The Theory of Statistics and Its Applications

By Dennis D. Cox
Rice University

Chapter 1

Measure Spaces.

Theoretical statistics relies heavily on probability theory, which in turn is based on measure theory. Thus, a student of advanced statistics needs to learn some measure theory. A proper introduction to measure theory is not provided here. Instead, definitions and concepts are given and the main theorems are stated without proof. The student should take a course on the subject, such as one based on the text *Probability and Measure*, by Billingsley. Other books on measure theory and probability are discussed at the end of Section 1.

Measure theory is a rather difficult and dry subject, and many statisticians believe it is unnecessary to learn measure theory in order to understand statistics. To counter these views, we offer the following list of benefits from studying measure theory:

(i) A good understanding of measure theory eliminates the artificial distinction between discrete and continuous random variables. Summations become an example of the abstract integral, so one need not dichotomize proofs into the discrete and continuous cases, but can cover both at once.

(ii) One can understand probability models which cannot be classified as either discrete or continuous. Such models do arise in practice, e.g. when censoring a continuous lifetime and in Generalized Random Effects Models such as the Beta-Binomial.

(iii) The measure theoretic statistics presented here provides a basis for understanding complex problems that arise in the statistical inference of stochastic processes and other areas of statistics.

(iv) Measure theory provides a unifying theme for much of statistics. As an example, consider the notion of likelihoods, which are rather mysterious in some ways, but at least from a formal point of view are measure theoretically quite simple. As with many mathematical theories, if one puts in the initial effort to understand the theory, one is rewarded with a deeper and clearer understanding of the subject.

(v) Certain fundamental notions (such as conditional expectation) are arguably not completely understandable except from a measure theoretic point of view.

Rather than spend more words on motivation, let us embark on the subject matter.

1.1 Measures.

A measure is a function μ defined for certain subsets A of a set Ω which assigns a nonnegative number $\mu(A)$ to each "measurable" set A. In probability theory, the probability is a measure, denoted by P instead of μ, which satisfies $P(\Omega) = 1$. In the context of probability theory, the subset A is called an event and Ω is called the sample space. Unfortunately, probability developed somewhat independently of measure theory, so there is a separate terminology, and we will frequently learn two names and notations for the same concept. To introduce the subject of measure theory and provide some insight into the technical difficulties (which are many), we begin with some examples which are pertinent to our study. More formal definitions are given below.

Example 1.1.1 Consider a die with 6 faces. Let the sample space Ω be the finite set of integers $\{ 1, 2, 3, 4, 5, 6 \}$ corresponding to the possible outcomes if we roll the die once and count the number of spots on the face that turns up. Let the collection of events be all subsets of Ω. We define a probability measure P on these events by $P(A) = \#(A)/6$ where $A \subset \Omega$ and $\#(A)$ denotes the number of elements or outcomes in A. It will turn out that $\#$ is a measure on Ω which will be very useful for us.

Recall that probability was invented to describe the long run frequencies involved with games of chance. Thus, in the context of this example, we expect that if the die is rolled many, many times, then the relative frequency of A (which is the ratio of the number of rolls where the outcome is in A to the total number of rolls) will be approximately $P(A) = \#(A)/6$. We will discuss these notions at great length, but in fact there is no way to mathematically prove this claim that the long run relative frequency equals $P(A)$. One must simply perform the "experiment" of rolling the die many, many times to see if it is a good approximation to reality.

□

Example 1.1.2 Let a random real number be chosen in the interval $[0, 1]$ such that the probability of the number lying in any subinterval $[a, b]$ $(0 \le a < b \le 1)$ is the length, i.e. $P([a, b]) = b - a$. Here, $\Omega = [0, 1]$. Such a random number is said to be *uniformly distributed* on the interval $[0, 1]$.

A possible physical way of constructing such a random real number is the following. Take a balanced wheel with a mark on it and a special fixed mark on the supporting structure. Spin the wheel with a great force so it makes many revolutions, and after it stops, measure in radians the counterclockwise angle between the fixed mark and the mark on the wheel, then divide the measured angle by 2π to obtain a real number in the range $[0, 1]$. Similarly to the previous example with rolling a die, for $[a, b] \subset [0, 1]$, we conjecture that the long run relative frequency of the number of times the measured angle divided by 2π is in $[a, b]$ is $b - a$. Again, this can only be verified empirically.

We clearly can extend the probability measure P from closed intervals to other subsets of $[0, 1]$. For instance, we can put $P((a, b)) = P([a, b)) = P((a, b]) = P([a, b]) = b - a$ for open intervals (a, b), left closed right open intervals $[a, b)$, and left open right closed intervals $(a, b]$, all with $0 \leq a \leq b \leq 1$. Also, if $[a_1, b_1]$, $[a_2, b_2]$, $\ldots$ is a finite or infinite sequence of disjoint closed intervals (one can also allow open or semi–open intervals), then we can put $P(\bigcup_i [a_i, b_i]) = \sum_i (b_i - a_i)$.

It turns out for technical reasons that in this case, one cannot define the probability measure of all subsets of $[0, 1]$. The reasons are very complicated and we shall generally not discuss the issue, although some insight is given in Remarks 1.1.4 below. The sets which are not “measurable” will never arise in our work. For an example, see p. 41 of Billingsley.

The probability measure of this example is related to a (nonprobability) measure: let $\mu = m$ be a measure on arbitrary intervals of real numbers which equals the length of the interval, i.e. $m((a, b)) = b - a$ for any open interval (a, b), $a < b$, and similarly for the other varieties of intervals. Here, $\Omega = I\!R$, the set of all real numbers, also denoted $(-\infty, \infty)$. This measure m is called *Lebesgue measure*, and will turn out to have many uses.

This probability measure (the uniform distribution on $[0, 1]$) plays a fundamental role in the computer generation of random numbers (or more correctly, pseudorandom numbers). Indeed, the basic (pseudo)random numbers generated by computers are uniformly distributed on $[0, 1]$, except for round-off. To obtain random numbers with other distributions, various transformations and other tricks are employed. We shall refer to the subject of computer generation of random numbers and their various applications as *Monte Carlo simulation.*

□

In the first example of rolling a die and counting spots, no technical difficulties arise, but in the second example of the uniformly distributed random number on $[0, 1]$, the sample space is so complex that it is not possible to define the probability or measure on all possible subsets. In general, it will be necessary to restrict the domain of definition of the measure to a collection of subsets of the sample space, and it will be necessary for this collection of subsets on which the measure is defined to satisfy certain properties.

1.1.1 σ–Fields.

Now we set forth the requisite properties for the class of sets on which a measure is defined.

Definition 1.1.1 *Let $\mathcal{F}$ be a collection of subsets of a set Ω. Then $\mathcal{F}$ is called a* sigma field *(or* sigma algebra*; written σ-field or σ-algebra) if and only if (abbreviated iff) it satisfies the following properties:*

(i) *The empty set $\emptyset \in \mathcal{F}$;*

(ii) *If $A \in \mathcal{F}$, then the complement $A^c \in \mathcal{F}$;*

(iii) *If A_1, A_2, $\ldots$ is a sequence of elements of $\mathcal{F}$, then their union $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.*

A pair $(\Omega, \mathcal{F})$ consisting of a set Ω and a σ-field of subsets $\mathcal{F}$ is called a measurable space*. The elements of $\mathcal{F}$ are called* measurable sets *or* events*.*

□

Remarks 1.1.1 (a) The set Ω is called the *sample space* in probability, but in general measure theory it is simply called the *underlying set* or *underlying space.*

(b) Since $\emptyset^c = \Omega$, it follows from (i) and (ii) that $\Omega \in \mathcal{F}$.

(c) Given any set Ω, the *trivial σ-field* is $\mathcal{F} = \{\emptyset, \Omega\}$. One easily verifies that this is a σ-field, and is in fact the smallest σ-field on Ω.

(d) Given any set Ω, the *power set*

$$\mathcal{P}(\Omega) = \{A : A \subset \Omega\}$$

consisting of all subsets of Ω is also a σ-field on Ω, and in fact is the largest σ-field on Ω. (Note: Many authors denote $\mathcal{P}(\Omega)$ by 2^{Ω}.)

(e) It follows from the definition that if $\mathcal{F}$ is a σ-field and A_1, A_2, $\ldots$ is a sequence in $\mathcal{F}$, then the intersection $\bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$. To see this, first note that $(A^c)^c = A$, so we have for any $A \subset \Omega$, $A \in \mathcal{F}$ iff $A^c \in \mathcal{F}$. Thus it suffices to show that $[\bigcap_{i=1}^{\infty} A_i]^c \in \mathcal{F}$. But by DeMorgan's law

$$\left[\bigcap_{i=1}^{\infty} A_i\right]^c = \bigcup_{i=1}^{\infty} A_i^c$$

and the latter event is in $\mathcal{F}$ since each $A_i^c \in \mathcal{F}$. This proves the result.

(f) Properties (i) and (iii) imply that the union of a finite sequence of sets from a σ-field is again in the σ-field. To see this, let $\mathcal{F}$ be a σ-field and A_1, A_2, $\ldots$, A_n a finite sequence from $\mathcal{F}$. Extend this to an infinite sequence by defining $A_k = \emptyset$ for $k > n$. Then

$$\cup_{i=1}^{n} A_i = \cup_{i=1}^{\infty} A_i \in \mathcal{F}.$$

(g) Property (iii) of the definition of a σ-field does not imply that the union of an arbitrary collection of sets from $\mathcal{F}$ is again in $\mathcal{F}$. There may exist collections which cannot be "listed" as a sequence. This is discussed in more detail in Remark 1.1.2 below.

□

Remarks 1.1.2 *A set A is called* countable *if it can be listed as a sequence, finite or infinite:*

$$A \;=\; \{a_1, a_2, \ldots\}.$$

We shall sometimes say that a set is countably infinite *to indicate that it is countable but not finite. There are in fact various "orders of infinity," and countable infinity is the smallest one. Here, we mean by "infinity" a counting number used to denote the number of elements in a set. We shall encounter another "infinity" shortly. Two sets A and B have the same number of elements if there is a one to one correspondence (or bijective map) between them. If there is no bijective map between A and B, but there is a one to one correspondence beween A and a subset of B, then B has more elements than A. Somewhat surprisingly, the set of rational numbers (real numbers which can be written as fractions) have the same number elements as its proper subset the integers (see Exercise 1.1.6). Any set which can be put in one to one correspondence with a subset of the natural numbers or "counting numbers" $I\!\!N = \{\ 1,\ 2,\ 3,\ \ldots\ \}$ is called* countable. *Such sets can be listed as a sequence $A = \{\ a_1,\ a_2,\ a_3,\ \ldots\ \}$. Property (iii) of Definition 1.1.1 is sometimes expressed as "a σ-field is closed under countable unions." It turns out the the set of all real numbers $I\!\!R$ (including irrational numbers like $\sqrt{2}$, π, and e) cannot be put into a one to one correspondence with the natural numbers, so it has "more" elements, and is said to be* uncountably infinite. *These issues are pertinent to the technical difficulties which make it impossible to extend Lebesgue measure to all subsets of $I\!\!R$, and hence require us to consider the notion of a σ-field (rather than just defining a measure on all subsets of the underlying space).*

□

It takes a certain amount of work to obtain σ-fields other than the trivial σ-field or the power set. A standard approach is to consider the smallest σ-field containing a given family of sets. We shall illustrate this concept. Let $A \subset \Omega$ be a nonempty proper subset of Ω (i.e. $\emptyset \neq A \neq \Omega$), then

$$\sigma(A) \;=\; \{\emptyset, A, A^c, \Omega\} \tag{1.1}$$

is a σ–algebra. If $\emptyset \neq B \neq \Omega$, $A \cap B \neq \emptyset$, and neither A nor B is a subset of the other, then one can obtain a σ-field $\sigma(\{A, B\})$ consisting of 16 elements.

Definition 1.1.2 *If $\mathcal{C}$ is any collection of subsets of Ω, then the σ-field generated by $\mathcal{C}$, denoted $\sigma(\mathcal{C})$, is the smallest σ-field containing $\mathcal{C}$. (Here, "smallest" means that if $\mathcal{F}$ is any σ-field containing $\mathcal{C}$, then $\sigma(\mathcal{C}) \subset \mathcal{F}$.)*

□

The following result shows that $\sigma(\mathcal{C})$ always exists and indicates how one may "construct" it.

Proposition 1.1.1 *Let Ω and $\mathcal{C}$ be as in Definition 1.1.2, and let $\Gamma = \{\mathcal{G} : \mathcal{G}$ is a σ-field on Ω and $\mathcal{C} \subset \mathcal{G}\}$. Then $\sigma(\mathcal{C}) = \bigcap_{\mathcal{G}\in\Gamma} \mathcal{G}$.*

Proof. Let $\mathcal{F}$ denote $\bigcap_{\mathcal{G}\in\Gamma} \mathcal{G}$. Since $\mathcal{P}(\Omega) \in \Gamma$, it follows $\Gamma \neq \emptyset$. By properities of set intesection, $\mathcal{F} \subset \mathcal{G}$ for all $\mathcal{G} \in \Gamma$. Thus, we need only verify that $\mathcal{F}$ is a σ-field.

We check properties (i) through (iii) of Definition 1.1.1. Property (i) follows since $\emptyset \in \mathcal{G}$ for all $\mathcal{G} \in \Gamma$, so $\emptyset$ is in $\bigcap_{\mathcal{G}\in\Gamma} \mathcal{G}$. Property (ii) follows similarly. Finally, suppose $A_1, A_2, \ldots$ is a sequence of elements of $\mathcal{F}$, then each $A_i \in \mathcal{G}$ for all $\mathcal{G} \in \Gamma$, so $\bigcup_{i=1}^{\infty} A_i \in \mathcal{G}$ for all $\mathcal{G} \in \Gamma$ by property (iii) of Definition 1.1.1 applied to each σ-field $\mathcal{G}$, and hence $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ by definition of intersection. This completes the proof.

□

On the real line $I\!R$ there is a special σ-field which we shall use often.

Definition 1.1.3 *The* Borel σ-field $\mathcal{B}$ *on $I\!R$ is the σ-field generated by the collection of all finite open intervals. In symbols,*

$$\mathcal{B} \;=\; \sigma\left(\{(a,b) : -\infty < a < b < \infty\}\right) \quad .$$

The elements of $\mathcal{B}$ are called Borel sets.

□

Proposition 1.1.2 *(a) The following are Borel sets: all open sets, all closed sets, all intervals (e.g. the half open interval $(a,b]$ or a semi-infinite interval (a,∞)), and all finite subsets of $I\!R$.*

(b) $\mathcal{B} = \sigma(\mathcal{O})$ where $\mathcal{O}$ is the collection of all open sets in $I\!R$, and $\mathcal{B} = \sigma(\mathcal{C})$ where $\mathcal{C}$ is the collection of all closed sets in $I\!R$.

Partial Proof. The proof that open sets are Borel sets depends on the following fact: any open set $U \subset I\!R$ can be expressed as a countable union of a of open intervals. Let

$$\mathcal{R} \;=\; \{(a,b) : a \text{ and } b \text{ are rational, and } (a,b) \subset U\}.$$

Since each $(a, b) \subset U$, clearly

$$\bigcup_{(a,b)\in\mathcal{R}} (a, b) \subset U.$$

If $x \in U$, then there is an open interval (c, d) such that $x \in (c, d) \subset U$ (this is the definition of an open set: every element of the set has a neighborhood contained in the set). We may find rational numbers a, b such that $c \leq a < x < b \leq d$, and then $(a, b) \in \mathcal{R}$ and so

$$x \in \bigcup_{(a,b)\in\mathcal{R}} (a, b).$$

As x was an arbitrary element of U, it follows that

$$U \subset \bigcup_{(a,b)\in\mathcal{R}} (a, b).$$

We have shown both inclusions, so

$$U \;=\; \bigcup_{(a,b)\in\mathcal{R}} (a, b).$$

Now, the collection rational numbers is countable (Exercise 1.1.6 (b)), and so also is the collections of ordered pairs of rational numbers (Exercise 1.1.6 (c)), and the collection $\mathcal{R}$ can be put in one to one correspondence with a subset of the collection of ordered pairs of rational numbers.

The remaining parts of the proof are left as an exercise (see Exercise 1.1.8).

□

1.1.2 Measures: Formal Definition.

Definition 1.1.4 *A measure space $(\Omega, \mathcal{F}, \mu)$ is a triple, consisting of an underlying set Ω, a σ-field $\mathcal{F}$, a function μ called the* measure *with domain $\mathcal{F}$ and satisfying the following:*

(i) $0 \;\leq\; \mu(A) \;\leq\; \infty$ *for all $A \in \mathcal{F}$;*

(ii) $\mu(\emptyset) \;=\; 0$*;*

(iii) *If A_1, A_2, ... is a sequence of disjoint elements of $\mathcal{F}$ (i.e. $A_i \cap A_j = \emptyset$ for all $i \neq j$), then*

$$\mu(\bigcup_{i=1}^{\infty} A_i) \;=\; \sum_{i=1}^{\infty} \mu(A_i) \quad .$$

A probability space *is a measure space $(\Omega, \mathcal{F}, P)$ for which $P(\Omega) = 1$. A measure on $(I\!R, \mathcal{B})$ is called a* Borel measure*.*

□

In probability theory, disjoint events are usually called *mutually exclusive.*

Remarks 1.1.3 Note that $\mu(A) = \infty$ is possible (unless $A = \emptyset$). In order for (iii) to make sense, we must know how to do arithmetic with ∞. For now, it suffices to know

$$\begin{aligned} \infty + x &= \infty \quad , \quad \text{for all } x \in I\!R \quad , \\ \infty + \infty &= \infty \quad . \end{aligned} \tag{1.2}$$

(Note that ∞ is not a real number.) Hence, if in (iii) in Definition 1.1.4, $\mu(A_j) = \infty$ for some j, then $\sum_i \mu(A_i) = \infty$. Of course, $\sum_i \mu(A_i)$ may equal ∞ even if all $\mu(A_i) < \infty$.

We will also need to know some multiplication rules for ∞ very soon, namely

$$\begin{aligned} a \cdot \infty &= \infty \quad , \quad \text{for all } a > 0 \quad , \\ 0 \cdot \infty &= 0 \quad . \end{aligned} \tag{1.3}$$

It is understood that addition and multiplication with ∞ is always commutative.

□

Now we consider some examples of measures.

Example 1.1.3 (Counting Measure) Let $(\Omega, \mathcal{F})$ be any measurable space. Let $A \in \mathcal{F}$ and define the measure $\#(A)$ = the number of elements of A. If A is an infinite set, then of course $\#(A) = \infty$. It is fairly easy to check that $(\Omega, \mathcal{F}, \#)$ is a measure space, i.e. that the three defining properties in Definition 1.1.4. (Whenever we say something like, "It is fairly easy to check ... ," the reader should take it upon himself or herself to check the claim!) Unless otherwise stated, we will use the power set for the σ–field when dealing with counting measure, i.e. $\mathcal{F} = \mathcal{P}(\Omega)$, the collection of all subsets of Ω.

□

Remarks 1.1.4 In the contexts of Definition 1.1.4 and Example 1.1.3, there is only one infinity, denoted ∞, which, together with its negative, is appended to the set of real numbers to "close" it. This is a different notion than that of infinity as a counting number discussed in Remarks 1.1.2. For counting measure in Example 1.1.3, "∞" is the extended real number ∞ since $\#$ is a measure. Property (iii) of Definition 1.1.4 is sometimes called the "countable additivity property" of a measure. Note that most of the unions and intersections above have been "countable" unions and intersections, i.e. we wrote $\bigcup_{i=1}^{\infty} A_i$ or something similar. The one exception is in the proof of Proposition 1.1.1. The intersection there (namely $\bigcap_{\mathcal{G}\in\Gamma} \mathcal{G}$) may be over an uncountable collection Γ of σ-fields. Further discussion of these issues can be found in Royden.

□

Example 1.1.4 (Unit Point Mass Measures) A special kind of measure which is often useful is the *unit point mass* at x. Given a measurable space $(\Omega, \mathcal{F})$ and $x \in \Omega$, put

$$\delta_x(A) \;=\; \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases}$$

It is easy to check that δ_x is a probability measure on $(\Omega, \mathcal{F})$.

Note that counting measure on $\{x_1, x_2, \ldots\}$ may be written in terms of unit point masses as

$$\# \;=\; \sum_i \delta_{x_i}.$$

See Proposition 1.1.5 for the fact that the sum of measures is a measure.

□

Elaborating somewhat on the last example, a useful intuitive interpretation of a general measure is as a "mass distribution". Indeed, think of $\Omega = I\!\!R$ as a very long rod with varying "mass density" and $\mu(A)$ as the amount of "mass" in a set $A \subset I\!\!R$. In the idealized situation of a unit mass at $x \in I\!\!R$ that takes up no space, we obtain δ_x. It could be argued that "mass theory" would be a better name for "measure theory".

Many of the measures we shall need for statistics are built from certain "elementary measures" using either product measures (Section 1.3) or densities (Section 1.4). The counting measure in Example 1.1.3 and the unit point masses of Example 1.1.4 will be useful in this context for defining discrete probability distributions, but the one given in the next theorem will be needed for the usual continuous distributions.

Theorem 1.1.3 (Existence of Lebesgue Measure) *There is a unique Borel measure m satisfying*

$$m([a,b]) \;=\; b - a \quad ,$$

for every finite closed interval $[a,b]$, $-\infty < a < b < \infty$.

□

The proof of this theorem is long and difficult. (See pp. 32–41 of Billingsley.) Intuitively, Lebesgue measure extends the notion of "length" to sets other than intervals. The content of the above theorem is that this notion can be extended to Borel sets, which is a very large collection of subsets of $I\!\!R$, although it does not include all subsets of $I\!\!R$. It is difficult to "construct" a subset of $I\!\!R$ which is not a Borel set, but it is sometimes difficult to prove that a set which is "obviously" a Borel set is in fact one. We shall generally ignore the issue, and in the future whenever we mention a subset of $I\!R$ it will be a Borel set, unless otherwise stated.

Thinking of Lebesgue measure as a "mass distribution," we see that it is a continuous distribution of mass which is "concentrated" on a line (the real line) and assigns one unit of mass per length.

The following results are proved on pp. 22-23 of Billingsley.

Proposition 1.1.4 (Basic Properties of Measures) *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space.*

(a) (Monotonicity) $A \subset B$ implies $\mu(A) \leq \mu(B)$, assuming A and B are in $\mathcal{F}$.

(b) (Subadditivity) If A_1, A_2, ... is any sequence of measurable sets, then

$$\mu(\bigcup A_i) \;\leq\; \sum \mu(A_i) \quad .$$

(c) If A_1, A_2, ... is a decreasing sequence of measurable sets (i.e. $A_1 \supset A_2 \supset \ldots$), and if $\mu(A_1) < \infty$, then

$$\mu(\bigcap_{i=1}^{\infty} A_i) \;=\; \lim_{i\to\infty} \mu(A_i) \quad .$$

□

The next proposition allows us to construct new measures from given measures.

Proposition 1.1.5 **(a)** *Let μ_1, μ_2, ... be a finite or infinite sequence of measures on $(\Omega, \mathcal{F})$. Suppose a_1, a_2, ... are nonnegative real numbers. Then $\mu = \sum_i a_i \mu_i$ is a measure on $(\Omega, \mathcal{F})$.*

(b) *Consider the same setup as in part (a). If each of the μ_i is a probability measure and if $\sum_i a_i = 1$, then μ is a probability measure.*

(c) *Let μ be a measure on $(\Omega, \mathcal{F})$ and let $A \in \mathcal{F}$. Define $\nu(B) = \mu(B \cap A)$ for all $B \in \mathcal{F}$. Then ν is a measure on $(\Omega, \mathcal{F})$.*

□

In the proposition, $\mu = \sum_i a_i \mu_i$ means that for any $A \in \mathcal{F}$, $\mu(A) = \sum_i a_i \mu_i(A)$. Note that if any $\mu(A_i) = \infty$, then our conventions from (1.3) are needed. The proof of this proposition is not difficult, and is an exercise (Exercise 1.1.10).

1.1.3 Distribution Functions.

If P is a probability measure (abbreviated *p.m.*) on $(I\!R, \mathcal{B})$, i.e. a *Borel p.m.*, then we can define

$$F(x) \;=\; P((-\infty, x]) \quad ,$$

for $-\infty < x < \infty$. F is called the *(cumulative) distribution function* (abbreviated *c.d.f.*) for P. (**Note:** Sometimes a probability measure itself is referred to as a *distribution*. In this text, we will never use "distribution" without the word "function" to refer to the c.d.f. Note that the probability measure and the c.d.f. are two different kinds of objects. One maps Borel sets to real numbers and the other maps real numbers to real numbers.)

Theorem 1.1.6 *(a) The c.d.f. of a Borel p.m. has the following properties:*

(i) $F(-\infty) = \lim_{x \to -\infty} F(x) = 0$.

(ii) $F(\infty) = \lim_{x \to \infty} F(x) = 1$.

(iii) *F is nondecreasing, i.e. if $x \leq y$ then $F(x) \leq F(y)$.*

(iv) *F is right continuous, i.e. $F(x+0) = \lim_{z \downarrow x} F(z) = F(x)$. (Here, $\lim_{z \downarrow x}$ means the limit is taken through values $z > x$.)*

(b) Suppose $F : I\!\!R \longrightarrow I\!\!R$ is any function satisfying (i) through (iv) above, then F is the c.d.f. of a unique Borel p.m. on $I\!\!R$.

□

The proof of part (a) follows easily from previously stated properties of measures, but the proof of part (b) is very difficult, similarly to the proof of Theorem 1.1.3, which also asserts the existence of a Borel measure. See Billingsley, Theorem 14.1, p. 190.

Figure 1.1 shows the graph of the c.d.f. F of the probability measure $P = \frac{1}{2}P_u + \frac{1}{2}\delta_{1/2}$ where P_u is the uniform distribution on $[0,1]$ introduced in Example 1.1.2. Note that there is a jump discontinuity in $F(x)$ at $x = 1/2$. The filled circle on the top branch of the graph indicates that $F(1/2) = 3/4$ whereas the open circle on the lower branch shows $F(1/2 - 0) = 1/4$. Here, we use $F(x-0)$ to denote the limit from the left or from below:

$$F(x-0) \;=\; \lim_{z \uparrow x} F(z) \;=\; \lim_{z \to x,\, z < x} F(z).$$

The magnitude of the jump $F(1/2) - F(1/2 - 0) = 1/2$ shows that the point $x = 1/2$ has a probability of $1/2$, i.e. $P(\{1/2\}) = 1/2$. Also, the c.d.f. is flat over an interval if that interval has no probability measure. In this case, the c.d.f. is flat over any interval $[a,b] \subset [1,\infty)$ or $[a,b] \subset (-\infty, 0]$.

The inverse of a c.d.f. or quantile function is very useful in statistics. Strictly speaking, $F^{-1}(u)$ is defined for all $u \in (0,1)$ if and only if F is continuous and strictly increasing. However, we can get around this and define a quantile function for an arbitrary distribution as shown next.

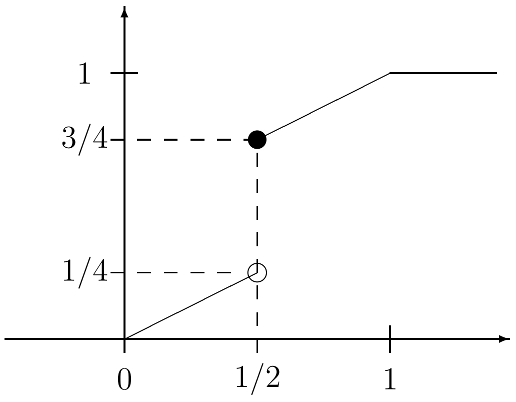

Figure 1.1: Graph of a c.d.f.

Definition 1.1.5 *Let F be any c.d.f. For $0 \leq \alpha \leq 1$, the* lower quantile function *and* upper quantile function *for F are given by*

$$\begin{aligned} F^{-}(\alpha) &= \inf\{\, x \,:\, F(x) \geq \alpha \,\}, \\ F^{+}(\alpha) &= \sup\{\, x \,:\, F(x) \leq \alpha \,\}, \end{aligned}$$

respectively. We adopt the conventions that

$$\inf \emptyset \;=\; +\infty \quad \textit{and} \quad \sup \emptyset \;=\; -\infty.$$

□

It is easy to show (see Exercise 1.1.15)

$$F^{-}(\alpha) \;\leq\; F^{+}(\alpha),\ \forall \alpha \in (0,1).$$

Also, $F^{-}(\alpha) = F^{+}(\alpha) = x$ if for all $\epsilon > 0$ there exist $x_1 \in (x-\epsilon, x)$ and $x_2 \in (x, x+\epsilon)$ with $F(x_1) < F(x) < F(x_2)$. When this happens, we say x is a *point of increase* for F. Furthermore, if F is continuous and strictly increasing, then $F^{-}(\alpha) = F^{+}(\alpha) = F^{-1}(\alpha)$ for all $\alpha \in (0,1)$.

The *median* of a c.d.f. is defined as

$$\mathrm{Med}(F) \;=\; \frac{1}{2}\left[F^{-}(1/2) \,+\, F^{+}(1/2) \right].$$

One can unambiguously define other quantiles as averages of the upper and lower quantiles.

We use the c.d.f. depicted in Figure 1.2 to illustrate the lower and upper quantile functions. Note that F is constant on the intervals $(-\infty, 0]$, $[1,2]$, and $[3.1, \infty)$. The reader should verify the following claims:

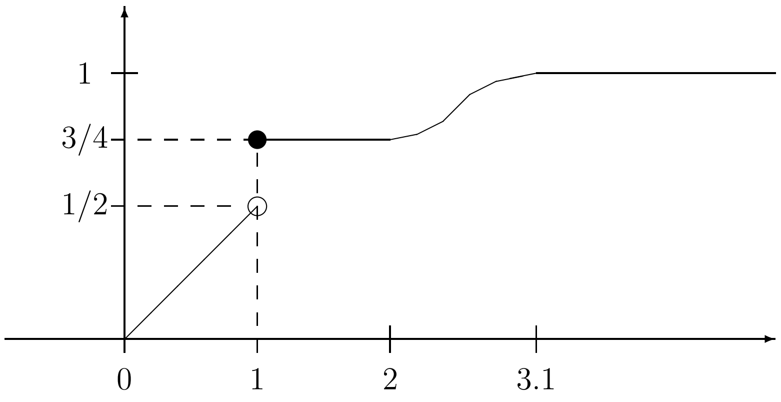

Figure 1.2: Example c.d.f. for discussion of quantile functions.

(i) For $\alpha = 0$, $F^-(\alpha) = -\infty$ and $F^+(\alpha) = 0$.

(ii) For $0 < \alpha < 1/2$, $F^-(\alpha) = F^+(\alpha) = F^{-1}(\alpha)$.

(iii) For $1/2 \leq \alpha < 3/4$, $F^-(\alpha) = F^+(\alpha) = 1$, but there is no value of x such that $F(x) = \alpha$ in this case.

(iv) $F^-(3/4) = 1$ and $F^+(3/4) = 2$.

(v) For $3/4 < \alpha < 1$, $F^-(\alpha) = F^+(\alpha) = F^{-1}(\alpha)$.

(vi) $F^-(1) = 3.1$ and $F^+(1) = \infty$.

1.1.4 Empirical Distributions.

Now we introduce an important notion in statistics. Suppose $(x_1, x_2, \ldots, x_n)$ is a *data set* where the x_i's are elements of some set Ω, e.g. $\Omega = I\!R$. We denote the data set as an ordered n–tuple and not a set since we wish to keep track of replications. That is to say, if $x_i = x_j$ for some $i \neq j$, then we count the value x_i twice (at least; more times if there are further replications), whereas replicated values do not count in a set. For instance $\{1, 2, 2\} = \{1, 2\}$, as sets, but $(1, 2, 2)$ $\neq (1, 2)$. Another way to think of an ordered n–tuple is as a finite sequence of length n. However, the ordering within the n–tuple is often not important in statistics. The *empirical distribution* of the data set is the probability measure

$$\begin{aligned} \hat{P}_n(A) &= \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}(A) \\ &= \frac{1}{n}\#\{i : x_i \in A\} \quad . \end{aligned} \tag{1.4}$$

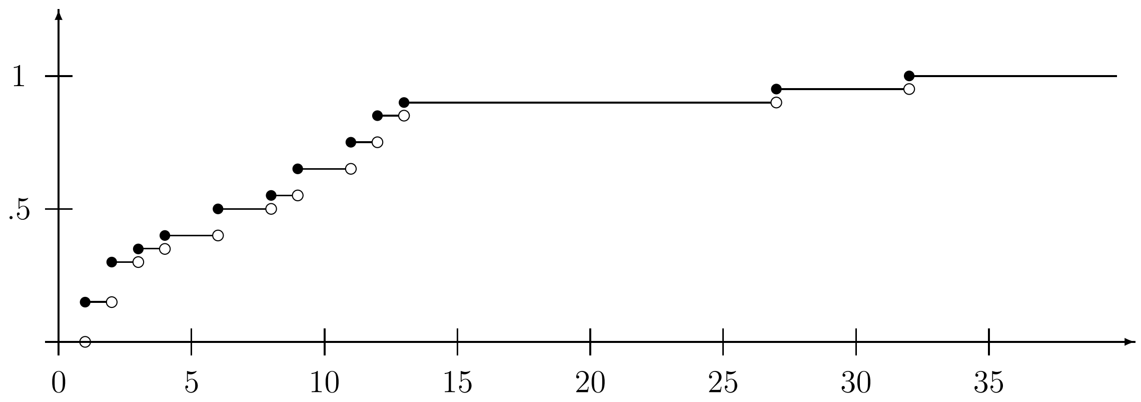

Figure 1.3: Empirical c.d.f. of times between airplane crashes.

Note that $\hat{P}_n(A)$ is the relative frequency of A, i.e. the proportion of observations that lie in A. We can recover the data set except for the ordering from $\hat{P}_n$, so if the order is immaterial, the empirical distribution is just as good as the original data set. If the x_i's are real valued, then the c.d.f. corresponding to $\hat{P}_n$ is called the *empirical c.d.f.* or *empirical distribution function* and will be denoted $\hat{F}_n$.

As an example, there were 21 major commercial airline crashes in the 15 year period ending 8 September 1994, and the times between crashes rounded off to the nearest month are $(2, 27, 13, 1, 12, 4, 2, 11, 9, 1, 3, 12, 2, 9, 1, 6, 8, 11, 6, 32)$, or in sorted order $(1, 1, 1, 2, 2, 2, 3, 4, 6, 6, 8, 9, 9, 11, 11, 12, 12, 13, 27, 32)$. The corresponding empirical distribution function is shown in Figure 1.3.

Empirical distributions are useful for many purposes. For one, they allow us to apply concepts from probability to data analysis. We have already introduced the notion of quantiles, and in dealing with real valued data we can define the *lower and upper α sample quantiles* as $\hat{F}_n^-(\alpha)$ and $\hat{F}_n^+(\alpha)$, respectively. Of course, the sample median is just $\mathrm{Med}(\hat{F}_n)$. In the airplane intercrash time data set introduced above, the median is $(6+8)/2 = 7$ months, since there are 20 observations in all and the 10^{th} and 11^{th} largest observations are 6 and 8 months. As we learn more about probability measures, we will acquire tools that can be used on data.

There is however a deeper connection between data and probability theory via empirical distributions. We have already mentioned the long run relative frequency interpretation of probability in Example 1.1.1. With the notation introduced here, we may state this as follows. Suppose we have a probability space $(\Omega, \mathcal{F}, P)$ which is to "model" some "experiment" that can be repeated indefinitely under identical conditions. It is understood that the "outcome" of each trial of the experiment is an element x of Ω. Given the series of outcomes $(x_1, x_2, \ldots, x_n)$ (where it is understood that x_i is the outcome from the i^{th} trial), we can form the empirical measure $\hat{P}_n$ as above. Then the model is "valid" if

$$\lim_{n\to\infty} \hat{P}_n(A) \; = \; P(A) \quad , \quad \text{for all } A \in \mathcal{F}. \tag{1.5}$$

That is, P "correctly models" the experiment if the long run relative frequencies converge to P. We shall refer to this as the *frequentist interpretation* of probability.

