James E. Gentle

A Companion for Mathematical Statistics

Preface: Mathematical Statistics

After teaching mathematical statistics for several years using chalk on a blackboard (and, later, smelly "dry erase markers" on a whiteboard) dwelling on the minutiae of theorems and examples, I decided to lecture from computer slides that provide an outline of the "big picture". Rather than spend class time doing the same proofs and going over the same examples, I decided that time would be better spent discussing the material from a different, higher-level perspective, and addressing the details only as I felt necessary, or as the students tell me is necessary.

It is of course expected that the student will read the primary textbook, as well as various other texts, and to work through all proofs and examples in the primary textbook. As a practical matter, obviously, even if I attempted to cover all of these in class, there just is not enough class time to do it.

After writing class slides (in LaTeX 2_ε, of course), mostly in bullet form, I began writing text around the bullets, and I put the notes on the class website. Later I decided that a single document with a subject index (see pages 479 through 488) would be useful to serve as a *Companion* for the study of mathematical statistics. Much of the present document reflects its origin as classroom notes; it contains many sentence fragments, and it lacks connective material in many places. (The connective material was (probably!) supplied orally during the lectures.) *Several sections are incomplete. That does not mean that the material is unimportant; it just means I have not had time to write up the material.*

The order of presentation has changed over the years that I have taught these courses. Initially, I began with Lehmann and Casella (1998) and covered it more-or-less sequentially until the first semester ended. Then in the second semester, I covered as much of Lehmann (1986) as I felt reasonable, again more-or-less sequentially, but trying to find a reasonable starting point to articulate with the material in Lehmann and Casella covered in the first semester. For the past few years I have used Shao (2003) as the primary text. The first time I followed Shao more-or-less sequentially, but each year I have

deviated a little more from that order, even though I continued to use it as the text.

This document is organized more closely to the order in which I cover the topics now. The exception is the coverage of the appendices. I cover some of this material first, especially Appendix D.2. Occasionally later I spend some class time on other material from the appendices, but I generally expect the appendices to be used for reference as they may be needed for the statistical topics being covered.

References are given to the related sections of Shao (2003), Lehmann and Casella (1998) ("TPE2"), and Lehmann and Romano (2005) ("TSH3"). One or the other of these have been required for CSI 972/973 each year. These texts state all of the important theorems, and in most cases, provide the proofs. They are also replete with examples. Full bibliographic citations for these references, as well as several other general resources are given in the Bibliography beginning on page 477. More specific references cited only in one chapter are given in the chapter Notes sections.

The purpose of this evolving document is not just to repeat all of the material in those other texts. Its purpose, rather, is to provide some additional background material, and to serve as an outline and a handy reference of terms and concepts. The nature of the propositions vary considerably; in some cases, a fairly trivial statement will be followed by a proof, and in other cases, a rather obtuse statement will not be supported by proof. In all cases, the student should understand why the statement is true (or, if it's not, immediately send me email to let me know of the error!).

I expect each student to read the primary textbook, and to work through the proofs and examples at a rate that matches the individual student's understanding of the individual problem. What one student thinks is rather obtuse, another student comprehends quickly, and then the tables are turned when a different problem is encountered. There is a lot of lonely work required, and this is why lectures that just go through the details are often not useful.

Notation

Adoption of notation is an overhead in communication. I try to minimize that overhead by using notation that is "standard", and using it locally consistently.

Examples of sloppy notation abound in mathematical statistics. Functions seem particularly susceptible to abusive notation. It is common to see "$f(x)$" and "$f(y)$" used in the same sentence to represent two different functions. (These are often two different PDFs, one for a random variable X and the other for a random variable Y. When I want to talk about two different things, I denote them by different symbols. When I want to talk about two different PDFs, I often use notation such as "$f_X(\cdot)$" and "$f_Y(\cdot)$". If $x = y$, which is of course very different from saying $X = Y$, then $f_X(x) = f_X(y)$ obviously.) For a function and a value of a function, there is a certain amount of ambiguity

that is almost necessary. I generally try to use notation such as "$f(x)$" to denote the value of the function f at x, and I use "f" or "$f(\cdot)$" to denote the function itself (although occasionally, I do use "$f(x)$" to represent the function — notice the word "try" in the previous paragraph). If, in the notation "$f(x)$", "x" denotes a real number, then "$f(A)$" does not make much sense if A is a set. For the image of A under f, I use "$f[A]$".

Appendix A provides a list of the common notation that I use. The reader is encouraged to look over that list both to see the notation itself and to get some idea of the objects that I discuss.

Easy Pieces

I recommend that all students develop a list of "easy pieces". These are propositions or examples and counterexamples that the student can state and prove or describe and work through *without resort to notes*. They may also include definitions, stated precisely.

Some easy pieces culled from the material presented in CSI 972 in Fall, 2007 are

- Let $\mathcal{C}$ be the class of all closed intervals in $\mathbb{R}$. Show that $\sigma(\mathcal{C}) = \mathcal{B}(\mathbb{R})$ (the real Borel σ-field).
- Define induced measure and prove that it is a measure. That is, prove: If $(\Omega, \mathcal{F}, \nu)$ is a measure space and $(\Lambda, \mathcal{G})$ is a measurable space, and f is a function from Ω to Λ that is measurable with respect to $\mathcal{F}/\mathcal{G}$, then the domain and range of the function $\nu \circ f^{-1}$ is $\mathcal{G}$ and it is a measure.
- Define the Lebesgue integral for a general Borel function.
- State and prove Fatou's lemma conditional on a sub-σ-field.
- State and prove the information inequality (CRLB) for a d-vector parameter. (Get the regularity conditions correct.)
- Give an example to distinguish the asymptotic bias from the limiting bias.
- State and prove Basu's theorem.
- Give an example of a function of some parameter in some family of distributions that is not U-estimable.

Make your own list of easy pieces.

"It Is Clear"

I tend to use the phrase "it is clear ..." often. (I only realized this recently, because someone pointed it out to me.) When I say "it is clear ...", I expect the reader to agree with me actively, not passively.

I use this phrase only when the statement is "clearly" true to me. I must admit, however, sometimes when I read the statement a few weeks later, it's not very clear! It may require many minutes of difficult reasoning. In any event, the reader should attempt to supply the reasoning for everything that I say is clear.

My Courses

The courses in mathematical statistics at George Mason University are CSI/STAT 972 and CSI/STAT 973. Until recently, the prerequisites for these courses did not include any advanced probability theory. Therefore, a substantial amount of CSI/STAT 972 was devoted to probability theory and a certain amount of the underlying measure theory. The courses now carry a prerequisite of a course in measure-theoretic-based probability theory. The coverage of probability as such in CSI/STAT 972 is consequently decreasing. Chapter 1 and the appendices address the prerequisite material briefly. Even after it can be assumed that all students in CSI/STAT 972 have the probability course prerequisite, however, some introductory coverage of probability will remain in CSI/STAT 972. Although Chapter 1 is on "probability", the focus is more on what is usually covered in "statistics" courses, such as families of disrtribtions, in particular, the exponential class of families.

My notes on these courses are available at

`http://mason.gmu.edu/~jgentle/csi9723/`

Acknowledgements

A work of this size is bound to have some errors (at least if I have anything to do with it!). Errors must first be detected and then corrected. I thank the students who have given me many corrections, either in the original lecture notes or in the later documents. Bruce McCullough over the years has detected several errors that I have been able to correct, and has given me many suggestions for improvements, not only in this book, but in other writing. I especially thank him.

I would appreciate any feedback – errors, comments, or suggestions. Email me at `jgentle@gmu.edu`

James E. Gentle
Fairfax County, Virginia
December 30, 2008

Contents

Appendices

1

Probability
(Shao Ch 1, Sec 5.2; TPE2 Ch 1; TSH3 Ch 2)

Probability theory provides the basis for mathematical statistics. This chapter covers important topics in probability theory at a fairly fast pace. Probability theory is based on measure theory, so the presentation in this chapter assumes familiarity with the material in Section D.2.

We begin in Section 1.1 with a statement of definitions and some basic properties. In some cases, proofs are given; in others, references are given; and in others, it is assumed that the reader supplies the reasoning.

Sections 1.2 and 1.3 are concerned with sequences of independent random variables. The limiting properties of such sequences are important. Many of the limiting properties can be studied using expansions in power series, which is the topic of Section 1.4.

Section 1.5 is devoted to conditional probability, which is not a fundamental concept, as probability itself is. Conditional probability, rather, is based on conditional expectation as the fundamental concept, so that is where we begin. This provides a more general foundation for conditional probability than we would have if we defined it more directly in terms of a measurable space. Conditional probability plays an important role in sequences that lack a simplifying assumption of independence. We discuss sequences that lack independence in Section 1.6. Many interesting sequences also do not have identical marginal distributions, but rather follow some kind of evolving model whose form depends on, but is not necessarily determined by, previous variates in the sequence.

The final section identifies and describes useful classes of probability distributions. These classes are important because they are good models of observable random phenomena, and because they are easy to work with. The properties of various statistical methods discussed in subsequent chapters depend on the underlying probability model, and some of the properties of the statistical methods can be worked out easily for particular models discussed in Section 1.7.

1.1 Some Important Probability Definitions and Facts

A probability distribution is built from a σ-field, say $\mathcal{F}$, defined on a sample space, say Ω, and a σ-finite probability measure, say P. Properties of the distribution and statistical inferences regarding it are derived and evaluated in the context of the "probability triple", $(\Omega, \mathcal{F}, P)$. Given a probability space $(\Omega, \mathcal{F}, P)$, a set $A \in \mathcal{F}$ is called an "event". In practice, the probability measure P is usually based either on the counting measure (defined on countable sets as their cardinality) or on the Lebesgue measure (the length of intervals).

In many ways the content of this section parallels that of Section D.2 (page 368) for more general measures.

1.1.1 Definitions of Probability and Probability Distributions

The basic ideas of probability are developed by consideration of a special measure and the measure space it is part of. We first consider this special function defined on subsets of the sample space, and then we consider a special type of function of the elements of the sample space, called a random variable. Random variables allow us to develop a theory of probability that is useful in statistical applications.

Probability Measure on Events: Definitions

Definition 1.1 (probability measure)
A measure ν whose domain is a σ-field defined on the sample space Ω with the property that $\nu(\Omega) = 1$ is called a probability measure. We often use P to denote such a measure.

Definition 1.2 (probability space)
If P in the measure space $(\Omega, \mathcal{F}, P)$ is a probability measure, the triple $(\Omega, \mathcal{F}, P)$ is called a probability space.

The elements in the probability space can be any kind of objects. They do not need to be numbers.

Definition 1.3 (probability of an event)
The probability of the event A is $P(A)$, also written as $\Pr(A)$.

The probability of A is $\int_A \mathrm{d}P$.

Definition 1.4 (independence)
We define independence in a probability space $(\Omega, \mathcal{F}, P)$ in three steps:

- **Independence of events** *within a collection of events.*
  Let $\mathcal{C}$ be a collection of events; that is, a collection of subsets of $\mathcal{F}$. The events in $\mathcal{C}$ are independent iff for any positive integer n and distinct events $A_1, \ldots, A_n$ in $\mathcal{C}$,

$$P(A_1 \cap \cdots \cap A_n) = P(A_1) \cdots P(A_n).$$

Sometimes people use the phrase "mutually independent" to try to emphasize that we are referring to independence of all events. We can have the situation in which all pairs within the collection are independent, but the collection is not independent; for example, in an experiment of tossing a coin twice, let
A be "heads on the first toss"
B be "heads on the second toss"
C be "exactly one head and one tail on the two tosses"
We see immediately that any pair is independent, but that the intersection is $\emptyset$.
BTW, the phrase "mutually independent" could be interpreted as "pairwise independent", so "mutually" really does not clarify anything.

- **Independence of collections of events** *(and, hence, of σ-fields).*
  For any index set $\mathcal{I}$, let $\mathcal{C}_i$ be a collection of sets with $\mathcal{C}_i \subset \mathcal{F}$. The collections $\mathcal{C}_i$ are independent iff the events in any collection of the form $\{A_i \in \mathcal{C}_i : i \in \mathcal{I}\}$ are independent.
- **Independence of functions** *(and, hence, of random variables).*
  (This also defines independence of any generators of σ-fields.)
  The Borel-measurable functions X_i, for $i \in \mathcal{I}$, are independent iff $\sigma(X_i)$ for $i \in \mathcal{I}$ are independent.

Definition 1.5 (exchangeability)
We define exchangeability in a probability space $(\Omega, \mathcal{F}, P)$ in three steps, similar to those in the definition of independence:

- **Exchangeability of events** *within a collection of events. Given a probability measure P, two events A and B are said to be exchangeable with respect to P if $P(A \cap B^{\mathrm{c}}) = P(A^{\mathrm{C}} \cap B)$. This definition can be extended to a collection of events $\mathcal{C}$ in an obvious manner.*
- **Exchangeability of collections of events** *(and, hence, of σ-fields).*
  For any index set $\mathcal{I}$, let $\mathcal{C}_i$ be a collection of sets with $\mathcal{C}_i \subset \mathcal{F}$. The collections $\mathcal{C}_i$ are exchangeable iff the events in any collection of the form $\{A_i \in \mathcal{C}_i : i \in \mathcal{I}\}$ are exchangeable.
- **Exchangeability of functions** *(and, hence, of random variables).*
  (This also defines exchangeable of any generators of σ-fields.)
  The Borel-measurable functions X_i, for $i \in \mathcal{I}$, are exchangeable iff $\sigma(X_i)$ for $i \in \mathcal{I}$ are exchangeable.

Independence implies exchangeability, but exchangeability does not imply independence. To see this, we first note that $P(A \cap B^{\mathrm{c}}) = P(A \cap B)$ iff A and B are independent; hence, independence implies exchangeability.

A simple urn example may illustrate the difference in exchangeability and independence. Suppose an urn contains 15 balls, 10 of which are red. We "randomly" draw balls from the urn without replacing them. Let R_i be the

event that a red ball is drawn on the i^{th} draw, and $\overline{R}_i$ be the event that a non-red ball is drawn. We see the following

$$\Pr(R_1) = \Pr(R_2) = \cdots = \Pr(R_{15}) = 2/3$$

and

$$\Pr(\overline{R}_1) = \Pr(\overline{R}_2) = \cdots = \Pr(\overline{R}_{15}) = 1/3.$$

Now

$$\Pr(R_2|R_1) = 5/7,$$

hence R_1 and R_2 are not independent. However,

$$\Pr(R_2 \cap \overline{R}_1) = \Pr(\overline{R}_2 \cap R_1) = 5/21.$$

hence R_1 and R_2 are exchangeable. In fact, we could extend the latter computations (by a binomial tree) to see that the elements of any subset of the 15 R_is is exchangeable.

Definition 1.6 (support of a probability measure)
If the probability measure P is defined with respect to the σ-field $\mathcal{F}$, $S \in \mathcal{F}$, and $P(S) = 1$, then S is called a support of the probability measure.

Random Variables and Probability Distributions: Definitions

Definition 1.7 (random variable)
Given a measurable space $(\Omega, \mathcal{F})$, a random variable is a real-valued measurable function, $X(\omega)$ or just X, defined on Ω.

(Recall that I often use "real" also to mean a vector over the reals. Although we will assume X is real, it does not have to be, and we could form a theory of probability and statistics that allowed X to be over a general field.) In our extended meaning of the symbol "$\in$" (see page 345), we write $X \in \mathcal{F}$.

Note that a constant is a random variable. If c is a constant and if $X = c$ a.s., then we call either c or X a *degenerate* random variable.

We often denote the image of X as $\mathcal{X}$.

Definition 1.8 (σ-field generated by a random variable)
As with any measurable function, we have a σ-field generated by a random variable. If $X : \Omega \mapsto B \subset \mathbb{R}^d$, then we can see that $\sigma(X^{-1}[B])$ is a sub-σ-field of $\mathcal{F}$. We call this the σ-field generated by X, and write it as $\sigma(X)$.

If X and Y are random variables defined on the same measurable space, we may write $\sigma(X, Y)$, with the obvious meaning. As with σ-fields generated by sets or functions discussed before, it is clear that $\sigma(X) \subset \sigma(X, Y)$. This idea of sub-$\sigma$-fields generated by random variables is important in the analysis of a sequence of random variables.

Notice that a random variable is defined in terms only of a measurable space $(\Omega, \mathcal{F})$ and a measurable space defined on the reals. No associated probability measure is necessary for the definition, but for meaningful applications of a random variable, we need some probability measure.

Definition 1.9 (probability distribution of a random variable)
Given the probability space $(\Omega, \mathcal{F}, P)$ and the random variable X defined on $(\Omega, \mathcal{F})$, the probability distribution of X is $P \circ X^{-1}$. The probability distribution is also just called the distribution or the law.

For a given random variable X, a probability distribution determines $\Pr(X \in A)$ for $A \in \mathcal{A}$. An underlying probability measure of course determines $\Pr(X \in A)$.

The *support of the distribution* (or of the random variable) is the closure of the smallest set $\mathcal{X}_S$ in the image of X such that $P(X^{-1}[\mathcal{X}_S]) = 1$.

Definition 1.10 (family of probability distributions)
A probability family or family of distributions, $\mathcal{P} = \{P_\theta, \theta \in \Theta \subset \mathbb{R}^d\}$, is a set of probability distributions of a random variable that is defined over Ω.

We call θ the *parameter* and Θ the parameter space. If the dimension of Θ is large (there is no precise meaning of "large" here), we may refrain from calling θ a parameter, because we want to refer to some statistical methods as "nonparametric". (In nonparametric methods, our analysis usually results in some general description of the distribution, rather than in a specification of the distribution.)

A family of distributions on a measurable space $(\Omega, \mathcal{F})$ with probability measures P_θ for $\theta \in \Theta$ is called a *parametric family* if $\Theta \subset \mathbb{R}^k$ for some fixed positive integer k and θ fully determines the measure.

We assume that every parametric family is *identifiable*; that is, $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ is an identifiable parametric family if it is a parametric family and for $\theta_1, \theta_2 \in \Theta$ if $\theta_1 \neq \theta_2$ then $P_{\theta_1} \neq P_{\theta_2}$.

A family that cannot be indexed in this way is called a nonparametric family.

An example of a parametric family of distributions for the measurable space $(\Omega = \{0, 1\}, \mathcal{F} = 2^\Omega)$ is that formed from the class of the probability measures $P_\pi(\{1\}) = \pi$ and $P_\pi(\{0\}) = 1 - \pi$. This is a parametric family, namely, the Bernoulli distributions. The measures are dominated by the counting measure.

An example of a nonparametric family over some measurable space $(\Omega, \mathcal{F})$ is $\mathcal{P}_c = \{P : P \ll \nu\}$, where ν is the Lebesgue measure.

Definition 1.11 (cumulative distribution function (CDF))
If $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), P)$ is a probability space, and F is defined by

$$F(x) = P((-\infty, x]) \quad \forall\, x \in \mathbb{R}^d,$$

then F is called a cumulative distribution function, or CDF.

The CDF is also called the distribution function, or DF. The CDF is particularly useful in the case $d = 1$.

The probability space completely determines F, and likewise, F completely determines P a.s.; hence, we often use the CDF and the probability measure interchangeably.

If the random variable is assumed to be in a family of distributions indexed by θ, we may use the notation $F_\theta(x)$ or $F(x;\theta)$.

For a given random variable X, $F(x) = \Pr(X \leq x)$. (If X is a vector-valued random variable, and x is a vector of the same order, $X \leq x$ is interpreted to mean that $X_i \leq x_i$ for each respective element.)

From the definition, four properties of a CDF are immediate:

- $\lim_{x\downarrow-\infty} F(x) = 0$.
- $\lim_{x\uparrow\infty} F(x) = 1$.
- $F(x_1) \leq F(x_2)$ if $x_1 \leq x_2$.
- $\lim_{\epsilon\downarrow 0} F(x+\epsilon) = F(x)$. (A CDF is continuous from the right.)

These four properties characterize a CDF, so they can serve as an alternate definition of a CDF, without reference to a probability distribution. Notice, for example, that the Cantor function (see Section D.2.4) is a CDF if we extend its definition to be 0 on $(-\infty, 0)$ and to be 1 on $(1, \infty)$.

Definition 1.12 (probability density function) *(PDF)*
The derivative of a CDF (or, equivalently, of the probability measure) with respect to an appropriate measure, if it exists, is called the probability density function, PDF.

The PDF is also called the density function.

We use these terms for either discrete random variables and the counting measure or for continuous random variables and Lebesgue measure. This we take as the general meaning of the term "discrete random variable": the probability measure is dominated by the counting measure; and likewise for a "continuous random variable": the probability measure is dominated by Lebesgue measure. Any simple CDF has a PDF wrt the counting measure, but not every continuous CDF has a PDF wrt Lebesgue measure (the Cantor function is a classic counterexample), but every absolutely continuous CDF does have a PDF wrt Lebesgue measure.

We can also think of mixtures of densities in the context of a continuous random variable with positive mass points.

The "appropriate measure" in the definition of PDF above must be σ-finite and must dominate the probability density function.

If a specific CDF is F_θ, we often write the corresponding PDF as f_θ:

$$f_\theta = \frac{\mathrm{d}F_\theta}{\mathrm{d}\nu}$$

If the specific probability measure is P_θ, we also often write the corresponding PDF as p_θ:

$$p_\theta = \frac{\mathrm{d}P_\theta}{\mathrm{d}\nu}.$$

The dominating measure for a given probability distribution is not unique, but use of a different dominating measure may change the representation of the distribution. For example, suppose that the support of a distribution is S, and so we write the PDF as

$$\frac{\mathrm{d}F_\theta}{\mathrm{d}\nu} = g(x,\theta)\mathrm{I}_S(x).$$

If we define a measure λ by $\lambda(A) = \int_A \mathrm{I}_S \mathrm{d}\nu \; \forall A \in \mathcal{F}$, then we could write the PDF as

$$\frac{\mathrm{d}F_\theta}{\mathrm{d}\lambda} = g(x,\theta).$$

Although we have already defined independence and exchangeability in general, it is useful to give equivalent definitions for random variables.

Definition 1.13 (independence of random variables)
The random variables $X_1, \ldots, X_k$ on $(\Omega, \mathcal{F}, P)$ are said to be independent iff for any sets $B_1, \ldots, B_k$ in the Borel σ-field,

$$P(X_1 \in B_1, \ldots, X_k \in B_k) = \prod_{i=1}^{k} P(X_i \in B_i).$$

Definition 1.14 (exchangeability of random variables)
The random variables $X_1, \ldots, X_k$ on $(\Omega, \mathcal{F}, P)$ are said to be exchangeable iff the joint distribution of $X_1, \ldots, X_k$ is the same as the joint distribution of $\Pi(\{X_1, \ldots, X_k\})$, for any Π, where $\Pi(A)$ denotes a permutation of the elements of the set A.

We can specify the distribution of a random variable by giving the CDF or the PDF. There are also a number of useful distributions that we give names. For example, the normal or Gaussian distribution, the binomial distribution, the chi-squared, and so on. Each of these distributions is actually a family of distributions. A specific member of the family is specified by specifying the value of each parameter associated with the family of distributions.

For some distributions, we introduce special symbols to denote the distribution. For example, we use $\mathrm{N}(\mu, \sigma^2)$ to denote a univariate normal distribution with parameters μ and σ^2 (the mean and variance). To indicate that a random variable has a normal distribution, we use notation of the form

$$X \sim \mathrm{N}(\mu, \sigma^2),$$

which here means that the random variable X has a normal distribution with parameters μ and σ^2.

In some cases, we also use special symbols to denote random variables with particular distributions. For example, we often use χ^2_ν to denote a random variable with a chi-squared distribution with ν degrees of freedom.

1.1.2 Definitions and Properties of Expected Values

First we define the expected value of a random variable:
Given a probability space $(\Omega, \mathcal{F}, P)$ and a d-variate random variable X defined on $\mathcal{F}$, we define the *expected value* of X with respect to P, which we denote by $\mathrm{E}(X)$ or for clarity by $\mathrm{E}_P(X)$, as

$$\mathrm{E}(X) = \int_{\mathbb{R}^d} X \,\mathrm{d}P. \tag{1.1}$$

Sometimes we limit this definition to integrable random variables X.

Look carefully at the integral. It is the integral of a function, X, over Ω with respect to a measure, P, over the σ-field that together with Ω forms the measurable space. To emphasize the meaning more precisely, we could write the integral in the definition as

$$\mathrm{E}(X) = \int_{\Omega} X(\omega) \,\mathrm{d}P(\omega).$$

We can also write the expectation in terms of the range of the random variable and an equivalent measure on range. If the CDF of the random variable is F, we have, in the abbreviated form of the first expression given in the definition,

$$\mathrm{E}(X) = \int x \,\mathrm{d}F,$$

or in the more precise form,

$$\mathrm{E}(X) = \int_{\mathbb{R}^d} x \,\mathrm{d}F(x).$$

If g is a Borel function, we define the expected value of $g(X)$ in the same way: $\mathrm{E}(g(X)) = \int_{\mathbb{R}^d} g(x) \,\mathrm{d}F(x)$.

If the PDF exists and is f, we also have

$$\mathrm{E}(X) = \int_{\mathbb{R}^d} x f(x) \,\mathrm{d}x.$$

From the definition of expectation, it is clear that if X and Y are random variables defined over the same probability space,

$$X \leq Y \text{ a.s.} \quad \Rightarrow \quad \mathrm{E}(X) \leq \mathrm{E}(Y). \tag{1.2}$$

Additionally,

$$g(X) \geq 0 \text{ a.s.} \quad \Rightarrow \quad \mathrm{E}(g(X)) \geq 0. \tag{1.3}$$

Expected Value and Probability

There are many interesting relationships between expected values and probabilities. For example, for a positive random variable X,

$$\mathrm{E}(X) = \int_0^\infty \Pr(X > t)\mathrm{d}t. \tag{1.4}$$

We can see this is a simple application of Fubini's theorem:

$$\begin{aligned}
\mathrm{E}(X) &= \int_0^\infty x \,\mathrm{d}F(x) \\
&= \int_0^\infty \int_{(0,x)} \mathrm{d}t \,\mathrm{d}F(x) \\
&= \int_0^\infty \int_{(t,\infty)} \mathrm{d}F(x)\mathrm{d}t \\
&= \int_0^\infty (1 - F(t))\mathrm{d}t \\
&= \int_0^\infty \Pr(X > t)\mathrm{d}t
\end{aligned}$$

This leads in general to the useful property for any given random variable X, if $\mathrm{E}(X)$ exists:

$$\mathrm{E}(X) = \int_0^\infty (1 - F(t))\mathrm{d}t - \int_{-\infty}^0 F(t)\mathrm{d}t. \tag{1.5}$$

Expected Value of the Indicator Function

We define the indicator function, $\mathrm{I}_A(x)$, as 1 if $x \in A$ and 0 otherwise. (This is also called the "characteristic function", but we use that term to refer to something else.) If X is an integrable random variable over A, then $\mathrm{I}_A(X)$ is an integrable random variable, and

$$\Pr(A) = \mathrm{E}(\mathrm{I}_A(X)). \tag{1.6}$$

When it is clear from the context, we may omit the X, and merely write $\mathrm{E}(\mathrm{I}_A)$.

Expected Value over a Measurable Set

The expected value of an integrable random variable over a measurable set $A \subset \mathbb{R}^d$ is

$$\mathrm{E}(X\mathrm{I}_A(X)) = \int_A X \,\mathrm{d}P.$$

We often denote this as $\mathrm{E}(X\mathrm{I}_A)$.

Entropy

Probability theory is developed from models that characterize uncertainty inherent in random events. Information theory is developed in terms of the information revealed by random events. The premise is that the occurrence of an event with low probability is more informative than an event of high probability. For a discrete random variable we can effectively associate a value of the random variable with an event, and we quantify information in such a way that the information revealed by a particular outcome decreases as the probability increases. We define the *self-information* of an event or the value of a discrete random variable with PDF p_X as $-\log_2(p_X(x))$. The logarithm to the base 2 comes from the basic representation of information in base 2, but we can equivalently use any base, and it is common to use the natural log in the definition of self-information.

We define the *entropy* of a discrete random variable X as the expected value of the self-information evaluated at the random variable,

$$H(X) = -\sum_x p_X(x)\log(p_X(x)). \tag{1.7}$$

We can see that the entropy is maximized if all outcomes are equally probable. In the case of a discrete random variable with two outcomes with probabilities π and $1-\pi$ (a Bernoulli random variable with parameter π), the entropy is $-\pi\log(\pi) - (1-\pi)\log(1-\pi)$. It is maximized when $\pi = 1/2$.

Although the definitions of information theory are generally given in the context of a countable sample space, they can be extended in an obvious way using equation (1.7) with p the PDF, whether it is dominated by a counting measure or not. It is just the expected value

$$\mathrm{E}(-\log(p_X(x))). \tag{1.8}$$

It should also be clear how to define the *joint entropy* $H(X,Y)$ in terms of the joint PDF $p_{X,Y}$.

Elementwise Moments

For the random variable X, $\mathrm{E}(X)$, if it exists, is called the *first moment of* X. For $r \geq 1$, the r^{th} *moment* of X, if it exists, is $\mathrm{E}(X^r)$. We often denote the r^{th} moment as μ'_r. The r^{th} moment is often called the r^{th} *raw* moment, because *central moments* or *moments about* $\mathrm{E}(X)$ are often more useful. The r^{th} *central moment* of X, if it exists, is $\mathrm{E}((X-\mathrm{E}(X))^r)$. We often denote the r^{th} central moment as μ_r. Note that $\mu'_1 \equiv \mu_1$. The first two central moments are usually the most important; μ_1 is called the *mean* and μ_2 is called the *variance*.$\mathrm{V}(\cdot)$ The variance of X is denoted by $\mathrm{V}(\cdot)$. (Note that for a d-vector random variable X, this is a d-vector whose elements correspond to the variances of the individual elements of X. A more important concept is the

variance-covariance matrix defined below. Also note that the term "variance" in the case of a random vector usually means the variance-covariance.) Because $(X-\mathrm{E}(X))^2 \geq 0$ a.s., we see that the variance is nonnegative. The square root of the variance is called the *standard deviation*. If it exists, $\mathrm{E}(|X|)$ is called the *first absolute moment* of X; and generally, if it exists, $\mathrm{E}(|X|^r)$ is called the r^{th} absolute moment.

Variance-Covariance

The variance-covariance of the random variable X, if it exists, is the expectation of the outer product,

$$\mathrm{V}(X) = \mathrm{E}\left((X - \mathrm{E}(X))(X - \mathrm{E}(X))^{\mathrm{T}}\right).$$

Although the rank of an outer product is no greater than 1, unless $X = \mathrm{E}(X)$ a.s., $\mathrm{V}(X)$ is nonnegative definite. (This follows from nonnegativity of the variance and the covariance inequality (1.43).)

Expected Value of General Measurable Functions

A real-valued measurable function g of a random variable X is itself a random variable, possibly with a different probability measure. Its expected value is defined in exactly the same way as above. If the probability triple associated with the random variable X is $(\Omega, \mathcal{F}, P)$ and $Y = g(X)$, we could identify a probability triple associated with Y. Being measurable, the relevant measurable space of $g(X)$ is $(\Omega, \mathcal{F})$, but the probability measure is not necessarily P. If we denote the probability triple associated with the random variable Y is $(\Omega, \mathcal{F}, Q)$, we may distinguish the defining integrals with respect to $\mathrm{d}P$ and $\mathrm{d}Q$ by E_P and E_Q.

We can also write the expected value of Y in terms of the CDF of the original random variable. The expected value of a real-valued measurable function g of a random variable X with CDF F is $\mathrm{E}(g(X)) = \int g(x)\mathrm{d}F(x)$.

Conditional Expectations and Conditional Distributions

Often the distribution of a random variable depends on the values taken on by another random variable. The expected value of the random variable depends on the values taken on by the other random variable. We will use conditional expectations to develop the concept of a conditional probability distribution.

We can study conditional distributions in some cases by beginning with the definition of conditional probability of an event under a restriction: If $\Pr(A) > 0$, the conditional probability of B written $\Pr(B|A)$, is defined by $\Pr(B|A) = \Pr(B \cap A)/\Pr(A)$. This approach is limited by the requirement $\Pr(A) > 0$. Defining conditional expectation first, and then defining the other

concepts of conditional probability in terms of conditional expectations avoids this problem.

We will just give the basic definitions and properties here, and then in Section 1.5 discuss the ideas further.

conditional expectation over a sub-σ-field

Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $\mathcal{A}$ be a sub-σ-field of $\mathcal{F}$, and let X be an integrable random variable over Ω. The *conditional expectation* of X given $\mathcal{A}$, denoted by $\mathrm{E}(X|\mathcal{A})$ is a random variable such that $\mathrm{E}(X|\mathcal{A})$ is a measurable function from $(\Omega, \mathcal{F})$ to $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and

$$\int_A \mathrm{E}(X|\mathcal{A})\,\mathrm{d}P = \int_A X\,\mathrm{d}P, \quad \forall\, A \in \mathcal{A}. \tag{1.9}$$

(The existence and uniqueness of this random variable follows from the Radon-Nikodym theorem (Shao, Theorem 1.4)). In terms of an indicator function, we have $\int_A \mathrm{E}(X|\mathcal{A})\,\mathrm{d}P = \mathrm{E}(X\mathrm{I}_A)$ for all $A \in \mathcal{A}$. Sometimes $\mathrm{E}(X\mathrm{I}_A)$ is written as $\mathrm{E}_A(X)$, but this notation is confusing, because as we note above, the subscript on the expectation operator usually refers to the probability measure used in the integration.

conditional expectation with respect to another measurable function

Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $\mathcal{A}$ be a sub-σ-field of $\mathcal{F}$, let X be an integrable random variable over Ω, and let Y be a measurable function from $(\Omega, \mathcal{F}, P)$ to any measurable space $(\Lambda, \mathcal{G})$. Then the *conditional expectation* of X given Y, denoted by $\mathrm{E}(X|Y)$ is defined as the conditional expectation of X given the sub-σ-field generated by Y, that is, $\mathrm{E}(X|\sigma(Y))$.

conditional probability

Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $\mathcal{A}$ be a sub-σ-field of $\mathcal{F}$, and let $B \in \mathcal{F}$. The *conditional probability* of B given $\mathcal{A}$, denoted by $\Pr(B|\mathcal{A})$ is defined as $\mathrm{E}(\mathrm{I}_B|\mathcal{A})$.

independence

X and Y are independent if the conditional distribution equals the marginal distribution.

$$\forall A,\ \Pr(X \in A|Y) = \Pr(X \in A).$$

(This means that we can factor the joint PDF or CDF.)

conditional independence

X and Y are conditionally independent given Z if the joint conditional distribution equals the joint marginal distribution.

$$\forall A,\ \Pr(X \in A|Y, Z) = \Pr(X \in A|Y).$$

1.1.3 Generating Functions

There are some functionals of the PDF or CDF that are useful to determine properties of the distribution.

moment-generating functions and characteristic functions

The moment-generating function (MGF) and characteristic function (CF) are *transforms* of the density function. The moment-generating function for the random variable X, which may not be well-defined (in which case we say it does not exist), is

$$\psi_X(t) = \mathrm{E}(\mathrm{e}^{tX}),$$

and the characteristic function, which is always well-defined is

$$\begin{aligned}\phi_X(t) &= \mathrm{E}(\mathrm{e}^{\mathrm{i}tX}) \\ &= \mathrm{E}(\cos(tX)) + \mathrm{i}\mathrm{E}(\sin(tX)).\end{aligned}$$

Both functions are nonnegative. If the MGF is finite for some $t \neq 0$, the CF can be obtained by replacing t in $\psi_X(t)$ by $\mathrm{i}t$ (where $\mathrm{i} = \sqrt{-1}$). The characteristic function is the Fourier transform of the density with argument $-t/(2\pi)$. Both transforms are defined for a vector-valued random variable X similarly, and the corresponding transforms are functions of a vector-valued variable t. The expression tX in the definitions above is replaced by $t^{\mathrm{T}}X$.

An interesting property of the MGF and the CF is that the (raw) moments of X can be obtained from their derivatives evaluated at 0. So we have, for example,

$$\left.\frac{\mathrm{d}^k\phi_X(t)}{\mathrm{d}t^k}\right|_{t=0} = (-1)^{k/2}\mathrm{E}(X^k).$$

For vector-valued random variables, the moments become tensors, but the first two moments are very simple: $\nabla\phi_X(t)|_{t=0} = \mathrm{E}(X)$ and $\nabla\nabla\phi_X(t)|_{t=0} = \mathrm{E}(X^{\mathrm{T}}X)$.

The CF or MGF completely determines the distribution. (This is the "Inversion Theorem", which is essentially the same theorem as the one often used in working with Fourier transforms.) Also, the limit of a sequence of CFs or MGFs determines the limiting distribution.

A nice use of CFs (or MGFs, if we are willing to assume that the MGF exists) is in the proof of a simple form of the central limit theorem that states that if $X_1, \ldots, X_n$ are i.i.d. with mean μ and variance $0 < \sigma^2 < \infty$, then $Y_n = (\sum X_i - n\mu)/\sqrt{n}\sigma$ has limiting distribution $\mathrm{N}(0, 1)$.

Proof. It will be convenient to define a function related to the CF: let $h(t) = \mathrm{e}^{\mu t}\phi_X(t)$; hence $h(0) = 1$, $h'(0) = 0$, and $h''(0) = \sigma^2$. Now expand h in a Taylor series about 0:

$$h(t) = h(0) + h'(0)\mathrm{i}t - \frac{1}{2}h''(\xi)t^2,$$

for some ξ between 0 and t. Substituting for $h(0)$ and $h'(0)$, and adding and subtracting $\sigma^2 t/2$ to this, we have

$$h(t) = 1 - \frac{\sigma^2 t^2}{2} - \frac{(h''(\xi) - \sigma^2)t^2}{2}.$$

This is the form we will find useful. Now, consider the CF of Y_n:

$$\begin{aligned}\phi_{Y_n}(t) &= \mathrm{E}\left(\exp\left(\mathrm{i}t\left(\frac{\sum X_i - n\mu)}{\sqrt{n}\sigma}\right)\right)\right) \\ &= \left(\mathrm{E}\left(\exp\left(\mathrm{i}t\left(\frac{X-\mu)}{\sqrt{n}\sigma}\right)\right)\right)\right)^n \\ &= \left(h\left(\frac{\mathrm{i}t}{\sqrt{n}\sigma}\right)\right)^n.\end{aligned}$$

From the expansion of h, we have

$$h\left(\frac{\mathrm{i}t}{\sqrt{n}\sigma}\right) = 1 - \frac{t^2}{2n} - \frac{(h''(\xi) - \sigma^2)t^2}{2n\sigma^2}.$$

So,

$$\phi_{Y_n}(t) = \left(1 - \frac{t^2}{2n} - \frac{(h''(\xi) - \sigma^2)t^2}{2n\sigma^2}\right)^n.$$

Now we need a well-known (but maybe forgotten) result: If $\lim_{n\to\infty} f(n) = 0$, then

$$\lim_{n\to\infty}\left(1 + \frac{a}{n} + \frac{f(n)}{n}\right)^b n = \mathrm{e}^{ab}.$$

Therefore, because $\lim_{n\to\infty} h''(\xi) = h''(0) = \sigma^2$, $\lim_{n\to\infty}\phi_{Y_n}(t) = \mathrm{e}^{-t^2/2}$, which is the CF of the $\mathrm{N}(0,1)$ distribution. ∎

This simple form of the CLT together with its proof is an *easy piece* that you should be able to prove relatively quickly.

cumulant-generating functions

The sequence of (raw) moments is very useful in characterizing a distribution, but often the central moments are to be preferred because, except for first, they are invariant to change in the first moment (the "location"). Another sequence of constants, which, except for the first, are invariant to change in the first moment, are called cumulants. Formally, for the random variable X with CF $\phi(t)$, the r^{th} *cumulant*, denoted κ_r, is

$$\frac{\mathrm{d}^r}{\mathrm{d}t^r}\log(\phi(t))\,|_{t=0}$$

if it exists. Thus, the cumulant-generating function is $\log(\phi(t))$, where $\phi(t)$ is the characteristic function.

Obviously, the cumulants and the moments are closely related. For the first few, for example,

$$\begin{aligned}\mu'_1 &= \kappa_1\\ \mu'_2 &= \kappa_2 + \kappa_1^2\\ \mu'_3 &= \kappa_3 + 3\kappa_2\kappa_1 + \kappa_1^3.\end{aligned}$$

frequency-generating functions and factorial-moment-generating functions

Frequency-generating functions or probability-generating functions and factorial-moment-generating functions are useful for discrete random variables.

For discrete random variable X taking values $x_1, x_2, \ldots$ with probabilities $0 < p_1, p_2, \ldots$, the *frequency-generating function* or *probability-generating function* is the polynomial

$$P(t) = \sum_{i=0}^{\infty} p_{i+1} t^i.$$

The probability of x_r is

$$\frac{\mathrm{d}^{r+1}}{\mathrm{d}t^{r+1}} P(t) \mid_{t=0}$$

The probability-generating function for the binomial distribution with parameters π and n, for example, is

$$P(t) = (\pi t + (1 - \pi))^n.$$

A discrete distribution with support $x_1, x_2, \ldots$ is equivalent to a discrete distribution with support $0, 1, \ldots$. For such a distribution another kind of moment is sometimes useful. It is the factorial moment, related to the r^{th} *factorial* of the real number y:

$$y^{[r]} = y(y-1)\cdots(y-(r-1)).$$

(We see that $y^{[y]} = y!$. It is, of course, not necessary that y be an integer, but factorials are generally more useful in the context of nonnegative integers.)

The r^{th} *factorial moment* of the random variable X above is

$$\mu'_{[r]} = \sum_{i=0}^{\infty} x_i^{[r]} p_i.$$

We see that $\mu'_{[1]} = \mu'_1 = \mu_1$.

The r^{th} central factorial moment, denoted $\mu_{[r]}$ is the r^{th} factorial moment about μ.

We see immediately that the factorial-moment-generating function is the same as the probability-generating function evaluated at $t+1$:

$$\begin{aligned} P(t+1) &= \sum_{j=0}^{\infty} p_{j+1}(t+1)^j \\ &= \sum_{j=0}^{\infty} p_{j+1} \sum_{i=1}^{j} \binom{j}{i} t^i \\ &= \sum_{i=0}^{\infty} \frac{t^i}{i!} \sum_{j=0}^{\infty} (p_{j+1} j(j-1)\cdots(j-i+1)) \\ &= \sum_{i=0}^{\infty} \frac{t^i}{i!} \mu'_{[i]}. \end{aligned}$$

1.1.4 Functionals of the CDF; Distribution "Measures"

Functionals are functions whose arguments are functions. The value of a functional may be any kind of object, a real number or another function, for example. The domain of a functional is a set of functions. I will use notation of the following form: for the functional, a capital Greek or Latin letter, Υ, M, etc.; for the domain, a calligraphic Latin letter, $\mathcal{G}$, $\mathcal{P}$, etc.; for a function, an italic letter, g, G, P, etc.; and for the value, the usual notation for functions, $\Upsilon(P)$ where $P \in \mathcal{P}$, for example.

Parameters of distributions as well as other interesting characteristics of distributions can often be defined in terms of functionals of the CDF. For example, the mean of a distribution, if it exists, may be written as the functional M of the CDF P:

$$M(P) = \int y \, \mathrm{d}P(y). \tag{1.10}$$

Viewing this as a Riemann–Stieltjes integral, for a discrete distribution, it reduces to a sum of the mass points times their associated probabilities. A functional operating on a CDF is called a *statistical functional* or *statistical function*. I will refer to the values of such functionals as *distributional measures*. (Although the distinction is not important, "M" in equation (1.10) a capital Greek letter mu. I usually—but not always—will use upper-case Greek letters to denote functionals, especially functionals of CDFs and in those cases, I usually will use the corresponding lower-case letters to represent the measures defined by the functionals.)

Linear functionals are often of interest. The functional M in equation (1.10), for example, is linear over the distribution function space of CDFs for which the integral exists.

It is important to recognize that a given functional may not exist at a given CDF. For example, if

$$P(y) = 1/2 + \tan^{-1}((y-\alpha)/\beta))/\pi \tag{1.11}$$

(that is, the distribution is Cauchy), then $M(P)$ does not exist. Without always using a phrase about existence, when I write an expression such as $M(P)$ or $\Upsilon(P)$ I generally imply the existence of the functional for the given P.

Also, for some parametric distributions, such as the family of beta distributions, there may not be a simple functional yields the parameter.

A functional of a CDF is generally a function of any parameters associated with the distribution. For example, if μ and σ are parameters of a distribution with CDF $P(y\,;\,\mu,\sigma)$ and Υ is some functional, we have

$$\Upsilon(P(y\,;\,\mu,\sigma)) = f(\mu,\sigma),$$

for some function f. If, for example, the M in equation (1.10) above is Υ and the P is the normal CDF $P(y\,;\,\mu,\sigma)$, then $\Upsilon(P(y\,;\,\mu,\sigma)) = \mu$.

Moments

For a univariate distribution with CDF P, if we denote the mean or the first moment of by μ, then, for $r \geq 2$, we define the r^{th} *central moment*, if it exists, as

$$\begin{aligned} \mu_r &= M_r(P) \\ &= \int (y-\mu)^r \mathrm{d}P(y). \end{aligned} \tag{1.12}$$

For a discrete distribution, this expression can be interpreted as a sum of the values at the mass points times their associated probabilities. If the μ in equation (1.12) is omitted, the corresponding moment is called "raw". (An additional comment on notation: Although the question of existence of moments is important, whenever I speak of a moment without further qualification, I will assume it exists.)

We define the r^{th} *standardized moment* as

$$\eta_r = \mu_r / \mu_2^{r/2}. \tag{1.13}$$

The first raw moment or the *mean*, is an indicator of the general "location" of the distribution. The second central moment or the *variance*, denoted as μ_2 or σ^2 is a measure of the "spread" of the distribution. The nonnegative square root, σ, is sometimes called the "scale" of the distribution. The third standardized moment, η_3, is an indicator of whether the distribution is skewed; it is called the *skewness coefficient*. If $\eta_3 \neq 0$, the distribution is asymmetric.

The fourth standardized moment, η_4 is called the *kurtosis coefficient*. It is an indicator of how "peaked" the distribution is, or how heavy the tails of the distribution are. (Actually, exactly what this standardized moment measures cannot be described simply. Because, for the random variable Y, we have

$$\eta_4 = \mathrm{V}\left(\frac{(Y-\mu)^2}{\sigma^2}\right) + 1,$$

it can be seen that the minimum value for η_4 is attained for a discrete distribution with mass points $-\sigma$ and σ. We might therefore interpret η_4 as a measure of variation about the two points $-\sigma$ and σ. This, of course, leads to the two characteristics mentioned above: peakedness and heaviness in the tails.)

For multivariate distributions, the first moment, which is a vector, is defined by equation (1.10), where the integral is taken over all components of the vector y. For $r = 2$, the matrix of joint moments is given by a simple extension of equation (1.12):

$$\Sigma(P) = \int (y-\mu)(y-\mu)^{\mathrm{T}} \, \mathrm{d}P(y).$$

Any of the marginal moments have a single component in the integrand but with the integration performed over all components. For multivariate distributions, the higher-order marginal moments are generally more useful than the higher-order joint moments.

Quantiles

Another useful distributional measure for describing a univariate distribution with CDF P is is the quantity y_π, such that

$$\Pr(Y \leq y_\pi) \geq \pi, \text{ and } \Pr(Y \geq y_\pi) \geq 1-\pi, \tag{1.14}$$

for $\pi \in (0,1)$. In this expression, y_π may not be unique. We define a similar quantity to be the unique π *quantile* as

$$\Xi_\pi(P) = \inf_y \{y, \text{ s.t. } P(y) \geq \pi\}. \tag{1.15}$$

For an absolutely continuous distribution, this is very simple:

$$\Xi_\pi(P) = P^{-1}(\pi). \tag{1.16}$$

I often use this notation for a quantile even when P^{-1} does not exist in a formal sense. The 0.5 quantile is an important one; it is called the *median*. For the Cauchy distribution, for example, the moment functionals do not exist, but the median does. An important functional for the Cauchy distribution is, therefore, $\Xi_{0.5}(P)$ because that is the location of the "middle" of the distribution.

For multivariate distributions, quantiles generalize to level curves or contours of the CDF. They are obviously much more complicated, and hence, less useful, than quantiles in a univariate distribution. The quantiles of the marginal univariate distributions may be of interest, however.

Quantiles can be used for measures of scale and of characteristics of the shape of a distribution. A measure of the scale of a distribution, for example, is the *interquartile range*:

$$\Xi_{0.75} - \Xi_{0.25}. \tag{1.17}$$

Various measures of skewness can be defined as

$$\frac{(\Xi_{1-\pi} - \Xi_{0.5}) - (\Xi_{0.5} - \Xi_{\pi})}{\Xi_{1-\pi} - \Xi_{\pi}}, \tag{1.18}$$

for $0 < \pi < 0.5$. For $\pi = 0.25$, this is called the *quartile skewness* or the *Bowley coefficient*. For $\pi = 0.125$, it is called the *octile skewness*. These can be especially useful with the measures based on moments do not exist. The extent of the peakedness and tail weight can be indicated by the ratio of interquantile ranges:

$$\frac{\Xi_{1-\pi_1} - \Xi_{\pi_1}}{\Xi_{1-\pi_2} - \Xi_{\pi_2}}. \tag{1.19}$$

These measures can be more useful than the kurtosis coefficient based on the fourth moment, because different choices of π_1 and π_2 emphasize different aspects of the distribution. In expression (1.19), $\pi_1 = 0.025$ and $\pi_2 = 0.125$ yield a good measure of tail weight, and $\pi_1 = 0.125$ and $\pi_2 = 0.25$ in expression (1.19) yield a good measure of peakedness.

L Functionals

Various modifications of the mean functional M in equation (1.10) are often useful, especially in robust statistics. A functional of the form

$$L_J(P) = \int yJ(y)\,\mathrm{d}P(y), \tag{1.20}$$

for some given function J, is called an L *functional*. If $J \equiv 1$, this is the mean functional. Often J is defined as a function of $P(y)$. A "trimmed mean", for example, is defined by an L functional with $J(y) = (\beta - \alpha)^{-1}\mathrm{I}_{(\alpha,\beta)}(P(y))$, for constants $0 \leq \alpha < \beta \leq 1$ and where I is the indicator function.

In this case, the L functional is often denoted as $T_{\alpha,\beta}$. Often β is taken to be $1 - \alpha$, so the trimming is symmetric in probability content.

M Functionals

Another family of functionals that generalize the mean functional are defined as a solution to the minimization problem

$$\int \rho(y, M_\rho(P))\,\mathrm{d}P(y) = \min_{\theta\in\Theta} \int \rho(y,\theta)\,\mathrm{d}P(y), \tag{1.21}$$

for some function ρ and where Θ is some open subset of $\mathbb{R}$. A functional defined as the solution to this optimization problem is called an M *functional*. (Note the similarity in names and notation: we call the M in equation (1.10) the mean functional; and we call the M_ρ in equation (1.21) the M functional.)

Two related functions that play important roles in the analysis of M functionals are

$$\psi(y,t) = \frac{\partial \rho(y,t)}{\partial t}, \tag{1.22}$$

and

$$\lambda_P(t) = \int \psi(y,t) \mathrm{d}P(y) = \frac{\partial}{\partial t} \int \rho(y,t) \mathrm{d}P(y) \tag{1.23}$$

If y is a scalar and $\rho(y,\theta) = (y-\theta)^2$ then $M_\rho(P)$ is the mean functional from equation (1.10). Other common functionals also yield solutions to the optimization problem (1.21); for example, for $\rho(y,\theta) = |y-\theta|$, $\Xi_{0.5}(P)$ from equation (1.15) is an M functional (possibly nonunique).

We often choose the ρ in an M functional to be a function of $y-\theta$, and to be convex and differentiable. In this case, the M functional is the solution to

$$\mathrm{E}(\psi(Y-\theta)) = 0, \tag{1.24}$$

where

$$\psi(y-\theta) = \mathrm{d}\rho(y-\theta)/\mathrm{d}\theta,$$

if that solution is in the interior of Θ.

1.1.5 Transformations of Random Variables

We often need to determine the distribution of some transformation of a given random variable or a set of random variables. In the simplest case, we have a random variable X, which may be a vector, with known distribution and we want to determine the distribution of $Y = h(X)$, where h is a full-rank transformation; that is, there is a function h^{-1} such that $X = h^{-1}(Y)$. In other cases, the function may not be full-rank, for example, X may be an n-vector, and $Y = \sum_{i=1}^n X_i$. There are some general approaches to the problem. Sometimes one method works best, and other times some other method works best.

method of CDFs

Given X with known CDF F_X and $Y = h(X)$ as above, we can write the CDF F_Y of Y as

$$\begin{aligned} F_Y(y) &= \Pr(Y \le y) \\ &= \Pr(h(X) \le y) \\ &= \Pr(X \le h^{-1}(y)) \\ &= F_X(h^{-1}(y)). \end{aligned}$$

method of change of variables

If X has density $p_X(x|\theta)$ and $Y = h(X)$, where h is a full-rank transformation (that is, there is a function h^{-1} such that $X = h^{-1}(Y)$), then

the density of Y is $p_Y(y|\theta) = p_X(h^{-1}(y)|\theta)|\mathrm{J}_{h^{-1}}(y)|$, where $\mathrm{J}_{h^{-1}}(y)$ is the Jacobian of the inverse transformation, and $|\cdot|$ is the determinant.

Why the inverse transformation? Think of the density as a differential; that is, it has a factor $\mathrm{d}x$, so in the density for Y, we want a factor $\mathrm{d}y$. Under pressure you may forget exactly how this goes, or want a quick confirmation of the transformation. You should be able to construct a simple example quickly. An easy one is the right-triangular distribution; that is, the distribution with density $p_X(x) = 2x$, for $0 < x < 1$. Let $y = 2x$, so $x = \frac{1}{2}y$. Sketch the density of Y, and think of what transformations are necessary to get the expression $p_Y(y) = \frac{1}{2}y$, for $0 < y < 2$.

Constant linear transformations are particularly simple. If X is an n-vector random variable with PDF f_X and A is an $n \times n$ constant matrix of full rank, the PDF of $Y = AX$ is $f_X|\det(A^{-1})|$.

In the change of variable method, we think of h a mapping of the range $\mathcal{X}$ of the random variable X to the range $\mathcal{Y}$ of the random variable Y, and the method works by expressing the probability content of small regions in $\mathcal{Y}$ in terms of the probability content of the pre-image of those regions in $\mathcal{X}$.

If the transformation is not one-to-one, we generally try to modify the method by identifying subregions in which there are one-to=one transformations.

convolutions

A simple application of the change of variables method is in the common situation of finding the distribution of the sum of two scalar random variables that are independent but not necessarily identically distributed.

Suppose X is a random variable with PDF f_X and Y is a random variable with PDF f_Y, and we want the density of $U = X + Y$. We form another variable $V = Y$ and the matrix

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$

so that we have a full-rank transformation, $(U, V)^{\mathrm{T}} = A(X, Y)^{\mathrm{T}}$ The inverse of the transformation matrix is

$$A^{-1} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix},$$

and the Jacobian is 1. Because X and Y are independent, their joint PDF is $f_{XY}(x, y) = f_X(x)f_Y(y)$, and the joint PDF of U and V is $f_{UV}(u, v) = f_X(u - v)f_Y(v)$; hence, the PDF of U is

$$f_U(u) = \int_{-\infty}^{\infty} f_X(u - v)f_Y(v)\mathrm{d}v.$$

We call f_U the *convolution* of f_X and f_Y. This form occurs often in applied mathematics.

method of MGFs or CFs

In this method, we write the MGF of Y as $\mathrm{E}(\mathrm{e}^{tY}) = \mathrm{E}(\mathrm{e}^{th(x)})$, or we write the CF in a similar way. If we can work out the expectation (with respect to the known distribution of X, we have the MGF or CF of Y, which determines its distribution.

The MGF or CF technique is particularly useful in the case when Y is the sum from a simple random sample.

1.1.6 Order Statistics

In a random sample of size n from a distribution with PDF f and CDF F given the i^{th} order statistic $X_{(i)} = a$, the $i-1$ random variables less than $X_{(i)}$ are i.i.d. as

$$f(x)/F(a);$$

the $n-i$ random variables greater than $X_{(i)}$ are i.i.d. as

$$f(x)/(1-F(a)).$$

Order statistics are not i.i.d.

The joint density of all order statistics is

$$n! \prod f(x_{(i)}) \mathrm{I}_{x_{(1)} \leq \cdots \leq x_{(n)}}(x_{(1)}, \ldots, x_{(n)})$$

The joint density of the i^{th} and j^{th} $(i < j)$ order statistics is

$$\frac{n!}{(i-1)!(j-i-1)!(n-j)!} \cdot$$

$$\Big(F(x_{(i)})\Big)^{i-1} \Big(F(x_{(j)}) - F(x_{(i)})\Big)^{j-i-1} \Big(1 - F(x_{(j)})\Big)^{n-j} f(x_{(i)}) f(x_{(j)}).$$

Understand the heuristics that lead to these formulas.

1.1.7 Useful Inequalities Involving Random Variables and Probability Distributions

Inequalities involving functions of events and random variables are important throughout the field of probability and statistics. Two important uses are for showing that one procedure is better than another and for showing that some sequence converges to a given object (a constant, a function, or a set).

In the following, for simplicity, we will assume $X \in \mathbb{R}$.

A simple, but surprisingly useful inequality states that if $\mathrm{E}(X^2) < \infty$, then the variance is the minimum expectation of the form $\mathrm{E}((X-c)^2)$ for any constant c. In other words, the minimum of $\mathrm{E}((X-c)^2)$ occurs at $c = \mathrm{E}(X)$). We see this by writing

$$\begin{aligned}\mathrm{E}((X-c)^2) &= \mathrm{E}((X-\mathrm{E}(X)+\mathrm{E}(X)-c)^2)\\ &= \mathrm{E}((X-\mathrm{E}(X))^2)+\mathrm{E}((\mathrm{E}(X)-c)^2)+\\ &\qquad 2\mathrm{E}((X-\mathrm{E}(X))(\mathrm{E}(X)-c))\\ &= \mathrm{E}((X-\mathrm{E}(X))^2)+\mathrm{E}((\mathrm{E}(X)-c)^2)\\ &\geq \mathrm{E}((X-\mathrm{E}(X))^2). \end{aligned} \tag{1.25}$$

We will use various inequalities often in the subsequent sections and chapters, so we collect a number of them in this section, where we have categorized them into four types depending of the kinds of expressions in the inequalities. These four types involve relations between

- $\Pr(X \in A)$ and $\mathrm{E}(f(X))$, e.g., Chebyshev
- $\mathrm{E}(f(X))$ and $f(\mathrm{E}(X))$, e.g., Jensen's
- $\mathrm{E}(f_1(X,Y))$ and $\mathrm{E}(f_2(X))$ and $\mathrm{E}(f_3(Y))$, e.g., covariance, Cauchy-Schwarz, information
- $\mathrm{V}(Y)$ and $\mathrm{V}\big(\mathrm{E}(Y|X)\big)$, e.g., Rao-Blackwell

Any special case of these involves an appropriate definition of A or f (e.g., nonnegative, convex, etc.)

A more general case of the inequalities is to replace distributions, and hence expected values, by conditioning on a sub-σ-field, $\mathcal{A}$.

For each type of inequality there is essentially a straightforward method of proof, which is important to know.

Some of these inequalities involve absolute values of the random variable. To work with these inequalities, it is useful to recall the triangle inequality for the absolute value of real numbers:

$$|x+y| \leq |x|+|y|. \tag{1.26}$$

We can prove this merely by considering all four cases for the signs of x and y.

This inequality generalizes immediately to $|\sum x_i| \leq \sum |x_i|$.

Expectations of absolute values of functions of random variables are functions of norms. (A *norm* is a function of x that (1) is positive unless $x = 0$ a.e., that (2) is equivariant to scalar multiplication, and that (3) satisfies the triangle inequality.) The important form $(\mathrm{E}(|X|^p))^{1/p}$, for $1 \leq p$ is an L_p norm, $\|X\|_p$. Some of the inequalities given below involving expectations of absolute values of random variables are essentially triangle inequalities and their truth establishes the expectation as a norm.

Some of the expectations discussed below are recognizable as familiar norms over vector spaces. For example, the expectation in Minkowski's inequality is essentially the L_p norm of a vector, which is defined for an n-vector x in a finite-dimensional vector space as $\|x\|_p \equiv (\sum |x_i|^p)^{1/p}$. Minkowski's inequality in this case is $\|x+y\|_p \leq \|x\|_p + \|y\|_p$. For $p = 1$, this is the triangle inequality for absolute values given above.

Inequalities Involving $\Pr(X \in A)$ and $\mathrm{E}(f(X))$

An important class of inequalities bound tail probabilities of a random variable, that is, limit the probability that the random variable will take on a value beyond some distance from the expected value of the random variable.

The important general form involving $\Pr(X \in A)$ and $\mathrm{E}(f(X))$ is Markov's inequality. Several others are special cases of it.

- **Markov's inequality**
  For $\epsilon > 0$, $k > 0$, and r.v. $X \ni \mathrm{E}(|X|^k)$ exists,

$$\Pr(|X| \geq \epsilon) \leq \frac{1}{\epsilon^k}\mathrm{E}(|X|^k) \tag{1.27}$$

  Proof. For a nonnegative random variable Y,

$$\mathrm{E}(Y) \geq \int_{y \geq \epsilon} y \, \mathrm{d}P(y) \geq \epsilon \int_{y \geq \epsilon} \mathrm{d}P(y) = \epsilon \Pr(Y \geq \epsilon).$$

  Now let $Y = |X|^k$. ∎
- **Chebyshev's inequality**
  For $\epsilon > 0$,

$$\Pr(|X - \mathrm{E}(X)| \geq \epsilon) \leq \frac{1}{\epsilon^2}\mathrm{V}(X) \tag{1.28}$$

  Proof. In Markov's inequality, let $k = 2$, and replace X by $X - \mathrm{E}(X)$. ∎
- **Chebyshev's inequality (another form)**
  For $f \ni f(x) \geq 0$ and $\epsilon > 0$,

$$\Pr(f(X) \geq \epsilon) \leq \frac{1}{\epsilon}\mathrm{E}(f(X)) \tag{1.29}$$

  Proof. Same as Markov's inequality; start with $\mathrm{E}(f(X))$. ∎
  Chebyshev's inequality is often useful for $\epsilon = \sqrt{\mathrm{V}(X)}$. There are also versions of Chebyshev's inequality for specific families of distributions.
  - **3σ rule for a unimodal random variable**
    If X is a random variable with a unimodal absolutely continuous distribution, and $\sigma = \sqrt{\mathrm{V}(X)}$, then

$$\Pr(|X - \mathrm{E}(X)| \geq 3\sigma) \leq \frac{4}{81}. \tag{1.30}$$

    See Dharmadhikari and Joag-Dev (1988).
  - **Normal tail probability**
    If $X \sim \mathrm{N}(\mu, \sigma^2)$, then

$$\Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{3k^2}. \tag{1.31}$$

    See DasGupta (2000).

- **Kolmogorov's inequality**
  This inequality is a generalization of Chebyshev's inequality that applies to finite partial sums of a sequence of independent random variables $X_1, X_2, \ldots$ over a common probability space such that for each, $\mathrm{E}(X_i) = 0$ and $\mathrm{E}(X_i^2) < \infty$. (The common probability space means that $\mathrm{E}(\cdot)$ has exactly the same meaning for each i.)
  For such a sequence, let $S_k = \sum_{i=1}^{k} X_i$). Then for any positive integer n and any $\epsilon > 0$,
  $$\Pr\left(\max_{1\le k\le n} |S_k| \ge \epsilon\right) \le \frac{1}{\epsilon^2}\mathrm{V}(S_n). \tag{1.32}$$
  This is a special case of Doob's submartingale inequality. It is also a special case of the Hájek-Rènyi inequality, which I state without proof:
- **The Hájek-Rènyi inequality**
  Let $X_1, X_2, \ldots$ be a sequence of independent random variables over a common probability space such that for each $\mathrm{E}(X_i^2) < \infty$. Then for any positive integer n and any $\epsilon > 0$,
  $$\Pr\left(\max_{1\le k\le n} c_k \left|\sum_{i=1}^{k}(X_i - \mathrm{E}(X_i)\right| \ge \epsilon\right) \le \frac{1}{\epsilon^2}\sum_{i=1}^{n} c_i^2\mathrm{V}(X_i),$$
  where $c_1 \ge \cdots \ge c_n > 0$ are constants.

Inequalities Involving $\mathrm{E}(f(X))$ and $f(\mathrm{E}(X))$

- **Jensen's inequality**
  For f a convex function over the support of the r.v. X (and all expectations shown exist),
  $$f(\mathrm{E}(X)) \le \mathrm{E}(f(X)). \tag{1.33}$$
  Proof. By the definition of convexity, f convex over $D \Rightarrow \exists\, c \ni \forall\, t \in D \ni c(x-t)+f(t) \le f(x)$. (Notice that $L(x) = c(x-t)+f(t)$ is a straight line through the point $(t, f(t))$. By the definition of convexity, f is convex if its value at the weighted average of two points does not exceed the weighted average of the function at those two points.) Now, given this, let $t = \mathrm{E}(X)$ and take expectations of both sides of the inequality. ∎
  If f is strictly convex, it is clear
  $$f(\mathrm{E}(X)) < \mathrm{E}(f(X)) \tag{1.34}$$
  unless $f(X) = \mathrm{E}(f(X))$ with probability 1.
  For a concave function, the inequality is reversed. (The negative of a concave function is convex.)
  Some simple examples for a nonconstant positive random variable X:

– Monomials of even power: for $k = 2, 4, 6, \ldots$,

$$\mathrm{E}(X)^k \leq \mathrm{E}(X^k).$$

This inequality implies the familiar fact that $\mathrm{E}(X) \geq 0$.

– Reciprocals:

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

– Logs:

$$\mathrm{E}(\log(X)) \leq \log(\mathrm{E}(X)).$$

The canonical picture is that of a quadratic function of a uniform random variable:

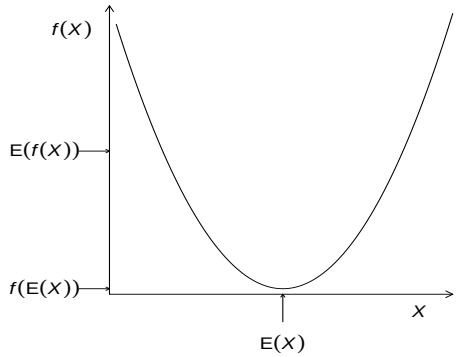


Fig. 1.1. Jensen's Inequality

Some other consequences of Jensen's inequality:

– Nonnegativity of the entropy distance (Kullback-Leibler information): If f and g are probability densities, $\mathrm{E}_f(\log(f(X)/g(X)))$, is the *entropy distance* between f and g with respect to g. It is also called the *Kullback-Leibler information* or *Kullback-Leibler distance*. It is nonnegative:

$$\mathrm{E}_f(\log(f(X)/g(X))) \geq 0. \tag{1.35}$$

Proof.

$$\begin{aligned}\mathrm{E}_f(\log(f(X)/g(X))) &= -\mathrm{E}_f(\log(g(X)/f(X))) \\ &\geq -\log(\mathrm{E}_f(g(X)/f(X))) \\ &= 0.\end{aligned}$$

A related fact applies to any nonnegative integrable functions f and g on a measure space with a σ-finite measure ν, for which $\int f\mathrm{d}\nu \geq \int g\mathrm{d}\nu > 0$:

$$\int f(\log(f/g))\mathrm{d}\nu \geq 0. \tag{1.36}$$

This can be proved as above by normalizing the functions, thus forming densities.

– An inequality important for showing the convergence of the EM algorithm:

$$\mathrm{E}_f(\log(f(X))) \geq \mathrm{E}_f(\log(g(X))), \tag{1.37}$$

for the PDFs f and g. This inequality is also sometimes called the "information inequality" (but see (1.45)). We see this by use of the entropy distance.

The strict form of Jensen's inequality (1.34) also applies to the consequent inequalities. For example, we have for the PDFs f and g,

$$\mathrm{E}_f(\log(f(X))) = \mathrm{E}_f(\log(g(X))) \Leftrightarrow f(X) = g(X) \text{ a.s.} \tag{1.38}$$

Proof(of $\Rightarrow$):
By the equality of $\mathrm{E}_f(\log(f(X)))$ and $\mathrm{E}_f(\log(g(X)))$ we have

$$\int_{\{f>0\}} g(x)\mathrm{d}x = 1,$$

and so for any A,

$$\begin{aligned}\int_A g(x)\mathrm{d}x &= \int_{A\cap\{f>0\}} g(x)\mathrm{d}x \\ &= \mathrm{E}_f\left(g(X)/f(X)|X \in A\cap\{f>0\}\right) \\ &= \Pr\left(X \in A\cap\{f>0\}\right) \\ &= \int_A f(x)\mathrm{d}x,\end{aligned}$$

hence $f(X) = g(X)$ a.s. ∎

Inequalities Involving $\mathbf{E}(f(X,Y))$ and $\mathbf{E}(g(X))$ and $\mathbf{E}(h(Y))$

In many of the inequalities in this section, the functions f, g, and h are norms. The inequalities hold for general L_p norms, and although we will consider the inequality relationship between expected values, similar inequalities often for real numbers, vectors, or random variables.

The inequalities are basically of two types:

- Hölder: $\mathrm{E}(|XY|) \leq \left(\mathrm{E}(|X|^p)\right)^{1/p}\left(\mathrm{E}(|Y|^q)\right)^{1/q}$

- Minkowski: $(\mathrm{E}(|X+Y|^p))^{1/p} \leq (\mathrm{E}(|X|^p))^{1/p} + (\mathrm{E}(|Y|^p))^{1/p}$

Hölder inequality is somewhat more basic, in that it is used in the proof of Minkowski's inequality.

Note that Minkowski's inequality has an interesting consequence: it means that $(\mathrm{E}(|\cdot|^p))^{1/p}$ is a norm.

Several other inequalities are special cases of these two.

In some inequalities in this section, the functions are second-degree monomials. The basic special inequality of this form is the Cauchy-Schwartz inequality, which then leads to one of the most important inequalities in applications in statistics, the covariance inequality. The covariance inequality, in turn, leads to fundamental bounds on the variances of estimators.

- **Hölder's inequality** (a general inequality relating $\mathrm{E}(f(X,Y))$ to $\mathrm{E}(g(X))$ and $\mathrm{E}(h(Y))$
  For $p, q > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$ (and all expectations shown exist),

$$\mathrm{E}(|XY|) \leq \Big(\mathrm{E}(|X|^p)\Big)^{1/p} \Big(\mathrm{E}(|Y|^q)\Big)^{1/q} \tag{1.39}$$

  p and q as in this inequality are called *dual* indices. Note that $q = p/(p-1)$.
  Proof. If $\mathrm{E}(|X|^p) = 0$ or $\mathrm{E}(|Y|^q) = 0$, then true because both sides $= 0$ wp1. Hence, assume both > 0.
  Now, for p and q as in hypothesis, $\forall a, b > 0,\ \exists\, s, t \ni\ a = \mathrm{e}^{s/p}$ and $b = \mathrm{e}^{t/q}$. Now e^x is convex, so $\mathrm{e}^{s/p+t/q} \leq \frac{1}{p}\mathrm{e}^s + \frac{1}{q}\mathrm{e}^t$, or $ab \leq a^p/p + b^q/q$.
  Now let

$$a = \left| \frac{X(\omega)}{\Big(\mathrm{E}(|X|^p)\Big)^{1/p}} \right| \quad \text{and} \quad b = \left| \frac{Y(\omega)}{\Big(\mathrm{E}(|Y|^q)\Big)^{1/q}} \right|$$

  and so

$$|X(\omega)Y(\omega)| \leq \Big(\mathrm{E}(|X|^p)\Big)^{1/p} \Big(\mathrm{E}(|Y|^q)\Big)^{1/q} \left(\frac{|X(\omega)|^p}{\mathrm{E}(|X|^p)} \frac{1}{p} + \frac{|Y(\omega)|^q}{\mathrm{E}(|Y|^q)} \frac{1}{q} \right).$$

  Now take expectations. (The notation $X(\omega), Y(\omega)$ is meant to emphasize how to take expectation of XY.) ∎
  We note a special case by letting $Y \equiv 1$:

$$\mathrm{E}(|X|) \leq \Big(\mathrm{E}(|X|^p)\Big)^{1/p},$$

  and with $p = 2$, we have a special case of the Cauchy-Schwarz inequality:

$$\mathrm{E}(|X|) \leq \Big(\mathrm{E}(X^2)\Big)^{1/2}.$$

  Other inequalities that derive from Hölder's inequality are the following.

– **Liapounov's inequality**
If $1 \leq r \leq s$

$$(\mathrm{E}(|X|^r))^{1/r} \leq (\mathrm{E}(|X|^s))^{1/s} \tag{1.40}$$

Proof. First, we observe this is true for $r = s$, and for $r = 1$ (in which it is a form of Jensen's inequality). If $1 < r < s$, replace $|X|$ in the special case of Hölder's inequality above with $|X|^r$, and let $s = pr$ for $1 < p$. This yields $(\mathrm{E}(|X|^r))^{1/r} \leq (\mathrm{E}(|X|^s))^{1/s}$. ∎

– **Schwarz inequality, or Cauchy-Schwarz inequality**

$$\mathrm{E}(|XY|) \leq \Big(\mathrm{E}(X^2)\mathrm{E}(Y^2)\Big)^{1/2} \tag{1.41}$$

Proof. Let $p = q = 2$ in Hölder's inequality. ∎
Another proof: For nonnegative r.v. X and Y and all t (real), $\mathrm{E}((tX + Y)^2) = t^2\mathrm{E}(X^2) + 2t\mathrm{E}(XY) + \mathrm{E}(Y^2) \geq 0$. Hence the discriminant of the quadratic formula ≤ 0. Now, for any r.v., take absolute value. ∎

– **Covariance inequality**
If the second moments of X and Y are finite, then

$$\Big(\mathrm{E}\big((X - \mathrm{E}(X))(Y - \mathrm{E}(Y))\big)\Big)^2 \leq \mathrm{E}\big((X - \mathrm{E}(X))^2\big)\,\mathrm{E}\big((Y - \mathrm{E}(Y))^2\big) \tag{1.42}$$

or

$$\big(\mathrm{Cov}(X, Y)\big)^2 \leq \mathrm{V}(X)\,\mathrm{V}(Y) \tag{1.43}$$

Notice that the covariance inequality is essentially the same as the Cauchy-Schwarz inequality.

The covariance inequality leads to useful lower bounds on the variances of estimators. These are of two types. One type includes the Hammersley-Chapman-Robbins inequality and its extension, the Kshirsagar inequality. The other type, which is based on Fisher information, requires some "regularity conditions".

– **Hammersley-Chapman-Robbins inequality**
Let X be a random variable in $\mathbb{R}^d$ with PDF $p(x;\theta)$ and let $\mathrm{E}_\theta(T(X) = g(\theta)$. Let μ be a fixed measure on $\mathcal{X} \subset \mathbb{R}^d$ such that $p(x;\theta) \ll \mu$. Now define $S(\theta)$ such that

$$\begin{aligned} p(x;\theta) &> 0 \text{ a.e. } x \in S(\theta) \\ p(x;\theta) &= 0 \text{ a.e. } x \notin S(\theta). \end{aligned}$$

Then

$$\mathrm{V}(T(X)) \geq \sup_{t \ni S(\theta) \supset S(\theta+t)} \frac{\big(g(\theta + t) - g(\theta)\big)^2}{\mathrm{E}_\theta\left(\left(\frac{p(X;\theta+t)}{p(X;\theta)}\right)^2\right)}. \tag{1.44}$$

This inequality follows from the covariance inequality, by first considering the case for an arbitrary t such that $g(\theta + t) \neq g(\theta)$. In that case
****

– **Kshirsagar inequality**

****

– **Information inequality**
Subject to some "regularity conditions" (see Section 1.7.3), if X has PDF $p(x;\theta)$,

$$\mathrm{V}(T(X)) \geq \frac{\left(\frac{\partial \mathrm{E}(T(X))}{\partial \theta}\right)^2}{\mathrm{E}_\theta\left(\left(\frac{\partial \log p(X;\theta)}{\partial \theta}\right)^2\right)} \tag{1.45}$$

The denominator of the quantity on the right side of the inequality is called the Fisher information, or just the information. Notice the similarity of this inequality to the Hammersley-Chapman-Robbins inequality, although the information inequality requires more conditions. Under the regularity conditions, which basically allow the interchange of integration and differentiation, the information inequality follows immediately from the covariance inequality.
In Section 4.1 we consider the multivariate form of this inequality. Our main interest will be in its application in unbiased estimation. If $T(X)$ is an unbiased estimator of a differentiable function $g(\theta)$, the right side of the inequality together with derivatives of $g(\theta)$ forms the Cramér-Rao lower bound and the Bhattacharyya lower bound.

- **Minkowski's inequality**
This is a triangle inequality for L_p norms and related functions.
For $1 \leq p$,

$$(\mathrm{E}(|X+Y|^p))^{1/p} \leq (\mathrm{E}(|X|^p))^{1/p} + (\mathrm{E}(|Y|^p))^{1/p} \tag{1.46}$$

Proof. Proof: First, observe the truth for $p = 1$ using the triangle inequality for the absolute value, $|x+y| \leq |x| + |y|$, giving $\mathrm{E}(|X+Y|) \leq \mathrm{E}(|X|) + \mathrm{E}(|Y|)$.
Now assume $p > 1$. Now,

$$\begin{aligned}\mathrm{E}(|X+Y|^p) &= \mathrm{E}(|X+Y||X+Y|^{p-1}) \\ &\leq \mathrm{E}(|X||X+Y|^{p-1}) + \mathrm{E}(|Y||X+Y|^{p-1}),\end{aligned}$$

where the inequality comes from the triangle inequality for absolute values. From Hölder's inequality on the terms above with $q = p/(p-1)$, we have

$$\mathrm{E}(|X+Y|^p) \leq (\mathrm{E}(|X|^p))^{1/p}(\mathrm{E}(|X+Y|^p))^{1/q} + \mathrm{E}(|Y|^p))^{1/p}(\mathrm{E}(|X+Y|^p))^{1/q}.$$

Now, if $\mathrm{E}(|X+Y|^p) = 0$, Minkowski's inequality holds. On the other hand, if $\mathrm{E}(|X+Y|^p) \neq 0$, it is positive, and so divide through by $(\mathrm{E}(|X+Y|^p))^{1/q}$, recalling again that $q = p/(p-1)$. ∎
Minkowski's inequality is a special case of two slightly tighter inequalities; one for $p \in [1,2]$ due to Esseen and von Bahr (1965), and one for $p \geq 2$ due to Marcinkiewicz and Zygmund (1937).

An inequality that derives from Minkowski's inequality, but which applies directly to real numbers or random variables, is the following.

– For $0 \leq p$,

$$|X+Y|^p \leq 2^p(|X|^p + |Y|^p) \tag{1.47}$$

This is true because $\forall\omega \in \Omega$, $\|X(\omega)+Y(\omega)\| \leq 2\max\{\|X(\omega)\|, \|Y(\omega)\|\}$, and so

$$\begin{aligned}\|X(\omega)+Y(\omega)\|^p &\leq \max\{2^p\|X(\omega)\|^p, 2^p\|Y(\omega)\|^p\} \\ &\leq 2^p\|X(\omega)\|^p + 2^p\|Y(\omega)\|^p.\end{aligned}$$

Inequalities Involving $\mathrm{V}(Y)$ and $\mathrm{V}\big(\mathrm{E}(Y|X)\big)$

- **Rao-Blackwell inequality**

$$\mathrm{V}\big(\mathrm{E}(Y|X)\big) \leq \mathrm{V}(Y) \tag{1.48}$$

This follows from the equality $\mathrm{V}(Y) = \mathrm{V}\big(\mathrm{E}(Y|X)\big) + \mathrm{E}\big(\mathrm{V}(Y|X)\big)$.

Multivariate Extensions

There are multivariate extensions of most of these inequalities. In some cases, the multivariate extensions apply to the minimum or maximum element of a vector.

Some inequalities involving simple inequalities are extended by conditions on vector norms, and the ones involving variances are usually extended by positive (or semi-positive) definiteness of the difference of two variance-covariance matrices.

1.2 Sequences of Events and of Random Variables

Countably infinite sequences play the main role in the definition of the basic concept of a σ-field, and consequently, in the development of a theory of probability. Sequences of sets correspond to sequences of events and, consequently, of sequences of random variables. Unions, intersections, and complements of sequences of sets are important for studying sequences of random variables. The material in this section depends heavily on the properties of sequences of sets discussed on page 349 and the following pages.

Two important types of sequences of probabilities are, similar to the analogous limits for sets on page 349,

$$\limsup_n \Pr(A_n) \equiv \inf_n \sup_{i \geq n} \Pr(A_i)$$

$$\liminf_n \Pr(A_n) \equiv \sup_n \inf_{i \geq n} \Pr(A_i).$$

$\limsup_n$ is often written as $\overline{\lim}_n$
$\liminf_n$ is often written as $\underline{\lim}_n$

We recall the intuitive interpretation of $\limsup_n A_n$ and $\liminf_n A_n$ (written also as A^* and A_*):
An element ω is in A^* iff for each n, there is some $i \geq n$ for which $\omega \in A_i$. This means that ω must lie in infinitely many of the A_n.
An element ω is in A_* iff there is some n such that for all $i \geq n$, $\omega \in A_i$. This means that ω must lie in all but finitely many of the A_n.

Similarly to the corresponding relationship between unions and intersections of sequences of sets, we have the relationships:

$$\Pr(\limsup_n A_n) \leq \limsup_n \Pr(A_n) \tag{1.49}$$

and

$$\Pr(\liminf_n A_n) \leq \liminf_n \Pr(A_n) \tag{1.50}$$

We see this by considering $B_n = \cup_{i=n}^{\infty} A_i$, so that $B_n \searrow \limsup_n A_n$, and likewise $C_n = \cap_{i=n}^{\infty} A_i$, so that $C_n \nearrow \liminf_n A_n$. We use the continuity of the measure to get $\Pr(A_n) \leq \Pr(B_n) \to \Pr(\limsup_n A_n)$ and $\Pr(A_n) \geq \Pr(C_n) \to \Pr(\liminf_n A_n)$.

The Borel-Cantelli Lemmas

Let A_n be a sequence of events and P be a probability measure.

- If $\sum_{n=1}^{\infty} P(A_n) < \infty$, then $P(\limsup_n A_n) = 0$.
  Proof. (First, notice that $P(\cup_{i=n}^{\infty} A_i)$ can be arbitrarily small if n is large enough.)
  From $\limsup_n A_n \subset \cup_{i=n}^{\infty} A_i$, we have

$$\begin{aligned} P(\limsup_n A_n) &\leq P(\cup_{i=n}^{\infty} A_i) \\ &\leq \sum_{i=n}^{\infty} P(A_i) \\ &\to 0 \quad \text{as } n \to \infty \quad \text{because } \sum_{n=1}^{\infty} P(A_n) < \infty. \end{aligned}$$

  ∎

- If $\sum_{n=1}^{\infty} P(A_n) = \infty$, then $P(\limsup_n A_n) = 1$.
  We can see this by a similar argument as above.

Recall a basic fact about probability (which we will discuss again from time to time):

$$\lim \Pr(A_i) \neq \Pr(\lim A_i). \tag{1.51}$$

Compare this with the fact from above:

$$\lim_{n\to\infty} \bigcup_{i=1}^{n} \left[a + \frac{1}{i},\ b - \frac{1}{i}\right] \neq \bigcup \lim_{i\to\infty} \left[a + \frac{1}{i},\ b - \frac{1}{i}\right].$$

Types of Convergence

The first important point to understand about asymptotic theory is that there are different kinds of convergence of a sequence of random variables, $\{X_n\}$.

One type of convergence applies directly to the function (the random variable). This is the strongest convergence.

One type of convergence applies to expected values of powers of the random variable. This is also a very strong convergence.

One type of convergence applies to probabilities of the random variable being within a range of another random variable.

One type of convergence applies to the distribution of the random variable. This is the weakest convergence.

Almost sure (a.s.)

We say that $\{X_n\}$ converges to X almost surely if

$$\lim_{n\to\infty} X_n = X \text{ a.s.}$$

We write

$$X_n \to_{\text{a.s.}} X.$$

The condition above can also be written as

$$\Pr(\lim_{n\to\infty} X_n = X) = 1.$$

For this reason, almost sure convergence is also called *convergence with probability 1*, and may be indicated by writing $X_n \to_{\text{wp1}} X$.

In r^{th} moment (in L_r)

For fixed $r > 0$, we say that $\{X_n\}$ converges to X in r^{th} moment if

$$\lim_{n\to\infty} \text{E}(\|X_n - X\|_r^r) = 0.$$

We write

$$X_n \to_{\text{L}_r} X.$$

For $r = 1$, this is called *convergence in mean*.

For $r = 2$, this is called *convergence in mean square*.

In probability

We say that $\{X_n\}$ converges to X in probability if for every $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr(\|X_n - X\| > \epsilon) = 0.$$

We write

$$X_n \to_{\mathrm{p}} X.$$

Notice the difference in convergence in probability and convergence with probability 1; in the former case the limit of probabilities is taken, in the latter the case a probability of a limit is evaluated.

In distribution (in law)

If $\{X_n\}$ have CDFs $\{F_n\}$ and X has CDF F, we say that $\{X_n\}$ converges to X in distribution or in law if at each point of continuity t of F,

$$\lim_{n\to\infty} F_n(t) = F(t).$$

We write

$$X_n \to_{\mathrm{d}} X.$$

If the sequence $\{X_n\}$ converges in distribution to X, we say that the sequence of CDFs $\{F_n\}$ *converges weakly* to the CDF of X, F. We write

$$F_n \to_{\mathrm{w}} F.$$

If $F_n \to_{\mathrm{w}} F$ and F is continuous in $\mathbb{R}^k$, then

$$\lim_{n\to\infty} \sup_{t\in\mathbb{R}^k} |F_n(t) - F(t)| = 0.$$

This is called "Pólya's theorem". We can establish this result by use of the δ-ϵ definition of continuity.

When a random variable converges in distribution to a distribution for which we have adopted a symbol such as $\mathrm{N}(\mu, \sigma^2)$, for example, we may use notation of the form

$$X_n \to_{\sim} \mathrm{N}(\mu, \sigma^2).$$

Because this notation only applies in this kind of situation, we often write it more simply as just

$$X_n \to \mathrm{N}(\mu, \sigma^2).$$

For certain distributions we have special symbols to represent a random variable. In such cases, we may use notation of the form

$$X_n \to_{\mathrm{d}} \chi^2_\nu,$$

which in this case indicates that the sequence $\{X_n\}$ converges in distribution to a random variable with a chi-squared distribution with ν degrees of freedom.

There are several necessary and sufficient conditions for convergence in distribution. A set of such conditions is given in the "portmanteau" theorem:

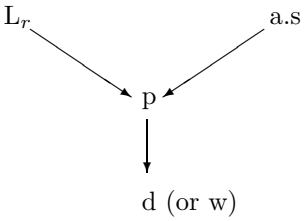
For the sequence of random variables X_n and the random variable X, all defined on a common probability space, the following are necessary and sufficient conditions that $X_n \to_{\mathrm{d}} X$.

- $\mathrm{E}(g(X_n)) \to \mathrm{E}(g(X))$ for all real bounded continuous functions g.
- $\mathrm{E}(g(X_n)) \to \mathrm{E}(g(X))$ for all real bounded Lipschitz functions g.
- $\mathrm{E}(g(X_n)) \to \mathrm{E}(g(X))$ for all real functions g such that $g(x) \to 0$ as $|x| \to \infty$.
- $\Pr(X_n \in B) \to \Pr(X \in B)$ for all Borel sets B such that $\Pr(X \in \partial B) = 0$.
- $\liminf \Pr(X_n \in S) \geq \Pr(X \in S)$ for all open sets S.
- $\limsup \Pr(X_n \in T) \leq \Pr(X \in T)$ for all closed sets T.

The proofs of the various parts of this theorem are in Billingsley (1995), among other resources.

Convergence in probability and convergence in distribution are essentially the same thing; it just depends on whether we are speaking of the sequence of random variables or the sequences of distributions of those random variables. In either case, we refer to this type of convergence as "weak convergence".

Almost sure convergence and convergence in r^{th} moment are both strong types of convergence, but they are not closely related to each other. We have the following logical relations:



(These relations are parts of "Theorem 1.8" in Shao – see proofs there.)

Just as for working with limits of unions and intersections of sets where we find it useful to identify sequences of sets that behave in some simple way (such as the intervals $[a+1/n, b-1/n]$ on page 358), it is also useful to identify sequences of random variables that behave in interesting but simple ways. A useful sequence is $\{U_n\}$, where $U_n \sim \mathrm{U}(0, 1/n)$. Other kinds of interesting sequences can be constructed as indicators of events. The events themselves may be defined by breaking a $\mathrm{U}(0,1)$ distribution into uniform distributions on partitions of $(0,1)$. For example, we for a positive integer k, we may form 2^k subintervals of $(0,1)$ for $j = 1, \ldots, 2^k$ as

$$\left(\frac{j-1}{2^k},\ \frac{j}{2^k}\right).$$

As k gets larger, the Lebesgue measure of these subintervals approaches 0 rapidly.

The following examples of sequences that show the lack of relationships between almost sure convergence and convergence in r^{th} moment come from Romano and Siegel (1986).

Let $U_n \sim \mathrm{U}(0, 1/n)$ and let $X_n = 2^n U_n$. Since $\Pr(\lim_{n\to\infty} X_n = 0) = 1$ (look at the complement of the subset of Ω for which X_n converges to 0; it is empty), $X_n \to_{\mathrm{a.s.}} 0$. However,

$$\mathrm{E}(|X_n - 0|^r) = \int_0^{1/n} (2^n)^r \mathrm{d}u = \frac{2^n r}{n}.$$

This diverges for every $r > 0$; hence $\{X_n\}$ does not converge to 0 in r^{th} moment for any r.

This example is also an example of a sequence that converges in probability (since a.s. convergence implies that), but does not converge in r^{th} moment.

Now consider an opposite case. We will use the partitioning of $(0, 1)$ referred to above. Let $U \sim \mathrm{U}(0, 1)$ and define

$$X_n = \begin{cases} 1 & \text{if } \dfrac{j_n - 1}{2^{k_n}} < U < \dfrac{j_n}{2^{k_n}} \\ 0 & \text{otherwise} \end{cases}$$

where $j_n = 1, \ldots, 2^{k_n}$ and $k_n \to \infty$ as $n \to \infty$. We see that $\mathrm{E}((X_n - 0)^2) = 1/(2^{k_n})$, hence $\{X_n\}$ converges in quadratic mean (or in mean square) to 0. We see, however, that $\lim_{n\to\infty} X_n$ does not exist (since for any value of U, X_n takes on each of the values 0 and 1 infinitely often). Therefore, $\{X_n\}$ cannot converge a.s. (to anything!).

This is another example of a sequence that converges in probability (since convergence in r^{th} moment implies that), but does not converge a.s.

Although convergence in distribution does not imply a.s. convergence, convergence in distribution does allow us to construct an a.s. convergent sequence. This is stated in Skorohod's theorem (part of Shao's "Theorem 1.8"), whose proof we will omit. (It's not hard, it's just long and complicated.)

Skorohod's Theorem: If for the random variables (vectors!) $X_1, X_2, \ldots$, we have $X_n \to_{\mathrm{p}} X$, then there exist random variables $Y_1, Y_2, \ldots$ on the same probability space with $P_{Y_n} = P_{X_n}$ and $P_Y = P_X$, such that $Y_n \to_{\mathrm{a.s.}} Y$.

Big O and Little o Almost Surely

For sequences of random variables X_n and Y_n defined on a common probability space, we identify different types of convergence, either almost sure or in probability.

- Big O almost surely, written $\mathrm{O}(Y_n)$ a.s.

$$X_n = \mathrm{O}(Y_n) \text{ a.s. iff } \Pr(\|X_n\| = \mathrm{O}(\|Y_n\|)) = 1$$

- Little o almost surely, written $\mathrm{o}(Y_n)$ a.s.

$$X_n = \mathrm{o}(Y_n) \text{ a.s. iff } \|X_n\|/(\|Y_n\|) \to_{\mathrm{a.s.}} 0.$$

Compare $X_n/Y_n \to_{\mathrm{a.s.}} 0$ for $X_n \in \mathbb{R}^m$ and $Y_n \in \mathbb{R}$.

Big O and Little o Weakly

We also have relationships in which one sequence converges to another in probability.

- Big O in probability, written $\mathrm{O_P}(Y_n)$.

  $$X_n = \mathrm{O_P}(Y_n) \text{ iff } \forall \epsilon > 0 \; \exists \text{ constant } C_\epsilon > 0 \ni \sup_n \Pr(\|X_n\| \geq C_\epsilon \|Y_n\|) < \epsilon.$$

  If $X_n = \mathrm{O_P}(1)$, X_n is said to be *bounded in probability*.
  It is clear that if $X_n \to_{\mathrm{d}} X$ for any random variable X, then $X_n = \mathrm{O_P}(1)$.
- Little o in probability, written $\mathrm{o_P}(Y_n)$.

  $$X_n = \mathrm{o_P}(Y_n) \text{ iff } \|X_n\|/\|Y_n\| \to_{\mathrm{p}} 0.$$

  If $X_n = \mathrm{o_P}(1)$, then X_n converges in probability to 0. If $X_n = \mathrm{o_P}(1)$, then also $X_n = \mathrm{O_P}(1)$.

Weak Convergence

Convergence in distribution, or weak convergence, plays a fundamental role in statistical inference. It is the type of convergence in the central limits (see Section 1.3.2) and it is the basis for the definition of asymptotic expectation (see Section 2.5.2), which, in turn is the basis for most of the concepts of asymptotic inference. (Asymptotic inference is not based on the limits of the properties of the statistics in a sequence, and in Section 2.5.3, beginning on page 116, we will consider some differences between "aysmptotic" properties and "limiting" properties.)

Theorem 1.9 in Shao gives conditions for convergence in distribution.

The *continuity theorem* (Theorem 1.9(ii)) concerns a sequence of random variables $X_1, X_2, \cdots$ (not necessarily independent) with characteristic functions $\phi_{X_1}, \phi_{X_2}, \cdots$ and a random variable X with characteristic function ϕ_X and states that $X_n \to_{\mathrm{d}} X$ iff $\phi_{X_n}(t) \to \phi_X(t) \; \forall t$. The "if" part of the continuity theorem is called the Lévy-Cramér theorem and the "only if" part is sometimes called the *first limit theorem*.

Another useful fact (Shao's Theorem 1.9(iii)) is called the Cramér-Wold "device". It states that $X_n \to_{\mathrm{d}} X$ iff $tX_n \to_{\mathrm{d}} tX \; \forall t$. (If $X_n, X \in \mathbb{R}^d$, the condition is $t^{\mathrm{T}}X_n \to_{\mathrm{d}} t^{\mathrm{T}}X \; \forall t \in \mathbb{R}^d$. This follows immediately from the continuity theorem.

********* tightness

Shao exercise 1.129: Let $\{F_n\}$ be a sequence of CDFs on $\mathbb{R}$. Let $G_n(x) = F_n(a_n x + c_n)$ and $H_n(x) = F_n(b_n x + d_n)$, where $\{a_n\}$ and $\{b_n\}$ are sequences of positive real numbers and $\{c_n\}$ and $\{d_n\}$ are sequences of real numbers. Now suppose that $G_n \to_{\mathrm{w}} G$ and $H_n \to_{\mathrm{w}} H$, where G and H are nondegenerate CDFs. Then

$$a_n/b_n \to a > 0 \quad \text{and} \quad (c_n - d_n/a_n \to b \in \mathbb{R},$$

and

$$H(ax+b) = G(x) \; \forall x \in \mathbb{R};$$

that is, the distributions are in a location-scale family. **********

Convergence of Functions

The next issue has to do with functions of convergent sequences. We consider a sequence $X_1, X_2, \ldots$ in $\mathbb{R}^k$ and a measurable function g from $(\mathbb{R}^k, \mathcal{B}^k)$ to $(\mathbb{R}^k, \mathcal{B}^k)$. If we know something about the convergence of $\{X_n\}$, can we say anything about the convergence of $\{g(X_n)\}$? We can if g is continuous. (Recall the "portmanteau" theorem for expected values and for convergence in distribution.) To speak about continuity of a function of random variables, we must add some kind of qualifier, such as a.s., which, of course, assumes a probability measure. (That consideration was not relevant for the expectations in the portmanteau theorem.)

So, for a given probability measure, say P_X, and a function g that is continuous a.s. w.r.t. P_X, the simple facts are

$$X_n \to_{\text{a.s.}} X \;\Rightarrow\; g(X_n) \to_{\text{a.s.}} g(X) \tag{1.52}$$

$$X_n \to_{\text{p}} X \;\Rightarrow\; g(X_n) \to_{\text{p}} g(X) \tag{1.53}$$

$$X_n \to_{\text{d}} X \;\Rightarrow\; g(X_n) \to_{\text{d}} g(X) \tag{1.54}$$

This is Theorem 1.10 in Shao.

The next question is about the convergence of sequences formed by addition and multiplication of two sequences of random variables. The answer is provided by Slutsky's theorem.

Slutsky's theorem (Theorem 1.11 in Shao) gives convergence in distribution for the case that one sequence converges in distribution to a random variable and another sequence converges in probability to a fixed real number. It tells us that sums, products, and quotients behave like we would expect (or hope):

$$X_n + Y_n \to_{\text{d}} X + c \tag{1.55}$$

$$X_n Y_n \to_{\text{d}} cX \tag{1.56}$$

$$X_n/Y_n \to_{\text{d}} X/c \text{ if } c \neq 0. \tag{1.57}$$

The next issue concerns the case when we have convergence in distribution as above, and we apply a function to the sequence, but not to the random variable itself. That is, we have $\{X_n\}$ converging in distribution to $Y + c$, and we ask about the convergence of $\{g(X_n)\}$. What we can say depends on the differentiability of g at c.

The useful fact is given as Theorem 1.12(i) in Shao:
Let $X_1, X_2, \ldots$ and Y be random variables (k-vectors) such that

$$a_n(X_n - c) \to_{\mathrm{d}} Y,$$

where c is a constant (k-vector) and $a_1, a_2, \ldots$ is a sequence of constant scalars such that $\lim_{n\to\infty} a_n = \infty$. Now let g be a function from $\mathbb{R}^k$ to $\mathbb{R}$ that is differentiable at c. Then

$$a_n(g(X_n) - g(c)) \to_{\mathrm{d}} (\nabla g(c))^{\mathrm{T}} Y. \tag{1.58}$$

Notice an important point in this theorem. We are given convergence of $(X_n - c)$ and we get convergence of $(g(X_n) - g(c))$ so long as g is differentiable at c.

There is an extension of this given as Theorem 1.12(ii) in Shao for powers of the a_n that has applications in the covariance of the random vector Y.

The most common application of Theorem 1.12(i) arises from the simple corollary (called "Corollary 1.1" in Shao) for the case when Y has the multivariate normal distribution $\mathrm{N}_k(0, \Sigma)$:

$$a_n(g(X_n) - g(c)) \to_{\mathrm{d}} Z, \tag{1.59}$$

where $Z \sim \mathrm{N}_k(0, (\nabla g(c))^{\mathrm{T}} \Sigma \nabla g(c))$.

One reason limit theorems are important is that they can provide approximations useful in statistical inference. For example, the convergence of the sequence above provides a method for setting approximate confidence sets using the normal distribution, so long as no element of $\nabla g(c)$ is zero. This method in asymptotic inference is called the *delta method*.

Expectations of Sequences and Sequences of Expectations

The monotonicity of the expectation operator (1.2) carries over to sequences. The three theorems that relate to the interchange of a Lebesgue integration operation and a limit operation stated on page 391 (monotone convergence, Fatou's lemma, and dominated convergence) apply immediately to expectations:

- monotone convergence
  For $0 \le X_1 \le X_2 \cdots$ a.s.

$$X_n \to_{\text{a.s.}} X \quad \Rightarrow \quad \mathrm{E}(X_n) \to_{\text{a.s.}} \mathrm{E}(X) \tag{1.60}$$

- Fatou's lemma

$$0 \le X_n \,\forall\, n \quad \Rightarrow \quad \mathrm{E}(\liminf_n X_n) \le \liminf_n \mathrm{E}(X_n) \tag{1.61}$$

- dominated convergence
  Given a fixed Y with $\mathrm{E}(Y) < \infty$,

$$|X_n| \le Y \,\forall\, n \text{ and } X_n \to_{\text{a.s.}} X \quad \Rightarrow \quad \mathrm{E}(X_n) \to_{\text{a.s.}} \mathrm{E}(X). \tag{1.62}$$

Another useful convergence result for expectations is the Helly-Bray theorem (or just the Helly theorem:
If g is a bounded and continuous function over the support of $\{X_n\}$, then

$$X_n \to_{\mathrm{d}} X \;\Leftrightarrow\; \mathrm{E}(g(X_n)) \to_{\mathrm{d}} \mathrm{E}(g(X)). \tag{1.63}$$

1.3 Limit Theorems

There are two general types of important limit theorems: laws of large numbers and central limit theorems. Laws of large numbers give limits for probabilities or for expectations of sequences of random variables. The convergence to the limits may be weak or strong.

Central limit theorems provide weak convergence results, but they do even more; they specify a limiting *normal distribution*. The first versions of both laws of large numbers and central limit theorem applied to sequences of binomial random variable.

1.3.1 Laws of Large Numbers

Bernoulli's theorem

The first law of large numbers was *Bernoulli's* (Jakob) *theorem*: If X_n has a binomial distribution with parameters n and π, then

$$X_n/n \to_{\mathrm{p}} \pi. \tag{1.64}$$

This follows from $\int_\Omega (X_n/n - \pi)^2 \mathrm{d}P = \pi(1-\pi)/n$, which means X_n/n converges in mean square to π, which in turn means that it converges in probability to π. This is a *weak law* because the convergence is in probability.

The weak law of large numbers for i.i.d. random variables

A generalization of the Bernoulli's theorem is the *weak law of large numbers (WLLN) for i.i.d. random variables*:

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables. There exists a sequence of real numbers $a_1, a_2, \ldots$ such that

$$n\Pr(|X_1| > n) \to 0 \quad \Longleftrightarrow \quad \frac{1}{n}\sum_{i=1}^{n} X_i - a_n \to_{\mathrm{p}} 0 \tag{1.65}$$

If this condition holds, we can choose $a_n = \mathrm{E}(X_i \mathrm{I}_{\{|X_1| \le n\}})$.

We will not prove this or the following limit theorems at this time.

The strong law of large numbers for i.i.d. random variables

If $\mathrm{E}(|X_1|) < \infty$, we can form a strong law in terms of $\mathrm{E}(X_1)$.

The *strong law of large numbers (SLLN) for i.i.d. random variables* states

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables such that $\mathrm{E}(|X_1|) < \infty$. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to_{\text{a.s.}} \mathrm{E}(X_1). \tag{1.66}$$

(Shao states this in a slightly different form.)
A slight generalization is the alternate conclusion

$$\frac{1}{n}\sum_{i=1}^{n} c_i(X_i - \mathrm{E}(X_1)) \to_{\text{a.s.}} 0,$$

for any bounded sequence of real numbers $c_1, c_2, \ldots$.

We can generalize these two limit theorems to the case of independence but not necessarily identical distributions, by putting limits on normalized p^{th} moments.

The weak law of large numbers for independent random variables with finite expectation

The *weak law of large numbers for independent random variables with finite expectation*: Let $X_1, X_2, \ldots$ be a sequence of independent random variables such for some constant $p \in [1, 2]$,

$$\lim_{n\to\infty} \frac{1}{n^p}\sum_{i=1}^{n} \mathrm{E}(|X_i|^p) = 0,$$

then

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathrm{E}(X_i)) \to_{\text{p}} 0. \tag{1.67}$$

The strong law of large numbers for independent random variables with finite expectation

The *strong law of large numbers for independent random variables with finite expectation*: Let $X_1, X_2, \ldots$ be a sequence of independent random variables such for some constant $p \in [1, 2]$,

$$\sum_{i=1}^{\infty} \frac{\mathrm{E}(|X_i|^p)}{i^p} < \infty,$$

then

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathrm{E}(X_i)) \to_{\text{a.s.}} 0. \tag{1.68}$$

1.3.2 Central Limit Theorems

Central limit theorems give conditions that imply that certain standardized sequences converge to a normal distribution. The simplest ones apply to i.i.d. random variables, and more complicated ones apply to independent random variables that are not necessarily identically distributed.

The central limit theorems require finite first and second moments.

The first central limit theorem, called the de Moivre Laplace central limit theorem followed soon after Bernoulli's theorem, and like Bernoulli's theorem, it applies to X_n that has a binomial distribution with parameters n and π.

The de Moivre Laplace central limit theorem

The *de Moivre Laplace central limit theorem* states that

$$\frac{X_n - n\pi}{\sqrt{n\pi(1-\pi)}} \to_{\mathrm{d}} \mathrm{N}(0,1). \tag{1.69}$$

This central limit theorem, called the de Moivre Laplace central limit theorem is a special case of the classical central limit theorem for i.i.d. random variables with finite mean and variance.

Notice that Bernoulli's theorem and the de Moivre Laplace central limit theorem, which are stated in terms of binomial random variables, apply to normalized limits of sums of Bernoulli random variables. This is the usual form of these kinds of limit theorems; that is, they apply to normalized limits of sums of random variables. The first generalizations apply to sums of i.i.d. random variables, and then further generalizations apply to sums of just independent random variables.

The central limit theorem for i.i.d. scalar random variables with finite mean and variance

Let $X_1, X_2, \ldots$ be a sequence of independent random variables that are identically distributed with mean μ and variance $\sigma^2 > 0$. Then

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} \to_{\mathrm{d}} \mathrm{N}(0,1). \tag{1.70}$$

The proof of this uses a limit of a characteristic function and the uniqueness of the characteristic function (see page 37).

Independent but not identical

The more general central limit theorems apply to a sequence of a particular type of finite subsequences. The variances of the sums of the subsequences is what is used to standardize the sequence so that it is convergent. We define the sequence and the subsequences follows.

Let $\{X_{nj}, j = 1, 2, \ldots, k_n\}$ be independent random variables with $0 < \sigma_n^2$, where $\sigma_n^2 - \mathrm{V}(\sum_{j=1}^{k_n} X_{nj})$ and $k_n \to \infty$ as $n \to \infty$.

A central limit theorem for independent scalar random variables with finite mean and variance

A more general central limit theorem is called *Lindeberg's central limit theorem.* It is stated in terms of a sequence of finite subsequences:

Let $\{X_{nj}, j = 1, 2, \ldots, k_n\}$ be independent random variables with $0 < \sigma_n^2$, where $\sigma_n^2 - \mathrm{V}(\sum_{j=1}^{k_n} X_{nj})$ and $k_n \to \infty$ as $n \to \infty$. If the *Lindeberg condition*,

$$\sum_{j=1}^{k_n} \mathrm{E}\left((X_{nj} - \mathrm{E}(X_{nj}))^2 \mathrm{I}_{\{|X_{nj} - \mathrm{E}X_{nj}| > \epsilon\sigma_n\}}(X_{nj})\right) = \mathrm{o}(\sigma_n^2) \text{ for any } \epsilon > 0, \tag{1.71}$$

holds, then

$$\frac{1}{\sigma_n} \sum_{j=1}^{k_n} (X_{nj} - \mathrm{E}(X_{nj})) \to_{\mathrm{d}} \mathrm{N}(0, 1). \tag{1.72}$$

This is proved, as Theorem 1.15, in Shao.

Lindeberg's condition requires that the sum of the second central moments over the full support minus the squared central differences near the mean is ultimately dominated by the variance of the sum. (That is to say, the sum of the tail components of the variance is dominated by the variance of the sum. This means that the distributions cannot be too heavy-tailed.) The requirement is in terms of an ϵ that tells how much of the central region to remove before computing the individual central moments.

Another approach is to compare the sum of higher order central moments. This yields a stronger condition; it is a condition on a power in terms of a positive addition δ to 2, rather than on a fixed power of 2 over an interval controlled by ϵ. *Liapounov's condition* applies to the order of $(2+\delta)$ moments for $\delta > 0$:

$$\sum_{j=1}^{k_n} \mathrm{E}\left(|X_{nj} - \mathrm{E}(X_{nj})|^{2+\delta}\right) = \mathrm{o}(\sigma_n^{2+\delta}) \quad \text{for some } \delta > 0. \tag{1.73}$$

As above, we assume k_n is a sequence such that $k_n \to \infty$ as $n \to \infty$. The more stringent Liapounov's condition implies Lindeberg's condition. It is sometimes easier to establish Liapounov's condition than Lindeberg's condition, however.

Lindeberg's condition (or Liapounov's condition, of course) implies *Feller's condition*, which is:

$$\lim_{n\to\infty} \max_{j \le k_n} \frac{\sigma_{nj}^2}{\sigma_{nj}^2} = 0, \tag{1.74}$$

under the assumption as above that k_n is a sequence such that $k_n \to \infty$ as $n \to \infty$. This condition comes up in the proof of Lindeberg's central limit theorem.

Multivariate central limit theorems for independent random variables with finite mean and variance

The central limit theorems stated above have multivariate extensions that are relatively straightforward. The complications arise from the variance-covariance matrices, which must replace the simple scalars σ_n^2.

The simplest situation is the i.i.d. case where each member of the sequence $\{X_n\}$ of random k-vectors has the finite variance-covariance matrix Σ. In that case, similar to equation (1.70) for i.i.d. scalar random variables, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \mathrm{E}(X_i)) \to_{\mathrm{d}} \mathrm{N}_k(0, \Sigma). \tag{1.75}$$

Another type of multivariate central limit theorem can be formed by thinking of the subsequences in equation (1.72) as multivariate random variables. Let $\{k_n\}$ be a sequence of constants such that $k_n \to \infty$ as $n \to \infty$. Let $X_{ni} \in \mathbb{R}^{m_i}$, where $m_i \leq m$ for some fixed integer m and for $i = 1, \ldots, k_n$, be independent with

$$\inf_{i,n} \lambda_{ni} > 0,$$

where λ_{ni} is the smallest eigenvalue of $\mathrm{V}(X_{ni})$. (Note that this is saying that variance-covariance matrix is positive definite for every n and i; but it's saying a little more than that.) Also suppose that for some $\delta > 0$, we have

$$\sup_{i,n} \mathrm{V}(\|X_{ni}\|^{2+\delta}) < \infty.$$

Now, let c_{ni} be a sequence in $\mathbb{R}^{m_i}$ with the property that it is diffuse:

$$\lim_{n\to\infty}\left(\max_{1\leq i\leq k_n}\|c_{ni}\|^2 \Big/ \sum_{i=1}^{k_n}\|c_{ni}\|^2\right) = 0.$$

Then we have something similar to equation (1.72):

$$\sum_{j=1}^{k_n} c_{ni}^{\mathrm{T}}(X_{nj} - \mathrm{E}(X_{nj})) \Big/ \left(\sum_{j=1}^{k_n}\mathrm{V}(c_{ni}^{\mathrm{T}}X_{nj})\right)^{1/2} \to_{\mathrm{d}} \mathrm{N}(0,1). \tag{1.76}$$

1.4 Power Series Expansions

Expansions in Taylor series are useful in studying asymptotic distributions. We have seen a simple example of the use of a first-order Taylor series for the asymptotic distributions of functions of random variables. This resulted in the delta method. Higher order Taylor series can be used to develop higher

order delta methods. (We will use a second order delta method on page 214.) Expansions are used to arrive at approximations that are of some order $\mathrm{O}(n^r)$ for some $r > 0$. These approximations immediately yield asymptotic equations or distributions.

Expansions of a given CDF in terms of another CDF, especially the normal CDF, yield very useful approximations. This is an instance of the more general method of representing a given function in terms of basis functions, as we discuss beginning on page 398.

A series using these Hermite polynomials is often called a Gram-Charlier series.

The first few Hermite polynomials are shown in equation (D.68) on page 401.

$$\begin{array}{ll} H_0^{\mathrm{e}}(t) = 1 & H_1^{\mathrm{e}}(t) = t \\ H_2^{\mathrm{e}}(t) = t^2 - 1 & H_3^{\mathrm{e}}(t) = t^3 - 3t \\ H_4^{\mathrm{e}}(t) = t^4 - 6t^2 + 3 & H_5^{\mathrm{e}}(t) = t^5 - 10t^3 + 15t \end{array}$$

******

1.5 Conditional Probability

The concept of conditional distributions provides the basis for the analysis of relationships among variables.

A simple way of developing the ideas begins by defining the conditional probability of event A, given event B. If $\Pr(B) \neq 0$, the conditional probability of event A given event B is

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)},$$

which leads to the useful multiplication rule

$$\Pr(A \cap B) = \Pr(B)\Pr(A|B).$$

We see from this that if A and B are independent

$$\Pr(A|B) = \Pr(A).$$

If we interpret all of this in the context of the probability space $(\Omega, \mathcal{F}, P)$, we can define a new "conditioned" probability space, $(\Omega, \mathcal{F}, P_B)$, where we define P_B by

$$P_B(A) = \Pr(A \cap B),$$

for any $A \in \mathcal{F}$. From this conditional probability space we could then proceed to develop "conditional" versions of the concepts discussed in the previous sections.

This approach, however, is not entirely satisfactory because of the requirement that $\Pr(B) \neq 0$.

Another approach is to make use of a concept of conditional expectation, and that is what we will proceed to do.

1.5.1 Conditional Expectation: Definition and Properties

The definition of conditional expectation of one random variable given another random variable is developed in two stages. First, we define conditional expectation over a sub-σ-field, and then we define conditional expectation with respect to another measurable function (a random variable, for example) in terms of the conditional expectation over the sub-σ-field generated by the inverse image of the function.

A major difference in conditional expectations and unconditional expectations is that conditional expectations may be nondegenerate random variables. When the expectation is conditioned on a random variable, relations involving the conditional expectations must be qualified as in probability, or with probablility 1.

Conditional expectation over a sub-σ-field

Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $\mathcal{A}$ be a sub-σ-field of $\mathcal{F}$, and let X be an integrable random variable over Ω. The *conditional expectation* of X given $\mathcal{A}$, denoted by, $\mathrm{E}(X|\mathcal{A})$, is a random variable from $(\Omega, \mathcal{F})$ to $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ such that $\int_A \mathrm{E}(X|\mathcal{A})\,\mathrm{d}P = \int_A X\,\mathrm{d}P$ for any $A \in \mathcal{A}$. (The existence and uniqueness of this random variable follows from the Radon-Nikodym theorem (Theorem 1.4 in Shao)).

Conditional expectation with respect to another measurable function

Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $\mathcal{A}$ be a sub-σ-field of $\mathcal{F}$, let X be an integrable random variable over Ω, and let Y be a measurable function from $(\Omega, \mathcal{F}, P)$ to any measurable space $(\Lambda, \mathcal{G})$. Then the *conditional expectation* of X given Y, denoted by $\mathrm{E}(X|Y)$ is defined as the conditional expectation of X given the sub-σ-field generated by Y, that is, $\mathrm{E}(X|\sigma(Y))$.

Sub-σ-fields generated by random variables, such as $\sigma(Y)$, play an important role in statistics. We can think of $\sigma(Y)$ as being the "information provided by Y". In an important type of time series, $Y_1, Y_2, \ldots$, we encounter a sequence $\sigma(Y_1) \subset \sigma(Y_2) \subset \cdots$.

Some properties of conditional expectations

Although the definition above may appear rather abstract, it is not too difficult to work with, and it yields the properties of conditional expectation that we have come to expect based on the limited definitions of elementary probability.

For example, we have the simple relationship with the unconditional expectation:

$$\mathrm{E}(\mathrm{E}(X|\mathcal{A})) = \mathrm{E}(X). \tag{1.77}$$

Also, if the individual conditional expectations exist, the conditional expectation is a linear operator:

$$\forall a \in \mathbb{R}, \mathrm{E}(aX + Y|\mathcal{A}) = a\mathrm{E}(X|\mathcal{A}) + \mathrm{E}(Y|\mathcal{A}) \text{ a.s.} \tag{1.78}$$

This fact follows immediately from the definition. For any $A \in \mathcal{A}$

$$\begin{aligned}\mathrm{E}(aX + Y|\mathcal{A}) &= \int_A aX + Y \,\mathrm{d}P \\ &= a\int_A X \,\mathrm{d}P + \int_A Y \,\mathrm{d}P \\ &= a\mathrm{E}(X|\mathcal{A}) + \mathrm{E}(Y|\mathcal{A})\end{aligned}$$

As with unconditional expectations, we have immediately from the definition:

$$X \leq Y \text{ a.s.} \quad \Rightarrow \quad \mathrm{E}(X|\mathcal{A}) \leq \mathrm{E}(Y|\mathcal{A}) \text{ a.s..} \tag{1.79}$$

We can establish conditional versions of the three theorems stated on page 39 that relate to the interchange of an integration operation and a limit operation (monotone convergence, Fatou's lemma, and dominated convergence). These extensions are fairly straightforward.

- monotone convergence:
  for $0 \leq X_1 \leq X_2 \cdots$ a.s.

$$X_n \to_{\text{a.s.}} X \quad \Rightarrow \quad \mathrm{E}(X_n|\mathcal{A}) \to_{\text{a.s.}} \mathrm{E}(X|\mathcal{A}). \tag{1.80}$$

- Fatou's lemma:

$$0 \leq X_n \,\forall\, n \quad \Rightarrow \quad \mathrm{E}(\liminf_n X_n|\mathcal{A}) \leq \liminf_n \mathrm{E}(X_n|\mathcal{A}) \text{ a.s..} \tag{1.81}$$

- dominated convergence:
  given a fixed Y with $\mathrm{E}(Y|\mathcal{A}) < \infty$,

$$|X_n| \leq Y \,\forall\, n \text{ and } X_n \to_{\text{a.s.}} X \quad \Rightarrow \quad \mathrm{E}(X_n|\mathcal{A}) \to_{\text{a.s.}} \mathrm{E}(X|\mathcal{A}). \tag{1.82}$$

Another useful fact is that if Y is $\mathcal{A}$-measurable and $|XY|$ and $|X|$ are integrable (notice this latter is stronger than what is required to define $\mathrm{E}(X|\mathcal{A})$), then

$$\mathrm{E}(XY|\mathcal{A}) = Y\mathrm{E}(X|\mathcal{A}) \text{ a.s.} \tag{1.83}$$

Conditional expectation can be viewed as a projection in a linear space defined by the square-integrable random variables over a given probability space and the inner product $\langle X, Y\rangle = \mathrm{E}(XY)$ and its induced norm.

Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $\mathcal{A}$ be a sub-σ-field of $\mathcal{F}$, and let $B \in \mathcal{F}$. The *conditional probability* of B given $\mathcal{A}$, denoted by $\Pr(B|\mathcal{A})$ is defined as $\mathrm{E}(\mathrm{I}_B|\mathcal{A})$

1.5.2 Some Useful Conditional Expectations

There are some conditional expectations that arise often, and which we should immediately recognize:

$$\mathrm{E}\big(\mathrm{E}(Y|X)\big) = \mathrm{E}(Y) \tag{1.84}$$

The student should realize that the expectation operator is based on a probability distribution, and so anytime we see "E", we need to ask "with respect to what probability distribution?" In notation such as that above, the distribution is implicit. The inner expectation on the left is with respect to the conditional distribution of X given Y, and so is a function of Y. The outer expectation is with respect to the marginal distribution of Y.

$$\mathrm{V}(Y) = \mathrm{V}\big(\mathrm{E}(Y|X)\big) + \mathrm{E}\big(\mathrm{V}(Y|X)\big) \tag{1.85}$$

This is intuitive, although you should be able to prove it formally. The intuitive explanation is: the total variation in Y is the sum of the variation of its mean given X and its average variation about X (or given X). (Think of SST = SSR + SSE in regression analysis.)

This equality implies the Rao-Blackwell inequality (drop the second term on the right).

1.5.3 Conditional Probability Distributions

We can now develop important concepts of joint and conditional probability distributions in terms of conditional expectations.

- Conditional probability.
  Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $\mathcal{A}$ be a sub-σ-field of $\mathcal{F}$, and let $B \in \mathcal{F}$. The *conditional probability* of B given $\mathcal{A}$, denoted by $\Pr(B|\mathcal{A})$ is defined as $\mathrm{E}(\mathrm{I}_B|\mathcal{A})$.
- Conditional distributions.
  For distributions with PDFs we can define conditional distributions very simply. The concept of a joint distribution with a PDF comes from the Radon-Nikodym derivative of a CDF over a product space. This is the familiar

  $$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)},$$

  which may be taken as a definition so long as $f_Y(y) > 0$. With this, we can work back to define a conditional expectation in agreement with that above.
  Another approach for defining joint and conditional distributions is given in Shao's "Theorem 1.7". In this we start with a probability space $(\mathbb{R}^m, \mathcal{B}^m, P_1)$ and define a probability measure on the measurable space $(\mathbb{R}^n \times \mathbb{R}^m, \sigma(\mathcal{B}^n \times \mathcal{B}^m))$. The existence of such a probability measure is given in the first part of this multi-theorem (which is proved in Billingsley).

For a random variable Y in $\mathbb{R}^m$, its (marginal) distribution is determined by P_1, which we denote as $P_Y(y)$. For $B \in \mathcal{B}^n$ and $C \in \mathcal{B}^m$, the conditional distribution is defined by identifying a probability measure, denoted as $P_{X|Y}(\cdot|y)$, on $(\mathbb{R}^n, \sigma(\mathcal{B}^n))$ for any fixed $y \in \mathbb{R}^m$.
The joint probability measure of (X, Y) over $\mathbb{R}^n \times \mathbb{R}^m$ is defined as

$$P_{XY} = \int_C P_{X|Y}(\cdot|y)\mathrm{d}P_Y(y),$$

where $C \in \mathcal{B}^m$.

- Conditional entropy.
  We define the conditional entropy of X given Y in two ways. The first meaning just follows the definition of entropy in equation (1.7) on page 10 with the conditional PDF $p_{X|Y}$ used in place of the marginal PDF p_X. This leads to an entropy that is a random variable or an entropy for a fixed value $Y = y$. In the more common usage, we define the conditional entropy of X given Y (which is also called the equivocation of X about Y) as the expected value of the term described above; that is,

$$H(X|Y) = -\sum_y p_Y(y) \sum_x p_{X|Y}(x|y) \log(p_{X|Y}(x|y)). \tag{1.86}$$

  As before, the basic definition is made in terms of a PDF dominated by a counting measure, but we extend it to any PDF.
  From the definition we see that

$$H(X|Y) = H(X, Y) - H(Y) \tag{1.87}$$

  or

$$H(X, Y) = H(X|Y) + H(Y).$$

  Compare the latter with equation (1.85).

1.6 Stochastic Processes

Many interesting statistical problems concern *stochastic processes*, which we can think of as a measurable function

$$X : \mathcal{I} \times \Omega \mapsto \mathbb{R}^d,$$

where $\mathcal{I}$ is some index set ($\mathcal{I}$ could be any ordered set). In many cases of interest, $d = 1$; that is, the process is *univariate*.

In the expression above, X is a random variable, and for each $i \in \mathcal{I}$, X_i is a random variable. If the stochastic process is viewed as evolving in time, we usually denote the index by t and we may denote the process as $\{X_t\}$ or as $X(t)$.

The sequences we discussed in Section 1.2 are of course stochastic processes. The sequences considered in that section did not have any particular structure, however. In some cases, we required that they have no structure; that is, that the elements in the sequence were independent. There are many special types of interesting stochastic processes with various structures, such as Markov chains, martingales, and other types of time series. In this section, we will just give some basic definitions, and then discuss briefly two important classes of stochastic process.

States, Times, Notation, and Basic Definitions

The smallest set of measure 1 is called the *state space* of a stochastic process; that is, the range of X is called the state space. Any point in the state space is called a *state*.

If the index set of a stochastic process is countable, we say the process is a *discrete time* stochastic process. We often denote the set of indexes as T, and we write the random variable at time $t \in T$ as X_t. We can index a discrete time process by $0, 1, 2, \ldots$, especially if there is a fixed starting point, although sometimes $\ldots, -2, -1, 0, 1, 2, \ldots$ is more appropriate.

In many applications, the index of a stochastic process ranges over a continuous interval. In that case, we often use a slightly different notation for the index set. Instead of T denoting that set, we consider index set to be the interval $[0, T]$, which of course could be transformed into any finite closed interval. If the index set is a real interval we say the process is a *continuous time* stochastic process. For continuous time stochastic processes, we sometimes use the notation $X(t)$, although we also use X_t.

For a stochastic process over a continuous index set $\mathcal{I}$ we must first be concerned about the continuity of the process in time. We can define continuity of a function (random variable) on $\mathcal{I}$ in the usual way at a given point ($\omega_0 \in \Omega$). Now we define sample continuity for a stochastic process. Sample continuity must depend on a probability measure (because it is not relevant over sets of probability 0), so first, assume a probability spaces $(\Omega, \mathcal{F}, P)$. Given a function

$$X : \mathcal{I} \times \Omega \mapsto \mathbb{R},$$

we say X is *sample continuous* if $X(\omega) : \mathcal{I} \mapsto \mathbb{R}$ is continuous for almost all ω (with respect to P). We also use the phrase *almost surely continuous* or, often, just *continuous*.

A property that seems to occur often in applications and, when it does, affords considerable simplifications for analyses is the conditional independence of the future on the past given the present. This property, called the *Markov property*, can be made precise by considering for $X(t)$ any set $t_0 < t_1 < \cdots < t_n < t$ and requiring for any x that

$$\Pr(X(t) \leq x \mid X(t_0), X(t_1), \ldots, X(t_n)) = \Pr(X(t) \leq x \mid X(t_n)). \tag{1.88}$$

If the marginal distribution of $X(t)$ is independent of t, the process is said to be *homogeneous*.

Many concepts are more easily defined for discrete time processes, although most have analogs for continuous time processes.

Given a discrete time stochastic process

$$X : \{0, 1, 2, \ldots\} \times \Omega \mapsto \mathbb{R},$$

a random variable

$$T : \Omega \mapsto \{0, 1, 2, \ldots\}$$

is called a *stopping time* if the event $\{T = t\}$ depends only on $X_0, \ldots, X_t$ for $n = 0, 1, 2, \ldots$.

A special stopping time is the *first passage time* defined (for discrete time processes) as

$$T_j = \min\{t \geq 1 : X_t = j\},$$

if this set is nonempty; otherwise, $T_j = \infty$.

Stopping times have several important characteristics, such as the fact that the Markov property holds at stopping times.

1.6.1 Probability Models for Stochastic Processes

A model for a stochastic process posits a sampling sequence over a sample space Ω. This yields a *path* or *trajectory*, $(\omega_1, \omega_2, \ldots)$. In continuous time we generally denote a path trajectory as $\omega(t)$. The sample space for the stochastic process becomes the set of paths. We denote this by Ω_T.

We think of a stochastic process in terms of a random variable, X_t, and an associated σ-field $\mathcal{F}_t$ in which X_t is measurable.

In many applications, we assume an evolution of information. If X_s is associated with the σ-field $\mathcal{F}_s$ and $s \leq t$, then $\mathcal{F}_s \subset \mathcal{F}_t$, and in this case, the sequence $\{\mathcal{F}_t\}$ is called a *filtration*. The stochastic process $\{X_t\}$ is said to be *adapted to* the filtration $\{\mathcal{F}_t\}$.

1.6.2 Markov Chains

The simplest stochastic processes is a sequence of i.i.d. random variables; that is, a sequence with no structure. In a simple, but useful structure we substitute a simple conditioning for independence. A sequence of random variables with the Markov property is called a *Markov process*. A Markov process in which the state space is countable is called a *Markov chain*. (The term "Markov chain" is also sometimes used to refer to any Markov process, as in the phrase "Markov chain Monte Carlo", in applications of which the state space is often continuous.)

The theory of Markov chains is usually developed first for *discrete-time* chains, that is, those with a countable index set, and then extended to *continuous-time* chains.

If the state space is countable, it is equivalent to $\mathcal{X} = \{1, 2, \ldots\}$. If X is a random variable from some sample space to $\mathcal{X}$, and

$$\pi_i = \Pr(X = i), \tag{1.89}$$

then the vector $\pi = (\pi_1, \pi_2, \ldots)$ defines a distribution of X on $\mathcal{X}$. Formally, we define a Markov chain (of random variables) $X_0, X_1, \ldots$ in terms of an initial distribution π and a conditional distribution for X_{t+1} given X_t. Let X_0 have distribution π, and given $X_t = j$, let X_{t+1} have distribution $(p_{ij}; i \in \mathcal{X})$; that is, p_{ij} is the probability of a transition from state j at time t to state i at time $t + 1$, and $K = (p_{ij})$ is called the *transition matrix* of the chain. The initial distribution π and the transition matrix K characterize the chain, which we sometimes denote as $Markov(\pi, K)$. It is clear that K is a stochastic matrix, and hence $\rho(K) = \|K\|_\infty = 1$, and $(1, 1)$ is an eigenpair of K.

If K does not depend on the time (and our notation indicates that we are assuming this), the Markov chain is stationary.

A discrete-time Markov chain $\{X_t\}$ with discrete state space $\{x_1, x_2, \ldots\}$ can be characterized by the probabilities $p_{ij} = \Pr(X_{t+1} = x_i \mid X_t = x_j)$. Clearly, $\sum_{i \in \mathcal{I}} p_{ij} = 1$. A vector such as p_{*j} whose elements sum to 1 is called a *stochastic vector* or a distribution vector.

Because for each j, $\sum_{i \in \mathcal{I}} p_{ij} = 1$, K is a *right stochastic matrix*.

The properties of a Markov chain are determined by the properties of the transition matrix. Transition matrices have a number of special properties, which we discuss in Section D.4.7, beginning on page 455.

(Note that many people who work with Markov chains define the transition matrix as the transpose of K above. This is not a good idea, because in applications with state vectors, the state vectors would naturally have to be row vectors. Until about the middle of the twentieth century, many mathematicians thought of vectors as row vectors; that is, a system of linear equations would be written as $xA = b$. Nowadays, almost all mathematicians think of vectors as column vectors in matrix algebra. Even in some of my previous writings, e.g., Gentle, 2007, I have called the transpose of K the transition matrix, and I defined a stochastic matrix in terms of the transpose. The transpose of a right stochastic matrix is a left stochastic matrix, which is what is commonly meant by the unqualified phrase "stochastic matrix". I think that it is time to adopt a notation that is more consistent with current matrix/vector notation. This is merely a change in notation; no concepts require any change.)

If we assume that X_t is a random variable taking values in $\{x_1, x_2, \ldots\}$ and with a PDF (or probability mass function) given by

$$\Pr(X_t = x_i) = \pi_i^{(t)}, \tag{1.90}$$

and we write $\pi^{(t)} = (\pi_1^{(t)}, \pi_1^{(t)}, \ldots)$, then the PDF at time $t + 1$ is

$$\pi^{(t+1)} = K\pi^{(t)}. \tag{1.91}$$

The properties of a Markov chain depend in large measure on whether the transition matrix is reducible or not.

Because 1 is an eigenvalue and the vector 1 is the eigenvector associated with 1, from equation (D.122), we have

$$\lim_{t\to\infty} K^t = 1\pi_s, \tag{1.92}$$

where π_s is the Perron vector of K^{T}.

This also gives us the *limiting distribution* for an irreducible, primitive Markov chain,

$$\lim_{t\to\infty} \pi^{(t)} = \pi_s.$$

The Perron vector has the property $\pi_s = K^{\mathrm{T}}\pi_s$ of course, so this distribution is the *invariant distribution* of the chain.

The definition means that $(1, 1)$ is an eigenpair of any stochastic matrix. It is also clear that if K is a stochastic matrix, then $\|K\|_\infty = 1$, and because $\rho(K) \leq \|K\|$ for any norm and 1 is an eigenvalue of K, we have $\rho(K) = 1$.

A stochastic matrix may not be positive, and it may be reducible or irreducible. (Hence, $(1, 1)$ may not be the Perron root and Perron eigenvector.)

If the state space is countably infinite, the vectors and matrices have infinite order; that is, they have "infinite dimension". (Note that this use of "dimension" is different from our standard definition that is based on linear independence.)

We write the initial distribution as $\pi^{(0)}$. A distribution at time t can be expressed in terms of $\pi^{(0)}$ and K:

$$\pi^{(t)} = K^t\pi^{(0)}. \tag{1.93}$$

K^t is often called the *t-step transition matrix*.

The transition matrix determines various relationships among the states of a Markov chain. State i is said to be *accessible* from state j if it can be reached from state j in a finite number of steps. This is equivalent to $(K^t)_{ij} > 0$ for some t. If state i is *accessible* from state j and state j is *accessible* from state i, states i and j are said to *communicate*. Communication is clearly an equivalence relation. The set of all states that communicate with each other is an *equivalence class*. States belonging to different equivalence classes do not communicate, although a state in one class may be accessible from a state in a different class. If all states in a Markov chain are in a single equivalence class, the chain is said to be *irreducible*.

The limiting behavior of the Markov chain is of interest. This of course can be analyzed in terms of $\lim_{t\to\infty} K^t$. Whether or not this limit exists depends on the properties of K.

Continuous Time Markov Chains

$$K(t) = \mathrm{e}^{tR}$$

R intensity rate. r_{ii} nonpositive, r_{ij} for $i \neq j$ nonnegative, $\sum_{i \in \mathcal{I}} r_{ij} = 0$ for all j.

1.6.3 Martingales

Martingales are an important class of stochastic processes. The concept of conditional expectation is important in developing a theory of martingales. Martingales are special sequences of random variables that have applications in various processes that evolve over time.

We say that $\{X_t : t \in T\}$ is a *martingale* relative to $\{\mathcal{D}_t : t \in T\}$ in some probability space $(\Omega, \mathcal{F}, P)$, if

$$X_s = \mathrm{E}(X_t|\mathcal{D}_t) \text{ for } s > t. \tag{1.94}$$

An alternate definition is in terms of the pairs $(X_t, \mathcal{F}_t)$; that is, the definition is for the random variable and an associated σ-field, rather than the random variable relative to some sequence of σ-fields, as above. We say the sequence $\{(X_t, \mathcal{F}_t) : t \in T\}$, where $\mathcal{F}_t \subset \mathcal{F}_{t+1} \subset \cdots$, is a *martingale* if $\mathrm{E}(X_n|\mathcal{F}_{n-1}) = X_{n-1}$ a.s. We often refer to this as a *forward martingale*, and define a *reverse martingale* analogously with the conditions $\mathcal{F}_t \supset \mathcal{F}_{t+1} \supset \cdots$ and $\mathrm{E}(X_{n-1}|\mathcal{F}_n) = X_n$ a.s.

