data appeared in "Student's" original paper on the t distribution

	X	Y	Z
	0.7	1.9	1.2
	−1.6	0.8	2.4
	−0.2	1.1	1.3
	−1.2	0.1	1.3
	−0.1	−0.1	0.0
	3.4	4.4	1.0
	3.7	5.5	1.8
	0.8	1.6	0.8
	0.0	4.6	4.6
	2.0	3.4	1.4
mean	0.75	2.33	1.58
s. d.			1.23

In each row, X and Y are the additional hours of sleep gained by one patient while using two different soporific drugs (there are 10 patients in the study). The third column is $Z = Y - X$, our usual trick of reducing data for paired comparison to differences. We want to test whether there is a significant difference between the two drugs. The null hypothesis is $\mu_Z = 0$ for a two-tailed test or $\mu_Z \le 0$ for a one-tailed test. The test statistic is

$$t = \frac{1.58}{1.23/\sqrt{10}} = 4.062$$

For a one-tailed test, the P-value is

$$P(T > 4.062) = 0.0014.$$

For a two-tailed test, the P-value is

$$P(|T| > 4.062) = 0.0028.$$

(All we could get from Table IIIa in Lindgren is that the one-tailed P-value is between 0.001 and 0.002 and hence the two-tailed P-value between 0.002 and 0.004. I used a computer to get exact P-values.)

Note that by the symmetry of the t distribution

$$P(T > t) = 2P(|T| > t)$$

so long as $t > 0$. Hence

A two-tailed P-value is twice the one-tailed P-value

so long as the one-tailed P-value is less than one-half. This has nothing to do with the t distribution in particular. It holds whenever the sampling distribution of the test statistic under H_0 is symmetric. By symmetry, two tails have twice the probability of one.

9.5.6 The Duality of Tests and Confidence Intervals

With all this theory we have somewhat lost track of the simple notion we started out with, that tests are just confidence intervals viewed from another angle. This section ties up the loose ends of that notion.

For this section only, forget P-values. Think of a test as a decision procedure that either accepts or rejects H_0 and has a specified significance level. That is the notion of tests that has a simple relationship to confidence intervals.

The word "duality" in the section heading is a fancy mathematical word for the relation between two concepts that are basically two sides of the same coin. Either can be used to define the other. Tests of equality-constrained null hypotheses have exactly that relation to confidence intervals.

> For any $100(1-\alpha)\%$ confidence interval for a parameter θ, the test that rejects $H_0 : \theta = \theta_0$ if and only if the confidence interval does not contain θ_0 has significance level α.
>
> Conversely, for any test with significance level α, the set of parameter values θ_0 such that $H_0 : \theta = \theta_0$ is accepted is a $100(1-\alpha)\%$ confidence interval for θ.

Example 9.5.7 (*t* Tests Yet Again).

$$\overline{X}_n - t_{\alpha/2} \frac{S_n}{\sqrt{n}} < \mu < \overline{X}_n + t_{\alpha/2} \frac{S_n}{\sqrt{n}}$$

is a $100(1-\alpha)\%$ confidence interval for μ assuming normal data. This confidence interval contains μ_0 if and only if

$$\left| \frac{\overline{X}_n - \mu_0}{S_n / \sqrt{n}} \right| < t_{\alpha/2}$$

which is the criterion for accepting H_0 in the usual two-tailed t test of $H_0 : \mu = \mu_0$. And it works both ways. If we start with the test and work backwards we get the confidence interval.

The duality is not exact if we use different asymptotic approximations for the test and the confidence interval. For example, the standard way to do a confidence interval for the binomial distribution involves the pivotal quantity Z_2 given by (9.51b) but the standard way to do a test involves the pivotal quantity Z_1 given by (9.51a). Z_1 and Z_2 are very close when n is large, but they are not identical. Thus the test and confidence interval will not have exact duality (they would if both were based on the same asymptotically pivotal quantity). However, we can say they have "approximate duality."

One-tailed tests do not seem at first sight to have such a simple duality relationship, but they do. In order to see it we have to change our view of one-tailed tests. All of the one-tailed tests we have considered can also be considered as equality-constrained tests. A test having a test statistic of the form (9.57) can be considered either a test of the hypotheses

$$\begin{aligned} H_0 &: \theta \le \theta_0 \\ H_A &: \theta > \theta_0 \end{aligned}$$

(which is the way we usually consider it) or a test of the hypotheses

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_A &: \theta > \theta_0 \end{aligned}$$

The latter way of thinking about the test changes the statistical model. Since H_0 and H_A partition the parameter space, the parameter space for the first test is $\Theta = \mathbb{R}$ and the parameter space for the second test is $\Theta = \{\, \theta \in \mathbb{R} : \theta \ge 0 \,\}$. But this is the only difference between the two procedures. They have the same test statistic, the same P-value for the same data, and the same decision for the same significance level and same data.

Now we can apply duality of tests and confidence intervals. Consider the t test yet again. The one-tailed t test accepts $H_0 : \mu = \mu_0$ at significance level α when

$$\frac{\overline{X}_n - \mu_0}{S_n / \sqrt{n}} < t_\alpha$$

The set of μ_0 values for which H_0 is accepted, the μ_0 such that

$$\overline{X}_n - t_\alpha \frac{S_n}{\sqrt{n}} < \mu_0 < +\infty,$$

is thus a $100(1 - \alpha)\%$ confidence interval for the true parameter value μ.

Thus we get "one-tailed" confidence intervals dual to one-tailed tests. Such intervals are not widely used, but there is nothing wrong with them. They are perfectly valid confidence intervals. Occasionally they are wanted in real applications.

9.5.7 Sample Size Calculations

All of the problems in this section and the preceding section (tests and confidence intervals), at least those that involve numbers, emulate the most common kind of data analysis, that which is done *after* the data have been collected. In this section we discuss the other kind, done *before* data have been collected.

We're not talking about some sort of magic. What we're talking about is part of most grant proposals and other preliminary work done before any large expensive scientific experiment is done. Of course you can't really analyze data that haven't been collected yet. But you can do *some* calculations that at least give *some* idea how they will come out.

The main issue of interest is whether the proposed sample size is large enough. We know that statistical precision varies as the square root of the sample size (the so-called square root law). So we can always get as precise a result as we please if only we expend enough time, money, and effort. The trouble is that the square root law means that twice the precision costs four times as much, ten times the precision costs a hundred times as much, and so forth. So generally, you must settle for less precision than you would like.

So suppose you are asking for several hundred thousand dollars to do an experiment with sample size 200. Before the funding agency gives you the money, one issue (among many others) that they will want to carefully consider is whether the precision you will get with $n = 200$ is worth the money. After all, if an experiment with $n = 200$ is unlikely to answer any questions of scientific interest because of lack of precision, they should fund some other projects with more promise.

These calculations *before the data are collected* look very different for tests and confidence intervals, so we will look at them separately. We'll do the simpler of the two first.

Confidence Intervals

A large sample confidence interval for the population mean is

$$\overline{X}_n \pm z_{\alpha/2} \frac{S_n}{\sqrt{n}}$$

which is just (9.24) repeated. Before we collect data we will know neither $\overline{X}_n$ nor S_n. We will know the n we propose to use.

Not knowing $\overline{X}_n$ is not a problem here. It will be close to the true population mean μ if n is large, but we don't know μ. In fact the whole point of the experiment (if this sort of confidence interval is of interest) is to estimate μ. So in a way it's a good thing we don't know μ. It gives us something to do.

The question we want to answer, or at least get some idea about, is how wide our confidence interval will be. If we expect it to be too wide to be of any scientific value, then we need a larger sample size or a completely different sort of experiment. So what we need to do is get some idea of the likely size of the plus-or-minus (also called "half-width" because it is half the width of the confidence interval).

Now the half-width depends on three things

- The confidence level (through α), which we know.
- The sample size n, which we know.
- The sample standard deviation S_n, which we do not know.

So in order to make progress we need a guess about the likely size of S_n. We might have some data from a similar experiment that will give us the likely size. Or we might have to just guess. Depending on who does the guessing and how much they know, we might call the guess anything from "expert opinion" to a "wild guess." But no matter where it comes from, we need some number to plug in for S_n.

Questions like this are often phrased backwards. Rather than what half-width will we likely get for a specified confidence level and sample size, one asks what sample size is necessary to get a specified half-width.

Example 9.5.8.
Suppose we are going to do an experiment to measure the expected life time of a new type of light bulb. The old light bulbs had a mean life of 700 hours with a standard deviation of 500 hours. The new light bulbs are supposed to last a lot longer, about 1000 hours, but let's use the same standard deviation in our sample size calculation. With no data yet on the new light bulbs you can call $s = 500$ a guess (we're guessing the s. d. of the new will be the same as the old) or you can call it an estimate based on preliminary data even though the data isn't about the exact same process.

So what sample size do we need to get a half-width of 100 hours for a 95% confidence interval? That's saying we want

$$100 = 1.96 \frac{500}{\sqrt{n}}$$

Solving for n gives

$$n = \left(\frac{1.96 \times 500}{100} \right)^2 = 96.04$$

Of course, a sample size must be a round number. Typically, one rounds up to be conservative, giving $n = 97$ as the answer.

This looks a lot more precise than it really is. Don't forget that we plugged in a guess for S_n. The actual experiment won't produce a confidence interval with a half-width of exactly 100 hours, because S_n won't come out to be exactly 500. We have, however, done the best we could with what we had to work with. Certainly, $n = 97$ is a lot better than complete cluelessness.

The Power of a Hypothesis Test

We will do a similar sort of calculation for a hypothesis test presently, but before we can even discuss such a calculation we need to learn a new term. This new term is called the *power* of a test. It is closely related to the *significance level* of a test, but not exactly the same thing.

First we explain the concepts in words to make the similarities and differences clear.

- The *significance level* of a test is the probability of rejecting the null hypothesis when it is in fact *true*.
- The *power* of a test is the probability of rejecting the null hypothesis when it is in fact *false*, that is when the alternative hypothesis is true.

Thus both level and power are probabilities of the same event "reject H_0" but probabilities under different assumed parameter values.

In symbols the *significance level* is

$$\alpha = P_\theta(\text{reject } H_0), \qquad \theta \in \Theta_0 \tag{9.60a}$$

This is our simpler definition of significance level given by (9.48), which assumes that the probability in (9.60a) does not actually depend on θ for θ in the null hypothesis. Our more general definition (Definition 9.5.2) is more complicated, but we won't worry about that here.

$$\pi(\theta) = P_\theta(\text{reject } H_0), \qquad \theta \in \Theta_A \tag{9.60b}$$

Note that the left hand side in (9.60b) is a function of θ, which we call the *power function* of the test. That's one important difference between level and power. Ideally, the level does not depend on the parameter, that's what the notation in (9.60a) indicates (as the following comment says, if it *did* depend on the parameter we would have to use a more complicated definition). And this is the case in simple situations.

Why don't we arrange for power to be constant too? Well that would defeat the whole purpose of the test. We want

- Low significance level, the lower α the better.

- High power, the higher $\pi(\theta)$ the better.

Why is that? Both level and power refer to the same event "reject H_0" but under different conditions. The level is the probability of a bad thing, erroneously rejecting the null hypothesis when it is true. The power is the probability of a good thing, correctly rejecting the null hypothesis when it is false. So we want low probability of the bad thing (level) and high probability of the good thing (power).

But as probabilities level and power are special cases of the same thing $P_\theta(\text{reject } H_0)$. So the only way power could be constant is if it were the same as the level, which is no good at all. That's not the way to get low level and high power.

What do power functions look like? Lindgren Section 9.10 gives some examples. Power functions are not all alike. It is not the case "seen one, you've seen them all." But fortunately, it is the case "seen two, you've seen them all." Power functions of upper-tailed tests look like Figure 9-10 in Lindgren and those of lower-tailed tests look like mirror images of that figure. Power functions of two-tailed tests look like Figure 9-11 in Lindgren.

Tests of Significance

We now return to sample size calculations. These too are usually phrased as "backwards" questions. What sample size n do we need to achieve a specified power for a specified level test?

Example 9.5.9.
Suppose in the situation explained in Example 9.5.8 we want to do a one-tailed test of whether the new light bulbs are better than the old in the sense of having longer life. We assume the mean life, 700 hours, of the old light bulbs is a known constant (it isn't really, but it is based on much more data than we intend to collect about the new light bulbs, so this isn't too bad an approximation to the right thing, which would be a two-sample test). We will also assume $S_n = 500$, as in Example 9.5.8 even this is only a guess (we need that here too as will be seen presently).

So what sample size is needed for $\alpha = 0.05$ and power 0.99 at the alternative hypothesis of interest, which is 1000 hours mean life for the new bulbs? Just to be perfectly clear, the hypotheses being tested are

$$\begin{aligned} H_0 &: \mu = 700 \text{ hours} \\ H_A &: \mu > 700 \text{ hours} \end{aligned}$$

and the alternative at which we want to calculate the power is $\mu = 1000$ hours.

The event "reject H_0" is in other notation $\overline{X}_n > c$, where c is the critical value for the test. So the first thing we must do is determine c. The test is based on the asymptotically standard normal quantity

$$Z = \frac{\overline{X}_n - \mu}{S_n / \sqrt{n}} \tag{9.61}$$

In power calculations we must be very, very careful with this object. The random variable (9.61) is asymptotically standard normal when μ is the true population mean and *this differs when calculating level and power!* Calculation of the level of a test is done *assuming the null hypothesis*. So in the level calculation we use $\mu = 700$. But calculation of the power is done *assuming the alternative hypothesis*. In particular, our power calculation here will be based on the particular alternative of interest $\mu = 1000$.

As usual, to get a one-tailed "z test" with level $\alpha = 0.05$ we use the critical value on the z scale 1.645. That is we reject H_0 when

$$Z = \frac{\overline{X}_n - 700}{500 / \sqrt{n}} > 1.645$$

Solving for $\overline{X}_n$, this is the same as

$$\overline{X}_n > 700 + 1.645 \frac{500}{\sqrt{n}} = c \tag{9.62}$$

So that finishes our calculation about level. All subsequent calculation assumes $\mu = 1000$. Notice that the critical value c depends on n. We don't get a number. We get a formula.

Now we need to calculate the power at $\mu = 1000$. What is $P(\overline{X}_n > c)$ when $\mu = 1000$. This is just like any other normal probability calculation. The main difficulty is not to get anything confused with the previous calculation. It only requires care and attention to detail. To calculate this probability using a normal table, we need to standardize the number being looked up, which is c.

$$P(\overline{X}_n > c) \approx 1 - \Phi\left(\frac{c - \mu}{S_n / \sqrt{n}}\right)$$

where, as usual, Φ is the standard normal c. d. f.We want this to be 0.99, hence we want the Φ term itself to be 0.01, and from the bottom row of Table IIIb in Lindgren (or from R or Mathematica) we see that we need the argument of Φ to be -2.33 to achieve that. Thus we get another equation

$$\frac{c - \mu}{S_n / \sqrt{n}} = \frac{c - 1000}{500 / \sqrt{n}} = -2.33 \tag{9.63}$$

(Note again, and this is the last time I'll say it, we are using $\mu = 1000$ here).

Plugging (9.62) in here gives

$$\frac{700 + 1.645 \frac{500}{\sqrt{n}} - 1000}{500 / \sqrt{n}} = -2.33$$

or

$$700 + 1.645 \frac{500}{\sqrt{n}} - 1000 = -2.33 \frac{500}{\sqrt{n}}$$

or

$$3.975 \frac{500}{\sqrt{n}} = 300$$

which comes to $n = 43.89$ or (rounded up) $n = 44$.

I concede that a messy calculation like this leaves me in doubt. Let's check that we actually got the right level and size with this critical value

```
> n <- 44
> crit <- 700 + 1.645 * 500 / sqrt(n)
> 1 - pnorm(crit, 700, 500 / sqrt(n))
[1] 0.04998491
> 1 - pnorm(crit, 1000, 500 / sqrt(n))
[1] 0.990227
```

The definition of `crit` is taken from (9.62). We must call it something other than c, because c is an R function name. The last two lines calculate $P(\overline{X} >$ `crit`$)$ under the null and the alternative. In both cases $\overline{X}_n$ is normal with standard deviation $\sigma/\sqrt{n}$, which is approximately $S_n/\sqrt{n}$. The difference between the two is the mean (oh, excuse me, I said I wasn't going to repeat this again, but here I go again), which is assumed to be $\mu = 700$ under the null and $\mu = 1000$ under the alternative.

If we were going to do a lot of these, we could clean up this calculation a bit and make a theorem out of it. It is clear from the way the calculation simplified at the end that some clean up is possible. But that wouldn't help us much. It would only apply to power calculation for tests involving means. Often one does power calculations for chi-square tests or F tests, which are much more complicated. We won't go into the details. We will let this subject go with these examples, which do illustrate the basic idea.

9.5.8 Multiple Tests and Confidence Intervals

All of Section 9.5 so far deals with the situation in which exactly one test is done on a data set. What if we want to do more than one test? Is the theory still valid?

No! If you take any complicated data set and keep doing different tests until one of them rejects the null hypothesis, this will eventually happen, but it proves absolutely nothing because this will always happen. If you keep going until you manage to think up a test that happens to reject H_0, then you will always eventually get a test to reject H_0.

What about situations between the ideal of just doing one test and doing a potentially infinite sequence of tests? What if you have several tests you want to do and will do no more even if none of them rejects H_0? Is there a valid way to do that?

Yes. In fact, many different ways have been proposed in the statistical literature.[5] Most only work with a specific kind of test. However, there is one procedure that is always applicable. It is the only one we will study.

Every known procedure for valid multiple testing conceptually combines the multiple tests into one big test. So you really do only one test. The null hypothesis for the combined test is that *all* the null hypotheses for the separate tests are true. The decision rule for the combined test is to reject the combined null hypothesis if any of the separate tests rejects its null hypothesis.

Suppose we have k tests with null hypotheses H_1, ..., H_k (we can't call them all H_0). The null hypothesis for the combined test is

$$H_0 = H_1 \text{ and } H_2 \text{ and} \cdots \text{and } H_k.$$

In terms of parameter sets, the logical "and" operation corresponds to set intersection. So if Θ_i is the set of parameter values corresponding to the null hypothesis H_i for the i-th separate test, then the parameter values corresponding to H_0 are

$$\Theta_0 = \Theta_1 \cap \Theta_2 \cap \cdots \cap \Theta_k.$$

The significance level is

$$P(\text{reject } H_0) = P(\text{reject } H_1 \text{ or reject } H_2 \text{ or } \ldots \text{ or reject } H_k)$$

assuming this does not depend on the parameter value (otherwise we would have to "sup" over Θ_0). Let E_i denote the event that the i-th test rejects its null hypothesis. If the i-th test is determined by a test statistic $T_i(\mathbf{X})$ and critical value c_i, then

$$E_i = \{\, \mathbf{X} : T_i(\mathbf{X}) \geq c_i \,\}$$

but the exact form of E_i doesn't matter, it is just the set of data values for which the i-th test rejects its null. Since the logical "or" operation corresponds to set union,

$$P(\text{reject } H_0) = P(E_1 \cup E_2 \cup \cdots \cup E_k). \tag{9.64}$$

But now we are stuck. In general we have no way to calculate (9.64). It is the correct significance level for the combined test. If we are to do the test properly, we must calculate it. But in general, especially when we have done a lot of tests with no simple pattern, there is no way to do this calculation.

The right hand side of (9.64) should be familiar. It appears in the addition rule for probability, which is (10) of Theorem 2 of Chapter 2 in Lindgren. But that rule has a condition, the E_i must be mutually exclusive, which never holds in multiple testing situations. So the addition rule is no help. However there is a rule with no conditions, *subadditivity of probability*

$$P(E_1 \cup E_2 \cup \cdots \cup E_k) \leq P(E_1) + P(E_2) + \cdots + P(E_k)$$

[5] There are whole books focused on this subject. A good one is *Simultaneous Statistical Inference* by Rupert G. Miller, Jr. (2nd ed., McGraw-Hill, 1981).

that holds for any events E_1, ..., E_k. This is (b) of Problem 2-22 in Lindgren. We can always apply this rule. Thus we get

$$\alpha = P(\text{reject } H_0) \le \sum_{i=1}^{k} P(\text{test } i \text{ rejects } H_i).$$

This at least provides an upper bound on the significance level. If we use the right hand side instead of α we at least get a conservative procedure. The true error rate, the probability of erroneously rejecting H_0 will be less that our upper bound.

If we adjust the separate tests so they all have the same individual significance level we get the following rules.

> *To do a combined test with significance level at most α, choose level α/k for the k separate tests.*

When we consider this in terms of P-values, the rule becomes

> *When you do k tests, multiply all P-values by k.*

This procedure is usually referred to as a *Bonferroni correction*, because a closely related inequality to subadditivity of probability is sometimes called Bonferroni's inequality.

Using the duality of tests and confidence intervals we immediately get the analogous procedure for multiple confidence intervals. There are two views we can take of multiple confidence intervals. If we have several confidence intervals, all with the same confidence level, for specificity say 95%, that does *not* mean there is 95% probability that they will all simultaneously cover. In fact if there are many intervals, there may be an extremely small probability of simultaneous coverage. Simultaneous confidence intervals is the dual concept of multiple tests. Bonferroni correction applied to confidence intervals says

> *To get simultaneous* $100(1-\alpha)\%$ *coverage for k confidence intervals choose confidence level* $100(1-\alpha/k)\%$ *for the separate intervals.*

Stargazing

Many scientific papers avoid P-values. They only indicate whether certain results are "statistically significant" or not at the 0.05 level and perhaps also at the 0.01 level. Such papers are full of tables like this one

1.13	−1.12	−1.30	1.16	−0.19
−1.18	0.12	0.02	−1.11	0.35
−0.49	−0.11	−0.45	−0.17	−1.66
2.70^{**}	0.03	0.14	−1.64	0.61
−0.35	1.80^{*}	2.65^{**}	−0.73	−1.32

* $P < 0.05$, ** $P < 0.01$

Although the asterisks are just footnote symbols, tables like this are so common that no one familiar with the literature needs to look at the footnote. One star means "significant" (statistically significant at the 0.05 level), and two stars means "highly significant" (statistically significant at the 0.01 level). The stars are supposed to indicate the interesting results. The unstarred numbers are garbage (uninteresting random noise).

Most such tables are completely bogus because *no correction was done for multiple testing.* If a Bonferroni correction (or some other correction for multiple testing) were done, there would be a lot fewer stars. And this would mean a lot fewer so-called significant results for scientists to woof about. Doing the tests honestly, with correction for multiple testing would take all the fun out of the game.

This practice has been disparagingly called "stargazing" by a sociologist (L. Guttman). It should have no place in real science. Yet it is widespread. In many scientific disciplines a paper is unusual if it *doesn't* have tables like this. Scientists being sheep, just like other people, they feel pressure to conform and use the stars. In many disciplines, tables like this are a form of what I call "honest cheating." The tests are bogus, but this is clearly admitted in the paper, so no one should be fooled. Actually, scientists never say anything so harsh as calling the procedure "bogus." That would offend their peers. The emit some academic weasel wording like "no correction was done for multiple testing." If you are statistically astute, you catch the implication of bogosity. Of course the naive reader completely misses the point, but the scientific literature isn't written to be readable by nonexperts.

To give the "honest cheaters" fair credit, they do have an argument for their failure to correct for multiple testing. If they did a Bonferroni correction, that would not be the exactly right thing to do, rather it would be the *conservative* thing to do. No correction is too liberal, Bonferroni is too conservative. The right thing would be somewhere in between, but we usually do not know how to do it. The "honest cheaters" admit they are making a mistake, but they assert that Bonferroni would *also* be a mistake (in the other direction). The trouble with this argument is that the right thing is a lot closer to Bonferroni correction than no correction. Doing many tests with no correction is always bogus.

Of course, tables full of stars are fine if Bonferroni or some other correction for multiple testing was done. Since this is so rare, all authors who do proper correction for multiple testing make it very clear that they did so. They don't want readers to assume they are as clueless and their results as meaningless as in the typical paper in their discipline.

Problems

9-1. The Laplace distribution defined in (9.1) does not have mean zero and variance one. Hence is not the *standard* Laplace distribution. What is the mean and variance of (9.1), and what would be the standard Laplace density (the one with mean zero and variance one)? If we use the standard Laplace density as the reference density of the Laplace location-scale family so that the parameters μ and σ would be the mean and standard deviation, what form would the densities have instead of (9.2)?

9-2. Show that the family of $\text{Gam}(\alpha, \lambda)$ distributions with α fixed and λ varying, taking on all values with $\lambda > 0$ is a scale family.

9-3. Suppose S_n^2 is the sample variance calculated from an i. i. d. normal random sample of size n.

(a) Calculate the bias of S_n as an estimator of the population variance σ.

(b) Find the constant k such that kS_n has the smallest mean square error as an estimator of σ.

9-4. Suppose U and V are statistics that are stochastically independent and are both unbiased estimators of a parameter θ. Write $\text{var}(U) = \sigma_U^2$ and $\text{var}(V) = \sigma_V^2$, and define another statistic $T = aU + (1 - a)V$ where a is an arbitrary but known constant.

(a) Show that T is an unbiased estimator of θ.

(b) Find the a that gives T the smallest mean square error.

9-5. The notes don't give any examples of estimators that are *not* consistent. Give an example of an inconsistent estimator of the population mean.

9-6. If $X \sim \text{Bin}(n, p)$, show that $\hat{p}_n = X/n$ is a consistent and asymptotically normal estimator of p, and give the asymptotic distribution of $\hat{p}_n$.

9-7. If X_1, X_2, $\ldots$ are i. i. d. from a distribution having a variance σ^2, show that both V_n and S_n^2 are consistent estimators of σ^2.

9-8. Suppose X_1, X_2, $\ldots$ are i. i. d. $\text{Geo}(p)$. Find a method of moments estimator for p.

9-9. Suppose X_1, X_2, $\ldots$ are i. i. d. $\text{Beta}(\alpha, 2)$.

(a) Find a method of moments estimator for α.

(b) Find the asymptotic normal distribution of your estimator.

9-10. Let X_1, X_2, $\ldots$, X_n be an i. i. d. sample from a $\text{Beta}(\theta, \theta)$ model, where θ is an unknown parameter. Find a method of moments estimator of θ.

9-11. Suppose X_1, X_2, $\ldots$ are i. i. d. $\text{Gam}(\alpha, \lambda)$. Find the asymptotic normal distribution of the method of moments estimator $\hat{\lambda}_n$ defined in (9.6b).

9-12. Calculate the ARE of $\overline{X}_n$ versus $\widetilde{X}_n$ as an estimator of the center of symmetry for

(a) The double exponential location-scale family having density given in Problem 7-6(b) of these notes. (Note that σ in in the formula for the densities given in that problem is *not* the standard deviation.)

(b) The $t(\nu)$ location-scale family, with densities given by

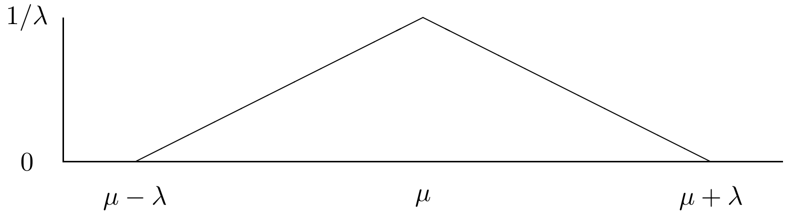
$$f_{\nu,\mu,\sigma}(x) = \frac{1}{\sigma} f_\nu\left(\frac{x-\mu}{\sigma}\right)$$

where f_ν is the $t(\nu)$ density given by (7.32). (Be careful to say things that make sense even considering that the $t(\nu)$ distribution does not have moments of all orders. Again σ is *not* the standard deviation.)

(c) The family of distributions called $\text{Tri}(\mu, \lambda)$ (for triangle) with densities

$$f_{\mu,\lambda}(x) = \frac{1}{\lambda}\left(1 - \frac{|x-\mu|}{\lambda}\right), \qquad |x-\mu| < \lambda$$

shown below



The parameter μ can be any real number, λ must be positive.

9-13. Let X_1, X_2, …, X_n be an i. i. d. sample from a $\mathcal{N}(\mu, \sigma^2)$ model, where μ and σ^2 are unknown parameters, and let S_n^2 denote the sample variance (defined as usual with $n-1$ in the denominator). Suppose $n = 5$ and $S_n^2 = 53.3$. Give an exact (not asymptotic) 95% confidence interval for σ^2.

9-14. In an experimental weight loss program five subjects were weighed before and after the 15 week treatment. The weights in pounds were as follows

	Subject				
	A	B	C	D	E
Before	225	216	215	225	186
After	193	206	171	223	156

If you want to use R on this problem, the data are in the file

```
http://www.stat.umn.edu/geyer/5102/prob9-14.dat
```

(a) Calculate a 95% confidence interval for the expected weight loss under the program.

(b) Describe the assumptions required to make this a valid confidence interval.

9-15. Suppose we have a sample with replacement of size n from a population and we are interested in the fraction p of the population having a certain property. For concreteness, say the property is that they intend to vote for Jones in an upcoming election. Let $\hat{p}_n$ denote the fraction of the sample having the property (intending to vote for Jones in the example).

(a) Show that

$$\frac{\hat{p}_n - p}{\sqrt{p(1-p)/n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1). \tag{9.65a}$$

(b) Show that $\hat{p}_n(1 - \hat{p}_n)$ is a consistent estimator of $p(1-p)$.

(c) Show that

$$\frac{\hat{p}_n - p}{\sqrt{\hat{p}_n(1-\hat{p}_n)/n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1). \tag{9.65b}$$

(d) Show that

$$\hat{p}_n \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}$$

is an asymptotic $100(1-\alpha)\%$ confidence interval for p, where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

(e) Find another asymptotic confidence interval for p based on the pivotal quantity in (9.65a) rather than the pivotal quantity in (9.65b).

9-16. Suppose we have two independent samples of size m and n from two different populations. We are interested in the fractions p and q of the populations that have a certain property (note: we are not using the $q = 1 - p$ convention here, p is the proportion of the first population having the property, and q is the proportion of the second population). We estimate these proportions by the sample proportions $\hat{p}_m$ and $\hat{q}_n$ which are the fractions of the first and second samples having the property. Show that

$$\hat{p}_m - \hat{q}_n \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_m(1-\hat{p}_m)}{m} + \frac{\hat{q}_n(1-\hat{q}_n)}{n}} \tag{9.66}$$

is an asymptotic $100(1-\alpha)\%$ confidence interval for $p - q$, where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

9-17. A physics lab is divided into 20 teams. Each team performs a measurement of the speed of light. Ten teams use one method and the other ten use another method. The average and standard deviation for the teams using each method was given in the following table (units are meters per second times 10^8).

	mean	standard deviation
Method 1	3.00013	0.00395
Method 2	2.99019	0.00853

If you want to use R on this problem, the data are in the file

```
http://www.stat.umn.edu/geyer/5102/prob9-17.dat
```

(a) Assuming that the measurements within each group of ten teams are independent and identically distributed around some unknown mean value (the speed of light as measured by that method), calculate a 95% confidence interval for the difference in the mean values for the two methods using Welch's approximation.

(b) Redo part (a) using the "pooled variance" t confidence interval that assumes both measurement methods have the same variance.

9-18. Suppose a sample of size 100 is assumed to be i. i. d. from a $\text{Gam}(\alpha, \lambda)$ model and the method of moments estimators of the parameters are $\hat{\alpha}_n = 5.23$ and $\hat{\lambda}_n = 21.3$. Find an asymptotic 95% confidence interval for α.

9-19. Suppose $V_{X,m}$ and $V_{Y,n}$ are sample variances and $M_{4,X,m}$ and $M_{4,Y,n}$ are the sample fourth central moments of independent samples from two populations having variances σ_X^2 and σ_Y^2, respectively. Find an asymptotic confidence interval for $\sigma_X^2 - \sigma_Y^2$.

9-20. Show that the "exact" confidence interval for the variance based on the chi-square distribution is asymptotically robust within the class of all distributions having fourth moments and satisfying $\mu_4 = 3\sigma^4$. That is, show that the assumption

$$\frac{nV_n}{\sigma_2} \sim \text{chi}^2(n-1)$$

implies a certain asymptotic limit for V_n and that this limit matches the *correct* asymptotic limit given by Theorem 7.17 only if $\mu_4 = 3\sigma^4$.

9-21. A *trimmed mean* is a point estimator of the form

$$\frac{X_{(k+1)} + \cdots + X_{(n-k)}}{n - 2k} \tag{9.67}$$

that is, the average of the data after the k lowest and k highest order statistics have been thrown away. If $0 \le \alpha < 1/2$ we say that (9.67) is a $100\alpha\%$ trimmed mean if $k = \lfloor n\alpha \rfloor$. We say that the median is a 50% trimmed mean.

For $0 < \alpha < 1/2$, find the breakdown point of the $100\alpha\%$ trimmed mean.

9-22. Given data $X_1, \ldots, X_n$, the *Walsh averages* consist of the n data items X_i and the $\binom{n}{2}$ averages $(X_i + X_j)/2$ for distinct pairs of indices i and j. The *Hodges-Lehmann estimator* of the center of symmetry of a symmetric distribution is the empirical median of the vector of Walsh averages.[6] Find the breakdown point of this estimator.

[6] Actually, there are lots of different Hodges-Lehmann estimators. This is the one associated with the Wilcoxon signed rank test.

9-23. Calculate the breakdown point of the MAD (median absolute deviation from the median) defined in Problem 7-1.

9-24. Assume $X_1, \ldots, X_n$ are i. i. d. $\mathcal{N}(\mu, \sigma^2)$. The observed value of $\overline{X}_n$ is 23.6, the observed value of S_n^2 is 103.2, and the sample size is $n = 20$. Perform a one-tailed test of the hypotheses $H_0 : \mu = 20$ versus $H_A : \mu > 20$, finding the P-value.

9-25. Two groups in physics lab have been measuring the density of aluminum at room temperature (20° C). They got the following summary statistics

	n	$\overline{X}_n$	S_n
Group I	10	2.792	0.241
Group II	8	2.538	0.313

(Units are grams per cubic centimeter.) Assume the measurements for group I are i. i. d. $\mathcal{N}(\mu_1, \sigma_1^2)$ and the measurements for group II are i. i. d. $\mathcal{N}(\mu_2, \sigma_2^2)$. We want to perform a test of $H_0 : \mu_1 = \mu_2$ versus $H_A : \mu_1 \neq \mu_2$. Perform Welch's approximate test, come as close as you can to the P-value.

If you want to use R on this problem, the data are in the file

```
http://www.stat.umn.edu/geyer/5102/prob9-25.dat
```

9-26. Suppose I have taken a random sample of size 100 of ears of corn from a field. My sample has mean ear length of 6.13 inches and standard deviation 1.44 inches. This gives me a 95% confidence interval for the true mean ear length all the corn in the field of 6.13 ± 0.28 inches.

Suppose I want a more accurate 95% confidence interval with a half-width (plus-or-minus) of 0.10 inches. What sample size do I need to get that?

9-27. Suppose I intend to collect data about the effect of coaching on SAT scores. The data will be SAT scores for individuals before and after taking a cram course. Suppose the test-retest variability without coaching is known to be about 50 points. How large a sample size do I need to have a power of 0.95 of detecting a true mean difference due to coaching as small as 10 points (the null hypothesis being no difference) at the 0.05 significance level? The test will be an upper-tailed test, since we expect that coaching cannot hurt.

9-28. For the data in Example 9.5.1 compute four confidence intervals, one for difference in each of the four rows of the table, so that your four intervals have 95% probability of *simultaneous* coverage.

Note: This problem can be done in R using the `prop.test` function, but getting the right confidence level is tricky. Be careful.

9-29. A problem on "stargazing." Suppose the twenty-five numbers in the table on p. 294 are all z-scores for different one-tailed, upper-tailed tests. The stars in the table do not reflect any correction for multiple testing. That is a z-score is declared "significant" (gets a star) if $z \geq 1.645$ and is declared "highly significant" (gets two stars) if $z \geq 2.326$. Here 1.645 and 2.326 are the one tailed 0.05 and 0.01 z critical values.

(a) What critical values should replace 1.645 and 2.326 in order to apply a Bonferroni correction to this multiple testing situation?

(b) What would the result of the Bonferroni correction be in terms of stars?

Chapter 10

Likelihood Inference

10.1 Likelihood

"Likelihood" is used as a technical term in statistics. It is not just a vague synonym for probability, as it is in everyday language. It is, however, closely related to probability.

Recall that we use *density* as a term that covers two of Lindgren's terms: p. f. (probability function) and p. d. f. (probability density function). In this chapter we will see one of the main reasons for our usage. The density will be used in exactly the same way, regardless of whether the data are discrete (so Lindgren would call it a p. f.) or continuous (so Lindgren would call it a p. d. f.). Also recall that a *statistical model* can be described by giving a parametric family of densities $\{ f_\theta : \theta \in \Theta \}$. This means that for each fixed parameter value θ in the parameter space Θ, there is a function $f_\theta(x)$ defined on the sample space that is nonnegative and sums or integrates to one, depending on whether the model is discrete or continuous.

A *likelihood* for the statistical model is defined by the same formula as the density, but the roles of x and θ are interchanged

$$L_x(\theta) = f_\theta(x). \tag{10.1}$$

Thus the likelihood is a different function of the parameter θ for each fixed value of the data x, whereas the density is a different function of x for each fixed value of θ. Likelihood is actually a slightly more general concept, we also call

$$L_x(\theta) = h(x) f_\theta(x) \tag{10.2}$$

a likelihood for the model when $h(x)$ is any nonzero function of x that does not contain the parameter θ. The reason for this extension of the notion is that all of the uses we make of the likelihood function will not be affected in any way by the presence or absence of $h(x)$. The way we make use of the extended definition is to simply drop terms in the density that do not contain the parameter.

Example 10.1.1 (Binomial Likelihood).
If $X \sim \text{Bin}(n, p)$, then

$$f_p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

This is also a likelihood $L_x(p)$. However, we are also free to drop the binomial coefficient, which does not contain the parameter p, writing

$$L_x(p) = p^x (1-p)^{n-x}. \tag{10.3}$$

When the data are an i. i. d. sample from a distribution with density $f_\theta(x)$, the joint density is

$$f_\theta(\mathbf{x}) = \prod_{i=1}^n f_\theta(x_i)$$

Hence this, thought of as a function of the parameter θ rather than the data $\mathbf{x}$, is also a likelihood. As usual we are allowed to drop multiplicative terms not containing the parameter.

When there are several parameters, the likelihood is a function of several variables (the parameters). Or, if we prefer, we can think of the likelihood as a function of a vector variable (the vector of parameters).

Example 10.1.2 (Normal Likelihood).
Suppose X_1, ..., X_n are i. i. d. $\mathcal{N}(\mu, \sigma^2)$, then the joint density of the data is

$$f_{\mu,\sigma}(\mathbf{x}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i-\mu)^2/2\sigma^2}$$

We can drop the $\sqrt{2\pi}$ terms. This gives

$$\begin{aligned}
L_{\mathbf{x}}(\mu, \sigma) &= \prod_{i=1}^n \frac{1}{\sigma} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \\
&= \frac{1}{\sigma^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i-\mu)^2\right) \qquad (10.4) \\
&= \frac{1}{\sigma^n} \exp\left(-\frac{nv_n + n(\bar{x}_n-\mu)^2}{2\sigma^2}\right)
\end{aligned}$$

where $\bar{x}_n$ is the empirical mean and v_n is the empirical variance, the last step using the empirical parallel axis theorem. Of course, we are free to use whichever form seems most convenient.

Example 10.1.3 (Normal Likelihood, Known Variance).
This is the same as the preceding example except now we assume σ^2 is a known constant, so μ is the only parameter. Now we are free to drop multiplicative terms not containing μ. Hence we can drop the σ^n term. We can also write

$$\exp\left(-\frac{nv_n + n(\bar{x}_n-\mu)^2}{2\sigma^2}\right) = \exp\left(-\frac{nv_n}{2\sigma^2}\right) \exp\left(-\frac{n(\bar{x}_n-\mu)^2}{2\sigma^2}\right)$$

and the first term on the right does not contain μ, hence

$$L_{\mathbf{x}}(\mu) = \exp\left(-\frac{n(\bar{x}_n - \mu)^2}{2\sigma^2}\right) \tag{10.5}$$

is a likelihood for this model. Note that this is a *different* statistical model from preceding problem because the parameter space is different. Hence it has a *different* likelihood function.

For the next several sections, we will only look at models with one parameter. We will return to multiparameter models in Section 10.4.

10.2 Maximum Likelihood

So far the only general method we have seen for constructing estimators is the method of moments. Now we are going to learn another method, even more general than the method of moments. It has a number of very desirable properties that will be developed as we go along. It is called the method of maximum likelihood. Roughly, the maximum likelihood estimator is the parameter value that maximizes the likelihood. For observed data $\mathbf{x}$ and likelihood $L_{\mathbf{x}}$ the *maximum likelihood estimator* (MLE) is defined to be the parameter value that maximizes the function $L_{\mathbf{x}}$ if the global maximum exists and is unique. If the global maximum does not exist or is not unique, then we have a problem defining the MLE. Mostly we will just deal with situations where there is a unique global maximizer. The MLE is denoted $\hat{\theta}(\mathbf{x})$ or sometimes just $\hat{\theta}$ when we want to leave the dependence on the data out of the notation. When we discuss asymptotics, we will often write it $\hat{\theta}_n(\mathbf{x})$ or $\hat{\theta}_n$ in order to indicate the dependence on the sample size n.

The *log likelihood* is the (natural) logarithm of the likelihood. It is denoted

$$l_{\mathbf{x}}(\theta) = \log L_{\mathbf{x}}(\theta).$$

We define $\log(0) = -\infty$. This makes sense because $\log(x) \to -\infty$ as $x \downarrow 0$.

Because the logarithm function is strictly increasing, a point maximizes the likelihood if and only if it maximizes the log likelihood. It is often simpler to maximize the log likelihood rather than the likelihood.

Example 10.2.1 (Binomial Model).
The log likelihood for the binomial distribution is found by taking logs in (10.3) giving

$$l_x(p) = x \log p + (n - x)\log(1 - p). \tag{10.6a}$$

Calculating derivatives at interior points p of the parameter space gives

$$\begin{aligned} l'_x(p) &= \frac{x}{p} - \frac{n-x}{1-p} \\ &= \frac{x - np}{p(1-p)} \qquad (10.6b) \\ l''_x(p) &= -\frac{x}{p^2} - \frac{n-x}{(1-p)^2}. \qquad (10.6c) \end{aligned}$$

The last equation shows that the log likelihood is a strictly concave function (Definition G.2.1 in Appendix G).

In the general case $0 < x < n$, the first derivative is zero at $\hat{p} = x/n$. By strict concavity, $\hat{p}$ is the unique global maximum. In the special cases $x = 0$ and $x = n$, there is no zero of the derivative, but the endpoints of the parameter space $(0 \le p \le 1)$, are local maxima. When $x = 0$

$$l'_x(p) = -\frac{n}{1-p}$$

so $l'_x(0) = -n$ which satisfies (G.2a). Similarly, when $x = n$ we have $l'_x(1) = n$ which satisfies the sufficient condition for $p = 1$ to be a local maximum. Thus in all three cases $\hat{p} = x/n$ is a local maximum of the log likelihood, hence the unique global maximum by strict concavity.

In this case, the MLE is the obvious estimator. By definition X is the sum of i. i. d. Bernoulli random variables, and $\hat{p}$ is the sample mean of these variables. It is also a method of moments estimator and an unbiased estimator, since $E(\hat{p}) = p$.

Example 10.2.2 (Normal Model, Known Variance).
The likelihood for this model is given by (10.5), hence the log likelihood is

$$l_n(\mu) = -\frac{n(\bar{x}_n - \mu)^2}{2\sigma^2} \qquad (10.7)$$

which is clearly maximized at

$$\hat{\mu}_n = \bar{x}_n$$

(since the log likelihood is zero there and negative elsewhere), so that is the MLE.

Example 10.2.3 (Cauchy Location Model).
The Cauchy location model has densities

$$f_\theta(x) = \frac{1}{\pi} \cdot \frac{1}{1 + (x - \theta)^2}$$

Hence the log likelihood for an i. i. d. sample of size n is

$$l_n(\theta) = -\sum_{i=1}^{n} \log\left(1 + (x_i - \theta)^2\right)$$

(we can drop the constant terms $1/\pi$). We can differentiate this

$$l'_n(\theta) = \sum_{i=1}^{n} \frac{2(x_i - \theta)}{1 + (x_i - \theta)^2} \tag{10.8}$$

but the result is horrible. We won't get anywhere by setting that equal to zero and trying to solve for θ.

Fortunately, computers can help. They can't give us a formula expressing the MLE as a function of θ. There isn't any such formula in terms of well-known elementary functions. But for any particular data set, the computer can maximize the likelihood and find the MLE. That's good enough in most applications. R, for example, has a function `nlm` that does nonlinear minimization of a function of several variables. We can use that to find the MLE (minimizing $-f$ maximizes f). First we make up some data

```
Rweb:> n <- 40
Rweb:> theta0 <- 0
Rweb:> x <- theta0 + rcauchy(n)  # make up data
Rweb:> summary(x)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
 -5.9590  -0.4615   0.1203  10.0100   1.1990 172.0000
```

In a real problem, of course, we would read in data obtained from the data collectors (though, to be honest, the Cauchy has such heavy tails that it's not used for real data). We have also run the `summary` command that gives the estimators we already know about, the sample mean and median, as well as some other interesting statistics. We know the sample mean is *not* a consistent estimator of θ because the expectation of the Cauchy distribution does not exist. We know the sample median is consistent and asymptotically normal [Problem 7-6(a)].

Then the following code finds the MLE. First it defines an R function `l` that evaluates minus the log likelihood. Then it hands that function to `nlm` to minimize. The `nlm` function uses an iterative algorithm and needs a starting point, which we supply as `median(x)`, the best estimator we know that we have a simple expression for (ignore the third argument to `nlm`, it's helpful but not necessary).

```
Rweb:> l <- function(theta) sum(log(1 + (x - theta)^2))
Rweb:> out <- nlm(l, median(x), fscale=n)
Rweb:> out$estimate
[1] 0.00276767
```

The result is the MLE. Notice that it is a lot closer to the true parameter value (which we know to be zero because we made up the data) than the median. This is no accident. We will eventually see that the MLE is a much better estimator here than the sample median.

10.3 Sampling Theory

10.3.1 Derivatives of the Log Likelihood

When we write $l_{\mathbf{X}}(\theta)$ rather than $l_{\mathbf{x}}(\theta)$, considering the subscript a random vector "big $\mathbf{X}$" rather than a fixed value "little $\mathbf{x}$" the log likelihood becomes a random function, and everything about it, including its derivatives, is also random. Specifically, $l_{\mathbf{X}}$ and its derivatives, $l'_{\mathbf{X}}$, $l''_{\mathbf{X}}$, and so forth, are random functions. The values of these functions at a specified point θ, that is, $l_{\mathbf{X}}(\theta)$, $l'_{\mathbf{X}}(\theta)$, $l''_{\mathbf{X}}(\theta)$, and so forth, are random variables. Note that each different value of θ gives a *different* random variable. Like every other random variable, they have probability distributions. As usual, when we are talking about randomness arising from or mimicking random sampling, we call these the sampling distributions of the random variables, in this case, of the log likelihood and its derivatives.

We are now going to change notation, suppressing the dependence of the log likelihood on the data $\mathbf{X}$ and emphasizing the dependence on the sample size n. As always, this is useful when we discuss asymptotics, and all of the distribution theory in likelihood inference is asymptotic theory. Thus we will write l_n, l'_n, and so forth rather than $l_{\mathbf{X}}$, $l'_{\mathbf{X}}$, and so forth.

The Score

The first derivative of the log likelihood

$$l'_n(\theta) = \frac{d}{d\theta} l_n(\theta)$$

is often called the *score function* or just the *score*. When we consider the score function to be random, $l'_n(\theta)$ is a random variable, a different random variable for each different value of θ.

The score function is important in maximum likelihood. We usually find the MLE by solving the equation $l'_n(\theta) = 0$. Of course, this doesn't work when the MLE is on the boundary of the parameter space or when the MLE doesn't exist, but it does work in the usual case and we have $l'_n(\hat{\theta}_n) = 0$. Note that this does *not* imply that $l'_n(\theta) = 0$ when θ is the true parameter value. Just the opposite! "The sample is not the population" implies that $\hat{\theta}_n$ is *not* θ. In fact, $l'_n(\theta)$ is a random variable and hence doesn't have any constant value.

Expected Fisher Information

The *Fisher information* for a statistical model is the variance of the score $l'_n(\theta)$

$$I_n(\theta) = \operatorname{var}_\theta\{l'_n(\theta)\}. \tag{10.9}$$

It is named after R. A. Fisher, who invented maximum likelihood and discovered many of the properties of maximum likelihood estimators and first called this concept "information." Lindgren calls this concept just "information" instead of "Fisher information," but the latter is standard terminology because more than one notion of "information" has been used in statistics (although Fisher information is by far the most important and the only one we will consider in this course).

Lindgren uses the notation $I_X(\theta)$ rather than $I_n(\theta)$ but admits this "could be misleading" because the Fisher information does *not* depend on the data X but rather on the *model*, that is on *which* variables we consider "data" rather than on the *values* of those variables. Note that the Fisher information is *not* a random quantity (because *no* unconditional expectation is a random quantity), another reason why the notation $I_X(\theta)$ is very misleading. Since Lindgren's notation is misleading, we will not use it.

Differentiating Under the Integral Sign

Any probability density satisfies

$$\int f_\theta(x)\,dx = 1 \tag{10.10}$$

(or the analogous equation with summation replacing integration if the data are discrete). Usually, although not always,[1] it is possible to take derivatives inside the integral sign, that is,

$$\frac{\partial^k}{\partial\theta^k}\int f_\theta(x)\,dx = \int \frac{\partial^k}{\partial\theta^k} f_\theta(x)\,dx. \tag{10.11}$$

Looking back at the right hand side of (10.10), we see that because the derivative of a constant is zero, that all of the derivatives in (10.11) are zero, that is,

$$\int \frac{\partial^k}{\partial\theta^k} f_\theta(x)\,dx = 0 \tag{10.12}$$

provided that differentiation under the integral sign is valid, that is, provided (10.11) holds.

The partial derivative notation will become unwieldy in the following proof, so we are going to introduce the following shorthand for (10.12)

$$\int f' = 0$$

and

$$\int f'' = 0$$

using primes to indicate partial derivatives with respect to θ (in likelihood theory we always differentiate with respect to parameters, never with respect to data)

[1] We will not worry about the precise technical conditions under which this operation is permitted. They can be found in advanced calculus books. The only condition we will mention is that the limits of integration in (10.11) must not contain the variable of differentiation θ. This will hold in all of the examples we consider.

and also suppressing the variable x entirely (although the integration is still with respect to the data x).

Now we write the log likelihood $l = \log f$, and using the chain rule we obtain

$$l' = \frac{f'}{f} \tag{10.13a}$$

$$l'' = \frac{f''}{f} - \left(\frac{f'}{f}\right)^2 \tag{10.13b}$$

Now note that for any random variable $g(X)$

$$E\{g(X)\} = \int g(x) f(x) \, dx$$

or in the shorthand we are using in this section $E(g) = \int gf$. What this says is that in order to change an expectation to an integral we need an extra f in the integrand. Thus taking expectations of (10.13a) and (10.13b) gives

$$\begin{aligned} E(l') &= \int f' \\ E(l'') &= \int f'' - E\left\{\left(\frac{f'}{f}\right)^2\right\} \end{aligned}$$

and we know from our previous discussion that (still assuming differentiability under the integral sign is valid) that the integrals here are zero, thus

$$E(l') = 0 \tag{10.13c}$$

$$E(l'') = -E\left\{\left(\frac{f'}{f}\right)^2\right\} \tag{10.13d}$$

Finally we note that we can use (10.13a) and (10.13c) to simplify (10.13d). $l' = f'/f$ is a random variable. It has mean zero by (10.13c). For any random variable having mean zero ordinary and central moments are the same, hence the variance is also the ordinary second moment. Thus the second term on the right hand side of (10.13d) is $\operatorname{var}(l')$. Thus we can rewrite (10.13d) as

$$E(l'') = -\operatorname{var}(l') \tag{10.13e}$$

Now we want to get rid of the shorthand, and restate our conclusions as a theorem using ordinary mathematical notation.

Theorem 10.1. *Provided* (10.10) *can be differentiated twice with respect to θ under the integral sign, that is* (10.12) *holds for $k = 1$ and $k = 2$,*

$$E_\theta\{l_n'(\theta)\} = 0 \tag{10.14a}$$

and

$$E_\theta\{l_n''(\theta)\} = -\operatorname{var}_\theta\{l_n'(\theta)\} \tag{10.14b}$$

for all values of θ for which the differentiation under the integral sign is permitted.

This is just what we proved in the preceding discussion.

Note that the variance on the right hand side of (10.14b) is Fisher information. This says that we can calculate Fisher information in two different ways, either the variance of the first derivative of the log likelihood or minus the expectation of the second derivative. You may use whichever seems simpler, your choice. First derivatives are sometimes simpler than second derivatives and sometimes not. Expectations are usually simpler than variances. My experience is that a majority of problems the second derivative calculation is simpler. But in a sizable minority of problems the first derivative calculation is simpler. Don't entirely ignore the first derivative method.

Example 10.3.1 (Binomial Model).
In Example 10.2.1 we found the second derivative of the log likelihood to be

$$l''_X(p) = -\frac{X}{p^2} - \frac{n - X}{(1 - p)^2}.$$

This is (10.6c) with "little x" changed to "big X" because we are now considering it a random quantity. Taking expectations using $E(X) = np$ gives

$$\begin{aligned} I_n(p) &= -E\{l''_X(p)\} \\ &= \frac{np}{p^2} + \frac{n - np}{(1 - p)^2} \\ &= \frac{n}{p(1 - p)} \end{aligned}$$

Example 10.3.2 (Normal Model, Known Variance).
In Example 10.2.2 we found the log likelihood (10.7) for this model Differentiating, we find

$$l''_n(\mu) = -\frac{n}{\sigma^2}$$

Since this happens not to depend on the data, it is nonrandom, hence is its own expectation. Thus it is minus the Fisher information, that is

$$I_n(\mu) = \frac{n}{\sigma^2}$$

The subscripts θ on the expectation and variance operators in (10.14a) and (10.14b) are important. If omitted, it would be possible to give these equations a reading that is false. The point is that there are two θ's involved. When they are different, the statement is simply false.

$$E_{\theta_1}\{l'_n(\theta_2)\} \neq 0$$

when $\theta_1 \neq \theta_2$. If (10.14a) is written with no subscript on the expectation operator

$$E\{l'_n(\theta)\} = 0,$$

then it is not clear what parameter value is meant and under the wrong assumption about what parameter is meant the equation is simply false.

The CLT for the Score

The log likelihood and its derivatives for an i. i. d. sample are sums of i. i. d. terms.

$$l_n(\theta) = \sum_{i=1}^n \log f_\theta(X_i) \tag{10.15a}$$

$$l_n'(\theta) = \sum_{i=1}^n \frac{\partial}{\partial\theta} \log f_\theta(X_i) \tag{10.15b}$$

$$l_n''(\theta) = \sum_{i=1}^n \frac{\partial^2}{\partial\theta^2} \log f_\theta(X_i) \tag{10.15c}$$

and so forth. The middle equation (10.15b) is the score. Also note that each term on the right hand side of (10.15a) is the log likelihood for a model having only X_i as data, that is, a log likelihood for sample size one.

Theorem 10.2.

$$I_n(\theta) = nI_1(\theta)$$

Proof. Because the X_i are assumed independent, the terms on the right hand side of (10.15b) are independent. Hence the variance of the sum is the sum of the variances. By the preceding comment, each term on the right hand side is a score for a sample of size one. □

In words, the theorem says the Fisher information for a sample of size n is equal to the Fisher information for a sample of size 1 multiplied by n.

Example 10.3.3 (Cauchy Location Model).
In Example 10.2.3 we found the first derivative of the log likelihood (the score) for this model in (10.8). The second derivative is

$$l_n''(\theta) = \sum_{i=1}^n \left(-\frac{2}{1 + (x_i - \theta)^2} + \frac{4(x_i - \theta)^2}{[1 + (x_i - \theta)^2]^2} \right) \tag{10.16}$$

We called the first derivative "horrible." This is even more of a mess, but we are only trying to integrate this, and there is a nice analytical (though fairly messy) indefinite integral. Note by Theorem 10.2 that to find the Fisher information, we only need to do the integral for sample size one. Mathematica has no trouble

```
In[1]:= <<Statistics‘ContinuousDistributions‘

In[2]:= dist = CauchyDistribution[theta, 1]

Out[2]= CauchyDistribution[theta, 1]

In[3]:= f[x_, theta_] = PDF[dist, x]
```

```
                 1
Out[3]= ----------------------
                             2
        Pi (1 + (-theta + x) )

In[4]:= Integrate[ D[ Log[f[x, theta]], {theta, 2} ] f[x, theta],
        {x, -Infinity, Infinity} ]

          1
Out[4]= -(-)
          2
```

We don't even need to do the differentiation ourselves. Let Mathematica do both differentiation and integration. The Fisher information is minus this

$$I_1(\theta) = \frac{1}{2} \tag{10.17}$$

And of course, by Theorem 10.2 the Fisher information for sample size n is just n times this.

In the proof of Theorem 10.2 we established that the right hand side of (10.15b) is the sum of i. i. d. terms, each of which is the score for a sample of size one and hence has mean zero by (10.14a) and variance $I_1(\theta)$ by (10.9), the definition of Fisher information.

Thus $l_n'(\theta)/n$ is the average of i. i. d. terms and the central limit theorem applies. Being precise, it says

$$\frac{1}{\sqrt{n}} l_n'(\theta) \xrightarrow{\mathcal{D}} \mathcal{N}\bigl(0, I_1(\theta)\bigr). \tag{10.18}$$

The $1/\sqrt{n}$ arises because we divide $l_n'(\theta)$ by n to get an average then we multiply by $\sqrt{n}$ as usual in the CLT. There is no mean subtracted off on the left hand side because the score has mean zero. The sloppy "double squiggle" version says

$$l_n'(\theta) \approx \mathcal{N}\bigl(0, I_n(\theta)\bigr).$$

Here we wrote $I_n(\theta)$ rather than $nI_1(\theta)$ for the variance (they are, of course, equivalent by Theorem 10.2). Note that the asymptotic mean and variance are no surprise, since by (10.14a) and the definition of Fisher information (10.9) these are the *exact* mean and variance of $l_n'(\theta)$. The only surprise (and it should be no surprise by now) is that the large sample distribution is normal (by the CLT).

Observed Fisher Information

The *observed Fisher information* is

$$J_n(\theta) = -l_n''(\theta). \tag{10.19}$$

For contrast $I_n(\theta)$ is sometimes called *expected* Fisher information to distinguish it from $J_n(\theta)$, although, strictly speaking, the "expected" is redundant.

Note that $J_n(\theta)$ is a random quantity, even though the notation does not explicitly indicate this. In contrast, expected Fisher information, like any other expected value is constant (nonrandom). The connection between observed and expected Fisher information is given by (10.14b), which says, using the notation (10.19) for observed Fisher information

$$E\{J_n(\theta)\} = I_n(\theta). \tag{10.20}$$

Example 10.3.4 (Cauchy Location Model).
In Example 10.3.3 we found the second derivative of the log likelihood for this model in (10.16). The observed Fisher information is just minus this.

$$J_n(\theta) = \sum_{i=1}^{n} \left(\frac{2}{1 + (x_i - \theta)^2} - \frac{4(x_i - \theta)^2}{[1 + (x_i - \theta)^2]^2} \right) \tag{10.21}$$

The LLN for Observed Fisher Information

Equation 10.20 gives us the expectation of the observed Fisher information. Generally, we do not know anything about its variance or any higher moments. Not knowing the variance, the CLT is of no use. But the LLN is still informative.

The analysis is just like the analysis of the sampling distribution of $l'_n(\theta)$ two sections back (but simpler because the LLN is simpler than the CLT). The right hand side of (10.15c) is the sum of i. i. d. terms, each of which is the second derivative of the log likelihood for a sample of size one and hence has mean $I_1(\theta)$ by (10.20).

Thus $J_n(\theta)/n$ is the average of i. i. d. terms and the law of large numbers applies. Being precise, it says

$$\frac{1}{n} J_n(\theta) \xrightarrow{P} I_1(\theta). \tag{10.22}$$

The sloppy "double squiggle" version would be

$$J_n(\theta) \approx I_n(\theta)$$

Note that this doesn't describe an *asymptotic distribution* for $J_n(\theta)$ because the right hand side is *constant* (as always the LLN gives less information than the CLT).

10.3.2 The Sampling Distribution of the MLE

If we expand l'_n using a Taylor series with remainder about the true parameter value θ_0, we get

$$l'_n(\theta) = l'_n(\theta_0) + l''_n(\theta_0)(\theta - \theta_0) + \tfrac{1}{2} l'''_n(\theta^*)(\theta - \theta_0)^2,$$

where θ^* in the remainder term is some point between θ and θ_0.