We hasten to mention that this is just one way of trying to "connect" probability models with the "real world," and has nothing to do *per se* with probability theory as a mathematical discipline (although the frequentist notion that probability is long run relative frequency was the motivation for founding probability theory in the first place). In fact, if one insists on being very realistic, there is no way to verify (1.5) since one cannot repeat any real "experiment" infinitely many times. In practice, one must be satisfied if $P(A)$ is a "good approximation" to $\hat{P}_n(A)$ for some relatively large number n and some class of sets A which is not the whole σ–field. Indeed, one can never know the "true model" or even if some real world process is "truly random" (and hence governable by the laws of probability). A useful maxim to keep in mind when doing statistics is, "All models are false but some are useful."

We should mention that probability theory is used to model other things, such as degrees of subjective belief. For instance, if I say, "the chance it will rain today is 25%," this may be based solely on my subjective belief about the likelihood of rain and have nothing to do with a probability model based on many observations. It could mean that if someone wants to bet me that it will rain, I will bet \$3 to his \$1.

We will have much occasion to return to the notion of empirical distributions and the philosophical issues raised here.

1.1.5 References

We shall "key" our discussion to Billingsley's text *Probability and Measure.* Other good texts for an introduction to measure theory are *Real Analysis* by Royden, *Real Analysis and Probability* by Ash, and *A Course in Probability Theory* by Chung. Several statistics texts give a sketchy introduction, such as *Linear Statistical Inference and Its Applications* by Rao and the Lehmann volumes, *Theory of Point Estimation* and *Testing Statistical Hypotheses.* One may also find a brief introduction to measure theory in *Principles of Mathematical Analysis* by Rudin.

Exercises for Section 1.1.

1.1.1 Prove DeMorgan's laws: If $\mathcal{A}$ is a collection of sets, then

$$\left[\bigcap_{A\in\mathcal{A}} A\right]^c = \bigcup_{A\in\mathcal{A}} A^c \quad , \tag{1.6}$$

and

$$\left[\bigcup_{A\in\mathcal{A}} A\right]^c = \bigcap_{A\in\mathcal{A}} A^c \quad .$$

Hint: Recall that

$$\bigcap_{A\in\mathcal{A}} A = \{x : x \in A \text{ for all } A \in \mathcal{A}\}.$$

Thus, x is an element of the l.h.s. of (1.6) if and only if there is at least one $A_1 \in \mathcal{A}$ such that x is not an element of A_1, in which case $x \in A_1^c$ and hence x is an element of the r.h.s. of (1.6) (since the union of a collection of sets is the set of elements which are in any one of them, and we have shown that x is in one of them, namely A_1.) This shows the l.h.s. of (1.6) is a subset of the r.h.s. (since any element of the l.h.s. has been shown to belong to the r.h.s.). Now show the r.h.s. is a subset of the l.h.s. and similarly for the second of DeMorgan's laws.

1.1.2 (a) Show that the trivial σ-field is the smallest possible σ-field on Ω in that (i) it is a σ-field, and (ii) it is contained in any other σ-field on Ω.

(b) Similarly to part (a), show that the power set is the largest σ-field on Ω.

1.1.3 Suppose A and B are nonempty subsets of Ω with $A \cup B \neq \Omega$, $A \cap B \neq \emptyset$, and neither is a subset of the other. Show that $\sigma(\{A, B\})$ contains 16 elements, and give an explicit listing of the elements. (Hint: partition Ω into $A_1 = A \cap B^c$, $A_2 = A \cap B$, $A_3 = A^c \cap B$, and $A_4 = A^c \cap B^c$.) What happens if $A \subset B$?

1.1.4 Suppose A_1, A_2, ... , A_n is a partition of Ω into nonempty sets (by a *partition*, we mean $A_i \cap A_j = \emptyset$ for all i and j, and $\bigcup_{i=1}^n A_i = \Omega$). Show that the σ-field generated by $\{\ A_1,\ A_2,\ \ldots\ ,\ A_n\ \}$ contains 2^n elements and give a description of the general element.

1.1.5 Verify that the unit point mass of Example 1.1.4 is a measure.

1.1.6 (a) Suppose $A = \{a_1, a_2, \ldots\}$ and $B = \{b_1, b_2, \ldots\}$ are countably infinite sets. Let $C = A \times B = \{(a, b) : a \in A \text{ and } b \in B\}$ be the Cartesian product. Show that C is countable. (Hint: Recognize the pattern in this listing of C as a sequence: $\{\ (a_1, b_1),\ (a_1, b_2),\ (a_2, b_1),\ (a_1, b_3),\ (a_2, b_2),\ (a_3, b_1),\ (a_1, b_4),\ \ldots\ \}$. We claim that C can be listed as $\{c_1, c_2, \ldots\}$ where $c_k = (a_i, b_j)$ with k being given by a function of (i, j).

(b) Show that the collection of rational numbers is a countable set. Hint: A rational number is of the form p/q, where p is an integer and q is a positive integer, and p and q have no nontrivial factors (so the fraction is in lowest terms).

1.1.7 Assuming initially that the only sets which you know are Borel sets are the finite closed intervals, show the following are Borel sets by using the properties of σ-fields.

(a) Any singleton set $\{a\}$.

(b) Any finite open interval (a, b).

(c) A semiopen interval $(a, b]$.

(d) A semi-infinite interval (a, ∞).

(e) A finite set.

(f) The set of rational numbers. You may assume the truth of Exercise 1.1.6 for this problem.

1.1.8 Complete the proof of Proposition 1.1.2.

1.1.9 Prove Proposition 1.1.4 using the defining properties of a measure.

1.1.10 Prove Proposition 1.1.5 (a) and (b).

1.1.11 Assuming initially only that you know that the Lebesgue measure of a finite closed interval is its length (as in Theorem 1.1.3), find the Lebesgue measure of each of the Borel sets in Exercise 1.1.7.

1.1.12 (a) Let $(\Omega, \mathcal{F})$ be a measurable space and $\Omega_0 \subset \Omega$. Put

$$\mathcal{F}_0 = \{A \cap \Omega_0 : A \in \mathcal{F}\} \quad .$$

Show that $(\Omega_0, \mathcal{F}_0)$ is a measurable space.

(b) Let $(\Omega, \mathcal{F})$ be a measurable space and $\Omega_0 \in \mathcal{F}$. Put

$$\mathcal{F}_1 = \{A : A \in \mathcal{F} \text{ and } A \subset \Omega_0\} \quad .$$

Show that $(\Omega_0, \mathcal{F}_1)$ is a measurable space.

(c) Let $\Omega_0 \in \mathcal{F}$ as in part (b) and define $\mathcal{F}_0$ and $\mathcal{F}_1$ as above. Show $\mathcal{F}_0 = \mathcal{F}_1$.

(d) Let $(\Omega, \mathcal{F})$, Ω_0, and $\mathcal{F}_1$ be as in part (b). Let μ be a measure on $(\Omega, \mathcal{F})$ and define

$$\mu_1(A) = \mu(A) \quad , \quad \text{for all } A \in \mathcal{F}_1. \tag{1.7}$$

Show that $(\Omega_0, \mathcal{F}_1, \mu_1)$ is a measure space.

(e) Let $(\Omega, \mathcal{F})$, Ω_0 be as in part (b), and μ as in part (d). Define

$$\mu_2(A) = \mu(A \cap \Omega_0) \quad , \quad \text{for all } A \in \mathcal{F}. \tag{1.8}$$

Show that $(\Omega, \mathcal{F}, \mu_2)$ is a measure space. Note that this proves Proposition 1.1.5 (c).

(f) Show that if $A \in \mathcal{F}$ and $A \subset \Omega_0$, then $\mu_1(A) = \mu_2(A)$.

(g) Assume that $0 < \mu(\Omega_0) < \infty$ and define

$$P_1(A) = \frac{\mu(A)}{\mu(\Omega_0)} \quad , \quad \text{for all } A \in \mathcal{F}_1, \tag{1.9}$$

$$P_2(A) = \frac{\mu(A \cap \Omega_0)}{\mu(\Omega_0)} \quad , \quad \text{for all } A \in \mathcal{F}. \tag{1.10}$$

Show that P_1 and P_2 are probability measures on suitable measurable spaces.

1.1.13 Let $\theta = [\theta_1, \theta_2]$ with $-\infty < \theta_1 < \theta_2 < \infty$ be a given closed interval. For Borel sets B let

$$P_\theta(B) = \frac{m(B \cap \theta)}{\theta_2 - \theta_1} \quad .$$

Show that for each such θ, $(I\!\!R, \mathcal{B}, P_\theta)$ is a probability space. (Note: The P_θ defined here is called the *uniform distribution* on the interval θ.)

1.1.14 Prove Theorem 1.1.6 (a).

1.1.15 Verify the claims made about the c.d.f. pictured in Figure 1.2.

1.1.16 Show the following facts about the lower and upper quantile functions.

(i) $F^-(\alpha) \leq F^+(\alpha)$ for all $\alpha \in (0, 1)$.

(ii) $F^-(\alpha) = F^+(\alpha)$ if and only if $F^-(\alpha)$ is a point of increase for F.

(iii) If $F(x)$ is continuous and strictly increasing for all $x \in I\!\!R$, then $F^-(\alpha) = F^+(\alpha) = F^{-1}(\alpha)$ for all $\alpha \in (0, 1)$.

(iv) For any c.d.f. F, $F^-(0) = -\infty$ and $F^+(1) = \infty$.

(v) $F(F^-(\alpha) - 0) \leq \alpha$ but $F(F^-(\alpha)) \geq \alpha$. Also the same statements hold if F^- is replaced with F^+.

(vi) $F^-(F(x)) \leq x \leq F^+(F(x))$.

1.1.17 (a) Suppose $\underline{x} = (x_1, \ldots, x_n)$ is a univariate sample (i.e. each x_i is a single real number) and let $\hat{F}_n$ be the corresponding empirical c.d.f. Let $\underline{y} = (y_1, \ldots, y_n)$ be the ordered x_i's, i.e. the same values appear in $\underline{y}$ as in $\underline{x}$ (replicated the same number of times) but $y_1 \leq y_2 \leq \ldots \leq y_{n-1} \leq y_n$. Show that $\hat{F}_n^-(i/n) = y_i$ and $\hat{F}_n^+(i/n) = y_{i+1}$. For what range of i is each of these equations valid.

(b) If $\underline{x}$ and $\underline{y}$ are as given in part (a), define the *sample median* of $\underline{x}$ by

$$\text{med}(\underline{x}) \;=\; \begin{cases} y_{(n+1)/2} & \text{if } n \text{ is odd,} \\ (y_{n/2} + y_{(n+1)/2})/2 & \text{if } n \text{ is even.} \end{cases}$$

Show $\text{med}(\underline{x}) = \hat{F}_n^-(1/2) = \hat{F}_n^+(1/2)$ if n is odd and $\text{med}(\underline{x}) = [\hat{F}_n^-(1/2) + \hat{F}_n^+(1/2)]/2$ if n is even, and in either event $\text{med}(\underline{x}) = \text{Med}(\hat{F}_n) = [F^-(1/2) + F^+(1/2)]/2$.

1.1.18 For the data set of times between airplane crashes, determine $F^+(\alpha)$ and $F^-(\alpha)$ for $\alpha = .25$, .5, and .75.

1.2 Measurable Functions and Integration.

In the previous section, a measure μ was defined as a real valued function μ defined on a class of subsets $\mathcal{F}$ of Ω. Now every set $A \subset \Omega$ is associated with a unique real valued function called the *indicator function of* A. It is given by

$$I_A(x) \;=\; \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases}$$

(Indicator are usually called *characteristic functions* by mathematicians, but we shall reserve the term "characteristic function" to mean something very different.) Thus, we may think of a measure as being defined on the class of indicator functions of sets $A \in \mathcal{F}$. Instead of writing $\mu(A)$, we could write "$\mu(I_A)$." In this section we define the abstract notion of integration which extends the definition of μ to a large class of real valued functions, i.e. we can define "$\mu(f)$," usually written $\int f\,d\mu$ (so that $\mu(A) = \int I_A\,d\mu$). First, we must consider the class of functions to which the definition will apply.

1.2.1 Measurable Functions.

Consider two sets Ω and Λ. Let $f : \Omega \longrightarrow \Lambda$ be any function and $A \subset \Lambda$. Then the *inverse image of* A *under* f is

$$f^{-1}(A) \;=\; \{\omega \in \Omega : f(\omega) \in A\} \quad .$$

Note that $f^{-1}(A)$ is defined even if the inverse function f^{-1} does not exist. The inverse image operation has some nice properties, such as

$$f^{-1}(A^c) \;=\; [f^{-1}(A)]^c \quad , \tag{1.11}$$

for any $A \subset \Lambda$. Also, if A_1, A_2, ... are subsets of Λ, then

$$f^{-1}\left(\bigcup_i A_i\right) \;=\; \bigcup_i f^{-1}(A_i) \quad . \tag{1.12}$$

To prove (1.12), take $\omega \in f^{-1}\left(\bigcup_i A_i\right)$. By definition of the inverse image, $f(\omega) \in \bigcup_i A_i$, and by definition of set union, $f(\omega) \in A_n$ for some n. From this latter, $\omega \in f^{-1}(A_n)$ some n, and this implies $\omega \in \bigcup_i f^{-1}(A_i)$. So we have shown that $f^{-1}\left(\bigcup_i A_i\right) \subset \bigcup_i f^{-1}(A_i)$. A similar argument shows the reverse inclusion.

A statement similar to (1.12) but with $\cap$ replacing $\cup$ holds as well (use DeMorgan's laws). One can also define the *forward image* for $B \subset \Omega$ by

$$f(B) \;=\; \{f(\omega) : \omega \in B\} \quad ,$$

but the forward image does not have the nice properties above and is not so useful.

If $\mathcal{A} \subset \mathcal{P}(\Lambda)$, i.e. $\mathcal{A}$ is a collection of subsets of Λ, then we write

$$f^{-1}(\mathcal{A}) \;=\; \{f^{-1}(A) : A \in \mathcal{A}\} \quad ,$$

to denote all inverse images of sets in $\mathcal{A}$.

Definition 1.2.1 *Let* $(\Omega, \mathcal{F})$ *and* $(\Lambda, \mathcal{G})$ *be measurable spaces and* $f : \Omega \longrightarrow \Lambda$ *a function. Then f is a* measurable function *iff* $f^{-1}(\mathcal{G}) \subset \mathcal{F}$*, in which case we shall write* $f : (\Omega, \mathcal{F}) \longrightarrow (\Lambda, \mathcal{G})$*. If* $\Lambda = I\!R$ *and* $\mathcal{G}$ *is the Borel* σ*-field, then we say* f *is* Borel measurable *or a* (real valued) Borel function*.*

In probability theory, a measurable function is called a random object *or* random element *and usually denoted* X*,* Y*, ..., and a real valued Borel function is called a* random variable *(abbreviated* r.v.*).*

□

If $f : (\Omega, \mathcal{F}) \longrightarrow (\Lambda, \mathcal{G})$ then one can check that $f^{-1}(\mathcal{G})$ is a σ-field on Ω, and since $f^{-1}(\mathcal{G}) \subset \mathcal{F}$, it is a *sub-*$\sigma$*-field* of $\mathcal{F}$.

Definition 1.2.2 *If* $f : (\Omega, \mathcal{F}) \rightarrow (\Lambda, \mathcal{G})$*, then the* σ–field generated by f *is* $f^{-1}(\mathcal{G})$ *and is denoted* $\sigma(f)$*.*

□

Just as all subsets of $I\!R$ that arise in "practice" are Borel sets, it is also the case that all real valued functions that arise in "practice" are Borel functions. Furthermore, any way of constructing new functions from Borel functions leads to Borel functions, in "practice".

Now we consider some examples of Borel functions. Let $(\Omega, \mathcal{F})$ be any measurable space and $A \in \mathcal{F}$, i.e. $A \subset \Omega$ is measurable. We will determine $\sigma(I_A)$. Let $B \subset I\!R$, then

$$I_A^{-1}(B) \;=\; \begin{cases} \emptyset & \text{if } 0, 1 \notin B, \\ A & \text{if } 1 \in B \text{ but } 0 \notin B, \\ A^c & \text{if } 0 \in B \text{ but } 1 \notin B, \\ \Omega & \text{if both } 1 \in B \text{ and } 0 \in B. \end{cases} \tag{1.13}$$

Since $I_A^{-1}(\mathcal{B}) = \{\emptyset,\ A,\ A^c,\ \Omega\} \subset \mathcal{F}$, it follows that I_A is a Borel function (note that we are using measurability of A here). In this example, $\sigma(I_A) = \sigma(\{A\})$.

The class of *simple functions* is obtained by taking linear combinations of indicators, i.e. a generic simple function has the form

$$\phi(\omega) \;=\; \sum_{i=1}^{n} a_i I_{A_i}(\omega) \quad ,$$

where A_1, A_2, ... A_n are in $\mathcal{F}$ and a_1, a_2, ... a_n are real numbers. Here, n is any finite positive integer. One can show directly that such a simple function is measurable, but if follows easily from the next result.

Proposition 1.2.1 *All functions below have domain Ω and range $I\!\!R$, and $(\Omega, \mathcal{F})$ is a measurable space.*

(a) f is Borel iff $f^{-1}((a, \infty)) \in \mathcal{F}$ for all $a \in I\!\!R$.

(b) If f and g are Borel, then so are $f+g$ and fg. Also, f/g is Borel provided $g(\omega) \neq 0$ for all $\omega \in \Omega$. In particular, linear combinations $af + bg$ $(a, b \in I\!\!R)$ of Borel functions are Borel.

(c) Suppose g, f_1, f_2, ... are Borel. Let

$$L \;=\; \{\,\omega \in \Omega : \lim_{n\to\infty} f_n(x) \textit{ exists}\,\} \quad .$$

Then L is a measurable set in Ω and the function

$$h(x) \;=\; \begin{cases} \lim_{n\to\infty} f_n(x) & \textit{if } x \in L, \\ \\ g(x) & \textit{if } x \notin L. \end{cases}$$

is Borel.

□

See Theorem 13.4, pp. 183–185 of Billingsley for a proof of the above result.

Proposition 1.2.2 *Let $\Omega \subset I\!\!R$ be a Borel set and let*

$$\mathcal{F} \;=\; \{\,\Omega \cap B \;:\; B \in \mathcal{B}\,\} \quad .$$

$\mathcal{F}$ is a σ-field on Ω and if $f : \Omega \longrightarrow I\!\!R$ is continuous at all points of Ω, then f is a Borel function.

□

The last proposition is Theorem 10.1, p. 156 of Billingsley. It is left to the reader to verify that $\mathcal{F}$ defined in the last proposition is a σ–field, and in fact the same one as given in Exercise 1.1.12 with Ω_0 replaced by Ω and Ω of that exercise replaced by $I\!\!R$.

Proposition 1.2.3 *Suppose*

$$f : (\Omega, \mathcal{F}) \longrightarrow (\Lambda, \mathcal{G})\ , \quad g : (\Lambda, \mathcal{G}) \longrightarrow (\Xi, \mathcal{H}) \quad .$$

Then the composite function $h = g \circ f$ is measurable $(\Omega, \mathcal{F}) \longrightarrow (\Xi, \mathcal{H})$. (Recall that $(g \circ f)(\omega) = g(f(\omega))$).

Proof. Let $C \in \mathcal{H}$, and we will show that $h^{-1}(C) \in \mathcal{F}$. Now we claim (see Exercise 1.2.1) that $h^{-1}(C) = (g \circ f)^{-1}(C) = f^{-1}(g^{-1}(C))$, and since $g^{-1}(C) \in \mathcal{G}$ by measurability of g, it follows that $f^{-1}(g^{-1}(C)) \in \mathcal{F}$ by measurability of f, which is what was needed.

□

1.2.2 Induced Measures.

Now we show how measurable functions can be used to construct measures. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, $(\Lambda, \mathcal{G})$ a measurable space, and $f : (\Omega, \mathcal{F}) \longrightarrow (\Lambda, \mathcal{G})$ a measurable function. Define a function $\mu \circ f^{-1}$ on $\mathcal{G}$ by

$$(\mu \circ f^{-1})(C) \; = \; \mu(f^{-1}(C)) \quad , \quad C \in \mathcal{G} \quad .$$

Note that $f^{-1}(C) \in \mathcal{F}$ so the r.h.s. above is well defined. We claim that $\mu \circ f^{-1}$ is a measure. The first two properties of Definition 1.1.4 are left as an exercise, so we verify the third. Let C_1, C_2, ... be a sequence of disjoint sets in $\mathcal{G}$, then $f^{-1}(C_1)$, $f^{-1}(C_2)$, ... are disjoint (by the analog of (1.12) for intersections) and are in $\mathcal{F}$, so

$$\mu(\bigcup_{i=1}^{\infty} f^{-1}(C_i)) \; = \; \sum_{i=1}^{\infty} (\mu \circ f^{-1})(C_i)$$

but by (1.12)

$$\bigcup_{i=1}^{\infty} f^{-1}(C_i) \; = \; f^{-1}(\bigcup_{i=1}^{\infty} C_i)$$

which shows $(\mu \circ f^{-1})(\bigcup_i C_i) = \sum (\mu \circ f^{-1})(C_i)$, and hence that $(\mu \circ f^{-1})$ satisfies property (iii) of Definition 1.1.4. We call $(\mu \circ f^{-1})$ *the measure induced by* f.

Just as a clarifying remark, note that measurable function $f : (\Omega, \mathcal{F}) \longrightarrow (\Lambda, \mathcal{G})$ pulls a σ-field backwards (i.e. $\sigma(f) \subset \mathcal{F}$ on the domain space) but takes a measure μ on $(\Omega, \mathcal{F})$ forwards (i.e. $\mu \circ f^{-1}$ is a measure on the range space $(\Lambda, \mathcal{G})$).

If $\mu = P$ is a probability measure and X is a r.v., then $P \circ X^{-1}$ is called the *distribution* of X or the *law* of X and is sometimes denoted P_X or Law$[X]$. It should be kept in mind that although we emphasize the role of the r.v. in this latter notation, the distribution of X also depends on the underlying p.m. P. There is much confusion that arises in probability because the underlying probability space (measure space) is typically suppressed in notation. Let X be a r.v. on the underlying probability space $(\Omega, \mathcal{F}, P)$. Let $B \subset I\!R$ be a Borel set. In general, we will use square brackets to denote events wherein the underlying measure space has been suppressed, as in

$$[X \in B] \; = \; \{\omega \in \Omega : X(\omega) \in B\} \; = \; X^{-1}(B).$$

Note the following:

$$P\{\omega \in \Omega : X(\omega) \in B\} \; = \; P[X \in B] \; = \; P \circ X^{-1}(B) \; = \; P_X(B). \qquad (1.14)$$

It would be technically incorrect to write "$P_X[X \in B]$," or "$P(B)$," in this context. Whenever in doubt, it is helpful to write out the event in detail as in the first expression in (1.14).

The notions of random variables and their distributions have profound import in applied statistics, at least from a conceptual point of view. Statistics is used in a wide range of other disciplines, from precise measurements in the physical sciences to more variable and fuzzy measurements in the biological and social sciences. It is amazing that one subject can have such a diversity of application. One of the main reasons it is so widely useful is that virtually all of the data that is collected in these various subjects is numerical. Whether the observation results from a randomly selected person recording his or her preference on a one to five scale, or from careful measurement on a replication of a chemical experiment, we deal with numerical data. Because of this, it is generally not necessary to concern oneself too much with the "underlying probability space" but only with the distribution of the random variable being observed. Thus, statisticians can concentrate mostly on the study of Borel probability measures on the reals, or a cartesian product of the real line with itself.

The notions of induced measures is also very important in Monte Carlo simulation. The following result is sometimes called the "fundamental theorem of simulation."

Proposition 1.2.4 *Let U be a random variable which is uniformly distributed on $[0,1]$, i.e. the c.d.f. of U is given by*

$$P[U \leq u] = \begin{cases} 0 & \textit{if } u < 0, \\ u & \textit{if } 0 \leq u \leq 1, \\ 1 & \textit{if } u > 1. \end{cases}$$

Let F be any c.d.f. Then the random variable $X = F^-(U)$ has c.d.f. F.

Proof. We shall claim that $\forall x \in I\!R$, the event $[X \leq x] = [U \leq F(x)]$. Then from the formula for the c.d.f. of U we obtain that $P[X \leq x] = P[U \leq F(x)] = F(x)$.

We first show that $u \leq F(x)$ is necessary and sufficient $F^-(u) \leq x$. Now $F^-(u) = \inf\{y : F(y) \geq u\}$, so if $u \leq F(x)$, then $x \in \{y : F(y) \geq u\}$ and hence $x \geq F^-(u)$ by definition of the infimum. This establishes the sufficiency. On the other hand, if $F(x) < u$, then by right continuity of F there is an $\epsilon > 0$ such that $F(x+\epsilon) < u$, and then $x+\epsilon \notin \{y : F(y) \geq u\}$, which implies $x+\epsilon \leq F^-(u)$, and we conclude that $x < F^-(u)$. By contraposition, this yields $F^-(u) \leq x$ implies $u \leq F(x)$.

Now to establish the original claim, note that $[X \leq x] = \{\omega \in \Omega : X(\omega) \leq x\}$, $= \{\omega \in \Omega : F^-(U(\omega)) \leq x\}$. By the previous paragraph, for all ω, $F^-(U(\omega)) \leq x$ if and only if $U(\omega) \leq F(x)$, which is to say $\{\omega \in \Omega : F^-(U(\omega)) \leq x\} = \{\omega \in \Omega : U(\omega) \leq F(x)\}$, as claimed.

□

Using this result, we can start with random numbers which are uniformly distributed on $[0,1]$ and generate random numbers with any distribution, provided we can compute F^-. In fact, in many cases one can generate random numbers from a given distribution more efficiently than the computation of $F^-(U)$.

1.2.3 The General Definition of an Integral.

It is sometimes useful to extend the real number system to include $\pm\infty$, and we define $\bar{I\!R} = [-\infty, +\infty] = I\!R \cup \{-\infty, \infty\}$ to be the *extended real number system.* Some arithmetic properties with $\pm\infty$ are defined by

$$\pm\infty + a \;=\; a + (\pm\infty) \;=\; \pm\infty\;, \quad a \in I\!R$$

$$a \cdot (\pm\infty) \;=\; (\pm\infty) \cdot a \;=\; \begin{cases} \pm\infty & \text{if } a > 0, \\ 0 & \text{if } a = 0, \\ \mp\infty & \text{if } a < 0. \end{cases} \tag{1.15}$$

$$\infty + \infty \;=\; \infty \quad , \quad \infty \cdot \infty \;=\; \infty \quad .$$

However, $\infty - \infty$ *is always undefined* (this is important!). Also

$$a/0 \;=\; \begin{cases} +\infty & \text{if } 0 < a \leq \infty, \\ \text{undefined} & \text{if } a = 0, \\ -\infty & \text{if } -\infty \leq a < 0. \end{cases} \tag{1.16}$$

Let $\bar{\mathcal{B}}$, the *extended Borel σ–field* on $\bar{I\!R}$ be the the collection of all sets of the form B, $B\cup\{\infty\}$, $B\cup\{-\infty\}$, or $B\cup\{\infty, -\infty\}$, where $B \in \mathcal{B}$ is a Borel set. (The student should check that $\bar{\mathcal{B}}$ is a σ-field.) A measurable function $f : (\Omega, \mathcal{F}) \longrightarrow (\bar{I\!R}, \bar{\mathcal{B}})$ is called an *extended Borel function.*

We would now like to define the integration of an extended Borel function w.r.t. a measure μ. The definition proceeds in a stepwise fashion, starting with functions for which the definition of the integral is obvious and proceeding to general extended Borel functions. If $A \in \mathcal{F}$, put

$$\int I_A(\omega)\, d\mu(\omega) \;=\; \mu(A) \quad . \tag{1.17}$$

Now let ϕ be a simple function, say

$$\phi(\omega) \;=\; \sum_{i=1}^{n} a_i I_{A_i}(\omega) \quad .$$

If either $\mu(A_i) < \infty$ for $1 \leq i \leq n$, or all a_i have the same sign (all positive or all negative) then we define

$$\int \phi\, d\mu \;=\; \int \phi(\omega)\, d\mu(\omega) \;=\; \sum_{i=1}^{n} a_i \mu(A_i) \quad . \tag{1.18}$$

We see this is what the definition should be if the integral has the usual linearity property, i.e.

$$\int \left[\sum_{i=1}^{n} a_i I_{A_i}(\omega) \right] d\mu(\omega) \;=\; \sum_{i=1}^{n} a_i \int I_{A_i}(\omega)\, d\mu(\omega) \quad .$$

There is a possible problem with (1.18), namely that different A_i and a_i can give the same simple function. For instance, with $\Omega = I\!R$, note that

$$I_{(0,2)}(x) \;+\; I_{(0,1)}(x) \;=\; 2I_{(0,1)}(x) \;+\; I_{[1,2)}(x) \quad .$$

However, one can show (with a tedious argument) that different representations of the same simple function in (1.18) give the same value to $\int \phi \, d\mu$ (see Royden's text, p. 76). Our restriction that either $\mu(A_i) < \infty$ for $1 \le i \le n$, or all a_i have the same sign comes from the fact that $\infty - \infty$ is undefined, so the r.h.s. of (1.18) would not be defined for instance if $\mu(A_1) = \mu(A_2) = \infty$ but $a_1 < 0 < a_2$.

Now we define the integral of a nonnegative extended Borel function. This is the most difficult step in the stepwise definition of the integral. If $f(\omega) \ge 0$ for all $\omega \in \Omega$, then

$$\begin{gathered} \int f d\mu \;=\; \int f(\omega)\, d\mu(\omega) \;= \\ \sup \Big\{ \int \phi d\mu \;:\; \phi \text{ is a simple function} \qquad (1.19) \\ \text{and } 0 \le \phi(\omega) \le f(\omega), \;\; \text{for all } \omega \in \Omega \,\Big\} \quad . \end{gathered}$$

In words, $\int f d\mu$ is the least upper bound of all integrals of nonnegative simple functions which are below f. Note that the set of such simple functions is nonempty since it contains $I_{\emptyset} = 0$. Note also that $\int f d\mu = \infty$ is possible.

For any $f : (\Omega, \mathcal{F}) \longrightarrow (\bar{I\!R}, \bar{\mathcal{B}})$ define the *positive part* and *negative part*, respectively, by

$$f_{+}(\omega) \;=\; \max\{f(\omega), 0\} \quad , \quad f_{-}(\omega) \;=\; \max\{-f(\omega), 0\} \quad . \qquad (1.20)$$

Note that for all $\omega \in \Omega$,

$$f_{+}(\omega) \;\ge\; 0 \quad , \quad f_{-}(\omega) \;\ge\; 0 \quad , \qquad (1.21)$$

$$f(\omega) \;=\; f_{+}(\omega) \;-\; f_{-}(\omega)\,, \qquad (1.22)$$

$$|f(\omega)| \;=\; f_{+}(\omega) \;+\; f_{-}(\omega) \quad . \qquad (1.23)$$

We say $\int f d\mu$ *exists* or *is defined* iff at least one of $\int f_{+} d\mu$ and $\int f_{-} d\mu$ is finite (these integrals are defined by (1.19)). If $\int f d\mu$ is defined, then it is given by

$$\int f d\mu \;=\; \int f(\omega)\, d\mu(\omega) \;=\; \int f_{+} d\mu \;-\; \int f_{-} d\mu \quad . \qquad (1.24)$$

Note that our requirement that at least one of the latter two integrals be finite avoids the undefined form $\infty - \infty$. We say f is *integrable* iff both $\int f_{+} d\mu$ and $\int f_{-} d\mu$ are finite.

Finally, we define the integral of f over the set $A \in \mathcal{F}$ as

$$\int_{A} f d\mu \;=\; \int I_{A} f d\mu\,, \qquad (1.25)$$

provided the latter integral is defined. If μ is a Borel measure (i.e. defined on the measurable space $(I\!\!R, \mathcal{B})$), and if $A = [a, b]$ is a closed interval, then we often will use the "limits of integration" notation as in

$$\int_{[a,b]} f d\mu \;=\; \int_a^b f d\mu \quad .$$

We also note that it is common to write $d\mu(\omega)$ as $\mu(d\omega)$, as in

$$\int f(\omega)\, d\mu(\omega) \;=\; \int f(\omega)\, \mu(d\omega) \quad . \tag{1.26}$$

To explain this notation, if $\sum a_i I_{A_i}$ is a simple function approximation to $f(x)$ so that $\int \sum a_i I_{A_i}\, d\mu = \sum a_i \mu(A_i) \doteq \int f\, d\mu$, then the values a_i will be approximately $f(\omega_i)$ for some $\omega_i \in A_i$ and the sets A_i will have small measure. If we write $d\omega_i$ to represent the "differential" set A_i, then we obtain notationally $\sum f(\omega_i)\mu(d\omega_i) \doteq \int f\, d\mu$. The notation $\mu(d\omega)$ is meant to remind us of the measure of these differential sets, which are multiplied by $f(\omega)$ and summed. We will sometimes use this notation when it helps to aid understanding.

In probability, $\int X(\omega)\, dP(\omega)$ is usually called the *expected value* or *expectation* of the r.v. X, and commonly denoted $E[X]$. If P is a Borel p.m. with c.d.f. F, then it is common to write $dF(x)$ in place of $dP(x)$.

If $f : (I\!\!R, \mathcal{B}) \longrightarrow (\bar{I\!\!R}, \bar{\mathcal{B}})$ then integrals w.r.t. Lebesgue measure (referred to as *Lebesgue integrals*) are generally written $\int f(x)\, dx$ rather than $\int f(x)\, dm(x)$ since they are the same as the usual Riemann integral when it is defined, as will be seen shortly. Thus, all of the integration theory that the reader has learned heretofore may be applied to the calculation of Lebesgue integrals. Note that the Lebesgue integral $\int f(x) dm(x)$ is in general an improper Riemann integral since it is over the entire real line $I\!\!R = (-\infty, \infty)$. Of course, an integral over a finite interval $\int_a^b f(x) dx = \int I_{[a,b]}(x) f(x) dm(x)$ will still be an improper Riemann integral if $f(x)$ tends to ±∞ at some point in $[a, b]$, and one must use an appropriate method for calculating its value as a Riemann or Lebesgue integral.

To introduce another class of examples of integrals that frequently arise in practice, let $\Omega = \{a_1,\, a_2,\, ...\}$ be a *discrete set*, by which we mean one that can be listed as a sequence, finite or infinite. Take $\mathcal{F} = \mathcal{P}(\Omega)$ as the σ-field, and $\mu = \#$, counting measure. Then one can easily see that that any $f : \Omega \longrightarrow I\!\!R$ is measurable, and it can be shown (see Exercise 1.2.26) that

$$\int f d\# \;=\; \sum_i f(a_i) \quad , \tag{1.27}$$

whenever the l.h.s. is defined. Thus, we see that measure theory includes the classical theory of summmation and Riemann integration.