We say that $\{X_t : t \in T\}$ is a *submartingale* relative to $\{\mathcal{D}_t : t \in T\}$ if

$$X_s \leq \mathrm{E}(X_t|\mathcal{D}_t) \text{ for } s > t. \tag{1.95}$$

The sequence of sub-σ-fields, which is a filtration, is integral to the definition of martingales.

Shao gives an interesting sequence of likelihood ratios that form a martingale (Example 1.25).

A common application of martingales is as a model for stock prices. As a concrete example, we can think of a random variable X_1 as an initial sum (say, of money), and a sequence of events in which $X_2, X_3, \ldots$ represents a sequence of sums with the property that each event is a "fair game"; that is, $\mathrm{E}(X_2|X_1) = X_1$ a.s., $\mathrm{E}(X_3|X_1, X_2) = X_2$ a.s., ... We can generalize this somewhat by letting $\mathcal{D}_n = \sigma(X_1, \ldots, X_n)$, and requiring that the sequence be such that $\mathrm{E}(X_n|\mathcal{D}_{n-1}) = X_{n-1}$ a.s.

Doob's Martingale Inequality

A useful property of submartingales is Doob's martingale inequality. This inequality is related to the Hájek-Rènyi inequality.

Let $\{X_t : t \in [0,T]\}$ be a submartingale relative to $\{\mathcal{D}_t : t \in [0,T]\}$ taking nonnegative real values; that is, $0 \leq X_s \leq \mathrm{E}(X_t|\mathcal{D}_t)$ for s,t. Then for any constant $\epsilon > 0$ and $r \geq 1$,

$$\Pr\left(\sup_{0\leq t\leq T} X_t \geq \epsilon\right) \leq \frac{1}{\epsilon^p}\mathrm{E}(|X_T|^p). \tag{1.96}$$

We also see that Kolmogorov's inequality is a special case of Doob's martingale inequality, because the partial sums in that inequality form a martingale.

Martingale Central Limit Theorem

In Lindeberg's central limit theorem, page 43, that applies to independent sequences, can we relax the independence assumption? We focus on the partial sums in equation (1.72). Let

$$Y_n = \begin{cases} \sum_{j=1}^{k_n}(X_{nj} - \mathrm{E}(X_{nj})) & \text{if } n \leq k_n \\ \sum_{j=1}^{k_n}(X_{k_n j} - \mathrm{E}(X_{k_n j})) & \text{if } n > k_n. \end{cases}$$

(The addends in Y_n are called a *triangular array.*) Now, assume $\{Y_n\}$ is a martingale.

Next, starting with a fixed value for each subsequence, say $X_{n0} = 0$, assume the sum of the normalized conditional variances converge to 1:

$$\frac{1}{\sigma_n}\sum_{j=1}^{k_n}\mathrm{E}\left((X_{nj} - \mathrm{E}(X_{nj}))^2|X_{n1},\ldots,X_{n,j-1}\right) \to_{\mathrm{p}} 1,$$

where, as before, $\sigma_n^2 = \mathrm{V}(\sum_{j=1}^{k_n} X_{nj})$. Then we have

$$\frac{1}{\sigma_n}\sum_{j=1}^{k_n}(X_{nj} - \mathrm{E}(X_{nj})) \to_{\mathrm{d}} \mathrm{N}(0,1), \tag{1.97}$$

which is the same as equation (1.72).

1.7 Families of Probability Distributions

Given a measurable space, $(\Omega, \mathcal{F})$, different choices of a probability measure lead to different probability triples, $(\Omega, \mathcal{F}, P)$. A set of measures $\mathcal{P} = \{P\}$ associated with a fixed $(\Omega, \mathcal{F})$ is called a family of distributions. Families can be defined in various ways. For example, for some $(\Omega, \mathcal{F})$, a very broad family is $\mathcal{P}_c = \{P : P \ll \nu\}$, where ν is the Lebesgue measure. An example of a very specific family for $\Omega = \{0,1\}$ and $\mathcal{F} = 2^\Omega$ is the probability measure

$P_\pi(\{1\}) = \pi$ and $P_\pi(\{0\}) = 1 - \pi$. The distributions in this family, the Bernoulli distributions, are dominated by the counting measure.

Certain families of distributions have proven to be very useful as models of observable random processes. Familiar families include the normal or Gaussian family of distributions, the Poisson family of distributions, the binomial family of distributions, and so on.

At this time you should become familiar with the families of distributions in Appendix C, beginning on page 337.

1.7.1 Characterizing a Family of Distributions

A probability family or family of distributions, $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$, is a set of probability distributions of a random variable that is defined over a given sample space Ω. The index of the distributions may be just that, an arbitrary index in some given set Θ which may be uncountable, or it may be some specific point in a given set Θ in which the value of θ carries some descriptive information about the distribution; for example, θ may be a 2-vector in which one element is the mean of the distribution and the other element is the variance of the distribution.

The distribution functions corresponding to the members of most interesting families of distributions that we will discuss below do not constitute a distribution function space as defined on page 403. This is because mixtures of distributions in most interesting families of distributions are not members of the same family. That is, distributions defined by convex linear combinations of CDFs generally are not members of the same family of distributions. On the other hand, often linear combinations of random variables do have distributions in the same family of distributions as that of the individual random variables. (The sum of two normals is normal; but a mixture of two normals is not normal.)

Likelihood Functions

The problem of fundamental interest in statistics is to identify a particular distribution within some family of distributions, given observed values of the random variable. Hence, in statistics, we may think of θ or P_θ as a *variable*. Given a PDF f_θ, which is a function whose argument is a value of a random variable x, we define a *likelihood function* as a function of θ for the fixed x:

$$L(\theta \,|\, x) = f_\theta(x).$$

In statistical applications we may faced with the problem of choosing between two distributions P_{θ_1} and P_{θ_2}. For a given value of x, we may base our choice on the two likelihoods, $L(\theta_1 \,|\, x)$ and $L(\theta_2 \,|\, x)$, perhaps using the *likelihood ratio*

$$\lambda(\theta) = \frac{L(\theta_2 \,|\, x)}{L(\theta_1 \,|\, x)}.$$

Parametric Families

A family of distributions on a measurable space $(\Omega, \mathcal{F})$ with probability measures P_θ for $\theta \in \Theta$ is called a *parametric family* if $\Theta \subset \mathbb{R}^d$ for some fixed positive integer d and θ fully determines the measure. In that case, we call θ the *parameter* and Θ the parameter space.

A family that cannot be indexed in this way is called a nonparametric family. In nonparametric methods, our analysis usually results in some general description of the distribution, rather than in a specification of the distribution.

The type of a family of distributions depends on the parameters that characterize the distribution. A "parameter" is a real number that can take on more than one value within a parameter space. If the parameter space contains only one point, the corresponding quantity characterizing the distribution is not a parameter.

Many common families are multi-parameter, and specialized subfamilies are defined by special values of one or more parameters. For example, in a very widely-used notation, the family of gamma distributions is characterized by three parameters, γ, called the "location"; β, called the "scale"; and α, called the "shape". Its PDF is $(\Gamma(\alpha))^{-1}\beta^{-\alpha}(x-\gamma)^{\alpha-1}\mathrm{e}^{-(x-\gamma)/\beta}\mathrm{I}_{[\gamma,\infty)}(x)$. This is sometimes called the "three-parameter gamma", because often γ is taken to be a fixed value, usually 0. As noted above, if a parameter is assigned a fixed value, then it ceases to be a parameter. This is important, because what are parameters determine the class of a particular family. For example, the three-parameter gamma is not a member of the exponential class; however, the standard two-parameter gamma, with γ fixed at 0, is a member of the exponential class.

Specific values of the parameters determine special subfamilies of distributions. For example, in the three-parameter gamma, if α is fixed at 1, the resulting distribution is the two-parameter exponential, and if, additionally, γ fixed at 0, the resulting distribution is what most people call an exponential distribution.

(Oddly, Lehmann many years ago chose the two-parameter exponential, with location and scale, to be "the exponential", and chose the two-parameter gamma, with shape and scale, to be "the gamma". The convenient result was that he could use the exponential as an example of a distribution that is not a member of the exponential class but is a member of the location-scale class, and he could use the gamma as an example of a distribution that is a member of the exponential class but is not a member of the location-scale class. Other authors in mathematical statistics, such as Shao, followed this terminology. It is not nonstandard; it is just odd to choose to include the location parameter in the definition of the exponential family of distributions, where it is very rarely used by anyone else, but not to include the location parameter in the definition of the gamma family of distributions, where occasionally other people use it.

As noted, of course, it is just so we can have convenient examples of specific types of families of distributions.)

Types of Families

We identify certain collections of families of distributions for which we can derive general results. Although I would prefer to call such a collection a "class", most people call it a "family", and so I will too, at least sometimes. Calling these collections of families "families" leads to some confusion, because we can have a situation such as "exponential family" with two different meanings.

The most important class is the exponential class, or "exponential family". This family has a number of useful properties that involve complete sufficient statistics, unbiased estimators, Bayes estimators, and maximum likelihood estimators.

Another important type of family of distributions is a group family, of which there are three important instances: a scale family, a location family, and a location-scale family.

There are various other types of families characterized by their shape or by other aspects useful in specific applications or that lead to optimal standard statistical procedures.

Mixture Families

In applications it is often the case that a single distribution models the observed data adequately. Sometimes two or more distributions from a single family of distributions provide a good fit of the observations, but in other cases, more than one distributional family is required to provide an adequate fit. In some cases most of the data seem to come from one population but a small number seem to be extreme outliers. Some distributions, such as a Cauchy, are said to be "outlier-generating", but often such distributions are difficult to work with (because they have infinite moments, for example). Mixtures of distributions, such as the ϵ-mixture distribution (see page 285), are often useful for modeling data with anomalous observations.

A mixture family can be defined in terms of a set of CDFs $\mathcal{P}_0$. The CDF of a mixture is $\sum w_i P_i$, where $P_i \in \mathcal{P}_0$, $0 \leq w_i \leq 1$, and $\sum w_i = 1$. The set $\mathcal{P}$ of all such mixture CDFs is called a distribution function space (see page 403).

1.7.2 Families Characterized by the Shape of the Probability Density

The general shape of a probability density may determine properties of statistical inference procedures. We can easily identify various aspects of a probability distribution that has a continuous density function. For discrete distributions, some of the concepts carry over in an intuitive fashion, and some do not apply.

In the following, we will use $p_\theta(x)$ to represent a continuous PDF for a distribution whose support is connected.

Symmetric family

A symmetric family is one for which for any given θ there is a constant τ that may depend on θ, such that $p_\theta(\tau + x) = p_\theta(\tau - x)$, for all x.

Unimodal family

The term "mode" is used in various ways; the most common of which is to mean a point x_0 such that $p_\theta(x_0) \geq p_\theta(x)$, for all x. In this sense, a family of distributions is unimodal if for any given θ the mode of the distribution exists and is unique. Another more common definition calls family of distributions unimodal if for any given θ, $p_\theta(x)$ is concave in x.

Totally positive family

A totally positive family of distributions is defined in terms of the total positivity of the PDF, treating it as a function of two variables, θ and x. In this sense, a family is totally positive of order r iff for all $x_1 < \cdots < x_n$ and $\theta_1 < \cdots < \theta_n$,

$$\begin{vmatrix} p_{\theta_1}(x_1) & \cdots & p_{\theta_1}(x_n) \\ \vdots & \vdots & \vdots \\ p_{\theta_n}(x_1) & \cdots & p_{\theta_n}(x_n) \end{vmatrix} \geq 0 \quad \forall\, n = 1, \ldots, r.$$

Logconcave family

If $\log p_\theta(x)$ is concave in x for any θ, the family is called a logconcave family. It is also called a *strongly unimodal family*. A strongly unimodal family is unimodal; that is, if $\log p_\theta(x)$ is concave in x $p_\theta(x)$ is concave in x (exercise!). Strong unimodality is a special case of total positivity. The relevance of strong unimodality is that the likelihood ratio is monotone in x iff the distribution is strongly unimodal. (See page 61 for a definition of monotone likelihood ratio. In Chapter 6 we will see how monotone likelihood ratios simplify the problem of testing statistical hypotheses.)

Heavy-tailed family

If for some constant b, either $x \geq b$ or $x \leq b$ implies $p(x) > c\exp(-x^2)$ where c is a positive constant, the distribution with PDF p is said to be heavy-tailed. Such a distribution is also called an outlier-generating distribution.

1.7.3 "Regular" Families

A reason for identifying a family of distributions is so that we can state interesting properties that hold for all distributions within the family. The statements that specify the family are the hypotheses for important theorems. These statements may be very specific: "if $X_1, X_2, \ldots$ is a random sample from a *normal distribution...*", or they may be more general: "if $X_1, X_2, \ldots$ is a random sample from a *distribution with finite second moment...*"

Some simple characterization such as "having finite second moment" is easy to state each time its need arises, so there is little to be gained by defining such a class of distributions. On the other hand, if the characteristics are more complicated to state in each theorem that refers to that family of distributions, it is worthwhile giving a name to the set of characteristics.

Because in statistical applications we are faced with the problem of choosing the particular distributions P_{θ_0} from a family of distributions, $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$, the behavior of the CDFs or PDFs as functions of θ are of interest. It may be important, for example, that the PDFs in this family be continuous with respect to θ or that derivatives of a specified order with respect to θ exist.

Conditions that characterize a set of objects for which a theorem applies are called "regularity conditions". I do not know the origin of this term, but it occurs in many areas of mathematics. In statistics there are a few sets of regularity conditions that define classes of interesting probability distributions.

Families Satisfying the Fisher Information Regularity Conditions

The most important set of regularity conditions in statistics are some that allow us to put a lower bound on the variance of an unbiased estimator (see inequality (1.45) and Section 4.1). Consider the family of distributions $\mathcal{P} = \{P_\theta; \theta \in \Theta\}$ that have densities p_θ.

There are generally three conditions that together are called the *Fisher information regularity conditions*:

- The parameter space Θ is an open interval (in one dimension, and a cross product of open intervals in multidimensions).
- The support is independent of θ; that is, all P_θ have a common support.
- For any x in the support and $\theta \in \Theta$, $\partial p_\theta(x)/\partial\theta$ exists and is finite.

The latter two conditions ensure that the operations of integration and differentiation can be interchanged.

Because the Fisher information regularity conditions are so important, the phrase "regularity conditions" is often taken to mean "Fisher information regularity conditions".

Families Satisfying the Le Cam Regularity Conditions

The Le Cam regularity conditions are the usual FI regularity conditions plus the requirement that the FI matrix be positive definite and for any fixed $\theta \in \Theta$, there exists a positive number c_θ and a positive function h_θ such that $\mathrm{E}(h_\theta(X)) < \infty$ and

$$\sup_{\gamma:\|\gamma-\theta\|<c_\theta} \left\| \frac{\partial^2 \log f_\gamma(x)}{\partial\gamma(\partial\gamma)^{\mathrm{T}}} \right\|_{\mathrm{F}} \leq h_\theta(x) \ a.e.$$

where $f_\theta(x)$ is a PDF w.r.t. a σ-finite measure.

Families with Monotone Likelihood Ratios

Let X be a random variable with distribution in the family $\mathcal{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}\}$ that is dominated by a σ-finite measure ν, and let $f_\theta(x) = \mathrm{d}P_\theta/\mathrm{d}\nu$. Let $y(x)$ be a scalar-valued function. The family $\mathcal{P}$ is said to have a *monotone likelihood ratio* in $y(x)$ iff for any $\theta_1 < \theta_2$, the likelihood ratio, $f_{\theta_2}(x)/f_{\theta_1}(x)$ is a nondecreasing function of $y(x)$ for all values of x for which $f_{\theta_1}(x)$ is positive.

We could of course reverse the inequality and/or require the ratio be nonincreasing.

Families with monotone likelihood ratios are of particular interest because they are easy to work with in testing composite hypotheses (see the discussion beginning on page 236).

The concept of a monotone likelihood ratio family can be extended to families of distributions with multivariate parameter spaces, but the applications in hypothesis testing are not as useful because we are usually interested in each element of the parameter separately.

1.7.4 The Exponential Class

The exponential class is a set of families of distributions that have some particularly useful properties for statistical inference. The important characteristic of a family of distributions in the exponential class is the way in which the parameter and the value of the random variable can be separated in the density function. Another important characteristic of the exponential family is that the support of a distribution in this family does not depend on any "unknown" parameter.

A member of a family of distributions in the exponential class is one with densities that can be written in the form

$$p_\theta(x) = \exp\left((\eta(\theta))^{\mathrm{T}} T(x) - \xi(\theta)\right) h(x), \tag{1.98}$$

where $\theta \in \Theta$.

Notice that all members of a family of distributions in the exponential class have the same support. Any restrictions on the range may depend on x through $h(x)$, but they cannot depend on the parameter.

A family of distributions in the exponential class is called an exponential family, but do not confuse an "exponential family" in this sense with *the* "exponential family", that is, the parametric family with density of the form $\frac{1}{b}\mathrm{e}^{-x/b}\,\mathrm{I}_{(0,\infty)}(x)$. (This is the usual form of the exponential family, and it is a member of the exponential class. In courses in mathematical statistics, it is common to define the exponential family to be the two-parameter family with density $\frac{1}{b}\mathrm{e}^{-(x-a)/b}\,\mathrm{I}_{(a,\infty)}(x)$. This two-parameter form is not used very often, but it is popular in courses in mathematical statistics because this exponential family is not an exponential family(!) because of the range dependency.)

The form of the expression depends on the σ-finite dominating measure that defines the PDF. If the expression above results from

$$p_\theta = \frac{\mathrm{d}P_\theta}{\mathrm{d}\nu}$$

and we define a measure λ by $\lambda(A) = \int_A h\mathrm{d}\nu \ \forall A \in \mathcal{F}$, then we could write the PDF as

$$\frac{\mathrm{d}P_\theta}{\mathrm{d}\lambda} = \exp\left((\eta(\theta))^{\mathrm{T}} T(x) - \xi(\theta)\right). \tag{1.99}$$

One thing that keeps a parametric family from being in the exponential class is dependence of the support of the distribution on a parameter.

The form of the expression also depends on the *parametrization*; that is, the particular choice of the form of the parameters. First, notice that the only identifiable parameters must be in the elements of $\eta(\theta)$. The other function of the parameters, $\xi(\theta)$, cannot introduce any more identifiable parameters; in fact, it can be written simply as

$$\xi(\theta) = \log\left(\int_{\mathcal{X}} \exp\left((\eta(\theta))^{\mathrm{T}} T(x) - \xi(\theta)\right) h(x)\mathrm{d}x\right).$$

The expression

$$\mu(\theta) = \mathrm{E}(T(x) \tag{1.100}$$

is called the *mean-value parameter*, and use of $\mu(\theta)$ is called the *mean-value parametrization*.

If a family of distributions has parameters α and β, we could equivalently say the family has parameters α and γ, where $\gamma = \alpha + \beta$; that is,

$$\begin{pmatrix} \alpha \\ \gamma \end{pmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}.$$

In this case of course we would have to replace $T(x) = (T(x)_1, T(x)_2)$

$$\widetilde{T}(x) = (T(x)_1 - T(x)_2, T(x)_2).$$

In fact, if $\eta(\theta) \in \mathbb{R}^d$, and D is any nonsingular $d \times d$ matrix, then with $\tilde{\eta} = D\eta(\theta)$, we can write an equivalent form of $(\eta(\theta))^{\mathrm{T}} T(x)$. To do so of course, we must transform $T(x)$ also. So $(\eta(\theta))^{\mathrm{T}} T(x) = \tilde{\eta}^{\mathrm{T}} \widetilde{T}(x)$, where $\widetilde{T}(x) = (D^{\mathrm{T}})^{-1} T(x)$.

In the expression for the density, it might be more natural to think of the parameter as η rather than θ; that way we would have an expression of form $\eta^{\mathrm{T}} T(x)$ rather than $(\eta(\theta))^{\mathrm{T}} T(x)$. We call the form

$$p_\theta(x) = \exp\left((\eta^{\mathrm{T}} T(x) - \zeta(\eta)\right) h(x) \tag{1.101}$$

the *canonical exponential form*, and we call

$$\mathrm{H} = \{\eta : \int \mathrm{e}^{\eta^{\mathrm{T}} T(x)} h(x)\mathrm{d}x < \infty\} \tag{1.102}$$

the *natural parameter space*. (Shao denotes this as Ξ; I use H, which is the upper-case form of η.) The conditions in equation (1.102) are necessary to ensure that a $\zeta(\eta)$ exists such that $p_\theta(x)$ is a PDF. Another characterization of H is

$$\mathrm{H} = \{\eta : \eta = \eta(\theta), \theta \in \Theta\}$$

(under the assumption that Θ is properly defined, of course).

We say the exponential family is of *full rank* if the natural parameter space contains an open set.

*** curved exponential

A PDF of the form $f(x;\theta)\mathrm{I}(x;\theta)$ (where $\mathrm{I}(x;\theta)$ is an indicator function such that for some given x_0, $\exists\theta_1,\theta_2 \in \Theta \ni \mathrm{I}(x;\theta_1) = 0, \mathrm{I}(x;\theta_2) = 1$) cannot be put in the form $c\exp(g(x;\theta))h(x)$ because $c\exp(g(x;\theta)) > 0$ a.e. (because the PDF must be bounded a.e.). Shao, following equation (2.7), presents a more complicated argument to show that $\mathrm{U}(0,\theta)$ is not a member of an exponential family.

- Some families that are exponential: the normal, the log-normal, the standard double exponential (with fixed mean), the binomial and multinomial, the Poisson, the negative binomial, the beta with fixed range (which includes the usual uniform) and the Dirichlet, and the usual fixed-range gamma (which includes the usual exponential).
- Some that are not: two-parameter exponential or three-parameter gamma whose range has an unknown lower limit, uniform with parametric ranges, double exponential with unknown mean, and Cauchy.

Properties of Exponential Families

Fisher information regularity conditions **** prove

Differentiate the identity

$$\int p_\theta(x) = \exp\left((\eta^{\mathrm{T}}T(x) - \zeta(\eta)\right)h(x)\mathrm{d}x = 1$$

w.r.t. η. Get

$$\mathrm{E}_\eta(T(X)) = \nabla\zeta(\eta). \tag{1.103}$$

Then differentiate (1.103) and get

$$\mathrm{V}_\eta(T(X)) = \mathrm{H}_\zeta(\eta), \tag{1.104}$$

where $\mathrm{H}_\zeta(\eta)$ is the matrix of second derivatives of ζ with respect to η.

It is often a simple matter to determine if a member of the exponential class of distributions is a monotone likelihood ratio family. If $\eta(\theta)$ and $T(x)$ in equation (1.98) for the PDF of a distribution in the exponential class are scalars, and if $\eta(\theta)$ is monotone in θ, then the family has a monotone likelihood ratio in $T(x)$.

1.7.5 Parametric-Support Families

Parametric-support families have simple range dependencies, that is, these are distributions whose supports depend on parameters. A distribution in any of these families has a PDF in the general form

$$p_\theta(x) = c(\theta)f(x)\mathrm{I}_{[f_1(\theta),f_2(\theta)]}(x).$$

Shao calls these families "truncation families". Most people use the term "truncated family" to refer to a family that is artificially truncated (for example, due to censoring). In his terminology, the three-parameter gamma would be a truncated distribution. In more standard terminology, a truncated gamma is the distribution formed from a two-parameter distribution with PDF $c(\Gamma(\alpha))^{-1}\beta^{-\alpha}x^{\alpha-1}\mathrm{e}^{-x/\beta}\mathrm{I}_{[\tau_1,\tau_2)}(x)$, where c is just the normalizing constant, which is a function of α, β, τ_1, and τ_2.

Parametric-support families, such as the family of two-parameter exponentials, are not exponential families. Exponential families, such as the family of one-parameter exponentials, are not parametric-support families.

1.7.6 Group Families

"Group" families are distributions that have a certain invariance with respect to a group of transformations on the random variable.

The most common group is the group of linear transformations, and this yields a location-scale group family, or just *location-scale family*, the general form of which is defined below.

A (multivariate) location-scale family of distributions is defined in terms of a given distribution on $(\mathbb{R}^k, \mathcal{B}^k)$ as all distributions for which the probability measure is invariant under linear transformations.

- Let P be a a probability measure on $(\mathbb{R}^k, \mathcal{B}^k)$. Let $\mathcal{V} \subset \mathbb{R}^k$ and let $\mathcal{M}_k$ be a collection of $k \times k$ symmetric positive definite matrices. The family

  $$\{P_{(\mu,\Sigma)} \ : \ P_{(\mu,\Sigma)}(B) = P(\Sigma^{1/2}(B-\mu)), \text{ for } \mu \in \mathcal{V}, \Sigma \in \mathcal{M}_k, B \in \mathcal{B}^k\}$$

  is called a *location-scale* family.

If the PDF of a distribution in a location-scale family is $f(x)$, the PDF of any other distribution in that family is $f((x-\mu)/\sigma)/\sigma$, for some choice of μ and σ; hence, we often use $f((x-\mu)/\sigma)/\sigma$ generically to represent the PDF of a distribution in a location-scale family.

Clearly, a location-scale family must have enough parameters and parameters of the right form in order for the location-scale transformation to result in a distribution in the same family. For example, a three-parameter gamma distribution is a location-scale family, but a two-parameter gamma (without the range dependency) is not.

Some standard parametric families that are location-scale group families: normal, double exponential, and Cauchy with parametric centers, and exponential and uniform with parametric ranges.

A family that is a member of the group family may also be a member of the exponential family.

1.7.7 Complete Families

A family of distributions $\mathcal{P}$ is said to be *complete* iff for any Borel function h that does not involve $P \in \mathcal{P}$

$$\mathrm{E}(h(X)) = 0 \; \forall P \in \mathcal{P} \quad \Longrightarrow \quad h(t) = 0 \text{ a.e. } \mathcal{P}.$$

A slightly weaker condition, "bounded completeness", is defined as above, but only for bounded Borel functions h.

Let

$$\mathcal{P}_1 = \{\text{distributions with densities of the form } (\sqrt{2\pi}\sigma)^{-1} \exp(x^2/(2\sigma^2))\}.$$

(This is the $\mathrm{N}(0, \sigma^2)$ family.) It is clear that $\mathrm{E}(h(X)) = 0$ for $h(x) = x$, yet clearly it is not the case that $h(t) = 0$ a.e.. Hence, this family, the family of normals with known mean, is not complete.

With some work, we can see that the family

$$\mathcal{P}_2 = \{\text{distributions with densities of the form } (\sqrt{2\pi}\sigma)^{-1} \exp((x-\mu)^2/(2\sigma^2))\}$$

is complete. Note that $\mathcal{P}_1 \subset \mathcal{P}_2$; and $\mathcal{P}_2$ is complete, but $\mathcal{P}_1$ is not. This is a common situation.

Going in the opposite direction, let $\mathcal{P}_2$ be the family of distributions w.r.t. which E is defined and $\mathcal{P}_2$ is complete. Now let $\mathcal{P}_2 \subset \mathcal{P}_1$, where all distributions in $\mathcal{P}_1$ have common support. Then $\mathcal{P}_1$ is complete.

Completeness of a Statistic

Complete families are defined in terms of properties of any Borel function of a random variable that does not involve the particular distribution. Such a function is called a statistic. Some particular statistics, such as $T(X)$ in the definition of exponential families above, may be of interest, and a particular interesting property of a given statistic is the one of general statistics that defines a complete family. We say a statistic $T(X)$ is *complete* iff for any Borel function h that does not involve $P \in \mathcal{P}$

$$\mathrm{E}(h(T(X))) = 0 \; \forall P \in \mathcal{P} \Rightarrow h(T(x)) = 0 \text{ a.e. } \mathcal{P}.$$

As above slightly weaker condition, "bounded completeness of a statistic", is defined in a similar manner, but only for bounded Borel functions h.

We will later define completeness of a class of statistics (in terms of an optimality property called admissibility), and in that context, define minimal completeness of the class.

The statistic $T(X)$ in the expression for the PDF of a member of an exponential family is complete if the family is of full rank.

In a parametric-support family, there may be a complete statistic. If so, it is usually an extreme order statistic.

Notes

We have developed the concept of probability by first defining a measurable space, then defining a measure, and finally defining a special measure as a probability measure. Alternatively, the concept of probability over a given measurable space could be stated as axioms. In this approach, there would be four axioms: nonnegativity, additivity over disjoint sets, probability of 1 for the sample space, and equality of the limit of probabilities of a monotonic sequence of sets to the probability of the limit of the sets. The axiomatic development of probability theory is due to Kolmogorov in the 1920s and 1930s. In Kolmogorov (1956), he starts with a sample space and a collection of subsets and gives six axioms that characterize a probability space. (Four axioms are the same or similar to those above, and the other two characterize the collection of subsets as a σ-field.)

Although the measurable spaces of Sections 1.1.1 and D.2 (beginning on page 368) do not necessarily consist of real numbers, we defined real-valued functions (random variables) that would be the basis of further development of probability theory. From the axioms characterizing probability (or equivalently from the definition of the concept of a probability measure), we developed expectation and various unifying objects such as distributions of random variables.

In order to develop an idea of conditional probability and conditional distributions, however, we began from a different starting point; we defined conditional expectation and then defined conditional probability.

An alternate approach to developing a probability theory can begin with a sample space and random variables defined on it. (Recall our definition of random variables did not require a definition of probability.) From this beginning, we can base a development of probability theory on expectation, rather than on a probability measure as we have done in this chapter. (This would be somewhat similar to our development of conditional probability from conditional expectation in Section 1.5.) We characterize an expectation operator E on a random variable X (and X_1 and X_2) by four axioms:

1. If $X \geq 0$, then $\mathrm{E}(X) \geq 0$.
2. If c is a constant in $\mathbb{R}$, then $\mathrm{E}(cX_1 + X_2) = c\mathrm{E}(X_1) + \mathrm{E}(X_2)$.
3. $\mathrm{E}(1) = 1$.

4. If a sequence of random variables $\{X_n(\omega)\}$ increases monotonically to a limit $\{X(\omega)\}$, then $\mathrm{E}(X) = \lim_{n\to\infty} \mathrm{E}(X_n)$.

(In these axioms, we have assumed a scalar-valued random variable, although with some modifications, we could have developed the axioms in terms of random variables in $\mathbb{R}^d$.) From these axioms, after defining the probability of a set as

$$\Pr(A) = \mathrm{E}(\mathrm{I}_A(\omega)),$$

we can develop the same probability theory as we did starting from a characterization of the probability measure. An interesting text that takes this approach is Whittle (2000).

Markov Chains

There are many other interesting properties of Markov chains that follow from various properties of nonnegative matrices (see Gentle, 2007, in the general references). For more information on the properties of Markov chains, we refer the interested reader to a text on Markov chains, such as Norris (1997).

Inequalities

Pages 633 to 687 of DasGupta (2008), in the general references, is a very extensive compendium of inequalities. None are proved there, but each is accompanied by a reference to a proof.

The Exponential Class

Extensive discussions of exponential families are provided by Barndorff-Nielson (1978) and Brown (1986).

Exercises in Shao

- For practice and discussion
  1.12, 1.14, 1.30, 1.31, 1.36, 1.38, 1.51, 1.53, 1.55, 1.60, 1.70, 1.85, 1.91, 1.97, 1.128, 1.161, 2.9, 2.13, 2.19, 2.23
  (Solutions in Shao, 2005)
- To turn in
  1.4, 1.5, 1,8, 1.18, 1.23, 1.43, 1.58, 1.78, 1.90, 1.101, 1.102, 1.103, 1.127, 1.158, 2.3, 2.4, 2.8, 2.20, 2.28

Additional References

Barndorff-Nielson, O. E. (1978), *Information and Exponential Families in Statistical Theory*, Wiley, Chichester.

Brown, Lawrence D. (1986), *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*, Institute of Mathematical Statistics, Hayward, California.

DasGupta, Anirban (2000), Best constants in Chebyshev inequalities with various applications, *Metrika* **51**, 185–200.

Dharmadhikari, Sudhakar, and Kumar Joag-Dev (1988), *Unimodality, Convexity, and Applications*, Academic Press, New York.

Kolmogorov, A. N. (1956, translated from the German), *Foundations of the Theory of Probability*, Chelsea, New York.

Norris, J. R. (1997), *Markov Chains*, Cambridge University Press, Cambridge, United Kingdom.

Whittle, Peter (2000), *Probability via Expectation*, fourth edition, Springer, New York.

2

Basic Statistical Concepts (Shao Ch 2, Sec 4.3, Sec 5.1, Sec 5.5; TPE2 Ch 1, Ch 5; TSH3 Ch 1, Ch 8)

In this chapter we give a brief overview of statistical inference. We begin with a high-level consideration of the issues in statistical inference and what kinds of approaches may be intellectually or heuristically sound. We then formulate statistical analysis as a decision problem, which allows us to approach statistical methods as optimization problems.

Statistical inference is based on an observed sample. Often the precision with which we can state conclusions depends on the size of the sample, so for any statistical procedure, it is of interest to know how the precision improves with increasing sample size. Although we cannot have an infinite sample size, we often carry the mathematical analysis to the limit. Another reason that we often consider limiting cases is that asymptotic properties are often more mathematically tractable than properties for finite samples.

This chapter provides the basic approach that will be followed in later chapters. Definitions are given of many of the basic concepts that will be discussed more thoroughly later.

2.1 Inferential Information in Statistics

In statistics, we generally assume that we have a *sample* of observations $X_1, \ldots, X_n$ on a random variable X. A *random sample*, which we will usually just call a "sample", is a set of i.i.d. random variables. We will often use X to denote a random sample on the random variable X. (This may sound confusing, but it is always clear from the context.) A *statistic* is any function of X that does not involve any unobservable values.

We assume that the sample arose from some distribution P_θ, which is a member of some family of probability distributions $\mathcal{P}$. We fully specify the family $\mathcal{P}$ (it can be a very large family), but we assume some aspects of P_θ are unknown. (If the distribution P_θ that yielded the sample is fully known, while there may be some interesting questions about probability, there are no interesting statistical questions.) Our objective in statistical inference is

to determine a specific $P_\theta \in \mathcal{P}$, or some subfamily $\mathcal{P}_\theta \subset \mathcal{P}$, that could likely have generated the sample.

The distribution may also depend on other observable variables. In general, we assume we have observations $X_1, \ldots, X_n$ on X, together with associated observations on any related variable Z or z. We denote the observed values as $(x_1, z_1), \ldots, (x_n, z_n)$, or just as $x_1, \ldots, x_n$. In this context, a *statistic* is any function that does not involve any unobserved values.

In statistical inference, we distinguish observable random variables and "parameters", but we are not always careful in referring to parameters. We think of two kinds of parameters; "known" and "unknown". A statistic is a function of observable random variables that does not involve any unknown parameters.

2.1.1 Types of Statistical Inference

There are three different types of inference related to the problem of determining the specific $P_\theta \in \mathcal{P}$: *point estimation*, *hypothesis tests*, and *confidence sets*. Hypothesis tests and confidence sets are associated with probability statements that depend on P_θ.

In parametric settings, each type of inference concerns a parameter, θ, in a parameter space, $\Theta \subset \mathbb{R}^k$. If Θ is not a closed set, it is more convenient to consider the closure of Θ, $\overline{\Theta}$, because sometimes a good estimator may actually be outside of the open set Θ. (If Θ is closed, $\overline{\Theta}$ is the same set, so we can always just consider $\overline{\Theta}$.)

A related problem in inference is *prediction*, in which in addition to the random variable X with the probability triple $(\Omega, \mathcal{F}, P)$ we have a measurable function Y that maps $(\Omega, \mathcal{F}, P)$ to $(\Lambda, \mathcal{G})$, and, given an observed value of Y we wish to predict X. The problem of predicting X is to find a Borel function g such that $\mathrm{E}(g(Y))$ is "close to" $\mathrm{E}(X)$.

We make inferences using observations on X and any covariate.

The Basic Paradigm of Point Estimation

A real-valued (to me, that does not necessarily mean a scalar) observable random variable X has a distribution that depends in some way on a real-valued parameter θ that takes a value in the set Θ. We assume we have observations $X_1, \ldots, X_n$ on X, together with associated observations on any related variable Z or z, $(x_1, z_1), \ldots, (x_n, z_n)$, or just as $x_1, \ldots, x_n$.

The object to be estimated is called the *estimand*. Although it may be an underlying natural parameter, sometimes it is a Borel function of that parameter. Some authors, such as Shao, use the symbol ϑ to indicate a relationship to θ, but also to indicate that the estimand may be a function of θ. We will use $g(\theta)$, or sometimes $g(\theta; z)$ when there is a covariate z, to represent the estimand. We want to estimate $g(\theta)$ or $g(\theta; z)$ using observations on X and any covariate. We denote the estimator as $T(X)$, or $T(X, z)$. We think of T as

a rule or formula. We also use T to denote a decision in hypothesis testing. We also denote the rule as $\delta(X)$, or $\delta(X, z)$, especially if the rule is randomized.

Optimal Point Estimators

We seek an estimator with "good" properties. "Good" can be defined in several ways.

One of the most commonly required desirable properties is *unbiasedness*. The estimator as $T(X)$ is unbiased for $g(\theta)$ if $\mathrm{E}(T(X)) = g(\theta)$ for any $\theta \in \Theta$. Unbiasedness as we have just defined it is a uniform property of the expected value. The *bias* of $T(X)$ for estimating $g(\theta)$ is $\mathrm{E}(T(X)) - g(\theta)$.

We can also define other types of unbiasedness in terms of other aspects of a probability distribution. For example, an estimator whose median is the estimand is said to be *median-unbiased*.

Unbiasedness has different definitions for other types of statistical inference (testing, see page 108, and determining confidence sets, see page 112), but the meanings are similar.

If two estimators are unbiased, we would reasonably prefer one with smaller variance.

Another measure of the goodness of a scalar estimator is the mean-squared error or MSE,

$$\mathrm{MSE}(T(x)) = \mathrm{E}((T(X) - g(\theta))^2), \tag{2.1}$$

which is the square of the bias plus the variance:

$$\mathrm{MSE}(T(x)) = (\mathrm{E}(T(X)) - g(\theta))^2 + \mathrm{V}(T(X)).$$

C. R. Rao gives an example that causes us to realize that we often face a dilemma in finding a good estimate. Suppose we have n observations $X_1, \ldots, X_n$ from a distribution with mean μ_1 and finite standard deviation σ. We wish to estimate μ_1. An obvious estimator is the sample mean $\overline{X}$. (We will see that this is generally a good estimator under most criteria.) The MSE of $\overline{X}$ is σ^2/n. Now, suppose we have m observations $Y_1, \ldots, Y_m$ from a different distribution with mean $\mu_2 = \mu_1 + \delta\sigma$ and the same standard deviation σ. Let

$$T = (n\overline{X} + m\overline{Y})/(n + m),$$

so we have

$$\mathrm{E}((T - \mu_1)^2) = \frac{\sigma^2}{n+m}\left(1 + \frac{m^2\delta^2}{n+m}\right).$$

Now if $\delta^2 < m^{-1} + n^{-1}$, then

$$\mathrm{MSE}(T) < \mathrm{MSE}(\overline{X});$$

that is, in this case, the MSE is improved by using spurious observations. If $\delta < 1$, just using a single spurious observation improves the MSE.

While the MSE gives us some sense of how "close" the estimator is to the estimand, another way of thinking about closeness is terms of the probability that $|T(X) - g(\theta)|$ is less than some small value ϵ. This type of measure is called *Pitman closeness*. Given two estimators $T_1(X)$ and $T_2(X)$ of $g(\theta)$, we say that $T_1(X)$ is *Pitman-closer* than $T_2(X)$, if

$$\Pr\big(|T_1(x) - g(\theta)| \leq |T_2(x) - g(\theta)| \,\big|\theta\big) \;\geq\; \frac{1}{2} \tag{2.2}$$

for all $\theta \in \Theta$ and for some $\theta_0 \in \Theta$

$$\Pr\big(|T_1(x) - g(\theta)| < |T_2(x) - g(\theta)| \,\big|\theta_0\big) \;\geq\; \frac{1}{2}.$$

We say that $T_1(X)$ is the *Pitman-closest* estimator, if $T_1(X)$ is *Pitman-closer* than $T(X)$ for any other statistic $T(X)$.

Pitman closeness is affected more by the properties of the distribution in the region of interest, rather than by the behavior of statistics in the tail regions. Measures such as MSE, for example, may be unduly affected by the properties of the distribution in the tail regions.

Although Pitman closeness is a useful measure in evaluating an estimator, the measure lacks the desirable property of transitivity; that is, $T_1(X)$ may be *Pitman-closer* than $T_2(X)$ and $T_2(X)$ *Pitman-closer* than $T_3(X)$, but yet $T_3(X)$ may be *Pitman-closer* than $T_1(X)$. It is easy to construct an example to illustrate that this is possible. Rather than trying to construct a realistic distribution and statistics, let us just consider three independent random variables T_1, T_2, and T_3 and assign probability distributions to them (following Colin Blyth, 1972):

$\Pr(T_1 = 3) = 1.0$
$\Pr(T_2 = 1) = 0.4,\ \Pr(T_2 = 4) = 0.6$
$\Pr(T_3 = 2) = 0.6,\ \Pr(T_3 = 5) = 0.4$

We see that

$$\Pr(T_1 < T_2) = 0.6,\ \Pr(T_2 < T_3) = 0.64,\ \Pr(T_3 < T_1) = 0.6.$$

Efron (1975) gives an example of an otherwise "good" estimator that is not as close in the Pitman sense as a biased estimator. Consider the problem of estimating the mean μ in a normal distribution $\mathrm{N}(\mu, 1)$, given a random sample $X_1, \ldots, X_n$. The usual estimator, the sample mean $\overline{X}$, is unbiased and has minimum variance among all unbiased estimators. Consider, however, the estimator

$$T(X) = \overline{X} - \Delta_n(\overline{X}), \tag{2.3}$$

where

$$\Delta_n(u) = \frac{\min(u\sqrt{n}, \Phi(-u\sqrt{n})}{2\sqrt{n}}, \quad \text{for } u \geq 0, \tag{2.4}$$

in which $\Phi(\cdot)$ is the standard normal CDF. This "shrinkage" of $\overline{X}$ toward 0 yields an estimator that is Pitman-closer to the population mean μ than the

sample mean $\overline{X}$. On page 100, we will encounter a more dramatic example of the effect of shrinking the sample mean in a multivariate normal distributional model.

If the goodness of an estimator does not depend on the parameter, we say the estimator is *uniformly* good (and, of course, in this statement we would be more precise in what we mean by "good"). All discussions of statistical inference are in the context of some family of distributions, and when we speak of a "uniform" property, we mean a property that holds for all members of the family.

There are several approaches to estimation of $g(\theta)$. We generally assume a specific estimate of $g(\theta)$, say $\widehat{g(\theta)}$, results in a specific distribution for X, or at least a specific family of distributions, with CDF $P_{\widehat{g(\theta)}}$. A good estimation scheme is one that specifies a distribution of X that corresponds in some sense to the observed values of X. We start on the problem by defining some computable, heuristic estimation procedure, and then analytically study the properties of that procedure under various scenarios, that is, under different assumed distributions.

Prediction

In addition to the three different types of inference related to the problem of determining the specific $P_\theta \in \mathcal{P}$, we may also want to *predict* the value that a random variable will realize.

In the prediction problem, we have a random variable X with the probability triple $(\Omega, \mathcal{F}, P)$ and a measurable function Y that maps $(\Omega, \mathcal{F}, P)$ to $(\Lambda, \mathcal{G})$. Given an observed value of Y we wish to predict X; that is, to find a Borel function g such that $\mathrm{E}((g(Y))^2) < \infty$ and $\mathrm{E}(g(Y))$ is "close to" $\mathrm{E}(X)$. A useful measure of closeness in the prediction problem is the *mean squared prediction error* or MSPE:

$$\mathrm{MSPE}(g) = \mathrm{E}((X - g(Y))^2). \tag{2.5}$$

Conditional expectation plays a major role in prediction. If $\mathrm{E}(X^2) < \infty$, it may be of interest to determine the best predictor in the sense of minimizing the mean squared prediction error. Letting $\mathcal{T}$ be the class of all functions $g(Y)$ such that $\mathrm{E}((g(Y))^2) < \infty$ and assuming $\mathrm{E}(X^2) < \infty$, we expand the mean-squared prediction error in a manner similar to the operations in inequality (1.25) on page 23:

$$\begin{aligned}
\mathrm{E}((X-g(Y))^2) &= \mathrm{E}((X-\mathrm{E}(X|Y)+\mathrm{E}(X|Y)-g(Y))^2) \\
&= \mathrm{E}((X-\mathrm{E}(X|Y))^2)+\mathrm{E}((\mathrm{E}(X|Y)-g(Y))^2)+ \\
&\quad 2\mathrm{E}((X-\mathrm{E}(X|Y))(\mathrm{E}(X|Y)-g(Y))) \\
&= \mathrm{E}((X-\mathrm{E}(X|Y))^2)+\mathrm{E}((\mathrm{E}(X|Y)-g(Y))^2)+ \\
&\quad 2\mathrm{E}\left(\mathrm{E}((X-\mathrm{E}(X|Y))(\mathrm{E}(X|Y)-g(Y)))|Y\right) \\
&= \mathrm{E}((X-\mathrm{E}(X|Y))^2)+\mathrm{E}((\mathrm{E}(X|Y)-g(Y))^2) \\
&\geq \mathrm{E}((X-\mathrm{E}(X|Y))^2). \qquad (2.6)
\end{aligned}$$

Statements of Probability Associated with Statistics

Although much of the development of inferential methods emphasize the expected value of statistics, often it is useful to consider the probabilities of statistics being in certain regions. Pitman closeness is an example of the use of probabilities associated with estimators. Two other approaches involve the probabilities of various sets of values that the statistics may take on. These approaches lead to statistical *tests of hypotheses* and determination of *confidence sets*. These topics will be discussed in Section 2.4, and more thoroughly in later chapters.

2.1.2 Sufficiency, Ancillarity, Minimality, and Completeness

There are important properties of statistics, such as sufficiency and complete sufficiency, that determine the usefulness of those statistics in statistical inference.

sufficiency

Let X be a sample from a population $P \in \mathcal{P}$. A statistic $T(X)$ is *sufficient* for $P \in \mathcal{P}$ if and only if the conditional distribution of X given T does not depend on P.

In general terms, this involves the conditional independence from the parameter of the distribution of any other function of the random variable, given the sufficient statistic. Sufficiency depends on

- $\mathcal{P}$, the family of distributions w.r.t. which E is defined. If a statistic is sufficient for $\mathcal{P}$, it may not be sufficient for a larger family, $\mathcal{P}_1$, where $\mathcal{P} \subset \mathcal{P}_1$.

Sufficiency may allow reduction of data without sacrifice of information.

We can establish sufficiency by the factorization criterion:

A necessary and sufficient condition for a statistic T to be sufficient for a family $\mathcal{P}$ of distributions of a sample X dominated by a σ-finite measure ν is that there exist nonnegative Borel functions g_P and h, where h does not depend on P, such that

$$\mathrm{d}P/\mathrm{d}\nu(x) = g_P(T(x))h(x) \quad \text{a.e. } \nu. \qquad (2.7)$$

An important consequence of sufficiency in an estimation problem with convex loss is the Rao-Blackwell theorem (see Section 2.3.2).

When the density can be written in the separable form $c(\theta)f(x)$, unless $c(\theta)$ is a constant, the support must be a function of θ, and a sufficient statistic must be an extreme order statistic. When the support depends on the parameter, the extreme order statistic(s) at the boundary of the support determined by the parameter carry the full information about the parameter.

ancillarity

Ancillarity is, in a way, the opposite of sufficiency: A statistic $U(X)$ is called *ancillary* for P (or θ) if the distribution of $U(X)$ does not depend on P (or θ).

If $\mathrm{E}(U(X))$ does not depend on P (or θ), then $U(X)$ is said to be *first-order ancillary* for P (or θ)

nuisance parameter

Often a probability model contains a parameter of no interest for inference. Such a parameter is called a *nuisance parameter*. A statistic to be used for inferences about the parameters of interest should be *ancillary* for a nuisance parameter.

minimal sufficiency

Let T be a given sufficient statistic for $P \in \mathcal{P}$. The statistic T is *minimal sufficient* if for any sufficient statistic for $P \in \mathcal{P}$, S, there is a measurable function h such that $T = h(S)$ a.s. $\mathcal{P}$.

Minimal sufficiency has a heuristic appeal: it relates to the greatest amount of data reduction.

An easy way of establishing minimality when the range does not depend on the parameter is by use of the following facts:

Let $\mathcal{P}$ be a family with densities $p_0, p_1, \ldots, p_k$, all with the same support. The statistic

$$T(X) = \left(\frac{p_1(X)}{p_0(X)}, \ldots, \frac{p_k(X)}{p_0(X)}\right) \tag{2.8}$$

is minimal sufficient.

This follows from the following corollary of the factorization theorem:

A necessary and sufficient condition for a statistic T to be sufficient for a family $\mathcal{P}$ of distributions of a sample X dominated be a σ-finite measure ν is that for any two densities p_1 and p_2 in $\mathcal{P}$, the ratio $p_1(x)/p_2(x)$ is a function only of $T(x)$.

Then, the other important fact is

Let $\mathcal{P}$ be a family of distributions with the common support, and let $\mathcal{P}_0 \subset \mathcal{P}$. If T is minimal sufficient for $\mathcal{P}_0$ and is sufficient for $\mathcal{P}$, then it is minimal sufficient for $\mathcal{P}$.

We see this by considering any statistic U that is sufficient for $\mathcal{P}$. It must also be sufficient for $\mathcal{P}_0$, and since T is minimal sufficient for $\mathcal{P}_0$, T is a function of U.

completeness

A sufficient statistic T is particularly useful in a complete family or a boundedly complete family of distributions. In this case, for every Borel (bounded) function h that does not involve P,

$$\mathrm{E}_P(h(T)) = 0 \,\forall\, P \in \mathcal{P} \Rightarrow h(t) = 0 \text{ a.e. } \mathcal{P}.$$

In a complete family, we often refer to the completeness of a statistic, rather than the completeness of the family. We often call a sufficient statistic in a complete family, a "complete sufficient" statistic.

Complete sufficiency depends on

- $\mathcal{P}$, the family of distributions w.r.t. which E is defined. If a statistic is complete and sufficient with respect to $\mathcal{P}$, and if it is sufficient for $\mathcal{P}_1$, where $\mathcal{P} \subset \mathcal{P}_1$ and all distributions in $\mathcal{P}_1$ have common support, then it is complete and sufficient for $\mathcal{P}_1$, because in this case, the condition a.s. $\mathcal{P}$ implies the condition a.s. $\mathcal{P}_1$.

Complete sufficiency is useful in UMVUE and for establishing independence using Basu's theorem.

It is important to remember that completeness and sufficiency are different properties; you can have either one without the other.

Sufficiency relates to a statistic and a sample. There is always a sufficient statistic: the sample itself.

There may or may not be a complete statistic within a given family.

If there is a complete statistic and it is sufficient, then it is minimal sufficient. That is, completeness implies minimality.

Complete sufficiency implies minimal sufficiency, but minimal sufficiency does not imply completeness. Consider a sample X of size 1 from $\mathrm{U}(\theta, \theta+1)$. Clearly, X is minimal sufficient. Any bounded periodic function $h(x)$ with period 1 that is not a.e. 0 serves to show that X is not complete. Let $h(x) = \sin(2\pi x)$. Then

$$\mathrm{E}(h(X)) = \int_\theta^{\theta+1} \mathrm{d}x = 0.$$

Clearly, however $h(X)$ is not 0 a.e., so X is not complete.

Basu's Theorem

Complete sufficiency, ancillarity, and independence are related.

Basu's theorem (Theorem 2.4 in Shao) states that if T is a boundedly *complete sufficient* statistic for θ, and if U is ancillary for θ, then T and U are independent.

An interesting example with $\mathrm{U}(\theta - 1/2,\ \theta + 1/2)$ shows the importance of completeness in Basu's theorem. This example also shows that minimality does not imply completeness. Let $X_1, \ldots, X_n$, with $n \geq 2$, be a random sample from $\mathrm{U}(\theta - 1/2,\ \theta + 1/2)$. It is clear that $T = \{X_{(1)}, X_{(n)}\}$ is sufficient; in fact, it is minimal sufficient. Now consider $U = X_{(n)} - X_{(1)}$,

which we easily see is ancillary. It is clear that T and U are not independent (U is a function of T).
Writing $U = h(T)$, where h is a measurable function, we can see the T is not complete (although it is minimal.)
If T were complete, then Basu's theorem would say that T and U are independent.

Sufficiency, Minimality, and Completeness in Various Families

We can use general properties of specific families of distributions to establish properties of statistics quickly and easily.

Complete sufficiency is often easy to show in exponential family or in distributions whose range depends on θ in a simple way. (We can relate any such range-dependent distribution to $\mathrm{U}(\theta_1, \theta_2)$.)

In general, proof of sufficiency is often easy, but proof of minimality or completeness is often difficult. We often must rely on the awkward use of the definitions of minimality and completeness. Completeness of course implies minimality.

Truncation families have simple range dependencies. A distribution in any of these families has a PDF in the general form

$$p_\theta(x) = c(\theta) f(x) \mathrm{I}_{S(\theta)}(x).$$

The two most useful examples of distributions whose support depends on the parameter are the uniform $\mathrm{U}(0, \theta)$ and the exponential $\mathrm{E}(\eta, 1)$. Many other distributions can be transformed into these; and, in fact, they can be transformed into each other. If $X_1, \ldots, X_n$ are i.i.d. $\mathrm{E}(\eta, 1)$, and $Y_i = \mathrm{e}^{-X_i}$ and $\theta = \mathrm{e}^{-\eta}$, then $Y_1, \ldots, Y_n$ are i.i.d. $\mathrm{U}(0, \theta)$; hence if we can handle one problem, we can handle the other. We can also handle distributions like $\mathrm{U}(\theta_1, \theta_2)$ and $\mathrm{E}(a, b)$, as well as some other related distributions, such as a shifted gamma.

We can show *completeness* using the fact that

$$\int_A |f| \, \mathrm{d}\mu = 0 \iff f = 0 \text{ a.e. on } A.$$

Another result we often need in going to a multiparameter problem is Fubini's theorem.

The sufficient statistic in the simple univariate case where $S(\theta) = (\theta_1, \theta_2)$ is $T(X) = (X_{(1)}, X_{(n)})$, as we can see from the the factorization theorem by writing the joint density of a sample as

$$c(\theta) f(x) \mathrm{I}_{(x_{(1)}, x_{(n)})}(x).$$

For example, for a distribution such as $\mathrm{U}(0, \theta)$ we see that $X_{(n)}$ is sufficient by writing the joint density of a sample as

$$\frac{1}{\theta} \mathrm{I}_{(0, x_{(n)})}.$$

The properties of a specific family of distributions are useful in identifying optimal methods of statistical inference. Exponential families are particularly useful for finding UMVU estimators. A group family is useful in identifying equivariant and invariant statistical procedures.

2.2 Statistical Inference: Approaches and Methods

If we assume that we have a random sample of observations $X_1, \ldots, X_n$ on a random variable X from some distribution P_θ, which is a member of some family of probability distributions $\mathcal{P}$, our objective in statistical inference is to determine a specific $P_\theta \in \mathcal{P}$, or some subfamily $\mathcal{P}_\theta \subset \mathcal{P}$, that could likely have generated the sample.

How should we approach this problem?

Five Approaches to Statistical Inference

Five approaches to statistical inference are

- use of the empirical cumulative distribution function (ECDF)
  for example, method of moments
- use of a likelihood function
  for example, maximum likelihood
- fitting expected values
  for example, least squares
- fitting a probability distribution
  for example, maximum entropy
- definition and use of a loss function
  for example, uniform minimum variance unbiased estimation.

We will briefly discuss the first four of these approaches in the following four subsections. The "decision theory" approach to statistical inference is based on a loss function, and we will discuss this important approach in Section 2.3.

Sometimes we must use approximations or second-order estimation in statistical inference. We discuss this briefly in Section 2.2.6, and later in Section 2.5, we discuss one type of approximate inference, asymptotic inference in more detail.

2.2.1 The Empirical Cumulative Distribution Function

From observations on a random variable, $X_1, \ldots, X_n$, we can form an empirical cumulative distribution function, or ECDF, that corresponds in a natural way with the CDF of the random variable.

For the sample, $X_1, \ldots, X_n$, the ECDF is defined as

$$P_n(x) = \frac{\#\{X_i \leq x\}}{n}. \tag{2.9}$$

The ECDF is a random simple function, and sometimes it is appropriate to treat the ECDF as a random variable. It is clear that the ECDF conditional on a given sample is itself a CDF. (Conditionally it is not a random variable.) It has the three properties that define a CDF:

- $\lim_{x\to-\infty} P_n(x) = 0$ and $\lim_{x\to\infty} P_n(x) = 1$.
- $P_n(x)$ is monotone increasing.
- $P_n(x)$ is continuous from the right.

The ECDF defines a discrete population with mass points at each value in the sample.

The ECDF is particularly useful in nonparametric inference.

Plug-In Measures

As discussed in Section 1.1.4, many distribution parameters and other measures can be represented as a statistical function, that is, as a functional of the CDF. The functional of the CDF that defines a parameter defines a plug-in estimator of that parameter when the functional is applied to the ECDF. A functional of a population distribution function, $\Theta(P)$, defining a parameter θ can usually be expressed as

$$\begin{aligned}\theta &= \Theta(P)\\ &= \int g(y)\,\mathrm{d}P(y).\end{aligned}$$

The plug-in estimator T is the same functional of the ECDF:

$$\begin{aligned}T &= T(P_n)\\ &= \Theta(P_n)\\ &= \int g(y)\,\mathrm{d}P_n(y).\end{aligned}$$

(In both of these expressions, we are using the integral in a general sense. In the second expression, the integral is a finite sum. It is also a countable sum in the first expression if the random variable is discrete. Note also that we use the same symbol to denote the functional and the random variable.)

We may base inferences on properties of the distribution with CDF P by identifying the corresponding properties of the ECDF P_n. In some cases, it may not be clear what we mean by "corresponding". If a property of a distribution can be defined by a functional on the CDF, the corresponding property is the same functional applied to the ECDF. This is the underlying idea of the method of moments, for example. In the method of moments, sample moments, which are moments of the discrete population represented

by the sample, are used for making inferences about population moments. The method-of-moments estimator of the population mean, $\mathrm{E}(X)$, is the sample mean, $\overline{X}$. The thing to be estimated is the functional M in equation (1.10), and the estimator is M applied to the ECDF:

$$M(P_n) = \sum X_i P_n(X_i).$$

The plug-in estimator $\Theta(P_n)$ in general is not unbiased for the associated statistical function $\Theta(P)$. A simple example is the variance, $\Theta(P) = \sigma^2 = \int (x - \int x \, \mathrm{d}P)^2 \, \mathrm{d}P$. The plug-in estimator $\Theta(P_n)$, which in this case is also a method-of-moments estimator, is $(n-1)S^2/n$, where

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

is the usual sample variance.

On the other hand, the plug-in estimator may have smaller MSE than an unbiased estimator, and, in fact, that is the case for the plug-in estimator of σ^2. Also, plug-in estimators often have good limiting and asymptotic properties, as we might expect based on convergence properties of the ECDF.

Convergence of the ECDF

The ECDF is one of the most useful statistics, especially in nonparametric and robust inference. It is essentially the same as the set of order statistics, so like them, it is a sufficient statistic. Its distribution at a point is binomial, and so its pointwise properties are easy to see. Its global relationship to the most fundamental measure of a probability model, however, accounts for its usefulness. The basic facts regard the convergence of the sup distance of the ECDF from the CDF, $\rho_\infty(F_n, F)$, to zero.

The Dvoretzky/Kiefer/Wolfowitz inequality provides a bound for the probability that the sup distance of the ECDF from the CDF exceeds a given value. In one-dimension, for any positive z, there is a positive constant C that does not depend on F, z, or n, such that

$$\Pr(\rho_\infty(F_n, F) > z) \leq C\mathrm{e}^{-2nz^2}. \tag{2.10}$$

Of course, in any event, $\rho_\infty(F_n, F) < 1$.

This inequality is useful in proving convergence results for the ECDF. Some important results are given in Theorems 5.1 and 5.2 in Shao. A simple and more common statement of the convergence is the so-called Glivenko-Cantelli theorem.

When we consider the convergence of metrics on functions, the arguments of the functions are sequences of random variables, yet the metric integrates

out the argument. One way of handling this is just to use the notation F_n and F, as Shao does. Another way is to use the notation $F_n(x, \omega)$ to indicate that the ECDF is a random variable, yet to allow it to have an argument just as the CDF does. I will use this notation occasionally, but usually I will just write $F_n(x)$. The randomness comes in the definition of $F_n(x)$, which is based on the random sample.

Theorem 2.1 (Glivenko-Cantelli) *If $X_1, \ldots, X_n$ be i.i.d. with CDF F and ECDF F_n, and if $D_n(\omega) = \rho_\infty(F_n, F) = \sup_x(|F_n(x, \omega) - F(x)|)$, then $D_n(\omega) \longrightarrow 0$ wp1.*

Proof. First, note by the SLLN and the binomial distribution of F_n, $\forall$ (fixed) x, $F_n(x, \omega) \longrightarrow F(x)$ wp1; that is,

$$\lim_{n\to\infty} F_n(x, \omega) = F(x)$$

$\forall x$, except $x \in A_x$, where $\Pr(A_x) = 0$.

The problem here is that A_x depends on x and so there are uncountably many such sets. The probability of their union may possibly be positive. So we must be careful.

We will work on the CDF and ECDF from the other side of x (the discontinuous side). Again, by the SLLN, we have

$$\lim_{n\to\infty} F_n(x-, \omega) = F(x-)$$

$\forall x$, except $x \in B_x$, where $\Pr(B_x) = 0$.
Now, let

$$\phi(u) = \inf\{x \ ; \ u \leq F(x)\} \quad \text{for } 0 < u \leq 1.$$

(Notice $F(\phi(u)-) \leq u \leq F(\phi(u))$. Sketch the picture.)

Now consider $x_{m,k} = \phi(k/m)$ for positive integers m and k with $1 \leq k \leq m$. (There are countably many $x_{m,k}$, and so when we consider $F_n(x_{m,k}, \omega)$ and $F(x_{m,k})$, there are countably many null-probability sets, $A_{x_{m,k}}$ and $B_{x_{m,k}}$, where the functions differ in the limit.)

We immediately have the three relations:

$$F(x_{m,k}-) - F(x_{m,k-1}) \leq m^{-1}$$

$$F(x_{m,1}-) \leq m^{-1}$$

and

$$F(x_{m,m}) \geq 1 - m^{-1},$$

and, of course, F is nondecreasing.

Now let $D_{m,n}(\omega)$ be the maximum over all $k = 1, \ldots, m$ of

$$|F_n(x_{m,k}, \omega) - F(x_{m,k})|$$

and

$$|F_n(x_{m,k}-,\omega) - F(x_{m,k}-)|.$$

(Compare $D_n(\omega)$.)
We now consider three ranges for x:

$$\begin{array}{c}(-\infty\ ,\ x_{m,1}) \\ [x_{m,k-1}\ ,\ x_{m,k}) \text{ for } k = 1,\ldots,m \\ [x_{m,m}\ ,\ \infty)\end{array}$$

Consider $[x_{m,k-1} \le x < x_{m,k})$. In this interval,

$$\begin{aligned} F_n(x,\omega) &\le F_n(x_{m,k}-,\omega) \\ &\le F(x_{m,k}-) + D_{m,n}(\omega) \\ &\le F(x) + m^{-1} + D_{m,n}(\omega) \end{aligned}$$

and

$$\begin{aligned} F_n(x,\omega) &\ge F_n(x_{m,k-1},\omega) \\ &\ge F(x_{m,k-1}) - D_{m,n}(\omega) \\ &\ge F(x) - m^{-1} - D_{m,n}(\omega) \end{aligned}$$

Hence, in these intervals, we have

$$\begin{aligned} D_{m,n}(\omega) + m^{-1} &\ge \sup_x |F_n(x,\omega) - F(x)| \\ &= D_n(\omega). \end{aligned}$$

We can get this same inequality in each of the other two intervals.

Now, $\forall m$, except on the unions over k of $A_{x_{m,k}}$ and $B_{x_{m,k}}$, $\lim_n D_{m,n}(\omega) = 0$, and so $\lim_n D_n(\omega) = 0$, except on a set of probability measure 0 (the countable unions of the $A_{x_{m,k}}$ and $B_{x_{m,k}}$.) Hence, we have the convergence wp1; i.e., a.s. convergence. ∎

The Bootstrap Principle

The ECDF plays a major role in a bootstrap method, in which the population of interest is studied by sampling from the population defined by a given sample from the population of interest. This is a method of *resampling*.

Resampling methods involve the use of many samples, each taken from a single sample that was taken from the population of interest. Inference based on resampling makes use of the conditional sampling distribution of a new sample (the "resample") drawn from a given sample. Statistical functions on the given sample, a finite set, can easily be evaluated. Resampling methods therefore can be useful even when very little is known about the underlying distribution.

A basic idea in bootstrap resampling is that, because the observed sample contains all the available information about the underlying population, the

observed sample can be considered *to be* the population; hence, the distribution of any relevant test statistic can be simulated by using random samples from the "population" consisting of the original sample.

Suppose that a sample $y_1, \ldots, y_n$ is to be used to estimate a population parameter, θ. For a statistic T that estimates θ, as usual, we wish to know the sampling distribution so as to correct for any bias in our estimator or to set confidence intervals for our estimate of θ. The sampling distribution of T is often intractable in applications of interest.

A basic bootstrapping method formulated by Efron (1979) uses the discrete distribution represented by the sample to study the unknown distribution from which the sample came. The basic tool is the empirical cumulative distribution function. The ECDF is the CDF of the finite population that is used as a model of the underlying population of interest.

For a parameter θ of a distribution with CDF P defined as $\theta = \Theta(P)$, we can form a plug-in estimator T as $T = T(P_n)$. Various properties of the distribution of T can be estimated by use of "bootstrap samples", each of the form $\{y_1^*, \ldots, y_n^*\}$, where the y_i^*'s are chosen from the original y_i's with replacement.

We define a *resampling vector*, p^*, corresponding to each bootstrap sample as the vector of proportions of the elements of the original sample in the given bootstrap sample. The resampling vector is a realization of a random vector P^* for which nP^* has an n-variate multinomial distribution with parameters n and $(1/n, \ldots, 1/n)$. The resampling vector has random components that sum to 1. For example, if the bootstrap sample $(y_1^*, y_2^*, y_3^*, y_4^*)$ happens to be the sample (y_2, y_2, y_4, y_3), the resampling vector p^* is

$$(0,\ 1/2,\ 1/4,\ 1/4).$$

The bootstrap replication of the estimator T is a function of p^*, $T(p^*)$. The resampling vector can be used to estimate the variance of the bootstrap estimator. By imposing constraints on the resampling vector, the variance of the bootstrap estimator can be reduced.

The *bootstrap principle* involves repeating the process that leads from a population CDF to an ECDF. Taking the ECDF P_n to be the CDF of a population, and resampling, we have an ECDF for the new sample, $P_n^{(1)}$. (In this notation, we could write the ECDF of the original sample as $P_n^{(0)}$.) The difference is that we know more about $P_n^{(1)}$ than we know about P_n. Our knowledge about $P_n^{(1)}$ comes from the simple discrete uniform distribution, whereas our knowledge about P_n depends on knowledge (or assumptions) about the underlying population.

The bootstrap resampling approach can be used to derive properties of statistics, regardless of whether any resampling is done. Most common uses of the bootstrap involve computer simulation of the resampling; hence, bootstrap methods are usually instances of computational inference.

2.2.2 Likelihood

Given a sample $X_1, \ldots, X_n$ from distributions with probability densities $p_i(x)$, where all PDFs are defined with respect to a common σ-finite measure, the *likelihood function* is

$$L_n(p_i\,;\,X) = \prod_{i=1}^{n} p_i(X_i). \tag{2.11}$$

(Any nonnegative function proportional to $L_n(p_i\,;\,X)$ is a likelihood function, but it is common to speak of $L_n(p_i\,;\,X)$ as "the" likelihood function.) We can view the sample either as a set of random variables or as a set of constants, the realized values of the random variables, in which case we usually use lower-case letters.

The *log-likelihood function* is the log of the likelihood:

$$l_{L_n}(p_i\,;\,x) = \log L_n(p_i \,|\, x_i), \tag{2.12}$$

It is a sum rather than a product.

The n subscript serves to remind us of the sample size, and this is often very important in use of the likelihood or log-likelihood function particularly because of their asymptotic properties. We often drop the n subscript, however.

In many cases of interest, the sample is from a single parametric family. If the PDF is $p(x\,;\,\theta)$ then the likelihood and log-likelihood functions are written as

$$L(\theta\,;\,x) = \prod_{i=1}^{n} p(x_i\,;\,\theta), \tag{2.13}$$

and

$$l(\theta\,;\,x) = \log L(\theta\,;\,x). \tag{2.14}$$

The Parameter Is the Variable

Note that the likelihood is a function of θ for a given x, while the PDF is a function of x for a given θ. We sometimes write the expression for the likelihood without the observations: $L(\theta)$. I like to think of the likelihood as a function of some dummy variable t, and write $L(t\,;\,x)$ or $l(t\,;\,x)$. While if we think of θ as a fixed, but unknown, value, it does not make sense to think of a function of that particular value, and if we have an expression in terms of that value, it does not make sense to perform operations such as differentiation with respect to that quantity.

The likelihood function arises from a probability density, but it is not a probability density function. It does not in any way relate to a "probability" associated with the parameters or the model.

Although non-statisticians will often refer to the "likelihood of an observation", in statistics, we use the term "likelihood" to refer to a model or a distribution *given observations*.

In a multiparameter case, we may be interested in only some of the parameters. There are two ways of approaching this, use of a profile likelihood or of a conditional likelihood.

If $\theta = (\theta_1, \theta_2)$, if θ_2 is fixed, the likelihood $L(\theta_1\,;\,\theta_2, x)$ is called a *profile likelihood* or *concentrated likelihood* of θ_1 for given θ_2 and x.

If the PDFs can be factored so that one factor includes θ_2 and some function of the sample, $S(x)$, and the other factor, given $S(x)$, is free of θ_2, then this factorization can be carried into the likelihood. Such a likelihood is called a *conditional likelihood* of θ_1 given $S(x)$.

Maximum Likelihood Estimation

The *maximum likelihood estimate* (MLE) of θ, $\hat{\theta}$, is defined as

$$\hat{\theta} = \arg\max_{\theta \in \overline{\Theta}} L(\theta\,;\,x). \tag{2.15}$$

The MLE in general is not unbiased for its estimand. A simple example is the MLE of the variance σ^2 in a normal distribution with unknown mean. The MLE for σ^2 in a normal distribution with unknown mean is the same as the plug-in estimator or method-of-moments estimator, $(n-1)S^2/n$, where S is the usual sample variance (the sum of squares divided by $n-1$). Note that the plug-in estimator (or method-of-moments estimator) is not based on an assumed underlying distribution, but the MLE is.

On the other hand, the MLE may have smaller MSE than an unbiased estimator, and, in fact, that is the case for the MLE of σ^2 in the case of a normal distribution with unknown mean.

Score Function

In statistical inference, we use the information in how the likelihood or log-likelihood would vary if θ were to change. For a likelihood function (and hence, a log-likelihood function) that is differentiable with respect to the parameter, a function that represents this change and plays an important role in statistical inference is the *score function*:

$$s_n(\theta\,;\,x) = \frac{\partial l(\theta\,;\,x)}{\partial \theta}. \tag{2.16}$$

Likelihood Equation

In statistical estimation, the point at which the likelihood attains its maximum (which is, of course, the same point at which the log-likelihood attains its maximum) is of interest. We will consider this approach to estimation more thoroughly in Chapter 5.

If the likelihood is differentiable with respect to the parameter, the roots of the score function are of interest. The score function equated to zero,

$$\frac{\partial l(\theta\,;\,x)}{\partial \theta} = 0, \tag{2.17}$$

is called the *likelihood equation*. The derivative of the likelihood equated to zero, $\partial L(\theta\,;\,x)/\partial\theta = 0$, is also called the likelihood equation.

Equation (2.17) is an *estimating equation*; that is, its solution, if it exists, is an estimator. (*Note that it is not necessarily MLE; it is a root of the likelihood equation, or RLE.* We will see in Chapter 5 that RLEs have desirable asymptotic properties.)

It is often useful to define an estimator as the solution of some estimating equation. We will see other examples of estimating equations in subsequent sections.

Likelihood Ratio

When we consider two different distributions for a sample x, we have two different likelihoods, say L_0 and L_1. (Note the potential problems in interpreting the subscripts; here the subscripts refer to the two different distributions. For example L_0 may refer to $L(\theta_0 \,|\, x)$ in a notation consistent with that used above.) In this case, it may be of interest to compare the two likelihoods in order to make an inference about the two possible distributions. A simple comparison, of course, is the ratio, and indeed

$$\frac{L(\theta_0\,;\,x)}{L(\theta_1\,;\,x)}, \tag{2.18}$$

or L_0/L_1 in the simpler notation, is called the *likelihood ratio* with respect to the two possible distributions. Although in most contexts we consider the likelihood to be a function of the parameter for given, fixed values of the observations, it may also be useful to consider the likelihood ratio to be a function of x. On page 61, we defined a family of distributions based on their having a "monotone" likelihood ratio. Monotonicity in this case is with respect to a function of x. In a family with a monotone likelihood ratio, for some scalar-valued function $y(x)$ and for any $\theta_1 < \theta_0$, the likelihood ratio is a nondecreasing function of $y(x)$ for all values of x for which $f_{\theta_1}(x)$ is positive.

The most important use of the likelihood ratio is as the basis for a statistical test.

Under certain conditions that we will detail later, with L_0 and L_1, with corresponding log-likelihoods l_0 and l_1, based on a random variable (that is, $L_i = L(p_i\,;\,X)$, instead of being based on a fixed x), the random variable

$$\begin{aligned} \lambda &= -2\log\left(\frac{L_0}{L_1}\right) \\ &= -2(l_0 - l_1) \end{aligned} \tag{2.19}$$

has an approximate chi-squared distribution with degrees of freedom whose number depends on the numbers of parameters. (We will discuss this more fully in Chapter 6.)

This quantity in a different setting is also called the *deviance*. We encounter the deviance in the analysis of generalized linear models, as well as in other contexts.

The likelihood ratio, or the log of the likelihood ratio, plays an important role in statistical inference. Given the data x, the log of the likelihood ratio is called the *support* of the hypothesis that the data came from the population that would yield the likelihood L_0 versus the hypothesis that the data came from the population that would yield the likelihood L_1. The support clearly is relative and ranges over $\mathbb{R}$. The support is also called the *experimental support*.

Likelihood Principle

The *likelihood principle* in statistical inference asserts that all of the information which the data provide concerning the relative merits of two hypotheses (two possible distributions that give rise to the data) is contained in the likelihood ratio of those hypotheses and the data. An alternative statement of the likelihood principle is that if for x and y,

$$\frac{L(\theta\,;\,x)}{L(\theta\,;\,y)} = c(x,y) \quad \forall\theta,$$

where $c(x,y)$ is constant for given x and y, then any inference about θ based on x should be in agreement with any inference about θ based on y.

2.2.3 Fitting Expected Values

Given a random sample $X_1,\ldots,X_n$ from distributions with probability densities $p(x_i;\theta)$, where all PDFs are defined with respect to a common σ-finite measure, if we have that $\mathrm{E}(X_i) = g_i(\theta)$, a reasonable approach to estimation of θ may be to choose a value $\hat{\theta}$ that makes the differences $\mathrm{E}(X_i) - g_i(\theta)$ close to zero.

We must define the sense in which the differences are close to zero. A simple way to do this is to define a nonnegative scalar-valued Borel function of scalars, $\rho(u,v)$, that is increasing in the absolute difference of its arguments. One simple choice is $\rho(u,v) = (u-v)^2$. We then define

$$S_n(\theta,x) = \sum_{i=1}^{n} \rho(x_i,\theta). \tag{2.20}$$

A reasonable estimator is

$$\hat{\theta} = \arg\min_{\theta \in \overline{\Theta}} S_n(\theta, x). \tag{2.21}$$

Compare this with the maximum likelihood estimate of θ, defined in equation (2.15).

If the X_i are i.i.d., then all $g_i(\theta)$ are the same, say $g(\theta)$.

In common applications, we have *covariates*, $Z_1, \ldots, Z_n$, and the $\mathrm{E}(X_i)$ have a constant form that depends on the covariate: $\mathrm{E}(X_i) = g(Z_i, \theta)$.

As with solving the maximization of the likelihood, the solution to the minimization problem (2.21) may be obtained by solving

$$\frac{\partial S_n(\theta\,;\,x)}{\partial \theta} = 0. \tag{2.22}$$

2.2.4 Fitting Probability Distributions

In an approach to statistical inference based on information theory, the true but unknown distribution is compared with information in the sample. The focus is on "information" of "entropy", in the sense discussed on page 10. The basic quantity is of the form $\mathrm{E}(-\log(\mathrm{d}P))$. The principle underlying methods of statistical inference using these concepts and quantities is called *maximum entropy*.

****

$$d(P, Q) = \int_{\mathbb{R}} \phi\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}Q, \tag{2.23}$$

if it exists, is called the *ϕ-divergence* from Q to P. The ϕ-divergence is also called the f-divergence.

The ϕ-divergence is in general not a metric because it is not symmetric. One function is taken as the base from which the other function is measured. The expression often has a more familiar form if both P and Q are dominated by Lebesgue measure and we write $p = \mathrm{d}P$ and $q = \mathrm{d}Q$.

A specific instance of ϕ-divergence is the *Kullback-Leibler measure*,

$$\int_{\mathbb{R}} p(x) \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x. \tag{2.24}$$

(Recall from page 26 that this quantity is nonnegative.)

functionals ***

***** Fisher information

***** move some material from SectionD.2.6 in 00b_basicmeasure.inc

2.2.5 Estimating Equations

Equation (2.22) is an *estimating equation*; that is, its solution, if it exists, is an estimator. Likewise, equation (2.17) is an estimating equation. *Note that*

the solution to equation (2.22) is not necessarily the solution to the minimization problem (2.21), nor is the solution to equation (2.17) necessarily a solution to the maximization problem (2.15). They are both merely roots of estimating equations. We will consider some asymptotic properties of solutions to estimating equations in Section 2.5.1 (consistency) and in Section 5.3.4 (asymptotic normality).

A common example of equation (2.22) and equation (2.17), in which we have covariates, is the set of "normal equations", encountered in linear regression.

Generalized Estimating Equations

We have called the likelihood equation, equation (2.17), and equation (2.22) estimating equations. Such equations arise often in statistical inference. There are also several modifications of the basic equations; for example, sometimes we cannot form a tractable likelihood, so we form some kind of "quasi-likelihood". We therefore consider a generalized class of estimating equations.

We consider an independent sample $X_1, \ldots, X_n$ of random vectors with orders $d_1, \ldots, d_n$, with $\sup d_i < \infty$. We assume the distributions of the X_i are defined with respect to a common parameter $\theta \in \Theta \subset \mathbb{R}^k$. We now define Borel functions $\psi_i(X_i, \gamma)$ and let

$$s_n(\gamma) = \sum_{i=1}^{n} \psi_i(X_i, \gamma) \quad \gamma \in \Theta. \tag{2.25}$$

We call

$$s_n(\gamma) = 0 \tag{2.26}$$

a *generalized estimating equation* (GEE) and its root(s) a GEE estimator. If we take $\psi_i(X_i, \gamma)$ as $\partial\rho(X_i, \gamma)/\partial\gamma$ note the similarity to equation (2.22).

The GEE is usually chosen so that

$$\mathrm{E}_\theta(s_n(\theta)) = 0. \tag{2.27}$$

The normal equations can serve as a prototype of a GEE. Notice that equation (2.27) holds for the left-hand side of the normal equations.

2.2.6 "Approximate" Inference

When the exact distribution of a statistic is known (based, of course, on an assumption of a given underlying distribution of a random sample), use of the statistic for inferences about the underlying distribution is called exact inference.

Often the exact distribution of a statistic is not known, or is too complicated for practical use. In that case, we may resort to approximate inference. There are basically three types of approximate inference.

One type occurs when a simple distribution is very similar to another distribution. For example, the Kumaraswamy distribution (with PDF $\alpha\beta x^{\alpha-1}(1-x)^{\beta-1}$ over $[0,1]$) may be used as an approximation to the beta distribution.

Another type of approximate inference, called computational inference, is used when an unknown distribution can be simulated by resampling of the given observations.

Asymptotic inference is probably the most commonly used type of approximate inference. In asymptotic approximate inference we are interested in the properties of T_n as the sample size increases. We focus our attention on the sequence $\{T_n\}$ for $n = 1, 2, \ldots$, and, in particular, consider the properties of $\{T_n\}$ as $n \to \infty$.

2.2.7 Statistical Inference in Parametric Families

A real-valued observable random variable X has a distribution that may depend in some way on a real-valued parameter θ that takes a value in the set Θ, called the *parameter space*. This random variable is used to model some observable phenomenon.

As the parameter ranges over Θ it determines a family of distributions, $\mathcal{P}$. We denote a specific member of that family as P_θ for some fixed value of θ.

We often want to make inferences about the value of θ or about some function or transformation of an underlying parameter θ. To generalize our object of interest, we often denote it as ϑ, or $g(\theta)$ or $g(\theta; z)$, where g is some Borel function.

Here are some general estimation procedures following the general approaches mentioned above:

- estimation based on the ECDF
  - estimate $g(\theta)$ so that the quantiles of $P_{\widehat{g(\theta)}}$ are close to the quantiles of the data
    How many and which quantiles to match?
    Use of a plug-in estimator from the empirical cumulative distribution function follows this approach, and in that case all quantiles from the data are used.
    This approach may involve questions of how to define sample quantiles.
    An example of this approach is the requirement of median-unbiasedness (one specific quantile).
  - estimate $g(\theta)$ so that the moments of $P_{\widehat{g(\theta)}}$ are close to the sample moments
    How many and which moments to match?
    Do the population moments exist?
    Method-of-moments estimators may have large variances; hence, while this method may be simple (and widely-used), it is probably not a good method generally.
    An example of this approach is the requirement of unbiasedness (one specific moment).

- use the likelihood
  - estimate $g(\theta)$ as $g(\hat{\theta})$, where $\hat{\theta}$ maximizes the likelihood function, $L(\theta, x, z)$.
    Maximum likelihood estimation is closely related to minimum-residual-norm estimation. For the normal distribution, for example, MLE is the same as LS, and for the double exponential distribution, MLE is the same as LAV.
    If there is a sufficient statistic, a MLE is a function of it. (This does not say that every MLE is a function of the sufficient statistic.)
    MLEs often have very good statistical properties. The are particularly easy to work with in exponential families.
- estimation by fitting expected values
  - estimate $g(\theta)$ so that residuals $\|x_i - \mathrm{E}_{\widehat{g(\theta)}}(X_i, z_i)\|$ are small.
    An example of this approach is least squares (LS) estimation (the Euclidean norm of the vector of residuals, or square root of an inner product of the vector with itself). If the expectation exists, least squares yields unbiasedness.
    Another example of this approach is least absolute values (LAV) estimation, in which the L_1 norm of the vector of residuals is minimized. This yields median-unbiasedness.
- define a loss function that depends on how much the estimator differs from the estimand, and then estimate $g(\theta)$ so as to minimize the expected value of the loss function (that is, the "risk") at points of interest in the sample space. (This is an approach based on "decision theory", which we introduce formally in Section 2.3. The specific types of estimators that result from this approach are the subjects of several later chapters.)
  - require unbiasedness and minimize the variance at all points in the sample space (this is UMVU estimation, which we discuss more fully in Chapter 4)
  - require equivariance and minimize the risk at all points in the sample space (this is MRE or MRI estimation, which we discuss more fully in Chapter 8)
  - minimize the maximize the maximum risk over the full sample space
  - define an a priori averaging function for the parameter, use the observed data to update the averaging function and minimize the risk defined by the updated averaging function.

2.3 The Decision Theory Approach to Statistical Inference

2.3.1 Decisions, Losses, Risks, and Optimal Actions

In the decision-theoretic approach to statistical inference, we call the inference a *decision* or an *action*, and we identify a *cost* or *loss* that depends on the

decision and the true (but unknown) state of nature modeled by $P \in \mathcal{P}$. Instead of loss, we could use its opposite, called *utility*.

Obviously, we try to take an action that minimizes the expected loss, or conversely maximizes the expected utility.

We call the set of allowable actions or decisions the *action space* or decision space, and we denote it as $\mathcal{A}$. We base the inference on the random variable X; hence, the decision is a mapping from $\mathcal{X}$, the range of X, to $\mathcal{A}$.

If we observe X, we take the action $T(X) = a \in \mathcal{A}$.

Decision Rules

Given a random variable X with associated measurable space $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$ and an action space $\mathcal{A}$ with a σ-field $\mathcal{F}_{\mathcal{A}}$, a decision rule is a function, T, from $\mathcal{X}$ to $\mathcal{A}$ that is measurable $\mathcal{F}_{\mathcal{X}}/\mathcal{F}_{\mathcal{A}}$.

Decision rules are often denoted by δ.

A *randomized decision rule* is a function is a mapping from $\mathcal{X}$ to the class of probability measures on $(\mathcal{A}, \mathcal{F}_{\mathcal{A}})$. A randomized decision rule can also be defined as a function δ over $\mathcal{X} \times \mathcal{F}_{\mathcal{A}}$ such that for every $A \in \mathcal{F}_{\mathcal{A}}$, $\delta(\cdot, A)$ is a Borel function, and for every $x \in \mathcal{X}$, $\delta(x, \cdot)$ is a probability measure on $(\mathcal{A}, \mathcal{F}_{\mathcal{A}})$.

To evaluate a randomized decision rule requires the realization of an additional random variable. In practice, this would be accomplished by simulation.

Shao uses δ to denote a randomized decision rule, but he usually uses an upper-case Latin letter to denote a nonrandomized decision rule.

Loss Function

A *loss function*, L, is a mapping from $\mathcal{P} \times \mathcal{A}$ to $[0, \infty)$. The value of the function at a given distribution P for the action a is $L(P, a)$.

If $\mathcal{P}$ indexed by θ, we can write the value of the function at a given value θ for the action a as $L(\theta, a)$.

The loss function is defined with respect to the objectives of the statistical inference in such a way that a small loss is desired.

Depending on Θ, $\mathcal{A}$, and our objectives, the loss function often is a function only of $a - g(\theta)$ or of $a/g(\theta)$; that is, we may have $L(\theta, a) = L_{\mathrm{l}}(a - g(\theta))$, or $L(\theta, a) = L_{\mathrm{s}}(a/g(\theta))$. For example,

$$L(\theta, a) = |g(\theta) - a|.$$

In this case, which might be appropriate for estimating $g(\theta)$,

$$\begin{aligned} L(\theta, a) &\geq 0 \quad \forall \theta, a \\ L(\theta, a) &= 0 \quad \text{if } a = g(\theta). \end{aligned}$$

Notice that the loss function is just a mathematical function associated with another function g. There are no assumed underlying random variables. It

does not matter what θ and a are; they are just mathematical variables, or placeholders, taking values in Θ and $\mathcal{A}$. In this case, the loss function generally should be nondecreasing in $|g(\theta) - a|$. A loss function that is convex has nice mathematical properties. (There is some heuristic appeal to convexity, but we like it because of its mathematical tractability. There are lots of other properties of statistical procedures that are deemed interesting for this nonreason.) A particularly nice loss function, which is strictly convex, is the "squared-error loss": $L_2(\theta, a) = (g(\theta) - a)^2$. Another loss function that is often appropriate is the "absolute-error loss": $L_1(\theta, a) = |g(\theta) - a|$. The "absolute-error loss", which is convex but not strictly convex, is not as mathematically tractable as the squared-error loss.

Any strictly convex loss function over an unbounded interval is unbounded. It is not always realistic to use an unbounded loss function. A common bounded loss function is the 0-1 loss function, which may be

$$\begin{aligned} L_{0-1}(\theta, a) &= 0 \quad \text{if } |g(\theta) - a| \leq \alpha(n) \\ L_{0-1}(\theta, a) &= 1 \quad \text{otherwise.} \end{aligned}$$

Risk Function

To choose an action rule T so as to minimize the loss function is not a well-defined problem. We can make the problem somewhat more precise by considering the expected loss based on the action $T(X)$, which we define to be the *risk*:

$$R(P, T) = \mathrm{E}\big(L(P, T(X))\big). \tag{2.28}$$

The risk depends on

- L
- P, the distribution w.r.t. which E is defined
- the decision rule; we may write $R(P, T)$ as $R_T(P)$.

The problem may still not be well defined. For example, to estimate $g(\theta)$ so as to minimize the risk function is still not a well-defined problem. We can make the problem precise either by imposing additional restrictions on the estimator or by specifying in what manner we want to minimize the risk.

Optimal Decision Rules

We compare decision rules based on their risk with respect to a given loss function and a given family of distributions. If a decision rule T_* has the property

$$R(P, T_*) \leq R(P, T) \quad \forall\, P \in \mathcal{P},$$

for all T, then T_* is called an *optimal* decision rule.

Often we limit the set of possible rules. If

$$R(P, T_*) \leq R(P, T) \quad \forall\, P \in \mathcal{P} \text{ and } \forall\, T \in \mathcal{T},$$

then T_* is called a $\mathcal{T}$*-optimal* decision rule.

Admissibility

Before considering specific definitions of a minimum-risk procedure, we define another general desirable property for a decision rule.

Given decision rules T_* and T. The rule T is said to *dominate* the rule T_* iff

$$R(P,T) \leq R(P,T_*) \quad \forall\, P \in \mathcal{P},$$

and

$$R(P,T) < R(P,T_*) \quad \text{for some } P \in \mathcal{P}.$$

A decision rule T_* is *admissible* if there does not exist a decision rule T that dominates T_*. Admissibility depends on

- L
- $\mathcal{P}$, the family of distributions w.r.t. which E is defined

For a given problem there may be no admissible estimator.

Often we limit the set of possible rules to a set $\mathcal{T}$. If the definition above is restricted to $T \in \mathcal{T}$, then T_* is called a $\mathcal{T}$*-admissible* decision rule.

Optimality implies admissibility.

Completeness of a Class of Decision Rules

We have defined completeness of distributions and of statistics. We now define completeness of a class of decision rules. A class of decision rules $\mathcal{T}$ is said to be *complete* if for any decision rule $T \notin \mathcal{T}$, there exists a rule in $\mathcal{T}$ that dominates T. A class is said to be *minimal complete* if it does not contain a complete proper subclass.

If two decision rules have identical risk functions, we would like to think of them as equivalent, but we do not want necessarily to include all such equivalent rules in a class of interest. We therefore define a class of rules $\mathcal{T}$ to be *essentially complete* if for any rule T there is a rule $T_0 \in \mathcal{T}$ such that $R(P,T_0) \leq R(P,T)\ \forall\, P$.

Let $\mathcal{T}$ be a class of decision rules and let $\mathcal{T}_0 \subset \mathcal{T}$. The class $\mathcal{T}_0$ is said to be $\mathcal{T}$*-complete* if $\forall T \in \mathcal{T} - \mathcal{T}_0, \exists T_0 \in \mathcal{T}_0$ that dominates T.

The class $\mathcal{T}_0$ is said to be $\mathcal{T}$*-minimal complete* if $\mathcal{T}_0$ is $\mathcal{T}$*-complete* and no proper subset of $\mathcal{T}_0$ is $\mathcal{T}$*-complete*.

It is easy to see (using the method of proving one set is equal to another by showing each is a subset of the other) that if a $\mathcal{T}$-minimal complete class exists, it is identical to the class of $\mathcal{T}$-admissible decision rule.

L-Unbiasedness

Admissibility involves the relationship between the expected values of the loss function with different decision rules at the same distribution in the family being considered. We can also consider the expected values taken at a given

point in the distribution space of the loss function of a given decision rule at the given value of the parameter compared with the loss at some other distribution. This leads to the concept of L-unbiasedness. A decision rule T is L-unbiased if for all P and $\tilde{P}$,

$$\mathrm{E}_P\big(L(\tilde{P}, T(X))\big) \geq \mathrm{E}_P\big(L(P, T(X))\big).$$

This is the basis for defining unbiasedness for statistical tests and confidence sets.

Unbiasedness for estimators has a simple definition. For squared error loss for estimating $g(\theta)$, if T is L-unbiased, then, and only then, it is unbiased.

Uniformly Minimizing the Risk

All discussions of statistical inference are in the context of some family of distributions, and when we speak of a "uniform" property, we mean a property that holds for all members of the family.

If we have the problem of estimating $g(\theta)$ under some given loss function L, it is often the case that for some specific value of θ, say θ_1, one particular estimator, say T_1, has the smallest expected loss, while for another value of θ, say θ_2, another estimator, say T_2, has a smaller expected loss. Neither T_1 nor T_2 is uniformly optimal.

The risk is a function of the parameter being estimated; therefore, to minimize the risk is not a well-posed problem. A solution is to seek a decision rule that is uniformly best within some restricted class of decision rules.

2.3.2 Approaches to Minimizing the Risk

We use the principle of minimum risk in the following restricted ways. In all cases, the approaches depend, among other things, on a given loss function.

- If there is a sufficient statistic and if the loss function is convex, we can condition any given statistic on the sufficient statistic. For a convex loss function, we have the Rao-Blackwell theorem:
  Let T be a sufficient statistic for $P \in \mathcal{P}$.
  Let T_0 be a statistic with finite expectation.
  Let $T_1 = \mathrm{E}(T_0|T)$.
  Then
  $$R(P, T_1) \leq R(P, T_0) \quad \forall\, P \in \mathcal{P}.$$
  If the loss function is strictly convex and T_0 is not a function of T, then T_0 is inadmissible.
  Finding a statistic with a smaller risk by this method is called "Rao-Blackwellization".
- We may first place a restriction on the estimator and then minimize risk subject to that restriction.
  For example:

  - require unbiasedness
    In this case, we can often eliminate θ from consideration; that is, we can uniformly minimize the risk.
    In a common situation we define loss as squared-error (because unbiased, this means variance), and this yields UMVU.
    Sufficiency and completeness play a major role in UMVUE.
    The information inequality is important in unbiased estimation.
    This approach is great for exponential families.
  - require equivariance
    This must be made more precise (unlike unbiasedness, "equivariance" requires more qualification).
    Equivariance implies independence of the risk from θ; we can uniformly minimize the risk by just minimizing it anywhere.
    This yields UMRE, or just MRE because uniformity is implied.
    This approach is especially useful for group families.
- We may minimize some global property of the risk ("global" over the values of θ).
  For example:
  - minimize "average" risk
    How to average? Let $\Lambda(\theta)$ be such that $\int_\Theta \mathrm{d}\Lambda(\theta) = 1$, then average risk is $\int_\Theta R(\theta, T)\mathrm{d}\Lambda(\theta)$.
    The estimator that minimizes the average risk w.r.t. $\Lambda(\theta)$, T_Λ, is called the Bayes estimator, and the minimum risk, $\int_\Theta R(\theta, T_\Lambda)\mathrm{d}\Lambda(\theta)$, is called the Bayes risk.
    The averaging function allows various interpretations, and it allows the flexibility of incorporating prior knowledge or beliefs. The regions over which $\Lambda(\theta)$ is large will be given more weight; therefore the estimator will be pulled toward those regions.
    In formal Bayes procedures, we call $\Lambda(\theta)$ the prior probability density for θ. We then form the joint distribution of θ and X, and then the conditional distribution of θ given X, called the posterior distribution. The Bayes estimator is determined by minimizing the risk, where the expectation is taken with respect to the posterior distribution. Because the Bayes estimator is determined by the posterior distribution, the Bayes estimator must be a function of a sufficient statistic.
  - minimize maximum risk
    The maximum risk may not exist, so we consider

$$\sup_{\theta\in\Theta} R(\theta, T(X)). \tag{2.29}$$

    The estimator that yields

$$\inf_T \sup_{\theta\in\Theta} R(\theta, T(X)) \tag{2.30}$$

    is the minimax estimator.

A comment about the supremum may be in order here. We mentioned earlier that in parametric inference, we often consider the closure of the parameter space, $\overline{\Theta}$, and in the maximum likelihood estimator in equation (2.15), for example, that allowed us to consider $\max\{\theta \in \overline{\Theta}\}$. We cannot do this in considering the "maximum" risk in equation (2.29) because we do not know how R behaves over $\overline{\Theta}$. (It could be discontinuous anywhere within $\overline{\Theta}$.)

– combinations of global criteria
We could consider various combinations of the global criteria. For example, we may see an estimator that generally minimizes the average risk, but such that its maximum risk is not so large. An intuitively reasonable bound on the maximum risk would bs some excess of the minimum maximum bound. This approach is called *restricted Bayes*, and results in the following constrained optimization problem:

$$\begin{aligned} &\min_T \int R(\theta, T)\mathrm{d}\Lambda(\theta) \\ &\text{s.t.} \quad \sup_{\theta\in\Theta} R(\theta, T(X)) \leq (M+\epsilon)\inf_T \sup_{\theta\in\Theta} R(\theta, T(X)) \end{aligned}$$

- We may combine various criteria.
It is often appropriate to combine criteria or to modify them. This often results in "better" estimators. For example, if for $\theta \in \Theta$, $g(\theta) \in [\gamma_1, \gamma_2]$, and $T(X)$ is an estimator of $g(\theta)$ such that $\Pr(T(X) \notin [\gamma_1, \gamma_2]) \neq 0$, then $T_*(X)$ defined as

$$T_*(X) = \begin{cases} T(X) \text{ if } T(X) \in [\gamma_1, \gamma_2] \\ \gamma_1 \quad \text{if } T(X) < \gamma_1 \\ \gamma_2 \quad \text{if } T(X) > \gamma_2 \end{cases}$$

dominates $T(X)$.
- We may focus on asymptotic criteria.
Sometimes we seek estimators that have good asymptotic properties, such as consistency.

Relationships Among Estimators

There are interesting connections among Bayes estimators and other estimation criteria:

- A Bayes estimator with a constant risk is minimax with respect to the same loss function and distribution.
- A unique Bayes estimator is admissible with respect to the same loss function and distribution.
- An admissible estimator is either Bayes or limiting Bayes.

We will discuss these further in Section 3.2.

Optimal Estimation under Squared-Error Loss

In estimation problems, squared-error loss functions are often the most logical (despite the examples above!). A squared-error loss function is strictly convex, so the useful properties of convex loss functions, such as those relating to the use of sufficient statistics (Rao-Blackwell, for example), hold for squared-error loss functions. Squared-error is of course the loss function in UMVU estimation, and so we use it often.

Squared-error loss functions yield nice properties for linear functions of estimands:

$$\text{If } T \text{ is } \left\{\begin{array}{c}\text{Bayes}\\ \text{UMVU}\\ \text{minimax}\\ \text{admissible}\end{array}\right\} \text{ for } g(\theta), \text{ then } aT+b \text{ is } \left\{\begin{array}{c}\text{Bayes}\\ \text{UMVU}\\ \text{minimax}\\ \text{admissible}\end{array}\right\} \text{ for } ag(\theta)+b,$$

where all properties are taken under squared-error loss.

If in a Bayesian setup, the prior distribution and the posterior distribution are in the same parametric family, then a squared-error loss yield Bayes estimators for $\mathrm{E}(X)$ that are linear in X. (If a prior distribution on the parameters together with a conditional distribution of the observables yield a posterior in the same parametric family as the prior, the prior is said to be *conjugate* with respect to the conditional distribution of the observables. We will consider various types of priors more fully in Chapter 3.)

Because we use squared-error loss functions so often, we must be careful not to assume certain common properties hold. Other types of loss functions can provide useful counterexamples.

2.3.3 Minimaxity and Admissibility

Minimax Estimators

Instead of uniform optimality properties for estimators restricted to be unbiased or equivariant or optimal average properties, we may just seek to find one with the smallest maximum risk. This is *minimax estimation*.

For a given estimation problem, the maximum risk may not exist, so we consider

$$\sup_{\theta\in\Omega} R(\theta, \delta(X)).$$

The estimator that yields

$$\inf_{\delta} \sup_{\theta\in\Omega} R(\theta, \delta(X))$$

is the *minimax estimator*.

Minimaxity, as with most optimality properties, depends on the loss function.

Minimax and Bayes Estimators

There are important, and not necessarily obvious, connections between minimax and Bayes estimators. One of the most important is given in Theorem 4.11 of Shao:

- A Bayes estimator with a constant risk is minimax with respect to the same loss function and distribution.

Hence, one way of finding a minimax estimator is to find a Bayes estimator with constant risk.

For a given loss function, and given distribution of the observable random variable, the minimax estimator is the Bayes estimator for "worst" prior distribution. (This is the implication of Theorem 4.12 in Shao.)

An Example

Theorem 4.14 in Shao provides a condition for identifying minimax estimators in one-parameter exponential families. The minimax estimator is not always the obvious one.

Lehmann gives a very interesting example of an UMVUE in a binomial (π, n) distrbution that is not minimax. The binomial is a complete one-parameter exponential family. The UMVUE of π is $T = X/n$, and under the squared-error loss, the risk, that is, the variance in this case is $\pi(1-\pi)/n$. The maximum risk is easily seen to be $1/(4n)$ (when $\pi = 1/2$). Now, consider the estimator

$$\delta^* = \frac{X}{n}\frac{n^{1/2}}{1+n^{1/2}} + \frac{1}{2(1+n^{1/2})}.$$

This has risk

$$\begin{aligned}
R(\delta^*, \pi) &= \mathrm{E}_\pi((\delta^* - \pi)^2) \\
&= \mathrm{E}_\pi\left(\left(\frac{X}{n}\frac{n^{1/2}}{1+n^{1/2}} + \frac{\pi n^{1/2}}{2(1+n^{1/2})} - \frac{\pi n^{1/2}}{2(1+n^{1/2})} + \frac{1}{2(1+n^{1/2})} - \pi\right)^2\right) \\
&= \left(\frac{n^{1/2}}{1+n^{1/2}}\right)^2 \mathrm{E}_\pi\left(\left(\frac{X}{n} - \pi\right)^2\right) + \left(\frac{\pi n^{1/2}}{1+n^{1/2}} + \frac{1}{2(1+n^{1/2})} - \pi\right)^2 \\
&= \left(\frac{n^{1/2}}{1+n^{1/2}}\right)^2 \frac{\pi(1-\pi)}{n} + \left(\frac{1-2\pi}{2(1+n^{1/2})}\right)^2 \\
&= \frac{1}{4(1+n^{1/2})^2},
\end{aligned}$$

which is constant.

The risk of δ^* is less than the maximum risk of T; therefore, T is not minimax.

Note also that δ^* is Bayes w.r.t. a beta prior. (Check this out and determine the beta hyperparameters.)

Furthermore, because δ^* is Bayes with a constant risk, it is a minimax estimator.

Admissible Estimators

An estimator δ_* is admissible if there does not exist an estimator δ that *dominates* δ_*, that is, such that

$$R(\theta, \delta) \leq R(\theta, \delta_*) \quad \forall \theta \in \Omega,$$

and

$$R(\theta, \delta) < R(\theta, \delta_*) \quad \text{for some } \theta \in \Omega.$$

A slightly more general form of admissibility is λ-admissibility:
An estimator δ_* is λ-admissible if it is admissible almost everywhere with respect to the measure λ defined over the sample space.

- A unique Bayes estimator is admissible with respect to the same loss function and distribution.
- An admissible estimator is either Bayes or limiting Bayes.

Inadmissible Estimators

Some estimators that have generally good properties or that are of a standard type, such as MLE or method of moments, may not be admissible. Sometimes, a randomized estimator can be constructed to show that a given estimator is not admissible.

The estimation of the mean of a normal distribution has interesting admissibility properties. It is relatively straightforward to show that $\overline{X}$ is admissible for estimating θ in $\mathrm{N}(\theta, 1)$ (and you should be able to do this; the variance is taken to be 1 without loss of generality). It can also be shown that $\overline{X}$ is admissible for estimating θ in $\mathrm{N}_2(\theta, I_2)$, and of course, in the simpler case of $n = 1$, X is admissible for estimating θ.

However, for $r > 2$, X is not admissible for estimating θ in $\mathrm{N}_r(\theta, I_r)$!

For $r > 2$, the estimator

$$\hat{\theta}_{\mathrm{J}} = \left(1 - c\frac{r-2}{\|X\|^2}\right) X \tag{2.31}$$

though biased, dominates X. This is called the James-Stein estimator.

The James-Stein estimator is generally shrunk toward 0. This type of adjustment is called Stein shrinkage. Choice of c allows for different amounts of bias and different amounts of reduction in the risk. (If you're familiar with ridge regression, compare it with this. The regularization parameter in ridge regression is similar to the c in this expression.)

This fact is related to the outlyingness of data in higher dimensions. Although for $r \leq 2$, the ordinary mean is admissible with respect to MSE, we have seen on page 72 that a certain shrunken estimator is Pitman-closer than the mean for a normal distribution with $r = 1$.

There are many other surprising cases of inadmissibility.

Consider the case of estimating θ in the finite population $\{1, \ldots, \theta\}$. Suppose we sample from this population with replacement, obtaining $X_1, \ldots, X_n$. Because $\mathrm{E}(\overline{X}) = (\theta + 1)/2$, the method of moments estimator of θ is $T = 2\overline{X} - 1$. This estimator is inadmissible (for any reasonable loss function), since $T^* = \max(X_{(n)}, T)$ is always at least as close to θ, and can be closer. (Note also that the MoM estimator of θ may produce a value that could never be the true value of θ.)

Consider another example from Lehmann for a one-parameter exponential family. Let X have the density

$$p_\theta(x) = \beta(\theta)\mathrm{e}^{\theta x}\mathrm{e}^{-|x|},$$

where $\theta \in (-1, 1)$ and $\beta(\theta) = 1 - \theta^2$ (so it integrates to 1). Consider a sample of size one, X, and the problem of estimating $g(\theta) = \mathrm{E}_\theta(X)$ with squared-error loss. Now,

$$\begin{aligned}\mathrm{E}_\theta(X) &= -\frac{\beta'(\theta)}{\beta(\theta)} \\ &= \frac{2\theta}{1-\theta^2},\end{aligned}$$

and

$$\begin{aligned}\mathrm{V}_\theta(X) &= \frac{\mathrm{d}}{\mathrm{d}\theta}\mathrm{E}_\theta(X) \\ &= 2\frac{1+\theta^2}{(1-\theta^2)^2};\end{aligned}$$

hence, the risk is

$$R(g(\theta), X) = 2\frac{1+\theta^2}{(1-\theta^2)^2}.$$

Now, consider the estimator $T_a = aX$. Its risk under squared-error is

$$\begin{aligned}R(\theta, T_a) &= \mathrm{E}_\theta(L(\theta, T_a)) \\ &= \mathrm{E}_\theta((g(\theta) - T_a)^2) \\ &= 2a^2\frac{1+\theta^2}{(1-\theta^2)^2} + 4(1-a^2)\frac{\theta^2}{(1-\theta^2)^2}.\end{aligned}$$

If $a = 0$, that is, it the estimator is the constant 0, the risk is $4\theta^2/(1-\theta^2)^2$, which is smaller than the risk for X for all $\theta \in (-1, 1)$!

The natural sufficient statistic in this one-parameter exponential family is inadmissible for its expectation!

Other Forms of Admissibility

We have defined admissibility in terms of a specific optimality criterion, namely minimum risk. Of course, the risk depends on the loss function, so admissibility depends on the particular loss function.

We can define admissibility in a similar fashion with respect to any optimality criterion; for example, the estimator $T(X)$ is *Pitman-admissible* for $g(\theta)$ if there does not exist an estimator that is *Pitman-closer* to $g(\theta)$.

2.3.4 Other Issues in Statistical Inference

We may also be interested in determining the conditional distribution of other functions of X, given $T(X)$ (Is $Y(X)$ sufficient? What statistics are ancillary?).

We generally seek estimators whose distributions have some optimal properties. We may also consider the performance over a class of probability families, $\{\mathcal{P}, \mathcal{Q}, \mathcal{R}, \ldots\}$. (This is called "robustness".)

Other approaches to estimation do not necessarily involve consideration of the probability distributions of the estimators, but are based on desirable heuristics: least squares, other minimal norms or minima of other functions of model residuals, maximum likelihood, etc.

2.4 Probability Statements in Statistical Inference

There are two instances in statistical inference in which statements about probability are associated with the decisions of the inferential methods. In hypothesis testing, under assumptions about the distributions, we base our inferential methods on probabilities of two types of errors. In confidence sets the decisions are associated with probability statements about coverage of the parameters.

In both of these types of inference, the basic set up is the standard one in statistical inference. We have a random sample of observations $X_1, \ldots, X_n$ on a random variable X that has a distribution P_θ, some aspects of which are unknown. We assume some family of probability distributions $\mathcal{P}$ such that $P_\theta \in \mathcal{P}$. We begin with some preassigned probability that, following the prescribed method of inference, we will arrive at set of distributions $\mathcal{P}_\theta$ that contain the distribution P_θ. Our objective is to determine such methods, and among a class of such methods, determine ones that have optimal properties with respect to reasonable criteria.

After having completed such a process, it may not be appropriate to characterize the relationship of the "true" unknown distribution P_θ to the set of $\mathcal{P}_\theta$ with any statement about "probability". Presumably, either $P_\theta \in \mathcal{P}_\theta$ or $P_\theta \notin \mathcal{P}_\theta$.

In these types of statistical inference, as we will describe below, we use the terms "significance level", "size", "confidence level", and "confidence coefficient" to describe our findings.

Approximations

In some cases, we have a tractable probability model, so that we can perform tests with exact levels of significance or determine exact confidence sets. In other cases the problem is not tractable analytically, so we must resort to approximations, which may be base on asymptotic distributions, or to estimates, which may be made using simulations.

Asymptotic inference uses asymptotic approximations. Computational inference uses probabilities estimated by simulation an assumed or hypothesized data generating process or by resampling of an observed sample.

2.4.1 Tests of Hypotheses

Given a set of data, X, and a family of possible distributions that gave rise to the data, $\mathcal{P}$, a common objective of statistical inference is to specify a particular member or subclass of $\mathcal{P}$ that "likely" generated X. For example, if $\mathcal{P} = \{\mathrm{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+\}$, given $X = x$, we may choose $\mathrm{N}(\bar{x}, s^2)$ as a good candidate for the population from which the data arose. This choice is based on statistical estimators that we know to be "good" ones.

In another kind of statistical inference, given a set of data X and a family of distributions $\mathcal{P}$, we are to decide whether the data "likely" came from some hypothesized subfamily $\mathcal{P}_0$ of $\mathcal{P}$. Our possible decisions are "yes" or "no". Rather than a general "no", a specific alternative may be hypothesized.

This kind of statistical inference is called "testing statistical hypotheses". We will discuss this topic more fully in Chapter 6. In Chapter 3 we discuss testing from a Bayesian perspective. Here, we just introduce some terms and consider some simple cases.

Statistical Hypotheses

The hypotheses concern a specific member $P \in \mathcal{P}$. This is the distribution that generated the observed data.

We have a *null hypothesis*

$$H_0 : \quad P \in \mathcal{P}_0$$

and an *alternative hypothesis*

$$H_1 : \quad P \in \mathcal{P}_1,$$

where $\mathcal{P}_0 \subset \mathcal{P}$, $\mathcal{P}_1 \subset \mathcal{P}$, and $\mathcal{P}_0 \cap \mathcal{P}_1 = \emptyset$. If $\mathcal{P}_0 \cup \mathcal{P}_1 = \mathcal{P}$, the alternative hypothesis is effectively "everything else".

An hypothesis that specifies exactly one distribution is called a *simple hypothesis*; otherwise it is called a *composite hypothesis*. H_0 above is a simple hypothesis if there is only one distribution in $\mathcal{P}_0$.

If the family of distributions is associated with a parameter space Θ, we may equivalently describe the tests as

$$H_0 : \theta \in \Theta_0$$

versus

$$H_1 : \theta \in \Theta_1.$$

An hypothesis $H : \theta \in \Theta_H$ in which $\#\Theta_H = 1$ is a simple hypothesis; if $\#\Theta_H > 1$ it is a composite hypothesis. Of course we are often interested in the case where $\Theta = \Theta_0 \cup \Theta_1$. An hypothesis of the form $H_0 : \theta = \theta_0$ is a simple hypothesis, while $H_i : \theta \geq \theta_0$ is a composite hypothesis.

Test Statistics and Critical Regions

A straightforward way of performing the test involves use of a test statistic, $T(X)$, computed from a random sample of data. Associated with $T(X)$ is a rejection region R, such that if the null hypothesis is true, for some preassigned (small) value, α,

$$\Pr(T(X) \in R) \leq \alpha.$$

We seek a statistic $T(X)$ such that $\Pr(T(X) \in R)$ is large if the null hypothesis is not true. Thus, R is a region of more "extreme" values of the test statistic if the null hypothesis is true.

If $T(X) \in R$, the null hypothesis is rejected. The rejection region is also called the critical region. The complement of the rejection region is called the acceptance region.

It is desirable that the test have a high probability of rejecting the null hypothesis if indeed the null hypothesis is not true.

p-Values

A procedure for testing that is mechanically equivalent to this is to compute the test statistic $T(X) \leftarrow t$ and then to determine the probability that $T(X)$ is more extreme than t. In this approach, the realized value of the test statistic determines a region R_t of more extreme values. The probability that the test statistic is in R_t if the null hypothesis is true, $\Pr(T \in R_t)$, is called the "p-value" or "significance level" of the realized test statistic. ***** discuss this! and size

In this framework we are testing one hypothesis versus another hypothesis. The two hypotheses are not treated symmetrically, however. We are still directly testing the null hypothesis. This asymmetry allows us to focus on two kinds of losses that we might incur. The losses relate to the two kinds of errors that we might make.

Test Rules

Instead of thinking of a test statistic T and a rejection region R, as above, we can formulate the testing procedure in a slightly different way. We can think of the test as a decision rule, $\delta(X)$, which is a statistic that relates more directly to the decision about the hypothesis. We sometimes refer to the statistic $\delta(X)$ as "the test", because its value is directly related to the outcome of the test; that is, there is no separately defined rejection region.

A *nonrandomized test procedure* is a rule $\delta(X)$ that assigns two decisions to two disjoint subsets, C_0 and C_1, of the range of X. In general, we require $C_0 \cup C_1$ be the support of X. We equate those two decisions with the real numbers d_0 and d_1, so $\delta(X)$ is a real-valued function,

$$\delta(x) = \begin{cases} d_0 \text{ for } x \in C_0 \\ d_1 \text{ for } x \in C_1. \end{cases}$$

For simplicity, we choose $d_0 = 0$ and $d_1 = 1$. Note for $i = 0, 1$,

$$\Pr(\delta(X) = i) = \Pr(X \in C_i).$$

We call C_1 the *critical region*, and generally denote it by just C. (It is not my intent to distinguish C from R above; they're both "critical regions". I have used C to denote a set of values of X, and R to denote a set of values of $T(X)$.)

If $\delta(X)$ takes the value 0, the decision is not to reject; if $\delta(X)$ takes the value 1, the decision is to reject. If the range of $\delta(X)$ is $\{0, 1\}$, the test is a nonrandomized test. Sometimes, however, it is useful to expand the range of $\delta(X)$ to be $[0, 1]$, where we can interpret a value of $\delta(X)$ as the probability that the null hypothesis is rejected. If it is not the case that $\delta(X)$ equals 0 or 1 a.s., we call the test a randomized test.

Power of the Test

We now can focus on the test under either hypothesis (that is, under either subset of the family of distributions) in a unified fashion. We define the *power function* of the test, for any given $P \in \mathcal{P}$ as

$$\beta(\delta, P) = \mathrm{E}_P(\delta(X)). \tag{2.32}$$

We also often use the notation $\beta_\delta(P)$ instead of $\beta(\delta, P)$. In general, the probability of rejection of the null hypothesis is called the power of the test.

An obvious way of defining optimality for tests is in terms of the power for distributions in the class of the alternative hypothesis; that is, we seek "most powerful" tests.

Errors

If $P \in \mathcal{P}_0$ and $\delta(X) = 1$, we make an error; that is, we reject a true hypothesis. We call that a "type I error". For a randomized test, we have the possibility of making a type I error if $\delta(X) > 0$. In general, if $P \in \mathcal{P}_0$, $\beta_\delta(P)$ is the probability of a type I error. Conversely, if $P \in \mathcal{P}_1$, then $1 - \beta_\delta(P)$ is the probability of a "type II error", that is failing to reject a false hypothesis.

Testing as a Decision Problem

For a statistical hypothesis that involves the distribution of the random variable X, a *nonrandomized test procedure* is a rule $\delta(X)$ that assigns two decisions to two disjoint subsets, C_0 and C_1, of the range of X. In general, we require $C_0 \cup C_1$ be the support of X. We equate those two decisions with the real numbers 0 and 1, so $\delta(X)$ is a real-valued function,

$$\delta(x) = \begin{cases} 0 \text{ for } x \in C_0 \\ 1 \text{ for } x \in C_1. \end{cases}$$

Note for $i = 0, 1$, $\Pr(\delta(X) = i) = \Pr(X \in C_i)$. We call C_1 the *critical region*, and generally denote it by just C.

We also write

$$\phi(x) = \Pr(\delta(X) = 1 \mid X = x).$$

Notice that this is the same as the power, except ϕ here is a function of the observations, while we think of the power as a function of the true distribution. Assuming only the two outcomes, we have

$$1 - \phi(x) = \Pr(\delta(X) = 0 \mid X = x).$$

For this decision problem, an obvious choice of a loss function is the 0-1 loss function:

$$\begin{aligned} L(\theta, i) &= 0 \quad \text{if } \theta \in \Theta_i \\ L(\theta, i) &= 1 \quad \text{otherwise.} \end{aligned}$$

It may be useful to consider a procedure with more than just two outcomes; in particular, a third outcome, γ, may make sense. In an application in analysis of data, this decision may suggest collecting more data; that is, it may correspond to "no decision", or, usually only for theoretical analyses, it may suggest that a decision be made randomly. We will, at least in the beginning, however, restrict our attention to procedures with just two outcomes.

For the two decisions and two state of nature case, there are four possibilities:

- the test yields 0 and H_0 is true (correct decision);
- the test yields 1 and H_1 is true (correct decision);

- the test yields 1 and H_0 is true (type I error); and
- the test yields 0 and H_1 is true (type II error).

We obviously want a test procedure that minimizes the probabiltiy of either type of error. It is clear that we can easily decrease the probability of one (if its probability is positive) at the cost of increasing the probability of the other.

We do not treat H_0 and H_1 symmetrically; H_0 is the *hypothesis* to be tested and H_1 is the *alternative*. This distinction is important in developing a methodology of testing.

We adopt the following approach for choosing δ (under the given assumptions on X, and the notation above):

1. Choose $\alpha \in (0, 1)$ and require that $\delta(X)$ be such that

$$\Pr(\delta(X) = 1 \mid \theta \in \Theta_0) \leq \alpha.$$

   α is called the *level of significance*.
2. Subject to this, find $\delta(X)$ so as to minimize

$$\Pr(\delta(X) = 0 \mid \theta \in \Theta_1).$$

The definition of significance level is not as ambiguous as it may appear at first glance.

One *chooses* α; that is the level of significance.

For some $\tilde{\alpha} > \alpha$, although $\Pr(T(X) = 1 \mid \theta \in \Theta_0) \leq \tilde{\alpha}$, we would not say that $\tilde{\alpha}$ is the level (or a level) of significance.

Notice that the restriction on the type I error in the first step applies $\forall \theta \in \Theta_0$. We call

$$\sup_{\theta \in \Theta_0} \Pr(\delta(X) = 1 \mid \theta)$$

the *size of the test*. If the size is less than the level of significance, the test is said to be *conservative*, and in that case, we often refer to α as the "nominal size".

Approximate Tests

If the distribution of the test statistic T or δ under the null hypothesis is known, the critical region or the p-value can be determined. If the distribution is not known, some other approach must be used. A common method is to use some approximation to the distribution. The objective is to approximate a quantile of δ under the null hypothesis. In asymptotic inference, the approximation is often based on an asymptotic distribution of the test statistic.

In computational inference, a Monte Carlo test may be used. In Monte Carlo tests the quantile of δ is estimated by simulation of the distribution.

Unbiased Tests

A test of $H_0 : P \in \mathcal{P}_0$ versus $H_1 : P \in \mathcal{P}_1$ is said to be *unbiased at level* α if the power function satisfies

$$\begin{aligned} \beta_T(P) &\leq \alpha \quad \text{for } P \in \mathcal{P}_0 \\ \beta_T(P) &\geq \alpha \quad \text{for } P \in \mathcal{P}_1 \end{aligned}$$

Uniformly Best Tests

The risk or the expected error in a test depends on the specific distribution within the family of distributions assumed. We may seek a test that has minimum expected errors of both types, or, in a practical approach to this objective, we may cap the probability of a type I error and seek the most powerful test for distributions within the class of the alternative hypothesis.

As we have seen in the estimation problem, optimality generally depends on the specific distribution, and it may not be possible to achieve it uniformly; that is, for all distributions within a given family.

We may then take the approach mentioned on page 96 for estimation and restrict the allowable tests in some way. We may require that the tests be unbiased, for example. That approach leads us to seek a UMPU test, that is, a uniformly most powerful unbiased test.

2.4.2 Confidence Sets

In a problem of statistical inference for a family of distributions $\mathcal{P}$, or equivalently, for a parameter space Θ, given a sample X, a level $1-\alpha$ *confidence set*, or *confidence region* (the terms are synonymous), is a is a random subset of $\mathcal{P}$, $A(X)$, such that

$$\Pr_P\left(A(X) \ni P\right) \geq 1-\alpha \quad \forall P \in \mathcal{P}.$$

More precisely, we call $A(X)$ a *random family of level* $1-\alpha$ *confidence sets*. This definition obviously leaves many issues to be examined because of the $\geq$ relationship. A family of $1-\alpha_1$ confidence sets is also a family of $1-\alpha_2$ confidence set for $\alpha_2 \geq \alpha_1$; and if $A(X)$ is a level $1-\alpha$ confidence set, then $B(X)$ is also a level $1-\alpha$ confidence set if $B(X) \supset A(X)$.

We call

$$\inf_{P \in \mathcal{P}} \Pr_P\left(A(X) \ni P\right)$$

the *confidence coefficient* of $A(X)$.

The confidence coefficient is also called the *coverage probability*.

Equivalently, we can define a random family $S(X)$ of $1-\alpha$ confidence sets for the parameter space Θ by

$$\Pr_\theta\left(S(X) \ni \theta\right) \geq 1-\alpha \quad \forall \theta \in \Theta.$$

A realization of a confidence set, say $A(x)$, is also called a confidence set. Although it may seem natural to state that the "probability that θ is in $A(x)$ is $1-\alpha$", this statement can be misleading unless a certain underlying probability structure is assumed.

We will introduce and discuss other terms in Chapter 7. In Chapter 3 we discuss confidence sets from a Bayesian perspective. Here, we just define the term and consider some simple cases.

Pivot Functions

For forming confidence sets, we often can use a function of the sample that also involves the parameter of interest, $f(T,\theta)$. The confidence set is then formed by separating the parameter from the sample values.

A class of functions that are particularly useful for forming confidence sets are called *pivotal* values, or pivotal functions. A function $f(T,\theta)$ is said to be a pivotal function if its distribution does not depend on any unknown parameters. This allows exact confidence intervals to be formed for the parameter θ.

Confidence Intervals

Our usual notion of a confidence leads to the definition of a $1-\alpha$ confidence interval for the (scalar) parameter θ as the random interval (T_L, T_U), that has the property

$$\Pr\left(T_L \leq \theta \leq T_U\right) \geq 1-\alpha. \tag{2.33}$$

This is also called a $(1-\alpha)100\%$ confidence interval. The interval (T_L, T_U) is not uniquely determined.

The concept extends easily to vector-valued parameters. Rather than taking vectors T_L and T_U, however, we generally define an ellipsoidal region, whose shape is determined by the covariances of the estimators.

A realization of the random interval, say (t_L, t_U), is also called a confidence interval.

In practice, the interval is usually specified with respect to an estimator of θ, T. If we know the sampling distribution of $T-\theta$, we may determine c_1 and c_2 such that

$$\Pr\left(c_1 \leq T-\theta \leq c_2\right) = 1-\alpha;$$

and hence

$$\Pr\left(T-c_2 \leq \theta \leq T-c_1\right) = 1-\alpha.$$

If either T_L or T_U is infinite or corresponds to a bound on acceptable values of θ, the confidence interval is one-sided. Suppose $\Theta = (a, b)$, where a or b may be infinite. In equation (2.33), if $T_L = a$, then T_U is called an *upper confidence bound*, and if $T_U = b$, then T_L is called a *lower confidence bound*. (It is better not to use the terms "upper confidence interval" or "lower confidence interval", because of the possible ambiguity in these terms.

For two-sided confidence intervals, we may seek to make the probability on either side of T to be equal, to make $c_1 = -c_2$, and/or to minimize $|c_1|$ or $|c_2|$. This is similar in spirit to seeking an estimator with small variance.

We can use a pivot function $f(T, \theta)$ to form confidence intervals for the parameter θ. We first form

$$\Pr\Big(f_{(\alpha/2)} \;\le\; f(T,\theta) \;\le\; f_{(1-\alpha/2)}\Big) \;=\; 1-\alpha,$$

where $f_{(\alpha/2)}$ and $f_{(1-\alpha/2)}$ are quantiles of the distribution of $f(T, \theta)$; that is,

$$\Pr(f(T,\theta) \le f_{(\pi)}) = \pi.$$

If, as in the case considered above, $f(T, \theta) = T - \theta$, the resulting confidence interval has the form

$$\Pr\Big(T - f_{(1-\alpha/2)} \;\le\; \theta \;\le\; T - f_{(\alpha/2)}\Big) \;=\; 1-\alpha.$$

For example, suppose $Y_1, Y_2, \ldots, Y_n$ is a random sample from a $\mathrm{N}(\mu, \sigma^2)$ distribution, and $\overline{Y}$ is the sample mean. The quantity

$$f(\overline{Y},\mu) = \frac{\sqrt{n(n-1)}\,(\overline{Y}-\mu)}{\sqrt{\sum\;\left(Y_i-\overline{Y}\right)^2}}$$

has a Student's t distribution with $n-1$ degrees of freedom, no matter what is the value of σ^2. This is one of the most commonly-used pivotal values.

The pivotal value can be used to form a confidence value for θ by first writing

$$\Pr\left(t_{(\alpha/2)} \;\le\; f(\overline{Y},\mu) \;\le\; t_{(1-\alpha/2)}\right) \;=\; 1-\alpha,$$

where $t_{(\pi)}$ is a percentile from the Student's t distribution. Then, after making substitutions for $f(\overline{Y}, \mu)$, we form the familiar confidence interval for μ:

$$\Big(\overline{Y} - t_{(1-\alpha/2)}\, S/\sqrt{n}, \quad \overline{Y} - t_{(\alpha/2)}\, S/\sqrt{n}\Big),$$

where S^2 is the usual sample variance, $\sum(Y_i - \overline{Y})^2/(n-1)$.

(Note the notation: $t_{(\pi)}$, or for clarity, $t_{\nu,(\pi)}$ is the π quantile of a Student's t distribution. That means that

$$\Pr(Y \le t_{\nu,(\pi)}) = \pi.$$

Other authors sometimes use a similar notation to mean the $1-\pi$ quantile and other times to mean the π quantiles. I mean the same authors use it both ways. I always use the notation in the way I indicate above. The reasons for the different symbols go back to the fact that $t_{\nu,(\pi)} = -t_{\nu,(1-\pi)}$, as for any distribution that is symmetric about 0.)

Other similar pivotal functions have F distributions. For example, consider the usual linear regression model in which the n-vector random variable Y has a $\mathrm{N}_n(X\beta, \sigma^2 I)$ distribution, where X is an $n \times m$ known matrix, and the m-vector β and the scalar σ^2 are unknown. A pivotal value useful in making inferences about β is

$$g(\widehat{\beta}, \beta) = \frac{\big(X(\widehat{\beta}-\beta)\big)^{\mathrm{T}} X(\widehat{\beta}-\beta)/m}{(Y-X\widehat{\beta})^{\mathrm{T}}(Y-X\widehat{\beta})/(n-m)},$$

where

$$\widehat{\beta} = (X^{\mathrm{T}}X)^{+}X^{\mathrm{T}}Y.$$

The random variable $g(\widehat{\beta}, \beta)$ for any finite value of σ^2 has an F distribution with m and $n-m$ degrees of freedom.

For a given parameter and family of distributions there may be multiple pivotal values. For purposes of statistical inference, such considerations as unbiasedness and minimum variance may guide the choice of a pivotal value to use. Alternatively, it may not be possible to identify a pivotal quantity for a particular parameter. In that case, we may seek an approximate pivot. A function is asymptotically pivotal if a sequence of linear transformations of the function is pivotal in the limit as $n \to \infty$.

If the distribution of T is known, c_1 and c_2 can be determined. If the distribution of T is not known, some other approach must be used. A common method is to use some numerical approximation to the distribution. Another method is to use bootstrap resampling.

Optimal Confidence Sets

We seek confidence sets that are "small" or "tight" in some way. We want the region of the parameter space that is excluded by the confidence set to be large; that is, we want the probability that the confidence set exclude parameters that are not supported by the observational evidence to be large. This is called "accuracy". We see most accurate confidence sets.

As with point estimation and tests of hypotheses, the risk in setting a confidence region depends on the specific distribution within the family of distributions assumed. We, therefore, seek *uniformly most accurate* confidence sets.

As in other cases where we seek uniform optimality, such procedures may not exist. We, therefore, may then take a similar approach for setting confidence regions, and restrict the allowable regions in some way. We may require that the confidence sets be unbiased, for example.

Unbiased Confidence Sets

A family of confidence sets $S(X)$ for θ is said to be *unbiased* (without regard to the level) if

$$\Pr_{\theta_0}(S(X) \ni \theta_1) \leq \Pr_{\theta_0}(S(X) \ni \theta_0) \; \forall \theta_0, \theta_1 \in \Theta.$$

Prediction Sets and Tolerance Sets

We often want to identify a set in which a future observation on a random variable has a high probability of occurring. This kind of set is called a *prediction set*.
For example, we may assume a given sample $X_1, \ldots, X_n$ is from a $\mathrm{N}(\mu, \sigma^2)$ and we wish to determine a measurable set $S(X)$ such that for a future observation X_{n+1}

$$\inf_{P \in \mathcal{P}} \Pr_P(X_{n+1} \in S(X)) \geq 1 - \alpha.$$

More generally, instead of X_{n+1}, we could define a prediction interval for any random variable V.

The difference in this and a confidence set for μ is that there is an additional source of variation. The prediction set will be larger, so as to account for this extra variation.

We may want to separate the statements about V and $S(X)$. A *tolerance set* attempts to do this.

Given a sample X, a measurable set $S(X)$, and numbers δ and α in $(0, 1)$, if

$$\inf_{P \in \mathcal{P}} (\inf_{P \in \mathcal{P}} \Pr_P(V \in S(X) | X) \geq \delta) \geq 1 - \alpha,$$

then $S(X)$ is called a δ-tolerance set for V with confidence level $1 - \alpha$.

2.5 Asymptotic Inference

In the standard problem in statistical inference, we are given some family of probability distributions, we take random observations on a random variable, and we use some function of the random sample to estimate some aspect of the underlying probability distribution or to test some statement about the probability distribution.

The approach to statistical inference that we would like to follow is to identify a reasonable statistic to use as an estimator or a test statistic, then work out its distribution under the given assumptions and under any null hypothesis, and, knowing that distribution, assess its goodness for the particular application and determine levels of confidence to associate with our inference. In many of interesting problems in statistical inference we cannot do this, usually because the distributions are not tractable.

There are two ways to proceed. One is to use computer simulation to *estimate* properties of our statistic. This approach is called *computational inference*. The other approach is to make some approximations, either of the underlying assumptions or for the unknown distribution.

Some approximations are just based on known similarities between two distributions. The most common kind of approximation, however, is based on the *asymptotic* properties of the statistic. This is *asymptotic inference*.

The Basic Setup and Notation

As usual in statistical inference, we have a family of probability distributions $\mathcal{P} = \{P_\theta\}$, where θ may be some parameter in a real-valued parameter space Θ ("parametric inference"), or θ may just be some index in an index set $\mathcal{I}$ to distinguish one distribution, P_{θ_1}, from another, P_{θ_2} ("nonparametric inference"). The parameter or the index is not observable; however, we assume $P_{\theta_1} \neq P_{\theta_2}$ if $\theta_1 \neq \theta_2$ ("identifiability").

We have an observable random variable X. We have a random sample, $X_1, \ldots, X_n$, which we may also denote by X; that is, we may use X not just as the random variable (that is, a Borel function on the sample space) but also as the sample: $X = X_1, \ldots, X_n$.

Both θ and X may be vectors. (I use "real-valued" to mean either a scalar (that is, an element in $\mathbb{R}$) or a real-valued vector (that is, an element in $\mathbb{R}^k$, where k is a positive integer possibly larger than 1)).

The canonical problem in parametric inference is to estimate $g(\theta)$ or to test some hypothesis concerning $g(\theta)$, where g is some real-valued measurable function. We denote our statistic (either an estimator or a test statistic) as $T_n(X)$, or just T_n.

2.5.1 Consistency

Consistency is a general term used for various types of asymptotic convergence. Unless it is clear from the context, we must qualify the word "consistency" with the type of convergence and with the type of inference. We speak of consistent point estimators, confidence sets, and tests.

We'll begin with consistency of point estimators. This relates to the convergence of the estimator $T_n(X)$ to the estimand $g(\theta)$.

Convergence is defined with respect to a distribution. In a problem of statistical inference we do not know the distribution, only the distributional family, $\mathcal{P}$. To speak of consistency, therefore, we require that the convergence be with respect to every distribution in $\mathcal{P}$.