Using $l''_n(\theta) = -J_n(\theta)$, we get

$$l'_n(\theta) = l'_n(\theta_0) - J_n(\theta_0)(\theta - \theta_0) + \tfrac{1}{2} l'''_n(\theta^*)(\theta - \theta_0)^2. \tag{10.23}$$

Now we assume that the MLE is in the interior of the parameter space, so it satisfies the "likelihood equation" $l'_n(\hat{\theta}_n) = 0$. Then if we plug in $\hat{\theta}_n$ for θ in (10.23), the left hand side is zero, and we get

$$0 = l'_n(\theta_0) - J_n(\theta_0)(\hat{\theta}_n - \theta_0) + \tfrac{1}{2} l'''_n(\theta^*_n)(\hat{\theta}_n - \theta_0)^2, \tag{10.24}$$

where now θ^*_n is some point between θ_0 and $\hat{\theta}_n$.

Now we want to multiply (10.24) by the appropriate constant so that the various terms converge to a nontrivial distribution. Looking at the CLT for $l'_n(\theta)$, equation (10.18) we see that the right constant is $1/\sqrt{n}$ ("constant" here means nonrandom, this is, of course, a function of n). That gives

$$0 = \frac{1}{\sqrt{n}} l'_n(\theta_0) - \frac{1}{n} J_n(\theta_0) \cdot \sqrt{n}(\hat{\theta}_n - \theta_0) + \frac{1}{2\sqrt{n}} l'''_n(\theta^*_n)(\hat{\theta}_n - \theta_0)^2. \tag{10.25}$$

In the middle term we wrote $1/\sqrt{n} = \sqrt{n}/n$ and put each piece with a different factor. We know the behavior of $J_n(\theta)/n$. It's given by (10.22). And we expect from our general experience with asymptotics so far that $\sqrt{n}(\hat{\theta}_n - \theta_0)$ will have a nontrivial asymptotic distribution.

The last term in (10.25) is a mess unlike anything we have ever seen. In order to make progress, we need to make an assumption that gets rid of that messy term. The appropriate assumption is the following.

$$\frac{1}{n} l'''_n(\theta^*_n)(\hat{\theta}_n - \theta) \xrightarrow{P} 0 \tag{10.26}$$

Then rearranging (10.25) gives

$$\sqrt{n}(\hat{\theta}_n - \theta)\left[1 - \frac{\frac{1}{2n} l'''_n(\theta^*_n)(\hat{\theta}_n - \theta)}{\frac{1}{n} J_n(\theta_0)}\right] = \frac{\frac{1}{\sqrt{n}} l'_n(\theta_0)}{\frac{1}{n} J_n(\theta_0)}.$$

Combining our assumption (10.26) with (10.22) and Slutsky's theorem, the messy second term in the square brackets converges in probability to zero (leaving only the unit term). Thus by another use of Slutsky's theorem $\sqrt{n}(\hat{\theta}_n - \theta)$ has the same asymptotic behavior as the right hand side, that is

$$\sqrt{n}\big(\hat{\theta}_n - \theta_0\big) \approx \frac{\frac{1}{\sqrt{n}} l'_n(\theta_0)}{\frac{1}{n} J_n(\theta_0)}.$$

Using (10.18) and (10.22) and Slutsky's theorem yet again, the right hand side converges in distribution to $Z/I_1(\theta)$ where $Z \sim \mathcal{N}\big(0, I_1(\theta)\big)$. Since a linear transformation of a normal is normal, $Z/I_1(\theta)$ is normal with mean

$$E\left\{\frac{Z}{I_1(\theta)}\right\} = \frac{E(Z)}{I_1(\theta)} = 0$$

and variance

$$\operatorname{var}\left\{\frac{Z}{I_1(\theta)}\right\} = \frac{\operatorname{var}(Z)}{I_1(\theta)^2} = \frac{1}{I_1(\theta)}$$

Thus

$$\sqrt{n}\big(\hat{\theta}_n - \theta_0\big) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, I_1(\theta_0)^{-1}\big) \tag{10.27}$$

This completes a proof of the asymptotic normality of the MLE.

Theorem 10.3. *Suppose the true parameter value θ_0 is in the interior of the parameter space, and suppose the assumptions about differentiability under the integral sign in Theorem 10.1 hold. Suppose we have i. i. d. sampling. And finally suppose that assumption* (10.26) *holds. Then* (10.27) *holds.*

Example 10.3.5 (Cauchy Location Model).
In Example 10.3.3 we found the expected Fisher information for this model (10.17). Inserting that into (10.27) we get

$$\sqrt{n}\big(\hat{\theta}_n - \theta_0\big) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, 2\big)$$

I hope you are suitably impressed at the magic of theoretical statistics here. The other examples we have done in this chapter don't need to use the theory. When the MLE turns out to be $\overline{X}_n$, we already know its asymptotic distribution. In fact, whenever the MLE turns out to be a simple function of any sample moments we could use the delta method to find its asymptotic distribution (though calculating Fisher information is usually easier than applying the delta method). Here we do not have an analytic expression for the MLE as a function of the data. Thus we cannot use the delta method or any other method we have covered (or for that matter any method we *haven't* covered). Fisher information gives us the asymptotic distribution of a random variable we can't even describe (except for the implicit description that it maximizes the likelihood).

Real-life applications of the method of maximum likelihood are more like this Cauchy example than any of the other examples or homework problems. For complicated data (and it seems that real scientific data sets get ever more complicated every year) often the only thing you can write down for the model is the likelihood function. You can't calculate anything else analytically. However, the computer can calculate the MLE and the observed Fisher information (See Example 10.3.6 for more on this), and you're in business. Nothing else from theoretical statistics works, just likelihood theory.

The difficulty with applying Theorem 10.3 is that it is rather hard to verify the conditions. Why should (10.26) hold? The truth is that for some models it does, and for some models it doesn't. The assumption is not as weird as it looks. If the MLE is consistent, then $\hat{\theta}_n - \theta \xrightarrow{P} 0$. Also $\frac{1}{n} l_n'''(\theta_0)$ converges in probability to some constant (its expectation) by the law of large numbers (assuming the expectation exists), because it too is the sum of i. i. d. terms. Then by Slutsky's theorem $\frac{1}{n} l_n'''(\theta_0)(\hat{\theta}_n - \theta)$ converges in probability to zero. Our assumption (10.26) differs only in having θ_n^* in place of θ_0 as the argument of l_n'''. If $\hat{\theta}_n$ converges to θ_0, then so does θ_n^*. Hence we might expect $\frac{1}{n} l_n'''(\theta_n^*)(\hat{\theta}_n - \theta)$

to also converge in probability to zero, which would imply (10.26). Thus the assumption is plausible. But actually showing it holds for any particular model is difficult mathematics, beyond the scope of this course.

What we are left with a rather annoying situation. The "standard asymptotics" of the MLE given by (10.27) usually holds. It holds for "nice" models. But we don't have any definition of "nice" that we can understand intuitively. In fact a half century of research by a lot of smart people has not found any simple definition of "nice" that will do the job here. So though the usual asymptotics usually hold, they don't always, so we are always beset with vague anxiety when using this stuff, or at least we *should* be anxious in order to be proper statisticians. The Alfred E. Newman philosophy "What, me worry?" just isn't appropriate, though to be honest, it does about as much good as the official philosophy that you can't use (10.27) until you have somehow verified the conditions of the theorem (using a lot of math far beyond the scope of this course) or had someone else (a good theoretical statistician) do it for you. Very few users of applied statistics actually do that. Recalling our slogan that asymptotics only produce heuristics and that if you are worried you simulate, one can see why. Even if you managed to verify the conditions of the theorem it still wouldn't tell you how large n would have to be to use the theorem on real data. You would still be in the position of having to simulate if worried.

10.3.3 Asymptotic Relative Efficiency

The MLE usually has the best possible asymptotic variance of any estimator, but a precise statement of this result is tricky, requiring a new concept. We say an estimator $\hat{\theta}_n$, whether or not it is the MLE, is *asymptotically efficient* if it satisfies (10.27). If it does better than that, if its asymptotic variance is less than $I_1(\theta_0)^{-1}$, we say it is *superefficient*. Since the true parameter value θ_0 is unknown, we also insist that it be efficient (at least) at every θ.

Using this new terminology, proving the MLE to be the best possible estimator is the same thing as proving that superefficient estimators do not exist. The trouble with that is that they *do* exist, although they are quite crazy. Here is an example of a superefficient estimator.

Suppose X_1, X_2, are i. i. d. normal with known variance σ^2. In Example 10.2.2 we found that the MLE of the mean μ is $\overline{X}_n$. Consider the estimator

$$X_n^* = \begin{cases} \overline{X}_n, & |\overline{X}_n| > n^{-1/4} \\ 0, & \text{otherwise} \end{cases}$$

If the true μ is not exactly zero, then $\overline{X}_n \xrightarrow{P} \mu$ by the LLN, and $P(\overline{X}_n \le n^{-1/4})$ converges to zero. Thus by Slutsky's theorem X_n^* and $\overline{X}_n$ have the same asymptotics.

But if the true μ is exactly zero, then the CLT says

$$\overline{X}_n \approx \frac{n^{-1/2} Z}{\sigma}$$

where Z is standard normal. Thus $P(\overline{X}_n \le n^{-1/4})$ converges to one because $n^{-1/2}$ is much smaller than $n^{-1/4}$ for large n. Thus in this case $P(X_n^* = 0)$ converges in probability to one. Thus we have two cases, our estimator obeys the usual asymptotics at almost all points

$$\sqrt{n}\,(X_n^* - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$$

but at just one point $\mu = 0$ it has superefficient asymptotics

$$\sqrt{n}\,(X_n^* - \mu) \xrightarrow{\mathcal{D}} 0.$$

Please don't complain about this example. It may seem to be a stupid theoretician trick, but that's the point. All superefficient estimators are stupid in this fashion.

Suppose we have an estimator θ_n^* of some parameter θ that is consistent and asymptotically normal, that is,

$$\sqrt{n}\,(\theta_n^* - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}\bigl(0, \tau^2(\theta)\bigr),$$

where we have written the asymptotic variance as $\tau^2(\theta)$ to indicate that, in general, it is a function of the true parameter value θ. If $\tau^2(\theta)$ is a continuous function of θ, the estimator cannot be superefficient. We must have

$$\tau^2(\theta) \ge I_1(\theta)^{-1}, \qquad \text{for all } \theta$$

Superefficiency can only occur discontinuously, as in the example. Thus superefficient estimators are a pathological phenomenon. They are no use in real applications, because you can never know whether such an estimator is superefficient at the true parameter value.

Ignoring superefficient estimators, the theorem says that it is not possible asymptotically to do better than the MLE. Of course you can do better for finite sample size n, but at least if n is moderately large you can't do much better. This fact creates a strong argument for using maximum likelihood estimators.

Example 10.3.6 (Cauchy Location Model).
In Example 10.3.5 we asymptotic distribution for the MLE in the Cauchy Location Model. The only other sensible estimator we know about is the sample median, whose asymptotic distribution was found in Problem 7-6(a). The asymptotic variance of the MLE is 2. The asymptotic variance of the sample median is $\pi^2/4$. The ARE is 0.81 or 1.23 depending on which way you write it. The MLE is more efficient (the MLE is always more efficient).

Example 10.3.7 (Normal Location Model).
In Example 10.2.2 we found that the sample mean is the MLE for the normal location model (normal, known variance). In Example 7.4.1 we found that the sample median was asymptotically less efficient than the sample mean for this model (asymptotic variance σ^2 for the mean and $\pi\sigma^2/2$ for the median). Now we find out why. The sample mean, being the MLE is better than *any* other estimator (barring weird superefficient estimators).

Example 10.3.8 (Laplace Location Model).
The Laplace (double exponential) location model has densities

$$f_\mu(x) = \frac{1}{2} e^{-|x-\mu|}$$

hence log likelihood

$$l_n(\mu) = -\sum_{i=1}^n |x_i - \mu| \tag{10.28}$$

Now this is a problem for which the tools we usually apply are no help. We can't take derivatives, because the absolute value function is nondifferentiable (having a kink at zero). However, we can use Corollary 7.7 (the characterization of the empirical median), which says, phrased in terms of the current context, that the maximizer of (10.28) is the sample median. Thus the sample median is the MLE. In Problem 7-6(b) we found the asymptotic variance of the sample median (now discovered also to be the MLE) to be one (in this parameterization). In problem 9-1 we found the variance of X to be two (in this parameterization). Hence the ARE of the mean to the median is either 1/2 or 2, depending on how you write it. Now we find out that it is no surprise the sample median is better. It's the MLE so it's better that any other estimator (not just better than the mean).

As an aside, note that we can't use Fisher information to calculate the asymptotic variance because it isn't defined, the log likelihood not being differentiable. The theorem that the MLE is more efficient, still holds though. So when we find out that the sample median is the MLE, we can use the theorem about the asymptotics of the sample median (Corollary 7.28) to calculate the asymptotic variance.

We summarize the results of the preceding three examples in a little table of asymptotic variances.

	MLE	median	mean
Cauchy	2	2.47	∞
Normal	1	1.57	1
Laplace	1	1	2

In the first line, all three estimators are different. The sample mean is useless, not even consistent (the LLN doesn't hold because the expectation of the Cauchy distribution doesn't exist). In the other two lines, the MLE is the same as one of the other estimators. In all three cases the MLE is best (by the theorem, the MLE is best in every case, not just these).

10.3.4 Estimating the Variance

One remaining problem with Theorem 10.3 is that the asymptotic variance $I_1(\theta_0)^{-1}$ is unknown because the true parameter value θ_0 is unknown. (If we knew θ_0, we wouldn't be bothered with estimating it!) But this is a minor problem. We just estimate the asymptotic variance using the "plug-in" principle by $I_1(\hat{\theta}_n)$. If I_1 is a continuous function of θ, then

$$I_1(\hat{\theta}_n) \xrightarrow{P} I_1(\theta_0), \qquad \text{as } n \to \infty$$

by the continuous mapping theorem. So we can use the left hand side as an approximation to the right hand side. Being a bit sloppy, we can write

$$\hat{\theta}_n \approx \mathcal{N}\left(\theta_0, \frac{1}{nI_1(\hat{\theta}_n)}\right).$$

Caution: There is a natural tendency, exhibited by many students, to get confused about whether one uses $I_1(\theta)$ or $I_n(\theta)$. They get either too many or too few n's or root n's in their standard errors. The error variance is

$$\frac{1}{nI_1(\hat{\theta}_n)} = \frac{1}{I_n(\hat{\theta}_n)}$$

In words, this can be called "inverse Fisher information" *if* (big if) one remembers which Fisher information is meant. It is the inverse of the Fisher information for the *actual problem at hand* (sample size n). Another way to remember which is the correct standard error is that the standard error must obey the "square root law," that is, it must decrease like $1/\sqrt{n}$. If one gets confused about the standard error, one gets ridiculous confidence intervals, too wide or too narrow by a factor of n or $\sqrt{n}$.

A second problem with the theorem is that the Fisher information $I_1(\theta)$ is defined by an expectation, which may be difficult or impossible to derive in closed form. In that case, we can use the observed Fisher information, substituting $J_n(\hat{\theta}_n)$ for $I_n(\hat{\theta}_n)$. These will typically be close to each other. Assumptions similar to those of Theorem 10.3 (ignoring the same sort of remainder term) imply

$$\frac{1}{n} J_n(\hat{\theta}_n) \xrightarrow{P} I_1(\theta_0), \qquad \text{as } n \to \infty.$$

Since $J_n(\theta)$ involves no expectations, only derivatives, it can be calculated whenever the likelihood itself can be calculated, and hence can almost always be used in calculating standard errors. Of course, one can use observed Fisher information even when the expected Fisher information can also be calculated. One can use whichever seems more convenient.

10.3.5 Tests and Confidence Intervals

These variance estimates are combined using the "plug-in" theorem to construct asymptotically pivotal quantities for tests and confidence intervals. If z_α is the $1 - \alpha$ quantile of the standard normal distribution, then

$$\hat{\theta}_n \pm z_{\alpha/2} \frac{1}{\sqrt{I_n(\hat{\theta}_n)}}$$

or

$$\hat{\theta}_n \pm z_{\alpha/2} \frac{1}{\sqrt{J_n(\hat{\theta}_n)}}$$

is an asymptotic $100(1 - \alpha)\%$ confidence interval for θ. We can use whichever is convenient. If we are doing a hypothesis test, then

$$T_n = \left(\hat{\theta}_n - \theta_0\right) \sqrt{I_n(\hat{\theta}_n)} \tag{10.29a}$$

or

$$T_n = \left(\hat{\theta}_n - \theta_0\right) \sqrt{J_n(\hat{\theta}_n)} \tag{10.29b}$$

is an asymptotically pivotal quantity (either is asymptotically standard normal) that can be used to construct a test. A two-tailed test of

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_A &: \theta \neq \theta_0 \end{aligned}$$

rejects H_0 when $|T_n| \ge z_{\alpha/2}$, a one-tailed test of

$$\begin{aligned} H_0 &: \theta \le \theta_0 \\ H_A &: \theta > \theta_0 \end{aligned}$$

rejects H_0 when $T_n \ge z_\alpha$, and a one-tailed test of

$$\begin{aligned} H_0 &: \theta \ge \theta_0 \\ H_A &: \theta < \theta_0 \end{aligned}$$

rejects H_0 when $T_n \le -z_\alpha$.

This should all seem very familiar. It is just like all other asymptotic confidence intervals and tests. The only novelty is using observed or expected Fisher information to calculate the asymptotic standard error.

Example 10.3.9 (A Problem in Genetics).
In his influential monograph *Statistical Methods for Research Workers*, first published in 1925, R. A. Fisher described the following problem in genetics. The data are counts of seedling plants classified according to their values of two traits, green or which leaf color and starchy or sugary carbohydrate content

	green	white
starchy	1997	904
sugary	906	32

The probability model for data such as this is the *multinomial distribution* (Section 5.4 in these notes). Data are assumed to be observations on an i. i. d. sample of individuals classified into k categories (here $k = 4$, the number of cells in the table). Because of the assumption the individuals are identically distributed, each has the same probability of falling in the i-th cell of the table, denote it p_i. Because probabilities sum to one, we must have

$$p_1 + \cdots + p_k = 1 \tag{10.30}$$

If the entries in the table (the category counts are denoted X_i), then the joint density of these random variables is given by equation (3) on p. 187 in Lindgren

$$f_{\mathbf{p}}(\mathbf{x}) = \binom{n}{x_1, \ldots, x_k} \prod_{i=1}^{k} p_i^{x_i} \tag{10.31}$$

As the boldface type in the notation on the left hand side indicates, this describes the distribution of the random vector $\mathbf{X} = (X_1, \ldots, X_k)$ which depends on the vector parameter $\mathbf{p} = (p_1, \ldots, p_k)$. This is the multivariate analog of the binomial distribution.

The components X_i of this random vector are dependent. In fact, if n is the sample size, they must add to n because each individual falls in some cell of the table

$$X_1 + \cdots + X_k = n \tag{10.32}$$

Thus there are "really" only $n - 1$ random variables, because one can be eliminated using (10.32), and only $n - 1$ parameters, because one can be eliminated using (10.30). But doing this elimination of variables and parameters spoils the symmetry of (10.31). It does not simplify the formulas but complicates them. The log likelihood for the multinomial model is

$$l_n(\mathbf{p}) = \sum_{i=1}^{k} x_i \log(p_i)$$

(we can drop the multinomial coefficient because it does not contain parameters).

Returning to the genetics, Fisher was actually interested in a one-parameter submodel of the multinomial model, having cell probabilities specified by a single parameter θ, given by

	green	white
starchy	$\frac{1}{4}(2 + \theta)$	$\frac{1}{4}(1 - \theta)$
sugary	$\frac{1}{4}(1 - \theta)$	$\frac{1}{4}\theta$

In order for this to make sense as a submodel of the multinomial model, the cell probabilities must add to one (10.30), which is easily checked. They must also all be greater than zero, which requires $0 \le \theta \le 1$. The parameter θ has a scientific interpretation. Under a specific genetic model $\sqrt{\theta}$ is the *recombination fraction*, which is a measure of the distance along the chromosome between the two genes controlling these two traits (assuming they are on the same chromosome, if not then $\theta = 1/4$).

Numbering the cells of the table from one to four, going across rows starting at the upper left corner, the log likelihood becomes

$$l_n(\theta) = x_1 \log(2 + \theta) + (x_2 + x_3) \log(1 - \theta) + x_4 \log(\theta) \tag{10.33}$$

(where we dropped some more terms involving $\log(\frac{1}{4})$ but not containing the parameter). And the score is

$$l'_n(\theta) = \frac{x_1}{2 + \theta} - \frac{x_2 + x_3}{1 - \theta} + \frac{x_4}{\theta} \tag{10.34}$$

In order to find the maximum likelihood estimate, we need to solve the equation $l'_n(\theta) = 0$. Multiplying through by the product of the denominators gives

$$x_1(1 - \theta)\theta - (x_2 + x_3)(2 + \theta)\theta + x_4(2 + \theta)(1 - \theta) = 0$$

or simplifying a bit

$$2x_4 + (x_1 - 2x_2 - 2x_3 - x_4)\theta - (x_1 + x_2 + x_3 + x_4)\theta^2 = 0$$

or using (10.32)

$$n\theta^2 - (x_1 - 2x_2 - 2x_3 - x_4)\theta - 2x_4 = 0.$$

This is a quadratic equation with solutions

$$\hat{\theta}_n = \frac{x_1 - 2x_2 - 2x_3 - x_4 \pm \sqrt{(x_1 - 2x_2 - 2x_3 - x_4)^2 + 8nx_4}}{2n}$$

Since the square root is larger than the first term of the numerator, choosing the minus sign always gives a negative solution, which is not in the set of allowed parameter values. Hence the only "solution of the likelihood equation" is

$$\hat{\theta}_n = \frac{x_1 - 2x_2 - 2x_3 - x_4 + \sqrt{(x_1 - 2x_2 - 2x_3 - x_4)^2 + 8nx_4}}{2n} \tag{10.35}$$

In order to check whether this is a local or global maximum we need to look at the second derivative of the log likelihood, which also be needed to calculate Fisher information,

$$l''_n(\theta) = -\frac{x_1}{(2 + \theta)^2} - \frac{x_2 + x_3}{(1 - \theta)^2} - \frac{x_4}{\theta^2} \tag{10.36}$$

Since this is negative for all θ in the interior of the parameter space $(0 < \theta < 1)$, the log likelihood is strictly concave and (10.35) is the unique global maximum of the log likelihood.

Plugging in the actual data from the table

$$x_1 - 2x_2 - 2x_3 - x_4 = 1997 - 2(904 + 906) - 32 = -1655$$

and

$$n = 1997 + 904 + 906 + 32 = 3839$$

$$
\begin{aligned}
\hat{\theta}_n &= \frac{-1655 + \sqrt{(-1655)^2 + 8 \cdot 3839 \cdot 32}}{2 \cdot 3839} \\
&= \frac{-1655 + \sqrt{1655^2 + 982784}}{7678} \\
&= 0.0357123
\end{aligned}
$$

To make a confidence interval we need the Fisher information, either observed or expected (we'll do both to show how it's done, but in a real application you would choose one or the other). Finding the variance of the score (10.34) is a bit tricky because the x_i are correlated. Thus in this example the calculation of expected Fisher information using the second derivative is a good deal easier than the calculation using the first derivative. The observed Fisher information is just minus (10.36)

$$
J_n(\theta) = \frac{x_1}{(2+\theta)^2} + \frac{x_2 + x_3}{(1-\theta)^2} + \frac{x_4}{\theta^2}
$$

and the expected Fisher information is its expectation. Since the marginal distribution of X_i is $\text{Bin}(n, p_i)$ (Section 5.4.5 of these notes) $E(X_i) = np_i$ and

$$
\begin{aligned}
I_n(\theta) &= \frac{np_1}{(2+\theta)^2} + \frac{n(p_2 + p_3)}{(1-\theta)^2} + \frac{np_4}{\theta^2} \\
&= n\left(\frac{\frac{1}{4}(2+\theta)}{(2+\theta)^2} + \frac{\frac{1}{2}(1-\theta)}{(1-\theta)^2} + \frac{\frac{1}{4}\theta}{\theta^2} \right) \\
&= \frac{n}{4}\left(\frac{1}{2+\theta} + \frac{2}{1-\theta} + \frac{1}{\theta} \right)
\end{aligned}
$$

Plugging the data into these formulas gives

$$
\begin{aligned}
J_n(\hat{\theta}_n) &= \frac{1997}{(2 + 0.0357123)^2} + \frac{904 + 906}{(1 - 0.0357123)^2} + \frac{32}{0.0357123^2} \\
&= 27519.2
\end{aligned}
$$

and

$$
\begin{aligned}
I_n(\hat{\theta}_n) &= \frac{3839}{4}\left(\frac{1}{2 + 0.0357123} + \frac{2}{1 - 0.0357123} + \frac{1}{0.0357123} \right) \\
&= 29336.5
\end{aligned}
$$

We may use either to construct confidence intervals. If we use the observed Fisher information, we get

$$
\hat{\theta}_n \pm \frac{1.96}{\sqrt{J_n(\hat{\theta}_n)}}
$$

as a 95% confidence interval for the unknown true θ. The "plus or minus" is $1.96/\sqrt{27519.2} = 0.011815$, so our 95% confidence interval is 0.036 ± 0.012. If we use expected Fisher information instead the “plus or minus” would be 0.011443, almost the same.

To illustrate a hypothesis test, the natural null hypothesis to test is that the genes are *unlinked* (not on the same chromosome), in which case $\theta = 1/4$. Thus we test

$$\begin{aligned} H_0 &: \theta = 1/4 \\ H_A &: \theta \neq 1/4 \end{aligned}$$

Under certain genetic models a one-tailed test would be appropriate, but Fisher doesn't give us enough information about the data to know which tail to test, we will do a two-tailed test. The test statistic is either (10.29a) or (10.29b) with $\theta_0 = 1/4$, depending on whether we want to use observed or expected Fisher information. These give

$$\begin{aligned} \left(\hat{\theta}_n - \theta_0\right) \sqrt{I_n(\hat{\theta}_n)} &= (0.0357123 - 0.25)\sqrt{27519.2} = -35.548 \\ \left(\hat{\theta}_n - \theta_0\right) \sqrt{J_n(\hat{\theta}_n)} &= (0.0357123 - 0.25)\sqrt{29336.5} = -36.703 \end{aligned}$$

In either case the test statistics are so large that we get zero for the P-value (not exactly zero, R gives 4×10^{-277} for one and 3×10^{-295} for the other, but the normal approximation has no validity whatsoever this far out in the tail, so $P \approx 0$ is a more sensible answer). What this says, is that there is clear evidence of linkage (genes on the same chromosome). Here the evidence is so strong that there's almost no doubt remaining.

One moral of the story is that you can use either observed or expected Fisher information, whichever you prefer, whichever seems easier. They won't always give *exactly* the same confidence interval or P-value. But they both give valid asymptotic approximations (neither is *exactly* right but both are *approximately* right for large n and neither is preferred over the other).

Another moral of the story is a lesson about what hypothesis tests say. The test above says there is no question that $\theta \neq 1/4$, and hence the genes are linked. It does not say anything else. It does not tell you anything about the value of θ other than that it is not the value hypothesized under the null hypothesis. If you want to know more, you must look at a confidence interval.

10.4 Multiparameter Models

All of the preceding theory goes through to the multiparameter case. It just becomes more complicated. In particular, the MLE $\hat{\boldsymbol{\theta}}_n$ is a vector (obviously we need a vector estimator of a vector parameter). Hence Fisher information, to describe the variance of a vector, must become a matrix. Basically, that's the whole story. We just need to fill in the details.

10.4.1 Maximum Likelihood

Example 10.4.1 (Two-Parameter Normal Model).
The log likelihood for the two-parameter normal model is found by taking logs in (10.4) giving

$$l_n(\mu, \sigma) = -n \log(\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \tag{10.37}$$

Actually, the variance σ^2 is a more sensible parameter to estimate than the standard deviation σ. So define $\varphi = \sigma^2$, which means $\sigma = \varphi^{1/2}$. Plugging this into (10.37) gives

$$l_n(\mu, \varphi) = -\frac{n}{2} \log(\varphi) - \frac{1}{2\varphi} \sum_{i=1}^{n} (x_i - \mu)^2 \tag{10.38}$$

Differentiating gives

$$\begin{aligned} \frac{\partial l_n(\mu, \varphi)}{\partial \mu} &= \frac{1}{\varphi} \sum_{i=1}^{n} (x_i - \mu) \\ &= \frac{n(\bar{x}_n - \mu)}{\varphi} \qquad \text{(10.39a)} \\ \frac{\partial l_n(\mu, \varphi)}{\partial \varphi} &= -\frac{n}{2\varphi} + \frac{1}{2\varphi^2} \sum_{i=1}^{n} (x_i - \mu)^2 \qquad \text{(10.39b)} \end{aligned}$$

We have two partial derivatives. They are the two components of the score vector. We find a point where the first derivative *vector* is zero by setting them both to zero and solving the simultaneous equations. In general, this is hard. Here it is actually easy, because it is clear that (10.39a) is zero when and only when $\mu = \bar{x}_n$. So that is the MLE of μ (just as we found when we assumed the variance was known). Plugging that solution into (10.39b) gives

$$-\frac{n}{2\varphi} + \frac{1}{2\varphi^2} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 = -\frac{n}{2\varphi} + \frac{n v_n}{2\varphi^2}$$

where v_n is the usual variance of the empirical distribution (note not s_n^2). And this is clearly zero when $\varphi = v_n$. So that is the MLE of the variance φ.

Since we have two parameters, it is tempting to say "the MLE's are ...," but we can also think of *the* parameter as a vector $\boldsymbol{\theta} = (\mu, \varphi)$ and the MLE of this vector parameter is

$$\hat{\boldsymbol{\theta}}_n = \begin{pmatrix} \hat{\mu}_n \\ \hat{\varphi}_n \end{pmatrix} = \begin{pmatrix} \bar{x}_n \\ v_n \end{pmatrix} \tag{10.40}$$

Actually, we are being a bit overconfident here. What we have found is a zero of the first derivative, in fact, the only zero. But we haven't shown yet that this is even a local maximum, much less a global maximum.

So we look at the second derivative matrix. This has components

$$\frac{\partial^2 l_n(\mu, \varphi)}{\partial \mu^2} = -\frac{n}{\varphi} \tag{10.41a}$$

$$\frac{\partial^2 l_n(\mu, \varphi)}{\partial \mu \partial \varphi} = -\frac{n(\bar{x}_n - \mu)}{\varphi^2} \tag{10.41b}$$

$$\frac{\partial^2 l_n(\mu, \varphi)}{\partial \varphi^2} = \frac{n}{2\varphi^2} - \frac{1}{\varphi^3} \sum_{i=1}^{n} (x_i - \mu)^2 \tag{10.41c}$$

or, more precisely, the second derivative is a 2×2 matrix having (10.41a) and (10.41c) as diagonal elements and off-diagonal elements (10.41b) (both the same because a second derivative matrix is symmetric). Unfortunately, this is not a negative definite matrix for all values of parameters and data, because (10.41c) is not always negative. Thus the log likelihood is not strictly concave, and the theory developed in Appendix G does not guarantee (10.40) is a global maximum.[2]

If we evaluate the second derivative matrix at the MLE, we get considerable simplification. When we plug in $\mu = \bar{x}_n$ (10.41b) is zero. Thus we get a diagonal matrix

$$\nabla^2 l_n(\hat{\boldsymbol{\theta}}_n) = \begin{pmatrix} -\frac{n}{\hat{\varphi}_n} & 0 \\ 0 & -\frac{n}{2\hat{\varphi}_n^2} \end{pmatrix} \tag{10.42}$$

the 1,1 component being (10.41a) with the MLE plugged in for the parameter, and the 2,2 component being (10.41c) simplified by using the fact that the sum in (10.41c) is $n v_n = n \hat{\varphi}_n$ when the MLE is plugged in for the parameter. Now (10.42) is negative definite (a diagonal matrix is negative definite if and only if each element on the diagonal is negative). So the theory we know does establish that (10.40) is a local maximum of the log likelihood.

Problems that work out so simply are quite rare. Usually there are no obvious solutions of the “likelihood equations” (first partial derivatives set equal to zero). In most problems the only way to find MLE’s is ask a competent computer.

Example 10.4.2 (Cauchy Location-Scale Model).
The Cauchy Location-Scale model has densities

$$f_{\mu,\sigma}(x) = \frac{1}{\pi} \cdot \frac{\sigma}{\sigma^2 + (x - \mu)^2}$$

(p. 191 in Lindgren). Here μ is the location parameter and σ is the scale parameter (of course we could use any other Greek letters for these parameters). We know that the Cauchy distribution is symmetric and hence μ is the population median. We used the sample median as an estimator of μ in Problem 7-6(a).

[2] It actually is a global maximum, but we won’t develop the theory needed to show that.

The log likelihood for an i. i. d. sample of size n is

$$l_n(\mu, \sigma) = n \log(\sigma) - \sum_{i=1}^{n} \log\left(\sigma^2 + (x_i - \mu)^2\right)$$

There's no point in differentiating. If we had to use the computer for the one-parameter Cauchy model, the two-parameter model certainly isn't going to be simple enough to do by hand. We proceed just as in the one-parameter problem (Example 10.2.3).

Define an R function that evaluates minus the log likelihood (minus because the `nlm` function minimizes rather than maximizes)

```
> l <- function(theta) {
+     mu <- theta[1]
+     sigma <- theta[2]
+     return(- n * log(sigma) + sum(log(sigma^2 + (x - mu)^2)))
+ }
```

Then we hand this to the `nlm` function as the function to be minimized. But first we have to figure out what to use as a starting point. The better starting point we have, the better chance that we find the right local minimum if more than one exists. We know a good estimator of μ, the sample median. What might be a good estimator of scale? Variance is no good. The Cauchy distribution doesn't even have a mean, much less a variance. The only other general estimator of scale that has even been mentioned is IQR (p. 202 in Lindgren) and we've never used it in any examples, nor do we know anything about its properties. However, it's the only idea we have, so let's try it. The reason that IQR works as a scale estimator is that the IQR of a general Cauchy distribution is just σ times the IQR of a standard Cauchy distribution (obvious from a picture of the density). The IQR of the standard Cauchy happens to be simple

```
> qcauchy(0.75) - qcauchy(0.25)
[1] 2
```

Thus half the IQR of the data is a sensible estimate of scale, and a good starting point for the optimization algorithm.

In real life, we would use real data. Here we just make up the data.

```
> n <- 80
> mu <- 0
> sigma <- 1
> x <- mu + sigma * rcauchy(n)  # make up data
> summary(x)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-38.2900  -0.6647   0.1869   0.6895   1.0630  54.1200
```

and then hand the data to the optimizer

```
out <- nlm(l, c(median(x), IQR(x) / 2), fscale=n, hessian=TRUE)
```

The result returned by the optimizer, saved in the variable `out`, which is a list with several components, only the interesting ones of which we show below

```
> out
$minimum
[1] 123.7644

$estimate
[1] 0.2162629 0.9229725

$gradient
[1] -6.470202e-05  5.941558e-05

$hessian
          [,1]      [,2]
[1,] 50.320659  2.273821
[2,]  2.273821 43.575535
```

- `estimate` is the optimum parameter value, that is

$$\hat{\boldsymbol{\theta}}_n = \begin{pmatrix} \hat{\mu}_n \\ \hat{\sigma}_n \end{pmatrix} = \begin{pmatrix} 0.2162629 \\ 0.9229725 \end{pmatrix}$$

- `minimum` is the optimal value of the objective function, that is

$$l_n(\hat{\boldsymbol{\theta}}_n) = -123.7644$$

  (recall that the objective function handed to `nlm` is *minus* the log likelihood).

- `gradient` is $-\nabla l_n(\hat{\boldsymbol{\theta}}_n)$ the first derivative vector of the objective function at the MLE. It should be zero to convergence tolerance of the optimization algorithm (and it is).

- `hessian` is $-\nabla^2 l_n(\hat{\boldsymbol{\theta}}_n)$, the second derivative matrix of the objective function at the MLE.

If we want to check that this is a local maximum of the log likelihood (hence a local *minimum* of the objective function passed to `nlm`) we check whether the `hessian` component is positive definite

```
> eigen(out$hessian)
$values
[1] 51.01558 42.88061

$vectors
           [,1]       [,2]
[1,] -0.9563346  0.2922741
[2,] -0.2922741 -0.9563346
```

Since both eigenvalues are positive, the Hessian is positive definite and the solution is a local maximum of the log likelihood.

Whether done by hand or done by computer, the process is much the same. Find a zero of the first derivative, and the second derivative tells you whether it is a local maximum or not (since the Hessian returned by the computer is the Hessian at the reported optimal value, it is of no use in checking whether you have a global maximum). As we will see in the next section, the Hessian is also observed Fisher information, and hence tells us important things about the asymptotic sampling distribution of the MLE.

10.4.2 Sampling Theory

We now have to repeat all of Section 10.3 giving the multiparameter analogs of everything in there. We will omit details that are basically the same as in the uniparameter case and concentrate on the differences.

In the multiparameter case the score is a vector $\nabla l_n(\boldsymbol{\theta})$. The *Fisher information* is, as in the uniparameter case, its variance

$$\mathbf{I}_n(\boldsymbol{\theta}) = \operatorname{var}_{\boldsymbol{\theta}}\{\nabla l_n(\boldsymbol{\theta})\}, \tag{10.43}$$

but now, since the variance of a random vector is a (nonrandom) matrix, the Fisher information is a matrix. Like any variance matrix, it is symmetric square and positive semi-definite. If $\nabla l_n(\boldsymbol{\theta})$ is not concentrated on a hyperplane (see Section 5.1.9 in these notes), then the Fisher information is actually positive definite.

As in the uniparameter case we sometimes call $\mathbf{I}_n(\boldsymbol{\theta})$ the "expected" Fisher information for contrast with "observed" Fisher information, but, strictly speaking the "expected" is redundant. "Fisher information" with no qualifying adjective always means $\mathbf{I}_n(\boldsymbol{\theta})$.

We still have the multiparameter analogs of the two important theorems about the score and Fisher information (Theorems 10.1 and 10.2).

Theorem 10.4. *Provided first and second order partial derivatives of* (10.10) *with respect to components of θ can be taken under the integral sign,*

$$E_{\boldsymbol{\theta}}\{\nabla l_n(\boldsymbol{\theta})\} = 0 \tag{10.44a}$$

and

$$E_{\boldsymbol{\theta}}\{\nabla^2 l_n(\boldsymbol{\theta})\} = -\operatorname{var}_{\boldsymbol{\theta}}\{\nabla l_n(\boldsymbol{\theta})\} \tag{10.44b}$$

for all values of $\boldsymbol{\theta}$ for which the differentiation under the integral sign is permitted.

Theorem 10.5.

$$\mathbf{I}_n(\boldsymbol{\theta}) = n\mathbf{I}_1(\boldsymbol{\theta})$$

The proofs of these two theorems are exactly the same as in the uniparameter case. One only has to use partial derivatives instead of ordinary derivatives to prove Theorem 10.4. Theorem 10.5 follows just as in the uniparameter case from the fact that the variance of a sum is the sum of the variance when the terms are independent, the multivariate version of which is (5.10) in Chapter 5 of these notes.

As in the uniparameter case, (10.44b) tells us we can calculate Fisher information in two quite different ways, either (10.43) or

$$\mathbf{I}_n(\boldsymbol{\theta}) = -E_{\boldsymbol{\theta}}\{\nabla^2 l_n(\boldsymbol{\theta})\}. \tag{10.45}$$

It's your choice. Use whichever seems easier.

Also as in the uniparameter case, Theorem 10.5 tells us we can calculate Fisher information using sample size one (which won't have any summations) and multiply by n.

Example 10.4.3 (Two-Parameter Normal Model).
The log likelihood and derivatives for this model were figured out in Example 10.4.1. The components of the (expected) Fisher information are the negative expectations of (10.41a), (10.41b), and (10.41c), that is,

$$\begin{aligned}
-E\left(\frac{\partial^2 l_n(\mu,\varphi)}{\partial \mu^2}\right) &= \frac{n}{\varphi} \\
-E\left(\frac{\partial^2 l_n(\mu,\varphi)}{\partial \mu \partial \varphi}\right) &= \frac{n[E(\overline{X}_n) - \mu]}{\varphi^2} \\
&= 0 \\
-E\left(\frac{\partial^2 l_n(\mu,\varphi)}{\partial \varphi^2}\right) &= -\frac{n}{2\varphi^2} + \frac{1}{\varphi^3}\sum_{i=1}^n E\{(x_i - \mu)^2\} \\
&= -\frac{n}{2\varphi^2} + \frac{n\varphi}{\varphi^3} \\
&= \frac{n}{2\varphi^2}
\end{aligned}$$

Thus the Fisher information is the diagonal matrix

$$\mathbf{I}_n(\boldsymbol{\theta}) = \begin{pmatrix} \frac{n}{\varphi} & 0 \\ 0 & \frac{n}{2\varphi^2} \end{pmatrix} \tag{10.46}$$

Observed Fisher information is also defined in the same way as in the uniparameter case, as minus the second derivative of the log likelihood

$$\mathbf{J}_n(\boldsymbol{\theta}) = -\nabla^2 l_n(\boldsymbol{\theta}). \tag{10.47}$$

Note that the second derivative is a matrix here, which is a good thing, because the expected Fisher information is also a matrix.

The CLT and the LLN apply in exactly the same way to the score and observed Fisher information. The analog of (10.18) is

$$\frac{1}{\sqrt{n}} \nabla l_n(\boldsymbol{\theta}) \xrightarrow{\mathcal{D}} \mathcal{N}\bigl(0, \mathbf{I}_1(\boldsymbol{\theta})\bigr) \tag{10.48}$$

(just the same, except for some boldface type). Of course, this is a multivariate CLT because $\nabla l_n(\boldsymbol{\theta})$ is a random vector rather than a random scalar. The sloppy "double squiggle" version is

$$\nabla l_n(\boldsymbol{\theta}) \approx \mathcal{N}\bigl(0, \mathbf{I}_n(\boldsymbol{\theta})\bigr).$$

The analog of (10.22) is

$$\frac{1}{n} \mathbf{J}_n(\boldsymbol{\theta}) \xrightarrow{P} \mathbf{I}_1(\boldsymbol{\theta}). \tag{10.49}$$

(again, just the same, except for some boldface type). It is, of course, a multivariate convergence in probability statement. The sloppy "double squiggle" version would be

$$\mathbf{J}_n(\boldsymbol{\theta}) \approx \mathbf{I}_n(\boldsymbol{\theta})$$

Still proceeding as in the univariate case, expanding ∇l_n using a Taylor series with remainder about the true parameter value $\boldsymbol{\theta}_0$, we get

$$\nabla l_n(\boldsymbol{\theta}) = \nabla l_n(\boldsymbol{\theta}_0) + \nabla^2 l_n(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \text{remainder}$$

This is a bit harder to interpret than the uniparameter analog. First, it is a vector equation, each term having k components if there are k parameters. As in the uniparameter case, we can use $\nabla^2 l_n(\boldsymbol{\theta}) = -\mathbf{J}_n(\boldsymbol{\theta})$ to rewrite this as

$$\nabla l_n(\boldsymbol{\theta}) = \nabla l_n(\boldsymbol{\theta}_0) - \mathbf{J}_n(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \text{remainder}, \tag{10.50}$$

but that still doesn't make it obvious what k-dimensional vector the middle term on the right hand side is supposed to be. Since $\mathbf{J}_n(\boldsymbol{\theta}_0)$ is a $k \times k$ matrix and $\boldsymbol{\theta} - \boldsymbol{\theta}_0$ is a k vector, this must be a matrix multiplication, which does indeed produce a k vector. We won't bother to write out the "remainder" term in detail.

Now we apply the same sort of argument we used in the uniparameter case. If the MLE is in the interior of the parameter space, the first derivative of the log likelihood will be zero at the MLE. So if we plug in the MLE for $\boldsymbol{\theta}$, the left hand side of (10.50) is zero, and we get

$$\nabla l_n(\boldsymbol{\theta}_0) \approx \mathbf{J}_n(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \tag{10.51}$$

(we dropped the remainder term, assuming as in the uniparameter case that it is asymptotically negligible, and replaced the equals sign with a "double squiggle" to indicate this is not an exact inequality). Again we divide through by $\sqrt{n}$ to make both sides the right size to converge in distribution to a nontrivial random variable

$$\frac{1}{\sqrt{n}} \nabla l_n(\boldsymbol{\theta}_0) \approx \frac{1}{n} \mathbf{J}_n(\boldsymbol{\theta}_0) \cdot \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \tag{10.52}$$

This is almost, but not quite, what we need. We need $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ by itself on one side. The way to do that is multiply through by the inverse of the matrix multiplying it.

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \approx \left(\frac{1}{n} \mathbf{J}_n(\boldsymbol{\theta}_0) \right)^{-1} \frac{1}{\sqrt{n}} \nabla l_n(\boldsymbol{\theta}_0) \tag{10.53}$$

Because matrix inversion is a continuous operation, the continuous mapping theorem and (10.49) imply that the first factor on the right hand side converges in probability to $\mathbf{I}_1(\boldsymbol{\theta}_0)^{-1}$. The convergence in distribution for second factor on the right hand side is given by (10.48). By Slutsky's theorem, the right hand side converges to the product, that is

$$\left(\frac{1}{n} \mathbf{J}_n(\boldsymbol{\theta}_0) \right)^{-1} \frac{1}{\sqrt{n}} \nabla l_n(\boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} \mathbf{I}_1(\boldsymbol{\theta}_0)^{-1} \mathbf{Y}$$

where

$$\mathbf{Y} \sim \mathcal{N}(0, \mathbf{I}_1(\boldsymbol{\theta}_0)^{-1}).$$

Since a linear transformation of a multivariate normal is multivariate normal (Theorem 12 of Chapter 12 in Lindgren), $\mathbf{I}_1(\boldsymbol{\theta}_0)^{-1}\mathbf{Y}$ is multivariate normal with mean vector

$$E\left\{ \mathbf{I}_1(\boldsymbol{\theta}_0)^{-1}\mathbf{Y} \right\} = \mathbf{I}_1(\boldsymbol{\theta}_0)^{-1} E(\mathbf{Y}) = 0$$

and variance matrix

$$\begin{aligned}
\operatorname{var}\left\{ \mathbf{I}_1(\boldsymbol{\theta}_0)^{-1}\mathbf{Y} \right\} &= \mathbf{I}_1(\boldsymbol{\theta}_0)^{-1} \operatorname{var}(\mathbf{Y}) \mathbf{I}_1(\boldsymbol{\theta}_0)^{-1} \\
&= \mathbf{I}_1(\boldsymbol{\theta}_0)^{-1} \mathbf{I}_1(\boldsymbol{\theta}_0) \mathbf{I}_1(\boldsymbol{\theta}_0)^{-1} \\
&= \mathbf{I}_1(\boldsymbol{\theta}_0)^{-1}
\end{aligned}$$

the middle equality being the formula for the variance of a (multivariate) linear transformation, (5.18b) in Chapter 5 of these notes.

Thus we have arrived at the multiparameter version of the "usual asymptotics" of maximum likelihood.

Theorem 10.6. *If the true parameter value $\boldsymbol{\theta}_0$ is in the interior of the parameter space, first and second order partial derivatives of* (10.10) *with respect to components of θ can be taken under the integral sign, and the difference of the two sides of* (10.52) *converges in probability to zero, then*

$$\sqrt{n}\big(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\big) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, \mathbf{I}_1(\boldsymbol{\theta}_0)^{-1}\big) \tag{10.54}$$

This looks just like the uniparameter version (10.27), except for some boldface type.

As in the uniparameter case, the conditions of the theorem are hard to verify. We often use it without any attempt to verify the conditions. As we remarked in regard to the uniparameter case, even if you have verified the conditions, that still doesn't prove that the normal approximation given by the theorem is good at the actual n in which you are interested (the sample size of the data in some real application). Hence our slogan about asymptotics only providing a heuristic applies. If you are worried about the validity of the asymptotics, check it with computer simulations.

One final point, just like in the uniparameter case, Theorem 10.6 must be combined with the plug-in theorem to get a useful result. Since we don't know the true parameter value θ_0, we don't know the asymptotic variance $\mathbf{I}_1(\boldsymbol{\theta}_0)^{-1}$ and must estimate it. Plugging either observed or expected Fisher information evaluated at the MLE gives

$$\hat{\boldsymbol{\theta}}_n \approx \mathcal{N}\big(\boldsymbol{\theta}_0, \mathbf{I}_n(\hat{\boldsymbol{\theta}}_n)^{-1}\big)$$

or

$$\hat{\boldsymbol{\theta}}_n \approx \mathcal{N}\big(\boldsymbol{\theta}_0, \mathbf{J}_n(\hat{\boldsymbol{\theta}}_n)^{-1}\big).$$

Example 10.4.4 (Two-Parameter Normal Model).
The MLE for this model is given by (10.40) in Example 10.4.1. The observed Fisher information evaluated at $\hat{\boldsymbol{\theta}}_n$ is given by minus (10.42) in the same example. The expected Fisher information is given by (10.46) in Example 10.4.3. When evaluated at the MLE, observed and expected Fisher information are the same

$$\mathbf{I}_n(\hat{\boldsymbol{\theta}}_n) = \mathbf{J}_n(\hat{\boldsymbol{\theta}}_n) = \begin{pmatrix} \frac{n}{\hat{\varphi}_n} & 0 \\ 0 & \frac{n}{2\hat{\varphi}_n^2} \end{pmatrix}$$

Inverting a diagonal matrix is easy. Just invert each term on the diagonal.

$$\mathbf{I}_n(\hat{\boldsymbol{\theta}}_n)^{-1} = \mathbf{J}_n(\hat{\boldsymbol{\theta}}_n)^{-1} = \begin{pmatrix} \frac{\hat{\varphi}_n}{n} & 0 \\ 0 & \frac{2\hat{\varphi}_n^2}{n} \end{pmatrix}$$

That's our estimate of the asymptotic variance of the MLE (vector).

This example doesn't tell us anything we didn't already know. It says the MLE's of μ and of $\varphi = \sigma^2$ are asymptotically independent, because the covariance is zero and uncorrelated jointly multivariate normal random variables are independent (Theorem 4 of Chapter 12 in Lindgren). This is no surprise, because we know that the MLE's $\overline{X}_n$ and V_n are actually independent (not just asymptotically) by the corollary to Theorem 10 of Chapter 7 in Lindgren. Since they are independent, their joint distribution is uninteresting (just the product of the marginals), and what this says about the marginals we also have long known

$$\hat{\mu}_n = \overline{X}_n \approx \mathcal{N}\left(\mu, \frac{V_n}{n}\right)$$

(which is just the CLT plus the plug-in theorem) and

$$\hat{\varphi}_n = V_n \approx \mathcal{N}\left(\sigma^2, \frac{2V_n^2}{n}\right)$$

This may not ring a bell right away. It is the asymptotic distribution of V_n worked out in Example 7.3.6 in these notes plus the plug-in theorem.

Example 10.4.5 (Cauchy Location-Scale Model).
The MLE and observed Fisher information for this model were found in Example 10.4.5. Of course they were not found by hand calculation, just by numerical optimization using the computer. To repeat the relevant bits of the computer output,

```
$estimate
[1] 0.2162629 0.9229725
```

is the MLE $\hat{\boldsymbol{\theta}}_n$, and

```
$hessian
          [,1]      [,2]
[1,] 50.320659  2.273821
[2,]  2.273821 43.575535
```

is the observed Fisher information evaluated at the MLE $\mathbf{J}_n(\hat{\boldsymbol{\theta}}_n)$.

Now to make confidence intervals for the parameters we need to calculate *inverse* Fisher information, because that is the asymptotic variance of the MLE. The R function that inverts matrices has the totally unintuitive name `solve` (because it also solves linear equations). Hence the inverse Fisher information $\mathbf{J}_n(\hat{\boldsymbol{\theta}}_n)^{-1}$ is given by

```
> solve(out$hessian)
             [,1]         [,2]
[1,]  0.019919520 -0.001039426
[2,] -0.001039426  0.023002887
```

The numbers on the diagonal are the asymptotic variances of the first component of the MLE ($\hat{\mu}_n$) and of the second component ($\hat{\sigma}_n$). So

```
> avar <- solve(out$hessian)
> out$estimate[1] + c(-1,1) * qnorm(0.975) * sqrt(avar[1,1])
[1] -0.0603596  0.4928854
> out$estimate[2] + c(-1,1) * qnorm(0.975) * sqrt(avar[2,2])
[1] 0.6257106 1.2202344
```

give asymptotic 95% confidence intervals for μ and σ, respectively. Note these are not *simultaneous* confidence intervals, because we did not do Bonferroni correction.

There are two important points illustrated by the last example.

- If you can write down the log likelihood for a model, you can do likelihood inference, without doing any derivatives or expectations. (In this example, the computer found the MLE and calculated the second derivative matrix by finite differences. We didn't do any differentiation. And because we used observed rather than expected Fisher information, we didn't need to do any integrals either.)

- In order to calculate asymptotic variances, you do need to invert the Fisher information matrix, but the computer easily does that too.

Honesty compels me to admit that this example is not as convincing as it might be, because Mathematica easily calculates the expected Fisher information and it is diagonal and so trivially invertible. In fact, all location-scale models with symmetric densities have diagonal Fisher information. You will just have to take my word for it that there are many interesting complicated models that arise in applied statistics (too complicated to discuss here) for which you can't do anything but write down the log likelihood and hand it to the computer to do the rest, just like in this example.

10.4.3 Likelihood Ratio Tests

One last subject before we are done with sampling theory: In this section we learn about a completely different kind of hypothesis test. All of the tests we studied in the last chapter (and in this chapter up to here) had null hypotheses that fixed the value of *one* parameter (the so-called *parameter of interest*) and said nothing about any other parameters (the so-called *nuisance parameters*). Now we are going to learn about tests with multiple parameters of interest.

A likelihood ratio test compares two models, which we will call the *little model* and the *big model*. The little model is a *submodel* of the big model. Another term commonly used to describe this situation is *nested* models (one is a submodel of the other).

Roughly speaking, the little and big models correspond to the null and alternative hypotheses for the test. To be precise, let Θ be the whole parameter space of the problem, which is the parameter space of the big model, and let Θ_0 be the parameter space of the little model. These are nested models if $\Theta_0 \subset \Theta$. The null hypothesis corresponds to the little model,

$$H_0 : \theta \in \Theta_0, \tag{10.55}$$

but the alternative hypothesis is not supposed to include the null and hence must correspond to the part of the big model that is not in the little model

$$H_0 : \theta \in \Theta_A, \tag{10.56}$$

where $\Theta_A = \Theta \setminus \Theta_0$ (the operation indicated here is called "set difference" and says that Θ_A consists of points in Θ that are not in Θ_0, that is, are in the big model but not in the little model).

This is almost enough in the way of preliminary discussion to describe the likelihood ratio test, but not quite. What we need is for the little model to be a *smooth* submodel of the big model. The simplest way to describe that is as follows. Suppose that the big model has m parameters, which we can also think of as a single vector parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$, so $\boldsymbol{\theta}$ is a point in m-dimensional Euclidean space $\mathbb{R}^m$, and Θ is a subset of $\mathbb{R}^m$. We need the little model to be a smooth k-dimensional surface in Θ. The simplest way to describe such a surface is by a differentiable map $\mathbf{g}$ from a subset Φ of k-dimensional Euclidean space $\mathbb{R}^k$ into Θ. We let Θ_0 be the *range* of the function $\mathbf{g}$, that is,

$$\Theta_0 = g(\Phi) = \{\, \mathbf{g}(\boldsymbol{\varphi}) : \boldsymbol{\varphi} \in \Phi \,\}$$

This gives us two ways to think of the little model. When we think of it as a submodel of the big model, it has parameter $\boldsymbol{\theta} \in \Theta_0$, which is an m-dimensional parameter, just like the parameter of the big model. In fact, since the models are nested, each parameter value $\boldsymbol{\theta}$ in the little model is also a possible parameter value of the big model. But this parameter cannot be varied freely. In order to do maximum likelihood, we must use the parameterization $\boldsymbol{\varphi}$. We say the big model has m free parameters, but the little model only has k free parameters.

To do the likelihood ratio test, we first find the MLE's for the two models. We denote the MLE in the big model $\hat{\boldsymbol{\theta}}_n$, and we denote the MLE in the little model by $\boldsymbol{\theta}_n^* = \mathbf{g}(\hat{\boldsymbol{\varphi}}_n)$. (We can't call them both "theta hat." We wouldn't be able to tell them apart.)

The *likelihood ratio test statistic* for comparing these two models, that is, for testing the hypotheses (10.55) and (10.56) is

$$2[l_n(\hat{\boldsymbol{\theta}}_n) - l_n(\boldsymbol{\theta}_n^*)]. \tag{10.57}$$

It is twice the log of the maximized likelihood ratios for the two models

$$2 \log \left(\frac{\max_{\boldsymbol{\theta} \in \Theta_A} L_n(\boldsymbol{\theta})}{\max_{\boldsymbol{\theta} \in \Theta_0} L_n(\boldsymbol{\theta})} \right), \tag{10.58}$$

where we are using the convention that big L is the likelihood and little l the log likelihood: $l_n(\boldsymbol{\theta}) = \log L_n(\theta)$. To see why these are the same, the parameter value at which the maximum in the numerator is achieved is by definition $\hat{\boldsymbol{\theta}}_n$, and the parameter value at which the maximum in the denominator is achieved is by definition $\boldsymbol{\theta}_n^*$, so (10.58) is equal to

$$2 \log \left(\frac{L_n(\hat{\boldsymbol{\theta}}_n)}{L_n(\boldsymbol{\theta}_n^*)} \right)$$

and by rule for the the log of a ratio, this is the same as (10.57).

Why this is interesting is the following, which for once we do not state as a formal theorem. If we assume

(i) the null hypothesis is correct, that is, the true parameter value has the form $\boldsymbol{\theta}_0 = \mathbf{g}(\boldsymbol{\varphi}_0)$ for some point $\boldsymbol{\varphi}_0$ in Φ, and

(ii) $\boldsymbol{\varphi}_0$ is an interior point of Φ, and $\boldsymbol{\theta}_0$ is an interior point of Θ,

then under all of the conditions required for the usual asymptotics of maximum likelihood (Theorem 10.6) plus a little bit more (we for once omit the gory details) the asymptotic distribution of (10.57) or (10.58) is $\text{chi}^2(m-k)$.

This is quite a remarkable property of maximum likelihood. When doing a *likelihood ratio test*, one using (10.57) or (10.58) as the test statistic, the asymptotic distribution of the test statistic does not depend on any details of the model. You simply calculate the MLE's in the big and little model, calculate the difference in log likelihoods, multiply by two, and compare to the chi-square distribution with the appropriate number of degrees of freedom.

Example 10.4.6 (Equality of Poisson Means).
Consider the following data, which are counts in regions of equal area in what is assumed to be a Poisson process, which makes the counts independent Poisson random variables.

$$26 \quad 37 \quad 31 \quad 30 \quad 26 \quad 42$$

The question we want to examine is whether the Poisson process is *homogeneous* or *inhomogeneous*. If homogeneous, the counts have mean $\mu = \lambda A$, and since the area A is assumed to be the same for each, the counts all have the same mean, and since the mean is the parameter of the Poisson distribution, that means they all have the same distribution. This is our null hypothesis. The counts X_i are i. i. d. $\text{Poi}(\mu)$. This is the little model. The big model allows unequal means $\mu_i = \lambda_i A$. So in this model the X_i are independent but not identically distributed $X_i \sim \text{Poi}(\mu_i)$. The little model has one free parameter (dimension $k = 1$, and the big model has m parameters if there are m counts in the data, here $m = 6$).

The MLE for the little model, data i. i. d. $\text{Poi}(\mu)$ has already been found in Problem 7-42(c) in Lindgren. It is the sample mean, which here we write $\hat{\mu} = \bar{x}_m$. (This is the $\hat{\varphi}$ parameter estimate in the general discussion.) The corresponding parameter in the big model is m-dimensional with all components equal

$$\boldsymbol{\mu}^* = \begin{pmatrix} \bar{x}_m \\ \vdots \\ \bar{x}_m \end{pmatrix}$$

(This is the $\boldsymbol{\theta}^*$ parameter estimate in the general discussion).