For another example of the integral, let δ_x denote a unit point mass measure on a measurable space $(\Omega, \mathcal{F})$. If $\phi = \sum_i c_i I_{A_i}$ is a simple function, then $\int \phi\, d\delta_x = \sum_i c_i \delta_x(A_i) = \sum_i c_i I_{A_i}(x) = \phi(x)$. Note that in general

$$\delta_x(A) \;=\; I_A(x).$$

If $f : \Omega \longrightarrow I\!R$ is measurable and $f \geq 0$, then for simple functions ϕ such that $0 \leq \phi \leq f$, we have $\int \phi \, d\delta_x = \phi(x) \leq f(x)$, and taking $\phi = f(x) I_{\{x\}}$, we get $\int \phi \, d\delta_x = \phi(x) = f(x)$, so $\int f \, d\delta_x = f(x)$. Finally, if $f : \Omega \longrightarrow I\!R$ is any measurable function, then $\int f \, d\delta_x = \int f_+ \, d\delta_x - \int f_- \, d\delta_x = f_+(x) - f_-(x) = f(x)$. We have established the formula

$$\int f \, d\delta_x \;=\; f(x). \tag{1.28}$$

For a linear combination of unit point mass measures, by Exercise 1.2.35 we have the following formula:

$$\text{If } \mu \;=\; \sum_i a_i \delta_{x_i}, \text{ then } \quad \int f \, d\mu \;=\; \sum_i a_i f(x_i). \tag{1.29}$$

We have already seen that the empirical distribution defined in (1.4) is very useful for motivating data analytic tools (such as sample quantiles) using notions from probability theory. Let $(x_1, x_2, \ldots, x_n)$ be a data set. In the context of integration, if g is a real valued function defined on the space Ω of possible observations, then by (1.29),

$$\int_\Omega g(x) \, d\hat{P}_n(x) \;=\; \frac{1}{n} \sum_{i=1}^n g(x_i). \tag{1.30}$$

We may interpret this as saying that the integral of g w.r.t. the empirical distribution is the sample average of $g(x_i)$. If P is the true probability model for the experiment (in the long run relative frequency sense of (1.5)), then a simple method for estimating an expectation $E[g(X)] = \int g(x) \, dP(x)$ is to use the sample average $\int g(x) \, d\hat{P}_n(x)$. Again, our association of a data set with the corresponding empirical distribution provides a connection of sorts between sample averages (which are widely used in data analysis) and concepts from measure theory and probability theory. Of course, we are interested if in some sense $\int g(x) \, d\hat{P}_n(x)$ converges to $E[g(X)]$ as $n \to \infty$. In general, under a suitable mathematical model, sample averages do tend to the theoretical expectations as sample size increases. Making this result rigorous has been a big concern of probabilists for centuries, and generally goes under the names of "law of large numbers" or "ergodic theorem." Refining the result leads to the central limit theorem, the law of the iterated logarithm, and numerous other results.

1.2.4 Riemann Integrals.

Our development here will closely follow that of Rudin, Chapter 6. We will begin by considering functions only on a finite interval $[a, b)$. The Riemann integral is defined as follows. A *step function* on $[a, b)$ is a function of the form

$$\psi(x) \;=\; \sum_{i=1}^n c_i I_{[a_{i-1}, a_i)}(x) \tag{1.31}$$

where $a = a_0 < a_1 < ... < a_{n-1} < a_n = b$. Note that step function is a simple function $\sum_i c_i I_{A_i}$ where the measurable sets A_i are required to be intervals. Now the intervals $[a_0, a_1)$, $[a_1, a_2)$, …, $[a_{n-1}, a_n)$ form a partition Π of $[a, b)$ consisting of finitely many intervals. The *mesh of* Π is $\max\{a_i - a_{i-1} : 1 \le i \le n\}$, i.e. the length of the longest interval in Π. Given such a partition and a bounded function $f : [a, b) \longrightarrow I\!R$, define

$$\begin{aligned} \mathcal{U}(f, \Pi) &= \sum_{i=1}^{n} \left(\sup_{[a_{i-1},a_i)} f \right) (a_i - a_{i-1}) \\ &= \int \bar{\psi}_{f,\Pi}(x)\, dx \end{aligned}$$

where $\bar{\psi}_{f,\Pi}$ is the step function

$$\bar{\psi}_{f,\Pi}(x) = \sum_{i=1}^{n} \left(\sup_{[a_{i-1},a_i)} f \right) I_{[a_{i-1},a_i)}(x).$$

The *upper Riemann integral* is defined by

$$\overline{\int}_a^b f(x)dx = \inf_{\Pi} \mathcal{U}(f, \Pi)$$

where the infimum is over partitions Π of $[a, b)$ into finitely many intervals. A similar definition for the lower Riemann integral holds, viz.

$$\begin{aligned} \mathcal{L}(f, \Pi) &= \sum_{i=1}^{n} \left(\inf_{[a_{i-1},a_i)} f \right) (a_i - a_{i-1}) = \int \underline{\psi}_{f,\Pi}(x)\, dx \\ \underline{\psi}_{f,\Pi}(x)) &= \sum_{i=1}^{n} \left(\inf_{[a_{i-1},a_i)} f \right) I_{[a_{i-1},a_i)}(x) \\ \underline{\int}_a^b f(x)dx &= \sup_{\Pi} \mathcal{L}(f, \Pi). \end{aligned}$$

The *Riemann integral exists* provided the upper and lower integrals are equal, and their common value is denoted $\mathcal{R} \int_a^b f(x)dx$ and called the *Riemann integral.*

Let f be a nonnegative bounded measurable function on the interval $[a, b)$, and then the Lebesgue integral $\int_{[a,b)} f(x)dm(x)$ exists. We will show that the value of the Lebesgue integral is between the lower and upper Riemann integrals, so if the Riemann integral exists, it must equal the Lebesgue integral. (Note: one can show that existence of the Riemann integral implies existence of the Lebesgue integral; see Exercise 1.2.29.) We may re-express the lower Riemann integral as

$$\underline{\int}_a^b f(x)dx = \sup_{\underline{\psi}_{f,\Pi}} \int \underline{\psi}_{f,\Pi}(x)\, dx$$

where the supremum is over step functions $\underline{\psi}_{f,\Pi} = \sum_i \left(\inf_{[a_{i-1},a_i)} f\right) I_{[a_{i-1},a_i)}(x)$ corresponding to some partition Π into finitely many intervals. Now notice that each of the step functions used in defining the lower Riemann integral are simple functions satisfying $0 \le \psi \le f$, on $[a,b)$ (set all functions to 0 outside of $[a,b)$ for purposes of this discussion). Hence, the class of step functions over which the suprememum is taken is a subclass of the simple functions over which one takes supremum to get the Lebesgue integral as in (1.19). Since the supremum of a subset is smaller than a superset,

$$\underline{\int_a^b} f(x)\,dx \;\le\; \int_{[a,b)} f(x)\,dm(x).$$

Now we show that the Lebesgue integral is bounded above by the upper Riemann integral. Each of the step functions $\bar{\psi}_{f,\Pi}$ that goes into the definition of the upper Riemann integral satisfies $f \le \bar{\psi}_{f,\Pi}$, so $\int_{[a,b)} f(x)dm(x) \le \int_{[a,b)} \bar{\psi}_{f,\Pi}dm(x)$ (see Proposition 1.2.5(c) below). Taking infimum over all such step functions gives $\int_{[a,b)} f(x)dm(x) \le \bar{\int}_a^b f(x)dx$, as claimed.

Not all Lebesgue integrable functions are Riemann integrable. For instance, let $f(x) = I_A(x)$ be the indicator of $A = \{x \in [0,1) : x \text{ is rational }\}$, then the Lebesgue integral is $m(A) = 0$. Now any interval $[a_{i-1}, a_i)$ with $a_{i-1} < a_i$ contains both rational and irrational numbers, so $\sup_{[a_{i-1},a_i)} I_A = 1$ and $\inf_{[a_{i-1},a_i)} I_A = 0$. Thus, $\bar{\psi}_{f,\Pi} \equiv 1$ on $[0,1)$ and $\underline{\psi}_{f,\Pi} \equiv 0$ on $[0,1)$ for all allowable partitions Π of $[0,1)$. Thus, $\bar{\int}_0^1 I_A(x)dx = 1$ and $\underline{\int}_0^1 I_A(x)dx = 0$, and hence the Riemann integral does not exist.

Thus we see that the Lebesgue integral is more general than the Riemann integral. The reason for the generality is not hard to see – the Riemann integral is derived as the limit of integrals of step functions, whereas the Lebesgue integral is derived as the limit of integrals of simple functions, and the class of simple functions is larger than the class of step functions. Now our discussion of the Riemann integral above was based on one of several definitions that appear in the literature. In Exercise 1.2.30, the reader is asked to consider another definition of the Riemann integral that is often used and to show it is equivalent to the definition above.

We close this subsection with a brief discussion of the situation of improper integrals. In the Riemann theory of integration, and integral is improper if it is over an inifinite interval or if the function is not bounded, and improper integrals are defined as limits of proper integrals. For instance,

$$\mathcal{R}\int_0^\infty f(x)\,dx \;=\; \lim_{b\to\infty} \mathcal{R}\int_0^b f(x)\,dx.$$

For the Lebesgue theory, there is no distinction between “proper” and “improper” integrals. An improper Riemann integral may exist, but its Lebesgue integral may fail to exist. For example consider

$$f(x) \;=\; \begin{cases} 1/n & \text{if } 2n-1 \le x < 2n, \\ \\ -1/n & \text{if } 2n \le x < 2n+1, \\ \\ 0 & \text{if } x < 1. \end{cases}$$

where n ranges over integers ≥ 1. Then for $b > 1$

$$\int_0^b f(x)\,dx \;=\; \begin{cases} [b-(2n-1)]/n & \text{if } 2n-1 \le b < 2n, \\ \\ [1-(b-2n)]/n & \text{if } 2n \le b < 2n+1. \end{cases}$$

Note that $\left|\int_0^b f(x)\,dx\right| \le 1/n \to 0$ as $b \to \infty$, so the improper Riemann integral $\mathcal{R}\int_0^\infty f(x)\,dx$ is 0. However, the Lebesgue integral does not exist since

$$\int f_+(x)\,dm(x) \;=\; \int f_-(x)\,dm(x) \;=\; \sum_{n=1}^{\infty} \frac{1}{n} \;=\; \infty.$$

Basically, the Lebesgue integral exists whenever the improper Riemann integral is absolutely convergent.

1.2.5 Properties of the Integral.

Proposition 1.2.5 (Basic properties of the integral.) *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and f, g extended Borel functions on Ω.*

(a) If $\int f d\mu$ exists and $a \in I\!R$, then $\int a f d\mu$ exists and equals $a \int f d\mu$.

(b) If $\int f d\mu$ and $\int g d\mu$ both exist and $\int f d\mu + \int g d\mu$ is defined (i.e. not of the form $\infty - \infty$), then $\int (f+g) d\mu$ is defined and equals $\int f d\mu + \int g d\mu$.

(c) If $f(\omega) \le g(\omega)$ for all $\omega \in \Omega$, then $\int f d\mu \le \int g d\mu$, provided the integrals exist.

(d) $|\int f d\mu| \le \int |f| d\mu$, provided $\int f d\mu$ exists.

□

The last result appears in Billingsley, Theorem 16.1, p. 209. Note in the above that our conventions regarding arithmetic with ±∞ will often be necessary.

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, and let $S(\omega)$ be a logical statement about a general element $\omega \in \Omega$. For instance, if f and g are extended Borel functions on Ω then "$f(\omega) \le g(\omega)$" is an example of such a logical statement, which may be true for some ω and false for other ω. We say S holds *μ–almost everywhere*, abbreviated *μ-a.e.* (or simply *a.e.* if the measure μ is clear from context), iff there is a set $N \in \mathcal{F}$ such that $\mu(N) = 0$ and $S(\omega)$ is true for $\omega \notin N$. Such a set N of μ-measure 0 is sometimes called a *μ–null set* or simply a *null set* if μ is clear. In probability, *P–almost surely* or *P–a.s.* means the same as P-a.e., with similar remarks about omitting the measure. With this terminology, we can give an extension of Proposition 1.2.5 (c) as in part (a) below.

Proposition 1.2.6 *(a) If f and g are extended Borel functions on $(\Omega, \mathcal{F}, \mu)$ and $f \leq g$ μ–a.e., then $\int f d\mu \leq \int g d\mu$, provided the integrals exist.*

(b) If $f \geq 0$ μ–a.e. and $\int f d\mu = 0$ then $f = 0$ μ–a.e.

□

The proof of part (b) is Exercise 1.2.32.

Perhaps the main use of the a.e. concept is with sequences of functions. A sequence f_1, f_2, ... often converges to a function f except on a null set. For example

$$f_n(x) \;=\; nI_{[0,1/n]}(x) \quad . \tag{1.32}$$

Then

$$\lim_{n\to\infty} f_n(x) \;=\; \begin{cases} \infty & \text{if } x = 0, \\ 0 & \text{if } x \neq 0. \end{cases}$$

Since $m(\{0\}) = 0$, we may say $f_n \to 0$ m-a.e., where of course m is Lebesgue measure. Notice that

$$\lim_{n\to\infty} \int f_n dm \;=\; 1 \;, \quad \int \lim_{n\to\infty} f_n dm \;=\; 0 \quad . \tag{1.33}$$

This example indicates that the interchange of $\lim_{n\to\infty}$ and $\int$ is not always permissable. One of the nice features of measure theory is that there are fairly simple conditions which can be used to justify this interchange. The following results appear in Billingsley, Theorem 15.1, p. 206, and Theorem 16.4, p. 213.

Theorem 1.2.7 *All functions here are extended Borel functions on $(\Omega, \mathcal{F}, \mu)$.*

(a) (Monotone convergence theorem.) Suppose f_n is an increasing sequence of nonnegative functions (i.e. $0 \leq f_1 \leq f_2 \leq ... \leq f_n \leq f_{n+1} \leq ...$) and $f(\omega) = \lim f_n(\omega)$. Then $\lim \int f_n d\mu = \int f d\mu$.

(b) (Lebesgue's dominated convergence theorem.) Suppose $f_n \to f$ a.e. as $n \to \infty$ and that there is an integrable function g such that for all n, $|f_n| \leq g$ a.e. Then $\lim \int f_n d\mu = \int f d\mu$.

A convenient notation for a sequence f_n satisfying the hypotheses of the monotone convergence theorem is $0 \leq f_n \uparrow f$. Note that for each ω the limit $f(\omega) = \lim_{n\to\infty} f_n(\omega)$ exists, but may be $+\infty$ (why?). In the dominated convergence theorem, g is called the *dominating function.* There is no unique dominating function, and one can be very flexible in choosing a convenient one. One of the conclusions of the dominated convergence theorem is that $\lim \int f_n d\mu$ exists. The next result in combination with the convergence theorems is a very powerful tool.

Proposition 1.2.8 *Let $f : (\Omega, \mathcal{F}) \longrightarrow (\bar{I\!R}, \bar{\mathcal{B}})$. Then there exists a sequence of simple functions ϕ_n on Ω such that $\phi_n \to f$ and $|\phi_n| \leq |f|$ for all n. If $f \geq 0$ then we may take $\phi_n \geq 0$ for all n.*

Further, if μ is a measure on $(\Omega, \mathcal{F})$ and $\int f d\mu$ is defined, then $\int \phi_n d\mu \to \int f d\mu$.

Proof. Suppose $f \geq 0$ for now and put

$$\phi_n(\omega) \; = \; \sum_{k=0}^{2^{2^n}-1} k2^{-n} I_{[k2^{-n},(k+1)2^{-n})}(f(\omega)) \quad + \; 2^n I_{[2^n,\infty]}(f(\omega)) \quad . \tag{1.34}$$

It is left an exercise to check the following:

(i) $0 \leq \phi_n(\omega) \leq f(\omega)$ for all ω and n.

(ii) If $f(\omega) < \infty$ then for all n such that $2^n > f(\omega)$, $|f(\omega) - \phi_n(\omega)| \leq 2^{-n}$.

(iii) If $f(\omega) = \infty$ then for all n, $\phi_n(\omega) = 2^n$.

(iv) $0 \leq \phi_1 \leq \phi_2 \leq ...$

These claims are easy to see from a picture with a small value of n. In either of case (ii) or (iii) we clearly have $\phi_n(\omega) \to f(\omega)$, and by the MCT, $\int \phi_n d\mu \to \int f d\mu$. This proves the the proposition for $f \geq 0$.

The claim for general f is easily established from $f = f_+ - f_-$, and nonnegativity of f_+ and f_-.

□

See Billingsley, Theorem 13.5, p. 185 for the above.

1.2.6 Applications.

As an example of an application of the above results, we have the following important theorems which are very useful in probability and statistics. One of the most important results is another method for constructing new measures out of old ones.

Theorem 1.2.9 (Measures Defined by Densities.) *Let $f : (\Omega, \mathcal{F}, \mu) \longrightarrow (\bar{I\!R}, \bar{\mathcal{B}})$ be nonnegative, and put*

$$\nu(A) \; = \; \int_A f d\mu \quad , \quad A \in \mathcal{F} \quad . \tag{1.35}$$

Show that ν is a measure on $(\Omega, \mathcal{F})$.

□

The proof is left as Exercise 1.2.38. In the context of this theorem, the function f is called *the density of ν with respect to (w.r.t.) μ.* Most of the probability measures we use in practice will be constructed through densities, either w.r.t. Lebesgue measure (so-called continuous distributions) or w.r.t. counting measure (discrete distributions). We will later provide necessary and sufficient conditions for when one measure has a density w.r.t another measure (the Radon-Nikodym theorem). This result has many ramifications in probability and statistics.

The next result is very subtle, and illustrates the power of abstract measure theory. As with the previous theorem, there will be many important consequences.

Theorem 1.2.10 (Change of variables.) *Suppose $f : (\Omega, \mathcal{F}, \mu) \longrightarrow (\Lambda, \mathcal{G})$ and $g : (\Lambda, \mathcal{G}) \longrightarrow (\bar{I\!R}, \bar{\mathcal{B}})$. Then*

$$\int_\Omega (g \circ f)(\omega)\, d\mu(\omega) \;=\; \int_\Lambda g(\lambda)\, d(\mu \circ f^{-1})(\lambda) \quad , \tag{1.36}$$

where this has the following interpretation: if either integral is defined, then so is the other and the two are equal.

Proof. First assume g is a nonnegative simple function, say

$$g(\lambda) \;=\; \sum_{i=1}^n a_i I_{A_i}(\lambda)$$

where $a_i \geq 0$ for all i. Then $g \circ f \geq 0$ so both integrals exist. Now

$$\int_\Omega (g \circ f) d\mu \;=\; \int \sum a_i I_{A_i}(f(\omega)) d\mu(\omega)$$

$$= \sum a_i \int I_{A_i}(f(\omega)) d\mu(\omega)$$

by Proposition 1.2.5 (b). Note that $I_A(f(\omega)) = 1$ iff $f(\omega) \in A$ iff $\omega \in f^{-1}(A)$ iff $I_{f^{-1}(A)}(\omega) = 1$, so $I_A \circ f = I_{f^{-1}(A)}$. Using this in the last display gives

$$\begin{aligned}
\int_\Omega (g \circ f) d\mu \;&=\; \sum a_i \int I_{f^{-1}(A_i)}(\omega) d\mu(\omega) \\
&=\; \sum a_i \mu(f^{-1}(A_i)) \\
&=\; \sum a_i (\mu \circ f^{-1})(A_i) \\
&=\; \sum a_i \int I_{A_i} d(\mu \circ f^{-1}) \\
&=\; \int g\, d(\mu \circ f^{-1}) \quad .
\end{aligned}$$

This completes the proof for nonnegative simple functions g.

Now suppose that $g \geq 0$. Then both integrals are still defined. Let ϕ_n be simple functions with $0 \leq \phi_n \uparrow g$ by Proposition 1.2.8. Then $\int \phi_n d(\mu \circ f^{-1}) \rightarrow$

$\int g\, d(\mu \circ f^{-1})$ by the monotone convergence theorem (Theorem 1.2.7 (a)). The argument above that $I_A \circ f = I_{f^{-1}(A)}$ shows that $\phi_n \circ f$ are nonnegative simple functions on Ω, and it is easy to see that $0 \leq \phi_n \circ f \uparrow g \circ f$, so $\int (\phi_n \circ f) d\mu \rightarrow \int (g \circ f) d\mu$ by the monotone convergence theorem. Since $\int (\phi_n \circ f) d\mu = \int \phi_n d(\mu \circ f^{-1})$ by the first part of the proof, we have $\int (g \circ f) d\mu = \int g\, d(\mu \circ f^{-1})$.

Now let g be a general extended Borel function on Λ and consider the positive and negative parts, g_+ and g_-, respectively. Note that $(g \circ f)_+ = g_+ \circ f$, and $(g \circ f)_- = g_- \circ f$, so by the preceding part of the proof for nonnegative functions,

$$\int (g \circ f)_+ d\mu \;=\; \int g_+ d(\mu \circ f^{-1}) \quad , \quad \int (g \circ f)_- d\mu \;=\; \int g_- d(\mu \circ f^{-1}) \quad .$$

Hence, if say $\int (g \circ f)_- d\mu < \infty$, so that the l.h.s. of (1.36) is defined, then $\int g_- d(\mu \circ f^{-1}) < \infty$ and the r.h.s. is defined, and

$$\begin{aligned} \int g \circ f d\mu &= \int (g \circ f)_+ d\mu - \int (g \circ f)_- d\mu \\ &= \int g_+ d(\mu \circ f^{-1}) - \int g_- d(\mu \circ f^{-1}) \\ &= \int g\, d(\mu \circ f^{-1}) \quad . \end{aligned}$$

A similar argument applies if $\int (g \circ f)_+ d\mu < \infty$, which is the other way the l.h.s. of (1.36) can exist. The r.h.s. of (1.36) exists just in case one of $\int g_+ d(\mu \circ f^{-1}) < \infty$ or $\int g_- d(\mu \circ f^{-1}) < \infty$, and the proof goes through without difficulty again.

□

The above result appears in Billingsley, Lemma 16.12, page 219. The technique of proof of the last theorem is important: start with simple functions, use Proposition 1.2.8 and Theorem 1.2.7 to extend to nonnegative functions, and finally to general functions using the decomposition into positive and negative parts.

We briefly indicate the importance of Theorem 1.2.10. Let $(\Omega, \mathcal{F}, P)$ be a probability space and X a r.v. defined thereon. If $E[X] = \int X dP$ exists, then one usually computes $E[X]$ by $\int_{I\!R} x\, dP_X(x)$ where $P_X = P \circ X^{-1} = \text{Law}[X]$ is the distribution of X. Thus, we compute an integral over the real line rather than an integral over the original probability space. If $g : I\!R \longrightarrow I\!R$, then $E[g(X)]$ is typically computed as $\int_{I\!R} g(x) dP_X(x)$ rather than $\int_{I\!R} y\, dP_{g(X)}(y)$, i.e. one integrates w.r.t. the distribution of the original r.v. X rather than w.r.t. the distribution of $g(X)$. Theorem 1.2.10 is used so often by statisticians without giving it any thought that it is sometimes referred to as “the law of the unconscious statistician.” It should be noted that calculation of $\mu \circ f^{-1}$ may be complicated, e.g. involving Jacobians, a subject treated in Chapter 2, Section 2.4

The next result is used frequently in statistics. It also appears in Billingsley, Theorem 16.8, p. 213.

Theorem 1.2.11 (Interchange of differentiation and integration.) *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and suppose $g(\omega, \theta)$ is a real valued function on the cartesian product space $\Omega \times (a,b)$ where (a,b) is a finite open interval in $I\!R$. Assume g satisfies the following:*

(i) *For each fixed $\theta \in (a,b)$, the function $f_\theta(\omega) = g(\omega,\theta)$ is a Borel function of ω and*

$$\int |g(\omega,\theta)|\, d\omega \;<\; \infty.$$

(ii) *There is a null set N such that for all $\omega \notin N$, the derivative $\partial g(\omega,\theta)/\partial\theta$ exists for all $\theta \in (a,b)$.*

(iii) *There is an integrable function $G : \Omega \longrightarrow \bar{I\!R}$ such that for all $\omega \notin N$ and all $\theta \in (a,b)$,*

$$\left| \frac{\partial g}{\partial \theta}(\omega,\theta) \right| \;\le\; G(\omega) \quad .$$

Then for each fixed $\theta \in (a,b)$, $\partial g(\omega,\theta)/\partial\theta$ is integrable w.r.t. μ and

$$\frac{d}{d\theta}\int_\Omega g(\omega,\theta)\, d\mu(\omega) \;=\; \int_\Omega \frac{\partial g}{\partial\theta}(\omega,\theta)\, d\mu(\omega) \quad .$$

Proof. For convenience, let $H(\theta) = \int g(\omega,\theta)\, d\mu(\omega)$. Suppose $\omega \notin N$, then by the mean value theorem (Theorem 5.10, p. 108 of Rudin), if $\theta \in (a,b)$ and $\theta + \delta \in (a,b)$, then

$$\frac{g(\omega,\theta+\delta) - g(\omega,\theta)}{\delta} \;=\; \frac{\partial g}{\partial\theta}(\omega, \theta + \alpha\delta)$$

for some $\alpha \in [0,1]$, and in particular for all $\omega \notin N$,

$$\left| \frac{g(\omega,\theta+\delta) - g(\omega,\theta)}{\delta} \right| \;\le\; G(\omega) \quad . \tag{1.37}$$

Now let η_n be any sequence in $I\!R$ converging to 0. Then by Proposition 1.2.6,

$$\frac{H(\theta+\eta_n) - H(\theta)}{\eta_n} \;=\; \int \frac{g(\omega,\theta+\eta_n) - g(\omega,\theta)}{\eta_n}\, d\mu(\omega) \quad .$$

Thus, we have for each fixed $\theta \in (a,b)$, the sequence of functions

$$f_n(\omega) \;=\; \frac{g(\omega,\theta+\eta_n) - g(\omega,\theta)}{\eta_n}$$

converges μ-a.e. to $\partial g(\omega,\theta)/\partial\theta$, and by (1.37), $|f_n| \le G$, μ-a.e., and G is μ-integrable by assumption. Hence, by the dominated convergence theorem (Theorem 1.2.7 (b)), $\int f_n d\mu \to \int [\partial g(\omega,\theta)/\partial\theta] d\mu$, i.e.

$$\lim_{n\to\infty} \frac{H(\theta+\eta_n) - H(\theta)}{\eta_n} \;=\; \int_\Omega \frac{\partial g}{\partial\theta}(\omega,\theta)\, d\mu(\omega) \quad .$$

Since the sequence $\eta_n \to 0$ was arbitrary, it follows that

$$\lim_{\delta \to 0} \frac{H(\theta+\delta) - H(\theta)}{\delta} = \int_\Omega \frac{\partial g}{\partial \theta}(\omega, \theta)\, d\mu(\omega) \quad .$$

(See e.g. Theorem 4.2, p. 84 of Rudin for this result which extends a limit from an arbitrary sequence to a continuous limit.) This last display states that $H(\theta)$ is differentiable and the derivative is the r.h.s. This proves the Theorem.

□

1.2.7 Final Notes.

(a) Most of the results of this section are standard measure theory, except the last two theorems. Theorem 1.2.10 may be found in Billingsley and Theorem 1.2.11 may be found in Ash. (b) The use of $dF(x)$ is place of $dP(x)$ when P is a Borel probability measure is more than an abuse of notation since it is derived from the theory of Stieltjes integrals (see Ash or Rudin). However, we shall not need that theory here.

Problems for Section 1.2.

1.2.1 Verify (1.11).

1.2.2 Show that the analogue of relation (1.11) for forward images does hold. Give a counterexample to the analogue of (1.12) for forward images to show it is not valid.

1.2.3 Verify the analogue of (1.12) for intersections by (a) using DeMorgan's laws, and (b) arguing directly as was done to show (1.12).

1.2.4 Let $f : I\!\!R \longrightarrow I\!\!R$ be given by $f(x) = \min\{|x|, 1\}$. Show that the inverse image of an arbitrary Borel set $B \subset I\!\!R$ is given by

$$f^{-1}(B) \ = \ \begin{cases} \{x : |x| \in B \cap [0,1)\} & \text{if } 1 \notin B, \\ \{x : |x| \in B \cap [0,1)\} \cup \{1\} & \text{if } 1 \in B. \end{cases}$$

1.2.5 Suppose A and B are disjoint, nonempty, proper subsets of Ω and let $\phi(\omega)$ $= -3I_A(\omega) + 3I_B(\omega)$. Find $\phi^{-1}(C)$ for $C \subset I\!\!R$. List all elements of $\sigma(\phi)$. Show $\sigma(\phi) = \sigma(\{A, B\})$.

1.2.6 Prove that $(g \circ f)^{-1}(A) = f^{-1}(g^{-1}(A))$ where $f : \Omega \longrightarrow \Lambda$, $g : \Lambda \longrightarrow \Xi$, and $A \subset \Xi$.

1.2.7 Let $\Omega = I\!\!R$ and $\mathcal{F} = \mathcal{B}$. Show that each of the following functions on Ω is a Borel function.

(i) $f(\omega) = |\sin[\exp(3\cos\omega)]|$.

(ii) $f(\omega) = 2I_{[0,\infty)}(\omega) - 1$.

(iii)

$$f(\omega) \ = \ \begin{cases} |\omega| & \text{if } \omega \text{ is rational,} \\ 0 & \text{otherwise.} \end{cases}$$

1.2.8 Let $A \subset \Omega$ be a measurable set in the measure space $(\Omega, \mathcal{F}, \mu)$, then show that the induced measure $\nu = \mu \circ I_A^{-1}$ is given by

$$\nu(B) \ = \ \begin{cases} 0 & \text{if neither } 1 \in B \text{ nor } 0 \in B, \\ \mu(A) & \text{if } 1 \in B \text{ but } 0 \notin B, \\ \mu(A^c) & \text{if } 0 \in B \text{ but } 1 \notin B, \\ \mu(\Omega) & \text{if both } 1 \in B \text{ and } 0 \in B. \end{cases}$$

Here, B is an arbitrary Borel set. Show that ν is the same as

$$\mu(A)\delta_1 \ + \ \mu(A^c)\delta_0 \quad ,$$

where δ_x is a unit point mass measure at x.

1.2.9 Verify that $\bar{\mathcal{B}}$ is a σ-field.

1.2.10 Show that Proposition 1.2.4 holds if we replace F^- by F^+.

1.2.11 Verify equations (1.21), (1.22), and (1.23).

1.2.12 Assuming $f : (\Omega, \mathcal{F}) \longrightarrow (\bar{I\!R}, \bar{\mathcal{B}})$, show that the following are equivalent:

(i) f is integrable w.r.t. μ;

(ii) $\int f d\mu$ is defined and $-\infty < \int f d\mu < \infty$;

(iii) $\int |f| d\mu \ < \ \infty$.

1.2.13 Let Ω be equipped with the power set $\mathcal{P}(\Omega)$ for the σ-field. Show that if $(\Lambda, \mathcal{G})$ is any other measurable space, and $f : \Omega \longrightarrow \Lambda$ is *any* function, then f is measurable.

1.2.14 Show Proposition 1.2.5 (d) follows from Proposition 1.2.5 (c).

1.2.15 Verify both relations in (1.33). Check that the sequence f_n there does not satisfy the hypothesis for either the dominated convergence theorem or the monotone convergence theorem.

1.2.16 Let N_1, N_2, ... be a sequence of null sets. Show that $\bigcup_i N_i$ is a null set.

1.2.17 Let $\mathcal{N}$ be an arbitrary collection of null sets. Show by counterexample that $\bigcup_{N \in \mathcal{N}} N$ is not necessarily a null set.

1.2.18 Suppose $(\Omega, \mathcal{F})$ and $(\Lambda, \mathcal{G})$ are measurable spaces and that $\mathcal{G} = \sigma(\mathcal{A})$ for some $\mathcal{A} \subset \mathcal{P}(\Lambda)$. Suppose $f : \Omega \longrightarrow \Lambda$ and $f^{-1}(\mathcal{A}) \subset \mathcal{F}$. Show that $f : (\Omega, \mathcal{F}) \longrightarrow (\Lambda, \mathcal{G})$, i.e., that f is measurable.

1.2.19 Let $f : \Omega \longrightarrow \Lambda$ and let $\mathcal{G}$ be a σ-field on Λ. Show that $f^{-1}(\mathcal{G})$ is a σ-field on Ω.

1.2.20 Suppose $f : \Omega \longrightarrow \Lambda$ is an arbitrary function and $\mathcal{F}$ is a σ-field on Ω. Let

$$\mathcal{C} \ = \ \{ C \ : \ C \subset \Omega \text{ and } f^{-1}(C) \in \mathcal{F} \} \quad .$$

Show that $\mathcal{C}$ is a σ-field on Λ. Is f measurable from $(\Omega, \mathcal{F}) \longrightarrow (\Lambda, \mathcal{C})$?

1.2.21 Suppose $\phi = \sum_{i=1}^n a_i I_{A_i}$ is a simple function on $(\Omega, \mathcal{F})$. Show that there is a unique representation $\phi = \sum_{i=1}^m b_i I_{B_i}$ where b_1, b_2, ... , b_m are distinct and nonzero, and B_1, B_2, ... , B_m are disjoint and nonempty. (Hint: $\{b_1, b_2, \ldots, b_m\} = \phi(\Omega) \setminus \{0\}$ and $B_i = \phi^{-1}(b_i)$.)

1.2.22 Let ϕ be as given in Exercise 1.2.21. Find $\sigma(\phi)$. You may wish to consider Exercise 1.2.5 first.

1.2.23 Let ϕ be as given in Exercises 1.2.21 and 1.2.22. Let μ be a measure on $(\Omega, \mathcal{F})$. Find $\mu \circ \phi^{-1}$.

1.2.24 For each of the following functions $f : I\!R \longrightarrow I\!R$, determine the induced measure $m \circ f^{-1}$.

(a) $f(x) = x + a$, some $a \in I\!R$.
(b) $f(x) = ax$, $a \neq 0$.
(c) Same as (b), but $a = 0$.

1.2.25 Suppose $\phi = \sum_{i=1}^{n} a_i I_{A_i} = \sum_{i=1}^{p} c_i I_{C_i}$ are two different representations of the same simple function. Show that the integrals of the two representations as given by (1.18) are equal. (Hint: use Exercise 1.2.21.)

1.2.26 Let $\Omega = \{a_1, a_2, ...\}$ be a discrete set (i.e. a set which can be listed as a finite or infinite sequence). Equip Ω with the $\mathcal{F} = \mathcal{P}(\Omega)$ σ-field, and let $\#$ be counting measure on Ω. Prove equation (1.27). (Hint: proceed stepwise through the definition of the integral, i.e. verify (1.27) first for indicator functions, then simple functions, nonnegative functions, and finally general functions for which the l.h.s. is defined.)

1.2.27 Let $(\Omega, \mathcal{F})$ be as in Exercise 1.2.26. Let μ be any measure on $(\Omega, \mathcal{F})$. Show that

$$\int f d\mu = \sum_i f(a_i)\mu(\{a_i\}) \quad , \tag{1.38}$$

provided the l.h.s. is defined.

1.2.28 Let $(\Omega, \mathcal{F})$ be as in Exercise 1.2.26. Let $f : \Omega \longrightarrow \Lambda$ be a one to one function. Find the induced measure $\# \circ f^{-1}$.

1.2.29 Suppose that $f \geq 0$ is bounded, $f(x) = 0$ for $x \notin [a, b)$ where $-\infty < a < b < \infty$, and the Riemann integral exists, i.e.

$$\overline{\int_a^b} f(x)dx = \underline{\int_a^b} f(x)dx.$$

Show that f is Borel measurable and hence that the Lebesgue integral exists and is finite.

(Hint: If the Riemann integral exists, you can show that f is a limit of step functions, and then apply Proposition 1.2.1 (c).)

1.2.30 Here we discuss another definition of Riemann integrals that is often used. Suppose f is a nonnegative bounded function on the finite interval $[a, b)$. Given a partition $\Pi = \{ [a_0, a_1), [a_1, a_2), \ldots, [a_{n-1}, a_n) \}$ as in the text, the *mesh of* Π is $|\Pi| = \max\{a_i - a_{i-1} : 1 \le i \le n\}$, i.e. the length of the longest interval in Π. Also, we the set of points $\Xi = \{\xi_1, \ldots, \xi_n\}$ is *compatible* with Π if $\xi_i \in [a_{i-1}, a_i)$, $1 \le i \le n$. For such a partition Π and Ξ compatible with Π, a *Riemann sum approximation* to $\int_a^b f(x)\, dx$ is

$$\mathcal{M}(f, \Pi, \Xi) \; = \; \sum_{i=1}^{n} f(\xi_i)[a_i - a_{i-1}].$$

We write

$$\mathcal{R}' \int_a^b f(x)\, dx \; = \; \lim_{|\Pi| \to 0} \mathcal{M}(f, \Pi, \Xi)$$

if the limit exists no matter what compatible set of points Ξ is used. What we mean by this is there exists a real number I such that given $\epsilon > 0$, there exists a $\delta > 0$ such that for all partitions Π of $[a, b)$ into finitely many intervals with $|\Pi| < \delta$ and for all sets of points Ξ compatible with Π,

$$|I - \mathcal{M}(f, \Pi, \Xi)| \; < \; \epsilon.$$

When such an I exists, it is the value of $\mathcal{R}' \int_a^b f(x)\, dx$.