The three most common kinds of consistency for point estimators are

weak consistency $T_n(X)$ is said to be *weakly consistent* for $g(\theta)$ iff

$$T_n(X) \to_{\mathrm{p}} g(\theta) \quad \text{w.r.t. any } P \in \mathcal{P}.$$

This kind of consistency involves a *weak convergence*. We often refer to weak consistency as just consistency, without the qualifier.

strong consistency $T_n(X)$ is said to be *strongly consistent* for $g(\theta)$ iff

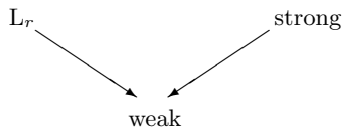
$$T_n(X) \to_{\text{a.s.}} g(\theta) \quad \text{w.r.t. any } P \in \mathcal{P}.$$

L_r-consistency $T_n(X)$ is said to be *L_r-consistent* for $g(\theta)$ iff

$$T_n(X) \to_{\mathrm{L}_r} g(\theta) \quad \text{w.r.t. any } P \in \mathcal{P}.$$

L_r-convergence applies to convergence in expectation: $\lim_{n\to\infty} \mathrm{E}(\|T_n(X) - g(\theta)\|_r^r) = 0$.
For $r = 1$, L_r-consistency is convergence in mean.
For $r = 2$, L_r-consistency is convergence in mean-squared error, and hence L_2-consistency is called consistency in mean-squared error.

There are relationships among these types of consistency that are similar to those among the types of convergence. We have



Another kind of consistency is a convergence in probability defined in terms of a divergent sequence.

a_n-consistency Given a sequence of positive constants $\{a_n\}$ with $\lim_{n\to\infty} a_n = \infty$, $T_n(X)$ is said to be *a_n-consistent* for $g(\theta)$ iff $a_n(T_n(X) - g(\theta)) = \mathrm{O_P}(1)$ w.r.t. any $P \in \mathcal{P}$, that is,

$$\forall \epsilon > 0 \; \exists \text{ constant } C_\epsilon > 0 \ni \sup_n \Pr(a_n \|T_n(X) - g(\theta)\| \geq C_\epsilon) < \epsilon.$$

Notice that this is a kind of weak consistency.
The most common kind of a_n-consistency that we consider is $\sqrt{n}$-consistency; that is, $a_n = \sqrt{n}$.

In asymptotic inference, we often encounter a sequence $\{a_n\}$ with

$$\lim_{n\to\infty} a_n = \infty,$$

like that in the definition of a_n-consistency above. We are interested the limiting behavior of such properties of statistics as the variance, the bias, and the mean-squared error. These quantities are defined in terms of expectations, so firstly, we need to be precise in our meaning of asymptotic expectation. In the following we will distinguish asymptotic expectation from limiting expectation. A related term is "approximate" expectation. This term is sometimes

used in different ways. For example, Shao (p. 135) uses the term "approximately unbiased" in reference to a limiting expectation. Other authors and I prefer the term "unbiased in the limit" to refer to this property. This property is different from asymptotically unbiased, as we will see.

Some points to remember:

- Consistency does not imply unbiasedness.
- $\mathrm{E}(|T_n - g(\theta)|) \to 0$ implies consistency.
- Consistency is often proved by use of a Chebyshev-type inequality.

Consistency of a Sequence to a Sequence

In some cases, rather than a fixed estimand $g(\theta)$, we are interested in a sequence of estimands $g_n(\theta)$. In such cases, it may not be adequate just to consider $|T_n - g_n(\theta)|$. This would not be meaningful if, for example, $g_n(\theta) \to 0$. This kind of situation occurs when $g_n(\theta)$ is the variance of the mean of a sample of size n from a population with finite variance. In such cases we could define any of the types of consistency defined above using the appropriate type of convergence in this expression,

$$|T_n/g_n(\theta) - 1| \to 0. \tag{2.34}$$

2.5.2 Asymptotic Expectation

Let $\{X_n\}$ be a sequence of random variables and $X_n \to_{\mathrm{d}} X$, with $\mathrm{E}(|X|) < \infty$. (Recall that this type of convergence is defined in terms of the convergence of the CDFs at each point of continuity t of the CDF of X, F: $\lim_{n\to\infty} F_n(t) = F(t)$, and an expectation can be defined in terms of a CDF.) Then an *asymptotic expectation* of $\{X_n\}$ is $\mathrm{E}(X)$. The resaon we call this "an" asymptotic expectation will become apparent below.

The *limiting expectation* is $\lim_{n\to\infty} \mathrm{E}(X_n)$. It is important to recognize the difference in limiting expectation and asymptotic expectation.

Because $\{X_n\}$ may converge to a degenerate random variable, it may be more useful to generalize the definition of asymptotic expectation slightly.

Let $\{X_n\}$ be a sequence of random variables, and let $\{a_n\}$ be a sequence of positive constants with $\lim_{n\to\infty} a_n = \infty$ or with $\lim_{n\to\infty} a_n = a > 0$, and such that $a_n X_n \to_{\mathrm{d}} X$, with $\mathrm{E}(|X|) < \infty$. Then *an asymptotic expectation* of $\{X_n\}$ is $\mathrm{E}(X/a_n)$.

Notice that this latter definition may allow us to address more general situations. For example, we may consider the asymptotic variance of a sequence of estimators $\sqrt{n}T_n(X)$. The asymptotic variance may be of the form $\mathrm{V}(T/n)$ (which we should not be tempted to say is just 0, because $n \to \infty$).

**** choice of sequence $\{a_n\}$ ********** Shao Proposition 2.3, p 136.

The multivariate generalization of asymptotic expectation is straightforward:

Let $\{X_n\}$ be a sequence of random k-vectors, and let $\{A_n\}$ be a sequence of $k\times k$ positive definite matrices such that either $\lim_{n\to\infty} A_n$ diverges (that is, in the limit has no negative diagonal elements and some diagonal elements that are positively infinite) or else $\lim_{n\to\infty} A_n = A$, where A is positive definite and such that $A_n X_n \to_{\mathrm{d}} X$, with $\mathrm{E}(|X|) < \infty$. Then *an asymptotic expectation* of $\{X_n\}$ is $\mathrm{E}(A_n^{-1}X)$.

2.5.3 Asymptotic Properties and Limiting Properties

After defining asymptotic expectation, we noted an alternative approach based on a limit of the expectations, which we distinguished by calling it the limiting expectation. These two types of concepts persist in properties of interest that are defined in terms of expectations, such as bias and variance and their combination, the mean-squared error.

One is based on the asymptotic distribution and the other is based on limiting moments. Although in some cases they may be the same, in general they are different, as we will see.

Asymptotic Bias and Limiting Bias

Now consider a sequence of estimators $\{T_n(X)\}$ for $g(\theta)$ in the family of distributions $\mathcal{P} = \{P_\theta\}$. Suppose $T_n(X) \to_{\mathrm{d}} T$ and $\mathrm{E}(|T|) < \infty$. We define the *asymptotic bias* of $\{T_n\}$ within the family $\mathcal{P}$ to be $\mathrm{E}(T) - g(\theta)$.

Notice that the bias may be a function of θ; that is, it may depend on the specific distribution within $\mathcal{P}$.

If the asymptotic bias is 0 for any distribution within $\mathcal{P}$, we say $\{T_n(X)\}$ is *asymptotically unbiased* for $g(\theta)$.

We also generalize the asymptotic bias to an asymptotic bias of $\{T_n\}$ given a sequence of positive constants $\{a_n\}$ with $\lim_{n\to\infty} a_n = \infty$ or with $\lim_{n\to\infty} a_n = a > 0$, and such that $a_n T_n(X) \to_{\mathrm{d}} T$. An asymptotic bias of $\{T_n\}$ is $\mathrm{E}(T - g(\theta))/a_n$.

We define the *limiting bias* of $\{T_n\}$ to be $\lim_{n\to\infty} \mathrm{E}((T_n) - g(\theta))$.

We can easily construct an estimator that is biased in any finite sample, but is unbiased in the limit. Suppose we want an estimator of the mean μ (which we assume is finite). Let

$$T_n = \overline{X}_n + \frac{c}{n},$$

for some $c \neq 0$. Now, the bias for any n is c/n. The limiting bias of T_n for μ, however, is 0, and since this does not depend on μ, we say it is unbiased in the limit.

To carry this further, suppose $X_1, \ldots, X_n \sim$ i.i.d. $\mathrm{N}(\mu, \sigma^2)$, and with T_n as above, form $\sqrt{n}(T_n - \mu) = \sqrt{n}(\overline{X}_n - \mu) + c/\sqrt{n}$. Now, we know $\sqrt{n}(\overline{X}_n - \mu) \to_{\mathrm{d}} \mathrm{N}(0, \sigma^2)$ and $c/\sqrt{n} \to 0$, so by Slutsky's theorem,

$$\sqrt{n}(T_n - \mu) \to_{\mathrm{d}} \mathrm{N}(0, \sigma^2).$$

Hence, the asymptotic bias of T_n for μ is also 0, and since this does not depend on μ, we say it is asymptotically unbiased.

To illustrate the difference in asymptotic bias and limiting bias, consider $X_1, \ldots, X_n \sim$ i.i.d. $\mathrm{U}(0, \theta)$, and the estimator $X_{(n)}$ (which we know to be sufficient for $g(\theta) = \theta$). We can work out the asymptotic distribution of $n(\theta - X_{(n)})$ to be exponential with parameter θ. (The distributions of the order statistics from the uniform distribution are betas. These distributions are interesting and you should become familiar with them.) Hence, $X_{(n)}$ is asymptotically biased. We see, however, that the limiting bias is $\lim_{n\to\infty} \mathrm{E}(X_{(n)} - \theta) = \frac{n-1}{n}\theta - \theta = 0$; that is, $X_{(n)}$ is unbiased in the limit.

Notice the role that the sequence $\{a_n\}$ plays. This would allow us to construct a sequence that is biased in the limit, but is asymptotically unbiased.

There are also, of course, relationships between consistency and limiting bias. Consider again $X_1, \ldots, X_n \sim$ i.i.d. $\mathrm{N}(\mu, \sigma^2)$, and an estimator of the mean

$$S_n = \overline{X}_n + \frac{c}{\sqrt{n}},$$

for some $c \neq 0$. (Notice this estimator is slightly different from that above.) As above, we see that this is unbiased in the limit (consistent in the mean), and furthermore, we have the mean-squared error

$$\begin{aligned} \mathrm{MSE}(S_n, \mu) &= \mathrm{E}((S_n - \mu)^2) \\ &= \frac{\sigma^2}{n} + \left(\frac{c}{\sqrt{n}}\right)^2 \end{aligned}$$

tending to 0, hence we see that this is consistent in mean-squared error. However, $\sqrt{n}(S_n - \mu) = \sqrt{n}(\overline{X}_n - \mu) + c$ has limiting distribution $\mathrm{N}(c, \sigma^2)$; hence S_n is asymptotically biased.

We also note that an estimator can be asymptotically unbiased but not consistent in mean-squared error. In the above example, we immediately see that X_1 is asymptotically unbiased for μ, but it is not consistent in mean-squared error for μ.

For a more interesting example, consider a distribution with slightly heavier tails than the normal, that is, the double exponential distribution with $\theta = 1$ (Shao, p.21), and the estimator of the mean

$$R_n(X) = \frac{X_{(n)} + X_{(1)}}{2}.$$

(This is the mid-range.) We can see that R_n is unbiased for any finite sample size (and hence, is unbiased in the limit); however, we can show that

$$\mathrm{V}(R_n) = \frac{\pi^2}{12},$$

and, hence, R_n is not consistent in mean-squared error.

Asymptotic and Limiting Variance and Efficiency

We define the asymptotic variance and the limiting variance in similar ways as in defining the asymptotic bias and limiting bias, and we also note that they are different from each other. We also define asymptotic mean-squared error and the limiting mean-squared error in a similar fashion. The limiting mean-squared error is of course related to consistency in mean-squared error.

Our interest in *asymptotic* (not "limiting") variance or mean-squared error is as they relate to optimal properties of estimators. The "efficiency" of an estimator is related to its mean-squared error.

Usually, rather than consider efficiency in a absolute sense, we use it to compare two estimators, and so speak of the *relative efficiency*. (When we restrict our attention to unbiased estimators, the mean-squared error is just the variance, and in that case we use the phrase *efficient* or *Fisher efficient* to refer to an estimator that attains its Cramér-Rao lower bound (the right-hand side of inequality (1.45) on page 30.) Notice the slight difference in "efficiency" and "efficient"; while one meaning of "efficiency" is a relative term that is not restricted to unbiased estimators (or other unbiased procedures, as we will see later), "efficient" is absolute. "Efficient" only applies to unbiased estimators, and an estimator either is or is not efficient. The state of being efficient, of course is called "efficiency". This is another meaning of the term. The phrase "Fisher efficiency" helps to emphasis this difference.)

As before, assume a family of distributions $\mathcal{P}$, a sequence of estimators $\{T_n\}$ of $g(\theta$, and a sequence of positive constants $\{a_n\}$ with $\lim_{n\to\infty} a_n = \infty$ or with $\lim_{n\to\infty} a_n = a > 0$, and such that $a_n T_n(X) \to_{\mathrm{d}} T$ and $0 < \mathrm{E}(T) < \infty$. We define the asymptotic mean-squared error of $\{T_n\}$ for estimating $g(\theta)$ w.r.t. $\mathcal{P}$ as an asymptotic expectation of $(T_n - g(\theta))^2$; that is, $\mathrm{E}((T - g(\theta))^2)/a_n$, which we denote as $\mathrm{AMSE}(T_n, g(\theta), \mathcal{P})$.

For comparing two estimators, we may use the *asymptotic relative efficiency*. The asymptotic relative efficiency of the estimators S_n and T_n for $g(\theta)$ w.r.t. $\mathcal{P}$ is defined as

$$\mathrm{ARE}(S_n, T_n) = \mathrm{AMSE}(S_n, g(\theta), \mathcal{P})/\mathrm{AMSE}(T_n, g(\theta), \mathcal{P}). \tag{2.35}$$

The ARE is essentially a scalar concept; for vectors, we usually do one at a time, ignoring covariances.

Asymptotic Significance

For use of asymptotic approximations for confidence sets and hypothesis testing, we need a concept of asymptotic significance. As for exact significance, the concepts in confidence sets and hypothesis tests are essentially the same.

We assume a family of distributions $\mathcal{P}$, a sequence of statistics $\{T_n\}$ $\{\delta(X_n)\}$ based on a random sample $X_1, \ldots, X_n$. The test statistic $\delta(\cdot)$ is

defined in terms the decisions; it takes the value 1 for the case of deciding to reject H_0 and conclude H_1, and the value 0 for the case of deciding not to reject H_0.

Tests

In hypothesis testing, the standard setup is that we have an observable random variable with a distribution in the family $\mathcal{P}$. Our hypotheses concern a specific member $P \in \mathcal{P}$. We have a null hypothesis

$$H_0 : \quad P \in \mathcal{P}_0$$

and an alternative hypothesis

$$H_1 : \quad P \in \mathcal{P}_1,$$

where $\mathcal{P}_0 \subset \mathcal{P}$, $\mathcal{P}_1 \subset \mathcal{P}$, and $\mathcal{P}_0 \cap \mathcal{P}_1 = \emptyset$.

Asymptotic size and asymptotic significance

Now, letting

$$\beta(\delta(X_n), P) = \Pr_P(\delta(X_n) = 1),$$

we define $\limsup_n \beta(\delta(X_n), P) \, \forall P \in \mathcal{P}_0$, if it exists, as the *asymptotic size* of the test. If $\limsup_n \beta(\delta(X_n), P) \leq \alpha \, \forall P \in \mathcal{P}_0$, then α is an *asymptotic significance level* of the test.

Consistency

$\delta(X_n)$ is *consistent* for the test iff $\limsup_n(1 - \beta(\delta(X_n), P)) = 0 \, \forall P \in \mathcal{P}_1$.

Chernoff consistency

$\delta(X_n)$ is *Chernoff-consistent* for the test iff $\delta(X_n)$ is consistent and furthermore, $\limsup_n \beta(\delta(X_n), P) = 0 \, \forall P \in \mathcal{P}_0$.

Confidence sets

Let $C(X)$ be a confidence set for $g(\theta)$.

Asymptotic significance level .

If $\liminf_n \Pr(g(\theta) \in C(X)) \geq 1-\alpha \, \forall P \in \mathcal{P}$, then $1-\alpha$ is an *asymptotic significance level* of $C(X)$.

Limiting confidence coefficient .

If $\liminf_n \Pr(g(\theta) \in C(X))$ exists $\forall P \in \mathcal{P}$, then it is the *limiting confidence coefficient* of $C(X)$.

"The" Asymptotic Distribution

In determining asymptotic confidence sets or asymptotic relative efficiencies, we need expressions that do not depend on unknown parameters. This fact determines which asymptotic expectations are useful.

The asymptotic expectation of some sequence of statistics, or of pivotal quantities, is determined by the sequence $\{a_n\}$ (used above in the definitions).

In the univariate delta method, for example, we find a quantity $a_n(g(X_n) - g(c))$ that converges in distribution to $\mathrm{N}(0, v)$, where v does not depend on

an unknown parameter. In that case, we can set a confidence interval based on the approximate distribution of $g(X_n)$ as $\mathrm{N}(g(c), v/a_n^2)$.

To speak of the asymptotic distribution of $a_n(g(X_n) - g(c))$ is clear; but to refer to "the" asymptotic distribution of $g(X_n)$ is somewhat less so.

Because it is the useful approximate distribution resulting from asymptotic expectations, we often say that "the asymptotic distribution" of $g(X_n)$ is $\mathrm{N}(g(c), v/a_n^2)$. You should recognize that "the" in this statement is somewhat arbitrary. It might be better to call it "the asymptotically approximate distribution that I'm going to use in this application".

Again, we should distinguish "asymptotic" from "limiting".

In the example of the delta method above, it is likely that

$$g(X_n)) \to_{\mathrm{d}} g(c);$$

that is, $g(X_n)$ converges in distribution to the constant $g(c)$; or the limiting distribution of $g(X_n)$ is degenerate at $g(c)$. "The" asymptotic variance is 0.

This would not be very useful in asymptotic inference. We therefore seek "an" asymptotic variance that is more useful. In asymptotic estimation using $g(X_n)$, we begin with an expression of the form $a_n(g(X_n) - g(c))$ that has a limiting distribution of the desired form (usually that means such that the variance does not involve any unknown parameters and it does not involve n). If this distribution is in a location-scale family, then we make the appropriate linear transformation (which probably results in a variance that does involve n).

We then often refer to this as the asymptotic distribution of $g(X_n)$. Often, as mentioned above, however, the limiting distribution of $g(X_n)$ is degenerate.

This is not to imply that asymptotic expectations are entirely arbitrary. Proposition 2.3 in Shao shows that there is a certain uniqueness in the asymptotic expectation. This proposition involves three cases regarding whether the expectation of $g(X_n)$ (without the a_n sequence) is 0. In the example above, we have a degenerate distribution, and hence the asymptotic expectation that defines the asymptotic variance is 0.

2.6 Variance Estimation

Statistical inferences that involve or are derived from statements of probability, such as hypothesis testing and setting confidence regions, require knowledge of the distribution of the statistic that is used. Often we know or can work out that distribution exactly, given the assumptions in the underlying probability model. In other cases we use approximate distributions. In either case, we are often faced with the problem of estimating the variance of a statistic.

In this section we first restrict our attention to the case in which the statistic of interest is a scalar; that is, the case in which the variance itself is

a scalar. We describe two general methods, the jackknife and the bootstrap, based on resampling. We then consider the more general problem of estimating the variance-covariance matrix for a vector statistic. The first consideration for estimators of a variance-covariance matrix is the meaning of consistency of a variance-covariance estimator. The jackknife and bootstrap can be used to estimate a variance-covariance matrix, and we also consider a "substitution" estimator.

2.6.1 Jackknife

Jackknife methods make use of systematic partitions of a dataset to estimate properties of an estimator computed from the full sample.

Suppose that we have a random sample, $Y_1, \ldots, Y_n$, from which we compute a statistic T as an estimator of a parameter θ in the population from which the sample was drawn. In the jackknife method, we partition the given dataset into r groups, each of size k. (For simplicity, we will assume that the number of observations n is kr.)

Now, we remove the j^{th} group from the sample and compute the estimator from the reduced sample. Let $T_{(-j)}$ denote the estimator computed from the sample with the j^{th} group of observations removed. (This sample is of size $n-k$.) The estimator $T_{(-j)}$ has properties similar to those of T. For example, if T is unbiased, so is $T_{(-j)}$. If T is not unbiased, neither is $T_{(-j)}$; its bias, however, is likely to be different.

The mean of the $T_{(-j)}$,

$$\overline{T}_{(\bullet)} = \frac{1}{r}\sum_{j=1}^{r} T_{(-j)}, \tag{2.36}$$

can be used as an estimate of θ. The $T_{(-j)}$ can also be used in some cases to obtain more information about the estimator T from the full sample. (For the case in which T is a linear functional of the ECDF, then $\overline{T}_{(\bullet)} = T$, so the systematic partitioning of a random sample will not provide any additional information.)

Consider the weighted differences in the estimate for the full sample and the reduced samples:

$$T_j^* = rT - (r-1)T_{(-j)}. \tag{2.37}$$

The T_j^* are called "pseudovalues". (If T is a linear functional of the ECDF and $k = 1$, then $T_j^* = T(x_j)$; that is, it is the estimator computed from the single observation, x_j.) We call the mean of the pseudovalues the "jackknifed" T and denote it as $\mathrm{J}(T)$:

$$\begin{aligned} \mathrm{J}(T) &= \frac{1}{r}\sum_{j=1}^{r} T_j^* \\ &= \overline{T}^*. \end{aligned} \tag{2.38}$$

We can also write $\mathrm{J}(T)$ as

$$\mathrm{J}(T) = T + (r-1)\left(T - \overline{T}_{(\bullet)}\right)$$

or

$$\mathrm{J}(T) = rT - (r-1)\overline{T}_{(\bullet)}. \tag{2.39}$$

In most applications of the jackknife, it is common to take $k = 1$, in which case $r = n$. It has been shown that this choice is optimal under certain assumptions about the population.

Jackknife Variance Estimate

Although the pseudovalues are not independent (except when T is a linear functional), we treat them as if they were independent, and use $\mathrm{V}(\mathrm{J}(T))$ as an estimator of the variance of T, $\mathrm{V}(T)$. The intuition behind this is simple: a small variation in the pseudovalues indicates a small variation in the estimator. The sample variance of the mean of the pseudovalues can be used as an estimator of $\mathrm{V}(T)$:

$$\widehat{\mathrm{V}(T)}_{\mathrm{J}} = \frac{\sum_{j=1}^{r}\left(T_j^* - \mathrm{J}(T)\right)^2}{r(r-1)}. \tag{2.40}$$

(Notice that when T is the mean and $k = 1$, this is the standard variance estimator.) From expression (2.40), it may seem more natural to take $\widehat{\mathrm{V}(T)}_{\mathrm{J}}$ as an estimator of the variance of $\mathrm{J}(T)$, and indeed it often is.

A variant of this expression for the variance estimator uses the original estimator T:

$$\frac{\sum_{j=1}^{r}(T_j^* - T)^2}{r(r-1)}. \tag{2.41}$$

How good a variance estimator is depends on the estimator T and on the underlying distribution. Monte Carlo studies indicate that $\widehat{\mathrm{V}(T)}_{\mathrm{J}}$ is often conservative; that is, it often overestimates the variance.

The alternate expression (2.41) is greater than or equal to $\widehat{\mathrm{V}(T)}_{\mathrm{J}}$, as is easily seen; hence, it is an even more conservative estimator.

2.6.2 Bootstrap

From a given sample $y_1, \ldots, y_n$, suppose that we have an estimator $T(y)$. The estimator T^* computed as the same function T, using a bootstrap sample (that is, $T^* = T(y^*)$), is a *bootstrap observation* of T.

The bootstrap estimate of some function of the estimator T is a plug-in estimate that uses the empirical distribution P_n in place of P. This is the bootstrap principle, and this bootstrap estimate is called the *ideal bootstrap*.

For the variance of T, for example, the ideal bootstrap estimator is the variance $\mathrm{V}(T^*)$. This variance, in turn, can be estimated from bootstrap samples. The bootstrap estimate of the variance, then, is the sample variance of T^* based on the m samples of size n taken from P_n:

$$\widehat{\mathrm{V}}(T) = \widehat{\mathrm{V}}(T^*) \tag{2.42}$$

$$= \frac{1}{m-1}\sum(T^{*j} - \overline{T}^*)^2, \tag{2.43}$$

where T^{*j} is the j^{th} bootstrap observation of T. This, of course, can be computed by Monte Carlo methods by generating m bootstrap samples and computing T^{*j} for each.

If the estimator of interest is the sample mean, for example, the bootstrap estimate of the variance is $\widehat{\mathrm{V}}(Y)/n$, where $\widehat{\mathrm{V}}(Y)$ is an estimate of the variance of the underlying population. (This is true no matter what the underlying distribution is, as long as the variance exists.) The bootstrap procedure does not help in this situation.

2.6.3 Consistency of Estimators of a Variance-Covariance Matrix

If the statistic is a vector, we need an estimator of the variance-covariance matrix. Because a variance-covariance matrix is positive definite, it is reasonable to consider only positive definite estimators a.s.

We first define what it means for such an estimator to be consistent. Because a variance-covariance matrix is positive definite and any positive definite matrix is a variance-covariance matrix (for some distribution), we can consider consistency of a sequence of positive definite matrices for a sequence of given positive definite matrices.

Let $\{V_n\}$ be a sequence of $k \times k$ positive definite matrices and $\widehat{V}_n$ be a positive definite matrix estimator of V_n for each n. Then $\widehat{V}_n$ is said to be consistent for V_n if

$$\left\| V_n^{-1/2}\widehat{V}_n V_n^{-1/2} - I_k \right\| \to_{\mathrm{p}} 0. \tag{2.44}$$

Also $\widehat{V}_n$ is said to be strongly consistent for V_n if

$$\left\| V_n^{-1/2}\widehat{V}_n V_n^{-1/2} - I_k \right\| \to_{\mathrm{a.s.}} 0. \tag{2.45}$$

Note the similarity of these expressions to expression (2.34). In many cases of interest $\|V_n\| \to 0$, so these expressions are not the same as $\left\|\widehat{V}_n - V_n\right\| \to 0$.

2.6.4 Methods of Estimating Variance-Covariance Matrices

The jackknife and bootstrap can be used to estimate a variance-covariance estimator. Another widely used type of estimator is called a substitution estimator or sandwich estimator.

Substitution Method

The idea in the "substitution method" for estimating V_n is to arrive at an expression for V_n that involves a simpler variance along with quantities that are known functions of the sample. Often that simpler variance can be estimated by an estimator with known desirable properties. An estimator of V_n in which the simpler estimator and the known sample functions are used is called a substitution estimator. A simple example is the estimator of the variance of $\widehat{\beta}$ in a linear regression. The variance-covariance matrix is $(Z^{\mathrm{T}}Z)^{-1}\sigma^2$ (in Shao's usual notation). A substitution estimator is one in which the regression MSE is substituted for σ^2.

The so-called "sandwich estimators" are often substitution estimators.

$$(Z^{\mathrm{T}}Z)^{-1}V(Z^{\mathrm{T}}Z)^{-1}$$

V is some variance-covariance estimator that probably includes a scalar multiple of $\hat{\sigma}^2$.

Theorem 5.15 in Shao (page 374) gives conditions for the consistency of substitution estimators.

Notes

The general problem of statistical inference, that is, the use of observed data for which we have a family of probability distributions to provide information about those probability distributions, is an "inverse problem". Nineteenth and twentieth century scientists who made inferences about probability models referred to the problem as one of "inverse probability". Statisticians in the early twentieth century also used this term. Although the maximum likelihood approach could be thought of as a method of inverse probability, R. A. Fisher, who developed likelihood methods, made a distinction between the methods and "inverse probability" as a general term fell into disuse.

Optimal Properties

Although unbiasedness is most often encountered in the context of point estimation, the term "unbiased" was actually first used by statisticians to refer to tests (Neyman and Pearson, 1936, cited in Lehmann, 1951), then used to refer to confidence regions (Neyman, 1937, cited in Lehmann, 1951), and lastly introduced to refer to point estimators (David and Neyman, 1938, cited in Lehmann, 1951). See Lehmann (1951) for general discussions, and see page 108 for unbiased tests and page 112 for unbiased confidence sets.

Ideas and approaches developed by engineers and physical scientists lead to statistical methods characterized by maximum entropy. Much of this work

dates back to Claude Shannon in the 1930's. E. T. Jaynes in the 1950's formalized the approach and incorporated it in a Bayesian framework. His posthumous book edited by G. Larry Bretthorst (Jaynes, 2003) is a very interesting discussion of a view toward probability that leads to a Bayesian maximum entropy principle for statistical inference. Pardo (2005), listed in the general references, gives an extensive overview of the use of functionals from information theory in statistical inference. In some disciplines, such as electrical engineering, this approach seems more natural.

Pitman's measure of closeness was introduced in 1937. The idea did not receive much attention until the article by Rao (1981), in which was given the definition we have used, which is slightly different from Pitman's. Pitman's original article was reproduced in a special issue of *Communications in Statistics* (Pitman, 1991) devoted to the topic of Pitman closeness. The lack of transitivity of Pitman's closeness follows from Arrow's "impossibility theorem", and is a natural occurrence in paired comparisons (see David, 1988). The example on page 72 is called a "cyclical triad".

David and Salem had considered estimators similar to (2.3) for a normal mean in 1973, and in David and Salem (1991) they generalized these shrunken estimators to estimators of the means that are Pitman-closer than the sample mean in a broad class of location families.

Variance Estimation

The idea of the jackknife goes back to Quenouille in 1949. It was popularized by John Tukey, and is currently widely-used, especially in sample surveys. Shao and Tu (1995) provide an extensive discussion.

The theory and methods of the bootstrap were largely developed by Efron, and Efron and Tibshirani (1993) introduce the principles and discuss many extensions and applications.

A sandwich-type estimator was introduced introduced by Eiker (1963) for estimation of the variance-covariance matrix of the least-squares estimator of the coefficient vector in linear regression in the case where the errors are uncorrelated, but possibly have different distributions. Huber (1967) used a similar kind of estimator as a robust estimator. White (1980) introduced a similar estimator for heteroscedastic situations in economics. The term "sandwich estimator" was introduced in the context of estimation of the variance-covariance matrix for the solution of a generalized estimation equation, and it is widely used in that type of problem.

Exercises in Shao

- For practice and discussion
  2.25, 2.30, 2.44, 2.56, 2.66, 2.74, 2.84, 2.93, 2.101, 2.115, 2.121, 4.89, 4.91
  (Solutions in Shao, 2005)

- To turn in
  2.33, 2.55, 2.63, 2.81, 2.116, 2.123

Additional References

Blyth, Colin R. (1972), Some probability paradoxes in choice from among random alternatives, *Journal of the American Statistical Association* **67**, 366–373.

Efron, B. (1975), Biased versus unbiased estimation, *Advances in Mathematics* **16**, 259–277.

Efron, Bradley, and Robert J. Tibshirani (1993), *An Introduction to the Bootstrap*, Chapman & Hall, New York.

Eiker, F. (1963), Asymptotic normality and consistency of the least squares estimators for families of linear regressions, *Advances in Mathematics* **16**, 259–277.

David, H. A. (1988), *The Method of Paired Comparisons*, second edition, Griffith/Oxford University Press, London.

David, H. T., and Shawki A. Salem (1991), Three shrinkage constructions for Pitman-closeness in the one-dimensional location case, *Communications in Statistics — Theory and Methods* **20**,3605–3627.

Huber, Peter J. (1967), The behavior of maximum likelihood estimates under nonstandard conditions, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume I, 221–233.

Jaynes, E. T. (2003), *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK.

Lehmann, E. L. (1951), A general concept of unbiasedness, *The Annals of Mathematical Statistics* **22**, 587–592.

Pitman, E. J. G. (1991), The "closest" estimates of statistical parameters, *Communications in Statistics — Theory and Methods* **20**,3423–3437.

Rao, C. R. (1981), Some comments on the minimum mean square as a criterion of estimation, *Statistics and Related Topics*, (edited by M. Csörgö, D. A. Dawson, J. N. K. Rao, and A. K. Md. E. Saleh), North-Holland, Amsterdam, 123–143.

Shao, Jun, and Dongsheng Tu (1995), *The Jackknife and Bootstrap*, Springer-Verlag, New York.

White, Halbert (1980), A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica* **48**, 817–838.

3

Bayesian Inference (Shao Sec 4.1, Sec 6.4.4, Sec 7.1.3; TPE2 Ch 4; TSH3 Sec 5.7)

In the decision-theoretic approach to statistical inference, we first quantify the loss relative to the true state of nature in any decision that we make. Of course, since we do not know the true state of nature, we must use an expectation of the loss; and in fact, that is exactly what we do: we take the expected value of the loss under some assumed family of distributions. This is the risk.

Then we see an inference procedure that minimizes the risk.

How to minimize?

Generally, we cannot minimize the risk uniformly; that is, for all states of nature. The optimal procedure at one state of nature or one value of the true parameter is different from the optimal procedure at another state.

We can put restrictions on our inference procedure (for example, we may require that an estimator be unbiased) and maybe we can get an optimal solution by minimizing the risk under the restriction.

We could also minimize some global property of the risk, such as its maximum value (a "minimax procedure") or we could minimize the "average" risk over the full parameter space Θ.

How to average?

Choose a function $\Pi(\theta)$, such that $\int_\Theta \mathrm{d}\Pi(\theta) = 1$. Then the average risk with respect to Π, for a given X is

$$r(\Pi, X) = \int_\Theta R(\theta, X)\mathrm{d}\Pi(\theta). \tag{3.1}$$

3.1 The Bayesian Paradigm

The function $\Pi(\theta)$ in equation (3.1) is effectively a CDF over Θ, and with that interpretation, the average risk is the conditional expectation with respect to the distribution. In our usual notation, we might denote Π as P_Θ to emphasize that it is the CDF of Θ. We likewise denote the CDF of the conditional distribution as $P_{\Theta|x}$.

The estimator that minimizes the average risk w.r.t. $\Pi(\theta)$, say $\delta_\Pi(X)$, is called the *Bayes estimator*. The minimum risk, which is achieved at that point, $\int_\Theta R(\theta, \delta_\Pi(X))\mathrm{d}\Pi(\theta)$, is called the *Bayes risk*.

The action that minimizes the posterior risk, that is, the action that achieves the Bayes risk, is called a *Bayes rule*.

The averaging function allows various interpretations, and it allows the flexibility of incorporating prior knowledge or beliefs. The regions over which $\Pi(\theta)$ is large will be given more weight; therefore the estimator will be pulled toward those regions.

Because the averaging function is essentially a probability density, we might formally consider the parameter to be a random variable. This is the paradigm of "Bayesian statistics".

In formal Bayes procedures, we call $\Pi(\theta)$ the prior probability density for θ, and to emphasize that perspective, we may write it as P_Θ. The distribution of Θ may depend on parameters. Such parameters are called "hyperparameters".

We next form the joint distribution of θ and X, and then the conditional distribution of θ given X, called the *posterior distribution*. The Bayes estimator is determined by minimizing the risk, where the expectation is taken with respect to the posterior distribution. Because the Bayes estimator is determined by the posterior distribution, the Bayes estimator must be a function of a sufficient statistic.

The relationships among the conditional, marginal, and joint distributions can be stated formally in the "Bayes formula" (Theorem 4.1 in Shao). We begin with a distributional family as usual, except we emphasize that it is a conditional distribution, given the parameter: $\mathcal{P} = \{P_{x|\theta} : \theta \in \Theta\}$, which we assume is dominated by a σ-finite measure ν.

Now, keeping the θ in the picture, we assume the density, $p_{X|\theta}(x) = \mathrm{d}P_{X|\theta}(x)/\mathrm{d}\nu$ is Borel on the product measure space, $(\mathcal{X} \times \Theta, \sigma(\mathcal{B}_\mathcal{X} \times \mathcal{B}_\Theta))$. We have that the posterior distribution $P_{\Theta|x}$ is dominated by P_Θ, and if $p_X(x) = \int_\Theta p_{X|\theta}(x)\mathrm{d}P_\Theta > 0$,

$$\frac{\mathrm{d}P_{\theta|x}}{\mathrm{d}P_\Theta} = \frac{p_{X|\theta}(x)}{p_X(x)}.$$

(Shao calls the marginal density of x $m(x)$.)

Furthermore, if λ is a σ-finite measure such that $P_\Theta \ll \lambda$ and $p_\Theta(\theta) = \frac{\mathrm{d}P_\Theta}{\mathrm{d}\lambda}$, then

$$\frac{\mathrm{d}P_{\Theta|x}}{\mathrm{d}\lambda} = \frac{p_{X|\theta}(x)p_\Theta(\theta)}{p_X(x)}.$$

Steps in a Bayesian Analysis

We can summarize the approach in a Bayesian statistical analysis as beginning with these steps:

1. identify the conditional distribution of the observable random variable; assuming the density exists, call it

$$p_{X|\theta}(x|\theta) \qquad \text{(Shao calls this } f_\theta(x)\text{)}$$

2. identify the prior (marginal) distribution of the parameter; assuming the density exists, call it

$$p_\Theta(\theta) \qquad \text{(Shao calls this } \pi(\theta)\text{)}$$

3. identify the joint distribution; if densities exist, it is

$$p_{X,\Theta}(x,\theta) = p_{X|\theta}(x|\theta)p_\Theta(\theta)$$

4. determine the marginal distribution of the observable; if densities exist, it is

$$p_X(x) = \int_\Theta p_{X,\Theta}(x,\theta)\mathrm{d}\theta \qquad \text{(Shao calls this } m(x)\text{)}$$

5. determine the posterior conditional distribution of the parameter given the observable random variable; this is the posterior; if densities exist, it is

$$p_{\Theta|x}(\theta|x) = p_{X,\Theta}(x,\theta)/p_X(x)$$

The posterior conditional distribution is then the basis for whatever decisions are to be made.

Interpretations of Probability Statements in Statistical Inference

Some methods of statistical inference are based on probabilities of a statistic taking on certain values given a specific member of a family of probability distributions; that is, perhaps, given a value of a parameter. The two main statistical methods that rely on statements of probability are hypothesis testing and setting of confidence regions. In these methods we assume a model P_θ for the state of nature and then consider probabilities of the form $\Pr(T(X) = 1|\theta)$ or $\Pr(T(X) \ni \theta|\theta)$. The proper interpretation of a confidence region, for example, is "[... given the assumptions, etc. ...] the probability that a random region formed in this manner includes true value of the parameter is ..."

These kinds of probability statements are somewhat awkward for use in interpreting the results of a statistical analysis.

Instead of a statement about $\Pr(\delta(X)|\theta)$, many people would prefer a statement about $\Pr(\Theta \in T(X)|X = x)$, that is,

$$\Pr(\Theta \in T(x))$$

even if they don't think of Θ as a random variable. In the Bayesian approach to testing and setting confidence regions, we do think of the parameter as a

random variable and so we can make statements about the probability of the parameter taking on certain values.

If the parameter is a random variable, point estimation of the parameter or testing an hypothesis that a parameter takes a specific value when the parameter is modeled as a continuous random variable does not make much sense. The idea of a point estimator that formally minimizes the Bayes risk, however, remains viable. Going beyond point estimation, the Bayesian paradigm provides a solid theoretical infrastructure for other aspects of statistical inference, such as confidence intervals and tests of hypotheses. The parameter random variable is different in a fundamental way from the other random variable in the estimation problem: the parameter random variable is not observable; the other random variable is — that is, we can observe and record realizations of this random variable of interest, and those observations constitute the sample which is the basis for the statistical inference.

The starting point in ordinary Bayesian inference is the conditional distribution of the observable random variable. (In a frequentist approach, this is just the distribution — not the "conditional" distribution.)

The prior density represents a probability distribution of the parameter assumed a priori, that is, without the information provided by a random sample. Bayesian inference depends on the conditional distribution of the parameter, given data from the random variable of interest.

Prior Distributions

The prior distribution obviously has an effect on Bayesian decisions. It is important, therefore, to choose a reasonable prior distribution that reflects our prior information or beliefs about the phenomenon of interest.

Various families of prior distributions can provide both flexibility in representing prior beliefs and computational simplicity. For many families of distributions of the observable random variable, there are corresponding families of prior distributions that yield a family of posterior distributions that is the same as the family of priors. We call a member of such a family of priors a *conjugate prior* with respect to the conditional distributional family of the observables.

Example 3.1 Binomial with Beta Prior
The Bayesian approach can be represented nicely by a problem in which we model the conditional distribution of an observable random variable as a binomial(π, n) distribution, conditional on π, of course. Suppose we assume π comes from a beta(α, β) prior distribution. This is a conjugate prior, as we will see.

We work out the density functions in the following order:
The conditional distribution of X given π has density (probability function)

$$f_{X|\pi}(x) = \binom{n}{x} \pi^x (1-\pi)^{n-x} \quad x = 1, \ldots, n.$$

The marginal (prior) distribution of Π has density

$$f_\Pi(\pi) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}, \quad \pi \in (0,1).$$

Suppose the hyperparameters in the beta prior as $\alpha = 3$ and $\beta = 5$. The prior, that is, the marginal distribution of Π, is as shown in Figure 3.1.

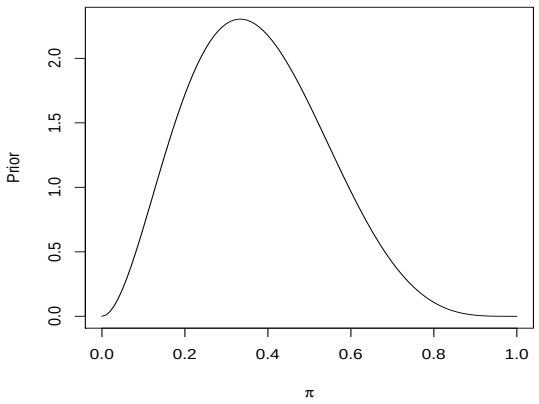


Fig. 3.1. Prior; $\alpha = 3$ and $\beta = 5$

The joint distribution of X and π has density

$$f_{X,\Pi}(x,\pi) = \binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{x+\alpha-1}(1-\pi)^{n-x+\beta-1}.$$

The marginal distribution of X is beta-binomial, with density

$$f_X(x) = \binom{n}{x}\frac{\Gamma(\alpha+\beta)\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(n+\alpha+\beta)}.$$

Finally, the conditional distribution of π given x (the posterior) has density,

$$f_{\Pi|x}(\pi) = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)}\pi^{x+\alpha-1}(1-\pi)^{n-x+\beta-1}.$$

Now, suppose n is 10 and we take one observation, and we observe $x = 2$. With the beta(3,5) prior, we get the posterior, that is, the conditional distribution of Π, as a beta with parameters $x+\alpha = 5$ and $n-x+\beta = 13$. The posterior density is shown in Figure 3.2.

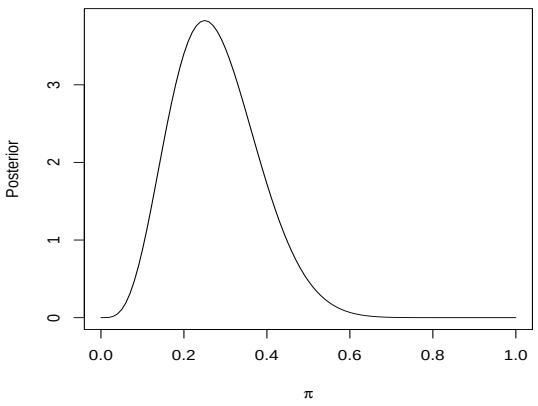


Fig. 3.2. Posterior after Observing $x = 2$

In a Bayesian analysis, sometimes it it worthwhile to consider the effects of various possible outcomes of the experiment. (Bayesians usually refer to the taking of observations as an "experiment"; most people would just refer to it as "taking a sample".)

Assessing the Problem Formulation

In any statistical analysis, the formulation of a model is important. In the example above, it must be reasonable to assume that the observable data follows some kind of binomial distribution. From first principles, this means that we are willing to assume that there is a set of n independent outcomes that may be 0 or 1, in each case with a constant probability π.

The purpose of our analysis is to gain more knowledge about π.

In the Bayesian approach taken in the example, we assume that while the n observations were being collected, some random variable Π had a fixed value of π. We are interested both in that value and in the conditional distribution of the random variable Π, given what we have observed. For a particular choice of hyperparameters characterizing the prior distribution on Π, we obtain the posterior distribution shown in Figure 3.2. Does this seem reasonable? What if instead of observing $x = 2$, we had observed some other value?

Without actually taking any observations, we can determine the posterior density. In Figure 3.3, we plot the posterior distribution of Π given various values of that we might have observed.

Assessing the effect on the posterior of various possible observations may give us some feeling of confidence in our choice of a prior distribution.

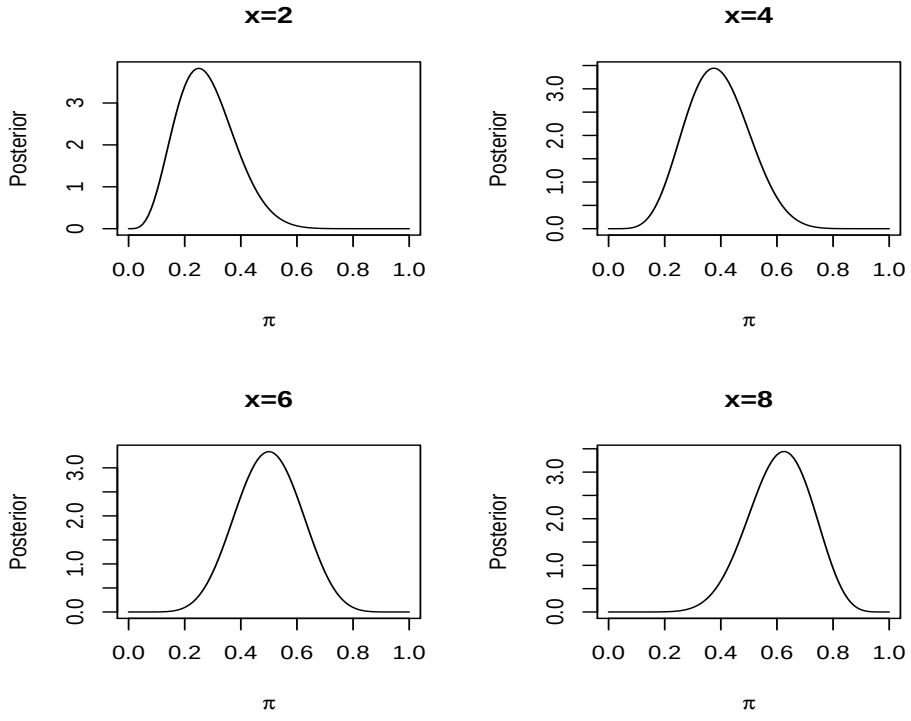


Fig. 3.3. Posteriors after Various Possible Observations

Choice of Hyperparameters

Usually in a Bayesian analysis, it is instructive to consider various priors and particularly various hyperparameters in some detail.

Of course, in most cases, we must also take into account the loss function. Recall the effects in this problem of different hyperparameter values on the point estimation problem (that is, the choice of the Bayes action to minimize the posterior risk) when the loss is squared error.

We might also consider what might be the effect of different hyperparameters. There are several possibilities we could consider. Let's just look at one possibility, which happens to be bimodal, as shown in Figure 3.4. In this case, we have chosen $\alpha = 0.1$ and $\beta = 0.2$. This would correspond to a general prior belief that π is probably either close to 0 or close to 1.

Now, again we might consider the effect of various observations on our belief about Π. We get the posteriors shown in Figure 3.5 for various possible values of the observations.

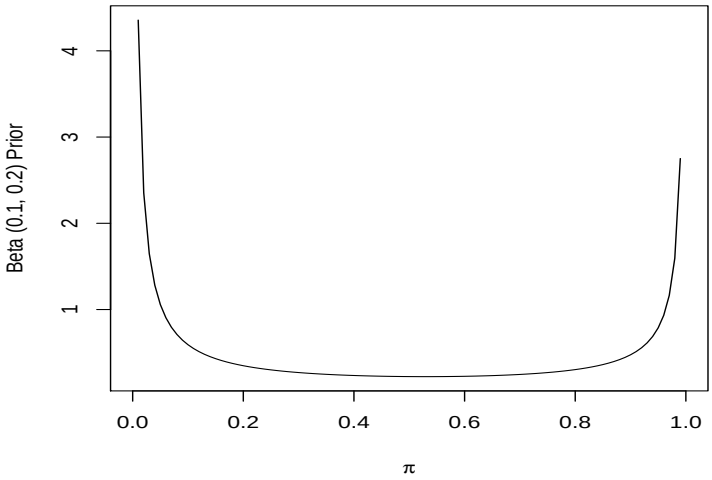


Fig. 3.4. Bimodal Prior

In each case in this example we see that our posterior belief is unimodal instead of bimodal, as our prior belief had been. Although in general a posterior may be multimodal, in the case of a binomial (π, n) distribution with a beta(α, β) prior, the posterior is unimodal, because as we have seen, the posterior is beta with parameters $x + \alpha$ and $n - x + \beta$, both of which cannot be less than 1.

Generalized Bayes Actions and Limits of Bayes Actions

Suppose we rewrite the risk slightly. A *generalized Bayes action* is one that minimizes

$$\int_{\Theta} L(\theta, \delta(x)) p_{X|\Theta}(x|\theta) \mathrm{d}\Pi(\theta),$$

if the integral exists, even if Π is not a distribution function. If Π is not a distribution function, it is called an *improper prior*.

An example is one in which $\mathrm{d}\Pi(\theta) = \mathrm{d}\nu$ where ν is Lebesgue and Θ is the reals, or some unbounded interval subset of the reals. This is a "noninformative" prior, in the sense that it gives equal weights to equal-length intervals for Θ.

Another type of noninformative prior is Jeffreys's noninformative prior, which is proportional to $\sqrt{\det(I(\theta))}$, where $\det(I(\theta))$ is the determinant of the Fisher information matrix. If Θ is the reals, or some unbounded interval subset of the reals, Jeffreys's noninformative prior is improper.

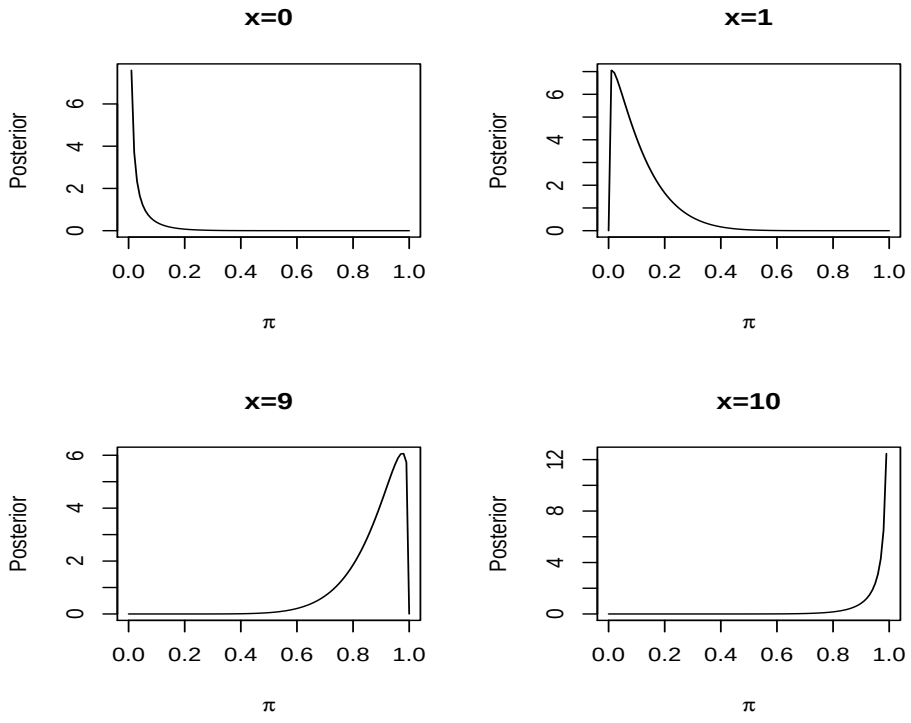


Fig. 3.5. Posteriors from the Bimodal Prior

Another variation on Bayes actions is the limit of the Bayes action as some hyperparameters approach some limits. These are called *limiting Bayes actions*.

3.2 Bayesian Estimation

In Bayesian estimation we begin with the standard steps on page 128.

After getting the posterior conditional distribution of the parameter given the observable random variable, for a given loss function L, we determine the estimator δ that minimizes the posterior risk, which is the expected value of the loss w.r.t. the posterior distribution on the parameter:

$$\int_{\Theta} L(\theta, \delta(x)) p_{\Theta|x}(\theta|x) \mathrm{d}\theta$$

The Bayes estimator depends on

- the conditional distribution of the observable
- the prior distribution of the parameter
- the function of the parameter to be estimated, that is, the estimand
- the loss function

The expected loss with respect to the posterior distribution of the parameter is the objective to be minimized.

A useful fact is that if the loss is squared error, the Bayes estimator is the posterior mean; that is, the expected value of the estimand, where the expected value is taken w.r.t. the posterior conditional distribution.

3.2.1 Properties of Bayes Estimators

Squared-error loss and a conjugate prior yield Bayes estimators for $\mathrm{E}(X)$ that are linear in X.

- A Bayes estimator with a constant risk is minimax with respect to the same loss function and distribution.
- A unique Bayes estimator is admissible with respect to the same loss function and distribution.
- An admissible estimator is either Bayes or limiting Bayes.

3.2.2 Examples

There are two standard examples of Bayesian analyses that serve as models. These examples should be in the student's bag of easy pieces. In both of these examples, the prior is conjugate.

Example 3.2 (Continuation of Example 3.1) Binomial with Beta Prior
A loss function was not used in deriving the posterior distribution, but to get the Bayes estimator, we must use a loss function.

Let us choose the loss to be squared-error. In this case we know the risk is minimized by choosing the estimate as $\delta(x) = \mathrm{E}(\Pi|x)$, where the expectation is taken w.r.t. the distribution with density $f_{\Pi|x}$.

We recognize the posterior conditional distribution as a beta$(x + \alpha, n - x + \beta)$, so we have the Bayes estimator

$$\frac{\alpha + X}{\alpha + \beta + n}.$$

We should study this estimator from various perspectives.

- First, we note that it is a weighted average of the mean of the prior and the standard UMVUE:

$$\left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\frac{\alpha}{\alpha+\beta} + \left(\frac{n}{\alpha+\beta+n}\right)\frac{X}{n}.$$

This is a useful insight, but we should not suppose that all Bayes estimators work that way.

- We see that the Bayes estimator is *unbiased* if and only if $\alpha = \beta = 0$ in the prior beta distribution. In this case, however, the integral of the prior density above does not converge. This is an improper prior, and the estimator is a generalized Bayes estimator, because it is not really a Bayes estimator.
- Because
$$\lim_{\alpha\to 0_+,\beta\to 0_+} \frac{\alpha + X}{\alpha + \beta + n} = \frac{X}{n},$$
and for $\alpha > 0$ and $\beta > 0$, the prior is proper, we see that the UMVUE is a *limit of Bayes estimators*.
- What about a Jeffreys's (noninformative) prior? The Jeffreys's prior in this case is proportional to $\sqrt{I(\pi)}$. Because the binomial is a member of the exponential family, we know $I(\pi) = 1/\mathrm{V}(T)$, where $\mathrm{E}(T) = \pi$. So $I(\pi) = n/\pi(1-\pi)$. Jeffreys's prior is therefore beta$(1/2, 1/2)$. The Bayes estimator corresponding to the noninformative prior is
$$\frac{X + \frac{1}{2}}{n+1}.$$
This is often used as an estimator of π in situations where $X = 1$ is rare.
- For the group invariant problem in which $g(X) = n - X$ and $\tilde{g}(\pi) = 1 - \pi$, we see that the Bayes estimator is *equivariant* if the prior is symmetric, that is, if $\alpha = \beta$.
- We can make an interesting *empirical Bayes* model from this example if we consider the observable random variable to be one of a set, X_k, each with conditional distribution binomial(π_k, n), where the π_k are all distributed independently as beta(α, β). An empirical Bayes procedure involves estimating α and β, and proceeding as before. Although any (reasonable) estimates of α and β would work, we generally use the MLEs. We get those by forming the conditional likelihood of x given α and β, and then maximizing to get $\hat{\alpha}$ and $\hat{\beta}$. (We do this numerically because it cannot be done in closed form. We get the conditional likelihood of x given α and β by first forming the joint of x and the π_k's, and integrating out the π_k's.) The Bayes estimator for π_k is
$$\frac{\hat{\alpha} + X_k}{\hat{\alpha} + \hat{\beta} + n}.$$
- If we put prior distributions on α and β, say gamma distributions with different parameters, we could form a hierarchical Bayes model and use iterative conditional simulated sampling to compute the estimates. (This type of approach is called Markov chain Monte Carlo, or specifically in this cases, Gibbs sampling. We discuss this approach in general in Section 3.3, and Gibbs sampling specifically beginning on page 146.) We would do this by working out the *full conditionals*.

- Could the Bayes estimator with this prior and squared-error loss function be minimax? Work out the risk,

$$\frac{1}{(\alpha+\beta+n)^2}\Big(n\pi(1-\pi)+\big(\alpha(1-\pi)-\beta\pi\big)^2\Big),$$

  and determine values of α and β such that it is constant. This will be minimax. The solution (to make it independent of π) is $\alpha=\beta=\sqrt{n}/2$. Notice what this does: it tends to push the estimator toward 1/2, which could have a maximum loss of 1/2, and we would expect that to be the minimum maximum.

The squared-error loss function is a very simple, and common loss function, of course. (In fact, the student must be very careful to remember that many simple properties of statistical methods depend on this special loss function.) What about other loss functions?

- Could we define a loss function so that the Bayes estimator is unbiased for a proper prior? Yes. Take

$$L(\pi,d)=\frac{(d-\pi)^2}{\pi(1-\pi)},$$

  and take a uniform(0,1) prior.
- For any loss function other than the squared-error, will the Bayes estimator be minimax? Yes, the loss function above yields this property. The Bayes estimator X/n has constant risk; therefore, it is minimax.

∎

The prior in this case is called a "conjugate" prior (when it exists; that is when $\alpha>0$ and $\beta>0$), because the posterior is in the same parametric family.

A conjugate prior and a squared-error loss function always yield Bayes estimators for $\mathrm{E}(X)$ that are linear in X, as we have seen in this specific case. Other priors may not be as easy to work with.

The statistics student should make the binomial/beta example above one of the "easy pieces" that can be pulled from memory. (This does not mean "rote memory" of the steps and equations; it means "process memory", which comes from understanding the process.)

Example 3.3 Normal with Inverted Chi-Squared and Conditional Normal Priors

For estimating both θ and σ^2 in $\mathrm{N}(\theta,\sigma^2)$, a conjugate prior family can be constructed by first defining a marginal prior on σ^2 and then a conditional prior on $\theta|\sigma^2$. From consideration of the case of known variance, we choose an inverted chi-squared distribution for the prior on σ^2:

$$\pi_{\sigma^2}(\sigma^2) \propto \frac{1}{\sigma}\sigma^{-(\nu_0/2+1)}\mathrm{e}^{(\nu_0\sigma_0^2)/(2\sigma^2)}$$

where we identify the parameters ν_0 and σ_0^2 as the degrees of freedom and the scale for σ^2. (People who work with simple Bayes procedures began calling the distribution of the reciprocal of a chi-squared random variable an "inverse" chi-squared distribution. Because "inverse" is used in the names of distributions in a different way ("inverse Gaussian"), I prefer the term inverted chi-squared, or, in general, inverted gamma.)

Given σ^2, let us choose a normal distribution for the conditional prior of $\theta|\sigma^2$:

$$\pi_{\theta|\sigma^2}(\theta;\sigma^2) \propto \exp\left(-\frac{1}{2}(\theta-\mu_0)^2/(\sigma^2/\kappa_0)\right),$$

where we identify the parameters μ_0 and σ^2/κ_0 as the location and the scale for θ.

Following the standard steps of forming the joint density of (X, θ, σ^2) and then the marginal of X, we get the joint posterior as

$$p_{\theta,\sigma^2|x}(\theta,\sigma^2;x) \propto \frac{1}{\sigma}\sigma^{-(\nu_0/2+1)}\exp\left(-\frac{1}{2\sigma^2}\left(\nu_0\sigma_0^2+\kappa_0(\theta-\mu_0)^2\right)\right).$$

For the estimators, we minimize the expected loss with respect to this probability distribution. ∎

Another way this problem is sometimes approached is to reparametrize the normal, and in place of σ^2, use $1/(2\tau)$.

3.3 Markov Chain Monte Carlo

Monte Carlo techniques often allow us to make statistical inferences when the statistical method involves intractable expressions.

Monte Carlo methods involve sampling, usually artificially, in the sense that the samples are generated on the computer.

The raw samples are from a $\mathrm{U}(0,1)$ distribution (or an approximate $\mathrm{U}(0,1)$ distribution, in the sense that the samples are generated on the computer).

A raw sample of uniforms, $U_1, U_2, \ldots$, is transformed into a sequence $\{X_j\}$ of (pseudo)random variables from a distribution of interest.

We often want the sequence $\{X_j\}$ to be i.i.d. As part of the transformation process, however, we may use a sequence $\{Y_i\}$ that has internal dependencies.

Inverse CDF

Assume that the CDF of the distribution of interest is F_X, and further, suppose that F_X is continuous and strictly monotone.

In that case, if X is a random variable with CDF F_X, then $U = F_X(X)$ has a U$(0, 1)$ distribution.

In the inverse CDF method, we transform each U_i to an X_i by

$$X_i = F_X^{-1}(U_i).$$

If F_X is not continuous or strictly monotone, we can modify this transformation slightly.

Acceptance/Rejection

To understand a slightly more complicated process that is often used, consider the problem of transforming an i.i.d. sequence $\{U_i\}$ of uniforms into an i.i.d. sequence $\{X_j\}$ from a distribution that has a probability density $p(\cdot)$.

We use an intermediate sequence $\{Y_k\}$ from a distribution that has a probability density $g(\cdot)$. (It could also be the uniform distribution.)

Further, suppose for some constant c that $h(x) = cg(x)$ is such that $h(x) \geq p(x)$.

1. Generate a variate y from the distribution having pdf g.
2. Generate independently a variate u from the uniform (0,1) distribution.
3. If $u \leq p(y)/h(y)$, then accept y as the variate, otherwise, reject y and return to step 1.

See Figure 3.6.

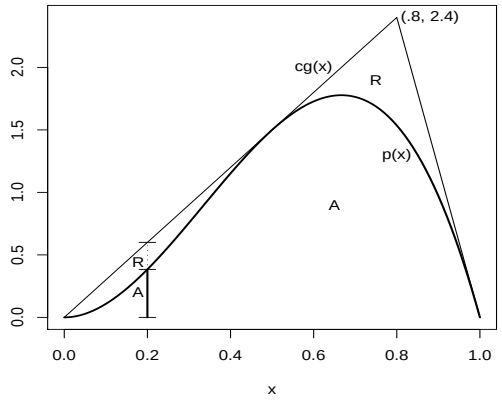


Fig. 3.6. Acceptance/Rejection

Why It Works

Let X be the random variable delivered. For any x, because Y (from the density g) and U are independent, we have

$$\begin{aligned}\Pr(X \le x) &= \Pr\left(Y \le x \,|\, U \le \frac{p(Y)}{cg(Y)}\right) \\ &= \frac{\int_{-\infty}^{x}\int_{0}^{p(t)/cg(t)} g(t)\,\mathrm{d}s\,\mathrm{d}t}{\int_{-\infty}^{\infty}\int_{0}^{p(t)/cg(t)} g(t)\,\mathrm{d}s\,\mathrm{d}t} \\ &= \int_{-\infty}^{x} p(t)\,\mathrm{d}t,\end{aligned}$$

the distribution function corresponding to p. Differentiating this quantity with respect to x yields $p(x)$.

The acceptance/rejection method can be visualized as choosing a subsequence from a sequence of independently and identically distributed (i.i.d.) realizations from the distribution with density g_Y in such a way the subsequence has density p_X.

i.i.d. from g_Y	y_i	y_{i+1}	y_{i+2}	y_{i+3}	$\cdots$	y_{i+k}	$\cdots$
accept?	no	yes	no	yes	$\cdots$	yes	$\cdots$
i.i.d. from p_X		x_j		x_{j+1}	$\cdots$	x_{j+l}	$\cdots$

Obviously, the closer $cg(x)$ is to $p(x)$, the faster the acceptance/rejection algorithm will be, if we ignore the time required to generate y from the dominating density g. A good majorizing function would be such that the l is almost as large as k.

Often, g is chosen to be a very simple density, such as a uniform or a triangular density. When the dominating density is uniform, the acceptance/rejection method is similar to the "hit-or-miss" method.

Variations of Acceptance/Rejection

There are many variations of the basic acceptance/rejection.

One is called *transformed rejection*. In the transformed acceptance/rejection method, the steps of the algorithm are combined and rearranged slightly.

There are various ways that acceptance/rejection can be used for discrete distributions.

It is clear from the description of the algorithm that the acceptance/rejection method also applies to multivariate distributions. (The uniform random number is still univariate, of course.)

Dependent Random Variables

The methods described above use a sequence of i.i.d. variates from the majorizing density. It is also possible to use a sequence from a conditional majorizing density.

A method using a nonindependent sequence is called a Metropolis method, and there are variations of these, with their own names.

There are two related cases:

Suppose $\{X_j : j = 0, 1, 2, \ldots\}$ is such that for $j = 1, 2, \ldots$ we know the conditional distributions of $X_j|X_0, \ldots, X_{j-1}$.

Alternatively, suppose we know the functional form (up to the normalizing constant) of the joint density of $X_1, X_2, \ldots, X_k$, and that we know the distribution of at least one $X_i|X_j(i \neq j)$.

Markov Chain Monte Carlo

If the density of interest, p, is the density of the stationary distribution of a Markov chain, correlated samples from the distribution can be generated by simulating the Markov chain.

This appears harder than it is.

A Markov chain is the basis for several schemes for generating random samples. The interest is not in the sequence of the Markov chain itself.

The elements of the chain are accepted or rejected in such a way as to form a different chain whose stationary distribution or limiting distribution is the distribution of interest.

Markov Chains

Markov chains are

An aperiodic, irreducible, positive recurrent Markov chain is associated with a *stationary distribution* or *invariant distribution*, which is the limiting distribution of the chain.

Convergence?

An algorithm based on a stationary distribution of a Markov chain is an *iterative method* because a sequence of operations must be performed until they *converge*; that is, until the chain has gone far enough to wash out any transitory phase that depends on where we start.

Several schemes for assessing convergence have been proposed. For example, we could use multiple starting points and then use an ANOVA-type test to compare variances within and across the multiple streams.

The Metropolis Algorithm

For a distribution with density p, the Metropolis algorithm, introduced by Metropolis et al. (1953) generates a random walk and performs an acceptance/rejection based on p evaluated at successive steps in the walk.

In the simplest version, the walk moves from the point y_i to a candidate point $y_{i+1} = y_i + s$, where s is a realization from $U(-a, a)$, and accepts y_{i+1} if

$$\frac{p(y_{i+1})}{p(y_i)} \geq u,$$

where u is an independent realization from $U(0, 1)$.

This method is also called the "heat bath" method because of the context in which it was introduced.

The random walk of Metropolis et al. is the basic algorithm of *simulated annealing*, which is currently widely used in optimization problems.

If the range of the distribution is finite, the random walk is not allowed to go outside of the range.

Example 3.4 Simulation of the von Mises Distribution with the Metropolis Algorithm
Consider, for example, the von Mises distribution, with density,

$$p(x) = \frac{1}{2\pi I_0(c)} e^{c\cos(x)}, \quad \text{for } -\pi \leq x \leq \pi,$$

where I_0 is the modified Bessel function of the first kind and of order zero.

The von Mises distribution is an easy one to simulate by the Metropolis algorithm. This distribution is often used by physicists in simulations of lattice gauge and spin models, and the Metropolis method is widely used in these simulations.

It is not necessary to know the normalizing constant, because it is canceled in the ratio. The fact that all we need is a nonnegative function that is proportional to the density of interest is an important property of this method.

If $c = 3$, after a quick inspection of the amount of fluctuation in p, we may choose $a = 1$. The R statements below implement the Metropolis algorithm to generate $n - 1$ deviates from the von Mises distribution.

Notice the simplicity of the algorithm in the R code. We did not need to determine a majorizing density, nor even evaluate the Bessel function that is the normalizing constant for the von Mises density.

```
n <- 1000
x <- rep(0,n)
a <-1
c <-3
yi <-3
j <-0
```

```
i <- 2
while (i < n) {
  i <- i + 1
  yip1 <- yi + 2*a*runif(1)- 1
  if (yip1 < pi & yip1 > - pi) {
  if (exp(c*(cos(yip1)-cos(yi))) > runif(1)) yi <- yip1
    else yi <- x[i-1]
  x[i] <- yip1
  }
}
```

The Metropolis method can be visualized as choosing a subsequence from a sequence of realizations from a random walk with density $g_{Y_{i+1}|Y_i}$ in such a way that the subsequence selected has density p_X.

random walk	y_i	$y_{i+1} = y_i + s_{i+1}$	$y_{i+3} = y_{i+1} + s_{i+2}$	$y_{i+2} = y_{i+2} + s_{i+3}$	$\cdots$
accept?	no	yes	no	yes	$\cdots$
i.i.d. from p_X		x_j		x_{j+1}	$\cdots$

A histogram is not affected by the sequence of the output in a large sample.

The Markov chain samplers generally require a "burn-in" period; that is, a number of iterations before the stationary distribution is achieved.

In practice, the variates generated during the burn-in period are discarded.

The number of iterations needed varies with the distribution, and can be quite large, sometimes several hundred.

The von Mises example is unusual; no burn-in is required. In general, convergence is much quicker for univariate distributions with finite ranges such as this one.

It is important to remember what convergence means; it does *not* mean that the sequence is independent from the point of convergence forward. The deviates are still from a Markov chain.

The Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm uses a more general chain for the acceptance/rejection step.

To generate deviates from a distribution with density p_X it uses deviates from a Markov chain with density $g_{Y_{t+1}|Y_t}$. The conditional density $g_{Y_{t+1}|Y_t}$ is chosen so that it is easy to generate deviates from it.

0. Set $k = 0$.
1. Choose $x^{(k)}$ in the range of p_X. (The choice can be arbitrary.)
2. Generate y from the density $g_{Y_{t+1}|Y_t}(y|x^{(k)})$.

3. Set r:
$$r = p_X(y)\frac{g_{Y_{t+1}|Y_t}(x^{(k)}|y)}{p_X(x^{(k)})g_{Y_{t+1}|Y_t}(y|x^{(k)})}$$
4. If $r \geq 1$, then
   4.a. set $x^{(k+1)} = y$;
   otherwise
   4.b. generate u from uniform(0,1) and
      if $u < r$, then
         4.b.i. set $x^{(k+1)} = y$,
      otherwise
         4.b.ii. set $x^{(k+1)} = x^{(k)}$.
5. If convergence has occurred, then
   5.a. deliver $x = x^{(k+1)}$;
   otherwise
   5.b. set $k = k + 1$, and go to step 2.

Compare the Metropolis-Hastings algorithm with the basic acceptance/rejection method.

The majorizing function in the Metropolis-Hastings algorithm is
$$\frac{g_{Y_{t+1}|Y_t}(x|y)}{p_X(x)\, g_{Y_{t+1}|Y_t}(y|x)}.$$

r is called the "Hastings ratio", and step 4 is called the "Metropolis rejection". The conditional density, $g_{Y_{t+1}|Y_t}(\cdot|\cdot)$ is called the "proposal density" or the "candidate generating density". Notice that because the majorizing function contains p_X as a factor, we only need to know p_X to within a constant of proportionality. As we have mentioned already, this is an important characteristic of the Metropolis algorithms.

As with the acceptance/rejection methods with independent sequences, the acceptance/rejection methods based on Markov chains apply immediately to multivariate random variables.

We can see why this algorithm works by using the same method as we used to analyze the acceptance/rejection method; that is, determine the CDF and differentiate.

The CDF is the probability-weighted sum of the two components corresponding to whether the chain moved or not. In the case in which the chain does move, that is, in the case of acceptance, for the random variable Z whose realization is y, we have
$$\begin{aligned}
\Pr(Z \leq x) &= \Pr\left(Y \leq x \,\middle|\, U \leq p(Y)\frac{g(x_i|Y)}{p(x_i)g(Y|x_i)}\right) \\
&= \frac{\int_{-\infty}^{x}\int_{0}^{p(t)g(x_i|t)/(p(x_i)g(t|x_i))} g(t|x_i)\,\mathrm{d}s\,\mathrm{d}t}{\int_{-\infty}^{\infty}\int_{0}^{p(t)g(x_i|t)/(p(x_i)g(t|x_i))} g(t|x_i)\,\mathrm{d}s\,\mathrm{d}t} \\
&= \int_{-\infty}^{x} p_X(t)\,\mathrm{d}t.
\end{aligned}$$

Gibbs Sampling

An iterative method, somewhat similar to the use of marginals and conditionals, can also be used to generate multivariate observations. It was first used for a a Gibbs distribution (Boltzmann distribution), and so is called the *Gibbs method*.

In the Gibbs method, after choosing a starting point, the components of the d-vector variate are generated one at a time conditionally on all others.

If p_X is the density of the d-variate random variable X, we use the conditional densities $p_{X_1|X_2,X_3,\cdots,X_d}$, $p_{X_2|X_1,X_3,\cdots,X_d}$, and so on.

At each stage the conditional distribution uses the most recent values of all the other components.

As with other MCMC methods, it may require a number of iterations before the choice of the initial starting point is washed out.

Gibbs sampling is often useful in higher dimensions. It depends on the convergence of a Markov chain to its stationary distribution, so a burn-in period is required.

0. Set $k = 0$.
1. Choose $x^{(k)} \in S$.
2. Generate $x_1^{(k+1)}$ conditionally on $x_2^{(k)}, x_3^{(k)}, \ldots, x_d^{(k)}$,
   Generate $x_2^{(k+1)}$ conditionally on $x_1^{(k+1)}, x_3^{(k)}, \ldots, x_d^{(k)}$,
   ...
   Generate $x_{d-1}^{(k+1)}$ conditionally on $x_1^{(k+1)}, x_2^{(k+1)}, \ldots, x_d^{(k)}$,
   Generate $x_d^{(k+1)}$ conditionally on $x_1^{(k+1)}, x_2^{(k+1)}, \ldots, x_{d-1}^{(k+1)}$.
3. If convergence has occurred, then
   3.a. deliver $x = x^{(k+1)}$;
   otherwise
   3.b. set $k = k + 1$, and go to step 2.

Example 3.5 Gibbs Sampling to Generate Independent Normals
Consider X_{t+1} normal with a mean of X_t and a variance of σ^2.

We will generate an i.i.d. sample from a standard normal distribution; that is, a normal with a mean of 0 and a variance of 1. In this example, the target distribution is simpler than the proposal.

We start with a x_0, chosen arbitrarily.

We take logs and cancel terms in the expression for r.

The following simple Matlab statements generate the sample.

```
x(1) = x0;
while i < n
  i = i + 1;
  yip1 = yi + sigma*randn;
  lr2 = yi^2 - yip1^2;
  if lr2 > 0
```

```
    yi = yip1;
  else
    u = rand;
    if lr2 > log(u)*2
      yi = yip1;
    else
      yi = x(i-1);
    end
  end
  x(i) = yi;
end
plot (x)
```

There are several variations of the basic Metropolis-Hastings algorithm. Two common related methods are Gibbs sampling and hit-and-run sampling. Those methods are particularly useful in multivariate simulation.

Markov chain Monte Carlo has become one of the most important tools in statistics in recent years. Its applications pervade Bayesian analysis, as well as many Monte Carlo procedures in the frequentist approach to statistical analysis.

Whenever a correlated sequence such as a Markov chain is used, variance estimation must be performed with some care. In the more common cases of positive autocorrelation, the ordinary variance estimators are negatively biased. The method of batch means or some other method that attempts to account for the autocorrelation should be used.

Convergence

Some of the most important issues in MCMC concern the rate of convergence, that is, the length of the burn-in, and the frequency with which the chain advances.

In many applications of simulation, such as studies of waiting times in queues, there is more interest in transient behavior than in stationary behavior.

This is usually not the case in use of MCMC methods. The stationary distribution is the only thing of interest.

The issue of convergence is more difficult to address in multivariate distributions. It is for multivariate distributions, however, that the MCMC method is most useful.

This is because the Metropolis-Hastings algorithm does not require knowledge of the normalizing constants, and the computation of a normalizing constant may be more difficult for multivariate distributions.

Various diagnostics have been proposed to assess convergence. Most of them use multiple chains in one way or another. Use of batch means from

separate streams can be used to determine when the variance has stabilized. A cusum plot on only one chain to help to identify convergence.

Various methods have been proposed to speed up the convergence.

Methods of assessing convergence is currently an area of active research.

The question of whether convergence has practically occurred in a finite number of iterations is similar in the Gibbs method to the same question in the Metropolis-Hastings method.

In either case, to determine that convergence has occurred is not a simple problem.

Once a realization is delivered in the Gibbs method, that is, once convergence has been deemed to have occurred, subsequent realizations can be generated either by starting a new iteration with $k = 0$ in step 0, or by continuing at step 1 with the current value of $x^{(k)}$.

If the chain is continued at the current value of $x^{(k)}$, we must remember that the subsequent realizations are not independent.

Effects of Dependence

This affects variance estimates (second order sample moments), but not means (first order moments).

In order to get variance estimates we may use means of batches of subsequences or use just every m^{th} (for some $m > 1$) deviate in step 3. (The idea is that this separation in the sequence will yield subsequences or a systematic subsample with correlations nearer 0.)

If we just want estimates of means, however, it is best not to subsample the sequence; that is, the variances of the estimates of means (first order sample moments) using the full sequence is smaller than the variances of the estimates of the same means using a systematic (or any other) subsample (so long as the Markov chain is stationary.)

To see this, let $\bar{x}_i$ be the mean of a systematic subsample of size n consisting of every m^{th} realization beginning with the i^{th} realization of the converged sequence. Now, we observe that

$$|\mathrm{Cov}(\bar{x}_i, \bar{x}_j)| \leq \mathrm{V}(\bar{x}_l)$$

for any positive i, j, and l less than or equal to m. Hence if $\bar{x}$ is the sample mean of a full sequence of length nm, then

$$\begin{aligned}\mathrm{V}(\bar{x}) &= \mathrm{V}(\bar{x}_l)/m + \sum_{i\neq j;i,j=1}^{m} \mathrm{Cov}(\bar{x}_i, \bar{x}_j)/m^2 \\ &\leq \mathrm{V}(\bar{x}_l)/m + m(m-1)\mathrm{V}(\bar{x}_l)/m \\ &= \mathrm{V}(\bar{x}_l).\end{aligned}$$

In the Gibbs method the components of the d-vector are changed systematically, one at a time. The method is sometimes called *alternating conditional sampling* to reflect this systematic traversal of the components of the vector.

Ordinary Monte Carlo and Iterative Monte Carlo

The general objective in Monte Carlo simulation is to calculate the expectation of some function g of a random variable X. In ordinary Monte Carlo simulation, the method relies on the fact that for independent, identically distributed realizations $X_1, X_2, \ldots$ from the distribution P of X,

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \to \mathrm{E}g(X)$$

almost surely as n goes to infinity. This convergence is a simple consequence of the law of large numbers.

In Monte Carlo simulation, the sample is simulated with a random number generator. When X is multivariate or a complicated stochastic process, however, it may be difficult or impossible to simulate independent realizations.

A Hierarchical Bayesian Model

Following custom, we use brackets to denote *densities*; $[X, Y], [X|Y]$, and $[X]$ represent the joint, conditional, and marginal densities, respectively.

In a hierarchical Bayesian model, the joint distribution of the data and parameters is

$$[X|\theta_1] \times [\theta_1|\theta_2] \times [\theta_2|\theta_3] \times \cdots \times \times [\theta_{k-1}|\theta_k] \times [\theta_k]$$

The thing of interest is $[\theta_1|X]$.
The hierarchical structure implies

$$\begin{aligned}[\theta_1|X, \theta_{i,(i\neq 1)}] &= [\theta_1|X, \theta_2] \\ &= [\theta_i|\theta_{i-1}, \theta_{i+1}] \\ &= [\theta_k|\theta_{k-1}]\end{aligned}$$

Gibbs sampling can be used to estimate the marginal posterior densities.

Example 3.6 Gibbs Sampling Example from Gelfand and Smith, *JASA*
The paper by Gelfand and Smith (1990) was very important in popularizing the Gibbs method.

Consider an exchangeable Poisson model in which independent counts are observed over differing periods of time.

The data are $\{(s_i, t_i)\}$. Each yields a rate r_i.

Assume $[s_i|\lambda_i] = \mathrm{P}(\lambda_i t_i)$.

Assume a gamma prior distribution on the λ_i's with density

$$\frac{1}{\beta^\alpha \Gamma(\alpha)} \lambda_i^{\alpha-1} e^{-\lambda_i/\beta}$$

Further, assume β has an inverse gamma distribution with density

$$\frac{1}{\beta^{\gamma+1}\Gamma(\gamma)}\delta^{\gamma}e^{-\delta/\beta}$$

Beginning with $X = (s_1, s_2, \ldots, s_k)$, the conditional distribution of λ_i given X, β, and $\lambda_{j(j\neq i)}$ is merely the gamma with parameters $\alpha + s_j$ and $\beta/(t_j + 1)$, and the conditional distribution of β given X and the λ_i's is an inverse gamma with parameters $\gamma + k\alpha$ and $\sum \lambda_i + \delta$.

The various parameters (α, δ, γ) have interpretions that can be used to select reasonable values.

The Gibbs sampling method would estimate the marginal density of λ_i by generating $\lambda_i^{(1)}$ from the appropriate gamma distribution, i.e., with parameters $\alpha + s_i$ and $\beta^{(0)}/(t_i + 1)$ for $i = 1, \ldots, k$, and then generating $\beta^{(1)}$ for the first iteration.

Continue this for k iterations.

Do it m times to have a density. ∎

Miscellaneous Results and Comments

Markov chain Monte Carlo has special applications when dealing with distributions that have densities known up to a constant of proportionality, that is densities specified as follows. Let h be a nonnegative integrable function that is not zero almost everywhere. Then h specifies a probability distribution, all we need to do to get the density f is normalize it.

$$f(x) = h(x)/c$$

where

$$c = \int h(x)d\mu(x)$$

The Hastings algorithm only uses h to simulate realizations from f, knowledge of the integral c is not required.

In Bayesian inference, h is the likelihood times the prior. This is always known, but the integral c is generally hard. MCMC permits easy simulations of realizations from the posterior (no knowledge of c necessary).

In most cases where there is complex dependence in the data, there is no simple probability model with c known, but it is easy to specify a model up to a constant of proportionality using an h. These are just very complicated exponential families.

Let t be a vector-valued statistic on the sample space and

$$h(x) = \exp(t(x)^{\mathrm{T}}\theta)$$

Then these specify a family of densities

$$f_\theta(x) = \exp(t(x)^{\mathrm{T}}\theta)/c(\theta).$$

In the expression

$$\exp(t(x)^{\mathrm{T}}\theta)/c(\theta),$$

$$c(\theta) = \int \exp(t(x)^{\mathrm{T}}\theta)d\mu(x),$$

but in MCMC it does not need to be known.

This is just an exponential family with canonical statistic $t(x)$ and canonical parameter θ.

Using Markov chain Monte Carlo we can simulate realizations from any distribution in the model, and using the simulations from any one distribution, we can calculate maximum likelihood estimates, bootstrap estimates of their sampling distribution and so forth.

There are also ways to get (randomized) significance tests with exact p-values using Markov chain Monte Carlo.

The output of the sampler is a Markov chain $X_1, X_2, \ldots$ whose equilibrium distribution is the distribution of interest, the one you want to sample from.

Averages with respect to that distribution are approximated by averages over the chain.

3.4 Bayesian Testing

In statistical hypothesis testing, the basic problem is to decide whether or not to reject a statement about the distribution of a random variable. The statement must be expressible in terms of membership in a well-defined class. The hypothesis can therefore be expressed by the statement that the distribution of the random variable X is in the class $\mathcal{P}_H = \{P_\theta : \theta \in \Theta_H\}$. An hypothesis of this form is called a statistical hypothesis.

The basic paradigm of statistical hypothesis testing was described in Section 2.4.1, beginning on page 103.

We usually formulate the testing problem as one of deciding between two statements:

$$H_0 \ : \ \theta \in \Theta_0$$

and

$$H_1 \ : \ \theta \in \Theta_1,$$

where $\Theta_0 \cap \Theta_1 = \emptyset$.

We do not treat H_0 and H_1 symmetrically; H_0 is the *hypothesis* to be tested and H_1 is the *alternative*. This distinction is important in developing a methodology of testing. We sometimes also refer to H_0 as the "null hypothesis" and to H_1 as the "alternative hypothesis".

In the Bayesian framework, we are interested in the probability that H_0 is true. The prior distribution provides an a priori probability, and the posterior distribution based on the data provides a posterior probability that H_0 is true. Clearly, we would choose to reject H_0 when the probability that it is true is small.

A First, Simple Example

Suppose we wish to test

$$H_0 : P = P_0$$

versus

$$H_1 : P = P_1,$$

and suppose that known probabilities p_0 and $p_1 = 1 - p_0$ can be assigned to H_0 and H_1 prior to the experiment. We see

- The overall probability of an error resulting from the use of the test δ is

$$p_0 \mathrm{E}_0(\delta(X)) + p_1 \mathrm{E}_1(1 - \delta(X)).$$

- The Bayes test that minimizes this probability is given by

$$\delta(x) = \begin{cases} 1 \text{ when } \hat{p}_1(x) > k\hat{p}_0(x) \\ \\ 0 \text{ when } \hat{p}_1(x) < k\hat{p}_0(x), \end{cases}$$

  for $k = p_0/p_1$.
- The conditional probability of H_i given $X = x$, that is, the posterior probability of H_i is

$$\frac{p_i \hat{p}_i(x)}{p_0 \hat{p}_0(x) + p_1 \hat{p}_1(x)}$$

  and the Bayes test therefore decides in favor of the hypothesis with the larger posterior probability.

Testing as an Estimation Problem

In the general setup above, we can define an indicator function $\mathrm{I}_{\Theta_0}(\theta)$. The testing problem, as we have described it, is the problem of estimating $\mathrm{I}_{\Theta_0}(\theta)$. Let us use a statistic $S(X)$ as an estimator of $\mathrm{I}_{\Theta_0}(\theta)$. The estimand is in $\{0, 1\}$, and so $S(X)$ should be in $\{0, 1\}$, or at least in $[0, 1]$.

Notice the relationship of $S(X)$ to $\delta(X)$. For the estimation approach using $S(X)$ to be equivalent to use of the test rule $\delta(X)$, it must be the case that

$$S(X) = 1 \Leftrightarrow \delta(X) = 0 \quad \text{(i.e., don't reject)}$$

and

$$S(X) = 0 \Leftrightarrow \delta(X) = 1 \quad \text{(i.e., reject)}$$

Following a decision-theoretic approach to the estimation problem, we define a loss function. In the classical framework of Neyman and Pearson, the loss function is 0-1. Under this loss, using $S(X) = s$ as the rule for the test we have

$$L(\theta, t) = \begin{cases} 0 \text{ if } s = \mathrm{I}_{\Theta_0}(\theta) \\ 1 \text{ otherwise.} \end{cases}$$

The Bayes estimator of $\mathrm{I}_{\Theta_0}(\theta)$ is the function that minimizes the posterior risk, $\mathrm{E}_{\Theta|x}(L(\Theta, s))$. The risk is just the posterior probability, so the Bayesian solution using this loss is

$$S(x) = \begin{cases} 1 \text{ if } \Pr(\theta \in \Theta_0|x) > \Pr(\theta \notin \Theta_0|x) \\ 0 \text{ otherwise,} \end{cases}$$

where $\Pr(\cdot)$ is evaluated with respect to the posterior distribution $P_{\Theta|x}$.

The 0-1-c Loss Function

In a Bayesian approach to hypothesis testing using the test $\delta(X) \in \{0, 1\}$, we often formulate a loss function of the form

$$L(\theta, d) = \begin{cases} c_d \text{ for } \theta \in \Theta_0 \\ b_d \text{ for } \theta \in \Theta_1 \end{cases}$$

where $c_1 > c_0$ and $b_0 > b_1$.

A common loss function has $c_0 = b_1 = 0$, $b_0 = 1$, and $c_1 = c > 0$.

This is called a 0-1-c loss function.

A Bayesian solution to hypothesis testing with a 0-1-c loss function is fairly easy to determine. The posterior risk for choosing $\delta(X) = 1$, that is, for rejecting the hypothesis, is

$$c\Pr(\Theta \in \Theta_{H_0}|X = x),$$

and the posterior risk for choosing $\delta(X) = 0$ is

$$\Pr(\Theta \in \Theta_{H_1}|X = x),$$

hence the optimal decision is to choose $\delta(X) = 1$ if

$$c\Pr(\Theta \in \Theta_{H_0}|X = x) < \Pr(\Theta \in \Theta_{H_1}|X = x),$$

which is the same as

$$\Pr(\Theta \in \Theta_{H_0}|X = x) < \frac{1}{1+c}.$$

In other words, the Bayesian approach says to reject the hypothesis if its posterior probability is small. The Bayesian approach has a simpler interpretation than the frequentist approach. It also makes more sense for other loss functions.

The Weighted 0-1 or a_0-a_1 Loss Function

Another approach to account for all possibilities and to penalize errors differently when the null hypothesis is true or false, is to define a weighted 0-1 loss function. Using the estimator $S(X) = s \in \{0, 1\}$, as above, we define

$$L(\theta, s) = \begin{cases} 0 & \text{if } s = \mathrm{I}_{\Theta_0}(\theta) \\ a_0 & \text{if } s = 0 \text{ and } \theta \in \Theta_0 \\ a_1 & \text{if } s = 1 \text{ and } \theta \notin \Theta_0. \end{cases}$$

This is sometimes called a a_0-a_1 loss. The 0-1-c loss and the a_0-a_1 loss could be defined either in terms of the test rule δ or the estimator S; I chose to do one one way and the other another way just for illustration.

The Bayes estimator of $\mathrm{I}_{\Theta_0}(\theta)$ using this loss is

$$S(x) = \begin{cases} 1 \text{ if } \Pr(\theta \in \Theta_0 | x) > \frac{a_1}{a_0 + a_1} \\ 0 \text{ otherwise,} \end{cases}$$

where again $\Pr(\cdot)$ is evaluated with respect to the posterior distribution. To see that this is the case, we write the posterior loss

$$\int_\Theta L(\theta, s) \mathrm{d}P_{\Theta|x} = a_0 \Pr(\theta \in \Theta_0 | x) \mathrm{I}_{\{0\}}(s) + a_1 \Pr(\theta \notin \Theta_0 | x) \mathrm{I}_{\{1\}}(s),$$

and then minimize it.

Under a a_0-a_1 loss, the null hypothesis H_0 is rejected whenever the posterior probability of H_0 is too small. The *acceptance level*, $a_1/(a_0 + a_1)$, is determined by the specific values chosen in the loss function. The Bayes test, which is the Bayes estimator of $\mathrm{I}_{\Theta_0}(\theta)$, depends only on a_0/a_1. The larger a_0/a_1 is the smaller the posterior probability of H_0 that allows for it to be accepted. This is consistent with the interpretation that the larger a_0/a_1 is the more important a wrong decision under H_0 is relative to H_1.

Examples

Let us consider two familiar easy pieces using a a_0-a_1 loss.

Example 3.7 Binomial with Uniform Prior
First, let $X|\pi \sim$ binomial(n, π) and assume a prior on Π of U$(0, 1)$ (a special case of the conjugate beta prior from Example 3.1). Suppose $\Theta_0 = [0, 1/2]$.

The posterior probability that H_0 is true is

$$\frac{(n+1)!}{x!(n-x)!} \int_0^{1/2} \pi^x (1-\pi)^{n-x} \mathrm{d}\pi.$$

This is computed and then compared to the acceptance level. (Note that the integral is a sum of fractions.)

∎

Example 3.8 Normal with Known Variance and Normal Prior on Mean
For another familiar example, consider $X|\theta \sim \mathrm{N}(\theta, \sigma^2)$, with σ^2 known, and $\theta \sim \mathrm{N}(\mu, \tau^2)$. We recall that $\Theta|x \sim \mathrm{N}(\mu(x), \omega^2)$, where

$$\mu(x) = \frac{\sigma^2\mu + \tau^2 x}{\sigma^2 + \tau^2} \quad \text{and} \quad \omega^2 = \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}.$$

To test H_0, we compute the posterior probability of H_0. Suppose the null hypothesis is $H_0 : \theta < 0$. Then

$$\begin{aligned}\Pr(H_0|x) &= \Pr\left(\frac{\theta - \mu(x)}{\omega} < \frac{-\mu(x)}{\omega}\right) \\ &= \Phi(-\mu(x)/\omega).\end{aligned}$$

The decision depends on the $a_1/(a_0 + a_1)$ quantile of $\mathrm{N}(0, 1)$. Let z_{a_0,a_1} be this quantile; that is, $\Phi(z_{a_0,a_1}) = a_1/(a_0 + a_1)$. The H_0 is accepted if

$$-\mu(x) > z_{a_0,a_1}\omega.$$

Rewriting this, we see that the null hypothesis is rejected x is greater than

$$-\frac{\sigma^2}{\tau^2}\mu - \left(1 + \frac{\sigma^2}{\tau^2}\right)\omega z_{a_0,a_1}.$$

∎

Notice a very interesting aspect of these tests. There is no predetermined acceptance level. The decision is based simply on the posterior probability that the null hypothesis is true.

A difficulty of the a_0-a_1 loss function, of course, is the choice of a_0 and a_1. Ideally, we would like to choose these based on some kind of utility considerations, but sometimes this takes considerable thought.

3.4.1 The Bayes Factor

Given a prior distribution P_Θ, let p_0 be the prior probability that H_0 is true, and p_1 be the prior probability that H_1 is true. The prior odds then is p_0/p_1. Similarly, let $\hat{p}_0$ be the posterior probability that H_0 is true given x, and $\hat{p}_1$ be the posterior probability that H_1 is true, yielding the posterior odds $\hat{p}_0/\hat{p}_1$.

The posterior probability of the event can be related to the relative odds. The posterior odds is

$$\frac{\hat{p}_0}{\hat{p}_1} = \frac{p_0}{p_1}\frac{p_{X|\theta}(x|\theta_0)}{\int p_{X|\theta}(x|\theta)\mathrm{d}P_\Theta}.$$

The term

$$\mathrm{BF}(x) = \frac{p_{X|\theta}(x|\theta_0)}{\int p_{X|\theta}(x|\theta)\mathrm{d}P_\Theta} \tag{3.2}$$

is called the *Bayes factor*. The Bayes factor obviously also depends on the prior $p_\Theta(\theta)$.

Rather than computing the posterior odds directly, we emphasize the Bayes factor, which for any stated prior odds yields the posterior odds. The Bayes factor is the posterior odds in favor of the hypothesis if $p_0 = 0.5$.

Note that, for the simple hypothesis versus a simple alternative, the Bayes factor simplifies to the likelihood ratio:

$$\frac{p_{X|\theta}(x|\theta_0)}{p_{X|\theta}(x|\theta_1)}.$$

One way of looking at this likelihood ratio is to use MLEs under the two hypotheses:

$$\frac{\sup_{\Theta_0} p_{X|\theta}(x|\theta)}{\sup_{\Theta_1} p_{X|\theta}(x|\theta)}.$$

This approach, however, assigns Dirac masses at the MLEs, $\hat{\theta}_0$ and $\hat{\theta}_1$.

The Bayes factor is more properly viewed as a Bayesian likelihood ratio,

$$\mathrm{BF}(x) = \frac{p_0 \int_{\Theta_0} p_{X|\theta}(x|\theta_0)\mathrm{d}\theta}{p_1 \int_{\Theta_1} p_{X|\theta}(x|\theta_1)\mathrm{d}\theta},$$

and, from a decision-theoretic point of view, it is entirely equivalent to the posterior probability of the null hypothesis. Under the a_0-a_1 loss function, H_0 is accepted when

$$\mathrm{BF}(x) > \frac{a_1}{a_0} / \frac{p_0}{p_1}$$

From this, we see that the Bayesian approach effectively gives an equal prior weight to the two hypotheses, $p_0 = p_1 = 1/2$ and then modifies the error penalties as $\tilde{a_i} = a_i p_i$, for $i = 0, 1$, or alternatively, incorporates the weighted error penalties directly into the prior probabilities:

$$\tilde{p_0} = \frac{a_0 p_0}{a_0 p_0 + a_1 p_1} \qquad \tilde{p_1} = \frac{a_1 p_1}{a_0 p_0 + a_1 p_1}.$$

The ratios such as likelihood ratios and relative odds that are used in testing carry the same information content if they are expressed as their reciprocals. These ratios can be thought of as evidence in favor of one hypothesis or model versus another hypothesis or model. The ratio provides a comparison of two alternatives, but there can be more than two alternatives under consideration. Instead of just H_0 and H_1 we may contemplate H_i and H_j, and follow the same steps using p_i/p_j. The Bayes factor then depends on i and j, and of course whether we use the odds ratio p_i/p_j or p_j/p_i. We therefore sometimes write the Bayes factor as $\mathrm{BF}_{ij}(x)$ where the subscript ij indicates use of the ratio p_i/p_j. In this notation, the Bayes factor (3.2) would be written as $\mathrm{BF}_{01}(x)$.

Jeffreys (1961) suggested a subjective "scale" to judge the evidence of the data in favor of or against H_0. Kass and Raftery (1995) discussed Jeffreys's scale and other issues relating to the Bayes factor. They modified his original scale (by combining two categories), and suggested

- if $0 < \log_{10}(\mathrm{BF}_{10}) < 0.5$, the evidence against H_0 is "poor",
- if $0.5 \leq \log_{10}(\mathrm{BF}_{10}) < 1$, the evidence against H_0 is "substantial",
- if $1 \leq \log_{10}(\mathrm{BF}_{10}) < 2$, the evidence against H_0 is "strong", and
- if $2 \leq \log_{10}(\mathrm{BF}_{10})$, the evidence against H_0 is "decisive".

Note that the Bayes factor is the reciprocal of the one we first defined in equation (3.2). While this scale makes some sense, the separations are of course arbitrary, and the approach is not based on a decision theory foundation. Given such a foundation, however, we still have the subjectivity inherent in the choice of a_0 and a_1, or in the choice of a significance level.

Kass and Raftery (1995) also gave an interesting example illustrating the Bayesian approach to testing of the "hot hand" hypothesis in basketball. They formulate the null hypothesis (that players do not have a "hot hand") as the distribution of good shots by a given player, Y_i, out of n_i shots taken in game i as binomial(n_i, π), for games $i = 1, \ldots, g$; that is, the probability for a given player, the probability of making a shot is constant in all games (within some reasonable period). A general alternative is $H_1 : Y_i \sim \text{binomial}(n_i, \pi_i)$. We choose a flat $\mathrm{U}(0,1)$ conjugate prior for the H_0 model. For the H_1 model, we choose a conjugate prior beta(α, β) with $\alpha = \xi/\omega$ and $\beta = (1-\xi)/\omega$. Under this prior, the prior expectation $\mathrm{E}(\pi_i|\xi, \omega)$ has an expected value of ξ, which is distributed as $\mathrm{U}(0,1)$ for fixed ω. The Bayes factor is is very complicated, involving integrals that cannot be solved in closed form. Kass and Raftery use this to motivate and to compare various methods of evaluating the integrals that occur in Bayesian analysis. One simple method is Monte Carlo.

Often, however, the Bayes factor can be evaluated relatively easily for a given prior, and then it can be used to investigate the sensitivity of the results to the choice of the prior, by computing it for another prior.

From Jeffreys's Bayesian viewpoint, the purpose of hypothesis testing is to evaluate the evidence in favor of a particular scientific theory. Kass and Raftery make the following points in the use of the Bayes factor in the hypothesis testing problem:

- Bayes factors offer a straightforward way of evaluating evidence in favor of a null hypothesis.
- Bayes factors provide a way of incorporating external information into the evaluation of evidence about a hypothesis.
- Bayes factors are very general and do not require alternative models to be nested.
- Several techniques are available for computing Bayes factors, including asymptotic approximations that are easy to compute using the output from standard packages that maximize likelihoods.

- In "nonstandard" statistical models that do not satisfy common regularity conditions, it can be technically simpler to calculate Bayes factors than to derive non-Bayesian significance tests.
- The Schwarz criterion (or BIC) gives a rough approximation to the logarithm of the Bayes factor, which is easy to use and does not require evaluation of prior distributions. The BIC is

$$\text{BIC} = -2\log(L(\theta_m|x)) + k\log n,$$

  where θ_m is the value of the parameters that specify a given model, k is the number of unknown or free elements in θ_m, and n is the sample size. The relationship is

$$\frac{-\text{BIC}/2 - \log(\text{BF})}{\log(\text{BF})} \to 0,$$

  as $n \to \infty$.
- When one is interested in estimation or prediction, Bayes factors may be converted to weights to be attached to various models so that a composite estimate or prediction may be obtained that takes account of structural or model uncertainty.
- Algorithms have been proposed that allow model uncertainty to be taken into account when the class of models initially considered is very large.
- Bayes factors are useful for guiding an evolutionary model-building process.
- It is important, and feasible, to assess the sensitivity of conclusions to the prior distributions used.

The Bayes Risk Set

A *risk set* can be useful in analyzing Bayesian procedures when the parameter space is finite. If

$$\Theta = \{\theta_1, \ldots, \theta_k\},$$

the risk set for a procedure T is a set in $\mathbb{R}^k$:

$$\{(z_1, ..., z_k) : z_i = R(\theta_i, T)\}.$$

In the case of 0-1 loss, the risk set is a subset of the unit hypercube; specifically, for $\Theta = \{0, 1\}$, it is a subset of the unit square: $[0, 1] \times [0, 1]$.

3.4.2 Bayesian Tests of a Simple Hypothesis

Although the test of a simple hypothesis versus a simple alternative, as in the first example in this section, is easy to understand and helps to direct our thinking about the testing problem, it is somewhat limited in application. In a more common application, we may have a dense parameter space Θ, and hypotheses that specify different subsets of Θ. A common situation is the "one-sided" test for $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$. We can usually develop

meaningful approaches to this problem, perhaps based on some boundary point of H_0. A "two-sided" test, in which, for example, the alternative specifies

$$\Theta_l = \{\theta : \theta < \theta_0\} \cup \Theta_u = \{\theta : \theta > \theta_0\},$$

presents more problems for the development of reasonable procedures.

In a Bayesian approach, when the parameter space Θ is dense, but either hypothesis is simple, there is a particularly troubling situation. This is because of the Bayesian interpretation of the problem as one in which a probability is to be associated with a statement about a specific value of a continuous random variable.

Consider the problem in a Bayesian approach to deal with an hypothesis of the form $H_0 : \Theta = \theta_0$, that is $\Theta_0 = \{\theta_0\}$; versus the alternative $H_1 : \Theta \neq \theta_0$.

A reasonable prior for Θ with a continuous support would assign a probability of 0 to $\Theta = \theta_0$.

One way of getting around this problem may be to modify the hypothesis slightly so that the null is a small interval around θ_0. This may make sense, but it is not clear how to proceed.

Another approach is, as above, to assign a positive probability, say p_0, to the event $\Theta = \theta_0$. Although it may not appear how to choose p_0, just as it would not be clear how to choose an interval around θ_0, we can at least proceed to simplify the problem following this approach. We can write the joint density of X and Θ as

$$p_{X,\Theta}(x,\theta) = \begin{cases} p_0 p_{X|\theta}(x|\theta_0) & \text{if } \theta = \theta_0, \\ (1-p_0) p_{X|\theta}(x|\theta) & \text{if } \theta \neq \theta_0. \end{cases}$$

There are a couple of ways of simplifying. Let us proceed by denoting the prior density of Θ over $\Theta \backslash \theta_0$ as λ. We can write the marginal of the data (the observable X) as

$$p_X(x) = p_0 p_{X|\theta}(x|\theta_0) + (1-p_0) \int p_{X|\theta}(x|\theta) \mathrm{d}\lambda(\theta).$$

We can then write the posterior density of Θ as

$$p_{\Theta|x}(\theta|x) = \begin{cases} p_1 & \text{if } \theta = \theta_0, \\ (1-p_1) \frac{p_{X|\theta}(x|\theta)}{p_X(x)} & \text{if } \theta \neq \theta_0, \end{cases}$$

where

$$p_1 = \frac{p_0 p_{X|\theta}(x|\theta_0)}{p_X(x)}.$$

This is the posterior probability of the event $\Theta = \theta_0$.

3.4.3 Interpretations of Probability Statements in Statistical Inference

In developing methods for hypothesis testing and for setting confidence regions, we can assume a model P_θ for the state of nature and develop procedures by consideration of probabilities of the form $\Pr(T(X) \circ C(\theta)|\theta)$, where $T(X)$ is a statistic, $C(\theta)$ is some region determined by the true (unknown) value of θ, and $\circ$ is some relationship. The forms of $T(X)$ and $C(\theta)$ vary depending on the statistical procedure. The procedure may be a test, in which case we may have $T(X) = 1$ or 0, according to whether the hypothesis is rejected or not, or it may by a procedure to define a confidence region, in which case $T(X)$ is a set. For example, if θ is given to be in Θ_H, and the procedure $T(X)$ is an α-level test of H, then $\Pr(T(X) = 1|\theta \in \Theta_H) \leq \alpha$. In a procedure to define a confidence set, we may be able to say $\Pr(T(X) \ni \theta) = 1 - \alpha$.

These kinds of probability statements are somewhat awkward, and a person without training in statistics may find them particularly difficult to interpret. Instead of a statement of the form $\Pr(T(X)|\theta)$, many people would prefer a statement of the form $\Pr(\Theta \in \Theta_H|X = x)$.

In order to make such a statement, however, we first must think of the parameter as a random variable and then we must formulate a conditional distribution for Θ, given $X = x$. The usual way is to use a model that has several components: a marginal (prior) probability distribution for the *unobservable random variable* Θ; a conditional probability distribution for the *observable random variable* X, given $\Theta = \theta$; and other assumptions about the distributions. We denote the prior density of Θ as p_Θ, and the conditional density of X as $p_{X|\theta}$. The procedure is to determine the conditional (posterior) distribution of Θ, given $X = x$.

If M is the model or hypothesis and D is the data, the difference is between $\Pr(D|M)$ (a "frequentist" interpretation), and $\Pr(M|D)$ (a "Bayesian" interpretation). People who support the latter interpretation will sometimes refer to the "prosecutor's fallacy" in which $\Pr(E|H)$ is confused with $\Pr(H|E)$, where E is some evidence and H is some hypothesis.

Some Bayesians do not think hypothesis testing is an appropriate statistical procedure. Because of the widespread role of hypothesis testing in science and in regulatory activities, however, statistical testing procedures must be made available.

3.4.4 Least Favorable Prior Distributions

In testing composite hypotheses, we often ask what is the "worst case" within the hypothesis. In a sense, this is the attempt to reduce the composite hypothesis to a simple hypothesis. This is the idea behind a p-value. In a Bayesian testing problem, this corresponds to a bound on the posterior probability.

Again, consider the problem of testing $H_0 : \Theta = \theta_0$ versus the alternative $H_1 : \Theta \neq \theta_0$.

3.4.5 Lindley's Paradox

Consider a null hypothesis H_0, the result of an experiment x, and a prior distribution that favors H_0 weakly. Lindley's paradox occurs when the result x is significant by a frequentist test, indicating sufficient evidence to reject H_0 at a given level, but the posterior probability of H_0 given x is high, indicating strong evidence that H_0 is in fact true.

This can happen at the same time when the prior distribution is the sum of a sharp peak at H_0 with probability p and a broad distribution with the rest of the probability $1 - p$. It is a result of the prior having a sharp feature at H_0 and no sharp features anywhere else.

Consider the problem of testing **********

3.5 Bayesian Confidence Sets

Statements about Probabilities

In developing methods for setting confidence regions, we have assumed a model P_θ for the state of nature and have developed procedures by consideration of probabilities of the form $\Pr(T(X) \circ S(\theta)|\theta)$, where $T(X)$ is a statistic, $S(\theta)$ is some region determined by the true (unknown) value of θ, and $\circ$ is some relationship. The forms of $T(X)$ and $S(\theta)$ vary depending on the statistical procedure. The procedure may by a procedure to define a confidence region, in which case $T(X)$ is a set. For example, in a procedure to define a confidence set, we may be able to say $\Pr(T(X) \ni \theta) = 1 - \alpha$.

These kinds of probability statements are somewhat awkward, and a person without training in statistics may find them particularly difficult to interpret. Instead of a statement of the form $\Pr(T(X)|\theta)$, many people would prefer a statement of the form $\Pr(\Theta \in \Theta_H|X = x)$.

The standard terminology for a Bayesian analogue of a confidence set is *credible set*.

If M is the model or hypothesis and D is the data, the difference is between

$$\Pr(D|M)$$

(a "frequentist" interpretation), and

$$\Pr(M|D)$$

(a "Bayesian" interpretation).

In order to make such a statement, however, we first must think of the parameter as a random variable and then we must formulate a conditional distribution for Θ, given $X = x$.

3.5.1 Credible Regions

In a Bayesian setup, we define a random variable Θ that corresponds to the parameter of interest, and the usual steps in a Bayesian analysis allows us to compute $\Pr(\Theta \in \Theta_{H_0}|X = x)$. The problem in setting a confidence region is an inverse problem; that is, for a given α, we determine C_α such that $\Pr(\Theta \in C_\alpha|X = x) = 1 - \alpha$. Of course there may be many sets with this property. We need some additional condition(s).

In the frequentist approach, we add the property that the region be the smallest possible. "Smallest" means with respect to some simple measure such as the usual Lebesgue measure; in the one-dimensional continuous case, we seek the shortest interval. In the Bayesian approach, we so something similar, except we use the posterior density as a measure.

The mechanics of determining credible regions begin with the standard Bayesian steps that yield the conditional distribution of the parameter given the observable random variable. If the density exists, we denote it as $p_{\Theta|x}$. At this point, we seek regions of θ in which $p_{\Theta|x}(\theta|x)$ is large. In general, the problem may be somewhat complicated, but in many situations of interest it is relatively straightforward. Just as in the frequentist approach, the identification of the region often depends on *pivotal* values, or pivotal functions. (Recall that a function $f(T, \theta)$ is said to be a pivotal function if its distribution does not depend on any unknown parameters.)

It is often straightforward to determine one with posterior probability content of $1 - \alpha$.

3.5.2 Highest Posterior Density Credible Regions

If the posterior density is $p_{\Theta|x}(\theta|x)$, we determine a number c such that the set

$$C_\alpha(x) = \{\theta \, : \, p_{\Theta|x}(\theta|x) \geq c_\alpha\} \tag{3.3}$$

is such that $\Pr(\Theta \in C_\alpha|X = x) = 1 - \alpha$. Such a region is called a level $1 - \alpha$ *highest posterior density* or HPD credible set.

We may impose other conditions. For example, in a one-dimensional continuous parameter problem, we may require that one endpoint of the interval be infinite (that is, we may seek a one-sided confidence interval).

An HPD region can be disjoint if the posterior is multimodal.

If the posterior is symmetric, all HPD regions will be symmetric about x.

For a simple example, consider a $\mathrm{N}(0, 1)$ prior distribution on Θ and a $\mathrm{N}(\theta, 1)$ distribution on the observable. The posterior given $X = x$ is $\mathrm{N}(x, 1)$. All HPD regions will be symmetric about x. In the case of a symmetric density, the HPD is the same as the centered equal-tail credible region; that is, the one with equal probabilities outside of the credible region. In that case, it is straightforward to determine one with posterior probability content of $1 - \alpha$.

3.5.3 Decision-Theoretic Approach

We can also use a specified loss function to approach the problem of setting a confidence region.

We choose a region so as to minimize the expected posterior loss.

For example, to form a two-sided interval in a one-dimensional continuous parameter problem, a reasonable loss function may be

$$L(\theta, [c_1, c_2]) = \begin{cases} k_1(c_1 - \theta) \text{ if } \theta < c_1, \\ 0 \qquad\qquad \text{if } c_1 \leq \theta \leq c_2, \\ k_2(\theta - c_2) \text{ if } \theta > c_2. \end{cases}$$

This loss function also leads to the interval between two quantiles of the posterior distribution.

It may not be HPD, and it may not be symmetric about some pivot quantity even if the posterior is symmetric.

3.5.4 Other Optimality Considerations

We may impose other conditions. For example, in a one-dimensional continuous parameter problem, we may require that one endpoint of the interval be infinite (that is, we may seek a one-sided confidence interval).

Alternatively, we may use the more fundamental concept of a loss function, and determine a credible set to minimize the expected loss.

Example 3.9 Credible Regions for the Binomial Parameter with a Beta Prior
Consider the problem of estimating π in a binomial(π, n) distribution with a beta(α, β) prior distribution, as in Example 3.2 on page 136.

Suppose we choose the hyperparameters in the beta prior as $\alpha = 3$ and $\beta = 5$. The prior, that is, the marginal distribution of Π, is as shown in Figure 3.1 and if n is 10 and we take one observation, $x = 2$ we have the conditional distribution of Π, as a beta with parameters $x + \alpha = 5$ and $n - x + \beta = 13$, as shown in Figure 3.2.

Now, given $x = 2$, and the original beta(3,5) prior, let's find an equal-tail 95% credible region. Here's some R code:

```
a<-3
b<-5
n<-10
x<-2
alpha<-0.05
lower<-qbeta(alpha/2,x+a,n-x+b)
upper<-qbeta(1-alpha/2,x+a,n-x+b)
pi<-seq(0,1,0.01)
plot(pi,dbeta(pi,x+a,n-x+b),type='l',
```

```
  main="95\% Credible Region with x=2",
  ylab="Posterior",xlab=expression(pi))
lines(c(lower,lower),c(0,dbeta(lower,x+a,n-x+b)))
lines(c(upper,upper),c(0,dbeta(upper,x+a,n-x+b)))
lines(c(0,1),c(0,0))
```

We get the credible region shown in Figure 3.7. The probability in each tail is 0.025.

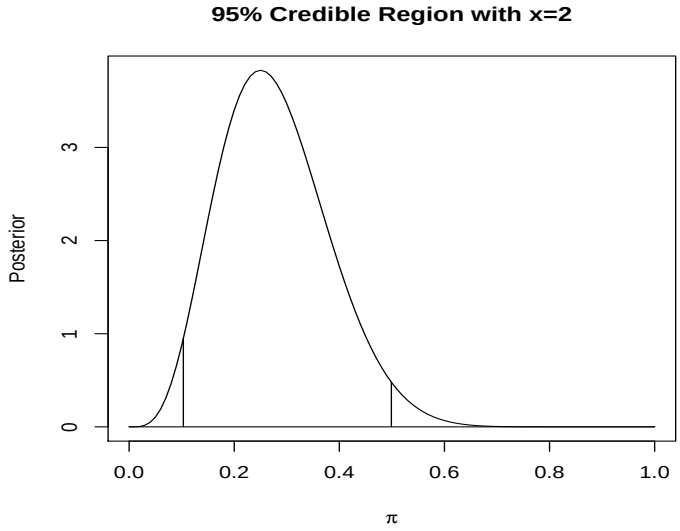


Fig. 3.7. 95% Credible Region after Observing $x = 2$

Because the posterior density is not symmetric, it is not an easy matter to get the HPD credible region.

The first question is whether the credible region is an interval. This depends on whether the posterior is unimodal. As we have already seen in Section 3.1, the posterior in this case is unimodal if $n > 0$, and so the credible region is indeed an interval.

We can determine the region iteratively by starting with the equal-tail credible region. At each step in the iteration we have a candidate lower bound and upper bound. We determine which one has the higher value of the density, and then shift the interval in that direction. We continue this process, keeping the total probability constant at each step. Doing this we get the credible region shown in Figure 3.8. The probability in the lower tail is 0.014 and that in the upper tail is 0.036. The density is 0.642 at each endpoint; that is, in equation (3.3), $c_\alpha = 0.642$.

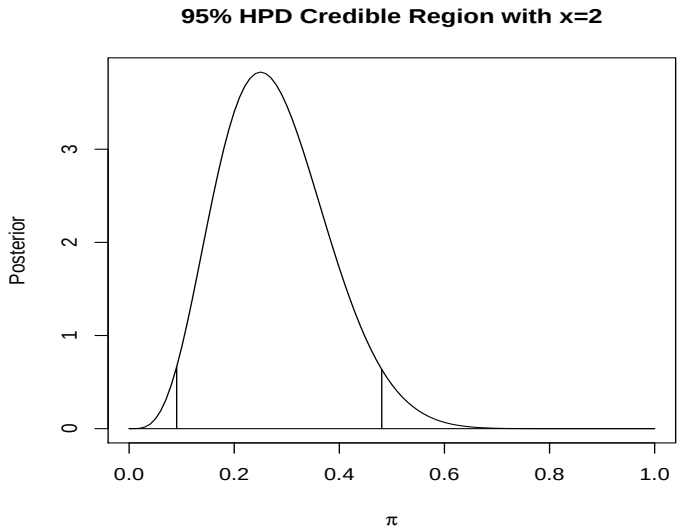


Fig. 3.8. HPD 95% Credible Region after Observing $x = 2$

Here's the R code that yielded the HPD:

```
a<-3
b<-5
n<-10
x<-2
alpha<-0.05
#     start by determining the equal-tail CR, using the posterior
lower<-qbeta(alpha/2,x+a,n-x+b)
upper<-qbeta(1-alpha/2,x+a,n-x+b)
#     set a tolerance for convergence
tol <- 0.005 # to get the density values to agree to 3 decimal places
a10 <- 0
a20 <- 0
a1 <- alpha/2
a2 <- 1-alpha/2
adj <- a1
d <- 1
while (abs(d)>tol){
#      determine difference in the density at the two candidate points
d <- dbeta(lower,x+a,n-x+b)-dbeta(upper,x+a,n-x+b)
#      halve the adjustment in each iteration
adj <- adj/2
```

```
#      if density at lower boundary is higher, shift interval to the left
s <- 1
if(d>0) s <- -1
a1 <- a1 + s*adj
a2 <- a2 + s*adj
lower<-qbeta(a1,x+a,n-x+b)
upper<-qbeta(a2,x+a,n-x+b)
}
```

∎

Notes

Berger (1985) and Robert (2001) provide extensive coverage of statistical inference from a Bayesian perspective. Both of these books compare the "frequentist" and Bayesian approaches and argue that the Bayesian paradigm is more solidly grounded. Many of the ideas in the Bayesian approach derive from Jeffreys (1961) book on probability, which emphasized a subjective view.

Ghosh and Sen (1991) have considered Pitman closeness in the context of a posterior distribution, and defined *posterior Pitman closeness* in terms of probabilities evaluated with respect to the posterior distribution. Interestingly, the posterior Pitman closeness is transitive, while as we have seen on page 72, Pitman closeness does not have the transitive property.

Bayesian methods for sampling from finite populations are discussed in Ghosh and Meeden (1998).

***** stuff to add:
improper priors
pseudo-Bayes factors
training sample
arithmetic intrinsic Bayes factor
geometric intrinsic Bayes factor
median intrinsic Bayes factor

Exercises in Shao

- For practice and discussion
  4.2(a)(b), 4.13, 4.14, 4.15, 4.19(b), 4.27, 4.30
  (Solutions in Shao, 2005)
- To turn in
  4.1(a)(b), 4.17, 4.18, 4.31, 4.32(a), 4.38(a)(b), 6.106, 6.107, 7.28, 7.29, 7.40

Additional References

Berger, James O. (1985), *Statistical Decision Theory and Bayesian Analysis*, second edition, Springer, New York.

Ghosh, Malay, and Glen Meeden (1998), *Bayesian Methods for Finite Population Sampling*, Chapman Hall/CRC, Boca Raton.

Ghosh, Malay, and Pranab Kumar Sen (1991), Bayesian Pitman closeness, *Communications in Statistics — Theory and Methods* **20**,3423–3437.

Jeffreys, Harold (1961), *Theory of Probability*, third edition, Oxford University Press, Oxford.

Robert, Christian P. (2001), *The Bayesian Choice*, second edition, Springer, New York.

4

Unbiased Point Estimation (Shao Ch 3, Sec 4.5; TPE2 Ch 2)

In a decision-theoretic approach to statistical inference, we seek a method that minimizes the risk no matter what is the true state of nature. In a problem of point estimation, for example, we seek an estimator $T(X)$ which for a given loss function $L(g(\theta), T(X))$ yields a minimum of $\mathrm{E}_\theta(L(g(\theta), T(X)))$.

For some specific value of θ, say θ_1, one particular estimator, say T_1, may have the smallest expected loss, while for another value of θ, say θ_2, another estimator, say T_2, may a smaller expected loss.

What we would like is an estimator with least expected loss no matter what is the value of θ; that is, we would like an estimator with uniformly minimum risk. Because the risk depends on the value of θ, however, we see that we cannot devise such an estimator. The optimal estimator would somehow involve θ. We would prefer a procedure that does not depend on the unknown quantity we are trying to estimate.

4.1 Uniformly Minimum Variance Unbiased Estimation

A property of a statistic that relates to a parameter, but does not depend on the value of the parameter, is unbiasedness. This leads us to require that the estimator of a given estimand, $g(\theta)$, be unbiased with respect to θ; that is,

$$\mathrm{E}_\theta(T(X)) = g(\theta) \; \forall \, \theta \in \Theta.$$

The requirement of unbiasedness cannot always be met. An estimand for which there is an unbiased estimator is said to be *U-estimable*. Remember unbiasedness refers to the entire parameter space. Consider, for example, the problem of estimating $1/\pi$ in binomial(n, π) for $\pi \in (0, 1)$. Suppose $T(X)$ is an unbiased estimator of $1/\pi$. Then

$$\sum_{x=0}^{n} T(x) \binom{n}{x} \pi^x (1-\pi)^{n-x} = 1/\pi.$$

Now as $\pi \to 0$, the left side tends to $T(0)$, which is finite, but the right side tends to ∞; hence, $1/\pi$ is not U-estimable. If, $1/\pi$ were U-estimable, the equation above would say that some polynomial in π is equal to $1/\pi$ for all $\pi \in (0, 1)$, and that clearly cannot be.

Another related example, but one that corresponds to a common parameter, is an estimator of the odds, $\pi/(1-\pi)$. We see that no unbiased estimator exists, for the same reason as in the previous example.

4.1.1 Unbiasedness and Squared-Error Loss

A squared-error loss function is particularly nice for an unbiased estimator, because in that case the expected loss is just the variance; that is, an unbiased estimator with minimum risk is an unbiased estimator with minimum variance.

If the unbiased estimator has minimum variance among all unbiased estimators within the parameter space, we say that such an estimator is a **uniformly** (for all values of θ) minimum variance unbiased estimator, that is, a UMVUE. (An unbiased estimator that has minimum variance among all unbiased estimators within a subspace of the parameter space is called a locally minimum variance unbiased estimator, or LMVUE.)

UMVU is a special case of uniform minimum risk (UMRU), which generally only applies to convex loss functions. In general, no UMRUE exists for bounded loss functions. Such loss functions cannot be (strictly) convex.

Uniformity (the first "U") means the MVU property is independent of the estimand. "Unbiasedness" is itself a uniform property, because it is defined in terms of an expectation for any distribution in the given family.

UMVU is closely related to complete sufficiency, which means that it is related to exponential families.

How to find an UMVUE

We generally find an UMVUE by beginning with a "good" estimator and manipulating it to make it UMVUE. It might be unbiased to begin with, and we reduce its variance while keeping it unbiased. It might not be unbiased to begin with but it might have some other desirable property, and we manipulate it to be unbiased.

One of the most useful facts is the Lehmann-Scheffé theorem, which says that if there is a complete sufficient statistic T for θ, and if $g(\theta)$ is U-estimable, then there is a unique UMVUE of $g(\theta)$ of the form $h(T)$, where h is a Borel function. (Notice that this follows from the Rao-Blackwell theorem. The uniqueness comes from the completeness, and of course, means unique a.e.) This fact leads to two methods:

- Find UMVU directly by finding $h(T)$ such that $\mathrm{E}_\theta(h(T)) = g(\theta)$.
  Example (Lehmann):

Given random sample of size n from Bernoulli(π). Want to estimate $g(\pi) = \pi(1-\pi)$. $T = \sum X_i$ is complete sufficient. The unbiasedness condition is

$$\sum_{t=0}^{n} \binom{n}{x} h(t)\pi^t(1-\pi)^{n-t} = \pi(1-\pi).$$

Rewriting this in terms of the odds $\rho = \pi/(1-\pi)$, we have, for all $\rho \in (0,\infty)$,

$$\begin{aligned}\sum_{t=0}^{n} \binom{n}{x} h(t)\rho^t &= \rho(1+\rho)^{n-2} \\ &= \sum_{t=1}^{n-1} \binom{n-2}{x-1}\rho^t.\end{aligned}$$

Now since for each t, the coefficient of ρ^t must be the same on both sides of the equation, we have

$$h(t) = \frac{t(n-t)}{n(n-1)}.$$

Also see examples 3.1 and 3.2 in Shao.

- If T_0 is unbiased, find UMVU as $h(T) = \mathrm{E}_\theta(T_0(X)|T)$. (This process is sometimes called "Rao-Blackwellization".)
  See example 3.3 in Shao.

An important property of unbiased estimators is the following.
If $T_0(X)$ is an unbiased estimator of $g(\theta)$, **all unbiased estimators** of $g(\theta)$ belong to an equivalence class defined as $\{T_0(X) - U(X)\}$, where $\mathrm{E}_\theta(U(X)) = 0$.

Unbiased estimators of 0 play a useful role in UMVUE problems.

We also see that useful estimators must have finite second moment, otherwise, we cannot minimize a variance by combining the estimators.

This leads to two methods if we have U such that $\mathrm{E}_\theta(U) = 0$ and $\mathrm{E}_\theta(U^2) < \infty$.

- Find UMVU by finding U to minimize $\mathrm{E}((T_0 - U)^2)$.
- If T is unbiased and has finite second moment, it is UMVU iff $\mathrm{E}(TU) = 0$ $\forall\theta \in \Theta$ and $\forall U \ni \mathrm{E}(U) = 0$ and $\mathrm{E}(U^2) < \infty$. (This is Theorem 3.2(i) in Shao.)

Theorem 3.2(ii) in Shao is similar to Theorem 3.2(i), but applies to functions of a sufficient statistic, $\widetilde{T}$.

Regularity Conditions

"Regularity conditions" apply to a *family* of distributions, $\mathcal{P} = \{P_\theta; \theta \in \Theta\}$, that have densities p_θ. There are generally three conditions that together are called the regularity conditions:

- The parameter space Θ is an open interval (in one dimension, and a cross product of open intervals in multidimensions).
- The support is independent of θ; that is, all P_θ have a common support.
- For any x in the support and $\theta \in \Theta$, $\partial p_\theta(x)/\partial\theta$ exists and is finite.

The latter two conditions ensure that the operations of integration and differentiation can be interchanged.

These three conditions play a major role in UMVUE. Fisher information is so important in minimum variance considerations, that these are sometimes called the Fisher information or FI regularity conditions.

4.1.2 Fisher Information

A fundamental question is how much information does a realization of the random variable X contain about the scalar parameter θ.

For a random variable X with PDF $p(x;\theta)$, we define the "information" (or "Fisher information") that X contains about θ as

$$I(\theta) = \mathrm{E}_\theta\left(\left(\frac{\partial \log p(X;\theta)}{\partial\theta}\right)\left(\frac{\partial \log p(X;\theta)}{\partial\theta}\right)^{\mathrm{T}}\right).$$

(Why does this definition make sense?) This is called Fisher information. (Another type of information is Shannon information, which for an event is the negative of the log of the probability of the event.)

Our information comes through the estimator $T(X)$. We are interested in the maximum information we can get.

Information is larger when there is larger relative variation in the density as the parameter changes, but the information available from an estimator is less when the estimator exhibits large variation (i.e., has large variance), so we want smaller variance.

There are several simple facts to know about $\log p(X;\theta)$:

$$\begin{aligned}\mathrm{E}\left(\frac{\partial \log(p(X,\theta))}{\partial\theta}\right) &= \int \frac{1}{p(x,\theta)}\frac{\partial p(x,\theta)}{\partial\theta}p(x,\theta)\mathrm{d}x\\ &= \frac{\partial}{\partial\theta}\int p(x,\theta)\mathrm{d}x\\ &= 0;\end{aligned}$$

therefore, the expectation in the information definition is also the variance:

$$\mathrm{E}\left(\left(\frac{\partial \log p(X;\theta)}{\partial\theta}\right)\left(\frac{\partial \log p(X;\theta)}{\partial\theta}\right)^{\mathrm{T}}\right) = \mathrm{V}\left(\frac{\partial \log(p(X,\theta))}{\partial\theta}\right).$$

We also have a relationship with the second derivative:

$$\mathrm{E}\left(\left(\frac{\partial \log p(X;\theta)}{\partial \theta}\right)\left(\frac{\partial \log p(X;\theta)}{\partial \theta}\right)^{\mathrm{T}}\right) = -\mathrm{E}\left(\frac{\partial^2 \log(p(X,\theta))}{\partial \theta^2}\right).$$

Consider the $\mathrm{N}(\mu, \sigma^2)$ distribution with $\theta = (\mu, \sigma)$ (which is simpler than for $\theta = (\mu, \sigma^2)$):

$$\log p_{(\mu,\sigma)}(x) = c - \log(\sigma) - (x-\mu)^2/(2\sigma^2).$$

We have

$$\frac{\partial}{\partial \mu} \log p_{(\mu,\sigma)}(x) = \frac{x-\mu}{\sigma^2}$$

and

$$\frac{\partial}{\partial \sigma} \log p_{(\mu,\sigma)}(x) = -\frac{1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3},$$

so

$$\begin{aligned} I(\theta) &= \mathrm{E}_\theta\left(\left(\frac{\partial \log p(X;\theta)}{\partial \theta}\right)\left(\frac{\partial \log p(X;\theta)}{\partial \theta}\right)^{\mathrm{T}}\right) \\ &= \mathrm{E}_{(\mu,\sigma^2)}\left(\begin{bmatrix} \frac{(X-\mu)^2}{(\sigma^2)^2} & \frac{X-\mu}{\sigma^2}\left(-\frac{1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3}\right) \\ \frac{x-\mu}{\sigma^2}\left(-\frac{1}{\sigma} + \frac{(X-\mu)^2}{\sigma^3}\right) & +\left(-\frac{1}{\sigma} + \frac{(X-\mu)^2}{\sigma^3}\right)^2 \end{bmatrix}\right) \\ &= \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}. \end{aligned}$$

Notice that the Fisher information matrix is dependent on the parametrization. The parametrization of the normal distribution in either the canonical exponential form or even $\theta = (\mu, \sigma^2)$ would result in a different Fisher information matrix.

This parametrization of the normal is rather unusual among common multiparameter distributions in that the information matrix is diagonal.

Consider the general canonical exponential form for a distribution in the exponential class:

$$p_\theta(x) = \exp\left((\eta^{\mathrm{T}} T(x) - \zeta(\eta)\right) h(x)$$

(See page 62).) If $\mu(\theta)$ is the mean-value parameter (see equation (1.100)), then

$$I(\theta) = V^{-1},$$

where

$$V = \mathrm{V}(T(X)).$$

********** prove this

The Fisher information for the two parameters $\theta = (\mu, \sigma)$ in a location-scale family with Lebesgue PDF

$$\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$$

has a particularly simple form:

$$I(\theta) = \frac{n}{\sigma^2}\begin{bmatrix} \int \left(\frac{f'(x)}{f(x)}\right)^2 f(x)\mathrm{d}x & \int x\left(\frac{f'(x)}{f(x)}\right)^2 f(x)\mathrm{d}x \\ \int x\left(\frac{f'(x)}{f(x)}\right)^2 f(x)\mathrm{d}x) & \int \left(\frac{xf'(x)}{f(x)}+1\right)^2 f(x)\mathrm{d}x \end{bmatrix}.$$

The prime on $f'(x)$ indicates differentiation with respect to x of course. (The information matrix is defined in terms of differentiation with respect to the parameters followed by an expectation.)

Another expression for the information matrix for a location-scale family is

$$I(\theta) = \frac{n}{\sigma^2}\begin{bmatrix} \int \frac{\left(f'(x)\right)^2}{f(x)}\mathrm{d}x & \int \frac{f'(x)\left(xf'(x)+f(x)\right)}{f(x)}\mathrm{d}x \\ \int \frac{f'(x)\left(xf'(x)+f(x)\right)}{f(x)}\mathrm{d}x & \int \frac{\left(xf'(x)+f(x)\right)^2}{f(x)}\mathrm{d}x \end{bmatrix}.$$

This is given in a slightly different form in Example 3.9 of Shao, which is Exercise 3.34, which is solved in his *Solutions*, using the form above, which is a more straightforward expression from the derivation that begins by defining the function $g(\mu, \sigma, x) = \log(f((x-\mu)/\sigma)/\sigma)$, and the proceeding with the definition of the information matrix.

Also we can see that in the location-scale family, if it is symmetric about the origin (that is about μ), the covariance term is 0.

Consider the gamma(α, β) distribution. We have for $x > 0$

$$\log p_{(\alpha,\beta)}(x) = -\log(\Gamma(\alpha)) - \alpha\log(\beta) + (\alpha-1)\log(x) - x/\beta.$$

This yields the Fisher information matrix

$$I(\theta) = \begin{bmatrix} \psi'(\alpha) & \frac{1}{\beta} \\ \frac{1}{\beta} & \frac{\alpha^2}{\beta^2} \end{bmatrix},$$

where $\psi(\alpha)$ is the digamma function, $\mathrm{d}\log(\Gamma(\alpha))/\mathrm{d}\alpha$, and $\psi'(\alpha)$ is the trigamma function, $\mathrm{d}\psi(\alpha)/\mathrm{d}\alpha$.

In the natural parameters, $\alpha - 1$ and $1/\beta$, obviously the Fisher information would be different. (Remember, derivatives are involved, so we cannot just substitute the transformed parameters in the information matrix.)

You should have in your repertoire of easy pieces the problem of working out the information matrix for $\theta = (\mu, \sigma)$ in the $\mathrm{N}(\mu, \sigma^2)$ distribution using all three methods; that is, (1) the expectation of the product of first derivatives with respect to the parameters, (2) the expectation of the second derivatives with respect to the parameters, and (3) the integrals of the derivatives with respect to the variable (which, in the first form above, is an expectation).

4.1.3 Lower Bounds on the Variance of Unbiased Estimators

The Information Inequality (CRLB) for Unbiased Estimators

How small can we get? For an unbiased estimator T of $g(\theta)$ in a family of densities satisfying the regularity conditions and such that T has a finite second moment, we have the matrix relationship

$$\mathrm{V}(T(X)) \succeq \left(\frac{\partial}{\partial\theta} g(\theta)\right)^{\mathrm{T}} (I(\theta))^{-1} \frac{\partial}{\partial\theta} g(\theta),$$

where we assume the existence of all quantities in the expression.