The MLE in the big model also seems obvious. The only data relevant to the mean μ_i is the count x_i, so we "really" have m separate problems, and the MLE is given by the special case $m = 1$ of the solution for the little model, that is, $\hat{\mu}_i = x_i$ and the vector MLE is just the data vector

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_m \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}$$

Plausible though this may be, it is not completely convincing. We should wade through the gory details to be sure this is really the MLE in the big model. The density for x_i is

$$f_{\mu_i}(x_i) = \frac{\mu_i^{x_i}}{x_i!} e^{-\mu_i}$$

The joint density is the product (because the x_i are independent)

$$f_{\boldsymbol{\mu}}(\mathbf{x}) = \prod_{i=1}^{m} \frac{\mu_i^{x_i}}{x_i!} e^{-\mu_i}$$

The $x_i!$ terms do not contain parameters and can be dropped from the likelihood

$$L(\boldsymbol{\mu}) = \prod_{i=1}^{m} \mu_i^{x_i} e^{-\mu_i}$$

So the log likelihood is

$$l(\boldsymbol{\mu}) = \sum_{i=1}^{m} \bigl(x_i \log(\mu_i) - \mu_i \bigr) \tag{10.59}$$

The derivatives are

$$\begin{aligned} \frac{\partial l(\boldsymbol{\mu})}{\partial \mu_i} &= \frac{x_i}{\mu_i} - 1 \\ \frac{\partial^2 l(\boldsymbol{\mu})}{\partial \mu_i^2} &= -\frac{x_i}{\mu_i^2} \\ \frac{\partial^2 l(\boldsymbol{\mu})}{\partial \mu_i \partial \mu_j} &= 0, \qquad i \neq j \end{aligned}$$

Since the second derivative is diagonal with negative diagonal elements, it is negative definite and the log likelihood is a strictly convex function. So the MLE is found by setting the first derivatives equal to zero and solving, which does indeed give $\hat{\mu}_i = x_i$.

The likelihood ratio test statistic is found by plugging $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{\mu}^*$ into the log likelihood (10.59), subtracting, and multiplying by two. Here's how to do it in R

```
> x <- c(26, 37, 31, 30, 26, 42)
> m <- length(x)
> mu.star <- mean(x)
> mu.hat <- x
> l <- function(mu, x) sum(x * log(mu) - mu)
> lrt <- 2 * (l(mu.hat, x) - l(mu.star, x))
> 1 - pchisq(lrt, m - 1)
[1] 0.2918324
```

There are $m - 1$ degrees of freedom in the chi-square distribution because the little model has one parameter and the big model has $m = 6$ parameters. The P-value $P = 0.29$ obviously is not close to statistical significance by any reasonable criterion. Thus we "accept" the little model, and conclude that the data may well be from a homogeneous Poisson process.

There are a couple issues about this example that may be bother you. First, what happened to n? How can we do a "large n" analysis, when there is no n? If m is supposed to be n, it certainly isn't large. This has to do with a special property of the Poisson distribution which we saw in getting a normal approximation. As it says in Appendix F,

$$\mathrm{Poi}(\mu) \approx \mathcal{N}(\mu, \mu)$$

if μ *is large*. There doesn't have to be any n for the CLT to apply. So asymptotics work here not because m is large, but because the means of the X_i are large.

As usual though, we have our dictum that asymptotics only provides a heuristic. If you are worried about the validity of the asymptotics, you simulate. This your author did, and the asymptotics provide a very good approximation here (details not shown, you'll have to take my word for it).

A test of this sort in which the whole point is to accept the little model is often called a *goodness of fit* test. When the little model is accepted, we say it seems to fit the data well. At least the test gives no evidence that the big model fits any better. Thus the principle of parsimony (other things being equal, simpler is better) says to choose the little model.

There is no difference between goodness of fit tests and any other kind of tests except which hypothesis you are rooting for. When you like the simpler model, you call the test a goodness of fit test and are happy when the null hypothesis is accepted, and you conclude that the little model fits just as well as the bigger, more complicated model to which it was compared. When you like the more complicated model, there is no special term for that situation, because that describes most tests. But then you are happy when the null hypothesis is rejected, and you conclude that the complexity of the big model is necessary to fit the data well.

Example 10.4.7 (A Problem in Genetics).
This revisits Example 10.3.9. Here we want to do a goodness of fit test. The little model is the model fit in Example 10.3.9. The big model to which we compare it is the general multinomial model. The log likelihood for the big model is given by the unnumbered displayed equation on p. 10.3.9

$$l_n(\mathbf{p}) = \sum_{i=1}^{k} x_i \log(p_i)$$

because of the constraint (10.30) there are actually only $k - 1$ free parameters in the big model, and in order to fit the model by maximum likelihood we must eliminate on of the parameters by writing it in terms of the others

$$p_k = 1 - p_1 + \cdots + p_{k=1} \tag{10.60}$$

Then we get

$$\begin{aligned}
\frac{\partial l_n(\mathbf{p})}{\partial p_i} &= \frac{x_i}{p_i} + \frac{x_k}{p_k} \cdot \frac{\partial p_k}{\partial p_i} \\
&= \frac{x_i}{p_i} - \frac{x_k}{p_k} \\
\frac{\partial^2 l_n(\mathbf{p})}{\partial p_i^2} &= - \frac{x_i}{p_i^2} - \frac{x_k}{p_k^2} \\
\frac{\partial^2 l_n(\mathbf{p})}{\partial p_i \partial p_j} &= - \frac{x_k}{p_k^2}, \qquad i \neq j
\end{aligned}$$

The second derivative matrix is rather messy. It is in fact negative definite, but we won't go through the details of showing this. Hence the log likelihood is strictly concave and the unique global maximum is found by setting the first derivative equal to zero and solving for the parameter vector $\mathbf{p}$. This gives us $k-1$ equations

$$\frac{x_i}{p_i} = \frac{x_k}{p_k} = \frac{x_k}{1 - p_1 + \cdots + p_{k-1}}$$

in the $k-1$ unknowns p_1, ..., p_{k-1}. It turns out that the expression on the right hand side here is not helpful. Rewriting the left hand equality gives

$$p_i = \frac{x_i p_k}{x_k}, \qquad i = 1, \ldots, k-1. \tag{10.61}$$

Now use the fact that probabilities sum to one and the x_i sum to n (10.30) and (10.32)

$$\begin{aligned}
1 &= \sum_{i=1}^{k} p_i \\
&= p_k + \sum_{i=1}^{k-1} \frac{x_i p_k}{x_k} \\
&= p_k + \frac{p_k}{x_k} \sum_{i=1}^{k-1} x_i \\
&= p_k + \frac{p_k}{x_k}(n - x_k) \\
&= \frac{n p_k}{x_k}
\end{aligned}$$

Solving for p_k gives

$$p_k = \frac{x_k}{n}$$

and plugging this back into (10.61) gives

$$p_i = \frac{x_i}{n}, \qquad i = 1, \ldots, k-1.$$

Putting both together gives $\hat{p}_i = x_i/n$ for all i, so that is the MLE in the big model.

Now we are ready to do the likelihood ratio test,

```
> x <- c(1997, 904, 906, 32)
> p.hat <- x / sum(x)
> theta <- 0.0357123
> p.star <- c((2 + theta) / 4, (1 - theta) / 4,
+    (1 - theta) / 4, theta / 4)
> l <- function(p) sum(x * log(p))
> lrt <- 2 * (l(p.hat) - l(p.star))
> 1 - pchisq(lrt, 2)
[1] 0.3644519
```

The MLE in the little model ($\hat{\theta} = 0.0357123$) was found in Example 10.3.9. The P-value for $P = 0.36$ shows no significant lack of fit of the small model (that is, we accept H_0 which is the small model, and this is tantamount to saying it fits the data well).

10.5 Change of Parameters

This section is more careful than the proceeding ones. It is so formal that it may be hard to see the forest for the trees. Hence before starting we present the "cartoon guide," which consists of two simple ideas

- a change of parameters does not affect likelihood inference, except that
- in calculating Fisher information some extra terms arise from the chain rule.

10.5.1 Invariance of Likelihood

What happens to the likelihood and log likelihood under a change of parameters? The answer to this question seems so obvious, that we have done the right thing in one of the preceding examples without making a point of it: in Example 10.4.1 we changed parameters from the standard deviation σ to the variance $\varphi = \sigma^2$. Here is an even simpler example.

Example 10.5.1 (Exponential Model).
Suppose X_1, X_2, ... are i. i. d. exponential. If we take the parameter to be the usual parameter λ, the likelihood was figured out in Problem 7-38(b) in Lindgren

$$\begin{aligned} L_n(\lambda) &= \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right) \\ &= \lambda^n e^{-\lambda n \bar{x}_n} \end{aligned}$$

So the log likelihood is

$$l_n(\lambda) = n \log(\lambda) - \lambda n \bar{x}_n \tag{10.62}$$

Now suppose that we are not interested in the parameter λ but in the mean $\mu = 1/\lambda$. Then what is the log likelihood? The answer is that we do the obvious: plug $\lambda = 1/\mu$ into (10.62). Giving

$$\tilde{l}_n(\mu) = -n \log(\mu) - \frac{n \bar{x}_n}{\mu} \tag{10.63}$$

for the log likelihood in this parameterization.

Because we are being very careful in this section, we have used different notation for the log likelihood. Recognizing that (10.62) and (10.63) define different functions, we denote one l_n and the other $\tilde{l}_n$. Formally, if we have two parameterizations related by an invertible transformation $\varphi = g(\theta)$ and $\theta = g^{-1}(\varphi)$, then the two log likelihoods are related by *invariance*

$$\tilde{l}_n(\varphi) = l_n(\theta), \qquad \text{when } \varphi = g(\theta). \tag{10.64}$$

The log likelihood has the same values at parameters representing the same probability distribution. In order to clearly show the effect of the change of parameter, we need to plug the condition in (10.64) into the invariance relation giving

$$\tilde{l}_n[g(\theta)] = l_n(\theta) \tag{10.65}$$

This clearly shows that $\tilde{l}_n$ and l_n are not the same function. Note that if we write $h = g^{-1}$ then an equivalent way to write (10.65) is

$$\tilde{l}_n(\varphi) = l_n[h(\varphi)]. \tag{10.66}$$

Also note that exactly the same formulas would hold in the multiparameter case, except that we would use some boldface type.

10.5.2 Invariance of the MLE

What happens to maximum likelihood estimates under a change of parameters?

Theorem 10.7 (Invariance of MLE's). *Suppose that $\varphi = g(\theta)$ is an invertible change of parameter. If $\hat{\theta}$ is the MLE for θ, then $\hat{\varphi} = g(\hat{\theta})$ is the MLE for φ.*

This is obvious from (10.65). If $\hat{\theta}_n$ maximizes l_n, then $\hat{\varphi}_n = g(\hat{\theta}_n)$ maximizes $\tilde{l}_n$. And vice versa: if $\hat{\varphi}_n$ maximizes $\tilde{l}_n$, then $\hat{\theta}_n = g^{-1}(\hat{\varphi}_n)$ maximizes l_n.

This theorem on invariance of maximum likelihood estimates seems obvious, and it is, but it shouldn't be ignored on that account. Other estimates do not possess such an invariance property. Method of moments estimates don't (at least not necessarily), and, as we shall see when we get to them, Bayes estimates don't either.

Example 10.5.2 (Exponential Model).
In Problem 7-42(b) in Lindgren we found $\hat{\lambda}_n = 1/\bar{x}_n$ as the MLE for the i. i. d. $\text{Exp}(\lambda)$ model. By the theorem, $1/\hat{\lambda}_n = \bar{x}_n$ is the MLE of the parameter $\mu = 1/\lambda$. We don't have to maximize (10.63) to find the MLE. We can get it from the theorem.

Example 10.5.3 (Normal Location-Scale Model).
In Example 10.4.1 we found $\hat{\varphi}_n = v_n$ for the MLE of the variance $\varphi = \sigma^2$. By the invariance theorem, the MLE of the standard deviation $\sigma = \sqrt{\varphi}$ is $\hat{\sigma}_n = \sqrt{v_n}$.

We also make the same remark here as in the preceding section, that exactly the same phenomenon holds in the multiparameter case. The formulas would even be exactly the same, except for some boldface type.

10.5.3 Invariance of Likelihood Ratio Tests

Theorem 10.8 (Invariance of the Likelihood Ratio Test). *The likelihood ratio test statistic* (10.57) *or* (10.58) *is unchanged by an invertible change of parameter.*

If g is an invertible change of parameter and $\hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}^*_n$ are the MLE's in the big model and the little model, respectively, then by Theorem 10.7 $\hat{\boldsymbol{\varphi}}_n = g(\hat{\boldsymbol{\theta}}_n)$ and $\boldsymbol{\varphi}^*_n = g(\boldsymbol{\theta}^*_n)$ are the MLE's in the transformed coordinates, and Theorem 10.8 asserts

$$2[l_n(\hat{\boldsymbol{\theta}}_n) - l_n(\boldsymbol{\theta}^*_n)] = 2[l_n(\hat{\boldsymbol{\varphi}}_n) - l_n(\boldsymbol{\varphi}^*_n)].$$

This is clear from the invariance of likelihood (10.64).

Again this seems obvious, and it is, but it is an important property not shared by other forms of inference. The value of the likelihood ratio test statistic, and hence the P-value for the test, does not depend on the parameterization.

10.5.4 Covariance of Fisher Information

What happens to observed and expected Fisher information under a change of parameters is a bit more complicated.

Theorem 10.9. *Suppose that $\boldsymbol{\varphi} = \mathbf{g}(\boldsymbol{\theta})$ is an invertible change of parameter with differentiable inverse $\boldsymbol{\theta} = \mathbf{h}(\boldsymbol{\varphi})$, and write*

$$\mathbf{H}(\boldsymbol{\varphi}) = \nabla \mathbf{h}(\boldsymbol{\varphi}).$$

Then $\mathbf{I}_n(\boldsymbol{\theta})$, the expected Fisher information for $\boldsymbol{\theta}$, and $\widetilde{\mathbf{I}}_n(\boldsymbol{\varphi})$, the expected Fisher information for $\boldsymbol{\varphi}$, are related by

$$\widetilde{\mathbf{I}}_n(\boldsymbol{\varphi}) = \mathbf{H}(\boldsymbol{\varphi}) \cdot \mathbf{I}_n[\mathbf{h}(\boldsymbol{\varphi})] \cdot \mathbf{H}(\boldsymbol{\varphi})'. \tag{10.67}$$

If $\hat{\boldsymbol{\theta}}_n = \mathbf{h}(\hat{\boldsymbol{\varphi}}_n)$ is an interior point of the parameter space, then $\mathbf{J}_n(\hat{\boldsymbol{\theta}}_n)$, the observed Fisher information for $\boldsymbol{\theta}$ evaluated at the MLE, and $\widetilde{\mathbf{J}}_n(\hat{\boldsymbol{\varphi}}_n)$, the observed Fisher information for $\boldsymbol{\varphi}$ evaluated at the MLE, are related by

$$\widetilde{\mathbf{J}}_n(\hat{\boldsymbol{\varphi}}_n) = \mathbf{H}(\hat{\boldsymbol{\varphi}}_n) \cdot \mathbf{J}_n(\hat{\boldsymbol{\theta}}_n) \cdot \mathbf{H}(\hat{\boldsymbol{\varphi}}_n)'. \tag{10.68}$$

Each of the terms in (10.67) and (10.68) is an $m \times m$ matrix if there are m parameters. Hence this theorem is not so easy to use, and we won't give any examples.

The uniparameter case is much simpler, and we record as in explicit corollary.

Corollary 10.10. *Suppose that $\varphi = g(\theta)$ is an invertible change of parameter with differentiable inverse $\theta = h(\varphi)$. Then $I_n(\theta)$, the expected Fisher information for θ, and $\widetilde{I}_n(\varphi)$, the expected Fisher information for φ, are related by*

$$\widetilde{I}_n(\varphi) = I_n[h(\varphi)] \cdot [h'(\varphi)]^2 \tag{10.69}$$

If $\hat{\theta}_n = h(\hat{\varphi}_n)$ is an interior point of the parameter space, then of the parameter space, then $J_n(\hat{\theta}_n)$, the observed Fisher information for θ evaluated at the MLE, and $\widetilde{J}_n(\hat{\varphi}_n)$, the observed Fisher information for φ evaluated at the MLE, are related by

$$\widetilde{J}_n(\hat{\varphi}_n) = J_n(\hat{\theta}_n) \cdot [h'(\hat{\varphi}_n)]^2 \tag{10.70}$$

Note the difference between (10.67) and (10.68). The transformation rule for expected Fisher information holds for all parameter values. The transformation rule for observed Fisher information holds only when it is evaluated at the MLE and the MLE is in the interior of the parameter space, not on the boundary.

Example 10.5.4 (Exponential Model).
Consider again the Exp(λ) model we looked at in Examples 10.5.1 and 10.5.2. In those examples we found the MLE's of λ and $\mu = 1/\lambda$ to be $\hat{\mu}_n = \bar{x}_n$ and $\hat{\lambda}_n = 1/\bar{x}_n$.

We also found in Problem 10-1(a) that the Fisher information for λ is

$$I_n(\lambda) = \frac{n}{\lambda^2}$$

Let us apply the corollary to find the Fisher information for μ.

The inverse transformation is

$$\lambda = h(\mu) = \frac{1}{\mu}$$

and the derivative is

$$h'(\mu) = -\frac{1}{\mu^2}$$

Thus (10.69) gives

$$\begin{aligned}
\widetilde{I}_n(\mu) &= I_n[h(\mu)] \cdot [h'(\mu)]^2 \\
&= I_n(1/\mu) \cdot \left(-\frac{1}{\mu^2}\right)^2 \\
&= \frac{1}{(1/\mu)^2} \cdot \frac{1}{\mu^4} \\
&= \frac{1}{\mu^2}
\end{aligned}$$

Thus we get for the asymptotic distribution

$$\hat{\mu}_n \approx \mathcal{N}\left(\mu, \frac{\mu^2}{n}\right)$$

Of course, we didn't need to do this calculation to find the asymptotic distribution. Since $\hat{mu}_n = \bar{x}_n$ the CLT gives its asymptotic distribution directly

$$\overline{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

We see that these are indeed the same because we know that for the exponential distribution $\mu = 1/\lambda$ and $\sigma^2 = 1/\lambda^2$ so $\sigma^2 = \mu^2$.

Problems

10-1. Suppose X_1, X_2, ... are i. i. d. Exp(λ). We found in Problem 7-42(b) in Lindgren that the MLE is $\hat{\lambda}_n = 1/\bar{x}_n$.

(a) Find the asymptotic distribution of $\hat{\lambda}_n$ using expected Fisher information, and check that this gives the same answer as the delta method (which was done in Example 8.2.1 in these notes).

(b) Find an asymptotic 95% confidence interval for λ, again using Fisher information (either observed or expected, your choice).

10-2. Suppose (X_i, Y_i), $i = 1$, ..., n are i. i. d. with joint density

$$f(x, y) = e^{-\theta x - y/\theta}, \qquad x > 0,\ y > 0.$$

(a) Find the MLE of θ.

(b) Find the observed and expected Fisher information (both) and asymptotic standard errors for the MLE based on each. Are they the same?

10-3. Let X_1, X_2, ..., X_n be an i. i. d. sample from a model having densities

$$f_\theta(x) = (\theta - 1)x^{-\theta}, \qquad 1 < x < \infty,$$

where $\theta > 1$ is an unknown parameter.

(a) Find the MLE of θ and prove that it is the global maximizer of the likelihood.

(b) Find the expected Fisher information for θ.

(c) Give an asymptotic 95% confidence interval for θ.

(d) Show that

$$\hat{\theta}_n = \frac{2\overline{X}_n - 1}{\overline{X}_n - 1}$$

is a method of moments estimator of θ.

(e) Use the delta method to calculate the asymptotic distribution of this method of moments estimator.

(f) Calculate the ARE for the two estimators.

10-4. Show that for data that are i. i. d. $\mathcal{U}(0, \theta)$ the MLE is $\hat{\theta}_n = X_{(n)}$, the maximum data value, the asymptotic distribution of which was found in Problem 7-7. (**Hint:** Careful! The solution involves the boundary of the sample space. Also Example 8.6a in Lindgren is similar and may give you some ideas.)

This shows that Theorem 10.3 doesn't always hold. The asymptotics

$$n\bigl(\theta - X_{(n)}\bigr) \xrightarrow{\mathcal{D}} \text{Exp}(1/\theta)$$

found in Problem 7-7 don't even remotely resemble the "usual asymptotics" of maximum likelihood.

10-5. Suppose x_1, ..., x_n are known numbers (not random), and we observe random variables Y_1, ..., Y_n that are independent but *not* identically distributed random variables having distributions

$$Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2),$$

where α, β, and σ^2 are unknown parameters.

(a) Write down the log likelihood for the parameters α, β, and $\varphi = \sigma^2$.

(b) Find the maximum likelihood estimates of these parameters.

(c) Find the expected Fisher information matrix for these parameters.

(**Caution:** In taking expectations remember only the Y_i are random. The x_i are known constants.)

10-6. Find the maximum likelihood estimates for the two-parameter gamma model with densities

$$f_{\alpha,\lambda}(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$$

This cannot be done in closed form, you will have to use R. Use the method of moments estimates derived in (9.6a) and (9.6b) in these notes the starting point supplied to the `nlm` function in R.

One way to write an R function that evaluates minus the log likelihood for this model, assuming the data are a vector `x` is

```
l <- function(theta) {
      alpha <- theta[1]
      lambda <- theta[2]
      return(- sum(log(dgamma(x, alpha, 1 / lambda))))
  }
```

Data for the problem are at the URL

`http://www.stat.umn.edu/geyer/5102/prob10-6.dat`

It helps the `nlm` function to add the optional argument `fscale=length(x)` to give it an idea of the approximate size of the log likelihood.

(a) Find the MLE (vector) for the data at this URL.

(b) Find the observed Fisher information at the MLE and show that it is a positive definite matrix.

(c) Find asymptotic 95% confidence intervals for the parameters α and λ. (They do not need to have simultaneous coverage, that is, you need not use Bonferroni correction).

10-7. Prove Corollary 10.10 directly. (It is just the one-parameter case of Theorem 10.9, so the corollary follows trivially from the theorem, but we didn't prove the theorem. So prove the corollary without using the theorem.)

Hint: Start with (10.66) and use the chain rule.

10-8. Suppose X_1, $\ldots$, X_n are i. i. d. Poi(μ). The probability that X_i is zero is $p = e^{-\mu}$. Note that the transformation $p = g(\mu) = e^{-\mu}$ is one-to-one because the exponential function is monotone. Hence we can also consider p a parameter of this distribution. It's just not the usual one.

(a) Find the MLE for μ and for p.

(b) Find the (expected) Fisher information for μ.

(c) Find the (expected) Fisher information for p. Corollary 10.10 may be helpful.

(d) Suppose we observe data $\overline{X}_n = 5.7$ and $n = 30$. Find a 95% confidence interval for the parameter p.

Chapter 11

Bayesian Inference

A *Bayesian* is a person who treats parameters as random variables, and a "frequentist" is a person who doesn't. The "frequentist" slogan that expresses this is "parameters are unknown constants, not random variables." This is supposed to explain why Bayesian inference is wrong. But it is a cheap rhetorical trick. Bayesians think that probability theory is a way to express lack of knowledge, so they *agree* that "parameters are unknown constants" and continue with "hence we describe our uncertainty about them with a probability model."

Slogans can be tossed back and forth forever with no change in positions. To see what the argument is about, we have to learn Bayesian inference.

11.1 Parametric Models and Conditional Probability

The Bayesian notion gives us another view of conditional probability.

> *Conditional probability distributions are no different from parametric families of distributions.*

For each fixed value of y, the conditional density $f(x \mid y)$, considered as a function of x alone, is a probability density. So long as the two properties

$$f(x \mid y) \geq 0, \qquad \text{for all } x \tag{11.1a}$$

and

$$\int f(x \mid y) \, dx = 1 \tag{11.1b}$$

(with the integral replaced by a sum in the discrete case) hold for all y, then this defines a conditional probability model. There is no other requirement. We also made this point when we considered conditional probability in Chapter 3 of these notes. In fact, (11.1a) and (11.1b) just repeat (3.5a) and (3.5b) from that chapter.

Last semester most of our time spent studying conditional probability involved deriving conditional densities from joint densities. When you are doing that, there is another requirement conditional densities must satisfy: "joint equals marginal times conditional"

$$f(x, y) = f(x \mid y) f_Y(y).$$

But that's a very different issue. If we are only interested in whether a formula defines a conditional density and have no interest in whether or not this conditional density is the one associated with a particular joint density, then the only requirements are that (11.1a) and (11.1b) hold for every possible value of y. These are, of course, the same requirements that apply to *any* probability density, conditional or unconditional, that it is nonnegative and integrate or sum, as the case may be, to one.

The point of the slogan about there being no difference between conditional probability and parametric families is that parametric families must satisfy the same two properties with y replaced by θ

$$f(x \mid \theta) \geq 0, \qquad \text{for all } x \tag{11.2a}$$

and

$$\int f(x \mid \theta) \, dx = 1 \tag{11.2b}$$

A frequentist may write a probability density $f_\theta(x)$ to emphasize that θ is not a random variable, but just an adjustable constant. We used this notation ourselves in the chapters on frequentist inference (9 and 10). A Bayesian always writes $f(x \mid \theta)$ to emphasize that θ is a random variable (a Bayesian is a person who treats parameters as random variables), and the density is being considered the conditional density of the random variable X given the random variable θ.

The multivariable or multiparameter case is no different except for some boldface type (and perhaps sums and integrals become multiple too). We still say that conditional probability and parametric models are different ways of looking at the same thing, and that the only requirement for a function to be a conditional density is that it be nonnegative and integrate (or sum) to one, integrating (or summing) with respect to the variable "in front of the bar."

11.2 Prior and Posterior Distributions

11.2.1 Prior Distributions

Bayesians use the same statistical models as frequentists. If X_1, X_2, ..., X_n are i. i. d. with density $f(x \mid \theta)$, then the joint distribution of the data is

$$f(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta).$$

When this is thought of as a function of the parameter rather than the data, it becomes the likelihood

$$L_{\mathbf{x}}(\theta) = f(\mathbf{x} \mid \theta) \tag{11.3}$$

or, more generally,

$$f_{\mathbf{X}}(\mathbf{x} \mid \theta) = c(\mathbf{x}) L_{\mathbf{x}}(\theta) \tag{11.4}$$

where $c(\mathbf{x})$ is any nonzero function of $\mathbf{x}$ alone, not a function of θ. These are just (10.1) and (10.2) repeated, the same definition of likelihood as in the preceding chapter. The only difference is that we are using the Bayesian notation $f(\mathbf{x} \mid \theta)$ rather than the frequentist notation $f_\theta(\mathbf{x})$ for densities. The point is that the Bayesian thinks of both $\mathbf{X}$ and θ as random variables and thinks of the density $f(\mathbf{x} \mid \theta)$ as a conditional density of $\mathbf{x}$ given θ.

So far Bayesians and non-Bayesians agree, except for notation. They part company when the Bayesian goes on to put a probability distribution on the parameter θ. In order to specify a joint distribution for $\mathbf{X}$ and θ, we need the marginal for θ. For reasons to be explained later, this is called the *prior* distribution of θ. Since we have already used the letter f for the density of $\mathbf{x}$ given θ, we (following Lindgren) will use g for the prior density.

We should take a brief time-out for a reminder that *mathematics is invariant under changes of notation.* Up to this point random variables have always been Roman letters, never Greek letters. You may have unconsciously made this a rule. If so, you will now have to unlearn it. For Bayesians, the parameters (usually Greek letters) are also random variables.

Example 11.2.1 (Exponential Data, Gamma Prior).
Suppose the data are one observation $X \sim \operatorname{Exp}(\lambda)$ so the conditional density of the data given the parameter is

$$f(x \mid \lambda) = \lambda e^{-\lambda x}, \tag{11.5a}$$

and suppose the prior distribution for λ is $\operatorname{Gam}(a, b)$. We call a and b *hyperparameters* of the prior. We can't use the usual notation α and λ for gamma distribution parameters for the hyperparameters, at least we can't use λ, because we are already using λ for something else, the parameter of the data distribution. Although a and b are parameters (of the prior), it would be too confusing to simply call them "parameters" as in "parameters of the distribution of the parameter." Hence the term "hyperparameter" which indicates parameters "one level up" (in the prior rather than the data distribution). Bayesians treat parameters (here λ) as random variables, but not hyperparameters (here a and b). The hyperparameters are constants chosen to give a particular prior distribution.

What is the prior density of λ? We usually write the gamma density as

$$f(x \mid \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \tag{11.5b}$$

but as we already noticed we can't use α and λ as the hyperparameters because we are already using λ for the parameter. The problem says we are to use a

and b for those parameters. This means that (11.5b) becomes

$$f(x \mid a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} \tag{11.5c}$$

That is what the notation $\text{Gam}(a, b)$ means. We also have to make another change. It is not X but λ that has this distribution. This means that (11.5c) becomes

$$f(\lambda \mid a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda} \tag{11.5d}$$

That is the prior density for λ. While we are making changes using "mathematics is invariant under changes of notation" we might as well make one more, changing the name of the function from f to g because we are already using f for the function in (11.5a). This means that (11.5d) becomes

$$g(\lambda \mid a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda} \tag{11.5e}$$

These sorts of changes of notation, changing α and λ to a and b to get (11.5c), then changing x to λ to get (11.5d), then changing f to g to get (11.5e) should be easy. If they throw you, practice until they become easy. The past experience with this course is that some students never understand these trivial manipulations. Hence they can't even get started right on any Bayesian problem, and hence completely botch all of them. So a word to the wise: if you haven't understood "mathematics is invariant under changes of notation" yet, get it now.

> ***A Sanity Check:*** *The prior density for θ is a function of θ, not some other variable (like x).*

Making this simple sanity check will save you from the worst errors: using (11.5b) or (11.5c) for the prior. Not that there are no other ways to screw up. If you decide to change x to λ first, paying no attention to the fact that there already is a λ in (11.5b), you get

$$f(\lambda \mid \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\lambda^2} \tag{11.5f}$$

and the problem is now irretrievably botched. There is no way to get from (11.5f) to the right formula (11.5e) except recognizing you've goofed and starting over.

11.2.2 Posterior Distributions

The *joint* distribution of data and parameters, that is, of the pair $(\mathbf{X}, \theta)$, is the conditional times the marginal $f(\mathbf{x} \mid \theta) g(\theta)$. The next step in Bayesian inference is to produce the conditional distribution of the parameter given the data. For this Lindgren uses yet a third letter h. We know how to find a conditional given a joint distribution: conditional = joint/marginal or written out in symbols

$$h(\theta \mid \mathbf{x}) = \frac{f(\mathbf{x} \mid \theta) g(\theta)}{p(\mathbf{x})} \tag{11.6}$$

where $p(\mathbf{x})$ is the marginal of $\mathbf{X}$, that is,

$$p(\mathbf{x}) = \int f(\mathbf{x} \mid \theta) g(\theta) \, d\theta \tag{11.7}$$

if θ is a continuous random variable. Of course, if θ is discrete we use the same formula except that the integral is replaced by a sum, but applications with discrete parameter spaces are very rare. We won't consider any in this course.

The conditional distribution $h(\theta \mid \mathbf{x})$ of the parameter given the data is called the *posterior* distribution of the parameter. The idea behind the terminology is that the prior represents knowledge (or conversely uncertainty) about the parameter before the data are observed and the posterior represents knowledge about the parameter after the data are observed. That agrees with our usual notion of conditional probability: $f(x \mid y)$ is what you use for the distribution of X after you observe Y. The only novelty is applying this notation to parameters (Greek letters) rather than to data (Roman letters). To a Bayesian these are both random variables, so there is no novelty. Bayesian inference is just an application of conditional probability. The only novelty is the notion of treating parameters as random variables in the first place.

If we use (11.4) to replace the conditional density $f(\mathbf{x} \mid \theta)$ by a constant times the likelihood, we see that this does not affect the calculation of the posterior, because (11.7) becomes

$$p(\mathbf{x}) = c(\mathbf{x}) \int L_{\mathbf{x}}(\theta) g(\theta) \, d\theta \tag{11.8}$$

so plugging both (11.4) and (11.8) into (11.6) gives

$$h(\theta \mid \mathbf{x}) = \frac{L_{\mathbf{x}}(\theta) g(\theta)}{\int L_{\mathbf{x}}(\theta) g(\theta) \, d\theta} \tag{11.9}$$

the factor $c(\mathbf{x})$ that appears in both the numerator and denominator cancels. Either of the formulas (11.6) or (11.9) is commonly called *Bayes' rule* or *Bayes' theorem*. Calling it a "theorem" seems a bit much, since it is a trivial rearrangement of the definition of conditional probability density. In fact, exactly this formula was introduced in the chapter on conditional probability of these notes (Section 3.4.5) last semester. The only difference is that we used Roman letters for the variables behind the bar, and now we are going to use Greek letters. Same mathematical idea, just different notation.

In the same chapter we also introduced the notion of unnormalized probability densities (Section 3.4.2) and calculation of conditional probabilities as a renormalization process (Sections 3.4.3 and 3.4.4). If you weren't in my section first semester (or if you have forgotten this material), you should review it.

A function is called an *unnormalized density* if it is nonnegative and has a finite nonzero integral, which is called the *normalizing constant* of the function. Using this notion, a simple way to express (11.9) is to say that $L_{\mathbf{x}}(\theta)g(\theta)$, thought of as a function of θ for fixed $\mathbf{x}$, is an unnormalized posterior density. The denominator in (11.9) is its normalizing constant. Another way to say it is the pseudo-mathematical expression

$$\text{posterior} \propto \text{likelihood} \times \text{prior}. \tag{11.10}$$

The symbol $\propto$, read "proportional to," expresses the notion that the right hand side is an unnormalized density.

Similarly, it does no harm if the prior density is unnormalized. Suppose we specify the prior by an unnormalized density $g_u(\theta)$. The normalized prior is then $g(\theta) = c g_u(\theta)$, where c is a nonzero constant. Plugging this into (11.9) gives

$$h(\theta \mid \mathbf{x}) = \frac{L_{\mathbf{x}}(\theta) g_u(\theta)}{\int L_{\mathbf{x}}(\theta) g_u(\theta) \, d\theta}.$$

The factor c appears in both the numerator and denominator and cancels. This is exactly the same as (11.9) except that g is replaced by g_u. Thus it makes no difference whether or not the prior is normalized. So we could re-express our slogan (11.10) as

$$\text{posterior} \propto \text{likelihood} \times \text{possibly unnormalized prior},$$

but that seems too verbose. We'll just use (11.10) with the tacit understanding that the prior density need not be normalized.

Example 11.2.2 (Example 11.2.1 Continued).
Since the hyperparameters a and b are constants, the first factor in the (correct!) prior density (11.5e) is constant and we can drop it, giving the unnormalized prior

$$g_u(\lambda \mid a, b) = \lambda^{a-1} e^{-b\lambda}. \tag{11.11}$$

Multiplying by the data distribution gives the unnormalized posterior

$$h_u(\lambda \mid x) = \lambda e^{-\lambda x} \lambda^{a-1} e^{-b\lambda} = \lambda^{a} e^{-(b+x)\lambda}. \tag{11.12}$$

Keep in mind that the random variable here is λ, the data x is fixed (because we are conditioning on it) and so are the hyperparameters a and b.

To normalize this density, we use our favorite trick of recognizing the unnormalized density of a brand name distribution. Clearly (11.12) has the same form as (11.11). The only difference is that a has been replaced by $a + 1$ and b has been replaced by $b + x$. Thus the posterior distribution of λ given x is $\text{Gam}(a + 1, b + x)$.

That constitutes a satisfactory answer to the problem. We don't even have to write down the density to specify the posterior distribution. If for some reason we do actually need the density, we can find it from the formula for the gamma density

$$h(\lambda \mid x) = \frac{(b+x)^{a+1}}{\Gamma(a+1)} \lambda^a e^{-(b+x)\lambda}. \tag{11.13}$$

What if we hadn't thought of the trick? Plodding on using (11.9), we would see that in order to find the denominator we would have to evaluate the integral

$$\int_0^\infty \lambda^a e^{-(b+x)\lambda} \, d\lambda$$

(Note well! The variable of integration is λ not x. The variable in a Bayesian problem is the parameter. If you proceed by reflexes rather than thinking during an exam, you are liable to write dx. If you remember this warning, you won't make that mistake.) This integral is rather hard to do unless we recognize its relationship to the gamma function or just find in a book. It is equation (4) on p. 173 in Lindgren. Evaluating the integral and plugging into (11.9) gives us (11.13) again.

Doing the calculation of the integral just rederives the normalizing constant of the gamma distribution, redoing the work on p. 173 in Lindgren. The trick saves you this extra work.

Example 11.2.3 (Binomial Data, Uniform Prior).
This example is the first Bayesian analysis ever done. It was discovered by Thomas Bayes and published posthumously in 1764 and gives Bayesian inference its name. Suppose the data are $X \sim \text{Bin}(n, p)$ and our prior distribution for p is $\mathcal{U}(0, 1)$.

Then the prior is $g(p) = 1$ for $0 < p < 1$, and the likelihood is (10.3). Since the prior is identically equal to one, the likelihood is also the unnormalized posterior

$$h(p \mid x) \propto p^x (1-p)^{n-x} \tag{11.14}$$

To normalize this density, we again use our favorite trick of recognizing the unnormalized density of a brand name distribution. In this case the factors p and $(1-p)$ should recall the beta distribution,[1] which has densities of the form

$$f(x \mid s, t) = \frac{\Gamma(s+t)}{\Gamma(s)\Gamma(t)} x^{s-1} (1-x)^{t-1} \tag{11.15}$$

(p. 175 in Lindgren). Comparing (11.15) with x changed to p with (11.14), we see that they are the same except for constants if $s = x + 1$ and $t = n - x + 1$.

[1] Why not the binomial distribution? That's the one that has p and $1 - p$ in the formula! The beta distribution has x and $1 - x$. If that's what you are thinking, you have again run afoul of "mathematics is invariant under changes of notation." The letters don't matter. A binomial distribution is a binomial distribution no matter whether you call the parameter p or x, and a beta distribution is a beta distribution no matter whether you call the random variable p or x. What matters is not which letter you use, but the role it plays. Here p is the random variable, the letter in front of the bar in the conditional density $h(p \mid x)$, hence we want to find a distribution having a density with factors p and $1 - p$ where p is the *random variable*. The beta distribution is the only one we know with that property.

Thus the posterior distribution of p given x is $\text{Beta}(x + 1, n - x + 1)$. No integration is necessary if we see the trick.

If you don't see the trick, you must integrate the right hand side of (11.14) to find the normalizing constant of the posterior. Again this integral is rather hard unless you recognize it as a beta integral, equation (14) on p. 175 in Lindgren. As in the preceding example, this just redoes the work of deriving the normalizing constant of a brand name distribution. The trick is easier.

Example 11.2.4 (Normal Data, Normal Prior on the Mean).
Assume X_1, $\ldots$, X_n are i. i. d. with unknown mean μ and known variance σ^2, and assume a normal prior distribution for the unknown parameter μ. This example is not very practical, because we rarely know σ^2, but it makes a good example. Inference for the case where both μ and σ^2 are unknown will be covered in Section 11.4.3.

Let us denote the prior distribution for μ by $\mathcal{N}(\mu_0, \sigma_0^2)$. As in Example 11.2.1 we can't use μ and σ for the hyperparameters, because we are already using these letters for parameters of the data distribution. Then an unnormalized prior density is

$$g(\mu) = \exp\left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right)$$

(we can drop the constant $\sqrt{2\pi}\sigma_0$). Combining this with the likelihood (10.5) gives the unnormalized posterior

$$\begin{aligned} h_u(\mu \mid x) &= L_{\mathbf{x}}(\mu) g(\mu) \\ &= \exp\left(-\frac{n(\bar{x}_n - \mu)^2}{2\sigma^2}\right) \exp\left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right) \\ &= \exp\left(-\frac{n(\bar{x}_n - \mu)^2}{2\sigma^2} - \frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right) \end{aligned}$$

It will considerably simplify notation in the rest of the problem if we introduce $\lambda = 1/\sigma^2$ and $\lambda_0 = 1/\sigma_0^2$. The technical term for reciprocal variance is *precision*, so λ is the precision of the data, and λ_0 is the precision hyperparameter. Then the unnormalized posterior becomes

$$h_u(\mu \mid x) = \exp\left(-\frac{n\lambda}{2}(\bar{x}_n - \mu)^2 - \frac{\lambda_0}{2}(\mu - \mu_0)^2\right) \tag{11.16}$$

Since the exponent is quadratic in μ, the posterior must be some normal distribution (this is the "e to a quadratic theorem, Theorem 5.10 of Chapter 5 of last semester's notes). To see which normal distribution, we could apply the theorem, but we will just do the calculation from first principles since the theorem is multivariate and we are only interested in the univariate case here (and, to be honest, I don't want to rewrite this, which was written last year before I wrote the theorem this year). To do the calculation we compare the exponent in (11.16) with the exponent of a normal density with mean a and precision b, which is $-b(\mu - a)^2/2$. Matching coefficients of μ and μ^2 gives the posterior mean a and posterior precision b. That is, we must have

$$\begin{aligned}
&b\mu^2 - 2ab\mu + \text{a constant} \\
&\qquad = (n\lambda + \lambda_0)\mu^2 - 2(n\lambda\bar{x}_n + \lambda_0\mu_0)\mu + \text{some other constant}
\end{aligned}$$

Hence

$$\begin{aligned}
b &= n\lambda + \lambda_0 \qquad (11.17a) \\
ab &= n\lambda\bar{x}_n + \lambda_0\mu_0
\end{aligned}$$

so

$$a = \frac{n\lambda\bar{x}_n + \lambda_0\mu_0}{n\lambda + \lambda_0} \qquad (11.17b)$$

and the posterior distribution of μ is normal with mean (11.17b) and precision (11.17a).

11.3 The Subjective Bayes Philosophy

That's more or less the story on the mechanics of Bayesian inference. There are some bells and whistles that we will add later, but this is the basic story. It's just conditional probability coupled with the notion of treating parameters as random variables. For the most part the calculations are no different from those we did when we studied conditional probability last semester. If you can get used to Greek letters as random variables, the rest is straightforward.

Here we take a time out from learning mechanics to learn enough of the Bayesian philosophy to understand this chapter. The Bayesian philosophy holds that all uncertainty can be described by means of probability distributions. This has far reaching implications. For one thing, it means that, since everyone is uncertain about many things, everyone has probability distributions inside their heads. These are the prior distributions that appear in Bayesian problems. A subjective Bayesian believes that everyone has a different prior distribution for any particular problem. An objective Bayesian believes there are ways in which different people can agree on a common prior distribution (by convention if no other way). We will only explain the subjective Bayesian view.

So in any particular problem, once the probability model for the data (and hence the likelihood) is decided, one then gets a prior by "eliciting" the prior distribution that represents the knowledge (or uncertainty, depending on how you look at it) of some expert. Then you apply Bayes rule, and you're done.

> *Once agreement is reached about being Bayesian and on the likelihood and prior, Bayesian inference is straightforward.*

Frequentist inference involves many technical difficulties, choice of point estimators, test statistics, and so forth. Bayesian inference involves no such difficulties.

Every inference is a straightforward conditional probability calculation, which if not doable with pencil and paper is usually doable by computer.

Getting the agreement mentioned in the slogan may be difficult. Bayesian inference is controversial. For a century roughly from 1860 to 1960, it was considered absurd, obviously completely wrong (and many other pejorative terms were applied). Now the pendulum of fashion has swung the other way, and Bayesian inference, if not the most popular, is at least very trendy in certain circles. But it takes some people a long time to get the word. Textbooks are often decades behind research. Opinions are passed from scientist to scientist without influence by statisticians and statistics courses. So there are still a lot of people out there who think Bayesian inference is a no-no.

Agreement to be Bayesian does not end philosophical arguments. There can be arguments about the appropriateness of the probability model for the data, but exactly the same arguments would arise if one wanted to use the same model for frequentist inference, so those arguments are not peculiar to Bayesian inference. And there can be arguments about the prior. Whose prior (what expert's opinion) is used? How it is elicited? Was it elicited correctly? The elicitation problem is made more difficult (or perhaps simpler, I'm not sure) by the fact that it does not really involve getting a probability distribution from inside someone's head down on paper. All psychological study of people's behavior involving probability and statistics has revealed no evidence for, and a good deal of evidence against, the notion that real people think in accord the rules of Bayesian inference.

What do we mean by "think in accord with the rules of Bayesian inference"? We will explain that, the so-called *Bayesian model of learning*, and that will end our discussion of philosophy.

Suppose data X_1, X_2, $\ldots$ are assumed to have a probability model with likelihood

$$L_{x_1,\ldots,x_n}(\theta) = \prod_{i=1}^{n} f(x_i \mid \theta).$$

(Note that we have gone back to our original notation of indicating the data by a subscript. The reason for it will become apparent presently.) And suppose we start with a prior $g(\theta)$ that represents our knowledge (or, looked at the other way, uncertainty) about the parameter θ before any data are observed.

Suppose the data arrive over the course of time, and intermediate analyses are done as the data arrive. For simplicity, we only consider two analyses, but it will be clear that everything said here extends to multiple analyses. For concreteness, say the analyses are done after observing m and n data values X_i, respectively, with $m < n$. In the first analysis, we derive a posterior distribution

$$h(\theta \mid x_1, \ldots, x_m) \propto L_{x_1,\ldots,x_m}(\theta) g(\theta) \tag{11.18}$$

that represents our knowledge (or, looked at the other way, uncertainty) about the parameter θ at this time and reflects both the information from the prior and from the data x_1, $\ldots$, x_m.

Now we take a time out for more philosophy. The distribution (11.18) can be thought of as *both* a prior and a posterior. It is a posterior, when we consider that it describes our knowledge *after* x_1, ..., x_m have been observed. It is a prior, when we consider that it describes our knowledge *before* x_{m+1}, ..., x_n are observed. So it should serve as our prior for the subsequent analysis of those data.

The likelihood for those data is

$$L_{x_{m+1},\ldots,x_n}(\theta) = \prod_{i=m+1}^{n} f(x_i \mid \theta).$$

and the posterior after observing those data is

$$h(\theta \mid x_1, \ldots, x_n) \propto L_{x_{m+1},\ldots,x_n}(\theta) h(\theta \mid x_1, \ldots, x_m) \tag{11.19}$$

Great! So what's the point? The point is that there is another way of thinking about this problem. If we ignore the fact that the data arrived in two clumps, we would analyze the whole data set at once, using the likelihood for all the data and the original prior $g(\theta)$. This would give a formula just like (11.18) except with m replaced by n. Now there is no philosophical reason why these two procedures (two-stage analysis and one-stage analysis) should give the same answers, but it is a remarkable fact that they do. No matter how you do a Bayesian analysis, so long as you correctly apply Bayes rule, you get the same answer (starting from the same prior).

Note that frequentist inference does not have this property. There is no way to use the results of a frequentist analysis of part of the data in subsequent analyses. In fact, the very fact of having done an analysis on part of the data *changes the answer* of the analysis of the complete data, because it gives rise to a need for correction for multiple testing. What we learned here is that Bayesian inference has (and needs) no analog of frequentist correction for multiple testing. So long as you apply Bayes rule correctly, you get the correct answer.

This same issue means that Bayesian inference can serve as a model of learning, but frequentist inference can't. The Bayesian notion of learning is just the transformation from prior to posterior via Bayes rule. The prior describes your knowledge before the data are observed, the posterior your knowledge after the data are observed, the difference is what you learned from the data.

11.4 More on Prior and Posterior Distributions

11.4.1 Improper Priors

We saw in the preceding section that it does no harm to use an unnormalized prior. We can go even further and drop the requirement that the prior be integrable. If $g(\theta)$ is a nonnegative function such that

$$\int g(\theta)\, d\theta = \infty$$

but the normalizing constant

$$\int L_{\mathbf{x}}(\theta) g(\theta) \, d\theta$$

is still finite, then (11.9) still defines a probability density.

Then we say we are using an *improper prior* and sometimes we say we are using the *formal Bayes rule*. It isn't really Bayes' rule, because neither the prior nor the joint distribution of parameter and data are proper probability distributions. But we use the same equation (11.9) and so the method has the same form ("formal" here doesn't mean "dressed for a prom," it means "having the same form").

Thus what we are doing here is philosophically bogus from the subjective point of view. An improper prior cannot represent "prior opinion" because it is not a probability distribution. That's why Lindgren, for example, makes a point of deriving the results with improper priors as limits of procedures using proper priors (Problem 7-83, for example). But not all results involving improper priors can be derived as such limits, so the limiting argument really contributes nothing to our understanding of improper priors. Hence our approach will be "just do it" with no worries about consequent philosophical difficulties.

Example 11.4.1 (Normal Data, Improper Prior).
Suppose X_1, X_2, $\ldots$, X_n are i. i. d. $\mathcal{N}(\mu, \sigma^2)$ where μ is a known number (not a parameter) and σ is an unknown parameter. We need a prior distribution for the unknown parameter σ, which we take to be the improper prior[2] with density $g(\sigma) = 1$ for all σ. This is improper because

$$\int_0^\infty d\sigma = \infty.$$

The likelihood is

$$L(\sigma) = \frac{1}{\sigma^n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right\}. \tag{11.20}$$

Since we are using a flat prior (11.20) is also the unnormalized posterior.

We now want to use our trick of recognizing the density of a known model, but (11.20) isn't proportional to any of the densities in Chapter 6 in Lindgren. It turns out, however, that a change of variable gives us a known family. Define a new parameter $\lambda = 1/\sigma^2$ (precision again). Then the likelihood becomes

$$L(\lambda) = \lambda^{n/2} \exp\left\{ -\frac{\lambda}{2} \sum_{i=1}^n (x_i - \mu)^2 \right\}. \tag{11.21}$$

[2] The reader should perhaps be warned that aficionados of improper priors consider this the wrong improper prior. The "natural" improper prior is $g(\sigma) = 1/\sigma$. The reasons why, however, are too complicated to explain here, and do not generalize to other problems. So we will use this one.

There is no use of the change-of-variable theorem (Theorem 8 of Chapter 3 in Lindgren) because λ is not a random variable in the *data model* or in the *likelihood*. There the X_i are the random variables.

We do, however, need to apply the change-of-variable theorem to the prior. The inverse transformation is

$$\sigma = h(\lambda) = \lambda^{-1/2},$$

and the change-of-variable theorem says the prior for λ is

$$g_\Lambda(\lambda) = g[h(\lambda)]|h'(\lambda)|$$

where

$$|h'(\lambda)| = \left| -\frac{1}{2}\lambda^{-3/2} \right| = \tfrac{1}{2}\lambda^{-3/2} \tag{11.22}$$

Since g, the prior for σ, is identically equal to 1, the prior for λ is

$$g_\Lambda(\lambda) = \tfrac{1}{2}\lambda^{-3/2}$$

The unnormalized posterior for λ is likelihood (11.21) times prior (11.22)

$$h(\lambda \mid \mathbf{x}) \propto \lambda^{n/2-3/2} \exp\left\{ -\frac{\lambda}{2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\}.$$

Considered as a function of λ (not the x_i) the right hand side must be an unnormalized density. Using the trick of recognizing an unnormalized brand name density, we see that the posterior distribution of λ is $\text{Gam}(a, b)$ with

$$\begin{aligned} a &= \frac{n-1}{2} \\ b &= \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2 \end{aligned}$$

Since the parameters a and b of the gamma distribution must be strictly positive, we need $n > 1$ in order to have a proper posterior ($b > 0$ is satisfied automatically). This check whether the posterior is integrable is always necessary when using an improper prior (and never necessary when using a proper prior). An improper posterior (one that doesn't integrate) is nonsense.

11.4.2 Conjugate Priors

The definition of *conjugate prior family* of distributions given on p. 247 in Lindgren, is fairly cryptic. A family $\mathcal{F}$ of probability distributions is conjugate to a probability model if the posterior is in $\mathcal{F}$ whenever the prior is in $\mathcal{F}$. How does one find such a family? One trivial example is the (nonparametric) family of *all* probability distributions on the parameter space. (The posterior is a probability distribution, hence is trivially in the family of all probability distributions.)

If that were all there was to the notion of conjugate families, it would be useless concept. The idea is to find a parametric conjugate family, one we recognize. Here is a recipe for such families, so-called *natural* conjugate families. If we have independent sampling, the likelihood satisfies

$$L_{x_1,\ldots,x_{m+n}}(\theta) \propto L_{x_1,\ldots,x_m}(\theta) L_{x_{m+1},\ldots,x_{m+n}}(\theta).$$

Thus we see that we can take a likelihood with some "made up" data as the prior and the unnormalized posterior will be the likelihood for the combination of real and "made up" data. In short, likelihoods form a conjugate family, so long as we include all sample sizes and all possible data values.

Usually we take a slightly larger family. If the likelihood is a well-defined positive function for noninteger values of the sample size, then we allow noninteger values. Similarly if the data are discrete, we also allow arbitrary data values so long as the resulting function is still well-defined and positive. It is clear that the result is a conjugate family so long as our possible "made up" data includes all possible actual data values.

Example 11.4.2 (Example 11.2.4 Continued).
In Example 11.2.4, we found a normal prior for μ resulted in a normal posterior. Thus family of normal distributions for the parameter μ is a conjugate prior family for a normal data model when the variance is known and μ is the only parameter.

11.4.3 The Two-Parameter Normal Distribution

The two-parameter normal model has data X_1, X_2, ... i. i. d. $\mathcal{N}(\mu, \sigma^2)$ with both μ and σ considered parameters. The likelihood is given by (10.4). As in Examples 11.2.4 and 11.4.1, the analysis becomes simpler if we use precision $\lambda = 1/\sigma^2$ as one of the parameters, giving

$$\begin{aligned} L_{\mathbf{x}}(\mu, \lambda) &= \lambda^{n/2} \exp\left\{-\tfrac{n}{2}\lambda[v_n + (\bar{x}_n - \mu)^2]\right\} \\ &= \lambda^{n/2} \exp\left\{-\tfrac{n}{2}\lambda[v_n + \bar{x}_n^2 - 2\bar{x}_n\mu + \mu^2]\right\} \end{aligned} \tag{11.23}$$

This has three bits of "made up data" to adjust: n, v_n, and $\bar{x}_n$. Replacing them with Greek letters α, β, and γ gives a conjugate family of priors with unnormalized densities

$$g(\mu, \lambda \mid \alpha, \beta, \gamma) = \lambda^{\alpha/2} \exp\left\{-\tfrac{1}{2}\alpha\lambda(\beta - 2\gamma\mu + \mu^2)\right\}. \tag{11.24}$$

Here α, β, and γ are hyperparameters of the prior, known constants, not random variables. Choosing the hyperparameters chooses a particular prior from the conjugate family to represent prior opinion about the parameters μ and λ.

The next task is to figure out the properties of the conjugate family we just discovered. With a little work we will be able to "factor" these distributions as joint = conditional × marginal and recognize the marginals and conditionals.

The conditional of μ given λ is clearly a normal distribution, because it is "e to a quadratic" (as a function of μ for fixed λ). To figure out which normal, we have to match up coefficients of powers of μ in the exponential. If $\mu \mid \lambda \sim \mathcal{N}(a, b)$, then we must have

$$\begin{aligned}(\mu - a)^2/b &= a^2/b - 2a\mu/b + \mu^2/b \\ &= \alpha\lambda(\beta - 2\gamma\mu + \mu^2) + \text{a constant}\end{aligned}$$

hence we can determine a and b by matching the coefficients of μ and μ^2 giving

$$a = \gamma \tag{11.25a}$$

$$b = \frac{1}{\alpha\lambda} \tag{11.25b}$$

To figure out the marginal of λ we have to do the "factorization" into conditional and marginal. The conditional $\mathcal{N}(a, b)$ with a and b given by (11.25a) and (11.25b) has density proportional to

$$b^{-1/2} \exp\left\{-\frac{1}{2b}(\mu - a)^2\right\} = \alpha^{1/2}\lambda^{1/2} \exp\left\{-\frac{1}{2}\alpha\lambda\left(\gamma^2 - 2\gamma\mu + \mu^2\right)\right\} \tag{11.26}$$

Thus the marginal of λ must have density proportional to (11.24) divided by (11.26), that is,

$$\lambda^{(\alpha-1)/2} \exp\left\{-\tfrac{1}{2}\alpha\lambda(\beta - \gamma^2)\right\}.$$

This is clearly proportional to a $\operatorname{Gam}(c, d)$ density with

$$c = (\alpha + 1)/2 \tag{11.27a}$$

$$d = \alpha\left(\beta - \gamma^2\right)/2 \tag{11.27b}$$

Thus we have discovered that our conjugate family can be "factored" as a product of normal and gamma distributions. The connection between the shape parameter $(\alpha + 1)/2$ of the gamma and the precision $\alpha\lambda$ of the normal seems arbitrary. Thus one usually allows the two to be varied independently, which gives a family with four hyperparameters.

Definition 11.4.1 (The Normal-Gamma Family of Distributions). *If a random variable X has a* $\operatorname{Gam}(\alpha, \beta)$ *distribution, and the conditional distribution of another random variable Y given $X = x$ is a* $\mathcal{N}(\gamma, \delta^{-1}x^{-1})$ *distribution, then we say the joint distribution of X and Y is* normal-gamma *with parameters α, β, γ, and δ. The parameter γ can be any real number, the rest must be strictly positive.*

Following the usual practice of making random variables Roman letters near the end of the alphabet, we have changed (λ, μ) to (X, Y) for this definition only. As we continue the Bayesian analysis we will go back to having the random variables being λ and μ. We have also redefined α, β, and γ and will no longer use the parameterization (11.24) for the normal-gamma family. The new parameterization given in the definition is standard.

Theorem 11.1. *If X_1, X_2, …, X_n are i. i. d. $\mathcal{N}(\mu, \lambda^{-1})$, then the normal-gamma family is a conjugate family of prior distributions for (λ, μ). If the prior distribution is*

$$\begin{aligned} \lambda &\sim \text{Gam}(\alpha_0, \beta_0) \\ \mu \mid \lambda &\sim \mathcal{N}(\gamma_0, \delta_0^{-1} \lambda^{-1}) \end{aligned}$$

then the posterior distribution is

$$\begin{aligned} \lambda &\sim \text{Gam}(\alpha_1, \beta_1) \\ \mu \mid \lambda &\sim \mathcal{N}(\gamma_1, \delta_1^{-1} \lambda^{-1}) \end{aligned}$$

where

$$\alpha_1 = \alpha_0 + \frac{n}{2} \tag{11.28a}$$

$$\beta_1 = \beta_0 + \frac{n}{2} \left(v_n + \frac{\delta_0 (\bar{x}_n - \gamma_0)^2}{\delta_0 + n} \right) \tag{11.28b}$$

$$\gamma_1 = \frac{\gamma_0 \delta_0 + n \bar{x}_n}{\delta_0 + n} \tag{11.28c}$$

$$\delta_1 = \delta_0 + n \tag{11.28d}$$

where $\bar{x}_n$ is the empirical mean and v_n is the empirical variance.

Proof. If (λ, μ) is normal-gamma with parameters α, β, γ, and δ, the unnormalized density is

$$\lambda^{\alpha - 1} \exp\{-\beta \lambda\} \cdot \lambda^{1/2} \exp\left\{ -\tfrac{1}{2} \delta \lambda (\mu - \gamma)^2 \right\}. \tag{11.29}$$

Putting subscripts of zero on the hyperparameters in (11.29) and multiplying by the likelihood (11.23) gives the unnormalized posterior

$$\lambda^{\alpha_0 + n/2 - 1/2} \exp\left\{ -\beta_0 \lambda - \tfrac{1}{2} \delta_0 \lambda (\mu - \gamma_0)^2 - \tfrac{n}{2} \lambda (v_n + \bar{x}_n^2 - 2 \bar{x}_n \mu + \mu^2) \right\} \tag{11.30}$$

To prove the theorem we have to show that this is equal to (11.29) with subscripts of one on the hyperparameters and that the relationship between the hyperparameters of prior and posterior is the one stated.

Comparing the exponent of λ in (11.29) and (11.30) gives (11.28a). The other three relationships between hyperparameters are found by equating the coefficients of λ, of $\lambda \mu$, and of $\lambda \mu^2$ in the exponential terms, which gives

$$-\beta_1 - \tfrac{1}{2} \delta_1 \gamma_1^2 = -\beta_0 - \tfrac{1}{2} \delta_0 \gamma_0^2 - \tfrac{n}{2} (v_n + \bar{x}_n^2) \tag{11.31a}$$

$$\gamma_1 \delta_1 = \gamma_0 \delta_0 + n \bar{x}_n \tag{11.31b}$$

$$-\tfrac{1}{2} \delta_1 = -\tfrac{1}{2} \delta_0 - \tfrac{n}{2} \tag{11.31c}$$

Equation (11.31c) immediately implies (11.28d). Plugging (11.28d) into (11.31b) gives (11.28c). Plugging (11.28c) and (11.28d) into (11.31a) gives

$$\beta_1 = \beta_0 + \tfrac{1}{2} \delta_0 \gamma_0^2 + \tfrac{n}{2} (v_n + \bar{x}_n^2) - \tfrac{1}{2} \frac{(\gamma_0 \delta_0 + n \bar{x}_n)^2}{\delta_0 + n}$$

which with a bit of formula manipulation is (11.28b). □

We also want to learn about the other "factorization" of the normal-gamma family into the marginal of μ times the conditional of λ given μ. This involves Student's t-distribution with noninteger degrees of freedom (Definition 7.3.2 in these notes).