Show that if the Riemann integral exists by this definition, then it exists under the definition given in the text, and the two integrals are equal.

Hint: It is easy to see that $\mathcal{L}(f, \Pi) \le \mathcal{M}(f, \Pi, \Xi) \le \mathcal{U}(f, \Pi)$, and hence if $\mathcal{R} \int_a^b f(x) dx$, then so does $\mathcal{R}' \int_a^b f(x) dx$ and the two are equal. For the converse, given Π, one can select Ξ so that $\mathcal{M}(f, \Pi, \Xi)$ is arbitrarily close to $\mathcal{L}(f, \Pi)$, and one can select another Ξ so that $\mathcal{M}(f, \Pi, \Xi)$ is arbitrarily close to $\mathcal{U}(f, \Pi)$.

1.2.31 Suppose f is an integrable function. Show $|f| < \infty$ a.e. (Hint: Suppose $f \ge 0$. If $\mu(\{\omega : f(\omega) = \infty\}) > 0$, then there is a sequence of simple functions ϕ_n with $0 \le \phi_n \le f$ and $\int \phi_n \, d\mu \to \infty$.)

1.2.32 Prove that $f \ge 0$ a.e. and $\int f d\mu = 0$ implies $f = 0$ a.e. (Hint: Show that $\mu\{\omega : f(\omega) > 0\} > 0$ implies $\mu\{\omega : f(\omega) > 1/n\} > 0$ for some n.).

1.2.33 Suppose a_{nm}, $n = 1, 2, \ldots$, $m = 1, 2, \ldots$ is a doubly indexed sequence of nonnegative real numbers. Show using the monotone convergence theorem that

$$\sum_{n=1}^{\infty} \sum_{m=1}^{\infty} a_{nm} \; = \; \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} a_{nm} \quad .$$

1.2.34 Suppose X, X_1, X_2, ... are r.v.'s on a common probability space, that $X_n \to X$ a.s., and that for all n, $P[|X_n| \le M] = 1$, where M is a fixed constant. Show $E[X_n] \to E[X]$. Be careful with sets of probability 0.

1.2.35 Suppose μ_1, μ_2, ... is a finite or infinite sequence of measures on $(\Omega, \mathcal{F})$. Define $\mu(A) = \sum_i \mu_i(A)$ for $A \in \mathcal{F}$. We know from Proposition 1.1.5 (a) that μ is a measure on $(\Omega, \mathcal{F})$. Show $\int f\, d\mu = \sum_i \int f d\mu_i$ whenever both sides are defined.

1.2.36 Let $\Omega = \{a_1, a_2, ...\}$ be a discrete set (see Exercise 1.2.26) and put

$$\mu = \sum_i \delta_{a_i}.$$

Identify μ as a measure we know by another name.

1.2.37 Let f_1, f_2, ... be a sequence of nonnegative functions on $(\Omega, \mathcal{F}, \mu)$. Show that $\int \sum_{i=1}^{\infty} f_i\, d\mu = \sum_{i=1}^{\infty} \int f_i d\mu$.

1.2.38 Prove Theorem 1.2.9. Show further that if ν has a density w.r.t. μ, and if $\mu(A) = 0$, then also $\nu(A) = 0$.

1.3 Measures on Product Spaces.

Given two measure spaces $(\Omega_2, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$, we shall show one can construct a "natural" measure on the cartesian product

$$\Omega_1 \times \Omega_2 \;=\; \{\, (\omega_1, \omega_2) \;:\; \omega_1 \in \Omega_1 \text{ and } \omega_2 \in \Omega_2 \,\} \quad .$$

This naturally extends to a finite product of n measure spaces defined through ordered n–tuples.

1.3.1 Basic Definitions and Results.

First, we define the appropriate σ–field on the product space.

Definition 1.3.1 *Given measurable spaces* $(\Omega_1, \mathcal{F}_1)$ *and* $(\Omega_2, \mathcal{F}_2)$*, let* $\mathcal{C} = \{\, A_1 \times A_2 \,:\, A_1 \in \mathcal{F}_1$ *and* $A_2 \in \mathcal{F}_2 \,\}$*. Elements of* $\mathcal{C}$ *are sometimes called* rectangle sets, cylinder sets *or* rectangles*. Then the* product σ–field *is given by*

$$\mathcal{F}_1 \times \mathcal{F}_2 \;=\; \sigma(\mathcal{C}) \quad .$$

□

Note that $\mathcal{F}_1 \times \mathcal{F}_2$ is not simply the cartesian product of $\mathcal{F}_1$ with $\mathcal{F}_2$, despite what the notation suggests. The elements of $\mathcal{F}_1 \times \mathcal{F}_2$ are subsets of $\Omega_1 \times \Omega_2$, whereas the cartesian product of $\mathcal{F}_1$ with $\mathcal{F}_2$ is a collection of ordered pairs of sets. We will denote the measurable space $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2)$ by $(\Omega_1, \mathcal{F}_1) \times (\Omega_2, \mathcal{F}_2)$. When several factor spaces are involved, we shall not hesitate to use the $\prod$ notation, as in

$$\prod_{i=1}^{n} \Omega_i \;=\; \Omega_1 \times \Omega_2 \times \ldots \times \Omega_n \quad ,$$

$$\prod_{i=1}^{n} \mathcal{F}_i \;=\; \sigma\left(\{ \prod_{i=1}^{n} A_i \;:\; \text{each } A_i \in \mathcal{F}_i \,\} \right) \quad .$$

The σ–field on $\mathbb{R}^n$ which is the product σ–field of n copies of $\mathcal{B}$ is called the *σ–field of n–dimensional Borel sets* and is denoted $\mathcal{B}_n$. As with the 1–dimensional case, all subsets of $\mathbb{R}^n$ that arise "in practice" are Borel sets, and all functions $f \;:\; \mathbb{R}^n \longrightarrow \mathbb{R}^m$ that arise "in practice" are measurable functions when the spaces are equipped with their Borel σ–fields.

A measure space $(\Lambda, \mathcal{G}, \mu)$ with $\Lambda \in \mathcal{B}_n$ and $\mathcal{G} = \{\, B \cap \Lambda \,:\, B \in \mathcal{B}_n \}$ is called a *Euclidean space*. Most applications of statistics involve Euclidean spaces.

Given measure spaces $(\Omega_2, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$, we wish to define a "natural" measure on the product measurable space $(\Omega_1, \mathcal{F}_1) \times (\Omega_2, \mathcal{F}_2)$. Before doing this, we will need a technical condition which arises frequently in measure theory. It will also be used in the next section.

Definition 1.3.2 *A measure space $(\Omega, \mathcal{F}, \mu)$ is called σ–*finite *iff there is an infinite sequence A_1, A_2, ... in $\mathcal{F}$ such that*

(i) *$\mu(A_i) < \infty$ for each i;*

(ii) *$\Omega = \bigcup_{i=1}^{\infty} A_i$.*

□

It is easy to check that $(\mathbb{R}, \mathcal{B}, m)$ is σ–finite and $(\mathbb{Z}, \mathcal{P}(\mathbb{Z}), \#)$ is also, where $\mathbb{Z}$ is the set of integers, positive and negative. Non–σ–finite measure spaces tend to be somewhat pathological (see e.g. Exercise 1.3.1) so little is lost by restricting attention to σ–finite spaces.

See Billingsley, Theorems 18.1 and 18.2, pp. 234–236 for the following important theorem.

Theorem 1.3.1 (Product Measure Theorem) *Let $(\Omega_2, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$ be σ–finite measure spaces. Then there exists a unique measure $\mu_1 \times \mu_2$ (called* product measure*) on $(\Omega_1, \mathcal{F}_1) \times (\Omega_2, \mathcal{F}_2)$ such that for all $A_1 \in \mathcal{F}_1$ and $A_2 \in \mathcal{F}_2$,*

$$(\mu_1 \times \mu_2)(A_1 \times A_2) \ = \ \mu_1(A_1) \times \mu_2(A_2) \quad . \tag{1.39}$$

□

Equation (1.39) states that the product measure of a rectangle set is the product of the measures of the factor sets. The theorem extends by induction to an arbitrary finite number of σ–finite measures. We will briefly explain in the case of 3 σ–finite measure spaces $(\Omega_i, \mathcal{F}_i, \mu_i)$, $i = 1, 2, 3$. The theorem tells us that there is a product measure $\nu = \mu_1 \times \mu_2$ on $(\Lambda, \mathcal{G}) = (\Omega_1, \mathcal{F}_1) \times (\Omega_2, \mathcal{F}_2)$. Now ν is also σ–finite; see Exercise 1.3.8. Therefore, applying again the Product Measure Theorem, there is a product measure $\nu \times \mu_3$ on $(\Lambda, \mathcal{G}) \times (\Omega_3, \mathcal{F}_3)$. Technically speaking, the underlying space in this last product is collections of ordered pairs of the form (λ, ω_3) where $\lambda = (\omega_1, \omega_2)$, but we may identify such an ordered pair $((\omega_1, \omega_2), \omega_3)$ with the ordered triple $(\omega_1, \omega_2, \omega_3)$. In this way, we can simply think of $\nu \times \mu_3$ as a measure on the underlying space $\Omega_1 \times \Omega_2 \times \Omega_3$.

The measure on $(\mathbb{R}^n, \mathcal{B}_n)$ obtained by taking the n–fold product of Lebesgue measure is called *n–dimensional Lebesgue measure* and is denoted m^n or simply m if n is clear from context. Note that equation (1.11) gives

$$m^2((a_1, b_1) \times (a_2, b_2)) \ = \ m((a_1, b_1)) \times m((a_2, b_2)) \ = \ (b_1 - a_1) \times (b_2 - a_2) \quad ,$$

where $(a_1, b_1) \times (a_2, b_2)$ is the two dimensional rectangle $\{ (x_1, x_2) : a_1 < x_1 < b_1$ and $a_2 < x_2 < b_2 \}$. Thus, the 2–dimensional Lebesgue measure of a rectangle is simply its area. Since any "nice" geometric figure in 2 dimensions can be "filled" up with disjoint rectangles, it follows that m^2 in general measures area. Similarly, m^3 measures volume in 3–dimensional space.

1.3.2 Integration with Product Measures.

The next theorem shows how to integrate with product measures using iterated integrals. See Billingsley, p. 238 (Theorem 18.3) for a proof.

Theorem 1.3.2 (Fubini's theorem) *Let $\Omega = \Omega_1 \times \Omega_2$, $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2$, and $\mu = \mu_1 \times \mu_2$ where μ_1 and μ_2 are σ–finite. If f is a Borel function on Ω whose integral w.r.t. μ exists, then*

$$\begin{aligned} \int_\Omega f(\omega)\, d\mu(\omega) &= \int_{\Omega_1\times\Omega_2} f(\omega_1,\omega_2)\, d(\mu_1\times\mu_2)(\omega_1,\omega_2) \\ &= \int_{\Omega_2} \left[\int_{\Omega_1} f(\omega_1,\omega_2)\, d\mu_1(\omega_1) \right] d\mu_2(\omega_2) \quad . \qquad (1.40) \end{aligned}$$

Part of the conclusion here is that

$$g(\omega_2) = \int_{\Omega_1} f(\omega_1,\omega_2)\, d\mu_1(\omega_1) \quad , \qquad (1.41)$$

exists μ_2–a.e. and defines a Borel function on Ω_2 whose integral w.r.t. μ_2 exists.

□

The first line in equation (1.40) is simply a change of notation from ω to (ω_1,ω_2), etc. The second line is the meaningful one. Note that in the inner integral of (1.40) (i.e. the r.h.s. of (1.41)), we integrate w.r.t. $d\mu_1(\omega_1)$ while holding ω_2 fixed (or constant), after which ω_1 "disappears" as a variable (as in the l.h.s. of (1.41)). The student is already familiar with such iterated integrals w.r.t. Lebesgue measure m^2. A "symmetry" argument shows that

$$\int_\Omega f\, d\mu = \int_{\Omega_1} \left[\int_{\Omega_2} f(\omega_1,\omega_2)\, d\mu_2(\omega_2) \right] d\mu_1(\omega_1) \quad , \qquad (1.42)$$

i.e. the integral may be evaluated in either order. The mathematical proof of this is actually a bit difficult. One notes that there is an "isomorphism" of measure spaces $(\Omega_1, \mathcal{F}_1, \mu_1) \times (\Omega_2, \mathcal{F}_2, \mu_2)$ and $(\Omega_2, \mathcal{F}_2, \mu_2) \times (\Omega_1, \mathcal{F}_1, \mu_1)$ given by the point map $(\omega_1,\omega_2) \mapsto (\omega_2,\omega_1)$. Then, $\int f(\omega_1,\omega_2)\, d(\mu_1\times\mu_2)(\omega_1,\omega_2) = \int f(\omega_1,\omega_2)\, d(\mu_2\times\mu_1)(\omega_2,\omega_1)$. Hopefully, the student finds obvious the appeal to "symmetry" and doesn't need the long winded argument to be convinced of (1.42).

Example 1.3.1 Let $\Omega_1 = \Omega_2 = I\!N = \{0,\ 1,\ 2,\ ...\}$ the natural numbers, and $\mu_1 = \mu_2 = \#$. One can check that $\# \times \#$ on $I\!N \times I\!N$ is $\#$ on $I\!N^2$. See Exercise 1.3.2. A function $f(n,m)$ on $I\!N^2$, otherwise known as a double sequence, is integrable w.r.t. $\#$ iff

$$\sum_{n=1}^{\infty} \sum_{m=1}^{\infty} |f(n,m)| < \infty \quad , \qquad (1.43)$$

and then

$$\begin{aligned} \int f d\# &= \int \left[\int f(n,m)\, d\#(n) \right] d\#(m) && (1.44) \\ &= \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} f(n,m) && (1.45) \\ &= \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} f(n,m) \quad . \end{aligned}$$

Thus we have shown a well known fact from advanced calculus: if a double series is absolutely summable (i.e. (1.43) holds), then it can be summed in either order. In fact, by Fubini's theorem, it suffices for either the sum of the positive terms to be finite or the sum of the negative terms to be finite. That some condition is required for interchanging the order of the summations is shown by Exercise 1.3.3. See also Exercise 1.2.33.

□

1.3.3 Random Vectors and Stochastic Independence.

We now explore the ramifications of this theory in probability. A function $\underline{X} : (\Omega, \mathcal{F}, P) \longrightarrow I\!\!R^n$ is called an *n–dimensional random vector*, or *random n–vector*, abbreviated r.v. Just as for $n = 1$, the induced measure $P \circ \underline{X}^{-1}$ on $I\!\!R^n$ is called the *distribution* or *law* of $\underline{X}$ and is denoted $P_{\underline{X}}$ or Law$[\underline{X}]$. We will write a vector as a column vector or as an ordered n–tuple, i.e.

$$(x_1, x_2, ..., x_n) = \left[\begin{array}{c} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{array} \right] \quad .$$

We need to use the r.h.s. of this last equation wherein $\underline{x}$ is represented as an $n \times 1$ matrix whenever we do matrix operations. Some authors insist on writing a vector as a row vector ($1 \times n$ matrix in this case), but this author is more accustomed to column vectors. Of course, it doesn't make any difference as long as the reader knows which convention is in use. The component functions X_1, X_2, ... X_n of a random n–vector $\underline{X}$ are random variables (Exercise 1.3.13), and their distributions on $I\!\!R^1$ are referred to as *marginal distributions*. The distribution of $\underline{X}$ on $I\!\!R^n$ is sometimes referred to as *the joint distribution of* X_1, X_2, ..., X_n.

The random variables X_1, X_2, ..., X_n are said to be *(jointly) independent* iff for all B_1, B_2, ..., $B_n \in \mathcal{B}$,

$$P\{X_1 \in B_1,\, X_2 \in B_2, \text{ ... and } X_n \in B_n\} = \prod_{i=1}^{n} P\{X_i \in B_i\} \quad . \qquad (1.46)$$

This definition extends to arbitrary random elements X_1, X_2, ..., X_n. This last displayed equation is equivalent to

$$(P \circ \underline{X}^{-1})(\prod_{i=1}^{n} B_i) \; = \; \prod_{i=1}^{n}(P \circ X_i^{-1})(B_i) \quad , \tag{1.47}$$

where $\underline{X} = (X_1, X_2, ..., X_n)$.

Proposition 1.3.3 *Let $\underline{X} = (X_1, X_2, ..., X_n)$ be a random vector. Then X_1, X_2, ..., X_n are independent iff*

$$Law[\underline{X}] \; = \; \prod_{i=1}^{n} Law[X_i] \quad .$$

Proof. Suppose X_1, X_2, ..., X_n are jointly independent, so (1.46) holds for all B_1, B_2, ..., $B_n \in \mathcal{B}$. Note that the l.h.s. of (1.46) is the joint distribution $P = \text{Law}[\underline{X}]$ evaluated at the rectangle set $B_1 \times B_2 \times ... \times B_n$, and (1.46) says that this equals the product of the corresponding measures of the factor sets. Since this holds for arbitrary rectangle sets, it follows that $P \circ \underline{X}^{-1} = \prod(P \circ X_i^{-1})$ by uniqueness in the Product Measure Theorem, as claimed.

Conversely, if $P \circ \underline{X}^{-1} = \prod(P \circ X_i^{-1})$, then (1.46) holds for all B_1, B_2, ..., B_n by the definition of the product measure, and hence X_1, X_2, ..., X_n are jointly independent.

□

We say X_1, X_2, ..., X_n are pairwise independent iff for all $i \neq j$, the pair X_i and X_j are independent. Joint independence implies pairwise independence, but the converse is false as the counterexample of Exercise 1.3.5 shows. The following result gives some useful consequences of independence.

Theorem 1.3.4 *Let X and Y be random elements defined on a common probability space.*

(a) If X and Y are independent, then so are $g(X)$ and $h(Y)$ where g and h are appropriately measurable functions.

(b) If g and h in part (a) are real valued, then

$$E[g(X)h(Y)] \; = \; E[g(X)]E[h(Y)] \tag{1.48}$$

provided all three functions are integrable.

Proof. Part (a) is left as Exercise 1.3.6. For (b), let $P = \text{Law}[g(X)]$ and $Q = \text{Law}[h(Y)]$. By part (a), $\text{Law}[(g(X), h(Y))] = P \times Q$. By the law of the unconscious statistician (Theorem 1.2.10),

$$\begin{aligned} E[g(X)h(Y)] &= \int_{R^2} g \cdot h \, d(P \times Q)(g,h) \\ &= \int_R \left[\int_R g \cdot h \, dP(g) \right] dQ(h) \\ &= \int_R \left[\int_R g \, dP(g) \right] h \, dQ(h) \\ &= \left[\int_R g dP(g) \right] \cdot \left[\int_R h \, dQ(h) \right] \\ &= E[g(X)] \cdot E[h(Y)] \quad . \end{aligned}$$

In the above, the second equality follows from Fubini's theorem (Theorem 1.3.2), the third from Proposition 1.2.5 (a) (note that h can be factored out of the integral w.r.t. $dP(g)$ as it is constant within this integral), the fourth from Proposition 1.2.5 (a) again ($\int g dP(g)$ is a constant), and the last equality from the law of the unconscious statistician applied to each factor integral. This completes the proof of (b).

□

Remarks 1.3.1 While we generally avoid checking measurability in this text, the following shows that measurability w.r.t. a product σ–field on the range space follows from measurability of the component functions w.r.t. the factor σ–fields. Suppose $f : \Omega \longrightarrow \Lambda_1 \times \Lambda_2$ is any function. Define the *projections*

$$\pi_i : \Lambda_1 \times \Lambda_2 \longrightarrow \Lambda_i \quad , \quad \pi_i(\lambda_1, \lambda_2) \;=\; \lambda_i \quad ,$$

and the *coordinate* or *component functions* of f by

$$f_i(\omega) \;=\; (\pi_i \circ f)(\omega) \;=\; \pi_i(f(\omega)) \quad .$$

So we may write f in ordered pair notation by $f(\omega) = (f_1(\omega), f_2(\omega))$.

Theorem 1.3.5 *Suppose $f : \Omega \longrightarrow \Lambda_1 \times \Lambda_2$ where $(\Omega, \mathcal{F})$, $(\Lambda_1, \mathcal{G}_1)$, and $(\Lambda_2, \mathcal{G}_2)$ are measurable spaces. Then f is measurable from $(\Omega, \mathcal{F})$ to $(\Lambda_1, \mathcal{G}_1) \times (\Lambda_2, \mathcal{G}_2)$ iff each coordinate function f_i is measurable (from $(\Omega, \mathcal{F})$ to $(\Lambda_i, \mathcal{G}_i)$), for $i = 1, 2$.*

Proof. That measurability of f implies measurability of the coordinate functions is Exercise 1.3.13. For the converse, assume the coordinate functions f_1 and f_2 are measurable. If $C_1 \times C_2$ is a rectangle set (i.e. each $C_i \in \mathcal{G}_i$), then

$$\begin{aligned} f^{-1}(C_1 \times C_2) &= \{\, \omega : f(\omega) \in C_1 \times C_2 \,\} \\ &= \{\, \omega : f_1(\omega) \in C_1 \text{ and } f_2(\omega) \in C_2 \,\} \\ &= f_1^{-1}(C_1) \cap f_2^{-1}(C_2) \\ &\in \mathcal{F} \quad , \end{aligned}$$

since each $f_i^{-1}(C_i) \in \mathcal{F}$. This shows that the inverse image of a rectangle set under f is measurable, but we must show this works for an arbitrary element of $\mathcal{G}_1 \times \mathcal{G}_2$. Now by Exercise 1.2.20,

$$\mathcal{C} \; = \; \{\, C \subset \Lambda_1 \times \Lambda_2 \, : \; f^{-1}(C) \in \mathcal{F} \,\}$$

is a σ–field on $\Lambda_1 \times \Lambda_2$, and we have just shown that $\mathcal{C}$ includes the rectangle sets. Since $\mathcal{G}_1 \times \mathcal{G}_2$ is the smallest σ–field which includes the rectangle sets (Definition 1.3.1), it follows that $\mathcal{G}_1 \times \mathcal{G}_2 \subset \mathcal{C}$, and hence that $f^{-1}(\mathcal{G}_1 \times \mathcal{G}_2) \subset \mathcal{C}$ by definition of $\mathcal{C}$. This shows f is measurable from $(\Omega, \mathcal{F})$ to $(\Lambda_1, \mathcal{G}_1) \, \times \, (\Lambda_2, \mathcal{G}_2)$.

□

As an application of this last result, suppose X_1, X_2, ..., X_n are r.v.'s defined on a common probability space. Then $\underline{X} = (X_1, \, X_2, \, ..., \, X_n)$ is automatically a random vector.

Exercises for Section 1.3.

1.3.1 Let $(\Omega, \mathcal{F}, \mu)$ be any measure space with $\mu(\Omega) > 0$ and define ν on $(\Omega, \mathcal{F})$ by

$$\nu(A) = \begin{cases} 0 & \text{if } \mu(A) = 0, \\ \infty & \text{if } \mu(A) \neq 0. \end{cases}$$

Show that ν is a measure and that $(\Omega, \mathcal{F}, \nu)$ is not σ-finite.

1.3.2 Show that the product of counting measures on two discrete sets is counting measure on the product set.

1.3.3 Let $f : I\!\!N \times I\!\!N \longrightarrow I\!\!R$ be given by

$$f(n, m) = \begin{cases} 1 & \text{if } n = m, \\ -1 & \text{if } n = m - 1, \\ 0 & \text{otherwise.} \end{cases}$$

Show that

$$\sum_{n=1}^{\infty} \left[\sum_{m=1}^{\infty} f(n, m) \right] \neq \sum_{n=1}^{\infty} \left[\sum_{m=1}^{\infty} f(n, m) \right] \quad .$$

Is f integrable w.r.t. $\# \times \#$? Determine the answer to this last question directly from the definition of integrability.

1.3.4 Show that (1.46) and (1.47) are equivalent.

1.3.5 Let $(X_1,\ X_2,\ X_3)$ have distribution on $I\!\!R^3$ which puts measure 1/4 on each of the points $(0, 0, 0)$, $(1, 1, 0)$, $(1, 0, 1)$, and $(0, 1, 1)$.

(a) Show that the distribution of any pair of the r.v.'s is the product of two copies of the measure on $I\!\!R$ which puts measure 1/2 at each point 0 and 1. Conclude that $(X_1,\ X_2,\ X_3)$ are pairwise independent.

(b) Show that $(X_1,\ X_2,\ X_3)$ are not jointly independent.

1.3.6 Prove Theorem 1.3.4 (a).

1.3.7 Suppose (1.48) holds for all bounded real valued functions g and h on the ranges of X and Y, respectively. Show X and Y are independent.

1.3.8 Show that the product of two σ-finite measure spaces is also σ-finite. Hint: Let $(\Omega_2, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$ be σ-finite, say $\Omega_1 = \bigcup_{n=1}^{\infty} A_n$ where $\mu_1(A_n) < \infty$, and $\Omega_2 = \bigcup_{m=1}^{\infty} B_m$ where $\mu_2(B_m) < \infty$. Let $C_{nm} = A_n \times B_m$. The collection $\{C_{nm} : n, m = 1, 2, \ldots\}$ is countable by Exercise 1.1.6 (a).

1.3.9 Define by induction the product of n measurable spaces and the product of n measure spaces.

1.3.10 Evaluate the integral $\int f\,d(\mu_1 \times \mu_2)$ when f, μ_1, and μ_2 are as given below.

(a) μ_1 and μ_2 are both m, Lebesgue measure, and

$$f(x_1, x_2) \;=\; \begin{cases} 1 & \text{if } x_1^2 + x_2^2 \le 1, \\ & \\ 0 & \text{otherwise.} \end{cases}$$

(b) μ_1 and μ_2 are the same as in (a) but

$$f(x_1, x_2) \;=\; \begin{cases} e^{-x_1/x_2} & \text{if } 0 \le x_1 < \infty \text{ and } 0 < x_2 < 1, \\ & \\ 0 & \text{otherwise.} \end{cases}$$

(c) μ_1 and μ_2 are counting measure on $I\!\!N$, and $f(n, m) = 2^{-(n+m)}$.

(d) μ_1 is Lebesgue measure, μ_2 is counting measure on $I\!\!N$, and

$$f(x, n) \;=\; \begin{cases} x^n & \text{if } 0 < x_1 \le 1/2, \\ & \\ 0 & \text{otherwise.} \end{cases}$$

(e) μ_1 is Lebesgue measure m^2 on $I\!\!R^2$, μ_2 is counting measure on $\{0, \pi, 2\pi\}$, and $f((x_1, x_2), x_3) = \exp[|x_1| - |x_2| + \cos(x_3)]$.

1.3.11 Generalize Theorem 1.3.4 to the case of n random elements.

1.3.12 Suppose X_1, X_2, ..., X_n are independent r.v.'s with c.d.f.'s given by F_1, F_2, ..., F_n, respectively. What is the c.d.f. of $Y = \max\{\, X_1, X_2, ..., X_n\}$?

1.3.13 Let $f = (f_1, f_2) : \; (\Omega, \mathcal{F}) \longrightarrow (\Lambda_1, \mathcal{G}_1) \times (\Lambda_2, \mathcal{G}_2)$. Show that each of the component functions f_1 and f_2 are measurable, i.e. that $f_1 : (\Omega, \mathcal{F}) \longrightarrow (\Lambda_1, \mathcal{G}_1)$.

1.3.14 Prove or disprove the following: Let $f : \Omega_1 \times \Omega_2 \longrightarrow \Lambda$ where $(\Omega_i, \mathcal{F}_i)$, $i = 1, 2$, and $(\Lambda, \mathcal{G})$ are measurable spaces. Suppose $f(\omega_1, \omega_2)$ is a measurable function of ω_1 for each fixed ω_2 and is also a measurable function of ω_2 for each fixed ω_1. Then f is measurable $(\Omega_1, \mathcal{F}_1) \times (\Omega_2, \mathcal{F}_2) \longrightarrow (\Lambda, \mathcal{G})$. You may assume the existence of non–Borel sets $V \subset I\!\!R$. (Hint: Let V be such a subset of $I\!\!R$ and consider the indicator of $D = \{(x_1, x_2) \in I\!\!R^2 : x_1 = x_2 \in V\}$.) Contrast this claim with Theorem 1.3.5.

1.3.15 Give an alternate solution to Exercise 1.2.37 using Fubini's theorem. State and prove a possible result when the f_i's are not required to be nonnegative.

1.3.16 Let μ_1 and μ_2 be σ-finite measures on $(\Omega, \mathcal{F})$, and let ν be a σ-finite measure on $(\Lambda, \mathcal{G})$.

(a) Show $\mu_1 + \mu_2$ is σ-finite. See Exercise 1.2.35 for the definition of $\mu_1 + \mu_2$. Hint: Similarly to Exercise 1.3.8, but consider $C_{nm} = A_n \cap B_m$.

(b) Show that $(\mu_1 + \mu_2) \times \nu = (\mu_1 \times \nu) + (\mu_2 \times \nu)$.

1.3.17 Let A be a measurable set on the product space $(\Omega_2, \mathcal{F}_1, \mu_1) \times (\Omega_2, \mathcal{F}_2, \mu_2)$ and for each $\omega_1 \in \Omega_1$ put

$$A(\omega_1) \;=\; \{\, \omega_2 \,:\, (\omega_1, \omega_2) \in A \,\} \quad .$$

(Note: Drawing a picture of an example when $\Omega_i = I\!R$ may be helpful here.) Show that $\mu_2(A(\omega_1)) = 0$ μ_1-a.e. iff $(\mu_1 \times \mu_2)(A) = 0$. Apply this to determine $m^2(\{(x_1, x_2) \,:\, x_1 = x_2\,\})$.

1.3.18 Find $m^n(\{\ (x_1,\ x_2,\ ...,\ x_n)\ :\ x_i = x_j$ for some $i \neq j\})$. (Hint: consider the special cases $n = 2$ and $n = 3$ first. In these cases, you can “see” the set.)

1.3.19 Determine whether or not the following function is integrable w.r.t. m^2:

$$f(x_1, x_2) \;=\; \begin{cases} (x_1 x_2)/(x_1^2 + x_2^2)^2 & \text{if } |x_1| \leq 1 \text{ and } |x_2| \leq 1, \\ \\ 0 & \text{otherwise.} \end{cases}$$

1.4 Densities and The Radon-Nikodym Theorem.

In this section, we introduce the basic notion of a "density," also called a "Radon-Nikodym derivative." The definition we will give here is very general.

1.4.1 Basic Definitions and Results.

Definition 1.4.1 *Let μ and ν be measures on $(\Omega, \mathcal{F})$. We say ν is* absolutely continuous *w.r.t. μ and write $\nu \ll \mu$ iff for all $A \in \mathcal{F}$, $\mu(A) = 0$ implies $\nu(A) = 0$. We sometimes say μ* dominates *ν, or that μ is a* dominating measure *for ν. We say ν and μ are equivalent (and write $\nu \simeq \mu$) iff both $\nu \ll \mu$ and $\mu \ll \nu$.*

□

In words, $\nu \ll \mu$ if the collection of μ-null sets is a subcollection of the collection of ν-null sets, i.e. ν "has more null sets than" μ. By Exercise 1.2.38, if $(\Omega, \mathcal{F}, \mu)$ is a measure space and $f : \Omega \longrightarrow [0, \infty]$ is Borel measurable, then

$$\nu(A) \;=\; \int_A f d\mu \tag{1.49}$$

defines a measure ν on the same measurable space $(\Omega, \mathcal{F})$. It is easy to show that $\nu \ll \mu$ (Exercise 1.4.1). It turns out that a converse is true also, provided μ is σ–finite. See Billingsley, Theorem 32.2, p. 443.

Theorem 1.4.1 (Radon-Nikodym Thoerem.) *Let $(\Omega, \mathcal{F}, \mu)$ be a σ–finite measure space and suppose $\nu \ll \mu$. Then there is a nonnegative Borel function f such that*

$$\nu(A) \;=\; \int_A f d\mu \quad , \textit{ for all } A \in \mathcal{F} \quad . \tag{1.50}$$

Furthermore, f is unique μ-a.e.; i.e. if $\nu(A) = \int_A g\, d\mu$ for all $A \in \mathcal{F}$, then $g = f$ μ-a.e.

□

The function f given in (1.50) is called the *Radon-Nikodym derivative* or *density* of ν w.r.t. μ, and is often denoted $d\nu/d\mu$. If $\mu = m$ is Lebesgue measure, then f is called a *Lebesgue density* or a *density of the continuous type.* We say a random variable X is a *continuous random variable* iff Law$[X]$ has a Lebesgue density, and we refer to this density at the *density of X* and will often write

$$f_X(x) \;=\; \frac{d\mathrm{Law}[X]}{dm}(x) \quad .$$

Similarly, if $\underline{X}$ is a random n–vector and Law$[\underline{X}] \ll m^n$, then we say $\underline{X}$ is a *continuous random vector* with a similar notation for its Lebesgue density, which is sometimes also called a density of the continuous type.

We may write (1.50) in the form

$$\nu(A) \;=\; \int_A 1\, d\nu \;=\; \int_A \frac{d\nu}{d\mu}\, d\mu \;=\; \int_A \frac{d\nu}{d\mu}(\omega)\, d\mu(\omega) \quad . \tag{1.51}$$

Notice how the $d\mu$'s "cancel" on the r.h.s. Also, the Radon-Nikodym derivative is only determined μ-a.e., i.e. we can change its value on a set of μ-measure 0 and not change the measure ν defined by the density. A particular choice for the function $d\nu/d\mu$ is called a *version* of the Radon-Nikodym derivative. Two versions of $d\nu/d\mu$ are equal μ-a.e. Another way we will sometimes indicate a Radon-Nikodym derivative is the following notation:

$$d\nu \;=\; f d\mu \text{ or } d\nu(x) \;=\; f(x)d\mu(x).$$

One can formally divide both sides of the equation by $d\mu$ to obtain $d\nu/d\mu = f$.

Radon-Nikodym derivatives or densities are widely used in probability and statistics. The student is no doubt familiar with the usage of the term "density" when refering to a Lebesgue density, but we may also have densities for discrete measures as the next example shows.

Example 1.4.1 Let $\Omega = \{a_1, a_2, ...\}$ be a discrete set (finite or infinite), and let μ be a measure on $(\Omega, \mathcal{P}(\Omega))$. Put

$$f(a) \;=\; \mu(\{a\}) \quad . \tag{1.52}$$

Then we claim that $d\mu/d\# = f$, where # is counting measure on Ω. To see this, note that by property (iii) of a measure (Definition 1.1.4),

$$\mu(A) \;=\; \sum_{a_i \in A} \mu(\{a_i\}) \;=\; \sum_{a_i \in A} f(a_i) \;=\; \int f d\# \quad .$$

The last equality follows from (1.27) In this context, it is sometimes said that f is a *density of the discrete type* for μ. If μ is a probability measure, the density of the discrete type is also sometimes called the *probability mass function.* If a random variable has a distribution which is dominated by counting measure, then it is called a *discrete random variable.*

□

Recall that a unit point mass measure at ω is given by

$$\delta_\omega(A) \;=\; \begin{cases} 1 & \text{if } \omega \in A, \\ & \\ 0 & \text{otherwise.} \end{cases} \tag{1.53}$$

Then a measure μ as given in (1.52) can be written as $\mu = \sum_i f(a_i)\delta_{a_i}$. The following example shows that point mass measures can be useful components of dominating measures for distributions which arise in applied statistics.

Example 1.4.2 Suppose a r.v. X is obtained by measuring the concentration of a chemical in water, but because of limitations of the measuring instrument, concentrations less than some amount x_0 are reported as x_0. (For instance, an instrument which measures lead concentration in water may register $x_0 = 10^{-5}$ grams per liter for any concentration at or below this value.) Suppose Y is the true concentration, then we might think of X as given by $X = \max\{x_0, Y\}$. Suppose Y has Lebesgue density $f_Y(y) = e^{-y}$, $y > 0$, and 0 otherwise. Then X does not have a Lebesgue density because $P[X = x_0] = 1 - e^{-x_0}$ but $m(\{x_0\}) = 0$ so we do not have $\text{Law}[X] \ll m$. But X does have a density w.r.t. the measure

$$\mu \;=\; m + \delta_{x_0} \quad ,$$

which is given by

$$f_X(x) \;=\; \frac{d\,\text{Law}[X]}{d\mu}(x) \;=\; \begin{cases} e^{-x} & \text{if } x > x_0, \\ 1 - e^{-x_0} & \text{if } x = x_0, \\ 0 & \text{otherwise.} \end{cases} \tag{1.54}$$

The details are left to Exercise 1.4.3.