Note the meaning of this relationship in the multiparameter case: it says that the matrix

$$\mathrm{V}(T(X)) - \left(\frac{\partial}{\partial\theta} g(\theta)\right)^{\mathrm{T}} (I(\theta))^{-1} \frac{\partial}{\partial\theta} g(\theta)$$

is nonnegative definite. (The zero matrix is nonnegative definite.)

This is called the information or the Cramér-Rao lower bound (CRLB). The CRLB results from the covariance inequality. The proof of the CRLB is an "easy piece" that every student should be able to provide quickly.

Consider a random sample $X_1, \ldots, X_n$, $n > 1$, from the $\mathrm{N}(\mu, \sigma^2)$ distribution. In this case, let's use the parametrization $\theta = (\mu, \sigma^2)$. The joint log density is

$$\log p_{(\mu,\sigma)}(x) = c - \frac{n}{2}\log(\sigma^2) - \sum_i (x_i - \mu)^2/(2\sigma^2).$$

The information matrix is diagonal, so the inverse of the information matrix is particularly simple:

$$I(\theta)^{-1} = \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{\sigma^4}{2(n-1)} \end{bmatrix}.$$

For the simple case of $g(\theta) = (\mu, \sigma^2)$, we have the unbiased estimator, $T(X) = (\overline{X}, \sum_{i=1}^n (X_i - \overline{X})^2/(n-1))$, and

$$\mathrm{V}(T(X)) = \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{\sigma^4}{2(n-1)} \end{bmatrix},$$

which is the same as the inverse of the information matrix. The estimators are Fisher efficient.

A more general information inequality (that is, without reference to unbiasedness) is

$$\mathrm{V}(T(X)) \succeq \left(\frac{\partial}{\partial\theta} \mathrm{E}(T(\theta))\right)^{\mathrm{T}} (I(\theta))^{-1} \frac{\partial}{\partial\theta} \mathrm{E}(T(\theta)).$$

It is important to know in what situations an unbiased estimator can achieve the CRLB. Notice this would depend on both $(p(X,\theta))$ and $g(\theta)$. The necessary and sufficient condition that an estimator T of $g(\theta)$ attain the CRLB is that $(T - g(\theta))$ be proportional to $\partial \log(p(X,\theta))/\partial\theta$ a.e.; that is, for some a that does not depend on X,

$$\frac{\partial \log(p(X,\theta))}{\partial\theta} = a(\theta)(T - g(\theta)) \quad \text{a.e.}$$

For example, there are unbiased estimators of the mean in the normal, Poisson, and binomial families that attain the CRLB. There is no unbiased estimator of θ that attains the CRLB in the family of distributions with densities proportional to $(1 + (x-\theta)^2)^{-1}$ (this is the Cauchy family).

If the CRLB is attained for an estimator of $g(\theta)$, it cannot be attained for any other (independent) function of θ. For example, there is no unbiased estimator of μ^2 in the normal distribution that achieves the CRLB.

If the CRLB is not sharp, that is, if it cannot be attained, there may be other (larger) bounds, for example the Bhattacharyya bound. These sharper bounds are usually based on higher-order derivatives.

4.2 U Statistics

In estimation problems, as we have seen, it is often fruitful to represent the estimand as some functional Υ of the CDF, P. The mean, for example, if it exists is

$$\Upsilon(P) = \int x \, \mathrm{d}P. \tag{4.1}$$

Given a random sample $X_1, \ldots, X_n$, we can form a plug-in estimator of $\Upsilon(P)$ by applying the functional to the ECDF.

In more complicated cases, the property of interest may be the quantile associated with π, that is, the unique value y_π defined by

$$\Xi_\pi(P) = \inf_y \{y : P(y) \geq \pi\}. \tag{4.2}$$

There is a basic difference in the functionals in equations (4.1) and (4.2). The first is an expected value, $\mathrm{E}(X_i)$ for each i. The second functional, however, cannot be written as an expectation. (Bickel and Lehmann, 1969, showed this.)

In the following, we will consider the class of statistical functions that can be written as an expectation of a function h of some subsample, $X_{i_1}, \ldots, X_{i_m}$, where $i_1, \ldots, i_m$ are distinct elements of $\{1, \ldots, n\}$:

$$\begin{aligned} \theta &= \Theta(P) \\ &= \mathrm{E}(h(X_{i_1}, \ldots, X_{i_m})). \end{aligned} \tag{4.3}$$

Such θs are called *expectation functionals*. In the case of Υ, h is the identity and $m = 1$.

Expectation functionals that relate to some parameter of interest are often easy to define. The simplest is just $\mathrm{E}(h(X_i))$. The utility of expectation functionals lies in the ease of working with them coupled with some useful general properties.

Note that without loss of generality we can assume that h is symmetric in its arguments because the X_is are i.i.d., and so even if h is not symmetric, any permutation $(i_1, \ldots, i_m)$ of the indexes has the same expectation, so we could form a function that is symmetric in the arguments and has the same expectation:

$$\bar{h}(X_1, \ldots, X_m) = \frac{1}{m!} \sum_{\text{all permutations}} h(X_{i_1}, \ldots, X_{i_m}).$$

Because of this, we will just need to consider h evaluated over the possible combinations of m items from the sample of size n. Furthermore, because the X_{i_j} are i.i.d., the properties of $h(X_{i_1}, \ldots, X_{i_m})$ are the same as the properties of $h(X_1, \ldots, X_m)$.

Now consider the estimation of an expectation functional θ, given a random sample $X_1, \ldots, X_n$, where $n \geq m$.

Clearly $h(X_1, \ldots, X_m)$ is an unbiased estimator of θ, and so is $h(X_{i_1}, \ldots, X_{i_m})$ for any m-tuple, $1 \leq i_1 < \cdots < i_m \leq n$; hence, we have that

$$U = \frac{1}{\binom{n}{m}} \sum_{\text{all combinations}} h(X_{i_1}, \ldots, X_{i_m}) \tag{4.4}$$

is unbiased for θ.

A statistic of this form is called a *U-statistic*. The U-statistic is a function of all n items in the sample. The function h, which is called the *kernel* of the U-statistic is a function of m arguments. The number of arguments of the kernel is called the *order of the kernel*. We also refer to the order of the kernel as the *order of the U-statistic*.

In the simplest U-statistic, the kernel is of order 1 and h is the identity, $h(x_i) = x_i$. This is just the sample mean, which we can immediately generalize by defining $h_r(x_i) = x_i^r$, yielding the first order U-statistic

$$U(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i^r,$$

the sample r^{th} moment.

Another simple U-statistic has the kernel of order 2

$$h(x_1, x_2) = \frac{1}{2}(x_1 - x_2)^2,$$

and is

$$U(X_1, \ldots, X_n) = \frac{2}{n(n-1)} \sum_{i<j}^{n} h(X_i, X_j). \tag{4.5}$$

This U-statistic is the sample variance S^2, which is unbiased for the population variance if it exists.

The quantile problem is related to an inverse problem in which the property of interest is the π; that is, given a value a, estimate $P(a)$. We can write an expectation functional and arrive at the U-statistic

$$\begin{aligned} U(X_1, \ldots, X_n) &= \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,a]}(X_i) \\ &= P_n(a), \end{aligned}$$

where P_n is the ECDF.

Occasionally, the kernel will include some argument computed from the full sample; that is, an m^{th} order kernel involves more than m items from the sample. An example of such a kernel is $h(X_i, \overline{X}) = (X_i - \overline{X})^2$. The U-statistic with this kernel is $\sum(X_i - \overline{X})^2/n = (n-1)S^2/n$. If the population mean is μ, the expected value of $(X_i - \mu)^2$ is the population variance, say σ^2, so at first glance, we might think that the expected value of this kernel is σ^2. Because X_i is included in $\overline{X}$, however, we have

$$\begin{aligned} \mathrm{E}\left(h(X_i, \overline{X})\right) &= \mathrm{E}\left(\left((n-1)X_i/n - \sum_{j\neq i} X_j/n\right)^2\right) \\ &= \mathrm{E}\left((n-1)^2 X_i^2/n^2 - 2(n-1)X_i \sum_{j\neq i} X_j/n^2 \right. \\ &\qquad \left. + \sum_{j\neq k\neq i\neq j} X_j X_k/n^2 + \sum_{j\neq i} X_j^2/n^2\right) \\ &= (n-1)^2\mu^2/n^2 + (n-1)^2\sigma^2/n^2 - 2(n-1)(n-1)\mu^2/n^2 \\ &\qquad +(n-1)(n-2)\mu^2/n^2 + (n-1)\mu^2/n^2 + (n-1)\sigma^2/n^2 \\ &= \frac{n-1}{n}\sigma^2. \end{aligned}$$

(We would, of course, expect this expectation to be less than σ^2, because the expectation of $(X_i - \mu)^2$, which does not have $(n-1)X_i/n$ subtracted out, is σ^2.)

If instead of the kernel h above, we used the kernel

$$g(X_i, \overline{X}) = \frac{n}{n-1}(X_i - \overline{X})^2,$$

we would have an expectation functional of interest; that is, one such that $\mathrm{E}(g(X_1, \ldots, X_m))$ is something of interest, namely σ^2.

A familiar second order U-statistic is *Gini's mean difference*, in which $h(x_1, x_2) = |x_2 - x_1|$, for $n \geq 2$,

$$U = \frac{1}{\binom{n}{2}} \sum_{i<j} |X_j - X_i|. \tag{4.6}$$

Another common second order U-statistic is the *one-sample Wilcoxon statistic*, in which $h(x_1, x_2) = \mathrm{I}_{(-\infty,0]}(x_1 + x_2)$, for $n \geq 2$,

$$U = \frac{1}{\binom{n}{2}} \sum_{i<j} \mathrm{I}_{(-\infty,0]}(X_i + X_j). \tag{4.7}$$

This is an unbiased estimator of $\Pr(X_1 + X_2 \leq 0)$.

We can generalize U-statistics in an obvious way to independent random samples from more than one population. We do not require that the number of elements used as arguments to the kernel be the same; hence, the order of the kernel is a vector whose number of elements is the same as the number of populations. A common U-statistic involving two populations is the *two-sample Wilcoxon statistic*. For this, we assume that we have two samples $X_{11}, \ldots, X_{1n_1}$ and $X_{21}, \ldots, X_{2n_2}$. The kernel is $h(x_{1i}, x_{2j}) = \mathrm{I}_{(-\infty,0]}(x_{2j} - x_{1i})$. The two-sample Wilcoxon statistic is

$$U = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \mathrm{I}_{(-\infty,0]}(X_{2j} - X_{1i}). \tag{4.8}$$

This is an unbiased estimator of $\Pr(X_{11} \leq X_{21})$.

4.2.1 Properties of U Statistics

U-statistics have a number of interesting properties. They are useful in nonparametric inference because of, among other reasons, they are asymptotically the same as the plug-in estimator that is based on the empirical CDF. Some of the important statistics used in modern computational statistical methods are U-statistics.

By conditioning on the order statistics, we can show that U-statistics are UMVUE for their expectations.

A sequence of adjusted kernels forms a martingale

If $\mathrm{E}((h(X_1, \ldots, X_m)^2) < \infty$, it is a simple matter to work out the variance of the corresponding U-statistic. ********

4.2.2 Projections of U Statistics

4.2.3 V Statistics

As we have seen, a U-statistic is an unbiased estimator of an expectation functional; specifically, if $\Theta(P) = \mathrm{E}(h(X_1, \ldots, X_m))$ the U-statistic with kernel h

is unbiased for $\Theta(P)$. Applying the functional Θ to the ECDF P_n, we have

$$\begin{aligned}\Theta(P_n) &= \frac{1}{n^m}\sum_{i_1=1}^{n}\cdots\sum_{i_m=1}^{n}(h(X_{i_1},\ldots,X_{i_m}))\\ &= V \quad \text{(say)},\end{aligned} \tag{4.9}$$

which we call the *V-statistic* associated with the kernel h, or equivalently associated with the U-statistic with kernel h. Recalling that $\Theta(P_n)$ in general is not unbiased for $\Theta(P)$, we do not expect a V-statistic to be unbiased in general. However, in view of the asymptotic properties of P_n, we might expect V-statistics to have good asymptotic properties.

A simple example is the variance, for which the U-statistic in equation (4.5) is unbiased. The V-statistic with the same kernel is

$$\begin{aligned}V &= \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - X_j)^2\\ &= \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i^2 + X_j^2 - 2X_iX_j\\ &= \frac{n-1}{n}S^2,\end{aligned}$$

where S^2 is the sample variance. This V-statistic is the same as the plug-in estimator of the population variance, and as with the plug-in estimator, there is no particular underlying distribution assumed. It is also the same as the MLE estimator given an assumed underlying normal distribution. The V-statistic is biased for the population variance; but as we have seen, it has a smaller MSE than the unbiased U-statistic.

4.3 Asymptotically Unbiased Estimation

There are many situations when an unbiased estimator does not exist, or when we cannot form one easily, or when a biased estimator has better MSE for any finite sample than an unbiased estimator. A biased estimator that is asymptotically unbiased, and for which there is no dominating unbiased estimator, is often considered optimal.

Three general kinds of estimators may be of this type: estimators based on the method of moments, functions of unbiased estimators, and V statistics. Some of these estimators arise as plug-in statistics in the ECDF, such as those based on the method of moments, and others from a general plug-in rule, in which individual estimators are used in different parts of the formula for the estimand, such as ratio estimators.

We would like for such biased estimators to have either limiting bias or asymptotic bias of zero.

Method of Moments Estimators

A simple example of an estimator based on the method of moments is $\widetilde{S}^2 = (n-1)S^2/n$ as an estimator of the population variance, σ^2. This is the second central moment of the sample, just as σ^2 is of the population. We have seen that, in certain conditions, the MSE of $\widetilde{S}^2$ is less than that of S^2, and while it is biased, its limiting and asymptotic bias is zero and is of order $1/n$.

Although the second central sample moment is biased, the raw sample moments are unbiased for the corresponding raw population moments, if they exist.

Ratio Estimators

Ratio estimators, that is, estimators composed of the ratio of two separate estimators, often arise in sampling applications. Another situation is when an estimator is based on a linear combination of observations with different variances. If we have some way of estimating the variances so we can form a weighted linear combination, the resulting estimator may be (will be!) biased, but its MSE may be better than the unweighted estimator. Also, it is often the case that the biased estimator is asymptotically normal and unbiased.

V-Statistics

The development of V-statistics can be based on the idea of applying the same functional to the ECDF F_n as the functional that defines the estimand when applied to the CDF F, and which is the basis for the U-statistics. Since the ECDF assigns probability $1/n$ to each point of the values $X_1, \ldots, X_n$, any m independent variables with CDF F_n take on each of the possible m-tuples $(X_{i_1}, \ldots, X_{i_m})$ with probability $1/n^m$. The plug-in estimator, call it V, of θ is therefore

$$V = \frac{1}{n^m} \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} h(X_{i_1}, \ldots, X_{i_m}).$$

Notice for $m = 1$, V is a U-statistic; but consider $m = 2$, as above. We have

$$U = \frac{1}{n(n-1)} \sum_{i} \sum_{j \neq i} h((X_i, X_j),$$

however

$$\begin{aligned} V &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} h((X_i, X_j) \\ &= \frac{1}{n^2} \sum_{i} \sum_{j \neq i} h((X_i, X_j) + \frac{1}{n^2} \sum_{i=1}^{n} h((X_i, X_i) \end{aligned}$$

While, as we have seen U is unbiased for θ, we see that V is biased:

$$\begin{aligned}\mathrm{E}(V) &= \frac{n-1}{n}\theta + \frac{1}{n}\mathrm{E}(h(X_1, X_1)) \\ &= \theta + \frac{1}{n}\left(\mathrm{E}(h(X_1, X_1)) - \theta\right).\end{aligned}$$

An example of a V statistic with $m = 2$ uses $h(x_1, x_2) = (x_1 - x_2)^2/2$ and results in $(n-1)S^2/m$ as an estimator of σ^2, which we have discussed above. This is of course asymptotically unbiased.

Theorem 3.16 in Shao shows that under certain general conditions, V statistics have limiting normal distributions and are asymptotically unbiased.

4.4 Asymptotic Efficiency

Often a statistical procedure does not have some desirable property for any finite sample size, but the procedure does have that property asymptotically. The asymptotic properties that are of most interest are those defined in terms of a sequence that has a limiting standard normal distribution, $\mathrm{N}(0, 1)$, or more generally, $\mathrm{N}_k(0, I_k)$. A standard normal distribution of a statistic is desirable because in that case, it is easy to associate statements of probabilities with values of the statistic. It is also desirable because it is often easy to work out the distribution of functions of a statistic that has a normal distribution.

It is important to remember the difference in an asymptotic property and a limiting property. An *asymptotic distribution* is the same as a *limiting distribution*, but other asymptotic properties are defined, somewhat arbitrarily, in terms of a limiting distribution of some function of the sequence of statistics and of a finite divergent or convergent sequence, a_n. This seems to mean that a particular asymptotic property, such as, say, the asymptotic variance, depends on what function of the sequence of statistics that we choose. Although there may be some degree of arbitrariness in "an" asymptotic expectation, there is a certain uniqueness, as expressed in Proposition 2.3 in Shao.

4.4.1 Asymptotic Relative Efficiency

We assume a family of distributions $\mathcal{P}$, a sequence of estimators $\{T_n\}$ of $g(\theta)$, and a sequence of constants $\{a_n\}$ with $\lim_{n\to\infty} a_n = \infty$ or with $\lim_{n\to\infty} a_n = a > 0$, and such that $a_n T_n(X) \to_{\mathrm{d}} T$ and $0 < \mathrm{E}(T) < \infty$. We define the asymptotic mean-squared error of $\{T_n\}$ for estimating $g(\theta)$ w.r.t. $\mathcal{P}$ as an asymptotic expectation of $(T_n - g(\theta))^2$; that is, $\mathrm{E}((T - g(\theta))^2)/a_n$, which we denote as $\mathrm{AMSE}(T_n, g(\theta), \mathcal{P})$.

For comparing two estimators, we may use the *asymptotic relative efficiency*, which for the estimators S_n and T_n of $g(\theta)$ w.r.t. $\mathcal{P}$ is

$$\mathrm{ARE}(S_n, T_n, \mathcal{P}) = \mathrm{AMSE}(S_n, g(\theta), \mathcal{P})/\mathrm{AMSE}(T_n, g(\theta), \mathcal{P}).$$

4.4.2 Asymptotically Efficient Estimators

Relative efficiency is a useful concept for comparing two estimators, whether or not they are unbiased. When we restrict our attention to unbiased estimators we use the phrase *Fisher efficient* to refer to an estimator that attains its Cramér-Rao lower bound. Again, notice the slight difference in "efficiency" and "efficient"; while one meaning of "efficiency" is a relative term that is not restricted to unbiased estimators (or other unbiased procedures, as we will see later), "efficient" is absolute. "Efficient" only applies to unbiased estimators, and an estimator either is or is not efficient. The state of being efficient, of course is called "efficiency". This is another meaning of the term. The phrase "Fisher efficiency" helps to emphasis this difference.

We consider the problem of estimating the k-vector θ based on a random sample $X_1, \ldots, X_n$. We denote the sequence of estimators as $\{\hat{\theta}_n\}$. Suppose

$$(V_n(\theta))^{-\frac{1}{2}} \left(\hat{\theta}_n - \theta\right) \rightarrow_{\mathrm{d}} \mathrm{N}_k(0, I_k),$$

where, for each n, $V_n(\theta)$ is a $k \times k$ positive definite matrix. From the definition of asymptotic expectation of $\left(\hat{\theta}_n - \theta\right)^2$, $V_n(\theta)$ is the asymptotic variance-covariance matrix of $\hat{\theta}_n$. Note that this matrix may depend on θ. We should note that for any fixed n, $V_n(\theta)$ is not necessarily the variance-covariance matrix of $\hat{\theta}_n$; that is, it is possible that $V_n(\theta) \neq \mathrm{V}(\hat{\theta}_n)$.

Just as we have defined Fisher efficiency for an unbiased estimator of fixed size, we define a sequence to be *asymptotically Fisher efficient* if the sequence is asymptotically unbiased, the Fisher information matrix $I_n(\theta)$ exists and is positive definite for each n, and $V_n(\theta) = (I_n(\theta))^{-1}$ for each n. The definition of asymptotically (Fisher) efficiency is often limited even further so as to apply only to estimators that are asymptotically normal. (Shao uses the restricted definition.)

Being asymptotically efficient does not mean for any fixed n that $\hat{\theta}_n$ is efficient. First of all, for fixed n, $\hat{\theta}_n$ may not even be unbiased; even if it is unbiased, however, it may not be efficient.

As we have emphasized many times, asymptotic properties are different from limiting properties. As a striking example of this, consider a very simple example from Romano and Siegel (1986). Let $X_1, \ldots, X_n \sim$ i.i.d $\mathrm{N}_1(\mu, 1)$, and consider a randomized estimator $\hat{\mu}_n$ of μ defined by

$$\hat{\mu}_n = \begin{cases} \overline{X}_n \text{ with probability } 1 - \frac{1}{n} \\ \\ n^2 \quad \text{with probability } \frac{1}{n}. \end{cases}$$

It is clear that $n^{1/2}(\hat{\mu}_n - \mu) \rightarrow_{\mathrm{d}} \mathrm{N}(0, 1)$, and furthermore, the Fisher information for μ is $n^{-1/2}$. The estimator $\hat{\mu}_n$ is therefore asymptotically Fisher efficient. The bias of $\hat{\mu}_n$, however, is

$$\mathrm{E}(\hat{\mu}_n - \mu) = \mu\left(1 - \frac{1}{n}\right) + n - \mu = n - \mu/n,$$

which tends to infinity, and the variance is

$$\begin{aligned}\mathrm{V}(\hat{\mu}_n) &= \mathrm{E}(\hat{\mu}^2) - (\mathrm{E}(\hat{\mu}))^2 \\ &= \left(1 - \frac{1}{n}\right)\frac{1}{n} + \left(\frac{1}{n}\right)n^4 - \left(\mu\left(1 - \frac{1}{n}\right) + n\right)^2 \\ &= n^3 + \mathrm{O}(n^2),\end{aligned}$$

which also tends to infinity. Hence, we have an asymptotically Fisher efficient estimator whose limiting bias and limiting variance are both infinite.

The example can be generalized to any estimator T_n of $g(\theta)$ such that $\mathrm{V}(T_n) = 1/n$ and $n^{1/2}(T_n - g(\theta)) \to_{\mathrm{d}} \mathrm{N}(0,1)$. From T_n form the estimator

$$\widetilde{T}_n = \begin{cases} T_n \text{ with probability } 1 - \frac{1}{n} \\ \\ n^2 \text{ with probability } \frac{1}{n}. \end{cases}$$

The estimator $\widetilde{T}_n$ is also asymptotically Fisher efficient but has infinite limiting bias and infinite limiting variance.

Asymptotic Efficiency and Consistency

Although asymptotic efficiency implies that the estimator is asymptotically unbiased, even if the limiting variance is zero, asymptotic efficiency does not imply consistency. The counterexample above shows this.

Likewise, of course, consistency does not imply asymptotic efficiency. There are many reasons. First, asymptotic efficiency is only defined in the case of asymptotic normality (of course, it is unlikely that a consistent estimator would not be asymptotically normal). More importantly, the fact that both the bias and the variance go to zero as required by consistency, is not very strong. There are many ways both of these can go to zero without requiring asymptotic unbiasedness or that the asymptotic variance satisfy the asymptotic version of the information inequality.

The Asymptotic Variance-Covariance Matrix

In the problem of estimating the k-vector θ based on a random sample $X_1, \ldots, X_n$ with the sequence of estimators as $\{\hat{\theta}_n\}$, if

$$(V_n(\theta))^{-\frac{1}{2}}\left(\hat{\theta}_n - \theta\right) \to_{\mathrm{d}} \mathrm{N}_k(0, I_k),$$

where, for each n, $V_n(\theta)$ is a $k \times k$ positive definite matrix, then $V_n(\theta)$ is the asymptotic variance-covariance matrix of $\hat{\theta}_n$. As we have noted, for any fixed n, $V_n(\theta)$ is not necessarily the variance-covariance matrix of $\hat{\theta}_n$.

If $V_n(\theta) = \mathrm{V}(\hat{\theta}_n)$, then under the information inequality regularity conditions that yield the CRLB, we know that

$$V_n(\theta) \geq (I_n(\theta))^{-1},$$

where $I_n(\theta)$ is the Fisher information matrix.

Superefficiency

Although if $V_n(\theta) \neq \mathrm{V}(\hat{\theta}_n)$, the CRLB says nothing about the relationship between $V_n(\theta)$ and $(I_n(\theta))^{-1}$, we might expect that $V_n(\theta) \geq (I_n(\theta))^{-1}$. That this is not necessarily the case is shown by a simple example given by Hodges in a lecture in 1951, published in Le Cam (1953) (see also Romano and Siegel, 1986).

Let $X_1, \ldots, X_n \sim$ i.i.d $\mathrm{N}_1(\mu, 1)$, and consider an estimator $\hat{\mu}_n$ of μ defined by

$$\hat{\mu}_n = \begin{cases} \overline{X}_n & \text{if } |\overline{X}_n| \geq n^{-1/4} \\ t\overline{X}_n & \text{otherwise,} \end{cases}$$

for some fixed t with $|t| < 1$.

What gives this example its kick is the dependence of the asymptotic distribution of $\hat{\mu}_n$ on μ. If $\mu \neq 0$, $\hat{\mu}_n$ has the same asymptotic distribution as $\overline{X}_n$, and obeys CRLB, both in its variance for finite n (even though it is biased) and in its asymptotic variance. However, if $\mu = 0$, $\hat{\mu}_n$ is still asymptotically unbiased, but the asymptotic variance of $\hat{\mu}_n$ is t^2/n, which is smaller than the inverse of asymptotic Fisher information, $1/n$.

A point in the parameter space at which this anomaly occurs is called a *point of superefficiency*. Le Cam has shown that under certain regularity conditions (that are slightly more stringent than the the information inequality regularity conditions, see page 60) the number of points of superefficiency is countable. (This is Theorem 4.16 in Shao.)

Superefficiency is not important in applications (that is, where n is finite) any decrease in mean-squared error at a point of superefficiency is accompanied by an increase in mean-squared error at nearby points (and, of course, if we knew the parameter was a point of superefficiency, we would probably not be estimating it.

4.5 Applications

4.5.1 Estimation in Linear Models

Many methods of statistical inference rely on samples of independent and identically distributed random variables.

Systematic and Random Components

In a simple variation on the i.i.d. requirement, we assume a model with two components, one "systematic" and one random. The most common form is one in which the random variable is the sum of a systematic component that determines its expected value and random component that is the value of an underlying unobservable random variable that has an expected value of 0. The systematic component may be a function of some additional variables z and parameters θ. If we represent the underlying unobservable random with expectation 0, as ϵ, we have

$$X = f(z, \theta) + \epsilon.$$

In this setup the mean of the random variable X is determined by the parameter θ and the values of the z variables, which are *covariates* (also called *regressors*, *carriers*, or *independent variables*). We generally treat the covariates as fixed variables, that is, whether or not we could also model the covariates as random variables, in the simplest cases, we will use their observed values without regard to their origin.

Regression Models

The model above is a regression model. In the simplest variation, the observable random variables are independent, and have distributions in the same location family: $\mathcal{P} = \{P_{f(z,\theta),P_\epsilon}\}$. The family $\mathcal{P}_\epsilon$ of distributions P_ϵ of the random component may be a parametric family, such as $\mathrm{N}(0, \sigma^2)$, or it may be a nonparametric family. Whatever other assumptions on P_ϵ, we assume $\mathrm{E}(\epsilon) = 0$.

Linear Models

Often we assume that the systematic component is a linear combination of the covariates. This setup is called a *linear model*, and is usually written in the form

$$Y = \beta^{\mathrm{T}} x + E,$$

where Y is the observable random variable, β is an unknown and unobservable p-vector of parameters, x is an observable p-vector of covariates, and E is an unobservable random variable with mean 0. The parameter space for β is $B \subset \mathbb{R}^p$.

An item of a random sample from this model may be denoted

$$Y_i = \beta^{\mathrm{T}} x_i + E_i,$$

and a random sample be written in the vector-matrix form

$$Y = X\beta + E,$$

where y and ϵ are n-vectors, X is an $n \times p$ matrix whose rows are the x_i^{T}, and β is the p-vector above. A sample of realizations may be written in the vector-matrix form

$$y = X\beta + \epsilon.$$

This is the most commonly used notation.

Shao's Notation

Shao uses X in place of Y; Z in place of x and also in place of X; and ϵ in place of E:

Inference in a Linear Model

Rather than formulating a decision problem and seeking a minimum risk estimator, for inference in a linear model, we usually begin with a different approach. Estimation in a linear model is most commonly developed based on two simple heuristics: least squares and unbiasedness.

The degree of β is p, meaning that the number of observations required for unbiased estimation of β is p. Inferences about characteristics of the distribution of ϵ require additional observations, however, and so we assume $n > p$ in the following.

Linear Least Squares

We define a *least squares* estimator (LSE) of β as

$$\hat{\beta} = \arg\min_{b \in B} \|X - Zb\|^2,$$

where $\|c\| = \|c\|_2 = \sqrt{c^{\mathrm{T}}c} = \sqrt{\sum_{i=1}^{p} c_i^2}$ for the p-vector c. A least squares estimator of β may or may not be unique. Whether or not β is unique, $\|X - Z\hat{\beta}\|^2/(n-p)$ is unique.

An LSE of β yields LSEs of other quantities. If $l \in \mathbb{R}^p$, then $l^{\mathrm{T}}\hat{\beta}$ is an LSE of $l^{\mathrm{T}}\beta$. Also, $\|X - Z\hat{\beta}\|^2/(n-p)$ is the LSE of $\mathrm{V}(\epsilon)$.

The least squares estimator is obtained by direct minimization of

$$\begin{aligned} s(b) &= \|X - Zb\|^2 \\ &= X^{\mathrm{T}}X - 2b^{\mathrm{T}}Z^{\mathrm{T}}X + b^{\mathrm{T}}Z^{\mathrm{T}}Zb. \end{aligned}$$

First of all, we note that $s(b)$ is differentiable, and

$$\frac{\partial^2}{\partial b^2} s(b) = Z^{\mathrm{T}}Z$$

is nonnegative definitive. We therefore know that at the minimum, $\partial s(b)/\partial b = 0$. This gives the *normal equations*:

$$Z^{\mathrm{T}}Zb = Z^{\mathrm{T}}X.$$

Gauss-Markov Theorem

The Gauss-Markov theorem provides a restricted optimality property for estimators of estimable functions of β under the condition that $\mathrm{E}(\epsilon) = 0$ and $\mathrm{V}(\epsilon) = \sigma^2 I$; that is, in addition to the assumption of zero expectation, which we have used above, we also assume that the elements of ϵ have constant variance and that their covariances are zero. (We are not assuming independence or normality.)

Given $X = Z\beta + \epsilon$ and $\mathrm{E}(\epsilon) = 0$ and $\mathrm{V}(\epsilon) = \sigma^2 I$, the Gauss-Markov theorem states that $l^{\mathrm{T}}\widehat{\beta}$ is the unique *best linear unbiased estimator* (BLUE) of the estimable function $l^{\mathrm{T}}\beta$.

"Linear" estimator in this context means a linear combination of X; that is, an estimator in the form $a^{\mathrm{T}}X$. It is clear that $l^{\mathrm{T}}\widehat{\beta}$ is linear, and we have already seen that it is unbiased for $l^{\mathrm{T}}\beta$. "Best" in this context means that its variance is no greater than any other estimator that fits the requirements. Hence, to prove the theorem, first let $a^{\mathrm{T}}X$ be any unbiased estimator of $l^{\mathrm{T}}\beta$, and write $l = Z^{\mathrm{T}}X\tilde{t}$ as above. Because $a^{\mathrm{T}}X$ is unbiased for any β, as we saw above, it must be the case that $a^{\mathrm{T}}Z = l^{\mathrm{T}}$. Recalling that $Z^{\mathrm{T}}Z\widehat{\beta} = Z^{\mathrm{T}}X$, we have

$$\begin{aligned}
\mathrm{V}(a^{\mathrm{T}}X) &= \mathrm{V}(a^{\mathrm{T}}X - l^{\mathrm{T}}\widehat{\beta} + l^{\mathrm{T}}\widehat{\beta}) \\
&= \mathrm{V}(a^{\mathrm{T}}X - \tilde{t}^{\mathrm{T}}Z^{\mathrm{T}}X + l^{\mathrm{T}}\widehat{\beta}) \\
&= \mathrm{V}(a^{\mathrm{T}}X - \tilde{t}^{\mathrm{T}}Z^{\mathrm{T}}X) + \mathrm{V}(l^{\mathrm{T}}\widehat{\beta}) + 2\mathrm{Cov}(a^{\mathrm{T}}X - \tilde{t}^{\mathrm{T}}Z^{\mathrm{T}}X,\ \tilde{t}^{\mathrm{T}}Z^{\mathrm{T}}X).
\end{aligned}$$

Now, under the assumptions on the variance-covariance matrix of ϵ, which is also the (conditional, given Z) variance-covariance matrix of y, we have

$$\begin{aligned}
\mathrm{Cov}(a^{\mathrm{T}}y - \tilde{t}^{\mathrm{T}}Z^{\mathrm{T}}X,\ l^{\mathrm{T}}\widehat{\beta}) &= (a^{\mathrm{T}} - \tilde{t}^{\mathrm{T}}Z^{\mathrm{T}})\sigma^2 I Z\tilde{t} \\
&= (a^{\mathrm{T}}Z - \tilde{t}^{\mathrm{T}}Z^{\mathrm{T}}Z)\sigma^2 I\tilde{t} \\
&= (l^{\mathrm{T}} - l^{\mathrm{T}})\sigma^2 I\tilde{t} \\
&= 0;
\end{aligned}$$

that is,

$$\mathrm{V}(a^{\mathrm{T}}X) = \mathrm{V}(a^{\mathrm{T}}X - \tilde{t}^{\mathrm{T}}Z^{\mathrm{T}}X) + \mathrm{V}(l^{\mathrm{T}}\widehat{\beta}).$$

This implies that

$$\mathrm{V}(a^{\mathrm{T}}X) \geq \mathrm{V}(l^{\mathrm{T}}\widehat{\beta});$$

that is, $l^{\mathrm{T}}\widehat{\beta}$ has minimum variance among the linear unbiased estimators of $l^{\mathrm{T}}\beta$. To see that it is unique, we consider the case in which $\mathrm{V}(a^{\mathrm{T}}X) = \mathrm{V}(l^{\mathrm{T}}\widehat{\beta})$; that is, $\mathrm{V}(a^{\mathrm{T}}X - \tilde{t}^{\mathrm{T}}Z^{\mathrm{T}}X) = 0$. For this variance to equal 0, it must be the case that $a^{\mathrm{T}} - \tilde{t}^{\mathrm{T}}Z^{\mathrm{T}} = 0$ or $a^{\mathrm{T}}X = \tilde{t}^{\mathrm{T}}Z^{\mathrm{T}}X = l^{\mathrm{T}}\widehat{\beta}$; that is, $l^{\mathrm{T}}\widehat{\beta}$ is the unique linear unbiased estimator that achieves the minimum variance.

If we assume further that $\epsilon \sim \mathrm{N}_n(0, \sigma^2 I)$, we can show that $l^{\mathrm{T}}\widehat{\beta}$ is the uniformly minimum variance unbiased estimator (UMVUE) for $l^{\mathrm{T}}\beta$. This is

because $(Z^{\mathrm{T}}X,\ (X - Z\widehat{\beta})^{\mathrm{T}}(X - Z\widehat{\beta}))$ is complete and sufficient for (β, σ^2). This line of reasoning also implies that $(X - Z\widehat{\beta})^{\mathrm{T}}(X - Z\widehat{\beta})/(n - r)$, where $r = \mathrm{rank}(Z)$, is UMVUE for σ^2.

4.5.2 Estimation in Survey Samples of Finite Populations

A substantial proportion of all applications of statistics deal with sample surveys in finite populations. Some aspects of this kind of application distinguish it from other areas of applied statistics. Särndal, Swensson, and Wretman (1997) provide a general coverage of the theory and methods. Valliant, Dorfman, and Royall (2000) provide a different perspective on some of the particular issues of inference in finite populations.

Finite Populations

We think of a finite population as being a finite set $\mathcal{P} = \{y_1, \ldots, y_N\}$. (Note that here we use "population" in a different way from the use of the term as a probability measure.) Our interest will be in making inferences about the population using a sample $X = \{X_1, \ldots, X_n\}$. In discussions of sampling it is common to use n to denote the size of the sample and N to denote the size of the population. Another common notation used in sampling is Y to denote the population total, $Y = \sum_{i=1}^{N} y_i$. The total is one of the most basic objectives in sampling applications.

The parameter that characterizes the population is $\theta = (y_1, \ldots, y_N)$. The parameter space, Θ, is the subspace of $\mathbb{R}^N$ containing all possible values of the y_i.

There are two approaches to the analysis of the problem. In one, which is the more common and which we will follow, $\mathcal{P}$ is essentially the sample space. In another approach $\mathcal{P}$ or θ is thought of as some random sample from a sample space or parameter space, called a "superpopulation".

The sample is completely determined by the set $\mathcal{S} = \{i_1, \ldots, i_n\}$ of indexes of $\mathcal{P}$ that correspond to elements in X. (Shao uses s where I use $\mathcal{S}$.)

"Sampling" can be thought of as selecting the elements of $\mathcal{S}$.

Probability-based inferences about $\mathcal{P}$ are determined by the method of selection of $\mathcal{S}$. This determines the probability of getting any particular $\mathcal{S}$, which we will denote by $p(\mathcal{S})$. If $p(\mathcal{S})$ is constant for all $\mathcal{S}$, we call the selected sample a *simple random sample*.

A sample may be collected *without replacement* or *with replacement*. (The meanings of these are just what the words mean. In sampling without replacement, the elements of $\mathcal{S}$ are distinct.) Sampling with replacement is generally easier to analyze, because it is the same as taking a random sample from a discrete uniform distribution. Sampling without replacement is more common and it is what we will assume throughout.

There are many variations on the method of collecting a sample. Both a general knowledge of the population and some consideration of the mechanical aspects of collecting the sample may lead to the use of *stratified sampling*, *cluster sampling*, *multi-stage sampling*, *systematic sampling*, or other variations.

Estimation

We are interested in "good" estimators, specifically UMVUEs, of estimable functions of θ. An interesting estimable function of θ is $Y = \sum_{i=1}^{N} \theta_i$.

The first important result is the Watson-Royall theorem (Shao's Theorem 3.13):
(i) if $p(\mathcal{S}) > 0$ for all $\mathcal{S}$, then the set of order statistics $X_{(1)}, \ldots, X_{(n)}$ is complete for all $\theta \in \Theta$.
and
(ii) if $p(\mathcal{S})$ is constant for all $\mathcal{S}$, then the order statistics $X_{(1)}, \ldots, X_{(n)}$ are sufficient for all $\theta \in \Theta$.

This theorem is somewhat similar to Examples 2.12 and 2.17, which applied to the family of distributions dominated by Lebesgue measure. The sufficiency is generally straightforward, and we expect it to hold in any i.i.d. case.

The completeness is a little more complicated, and Shao's proof is worth looking at. The set of order statistics may be complete in some family, such as *the* family of distributions dominated by Lebesgue measure, but may not be complete in some subfamily, such as the family of normal distributions with mean 0.

After we have (i) and (ii), we have
(iii): For any estimable function of θ, its unique UMVUE is the unbiased estimator $T(X_1, \ldots, X_n)$ that is symmetric in its arguments. (The symmetry makes the connection to the order statistics.)

Consider estimation of $Y = g(\theta) = \sum_{i=1}^{N} y_i$, from the simple random sample $X_1, \ldots, X_n$. We can see easily that $\widehat{Y} = N\overline{X}$ is the UMVUE. This statistic was considered in Example 2.27, where Shao showed that the variance of $\widehat{Y}$ is composed of three parts, an expansion factor N^2/n, a finite population correction factor $(1 - n/N)$, and the variance of a selection from a finite population,

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(y_i - \frac{Y}{N} \right)^2 .$$

It is a simple exercise (which I did not assign, but you should work out) to show that the sample variance S^2 is unbiased for σ^2, and then from this we have immediately the UMVUE of $\mathrm{V}(\widehat{Y})$.

Horvitz-Thompson Estimation

The properties of any statistic derived from a sample $X_1, \ldots, X_n$ depend on the sampling design; that is, on how the items in the sample were selected. The two main properties of the design are the probability that a specific population item, say y_i, is selected, and the probability that two specific population items, say y_i and y_j are both selected. Probabilities of combinations of larger sets may also be of interest, but we can work out simple expectations and variances just based on these two kinds of probabilities.

Let π_i be the probability that y_i is included in the sample, and let π_{ij} be the probability that both y_i and y_j are included.

If $\pi_i > 0$ for all i, the *Horvitz-Thompson estimator* of the population total is

$$\widehat{Y}_{\mathrm{HT}} = \sum_{i \in \mathcal{S}} \frac{y_i}{\pi_i}.$$

It is easy to see that $\widehat{Y}_{\mathrm{HT}}$ is unbiased for Y.

The variance of the Horvitz-Thompson estimator depends on the π_{ij} as well as the π_i. It is given in equation (3.48). (The "advanced arithmetic" used in the derivation of this formula is one thing that turns graduate students off from pursuing research in sampling.) Expressions for other sampling estimators are often shown in a similar manner. The other main thing that is used in working out variances of sampling estimators is linearization, especially when the estimator involves a ratio.

Notes

Unbiasedness

Although unbiasedness has a heuristic appeal, one of the main reasons for requiring it is to be able to obtain uniformly minimum risk estimators for squared error loss functions. For absolute error loss functions, a corresponding approach would be to require median unbiasedness.

Exercises in Shao

- For practice and discussion
  3.6, 3.19, 3.33, 3.34, 3.60, 3.70, 3.106, 3.107, 3.111 (Solutions in Shao, 2005)
- To turn in
  3.3, 3.16, 3.32(a)(b)(c), 3.35(a)(b)(c), 3.44, 3.52, 3.91, 3.109, 3.114

Additional References

Bickel, P. J., and E. L. Lehmann (1969), Unbiased estimation in convex families, *Annals of Mathematical Statistics* **40**, 1523–1535.

Le Cam, L. (1953), On some asymptotic properties of maximum likelihood estimates and related Bayes estimates, *University of California Publications in Statistics* **1**, 277–330. (Though this is a widely-cited paper and a review, MR0054913, appeared in *Mathematical Review*, no such serial as *University of California Publications in Statistics* seems to be widely available.)

Särndal, Carl-Erik; Bengt Swensson; and Jan Wretman (1997) *Model Assisted Survey Sampling*, Springer, New York.

Valliant, Richard; Alan H. Dorfman; and Richard M. Royall (2000), *Finite Population Sampling and Inference: A Prediction Approach*, John Wiley & Sons, New York.

5

Maximum Likelihood Estimation (Shao Sec 4.4, Sec 4.5, Sec 5.4; TPE2 Ch 6)

5.1 The Likelihood Function and Its Use in Parametric Estimation

One of the most commonly-used approaches to statistical estimation is *maximum likelihood*. The concept has an intuitive appeal, and the estimators based on this approach have a number of desirable mathematical properties, at least for broad classes of distributions.

The Likelihood Function

Given a sample $x_1, \ldots, x_n$ from distributions with probability densities $p_i(x)$ with respect to a common σ-finite measure, the *likelihood function* is

$$L_n(p_i\,;\,x) = \prod_{i=1}^{n} p_i(x_i).$$

(Any nonnegative function proportional to $L_n(p_i\,;\,x)$ is a likelihood function, but it is common to speak of $L_n(p_i\,;\,x)$ as "the" likelihood function.) We can view the sample either as a set of constants, or as a sample of random variables.

The domain of the likelihood is some class of distributions specified by their probability densities, $\{p\}$, where all PDFs are with respect to a common σ-finite measure.

If the sample is from a distribution with probability density $p_\theta(x)$, a reasonable estimate of p_θ is the nonnegative function p_{θ_*} that has an integral of 1 over the support of $p_\theta(x)$ for which the likelihood function,

$$L_n(p_\theta\,;\,x) = \prod_{i=1}^{n} p_\theta(x_i), \tag{5.1}$$

is maximized. If this function is unbounded above, the maximum does not exist.

This approach is called maximum likelihood, or ML.

Notice that this is an ill-posed problem. (Why?) In nonparametric models, it may be difficult to resolve this quandary. (And, of course, in such models, we may not know the support of the distribution, so a first step in nonparametric estimation may be to change the normalization requirement to be that the function has an integral of 1 over $[\min(x_i), \max(x_i)]$.) An example of a likelihood function that is not very useful without some modification is in nonparametric probability density estimation. Suppose we assume that a sample comes from a distribution with continuous PDF $p(x)$. The likelihood is $\prod_{i=1}^{n} p(x_i)$. Even under the assumption of continuity, there is no solution.

Another example in which the likelihood function is not very meaningful, due to C. R. Rao, is the case of N balls labeled $1, \ldots, N$ and also labeled with distinct real numbers $\theta_1, \ldots, \theta_N$ (with N known). For a sample without replacement of size $n < N$ where we observe $(x_i, y_i) = (\text{label}, \theta_{\text{label}})$, what is the likelihood function? It is either 0, if the label and θ_{label} for at least one observation is inconsistent, or $\binom{N}{n}^{-1}$, otherwise; and, of course, we don't know! This likelihood function is not informative, and could not be used, for example, for estimating $\theta = \theta_1 + \cdots + \theta_N$. (There is a pretty good estimator of θ; it is $N(\sum y_i)/n$.)

Although the likelihood function has an intuitive appeal, one of the reasons that it is important important in statistical inference is its asymptotic properties. For that reason, it is very common to use the n subscript. In the following, however, we will often find it convenient to drop the n.

What Likelihood Is Not

First of all:

- Likelihood is not a probability.
- Likelihood is not a probability density.

Although non-statisticians will often refer to the "likelihood of an observation", in statistics, we use the term "likelihood" to refer to a model or a distribution *given observations*.

The Log-Likelihood Function

The *log-likelihood function*,

$$l_L(p_\theta\,;\,x) = \log L_n(p_\theta\,;\,x), \tag{5.2}$$

is a sum rather than a product. The form of the log-likelihood in the exponential family is particularly simple:

$$l_L(\theta\,;\,x) = \sum_{i=1}^{n} \theta^{\mathrm{T}} g(x_i) - n\,a(\theta) + c,$$

where c depends on the x_i, but is constant with respect to the variable of interest.

The logarithm is monotone, so the optimization problem (5.1) can be solved by solving the maximization problem with the log-likelihood function:

$$\max_{\theta} l_L(\theta\,;\,x). \tag{5.3}$$

We will often work with the likelihood and log-likelihood as if there is only one observation. (A general definition of a likelihood function is any nonnegative function that is proportional to the density or the probability mass function; that is, it is the same as the density or the probability mass function except that the arguments are switched, and its integral or sum over the domain of the random variable need not be 1.)

5.1.1 Parametric Estimation

Let us assume a parametric model; that is, a family of densities $\mathcal{P} = \{p_\theta(x)\}$ where $\theta \in \Theta$, a known parameter space. In the parametric case, it is usually more convenient to write $p_\theta(x)$ as $p(x\,;\,\theta)$. Let us also assume the "regular case", which is guaranteed by the Fisher information regularity conditions. The important regularity conditions are that $p(x\,;\,\theta)$ is twice differentiable with respect to θ and that there is a common support of all distributions in $\mathcal{P}$. (Without this latter assumption, the varying support may effectively allow the data to provide information that allows parameters to be estimated with greater efficiency than is attainable in the regular case. Recall the discussion on superefficiency on page 185.)

For a sample $X_1, \ldots, X_n$ from a distribution with probability density $p(x\,;\,\theta)$, we write the likelihood function as a function of a variable in place of the parameter:

$$L(t\,;\,x) = \prod_{i=1}^{n} p(x_i\,;\,t). \tag{5.4}$$

For a discrete distribution, the likelihood is defined with the probability mass function in place of the density in equation (5.4).

It is important to specify the domain of the likelihood function. If Θ is the domain of L in equation (5.4), we want to maximize L for $t \in \Theta$.

Note the reversal in roles of variables and parameters. We sometimes write the expression for the likelihood without the observations: $L(\theta)$.

While I really like to write the likelihood as a variable of something other than the parameter, which I think of as fixed, I usually write it like everyone else: $L(\theta\,;\,x) = \prod_{i=1}^{n} p(x_i\,;\,\theta)$.

The data, that is, the realizations of the variables in the density function, are considered as fixed and the parameters are considered as variables of the optimization problem,

$$\max_{\theta} L(\theta\,;\, x). \tag{5.5}$$

If Θ is not a closed set, the maximum may not exist, so we consider the closure of Θ, $\overline{\Theta}$. (If Θ is closed $\overline{\Theta}$ is the same set, so we can always just consider $\overline{\Theta}$.)

The *maximum likelihood estimate* of θ, written $\hat{\theta}$, is defined as

$$\hat{\theta} = \arg\max_{\theta \in \overline{\Theta}} L(\theta\,;\, x),$$

if it exists. If the maximum does not exist because the likelihood is unbounded from above, then the argmax does not exist, and the maximum likelihood estimate does not exist.

Notice that $\hat{\theta}$ is a function of the observations, x. If $\hat{\theta}(x)$ is a Borel function, then $\hat{\theta}$ is called a *maximum likelihood estimator* of θ.

We use "MLE" to denote maximum likelihood estimate or estimator, or the method of maximum likelihood estimation. The proper word can be determined from the context. If in a statement is about a maximum likelihood estimate or estimator, and the term MLE is used, then the statement can be assumed to apply to both the estimate and the estimator.

If $\hat{\theta}$ is an MLE of θ, and g is a Borel function, then $g(\hat{\theta})$ is an MLE of the estimand $g(\theta)$.

In some cases the MLE occurs at a stationary point, which can be identified by differentiation. That is not always the case, however. A standard example in which the MLE does not occur at a stationary point is a distribution in which the range depends on the parameter, and the simplest such distribution is the uniform $\mathrm{U}(0, \theta)$.

In some cases the MLE may not be in Θ, for example in the Bernoulli cases, $\Theta = (0, 1)$, but it is possible that $\hat{\theta} = 0$ or 1. These values are in $\overline{\Theta}$, of course, and so either of them could be chosen as an MLE. This solution is preferable to saying that an MLE does not exist. It does, however, ignore the problem of continuity of $L(\theta\,;\, x)$ over $\overline{\Theta}$, and it allows an estimated PDF that is degenerate.

Often estimation based on least squares is the same as MLE. We almost always expect this to be the case when the underlying probability distribution is normal. There are situations, in analysis of mixed linear models, for example, in which the least squares approach leads to estimates of certain elements of the parameter θ that are not in $\overline{\Theta}$; specifically, some estimates of variance components are negative. These are not MLEs, by the definition above. One solution to this problem is called REML, "restricted maximum likelihood". This results from maximization restricted to $\overline{\Theta}$, which means it is really just ML. In general, we often encounter the phrase "constrained maximum likelihood". In most cases, this just means constrained to the closure of

the parameter space; that is, it is just regular ML. The problem of obtaining an MLE is often a *constrained optimization problem*.

Estimation of Parameters in an Open Parametric Range

Consider a distribution with PDF

$$p_X(x) = h(x,\theta)\mathrm{I}_{S(\theta)}(x) \tag{5.6}$$

where $S(\theta)$ is open. In this case, the likelihood has the form

$$L(\theta\,;\,x) = h(x,\theta)\mathrm{I}_{R(x)}(\theta),$$

where $R(x)$ is open. It is quite possible that $\sup L(\theta\,;\,x)$ will occur on $\overline{R}(x)$. Consider, for example, $X_1,\ldots,X_n$ i.i.d. with PDF

$$p_X(x) = \frac{1}{\theta}\mathrm{I}_{(0,\theta)}(x),$$

where $\Theta = \mathbb{R}$.

The likelihood is

$$L(\theta\,;\,x) = \frac{1}{\theta}\mathrm{I}_{(x_{(n)},\infty)}(\theta).$$

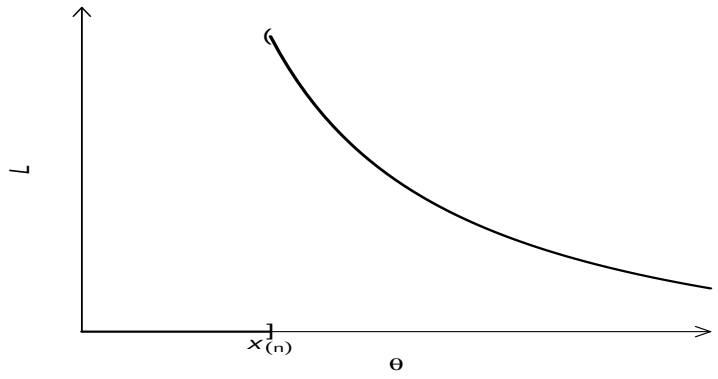


Fig. 5.1. Discontinuous Likelihood

The maximum of the likelihood does not exist. The supremum of the likelihood occurs at $x_{(n)}$ and it is finite.

We reasonably want to call $x_{(n)}$ the MLE of θ.

Because the distribution has a PDF w.r.t. a σ-finite measure ν such that $\nu(A) = 0$ if A does not contain an open set (Lebesgue has this property, for example), we can define an a.e. equivalent PDF:

$$\bar{p}_X(x) = \frac{1}{\theta} \mathrm{I}_{[0,\theta]}(x),$$

where $\Theta = \mathbb{R}$. This yields the likelihood

$$\bar{L}(\theta\,;\, x) = \frac{1}{\theta} \mathrm{I}_{[x_{(n)},\infty)}(\theta),$$

which is continuous from the right. Now, the max occurs at $x_{(n)}$ and so $x_{(n)}$ is the MLE of θ, as we would want it to be.

This is the general approach. Given a PDF of the form (5.6) w.r.t. a continuous measure like Lebesgue, we could just say the MLE does not exist, but that would not be satisfactory.

Hence, we form an a.e. equivalent PDF,

$$\bar{p}_X(x) = h(x) \mathrm{I}_{\overline{S(\theta)}}(x).$$

(Notice that substituting $\overline{S}(\theta)$ for $S(\theta)$ is very different from substituting $\overline{\Theta}$ for Θ in the definition of an MLE. The parameter space may or may not be open.)

This approach is cleaner than solving the logical problem by defining the MLE in terms of the sup rather than the max. A definition in terms of the sup may not address problems that could arise due to various types of discontinuity of $L(\theta\,;\, x)$ at the boundary of $S(\theta)$.

Derivatives of the Likelihood and Log-Likelihood

In the regular case, the likelihood function and consequently the log-likelihood are twice differentiable within Θ°.

The derivatives of the log-likelihood function relate directly to useful concepts in statistical inference. If it exists, the derivative of the log-likelihood is the relative rate of change, with respect to the parameter placeholder θ, of the probability density function at a fixed observation. If θ is a scalar, some positive function of the derivative, such as its square or its absolute value, is obviously a measure of the effect of change in the parameter, or of change in the estimate of the parameter. More generally, an outer product of the derivative with itself is a useful measure of the changes in the components of the parameter:

$$\nabla l_L\big(\theta^{(k)}\,;\, x\big) \left(\nabla l_L\big(\theta^{(k)}\,;\, x\big)\right)^{\mathrm{T}}.$$

Notice that the average of this quantity with respect to the probability density of the random variable X,

$$I(\theta_1\,;\,X) = \mathrm{E}_{\theta_1}\left(\nabla l_L\big(\theta^{(k)}\,;\,X\big)\,\left(\nabla l_L\big(\theta^{(k)}\,;\,X\big)\right)^{\mathrm{T}}\right), \tag{5.7}$$

is the *information matrix* for an observation on Y about the parameter θ.

If θ is a scalar, the square of the first derivative is the negative of the second derivative,

$$\left(\frac{\partial}{\partial\theta} l_L(\theta\,;\,x)\right)^2 \;=\; -\frac{\partial^2}{\partial\theta^2} l_L(\theta\,;\,x),$$

or, in general,

$$\nabla l_L\big(\theta^{(k)}\,;\,x\big)\,\left(\nabla l_L\big(\theta^{(k)}\,;\,x\big)\right)^{\mathrm{T}} \;=\; -\,\mathrm{H}_{l_L}\big(\theta^{(k)}\,;\,x\big). \tag{5.8}$$

Evaluation of the Maximum of the Likelihood or Log-Likelihood

If the log-likelihood is twice differentiable and if the range does not depend on the parameter, Newton's method could be used to solve (5.3). Newton's equation

$$\mathrm{H}_{l_L}(\theta^{(k-1)}\,;\,x)\,d^{(k)} = \nabla l_L(\theta^{(k-1)}\,;\,x) \tag{5.9}$$

is used to determine the step direction in the k^{th} iteration. A quasi-Newton method uses a matrix $\widetilde{H}_{l_L}(\theta^{(k-1)})$ in place of the Hessian $\mathrm{H}_{l_L}(\theta^{(k-1)})$. (See notes on optimization in the Appendix.)

Equation (5.8) is interesting because the second derivative, or an approximation of it, is used in a Newton-like method to solve the maximization problem.

The Likelihood Equation

In the regular case, with the likelihood log-likelihood function differentiable within Θ°, we call

$$\nabla L(\theta\,;\,x) = 0$$

or

$$\nabla l_L(\theta\,;\,x) = 0$$

the *likelihood equations*. If the maximum occurs within Θ°, then every MLE is a root of the likelihood equations.

A likelihood equation is sometimes called an "estimating equation". Similar equations are called "generalized estimating equations", or GEEs.

Any root of the likelihood equations, which is called an RLE, may be an MLE. A theorem from functional analysis, usually proved in the context of numerical optimization, states that if θ_* is an RLE and $\mathrm{H}_{l_L}(\theta_*\,;\,x)$ is negative

definite, then there is a *local maximum* at θ_*. This may allow us to determine that an RLE is an MLE. There are, of course, other ways of determining whether an RLE is an MLE. In MLE, the determination that an RLE is actually an MLE is an important step in the process.

There are interesting open questions associated with determining if an RLE yields a global maximum. (See, e.g., Christophe Biernacki, 2005, Testing for a global maximum of the likelihood, *JCGS* **14**, 657–674.)

Easy piece: Determine the MLEs of μ and σ^2 in $\mathrm{N}(\mu, \sigma^2)$. (Don't forget to prove that your solution is actually a maximum.)

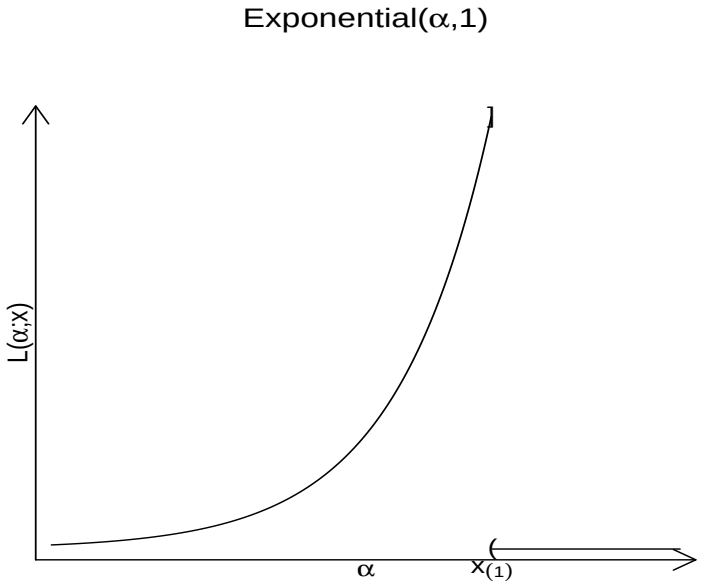
Nondifferentiable Likelihood Functions

The definition of MLEs does not depend on the existence of a likelihood equation. The likelihood function may not be differentiable with respect to the parameter, as in the case of the hypergeometric distribution, in which the parameter must be an integer (see Shao, Example 4.32).

Another example in which the derivative is not useful in finding the MLE is in a parametric-support family. For example, assume $X_1, \ldots, X_n \sim$ i.i.d. exponential$(\alpha, 1)$. The likelihood is

$$L(\alpha\,;\,x) = \mathrm{e}^{-\sum(x_i-\alpha)}\mathrm{I}_{(-\infty,x_{(1)}]}(\alpha).$$

Setting the derivative to 0 is not a useful way to find a stationary point. (Note that the derivative of the indicator function is the Dirac delta function.) In fact, the max does not occur at a stationary point. The MLE of α is $x_{(1)}$.



5.1.2 Properties of MLEs

As we have mentioned, MLEs have a nice intuitive property, and they also often have good asymptotic properties, as we will see later.

If there is a sufficient statistic and an MLE exists, then an MLE is a function of the sufficient statistic. We can see this very easily by use of the factorization theorem. (Note that this statement hints at two issues: existence of an MLE, and nonuniqueness of an MLE.)

Other properties are not always desirable.

First of all we note that an MLE may be biased. The most familiar example of this is the MLE $\hat{\sigma}^2$ in $\mathrm{N}(\mu, \sigma^2)$. Another example is the MLE of the location parameter in the exponential.

If two samples provide the same MLE for θ, say $\hat{\theta}$, then combining the two samples provides the same estimate, $\hat{\theta}$. We can see this must be the case because the likelihood for the combined sample is just the product of the likelihoods. Suppose for our two samples, we have $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$, with $\hat{\theta}_2^{(1)} = \hat{\theta}_2^{(2)}$, but $\hat{\theta}_1^{(1)} \neq \hat{\theta}_1^{(2)}$. In this case, after combining the two samples, we may get $\hat{\theta}_2^{(1+2)} \neq \hat{\theta}_2^{(1)}$. As an example, due to Romano and Siegel, consider two samples from a normal distribution, $S_1 = \{9, 10, 11\}$ and $S_2 = \{29, 30, 31\}$.

We get $(\hat{\sigma}_2^2)^{(1)} = (\hat{\sigma}_2^2)^{(2)} = 2/3$, but $(\hat{\sigma}_2^2)^{(1+2)} = 200/3$. (Look at the combined dataset, with mean 20.)

Nonuniqueness

There are many cases in which the MLEs are not unique (and I'm not just referring to RLEs). An example is the Cauchy distribution with location parameter θ. The likelihood equation is

$$\sum_{i=1}^{n} \frac{2(x_i - \theta)}{1 + (x_i - \theta)^2}.$$

This may have multiple roots (depending on the sample), and so the one yielding the maximum would be the MLE. Depending on the sample, however, multiple roots can yield the same value of the likelihood function.

Another example in which the MLE is not unique is $\mathrm{U}(\theta - 1/2, \theta + 1/2)$. The likelihood function is

$$\mathrm{I}_{(x_{(n)}-1/2,(x_{(1)}+1/2)}(\theta).$$

Where is it maximized? (Any value between $x_{(n)} - 1/2$ and $(x_{(1)} + 1/2$.)

Nonexistence and Other Properties

We have already mentioned situations in which the likelihood approach does not seem to be the logical way, and have seen that sometimes in nonparametric problems, the MLE does not exist. This often happens when there are more "things to estimate" than there are observations. This can also happen in parametric problems. Consider $\mathrm{N}(\mu, \sigma^2)$ with one observation. The likelihood function is

$$-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x-\mu)^2}{\sigma^2}.$$

It is clear that no MLE exists because the function becomes unbounded as σ^2 tends to zero and μ is fixed.

In this case, some people prefer to say that the *likelihood function does not exist*; that is, they suggest that the definition of a likelihood function include boundedness.

There are other interesting examples in which MLEs do not have desirable (or expected) properties. For example,

- an MLE may be discontinuous
- an MLE may not be a function of a sufficient statistic (if the MLE is not unique)
- an MLE may not satisfy the likelihood equation
- the likelihood equation has a unique root, but no MLE exists

- an MLE may not be a MOM estimator; in particular an MLE of the population mean may not be the sample mean.

Although the MLE approach is usually an intuitively logical one, it is not based on a formal decision theory, so it is not surprising that MLEs may not possess certain desirable properties.

5.1.3 MLE and the Exponential Class

If X has a distribution in the exponential class and we write its density in the natural or canonical form, the likelihood has the form

$$L(\eta\,;\,x) = \exp(\eta^{\mathrm{T}}T(x) - \zeta(\eta))h(x).$$

The log-likelihood equation is particularly simple:

$$T(x) - \frac{\partial\zeta(\eta)}{\partial\eta} = 0.$$

Newton's method for solving the likelihood equation is

$$\eta^{(k)} = \eta^{(k-1)} - \left(\frac{\partial^2\zeta(\eta)}{\partial\eta(\partial\eta)^{\mathrm{T}}}\,|_{\eta=\eta^{(k-1)}}\right)^{-1}\left(T(x) - \frac{\partial\zeta(\eta)}{\partial\eta}\,|_{\eta=\eta^{(k-1)}}\right)$$

Note that the second term includes the Fisher information matrix for η. (The expectation is constant.) (Note that the FI is not for a *distribution*; it is for a *parametrization*.)

We have

$$\mathrm{V}(T(X)) = \frac{\partial^2\zeta(\eta)}{\partial\eta(\partial\eta)^{\mathrm{T}}}\,|_{\eta=\eta}\,.$$

Note $\eta = \eta$ (true).

If we have a full-rank member of the exponential class then V is positive definite, and hence there is a unique maximum.

If we write

$$\mu(\eta) = \frac{\partial\zeta(\eta)}{\partial\eta},$$

in the full-rank case, μ^{-1} exists and so we have the solution to the likelihood equation:

$$\hat{\eta} = \mu^{-1}(T(x)).$$

So MLE is very nice for the exponential class.

We also see that MLE and LSE are equivalent for normal distributions. In particular, in the linear model $X = Z\beta + \epsilon$ with normal errors $\hat{\beta}$ is LSE and MLE.

5.1.4 Variations on the Likelihood

There are situations in which a likelihood equation either cannot be written or else it is not solvable. This may happen because of too many parameters, for example. In such cases an approximate likelihood equation may be more appropriate. In other cases, there may be a nuisance parameter that complicates the computation of the MLE for the parameter of interest. In both kind of these situations, we use approximate likelihood methods.

In a multiparameter case, $\theta = (\theta_1, \theta_2)$, we may be interested in only some of the parameters, or in some function of the parameters, perhaps a transformation into a lower-dimensional space. There are various ways of approaching this.

Profile Likelihood

If $\theta = (\theta_1, \theta_2)$ and our interest is only in θ_1, the simplest way of handling this is just to consider θ_2 to be fixed, perhaps at several different values, one at a time. If θ_2 is fixed, the likelihood $L(\theta_1\,;\,\theta_2, x)$ is called a *profile likelihood* or *concentrated likelihood* of θ_1 for given θ_2 and x.

In some cases, it turns out that the estimation of a subset of the parameters does not depend on the value of some other subset. A good method of estimation of β in a linear model $X = Z\beta + \epsilon$ where the residuals ϵ have a common variance σ^2 and zero correlation can be performed equally well no matter what the value of σ^2 is. (The Gauss-Markov theorem tells us that the least-squares method yields a good estimator.) If the residuals are independently distributed as normals with a common variance, we can formulate the problem as a problem in maximum likelihood estimation. The MLE for β (which just happens to be the same as the LSE) can be thought of in terms of a profile likelihood, because a particular value of σ^2 could be chosen a priori. (This is of course not necessary because the maximum or the likelihood with respect to β occurs at the same point regardless of the value of σ^2.)

Induced Likelihood

We can approach the problem of estimation of a function of the parameters by means of an *induced likelihood*. Given a family of PDFs with respect to a common σ-finite measure that are indexed by a parameter $\theta \in \Theta \subset \mathbb{R}^k$. Let $L(\theta\,;\,x)$ be the likelihood of θ given data x. Now let h be a Borel function from Θ onto $\Lambda \subset \mathbb{R}^p$, with $1 \leq p \leq k$. Let

$$\widetilde{L}(\lambda\,;\,x) = \sup_{\theta:h(\theta)=\lambda} L(\theta\,;\,x).$$

Then $\widetilde{L}(\lambda\,;\,x)$ is called the induced likelihood for the transformed parameter $\lambda \in \Lambda$. It turns out that if $\hat{\theta}$ is an MLE of θ then $\hat{\lambda} = h(\hat{\theta})$ maximizes $\widetilde{L}(\lambda\,;\,x)$. (This is Exercise 4.6.95 in Shao.)

As we have noted before, we also speak of an MLE of a function of a parameter without forming an induced likelihood function. If $\hat{\theta}$ is an MLE of θ, and g is a Borel function, then $g(\hat{\theta})$ is called an MLE of the estimand $g(\theta)$, even if does not formally maximize a likelihood in $g(\theta)$.

Conditional Likelihood

When there is a nuisance parameter for which we have a sufficient statistic, a simple approach is to use the PDF conditional on the sufficient statistic to form the likelihood function for the parameter of interest. After doing this, the MLE procedure continues as in the usual case. If the PDFs can be factored so that one factor includes θ_2 and some function of the sample, $S(x)$, and the other factor, given $S(x)$, is free of θ_2, then this factorization can be carried into the likelihood. Such a likelihood is called a *conditional likelihood* of θ_1 given $S(x)$.

Conditional likelihood methods often arise in applications in which the parameters of two different distributions are to be compared; that is, when only their relative values are of interest. Suppose $\mu = (\mu_1, \mu_2)$ and let $\theta_1 = \mu_1/\mu_2$. Although our interest is in θ_1, we may not be able to write the lieklihood as a function of θ_1. If, however, we can find θ_2 for which we have a sufficient statistic, $T_2(X)$, and we can factor the likelihood using the factorization theorem so that the factor corresponding to conditional distribution of X given $T_2(X)$ does not depend on θ_2. This factor, as a function of θ_1, is the conditional likelihood function.

Sometimes a profile likelihood can be thought of as a particularly simple conditional likelihood. The linear model estimation problem referred to above could be formulated as a conditional likelihood. The actual form of the likelihood would be more complicated, but the solution is equivalent to the solution in which we think of the likelihood as a profile likelihood.

Conditional Likelihood for the Exponential Class

If X has a distribution in the exponential class with $\theta = (\eta_1, \eta_2)$, and its likelihood can be written in the form

$$L(\theta\,;\,x) = \exp(\eta_1^{\mathrm{T}} T_1(x) + \eta_2^{\mathrm{T}} T_2(x) - \zeta(\eta_1, \eta_2))h(x),$$

or, in the log-likelihood form,

$$l_L(\theta\,;\,x) = \eta_1^{\mathrm{T}} T_1(x) + \eta_2^{\mathrm{T}} T_2(x) - \zeta(\eta_1, \eta_2) + c(x),$$

we can easily write the conditional log-likelihood:

$$l_L(\eta_1\,;\,x\,;\,T_2) = \eta_1^{\mathrm{T}} T_1(x) + \tilde{\zeta}(\eta_1, T_2) + c(x).$$

Notice that this decomposition can be achieved iff η_1 is a linear function of θ.

If our interest is only in η_1, we only determine the argument that maximizes the function

$$\eta_1^{\mathrm{T}} T_1(x) + \tilde{\zeta}(\eta_1, T_2),$$

which is does not depend on η_2.

Quasi-likelihood Methods

Another way we deal with nuisance parameters in maximum likelihood estimation is by making some simplifying approximations. One type of simplification is to reduce the dimensionality of the nuisance parameters by assuming some relationship among them. This yields a "quasi-likelihood" function. This may allow us to solve what otherwise might be a very difficult problem. In some cases it may not affect the MLE for the parameters of interest. A common application in which quasi-likelihood methods are useful is in estimation of parameters in a generalized linear model.

5.2 EM Methods

Although EM methods do not rely on missing data, they can be explained most easily in terms of a random sample that consists of two components, one observed and one unobserved or missing.

Missing Data

A simple example of missing data occurs in life-testing, when, for example, a number of electrical units are switched on and the time when each fails is recorded.

In such an experiment, it is usually necessary to curtail the recordings prior to the failure of all units.

The failure times of the units still working are unobserved, but the number of censored observations and the time of the censoring obviously provide information about the distribution of the failure times.

Mixtures

Another common example that motivates the EM algorithm is a finite mixture model.

Each observation comes from an unknown one of an assumed set of distributions. The missing data is the distribution indicator.

The parameters of the distributions are to be estimated. As a side benefit, the class membership indicator is estimated.

Applications of EM Methods

The missing data can be missing observations on the same random variable that yields the observed sample, as in the case of the censoring example; or the missing data can be from a different random variable that is related somehow to the random variable observed.

Many common applications of EM methods involve missing-data problems, but this is not necessary.

Often, an EM method can be constructed based on an artificial "missing" random variable to supplement the observable data.

Example 1

One of the simplest examples of the EM method was given by Dempster, Laird, and Rubin (1977).

Consider the multinomial distribution with four outcomes, that is, the multinomial with probability function,

$$p(x_1, x_2, x_3, x_4) = \frac{n!}{x_1!x_2!x_3!x_4!}\pi_1^{x_1}\pi_2^{x_2}\pi_3^{x_3}\pi_4^{x_4},$$

with $n = x_1+x_2+x_3+x_4$ and $1 = \pi_1+\pi_2+\pi_3+\pi_4$. Suppose the probabilities are related by a single parameter, θ, with $0 \leq \theta \leq 1$:

$$\begin{aligned}
\pi_1 &= \frac{1}{2} + \frac{1}{4}\theta \\
\pi_2 &= \frac{1}{4} - \frac{1}{4}\theta \\
\pi_3 &= \frac{1}{4} - \frac{1}{4}\theta \\
\pi_4 &= \frac{1}{4}\theta.
\end{aligned}$$

Given an observation (x_1, x_2, x_3, x_4), the log-likelihood function is

$$l(\theta) = x_1 \log(2+\theta) + (x_2 + x_3)\log(1-\theta) + x_4 \log(\theta) + c$$

and

$$\mathrm{d}l(\theta)/\mathrm{d}\theta = \frac{x_1}{2+\theta} - \frac{x_2+x_3}{1-\theta} + \frac{x_4}{\theta}.$$

The objective is to estimate θ.

Dempster, Laird, and Rubin used $n = 197$ and $x = (125, 18, 20, 34)$. (For this simple problem, the MLE of θ can be determined by solving a simple polynonial equation, but let's proceed with an EM formulation.)

To use the EM algorithm on this problem, we can think of a multinomial with five classes, which is formed from the original multinomial by splitting

the first class into two with associated probabilities $1/2$ and $\theta/4$. The original variable x_1 is now the sum of u_1 and u_2. Under this reformulation, we now have a maximum likelihood estimate of θ by considering $u_2 + x_4$ (or $x_2 + x_3$) to be a realization of a binomial with $n = u_2 + x_4 + x_2 + x_3$ and $\pi = \theta$ (or $1 - \theta$). However, we do not know u_2 (or u_1). Proceeding as if we had a five-outcome multinomial observation with two missing elements, we have the log-likelihood for the complete data,

$$l_c(\theta) = (u_2 + x_4)\log(\theta) + (x_2 + x_3)\log(1 - \theta),$$

and the maximum likelihood estimate for θ is

$$\frac{u_2 + x_4}{u_2 + x_2 + x_3 + x_4}.$$

The E-step of the iterative EM algorithm fills in the missing or unobservable value with its expected value given a current value of the parameter, $\theta^{(k)}$, and the observed data. Because $l_c(\theta)$ is linear in the data, we have

$$\mathrm{E}\left(l_c(\theta)\right) = \mathrm{E}(u_2 + x_4)\log(\theta) + \mathrm{E}(x_2 + x_3)\log(1 - \theta).$$

Under this setup, with $\theta = \theta^{(k)}$,

$$\begin{aligned}\mathrm{E}_{\theta^{(k)}}(u_2) &= \frac{1}{4}x_1\theta^{(k)} \,/\, \left(\frac{1}{2} + \frac{1}{4}x_1\theta^{(k)}\right) \\ &= u_2^{(k)}.\end{aligned}$$

We now maximize $\mathrm{E}_{\theta^{(k)}}\left(l_c(\theta)\right)$. This maximum occurs at

$$\theta^{(k+1)} = (u_2^{(k)} + x_4)/(u_2^{(k)} + x_2 + x_3 + x_4).$$

The following Matlab statements execute a single iteration.

```
function [u2kp1,tkp1] = em(tk,x)
u2kp1 = x(1)*tk/(2+tk);
tkp1 = (u2kp1 + x(4))/(sum(x)-x(1)+u2kp1);
```

Example 2: A Variation of the Life-Testing Experiment Using an Exponential Model

Consider an experiment described by Flury and Zoppè (2000). It is assumed that the lifetime of light bulbs follows an exponential distribution with mean θ. To estimate θ, n light bulbs were tested until they all failed. Their failure times were recorded as $x_1, \ldots, x_n$. In a separate experiment, m bulbs were tested, but the individual failure times were not recorded. Only the number of bulbs, r, that had failed at time t was recorded.

The missing data are the failure times of the bulbs in the second experiment, $u_1, \ldots, u_m$. We have

$$l_c(\theta\ ;\ x, u) = -n(\log\theta + \bar{x}/\theta) - \sum_{i=1}^{m}(\log\theta + u_i/\theta).$$

The expected value for a bulb still burning is

$$t + \theta$$

and the expected value of one that has burned out is

$$\theta - \frac{t\mathrm{e}^{-t/\theta^{(k)}}}{1 - \mathrm{e}^{-t/\theta^{(k)}}}.$$

Therefore, using a provisional value $\theta^{(k)}$, and the fact that r out of m bulbs have burned out, we have $\mathrm{E}_{U|x,\theta^{(k)}}(l_c)$ as

$$\begin{aligned} q^{(k)}(x,\theta) = &-(n+m)\log\theta \\ &-\frac{1}{\theta}\left(n\bar{x} + (m-r)(t+\theta^{(k)}) + r(\theta^{(k)} - th^{(k)})\right), \end{aligned}$$

where $h^{(k)}$ is given by

$$h^{(k)} = \frac{\mathrm{e}^{-t/\theta^{(k)}}}{1 - \mathrm{e}^{-t/\theta^{(k)}}}.$$

The k^{th} M step determines the maximum with respect to the variable θ, which, given $\theta^{(k)}$, occurs at

$$\theta^{(k+1)} = \frac{1}{n+m}\left(n\bar{x} + (m-r)(t+\theta^{(k)}) + r(\theta^{(k)} - th^{(k)})\right). \tag{5.10}$$

Starting with a positive number $\theta^{(0)}$, equation (5.10) is iterated until convergence. The expectation $q^{(k)}$ does not need to be updated explicitly.

To see how this works, let's generate some artificial data and try it out. Some R code to implement this is:

```
# Generate data from an exponential with theta=2,
# and with the second experiment truncated at t=3.
# Note that R uses a form of the exponential in
# which the parameter is a multiplier; i.e., the R
# parameter is 1/theta.
# Set the seed, so computations are reproducible.
set.seed(4)
n <- 100
m <- 500
theta <- 2
t <- 3
x <- rexp(n,1/theta)
r<-min(which(sort(rexp(m,1/theta))>=3))-1
```

Some R code to implement the EM algorithm:

```
# We begin with  theta=1.
# (Note theta.k is set to theta.kp1 at
#  the beginning of the loop.)
theta.k<-.01
theta.kp1<-1
# Do some preliminary computations.
n.xbar<-sum(x)
# Then loop and test for convergence
   theta.k <- theta.kp1
   theta.kp1 <- (n.xbar +
                 (m-r)*(t+theta.k) +
                 r*(theta.k-
                     t*exp(-t/theta.k)/(1-exp(-t/theta.k))
                   )
                )/(n+m)
```

The value of θ stabilizes to less than 0.1% change at 1.912 in 6 iterations.

This example is interesting because if we assume that the distribution of the light bulbs is uniform, $\mathrm{U}(0,\theta)$ (such bulbs are called "heavybulbs"!), the EM algorithm cannot be applied.

Maximum likelihood methods must be used with some care whenever the range of the distribution depends on the parameter.

In this case, however, there is another problem. It is in computing $q^{(k)}(x,\theta)$, which does not exist for $\theta < \theta^{(k-1)}$.

Example 3: Estimation in a Normal Mixture Model

A two-component normal mixture model can be defined by two normal distributions, $\mathrm{N}(\mu_1,\sigma_1^2)$ and $\mathrm{N}(\mu_2,\sigma_2^2)$, and the probability that the random variable (the observable) arises from the first distribution is w.

The parameter in this model is the vector $\theta = (w,\mu_1,\sigma_1^2,\mu_2,\sigma_2^2)$. (Note that w and the σs have the obvious constraints.)

The pdf of the mixture is

$$p(y;\theta) = wp_1(y;\mu_1,\sigma_1^2) + (1-w)p_2(y;\mu_2,\sigma_2^2),$$

where $p_j(y;\mu_j,\sigma_j^2)$ is the normal pdf with parameters μ_j and σ_j^2. (I am just writing them this way for convenience; p_1 and p_2 are actually the same parametrized function of course.)

In the standard formulation with $C = (X,U)$, X represents the observed data, and the unobserved U represents class membership.

Let $U = 1$ if the observation is from the first distribution and $U = 0$ if the observation is from the second distribution.

The unconditional $\mathrm{E}(U)$ is the probability that an observation comes from the first distribution, which of course is w.

Suppose we have n observations on X, $x_1, \ldots, x_n$.

Given a provisional value of θ, we can compute the conditional expected value $\mathrm{E}(U|x)$ for any realization of X. It is merely

$$\mathrm{E}(U|x, \theta^{(k)}) = \frac{w^{(k)} p_1(x; \mu_1^{(k)}, \sigma_1^{2^{(k)}})}{p(x; w^{(k)}, \mu_1^{(k)}, \sigma_1^{2^{(k)}}, \mu_2^{(k)}, \sigma_2^{2^{(k)}})}$$

The M step is just the familiar MLE of the parameters:

$$w^{(k+1)} = \frac{1}{n} \sum \mathrm{E}(U|x_i, \theta^{(k)})$$

$$\mu_1^{(k+1)} = \frac{1}{n w^{(k+1)}} \sum q^{(k)}(x_i, \theta^{(k)}) x_i$$

$$\sigma_1^{2^{(k+1)}} = \frac{1}{n w^{(k+1)}} \sum q^{(k)}(x_i, \theta^{(k)})(x_i - \mu_1^{(k+1)})^2$$

$$\mu_2^{(k+1)} = \frac{1}{n(1 - w^{(k+1)})} \sum q^{(k)}(x_i, \theta^{(k)}) x_i$$

$$\sigma_2^{2^{(k+1)}} = \frac{1}{n(1 - w^{(k+1)})} \sum q^{(k)}(x_i, \theta^{(k)})(x_i - \mu_2^{(k+1)})^2$$

(Recall that the MLE of σ^2 has a divisor of n, rather than $n - 1$.)

To see how this works, let's generate some artificial data and try it out. Some R code to implement this is:

```
# Normal mixture.  Generate data from nomal mixture with w=0.7,
# mu_1=0, sigma^2_1=1,  mu_2=1, sigma^2_2=2.
# Note that R uses sigma, rather than sigma^2 in rnorm.
#  Set the seed, so computations are reproducible.
set.seed(4)
n <- 300
w <- 0.7
mu1 <- 0
sigma21 <- 1
mu2 <- 5
sigma22 <- 2
x <- ifelse(runif(n)<w,
rnorm(n,mu1,sqrt(sigma21)),rnorm(n,mu2,sqrt(sigma22)))
```

First, assume that μ_1, σ_1^2, μ_2, and σ_2^2 are all known:

```
# Initialize.
theta.k<-.1
theta.kp1<-.5
```

```
# Then loop over the following
   theta.k <- theta.kp1
   tmp <- theta.k*dnorm(x, mu1,sqrt(sigma21))
   ehat.k <- tmp/(tmp+(1-theta.k)*dnorm(x, mu2,sqrt(sigma22)))
   theta.kp1<- mean(ehat.k)
```

This converges very quickly to 0.682, at which point the parameter estimate changes less than 0.1%.

5.3 Asymptotic Properties of MLEs, RLEs, and GEE Estimators

The argmax of the likelihood function, that is, the MLE of the argument of the likelihood function, is obviously an important statistic.

In many cases, a likelihood equation exists, and often in those cases, the MLE is a root of the likelihood equation. In some cases there are roots of the likelihood equation (RLEs) that may or may not be an MLE.

5.3.1 Asymptotic Efficiency of MLEs and RLEs

One of the most important properties of roots of the likelihood equation, given the Le Cam regularity conditions (see page 60), is asymptotic efficiency. The regularity conditions are the same as those for Le Cam's theorem on the countability of superefficient estimators (Shao's theorem 4.16).

There are two parts to Shao's Theorem 4.17.

The first says that there is a sequence of estimators $\hat{\theta}_n$ such that

$$\Pr(s_n(\hat{\theta}_n) = 0) \to 1,$$

where $s_n(\hat{\theta}_n)$ is the score function $s_n(\gamma) = \partial L(\gamma)/\partial\gamma$ evaluated at $\hat{\theta}_n$, and

$$\hat{\theta}_n \to_{\mathrm{p}} \theta.$$

The second part says that any consistent sequence of RLEs is asymptotically efficient.

5.3.2 Examples

Let's consider a couple of examples from Shao.

The E(α, θ) Distribution

We have a random sample $X_1, \ldots, X_n$ from E(α, θ), with α and θ unknown. The likelihood function is

$$L(\alpha, \theta; X) = \theta^{-n} \exp\left(-\frac{1}{\theta}\sum(X_i - \alpha)\right) \mathrm{I}_{(0,X_{(1)}]}(\alpha)\mathrm{I}_{(0,\infty)}(\theta).$$

This is 0 when $\alpha > X_{(1)}$, but it is increasing in α on $(0, X_{(1)}]$ independently of θ. Hence, the MLE of α is $X_{(1)}$.

Now, we substitute this back into $L(\alpha, \theta; X)$ and maximize w.r.t. θ, i.e. solve

$$\max_{\theta}\left(\theta^{-n} \exp\left(-\frac{1}{\theta}\sum(X_i - X_{(1)})\right)\right).$$

We do this by forming and solving the likelihood equation, noting that it yields a maximum within the parameter space. We get

$$\hat{\theta} = \frac{1}{n}\sum(X_i - X_{(1)}).$$

In Exercise 3.6 (one of the practice exercises assigned last semester) we found the UMVUEs:

$$T_\alpha = X_{(1)} - \frac{1}{n(n-1)}\sum(X_i - X_{(1)}).$$

and

$$T_\theta = \frac{1}{n-1}\sum(X_i - X_{(1)}).$$

(Recall that we find a complete sufficient statistic and then manipulate it to be unbiased.) Notice the similarity of these to the MLEs, which are biased.

Now let's consider the ARE of the MLE to the UMVUE for these two parameters. (Remember that the ARE is the ratio of two asymptotic expectations — not the asymptotic expectation of a ratio, and certainly not the limit of a ratio; although of course sometimes these three things are the same.)

- ARE(MLE,UMVUE) for θ.
  This is an easy case, because the estimators always differ by the ratio $n/(n-1)$; hence the ARE is 1.
  The distributions for $\hat{\theta}$ and T_θ are relatively easy. From Exercise 1.78, we have that the distribution of $\sum X_i$ is $\Gamma(n, \theta)$ if $X_i \sim \mathrm{E}(0, \theta)$, hence for T_θ above, if we let $Y = 2(n-1)T_\theta/\theta$, we have $Y \sim \chi^2_{2(n-1)}$.
- ARE(MLE,UMVUE) for α.
  We must work out the asymptotic expectations of U^2 and V^2 where $U = \hat{\alpha} - \alpha$ and $V = T_\alpha - \alpha$. We get immediately that $nU = n(\hat{\alpha} - \alpha) = n(X_{(1)} - \alpha)$ has a E$(0, \theta)$ distribution. Now

$$nV = n(X_{(1)} - \alpha) - \frac{1}{n(n-1)}\sum(X_i - X_{(1)}),$$

and because $\frac{1}{n(n-1)}\sum(X_i - X_{(1)}) \to_{\mathrm{p}} \theta$, we have $nV \to_{\mathrm{d}} W - \theta$, where $W \sim \mathrm{E}(0, \theta)$. Therefore, the ARE, which is $\mathrm{E}(V^2)/\mathrm{E}(U^2)$ is $\theta/(\theta + \theta^2)$.

The Bernoulli Distribution

Consider a Bernoulli distribution with unknown $\pi \in (0, 1)$, and suppose we are interested in estimating $g(\pi) = \pi(1-\pi)$. The MLE of $g(\pi)$ is $T_n = \overline{X}(1-\overline{X})$. (Why?)

Now, let's look at its asymptotic distributions.

From the central limit theorem, $\sqrt{n}(\overline{X} - \pi) \to \mathrm{N}(0, g(\pi))$.

If $\pi \neq 1/2$, $g'(\pi) \neq 0$, we can use the delta method and the CLT to get

$$\sqrt{n}(g(\pi) - T_n) \to \mathrm{N}(0, \pi(1-\pi)(1-2\pi)^2).$$

(I have written $(g(\pi) - T_n)$ instead of $(T_n - g(\pi))$ so the expression looks more similar to one we get next.) If $\pi = 1/2$, this is a degenerate distribution. (The limiting variance actually is 0, but the degenerate distribution is not very useful.)

Let's take a different approach for the case $\pi = 1/2$. We have from the CLT, $\sqrt{n}(\overline{X} - \frac{1}{2}) \to \mathrm{N}(0, \frac{1}{4})$. Hence, if we scale and square, we get $4n(\overline{X} - \frac{1}{2})^2 \to_{\mathrm{d}} \chi^2_1$, or

$$4n(g(\pi) - T_n) \to_{\mathrm{d}} \chi^2_1.$$

This is a specific instance of a *second order delta method*. Following the same methods using a Taylor series expansion as in Section 1.2, for the univariate case, we can see that

$$\left.\begin{array}{c} \sqrt{n}(S_n - c) \to \mathrm{N}(0, \sigma^2) \\ g'(c) = 0 \\ g''(c) \neq 0 \end{array}\right\} \Longrightarrow 2n\frac{(g(S_n) - g(c))}{\sigma^2 g''(c)} \to_{\mathrm{d}} \chi^2_1. \tag{5.11}$$

5.3.3 Inconsistent MLEs

In previous sections, we have seen that sometimes MLEs do not have some statistical properties that we usually expect of good estimators.

The discussion in this section has focused on MLEs (or RLEs) that are consistent. Even for MLEs, however, they may not be consistent.

Rational, Irrational Estimand

Consider an example from Romano and Siegel:
Let $X_1, \ldots, X_n$ be a sample from $\mathrm{N}(\theta, 1)$. Define the estimand $g(\theta)$ as

$$g(\theta) = \begin{cases} -\theta \text{ if } \theta \text{ is irrational} \\ \theta \text{ if } \theta \text{ is rational.} \end{cases}$$

Because $\overline{X}_n$ is the MLE of θ, $g(\overline{X}_n)$ is the MLE of $g(\theta)$. Now $\overline{X}_n \sim \mathrm{N}(\theta, 1/n)$ and so is almost surely irrational; hence, $g(\overline{X}_n) = -\overline{X}_n$ a.s. Now, by the SLLN, we have $g(\overline{X}_n) = -\theta$ a.s. Hence, if θ is a rational number $\neq 0$, then

$$g(\overline{X}_n) \to_{\text{a.s.}} -\theta \neq \theta = g(\theta).$$

Mixture

While that example may seem somewhat contrived, consider an example due to Ferguson:
Let $X_1, \ldots, X_n$ be a sample from from the distribution with PDF w.r.t. Lebesgue measure

$$p_X(x;\theta) = (1-\theta)p_{\mathrm{T}}(x;\theta,\delta(\theta)) + \theta p_{\mathrm{U}}(x),$$

where $\theta \in [0,1]$, $\delta(\theta)$ is a continuous decreasing function of θ with $\delta(0) = 1$ and $0 < \delta(\theta) \leq 1-\theta$ for $0 < \theta < 1$, and

$$p_{\mathrm{T}}(x;\theta,\delta(\theta)) = \frac{1}{\delta(\theta)}\left(1 - \frac{|x-\theta|}{\delta(\theta)}\right)\mathrm{I}_{[\theta-\delta(\theta),\theta+\delta(\theta)]}(x)$$

and

$$p_{\mathrm{U}}(x) = \frac{1}{2}\mathrm{I}_{[-1,1]}(x).$$

The distribution is a mixture of a triangular distribution centered on θ and the $\mathrm{U}(-1,1)$ distribution.

Note that the densities are continuous in θ for any x and is defined on $[0,1]$ and therefore an MLE exists.

Let $\hat{\theta}_n = \hat{\theta}(X_1, \ldots, X_n)$ denote any MLE of θ. Now, if $\theta < 1$, then

$$p_X(x;\theta) \leq (1-\theta)/\delta(\theta) + \theta/2 < 1/\delta(\theta) + \frac{1}{2},$$

and so for any $\alpha < 1$

$$\max_{0\leq\theta\leq\alpha} \frac{l_n(\theta)}{n} \leq \log\left(\frac{1}{\delta(\theta)} + \frac{1}{2}\right) < \infty.$$

Now, if we could choose $\delta(\theta)$ so that

$$\max_{0\leq\theta\leq 1} \frac{l_n(\theta)}{n} \to_{\text{a.s.}} \infty,$$

then $\hat{\theta}_n$ will eventually be greater than α for any $\alpha < 1$, and so the MLE is not consistent.

So, can we choose such a $\delta(\theta)$?

Let

$$M_n = \max(X_1, \ldots, X_n),$$

hence $M_n \to_{\text{a.s.}} \infty$, and

$$\begin{aligned}\max_{0\le\theta\le1} \frac{l_n(\theta)}{n} &\ge \frac{l_n(M_n)}{n} \\ &\ge \frac{n-1}{n}\log\left(\frac{M_n}{2}\right) + \frac{1}{n}\log\left(\frac{1-M_n}{\delta(M_n)}\right),\end{aligned}$$

and so

$$\liminf_n \max_{0\le\theta\le1} \frac{l_n(\theta)}{n} \ge \log\left(\frac{1}{2}\right) + \liminf_n \log\left(\frac{1-M_n}{\delta(M_n)}\right) \text{ a.s.}$$

So we need to choose $\delta(\theta)$ so that the last limit is infinite a.s. Now, $\forall\theta M_n \to_{\text{a.s.}} \infty$, and the slowest rate is for $\theta = 1$, because that distribution has the smallest mass in a sufficiently small neighborhood of 1. Therefore, all we need to do is choose $\delta(\theta) \to 0$ as $\theta \to 1$ fast enough so that the limit is infinite a.s. when $\theta = 0$.

So now for $0 < \epsilon < 1$,

$$\begin{aligned}\sum_n \Pr_{\theta=0}(n^{1/4}(1-M_n) > \epsilon) &= \sum_n \Pr_{\theta=0}(M_n < 1 - \epsilon n^{-1/4}) \\ &= \sum_n \left(1 - \epsilon^2\frac{n^{-1/4}}{2}\right)^n \\ &\le \sum_n \exp\left(-\epsilon^2\frac{n^{-1/4}}{2}\right) \\ &< \infty.\end{aligned}$$

Hence, by the Borel-Cantelli lemma, $n^{1/4}(1-M_n) \to 0$ a.s. Finally, choosing

$$\delta(\theta) = (1-\theta)\exp\left(-(1-\theta)^{-4}+1\right),$$

we have a function that satisfies the requirements above (it is continuous decreasing with $\delta(0) = 1$ and $0 < \delta(\theta) \le 1-\theta$ for $0 < \theta < 1$) and it is such that

$$\begin{aligned}\frac{1}{n}\log\left(\frac{1-M_n}{\delta(M_n)}\right) &= \frac{1}{n(1-M_n)^4} - \frac{1}{n} \\ &\to_{\text{a.s.}} \infty.\end{aligned}$$

This says that *any* MLE must tend to 1 a.s.

Consistency of GEE Estimators

5.3.4 Asymptotic Normality of GEE Estimators

The class of estimators arising from the generalized estimating equations (2.25) and (2.26), under very general assumptions have an asymptotic normal distribution. This is Theorem 5.13 in Shao.

$$\sqrt{n}(\hat{\theta}_n - \theta) \to_{\mathrm{d}} \mathrm{N}(0, \sigma_F^2),$$

where $\{\hat{\theta}_n\}$ is a sequence of GEE estimators and

$$\sigma_F^2 = \int (\psi(x,\theta))^2 \mathrm{d}F(x)/(\psi'(x,\theta))^2.$$

5.4 Application: Maximimum Likelihood Estimation in Generalized Linear Models

In a very useful model for statistical applications, with observed data X as an n-vector and Z as an $n \times p$ matrix, we have the relationship of the form

$$X = f(Z;\theta) + \epsilon,$$

where θ is a vector of *parameters* used in the specification of the function f, and ϵ is a deviation, usually assumed to be a random variable.

The expression "$f(\cdot)$" represents a systematic effect related to the values of "Z", and "ϵ" represents a random effect, an unexplained effect, or simply a "residual" that is added to the systematic effect.

A model in which the parameters are additively separable and with an additive random effect is sometimes called an additive model:

$$X = f(Z)\theta + \epsilon.$$

A simple version of this is called a linear (additive) model:

$$X = Z\beta + \epsilon, \tag{5.12}$$

where β is a p-vector of parameters.

Either form of the additive model can be generalized with a "link function" to be a *generalized additive model*.

In the following, we will concentrate on the linear model, $X = Z\beta + \epsilon$, and we will discuss the link function and the generalization of the linear model, which is called a generalized linear model (GLM or GLIM).

5.4.1 Linear Models

First, consider the random component in the model, and assume that the elements are independent.

Let us assume that the distribution of the residual has a first moment and that it is known. In that case, we can take its mean to be 0, otherwise, we can incorporate it into $Z\beta$. (If the first moment does not exist, we can work with

the median.) Hence, assuming the mean of the residual exists, the model can be written as

$$\mathrm{E}(X) = Z\beta,$$

that is, the expected value of X is the systematic effect in the model. More generally, we can think of the model as being a location family with PDF

$$p_\epsilon(\epsilon) = p_\epsilon(x + Z\beta),$$

w.r.t. a given σ-finite measure.

The Exponential Class

Let us now assume the happy case in which each X_i is distributed as a member of the exponential class. (We are also assuming that they are independent.) In the canonical formulation of the exponential form,

$$\exp\left(\eta_i^{\mathrm{T}} T(x_i) - \zeta(\eta_i)\right) h(x_i), \tag{5.13}$$

we have seen that the MLE of the natural parameter η_i has a particularly simple form; it is just

$$\left(\frac{\partial \zeta}{\partial \eta_i}\right)^{-1}(\eta_i)\,\big|_{\eta_i = T(x_i)}\,.$$

In the linear model (5.12), if $\epsilon \sim \mathrm{N}(0, \sigma^2)$, as we usually assume, we can easily identify η_i, $T(x_i)$, and $\zeta(\eta_i)$ in equation (5.13), and of course, $h(x_i) \equiv 1$. This is a location-scale family.

Location-Scale Family with the Scale a Nuisance Parameter

We can also consider a useful family of distributions in which the scale parameter is a nuisance parameter. We extend the distributional family of equation (5.13) slightly as a scale family with scale parameter $\phi_i > 0$, but we will not restrict the distribution to be in the exponential class if ϕ_i is unknown.

We now have the probability density of X_i as

$$p_{X_i}(x_i|\eta_i, \phi_i) = \exp\left(\frac{\eta_i^{\mathrm{T}} T(x_i) - \zeta(\eta_i)}{\phi_i}\right) h(x_i, \phi_i), \tag{5.14}$$

where ζ and h are known functions and w.r.t. the same σ-finite measure. The scale ϕ_i is a nuisance parameter. (Notice, we could have made this slightly more general, by considering the scale to be $\xi(\phi_i)$, some more function of the basic parameter ϕ_i.)

The Likelihood

The objective is to fit the model, that is, to estimate the parameters. The model parameters are usually estimated either by a maximum likelihood method or by minimizing some function of the residuals.

The likelihood, as a function of the parameters given realizations of the random variable, is

$$L(\eta_i, \phi_i|x_i) = \exp\left\{\frac{x_i\eta_i - \zeta(\eta_i)}{\phi_i} + c(x_i, \phi_i)\right\}.$$

As usual, we let

$$l(\eta_i, \phi_i|x_i) = \log(L(\eta_i, \phi_i|x_i)).$$

Moments of $\boldsymbol{X}$

Note, that if the FI regularity conditions are satisfied in the distribution with PDF (5.14) where ϕ_i is known and θ is a function of η_i and ϕ_i, then

$$\frac{\partial}{\partial\theta}\mathrm{E}(l) = \mathrm{E}\left(\frac{\partial l}{\partial\theta}\right),$$

and so

$$\mathrm{E}\left(\frac{\partial l}{\partial\theta}\right) = 0,$$

and

$$\mathrm{E}\left(\frac{\partial^2 l}{\partial\theta^2}\right) + \mathrm{E}\left(\frac{\partial l}{\partial\theta}\right)^2 = 0.$$

Hence, for a random variable X_i with this density,

$$\mathrm{E}(X) = \frac{d}{d\eta_i}\zeta(\eta_i)$$

and

$$\mathrm{V}(X) = \frac{d^2}{d\eta_i^2}\zeta(\eta_i)\phi_i.$$

So, in terms of the individual observations, it is convenient to write

$$\mu_i = \mathrm{E}(X_i) = \frac{d}{d\eta_i}\zeta(\eta_i).$$

5.4.2 Generalized Linear Models

A model as in equation (5.12) has limitations. Suppose, for example, that we are interested in modeling a response that is binary, for example, two states of a medical patient, "diseased" or "disease-free". As usual, we set up a random variable to map the sample space to $\mathbb{R}$:

$$X : \{\text{disease-free,diseased}\} \mapsto \{0, 1\}.$$

The linear model $X = Z\beta + \epsilon$ does not make sense. It is continuous and unbounded.

A more useful model may address $\Pr(X = 0)$.

To make this more concrete, consider the situation in which several groups of subjects are each administered a given dose of a drug, and the number responding in each group is recorded. The data consist of the counts x_i responding in the i^{th} group, which received a level z_i of the drug.

A basic model is

$$\mathrm{P}(X_i = 0|z_i) = 1 - \pi_i$$

$$\mathrm{P}(X_i = 1|z_i) = \pi_i$$

The question is how does π depend on z?

A linear dependence, $\pi = \beta_0 + \beta_1 z$ does not fit well in this kind of situation – unless we impose restrictions, π would not be between 0 and 1.

We can try a transformation to $[0, 1]$.

Suppose we impose an invertible function on

$$\eta = \beta_0 + \beta_1 z$$

that will map it into $[0, 1]$:

$$\pi = h(\eta),$$

or

$$g(\pi) = \eta.$$

We call this a *link function*.

A common model following this setup is

$$\pi_z = \Phi(\beta_0 + \beta_1 z),$$

where Φ is the normal cumulative distribution function, and β_0 and β_1 are unknown parameters to be estimated. This is called a *probit model*. The link function in this case is Φ^{-1}.

The related *logit model*, in which the log odds ratio $\log(\pi/(1 - \pi))$ is of interest, has as link function

$$\eta = \log\left(\frac{\pi}{1 - \pi}\right).$$

Other possibilities are the complementary log-log function

$$\eta = \log\{-\log(1-\pi)\},$$

and the log-log function,

$$\eta = -\log\{-\log(\pi)\}.$$

Link Functions

The link function relates the systematic component to the mean of the random variable.

In the case of the linear model, let η_i be the systematic component for a given value of the independent variable,

$$\eta_i = \beta_0 + \beta_1 z_{1i} + \cdots \beta_p z_{pi},$$

and let $\mu_i = \mathrm{E}(X)$, as before. Let g be the link function:

$$\eta_i = g(\mu_i).$$

In this case, the link function is linear in a set of parameters, β_j, and it is usually more natural to think in terms of these parameters rather than θ,

$$g(\frac{d}{d\theta}b(\theta_i)) = g(\mu_i) = \eta_i = x_i'\beta.$$

The generalized linear model can now be thought of as consisting of three parts:

1. the systematic component
2. the random component
3. the link between the systematic and random components.

In the context of generalized linear models, a standard linear model has a systematic component of

$$\beta_0 + \beta_1 x_{1i} + \cdots \beta_m x_{mi},$$

a random component that is an identical and independent normal distribution for each observation, and a link function that is the identity.

5.4.3 Fitting Generalized Linear Models

Our initial objective is to fit the model, that is, to determine estimates of the β_j.

The model parameters are usually determined either by a maximum likelihood method or by minimizing some function of the residuals. One approach

is to use the link function and do a least squares fit of η using the residuals $y_i - \mu_i$. It is better, however, to maximize the likelihood or, alternatively, the log-likelihood,

$$l(\theta, \phi|y) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi).$$

The most common method of optimizing this function is "Fisher scoring", which is a method like Newton's method, except that some quantities are replaced by their expected values.

Newton's Method

Problem: find the maximum of $f(x)$.

Minimum: maximum $-f(x)$.

What is $f(\cdot)$ like?

Suppose can expand in Taylor series:

$$f(x) = f(x_0) + (x - x_0)'\nabla f(x_0) + \frac{1}{2!}(x - x_0)'H_f(x_0)(x - x_0) + \cdots$$

Now, suppose $f(\cdot)$ is a quadratic (i.e., no $\cdots$ above).

Suppose its maximum (or minimum) is at x_1. Then $\nabla f(x_1) = 0$, or

$$\nabla\{(x_1 - x_0)'\nabla f(x_0)\} + \nabla\{\frac{1}{2!}(x_1 - x_0)'H_f(x_0)(x_1 - x_0)\} = 0,$$

i.e.,

$$\nabla f(x_0) + H_f(x_0)(x_1 - x_0) = 0.$$

If H_f is nonsingular, we have

$$x_1 = x_0 - H_f^{-1}(x_0)\nabla f(x_0).$$

We can build a sequence of approximations, by expanding f about x_1, and at each stage assuming a quadratic fit.

This is also called Newton-Raphson. (Note we have to evaluate and invert H at each step. Invertible? Positive-definite? How to approximate it?)

Fisher Scoring on the Log-Likelihood

The Hessian in Newton's method is replaced by its expected value. The iterates then are

$$\hat{\theta}_{k+1} = \hat{\theta}_k - H_l^{-1}(\hat{\theta}_k|x)\nabla l(\hat{\theta}_k|x)$$

In the generalized linear model, where the likelihood is linked to the parameters that are really of interest, this still must be cast in terms that will yield values for $\hat{\beta}$.

Analysis of Deviance

Our approach to modeling involves using the observations (including the realizations of the random variables) as fixed values and treating the parameters as variables (not random variables, however). The original model was then encapsulated into a likelihood function, $L(\theta|y)$, and the principle of fitting the model was maximization of the likelihood with respect to the parameters. The log likelihood, $l(\theta|x)$, is usually used.

In model fitting an important issue is how well does the model fit the data? How do we measure the fit? Maybe use residuals. (Remember, some methods of model fitting work this way; they minimize some function of the residuals.) We compare different models by means of the measure of the fit based on the residuals. We make inference about parameters based on changes in the measure of fit.

Using the likelihood approach, we make inference about parameters based on changes in the likelihood. Likelihood ratio tests are based on this principle.

A convenient way of comparing models or making inference about the parameters is with the *deviance function*, which is a likelihood ratio:

$$D(y|\hat{\theta}) = 2[l(\theta_{\max}|y) - l(\hat{\theta}|y)],$$

where $\hat{\theta}$ is the fit of a potential model.

For generalized linear models the analysis of deviance plays a role similar to that of the analysis of sums of squares (analysis of "variance") in linear models.

Under appropriate assumptions, when θ_1 is a subvector of θ_2, the difference in deviances of two models, $D(y|\hat{\theta_2}) - D(y|\hat{\theta_1})$ has an asymptotic chi-squared distribution with degrees of freedom equal to the difference in the number of parameters.

For models with a binary response variable, we need a different measure of residuals. Because we are measuring the model fit in terms of the deviance, D, we may think of the observations as each contributing a quantity d_i, such that $\sum d_i = D$. (Exactly what that value is depends on the form of the systematic component and the link function that are in the likelihood.) The quantity

$$r_i = \text{sign}(y_i - \hat{\mu_i})\sqrt{d_i}$$

increases in $(y_i - \hat{\mu_i})$ and $\sum r_i^2 = D$. We call r_i the *deviance residual*.

For the logit model,

$$r_i = \text{sign}(y_i - \hat{\mu_i})\sqrt{-2[y_i \log(\hat{\pi_i}) + (1 - y_i)\log(1 - \hat{\pi_i})]}$$

5.4.4 Generalized Additive Models

The mechanical process of dealing with generalized additive models parallels that of dealing with generalized linear models. There are some very important

differences, however. The most important is probably that the distribution of the deviances is not worked out.

The meaning of degrees of freedom is also somewhat different.

So, first, we work out an analogous concept for degrees of freedom.

The response variable is Bernoulli (or binomial). We model the log odds ratios as

$$\begin{aligned}\log(\frac{\pi_i}{1-\pi_i}) &= \eta_i \\ &= \beta_0 + \beta_1 x_{1i} + \cdots + \beta_6 x_{6i} \\ &= x_i'\beta.\end{aligned}$$

For a binomial with number m_i, we write the log-likelihood,

$$l(\pi|y) = \sum_{i=1}^{n}\{y_i \log(\pi_i/(1-\pi_i)) + m_i \log(1-\pi_i)\},$$

where a constant involving m_i and y_i has been omitted. Substituting, we have,

$$l(\beta|y) = \sum_{i=1}^{n} y_i x_i'\beta - \sum_{i=1}^{n} m_i \log\{1+\exp(x_i'\beta)\}.$$

The log likelihood depends on y only through $X'y$.

$$\frac{\partial l}{\partial \pi_i} = \frac{y_i - m_i\pi_i}{\pi_i(1-\pi_i)}$$

Using the chain rule, we have

$$\begin{aligned}\frac{\partial l}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{y_i - m_i\pi_i}{\pi_i(1-\pi_i)} \frac{\partial \pi_i}{\partial \beta_j} \\ &= \sum_{i=1}^{n} \frac{y_i - m_i\pi_i}{\pi_i(1-\pi_i)} \frac{d\pi_i}{d\eta_i} x_{ij}\end{aligned}$$

The Fisher information is

$$\begin{aligned}-\mathrm{E}\left(\frac{\partial^2 l}{\partial\beta_j\partial\beta_k}\right) &= \sum_{i=1}^{n} \frac{m_i}{\pi_i(1-\pi_i)} \frac{\partial \pi_i}{\partial \beta_j} \frac{\partial \pi_i}{\partial \beta_k} \\ &= \sum_{i=1}^{n} \frac{m_i(d\pi_i/d\eta_i)^2}{\pi_i(1-\pi_i)} x_{ij}x_{ik} \\ &= (X'WX)_{jk},\end{aligned}$$

where W is a diagonal matrix of weights,

$$\frac{m_i(d\pi_i/d\eta_i)^2}{\pi_i(1-\pi_i)}$$

Notice

$$\frac{d\pi_i}{d\eta_i} = \pi_i(1-\pi_i),$$

so we have the simple expression,

$$\frac{\partial l}{\partial \beta} = X'(y - m\pi)$$

in matrix notation, and for the weights we have,

$$m_i\pi_i(1-\pi_i)$$

Use Newton's method,

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} - H_l^{-1}(\hat{\beta}^{(k)})\nabla l(\hat{\beta}^{(k)}),$$

in which H_l is replaced by

$$\mathrm{E}\left(-\frac{\partial^2 l}{\partial\beta\partial\beta'}.\right)$$

Using $\hat{\beta}^{(k)}$, we form $\hat{\pi}^{(k)}$ and $\hat{\eta}^{(k)}$, and then, an adjusted $y^{(k)}$,

$$y_i^{(k)} = \hat{\eta}^{(k)} + \frac{(y - m_i\hat{\pi}_i^{(k)})}{m_i}\frac{d\eta_i}{d\pi_i}$$

This leads to

$$\hat{\beta}^{(k+1)} = (X'W^{(k)}X)^{-1}X'W^{(k)}y^{(k)},$$

and it suggests an iteratively reweighted least squares (IRLS) algorithm.

Residuals

For models with a binary response variable, we need a different measure of residuals. Because we are measuring the model fit in terms of the deviance, D, we may think of the observations as each contributing a quantity d_i, such that $\sum d_i = D$. (Exactly what that value is depends on the form of the systematic component and the link function that are in the likelihood.) The quantity

$$r_i^D = \mathrm{sign}(y_i - \hat{\mu_i})\sqrt{d_i}$$

increases in $(y_i - \hat{\mu_i})$ and $\sum(r_i^D)^2 = D$. We call r_i^D the *deviance residual*.

For the logit model,

$$r_i^D = \mathrm{sign}(y_i - \hat{\mu_i})\sqrt{-2[y_i\log(\hat{\pi_i}) + (1-y_i)\log(1-\hat{\pi_i})]}.$$

Another kind of residual is called the "working" residual. It is

$$r_i^W = (y_i - \hat{\mu_i})\frac{\partial\hat{\eta_i}}{\partial\hat{\mu_i}},$$

where the derivatives are evaluated at the final iteration of the scoring algorithm.

In the logistic regression model, these working residuals are

$$\frac{y_i - \hat{\pi}_i}{\hat{\pi}_i(1-\hat{\pi}_i)}$$

Residuals can be standardized by taking into account their different standard deviations that result from the influence.

This is the same kind of concept as influence in linear models. Here, however, we have

$$\hat{\beta}^{(k+1)} = (X'W^{(k)}X)^{-1}X'W^{(k)}y^{(k)},$$

where the weights are

$$m_i\hat{\pi}_i^{(k)}(1-\hat{\pi}_i^{(k)}).$$

One measure is the diagonal of the hat matrix:

$$W^{\frac{1}{2}}X(X'WX)^{-1}X'W^{\frac{1}{2}}$$

In the case of generalized linear models, the hat matrix is only the prediction transformation matrix for the linear, systematic component.

Data consisting of counts, for example, the number of certain events within a fixed period of time, give rise naturally to a Poisson model. The relationship between the mean and the covariates is often assumed to be multiplicative, giving rise to a log-linear model,

$$\log(\mu) = \eta = x'\beta.$$

Another possibility for count data is that the covariates have an additive effect and the direct relation

$$\mu = x'\beta$$

can be used.

Notice that the mean of the binomial and the Poisson distributions determine the variance.

In practice the variance of discrete response data, such as binomial or Poisson data, is observed to exceed the nominal variance that would be determined by the mean.

This phenomenon is referred to as "over-dispersion". There may be logical explanations for over-dispersion, such as additional heterogeneity over and above what is accounted for by the covariates, or some more complicated variance structure arising from correlations among the responses.

Quasi-likelihood Methods in Generalized Linear Models

Over-dispersion in the generalized linear model can often be accounted for by the nuisance parameter ϕ in the likelihood. For example, we modify the simple binomial model so the variance is

$$\mathrm{V}(y_i|x_i) = \phi\frac{\pi_i(1-\pi_i)}{n_i}.$$

Notice the multiplier ϕ is constant, while π depends on the covariates and n depends on the group size. This of course leads to a more complicated likelihood function, but it may not be necessary to use the actual likelihood.

Quasi-likelihood and need not correspond to any particular distribution; rather quasi can be used to combine any available link and variance function.

Wedderburn (1974) introduced a quasi-likelihood function to allow

$$\mathrm{E}(y|x) = \mu = h(x'\beta)$$

and

$$\mathrm{V}(y|x) = \sigma^2(\mu) = \phi v(\mu),$$

where ϕ is the (nuisance) dispersion parameter in the likelihood and $v(\mu)$ is a variance function that is entirely separate from the likelihood.

McCullagh (1983) extended the concept of quasi-likelihood to allow for a variance-covariance matrix V, and derived an asymptotic theory for the resulting estimators. There may or may not be a true likelihood function with corresponding mean and variance (see Morris, 1982).

Quasi-likelihood methods require only specification of a relationship between the mean and variance of the response.

Fahrmeir and Tutz (1994), following ideas of Gourieroux, Montfort, and Trognon (1984), assume that the mean is indeed given by $\mu = h(x'\beta)$, but that the variance is

$$\mathrm{V}(y|x) = \sigma_0^2(\mu),$$

which may be different from $\sigma^2(\mu) = \phi v(\mu)$. They take $\phi v(\mu)$ to be a "working variance". Assuming that the responses are independent, they write a "quasi-score function",

$$s(\beta) = \sum_i x_i D_i(\beta)\sigma_i^2(\beta)(y_i - \mu_i(\beta)),$$

where $\mu_i(\beta)$ is the correct mean, i.e., $h(x_i'\beta)$ and $D_i(\beta)$ is the derivative of h, but $\sigma_i^2(\beta)$ is a working variance, $\phi v(\mu_(\beta))$, with v arbitrary. Obviously, to use the quasi-score function for computing estimates, v must be somewhat close to the true variance of the data.

In the ordinary quasi-likelihood methods, the variance function is assumed known (or arbitrary, but not estimated directly). Nelder and Pregibon (1987)

developed an extended quasi-likelihood approach, in which the variance function is also studied.

Green and Silverman (1994), using a roughness penalty approach, developed quasi-likelihood methods for semiparametric generalized linear models, in which the penalized quasi-likelihood estimates are penalized least squares estimates with a weight function corresponding to the inverse of the variance-covariance matrix, V.

Morgenthaler (1992) and Jung (1996) considered quasi-likelihood methods of estimation for generalized linear models using least absolute deviations and other robust fitting methods.

Notes

Unbiasedness and Consistency

While many MLEs are biased, most of the ones encountered in common situations are at least consistent in mean-squared error. Neyman and Scott (1948) give an example, which is a simplified version of an example due to Wald, of an MLEs that is not consistent. The problem is the standard one-way ANOVA model with two observations per class. The asymptotics are in the number of classes, and hence, of course in the number of observations. The model is $X_{ij} \sim \mathrm{N}(\mu_j, \sigma^2)$ with $i = 1, 2$ and $j = 1, 2, \ldots$. The asymptotic (and constant) expectation of the MLE of σ^2 is $\sigma^2/2$. This example certainly shows that MLEs may behave very poorly, but its special property should be recognized. The dimension of the parameter space is growing at the same rate as the number of observations.

Exercises in Shao

- For practice and discussion
  4.96(a)(g)(h), 4.107, 4.151, 5.20, 5.21 (Solutions in Shao, 2005)
- To turn in
  4.94, 4.95, 4.97, 4.109, 4.120, 4.152, 5.90

Additional References

Neyman, J., and E. L. Scott (1948) Consistent estimates based on partially consistent observations, *Econometrica* **16**, 1–32.

6

Testing Statistical Hypotheses (Shao Ch 6; TSH3 Ch 3, 4, 5)

In statistical hypothesis testing, the basic problem is to decide whether or not to reject a statement about the distribution of a random variable. The statement must be expressible in terms of membership in a well-defined class. The hypothesis can therefore be expressed by the statement that the distribution of the random variable X is in the class $\mathcal{P}_H = \{P_\theta : \theta \in \Theta_H\}$. An hypothesis of this form is called a statistical hypothesis.

The basic paradigm of statistical hypothesis testing was described in Section 2.4.1, beginning on page 103. We first review some of those ideas, and then in Sections 6.1 and 6.2 we consider the issue of optimality of tests. As we saw in the point estimation problem, it is often not possible to develop a procedure that is *uniformly* optimal. As with the estimation problem, we can impose restrictions, such as unbiasedness or invariance (Section 8.3), or we can define uniformity in terms of a global averaging (Section 3.4). If we impose restrictions, we then proceed to find uniformly most powerful tests under those restrictions. We discuss uniformly most powerful unbiased tests in Section 6.3. In Sections 6.4, 6.6, and 6.7, we discuss general methods for constructing tests.

This kind of statement is usually broken into two pieces, one part an assumption, "assume the distribution of X is in the class $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$", and the other part the hypothesis, "$\theta \in \Theta_H$, where $\Theta_H \subset \Theta$." Given the assumptions, and the definition of Θ_H, we often denote the hypothesis as H, and write it as

$$H : \theta \in \Theta_H.$$

While, in general, to reject the hypothesis H would mean to decide that $\theta \notin \Theta_H$, it is generally more convenient to formulate the testing problem as one of deciding between two statements:

$$H_0 : \theta \in \Theta_0$$

and

$$H_1 : \theta \in \Theta_1,$$

where $\Theta_0 \cap \Theta_1 = \emptyset$.

We do not treat H_0 and H_1 symmetrically; H_0 is the *hypothesis* (or "null hypothesis") to be tested and H_1 is the *alternative*. This distinction is important in developing a methodology of testing.

To test the hypotheses means to choose one or the other; that is, to make a decision, d. A *nonrandomized test procedure* is a rule $\delta(X)$ that assigns two decisions to two disjoint subsets, C_0 and C_1, of the range of X. In general, we require $C_0 \cup C_1$ be the support of X. We equate those two decisions with the real numbers d_0 and d_1, so $\delta(X)$ is a real-valued function,

$$\delta(x) = \begin{cases} d_0 \text{ for } x \in C_0 \\ d_1 \text{ for } x \in C_1. \end{cases}$$

For simplicity, we choose $d_0 = 0$ and $d_1 = 1$. Note for $i = 0, 1$,

$$\Pr(\delta(X) = i) = \Pr(X \in C_i).$$

We call C_1 the *critical region*, and generally denote it by just C. (It is not my intent to distinguish C from R above; they're both "critical regions". I have used C to denote a set of values of X, and R to denote a set of values of $T(X)$.

If $\delta(X)$ takes the value 0, the decision is not to reject; if $\delta(X)$ takes the value 1, the decision is to reject. If the range of $\delta(X)$ is $\{0, 1\}$, the test is a nonrandomized test. Sometimes, however, it is useful to expand the range of $\delta(X)$ to be $[0, 1]$, where we can interpret a value of $\delta(X)$ as the probability that the null hypothesis is rejected. If it is not the case that $\delta(X)$ equals 0 or 1 a.s., we call the test a randomized test.

Our standard approach in hypothesis testing is to control the level of the probability of a type I error under the assumptions, and to try to find a test subject to that significance level that has a small probability of a type II error. We call the maximum allowable probability of a type I error the "significance level", and usually denote it by α. We will call the probability of rejecting the null hypothesis the power of the test, and will denote it by β. If the alternate hypothesis is the true state of nature, the power is one minus the probability of a type II error.

It is clear that we can easily decrease the probability of one (if its probability is positive) at the cost of increasing the probability of the other.

Hence, in one approach to hypothesis testing under the given assumptions on X, and the notation above), we choose $\alpha \in (0, 1)$ and require that $\delta(X)$ be such that

$$\Pr(\delta(X) = 1 \mid \theta \in \Theta_0) \leq \alpha. \tag{6.1}$$

and, subject to this, find $\delta(X)$ so as to minimize

$$\Pr(\delta(X) = 0 \mid \theta \in \Theta_1). \tag{6.2}$$

Optimality of a test T is defined in terms of this constrained optimization problem.

Notice that the restriction on the type I error applies $\forall\theta \in \Theta_0$. We call

$$\sup_{\theta\in\Theta_0} \Pr(\delta(X) = 1 \mid \theta) \tag{6.3}$$

the *size of the test*. If the size is less than the level of significance, the test is said to be *conservative*, and in that case, we often refer to α as the "nominal size".

Note that there is a difference in *choosing* the test procedure, and in *using* the test. The question of the choice of α comes back. Does it make sense to choose α first, and then proceed to apply the test just to end up with a decision d_0 or d_1? It is not likely that this rigid approach would be very useful for most objectives. In statistical data analysis our objectives are usually broader than just deciding which of two hypotheses appears to be true. On the other hand, if we have a well-developed procedure for testing the two hypotheses, the decision rule in this procedure could be very useful in data analysis. One common approach is to use the *functional form* of the rule, but not to predefine the critical region. Then, *given the same setup* of null hypothesis and alternative, to collect data $X = x$, and to determine the smallest value $\hat{\alpha}(x)$ at which the null hypothesis would be rejected. The value $\hat{\alpha}(x)$ is called the *p-value* of x associated with the hypotheses. The p-value indicates the strength of the evidence of the data against the null hypothesis.

We call the probability of rejecting H_0 the *power* of the test, and denote it by β, or for the particular test $\delta(X)$, β_T. (Some authors define the power only in the case that H_1 is true, but we do not restrict the definition in that way.) The power in the case that H_1 is true is 1 minus the probability of a type II error. The probability of a type II error is generally a function of the true distribution of the sample P_θ, and hence so is the power, which we may emphasize by the notation $\beta_\delta(P_\theta)$ or $\beta_\delta(\theta)$. Thus, the second step above is equivalent to maximizing the power within Θ_1. Because the probability of a type II error is generally a function of θ, what does the second step to minimize the probability mean; that is, to minimize it *for what values of* θ? Ideally, we would like a procedure that yields the minimum for all values of θ; that is, one that is most powerful for all values of θ. We call such a procedure a *uniformly most powerful* or UMP test. For a given problem, finding such procedures, or establishing that they do not exist, will be one of our primary objectives.

Decision Theoretic Approach

As in any decision-theoretic formulation of a statistical procedure, we seek to minimize the *risk*:

$$R(P, \delta) = \mathrm{E}\big(L(P, \delta(X))\big). \tag{6.4}$$

In the case of the 0-1 loss function and the four possibilities, the risk is just the probabiltiy of either type of error.

We obviously want a test procedure that minimizes the risk.

Randomized Tests

Just as in the theory of point estimation, we found randomized procedures useful for establishing properties of estimators or as counterexamples to some statement about a given estimator, we can use randomized test procedures to establish properties of tests. (While randomized estimators rarely have application in practice, randomized test procedures can actually be used to increase the power of a conservative test. Use of a randomized test in this way would not make much sense in real-world data analysis, but if there are regulatory conditions to satisfy, it might be useful.)

Let us define a random experiment R that has two outcomes, r_0 and r_1, such that

$$\Pr(R = r_0) = 1 - \phi(x)$$

and so

$$\Pr(R = r_1) = \phi(x).$$

We could also extend this to more than two outcomes, as we indicated above for δ. This is a randomized test if we equate r_0 with d_0 and r_1 with d_1.

6.1 Optimal Tests

Optimal tests are those that minimize the risk (6.4). The risk considers the total expected loss. In the testing problem, we generally prefer to restrict the probability of a type I error as in inequality (6.1) and then, subject to that, minimize the probability of a type II error as in equation (6.2), which is equivalent to maximizing the power under the alternative hypothesis.

An Optimal Test in a Simple Situation

First, consider the problem of picking the optimal critical region C in a problem of testing the hypothesis that a discrete random variable has the probabiltiy mass function $p_0(x)$ versus the alternative that it has the probabiltiy mass function $p_1(x)$.

We will develop an optimal test for any given significance level based on one observation. For $x \ni p_0(x) > 0$, let

$$r(x) = \frac{p_1(x)}{p_0(x)},$$

and label the values of x for which r is defined so that $r(x_1) \geq r(x_2) \geq \cdots$.

Let N be the set of x for which $p_0(x) = 0$ and $p_1(x) > 0$. Assume that there exists a j such that

$$\sum_{i=1}^{j} p_0(x_i) = \alpha.$$

If S is the set of x for which we reject the test, we see that the significance level is

$$\sum_{x_i \in S} p_0(x_i).$$

and the power over the region of the alternative hypothesis is

$$\sum_{x_i \in S} p_1(x_i).$$

Then it is clear that if $C = \{x_1, \ldots, x_j\} \cup N$, then $\sum_{x \in S} p_1(x)$ is maximized over all sets C subject to the restriction on the size of the test.

If there does not exist a j such that $\sum_{i=1}^{j} p_0(x_i) = \alpha$, the rule is to put $x_1, \ldots, x_j$ in C so long as

$$\sum_{i=1}^{j} p_0(x_i) = \alpha^* < \alpha.$$

We then define a randomized auxilliary test R

$$\begin{aligned} \Pr(R = r_1) &= \phi(x_{j+1}) \\ &= (\alpha - \alpha^*)/p_0(x_{j+1}) \end{aligned}$$

It is clear in this way that $\sum_{x \in S} p_1(x)$ is maximized subject to the restriction on the size of the test.

Example: Two Discrete Distributions

Consider two distributions with support on a subset of $\{0, 1, 2, 3, 4, 5\}$. Let $p_0(x)$ and $p_1(x)$ be the probability mass functions. Based on one observation, we want to test $H_0 : p_0(x)$ is the mass function versus $H_1 : p_1(x)$ is the mass function.

Suppose the distributions are as shown in Table 6.1, where we also show the values of r and the labels on x determined by r.

Table 6.1. Formulas for Some Vector Derivatives

x	0	1	2	3	4	5
p_0	.05	.10	.15	0	.50	.20
p_1	.15	.40	.30	.05	.05	.05
r	3	4	2	-	1/10	2/5
label	2	1	3	-	5	4

Thus, for example, we see $x_1 = 1$ and $x_2 = 0$. Also, $N = \{3\}$.

For given α, we choose C such that

$$\sum_{x \in C} p_0(x) \leq \alpha$$

and so as to maximize

$$\sum_{x \in C} p_1(x).$$

We find the optimal C by first ordering $r(x_{i_1}) \geq r(x_{i_2}) \geq \cdots$ and then satisfying $\sum_{x \in C} p_0(x) \leq \alpha$. The ordered possibilities for C in this example are

$$\{1\} \cup \{3\}, \quad \{1, 0\} \cup \{3\}, \quad \{1, 0, 2\} \cup \{3\}, \quad \cdots .$$

Notice that including N in the critical region does not cost us anything (in terms of the type I error that we are controlling).

Now, for any given significance level, we can determine the optimum test based on one observation.

- Suppose $\alpha = .10$. Then the optimal critical region is $C = \{1, 3\}$, and the power for the null hypothesis is $\beta_\delta(p_1) = .45$.
- Suppose $\alpha = .15$. Then the optimal critical region is $C = \{0, 1, 3\}$, and the power for the null hypothesis is $\beta_\delta(p_1) = .60$.
- Suppose $\alpha = .05$. We cannot put 1 in C, with probability 1, but if we put 1 in C with probability 0.5, the α level is satisfied, and the power for the null hypothesis is $\beta_\phi(p_1) = .22$.
- Suppose $\alpha = .05$. We cannot put 1 in C, with probability 1, but if we put 1 in C with probability 0.5, the α level is satisfied, and the power for the null hypothesis is $\beta_\phi(p_1) = .25$.
- Suppose $\alpha = .20$. We choose $C = \{0, 1, 3\}$ with probability 2/3 and $C = \{0, 1, 2, 3\}$ with probability 1/3. The α level is satisfied, and the power for the null hypothesis is $\beta_\phi(p_1) = .75$.

All of these tests are most powerful based on one observations for the given values of α.

Now, how would we extend this idea to tests based on two observations. We see immediately that the ordered critical regions are

$$C_1 = \{(1, 3)\} \times \{(1, 3)\}, \quad C_1 \cup \{(1, 3)\} \times \{(0, 3)\}, \quad \cdots .$$

Extending this direct enumeration would be tedious, but, at this point we have grasped the implication: the ratio of the likelihoods is the basis for the most powerful test. The is the Neyman-Pearson Fundamental Lemma.

The Neyman-Pearson Fundamental Lemma

The example in the previous section illustrates the way we can approach the problem of testing any simple hypothesis against another simple hypothesis.

Thinking of the hypotheses in terms of a parameter θ that indexes these two densities by θ_0 and θ_1, for a sample $X = x$, we have the likelihoods

associated with the two hypotheses as $L(\theta_0; x)$ and $L(\theta_1; x)$. We may be able to define an α-level critical region for nonrandomized tests in terms of the ratio of these likelihoods: Let us assume that a positive number k exists such that there is a subset of the sample space C with complement with respect to the sample space $\bar{C}$, such that

$$\begin{aligned} \frac{L(\theta_1; x)}{L(\theta_0; x)} &\geq k \quad \forall x \in C \\ \frac{L(\theta_1; x)}{L(\theta_0; x)} &\leq k \quad \forall x \in \bar{C} \end{aligned} \tag{6.5}$$

and

$$\alpha = \Pr(X \in C \mid H_0).$$

(Notice that such a k and C may not exist.)

The Neyman-Pearson Fundamental Lemma tells us this test based on the *likelihood ratio* is the most powerful nonrandomized test of the simple null H_0 that specifies the density p_0 for X versus the simple alternative H_1 that specifies the density p_1. Let's consider the form of the Lemma that does not involve a randomized test; that is, in the case that an exact α-level nonrandomized test exists, as assumed above. Let k and C be as above. Then the Neyman-Pearson Fundamental Lemma states that C is the best critical region of size α for testing H_0 versus H_1.

Proof. Let A be any critical region of size α. We want to prove

$$\int_C L(\theta_1) - \int_A L(\theta_1) \geq 0.$$

We can write this as

$$\begin{aligned} \int_C L(\theta_1) - \int_A L(\theta_1) &= \int_{C\cap A} L(\theta_1) + \int_{C\cap \bar{A}} L(\theta_1) - \int_{A\cap C} L(\theta_1) - \int_{A\cap \bar{C}} L(\theta_1) \\ &= \int_{C\cap \bar{A}} L(\theta_1) - \int_{A\cap \bar{C}} L(\theta_1). \end{aligned}$$

By the given condition, $L(\theta_1; x) \geq kL(\theta_0; x)$ at each $x \in C$, so

$$\int_{C\cap \bar{A}} L(\theta_1) \geq k \int_{C\cap \bar{A}} L(\theta_0),$$

and $L(\theta_1; x) \leq kL(\theta_0; x)$ at each $x \in \bar{C}$, so

$$\int_{A\cap \bar{C}} L(\theta_1) \leq k \int_{A\cap \bar{C}} L(\theta_0).$$

Hence

$$\int_C L(\theta_1) - \int_A L(\theta_1) \geq k \left(\int_{C\cap \bar{A}} L(\theta_0) - \int_{A\cap \bar{C}} L(\theta_0) \right).$$

But

$$\begin{aligned}\int_{C\cap\bar{A}} L(\theta_0) - \int_{A\cap\bar{C}} L(\theta_0) &= \int_{C\cap\bar{A}} L(\theta_0) + \int_{C\cap A} L(\theta_0) - \int_{C\cap A} L(\theta_0) - \int_{A\cap\bar{C}} L(\theta_0) \\ &= \int_C L(\theta_0) - \int_A L(\theta_0) \\ &= \alpha - \alpha \\ &= 0.\end{aligned}$$

Hence, $\int_C L(\theta_1) - \int_A L(\theta_1) \geq 0$. ∎

This simple statement of the Neyman-Pearson Lemma and its proof should be in your bag of easy pieces.