Theorem 11.2. *If*

$$\begin{aligned} \lambda &\sim \operatorname{Gam}(\alpha, \beta) \\ \mu \mid \lambda &\sim \mathcal{N}(\gamma, \delta^{-1}\lambda^{-1}) \end{aligned}$$

then

$$\begin{aligned} (\mu - \gamma)/d &\sim t(\nu) \\ \lambda \mid \mu &\sim \operatorname{Gam}(a, b) \end{aligned}$$

where

$$\begin{aligned} a &= \alpha + \tfrac{1}{2} \\ b &= \beta + \tfrac{1}{2}\delta(\mu - \gamma)^2 \\ \nu &= 2\alpha \\ d &= \sqrt{\frac{\beta}{\alpha\delta}} \end{aligned}$$

Proof. The unnormalized joint density for μ and λ is given by (11.29). The conditional distribution of λ given μ is clearly the $\operatorname{Gam}(a, b)$ asserted by the theorem. This has density

$$\frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}. \tag{11.32}$$

We may ignore the factor $\Gamma(a)$, which is a constant, but we must keep b^a, which contains μ. The unnormalized marginal for μ is thus (11.29) divided by (11.32), which is b^{-a} or

$$h(\mu) = \frac{1}{[\beta + \frac{1}{2}\delta(\mu - \gamma)^2]^{\alpha + 1/2}} \tag{11.33}$$

Hence we see that some linear function of μ has a t distribution with $\nu = 2\alpha$ degrees of freedom. To determine the linear function we must equate coefficients of powers of μ in

$$\beta + \tfrac{1}{2}\delta(\mu - \gamma)^2 = k\left(1 + \frac{(\mu - c)^2}{d^2\nu}\right)$$

which is derived by plugging in $(\mu - c)/d$ for x in (7.32) and matching the $1 + x^2/\nu$ term in the denominator to the term in square brackets in (11.33). In order for $(\mu - c)/d$ to have a $t(\nu)$ distribution, these two terms must be proportional, hence equal when multiplied by k, a yet to be determined constant of proportionality.

Hence

$$\begin{aligned} \beta + \frac{\delta \gamma^2}{2} &= k + \frac{k c^2}{d^2 \nu} \\ - \delta \gamma &= - \frac{2 k c}{d^2 \nu} \\ \frac{\delta}{2} &= \frac{k}{d^2 \nu} \end{aligned}$$

Solving for k, c, and d we get $c = \gamma$, $k = \beta$, and

$$d^2 = \frac{2 \beta}{\delta \nu} = \frac{\beta}{\delta \alpha}$$

which finishes the proof of the theorem. □

Example 11.4.3 (Improper Prior for the Normal).
Bayesian inference for the two-parameter normal is quite complicated. Choosing a conjugate prior involves specifying four hyperparameters. The hyperparameters of the posterior are complicated functions of the hyperparameters of the prior and the data.

Here we analyze a simple case. Choose $\beta = \delta = 0$ in (11.29). Then the unnormalized prior is just $\lambda^{\alpha - 1/2}$. This is, of course, an improper prior. The posterior is normal-gamma with parameters

$$\begin{aligned} \alpha_1 &= \alpha + \frac{n}{2} \\ \beta_1 &= \frac{n v_n}{2} \\ \gamma_1 &= \bar{x}_n \\ \delta_1 &= n \end{aligned}$$

and is a proper distribution so long as $\alpha > -n/2$.

The posterior marginal distribution of λ is $\text{Gam}(\alpha_1, \beta_1)$ and the posterior marginal distribution of $(\mu - \bar{x}_n)/\sqrt{v_n/\nu}$ is $t(\nu)$, where $\nu = 2\alpha_1 = n + 2\alpha$.

Suppose we decide on the value $\alpha = -\frac{1}{2}$ for the remaining hyperparameter. Then $\nu = n - 1$. And since $v_n/(n-1) = s_n^2/n$, we get a marginal posterior distribution

$$\frac{\mu - \bar{x}_n}{s_n/\sqrt{n}} \sim t(n-1)$$

Thus for this particular improper prior, the marginal posterior distribution of this quantity agrees with its sampling distribution.

The agreement of Bayesian posterior and frequentist sampling distributions leads to numerically identical though philosophically different inferences. But no great message should be read into this. No Bayesian with a proper prior would get the "same" inference as the frequentist. A Bayesian with a subjective prior would not get the same inference, because subjective priors representing prior knowledge about the parameters are supposed to be proper.

11.5 Bayesian Point Estimates

Most Bayesians are not much interested in point estimates of parameters. To them a parameter is a random variable, and what is important is its distribution. A point estimate is a meager bit of information as compared, for example, to a plot of the posterior density.

Frequentists too are not much interested in point estimates for their own sake, but frequentists find many uses for point estimates as tools for constructing tests and confidence intervals. All asymptotic arguments start with calculating the asymptotic distribution of some point estimate. They may also require a point estimate of the asymptotic standard deviation for use in the "plug-in" theorem. Bayesians do not need point estimates for any of these purposes. All Bayesian inference starts with calculating the posterior distribution. To go from there to a point estimate is to throw away most of the information contained in the posterior.

Still, point estimates are easy to calculate (some of them, at least) and easy to discuss. So they are worth some study. Bayesians use three main kinds of point estimates: the posterior mean, median, and mode. The first two we have already met.

Definition 11.5.1 (Mode).
A mode *of a random variable having a continuous density is a local maximum of the density. The variable is* unimodal *if it has one mode,* bimodal *if two, and* multimodal *if more than one.*

When we say the *mode (rather than* a *mode) in reference to a multimodal distribution, we mean the highest mode (if one is higher than the others).*

All of the brand name continuous distributions introduced in Chapter 6 in Lindgren are unimodal. The normal distribution is unimodal, and the mode is the mean. In fact this is obviously true (draw a picture) for any symmetric unimodal distribution.

> *For a symmetric unimodal distribution, the mean (if it exists), the median, and the mode are all equal to the center of symmetry.*

The gamma distribution, and its special cases the exponential and chi-square, are not symmetric, but are unimodal. For them the mean, median, and mode are three different points.

Example 11.5.1 (Mode of the Gamma Distribution).
It does not matter if we use an unnormalized density. Multiplying by a constant changes only the vertical scale, not the position of the mode. An unnormalized $\text{Gam}(\alpha, \lambda)$ density is

$$f(x) = x^{\alpha-1} e^{-\lambda x}.$$

As in maximum likelihood, it is often easier to maximize the log density, which must have the same mode. The log density is

$$g(x) = \log f(x) = (\alpha - 1)\log(x) - \lambda x$$

Differentiating gives

$$g'(x) = \frac{\alpha - 1}{x} - \lambda \tag{11.34}$$

$$g''(x) = -\frac{\alpha - 1}{x^2} \tag{11.35}$$

From (11.35) we see that the log density is strictly concave, hence the local maximum is unique if it exists. Solving $g'(x) = 0$, we get

$$x = (\alpha - 1)/\lambda \tag{11.36}$$

for the mode when $\alpha \geq 1$. When $\alpha < 1$ (11.36) is negative and hence not in the sample space. In that case (11.34) is negative for all x, hence $g(x)$ is strictly decreasing and the only local maximum occurs at $x = 0$. The mode is a bit weird because $f(x) \to \infty$ as $x \to 0$ in this case. But we still call it the mode.

Frequentists are not much interested in the mode as a point estimate of location, because it is very hard to estimate and may be far from the main mass of the distribution, even when the distribution is unimodal (but not symmetric). For example, consider the $\text{Exp}(\lambda)$ distribution. The mean is $1/\lambda$, the median is $\log(2)/\lambda = 0.693/\lambda$ (Problem 6-47 in Lindgren), and the mode is zero.

Bayesians are interested in the posterior mode because of its analogy to maximum likelihood. As we saw in the preceding example, it does not matter if we use an unnormalized objective function in determining the mode, since normalization only changes the vertical scale and does not change the position of the mode. An unnormalized posterior is likelihood times prior. Thus we find the posterior mode by maximizing $L_{\mathbf{x}}(\theta)g(\theta)$, considered as a function of θ for fixed x. If we use a flat prior, this is the same as maximum likelihood. If we do not use a flat prior, then the posterior mode will be different from the MLE. But in either case the posterior mode can be calculated directly from the unnormalized posterior $L_{\mathbf{x}}(\theta)g(\theta)$. There is no need to calculate the normalizing constant, integral in (11.9), if all we want is the posterior mode.

Example 11.5.2 (Example 11.2.1 and Example 11.2.2 Continued).
In Example 11.2.2 we found a $\text{Gam}(a + 1, b + x)$ posterior distribution, where x was the data and a and b hyperparameters of the prior. The mode of this distribution, hence the posterior mode is given by (11.36) with $a + 1$ plugged in for α and $b + x$ plugged in for λ. Hence the posterior mode is

$$\lambda^* = \frac{a}{b + x}$$

For comparison the MLE is

$$\hat{\lambda} = \frac{1}{x}$$

and this is the posterior mode for a flat prior (rather than the gamma prior used in Examples 11.2.1 and 11.2.2). The posterior mean is

$$E(\lambda \mid x) = \frac{a + 1}{b + x}$$

The posterior median is hard to calculate. There is no closed form expression for it as a function of a, b, and x. For any fixed values of a, b, and x, we could use tables of the incomplete gamma function (not in Lindgren but in reference books) or a computer statistics package to calculate the posterior median, but we cannot exhibit a formula like those for the other estimators.

11.6 Highest Posterior Density Regions

This section covers the Bayesian competitor to confidence intervals, which go under the name "highest posterior density regions." A *highest posterior density (HPD) region* is a level set of the posterior, that is a set of the form

$$\{\, \theta : h(\theta \mid \mathbf{x}) > \alpha \,\}$$

for some $\alpha > 0$, that has a specified posterior probability, i. e.,

$$P\{h(\theta \mid \mathbf{X}) > \alpha \mid \mathbf{X} = \mathbf{x}\} = \beta$$

Note that, *as always when we are being Bayesians,* we are conditioning on the data $\mathbf{X}$, what is random here is the parameter θ. The idea behind the HPD region is that all of the points included in the region should be more probable (in the sense of higher posterior density) than those not in the region.

Example 11.6.1 (Examples 11.2.4 and 11.4.2 Continued).
In Example 11.2.4 we saw that if the data are i. i. d. normal with mean μ and precision λ with λ known and the prior for μ was normal with mean μ_0 and precision λ_0, then the posterior is normal with mean (11.17b) and precision (11.17a). By the symmetry of the normal distribution, the 95% HPD region is a symmetric interval centered at the posterior mean. The same logic we use to figure out critical values for confidence intervals tells us the half width of the interval is 1.96 posterior standard deviations, that is, the 95% HPD region for μ is

$$\frac{n\lambda\bar{x}_n + \lambda_0\mu_0}{n\lambda + \lambda_0} \pm 1.96\sqrt{\frac{1}{n\lambda + \lambda_0}}$$

(recall that $n\lambda + \lambda_0$ is the *precision* not the variance, so the standard deviation is the square root of its reciprocal).

Comparing this with the frequentist 95% confidence interval, which is

$$\bar{x}_n \pm 1.96\sqrt{\frac{1}{n\lambda}}$$

(recall that $\sigma^2 = 1/\lambda$), we see that in general the two may be rather different, although they do become very close in the limit as $\lambda_0 \to 0$. The case $\lambda_0 = 0$ does not correspond to any normal prior, but is the what results from using a flat, improper prior (Problem 7-82 in Lindgren). Thus the frequentist and the Bayesian produce the same interval, albeit with different philosophical interpretation, when (and only when) the Bayesian uses the flat improper prior.

Otherwise, they disagree. The disagreement will be slight if the Bayesian's prior is very diffuse (this means the prior variance is very large, hence the prior precision λ_0 is very small). If the Bayesian's prior is fairly precise, the disagreement may be substantial and the 95% HPD region much shorter than the 95% confidence interval.

Example 11.6.2 (Marginal t Posterior for μ).
When the data are i. i. d. normal with both mean and variance unknown parameters and we use a conjugate prior, then Theorems 11.1 and 11.2 tell us that the marginal posterior for μ is a location-scale transform of a t distribution with noninteger degrees of freedom. More precisely, Theorem 11.2 says that $(\mu - \gamma)/d$ has a $t(\nu)$ distribution, where γ is a hyperparameter of the posterior and d and ν are defined (in the theorem) in terms of the other hyperparameters of the posterior α, β, and δ, and Theorem 11.1 gives the relation between the hyperparameters of the posterior and the hyperparameters of the prior and the data.

What is new in this example is that we want to figure out the HPD region for μ. This is easily done by the same logic that gives frequentist confidence intervals. By the symmetry of the t distribution, the HPD region is the set of μ values satisfying

$$\left| \frac{\mu - \gamma}{d} \right| < t_{\alpha/2}$$

where $t_{\alpha/2}$ is the $1 - \alpha/2$ quantile of the $t(\nu)$ distribution. So the HPD region is $\gamma \pm t_{\alpha/2} d$.

This resembles the frequentist confidence interval in being

$$\text{something} \pm t_{\alpha/2} \times \text{something else},$$

but the "something" is not $\overline{X}_n$, the "something else" is not $S_n/\sqrt{n}$ and the degrees of freedom ν is not $n - 1$ except for the particular improper prior used in Example 11.4.3.

Calculating HPD regions is not so easy when the posterior is not symmetric. Then it is generally necessary to do a computer search to find the endpoints of the region.

Example 11.6.3.
This continues Example 8.8b in Lindgren, which in class we gave both exact and asymptotic "frequentist" analyses. The data X_1, $\ldots$, X_n are i. i. d. Exp$(1/\theta)$. In order to be Bayesians we need a prior for θ, which we take to be $g(\theta) = 1/\theta$, an improper prior. The likelihood is

$$L(\theta) = \theta^{-n} \exp\left\{ -\frac{1}{\theta} \sum_{i=1}^{n} X_i \right\}$$

so the unnormalized posterior is

$$h(\theta) = L(\theta) g(\theta) = \theta^{-n-1} \exp\left\{ -\frac{1}{\theta} \sum_{i=1}^{n} X_i \right\} \tag{11.37}$$

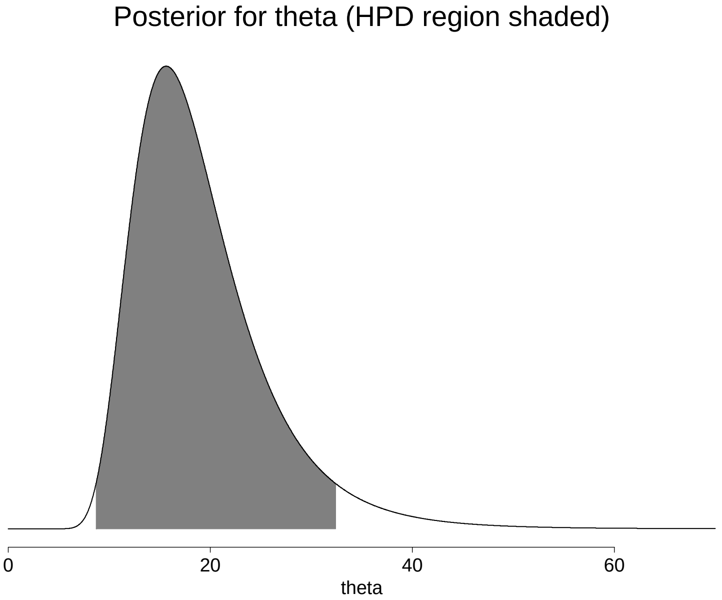
Lindgren would write $h(\theta \mid \mathbf{X})$ but we will temporarily suppress the data in the notation.

This is not a recognizable distribution, but it looks like $\lambda = 1/\theta$ has a gamma distribution. Let's check. Solving for θ the change of variable is $\theta = 1/\lambda = w(\lambda)$. The derivative of this transformation is $w'(\lambda) = -1/\lambda^2$, hence the unnormalized posterior for λ is

$$\begin{aligned} h[w(\lambda)] \cdot |w'(\lambda)| &= \lambda^{n+1} \exp\left\{ -\lambda \sum_{i=1}^{n} X_i \right\} \cdot \frac{1}{\lambda^2} \\ &= \lambda^{n-1} \exp\left\{ -\lambda \sum_{i=1}^{n} X_i \right\} \end{aligned}$$

which we recognize as $\text{Gam}\,(n, \sum_i X_i)$. Thus we can use the known distribution of λ to calculate probabilities about θ.

Finding the HPD region is not so easy. There is no way to calculate the endpoints or look them up in a table. There is a simple method using a computer. Plot the unnormalized posterior (11.37). For a specific numerical example we used before $\sum_i X_i = 172.0$. The unnormalized posterior for θ is the curve plotted below.



The shaded area is the probability of the HPD region. The region itself is the range of θ values covered $(8.70, 32.41)$. The posterior probability of the HPD region is 95% (this is the Bayesian analog of the "confidence" in a confidence interval).

The HPD region was determined from the plot as follows. The curve is actually plotted on a grid of points, spaced .01 apart on the θ axis. The sum of the y-values for these points approximates the integral for the normalizing constant of the unnormalized density $h(\theta)$ given by (11.37). The points with the highest y-values that constitute 95% of the sum are easily found by the computer and give a good approximation to the HPD region. The actual probability of the region, calculated using the gamma distribution of λ, is 94.99% (pretty close), and the heights of the unnormalized density (11.37) at the endpoints are 1.200×10^{-19} and 1.19610^{-19}. So we have come pretty close to a level set of the posterior.

11.7 Bayes Tests

The Bayesian view of one-tailed tests is fairly straightforward. Strangely, two-tailed tests, which the frequentist finds to be a minor variant of one-tailed tests, the Bayesian finds incredibly complicated and somewhat bizarre, so much so that Lindgren just avoids the subject. He blames the problem, calling it a "mathematical idealization," which is a meaningless criticism since so is everything else in statistics and every other mathematical subject.

A Bayesian one-tailed test is simple. The null and alternative hypotheses, being subsets of the parameter space are events, because the Bayesian considers parameters to be random variables. Hence they have probabilities (both prior and posterior). The test is done by calculating the posterior probabilities of the hypotheses and seeing which is bigger.

Example 9.5a in Lindgren provides an example of this. The data $Y \sim \text{Bin}(n, p)$ with $n = 15$ and the observed value of the data $y = 12$. The parameter p is unknown and given a $\mathcal{U}(0, 1)$ prior distribution. The hypotheses are

$$
\begin{aligned}
H_0 &: p \le \tfrac{1}{2} \\
H_A &: p > \tfrac{1}{2}
\end{aligned}
$$

Lindgren calculates $P(H_0 \mid Y = 12) = 0.0106$ and hence by the complement rule $P(H_A \mid Y = 12) = 1 - P(H_0 \mid Y = 12) = 0.9894$.

For comparison, Lindgren gives the P-value of the frequentist test, which is $P(Y \ge 12 \mid p = \frac{1}{2}) = 0.0176$. Both the Bayesian and frequentist tests strongly favor the alternative by conventional standards of evidence, and the P-value and Bayesian posterior probability of the null hypothesis are fairly similar, though different in philosophical interpretation. The frequentist says "probability of the null hypothesis" is a meaningless phrase because parameters are not random. The Bayesian says this probability exists and is 0.0106. The P-value is quite a different philosophical animal. It is the probability of seeing data at least as extreme as the actual observed data under the assumption that the null hypothesis is true. As we saw with confidence intervals and HPD regions, the numbers are slightly different, but the philosophical interpretations are wildly different.

The Bayesian two-tailed test runs into a problem. The null and alternative hypotheses are still subsets of the parameter space, hence are still events (to the Bayesian), and hence still have probabilities. The trouble is that when the hypotheses are

$$\begin{aligned} H_0 &: p = \tfrac{1}{2} \\ H_A &: p \neq \tfrac{1}{2} \end{aligned}$$

and the prior is continuous, the null hypothesis, being a single point, has probability zero. Thus if we use the same prior as we used for the one-tailed test, or any continuous prior, the posterior probability will be zero. But this is only because the prior probability is zero. As we saw in Problem 7-84 in Lindgren, whenever the prior probability is zero, the posterior probability will be zero too. So we haven't learned anything by doing such a test. Our mind was made up before we observed any data that H_0 was impossible and no data can change our minds. Might as well not bother to collect data or analyze it.

As Lindgren says on p. 313 a way out of this dilemma is to make the null hypothesis an interval, say

$$\begin{aligned} H_0 &: \tfrac{1}{2} - \epsilon \le p \le \tfrac{1}{2} + \epsilon \\ H_A &: p < \tfrac{1}{2} - \epsilon \text{ or } \tfrac{1}{2} + \epsilon < p \end{aligned}$$

for some $\epsilon > 0$. But this only adds to the problems. True the prior and posterior probabilities are now no longer zero, but where did ϵ come from? This "solution" has raised more questions than it answers. Furthermore, the posterior probability will still converge to zero if we let ϵ go to zero (by continuity of probability, Theorem 4 of Chapter 2 in Lindgren) so our analysis will depend very strongly on the choice of ϵ. We've only added to our troubles in finding a sensible Bayesian analysis.

The choice of ϵ is so problematic that most Bayesians that bother with two-tailed tests at all use a different solution to the dilemma. It is also weird, but less weird. The solution is to choose a prior that is not continuous, and puts some probability on the point null hypothesis $\Theta_0 = \{\frac{1}{2}\}$. For example, continuing the use of a uniform prior as much as possible, consider the prior distribution defined as follows

$$\begin{aligned} P(H_0) &= \alpha \\ P(H_A) &= 1 - \alpha \\ p \mid H_A &\sim \mathcal{U}(0, 1) \end{aligned}$$

This is a mixture of a distribution concentrated at one point ($\frac{1}{2}$) and a uniform distribution.

Allowing such a prior takes us out of the theory we know. If the prior is continuous, we calculate expectations, marginals, etc. by integrating. If it is discrete, by summing. If it is neither discrete nor continuous, we don't know what to do. Fortunately we can describe what to do in this very simple case where the distribution is continuous except for a single atom and avoid the complexity of the general situation.

In order to apply Bayes rule, we need to calculate the marginal probability of the observed data $P(Y = y)$ in the binomial example. We can do the calculation in two parts, using what Lindgren calls the "law of total probability" (Theorem 3 of Chapter 2), which in this case says

$$P(Y = y) = P(Y = y \,\text{and}\, p = \tfrac{1}{2}) + P(Y = y \,\text{and}\, p \neq \tfrac{1}{2}).$$

First

$$\begin{aligned} P(Y = y \,\text{and}\, p = \tfrac{1}{2}) &= P(Y = y \mid p = \tfrac{1}{2}) P(p = \tfrac{1}{2}) \\ &= \alpha \binom{n}{y} \left(\frac{1}{2}\right)^y \left(\frac{1}{2}\right)^{n-y} \\ &= \alpha \binom{n}{y} \left(\frac{1}{2}\right)^n \end{aligned}$$

Second

$$\begin{aligned} P(Y = y \,\text{and}\, p \neq \tfrac{1}{2}) &= P(Y = y \mid p \neq \tfrac{1}{2}) P(p \neq \tfrac{1}{2}) \\ &= (1 - \alpha) \binom{n}{y} \int_0^1 p^y (1 - p)^{n-y} \, dp \\ &= (1 - \alpha) \binom{n}{y} B(y + 1, n - y + 1) \\ &= \frac{1 - \alpha}{n + 1} \end{aligned}$$

Putting these together, the marginal probability is

$$P(Y = y) = \alpha \binom{n}{y} \left(\frac{1}{2}\right)^n + \frac{1 - \alpha}{n + 1}$$

Hence the probability of H_0 is

$$P(p = \tfrac{1}{2} \mid Y = y) = \frac{P(p = \frac{1}{2} \,\text{and}\, Y = y)}{P(Y = y)} = \frac{\alpha \binom{n}{y} \left(\frac{1}{2}\right)^n}{\alpha \binom{n}{y} \left(\frac{1}{2}\right)^n + \frac{1-\alpha}{n+1}} \tag{11.38}$$

That's how a Bayesian two-tailed test is done.

Some practical or philosophical (non-mathematical anyway) issues remain. The probability $P(H_0)$ given by (11.38) still depends strongly on the prior probability of H_0 (that is, α). This means that no two Bayesians will produce the same answer, since each will have a different prior probability (they will agree on the formula but plug in different numbers for α).

In order to eliminate this source of disagreement, we need a new notion, which is called the "Bayes factor" for the test. It is the ratio of the posterior to prior odds. Recall that *odds* are probabilities expressed as a ratio rather than a fraction. If the probability of an event is p, then the odds are $p/(1 - p)$. Here the prior odds of H_0 are $\alpha/(1 - \alpha)$ and the posterior odds are

$$\frac{P(p = \frac{1}{2} \mid Y = y)}{P(p \neq \frac{1}{2} \mid Y = y)} = \frac{\binom{n}{y} \left(\frac{1}{2}\right)^n}{\frac{1}{n+1}} \cdot \frac{\alpha}{1 - \alpha}$$

Hence the Bayes factor is the first term on the right hand side. Notice that it does not depend on α at all, although it still does depend on prior probabilities, since it depends on the choice of a prior that is uniform on the alternative hypothesis. The Bayes factor eliminates some, but not all, of the dependence on the prior.

Now let us plug in a few numbers, to get a concrete example. Continuing the example above with observed data $y = 12$, the Bayes factor is

$$\binom{15}{12} \left(\frac{1}{2} \right)^{15} \cdot (15 + 1) = 0.2221$$

For comparison, the two-tailed P-value is twice the one tailed $P = 0.035$ (because the distribution of the test statistic Y is symmetric under H_0, the binomial distribution is only symmetric when $p = \frac{1}{2}$ but that's what H_0 asserts).

Notice the big difference between the Bayesian and frequentist analyses. Frequentists are impressed with the evidence against H_0, at least those frequentists who think $P < 0.05$ implies "statistical significance." Bayesians are unimpressed. The data only lower the odds in favor of H_0 by a factor between 4 and 5 ($1/0.2221 = 4.5$). If the prior odds in favor of H_0 were even (1 to 1), then the posterior odds in favor of H_0 are now 0.222, and the posterior probability of H_0 is $0.222/(1 + 0.222) = .182$, still almost one chance in 5 that H_0 is true.

It shouldn't be any surprise that the frequentist and Bayesian answers turn out so different. They don't purport to resemble each other in any way. The only connection between the two is that they are competitors, different approaches to the same issue, saying something about whether H_0 or H_A is correct. The situation we saw here is typical, the Bayesian is always less impressed by evidence against H_0 and "accepts" H_0 less often than the frequentist.[3] This gives the Bayesians a problem with selling Bayes factors. Users of tests generally want to reject H_0. They didn't collect their data with the idea that there was nothing interesting in it (which is what H_0 usually says). Thus they are reluctant to switch to a procedure that makes rejecting H_0 even harder. Of course the Bayesian argues that frequentist tests are too lenient, but since frequentist tests are widely accepted and everyone is used to them, this sales job is an uphill battle.

Now let us go back and redo the calculation above abstractly so we get a general formula for the Bayes factor. Suppose we are doing a problem with likelihood $L_x(\theta)$ and put prior probability α on the point null hypothesis $H_0 : \theta = \theta_0$ and $1 - \alpha$ on the alternative hypothesis $H_A : \theta \neq \theta_0$ distributed according to the conditional density $g(\theta)$. Unlike the situation in most Bayesian inference, we must have $g(\theta)$ a proper probability density (not improper, not unnormalized).

As in the example above, the marginal probability of the data is

$$\begin{aligned} P(X = x) &= P(X = x \,\text{and}\, H_0) + P(X = x \,\text{and}\, H_A) \\ &= P(X = x \mid H_0)\alpha + P(X = x \mid H_A)(1 - \alpha). \end{aligned}$$

[3] Berger and Sellke, "Testing a point null hypothesis: The irreconcilability of P values and evidence" (with discussion), *Journal of the American Statistical Association*, 82:112-122, 1987.

The posterior probability of H_0 is

$$\begin{aligned} P(H_0 \mid X = x) &= \frac{P(X = x \,\text{and}\, H_0)}{P(X = x)} \\ &= \frac{P(X = x \mid H_0)\alpha}{P(X = x \mid H_0)\alpha + P(X = x \mid H_A)(1 - \alpha)} \end{aligned}$$

and the posterior odds in favor of H_0 are

$$\frac{P(X = x \mid H_0)}{P(X = x \mid H_A)} \cdot \frac{\alpha}{1 - \alpha}$$

Thus the Bayes factor is the first term above, the ratio of prior to posterior odds

$$\frac{P(X = x \mid H_0)}{P(X = x \mid H_A)}$$

To proceed we need the density of the data, which as always is proportional to the likelihood, $f(x \mid \theta) = c(x)L_x(\theta)$. Then

$$P(X = x \mid H_0) = c(x)L_x(\theta_0)$$

and

$$P(X = x \mid H_A) = c(x) \int L_x(\theta)g(\theta)\,d\theta$$

So the Bayes factor in favor of H_0 is

$$\frac{L_x(\theta_0)}{\int L_x(\theta)g(\theta)\,d\theta}$$

Notice several things. First, the factor $c(x)$ appears in both the numerator and denominator of the Bayes factor and hence cancels, not appearing in the result. Second, the prior on the alternative $g(\theta)$ appears *only* in the denominator. That's why it must be a proper density. If it were unnormalized or improper, that would introduce an arbitrary constant that would not cancel into the Bayes factor, rendering it meaningless.

11.8 Bayesian Asymptotics

For large sample sizes, frequentist and Bayesian procedures (most of them anyway) give approximately the same answers. This is the result of a theorem that we will not state precisely. Under the same conditions required for the usual asymptotics of maximum likelihood plus one additional condition (that usually holds, but we won't describe since it is fairly technical) the asymptotic posterior distribution is "the same" as the asymptotic sampling distribution of the MLE. We put "the same" in quotes because the philosophical interpretation is radically different, but the asymptotic distribution is the same in both cases. Here's what we mean. The asymptotic sampling distribution of the MLE $\hat{\theta}_n$ is

$$\hat{\theta}_n \approx \mathcal{N}\left(\theta, \frac{1}{I_n(\theta)}\right)$$

where θ is the true parameter value. Of course we don't know θ (that's why we are estimating it) so we don't know the asymptotic variance $1/I_n(\theta)$. But we can consistently estimate it, plugging in $\hat{\theta}_n$ for θ giving

$$\hat{\theta}_n \approx \mathcal{N}\left(\theta, \frac{1}{I_n(\hat{\theta}_n)}\right) \tag{11.39}$$

Equation (11.39) is fairly sloppy notation. Strictly speaking, we should write

$$\sqrt{I_n(\hat{\theta}_n)}\big(\hat{\theta}_n - \theta\big) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1) \tag{11.40}$$

but it is clear what is meant. The asymptotic posterior distribution of the parameter θ is

$$\theta \approx \mathcal{N}\left(\hat{\theta}_n, \frac{1}{I_n(\hat{\theta}_n)}\right) \tag{11.41}$$

Comparing with (11.39) we see that they differ only in the interchange of θ and $\hat{\theta}_n$. The frequentist considers θ fixed and $\hat{\theta}_n$ random and the asymptotic sampling distribution of $\hat{\theta}_n$ to be a normal distribution centered at the unknown true parameter value θ. The Bayesian considers $\hat{\theta}_n$ fixed (Bayesians *condition* on the data) and θ random and the asymptotic posterior distribution of θ to be a normal distribution centered at the MLE $\hat{\theta}_n$.

It is an important point that the asymptotic posterior distribution does *not* depend on the prior distribution of the parameter so long as the prior density is continuous and nonzero at the true parameter value. The catch phrase that expresses this is that the likelihood "outweighs" the prior for large sample sizes. Thus for large (perhaps very large) sample sizes all Bayesians agree (priors don't matter) and they also agree with the frequentists.

At least they agree about most things. Frequentist asymptotic confidence intervals will also be Bayesian asymptotic HPD regions. Frequentist asymptotic P-values for one-tailed tests will also be Bayesian asymptotic posterior probabilities of the null hypothesis for the same tests. One thing that will stay different is two-tailed tests. For them the posterior probabilities do not go away asymptotically and the frequentist and Bayesian do not get the same results no matter how large the sample size.

Problems

11-1. Suppose we observe $X \sim \text{Poi}(\mu)$ and we want to do a Bayesian analysis with prior distribution $\text{Gam}(\alpha, \beta)$ for μ, where α and β are known numbers expressing our prior opinion about probable values of μ.

(a) Find the posterior distribution of μ.

(b) Find the posterior mean of μ.

11-2. Suppose X is a single observation from a $\text{Gam}(\alpha, \lambda)$ distribution, where α is a known constant. Suppose our prior distribution for λ is $\text{Gam}(\alpha_0, \lambda_0)$, where the hyperparameters α_0 and λ_0 are also known constants.

(a) Find the posterior distribution for λ given X.

(b) Find the posterior mean $E(\lambda \mid X)$.

(c) Find the posterior mode of λ.

11-3. Using the same improper prior as was used in Example 11.4.3, show that the posterior marginal distribution of $(n-1)S_n^2\lambda$ is the same as its sampling distribution. More precisely stated, show that the frequentist sampling distribution of $(n-1)S_n^2\lambda$ with λ considered a nonrandom constant is the same as the Bayesian marginal posterior distribution of $(n-1)S_n^2\lambda$ with λ considered random and $S_n^2 = s_n^2$ fixed at the observed value.

11-4. Find the posterior mean and variance of μ when the data are i. i. d. normal and the prior is a general normal-gamma prior. Say for which values of the hyperparameters the posterior mean and variance of μ exist.

11-5. Suppose X_1, ..., X_n are i. i. d. $\mathcal{N}(\mu, 4)$, the prior distribution for μ is $\mathcal{N}(10, 9)$, and the sample mean of a sample of size 10 is $\overline{X}_n = 12$. Calculate a 90% HPD region for μ (note not 95%).

11-6. Suppose $X_1, \ldots, X_n$ are i. i. d. $\mathcal{N}(\mu, \lambda^{-1})$, the prior distribution for (μ, λ) is the conjugate normal-gamma prior with

$$\begin{aligned} \lambda &\sim \text{Gam}(3, 3) \\ \mu \mid \lambda &\sim \mathcal{N}(10, 16\lambda^{-1}) \end{aligned}$$

the sample mean of a sample of size 15 is $\overline{X}_n = 12$ and the sample variance is $S_n^2 = 50$ (note not V_n). Calculate a 95% HPD region for μ.

11-7. Suppose $X \sim \text{Bin}(n, p)$, where p is an unknown parameter. Find a formula giving the Bayes factor for the two-tailed test of

$$\begin{aligned} H_0 &: p = p_0 \\ H_A &: p \neq p_0 \end{aligned}$$

when the prior distribution for p given H_A is $\text{Beta}(s, t)$, where s and t are known constants. Hint: this is just like the test worked out in Section 11.7 except for the prior.

11-8. Suppose X_1, ..., X_n are i. i. d. $\mathcal{N}(\mu, \lambda^{-1})$, where μ is an unknown parameter and the precision λ is a known constant. Find a formula giving the Bayes factor for the two-tailed test of

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_A &: \mu \neq \mu_0 \end{aligned}$$

when the prior distribution for μ given H_A is $\mathcal{N}(\mu_0, \lambda_0^{-1})$, where μ_0 and λ_0 are known constants.

11-9. Suppose the setup described at the end of Section 11.3. Verify that the posterior (11.19) from the two-stage analysis described in that section is the same as the posterior from analyzing all the data at once, which would be (11.18) with m replaced by n.

11-10. Suppose X_1, X_2, ... X_n are i. i. d. from the distribution with density

$$f(x) = \theta x^{-\theta - 1}, \qquad x > 1,$$

where $\theta > 0$ is an unknown parameter. Suppose our prior distribution for the parameter θ is $\text{Exp}(\lambda)$, where λ is a known number (hyperparameter of the prior).

(a) Find the posterior density of θ.

(b) Find the posterior mean of θ.

(c) Find the posterior mode of θ.

Chapter 12

Regression

12.1 The Population Regression Function

12.1.1 Regression and Conditional Expectation

Recall from last semester (Section 3.3.2 of these notes) that "regression function" is another name for conditional expectation. Recall that a conditional expectation *is not* a function of the variable or variables "in front of the bar" and *is* a function of the variable or variables "behind the bar." Thus $E(Y \mid X)$ is not a function of Y and is a function of X, so we can write

$$h(X) = E(Y \mid X).$$

This function h is an ordinary function. When we wish to emphasize this and write it as a function of an ordinary variable, we write

$$h(x) = E(Y \mid x),$$

but the meaning is the same in either case. This function h is called the *regression function of Y on X*, the reason for the long name being that

$$g(Y) = E(X \mid Y)$$

defines a different function, the *regression function of X on Y*.

When we developed this terminology last semester, we had not begun systematic study of random vectors. Now we want to generalize this to allow a vector variable "behind the bar" leaving the variable in "front of the bar" a scalar. Then the regression function is a scalar function of a vector variable

$$h(\mathbf{X}) = E(Y \mid \mathbf{X})$$

which we can also think of as a function of several variables

$$h(X_1, \ldots, X_k) = E(Y \mid X_1, \ldots, X_k).$$

12.1.2 Best Prediction

There is a connection between conditional expectation (or the regression function) and prediction, which is given by the following theorem, which is Theorem 3.6 in last semester's notes improved to have a random vector "behind the bar." The proof is exactly the same as for Theorem 3.6 except for boldface type for $\mathbf{X}$, which does not make any essential difference.

Theorem 12.1 (Best Prediction). *For predicting a random variable Y given the value of a random vector $\mathbf{X}$, the predictor function $a(\mathbf{X})$ that minimizes the expected squared prediction error*

$$E\{[Y - a(\mathbf{X})]^2\}$$

is the conditional expectation $a(\mathbf{X}) = E(Y \mid \mathbf{X})$.

This theorem is analogous to theorem about the characterization of the mean (Corollary 7.2 in these notes). Together these two theorems say

- The best estimate of the the value of a random variable Y, where "best" means minimizing expected squared prediction error, is the mean $E(Y)$, when no other information is available.

- The best estimate of the the value of a random variable Y given the value of a random vector $\mathbf{X}$, where "best" means minimizing expected squared prediction error, is the conditional mean $E(Y \mid \mathbf{X})$.

The theorem gives yet another name for $E(Y \mid \mathbf{X})$. In addition to *conditional expectation* and the *regression function*, we also call it the *best predictor* (BP). Sometimes the best predictor is called the *best unbiased predictor* (BUP) because it is unbiased in the sense that its expectation is the mean of Y. This is a consequence of the iterated expectation property (Axiom CE2 for conditional expectation in Chapter 3 of these notes).

$$E\{E(Y \mid \mathbf{X})\} = E(Y).$$

Since the best predictor is always unbiased, it is irrelevant whether or not you bother to mention that it is unbiased. BP and BUP mean the same thing.

We give no examples because our interest in BP is mostly abstract. If you know the regression function, then you use it to give the best prediction. But when we are doing statistics, we don't know the regression function, because it depends on the true distribution of the data, and that depends on unknown parameters. Thus when doing statistics, the regression function isn't something we calculate, it's something we *estimate*. And often, as we will see in the next section, we don't even try to use the regression function (use best prediction), because it's too hard.

12.1.3 Best Linear Prediction

A widely used simplification of the regression problem restricts the allowable predictions to linear predictions, functions of the form

$$h(\mathbf{X}) = \alpha + \boldsymbol{\beta}'\mathbf{X} = \alpha + \sum_{i=1}^{n} \beta_i X_i. \tag{12.1}$$

(where α and β_1, …, β_n are constants). The function of this form that has the smallest mean square prediction error

$$E\{(Y - \alpha - \boldsymbol{\beta}'\mathbf{X})^2\} \tag{12.2}$$

is called the *best linear predictor* (BLP).

It should be understood, that using BLP is the Wrong Thing (not BP) unless the regression function just happens to be linear. The reason for doing the Wrong Thing is presumably because the Right Thing (using BP) is too hard, or we are too ignorant, or something of the sort.

Estimating the regression function is hard, but can be done. The main reason for the widespread use of linear prediction is that it had a 150 year head start (the development of linear regression theory started around 1800, whereas the development of nonlinear regression theory didn't really take off until the 1970's and is still a very active research area). So people understand linear regression much better, there is a long history of use in the various sciences, and so forth. Hence we will study it because of its popularity. (You should keep in mind though, that it is usually the Wrong Thing, decidedly not "best" despite the name).

We are now going to do a "stupid math trick" that simplifies notation at the expense of some mystification. Expression (12.1) is needlessly complicated (says the mathematician) by having two kinds of coefficients: α and the β_i. Only one kind is actually needed. We can consider (12.1) a special case of

$$h(\mathbf{X}) = \boldsymbol{\beta}'\mathbf{X} = \sum_{i=1}^{n} \beta_i X_i, \tag{12.3}$$

because if we make X_1 the constant random variable $X_1 = 1$, then (12.3) becomes

$$h(\mathbf{X}) = \beta_1 + \sum_{i=2}^{n} \beta_i X_i,$$

and this describes the same family of predictor functions as (12.1). Only the notation has changed (what was α is now β_1, what was β_i is now β_{i+1} for $i > 1$).

Thus we see that, although the simpleminded notion of the relationship between our two expressions for a linear prediction function is that (12.3) is a special case of (12.1) obtained by taking $\alpha = 0$, the really sophisticated notation is just the reverse, that (12.1) is a special case of (12.3) obtained by taking one of the $X_i = 1$. Having seen this, we will just use the mathematically simpler form (12.3) without any assertion that any of the X_i are constant. Understanding the general case tells us about the special case.

Theorem 12.2 (Best Linear Prediction). *For predicting a random variable Y given the value of a random vector $\mathbf{X}$, the linear predictor function* (12.3) *that minimizes the expected squared prediction error*

$$E\{(Y - \boldsymbol{\beta}'\mathbf{X})^2\} \tag{12.4}$$

is defined by

$$\boldsymbol{\beta} = E(\mathbf{X}\mathbf{X}')^{-1}E(Y\mathbf{X}) \tag{12.5}$$

assuming the inverse exists.[1]

Proof. The m. s. p. e. (12.4) is a quadratic function of β_1, $\ldots$, β_n. Since it is nonnegative it is a positive semi-definite quadratic function and hence has a global minimum where the first derivative is zero.

The proof is simpler if we rewrite (12.4) in non-matrix notation as

$$Q(\boldsymbol{\beta}) = E\left\{\left(Y - \sum_{i=1}^n \beta_i X_i\right)\left(Y - \sum_{j=1}^n \beta_j X_j\right)\right\}$$

and further simplify using linearity of expectation

$$Q(\boldsymbol{\beta}) = E(Y^2) + \sum_{i=1}^n \sum_{j=1}^n \beta_i \beta_j E(X_i X_j) - 2\sum_{i=1}^n \beta_i E(Y X_i)$$

The first derivative vector has components

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_k} = 2\sum_{i=1}^n \beta_i E(X_i X_k) - 2E(Y X_k) \tag{12.6}$$

These are n linear equations in n unknowns β_1, $\ldots$, β_k. They always have a solution (not necessarily unique), but in general there is no nice expression for the solution, except that we can always write the solution of any set of linear equations in terms of a matrix inverse (if the inverse exists, or a generalized inverse if not). Before we can do that, we need to put (12.6) back in matrix notation, using the fact that $E(\mathbf{X}\mathbf{X}')$ is a matrix with components $E(X_i X_j)$ so (12.6) can be rewritten

$$\nabla Q(\boldsymbol{\beta}) = 2E(\mathbf{X}\mathbf{X}')\boldsymbol{\beta} - 2E(Y\mathbf{X}) \tag{12.7}$$

Hence the equations to be solved are (12.7) set to zero, that is

$$E(\mathbf{X}\mathbf{X}')\boldsymbol{\beta} = E(Y\mathbf{X}) \tag{12.8}$$

Multiplying both sides on the left by $E(\mathbf{X}\mathbf{X}')^{-1}$ gives (12.5). □

[1] If the inverse does not exist, it can be replaced by a so-called "generalized inverse" and the same formula still produces a best linear predictor, but a generalized inverse is non-unique, so the $\boldsymbol{\beta}$ produced by the formula is non-unique. However, every such $\boldsymbol{\beta}$ gives the same prediction $\boldsymbol{\beta}'\mathbf{X}$ for the same value of $\mathbf{X}$. The nonuniqueness arises because $\mathbf{X}$ is a degenerate random vector (concentrated on a hyperplane). We will ignore this issue henceforth and assume the inverse exists.

For convenience we give the special case in which there is one constant predictor variable and one non-constant predictor, in which case we write the linear predictor function as $\alpha + \beta X$.

Corollary 12.3. *For predicting a random scalar Y given the value of a random scalar X, the linear predictor function of the form $\alpha + \beta X$ that minimizes the expected squared prediction error is defined by*

$$\alpha = \mu_Y - \beta \mu_X \tag{12.9a}$$

$$\beta = \rho \frac{\sigma_Y}{\sigma_X} \tag{12.9b}$$

This looks so different from the form given in the theorem that it is simpler to just derive it separately, which is done on p. 426 in Lindgren. The reason we have called it a "corollary" is to remind you that it is, despite appearances, a special case of the theorem.

As with BP and BUP, sometimes that BLP is BLUP (best linear unbiased prediction). In general the BLP is not unbiased, but when one of the predictors is constant (when we are using our "stupid math trick") it is.

BLP is BLUP when one of the predictors is constant.

In particular, there is no difference between BLP and BLUP in Corollary 12.3.

The proof of this assertion is a direct consequence of (12.8) in the proof of the theorem. This one vector equation is equivalent to n scalar equations, which are sometimes called the "normal equations." If $X_k = 1$ with probability one, then the k-th normal equation

$$E(X_k \mathbf{X}')\boldsymbol{\beta} = E(Y X_k)$$

becomes

$$E(Y) = E(\mathbf{X})'\boldsymbol{\beta} = \boldsymbol{\beta}' E(\mathbf{X}) = E(\boldsymbol{\beta}'\mathbf{X})$$

and this says that the prediction $\boldsymbol{\beta}'\mathbf{X}$ is unbiased for Y.

Example 12.1.1 (A Pretty Bad "Best" Linear Prediction).
In Example 3.5.1 in Chapter 3 of these notes we considered positive scalar random variables X and Y having joint density

$$f(x, y) = \tfrac{1}{2}(x + y)e^{-x-y}, \qquad x > 0,\ y > 0.$$

There we found the best predictor of X given Y is

$$a(Y) = E(X \mid Y) = \frac{2 + Y}{1 + Y}, \qquad Y > 0.$$

This is a fairly nonlinear function so we don't expect BLP to do very well, and it doesn't.

Direct calculation using gamma integrals gives

$$
\begin{aligned}
E(X) &= \frac{1}{2} \int_0^\infty \int_0^\infty (x^2 + xy) e^{-x-y} \, dx \, dy \\
&= \frac{1}{2} \int_0^\infty (2 + y) e^{-y} \, dy \\
&= \frac{3}{2} \\
E(X^2) &= \frac{1}{2} \int_0^\infty \int_0^\infty (x^3 + x^2 y) e^{-x-y} \, dx \, dy \\
&= \frac{1}{2} \int_0^\infty \int_0^\infty (6 + 2y) e^{-y} \, dy \\
&= 4 \\
E(XY) &= \frac{1}{2} \int_0^\infty \int_0^\infty (x^2 y + x y^2) e^{-x-y} \, dx \, dy \\
&= \frac{1}{2} \int_0^\infty \int_0^\infty (2y + y^2) e^{-y} \, dy \\
&= 2
\end{aligned}
$$

By symmetry $E(X) = E(Y)$ and $E(X^2) = E(Y^2)$. So

$$
\begin{aligned}
\operatorname{var}(X) &= E(X^2) - E(X)^2 = \frac{7}{4} \\
\operatorname{var}(Y) &= \operatorname{var}(X) \\
\operatorname{cov}(X, Y) &= E(XY) - E(X)E(Y) = -\frac{1}{4} \\
\operatorname{cor}(X, Y) &= \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X) \operatorname{var}(Y)}} = -\frac{1}{7}
\end{aligned}
$$

So the BLP corollary gives

$$
\begin{aligned}
\beta &= -\frac{1}{7} \\
\alpha &= \frac{12}{7}
\end{aligned}
$$

and best linear prediction is

$$
a_{\text{blp}}(Y) = \frac{12 - Y}{7}, \qquad Y > 0.
$$

Note that for $Y > 12$ the BLP is negative, whereas the variable X it is predicting is necessarily positive. So the prediction isn't very good. The theorem asserts that this prediction is the best of all linear predictions. The problem is that no linear prediction is very good, even the best of them.

The BLP isn't the BP. Sometimes the BLP is a very bad predictor.

The theorem describes the BLP (or BLUP when one of the predictors is constant) in the case where you *know* the "population" distribution, the true distribution of $\mathbf{X}$ and Y. But when we are doing statistics, we don't know the true distribution of the data, because it depends on unknown parameters. Thus when doing statistics, the BLP or BLUP isn't something we calculate, it's something we *estimate*. It's the true unknown "population" function that we are trying to estimate from a sample.

12.2 The Sample Regression Function

Recall the empirical distribution introduced in Section 7.1 of these notes. It is the distribution that puts probability $1/n$ at each of n points. There we were interested in the case where the points were scalars. Here we are interested in the case where the points are vectors, but there is no real difference except for boldface. The empirical expectation operator associated with the vector points $\mathbf{x}_1$, $\ldots$, $\mathbf{x}_n$ is defined by

$$E_n\{g(\mathbf{X})\} = \frac{1}{n} \sum_{i=1}^{n} g(\mathbf{x}_i). \tag{12.10}$$

which is just (7.2) with some type changed to boldface.

In regression, we are interested in vectors of a rather special form, consisting of a scalar "response" variable y and a vector "predictor" variable $\mathbf{x}$. Suppose we have observed a sample of predictor-response pairs $(\mathbf{x}_i, y_i)$, then the corresponding empirical expectation formula is

$$E_n\{g(Y, \mathbf{X})\} = \frac{1}{n} \sum_{i=1}^{n} g(y, \mathbf{x}_i). \tag{12.11}$$

In particular, the empirical mean square prediction error for a linear predictor of the form described in the theorem is

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}' \mathbf{x}_i)^2 \tag{12.12}$$

Now the theorem applied to the empirical distribution gives the following. The empirical BLP is

$$\hat{\boldsymbol{\beta}} = E_n(\mathbf{X}\mathbf{X}')^{-1} E_n(Y\mathbf{X}) \tag{12.13}$$

where

$$E_n(\mathbf{X}\mathbf{X}') = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \tag{12.14}$$

$$E_n(Y\mathbf{X}) = \frac{1}{n} \sum_{i=1}^{n} y_i \mathbf{x}_i \tag{12.15}$$

However, this is not the usual form in which the empirical analogue of the theorem is stated. The usual form involves yet another "stupid math trick." The formulas above have some explicit sums, those involved in the empirical expectation, and some implicit sums, those involved in matrix multiplications. The "stupid math trick" we are now introducing makes all the sums implicit (matrix multiplications).

To understand the trick, we need a closer look at the predictor variables. The subscripts on the $\mathbf{x}_i$ do not denote components, but different vectors in the sample. Each $\mathbf{x}_i$ is a vector and has components, say

$$\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$$

(Before we were writing n as the dimension of the vectors. Now we are using n for the sample size. So the dimension must be a different letter, here p.) Thus the "predictor" part of the observed data are np variables

$$x_{ij}, \qquad i = 1, \ldots, n \text{ and } j = 1, \ldots, p.$$

which if we like, we can think of as an $n \times p$ matrix $\mathbf{X}$ (we've introduced a new notation here: $\mathbf{X}$ with no subscripts will henceforth be an $n \times p$ matrix). This matrix is very important in the theory of linear regression. It is commonly called the *design matrix*. The reason for the name is that if the data are from a designed experiment, then the design matrix incorporates everything about the design that is involved in linear regression theory. If the data are not from a designed experiment, then the name is inappropriate, but everyone uses it anyway. The relationship between the design matrix $\mathbf{X}$ and the predictor vectors $\mathbf{x}_1$, ..., $\mathbf{x}_p$ is that the predictor vectors are the columns of the design matrix.

Now write $\boldsymbol{\mu}$ as the n-dimensional vector of all the theoretical predictions (the conditional expectation of Y_i given all the $\mathbf{x}$'s), which has components

$$\mu_i = \boldsymbol{\beta}'\mathbf{x}_i = \sum_{j=1}^{p} X_{ij}\beta_j$$

This sum can be written as a matrix multiplication

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}. \tag{12.16}$$

because the dimensions match

$$\underset{n \times 1}{\boldsymbol{\mu}} = \underset{n \times p}{\mathbf{X}} \quad \underset{p \times 1}{\boldsymbol{\beta}}$$

Now we want to also write the sum in (12.12) as a matrix multiplication. The way we do this is to note that for any vector $\mathbf{z} = (z_1, \ldots, z_n)$

$$\mathbf{z}'\mathbf{z} = \sum_{i=1}^{n} z_i^2.$$

Applying this with $\mathbf{z} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ gives

$$\frac{1}{n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{12.17}$$

as another expression for the empirical m. s. p. e. (12.12).

Now we also want to rewrite (12.14) and (12.15) using this trick. When we write these equations out explicitly using all the subscripts, we see that (12.14) is a matrix with (j, k)-th element

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik}$$

which is seen to be the j, k element of $\mathbf{X}'\mathbf{X}/n$. Similarly, (12.15) is a vector with j-th element

$$\frac{1}{n}\sum_{i=1}^{n} y_i X_{ij}$$

which is seen to be the j-th element of $\mathbf{y}'\mathbf{X}/n$ or of $\mathbf{X}'\mathbf{y}/n$. Putting these together we get the following very compact matrix notation for (12.13)

$$\hat{\boldsymbol{\beta}} = \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\frac{1}{n}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

This is the “usual” way the empirical version of the BLP theorem is written

Corollary 12.4 (Multiple Linear Regression). *The $\boldsymbol{\beta}$ that minimizes* (12.17) *is*

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{12.18}$$

For completeness, we also record the empirical analog of Corollary 12.3

Corollary 12.5 (Simple Linear Regression). *The values α and β that minimize the empirical expected squared prediction error*

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$$

are

$$\hat{\alpha} = \bar{y} - \beta\bar{x} \tag{12.19a}$$

$$\hat{\beta} = r\frac{s_y}{s_x} \tag{12.19b}$$

Fortunately, we do not have to do the calculations described by these corollaries by hand. Many calculators will do the “simple case” of Corollary 12.5. Any computer statistics package will do the “multiple” case of Corollary 12.4.

Here’s an example using R.

Example 12.2.1 (Multiple Regression).
We use the data in the URL

```
http://www.stat.umn.edu/geyer/5102/ex12.2.1.dat
```

The R command that does multiple linear regression is `lm` (for “linear model”). This data set has three variables `x1`, `x2`, and `y`. In R each is a vector, and they all have the same length (in this particular data set $n = 100$). The response is `y`, and the predictor variables are `x1` and `x2`. The specific R commands that do the regression and print the results are

```
out <- lm(y ~ x1 + x2)
summary(out)
```

The first command doesn’t print anything (it just returns the dataset `out`), the latter prints the fairly voluminous output

```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
     Min       1Q   Median       3Q      Max
-121.338  -32.564    5.525   35.309  124.846

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  94.4803    10.9616   8.619 1.27e-13 ***
x1            0.8503     0.5606   1.517    0.133
x2            1.3599     0.5492   2.476    0.015 *
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1

Residual standard error: 54.22 on 97 degrees of freedom
Multiple R-Squared: 0.5875,     Adjusted R-squared: 0.579
F-statistic: 69.08 on 2 and 97 degrees of freedom,      p-value:     0
```

most of which we won’t explain now (and a fair amount of which we won’t explain ever).

The first thing we will explain is what model was fit, and where to find the estimates of the β’s in the printout. R always assumes by default that you want a constant predictor. Hence the model fit here has *three* predictors, not just the two explicitly mentioned. Hence it also has three corresponding parameters. We can write the model as

$$h(\mathbf{x}) = \alpha + \beta_1 x_1 + \beta_2 x_2.$$

Information about the parameter estimates is found in the section labeled `Coefficients:` (α and the β_i and their estimates are often called *regression* *coefficients* because they are the coefficients of the predictor variables in the definition of the regression function). The estimates are given in the column labeled `Estimate` in that section, which we repeat here

```
             Estimate
(Intercept)  94.4803
x1            0.8503
x2            1.3599
```

The three estimates are $\hat{\alpha}$, $\hat{\beta}_1$ and $\hat{\beta}_2$ respectively. The coefficients of the non-constant predictors are labeled by the names of the variables they multiply. The coefficient of the constant predictor is labeled `(Intercept)` because α is usually called the "y-intercept" in elementary math.

If you actually wanted to do regression *without* a constant predictor, you would need to know the magic incantation that makes R do this. In fact, it has two[2]

```
  out <- lm(y ~ x1 + x2 + 0)
  out <- lm(y ~ x1 + x2 - 1)
```

So that covers the mechanics of doing linear regression. Let the computer do it!

12.3 Sampling Theory

12.3.1 The Regression Model

In order to have sampling theory, we need a probability model. The probability model usually adopted assumes that we observe pairs $(Y_i, \mathbf{X}_i)$, $i = 1$, 2, ... The Y_i are scalars, and the $\mathbf{X}_i$ are vectors. The $\mathbf{X}_i$ may or may not be random, but if random we condition on them, meaning we *treat* them as if *not* random. Thus we will henceforth write them as lower case $\mathbf{x}_i$.

The *linear regression model* (sometimes just called the *linear model*) is that the means of the Y_i are linear functions of the $\mathbf{x}_i$

$$E(Y_i) = \boldsymbol{\beta}'\mathbf{x}_i. \tag{12.20}$$

Note that this formula assumes the $\mathbf{x}_i$ are constants. If we didn't assume that, we would have to write (12.20) as

$$E(Y_i \mid \mathbf{X}_i) = \boldsymbol{\beta}'\mathbf{X}_i \tag{12.21}$$

(this is the last time we will note the distinction between the two approaches).

We also usually write

$$Y_i = \boldsymbol{\beta}'\mathbf{x}_i + e_i, \qquad i = 1, \ldots, n,$$

[2] The reason for two is that the R team like the first, and the second is provided for backwards compatibility with the S language, which R is more or less a clone of. I find neither very intuitive.

which can be written as a single vector equation

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{12.22}$$

where $\mathbf{e} = (e_1, \ldots, e_n)$.

This equation has no statistical content. We are just defining variables e_i to be the deviations of the Y_i from their means. The e_i are usually called "errors." Despite the lower case letter, they are random variables ("big E" is a frozen letter, reserved for expectation operators). In order for (12.20) to hold, the errors must have mean zero.

To further specify the distribution, we can describe the distribution of either the Y_i or the e_i. The latter is simpler. There are two different types of assumptions that can be made about the errors: strong and weak. The weak assumption, used in the following section, describes only the first two moments. The weak assumption says

$$\begin{aligned} E(\mathbf{e}) &= 0 \\ \operatorname{var}(\mathbf{e}) &= \sigma^2 \mathbf{I} \end{aligned} \tag{12.23}$$

or in words, the errors

- have mean zero,
- are uncorrelated, and
- have constant variance

(that is, they all have the same variance).[3]

The strong assumption actually gives the distribution of the errors

$$\mathbf{e} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}) \tag{12.24}$$

It is a special case of the weak assumption. It says the errors have the mean and variance specified by (12.23), and in addition that they are multivariate normal. Note that by Theorem 5.13 of Chapter 5 in last semester's notes (uncorrelated implies independent for jointly multivariate normal random variables), an equivalent way to state the strong assumption is that the e_i are i. i. d. $\mathcal{N}(0, \sigma^2)$.

Thus the weak assumption only makes the errors uncorrelated (which does *not* imply they are independent if they are not multivariate normal), whereas the strong assumption makes the errors both independent and normally distributed.

Both the weak and strong assumption make the same assumption (12.20) about the means of the Y_i. Another way to describe this part of the model assumptions is by saying that we are assuming that the true population regression function is linear. It is clear from (12.21) that

$$h(\mathbf{x}) = \boldsymbol{\beta}'\mathbf{x} \tag{12.25}$$

[3] The assumption of constant variance is so important that some statistician invented a big word to describe it: *homoscedasticity.* Violation of the assumption (different variances) is called *heteroscedasticity.* But we'll just say "constant variance" and "non-constant variance."

is assumed to be the regression function of Y on $\mathbf{X}$. We often call this the *population regression function* in our usual abuse of terminology that talks about i. i. d. variables as a "sample" from an infinite "population." What we mean is only the assertion (12.20) or (12.21) that this is the true unknown regression function.

This is the usual statistical set-up. The $\mathbf{x}_i$ and Y_i are observed data; the β_i are unknown parameters. The best we can do is to *estimate* the unknown parameters with *estimates* $\hat{\beta}_i$ that are functions of the data and study the statistical properties of the estimates (and eventually get confidence intervals, hypothesis tests, and the rest of the paraphernalia of statistical inference).

The estimators we are going to study are the empirical BLUP estimates described by Corollaries 12.4 and 12.5. The name "empirical BLUP" we used for them is nonstandard. They have accumulated a lot of names over the years. One common name for the $\hat{\beta}_i$ is *sample regression coefficients*. (And the analogous term for the β_i is *population regression coefficients*.) Another common name, in use for 200 years, for the $\hat{\beta}_i$ is *least squares estimates* because they minimize the empirical m. s. p. e. (12.12).