□

In the last example we said "X does have a density ..." when we really meant "Law$[X]$ does have a density ..." This is a common abuse of terminology one sees in probability and statistics.

Proposition 1.4.2 (Calculus with Radon-Nikodym derivatives.) *Let $(\Omega, \mathcal{F})$ be a measurable space with measures μ, ν, ν_1, ν_2, and λ. Assume μ and λ are σ–finite.*

(a) If $\nu \ll \mu$ and $f \geq 0$, then

$$\int f d\nu \;=\; \int f \left(\frac{d\nu}{d\mu} \right) d\mu \quad .$$

(b) If $\nu_i \ll \mu$, then $\nu_1 + \nu_2 \ll \mu$ and

$$\frac{d(\nu_1 + \nu_2)}{d\mu} \;=\; \frac{d\nu_1}{d\mu} + \frac{d\nu_2}{d\mu} \quad , \quad \mu - a.e.$$

(c) (Chain rule.) If $\nu \ll \mu \ll \lambda$, then

$$\frac{d\nu}{d\lambda} \;=\; \frac{d\nu}{d\mu}\frac{d\mu}{d\lambda} \quad , \quad \lambda - a.e.$$

In particular, if $\mu \simeq \nu$ then

$$\frac{d\nu}{d\mu} \;=\; \left(\frac{d\mu}{d\nu} \right)^{-1} \quad , \quad \mu - \ (or\ \nu-) a.e.$$

Partial Proof. (a) The result is obviously true for indicators. Proceed to simple functions, then take limits using the Monotone Convergence Theorem and Proposition 1.2.8.

(b) Note that $\nu_1 + \nu_2$ is a measure by Exercise 1.2.35 (a). Now $\nu_i \ll \mu$ for $i = 1, 2$ implies $\nu_1 + \nu_2 \ll \mu$ (Exercise 1.4.6). If $A \in \mathcal{F}$ then

$$\begin{aligned}(\nu_1 + \nu_2)(A) &= \nu_1(A) + \nu_2(A) \\ &= \int_A \frac{d\nu_1}{d\mu}\, d\mu + \int_A \frac{d\nu_2}{d\mu}\, d\mu \\ &= \int_A \left[\frac{d\nu_1}{d\mu} + \frac{d\nu_2}{d\mu} \right] d\mu \quad .\end{aligned}$$

The first equality follows from the definition of $\nu_1 + \nu_2$, the second from the definition of $d\nu_i/d\mu$, and the third from linearity of the integral. By uniqueness of the Radon-Nykodym derivative, the integrand $(d\nu_1/d\mu) + (d\nu_2/d\mu)$ in the last displayed expression must be a version of $d(\nu_1 + \nu_2)/d\mu$, as required.

(c) This follows from a similar appeal to uniqueness of the Radon-Nykodym derivative along with part (a). See Exercise 1.4.9.

□

Note how the $d\mu$ "cancels" from the r.h.s. of the equations in part (a) to give the l.h.s. Part (a) of the last proposition is familiar in the context of probability and statistics in the following way: if X is a continuous r.v. with Lebesgue density f and g is a Borel measurable function $I\!R \longrightarrow I\!R$, then

$$E[g(X)] = \int_R g \, d\,\mathrm{Law}[X] = \int_{-\infty}^{\infty} g(x) f(x)\, dx \quad .$$

Note that the first equality is the law of the unconscious statistician (Theorem 1.2.10).

1.4.2 Densities w.r.t. Product Measures.

Proposition 1.4.3 *Let $(\Omega_i, \mathcal{F}_i, \mu_i)$ and $(\Omega_i, \mathcal{F}_i, \nu_i)$, $i = 1, 2$ be σ–finite measure spaces, with $\nu_i \ll \mu_i$, $i = 1, 2$. Then $\nu_1 \times \nu_2 \ll \mu_1 \times \mu_2$ and*

$$\frac{d(\nu_1 \times \nu_2)}{d(\mu_1 \times \mu_2)}(\omega_1, \omega_2) = \left[\frac{d\nu_1}{d\mu_1}(\omega_1) \right] \left[\frac{d\nu_2}{d\mu_1}(\omega_2) \right] \quad , \quad \mu_1 \times \mu_2 - a.e.$$

Proof. Let $A_i \in \mathcal{F}_i$ for $i = 1, 2$. Then

$$\begin{aligned}(\nu_1 \times \nu_2)(A_1 \times A_2) &= \nu_1(A_1)\nu_2(A_2) \quad , \\ &= \int_{A_1} \frac{d\nu_1}{d\mu_1}(\omega_1)\, d\mu_1(\omega_1) \int_{A_2} \frac{d\nu_2}{d\mu_2}(\omega_2)\, d\mu_2(\omega_2)\end{aligned}$$

$$\begin{aligned}
&= \int_{A_1}\int_{A_2} \frac{d\nu_1}{d\mu_1}(\omega_1)\frac{d\nu_2}{d\mu_2}(\omega_2)\, d\mu_2(\omega_2)d\mu_1(\omega_1) \\
&= \int_{\Omega_1}\int_{\Omega_2} I_{A_1}(\omega_1)I_{A_2}(\omega_2)\frac{d\nu_1}{d\mu_1}(\omega_1)\frac{d\nu_2}{d\mu_2}(\omega_2)\, d\mu_2(\omega_2)d\mu_1(\omega_1) \\
&= \int_{\Omega_1}\int_{\Omega_2} I_{A_1\times A_2}(\omega_1,\omega_2)\frac{d\nu_1}{d\mu_1}(\omega_1)\frac{d\nu_2}{d\mu_2}(\omega_2)\, d\mu_2(\omega_2)d\mu_1(\omega_1) \\
&= \int_{\Omega_1\times\Omega_2} I_{A_1\times A_2}(\omega)\frac{d\nu_1}{d\mu_1}(\omega_1)\frac{d\nu_2}{d\mu_2}(\omega_2)\, d(\mu_1\times\mu_2)(\omega_1,\omega_2) \\
&= \int_{A_1\times A_2} \frac{d\nu_1}{d\mu_1}(\omega_1)\frac{d\nu_2}{d\mu_2}(\omega_2)\, d(\mu_1\times\mu_2)(\omega_1,\omega_2) \quad ,
\end{aligned}$$

where the second to last equality follows from Fubini's theorem. By the uniqueness part of the Product Measure Theorem (Theorem 1.3.1), it follows that the measure

$$\nu(C) \;=\; \int_C \frac{d\nu_1}{d\mu_1}(\omega_1)\frac{d\nu_2}{d\mu_2}(\omega_2)\, d(\mu_1\times\mu_2)(\omega_1,\omega_2) \quad ,$$

defined on $(\Omega_1,\mathcal{F}_1,\mu_1)\times(\Omega_2,\mathcal{F}_2,\mu_2)$ is in fact $\nu_1\times\nu_2$. Now $\nu \ll \mu_1\times\mu_2$ by Exercise 1.4.1, and so by the uniqueness part of the Radon-Nikodym theorem,

$$\frac{d\nu}{d(\mu_1\times\mu_2)}(\omega_1,\omega_2) \;=\; \left[\frac{d\nu_1}{d\mu_1}(\omega_1)\right]\left[\frac{d\nu_2}{d\mu_1}(\omega_2)\right] \quad , \quad \mu_1\times\mu_2 - \text{a.e.}$$

□

Remarks 1.4.1 The last result implies that if X_1 and X_2 are independent continuous r.v.'s with (Lebesgue) densities f_1 and f_2, then the joint distribution of (X_1, X_2) is also continuous (i.e. Law$[(X_1,X_2)] \ll m^2$) and the joint density f w.r.t. m^2 is the product of the marginal densities, i.e. $f(x_1,x_2) = f_1(x_1)f_2(x_2)$. Of course, this remark (and the preceding Proposition) can be extended to more than two r.v.'s (measures, respectively) by induction. The converse of this remark is also true (Exercise 1.4.11).

□

Under independence, we can construct the joint density w.r.t. the product of the dominating measures from the marginal densities by simple multiplication. In general, there is no such nice relationship between the joint and the marginal densities, but we can always recover the marginal densities from the joint density.

Proposition 1.4.4 (Marginalization of a density) *Let $(\Omega_i,\mathcal{F}_i,\mu_i)$, $i = 1,2$ be σ–finite measure spaces, and suppose $\nu \ll \mu_1\times\mu_2$. Let $\pi_1 : \Omega_1\times\Omega_2 \to \Omega_1$ be the coordinate projection given by $\pi_1(\omega_1,\omega_2) = \omega_1$, and similarly for π_2. Then $\nu\circ\pi_i^{-1} \ll \mu_i$, $i = 1,2$, and*

$$\frac{d(\nu\circ\pi_1^{-1})}{d\mu_1}(\omega_1) \;=\; \int_{\Omega_2} \frac{d\nu}{d(\mu_1\times\mu_2)}(\omega_1,\omega_2)\, d\mu_2(\omega_2) \quad .$$

Proof. For notational simplicity put

$$\begin{aligned} \nu_1 &= \nu \circ \pi_1^{-1} \quad , \\ f(\omega_1, \omega_2) &= \frac{d\nu}{d(\mu_1 \times \mu_2)}(\omega_1, \omega_2) \quad , \\ f_1(\omega_1) &= \int_{\Omega_2} f(\omega_1, \omega_2)\, d\mu_2(\omega_2) \quad . \end{aligned}$$

Note that ν_1 is a measure on $(\Omega_1, \mathcal{F}_1)$. Our goal in this proof is to show that $\nu_1 \ll \mu_1$, and then that $d\nu_1/d\mu_1 = f_1$, μ_1–a.e.

Now if $A \in \mathcal{F}_1$ then $\pi_1^{-1}(A) = A \times \Omega_2$ is a rectangle set. Thus

$$\begin{aligned} \nu_1(A) &= \nu\left(\pi_1^{-1}(A)\right) \\ &= \nu(A \times \Omega_2) \\ &= \int_{A \times \Omega_2} f(\omega_1, \omega_2)\, d(\mu_1 \times \mu_2)(\omega_1, \omega_2) \\ &= \int_A \left[\int_{\Omega_2} f(\omega_1, \omega_2)\, d\mu_2(\omega_2)\right] d\mu_1(\omega_1) \\ &= \int_A f_1(\omega_1)\, d\mu_1(\omega_1) \quad . \end{aligned}$$

Note that Fubini's theorem was used at the fourth line. The other steps in the above calculation follow from plugging in definitions. Now if $\mu_1(A) = 0$, then the last integral above is 0, so $\nu_1(A) = 0$ and we have that $\nu_1 \ll \mu_1$. Furthermore, since $A \in \mathcal{F}_1$ was arbitrary, the last display shows that we can calculate the ν_1 measure of a set by integrating w.r.t. $d\mu_1$ the function f_1 over the set. Hence, by the uniqueness part of the Radon–Nikodym theorem, $d\nu_1/d\mu_1 = f_1$, μ_1–a.e.

□

See Exercise 1.4.12 for specialization of this last result to probability theory.

1.4.3 Support of a Measure.

Before introducing the next important concept from measure theory, we briefly review the topology of Euclidean spaces. This is discussed at much greater length in Rudin's book, *Principles of Mathematical Analysis.* Let $x \in I\!R^n$, then a *neighborhood* of x is any ball (or sphere) of positive radius ϵ centered at x. A ball of positive radius ϵ centered at x is a set of the form

$$B(x, \epsilon) \;=\; \{\, y \in I\!R^n \,:\, \|x - y\| < \epsilon \,\} \quad .$$

Here, $\|\cdot\|$ denotes the *norm* on $I\!R^n$ given by

$$\|x\| \;=\; \|(x_1, ..., x_n)\| \;=\; \sqrt{x_1^2 + ...x_n^2} \quad .$$

A set $A \subset I\!R^n$ is called *open* iff for every $x \in A$, there is some $\epsilon > 0$ such that $B(x, \epsilon) \subset A$. A set $C \subset I\!R^n$ is called *closed* iff it is the complement of an open set. One can show that a union of open sets is also open, and hence that an intersection of closed sets is also closed. Also, the sets $I\!R^n$ and $\emptyset$ are both open and closed. Thus, any set $D \subset I\!R^n$ is contained in some closed set (namely $I\!R^n$), and the intersection of all closed sets which contain D is also a closed set, namely the smallest closed set containing D. This set is called the *closure* of D and denoted $\bar{D}$. $\bar{D}$ is also given by the following

$$\begin{aligned} \bar{D} \quad = \quad & \{ \lim_n x_n \ : \ x_1, x_2, ..., x_n, ... \text{ is a sequence of} \\ & \qquad \text{points in } D \text{ for which the } \lim_n \text{ exists.} \} \end{aligned} \qquad (1.55)$$

Otherwise said, $\bar{D}$ is the set of limit points of D. Now we briefly explore a concept related to absolutely continuity.

Definition 1.4.2 *Suppose ν is a measure on $(I\!R^n, \mathcal{B}_n)$. The* support *of ν is the set*

$$supp(\nu) \ = \ \{ \, x \in I\!R^n \ : \ \nu(B(x,\epsilon)) > 0 \ \textit{for all} \ \epsilon > 0 \, \} \quad .$$

□

One can show that supp(ν) is a closed set (Exercise 1.4.13), and if ν is a p.m., then supp(ν) is the smallest closed set with probability 1 (Exercise 1.4.14).

Proposition 1.4.5 *Suppose μ and ν are Borel measures on $I\!R^n$, μ is σ–finite, and $\nu \ll \mu$. Then supp(ν) $\subset \bar{S}$ where*

$$S \ = \ \{ \, x \in supp(\mu) \ : \ \frac{d\nu}{d\mu}(x) > 0 \, \} \quad .$$

Proof. Let $x \in$ supp(ν), then for any $\epsilon > 0$ we have

$$\nu(B(x,\epsilon)) \ = \ \int_{B(x,\epsilon)} \frac{d\nu}{d\mu} \, d\mu \ > \ 0 \quad .$$

In particular, the nonnegative function $I_{B(x,\epsilon)}(y) \cdot (d\nu/d\mu)(y)$ cannot be identically 0 on $B(x,\epsilon)$, i.e. $[d\nu/d\mu](y) > 0$ for some $y \in B(x,\epsilon)$.

Now let A_n be the sequence of balls $B(x, 1/n)$ and $y_n \in A_n$ such that $[d\nu/d\mu](y_n) > 0$. One checks that $y_n \to x$, i.e. x is a limit point of S, so $x \in \bar{S}$, as asserted.

□

Remarks 1.4.2 As a corollary to this and Exercise 1.4.16, $\nu \ll \mu$ σ–finite implies $\mathrm{supp}(\nu) \subset \mathrm{supp}(\mu)$. The converse is false, i.e. $\mathrm{supp}(\nu) \subset \mathrm{supp}(\mu)$ does not imply $\nu \ll \mu$. Also, we cannot in general claim $\mathrm{supp}(\nu) = \bar{S}$ in Proposition 1.4.5 (see Exercise 1.4.15). One does however have the next result.

□

Proposition 1.4.6 *Let $U \subset \mathbb{R}^n$ be open. Suppose*

(i) *μ is Lebesgue measure restricted to U, i.e. $\mu(B) = m(B \cap U)$ for all $B \in \mathcal{B}_n$;*

(ii) *$\nu \ll \mu$;*

(iii) *the version of $f = d\nu/d\mu$ is continuous on U;*

Then $supp(\nu) = \bar{S}$ where

$$S \;=\; \{\, x \in U \;:\; f(x) > 0 \,\}. \tag{1.56}$$

Proof. Now f continuous on U and $f(x) > 0$ for some $x \in U$ implies there is a $\epsilon > 0$ such that $f(y) > \epsilon$ for all y in some neighborhood $B(x, \delta_0)$ of x. Hence, for $x \in S$ where S is given in (1.56), we have for all $\delta > 0$ that

$$\nu\,(B(x,\delta)) \;\geq\; \epsilon m^n\,(\; B(\,x\,,\,\min\{\delta,\delta_0\}\,)\,) \qquad .$$

Since the r.h.s. above is positive, it follows that $S \subset \mathrm{supp}(\nu)$.

On the other hand, if $x \in \mathrm{supp}(\nu)$, then for all $\delta > 0$

$$0 \;<\; \int_{B(x,\delta)} f(y)\, dm(y)$$

so in particular, for all $\delta > 0$ there is a $y \in B(x, \delta)$ with $f(y) > 0$ and we can find a sequence $y_n \in S$ with $y_n \to x$. Thus, $x \in \bar{S}$, and we have shown that $\mathrm{supp}(\nu) \subset \bar{S}$. Since $\bar{S}$ is the smallest closed set containg S and $\mathrm{supp}(\nu)$ is a closed set containing S (by the first part of the proof), it follows that $\mathrm{supp}(\nu) = \bar{S}$.

□

Example 1.4.3 Consider the exponential distribution with Lebesgue density

$$f(x) \;=\; \begin{cases} e^{-x} & \text{if } x \geq 0, \\ \\ 0 & \text{otherwise.} \end{cases}$$

We cannot apply the previous proposition to this version of the density, but we can apply it to

$$f(x) = \begin{cases} e^{-x} & \text{if } x > 0, \\ \\ 0 & \text{otherwise.} \end{cases}$$

which is another version (that agrees with the first version except on the set $\{0\}$, which has Lebesgue measure 0). In this second version, the density is positive on the open set $(0, \infty)$, and so the support is the (closed) set $[0, \infty)$.

□

Exercises for Section 1.4.

1.4.1 Let ν be defined as in (1.49). Show that $\nu \ll \mu$.

1.4.2 True or false: If μ is a σ–finite measure and $\nu \ll \mu$, then ν is σ–finite.

1.4.3 Verify equation (1.54).

1.4.4 A lightbulb has an exponential lifetime T with mean λ, i.e. T is a r.v. with Lebesgue density $f_T(t) = \lambda^{-1} \exp[-t/\lambda] I_{(0,\infty)}(t)$. However, we only observe $Y = \min\{T, \tau\}$ where τ, the maximum time allowed for the experiment, is a positive constant. Find a σ–finite measure μ such that $\text{Law}[Y] \ll \mu$ and give the density $d\,\text{Law}[Y]/d\mu$.

1.4.5 Let P_λ be the *Poisson distribution* on $I\!N = \{0, 1, 2, \ldots\}$ given by

$$P_\lambda(\{n\}) \;=\; e^{-\lambda}\lambda^n/n!\;, \quad n \in I\!N \quad .$$

Show that $P_\lambda \ll P_1$ and find dP_λ/dP_1. Here, P_1 is the Poisson distribution with $\lambda = 1$.

1.4.6 Verify that $\nu_i \ll \mu$ for $i = 1, 2$ implies $\nu_1 + \nu_2 \ll \mu$.

1.4.7 Show that if f is not restricted to be nonnegative in Proposition 1.4.2 (a), then existence of either integral implies existence of the other and equality of the two.

1.4.8 Fill in the details of the proof of Proposition 1.4.2 (a).

1.4.9 Fill in the details of the proof of Proposition 1.4.2 (c).

1.4.10 Verify that Proposition 1.4.3 and Remark 1.4.1 extend to n measures and r.v.'s, respectively.

1.4.11 (a) Show the following converse of Remark 1.4.1: Suppose (X_1, X_2) have joint density w.r.t. a product measure $\mu_1 \times \mu_2$ which factors into the product of the marginals, as in $f(x_1, x_2) = f_1(x_1) f_2(x_2)$, then X_1 and X_2 are independent.

(b) Extend to more than two r.v.'s.

1.4.12 (a) Show that Proposition 1.4.4 has the following application in probability theory: Let X and Y be r.v.'s with joint distribution $\text{Law}[X, Y] \ll \mu_1 \times \mu_2$ where μ_1 and μ_2 are σ–finite. Write the joint density as

$$f_{XY}(x, y) \;=\; \frac{d\text{Law}[X, Y]}{d(\mu_1 \times \mu_2)}(x, y) \quad .$$

Show that $\text{Law}[X] \ll \mu_1$ and the marginal density is given by

$$f_X(x) \;=\; \int f_{XY}(x, y)\, d\mu_2(y) \quad .$$

(b) Let $\underline{X}$ be a random $(n + k)$–vector with Lebesgue density f. Let $\underline{Y} = (X_1, \ldots, X_n)$ be the random n-vector given by the first n components of $\underline{X}$. Show $\text{Law}[\underline{Y}] \ll m^n$ and find a formula for the Lebesgue density of Y.

1.4.13 Show that $\mathrm{supp}(\nu)$ is a closed set. Hint: Show the complement is open.

1.4.14 Let P be a Borel p.m. on $\mathbb{R}^n$. Show that $\mathrm{supp}(P)$ is the smallest closed set C such that $P(C) = 1$.

1.4.15 Suppose ν has Lebesgue density given by

$$f(x) = \begin{cases} 1 & \text{if } 0 < x < 1 \text{ or } x = 2, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find $\mathrm{supp}(\nu)$.
(b) Show that $\mathrm{supp}(\nu)$ is a proper subset of $\bar{S}$ given in Proposition 1.4.5.

1.4.16 Suppose $\nu \ll \mu$ σ–finite are Borel measures and μ is σ–finite. Let f be a version of $d\nu/d\mu$. Show that the following is also a version of $d\nu/d\mu$:

$$g(x) = \begin{cases} f(x) & \text{if } x \in \mathrm{supp}(\mu), \\ 0 & \text{otherwise.} \end{cases}$$

1.4.17 Suppose μ is any Borel measure on $\mathbb{R}$ with $\mathrm{supp}(\mu) \subset \mathbb{N}$. Show that $\mu \ll \#$, counting measure on $\mathbb{N}$, and find $d\mu/d\#$.

1.4.18 P is a Borel p.m. on $\mathbb{R}$ with c.d.f. F. Show that $\mathrm{supp}(P) = [a, b]$, a finite closed interval, iff $F(a - 0) = 1 - F(b) = 0$ and F is strictly increasing on $[a, b]$.

1.4.19 (a) Show that $\mathrm{supp}(\nu_1 \times \nu_2) = \mathrm{supp}(\nu_1) \times \mathrm{supp}(\nu_2)$ for σ–finite Borel measures ν_1 and ν_2.

(b) Suppose ν is a Borel measure on $\mathbb{R}^2$ but $\mathrm{supp}(\nu)$ is not a product set, i.e. not of the form of a Cartesian product $B_1 \times B_2$ for Borel sets $B_i \subset \mathbb{R}$. Show that ν cannot be a product measure.

1.4.20 Suppose $h : U \longrightarrow V$ is one to one and $h(U) = V$, where U and V are open sets in $\mathbb{R}^n$. Suppose $h(W)$ is an open set for each open $W \subset U$. Show that $h^{-1} : V \longrightarrow U$ is Borel measurable.

1.5 Conditional Expectation.

Suppose $X : (\Omega, \mathcal{F}, P) \longrightarrow (I\!\!R, \mathcal{B})$ is a r.v. and $Y : (\Omega, \mathcal{F}, P) \longrightarrow (\Lambda, \mathcal{G})$ is any random element. Knowing the value of Y tells us something about the particular outcome ω which occurred, and hence possibly also something about the value of X, i.e. $X(\omega)$. It is often of interest to find the "best predictor" or "estimator" of X based on the observed value of Y. By "based on the observed value of Y", we mean this predictor is a function of Y. For mathematical convenience, we take "best" to mean "mimimizes the mean squared prediction error," which is defined to be

$$\text{MSPE}(Z) \;=\; E[(Z - X)^2] \quad . \tag{1.57}$$

It will be necessary to assume $E[X^2] < \infty$, and then to restrict attention to Z's which are (Borel measurable) functions of Y and satisfy $E[Z^2] < \infty$. This latter requirement on Z is needed to guarantee that $E[(Z - X)^2] < \infty$.

Suppose Z_* minimizes MSPE, and let W be any other r.v. which is a function of Y with $E[W^2] < \infty$. Then consider the quadratic function of $t \in I\!\!R$ given by

$$\begin{aligned} M(t) &= \text{MSPE}(Z_* + tW) \\ &= E[\{(Z_* - X) + tW\}^2] \\ &= E[(Z_* - X)^2] \,+\, 2tE[(Z_* - X)W] \,+\, t^2E[W^2] \quad . \end{aligned}$$

Since Z_* minimizes MSPE, $M(t)$ has its minimum at $t = 0$. Since a quadratic function of t of the form $M(t) = at^2 + bt + c$ has its minimum at $t = -b/2a$, it follows that $b = 0$, i.e.,

$$E[W(Z_* - X)] \;=\; 0 \quad ,$$

that is

$$E[WX] \;=\; E[WZ_*] \quad . \tag{1.58}$$

Now W was an arbitrary function of Y with $E[W^2] < \infty$. In particular, (1.58) holds for all indicators $W = I_A(Y)$, where $A \subset \Lambda$ is $\mathcal{G}$–measurable, since indicators trivially have finite second moments. Conversely, if

$$E[I_A(Y)X] \;=\; E[I_A(Y)Z_*] \,, \quad \text{for all } A \in \mathcal{G} \quad , \tag{1.59}$$

then (1.58) holds for any $W = h(Y)$ with h real valued and W having finite second moment. This follows by taking a sequence of simple functions on $I\!\!R$ converging to h (Exercise 1.5.5). Since $I_A(Y(\omega)) = I_{Y^{-1}(A)}(\omega)$, and $Y^{-1}(A)$ is a generic element of $\sigma(Y)$, (1.59) is equivalent to

$$E[I_C X] \;=\; E[I_C Z_*] \,, \quad \text{for all } C \in \sigma(Y) \quad , \tag{1.60}$$

Now (1.60) provides us with a possibly useful characterization of the "best" predictor of X which is a function of Y. We will denote the Z_* which is a function of Y and satisfies (1.60) by $E[X|Y]$, referred to as the conditional expectation of

X given Y. There are several issues to deal with here, such as does such a Z_* exist, and is it unique? Is such an abstract characterization really useful? In the end, we will make the definition even more abstract by focusing attention on the σ–field rather than Y.

1.5.1 Characterization of Measurable Transformations of a Random Element.

Recall that we wanted $E[X|Y]$ to be a function of Y satisfying other conditions (namely (1.60)). The next result is a very useful characterization of the class of r.v.'s which are functions of Y.

Theorem 1.5.1 *Suppose $Y : (\Omega, \mathcal{F}) \longrightarrow (\Lambda, \mathcal{G})$ and $Z : (\Omega, \mathcal{F}) \longrightarrow (I\!\!R^n, \mathcal{B}_n)$. Then Z is $\sigma(Y)$-measurable if and only if there is a Borel function $h : (\Lambda, \mathcal{G}) \longrightarrow (I\!\!R^n, \mathcal{B}_n)$ such that $Z = h(Y)$.*

Remarks 1.5.1 To say "Z is $\sigma(Y)$–measurable" means $Z : (\Omega, \sigma(Y)) \longrightarrow (I\!\!R^n, \mathcal{B}_n)$, i.e. $\sigma(Z) = Z^{-1}(\mathcal{B}_n) \subset \sigma(Y)$. Note that $\sigma(Y)$ is a sub–σ–field of $\mathcal{F}$. The theorem may be summarized pictorially as follows:

$$\begin{tikzcd}
 & (\Lambda, \mathcal{G}) \arrow[dd, dashed, "h \text{ exists iff } \sigma(Z) \subset \sigma(Y)"] \\
(\Omega, \mathcal{F}) \arrow[ur, "Y"] \arrow[dr, "Z"'] & \\
 & (I\!\!R^n, \mathcal{B}_n)
\end{tikzcd}$$

□

Proof. Assume that $\sigma(Z) \subset \sigma(Y)$ and we will show the existence of such an h. Also, assume for now $n = 1$. We proceed in steps, as usual.

Step 1. If Z is a simple function, say

$$Z = \sum_{i=1}^{m} a_i I_{A_i} \quad ,$$

where the sets A_i are disjoint and coefficients a_i are distinct and nonzero, i.e.

$$a_i \neq a_j \text{ if } i \neq j \quad .$$

See Exercise 1.2.21. Then $A_i = Z^{-1}(\{a_i\}) \in \sigma(Z)$ and hence also $A_i \in \sigma(Y)$, $1 \leq i \leq m$, i.e. $A_i = Y^{-1}(C_i)$ for some $C_i \in \mathcal{G}$ since all $A_i \in \sigma(Y)$ are of this form by definition of $\sigma(Y)$. Put

$$h \;=\; \sum_{i=1}^{m} a_i I_{C_i} \quad .$$

Then

$$h(Y(\omega)) \;=\; \sum_{i=1}^{m} a_i I_{C_i}(Y(\omega)) \;=\; \sum_{i=1}^{m} a_i I_{Y^{-1}(C_i)}(\omega) \;=\; \sum_{i=1}^{m} a_i I_{A_i}(\omega) \quad .$$

This completes the proof if Z is a simple function.

Step 2. If Z is not simple, then there exist simple functions Z_n such that $Z_n(\omega) \to Z(\omega)$ for all $\omega \in \Omega$ by Proposition 1.2.8. By Step 1, each $Z_n = g_n(Y)$ for some $g_n : (\Lambda, \mathcal{G}) \longrightarrow (I\!R^n, \mathcal{B}_n)$. Now put $L = \{\lambda \in \Lambda : \lim_n g_n(\lambda) \text{ exists } \}$. Let $h_n = g_n I_L$. Clearly there is a function $h = \lim_n h_n$ (since if $\lambda \in L$ then $h_n(\lambda)$ $= g_n(\lambda)$ and the sequence of real numbers $g_n(\lambda)$ has a limit by definition of L, and if $\lambda \notin L$ then $h_n(\lambda) = 0$, which has the limit 0 as $n \to \infty$), and h is measurable by Proposition 1.2.1 (c).

We will show $Z(\omega) = h(Y(\omega))$ for all $\omega \in \Omega$. Note that $Y(\omega) \in L$ because $g_n(Y(\omega)) = Z_n(\omega) \to Z(\omega)$. so by definition of h_n, $h_n(Y(\omega)) = g_n(Y(\omega)) \to Z(\omega)$, but $h_n(Y(\omega)) \to h(Y(\omega))$ by definition of h, so $Z(\omega) = h(Y(\omega))$. This finishes Step 2.

Finally, to remove the restriction $n = 1$, use the result for $n = 1$ on each component of $\underline{Z} = (Z_1, ..., Z_n)$ and apply Theorem 1.3.5 to conclude that $\underline{Z}$ is $\sigma(Y)$ measurable when each component is $\sigma(Y)$ measurable.

To prove the converse, assuming $Z = h(Y) = h \circ Y$ for some $h : (\Lambda, \mathcal{G}) \longrightarrow (I\!R^n, \mathcal{B}_n)$, we have $Z^{-1}(B) = (h \circ Y)^{-1}(B) = Y^{-1}(h^{-1}(B))$ by Exercise 1.2.6. If $B \in \mathcal{B}_n$, then $h^{-1}(B) \in \mathcal{G}$, so it follows that $Y^{-1}(h^{-1}(B)) \in \sigma(Y)$. This shows $\sigma(Z) \subset \sigma(Y)$.

□

1.5.2 Formal Definition of Conditional Expectation.

We have shown that equation (1.60) is necessary for Z_* to be the “best” predictor of X based on Y. One can show that it is also sufficient (Exercise 1.5.10). Realizing that (1.60) characterizes the “best” such predictor when X has finite second moment allows us to generalize this notion of “best” predictor when X has only first moment. Also, notice that it only depends on the σ–field $\sigma(Y)$, so we can generalize the definition of conditional expectation to the situation where the given "information" is in the form of a σ–field (which may not often be the case in practical applications).

Definition 1.5.1 *Let $X : (\Omega, \mathcal{F}, P) \longrightarrow (I\!R, \mathcal{B})$ be a r.v. with $E[|X|] < \infty$, and suppose $\mathcal{G}$ is a sub–σ–field of $\mathcal{F}$. Then the* conditional expectation *of X given $\mathcal{G}$, denoted $E[X|\mathcal{G}]$, is the essentially unique r.v. Z_* satisfying*

(i) *Z_* is $\mathcal{G}$ measurable;*

(ii) *$\int_A Z_* dP = \int_A X dP$, for all $A \in \mathcal{G}$.*

Here, "essentially unique" means that if Z is any other r.v. on $(\Omega, \mathcal{F}, P)$ satisfying (i) and (ii), then $Z = Z_*$ a.s. Such Z's satisfying (i) and (ii) are called versions *of $E[X|\mathcal{G}]$.*

If Y is a random element on $(\Omega, \mathcal{F})$, then $E[X|Y] = E[X|\sigma(Y)]$. If $B \in \mathcal{F}$ then the conditional probability *of B given $\mathcal{G}$ is $P[B|\mathcal{G}] = E[I_B|\mathcal{G}]$.*

□

Remarks 1.5.2 (a) It is shown in the next theorem that such a Z_* exists and is essentially unique.

(b) Note that $E[X|\mathcal{G}]$ is a r.v., i.e. a mapping from $(\Omega, \mathcal{F}) \longrightarrow (I\!R, \mathcal{B})$. Thus, $E[X|\mathcal{G}](\omega) \in I\!R$ for each $\omega \in \Omega$.

(c) Since $E[|X|] < \infty$ we also have $E[|I_A X|] < \infty$ for all $A \in \mathcal{G}$ so the r.h.s. of (ii) is defined and is a finite real number.

(d) Note that (ii) is the same as (1.60) with $\mathcal{G} = \sigma(Y)$, which justifies in some sense writing $E[X|Y] = E[X|\sigma(Y)]$. From a probabilistic point of view, one can say that $\sigma(Y)$ "contains the information in Y" useful for prediction of any r.v. X. Note that from the observed value $Y(\omega)$ one can only determine whether or not $\omega \in A$ if $A \in \sigma(Y)$.

We try to clarify this even more with the following. If $\sigma(W) = \sigma(Y)$, then we know from Theorem 1.5.1 that $W = h(Y)$ for some function h, and $Y = g(W)$ for some function g, so $g = h^{-1}$. Thus, if we know the value of $W(\omega)$, then we know $Y(\omega)$ and conversely, so it is reasonable to say W and Y "contain the same information". If it is only true that $W = h(Y)$ for some function h, then knowing Y we can determine W, but not conversely in general, so it is reasonable that Y "contains more information than" W, which means that $\sigma(W) \subset \sigma(Y)$ as follows again from Theorem 1.5.1.

□

Theorem 1.5.2 (Existence and uniqueness of conditional expectation.) *There is an essentially unique r.v. Z_* satisfying (i) and (ii) of Definition 1.5.1.*

Proof. First assume $X \geq 0$. Define a measure ν on $(\Omega, \mathcal{F})$ by

$$\nu(A) \;=\; \int_A X \, dP \quad , \qquad \text{for all } A \in \mathcal{F} \; .$$

Note that $\nu \ll P$ and $d\nu/dP = X$, a.s. Let ν_0 and P_0 denote the restrictions of ν and P to $\mathcal{G}$, i.e. ν_0 is the measure on $(\Omega, \mathcal{G})$ given by $\nu_0(A) = \nu(A)$ for all $A \in \mathcal{G}$. Then we still have $\nu_0 \ll P_0$, but not necessarily that $d\nu_0/dP_0 = X$ since X is not necessarily $\mathcal{G}$-measurable, i.e. we may not have $\sigma(X) \subset \mathcal{G}$. However, by the Radon-Nikodym theorem (note that P_0 is trivially σ–finite) we have that there is a r.v. $Z_* = d\nu_0/dP_0$, P_0-a.s. such that Z_* is $\mathcal{G}$-measurable (i.e. property (i) of the definition holds) and

$$\nu_0(A) \;=\; \int_A Z_* \, dP_0 \quad , \qquad \text{for all } A \in \mathcal{G} \; .$$

Since $\nu_0(A) = \nu(A) = \int_A X dP$, we have

$$\int_A X \, dP \;=\; \int_A Z_* \, dP_0 \quad , \qquad \text{for all } A \in \mathcal{G} \; . \tag{1.61}$$

Now we claim that for any r.v. W on $(\Omega, \mathcal{G}, P_0)$, $\int W dP_0 = \int W dP$. (Note that W is automatically a r.v. on $(\Omega, \mathcal{F}, P)$.) This is certainly true if W is an indicator by definition of P_0, and then it follows immediately for simple functions by linearity of integrals. For $W \geq 0$, consider a sequence of $\mathcal{G}$-measurable simple functions $0 \leq \phi_n \uparrow W$ as in Proposition 1.2.8 and apply monotone convergence. Finally, the general case (which we do not actually need here) follows from linearity and the decomposition of W into its positive and negative parts.