The Lemma applies more generally by use of a random experiment so as to achieve the level α. Shao gives a clear statement and proof.

Generalizing the Optimal Test to Hypotheses of Intervals

Although it applies to a simple alternative (and hence "uniform" properties do not make much sense), the Neyman-Pearson Lemma gives us a way of determining whether a *uniformly most powerful* (UMP) test exists, and if so how to find one. We are often interested in testing hypotheses in which either or both of Θ_0 and Θ_1 are continuous regions of $\mathbb{R}$ (or $\mathbb{R}^k$).

We must look at the likelihood ratio as a function both of θ and x. The question is whether, for given θ_0 and any $\theta_1 > \theta_0$ (or equivalently any $\theta_1 < \theta_0$), the likelihood is monotone in some function of x; that is, whether the family of distributions of interest is parameterized by a scalar in such a way that it has a *monotone likelihood ratio* (see page 61). In that case, it is clear that we can extend the test in (6.5) to test to be uniformly most powerful for testing $H_0 : \theta = \theta_0$ against an alternative $H_1 : \theta > \theta_0$ (or $\theta_1 < \theta_0$).

The exponential class of distributions is important because UMP tests are easy to find for families of distributions in that class. Discrete distributions are especially simple, but there is nothing special about them. As an example, work out the test for $H_0 : \lambda \geq \lambda_0$ versus the alternative $H_1 : \lambda < \lambda_0$ in a one-parameter exponential distribution. (The one-parameter exponential distribution, with density over the positive reals $\lambda e^{-\lambda x}$ is a member of the exponential class. Recall that the two-parameter exponential distribution used is not a member of the exponential family.)

Two easy pieces you should have are construction of a UMP for the hypotheses in the one-parameter exponential (above), and the construction of a UMP for testing $H_0 : \pi \geq \pi_0$ versus the alternative $H_1 : \pi < \pi_0$ in a binomial(n, π) distribution.

Use of Sufficient Statistics

It is a useful fact that if there is a sufficient statistic $S(X)$ for θ, and $\tilde{\delta}(X)$ is an α-level test for an hypothesis specifying values of θ, then there exists an

α-level test for the same hypothesis, $\delta(S)$ that depends only on $S(X)$, and which has power at least as great as that of $\tilde{\delta}(X)$. We see this by factoring the likelihoods.

Nuisance Parameters and Similar Regions

A situation in which a sufficient statistic may be important is when there are additional parameters not specified by the hypotheses being tested. In this situation we have $\theta = (\theta_s, \theta_u)$, and the hypothesis may be of the form

$$H_0 \ : \ \theta_s = \theta_{s0}.$$

The problem is that the performance of the test, that is, $\mathrm{E}(\delta(X))$ may depend on θ_u. There is nothing we can do about this. We might, however, seek to do something about it in some cases; in particular, we might try to eliminate the dependency under the null hypothesis. We think it is important to control the size of the test; hence, we require, for given α,

$$\mathrm{E}_{H_0}(\delta(X)) \leq \alpha.$$

Is this possible? It certainly is if $\alpha = 1$; that is if the rejection region is the entire sample space. Are there regions similar to the sample space in this regard? Maybe. If a critical region is such that $\mathrm{E}_{H_0}(\delta(X)) \leq \alpha$ for all values of θ the region is called an α-level *similar region* with respect to θ (or more precisely with respect to θ_u), and the test is called an α-level *similar test*.

6.2 Uniformly Most Powerful Tests

The probability of a type I error is limited to α or less. We seek a procedure that yields the minimum probability of a type II error. This would be a "most powerful" test. Ideally, the test would be most powerful for all values of θ; that is, one that is most powerful for all values of θ. We call such a procedure a *uniformly most powerful* or UMP test. For a given problem, finding such procedures, or establishing that they do not exist, will be one of our primary objectives. The Neyman-Pearson Lemma gives us a way of determining whether a UMP test exists, and if so how to find one. The main issue is the likelihood ratio as a function of the parameter in the region specified by a composite H_1. If the likelihood ratio is monotone, then we have a UMP based on the ratio.

You should be able to develop UMP tests for a variety of problems using this fact. You should also be able to identify when a UMP test cannot exist.

UMP Tests when There Are Nuisance Parameters

Let U be a sufficient statistic for θ_u, suppose an α-level test $\delta(x)$ of an hypothesis H_0 specifying θ_s is such that

$$\mathrm{E}_{H_0}(\delta(X)|U) = \alpha.$$

(This condition on the critical region is called "Neyman structure".) Because U is sufficient, taking the expectation over U, we have

$$\begin{aligned}\mathrm{E}_{H_0}(\mathrm{E}_{H_0}(\delta(X)|U)) &= \mathrm{E}_{H_0}(\delta(X)) \\ &= \alpha\end{aligned}$$

and so $\delta(X)$ is a similar test of size α.

Now suppose that U is boundedly complete sufficient for θ_u. If

$$\mathrm{E}_{H_0}(\mathrm{E}_{H_0}(\delta(X)|U)) = \alpha,$$

then the test has Neyman structure.

While the power may still depend on θ_u, this fact may allow us to determine UMP tests of given size without regard to the nuisance parameters.

Unfortunately, in the presence of nuisance parameters, usually UMP tests do not exist. (We must add other restrictions on the tests, as we see below.)

Nonexistence of UMP Tests

One of the most interesting cases in which a UMP test cannot exist is the two-sided null hypothesis

$$H_0 : \theta = \theta_0$$

versus the alternative

$$H_1 : \theta \neq \theta_0.$$

This is easy to see, and you should reason through this statement to see that it is true.

So what can we do?

There are basically two ways of approaching the problem. We can add a requirement on the uniformity or we can introduce an additional criterion.

In point estimation when we realized we could not have an estimator that would uniformly minimize the risk, we required unbiasedness or invariance, or we added some global property of the risk, such as minimum averaged risk or minimum maximum risk. We might introduce similar criteria for the testing problem.

First, let's consider a desirable property that we will call *unbiasedness*.

6.3 UMP Unbiased Tests

Recall that there are a couple of standard definitions of unbiasedness.

- If a random variable X has a distribution with parameter θ, for a point estimator $T(X)$ of an estimand $g(\theta)$ to be unbiased means that

$$\mathrm{E}_\theta(T(X)) = g(\theta).$$

  Although no loss function is specified in this meaning of unbiasedness, we know that such an estimator minimizes the risk based on a squared-error loss function. (This last statement is not iff. Under squared-error loss the conditions of minimum risk and unbiasedness defined in this way are equivalent if g is continuous and not constant over any open subset of the parameter space and if $\mathrm{E}_\theta(T(X))$ is a continuous function of θ.)
- Another definition of unbiasedness is given with direct reference to a loss function. This is sometimes called L-unbiasedness. The estimator (or more generally, the procedure) $T(X)$ is said to be L-unbiased under the loss function L, if for all θ and $\tilde{\theta}$,

$$\mathrm{E}_\theta(L(\theta, T(X))) \leq \mathrm{E}_\theta(L(\tilde{\theta}, T(X))).$$

  Notice the subtle differences in this property and the property of an estimator that may result from an approach in which we seek a minimum-risk estimator; that is, an approach in which we seek to solve the minimization problem,

$$\min_T \mathrm{E}_\theta(L(\theta, T(X)))$$

  for all θ. This latter problem does not have a solution. (Recall the approach is to add other restrictions on $T(X)$.)
  L-unbiasedness under a squared-error also leads to the previous definition of unbiasedness.

Unbiasedness in hypothesis testing is the property that the test is more likely to reject the null hypothesis at any point in the parameter space specified by the alternative hypothesis than it is at any point in the parameter space specified by the null hypothesis. More formally, the α-level test δ with power function $\beta(\theta) = \mathrm{E}_\theta(\delta(X))$ of the hypothesis $H_0 : \theta \in \Theta_{H_0}$ versus $H_1 : \theta \in \Theta_{H_1}$ is said to be *unbiased* if

$$\beta(\theta) \leq \alpha \quad \forall\, \theta \in \Theta_{H_0}$$

and

$$\beta(\theta) \geq \alpha \quad \forall\, \theta \in \Theta_{H_1}.$$

Notice that this unbiasedness depends not only on the hypotheses, but also on the significance level.

This definition of unbiasedness for a test is L-unbiasedness if the loss function is 0-1.

Notice that if an α-level UMP test exists, it is unbiased, because its power is at least as great as the power of the constant test (for all x), $\delta(x) = \alpha$.

Similar UMPU tests remain so in the presence of nuisance parameters.

6.4 Likelihood Ratio Tests

We see that the Neyman-Pearson Lemma leads directly to use of the ratio of the likelihoods in constructing tests. Now we want to generalize this approach and to study the properties of tests based on that ratio. Although as we have emphasized, the likelihood is a function of the distribution rather than of the random variable, we want to study its properties under the distribution of the random variable. Using the idea of the ratio as in the test (6.5) of $H_0 : \theta \in \Theta_0$, but inverting that ratio and including both hypotheses in the denominator, we define the *likelihood ratio* as

$$\lambda(X) = \frac{\sup_{\theta\in\Theta_0} L(\theta; X)}{\sup_{\theta\in\Theta} L(\theta; X)}. \tag{6.6}$$

The test, similarly to (6.5), rejects H_0 if $\lambda(X) < c$, where c is some value in $[0, 1]$. (The inequality goes in the opposite direction because we have inverted the ratio.) Tests such as this are called *likelihood ratio tests*. (We should note that there are other definitions of a likelihood ratio; in particular, in TSH3 its denominator is the sup over the alternative hypothesis. If the alternative hypothesis does not specify $\Theta - \Theta_0$, such a definition requires specification of both H_0, and H_1; whereas (6.6) requires specification only of H_0.)

The likelihood ratio may not exist, but if if is well defined, clearly it is in the interval $[0, 1]$, and values close to 1 provide evidence that the null hypothesis is true, and values close to 0 provide evidence that it is false.

6.5 Sequential Probability Ratio Tests

Wald (1945)

6.6 Asymptotic Likelihood Ratio Tests

Some of the most important properties of LR tests are asymptotic ones.

There are various ways of using the likelihood to build practical tests. Some are asymptotic tests that use MLEs (or RLEs).

Asymptotic Properties of Tests

For use of asymptotic approximations for hypothesis testing, we need a concept of asymptotic significance, as discussed on page 118.

We assume a family of distributions $\mathcal{P}$, a sequence of statistics $\{\delta_n\}$ based on a random sample $X_1, \ldots, X_n$. In hypothesis testing, the standard setup is that we have an observable random variable with a distribution in the family $\mathcal{P}$. Our hypotheses concern a specific member $P \in \mathcal{P}$. We have a null hypothesis

$$H_0 : \quad P \in \mathcal{P}_0$$

and an alternative hypothesis

$$H_1 : \quad P \in \mathcal{P}_1,$$

where $\mathcal{P}_0 \subset \mathcal{P}$, $\mathcal{P}_1 \subset \mathcal{P}$, and $\mathcal{P}_0 \cap \mathcal{P}_1 = \emptyset$.

We often define the test statistic δ in regard to the decisions, which we denote by 1 for the case of deciding to reject H_0 and conclude H_1, and by 0 for the case of deciding not to reject H_0.

Letting

$$\beta(\delta_n, P) = \Pr(\delta_n = 1),$$

we define $\limsup_n \beta(\delta_n, P) \, \forall P \in \mathcal{P}_0$, if it exists, as the *asymptotic size* of the test.

If $\limsup_n \beta(\delta_n, P) \leq \alpha \, \forall P \in \mathcal{P}_0$, then α is an *asymptotic significance level* of the test.

δ_n is *consistent* for the test iff $\limsup_n \beta(\delta_n, P) - 0 \, \forall P \in \mathcal{P}_1$.

δ_n is *Chernoff-consistent* for the test iff δ_n is consistent and furthermore, $\limsup_n \beta(\delta_n, P) - 0 \, \forall P \in \mathcal{P}_0$.

The asymptotic distribution of a maximum of a likelihood is a chi-squared and the ratio of two is asymptotically an F.

Regularity Conditions

The interesting asymptotic properties of LR tests depend on the Le Cam regularity conditions, which go slightly beyond the Fisher information regularity conditions. (See page 60.)

These are the conditions to ensure that superefficiency can only occur over a set of Lebesgue measure 0 (Shao Theorem 4.16), the asymptotic efficiency of RLEs (Shao Theorem 4.17), and the chi-squared asymptotic significance of LR tests (Shao Theorem 6.5).

Asymptotic Significance of LR Tests

We consider a general form of the null hypothesis,

$$H_0 : R(\theta) = 0 \tag{6.7}$$

versus the alternative

$$H_1 : R(\theta) \neq 0, \tag{6.8}$$

for a continuously differential function $R(\theta)$ from $\mathbb{R}^k$ to $\mathbb{R}^r$. (Shao's notation, $H_0 : \theta = g(\vartheta)$ where ϑ is a $(k-r)$-vector, although slightly different, is equivalent.)

The key result is Theorem 6.5 in Shao, which, assuming the Le Cam regularity conditions, says that under H_0,

$$-2\log(\lambda_n) \rightarrow_{\mathrm{d}} \chi^2_r,$$

where χ^2_r is a random variable with a chi-squared distribution with r degrees of freedom and r is the number of elements in $R(\theta)$. (In the simple case, r is the number of equations in the null hypothesis.)

This allows us to determine the asymptotic significance of an LR test. It is also the basis for constructing asymptotically correct confidence sets, as we discuss beginning on page 259.

6.7 Wald Tests and Score Tests

There are two types of tests that arise from likelihood ratio tests. These are called Wald tests and score tests. Score tests are also called Rao test or Lagrange multiplier tests. Buse (1982) gives an interesting exposition of the three types of tests (*The American Statistician* **36**, pages 153–157).

These tests are asymptotically equivalent. They are consistent under the Le Cam regularity conditions, and they are Chernoff-consistent if α is chosen so that as $n \rightarrow \infty$, $\alpha \rightarrow 0$ and $\chi^2_{r,\alpha_n} = \mathrm{o}(n)$.

Wald Tests

For the hypostheses (6.7) and (6.8), the *Wald test* uses the test statistics

$$W_n = \Big(R(\hat{\theta})\Big)^\top \left(\Big(S(\hat{\theta})\Big)^\top \Big(I_n(\hat{\theta})\Big)^{-1} S(\hat{\theta})\right)^{-1} R(\hat{\theta}), \tag{6.9}$$

where $S(\theta) = \partial R(\theta)/\partial\theta$ and $I_n(\theta)$ is the Fisher information matrix, and these two quantities are evaluated at an MLE or RLE $\hat{\theta}$. The test rejects the null hypothesis when this value is large.

Notice that for the simple hypothesis $H_0 : \theta = \theta_0$, this simplifies to

$$(\hat{\theta} - \theta_0)^\top I_n(\hat{\theta})(\hat{\theta} - \theta_0).$$

An asymptotic test can be constructed because $W_n \rightarrow_{\mathrm{d}} Y$, where $Y \sim \chi^2_r$ and r is the number of elements in $R(\theta)$. This is proven in Theorem 6.6 of Shao, page 434.

The test rejects at the α level if $W_n > \chi^2_{r,1-\alpha}$, where $\chi^2_{r,1-\alpha}$ is the $1-\alpha$ quantile of the chi-squared distribution with r degrees of freedom. (Note that Shao denotes this quantity as $\chi^2_{r,\alpha}$.)

Score Tests

A related test is the Rao *score test*, sometimes called a *Lagrange multiplier test*. It is based on a MLE or RLE $\tilde{\theta}$ under the restriction that $R(\theta) = 0$ (whence the Lagrange multiplier), and rejects H_0 when the following is large:

$$R_n = (s_n(\tilde{\theta}))^\top \left(I_n(\tilde{\theta})\right)^{-1} s_n(\tilde{\theta}), \tag{6.10}$$

where $s_n(\theta) = \partial l_L(\theta)/\partial\theta$, and is called the *score function*.

The information matrix can either be the Fisher information matrix (that is, the expected values of the derivatives) evaluated at the RLEs or the "observed" information matrix in which instead of expected values, the observed values are used.

An asymptotic test can be constructed because $R_n \to_{\mathrm{d}} Y$, where $Y \sim \chi^2_r$ and r is the number of elements in $R(\theta)$. This is proven in Theorem 6.6 (ii) of Shao.

The test rejects at the α level if $R_n > \chi^2_{r,1-\alpha}$, where $\chi^2_{r,1-\alpha}$ is the $1-\alpha$ quantile of the chi-squared distribution with r degrees of freedom.

Example of Tests

Consider a general regression model:

$$X_i = f(z_i, \beta) + \epsilon, \quad \text{where } \epsilon_i \sim \text{iid}\, \mathrm{N}(0, \sigma^2).$$

For given $k \times r$ matrix L, we want to test

$$H_0 : L\beta = \beta_0$$

Let X be the sample (it's an n-vector). Let Z be the matrix whose rows are the z_i.

The log likelihood is

$$\log \ell(\beta; X) = c(\sigma^2) - \frac{1}{2\sigma^2}(X - f(Z,\beta))^{\mathrm{T}}(X - f(Z,\beta)).$$

The MLE is the LSE, $\hat{\beta}$.

Let $\tilde{\beta}$ be the maximizer of the log likelihood under the restriction $L\beta = \beta_0$.

The likelihood ratio is the same as the difference in the log likelihoods.

The maximum of the unrestricted log likelihood (minus a constant) is the minimum of the residuals:

$$\frac{1}{2\sigma^2}(X - f(Z,\hat{\beta}))^{\mathrm{T}}(X - f(Z,\hat{\beta})) = \frac{1}{2\sigma^2}\mathrm{SSE}(\hat{\beta})$$

and likewise, for the restricted:

$$\frac{1}{2\sigma^2}(X - f(Z,\tilde{\beta}))^{\mathrm{T}}(X - f(Z,\tilde{\beta})) = \frac{1}{2\sigma^2}\mathrm{SSE}(\tilde{\beta}).$$

Now, the difference,

$$\frac{\mathrm{SSE}(\hat{\beta}) - \mathrm{SSE}(\tilde{\beta})}{\sigma^2},$$

has an asymptotic $\chi^2(r)$ distribution. (Note that the 2 goes away.)

We also have that

$$\frac{\mathrm{SSE}(\hat{\beta})}{\sigma^2}$$

has an asymptotic $\chi^2(n-k)$ distribution.

So for the likelihood ratio test we get an "F-type statistic":

$$\frac{(\mathrm{SSE}(\hat{\beta}) - \mathrm{SSE}(\tilde{\beta}))/r}{\mathrm{SSE}(\hat{\beta})/(n-k)}.$$

Use unrestricted MLE $\hat{\beta}$ and consider $L\hat{\beta} - \beta_0$.

$$\mathrm{V}(\hat{\beta}) \to \left(\mathrm{J}_{f(\hat{\beta})}^{\mathrm{T}}\mathrm{J}_{f(\hat{\beta})}\right)^{-1}\sigma^2,$$

and so

$$\mathrm{V}(L\hat{\beta}) \to L\left(\mathrm{J}_{f(\hat{\beta})}^{\mathrm{T}}\mathrm{J}_{f(\hat{\beta})}\right)^{-1}L^{\mathrm{T}}\sigma^2,$$

where $\mathrm{J}_{f(\hat{\beta})}$ is the $n \times k$ Jacobian matrix.

Hence, we can write an asymptotic $\chi^2(r)$ statistic as

$$(L\hat{\beta} - \beta_0)^{\mathrm{T}}\left(L\left(\mathrm{J}_{f(\hat{\beta})}^{\mathrm{T}}\mathrm{J}_{f(\hat{\beta})}\right)^{-1}L^{\mathrm{T}}s^2\right)^{-1}(L\hat{\beta} - \beta_0)$$

We can form a Wishart-type statistic from this.

If $r = 1$, L is just a vector (the linear combination), and we can take the square root and from a "pseudo t":

$$\frac{L^{\mathrm{T}}\hat{\beta} - \beta_0}{s\sqrt{L^{\mathrm{T}}\left(\mathrm{J}_{f(\hat{\beta})}^{\mathrm{T}}\mathrm{J}_{f(\hat{\beta})}\right)^{-1}L}}.$$

Get MLE with the restriction $L\beta = \beta_0$ using a Lagrange multiplier, λ of length r.

Minimize

$$\frac{1}{2\sigma^2}(X - f(Z,\beta))^{\mathrm{T}}(X - f(Z,\beta)) + \frac{1}{\sigma^2}(L\beta - \beta_0)^{\mathrm{T}}\lambda.$$

Differentiate and set = 0:

$$-\mathrm{J}_{f(\hat{\beta})}^{\mathrm{T}}(X - f(Z,\hat{\beta})) + L^{\mathrm{T}}\lambda = 0$$
$$L\hat{\beta} - \beta_0 = 0.$$

$\mathrm{J}_{f(\hat{\beta})}^{\mathrm{T}}(X - f(Z,\hat{\beta}))$ is called the *score* vector. It is of length k.

Now $\mathrm{V}(X - f(Z,\hat{\beta})) \to \sigma^2 I_n$, so the variance of the score vector, and hence, also of $L^{\mathrm{T}}\lambda$, goes to $\sigma^2 \mathrm{J}_{f(\beta)}^{\mathrm{T}}\mathrm{J}_{f(\beta)}$.
(Note this is the true β in this expression.)

Estimate the variance of the score vector with $\tilde{\sigma}^2 \mathrm{J}_{f(\tilde{\beta})}^{\mathrm{T}}\mathrm{J}_{f(\tilde{\beta})}$,

where $\tilde{\sigma}^2 = \mathrm{SSE}(\tilde{\beta})/(n-k+r)$.

Hence, we use $L^{\mathrm{T}}\tilde{\lambda}$ and its estimated variance (previous slide).

Get

$$\frac{1}{\tilde{\sigma}^2}\tilde{\lambda}^{\mathrm{T}}L\left(\mathrm{J}_{f(\tilde{\beta})}^{\mathrm{T}}\mathrm{J}_{f(\tilde{\beta})}\right)^{-1}L^{\mathrm{T}}\tilde{\lambda}$$

It is asymptotically $\chi^2(r)$.

This is the Lagrange multiplier form.

Another form:

Use $\mathrm{J}_{f(\tilde{\beta})}^{\mathrm{T}}(X - f(Z,\tilde{\beta}))$ in place of $L^{\mathrm{T}}\tilde{\lambda}$.

Get

$$\frac{1}{\tilde{\sigma}^2}(X - f(Z,\tilde{\beta}))^{\mathrm{T}}\mathrm{J}_{f(\tilde{\beta})}\left(\mathrm{J}_{f(\tilde{\beta})}^{\mathrm{T}}\mathrm{J}_{f(\tilde{\beta})}\right)^{-1}\mathrm{J}_{f(\tilde{\beta})}^{\mathrm{T}}(X - f(Z,\tilde{\beta}))$$

This is the score form. Except for the method of computing it, it is the same as the Lagrange multiplier form.

This is the SSReg in the AOV for a regression model.

An Anomalous Score Test

An interesting example of the use of a score test is discussed by Morgan, Palmer, and Ridout, Verbeke and Molenberghs, and Freedman in *The American Statistician* **61** (2007), pages 285–295.

Morgan, Palmer, and Ridout (MPR) illustrate some interesting issues using a simple example of counts of numbers of stillbirths in each of a sample of litters of laboratory animals.

MPR suggest that a zero-inflated Poisson is an appropriate model. This distribution is an ω mixture of a point mass at 0 and a Poisson distribution. The CDF (in a notation we will use often later) is

$$P_{0,\omega}(x|\lambda) = (1-\omega)P(x|\lambda) + \omega \mathrm{I}_{[0,\infty)}(x),$$

where $P(x)$ is the Poisson CDF with parameter λ.

(Write the PDF (under the counting measure). Is this a reasonable probability model? What are the assumptions? Do the litter sizes matter?)

If we denote the number of litters in which the number of observed stillbirths is i by n_i, the log-likelihood function is

$$l(\omega,\lambda) = n_0 \log\left(\omega + (1-\omega)\mathrm{e}^{-\lambda}\right) + \sum_{i=1}^{\infty} n_i \log(1-\omega) - \sum_{i=1}^{\infty} n_i\lambda + \sum_{i=1}^{\infty} i n_i \log(\lambda) + c.$$

Suppose we want to test the null hypothesis that $\omega = 0$.

The score test has the form

$$s^{\mathrm{T}} J^{-1} s,$$

where s is the score vector and J is either the observed or the expected information matrix. For each we substitute $\omega = 0$ and $\lambda = \hat{\lambda}_0$, where $\hat{\lambda}_0 = \sum_{i=1}^{\infty} i n_i / n$ with $n = \sum_{i=0}^{\infty} n_i$, which is the MLE when $\omega = 0$.

Let

$$n_+ = \sum_{i=1}^{\infty} n_i$$

and

$$d = \sum_{i=0}^{\infty} i n_i.$$

The frequency of 0s is important. Let

$$f_0 = n_0/n.$$

Taking the derivatives and setting $\omega = 0$, we have

$$\frac{\partial l}{\partial \omega} = n_0 \mathrm{e}^{\lambda} - n,$$

$$\frac{\partial l}{\partial \lambda} = -n + d/\lambda,$$

$$\frac{\partial^2 l}{\partial \omega^2} = -n - n_0 \mathrm{e}^{2\lambda} + n_0 \mathrm{e}^{\lambda},$$

$$\frac{\partial^2 l}{\partial \omega \lambda} = n_0 \mathrm{e}^{\lambda},$$

and

$$\frac{\partial^2 l}{\partial \lambda^2} = -d/\lambda^2.$$

So, substituting the observed data and the restricted MLE, we have observed information matrix

$$O(0,\hat{\lambda}_0) = n \begin{bmatrix} 1 + f_0 \mathrm{e}^{2\hat{\lambda}_0} - 2 f_0 \mathrm{e}^{\hat{\lambda}_0} & -f_0 \mathrm{e}^{\hat{\lambda}_0} \\ -f_0 \mathrm{e}^{\hat{\lambda}_0} & 1/\hat{\lambda}_0 \end{bmatrix}.$$

Now, for the expected information matrix when $\omega = 0$, we first observe that $\mathrm{E}(n_0) = n\mathrm{e}^{-\lambda}$, $\mathrm{E}(d) = n\lambda$, and $\mathrm{E}(n_+) = n(1-\mathrm{e}^{-\lambda})$; hence

$$I(0,\hat{\lambda}_0) = n\begin{bmatrix} \mathrm{e}^{\hat{\lambda}_0}-1 & -1 \\ -1 & 1/\hat{\lambda}_0 \end{bmatrix}.$$

Hence, the score test statistic can be written as

$$\kappa(\hat{\lambda}_0)(n_0\mathrm{e}^{\hat{\lambda}_0}-n)^2,$$

where $\kappa(\hat{\lambda}_0)$ is the (1,1) element of the inverse of either $O(0,\hat{\lambda}_0)$ or $I(0,\hat{\lambda}_0)$.

Inverting the matrices (they are 2×2), we have as the test statistic for the score test, either

$$s_I = \frac{n\mathrm{e}^{-\hat{\lambda}_0}(1-\theta)^2}{1-\mathrm{e}^{-\hat{\lambda}_0}-\hat{\lambda}_0\mathrm{e}^{-\hat{\lambda}_0}}$$

or

$$s_O = \frac{n\mathrm{e}^{-\hat{\lambda}_0}(1-\theta)^2}{\mathrm{e}^{-\hat{\lambda}_0}+\theta-2\theta\mathrm{e}^{-\hat{\lambda}_0}\theta^2\hat{\lambda}_0\mathrm{e}^{-\hat{\lambda}_0}},$$

where $\theta = f_0\mathrm{e}^{\hat{\lambda}_0}$, which is the ratio of the observed proportion of 0 counts to the estimated probability of a zero count under the Poisson model. (If n_0 is actually the number expected under the Poisson model, then $\theta = 1$.)

Now consider the actual data reported by MPR for stillbirths in each litter of a sample of 402 litters of laboratory animals.

No. stillbirths	0	1	2	3	4	5	6	7	8	9	10	11
No. litters	314	48	20	7	5	2	2	1	2	0	0	1

For these data, we have $n = 402$, $d = 185$, $\hat{\lambda}_0 = 0.4602$, $\mathrm{e}^{-\hat{\lambda}_0} = 0.6312$, and $\theta = 1.2376$.

What is interesting is the difference in s_I and s_O.

In this particular example, if all n_i for $i \geq 1$ are held constant at the observed values, but different values of n_0 are considered, as n_0 increases the ratio s_I/s_O increases from about 1/4 to 1 (when the n_0 is the expected number under the Poisson model; i.e., $\theta = 1$), and then decreases, actually becoming negative (around $n_0 = 100$).

This example illustrates an interesting case. The score test is inconsistent because the observed information generates negative variance estimates at the MLE under the null hypothesis. (The score test can also be inconsistent if the expected likelihood equation has spurious roots.)

6.8 Nonparametric Tests

Notes

Anomalies of the Score Test

Morgan et al. (2007) give an interesting example in which the score test does not perform as we might expect.

Exercises in Shao

- For practice and discussion
  6.2, 6.3, 6.4, 6.6, 6.10, 6.17, 6.20, 6.29, 6.37, 6.51, 6.52, 6.58, 6.93, 6.98, 6.123 (Solutions in Shao, 2005)
- To turn in
  6.1, 6.5(a),(b)(c), 6.12, 6.21, 6.23, 6.27(a)(b)(c), 6.38, 6.52(a)(b), 6.92(a)(b), 6.99(a)(b)

Additional References

Freedman, David A. (2007), How can the score test be inconsistent? *The American Statistician* **61**, 291–295.

Morgan, B. J. T.; K. J. Palmer; and M. S. Ridout (2007), Negative score test statistic, *The American Statistician* **61**, 285–288.

Verbeke, Geert, and Geert Molenberghs (2007), What can go wrong with the score test? *The American Statistician* **61**, 289–290.

Wald, Abraham (1945). Sequential Tests of Statistical Hypotheses, *The Annals of Mathematical Statistics* **16**, 117–186.

7

Confidence Sets
(Shao Ch 7; TSH3 Ch 3, Ch 5)

For statistical confidence sets, the basic problem is to use a random sample X from an unknown distribution P to determine a random subfamily $A(X)$ of a given family of distributions $\mathcal{P}$ such that

$$\Pr_P(A(X) \ni P) \geq 1-\alpha \quad \forall P \in \mathcal{P}, \tag{7.1}$$

for some given α. The set $A(X)$ is called a $1-\alpha$ *confidence set* or *confidence region*. The "confidence level" is $1-\alpha$, so we sometimes call it a "level $1-\alpha$ confidence set". Notice that α is given a priori. We call

$$\inf_{P\in\mathcal{P}} \Pr_P(A(X) \ni P) \tag{7.2}$$

the *confidence coefficient* of $A(X)$.

If the confidence coefficient of $A(X)$ is $> 1-\alpha$, then $A(X)$ is said to be a *conservative* $1-\alpha$ confidence set.

We generally wish to determine a region with a given confidence coefficient, rather than with a given significance level.

If the distributions are characterized by a parameter θ in a given parameter space Θ an equivalent $1-\alpha$ confidence set for θ is a random subset $S(X)$ such that

$$\Pr_\theta(S(X) \ni \theta) \geq 1-\alpha \quad \forall \theta \in \Theta. \tag{7.3}$$

The basic paradigm of statistical confidence sets was described in Section 2.4.2, beginning on page 108. We first review some of those basic ideas in Section 7.1, starting first with simple interval confidence sets. Then in Section 7.2 we discuss optimality of confidence sets.

As we have seen in other problems in statistical inference, it is often not possible to develop a procedure that is *uniformly* optimal. As with the estimation problem, we can impose restrictions, such as unbiasedness, we discuss in Section 7.2, or equivariance, which we discuss in Section 8.3.

We can define optimality in terms of a global averaging over the family of distributions of interest. If the the global averaging is considered to be a

true probability distribution, then the resulting confidence intervals can be interpreted differently, and it can be said that the probability that the distribution of the observations is in some fixed family is some stated amount. The HPD Bayesian credible regions discussed in Section 3.5 can also be thought of as optimal sets that address similar applications in which confidence sets are used.

Because determining an exact $1-\alpha$ confidence set requires that we know the exact distribution of some statistic, we often have to form approximate confidence sets. There are three common ways that we do this as discussed in Section 2.2.6. In Section 7.3 we discuss asymptotic confidence sets, and in Section 7.4, bootstrap confidence sets.

7.1 Introduction: Construction and Properties

Confidence Intervals

Our usual notion of a confidence interval relies on a frequency approach to probability, and it leads to the definition of a $1-\alpha$ confidence interval for the (scalar) parameter θ as the random interval (T_L, T_U), that has the property

$$\Pr(T_L \leq \theta \leq T_U) = 1-\alpha.$$

This is also called a $(1-\alpha)100\%$ confidence interval. The interval (T_L, T_U) is not uniquely determined.

The concept extends easily to vector-valued parameters. A simple extension would be merely to let T_L and T_U, and let the confidence region be hyperrectangle defined by the cross products of the intervals. Rather than taking vectors T_L and T_U, however, we generally define other types of regions; in particular, we often take an ellipsoidal region whose shape is determined by the covariances of the estimators.

A realization of the random interval, say (t_L, t_U), is also called a confidence interval. Although it may seem natural to state that the "probability that θ is in (t_L, t_U) is $1-\alpha$", this statement can be misleading unless a certain underlying probability structure is assumed.

In practice, the interval is usually specified with respect to an estimator of θ, $T(X)$. If we know the sampling distribution of $T-\theta$, we may determine c_1 and c_2 such that

$$\Pr(c_1 \leq T-\theta \leq c_2) = 1-\alpha;$$

and hence

$$\Pr(T-c_2 \leq \theta \leq T-c_1) = 1-\alpha.$$

If either T_L or T_U is infinite or corresponds to a bound on acceptable values of θ, the confidence interval is one-sided. For two-sided confidence intervals,

we may seek to make the probability on either side of T to be equal. This is called an *equal-tail* confidence interval. We may, rather, choose to make $c_1 = -c_2$, and/or to minimize $|c_2 - c_1|$ or $|c_1|$ or $|c_2|$. This is similar in spirit to seeking an estimator with small variance.

Prediction Sets

We often want to identify a set in which a future observation on a random variable has a high probability of occurring. This kind of set is called a *prediction set*.
For example, we may assume a given sample $X_1, \ldots, X_n$ is from a $\mathrm{N}(\mu, \sigma^2)$ and we wish to determine a measurable set $C(X)$ such that for a future observation X_{n+1}

$$\inf_{P \in \mathcal{P}} \Pr_P(X_{n+1} \in C(X)) \geq 1 - \alpha.$$

More generally, instead of X_{n+1}, we could define a prediction interval for any random variable V.

The difference in this and a confidence set for μ is that there is an additional source of variation. The prediction set will be larger, so as to account for this extra variation.

We may want to separate the statements about V and $S(X)$. A *tolerance set* attempts to do this.

Given a sample X, a measurable set $S(X)$, and numbers δ and α in $(0, 1)$, if

$$\inf_{P \in \mathcal{P}} \left(\inf_{P \in \mathcal{P}} \Pr_P(V \in S(X) | X) \geq \delta \right) \geq 1 - \alpha,$$

then $S(X)$ is called a δ-tolerance set for V with confidence level $1 - \alpha$.

Randomized confidence Sets

For discrete distributions, as we have seen, sometimes to achieve a test of a specified size, we had to use a randomized test.

Confidence sets may have exactly the same problem – and solution – in forming confidence sets for parameters in discrete distributions. We form *randomized confidence sets*. The idea is the same as in randomized tests, and we will discuss randomized confidence sets in the context of hypothesis tests below.

Pivot Functions

A straightforward way to form a confidence interval is to use a function of the sample that also involves the parameter of interest, but that does not involve any nuisance parameters. The confidence interval is then formed by separating the parameter from the sample values.

A class of functions that are particularly useful for forming confidence intervals are called *pivotal* values, or pivotal functions. A function $f(T,\theta)$ is said to be a pivotal function if its distribution does not depend on any unknown parameters. This allows exact confidence intervals to be formed for the parameter θ. We first form

$$\Pr\Big(f_{(\alpha/2)} \;\le\; f(T,\theta) \;\le\; f_{(1-\alpha/2)}\Big) \;=\; 1-\alpha,$$

where $f_{(\alpha/2)}$ and $f_{(1-\alpha/2)}$ are quantiles of the distribution of $f(T,\theta)$; that is,

$$\Pr(f(T,\theta) \le f_{(\pi)}) = \pi.$$

If, as in the case considered above, $f(T,\theta) = T-\theta$, the resulting confidence interval has the form

$$\Pr\Big(T - f_{(1-\alpha/2)} \;\le\; \theta \;\le\; T - f_{(\alpha/2)}\Big) \;=\; 1-\alpha.$$

For example, suppose $Y_1, Y_2, \ldots, Y_n$ is a random sample from a $\mathrm{N}(\mu,\sigma^2)$ distribution, and $\overline{Y}$ is the sample mean. The quantity

$$f(\overline{Y},\mu) = \frac{\sqrt{n(n-1)}\,(\overline{Y}-\mu)}{\sqrt{\sum\;\big(Y_i-\overline{Y}\big)^2}}$$

has a Student's t distribution with $n-1$ degrees of freedom, no matter what is the value of σ^2. This is one of the most commonly-used pivotal values.

The pivotal value can be used to form a confidence interval for θ by first writing

$$\Pr\big(t_{(\alpha/2)} \;\le\; f(\overline{Y},\mu) \;\le\; t_{(1-\alpha/2)}\big) \;=\; 1-\alpha,$$

where $t_{(\pi)}$ is a percentile from the Student's t distribution. Then, after making substitutions for $f(\overline{Y},\mu)$, we form the familiar confidence interval for μ:

$$\Big(\overline{Y} - t_{(1-\alpha/2)}\, s/\sqrt{n}, \quad \overline{Y} - t_{(\alpha/2)}\, s/\sqrt{n}\Big),$$

where s^2 is the usual sample variance, $\sum(Y_i-\overline{Y})^2/(n-1)$.

Other similar pivotal values have F distributions. For example, consider the usual linear regression model in which the n-vector random variable Y has a $\mathrm{N}_n(X\beta,\sigma^2 I)$ distribution, where X is an $n\times m$ known matrix, and the m-vector β and the scalar σ^2 are unknown. A pivotal value useful in making inferences about β is

$$g(\widehat{\beta},\beta) = \frac{\big(X(\widehat{\beta}-\beta)\big)^{\mathrm{T}} X(\widehat{\beta}-\beta)/m}{(Y-X\widehat{\beta})^{\mathrm{T}}(Y-X\widehat{\beta})/(n-m)},$$

where

$$\widehat{\beta} = (X^{\mathrm{T}}X)^{+}X^{\mathrm{T}}Y.$$

The random variable $g(\widehat{\beta}, \beta)$ for any finite value of σ^2 has an F distribution with m and $n - m$ degrees of freedom.

For a given parameter and family of distributions there may be multiple pivotal values. For purposes of statistical inference, such considerations as unbiasedness and minimum variance may guide the choice of a pivotal value to use.

Approximate Pivot Values

It may not be possible to identify a pivotal quantity for a particular parameter. In that case, we may seek an approximate pivot. A function is asymptotically pivotal if a sequence of linear transformations of the function is pivotal in the limit as $n \to \infty$.

*** nuisance parameters ***** find consistent estimator

If the distribution of T is known, c_1 and c_2 can be determined. If the distribution of T is not known, some other approach must be used. A common method is to use some numerical approximation to the distribution. Another method is to use bootstrap samples from the ECDF.

Relation to Acceptance Regions of Hypothesis Tests

A test at the α level has a very close relationship with a $1-\alpha$ level confidence set.

When we test the hypothesis $H_0 : \theta \in \Theta_{H_0}$ at the α level, we form a critical region for a test statistic or rejection region for the values of the observable X. This region is such that the probability that the test statistic is in it is $\leq \alpha$.

For any given $\theta_0 \in \Theta$, consider the nonrandomized test T_{θ_0} for testing the simple hypothesis $H_0 : \theta = \theta_0$, against some alternative H_1. We let $A(\theta_0)$ be the set of all x such that the test statistic is not in the critical region; that is, $A(\theta_0)$ is the acceptance region.

Now, for any θ and any value x in the range of X, we let

$$C(x) = \{\theta : x \in A(\theta)\}.$$

For testing $H_0 : \theta = \theta_0$ at the α significance level, we have

$$\sup \Pr(X \notin A(\theta_0) \,|\, \theta = \theta_0) \leq \alpha;$$

that is,

$$1 - \alpha \leq \inf \Pr(X \in A(\theta_0) \,|\, \theta = \theta_0) = \inf \Pr(C(X) \ni \theta_0 \,|\, \theta = \theta_0).$$

This holds for any θ_0, so

$$\begin{aligned} \inf_{P \in \mathcal{P}} \Pr_P(C(X) \ni \theta) &= \inf_{\theta_0 \in \Theta} \inf \Pr_P(C(X) \ni \theta_0 \,|\, \theta = \theta_0) \\ &\geq 1 - \alpha. \end{aligned}$$

Hence, $C(X)$ is a $1-\alpha$ level confidence set for θ.

If the size of the test is α, the inequalities are equalities, and so the confidence coefficient is $1-\alpha$.

For example, suppose $Y_1, Y_2, \ldots, Y_n$ is a random sample from a $\mathrm{N}(\mu, \sigma^2)$ distribution, and $\overline{Y}$ is the sample mean.

To test $H_0 : \mu = \mu_0$, against the universal alternative, we form the test statistic

$$T(X) = \frac{\sqrt{n(n-1)}\,(\overline{Y} - \mu_0)}{\sqrt{\sum \left(Y_i - \overline{Y}\right)^2}}$$

which, under the null hypothesis, has a Student's t distribution with $n-1$ degrees of freedom.

An acceptance region at the α level is

$$\left(t_{(\alpha/2)},\ t_{(1-\alpha/2)}\right),$$

and hence, putting these limits on $T(X)$ and inverting, we get

$$\left(\overline{Y} - t_{(1-\alpha/2)}\, s/\sqrt{n}, \quad \overline{Y} - t_{(\alpha/2)}\, s/\sqrt{n}\right),$$

which is a $1-\alpha$ level confidence interval.

The test has size α and so the confidence coefficient is $1-\alpha$.

Randomized confidence Sets

To form a $1-\alpha$ confidence level set, we form a nonrandomized confidence set (which may be null) with $1-\alpha_1$ confidence level, with $0 \leq \alpha_1 \leq \alpha$, and then we define a random experiment with some event that has a probability of $\alpha - \alpha_1$.

7.2 Optimal Confidence Sets

We often evaluate a confidence set using a family of distributions that *does not include the true parameter*.

For example, "accuracy" is the (true) probability of the set including an incorrect value.

The "volume" (or "length") of a confidence set is the Lebesgue measure of the set:

$$\mathrm{vol}(C(x)) = \int_{C(x)} \mathrm{d}\tilde{\theta}.$$

This may not be finite.

If the volume is finite, we have (Theorem 7.6 in Shao)

$$\mathrm{E}_\theta(\mathrm{vol}(C(x))) = \int_{\theta \neq \tilde{\theta}} \Pr_\theta(C(x) \ni \tilde{\theta}) \mathrm{d}\tilde{\theta}.$$

We see this by a simple application of Fubini's theorem to handle the integral over the product space, and then an interchange of integration:

Want to minimize volume (if appropriate; i.e., finite.)

Want to maximize accuracy.

Uniformly most accurate $1-\alpha$ level set:
$\Pr_\theta(C(X) \ni \tilde{\theta})$ is minimum among all $1-\alpha$ level sets and $\forall \tilde{\theta} \neq \theta$.

This definition of UMA may not be so relevant in the case of a one-sided confidence interval.

If $\widetilde{\Theta}$ is a subset of Θ that does not include θ, and

$$\Pr_\theta(C(X) \ni \tilde{\theta}) \leq \Pr_\theta(C_1(X) \ni \tilde{\theta})$$

for any $1-\alpha$ level set $C_1(X)$ and $\forall \tilde{\theta} \in \widetilde{\Theta}$, then $C(X)$ is said to be $\widetilde{\Theta}$-uniformly most accurate.

A confidence set formed by inverting a nonrandomized UMP test is UMA.

We see this easily from the definitions of UMP and UMA. (This is Theorem 7.4 in Shao.)

With tests, sometimes no UMP exists, and hence we added a criterion, such as unbiasedness or invariance.

Likewise, sometimes we cannot form a UMA confidence interval, so we add some criterion.

We define unbiasedness in terms of a subset $\widetilde{\Theta}$ that does not include the true θ.

A $1-\alpha$ level confidence set $C(X)$ is said to be $\widetilde{\Theta}$-unbiased if

$$\Pr_\theta(C(X) \ni \tilde{\theta}) \leq 1-\alpha \quad \forall \tilde{\theta} \in \widetilde{\Theta}.$$

If $\widetilde{\Theta} = \{\theta\}^c$, we call the set *unbiased*.

A $\widetilde{\Theta}$-unbiased set that is uniformly more accurate ("more" is defined similarly to "most") than any other $\widetilde{\Theta}$-unbiased set is said to be a *uniformly most accurate unbiased* (UMAU) set.

Accuracy of Confidence Regions

Confidence regions can be thought of a a family of tests of hypotheses of the form $\theta \in H_0(\tilde{\theta})$ versus $\theta \in H_1(\tilde{\theta})$. A confidence region of size $1-\alpha$ is equivalent to a critical region $S(X)$ such that

$$\Pr\left(S(X) \ni \tilde{\theta}\right) \geq 1-\alpha \quad \forall\, \theta \in H_0\left(\tilde{\theta}\right).$$

The power of the related tests is just

$$\Pr\left(S(X) \ni \tilde{\theta}\right)$$

for any θ. In testing hypotheses, we are concerned about maximizing this for $\theta \in H_1(\tilde{\theta})$.

This is called the *accuracy* of the confidence region, and so in this terminology, we seek the *most accurate* confidence region, and, of course, the *uniformly most accurate* confidence refion. Similarly to the case of UMP tests, the uniformly most accurate confidence refion may or may not exist.

The question of existence of uniformly most accurate confidence intervals also depends on whether or not there are nuisance parameters. Just as with UMP tests, in the presence of nuisance parameters, usually uniformly most accurate confidence intervals do not exist. (We must add other restrictions on the intervals, as we see below.) The nonexistence of uniformly most accurate confidence regions can also be addressed by imposing unbiasedness.

The concept of unbiasedness in tests carries over immediately to confidence regions. A family of confidence regions of size $1-\alpha$ is said to be *unbiased* if

$$\Pr\left(S(X) \ni \tilde{\theta}\right) \leq 1-\alpha \quad \forall\, \theta \in H_1\left(\tilde{\theta}\right).$$

In the case of nuisance parameters θ_u, unbiasedness means that this holds for all values of the nuisance parameters. In this case, similar regions and Neyman structure also are relevant, just as in the case of testing.

Volume of a Confidence Set

If there are no nuisance parameters, the expected volume of a confidence set is usually known a priori, e.g., for μ in $\mathrm{N}(\mu, 1)$.

What about a confidence set for μ in $\mathrm{N}(\mu, \sigma^2)$, with σ^2 unknown?

The expected length is proportional to σ, and can be very long. (This is a consequence of the fact that two normal distributions $\mathrm{N}(\mu_1, \sigma^2)$ and $\mathrm{N}(\mu_2, \sigma^2)$ become indistinguishable as $\sigma \to \infty$.

The length of the considence interval is inversely proportional to $\sqrt{n}$. How about a sequential procedure?

A Sequential Procedure for a Confidence Set

Let $X_1, X_2, \ldots$ be i.i.d. from $\mathrm{N}(\mu, \sigma^2)$.

Fix n_0. Let $\bar{x}_0 = \sum_{i=1}^{n_0} x_i/n_0$ and $s_0^2 = \sum_{i=1}^{n_0}(x_i - \bar{x}_0)^2/(n_0 - 1)$.

Now, for measurable function of s, a and b, for $n \geq n_0$, let

$$a_1 = \cdots = a_{n_0} = a$$

and

$$a_{n_0+1} = \cdots = a_n = b.$$

Then

$$Y = \frac{\sum_{i=1}^{n} a_i(X_i - \mu)}{\sqrt{S_0^2 \sum_{i=1}^{n} a_i^2}}$$

has a Student's t distribution with $n_0 - 1$ df.

Controlling the Volume

Now compute s^2 from an initial sample of size n_0. Let c be a given positive constant. Now choose $n - n_0$ additional observations where

$$n = \max\left(n_0 + 1, \left[\frac{s^2}{c}\right] + 1\right).$$

Then there exists numbers $a_1, \ldots, a_n$ with $a_1 = \cdots = a_{n_0}$ and $a_{n_0+1} = \cdots = a_n$ such that $\sum_{i=1}^n a_i = 1$ and $\sum_{i=1}^n a_i^2 = c/s^2$.

And so (from above), $\sum_{i=1}^n a_i(X_i - \mu)/\sqrt{c}$ has a Student's t distribution with $n_0 - 1$ df.

Therefore, given $X_1, \ldots, X_{n_0}$ the expected length of the confidence interval can be controlled.

An Example of a Confidence Interval in Regression

Consider $\mathrm{E}(Y|x) = \beta_0 + \beta_1 x$. Want to estimate the point at which $\beta_0 + \beta_1 x$ has a preassigned value; for example find dosage $x = -\beta_0/\beta_1$ at which $\mathrm{E}(Y|x) = 0$.

This is equivalent to finding the value $v = (x - \bar{x})/\sqrt{\sum(x_i - \bar{x})^2}$ at which

$$y = \gamma_0 + \gamma_1 v = 0.$$

So we want to find $v = -\gamma_0/\gamma_1$.

The most accurate unbiased confidence sets for $-\gamma_0/\gamma_1$ can be obtained from UMPU tests of the hypothesis $-\gamma_0/\gamma_1 = v_0$.

Acceptance regions of these tests are given by

$$\frac{|v_0 \sum v_i Y_i + \bar{Y}|/\sqrt{\frac{1}{n} + v_0^2}}{\sqrt{\left(\sum(Y_i - \bar{Y})^2 - (\sum v_i Y_i)^2\right)/(n-2)}} \leq c$$

where

$$\int_{-c}^{c} p(y)\mathrm{d}y = 1 - \alpha,$$

where p is the PDF of t with $n - 2$ df.

So square and get quadratic inequalities in v:

$$v^2\left(c^2 s^2 - (\sum v_i Y_i)^2\right) - 2v\bar{Y}\sum v_i Y_i + \frac{1}{n}(c^2 x^2 - n\bar{Y}) \geq 0.$$

Now let $\underline{v}$ and $\overline{v}$ be the roots of the equation.

So the confidence statement becomes

$$\underline{v} \leq \frac{\gamma_0}{\gamma_1} \leq \overline{v} \quad \text{if} \quad \frac{|\sum v_i Y_i|}{s} > c,$$

$$\frac{\gamma_0}{\gamma_1} < \underline{v} \quad \text{or} \quad \frac{\gamma_0}{\gamma_1} > \overline{v} \quad \text{if} \quad \frac{|\sum v_i Y_i|}{s} < c,$$

and if $= c$, no solution.

If $y = \gamma_0 + \gamma_1 v$ is nearly parallel to the v-axis, then the intercept with the v-axis will be large in absolute value and its sign is sensitive to a small change in the angle.

Suppose in the quadratic that $n\bar{Y}^2 + (\sum v_i Y_i)^2 < c^2 s^2$, then there is no real solution.

For the confidence levels to remain valid, the confidence interval must be the whole real line.

7.3 Asymptotic Confidence Sets

It is often difficult to determine sets with a specified confidence coefficient or significance level, or with other specified properties.

In such cases it may be useful to determine a set that "approximately" meets the specified requirements.

What does "approximately" mean?

- uses numerical approximations
- uses approximate distributions
- uses a random procedure
- uses asymptotics

Asymptotic Confidence Sets

We assume a random sample $X_1, \ldots, X_n$ from $P \in \mathcal{P}$

An *asymptotic significance level* of a confidence set $C(X)$ for $g(\theta)$ is $1 - \alpha$ if

$$\liminf_n \Pr(C(X) \ni \theta) \geq 1 - \alpha \quad \text{for any } P \in \mathcal{P}.$$

The *limiting confidence coefficient* of a confidence set $C(X)$ for θ is

$$\lim_n \inf_{P \in \mathcal{P}} \Pr(C(X) \ni \theta)$$

if it exists.

Example (Shao). Suppose $X_1, \ldots, X - n$ are i.i.d. from a distribution with CDF P_X and finite mean μ and variance σ^2. Suppose σ^2 is known, and we want to form a $1 - \alpha$ level confidence interval for μ. Unless P_X is specified, we can only seek a confidence interval with asymptotic significance level $1 - \alpha$. We have an asymptotic pivot $T(X, \mu) = (\overline{X} - \mu)/\sigma$, and $\sqrt{n}T$ has an asymptotic $\mathrm{N}(0, 1)$ distribution. We then form an interval

$$\begin{aligned} C(X) &= (C_1(X),\ C_2(X)) \\ &= (\overline{X} - \sigma z_{1-\alpha/2}/\sqrt{:}n,\ \overline{X} + \sigma z_{1-\alpha/2}/\sqrt{:}n), \end{aligned}$$

where $z_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$ and Φ is the $\mathrm{N}(0,1)$ CDF. Now consider $\Pr(\mu \in C(X))$. We have

Asymptotic Correctness

A confidence set $C(X)$ for θ is $1-\alpha$ *asymptotically correct* if

$$\lim_n \Pr(C(X) \ni \theta) = 1 - \alpha.$$

A confidence set $C(X)$ for θ is $1-\alpha$ l^{th}*-order asymptotically correct* if it is $1-\alpha$ asymptotically correct and

$$\lim_n \Pr(C(X) \ni \theta) = 1 - \alpha + \mathrm{O}(n^{-l/2}).$$

Asymptotic Accuracy of Confidence Regions

****

Constructing Asymptotic Confidence Sets

There are two straightforward ways of constructing good asymptotic confidence sets.

One is based on an *asymptotically pivotal function*, that is one whose limiting distribution does not depend on any parameters other than the one of the confidence set.

Another method is to invert the acceptance region of a test. The properties of the test carry over to the confidence set.

The likelihood yields statistics with good asymptotic properties (for testing or for confidence sets).

See Example 7.24.

Woodruff's interval

7.4 Bootstrap Confidence Sets

A method of forming a confidence interval for a parameter θ is to find a pivotal quantity that involves θ and a statistic T, $f(T,\theta)$, and then to rearrange the terms in a probability statement of the form

$$\Pr\left(f_{(\alpha/2)} \leq f(T,\theta) \leq f_{(1-\alpha/2)}\right) = 1-\alpha. \tag{7.4}$$

When distributions are difficult to work out, we may use bootstrap methods for estimating and/or approximating the percentiles, $f_{(\alpha/2)}$ and $f_{(1-\alpha/2)}$.

Basic Intervals

For computing confidence intervals for a mean, the pivotal quantity is likely to be of the form $T - \theta$.

The simplest application of the bootstrap to forming a confidence interval is to use the sampling distribution of $T^* - T_0$ as an approximation to the sampling distribution of $T - \theta$; that is, instead of using $f(T, \theta)$, we use $f(T^*, T_0)$, where T_0 is the value of T in the given sample.

The percentiles of the sampling distribution determine $f_{(\alpha/2)}$ and $f_{(1-\alpha/2)}$ in

$$\Pr\left(f_{(\alpha/2)} \leq f(T, \theta) \leq f_{(1-\alpha/2)}\right) = 1 - \alpha.$$

Monte Carlo to Get Basic Intervals

If we cannot determine the sampling distribution of $T^* - T$, we can easily estimate it by Monte Carlo methods.

For the case $f(T, \theta) = T - \theta$, the probability statement above is equivalent to

$$\Pr\left(T - f_{(1-\alpha/2)} \leq \theta \leq T - f_{(\alpha/2)}\right) = 1 - \alpha. \tag{7.5}$$

The $f_{(\pi)}$ may be estimated from the percentiles of a Monte Carlo sample of $T^* - T_0$.

Bootstrap-t Intervals

Methods of inference based on a normal distribution often work well even when the underlying distribution is not normal.

A useful approximate confidence interval for a location parameter can often be constructed using as a template the familiar confidence interval for the mean of a normal distribution,

$$\left(\overline{Y} - t_{(1-\alpha/2)}\, s/\sqrt{n}, \quad \overline{Y} - t_{(\alpha/2)}\, s/\sqrt{n}\right),$$

where $t_{(\pi)}$ is a percentile from the Student's t distribution, and s^2 is the usual sample variance.

A confidence interval for any parameter constructed in this pattern is called a *bootstrap-t interval*. A bootstrap-t interval has the form

$$\left(T - \widehat{t}_{(1-\alpha/2)}\sqrt{\widehat{\mathrm{V}}(T)}, \quad T - \widehat{t}_{(\alpha/2)}\sqrt{\widehat{\mathrm{V}}(T)}\right). \tag{7.6}$$

In the interval

$$\left(T - \widehat{t}_{(1-\alpha/2)}\sqrt{\widehat{\mathrm{V}}(T)}, \quad T - \widehat{t}_{(\alpha/2)}\sqrt{\widehat{\mathrm{V}}(T)}\right)$$

$\widehat{t}_{(\pi)}$ is the estimated percentile from the studentized statistic,

$$\frac{T^* - T_0}{\sqrt{\widehat{\mathrm{V}}(T^*)}}.$$

For many estimators T, no simple expression is available for $\widehat{\mathrm{V}}(T)$.

The variance could be estimated using a bootstrap. This bootstrap nested in the bootstrap to determine $\widehat{t}_{(\pi)}$ increases the computational burden multiplicatively.

If the underlying distribution is normal and T is a sample mean, the interval in expression (7.6) is an exact $(1-\alpha)100\%$ confidence interval of shortest length.

If the underlying distribution is not normal, however, this confidence interval may not have good properties. In particular, it may not even be of size $(1-\alpha)100\%$. An asymmetric underlying distribution can have particularly deleterious effects on one-sided confidence intervals.

If the estimators T and $\widehat{\mathrm{V}}(T)$ are based on sums of squares of deviations, the bootstrap-t interval performs very poorly when the underlying distribution has heavy tails. This is to be expected, of course. Bootstrap procedures can be no better than the statistics used.

Bootstrap Percentile Confidence Intervals

Given a random sample $(y_1, \ldots, y_n)$ from an unknown distribution with CDF P, we want an interval estimate of a parameter, $\theta = \Theta(P)$, for which we have a point estimator, T.

If T^* is a bootstrap estimator for θ based on the bootstrap sample $(y_1^*, \ldots, y_n^*)$, and if $G_{T^*}(t)$ is the distribution function for T^*, then the exact upper $1-\alpha$ confidence limit for θ is the value $t^*_{(1-\alpha)}$, such that $G_{T^*}(t^*_{(1-\alpha)}) = 1-\alpha$.

This is called the *percentile upper confidence limit*.

A lower limit is obtained similarly, and an interval is based on the lower and upper limits.

Monte Carlo for Bootstrap Percentile Confidence Intervals

In practice, we generally use Monte Carlo and m bootstrap samples to estimate these quantities.

The probability-symmetric bootstrap percentile confidence interval of size $(1-\alpha)100\%$ is thus

$$\left(t^*_{(\alpha/2)}, \quad t^*_{(1-\alpha/2)}\right),$$

where $t^*_{(\pi)}$ is the $[\pi m]^{\text{th}}$ order statistic of a sample of size m of T^*.

(Note that we are using T and t, and hence T^* and t^*, to represent estimators and estimates in general; that is, $t^*_{(\pi)}$ here does not refer to a percentile of the Student's t distribution.)

This percentile interval is based on the ideal bootstrap and may be estimated by Monte Carlo simulation.

Confidence Intervals Based on Transformations

Suppose that there is a monotonically increasing transformation g and a constant c such that the random variable

$$W = c(g(T^*) - g(\theta)) \tag{7.7}$$

has a symmetric distribution about zero. Here $g(\theta)$ is in the role of a mean and c is a scale or standard deviation.

Let H be the distribution function of W, so

$$G_{T^*}(t) = H\left(c(g(t) - g(\theta))\right) \tag{7.8}$$

and

$$t^*_{(1-\alpha/2)} = g^{-1}\left(g(t^*) + w_{(1-\alpha/2)}/c\right), \tag{7.9}$$

where $w_{(1-\alpha/2)}$ is the $(1-\alpha/2)$ quantile of W. The other quantile $t^*_{(\alpha/2)}$ would be determined analogously.

Instead of approximating the ideal interval with a Monte Carlo sample, we could use a transformation to a known W and compute the interval that way. Use of an exact transformation g to a known random variable W, of course, is just as difficult as evaluation of the ideal bootstrap interval. Nevertheless, we see that forming the ideal bootstrap confidence interval is equivalent to using the transformation g and the distribution function H.

Because transformations to approximate normality are well-understood and widely used, in practice, we generally choose g as a transformation to normality. The random variable W above is a standard normal random variable, Z. The relevant distribution function is Φ, the normal CDF. The normal approximations have a basis in the central limit property. Central limit approximations often have a bias of order $\mathrm{O}(n^{-1})$, however, so in small samples, the percentile intervals may not be very good.

Bias in Intervals Due to Bias in the Estimator

It is likely that the transformed statistic $g(T^*)$ in equation (7.7) is biased for the transformed θ, even if the untransformed statistic is unbiased for θ.

We can account for the possible bias by using the transformation

$$Z = c(g(T^*) - g(\theta)) + z_0,$$

and, analogous to equation (7.8), we have

$$G_{T^*}(t) = \Phi\left(c(g(t) - g(\theta)) + z_0\right).$$

The bias correction z_0 is $\Phi^{-1}\left(G_{T^*}(t)\right)$.

Bias in Intervals Due to Lack of Symmetry

Even when we are estimating θ directly with T^* (that is, g is the identity), another possible problem in determining percentiles for the confidence interval is the lack of symmetry of the distribution about z_0.

We would therefore need to make some adjustments in the quantiles instead of using equation (7.9) without some correction.

Correcting the Bias in Intervals

Rather than correcting the quantiles directly, we may adjust their levels.

For an interval of confidence $(1-\alpha)$, instead of $(t^*_{(\alpha/2)},\ \ t^*_{(1-\alpha/2)})$, we take

$$\left(t^*_{(\alpha_1)},\quad t^*_{(\alpha_2)}\right),$$

where the adjusted probabilities α_1 and α_2 are determined so as to reduce the bias and to allow for the lack of symmetry.

As we often do, even for a nonnormal underlying distribution, we relate α_1 and α_2 to percentiles of the normal distribution.

To allow for the lack of symmetry—that is, for a scale difference below and above z_0—we use quantiles about that point.

Efron (1987), who developed this method, introduced an "acceleration", a, and used the distance $a(z_0 + z_{(\pi)})$.

Using values for the bias correction and the acceleration determined from the data, Efron suggested the quantile adjustments

$$\alpha_1 = \Phi\left(\widehat{z}_0 + \frac{\widehat{z}_0 + z_{(\alpha/2)}}{1 - \widehat{a}(\widehat{z}_0 + z_{(\alpha/2)})}\right)$$

and

$$\alpha_2 = \Phi\left(\widehat{z}_0 + \frac{\widehat{z}_0 + z_{(1-\alpha/2)}}{1 - \widehat{a}(\widehat{z}_0 + z_{(1-\alpha/2)})}\right).$$

Use of these adjustments to the level of the quantiles for confidence intervals is called the bias-corrected and accelerated, or "BC_a", method.

This method automatically takes care of the problems of bias or asymmetry resulting from transformations that we discussed above.

Note that if $\widehat{a} = \widehat{z}_0 = 0$, then $\alpha_1 = \Phi(z_{(\alpha)})$ and $\alpha_2 = \Phi(z_{(1-\alpha)})$. In this case, the BC_a is the same as the ordinary percentile method.

The problem now is to estimate the bias correction z_0 and the acceleration a from the data.

Estimating the Correction

The bias-correction term z_0 is estimated by correcting the percentile near the median of the m bootstrap samples:

$$\widehat{z}_0 = \Phi^{-1}\left(\frac{1}{m}\sum_j \mathrm{I}_{(-\infty,T]}\left(T^{*j}\right)\right).$$

The idea is that we approximate the bias of the median (that is, the bias of a central quantile) and then adjust the other quantiles accordingly.

Estimating the Acceleration

Estimating a is a little more difficult. The way we proceed depends on the form the bias may take and how we choose to represent it.

Because one cause of bias may be skewness, Efron (1987) adjusted for the skewness of the distribution of the estimator in the neighborhood of θ.

The skewness is measured by a function of the second and third moments of T.

We can use the jackknife to estimate the second and third moments of T. The expression is

$$\widehat{a} = \frac{\sum\left(\mathrm{J}(T) - T_{(i)}\right)^3}{6\left(\sum\left(\mathrm{J}(T) - T_{(i)}\right)^2\right)^{3/2}}. \tag{7.10}$$

Bias resulting from other departures from normality, such as heavy tails, is not addressed by this adjustment.

There are R and S-Plus functions to compute BC_a confidence intervals.

Comparisons of Bootstrap-t and BC_a Intervals

It is difficult to make analytic comparisons between these two types of bootstrap confidence intervals.

In some Monte Carlo studies, it has been found that, for moderate and approximately equal sample sizes,the coverage of BC_a intervals is closest to the nominal confidence level; however, for samples with very different sizes, the bootstrap-t intervals were better in the sense of coverage frequency.

Because of the variance of the components in the BC_a method, it generally requires relatively large numbers of bootstrap samples. For location parameters, for example, we may need $m = 1,000$.

Other Bootstrap Confidence Intervals

Another method for bootstrap confidence intervals is based on a delta method approximation for the standard deviation of the estimator.

This method yields *approximate bootstrap confidence*, or ABC, intervals.

Terms in the Taylor series expansions are used for computing $\widehat{a}$ and $\widehat{z}_0$ rather than using bootstrap estimates for these terms.

As with the BC_a method, bias resulting from other departures from normality, such as heavy tails, is not addressed.

There are R and S-Plus functions to compute ABC confidence intervals.
****

7.5 Simultaneous Confidence Sets

If $\theta = (\theta_1, \theta_2)$ a $1-\alpha$ level confidence set for θ is a region in $\mathbb{R}^2$, $C(X)$, such that $\Pr_\theta(C(X) \ni \theta) \geq 1-\alpha$.

Now consider the problem of separate intervals (or sets) in $\mathbb{R}^1$, $C_(X)$ and $C_2(X)$, such that $\Pr_\theta(C_1(X) \ni \theta_1 \quad \text{and} \quad C_2(X) \ni \theta_2) \geq 1-\alpha$.

These are called $1-\alpha$ simultaneous confidence intervals.

This is equivalent to $C(X) = C_1(X) \times C_2(X)$ in the case above. Or, in general $\times C_i(X)$.

(Of course, we may want to minimize expected area or some other geometric measures of $C(X)$.)

There are several methods. In linear models, many methods depend on contrasts, e.g., Scheffé's intervals or Tukey's intervals.

General conservative procedures depend depend on inequalities of probabilities of events.

7.5.1 Bonferroni's Confidence Intervals

A common conservative procedure called a Bonferroni method is based on the inequality

$$\Pr\left(\cup A_i\right) \leq \sum \Pr\left(A_i\right),$$

for any events $A_1, \ldots, A_k$. For each component of θ, θ_t, we choose α_t with $\sum \alpha_t = \alpha$, and we let $C_t(X)$ be a level $1-\alpha_t$ confidence interval. It is easy to see that these are of level $1-\alpha$ because

$$\begin{aligned}
\inf \Pr(C_t(X) \ni \theta_t \, \forall t) &= \Pr(\cap\{C_t(X) \ni \theta_t\}) \\
&= 1 - \Pr\left((\cap\{\theta_t \in C_t(X)\})^c\right) \\
&= 1 - \Pr\left(\cup\{\theta_t \notin C_t(X)\}\right) \\
&\geq 1 - \sum \Pr\left(\{\theta_t \notin C_t(X)\}\right) \\
&\geq 1 - \sum \alpha_t \\
&= 1 - \alpha.
\end{aligned}$$

7.5.2 Scheffé's Confidence Intervals

7.5.3 Tukey's Confidence Intervals

Notes

Exercises in Shao

- For practice and discussion
  7.9, 7.18, 7.29, 7.31, 7.48, 7.60, 7.63, 7.67 (Solutions in Shao, 2005)
- To turn in
  7.1, 7.2, 7.22(a), 7.22(b), 7.44, 7.79, 7.82, 7.93, 7.95, 7.101

8

Equivariant Statistical Procedures (Shao Sec 4.2, Sec 6.3; TPE2 Ch 3; TSH3 Ch 6)

8.1 Transformations

Statistical decisions or actions based on data should not be affected by simple transformations on the data or by reordering of the data, so long as these changes on the data are reflected in the statement of the decision; that is, the actions should be *invariant*. If the action is a yes-no decision, such as in hypothesis testing, it should be completely invariant. If a decision is a point estimate, its value is not unaffected, but it should be *equivariant*, in the sense that it reflects the transformations in a meaningful way.

We can formalize this principle by defining appropriate classes of transformations, and then specifying rules that statistical decision functions must satisfy. We identify "reasonable" classes of transformations on the sample space and the corresponding transformations on other components of the statistical decision problem. We will limit consideration to transformations that are one-to-one and onto. Such transformations can easily be identified as members of a group, which I will define more precisely below.

In this section we will consider only parametric inference; that is, we will consider distributions $P_{X|\theta}$ for $\theta \in \Theta$. We are interested in what happens under a transformation of the random variable $g(X)$; in particular, is there a transformation of the parameter $\bar{g}(\theta)$ such that $P_{g(X)|\bar{g}(\theta)}$ is a member of the same distributional family, and will the same optimal methods of inference for $P_{X|\theta}$ remain optimal for $P_{g(X)|\bar{g}(\theta)}$.

First, consider two simple cases.

If the conditional distribution of $X - \theta$, given $\theta = \theta_0$, is the same for all $\theta_0 \in \Theta$, then θ is called a *location parameter*. The family of distributions $P_{X|\theta}$ is called a *location family*.

If $\Theta \subset \mathbb{R}_+$, and if the conditional distribution of X/θ, given $\theta = \theta_0$, is the same for all $\theta_0 \in \Theta$, then θ is called a *scale parameter*. The family of distributions $P_{X|\theta}$ is called a *scale family*.

More generally, given a distribution with parameter θ, that distribution together with a *group* of transformations on θ forms a "group family" of distributions.

8.1.1 Transformation Groups

Definition. A *group* $\mathcal{G}$ is a nonempty set G together with a binary operation $\circ$ such that

- $g_1, g_2 \in G \Rightarrow g_1 \circ g_2 \in G$ (closure);
- $\exists\, e \in G \ni \forall\, g \in G,\ e \circ g = g$ (identity);
- $\forall\, g \in G\ \exists\, g^{-1} \in G \ni g^{-1} \circ g = e$ (inverse);
- $g_1, g_2, g_3 \in G \Rightarrow g_1 \circ (g_2 \circ g_3) = (g_1 \circ g_2) \circ g_3$ (associativity).

Notice that the binary operation need not be commutative, but from these defining properties, we can easily see that $g \circ e = e \circ g$ and $g \circ g^{-1} = g^{-1} \circ g$. (We see, for example, that $g \circ e = e \circ g$ by writing $g \circ e = g \circ (e \circ e) = g \circ (e \circ (g^{-1} \circ g)) = g \circ ((e \circ g^{-1}) \circ g) = g \circ (g^{-1} \circ g) = (g \circ g^{-1}) \circ g) = e \circ g$.) We can also see that e is unique, and for a given g, g^{-1} is unique.

A group $\mathcal{G}$ is a structure of the form $(G, \circ)$. (Sometimes the same symbol that is used to refer to the set is used to refer to the group. The expression $g \in \mathcal{G}$ is interpreted to mean $g \in G$.) Any subset of the set on which the group is defined that is closed and contains the identity and all inverses forms a group with the same operation as the original group. This subset together with the operation is called a *subgroup*. We use the standard terminology of set operations for operations on groups.

A set G_1 together with an operation $\circ$ defined on G_1 *generates* a group $\mathcal{G}$ that is the smallest group $(G, \circ)$ such that $G_1 \subset G$. If $\mathcal{G}_1$ and $\mathcal{G}_2$ are groups over G_1 and G_2 with a common operation $\circ$, the group generated by G_1 and G_2 is $(G, \circ)$, where G is the smallest set containing G_1 and G_2 so that $(G, \circ)$ is a group. Notice that the G may contain elements that are in neither G_1 nor G_2.

If the elements of a set are transformations, function composition is a binary operation. A set of one-to-one and onto functions with common domain together with the operation of function composition is a group, referred to as a transformation group. A transformation group has an associated set that is the common domain of the transformations. This is called the domain of the transformation group.

Both function composition and a function-argument pair are often indicated by juxtaposition with no symbol for the operator or operation. For example, the expression $g^*Tg^{-1}X$ means function composition on the argument X. Remember what the arguments of these individual functions are. We can write

$$\begin{aligned} g^*(T(g^{-1}(\tilde{X}))) &= g^*(T(X)) && (8.1)\\ &= g^*(a) && (8.2)\\ &= \tilde{a}. && (8.3) \end{aligned}$$

Invariant Functions

A function f is said to be *invariant* under the transformation group $\mathcal{G}$ with domain $\mathcal{X}$ if for all $x \in \mathcal{X}$ and $g \in \mathcal{G}$,

$$f(g(x)) = f(x). \tag{8.4}$$

We also use the phrases "invariant over ..." and "invariant with respect to ..." to denote this kind of invariance.

A transformation group $\mathcal{G}$ defines an equivalence relation (identity, symmetry, and transitivity) for elements in its domain, $\mathcal{X}$. If $x_1, x_2 \in \mathcal{X}$ and there exists a g in $\mathcal{G}$ such that $g(x_1) = x_2$, then we say x_1 and x_2 are equivalent under $\mathcal{G}$, and we write

$$x_1 \equiv x_2 \bmod G. \tag{8.5}$$

Sets of equivalent points are called *orbits* of $\mathcal{G}$. (In other contexts such sets are called "residue classes".) It is clear that a function that is *invariant* under the transformation group $\mathcal{G}$ must be constant over the orbits of G. A transformation group $\mathcal{G}$ is said to be *transitive* over the set $\mathcal{X}$ if for any $x_1, x_2 \in \mathcal{X}$, there exists a g in $\mathcal{G}$ such that $g(x_1) = x_2$. (This terminology is not standard.) In this case the whole domain is an orbit.

An invariant function M over $\mathcal{G}$ is called *maximal invariant* over $\mathcal{G}$ if

$$M(x_1) = M(x_2) \quad \Rightarrow \quad \exists\, g \in \mathcal{G} \ni g(x_1) = x_2. \tag{8.6}$$

Maximal invariance can be used to characterize invariance. If M is maximal invariant under $\mathcal{G}$, then the function f is invariant under $\mathcal{G}$ if and only if it depends on x only through M; that is, if and only if there exists a function h such that for all x, $f(x) = h(M(x))$.

Any invariant function with respect to a transitive group is maximal invariant.

The underlying concept of maximal invariance is similar to the concept of sufficiency. A sufficient statistic may reduce the sample space; a maximal invariant statistic may reduce the parameter space. (Maximal invariant statistics have some technical issues regarding measurability, however; X being measurable does not guarantee $M(X)$ is measurable under the same measure.)

Equivariant Functions

A function f is said to be *equivariant* under the transformation group $\mathcal{G}$ with domain $\mathcal{X}$ if for all $x \in \mathcal{X}$ and $g \in \mathcal{G}$,

$$f(g(x)) = g(f(x)). \tag{8.7}$$

We also use the phrases "equivariant over ..." and "equivariant with respect to ..." to denote this kind of equivariance.

8.1.2 Invariant and Equivariant Statistical Procedures

We will denote a general statistical decision function by T. This is a function from the sample space (actually from the range $\mathcal{X}$ of the random variable X) to the decision space $\mathcal{A}$ (which can be considered a subset of the reals); that is, $T : \mathcal{X} \mapsto \mathcal{A} \subseteq \mathbb{R}$. We write

$$T(X) = a. \tag{8.8}$$

We are interested in the invariance or equivariance of $T(x)$ in the context of certain transformations. The context in which this has meaning is somewhat limited. It has meaning in group families when the loss function is of an appropriate type.

An estimator that changes appropriately (in ways that we will specify more precisely below) so that the risk is invariant under changes in the random variable is said to be equivariant. In testing statistical hypotheses, we often denote the statistical decision function by ϕ, and define the decision space as $[0, 1]$. We interpret $\phi(x)$ as the probability of rejecting the hypothesis for a given $x \in \mathcal{X}$. A test that does not change under changes in the random variable is said to be invariant. We emphasize that invariance or equivariance has meaning only in special contexts; both the family of distributions and the form of the loss function must have properties that are similar in certain ways.

The invariance or equivariance of interest is with respect to a given class of transformations. The most common class of interest is the group of linear transformations of the form $\tilde{x} = Ax + c$. A family of distributions whose probability measures accommodate a group of transformations in a natural way is called a *group family*. The group families of interest have a certain invariance with respect to a group of linear transformations on the random variable. We call such a group family a *location-scale family*. Formally, let P be a a probability measure on $(\mathbb{R}^k, \mathcal{B}^k)$, let $\mathcal{V} \subset \mathbb{R}^k$, and let $\mathcal{M}_k$ be a collection of $k \times k$ symmetric positive definite matrices. The family

$$\{P_{(\mu,\Sigma)} \ : \ P_{(\mu,\Sigma)}(B) = P(\Sigma^{1/2}(B - \mu)), \text{ for } \mu \in \mathcal{V}, \Sigma \in \mathcal{M}_k, B \in \mathcal{B}^k\} \tag{8.9}$$

is called a *location-scale* family.

Some standard parametric families that are group families: normal, double exponential, exponential and uniform (even with parametric ranges), and Cauchy.

A location-scale family of distributions can be defined in terms of a given distribution on $(\mathbb{R}^k, \mathcal{B}^k)$ as all distributions for which the probability measure is invariant under linear transformations.

Whatever parameter θ may characterize the distribution, we often focus on just μ and Σ, as above, or in the univariate case, μ and σ. (In most other cases our object of interest has been a transformation on the parameter space, $g(\theta)$. In the following, we will often denote a basic transformation of the *probability space* as $g(\cdot)$, and we may denote a corresponding transformation on the parameter space, as $\bar{g}(\cdot)$.)

Transformations on the Sample Space, the Parameter Space, and the Decision Space

To study invariance of statistical procedures we will now identify three groups of transformations $\mathcal{G}$, $\overline{\mathcal{G}}$, and $\mathcal{G}^*$, and the relationships among the groups. This notation is widely used in mathematical statistics, maybe with some slight modifications.

- Let $\mathcal{G}$ be a group of transformations that map the probability space onto itself. We write
$$g(X) = \tilde{X}. \tag{8.10}$$
Note that X and $\tilde{X}$ are random variables, so the domain and the range of the mapping are subsets of *probability spaces*; the random variables are based on the same underlying measure, so the probability spaces are the same; the transformation is a member of a transformation group, so the domain and the range are equal and the transformations are one-to-one.
$$g : \mathcal{X} \mapsto \mathcal{X}, \quad 1:1 \text{ and onto}$$

- For given $g \in \mathcal{G}$ above, let $\bar{g}$ be a 1:1 function that maps the parameter space onto itself, $\bar{g} : \Theta \mapsto \Theta$, in such a way that for any set A,
$$\Pr_\theta(g(X) \in A) = \Pr_{\bar{g}(\theta)}(X \in A).$$
If this is the case we say g *preserves* Θ. Any two functions that preserve the parameter space form a group of functions that preserve the parameter space. The set of all such $\bar{g}$ together with the induced structure is a group, $\bar{\mathcal{G}}$. We write
$$\bar{g}(\theta) = \tilde{\theta}. \tag{8.11}$$
$$\bar{g} : \Theta \mapsto \Theta, \quad 1:1 \text{ and onto}$$
We may refer to $\bar{\mathcal{G}}$ as the *induced* group under $\mathcal{G}$.
- For each $g \in \mathcal{G}$ above, there is a 1:1 function g^* that maps the decision space onto itself, $g^* : \mathcal{A} \mapsto \mathcal{A}$. The set of all such g^* together with the induced structure is a group, $\mathcal{G}^*$.
We write
$$g^*(a) = \tilde{a}. \tag{8.12}$$
$$g^* : \mathcal{A} \mapsto \mathcal{A}, \quad 1:1 \text{ and onto.}$$
The relationship between $\mathcal{G}$ and $\mathcal{G}^*$ is a homomorphism; that is, for $g \in \mathcal{G}$ and $g^* \in \mathcal{G}^*$, if $g^* = k(g)$, then $k(g_1 \circ g_2) = k(g_1) \circ k(g_2)$.

We are interested in a probability space, $(\Omega, \mathcal{F}, \mathcal{P}_\Theta)$, that is invariant to a class of transformations $\mathcal{G}$; that is, one in which $\mathcal{P}_\Theta$ is a group family with respect to $\mathcal{G}$. The induced groups $\bar{\mathcal{G}}$ and $\mathcal{G}^*$ determine the transformations to be applied to the parameter space and the action space.

Invariance of the Loss Function

In most statistical decision problems, we assume a *symmetry* or *invariance* or *equivariance* of the problem before application of any of these transformations, and the problem that results from applying all of the transformations. For given classes of transformations, we consider loss functions that are invariant to those transformations; that is, we require that the loss function have the property

$$\begin{aligned} L(\tilde{\theta}, \tilde{a}) &= L(\bar{g}(\theta), g^*(a)) \\ &= L(\theta, a). \end{aligned} \tag{8.13}$$

This means that a good statistical procedure, T, for the original problem is good for the transformed problem. Note that this is an *assumption* about the class of meaningful loss functions for this kind of statistical problem.

From this assumption about the loss function, we have the risk property

$$\mathrm{E}_\theta(g(X)) = \mathrm{E}_{\bar{g}(\theta)}(X). \tag{8.14}$$

We have seen cases in which, for a univariate function of the parameter, the loss function is a function only of $a - g(\theta)$ or of $a/g(\theta)$; that is, we may have $L(\theta, a) = L_{\mathrm{l}}(a - g(\theta))$, or $L(\theta, a) = L_{\mathrm{s}}(a/g(\theta))$. In order to develop equivariant procedures for a general location-scale family $P_{(\mu,\Sigma)}$ we need a loss function of the form

$$L((\mu, \Sigma), a) = L_{\mathrm{ls}}(\Sigma^{1/2}(a - \mu)). \tag{8.15}$$

Invariance of Statistical Procedures

The basic idea underlying invariance of statistical procedures naturally is invariance of the risk under the given transformations.

We seek a statistical procedure $T(x)$ that is an invariant function under the transformations. Because if there is a maximal invariant function M all invariant functions are dependent on M, our search for optimal invariant procedures can use M.

A probability model may be defined in different ways. There may be an equivalence between two different models that is essentially a result of a reparametrization: $\tilde{\theta} = \bar{g}(\theta)$. A random variable in the one model may be a function of the random variable in the other model: $\widetilde{X} = g(X)$. There are two ways of thinking of estimation under a reparametrization, both in the context of an estimator $T(X)$ of $h(\theta)$, and with the transformations defined above:

- functional, $g^*(T(X))$ estimates $g^*(h(\theta))$;
- formal, $T(g(X))$ estimates $g^*(h(\theta))$.

Functional equivariance is trivial. This is the equivariance we expect under a simple change of units, for example. If X is a random variable that models physical temperatures in some application, it should not make any real difference whether the temperatures are always measured in degrees Celsius or degrees Faharenheit. The random variable itself does not include units, of course (it is a real number). If the meausrements are made in degrees Celsius at a time when X is the random variable used to model the distribution of the data and the estimator $T(X)$ and the estimand $h(\theta)$ relates to X in a linear fashion (if $h(\theta)$ is the mean of X, for example), and later in a similar application the meausrements are made in degrees Faharenheit, applying $g^*(t) = 9t/5 + 32$ to both $T(X)$ and $h(\theta)$ preserves the interpretation of the model.

Formal equivariance, however, is not meaningful unless the problem itself has fundamentally symmetric properties; the family of probability distributions is closed under some group of transformations on the sample space one on the parameter space. In this case, we need a corresponding transformation on the decision space. The statistical procedure is equivariant it the functional eequivariance is the same as the formal equivariance; that is,

$$T(g(X)) = g^*(T(X)). \tag{8.16}$$

Optimality

Equivariance can be combined with other properties such as minimum risk or most powerfulness. As we have seen, there are situations where we cannot obtain these properties uniformly. By restricting attention to procedures with properties such as equivariance or unbiasedness, we may be able to achieve uniformly best procedures. With unbiasedness, we seek UMVU estimators and UMPU tests. Within a collection of equivariant estimators, we would choose the one with some optimality property such as minimum risk.

The simplest and most interesting transformations are translations and scalings, and the combinations of these two, that is linear transformations. Consequently, the two most common types of invariant inference problems are those that are location invariant (or equivariant) and those that are scale invariant (or equivariant). A location invariant procedure is not invariant to scale transformations, but a scale invariant procedure is invariant to location transformations.

8.2 Equivariant Point Estimation

If the estimand under the untransformed problem is θ, the estimand after the transformations is $\bar{g}(\theta)$. If $T(X)$ is an estimator of θ, equivariance of the estimator requires that $g^*(T(X)) = T(g(X))$ be an estimator of $\bar{g}(\theta)$ with the same risk.

The properties of the estimator in the untransformed problem are preserved under the transformations. An estimator that is equivariant except possibly on a set of zero probability is said to be *almost equivariant*.

Within a collection of equivariant estimators, we would choose the one with minimum risk. This is MRE estimation, and the estimator is an MREE. (Some authors call it MRI and MRIE.)

By the definition of "equivariance" in this context, the MRE estimator is UMRE, so the concept of uniformity does not arise as a separate issue here.

An estimator that is equivariant except possibly on a set of zero probability is said to be almost equivariant.

To find an MREE, the methods are very similar for location equivariance and scale equivariance. One method in both cases is first to characterize all equivariant estimators in terms of a given one, and then identify the one that minimizes the risk.

Location Equivariant Estimation

In location equivariant estimation, the basic transformation is a translation on both the random variable and the location parameter: $\widetilde{X} = X + c$ and $\tilde{\mu} = \mu + c$. The estimand of interest is μ. A reasonable loss function must have the property (8.15), that is, $L(\mu+c, a+c) = L(\mu, a)$ for any c, μ and a; hence, $L(\mu, a)$ is a function only of $(a - \mu)$:

$$L(\mu, a) = \rho(a - \mu). \tag{8.17}$$

(To repeat the argument that led to equation (8.15) and to see it in this particular case, let $\mu = -c$, and so we have $L(0, a) = L(0, a - \mu)$, and this equality must continue to hold as μ and c move in tandem.)

The estimator must have the property (8.16), that is,

$$T(x + a) = T(x) + a. \tag{8.18}$$

If T_0 is a location equivariant estimator, then any location equivariant estimator must be of the form $T(x) = T_0(x) + u(x)$, for any function u such that u is invariant to translations: $u(x + a) = u(x)$. (Notice the difference in "invariant" and "equivariant".) In particular, if $n > 1$,

$$T(x) = T_0(x) + u(y), \tag{8.19}$$

where

$$y_i = x_i - x_n \quad \text{for } i = 1, \ldots, n - 1,$$

and if $n = 1$, any location equivariant estimator must be of the form

$$T(x) = T_0(x) + c, \tag{8.20}$$

where c is a constant. With this knowledge, we can seek an estimator with minimum risk.

If we have a location equivariant estimator T_0 with finite risk, we determine the MREE (if it exists) as

$$T_*(x) = T_0(x) - u_*(y), \tag{8.21}$$

where $u_*(y)$ minimizes the conditional risk at $c = 0$:

$$\mathrm{E}_0\Big(\rho(T_0(x) - u(y)) \,|\, y\Big). \tag{8.22}$$

Note that the loss function has a special form of equation (8.17). In particular, for squared-error loss, which is of this form, we have

$$u_*(y) = \mathrm{E}_0\big(T_0(x) \,|\, y\big), \tag{8.23}$$

and in this case, if a UMVUE exists and is equivariant, it is MRE.

An equivariant estimator under a squared-error loss is called a *Pitman estimator*. For a sample $X_1, \ldots, X_n$ from a location family with joint PDF p, this Pitman estimator (that is, $T_*(x)$ in equation (8.21), with $u_*(x)$ from equation (8.23)), can be written as

$$T_*(x) = \frac{\int t\, p(X_1 - t, \ldots, X_n - t)\mathrm{d}t}{\int p(X_1 - t, \ldots, X_n - t)\mathrm{d}t}. \tag{8.24}$$

(This is Theorem 4.6 in Shao.)

A location equivariant estimator is not invariant to scale transformations.

Scale Equivariant Estimation

In scale equivariant estimation, the basic transformation is a multiplication on both the random variable and the a power nonzero power of the scale parameter: $\widetilde{X} = rX$, for $r > 0$, and $\tilde{\sigma} = r^h\sigma^h$. This development parallels that for location equivariant estimation in the previous section.

The estimand of interest is σ. A reasonable loss function must have the property (8.15), $L(r\sigma, r^h a) = L(\sigma, a)$, hence,

$$L(\sigma, a) = \gamma(a/\sigma^h), \tag{8.25}$$

and the estimator must have the property

$$T(rx) = r^h T(x). \tag{8.26}$$

If T_0 is a scale equivariant estimator, then any scale equivariant estimator must be of the form

$$T(x) = \frac{T_0(x)}{u(z)}, \tag{8.27}$$

where

$$z_i = \frac{x_1}{x_n}, \text{ for } i = 1, \ldots, n-1, \quad \text{and } z_n = \frac{x_n}{|x_n|}.$$

If we have a scale equivariant estimator T_0 with finite risk, we determine the MREE (if it exists) as

$$T_*(x) = T_0(x)/u_*(z), \tag{8.28}$$

where $u_*(z)$ minimizes the conditional risk at $r = 1$:

$$\mathrm{E}_1\Big(\gamma(T_0(x)/u(z)) \,|\, z\Big). \tag{8.29}$$

Note that the loss function has a special form. In the scale equivariant estimation problem, there are a couple of special loss functions. One is a squared error of the form

$$\gamma(a/\sigma^h) = \frac{(a - \sigma^h)^2}{\sigma^{2h}}, \tag{8.30}$$

in which case

$$u_*(z) = \frac{\mathrm{E}_1\Big((T_0(x))^2 \,|\, y\Big)}{\mathrm{E}_1\Big(T_0(x) \,|\, y\Big)}, \tag{8.31}$$

and the estimator is a Pitman estimator.

Another special loss functions is of the form

$$\gamma(a/\sigma^h) = a/\sigma^h - \log(a/\sigma^h) - 1, \tag{8.32}$$

called "Stein's loss", in which case

$$u_*(z) = \mathrm{E}_1(T_0(x) \,|\, y). \tag{8.33}$$

Stein's loss has the interesting property that it is the only scale-invariant loss function for which the UMVUE is also the MREE (difficult proof).

A scale equivariant estimator is invariant to location transformations; that is, if T is scale invariant, then $T(x + a) = T(x)$.

Location-Scale Equivariant Estimation

Location-scale equivariance involves the combination of the two separate developments. The basic transformations are location and scale: $\widetilde{X} = bX + a$ and $\tilde{\theta} = b\theta + a$.

The estimator must have the property

$$T(bx + a) = b^r T(x) + a. \tag{8.34}$$

Analysis of these estimators does not involve anything fundamentally different from combinations of the ideas discussed separately for the location and scale cases.

Equivariant Estimation in a Normal Family

MRE estimation has particular relevance to the family of normal distributions, which is a location-scale group family. The standard estimators of the location and the scale are equivariant, and furthermore are independent of each other. An interesting fact is that in location families that have densities with respect to Lebesgue measure and with finite variance, the risk of a MRE location estimator with squared error loss is larger in the normal family than in any other such family.

8.3 Invariant Tests and Equivariant Confidence Regions

8.3.1 Invariant Tests

We generally want statistical procedures to be invariant to various transformations of the problem. For example, if the observables X are transformed in some way, it should be possible to transform a "good" test for a certain hypothesis in some obvious way so that the test remains "good" using the transformed data. (This of course means that the hypothesis is also transformed.)

To address this issue more precisely, we consider transformation groups $\mathcal{G}$, $\overline{\mathcal{G}}$, and $\mathcal{G}^*$, defined and discussed beginning on page 271.

We are often able to define optimal tests under the restriction of invariance.

A test δ is said to be invariant under G, whose domain is the sample space $\mathcal{X}$, if for all $x \in \mathcal{X}$ and $g \in G$,

$$\delta(g(x)) = \delta(x). \tag{8.35}$$

(This is just the definition of an invariant function, equation (8.4).)

We seek most powerful invariant tests. (They are invariant because the accept/reject decision does not change.) Because of the meaning of "invariance" in this context, the most powerful invariant test is uniformly most powerful (UMPI), just as we saw in the case of the equivariant minimum risk estimator. The procedure for finding UMPI (or just MPI) tests is similar to the procedure used in the estimation problem. For a given class of transformations, we first attempt to characterize the form of ϕ, and then to determine the most powerful test of that form. Because of the relationship of invariant functions to a maximal invariant function, we may base our procedure on a maximal invariant function.

As an example, consider the group G of translations, for $x = (x_1, \ldots, x_n)$:

$$g(x) = (x_1 + c, \ldots, x_n + c).$$

Just as before, we see that for $n > 1$, the set of differences

$$y_i = x_i - x_n \quad \text{for } i = 1, \ldots, n-1,$$

is invariant under G. This function is also maximal invariant. For x and $\tilde{x}$, let $y(x) = y(\tilde{x})$. So we have for $i = 1, \ldots, n-1$,

$$\begin{aligned} \tilde{x_i} - \tilde{x_n} &= x_i - x_n \\ &= (x_i + c) - (x_n + c) \\ &= g(x), \end{aligned}$$

and therefore the function is maximal invariant. Now, suppose we have the sample $X = (X_1, \ldots, X_n)$ and we wish to test the hypothesis that the density of X is $p_0(x_1 - \theta, \ldots, x_n - \theta)$ versus the alternative that it is $p_1(x_1 - \theta, \ldots, x_n - \theta)$. This testing problem is invariant under the group G of translations, with the induced group of transformations $\overline{G}$ of the parameter space (which are translations also). Notice that there is only one orbit of $\overline{G}$, the full parameter space. The most powerful invariant test will be based on $Y = (X_1 - X_n, \ldots, X_{n-1} - X_n)$. The density of Y under the null hypothesis is given by

$$\int p_0(y_1 + z, \ldots, y_{n-1} + z, z)\mathrm{d}z,$$

and the density of Y under the alternate hypothesis is similar. Because both densities are independent of θ, we have two simple hypotheses, and the Neyman-Pearson lemma gives us the UMP test among the class of invariant tests. The rejection criterion is

$$\frac{\int p_1(y_1 + u, \ldots, y_n + u)\mathrm{d}u}{\int p_0(y_1 + u, \ldots, y_n + u)\mathrm{d}u} > c,$$

for some c.

You should look over similar location and/or scale invariant tests for the hypotheses about the parameters of a normal distribution and location and/or scale invariant permutation tests using paired comparisons. The basic idea is the same.

As we might expect, there are cases in which invariant procedures do not exist. For $n = 1$ there are no invariant functions under G in the translation example above. In such situations, obviously, we cannot seek UMP invariant tests.

8.3.2 Equivariant Confidence Sets

The connection we have seen between a $1-\alpha$ confidence region $S(x)$, and the acceptance region of a α-level test, $A(\theta)$, that is

$$S(x) \ni \theta \quad \Leftrightarrow \quad x \in A(\theta),$$

can often be used to relate UMP invariant tests to best equivariant confidence sets.

Equivariance for confidence sets is defined similarly to equivariance in other settings.

Under the notation developed above, for the group of transformations G and the induced transformation groups G^* and $\overline{G}$, a confidence set $S(x)$ is *equivariant* if for all $x \in \mathcal{X}$ and $g \in G$,

$$g^*(S(x)) = S(g(x)).$$

The uniformly most powerful property of the test corresponds to uniformly minimizing the probability that the confidence set contains incorrect values, and the invariance corresponds to equivariance.

An equivariant set that is $\widetilde{\Theta}$-uniformly more accurate ("more" is defined similarly to "most") than any other equivariant set is said to be a *uniformly most accurate equivariant* (UMAE) set.

There are situations in which there do not exist confidence sets that have uniformly minimum probability of including incorrect values. In such cases, we may retain the requirement for equivariance, but impose some other criterion, such as expected smallest size (w.r.t. Lebesgue measure) of the confidence interval.

8.3.3 Invariance/Equivariance and Unbiasedness and Admissibility

In some problems, the principles of invariance and unbiasedness are completely different; and in some cases, one may be relevant and the other totally irrelevant. In other cases there is a close connection between the two.

For the testing problem, the most interesting relationship between invariance and unbiasedness is that if a unique up to sets of measure zero UMPU test exists, and a UMPI test up to sets of measure zero exists, then the two tests are the same up to sets of measure zero. (To be proven later.)

Admissibility of a statistical procedure means that there is no procedure that is at least as "good" as the given procedure everywhere, and better than the given procedure where. In the case of testing "good" means "powerful", and, of course, everything depends on the level of the test.

A UMPU test is admissible, but a UMPI test is not necessarily admissible.

Notes

Exercises in Shao

- For practice and discussion
  4.47, 4.52 (Solutions in Shao, 2005)
- To turn in
  4.57(a), 6.63, 6.69(a), 6.72, 6.74, 7.58, 7.59

9

Robust Inference (Shao Sec 5.1, Sec 5.2, Sec 5.3; Staudte-Sheather)

A major concern is how well the statistical model corresponds to the data-generating process. Analyses based on an inappropriate model are likely to yield misleading conclusions. An important concern in the field of *robust statistics* is the consequence of the differences in the model and the data-generating process. A major objective of the field of robust statistics is to identify or develop procedures that yield useful conclusions even when the data-generating process differs in certain ways from the statistical model. Such procedures are *robust* to departures within a certain class from the assumed model.

Our study of robust procedures begins with development of measures of differences in probability distributions.

9.1 Statistical Functions

Functionals are functions whose arguments are functions. The value of a functional may be any kind of object, a real number or another function, for example. The domain of a functional is a set of functions. I will use notation of the following form: for the functional, a capital Greek or Latin letter, Υ, M, etc.; for the domain, a calligraphic Latin letter, $\mathcal{F}$, $\mathcal{P}$, etc.; for a function, an italic letter, f, F, g, etc.; and for the value, the usual notation for functions, $\Upsilon(F)$ where $F \in \mathcal{F}$, for example.

Functionals have important uses in statistics. Functionals of CDFs can be used as measures of the differences between two distributions. They can also be used to define distributional measures of interest, and to define estimators of those measures. The functionals used to measure differences between two distributions can then be used to evaluate the statistical properties of estimators that are defined in terms of functionals.

We often measure the difference in functions by a special kind of functional called a norm.

Distances between Probability Distributions

The difference in two probability distributions may be measured in terms of a distance between the cumulative distribution functions such as the Hellinger distance or the Kullback-Leibler measure as described on page 396, or it may be measured in terms of differences in probabilities or differences in expected values.

We are usually interested in using samples to make inferences about the distances between probability distributions. If we measure the distance between probability distributions in terms of a distance between the cumulative distribution functions, we may use the ECDFs from the samples. If the comparison is between a distribution of a sample and some family of distributions, we use the ECDF from the sample and a CDF from the family; If the comparison is between the distributions of two samples, we use the ECDFs from the samples.

It is important to note that even though the measure of the difference between two CDFs may be small, there may be very large differences in properties of the probability distributions. For example, consider the difference between the CDF of a standard Cauchy and a standard normal. The sup difference is about 0.1256. (It occurs near ± 1.85.) The sup dif between the ECDFs for samples of size 20 will often be between 0.2 and 0.3. (That is a significance level of 0.83 and 0.34 on a KS test.)

The Kolmogorov Distance; An L_∞ Metric

If g and f are CDFs, the L_∞ norm of their difference is called the *Kolmogorov distance* between the two distributions. We sometimes write the Kolmogorov distance between two CDFs P_1 and P_2, as $\rho_{\mathrm{K}}(P_1, P_2)$:

$$\rho_{\mathrm{K}}(P_1, P_2) = \sup |P_1 - P_2|. \tag{9.1}$$

If one or both of P_1 and P_2 are ECDFs we can compute the Kolmogorov distance fairly easily using the order statistics.

The Lévy Metric

Another measure of the distance between two CDFs is the *Lévy distance*, defined for the CDFs P_1 and P_2 as

$$\rho_{\mathrm{L}}(P_1, P_2) = \inf\{\epsilon, \text{ s.t. } \forall x, P_1(x - \epsilon) - \epsilon \leq P_2(x) \leq P_1(x + \epsilon) + \epsilon\}.$$

It can be shown that $\rho_{\mathrm{L}}(P_1, P_2)$ is a metric over the set of distribution functions. It can also be shown that for any CDFs P_1 and P_2,

$$\rho_{\mathrm{L}}(P_1, P_2) \leq \rho_{\mathrm{K}}(P_1, P_2) \leq 1.$$

The Mallows Metric, or the "Earth Movers' Distance"

Another useful measure of the distance between two CDFs is the *Mallows distance*. For the CDFs P_1 and P_2, with random variables Y_1 having CDF P_1 and Y_2 having CDF P_2, if $\mathrm{E}(\|Y_1\|^p)$ and $\mathrm{E}(\|Y_2\|^p)$ are finite, this distance is

$$\rho_{\mathrm{M}_p}(P_1, P_2) = \inf(\mathrm{E}(\|Y_1 - Y_2\|^p))^{1/p},$$

where the infimum is taken over all joint distributions with marginals P_1 and P_2.

For scalar-valued random variables, we can show that

$$\rho_{\mathrm{M}_p}(P_1, P_2) = (\mathrm{E}(\|P_1^{-1}(U) - P_2^{-1}(U)\|^p))^{1/p},$$

where U is a random variable with a $\mathrm{U}(0,1)$ distribution, and the "inverse CDF" is defined as

$$P_i^{-1}(t) = \inf_x\{x : P_i(x) \geq t\}.$$

The inverse CDF has many useful applications in statistics. The basic fact is that if X is an absolutely continuous random variable with CDF P, then $U = P(X)$ has a $\mathrm{U}(0,1)$ distribution. A discrete random variable has a similar property when we "spread out" the probability between the mass points. (One of the most common applications is in random number generation, because the basic pseudorandom variable that we can simulate has a $\mathrm{U}(0,1)$ distribution.)

The first question we might consider given this definition is whether the infimum exists, and then it is not clear whether this is indeed a metric. (The triangle inequality is the only hard question.) Bickel and Freedman (1981) answered both of these questions in the affirmative. The proof is rather complicated for vector-valued random variables; for scalar-valued random variables, there is a simpler proof in terms of the inverse CDF.

If P_1 and P_2 are univariate ECDFs based on the same number of observations, we have

$$\rho_{\mathrm{M}_p}(P_{n1}, P_{n2}) = \left(\frac{1}{n}\sum(|y_{1(i)} - y_{2(i)}|^p)\right)^{1/p}.$$

Functionals of the CDF; Distribution Measures

We have often defined classes of probability distributions indexed by a *parameter*. This is the basic approach in parametric inference. Often, however, we want to consider some general property or measure of a distribution without identifying that measure as an index, or characterization of the distribution.

While the cumulative distribution function is the most basic function for describing a probability distribution or a family of distributions, there are a number of other, simpler descriptors of probability distributions that are useful. Many of these are expressed as functionals of the CDF. For example,

the mean of a distribution, if it exists, may be written as the functional M of the CDF P:

$$M(P) = \int y \, \mathrm{d}P(y). \tag{9.2}$$

Estimators Based on Statistical Functions

A natural way of estimating a distributional measure that is defined in terms of a statistical function of the CDF is to use the same statistical function on the ECDF. This leads us to a plug-in estimator.

For example, the functional M defining the mean of a distribution in equation (9.2), if it exists, can be applied to the ECDF to yield the sample mean:

$$\begin{aligned} M(P_n) &= \int y \, \mathrm{d}P_n(y) \\ &= \sum y_i \frac{1}{n} \\ &= \bar{y}. \end{aligned}$$

As we know, this is a "good" estimator of the population mean. Likewise, corresponding to the central moments defined in equation (1.12), we have the sample central moments by applying M_r to the ECDF. Notice that this yields $(n-1)s^2/n$ as the second centralized sample moment.

Estimators based on statistical functions play major roles throughout nonparametric and semiparametric inference. They are also important in robust statistics. In robustness studies, we first consider the sensitivity of the statistical function to perturbations in distribution functions. Statistical functions that are relatively insensitive to perturbations in distribution functions when applied to a ECDF should yield robust estimators.

These kinds of plug-in estimators should generally have good asymptotic properties *relative to the corresponding population measures* because of the global asymptotic properties of the ECDF.

9.2 Robust Inference

Although the statistical functions we have considered have intuitive interpretations, the question remains as to what are the most useful distributional measures by which to describe a given distribution. In a simple case such as a normal distribution, the choices are obvious. For skewed distributions, or distributions that arise from mixtures of simpler distributions, the choices of useful distributional measures are not so obvious. A central concern in robust statistics is how a functional of a CDF behaves as the distribution is perturbed. If a functional is rather sensitive to small changes in the distribution, then one has more to worry about if the observations from the process of interest are contaminated with observations from some other process.

9.2.1 Sensitivity of Statistical Functions to Perturbations in the Distribution

One of the most interesting things about a function (or a functional) is how its value varies as the argument is perturbed. Two key properties are *continuity* and *differentiability*.

For the case in which the arguments are functions, the cardinality of the possible perturbations is greater than that of the continuum. We can be precise in discussions of continuity and differentiability of a functional Υ at a point (function) F in a domain $\mathcal{F}$ by defining another set $\mathcal{D}$ consisting of difference functions over $\mathcal{F}$; that is the set the functions $D = F_1 - F_2$ for $F_1, F_2 \in \mathcal{F}$.

Three kinds of functional differentials are defined on page 404.

Perturbations

In statistical applications using functionals defined on the CDF, we are interested in how the functional varies for "nearby" CDFs in the distribution function space.

A simple kind of perturbation of a given distribution is to form a mixture distribution with the given distribution as one of the components of the mixture.

We often consider a simple type of function in the neighborhood of the CDF. These are CDFs formed by adding a single mass point to the given distribution.

For a given CDF $P(y)$, we can define a simple perturbation as

$$P_{x,\epsilon}(y) = (1-\epsilon)P(y) + \epsilon \mathrm{I}_{[x,\infty)}(y), \tag{9.3}$$

where $0 \leq \epsilon \leq 1$. We will refer to this distribution as an ϵ-mixture distribution, and to the distribution with CDF P as the reference distribution. (This, of course, is the distribution of interest, so I often refer to it without any qualification.)

A simple interpretation of the perturbation in equation (9.3) is that it is the CDF of a mixture of a distribution with CDF P and a degenerate distribution with a single mass point at x, which may or may not be in the support of the distribution. The extent of the perturbation depends on ϵ; if $\epsilon = 0$, the distribution is the reference distribution.

If the distribution with CDF P is continuous with PDF p, the PDF of the mixture is

$$\mathrm{d}P_{x,\epsilon}(y)/\mathrm{d}y = (1-\epsilon)p(y) + \epsilon\delta(x-y),$$

where $\delta(\cdot)$ is the Dirac delta function. If the distribution is discrete, the probability mass function has nonzero probabilities (scaled by $(1-\epsilon)$) at each of the mass points associated with P together with a mass point at x with probability ϵ.

Figure 9.1 shows in the left-hand graph the PDF of a continuous reference distribution (solid line) and the PDF of the ϵ-mixture distribution (dotted line together with the mass point at x). (In Figure 9.1, although the specifics are not important, the reference distribution is a standard normal, $x = 1$, and $\epsilon = 0.1$.) A statistical function evaluated at $P_{x,\epsilon}$ compared to the function

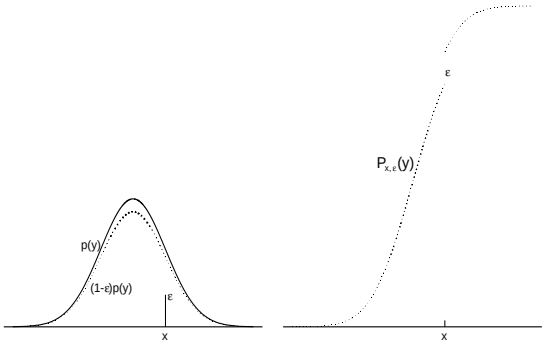


Fig. 9.1. PDFs and the CDF of the ϵ-Mixture Distribution

evaluated at P allows us to determine the effect of the perturbation on the statistical function. For example, we can determine the mean of the distribution with CDF $P_{x,\epsilon}$ in terms of the mean μ of the reference distribution to be $(1-\epsilon)\mu + \epsilon x$. This is easily seen by thinking of the distribution as a mixture. Formally, using the M in equation (9.2), we can write

$$\begin{aligned} M(P_{x,\epsilon}) &= \int y \, \mathrm{d}((1-\epsilon)P(y) + \epsilon \mathrm{I}_{[x,\infty)}(y)) \\ &= (1-\epsilon)\int y \, \mathrm{d}P(y) + \epsilon \int y\delta(y-x)\, \mathrm{d}y \\ &= (1-\epsilon)\mu + \epsilon x. \end{aligned} \tag{9.4}$$

For a discrete distribution we would follow the same steps using summations (instead of an integral of y times a Dirac delta function, we just have a point mass of 1 at x), and would get the same result.

The π quantile of the mixture distribution, $\Xi_\pi(P_{x,\epsilon}) = P^{-1}_{x,\epsilon}(\pi)$, is somewhat more difficult to work out. This quantile, which we will call q, is shown relative to the π quantile of the continuous reference distribution, y_π, for two cases in Figure 9.2. (In Figure 9.2, although the specifics are not important, the reference distribution is a standard normal, $\pi = 0.7$, so $y_\pi = 0.52$, and

$\epsilon = 0.1$. In the left-hand graph, $x_1 = -1.25$, and in the right-hand graph, $x_2 = 1.25$.)

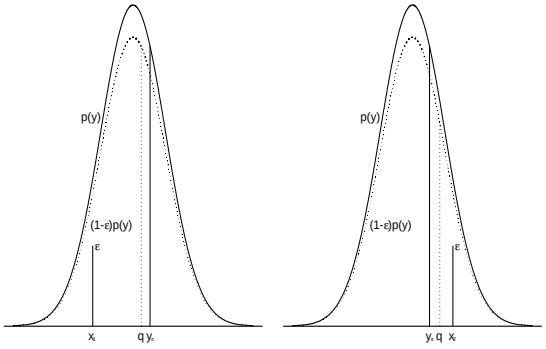


Fig. 9.2. Quantile of the ϵ-Mixture Distribution

We see that in the case of a continuous reference distribution (implying P is strictly increasing, as in the right-hand graph in Figure 9.1),

$$P^{-1}_{x,\epsilon}(\pi) = \begin{cases} P^{-1}\left(\frac{\pi-\epsilon}{1-\epsilon}\right), & \text{for } (1-\epsilon)P(x) + \epsilon < \pi, \\ x, & \text{for } (1-\epsilon)P(x) \leq \pi \leq (1-\epsilon)P(x) + \epsilon, \\ P^{-1}\left(\frac{\pi}{1-\epsilon}\right), & \text{for } \pi < (1-\epsilon)P(x). \end{cases} \tag{9.5}$$

The conditions in equation (9.5) can also be expressed in terms of x and quantiles of the reference distribution. For example, the first condition is equivalent to $x < \frac{y_{\pi-\epsilon}}{1-\epsilon}$.

The Influence Function

The extent of the perturbation depends on ϵ, and so we are interested in the relative effect; in particular, the relative effect as ϵ approaches zero. Davies and Gather (2004) discuss and give several examples of this kind of perturbation to study the sensitivity of a functional to perturbations of the CDF at a given point x.

The *influence function* for the functional Υ and the CDF P, defined at x as

$$\phi_{\Upsilon,P}(x) = \lim_{\epsilon \downarrow 0} \frac{\Upsilon(P_{x,\epsilon}) - \Upsilon(P)}{\epsilon} \tag{9.6}$$

if the limit exists, is a measure of the sensitivity of the distributional measure defined by Υ to a perturbation of the distribution at the point x. The influence function is also called the influence curve, and denoted by IC. The limit in equation (9.6) is the right-hand Gâteaux derivative of the functional Υ at P and x.

Staudte and Sheather (1990) point out that the influence function can also be expressed as the limit of the derivative of $\Upsilon(P_{x,\epsilon})$ with respect to ϵ:

$$\phi_{\Upsilon,P}(x) = \lim_{\epsilon \downarrow 0} \frac{\partial}{\partial \epsilon} \Upsilon(P_{x,\epsilon}). \tag{9.7}$$

This form is often more convenient for evaluating the influence function.

Some influence functions are easy to work out, for example, the influence function for the functional M in equation (9.2) that defines the mean of a distribution, which we denote by μ. The influence function for this functional operating on the CDF P at x is

$$\begin{aligned} \phi_{\mu,P}(x) &= \lim_{\epsilon \downarrow 0} \frac{M(P_{x,\epsilon}) - M(P)}{\epsilon} \\ &= \lim_{\epsilon \downarrow 0} \frac{(1-\epsilon)\mu + \epsilon x - \mu}{\epsilon} \\ &= x - \mu. \end{aligned} \tag{9.8}$$

We note that the influence function of a functional is a type of derivative of the functional, $\partial M(P_{x,\epsilon})/\partial \epsilon$. The influence function for other moments can be computed in the same way as the steps in equation (9.8).

Note that the influence function for the mean is unbounded in x; that is, it increases or decreases without bound as x increases or decreases without bound. Note also that this result is the same for multivariate or univariate distributions.

The influence function for a quantile is more difficult to work out. The problem arises from the difficulty in evaluating the quantile. As I informally described the distribution with CDF $P_{x,\epsilon}$, it is a mixture of some given distribution and a degenerate discrete distribution. Even if the reference distribution is continuous, the CDF of the mixture, $P_{x,\epsilon}$, does not have an inverse over the full support (although for quantiles we will write $P^{-1}_{x,\epsilon}$).

Let us consider a simple instance: a univariate continuous reference distribution, and assume $p(y_\pi) > 0$. We approach the problem by considering the PDF, or the probability mass function.

In the left-hand graph of Figure 9.2, the total probability mass up to the point y_π is $(1-\epsilon)$ times the area under the curve, that is, $(1-\epsilon)\pi$, plus the mass at x_1, that is, ϵ. Assuming ϵ is small enough, the π quantile of the ϵ-mixture distribution is the $\pi - \epsilon$ quantile of the reference distribution, or $P^{-1}(\pi - \epsilon)$.

It is also the π quantile of the scaled reference distribution; that is, it is the value of the function $(1-\epsilon)p(x)$ that corresponds to the proportion π of the total probability $(1-\epsilon)$ of that component. Use of equation (9.5) directly in equation (9.6) is somewhat messy. It is more straightforward to differentiate $P^{-1}_{x_1,\epsilon}$ and take the limit as in equation (9.7). For fixed $x < y_\pi$, we have

$$\frac{\partial}{\partial\epsilon}P^{-1}\left(\frac{\pi-\epsilon}{1-\epsilon}\right) = \frac{1}{p\left(P^{-1}\left(\frac{\pi-\epsilon}{1-\epsilon}\right)\right)}\frac{(\pi-1)(1-\epsilon)}{(1-\epsilon)^2}.$$

Likewise, we take the derivatives for the other cases in equation (9.5), and then take limits. We get

$$\phi_{\Xi_\pi,P}(x) = \begin{cases} \dfrac{\pi-1}{p(y_\pi)}, \text{ for } x < y_\pi, \\ 0, \qquad \text{for } x = y_\pi, \\ \dfrac{\pi}{p(y_\pi)}, \text{ for } x > y_\pi. \end{cases} \tag{9.9}$$

Notice that the actual value of x is not in the influence function; only whether x is less than, equal to, or greater than the quantile. Notice also that, unlike influence function for the mean, the influence function for a quantile is bounded; hence, a quantile is less sensitive than the mean to perturbations of the distribution. Likewise, quantile-based measures of scale and skewness, as in equations (1.17) and (1.18), are less sensitive than the moment-based measures to perturbations of the distribution.

The functionals L_J and M_ρ defined in equations (1.20) and (1.21), depending on J or ρ, can also be very insensitive to perturbations of the distribution.

The mean and variance of the influence function at a random point are of interest; in particular, we may wish to restrict the functional so that

$$\mathrm{E}(\phi_{\Upsilon,P}(X)) = 0$$

and

$$\mathrm{E}\left((\phi_{\Upsilon,P}(X))^2\right) < \infty.$$

9.2.2 Sensitivity of Estimators Based on Statistical Functions

If a distributional measure of interest is defined on the CDF as $\Upsilon(P)$, we are interested in the performance of the plug-in estimator $\Upsilon(P_n)$; specifically, we are interested in $\Upsilon(P_n) - \Upsilon(P)$. This turns out to depend crucially on the differentiability of Υ. If we assume Gâteaux differentiability, from equation (D.76), we can write

$$\begin{aligned}\sqrt{n}\,(\Upsilon(P_n) - \Upsilon(P)) &= \Lambda_P(\sqrt{n}(P_n - P)) + R_n \\ &= \frac{1}{\sqrt{n}} \sum_i \phi_{\Upsilon,P}(Y_i) + R_n\end{aligned}$$

where the remainder $R_n \to 0$.

We are interested in the stochastic convergence. First, we assume $\mathrm{E}(\phi_{\Upsilon,P}(X)) = 0$ and $\mathrm{E}\left((\phi_{\Upsilon,P}(X))^2\right) < \infty$. Then the question is the stochastic convergence of R_n. Gâteaux differentiability does not guarantee that R_n converges fast enough. However, ρ-Hadamard differentiability, does imply that that R_n is $\mathrm{o}_P(1)$, because it implies that norms of functionals (with or without random arguments) go to 0. We can also get that R_n is $\mathrm{o}_P(1)$ by assuming Υ is ρ-Fréchet differentiable and that $\sqrt{n}\rho(P_n, P)$ is $\mathrm{O}_P(1)$. In either case, that is, given the moment properties of $\phi_{\Upsilon,P}(X)$ and R_n is $\mathrm{o}_P(1)$, we have by Slutsky's theorem (Shao, page 60),

$$\sqrt{n}\,(\Upsilon(P_n) - \Upsilon(P)) \to_d \mathrm{N}(0, \sigma^2_{\Upsilon,P}),$$

where $\sigma^2_{\Upsilon,P} = \mathrm{E}\left((\phi_{\Upsilon,P}(X))^2\right)$.

For a given plug-in estimator based on the statistical function Υ, knowing $\mathrm{E}\left((\phi_{\Upsilon,P}(X))^2\right)$ (and assuming $\mathrm{E}(\phi_{\Upsilon,P}(X)) = 0$) provides us an estimator of the asymptotic variance of the estimator.

The influence function is also very important in leading us to estimators that are robust; that is, to estimators that are relatively insensitive to departures from the underlying assumptions about the distribution. As mentioned above, the functionals L_J and M_ρ, depending on J or ρ, can be very insensitive to perturbations of the distribution; therefore estimators based on them, called L-estimators and M-estimators, can be robust. A class of L-estimators that are particularly useful are linear combinations of the order statistics. Because of the sufficiency and completeness of the order statistics in many cases of interest, such estimators can be expected to exhibit good statistical properties.

Another class of estimators similar to the L-estimators are those based on ranks, which are simpler than order statistics. These are not sufficient – the data values have been converted to their ranks – nevertheless they preserve a lot of the information. The fact that they lose some information can actually work in their favor; they can be robust to extreme values of the data.

A functional to define even a simple linear combination of ranks is rather complicated. As with the L_J functional, we begin with a function J, which in this case we require to be strictly increasing, and also, in order to ensure uniqueness, we require that the CDF P be strictly increasing. The R_J functional is defined as the solution to the equation

$$\int J\left(\frac{P(y) + 1 - P(2R_J(P) - y)}{2}\right) \mathrm{d}P(y) = 0. \tag{9.10}$$

A functional defined as the solution to this optimization problem is called an R_J *functional*, and an estimator based on applying it to a ECDF is called an R_J *estimator* or just an R-estimator.

Notes

Adaptive Procedures

Exercises in Shao

- For practice and discussion
  5.5, 5.59, 5.61, 5.63, 5.74, 5.86, 5.111 (Solutions in Shao, 2005)
- To turn in
  5.3, 5.9, 5.24, 5.27, 5.39, 5.96

Additional References

Hogg, Robert V. (1974), Adaptive robust procedures: A partial review and some suggestions for future applications and theory (with discussion), *Journal of the American Statistical Association* **69**, 909–927.

Hogg, Robert V., and Russell V. Lenth (1984), A review of some adaptive statistical techniques, *Communications in Statistics — Theory and Methods* **13**, 1551–1579.

10

Nonparametric Estimation of Functions (Shao Sec 5.1; Scott)

10.1 Estimation of Functions

An interesting problem in statistics, and one that is generally difficult, is the estimation of a continuous function such as a probability density function. The statistical properties of an estimator of a function are more complicated than statistical properties of an estimator of a single parameter or even of a countable set of parameters. In this chapter we will discuss the properties of an estimator in the general case of a real scalar-valued function over real vector-valued arguments (that is, a mapping from $\mathbb{R}^d$ into $\mathbb{R}$). One of the most common situations in which these properties are relevant is in nonparametric probability density estimation.

First, we say a few words about notation. We may denote a function by a single letter, f, for example, or by the function notation, $f(\cdot)$ or $f(x)$. When $f(x)$ denotes a function, x is merely a placeholder. The notation $f(x)$, however, may also refer to the value of the function at the point x. The meaning is usually clear from the context.

Using the common "hat" notation for an estimator, we use $\widehat{f}$ or $\widehat{f}(x)$ to denote the estimator of f or of $f(x)$. Following the usual terminology, we use the term "estimator" to denote a random variable, and "estimate" to denote a realization of the random variable. The hat notation is also used to denote an estimate, so we must determine from the context whether $\widehat{f}$ or $\widehat{f}(x)$ denotes a random variable or a realization of a random variable. The estimate or the estimator of the value of the function at the point x may also be denoted by $\widehat{f(x)}$. Sometimes, to emphasize that we are estimating the ordinate of the function rather than evaluating an estimate of the function, we use the notation $\widehat{f(x)}$. In this case also, we often make no distinction in the notation between the realization (the estimate) and the random variable (the estimator). We must determine from the context whether $\widehat{f}(x)$ or $\widehat{f(x)}$ denotes a random variable or a realization of a random variable. In most of the following discussion, the hat notation denotes a random variable that

depends on the underlying random variable that yields the sample from which the estimator is computed.

The usual optimality properties that we use in developing a theory of estimation of a finite-dimensional parameter must be extended for estimation of a general function. As we will see, two of the usual desirable properties of point estimators, namely unbiasedness and maximum likelihood, cannot be attained in general by estimators of functions.

There are many similarities in *estimation* of functions and *approximation* of functions, but we must be aware of the fundamental differences in the two problems. Estimation of functions is similar to other estimation problems: we are given a sample of observations; we make certain assumptions about the probability distribution of the sample; and then we develop estimators. The estimators are random variables, and how useful they are depends on properties of their distribution, such as their expected values and their variances. Approximation of functions is an important aspect of numerical analysis. Functions are often approximated to interpolate functional values between directly computed or known values. Functions are also approximated as a prelude to quadrature. Methods for estimating functions often use methods for approximating functions.

10.1.1 General Methods for Estimating Functions

In the problem of function estimation, we may have observations on the function at specific points in the domain, or we may have indirect measurements of the function, such as observations that relate to a derivative or an integral of the function. In either case, the problem of function estimation has the competing goals of providing a good fit to the observed data and predicting values at other points. In many cases, a smooth estimate satisfies this latter objective. In other cases, however, the unknown function itself is not smooth. Functions with different forms may govern the phenomena in different regimes. This presents a very difficult problem in function estimation, and it is one that we will not consider in any detail here.

There are various approaches to estimating functions. Maximum likelihood has limited usefulness for estimating functions because in general the likelihood is unbounded. A practical approach is to assume that the function is of a particular form and estimate the parameters that characterize the form. For example, we may assume that the function is exponential, possibly because of physical properties such as exponential decay. We may then use various estimation criteria, such as least squares, to estimate the parameter. An extension of this approach is to assume that the function is a mixture of other functions. The mixture can be formed by different functions over different domains or by weighted averages of the functions over the whole domain. Estimation of the function of interest involves estimation of various parameters as well as the weights.

Another approach to function estimation is to represent the function of interest as a linear combination of basis functions, that is, to represent the function in a series expansion. The basis functions are generally chosen to be orthogonal over the domain of interest, and the observed data are used to estimate the coefficients in the series.

It is often more practical to estimate the function value at a given point. (Of course, if we can estimate the function at any given point, we can effectively have an estimate at all points.) One way of forming an estimate of a function at a given point is to take the average at that point of a filtering function that is evaluated in the vicinity of each data point. The filtering function is called a kernel, and the result of this approach is called a kernel estimator.

In the estimation of functions, we must be concerned about the properties of the estimators at specific points and also about properties over the full domain. Global properties over the full domain are often defined in terms of integrals or in terms of suprema or infima.

Kernel Methods

Another approach to function estimation and approximation is to use a *filter* or *kernel* function to provide local weighting of the observed data. This approach ensures that at a given point the observations close to that point influence the estimate at the point more strongly than more distant observations. A standard method in this approach is to convolve the observations with a unimodal function that decreases rapidly away from a central point. This function is the filter or the kernel. A kernel has two arguments representing the two points in the convolution, but we typically use a single argument that represents the distance between the two points.

Some examples of univariate kernel functions are shown below.

$$\begin{array}{lll} \text{uniform:} & K_{\mathrm{u}}(t) = 0.5, & \text{for } |t| \leq 1, \\ \text{quadratic:} & K_{\mathrm{q}}(t) = 0.75(1-t^2), & \text{for } |t| \leq 1, \\ \text{normal:} & K_{\mathrm{n}}(t) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-t^2/2}, & \text{for all } t. \end{array}$$

The kernels with finite support are defined to be 0 outside that range. Often, multivariate kernels are formed as products of these or other univariate kernels.

In kernel methods, the locality of influence is controlled by a *window* around the point of interest. The choice of the size of the window is the most important issue in the use of kernel methods. In practice, for a given choice of the size of the window, the argument of the kernel function is transformed to reflect the size. The transformation is accomplished using a positive definite matrix, V, whose determinant measures the volume (size) of the window.

To estimate the function f at the point x, we first decompose f to have a factor that is a probability density function, p,

$$f(x) = g(x)p(x).$$

For a given set of data, $x_1, \ldots, x_n$, and a given scaling transformation matrix V, the kernel estimator of the function at the point x is

$$\widehat{f(x)} = (n|V|)^{-1} \sum_{i=1}^{n} g(x) K\left(V^{-1}(x - x_i)\right). \tag{10.1}$$

In the univariate case, the size of the window is just the width h. The argument of the kernel is transformed to s/h, so the function that is convolved with the function of interest is $K(s/h)/h$. The univariate kernel estimator is

$$\widehat{f(x)} = \frac{1}{nh} \sum_{i=1}^{n} g(x) K\left(\frac{x - x_i}{h}\right).$$

10.1.2 Pointwise Properties of Function Estimators

The statistical properties of an estimator of a function at a given point are analogous to the usual statistical properties of an estimator of a scalar parameter. The statistical properties involve expectations or other properties of random variables. In the following, when we write an expectation, $\mathrm{E}(\cdot)$, or a variance, $\mathrm{V}(\cdot)$, the expectations are usually taken with respect to the (unknown) distribution of the underlying random variable. Occasionally, we may explicitly indicate the distribution by writing, for example, $\mathrm{E}_p(\cdot)$, where p is the density of the random variable with respect to which the expectation is taken.

Bias

The bias of the estimator of a function value at the point x is

$$\mathrm{E}\big(\widehat{f}(x)\big) - f(x).$$

If this bias is zero, we would say that the estimator is unbiased at the point x. If the estimator is unbiased at every point x in the domain of f, we say that the estimator is pointwise unbiased. Obviously, in order for $\widehat{f}(\cdot)$ to be pointwise unbiased, it must be defined over the full domain of f.

Variance

The variance of the estimator at the point x is

$$\mathrm{V}\big(\widehat{f}(x)\big) = \mathrm{E}\left(\left(\widehat{f}(x) - \mathrm{E}\big(\widehat{f}(x)\big)\right)^2\right).$$

Estimators with small variance are generally more desirable, and an optimal estimator is often taken as the one with smallest variance among a class of unbiased estimators.

Mean Squared Error

The mean squared error, MSE, at the point x is

$$\mathrm{MSE}\big(\widehat{f}(x)\big) = \mathrm{E}\Big(\big(\widehat{f}(x) - f(x)\big)^2\Big). \tag{10.2}$$

The mean squared error is the sum of the variance and the square of the bias:

$$\begin{aligned}\mathrm{MSE}\big(\widehat{f}(x)\big) &= \mathrm{E}\Big(\big(\widehat{f}(x)\big)^2 - 2\widehat{f}(x)f(x) + \big(f(x)\big)^2\Big) \\ &= \mathrm{V}\big(\widehat{f}(x)\big) + \Big(\mathrm{E}\big(\widehat{f}(x)\big) - f(x)\Big)^2. \end{aligned} \tag{10.3}$$

Sometimes, the variance of an unbiased estimator is much greater than that of an estimator that is only slightly biased, so it is often appropriate to compare the mean squared error of the two estimators. In some cases, as we will see, unbiased estimators do not exist, so rather than seek an unbiased estimator with a small variance, we seek an estimator with a small MSE.

Mean Absolute Error

The mean absolute error, MAE, at the point x is similar to the MSE:

$$\mathrm{MAE}\big(\widehat{f}(x)\big) = \mathrm{E}\Big(\big|\widehat{f}(x) - f(x)\big|\Big). \tag{10.4}$$

It is more difficult to do mathematical analysis of the MAE than it is for the MSE. Furthermore, the MAE does not have a simple decomposition into other meaningful quantities similar to the MSE.

Consistency

Consistency of an estimator refers to the convergence of the expected value of the estimator to what is being estimated as the sample size increases without bound. A point estimator T_n, based on a sample of size n, is consistent for θ if

$$\mathrm{E}(T_n) \to \theta \quad \text{as } n \to \infty.$$

The convergence is stochastic, of course, so there are various types of convergence that can be required for consistency. The most common kind of convergence considered is weak convergence, or convergence in probability.

In addition to the type of stochastic convergence, we may consider the convergence of various measures of the estimator. In general, if m is a function (usually a vector-valued function that is an elementwise norm), we may define consistency of an estimator T_n in terms of m if

$$\mathrm{E}(m(T_n - \theta)) \to 0. \tag{10.5}$$

For an estimator, we are often interested in *weak convergence in mean square* or *weak convergence in quadratic mean*, so the common definition of consistency of T_n is

$$\mathrm{E}\big((T_n-\theta)^{\mathrm{T}}(T_n-\theta)\big)\to 0,$$

where the type of convergence is convergence in probability. Consistency defined by convergence in mean square is also called L_2 consistency.

If convergence does occur, we are interested in the rate of convergence. We define rate of convergence in terms of a function of n, say $r(n)$, such that

$$\mathrm{E}(m(T_n-\theta))=\mathrm{O}(r(n)).$$

A common form of $r(n)$ is n^α, where $\alpha<0$. For example, in the simple case of a univariate population with a finite mean μ and finite second moment, use of the sample mean $\bar{x}$ as the estimator T_n, and use of $m(z)=z^2$, we have

$$\begin{aligned}\mathrm{E}(m(\bar{x}-\mu)) &= \mathrm{E}\big((\bar{x}-\mu)^2\big)\\ &= \mathrm{MSE}(\bar{x})\\ &= \mathrm{O}\big(n^{-1}\big).\end{aligned}$$

In the estimation of a function, we say that the estimator $\widehat{f}$ of the function f is *pointwise consistent* if

$$\mathrm{E}\big(\widehat{f}(x)\big)\to f(x) \tag{10.6}$$

for every x the domain of f. Just as in the estimation of a parameter, there are various kinds of pointwise consistency in the estimation of a function. If the convergence in expression (10.6) is in probability, for example, we say that the estimator is weakly pointwise consistent. We could also define other kinds of pointwise consistency in function estimation along the lines of other types of consistency.

10.1.3 Global Properties of Estimators of Functions

Often, we are interested in some measure of the statistical properties of an estimator of a function over the full domain of the function. The obvious way of defining statistical properties of an estimator of a function is to integrate the pointwise properties discussed in the previous section.

Statistical properties of a function, such as the bias of the function, are often defined in terms of a norm of the function.

For comparing $\widehat{f}(x)$ and $f(x)$, the L_p norm of the error is

$$\left(\int_D \big|\widehat{f}(x)-f(x)\big|^p\,\mathrm{d}x\right)^{1/p}, \tag{10.7}$$

where D is the domain of f. The integral may not exist, of course. Clearly, the estimator $\widehat{f}$ must also be defined over the same domain.

Three useful measures are the L_1 norm, also called the *integrated absolute error*, or IAE,

$$\mathrm{IAE}(\widehat{f}\,) = \int_D \left|\widehat{f}(x) - f(x)\right| \,\mathrm{d}x, \tag{10.8}$$

the square of the L_2 norm, also called the *integrated squared error*, or ISE,

$$\mathrm{ISE}(\widehat{f}\,) = \int_D \left(\widehat{f}(x) - f(x)\right)^2 \,\mathrm{d}x, \tag{10.9}$$

and the L_∞ norm, the *sup absolute error*, or SAE,

$$\mathrm{SAE}(\widehat{f}\,) = \sup\left|\widehat{f}(x) - f(x)\right|. \tag{10.10}$$

The L_1 measure is invariant under monotone transformations of the coordinate axes, but the measure based on the L_2 norm is not.

The L_∞ norm, or SAE, is the most often used measure in general function approximation. In statistical applications, this measure applied to two cumulative distribution functions is the *Kolmogorov distance*. The measure is not so useful in comparing densities and is not often used in density estimation.

Other measures of the difference in $\widehat{f}$ and f over the full range of x are the Kullback-Leibler measure,

$$\int_D \widehat{f}(x) \log\left(\frac{\widehat{f}(x)}{f(x)}\right) \,\mathrm{d}x,$$

and the Hellinger distance,

$$\left(\int_D \left(\widehat{f}^{\,1/p}(x) - f^{1/p}(x)\right)^p \,\mathrm{d}x\right)^{1/p}.$$

For $p = 2$, the Hellinger distance is also called the Matusita distance.