When we plug the estimates into (12.25) we get

$$\hat{h}(\mathbf{x}) = \hat{\boldsymbol{\beta}}'\mathbf{x}, \tag{12.26}$$

which is the *sample regression function*.[4]

It is important here, as everywhere else in (frequentist) statistics to keep the slogan about *the sample is not the population* firmly in mind. Even assuming that the population regression function is linear so (12.25) gives best predictions, the sample regression function (12.26) does *not* give best predictions because the sample is not the population. How far off they are is the job of sampling theory to describe.

12.3.2 The Gauss-Markov Theorem

This section uses only the "weak" distributional assumptions (12.23). Normality is not used. The content of the Gauss-Markov theorem is simply stated as the least squares estimates are *best linear unbiased estimates* (BLUE).

Before we can even state the theorem properly we need to explain what "best" means in this context. For unbiased estimates mean square error is the same as variance. So best means smallest variance. The problem is that the estimate $\hat{\boldsymbol{\beta}}$ is a vector, so its variance is a matrix. What does it mean for one matrix to be "smaller" than another?

In general there is no sensible definition of a "less than" relation for matrices. Recall, though that variance matrices have the special property of being positive semi-definite (Corollary 5.5 of Chapter 5 of these notes). There is a natural

[4] You also hear people say a lot of other pairs: *sample regression thingummy* and *population regression thingummy* for instances of "thingummy" other than *function* and *coefficients*, such as *equation*, *line* (thinking of the graph of a linear function being a line in the "simple" case of one non-constant predictor), and so forth.

partial order for positive semi-definite matrices. We say $A \leq B$ if $B - A$ is a positive semi-definite matrix.

To understand what this means, look at the proof of why covariance matrices are positive semi-definite. A matrix $\mathbf{M}$ is positive semi-definite if

$$\mathbf{c}'\mathbf{M}\mathbf{c} \geq 0, \qquad \text{for every vector } \mathbf{c}.$$

We also know that for any random vector $\mathbf{X}$ having variance matrix $\mathbf{M}$, the variance of a scalar linear function is given by

$$\operatorname{var}(a + \mathbf{c}'\mathbf{X}) = \mathbf{c}'\mathbf{M}\mathbf{c} \tag{12.27}$$

by (5.19b) from Chapter 5 of these notes. Since variances are nonnegative, this shows $\mathbf{M}$ is positive semi-definite.

Now consider two random vectors $\mathbf{X}$ and $\mathbf{Y}$ with variance matrices $\mathbf{M}_\mathbf{X}$ and $\mathbf{M}_\mathbf{Y}$, respectively. We say that $\mathbf{M}_\mathbf{X} \leq \mathbf{M}_\mathbf{Y}$ if and only if $\mathbf{M}_\mathbf{Y} - \mathbf{M}_\mathbf{X}$ is a positive semi-definite matrix (that's the definition of the partial order). This means

$$\mathbf{c}'(\mathbf{M}_\mathbf{Y} - \mathbf{M}_\mathbf{X})\mathbf{c} \geq 0, \qquad \text{for every vector } \mathbf{c},$$

and this is equivalent to

$$\mathbf{c}'\mathbf{M}_\mathbf{X}\mathbf{c} \leq \mathbf{c}'\mathbf{M}_\mathbf{Y}\mathbf{c}, \qquad \text{for every vector } \mathbf{c},$$

and by (12.27) this is also equivalent to

$$\operatorname{var}(a + \mathbf{c}'\mathbf{X}) \leq \operatorname{var}(a + \mathbf{c}'\mathbf{Y}), \qquad \text{for every vector } \mathbf{c}. \tag{12.28}$$

This characterization tells us what the partial order means. The variance matrices are ordered $\operatorname{var}(\mathbf{X}) \leq \operatorname{var}(\mathbf{Y})$ if and only if the variance of *every scalar-valued linear function* of $\mathbf{X}$ is no greater than the variance of the same function of $\mathbf{Y}$. That's a strong condition!

Now that we have got this rather complicated definition of "best" explained, the theorem itself is very simple.

Theorem 12.6 (Gauss-Markov). *Under the assumptions* (12.22) *and* (12.23), *the least squares estimate* (12.18) *is an unbiased estimate of* $\boldsymbol{\beta}$. *Furthermore it is the best linear unbiased estimate, where "best" means smallest variance.*

In short, the least squares estimate is BLUE.

Proof. The first assertion is that $\hat{\boldsymbol{\beta}}$ given by (12.18) is an unbiased for $\boldsymbol{\beta}$. This is trivial

$$E(\hat{\boldsymbol{\beta}}) = E\left\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\right\} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{Y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

(just linearity of expectation and the definition of matrix inverse).

Consider an unbiased but otherwise completely arbitrary estimate, which will have the form $\boldsymbol{\beta}^* = \mathbf{A}\mathbf{Y}$ for some constant matrix $\mathbf{A}$. (Saying $\mathbf{A}$ is constant means $\mathbf{A}$ can depend on $\mathbf{X}$ but not on $\mathbf{Y}$. Saying $\boldsymbol{\beta}^*$ is an estimate means $\mathbf{A}$ cannot depend on the parameters $\boldsymbol{\beta}$ and σ^2.) It simplifies the proof somewhat if we define

$$\mathbf{B} = \mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

so

$$\boldsymbol{\beta}^* = \left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{B}\right]\mathbf{Y} = \hat{\boldsymbol{\beta}} + \mathbf{B}\mathbf{Y}$$

The condition that $\boldsymbol{\beta}^*$ be unbiased is then

$$\boldsymbol{\beta} = E(\boldsymbol{\beta}^*) = E(\hat{\boldsymbol{\beta}}) + \mathbf{B}E(\mathbf{Y}) = \boldsymbol{\beta} + \mathbf{B}\mathbf{X}\boldsymbol{\beta}$$

which simplifies to

$$\mathbf{B}\mathbf{X}\boldsymbol{\beta} = 0$$

Unbiasedness means this must hold for all possible values of $\boldsymbol{\beta}$. Hence $\mathbf{B}\mathbf{X} = 0$.

Now we calculate

$$\operatorname{var}(\boldsymbol{\beta}^*) = \operatorname{var}(\hat{\boldsymbol{\beta}} + \mathbf{B}\mathbf{Y}) = \operatorname{var}(\hat{\boldsymbol{\beta}}) + \operatorname{var}(\mathbf{B}\mathbf{Y}) + 2\operatorname{cov}(\hat{\boldsymbol{\beta}}, \mathbf{B}\mathbf{Y}) \qquad (12.29)$$

The formula for the variance of a sum here is (5.9) from Chapter 5.

I now claim that the covariance is zero (meaning we haven't proved that *yet*, but we want to look at its consequences to see why it is worth proving), from which the BLUE assertion follows immediately, because then (12.29) becomes

$$\operatorname{var}(\boldsymbol{\beta}^*) = \operatorname{var}(\hat{\boldsymbol{\beta}}) + \operatorname{var}(\mathbf{B}\mathbf{Y})$$

and, since $\operatorname{var}(\mathbf{B}\mathbf{Y})$ like any variance matrix must be positive semi-definite, this implies that $\operatorname{var}(\boldsymbol{\beta}^*) - \operatorname{var}(\hat{\boldsymbol{\beta}})$ is positive semi-definite, which according to the definition of partial order for matrices is the same as $\operatorname{var}(\hat{\boldsymbol{\beta}}) \le \operatorname{var}(\boldsymbol{\beta}^*)$, which is the "$\hat{\boldsymbol{\beta}}$ is best" assertion of the theorem.

Thus we have a proof that is complete except for the unproved claim that the covariance term in (12.29) is zero. So we now prove that claim. Again we calculate, using

$$\operatorname{cov}(\hat{\boldsymbol{\beta}}, \mathbf{B}\mathbf{Y}) = E\left\{\hat{\boldsymbol{\beta}}(\mathbf{B}\mathbf{Y})'\right\} - E(\hat{\boldsymbol{\beta}})E(\mathbf{B}\mathbf{Y})' = E\left\{\hat{\boldsymbol{\beta}}(\mathbf{B}\mathbf{Y})'\right\}$$

because $E(\mathbf{B}\mathbf{Y}) = \mathbf{B}\mathbf{X}\boldsymbol{\beta} = 0$. And

$$\begin{aligned} E\left\{\hat{\boldsymbol{\beta}}(\mathbf{B}\mathbf{Y})'\right\} &= E\left\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{B}'\right\} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{Y}\mathbf{Y}')\mathbf{B}' \end{aligned}$$

Now

$$E(\mathbf{Y}\mathbf{Y}') = \operatorname{var}(\mathbf{Y}) + E(\mathbf{Y})E(\mathbf{Y})' = \sigma^2\mathbf{I} + \mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}'\mathbf{X}'$$

so

$$\begin{aligned} E\left\{\hat{\boldsymbol{\beta}}(\mathbf{B}\mathbf{Y})'\right\} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I} + \mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}'\mathbf{X}')\mathbf{B}' \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{B}' + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}'\mathbf{X}'\mathbf{B}' \end{aligned}$$

And both terms contain $\mathbf{X}'\mathbf{B}' = (\mathbf{B}\mathbf{X})' = 0$. □

This is a very famous theorem. All students should know about it. But for all its supposed theoretical importance, you can't actually do anything with it. It's only good for theoretical woofing. Saying "least squares estimates are BLUE" intimidates people who haven't had a course like this.

The BLUE assertion of the theorem isn't even that important. After all, it doesn't claim that the least squares estimates are *best.* It only claims that they are *best linear unbiased*, which means best among linear and unbiased estimators. Presumably there are nonlinear or biased estimators that are better. Otherwise we could prove a stronger theorem. You have to read between the lines. It sounds like a claim that least squares estimates are the best, but when you decode the qualifications, it actually suggests that they aren't the best.

Moreover, the whole analysis is based on the assumption that the linear model is correct, that the true unknown population regression function is *linear*, that is, has the form (12.25). If the true unknown population regression function is *not* linear, then the least squares estimates are not even unbiased, much less BLUE.

12.3.3 The Sampling Distribution of the Estimates

The Regression Coefficients

We now turn to a much more straightforward problem: what is the sampling distribution of $\hat{\boldsymbol{\beta}}$. In order to have a sampling distribution, we need to specify the whole distribution of the data (not just two moments like we used in the Gauss-Markov theorem). Thus we now switch to the strong linear regression assumptions (12.24).

The least squares estimates are a linear transformation of the data $\mathbf{Y}$ by (12.18), hence if the data are multivariate normal, so are the estimates. A multivariate normal distribution is determined by its mean vector and variance matrix, so we only need to calculate the mean and variance to figure out the distribution.

Theorem 12.7. *Under the assumptions* (12.22) *and* (12.24)

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right). \tag{12.30}$$

Proof. We already showed that $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ under the weak assumptions in the proof of the Gauss-Markov theorem. Since the strong assumptions are stronger, this holds here too.

Now

$$\begin{aligned}
\operatorname{var}(\hat{\boldsymbol{\beta}}) &= \operatorname{var}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\right) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \operatorname{var}(\mathbf{Y}) \left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)' \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \operatorname{var}(\mathbf{Y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}$$

because $\text{var}(\mathbf{Y}) = \text{var}(\mathbf{e}) = \sigma^2 \mathbf{I}$. □

Please note the assumptions of the theorem. If we assume

- the true regression function is linear (12.22) and
- the errors are independent, identically distributed, and *exactly* normally distributed (12.24),

then (12.30) gives the *exact* (not asymptotic) sampling distribution of the least squares estimates. If any of these assumptions are not exactly correct, then it doesn't.

The Error Variance

Unfortunately, the theorem by itself is useless for inference because the distribution contains an unknown parameter σ^2. To make progress, we need an estimate of this parameter and knowledge of its sampling distribution.

If we observed the actual errors e_i, the natural estimate of their variance would be

$$\frac{1}{n} \sum_{i=1}^{n} e_i^2$$

We don't subtract off their mean, because we know $E(e_i) = 0$.

Unfortunately, we do not observe the errors, and must estimate them. Since

$$e_i = y_i - \boldsymbol{\beta}' \mathbf{x}_i$$

the natural estimate is

$$\hat{e}_i = y_i - \hat{\boldsymbol{\beta}}' \mathbf{x}_i \tag{12.31}$$

Because the sample is not the population, these are not the right thing. Hence we should call them not "errors" but "estimated errors." The usual name, however, for (12.31) is *residuals*. We often rewrite (12.31) as a vector equation

$$\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}. \tag{12.32}$$

Plugging the residuals in for the errors in our "natural estimate" gives

$$\frac{1}{n} \hat{\mathbf{e}}' \hat{\mathbf{e}} = \frac{1}{n} \sum_{i=1}^{n} \hat{e}_i^2 \tag{12.33}$$

as a sensible estimate of σ^2, and it turns out this is the MLE (p. 491 in Lindgren). However, this is not the estimator commonly used, because it is biased. The commonly used estimator is

$$\hat{\sigma}^2 = \frac{1}{n - p} \sum_{i=1}^{n} \hat{e}_i^2 \tag{12.34}$$

where p is the number of predictor variables (and regression coefficients). As we shall see, this estimate turns out to be unbiased.

The sum in either of these estimators referred to often enough that it needs a name. It is called the *sum of squares of the residuals* (SSResid) or the *residual sum of squares*

$$\text{SSResid} = \sum_{i=1}^{n} \hat{e}_i^2 = \hat{\mathbf{e}}'\hat{\mathbf{e}}. \tag{12.35}$$

Theorem 12.8. *Under the assumptions* (12.22) *and* (12.24) SSResid *is independent of* $\hat{\boldsymbol{\beta}}$*, and*

$$\frac{\text{SSResid}}{\sigma^2} \sim \text{chi}^2(n-p),$$

where n is the number of observations and p the number of regression coefficients.

From (12.34) and (12.35) we see that the theorem is equivalent to

Corollary 12.9. *Under the assumptions* (12.22) *and* (12.24) $\hat{\sigma}^2$ *is independent of* $\hat{\boldsymbol{\beta}}$*, and*

$$\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \text{chi}^2(n-p),$$

where n and p are as in the theorem.

We will have to defer the proof of the theorem for a bit while we develop a deeper understanding of what linear regression does. First we will look at what we can do with the theorem.

12.3.4 Tests and Confidence Intervals for Regression Coefficients

The main thing we can do with these theorems is make pivotal quantities having Student's t distribution. Recall the definition of Student's t distribution from Section 7.3.5 of these notes: the ratio of a standard normal and the square root of an independent chi-square divided by its degrees of freedom. The vector $\hat{\boldsymbol{\beta}}$ of sample regression coefficients is multivariate normal by Theorem 12.7. To simplify notation define

$$\mathbf{M} = (\mathbf{X}'\mathbf{X})^{-1}.$$

Then the variance of $\hat{\boldsymbol{\beta}}$ is $\sigma^2\mathbf{M}$. Hence a particular sample regression coefficient $\hat{\beta}_k$ has variance $\sigma^2 m_{kk}$ (where, as usual, the elements of $\mathbf{M}$ are denoted m_{ij}). The mean of $\hat{\beta}_k$ is β_k. Thus

$$Z_k = \frac{\hat{\beta}_k - \beta_k}{\sigma\sqrt{m_{kk}}}$$

is standard normal. By Corollary 12.9, $\hat{\sigma}^2/\sigma^2$ is an independent chi-square divided by its degrees of freedom, hence

$$T_k = \frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}\sqrt{m_{kk}}} \sim t(n-p). \tag{12.36}$$

Since the right hand side does not contain unknown parameters, this is a *pivotal quantity*. Hence it can be used for exact confidence intervals and tests about the unknown parameter β_k.

The only difficulty in using this pivotal quantity is calculating the denominator $\hat{\sigma}\sqrt{m_{kk}}$. Because it involves a matrix inverse, there is no simple formula for calculating it. You must use a computer. When using R to do the regression, it always calculates this quantity and prints it out.

Example 12.3.1.
This continues Example 12.2.1. Again we repeat part of the printout from that example

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  94.4803    10.9616   8.619 1.27e-13 ***
x1            0.8503     0.5606   1.517    0.133
x2            1.3599     0.5492   2.476    0.015 *
```

Recall from Example 12.2.1 that we explained that the column labeled `Estimate` from this table gives the sample regression coefficients $\hat{\beta}_k$. Now we explain the rest of the table. The column labeled `Std. Error` gives the denominators $\hat{\sigma}\sqrt{m_{kk}}$ of the t pivotal quantities involving the regression coefficients. The label follows the widespread convention of calling an estimated standard deviation a "standard error." The standard deviation of $\hat{\beta}_k$ is $\sigma\sqrt{m_{kk}}$, which involves an unknown parameter. Estimating it by plugging in $\hat{\sigma}$ for σ gives the *standard error*.

The column labeled `t value` gives the value of the t statistic (12.36) for testing the null hypothesis $\beta_k = 0$. This means that it is the value of (12.36) with zero plugged in for β_k. Let's check this. Looking at the last row, for example $1.3599/0.5492 = 2.476147$, and we see that the third column is indeed the first column divided by the second.

The column labeled `Pr(>|t|)` gives the P-value for the two-tailed test of $\beta_k = 0$ (that is, the alternative is $H_A : \beta_k \neq 0$). Let's also check this. The degrees of freedom of the relevant t distribution are $n - p$, where $n = 100$ and $p = 3$ (there are three regression coefficients including the intercept). Actually, we do not even have to do this subtraction. The degrees of freedom are also given in the R printout in the line

```
Residual standard error: 54.22 on 97 degrees of freedom
```

The P-value corresponding to the t statistic 2.476 in the bottom row of the table is

```
> 2 * (1 - pt(2.476, 97))
[1] 0.01501905
```

and this does indeed agree with the number in the fourth column of the table.

Thus R makes the test with null hypothesis $\beta_k = 0$ easy. It prints the P-value for the two-tailed test and, of course, the P-value for a one-tailed test, if desired, would be half the two-tailed P-value.

The confidence intervals are a bit harder. They have the form

$$\text{estimate} \pm \text{critical value} \times \text{standard error}$$

and all the pieces except the critical value are given in the printout, but that is easily looked up. The critical value for a 95% confidence interval is

```
> qt(0.975, 97)
[1] 1.984723
```

Thus a 95% confidence interval for β_2 (using numbers from the bottom row of the table in the printout) is

```
> 1.3599 + c(-1,1) * 1.984723 * 0.5492
[1] 0.2698901 2.4499099
```

One final warning: with three regression coefficients here, you can do three confidence intervals or three tests. But doing that without correction for multiple testing (Bonferroni correction, for example) is *bogus*. In fact, R's attempt to be helpful by providing the "stars" necessary for "stargazing" is just the bogosity we warned about in Section 9.5.8. So unless there is a strong tradition of stargazing in your scientific subdiscipline, so strong that you just have to do it no matter how bogus, ignore the stars. You can turn off the printing of stars by inserting the command

```
> options(show.signif.stars=FALSE)
```

before the `summary(out)` command.

12.3.5 The Hat Matrix

Also of interest besides the sample regression coefficients is the estimate of the regression function itself

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

The matrix that multiplies $\mathbf{y}$ on the right hand side

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'. \tag{12.37}$$

was dubbed by someone suffering a fit of cuteness the *hat matrix* because it puts the "hat" on $\mathbf{y}$, that is, $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ (actually, I enjoy terminology like this, I just don't care to defend it against stuffed shirts who think scientific terminology should be very serious and boring).

Lemma 12.10. *The hat matrix* (12.37) *is the orthogonal projection onto the subspace spanned by the predictor vectors (the columns of* $\mathbf{X}$*).*

The notion of an orthogonal projection matrix is defined in Section H.1 in Appendix H. Another name for the subspace mentioned in the theorem is just the range of the linear transformation represented by the matrix $\mathbf{X}$, which we write range($\mathbf{X}$). The theorem asserts that this is also the range of the linear transformation represented by the hat matrix $\mathbf{H}$, that is, range($\mathbf{X}$) = range($\mathbf{H}$).

Proof. That the hat matrix is symmetric is obvious from the formula and the rule that the transpose of a matrix product is the product of the transposes in reverse order. That the hat matrix is idempotent is verified by just looking at the formula for $\mathbf{H}^2$.

So the only thing left to verify is that $\mathbf{H}$ actually maps onto range($\mathbf{X}$). We need to show that an arbitrary element of range($\mathbf{X}$), which has the form $\mathbf{X}\boldsymbol{\beta}$ for an arbitrary vector $\boldsymbol{\beta}$, is equal to $\mathbf{H}\mathbf{y}$ for some vector $\mathbf{y}$. It is easily verified that $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$ does the job. □

With this lemma we can finally finish the proof of the theorem that gives t statistics.

Proof of Theorem 12.8. First observe that $\mathbf{H}$ and $\mathbf{I} - \mathbf{H}$ are orthogonal projections that are orthogonal to each other (see Section H.1 in Appendix H for definitions). Define $\mathbf{Z} = \mathbf{e}/\sigma$, then $\mathbf{Z}$ is a multivariate standard normal random vector (that is, the components are i. i. d. standard normal). Theorem H.3 in Appendix H says that $\mathbf{HZ}$ and $(\mathbf{I} - \mathbf{H})\mathbf{Z}$ are independent multivariate normal random vectors and their squared lengths are chi-square random variables. The next step is to see what these vectors are in terms of the variables we have been using.

$$\mathbf{HZ} = \frac{1}{\sigma}\mathbf{He} = \frac{1}{\sigma}\mathbf{H}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \frac{1}{\sigma}(\hat{\mathbf{y}} - \boldsymbol{\mu})$$

where $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, originally defined in (12.16), is the vector of means of the response variables (the fact that $\mathbf{HX}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}$ was verified in the preceding proof). And

$$(\mathbf{I} - \mathbf{H})\mathbf{Z} = \frac{1}{\sigma}(\mathbf{I} - \mathbf{H})\mathbf{e} = \frac{1}{\sigma}(\mathbf{I} - \mathbf{H})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \frac{1}{\sigma}(\mathbf{y} - \hat{\mathbf{y}})$$

Thus we see that

$$\hat{\mathbf{y}} = \boldsymbol{\mu} + \sigma\mathbf{HZ}$$

and

$$\frac{\text{SSResid}}{\sigma^2} = \|(\mathbf{I} - \mathbf{H})\mathbf{Z}\|^2 \tag{12.38}$$

As we said above, Theorem H.3 in Appendix H says that these are independent random variables, and the latter has a chi-square distribution with degrees of freedom rank($\mathbf{I} - \mathbf{H}$).

That almost finishes the proof. There are two loose ends. We were supposed to show that SSResid is independent of $\hat{\boldsymbol{\beta}}$, but what we showed above is that it is independent of $\hat{\mathbf{y}}$. However, $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{y}}$, so independence of $\hat{\mathbf{y}}$ implies independence of $\hat{\boldsymbol{\beta}}$.

The last loose end is that we need to calculate the rank of $\mathbf{I}-\mathbf{H}$. Since $\mathbf{I}-\mathbf{H}$ is the projection on the subspace orthogonally complementary to range($\mathbf{H}$), its rank is $n - p$, where n is the dimension of the whole space and $p = \text{rank}(\mathbf{H})$. Lemma 12.10 asserts that $\text{rank}(\mathbf{H}) = \text{rank}(\mathbf{X})$, which is number of predictor variables. So we are done. □

12.3.6 Polynomial Regression

The reader may have been lead from what has been said so far to think that linear regression is only useful for fitting *linear* regression functions. No! It's much more useful than that. Here is a slogan that captures the issue.

> *It's called "linear regression" because it's linear in the β's, not because it's linear in the x's.*

Here's what the slogan means. Suppose I have a function that is linear in the β's but not linear in the x's, for example

$$h(x) = \beta_1 \sin(x) + \beta_2 \log(x) + \beta_3 x^2.$$

We can put this in linear regression form by simply making up new predictor variables

$$\begin{aligned} x_1 &= \sin(x) \\ x_2 &= \log(x) \\ x_3 &= x^2 \end{aligned}$$

It matters not a bit that these new variables are dependent, all functions of the original predictor variable x. In terms of these "made up" predictor variables our regression function is now linear in both the β's and the x's

$$h(x) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

and is in the form required by the assumptions for linear regression.

> *You can make up as many predictors as you please.*

This section describes one way to "make up predictors."

Example 12.3.2 (Polynomial Regression).
Look at Figure 12.1 which is a plot of some regression data found in the data set at the URL

`http://www.stat.umn.edu/geyer/5102/ex12.3.2.dat`

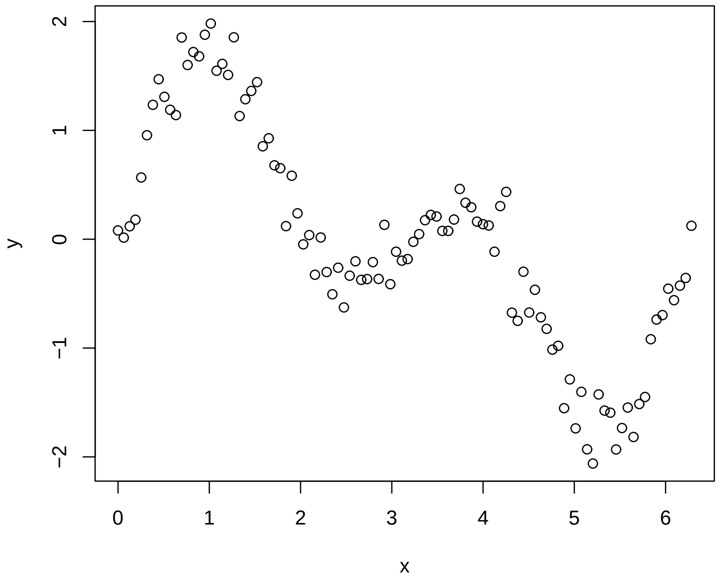


Figure 12.1: Some regression data.

A mere glance at the plot shows that y is decidedly not a linear function of x, not even close. However, the slogan says there is nothing that prevents us from making up more predictor variables. The data set itself has no other variables in it, just x and y. So any new predictor variables we make up must be functions of x. What functions?

Since we are told nothing about the data, we have no guidance as to what functions to make up. In a real application, there might be some guidance from scientific theories that describe the data. Or there might not. Users of linear regression often have no preconceived ideas as to what particular functional form the regression function may have. The title of this section suggests we try a polynomial, that is we want to use a regression function of the form

$$E(Y \mid X) = \sum_{i=0}^{k} \beta_i X^i. \tag{12.39}$$

This is a linear regression function if we consider that we have $k + 1$ different predictor variables $X^0 = 1$, $X^1 = X$, X^2, $\ldots$, X^k. What we have done is "made up" new predictor variables, which are the higher powers of X. Here's how R does it.

```
> out <- lm(y ~ x + I(x^2) + I(x^3) + I(x^4) + I(x^5) + I(x^6))
> summary(out)

Call:
```

```
lm(formula = y ~ x + I(x^2) + I(x^3) + I(x^4) + I(x^5) + I(x^6))

Residuals:
      Min        1Q    Median        3Q       Max
-0.577040 -0.192875 -0.004183  0.196342  0.734926

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.654541   0.192106  -3.407 0.000971 ***
x            7.868747   0.868968   9.055 2.04e-14 ***
I(x^2)      -8.408605   1.228159  -6.847 8.00e-10 ***
I(x^3)       3.334579   0.741135   4.499 1.97e-05 ***
I(x^4)      -0.554590   0.215506  -2.573 0.011651 *
I(x^5)       0.029160   0.029832   0.977 0.330884
I(x^6)       0.000576   0.001577   0.365 0.715765
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1

Residual standard error: 0.308 on 93 degrees of freedom
Multiple R-Squared: 0.9122,     Adjusted R-squared: 0.9065
F-statistic: 160.9 on 6 and 93 degrees of freedom,      p-value:     0
```

The function `I()` is used to give arithmetic expressions their literal meaning in the model formula. If you leave it out, R doesn't do the right thing.

Why start with a sixth degree polynomial? No particular reason. We'll examine this issue later. For now, just accept it.

How do we interpret this mess? The *naive* way is to pay attention to the stars (of course, you wouldn't be that naive now, after all our harping on the bogosity of stargazing, would you?). They seem to say that the coefficients up the x^4 term are statistically significant, and the coefficients of the two higher powers are not. So we should try a fourth degree polynomial next.

Stargazing violates the "do *one* test" dogma. To do the right thing we must do only one test. The obvious coefficient to test is the one for the highest power of x. Clearly it is not statistically significant. Thus we can accept the null hypothesis of that test and set the corresponding regression coefficient equal to zero. And that is the end of the conclusions we can draw from this regression!

However, that conclusion leaves us with a new model to fit. The part of the printout about the regression coefficients for that model is

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.687409   0.168940  -4.069 9.83e-05 ***
x            8.112834   0.552844  14.675  < 2e-16 ***
I(x^2)      -8.808811   0.552153 -15.954  < 2e-16 ***
I(x^3)       3.592480   0.224068  16.033  < 2e-16 ***
I(x^4)      -0.631962   0.039391 -16.043  < 2e-16 ***
I(x^5)       0.040017   0.002495  16.039  < 2e-16 ***
```

Surprise! The coefficient of x^5 wasn't statistically significant before, but now it is. In fact, it wasn't even close to significance before, and now it is extremely significant. What's happening?

A long answer could get very complicated. All of the β's are correlated with each other, so it is very hard to tell what is going on. A short answer is that we shouldn't have been surprised. None of the theory we have developed so far gives any positive reason why this can't happen. An even shorter answer is the following slogan.

> *If you want to know anything about a model, you* ***must*** *fit that model. You can't tell* ***anything*** *about one model by looking at the regression output for* ***some other model.***

Guessing that the coefficient of x^5 wouldn't be significant in this model from looking at the output of the other model is a mugs game. If that's not a clear example of the bogosity of stargazing, I don't know what is.

Now let us add the sample regression function to the plot. The vector $\hat{\mathbf{y}}$, which is the sample regression function evaluated at the predictor values in the data set is in the component `fitted.values` of the list returned by `lm` function (what we usually store in the variable `out`). The R commands

```
> out <- lm(y ~ x + I(x^2) + I(x^3) + I(x^4) + I(x^5))
> plot(x, y)
> lines(x, out$fitted.values)
```

Figure 12.2 shows this plot. The fit isn't bad, but it isn't great either. I think I could draw a better fitting curve by hand. How can that be? Isn't linear regression BLUE? How can I do better than the "best?" Linear regression is BLUE only if the model assumptions are true. In particular, in this case, its BLUE only if the true unknown regression function is a polynomial of degree five. Doesn't appear to be. So much for the optimality of linear regression. It's only optimal in toy problems for which the answer is known. For real data where you don't know the true population regression function, it isn't.

Before leaving this section we should ask and answer the following question.

> *What's so special about polynomials? Nothing!*
> *Made up predictors can be* ***any*** *functions of the original predictors.*

Problem 12-4 explores using sines and cosines instead of polynomials.

We revisit the issue we started with, just to make sure everything is clear: is this linear regression or not? It was certainly done with a linear regression computer program, and looked at abstractly enough, it is linear regression. As we explained at the beginning (12.39) is in the form of a linear population regression function, if we consider it a function of $k + 1$ variables, X^0, …, X^k instead of just the one variable X. But if we consider it a function of just one variable X, the graph of which is the line in Figure 12.2, it isn't linear. Thus we see that linear regression is more versatile than it appears at first sight. It also does nonlinear regression by making it a special case of linear regression (quite a trick, that).

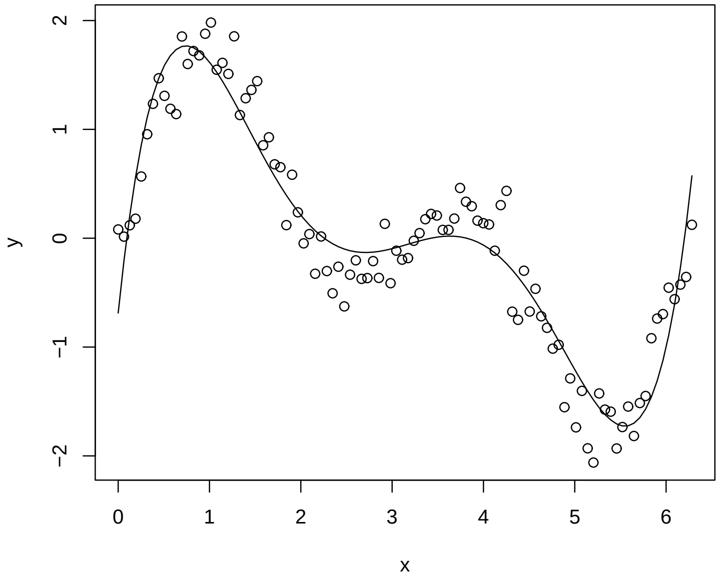


Figure 12.2: The same regression data plotted in Figure 12.1 with the best fitting polynomial of degree five added.

12.3.7 The F-Test for Model Comparison

This section deals with the regression analog of the likelihood ratio test. Suppose we have two nested regression models, and we want do a test comparing them. The null and alternative hypotheses are exactly as extensively discussed in Section 10.4.3 (Likelihood Ratio Tests). We could, in fact, just use the likelihood ratio test exactly as described in that section. It would provide an asymptotically valid test, approximately correct when the sample size is large. However, likelihood ratio tests are not traditionally used in regression. What is used, what we will develop in this section, are exact tests in which the test statistic has an F distribution. For large sample sizes, these F tests give P-values very close to the likelihood ratio test. The difference is with small sample sizes, where the likelihood ratio test is not valid, but the F tests are valid under the "strong" assumption that the errors are i. i. d. normal. (The F tests are, of course, not exact when that assumption is violated.)

Before we can state the theorem we need to see what the condition that models be "nested" means in the regression context. As with many other things about linear regression, there is a simple notion, useful when you are actually doing regression, and a mathematically sophisticated notion, useful in proofs. The simple notion is that the little model is just like the big model except that some of the regression coefficients in the big model are fixed in the little model, usually at zero.

For example, if the models under consideration are the polynomial regression models considered in the preceding section, then the big model might be all sixth degree polynomials, having regression functions of the form (12.39) with $k = 6$ and the little model might be all third degree polynomials having regression functions of the same form with $k = 3$. The little model is clearly obtained from the big model by setting $\beta_4 = \beta_5 = \beta_6 = 0$. Also clearly, the big model has three more parameters than the little model.

The mathematically sophisticated notion is that if $\mathbf{X}_{\text{big}}$ is the design matrix of the big model and $\mathbf{X}_{\text{little}}$ is the design matrix of the little model, then the models are nested if $\operatorname{range}(\mathbf{X}_{\text{little}}) \subset \operatorname{range}(\mathbf{X}_{\text{big}})$, or, what is equivalent because we know the range of the design matrix is the same as the range of the hat matrix, $\operatorname{range}(\mathbf{H}_{\text{little}}) \subset \operatorname{range}(\mathbf{H}_{\text{big}})$, where $\mathbf{H}_{\text{little}}$ and $\mathbf{H}_{\text{big}}$ are the corresponding hat matrices.

While we are at it, we generalize to a sequence of nested models. Suppose $\mathbf{H}_i$, $i = 1, \ldots, k$ are the hat matrices of a sequence of regression models. Then we say the sequence is *nested* if

$$\operatorname{range}(\mathbf{H}_i) \subset \operatorname{range}(\mathbf{H}_{i+1}), \qquad i = 1, \ldots, k-1 \tag{12.40}$$

Theorem 12.11. *Let* SSResid$_i$ *denote the residual sum of squares for the i-th model in a sequence of k nested regression models. Assume the smallest model is true, that is,*

$$E(Y) = \mathbf{X}_1 \boldsymbol{\beta}$$

where $\mathbf{X}_1$ is the design matrix for the smallest model, and assume the errors satisfy (12.24)*, then* $\text{SSResid}_k/\sigma^2$ *and*

$$\frac{\text{SSResid}_i - \text{SSResid}_{i+1}}{\sigma^2}, \qquad i = 1, \ldots, k-1$$

are independent random variables, and

$$\frac{\text{SSResid}_k}{\sigma^2} \sim \text{chi}^2(n - p_k) \tag{12.41a}$$

and

$$\frac{\text{SSResid}_i - \text{SSResid}_{i+1}}{\sigma^2} \sim \text{chi}^2(p_{i+1} - p_i), \tag{12.41b}$$

where p_i is the dimension (number of regression coefficients) of the i-th model.

Proof. Assertion (12.41a) does not have to be proved, since it is just the assertion of Theorem 12.8 applied to the k-th model. In the proof of Theorem 12.8, in equation (12.38) we derived a formula giving SSResid in terms of the hat matrix. If we add subscripts to make it apply to the current situation, it becomes

$$\frac{\text{SSResid}_i}{\sigma^2} = \|(\mathbf{I} - \mathbf{H}_i)\mathbf{Z}\|^2 \tag{12.42}$$

where as in the proof of Theorem 12.8, we are defining $\mathbf{Z} = \mathbf{e}/\sigma$. Now note that

$$\begin{aligned} \|(\mathbf{H}_{i+1} - \mathbf{H}_i)\mathbf{Z}\|^2 &= \mathbf{Z}'(\mathbf{H}_{i+1} - \mathbf{H}_i)^2\mathbf{Z} \\ &= \mathbf{Z}'(\mathbf{H}_{i+1}^2 - \mathbf{H}_{i+1}\mathbf{H}_i - \mathbf{H}_i\mathbf{H}_{i+1} + \mathbf{H}_i^2)\mathbf{Z} \\ &= \mathbf{Z}'(\mathbf{H}_{i+1} - \mathbf{H}_i)\mathbf{Z} \\ &= \mathbf{Z}'(\mathbf{I} - \mathbf{H}_i)\mathbf{Z} - \mathbf{Z}'(\mathbf{I} - \mathbf{H}_{i+1})\mathbf{Z} \\ &= \frac{\text{SSResid}_i - \text{SSResid}_{i+1}}{\sigma^2} \end{aligned} \tag{12.43}$$

where in the middle we use the fact that the hat matrices are idempotent and $\mathbf{H}_{i+1}\mathbf{H}_i = \mathbf{H}_i\mathbf{H}_{i+1} = \mathbf{H}_i$, which comes from Lemma H.1 in Appendix H.

Now we want to apply Theorem H.3 in the same appendix. This will show both the asserted independence and the chi-square distributions if we can show the following. The matrices

$$\mathbf{I} - \mathbf{H}_k \tag{12.44a}$$

and

$$\mathbf{H}_{i+1} - \mathbf{H}_i, \qquad i = 1, \ldots, k-1 \tag{12.44b}$$

are an orthogonal set of orthogonal projections, and

$$\operatorname{rank}(\mathbf{H}_{i+1} - \mathbf{H}_i) = p_{i+1} - p_i. \tag{12.44c}$$

Note that we can avoid treating (12.44a) as a special case by defining $H_{k+1} = \mathbf{I}$.

First we have to show that $\mathbf{H}_{i+1} - \mathbf{H}_i$ is an orthogonal projection. It is clearly symmetric, and idempotence was already shown in (12.43).

Then we have to show

$$(\mathbf{H}_{i+1} - \mathbf{H}_i)(\mathbf{H}_{j+1} - \mathbf{H}_j) = 0, \qquad i < j.$$

This also follows directly from Lemma H.1 in Appendix H.

$$\begin{aligned} (\mathbf{H}_{i+1} - \mathbf{H}_i)(\mathbf{H}_{j+1} - \mathbf{H}_j) &= \mathbf{H}_{i+1}\mathbf{H}_{j+1} - \mathbf{H}_{i+1}\mathbf{H}_j - \mathbf{H}_i\mathbf{H}_{j+1} + \mathbf{H}_i\mathbf{H}_j \\ &= \mathbf{H}_{i+1} - \mathbf{H}_{i+1} - \mathbf{H}_i + \mathbf{H}_i \end{aligned}$$

The only bit remaining to prove is (12.44c). Note that $p_i = \operatorname{rank}(\mathbf{H}_i)$, so this is the same thing as

$$\operatorname{rank}(\mathbf{H}_{i+1} - \mathbf{H}_i) = \operatorname{rank}(\mathbf{H}_{i+1}) - \operatorname{rank}(\mathbf{H}_i)$$

By definition $\operatorname{rank}(\mathbf{H}_{i+1} - \mathbf{H}_i)$ is the dimension of $\operatorname{range}(\mathbf{H}_{i+1} - \mathbf{H}_i)$. We now claim that this is the orthogonal complement of $\operatorname{range}(\mathbf{H}_i)$ in $\operatorname{range}(\mathbf{H}_{i+1})$. Consider an arbitrary vector $\mathbf{y}$ in $\operatorname{range}(\mathbf{H}_{i+1})$. Then

$$(\mathbf{H}_{i+1} - \mathbf{H}_i)\mathbf{y} = \mathbf{y} - \mathbf{H}_i\mathbf{y}$$

which is a vector orthogonal to $\operatorname{range}(\mathbf{H}_i)$. Since every vector in $\operatorname{range}(\mathbf{H}_{i+1} - \mathbf{H}_i)$ is orthogonal to every vector in $\operatorname{range}(\mathbf{H}_i)$, this implies that a basis for one is orthogonal to a basis for the other. Hence the union of the bases is a basis for $\operatorname{range}(\mathbf{H}_{i+1})$ and the dimensions add, which is what was to be proved. □

Corollary 12.12. *If* SSResid_{little} *and* SSResid_{big} *are the residual sums of squares for nested models of dimension* p_{little} *and* p_{big}*, respectively, and the "strong" model assumptions* (12.22) *and* (12.24) *hold for the little model, then*

$$\frac{\text{SSResid}_{little} - \text{SSResid}_{big}}{p_{big} - p_{little}} \cdot \frac{n - p_{big}}{\text{SSResid}_{big}} \sim F(p_{big} - p_{little}, n - p_{big}).$$

Proof. The theorem says the two random variables involving residual sums of squares are independent chi-square random variables. Dividing each by its degrees of freedom and forming the ratio makes an F random variable. □

Example 12.3.3 (Multivariable Polynomial Regression).
Let us consider whether a quadratic model or higher polynomial would fit the data of Example 12.2.1 better than the linear model used in that example. The most general quadratic model has six terms

$$E(Y \mid x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_1 x_2 + \beta_5 x_2^2$$

The R command to fit this model is

```
out <- lm(y ~ x1 + x2 + I(x1^2) + I(x1 * x2) + I(x2^2))
summary(out)
```

The part of the output that describes the regression coefficients is shown below.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 91.011781  16.984368   5.359 5.95e-07 ***
x1           0.460700   1.540967   0.299    0.766
x2           1.921907   1.399151   1.374    0.173
I(x1^2)      0.013292   0.052642   0.252    0.801
I(x1 * x2)  -0.020873   0.097794  -0.213    0.831
I(x2^2)      0.005785   0.047867   0.121    0.904
```

It says, if we are naive enough to believe the "stars" (which of course we aren't), that none of the regression coefficients except the one for the constant predictor is interesting. Of course this contradicts Example 12.2.1 where we found that the coefficient of x_2 was "significant" (yet another case illustrating how misleading stargazing is).

In order to compare this quadratic model with the linear model fit in Example 12.2.1 we should do the F-test described in this section. R provides a way to do this easily. First fit the two models, saving both results.

```
out.lin <- lm(y ~ x1 + x2)
out.quad <- lm(y ~ x1 + x2 + I(x1^2) + I(x1 * x2) + I(x2^2))
```

Then the function `anova` computes a so-called "analysis of variance" (ANOVA) table for the model comparison.

```
anova(out.lin, out.quad)
```

The output of the `anova` function is

```
Analysis of Variance Table

Model 1: y ~ x1 + x2
Model 2: y ~ x1 + x2 + I(x1^2) + I(x1 * x2) + I(x2^2)
  Res.Df Res.Sum Sq Df Sum Sq F value Pr(>F)
1     97     285119
2     94     284370  3    749  0.0825 0.9694
```

The last three lines of the printout here are a so-called "analysis of variance" table. Back in the stone age, when people calculated this stuff without computers, someone decided it was helpful to lay out the arithmetic in calculating the F statistic this way. Nowadays only the final result $F = 0.0825$ and the P-value of the test $P = 0.9694$ are interesting. But these tables give old timers a warm fuzzy feeling, so computers still print them out.

Since R calculates everything, there is nothing left for you to do except interpreting the P-value. Low P-values are evidence in favor of the alternative, high P-values in favor of the null. This one is certainly high, much higher than one would expect by chance if the null hypothesis is true. Thus we accept the null hypothesis (here, the linear model). We say it fits just as well as the quadratic model. The extra terms in the quadratic model add no predictive or explanatory value.

Thus we should examine the linear model having only x_1, x_2, and the constant predictor. But we already did this in Example 12.3.1. The printout for that example apparently shows that x_1 can also be dropped.

12.3.8 Intervals for the Regression Function

Confidence Intervals

An important problem is estimating the regression function itself, either at some specified $\mathbf{x}$ value, or at all $\mathbf{x}$ values. As everywhere else the sample is not the population. What we want to know is

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta},$$

the population mean vector. But as the Greek letters indicate, we don't know $\boldsymbol{\beta}$ hence don't know $\mathbf{X}\boldsymbol{\beta}$. What we do know is the corresponding sample quantity

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}},$$

the vector of "predicted values." What is the relation between the two?

More generally, we can consider the regression function as a function of the predictor vector $\mathbf{x} = (x_1, \ldots, x_p)$

$$E(Y \mid \mathbf{x}) = \sum_{i=1}^{p} \beta_i x_i = \mathbf{x}'\boldsymbol{\beta}$$

that we can evaluate at arbitrary $\mathbf{x}$ values, not just at so-called "design points" ($\mathbf{x}$ values occurring in the data set being analyzed). If we write this function as

$$h(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}, \tag{12.45}$$

it is obvious that the most sensible estimate is

$$\hat{h}(\mathbf{x}) = \mathbf{x}'\hat{\boldsymbol{\beta}}. \tag{12.46}$$

Before we become bogged down in details, it is worthwhile to get the big picture firmly in mind. The sample regression coefficient vector $\hat{\boldsymbol{\beta}}$ is multivariate normal and independent of the error sum of squares SSResid. The regression function estimate (12.46), being a linear transformation of a multivariate normal random vector, is a normal random scalar. Hence we can combine it with the error sum of squares to make t-confidence intervals and t-tests. All we need to do is work out the details.

We have already said that (12.46) is normal. Clearly its mean is (12.45). Hence it is an unbiased estimate of the population regression function. By Corollary 5.4 in Chapter 5 of these notes, the variance of (12.46) is $\mathbf{x}' \operatorname{var}(\hat{\boldsymbol{\beta}})\mathbf{x}$, and plugging in the variance of the sample regression coefficient vector from Theorem 12.7 gives

$$\operatorname{var}(\mathbf{x}'\hat{\boldsymbol{\beta}}) = \sigma^2 \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}. \tag{12.47}$$

If you are confused about what big $\mathbf{X}$ and little $\mathbf{x}$ are here, $\mathbf{X}$ is the design matrix for the original data set and $\mathbf{x}$ is one possible value of the predictor vector. If $\mathbf{x}$ is a value that occurs in the original data, then it is one row of the design matrix $\mathbf{X}$, otherwise $\mathbf{X}$ and $\mathbf{x}$ are unrelated. The vector $\mathbf{x}$ can be any vector of predictor values for any individual, whether one in the original data set or some other individual.

Of course, we have the usual problem that we don't know σ^2 and have to plug in the estimate $\hat{\sigma}^2$ given by (12.34). This gives a t confidence interval

$$\mathbf{x}'\hat{\boldsymbol{\beta}} \pm t_{\alpha/2}\hat{\sigma}\sqrt{\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}}$$

where, as usual, the degrees of freedom for the t distribution used to calculate the critical value is $n - p$.

Fortunately, we don't need to do any of these calculations. R has a function `predict` that does them all. Suppose we have a data set with three predictors `x1`, `x2`, and `x3`, and a response `y`, then, as usual,

```
out <- lm(y ~ x1 + x2 + x3)
```

fits the model. Now

```
predict(out, data.frame(x1=1, x2=1, x3=1), interval="confidence")
```

produces a 95% confidence interval for the value of the population regression function at $\mathbf{x} = (1, 1, 1)$. The output is (for data which are not shown)

```
          fit      lwr      upr
[1,] 3.623616 2.022624 5.224608
```

The component labeled `fit` is the estimated value of the population regression function $\mathbf{x}'\hat{\boldsymbol{\beta}}$. The component labeled `lwr` is the lower endpoint of the 95% confidence interval and the component labeled `upr` is the upper endpoint. For different confidence level use the optional argument `level`, for example,

```
predict(out, data.frame(x1=1, x2=1, x3=1), interval="confidence",
level=.90)
```

Prediction Intervals

Actually, one rarely wants a confidence interval of the kind described in the preceding section. One usually wants a very closely related interval called a *prediction interval*. The idea is this. What is the point of knowing the population regression $E(Y \mid \mathbf{x})$? It gives BLUP (best linear unbiased predictions) for the response Y (with the usual proviso, the model must be correct). However, these predictions, even if they used the true population regression function would not be exactly correct, because Y is observed "with error." If we write $\mu = E(Y \mid \mathbf{x})$, then $Y \sim \mathcal{N}(\mu, \sigma^2)$ under the "strong" linear regression assumptions. So our "best" estimate will be wrong by about σ (the error standard deviation), sometimes more, sometimes less, because Y is a random variable.

Of course, we don't know the population regression function and are forced to substitute the sample regression function. We can write

$$Y = \mathbf{x}'\boldsymbol{\beta} + e$$

for the population regression model, but we can't use that. Let us rewrite this

$$Y = \mathbf{x}'\hat{\boldsymbol{\beta}} + \mathbf{x}'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + e.$$

The first term on the right, is what we use for prediction. The second two terms are unknown errors (the middle term being unknown because we don't know the true population regression coefficient vector $\boldsymbol{\beta}$). However we do know the sampling distribution of the sum of the two terms on the right. Being a linear transformation of jointly normal random variables, it is normal with mean

$$E\{\mathbf{x}'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + e\} = 0$$

and variance

$$\operatorname{var}\{\mathbf{x}'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + e\} = \operatorname{var}\{\mathbf{x}'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\} + \operatorname{var}(e) = \sigma^2 + \sigma^2 \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x},$$

where we used (12.47) to get the variance of $\mathbf{x}'\hat{\boldsymbol{\beta}}$. Here the first equality assumes that e and $\hat{\boldsymbol{\beta}}$ are independent random variables, which will be the case if Y here refers to a "new" individual, *not* one in the original data set used to calculate the sample regression function. Thus the "prediction" interval is almost exactly like the "confidence" interval with just a slight change in the formula

$$\mathbf{x}'\hat{\boldsymbol{\beta}} \pm t_{\alpha/2}\hat{\sigma}\sqrt{1 + \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}}$$

The "1+" in the square root is only difference between the two formulas. R also makes this interval convenient to calculate. Just do the same thing as for the confidence interval but use `interval="prediction"` as the argument specifying the type of interval wanted.

Example 12.3.4.
This continues Example 12.3.2. What we want to do here is add lines indicating the prediction intervals to Figure 12.2. The following code

```
out <- lm(y ~ x + I(x^2) + I(x^3) + I(x^4) + I(x^5))
pout <- predict(out, data.frame(x=x), interval="prediction")
plot(x, y, ylim=range(pout))
lines(x, out$fitted.values)
lines(x, pout[ , "lwr"], lty=2)
lines(x, pout[ , "upr"], lty=2)
```

(This uses a bit of magic of optional arguments. The `ylim=range(pout)` argument to the `plot` command leaves room for the confidence intervals. The `lty=2` says to use a line type different from the default. Supplying a whole vector `data.frame(x=x)` to the `predict` function, produces all the prediction intervals in one statement. Using labels as subscripts, as in `pout[ , "lwr"]` is another R idiom we won't try to explain.)

12.4 The Abstract View of Regression

> *Regression coefficients are meaningless. Only regression functions and fitted values are meaningful.*

The idea of a regression problem is to estimate a regression function. When there are several predictors, there is no unique way to express the regression function as a linear function of the predictors.

> *Any linearly independent set of linear combinations of predictor variables makes for an equivalent regression problem.*

Suppose $\mathbf{X}$ is a design matrix for a regression problem. The columns of $\mathbf{X}$ correspond to the predictor variables. Using linear combinations of predictors is like using a design matrix

$$\mathbf{X}^* = \mathbf{X}\mathbf{A},$$

where A is an invertible $p \times p$ matrix (where, as usual, there are p predictors, including the constant predictor, if there is one). The requirement that $\mathbf{A}$ be invertible is necessary so that $\mathbf{X}^*$ will have rank p if $\mathbf{X}$ does. Then

$$(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1} = [(\mathbf{X}\mathbf{A})'(\mathbf{X}\mathbf{A})]^{-1} = [\mathbf{A}'\mathbf{X}'\mathbf{X}\mathbf{A}]^{-1} = \mathbf{A}^{-1}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{A}')^{-1}$$

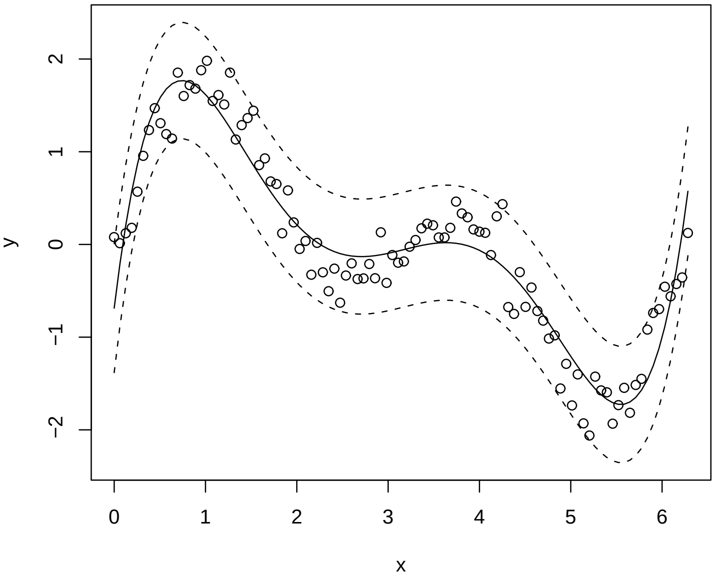


Figure 12.3: The same regression data plotted in Figure 12.1 with the best fitting polynomial of degree five added and pointwise prediction intervals.

because the inverse of a product is the product of the inverses in reverse order. (We can't apply this rule to $\mathbf{X}$ itself because $\mathbf{X}$ is not invertible. Only the product $\mathbf{X}'\mathbf{X}$ is invertible).

The regression coefficients for the "starred problem" are different

$$\begin{aligned} \hat{\boldsymbol{\beta}}^* &= (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{y} \\ &= \mathbf{A}^{-1}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{A}')^{-1}\mathbf{A}'\mathbf{X}'\mathbf{y} \\ &= \mathbf{A}^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \mathbf{A}^{-1}\hat{\boldsymbol{\beta}} \end{aligned}$$

because $(\mathbf{A}')^{-1}\mathbf{A}'$ is the identity and using the definition (12.18) of $\hat{\boldsymbol{\beta}}$.

Although the *regression coefficients* are different, the *fitted values* are not different!

$$\hat{\mathbf{y}}^* = \mathbf{X}^*\hat{\boldsymbol{\beta}}^* = \mathbf{X}\mathbf{A}\mathbf{A}^{-1}\hat{\boldsymbol{\beta}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\mathbf{y}}$$

This means the hat matrices for the two problems are also the same (as is easily checked).

Since F tests for model comparison depend only on residual sums of squares, which depend only on fitted values, no F test is changed by replacing an "unstarred" problem with a "starred" problem in the manner described above. Nothing of statistical significance changes. There is no such thing as a test of whether the "starred" or the "unstarred" model fit the data better. Both always fit exactly the same, no better and no worse. They have the same residual sum of squares. In a sophisticated abstract sense they are two equivalent descriptions of the *same* model.

But when you look at the regression coefficients, for example, at the table the R `summary` command prints out giving the coefficients, standard errors, t statistics, and P-values, the "starred" and "unstarred" models look *very* different. The regression coefficients seem to have nothing to do with each other (though, of course, they are actually related by the linear relationship $\boldsymbol{\beta}^* = \mathbf{A}^{-1}\boldsymbol{\beta}$, this is impossible to visualize if $\mathbf{A}$ has complicated structure).

So whenever you see someone taking regression coefficients very seriously, remember that they are actually meaningless. The discussion would be better phrased in terms of prediction, predicted (fitted) values, and regression functions.

> *Regression is for prediction, not for explanation.*

Of course, what scientists mostly want from regression is explanation not prediction. But what we are saying that what they want and what regression actually delivers are two different things.

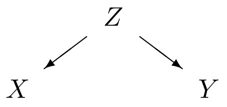
Another related slogan that is a bit off the subject, but worth throwing into the discussion for the sake of completeness, is

> *Correlation is not causation, and regression isn't either.*

What is meant by the slogan "correlation is not causation" is that mere correlation doesn't show a causative relationship between variables. This is clear from the fact that correlation is a symmetric relation (the correlation of x and y is the same as the correlation of y and x), but causal relationships are not symmetric ("x causes y" is not the same as "y causes x"). If we denote causal relationships by arrows, there are two possible causal relationships involving x and y

$$X \longrightarrow Y \qquad \text{or} \qquad X \longleftarrow Y$$

If we admit other variables into consideration, there are many possible causal relationships, for example



Here neither "x causes y" nor "y causes x" holds. Both are controlled by a third variable z. So mere existence of a correlation does not entitle us to say anything about underlying causal relationships.

Now regression is just correlation looked at from a different angle. This is clear in the "simple" case (one non-constant predictor) where the slope of the regression line $\hat{\beta}$ is related to the correlation coefficient r by

$$\hat{\beta} = r \frac{s_y}{s_x}.$$

In general, the relationship between regression and correlation is less transparent, but the regression coefficients and fitted values are functions of the sample first and second moments of the response and predictor variables (including covariances). This is clear from the formulation of least squares estimates as functions of "empirical" second moments given (12.13), (12.14), and (12.15).

Regression does not "explain" the relationship between the variables. There is no way to tell which way the causal arrow goes between the variables, or even if there is any direct causal relationship. What scientists *want* is to find causal relationships. Often, as in many social sciences, there is no way do do controlled experiments and regression is the only tool available to explore relationships between variables. Scientists want so hard to find causal relationships that they often forget the slogans above (or pay lip service to them while ignoring their content).

There is even a school of regression use called *causal modeling* that claims to use regression and similar tools to find causal relationships. But the theorems of that school are of the "ham and eggs" variety (if we had some ham, we'd have ham and eggs, if we had some eggs). First they assume there are *no* unmeasured variables that have any causal relationships with *any* measured variables (predictor or response). That is, they assume there are no variables like Z in the picture above involving three variables. Then they assume that there are non-statistical reasons for deciding which way the arrow goes in the other picture. Then they have a theorem that says causal relationships can be determined. But in the "simple" case (only two variables X and Y) this is a pure tautology

> *If we assume X causes Y, then we can conclude X causes Y*

(well, duh!) When there are more than two variables, so-called causal modeling can yield conclusions that are not purely tautological, but they are always based on exceedingly strong assumptions (no unmeasured variables with causal connection to measured variables) that are always known to be false without a doubt. There is no real escape from "correlation is not causation, and regression isn't either."

12.5 Categorical Predictors (ANOVA)

> *ANOVA is just regression with all predictor variables categorical.*

12.5.1 Categorical Predictors and Dummy Variables

When a predictor variable is categorical, there is no sense in which there can be one regression coefficient that applies to the variable. If x_1 is a variable taking values in the set $\{\text{Buick}, \text{Toyota}, \text{Mercedes}\}$, then $\beta_1 x_1$ makes no sense because the values of x_1 are not numbers. However, categorical predictor variables are easily incorporated into the regression framework using the device of so-called *dummy variables*.

If x is a categorical predictor variable taking values in an arbitrary set S, then the *dummy variables* associated with x are the indicator random variables (zero-one valued)

$$I_{\{s\}}(x), \qquad s \in S.$$

For example, if we have a predictor variable x associated with the predictor vector

$$\mathbf{x} = \begin{pmatrix} \text{Buick} \\ \text{Buick} \\ \text{Mercedes} \\ \text{Buick} \\ \text{Toyota} \\ \text{Mercedes} \end{pmatrix}$$

then this is associated with three dummy variable vectors that make up three columns of the design matrix

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{12.48}$$

the first column is the indicator of the category *Buick*, the second column the indicator of the category *Toyota*, the third column the indicator of the category *Mercedes*.

Suppose we fit a model with this design matrix (no constant predictor). Call the three predictor vectors, the columns of (12.48), $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$. Then the regression model is

$$\mathbf{y} = \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \beta_3 \mathbf{x}_3 + \mathbf{e}$$

Note that exactly one of the three predictor vectors is nonzero, that is, if we write the scalar equations with one more subscript,

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i,$$

then this is the same as

$$y_i = \beta_k + e_i,$$

if the i-th individual is in category k. Thus the regression coefficient β_k is just the population mean for category k.

For technical reasons, to be explained presently, we often drop one of the predictors (it doesn't matter which) and add a constant predictor. This gives us a different design matrix. If we drop the *Mercedes* dummy variable in the example, this gives us a design matrix

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \tag{12.49}$$

Now the first column is the constant predictor, the second column is the *Buick* predictor and the third is the *Toyota* predictor.

Although this seems at first sight to change the model, in the abstract sense discussed in the preceding section, it does not. The constant predictor is the sum of the rows of the original design matrix (12.48). Thus (12.48) and (12.49) are abstractly equivalent design matrices: "starred" and "unstarred" matrices in the notation of the preceding section. The two models both fit three parameters which determine estimates of the population means for the three categories. In the representation using design matrix (12.48) the regression coefficients *are* just the sample means for the three categories. In the other representation, they aren't, but the *fitted value* for each individual *will* be the sample mean for the individual's category.

Categorical predictors are so important that R makes it easy to fit models involving them

```
x <- c("Buick", "Buick", "Mercedes", "Buick", "Toyota", "Mercedes")
y <- c(0.9, 1.0, 1.9, 1.1, 3.0, 2.1)
xf <- factor(x)
out <- lm(y ~ xf)
summary(out)
```

The first two statements define the predictor `x` and the response `y`. The last two are the usual R commands for fitting a regression model and printing out various information about it. The only new wrinkle is the command in the middle that makes a special "factor" variable `xf` that will be dealt with in the right way by the `lm` command. We don't have to set up the dummy variables ourselves. R will do it automagically whenever a "factor" (i. e., categorical) variable appears in the regression formula. The reason why we have to run the original predictor variable `x` through the `factor` function is because if the category labels are numeric (instead of text as in this example) there is no way for R to tell whether we want the variable treated as quantitative or categorical unless we tell it. The `factor` function is the way we tell R we want a variable treated as categorical. We could also have compressed the two lines above involving the `factor` and `lm` functions into one

```
three <- lm(y ~ factor(x))
```

Either way, we get the following table of regression coefficients in the output of the `summary` command. Only the labels of the regression coefficients differ. All the numbers are identical.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.00000    0.06667  15.000 0.000643 ***
xfMercedes   1.00000    0.10541   9.487 0.002483 **
xfToyota     2.00000    0.13333  15.000 0.000643 ***
```

These regression coefficients are not easy to interpret. Their interpretation depends on the actual design matrix R uses, which is neither of the design matrices (12.48) or (12.49) described above. However, we shouldn't let this bother us in the slightest. The slogan at the beginning of the preceding section tells us that regression coefficients are meaningless anyway. They are especially meaningless here. What is important are the fitted values

```
> predict(out)
1 2 3 4 5 6
1 1 2 1 3 2
```

Comparing this with the definition of x. We see that individuals 1, 2, and 4 are in the *Buick* category. All have predicted value 1. Thus that is the sample mean for the *Buick* category. Similarly, the sample means for the *Toytota* and *Mercedes* categories are 3 and 2, respectively.