Hence, from (1.61) we have

$$\int_A X \, dP \;=\; \int_A Z_* \, dP \quad , \qquad \text{for all } A \in \mathcal{G} \; , \tag{1.62}$$

which is property (ii).

If Z' is any other r.v. satisfying (i) and (ii), then $Z' = d\nu_0/dP_0 = Z_*$, P_0-a.s. by the essential uniqueness of Radon-Nikodym derivatives. Note that P_0-a.s. implies P-a.s. since a P_0-null set is just a P-null set which happens to belong to $\mathcal{G}$.

If we drop the restriction that $X \geq 0$ but require $E[|X|] < \infty$, then apply the previous argument to X_+ and X_- to obtain essentially unique r.v.'s Z_{*+} and Z_{*-} which are $\mathcal{G}$-measurable and satisfy

$$\int_A X_+ dP \;=\; \int_A Z_{*+} dP \quad , \quad \int_A X_- dP \;=\; \int_A Z_{*-} dP \quad , \qquad \text{for all } A \in \mathcal{G} \; .$$

We claim Z_{*+} and Z_{*-} are both finite a.s. so that the r.v. $Z_* = Z_{*+} - Z_{*-}$ is defined a.s. (i.e. it can be of the form $\infty - \infty$ only on a null set, and we may define it arbitrarily there). Now X_+ and X_- are both finite a.s. (Exercise 1.2.31), and if say $A = [Z_{*+} = \infty]$ satisfied $P(A) > 0$, then since $A = Z_{*+}^{-1}(\{\infty\}) \in \mathcal{G}$, $\int_A X_+ dP = \int_A Z_{*+} dP = \infty$. However, since X is integrable, $\int_A X_+ dP \leq \int X_+ dP < \infty$, a contradiction. This establishes the claim for Z_{*+} and the claim that $Z_{*-} < \infty$ a.s. follows similarly.

Verification of properties (i) and (ii) is easy. If Z' is any other r.v. satisfying (i) and (ii), then let $D = Z_* - Z'$. Then D is $\mathcal{G}$–measurable by Proposition 1.2.1 (b), so $A = [D \geq 0]$ is in $\mathcal{G}$. Since both Z_* and Z' satisfy (ii)

$$\int_\Omega I_A D \, dP \; = \; \int_A Z_* \, dP \; - \; \int_A Z' \, dP \; = \; \int_A X \, dP \; - \; \int_A X \, dP \; = \; 0.$$

However, $I_A D$ is a nonnegative function, so by Proposition 1.2.6 (b), $I_A D = 0$, a.s. A similar argument shows $I_{A^c} D = 0$, a.s., and hence $Z_* = Z'$, a.s., which completes the proof.

□

The proof may also be found in Billingsley, p. 466 ff.

Now we introduce another object sometimes known as the "conditional expectation." Let $Y : (\Omega, \mathcal{F}, P) \longrightarrow (\Lambda, \mathcal{G})$ be any random element, and let X be an integrable r.v. Now

$$Z(\omega) \; = \; E[X|Y](\omega) \quad \omega \in \Omega \, ,$$

is a r.v. on Ω which is $\sigma(Y)$-measurable by definition. Hence, by Theorem 1.5.1 there is a function

$$h : (\Lambda, \mathcal{G}) \longrightarrow (I\!R, \mathcal{B})$$

such that $Z(\omega) = h(Y(\omega)) = (h \circ Y)(\omega)$. Furthermore, this function is Law[Y]-essentially unique in the sense that $Z = h' \circ Y$ for some other $h' : (\Lambda, \mathcal{G}) \longrightarrow (I\!R, \mathcal{B})$ implies that $h' = h$ Law[Y]-a.s., i.e. $P[h(Y) \; = \; h'(Y)] = 1$. Any such version is defined to be *the conditional expectation of* X *given* $Y = y$, and denoted

$$E[X|Y = y] \; = \; h(y) \; , \quad y \in \Lambda \; .$$

The following picture may help the student keep matters clear:

$$\begin{tikzcd}
 & (\Lambda, \mathcal{G}) \arrow[dd, "{E[X|Y = \,\cdot\,]}"] \\
(\Omega, \mathcal{F}) \arrow[ur, "Y"] \arrow[dr, "{E[X|Y]}"'] & \\
 & (\mathbb{R}, \mathcal{B})
\end{tikzcd}$$

The notations here are very confusing for many students, so we will try to explain some of the subtleties. One difficulty is that $E[X|Y = y]$ is a function of $y \in \Lambda$ in our setup, and the argument of the function y does not appear in a convenient place. Indeed, in the defining equation above $E[X|Y = y] = h(y)$ where h is the function such that $h(Y) = E[X|Y]$, if we substitute the random object Y for y we obtain the seemingly nonsensical "$E[X|Y] = E[X|Y = Y]$." The following may be a little clearer:

$$E[X|Y](\omega) \;=\; E[X|Y = Y(\omega)] \quad . \tag{1.63}$$

The argument of the function $E[X|Y = \,\cdot\,]$ is whatever appears on the r.h.s. of the equals sign "=" after the conditioning bar "|". We do not call $E[X|Y = \cdot]$ a random variable in general since it is not a function defined on the underlying probability space $(\Omega, \mathcal{F}, P)$, although it is a function on the probability space $(\Lambda, \mathcal{G}, \mathrm{Law}[Y])$, so technically we could call it a random variable.

1.5.3 Examples of Conditional Expectations.

The definition of $E[X|\mathcal{G}]$ is very unsatisfactory from an intuitive point of view, although it turns out to be very convenient from a formal mathematical point of view. In order to make it more appealing intuitively, we shall verify that it gives the "right answer" in a number of circumstances with which the student is already familiar.

Proposition 1.5.3 *Suppose A_1, A_2, ..., A_n are events which partition Ω (i.e. the A_i are mutually exclusive and $\Omega = \bigcup_{i=1}^n A_i$). Suppose $P(A_i) > 0$ for each i and a_1, a_2, ..., a_n are distinct real numbers. Let*

$$Y \;=\; \sum_{i=1}^{n} a_i I_{A_i}$$

be a simple r.v. If X is an integrable r.v., then

$$E[X|Y] \;=\; \sum_{i=1}^{n} \frac{\int_{A_i} X dP}{P(A_i)} I_{A_i} \quad , \qquad a.s. \tag{1.64}$$

Remarks 1.5.3 Consider the elementary case $n = 2$ and $X = I_B$ for some event B. Write $A = A_1$ and $A^c = A_2$. The values of a_1 and a_2 are irrelevant, as long as they are distinct, since any such Y contains the same “information”, namely the σ–field $\sigma(Y) = \{\emptyset,\, A,\, A^c,\, \Omega\}$ (see the example after Definition 1.2.2). We may take $Y = I_A$ for simpicity. Then, according to (1.64),

$$E[X|Y] \;=\; P[B|Y] \;=\; \frac{\int_A I_B dP}{P(A)} I_A + \frac{\int_{A^c} I_B dP}{P(A^c)} I_{A^c} \quad , \quad \text{a.s.} \tag{1.65}$$

That is, almost surely

$$P[B|Y](\omega) \;=\; \begin{cases} P(A \cap B)/P(A) & \text{if } \omega \in A, \\ \\ P(A^c \cap B)/P(A^c) & \text{if } \omega \in A^c. \end{cases}$$

Note that for $\omega \in A$, $P[B|I_A](\omega) = P[B|A] = P(A \cap B)/P(A)$, with probability 1, where $P[B|A]$ denotes the “classical” or “elementary” conditional probability of B given A. Similarly, for $\omega \in A^c$, $P[B|I_A](\omega) = P[B|A^c]$ a.s. Thus, we have

$$P[B|I_A] \;=\; P[B|A] I_A \;+\; P[B|A^c] I_{A^c}.$$

Note that we have mixed meanings for conditional probability in the last display. The l.h.s. is the “sophisticated” type of conditional probability defined in Definition 1.5.1, whereas both conditional probabilities on the r.h.s. are of the elementary variety. It will always be clear when we intend elementary conditional probability (which is a fixed number) rather than our more sophisticated kind (which is a random variable) since the second member of the conditional probability operator will be a set in the case of elementary conditional probability but will be a random variable or a σ-field for the more sophisticated variety.

□

Remarks 1.5.4 We may also express the result of Proposition 1.5.3 in terms of the “other” kind of conditional expectation. The reader should be able to check that

$$E[X|Y = y] \;=\; \sum_{i=1}^{n} \frac{\int_{A_i} X dP}{P(A_i)} I_{\{a_i\}}(y) \quad , \quad \text{Law}[Y] - \text{a.s.}$$

□

Proof. Let Z denote the proposed $E[X|Y]$ on the r.h.s. of (1.64). Since $A_i = Y^{-1}(\{a_i\})$, it follows that Z is $\sigma(Y)$-measurable. In fact, one can show that $\sigma(Y)$ is the collection of all unions of the A_i. For instance, if $B \in \mathcal{B}$, then

$$Y^{-1}(B) \;=\; \bigcup_{\{i:a_i \in B\}} A_i \quad .$$

See also Exercises 1.2.21 and 1.2.22. Hence, if $A \in \sigma(Y)$, say $A = Y^{-1}(B)$ for $B \in \mathcal{B}$, then

$$\int_A X\,dP \;=\; \int_{Y^{-1}(B)} X\,dP \;=\; \sum_{\{i:a_i \in B\}} \int_{A_i} X\,dP \;. \tag{1.66}$$

Now,

$$\int_A Z\,dP \;=\; \sum_{\{i:a_i \in B\}} \int_{A_i} \sum_{j=1}^{n} \frac{\int_{A_j} X(\omega_1) dP(\omega_1)}{P(A_j)} I_{A_j}(\omega) dP(\omega)$$

$$= \sum_{\{i:a_i \in B\}} \sum_{j=1}^{n} \frac{\int_{A_j} X(\omega_1) dP(\omega_1)}{P(A_j)} \int_{A_i} I_{A_j}(\omega) dP(\omega) \quad .$$

Note in the last expression that when i is fixed in the outer summation, then $\int_{A_i} I_{A_j} dP$ is nonzero only when $j = i$ since otherwise A_i and A_j are disjoint. If $i = j$ then this integral is $P(A_i)$. Hence,

$$\int_A Z\,dP \;=\; \sum_{\{i:a_i \in B\}} \frac{\int_{A_i} X dP}{P(A_i)} \int_{A_i} I_{A_i}(\omega) dP(\omega) \;=\; \sum_{\{i:a_i \in B\}} \int_{A_i} X\,dP \quad . \tag{1.67}$$

This shows the proposed $E[X|Y]$ satisfies property (ii) of the definition by (1.66) and (1.67).

□

One virtue of the abstract definition of conditional expectation is that it allows us to make sense of $P[B|X]$ even when $P[X = x] = 0$ for any single value x. The next result makes this clearer.

Proposition 1.5.4 *Suppose $X : (\Omega, \mathcal{F}, P) \longrightarrow (\Lambda_1, \mathcal{G}_1)$ and $Y : (\Omega, \mathcal{F}, P) \longrightarrow (\Lambda_2, \mathcal{G}_2)$ are random elements and μ_i is a σ-finite measure on $(\Lambda_i, \mathcal{G}_i)$ for $i = 1, 2$ such that $Law[X, Y] \ll \mu_1 \times \mu_2$. Let $f(x, y)$ denote the corresponding joint density. Let $g(x, y)$ be any Borel function $\Lambda_1 \times \Lambda_2 \longrightarrow I\!R$ such that $E|g(X, Y)| < \infty$. Then*

$$E[g(X,Y)|Y] \;=\; \frac{\int_{\Lambda_1} g(x,Y)\, f(x,Y)\, d\mu_1(x)}{\int_{\Lambda_1} f(x,Y)\, d\mu_1(x)} \;, a.s. \tag{1.68}$$

Remarks 1.5.5 Note that the denominator is $f_Y(Y)$, which is the marginal density of Y w.r.t. μ_2. See Exercise 1.4.12. Define the *conditional density* of X given Y by

$$f_{X|Y}(x|y) \;=\; \frac{f_{XY}(x,y)}{f_Y(y)} \quad . \tag{1.69}$$

Note that for fixed y this is a density w.r.t. μ_1 for a probability measure on Λ_1, where X is a random element taking values in Λ_1. Then (1.68) may be rewritten as

$$E[g(X,Y)|Y] \;=\; \int_{\Lambda_1} g(x,Y)\, f_{X|Y}(x|Y)\, d\mu_1(x) \quad ,$$

or in terms of the "other" conditional expectation,

$$E[g(X,Y)|Y=y] \;=\; \int_{\Lambda_1} g(x,y)\, f_{X|Y}(x|y)\, d\mu_1(x) \quad . \tag{1.70}$$

□

Proof. It follows from Fubini's theorem that both of the functions

$$\int_{\Lambda_1} g(x,y)\, f(x,y)\, d\mu_1(x) \text{ and } \int_{\Lambda_1} f(x,y)\, d\mu_1(x)$$

are measurable functions of y, and the second is positive Law$[Y]$-a.s. (the set of y values where it is 0 has Law$[Y]$ measure 0). If we define $h(y)$ to be the quotient of the first over the second (i.e. $h(Y)$ is the function of Y on the r.h.s. of (1.68)), then $h(y)$ is defined Law$[Y]$-a.s. and is measurable from $\Lambda_2 \longrightarrow I\!R$. As the r.h.s. of (1.68) equals $h(Y) = h \circ Y$, it follows that the r.h.s. is $\sigma(Y)$-measurable. This is property (i) of Definition 1.5.1.

Now we check the second property of Definition 1.5.1. Let $B \in \mathcal{B}_n$ so $Y^{-1}(B)$ is a generic element of $\sigma(Y)$. Then

$$\int_{Y^{-1}(B)} h(Y)\, dP \;=\; \int_B h(y)\, dP_Y(y) \;=\; \int_B h(y) f_Y(y)\, d\mu_2(y)$$

$$= \int_B \frac{\int g(x,y) f(x,y)\, d\mu_1(x)}{f_Y(y)} f_Y(y)\, d\mu_2(y) \;=\; \int_B \int g(x,y) f(x,y)\, d\mu_1(x) d\mu_2(y)$$

$$= \int_{\Lambda_1 \times B} g(x,y) f(x,y)\, d(\mu_1 \times \mu_2)(x,y) \;=\; \int_{Y^{-1}(B)} g(X,Y)\, dP \quad .$$

This proves the theorem.

□

1.5.4 Conditional Distributions.

It would seem that conditional expectation is difficult enough, and might hope that further complexity could be avoided. However, we shall have much need for something even more complicated.

Definition 1.5.2 *Let* $X : (\Omega, \mathcal{F}, P) \longrightarrow (\Lambda_1, \mathcal{G}_1)$ *and* $Y : (\Omega, \mathcal{F}, P) \longrightarrow (\Lambda_2, \mathcal{G}_2)$ *be random elements. A* family of regular conditional (probability) distribution(s) for X given $Y = y$*, or more simply called* the conditional distribution of X given $Y = y$*, is a function* $p : \mathcal{G}_1 \times \Lambda \longrightarrow [0, 1]$ *satisfying*

(i) *for all* $B \in \mathcal{G}_1$*,* $p(B, y) = P[X \in B|Y = y]$ *for* P_Y*-almost all* $y \in \Lambda$*, i.e.* $p(B, \cdot)$ *is a version of* $P[X \in B|Y = \cdot]$ *for fixed* $B \in \mathcal{G}_1$*.*

(ii) $p(\cdot, y)$ *is a p.m. on* $(\Lambda_1, \mathcal{G}_1)$ *for all* $y \in \Lambda_2$*.*

When such a $p(B, y)$ *exists, we shall write it as* $P_{X|Y}(B|Y = y)$*, or we will denote the p.m.* $P_{X|Y}(\cdot|Y = y)$ *by Law*$[X|Y = y]$*.*

□

Proposition 1.5.5 *Suppose that the assumptions of Proposition 1.5.4 hold. Then we have*

(i) *the family of regular conditional distributions Law*$[X|Y = y]$ *exists;*

(i) *Law*$[X|Y = y] \ll \mu_1$ *for Law*$[Y]$*–almost all values of* y*;*

(iii) *the Radon-Nikodym derivatives are given by*

$$\frac{dLaw[X|Y = y]}{d\mu_1}(x) \;=\; f_{X|Y}(x|y), \quad \mu_1 \times \mu_2 - a.e.,$$

where $f_{X|Y}(x|y)$ *is the conditional density given in Proposition 1.5.4.*

Proof. From the proof of Proposition 1.5.4, for all $B \in \mathcal{G}_1$,

$$P[X \in B|Y = y] \;=\; \int I_B(x) f_{X|Y}(x|y)\, d\mu_1(x) \;.$$

This verifies (i) of Definition 1.5.2. Condition (ii) of the definition follows since $f_{X|Y}(x|y)$ is a probability density w.r.t. $d\mu_1(x)$ for each fixed $y \in \Lambda_2$, i.e. $f_{X|Y}(x|y) \geq 0$ for all x and y, and for all y, $\int f_{X|Y}(x|y)\, d\mu_1(x) = 1$.

□

Remarks 1.5.6 The reader may find the definition and previous result very puzzling. After all, is it not obvious that conditional probability distributions exist? The answer is, "No," but it is also not obvious why they should not automatically exist. To explain, suppose $(\Omega, \mathcal{F}, P)$ is a probability space and $\mathcal{G}$ is a sub–σ–field of $\mathcal{F}$. Then for each event $A \in \mathcal{F}$, the conditional probability $P[A|\mathcal{G}] = E[I_A|\mathcal{G}]$ is an *almost surely* uniquely defined r.v. Fix $\omega \in \Omega$. Does it follow that $P[A|\mathcal{G}](\omega)$ is a probability measure when considered as a function of the event A? Given that $P[A|\mathcal{G}](\cdot)$ may be modified arbitrarily on P-null sets (as long as it is done in a $\mathcal{G}$-measurable way), clearly we may not use any version of the family of r.v.'s $\{P[A|\mathcal{G}](\cdot) : A \in \mathcal{F}\}$ and obtain a family of probability measures $\{P[\cdot|\mathcal{G}](\omega) : \omega \in \Omega\}$. In general, such versions of $P[A|\mathcal{G}](\cdot)$ may not exist. Like a number of issues in measure theory, (e.g. the existence of subsets of $I\!R$ which are not Borel measurable) the nonexistence of conditional probability distributions is a technical detail which is of little importance in statistics. The next theorem shows that conditional *distributions* exist for the settings we shall encounter in this book. For further discussion of the difficulties involved with obtaining a family of conditional probability distributions (including counterexamples wherein they don't exist), see Ash or Brieman. Exercise 33.13, p. 464 of Billingsley provides a specific example.

□

Theorem 1.5.6 *Suppose $Y : (\Omega, \mathcal{F}, P) \longrightarrow (\Lambda, \mathcal{G})$ is a random element and $\underline{X}$ is a random n–vector on $(\Omega, \mathcal{F}, P)$. Then there exists a family $\langle P_{\underline{X}|Y}(B|Y = y) : y \in \Lambda,\ B \in \mathcal{B}_n\rangle$ such that*

(i) *for all $B \in \mathcal{B}_n$, $P_{\underline{X}|Y}(B|Y = y) = P[\underline{X} \in B|Y = y]$ for P_Y-almost all $y \in \Lambda$, i.e. $P_{\underline{X}|Y}(B|Y = y)$ is a version of $P[\underline{X} \in B|Y = y]$.*

(ii) *$P_{\underline{X}|Y}(\cdot|Y = y)$ is a Borel p.m. on $I\!R^n$ for all $y \in \Lambda$.*

Furthermore, if $E[|h(\underline{X}, Y)|] < \infty$, then

$$E[h(\underline{X}, Y)|Y = y] \ = \ \int_{R^n} h(\underline{x}, y)\, dP_{\underline{X}|Y}(\underline{x}|Y = y) \ . \tag{1.71}$$

□

The proof of this theorem may be found in Breiman. It also follows from Theorems 33.3, p. 460, and Theorem 34.5, p. 471 of Billingsley. Comparison of Proposition 1.5.5 and Theorem 1.5.6 demonstrates the usual situation in statistics: in spite of the difficulty of proving a general result like Theorem 1.5.6, with a few more "concrete" assumptions as in Proposition 1.5.5, one can "barehandedly" construct the conditional distribution.

Remarks 1.5.7 (a) Now we outline the usual procedure for rigorously "deriving" a conditional distribution. One typically has a "candidate" for the conditional distribution $P_{X|Y}$, and it is necessary to verify that it satisfies the defining properties. The "candidate" comes from previous experience with elementary conditional probabilities or conditional densities, or from intuition. A candidate for $P_{X|Y}$ must be a function of the form $p(B, y)$ where B varies over measurable sets in the range of X and y varies over elements in the range of Y. Then there are basically three conditions that must be verified:

(1) $\forall y \in \Lambda$, $p(\cdot, y)$ is a p.m. on $(\Lambda_1, \mathcal{G}_1)$;

(2) $\forall B \in \mathcal{G}_1$, $p(B, \cdot)$ is measurable $(\Lambda_2, \mathcal{G}_2) \longrightarrow (I\!R, \mathcal{B})$;

(3) $\forall A \in \mathcal{G}_2$ and $\forall B \in \mathcal{G}_1$,

$$P\,[\,Y \in A \,\&\, X \in B\,] \;=\; \int_A p(B, y)\, d\mathrm{Law}[Y](y) \quad .$$

Now condition (1) here is simply a restatement of condition (ii) in Definition 1.5.2, and conditions (2) and (3) together amount to condition (i) in Definition 1.5.2. Note that (2) means that $p(A, Y)$ is a $\sigma(Y)$ measurable r.v. as required in item (i) of the definition of conditional expectation (Definition 1.5.1). We will show that (3) here is simply a restatement of the integral condition in item (ii) of Definition 1.5.1. Now according to that condition in Definition 1.5.1, we should have

$$\forall A \in \mathcal{G}_2, \quad \int_{[Y \in A]} p(B, Y(\omega))\, dP(\omega) \;=\; \int_{[Y \in A]} I_{[X \in B]}(\omega)\, dP(\omega).$$

To explain, $[Y \in A]$, which is another way of denoting $\{\omega \in \Omega : Y(\omega) \in A\} = Y^{-1}(A)$ is a generic element of $\sigma(Y)$. Also, recall that $P[C|\mathcal{G}] = E[I_C|\mathcal{G}]$, so we use an indicator for "X" in Definition 1.5.1. Now

$$\begin{aligned}
\int_{[Y \in A]} I_{[X \in B]}\, dP \quad &= \quad \int I_{[Y \in A]} I_{[X \in B]}\, dP \\
&= \quad \int I_{[Y \in A] \cap [X \in B]}\, dP \\
&= \quad P\,[\,Y \in A \,\&\, \underline{X} \in B\,]\,.
\end{aligned}$$

Also, by the Law of the Unconscious Statistician,

$$\int_{[Y \in A]} p(B, Y(\omega))\, dP(\omega) \;=\; \int_A p(B, y)\, d\mathrm{Law}[Y](y) \quad .$$

This completes the verification that (3) here is the same as condition (ii) in Definition 1.5.1.

Condition (1) here is usually easy to check. We generally regard condition (2) as automatic – any function $p(B, y)$ that you can "write down" (e.g. as a formula in y) is measurable. So, any difficulties usually come in verification of condition (3).

(b) Note that x is the only variable of integration in (1.71), and both sides are functions of y. This should be clear because "x" occupies the site in the function where the measurable set would go when evaluating its measure. The notation is not entirely desirable, and it is perhaps preferable to write

$$E[h(X,Y)|Y=y] \;=\; \int_{\Lambda_1} h(x,y)\, P_{X|Y}(dx|Y=y)\ . \tag{1.72}$$

This makes clearer the variable of integration, and it is more consistent perhaps that a "differential set" dx should occupy the set argument than a regular variable. See equation (1.26) and the remarks there. However, putting the "d" in front of the measure is much more convenient for the mnemonics of Radon-Nikodym derivatives, which is why we chose this convention. We shall use the convention as in (1.72) for clarity on occasion.

□

1.5.5 Results on Conditional Expectation.

Theorem 1.5.7 (Basic Properties of Conditional Expectation.) *Let X, X_1, and X_2 be integrable r.v.'s on $(\Omega, \mathcal{F}, P)$, and let $\mathcal{G}$ be a fixed sub–σ–field of $\mathcal{F}$.*

(a) If $X = k$ a.s., k a constant, then $E[X|\mathcal{G}] = k$ a.s.

(b) If $X_1 \leq X_2$ a.s., then $E[X_1|\mathcal{G}] \leq E[X_2|\mathcal{G}]$ a.s.

(c) If a_1, $a_2 \in I\!R$, then

$$E[a_1X_1 + a_2X_2|\mathcal{G}] \;=\; a_1E[X_1|\mathcal{G}] \,+\, a_2E[X_2|\mathcal{G}] \quad , \ a.s.$$

(d) (The Law of Total Expectation.) $E[E[X|\mathcal{G}]] = E[X]$.

(e) $E[X|\{\emptyset, \Omega\}] = E[X]$.

(f) If $\sigma(X) \subset \mathcal{G}$, then $E[X|\mathcal{G}] = X$ a.s.

(g) (Law of Successive Conditioning.) If $\mathcal{G}_\infty$ is a sub–σ–field of $\mathcal{G}$, then

$$E[E[X|\mathcal{G}]|\mathcal{G}_\infty] \;=\; E[E[X|\mathcal{G}_\infty]|\mathcal{G}] \;=\; E[X|\mathcal{G}_\infty]\, a.s. \quad .$$

(h) If $\sigma(X_1) \subset \mathcal{G}$ and $E|X_1X_2| < \infty$, then $E[X_1X_2|\mathcal{G}] = X_1E[X_2|\mathcal{G}]$ a.s.

Partial Proofs and Remarks. Part (a) follows from (f). For (b), it suffices to show that $X \geq 0$ a.s. implies $E[X|\mathcal{G}] \geq 0$ a.s. by taking $X = X_2 - X_1$, but this was shown in the proof of Theorem 1.5.2. In (c), we show that $E[aX|\mathcal{G}] = aE[X|\mathcal{G}]$ a.s. The proof of additivity (that $E[X_1 + X_2|\mathcal{G}] = E[X_1|\mathcal{G}] + E[X_2|\mathcal{G}]$) is left as Exercise 1.5.6. Clearly $aE[X|\mathcal{G}]$ is $\mathcal{G}$-measurable, and for $A \in \mathcal{G}$,

$$\int_A aE[X|\mathcal{G}]\, dP \;=\; a\int_A E[X|\mathcal{G}]\, dP \;=\; a\int_A X\, dP \;=\; \int_A (aX)\, dP$$

which proves the result. Part (d) follows by taking $A = \Omega \in \mathcal{G}$ in the part (ii) of Definition 1.5.1. Parts (e), (f), and (g) are likewise elementary.

Note that (h) says that if X_1 is $\mathcal{G}$-measurable, then we may treat it the same as a constant when computing $E[X_1 X_2|\mathcal{G}]$. Clearly $X_1 E[X_2|\mathcal{G}]$ is $\mathcal{G}$-measurable. We will verify property (ii) of the definition only when X_1 is an $\mathcal{G}$-measurable simple function, say $X_1 = \sum a_i I_{A_i}$ for $A_i \in \mathcal{G}$. In this case, for $A \in \mathcal{G}$,

$$\int_A X_1 E[X_2|\mathcal{G}]\, dP \;=\; \sum a_i \int_{A\cap A_i} E[X_2|\mathcal{G}]\, dP \;=\; \sum a_i \int_{A\cap A_i} X_2\, dP \;=\; \int_A X_1 X_2\, dP\ .$$

The second equality follows since $A \cap A_i \in \mathcal{G}$.

□

Theorem 1.5.8 (Convergence Theorems for Conditional Expectation.) *Let X, X_1, X_2, ... be integrable r.v.'s on $(\Omega, \mathcal{F}, P)$ and let $\mathcal{G}$ be a sub-σ–field of $\mathcal{F}$.*

(a) (Monotone Convergence Theorem.) If $0 \le X_i \uparrow X$ a.s. then

$$E[X_i|\mathcal{G}] \uparrow E[X|\mathcal{G}] \text{ a.s.}$$

(b) (Dominated Convergence Theorem.) Suppose there is an integrable r.v. Y such that $X_i \le Y$ a.s. for all i, and suppose that $X_i \to X$ a.s. Then

$$E[X_i|\mathcal{G}] \;\to\; E[X|\mathcal{G}] \text{ a.s.}$$

Partial Proof. (a) Clearly $\lim E[X_i|\mathcal{G}]$ is a $\mathcal{G}$-measurable r.v. by Proposition 1.2.1 (c). If $A \in \mathcal{G}$ then $I_A E[X_i|\mathcal{G}]$ is a nonnegative increasing sequence of functions so by two applications of the ordinary Monotone Convergence Theorem,

$$\int_A \lim\, E[X_i|\mathcal{G}]\, dP \;=\; \lim \int_A E[X_i|\mathcal{G}]\, dP \;=\; \lim \int_A X_i\, dP \;=\; \int_A X\, dP \quad .$$

The result follows from the essential uniqueness of conditional expectations.

□

The foregoing results may be found in Billinsgley, pp. 468-470.

Theorem 1.5.9 (Conditional Expectation and Independence.) *Suppose X is an integrable r.v. and Y_1 and Y_2 are random vectors with (X, Y_1) independent of Y_2. Then*

$$E[X|Y_1, Y_2] \;=\; E[X|Y_1] \text{ a.s.}$$

In particular,

$$E[X|Y_2] \;=\; E[X] \text{ a.s.}$$

□

The proof is given in the standard measure theoretic probability texts.

Remarks 1.5.8 From an intuitive point of view, Y_2 provides no information about X if they are independent, so it is reasonable that the conditional expectation of X given Y_2 not depend on Y_2.

□

Proposition 1.5.5 and Theorem 1.5.6 were concerned with the construction of conditional distributions from joint distributions. We will frequently be interested in the "reverse" construction, i.e. we will be given the conditional distribution $P_{\underline{X}|\underline{Y}}$ and the marginal $P_{\underline{Y}}$, and we will want to construct the joint distribution. The following is proved in Theorem 2.6.2 of Ash. It appears as problem 18.25 on p. 247 of Billingsley.

Theorem 1.5.10 (Two Stage Experiment Theorem.) *Let $(\Lambda, \mathcal{G})$ be a measurable space and suppose $p : \mathcal{B}_n \times \Lambda \longrightarrow I\!\!R$ satisfies the following:*

(i) *$p(B, \cdot)$ is Borel measurable for each fixed $B \in \mathcal{B}_n$;*

(ii) *$p(\cdot, y)$ is a Borel p.m. for each fixed $y \in \Lambda$.*

Let ν be any p.m. on $(\Lambda, \mathcal{G})$. Then there is a unique p.m. P on $(I\!\!R^n \times \Lambda, \mathcal{B}_n \times \mathcal{G})$ such that

$$P(B \times C) \ = \ \int_C p(B, y)\, d\nu(y) \quad , \textit{ for all } B \in \mathcal{B}_n \textit{ and } C \in \mathcal{G}. \tag{1.73}$$

Furthermore, if $\underline{X}(\underline{x}, y) = \underline{x}$ and $Y(\underline{x}, y) = y$ define the random coordinate elements on $I\!\!R^n \times \Lambda$, then $Law[Y] = \nu$ and $Law[\underline{X}|Y = y] = p(\cdot, y)$.

□

Remarks 1.5.9 The reason for the name of the theorem is as follows. If Y is selected in stage 1 according to ν, and given $Y = y$, we then select $\underline{X}$ at stage 2 according to the distribution $p(\cdot, y)$, then the combined two stage experiment results in a random element $(\underline{X}, Y)$ with the joint distribution indicated. However "obvious" the result appears when stated this way, the proof is nontrivial. In fact, it is a nontrivial extension of the product measure theorem.

□

As in the case of the existence of conditional distributions, the two stage experiment theorem is "easy" when one has densities.

Proposition 1.5.11 *Let $(\Omega, \mathcal{F}, \mu)$ and $(\Lambda, \mathcal{G}, \nu)$ be σ-finite measure spaces. Let $g : (\Lambda, \mathcal{G}) \longrightarrow (I\!R, \mathcal{B})$ be a probability density function w.r.t. ν. Suppose $h : \Omega \times \Lambda \longrightarrow I\!R$ is Borel measurable (under the product σ-field on $\Omega \times \Lambda$) and for each fixed $y \in \Lambda$, $h(\cdot, y)$ is a probability density function on Ω w.r.t. μ. Then there is a unique p.m. P on $(\Omega \times \Lambda, \mathcal{F} \times \mathcal{G})$ such that*

$$P(B \times C) \;=\; \int_C \int_B h(x,y)g(y)\, d\mu(x) d\nu(y) \quad , \text{ for all } B \in \mathcal{F} \text{ and } C \in \mathcal{G}. \tag{1.74}$$

Furthermore, letting $X(x,y) = x$ and $Y(x,y) = y$ define the random coordinate elements on $\Omega \times \Lambda$, then the following hold:

(i) *$Law[Y] << \nu$ and $dLaw[Y]/d\nu = g$, ν-a.e.*

(ii) *$Law[X|Y = y]$ has conditional density*

$$\frac{dLaw[X|Y = y]}{d\mu}(x) \;=\; f_{X|Y}(x|y) \;=\; h(x,y). \tag{1.75}$$

(iii) *$Law[X, Y] \ll \mu \times \nu$ and the joint density is*

$$\frac{dLaw[X, Y]}{d(\mu \times \nu)}(x,y) \;=\; h(x,y)g(y).$$

Proof. We give a sketch of the proof. It is easy to verify that $h(x,y)g(y)$ is a probability density function w.r.t. $\mu \times \nu$, and hence that there is a unique P such that (1.74) holds. The derivation of the conditional density then follows as in the proof of Proposition 1.5.4 (see also Remark 1.5.5). The formulae for the other densities (i.e. the purported marginal density for Y and the puported joint density) follow from the observation that they satisfy the defining property of being the appropriate density.

□

Finally, we close this chapter with a result which has far reaching and controversial applications in Statistics.

Theorem 1.5.12 (Bayes Formula) *Suppose $\Theta : (\Omega, \mathcal{F}, P) \longrightarrow (\Lambda_2, \mathcal{G}_2)$ is a random element and let λ be a σ-finite measure on $(\Lambda_2, \mathcal{G}_2)$ such that $Law[\Theta] \ll \lambda$. Denote the corresponding density by*

$$\pi(\theta) \;=\; \frac{dLaw[\Theta]}{d\lambda}(\theta).$$

Let μ be a σ–finite measure on $(\Lambda_1, \mathcal{G}_1)$. Suppose that for each $\theta \in \Lambda_1$ there is given a probability density function w.r.t. μ denoted $f(\cdot|\theta)$. Denote by X a *random element taking values in Λ_1 with $dLaw[X|\Theta = \theta]/d\mu = f(\cdot|\theta)$. Then there is a version of $Law[\Theta|X = x]$ given by*

$$\pi(\theta|x) \;=\; \frac{dLaw[\Theta|X = x]}{d\lambda}(\theta) \;=\; \frac{f(x|\theta)\pi(\theta)}{\int_{\Lambda_2} f(x|\vartheta)\pi(\vartheta)\,d\lambda(\vartheta)} \;.$$

Proof. By Proposition 1.5.11 with $h(x, \theta) = f(x|\theta)$, there is a joint distribution for (X, Θ) for which $\pi(\theta)$ is the marginal density of θ w.r.t. λ, $f(x|\theta)\pi(\theta)$ is the joint density for (X, Θ), and $f(x|\theta)$ is the conditional density for X given $\Theta = \theta$. There only remains to verify the formula for the conditional density of Θ given $X = x$. But the marginal density for X is the joint density with θ integrated out, i.e. $\int_{\Lambda_2} f(x|\theta)\pi(\theta)\,d\lambda(\theta)$. Thus, one recognizes the r.h.s. of the formula as the joint density divided by the marginal for X, i.e. the conditional density for Θ given $X = x$.