Integrated Bias and Variance

We now want to develop global concepts of bias and variance for estimators of functions. Bias and variance are statistical properties that involve expectations of random variables. The obvious global measures of bias and variance are just the pointwise measures integrated over the domain. In the case of the bias, of course, we must integrate the absolute value, otherwise points of negative bias could cancel out points of positive bias.

The estimator $\widehat{f}$ is pointwise unbiased if

$$\mathrm{E}\big(\widehat{f}(x)\big) = f(x) \quad \text{for all } x \in \mathbb{R}^d.$$

Because we are interested in the bias over the domain of the function, we define the *integrated absolute bias* as

$$\mathrm{IAB}(\widehat{f}) = \int_D \left| \mathrm{E}(\widehat{f}(x)) - f(x) \right| \mathrm{d}x \tag{10.11}$$

and the *integrated squared bias* as

$$\mathrm{ISB}(\widehat{f}) = \int_D \left(\mathrm{E}(\widehat{f}(x)) - f(x) \right)^2 \mathrm{d}x. \tag{10.12}$$

If the estimator is unbiased, both the integrated absolute bias and integrated squared bias are 0. This, of course, would mean that the estimator is pointwise unbiased almost everywhere. Although it is not uncommon to have unbiased estimators of scalar parameters or even of vector parameters with a countable number of elements, it is not likely that an estimator of a function could be unbiased at almost all points in a dense domain. ("Almost" means all except possibly a set with a probability measure of 0.)

The *integrated variance* is defined in a similar manner:

$$\begin{aligned} \mathrm{IV}(\widehat{f}) &= \int_D \mathrm{V}(\widehat{f}(x)) \, \mathrm{d}x \\ &= \int_D \mathrm{E}\Big(\big(\widehat{f}(x) - \mathrm{E}(\widehat{f}(x))\big)^2 \Big) \, \mathrm{d}x. \end{aligned} \tag{10.13}$$

Integrated Mean Squared Error and Mean Absolute Error

As we suggested above, global unbiasedness is generally not to be expected. An important measure for comparing estimators of funtions is, therefore, based on the mean squared error.

The *integrated mean squared error* is

$$\begin{aligned} \mathrm{IMSE}(\widehat{f}) &= \int_D \mathrm{E}\Big(\big(\widehat{f}(x) - f(x)\big)^2 \Big) \, \mathrm{d}x \\ &= \mathrm{IV}(\widehat{f}) + \mathrm{ISB}(\widehat{f}) \end{aligned} \tag{10.14}$$

(compare equations (10.2) and (10.3)).

If the expectation integration can be interchanged with the outer integration in the expression above, we have

$$\begin{aligned} \mathrm{IMSE}(\widehat{f}) &= \mathrm{E}\left(\int_D \left(\widehat{f}(x) - f(x) \right)^2 \mathrm{d}x \right) \\ &= \mathrm{MISE}(\widehat{f}), \end{aligned}$$

the *mean integrated squared error*. We will assume that this interchange leaves the integrals unchanged, so we will use MISE and IMSE interchangeably.

Similarly, for the *integrated mean absolute error*, we have

$$\begin{aligned}\mathrm{IMAE}(\widehat{f}) &= \int_D \mathrm{E}\Big(\big|\widehat{f}(x) - f(x)\big|\Big)\,\mathrm{d}x \\ &= \mathrm{E}\left(\int_D \Big|\widehat{f}(x) - f(x)\Big|\,\mathrm{d}x\right) \\ &= \mathrm{MIAE}(\widehat{f}),\end{aligned}$$

the *mean integrated absolute error*.

Mean SAE

The *mean sup absolute error*, or MSAE, is

$$\mathrm{MSAE}(\widehat{f}) = \int_D \mathrm{E}\Big(\sup\big|\widehat{f}(x) - f(x)\big|\Big)\,\mathrm{d}x. \tag{10.15}$$

This measure is not very useful unless the variation in the function f is relatively small. For example, if f is a density function, $\widehat{f}$ can be a "good" estimator, yet the MSAE may be quite large. On the other hand, if f is a cumulative distribution function (monotonically ranging from 0 to 1), the MSAE may be a good measure of how well the estimator performs. As mentioned earlier, the SAE is the *Kolmogorov distance*. The Kolmogorov distance (and, hence, the SAE and the MSAE) does poorly in measuring differences in the tails of the distribution.

Large-Sample Statistical Properties

The pointwise consistency properties are extended to the full function in the obvious way. In the notation of expression (10.5), consistency of the function estimator is defined in terms of

$$\int_D \mathrm{E}\Big(m\big(\widehat{f}(x) - f(x)\big)\Big)\,\mathrm{d}x \to 0.$$

The estimator of the function is said to be *mean square consistent* or L_2 *consistent* if the MISE converges to 0; that is,

$$\int_D \mathrm{E}\Big(\big(\widehat{f}(x) - f(x)\big)^2\Big)\,\mathrm{d}x \;\to\; 0.$$

If the convergence is weak, that is, if it is convergence in probability, we say that the function estimator is weakly consistent; if the convergence is strong, that is, if it is convergence almost surely or with probability 1, we say the function estimator is strongly consistent.

The estimator of the function is said to be L_1 *consistent* if the mean integrated absolute error (MIAE) converges to 0; that is,

$$\int_D \mathrm{E}\Big(\Big|\widehat{f}(x) - f(x)\Big|\Big)\,\mathrm{d}x \;\to\; 0.$$

As with the other kinds of consistency, the nature of the convergence in the definition may be expressed in the qualifiers "weak" or "strong".

As we have mentioned above, the integrated absolute error is invariant under monotone transformations of the coordinate axes, but the L_2 measures are not. As with most work in L_1, however, derivation of various properties of IAE or MIAE is more difficult than for analogous properties with respect to L_2 criteria.

If the MISE converges to 0, we are interested in the rate of convergence. To determine this, we seek an expression of MISE as a function of n. We do this by a Taylor series expansion.

In general, if $\widehat{\theta}$ is an estimator of θ, the Taylor series for $\mathrm{ISE}(\widehat{\theta})$, equation (10.9), about the true value is

$$\mathrm{ISE}\big(\widehat{\theta}\big) = \sum_{k=0}^{\infty} \frac{1}{k!} \Big(\widehat{\theta} - \theta\Big)^k \, \mathrm{ISE}^{k'}(\theta), \tag{10.16}$$

where $\mathrm{ISE}^{k'}(\theta)$ represents the k^{th} derivative of ISE evaluated at θ.

Taking the expectation in equation (10.16) yields the MISE. The limit of the MISE as $n \to \infty$ is the *asymptotic mean integrated squared error*, AMISE. One of the most important properties of an estimator is the order of the AMISE.

In the case of an unbiased estimator, the first two terms in the Taylor series expansion are zero, and the AMISE is

$$\mathrm{V}(\widehat{\theta})\, \mathrm{ISE}''(\theta)$$

to terms of second order.

Other Global Properties of Estimators of Functions

There are often other properties that we would like an estimator of a function to possess. We may want the estimator to weight given functions in some particular way. For example, if we know how the function to be estimated, f, weights a given function r, we may require that the estimate $\widehat{f}$ weight the function r in the same way; that is,

$$\int_D r(x)\widehat{f}(x)\mathrm{d}x = \int_D r(x)f(x)\mathrm{d}x.$$

We may want to restrict the minimum and maximum values of the estimator. For example, because many functions of interest are nonnegative, we may want to require that the estimator be nonnegative.

We may want to restrict the variation in the function. This can be thought of as the "roughness" of the function. A reasonable measure of the variation is

$$\int_D \left(f(x) - \int_D f(x)\mathrm{d}x \right)^2 \mathrm{d}x.$$

If the integral $\int_D f(x)\mathrm{d}x$ is constrained to be some constant (such as 1 in the case that $f(x)$ is a probability density), then the variation can be measured by the square of the L_2 norm,

$$\mathcal{S}(f) = \int_D \big(f(x)\big)^2 \mathrm{d}x. \tag{10.17}$$

We may want to restrict the derivatives of the estimator or the smoothness of the estimator. Another intuitive measure of the roughness of a twice-differentiable and integrable univariate function f is the integral of the square of the second derivative:

$$\mathcal{R}(f) = \int_D \big(f''(x)\big)^2 \mathrm{d}x. \tag{10.18}$$

Often, in function estimation, we may seek an estimator $\widehat{f}$ such that its roughness (by some definition) is small.

10.2 Nonparametric Estimation of CDFs and PDFs

10.2.1 Nonparametric Probability Density Estimation

Estimation of a probability density function is similar to the estimation of any function, and the properties of the function estimators that we have discussed are relevant for density function estimators. A density function $p(y)$ is characterized by two properties:

- it is nonnegative everywhere;
- it integrates to 1 (with the appropriate definition of "integrate").

In this chapter, we consider several nonparametric estimators of a density; that is, estimators of a general nonnegative function that integrates to 1 and for which we make no assumptions about a functional form other than, perhaps, smoothness.

It seems reasonable that we require the density estimate to have the characteristic properties of a density:

- $\widehat{p}(y) \geq 0$ for all y;
- $\int_{\mathbb{R}^d} \widehat{p}(y)\,\mathrm{d}y = 1$.

A probability density estimator that is nonnegative and integrates to 1 is called a *bona fide* estimator.

Rosenblatt has shown that no unbiased bona fide estimator can exist for all continuous p. Rather than requiring an unbiased estimator that cannot be a bona fide estimator, we generally seek a bona fide estimator with small mean

squared error or a sequence of bona fide estimators $\widehat{p}_n$ that are asymptotically unbiased; that is,

$$\mathrm{E}_p(\widehat{p}_n(y)) \to p(y) \quad \text{for all } y \in \mathbb{R}^d \text{ as } n \to \infty.$$

The Likelihood Function

Suppose that we have a random sample, $y_1, \ldots, y_n$, from a population with density p. Treating the density p as a variable, we write the likelihood functional as

$$L(p;\, y_1, \ldots, y_n) = \prod_{i=1}^{n} p(y_i).$$

The *maximum likelihood method* of estimation obviously cannot be used directly because this functional is unbounded in p. We may, however, seek an estimator that maximizes some modification of the likelihood. There are two reasonable ways to approach this problem. One is to restrict the domain of the optimization problem. This is called *restricted maximum likelihood*. The other is to *regularize* the estimator by adding a penalty term to the functional to be optimized. This is called *penalized maximum likelihood*.

We may seek to maximize the likelihood functional subject to the constraint that p be a bona fide density. If we put no further restrictions on the function p, however, infinite Dirac spikes at each observation give an unbounded likelihood, so a maximum likelihood estimator cannot exist, subject only to the restriction to the bona fide class. An additional restriction that p be Lebesgue-integrable over some domain D (that is, $p \in L^1(D)$) does not resolve the problem because we can construct sequences of finite spikes at each observation that grow without bound.

We therefore must restrict the class further. Consider a finite dimensional class, such as the class of step functions that are bona fide density estimators. We assume that the sizes of the regions over which the step function is constant are greater than 0.

For a step function with m regions having constant values, $c_1, \ldots, c_m$, the likelihood is

$$\begin{aligned} L(c_1, \ldots, c_m;\, y_1, \ldots, y_n) &= \prod_{i=1}^{n} p(y_i) \\ &= \prod_{k=1}^{m} c_k^{n_k}, \end{aligned} \tag{10.19}$$

where n_k is the number of data points in the k^{th} region. For the step function to be a bona fide estimator, all c_k must be nonnegative and finite. A maximum therefore exists in the class of step functions that are bona fide estimators.

If v_k is the measure of the volume of the k^{th} region (that is, v_k is the length of an interval in the univariate case, the area in the bivariate case, and so on), we have

$$\sum_{k=1}^{m} c_k v_k = 1.$$

We incorporate this constraint together with equation (10.19) to form the Lagrangian,

$$L(c_1, \ldots, c_m) + \lambda \left(1 - \sum_{k=1}^{m} c_k v_k\right).$$

Differentiating the Lagrangian function and setting the derivative to zero, we have at the maximum point $c_k = c_k^*$, for any λ,

$$\frac{\partial L}{\partial c_k} = \lambda v_k.$$

Using the derivative of L from equation (10.19), we get

$$n_k L = \lambda c_k^* v_k.$$

Summing both sides of this equation over k, we have

$$nL = \lambda,$$

and then substituting, we have

$$n_k L = nL c_k^* v_k.$$

Therefore, the maximum of the likelihood occurs at

$$c_k^* = \frac{n_k}{n v_k}.$$

The restricted maximum likelihood estimator is therefore

$$\begin{aligned} \widehat{p}(y) &= \frac{n_k}{n v_k}, \text{ for } y \in \text{region } k, \\ &= 0, \qquad \text{otherwise.} \end{aligned} \tag{10.20}$$

Instead of restricting the density estimate to step functions, we could consider other classes of functions, such as piecewise linear functions.

We may also seek other properties, such as smoothness, for the estimated density. One way of achieving other desirable properties for the estimator is to use a penalizing function to modify the function to be optimized. Instead of the likelihood function, we may use a penalized likelihood function of the form

$$L_{\mathrm{p}}(p; y_1, \ldots, y_n) = \prod_{i=1}^{n} p(y_i) \mathrm{e}^{-\mathcal{T}(p)},$$

where $\mathcal{T}(p)$ is a transform that measures some property that we would like to minimize. For example, to achieve smoothness, we may use the transform $\mathcal{R}(p)$ of equation (10.18) in the penalizing factor. To choose a function $\hat{p}$ to maximize $L_{\mathrm{p}}(p)$ we would have to use some finite series approximation to $\mathcal{T}(\hat{p})$.

For densities with special properties there may be likelihood approaches that take advantage of those properties.

10.2.2 Histogram Estimators

Let us assume finite support D, and construct a fixed partition of D into a grid of m nonoverlapping bins T_k. (We can arbitrarily assign bin boundaries to one or the other bin.) Let v_k be the volume of the k^{th} bin (in one dimension, v_k is a length and in this simple case is often denoted h_k; in two dimensions, v_k is an area, and so on). The number of such bins we choose, and consequently their volumes, depends on the sample size n, so we sometimes indicate that dependence in the notation: $v_{n,k}$. For the sample $y_1, \ldots, y_n$, the histogram estimator of the probability density function is defined as

$$\begin{aligned}\widehat{p}_H(y) &= \sum_{k=1}^{m} \frac{1}{v_k} \frac{\sum_{i=1}^{n} \mathrm{I}_{T_k}(y_i)}{n} \mathrm{I}_{T_k}(y), \quad \text{for } y \in D,\\ &= 0, \quad \text{otherwise.}\end{aligned}$$

The histogram is the restricted maximum likelihood estimator (10.20).

Letting n_k be the number of sample values falling into T_k,

$$n_k = \sum_{i=1}^{n} \mathrm{I}_{T_k}(y_i),$$

we have the simpler expression for the histogram over D,

$$\widehat{p}_H(y) = \sum_{k=1}^{m} \frac{n_k}{n v_k} \mathrm{I}_{T_k}(y). \tag{10.21}$$

As we have noted already, this is a bona fide estimator:

$$\widehat{p}_H(y) \geq 0$$

and

$$\begin{aligned}\int_{\mathbb{R}^d} \widehat{p}_H(y)\mathrm{d}y &= \sum_{k=1}^{m} \frac{n_k}{n v_k} v_k\\ &= 1.\end{aligned}$$

Although our discussion generally concerns observations on multivariate random variables, we should occasionally consider simple univariate observations. One reason why the univariate case is simpler is that the derivative is a

scalar function. Another reason why we use the univariate case as a model is because it is easier to visualize. The density of a univariate random variable is two-dimensional, and densities of other types of random variables are of higher dimension, so only in the univariate case can the density estimates be graphed directly.

In the univariate case, we assume that the support is the finite interval $[a, b]$. We partition $[a, b]$ into a grid of m nonoverlapping bins $T_k = [t_{n,k}, t_{n,k+1})$ where

$$a = t_{n,1} < t_{n,2} < \ldots < t_{n,m+1} = b.$$

The univariate histogram is

$$\widehat{p}_H(y) = \sum_{k=1}^{m} \frac{n_k}{n(t_{n,k+1} - t_{n,k})} \mathrm{I}_{T_k}(y). \tag{10.22}$$

If the bins are of equal width, say h (that is, $t_k = t_{k-1} + h$), the histogram is

$$\widehat{p}_H(y) = \frac{n_k}{nh}, \quad \text{for } y \in T_k.$$

This class of functions consists of polynomial splines of degree 0 with fixed knots, and the histogram is the maximum likelihood estimator over the class of step functions. Generalized versions of the histogram can be defined with respect to splines of higher degree. Splines with degree higher than 1 may yield negative estimators, but such histograms are also maximum likelihood estimators over those classes of functions.

The histogram as we have defined it is sometimes called a "density histogram", whereas a "frequency histogram" is not normalized by the n.

Some Properties of the Histogram Estimator

The histogram estimator, being a step function, is discontinuous at cell boundaries, and it is zero outside of a finite range. It is sensitive both to the bin size and to the choice of the origin.

An important advantage of the histogram estimator is its simplicity, both for computations and for analysis. In addition to its simplicity, as we have seen, it has two other desirable global properties:

- It is a bona fide density estimator.
- It is the unique maximum likelihood estimator confined to the subspace of functions of the form

$$\begin{aligned} g(t) &= c_k, \text{ for } t \in T_k, \\ &= 0, \text{ otherwise}, \end{aligned}$$

  and where $g(t) \geq 0$ and $\int_{\cup_k T_k} g(t)\,\mathrm{d}t = 1$.

Pointwise and Binwise Properties

Properties of the histogram vary from bin to bin. From equation (10.21), the expectation of the histogram estimator at the point y in bin T_k is

$$\mathrm{E}(\widehat{p}_H(y)) = \frac{p_k}{v_k}, \tag{10.23}$$

where

$$p_k = \int_{T_k} p(t)\,\mathrm{d}t \tag{10.24}$$

is the probability content of the k^{th} bin.

Some pointwise properties of the histogram estimator are the following:

- The **bias** of the histogram at the point y within the k^{th} bin is

$$\frac{p_k}{v_k} - p(y). \tag{10.25}$$

  Note that the bias is different from bin to bin, even if the bins are of constant size. The bias tends to decrease as the bin size decreases. We can bound the bias if we assume a regularity condition on p. If there exists γ such that for any $y_1 \neq y_2$ in an interval

$$|p(y_1) - p(y_2)| < \gamma\|y_1 - y_2\|,$$

  we say that p is Lipschitz-continuous on the interval, and for such a density, for any ξ_k in the k^{th} bin, we have

$$\begin{aligned} |\mathrm{Bias}(\widehat{p}_H(y))| &= |p(\xi_k) - p(y)| \\ &\leq \gamma_k\|\xi_k - y\| \\ &\leq \gamma_k v_k. \end{aligned} \tag{10.26}$$

- The **variance** of the histogram at the point y within the k^{th} bin is

$$\begin{aligned} \mathrm{V}\big(\widehat{p}_H(y)\big) &= \mathrm{V}(n_k)/(nv_k)^2 \\ &= \frac{p_k(1-p_k)}{nv_k^2}. \end{aligned} \tag{10.27}$$

  This is easily seen by recognizing that n_k is a binomial random variable with parameters n and p_k. Notice that the variance decreases as the bin size increases. Note also that the variance is different from bin to bin. We can bound the variance:

$$\mathrm{V}(\widehat{p}_H(y)) \leq \frac{p_k}{nv_k^2}.$$

By the mean-value theorem, we have $p_k = v_k p(\xi_k)$ for some $\xi_k \in T_k$, so we can write

$$\mathrm{V}(\widehat{p}_H(y)) \leq \frac{p(\xi_k)}{nv_k}.$$

Notice the tradeoff between bias and variance: *as h increases the variance, equation (10.27), decreases, but the bound on the bias, equation (10.26), increases.*

- The **mean squared error** of the histogram at the point y within the k^{th} bin is

$$\mathrm{MSE}\big(\widehat{p}_H(y)\big) = \frac{p_k(1-p_k)}{nv_k^2} + \left(\frac{p_k}{v_k} - p(y)\right)^2. \tag{10.28}$$

For a Lipschitz-continuous density, within the k^{th} bin we have

$$\mathrm{MSE}\big(\widehat{p}_H(y)\big) \leq \frac{p(\xi_k)}{nv_k} + \gamma_k^2 v_k^2. \tag{10.29}$$

We easily see that the histogram estimator is L_2 pointwise consistent for a Lipschitz-continuous density if, as $n \to \infty$, for each k, $v_k \to 0$ and $nv_k \to \infty$. By differentiating, we see that the minimum of the bound on the MSE in the k^{th} bin occurs for

$$h^*(k) = \left(\frac{p(\xi_k)}{2\gamma_k^2 n}\right)^{1/3}. \tag{10.30}$$

Substituting this value back into MSE, we obtain the order of the optimal MSE at the point x,

$$\mathrm{MSE}^*\big(\widehat{p}_H(y)\big) = \mathrm{O}\big(n^{-2/3}\big).$$

Asymptotic MISE (or AMISE) of Histogram Estimators

Global properties of the histogram are obtained by summing the binwise properties over all of the bins.

The expressions for the integrated variance and the integrated squared bias are quite complicated because they depend on the bin sizes and the probability content of the bins. We will first write the general expressions, and then we will assume some degree of smoothness of the true density and write approximate expressions that result from mean values or Taylor approximations. We will assume rectangular bins for additional simplification. Finally, we will then consider bins of equal size to simplify the expressions further.

First, consider the integrated variance,

$$\begin{aligned}
\mathrm{IV}\big(\widehat{p}_H\big) &= \int_{\mathbb{R}^d} \mathrm{V}(\widehat{p}_H(t))\,\mathrm{d}t \\
&= \sum_{k=1}^{m} \int_{T_k} \mathrm{V}(\widehat{p}_H(t))\,\mathrm{d}t \\
&= \sum_{k=1}^{m} \frac{p_k - p_k^2}{nv_k} \\
&= \sum_{k=1}^{m} \left(\frac{1}{nv_k} - \frac{\sum p(\xi_k)^2 v_k}{n} \right) + \mathrm{o}(n^{-1})
\end{aligned}$$

for some $\xi_k \in T_k$, as before. Now, taking $\sum p(\xi_k)^2 v_k$ as an approximation to the integral $\int (p(t))^2\,\mathrm{d}t$, and letting $\mathcal{S}$ be the functional that measures the variation in a square-integrable function of d variables,

$$\mathcal{S}(g) = \int_{\mathbb{R}^d} (g(t))^2\,\mathrm{d}t, \tag{10.31}$$

we have the integrated variance,

$$\mathrm{IV}\big(\widehat{p}_H\big) \approx \sum_{k=1}^{m} \frac{1}{nv_k} - \frac{\mathcal{S}(p)}{n}, \tag{10.32}$$

and the asymptotic integrated variance,

$$\mathrm{AIV}\big(\widehat{p}_H\big) = \sum_{k=1}^{m} \frac{1}{nv_k}. \tag{10.33}$$

The measure of the variation, $\mathcal{S}(p)$, is a measure of the roughness of the density because the density integrates to 1.

Now, consider the other term in the integrated MSE, the integrated squared bias. We will consider the case of rectangular bins, in which $h_k = (h_{k_1}, \ldots, h_{k_d})$ is the vector of lengths of sides in the k^{th} bin. In the case of rectangular bins, $v_k = \Pi_{j=1}^{d} h_{k_j}$.

We assume that the density can be expanded in a Taylor series, and we expand the density in the k^{th} bin about $\bar{t}_k$, the midpoint of the rectangular bin. For $\bar{t}_k + t \in T_k$, we have

$$p(\bar{t}_k + t) = p(\bar{t}_k) + t^{\mathrm{T}} \nabla p(\bar{t}_k) + \frac{1}{2} t^{\mathrm{T}} \mathrm{H}_p(\bar{t}_k) t + \cdots, \tag{10.34}$$

where $\mathrm{H}_p(\bar{t}_k)$ is the Hessian of p evaluated at $\bar{t}_k$.

The probability content of the k^{th} bin, p_k, from equation (10.24), can be expressed as an integral of the Taylor series expansion:

$$\begin{aligned} p_k &= \int_{\bar{t}_k+t\in T_k} p(\bar{t}_k + t)\,\mathrm{d}t \\ &= \int_{-h_{kd}/2}^{h_{kd}/2} \cdots \int_{-h_{k1}/2}^{h_{k1}/2} \left(p(\bar{t}_k) + t^{\mathrm{T}}\nabla p(\bar{t}_k) + \ldots\right) \mathrm{d}t_1 \cdots \mathrm{d}t_d \\ &= v_k p(\bar{t}_k) + \mathrm{O}\Big(h_{k*}^{d+2}\Big), \end{aligned} \tag{10.35}$$

where $h_{k*} = \min_j h_{kj}$. The bias at a point $\bar{t}_k + t$ in the k^{th} bin, after substituting equations (10.34) and (10.35) into equation (10.25), is

$$\frac{p_k}{v_k} - p(\bar{t}_k + t) = -t^{\mathrm{T}}\nabla p(\bar{t}_k) + \mathrm{O}\big(h_{k*}^2\big).$$

For the k^{th} bin the integrated squared bias is

$$\begin{aligned} &\mathrm{ISB}_k(\widehat{p}_{\mathrm{H}}) \\ &= \int_{T_k} \left(\big(t^{\mathrm{T}}\nabla p(\bar{t}_k)\big)^2 - 2\mathrm{O}\big(h_{k*}^2\big) t^{\mathrm{T}}\nabla p(\bar{t}_k) + \mathrm{O}\big(h_{k*}^4\big) \right) \mathrm{d}t \\ &= \int_{-h_{kd}/2}^{h_{kd}/2} \cdots \int_{-h_{k1}/2}^{h_{k1}/2} \sum_i \sum_j t_{ki} t_{kj} \nabla_i p(\bar{t}_k) \nabla_j p(\bar{t}_k)\,\mathrm{d}t_1 \cdots \mathrm{d}t_d + \mathrm{O}\Big(h_{k*}^{4+d}\Big). \end{aligned} \tag{10.36}$$

Many of the expressions above are simpler if we use a constant bin size, v, or $h_1, \ldots, h_d$. In the case of constant bin size, the asymptotic integrated variance in equation (10.33) becomes

$$\mathrm{AIV}\big(\widehat{p}_H\big) = \frac{m}{nv}. \tag{10.37}$$

In this case, the integral in equation (10.36) simplifies as the integration is performed term by term because the cross-product terms cancel, and the integral is

$$\frac{1}{12}(h_1 \cdots h_d) \sum_{j=1}^{d} h_j^2 \big(\nabla_j p(\bar{t}_k)\big)^2. \tag{10.38}$$

This is the asymptotic squared bias integrated over the k^{th} bin.

When we sum the expression (10.38) over all bins, the $\big(\nabla_j p(\bar{t}_k)\big)^2$ become $\mathcal{S}\big(\nabla_j p\big)$, and we have the asymptotic integrated squared bias,

$$\mathrm{AISB}\big(\widehat{p}_H\big) = \frac{1}{12} \sum_{j=1}^{d} h_j^2 \mathcal{S}\big(\nabla_j p\big). \tag{10.39}$$

Combining the asymptotic integrated variance, equation (10.37), and squared bias, equation (10.39), for the histogram with rectangular bins of constant size, we have

$$\mathrm{AMISE}\big(\widehat{p}_H\big) = \frac{1}{n(h_1 \cdots h_d)} + \frac{1}{12}\sum_{j=1}^{d} h_j^2 \mathcal{S}\big(\nabla_j p\big). \tag{10.40}$$

As we have seen before, smaller bin sizes increase the variance but decrease the squared bias.

Bin Sizes

As we have mentioned and have seen by example, the histogram is very sensitive to the bin sizes, both in appearance and in other properties. Equation (10.40) for the AMISE assuming constant rectangular bin size is often used as a guide for determining the bin size to use when constructing a histogram. This expression involves $\mathcal{S}\big(\nabla_j p\big)$ and so, of course, cannot be used directly. Nevertheless, differentiating the expression with respect to h_j and setting the result equal to zero, we have the bin width that is optimal with respect to the AMISE,

$$h_{j*} = \mathcal{S}\big(\nabla_j p\big)^{-1/2} \left(6 \prod_{i=1}^{d} \mathcal{S}\big(\nabla_i p\big)^{1/2}\right)^{\frac{1}{2+d}} n^{-\frac{1}{2+d}}. \tag{10.41}$$

Substituting this into equation (10.40), we have the optimal value of the AMISE

$$\frac{1}{4}\left(36 \prod_{i=1}^{d} \mathcal{S}\big(\nabla_i p\big)^{1/2}\right)^{\frac{1}{2+d}} n^{-\frac{2}{2+d}}. \tag{10.42}$$

Notice that the optimal rate of decrease of AMISE for histogram estimators is $\mathrm{O}(n^{-\frac{2}{2+d}})$. Although histograms have several desirable properties, this order of convergence is not good compared to that of some other bona fide density estimators, as we will see in later sections.

The expression for the optimal bin width involves $\mathcal{S}\big(\nabla_j p\big)$, where p is the unknown density. An approach is to choose a value for $\mathcal{S}\big(\nabla_j p\big)$ that corresponds to some good general distribution. A "good general distribution", of course, is the normal with a diagonal variance-covariance matrix. For the d-variate normal with variance-covariance matrix $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$,

$$\mathcal{S}\big(\nabla_j p\big) = \frac{1}{2^{d+1}\pi^{d/2}\sigma_j^2 |\Sigma|^{1/2}}.$$

For a univariate normal density with variance σ^2,

$$\mathcal{S}(p') = 1/(4\sqrt{\pi}\sigma^3),$$

so the optimal constant one-dimensional bin width under the AMISE criterion is

$$3.49\sigma n^{-1/3}.$$

In practice, of course, an estimate of σ must be used. The sample standard deviation s is one obvious choice. A more robust estimate of the scale is based on the sample interquartile range, r. The sample interquartile range leads to a bin width of $2rn^{-1/3}$.

The AMISE is essentially an L_2 measure. The L_∞ criterion—that is, the sup absolute error (SAE) of equation (10.10)—also leads to an asymptotically optimal bin width that is proportional to $n^{-1/3}$.

One of the most commonly used rules is for the number of bins rather than the width. Assume a symmetric binomial model for the bin counts, that is, the bin count is just the binomial coefficient. The total sample size n is

$$\sum_{k=0}^{m-1} \binom{m-1}{k} = 2^{m-1},$$

and so the number of bins is

$$m = 1 + \log_2 n.$$

Bin Shapes

In the univariate case, histogram bins may vary in size, but each bin is an interval. For the multivariate case, there are various possibilities for the shapes of the bins. The simplest shape is the direct extension of an interval, that is a hyperrectangle. The volume of a hyperrectangle is just $v_k = \prod h_{kj}$. There are, of course, other possibilities; any tessellation of the space would work. The objects may or may not be regular, and they may or may not be of equal size. Regular, equal-sized geometric figures such as hypercubes have the advantages of simplicity, both computationally and analytically. In two dimensions, there are three possible regular tessellations: triangles, squares, and hexagons.

For hyperrectangles of constant size, the univariate theory generally extends fairly easily to the multivariate case. The histogram density estimator is

$$\widehat{p}_H(y) = \frac{n_k}{nh_1h_2\cdots h_d}, \quad \text{for } y \in T_k,$$

where the h's are the lengths of the sides of the rectangles. The variance within the k^{th} bin is

$$\mathrm{V}(\widehat{p}_H(y)) = \frac{np_k(1-p_k)}{(nh_1h_2\cdots h_d)^2}, \quad \text{for } y \in T_k,$$

and the integrated variance is

$$\mathrm{IV}(\widehat{p}_H) \approx \frac{1}{nh_1h_2\cdots h_d} - \frac{\mathcal{S}(f)}{n}.$$

Other Density Estimators Related to the Histogram

There are several variations of the histogram that are useful as probability density estimators. The most common modification is to connect points on the histogram by a continuous curve. A simple way of doing this in the univariate case leads to the *frequency polygon*. This is the piecewise linear curve that connects the midpoints of the bins of the histogram. The endpoints are usually zero values at the midpoints of two appended bins, one on either side.

The *histospline* is constructed by interpolating knots of the empirical CDF with a cubic spline and then differentiating it. More general methods use splines or orthogonal series to fit the histogram.

As we have mentioned and have seen by example, the histogram is somewhat sensitive in appearance to the location of the bins. To overcome the problem of location of the bins, a density estimator that is the average of several histograms with equal bin widths but different bin locations can be used. This is called the *average shifted histogram*, or ASH. It also has desirable statistical properties, and it is computationally efficient in the multivariate case.

10.2.3 Kernel Estimators

Kernel methods are probably the most widely used technique for building nonparametric probability density estimators. They are best understood by developing them as a special type of histogram. The difference is that the bins in kernel estimators are centered at the points at which the estimator is to be computed. The problem of the choice of location of the bins in histogram estimators does not arise.

Rosenblatt's Histogram Estimator; Kernels

For the one-dimensional case, Rosenblatt defined a histogram that is shifted to be centered on the point at which the density is to be estimated. Given the sample $y_1, \ldots, y_n$, Rosenblatt's histogram estimator at the point y is

$$\widehat{p}_R(y) = \frac{\#\{y_i \text{ s.t. } y_i \in (y - h/2, \quad y + h/2]\}}{nh}. \tag{10.43}$$

This histogram estimator avoids the ordinary histogram's constant-slope contribution to the bias. This estimator is a step function with variable lengths of the intervals that have constant value.

Rosenblatt's centered histogram can also be written in terms of the ECDF:

$$\widehat{p}_R(y) = \frac{P_n(y + h/2) - P_n(y - h/2)}{h},$$

where, as usual, P_n denotes the ECDF. As seen in this expression, Rosenblatt's estimator is a centered finite-difference approximation to the derivative of the

empirical cumulative distribution function (which, of course, is not differentiable at the data points). We could, of course, use the same idea and form other density estimators using other finite-difference approximations to the derivative of P_n.

Another way to write Rosenblatt's shifted histogram estimator over bins of length h is

$$\widehat{p}_R(y) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{y - y_i}{h}\right), \tag{10.44}$$

where $K(t) = 1$ if $|t| < 1/2$ and $= 0$ otherwise. The function K is a kernel or filter. In Rosenblatt's estimator, it is a "boxcar" function, but other kernel functions could be used.

The estimator extends easily to the multivariate case. In the general kernel estimator, we usually use a more general scaling of $y - y_i$,

$$V^{-1}(y - y_i),$$

for some positive-definite matrix V. The determinant of V^{-1} scales the estimator to account for the scaling within the kernel function. The general kernel estimator is given by

$$\widehat{p}_K(y) = \frac{1}{n|V|} \sum_{i=1}^{n} K\left(V^{-1}(y - y_i)\right), \tag{10.45}$$

where the function K is called the *kernel*, and V is the *smoothing matrix*. The determinant of the smoothing matrix is exactly analogous to the bin volume in a histogram estimator. The univariate version of the kernel estimator is the same as Rosenblatt's estimator (10.44), but in which a more general function K is allowed.

In practice, V is usually taken to be constant for a given sample size, but, of course, there is no reason for this to be the case, and indeed it may be better to vary V depending on the number of observations near the point y. The dependency of the smoothing matrix on the sample size n and on y is often indicated by the notation $V_n(y)$.

Properties of Kernel Estimators

The appearance of the kernel density estimator depends to some extent on the support and shape of the kernel. Unlike the histogram estimator, the kernel density estimator may be continuous and even smooth.

It is easy to see that if the kernel satisfies

$$K(t) \geq 0, \tag{10.46}$$

and

$$\int_{\mathbb{R}^d} K(t)\,\mathrm{d}t = 1 \tag{10.47}$$

(that is, if K is a density), then $\widehat{p}_K(y)$ is a bona fide density estimator.

There are other requirements that we may impose on the kernel either for the theoretical properties that result or just for their intuitive appeal. It also seems reasonable that in estimating the density at the point y, we would want to emphasize the sample points near y. This could be done in various ways, but one simple way is to require

$$\int_{\mathbb{R}^d} tK(t)\,\mathrm{d}t = 0. \tag{10.48}$$

In addition, we may require the kernel to be symmetric about 0.

For multivariate density estimation, the kernels are usually chosen as a radially symmetric generalization of a univariate kernel. Such a kernel can be formed as a product of the univariate kernels. For a product kernel, we have for some constant σ_K^2,

$$\int_{\mathbb{R}^d} tt^{\mathrm{T}}K(t)\,\mathrm{d}t = \sigma_K^2 I_d, \tag{10.49}$$

where I_d is the identity matrix of order d. We could also impose this as a requirement on any kernel, whether it is a product kernel or not. This makes the expressions for bias and variance of the estimators simpler. The spread of the kernel can always be controlled by the smoothing matrix V, so sometimes, for convenience, we require $\sigma_K^2 = 1$.

In the following, we will assume the kernel satisfies the properties in equations (10.46) through (10.49).

The pointwise properties of the kernel estimator are relatively simple to determine because the estimator at a point is merely the sample mean of n independent and identically distributed random variables. The expectation of the kernel estimator (10.45) at the point y is the convolution of the kernel function and the probability density function,

$$\begin{aligned}\mathrm{E}\left(\widehat{p}_K(y)\right) &= \frac{1}{|V|}\int_{\mathbb{R}^d} K\left(V^{-1}(y-t)\right)p(t)\,\mathrm{d}t \\ &= \int_{\mathbb{R}^d} K(u)p(y-Vu)\,\mathrm{d}u,\end{aligned} \tag{10.50}$$

where $u = V^{-1}(y-t)$ (and, hence, $\mathrm{d}u = |V|^{-1}\mathrm{d}t$).

If we approximate $p(y-Vu)$ about y with a three-term Taylor series, using the properties of the kernel in equations (10.46) through (10.49) and using properties of the trace, we have

$$\begin{aligned}\mathrm{E}\left(\widehat{p}_K(y)\right) &\approx \int_{\mathbb{R}^d} K(u)\left(p(y) - (Vu)^{\mathrm{T}}\nabla p(y) + \frac{1}{2}(Vu)^{\mathrm{T}}\mathrm{H}_p(y)Vu\right)\,\mathrm{d}u \\ &= p(y) - 0 + \frac{1}{2}\mathrm{trace}\left(V^{\mathrm{T}}\mathrm{H}_p(y)V\right).\end{aligned} \tag{10.51}$$

To second order in the elements of V (that is, $\mathrm{O}(|V|^2)$), the bias at the point y is therefore

$$\frac{1}{2}\mathrm{trace}\left(VV^{\mathrm{T}}\mathrm{H}_p(y)\right). \tag{10.52}$$

Using the same kinds of expansions and approximations as in equations (10.50) and (10.51) to evaluate $\mathrm{E}\left((\widehat{p}_K(y))^2\right)$ to get an expression of order $\mathrm{O}(|V|/n)$, and subtracting the square of the expectation in equation (10.51), we get the approximate variance at y as

$$\mathrm{V}\left(\widehat{p}_K(y)\right) \approx \frac{p(y)}{n|V|}\int_{\mathbb{R}^d}(K(u))^2\,\mathrm{d}u,$$

or

$$\mathrm{V}\left(\widehat{p}_K(y)\right) \approx \frac{p(y)}{n|V|}\mathcal{S}(K). \tag{10.53}$$

Integrating this, because p is a density, we have

$$\mathrm{AIV}\big(\widehat{p}_K\big) = \frac{\mathcal{S}(K)}{n|V|}, \tag{10.54}$$

and integrating the square of the asymptotic bias in expression (10.52), we have

$$\mathrm{AISB}\big(\widehat{p}_K\big) = \frac{1}{4}\int_{\mathbb{R}^d}\left(\mathrm{trace}\left(V^{\mathrm{T}}\mathrm{H}_p(y)V\right)\right)^2\mathrm{d}y. \tag{10.55}$$

These expressions are much simpler in the univariate case, where the smoothing matrix V is the smoothing parameter or window width h. We have a simpler approximation for $\mathrm{E}\left(\widehat{p}_K(y)\right)$ than that given in equation (10.51),

$$\mathrm{E}\left(\widehat{p}_K(y)\right) \approx p(y) + \frac{1}{2}h^2p''(y)\int_{\mathbb{R}}u^2K(u)\,\mathrm{d}u,$$

and from this we get a simpler expression for the AISB. After likewise simplifying the AIV, we have

$$\mathrm{AMISE}\big(\widehat{p}_K\big) = \frac{\mathcal{S}(K)}{nh} + \frac{1}{4}\sigma_K^4h^4\mathcal{R}(p), \tag{10.56}$$

where we have left the kernel unscaled (that is, $\int u^2K(u)\,\mathrm{d}u = \sigma_K^2$).

Minimizing this with respect to h, we have the optimal value of the smoothing parameter

$$\left(\frac{\mathcal{S}(K)}{n\sigma_K^4\mathcal{R}(p)}\right)^{1/5}. \tag{10.57}$$

Substituting this back into the expression for the AMISE, we find that its optimal value in this univariate case is

$$\frac{5}{4}\mathcal{R}(p)(\sigma_K\mathcal{S}(K))^{4/5}\,n^{-4/5}. \tag{10.58}$$

The AMISE for the univariate kernel density estimator is thus $O(n^{-4/5})$. Recall that the AMISE for the univariate histogram density estimator is $O(n^{-2/3})$.

We see that the bias and variance of kernel density estimators have similar relationships to the smoothing matrix that the bias and variance of histogram estimators have. As the determinant of the smoothing matrix gets smaller (that is, as the window of influence around the point at which the estimator is to be evaluated gets smaller), the bias becomes smaller and the variance becomes larger. This agrees with what we would expect intuitively.

Choice of Kernels

Standard normal densities have these properties described above, so the kernel is often chosen to be the standard normal density. As it turns out, the kernel density estimator is not very sensitive to the form of the kernel.

Although the kernel may be from a parametric family of distributions, in kernel density estimation, we do not estimate those parameters; hence, the kernel method is a nonparametric method.

Sometimes, a kernel with finite support is easier to work with. In the univariate case, a useful general form of a compact kernel is

$$K(t) = \kappa_{rs}(1 - |t|^r)^s \mathrm{I}_{[-1,1]}(t),$$

where

$$\kappa_{rs} = \frac{r}{2\mathrm{B}(1/r, s+1)}, \quad \text{for } r > 0,\ s \geq 0,$$

and $\mathrm{B}(a, b)$ is the complete beta function.

This general form leads to several simple specific cases:

- for $r = 1$ and $s = 0$, it is the rectangular kernel;
- for $r = 1$ and $s = 1$, it is the triangular kernel;
- for $r = 2$ and $s = 1$ ($\kappa_{rs} = 3/4$), it is the "Epanechnikov" kernel, which yields the optimal rate of convergence of the MISE (see Epanechnikov, 1969);
- for $r = 2$ and $s = 2$ ($\kappa_{rs} = 15/16$), it is the "biweight" kernel.

If $r = 2$ and $s \to \infty$, we have the Gaussian kernel (with some rescaling).

As mentioned above, for multivariate density estimation, the kernels are often chosen as a product of the univariate kernels. The product Epanechnikov kernel, for example, is

$$K(t) = \frac{d+2}{2c_d}(1 - t^{\mathrm{T}}t)\mathrm{I}_{(t^{\mathrm{T}}t \leq 1)},$$

where

$$c_d = \frac{\pi^{d/2}}{\Gamma(d/2+1)}.$$

We have seen that the AMISE of a kernel estimator (that is, the sum of equations (10.54) and (10.55)) depends on $\mathcal{S}(K)$ and the smoothing matrix V. As we mentioned above, the amount of smoothing (that is, the window of influence) can be made to depend on σ_K. We can establish an approximate equivalence between two kernels, K_1 and K_2, by choosing the smoothing matrix to offset the differences in $\mathcal{S}(K_1)$ and $\mathcal{S}(K_2)$ and in σ_{K_1} and σ_{K_2}.

Computation of Kernel Density Estimators

If the estimate is required at one point only, it is simplest just to compute it directly. If the estimate is required at several points, it is often more efficient to compute the estimates in some regular fashion.

If the estimate is required over a grid of points, a fast Fourier transform (FFT) can be used to speed up the computations.

10.2.4 Choice of Window Widths

An important problem in nonparametric density estimation is to determine the smoothing parameter, such as the bin volume, the smoothing matrix, the number of nearest neighbors, or other measures of locality. In kernel density estimation, the window width has a much greater effect on the estimator than the kernel itself does.

An objective is to choose the smoothing parameter that minimizes the MISE. We often can do this for the AMISE, as in equation (10.41) on page 312. It is not as easy for the MISE. The first problem, of course, is just to estimate the MISE.

In practice, we use cross validation with varying smoothing parameters and alternate computations between the MISE and AMISE.

In univariate density estimation, the MISE has terms such as $h^\alpha \mathcal{S}(p')$ (for histograms) or $h^\alpha \mathcal{S}(p'')$ (for kernels). We need to estimate the roughness of a derivative of the density.

Using a histogram, a reasonable estimate of the integral $\mathcal{S}(p')$ is a Riemann approximation,

$$\begin{aligned}\widehat{\mathcal{S}}(p') &= h \sum \big(\widehat{p}'(t_k)\big)^2 \\ &= \frac{1}{n^2 h^3} \sum (n_{k+1} - n_k)^2,\end{aligned}$$

where $\widehat{p}'(t_k)$ is the finite difference at the midpoints of the k^{th} and $(k+1)^{\text{th}}$ bins; that is,

$$\widehat{p}'(t_k) = \frac{n_{k+1}/(nh) - n_k/(nh)}{h}.$$

This estimator is biased. For the histogram, for example,

$$\mathrm{E}(\widehat{\mathcal{S}}(p')) = \mathcal{S}(p') + 2/(nh^3) + \ldots$$

A standard estimation scheme is to correct for the $2/(nh^3)$ term in the bias and plug this back into the formula for the AMISE (which is $1/(nh) + h^2\mathcal{S}(r')/12$ for the histogram).

We compute the estimated values of the AMISE for various values of h and choose the one that minimizes the AMISE. This is called *biased cross validation* because of the use of the AMISE rather than the MISE.

These same techniques can be used for other density estimators and for multivariate estimators, although at the expense of considerably more complexity.

10.2.5 Orthogonal Series Estimators

A continuous real function $p(x)$, integrable over a domain D, can be represented over that domain as an infinite series in terms of a complete spanning set of real orthogonal functions $\{f_k\}$ over D:

$$p(x) = \sum_k c_k f_k(x). \tag{10.59}$$

The orthogonality property allows us to determine the coefficients c_k in the expansion (10.59):

$$c_k = \langle f_k, p \rangle. \tag{10.60}$$

Approximation using a truncated orthogonal series can be particularly useful in estimation of a probability density function because the orthogonality relationship provides an equivalence between the coefficient and an expected value. Expected values can be estimated using observed values of the random variable and the approximation of the probability density function. Assume that the probability density function p is approximated by an orthogonal series $\{q_k\}$ with weight function $w(y)$:

$$p(y) = \sum_k c_k q_k(y).$$

From equation (10.60), we have

$$\begin{aligned} c_k &= \langle q_k, p \rangle \\ &= \int_D q_k(y)p(y)w(y)\mathrm{d}y \\ &= \mathrm{E}(q_k(Y)w(Y)), \end{aligned} \tag{10.61}$$

where Y is a random variable whose probability density function is p.

The c_k can therefore be unbiasedly estimated by

$$\widehat{c}_k = \frac{1}{n}\sum_{i=1}^{n} q_k(y_i)w(y_i).$$

The orthogonal series estimator is therefore

$$\widehat{p}_S(y) = \frac{1}{n}\sum_{k=0}^{j}\sum_{i=1}^{n} q_k(y_i)w(y_i)q_k(y) \tag{10.62}$$

for some truncation point j.

Without some modifications, this generally is not a good estimator of the probability density function. It may not be smooth, and it may have infinite variance. The estimator may be improved by shrinking the $\widehat{c}_k$ toward the origin. The number of terms in the finite series approximation also has a major effect on the statistical properties of the estimator. Having more terms is not necessarily better. One useful property of orthogonal series estimators is that the convergence rate is independent of the dimension. This may make orthogonal series methods more desirable for higher-dimensional problems.

There are several standard orthogonal series that could be used. The two most commonly used series are the Fourier and the Hermite. Which is preferable depends on the situation.

The Fourier series is commonly used for distributions with bounded support. It yields estimators with better properties in the L_1 sense.

For distributions with unbounded support, the Hermite polynomials are most commonly used.

Notes

Exercises in Shao

- For practice and discussion
  5.15, 5.16, 5.17, 5.23 (Solutions in Shao, 2005)
- To turn in
  5.18, 5.19

Additional References

Scott, David W. (1992), *Multivariate Density Estimation*, John Wiley & Sons, New York.

Appendices

A

Notation and Definitions

All notation used in this work is "standard". I have opted for simple notation, which, of course, results in a one-to-many map of notation to object classes. Within a given context, however, the overloaded notation is generally unambiguous. I have endeavored to use notation consistently.

This appendix is not intended to be a comprehensive listing of definitions.

A.1 General Notation

Uppercase italic Latin and Greek letters, such as A, B, E, Λ, etc., are generally used to represent sets, random variables, and matrices. Realizations of random variables and placeholders in functions associated with random variables are usually represented by lowercase letters corresponding to the uppercase letters; thus, ϵ may represent a realization of the random variable E.

Parameters in models (that is, unobservables in the models) are generally represented by Greek letters. Uppercase Latin and Greek letters are also used to represent cumulative distribution functions. Symbols whose meaning is context-independent are usually written in an upright font, whereas symbols representing variables are written in a slant or italic font; for example, Γ is used to represent the gamma function, while $\mathit{\Gamma}$ may be used to represent a variable or a parameter. An upright font is also used to represent a special object, such as a sample space or a parameter space.

Lowercase Latin and Greek letters are used to represent ordinary scalar or vector variables and functions. **No distinction in the notation is made between scalars and vectors**; thus, β may represent a vector and β_i may represent the i^{th} element of the vector β. In another context, however, β may represent a scalar. All vectors are considered to be column vectors, although we may write a vector as $x = (x_1, x_2, \ldots, x_n)$. Transposition of a vector or a matrix is denoted by the superscript "T".

Uppercase calligraphic Latin letters, such as $\mathcal{D}$, $\mathcal{V}$, and $\mathcal{W}$, are generally used to represent special collections of sets, vector spaces, or transforms (functionals).

A single symbol in an italic font is used to represent a single variable. A Roman font or a special font is often used to represent a standard operator or a standard mathematical structure. Sometimes a string of symbols in a Roman font is used to represent an operator (or a standard function); for example, $\exp(\cdot)$ represents the exponential function. But a string of symbols in an italic font on the same baseline should be interpreted as representing a composition (probably by multiplication) of separate objects; for example, *exp* represents the product of e, x, and p. Likewise a string of symbols in a Roman font (usually a single symbol) is used to represent a fundamental constant; for example, e represents the base of the natural logarithm, while e represents a variable.

Subscripts generally represent indexes to a larger structure; for example, x_{ij} may represent the $(i,j)^{\text{th}}$ element of a matrix, X. A subscript in parentheses represents an order statistic. A superscript in parentheses represents an iteration; for example, $x_i^{(k)}$ may represent the value of x_i at the k^{th} step of an iterative process.

x_i	The i^{th} element of a structure (including a sample, which is a multiset).
$x_{(i)}$	The i^{th} order statistic.
$x^{(i)}$	The value of x at the i^{th} iteration.

Some important mathematical structures and other objects are:

$\mathbb{R}$	The field of reals or the set over which that field is defined.
$\mathbb{R}^*$	The "extended reals"; $\mathbb{R}^* = \mathbb{R} \cup \{-\infty, \infty\}$.
$\mathbb{R}^d$	The usual d-dimensional vector space over the reals or the set of all d-tuples with elements in $\mathbb{R}$.
$\mathbb{Z}$	The ring of integers or the set over which that ring is defined.
e	The base of the natural logarithm. This is a constant; e may be used to represent a variable. (Note the difference in the font.)

i	The imaginary unit, $\sqrt{-1}$. This is a constant; i may be used to represent a variable. (Note the difference in the font.)

A.2 General Mathematical Functions and Operators

Functions such as sin, max, span, and so on that are commonly associated with groups of Latin letters are generally represented by those letters in a Roman font.

Operators such as d (the differential operator) that are commonly associated with a Latin letter are generally represented by that letter in a Roman font.

Note that some symbols, such as $|\cdot|$, are overloaded; such symbols are generally listed together below.

$\times$	Cartesian or cross product of sets, or multiplication of elements of a field or ring.
$\lvert x\rvert$	The modulus of the real or complex number x; if x is real, $\lvert x\rvert$ is the absolute value of x.
$\lceil x\rceil$	The ceiling function evaluated at the real number x: $\lceil x\rceil$ is the largest integer less than or equal to x.
$\lfloor x\rfloor$	The floor function evaluated at the real number x: $\lfloor x\rfloor$ is the smallest integer greater than or equal to x.
$x!$	The factorial of x. If x is a positive integer, $x! = x(x-1)\cdots 2\cdot 1$.
$x^{[r]}$	The r^{th} factorial of x. If x is a positive integer, $x^{[r]} = x(x-1)\cdots 2\cdot(x-(r-1))$.
$\mathrm{O}(f(n))$	Big O; $g(n) = \mathrm{O}(f(n))$ means $g(n)/f(n) \to c$ as $n \to \infty$, where c is a nonzero finite constant. In particular, $g(n) = \mathrm{O}(1)$ means $g(n)$ is bounded.
$\mathrm{o}(f(n))$	Little o; $g(n) = \mathrm{o}(f(n))$ means $g(n)/f(n) \to 0$ as $n \to \infty$. In particular, $g(n) = \mathrm{o}(1)$ means $g(n) \to 0$.
$\mathrm{o}_P(f(n))$	Convergent in probability; $X(n) = \mathrm{o}_P(f(n))$ means that for any positive ϵ, $\Pr(\lvert X(n) - f(n)\rvert > \epsilon) \to 0$ as $n \to \infty$.

d	The differential operator.
Δ	A perturbation operator; Δx represents a perturbation of x and not a multiplication of x by Δ, even if x is a type of object for which a multiplication is defined.
$\Delta(\cdot,\cdot)$	A real-valued difference function; $\Delta(x,y)$ is a measure of the difference of x and y. For simple objects, $\Delta(x,y) = \lvert x-y\rvert$. For more complicated objects, a subtraction operator may not be defined, and Δ is a generalized difference.
$\tilde{x}$	A perturbation of the object x; $\Delta(x,\tilde{x}) = \Delta x$.
$\tilde{x}$	An average of a sample of objects generically denoted by x.
$\bar{x}$	The mean of a sample of objects generically denoted by x.

Special Functions

Good general references on special functions in mathematics are Abramowitz and Stegun (1964) and Thompson(1997). The venerable book edited by Abramowitz and Stegun has been kept in print by Dover Publications.

$\log x$	The natural logarithm evaluated at x.
$\sin x$	The sine evaluated at x (in radians) and similarly for other trigonometric functions.

$\Gamma(\alpha)$ The complete gamma function:

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}\mathrm{e}^{-t}\mathrm{d}t.$$

(This is called Euler's integral.) Integration by parts immediately gives the replication formula $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$, and so if α is a positive integer, $\Gamma(\alpha+1) = \alpha!$, and more generally, $\Gamma(\alpha+1)$ defines $\alpha!$.
Direct evaluation of the integral yields $\Gamma(1/2) = \sqrt{\pi}$. Using this and the replication formula, with some manipulation we get for the positive integer j

$$\Gamma(j+1/2) = \frac{1\cdot 2\cdots(2j-1)}{2^j}\sqrt{\pi}.$$

The notation $\Gamma_d(\alpha)$ denotes the multivariate gamma function, where α is a d-vector. (In other literature this notation denotes the incomplete univariate gamma function.)

Associated with the gamma function are some other useful functions:

$\psi(\alpha)$ The digamma function:

$$\psi(\alpha) = \mathrm{d}\log(\Gamma(\alpha))/\mathrm{d}\alpha.$$

$\psi'(\alpha)$ The trigamma function,

$$\psi'(\alpha) = \mathrm{d}\psi(\alpha)/\mathrm{d}\alpha.$$

More general are the polygamma functions, for $n = 1, 2, \ldots$, $\psi^{(n)}(\alpha) = \mathrm{d}^{(n)}\psi(\alpha)/(\mathrm{d}\alpha)^{(n)}$, which for a fixed n, is called the $(n+2)$-gamma function.

$\gamma(\alpha, x)$ The incomplete gamma function,

$$\Gamma(x) = \int_0^\infty t^{\alpha-1}\mathrm{e}^{-t}\mathrm{d}t.$$

$P(\alpha, x)$ The regularized incomplete gamma function, which is the CDF of the standard gamma distribution,

$$P(\alpha, x) = \gamma(\alpha, x)/\Gamma(\alpha).$$

$\mathrm{B}(\alpha, \beta)$ The beta function,

$$\mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}.$$

A.3 Sets, Measure, and Probability

The notation listed below does not always represent the things associated with it here, but for these objects, I generally use either this notation or other symbols in the same font.

A° The set of interior points of the set A.

$\overline{A}$ The set of closure points of the set A.

∂A The set of closure points of the set A. We have $\partial A = \overline{A} - A^\circ$.

Ω Sample space; the universal set in a given probability distribution.

$\mathcal{F}$ A σ-field.

$\mathcal{B}$ The Borel σ-field.

$\mathcal{B}_I$ The Borel σ-field restricted to the interval I; that is, the σ-field generated by all open intervals contained in I and $\Omega = I$.

$(\Omega, \mathcal{F})$ A measurable space: the sample space Ω and the σ-field $\mathcal{F}$.

$(\Omega, \mathcal{F}, \nu)$ A measure space: the sample space Ω, the σ-field $\mathcal{F}$, and the measure ν defined over the sets in $\mathcal{F}$.

$\lambda \ll \nu$ — The measure ν dominates the measure λ; that is, λ is *absolutely continuous with respect to* ν*:*

$$\nu(A) = 0 \quad \Rightarrow \quad \lambda(A) = 0,$$

for any set in the domain of both λ *and* ν*.*

$(\Omega, \mathcal{F}, P)$ — The "probability triple": the sample space Ω, the σ-field $\mathcal{F}$, and the probability measure P.

$\mathcal{P}$ — A family of probability distributions.

Θ — Parameter space.

$\mathcal{X}$ — The range of a random variable.

A.4 Linear Spaces and Matrices

$\mathcal{V}(G)$ — For the set of vectors (all of the same order) G, the vector space generated by that set.

$\mathcal{V}(X)$ — For the matrix X, the vector space generated by the columns of X.

$\dim(\mathcal{V})$ — The dimension of the vector space $\mathcal{V}$; that is, the maximum number of linearly independent vectors in the vector space.

$\text{span}(Y)$ — For Y either a set of vectors or a matrix, the vector. space $\mathcal{V}(Y)$

$\text{tr}(A)$ — The trace of the square matrix A, that is, the sum of the diagonal elements.

$\text{rank}(A)$ — The rank of the matrix A, that is, the maximum number of independent rows (or columns) of A.

$\rho(A)$ — The spectral radius of the matrix A (the maximum absolute value of its eigenvalues).

$A > 0$
$A \geq 0$ — If A is a matrix, this notation means, respectively, that each element of A is positive or nonnegative.

$A \succ 0$ $A \succeq 0$	This notation means that A is a symmetric matrix and that it is, respectively, positive definite or nonnegative definite.
A^{T}	For the matrix A, its transpose (also used for a vector to represent the corresponding row vector).
A^{H}	The conjugate transpose, also called the adjoint, of the matrix A; $A^{\mathrm{H}} = \bar{A}^{\mathrm{T}} = \overline{A^{\mathrm{T}}}$.
A^{-1}	The inverse of the square, nonsingular matrix A.
$A^{-\mathrm{T}}$	The inverse of the transpose of the square, nonsingular matrix A.
A^{+}	The g_4 inverse, the Moore-Penrose inverse, or the pseudoinverse of the matrix A.
A^{-}	A g_1, or generalized, inverse of the matrix A.
$A^{\frac{1}{2}}$	The square root of a nonnegative definite or positive definite matrix A; $(A^{\frac{1}{2}})^2 = A$.
$A^{-\frac{1}{2}}$	The square root of the inverse of a positive definite matrix A; $(A^{-\frac{1}{2}})^2 = A^{-1}$.

Norms and Inner Products

L_p	For real $p \geq 1$, a norm formed by accumulating the p^{th} powers of the moduli of individual elements in an object and then taking the $(1/p)^{\text{th}}$ power of the result.
$\|\cdot\|$	In general, the norm of the object $\cdot$.
$\|\cdot\|_p$	In general, the L_p norm of the object $\cdot$.
$\|x\|_p$	For the vector x, the L_p norm $$\|x\|_p = \left(\sum |x_i|^p\right)^{\frac{1}{p}}.$$

$\|X\|_p$ For the matrix X, the L_p norm

$$\|X\|_p = \max_{\|v\|_p=1} \|Xv\|_p.$$

$\|X\|_F$ For the matrix X, the Frobenius norm

$$\|X\|_F = \sqrt{\sum_{i,j} x_{ij}^2}.$$

$\langle x, y\rangle$ The inner product or dot product of x and y.

$\kappa_p(A)$ The L_p condition number of the nonsingular square matrix A with respect to inversion.

Notation Relating to Matrix Determinants

$|A|$ The determinant of the square matrix A, $|A| = \det(A)$.

$\det(A)$ The determinant of the square matrix A, $\det(A) = |A|$.

$|A_{(i_1,\ldots,i_k)}|$ A principal minor of a square matrix A; in this case, it is the minor corresponding to the matrix formed from rows $i_1, \ldots, i_k$ and columns $i_1, \ldots, i_k$ from a given matrix A.

$|A_{-(i)(j)}|$ The minor associated with the $(i,j)^{\text{th}}$ element of a square matrix A.

$a_{(ij)}$ The cofactor associated with the $(i,j)^{\text{th}}$ element of a square matrix A; that is, $a_{(ij)} = (-1)^{i+j}|A_{-(i)(j)}|$.

$\mathrm{adj}(A)$ The adjugate, also called the classical adjoint, of the square matrix A: $\mathrm{adj}(A) = (a_{(ji)})$; that is, the matrix of the same size as A formed from the cofactors of the elements of A^{T}.

Matrix-Vector Differentiation

$\mathrm{d}t$	The differential operator on the scalar, vector, or matrix t. This is an operator; d may be used to represent a variable. (Note the difference in the font.)
g_f or ∇f	For the scalar-valued function f of a vector variable, the vector whose i^{th} element is $\partial f/\partial x_i$. This is the gradient, also often denoted as g_f.
∇f	For the vector-valued function f of a vector variable, the matrix whose element in position (i, j) is $$\frac{\partial f_j(x)}{\partial x_i}.$$ This is also written as $\partial f^{\mathrm{T}}/\partial x$ or just as $\partial f/\partial x$. This is the transpose of the Jacobian of f.
J_f	For the vector-valued function f of a vector variable, the Jacobian of f denoted as J_f. The element in position (i, j) is $$\frac{\partial f_i(x)}{\partial x_j}.$$ This is the transpose of (∇f): $\mathrm{J}_f = (\nabla f)^{\mathrm{T}}$.
H_f or $\nabla\nabla f$ or $\nabla^2 f$	The Hessian of the scalar-valued function f of a vector variable. The Hessian is the transpose of the Jacobian of the gradient. Except in pathological cases, it is symmetric. The element in position (i, j) is $$\frac{\partial^2 f(x)}{\partial x_i \partial x_j}.$$ The symbol $\nabla^2 f$ is sometimes also used to denote the trace of the Hessian, in which case it is called the Laplacian.

B

Important Probability Distributions

C

Notation and Definitions

Development of stochastic models is facilitated by identifying a few probability distributions that seem to correspond to a variety of data-generating processes, and then studying the properties of these distributions. In the following tables, I list some of the more useful distributions, both discrete distributions and continuous ones. The names listed are the most common names, although some distributions go by different names, especially for specific values of the parameters. In the first column, following the name of the distribution, the parameter space is specified. Also, given in the first column is the root name of the computer routines in both R and IMSL that apply to the distribution. In the last column, the PDF (or probability mass function) and the mean and variance are given.

There are two very special continuous distributions, for which I use special symbols: the unit uniform, designated $\mathrm{U}(0,1)$, and the normal (or Gaussian), denoted by $\mathrm{N}(\mu, \sigma^2)$. Notice that the second parameter in the notation for the normal is the variance. Sometimes, such as in the functions in R, the second parameter of the normal distribution is the standard deviation instead of the variance.

Except for the uniform and the normal, I designate distributions by a name followed by symbols for the parameters, for example, binomial(n, π) or gamma(α, β). Some families of distributions are subfamilies of larger families. For example, the usual gamma family of distributions is a the two-parameter subfamily of the three-parameter gamma.

Evans, Hastings, and Peacock (2000) give general descriptions of 40 probability distributions. Leemis and McQueston (2008) provide a compact graph of the relationships among a large number of probability distributions.

Table C.1. Discrete Distributions (PDFs are w.r.t counting measure)

discrete uniform	PDF	$\frac{1}{m}, \quad y = a_1, \ldots, a_m$
$a_1, \ldots, a_m \in \mathbb{R}$	mean	$\sum a_i/m$
R: `sample`; IMSL: `und`	variance	$\sum(a_i - \bar{a})^2/m$, where $\bar{a} = \sum a_i/m$
binomial	PDF	$\binom{n}{y}\pi^y(1-\pi)^{n-y}, \quad y = 0, 1, \ldots, n$
$n = 1, 2, \ldots; \quad \pi \in (0, 1)$	mean	$n\pi$
R: `binom`; IMSL: `bin`	variance	$n\pi(1-\pi)$
Bernoulli	PDF	$\pi^y(1-\pi)^{1-y}, \quad y = 0, 1$
$\pi \in (0, 1)$	mean	π
(special binomial)	variance	$\pi(1-\pi)$
Poisson	PDF	$\theta^y \mathrm{e}^{-\theta}/y!, \quad y = 0, 1, 2, \ldots$
$\theta > 0$	mean	θ
R: `pois`; IMSL: `poi`	variance	θ
hypergeometric	PDF	$\dfrac{\binom{M}{y}\binom{L-M}{N-y}}{\binom{L}{N}}$, $y = \max(0, N-L+M), \ldots, \min(N, M)$
$L = 1, 2, \ldots;$	mean	NM/L
$M = 1, 2, \ldots, L;\ N = 1, 2, \ldots, L$	variance	$((NM/L)(1-M/L)(L-N))/(L-1)$
R: `hyper`; IMSL: `hyp`		
negative binomial	PDF	$\binom{y+r-1}{r-1}\pi^r(1-\pi)^y, \quad$ y=0,1,2,...
$r > 0;\ \pi \in (0, 1)$	mean	$r(1-\pi)/\pi$
R: `nbinom`; IMSL: `nbn`	variance	$r(1-\pi)/\pi^2$
geometric	PDF	$\pi(1-\pi)^y, \quad$ y=0,1,2,...
$\pi \in (0, 1)$	mean	$(1-\pi)/\pi$
(special negative binomial)	variance	$(1-\pi)/\pi^2$
logarithmic	PDF	$-\dfrac{\pi^y}{y\log(1-\pi)}, \quad$ y=1,2,3,...
$\pi \in (0, 1)$	mean	$-\pi/((1-\pi)\log(1-\pi))$
IMSL: `lgr`	variance	$-\pi(\pi + \log(1-\pi))/((1-\pi)^2(\log(1-\pi))^2)$
multinomial	PDF	$\dfrac{n!}{\prod \pi_i!}\prod \pi_i^{y_i},\ y_i = 0, 1, \ldots, n,\ \sum y_i = n$
$n = 1, 2, \ldots,\ \pi_i \in (0, 1),\ \sum \pi_i = 1$	means	$n\pi_i$
R: `multinom`; IMSL: `mtn`	variances	$n\pi_i(1-\pi_i)$
	covariances	$-n\pi_i\pi_j$

Table C.2. Continuous Distributions (PDFs are w.r.t Lebesgue measure)

uniform	PDF	$\frac{1}{\theta_2-\theta_1}\,\mathrm{I}_{(\theta_1,\theta_2)}(y)$
$\theta_1 < \theta_2 \in \mathbb{R}$	mean	$(\theta_2+\theta_1)/2$
R: `unif`; IMSL: `unf`	variance	$(\theta_2^2 - 2\theta_1\theta_2 + \theta_1^2)/12$
normal	PDF	$\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-(y-\mu)^2/2\sigma^2}$
$\mu \in \mathbb{R}$; $\sigma > 0 \in \mathbb{R}$	mean	μ
R: `norm`; IMSL: `nor`	variance	σ^2
multivariate normal	PDF	$\frac{1}{(2\pi)^{d/2}\lvert\Sigma\rvert^{1/2}}\mathrm{e}^{-(y-\mu)^{\mathrm{T}}\Sigma^{-1}(y-\mu)/2}$
$\mu \in \mathbb{R}^d$; $\Sigma \succ 0 \in \mathbb{R}^{d\times d}$	mean	μ
R: `mvrnorm`; IMSL: `mvn`	covariance	Σ
chi-squared	PDF	$\frac{1}{\Gamma(\nu/2)2^{\nu/2}}y^{\nu/2-1}\mathrm{e}^{-y/2}\,\mathrm{I}_{(0,\infty)}(y)$
$\nu > 0$	mean	ν
R: `chisq`; IMSL: `chi`	variance	2ν
t	PDF	$\frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}}(1+y^2/\nu)^{-(\nu+1)/2}$
$\nu > 0$	mean	0
R: `t`; IMSL: `stt`	variance	$\nu/(\nu-2)$, for $\nu > 2$
F	PDF	$\frac{\nu_1^{\nu_1/2}\nu_2^{\nu_2/2}\Gamma(\nu_1+\nu_2)y^{\nu_1/2-1}}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)(\nu_2+\nu_1 y)^{(\nu_1+\nu_2)/2}}\,\mathrm{I}_{(0,\infty)}(y)$
$\nu_1 > 0$; $\nu_2 > 0$	mean	$\nu_2/(\nu_2-2)$, for $\nu_2 > 2$
R: `f`; IMSL: `f`	variance	$2\nu_2^2(\nu_1+\nu_2-2)/(\nu_1(\nu_2-2)^2(\nu_2-4))$, for $\nu_2 > 4$
lognormal	PDF	$\frac{1}{\sqrt{2\pi}\sigma}y^{-1}\mathrm{e}^{-(\log(y)-\mu)^2/2\sigma^2}\,\mathrm{I}_{(0,\infty)}(y)$
$\mu \in \mathbb{R}$; $\sigma > 0 \in \mathbb{R}$	mean	$\mathrm{e}^{\mu+\sigma^2/2}$
R: `lnorm`; IMSL: `lnl`	variance	$\mathrm{e}^{2\mu+\sigma^2}(\mathrm{e}^{\sigma^2}-1)$
gamma	PDF	$\frac{1}{\Gamma(\alpha)\beta^\alpha}y^{\alpha-1}\mathrm{e}^{-y/\beta}\,\mathrm{I}_{(0,\infty)}(y)$
$\alpha > 0$, $\beta > 0 \in \mathbb{R}$	mean	$\alpha\beta$
R: `gamma`; IMSL: `gam`	variance	$\alpha\beta^2$
exponential	PDF	$\lambda\mathrm{e}^{-\lambda y}\,\mathrm{I}_{(0,\infty)}(y)$
$\lambda > 0 \in \mathbb{R}$	mean	$1/\lambda$
R: `exp`; IMSL: `exp`	variance	$1/\lambda^2$
double exponential	PDF	$\frac{1}{2}\lambda\mathrm{e}^{-\lambda\lvert y-\mu\rvert}$
$\mu \in \mathbb{R}$; $\lambda > 0 \in \mathbb{R}$	mean	μ
(folded exponential)	variance	$2/\lambda^2$

Table C.2. Continuous Distributions (continued)

Weibull	PDF	$\frac{\alpha}{\beta} y^{\alpha-1} \mathrm{e}^{-y^\alpha/\beta} \, \mathrm{I}_{(0,\infty)}(y)$
$\alpha > 0,\ \beta > 0 \in \mathbb{R}$	mean	$\beta^{1/\alpha}\Gamma(\alpha^{-1}+1)$
R: `weibull`; IMSL: `wib`	variance	$\beta^{2/\alpha}\left(\Gamma(2\alpha^{-1}+1) - (\Gamma(\alpha^{-1}+1))^2\right)$
Cauchy	PDF	$\frac{1}{\pi\beta\left(1+\left(\frac{y-\gamma}{\beta}\right)^2\right)}$
$\gamma \in \mathbb{R};\ \beta > 0 \in \mathbb{R}$	mean	does not exist
R: `cauchy`; IMSL: `chy`	variance	does not exist
beta	PDF	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} \, \mathrm{I}_{(0,1)}(y)$
$\alpha > 0,\ \beta > 0 \in \mathbb{R}$	mean	$\alpha/(\alpha+\beta)$
R: `beta`; IMSL: `beta`	variance	$\alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$
logistic	PDF	$\frac{\mathrm{e}^{-(y-\mu)/\beta}}{\beta(1+\mathrm{e}^{-(y-\mu)/\beta})^2}$
$\mu \in \mathbb{R};\ \beta > 0 \in \mathbb{R}$	mean	μ
R: `logis`	variance	$\beta^2\pi^2/3$
Pareto	PDF	$\frac{\alpha\gamma^\alpha}{y^{\alpha+1}} \, \mathrm{I}_{(\gamma,\infty)}(y)$
$\alpha > 0,\ \gamma > 0 \in \mathbb{R}$	mean	$\alpha\gamma/(\alpha-1)$ for $\alpha > 1$
	variance	$\alpha\gamma^2/((\alpha-1)^2(\alpha-2))$ for $\alpha > 2$
von Mises	PDF	$\frac{1}{2\pi \mathrm{I}_0(\kappa)} \mathrm{e}^{\kappa\cos(x-\mu)} \, \mathrm{I}_{(\mu-\pi,\mu+\pi)}(y)$
$\mu \in \mathbb{R};\ \kappa > 0 \in \mathbb{R}$	mean	μ
IMSL: `vms`	variance	$1 - (\mathrm{I}_1(\kappa)/\mathrm{I}_0(\kappa))^2$

D

Basic Mathematical Ideas and Tools

Statistics is grounded in mathematics. All of mathematics is important and it is difficult to identify which particular areas of mathematics, and at what levels, must be mastered by statisticians. Of course, statistics is a large field, and statisticians working in different areas need different kinds and different levels of mathematical competence.

The attitudes and methods of mathematics pervade mathematical statistics. We study objects. These objects may be structures, such as groups and fields, or functionals, such as integrals, estimators, or tests. We want to understand the properties of these objects. We identify, describe, and name these properties in fixed statements, with labels, such as the "Neyman-Pearson Lemma", or the "Dominated Convergence Theorem". We identify limits to the properties of an object or boundary points on the characteristics of the object by means of "counterexamples".

Our understanding and appreciation of a particular object is enhanced by comparing the properties of the given object with similar objects. The properties of objects of the same class as the given object are stated in theorems (or "lemmas", or "corollaries", or "propositions" — unless you understand the difference, just call them all "theorems"; clearly, many otherwise competent mathematical statisticians have no idea what these English words mean). The hypotheses of the theorems define the various classes of objects. Objects that do not satisfy all of the hypotheses of a given theorem provide us insight into these hypotheses. These kinds of objects are called counterexamples for the conclusions of the theorem. For example, the Lebesgue integral and the Riemann integral are similar objects. How are they different? First, we should look at the big picture: in the Lebesgue integral, we begin with a partitioning of the range of the function; in the Riemann integral, we begin with a partitioning of the domain of the function. What about some specific properties? Some important properties of the Lebesgue integral are codified in the Big Four Theorems: Fatou's lemma, the monotone convergence theorem, the dominated convergence theorem, and the bounded convergence theorem. None of these hold for the Riemann integral; that is, the Riemann integral provides

counterexamples for the conclusions of these theorems. To understand these two objects, we need to be able to prove the four theorems (they're related), and to construct counterexamples to show that they do not hold for the Riemann integral. The specifics here are not as important as the understanding of the attitude of mathematics.

A reasoning system depends on both *objects* and *methods*. There are many standard methods we use in mathematical statistics. It may seem that most methods are ad hoc, but it is useful to identify common techniques and have a ready tool kit of methods with general applicability. In Section D.1.4, we describe some standard techniques that every statistician should have in a toolkit.

The purpose of this appendix is to provide some general mathematical background for the theory of statistics. Beyond the general basics covered in Section D.1, the statistician needs grounding in linear algebra to the extent covered in Section D.4, in measure theory to the extent covered in Section D.2, in stochastic calculus to the extent covered in Section D.3, and in methods of optimization to the extent covered in Section D.5.

Notation

I must first of all point our a departure from the usual notation and terminology in regard to the real numbers. I use $\mathbb{R}$ to denote the scalar real number system in which the elements of the underlying set are singleton numbers. Much of the underlying theory is based on $\mathbb{R}$, but my main interest is usually $\mathbb{R}^d$, for some fixed positive integer d. The elements of the underlying set for $\mathbb{R}^d$ are d-tuples, or vectors. I sometimes emphasize the difference by the word "scalar" or "vector". I do not, however, distinguish in the notation for these elements from the notation for the singleton elements of $\mathbb{R}$; thus, the symbol x may represent a scalar or a vector, and a "random variable" X may be a scalar random variable or a vector random variable.

This unified approach requires a generalized interpretation for certain functions and relational operators; for example, $|x|$, $|x|^p$, e^x, and $x < y$. If $x = (x_1, \ldots, x_d)$ and $y = (y_1, \ldots, y_d)$, then

$$|x| \stackrel{\text{def}}{=} (|x_1|, \ldots, |x_d|) \tag{D.1}$$

$$|x|^p \stackrel{\text{def}}{=} (|x_1|^p, \ldots, |x_d|^p) \tag{D.2}$$

$$\mathrm{e}^x \stackrel{\text{def}}{=} (\mathrm{e}^{x_1}, \ldots, \mathrm{e}^{x_d}) \tag{D.3}$$

and

$$x < y \iff x_1 < y_1, \ldots, x_d < y_d, \tag{D.4}$$

that is, these functions and relations are applied elementwise. For more complicated objects, such as matrices, the indicated operations may have different meanings.

D.1 Some Basic Mathematical Concepts

We first need to develop some basic definitions and properties of sets. Given simple operations on sets, we expand the concepts to include various functions and structures over sets. The most important set and the one for which we define many functions is the set of reals, which I denote as $\mathbb{R}$ or $\mathbb{R}^d$.

In the following, and in all of my writing, I try to be very consistent in use of notation. Occasionally, I will mention alternative notation that some people use.

D.1.1 Sets and Spaces

We use the term *set* without a formal definition to denote a collection of things, called *elements* or *points*. If every element of a set A_2 is also in a set A_1, we say A_2 is a *subset* of A_1, and write $A_2 \subset A_1$. If A_2 is subset of A_1, but A_1 is not a subset of A_2, we say A_2 is a *proper subset* of A_1. (Some people use the notation $A_2 \subset A_1$ to mean A_2 is a proper subset of A_1, and use the notation $A_2 \subseteq A_1$ to mean that A_2 is a subset of A_1. I use $A_2 \subset A_1$ for either case.)

Given sets A_1 and A_2, their *union*, written $A_1 \cup A_2$, is the set consisting of all elements that are in A_1 or A_2; and their *intersection*, written $A_1 \cap A_2$, is the set consisting of all elements that are in both A_1 and A_2.

In working with sets, it is useful to define an *empty set*. This is the set that contains no elements. We often denote it as $\emptyset$.

The *cardinality* of a set is an indicator of how many elements the set contains. If the number of elements in a set is a finite integer, that number is the cardinality of the set. If the elements of a set can be put into a one-to-one correspondence with a sequence of positive integers, the set is said to be *countable*. If it is countable but its cardinality is not a finite integer, then the set is said to be *countably infinite*. Any interval of $\mathbb{R}$ is *uncountably infinite*. Its cardinality is said to be the cardinality of the continuum.

In any particular application, we can conceive of a set of "everything", or a "universe of discourse". In general, we call this the *universal set*. (Later we will call it the sample space.) If A is the universal set, then when we speak of the set A_1, we imply $A_1 \subset A$. This also leads naturally to the concept of the complement of a set. The *complement* of A_1, written A_1^c, is the set of all elements in A that are not in A_1, which we can also write as $A - A_1$. More generally, given the sets A_1 and A_2, we write $A_1 - A_2$ (some people write $A_1 \backslash A_2$ instead) to represent *difference* of A_1 and A_2; that is, the complement of A_2 in A_1: $A_1 - A_2 = A_1 \cap A_2^c$. If $A_2 \subset A_1$, the difference $A_1 - A_2$ is called the *proper difference*.

The *symmetric difference* of A_1 and A_2, written $A_1 \Delta A_2$, is $A_1 - A_2 \cup A_2 - A_1$:

$$A_1 \Delta A_2 \stackrel{\text{def}}{=} A_1 - A_2 \cup A_2 - A_1. \tag{D.5}$$

Two useful relationships, known as De Morgan's laws, are

$$(A_1 \cup A_2)^c = A_1^c \cap A_2^c \tag{D.6}$$

and

$$(A_1 \cap A_2)^c = A_1^c \cup A_2^c. \tag{D.7}$$

Product sets

The cartesian product (or direct product or cross product) of two sets A and B, written $A \times B$, is the set of all doubletons, (a_i, b_j), where $a_i \in A$ and $b_j \in B$. The set $A \times B$ is called a product set.

Obviously, $A \times B \neq B \times A$ in general, and $\emptyset \times A = A \times \emptyset = \emptyset$.

The concept of product sets can be extended to more than two sets in a natural way.

One statement of the Axiom of Choice is that the cartesian product of any non-empty collection of non-empty sets is non-empty.

Functions

A *function* is a set of doubletons, or pairs of elements, such that no two different pairs have the same first element.

We use "function" and "mapping" synonymously, although the latter term is sometimes interpreted more generally. To say that f is a mapping from Ω to Λ, written

$$f : \Omega \mapsto \Lambda,$$

means that for every $\omega \in \Omega$ there is a pair in f whose first member is ω. We use the notation $f(\omega)$ to represent the second member of the pair in f whose first member is ω, and we call ω the argument of the function. We call Ω the domain of the function and we call $\{\lambda | \lambda = f(\omega) \text{ for some } \omega \in \Omega\}$ the range of the function.

Variations include functions that are *onto*, meaning that for every $\lambda \in \Lambda$ there is a pair in f whose second member is λ; and functions that are *one-to-one*, written $1:1$, meaning that no two pairs have the same second member. A function that is one-to-one and onto is called a *bijection*. A function f that is one-to-one has an inverse, written f^{-1}, that is a function from Λ to Ω, such that if $f(\omega_0) = \lambda_0$, then $f^{-1}(\lambda_0) = \omega_0$.

If $(a, b) \in f$, we may write $a = f^{-1}(b)$, although sometimes this notation is restricted to the cases in which f is one-to-one. (There are some subtleties here; if f is not one-to-one, if the members of the pairs in f are reversed, the resulting set is not a function. We say f^{-1} does not exist; yet we may write $a = f^{-1}(b)$, with the meaning above.)

If $A \subset \Omega$, the *image* of A, denoted by $f[A]$, or just by $f(A)$, is the set of all $\lambda \in \Lambda$ for which $\lambda = f(\omega)$ for some $\omega \in \Omega$. (The notation $f[A]$ is preferable,

but we will often just use $f(A)$.) Similarly, if $\mathcal{C}$ is a collection of sets (see below), the notation $f[\mathcal{C}]$ denotes the collection of sets $\{f[C] : C \in \mathcal{C}\}$.

For the function f that maps from Ω to Λ, Ω is called the *domain* of the function and and $f[\Omega]$ is called the *range* of the function.

For a subset B of Λ, the *inverse image* or the *preimage* of B, denoted by $f^{-1}[B]$, or just by $f^{-1}(B)$, is the set of all $\omega \in \Omega$ such that $f(\omega) \in B$. The notation f^{-1} used in this sense must not be confused with the inverse function f^{-1} (if the latter exists). We use this notation for the inverse image whether or not the inverse of the function exists. Notice that the inverse image of a set may not generate the set; that is, $f[f^{-1}[B]] \subset B$. We also write $f[f^{-1}[B]]$ as $f \circ f^{-1}[B]$. The set $f[f^{-1}[B]]$ may be a proper subset of B; that is, there may be an element λ in B for which there is no $\omega \in \Omega$ such that $f(\omega) = \lambda$. If f is bijective, then $f[f^{-1}[B]] = B$.

Collections of Sets

Collections of sets are usually called "collections", rather than "sets". We usually denote collections of sets with upper-case calligraphic letters, e.g., $\mathcal{B}$, $\mathcal{F}$, etc.

The usual set operators and set relations are used with collections of sets, and generally have the same meaning. Thus if $\mathcal{F}_1$ is a collection of sets that conatains the set A, we write $A \in \mathcal{F}_1$, and if $\mathcal{F}_2$ is also a collection of sets, we denote the collection of all sets that are in either $\mathcal{F}_1$, or $\mathcal{F}_2$ as $\mathcal{F}_1 \cup \mathcal{F}_2$.

The collection of all subsets of a given set is called the *power set* of the given set. An axiom of naive set theory postulates the existence of the power set for any given set. We denote the power set for a set S as 2^S.

Partitions; Disjoint Sets

A *partition* of a set S is a collection of disjoint subsets of S whose union is S. Partitions of sets play an important role.

A collection of sets $\mathcal{A}$ is said to *cover* a set S if $S \subset \cup_{A_i \in \mathcal{A}} A_i$.

Given a finite collection $\mathcal{A} = \{A_1, \ldots, A_n\}$ that covers a set S, a partition of S can be formed by removing some of the intersections of sets in $\mathcal{A}$. For example, if $S \subset A_1 \cup A_2$, then $\{A_1 \cap S,\ (A_2 \cap S) - (A_1 \cap A_2)\}$ is a partition of S.

Another simple example is a partition of the union $A_1 \cup A_2$. A simple partition of the union is just $\{A_1 - (A_1 \cap A_2),\ A_2\}$, but a more useful partition, because it is easily generalizable, uses three sets:

$$\{A_1 - (A_1 \cap A_2),\ A_2 - (A_1 \cap A_2),\ (A_1 \cap A_2)\}. \tag{D.8}$$

A_1 A_2

Given any finite union of sets $\cup_{i=1}^{n} A_i$, we can obtain similar partitions. This leads to the inclusion-exclusion formula for measures of sets that has common applications in deriving properties of measures.

It is often of interest to determine the smallest partition (that is, the partition with the smallest number of sets) of the universal set formed by sets in a given collection. For example, consider the collection $\mathcal{A} = \{A_1, A_2\}$. If neither A_1 nor A_2 is a subset of the other, then the partition

$$\{A_1 \cap A_2,\ A_1 - A_2,\ A_2 - A_1,\ (A_1 \cup A_2)^c\}$$

consists of the "smallest" collection of subsets that can be identified with operations on A_1 and A_2.

If $A_1 \subset A_2$, then $A_1 - A_2 = \emptyset$ and so the smallest partition is

$$\{A_1,\ A_2 - A_1,\ A_2^c\}.$$

Ordered sets

A set A is said to be *partially ordered* if there exists a relation $\leq$ on $A \times A$ such that $\forall a, b, c \in A$:

- $a \leq a$ (it is reflexive)
- $a \leq b, b \leq c \Rightarrow a \leq c$ (it is transitive)
- $a \leq b, b \leq a \Rightarrow a = b$ (it is antisymmetric)

A set A is called *ordered* if it is partially ordered and every pair of elements $a, b \in A$ can be compared with each other by the partial ordering relation. The real numbers are ordered.

A set A is called *well-ordered* if it is an ordered set for which every non-empty subset contains a smallest element. The positive integers are well-ordered (obviously). By the Axiom of Choice, every set (e.g., the reals) can be well-ordered.

Spaces

In any application it is generally useful to define some "universe of discourse" that is the set of all elements that will be considered in a given problem. Given a universe or universal set, which we often denote by the special symbol Ω (note the font), we then define various mathematical structures on Ω. These structures, which we often call "spaces", are formed by specifying certain types of collections of subsets of Ω and/or by defining operations on the elements of Ω or on the subsets in the special collection of subsets. In probability and statistics, we will call the universal set the *sample space*.

Some of the general structures that we will find useful are *topological spaces*, which are defined in terms of the type of collection of subsets of the universal set, and *metric spaces* and *linear spaces*, which are defined in terms

of operations on elements of the universal set. We will discuss these below, and then in Section D.1.3, we will discuss some properties of the special spaces in which the universal set is the set of real numbers. In Section D.2, we will discuss various types of collections of subsets of the universal set, and then for a particular type of collection, called a σ-field, we will discuss a special type of space, called a *measurable space*, and then, with the addition of a real-valued set function, we will define a *measure space*. A particular type of measure space is a *probability space*.

Topologies

One of the simplest structures based on the nonempty universal set Ω is a *topological space* or a *topology*, which is formed by a collection $\mathcal{T}$ of subsets of Ω with the following properties:

(t_1) $\emptyset, \Omega \in \mathcal{T}$, and
(t_2) $A, B \in \mathcal{T} \Rightarrow A \cap B \in \mathcal{T}$, and
(t_3) $\mathcal{A} \subset \mathcal{T} \Rightarrow \cup\{A : A \in \mathcal{A}\} \in \mathcal{T}$.

Members of a a topological space are called *open sets*. (This definition of open sets is more abstract than one we will give below after defining metrics and neighborhoods. That other definition is the one we will use for sets of real numbers. The corresponding topology is then defined as the collection of all open sets according to the definition of openness in that context.)

Properties of Ω that can be expressed in terms of a topology are called its *topological properties*.

Without imposing any additional structure on a topological space, we can define several useful concepts.

Let $(\Omega, \mathcal{T})$ be a topological space. A set $A \subset \Omega$ is said to be *closed* iff $\Omega \cap A^c \in \mathcal{T}$. For the set $A \subset \Omega$, the *closure* of A is the set $\overline{A} = \cap\{B : B \text{ is closed, and } A \subset B \subset \Omega\}$. (Notice that every $y \in A$ is a point of closure of A, and that A is closed iff $A = \overline{A}$.) For the set $A \subset \Omega$, the *interior* of A is the set $A^\circ = \cup\{U : U \text{ is open, and } U \subset A\}$. The *boundary* of the set $A \subset \Omega$ is the set $\partial A = \overline{A} \cap \overline{A^c}$. A *neighborhood of a point* $\omega \in \Omega$ is any set $U \in \mathcal{T}$ such that $x \in U$. Notice that Ω is a neighborhood of each point. The space $(\Omega, \mathcal{T})$ is called a *Hausdorff space* iff each pair of distinct points of Ω have disjoint neighborhoods. For $x \in U$ and $A \subset \Omega$, we say that x is a *limit point* of A iff for each neighborhood U of x, $(U \cap \{x\}^c \cap A \neq \emptyset$. The topological space $(\Omega, \mathcal{T})$ is said to be *connected* iff there do not exist two disjoint open sets A and B such that $A \cup B = \Omega$. We can also speak of a subset of Ω as being connected, using this same condition.

Metrics

A useful structure can be formed by introduction a function that maps the product set $\Omega \times \Omega$ into the nonnegative reals. Given a space Ω, a *metric* over Ω is a function ρ such that for $x, y, z \in \Omega$

- $\rho(x, y) = 0$ if and only if $x = y$
- $\rho(x, y) = \rho(y, x)$
- $\rho(x, y) \leq \rho(x, z) + \rho(z, x)$

The structure (Ω, ρ) is called a *metric space*.

The concept of a metric allows us to redefine the topological properties introduced above in terms of the metric, which we do in the following sections. The definitions in terms of a metric are generally more useful, and also a metric allows us to define additional important properties, such as continuity.

A common example of a metric space is the set $\mathbb{R}$ together with $\rho(x, y) = |x - y|$.

Neighborhoods

The concept of a metric allows us to define a *neighborhood of a point* in a set. For a point $x \in \Omega$, a metric ρ on Ω, and any positive number ϵ, an ϵ-*neighborhood* of x, denoted by $\mathcal{N}_\rho(x, \epsilon)$, is the set of $y \in \Omega$ whose distance from x is less than ϵ; that is,

$$\mathcal{N}_\rho(x, \epsilon) \stackrel{\text{def}}{=} \{y \, : \, \rho(x, y) < \epsilon\}. \tag{D.9}$$

Notice that the meaning of a neighborhood depends on the metric, but in any case it is an open set. Usually, we assume that a metric is given and just denote the neighborhood as $\mathcal{N}(x, \epsilon)$.

The concept of a neighborhood allows us to give a more meaningful definition of open sets and to define such things as continuity.

Open Sets

The specification of a topology defines the open sets of the structure and consequently neighborhoods of point. It is often a more useful approach to define first a metric, then to define neighborhoods as above, and then to define open sets in terms of neighborhoods. In this approach, a subset G of Ω is said to be *open* if each member of G has a neighborhood that is contained in G.

Note that with each metric space (Ω, ρ), we can associate a topological space $(\Omega, \mathcal{T})$, where $\mathcal{T}$ is the collection of open sets in (Ω, ρ).

We note that $(\mathbb{R}, \rho)$ is a Hausdorff space because, given $x, y \in \mathbb{R}$ and $x \neq y$ we have $\rho(x, y) > 0$ and so $\mathcal{N}(x, \rho(x, y)/2)$ and $\mathcal{N}(y, \rho(x, y)/2)$ are disjoint open sets.

We also note that $\mathbb{R}$ is connected, as is any interval in $\mathbb{R}$.

We will defer further discussion of openness and related concepts to page 356 in Section D.1.3 where we discuss the real number system.

Continuous Functions

A function f from the metric space Ω with metric ρ to the metric space Λ with metric τ is said to be *continuous* at the point $\omega_0 \in \Omega$ if for any $\epsilon > 0$ there is a $\delta > 0$ such that f maps $\mathcal{N}_\rho(\omega_0, \epsilon)$ into $\mathcal{N}_\tau(f(\omega_0), \delta)$. Usually, we assume that the metrics are given and, although they may be different, we denote the neighborhood without explicit reference to the metrics. Thus, we write $f[\mathcal{N}(\omega_0, \epsilon)] \subset \mathcal{N}(f(\omega_0), \delta)$.

We will discuss various types of continuity of real-valued functions over real domains in Section D.2.4 beginning on page 385.

Sequences of Sets

De Morgan's laws (D.6) and (D.7) express important relationships between unions, intersections, and complements.

Two important types of unions and intersections of sequences of sets are called the *lim sup* and the *lim inf* and are defined as

$$\limsup_n A_n \stackrel{\text{def}}{=} \cap_{n=1}^{\infty} \cup_{i=n}^{\infty} A_i \tag{D.10}$$

and

$$\liminf_n A_n \stackrel{\text{def}}{=} \cup_{n=1}^{\infty} \cap_{i=n}^{\infty} A_i. \tag{D.11}$$

We often use the alternative notation A^* or $\overline{\lim}_n$, and A_* or $\underline{\lim}_n$:

$$A^* \stackrel{\text{def}}{=} \overline{\lim}_n \stackrel{\text{def}}{=} \limsup_n A_n \tag{D.12}$$

and

$$A_* \stackrel{\text{def}}{=} \underline{\lim}_n \stackrel{\text{def}}{=} \liminf_n A_n. \tag{D.13}$$

We can interpret A^* and A_* in an intuitive fashion:
An element ω is in A^* iff for each n, there is some $i \geq n$ for which $\omega \in A_i$. This means that ω must lie in infinitely many of the A_n.
An element ω is in A_* iff there is some n such that for all $i \geq n$, $\omega \in A_i$. This means that ω must lie in all but finitely many of the A_n.

Convergence of Sequences of Sets

We define convergence of a sequence of sets in terms of lim sup and lim inf.

The sequence of sets $\{A_n\}$ is said to *converge* if

$$\limsup_n A_n = \liminf_n A_n, \tag{D.14}$$

and this set is said to be the *limit* of the sequence.

A sequence of sets $\{A_n\}$ is said to be *increasing* if

$$A_n \subset A_{n+1} \,\forall\, n, \tag{D.15}$$

and is said to be *decreasing* if

$$A_{n+1} \subset A_n \,\forall\, n. \tag{D.16}$$

In either case, the sequence is said to be *monotone*.

An increasing sequence $\{A_n\}$ converges to $\cup_{n=1}^{\infty} A_n$.

A decreasing sequence $\{A_n\}$ converges to $\cap_{n=1}^{\infty} A_n$.

Some Basic Facts about lim sup and lim inf

Two simple relationships that follow immediately from the definitions:

$$\limsup_n A_n \subset \cup_{i=n}^{\infty} A_i \tag{D.17}$$

and

$$\cap_{i=n}^{\infty} A_i \subset \liminf_n A_n. \tag{D.18}$$

Another important fact is

$$\liminf_n A_n \subset \limsup_n A_n. \tag{D.19}$$

To see this, consider any $\omega \in \liminf_n A_n$:

$$\omega \in \cup_{n=1}^{\infty} \cap_{i=n}^{\infty} A_i \quad \Longleftrightarrow \quad \exists n \text{ such that } \forall i \geq n,\ \omega \in A_i,$$

so $\omega \in \limsup_n A_n$.

A similar relation for any $\omega \in \limsup_n A_n$ is

$$\omega \in \cap_{n=1}^{\infty} \cup_{i=n}^{\infty} A_i \quad \Longleftrightarrow \quad \forall n\, \exists i \geq n \text{ such that } \omega \in A_i.$$

Examples

1. Consider the alternating-constant series: $A_{2n} = B$ and $A_{2n+1} = C$. Then $\liminf_n A_n = B \cap C$ and $\limsup_n A_n = B \cup C$.
2. Let the sample space be $\mathbb{R}$, and let $A_{2n} = (-n, n)$ and $A_{2n+1} = (0, 1/n)$. Then $\liminf_n A_n = \emptyset$ and $\limsup_n A_n = \mathbb{R}$.
3. Consider
$$A_n = \begin{cases} \left(\frac{1}{n},\ \frac{3}{4} - \frac{1}{n}\right) & \text{for } n = 1, 3, 5, \ldots \\ \left(\frac{1}{4} - \frac{1}{n},\ 1 + \frac{1}{n}\right) & \text{for } n = 2, 4, 6, \ldots \end{cases}$$
We have $\liminf_n A_n = \left[\frac{1}{4}, \frac{3}{4}\right)$ and $\limsup_n A_n = (0, 1]$.

D.1.2 Linear Spaces

An interesting class of spaces are those that have a closed addition operation for all elements and an additive identity, and for which we define a multiplication of real numbers and elements of the space. We denote the addition operation by "+", the additive identity by "0", and the multiplication of a real number and an element of the space by juxtaposition. A structure $\mathcal{S}, +$ is called a *linear space* if for any $x, y \in \mathcal{S}$ and any $a \in \mathbb{R}$, $ax + y \in \mathcal{S}$. The "axpy operation", $ax + y$, is the fundamental operation in linear spaces.

Linear Combinations, Linear Independence, and Basis Sets

Given $x_1, x_2, \ldots \in \mathcal{S}$ and $c_1, c_2, \ldots \in \mathbb{R}$, $\sum_i c_i x_i$ is called a *linear combination*.

A set of elements $x_1, x_2, \ldots \in \mathcal{S}$ are said to be *linearly independent* if $\sum_i c_i x_i = 0$ for $c_1, c_2, \ldots \in \mathbb{R}$ implies that $c_1 = c_2 = \cdots = 0$.

Given a linear space $\mathcal{S}$ and a set $B = \{b_i\}$ of linearly independent elements of $\mathcal{S}$ if for any element $x \in \mathcal{S}$, there exist $c_1, c_2, \ldots \in \mathbb{R}$ such that $x = \sum_i c_i b_i$, then B is called a *basis set* of $\mathcal{S}$.

Inner Products

If $\mathcal{S}$ is a linear space, an *inner product* is a mapping from $\mathcal{S} \times \mathcal{S}$ to $\mathbb{R}$. We denote an inner product by $\langle x, y \rangle$. It has the following properties for all x, y, and z in $\mathcal{S}$

1. Nonnegativity and mapping of the identity:
   if $x \neq 0$, then $\langle x, x \rangle > 0$ and $\langle 0, x \rangle = \langle x, 0 \rangle = \langle 0, 0 \rangle = 0$.
2. Commutativity:
   $\langle x, y \rangle = \langle y, x \rangle$.
3. Factoring of scalar multiplication in inner products:
   $\langle ax, y \rangle = a\langle x, y \rangle$ for real a.
4. Relation of vector addition to addition of inner products:
   $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$.

Inner products are often called dot products, although "dot product" is often used to mean a specific inner product.

A useful property of inner products is the *Cauchy-Schwarz inequality*:

$$\langle x, y \rangle \leq \langle x, x \rangle^{\frac{1}{2}} \langle y, y \rangle^{\frac{1}{2}}.$$

The proof of this is a classic: form the nonnegative polynomial in t:

$$0 \leq \langle tx + y, tx + y \rangle = \langle x, x \rangle t^2 + 2\langle x, y \rangle t + \langle y, y \rangle,$$

and then, because there can be at most one real root in t, require that the discriminant

$$(2\langle x, y \rangle)^2 - 4\langle x, x \rangle \langle y, y \rangle$$

be nonpositive.

Norms

A *norm* is a function, $\|\cdot\|$, from $\mathcal{S}$ to $\mathbb{R}$ that satisfies the following three conditions for all x and y in $\mathcal{S}$.

1. Nonnegativity and mapping of the identity:
   if $x \neq 0$, then $\|x\| > 0$, and $\|0\| = 0$
2. Relation of scalar multiplication to real multiplication:
   $\|ax\| = |a|\,\|x\|$ for real a
3. Triangle inequality:
   $\|x + y\| \leq \|x\| + \|y\|$

If, in the first condition, the requirement $\|x\| > 0$ is replaced by $\|x\| \geq 0$, the resulting function is called a *pseudonorm*. In some contexts, we may need to qualify the implication implicit in the first condition by "almost everywhere"; that is, $\|x\| = 0 \Rightarrow x = 0$ almost everywhere. Of course the other alternative is just to use a pseudonorm in such a context, but it does not carry the same properties as a norm even with the weakened condition of almost everywhere.

A linear space together with a norm is called a *normed linear space*.

Norms and Metrics Induced by an Inner Product

If $\langle\cdot,\cdot\rangle$ is an inner product on $\mathcal{S}$, let $\|x\| = \sqrt{\langle x, x\rangle}$, for all x in $\mathcal{S}$. We can show that $\|\cdot\|$ satisfies the definition of a norm. This is called the norm induced by that inner product.

For x and y in the normed linear space $(\mathcal{S}, \|\cdot\|)$ the function $\rho(x, y) = \|x - y\|$ is a metric, as we can easily see from the definition of metric on page 347. This metric is said to be induced by the norm $\|\cdot\|$.

Countable Sequences and Complete Spaces

Countable sequences of elements of a linear space, $\{x_i\}$, for $i = 1, 2, \ldots$, are often of interest. The limit of the sequence, that is, $\lim_{i\to\infty} x_i$, is of interest. The first question, of course, is whether it exists. An oscillating sequence such as $-1, +1, -1, +1, \ldots$ does not have a limit. Likewise, a divergent sequence such as $1, 2, 3, \ldots$ does not have a limit. If the sequence has a finite limit, we say the sequence *converges*, but the next question is whether it converges to a point in the given linear space.

Let $A = \{x_i \,|\, i = 1, 2, \ldots; x_i \in \mathbb{R}^d\}$. If for every $\epsilon > 0$, there exists a constant n_ϵ such that

$$\|x_n - x_m\| < \epsilon \quad \forall m, n > n_\epsilon,$$

then A is called a *Cauchy sequence*. A sequence of elements of a linear space converges only if it is a Cauchy sequence.

A normed linear space is said to be *complete* if every Cauchy sequence in the space converges to a point in the space. Such a space is called a *Banach space*.

A Banach space whose metric arises from an inner product is called a *Hilbert space*.

D.1.3 The Real Number System

The most important sets we will work with are sets of real numbers or product sets of real numbers. We assume the operations of the field of the real numbers, that is, ordinary addition and multiplication, along with the usual derived operations.

We denote the full set of real numbers, that is, the "reals", by $\mathbb{R}$ and the set of positive real numbers by $\mathbb{R}_+$. For a positive integer d, we denote the product set $\prod_{i=1}^{d} \mathbb{R}$ as $\mathbb{R}^d$.

The simplest metric on $\mathbb{R}$ is the absolute value of the difference of two numbers; that is, for $x, y \in \mathbb{R}$,

$$\rho(x, y) = |x - y|.$$

This allows us to define neighborhoods and open sets.

Metrics on $\mathbb{R}^n$ are usually defined in terms of norms of differences that generalize the simple metric on $\mathbb{R}$. A simple extension of the absolute value metric is the Euclidean distance:

$$\|x - y\|_2 = \left(\sum_{i=1}^{n} (x_i - y_i)^2 \right)^{1/2}. \tag{D.20}$$

We often write the Euclidean distance between x and y as $\|x - y\|$.

The reals do not include the two special elements ∞ and $-\infty$, although we sometimes speak of the "extended reals", which we denote and define by

$$\mathbb{R}^* \stackrel{\text{def}}{=} \mathbb{R} \cup \{-\infty, \infty\}. \tag{D.21}$$

These two elements have a number of special properties such as $\forall x \in \mathbb{R}, -\infty < x < \infty$ and $x \pm \infty = \pm\infty$.

The finite reals, that is, the reals without ∞ and $-\infty$, are generally more useful, and by not including the infinities in the reals, we make the discussions simpler.

We denote the full set of integers by $\mathbb{Z}$, and the set of positive integers by $\mathbb{Z}_+$. The positive integers are also called the natural numbers. Integers are reals and so $\mathbb{Z} \subset \mathbb{R}$ and $\mathbb{Z}_+ \subset \mathbb{R}_+$.

Sets, Sequences, and Limits of Reals

A useful limit of sequences of reals that we will encounter from time to time is

$$\lim_{n\to\infty}\left(1+\frac{c}{n}\right)^n = \mathrm{e}^c. \tag{D.22}$$

We can prove this easily using some simple properties of the logarithm function, which we define as $L(t) = \int_1^t (1/x)\mathrm{d}x$ for $t > 0$. We first observe that L is continuous and increasing, $L(1) = 0$, that L' exists at 1, $L'(1) = 1$, and $nL(x) = L(x^n)$. For a fixed constant $c \neq 0$ we can write the derivative at 1 as

$$\lim_{n\to\infty}\frac{L(1+c/n)-L(1)}{c/n} = 1,$$

which, because $L(1) = 0$, we can rewrite as $\lim_{n\to\infty} L((1+c/n)^n) = c$. Since L is continuous and increasing $\lim_{n\to\infty}(1+c/n)^n$ exists and is the value of x such that $L(x) = c$; that is, it is e^c.

A related limit for a function $g(n)$ that has the limit $\lim_{n\to\infty} = b$ is

$$\lim_{n\to\infty}\left(1+\frac{cg(n)}{n}\right)^n = \mathrm{e}^{bc}, \tag{D.23}$$

which can be shown easily by use of the limit above, and the bounds

$$\left(1+\frac{c(b-\epsilon)}{n}\right)^n \leq \left(1+\frac{cg(n)}{n}\right)^n \leq \left(1+\frac{c(b+\epsilon)}{n}\right)^n,$$

for $c > 0$ and any $\epsilon > 0$, which arise from the bounds $b - \epsilon < g(n) < b + \epsilon$ for n sufficiently large. Taking limits, we get

$$\mathrm{e}^{c(b-\epsilon)} \leq \lim_{n\to\infty}\left(1+\frac{cg(n)}{n}\right)^n \leq \mathrm{e}^{c(b+\epsilon)},$$

and since ϵ was arbitrary, we have the desired conclusion under the assumption that $c > 0$. We get the same result (with bounds reversed) for $c < 0$.

Another related limit is for a function $g(n)$ that has the limit $\lim_{n\to\infty} = 0$, and constants b and c with $c \neq 0$ is

$$\lim_{n\to\infty}\left(1+\frac{c}{n}+\frac{g(n)}{n}\right)^b n = \mathrm{e}^{bc}. \tag{D.24}$$

An important property of the reals is that a sequence of reals converges if and only if it is a Cauchy sequence. (The "if" part means that the reals are complete.)

Big O and Little o Notation

We are often interested in the convergence of a sequence to another sequence. For sequences of real numbers, we have the standard definitions:

Big O, written $\mathrm{O}(a_n)$.
$b_n = \mathrm{O}(a_n)$ means $b_n/a_n \to c$ as $n \to \infty$, where c is a nonzero finite constant.
In particular, $b_n = \mathrm{O}(1)$ means b_n is bounded.
Little o, written $\mathrm{o}(a_n)$.
$b_n = \mathrm{o}(a_n)$ means $b_n/a_n \to 0$ as $n \to \infty$.
In particular, $b_n = \mathrm{o}(1)$ means $b_n \to 0$.

A slightly different definition requires that the ratios never exceed a given c. They are equivalent for sequences all of whose elements are finite.

Sums of Sequences of Reals

Sums of countable sequences of real numbers $\{x_i\}$, for $i = 1, 2, \ldots$, are often of interest. A sum of a countable sequence of real numbers is called a (real) *series*. The usual question is what is $\lim_{n\to\infty} \sum_{i=1}^{n} x_i$. If this limit is finite, the series is said to converge.

A useful way to investigate sums of sequences of reals is by use of partial sums. When we are interested in $\sum x_i$, we form the partial sum,

$$S_k = \sum_{i=1}^{k} x_i,$$

where k is some integer. Clearly, assuming the x_is are finite, S_k is finite. The use of partial sums can be illustrated by considering the geometric series, which is the sum of the geometric progression, $a, ar, ar^2, \ldots$. Let

$$S_k = \sum_{i=0}^{k} ar^i.$$

Multiplying both sides by r and subtracting the resulting equation, we have

$$(1 - r)S_k = a(1 - r^{k+1}),$$

which yields for the partial sum

$$S_k = a\frac{1 - r^{k+1}}{1 - r}.$$

This formula is useful for finite sums, but its main use is for the series. If $|r| < 1$, then

$$\sum_{i=0}^{\infty} ar^i = \lim_{k\to\infty} S_k = \frac{a}{1-r}.$$

If $|r| > 1$, then the series diverges.

Another important fact about series, called Kronecker's lemma, is useful in proofs of theorems about sums of independent random variables, such as the strong law of large numbers:

Theorem D.1.1 *Let $\{x_i \,|\, i = 1, 2, \ldots\}$ and $\{a_i \,|\, i = 1, 2, \ldots\}$ be sequences of real numbers such that $\sum_{i=1}^{\infty} x_i$ exists (and is finite), and $0 < a_1 \leq a_2 \leq \ldots$ and $a_n \to \infty$. Then*

$$\lim_{n\to\infty} \frac{1}{a_n} \sum_{i=1}^{n} a_i x_i = 0.$$

Proof. Form the partial sums in x_i, S_k and S_n, with $k < n$. We have

$$\frac{1}{a_n} \sum_{i=1}^{n} a_i x_i = S_n - \frac{1}{a_n} \sum_{i=1}^{n-1} (a_{i+1} - a_i) S_k.$$

Let $s = \sum_{i=1}^{\infty} x_i$, and for any $\epsilon > 0$, let N be such that for $n > N$, $|S_n - s| < \epsilon$. We can now write the left-hand side of the equation above as

$$\begin{aligned}
&S_n - \frac{1}{a_n} \sum_{i=1}^{N-1} (a_{i+1} - a_i) S_k - \frac{1}{a_n} \sum_{i=N}^{n-1} (a_{i+1} - a_i) S_k \\
&\quad = S_n - \frac{1}{a_n} \sum_{i=1}^{N-1} (a_{i+1} - a_i) S_k - \frac{1}{a_n} \sum_{i=N}^{n-1} (a_{i+1} - a_i) s - \frac{1}{a_n} \sum_{i=N}^{n-1} (a_{i+1} - a_i)(S_k - s) \\
&\quad = S_n - \frac{1}{a_n} \sum_{i=1}^{N-1} (a_{i+1} - a_i) S_k - \frac{a_n - a_N}{a_n} s - \frac{1}{a_n} \sum_{i=N}^{n-1} (a_{i+1} - a_i)(S_k - s) \qquad .
\end{aligned}$$

Now, consider $\lim_{n\to\infty}$. The first term goes to s, which cancels with the third term. The second term goes to zero (because the sum is a fixed value). Since the sequence $\{a_i\}$ is nondecreasing, the last term is bounded by $\frac{a_n - a_N}{a_n}\epsilon$, which is less than or equal to ϵ, which was any positive number. ∎

Open, Closed, Compact

For sets of reals we can define some useful concepts based on the underlying arithmetic on the sets. (I use the term "reals" to apply to elements of $\mathbb{R}^d$.) These concepts yield important kinds of sets, such as open, closed, compact, and convex sets.

A set A of reals is called *open* if for each $x \in A$, there exists a $\delta > 0$ such that for each y with $\|x - y\| < \delta$ belongs to A.

If A is a set of reals and if for a given $x \in A$, there exists a $\delta > 0$ such that for each y with $\|x - y\| < \delta$ belongs to A, then x is called an *interior point* of A.

We denote the set of all interior points of A as A° and call it the *interior* of A. Clearly A° is open, and, in fact, it is the union of all open subsets of A.

A real number(vector) x is called a *point of closure* of a set A of real numbers(vectors) if for every $\delta > 0$ there exists a y in A such that $\|x-y\| < \delta$. (Notice that every $y \in A$ is a point of closure of A.)

We denote the set of points of closure of A by $\overline{A}$.

A set A is called *closed* if $A = \overline{A}$.

The *boundary* of the set A, denoted ∂A, is the set of points of closure of A that are not interior points of A; that is,

$$\partial A = \overline{A} - A^\circ. \tag{D.25}$$

An important type of set of reals is an *interval*. We denote an *open interval* with parentheses; for example (a, b) is the set of all real x such that $a < x < b$. We denote a *closed interval* with square brackets; for example $[a, b]$ is the set of all real x such that $a \leq x \leq b$. We also have "half-open" or "half-closed" intervals, with obvious meanings.

The *maximum* of a well-ordered set is the largest element of the set, if it exists; likewise, the *minimum* of a well-ordered set is the smallest element of the set, if it exists. The maximum and/or the minimum may not exist if the set has an infinite number of elements. This can happen in two ways: one, the set may have no bound; and another, the bound may not be in the set, in which case, we speak of the *supremum*, or the smallest upper bound, of the set, or the *infimum*, or the largest lower bound, of the set.
Let $A = \{x \,|\, x = 1/i, i \in \mathbb{Z}_+\}$. Then $\inf(A) = 0$.

A set A is said to be *compact* if each collection of open sets that covers A contains a finite subcollection of open sets that covers A.

Heine-Borel theorem: A closed and bounded set of real numbers is compact. (See Royden, 1988, for example.)

The following properties of unions and intersections of open and closed sets are easy to show from the definitions:

- The intersection of a finite collection of open sets is open.
- The union of a countable collection of open sets is open.
- The union of a finite collection of closed sets is closed.
- The intersection of a countable collection of closed sets is closed.

Intervals in $\mathbb{R}$

A very important type of set is an *interval* in $\mathbb{R}$, which is a connected subset of $\mathbb{R}$. Intervals are the basis for building important structures on $\mathbb{R}$.

The main kinds of intervals have forms such as
$(-\infty, a)$, $(-\infty, a]$, (a, b), $[a, b]$, $(a, b]$, $[a, b)$, (b, ∞), and $[b, \infty)$.

Of these, $(-\infty, a)$, (a, b), and (b, ∞) are open;
$[a, b]$ is closed and $\overline{(a, b)} = [a, b]$;

$(a, b]$ and $[a, b)$ are neither (they are "half-open");
$(-\infty, a]$ and $[b, \infty)$ are closed, although in a special way that sometimes requires special treatment.

A finite closed interval is compact (by the Heine-Borel theorem); but an open or half-open interval is not, as we see below.

The following facts for real intervals are special cases of the properties of unions and intersections of open and closed sets we listed above, which can be shown from the definitions:

- $\cap_{i=1}^{n}(a_i, b_i) = (a, b)$ (that is, some open interval)
- $\cup_{i=1}^{\infty}(a_i, b_i)$ is an open set
- $\cup_{i=1}^{n}[a_i, b_i]$ is a closed set
- $\cap_{i=1}^{\infty}[a_i, b_i] = [a, b]$ (that is, some closed interval)

Two types of interesting intervals are

$$\left(a - \frac{1}{i},\ b + \frac{1}{i}\right) \tag{D.26}$$

and

$$\left[a + \frac{1}{i},\ b - \frac{1}{i}\right]. \tag{D.27}$$

Sequences of intervals of these two forms are worth remembering because they illustrate interesting properties of intersections and unions of infinite sequences. Infinite intersections and unions behave differently with regard to collections of open and closed sets. For finite intersections and unions we know that $\cap_{i=1}^{n}(a_i, b_i)$ is an open interval, and $\cup_{i=1}^{n}[a_i, b_i]$ is a closed set.

First, observe that

$$\lim_{i\to\infty}\left(a - \frac{1}{i},\ b + \frac{1}{i}\right) = [a, b] \tag{D.28}$$

and

$$\lim_{i\to\infty}\left[a + \frac{1}{i},\ b - \frac{1}{i}\right] = [a, b]. \tag{D.29}$$

Now for finite intersections of the open intervals and finite unions of the closed intervals, that is, for finite k, we have

$$\bigcap_{i=1}^{k}\left(a - \frac{1}{i},\ b + \frac{1}{i}\right) \quad \text{is open}$$

and

$$\bigcup_{i=1}^{k}\left[a + \frac{1}{i},\ b - \frac{1}{i}\right] \quad \text{is closed.}$$

Infinite intersections and unions behave differently with regard to collections of open and closed sets. With the open and closed intervals of the special forms, for infinite intersections and unions, we have the important facts:

$$\bigcap_{i=1}^{\infty}\left(a-\frac{1}{i},\ b+\frac{1}{i}\right)=[a,b] \tag{D.30}$$

and

$$\bigcup_{i=1}^{\infty}\left[a+\frac{1}{i},\ b-\frac{1}{i}\right]=(a,b). \tag{D.31}$$

Iff $x \in A_i$ for some i, then $x \in \cup A_i$. So if $x \notin A_i$ for any i, then $x \notin \cup A_i$. (This is why the union of the closed intervals above is not a closed interval.)

Likewise, we have

$$\begin{aligned}\bigcup_{i=1}^{\infty}\left[a+\frac{1}{i},\ b\right] &= \bigcap_{i=1}^{\infty}\left(a,\ b+\frac{1}{i}\right)\\ &= (a,b].\end{aligned} \tag{D.32}$$

From this we see that

$$\lim_{n\to\infty}\bigcup_{i=1}^{n}\left[a+\frac{1}{i},\ b-\frac{1}{i}\right] \neq \bigcup \lim_{i\to\infty}\left[a+\frac{1}{i},\ b-\frac{1}{i}\right].$$

The equations for (a,b) and $(a,b]$ above show that open intervals and the half-open intervals are not compact, because no finite collection of sets in the unions cover the intervals.

Convexity

Another useful concept for real sets and for real functions of real numbers is *convexity*.

A set $A \subset \mathbb{R}^d$ is *convex* iff for $x, y \in A$, $\forall a \in [0,1]$, $ax + (1-a)y \in A$.

Convex Functions

A function $f : D \subset \mathbb{R}^d \mapsto \mathbb{R}$, where D is convex, is *convex* iff for $x, y \in D$, $\forall a \in [0,1]$,

$$f(ax + (1-a)y) \leq af(x) + (1-a)f(y).$$

A function is *strictly convex* if the inequality above is strict.

A useful theorem that characterizes convexity of twice differentiable functions is the following:

> If the function f is twice differentiable over an open convex set D, then f is convex iff the Hessian, H_f, is nonnegative definite at all points in D. Iff it is positive definite, f is strictly convex.

The composition of a convex function and a convex function is convex. We see that this is true by letting f and g be any convex functions for which $f \circ g$ is defined. Now let a be any real number in $[0,1]$. Then $f \circ g(ax + (1-a)y) \leq f(ag(x) + (1-a)g(y)) \leq af \circ g(x) + (1-a)f \circ g(y)$.

A function f is *concave* iff $-f$ is convex.

Subharmonic Functions

Convexity of a function is defined in terms of the average of the function at two points, compared to the function at the average of the two points. We can extend that basic idea to the average of the function over a sphere compared to the function at the sphere. (The average of the function over a sphere is defined in terms of the ratio of a measure of the function image to the surface of the spheres. The measures are integrals.)

A function $f : D \subset \mathbb{R}^d \mapsto \mathbb{R}$, where D is convex, is *subharmonic* over D, iff for every point $x_0 \in D$ and for every $r > 0$, the average of f over the surface of the sphere $S_r(x_0) = \{x : \|x - x_0\| = r\}$ is greater than or equal to $f(x_0)$.

In one dimension, subharmonic and convex are the same property.

A function f is *superharmonic* if $-f$ is subharmonic. A function is *harmonic* if it is both superharmonic and subharmonic.

A useful theorem that characterizes harmonicity of twice differentiable functions is the following:

> If the function f is twice differentiable over an open convex set D, then f is subharmonic iff the Laplacian, $\nabla^2 f$, (which is just the trace of H_f) is nonnegative at all points in D. The function is harmonic if the Laplacian is 0, and superharmonic if the Laplacian is nonpositive.

The relatively simple Laplacian operator considers curvature only in the orthogonal directions corresponding to the principal axes; if the function is twice differentiable everywhere, however, this is sufficient to characterize the (sub-, super-) harmonic property. These properties are of great importance in multidimensional loss functions.

Harmonicity is an important concept in potential theory. It arises in field equations in physics. The basic equation $\nabla^2 f = 0$, which implies f is harmonic, is called Laplace's equation. Another basic equation in physics is $\nabla^2 f = -c\rho$, where $c\rho$ is positive, which implies f is superharmonic. This is called Poisson's equation, and is the basic equation in a potential (electrical, gravitational, etc.) field. A superharmonic function is called a potential for this reason. These PDE's, which are of the elliptical type, govern the diffusion of energy or mass as the domain reaches equilibrium. Laplace's equation represents a steady diffusion and Poisson's equation models an unsteady diffusion, that is, diffusion with a source or sink.

A probability density function that is superharmonic is unimodal. If the function is twice differentiable, unimodality can be characterized by the Laplacian. For densities that are not twice differentiable, negative curvature along the principal axes is sometimes called orthounimodality.

Consider $f(x) = \exp\left(\sum_{j=1}^k x_j^2\right)$. This function is twice differentiable, and we have

$$\nabla^2 \exp\left(\sum_{j=1}^{k} x_j^2\right) = \sum_{i=1}^{k}(4x_i^2 - 2)\exp\left(\sum_{j=1}^{k} x_j^2\right).$$

The exponential term is positive, so the condition depends on $\sum_{i=1}^{k}(4x_i^2 - 2)$. If $\sum_{i=1}^{k} x_i^2 < 1/2$, it is superharmonic; if $\sum_{i=1}^{k} x_i^2 = 1/2$, it is harmonic; if $\sum_{i=1}^{k} x_i^2 > 1/2$, it is subharmonic.

D.1.4 Some Useful Basic Mathematical Operations

Here are some mathematical operations that should be in fast memory.

Completing the Square

Squared binomials occur frequently in statistical theory, often in a loss function or as the exponential argument in the normal density function. Often in an algebraic manipulation, we have an expression of the form $ax^2 + bx$, and we want an expression in the form $(cx + d)^2 + e$ for the same quantity. This form can be achieved by adding and subtracting $b^2/(4a)$, so as to have $(\sqrt{a}x + b/(2\sqrt{a}))^2 - b^2/(4a)$:

$$ax^2 + bx = (\sqrt{a}x + b/(2\sqrt{a}))^2 - b^2/(4a) \tag{D.33}$$

We have a similar operation for vectors and positive definite matrices. If A is a positive definite matrix (meaning that $A^{-\frac{1}{2}}$ exists) and x and b are matrices, we can complete the square of $x^{\mathrm{T}}Ax + x^{\mathrm{T}}b$ in a similar fashion: we add and subtract $b^{\mathrm{T}}A^{-1}b/4$. This gives

$$\left(A^{\frac{1}{2}}x + A^{-\frac{1}{2}}b/2\right)^{\mathrm{T}}\left(A^{\frac{1}{2}}x + A^{-\frac{1}{2}}b/2\right) - b^{\mathrm{T}}A^{-1}b/4$$

or

$$\left(x + A^{-1}b/2\right)^{\mathrm{T}} A\left(x + A^{-1}b/2\right) - b^{\mathrm{T}}A^{-1}b/4. \tag{D.34}$$

Use of Known Integrals and Series

The standard families of probability distributions provide a compendium of integrals and series with known values. The student should immediately learn the following three basic continuous distributions and the associated integrals:

- over $\mathbb{R}$; the normal integral:

$$\int_{-\infty}^{\infty} \mathrm{e}^{-(x-\mu)^2/2\sigma^2}\mathrm{d}x = \sqrt{2\pi}\sigma, \tag{D.35}$$

  for $\sigma > 0$, and its multivariate extension,

- over $\mathbb{R}^d$; Aitken's integral:

$$\int_{\mathbb{R}^d} \mathrm{e}^{-(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)/2}\,\mathrm{d}x = (2\pi)^{d/2}|\Sigma|^{1/2}, \tag{D.36}$$

  for positive definite Σ^{-1}.
- over $\mathbb{R}_+$; the gamma integral (called the complete gamma function):

$$\int_0^\infty \frac{1}{\gamma^\alpha} x^{\alpha-1}\mathrm{e}^{-x/\gamma}\mathrm{d}x = \Gamma(\alpha), \tag{D.37}$$

  for $\alpha, \gamma > 0$.
- over $(0, 1)$; the beta integral:

$$\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\mathrm{d}x = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \tag{D.38}$$

  for $\alpha, \beta > 0$.

There are four simple series that should also be immediately recognizable:

- over $0, \ldots, n$; the binomial series:

$$\sum_{x=0}^{n} \frac{\Gamma(n+1)}{\Gamma(x+1)\Gamma(n-x+1)}\pi^x(1-\pi)^{n-x} = 1, \tag{D.39}$$

  for $\pi > 0$ and $n \geq 1$.
- over $0, 1, 2, \ldots$; the geometric series:

$$\sum_{x=0}^{\infty}(1-\pi)^x = \pi^{-1} \tag{D.40}$$

  for $\pi > 0$.
- over $\max(0, N-L+M), \ldots, \min(N, M)$; the hypergeometric series:

$$\sum_{x=\max(0,N-L+M)}^{\min(N,M)} \binom{M}{x}\binom{L-M}{N-x} = \binom{L}{n}, \tag{D.41}$$

  for $1 \leq L$, $0 \leq N \leq L$, and $0 \leq M \leq L$.
- over $0, 1, 2, \ldots$; the Poisson series:

$$\sum_{x=0}^{\infty} \frac{\theta^x}{x!} = \mathrm{e}^\theta, \tag{D.42}$$

  for $\theta > 0$.

Note that for $0 \leq x \leq n$,

$$\frac{\Gamma(n+1)}{\Gamma(x+1)\Gamma(n-x+1)} = \binom{n}{x}. \tag{D.43}$$

For computing expected values or evaluating integrals or sums, the trick often is to rearrange the integral or the sum so that it is in the form of the original integrand or summand with different parameters.

As an example, consider the integral that is the q^{th} raw moment of a gamma(α, β) random variable:

$$\int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} x^q x^{\alpha-1} \mathrm{e}^{-x/\beta} \mathrm{d}x.$$

We use the known value of the integral of the density:

$$\int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \mathrm{e}^{-x/\beta} \mathrm{d}x = 1.$$

So

$$\begin{aligned}
\int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} x^q x^{\alpha-1} \mathrm{e}^{-x/\beta} \mathrm{d}x &= \int_0^\infty \frac{1}{\Gamma(\alpha)} \frac{\Gamma(q+\alpha)\beta^q}{\Gamma(q+\alpha)\beta^{q+\alpha}} x^{(q+\alpha)-1} \mathrm{e}^{-x/\beta} \mathrm{d}x \\
&= \frac{\Gamma(q+\alpha)\beta^q}{\Gamma(\alpha)} \int_0^\infty \frac{1}{\Gamma(q+\alpha)\beta^{q+\alpha}} x^{(q+\alpha)-1} \mathrm{e}^{-x/\beta} \mathrm{d}x \\
&= \frac{\Gamma(q+\alpha)\beta^q}{\Gamma(\alpha)}
\end{aligned}$$

Another example is a series of the form

$$\sum_{x=0}^\infty x^q \, \theta^x \, \frac{\mathrm{e}^{-\theta}}{x!}.$$

We recognize in this the known series that corresponds to the probability function associated with the Poisson distribution:

$$\sum_{x=0}^\infty \theta^x \, \frac{\mathrm{e}^{-\theta}}{x!} = 1,$$

and realize that evaluation of the series involves a manipulation of x^q and $x!$. For $q = 1$, we have

$$\begin{aligned}
\sum_{x=0}^\infty x \, \theta^x \, \frac{\mathrm{e}^{-\theta}}{x!} &= \theta \sum_{x=1}^\infty \theta^{(x-1)} \, \frac{\mathrm{e}^{-\theta}}{(x-1)!} \\
&= \theta.
\end{aligned}$$

For $q = 2$, we form two sums so that we can get expressions involving the basic probability function:

$$\begin{aligned}\sum_{x=0}^{\infty} x^2\,\theta^x\,\frac{\mathrm{e}^{-\theta}}{x!} &= \sum_{x=2}^{\infty} x(x-1)\,\theta^x\,\frac{\mathrm{e}^{-\theta}}{x!} + \sum_{x=1}^{\infty} x\,\theta^x\,\frac{\mathrm{e}^{-\theta}}{x!} \\ &= \theta^2 \sum_{x=2}^{\infty} \theta^{(x-2)}\,\frac{\mathrm{e}^{-\theta}}{(x-2)!} + \theta \sum_{x=1}^{\infty} \theta^{(x-1)}\,\frac{\mathrm{e}^{-\theta}}{(x-1)!} \\ &= \theta^2 + \theta.\end{aligned}$$

Expansion in a Taylor Series

One of the most useful tools in analysis is the Taylor series expansion of a function about a point a:

$$f(x) = f(a) + (x-a)f' + \frac{1}{2!}(x-a)^2 f'' + \cdots \tag{D.44}$$

For a function of m variables, this is

$$f(x_1,\ldots,x_m) = \sum_{j=0}^{\infty} \left(\frac{1}{j!} \left(\sum_{k=1}^{m} (x_k - a_k) \frac{\partial}{\partial x_k} \right)^j f(x_1,\ldots,x_m) \right)_{(x_1,\ldots,x_m)=(a_1,\ldots,a_m)}. \tag{D.45}$$

Orthogonalizing Vectors

Given a set of nonnull, linearly independent vectors, $x_1, x_2, \ldots$, it is easy to form orthonormal vectors, $\tilde{x}_1, \tilde{x}_2, \ldots$, that span the same space. This can be done with respect to any inner product and the norm defined by the inner product. The most common inner product for vectors of course is $\langle x_i, x_j \rangle = x_i^{\mathrm{T}} x_j$, and the Euclidean norm, $\|x\| = \sqrt{\langle x, x \rangle}$, which we often write without the subscript.

$$\begin{aligned}\tilde{x}_1 &= \frac{x_1}{\|x_1\|} \\ \tilde{x}_2 &= \frac{(x_2 - \langle \tilde{x}_1, x_2 \rangle \tilde{x}_1)}{\|x_2 - \langle \tilde{x}_1, x_2 \rangle \tilde{x}_1)\|} \\ \tilde{x}_3 &= \frac{(x_3 - \langle \tilde{x}_1, x_3 \rangle \tilde{x}_1 - \langle \tilde{x}_2, x_3 \rangle \tilde{x}_2)}{\|x_3 - \langle \tilde{x}_1, x_3 \rangle \tilde{x}_1 - \langle \tilde{x}_2, x_3 \rangle \tilde{x}_2\|} \\ &\text{etc.}\end{aligned}$$

These are called *Gram-Schmidt transformations*. These transformations also apply to other kinds of objects, such as functions, for which we can define an inner product. *(Note: the third expression above, and similar expressions for subsequent vectors may be numerically unstable. See Gentle (2007), pages 27–29 and 432, for a discussion of numerical issues.)*

Optimization

Many statistical methods depend on maximizing something (e.g., MLE), or minimizing something, generally a risk (e.g., UMVUE, MRE) or something subjective with intuitive appeal (e.g., squared deviations from observed values, "least squares") that may or may not have optimal statistical properties.

- When looking for an optimal solution, it is important to consider the problem carefully, and not just immediately differentiate something and set it equal to 0.
- A practical optimization problem often has constraints of some kind.

$$\begin{aligned} \min_{\alpha} \quad & f(x,\alpha) \\ \text{s.t. } & g(x,\alpha) \leq b. \end{aligned}$$

- If the functions are differentiable, and if the minimum occurs at an interior point, use of the Lagrangian is usually the way to solve the problem.
- With the dependence on x suppressed, the Lagrangian is

$$L(\alpha,\lambda) = f(\alpha) + \lambda^{\mathrm{T}}(g(\alpha) - b).$$

- Differentiating the Lagrangian and setting to 0, we have a system of equations that defines a stationary point, α_*.
- We check to insure that it is a minimum by evaluating the Hessian,

$$\nabla\nabla f(\alpha)\Big|_{\alpha=\alpha_*}.$$

  If this is positive definite, there is a local minimum at α_*.

Mathematical Proofs

A conditional statement in mathematics has the form "if A then B", or "$A \Rightarrow B$", where A and B are either simple declarative statements or conditional statements. A conditional statement is either a *definition*, an *axiom*, or a *proposition*. A proposition requires a proof. (A proposition that has a proof is sometimes called a "lemma", a "theorem", or a "corollary". While these terms have meanings, the meanings are rather vague or subjective, and many authors' usage of the different terms serves no purpose other than to annoy the reader. If a proposition has no known proof, it is sometimes called a "conjecture".)

There are various types of proofs for theorems. Some are "better" than others. (See Aigner and Ziegler, 2004 for discussions of different types of proof.) The "best" proof of a proposition is a *direct* proof, which is a sequence of statements "if A then A_1, if $A_1 \ldots, \ldots$ then B", where each statement in the sequence is an axiom or a previously proven proposition. A direct proof is

called *deductive*, because each of the steps after the first is deduced from the preceding step.

Two useful types of *indirect* proofs are *contradiction* and *induction*.

In a proof of "$A \Rightarrow B$" by contradiction, we assume "A", and suppose "not B". Then we ultimately arrive at a conclusion that contradicts an axiom or a previously proven proposition. The means that the supposition "not B" cannot be true, and hence that "B" is true.