If we actually wanted to force the regression coefficients to be the sample means, we could do that.

```
  x1 <- as.numeric(x == "Buick")
  x2 <- as.numeric(x == "Toyota")
  x3 <- as.numeric(x == "Mercedes")
  out.too <- lm(y ~ x1 + x2 + x3 + 0)
  summary(out.too)
```

Gives the output

```
Coefficients:
   Estimate Std. Error t value Pr(>|t|)
x1  1.00000    0.06667   15.00 0.000643 ***
x2  3.00000    0.11547   25.98 0.000125 ***
x3  2.00000    0.08165   24.50 0.000149 ***
```

The design matrix for this regression is (12.48)

```
> cbind(x1, x2, x3)
     x1 x2 x3
[1,]  1  0  0
[2,]  1  0  0
[3,]  0  0  1
[4,]  1  0  0
[5,]  0  1  0
[6,]  0  0  1
```

But we don't need to worry about this because "regression coefficients are meaningless." Either regression gives the same predicted values. They agree about every statistically meaningful quantity.

Now we return to the promised explanation of the technical reason why a design matrix like (12.49) is preferred to one like (12.48). Suppose we have *two* categorical predictors, say

$$\begin{pmatrix} \text{Buick} & \text{red} \\ \text{Buick} & \text{yellow} \\ \text{Mercedes} & \text{red} \\ \text{Buick} & \text{yellow} \\ \text{Toyota} & \text{red} \\ \text{Mercedes} & \text{yellow} \end{pmatrix}$$

Now there are *five* dummy variables

$$\begin{pmatrix} 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix}$$

The first three columns are the same as in (12.48), the fourth column the indicator of the category *red*, and the fifth column the indicator of the category *yellow*. But (and this is a very important "but") this design matrix does not have full rank, because the first three columns add to the predictor vector that is all ones, and so do the last two columns. The rank is only 4, not 5. In order to have uniquely determined regression coefficients, we must have an $n \times 4$ design matrix. The simple way to achieve this is to drop one dummy variable from each set, it doesn't matter which, and add the constant predictor. This gives us something like

$$\begin{pmatrix} 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}$$

here we have the *constant*, *Buick*, *Toyota*, and *red* dummy variables. We've kept all but one of each set (two cars, one color). R does this automagically, we don't have to do anything special. With `x` and `y` defined as above

```
z <- c("red", "yellow", "red", "yellow", "red", "yellow")
out <- lm(y ~ factor(x) + factor(z))
summary(out)
```

produces the following table of regression coefficients

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.88571    0.04518  19.606  0.00259 **
factor(x)Mercedes  1.02857    0.04949  20.785  0.00231 **
factor(x)Toyota    2.11429    0.06999  30.210  0.00109 **
factor(z)yellow    0.17143    0.04949   3.464  0.07418 .
```

It always does the right thing, provided you remember to tell it that categorical variables are "factors." (Well perhaps we should have said it always does *a* rather than *the* right thing. It didn't keep the same dummy variables, that we suggested. But it did keep two cars and one color, which is all that matters.)

No problem arises in mixing quantitative and categorical random variables. Just do it (remembering to tell R which predictors are categorical)!

Example 12.5.1.
Suppose we have a data set like

```
http://www.stat.umn.edu/geyer/5102/ex12.5.1.dat
```

which has one categorical predictor variable `sex`, one quantitative predictor `x` and a response `y`. Suppose we want to fit parallel regression lines for each of the categories, as in Figure 12.4. We will see how to make such a plot below, but first we need to discuss how to fit the regression model. If we let z denote the dummy variable indicating one of the two category values, the regression model we want has the form

$$\mathbf{y} = \alpha + \beta\mathbf{z} + \gamma\mathbf{x} + \mathbf{e}.$$

Here γ is the slope of both regression lines in the figure, α is the y-intercept of one of the lines, and $\alpha + \beta$ is the y-intercept of the other. Now we see that

```
out <- lm(y ~ factor(sex) + x)
summary(out)
```

fits this regression model. The part of the printout concerning the regression coefficients is

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.9253     0.2352  20.938  < 2e-16 ***
factor(sex)  -2.0633     0.1945 -10.606  < 2e-16 ***
x             1.0688     0.3316   3.223  0.00173 **
```

Figure 12.4 was made with the following commands.

```
f <- sex == "female"
plot(x, y, type="n")
points(x[f], y[f], pch="f")
points(x[!f], y[!f], pch="m")
lines(x[f], predict(out)[f])
lines(x[!f], predict(out)[!f])
```

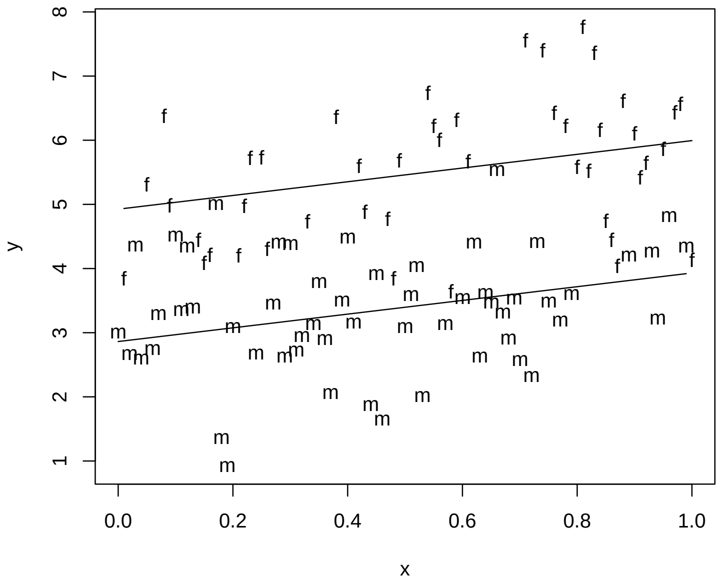


Figure 12.4: Scatter plot with parallel regression lines. The letters "f" and "m" indicate points in the two categories. The upper line is the regression line for the female category, the lower for the male category.

(If this seems a little magical, never mind. Doing fancy things with R graphics is complicated and beyond the scope of this course, not theoretical statistics.)

Great! But how about some statistics, say a test or a confidence interval? One question that is interesting is whether the true population slopes of the regression lines are the same or different. In order to find out about the "big model" that allows different slopes, we need to fit that model.

One obvious way to fit it is to divide the data, and fit a regression line to each category separately. There will be two regression coefficients (slope and intercept) for each category, making four in all. But this won't be useful for doing the test. We need fit a model to all the data that has the same predicted values (is abstractly the same regression). A little thought about dummy variables tells us that the following model will do what we want

$$\mathbf{y} = \alpha + \beta\mathbf{z} + \gamma\mathbf{x} + \delta\mathbf{x} \cdot \mathbf{z} + \mathbf{e}.$$

Here γ is the slope of one regression line and $\gamma + \delta$ is the slope of the other. As before, α is the y-intercept of one of the lines, and $\alpha + \beta$ is the y-intercept of the other. Thus something like

```
out.too <- lm(y ~ factor(sex) + x + I(factor(sex) * x)) # bogus!
```

would seem to be what is wanted. But actually, the much simpler

```
out.too <- lm(y ~ factor(sex) * x)
```

works. R assumes we want the so-called "main effects" sex and x whenever we specify the "interaction" sex * x. Also we do not need to enclose the multiplication in the I() function, because the * here doesn't really indicate multiplication. Rather it is a magic character indicating "interaction" that R recognizes in model formula and treats specially (just like + is magic). In fact, the more complicated form *doesn't* work. One must use the simple form. The part of the printout of the summary(out.too) command that is the table of regression coefficients is

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.7479     0.3013  15.757  < 2e-16 ***
factor(sex)    -1.7464     0.3885  -4.495 1.92e-05 ***
x               1.3822     0.4698   2.942  0.00408 **
factor(sex).x  -0.6255     0.6637  -0.943  0.34825
```

The four regression coefficients are α, β, γ, δ in the discussion above (in that order). A test of whether the two lines have the same slope or not, is just a test of $\delta = 0$. Hence we can read the P-value right off the printout: $P = 0.348$ (two-tailed). There is no statistically significant difference in the slopes of the two regression lines. Thus we are free to adopt the simpler model fit before with only three parameters.

If we wished to next ask the question whether a single line would fit the data (a two-parameter model), we could read the P-value for that test off the printout of the three-parameter model: $P < 2 \times 10^{-16}$ (two-tailed, though it doesn't matter for a P-value this low). Hence there is a highly statistically significant difference between the intercepts for the two categories.

12.5.2 ANOVA

Often, regression with all predictors categorical is called *analysis of variance* (ANOVA). Most textbooks, Lindgren (Sections 14.7 through 14.10) give this special case very special treatment. We won't bother, being content with the slogan that began this section.

We will just redo Example 14.7a in Lindgren to show how to do it in R

```
analyst <- c(1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 4, 4)
yield <- c(8, 5, -1, 6, 5, 3, 7, 12, 5, 3, 10, 4, -2, 1, 1, 6, 10, 7)
out <- aov(yield ~ factor(analyst))
summary(out)
```

produces the same table as in Lindgren, the only difference is that R adds the P-value for the F test.

The aov function "cuts to the chase." In ANOVA you almost always want to do F tests for models that include all the dummy variables for a given category or none. It just goes straight to the analysis of variance table for such comparisons.

Here where there is just one categorical variable, there is just one test. In so-called two-way ANOVA (Section 14.9 in Lindgren) there are two tests, one for each category. If an interaction is added (p. 533 ff. in Lindgren) that adds another test. And so forth.

12.6 Residual Analysis

An important topic that has been ignored so far is how one checks whether the model assumptions are plausible. There is, of course, never any way to prove they are correct, but it is generally accepted that one should make some effort to show that the model assumptions are not completely ridiculous.

This has not always been the case. Back in the stone age, when computers didn't come with video monitors and statistics programs just produced printout, techniques in this section were not used. People did regression with no useful checks of model assumptions, hence often when it was completely ridiculous, although they had no awareness of the ridiculosity.

Nowadays, you can install R (or similar software) on any computer and easily make diagnostic plots that will reveal some violations of model assumptions. (And miss some. There is no magic that will reveal *all* violations.) We can only scratch the surface of this area. Books on regression have much more.

For a start, we divide the (strong) assumptions into two classes.

- Assumptions about **errors**.
    - independent
    - normal
    - homoscedastic (same variance)
- The assumption about **the regression function**.

These two classes of assumptions are treated quite differently. The assumption of a particular form for the regression function is checked using F tests for model comparison. If a particular model is wrong, a larger model may be right. Presumably, some large enough model will be right. The only problem is to find it. So when we said we had been ignoring model checking, that wasn't quite right. We haven't ignored this part (although we will have a bit more to say about it later.)

To be precise, we should modify the last paragraph to say that F tests check the assumption about the regression function, *if the other assumptions are correct*. If the error assumptions don't hold, then the F statistic doesn't have an F distribution, and there's no way to interpret it.

Thus logically, the error assumptions come first, but there is a slight problem with this. We don't see the errors, so we can't check them. We do have error estimates $\hat{e}_i$, but they depend on the model we fit, which depends on the assumed regression function.

Some misguided souls attempt to avoid this dilemma by applying their checks to the responses y_i, but this is entirely misguided. The responses y_i are not identically distributed, either conditionally or unconditionally, so there is no point in looking at their distribution. Moreover the *marginal* distribution of the responses y_i is not assumed to be normal in regression theory, only the *conditional* distribution given the predictor variables. Hence there is *no* useful conclusion about regression that can be derived from checking whether the responses appear normally distributed. If they appear normal, that doesn't prove anything. If they appear non-normal, that doesn't prove anything either. I stress this because I often see naive users of regression looking at histograms of the response variable, and asking what it means. Then I trot out a slogan.

*Normality checks must be applied to **residuals**, not responses.*

Thus to return to our dilemma. We can't check the assumptions about the regression function until we have checked the error assumptions. But we can't check the error assumptions without knowing the correct regression function. The only way to proceed is to apply checks about error assumptions to residuals from a model that is large enough so that one can reasonably hope it is correct. So always apply these checks to residuals from the *largest* model under consideration, not any smaller model. That doesn't really avoid the dilemma, but it's the best one can do.

So what is the distribution of the residuals? The *errors* are i. i. d. normal (at least, that's the assumption we want check), but the residuals aren't.

Theorem 12.13. *Under the assumptions* (12.22) *and* (12.24)

$$\hat{\mathbf{e}} \sim \mathcal{N}\big(0, \sigma^2(\mathbf{I} - \mathbf{H})\big),$$

where $\mathbf{H}$ *is the "hat" matrix* (12.37).

The proof is left as an exercise.

The theorem says the residuals are jointly multivariate normal, but are neither independent nor identically distributed. Hence they do have the *normality* property assumed for the errors, but not the *independence* or *constant variance* properties.

It is a sad fact that there is no sensible test for independence of the errors. Even if we observed the errors (which we don't), there would be no test that could, even in principle tell whether they were independent. The problem is that there are too many ways that random variables can be dependent, and no test can rule them all out. If you test for some particular form of dependence, and the test accepts the null hypothesis, that does not prove independence. Some other form of dependence may be there. Thus the independence assumption is usually not checked. We have to proceed on hope here.

In checking the other properties, the lack of identical distribution is a problem. How can we check if the residuals are normal, if each one has to be checked against a different normal distribution? The obvious solution is to standardize the residuals. The random quantities

$$\frac{\hat{e}_i}{\sigma\sqrt{1 - h_{ii}}} \tag{12.50}$$

are not independent because the $\hat{e}_i$ are correlated, but they are (marginally) identically distributed, in fact, standard normal (under the "strong" regression assumptions). Hence, plugging in $\hat{\sigma}$ for σ gives quantities

$$t_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}} \tag{12.51}$$

that are identically $t(n - p)$ distributed. The quantities (12.51) are called *internally studentized residuals* and are often used to check whether the residuals are normal.

We, following R, are going to ignore them and look at a better idea, so-called *externally studentized residuals.* But the path to get there is long. It will require some patience to follow.

12.6.1 Leave One Out

A problem with residuals and internally studentized residuals is that in the i-th residual

$$\hat{e}_i = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}$$

the data y_i is used twice because $\hat{\boldsymbol{\beta}}$ depends on all the y's including y_i. A better, more honest, estimate of the error e_i, is

$$\hat{e}_{(i)} = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{(i)}$$

where $\hat{\boldsymbol{\beta}}_{(i)}$ is the regression estimate obtained by dropping the i-th case from the data. This is called a *leave-one-out residual.* Note that subscripts in parentheses do not indicate order statistics (as they did in Chapter 7 of these notes). In this section the indicate various quantities associated with a leave-one-out regression.

It would seem that leave-one-out residuals would be a big pain. It would require doing n regressions rather than just one to calculate these residuals. It is a very interesting fact about regression that this not so. The leave-one-out residuals can be calculated from the original regression using all the data. We will now see how to do this. Our analysis will also derive the distribution of the leave-one-out residuals.

Lemma 12.14. *If* $\mathbf{A}$ *is a symmetric matrix,* $\mathbf{a}$ *is a vector, and* $\mathbf{a}'\mathbf{A}\mathbf{a} \neq 1$*, then*

$$(\mathbf{A} - \mathbf{a}\mathbf{a}')^{-1} = \mathbf{A}^{-1} + \frac{\mathbf{A}^{-1}\mathbf{a}\mathbf{a}'\mathbf{A}^{-1}}{1 - \mathbf{a}'\mathbf{A}^{-1}\mathbf{a}}$$

Proof. We merely need to multiply $\mathbf{A} - \mathbf{a}\mathbf{a}'$ by the formula the lemma asserts is its inverse and check that we get the identity.

$$\begin{aligned}
(\mathbf{A} - \mathbf{a}\mathbf{a}')\left(\mathbf{A}^{-1} + \frac{\mathbf{A}^{-1}\mathbf{a}\mathbf{a}'\mathbf{A}^{-1}}{1 - \mathbf{a}'\mathbf{A}^{-1}\mathbf{a}}\right) &= I - \mathbf{a}\mathbf{a}'\mathbf{A}^{-1} + \frac{\mathbf{a}\mathbf{a}'\mathbf{A}^{-1} - \mathbf{a}\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}\mathbf{a}'\mathbf{A}^{-1}}{1 - \mathbf{a}'\mathbf{A}^{-1}\mathbf{a}} \\
&= I - \mathbf{a}\mathbf{a}'\mathbf{A}^{-1} + \frac{(1 - \mathbf{a}'\mathbf{A}^{-1}\mathbf{a})\mathbf{a}\mathbf{a}'\mathbf{A}^{-1}}{1 - \mathbf{a}'\mathbf{A}^{-1}\mathbf{a}} \\
&= I - \mathbf{a}\mathbf{a}'\mathbf{A}^{-1} + \mathbf{a}\mathbf{a}'\mathbf{A}^{-1} \\
&= I
\end{aligned}$$

The only tricky bit is the second equality, which results from the realization that the factor $\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}$ in $\mathbf{a}\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}\mathbf{a}'\mathbf{A}^{-1}$ is a *scalar* an hence can be factored out. □

Lemma 12.15. *For any matrix* $\mathbf{X}$*, let* $\mathbf{x}_i$ *denote the (column) vector corresponding to the i-th row of* $\mathbf{X}$ *and* $\mathbf{X}_{(i)}$ *the matrix obtained by deleting the i-th row from* $\mathbf{X}$*, then*

$$\mathbf{X}'\mathbf{X} = \mathbf{X}'_{(i)}\mathbf{X}_{(i)} + \mathbf{x}_i\mathbf{x}'_i$$

Proof. Obvious. Just write out in detail what the formulas mean. □

Lemma 12.16. *With the notation in the preceding lemma, if*

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

has elements h_{ij}*, then*

$$h_{ii} = \mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$$

Proof. Obvious. Just write out in detail what the formulas mean. □

Corollary 12.17.

$$\left(\mathbf{X}'_{(i)}\mathbf{X}_{(i)}\right)^{-1} = (\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}}{1 - h_{ii}} \tag{12.52}$$

Theorem 12.18. *Let* $\mathbf{y}$ *have elements* y_i *and let* $\mathbf{y}_{(i)}$ *denote the vector obtained by deleting the i-th element from* $\mathbf{y}$*. Define*

$$\begin{aligned}
\hat{y}_i &= \mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
\hat{y}_{(i)} &= \mathbf{x}'_i\left(\mathbf{X}'_{(i)}\mathbf{X}_{(i)}\right)^{-1}\mathbf{X}'_{(i)}\mathbf{y}_{(i)}
\end{aligned}$$

and

$$\begin{aligned}
\hat{e}_i &= y_i - \hat{y}_i \\
\tilde{e}_i &= y_i - \hat{y}_{(i)}
\end{aligned}$$

Then

$$\hat{e}_{(i)} = \frac{\hat{e}_i}{1 - h_{ii}} \tag{12.53}$$

Proof. Using the corollary, and

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'_{(i)}\mathbf{y}_{(i)} + y_i \mathbf{x}_i \tag{12.54}$$

which is proved like Lemma 2.39,

$$\begin{aligned}
\hat{y}_{(i)} &= \mathbf{x}_i' \left((\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}}{1 - h_{ii}} \right) (\mathbf{X}'\mathbf{y} - y_i \mathbf{x}_i) \\
&= \hat{y}_i - h_{ii} y_i + \frac{h_{ii}\hat{y}_i - h_{ii}^2 y_i}{1 - h_{ii}} \\
&= \frac{\hat{y}_i - h_{ii} y_i}{1 - h_{ii}}
\end{aligned}$$

And

$$\hat{e}_{(i)} = \frac{y_i - \hat{y}_i}{1 - h_{ii}} = \frac{\hat{e}_i}{1 - h_{ii}}$$

which is the assertion (12.53) of the theorem. □

Thus we finally arrive at the definition of the leave-one-out residuals in terms of the ordinary residuals (12.53). At first sight, this doesn't seem to do much because the leave-one-out residuals are just a constant times the ordinary residuals (a different constant for each residual, but "constant" here means non-random rather than "same") hence when standardized are exactly the same (12.50). However, a bit deeper thought says that the "plug-in" step that follows is different. Instead of (12.51) we should plug in the standard error for the *leave-one-out* regression obtaining

$$t_{(i)} = \frac{\hat{e}_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}} \tag{12.55}$$

where $\hat{\sigma}_{(i)}$ is the estimate of σ obtained by dropping the i-th case from the data. These residuals (12.55) are called *externally studentized residuals.*

These residuals are identically $t(n - 1 - p)$ distributed, because $\hat{\sigma}_{(i)}$ is based on $n - 1$ data points and p predictors. They are exactly the t statistics for the test of whether y_i is data from the model by whether the prediction interval for y_i based on the other $n - 1$ data points covers y_i. (This is not obvious, since we didn't derive them that way.)

We are not quite finished with our theoretical derivation. We still need a formula for $\hat{\sigma}_{(i)}$ that doesn't require a new regression procedure.

Lemma 12.19.

$$\hat{\sigma}_{(i)}^2 = \hat{\sigma}^2 \frac{n - p - t_i^2}{n - p - 1}$$

where the t_i are the internally studentized residuals (12.51).

Proof. By definition

$$\begin{aligned}\hat\sigma^2 &= \frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{n-p} \\ &= \frac{\mathbf{y}'(\mathbf{I}-\mathbf{H})^2\mathbf{y}}{n-p} \\ &= \frac{\mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y}}{n-p}\end{aligned}$$

because $\hat{\mathbf{e}} = (\mathbf{I}-\mathbf{H})\mathbf{y}$ and $\mathbf{I}-\mathbf{H}$ is symmetric and idempotent.

Hence the analogous formula

$$\hat\sigma^2_{(i)} = \frac{\mathbf{y}'_{(i)}\bigl(\mathbf{I}-\mathbf{H}_{(i)}\bigr)\mathbf{y}_{(i)}}{n-p-1}$$

holds for the leave-one-out regression. Now

$$\mathbf{y}'_{(i)}\bigl(\mathbf{I}-\mathbf{H}_{(i)}\bigr)\mathbf{y}_{(i)} = \mathbf{y}'_{(i)}\mathbf{y}_{(i)} - \mathbf{y}'_{(i)}\mathbf{H}_{(i)}\mathbf{y}_{(i)}$$

and the first term is

$$\mathbf{y}'_{(i)}\mathbf{y}_{(i)} = \mathbf{y}'\mathbf{y} - y_i^2. \tag{12.56}$$

The second term is more complicated but can be calculated using (12.54) and (12.52).

$$\begin{aligned}\mathbf{y}'_{(i)}\mathbf{H}_{(i)}\mathbf{y}_{(i)} &= \mathbf{y}'_{(i)}\mathbf{X}_{(i)}\left(\mathbf{X}'_{(i)}\mathbf{X}_{(i)}\right)^{-1}\mathbf{X}'_{(i)}\mathbf{y}_{(i)} \\ &= (\mathbf{y}'\mathbf{X} - y_i\mathbf{x}'_i)\left((\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}}{1-h_{ii}}\right)(\mathbf{X}'\mathbf{y} - y_i\mathbf{x}_i) \\ &= \mathbf{y}'\mathbf{H}\mathbf{y} - 2y_i\hat y_i + h_{ii}y_i^2 + \frac{\hat y_i^2 - 2h_{ii}y_i\hat y_i + h_{ii}^2y_i^2}{1-h_{ii}} \\ &= \mathbf{y}'\mathbf{H}\mathbf{y} + \frac{\hat y_i^2 - 2y_i\hat y_i + h_{ii}y_i^2}{1-h_{ii}}\end{aligned}$$

Subtracting this from (12.56) gives

$$\begin{aligned}(n-p-1)\hat\sigma^2_{(i)} &= \mathbf{y}'\mathbf{y} - y_i^2 - \mathbf{y}'\mathbf{H}\mathbf{y} - \frac{\hat y_i^2 - 2y_i\hat y_i + h_{ii}y_i^2}{1-h_{ii}} \\ &= (n-p)\hat\sigma^2 - y_i^2 - \frac{\hat y_i^2 - 2y_i\hat y_i + h_{ii}y_i^2}{1-h_{ii}} \\ &= (n-p)\hat\sigma^2 - \frac{\hat y_i^2 - 2y_i\hat y_i + y_i^2}{1-h_{ii}} \\ &= (n-p)\hat\sigma^2 - \frac{\hat e_i^2}{1-h_{ii}} \\ &= (n-p)\hat\sigma^2 - \hat\sigma^2 t_i^2\end{aligned}$$

and solving for $\hat\sigma^2_{(i)}$ gives the assertion of the lemma. □

Linear algebra is really remarkable. Or perhaps it is least squares that is so remarkable. There is something magic, that such a complicated calculation yields such a simple result. Whatever the reason, we now have two simple formulas for identically distributed, standardized residual estimates. Different computer statistical software packages make different choices about which residuals to use. We agree with the R team that the externally studentized residuals are the ones to use.

12.6.2 Quantile-Quantile Plots

How does one check that externally studentized residuals are $t(n-p-1)$ distributed? A widely used method uses a quantile-quantile (Q-Q) plot. Q-Q plots can be applied to any random quantities, not just residuals. So temporarily forget residuals.

Let X_1, X_2, $\ldots$, X_n be data assumed to be i. i. d., and suppose we want to check whether their distribution has a particular distribution with c. d. f. F. A Q-Q plot is a plot of the order statistics $X_{(k)}$ of the data[5] against quantities that are reasonable theoretical positions of these order statistics. We can't be more precise than that, because there are many different proposals about what positions should be used. Two are

$$F^{-1}\left(\frac{k-\frac{1}{2}}{n}\right) \tag{12.57}$$

and

$$F^{-1}\left(\frac{k}{n+1}\right) \tag{12.58}$$

Some more proposals will be discussed below.

We know (Theorem 9 of Chapter 3 in Lindgren) that if F is continuous, then the variables

$$U_i = F(X_i)$$

are i. i. d. $\mathcal{U}(0,1)$, hence of course, the order statistics $U_{(k)}$ are order statistics of a sample of size n from the $\mathcal{U}(0,1)$ distribution, and

$$X_{(k)} = F^{-1}(U_{(k)}).$$

The reason why this is important is that we know the distribution of the $U_{(k)}$

$$X_{(k)} \sim \text{Beta}(k, n-k+1) \tag{12.59}$$

(p. 217 in Lindgren). Hence

$$E\{U_{(k)}\} = \frac{k}{n+1}.$$

[5] Here the $X_{(k)}$ indicate order statistics as in Chapter 7 of these notes, not the leave-one-out quantities of the preceding section. In fact, since the X_i here have nothing to do with regression, the parenthesized subscripts couldn't possibly indicate leave one out.

That is the origin of (12.58). Of course, this doesn't prove that (12.58) is the Right Thing. Far from it. We know that in general

$$g\bigl(E\{X\}\bigr) = E\bigl\{g(X)\bigr\} \tag{12.60}$$

is generally *false*. The only condition we know that makes (12.60) hold is that g be a *linear* function. Now inverse c. d. f.'s are never linear except for the special case of the uniform distribution. Hence

$$E\left\{X_{(k)}\right\} = E\left\{F^{-1}\bigl(U_{(k)}\bigr)\right\} \neq F^{-1}\left(E\bigl\{U_{(k)}\bigr\}\right) = F^{-1}\left(\frac{k}{n+1}\right)$$

Thus, although (12.58) has some theoretical woof associated with it, it does not do exactly the right thing. We can only consider (12.58) *a thing to do* (as opposed to *the* thing to do). Hence the other proposals.

The proposal (12.57) has less theory behind it. It is based on the idea that n in the denominator rather than $n+1$ is more natural (no theoretical reason for this). Unit spacing between the points also seems natural. Then the requirement that they be placed symmetrically in the interval $(0,1)$ determines the form $(k - \frac{1}{2})/n$.

Another proposal often seen is so-called *normal scores*. These are $E\bigl\{X_{(k)}\bigr\}$ when the X_i have a standard normal distribution. The are, however, hard to compute. Some statistics packages have them, but not all. R doesn't. Of course, these are only useful when the distribution of interest is the normal distribution. The analogous quantities could be defined for any distribution, but software and tables exist only for the normal.

A proposal that does work for all distributions would be to put the *medians* of the beta distributions (12.59) through the inverse c. d. f., that is, if ζ_k is the median of the $\text{Beta}(k, n-k+1)$ distribution use $F^{-1}(\zeta_k)$ as the plotting points. This proposal has the virtue of coming from a correct theoretical argument. The median of $X_{(k)}$ is indeed $F^{-1}(\zeta_k)$, because medians (as opposed to means) do indeed go through quantile transforms.

In practice all of these proposals produce almost the same picture. So nobody worries about the differences, and does what seems simplest. The R function `qqnorm` does a Q-Q plot against the normal distribution and uses the proposal (12.57) for the plotting points. Here's how to do a Q-Q plot against a normal distribution of an arbitrary data vector `x` in R.

```
qqnorm(x)
qqline(x)
```

The first command does the Q-Q plot. The second puts on a line about which the points should cluster. Since we don't know the parameters μ and σ^2 of the population from which `x` was drawn, we don't know in advance which line the points should cluster about (`qqnorm` plots against the *standard* normal. If the data are standard normal, then the points cluster about the line with intercept zero and slope one. If the data are normal (but not standard normal), then the points cluster about the line with intercept μ and slope σ. So if the points cluster

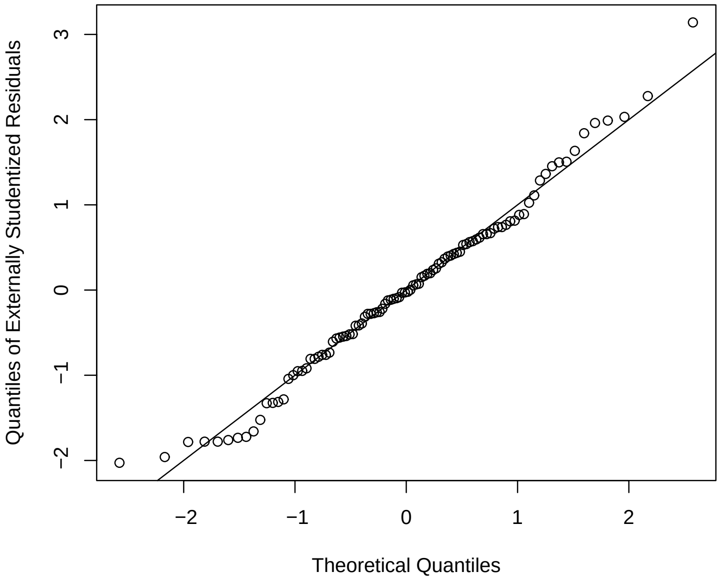


Figure 12.5: A Q-Q Plot.

about *any* line, we conclude they are approximately normally distributed. R picks a reasonable line, one that most of the points cluster about.

Here's how to do a Q-Q plot of the externally studentized residuals in R, assuming we have already fit a linear model and put the result of the `lm` command in a variable `out`

```
qqnorm(rstudent(out))
abline(0, 1)
```

The second command draws a line with intercept zero and slope one. If the residuals are approximately standard normal,[6] then the points in the plot should lie near this line.

Example 12.6.1.
This looks at the residuals from the fifth degree polynomial fit to the data of Example 12.3.2. Figure 12.5 shows the Q-Q plot of the externally studentized residuals.

It's not clear what one is supposed to make of a Q-Q plot. The points never lie *exactly* on the line because of chance variation (the sample is not the population). So how far off the line do they have to be before you should question

[6] Why standard normal when the externally studentized residuals have marginal distribution $t(n-p-1)$? Because they are not independent samples from this distribution. In fact the random parts of the denominators $\hat{\sigma}_{(i)}$ are highly correlated. Thus they look much more like a random sample from a normal than from a t distribution.

the regression assumptions? One sensible recommendation is to "calibrate" your eye by looking at several Q-Q plots for data simulated from the correct distribution (here standard normal).

```
qqnorm(rnorm(n))
abline(0, 1)
```

where `n` is the number of cases in the data. Repeat until you get an idea how much variation you should expect.

But isn't there a hypothesis test we should apply? There are hypothesis tests one can do. The trouble with them is that, when n is large, they tend to find "statistically significant" departures from normality that are fairly minor in their effects. Linear regression is somewhat robust against minor departures from normality of the error distribution. So small departures, "statistically significant" though they may be, can be ignored. What we are looking for here is really obvious and serious nonnormality.

12.7 Model Selection

This section discusses the problem of choosing among many models. Although our examples will be regression problems and some of the discussion and methods will be specific to linear regression, the problem is general. When many models are under consideration, how does one choose the best? Thus we also discuss methods that apply outside the regression context.

There are two main issues.

- Non-nested models
- Many models.

The only methods for model comparison we have studied, the F test for comparison of linear regression models and the likelihood ratio test for comparison of general models, are valid only for comparing two *nested* models. We also want to test *non-nested* models, and for that we need new theory. When more than two models are under consideration, the issue of correction for multiple testing arises. If there are only a handful of models, Bonferroni correction (or some similar procedure) may suffice. When there are many models, a conservative correction like Bonferroni is too conservative. Let's consider how many models might be under consideration in a model selection problem. Consider a regression problem with k predictor variables.

- [**Almost the worst case**] There are 2^k possible submodels formed by choosing a subset of the k predictors to include in the model (because a set with k elements has 2^k subsets).
- [**Actually the worst case**] That doesn't consider all the new predictors that one might "make up" using functions of the old predictors. Thus there are potentially infinitely many models under consideration.

Bonferroni correction for infinitely many tests is undefined. Even for 2^k tests with k large, Bonferroni is completely pointless. It would make nothing statistically significant.

12.7.1 Overfitting

> *Least squares is **good** for model fitting, but **useless** for model selection.*

Why? A bigger model *always* has a smaller residual sum of squares, just because a minimum taken over a larger set is smaller.[7] Thus least squares, taken as a criterion for model selection says "always choose the biggest model." But this is silly. Consider what the principle of choosing the biggest model says about polynomial regression.

Example 12.7.1 (Overfitting in Polynomial Regression).
Consider regression data shown in Figure 12.6. Couldn't ask for nicer data for simple linear regression. The data appear to fit the "simple" model (one non-constant predictor, which we take to be x itself).

But now consider what happens when we try to be a bit more sophisticated. How do we know that the "simple" model is o. k.? Perhaps we should consider some more complicated models. How about trying polynomial regression? But if we consider all possible polynomial models, that's an infinite number of models (polynomials of all orders).

Although an infinite number of models are potentially under consideration, the fact that the data set is finite limits the number of models that actually need to be considered to a finite subset. You may recall from algebra that any set of n points in the plane having n different x values can be interpolated (fit exactly) by a polynomial of degree $n - 1$. A polynomial that fits exactly has residual sum of squares zero (fits perfectly). Can't do better than that by the least squares criterion! Thus all polynomials of degree at least $n - 1$ will give the same fitted values and zero residual sum of squares. Although they may give different predicted values for x values that do not occur in the data, they give the same predicted values at those that do. Hence the polynomials with degree at least $n - 1$ cannot be distinguished by least squares. In fact the polynomials with degree more than $n - 1$ cannot be fit at all, because they

[7] If g is any real-valued function and A and B are two subsets of the domain of g with $A \subset B$ then

$$\inf_{x \in A} g(x) \geq \inf_{y \in B} g(y)$$

simply because every x that occurs on the left hand side also occurs as a y on the right hand side because $A \subset B$. Taking g to be the least squares criterion for a regression model, and A and B the parameter spaces for two nested models gives the result that the larger model always has the smaller residual sum of squares.

Note this is exactly analogous to what happens with maximum likelihood. The larger model always has the larger log likelihood. The reasoning is exactly the same except for the inequality being reversed because of maximizing in maximum likelihood rather than minimizing in least squares.

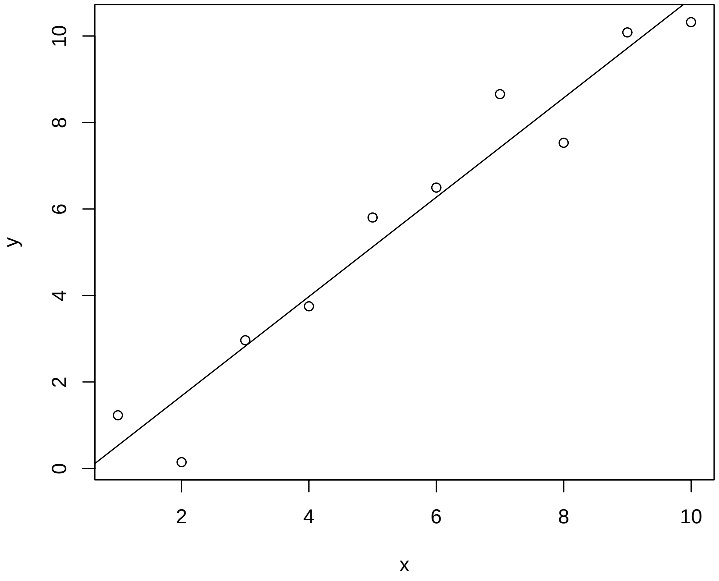


Figure 12.6: Some regression data. With fitted "simple" regression function of the form $y = \hat{\alpha} + \hat{\beta}x$.

have more parameters than there are equations to determine them. $\mathbf{X}'\mathbf{X}$ is a singular matrix and cannot inverted to give unique estimates of the regression coefficients.

This is a general phenomenon, which occurs in all settings, not just with linear regression. Models with more parameters than there are data points are underdetermined. Many different parameter vectors give the same likelihood, or the same empirical moments for the method of moments, or the same for whatever criterion is being used for parameter estimation. Thus in general, even when an infinite number of models are theoretically under consideration, only n models are practically under consideration, where n is the sample size.

Hence the "biggest model" that the least squares criterion selects is the polynomial of degree $n - 1$. What does it look like? Figure 12.7 shows both the best fitting (perfectly fitting!) polynomial of degree $n - 1 = 9$ and the least squares regression line from the other figure.

How well does the biggest model do? It fits the observed data perfectly, but it's hard to believe that it would fit *new* data from the same population as well. The extreme oscillations near the ends of the range of the data are obviously nonsensical, but even the smaller oscillations in the middle seem to be tracking random noise rather than any real features of the population regression function. Of course, we don't actually know what the true population regression function is. It could be either of the two functions graphed in the figure, or it could be some other function. But it's hard to believe, when the linear function fits so

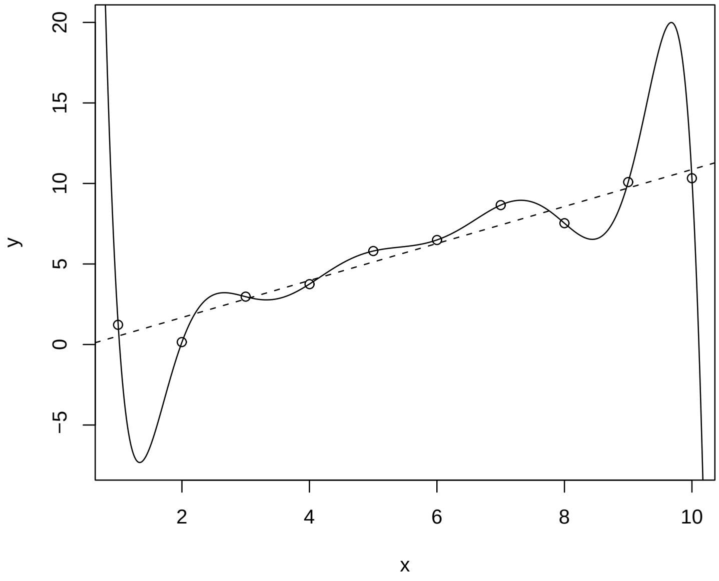


Figure 12.7: Some regression data. With fitted linear regression function (dashed line) and ninth degree polynomial regression function (solid curve).

well, that something as complicated as the ninth degree polynomial is close to the true regression function. We say it "overfits" the data, meaning it's *too* close to the data and not close enough to the true population regression function.

12.7.2 Mean Square Error

So what criterion should we use for model selection if residual sum of squares is no good? One theoretical criterion is mean square error. In the regression setting, it is unclear what quantities mean square error should apply to. Do we use the mean square error of the parameter estimates? We haven't even defined mean square error for vector quantities. That alone suggests we should avoid looking at m. s. e. of regression coefficients. There is also our slogan that regression coefficients are meaningless. Hence we should look at estimates of the regression function.

But here too, there are still issues that need to be clarified. The regression function $h(\mathbf{x}) = E(Y \mid \mathbf{x})$ is a scalar function of $\mathbf{x}$. (Hence we don't need to worry about m. s. e. of a vector quantity). But it is a function of the predictor value $\mathbf{x}$.

$$\operatorname{mse}\{\hat{h}(\mathbf{x})\} = \text{variance} + \text{bias}^2 = \operatorname{var}\{\hat{h}(\mathbf{x})\} + \Big(E\{\hat{h}(\mathbf{x})\} - h(\mathbf{x})\Big)^2 \qquad (12.61)$$

where $h(\mathbf{x})$ is the true population regression function and $\hat{h}(\mathbf{x})$ is an estimate (we are thinking of least squares regression estimates, but (12.61) applies to any estimate). ("Bias?" did I hear someone say? Aren't linear regression estimates *unbiased*? Yes, the are when the model is *correct*. Here we are considering cases when the model is too small to contain the true regression function.)

To make a criterion that we can minimize to find the best model, we need a single scalar quantity, not a function. There are several things we could do to make a scalar quantity from (12.61). We could *integrate* it over some range of values, obtaining so-called *integrated mean squared error*. A simpler alternative is to sum it over the *design points*, the $\mathbf{x}$ values occurring in the data set under discussion. We'll discuss only the latter.

As we did before, write $\boldsymbol{\mu}$ for the true population means of the responses given the predictors, defined by $\mu_i = h(\mathbf{x}_i)$. Let m index models. The m-th model will have design matrix $\mathbf{X}_m$ and hat matrix $\mathbf{H}_m$. The expected value of the regression predictions $\hat{\mathbf{y}} = \mathbf{H}_m \mathbf{y}$ under the m-th model is

$$E(\hat{\mathbf{y}}) = \mathbf{H}_m \boldsymbol{\mu}.$$

Hence the bias is

$$E(\mathbf{y}) - E(\hat{\mathbf{y}}) = (\mathbf{I} - \mathbf{H}_m)\boldsymbol{\mu}. \tag{12.62}$$

Note that if the model is correct, then $\boldsymbol{\mu}$ is in the range of $\mathbf{H}_m$ and the bias is zero. Models that are incorrect have nonzero bias. (12.62) is, of course, a vector. Its i-th element gives the bias of $\hat{y}_i$. What we decided to study was the sum of the m. s. e.'s at the design points, the "bias" part of which is just the sum of squares of the elements of (12.62), which is the same thing as the squared length of this vector

$$\text{bias}^2 = \|(\mathbf{I} - \mathbf{H}_m)\boldsymbol{\mu}\|^2 = \boldsymbol{\mu}'(\mathbf{I} - \mathbf{H}_m)^2\boldsymbol{\mu} = \boldsymbol{\mu}'(\mathbf{I} - \mathbf{H}_m)\boldsymbol{\mu}. \tag{12.63}$$

The variance is

$$\begin{aligned}
\operatorname{var}(\hat{\mathbf{y}}) &= E\left\{\left\|\hat{\mathbf{y}} - E(\hat{\mathbf{y}})\right\|^2\right\} \\
&= E\left\{\left\|\hat{\mathbf{y}} - \mathbf{H}_m\boldsymbol{\mu}\right\|^2\right\} \\
&= E\left\{\left\|\mathbf{H}_m(\mathbf{y} - \boldsymbol{\mu})\right\|^2\right\} \\
&= E\left\{\mathbf{H}_m(\mathbf{y} - \boldsymbol{\mu})'(\mathbf{y} - \boldsymbol{\mu})\mathbf{H}_m\right\} \\
&= \sigma^2\mathbf{H}_m^2 \\
&= \sigma^2\mathbf{H}_m
\end{aligned}$$

This is, of course, a matrix. What we want though is just the sum of the diagonal elements. The i-th diagonal element is the variance of $\hat{y}_i$, and our decision to focus on the sum of the m. s. e.'s at the design points says we want the sum of these. The sum of the diagonal elements of a square matrix $\mathbf{A}$ is called its *trace*, denoted $\operatorname{tr}(\mathbf{A})$. Thus the "variance" part of the m. s. e. is

$$\text{variance} = \sigma^2 \operatorname{tr}(\mathbf{H}_m). \tag{12.64}$$

And

$$\text{MSE}_m = \sum_{i=1}^{n} \text{mse}(\hat{y}_i) = \sigma^2 \operatorname{tr}(\mathbf{H}_m) + \boldsymbol{\mu}'(\mathbf{I} - \mathbf{H}_m)\boldsymbol{\mu}.$$

12.7.3 The Bias-Variance Trade-Off

Generally, one cannot reduce both bias and variance at the same time. Bigger models have less bias but *more* variance. Smaller models have less variance but *more* bias. This is called the "bias-variance trade-off."

Example 12.7.2.
We again use the data for Example 12.3.2. This example, however, will be purely theoretical. We will look at various polynomial regression models, using the x values, but ignoring the y values. Instead we will do a purely theoretical calculation using the parameter values that were actually used to simulate the y values in the data

$$\mu_i = \sin(x_i) + \sin(2x_i) \tag{12.65a}$$

$$\sigma = 0.2 \tag{12.65b}$$

Since the true population regression curve is a trigonometric rather than a polynomial function, *no* polynomial is unbiased. This is typical of real applications. No model under consideration is exactly correct.

Table 12.1 shows the results of the theoretical calculations for the data in this example. We see that the fifth degree polynomial chosen in Example 12.3.2 is not the best. The ninth degree polynomial is the best. Figure 12.8 shows both the true regression function used to simulate the response values (12.65a) and the sample ninth-degree polynomial regression function. We see that the sample regression function does not estimate the true regression function perfectly (because the sample is not the population), but we now know from the theoretical analysis in this example that no polynomial will fit better. A lower degree polynomial will have less variance. A higher degree will have less bias. But the bias-variance trade-off will be worse for either.

12.7.4 Model Selection Criteria

The theory discussed in the preceding section gives us a framework for discussing model selection. We want the model with the smallest m. s. e., the model which makes the optimal bias-variance trade-off.

Unfortunately, this is useless in practice. Mean square error is a theoretical quantity. It depends on the unknown true regression function and the unknown error variance. Furthermore, there is no obvious way to estimate it. Without knowing which models are good, which is exactly the question we are trying to resolve, we can't get a good estimate of the true regression function (and without that we can't estimate the error variance well either).

Table 12.1: Bias and Variance for Different Polynomial Regression Models. The second column gives the variance (12.64) and the third column gives the "bias^2" (12.63) for polynomial regression models of different degrees. The last column gives their sum (mean squared error). The model with the smallest m. s. e. (degree 7) is the best. The calculations are for the situation with true regression function (12.65a) and true error standard deviation (12.65b).

degree	variance	bias^2	m. s. e.
0	0.04	99.0000	99.0400
1	0.08	33.3847	33.4647
2	0.12	33.3847	33.5047
3	0.16	29.0924	29.2524
4	0.20	29.0924	29.2924
5	0.24	4.8552	5.0952
6	0.28	4.8552	5.1352
7	0.32	0.1633	0.4833
8	0.36	0.1633	0.5233
9	0.40	0.0019	0.4019
10	0.44	0.0019	0.4419
11	0.48	9.8×10^{-6}	0.4800
12	0.52	9.8×10^{-6}	0.5200
13	0.56	2.6×10^{-8}	0.5600
14	0.60	2.6×10^{-8}	0.6000
15	0.64	3.9×10^{-11}	0.6400
16	0.68	3.9×10^{-11}	0.6800
17	0.72	4.9×10^{-14}	0.7200
18	0.76	4.9×10^{-14}	0.7600
19	0.80	1.2×10^{-14}	0.8000
20	0.84	1.4×10^{-14}	0.8400

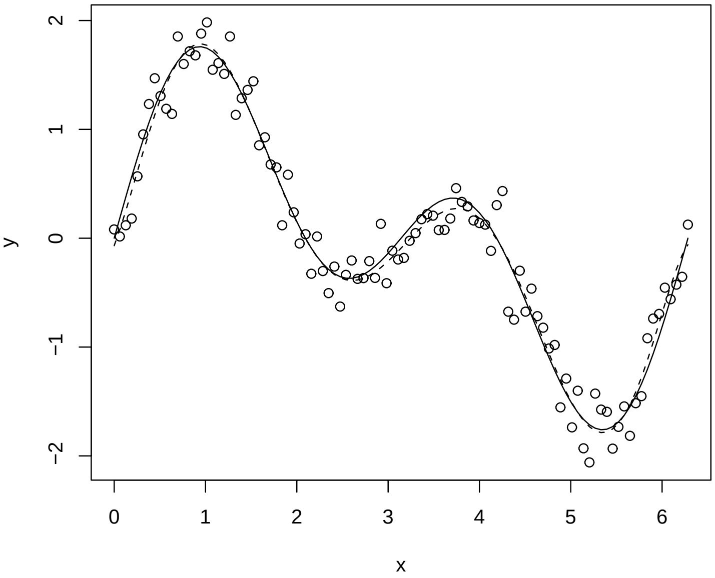


Figure 12.8: Some regression data with true regression function (solid line) and sample regression function (dashed line).

There are several quantities that have been proposed in the literature as estimates of m. s. e. No one is obviously right. We shall not go into the theory justifying them (it is merely heuristic anyway). One fairly natural quantity is

$$\sum_{i=1}^{n} \hat{e}_{(i)}^2 \tag{12.66}$$

Unlike SSResid (the sum of the $\hat{e}_i$), this does not always favor the biggest model. Models that overfit tend to do a bad job of even their leave-one-out predictions. (12.66) is called the *predicted residual sum of squares* (PRESS) or the *cross-validated sum of squares* (CVSS), cross-validation being another term used to describe the leave-one-out idea.

The idea of CVSS or other criteria to be described presently is to pick the model with the smallest value of the criterion. This will not necessarily be the model with the smallest m. s. e., but it is a reasonable estimate of it.

Another criterion somewhat easier to calculate is Mallows' C_p defined by

$$\begin{aligned} C_p &= \frac{\text{SSResid}_p}{\hat{\sigma}^2} + 2p - n \\ &= \frac{\text{SSResid}_p - \text{SSResid}_k}{\hat{\sigma}^2} + p - (k - p) \\ &= (k - p)(F_{k-p,n-k} - 1) + p \end{aligned} \tag{12.67}$$

where SSResid_p is the sum of squares of the residuals for some model with p predictors (including the constant predictor, if present), $\hat{\sigma}^2 = \text{SSResid}_k/(n-k)$ is the estimated error variance for the largest model under consideration, which has k predictors, and $F_{p,k}$ is the F statistic for the F test for comparison of these two models. The F statistic is about one in size if the small model is correct, in which case $C_p \approx p$. This gives us a criterion for finding reasonably fitting models. When many models are under consideration, many of them may have $C_p \approx p$ or smaller. All such models must be considered reasonably good fits. Any might be the correct model or as close to correct as any model under consideration. The quantity estimated by C_p is the mean square error of the model with p predictors, divided by σ^2.

An idea somewhat related to Mallows' C_p, but applicable outside the regression context is the *Akaike information criterion* (AIC).

$$-2 \cdot (\text{log likelihood}) + 2p \tag{12.68}$$

where as in Mallows' C_p, the number of parameters in the model is p. Although (12.67) and (12.68) both have the term $2p$, they are otherwise different. The log likelihood for a linear regression model, with the MLE's plugged in for the parameters is

$$-\frac{n}{2}[1 + \log(\hat{\sigma}^2)]$$

[equation (12) on p. 511 in Lindgren]. Thus for a regression model

$$\text{AIC} = n + n\log(\hat{\sigma}_p^2) + 2p$$

where we have put a subscript p on the estimated error variance to indicate clearly that it is the estimate from the model with p predictors, not the error estimate from the largest model (k predictors used in C_p).

Finally, we add one last criterion, almost the same as AIC, called the *Bayes information criterion* (BIC)

$$-2 \cdot (\text{log likelihood}) + p\log(n) \tag{12.69}$$

In the regression context, this becomes

$$\text{BIC} = n + n\log(\hat{\sigma}_p^2) + p\log(n)$$

Neither AIC nor BIC have a rigorous theoretical justification applicable to a wide variety of models. Both were derived for special classes of models that were easy to analyze and both involve some approximations. Neither can be claimed to be the right thing (nor can anything else). As the "Bayes" in BIC indicates, the BIC criterion is intended to approximate using Bayes tests instead of frequentist tests (although its approximation to true Bayes tests is fairly crude). Note that BIC penalizes models with more parameters more strongly than AIC ($p\log n$ versus $2p$). So BIC always selects a smaller model than AIC.

This gives us four criteria for model selection. There are arguments in favor of each. None of the arguments are completely convincing. All are widely used.

Table 12.2: Model Selection Criteria. Four model selection criteria, CVSS, Mallows' C_p, AIC, and BIC applied to the data of Example 12.3.2.

p	CVSS	C_p	AIC	BIC
1	102.44	2327.01	102.40	105.01
2	40.44	832.80	10.45	15.66
3	41.79	834.29	13.42	21.24
4	36.06	706.84	1.44	11.86
5	37.01	707.20	4.28	17.31
6	10.72	125.33	−124.48	−108.85
7	11.36	127.03	−121.56	−103.32
8	4.52	8.22	−202.20	−181.36
9	4.73	9.75	−199.62	−176.17
10	4.91	11.50	−196.79	−170.74
11	5.22	13.50	−193.67	−165.02
12	5.71	15.49	−190.55	−159.29
13	5.10	14.58	−190.64	−156.77
14	5.74	16.06	−188.07	−151.59
15	5.08	14.62	−188.89	−149.82
16	4.89	16.00	−186.44	−144.76

Example 12.7.3.
We use the data for Example 12.3.2 yet again. Now we fit polynomials of various degrees to the data, and look at our four criteria. The results are shown in Table 12.2. In this example, all four criteria select the same model, the model with $p = 8$ predictors, which is the polynomial of degree 7. This is not the model with the smallest m. s. e. discovered by the theoretical analysis. The criteria do something sensible, but as everywhere else in statistics, there are errors (the sample is not the population).

12.7.5 All Subsets Regression

We know return to the situation in which there are k predictors including the constant predictor and 2^{k-1} models under consideration (the constant predictor is usually included in all models). If k is large and one has no non-statistical reason (e. g., a practical or scientific reason) that cuts down the number of models to be considered, then one must fit them all. Fortunately, there are fast algorithms that allow a huge number of models to be fit or at least quickly checked to see that they are much worse than other models of the same size.

There is a contributed package to R that contains a function `leaps` that does this.

Example 12.7.4 (2^k subsets).
The data set in the URL

`http://www.stat.umn.edu/geyer/5102/ex12.7.4.dat`

consists of multivariate normal data with one response variable y and 20 predictor variables x_1, …, x_{20}. The predictors are correlated. The distribution from which they were simulated has all correlations equal to one-half. The actual (sample) correlations, of course, are all different because of chance variation.

The true population regression function (the one used to simulate the y values) was

$$y = x_1 + x_2 + x_3 + x_4 + x_5 + e \tag{12.70}$$

with error variance $\sigma^2 = 1.5^2$. If we fit the model

```
foo <- as.matrix(X)
x <- foo[ , -1]
out <- lm(y ~ x)
summary(out)
```

the table of information about the regression coefficients is

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.07879    0.17449  -0.452  0.65286
xx1          1.11887    0.22303   5.017 3.17e-06 ***
xx2          0.79401    0.26519   2.994  0.00367 **
xx3          0.45118    0.25848   1.746  0.08478 .
xx4          0.59879    0.23037   2.599  0.01115 *
xx5          1.09573    0.24277   4.513 2.20e-05 ***
xx6          0.26067    0.24220   1.076  0.28509
xx7         -0.15959    0.21841  -0.731  0.46712
xx8         -0.50182    0.23352  -2.149  0.03470 *
xx9          0.14047    0.22888   0.614  0.54116
xx10         0.37689    0.22831   1.651  0.10275
xx11         0.39805    0.21722   1.832  0.07065 .
xx12         0.38825    0.22396   1.734  0.08689 .
xx13        -0.07910    0.23553  -0.336  0.73788
xx14         0.26716    0.20737   1.288  0.20138
xx15        -0.12016    0.23073  -0.521  0.60398
xx16         0.08592    0.22372   0.384  0.70195
xx17         0.31296    0.22719   1.378  0.17224
xx18        -0.24605    0.23355  -1.054  0.29531
xx19         0.10221    0.21503   0.475  0.63586
xx20        -0.45956    0.23698  -1.939  0.05604 .
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1
```

We've left in the significance codes, bogus though they may be (more on this later), so you can easily spot the regression coefficients that the least squares fit indicates may be important. If we go only with the strongest evidence (two or three stars) we get as "significant" three of the five truly important regression coefficients [recall from (12.70) that the true nonzero regression coefficients are β_1 through β_5]. The other two are missed.

If we use a less stringent standard, say one or more stars, we do pick up another truly nonzero regression coefficient, but we also pick up a false one. Thus we now have both false negatives (we still missed β_3) and false positives (we've picked up β_8).

With the least stringent standard, all the coefficients marked by any of the "significance codes" we now have no false negatives (all five of the truly nonzero regression coefficients are now declared "significant") but we have four false positives.

No matter how you slice it, least squares regression doesn't pick the right model. Of course, this is no surprise. It's just "the sample is not the population." But it does show that the results of such model selection procedures must be treated with skepticism.

Actually, we haven't even started a sensible model selection procedure. Recall the slogan that if you want to know how good a model fits, you have to fit that model. So far we haven't fit any of the models we've discussed. We're fools to think we can pick out the good submodels just by looking at printout for the big model.

There is a function `leaps` in the `leaps` contributed package[8] that fits a huge number of models. By default, it finds the 10 best models of each size (number of regression coefficients) for which there are 10 or more models and finds all the models of other sizes.