□

Exercises for Section 1.5.

1.5.1 Let $\Omega = [0,1]$, $\mathcal{F}$ = the Borel subsets of $[0,1]$, and P be the uniform distribution on $[0,1]$. Define a r.v. Y on this probability space by

$$Y(\omega) = \begin{cases} 1/2 & \text{if } \omega \leq 1/2, \\ \omega & \text{if } \omega > 1/2. \end{cases}$$

Let X be the r.v. given by

$$X = I_{[0,1/4)} - I_{[1/4,1/2)} + I_{[1/2,3/4)} - I_{[3/4,1]} \ .$$

Find both $E[X|Y]$ and $E[X|Y=y]$. Justify your answers.

1.5.2 Verify Remarks 1.5.4 and equation 1.70.

1.5.3 Let X and Y be random variables with the joint distributions as given below. In each case, determine $E[X|Y=y]$.

(a) Law$[X,Y]$ has density w.r.t. m^2 given by

$$f(x,y) = \frac{1}{\pi} I_D(x,y)$$

where $D = \{(x,y) : x^2 + y^2 \leq 1\}$.

(b)

$$\text{Law}[X,Y] = \frac{1}{4}\delta_{(0,0)} + \frac{1}{4}\delta_{(0,1)} + \frac{1}{4}\delta_{(1,1)} + \frac{1}{16}\delta_{(1,2)} + \frac{1}{16}\delta_{(2,2)} + \frac{1}{8}\delta_{(3,2)} \ .$$

1.5.4 Can you give $E[X|Y]$ in Exercise 1.5.3?

1.5.5 Suppose (1.59) holds. Show that (1.58) holds for all $W : (\Omega, \sigma(Y)) \longrightarrow (I\!R, \mathcal{B})$ with finite second moment.

1.5.6 In the setting of Theorem 1.5.7, show that $E[X_1 + X_2|\mathcal{G}] = E[X_1|\mathcal{G}] + E[X_2|\mathcal{G}]$. Also, give the proofs of parts (e), (f), and (g) of Theorem 1.5.7, and complete the proof of part (h).

1.5.7 Assuming $E[X^2] < \infty$, the *conditional variance* is given by

$$\text{Var}[X|\mathcal{G}] = E[(X - E[X|\mathcal{G}])^2|\mathcal{G}] \ .$$

(a) Show $\text{Var}[X|\mathcal{G}] = E[X^2|\mathcal{G}] - E[X|\mathcal{G}]^2$.

(b) Show $E[\text{Var}[X|\mathcal{G}]] = \text{Var}[X] - \text{Var}[E[X|\mathcal{G}]]$. Conclude that $E[\text{Var}[X|\mathcal{G}]] \leq \text{Var}[X]$.

(c) Find a simple necessary and sufficient condition for $\text{Var}[X|\mathcal{G}] = 0$ a.s.

(d) Suppose $E[Y^2] < \infty$ and $E[X^2|\mathcal{G}] = Y^2$ a.s., $E[X|\mathcal{G}] = Y$ a.s. Show $X = Y$ a.s.

1.5.8 Prove Theorem 1.5.8 (b).

1.5.9 State and prove analogues of Theorems 1.5.7 and 1.5.8 for the conditional expectations of the type $E[X|Y = y]$.

1.5.10 Suppose $E[X^2] < \infty$. Let $\mathbf{Z} = \{Z : Z = h(Y)$ for some measurable h and $E[Z^2] < \infty\}$. Show that $E[X|Y]$ is the essentially unique element of $\mathbf{Z}$ which minimizes MSPE(Z) as defined in (1.57). Hint: You don't know at this point that such a minimizer exists. Start out with any $Z \in \mathbf{Z}$ and show that MSPE(Z) – MSPE($E[X|Y]$) ≥ 0 by using properties of conditional expectation.

1.5.11 Let X be a random variable. Show that Law$[X|X = x] = \delta_x$. Explain why this is intuitively correct.

1.5.12 Let $\underline{X} = (X_1, X_2)$ be a random 2-vector. What is Law$[\underline{X}|X_1 = x_1]$?

1.5.13 Assume X is a r.v. which is either discrete, i.e. supp$[X] = \{x_1,\ x_2,$ ...$\}$ is either a finite or infinite sequence and $f(x_i) = P[X = x_i]$ is the density function w.r.t. # on supp$[X]$, or X has a Lebesgue density $f(x)$. Let $Y = X^2$. Show that for $y \geq 0$, Law$[X|Y = y]$ has a discrete distribution on $\{\sqrt{y},\ -\sqrt{y}\}$ and determine the distribution.

1.5.14 Assume X is a r.v. which is either discrete or has a Lebesgue density $f(x)$ as in Exercise 1.5.13, and let $Y = \sin(X)$. What is Law$[X|Y = y]$? (Note: The answer to this one is real messy.)

Chapter 2

Probability Measures on Euclidean Spaces

In this chapter, we consider the detailed properties of probability models on Euclidean spaces. These are widely used as models in statistics.

2.1 Moments and Moment Inequalities.

In general, a *moment* refers to an expectation of a random variable or a function of that r.v. The *mean* or first moment is $E[X]$, often denoted μ_X or just μ. It will always be clear from context whether we mean μ to denote the mean of a r.v. or a measure. The *k'th moment* is $E[X^k]$, sometime denoted μ_k, and the *k'th central moment* is $E[(X-\mu)^k]$. The second central moment is called the *variance* and is also denoted $\mathrm{Var}[X]$, σ_X^2, or just σ^2. Of course, one can show $\mathrm{Var}[X] = E[X^2] - (E[X])^2 = \mu_2 - \mu_1^2$. Similar formulae can be obtained for other central moments in terms of noncentral moments.

Let $\underline{X}$ be a random n-vector, say $\underline{X} = (X_1, X_2, \ldots X_n)$. We say $E[\underline{X}]$ *exists* iff each $E[X_i]$ exists, $1 \leq i \leq n$, and then we define $E[\underline{X}] = (E[X_1], E[X_2], \ldots E[X_n])$. Similarly, $\underline{X}$ *is integrable* iff each component r.v. is integrable. Integrability of $\underline{X}$ is equivalent to $E\|\underline{X}\| < \infty$ (Exercise 2.1.1). If $\underline{X}$ is integrable then $\underline{\mu} = E[\underline{X}] \in \mathbb{R}^n$.

Now assume further that $\underline{X}$ has *finite second moments*, i.e. each component of $\underline{X}$ has finite second moment, which is the same as $E[\|\underline{X}\|^2] < \infty$. Finiteness of the second moment implies $\underline{X}$ has *finite first moments*, i.e. $E[\|\underline{X}\|] < \infty$. This follows from the following calculation.

$$\begin{aligned} E[\|\underline{X}\|] &= \int_{R^n} \|\underline{x}\| \, dP_{\underline{X}}(\underline{x}) \\ &= \int_{\{\underline{x}:\|\underline{x}\|<1\}} \|\underline{x}\| \, dP_{\underline{X}}(\underline{x}) + \int_{\{\underline{x}:\|\underline{x}\|\geq 1\}} \|\underline{x}\| \, dP_{\underline{X}}(\underline{x}) \\ &\leq \int_{\|\underline{x}\|<1} 1 \, dP_{\underline{X}}(\underline{x}) + \int_{\|\underline{x}\|\geq 1} \|\underline{x}\|^2 \, dP_{\underline{X}}(\underline{x}) \end{aligned}$$

$$\leq \quad P[\|\underline{X}\| < 1] \,+\, E\|\underline{X}\|^2 \,<\, \infty \quad . \tag{2.1}$$

In general, we say $\underline{X}$ has finite p'th moment, $0 < p < \infty$, if $E[\|\underline{X}\|^p] < \infty$. If $\underline{X}$ has finite p'th moment, then all smaller moments are also finite (Exercise 2.1.5).

2.1.1 Elementary Moment Bounds.

We now recall a couple of inequalities (attributed to Russian probabilists) which allow us to estimate probabilities using moments.

Proposition 2.1.1 (Markov's Inequality.) *Suppose $X \geq 0$ a.s., then for all $\epsilon > 0$,*

$$P[\,X > \epsilon\,] \;\leq\; \frac{E[X]}{\epsilon} \quad .$$

Note: $P[\,X > \epsilon\,]$ is a shorthand way to write $P(\{\,\omega \in \Omega \,:\, X(\omega) > \epsilon\,\})$, which is equal to $(P \circ X^{-1})((\epsilon, \infty))$.. There are many other inequalities that are trivial corollaries of this one, e.g. $P[\,|X| > \epsilon\,] \leq E[X^2]/\epsilon^2$. We will give the proof in some detail, although it is really quite elementary. The student should be able to reproduce the proof and completely justify each detail.

Proof.

$$E[X] \;=\; \int_0^\infty x\, dP_X(x)$$

by the Law of the Unconscious Statistician (Theorem 1.2.10),

$$= \int_{[0,\epsilon)} x\, dP_X(x) \,+\, \int_{[\epsilon,\infty]} x\, dP_X(x)$$

by additivity of the integral (Proposition 1.2.2 (b)) as applied to $x = I_{[0,\epsilon)}(x)x + I_{[\epsilon,\infty]}(x)x$,

$$\geq \int_{[\epsilon,\infty]} x\, dP_X(x)$$

since $I_{[0,\epsilon)}(x)x \geq 0$ and so also is its integral (Proposition 1.2.5 (c)),

$$\geq \int_{[\epsilon,\infty]} \epsilon\, dP_X(x)$$

by Proposition 1.2.5 (c) because $I_{[\epsilon,\infty]}(x)x \geq I_{[\epsilon,\infty]}(x)\epsilon$,

$$= \;\epsilon P[\,X > \epsilon\,]$$

by Proposition 1.2.5 (a) and the fact that $\int I_A\, dP = P(A)$.

□

Proposition 2.1.2 (Chebyshev's Inequality.) *Suppose X is a r.v. with $E[X^2] < \infty$. Let $\mu = E[X]$ and $\sigma^2 = Var[X]$. Then for any $k > 0$,*

$$P[\,|X - \mu| \geq k\sigma\,] \;\leq\; \frac{1}{k^2} \quad .$$

Proof. Apply Markov's inequality to the r.v. $(X-\mu)^2$ with $\epsilon = (k\sigma)^2$.

□

2.1.2 Convexity and Jensen's Inequality.

Now we turn to a much more subtle inequality, after some definitions. A set $K \subset I\!R^n$ is called *convex* iff for any finite subset

$$\{\,\underline{x}_1, \underline{x}_2, ..., \underline{x}_m\} \;\subset\; K$$

and any real numbers p_1, p_2, ... p_m with

$$p_i \geq 0 \; \forall i \quad , \quad \sum_{i=1}^{m} p_i \;=\; 1 \quad , \tag{2.2}$$

we have

$$\sum_{i=1}^{m} p_i \underline{x}_i \;\in\; K \quad . \tag{2.3}$$

A linear combination such as the l.h.s. of (2.3) with the coefficients satisfying (2.2) is called a *convex combination*. Thus, a set K is convex iff it is closed under taking convex combinations. It can be shown that it suffices to take $m = 2$ in (2.3), i.e. if one shows that (2.3) holds with $m = 2$ then it holds for all m. Assuming $m = 2$, one can see geometrically that as p_1 and p_2 vary over values satisfying (2.2), the set of vectors $p_1\underline{x}_1 + p_2\underline{x}_2$ so obtained is the line segment between $\underline{x}_1$ and $\underline{x}_2$. Thus, a set K is convex iff for every two points in K, the line segment between the two points is contained in K. See Rudin's *Principles* for more discussion of convex sets and related notions.

Let $f : K \longrightarrow I\!R$ where K is a convex subset of $I\!R^n$. Then f is called a *convex function* iff

$$\{\,\underline{x}_1, \underline{x}_2, ..., \underline{x}_m\} \;\subset\; K$$

and

$$p_i \geq 0 \; \forall i \quad , \quad \sum_{i=1}^{m} p_i \;=\; 1 \quad , \tag{2.4}$$

implies

$$f\left(\sum_{i=1}^{m} p_i \underline{x}_i\right) \;\leq\; \sum_{i=1}^{m} p_i f(\underline{x}_i) \quad . \tag{2.5}$$

Taking $m = 2$, one can see that (2.5) implies that the line segment in $\mathbb{R}^{n+1}$ between $(\underline{x}_1, f(\underline{x}_1))$ and $(\underline{x}_2, f(\underline{x}_2))$ lies on or above the graph of $f(\underline{x})$. The function f is called *strictly convex* iff

$$p_i > 0 \, \forall i \quad , \quad \sum_{i=1}^{m} p_i = 1 \quad , \tag{2.6}$$

implies

$$f\left(\sum_{i=1}^{m} p_i \underline{x}_i \right) < \sum_{i=1}^{m} p_i f(\underline{x}_i) \quad . \tag{2.7}$$

We wish to give an easily checked sufficient condition for convexity of a function, but some definitions are needed first. Suppose $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ has continuous second order partial derivatives. The Hessian matrix $D^2 f(\underline{x}) = H(\underline{x})$ is given by

$$H_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j} \quad . \tag{2.8}$$

Note that H is actually a mapping of n-vectors to $n \times n$ matrices. If B is an $n \times m$ matrix with (i, j) entry B_{ij}, the the *transpose* of B, denotes B^t, is an $m \times n$ matrix obtained by interchanging rows and columns, i.e. the (i, j) entry of B^t is B_{ji}. A $n \times n$ matrix A is *symmetric* iff $A^t = A$, which is the same as $A_{ij} = A_{ji}$ for all i and j. Observe that our assumption of continuity of the second order partial derivatives of f implies equality of mixed partials (i.e. $\partial^2 f/\partial x_i \partial x_j = \partial^2 f/\partial x_j \partial x_i$; Reference???), and hence that the Hessian is symmetric. A symmetric matrix A is called *nonnegative definite* iff $y^t A y \geq 0$ for all n-vectors y. Note that y^t is an $1 \times n$ matrix, and

$$y^t A y = \sum_{i=1}^{n} \sum_{j=1}^{n} y_i A_{ij} y_j \quad .$$

A symmetric matrix A is called *strictly positive definite* iff $y^t A y > 0$ for all nonzero n-vectors y. (Note: our terminology is nonambiguous, but "positive definite" is used by some authors to mean "nonnegative definite" and by other authors to mean "strictly positive definite." Some authors also use "positive semidefinite" to mean "nonnegative definite.")

Theorem 2.1.3 *Suppose $f : K \longrightarrow \mathbb{R}$ where $K \subset \mathbb{R}^n$ is a convex, open set and f has continuous second order partial derivatives on K.*

(a) If the Hessian matrix $H(\underline{x})$ is nonnegative definite for all $\underline{x} \in K$, then f is convex.

(b) If $H(\underline{x})$ is strictly positive definite for all $\underline{x} \in K$, then f is strictly convex.

Partial Proof. Fix arbitrary $\underline{x}_0$ and $\underline{x}_1$ in K, and consider

$$g(p) = (1 - p) f(\underline{x}_0) + p f(\underline{x}_1), \tag{2.9}$$

for $p \in (0,1)$. It suffices to check that g is convex or strictly convex, which is a one dimensional problem. This illustrates a common theme in convex analysis: general problems involving convex functions can often be reduced to problems involving functions of a single real variable.

□

Simple examples of convex functions are

$$f(\underline{x}) \;=\; \|\underline{x}\|^p \quad , \quad \underline{x} \in I\!R^n \ , \ \text{where } p \geq 1 \ ; \tag{2.10}$$

$$f(x) \;=\; x^{-p} \quad , \quad x \in (0,\infty) \quad , \ \text{where } p \geq 0 \ ; \tag{2.11}$$

$$f(x) \;=\; e^{ax} \quad , \quad x \in I\!R \ ; \tag{2.12}$$

$$f(\underline{x}) \;=\; \underline{x}^t Q \underline{x} \ , \quad \underline{x} \in I\!R^n \quad , \tag{2.13}$$

$$\text{where } Q \text{ is nonnegative definite.}$$

If $p > 1$ in (2.10), $p > 0$ in (2.11), $a \neq 0$ in (2.12), and Q strictly positive definite in (2.13), then f is strictly convex in all of these examples.

A real valued function f defined on a convex set K is called *concave* if $-f$ is convex, and similarly f is *strictly concave* if $-f$ is strictly convex. Some examples:

$$f(\underline{x}) \;=\; \|\underline{x}\|^p \quad , \quad \underline{x} \in [0,\infty) \ , \ \text{where } 0 \leq p \leq 1 \ ; \tag{2.14}$$

$$f(x) \;=\; \log x \quad , \quad x \in (0,\infty) \quad . \tag{2.15}$$

The power functions in (2.14) are strictly concave if $0 < p < 1$, and the log function is strictly concave. Most functions of a single real variable one encounters are altenately convex and concave with the intervals of convexity and concavity separated by points of inflection. In two or more dimensions, a function may be neither convex nor concave in a nontrivial region, e.g. $f(x,y) = x^2 - y^2$ is neither convex nor concave anywhere in $I\!R^2$.

Equations (2.4) and (2.5) may interpreted probabilistically. Let $\underline{X}$ be a discrete random n-vector with distribution given by

$$P[\underline{X} = \underline{x}_i] \;=\; p_i \ , \tag{2.16}$$

i.e. Law$[\underline{X}] = \sum_{i=1}^m p_i \delta_{\underline{x}_i}$. Note that by (2.4), this latter summation is a p.m. Then the r.h.s. of (2.5) is $E[f(\underline{X})]$, and the l.h.s. is $f(E[\underline{X}])$. With somewhat more effort, one can show

Theorem 2.1.4 (Jensen's Inequality.) *Let f be a convex function on a convex set $K \subset I\!R^n$ and suppose $\underline{X}$ is a random n-vector with $E\|\underline{X}\| < \infty$ and $X \in K$ a.s. Then $E[\underline{X}] \in K$ and*

$$f(E[\underline{X}]) \;\leq\; E[f(\underline{X})] \quad .$$

Furthermore, if f is strictly convex and Law$[\underline{X}]$ is nondegenerate (i.e. $\underline{X}$ is not a.s. equal to a constant, or equivalently Law$[\underline{X}]$ is not a unit point mass), then strict inequality holds in the above.

□

A proof may be found in Billingsley on p. 283.

2.1.3 The Covariance Matrix.

Now we look more closely at some "quadratic" moments.

Theorem 2.1.5 (Cauchy-Schwarz Inequality.) *For any r.v.'s X and Y,*

$$(E|XY|)^2 \ \leq \ E[X^2]E[Y^2] \, .$$

Assume the l.h.s. is finite. Then equality holds iff either $X = 0$ a.s. or $Y = cX$ a.s. for some constant c.

□

A proof may be found in Billingsley on p. 283.

Let $\underline{X} = (X_1, X_2, \ldots X_n)$ be a random n-vector with finite second moments. We have by Cauchy-Schwarz that

$$E[|(X_i - \mu_i)(X_j - \mu_j)|] \ \leq \ \left\{ E[(X_i - \mu_i)^2]E[(X_j - \mu_j)^2] \right\}^{1/2}$$

and

$$E[(X_i - \mu_i)^2] \ = \ \mathrm{Var}[X_i] \ = \ E[X_i^2] - \mu_i^2 \in [0, \infty) \quad ,$$

so $(X_i - \mu_i)(X_j - \mu_j)$ is integrable.

Assuming $E[\|\underline{X}\|^2] < \infty$, we define the *covariance matrix* $V = \mathrm{Cov}[\underline{X}]$ by

$$V_{ij} \ = \ E[(X_i - \mu_i)(X_j - \mu_j)] \quad , \quad 1 \leq i, j \leq n \quad ,$$

or, in a more compact matrix notation,

$$V \ = \ E[(\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})^t] \quad .$$

Note that V is an $n \times n$ matrix with real (i.e. finite) entries. In fact, by the above

$$|V_{ij}| \ \leq \ \sqrt{\mathrm{Var}[X_i]\mathrm{Var}[X_j]} \quad . \tag{2.17}$$

Also, one can further check that $\mathrm{Cov}[\underline{X}]$ is symmetric and nonnegative definite (Exercise 2.1.6). If A is any $m \times n$ matrix and $\underline{b} \in I\!R^m$, then (Exercise 2.1.8)

$$\mathrm{Cov}[A\underline{X} + \underline{b}] \ = \ A\mathrm{Cov}[\underline{X}]A^t \quad . \tag{2.18}$$

If $\underline{X}$ is a random n-vector and $\underline{Y}$ is a random m-vector, then the *covariance between $\underline{X}$ and $\underline{Y}$* is

$$\mathrm{Cov}[\underline{X}, \underline{Y}] \ = \ E[(\underline{X} - E[\underline{X}])(\underline{Y} - E[\underline{Y}])^t] \quad ,$$

which is an $n \times m$ matrix. Note that

$$\mathrm{Cov}[\underline{X}] \;=\; \mathrm{Cov}[\underline{X}, \underline{X}] \quad , \qquad \mathrm{Cov}[\underline{Y}, \underline{X}] \;=\; \mathrm{Cov}[\underline{X}, \underline{Y}]^t \quad . \tag{2.19}$$

If $\underline{Z}$ is the random $n+m$–vector $(\underline{X}, \underline{Y})$, then

$$\mathrm{Cov}[\underline{Z}] \;=\; \left[\begin{array}{cc} \mathrm{Cov}[\underline{X}] & \mathrm{Cov}[\underline{X}, \underline{Y}] \\ & \\ \mathrm{Cov}[\underline{Y}, \underline{X}] & \mathrm{Cov}[\underline{Y}] \end{array} \right] \quad . \tag{2.20}$$

If $\mathrm{Cov}[\underline{X}, \underline{Y}] = 0$ (where the latter is a matrix of zeroes), then we say $\underline{X}$ and $\underline{Y}$ are *uncorrelated*. One can show that if $\underline{X}$ and $\underline{Y}$ are independent then they are uncorrelated, provided both have finite second moments, but the converse is false (Exercise 2.1.10).

Now we introduce some matrix theory which is extremely useful in many areas of statistics. Recall that a square matrix U is called *orthogonal* iff $U^{-1} = U^t$. Assuming U is $n \times n$, then U is an orthogonal matrix if and only if the columns of U form an orthonormal basis for $I\!R^n$. A square matrix D is *diagonal* if the off diagonal entries are zero, i.e. $D_{ij} = 0$ if $i \neq j$. It will be convenient to write $D = \mathrm{diag}[\underline{d}]$, where $\underline{d}$ is the vector of diagonal entries, i.e.

$$D \;=\; \left[\begin{array}{cccccc} d_1 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & d_2 & & & & \cdot \\ \cdot & & \cdot & & & \cdot \\ \cdot & & & \cdot & & \cdot \\ \cdot & & & & \cdot & 0 \\ 0 & \cdot & \cdot & \cdot & 0 & d_n \end{array} \right] \quad .$$

Theorem 2.1.6 (Spectral Decomposition of a Symmetric Matrix.) *Let A be a symmetric matrix. Then there is an orthogonal matrix U and a diagonal matrix Λ such that $A = U \Lambda U^t$.*

□

For a proof, see pp. 39-40 of Rao, *Linear Statistical Inference and Its Applications.* This decomposition is sometimes called the *orthogonal-diagonal-orthogonal* or *eigenvector-eigenvalue*, decomposition. In fact, the diagonal entries of Λ are the eigenvalues of A, and the columns of U are the corresponding eigenvectors (Exercise 2.1.11).

Proposition 2.1.7 *Suppose $\underline{X}$ is a random n-vector with finite second moments. Then there is an orthogonal matrix U such that $Cov[U^t \underline{X}]$ is a diagonal matrix.*

Proof. Since $V = \mathrm{Cov}[\underline{X}]$ is symmetric there is an orthogonal matrix U and a diagonal matrix Λ such that $V = U \Lambda U^t$. Then multiplying on the left by U^t and on the right by U and using the defining property of an orthogonal matrix, $\Lambda = U^t V U$. The result now follows from (2.18) with $A = U^t$.

□

Assume $\underline{X}$ is a random n-vector with finite second moments and put $\mu = E[\underline{X}]$, $V = \mathrm{Cov}[\underline{X}]$. Write $V = U\Lambda U^t$, where U is orthogonal and Λ is diagonal, as in the proof of the last result. Since V is nonnegative definite, the eigenvalues (which are the diagonal entries of Λ) are nonnegative (Exercise 2.1.12). Assume that the number of positive eigenvalues is r, so there are $n - r$ zero eigenvalues. We may reorder the diagonal entries of Λ (as long as we correspondingly reorder the columns of U), and it is convenient to assume

$$\Lambda \;=\; \mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_r, 0, 0, ..., 0) \quad ,$$

where

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r > 0 \;.$$

Now write $\underline{u}_j = (u_{1j}, u_{2j}, \ldots, u_{nj})$ for the j^{th} column of U. The *null space* of V (which is defined to be the set of vectors $\underline{x}$ such that $V\underline{x} = \underline{0}$) is given by

$$\mathbf{N}(V) \;=\; \mathrm{span}[\underline{u}_{r+1}, \underline{u}_{r+2}, ..., \underline{u}_n] \;, \tag{2.21}$$

and the *column space* or *range* (which is defined to be $\{V\underline{x} : \underline{x} \in I\!\!R^n\}$) is given by

$$\begin{aligned}\mathbf{R}(V) \;&=\; \mathrm{span}[\underline{u}_1, \underline{u}_2, ..., \underline{u}_r] \qquad (2.22)\\ &=\; \{\sum_{i=1}^{r} a_i\underline{u}_i \;:\; a_i \in I\!\!R \text{ for all } i\,\} \quad .\end{aligned}$$

Here, the *span* of a collection of vectors is the set of all linear combinations of the given collection, i.e. the smallest linear subspace which includes the given collection. Also, $r = \mathrm{rank}(V)$, the dimension of the range of V, is known as the *rank* of the linear transformation V. Equations (2.21) and (2.22) follow since any $\underline{x} \in I\!\!R^n$ may be expanded as

$$\underline{x} \;=\; \sum_{i=1}^{n} (\underline{x}^t\underline{u}_i)\,\underline{u}_i \;, \tag{2.23}$$

because $\{\underline{u}_i : 1 \leq i \leq n\}$ form an orthornormal basis for $I\!\!R^n$. Thus,

$$V\underline{x} \;=\; \sum_{i=1}^{n} (\underline{x}^t\underline{u}_i)\,V\underline{u}_i \;=\; \sum_{i=1}^{n} \lambda_i(\underline{x}^t\underline{u}_i)\,\underline{u}_i \;=\; \sum_{i=1}^{r} \lambda_i(\underline{x}^t\underline{u}_i)\,\underline{u}_i \quad . \tag{2.24}$$

Thus, $V\underline{x} = 0$ iff $\underline{x}^t\underline{u}_i = 0$ for $1 \leq i \leq r$, which is true if and only if $\underline{x} \in \mathrm{span}[\underline{u}_{r+1}, \underline{u}_{r+2}, ..., \underline{u}_n]$. Also, $\underline{y} = V\underline{x}$ for some $\underline{x} \in I\!\!R^n$ iff $\underline{y}$ has the form of the last expression given in (2.24), which is true if and only if $\underline{y} \in \mathrm{span}[\underline{u}_1, \underline{u}_2, ..., \underline{u}_r]$. Note that in this latter case we may take $\underline{y} = V\underline{x}$ where

$$\underline{x} \;=\; \sum_{i=1}^{r} \lambda_i^{-1}(\underline{y}^t\underline{u}_i)\,\underline{u}_i \;=\; V^{-}\underline{y} \quad . \tag{2.25}$$

Here, the last equation defines the linear transformation V^-, which is known as the *Moore-Penrose generalized inverse* of V. Note that $V^-\underline{y}$ is just one of infinitely many $\underline{x}$'s satisfying $V\underline{x} = y$ when rank$(V) < n$. If rank$(V) = n$, i.e. V is nonsingular, then $V^- = V^{-1}$.

Proposition 2.1.8 *If $\underline{X}$ is a random n–vector with $E[\|\underline{X}\|^2] < \infty$, $\underline{\mu} = E[\underline{X}]$, and $V = Cov[\underline{X}]$, then*

$$P[\underline{X} \in \mathbf{R}(V) + \underline{\mu}] \ = \ 1 \ .$$

where $\mathbf{R}(V) + \underline{\mu} = \{\underline{y} + \underline{\mu} : \underline{y} \in \mathbf{R}(V)\}$.

Proof. Let $\underline{Y} = \underline{X} - \underline{\mu}$, so $\underline{Y} \in \mathbf{R}(V)$ iff $\underline{X} \in \mathbf{R}(V) + \underline{\mu}$. Write

$$\underline{Y} \ = \ \sum_{i=1}^{n} Y_i \underline{u}_i \quad , \quad Y_i \ = \ \underline{Y}^t \underline{u}_i \quad ,$$

where the $\underline{u}_i$ are the same as in (2.21) and (2.22). Then Cov$[\underline{X}]$ = Cov$[\underline{Y}]$ and

$$E[Y_i^2] \ = \ \underline{u}_i^t V \underline{u}_i \ = \ \lambda_i \quad . \tag{2.26}$$

See Exercise 2.1.13. Hence, $E[Y_i^2] = 0$ iff $i > r$ by Exercise 2.1.2, which is equivalent to $Y_i = 0$ a.s. iff $i > r$, hence

$$\underline{Y} \ = \ \sum_{i=1}^{r} Y_i \underline{u}_i \quad , \quad \text{a.s.}$$

which implies $\underline{Y} \in \mathbf{R}(V)$.

□

Proposition 2.1.9 *Let $\underline{X}$ be as in Proposition 2.1.8. If Law$[\underline{X}] << m^n$ then rank$[Cov[\underline{X}]] = n$.*

Proof. If rank$[V] = r < n$, then $\mathbf{R}(V)$ is a proper linear subspace of $I\!\!R^n$, and $\mathbf{R}(V) + \underline{\mu}$ is a proper *linear manifold*, i.e. a translate of a proper linear subspace. Such a set is closed, hence a Borel set, and we claim its Lebesgue measure is 0.

We have that

$$\mathbf{R}(V) + \underline{\mu} \subset \{\underline{x} \in I\!\!R^n : (\underline{x} - \underline{\mu})^t \underline{u}_n = 0\} \ := \ B. \tag{2.27}$$

See Exercise 2.1.22. Applying Fubini's Theorem (Theorem 1.3.2) and the fact that $m^n = m^{n-1} \times m$ by definition, we have

$$\begin{aligned} m^n(B) \quad &= \quad \int_{R^n} I_B(\underline{x}) \, d\underline{x} \\ &= \quad \int_{R^{n-1}} \int_R I_B(\underline{y}, x_n) \, dx_n d\underline{y} \end{aligned} \tag{2.28}$$

where $\underline{y} = (x_1, ..., x_{n-1})$. Now $I_B(\underline{y}, x_n) = 1$ iff

$$(x_n - \mu_n)u_{nn} \;=\; -\sum_{i=1}^{n-1}(y_j - \mu_j)u_{nj}$$

where $\underline{u}_n = (u_{1n},\, u_{2n},\, ...,\, u_{nn})$. Assuming $u_{nn} \neq 0$, then for fixed $\underline{y} \in I\!R^{n-1}$, $I_B(\underline{y}, x_n) \neq 0$ iff

$$x_n \;=\; \mu_n \;-\; u_{nn}^{-1}\sum_{i=1}^{n-1}(y_i - \mu_i)u_{in}$$

which is only a single point. Hence, the inner integral in (2.28) is 0. If $u_{nn} = 0$, then some component of $\underline{u}_n$ is nonzero (since $\underline{u}_n$ is an element of an orthornormal basis for $I\!R^n$), say $u_{nj} \neq 0$, and then replace x in (2.28) with u_{nj} and $\underline{y}$ with the remaining components of $\underline{x}$.

□

Exercises for Section 2.1.

2.1.1 Show that the random n-vector $\underline{X}$ is integrable iff $E\|\underline{X}\| < \infty$. (Hint: Show that for any vector $\underline{x}$,

$$\max\{|x_i| \, : \, 1 \leq i \leq n\} \; \leq \; \|\underline{x}\| \; \leq \; \sum_{i=1}^{n} |x_i| \quad .)$$

2.1.2 Suppose X is a random variable with finite second moment. Show Var$[X]$ $= 0$ if and only if X is degenerate (i.e. there is a constant a such that $P[X = a]$ $= 1$, which is the same as Law$[X] = \delta_a$). Show that necessarily $a = E[X]$.

2.1.3 Verify convexity of each of the functions given in (2.10) through (2.13). Also verify the strict convexity claims under the stricter conditions mentioned below (2.13). Similarly, verify the concavity and strict concavity claims for the functions in (2.14) and (2.15).

2.1.4 Verify $E[X]^2 \leq E[X^2]$ using both Jensen's inequality and the Cauchy-Schwarz inequality. Use the equality conditions in both inequalities to develop necessary and sufficient conditions for $E[X]^2 = E[X^2]$.

2.1.5 Let $0 \leq p < q < \infty$ and let $\underline{X}$ be any random vector. Show that $E[\|\underline{X}\|^p]$ $\leq P[\|\underline{X}\| \leq 1] + E[\|\underline{X}\|^q]$. In particular, if $\underline{X}$ has moment of order q (i.e. $E[\|\underline{X}\|^q] < \infty$), then it also has moment of any smaller order.

2.1.6 Show that for any random n-vector $\underline{X}$ with finite second moment, Cov$[\underline{X}]$ is symmetric and nonnegative definite.

2.1.7 Assuming $\underline{X}$ and $\underline{Y}$ have finite second moments, show

$$\text{Cov}[\underline{X}, \underline{Y}] \; = \; E[\underline{X}\underline{Y}^t] \; - \; E[\underline{X}]E[\underline{Y}]^t \quad .$$

2.1.8 Assuming $\underline{X}$ is a random n-vector with finite second moment, A is any $m \times n$ matrix, and $\underline{b} \in I\!R^m$, show that $A\underline{X} + \underline{b}$ has finite second moment and verify that (2.18) holds.

2.1.9 Let $\underline{X}$ be a random n-vector with finite second moment and $\underline{Y}$ be a random m-vector with finite second moment. Let A be a $k \times n$ matrix and B be a $j \times m$ matrix. Find Cov$[A\underline{X}, B\underline{Y}]$.

2.1.10 Let $\underline{X}$ and $\underline{Y}$ be random vectors with finite second moments. Show that independence of $\underline{X}$ and $\underline{Y}$ implies they are uncorrelated, but that the converse is false.

2.1.11 Let $A = U\Lambda U^t$ be the spectral decomposition of the symmetric matrix A as given in Theorem 2.1.6. Let $\lambda_i = \Lambda_{ii}$ be the i'th diagonal entry of Λ, and let $\underline{u}_j$ be the j'th column of U. Show that $A\underline{u}_j = \lambda_j \underline{u}_j$, i.e. that λ_j is an eigenvalue of A and $\underline{u}_j$ is the corresponding eigenvector.

2.1.12 Let A be as in Exercise 2.1.11. Show that A is nonnegative definite if and only if $\lambda_i \geq 0$ for all i, and A is strictly positive definite if and only if $\lambda_i > 0$ for all i.

2.1.13 Verify (2.26).

2.1.14 Let $\underline{X}$ be a random n-vector with distribution

$$\mathrm{Law}[\underline{X}] \; = \; \frac{1}{3}\left\{ \, \delta_{(0,0,0)} + \delta_{(0,1,1)} + \delta_{(1,0,0)} \, \right\} \quad .$$

(a) Find a 3×3 matrix A and a 3-vector $\underline{b}$ such that (i) $E[A\underline{X} + \underline{b}] = \underline{0}$ and (ii) $\mathrm{Cov}[A\underline{X} + \underline{b}] = \mathrm{diag}(1, 1, 0)$.