A proof by induction may be appropriate when we can index a sequence of statements by $n \in \mathbb{Z}_+$, that is, S_n, and the statement we wish to prove is that S_n is true for all $n \geq m \in \mathbb{Z}_+$. We first show that S_m is true. (Here is where a proof by induction requires some care; this statement must be nontrivial; that is, it must be a legitimate member of the sequence of statements.) Then we show that for $n \geq m$, $S_n \Rightarrow S_{n+1}$, in which case we conclude that S_n is true for all $n \geq m \in \mathbb{Z}_+$.

Another useful type of deductive proof for "$A \Rightarrow B$" is a contrapositive proof; that is, a proof of "not $B \Rightarrow$ not A".

Standard Procedures in Proofs

If the conclusion is that two sets A and B are equal, show that $A \subset B$ and $B \subset A$. To do this (for the first one), choose any $x \in A$ and show $x \in B$. The same technique is used to show that two collections of sets, for example, two σ-fields, are equal.

To show that a sequence converges, use partial sums and an ϵ bound.

To show that a series converges, show that the sequence is a Cauchy sequence.

The standard procedures may not always work, but try them first.

Use the mathematical operations, such as series expansions, discussed above.

Notes and Additional References for Section D.1

It is important that the student fully understand the concept of a mathematical proof. Solow (2002) discusses the basic ideas, and Aigner and Ziegler (2004), whose title comes from a favorite phrase of Paul Erdös, give many well-constructed proofs of common facts. Khuri (2003) presents the important facts and techniques in advanced calculus.

Additional References

Aigner and Ziegler (2004), *Proofs from THE BOOK*, third edition, Springer-Verlag, Berlin.

Khuri, André I. (2003), *Advanced Calculus with Applications in Statistics*, second edition, John Wiley & Sons, Inc., New York.

Royden, H. L. (1988), *Real Analysis*, third edition, MacMillan, New York.
Solow, Daniel (2003), *How to Read and Do Proofs*, third edition, John Wiley & Sons, Inc., New York.

D.2 Measure, Integration, and Functional Analysis

Measure and integration and the probability theory built on those topics are major fields in mathematics. The objective of this setion is just to get enough measure theory to support the probability theory necessary for a solid foundation in statistical inference.

Notice that much of the development is for abstract objects; for each of these, however, there is a concrete instance that is relevant in probability theory.

We begin with some definitions leading up to measurable spaces, abstract measurable spaces in Section D.2.1, and real measurable spaces in Section D.2.2. In Section D.2.3, we finally define a measure and discuss some properties of measures, again first in an abstract setting, and then in Section D.2.4, for real-valued functions.

Finally in Section D.2.5, we discuss integration.

D.2.1 Basic Concepts of Measure Theory

Sample Space

A sample space is a nonempty set. It is the "universe of discourse" in a given problem. It is often denoted by Ω.

An important sample space is $\mathbb{R}$, the set of reals.

Collections of Sets

For a given set Ω, there are some important types of collections of subsets.

Definition D.2.1 (π-system) *A nonempty collection of subsets, $\mathcal{P}$, is called a π-system iff*

(π_1) $A, B \in \mathcal{P} \Rightarrow A \cap B \in \mathcal{P}$.

Definition D.2.2 (ring) *A nonempty collection of subsets, $\mathcal{R}$, is called a ring iff*

(r_1) $A, B \in \mathcal{R} \Rightarrow A \cup B \in \mathcal{R}$.
(r_2) $A, B \in \mathcal{R} \Rightarrow A - B \in \mathcal{R}$.

The term "ring" also applies to a mathematical structure consisting of a set and two operations on the set satisfying certain properties. The prototypic ring is the set of integers with ordinary addition and multiplication.

Definition D.2.3 (field) *A collection of subsets, $\mathcal{F}$ is called a field iff*

(f_1) $\Omega \in \mathcal{F}$, *and*
(f_2) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$, *and*
(f_3) $A, B \in \mathcal{F} \Rightarrow A \cup B \in \mathcal{F}$.

Notice that property (f_3) is equivalent to

(f_3') $A_1, A_2, \ldots A_n \in \mathcal{F} \Rightarrow \cup_{i=1}^n A_i \in \mathcal{F}$;

that is, $\mathcal{F}$ is closed under finite unions. The next systems we describe are closed under countable unions.

Notice that a field is nonempty by definition. It contains at least one set, Ω, and if Ω is nonempty, it contains two sets, Ω and $\emptyset$.

The term "field" also applies to a mathematical structure consisting of a set and two operations on the set satisfying certain properties. The prototypic field is the set of real numbers with ordinary addition and multiplication.

Definition D.2.4 (Dynkin system) *A collection of subsets, $\mathcal{D}$, is called a Dynkin system iff*

(D_1) $\Omega \in \mathcal{D}$, *and*
(D_2) $A, B \in \mathcal{D}$ *and* $A \subset B \Rightarrow B - A \in \mathcal{D}$, *and*
(D_3) $A_1, A_2, \ldots \in \mathcal{D}$ *with* $A_1 \subset A_2 \subset \cdots \Rightarrow \cup_i A_i \in \mathcal{D}$.

Definition D.2.5 (λ-system) *A collection of subsets, $\mathcal{L}$, is called a λ-system iff*

(λ_1) $\Omega \in \mathcal{L}$, *and*
(λ_2) $A \in \mathcal{L} \Rightarrow A^{\mathrm{c}} \in \mathcal{L}$, *and*
(λ_3) $A_1, A_2, \ldots \in \mathcal{L}$ *and* $A_i \cap A_j = \emptyset$ *for* $i \neq j \Rightarrow \cup_i A_i \in \mathcal{L}$.

We can see that the first and third properties imply that the second property is equivalent to

(λ_2') $A, B \in \mathcal{L}$ and $A \subset B \Rightarrow B - A \in \mathcal{L}$.

To see this, first assume the three poperties that characterize a λ-system $\mathcal{L}$, and $A, B \in \mathcal{L}$ and $A \subset B$. We first see that this implies $B^{\mathrm{c}} \in \mathcal{L}$ and so the disjoint union $A \cup B^{\mathrm{c}} \in \mathcal{L}$. This implies that the complement $A \cup B^{\mathrm{c}})^{\mathrm{c}} \in \mathcal{L}$. But $(A \cup B^{\mathrm{c}})^{\mathrm{c}} = B - A$; hence, we have the alternative property (λ_2'). Conversely, assume this alternative property together with the first property (λ_1). Hence, $A \in \mathcal{L} \Rightarrow \Omega - A \in \mathcal{L}$, but $\Omega - A = A^{\mathrm{c}}$; that is, $A^{\mathrm{c}} \in \mathcal{L}$.

This means that the second property that characterizes a λ-system could be replaced by the second property of a Dynkin system. In a similar manner, we can show that the third properties are equivalent; hence, although the definitions are different, a Dynkin system is a λ-system, and a λ-system is a Dynkin system.

Definition D.2.6 (σ-field) *A collection of subsets, $\mathcal{F}$, of a given sample space, Ω, is called a σ-field iff*

(σ_1) $\Omega \in \mathcal{F}$
(σ_2) $A \in \mathcal{F} \Rightarrow A^{\mathrm{c}} \in \mathcal{F}$
(σ_3) $A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \cup_i A_i \in \mathcal{F}$.

A σ-field is also called a σ-algebra or a a σ-ring.

A field with the properties (σ_1) and (σ_2), but with (σ_3) replaced with the property

(δ_3) $A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \cap_i A_i \in \mathcal{F}$

is called a δ-field, a δ-algebra or a δ-ring. It is clear, however, that δ-field and σ-field are equivalent concepts. We will use the latter term.

We note that for a σ-field $\mathcal{F}$ if $A_1, A_2, \ldots \in \mathcal{F}$ then $\limsup_n A_n \in \mathcal{F}$ and $\liminf_n A_n \in \mathcal{F}$. This is because for any n, $\cup_{i=n}^{\infty} A_i \in \mathcal{F}$ and $\cap_{i=n}^{\infty} A_i \in \mathcal{F}$.

A σ-field is the most important type of field, and we will discuss this type in more detail below.

Notice that the definition of a π-system must specify that the collection is nonempty; the definitions of the other systems ensure that the collections are nonempty without saying so explicitly. The exact definitions of these systems can be modified in various simple ways. For example, in the definitions of a field, a λ-system, and a σ-field the requirement that Ω be in the system could be replaced by the requirement that $\emptyset$ be in the system, because closure with respect to complementation guarantees the inclusion of Ω. (The requirement that Ω be in a Dynkin system, however, could not be replaced by the requirement that $\emptyset$ be in the system.) The closure property for unions in a field or a σ-field could be replaced by the requirement that the system be closed for intersections of the same kind as the unions.

σ-Field Generated by a Collection of Sets

Given any sample space Ω and any collection $\mathcal{C}$ of subsets of Ω, the "smallest" σ-field over Ω of which $\mathcal{C}$ is a subset is called the σ-field generated by $\mathcal{C}$, and is denoted by $\sigma(\mathcal{C})$. It is the *minimal* σ-field that contains $\mathcal{C}$. The σ-field generated by $\mathcal{C}$ is the intersection of all σ-fields that contain $\mathcal{C}$.

σ-fields can contain a very large number of subsets. If k is the maximum number of sets that partition Ω that can be formed by operations on the sets in $\mathcal{A}$, then the number of sets in the σ-field is 2^k. (What is the "largest" σ-field over Ω?)

Other special collections of subsets can also be generated by a given collection. For example, given a collection $\mathcal{C}$ of subsets of a sample space Ω, we can form a π-system by adding (only) enough subsets to make the collection closed with respect to intersections. This system generated by $\mathcal{C}$ is the *minimal* π-system that contains $\mathcal{C}$. This π-system, denoted by $\pi(\mathcal{C})$ is the intersection of all π-systems that contain $\mathcal{C}$. Likewise, we define the λ-system generated by $\mathcal{C}$ as the minimal λ-system that contains $\mathcal{C}$, and we denote it by $\lambda(\mathcal{C})$.

Examples:
1.
The trivial σ-field is $\{\emptyset, \Omega\}$.

2.
The second most trivial σ-field is $\{\emptyset, A, A^c, \Omega\}$, for some $A \neq \emptyset$ and $A \neq \Omega$.
3.
If $\mathcal{A} = \{A\}$, with respect to the sample space Ω, $\sigma(\mathcal{A}) = \{\emptyset, A, A^c, \Omega\}$.
4.
If $\mathcal{A} = \{A_1, A_2\}$ and neither A_1 nor A_2 is a subset of the other, with respect to the sample space Ω, there are 4 "smallest" sets that partition $\mathcal{A}$. These are called *atoms*. They are

$$\{A_1 \cap A_2,\ A_1 - A_2,\ A_2 - A_1,\ (A_1 \cup A_2)^c\}.$$

Hence, there are 16 sets in $\sigma(\mathcal{A})$. These can be written simply as the binary combinations of all above: (0000), (0001), (0010), ... Following this order, using the partition above, the sets are (after simplification):

$$\begin{aligned}\sigma(\mathcal{A}) = \{&\emptyset,\ (A_1 \cup A_2)^c,\ A_2 - A_1,\ A_1^c,\\ &A_1 - A_2,\ A_2^c,\ A_1 \Delta A_2,\ (A_1 \cap A_2)^c,\\ &A_1 \cap A_2,\ (A_1 \Delta A_2)^c,\ A_2,\ (A_1 - A_2)^c,\\ &A_1,\ (A_2 - A_1)^c,\ A_1 \cup A_2,\ \Omega\}.\end{aligned}$$

5.
If $\mathcal{A} = \{A_1, A_2\}$ and $A_1 \subset A_2$, with respect to the sample space Ω, there are 8 sets in $\sigma(\mathcal{A})$.
6.
For the sample space Ω, $\sigma(\{\Omega\}) = \{\emptyset, \Omega\}$. This is the trivial σ-field.
Why do we have the braces in $\sigma(\{\Omega\})$?
We do often abuse the notation, however; if the argument of $\sigma(\dot{})$ is a singleton, we sometimes omit the braces. For example, the second most trivial σ-field is that generated by a single set, say A: $\sigma(A) = \{\emptyset, A, A^c, \Omega\}$. We may also abuse the notation even further by writing a collection of subsets without putting them in braces. So, for example, $\sigma(A, B)$ may be used instead of $\sigma(\{A.B\})$. It is obvious that $\sigma(A) \subset \sigma(A, B)$.
7.
For the sample space Ω, the power set 2^Ω is a σ-field. It is the "largest" σ-field over Ω.

Borel σ-Fields

Given a topological space $(\Omega, \mathcal{T})$ (that is, a space defined by the collection $\mathcal{T}$ of open subsets of Ω), we call the σ-field generated by $\mathcal{T}$ the *Borel σ-field* on $(\Omega, \mathcal{T})$. We often denote this as $\mathcal{B}(\Omega, \mathcal{T})$, or as $\mathcal{B}(\Omega)$, or just as $\mathcal{B}$.

Relations of σ-Fields to Other Structures

A σ-field is a π-system, a field, and a λ-system.

A field is not necessarily a σ-field. The classic example is this: Let Ω be a countably infinite set, and let $\mathcal{F}$ consist of all finite subsets of Ω along with all subsets of Ω whose complements are finite. We see immediately that $\mathcal{F}$ is a field. To see that it is not a σ-field, all we must do is choose a set A that is countably infinite and has infinite complement. One such set can be constructed from a sequence $\omega_1, \omega_2, \ldots$, and let $A = \{\omega_1, \omega_3, \ldots\}$. Therefore $\{\omega_{2i-1}\} \in \mathcal{F}$, but $A \notin \mathcal{F}$ even though $A = \cup_i \{\omega_{2i-1}\}$.

Theorem D.2.1 *A class that is both a π-system and a λ-system is a σ-field.*

Proof. Because it is a λ-system, the class contains $\emptyset$ and is closed under formation of complements, and because it is a π-system, it is closed under finite intersections. It is therefore a field. Now, suppose that it contains sets A_i, for $i = 1, 2, \ldots$. The class then contains the sets $B_i = A_i \cap A_1^c \cap \cdots \cap A_{i-1}^c$, which are necessarily disjoint. Because it is a λ-system, it contains $\cup_i B_i$. But $\cup_i B_i = \cup_i A_i$, and since it contains $\cup_i A_i$ it is a σ-field. ∎

We can see in a similar fashion, that a class that is both a π-system and a Dynkin system is a σ-field.

A useful fact is known as *Dynkin's π-λ theorem*.

Theorem D.2.2 (Dynkin's π-λ theorem) *If $\mathcal{P}$ is a π-system and $\mathcal{L}$ is a λ-system, and if $\mathcal{P} \subset \mathcal{L}$, then*

$$\sigma(\mathcal{P}) \subset \mathcal{L}.$$

Proof. We use the given notation and assume the hypothesis. Let $\mathcal{L}_\mathcal{P}$ be the λ-system generated by $\mathcal{P}$; that is,

$$\mathcal{L}_\mathcal{P} = \lambda(\mathcal{P}).$$

$\mathcal{L}_\mathcal{P}$ is the intersection of every λ-system that contains $\mathcal{P}$, and it is contained in every λ-system that contains $\mathcal{P}$. Thus, we have

$$\mathcal{P} \subset \mathcal{L}_\mathcal{P} \subset \mathcal{L}.$$

It will now suffice to show that $\mathcal{L}_\mathcal{P}$ is also a π-system, because from the result above, if it is both a π-system and a λ-system it is a σ-field, and it contains $\mathcal{P}$ so it must be the case that $\sigma(\mathcal{P}) \subset \mathcal{L}_\mathcal{P}$ because $\sigma(\mathcal{P})$ is the minimal σ-field that contains $\mathcal{P}$.

Now define a collection of sets whose intersection with a given set is a member of $\mathcal{L}_\mathcal{P}$. For any set A, let

$$\mathcal{L}_A = \{B : A \cap B \in \mathcal{L}_\mathcal{P}\}.$$

Later in the proof, for some given set B, we use the symbol "$\mathcal{L}_B$" to denote the collection of sets whose intersection with B is a member of $\mathcal{L}_\mathcal{P}$.

If $A \in \mathcal{L}_\mathcal{P}$, then $\mathcal{L}_A$ is a λ-system, as we see by checking the conditions:

(λ_1) $A \cap \Omega = A \in \mathcal{L}_\mathcal{P}$ so $\Omega \in \mathcal{L}_A$

(λ_2') If $B_1, B_2 \in \mathcal{L}_A$ and $B_1 \subset B_2$, then $\mathcal{L}_\mathcal{P}$ contains $A \cap B_1$ and $A \cap B_2$, and hence contains the difference $(A \cap B_2) - (A \cap B_1) = A \cap (B_2 - B_1)$; that is, $B_2 - B_1 \in \mathcal{L}_A$.

(λ_3) If $B_1, B_2, \ldots \in \mathcal{L}_A$ and $B_i \cap B_j = \emptyset$ for $i \neq j$, then $\mathcal{L}_\mathcal{P}$ contains the disjoint sets $(A \cap B_1), (A \cap B_2), \ldots$ and hence their union $A \cap (\cup_i B_i)$, which in turn implies $\cup_i B_i \in \mathcal{L}_A$.

Now because $\mathcal{P}$ is a π-system,

$$\begin{aligned} A, B \in \mathcal{P} &\Rightarrow A \cap B \in \mathcal{P} \\ &\Rightarrow B \in \mathcal{L}_A \\ &\Rightarrow \mathcal{P} \subset \mathcal{L}_A \\ &\Rightarrow \mathcal{L}_\mathcal{P} \subset \mathcal{L}_A. \end{aligned}$$

(The last implication follows from the minimality of $\mathcal{L}_\mathcal{P}$ and because $\mathcal{L}_A$ is a λ-system containing $\mathcal{P}$.)

Using a similar argument as above, we have $A \in \mathcal{P}$ and $B \cap B \in \mathcal{L}_\mathcal{P}$ also imply $A \in \mathcal{L}_B$ (here $\mathcal{L}_B$is in the role of $\mathcal{L}_A$ above) and we have

$$A \in \mathcal{L}_B \iff B \in \mathcal{L}_A.$$

Continuing as above, we also have $\mathcal{P} \subset \mathcal{L}_B$ and $\mathcal{L}_\mathcal{P} \subset \mathcal{L}_B$.

Now, to complete the proof, let $B, C \in \mathcal{L}_\mathcal{P}$. This means that $C \in \mathcal{L}_B$, which from the above means that $B \cap C \in \mathcal{L}_\mathcal{P}$; that is, $\mathcal{L}_\mathcal{P}$ is a π-system, which, as we noted above is sufficient to imply the desired conclusion: $\sigma(\mathcal{P}) \subset \mathcal{L}_\mathcal{P} \subset \mathcal{L}$. ∎

Dynkin's π-λ theorem immediately implies that if $\mathcal{P}$ is a π-system then

$$\sigma(\mathcal{P}) = \lambda(\mathcal{P}).$$

Operations on σ-Fields

The usual set operators and set relations are used with collections of sets, and generally have the same meaning. If the collections of sets are σ-fields, the operaton on the collections may not yield a collection that is a σ-field, however.

Given σ-fields $\mathcal{F}_1$ and $\mathcal{F}_2$ defined with respect to a common sample space, the intersection, $\mathcal{F}_1 \cap \mathcal{F}_2$, is a σ-field. (You should work thorough an easy proof of this.) The union, $\mathcal{F}_1 \cup \mathcal{F}_2$, however, may not be a σ-field. A simple counterexample with $\Omega = \{a, b, c\}$ is

$$\mathcal{F}_1 = \{\{a\}, \{b, c\}, \emptyset, \Omega\} \quad \text{and} \quad \mathcal{F}_2 = \{\{b\}, \{a, c\}, \emptyset, \Omega\}.$$

You should show that $\mathcal{F}_1 \cup \mathcal{F}_2$ is not a σ-field.

Sub-σ-Fields

A subset of a σ-field may or may not be a σ-field. If it is, it is called a sub-σ-field.

Increasing sequences of σ-fields, $\mathcal{F}_1 \subset \mathcal{F}_2 \cdots$, are often of interest.

Given a σ-field $\mathcal{F}$, an interesting sub-σ-field can be formed by taking a specific set C in $\mathcal{F}$, and forming its intersection with all of the other sets in $\mathcal{F}$. We often denote this sub-σ-field as $\mathcal{F}_C$:

$$\mathcal{F}_C = \{C \cap A : A \in \mathcal{F}\}.$$

You should verify the three defining properties of a σ-field for $\mathcal{F}_C$.

Measurable Space: The Structure $(\Omega, \mathcal{F})$

If Ω is a sample space, and $\mathcal{F}$ is a σ-field over Ω, the double $(\Omega, \mathcal{F})$ is called a *measurable space*.

(Notice that no *measure* is required.)

Measurable spaces are fundamental objects in our development of a theory of measure and its extension to probability theory.

Subspaces

Given a measurable space $(\Omega, \mathcal{F})$, and a set $C \in \mathcal{F}$, we have seen how to form a sub-σ-field $\mathcal{F}_C$. This immediately yields a sub-measurable-space $(C, \mathcal{F}_C)$, if we take the sample space to be $\Omega \cap C = C$.

Functions and Images

A function is a set of ordered pairs such that no two pairs have the same first element. If (a, b) is an ordered pair in f, we write $b = f(a)$, a is called an argument of the function, and b is called the corresponding value of the function. If the arguments of the function are sets, the function is called a set function. The set of all arguments of the function is called the domain of the function, and the set of all values of the function is called the range of the function.

We will be interested in a function, say f, that maps one measurable space $(\Omega, \mathcal{F})$ to another measurable space $(\Lambda, \mathcal{G})$. We may write $f : (\Omega, \mathcal{F}) \mapsto (\Lambda, \mathcal{G})$, or just $f : \Omega \mapsto \Lambda$ because the argument of the function is an element of Ω (in fact, *any* element of Ω) and the value of the function is an element of Λ. It may not be the case that all elements of Λ are values of f. If it is the case that for every element $\lambda \in \Lambda$, there is an element $\omega \in \Omega$ such that $f(\omega) = \lambda$, then the function is said to be "onto" Λ. The σ-fields in the measurable spaces determine certain properties of the function.

If $(a, b) \in f$, we may write $a = f^{-1}(b)$, although sometimes this notation is restricted to the cases in which f is one-to-one. (There are some subtleties here; if f is not one-to-one, if the members of the pairs in f are reversed, the resulting set is not a function. We say f^{-1} does not exist; yet we may write $a = f^{-1}(b)$, with the meaning above.)

If $A \subset \Omega$, the *image* of A, denoted by $f[A]$, is the set of all $\lambda \in \Lambda$ for which $\lambda = f(\omega)$ for some $\omega \in \Omega$. Likewise, if $\mathcal{C}$ is a collection os subsets of Ω, the *image* of $\mathcal{C}$, denoted by $f[\mathcal{C}]$, or just by $f(\mathcal{C})$, is the set of all subsets of Λ that are images of the subsets of $\mathcal{C}$. (While I prefer the notation "$[\cdot]$" when the argument of the function is a set — unless the funtion is a set function — in most cases I will just use the "$(\cdot)$", which applies more properly to an element.)

For a subset B of Λ, the *inverse image* or the *preimage* of B, denoted by $f^{-1}[B]$, is the set of all $\omega \in \Omega$ such that $f(\omega) \in B$. We also write $f[f^{-1}[B]]$ as $f \circ f^{-1}[B]$. The set $f[f^{-1}[B]]$ may be a proper subset of B; that is, there may be an element λ in B for which there is no $\omega \in \Omega$ such that $f(\omega) = \lambda$. If there is no element $\omega \in \Omega$ such that $f(\omega) \in B$, then $f^{-1}[B] = \emptyset$.

The following are useful facts (or conventions):

- $f[\emptyset] = \emptyset$ and $f^{-1}[\emptyset] = \emptyset$.
  We can take this as a convention.
- For $B \subset \Lambda$, $f^{-1}[B^c] = (f^{-1}[B])^c$ (where $B^c = \Lambda - B$, and $(f^{-1}[B])^c = \Omega - f^{-1}[B]$).
  We see this in the standard way by showing that each is a subset of the other.
  Let ω be an arbitrary element of Ω.
  Suppose $\omega \in f^{-1}[B^c]$. Then $f(\omega) \in B^c$, so $f(\omega) \notin B$, hence $\omega \notin f^{-1}[B]$, and so $\omega \in (f^{-1}[B])^c$. We have $f^{-1}[B^c] \subset (f^{-1}[B])^c$.
  Now suppose $\omega \in (f^{-1}[B])^c$. Then $\omega \notin f^{-1}[B]$, so $f(\omega) \notin B$, hence $f(\omega) \in B^c$, and so $\omega \in f^{-1}[B^c]$. We have $(f^{-1}[B])^c \subset f^{-1}[B^c]$.
- Let $A_1, A_2, \ldots \subset \Lambda$ and suppose $(\cup_{i=1}^{\infty} A_i) \subset \Lambda$, then $f^{-1}[\cup_{i=1}^{\infty} A_i] = \cup_{i=1}^{\infty} f^{-1}(A_i)$.
  Again, let λ be an arbitrary element of Λ.
  Suppose $\lambda \in f^{-1}[\cup_{i=1}^{\infty} A_i]$. Then $f(\lambda) \in \cup_{i=1}^{\infty} A_i$, so for some j, $f(\lambda) \in A_j$ and $\lambda \in f^{-1}[A_j]$; hence $\lambda \in \cup_{i=1}^{\infty} f^{-1}[A_i]$. We have $f^{-1}[\cup_{i=1}^{\infty} A_i] \subset \cup_{i=1}^{\infty} f^{-1}[A_i]$.
  Now suppose $\lambda \in \cup_{i=1}^{\infty} f^{-1}[A_i]$. Then for some j, $\lambda \in f^{-1}[A_j]$, so $f(\lambda) \in A_j$ and $f(\lambda) \in \cup_{i=1}^{\infty} A_i$; hence $\lambda \in f^{-1}[\cup_{i=1}^{\infty} A_i]$. We have $\cup_{i=1}^{\infty} f^{-1}[A_i] \subset f^{-1}[\cup_{i=1}^{\infty} A_i])$.

Measurable Function

If $(\Omega, \mathcal{F})$ and $(\Lambda, \mathcal{G})$ are measurable spaces, and f is a mapping from Ω to Λ, with the property that $\forall A \in \mathcal{G}, f^{-1}[A] \in \mathcal{F}$, then f is a *measurable* function with respect to $\mathcal{F}$ and $\mathcal{G}$. It is also said to be measurable $\mathcal{F}/\mathcal{G}$. Note that

$$\{f^{-1}[A] : A \in \mathcal{G}\} = f^{-1}[\mathcal{G}],$$

where we have extended the notation "$f[\cdot]$" to collections of sets, in the obvious way.

For a real-valued function, that is, a mapping from Ω to $\mathbb{R}$, or in other cases where there is an "obvious" σ-field, we often just say that the function is measurable with respect to $\mathcal{F}$. In any event, the role of $\mathcal{F}$ is somewhat more important. We use the notation $f \in \mathcal{F}$ to denote the fact that f is measurable with respect to $\mathcal{F}$. (Note that this is an abuse of the notation, because f is not one of the sets in the collection $\mathcal{F}$.)

Given the measurable spaces $(\Omega, \mathcal{F})$ and $(\Lambda, \mathcal{G})$ and a mapping f from Ω to Λ, we also call call f a mapping from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$.

Note that a measurable function $f(\cdot)$ does not depend on a measure. The domain of $f(\cdot)$ has no relationship to $\mathcal{F}$, except through the range of $f(\cdot)$ that happens to be in the subsets in $\mathcal{G}$.

σ-Field Generated by a Measurable Function

If f is a measurable function from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$, then we can see that $f^{-1}[\mathcal{G}]$ is a sub-σ-field of $\mathcal{F}$. We call this the σ-field generated by f, and write it as $\sigma(f)$.

For measurable functions f and g from and to the same measurable spaces, we may write $\sigma(f, g)$, with the obvious meaning. As with σ-fields generated by sets discussed above, it is clear that $\sigma(f) \subset \sigma(f, g)$.

For measurable functions f and g from $(\Omega, \mathcal{F})$ to $(\Omega, \mathcal{F})$, it is clear that

$$\sigma(g \circ f) \subset \sigma(f). \tag{D.46}$$

Cartesian Products

The cartesian product of two sets A and B, written $A \times B$, is the set of all doubletons, (a_i, b_j), where $a_i \in A$ and $b_j \in B$. The cartesian product of two collections of sets is usually interpreted as the collection consisting of all possible cartesian products of the elements of each, e.g., if $\mathcal{A} = \{A_1, A_2\}$ and $\mathcal{B} = \{B_1, B_2\}$

$$\mathcal{A} \times \mathcal{B} = \{A_1 \times B_1,\ A_1 \times B_2,\ A_2 \times B_1,\ A_2 \times B_2\},$$

that is,

$$\{\{(a_{1i}, b_{1j}) \,|\, a_{1i} \in A_1, b_{1j} \in B_1\},\ \{(a_{1i}, b_{2j}) \,|\, a_{1i} \in A_1, b_{2j} \in B_2\},$$

$$\{(a_{2i}, b_{1j}) \,|\, a_{2i} \in A_2, b_{1j} \in B_1\},\ \{(a_{2i}, b_{2j}) \,|\, a_{2i} \in A_2, b_{2j} \in B_2\}\}.$$

The cartesian product of two collections of sets is not a very useful object, because, as we see below, important characteristics of the collections, such as being σ-fields do not carry over to the product.

Two measurable spaces $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ can be used to form a *cartesian product measurable space* with sample space $\Omega_1 \times \Omega_2$. The product of the σ-fields is not necessarily a σ-field. A simple counterexample is the same as we have used before with $\Omega = \{a, b, c\}$. Let

$$\mathcal{F}_1 = \{\{a\}, \{b, c\}, \emptyset, \Omega\} \quad \text{and} \quad \mathcal{F}_2 = \{\{b\}, \{a, c\}, \emptyset, \Omega\}.$$

The product $\mathcal{F}_1 \times \mathcal{F}_2$ contains 8 sets of doubletons, two of which are $\{(a, b)\}$ and $\{(b, b), (c, b)\}$; however, we see that their union $\{(a, b), (b, b), (c, b)\}$ is not a member of $\mathcal{F}_1 \times \mathcal{F}_2$; hence, $\mathcal{F}_1 \times \mathcal{F}_2$ is not a σ-field.

As another example, let $\Omega = \mathbb{R}$, let $\mathcal{F} = \sigma(\mathbb{R}_+) = \{\emptyset, \mathbb{R}_+, \mathbb{R} - \mathbb{R}_+, \mathbb{R}\}$, let $\mathcal{G}_1 = \sigma(\mathbb{R}_+ \times \mathbb{R}_+)$, and let $\mathcal{G}_2 = \sigma(\{F_i \times F_j \ : \ F_i, F_j \in \mathcal{F}\})$. We see that $\mathcal{G}_1 \neq \mathcal{G}_2$, because, for example, $\mathbb{R}_+ \times \mathbb{R}$ is in $\mathcal{G}_2$ but it is not in $\mathcal{G}_1$.

Given $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$, we define the cartesian product measurable space as $(\Omega_1 \times \Omega_2, \sigma(\mathcal{F}_1 \times \mathcal{F}_2))$.

The collection $\sigma(\mathcal{F}_1 \times \mathcal{F}_2)$ is a product σ-field.

Product measure spaces provide us the basis for developing a probability theory for vectors and multivariate distributions.

D.2.2 Sets in $\mathbb{R}$

First, recall some important definitions:

A set A of real numbers is called *open* if for each $x \in A$, there exists a $\delta > 0$ such that for each y with $|x - y| < \delta$ belongs to A.
A real number x is called a *point of closure* of a set A of real numbers if for every $\delta > 0$ there exists a y in A such that $|x - y| < \delta$. (Notice that every $y \in A$ is a point of closure of A.)
We denote the set of points of closure of A by $\overline{A}$.
A set A is called *closed* if $A = \overline{A}$.

Some simple facts follow:

- The intersection of a finite collection of open sets is open.
- The union of a countable collection of open sets is open.
- The union of a finite collection of closed sets is closed.
- The intersection of a countable collection of closed sets is closed.

Notice what is *not* said above (where we use the word "finite").

Intervals in $\mathbb{R}$

A very important type of set is an *interval* in $\mathbb{R}$. Intervals are the basis for building important structures on $\mathbb{R}$.

All intervals are Borel sets.

The main kinds of intervals have forms such as
$(-\infty, a)$, $(-\infty, a]$, (a, b), $[a, b]$, $(a, b]$, $[a, b)$, (b, ∞), and $[b, \infty)$.

$(-\infty, a)$, (a, b), and (b, ∞) are open;
$[a, b]$ is closed;
$(a, b]$ and $[a, b)$ are neither (they are "half-open").
$(-\infty, a]$ and $[b, \infty)$ are closed, although in a special way that sometimes requires special treatment.

$\overline{(a, b)} = [a, b]$.

Some simple facts:

- $\cap_{i=1}^{n}(a_i, b_i) = (a, b)$ (some open interval)
- $\cup_{i=1}^{\infty}(a_i, b_i)$ is an open set
- $\cup_{i=1}^{n}[a_i, b_i]$ is a closed set
- $\cap_{i=1}^{\infty}[a_i, b_i] = [a, b]$ (some closed interval)

Notice what is *not* said above (where limit is n rather than ∞).
(Also note that intersection of intervals are intervals, but of course unions may not be.)

In passing, we might note a simple version of the Heine-Borel theorem: If $[a, b] \subset \cup_{i=1}^{\infty}(a_i, b_i)$, then for some finite n, $[a, b] \subset \cup_{i=1}^{n}(a_i, b_i)$.

Two types of interesting intervals are

$$\left(a - \frac{1}{i},\ b + \frac{1}{i}\right)$$

and

$$\left[a + \frac{1}{i},\ b - \frac{1}{i}\right].$$

Notice, of course, that

$$\lim_{i\to\infty}\left(a - \frac{1}{i},\ b + \frac{1}{i}\right) = [a, b]$$

and

$$\lim_{i\to\infty}\left[a + \frac{1}{i},\ b - \frac{1}{i}\right] = [a, b].$$

Some Basic Facts about Sequences of Unions

Infinite intersections and unions behave differently with regard to collections of open and closed sets. Recall our earlier statements that were limited to finite intersections and unions: $\cap_{i=1}^{n}(a_i, b_i)$ is an open interval, and $\cup_{i=1}^{n}[a_i, b_i]$ is a closed set.

Now, with our open and closed intervals of the special forms, for infinite intersections and unions, we have the important facts:

$$[a, b] = \bigcap_{i=1}^{\infty}\left(a - \frac{1}{i},\ b + \frac{1}{i}\right)$$

$$(a,b) = \bigcup_{i=1}^{\infty} \left[a + \frac{1}{i},\ b - \frac{1}{i}\right]$$

Iff $x \in A_i$ for some i, then $x \in \cup A_i$.

So if $x \notin A_i$ for any i, then $x \notin \cup A_i$. (This is why the union of the closed intervals above is not a closed interval.)

Likewise, we have

$$(a,b] = \bigcup_{i=1}^{\infty} \left[a + \frac{1}{i},\ b\right]. = \bigcap_{i=1}^{\infty} \left(a,\ b + \frac{1}{i}\right).$$

Recall a basic fact about probability (which we will discuss again from time to time):

$$\lim \Pr(A_i) \neq \Pr(\lim A_i).$$

Compare this with the fact from above:

$$\lim_{n\to\infty} \bigcup_{i=1}^{n} \left[a + \frac{1}{i},\ b - \frac{1}{i}\right] \neq \bigcup \lim_{i\to\infty} \left[a + \frac{1}{i},\ b - \frac{1}{i}\right].$$

The Borel σ-Field on the Reals

Consider the sample space $\mathbb{R}$, and let $\mathcal{C}$ be the collection of all open intervals in $\mathbb{R}$. The σ-field $\sigma(\mathcal{C})$ is called the Borel σ-field over $\mathbb{R}$, and is denoted by $\mathcal{B}(\mathbb{R})$. When our primary interest is just the scalar reals $\mathbb{R}$, we often call this Borel σ-field just the Borel field, and i denoted it by $\mathcal{B}$.

Borel Sets

Any set in $\mathcal{B}$ is called a Borel set. The following are Borel sets:

- $\mathbb{R}$
- $\emptyset$
- any countable set; in particular, any finite set, $\mathbb{Z}$, $\mathbb{Z}_+$ (the *natural numbers*), and the set of all rational numbers
- hence, from the foregoing, the set of all irrational numbers (which is uncountable)
- any interval, open, closed, or neither
- the Cantor set
  We see this by writing the Cantor set as $\cap_{i=1}^{\infty} C_i$, where

  $$C_1 = [0,1/3] \cup [2/3,1], \quad C_2 = [0,1/9] \cup [2/9,1/3] \cup [2/3,7/9] \cup [8/9,1], \quad \ldots,$$

  and realizing that each of these is Borel.
- any union of any of the above

So, are all subsets of $\mathbb{R}$ Borel sets?

No. Interestingly enough, the cardinality of $\mathcal{B}$ can be shown to be the same as that of $\mathbb{R}$, and the cardinality of the collection of all subsets of $\mathbb{R}$, that is, the cardinality of the power set, $2^{\mathbb{R}}$, is much larger – which means there are *many* subsets of $\mathbb{R}$ that are not Borel sets. Construction of a non-Borel set uses the Axiom of Choice, which can be used to construct some truly weird sets.

Equivalent Definitions of the Borel σ-Field

The facts that unions of closed sets may be open and that intersections of open intervals may be closed allow us to characterize the Borel σ-field $\mathcal{B}$ in various ways. The canonical definition is that $\mathcal{B} = \sigma(\mathcal{C})$, where $\mathcal{C}$ is the collection of all finite open intervals.

If $\mathcal{D}$ is the collection of all finite *closed* intervals then $\mathcal{B} = \sigma(\mathcal{D})$.

Proof.

To show that the σ-fields generated by two collections $\mathcal{C}$ and $\mathcal{D}$ are the same, we use the fact that a σ-field is closed with respect to countable intersections (remember the usual definition requires that it be closed with respect to countable unions) and then we show that (1) $C \in \mathcal{C} \Rightarrow C \in \sigma(\mathcal{D})$ and (2) $D \in \mathcal{D} \Rightarrow D \in \sigma(\mathcal{C})$.

Hence, (1) assume D =[a,b]$\in \mathcal{D}$. Now, consider the sequence of sets $B_i = (a - 1/i,\ b + 1/i)$. These open intervals are in $\mathcal{B}$, and hence, $\bigcap_{i=1}^{\infty}(a - 1/i,\ b + 1/i) = [a, b] \in \mathcal{B}$.

Next, (2) let (a, b) be any set in the generator collection of $\mathcal{B}$, and consider the sequence of sets $D_i = [a + 1/i,\ b - 1/i]$, which are in $\mathcal{D}$. By definition, we have $\bigcup_{i=1}^{\infty}[a + 1/i,\ b - 1/i] = (a, b) \in \sigma(\mathcal{D})$

Likewise, if $\mathcal{A}$ is the collection of all intervals of the form (a, ∞), then $\mathcal{B} = \sigma(\mathcal{A})$. The proof of this is similar to that of the previous statement.

We also get the same Borel field by using other collections of intervals as the generator collections.

The σ-Field $\mathcal{B}_{[0,1]}$

We are often interested in some subspace of $\mathbb{R}^d$, for example an interval (or rectangle). One of the most commonly-used intervals in $\mathbb{R}$ is $[0, 1]$.

For the sample space $\Omega = [0, 1]$, the most useful σ-field consists of the collection of all sets of the form $[0, 1] \cap B$, where $B \in \mathcal{B}(\mathbb{R})$. We often denote this σ-field as $\mathcal{B}_{[0,1]}$.

The σ-field formed in this way is the same as the σ-field generated by all open intervals on $[0, 1]$; that is, $\mathcal{B}([0, 1])$. (The reader should show this, of course.)

Product Borel σ-Fields

For the product measurable space generated by $\mathbb{R}^d$, a σ-field is $\sigma(\mathcal{B}^d)$, which we denote as $\mathcal{B}(\mathbb{R}^d)$, or occasionally in a slight abuse of notation, merely as $\mathcal{B}^d$. (Recall that a product of σ-fields is not necessarily a σ-field.)

We denote this product measurable space as $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, or in the abused notation, $(\mathbb{R}^d, \mathcal{B}^d)$.

It can be shown that $\mathcal{B}(\mathbb{R}^d)$ is the same as the σ-field generated by all open intervals (or "hyperrectangles") in $\mathbb{R}^d$.

Borel Function

A measurable function from $(\Omega, \mathcal{F})$ to $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is said to be *Borel measurable* with respect to $\mathcal{F}$. A function that is Borel measurable is called a *Borel function*.

Random Variable

A measurable function from any measurable space $(\Omega, \mathcal{F})$ to the measurable space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is called a *random variable*, or, to be more specific, a d-variate *random variable*.That is, "Borel function" and "random variable" are synonymous. (Most authors define a random variable only for the case $d = 1$, and for the case of $d \geq 1$, call the Borel function a "random vector". I see no reason for this distinction.)

This is the definition of the phrase *random variable*. Notice that the words "random" and "variable" do not carry any separate meaning.

D.2.3 Measure

A measure is a scalar real-valued nonnegative set function whose domain is a σ-field with the properties that the measure of the null set is 0 and the measure of the union of any collection of disjoint sets is the sum of the measures of the sets.

From this simple definition, several properties derive. For example, if ν is a measure with domain $\mathcal{F}$ then

- $\nu(\emptyset) = 0$
- if $A_1 \subset A_2 \in \mathcal{F}$, then $\nu(A_1) \leq \nu(A_2)$ (this is "monotonicity")
- if $A_1, A_2, \ldots \in \mathcal{F}$, then $\nu(\cup_i A_i) \leq \sum_i \nu(A_i)$ (this is "subadditivity")
- if $A_1 \subset A_2 \subset \ldots \in \mathcal{F}$, then $\nu(\cup_{i=1}^{\infty} A_i) = \lim_{i \to \infty} \nu(A_i)$ (this is "continuity from below"; think of the A_is as nested intervals).
  We can see this by defining a sequence of disjoint sets, $B_1, B_2, \ldots$ as $B_j = A_{j+1} - A_j$, so $\cup_{j=1}^{i} B_j = A_i$, and

$$\cup_{i=1}^{\infty} B_i = \cup_{i=1}^{\infty} A_i$$

Hence,

$$\begin{aligned}\nu\left(\cup_{i=1}^{\infty} A_i\right) &= \nu\left(\cup_{i=1}^{\infty} B_i\right)\\ &= \sum_{i=1}^{\infty} \nu(B_i)\\ &= \lim_{i\to\infty} \sum_{j=1}^{i} \nu(B_j)\\ &= \lim_{i\to\infty} \nu\left(\cup_{j=1}^{i} B_j\right)\\ &= \lim_{i\to\infty} \nu(A_i).\end{aligned}$$

A measure ν such that $\nu(\Omega) < \infty$ is called a *finite measure*.

Sequences of nested intervals are important. We denote a sequence $A_1 \subset A_2 \subset \ldots$ with $A = \cup_{i=1}^{\infty} A_i$, as $A_i \nearrow A$. (This same notation is used for a sequence of real numbers x_i such that $x_1 \leq x_2 \cdots$ and $\lim x_i = x$; that is, in that case, we write $x_i \nearrow x$.)

Continuity from below is actually a little stronger than what is stated above, because the sequence of values of the measure is also monotonic: for $A_i \in \mathcal{F}, A_i \nearrow A \Rightarrow \nu(A_i) \nearrow \nu(A)$

A similar sequence is $A_1 \supset A_2 \supset \ldots$ with $A = \cup_{i=1}^{\infty} A_i$. We denote this as $A_i \searrow A$. Continuity from above is the fact that for $A_i \in \mathcal{F}$, if $A_i \searrow A$ and $\nu(A_1) < \infty$, then $\nu(A_i) \searrow \nu(A)$.

Notice that the definition of a measure does not preclude the possibility that the measure is identically 0. This often requires us to specify "nonzero measure" in order to discuss nontrivial properties. Another possibility, of course, would be just to specify $\nu(\Omega) > 0$ (remember $\Omega \neq \emptyset$ in a measurable space).

To evaluate $\nu(\cup_i A_i)$ we form disjoint sets by intersections. For example, we have $\nu(A_1 \cup A_2) = \nu(A_1) + \nu(A_2) - \nu(A_1 \cap A_2)$. This is an application of the simplest form of the inclusion-exclusion formula (see page 345). If there are three sets, we take out all pairwise intersections and then add back in the triple intersection. We can easily extend this (the proof is by induction) so that, in general, we have

$$\begin{aligned}\nu(\cup_i^n A_i) = &\sum_{1\leq i\leq n} \nu(A_i) - \sum_{1\leq i<j\leq n} \nu(A_i \cap A_j) +\\ &\sum_{1\leq i<j<k\leq n} \nu(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n+1} \nu(A_1 \cap \cdots \cap A_n).\end{aligned}$$

Lebesgue measure

If $\mathcal{F}$ is the Borel σ-field, the most common measure is the *Lebesgue measure*, which is defined by the relation

$$\nu((a, b)) = b - a.$$

counting measure

If $\mathcal{F}$ is a countable σ-field, the most common measure is the *counting measure*, which is defined by the relation

$$\nu(A) = \#(A).$$

Dirac measure

If $\mathcal{F}$ is the collection of all subsets of Ω, a useful measure at a fixed point $\omega \in \Omega$ is the *Dirac measure* concentrated at ω, usually denoted by δ_ω, and defined by

$$\delta_\omega(A) = \begin{cases} 1 \text{ if } \omega \in A \\ 0 \text{ otherwise.} \end{cases}$$

Measurable Set

If ν is a measure with domain $\mathcal{F}$ then every set in $\mathcal{F}$ is said to be *ν-measurable*, or just *measurable*.

Note that unlike other terms above that involve "measurable", this term is defined in terms of a given measure.

Measure Space: The Structure $(\Omega, \mathcal{F}, \nu)$

If Ω is a sample space, $\mathcal{F}$ is a σ-field over Ω, and ν is a measure with domain $\mathcal{F}$, the triple $(\Omega, \mathcal{F}, \nu)$ is called a *measure space* (compare *measurable space*, above).

The elements in the measure space can be any kind of objects. They do not need to be numbers.

If $(\Omega, \mathcal{F}, \nu)$ is a measure space, and for some set $C \in \mathcal{F}$, $(C, \mathcal{F}_C)$ is a sub measurable space as described above, then the function ν restricted to $\mathcal{F}_C$ is a measure and $(C, \mathcal{F}_C, \nu)$ is a measure space. It corresponds in a natural way to the subsetting operations.

Any set $A \in \mathcal{F}$ such that $\nu(A^c) = 0$ is called a *support of the measure* and ν is said to be *concentrated* on A. If ν is a finite measure then A is a support iff $\mu(A) = \nu(\Omega)$.

σ-Finite Measure

A measure ν is *σ-finite* on $(\Omega, \mathcal{F})$ iff there exists a sequence $A_1, A_2, \ldots$ in $\mathcal{F}$ such that $\cup_i A_i = \Omega$ and $\nu(A_i) < \infty$ for all i.

A finite measure is σ-finite.

Although it is not finite, the Lebesgue measure is σ-finite, as can be seen from the sequence of open intervals $(-i, i)$.

The counting measure is σ-finite iff Ω is countable. If Ω is finite, the counting measure is finite.

Almost Everywhere (a.e.)

Given a measure space, a property that holds for any element of the σ-field with positive measure is said to hold *almost everywhere*, or a.e.

Probability Measure

A measure whose domain is a σ-field defined on the sample space Ω with the property that $\nu(\Omega) = 1$ is called a *probability measure*. We often use P to denote such a measure.

A property that holds a.e. with respect to a probability measure is said to hold *almost surely*, or a.s.

Probability Space

If P in the measure space $(\Omega, \mathcal{F}, P)$ is a probability measure, the triple $(\Omega, \mathcal{F}, P)$ is called a *probability space*. A set $A \in \mathcal{F}$ is called an "event".

Induced Measure

If $(\Omega, \mathcal{F}, \nu)$ is a measure space and $(\Lambda, \mathcal{G})$ is a measurable space, and f is a function from Ω to Λ that is measurable with respect to $\mathcal{F}$, then the domain and range of the function $\nu \circ f^{-1}$ is $\mathcal{G}$ and it is a measure.

The measure $\nu \circ f^{-1}$ is called an *induced measure* on $\mathcal{G}$. (It is induced from the measure space $(\Omega, \mathcal{F}, \nu)$.)

Radon Measure

For a topological measurable space $(\Omega, \mathcal{F})$ with a metric (that is, a space in which the concept of a compact set is defined), a measure μ such that for every compact set $B \in \mathcal{F}$, $\mu(B) < \infty$ is called a *Radon measure*.

A Radon measure is σ-finite, although it is not necessarily finite.

The Lebesgue and Dirac measures are Radon measures.

Product Measures

Given measure spaces $(\Omega_1, \mathcal{F}_1, \nu_1)$ and $(\Omega_2, \mathcal{F}_2, \nu_2)$, we define the cartesian product measure space as $(\Omega_1 \times \Omega_2, \sigma(\mathcal{F}_1 \times \mathcal{F}_2), \nu_1 \times \nu_2)$, where the product measure $\nu_1 \times \nu_2$ is defined on the the product σ-field $\sigma(\mathcal{F}_1 \times \mathcal{F}_2)$ to have the property for $A_1 \in \mathcal{F}_1$ and $A_2 \in \mathcal{F}_2$

$$\nu_1 \times \nu_2(A_1 \times A_2) = \nu_1(A_1)\nu_2(A_2).$$

It can be shown that the measure with this property is unique.

D.2.4 Real-Valued Functions over Real Domains

In the foregoing we have given special consideration to real-valued functions; random variables are real-valued functions; measurable real-valued functions are called Borel functions. In the following we will define integrals and derivatives of real-valued functions.

For real-valued functions over real domains we identify some additional properties. We have already defined similar properties for functions that are measures (real-valued set functions), and we could define them for other functions from one measure space to another. The definitions for real-valued functions over real domains are simpler and also more useful.

One important function is the indicator function.

Definition D.2.7 (indicator function) *The indicator function, denoted* $\mathrm{I}_S(x)$ *for a given set* S*, is defined by* $\mathrm{I}_S(x) = 1$ *if* $x \in S$ *and* $\mathrm{I}_S(x) = 0$ *otherwise.*

Notice that $\mathrm{I}_S^{-1}[A] = \emptyset$ if $0 \notin A$ and $1 \notin A$; $\mathrm{I}_S^{-1}[A] = S$ if $0 \notin A$ and $1 \in A$; $\mathrm{I}_S^{-1}[A] = S^{\mathrm{c}}$ if $0 \in A$ and $1 \notin A$; and $\mathrm{I}_S^{-1}[A] = \Omega$ if $0 \in A$ and $0 \in A$. Hence, $\sigma(\mathrm{I}_S)$ is the second most trivial σ-field we referred to earlier; i.e., $\sigma(S) = \{\emptyset, S, S^{\mathrm{c}}, \Omega\}$.

Is I_S measurable? (What do we need to know to answer that?)

Definition D.2.8 (simple function) *If* $A_1, \ldots, A_k$ *are measurable subsets of* Ω *and* $a_1, \ldots, a_k$ *are constant real numbers, a function* φ *is a simple function if for* $\omega \in \Omega$*,*

$$\varphi(\omega) = \sum_{i=1}^{k} a_i \mathrm{I}_{A_i}(\omega),$$

where $\mathrm{I}_S(x)$ *is the indicator function.*

Continuity is an important property of some functions. For real functions on a real domain, we typically define continuity in terms of the Euclidean distance between two points in the domain and the Euclidean distance between the corresponding function values.

There are various types of continuity, and two functions help to illustrate the differences.

Definition D.2.9 (continuous function) *Let* f *be a real-valued function whose domain is a set* $D \subset \mathbb{R}^d$*. We say that* f *is continuous at the point* $x \in D$ *if, given* $\epsilon > 0, \exists \delta \ni \forall y \in D \ni \|x - y\| < \delta, \|f(x) - f(y)\| < \epsilon$.

Here, the norms are the Euclidean norms. Notice that the order of $f(x)$ may be different from the order of x.

The δ in the definition may depend on x as well as on ϵ.

If f is continuous at each point in a subset of its domain, we say it is continuous on that subset.

If f is continuous at each point in its domain, we say that f is *continuous*.

From this definition we have an immediate useful fact about continuous functions: the inverse image of an open set is open.

Definition D.2.10 (uniformly continuous function) *Let f be a real-valued function whose domain includes a set $D \subset \mathbb{R}^d$. We say that f is uniformly continuous over D if, given $\epsilon > 0, \exists \delta \ni \forall x, y \in D \ni \|x - y\| < \delta, \|f(x) - f(y)\| < \epsilon$.*

Continuity is a point-wise property, while uniform continuity is a global property.

The function $f(x) = 1/x$ is continuous on $(0, \infty)$, but is not uniformly continuous over that interval. This function is, however, uniformly continuous over any closed and bounded subinterval of $(0, \infty)$. The Heine-Cantor theorem, in fact, states that any function that is continuous over a compact set is uniformly continuous over that set.

If $\{x_n\}$ is a Cauchy sequence in the domain of a a uniformly continuous function f, then $\{f(x_n)\}$ is also a Cauchy sequence.

If a function f is uniformly continuous over a finite interval (a, b), then f is bounded over (a, b).

Definition D.2.11 (absolutely continuous function) *Let f be a real-valued function defined on $[a, b]$ (its domain may be larger). We say that f is absolutely continuous on $[a, b]$ if, given $\epsilon > 0$, there exists a δ such that for every finite collection of nonoverlapping open rectangles $(x_i, y_i) \subset [a, b]$ with*

$$\sum_{i=1}^{n} \|x_i - y_i\| < \delta,$$

$$\sum_{i=1}^{n} \|f(x_i) - f(y_i)\| < \epsilon.$$

We also speak of local absolutely continuity in the obvious way.

If f is absolutely continuous over D, it is uniformly continuous on D, but the converse is not true. The Cantor function, defined over the interval $[0, 1]$, is an example of a function that is continuous everywhere, and hence, uniformly continuous on that compact set, but not absolutely continuous. The Cantor function takes different values over the different intervals used in the construction of the Cantor set. Let $f_0(x) = x$, and then for $n = 0, 1, \ldots$, let

$$\begin{aligned} f_{n+1}(x) &= 0.5 f_n(3x) \quad \text{for } 0 \le x < 1/3 \\ f_{n+1}(x) &= 0.5 \quad \text{for } 1/3 \le x < 2/3 \\ f_{n+1}(x) &= 0.5 + 0.5 f_n(3(x - 2/3)) \quad \text{for } 2/3 \le x \le 1. \end{aligned}$$

The Cantor function is $\lim_{n\to\infty} f_n$.

The Cantor function has a derivative of 0 almost everywhere, but has no derivative at any member of the Cantor set.

An absolutely continuous function is of bounded variation; it has a derivative almost everywhere; and if the derivative is 0 a.e., the function is constant.

Definition D.2.12 (Lipschitz-continuous function) *Let f be a real-valued function whose domain is an interval $D \subset \mathbb{R}^d$. We say that f is Lipschitz-continuous if for any $y_1, y_2 \in D$ and $y_1 \neq y_2$, there exists γ such that*

$$\|f(y_1) - f(y_2)\| \leq \gamma \|y_1 - y_2\|.$$

The smallest γ for which the inequality holds is called the Lipschitz constant.

We also speak of local Lipschitz-continuity in the obvious way.

Every Lipschitz-continuous function is absolutely continuous.

Lipschitz-continuity plays an important role in nonparametric function estimation.

The graph of a scalar-valued Lipschitz-continuous function f over $D \subset \mathbb{R}$ has the interesting geometric property that the entire graph of $f(x)$ lies between the lines $y = f(c) \pm \gamma(x - c)$ for any $c \in D$.

The following theorem is useful because it allows us to build up any measurable real-valued function from a sequence of simple functions.

Theorem D.2.3 *Every measurable real-valued function can be represented at any point as the limit of a sequence of simple functions.*

Proof. Let f be real and measurable. Now, if $f(\omega) \geq 0$, there exists a sequence $\{f_n\}$ of simple functions such that

$$0 \leq f_n(\omega) \nearrow f(\omega) \quad \text{a.s.},$$

and if $f(\omega) \leq 0$, there exists a sequence $\{f_n\}$ of simple functions such that

$$0 \geq f_n(\omega) \searrow f(\omega) \quad \text{a.s.}$$

The sequence is

$$f_n(\omega) = \begin{cases} -n & \text{if} \quad f(\omega) \leq -n, \\ -(k-1)2^{-n} & \text{if} \quad -k2^{-n} < f(\omega) \leq -(k-1)2^{-n}, \text{ for } 1 \leq k \leq n2^{-n}, \\ (k-1)2^{-n} & \text{if} \quad (k-1)2^{-n} < f(\omega) < k2^{-n}, \text{ for } 1 \leq k \leq n2^{-n}, \\ n & \text{if} \quad n \leq f(\omega). \end{cases}$$

∎

As a corollary of Theorem D.2.3, we have that for a nonnegative random variable X, there exists a sequence of simple (degenerate) random variables $\{X_n\}$ such that

$$0 \leq X_n \nearrow X \quad \text{a.e.}$$

D.2.5 Integration

Integrals are some of the most important functionals of real-valued functions. Integrals and the action of integration are defined using measures. Integrals of nonnegative functions are themselves measures. There are various types of integrals, Lebesgue, Riemann, Riemann-Stieltjes, Ito, and so on. The most important is the Lebesgue, and when we use the term "integral" without qualification that will be the integral meant.

The Lebesgue Integral of a Function with Respect to a Given Measure: The Definition

An *integral of a function f with respect to a given measure ν*, if it exists, is a functional whose value is an average of the function weighted by the measure. It is denoted by $\int f \,\mathrm{d}\nu$. The function f is called the *integrand*.

The integral is defined over the sample space of a given measure space, say $(\Omega, \mathcal{F}, \nu)$. This is called the *domain* of the integral. We often may consider integrals over different domains formed from a sub measure space, $(D, \mathcal{F}_D, \nu)$ for some set $D \in \mathcal{F}$, as described above. We often indicate the domain explicitly by notation such as $\int_D f \,\mathrm{d}\nu$.

If the domain is a real interval $[a, b]$, we often write the restricted interval as $\int_a^b f \,\mathrm{d}\nu$. If ν is the Lebesgue measure, this integral is the same as the integral over the interval (a, b).

We also write an integral in various equivalent ways. For example if the integrand is a function of real numbers and our measure is the Lebesgue measure, we may write the integral over the interval (a, b) as $\int_a^b f(x) \,\mathrm{d}x$.

We build the definition of an integral of a function in three steps.

1. simple function.
   If f is a simple function defined as $f(\omega) = \sum_{i=1}^k a_i \mathrm{I}_{A_i}(\omega)$, where the A_is are measurable with respect to ν, then
   $$\int f \,\mathrm{d}\nu = \sum_{i=1}^k a_i \nu(A_i).$$
   (Note that a simple function over measurable A_is is measurable.)
2. nonnegative Borel function.
   We define the integral of a nonnegative Borel function in terms of the supremum of a collection of simple functions.
   Let f be a nonnegative Borel function with respect to ν on Ω, and let S_f be the collection of all nonnegative simple functions such that
   $$\varphi \in S_f \Rightarrow \varphi(\omega) \leq f(\omega) \forall\, \omega \in \Omega$$
   We define the integral of f with respect to ν as
   $$\int f \,\mathrm{d}\nu = \sup \left\{ \int \varphi \,\mathrm{d}\nu | \varphi \in S_f \right\}.$$
3. general Borel function.
   For a general Borel function f, we form two nonnegative Borel functions f_+ and f_- such that $f = f_+ - f_-$:
   $$f_+(\omega) = \max\{f(\omega), 0\}$$
   $$f_-(\omega) = \max\{-f(\omega), 0\}$$

We define the integral of f with respect to ν as the difference of the integrals of the two nonnegative functions:

$$\int f \,\mathrm{d}\nu = \int f_+ \,\mathrm{d}\nu - \int f_- \,\mathrm{d}\nu,$$

so long as either $\int f_+ \,\mathrm{d}\nu$ or $\int f_- \,\mathrm{d}\nu$ is finite because $\infty - \infty$ is not defined.

So for what kind of function would the Lebesgue integral not be defined? Consider $f(x) = \sin(x)/x$. We see that $\int_{\mathbb{R}_+} f \,\mathrm{d}\nu$ is not covered by the definition (because both the positive part and the negative of the negative part is ∞).

We define the *integral over a domain* for $A \subset \Omega$ as

$$\int_A f \,\mathrm{d}\nu = \int \mathrm{I}_A f \,\mathrm{d}\nu$$

If $f = f_+ - f_-$, where f_+ and f_- are nonnegative Borel functions, and both $\int f_+ \,\mathrm{d}\nu$ and $\int f_- \,\mathrm{d}\nu$ are finite, we say f is *integrable*.

Note that being Borel does not imply that a function is integrable.

A random variable is not necessarily integrable.

Given a probability space $(\Omega, \mathcal{F}, P)$ and a random variable with respect to $\mathcal{F}$, X, we define the *expected value* of X with respect to P as

$$\int X \,\mathrm{d}P,$$

and denote it as $\mathrm{E}(X)$ or for clarity, $\mathrm{E}_P(X)$.

Sometimes we limit the definition of expected value to integrable random variables X.

Notation

There are various equivalent notations for denoting an integral. If x is assumed to range over a the reals and $g(x)$ is a real-valued function then

$$\int g(x) \,\mathrm{d}x$$

may be used to denote $\int g \,\mathrm{d}\nu$, where in the former notation, we assume the measuer ν.

If the measure is a probability measure P with associated CDF F, all of the following notations are equivalent:
$\int g \,\mathrm{d}P$, $\int g(x) \,\mathrm{d}P$, $\int g \,\mathrm{d}F$, $\int g(x) \,\mathrm{d}F(x)$

Measures Defined by Integrals

The integral over a domain together with a nonnegative Borel function leads to an induced measure: If a given measure space $(\Omega, \mathcal{F}, \nu)$ and a given nonnegative Borel function f, let $\lambda(A)$ for $A \subset \Omega$ be defined as

$$\lambda(A) = \int_A f \, \mathrm{d}\nu.$$

Then $\lambda(A)$ is a measure over $(\Omega, \mathcal{F})$. (Exercise.)

If $f \equiv 1$ the integral with respect to a given measure defines the same measure. This leads to the representation of the probability of an event as an integral. Given a probability space $(\Omega, \mathcal{F}, P)$, $\int_A \mathrm{d}P$ is the *probability of* A, written $P(A)$ or $\Pr(A)$.

The properties of a measure defined by an integral depend on the properties of the underlying measure space and the function. For example, in $\mathbb{R}$ with a Lebesgue measure the measure for Borel sets of positive reals defined by

$$\lambda(A) = \int_A \frac{1}{x} \, \mathrm{d}x$$

has an interesting invariance property. For any positive real number a, let $aA = Aa = \{x : x/a \in A\}$. Then we have $\lambda(aA) = \lambda(A)$. A measure with this kind of invariance is called a *Haar invariant measure*, or just a *Haar measure*. More interesting Haar measures are those defined over nonsingular $n \times n$ real matrices,

$$\mu(D) = \int_D \frac{1}{|\det(X)|^n} \, \mathrm{d}X,$$

or over matrices in the orthogonal group. (See Gentle, 2007, pages 169–171.)

Properties of the Lebesgue Integral

The definition of an integral immediately yields some important properties of integrals:

- linearity: for real a and Borel f and g, $\int af + g \, \mathrm{d}\nu = a \int f \, \mathrm{d}\nu + \int g \, \mathrm{d}\nu$.
- $\int |f| \, \mathrm{d}\nu$ is a norm: $\int |f| \, \mathrm{d}\nu = 0 \Rightarrow f = 0$ a.e.
  This fact together with the linearity means that $\int |f| \, \mathrm{d}\nu$ is a *norm* for functions. A more general norm based on the integral is $(\int |f|^p \, \mathrm{d}\nu)^{1/p}$ for $1 \leq p$. (Notice that there is a slight difference here from the usual definition of a norm. It seems reasonable to allow the implication of $\int |f| \, \mathrm{d}\nu = 0$ to be only almost everywhere. Strictly speaking, without this weakened form of equality to 0, $\int |f| \, \mathrm{d}\nu$ is only a pseudonorm.)
- finite monotonicity: for integrable f and g, $f \leq g$ a.e. $\Rightarrow \int f \, \mathrm{d}\nu \leq \int g \, \mathrm{d}\nu$.

There are some conditions for interchange of an integration operation and a limit operation that are not so obvious. The following three theorems are closely related. Monotone convergence is the simplest, and Fatou's lemma is a fairly simple extension. Dominated convergence is another extension. The proofs are in Shao (although he should prove (iii) (monotone convergence) first).

- monotone convergence: if $0 \leq f_1 \leq f_2$ and $\lim_{n\to\infty} f_n = f$ a.e., then

$$\int \lim_{n\to\infty} f_n \, \mathrm{d}\nu = \lim_{n\to\infty} \int f_n \, \mathrm{d}\nu. \tag{D.47}$$

- Fatou's lemma (follows from finite monotonicity and monotone convergence): if $0 \leq f_n$ then

$$\int \liminf_n f_n \, \mathrm{d}\nu \leq \liminf_n \int f_n \, \mathrm{d}\nu. \tag{D.48}$$

- dominated convergence: if $\lim_{n\to\infty} f_n = f$ a.e. and there exists an integrable function g such that $|f_n| \leq g$ a.e., then

$$\int \lim_{n\to\infty} f_n \, \mathrm{d}\nu = \lim_{n\to\infty} \int f_n \, \mathrm{d}\nu. \tag{D.49}$$

change of variables

Consider two measurable spaces $(\Omega, \mathcal{F})$ and $(\Lambda, \mathcal{G})$, let f be a measurable function from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$, and let ν be a measure on $\mathcal{F}$. As we have seen, $\nu \circ f^{-1}$ is an induced measure on $\mathcal{G}$. Now let g be a Borel function on $(\Lambda, \mathcal{G})$. Then the integral of $g \circ f$ over Ω with respect to ν is the same as the integral of g over Λ with respect to $\nu \circ f^{-1}$:

$$\int_\Omega g \circ f \, \mathrm{d}\nu = \int_\Lambda g \, \mathrm{d}(\nu \circ f^{-1})$$

integration in a product space (Fubini's theorem)

Given two measure spaces $(\Omega_1, \mathcal{F}_1, \nu_1)$ and $(\Omega_2, \mathcal{F}_2, \nu_2)$ and a Borel function f on $\Omega_1 \times \Omega_2$, the integral over Ω_1, if it exists, is a function of $\omega_2 \in \Omega_2$ a.e., and likewise, the integral over Ω_2, if it exists, is a function of $\omega_2 \in \Omega_1$ a.e. Fubini's theorem shows that if one of these marginal integrals, say

$$g(\omega_2) = \int_{\Omega_1} f(\omega_1, \omega_2) \, \mathrm{d}\nu_1,$$

exists a.e., then the natural extension of an integral to a product space, resulting in the *double integral*,

$$\int_{\Omega_1 \times \Omega_2} f(\omega_1, \omega_2) \, \mathrm{d}\nu_1 \times \mathrm{d}\nu_2$$

is the same as the *iterated integral*,

$$\int_{\Omega_2} \left(\int_{\Omega_1} f(\omega_1, \omega_2) \, \mathrm{d}\nu_1 \right) \mathrm{d}\nu_2.$$

integration by parts

If f and g are bounded on the interval $[a, b]$ and have no common points of disconuity in that interval, then

$$\int_{[a,b]} f(x) \mathrm{d}g(x) = f(b)g(b) - f(a)g(a) - \int_{[a,b]} g(x) \mathrm{d}f(x).$$

This is proved using Fubini's theorem.

absolute continuity of a measure w.r.t. a measure; dominating measure

Given measures ν and λ on the same measurable space, $(\Omega, \mathcal{F})$, if $\forall A \in \mathcal{F}$

$$\nu(A) = 0 \quad \Rightarrow \quad \lambda(A) = 0,$$

then λ is said to be *absolutely continuous* with respect to ν.

In this case we also say that λ is *dominated* by ν.

We denote that λ is dominated by ν by

$$\lambda \ll \nu.$$

As an example, let $(\Omega, \mathcal{F}, \nu)$ be a measure space and let f be a nonnegative Borel function on Ω. Define the measure λ by

$$\lambda(A) = \int_A f \, \mathrm{d}\nu$$

for any $A \in \mathcal{F}$. Then $\nu(A) = 0 \Rightarrow \lambda(A) = 0$, and so λ is absolutely continuous with respect to ν.

If $\lambda \ll \nu$ and $\nu \ll \lambda$, then λ and ν are *equivalent*, and we write $\lambda \equiv \nu$.

If λ is finite (that is, $\lambda(A) < \infty \forall A \in \mathcal{F}$) the absolute continuity of λ with respect to ν can be characterized by an ϵ-δ relationship as used in the definition of absolute continuity of functions: Given that λ is finite, λ is absolutely continuous with respect to ν iff for any $A \in \mathcal{F}$ and for any $\epsilon > 0$, there exists a δ such that

$$\nu(A) < \delta \quad \Rightarrow \quad \lambda(A) < \epsilon.$$

Radon-Nikodym theorem

Given two measures ν and λ on the same measurable space, $(\Omega, \mathcal{F})$, such that $\lambda \ll \nu$ and ν is σ-finite. Then there exists a unique a.e. nonnegative Borel function f on Ω such that $\lambda(A) = \int_A f \mathrm{d}\nu \; \forall \, A \in \mathcal{F}$.

Uniqueness a.e. means that if also, for some g, $\lambda(A) = \int_A g \mathrm{d}\nu \; \forall \, A \in \mathcal{F}$ then $f = g$ a.e.

A proof of the Radon-Nikodym theorem is given in Billingsley (1995), page 422.

The Riemann Integral

The Riemann integral is one of the simplest integrals. We can define the Riemann integral of a real function f over the interval $(a, b]$ in terms of the Lebesgue measure λ as the real number r such that for any $\epsilon > 0$, there exists a δ such that

$$\left| r - \sum_i f(x_i)\lambda(I_i) \right| < \epsilon$$

where $\{I_i\}$ is any finite partition of $(a, b]$ such that for each i, $\lambda(I_i) < \delta$ and $x_i \in I_i$. If the Riemann integral exists, it is the same as the Lebesgue integral.

A classic example for which the Lebesgue integral exists, but the Riemann integral does not, is the function g defined over $(0, 1]$ as $g(x) = 1$ is x is rational, and $g(x) = 0$ otherwise. The Lebesgue integral $\int_0^1 g(x)\,\mathrm{d}x$ exists and equals 0, because $g(x) = 0$ a.e. The Riemann integral, on the other had does not exist because for an arbitrary partition $\{I_i\}$, the integral is 1 if $x_i \in I_i$ is taken as a rational, and the integral is 0 if $x_i \in I_i$ is taken as a irrational.

The Riemann integral lacks the three convergence properties of the Lebesgue integral given on page 391.

Derivatives

If $\lambda(A) = \int_A f\mathrm{d}\nu \; \forall \, A \in \mathcal{F}$, then f is called the *Radon-Nikodym derivative.* or just the derivative, of λ with respect to ν, and we write $f = \mathrm{d}\lambda/\mathrm{d}\nu$.

Notice an important property of the derivative: If $\mathrm{d}\lambda/\mathrm{d}\nu > 0$ over A, but $\lambda(A) = 0$, then $\nu(A) = 0$.

With this definition of a derivative, we have the familiar properties for measures λ, λ_1, λ_2, μ, and ν on the same measurable space, $(\Omega, \mathcal{F})$:

1. If $\lambda \ll \nu$, with ν σ-finite, and $f \geq 0$, then

$$\int f\mathrm{d}\lambda = \int f\frac{\mathrm{d}\lambda}{\mathrm{d}\nu}\mathrm{d}\nu.$$

2. If $\lambda_1 \ll \nu$ and $\lambda_1 + \lambda_2 \ll \nu$, with ν σ-finite, then

$$\frac{\mathrm{d}(\lambda_1 + \lambda_2)}{\mathrm{d}\nu} = \frac{\mathrm{d}\lambda_1}{\mathrm{d}\nu} + \frac{\mathrm{d}\lambda_2}{\mathrm{d}\nu} \quad \text{a.e. } \nu.$$

3. If $\lambda \ll \mu \ll \nu$, with μ and ν σ-finite, then

$$\frac{\mathrm{d}\lambda}{\mathrm{d}\nu} = \frac{\mathrm{d}\lambda}{\mathrm{d}\mu}\frac{\mathrm{d}\mu}{\mathrm{d}\nu} \quad \text{a.e. } \nu.$$

   If $\lambda \equiv \nu$, then

$$\frac{\mathrm{d}\lambda}{\mathrm{d}\nu} = \left(\frac{\mathrm{d}\nu}{\mathrm{d}\lambda}\right)^{-1} \quad \text{a.e. } \nu \text{ and } \mu.$$

4. If $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ are measurable spaces, λ_1 and ν_1, with $\lambda_1 \ll \nu_1$, are measures on $(\Omega_1, \mathcal{F}_1)$, λ_2 and ν_2, with $\lambda_2 \ll \nu_2$, are measures on $(\Omega_2, \mathcal{F}_2)$, and ν_1 and ν_2 are σ-finite, then for $\omega_1 \in \Omega_1$ and $\omega_2 \in \Omega_2$

$$\frac{\mathrm{d}(\lambda_1 + \lambda_2)}{\mathrm{d}(\nu_1 + \nu_2)}(\omega_1, \omega_2) = \frac{\mathrm{d}\lambda_1}{\mathrm{d}\nu_1}(\omega_1)\frac{\mathrm{d}\lambda_2}{\mathrm{d}\nu_2}(\omega_2) \quad \text{a.e. } \nu_1 \times \nu_2.$$

D.2.6 Real Function Spaces

Two of the most important linear spaces are finite-dimensional vector spaces and function spaces. We will discuss function spaces now, and cover vector spaces along with matrices in Section D.4.

Function spaces are sets of functions over a common domain. The standard operations of linear spaces are pointwise addition and scalar multiplication of function values.

Given the measure space $(\Omega, \mathcal{F}, \nu)$ and the real number $p \geq 1$, we denote by $\mathcal{L}^p(\nu)$ the space of all measurable functions f on Ω for which $\int |f|^p \mathrm{d}\nu < \infty$. Although the measure ν is needed to define the integral, we often drop the ν in $\mathcal{L}^p(\nu)$. If the integral is taken only over some $D \in \mathcal{F}$, we may denote the space as $\mathcal{L}^p(D)$.

An important fact about the $\mathcal{L}^p$ spaces is that they are Banach spaces (that is, among other things, they are complete). This fact is called the Riesz-Fischer theorem and is proved in most texts on real analysis.

Inner Products of Functions

The inner product (or dot product) of the real functions f and g over the domain D, denoted by $\langle f, g\rangle_D$ or usually just by $\langle f, g\rangle$, is defined as

$$\langle f, g\rangle_D = \int_D f(x)g(x)\,\mathrm{d}x \tag{D.50}$$

if the (Lebesgue) integral exists. (For complex functions, we define the inner product as $\int_D f(x)\bar{g}(x)\,\mathrm{d}x$, where $\bar{g}$ is the complex conjugate of g. In the following, we only consider real-valued functions of real arguments.)

$$\langle f, g\rangle_{(\mu;D)} = \int_D f(x)\bar{g}(x)w(x)\,\mathrm{d}x,$$

To avoid questions about integrability, we generally restrict attention to functions whose dot products with themselves exist; that is, to functions that are square Lebesgue integrable over the region of interest. These functions are members of the space $\mathcal{L}^2(D)$.

The standard properties, such as linearity and the Cauchy-Schwarz inequality, obviously hold for the inner products of functions.

We sometimes define function inner products with respect to a weight function, $w(x)$, or with respect to the measure μ, where $\mathrm{d}\mu = w(x)\mathrm{d}x$,

$$\langle f, g\rangle_{(\mu;D)} = \int_D f(x)g(x)w(x)\,\mathrm{d}x,$$

if the integral exists. Often, both the weight and the range are assumed to be fixed, and the simpler notation $\langle f, g\rangle$ is used.

Norms of Functions

The norm of a function f, denoted generically as $\|f\|$, is a mapping into the nonnegative reals that satisfies the properties of the definition of a norm given on page 352. A norm of a function $\|f\|$ is often defined as some nonnegative, strictly increasing function of the inner product of f with itself, $\langle f, f\rangle$. Not all norms are defined in terms of inner products, however.

The most common type of norm for a real-valued function is the L_p norm, denoted as $\|f\|_p$, which is defined similarly to the L_p vector norm as

$$\|f\|_p = \left(\int_D |f(x)|^p w(x)\,\mathrm{d}x\right)^{1/p}, \tag{D.51}$$

if the integral exists. It is clear that $\|f\|_p$ satisfies the properties that define a norm.

The space of functions for which the integrals in (D.51) exist is often denoted by $\mathcal{L}^p_{(w;D)}$, or just $\mathcal{L}^p$. A common value of p is 2, as noted above.

Often μ is taken as Lebesgue measure, and $w(x)\mathrm{d}x$ becomes $\mathrm{d}x$. This is a uniform weighting.

A common L_p function norm is the L_2 norm, which is often denoted simply by $\|f\|$. This norm is related to the inner product:

$$\|f\|_2 = \langle f, f\rangle^{1/2}. \tag{D.52}$$

The space consisting of the set of functions whose L_2 norms over $\mathbb{R}$ exist together with this norm, that is, $\mathcal{L}^2(\mathbb{R})$, is a Hilbert space.

Another common L_p function norm is the L_∞ norm, especially as a measure of the difference between two functions. This norm, which is called the *Chebyshev norm* or the *uniform norm*, is the limit of equation (D.51) as $p \to \infty$. This norm has the simpler relationship

$$\|f\|_\infty = \sup|f(x)w(x)|. \tag{D.53}$$

To emphasize the measure of the weighting function, the notation $\|f\|_\mu$ is sometimes used. (The ambiguity of the possible subscripts on $\|\cdot\|$ is usually resolved by the context.) For functions over finite domains, the weighting function is most often the identity.

Another type of function norm, called the *total variation*, is a measure of the amount of variability of the function.

****

A *normal function* is one whose norm is 1. (Analogously to the terminology for vectors, we also call a normal function a *normalized function*.) Although this term can be used with respect to any norm, it is generally reserved for the L_2 norm (that is, the norm arising from the inner product). A function whose integral (over a relevant range, usually $\mathbb{R}$) is 1 is also called a normal function. (Although this latter definition is similar to the standard one, the latter is broader because it may include functions that are not square-integrable.) Density and weight functions are often normalized (that is, scaled to be normal).

Metrics in Function Spaces

Statistical properties such as bias and consistency are defined in terms of the difference of the estimator and what is being estimated. For an estimator of a function, first we must consider some ways of measuring this difference. These are general measures for functions and are not dependent on the distribution of a random variable. How well one function approximates another function is usually measured by a norm of the difference in the functions over the relevant range.

The most common measure of the difference between two functions, $g(x)$ and $f(x)$, is a norm of the function

$$e(x) = g(x) - f(x).$$

When one function is an estimate or approximation of the other function, we may call this difference the "error".

If g approximates f, $\|g-f\|_\infty$ is likely to be the norm of interest. This is the norm most often used in numerical analysis when the objective is interpolation or quadrature. In problems with noisy data, or when g may be very different from f, $\|g - f\|_2$ may be the more appropriate norm. This is the norm most often used in estimating probability density functions.

Distances between Functions

For comparing two functions g and f we can use a metric based on a norm of their difference, $\|g - f\|$. We often prefer to use a pseudometric, which is the same as a matric except that $\rho(g, f) = 0$ if and only if $g = f$ a.e. (We usually just use this interpretation and call it a metric, however.)

For functions, the norm is often taken as the L_∞ norm. If P and Q are CDFs, $\|P - Q\|_\infty$ is called the Kolmogorov distance.

Sometimes the difference in two functions is defined asymmetrically. A general class of divergence measures for comparing CDFs was introduced independently by Ali and Silvey (1966) and Csiszár (1967). The measure is

based on a convex function ϕ of the a term similar to the "odds". If P is absolutely continuous with respect to Q and ϕ is a convex function,

$$d(P,Q) = \int_{\mathbb{R}} \phi\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}Q, \tag{D.54}$$

if it exists, is called the *ϕ-divergence* from Q to P. The ϕ-divergence is also called the *f-divergence*.

The ϕ-divergence is in general not a metric because it is not symmetric. One function is taken as the base from which the other function is measured. The expression often has a more familiar form if both P and Q are dominated by Lebesgue measure and we write $p = \mathrm{d}P$ and $q = \mathrm{d}Q$.

A specific instance of ϕ-divergence is the *Kullback-Leibler measure*,

$$\int_{\mathbb{R}} p(x) \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x. \tag{D.55}$$

The Kullback-Leibler measure is not a metric.

Another specific instance of a function of ϕ-divergence, which happens to be a metric, is the *Hellinger distance* given by

$$\left(\int_{\mathbb{R}} \left(q^{1/r}(x) - p^{1/r}(x)\right)^r \mathrm{d}x\right)^{1/r}. \tag{D.56}$$

This turns out to be the $1/r$ power of a ϕ-divergence. The most common case has $r = 2$, and in this case the Hellinger distance is also called the *Matusita distance*. The Matusita distance is the square root of a ϕ-divergence with $\phi(t) = (\sqrt{t} - 1)^2$.

Various forms of ϕ-divergence are used in goodness-of-fit analyses. The Pearson chi-squared discrepancy measure, for example, has $\phi(t) = (t-1)^2$:

$$\int_{\mathbb{R}} \frac{(q(x) - p(x))^2}{q(x)} \mathrm{d}x. \tag{D.57}$$

See the discussion beginning on page 282 for other applications in which two probability distributions are compared.

Basis Sets in Function Spaces

If each function in a linear space can be expressed as a linear combination of the functions in a set G, then G is said to be a *generating set*, a *spanning set*, or a *basis set* for the linear space. (These three terms are synonymous.) The basis sets for finite-dimensional vector spaces are finite; for most function spaces of interest, the basis sets are infinite.

A set of functions $\{q_k\}$ is *orthogonal over the domain D with respect to the nonnegative weight function $w(x)$* if the inner product with respect to $w(x)$ of q_k and q_l, $\langle q_k, q_l \rangle$, is 0 if $k \neq l$; that is,

$$\int_D q_k(x)\bar{q}_l(x)w(x)\mathrm{d}x = 0 \quad k \neq l. \tag{D.58}$$

If, in addition,

$$\int_D q_k(x)\bar{q}_k(x)w(x)\mathrm{d}x = 1,$$

the functions are called *orthonormal*.

In the following, we will be concerned with real functions of real arguments, so we can take $\bar{q}_k(x) = q_k(x)$.

The weight function can also be incorporated into the individual functions to form a different set,

$$\tilde{q}_k(x) = q_k(x)w^{1/2}(x).$$

This set of functions also spans the same function space and is orthogonal over D with respect to a constant weight function.

Basis sets consisting of orthonormal functions are generally easier to work with and can be formed from any basis set. Given two nonnull, linearly independent functions, q_1 and q_2, two orthonormal vectors, $\tilde{q}_1$ and $\tilde{q}_2$, that span the same space can be formed as

$$\begin{aligned}
\tilde{q}_1(\cdot) &= \frac{1}{\|q_1\|}\, q_1(\cdot), \\
\tilde{q}_2(\cdot) &= \frac{1}{\|q_2 - \langle \tilde{q}_1, q_2\rangle \tilde{q}_1\|}\, \big(q_2(\cdot) - \langle \tilde{q}_1, q_2\rangle \tilde{q}_1(\cdot)\big).
\end{aligned} \tag{D.59}$$

These are the Gram-Schmidt function transformations. They can easily be extended to more than two functions to form a set of orthonormal functions from any set of linearly independent functions.

Series Expansions in Basis Functions

Our objective is to represent a function of interest, $f(x)$, over some domain $D \subset \mathbb{R}$, as a linear combination of "simpler" functions, $q_0(x), q_1(x), \ldots$:

$$f(x) = \sum_{k=0}^{\infty} c_k q_k(x). \tag{D.60}$$

There are various ways of constructing the q_k functions. If they are developed through a linear operator on a function space, they are called *eigenfunctions*, and the corresponding c_k are called eigenvalues.

We choose a set $\{q_k\}$ that spans some class of functions over the given domain D. A set of orthogonal basis functions is often the best choice because they have nice properties that facilitate computations and a large body of theory about their properties is available.

If the function to be estimated, $f(x)$, is continuous and integrable over a domain D, the orthonormality property allows us to determine the coefficients c_k in the expansion (D.60):

$$c_k = \langle f, q_k \rangle. \tag{D.61}$$

The coefficients $\{c_k\}$ are called the *Fourier coefficients* of f with respect to the orthonormal functions $\{q_k\}$.

In applications, we *approximate* the function using a truncated orthogonal series. The error due to finite truncation at j terms of the infinite series is the residual function $f - \sum_{k=1}^{j} c_k f_k$. The *mean squared error* over the domain D is the scaled, squared L_2 norm of the residual,

$$\frac{1}{d} \left\| f - \sum_{k=0}^{j} c_k q_k \right\|^2, \tag{D.62}$$

where d is some measure of the domain D. (If the domain is the interval $[a, b]$, for example, one choice is $d = b - a$.)

A very important property of Fourier coefficients is that they yield the minimum mean squared error for a given set of basis functions $\{q_i\}$; that is, for any other constants, $\{a_i\}$, and any j,

$$\left\| f - \sum_{k=0}^{j} c_k q_k \right\|^2 \leq \left\| f - \sum_{k=0}^{j} a_k q_k \right\|^2. \tag{D.63}$$

In applications of statistical data analysis, after forming the approximation, we then *estimate* the coefficients from equation (D.61) by identifying an appropriate probability density that is a factor of the function of interest, f. (Note again the difference in "approximation" and "estimation".) Expected values can be estimated using observed or simulated values of the random variable and the approximation of the probability density function.

The basis functions are generally chosen to be easy to use in computations. Common examples include the Fourier trigonometric functions $\sin(kt)$ and $\cos(kt)$ for $k = 1, 2, \ldots$, orthogonal polynomials such as Legendre, Hermite, and so on, splines, and wavelets.

Orthogonal Polynomials

The most useful type of basis function depends on the nature of the function being estimated. The orthogonal polynomials are useful for a very wide range of functions. Orthogonal polynomials of real variables are their own complex conjugates. It is clear that for the k^{th} polynomial in the orthogonal sequence, we can choose an a_k that does not involve x, such that

$$q_k(x) - a_k x q_{k-1}(x)$$

is a polynomial of degree $k-1$.

Because any polynomial of degree $k-1$ can be represented by a linear combination of the first k members of any sequence of orthogonal polynomials, we can write

$$q_k(x) - a_k x q_{k-1}(x) = \sum_{i=0}^{k-1} c_i q_i(x).$$

Because of orthogonality, all c_i for $i < k-2$ must be 0. Therefore, collecting terms, we have, for some constants a_k, b_k, and c_k, the three-term recursion that applies to any sequence of orthogonal polynomials:

$$q_k(x) = (a_k x + b_k) q_{k-1}(x) - c_k q_{k-2}(x), \quad \text{for } k = 2, 3, \ldots. \tag{D.64}$$

This recursion formula is often used in computing orthogonal polynomials. The coefficients in this recursion formula depend on the specific sequence of orthogonal polynomials, of course.

This three-term recursion formula can also be used to develop a formula for the sum of products of orthogonal polynomials $q_i(x)$ and $q_i(y)$:

$$\sum_{i=0}^{k} q_i(x) q_i(y) = \frac{1}{a_{k+1}} \frac{q_{k+1}(x) q_k(y) - q_k(x) q_{k+1}(y)}{x - y}. \tag{D.65}$$

This expression, which is called the Christoffel-Darboux formula, is useful in evaluating the product of arbitrary functions that have been approximated by finite series of orthogonal polynomials.

There are several widely used complete systems of univariate orthogonal polynomials. The different systems are characterized by the one-dimensional intervals over which they are defined and by their weight functions. The Legendre, Chebyshev, and Jacobi polynomials are defined over $[-1, 1]$ and hence can be scaled into any finite interval. The weight function of the Jacobi polynomials is more general, so a finite sequence of them may fit a given function better, but the Legendre and Chebyshev polynomials are simpler and so are often used. The Laguerre polynomials are defined over the half line $[0, \infty)$, and the Hermite polynomials are defined over the reals, $(-\infty, \infty)$.

Any of these systems of polynomials can be developed easily by beginning with the basis set $1, x, x^2, \ldots$ and orthogonalizing them by use of equations (D.59) and their extensions.

Table D.1 summarizes the ranges and weight functions for these standard orthogonal polynomials.

The *Legendre polynomials* have a constant weight function and are defined over the interval $[-1, 1]$. The first few (unnormalized) Legendre polynomials are

$$\begin{array}{ll} P_0(t) = 1 & P_1(t) = t \\ P_2(t) = (3t^2 - 1)/2 & P_3(t) = (5t^3 - 3t)/2 \\ P_4(t) = (35t^4 - 30t^2 + 3)/8 & P_5(t) = (63t^5 - 70t^3 + 15t)/8 \end{array} \tag{D.66}$$

Table D.1. Orthogonal Polynomials

Polynomial Series	Range	Weight Function
Legendre	$[-1,\ 1]$	1 (uniform)
Chebyshev	$[-1,\ 1]$	$(1-x^2)^{1/2}$ (symmetric beta)
Jacobi	$[-1,\ 1]$	$(1-x)^{\alpha}(1+x)^{\beta}$ (beta)
Laguerre	$[0,\ \infty)$	$x^{\alpha-1}\mathrm{e}^{-x}$ (gamma)
Hermite	$(-\infty,\ \infty)$	$\mathrm{e}^{-x^2/2}$ (normal)

The normalizing constant for the k^{th} Legendre polynomial is determined by noting

$$\int_{-1}^{1}(P_k(t))^2\mathrm{d}t = \frac{2}{2k+1}.$$

The recurrence formula for the Legendre polynomials is

$$P_k(t) = \frac{2k-1}{k}tP_{k-1}(t) - \frac{k-1}{k}P_{k-2}(t). \tag{D.67}$$

The *Hermite polynomials* are orthogonal with respect to a Gaussian, or standard normal, weight function. A series using these Hermite polynomials is often called a Gram-Charlier series. See Section 1.4.

The first few Hermite polynomials are

$$\begin{array}{ll} H_0^{\mathrm{e}}(t) = 1 & H_1^{\mathrm{e}}(t) = t \\ H_2^{\mathrm{e}}(t) = t^2 - 1 & H_3^{\mathrm{e}}(t) = t^3 - 3t \\ H_4^{\mathrm{e}}(t) = t^4 - 6t^2 + 3 & H_5^{\mathrm{e}}(t) = t^5 - 10t^3 + 15t \end{array} \tag{D.68}$$

These are not the standard Hermite polynomials, but they are the ones most commonly used by statisticians because the weight function is proportional to the normal density.

The recurrence formula for the Hermite polynomials is

$$H_k^{\mathrm{e}}(t) = tH_{k-1}^{\mathrm{e}}(t) - (k-1)H_{k-2}^{\mathrm{e}}(t). \tag{D.69}$$

Multivariate Orthogonal Polynomials

Multivariate orthogonal polynomials can be formed easily as tensor products of univariate orthogonal polynomials. The tensor product of the functions $f(x)$ over D_x and $g(y)$ over D_y is a function of the arguments x and y over $D_x \times D_y$:

$$h(x,y) = f(x)g(y).$$

If $\{q_{1,k}(x_1)\}$ and $\{q_{2,l}(x_2)\}$ are sequences of univariate orthogonal polynomials, a sequence of bivariate orthogonal polynomials can be formed as

$$q_{kl}(x_1, x_2) = q_{1,k}(x_1)q_{2,l}(x_2). \tag{D.70}$$

These polynomials are orthogonal in the same sense as in equation (D.58), where the integration is over the two-dimensional domain. Similarly as in equation (D.60), a bivariate function can be expressed as

$$f(x_1, x_2) = \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} c_{kl} q_{kl}(x_1, x_2), \tag{D.71}$$

with the coefficients being determined by integrating over both dimensions.

Although obviously such product polynomials, or radial polynomials, would emphasize features along coordinate axes, they can nevertheless be useful for representing general multivariate functions. Often, it is useful to apply a rotation of the coordinate axes.

The weight functions, such as those for the Jacobi polynomials, that have various shapes controlled by parameters can also often be used in a mixture model of the function of interest. The weight function for the Hermite polynomials can be generalized by a linear transformation (resulting in a normal weight with mean μ and variance σ^2), and the function of interest may be represented as a mixture of general normals.

Function Decomposition and Estimation of the Coefficients in an Orthogonal Expansion

We first decompose the function of interest to have a factor that is a probability density function, p,

$$f(x) = g(x)p(x). \tag{D.72}$$

We have

$$\begin{aligned} c_k &= \langle f, q_k \rangle \\ &= \int_D q_k(x)g(x)p(x)\mathrm{d}x \\ &= \mathrm{E}(q_k(X)g(X)), \end{aligned} \tag{D.73}$$

where X is a random variable whose probability density function is p.

If we can obtain a random sample, $x_1, \ldots, x_n$, from the distribution with density p, the c_k can be unbiasedly estimated by

$$\widehat{c}_k = \frac{1}{n} \sum_{i=1}^{n} q_k(x_i)g(x_i).$$

The series estimator of the function for all x therefore is

$$\widehat{f}(x) = \frac{1}{n}\sum_{k=0}^{j}\sum_{i=1}^{n} q_k(x_i)g(x_i)q_k(x) \tag{D.74}$$

for some truncation point j.

The random sample, $x_1, \ldots, x_n$, may be an observed dataset, or it may be the output of a random number generator.

Distribution Function Spaces

In probability and statistics, one of the most important kinds of function is a cumulative distribution function, or CDF, defined on page 5 both in terms of a probability distribution and in terms of four characterizing properties.

A set of CDFs cannot constitute a linear space, because of the restrictions on the functions. Instead, we will define a *distribution function space* that has similar properties. If $\mathcal{P}$ is a set of CDFs such that for any $w \in [0,1]$ and $P_1, P_2 \in \mathcal{P}$, $(1-w)P_1 + wP_2 \in \mathcal{P}$, then $\mathcal{P}$ is a distribution function space.

The ϵ-mixture distribution defined on page 285 is a simple example of a distribution function space. In that space, one of the CDFs is degenerate.

Important distribution function spaces are those consisting of CDFs P such that for given $p \geq 1$

$$\int \|t\|^p \mathrm{d}P < \infty. \tag{D.75}$$

Such a distribution function space is denoted by $\mathcal{P}^p$. (Constrast this with the $\mathcal{L}^p$ space.) It is clear that $\mathcal{P}^{p_1} \subset \mathcal{P}^{p_2}$ if $p_1 \geq p_2$.

Spaces of distribution functions are useful in robustness studies, but most of the interesting families of probability distributions as discussed in Section 1.7 do not generate distribution function spaces.

Functionals

Functionals are functions whose arguments are functions. The value of a functional may be any kind of object, a real number or another function, for example. The domain of a functional is a set of functions.

If $\mathcal{F}$ is a linear space of functions, that is, if $\mathcal{F}$ is such that $f \in \mathcal{F}$ and $g \in \mathcal{F}$ implies $(af + g) \in \mathcal{F}$ for any real a, then the functional Υ defined on $\mathcal{F}$ is said to be *linear* if $\Upsilon(af + g) = a\Upsilon(f) + \Upsilon(g)$.

A similar expression defines linearity of a functional over a distribution function space $\mathcal{P}$: Υ defined on $\mathcal{P}$ is linear if $\Upsilon((1-w)P_1 + wP_2) = (1-w)\Upsilon(P_1) + w\Upsilon(P_2)$ for $w \in [0,1]$ and $P_1, P_2 \in \mathcal{P}$.

Functionals of CDFs have important uses in statistics as measures of the differences between two distributions or to define distributional measures of interest. A functional applied to a ECDF is a plug-in estimator of the distributional measure defined by the same functional applied to the corresponding CDF.

Derivatives of Functionals

For the case in which the arguments are functions, the cardinality of the possible perturbations is greater than that of the continuum. We can be precise in discussions of continuity and differentiability of a functional Υ at a point (function) F in a domain $\mathcal{F}$ by defining another set $\mathcal{D}$ consisting of difference functions over $\mathcal{F}$; that is the set the functions $D = F_1 - F_2$ for $F_1, F_2 \in \mathcal{F}$.

The concept of differentiability for functionals is necessarily more complicated than for functions over real domains. For a functional Υ over the domain $\mathcal{F}$, we define three levels of differentiability at the function $F \in \mathcal{F}$. All definitions are in terms of a domain $\mathcal{D}$ of difference functions over $\mathcal{F}$, and a linear functional Λ_F defined over $\mathcal{D}$ in a neighborhood of F. The first type of derivative is very general. The other two types depend on a metric ρ on $\mathcal{F} \times \mathcal{F}$ induced by a norm $\|\cdot\|$ on $\mathcal{F}$.

Definition D.2.13 (Gâteaux differentiable.) *Υ is Gâteaux differentiable at F iff there exists a linear functional $\Lambda_F(D)$ over $\mathcal{D}$ such that for $t \in \mathbb{R}$ for which $F + tD \in \mathcal{F}$,*

$$\lim_{t\to 0}\left(\frac{\Upsilon(F+tD)-\Upsilon(F)}{t} - \Lambda_F(D)\right) = 0. \tag{D.76}$$

Definition D.2.14 (ρ-Hadamard differentiable.) *For a metric ρ induced by a norm, Υ is ρ-Hadamard differentiable at F iff there exists a linear functional $\Lambda_F(D)$ over $\mathcal{D}$ such that for any sequence $t_j \to 0 \in \mathbb{R}$ and sequence $D_j \in \mathcal{D}$ such that $\rho(D_j, D) \to 0$ and $F + t_j D_j \in \mathcal{F}$,*

$$\lim_{j\to\infty}\left(\frac{\Upsilon(F+t_jD_j)-\Upsilon(F)}{t_j} - \Lambda_F(D_j)\right) = 0. \tag{D.77}$$

Definition D.2.15 (ρ-Fréchet differentiable.) *Υ is ρ-Fréchet differentiable at F iff there exists a linear functional $\Lambda(D)$ over $\mathcal{D}$ such that for any sequence $F_j \in \mathcal{F}$ for which $\rho(F_j, F) \to 0$,*

$$\lim_{j\to\infty}\left(\frac{\Upsilon(F_j)-\Upsilon(F)-\Lambda_F(F_j-F)}{\rho(F_j,F)}\right) = 0. \tag{D.78}$$

The linear functional Λ_F in each case is called, respectively, the [*Gâteaux* | *ρ-Hadamard* | *ρ-Fréchet*] *differential* of Υ at F.

Notes and Additional References for Section D.2

Measure theory is the most important element of analysis. There are many classic and standard texts on the subject, and it would be difficult to select "best" ones. Many treat measure theory in the context of probability theory, and some of those are listed in the general bibliography beginning on page 477. I will just list a few more that I have found useful. Although they are relatively old, all, except possibly Hewitt and Stromberg (1965) (from which I first began learning real analysis), are still readily available.

Additional References

Doob, J. L. (1994), *Measure Theory*, Springer-Verlag, New York.

Hewitt, Edwin, and Karl Stromberg (1965), *Real and Abstract Analysis*, Springer-Verlag, Berlin. (A second edition was published in 1969.)

Kolmogorov, A. N., and S. V. Fomin (1954 and 1960, translated from the Russian), *Elements of the Theory of Functions and Functional Analysis*, in two volumes, Gaylock Press, Rochester, NY. Reprinted in one volume (1999) by Dover Publications, Inc., Mineola, NY.

Royden, H. L. (1988), *Real Analysis*, third edition, MacMillan, New York.

D.3 Stochastic Calculus

The other sections in this appendix generally cover prerequisite material for the rest of the book. This section, on the other hand, depends on some of the material in Chapter 1, and is closely interrelated with the material in Section 1.6.

D.3.1 Continuous Time Stochastic Processes

We consider a stochastic process $\{B_t\}$ in which we generally associate the index t with time. We often write $\{B(t)\}$ in place of $\{B_t\}$, but for all practical purposes, the notation is equivalent.

In a very important class of stochastic processes, the differences between the values at two time points have normal distributions and the difference between two points is independent of the difference between two nonoverlapping points.

Wiener Processes

Suppose in the sequence $B_0, B_1, \ldots$, the distribution of $B_{t+1} - B_t$ is normal with mean 0 and standard deviation 1. In this case, the distribution of $B_{t+2} - B_t$ is normal with mean 0 and standard deviation $\sqrt{2}$, and the distribution of $B_{t+0.5} - B_t$ is normal with mean 0 and standard deviation $\sqrt{0.5}$. More generally, the distribution of the change ΔB in time Δt has a standard deviation of $\sqrt{\Delta t}$

This kind of process with the Markovian property and with a normal distribution of the changes leads to a Brownian motion or a Wiener process.

Consider a process of changes ΔB characterized by two properties:

- The change ΔB during a small period of time Δt is given by

$$\Delta B = Z\sqrt{\Delta t},$$

  where Z is a random variable with a $\mathrm{N}(0,1)$ distribution.
- The values of ΔB for any two short intervals of time Δt are independent.

Now, consider N time periods, and let $T = N\Delta t$. We have

$$B(T) - B(0) = \sum_{i=1}^{N} Z_i\sqrt{\Delta t}.$$

The fact that we have $\sqrt{\Delta t}$ in this equation has important implications that we will return to later.

As in ordinary calculus, we consider $\Delta B/\Delta t$ and take the limit as $\Delta t \to 0$, which we call $\mathrm{d}B/\mathrm{d}t$, and we have the differential equation

$$\mathrm{d}B = Z\mathrm{d}t.$$

A random variable formed as $\mathrm{d}B$ above is called a *stochastic differential*.

A stochastic differential arising from a process of changes ΔB with the two properties above is called a *Wiener process* or a *Brownian motion*. In the following, we will generally use the phrase "Wiener process".

We can use the Wiener process to develop a *generalized Wiener process*:

$$\mathrm{d}S = \mu\mathrm{d}t + \sigma\mathrm{d}B,$$

where μ and σ are constants.

Properties of the Discrete Process Underlying the Wiener Process

With $\Delta B = Z\sqrt{\Delta t}$ and $Z \sim \mathrm{N}(0,1)$, we immediately have

$$\begin{aligned}
\mathrm{E}(\Delta B) &= 0 \\
\mathrm{E}\left((\Delta B)^2\right) &= \mathrm{V}(\Delta B) + (\mathrm{E}(\Delta B))^2 \\
&= \Delta t \\
\mathrm{E}\left((\Delta B)^3\right) &= 0 \\
\mathrm{E}\left((\Delta B)^4\right) &= \mathrm{V}\left((\Delta B)^2\right) + \left(\mathrm{E}\left((\Delta B)^2\right)\right)^2 \\
&= 3(\Delta t)^2
\end{aligned}$$

Because of independence, for $\Delta_i B$ and $\Delta_j B$ representing changes in two nonoverlapping intervals of time,

$$\mathrm{E}((\Delta_i B)(\Delta_j B)) = \mathrm{cov}(\Delta_i B, \Delta_j B) = 0.$$

The Wiener process is a random variable; that is, it is a real-valued mapping from a sample space Ω. We sometimes use the notation $B(\omega)$ to emphasize this fact.

The Wiener process is a function in continuous time. We sometimes use the notation $B(t,\omega)$ to emphasize the time dependency.

Most of the time we drop the "ω". Also, sometimes we write B_t instead of $B(t)$.

All of these notations are equivalent.

There two additional properties of a Wiener process or Brownian motion that we need in order to have a useful model. We need an initial value, and we need it to be continuous in time.

Because the Wiener process is a random variable, the values it takes are those of a function at some point in the underlying sample space, Ω. Therefore, when we speak of $B(t)$ at some t, we must speak in terms of probabilities of values or ranges of values.

When we speak of a particular value of $B(t)$, unless we specify a specific point $\omega_0 \in \Omega$, the most we can say is that the values occurs almost surely.

- We assume $B(t) = 0$ almost surely at $t = 0$.
- We assume $B(t)$ is almost surely continuous in t.

These two properties together with the limiting forms of the two properties given at the beginning define a Wiener process or Brownian motion.

(There is a theorem due to Kolmogorov that states that given the first three properties, there exists a "version" that is absolutely continuous in t.)

From the definition, we can see immediately that

- the Wiener process is Markovian
- the Wiener process is a martingale.

Generalized Wiener Processes

A Wiener process or Brownian motion is a model for changes. It models diffusion.

If the process drifts over time (in a constant manner), we can add a term for the drift, $a\mathrm{d}t$.

More generally, a model for the state of a process that has both a Brownian diffusion and a drift is a generalized Wiener process:

$$\mathrm{d}S = a\mathrm{d}t + b\mathrm{d}B,$$

where a and b are constants. A generalized Wiener process is a type of a more general "drift-diffusion process".

While the expected value of the Wiener process at any time is 0, the expected value of the state S is not necessarily 0. Likewise, the variance is affected by b. Both the expected value and the variance of S are functions of time.

One of the most interesting properties of a Wiener process is that its first variation is infinite. It is infinitely "wiggly". We can see this by generating normal processes over varying length time intervals, as in Figure D.1.

Variation of Functionals

The variation of a functional is a measure of its rate of change. It is similar in concept to an integral of a derivative of a function.

For studying variation, we will be interested only in functions from the interval $[0, T]$ to $\mathbb{R}$.

To define the variation of a general function $f : [0, T] \mapsto \mathbb{R}$, we form N intervals $0 = t_0 \leq t_1 \leq \cdots \leq t_N = T$. The intervals are not necessarily of equal length, so we define Δ as the maximum lenght of any interval; that is,

$$\Delta = \max(t_i - t_{i-1}).$$

Now, we denote the p^{th} variation of f as $\mathrm{V}^p(f)$ and define it as

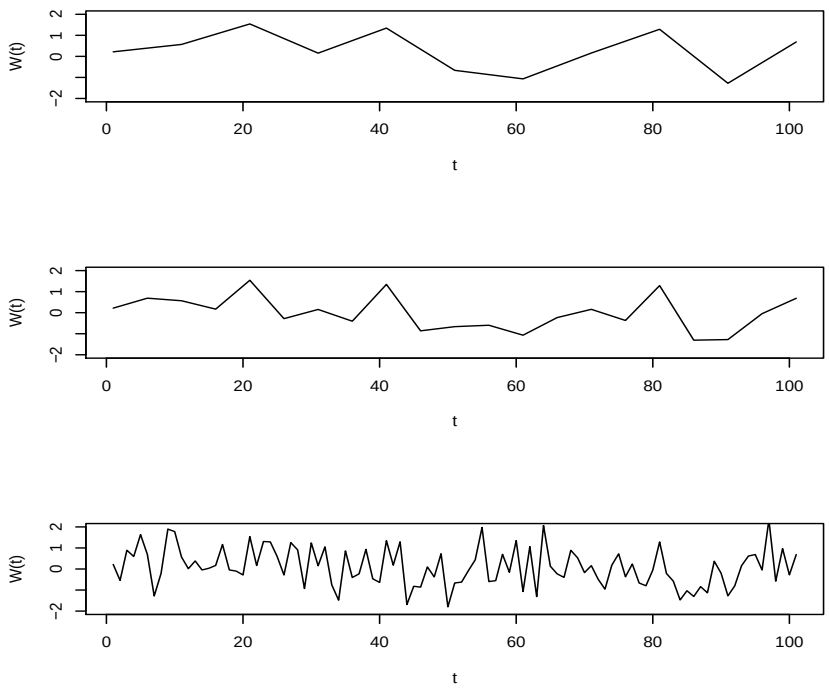


Fig. D.1. A Wiener Process Observed at Varying Length Intervals

$$\mathrm{V}^p(f) = \lim_{\Delta \to 0} \sum_{i=1}^{N} |f(t_i) - f(t_{i-1})|^p.$$

(Notice that $\Delta \to 0$ implies $N \to \infty$.)

With equal intervals, Δt, for the first variation, we can write

$$\begin{aligned} \mathrm{V}^1(f) &= \lim_{\Delta t \to 0} \sum_{i=1}^{N} |f(t_i) - f(t_{i-1})| \\ &= \lim_{N \to \infty} \sum_{i=0}^{N-1} \Delta t \frac{|f(t_i + \Delta t) - f(t_i)|}{\Delta t}, \end{aligned}$$

from which we can see that for a differentiable function $f : [0, T] \mapsto \mathbb{R}$,

$$\mathrm{V}^1(f) = \int_0^T \left| \frac{\mathrm{d}f}{\mathrm{d}t} \right| \mathrm{d}t.$$

The notation $FV(f)$, or more properly, $\mathrm{FV}(f)$, is sometimes used instead of $\mathrm{V}^1(f)$.

Again, with equal intervals, Δt, for the second variation, we can write

$$\begin{aligned}\mathrm{V}^2(f) &= \lim_{\Delta t\to 0}\sum_{i=1}^{N}(f(t_i)-f(t_{i-1}))^2 \\ &= \lim_{\Delta t\to 0}\Delta t \lim_{N\to\infty}\sum_{i=0}^{N-1}\Delta t\left(\frac{|f(t_i+\Delta t)-f(t_i)|}{\Delta t}\right)^2.\end{aligned}$$

For a differentiable function $f : [0,T] \mapsto \mathbb{R}$, we have

$$\mathrm{V}^2(f) = \lim_{\Delta t\to 0}\Delta t\int_0^T \left|\frac{\mathrm{d}f}{\mathrm{d}t}\right|^2 \mathrm{d}t.$$

The integrand is bounded, therefore this limit is 0, and we conclude that the second variation of a differentiable function is 0.

If X is a stochastic functional, then $\mathrm{V}^p(X)$ is also stochastic. If it converges to a deterministic quantity, the nature of the convergence must be considered.

First and Second Variation of a Wiener Process

Two important properties of a Wiener process on $[0,T]$ are

- $\mathrm{V}^2(B) = T$ a.s., which as we have seen, implies that $B(t)$ is not differentiable.
- $\mathrm{V}^1(B) = \infty$ a.s.

Notice that because B is a random variable we must temper our statement with a phrase about the probability or expected value.

We now prove these for the quadratic mean instead of a.s. We start with the first one, because it will imply the second one. Let

$$\begin{aligned}X_N &= \sum_{n=0}^{N-1}(B(t_{n+1})-B(t_n))^2 \\ &= \sum_{n=0}^{N-1}(\Delta_n B)^2 \quad \text{note notation}\end{aligned}$$

We want to show

$$\mathrm{E}\left((X_N-T)^2\right)\to 0 \quad \text{as } |\Delta t|\to 0.$$

Now,

$$\mathrm{E}\left((X_N-T)^2\right) = \mathrm{E}\left(X_N^2\right) - 2T\mathrm{E}(X_N) + T^2 = \mathrm{E}\left(X_N^2\right) - T^2.$$

So now we want to show $\mathrm{E}\left(X_N^2\right) = T^2$.

$$\begin{aligned}
\mathrm{E}\left(X_N^2\right) &= \mathrm{E}\left(\sum_{i=0}^{N-1}(\Delta_i B)^2 \sum_{j=0}^{N-1}(\Delta_j B)^2\right) \\
&= \mathrm{E}\left(\sum_{i=0}^{N-1}(\Delta_i B)^4\right) + \mathrm{E}\left(\sum_{i\neq j}(\Delta_i B)^2(\Delta_j B)^2\right) \\
&= \sum_{i=0}^{N-1}(\Delta_i t)^2 + \sum_{i\neq j}(\Delta_i t)(\Delta_j t).
\end{aligned}$$

Because $|\Delta t| \to 0$ (or, if we allow different size intervals, $\sup|\Delta_i t| \to 0$), we have

$$\sum_{i=0}^{N-1}(\Delta_i t)^2 \to 0.$$

So the first term goes to 0; now consider $\sum_{i\neq j}(\Delta_i t)(\Delta_j t)$.

$$\begin{aligned}
\sum_{i\neq j}(\Delta_i t)(\Delta_j t) &= \sum_{i=0}^{N-1}(\Delta_i t)\left(\sum_{j=0}^{i-1}(\Delta_j t) + \sum_{j=i+1}^{N-1}(\Delta_j t)\right) \\
&= \sum_{i=0}^{N-1}(\Delta_i t)(T - \Delta_i t) \\
&= T\sum_{i=0}^{N-1}(\Delta_i t) - \sum_{i=0}^{N-1}(\Delta_i t)^2 \\
&= T^2 - 0.
\end{aligned}$$

So now we have $\mathrm{E}\left((X_N - T)^2\right) \to 0$, or $X_N \to_{\mathrm{L}_2} T$ as $|\Delta t| \to 0$; that is, $\mathrm{V}^2(B) = T$ in quadratic mean, or in L_2 norm.

(I just realized that I had stated a.s. convergence, and I proved L_2 *convergence. One does not imply the other, but a.s. is also true in this case.)*

Now, although we have already seen that since the second variation is nonzero, B cannot be differentiable.

But also because of the continuity of B in t, it is easy to see that the first variation diverges if the second variation converges to a finite value. This is because

$$\sum_{n=0}^{N-1}(B(t_{n+1}) - B(t_n))^2 \leq \sup|B(t_{n+1}) - B(t_n)| \sum_{n=0}^{N-1}|B(t_{n+1}) - B(t_n)|$$

In the limit the term on the left is $T > 0$, and the term on the right is 0 times $\mathrm{V}^1(B)$; therefore $\mathrm{V}^1(B) = \infty$.

Properties of Stochastic Differentials

Although B and $\mathrm{d}B$ are random variables, the product $\mathrm{d}B\mathrm{d}B$ is deterministic.

We can see this by considering the stochastic process $(\Delta B)^2$. We have seen that $\mathrm{V}\left((\Delta B)^2\right) = 2(\Delta t)^2$, so the variance of this process is $2(\Delta t)^2$; that is, as $\Delta t \to 0$, the variance of this process goes to 0 faster, as $(\Delta t)^2$.

Also, as we have seen, $\mathrm{E}\left((\Delta B)^2\right) = \Delta t$, and so $(\Delta B)^2$ goes to Δt at the same rate as $\Delta t \to 0$. That is,

$$(\Delta B)(\Delta B) \to_{\text{a.s.}} \Delta t \quad \text{as } \Delta t \to 0.$$

The convergence of $(\Delta B)(\Delta B)$ to Δt as $\Delta t \to 0$ yields

$$\mathrm{d}B\mathrm{d}B = \mathrm{d}t.$$

(This equality is almost sure.) But $\mathrm{d}t$ is a deterministic quantity.

This is one of the most remarkable facts about a Wiener process.

Multidimensional Wiener Processes

If we have two Wiener processes B_1 and B_2, with $\mathrm{V}(\mathrm{d}B_1) = \mathrm{V}(\mathrm{d}B_2) = \mathrm{d}t$ and $\mathrm{cov}(\mathrm{d}B_1, \mathrm{d}B_2) = \rho\mathrm{d}t$ (that is, $\mathrm{corr}(\mathrm{d}B_1, \mathrm{d}B_2) = \rho$), then by a similar argument as before, we have $\mathrm{d}B_1\mathrm{d}B_2 = \rho\mathrm{d}t$, almost surely.

Again, this is deterministic.

The results of course extend to any vector of Wiener processes $(B_1, \ldots, B_d)$.

If $(B_1, \ldots, B_d)$ arise from

$$\Delta B_i = X_i\sqrt{\Delta t},$$

where the vector of Xs has a multivariate normal distribution with mean 0 and variance-covariance matrix Σ, then the variance-covariance matrix of $(\mathrm{d}B_1, \ldots, \mathrm{d}B_d)$ is $\Sigma\mathrm{d}t$, which is deterministic.

Starting with $(Z_1, \ldots, Z_d)$ i.i.d. $\mathrm{N}(0, 1)$ and forming the Wiener processes $B = (B_1, \ldots, B_d)$ beginning with

$$\Delta B_i = Z_i\sqrt{\Delta t},$$

we can form a vector of Wiener processes $B = (B_1, \ldots, B_d)$ with variance-covariance matrix $\Sigma\mathrm{d}t$ for $\mathrm{d}B = (\mathrm{d}B_1, \ldots, \mathrm{d}B_d)$ by the transformation

$$B = \Sigma^{1/2}B,$$

or equivalently by

$$B = \Sigma_C B,$$

where Σ_C is a Cholesky factor of Σ, that is, $\Sigma_C^{\mathrm{T}}\Sigma_C = \Sigma$.

Recall, for a fixed matrix A,

$$\mathrm{V}(AY) = A^{\mathrm{T}}\mathrm{V}(Y)A,$$

so from above, for example,

$$\mathrm{V}(\mathrm{d}B) = \Sigma_C^{\mathrm{T}}\mathrm{V}(\mathrm{d}B)\Sigma_C = \Sigma_C^{\mathrm{T}}\mathrm{diag}(\mathrm{d}t)\Sigma_C = \Sigma\mathrm{d}t.$$

D.3.2 Integration with Respect to Stochastic Differentials

Stochastic Integrals with Respect to Wiener Processes

The stochastic differentials such as $\mathrm{d}B$ naturally lead us to consider integration with respect to stochastic differentials, that is, stochastic integrals.

If B is a Wiener process on $[0, T]$, we may be interested in an integral of the form

$$\int_0^T g(Y(t), t)\mathrm{d}B,$$

where $Y(t)$ is a stochastic process (that is, Y is a random variable) and g is some function.

The problem with developing a definition of this integral following the same steps as in the definition of a Riemann integral, that is, as a limit of sequences of sums of areas of rectangles, is that because the sides of these rectangles, Y and $\mathrm{d}B$, are random variables, there are different kinds of convergence of a limit.

Also, the convergence of products of $Y(t)$ depend on where $Y(t)$ is evaluated.

The Ito Integral

We begin developing a definition of

$$\int_0^T g(Y(t), t)\mathrm{d}B,$$

by considering how the Riemann integral is defined in terms of the sums

$$I_n(t) = \sum_{i=0}^{n-1} g(Y(\tau_i), \tau_i)(B(t_{i+1}) - B(t_i)),$$

where $0 = t_0 \leq \tau_0 \leq t_1 \leq \tau_1 \leq \cdots \leq \tau_{n-1} \leq t_n = T$.

As in the Riemann case we will define the integral in terms of a limit as the mesh size goes to 0.

First, the existence depends on a finite expectation that is similar to a variance. We assume

$$\mathrm{E}\left(\int_0^T g(Y(t),t)\mathrm{d}t\right) < \infty.$$

The convergence must be qualified because the intervals are random variables; furthermore, (although it is not obvious!) the convergence depends on where τ_i is in the interval $[t_i, t_{i+1}]$.

The first choice in the definition of the Ito stochastic integral is to choose $\tau_i = t_i$. Other choices, such as choosing τ_i to be at the midpoint of the integral, lead to different types of stochastic integrals.

Next is the definition of the type of convergence. In the Ito stochastic integral, the convergence is in mean square, that is L_2 convergence.

With the two choices me have made, we take

$$I_n(t) = \sum_{i=0}^{n-1} g(Y(t_i), t_i)(B(t_{i+1}) - B(t_i)),$$

and the Ito integral is defined as

$$I(t) = \text{ms-lim}_{n\to\infty} I_n(t).$$

This integral based on a Wiener process is used throughout financial analysis.

Note that this integral is a random variable; in fact, it is a stochastic process. This is because of the fact that the differentials are from a Wiener process.

Also, because the integral is defined by a Wiener process, it is a martingale.

Ito Processes

An Ito process is a generalized Wiener process $\mathrm{d}X = a\mathrm{d}t + b\mathrm{d}B$, in which the parameters a and b are functions of the underlying variable X and of time t (of course, X is also a function of t).

The functions a and b must be measurable with respect to the filtration generated by $B(t)$ (that is, to the sequence of smallest σ-fields with respect to which $B(t)$ is measurable. (This is expressed more simply by saying a and b are adapted to the filtration generated by $B(t)$.)

The Ito process is of the form

$$\mathrm{d}X(t) = a(X(t),t)\mathrm{d}t + b(X(t),t)\mathrm{d}B.$$

The Ito integral (or any other stochastic integral) gives us a solution to this stochastic differential equation:

$$X(T) = X(0) + \int_0^T a(X(t),t)\mathrm{d}t + \int_0^T b(X(t),t)\mathrm{d}B(t).$$

(The differential in the first integral is deterministic although the integrand is stochastic. The second integral, however, is a stochastic integral. Other definitions of this integral would require modifications in the interpretation of properties of the Ito process.)

We are often interested in multidimensional Ito processes. Their second-order properties (variances and covariances) behave very similarly to those of Wiener processes, which we discussed earlier.

Geometric Brownian Motion

The Ito process would be much easier to work with if $\mu(\cdot)$ and $\sigma(\cdot)$ did not depend on the value of the state; that is, if we we use the model

$$\frac{\mathrm{d}S(t)}{S(t)} = \mu(t)\mathrm{d}t + \sigma(t)\mathrm{d}B,$$

where I have switched to "$S(t)$" because I'm thinking of the price of a stock.

The Ito process would be even easier to work with if $\mu(\cdot)$ and $\sigma(\cdot)$ were constant; that is, if we we just use the model

$$\frac{\mathrm{d}S(t)}{S(t)} = \mu\mathrm{d}t + \sigma\mathrm{d}B.$$

This model is called a geometric Brownian motion, and is widely used in modeling prices of various financial assets. ("Geometric" refers to series with multiplicative changes, as opposed to "arithmetic series" that have additive changes).

The geometric Brownian motion model is similar to other common statistical models:

$$\frac{\mathrm{d}S(t)}{S(t)} = \mu\mathrm{d}t + \sigma\mathrm{d}B(t)$$

or

$$\text{response} = \text{systematic component} + \text{random error}.$$

Without the stochastic component, the differential equation has the simple solution

$$S(t) = c\mathrm{e}^{\mu t},$$

from which we get the formula for continuous compounding for a rate μ.

Ito's Lemma

We can formalize the preceding discussion using Ito's lemma.

Suppose X follows an Ito process,

$$\mathrm{d}X(t) = a(X,t)\mathrm{d}t + b(X,t)\mathrm{d}B(t),$$

where $\mathrm{d}B$ is a Wiener process. Let G be an infinitely differentiable function of X and t. Then G follows the process

$$\mathrm{d}G(t) = \left(\frac{\partial G}{\partial X}a(X,t) + \frac{\partial G}{\partial t} + \frac{1}{2}\frac{\partial^2 G}{\partial X^2}b^2\right)\mathrm{d}t + \frac{\partial G}{\partial X}b(X,t)\mathrm{d}B(t). \tag{D.79}$$

Thus, Ito's lemma provides a formula that tells us that G also follows an Ito process.

The drift rate is

$$\frac{\partial G}{\partial X}a(X,t) + \frac{\partial G}{\partial t} + \frac{1}{2}\frac{\partial^2 G}{\partial X^2}b^2$$

and the volatility is

$$\frac{\partial G}{\partial X}b(X,t).$$

This allows us to work out expected values and standard deviations of G over time.

First, suppose that G is infinitely of X and an unrelated variable y, and consider a Taylor series expansion for ΔG:

$$\Delta G = \frac{\partial G}{\partial X}\Delta X + \frac{\partial G}{\partial y}\Delta y + \frac{1}{2}\left(\frac{\partial^2 G}{\partial X^2}(\Delta X)^2 + \frac{\partial^2 G}{\partial y^2}(\Delta y)^2 + 2\frac{\partial^2 G}{\partial X \partial y}\Delta X \Delta y\right) + \cdots \tag{D.80}$$

In the limit as ΔX and Δy tend to zero, this is the usual "total derivative"

$$\mathrm{d}G = \frac{\partial G}{\partial X}\mathrm{d}X + \frac{\partial G}{\partial y}\mathrm{d}y, \tag{D.81}$$

in which the terms in ΔX and Δy have dominated and effectively those in $(\Delta X)^2$ and $(\Delta y)^2$ and higher powers have disappeared.

Now consider an X that follows an Ito process,

$$\mathrm{d}X(t) = a(X,t)\mathrm{d}t + b(X,t)\mathrm{d}B(t),$$

or

$$\Delta X(t) = a(X,t)\Delta t + b(X,t)Z\sqrt{\Delta t}.$$

Now let G be a function of both X and t, and consider the analogue to equation (D.80). The factor $(\Delta X)^2$, which could be ignored in moving to equation (D.81), now contains a term with the factor Δt, which cannot be ignored. We have

$$(\Delta X(t))^2 = b(X,t)^2 Z^2 \Delta t + \text{ terms of higher degree in } \Delta t.$$

Consider the Taylor series expansion

$$\Delta G = \frac{\partial G}{\partial X}\Delta X + \frac{\partial G}{\partial t}\Delta t + \frac{1}{2}\left(\frac{\partial^2 G}{\partial X^2}(\Delta X)^2 + \frac{\partial^2 G}{\partial t^2}(\Delta t)^2 + 2\frac{\partial^2 G}{\partial X \partial t}\Delta X \Delta t\right) + \cdots \tag{D.82}$$

We have seen, under the assumptions of Brownian motion, $(\Delta X(t))^2$ or, equivalently, $Z^2\Delta t$, is nonstochastic; that is, we can treat $Z^2\Delta t$ as equal to its expected value as Δt tends to zero. Therefore, when we substitute for $\Delta X(t)$, and take limits in equation (D.82) as ΔX and Δt tend to zero, we get

$$\mathrm{d}G(t) = \frac{\partial G}{\partial X}\mathrm{d}X + \frac{\partial G}{\partial t}\mathrm{d}t + \frac{1}{2}\frac{\partial^2 G}{\partial X^2}b^2\mathrm{d}t \tag{D.83}$$

or, after substituting for $\mathrm{d}X$ and rearranging, we have Ito's formula

$$\mathrm{d}G(t) = \left(\frac{\partial G}{\partial X}a(X,t) + \frac{\partial G}{\partial t} + \frac{1}{2}\frac{\partial^2 G}{\partial X^2}b^2\right)\mathrm{d}t + \frac{\partial G}{\partial X}b(X,t)\mathrm{d}B(t).$$

Equation (D.83) is also called Ito's formula. Compare equation (D.83) with equation (D.81).

Multivariate Processes

There is a multivariate version of Ito's formula for a multivariate Ito process. The multivariate Ito process has the form

$$\mathrm{d}X(t) = a(X,t)\mathrm{d}t + B(X,t)\mathrm{d}B(t), \tag{D.84}$$

where $\mathrm{d}X(t)$, $a(X,t)$, and $\mathrm{d}B(t)$ are vectors and $B(X,t)$ is a matrix.

The elements of $\mathrm{d}B(t)$ can come from independent Wiener processes, or from correlated Wiener processes. I think it is easier to work with independent Wiener processes and incorporate any correlations into the $B(X,t)$ matrix. Either way is just as general.

We write the individual terms in a multivariate Ito process in the form

$$\mathrm{d}X_i(t) = a_i(X,t)\mathrm{d}t + b_i(X,t)\mathrm{d}B_i(t), \tag{D.85}$$

where the $B_i(t)$ are Wiener processes with

$$\mathrm{corr}(\mathrm{d}B_i(t), \mathrm{d}B_j(t)) = \rho_{ij}, \tag{D.86}$$

for some constants ρ_{ij}. Note that a_i and b_i are functions of all X_j, so the processes are coupled not just through the ρ_{ij}.

Recall that $\mathrm{V}(\mathrm{d}B_i(t)) = \mathrm{V}(\mathrm{d}B_i(t)) = \mathrm{d}t$, and hence $\mathrm{cov}(\mathrm{d}B_i(t), \mathrm{d}B_j(t)) = \rho_{ij}\mathrm{d}t$.

Also recall that $(\mathrm{d}B_i(t))^2 =_{\mathrm{a.s.}} \mathrm{E}((\mathrm{d}B_i(t))^2) = \mathrm{d}t$; i.e., $(\mathrm{d}B_i(t))^2$ is nonstochastic. Likewise, $\mathrm{d}B_i(t)\mathrm{d}B_i(t) =_{\mathrm{a.s.}} \rho_{ij}\mathrm{d}t$.

Given an infinitely differential function G of the vector $X = (X_1, \ldots, X_d)$ and the scalar t, Ito's formula in the form of equation (D.83), derived in the same way as for the univariate case, is

$$\mathrm{d}G(t) = \sum_{i=1}^{d} \frac{\partial G}{\partial X_i} \mathrm{d}X_i(t) + \frac{\partial G}{\partial t}\mathrm{d}t + \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d} \frac{\partial G^2}{\partial X_i \partial X_j}\rho_{ij} b_i(X,t) b_j(X,t)\mathrm{d}t. \tag{D.87}$$

The form of equation (D.79) is obtained by substituting for $\mathrm{d}X_i(t)$.

Solution of Stochastic Differential Equations

The solution of a differential equation is obtained by integrating both sides and allowing for constant terms. Constant terms are evaluated by satisfying known boundary conditions, or initial values.

In a stochastic differential equation (SDE), we must be careful in how the integration is performed, although different interpretations may be equally appropriate.

For example, the SDE that defines an Ito process

$$\mathrm{d}X(t) = a(X,t)\mathrm{d}t + b(X,t)\mathrm{d}B(t),$$

when integrated from time t_0 to T yields

$$X(T) - X(t_0) = \int_{t_0}^{T} a(X,t)\mathrm{d}t + \int_{t_0}^{T} b(X,t)\mathrm{d}B(t).$$

The second integral is a stochastic integral. We will interpret it as an Ito integral.

The nature of $a(X,t)$ and $b(X,t)$ determine the complexity of the solution to the SDE.

In the Ito process

$$\mathrm{d}S(t) = \mu(t)S(t)\mathrm{d}t + \sigma(t)S(t)\mathrm{d}B(t),$$

using Ito's formula for the log as before, we get the solution

$$S(T) = S(t_0)\exp\left(\int_{t_0}^{T}\left(\mu(t) - \frac{1}{2}\sigma(t)^2\right)\mathrm{d}t + \int_{t_0}^{T}\sigma(t)\mathrm{d}B(t)\right).$$

In the simpler version of a geometric Brownian motion model, in which μ and σ are constants, we have

$$S(T) = S(t_0)\exp\left(\left(\mu - \frac{1}{2}\sigma^2\right)\Delta t + \sigma\Delta B\right).$$

Given a solution of a differential equation we may determine the mean, variance and so on by taking expectations of the random component in the solution.

Sometimes, however, it is easier just to develop an ordinary (nonstochastic) differential equation for the moments. We do this from an Ito process

$$\mathrm{d}X(t) = a(X,t)\mathrm{d}t + b(X,t)\mathrm{d}B(t),$$

by using Ito's formula on the powers of the variable. So we have

$$\mathrm{d}X^p(t) = \left(pX(t)^{p-1}a(X,t) + \frac{1}{2}p(p-1)X(t)^{p-2}b(X,t)^2\right)\mathrm{d}t + pX(t)^{p-1}b(X,t)\mathrm{d}B(t).$$

** exercise

Taking expectations of both sides, we have an ordinary differential equation in the expected values.

Jump Processes

We have assumed that a Wiener process is continuous in time (almost surely). A jump process is one that is discontinuous in time.

In financial modeling, we often use a compound process that consists of some smooth process coupled with a jump process. The parameters controlling the frequency of jumps may also be modeled as a stochastic process. The *amount* of the jump is usually modeled as a random variable.

The most important jump processes are Poisson processes.

A Poisson process is a sequence of events in which the probability of k events (where $k = 0, 1, \ldots$) in an interval of length Δt, denoted by $g(k, \Delta t)$ satisfies the following conditions:

- $g(1, \Delta t) = \lambda \Delta t + \mathrm{o}(\Delta t)$, where λ is a positive constant and $(\Delta t) > 0$. (The little o notation, $\mathrm{o}(s)$, stands for any function $h(s)$ such that $\lim_{s\to\infty}(h(s)/s) = 0$; for example, the function s^2 is $\mathrm{o}(s)$. We also have $\mathrm{o}(s) + \mathrm{o}(s) = \mathrm{o}(s)$.)
- $\sum_{k=2}^{\infty} g(k, \Delta t) = \mathrm{o}(\Delta t)$.
- The numbers of changes in nonoverlapping intervals are stochastically independent.

This axiomatic characterization of a Poisson process leads to a differential equation whose solution (using mathematical induction) is

$$g(k, \Delta t) = \frac{(\lambda \Delta t)^k \mathrm{e}^{-\lambda \Delta t}}{k!}, \quad \text{for } k = 1, 2, \ldots$$

which, in turn leads to the familiar probability function for a Poisson distribution

$$p_K(k) = \frac{(\theta)^k \mathrm{e}^{-\theta}}{k!}, \quad \text{for } k = 0, 1, 2, \ldots$$

We merely add a pure jump process $\mathrm{d_j}S(t)$ to our drift-diffusion process,

$$\mathrm{d}S(t) = \mu(S(t), t)\mathrm{d}t + \sigma(S(t), t)\mathrm{d}B(t).$$

After rearranging terms, this yields

$$\begin{aligned} \mathrm{d}S(t) = &\left(\mu(S(t),t) + (\lambda(S(t),t)\int_{\mathcal{Z}} z\, p_Z(z;S(t))\mathrm{d}z\right)\mathrm{d}t \\ &+\sigma(S(t),t)\mathrm{d}B(t) \\ &+\mathrm{d_j}J_S(t). \end{aligned}$$

There are two stochastic terms, $\mathrm{d}B(t)$ and $\mathrm{d_j}J_S(t)$.

We will assume that they are independent.

Note that I suppressed the $\mathrm{d_j}$ on the left hand side, although, clearly, this is a discontinuous process, both because of the compensated process and the discontinuity in the drift.

Ito's Formula in Jump-Diffusion Processes

Now suppose we are interested in a process defined by a function g of $S(t)$ and t. This is where Ito's formula is used.

The simple approach is to apply Ito's formula directly to the drift-diffusion part and then consider $\mathrm{d_j}g(t)$ separately. (We have absorbed $S(t)$ into t in the notation $g(t)$.)

As before,we consider the random variable of the magnitude of the change, Δg and write the process as a systematic component plus a random component

$$\begin{aligned} \mathrm{d_j}g(t) &= g(t) - g(t^-) \\ &= \left(\lambda(S(t),t)\int_{\mathcal{D}(\Delta g)} p_{\Delta g}(\Delta g; g(t))\mathrm{d}\Delta g\right)\mathrm{d}t + \mathrm{d_j}J_g(t) \end{aligned}$$

where the random component $\mathrm{d_j}J_g(t)$ is a compensated process as before.