It uses the inequality that a bigger model always has a smaller sum of squares to eliminate many models. Suppose we have already found 10 models of size p with SSResid less than 31.2. Suppose there was a model of size $p+1$ that we fit and found its SSResid was 38.6. Finally suppose $\hat{\sigma}^2$ for the big model is 2.05. Now the C_p for the 10 best models of size p already found is

$$C_p = \frac{\text{SSResid}_p}{\hat{\sigma}^2} + 2p - n < \frac{31.2}{2.05} + 2p - n$$

and the C_p for any submodel of size p of the model with SSResid = 38.6 (i. e., models obtained by dropping one predictor from that model) has

$$C_p \geq \frac{38.6}{2.05} + 2p - n$$

This means that no such model can be better than the 10 already found, so they can be rejected even though we haven't bothered to fit them. Considerations of this sort make it possible for `leaps` to pick the 10 best of each size without fitting all or even a sizable fraction of the models of each size. Thus it manages to do in minutes what it couldn't do in a week if it actually had to fit all 2^k models. The reason why `leaps` uses C_p as its criterion rather than one of the

[8] This has to be installed separately. It doesn't come with the "base" package. Like everything else about R, it can be found at `http://cran.r-project.org`.

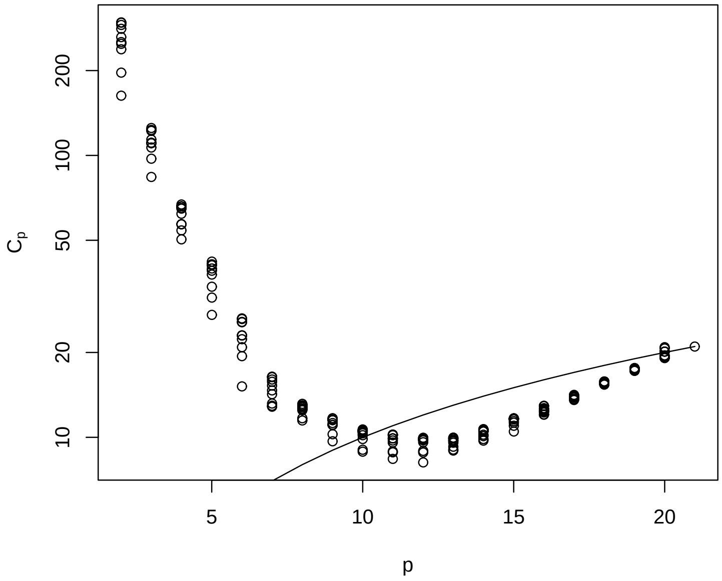


Figure 12.9: C_p plot. Plot of C_p versus p. The dots are the C_p for the 10 best models for each p. The curve is the line $C_p = p$ (curved because of the log scale for C_p).

others is that it is a simple function of SSResid and hence to these inequalities that permit its efficient operation.

We run the leaps function as follows, with the design matrix `x` defined as above,[9]

```
library(leaps)
outs <- leaps(x, y, strictly.compatible=FALSE)
plot(outs$size, outs$Cp, log="y", xlab="p", ylab=expression(C[p]))
lines(outs$size, outs$size)
```

Figure 12.9 shows the plot made by the two plot commands. Every model with $C_p < p$, corresponding to the dots below the line is "good." There are a *huge* number of perfectly acceptable models, because for the larger p there are many more than 10 good models, which are not shown.

The best model according to the C_p criterion is one with $p = 12$, so 11 non-constant predictors, which happen to be x_1 through x_5 (the truly significant predictors) plus x_8, x_{10}, x_{11}, x_{12}, x_{14}, and x_{20}. We can get its regression output as follows.

[9] For some reason, `leaps` doesn't take formula expressions like `lm` does. The reason is probably historical. The equivalent S-plus function doesn't either, because it was written before S had model formulas and hasn't changed. The `strictly.compatible=FALSE` tells R not to be bug-for-bug compatible with S-plus.

```
foo <- x[ , outs$which[outs$Cp == min(outs$Cp)]]
out.best <- lm(y ~ foo)
summary(out.best)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.04925    0.15756  -0.313 0.755316
foox1        1.07411    0.19981   5.376 6.20e-07 ***
foox2        0.88230    0.24483   3.604 0.000519 ***
foox2        0.43560    0.22969   1.896 0.061177 .
foox4        0.72393    0.19912   3.636 0.000466 ***
foox5        1.15609    0.22367   5.169 1.46e-06 ***
foox8       -0.35752    0.21251  -1.682 0.096046 .
foox10       0.43501    0.21885   1.988 0.049957 *
foox11       0.34579    0.20295   1.704 0.091940 .
foox12       0.38479    0.19811   1.942 0.055301 .
foox14       0.28910    0.18838   1.535 0.128455
foox20      -0.49878    0.21736  -2.295 0.024124 *
```

Note that the “stargazing” doesn’t correspond with the notion of the best model by the C_p criterion. One of these coefficients doesn’t even have a dot (so for it $P > 0.10$), and four others only have dots ($0.05 < P < 0.10$). Considering them separately, this would lead us to drop them. But that would be the Wrong Thing (multiple testing without correction). The `leaps` function does as close to the Right Thing as can be done. The only defensible improvement would be to change the criterion, to BIC perhaps, which would choose a smaller “best” model because it penalizes larger models more. However BIC wouldn’t have the nice inequalities that make `leaps` so efficient, which accounts for the use of C_p.

I hope you can see from this analysis that model selection when there are a huge number of models under consideration and no extra-statistical information (scientific, practical, etc.) that can be used to cut down the number is a mug’s game. The best you can do is not very good. The only honest conclusion is that a huge number of models are about equally good, as good as one would expect the correct model to be ($C_p \approx p$).

Thus it is silly to get excited about exactly which model is chosen as the “best” by some model selection procedure (any procedure)! When many models are equally good, the specific features of any one of them can’t be very important.

All of this is related to our slogan about “regression is for prediction, not explanation.” All of the models with $C_p < p$ predict about equally well. So if regression is used for *prediction*, the model selection problem is not serious. Just pick any one of the many good models and use it. For *prediction* it doesn’t matter which good prediction is used. But if regression is used for *explanation*, the model selection problem is insoluble. If you can’t decide which model is “best” and are honest enough to admit that lots of other models are equally good, then how can you claim to have found the predictors which “explain” the response? Of course, if you really understand "correlation is not causation, and regression isn't either," then you know that such "explanations" are bogus anyway, even in the "simple" case (one non-constant predictor) where the model selection problem does not arise.

Problems

12-1. Prove the assertion $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{y}}$ in the proof of Theorem 12.8.

12-2. For the data in Example 12.2.1 give 90% confidence intervals for α and β_1 (the intercept and coefficient for x_1, the coefficient for x_2 was done in Example 12.3.1

12-3. Fill in the details of the proof of Lemma 12.10.

12-4. For the data in Example 12.3.2 in the URL

`http://www.stat.umn.edu/geyer/5102/ex12.3.2.dat`

try fitting some Fourier series models. A Fourier series is sum of sines and cosines of multiples of one fundamental frequency

$$f(x) = a + \sum_{k=1}^{m} b_k \sin(2\pi kx/L) + \sum_{k=1}^{m} c_k \cos(2\pi kx/L)$$

where L is a known constant (the wavelength of the lowest frequency sine and cosine terms) and a, the b_k, and the c_k are adjustable constants. These adjustable constants will be the regression coefficients you fit using linear regression. A Fourier series is always periodic with period L. Since the variable x in this data set does just happen to take values evenly spaced between zero and 2π, and inspection of Figure 12.1 suggests the true regression may be periodic with this period, I recommend using $L = 2\pi$, which gives a regression function of the form

$$E(Y|X) = \alpha + \sum_{k=1}^{m} \beta_k \sin(kX) + \sum_{k=1}^{m} \gamma_k \cos(kX)$$

and we have changed the coefficients to Greek letters to indicate that they are the population regression coefficients, which are unknown constants that we have to estimate.

Use linear regression to find a sample regression function that seems to fit the data better than the polynomial found in Example 12.3.2. (It's up to you to figure out how high m should be and whether all the terms up to order m should be included. You don't have to find the "best" model, whatever that means, just a good model.) Hand in a plot of your fitted sample regression function with the data points also plotted (like Figure 12.2 and the output from the R `summary` command showing the regression fit. (The R functions for sine and cosine are `sin` and `cos`. The R for $\sin(2x)$ is `sin(2 * x)`, because you need the `*` operator for multiplication. You will also need to wrap such terms in the `I()` function, like `I(sin(2 * x))`.)

12-5. The data set in the URL

```
http://www.stat.umn.edu/geyer/5102/prob12-5.dat
```

has three variables `x1` and `x2` (the predictor variables) and `y` (the response variable).

Fit three models to this data set (1) a "linear" model fitting a polynomial of degree one in the two predictor variables, (2) a "quadratic" model fitting a polynomial of degree two, and (3) a "cubic" model fitting a polynomial of degree three. Don't forget the terms of degree two and three containing products of powers of the two predictor variables.

Print out the ANOVA table for comparing these three models and interpret the P-values in the table. Which model would you say is the best fitting, and why?

12-6. The data set in the URL

```
http://www.stat.umn.edu/geyer/5102/prob12-6.dat
```

has two variables `x` (the predictor variable) and `y` (the response variable). As a glance at a scatter plot of the data done in R by

```
plot(x, y)
```

shows, the relationship between x and y does not appear to be linear. However, it does appear that a so-called *piecewise linear* function with a *knot* at 11 may fit the data well. The means a function having the following three properties.

- It is linear on the interval $x \le 11$.
- It is linear on the interval $x \ge 11$.
- These two linear functions agree at $x = 11$.

Figure out how to fit this model using linear regression. (For some choice of predictor variables, which are functions of x, the regression function of the model is the piecewise linear function described above. Your job is to figure out what predictor variables do this.)

(a) Describe your procedure. What predictors are you using? How many regression coefficients does your procedure have?

(b) Use R to fit the model. Report the parameter estimates (regression coefficients and residual standard error).

The following plot

```
plot(x, y)
lines(x, out$fitted.values)
```

puts a line of predicted values ($\hat{\mathbf{y}}$ in the notation used in the notes and in Lindgren) on the scatter plot. It may help you see when you have got the right thing. You do not have to turn in the plot.

Hint: The `ifelse` function in R defines vectors whose values depend on a condition, for example

```
ifelse(x <= 11, 1, 0)
```

defines the indicator function of the interval $x \leq 11$. (This is *not* one of the predictor variables you need for this problem. It's a hint, but not that much of a hint. The `ifelse` function may be useful, this particular instance is not.)

12-7. For the data in the URL

```
http://www.stat.umn.edu/geyer/5102/ex12.3.2.dat
```

(a) Find the 95% percent prediction interval for an individual with x value 5 using the fifth degree polynomial model fit in Examples 12.3.2 and 12.3.4 (this interval can be read off Figure 12.3, but get the exact numbers from R).

(b) Find the 95% percent confidence interval for the population regression function at the same x value for the same model.

12-8. Prove Theorem 12.13. (Use $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, don't reprove it.)

12-9. The data set

```
http://www.stat.umn.edu/geyer/5102/prob12-9.dat
```

contains three variables, `x`, `y`, and `z`. Each is an i. i. d. sample from some distribution. The three variables are independent of each other (this is not a regression problem). Make Q-Q plots of the variables. One is normal. Which one? Describe the features of the other two plots that make you think the variables plotted are not normal.

It is an interesting comment on the usefulness of Q-Q plots that this problem is essentially undoable at sample size 50 (no differences are apparent in the plots). It's not completely obvious at the sample size 100 used here. Fairly large sample sizes are necessary for Q-Q plots to be useful.

12.8 Bernoulli Regression

As we said in Section 12.5

- Categorical *predictors* are no problem for linear regression. Just use "dummy variables" and proceed normally.

but

- Categorical *responses* do present a problem. Linear regression assumes normally distributed responses. Categorical variables can't be normally distributed.

So now we learn how to deal with at least one kind of categorical response, the simplest, which is Bernoulli.

Suppose the responses are

$$Y_i \sim \text{Ber}(p_i) \tag{12.71}$$

contrast this with the assumptions for linear regression which we write as

$$Y_i \sim \mathcal{N}(\mu_i, \sigma^2) \tag{12.72}$$

and

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} \tag{12.73}$$

Equations (12.72) and (12.73) express the same assumptions as (12.22) and (12.24). We have just rewritten the "strong" regression assumptions in order to bring out the analogy with what we want to do with Bernoulli regression.

The analogy between (12.71) and (12.72) should be clear. Both assume the data are independent, but not identically distributed. The responses Y_i have distributions in the same family, but not the same parameter values. So all we need to finish the specification of a regression-like model for Bernoulli is an equation that takes the place of (12.73).

12.8.1 A Dumb Idea (Identity Link)

We could use (12.73) with the Bernoulli model, although we have to change the symbol for the parameter from $\boldsymbol{\mu}$ to $\mathbf{p}$

$$\mathbf{p} = \mathbf{X}\boldsymbol{\beta}.$$

This means, for example, in the "simple" linear regression model (with one constant and one non-constant predictor x_i)

$$p_i = \alpha + \beta x_i. \tag{12.74}$$

Before we further explain this, we caution that this is universally recognized to be a dumb idea, so don't get too excited about it.

Now nothing is normal, so least squares, t and F tests, and so forth make no sense. But maximum likelihood, the asymptotics of maximum likelihood estimates, and likelihood ratio tests do make sense.

Hence we write down the log likelihood

$$l(\alpha, \beta) = \sum_{i=1}^{n} \bigl[y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \bigr]$$

and its derivatives

$$\begin{aligned}
\frac{\partial l(\alpha, \beta)}{\partial \alpha} &= \sum_{i=1}^{n} \left[\frac{y_i}{p_i} - \frac{1 - y_i}{1 - p_i} \right] \\
\frac{\partial l(\alpha, \beta)}{\partial \beta} &= \sum_{i=1}^{n} \left[\frac{y_i}{p_i} - \frac{1 - y_i}{1 - p_i} \right] x_i
\end{aligned}$$

and set equal to zero to solve for the MLE's. Fortunately, even for this dumb idea, R knows how to do the problem.

Example 12.8.1 (Bernoulli Regression, Identity Link).
We use the data in

```
http://www.stat.umn.edu/geyer/5102/ex12.8.1.dat
```

which has three variables `x`, `y`, and `z`. For now we will just use the first two. The response `y` is Bernoulli (zero-one-valued). We will do a Bernoulli regression using the model assumptions described above, of `y` on `x`. The following code

```
out <- glm(y ~ x, family=quasi(variance="mu(1-mu)"),
    start=c(0.5, 0))
summary(out, dispersion=1)
```

does the regression and prints out a summary. We have to apologize for the rather esoteric syntax, which results from our choice of introducing Bernoulli regression via this rather dumb example. The printout is

```
Call:
glm(formula = y ~ x, family = quasi(variance = "mu(1-mu)"),
    start = c(0.5, 0))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5443  -1.0371  -0.6811   1.1221   1.8214

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.34750    0.19320  -1.799    0.072 .
x            0.01585    0.00373   4.250 2.14e-05 ***
```

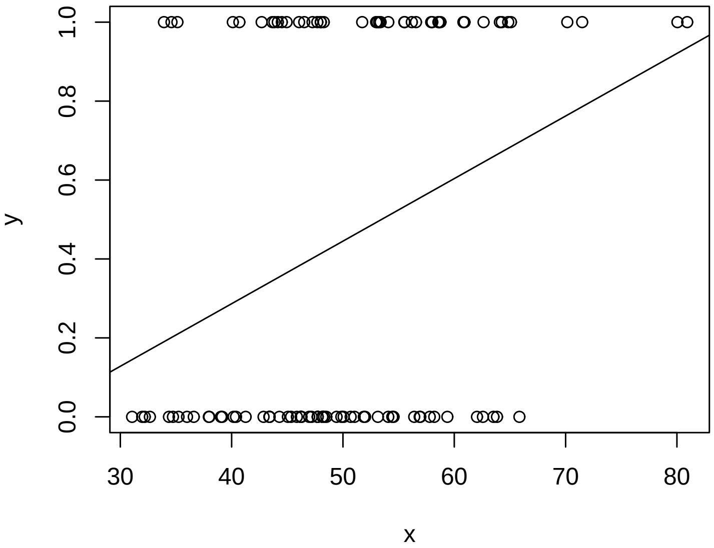


Figure 12.10: Scatter plot and regression line for Example 12.8.1 (Bernoulli regression with an identity link function).

```
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1

(Dispersion parameter for quasi family taken to be 1)

    Null deviance: 137.19  on 99  degrees of freedom
Residual deviance: 126.96  on 98  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 3
```

As usual, our main interest is in the table labeled `Coefficients:`, which says the estimated regression coefficients (the MLE's) are $\hat{\alpha} = -0.34750$ and $\hat{\beta} = 0.01585$. This table also gives standard errors, test statistics ("z values") and P-values for the two-tailed test of whether the true value of the coefficient is zero.

The scatter plot with regression line for this regression is somewhat unusual looking. It is produced by the code

```
plot(x, y)
curve(predict(out, data.frame(x=x)), add=TRUE)
```

and is shown in Figure 12.10. The response values are, of course, being Bernoulli, either zero or one, which makes the scatter plot almost impossible to interpret (it is clear that there are more ones for high x values than for low, but it's impossible to see much else, much less to visualize the correct regression line).

That finishes our discussion of the example. So why is it "dumb"? One reason is that nothing keeps the parameters in the required range. The p_i, being probabilities must be between zero and one. The right hand side of (12.74), being a linear function may take any values between $-\infty$ and $+\infty$. For the data set used in the example, it just happened that the MLE's wound up in $(0, 1)$ without constraining them to do so. In general that won't happen. What then? R being semi-sensible will just crash (produce error messages rather than estimates).

There are various ad-hoc ways one could think to patch up this problem. One could, for example, truncate the linear function at zero and one. But that makes a nondifferentiable log likelihood and ruins the asymptotic theory. The only simple solution is to realize that linearity is no longer simple and give up linearity.

12.8.2 Logistic Regression (Logit Link)

What we need is an assumption about the p_i that will always keep them between zero and one. A great deal of thought by many smart people came up with the following general solution to the problem. Replace the assumption (12.73) for linear regression with the following two assumptions

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} \tag{12.75}$$

and

$$p_i = h(\eta_i) \tag{12.76}$$

where g is a smooth invertible function that maps $\mathbb{R}$ into $(0, 1)$ so the p_i are always in the required range. We now stop for some important terminology.

- The vector $\boldsymbol{\eta}$ in (12.75) is called the *linear predictor.*

- The function h is called the *inverse link function* and its inverse $g = h^{-1}$ is called the *link function.*

The most widely used (though not the only) link function for Bernoulli regression is the *logit* link defined by

$$g(p) = \operatorname{logit}(p) = \log\left(\frac{p}{1-p}\right) \tag{12.77a}$$

$$h(\eta) = g^{-1}(\eta) = \frac{e^\eta}{e^\eta + 1} \tag{12.77b}$$

The right hand equality in (12.77a) defines the so-called *logit* function, and, of course, the right hand inequality in (12.77b) defines the inverse logit function.

For generality, we will not at first use the explicit form of the link function writing the log likelihood

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \bigl[y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \bigr]$$

where we are implicitly using (12.75) and (12.76) as part of the definition. Then

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \left[\frac{y_i}{p_i} - \frac{1 - y_i}{1 - p_i} \right] \frac{\partial p_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}$$

where the two partial derivatives on the right arise from the chain rule and are explicitly

$$\begin{aligned} \frac{\partial p_i}{\partial \eta_i} &= h'(\eta_i) \\ \frac{\partial \eta_i}{\partial \beta_j} &= x_{ij} \end{aligned}$$

where x_{ij} denotes the i, j element of the design matrix $\mathbf{X}$ (the value of the j-th predictor for the i-th individual). Putting everything together

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \left[\frac{y_i}{p_i} - \frac{1 - y_i}{1 - p_i} \right] h'(\eta_i) x_{ij}$$

These equations also do not have a closed form solution, but are easily solved numerically by R

Example 12.8.2 (Bernoulli Regression, Logit Link).
We use the same data in Example 12.8.1. The R commands for logistic regression are

```
out <- glm(y ~ x, family=binomial)
summary(out)
```

Note that the syntax is a lot cleaner for this (logit link) than for the "dumb" way (identity link). The `Coefficients:` table from the printout (the only part we really understand) is

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.56633    1.15871  -3.078  0.00208 **
x            0.06607    0.02257   2.927  0.00342 **
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1
```

The regression function for this "logistic regression" is shown in Figure 12.11, which appears later, after we have done another example.

12.8.3 Probit Regression (Probit Link)

Another widely used link function for Bernoulli regression is the *probit* function, which is just another name for the standard normal inverse c. d. f. That is, the link function is $g(p) = \Phi^{-1}(p)$ and the inverse link function is $g^{-1}(\eta) = \Phi(\eta)$. The fact that we do not have closed-form expressions for these functions and must use table look-up or computer programs to evaluate them is no problem. We need computers to solve the likelihood equations anyway.

Example 12.8.3 (Bernoulli Regression, Probit Link).
We use the same data in Example 12.8.1. The R commands for probit regression are

```
out <- glm(y ~ x, family=binomial(link="probit"))
summary(out)
```

The `Coefficients:` table from the printout is

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.20896    0.68646  -3.218  0.00129 **
x            0.04098    0.01340   3.058  0.00223 **
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1
```

Note that there is a huge difference in the regression coefficients for our three examples, but this should be no surprise because the coefficients for the three regressions are not comparable. Because the regressions involve different link functions, the *meaning* of the regression coefficients are not the same. Comparing them is like comparing apples and oranges, as the saying goes. Thus Bernoulli regression in particular and generalized linear models in general give us yet another reason why *regression coefficients are meaningless.* Note that Figure 12.11 shows that the estimated regression functions $E(Y \mid X)$ are almost identical for the logit and probit regressions despite the regression coefficients being wildly different. Even the linear regression function used in our first example is not so different, at least in the middle of the range of the data, from the other two.

> *Regression functions (response predictions) have a direct probabilistic interpretation $E(Y \mid X)$.*
>
> *Regression coefficients don't.*

The regression function $E(Y \mid X)$ for all three of our Bernoulli regression examples, including this one, are shown in Figure 12.11, which was made by the following code, assuming that the results of the `glm` function for the three examples were saved in `out.quasi`, `out.logit`, and `out.probit`, respectively, rather than all three in `out`.

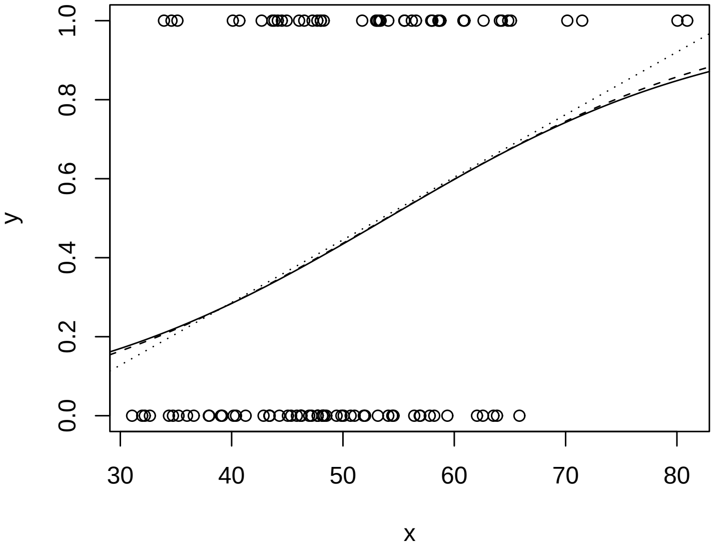


Figure 12.11: Scatter plot and regression functions for Examples 12.8.1, 12.8.2, and 12.8.3. Solid line: regression function for logistic regression (logit link). Dashed line: regression function for probit regression (probit link). Dotted line: regression function for no-name regression (identity link).

```
plot(x, y)
curve(predict(out.logit, data.frame(x=x), type="response"),
   add=TRUE, lty=1)
curve(predict(out.probit, data.frame(x=x), type="response"),
   add=TRUE, lty=2)
curve(predict(out.quasi, data.frame(x=x)), add=TRUE, lty=3)
```

The `type="response"` argument says we want the predicted mean values $g(\boldsymbol{\eta})$, the default being the linear predictor values $\boldsymbol{\eta}$. The reason why this argument is not needed for the last case is because there is no difference with an identity link.

12.9 Generalized Linear Models

A *generalized linear model* (GLM) is a rather general (duh!) form of model that includes ordinary linear regression, logistic and probit regression, and lots more. We keep the regression-like association (12.75) between the regression coefficient vector $\boldsymbol{\beta}$ and the *linear predictor* vector $\boldsymbol{\eta}$ that we used in Bernoulli regression. But now we generalize the probability model greatly. We assume the responses Y_i are independent but not identically distributed with densities of the form

$$f(y \mid \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi / w_i} - c(y, \phi)\right) \tag{12.78}$$

We assume

$$Y_i \sim f(\,\cdot \mid \theta_i, \phi),$$

that is, the *canonical parameter* θ_i is different for each case and is determined (in a way yet to be specified) by the linear predictor η_i but the so-called *dispersion parameter* ϕ is the same for all Y_i. The *weight* w_i is a known positive constant, not a parameter. Also $\phi > 0$ is assumed ($\phi < 0$ would just change the sign of some equations with only trivial effect). The function b is a smooth function but otherwise arbitrary. Given b the function c is determined by the requirement that f integrate to one (like any other probability density).

The log likelihood is thus

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\frac{y_i\theta_i - b(\theta_i)}{\phi / w_i} - c(y_i, \phi)\right) \tag{12.79}$$

Before we proceed to the likelihood equations, let us first look at what the identities derived from differentiating under the integral sign (10.14a) and (10.14b) and their multiparameter analogs (10.44a) and (10.44b) tell us about this model. Note that these identities are exact, not asymptotic, and so can be applied to sample size one and to any parameterization. So let us differentiate one term of (12.79) with respect to its θ parameter

$$\begin{aligned}
l(\theta, \phi) &= \frac{y\theta - b(\theta)}{\phi / w} - c(y, \phi) \\
\frac{\partial l(\theta, \phi)}{\partial \theta} &= \frac{y - b'(\theta)}{\phi / w} \\
\frac{\partial^2 l(\theta, \phi)}{\partial \theta^2} &= -\frac{b''(\theta)}{\phi / w}
\end{aligned}$$

Applied to this particular situation, the identities from differentiating under the integral sign are

$$\begin{aligned}
E_{\theta,\phi}\left\{\frac{\partial l(\theta, \phi)}{\partial \theta}\right\} &= 0 \\
\operatorname{var}_{\theta,\phi}\left\{\frac{\partial l(\theta, \phi)}{\partial \theta}\right\} &= -E_{\theta,\phi}\left\{\frac{\partial^2 l(\theta, \phi)}{\partial \theta^2}\right\}
\end{aligned}$$

or

$$\begin{aligned}
E_{\theta,\phi}\left\{\frac{Y - b'(\theta)}{\phi / w}\right\} &= 0 \\
\operatorname{var}_{\theta,\phi}\left\{\frac{Y - b'(\theta)}{\phi / w}\right\} &= \frac{b''(\theta)}{\phi / w}
\end{aligned}$$

From which we obtain

$$E_{\theta,\phi}(Y) = b'(\theta) \tag{12.80a}$$

$$\operatorname{var}_{\theta,\phi}(Y) = b''(\theta)\frac{\phi}{w} \tag{12.80b}$$

From this we derive the following lemma.

Lemma 12.20. *The function b in* (12.78) *has the following properties*

(i) *b is strictly convex,*

(ii) *b' is strictly increasing,*

(iii) *b'' is strictly positive,*

unless $b''(\theta) = 0$ for all θ and the distribution of Y is concentrated at one point for all parameter values.

Proof. Just by ordinary calculus (iii) implies (ii) implies (i), so we need only prove (iii). Equation (12.80b) and the assumptions $\phi > 0$ and $w > 0$ imply $b''(\theta) \ge 0$. So the only thing left to prove is that if $b''(\theta^*) = 0$ for any one θ^*, then actually $b''(\theta) = 0$ for all θ. By (12.80b) $b''(\theta^*) = 0$ implies $\operatorname{var}_{\theta^*,\phi}(Y) = 0$, so the distribution of Y for the parameter values θ^* and ϕ is concentrated at one point. But now we apply a trick using the distribution at θ^* to calculate for other θ

$$\begin{aligned} f(y \mid \theta, \phi) &= \frac{f(y \mid \theta, \phi)}{f(y \mid \theta^*, \phi)} f(y \mid \theta^*, \phi) \\ &= \exp\left(\frac{y\theta - b(\theta)}{\phi/w_i} - \frac{y\theta^* - b(\theta^*)}{\phi/w_i} \right) f(y \mid \theta^*, \phi) \end{aligned}$$

The exponential term is strictly positive, so the only way the distribution of Y can be concentrated at one point and have variance zero for $\theta = \theta^*$ is if the distribution is concentrated at the same point and hence has variance zero for all other θ. And using (12.80b) again, this would imply $b''(\theta) = 0$ for all θ. □

The "unless" case in the lemma is uninteresting. We never use probability models for data having distributions concentrated at one point (that is, constant random variables). Thus (i), (ii), and (iii) of the lemma hold for any GLM we would actually want to use. The most important of these is (ii) for a reason that will be explained when we return to the general theory after the following example.

Example 12.9.1 (Binomial Regression).
We generalize Bernoulli regression just a bit by allowing more than one one Bernoulli variable to go with each predictor value $\mathbf{x}_i$. Adding those Bernoullis gives a binomial response, that is, we assume

$$Y_i \sim \operatorname{Bin}(m_i, p_i)$$

where m_i is the number of Bernoulli variables with predictor vector $\mathbf{x}_i$. The density for Y_i is

$$f(y_i \mid p_i) = \binom{m_i}{y_i} p_i^{y_i}(1 - p_i)^{1-y_i}$$

we try to match this up with the GLM form. So first we write the density as an exponential

$$\begin{aligned} f(y_i \mid p_i) &= \exp\left[y_i \log(p_i) + (m_i - y_i)\log(1 - p_i) + \log\binom{m_i}{y_i}\right] \\ &= \exp\left[y_i \log\left(\frac{p_i}{1 - p_i}\right) + m_i \log(1 - p_i) + \log\binom{m_i}{y_i}\right] \\ &= \exp\left\{m_i\bigl[\bar{y}_i\theta_i - b(\theta_i)\bigr] + \log\binom{m_i}{y_i}\right\} \end{aligned}$$

where we have defined

$$\begin{aligned} \bar{y}_i &= y_i/m_i \\ \theta_i &= \operatorname{logit}(p_i) \\ b(\theta_i) &= -\log(1 - p_i) \end{aligned}$$

So we see that

- The *canonical parameter* for the binomial model is $\theta = \operatorname{logit}(p)$. That explains why the logit link is popular.
- The *weight* w_i in the GLM density turns out to be the number of Bernoullis m_i associated with the i-th predictor value. So we see that the weight allows for grouped data like this.
- There is nothing like a dispersion parameter here. For the binomial family the dispersion is known; $\phi = 1$.

Returning to the general GLM model (a doubly redundant redundancy), we first define yet another parameter, the *mean value parameter*

$$\mu_i = E_{\theta_i,\phi}(Y_i) = b'(\theta_i).$$

By (ii) of Lemma 12.20 b' is a strictly increasing function, hence an invertible function. Thus the mapping between the canonical parameter θ and the mean value parameter μ is an invertible change of parameter. Then by definition of "link function" the relation between the mean value parameter μ_i and the linear predictor η_i is given by the link function

$$\eta_i = g(\mu_i).$$

The link function g is required to be a strictly increasing function, hence an invertible change of parameter.

If, as in logistic regression we take the linear predictor to be the canonical parameter, that determines the link function, because $\eta_i = \theta_i$ implies $g^{-1}(\theta) = b'(\theta)$. In general, as is the case in probit regression, the link function g and the function b' that connects the canonical and mean value parameters are unrelated.

It is traditional in GLM theory to make primary use of the mean value parameter and not use the canonical parameter (unless it happens to be the same as the linear predictor). For that reason we want to write the variance as a function of μ rather than θ

$$\operatorname{var}_{\theta_i,\phi}(Y_i) = \frac{\phi}{w} V(\mu_i) \tag{12.81}$$

where

$$V(\mu) = b''(\theta) \qquad \text{when} \qquad \mu = b'(\theta)$$

This definition of the function V makes sense because the function b' is an invertible mapping between mean value and canonical parameters. The function V is called the *variance function* even though it is only proportional to the variance, the complete variance being $\phi V(\mu)/w$.

12.9.1 Parameter Estimation

Now we can write out the log likelihood derivatives

$$\begin{aligned}
\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} &= \sum_{i=1}^n \left(\frac{y_i - b'(\theta_i)}{\phi / w_i} \right) \frac{\partial \theta_i}{\partial \beta_j} \\
&= \sum_{i=1}^n \left(\frac{y_i - \mu_i}{\phi / w_i} \right) \frac{\partial \theta_i}{\partial \beta_j}
\end{aligned}$$

In order to completely eliminate θ_i we need to calculate the partial derivative. First note that

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i)$$

so by the inverse function theorem

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b''(\theta_i)} = \frac{1}{V(\mu_i)}$$

Now we can write

$$\frac{\partial \theta_i}{\partial \beta_j} = \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \frac{1}{V(\mu_i)} h'(\eta_i) x_{ij} \tag{12.82}$$

where $h = g^{-1}$ is the inverse link function. And we finally arrive at the likelihood equations expressed in terms of the mean value parameter and the linear predictor

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^n \left(\frac{y_i - \mu_i}{V(\mu_i)} \right) w_i h'(\eta_i) x_{ij}$$

These are the equations the computer sets equal to zero and solves to find the regression coefficients. Note that the dispersion parameter ϕ appears only multiplicatively. So it cancels when the partial derivatives are set equal to zero. Thus the regression coefficients can be estimated without estimating the dispersion (just as in linear regression).

Also as in linear regression, the dispersion parameter is not estimated by maximum likelihood but by the method of moments. By (12.81)

$$E\left\{\frac{w_i(Y_i - \mu_i)^2}{V(\mu_i)}\right\} = \frac{w_i}{V(\mu_i)} \operatorname{var}(Y_i) = \phi$$

Thus

$$\frac{1}{n} \sum_{i=1}^{n} \frac{w_i(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

would seem to be an approximately unbiased estimate of ϕ. Actually it is not because $\hat{\boldsymbol{\mu}}$ is not $\boldsymbol{\mu}$, and

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{w_i(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

is closer to unbiased where p is the rank of the design matrix $\mathbf{X}$. We won't bother to prove this. The argument is analogous to the reason for $n - p$ in linear regression.

12.9.2 Fisher Information, Tests and Confidence Intervals

The log likelihood second derivatives are

$$\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\beta_j \partial\beta_k} = \sum_{i=1}^{n} \left(\frac{y_i - b'(\theta_i)}{\phi/w_i}\right) \frac{\partial^2 \theta_i}{\partial\beta_j \partial\beta_k} - \sum_{i=1}^{n} \left(\frac{b''(\theta_i)}{\phi/w_i}\right) \frac{\partial\theta_i}{\partial\beta_j} \frac{\partial\theta_i}{\partial\beta_k}$$

This is rather a mess, but because of (12.80a) the expectation of the first sum is zero. Thus the j, k term of the expected Fisher information is, using (12.82) and $b'' = V$,

$$\begin{aligned} -E\left\{\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\beta_j \partial\beta_k}\right\} &= \sum_{i=1}^{n} \left(\frac{b''(\theta_i)}{\phi/w_i}\right) \frac{\partial\theta_i}{\partial\beta_j} \frac{\partial\theta_i}{\partial\beta_k} \\ &= \sum_{i=1}^{n} \left(\frac{V(\mu_i)}{\phi/w_i}\right) \frac{1}{V(\mu_i)} h'(\eta_i) x_{ij} \frac{1}{V(\mu_i)} h'(\eta_i) x_{ik} \\ &= \frac{1}{\phi} \sum_{i=1}^{n} \left(\frac{w_i h'(\eta_i)^2}{V(\mu_i)}\right) x_{ij} x_{ik} \end{aligned}$$

We can write this as a matrix equation if we define $\mathbf{D}$ to be the diagonal matrix with i, i element

$$d_{ii} = \frac{1}{\phi} \sum_{i=1}^{n} \frac{w_i h'(\eta_i)^2}{V(\mu_i)}$$

Then

$$\mathbf{I}(\boldsymbol{\beta}) = \mathbf{X}'\mathbf{D}\mathbf{X}$$

is the expected Fisher information matrix. From this standard errors for the parameter estimates, confidence intervals, test statistics, and so forth can be derived using the usual likelihood theory. Fortunately, we do not have to do all of this by hand. R knows all the formulas and computes them for us.

12.10 Poisson Regression

The Poisson model is also a GLM. We assume responses

$$Y_i \sim \text{Poi}(\mu_i)$$

and connection between the linear predictor and regression coefficients, as always, of the form (12.75). We only need to identify the link and variance functions to get going. It turns out that the canonical link function is the log function (Problem 12-13). The Poisson distribution distribution has the relation

$$\text{var}(Y) = E(Y) = \mu$$

connecting the mean, variance, and mean value parameter. Thus the variance function is $V(\mu) = \mu$, the dispersion parameter is known ($\phi = 1$), and the weight is also unity ($w = 1$).

Example 12.10.1 (Poisson Regression).
The data set

```
http://www.stat.umn.edu/geyer/5102/ex12.10.1.dat
```

simulates the hourly counts from a not necessarily homogeneous Poisson process. The variables are `hour` and `count`, the first counting hours sequentially throughout a 14-day period (running from 1 to $14 \times 24 = 336$) and the second giving the count for that hour.

The idea of the regression is to get a handle on the mean as a function of time if it is not constant. Many time series have a daily cycle. If we pool the counts for the same hour of the day over the 14 days of the series, we see a clear pattern in the histogram (Figure 12.12). In contrast, if we pool the counts for each day of the week, the histogram is fairly even (not shown). Thus it seems to make sense to model the mean function as being periodic with period 24 hours, and the obvious way to do that is to use trigonometric functions. Let us do a bunch of fits

```
w <- hour / 24 * 2 * pi
out1 <- glm(count ~ I(sin(w)) + I(cos(w)), family=poisson)
summary(out1)
out2 <- glm(count ~ I(sin(w)) + I(cos(w)) + I(sin(2 * w))
   + I(cos(2 * w)), family=poisson)
```

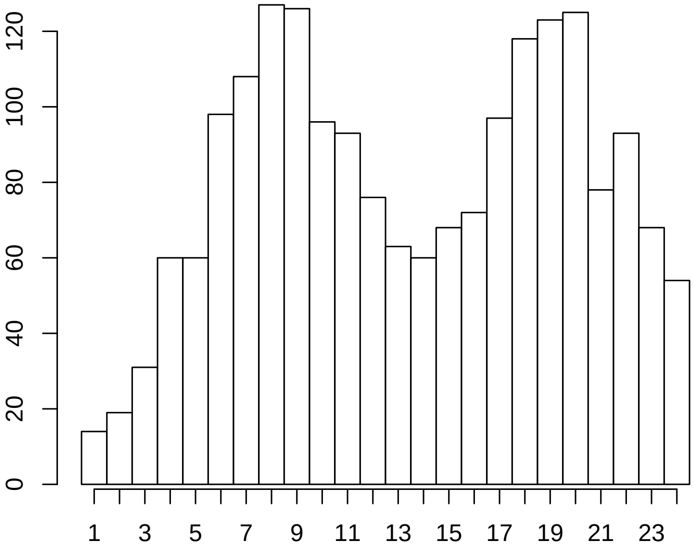


Figure 12.12: Histogram of the total count in each hour of the day for the data for Example 12.10.1.

```
summary(out2)
out3 <- glm(count ~ I(sin(w)) + I(cos(w)) + I(sin(2 * w))
   + I(cos(2 * w)) + I(sin(3 * w)) + I(cos(3 * w)),
   family=poisson)
summary(out3)
```

The `Coefficient:` tables from the printouts are

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.73272    0.02310   75.02  < 2e-16 ***
I(sin(w))   -0.10067    0.03237   -3.11  0.00187 **
I(cos(w))   -0.21360    0.03251   -6.57 5.02e-11 ***

              Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.65917    0.02494  66.521  < 2e-16 ***
I(sin(w))     -0.13916    0.03128  -4.448 8.65e-06 ***
I(cos(w))     -0.28510    0.03661  -7.788 6.82e-15 ***
I(sin(2 * w)) -0.42974    0.03385 -12.696  < 2e-16 ***
I(cos(2 * w)) -0.30846    0.03346  -9.219  < 2e-16 ***

               Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.655430   0.025149  65.826  < 2e-16 ***
I(sin(w))     -0.151196   0.032530  -4.648 3.35e-06 ***
I(cos(w))     -0.301336   0.038244  -7.879 3.29e-15 ***
```

```
I(sin(2 * w)) -0.439789   0.034461 -12.762  < 2e-16 ***
I(cos(2 * w)) -0.312843   0.033919  -9.223  < 2e-16 ***
I(sin(3 * w)) -0.063440   0.033803  -1.877   0.0606 .
I(cos(3 * w))  0.004311   0.033630   0.128   0.8980
```

with the usual "`Signif. codes`". It seems from the pattern of "stars" that maybe it is time to stop. A clearer indication is given by the so-called *analysis of deviance* table, "deviance" being another name for the likelihood ratio test statistic (twice the log likelihood difference between big and small models), which has an asymptotic chi-square distribution by standard likelihood theory.

```
anova(out1, out2, out3, test="Chisq")
```

prints out

```
Analysis of Deviance Table

Model 1: count ~ I(sin(w)) + I(cos(w))
Model 2: count ~ I(sin(w)) + I(cos(w)) + I(sin(2 * w)) + I(cos(2 * w))
Model 3: count ~ I(sin(w)) + I(cos(w)) + I(sin(2 * w)) + I(cos(2 * w)) +
    I(sin(3 * w)) + I(cos(3 * w))
  Resid. Df Resid. Dev  Df Deviance P(>|Chi|)
1       333     651.10
2       331     399.58   2   251.52 2.412e-55
3       329     396.03   2     3.55      0.17
```

The approximate P-value for the likelihood ratio test comparing models 1 and 2 is $P \approx 0$, which clearly indicates that model 1 should be rejected. The approximate P-value for the likelihood ratio test comparing models 2 and 3 is $P = 0.17$, which fairly clearly indicates that model 1 should be accepted and that model 3 is unnecessary. $P = 0.17$ indicates exceedingly weak evidence favoring the larger model. Thus we choose model 2.

The following code

```
hourofday <- (hour - 1) %% 24 + 1
plot(hourofday, count, xlab="hour of the day")
curve(predict(out2, data.frame(w=x/24*2*pi), type="response"),
   add=TRUE)
```

draws the scatter plot and estimated regression function for model 2 (Figure 12.13).

I hope all readers are impressed by how magically statistics works in this example. A glance at Figure 12.13 shows

- Poisson regression is obviously doing more or less the right thing,

- there is no way one could put in a sensible regression function without using theoretical statistics. The situation is just too complicated.

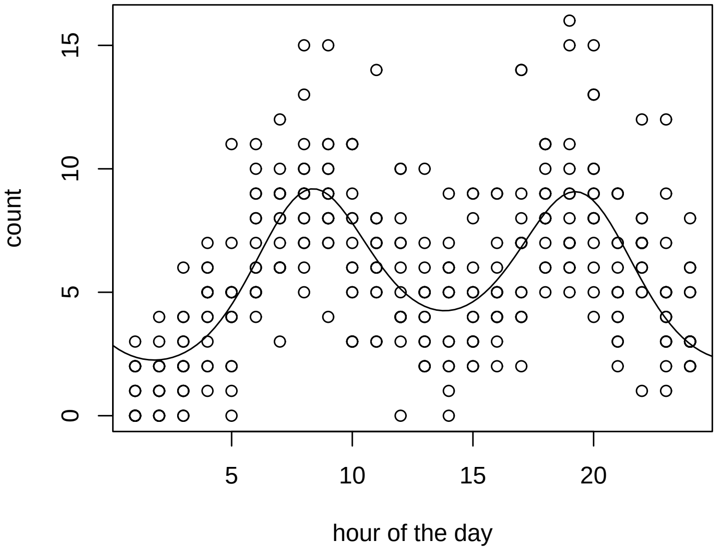


Figure 12.13: Scatter plot and regression curve for Example 12.10.1 (Poisson regression with log link function). The regression function is trigonometric on the scale of the linear predictor with terms up to the frequency 2 per day.

12.11 Overdispersion

So far we have seen only models with unit dispersion parameter ($\phi = 1$). This section gives an example with $\phi \neq 1$ so we can see the point of the dispersion parameter.

The reason $\phi = 1$ for binomial regression is that the mean value parameter $p = \mu$ determines the variance $mp(1-p) = m\mu(1-\mu)$. Thus the variance function is

$$V(\mu) = \mu(1 - \mu) \tag{12.83}$$

and the weights are $w_i = m_i$, the sample size for each binomial variable (this was worked out in detail in Example 12.9.1).

But what if the model is wrong? Here is another model. Suppose

$$Y_i \mid W_i \sim \text{Bin}(m_i, W_i)$$

where the W_i are i. i. d. random variables with mean μ and variance τ^2. Then by the usual rules for conditional probability (Axiom CE2 and Theorem 3.7 in Chapter 3 of these notes)

$$E(Y_i) = E\{E(Y_i \mid W_i)\} = E(m_i W_i) = m_i \mu$$

and

$$\begin{aligned}
\operatorname{var}(Y_i) &= E\{\operatorname{var}(Y_i \mid W_i)\} + \operatorname{var}\{E(Y_i \mid W_i)\} \\
&= E\{m_i W_i (1 - W_i)\} + \operatorname{var}\{m_i W_i\} \\
&= m_i \mu - m_i E(W_i^2) + m_i^2 \tau^2 \\
&= m_i \mu - m_i (\tau^2 + \mu^2) + m_i^2 \tau^2 \\
&= m_i \mu (1 - \mu) + m_i (m_i - 1) \tau^2
\end{aligned}$$

This is clearly larger than the formula $m_i \mu(1-\mu)$ one would have for the binomial model. Since the variance is always larger than one would have under the binomial model.

So we know that if our response variables Y_i are the sum of a random mixture of Bernoullis rather than i. i. d. Bernoullis, we will have overdispersion. But how to model the overdispersion? The GLM model offers a simple solution. Allow for general ϕ so we have, defining $\overline{Y}_i = Y_i / m_i$

$$\begin{aligned}
E(\overline{Y}_i) &= \mu_i \\
\operatorname{var}(\overline{Y}_i) &= \frac{\phi}{m_i} \mu_i (1 - \mu_i) \\
&= \frac{\phi}{m_i} V(\mu_i)
\end{aligned}$$

where V is the usual binomial variance function (12.83).

Example 12.11.1 (Overdispersed Binomial Regression).
The data set

```
http://www.stat.umn.edu/geyer/5102/ex12.11.1.dat
```

contains some data for an overdispersed binomial model. The commands

```
y <- cbind(succ, fail)
out.binom <- glm(y ~ x, family=binomial)
summary(out.binom)
out.quasi <- glm(y ~ x, family=quasibinomial)
summary(out.quasi)
```

fit both the binomial model (logit link and $\phi = 1$) and the "quasi-binomial" (logit link again but ϕ is estimated with the method of moments estimator as explained in the text). Both models have exactly the same maximum likelihood regression coefficients, but because the dispersions differ, the standard errors, z-values, and P-values differ.

The relevant part of the binomial output

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.92155    0.35260  -5.450 5.05e-08 ***
```

```
x            0.07436    0.01227   6.062 1.35e-09 ***
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1

(Dispersion parameter for binomial family taken to be 1)
```

and the relevant part of the quasi-binomial output

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.92155    0.41569  -4.623 2.88e-05 ***
x            0.07436    0.01446   5.141 4.97e-06 ***
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1

(Dispersion parameter for quasibinomial family taken to be 1.38992)
```

Your humble author finds this a bit unsatisfactory. If the data are really overdispersed, then the standard errors and so forth from the latter output are the right ones to use. But since the dispersion was not estimated by maximum likelihood, there is no likelihood ratio test for comparing the two models. Nor could your author find any other test in a brief examination of the literature. Apparently, if one is worried about overdispersion, one should use the model that allows for it. And if not, not. But that's not the way we operate in the rest of statistics. I suppose I need to find out more about overdispersion.

Problems

12-10. Show that (12.77a) and (12.77b) do indeed define a pair of inverse functions.

12-11. Do calculations similar to Example 12.9.1 for the normal problem

$$Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$$

identifying (a) the canonical parameter θ, the dispersion parameter ϕ, and the weight w_i.

12-12. The data set

`http://www.stat.umn.edu/geyer/5102/ex12.8.1.dat`

contains another predictor vector `z` besides the ones we used in Examples 12.8.1, 12.8.2, and 12.8.3. Perform the logistic regression of `y` on `x` and `z`. Perform a test comparing this new model and the one fit in Example 12.8.2 giving the P-value for the test and the conclusion as to which model the test accepts.

12-13. Do calculations similar to Example 12.9.1 for the Poisson model, showing that the canonical parameter for the Poisson distribution is $\theta = \log(\mu)$.

Appendix A

Greek Letters

Table A.1: Table of Greek Letters (Continued on following page.)

name	capital letter	small letter	pronunciation	sound
alpha	A	α	AL-fah	short a
beta	B	β	BAY-tah	b
gamma	Γ	γ	GAM-ah	g
delta	Δ	δ	DEL-tah	d
epsilon	E	ϵ	EP-si-lon	e
zeta	Z	ζ	ZAY-tah	z
eta	H	η	AY-tah	long a
theta	Θ	θ or ϑ	THAY-thah	soft th (as in thin)
iota	I	ι	EYE-oh-tah	i
kappa	K	κ	KAP-ah	k
lambda	Λ	λ	LAM-dah	l
mu	M	μ	MYOO	m
nu	N	ν	NOO	n
xi	Ξ	ξ	KSEE	x (as in box)
omicron	O	o	OH-mi-kron	o
pi	Π	π	PIE	p
rho	R	ρ	RHOH	rh[1]
sigma	Σ	σ	SIG-mah	s
tau	T	τ	TAOW	t
upsilon	Υ	υ	UP-si-lon	u

[1] The sound of the Greek letter ρ is not used in English. English words, like *rhetoric* and *rhinoceros* that are descended from Greek words beginning with ρ have English pronunciations beginning with an “r” sound rather than “rh” (though the spelling reminds us of the Greek origin).

Table A.2: Table of Greek Letters (Continued.)

	capital	small		
name	letter	letter	pronunciation	sound
phi	Φ	ϕ or φ	FIE	f
chi	X	χ	KIE	guttural ch[2]
psi	Ψ	ψ	PSY	ps (as in stops)[3]
omega	Ω	ω	oh-MEG-ah	o

[2] The sound of the Greek letter χ is not used in English. It is heard in the German *Buch* or Scottish *loch*. English words, like *chemistry* and *chorus* that are descended from Greek words beginning with χ have English pronunciations beginning with a “k” sound rather than “guttural ch” (though the spelling reminds us of the Greek origin).

[3] English words, like *pseudonym* and *psychology* that are descended from Greek words beginning with ψ have English pronunciations beginning with an “s” sound rather than “ps” (though the spelling reminds us of the Greek origin).

Appendix B

Summary of Brand-Name Distributions

B.1 Discrete Distributions

B.1.1 The Discrete Uniform Distribution

The Abbreviation $\mathcal{DU}(S)$.

The Sample Space Any finite set S.

The Density

$$f(x) = \frac{1}{n}, \qquad x \in S,$$

where $n = \text{card}(S)$.

Specialization The case in which the sample space consists of consecutive integers $S = \{m, m+1, \ldots, n\}$ is denoted $\mathcal{DU}(m, n)$.

Moments If $X \sim \mathcal{DU}(1, n)$, then

$$\begin{aligned} E(X) &= \frac{n+1}{2} \\ \text{var}(X) &= \frac{n^2 - 1}{12} \end{aligned}$$

B.1.2 The Binomial Distribution

The Abbreviation $\text{Bin}(n, p)$

The Sample Space The integers 0, …, n.

The Parameter p such that $0 < p < 1$.

The Density

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, \ldots, n.$$

Moments

$$\begin{aligned} E(X) &= np \\ \operatorname{var}(X) &= np(1-p) \end{aligned}$$

Specialization

$$\operatorname{Ber}(p) = \operatorname{Bin}(1, p)$$

B.1.3 The Geometric Distribution, Type II

Note This section has changed. The roles of p and $1 - p$ have been reversed, and the abbreviation Geo(p) is no longer used to refer to this distribution but the distribution defined in Section B.1.8. All of the changes are to match up with Chapter 6 in Lindgren.

The Abbreviation No abbreviation to avoid confusion with the other type defined in Section B.1.8.

Relation Between the Types If $X \sim \operatorname{Geo}(p)$, then $Y = X - 1$ has the distribution defined in this section.

X is the number of *trials* before the first success in an i. i. d. sequence of Ber(p) random variables. Y is the number of *failures* before the first success.

The Sample Space The integers 0, 1, ….

The Parameter p such that $0 < p < 1$.

The Density

$$f(x) = p(1-p)^x, \qquad x = 0, 1, \ldots.$$

Moments

$$\begin{aligned} E(X) &= \frac{1}{p} - 1 = \frac{1-p}{p} \\ \operatorname{var}(X) &= \frac{1-p}{p^2} \end{aligned}$$

B.1.4 The Poisson Distribution

The Abbreviation Poi(μ)

The Sample Space The integers 0, 1, ...

The Parameter μ such that $\mu > 0$.

The Density

$$f(x) = \frac{\mu^x}{x!} e^{-\mu}, \qquad x = 0, 1, \ldots.$$

Moments

$$\begin{aligned} E(X) &= \mu \\ \operatorname{var}(X) &= \mu \end{aligned}$$

B.1.5 The Bernoulli Distribution

The Abbreviation Ber(p)

The Sample Space The integers 0 and 1.

The Parameter p such that $0 < p < 1$.

The Density

$$f(x) = \begin{cases} p, & x = 1 \\ 1 - p & x = 0 \end{cases}$$

Moments

$$\begin{aligned} E(X) &= p \\ \operatorname{var}(X) &= p(1 - p) \end{aligned}$$

Generalization

$$\operatorname{Ber}(p) = \operatorname{Bin}(1, p)$$

B.1.6 The Negative Binomial Distribution, Type I

The Abbreviation NegBin(k, p)

The Sample Space The integers k, $k + 1$, ...

The Parameter p such that $0 < p < 1$.

The Density

$$f(x) = \binom{x-1}{k-1} p^k (1-p)^{x-k}, \qquad x = k, k+1, \ldots.$$

Moments

$$\begin{aligned} E(X) &= \frac{k}{p} \\ \operatorname{var}(X) &= \frac{k(1-p)}{p^2} \end{aligned}$$

Specialization

$$\operatorname{Geo}(p) = \operatorname{NegBin}(1, p)$$

B.1.7 The Negative Binomial Distribution, Type II

The Abbreviation No abbreviation to avoid confusion with the other type defined in Section B.1.6.

Relation Between the Types If $X \sim \operatorname{NegBin}(k, p)$, then $Y = X - k$ has the distribution defined in this section.

X is the number of *trials* before the k-th success in an i. i. d. sequence of $\operatorname{Ber}(p)$ random variables. Y is the number of *failures* before the k-th success.

The Sample Space The integers 0, 1, ...

The Parameter p such that $0 < p < 1$.

The Density

$$f(y) = \binom{y+k-1}{k-1} p^k (1-p)^{y}, \qquad y = 0, 1, \ldots.$$

Moments

$$\begin{aligned} E(X) &= \frac{k}{p} - k = \frac{k(1-p)}{p} \\ \operatorname{var}(X) &= \frac{k(1-p)}{p^2} \end{aligned}$$

B.1.8 The Geometric Distribution, Type I

The Abbreviation $\operatorname{Geo}(p)$

The Sample Space The integers 1, 2, ...

The Parameter p such that $0 < p < 1$.

The Density

$$f(x) = p(1 - p)^{x-1}, \qquad x = 1, 2, \ldots.$$

Moments

$$E(X) = \frac{1}{p}$$

$$\operatorname{var}(X) = \frac{1 - p}{p^2}$$

Generalization

$$\operatorname{Geo}(p) = \operatorname{NegBin}(1, p)$$

B.2 Continuous Distributions

B.2.1 The Uniform Distribution

The Abbreviation $\mathcal{U}(S)$.

The Sample Space Any subset S of $\mathbb{R}^d$.

The Density

$$f(x) = \frac{1}{c}, \qquad x \in S,$$

where

$$c = m(S) = \int_S dx$$

is the measure of S (length in $\mathbb{R}^1$, area in $\mathbb{R}^2$, volume in $\mathbb{R}^3$, and so forth).

Specialization The case having $S = (a, b)$ in $\mathbb{R}^1$ and density

$$f(x) = \frac{1}{b - a}, \qquad a < x < b$$

is denoted $\mathcal{U}(a, b)$.

Moments If $X \sim \mathcal{U}(a, b)$, then

$$E(X) = \frac{a + b}{2}$$

$$\operatorname{var}(X) = \frac{(b - a)^2}{12}$$

B.2.2 The Exponential Distribution

The Abbreviation Exp(λ).

The Sample Space The interval $(0, \infty)$ of the real numbers.

The Parameter λ such that $\lambda > 0$.

The Density

$$f(x) = \lambda e^{-\lambda x}, \qquad x > 0.$$

Moments

$$\begin{aligned} E(X) &= \frac{1}{\lambda} \\ \operatorname{var}(X) &= \frac{1}{\lambda^2} \end{aligned}$$

Generalization

$$\operatorname{Exp}(\lambda) = \operatorname{Gam}(1, \lambda)$$

B.2.3 The Gamma Distribution

The Abbreviation Gam(α, λ).

The Sample Space The interval $(0, \infty)$ of the real numbers.

The Parameters α and λ such that $\alpha > 0$ and $\lambda > 0$.

The Density

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha - 1} e^{-\lambda x}, \qquad x > 0.$$

where $\Gamma(\alpha)$ is the gamma function (Section B.3.1 below).

Moments

$$\begin{aligned} E(X) &= \frac{\alpha}{\lambda} \\ \operatorname{var}(X) &= \frac{\alpha}{\lambda^2} \end{aligned}$$

Specialization

$$\begin{aligned} \operatorname{Exp}(\lambda) &= \operatorname{Gam}(1, \lambda) \\ \operatorname{chi}^2(k) &= \operatorname{Gam}\left(\tfrac{k}{2}, \tfrac{1}{2}\right) \end{aligned}$$

B.2.4 The Beta Distribution

The Abbreviation Beta(s, t).

The Sample Space The interval $(0, 1)$ of the real numbers.

The Parameters s and t such that $s > 0$ and $t > 0$.

The Density

$$f(x) = \frac{1}{B(s,t)} x^{s-1}(1-x)^{t-1} \qquad 0 < x < 1.$$

where $B(s, t)$ is the *beta function* defined by

$$B(s,t) = \frac{\Gamma(s)\Gamma(t)}{\Gamma(s+t)} \tag{B.1}$$

Moments

$$\begin{aligned} E(X) &= \frac{s}{s+t} \\ \operatorname{var}(X) &= \frac{st}{(s+t)^2(s+t+1)} \end{aligned}$$

B.2.5 The Normal Distribution

The Abbreviation $\mathcal{N}(\mu, \sigma^2)$.

The Sample Space The real line $\mathbb{R}$.

The Parameters μ and σ^2 such that $\sigma^2 > 0$.

The Density

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad x \in \mathbb{R}.$$

Moments

$$\begin{aligned} E(X) &= \mu \\ \operatorname{var}(X) &= \sigma^2 \\ \mu_4 &= 3\sigma^4 \end{aligned}$$

B.2.6 The Chi-Square Distribution

The Abbreviation chi$^2(k)$.

The Sample Space The interval $(0, \infty)$ of the real numbers.

The Parameter A positive integer k.

The Density

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}, \qquad x > 0.$$

Moments

$$\begin{aligned} E(X) &= k \\ \operatorname{var}(X) &= 2k \end{aligned}$$

Generalization

$$\operatorname{chi}^2(k) = \operatorname{Gam}\left(\tfrac{k}{2}, \tfrac{1}{2}\right)$$

B.2.7 The Cauchy Distribution

The Abbreviation Cauchy(μ, σ).

The Sample Space The real line $\mathbb{R}$.

The Parameters μ and σ such that $\sigma > 0$.

The Density

$$f(x) = \frac{1}{\pi} \cdot \frac{\sigma}{\sigma^2 + (x - \mu)^2}, \qquad x \in \mathbb{R}.$$

Moments None: $E(|X|) = \infty$.

B.2.8 Student's t Distribution

The Abbreviation $t(\nu)$.

The Sample Space The real line $\mathbb{R}$.

The Parameters ν such that $\nu > 0$, called the "degrees of freedom" of the distribution.

The Density

$$f_\nu(x) = \frac{1}{\sqrt{\nu}} \cdot \frac{1}{B(\frac{\nu}{2}, \frac{1}{2})} \cdot \frac{1}{\left(1 + \frac{x^2}{\nu}\right)^{(\nu+1)/2}}, \qquad -\infty < x < +\infty$$

where $B(s, t)$ is the beta function defined by (B.1).

Moments If $\nu > 1$, then

$$E(X) = 0.$$

Otherwise the mean does not exist. If $\nu > 2$, then

$$\operatorname{var}(X) = \frac{\nu}{\nu - 2}.$$

Otherwise the variance does not exist.

Specialization

$$t(1) = \text{Cauchy}(0, 1)$$

and in a manner of speaking

$$t(\infty) = \mathcal{N}(0, 1)$$

(see Theorem 7.21 of Chapter 7 of these notes).

B.2.9 Snedecor's F Distribution

The Abbreviation $F(\mu, \nu)$.

The Sample Space The interval $(0, \infty)$ of the real numbers.

The Parameters μ and ν such that $\mu > 0$ and $\nu > 0$, called the "numerator degrees of freedom" of the the "denominator degrees of freedom" of the distribution, respectively.

The Density Not derived in these notes.

Moments If $\nu > 2$, then

$$E(X) = \frac{\nu}{\nu - 2}.$$

Otherwise the mean does not exist.

The variance is not derived in these notes.

Relation to the Beta Distribution

$$X \sim F(\mu, \nu)$$

if and only if

$$W \sim \text{Beta}\left(\frac{\mu}{2}, \frac{\nu}{2}\right),$$

where

$$W = \frac{\frac{\mu}{\nu} X}{1 + \frac{\mu}{\nu} X}$$

B.3 Special Functions

B.3.1 The Gamma Function

The Definition

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} \, dx, \qquad \alpha > 0 \tag{B.2}$$

The Recursion Relation

$$\Gamma(\alpha + 1) = \alpha \Gamma(\alpha) \tag{B.3}$$

Known Values

$$\Gamma(1) = 1$$

and hence using the recursion relation

$$\Gamma(n + 1) = n!$$

for any nonnegative integer n.

Also

$$\Gamma(\tfrac{1}{2}) = \sqrt{\pi}$$

and hence using the recursion relation

$$\begin{aligned} \Gamma(\tfrac{3}{2}) &= \tfrac{1}{2}\sqrt{\pi} \\ \Gamma(\tfrac{5}{2}) &= \tfrac{3}{2} \cdot \tfrac{1}{2}\sqrt{\pi} \\ \Gamma(\tfrac{7}{2}) &= \tfrac{5}{2} \cdot \tfrac{3}{2} \cdot \tfrac{1}{2}\sqrt{\pi} \end{aligned}$$

and so forth.

B.3.2 The Beta Function

The function $B(s, t)$ defined by (B.1).

B.4 Discrete Multivariate Distributions

B.4.1 The Multinomial Distribution

The Abbreviation $\text{Multi}_k(n, \mathbf{p})$ or $\text{Multi}(n, \mathbf{p})$ if the dimension k is clear from context.

The Sample Space

$$S = \{\, \mathbf{y} \in \mathbb{N}^k : y_1 + \cdots y_k = n \,\}$$

where $\mathbb{N}$ denotes the "natural numbers" 0, 1, 2, $\ldots$.

The Parameter $\mathbf{p} = (p_1, \ldots, p_k)$ such that $p_i \geq 0$ for all i and $\sum_i p_i = 1$.

The Density

$$f(\mathbf{y}) = \binom{n}{y_1, \ldots, y_k} \prod_{j=1}^{k} p_j^{y_j}, \qquad \mathbf{y} \in S$$

Moments

$$\begin{aligned} E(\mathbf{Y}) &= n\mathbf{p} \\ \operatorname{var}(\mathbf{Y}) &= \mathbf{M} \end{aligned}$$

where $\mathbf{M}$ is the $k \times k$ matrix with elements

$$m_{ij} = \begin{cases} np_i(1 - p_i), & i = j \\ -np_i p_j & i \neq j \end{cases}$$

Specialization The special case $n = 1$ is called the multivariate Bernoulli distribution

$$\operatorname{Ber}_k(\mathbf{p}) = \operatorname{Bin}_k(1, \mathbf{p})$$

but for once we will not spell out the details with a special section for the multivariate Bernoulli. Just take $n = 1$ in this section.

Marginal Distributions Distributions obtained by collapsing categories are again multinomial (Section 5.4.5 in these notes).

In particular, if $\mathbf{Y} \sim \operatorname{Multi}_k(n, \mathbf{p})$, then

$$(Y_1, \ldots, Y_j, Y_{j+1} + \cdots + Y_k) \sim \operatorname{Multi}_{j+1}(n, \mathbf{q}) \tag{B.4}$$

where

$$\begin{aligned} q_i &= p_i, & i \leq j \\ q_{j+1} &= p_{j+1} + \cdots p_k \end{aligned}$$

Because the random vector in (B.4) is degenerate, this equation also gives implicitly the marginal distribution of Y_1, …, Y_j

$$\begin{aligned} &f(y_1, \ldots, y_j) \\ &\quad = \binom{n}{y_1, \ldots, y_j, n - y_1 - \cdots - y_j} p_1^{y_1} \cdots p_j^{y_j} (1 - p_1 - \cdots - p_j)^{n - y_1 - \cdots - y_j} \end{aligned}$$

Univariate Marginal Distributions If $\mathbf{Y} \sim \operatorname{Multi}(n, \mathbf{p})$, then

$$Y_i \sim \operatorname{Bin}(n, p_i).$$

Conditional Distributions If $\mathbf{Y} \sim \text{Multi}_k(n, \mathbf{p})$, then

$$(Y_1, \ldots, Y_j) \mid (Y_{j+1}, \ldots, Y_k) \sim \text{Multi}_j(n - Y_{j+1} - \cdots - Y_k, \mathbf{q}),$$

where

$$q_i = \frac{p_i}{p_1 + \cdots + p_j}, \qquad i = 1, \ldots, j.$$

B.5 Continuous Multivariate Distributions

B.5.1 The Uniform Distribution

The uniform distribution defined in Section B.2.1 actually made no mention of dimension. If the set S on which the distribution is defined lies in $\mathbb{R}^n$, then this is a multivariate distribution.

Conditional Distributions Every conditional distribution of a multivariate uniform distribution is uniform.

Marginal Distributions No regularity. Depends on the particular distribution. Marginals of the uniform distribution on a rectangle with sides parallel to the coordinate axes are uniform. Marginals of the uniform distribution on a disk or triangle are not uniform.

B.5.2 The Standard Normal Distribution

The distribution of a random vector $\mathbf{Z} = (Z_1, \ldots, Z_k)$ with the Z_i i. i. d. standard normal.

Moments

$$\begin{aligned} E(\mathbf{Z}) &= 0 \\ \operatorname{var}(\mathbf{Z}) &= \mathbf{I}, \end{aligned}$$

where $\mathbf{I}$ denotes the $k \times k$ identity matrix.

B.5.3 The Multivariate Normal Distribution

The distribution of a random vector $\mathbf{X} = \mathbf{a} + \mathbf{B}\mathbf{Z}$, where $\mathbf{Z}$ is multivariate standard normal.

Moments

$$\begin{aligned} E(\mathbf{X}) &= \boldsymbol{\mu} = \mathbf{a} \\ \operatorname{var}(\mathbf{X}) &= \mathbf{M} = \mathbf{B}\mathbf{B}' \end{aligned}$$

The Abbreviation $\mathcal{N}_k(\boldsymbol{\mu}, \mathbf{M})$ or $\mathcal{N}(\boldsymbol{\mu}, \mathbf{M})$ if the dimension k is clear from context.

The Sample Space If $\mathbf{M}$ is positive definite, the sample space is $\mathbb{R}^k$.

Otherwise, X is concentrated on the intersection of hyperplanes determined by null eigenvectors of $\mathbf{M}$

$$S = \{\, \mathbf{x} \in \mathbb{R}^k : \mathbf{z}'\mathbf{x} = \mathbf{z}'\boldsymbol{\mu} \text{ whenever } \mathbf{M}\mathbf{z} = 0 \,\}$$

The Parameters The mean vector $\boldsymbol{\mu}$ and variance matrix $\mathbf{M}$.

The Density Only exists if the distribution is nondegenerate ($\mathbf{M}$ is positive definite). Then

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} \det(\mathbf{M})^{1/2}} \exp\left(-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\mathbf{M}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right), \qquad \mathbf{x} \in \mathbb{R}^k$$

Marginal Distributions All are normal. If

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$$

is a partitioned random vector with (partitioned) mean vector

$$E(\mathbf{X}) = \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}$$

and (partitioned) variance matrix

$$\operatorname{var}(\mathbf{X}) = \mathbf{M} = \begin{pmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{pmatrix}$$

and $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{M})$, then

$$\mathbf{X}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{M}_{11}).$$

Conditional Distributions All are normal. If $\mathbf{X}$ is as in the preceding section and $\mathbf{X}_2$ is nondegenerate, then the conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_2$ is normal with

$$\begin{aligned} E(\mathbf{X}_1 \mid \mathbf{X}_2) &= \boldsymbol{\mu}_1 + \mathbf{M}_{12}\mathbf{M}_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2) \\ \operatorname{var}(\mathbf{X}_1 \mid \mathbf{X}_2) &= \mathbf{M}_{11} - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{21} \end{aligned}$$

If $\mathbf{X}_2$ is degenerate so $\mathbf{M}_{22}$ is not invertible, then the conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_2$ is still normal and the same formulas work if $\mathbf{M}_{22}^{-1}$ is replaced by a generalized inverse.

B.5.4 The Bivariate Normal Distribution

The special case $k = 2$ of the preceeding section.

The Density

$$
\begin{aligned}
f(x, y) = {} & \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \times \\
& \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right)
\end{aligned}
$$

Marginal Distributions

$$
Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)
$$

Conditional Distributions The conditional distribution of X given Y is normal with

$$
\begin{aligned}
E(X \mid Y) &= \mu_X + \rho\frac{\sigma_X}{\sigma_Y}(Y - \mu_Y) \\
\operatorname{var}(X \mid Y) &= \sigma_X^2(1-\rho^2)
\end{aligned}
$$

where $\rho = \operatorname{cor}(X, Y)$.

Appendix C

Addition Rules for Distributions

"Addition rules" for distributions are rules of the form: if X_1, ..., X_k are independent with some specified distributions, then $X_1 + \cdots + X_k$ has some other specified distribution.

Bernoulli If X_1, ..., X_k are i. i. d. $\text{Ber}(p)$, then

$$X_1 + \cdots + X_k \sim \text{Bin}(k, p). \tag{C.1}$$

- All the Bernoulli distributions must have the *same* success probability p.

Binomial If X_1, ..., X_k are independent with $X_i \sim \text{Bin}(n_i, p)$, then

$$X_1 + \cdots + X_k \sim \text{Bin}(n_1 + \cdots + n_k, p). \tag{C.2}$$

- All the binomial distributions must have the *same* success probability p.
- (C.1) is the special case of (C.2) obtained by setting $n_1 = \cdots = n_k = 1$.

Geometric If X_1, ..., X_k are i. i. d. $\text{Geo}(p)$, then

$$X_1 + \cdots + X_k \sim \text{NegBin}(k, p). \tag{C.3}$$

- All the geometric distributions must have the *same* success probability p.

Negative Binomial If X_1, ..., X_k are independent with $X_i \sim \text{NegBin}(n_i, p)$, then

$$X_1 + \cdots + X_k \sim \text{NegBin}(n_1 + \cdots + n_k, p). \tag{C.4}$$

- All the negative binomial distributions must have the *same* success probability p.
- (C.3) is the special case of (C.4) obtained by setting $n_1 = \cdots = n_k = 1$.

Poisson If X_1, …, X_k are independent with $X_i \sim \text{Poi}(\mu_i)$, then

$$X_1 + \cdots + X_k \sim \text{Poi}(\mu_1 + \cdots + \mu_k). \tag{C.5}$$

Exponential If X_1, …, X_k are i. i. d. $\text{Exp}(\lambda)$, then

$$X_1 + \cdots + X_k \sim \text{Gam}(n, \lambda). \tag{C.6}$$

- All the exponential distributions must have the *same* rate parameter λ.

Gamma If X_1, …, X_k are independent with $X_i \sim \text{Gam}(\alpha_i, \lambda)$, then

$$X_1 + \cdots + X_k \sim \text{Gam}(\alpha_1 + \cdots + \alpha_k, \lambda). \tag{C.7}$$

- All the gamma distributions must have the *same* rate parameter λ.
- (C.6) is the special case of (C.7) obtained by setting $\alpha_1 = \cdots = \alpha_k = 1$.

Chi-Square If X_1, …, X_k are independent with $X_i \sim \text{chi}^2(n_i)$, then

$$X_1 + \cdots + X_k \sim \text{chi}^2(n_1 + \cdots + n_k). \tag{C.8}$$

- (C.8) is the special case of (C.7) obtained by setting

$$\alpha_i = n_i/2 \quad \text{and} \quad \lambda_i = 1/2, \qquad i = 1, \ldots, k.$$

Normal If X_1, …, X_k are independent with $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, then

$$X_1 + \cdots + X_k \sim \mathcal{N}(\mu_1 + \cdots + \mu_k, \sigma_1^2 + \cdots + \sigma_k^2). \tag{C.9}$$

Linear Combination of Normals If X_1, …, X_k are independent with $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ and a_1, …, a_k are constants, then

$$\sum_{i=1}^k a_i X_i \sim \mathcal{N}\left(\sum_{i=1}^k a_i \mu_i, \sum_{i=1}^k a_i^2 \sigma_i^2\right). \tag{C.10}$$

- (C.9) is the special case of (C.10) obtained by setting $a_1 = \cdots = a_k = 1$.

Cauchy If X_1, …, X_k are independent with $X_i \sim \text{Cauchy}(\mu, \sigma)$, then

$$X_1 + \cdots + X_k \sim \text{Cauchy}(n\mu, n\sigma). \tag{C.11}$$

Appendix D

Relations Among Brand Name Distributions

D.1 Special Cases

First there are the special cases, which were also noted in Appendix B.

$$
\begin{aligned}
\operatorname{Ber}(p) &= \operatorname{Bin}(1, p) \\
\operatorname{Geo}(p) &= \operatorname{NegBin}(1, p) \\
\operatorname{Exp}(\lambda) &= \operatorname{Gam}(1, \lambda) \\
\operatorname{chi}^2(k) &= \operatorname{Gam}\left(\tfrac{k}{2}, \tfrac{1}{2}\right)
\end{aligned}
$$

The main point of this appendix are the relationships that involve more theoretical issues.

D.2 Relations Involving Bernoulli Sequences

Suppose X_1, X_2, $\ldots$ are i. i. d. $\operatorname{Ber}(p)$ random variables.

If n is a positive integer and

$$Y = X_1 + \cdots + X_n$$

is the number of "successes" in the n Bernoulli trials, then

$$Y \sim \operatorname{Bin}(n, p).$$

On the other hand, if y is positive integer and N is the trial at which the y-th success occurs, that is the random number N such that

$$
\begin{aligned}
X_1 + \cdots + X_N &= y \\
X_1 + \cdots + X_k &< y, \qquad k < N,
\end{aligned}
$$

then

$$N \sim \operatorname{NegBin}(y, p).$$

D.3 Relations Involving Poisson Processes

In a one-dimensional homogeneous Poisson process with rate parameter λ, the counts are Poisson and the waiting and interarrival times are exponential. Specifically, the number of points (arrivals) in an interval of length t has the $\text{Poi}(\lambda t)$ distribution, and the waiting times and interarrival times are independent and indentically $\text{Exp}(\lambda)$ distributed.

Even more specifically, let X_1, X_2, $\ldots$ be i. i. d. $\text{Exp}(\lambda)$ random variables. Take these to be the waiting and interarrival times of a Poisson process. This means the arrival times themselves are

$$T_k = \sum_{i=1}^{k} X_i$$

Note that

$$0 < T_1 < T_2 < \cdots$$

and

$$X_i = T_i - T_{i-1}, \qquad i > 1$$

so these are the interarrival times and $X_1 = T_1$ is the waiting time until the first arrival.

The characteristic property of the Poisson process, that counts have the Poisson distribution, says the number of points in the interval $(0, t)$, that is, the number of T_i such that $T_i < t$, has the $\text{Poi}(\lambda t)$ distribution.

D.4 Normal, Chi-Square, t, and F

D.4.1 Definition of Chi-Square

If Z_1, Z_2, $\ldots$ are i. i. d. $\mathcal{N}(0, 1)$, then

$$Z_1^2 + \ldots + Z_n^2 \sim \text{chi}^2(n).$$

D.4.2 Definition of t

If Z and Y are independent and

$$\begin{aligned} Z &\sim \mathcal{N}(0, 1) \\ Y &\sim \text{chi}^2(\nu) \end{aligned}$$

then

$$\frac{Z}{\sqrt{Y/\nu}} \sim t(\nu)$$

D.4.3 Definition of F

If X and Y are independent and

$$\begin{aligned} X &\sim \text{chi}^2(\mu) \\ Y &\sim \text{chi}^2(\nu) \end{aligned}$$

then

$$\frac{X/\mu}{Y/\nu} \sim F(\mu, \nu)$$

D.4.4 t as a Special Case of F

If

$$T \sim t(\nu),$$

then

$$T^2 \sim F(1, \nu).$$

Appendix E

Eigenvalues and Eigenvectors

E.1 Orthogonal and Orthonormal Vectors

If $\mathbf{x}$ and $\mathbf{y}$ are vectors of the same dimension, we say they are *orthogonal* if $\mathbf{x}'\mathbf{y} = 0$. Since the transpose of a matrix product is the product of the transposes in reverse order, an equivalent condition is $\mathbf{y}'\mathbf{x} = 0$. Orthogonality is the n-dimensional generalization of perpendicularity. In a sense, it says that two vectors make a right angle.

The *length* or *norm* of a vector $\mathbf{x} = (x_1, \ldots, x_n)$ is defined to be

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}'\mathbf{x}} = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

Squaring both sides gives

$$\|\mathbf{x}\|^2 = \sum_{i=1}^{n} x_i^2,$$

which is one version of the Pythagorean theorem, as it appears in analytic geometry.

Orthogonal vectors give another generalization of the Pythagorean theorem. We say a set of vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ is *orthogonal* if

$$\mathbf{x}_i'\mathbf{x}_j = 0, \qquad i \neq j. \tag{E.1}$$

Then

$$\begin{aligned}\|\mathbf{x}_1 + \cdots + \mathbf{x}_k\|^2 &= (\mathbf{x}_1 + \cdots + \mathbf{x}_k)'(\mathbf{x}_1 + \cdots + \mathbf{x}_k) \\ &= \sum_{i=1}^k \sum_{j=1}^k \mathbf{x}_i'\mathbf{x}_j \\ &= \sum_{i=1}^k \mathbf{x}_i'\mathbf{x}_i \\ &= \sum_{i=1}^k \|\mathbf{x}_i\|^2\end{aligned}$$

because, by definition of orthogonality, all terms in the second line with $i \neq j$ are zero.

We say an orthogonal set of vectors is *orthonormal* if

$$\mathbf{x}_i'\mathbf{x}_i = 1. \tag{E.2}$$

That is, a set of vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ is orthonormal if it satisfies both (E.1) and (E.2).

An orthonormal set is automatically linearly independent because if

$$\sum_{i=1}^k c_i \mathbf{x}_i = 0,$$

then

$$0 = \mathbf{x}_j' \left(\sum_{i=1}^k c_i \mathbf{x}_i \right) = c_j \mathbf{x}_j'\mathbf{x}_j = c_j$$

holds for all j. Hence the only linear combination that is zero is the one with all coefficients zero, which is the definition of linear independence.

Being linearly independent, an orthonormal set is always a *basis* for whatever subspace it spans. If we are working in n-dimensional space, and there are n vectors in the orthonormal set, then they make up a basis for the whole space. If there are $k < n$ vectors in the set, then they make up a basis for some proper subspace.

It is always possible to choose an orthogonal basis for any vector space or subspace. One way to do this is the Gram-Schmidt orthogonalization procedure, which converts an arbitrary basis $\mathbf{y}_1$, $\ldots$, $\mathbf{y}_n$ to an orthonormal basis $\mathbf{x}_1$, $\ldots$, $\mathbf{x}_n$ as follows. First let

$$\mathbf{x}_1 = \frac{\mathbf{y}_1}{\|\mathbf{y}_1\|}.$$

Then define the $\mathbf{x}_i$ in order. After $\mathbf{x}_1$, $\ldots$, $\mathbf{x}_{k-1}$ have been defined, let

$$\mathbf{z}_k = \mathbf{y}_k - \sum_{i=1}^{k-1} \mathbf{x}_i \mathbf{x}_i' \mathbf{y}$$

and

$$\mathbf{x}_k = \frac{\mathbf{z}_k}{\|\mathbf{z}_k\|}.$$

It is easily verified that this does produce an orthonormal set, and it is only slightly harder to prove that none of the $\mathbf{x}_i$ are zero because that would imply linear dependence of the $\mathbf{y}_i$.

E.2 Eigenvalues and Eigenvectors

If $\mathbf{A}$ is any matrix, we say that λ is a *right eigenvalue* corresponding to a *right eigenvector* $\mathbf{x}$ if

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$$

Left eigenvalues and eigenvectors are defined analogously with "left multiplication" $\mathbf{x}'\mathbf{A} = \lambda\mathbf{x}'$, which is equivalent to $\mathbf{A}'\mathbf{x} = \lambda\mathbf{x}$. So the right eigenvalues and eigenvectors of $\mathbf{A}'$ are the left eigenvalues and eigenvectors of $\mathbf{A}$. When $\mathbf{A}$ is symmetric ($\mathbf{A}' = \mathbf{A}$), the "left" and "right" concepts are the same and the adjectives "left" and "right" are unnecessary. Fortunately, this is the most interesting case, and the only one in which we will be interested. From now on we discuss only eigenvalues and eigenvectors of *symmetric* matrices.

There are three important facts about eigenvalues and eigenvectors. Two elementary and one very deep. Here's the first (one of the elementary facts).

Lemma E.1. *Eigenvectors corresponding to distinct eigenvalues are orthogonal.*

This means that if

$$\mathbf{A}\mathbf{x}_i = \lambda_i\mathbf{x}_i \tag{E.3}$$

then

$$\lambda_i \neq \lambda_j \qquad \text{implies} \qquad \mathbf{x}_i'\mathbf{x}_j = 0.$$

Proof. Suppose $\lambda_i \neq \lambda_j$, then at least one of the two is not zero, say λ_j. Then

$$\mathbf{x}_i'\mathbf{x}_j = \frac{\mathbf{x}_i'\mathbf{A}\mathbf{x}_j}{\lambda_j} = \frac{(\mathbf{A}\mathbf{x}_i)'\mathbf{x}_j}{\lambda_j} = \frac{\lambda_i\mathbf{x}_i'\mathbf{x}_j}{\lambda_j} = \frac{\lambda_i}{\lambda_j}\cdot\mathbf{x}_i'\mathbf{x}_j$$

and since $\lambda_i \neq \lambda_j$ the only way this can happen is if $\mathbf{x}_i'\mathbf{x}_j = 0$. □

Here's the second important fact (also elementary).

Lemma E.2. *Every linear combination of eigenvectors corresponding to the same eigenvalue is another eigenvector corresponding to that eigenvalue.*

This means that if

$$\mathbf{A}\mathbf{x}_i = \lambda\mathbf{x}_i$$

then

$$\mathbf{A}\left(\sum_{i=1}^{k} c_i\mathbf{x}_i\right) = \lambda\left(\sum_{i=1}^{k} c_i\mathbf{x}_i\right)$$

Proof. This is just linearity of matrix multiplication. □

The second property means that all the eigenvectors corresponding to one eigenvalue constitute a subspace. If the dimension of that subspace is k, then it is possible to choose an orthonormal basis of k vectors that span the subspace. Since the first property of eigenvalues and eigenvectors says that (E.1) is also satisfied by eigenvectors corresponding to different eigenvalues, all of the eigenvectors chosen this way form an orthonormal set.

Thus our orthonormal set of eigenvectors spans a subspace of dimension m which contains all eigenvectors of the matrix in question. The question then arises whether this set is *complete*, that is, whether it is a basis for the whole space, or in symbols whether $m = n$, where n is the dimension of the whole space ($\mathbf{A}$ is an $n \times n$ matrix and the $\mathbf{x}_i$ are vectors of dimension n). It turns out that the set *is* always complete, and this is the third important fact about eigenvalues and eigenvectors.

Lemma E.3. *Every real symmetric matrix has an orthonormal set of eigenvectors that form a basis for the space.*

In contrast to the first two facts, this is deep, and we shall not say anything about its proof, other than that about half of the typical linear algebra book is given over to building up to the proof of this one fact.

The "third important fact" says that *any* vector can be written as a linear combination of eigenvectors

$$\mathbf{y} = \sum_{i=1}^{n} c_i \mathbf{x}_i$$

and this allows a very simple description of the action of the linear operator described by the matrix

$$\mathbf{A}\mathbf{y} = \sum_{i=1}^{n} c_i \mathbf{A}\mathbf{x}_i = \sum_{i=1}^{n} c_i \lambda_i \mathbf{x}_i \tag{E.4}$$

So this says that *when we use an orthonormal eigenvector basis*, if $\mathbf{y}$ has the representation $(c_1, \ldots, c_n)$, then $\mathbf{A}y$ has the representation $(c_1\lambda_1, \ldots, c_n\lambda_n)$. Let $\mathbf{D}$ be the representation in the orthonormal eigenvector basis of the linear operator represented by $\mathbf{A}$ in the standard basis. Then our analysis above says the i-the element of $\mathbf{Dc}$ is $c_i\lambda_i$, that is,

$$\sum_{j=1}^{n} d_{ij} c_j = \lambda_i c_i.$$

In order for this to hold for all real numbers c_i, it must be that $\mathbf{D}$ is diagonal

$$\begin{aligned} d_{ii} &= \lambda_i \\ d_{ij} &= 0, \qquad i \neq j \end{aligned}$$

In short, using the orthonormal eigenvector basis *diagonalizes* the linear operator represented by the matrix in question.

There is another way to describe this same fact without mentioning bases. Many people find it a simpler description, though its relation to eigenvalues and eigenvectors is hidden in the notation, no longer immediately apparent. Let $\mathbf{O}$ denote the matrix whose columns are the orthonormal eigenvector basis ($\mathbf{x}_1$, $\ldots$, $\mathbf{x}_n$), that is, if o_{ij} are the elements of $\mathbf{O}$, then

$$\mathbf{x}_i = (o_{1i}, \ldots, o_{ni}).$$

Now (E.1) and (E.2) can be combined as one matrix equation

$$\mathbf{O}'\mathbf{O} = \mathbf{I} \tag{E.5}$$

(where, as usual, $\mathbf{I}$ is the $n \times n$ identity matrix). A matrix $\mathbf{O}$ satisfying this property is said to be *orthogonal.* Another way to read (E.5) is that it says $\mathbf{O}' = \mathbf{O}^{-1}$ (an orthogonal matrix is one whose inverse is its transpose). The fact that inverses are two-sided ($\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ for any invertible matrix $\mathbf{A}$) implies that $\mathbf{O}\mathbf{O}' = \mathbf{I}$ as well.

Furthermore, the eigenvalue-eigenvector equation (E.3) can be written out with explicit subscripts and summations as

$$\sum_{j=1}^{n} a_{ij} o_{jk} = \lambda_k o_{ik} = o_{ik} d_{kk} = \sum_{j=1}^{n} o_{ij} d_{jk}$$

(where $\mathbf{D}$ is the the diagonal matrix with eigenvalues on the diagonal defined above). Going back to matrix notation gives

$$\mathbf{A}\mathbf{O} = \mathbf{O}\mathbf{D} \tag{E.6}$$

The two equations (E.3) and (E.6) may not look much alike, but as we have just seen, they say exactly the same thing in different notation. Using the orthogonality property ($\mathbf{O}' = \mathbf{O}^{-1}$) we can rewrite (E.6) in two different ways.

Theorem E.4 (Spectral Decomposition). *Any real symmetric matrix* $\mathbf{A}$ *can be written*

$$\mathbf{A} = \mathbf{O}\mathbf{D}\mathbf{O}' \tag{E.7}$$

where $\mathbf{D}$ *is diagonal and* $\mathbf{O}$ *is orthogonal.*

Conversely, for any real symmetric matrix $\mathbf{A}$ *there exists an orthogonal matrix* $\mathbf{O}$ *such that*

$$\mathbf{D} = \mathbf{O}'\mathbf{A}\mathbf{O}$$

is diagonal.

(The reason for the name of the theorem is that the set of eigenvalues is sometimes called the *spectrum* of $\mathbf{A}$). The spectral decomposition theorem says nothing about eigenvalues and eigenvectors, but we know from the discussion above that the diagonal elements of $\mathbf{D}$ are the eigenvalues of $\mathbf{A}$, and the columns of $\mathbf{O}$ are the corresponding eigenvectors.

E.3 Positive Definite Matrices

Using the spectral theorem, we can prove several interesting things about positive definite matrices.

Corollary E.5. *A real symmetric matrix* $\mathbf{A}$ *is positive semi-definite if and only if its spectrum is nonnegative. A real symmetric matrix* $\mathbf{A}$ *is positive definite if and only if its spectrum is strictly positive.*

Proof. First suppose that $\mathbf{A}$ is positive semi-definite with spectral decomposition (E.7). Let $\mathbf{e}_i$ denote the vector having elements that are all zero except the i-th, which is one, and define $\mathbf{w} = \mathbf{O}\mathbf{e}_i$, so

$$0 \le \mathbf{w}'\mathbf{A}\mathbf{w} = \mathbf{e}_i'\mathbf{O}'\mathbf{O}\mathbf{D}\mathbf{O}'\mathbf{O}\mathbf{e}_i = \mathbf{e}_i'\mathbf{D}\mathbf{e}_i = d_{ii} \tag{E.8}$$

using $\mathbf{O}'\mathbf{O} = I$. Hence the spectrum is nonnegative.

Conversely, suppose the d_{ii} are nonnegative. Then for any vector $\mathbf{w}$ define $\mathbf{z} = \mathbf{O}'\mathbf{w}$, so

$$\mathbf{w}'\mathbf{A}\mathbf{w} = \mathbf{w}'\mathbf{O}\mathbf{D}\mathbf{O}'\mathbf{w} = \mathbf{z}'\mathbf{D}\mathbf{z} = \sum_i d_{ii} z_i^2 \ge 0$$

Hence $\mathbf{A}$ is positive semi-definite.

The assertions about positive definiteness are proved in almost the same way. Suppose that $\mathbf{A}$ is positive definite. Since $\mathbf{e}_i$ is nonzero, $\mathbf{w}$ in (E.8) is also nonzero because $\mathbf{e}_i = \mathbf{O}'\mathbf{w}$ would be zero (and it isn't) if $\mathbf{w}$ were zero. Thus the inequality in (E.8) is actually strict. Hence the spectrum of is strictly positive.

Conversely, suppose the d_{ii} are strictly positive. Then for any nonzero vector $\mathbf{w}$ define $\mathbf{z} = \mathbf{O}'\mathbf{w}$ as before, and again note that $\mathbf{z}$ is nonzero because $\mathbf{w} = \mathbf{O}\mathbf{z}$ and $\mathbf{w}$ is nonzero. Thus $\mathbf{w}'\mathbf{A}\mathbf{w} = \mathbf{z}'\mathbf{D}\mathbf{z} > 0$, and hence $\mathbf{A}$ is positive definite. □

Corollary E.6. *A positive semi-definite matrix is invertible if and only if it is positive definite.*

Proof. It is easily verified that the product of diagonal matrices is diagonal and the diagonal elements of the product are the products of the diagonal elements of the multiplicands. Thus a diagonal matrix $\mathbf{D}$ is invertible if and only if all its diagonal elements d_{ii} are nonzero, in which case $\mathbf{D}^{-1}$ is diagonal with diagonal elements $1/d_{ii}$.

Since $\mathbf{O}$ and $\mathbf{O}'$ in the spectral decomposition (E.7) are invertible, $\mathbf{A}$ is invertible if and only if $\mathbf{D}$ is, hence if and only if its spectrum is nonzero, in which case

$$\mathbf{A}^{-1} = \mathbf{O}\mathbf{D}^{-1}\mathbf{O}'.$$

By the preceding corollary the spectrum of a positive semi-definite matrix is nonnegative, hence nonzero if and only if strictly positive, which (again by the preceding corollary) occurs if and only if the matrix is positive definite. □

Corollary E.7. *Every real symmetric positive semi-definite matrix* $\mathbf{A}$ *has a symmetric square root*

$$\mathbf{A}^{1/2} = \mathbf{O}\mathbf{D}^{1/2}\mathbf{O}' \tag{E.9}$$

where (E.7) *is the spectral decomposition of* $\mathbf{A}$ *and where* $\mathbf{D}^{1/2}$ *is defined to be the diagonal matrix whose diagonal elements are* $\sqrt{d_{ii}}$*, where* d_{ii} *are the diagonal elements of* $\mathbf{D}$.

Moreover, $\mathbf{A}^{1/2}$ *is positive definite if and only if* $\mathbf{A}$ *is positive definite.*

Note that by Corollary E.5 all of the diagonal elements of $\mathbf{D}$ are nonnegative and hence have real square roots.

Proof.

$$\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{O}\mathbf{D}^{1/2}\mathbf{O}'\mathbf{O}\mathbf{D}^{1/2}\mathbf{O}' = \mathbf{O}\mathbf{D}^{1/2}\mathbf{D}^{1/2}\mathbf{O}' = \mathbf{O}\mathbf{D}\mathbf{O}' = \mathbf{A}$$

because $\mathbf{O}'\mathbf{O} = \mathbf{I}$ and $\mathbf{D}^{1/2}\mathbf{D}^{1/2} = \mathbf{D}$.

From Corollary E.5 we know that $\mathbf{A}$ is positive definite if and only if all the d_{ii} are strictly positive. Since (E.9) is the spectral decomposition of $\mathbf{A}^{1/2}$, we see that $\mathbf{A}^{1/2}$ is positive definite if and only if all the $\sqrt{d_{ii}}$ are strictly positive. Clearly $d_{ii} > 0$ if and only if $\sqrt{d_{ii}} > 0$. □

Appendix F

Normal Approximations for Distributions

F.1 Binomial Distribution

The $\mathrm{Bin}(n, p)$ distribution is approximately normal with mean np and variance $np(1-p)$ if n is large.

F.2 Negative Binomial Distribution

The $\mathrm{NegBin}(n, p)$ distribution is approximately normal with mean n/p and variance $n(1-p)/p^2$ if n is large.

F.3 Poisson Distribution

The $\mathrm{Poi}(\mu)$ distribution is approximately normal with mean μ and variance μ if μ is large.

F.4 Gamma Distribution

The $\mathrm{Gam}(\alpha, \lambda)$ distribution is approximately normal with mean α/λ and variance α/λ^2 if α is large.

F.5 Chi-Square Distribution

The $\mathrm{chi}^2(n)$ distribution is approximately normal with mean n and variance $2n$ if n is large.

Appendix G

Maximization of Functions

This appendix contains no statistics. It just reviews some facts from calculus about maximization of functions.

First we distinguish between local and global maxima.[1] A point x is a *global maximum* of a function f if

$$f(x) \geq f(y), \qquad \text{for all } y \text{ in the domain of } f.$$

In words, $f(x)$ is greater than or equal to $f(y)$ for all other y.

Unfortunately, calculus isn't much help in finding global maxima, hence the following definition, which defines something calculus is much more helpful in finding. A point x is a *local maximum* of the function f if

$$f(x) \geq f(y), \qquad \text{for all } y \text{ in some neighborhood of } x.$$

The point is that saying x is a local maximum doesn't say anything at all about whether a global maximum exists or whether x is also a global maximum.

> *Every global maximum is a local maximum, but not all local maxima are global maxima.*

G.1 Functions of One Variable

The connection between calculus and local maxima is quite simple.

Theorem G.1. *Suppose f is a real-valued function of one real variable and is twice differentiable at the point x. A sufficient condition that x be a local maximum of f is*

$$f'(x) = 0 \quad \textit{and} \quad f''(x) < 0. \tag{G.1}$$

[1] An irregular plural following the Latin rather than the English pattern. Singular: *maximum.* Plural: *maxima.*

In words, to find a local maximum, find a point x where the derivative is zero. Then check the second derivative. If $f''(x)$ is negative, then x is a local maximum. If $f''(x)$ is positive, then x is definitely not a local maximum (in fact it's a local minimum). If $f''(x)$ is zero, you are not sure. Consider $f(x) = x^3$ and $g(x) = -x^4$. Both have second derivative zero at $x = 0$, but f is strictly increasing (draw a graph) and hence does not have a maximum (local or global), whereas g does have a local maximum at $x = 0$.

That takes care of local maxima that occur at interior points of the domain of the function being maximized. What about local maxima that occur at boundary points? Here the situation becomes more complicated.

Our first problem is that ordinary derivatives don't exist, but there still may be one-sided derivatives. In the following discussion all the derivatives are one-sided.

Theorem G.2. *Suppose f is a twice differentiable real-valued function defined on a closed interval of the real line. A sufficient condition that a lower boundary point x of the interval be a local maximum of f is*

$$f'(x) < 0. \tag{G.2a}$$

Another sufficient condition is

$$f'(x) = 0 \quad \textit{and} \quad f''(x) < 0. \tag{G.2b}$$

If x is an upper boundary point, the conditions are the same except the inequality in (G.2a) *is reversed: $f'(x) > 0$.*

This theorem is so complicated that you're excused if you want to ignore it and just draw a graph of the function. The main point of the theorem is that a local maximum can occur at a boundary point when the derivative is not zero. Consider the function $f(x) = -x$ defined on the interval $0 \le x < +\infty$. Since f is strictly decreasing, the only global (and local) maximum occurs at $x = 0$ where $f'(x) = -1$. This satisfies the condition (G.2a) of the theorem, but notice the derivative is not zero.

Thus it is *not* enough to just look for points where the first derivative is zero. You also have to check the boundary points, where the more complicated test applies.

Are we done with maximization theory? Not if we are interested in global maxima. Even if you find all the local maxima, it is not necessarily true that a global maximum exists. Consider the function $f(x) = x^3 - x$ graphed in Figure G.1. The first derivative is

$$f'(x) = 3x^2 - 1,$$

which has zeros at $\pm 1/\sqrt{3}$. From the graph it is clear that $-1/\sqrt{3}$ is a local maximum and $+1/\sqrt{3}$ is a local *minimum*. But there is no global maximum since $f(x) \to +\infty$ as $x \to +\infty$.

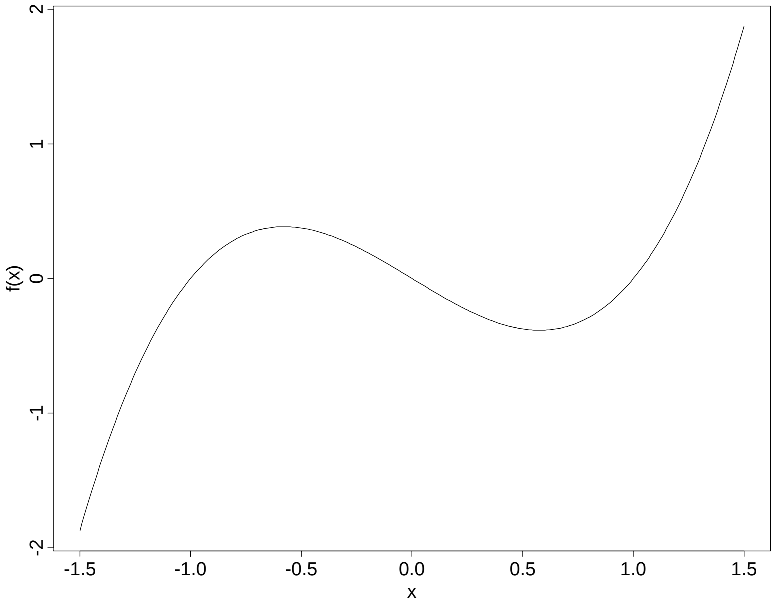


Figure G.1: Graph of $f(x) = x^3 - x$.

If a global maximum exists, then it is also a local maximum. So if you find all local maxima, they must include any global maxima. But the example shows that local maxima can exist when there is no global maximum. Thus calculus can help you find local maxima, but it is no help in telling you which of them are global maxima or whether any of them are global maxima.

G.2 Concave Functions of One Variable

There is one situation in which maximization is much simpler.

Definition G.2.1 (Strictly Concave Functions).
A continuous real-valued function f defined on an interval of the real line and twice differentiable at interior points of the interval is strictly concave *if the inequality $f''(x) < 0$ holds at every interior point x of the interval.*

There is a more general definition of "concave function" that does not require differentiability, but we will not use it.[2]

The concavity property $f''(x) < 0$ is easily recognized from a graph. It says the function curves downward at every point. Figure G.2 is an example.

Strictly concave functions are very special. For them, there is no difference between local and global maxima.

[2] Functions that are twice differentiable are concave but not strictly concave if $f''(x) \leq 0$ at all interior points and $f''(x) = 0$ at some points. But we won't have any use for this concept.

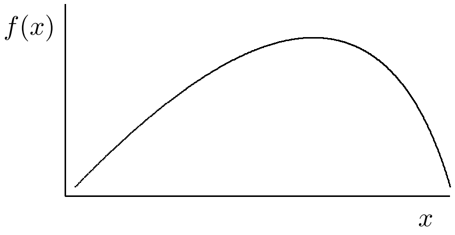


Figure G.2: A Concave Function.

Theorem G.3. *A continuous real-valued strictly concave function f defined on an interval of the real line and twice differentiable at interior points of the interval has at most one local maximum. If it has a local maximum, then that point is also the global maximum.*

Also, f' has at most one zero. If it has a zero, this is the global maximum.

The theorem says the the situation for strictly concave functions is very simple. If you find a local maximum or a zero of the first derivative, then you have found the unique global maximum.

To summarize, there is a very important distinction between general functions and strictly concave ones. The derivative tests are almost the same, but there is a subtle difference. If

$$f'(x) = 0 \text{ and } f''(x) < 0$$

then you know x is a local maximum of f, but you don't know that x is a global maximum or whether there is any global maximum. But if

$$f'(x) = 0 \text{ and } f''(y) < 0 \text{ for all } y$$

then you know that f is strictly concave and hence that x is the unique global maximum. The only difference is whether you just check the sign of the second derivative only at the point x or at all points y in the domain of f.

G.3 Functions of Several Variables

In Chapter 5 (p. 22 of the notes) we learned about the first derivative of a vector-valued function of a vector variable. This derivative was used in the multivariable delta method.

To develop the multivariable analog of the theory of the preceeding sections, we need to develop first and *second* derivatives of a scalar-valued function of a vector variable. (Fortunately, we don't need second derivatives of *vector*-valued functions of a vector variable. They're a mess.)

According to the theory developed in Chapter 5, if $\mathbf{g}$ is a function that maps vectors of dimension n to vectors of dimension m, then its derivative at the point $\mathbf{x}$ is the $m \times n$ matrix $\nabla \mathbf{g}(\mathbf{x})$ having elements

$$g_{ij} = \frac{\partial g_i(\mathbf{x})}{\partial x_j}$$

Here we are interested in the case $m = 1$ (a *scalar*-valued function) so the derivative is a $1 \times n$ matrix (a row vector).

Thus if f is a real-valued ("scalar-valued" means the same thing) function of a vector variable of dimension n, its first derivative is the row vector $\nabla f(\mathbf{x})$ having elements

$$\frac{\partial f(\mathbf{x})}{\partial x_i}, \qquad i = 1, \ldots, n.$$

It is pronounced "del f".

So what might the second derivative be? It is clear from the pattern that it should involve partial derivatives, in this case second derivatives. There are a lot of them. If f is a real-valued function of a vector variable of dimension n, its second derivative is the $n \times n$ matrix $\nabla^2 f(\mathbf{x})$ having elements

$$\frac{\partial f(\mathbf{x})}{\partial x_i \partial x_j}, \qquad i = 1, \ldots, n \text{ and } j = 1, \ldots, n.$$

It is pronounced "del squared f". Note that by the properties of partial derivatives

$$\frac{\partial f(\mathbf{x})}{\partial x_i \partial x_j} = \frac{\partial f(\mathbf{x})}{\partial x_j \partial x_i}$$

the second derivative matrix is a *symmetric* matrix.

Before we can state the multivariate analogue of Theorem G.2, we need to develop one more concept. Recall from Section 5.1.8 of last semester's notes (or look at Section 12.5 in Lindgren) that a symmetric square matrix $\mathbf{A}$ is *positive semidefinite* (Lindgren says *nonnegative definite*) if

$$\mathbf{c}'\mathbf{A}\mathbf{c} \geq 0, \qquad \text{for all vectors } \mathbf{c}, \tag{G.3a}$$

and $\mathbf{A}$ is *positive definite* if the inequality is strict

$$\mathbf{c}'\mathbf{A}\mathbf{c} > 0, \qquad \text{for all nonzero vectors } \mathbf{c}. \tag{G.3b}$$

As a shorthand we write $\mathbf{A} \geq 0$ to indicate (G.3a) and $A > 0$ to indicate (G.3b). No confusion should arise, because these can't have any other meaning (matrices aren't naturally ordered). We also write $\mathbf{A} \leq 0$ and $\mathbf{A} < 0$ to mean that $-\mathbf{A}$ is positive semidefinite or positive definite, respectively. When $\mathbf{A} \leq 0$ we say that it is *negative semidefinite*, and when $\mathbf{A} < 0$ we say that it is *negative definite*.

The place where positive (semi)definiteness arose last semester was that fact that every variance matrix is positive semidefinite (Corollary 5.5 in these notes, Theorem 9 of Chapter 12 in Lindgren) and actually positive definite if the random variable in question is not concentrated on a hyperplane (Corollary 5.6 in last semester's notes).

With these concepts we can now state the multivariate analogue of Theorem G.2.

Theorem G.4. *Suppose f is a real-valued function of a vector variable and is twice differentiable at the point $\mathbf{x}$. A sufficient condition that $\mathbf{x}$ be a local maximum of f is*

$$\nabla f(\mathbf{x}) = 0 \quad \textit{and} \quad \nabla^2 f(\mathbf{x}) < 0, \tag{G.4}$$

Recall from the discussion just before the theorem that the last part of the condition means $\nabla^2 f(\mathbf{x})$ is a *negative definite* matrix.

Unfortunately, the condition that a matrix is negative definite is impossible to check by hand except in a few special cases. However, it is fairly easy to check by computer. Compute the eigenvalues of the matrix (either R or Mathematica can do this) if all the eigenvalues are positive (resp. nonnegative), then the matrix is positive definite (resp. positive semidefinite), and if all the eigenvalues are negative (resp. nonpositive), then the matrix is negative definite (resp. negative semidefinite).

In fact, the first condition of the theorem isn't very easy to handle either except in very special cases. It's hard to find an $\mathbf{x}$ such that $\nabla f(\mathbf{x})$ holds. Recall this is a *vector* equation so what we are really talking about is solving n equations in n unknowns. Since these are generally *nonlinear* equations, there is no general method of finding a solution. In fact, it is much easier if you don't use first derivative information alone. The way to find the maximum of a function is to have the computer go uphill until it can't make any more progress.

Fortunately R has a function that minimizes functions of several variables (and maximizing f is equivalent to minimizing $-f$). So that can be used to solve all such problems.

Example G.3.1.
Consider minimizing the function[3]

$$f(x_1, x_2) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$

Here's how we minimize f using R

```
> f <- function(x) 100 * (x[2] - x[1]^2)^2 + (1 - x[1])^2
> out <- nlm(f, c(0,0), hessian=TRUE)
```

The first line defines an R function `f` of one variable `x` which is in this case a vector with two components `x[1]` and `x[2]`. The `nlm` function minimizes the function `f` using the second argument `c(0,0)` as the starting point for its iterative procedure. The starting point also specifies the dimension of the problem. The `x` that `nlm` will pass to `f` will have the same length as this starting point (in this case length 2). The argument `hessian=TRUE` tells `nlm` that we

[3] This function has no relevance to either probability or statistics. It's just a commonly used test case books about optimization. It's called Rosenbrock's function.

want the second derivative matrix too (Hessian matrix is another name for second derivative matrix)[4]

The result `out` returned by the minimization procedure is a list of components, only the first four of which are interesting (we omit the rest).

```
> out
$minimum
[1] 4.023726e-12

$estimate
[1] 0.999998 0.999996

$gradient
[1] -7.328278e-07  3.605688e-07

$hessian
          [,1]      [,2]
[1,]  802.2368 -400.0192
[2,] -400.0192  200.0000
```

The component `estimate` is the point $\mathbf{x}$ at which the function is minimized (or at least at which `nlm` claims it is minimized), and the component `minimum` is the value $f(\mathbf{x})$ of the function at that point. The component `gradient` is the first derivative ("gradient" is another name for a derivative vector). Notice that it is as close to zero as computer arithmetic allows. And the component `hessian` is, as we said above, the second derivative matrix. To check whether the second derivative matrix is positive definite, calculate eigenvalues

```
> eigen(out$hessian)
$values
[1] 1001.8055799    0.4312236

$vectors
           [,1]       [,2]
[1,] -0.8948213 -0.4464245
[2,]  0.4464245 -0.8948213
```

Since both eigenvalues (the elements of the `values`) component of the result list returned by the `eigen` function are positive this is a positive definite matrix, from which we conclude that the point found is a local minimum.

Positive definite? Doesn't the theorem say the second derivative should be *negative definite*? It does. This is the difference between maximization

[4] Unlike Mathematica, R doesn't know any calculus, so it calculates derivatives by finite differences

$$\frac{df(x)}{dx} \approx \frac{f(x+h) - f(x)}{h}$$

for small h, and similarly for partial derivatives. For second derivatives, apply this idea twice.

and minimization. Since maximizing f is equivalent to minimizing $-f$ and $\nabla^2 f(\mathbf{x}) = -\nabla^2(-f(\mathbf{x}))$, the condition is

negative definite Hessian at a local *maximum*,
positive definite Hessian at a local *minimum*.

Unfortunately, in statistics we are often interested in maximization, but most optimization theory and optimization software uses minimization, so we're always having to convert between the two.

G.4 Concave Functions of Several Variables

Definition G.4.1 (Convex Sets).
A subset of $\mathbb{R}^n$ is convex *if for any two points in the set the line segment between them lies entirely in the set.*

Definition G.4.2 (Strictly Concave Functions).
A continuous real-valued function f defined on a convex subset of $\mathbb{R}^n$ with a nonempty interior and twice differentiable at interior points of the subset is strictly concave *if the inequality $\nabla^2 f(\mathbf{x}) < 0$ holds at every interior point $\mathbf{x}$.*

As in the single-variable case, strictly concave functions are very special.

Theorem G.5. *A continuous real-valued strictly concave function f defined on a convex subset of $\mathbb{R}^n$ with a nonempty interior and twice differentiable at interior points of the interval has at most one local maximum. If it has a local maximum, then that point is also the global maximum.*

Also, ∇f has at most one zero. If it has a zero, this is the global maximum.

The theorem says the the situation for strictly concave functions is very simple. If you find a local maximum or a zero of the first derivative, then you have found the unique global maximum.

To summarize, there is a very important distinction between general functions and strictly concave ones. The derivative tests are almost the same, but there is a subtle difference. If

$$\nabla f(\mathbf{x}) = 0 \text{ and } \nabla^2 f(\mathbf{x}) < 0$$

then you know x is a local maximum of f, but you don't know that x is a global maximum or whether there is any global maximum. But if

$$\nabla f(\mathbf{x}) = 0 \text{ and } \nabla^2 f(\mathbf{y}) < 0 \text{ for all } \mathbf{y}$$

then you know that f is strictly concave and hence that $\mathbf{x}$ is the unique global maximum. The only difference is whether you just check negative definiteness the second derivative only at the point $\mathbf{x}$ or at all points $\mathbf{y}$ in the domain of f.

Appendix H

Projections and Chi-Squares

H.1 Orthogonal Projections

A matrix $\mathbf{A}$ is said to be an *orthogonal projection* if it is *symmetric* ($\mathbf{A}' = \mathbf{A}$) and *idempotent* ($\mathbf{A}^2 = \mathbf{A}$). The linear transformation represented by the matrix maps onto the subspace range($\mathbf{A}$). We say that $\mathbf{A}$ is the orthogonal projection onto range($\mathbf{A}$). The *rank* of $\mathbf{A}$, denoted rank($\mathbf{A}$) is the dimension of its range.

A typical element of range($\mathbf{A}$) has the form $\mathbf{y} = \mathbf{A}\mathbf{z}$ for an arbitrary vector $\mathbf{z}$. The idempotence property implies

$$\mathbf{A}\mathbf{y} = \mathbf{y}, \qquad \mathbf{y} \in \operatorname{range}(\mathbf{A}),$$

that is, the linear transformation represented by $\mathbf{A}$ behaves like the identity mapping on range($\mathbf{A}$). Any idempotent matrix (symmetric or not) has this property, and all such matrices are called projections.

The reason why the symmetric projections $\mathbf{A}$ are called *orthogonal* projections is because the vector from $\mathbf{y}$ to its projection $\mathbf{A}\mathbf{y}$ is orthogonal to the subspace range($\mathbf{A}$), which means

$$(\mathbf{y} - \mathbf{A}\mathbf{y})'(\mathbf{A}\mathbf{z}) = \mathbf{y}'(\mathbf{I} - \mathbf{A})'\mathbf{A}\mathbf{z} = \mathbf{y}'(\mathbf{I} - \mathbf{A})\mathbf{A}\mathbf{z} = 0, \qquad \text{for all vectors } \mathbf{y} \text{ and } \mathbf{z},$$

which is equivalent to

$$(\mathbf{I} - \mathbf{A})\mathbf{A} = 0. \tag{H.1}$$

But this is just the same thing as the idempotence property.

Since an orthogonal projection is symmetric, it has a spectral decomposition (E.7). Combining the spectral decomposition with the idempotence property gives

$$\mathbf{A} = \mathbf{O}\mathbf{D}\mathbf{O}' = \mathbf{A}^2 = \mathbf{O}\mathbf{D}\mathbf{O}'\mathbf{O}\mathbf{D}\mathbf{O}' = \mathbf{O}\mathbf{D}^2\mathbf{O}'$$

(because $\mathbf{O}'\mathbf{O} = \mathbf{I}$). Multiplying this by $\mathbf{O}$ on the right and $\mathbf{O}'$ on the left gives $\mathbf{D} = \mathbf{D}^2$ (again using $\mathbf{O}'\mathbf{O} = \mathbf{O}\mathbf{O}' = \mathbf{I}$). Since $\mathbf{D}$ is diagonal, so is $\mathbf{D}^2$, and the diagonal elements of $\mathbf{D}^2$ are just the squares of the corresponding diagonal elements of $\mathbf{D}$. Thus the diagonal elements of $\mathbf{D}$ must be idempotent numbers, satisfying $x^2 = x$, and the only such numbers are zero and one.

Hence we have another characterization of orthogonal projections

> *An orthogonal projection is a symmetric matrix having all eigenvalues either zero or one. Its rank is the number of nonzero eigenvalues.*

(The comment about the rank follows from the fact that $\mathbf{D}$ clearly has this rank, since it maps an arbitrary vector to one having this many nonzero components, and the fact that an orthogonal matrix, being invertible maps one subspace to another of the same dimension.)

We say that a pair of orthogonal projections $\mathbf{A}$ and $\mathbf{B}$ are *orthogonal* to each other if $\mathbf{AB} = 0$. Using the fact that the transpose of a product is the product of the transposes in reverse order, we see that for any symmetric matrices $\mathbf{A}$ and $\mathbf{B}$ (projections or not)

$$(\mathbf{AB})' = \mathbf{B}'\mathbf{A}' = \mathbf{BA} \tag{H.2}$$

Thus for orthogonal projections $\mathbf{AB} = 0$ implies $\mathbf{BA} = 0$ and vice versa.

The terminology here may be a bit confusing, because we are using "orthogonal" to mean two slightly different but closely related things. When applied to one matrix, it means symmetric and idempotent. When applied to two matrices, it means the product is zero. The relationship between the two usages is as follows. If $\mathbf{A}$ is an orthogonal projection, then so is $\mathbf{I} - \mathbf{A}$, because

$$(\mathbf{I} - \mathbf{A})^2 = \mathbf{I}^2 - 2\mathbf{IA} + \mathbf{A}^2 = \mathbf{I} - 2\mathbf{A} + \mathbf{A} = \mathbf{I} - \mathbf{A}.$$

Then (H.1) says these two orthogonal projections (in the first usage) are orthogonal to each other (in the second usage).

We say a set $\{\, \mathbf{A}_i : i = 1, \ldots, k \,\}$ of orthogonal projections is *orthogonal* (that is, it is an *orthogonal* set of *orthogonal* projections) if $\mathbf{A}_i\mathbf{A}_j = 0$, when $i \neq j$.

Another useful fact about orthogonal projections is the following.

Lemma H.1. *If orthogonal projections* $\mathbf{A}$ *and* $\mathbf{B}$ *satisfy*

$$\operatorname{range}(\mathbf{A}) \subset \operatorname{range}(\mathbf{B}), \tag{H.3}$$

then

$$\mathbf{A} = \mathbf{AB} = \mathbf{BA}. \tag{H.4}$$

Proof. $\mathbf{A} = \mathbf{BA}$ follows from the fact that $\mathbf{B}$ behaves like the identity map on range($\mathbf{B}$) which includes range($\mathbf{A}$). But this implies that $\mathbf{BA}$ is a symmetric matrix, hence (H.2) implies the other equality in (H.4). □

H.2 Chi-Squares

Theorem H.2. *Suppose* $\mathbf{X} = (X_1, \ldots, X_n)$ *is a multivariate normal random vector with mean vector zero and variance matrix* $\mathbf{M}$ *that is an orthogonal projection having rank* k*, then*

$$\mathbf{X}'\mathbf{X} = \sum_{i=1}^{n} X_i^2 \sim \text{chi}^2(k).$$

Proof. Being an orthogonal projection, the variance matrix has a spectral decomposition $\mathbf{M} = \mathbf{ODO}'$ in which the diagonal elements of $\mathbf{D}$ are k ones and $n - k$ zeros. By reordering the indices, we can arrange the first k to be ones.

Define $\mathbf{Y} = \mathbf{O}'\mathbf{X}$. Then

$$\text{var}(Y) = \mathbf{O}'\mathbf{MO} = \mathbf{D}.$$

Thus the components of $\mathbf{Y}$ are uncorrelated (because $\mathbf{D}$ is diagonal) and hence independent ($\mathbf{Y}$ being a linear transformation of a multivariate normal is multivariate normal, and uncorrelated implies independent for multivariate normal). The first k components are standard normal, and the last $n-k$ are concentrated at zero (because their variance is zero). Thus

$$\mathbf{Y}'\mathbf{Y} = \sum_{i=1}^{k} Y_i \sim \text{chi}^2(k).$$

But

$$\mathbf{Y}'\mathbf{Y} = (\mathbf{O}'\mathbf{X})'(\mathbf{O}'\mathbf{X}) = \mathbf{X}'\mathbf{OO}'\mathbf{X} = \mathbf{X}'\mathbf{X}$$

So $\mathbf{X}'\mathbf{X}$ also has this distribution, which is what the theorem asserts. □

Note that by the definition of the length (norm) of a vector

$$\|\mathbf{X}\|^2 = \mathbf{X}'\mathbf{X}$$

so we sometimes call the random variable described by the theorem $\|\mathbf{X}\|^2$.

Theorem H.3. *Suppose* $\mathbf{Z} = (Z_1, \ldots, Z_n)$ *is a multivariate standard normal random vector (that is, the* Z_i *are i. i. d. standard normal) and* $\{\, \mathbf{A}_i : i = 1, \ldots, k \,\}$ *is an orthogonal set of orthogonal projections, then*

$$\mathbf{Y}_i = \mathbf{A}_i\mathbf{Z}, \qquad i = 1, \ldots, k,$$

are independent random variables, and

$$\mathbf{Y}_i'\mathbf{Y}_i \sim \text{chi}^2(\text{rank}(\mathbf{A}_i)).$$

Proof. First note that the $\mathbf{Y}_i$ are jointly multivariate normal (because they are linear transformations of the same multivariate normal random vector $\mathbf{Z}$). Thus by the corollary to Theorem 13 of Chapter 12 in Lindgren they are independent if uncorrelated. Hence we calculate their covariance matrices

$$\begin{aligned} \operatorname{cov}(\mathbf{Y}_i, \mathbf{Y}_j) &= \operatorname{cov}(\mathbf{A}_i \mathbf{Z}, \mathbf{A}_j \mathbf{Z}) \\ &= E\{\mathbf{A}_i \mathbf{Z} (\mathbf{A}_j \mathbf{Z})'\} \\ &= E\{\mathbf{A}_i \mathbf{Z} \mathbf{Z}' \mathbf{A}_j\} \\ &= \mathbf{A}_i E(\mathbf{Z} \mathbf{Z}') \mathbf{A}_j \\ &= \mathbf{A}_i \operatorname{var}(\mathbf{Z}) \mathbf{A}_j \\ &= \mathbf{A}_i \mathbf{A}_j \end{aligned}$$

and this is zero when $i \neq j$ by assumption. That proves the independence assertion.

The chi-square assertion follows from Theorem H.2, because

$$\operatorname{var}(\mathbf{Y}_i) = \operatorname{cov}(\mathbf{Y}_i, \mathbf{Y}_i) = \mathbf{A}_i \mathbf{A}_i = \mathbf{A}_i$$

because $\mathbf{A}_i$ is idempotent. □