Putting this all together we have

$$\begin{aligned} \mathrm{d}g(t) = &\left(\frac{\partial g}{\partial t} + \mu\frac{\partial g}{\partial S} + \frac{1}{2}\sigma^2\frac{\partial^2 g}{\partial S^2}\right. \\ = &\quad \left. +\lambda(t)\int_{\mathcal{D}(\Delta g)} \Delta g\, p_{\Delta g}(\Delta g; g(t))\mathrm{d}\Delta g\right)\mathrm{d}t \\ = &\quad +\frac{\partial g}{\partial S}\sigma\mathrm{d}B(t) \\ = &\quad +\mathrm{d_j}J_g(t). \end{aligned}$$

We must remember that this is a discontinuous process.

Notes and Additional References for Section D.3

Stochastic calculus is widely used in models of prices of financial assets, and many of the developments in the general theory have come from that area of application.

Additional References

Øksendal, Bernt (1998), *Stochastic Differential Equations. An Introduction with Applications*, fifth edition, Springer, Heidelberg.

Steele, . Michael (2001), *Stochastic Calculus and Financial Applications*, Springer-Verlag, New York.

D.4 Some Basics of Linear Algebra: Matrix/Vector Definitions and Facts

In the following we will assume the usual axioms for the reals, $\mathbb{R}$. We will be concerned with two linear structures on $\mathbb{R}$. We denote one as $\mathbb{R}^n$, and call its members *vectors*. We denote another as $\mathbb{R}^{n\times m}$, and call its members *matrices*. For both structures we have scalar multiplication (multiplication of a member of the structure by a member of $\mathbb{R}$), an addition operation, an additive identity, and additive inverses for all elements. The addition operation is denoted by "+" and the additive identity by "0", which are the same two symbols used similarly in $\mathbb{R}$. We also have various types of multiplication operations, all with identities, and some with inverses. In addition, we define various real-valued functions over these structures, the most important of which are inner products and norms.

Both $\mathbb{R}^n$ and $\mathbb{R}^{n\times m}$ with addition and multiplication operations are *linear spaces*.

In this section, we abstract some of the basic material on linear algebra from Gentle (2007).

D.4.1 Inner Products, Norms, and Metrics

Although various inner products could be defined in $\mathbb{R}^n$, "the" inner product or dot product for vectors x and y in $\mathbb{R}^n$ is defined as $\sum_{i=1}^n x_i y_i$, and is often written as $x^{\mathrm{T}}y$. It is easy to see that this satisfies the definition of an inner product. *Note that this is different from the notation in Shao; Shao uses x^τ in place of x^{T}.*

Two elements $x, y \in \mathbb{R}^n$ are said to be *orthogonal* if $\langle x, y\rangle = 0$.

An element $x \in \mathbb{R}^n$ is said to be *normal* or *normalized* if $\langle x, x\rangle = 1$. Any $x \neq 0$ can be normalized, that is, mapped to a normal element, $x/\langle x, x\rangle$. A set of normalized elements that are pairwise orthogonal is called an *orthonormal* set. (On page 364 we discuss a method of forming a set of orthogonal vectors.)

Various inner products could be defined in $\mathbb{R}^{n\times m}$, but "the" inner product or dot product for matrices A and B in $\mathbb{R}^{n\times m}$ is defined as $\sum_{j=1}^m a_j^{\mathrm{T}} b_j$, where a_j is the vector whose elements are those from the j^{th} column of A, and likewise for b_j. Again, it is easy to see that this satisfies the definition of an inner product.

Norms and Metrics

There are various norms that can be defined on $\mathbb{R}^n$. An important class of norms are the L_p norms, defined for $p \geq 1$ by

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}. \tag{D.88}$$

It is easy to see that this satisfies the definition of a norm.

The norm in $\mathbb{R}^n$ induced by the inner product (that is, "the" inner product) is the Euclidean norm or the L_2 norm:

$$\|x\|_2 = \sqrt{\langle x, x\rangle} = \sqrt{\sum_{i=1}^{n} x_i^2}. \tag{D.89}$$

This is the only L_p norm induced by an inner product.

The norm in $\mathbb{R}^{n\times m}$ induced by the inner product exists only for $n = m$. In that case it is $\|A\| = \sum_{j=1}^{n} a_j^{\mathrm{T}} a_j = \sum_{j=1}^{n}\sum_{i=1}^{n} a_{ij}^2$. Note that this is not the L_2 matrix norm; it is the Frobenius norm (see below).

The most common and useful metrics in $\mathbb{R}^n$ and $\mathbb{R}^{n\times m}$ are those induced by the norms. For $\mathbb{R}^n$ the L_2 norm is the most common, and a metric for $x, y \in \mathbb{R}^n$ is defined as

$$\rho(x, y) = \|x - y\|_2. \tag{D.90}$$

This metric is called the Euclidean distance.

D.4.2 Matrices and Vectors

Vectors are n-tuples and matrices are n by m rectangular arrays. We will be interested in vectors and matrices whose elements are real numbers. We denote the set of such vectors as $\mathbb{R}^n$ and the set of such matrics as $\mathbb{R}^{n\times m}$.

We generally denote a member of $\mathbb{R}^{n\times m}$ by an upper case letter. A member of $\mathbb{R}^{n\times m}$ consists of nm elements, which we denote by use of two subscripts. We often use a lower-case letter with the two subscripts. For example, for a matrix A, we denote the elements as A_{ij} or a_{ij} with $i = 1, \ldots, n$ and $j = 1, \ldots, m$.

The transpose of a matrix A in $\mathbb{R}^{n\times m}$ is a matrix in $\mathbb{R}^{m\times n}$ denoted by A^{T} such that $(A^{\mathrm{T}})_{ij} = A_{ji}$. Note that this is consistent with the use of $^{\mathrm{T}}$ above for vectors. *Note that this is different from the notation in Shao; Shao uses A^{τ}.*

If $n = m$ the matrix is *square*.

We define (Cayley) multiplication of the matrix $A \in \mathbb{R}^{n\times m}$ and the matrix $B \in \mathbb{R}^{m\times p}$ as $C = AB \in \mathbb{R}^{n\times p}$, where $c_{ij} = \sum_{k=1}^{m} a_{ik} b_{kj}$.

If x and y are n-vectors, in most cases, we can consider them to be $n \times 1$ matrices. Hence, $x^{\mathrm{T}}y$ is a 1×1 matrix and xy^{T} is an $n \times n$ matrix.

We see from the definition that $x^{\mathrm{T}}y$ is an inner product. This inner product is also called the dot product. The product xy^{T} is called the outer product.

As above, we see that $\sqrt{x^{\mathrm{T}}x}$ is a norm (it is the induced norm). We sometimes denote this norm as $\|x\|_2$, because it is $\left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2}$. We call it the Euclidean norm and also the L_2 norm. More generally, for $p \geq 1$, we define the L_p norm for the n-vector x as $\left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$.

We denote the L_p norm of x as $\|x\|_p$. We generally denote the Euclidean or L_2 norm simply as $\|x\|$.

The sum of the diagonal elements of a square matrix is called the *trace* of the matrix. We use the notation "tr(A)" to denote the trace of the matrix A:

$$\mathrm{tr}(A) = \sum_i a_{ii}.$$

$$\mathrm{tr}(A) = \mathrm{tr}(A^{\mathrm{T}}).$$

For a scalar c and an $n \times n$ matrix A,

$$\mathrm{tr}(cA) = c\,\mathrm{tr}(A).$$

If A and B are such that both AB and BA are defined,

$$\mathrm{tr}(AB) = \mathrm{tr}(BA).$$

If x is a vector, we have

$$\|x\|^2 = x^{\mathrm{T}}x = \mathrm{tr}(x^{\mathrm{T}}x) = \mathrm{tr}(xx^{\mathrm{T}}).$$

If x is a vector and A a matrix, we have

$$x^{\mathrm{T}}Ax = \mathrm{tr}(x^{\mathrm{T}}Ax) = \mathrm{tr}(Axx^{\mathrm{T}}).$$

Properties, Concepts, and Notation Associated with Matrices and Vectors

linear independence A set of vectors $x_1, \ldots, x_n \in \mathbb{R}^n$ is linearly independent if $\sum_{i=1}^n a_i x_i = 0$ implies $a_i = 0$ for $i = 1, \ldots, n$.

rank of a matrix The rank of a matrix is the maximum number of rows or columns that are linearly independent. (The maximum number of rows that are linearly independent is the same as the maximum number of columns that are linearly independent.) For the matrix A, we write $\mathrm{rank}(A)$. We adopt the convention that $\mathrm{rank}(A) = 0 \Leftrightarrow A = 0$ (the zero matrix). $A \in \mathbb{R}^{n\times m}$ is said to be full rank iff $\mathrm{rank}(A) = \min(n, m)$.
An important fact is

$$\mathrm{rank}(AB) \leq \min(\mathrm{rank}(A), \mathrm{rank}(B)),$$

and a consequence of this is that the rank of an outer product is less than or equal to 1.

determinant of a square matrix The determinant is a real number. We write $|A|$ or $\det(A)$. $|A| \neq 0$ iff A is square and of full rank.

identity matrix $I \in \mathbb{R}^{n\times n}$ and $I[i, j] = 0$ if $i \neq j$ and $I[i, j] = 1$ if $i \neq i$; that is $I[i, j] = \delta_{ij}$, where δ_{ij} is the Kronecker delta. We write the identity as I_n or just I.

inverse of a matrix For $A \in \mathbb{R}^{n\times n}$, if a matrix $B \in \mathbb{R}^{n\times n}$ exists, such that $AB = I$, then B is the inverse of A, and is written A^{-1}. A matrix has an inverse iff it is square and of full rank.

generalized inverse of a matrix For $A \in \mathbb{R}^{n\times m}$, a matrix $B \in \mathbb{R}^{m\times n}$ such that $ABA = A$ is called a generalized inverse of A, and is written A^-. If A is nonsingular (square and full rank), then obviously $A^- = A^{-1}$.

pseudoinverse or Moore-Penrose inverse of a matrix For $A \in \mathbb{R}^{n\times m}$, the matrix $B \in \mathbb{R}^{m\times n}$ such that $ABA = A$, $BAB = B$, $(AB)^{\mathrm{T}} = AB$, and $(BA)^{\mathrm{T}} = BA$ is called the pseudoinverse of A, and is written A^+.

orthogonal matrix For $A \in \mathbb{R}^{n\times m}$, if $A^{\mathrm{T}}A = I_m$, that is, if the columns are orthonormal and $m \leq n$, or $AA^{\mathrm{T}} = I_n$, that is, if the rows are orthonormal and $n \leq m$, then A is said to be orthogonal.

quadratic forms For $A \in \mathbb{R}^{n\times n}$ and $x \in \mathbb{R}^n$, the scalar $x^{\mathrm{T}}Ax$ is called a quadratic form.

nonnegative definite matrix For $A \in \mathbb{R}^{n\times n}$ and any $x \in \mathbb{R}^n$, if $x^{\mathrm{T}}Ax \geq 0$, then is said to be nonnegative definite. We generally restrict the definition to symmetric matrices. This is essentially without loss of generality because if a matrix is nonnegative definite, then there is a similar symmetric matrix. (Two matrices are said to be *similar* if they have exactly the same eigenvalues.) We write $A \succeq 0$ to denote that A is nonnegative definite.
Note that this is different from Shao.

positive definite matrix For $A \in \mathbb{R}^{n\times n}$ and any $x \in \mathbb{R}^n$, if $x^{\mathrm{T}}Ax \geq 0$ and $x^{\mathrm{T}}Ax = 0$ implies $x = 0$, then is said to be positive definite. As with nonnegative definite matrices, we generally restrict the definition of positive definite matrices to symmetric matrices. We write $A \succ 0$ to denote that A is positive definite.
Note that this is different from Shao.

eigenvalues and eigenvectors If $A \in \mathbb{R}^{n\times n}$, v is an n-vector (complex), and c is a scalar (complex), and $Av = cv$, then c is an eigenvalue of A and v is an eigenvector of A associated with c. All eigenvalues and eigenvectors of a (real) symmetric matrix are real. The eigenvalues of a nonnegative definite matrix are all nonnegative, and the eigenvalues of a positive definite matrix are all positive.

Matrix Factorizations

There are a number of useful ways of factorizing a matrix.

- the LU (and LR and LDU) factorization of a general matrix:
- the QR factorization of a general matrix,
- the similar canonical factorization or "diagonal factorization" of a diagonalizable matrix (which is necessarily square):

$$A = VCV^{-1},$$

where V is a matrix whose columns correspond to the eigenvectors of A and is nonsingular, and C is a diagonal matrix whose entries are the eigenvalues corresponding to the columns of V.

- the singular value factorization of a general $n \times m$ matrix A:

$$A = UDV^{\mathrm{T}},$$

where U is an $n \times n$ orthogonal matrix, V is an $m \times m$ orthogonal matrix, and D is an $n \times m$ diagonal matrix with nonnegative entries. (An $n \times m$ diagonal matrix has $\min(n, m)$ elements on the diagonal, and all other entries are zero.)

- the square root of a nonnegative definite matrix A (which is necessarily symmetric):

$$A = A^{1/2}A^{1/2}$$

- the Cholesky factorization of a nonnegative definite matrix:

$$A = A_{\mathrm{C}}^{\mathrm{T}}A_{\mathrm{C}},$$

where A_{C} is an upper triangular matrix with nonnegative diagonal elements.

Spectral Decomposition

For a symmetric matrix A, we can always write $A = VCV^{\mathrm{T}}$, as above. This is called the spectral decomposition, and is unique except for the ordering and the choice of eigenvectors for eigenvalues with multiplicities greater than 1. We can also write

$$A = \sum_i c_i P_i,$$

where the P_i are the outer products of the eigenvectors,

$$P_i = v_i v_i^{\mathrm{T}},$$

and are called spectral projectors.

Matrix Norms

A matrix norm is generally required to satisfy one more property in addition to those listed above for the definition of a norm. It is the consistency property: $\|AB\| \leq \|A\| \, \|B\|$. The L_p matrix norm for the $n \times m$ matrix A is defined as

$$\|A\|_p = \max_{\|x\|_p=1} \|Ax\|_p.$$

The L_2 matrix norm has the interesting relationship

$$\|A\|_2 = \sqrt{\rho(A^{\mathrm{T}}A)},$$

where $\rho(\cdot)$ is the spectral radius (the modulus of the eigenvalue with the maximum modulus).

The "usual" matrix norm is the Frobenius norm:

$$\|A\|_{\mathrm{F}} = \sqrt{\sum_{i,j} a_{ij}^2}.$$

Idempotent and Projection Matrices

A matrix A such that $AA = A$ is called an *idempotent matrix*. An idempotent matrix is square, and it is either singular or it is the identity matrix. (It must be square in order to be conformable for the indicated multiplication. If it is not singular, we have $A = (A^{-1}A)A = A^{-1}(AA) = A^{-1}A = I$; hence, an idempotent matrix is either singular or it is the identity matrix.)

A useful idempotent matrix often encountered in statistical linear models is Z^-Z.

If A is idempotent and $n \times n$, then $(I - A)$ is also idempotent, as we see by multiplication.

In this case, we also have

$$\mathrm{rank}(I - A) = n - \mathrm{rank}(A).$$

Because the eigenvalues of A^2 are the squares of the eigenvalues of A, all eigenvalues of an idempotent matrix must be either 0 or 1. The number of eigenvalues that are 1 is the rank of the matrix. We therefore have for an idempotent matrix A,

$$\mathrm{tr}(A) = \mathrm{rank}(A).$$

Because $AA = A$, any vector in the column space of A is an eigenvector of A.

For a given vector space $\mathcal{V}$, a symmetric idempotent matrix A whose columns span $\mathcal{V}$ is said to be a *projection matrix* onto $\mathcal{V}$; in other words, a matrix A is a projection matrix onto $\mathrm{span}(A)$ if and only if A is symmetric and idempotent.

It is easy to see that for any vector x, if A is a projection matrix onto $\mathcal{V}$, the vector Ax is in $\mathcal{V}$, and the vector $x - Ax$ is in $\mathcal{V}^\perp$ (the vectors Ax and $x - Ax$ are orthogonal). For this reason, a projection matrix is sometimes called an "orthogonal projection matrix". Note that an orthogonal projection matrix is not an orthogonal matrix, however, unless it is the identity matrix. Stating this in alternate notation, if A is a projection matrix and $A \in \mathbb{R}^{n\times n}$, then A maps $\mathbb{R}^n$ onto $\mathcal{V}(A)$, and $I - A$ is also a projection matrix (called the *complementary projection matrix* of A), and it maps $\mathbb{R}^n$ onto the orthogonal complement, $\mathcal{N}(A)$. These spaces are such that $\mathcal{V}(A) \oplus \mathcal{N}(A) = \mathbb{R}^n$.

Useful projection matrices often encountered in statistical linear models are A^+A and AA^+.

If x is a general vector in $\mathbb{R}^n$, that is, if x has order n and belongs to an n-dimensional space, and A is a projection matrix of rank $r \leq n$, then Ax has order n and belongs to span(A), which is an r-dimensional space.

Because a projection matrix is idempotent, the matrix projects any of its columns onto itself, and of course it projects the full matrix onto itself: $AA = A$. More generally, if x and y are vectors in span(A) and a is a scalar, then

$$A(ax + y) = ax + y.$$

(To see this, we merely represent x and y as linear combinations of columns (or rows) of A and substitute in the equation.)

The projection of a vector y onto a vector x is

$$\frac{x^{\mathrm{T}}y}{x^{\mathrm{T}}x}x.$$

The projection matrix to accomplish this is the "outer/inner products matrix",

$$\frac{1}{x^{\mathrm{T}}x}xx^{\mathrm{T}}.$$

The outer/inner products matrix has rank 1. It is useful in a variety of matrix transformations. If x is normalized, the projection matrix for projecting a vector on x is just xx^{T}. The projection matrix for projecting a vector onto a unit vector e_i is $e_i e_i^{\mathrm{T}}$, and $e_i e_i^{\mathrm{T}} y = (0, \ldots, y_i, \ldots, 0)$.

Inverses of Matrices

Often in applications we need inverses of various sums of matrices. If A and B are full rank matrices of the same size, the following relationships are easy to show.

$$\begin{aligned}
(I + A^{-1})^{-1} &= A(A + I)^{-1} \\
(A + BB^{\mathrm{T}})^{-1}B &= A^{-1}B(I + B^{\mathrm{T}}A^{-1}B)^{-1} \\
(A^{-1} + B^{-1})^{-1} &= A(A + B)^{-1}B \\
A - A(A + B)^{-1}A &= B - B(A + B)^{-1}B \\
A^{-1} + B^{-1} &= A^{-1}(A + B)B^{-1} \\
(I + AB)^{-1} &= I - A(I + BA)^{-1}B \\
(I + AB)^{-1}A &= A(I + BA)^{-1}
\end{aligned}$$

From the relationship $\det(AB) = \det(A)\det(B)$ for square matrices mentioned earlier, it is easy to see that for nonsingular A,

$$\det(A) = 1/\det(A^{-1}).$$

For a square matrix A, $\det(A) = 0$ if and only if A is singular.

Partitioned Matrices

We often find it useful to partition a matrix into submatrices, and we usually denote those submatrices with capital letters with subscripts indicating the relative positions of the submatrices. Hence, we may write

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

where the matrices A_{11} and A_{12} have the same number of rows, A_{21} and A_{22} have the same number of rows, A_{11} and A_{21} have the same number of columns, and A_{12} and A_{22} have the same number of columns.

The term "submatrix" is also sometimes used to refer to a matrix formed from another one by deleting various rows and columns of the given matrix. In this terminology, B is a submatrix of A if for each element b_{ij} there is an a_{kl} with $k \geq i$ and $l \geq j$, such that $b_{ij} = a_{kl}$; that is, the rows and/or columns of the submatrix are not contiguous in the original matrix.

A submatrix whose principal diagonal elements are elements of the principal diagonal of the given matrix is called a *principal submatrix*; A_{11} is a principal submatrix in the example above, and if A_{22} is square it is also a principal submatrix. Sometimes the term "principal submatrix" is restricted to square submatrices.

A principal submatrix that contains the $(1, 1)$ and whose rows and columns are contiguous in the original matrix is called a *leading principal submatrix*. A_{11} is a principal submatrix in the example above.

Multiplication and other operations with matrices, such as transposition, are carried out with their submatrices in the obvious way. Thus,

$$\begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} A_{11}^{\mathrm{T}} & A_{21}^{\mathrm{T}} \\ A_{12}^{\mathrm{T}} & A_{22}^{\mathrm{T}} \\ A_{13}^{\mathrm{T}} & A_{23}^{\mathrm{T}} \end{bmatrix},$$

and, assuming the submatrices are conformable for multiplication,

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{bmatrix}.$$

Sometimes a matrix may be partitioned such that one partition is just a single column or row, that is, a vector or the transpose of a vector. In that case, we may use a notation such as

$$[X \; y]$$

or

$$[X \,|\, y],$$

where X is a matrix and y is a vector. We develop the notation in the obvious fashion; for example,

$$[X\ y]^{\mathrm{T}}[X\ y] = \begin{bmatrix} X^{\mathrm{T}}X & X^{\mathrm{T}}y \\ y^{\mathrm{T}}X & y^{\mathrm{T}}y \end{bmatrix}.$$

Partitioned matrices may also have useful patterns. A "block diagonal" matrix is one of the form

$$\begin{bmatrix} \mathtt{X} & \mathtt{0} & \cdots & \mathtt{0} \\ \mathtt{0} & \mathtt{X} & \cdots & \mathtt{0} \\ & & \ddots & \\ \mathtt{0} & \mathtt{0} & \cdots & \mathtt{X} \end{bmatrix},$$

where `0` represents a submatrix with all zeros, and `X` represents a general submatrix, with at least some nonzeros. The diag(·) function previously introduced for a vector is also defined for a list of matrices:

$$\mathrm{diag}(A_1, A_2, \ldots, A_k)$$

denotes the block diagonal matrix with submatrices $A_1, A_2, \ldots, A_k$ along the diagonal and zeros elsewhere.

Inverses of Partitioned Matrices

If A is nonsingular, and can be partitioned as

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

where both A_{11} and A_{22} are nonsingular, it is easy to see that the inverse of A is given by

$$A^{-1} = \begin{bmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}Z^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}Z^{-1} \\ -Z^{-1}A_{21}A_{11}^{-1} & Z^{-1} \end{bmatrix},$$

where $Z = A_{22} - A_{21}A_{11}^{-1}A_{12}$. In this partitioning Z is called the *Schur complement* of A_{11} in A.

If

$$A = [Xy]^{\mathrm{T}}[Xy]$$

and is partitioned as into $X^{\mathrm{T}}X$ and $y^{\mathrm{T}}y$ on the diagonal, and X is of full column rank, then the Schur complement of $X^{\mathrm{T}}X$ in $[Xy]^{\mathrm{T}}[Xy]$ is

$$y^{\mathrm{T}}y - y^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y.$$

This particular partitioning is useful in linear regression analysis, where this Schur complement is the residual sum of squares.

Gramian Matrices and Generalized Inverses

A matrix of the form $Z^{\mathrm{T}}Z$ is called a *Gramian matrix*. Such matrices arise often in statistical applications.

Some interesting properties of a Gramian matrix $Z^{\mathrm{T}}Z$ are

- $Z^{\mathrm{T}}Z$ is symmetric;
- $Z^{\mathrm{T}}Z$ is of full rank if and only if Z is of full column rank, or, more generally,

$$\mathrm{rank}(Z^{\mathrm{T}}Z) = \mathrm{rank}(Z);$$

- $Z^{\mathrm{T}}Z$ is nonnegative definite, and positive definite if and only if Z is of full column rank;
- $Z^{\mathrm{T}}Z = 0 \implies Z = 0$.

The generalized inverses of $Z^{\mathrm{T}}Z$ have useful properties. First, we see from the definition, for any generalized inverse, $(Z^{\mathrm{T}}Z)^-$ that $((Z^{\mathrm{T}}Z)^-)^{\mathrm{T}}$ is also a generalized inverse of $Z^{\mathrm{T}}Z$. (Note that $(Z^{\mathrm{T}}Z)^-$ is not necessarily symmetric.)

Another useful property of a Gramian matrix is that for any matrices B and C (that are conformable for the operations indicated),

$$BZ^{\mathrm{T}}Z = CZ^{\mathrm{T}}Z \iff BZ^{\mathrm{T}} = CZ^{\mathrm{T}}.$$

The implication from right to left is obvious, and we can see the left to right implication by writing

$$(BZ^{\mathrm{T}}Z - CZ^{\mathrm{T}}Z)(B^{\mathrm{T}} - C^{\mathrm{T}}) = (BZ^{\mathrm{T}} - CZ^{\mathrm{T}})(BZ^{\mathrm{T}} - CZ^{\mathrm{T}})^{\mathrm{T}},$$

and then observing that if the left side is null, then so is the right side, and if the right side is null, then $BZ^{\mathrm{T}} - CZ^{\mathrm{T}} = 0$. Similarly, we have

$$Z^{\mathrm{T}}ZB = Z^{\mathrm{T}}ZC \iff Z^{\mathrm{T}}B = Z^{\mathrm{T}}C.$$

Also,

$$Z(Z^{\mathrm{T}}Z)^-Z^{\mathrm{T}}Z = Z.$$

This means that $(Z^{\mathrm{T}}Z)^-Z^{\mathrm{T}}$ is a generalized inverse of Z

An important property of $Z(Z^{\mathrm{T}}Z)^-Z^{\mathrm{T}}$ is its invariance to the choice of the generalized inverse of $Z^{\mathrm{T}}Z$. Suppose G is any generalized inverse of $Z^{\mathrm{T}}Z$. Then we have

$$ZGZ^{\mathrm{T}} = Z(Z^{\mathrm{T}}Z)^-Z^{\mathrm{T}};$$

that is, $Z(Z^{\mathrm{T}}Z)^-Z^{\mathrm{T}}$ is invariant to choice of generalized inverse.

The squared norm of the residual vector obtained from any generalized inverse of $Z^{\mathrm{T}}Z$ has some interesting properties. First, just by direct multiplication, we get the "Pythagorean property" of the norm of the predicted values and the residuals:

$$\|X - Z\beta\|^2 = \|X - Z\widehat{\beta}\|^2 + \|Z\widehat{\beta} - Z\beta\|^2$$

where $\widehat{\beta} = (Z^{\mathrm{T}}Z)^{-}Z^{\mathrm{T}}X$ for any generalized inverse. We also have

$$\mathrm{E}(Z\widehat{\beta}) = Z\beta,$$

and

$$\mathrm{E}((Z\widehat{\beta} - Z\beta)^{\mathrm{T}}(Z\widehat{\beta} - Z\beta)) = \mathrm{V}(Z\widehat{\beta}).$$

Because for any vector y, we have

$$\|y\|^2 = y^{\mathrm{T}}y = \mathrm{tr}(y^{\mathrm{T}}y),$$

we can derive an interesting expression for $\mathrm{E}(\|X - Z\beta\|^2)$ (Shao, p. 187):

$$\begin{aligned}
\mathrm{E}(\|X - Z\widehat{\beta}\|^2) &= \mathrm{tr}(\mathrm{E}(\|X - Z\widehat{\beta}\|^2)) \\
&= \mathrm{tr}(\mathrm{E}((X - Z\beta)^{\mathrm{T}}(X - Z\beta)) - \mathrm{E}((Z\widehat{\beta} - Z\beta)^{\mathrm{T}}(Z\widehat{\beta} - Z\beta))) \\
&= \mathrm{tr}(\mathrm{V}(X) - \mathrm{V}(Z\widehat{\beta}) \\
&= n\sigma^2 - \mathrm{tr}((Z(Z^{\mathrm{T}}Z)^{-}Z^{\mathrm{T}})\sigma^2 I(Z(Z^{\mathrm{T}}Z)^{-}Z^{\mathrm{T}})) \\
&= \sigma^2(n - \mathrm{tr}((Z^{\mathrm{T}}Z)^{-}Z^{\mathrm{T}}Z)).
\end{aligned}$$

The trace in the latter expression is the "regression degrees of freedom".

The Moore-Penrose Inverse

The Moore-Penrose inverse, or the pseudoinverse, of Z has an interesting relationship with a generalized inverse of $Z^{\mathrm{T}}Z$:

$$ZZ^{+} = Z(Z^{\mathrm{T}}Z)^{-}Z^{\mathrm{T}}.$$

This can be established directly from the definition of the Moore-Penrose inverse.

D.4.3 Vector/Matrix Derivatives and Integrals

The operations of differentiation and integration of vectors and matrices are logical extensions of the corresponding operations on scalars. There are three objects involved in this operation:

- the variable of the operation;
- the operand (the function being differentiated or integrated); and
- the result of the operation.

In the simplest case, all three of these objects are of the same type, and they are scalars. If either the variable or the operand is a vector or a matrix, however, the structure of the result may be more complicated. This statement will become clearer as we proceed to consider specific cases.

In this section, we state or show the form that the derivative takes in terms of simpler derivatives. We state high-level rules for the nature of the

differentiation in terms of simple partial differentiation of a scalar with respect to a scalar. We do not consider whether or not the derivatives exist. In general, if the simpler derivatives we write that comprise the more complicated object exist, then the derivative of that more complicated object exists. Once a shape of the derivative is determined, definitions or derivations in ϵ-δ terms could be given, but we will refrain from that kind of formal exercise. The purpose of this section is not to develop a calculus for vectors and matrices but rather to consider some cases that find wide applications in statistics. For a more careful treatment of differentiation of vectors and matrices see Gentle (2007).

Basics of Differentiation

It is useful to recall the heuristic interpretation of a derivative. A derivative of a function is the infinitesimal rate of change of the function with respect to the variable with which the differentiation is taken. If both the function and the variable are scalars, this interpretation is unambiguous. If, however, the operand of the differentiation, Φ, is a more complicated function, say a vector or a matrix, and/or the variable of the differentiation, Ξ, is a more complicated object, the changes are more difficult to measure. Change in the value both of the function,

$$\delta\Phi = \Phi_{\text{new}} - \Phi_{\text{old}},$$

and of the variable,

$$\delta\Xi = \Xi_{\text{new}} - \Xi_{\text{old}},$$

could be measured in various ways, by using various norms, for example. (Note that the subtraction is not necessarily ordinary scalar subtraction.)

Furthermore, we cannot just divide the function values by $\delta\Xi$. We do not have a definition for division by that kind of object. We need a mapping, possibly a norm, that assigns a positive real number to $\delta\Xi$. We can define the change in the function value as just the simple difference of the function evaluated at the two points. This yields

$$\lim_{\|\delta\Xi\|\to 0} \frac{\Phi(\Xi + \delta\Xi) - \Phi(\Xi)}{\|\delta\Xi\|}. \tag{D.91}$$

So long as we remember the complexity of $\delta\Xi$, however, we can adopt a simpler approach. Since for both vectors and matrices, we have definitions of multiplication by a scalar and of addition, we can simplify the limit in the usual definition of a derivative, $\delta\Xi \to 0$. Instead of using $\delta\Xi$ as the element of change, we will use $t\Upsilon$, where t is a scalar and Υ is an element to be added to Ξ. The limit then will be taken in terms of $t \to 0$. This leads to

$$\lim_{t\to 0} \frac{\Phi(\Xi + t\Upsilon) - \Phi(\Xi)}{t} \tag{D.92}$$

as a formula for the derivative of Φ with respect to Ξ.

The expression (D.92) may be a useful formula for evaluating a derivative, but we must remember that it is not the derivative. The type of object of this formula is the same as the type of object of the function, Φ; it does not accommodate the type of object of the argument, Ξ, unless Ξ is a scalar. As we will see below, for example, if Ξ is a vector and Φ is a scalar, the derivative must be a vector, yet in that case the expression (D.92) is a scalar.

The expression (D.91) is rarely directly useful in evaluating a derivative, but it serves to remind us of both the generality and the complexity of the concept. Both Φ and its arguments could be functions, for example. In functional analysis, various kinds of functional derivatives are defined, such as a Gâteaux derivative. These derivatives find applications in developing robust statistical methods. Here we are just interested in the combinations of three possibilities for Φ, namely scalar, vector, and matrix, and the same three possibilities for Ξ and Υ.

Continuity

It is clear from the definition of continuity that for the derivative of a function to exist at a point, the function must be continuous at that point. A function of a vector or a matrix is continuous if it is continuous for each element of the vector or matrix. Just as scalar sums and products are continuous, vector/matrix sums and all of the types of vector/matrix products we have discussed are continuous. A continuous function of a continuous function is continuous.

Many of the vector/matrix functions we have discussed are clearly continuous. For example, the L_p vector norms are continuous over the nonnegative reals but not over the reals unless p is an even (positive) integer. The determinant of a matrix is continuous, as we see from the definition of the determinant and the fact that sums and scalar products are continuous. The fact that the determinant is a continuous function immediately yields the result that cofactors and hence the adjugate are continuous. From the relationship between an inverse and the adjugate, we see that the inverse is a continuous function.

Notation and Properties

We write the differential operator with respect to the dummy variable x as $\partial/\partial x$ or $\partial/\partial x^{\mathrm{T}}$. We usually denote differentiation using the symbol for "partial" differentiation, ∂, whether the operator is written ∂x_i for differentiation with respect to a specific scalar variable or ∂x for differentiation with respect to the array x that contains all of the individual elements. Sometimes, however, if the differentiation is being taken with respect to the whole array (the vector or the matrix), we use the notation $\mathrm{d}/\mathrm{d}x$.

The operand of the differential operator $\partial/\partial x$ is a function of x. (If it is not a function of x — that is, if it is a constant function with respect to x — then the operator evaluates to 0.) The result of the operation, written

$\partial f/\partial x$, is also a function of x, with the same domain as f, and we sometimes write $\partial f(x)/\partial x$ to emphasize this fact. The value of this function at the fixed point x_0 is written as $\partial f(x_0)/\partial x$. (The derivative of the constant $f(x_0)$ is identically 0, but it is not necessary to write $\partial f(x)/\partial x|_{x_0}$ because $\partial f(x_0)/\partial x$ is interpreted as the value of the function $\partial f(x)/\partial x$ at the fixed point x_0.)

If $\partial/\partial x$ operates on f, and $f : S \to T$, then $\partial/\partial x : S \to U$. The nature of S, or more directly the nature of x, whether it is a scalar, a vector, or a matrix, and the nature of T determine the structure of the result U. For example, if x is an n-vector and $f(x) = x^{\mathrm{T}}x$, then

$$f : \mathbb{R}^n \to \mathbb{R}$$

and

$$\partial f/\partial x : \mathbb{R}^n \to \mathbb{R}^n,$$

as we will see. The outer product, $h(x) = xx^{\mathrm{T}}$, is a mapping to a higher rank array, but the derivative of the outer product is a mapping to an array of the same rank; that is,

$$h : \mathbb{R}^n \to \mathbb{R}^{n\times n}$$

and

$$\partial h/\partial x : \mathbb{R}^n \to \mathbb{R}^n.$$

(Note that "rank" here means the number of dimensions. This term is often used in this way in numerical software. See Gentle, 2007, page 5.)

As another example, consider $g(\cdot) = \det(\cdot)$, so

$$g : \mathbb{R}^{n\times n} \mapsto \mathbb{R}.$$

In this case,

$$\partial g/\partial X : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^{n\times n};$$

that is, the derivative of the determinant of a square matrix is a square matrix, as we will see later.

Higher-order differentiation is a composition of the $\partial/\partial x$ operator with itself or of the $\partial/\partial x$ operator and the $\partial/\partial x^{\mathrm{T}}$ operator. For example, consider the familiar function in linear least squares

$$f(b) = (y - Xb)^{\mathrm{T}}(y - Xb).$$

This is a mapping from $\mathbb{R}^m$ to $\mathbb{R}$. The first derivative with respect to the m-vector b is a mapping from $\mathbb{R}^m$ to $\mathbb{R}^m$, namely $2X^{\mathrm{T}}Xb - 2X^{\mathrm{T}}y$. The second derivative with respect to b^{T} is a mapping from $\mathbb{R}^m$ to $\mathbb{R}^{m\times m}$, namely, $2X^{\mathrm{T}}X$.

We see from expression (D.91) that differentiation is a linear operator; that is, if $\mathcal{D}(\Phi)$ represents the operation defined in expression (D.91), Ψ is another function in the class of functions over which $\mathcal{D}$ is defined, and a is a scalar that does not depend on the variable Ξ, then $\mathcal{D}(a\Phi + \Psi) = a\mathcal{D}(\Phi) + \mathcal{D}(\Psi)$. This yields the familiar rules of differential calculus for derivatives of sums

or constant scalar products. Other usual rules of differential calculus apply, such as for differentiation of products and composition (the chain rule). We can use expression (D.92) to work these out. For example, for the derivative of the product $\Phi\Psi$, after some rewriting of terms, we have the numerator

$$\begin{aligned}&\Phi(\Xi)\big(\Psi(\Xi+t\Upsilon)-\Psi(\Xi)\big)\\&+\Psi(\Xi)\big(\Phi(\Xi+t\Upsilon)-\Phi(\Xi)\big)\\&+\big(\Phi(\Xi+t\Upsilon)-\Phi(\Xi)\big)\big(\Psi(\Xi+t\Upsilon)-\Psi(\Xi)\big).\end{aligned}$$

Now, dividing by t and taking the limit, assuming that as

$$t \to 0,$$

$$(\Phi(\Xi+t\Upsilon)-\Phi(\Xi)) \to 0,$$

we have

$$\mathcal{D}(\Phi\Psi) = \mathcal{D}(\Phi)\Psi + \Phi\mathcal{D}(\Psi), \tag{D.93}$$

where again $\mathcal{D}$ represents the differentiation operation.

Differentials

For a differentiable scalar function of a scalar variable, $f(x)$, the *differential of f at c with increment u* is $u\mathrm{d}f/\mathrm{d}x|_c$. This is the linear term in a truncated Taylor series expansion:

$$f(c+u) = f(c) + u\frac{\mathrm{d}}{\mathrm{d}x}f(c) + r(c,u). \tag{D.94}$$

Technically, the differential is a function of both x and u, but the notation $\mathrm{d}f$ is used in a generic sense to mean the differential of f. For vector/matrix functions of vector/matrix variables, the differential is defined in a similar way. The structure of the differential is the same as that of the function; that is, for example, the differential of a matrix-valued function is a matrix.

Types of Differentiation

In the following sections we consider differentiation with respect to different types of objects first, and we consider differentiation of different types of objects.

Differentiation with Respect to a Scalar

Differentiation of a structure (vector or matrix, for example) with respect to a scalar is quite simple; it just yields the ordinary derivative of each element of the structure in the same structure. Thus, the derivative of a vector or a matrix with respect to a scalar variable is a vector or a matrix, respectively, of the derivatives of the individual elements.

Differentiation with respect to a vector or matrix, which we will consider below, is often best approached by considering differentiation with respect to the individual elements of the vector or matrix, that is, with respect to scalars.

Derivatives of Vectors with Respect to Scalars

The derivative of the vector $y(x) = (y_1, \ldots, y_n)$ with respect to the scalar x is the vector

$$\partial y/\partial x = (\partial y_1/\partial x, \ldots, \partial y_n/\partial x). \tag{D.95}$$

The second or higher derivative of a vector with respect to a scalar is likewise a vector of the derivatives of the individual elements; that is, it is an array of higher rank.

Derivatives of Matrices with Respect to Scalars

The derivative of the matrix $Y(x) = (y_{ij})$ with respect to the scalar x is the matrix

$$\partial Y(x)/\partial x = (\partial y_{ij}/\partial x). \tag{D.96}$$

The second or higher derivative of a matrix with respect to a scalar is likewise a matrix of the derivatives of the individual elements.

Derivatives of Functions with Respect to Scalars

Differentiation of a function of a vector or matrix that is linear in the elements of the vector or matrix involves just the differentiation of the elements, followed by application of the function. For example, the derivative of a trace of a matrix is just the trace of the derivative of the matrix. On the other hand, the derivative of the determinant of a matrix is not the determinant of the derivative of the matrix (see below).

Higher-Order Derivatives with Respect to Scalars

Because differentiation with respect to a scalar does not change the rank of the object ("rank" here means rank of an array or "shape"), higher-order derivatives $\partial^k/\partial x^k$ with respect to scalars are merely objects of the same rank whose elements are the higher-order derivatives of the individual elements.

Differentiation with Respect to a Vector

Differentiation of a given object with respect to an n-vector yields a vector for each element of the given object. The basic expression for the derivative, from formula (D.92), is

$$\lim_{t \to 0} \frac{\Phi(x + ty) - \Phi(x)}{t} \tag{D.97}$$

for an arbitrary conformable vector y. The arbitrary y indicates that the derivative is omnidirectional; it is the rate of change of a function of the vector in any direction.

Derivatives of Scalars with Respect to Vectors; The Gradient

The derivative of a scalar-valued function with respect to a vector is a vector of the partial derivatives of the function with respect to the elements of the vector. If $f(x)$ is a scalar function of the vector $x = (x_1, \ldots, x_n)$,

$$\frac{\partial f}{\partial x} = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n} \right), \tag{D.98}$$

if those derivatives exist. This vector is called the *gradient* of the scalar-valued function, and is sometimes denoted by $\mathrm{g}_f(x)$ or $\nabla f(x)$, or sometimes just g_f or ∇f:

$$\mathrm{g}_f = \nabla f = \frac{\partial f}{\partial x}. \tag{D.99}$$

The notation g_f or ∇f implies differentiation with respect to "all" arguments of f, hence, if f is a scalar-valued function of a vector argument, they represent a vector.

This derivative is useful in finding the maximum or minimum of a function. Such applications arise throughout statistical and numerical analysis.

Inner products, bilinear forms, norms, and variances are interesting scalar-valued functions of vectors. In these cases, the function Φ in equation (D.97) is scalar-valued and the numerator is merely $\Phi(x + ty) - \Phi(x)$. Consider, for example, the quadratic form $x^{\mathrm{T}}Ax$. Using equation (D.97) to evaluate $\partial x^{\mathrm{T}}Ax/\partial x$, we have

$$\begin{aligned} &\lim_{t\to 0} \frac{(x+ty)^{\mathrm{T}}A(x+ty) - x^{\mathrm{T}}Ax}{t} \\ &= \lim_{t\to 0} \frac{x^{\mathrm{T}}Ax + ty^{\mathrm{T}}Ax + ty^{\mathrm{T}}A^{\mathrm{T}}x + t^2y^{\mathrm{T}}Ay - x^{\mathrm{T}}Ax}{t} \\ &= y^{\mathrm{T}}(A + A^{\mathrm{T}})x, \end{aligned} \tag{D.100}$$

for an arbitrary y (that is, "in any direction"), and so $\partial x^{\mathrm{T}}Ax/\partial x = (A+A^{\mathrm{T}})x$.

This immediately yields the derivative of the square of the Euclidean norm of a vector, $\|x\|_2^2$, and the derivative of the Euclidean norm itself by using the chain rule. Other L_p vector norms may not be differentiable everywhere because of the presence of the absolute value in their definitions. The fact that the Euclidean norm is differentiable everywhere is one of its most important properties.

The derivative of the quadratic form also immediately yields the derivative of the variance. The derivative of the correlation, however, is slightly more difficult because it is a ratio.

The operator $\partial/\partial x^{\mathrm{T}}$ applied to the scalar function f results in $\mathrm{g}_f^{\mathrm{T}}$.

The second derivative of a scalar-valued function with respect to a vector is a derivative of the first derivative, which is a vector. We will now consider derivatives of vectors with respect to vectors.

Derivatives of Vectors with Respect to Vectors; The Jacobian

The derivative of an m-vector-valued function of an n-vector argument consists of nm scalar derivatives. These derivatives could be put into various structures. Two obvious structures are an $n \times m$ matrix and an $m \times n$ matrix. For a function $f : S \subset \mathbb{R}^n \to \mathbb{R}^m$, we define $\partial f^{\mathrm{T}}/\partial x$ to be the $n \times m$ matrix, which is the natural extension of $\partial/\partial x$ applied to a scalar function, and $\partial f/\partial x^{\mathrm{T}}$ to be its transpose, the $m \times n$ matrix. Although the notation $\partial f^{\mathrm{T}}/\partial x$ is more precise because it indicates that the elements of f correspond to the columns of the result, we often drop the transpose in the notation. We have

$$\begin{aligned}
\frac{\partial f}{\partial x} &= \frac{\partial f^{\mathrm{T}}}{\partial x} \quad \text{by convention} \\
&= \left[\frac{\partial f_1}{\partial x} \cdots \frac{\partial f_m}{\partial x}\right] \\
&= \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_2} \\ & \cdots & & \\ \frac{\partial f_1}{\partial x_n} & \frac{\partial f_2}{\partial x_n} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}
\end{aligned} \tag{D.101}$$

if those derivatives exist. This derivative is called the *matrix gradient* and is denoted by G_f or ∇f for the vector-valued function f. (Note that the ∇ symbol can denote either a vector or a matrix, depending on whether the function being differentiated is scalar-valued or vector-valued.)

The $m \times n$ matrix $\partial f/\partial x^{\mathrm{T}} = (\nabla f)^{\mathrm{T}}$ is called the *Jacobian* of f and is denoted by J_f:

$$\mathrm{J}_f = \mathrm{G}_f^{\mathrm{T}} = (\nabla f)^{\mathrm{T}}. \tag{D.102}$$

The absolute value of the determinant of the Jacobian appears in integrals involving a change of variables. (Occasionally, the term "Jacobian" is used to refer to the absolute value of the determinant rather than to the matrix itself.)

To emphasize that the quantities are functions of x, we sometimes write $\partial f(x)/\partial x$, $\mathrm{J}_f(x)$, $\mathrm{G}_f(x)$, or $\nabla f(x)$.

Derivatives of Matrices with Respect to Vectors

The derivative of a matrix with respect to a vector is a three-dimensional object that results from applying equation (D.98) to each of the elements of the matrix. For this reason, it is simpler to consider only the partial derivatives of the matrix Y with respect to the individual elements of the vector x; that is, $\partial Y/\partial x_i$. The expressions involving the partial derivatives can be thought of as defining one two-dimensional layer of a three-dimensional object.

Using the rules for differentiation of powers that result directly from the definitions, we can write the partial derivatives of the inverse of the matrix Y as

$$\frac{\partial}{\partial x}Y^{-1} = -Y^{-1}\left(\frac{\partial}{\partial x}Y\right)Y^{-1}. \tag{D.103}$$

Beyond the basics of differentiation of constant multiples or powers of a variable, the two most important properties of derivatives of expressions are the linearity of the operation and the chaining of the operation. These yield rules that correspond to the familiar rules of the differential calculus. A simple result of the linearity of the operation is the rule for differentiation of the trace:

$$\frac{\partial}{\partial x}\mathrm{tr}(Y) = \mathrm{tr}\left(\frac{\partial}{\partial x}Y\right).$$

Higher-Order Derivatives with Respect to Vectors; The Hessian

Higher-order derivatives are derivatives of lower-order derivatives. As we have seen, a derivative of a given function with respect to a vector is a more complicated object than the original function. The simplest higher-order derivative with respect to a vector is the second-order derivative of a scalar-valued function. Higher-order derivatives may become uselessly complicated.

In accordance with the meaning of derivatives of vectors with respect to vectors, the second derivative of a scalar-valued function with respect to a vector is a matrix of the partial derivatives of the function with respect to the elements of the vector. This matrix is called the *Hessian*, and is denoted by H_f or sometimes by $\nabla\nabla f$ or $\nabla^2 f$:

$$\mathrm{H}_f = \frac{\partial^2 f}{\partial x \partial x^{\mathrm{T}}} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_m} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_m} \\ & \cdots & & \\ \frac{\partial^2 f}{\partial x_m \partial x_1} & \frac{\partial^2 f}{\partial x_m \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_m^2} \end{bmatrix}. \tag{D.104}$$

To emphasize that the Hessian is a function of x, we sometimes write $\mathrm{H}_f(x)$ or $\nabla\nabla f(x)$ or $\nabla^2 f(x)$.

Summary of Derivatives with Respect to Vectors

As we have seen, the derivatives of functions are complicated by the problem of measuring the change in the function, but often the derivatives of functions with respect to a vector can be determined by using familiar scalar differentiation. In general, we see that

- the derivative of a scalar (a quadratic form) with respect to a vector is a vector and

- the derivative of a vector with respect to a vector is a matrix.

Table D.2 lists formulas for the vector derivatives of some common expressions. The derivative $\partial f/\partial x^{\mathrm{T}}$ is the transpose of $\partial f/\partial x$.

Table D.2. Formulas for Some Vector Derivatives

$f(x)$	$\partial f/\partial x$
ax	a
$b^{\mathrm{T}}x$	b
$x^{\mathrm{T}}b$	b^{T}
$x^{\mathrm{T}}x$	$2x$
xx^{T}	$2x^{\mathrm{T}}$
$b^{\mathrm{T}}Ax$	$A^{\mathrm{T}}b$
$x^{\mathrm{T}}Ab$	$b^{\mathrm{T}}A$
$x^{\mathrm{T}}Ax$	$(A+A^{\mathrm{T}})x$
	$2Ax$, if A is symmetric
$\exp(-\frac{1}{2}x^{\mathrm{T}}Ax)$	$-\exp(-\frac{1}{2}x^{\mathrm{T}}Ax)Ax$, if A is symmetric
$\|x\|_2^2$	$2x$
$\mathrm{V}(x)$	$2x/(n-1)$

In this table, x is an n-vector, a is a constant scalar, b is a constant conformable vector, and A is a constant conformable matrix.

Differentiation with Respect to a Matrix

The derivative of a function with respect to a matrix is a matrix with the same shape consisting of the partial derivatives of the function with respect to the elements of the matrix. This rule defines what we mean by differentiation with respect to a matrix.

By the definition of differentiation with respect to a matrix X, we see that the derivative $\partial f/\partial X^{\mathrm{T}}$ is the transpose of $\partial f/\partial X$. For scalar-valued functions, this rule is fairly simple. For example, consider the trace. If X is a square matrix and we apply this rule to evaluate $\partial\,\mathrm{tr}(X)/\partial X$, we get the identity matrix, where the nonzero elements arise only when $j=i$ in $\partial(\sum x_{ii})/\partial x_{ij}$. If AX is a square matrix, we have for the (i,j) term in $\partial\,\mathrm{tr}(AX)/\partial X$, $\partial\sum_i\sum_k a_{ik}x_{ki}/\partial x_{ij} = a_{ji}$, and so $\partial\,\mathrm{tr}(AX)/\partial X = A^{\mathrm{T}}$, and likewise, inspecting $\partial\sum_i\sum_k x_{ik}x_{ki}/\partial x_{ij}$, we get $\partial\,\mathrm{tr}(X^{\mathrm{T}}X)/\partial X = 2X^{\mathrm{T}}$. Likewise for the scalar-valued $a^{\mathrm{T}}Xb$, where a and b are conformable constant vectors, for $\partial\sum_m(\sum_k a_k x_{km})b_m/\partial x_{ij} = a_i b_j$, so $\partial a^{\mathrm{T}}Xb/\partial X = ab^{\mathrm{T}}$.

Now consider $\partial|X|/\partial X$. Using an expansion in cofactors, the only term in $|X|$ that involves x_{ij} is $x_{ij}(-1)^{i+j}|X_{-(i)(j)}|$, and the cofactor $(x_{(ij)}) =$

$(-1)^{i+j}|X_{-(i)(j)}|$ does not involve x_{ij}. Hence, $\partial|X|/\partial x_{ij} = (x_{(ij)})$, and so $\partial|X|/\partial X = (\mathrm{adj}(X))^{\mathrm{T}}$. We can write this as $\partial|X|/\partial X = |X|X^{-\mathrm{T}}$.

The chain rule can be used to evaluate $\partial \log|X|/\partial X$.

Applying the rule stated at the beginning of this section, we see that the derivative of a matrix Y with respect to the matrix X is

$$\frac{\mathrm{d}Y}{\mathrm{d}X} = Y \otimes \frac{\mathrm{d}}{\mathrm{d}X}. \tag{D.105}$$

Table D.3 lists some formulas for the matrix derivatives of some common expressions. The derivatives shown in Table D.3 can be obtained by evaluating expression (D.105), possibly also using the chain rule.

Table D.3. Formulas for Some Matrix Derivatives

General X

$f(X)$	$\partial f/\partial X$
$a^{\mathrm{T}}Xb$	ab^{T}
$\mathrm{tr}(AX)$	A^{T}
$\mathrm{tr}(X^{\mathrm{T}}X)$	$2X^{\mathrm{T}}$
BX	$I_n \otimes B$
XC	$C^{\mathrm{T}} \otimes I_m$
BXC	$C^{\mathrm{T}} \otimes B$

Square and Possibly Invertible X

$f(X)$	$\partial f/\partial X$
$\mathrm{tr}(X)$	I_n
$\mathrm{tr}(X^k)$	kX^{k-1}
$\mathrm{tr}(BX^{-1}C)$	$-(X^{-1}CBX^{-1})^{\mathrm{T}}$
$\lvert X\rvert$	$\lvert X\rvert X^{-\mathrm{T}}$
$\log\lvert X\rvert$	$X^{-\mathrm{T}}$
$\lvert X\rvert^k$	$k\lvert X\rvert^k X^{-\mathrm{T}}$
$BX^{-1}C$	$-(X^{-1}C)^{\mathrm{T}} \otimes BX^{-1}$

In this table, X is an $n \times m$ matrix, a is a constant n-vector, b is a constant m-vector, A is a constant $m \times n$ matrix, B is a constant $p \times n$ matrix, and C is a constant $m \times q$ matrix.

There are some interesting applications of differentiation with respect to a matrix in maximum likelihood estimation. Depending on the structure of the parameters in the distribution, derivatives of various types of objects may

be required. For example, the determinant of a variance-covariance matrix, in the sense that it is a measure of a volume, often occurs as a normalizing factor in a probability density function; therefore, we often encounter the need to differentiate a determinant with respect to a matrix.

Optimization of Functions

Because a derivative measures the rate of change of a function, a point at which the derivative is equal to 0 is a stationary point, which may be a maximum or a minimum of the function. Differentiation is therefore a very useful tool for finding the optima of functions, and so, for a given function $f(x)$, the gradient vector function, $\mathrm{g}_f(x)$, and the Hessian matrix function, $\mathrm{H}_f(x)$, play important roles in optimization methods.

We may seek either a maximum or a minimum of a function. Since maximizing the scalar function $f(x)$ is equivalent to minimizing $-f(x)$, we can always consider optimization of a function to be minimization of a function. Thus, we generally use terminology for the problem of finding a minimum of a function. Because the function may have many ups and downs, we often use the phrase *local minimum* (or local maximum or local optimum).

Except in the very simplest of cases, the optimization method must be iterative, moving through a sequence of points, $x^{(0)}, x^{(1)}, x^{(2)}, \ldots$, that approaches the optimum point arbitrarily closely. At the point $x^{(k)}$, the direction of *steepest descent* is clearly $-\mathrm{g}_f(x^{(k)})$, but because this direction may be continuously changing, the steepest descent direction may not be the best direction in which to seek the next point, $x^{(k+1)}$.

In the following subsection we describe some specific methods of optimization in the context of vector/matrix differentiation. We will discuss optimization in somewhat more detail in Section D.5.

Stationary Points of Functions

The first derivative helps only in finding a stationary point. The matrix of second derivatives, the Hessian, provides information about the nature of the stationary point, which may be a local minimum or maximum, a saddlepoint, or only an inflection point.

The so-called second-order optimality conditions are the following (see a general text on optimization for their proofs).

- If (but not only if) the stationary point is a local minimum, then the Hessian is nonnegative definite.
- If the Hessian is positive definite, then the stationary point is a local minimum.
- Likewise, if the stationary point is a local maximum, then the Hessian is nonpositive definite, and if the Hessian is negative definite, then the stationary point is a local maximum.

- If the Hessian has both positive and negative eigenvalues, then the stationary point is a saddlepoint.

Newton's Method

We consider a differentiable scalar-valued function of a vector argument, $f(x)$. By a Taylor series about a stationary point x_*, truncated after the second-order term

$$f(x) \approx f(x_*) + (x - x_*)^{\mathrm{T}} \mathrm{g}_f(x_*) + \frac{1}{2}(x - x_*)^{\mathrm{T}} \mathrm{H}_f(x_*)(x - x_*), \tag{D.106}$$

because $\mathrm{g}_f(x_*) = 0$, we have a general method of finding a stationary point for the function $f(\cdot)$, called Newton's method. If x is an m-vector, $\mathrm{g}_f(x)$ is an m-vector and $\mathrm{H}_f(x)$ is an $m \times m$ matrix.

Newton's method is to choose a starting point $x^{(0)}$, then, for $k = 0, 1, \ldots$, to solve the linear systems

$$\mathrm{H}_f(x^{(k)})p^{(k+1)} = -\mathrm{g}_f(x^{(k)}) \tag{D.107}$$

for $p^{(k+1)}$, and then to update the point in the domain of $f(\cdot)$ by

$$x^{(k+1)} = x^{(k)} + p^{(k+1)}. \tag{D.108}$$

The two steps are repeated until there is essentially no change from one iteration to the next. If $f(\cdot)$ is a quadratic function, the solution is obtained in one iteration because equation (D.106) is exact. These two steps have a very simple form for a function of one variable.

Linear Least Squares

In a least squares fit of a linear model

$$y = X\beta + \epsilon, \tag{D.109}$$

where y is an n-vector, X is an $n \times m$ matrix, and β is an m-vector, we replace β by a variable b, define the residual vector

$$r = y - Xb, \tag{D.110}$$

and minimize its Euclidean norm,

$$f(b) = r^{\mathrm{T}} r, \tag{D.111}$$

with respect to the variable b. We can solve this optimization problem by taking the derivative of this sum of squares and equating it to zero. Doing this, we get

$$\begin{aligned}\frac{\mathrm{d}(y-Xb)^{\mathrm{T}}(y-Xb)}{\mathrm{d}b} &= \frac{\mathrm{d}(y^{\mathrm{T}}y - 2b^{\mathrm{T}}X^{\mathrm{T}}y + b^{\mathrm{T}}X^{\mathrm{T}}Xb)}{\mathrm{d}b} \\ &= -2X^{\mathrm{T}}y + 2X^{\mathrm{T}}Xb \\ &= 0,\end{aligned}$$

which yields the normal equations

$$X^{\mathrm{T}}Xb = X^{\mathrm{T}}y.$$

The solution to the normal equations is a stationary point of the function (D.111). The Hessian of $(y-Xb)^{\mathrm{T}}(y-Xb)$ with respect to b is $2X^{\mathrm{T}}X$ and

$$X^{\mathrm{T}}X \succeq 0.$$

Because the matrix of second derivatives is nonnegative definite, the value of b that solves the system of equations arising from the first derivatives is a local minimum of equation (D.111).

Quasi-Newton Methods

All gradient-descent methods determine the path $p^{(k)}$ to take in the k^{th} step by a system of equations of the form

$$R^{(k)}p^{(k)} = -\mathrm{g}_f\big(x^{(k-1)}\big).$$

In the steepest-descent method, $R^{(k)}$ is the identity, I, in these equations. For functions with eccentric contours, the steepest-descent method traverses a zigzag path to the minimum. In Newton's method, $R^{(k)}$ is the Hessian evaluated at the previous point, $\mathrm{H}_f\big(x^{(k-1)}\big)$, which results in a more direct path to the minimum. Aside from the issues of consistency of the resulting equation and the general problems of reliability, a major disadvantage of Newton's method is the computational burden of computing the Hessian, which requires $\mathrm{O}(m^2)$ function evaluations, and solving the system, which requires $\mathrm{O}(m^3)$ arithmetic operations, at each iteration.

Instead of using the Hessian at each iteration, we may use an approximation, $B^{(k)}$. We may choose approximations that are simpler to update and/or that allow the equations for the step to be solved more easily. Methods using such approximations are called *quasi-Newton* methods or *variable metric* methods.

Because

$$\mathrm{H}_f\big(x^{(k)}\big)\big(x^{(k)} - x^{(k-1)}\big) \approx \mathrm{g}_f\big(x^{(k)}\big) - \mathrm{g}_f\big(x^{(k-1)}\big),$$

we choose $B^{(k)}$ so that

$$B^{(k)}\big(x^{(k)} - x^{(k-1)}\big) = \mathrm{g}_f\big(x^{(k)}\big) - \mathrm{g}_f\big(x^{(k-1)}\big). \tag{D.112}$$

This is called the *secant condition*.

We express the secant condition as

$$B^{(k)}s^{(k)} = y^{(k)}, \tag{D.113}$$

where

$$s^{(k)} = x^{(k)} - x^{(k-1)}$$

and

$$y^{(k)} = \mathrm{g}_f(x^{(k)}) - \mathrm{g}_f(x^{(k-1)}),$$

as above.

The system of equations in (D.113) does not fully determine $B^{(k)}$ of course. Because $B^{(k)}$ should approximate the Hessian, we may require that it be symmetric and positive definite.

The most common approach in quasi-Newton methods is first to choose a reasonable starting matrix $B^{(0)}$ and then to choose subsequent matrices by additive updates,

$$B^{(k+1)} = B^{(k)} + B_a^{(k)}, \tag{D.114}$$

subject to preservation of symmetry and positive definiteness. An approximate Hessian $B^{(k)}$ may be used for several iterations before it is updated; that is, $B_a^{(k)}$ may be taken as 0 for several successive iterations.

Multiparameter Likelihood Functions

For a sample $y = (y_1, \ldots, y_n)$ from a probability distribution with probability density function $p(\cdot; \theta)$, the *likelihood function* is

$$L(\theta; y) = \prod_{i=1}^{n} p(y_i; \theta), \tag{D.115}$$

and the *log-likelihood function* is $l(\theta; y) = \log(L(\theta; y))$. It is often easier to work with the log-likelihood function.

The log-likelihood is an important quantity in information theory and in unbiased estimation. If Y is a random variable with the given probability density function with the r-vector parameter θ, the *Fisher information* matrix that Y contains about θ is the $r \times r$ matrix

$$I(\theta) = \mathrm{Cov}_\theta \left(\frac{\partial l(t, Y)}{\partial t_i}, \frac{\partial l(t, Y)}{\partial t_j} \right), \tag{D.116}$$

where Cov_θ represents the variance-covariance matrix of the functions of Y formed by taking expectations for the given θ. (I use different symbols here because the derivatives are taken with respect to a *variable*, but the θ in Cov_θ cannot be the variable of the differentiation. This distinction is somewhat pedantic, and sometimes I follow the more common practice of using the same symbol in an expression that involves both Cov_θ and $\partial l(\theta, Y)/\partial \theta_i$.)

For example, if the distribution is the d-variate normal distribution with mean d-vector μ and $d \times d$ positive definite variance-covariance matrix Σ, the likelihood, equation (D.115), is

$$L(\mu, \Sigma; y) = \frac{1}{\left((2\pi)^{d/2}|\Sigma|^{1/2}\right)^n} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^{\mathrm{T}}\Sigma^{-1}(y_i - \mu)\right).$$

(Note that $|\Sigma|^{1/2} = |\Sigma^{\frac{1}{2}}|$. The square root matrix $\Sigma^{\frac{1}{2}}$ is often useful in transformations of variables.)

Anytime we have a quadratic form that we need to simplify, we should recall the useful fact: $x^{\mathrm{T}}Ax = \mathrm{tr}(Axx^{\mathrm{T}})$. Using this, and because, as is often the case, the log-likelihood is easier to work with, we write

$$l(\mu, \Sigma; y) = c - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\sum_{i=1}^{n}(y_i - \mu)(y_i - \mu)^{\mathrm{T}}\right), \qquad \text{(D.117)}$$

where we have used c to represent the constant portion. Next, we use the "Pythagorean equation" on the outer product to get

$$l(\mu, \Sigma; y) = c - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})^{\mathrm{T}}\right) - \frac{n}{2}\mathrm{tr}\left(\Sigma^{-1}(\bar{y} - \mu)(\bar{y} - \mu)^{\mathrm{T}}\right). \qquad \text{(D.118)}$$

In maximum likelihood estimation, we seek the maximum of the likelihood function (D.115) with respect to θ while we consider y to be fixed. If the maximum occurs within an open set and if the likelihood is differentiable, we might be able to find the maximum likelihood estimates by differentiation. In the log-likelihood for the d-variate normal distribution, we consider the parameters μ and Σ to be variables. To emphasize that perspective, we replace the parameters μ and Σ by the variables $\hat{\mu}$ and $\widehat{\Sigma}$. Now, to determine the maximum, we could take derivatives with respect to $\hat{\mu}$ and $\widehat{\Sigma}$, set them equal to 0, and solve for the maximum likelihood estimates. Some subtle problems arise that depend on the fact that for any constant vector a and scalar b, $\Pr(a^{\mathrm{T}}X = b) = 0$, but we do not interpret the likelihood as a probability.

Often in working out maximum likelihood estimates, students immediately think of differentiating, setting to 0, and solving. As noted above, this requires that the likelihood function be differentiable, that it be concave, and that the maximum occur at an interior point of the parameter space. Keeping in mind exactly what the problem is — one of finding a maximum — often leads to the correct solution more quickly.

Vector Random Variables

The simplest kind of vector random variable is one whose elements are independent. Such random vectors are easy to work with because the elements

can be dealt with individually, but they have limited applications. More interesting random vectors have a multivariate structure that depends on the relationships of the distributions of the individual elements. The simplest nondegenerate multivariate structure is of second degree; that is, a covariance or correlation structure. The probability density of a random vector with a multivariate structure generally is best represented by using matrices. In the case of the multivariate normal distribution, the variances and covariances together with the means completely characterize the distribution. For example, the fundamental integral that is associated with the d-variate normal distribution, sometimes called Aitken's integral, equation (D.36) on page 362, provides that constant. The rank of the integral is the same as the rank of the integrand. ("Rank" is used here in the sense of "number of dimensions".) In this case, the integrand and the integral are scalars.

Equation (D.36) is a simple result that follows from the evaluation of the individual single integrals after making the change of variables $y_i = x_i - \mu_i$. If Σ^{-1} is positive definite, Aitken's integral can also be evaluated by writing $P^{\mathrm{T}}\Sigma^{-1}P = I$ for some nonsingular matrix P. Now, after the translation $y = x - \mu$, which leaves the integral unchanged, we make the linear change of variables $z = P^{-1}y$, with the associated Jacobian $|\det(P)|$. From $P^{\mathrm{T}}\Sigma^{-1}P = I$, we have $|\det(P)| = (\det(\Sigma))^{1/2} = |\Sigma|^{1/2}$ because the determinant is positive. Aitken's integral therefore is

$$\begin{aligned}\int_{\mathbb{R}^d} \mathrm{e}^{-y^{\mathrm{T}}\Sigma^{-1}y/2}\,\mathrm{d}y &= \int_{\mathbb{R}^d} \mathrm{e}^{-(Pz)^{\mathrm{T}}\Sigma^{-1}Pz/2}\,(\det(\Sigma))^{1/2}\mathrm{d}z \\ &= \int_{\mathbb{R}^d} \mathrm{e}^{-z^{\mathrm{T}}z/2}\,\mathrm{d}z\,(\det(\Sigma))^{1/2} \\ &= (2\pi)^{d/2}(\det(\Sigma))^{1/2}.\end{aligned}$$

The expected value of a function f of the vector-valued random variable X is

$$\mathrm{E}(f(X)) = \int_{D(X)} f(x)p_X(x)\,\mathrm{d}x, \tag{D.119}$$

where $D(X)$ is the support of the distribution, $p_X(x)$ is the probability density function evaluated at x, and $x\ \mathrm{d}x$ are dummy vectors whose elements correspond to those of X. Interpreting $\int_{D(X)} \mathrm{d}x$ as a nest of univariate integrals, the result of the integration of the vector $f(x)p_X(x)$ is clearly of the same type as $f(x)$. For example, if $f(x) = x$, the expectation is the mean, which is a vector. For the normal distribution, we have

$$\begin{aligned}\mathrm{E}(X) &= (2\pi)^{-d/2}|\Sigma|^{-1/2}\int_{\mathbb{R}^d} x\mathrm{e}^{-(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)/2}\,\mathrm{d}x \\ &= \mu.\end{aligned}$$

For the variance of the vector-valued random variable X,

$$\mathrm{V}(X),$$

the function f in expression (D.119) above is the matrix $(X - \mathrm{E}(X))(X - \mathrm{E}(X))^{\mathrm{T}}$, and the result is a matrix. An example is the normal variance:

$$\begin{aligned}\mathrm{V}(X) &= \mathrm{E}\left((X - \mathrm{E}(X))(X - \mathrm{E}(X))^{\mathrm{T}}\right) \\ &= (2\pi)^{-d/2}|\Sigma|^{-1/2}\int_{\mathbb{R}^d}\left((x-\mu)(x-\mu)^{\mathrm{T}}\right)\mathrm{e}^{-(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)/2}\,\mathrm{d}x \\ &= \Sigma.\end{aligned}$$

D.4.4 Least Squares Solutions of Overdetermined Linear Systems

Consider the linear system

$$x = Zb,$$

where x and Z are given, b is unknown, and $x \in \mathbb{R}^n$, $Z \in \mathbb{R}^{n\times p}$, and $b \in \mathbb{R}^p$.

Under certain conditions for x and Z, there is no value of b that makes the system a system of equations. We then write the system as

$$x \approx Zb,$$

or

$$x = Zb + r.$$

A common example in which there may not be a solution is when $n > p$. (In such a case, there will not be a solution is $\mathrm{rank}([Z|b]) > p$.) When there is no solution, we may seek a value of b such that $x - Zb$ is small. The most common definition of "small" in this setting is small in $\|x - Zb\|_2$. (We usually drop the subscript on the norm.) The solution in that case, that is, the value of b such that $\|x - Zb\|$ is minimized, is the "least squares" solution.

At the minimum of $\|x - Zb\|$, we have the "normal equations"

$$Z^{\mathrm{T}}Zb = Z^{\mathrm{T}}x.$$

The coefficient matrix in these equations has a special form; it is a *Gramian* matrix. A unique solution to these equations is

$$b = (Z^{\mathrm{T}}Z)^{+}Z^{\mathrm{T}}x;$$

that is, the solution arising from the Moore-Penrose inverse.

D.4.5 Linear Statistical Models

In statistical applications, we often assume that the distribution of some random variable depends in a linear fashion on some covariate. This setup leads to linear regression analysis and to the analysis of variance in general linear classification models.

Shao expresses the linear statistical model as

$$X = Z\beta + \epsilon,$$

with various assumptions about the distribution of ϵ.

In all useful assumptions about the distribution of ϵ, we have $\mathrm{E}(\epsilon) = 0$, so under any of the assumptions

$$\mathrm{E}(X) = Z\beta.$$

In statistical inference, we can think of β either as an unobservable random variable or as an unknown constant. If we think of it as an unknown constant and we want to determine a value of it that optimizes some objective function (such as a likelihood or a sum of squares), then we first must substitute a variable for the constant. (Although we often skip over this step, it is important conceptually.) In the context of the least squares discussion above, we may consider $\|x - Zb\|_2$, where the variable b is in place of the unknown model parameter β.

Shao uses $\widehat{\beta}$ to denote any solution to the normal equations formed from the linear system $X = Z\beta$, that is

$$\widehat{\beta} = (Z^{\mathrm{T}}Z)^{-}Z^{\mathrm{T}}X.$$

Notice that if Z is not of full rank, $\widehat{\beta}$ is not unique. (Other authors Use the notation $\widehat{\beta}$ to represent the unique solution to the normal equations arising from the Moore-Penrose inverse. We will discuss that particular solution below.)

Linear U-Estimability

One of the most important questions for statistical inference involves estimating or testing some linear combination of the elements of the parameter β; for example, we may wish to estimate $\beta_1 - \beta_2$ or to test the hypothesis that $\beta_1 - \beta_2 = c_1$ for some constant c_1. In general, we will consider the linear combination $l^{\mathrm{T}}\beta$. Whether or not it makes sense to estimate such a linear combination depends on whether there is a function of the observable random variable X such that $g(\mathrm{E}(X)) = l^{\mathrm{T}}\beta$.

We generally restrict our attention to linear functions of $\mathrm{E}(X)$ and formally define a linear combination $l^{\mathrm{T}}\beta$ to be (linearly) *U-estimable* if there exists a vector t such that

$$t^{\mathrm{T}}\mathrm{E}(X) = l^{\mathrm{T}}\beta$$

for any β.

It is clear that if X is of full column rank, $l^{\mathrm{T}}\beta$ is linearly estimable for any l or, more generally, $l^{\mathrm{T}}\beta$ is linearly estimable for any $l \in \mathrm{span}(Z^{\mathrm{T}})$. (The t vector is just the normalized coefficients expressing l in terms of the columns of Z.)

Estimability depends only on the simplest distributional assumption about the model; that is, that $\mathrm{E}(\epsilon) = 0$. Under this assumption, we see that the

estimator $\widehat{\beta}$ based on the least squares fit of β is unbiased for the linearly estimable function $l^{\mathrm{T}}\beta$. Because $l \in \mathrm{span}(Z^{\mathrm{T}}) = \mathrm{span}(Z^{\mathrm{T}}Z)$, we can write $l = Z^{\mathrm{T}}Z\tilde{t}$. Now, we have

$$\begin{aligned}\mathrm{E}(l^{\mathrm{T}}\widehat{\beta}) &= \mathrm{E}(l^{\mathrm{T}}(Z^{\mathrm{T}}Z)^{+}Z^{\mathrm{T}}X)\\ &= \tilde{t}^{\mathrm{T}}Z^{\mathrm{T}}Z(Z^{\mathrm{T}}Z)^{+}Z^{\mathrm{T}}Z\beta\\ &= \tilde{t}^{\mathrm{T}}Z^{\mathrm{T}}Z\beta\\ &= l^{\mathrm{T}}\beta.\end{aligned}$$

Although we have been taking $\widehat{\beta}$ to be $(Z^{\mathrm{T}}Z)^{+}Z^{\mathrm{T}}X$, the equations above follow for other least squares fits, $b = (Z^{\mathrm{T}}Z)^{-}Z^{\mathrm{T}}X$, for any generalized inverse. In fact, the estimator of $l^{\mathrm{T}}\beta$ is invariant to the choice of the generalized inverse. This is because if $b = (Z^{\mathrm{T}}Z)^{-}Z^{\mathrm{T}}X$, we have $Z^{\mathrm{T}}Zb = Z^{\mathrm{T}}X$, and so

$$l^{\mathrm{T}}\widehat{\beta} - l^{\mathrm{T}}b = \tilde{t}^{\mathrm{T}}Z^{\mathrm{T}}Z(\widehat{\beta} - b) = \tilde{t}^{\mathrm{T}}(Z^{\mathrm{T}}X - Z^{\mathrm{T}}X) = 0.$$

The Gauss-Markov Theorem

The Gauss-Markov theorem provides a restricted optimality property for estimators of estimable functions of β under the condition that $\mathrm{E}(\epsilon) = 0$ and $\mathrm{V}(\epsilon) = \sigma^2 I$; that is, in addition to the assumption of zero expectation, which we have used above, we also assume that the elements of ϵ have constant variance and that their covariances are zero. (We are not assuming independence or normality.)

Given $X = Z\beta + \epsilon$ and $\mathrm{E}(\epsilon) = 0$ and $\mathrm{V}(\epsilon) = \sigma^2 I$, the Gauss-Markov theorem states that $l^{\mathrm{T}}\widehat{\beta}$ is the unique *best linear unbiased estimator* (BLUE) of the estimable function $l^{\mathrm{T}}\beta$.

"Linear" estimator in this context means a linear combination of X; that is, an estimator in the form $a^{\mathrm{T}}X$. It is clear that $l^{\mathrm{T}}\widehat{\beta}$ is linear, and we have already seen that it is unbiased for $l^{\mathrm{T}}\beta$. "Best" in this context means that its variance is no greater than any other estimator that fits the requirements. Hence, to prove the theorem, first let $a^{\mathrm{T}}X$ be any unbiased estimator of $l^{\mathrm{T}}\beta$, and write $l = Z^{\mathrm{T}}Z\tilde{t}$ as above. Because $a^{\mathrm{T}}X$ is unbiased for any β, as we saw above, it must be the case that $a^{\mathrm{T}}Z = l^{\mathrm{T}}$. Recalling that $Z^{\mathrm{T}}Z\widehat{\beta} = Z^{\mathrm{T}}X$, we have

$$\begin{aligned}\mathrm{V}(a^{\mathrm{T}}X) &= \mathrm{V}(a^{\mathrm{T}}X - l^{\mathrm{T}}\widehat{\beta} + l^{\mathrm{T}}\widehat{\beta})\\ &= \mathrm{V}(a^{\mathrm{T}}X - \tilde{t}^{\mathrm{T}}Z^{\mathrm{T}}X + l^{\mathrm{T}}\widehat{\beta})\\ &= \mathrm{V}(a^{\mathrm{T}}X - \tilde{t}^{\mathrm{T}}Z^{\mathrm{T}}X) + \mathrm{V}(l^{\mathrm{T}}\widehat{\beta}) + 2\mathrm{Cov}(a^{\mathrm{T}}X - \tilde{t}^{\mathrm{T}}Z^{\mathrm{T}}X,\ \tilde{t}^{\mathrm{T}}Z^{\mathrm{T}}X).\end{aligned}$$

Now, under the assumptions on the variance-covariance matrix of ϵ, which is also the (conditional, given Z) variance-covariance matrix of X, we have

$$\begin{aligned}\mathrm{Cov}(a^{\mathrm{T}}X - \tilde{t}^{\mathrm{T}}Z^{\mathrm{T}}X,\ l^{\mathrm{T}}\widehat{\beta}) &= (a^{\mathrm{T}} - \tilde{t}^{\mathrm{T}}Z^{\mathrm{T}})\sigma^2 I Z\tilde{t} \\ &= (a^{\mathrm{T}}Z - \tilde{t}^{\mathrm{T}}Z^{\mathrm{T}}Z)\sigma^2 I\tilde{t} \\ &= (l^{\mathrm{T}} - l^{\mathrm{T}})\sigma^2 I\tilde{t} \\ &= 0;\end{aligned}$$

that is,

$$\mathrm{V}(a^{\mathrm{T}}X) = \mathrm{V}(a^{\mathrm{T}}X - \tilde{t}^{\mathrm{T}}Z^{\mathrm{T}}X) + \mathrm{V}(l^{\mathrm{T}}\widehat{\beta}).$$

This implies that

$$\mathrm{V}(a^{\mathrm{T}}X) \geq \mathrm{V}(l^{\mathrm{T}}\widehat{\beta});$$

that is, $l^{\mathrm{T}}\widehat{\beta}$ has minimum variance among the linear unbiased estimators of $l^{\mathrm{T}}\beta$. To see that it is unique, we consider the case in which $\mathrm{V}(a^{\mathrm{T}}X) = \mathrm{V}(l^{\mathrm{T}}\widehat{\beta})$; that is, $\mathrm{V}(a^{\mathrm{T}}X - \tilde{t}^{\mathrm{T}}Z^{\mathrm{T}}X) = 0$. For this variance to equal 0, it must be the case that $a^{\mathrm{T}} - \tilde{t}^{\mathrm{T}}Z^{\mathrm{T}} = 0$ or $a^{\mathrm{T}}X = \tilde{t}^{\mathrm{T}}Z^{\mathrm{T}}X = l^{\mathrm{T}}\widehat{\beta}$; that is, $l^{\mathrm{T}}\widehat{\beta}$ is the unique linear unbiased estimator that achieves the minimum variance.

If we assume further that $\epsilon \sim \mathrm{N}_n(0, \sigma^2 I)$, we can show that $l^{\mathrm{T}}\widehat{\beta}$ is the uniformly minimum variance unbiased estimator (UMVUE) for $l^{\mathrm{T}}\beta$. This is because $(Z^{\mathrm{T}}X,\ (X - Z\widehat{\beta})^{\mathrm{T}}(X - Z\widehat{\beta}))$ is complete and sufficient for (β, σ^2). This line of reasoning also implies that $(X - Z\widehat{\beta})^{\mathrm{T}}(X - Z\widehat{\beta})/(n - r)$, where $r = \mathrm{rank}(Z)$, is UMVUE for σ^2.

Optimal Properties of the Moore-Penrose Inverse

The solution corresponding to the Moore-Penrose inverse is unique because that generalized inverse is unique.

That solution is interesting for another reason, however: the $\widehat{\beta}$ from the Moore-Penrose inverse has the minimum L_2-norm of all solutions.

To see that this solution has minimum norm, first factor Z, as

$$Z = QRU^{\mathrm{T}},$$

and form the Moore-Penrose inverse as

$$Z^{+} = U \begin{bmatrix} R_1^{-1} & 0 \\ 0 & 0 \end{bmatrix} Q^{\mathrm{T}}.$$

Now let

$$\widehat{\beta} = Z^{+}X.$$

This is a least squares solution (that is, we have chosen a specific least squares solution).

Now, let

$$Q^{\mathrm{T}}X = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix},$$

where c_1 has exactly r elements and c_2 has $n - r$ elements, and let

$$U^{\mathrm{T}}b = \begin{pmatrix} t_1 \\ t_2 \end{pmatrix},$$

where b is the variable in the norm $\|X - Zb\|_2$ that we seek to minimize, and where t_1 has r elements.

Because multiplication by an orthogonal matrix does not change the norm, we have

$$\begin{aligned} \|X - Zb\|_2 &= \|Q^{\mathrm{T}}(X - ZUU^{\mathrm{T}}b)\|_2 \\ &= \left\| \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} - \begin{bmatrix} R_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{pmatrix} t_1 \\ t_2 \end{pmatrix} \right\|_2 \\ &= \left\| \begin{pmatrix} c_1 - R_1 t_1 \\ c_2 \end{pmatrix} \right\|_2. \end{aligned}$$

The residual norm is minimized for $t_1 = R_1^{-1}c_1$ and t_2 arbitrary. However, if $t_2 = 0$, then $\|t\|_2$ is also minimized. Because $U^{\mathrm{T}}b = t$ and U is orthogonal, $\|b\|_2 = \|t\|_2 = \|t_1\|_2 + \|t_2\|_2$, and so with $t_2 = 0$, that is, with $b = \widehat{\beta}$, $\|\widehat{\beta}\|_2$ is the minimum among the norms of all least squares solutions, $\|b\|_2$.

D.4.6 Cochran's Theorem

There are various facts that are sometimes called *Cochran's theorem*. The simplest one concerns k symmetric idempotent $n \times n$ matrices, $A_1, \ldots, A_k$, such that

$$I_n = A_1 + \cdots + A_k.$$

Under these conditions, we have

$$A_i A_j = 0 \text{ for all } i \neq j.$$

Proof

For an arbitrary j, for some matrix V, we have

$$V^{\mathrm{T}} A_j V = \mathrm{diag}(I_r, 0),$$

where $r = \mathrm{rank}(A_j)$. Now

$$\begin{aligned} I_n &= V^{\mathrm{T}} I_n V \\ &= \sum_{i=1}^{k} V^{\mathrm{T}} A_i V \\ &= \mathrm{diag}(I_r, 0) + \sum_{i \neq j} V^{\mathrm{T}} A_i V, \end{aligned}$$

which implies

$$\sum_{i \neq j} V^{\mathrm{T}} A_i V = \mathrm{diag}(0, I_{n-r}).$$

Now for each i, $V^{\mathrm{T}} A_i V$ is idempotent, and because the diagonal elements of a symmetric idempotent matrix are all nonnegative, and hence the equation implies implies that for each $i \neq j$, the first r diagonal elements are 0. Furthermore, since these diagonal elements are 0, all elements in the first r rows and columns are 0. We have, therefore, for each $i \neq j$,

$$V^{\mathrm{T}} A_i V = \mathrm{diag}(0, B_i)$$

for some $(n-r) \times (n-r)$ symmetric idempotent matrix B_i. Now, for any $i \neq j$, consider $A_i A_j$ and form $V^{\mathrm{T}} A_i A_j V$. We have

$$\begin{aligned} V^{\mathrm{T}} A_i A_j V &= (V^{\mathrm{T}} A_i V)(V^{\mathrm{T}} A_j V) \\ &= \mathrm{diag}(0, B_i)\mathrm{diag}(I_r, 0) \\ &= 0. \end{aligned}$$

Because V is nonsingular, this implies the desired conclusion; that is, that $A_i A_j = 0$ for any $i \neq j$.

We can now extend this result to an idempotent matrix in place of I; that is, for an idempotent matrix A with $A = A_1 + \cdots + A_k$. Let $A_1, \ldots, A_k$ be $n \times n$ symmetric matrices and let

$$A = A_1 + \cdots + A_k.$$

Then any two of the following conditions imply the third one:

(a). A is idempotent.
(b). A_i is idempotent for $i = 1, \ldots, k$.
(c). $A_i A_j = 0$ for all $i \neq j$.

This is also called *Cochran's theorem*. (The theorem also applies to nonsymmetric matrices if condition (c) is augmented with the requirement that $\mathrm{rank}(A_i^2) = \mathrm{rank}(A_i)$ for all i. We will restrict our attention to symmetric matrices, however, because in most applications of these results, the matrices are symmetric.)

First, if we assume properties (a) and (b), we can show that property (c) follows for the special case $A = I$.

Now, let us assume properties (b) and (c) and show that property (a) holds. With properties (b) and (c), we have

$$\begin{aligned} AA &= (A_1 + \cdots + A_k)(A_1 + \cdots + A_k) \\ &= \sum_{i=1}^{k} A_i A_i + \sum_{i \neq j} \sum_{j=1}^{k} A_i A_j \\ &= \sum_{i=1}^{k} A_i \\ &= A. \end{aligned}$$

Hence, we have property (a); that is, A is idempotent.

Finally, let us assume properties (a) and (c). Property (b) follows immediately from

$$A_i^2 = A_i A_i = A_i A = A_i A A = A_i^2 A = A_i^3$$

and the fact that $A^{p+1} = A^p \Longrightarrow A$ is idempotent.

Any two of the properties (a) through (c) also imply a fourth property for $A = A_1 + \cdots + A_k$ when the A_i are symmetric:

(d). $\text{rank}(A) = \text{rank}(A_1) + \cdots + \text{rank}(A_k)$.

We first note that any two of properties (a) through (c) imply the third one, so we will just use properties (a) and (b). Property (a) gives

$$\text{rank}(A) = \text{tr}(A) = \text{tr}(A_1 + \cdots + A_k) = \text{tr}(A_1) + \cdots + \text{tr}(A_k),$$

and property (b) states that the latter expression is $\text{rank}(A_1) + \cdots + \text{rank}(A_k)$, thus yielding property (d).

There is also a partial converse: properties (a) and (d) imply the other properties.

One of the most important special cases of Cochran's theorem is when $A = I$ in the sum:

$$I_n = A_1 + \cdots + A_k.$$

The identity matrix is idempotent, so if $\text{rank}(A_1) + \cdots + \text{rank}(A_k) = n$, all the properties above hold. (See Gentle, 2007, pages 283–285.)

The most important statistical application of Cochran's theorem is for the distribution of quadratic forms of normally distributed random vectors.

In applications of linear models, a quadratic form involving Y is often partitioned into a sum of quadratic forms. Assume that Y is distributed as $\text{N}_d(\mu, I_d)$, and for $i = 1, \ldots k$, let A_i be a $d \times d$ symmetric matrix with rank r_i such that $\sum_i A_i = I_d$. This yields a partition of the total sum of squares $Y^{\text{T}}Y$ into k components:

$$Y^{\text{T}}Y = Y^{\text{T}}A_1 Y + \cdots + Y^{\text{T}}A_k Y.$$

One of the most important results in the analysis of linear models states that the $Y^{\text{T}}A_i Y$ have independent noncentral chi-squared distributions $\chi^2_{r_i}(\delta_i)$ with $\delta_i = \mu^{\text{T}}A_i\mu$ if and only if $\sum_i r_i = d$.

This distribution result is also called Cochran's theorem, and it is implied by the results above. (See Gentle, 2007, pages 324–325.)

D.4.7 Transition Matrices

An important use of matrices in statistics is in models of transitions of a stochastic process from one state to another. In a discrete-state Markov chain, for example, the probability of going from state j to state i may be represented as elements of a *transition matrix*, which can any square matrix with

nonnegative elements and such that the sum of the elements in any column is 1. Any square matrix with nonnegative elements whose columns each sum to 1 is called a *right stochastic matrix*.

(Note that many people who work with Markov chains define the transition matrix as the transpose of K above. This is not a good idea, because in applications with state vectors, the state vectors would naturally have to be row vectors. Until about the middle of the twentieth century, many mathematicians thought of vectors as row vectors; that is, a system of linear equations would be written as $xA = b$. Nowadays, almost all mathematicians think of vectors as column vectors in matrix algebra. Even in some of my previous writings, e.g., Gentle, 2007, I have called the transpose of K the transition matrix, and I defined a stochastic matrix in terms of the transpose. I think that it is time to adopt a notation that is more consistent with current matrix/vector notation. This is merely a change in notation; no concepts require any change.)

There are various properties of transition matrices that are important for studying Markov chains.

Irreducible Matrices

Any nonnegative square matrix that can be permuted into the form

$$\begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix}$$

with square diagonal submatrices is said to be *reducible*; a matrix that cannot be put into that form is *irreducible*. An alternate term for reducible is *decomposable*, with the associated term *indecomposable*.

We see from the definition that a positive matrix is irreducible.

We now consider irreducible square nonnegative matrices. This class includes positive matrices.

Irreducible matrices have several interesting properties. An $n \times n$ nonnegative matrix A is irreducible if and only if $(I + A)^{n-1}$ is a positive matrix; that is,

$$A \text{ is irreducible} \iff (I + A)^{n-1} > 0. \tag{D.120}$$

To see this, first assume $(I+A)^{n-1} > 0$; thus, $(I+A)^{n-1}$ clearly is irreducible. If A is reducible, then there exists a permutation matrix E_π such that

$$E_\pi^{\mathrm{T}} A E_\pi = \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix},$$

and so

$$\begin{aligned} E_\pi^{\mathrm{T}}(I + A)^{n-1} E_\pi &= \left(E_\pi^{\mathrm{T}}(I + A)E_\pi\right)^{n-1} \\ &= \left(I + E_\pi^{\mathrm{T}} A E_\pi\right)^{n-1} \\ &= \begin{bmatrix} I_{n_1} + B_{11} & B_{12} \\ 0 & I_{n_2} + B_{22} \end{bmatrix}. \end{aligned}$$

This decomposition of $(I+A)^{n-1}$ cannot exist because it is irreducible; hence we conclude A is irreducible if $(I+A)^{n-1} > 0$. We can see that $(I+A)^{n-1}$ must be a positive matrix by first observing that the $(i,j)^{\text{th}}$ element of $(I+A)^{n-1}$ can be expressed as

$$\left((I+A)^{n-1}\right)_{ij} = \left(\sum_{k=0}^{n-1} \binom{n-1}{k} A^k\right)_{ij}. \tag{D.121}$$

Hence, for $k = 1, \ldots, n-1$, we consider the $(i,j)^{\text{th}}$ entry of A^k. Let $a_{ij}^{(k)}$ represent this quantity.

Given any pair (i,j), for some $l_1, l_2, \ldots, l_{k-1}$, we have

$$a_{ij}^{(k)} = \sum_{l_1, l_2, \ldots, l_{k-1}} a_{1l_1} a_{l_1 l_2} \cdots a_{l_{k-1} j}.$$

Now $a_{ij}^{(k)} > 0$ if and only if $a_{1l_1}, a_{l_1 l_2}, \ldots, a_{l_{k-1} j}$ are all positive; that is, if there is a path $v_1, v_{l_1}, \ldots, v_{l_{k-1}}, v_j$ in $\mathcal{G}(A)$. If A is irreducible, then $\mathcal{G}(A)$ is strongly connected, and hence the path exists. So, for any pair (i,j), we have from equation (D.121) $\left((I+A)^{n-1}\right)_{ij} > 0$; that is, $(I+A)^{n-1} > 0$.

The positivity of $(I+A)^{n-1}$ for an irreducible nonnegative matrix A is a very useful property because it allows us to extend some conclusions of the Perron theorem to irreducible nonnegative matrices.

Properties of Square Irreducible Nonnegative Matrices; the Perron-Frobenius Theorem

If A is a square irreducible nonnegative matrix, then we have the following properties. These following properties are the conclusions of the *Perron-Frobenius theorem*.

1. $\rho(A)$ is an eigenvalue of A. This eigenvalue is called the *Perron root*, as before.
2. The Perron root $\rho(A)$ is simple. (That is, the algebraic multiplicity of the Perron root is 1.)
3. The dimension of the eigenspace of the Perron root is 1. (That is, the geometric multiplicity of $\rho(A)$ is 1.)
4. The eigenvector associated with $\rho(A)$ is positive. This eigenvector is called the *Perron vector*, as before.

The relationship (D.120) allows us to prove properties 1 and 4.

The one property of square positive matrices that does not carry over to square irreducible nonnegative matrices is that $r = \rho(A)$ is the only eigenvalue on the spectral circle of A. For example, the small irreducible nonnegative matrix

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

has eigenvalues 1 and -1, and so both are on the spectral circle.

It turns out, however, that square irreducible nonnegative matrices that have only one eigenvalue on the spectral circle also have other interesting properties that are important, for example, in Markov chains. We therefore give a name to the property:

A square irreducible nonnegative matrix is said to be *primitive* if it has only one eigenvalue on the spectral circle.

In modeling with Markov chains and other applications, the limiting behavior of A^k is an important property.

If A is a primitive matrix, then we have the useful result

$$\lim_{k\to\infty}\left(\frac{A}{\rho(A)}\right)^k = vw^{\mathrm{T}}, \tag{D.122}$$

where v is an eigenvector of A associated with $\rho(A)$ and w is an eigenvector of A^{T} associated with $\rho(A)$, and w and v are scaled so that $w^{\mathrm{T}}v = 1$. (Such eigenvectors exist because $\rho(A)$ is a simple eigenvalue. They also exist because they are both positive. Note that A is not necessarily symmetric, and so its eigenvectors may include imaginary components; however, the eigenvectors associated with $\rho(A)$ are real, and so we can write w^{T} instead of w^{H}.)

To see equation (D.122), we consider $\left(A - \rho(A)vw^{\mathrm{T}}\right)$. First, if (c_i, v_i) is an eigenpair of $\left(A - \rho(A)vw^{\mathrm{T}}\right)$ and $c_i \neq 0$, then (c_i, v_i) is an eigenpair of A. We can see this by multiplying both sides of the eigen-equation by vw^{T}:

$$\begin{aligned} c_i vw^{\mathrm{T}}v_i &= vw^{\mathrm{T}}\left(A - \rho(A)vw^{\mathrm{T}}\right)v_i \\ &= \left(vw^{\mathrm{T}}A - \rho(A)vw^{\mathrm{T}}vw^{\mathrm{T}}\right)v_i \\ &= \left(\rho(A)vw^{\mathrm{T}} - \rho(A)vw^{\mathrm{T}}\right)v_i \\ &= 0; \end{aligned}$$

hence,

$$\begin{aligned} Av_i &= \left(A - \rho(A)vw^{\mathrm{T}}\right)v_i \\ &= c_i v_i. \end{aligned}$$

Next, we show that

$$\rho\left(A - \rho(A)vw^{\mathrm{T}}\right) < \rho(A). \tag{D.123}$$

If $\rho(A)$ were an eigenvalue of $\left(A - \rho(A)vw^{\mathrm{T}}\right)$, then its associated eigenvector, say w, would also have to be an eigenvector of A, as we saw above. But since as an eigenvalue of A the geometric multiplicity of $\rho(A)$ is 1, for some scalar s, $w = sv$. But this is impossible because that would yield

$$\begin{aligned} \rho(A)sv &= \left(A - \rho(A)vw^{\mathrm{T}}\right)sv \\ &= sAv - s\rho(A)v \\ &= 0, \end{aligned}$$

and neither $\rho(A)$ nor sv is zero. But as we saw above, any eigenvalue of $\left(A - \rho(A)vw^{\mathrm{T}}\right)$ is an eigenvalue of A and no eigenvalue of $\left(A - \rho(A)vw^{\mathrm{T}}\right)$ can be as large as $\rho(A)$ in modulus; therefore we have inequality (D.123).

Finally, with w and v as defined above, and with the eigenvalue $\rho(A)$,

$$\left(A - \rho(A)vw^{\mathrm{T}}\right)^k = A^k - (\rho(A))^k vw^{\mathrm{T}}, \tag{D.124}$$

for $k = 1, 2, \ldots$.

Dividing both sides of equation (D.124) by $(\rho(A))^k$ and rearranging terms, we have

$$\left(\frac{A}{\rho(A)}\right)^k = vw^{\mathrm{T}} + \frac{\left(A - \rho(A)vw^{\mathrm{T}}\right)}{\rho(A)}. \tag{D.125}$$

Now

$$\rho\left(\frac{\left(A - \rho(A)vw^{\mathrm{T}}\right)}{\rho(A)}\right) = \frac{\rho\left(A - \rho(A)vw^{\mathrm{T}}\right)}{\rho(A)},$$

which is less than 1; hence, we have

$$\lim_{k\to\infty}\left(\frac{\left(A - \rho(A)vw^{\mathrm{T}}\right)}{\rho(A)}\right)^k = 0;$$

so, taking the limit in equation (D.125), we have equation (D.122).

Applications of the Perron-Frobenius theorem are far-ranging.

D.5 Optimization

Optimization problems — maximization or minimization — arise in many areas of statistics. Statistical estimation and modeling both are usually special types of optimization problems. In a common method of statistical estimation, we *maximize* a likelihood, which is a function proportional to a probability density at the point of the observed data. In another method of estimation and in standard modeling techniques, we *minimize* a norm of the residuals. The best fit of a model is often defined in terms of a minimum of a norm, such as least squares. Other uses of optimization in statistical applications occur prior to collection of data, for example, when we design an experiment or a survey so as to minimize experimental or sampling errors.

When a statistical method is based on the solution of an optimization problem, to formulate that problem unambiguously helps us both to understand the method and to decide whether the method is appropriate to the purposes for which it is applied.

Some of the simpler and more common optimization problems in statistics can be solved easily, often by solving a system of linear equations. Many other problems, however, do not have closed-form solutions, and the solutions must be approximated by iterative methods.

D.5.1 Overview of Optimization

Optimization means to find a **maximum** or a **maximum** of an **objective function**, $f : D \subset \mathbb{R}^d \mapsto \mathbb{R}$.

Local optimization means optimization within a some subset of the the domain of the objective function **Global** optimzation results in the optimum of all local optima.

In **unconstrained** optimization, we take all points in D to be feasible.

Important Properties of the Objective Function

- domain dense or not
- differentiable or not
  - to what order
  - easy or hard to compute
- concave (or convex) or neither
  - if neither, there may be **local** optima

In the following, let $f(x)$ be the objective function, and assume we want to maximize it.
(To minimize, $f(x) \leftarrow -f(x)$ and convex $\leftarrow$ concave.)

Methods

- **Analytic**: yields closed form for all local maxima.
- **Iterative**: for $k = 1, 2, \ldots$, given $x^{(k-1)}$ choose $x^{(k)}$ so that $f(x^{(k)}) \rightarrow$local maximum of f.
  We need
  - a **starting point:** $x^{(0)}$;
  - a method to **choose** $\tilde{x}$ with good prospects of being $x^{(k)}$;
  - a method to **decide** whether $\tilde{x}$ should be $x^{(k)}$.

  How we **choose** and **decide** determines the differences between optimization algorithms.

How to **choose** may be based on derivative information or on some systematic method of exploring the domain.

How to **decide** may be based on a deterministic criterion, such as requiring $f(x^{(k)}) > f(x^{(k-1)})$,
or the decision may be randomized.

Metamethods: General Tools

- Transformations (for either analytic or iterative methods).
- Any trick you can think of (for either analytic or iterative methods), e.g., alternating conditional optimization.
- Conditional bounding functions (for iterative methods).

Convergence of Iterative Algorithms

In an iterative algorithm, we have a sequence $\left\{\left(f\left(x^{(k)}\right), x^{(k)}\right)\right\}$.

The first question is *whether* the sequence converges to the correct solution.

If there is a unique maximum, and if x_* is the point at which the maximum occurs, the first question can be posed more precisely as, given ϵ_1 does there exist M_1 such that for $k > M_1$,

$$\left|f(x^{(k)}) - f(x_*)\right| < \epsilon_1;$$

or, alternatively, for a given ϵ_2 does there exist M_2 such that for $k > M_2$,

$$\left\|x^{(k)} - x_*\right\| < \epsilon_2.$$

Recall that $f : \mathbb{R}^d \mapsto \mathbb{R}$, so $|\cdot|$ above is the absolute value, while $\|\cdot\|$ is some kind of vector norm.

There are complications if x_* is not unique as the point at which the maximum occurs.

Similarly, there are comlications if x_* is merely a point of local maximum.

Assessing Convergence

In practice, we must *decide* when convergence has occurred; that is, whether the iterations have become close enough to the solution. Since we don't know the solution, we cannot base this decision on the convergence criteria above.

We put faith in our algorithm, and decide convergence has occurred if, for some $e_1, e_2 > 0$, either

$$|f(x^{(k)}) - f(x^{(k-1)})| \leq e_1$$

or

$$\|x^{(k)} - x^{(k-1)}\| \leq e_2,$$

or both.

Notice, that lacking any known value, we trust the algorithm to do the right thing; both $x^{(k)}$ and $x^{(k-1)}$ are just values in the algorithmic sequence. The fact that this particular sequence — or any sequence, even ones yielding nonincreasing function values — converges does not really get at the question of whether $x^{(k)} \to x_*$.

Note also the subtle change from "$<$" to "$\leq$".

For some special class of functions $\nabla f(x)$ may exist, and we may know that at the solution, $\nabla f(x_*) = 0$. In these cases, we may have another criterion for deciding convergence has occurred:

$$\|\nabla f(x_*)\| \leq e_3.$$

Rate of Convergence of Iterative Algorithms

If the answer to the first question is "yes", that is, if the algorithmic sequence converges, the next question is *how fast* the sequence converges. (We address this question assuming it converges to the correct solution.)

The rate of convergence is a measure of how fast the "error" decreases. Any of three quantities we mentioned in discussing convergence, $f(x^{(k)}) - f(x_*)$, $x^{(k)} - x_*$, or $\nabla f(x_*)$, could be taken to be the error. If we take

$$e_k = x^{(k)} - x_*$$

to be the error at step k, we might define the magnitude of the error as $\|e_k\|$ (for some norm $\|\cdot\|$). If the algorithm converges to the correct solution,

$$\lim_{k\to\infty} \|e_k\| = 0.$$

Our interest is in how fast $\|e_k\|$ decreases.

Sometimes there is no reasonable way of quantifying the rate at which this quantity decreases.

In the happy case (and a common case for simple algorithms), if there exist $r > 0$ and $c > 0$ such that

$$\lim_{k\to\infty} \frac{\|e_k\|}{\|e_{k-1}\|^r} = c,$$

we say the **rate of convergence** is r and the **rate constant** is c.

The Steps in Iterative Algorithms For a Special Class of Functions

The steps in iterative algorithms are often based on some analytic relationship between $f(x)$ and $f(x^{(k-1)})$. For a continuously differentiable function, the most common relationship is the Taylor series expansion:

$$\begin{aligned} f(x) = f(x^{(k-1)}) \; + \\ (x - x^{(k-1)})^{\mathrm{T}} \nabla f(x^{(k-1)}) \; + \\ \frac{1}{2}(x - x^{(k-1)})^{\mathrm{T}} \nabla^2 f(x^{(k-1)})(x - x^{(k-1)}) \; + \\ \cdots \end{aligned}$$

Note this limitation: "For a continuously differentiable function, ...".
We cannot use this method on just any old function.

In the following, we will consider only this restricted class of functions.

Steepest Ascent (Descent)

The steps are defined by truncating the Taylor series. A truncation to two terms yields the steepest ascent direction. For a steepest ascent step, we find $x^{(k)}$ along the path $\nabla f(x^{(k-1)})$ from $x^{(k-1)}$.

If $\nabla f(x^{(k-1)}) \geq 0$, then moving along the path $\nabla f(x^{(k-1)})$ can increase the function value. If $f(x)$ is bounded above (i.e., if the maximization problem makes sense), then at some point along this path, the function begins to decrease.

This point is not necessarily the maximum of the function, of course. Finding the maximum along the path, is a one-dimensional "line search".

After moving to the point $x^{(k)}$ in the direction of $\nabla f(x^{(k-1)})$, if $\nabla f(x^{(k)}) = 0$, we are at a stationary point. This may not be a maximum, but it is as good as we can do using steepest ascent from $x^{(k-1)}$. (In practice, we check $\|\nabla f(x^{(k)})\| \leq \epsilon$, for some norm $\|\cdot\|$ and some positive ϵ.)

If $\nabla f(x^{(k)}) < 0$ (remember we're maximizing the function), we change directions and move in the direction of $\nabla f(x^{(k)}) < 0$.

Knowing that we will probably be changing direction anyway, we often truncate the line search before we find the best $x^{(k)}$ in the direction of $\nabla f(x^{(k-1)})$.

Newton's Method

At the maximum x_*, $\nabla f(x_*) = 0$.
"Newton's method" for optimization is based on this fact.

Newton's method for optimization just solves the system of equations $\nabla f(x) = 0$ using

Newton's iterative method for solving equations:
to solve the system of n equations in n unknowns, $g(x) = 0$, we move from point $x^{(k-1)}$ to point $x^{(k)}$ by

$$x^{(k)} = x^{(k-1)} - \left(\nabla g(x^{(k-1)})^{\mathrm{T}}\right)^{-1} g(x^{(k-1)}).$$

Hence, applying this to solving $\nabla f(x^{(k-1)}) = 0$, we have the k^{th} step in Newton's method for optimization:

$$x^{(k)} = x^{(k-1)} - \nabla^2 f(x^{(k-1)})^{-1} \nabla f(x^{(k-1)}).$$

The direction of the step is $d_k = x^{(k)} - x^{(k-1)}$.

For numerical reasons, it is best to think of this as the problem of solving the equations

$$\nabla^2 f(x^{(k-1)}) d_k = -\nabla f(x^{(k-1)}),$$

and then taking $x^{(k)} = x^{(k-1)} + d_k$.

The Hessian

The Hessian $H(x) = \nabla^2 f(x)$ clearly plays an important role in Newton's method; if it is singular, the Newton step based on the solution to

$$\nabla^2 f(x^{(k-1)}) d_k = -\nabla f(x^{(k-1)}),$$

is undetermined.

The relevance of the Hessian goes far beyond this, however. The Hessian reveals important properties of the shape of the surface $f(x)$ at $x^{(k-1)}$.

The shape is especially interesting at a stationary point; that is a point x_* at which $\nabla f(x) = 0$.

If the Hessian is negative definite at x_*, $f(x_*)$ is a local maximum.

If the Hessian is positive definite at x_*, $f(x_*)$ is a local maximum.

If the Hessian is nonsingular, but neither negative definite nor positive definite at x_*, it is a saddlepoint.

If the Hessian is singular, the stationary point is none of the above.

In minimization problems, such as least squares, we hope the Hessian is positive definite, in which case the function is concave. In least squares fitting of the standard linear regression model, the Hessian is the famous $X^{\mathrm{T}}X$ matrix.

In maximization problems, such as MLE, it is particularly interesting to know whether $H(x)$ is negative definite everywhere (or -$H(x)$ is positive definite everywhere). In this case, the function is convex.

When $H(x)$ (in minimization problems or -$H(x)$ in maximization problems) is positive definite but nearly singular, it may be helpful to regularize the problem by adding a diagonal matrix with positive elements: $H(x) + D$.

One kind of regularization is ridge regression, in which the Hessian is replaced by $X^{\mathrm{T}}X + dI$.

Modifications of Newton's Method

In the basic Newton step, the direction d_k from $x^{(k-1)}$ is the best direction, but the point $d_k + x^{(k-1)}$ may not be the best point. In fact, the algorithm can often be speeded up by not going quite that far; that is, by "damping" the Newton step and taking $x^{(k)} = \alpha_k d_k + x^{(k-1)}$. This is a line search, and there are several ways of doing this. In the context of least squares, a common way of damping is the **Levenberg-Marquardt** method.

Rather than finding $\nabla^2 f(x^{(k-1)})$, we might find an approximate Hessian at $x^{(k-1)}$, $\widetilde{H}_k$, and then solve

$$\widetilde{H}_k d_k = -\nabla f(x^{(k-1)}).$$

This is called a **quasi-Newton** method.

In MLE, we may take the objective function to be the log likelihood, with the variable θ. In this case, the Hessian, $H(\theta)$, is $\partial^2 \log L(\theta;\ x)/\partial\theta(\partial\theta)^{\mathrm{T}}$. Under very general regularity conditions, the expected value of $H(\theta)$ is the negative of the expected value of $(\partial \log L(\theta;\ x)/\partial\theta)(\partial \log L(\theta;\ x)\partial\theta)^{\mathrm{T}}$, which is the Fisher information matrix, $I(\theta)$. This quantity plays an important role in statistical estimation. In MLE it is often possible to compute $I(\theta)$, and take the Newton step as

$$I(\theta^{(k)})d_k = \nabla \log L(\theta^{(k-1)};\ x).$$

This quasi-Newton method is called **Fisher scoring**.

More Modifications of Newton's Method

The method of solving the Newton or quasi-Newton equations may itself be iterative, such as a conjugate gradient or Gauss-Seidel method. (These are "inner loop iterations".) Instead of continuing the inner loop iterations to the solution, we may stop early. This is called a **truncated Newton method**.

The best gains in iterative algorithms often occur in the first steps. When the optimization is itself part of an iterative method, we may get an acceptable approximate solution to the optimization problem by stopping the optimization iterations early. Sometimes we may stop the optimization after just one iteration. If Newton's method is used, this is called a **one-step Newton method**.

D.5.2 Alternating Conditional Optimization

The computational burden in a single iteration for solving the optimization problem can sometimes be reduced by more than a linear amount by separating x into two subvectors. The optimum is then computed by alternating between computations involving the two subvectors, and the iterations proceed in a zigzag path to the solution.

Each of the individual sequences of iterations is simpler than the sequence of iterations on the full x.

For the problem

$$\min_x f(x)$$

if $x = (x_1, x_2)$ that is, x is a vector with at least two elements, and x_1 and x_2 may be vectors), an iterative alternating conditional optimization algorithm may start with $x_2^{(0)}$, and then for $k = 0, 1, \ldots$,

1. $x_1^{(k)} = \arg\min_{x_1} f\left(x_1, x_2^{(k-1)}\right)$
2. $x_2^{(k)} = \arg\min_{x_2} f\left(x_1^{(k)}, x_2\right)$

Use of Conditional Bounding Functions: MM Methods

In an iterative method to maximize $f(x)$, the idea, given $x^{(k-1)}$ at step k, is to try to find a function $g\left(x;\ x^{(k-1)}\right)$ with these properties:

- is easy to work with (that is, is easy to maximize)
- $g\left(x;\ x^{(k-1)}\right) \leq f(x) \quad \forall x$
- $g\left(x^{(k-1)};\ x^{(k-1)}\right) = f\left(x^{(k-1)}\right)$

If we can find $x^{(k)} \ni g\left(x^{(k)};\ x^{(k-1)}\right) > g\left(x^{(k-1)};\ x^{(k-1)}\right)$, we have the "sandwich inequality":

$$f\left(x^{(k)}\right) \geq g\left(x^{(k)};\ x^{(k-1)}\right) > g\left(x^{(k-1)};\ x^{(k-1)}\right) = f\left(x^{(k-1)}\right).$$

An equivalent (but more complicated) method for seeing this inequality uses the fact that

$$f\left(x^{(k)}\right) - g\left(x^{(k)};\ x^{(k-1)}\right) \geq f\left(x^{(k-1)}\right) - g\left(x^{(k)};\ x^{(k-1)}\right).$$

(From the properties above,

$$g\left(x;\ x^{(k-1)}\right) - f\left(x\right)$$

attains its maximum at $x^{(k-1)}$.)

Hence,

$$\begin{aligned} f\left(x^{(k)}\right) &= g\left(x^{(k)};\ x^{(k-1)}\right) + f\left(x^{(k)}\right) - g\left(x^{(k)};\ x^{(k-1)}\right) \\ &> g\left(x^{(k-1)};\ x^{(k-1)}\right) + f\left(x^{(k-1)}\right) - g\left(x^{(k-1)};\ x^{(k-1)}\right) \\ &= f\left(x^{(k-1)}\right). \end{aligned}$$

The relationship between $f\left(x^{(k)}\right)$ and $f\left(x^{(k-1)}\right)$, that is, whether we have "$>$" or "$\geq$" in the inequalities, depends on the relationship between $g\left(x^{(k)};\ x^{(k-1)}\right)$ and $g\left(x^{(k-1)};\ x^{(k-1)}\right)$.

We generally require $g\left(x^{(k)};\ x^{(k-1)}\right) > g\left(x^{(k-1)};\ x^{(k-1)}\right)$.
Clearly, the best step would be

$$x^{(k)} = \arg\min_x g\left(x;\ x^{(k-1)}\right),$$

but the overall efficiency of the method may be better if we don't work too hard to find the maximum, but just accept some $x^{(k)}$ that satisfies $g\left(x^{(k)};\ x^{(k-1)}\right) \leq g\left(x^{(k-1)};\ x^{(k-1)}\right)$.

After moving to $x^{(k)}$, we must find a new $g\left(x;\ x^{(k)}\right)$.

Equivalent notations:

$$g\left(x;\ x^{(k-1)}\right) \leftrightarrow g^{(k)}(x) \leftrightarrow g_k(x)$$

Note the logical difference in k and $k-1$, although both determine the same g.

The g that we maximize is a "minorizing" function.
Thus, we Minorize then Maximize: MM.
Alternatively, we Majorize then Minimize: MM.
Reference: Lang, Hunter, and Yang (2000).

Maximization in Alternating Algorithms

In alternating multiple step methods such as alternating conditional maximization methods and methods that use a conditional bounding function, at least one of the alternating steps involves maximization of some function.

As we indicated in discussing conditional bounding functions, instead of finding a point that actually maximizes the function, which may be a difficult task, we may just find a point that increases the value of the function. Under this weaker condition, the methods still work.

We may relax the requirement even further, so that for some steps we only require that the function not be decreased. So long as we maintain the requirement that the function actually be increased in a sufficient number of steps, the methods still work.

The most basic requirement is that $g\left(x^{(k)};\ x^{(k)}\right) \geq g\left(x^{(k-1)};\ x^{(k-1)}\right)$. (Even this requirement is relaxed in the class of optimization algorithms based on annealing. A reason for relaxing this requirement may be to avoid getting trapped in local optima.)

Applications and the Special Case of EM Methods

In maximum likelihood estimation, the objective function is the likelihood, $L_X(\theta; x)$ or the log-likelihood, $l_X(\theta; x)$. (Recall that a likelihood depends on a known distributional form for the data; that is why we use the notation $L_X(\theta; x)$ and $l_X(\theta; x)$, where "X" represents the random variable of the distribution.)

The variable for the optimization is θ; thus in an iterative algorithm, we find $\theta^{(1)}, \theta^{(2)}, \ldots$.

One type of alternating method is based on conditional optimization and a conditional bounding function alternates between updating $\theta^{(k)}$ using maximum likelihood and conditional expected values. This method is called the *EM method* because the alternating steps involve an expectation and a maximization.

Given $\theta^{(k-1)}$ we seek a function $q_k(x, \theta)$ that has a known relationship with $l_X(\theta;\ x)$, and then we determine $\theta^{(k)}$ to maximize $q_k(x, \theta)$ (subject to any constraints on acceptable values of θ).

The minorizing function $q_k(x, \theta)$ is formed as a conditional expectation of a joint likelihood. In addition to the data we have observed, call it X, we assume we have some unobserved data U.

Thus, we have "complete" data $C = (X, U)$ given the actual observed data X, and the other component, U, of C that is not observed.

Let $L_C(\theta; c)$ be the likelihood of the complete data, and let $L_X(\theta; x)$ be the likelihood of the observed data, with similar notation for the log-likelihoods. We refer to $L_C(\theta; c)$ as the "complete likelihood".

There are thus two likelihoods, one based on the complete (but unknown) sample, and one based only on the observed sample.

We wish to estimate the parameter θ, which figures in the distribution of both components of C.

The conditional likelihood of C given X is

$$L_{C|X}(\theta; c|x) = L_C(\theta;\ x, u)/L_X(\theta;\ x),$$

or

$$l_{C|X}(\theta; c|x) = l_C(\theta;\ x, u) - l_X(\theta;\ x).$$

Note that the conditional of C given X is the same as the conditional of U given X, and we may write it either way, either $C|X$ or $U|X$.

Because we do not have all the observations, $L_{C|X}(\theta; c|x)$ and $L_C(\theta;\ c)$ have

- unknown variables (the unobserved U)
- the usual unknown parameter.

Hence, we cannot follow the usual approach of maximizing the likelihood with given data.

We concentrate on the unobserved or missing data first.

We use a provisional value of $\theta^{(k-1)}$ to approximate the complete likelihood based on the expected value of U given $X = x$.

The expected value of the likelihood, which will generally be a function of both θ and $\theta^{(k-1)}$, minorizes the objective function of interest, $L_X(\theta;\ x)$, as we will see.

We then maximize this minorizing function with respect to θ to get $\theta^{(k)}$.

Let $L_C(\theta\ ;\ x, u)$ and $l_C(\theta\ ;\ x, u)$ denote, respectively, the likelihood and the log-likelihood for the complete sample. The objective function, that is, the likelihood for the observed X, is

$$L_X(\theta\ ;\ x) = \int L_C(\theta\ ;\ x, u)\,\mathrm{d}u,$$

and $l_X(\theta\ ;\ x) = \log L_X(\theta\ ;\ x)$.

After representing the function of interest, $L_X(\theta\ ;\ x)$, as an integral, the problem is to determine this function; that is, to average over U. (This is what the integral does, but we do not know what to integrate.) The average over U is the expected value with respect to the marginal distribution of U.

This is a standard problem in statistics: we estimate an expectation using observed data.

In this case, however, even the values that we average to estimate the expectation depends on θ, so we use a provisional value of θ.

We begin with a provisional value of θ, call it $\theta^{(0)}$.

Given any provisional value $\theta^{(k-1)}$, we will compute a provisional value $\theta^{(k)}$ that increases (or at least does not decrease) the conditional expected value of the complete likelihood.

EM methods were first discussed systematically by Dempster, Laird, and Rubin (1977).

The EM approach to maximizing $L_X(\theta\ ;\ x)$ has two alternating steps. The steps are iterated until convergence.

E step : compute $q_k(x, \theta) = \mathrm{E}_{U|x,\theta^{(k-1)}}\big(l_C(\theta;\ x, U)\big)$.

M step : determine $\theta^{(k)}$ to maximize $q_k(x, \theta)$, or at least to increase it (subject to any constraints on acceptable values of θ).

Convergence of the EM Method

Is $l_X(\theta^{(k)};\ x) \geq l_X(\theta^{(k-1)};\ x)$?
(If it is, of course, then $L_X(\theta^{(k)};\ x) \geq L_X(\theta^{(k-1)};\ x)$, because the log is monotone increasing.)

The sequence $\theta^{(1)}, \theta^{(2)}, \ldots$ converges to a local maximum of the observed-data likelihood $L(\theta\ ;\ x)$ under fairly general conditions. (It can be very slow to converge, however.)

Why EM Works

The real issue is whether the EM sequence

$$\begin{aligned}\{\theta^{(k)}\} \rightarrow & \arg\max_\theta l_X(\theta;\ x) \\ & (= \arg\max_\theta L_X(\theta;\ x)).\end{aligned}$$

If $l_X(\cdot)$ is bounded (and it better be!), this is essentially equivalent to to asking if

$$l_X(\theta^{(k)};\ x) \geq l_X(\theta^{(k-1)};\ x).$$

(So long as in a sufficient number of steps the inequality is strict.)

Using an equation from before, we first write

$$l_X(\theta;\ X) = l_C(\theta;\ (X,U)) - l_{U|X}(\theta; U|X),$$

and then take the conditional expectation of functions of U given x and under the assumption that θ has the provisional value $\theta^{(k-1)}$:

$$\begin{aligned} l_X(\theta;\ X) &= \mathrm{E}_{U|x,\theta^{(k-1)}}\big(l_C(\theta;\ (x,U))\big) - \mathrm{E}_{U|x,\theta^{(k-1)}}\big(l_{U|X}(\theta; U|x)\big) \\ &= q_k(x,\theta) - h_k(x,\theta), \end{aligned}$$

where

$$h_k(x,\theta) = \mathrm{E}_{U|x,\theta^{(k-1)}}\big(l_{U|X}(\theta; U|x)\big).$$

Now, consider

$$l_X(\theta^{(k)};\ X) - l_X(\theta^{(k-1)};\ X).$$

This has two parts:

$$q_k(x,\theta^{(k)}) - q_k(x,\theta^{(k-1)})$$

and

$$-\left(h_k(x,\theta^{(k)}) - h_k(x,\theta^{(k-1)})\right).$$

The first part is nonnegative from the M part of the k^{th} step.

What about the second part? We will show that it is nonnegative also (or without the minus sign it is nonpositive).

For the other part, for given $\theta^{(k-1)}$ and any θ, ignoring the minus sign, ...

$$\begin{aligned} & h_k(x,\theta) - h_k(x,\theta^{(k-1)}) \\ &= \mathrm{E}_{U|x,\theta^{(k-1)}}\big(l_{U|X}(\theta; U|x)\big) - \mathrm{E}_{U|x,\theta^{(k-1)}}\big(l_{U|X}(\theta^{(k-1)}; U|x)\big) \\ &= \mathrm{E}_{U|x,\theta^{(k-1)}}\left(\log\left(L_{U|x}(\theta; U|x)/L_{U|X}(\theta^{(k-1)}; U|x)\right)\right) \\ &\leq \log\left(\mathrm{E}_{U|x,\theta^{(k-1)}}\left(L_{U|X}(\theta; U|x)/L_{U|X}(\theta^{(k-1)}; U|x)\right)\right) \\ &\qquad \text{(by Jensen's inequality)} \\ &= \log\int_{\mathcal{D}(U)} \frac{L_{U|X}(\theta; U|x)}{L_{U|X}(\theta^{(k-1)}; U|x)}\, L_{U|X}(\theta^{(k-1)}; U|x)\ \mathrm{d}u \\ &= \log\int_{\mathcal{D}(U)} L_{U|X}(\theta; U|x)\ \mathrm{d}u \\ &= 0. \end{aligned}$$

So the second term is also nonnegative, and hence,

$$l_X(\theta^{(k)};\ x) \geq l_X(\theta^{(k-1)};\ x).$$

A Minorizing Function in EM Algorithms

With $l_X(\theta;\ x) = q_k(x,\theta) - h_k(x,\theta)$, and $h_k(x,\theta) \leq h_k(x,\theta^{(k-1)})$ from the previous pages, we have

$$l_X(\theta^{(k-1)};\ x) - q_k(x,\theta^{(k-1)}) \leq l_X(\theta;\ x) - q_k(x,\theta);$$

and so

$$q_k(x,\theta) + c(x,\theta^{(k-1)}) \leq l_X(\theta;\ x),$$

where $c(x,\theta^{(k-1)})$ is constant with respect to θ.

Therefore for given $\theta^{(k-1)}$ and any x,

$$g(\theta) = l_X(\theta^{(k-1)};\ X) - q_k(x,\theta^{(k-1)})$$

is a minorizing function for $l_X(\theta;\ x)$.

Alternative Ways of Performing the Computations

There are two kinds of computations that must be performed in each iteration:

- E step : compute $q_k(x,\theta) = \mathrm{E}_{U|x,\theta^{(k-1)}}\big(l_c(\theta;\ x,U)\big)$.
- M step : determine $\theta^{(k)}$ to maximize $q_k(x,\theta)$, subject to any constraints on acceptable values of θ.

There are obviously various ways to perform each of these computations.

A number of papers since 1977 have suggested specific methods for the computations.

For each specification of a method for doing the computations or each little modification, a new name is given, just as if it were a new idea:
GEM, ECM, ECME, AECM, GAECM, PXEM, MCEM, AEM, EM1, SEM

A general reference for EM methods is Ng, Krishnan, and McLachlan (2004).

E Step

There are various ways the expectation step can be carried out.

In the happy case of an exponential family or some other nice distributions, the expectation can be computed in closed form. Otherwise, computing the expectation is a numerical quadrature problem. There are various procedures for quadrature, including Monte Carlo.

Some people have called an EM method that uses Monte Carlo to evaluate the expectation an MCEM method. (If a Newton-Cotes method is used, however, we do not call it an NCEM method!) The additional Monte Carlo computations add a lot to the overall time required for convergence of the EM method.

An additional problem in using Monte Carlo in the expectation step may be that the distribution of C is difficult to simulate. The convergence criterion for optimization methods that involve Monte Carlo generally should be tighter than for deterministic methods.

M Step

For the maximization step, there are even more choices.

The first thing to note, as we mentioned earlier for alternating algorithms generally, is that rather than maximizing q_k, we can just require that the overall sequence increase.

Dempster, Laird, and Rubin (1977) suggested requiring only an increase in the expected value; that is, take $\theta^{(k)}$ so that

$$q_k(u, \theta^{(k)}) \geq q_{k-1}(u, \theta^{(k-1)}).$$

They called this a generalized EM algorithm, or GEM. (Even in the paper that introduced the "EM" acronym, another acronym was suggested for a variation.) If a one-step Newton method is used to do this, some people have called this a EM1 method.

Meng and Rubin (1993) describe a GEM algorithm in which the M-step is an alternating conditional maximization; that is, if $\theta = (\theta_1, \theta_2)$, first $\theta_1^{(k)}$ is determined to maximize q subject to the constraint $\theta_2 = \theta_2^{(k-1)}$; then $\theta_2^{(k)}$ is determined to maximize q_k subject to the constraint $\theta_1 = \theta_1^{(k)}$. This sometimes simplifies the maximization problem so that it can be done in closed form. They call this an expectation conditional maximization method, ECM.

Alternate Ways of Terminating the Computations

In any iterative algorithm, we must have some way of deciding to terminate the computations. (The generally-accepted definition of "algorithm" requires that it terminate. In any event, of course, we want the computations to cease at some point.)

One way of deciding to terminate the computations is based on convergence; if the computations have converged we quit. In addition, we also have some criterion by which we decide to quit anyway.

In an iterative optimization algorithm, there are two obvious ways of deciding when convergence has occurred. One is when the decision variables (the estimates in MLE) are no longer changing appreciably, and the other is when the value of the objective function (the likelihood) is no longer changing appreciably.

Convergence

It is easy to think of cases in which the objective function converges, but the decision variables do not. All that is required is that the objective function is flat over a region at its maximum. In statistical terms, the corresponds to unidentifiability.

The Variance of Estimators Defined by the EM Method

As is usual for estimators defined as solutions to optimization problems, we may have some difficulty in determining the statistical properties of the estimators.

Louis (1982) suggested a method of estimating the variance-covariance matrix of the estimator by use of the gradient and Hessian of the complete-data log-likelihood, $l_{L_c}(\theta\ ;\ u, v)$.

Meng and Rubin (1991) use a "supplemented" EM method, SEM, for estimation of the variance-covariance matrix.

Kim and Taylor (1995) also described ways of estimating the variance-covariance matrix using computations that are part of the EM steps.

It is interesting to note that under certain assumptions on the distribution, the iteratively reweighted least squares method can be formulated as an EM method (see Dempster, Laird, and Rubin, 1980).

Missing Data

Although EM methods do not rely on missing data, they can be explained most easily in terms of a random sample that consists of two components, one observed and one unobserved or missing.

A simple example of missing data occurs in life-testing, when, for example, a number of electrical units are switched on and the time when each fails is recorded.

In such an experiment, it is usually necessary to curtail the recordings prior to the failure of all units.

The failure times of the units still working are unobserved, but the number of censored observations and the time of the censoring obviously provide information about the distribution of the failure times.

Mixtures

Another common example that motivates the EM algorithm is a finite mixture model.

Each observation comes from an unknown one of an assumed set of distributions. The missing data is the distribution indicator.

The parameters of the distributions are to be estimated. As a side benefit, the class membership indicator is estimated.

Applications of EM Methods

The missing data can be missing observations on the same random variable that yields the observed sample, as in the case of the censoring example; or the missing data can be from a different random variable that is related somehow to the random variable observed.

Many common applications of EM methods involve missing-data problems, but this is not necessary.

Often, an EM method can be constructed based on an artificial "missing" random variable to supplement the observable data.

Notes and Additional References for Section D.5

There is an extensive literature on optimization, much of it concerned with practical numerical algorithms. Software for optimization is widely available, both in special-purpose programs and in general-purpose packages such as R and Matlab.

Additional References

Dempster, A. P.; N. M. Laird; and D. B. Rubin (1977), Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society, Series B* **45**, 51–59.

Gentle, James E. (2009), *Optimization Methods for Applications in Statistics*, Springer-Verlag, New York.

Kim, Dong K., and Jeremy M. G. Taylor (1995), The restricted EM algorithm for maximum likelihood estimation under linear restrictions on the parameters, *Journal of the American Statistical Association* **90**, 708–716.

Louis, Thomas A. (1982), Finding the observed information matrix when using the EM algorithm, *Journal of the the Royal Statistical Society B* **44**, 226–233.

Meng, Xiao-Li, and Donald B. Rubin (1991), Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm, *Journal of the American Statistical Association* **86**, 899–909.

Meng, X.-L., and D. B. Rubin (1993), Maximum likelihood estimation via the ECM algorithm: a general framework, *Biometrika* **80**, 267–278.

Ng, Shu Kay; Thriyambakam Krishnan, and Geoffrey J. McLachlan (2004), The EM algorithm, *Handbook of Computational Statistics; Concepts and Methods* (edited by James E. Gentle, Wolfgang Härdle, and Yuichi Mori), Springer, Berlin, 137–168.

Notes

Gelbaum and Olmsted (1990, 2003) have remarked that mathematics is built on two types of things: theorems and counterexamples. Counterexamples help

us to understand the principles in a way that we might miss if we only considered theorems. Counterexamples delimit the application of a theorem. They help us understand why each part of the hypothesis of a theorem is important.

The book by Romano and Siegel (1986), which is listed in the general references, is replete with examples that illustrate the "edges" of statistical properties.

Books of this general type concerned with other areas of mathematics are listed below.

Additional References

Gelbaum, Bernard R., and John M. H. Olmsted (1990), *Theorems and Counterexamples in Mathematics*, Springer, New York.

Gelbaum, Bernard R., and John M. H. Olmsted (2003), *Counterexamples in Analysis*, (corrected reprint of the second printing published by Holden-Day, Inc., San Francisco, 1965), Dover Publications, Inc., Mineola, New York.

Rajwade, A. R., and A. K. Bhandari (2007), *Surprises and Counterexamples in Real Function Theory*, Hindustan Book Agency, New Delhi.

Steen, Lynn Arthur, and J. Arthur Seebach, Jr. (1995), *Counterexamples in Topology* (reprint of the second edition published by Springer-Verlag, New York, 1978), Dover Publications, Inc., Mineola, New York.

Stoyanov, Jordan M. (1987), *Counterexamples in Probability*, John Wiley & Sons, Ltd., Chichester, United Kingdom.

Wise, Gary L., and Eric B. Hall (1993), *Counterexamples in Probability and Real Analysis*, The Clarendon Press, Oxford University Press, New York.

Bibliography

The references listed in this bibliography are relevant to broad areas of mathematical statistics. References more specific to topics in individual chapters are listed at the ends of those chapters.

Abramowitz, Milton, and Irene A. Stegun (Editors) (1964), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, National Bureau of Standards (NIST), Washington. (Reprinted by Dover Publications, New York, 1974. Work on an updated version is occurring at NIST; see http://dlmf.nist.gov/ for the current status.)

Ash, Robert B., and Catherine A. Doleans-Dade (1999), *Probability & Measure Theory*, second edition, Academic Press, New York.

Athreya, Krishna B., and Soumen N. Lahiri (2006), *Measure Theory and Probability Theory*, Springer, New York.

Barndorff-Nielson, O. E., and D. R. Cox (1994), *Inference and Asymptotics*, Chapman and Hall, New York.

Bickel, Peter J., and David A. Freedman (1981), Some asymptotic theory for the bootstrap, *The Annals of Statistics* **9**, 1196–1217.

Billingsley, Patrick (1995), *Probability and Measure*, third edition, John Wiley & Sons, New York.

Breiman, Leo (1968), *Probability*, Addison-Wesley, New York.

DasGupta, Anirban (2008), *Asymptotic Theory of Statistics and Probability*, Springer, New York.

Davies, Laurie, and Ursula Gather (2004), Robust statistics, *Handbook of Computational Statistics; Concepts and Methods* (edited by James E. Gentle, Wolfgang Härdle, and Yuichi Mori), Springer, Berlin, 653–695.

Dudley, R. M. (2002), *Real Analysis and Probability*, second edition, Cambridge University Press, Cambridge, United Kingdom.

Evans, Merran; Nicholas Hastings; and Brian Peacock (2000), *Statistical Distributions*, third edition, John Wiley & Sons, New York.

Gentle, James E. (2007), *Matrix Algebra: Theory, Computations, and Applications in Statistics*, Springer, New York.

Hall, Peter (1992), *The Bootstrap and Edgeworth Expansion*, Springer, New York.

Leemis, Lawrence M., and Jacquelyn T. McQueston (2008), Univariate distribution relationships, *The American Statistician* **62**, 45–53.

Lehmann, E. L. (1999), *Elements of Large-Sample Theory*, Springer, New York.

Lehmann, E. L., and George Casella (1998), *Theory of Point Estimation*, second edition, Springer, New York.

Lehmann, E. L., and Joseph P. Romano (2005), *Testing Statistical Hypotheses*, third edition, Springer, New York.

Pardo, Leandro (2005), *Statistical Inference Based on Divergence Measures*, Chapman and Hall, New York.

Romano, Joseph P., and Andrew F. Siegel (1986), *Counterexamples in probability and statistics*, Wadsworth & Brooks/Cole, Monterey, California.

Schervish, Mark J. (1995), *Theory of Statistics*, Springer, New York.

Serfling, Robert J. (1980), *Approximation Theorems of Mathematical Statistics*, John Wiley & Sons, New York.

Shao, Jun (2003), *Mathematical Statistics*, second edition, Springer, New York.

Shao, Jun (2005), *Mathematical Statistics: Exercises and Solutions*, Springer, New York.

Staudte, Robert G., and Simon J. Sheather (1990), *Robust Estimation and Testing*, John Wiley & Sons, New York.

Thompson, William J. (1997), *Atlas for Computing Mathematical Functions: An Illustrated Guide for Practitioners with Programs in Fortran 90 and Mathematica*, John Wiley & Sons, New York.

Various authors (2002), Chapter 4, Theory and Methods of Statistics, in *Statistics in the 21st Century*, edited by Adrian E. Raftery, Martin A. Tanner, and Martin T. Wells, Chapman and Hall, New York.

Index