(b) Let A be any 3×3 matrix and $\underline{b}$ any 3-vector such that (i) and (ii) of part (a) hold. Put $Y = A\underline{X} + \underline{b}$. Show that Y_1 and Y_2 are uncorrelated but not independent.

2.1.15 Let $\underline{X}$ be as in Proposition 2.1.8.

(a) Show that $\mathrm{supp}[\mathrm{Law}[\underline{X}]] \subset \mathbf{R}(V) + \underline{\mu}$.

(b) Let M be any linear manifold. Show that $P[\underline{X} \in M] = 1$ implies $\mathbf{R}(V) + \underline{\mu} \subset M$.

(c) Show that $\dim(M) < \mathrm{rank}(V)$ implies $P[\underline{X} \in M] < 1$ for any linear manifold M.

2.1.16 Prove the Cauchy-Schwarz inequality (Theorem 2.1.5). Remember that any one of the expectations may be infinite.

2.1.17 Suppose $1 \leq p \leq q < \infty$, and that $E[|X|^q] < \infty$. Show $(E[|X|^p])^{1/p} \leq (E[|X|^q])^{1/q}$. This is known as *Lyapunov's Inequality.* Give necessary and sufficient conditions for equality.

2.1.18 Let $\underline{X}$ be a random n-vector with finite second moments and put $V = \mathrm{Cov}[\underline{X}]$. Let $V = U\Lambda U^t$ be the spectral decomposition of V. Assume V has full rank. Put

$$A \; = \; U\mathrm{diag}[\lambda_1^{-1/2}, \lambda_2^{-1/2}, ..., \lambda_n^{-1/2}]U^t \quad .$$

(a) Show that A is symmetric, positive definite, and $A^2 = V^{-1}$. We will write

$$A \; = \; V^{-1/2} \quad . \tag{2.29}$$

(b) Show that $\mathrm{Cov}[A\underline{X}] = I$.

(c) Suppose B is an $n \times n$ matrix such that $\mathrm{Cov}[B\underline{X}] = I$. Show that $BB^t = V^{-1}$. Show that if B is symmetric, then $B = V^{-1/2}$. Show that in general, $B = WV^{-1/2}$ where W is orthogonal.

2.1.19 (Cholesky decomposition) Let $\underline{X}$ and V be as in the previous exercise. Let

$$Y_1 \;=\; (X_1 - \mu_1)/\sqrt{\mathrm{Var}[X_1]} \quad . \tag{2.30}$$

Assuming Y_1, Y_2, ..., Y_{m-1} have been defined, define

$$Y_m \;=\; \frac{\left\{ (X_m - \mu_m) - \sum_{i=1}^{m-1} E[X_m Y_i] Y_i \right\}}{\sqrt{\left\{ (X_m - \mu_m) - \sum_{i=1}^{m-1} E[X_m Y_i] Y_i \right\}}} \quad . \tag{2.31}$$

Let $\underline{Y} = (Y_1, Y_2, ..., Y_n)$. Show the following results.

(a) $E[\underline{Y}] = \underline{0}$ and $\mathrm{Cov}[\underline{Y}] = I$.

(b) $\underline{Y} = B\underline{X}$ where B is a lower triangular matrix (i.e. $B_{ij} = 0$ if $j > i$) with positive diagonal entries.

(c) B is the only lower triangular matrix with positive diagonal entries such that $BB^t = V^{-1}$.

Remarks 2.1.1 B^t given above is called the *Cholesky factor* of V^{-1}. For computational purposes, B is perhaps the most useful matrix satisfying the property in (c). In particular, note that B is easy to compute from the recursive formula given in (2.30) and (2.31) and B^{-1} is also lower triangular and easy to compute from B. In solving this problem, you may find it useful to compare the recursive definition of $\underline{Y}$ with the Gram-Schmidt orthonormalization procedure of linear algebra.

2.1.20 Suppose P and Q are p.m.'s on a measurable space $(\Omega, \mathcal{F})$ with $Q \ll P$. Define the *Kullback-Leibler Information* between Q and P by

$$K(Q,P) \;=\; \int_\Omega \log\left(\frac{dQ}{dP}\right) dQ \quad .$$

Show the following.

(a) K is well defined, i.e. the integral defining K exists. (Hint: First show $\log t \leq t - 1$ for $0 \leq t < \infty$. Equality holds only if $t = 1$.)

(b) $0 \leq K(Q,P) \leq \infty$.

(c) $K(Q,P) = 0$ iff $P = Q$.

2.1.21 Prove Theorem 2.1.3.

2.1.22 Verify (2.27).

2.2 Characteristic and Moment Generating Functions.

In this section we consider some other useful special expectations w.r.t. a p.m. on a Euclidean space.

2.2.1 General Results.

Definition 2.2.1 *The* characterstic function *(abbreviated ch.f.) of a random n-vector $\underline{X}$ is the complex valued function $\phi_{\underline{X}} : I\!R^n \longrightarrow \mathbb{C}$ given by*

$$\begin{aligned} \phi_{\underline{X}}(\underline{u}) &= E[\exp(i\underline{u}^t\underline{X})] \\ &= E[\cos(\underline{u}^t\underline{X})] + iE[\sin(\underline{u}^t\underline{X})] \quad . \end{aligned}$$

(Here, $i = \sqrt{-1}$ is the imaginary unit.) The moment generating function *(abbreviated m.g.f.) is given by*

$$\psi_{\underline{X}}(\underline{u}) = E[\exp(\underline{u}^t\underline{X})] \quad , \quad \underline{u} \in I\!R^n \quad .$$

We say the m.g.f. exists in a neighborhood of $\underline{0}$ *iff there is an $\epsilon > 0$ such that $\psi_{\underline{X}}(\underline{u}) < \infty$ for all $\underline{u} \in I\!R^n$ such that $\|\underline{u}\| < \epsilon$.*

□

The ch.f. is defined and finite for all $\underline{u} \in I\!R^n$, since it is the expectation of a bounded continuous function (or more simply, its real and imaginary components are bounded and continuous). In fact, $|\phi_{\underline{X}}(\underline{u})| \leq 1$ for all $\underline{u} \in I\!R^n$ since $|\exp(it)|$ ≤ 1 for all $t \in I\!R$ (Exercise 2.2.1). The m.g.f. is defined for all $\underline{u}$ but may be ∞ everywhere except $\underline{u} = \underline{0}$. Many of the results for ch.f.'s given in Chung or Ash for r.v.'s carry over to random vectors as well, and also to the m.g.f. Some of the results of most interest to us are in the next proposition. Further discussion and proofs may be found in Billinsgley, pp. 352-356.

Theorem 2.2.1 *Let $\underline{X}$ be a random n–vector with ch.f. ϕ and m.g.f. ψ.*

(a) **(Continuity)** *ϕ is uniformly continuous on $I\!R^n$, and ψ is continuous at every point $\underline{u}$ such that $\psi(\underline{v}) < \infty$ for all $\underline{v}$ in a neighborhood of $\underline{u}$.*

(b) **(Relation to moments)** *If $\underline{X}$ is integrable, then the gradient*

$$\nabla\phi = \left(\frac{\partial\phi}{\partial u_1}, \frac{\partial\phi}{\partial u_2}, ..., \frac{\partial\phi}{\partial u_n} \right) \quad ,$$

is defined at $\underline{u} = \underline{0}$ and equals $iE[\underline{X}]$. Also, $\underline{X}$ has finite second moments iff the Hessian $D^2\phi(\underline{u}) = H(\underline{u})$ of ϕ exists at $\underline{u} = \underline{0}$ and then $H(\underline{0}) = -E[\underline{X}\underline{X}^t]$.

If ψ is finite in a neighborhood of $\underline{0}$, then $E[\|\underline{X}\|^p] < \infty$ for all $p \geq 1$. Further, $\nabla\psi(\underline{0}) = E[\underline{X}]$, and $D^2\psi(\underline{0}) = E[\underline{X}\underline{X}^t]$.

(c) **(Linear Transformation Formulae.)** *Let $\underline{Y} = A\underline{X} + \underline{b}$ for some $m \times n$ matrix A and some m–vector $\underline{b}$. Then for all $\underline{v} \in I\!\!R^m$,*

$$\begin{aligned} \phi_{\underline{Y}}(\underline{v}) &= \exp(i\underline{v}^t\underline{b})\phi_{\underline{X}}(A^t\underline{v}) \\ \psi_{\underline{Y}}(\underline{v}) &= \exp(\underline{v}^t\underline{b})\psi_{\underline{X}}(A^t\underline{v}) \quad . \end{aligned}$$

(d) **(Uniqueness.)** *If $\underline{Y}$ is a random n-vector and if $\phi_{\underline{X}}(\underline{u}) = \phi_{\underline{Y}}(\underline{u})$ for all $\underline{u} \in I\!\!R^n$, then Law$[\underline{X}]$ = Law$[\underline{Y}]$. If both $\psi_{\underline{X}}$ and $\psi_{\underline{Y}}$ are defined and equal in a neighborhood of $\underline{0}$, then Law$[\underline{X}]$ = Law$[\underline{Y}]$.*

(e) **Ch.f. for Sums of Independent R.V.'s** *Suppose $\underline{X}$ and $\underline{Y}$ are independent random p-vectors and let $\underline{Z} = \underline{X} + \underline{Y}$. Then*

$$\phi_{\underline{Z}}(\underline{u}) = \phi_{\underline{X}}(\underline{u})\phi_{\underline{Y}}(\underline{u}).$$

Partial Proof. (b) For the second part of (b), assume the m.g.f. is defined in a neighborhood of $\underline{0}$, say

$$\psi(\underline{u}) < \infty \text{ for } \|\underline{u}\| < \epsilon \quad . \tag{2.32}$$

It suffices to prove the result for $p \geq 2$ (see Exercise 2.1.5). For $E[\|\underline{X}\|^p] < \infty$ when $p \geq 2$, it suffices for $E[|X_i|^p] < \infty$, $1 \leq i \leq n$, since by convexity

$$\begin{aligned} \|\underline{X}\|^p &= n^{p/2}\left(\frac{1}{n}\sum_{i=1}^n X_i^2\right)^{p/2} \qquad (2.33) \\ &\leq n^{p/2}\frac{1}{n}\sum_{i=1}^n (X_i^2)^{p/2} \\ &= n^{p/2-1}\sum_{i=1}^n |X_i|^p \quad . \end{aligned}$$

Now $\psi_{X_i}(\pm\epsilon/2) = E[\exp(\pm\epsilon X_i/2)] < \infty$ by taking $\underline{u} = (\pm\epsilon/2,\ 0, 0,\ ...,\ 0)$ in (2.32). Since exponential functions "grow faster" than power functions, there is some $M > 0$ such that $|x|^p \leq \exp(\epsilon|x|/2)$ for all $|x| > M$. Hence,

$$\begin{aligned} E[|X_i|^p] &= \int_{-\infty}^{\infty} |x|^p \, dP_{X_i}(x) \\ &\leq \int_{-\infty}^{-M} \exp[-\epsilon|x|/2] \, dP_{X_i}(x) + \int_{-M}^{+M} |x|^p \, dP_{X_i}(x) + \\ &\qquad \int_{+M}^{\infty} \exp[\epsilon|x|/2] \, dP_{X_i}(x) \\ &\leq \psi_{X_i}(-\epsilon/2) + M^p + \psi_{X_i}(\epsilon/2) < \infty \quad . \end{aligned}$$

We show that $\partial\psi/\partial u_1$ exists and can be computed by differentiation under the integral sign. (An extension of this argument will show that ψ has partial derivatives of *all* orders on the interior of $\{\underline{u} : \psi(\underline{u}) < \infty\}$, and can be computed by differentiation under the integral sign.) For simplicity, assume $n = 2$. Fix u_2 and let

$$\delta \;=\; \sqrt{\epsilon^2 - u_2^2} \quad .$$

Then for $|u_1| < \delta$, $\|\underline{u}\|^2 = u_1^2 + u_2^2 < \epsilon^2$. Now take any $\delta_0 < \delta$ and $\delta_1 \in (\delta_0, \delta)$. Put

$$g(\underline{x}, u_1) \;=\; \exp[u_1 x_1 + u_2 x_2] \quad .$$

Then

$$\frac{\partial}{\partial u_1} g(\underline{x}, u_1) \;=\; x_1 e^{u_1 x_1 + u_2 x_2} \quad .$$

Since the exponential $\exp[(\delta_1 - \delta_0)|x_1|]$ "grows faster" than $|x_1|$ as $|x_1| \to \infty$, there is a constant $M > 0$ such that

$$|x_1| \;\leq\; M e^{(\delta_1 - \delta_0)|x_1|}$$

for all x_1. Also, if $|u_1| < \delta_0$, then

$$0 \;<\; e^{u_1 x_1} < e^{\delta_0 |x_1|} \quad .$$

Combining these last two estimates we have

$$\begin{aligned} \left| \frac{\partial}{\partial u_1} g(\underline{x}, u_1) \right| \;&=\; |x_1 e^{u_1 x_1 + u_2 x_2}| \\ &\leq\; (M e^{(\delta_1 - \delta_0)|x_1|}) e^{\delta_0 |x_1|} e^{u_2 x_2} \\ &\leq\; G(\underline{x}) \end{aligned}$$

where

$$G(\underline{x}) \;=\; M(e^{\delta_1 x_1 + u_2 x_2} \;+\; e^{-\delta_1 x_1 + u_2 x_2}) \quad .$$

We have used the fact that $e^{a|t|} \leq e^{at} + e^{-at}$ for all $a > 0$ and all $t \in I\!R$ in choosing a dominating function G. Since $\delta_0 < \epsilon$ and $\delta_1^2 + u_2^2 < \epsilon^2$, we have

$$\begin{aligned} \int_{R^2} G(\underline{x})\, dP_{\underline{X}}(\underline{x}) \;&=\; M[\psi(\delta_1, u_2) + \psi(-\delta_1, u_2)] \\ &<\; \infty \quad . \end{aligned}$$

Theorem 1.2.11 applies with $\theta = u_1$ and the open interval $(a, b) = (-\delta_0, \delta_0)$. Also, $d\mu(\omega) = dP_{\underline{X}}(\underline{x})$. (Note that u_2 is fixed throughout this argument. Also, we have chosen a convenient dominating function $G(\underline{x})$ whose integral is easy to bound using the m.g.f.) We obtain then from Theorem 1.2.11,

$$\begin{aligned} \frac{\partial}{\partial u_1} \psi(\underline{u}) \;&=\; \frac{\partial}{\partial u_1} \int g(\underline{x}, u_1) dP_{\underline{X}}(\underline{x}) \\ &=\; \int \frac{\partial}{\partial u_1} g(\underline{x}, u_1) dP_{\underline{X}}(\underline{x}) \\ &=\; \int x_1 e^{x_1 u_1 + x_2 u_2} dP_{\underline{X}}(\underline{x}) \quad . \end{aligned}$$

Hence,

$$\frac{\partial \psi}{\partial u_1}\Big|_{\underline{u}=0} \;=\; \int x_1 dP_{\underline{X}}(\underline{x}) \;=\; E[X_1] \quad .$$

This shows one component of the equation $\nabla \psi(\underline{0}) = E[\underline{X}]$, and the others follow similarly. A similar argument shows

$$\frac{\partial^2 \psi}{\partial u_i \partial u_j} \;=\; \int x_i x_j \exp[\underline{u}^t \underline{x}] dP_{\underline{X}}(\underline{x})$$

so the Hessian at $\underline{u} = \underline{0}$ is $E[\underline{X}\underline{X}^t]$.

□

A slight extension of the argument above can be used to prove the following theorem, which has many applications in statistics. See Theorem 9, pp. 52-54 of Lehmann's *Testing* book for the complete proof. If $z = x+iy$ is a complex number with $x \in I\!\!R$ and $y \in I\!\!R$, then $x = \text{Real}[z]$ and $y = \text{Imag}[z]$ are called the *real* and *imaginary parts*, respectively. If $\underline{z} \in \mathbb{C}^n$, i.e. $\underline{z}$ is an n–tuple of complex numbers (or an n–vector with complex components), say $\underline{z} = (z_1, \ldots, z_n)$, then Real[$\underline{z}$] = (Real[$z_1$], ..., Real[$z_n$]) is the vector of real parts, and similarly for Imag[$\underline{z}$]. Recall that for $D \subset I\!\!R^n$, the *interior* of D is $\text{int}[D] = \{\underline{x} \in D : B(\underline{x}, \epsilon) \subset D$ for some $\epsilon > 0\}$, i.e. the points $\underline{x}$ in D for which an entire neighborhood $B(\underline{x}, \epsilon)$ of $\underline{x}$ (otherwise known as an ϵ ball centered at $\underline{x}$) is contained in D. One can easily show that int[D] is the largest open subset of D.

Now we briefly review some complex analysis. A complex valued function g of a complex variable (i.e. $g : \mathbb{C} \longrightarrow \mathbb{C}$) is *analytic* at $z \in \mathbb{C}$ iff it is differentiable in a neighborhhood of z. One remarkable result from complex analysis is that a function which is analytic in an open set of $\mathbb{C}$ is in fact infinitely differentiable in that open set. (See e.g. Ahlfors, *Complex Analysis*, pp. 120-122.) (Here, an open subset of $\mathbb{C}$ is the same as an open subset of $I\!\!R^2$ when we identify $\mathbb{C}$ with $I\!\!R^2$ via $x + iy \leftrightarrow (x, y)$. We will mainly consider a "strip" of the form $\{x + iy : -\epsilon < x < \epsilon,\ -\infty < y < \infty\} = \{z \in \mathbb{C} : |\text{Real}(z)| < \epsilon\}$.)

Theorem 2.2.2 *Suppose $f : \Omega \longrightarrow \mathbb{C}$ is any bounded Borel function on a measure space $(\Omega, \mathcal{F}, \mu)$. Let $\underline{T} : (\Omega, \mathcal{F}) \longrightarrow (I\!\!R^n, \mathcal{B}_n)$ and let $\underline{\theta} \in \mathbb{C}^n$. Let*

$$B(\underline{\theta}) \;=\; \int_\Omega f(\omega) \exp[\underline{\theta}^t \underline{T}(\omega)]\, d\mu(\omega) \quad .$$

For $1 \leq j \leq n$ and $(\xi_1, \ldots, \xi_{j-1}, \xi_{j+1}, \ldots, \xi_n) \in I\!\!R^j \times I\!\!R^{n-j-1}$, define the set

$$W_j(\xi_1, \ldots, \xi_{j-1}, \xi_{j+1}, \ldots, \xi_n) \;=$$

$$\{\xi_j \in I\!\!R : \int_\Omega |f(\omega)| \exp[\underline{\xi}^t \underline{T}(\omega)]\, d\mu(\omega) \;<\; \infty\} \quad ,$$

where $\underline{\xi} = (\xi_1, ..., \xi_{j-1}, \xi_j, \xi_{j+1}, ..., \xi_n) \in I\!R^n$ *in the above. If* $\theta_k \in W$ *are fixed for* $k \neq j$*, then* B *is an analytic function in* $\{\theta_j : Real[\theta_j] \in int[W_j]\}$*, where* W_j *is as given above with* $\xi_k = Real[\theta_k]$ *for* $k \neq j$*. Further, any order partial derivative of* B *can be computed by differentiation under the integral sign.*

□

Remarks 2.2.1 (a) The fact that Real$[\theta_j] \in$ int$[W_j]$ allows us to use a dominating function as in the proof of Theorem 2.2.1 (b) above.

(b) Another remarkable fact from complex analysis is the following: Suppose f and g are both analytic functions in the open strip $\{z \in C\!\!\!\!C : |\text{Real}(z)| < \epsilon\}$, and that $\{z_n : n \in I\!N\}$ is an infinite sequence of distinct values which converges to a limit in the strip, say $z_n \to z$ with $|\text{Real}(z)| < \epsilon$. Then if $g(z_n) = f(z_n)$ for all n, we have $f = g$ everywhere on the strip.

Now suppose X is a r.v. with $\psi_X(u) < \infty$ for all $|u| < \epsilon$. Then ψ_X can be extended to an analytic function in the strip $\{z : |\text{Real}[z]| < \epsilon\}$, which contains the imaginary axis. (This is an example of analytic continuation, which is discussed at length in Ahlfors, p. 285 ff.) Hence, $\phi_X(u) = \psi_X(iu)$ by the previous theorem. Note that under these conditions, it is possible to obtain a stronger uniqueness condition than in Theorem 2.2.1 (d), namely if both $\psi_X(u) < \infty$ and $\psi_Y(u) < \infty$ for all $|u| < \epsilon$, and $\psi_X(z_n) = \psi_Y(z_n)$ for any distinct sequence of complex numbers in the strip $\{z : |\text{Real}[z]| < \epsilon\}$ with a limit in that strip, then Law$[X]$ = Law$[Y]$.

(c) Another useful fact about analytic functions is that they can be expanded in power series, i.e. suppose g is a complex function of a complex variable and $\rho > 0$ is such that g is analytic in the disk (or "ball") $\{z \in C\!\!\!\!C : |z - z_0| < \rho\}$. Then

$$g(z) \;=\; \sum_{j=0}^{\infty} \frac{1}{j!} g^{(j)}(z_0)(z - z_0)^j \quad , \text{ for } |z - z_0| < \rho \quad .$$

Further, derivatives of g may be computed by differentiating under the summation sign, for $|z - z_0| < \rho$. Using this fact along with Theorem 2.2.2, one can show that if X is a r.v. with $\psi_X(u) < \infty$ for all $|u| < \epsilon$, then

$$\begin{aligned} \psi_X(u) \;&=\; \sum_{r=0}^{\infty} \frac{d^r \psi_X}{du^r}(0)\, \frac{1}{r!}\, u^r \qquad (2.34) \\ &=\; \sum_{r=0}^{\infty} \frac{1}{r!}\, E[X^r]\, u^r \quad . \end{aligned}$$

Thus, we can read off the moments of X from the power series expansion of the m.g.f.

Now we consider the multivariate case with a random n-vector $\underline{X}$. First, we will introduce some notations that will make it easier to present the material. An n–vector $\underline{r}$ with nonnegative integer components is called a *multi–index* i.e.

$\underline{r} = (r_1, r_2, \ldots, r_n)$ with each $r_i \in I\!N$. We can use a "vector" exponential notation for a monomial as in

$$\underline{x}^{\underline{r}} \; = \; \prod_{j=1}^{n} x_j^{r_j} \quad ,$$

where $\underline{x} \in I\!R^n$. Thus, by analogy with the univariate case, we may call

$$\mu_{\underline{r}} \; = \; E[\underline{X}^{\underline{r}}] \; = \; E[\prod_{j=1}^{n} X_j^{r_j}]$$

the $\underline{r}$*'th moment of the random n–vector* $\underline{X}$. The "multi-index" factorial is defined by

$$\underline{r}! \; = \; \prod_{j=1}^{n} r_j! \quad .$$

The *order* of the multi-index $\underline{r}$ is

$$|\underline{r}| \; = \; \sum_{j=1}^{n} r_j \quad .$$

We can also define an $\underline{r}$*'th order derivative* by

$$D^{\underline{r}} \; = \; \frac{\partial^{|\underline{r}|}}{\prod_{j=1}^{n} \partial u_j^{r_j}} \quad .$$

Note that this is a partial differential operator of order $|\underline{r}|$. With these notations, one can show that the power series expansion about $\underline{0}$ for a complex function g of n complex variables which is analytic in each variable is given by

$$g(\underline{z}) \; = \; \sum_{\underline{r}} \frac{1}{\underline{r}!} D^{\underline{r}} g(\underline{0})\, \underline{z}^{\underline{r}} \quad .$$

Thus, if $\underline{X}$ is a random n-vector with $\psi_{\underline{X}}(\underline{u}) < \infty$ for all $\|\underline{u}\| < \epsilon$, then

$$\begin{aligned} \psi_{\underline{X}}(\underline{u}) \quad &= \quad \sum_{\underline{r}} \frac{1}{\underline{r}!} D^{\underline{r}} \psi_{\underline{X}}(\underline{0})\, \underline{u}^{\underline{r}} && (2.35) \\ &= \quad \sum_{\underline{r}} \frac{1}{\underline{r}!} E[\underline{X}^{\underline{r}}]\, \underline{u}^{\underline{r}} \quad , \end{aligned}$$

where the series converges in a neighborhood of $\underline{u} = \underline{0}$.

(d) Let $\underline{X}$ and $\psi_{\underline{X}}$ be as in part (b). Consider the *cumulant generating function* given by

$$K(\underline{u}) \; = \; \log\, \psi_{\underline{X}}(\underline{u}) \quad . \qquad (2.36)$$

Then the $\underline{r}$'th cumulant of $\underline{X}$ is

$$\kappa_{\underline{r}} \; = \; \frac{\partial^{|\underline{r}|} K}{\prod_{j=1}^{n} \partial u_j^{r_j}}(\underline{0}) \; = \; D^{\underline{r}} K(\underline{0}) \quad . \qquad (2.37)$$

One can show by comparison of the terms of the power series that if $n = 1$ (Exercise 2.2.3), then

$$\kappa_0 \;=\; 0 \quad , \kappa_1 \;=\; E[X] \quad , \kappa_2 \;=\; \mathrm{Var}[X] \quad . \tag{2.38}$$

For higher dimensional random vectors, we still have $\kappa_{\underline{0}} = 0$, and

$$E[X_i] \;=\; \kappa_{\underline{r}} \quad , \text{ with } r_i = 1 \text{ and } r_j = 0 \text{ if } j \neq i \quad . \tag{2.39}$$

Also, if $V = \mathrm{Cov}[\underline{X}]$, then

$$V_{ij} \;=\; \kappa_{\underline{r}} \quad , \text{ with } r_i = r_j = 1 \text{ and } r_k = 0 \text{ if } k \neq i \text{ or } k \neq j \quad . \tag{2.40}$$

See Exercise 2.2.4.

□

Exercises for Section 2.2.

2.2.1 Let $f : (\Omega, \mathcal{F}, \mu) \longrightarrow (I\!\!R^2, \mathcal{B}_2)$ and define a complex valued function $g : \Omega \longrightarrow \mathbb{C}$ by $g(\omega) = f_1(\omega) + if_2(\omega)$. The modulus of g is $|g| = \sqrt{f_1^2 + f_2^2}$. Also, $\int gd\mu = \int f_1 d\mu + i \int f_2 d\mu$. Show that $|\int gd\mu| \leq \int |g|d\mu$. Conclude that the ch.f. is bounded by 1 in modulus.

2.2.2 Prove Theorem 2.2.1 (c).

2.2.3 Verify equation (2.38).

2.2.4 Verify equations (2.39) and (2.40).

2.2.5 Let X be a r.v. with values in $I\!\!N$ and define the *probability generating function* (abbreviated p.g.f.)

$$\gamma(z) = \sum_{n \in I\!\!N} P[X = n]\, z^n \quad .$$

(a) Show that the series defining γ converges absolutely for complex numbers z with $|z| \leq 1$.

(b) Give formulae relating the ch.f. and m.g.f. to the p.g.f.

(c) Show that the m.g.f. for X is finite in a neighborhood of the origin if and only if there is some real $x > 1$ where the $\gamma < \infty$.

(d) Under what circumstances and by what formulae can one recover the moments of X from γ?

2.3 Common Distributions Used in Statistics.

In this section, we introduce some of the commonly used families of distributions. We assume throughout this section that we will observe a random element X (called the *observable*) taking values in a measurable *observation space* $(\Xi, \mathcal{G})$. In fact, X will almost always be a random n–vector.

Definition 2.3.1 *Let* $\mathbf{P}$ *be a family of probability measures on* $(\Xi, \mathcal{G})$*. We say* $\mathbf{P}$ *is a* parametric family *with* parameter θ *and* parameter space Θ *if for each* θ *there is a unique* $P_\theta \in \mathbf{P}$ *associated with* θ*. The mapping* $\theta \mapsto P_\theta$ *is called the* parameterization*.*

□

In some sense, the parameterization is just a way of labelling a probability distribution. Notice that the parameter space Θ is not assumed to have any "structure" (e.g. like being a measurable space). One can always "parameterize" a family of probability measures: given a family of probability measures $\mathbf{P}$ let $\theta = P$ for $P \in \mathbf{P}$. Most often, Θ will be a nice subset of $I\!R^p$ for some p and the parameterization will be natural or convenient.

Definition 2.3.2 *A family* $\{P_\theta : \theta \in \Theta\}$ *is called* identifiable *iff* $\theta_1 \neq \theta_2$ *implies* $P_{\theta_1} \neq P_{\theta_2}$*.*

□

Note that $P_{\theta_1} \neq P_{\theta_2}$ means for some measurable set A, $P_{\theta_1}(A) \neq P_{\theta_2}(A)$. In general, we want to use only identifiable parameterizations. If the parameter is not identifiable there will be differences in parameter values which are not statistically meaningful since we cannot determine them from the distribution of the observable. In general, if we have a nonidentifiable parameterization, we will reparameterize to obtain identifiability.

Definition 2.3.3 *Suppose* μ *is a* σ*–finite measure on* $(\Xi, \mathcal{G})$ *and* $\mathbf{P}$ *is a family of probability measures on* $(\Xi, \mathcal{G})$*. Then we say* $\mathbf{P}$ *is* dominated *by* μ *and write* $\mathbf{P} \ll \mu$ *iff every* $P \in \mathbf{P}$ *satisfies* $P \ll \mu$*. Assuming* $\mathbf{P} = \{P_\theta : \theta \in \Theta\}$*, we will write*

$$f_\theta(x) \ = \ \frac{dP_\theta}{d\mu}(x)$$

for the Radon-Nikodym derivative of P_θ *w.r.t.* μ*.*

□

Typically we will deal with families dominated by either Lebesgue measure on $I\!R^n$ (the so called "continuous distributions" from more elementary courses) or by counting measure on some $I\!N^n$ (the so called "discrete distributions").

2.3.1 Exponential Families.

Next we introduce the perhaps the most important general class of statistical models.

Definition 2.3.4 *A dominated family* $\mathbf{P} = \{P_\theta : \theta \in \Theta\} \ll \mu$ *on* $(\Xi, \mathcal{G})$ *with* μ *σ–finite is called an* exponential family *iff the densities w.r.t.* μ *can be written in the form*

$$f_\theta(x) \;=\; \exp[\underline{\eta}(\theta)^t \underline{T}(x) - B(\theta)]h(x) \tag{2.41}$$

where $\underline{T} : (\Xi, \mathcal{G}) \longrightarrow (I\!\!R^p, \mathcal{B}_p)$, $\underline{\eta} : \Theta \longrightarrow I\!\!R^p$, $h : (\Xi, \mathcal{G}) \longrightarrow (I\!\!R, \mathcal{B})$, *and* $B : \Theta \longrightarrow I\!\!R$ *is the "normalizing constant" in logarithmic form*

$$B(\theta) \;=\; \log \int_\Xi \exp[\underline{\eta}(\theta)^t \underline{T}(x)] h(x)\, d\mu(x) \; .$$

It is specifically required that $B(\theta)$ *be finite.*

□

Example 2.3.1 Let $\mathbf{P}$ be the normal family on $I\!\!R$, $\{N(\mu, \sigma^2) : \mu \in I\!\!R$ and $\sigma^2 > 0\}$. Then $\mu = m$, Lebesgue measure, may be taken as the dominating measure and this is σ–finite. The density may be written in the form

$$\begin{aligned} f_{\mu,\sigma^2}(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-1}{2\sigma^2}(x-\mu)^2\right] \\ &= \exp\left[\frac{-1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \left(\frac{\mu^2}{2\sigma^2} + \log\sigma\right)\right]\left(\frac{1}{\sqrt{2\pi}}\right) \end{aligned}$$

which is an exponential family with $p = 2$ and

$$\eta_1(\mu, \sigma^2) \;=\; \frac{-1}{2\sigma^2} \quad , \quad \eta_2(\mu, \sigma^2) \;=\; \frac{\mu}{\sigma^2}$$

$$T_1(x) \;=\; x^2 \quad , \quad T_2(x) \;=\; x$$

$$B(\mu, \sigma^2) \;=\; \frac{\mu^2}{2\sigma^2} + \log\sigma \quad , \quad h(x) \;=\; \frac{1}{\sqrt{2\pi}}$$

Notice that determining if a given family is an exponential family is simply a matter of algebraically putting the density in the right form. Also, the various components of the exponential family form are not unique except for $h(x)$ and $B(\theta)$, but even $h(x)$ can change if we change the dominating measure μ (see Remark 2.3.1 (a) below). For instance we can multiply η_1 by 2 and divide T_1 by 2 and the density remains unchanged.

□

We wish to define the multivariate generalization of the normal distribution. We will use moment generating function.

Definition 2.3.5 *Given $\underline{\mu} \in I\!R^n$ and V an $n \times n$ nonnegative definite matrix, let $N(\underline{\mu}, V)$ denote the Borel p.m. on $I\!R^n$ with m.g.f.*

$$\psi(\underline{u}) \;=\; \exp[\underline{\mu}^t \underline{u} + \frac{1}{2}\underline{u}^t V \underline{u}] \quad . \tag{2.42}$$

This p.m. is called the (Multivariate) Normal Distribution *with mean $\underline{\mu}$ and covariance V.*

□

Of course, there is the question of whether or not a p.m. exists with the given m.g.f. An m.g.f. is a very special kind of function (it satisfies more properties than just being analytic), so one cannot in general write down any old function and say it is the m.g.f. of a probability distribution. Assuming the student is familiar with the normal distribution in one dimension, it is not hard to show that there is a p.m. with m.g.f. as given above, and it is unique. Further, the mean and the covariance are $\underline{\mu}$ and V, respectively. It is convenient to introduce the normal distribution this way rather than through a Lebesgue density since if V is not positive definite then a Lebesgue density does not exist. If V is strictly positive definite then the Lebesgue density is given by

$$f(\underline{x}) \;=\; \frac{1}{(2\pi)^{n/2}\det(V)^{1/2}} \exp[-\frac{1}{2}(\underline{x} - \underline{\mu})^t V^{-1}(\underline{x} - \underline{\mu})] \quad . \tag{2.43}$$

When the covariance matrix is nonsingular (so the Lebesgue density exists) then we say the corresponding multivariate normal distribution is nonsingular, and otherwise we say the multivariate normal distribution is singular (as it is singular w.r.t. Lebesgue measure). In Exercises 2.3.3 and 2.4.4 the student is asked to verify these claims, and to show that the nonsingular multivariate normal distributions for a given dimension form an exponential family.

Example 2.3.2 For $\alpha > 0$ and $\beta > 0$, the *Gamma*(α, β) density is given by

$$f_{\alpha\beta}(x) \;=\; \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} \quad , \quad x > 0$$

w.r.t. Lebesgue measure. Here,

$$\Gamma(\alpha) \;=\; \int_0^\infty x^{\alpha-1} e^{-x}\, dx$$

is a well known special function. One can put the density in the form

$$f_{\alpha\beta}(x) \;=\; \exp\left\{ \alpha \log(x) \;+\; \frac{-1}{\beta} x \;-\; [\,\alpha \log\beta + \log\Gamma(\alpha)\,] \right\} \left[x^{-1} I_{(0,\infty)}(x) \right]$$

which shows that the family **Gamma** $= \{Gamma(\alpha, \beta) : \alpha > 0 \text{ and } \beta > 0 \ \}$ is an exponential family with $p = 2$ and

$$\eta_1(\alpha,\beta) \ = \ \alpha \quad , \quad \eta_2(\alpha,\beta) \ = \ \frac{-1}{\beta}$$

$$T_1(x) \ = \ \log(x) \quad , \quad T_2(x) \ = \ x$$

$$B(\alpha,\beta) \ = \ \alpha\log\beta + \log\Gamma(\alpha) \quad , \quad h(x) \ = \ x^{-1}I_{(0,\infty)}(x)$$

A subfamily of **Gamma** which is also an exponential family is $\{Exp(\beta) : \beta > 0\}$ $= \{Gamma(1,\beta) : \beta > 0\}$ where the $Exp(\beta)$ distribution has Lebesgue density

$$f_\beta(x) \ = \ \beta^{-1}e^{-x/\beta} \quad , \quad x > 0.$$

The family of distributions $\{Exp(\beta) : \beta > 0\}$ is an *exponential subfamily* of the Gamma family. We will call this subfamily the family of exponential distributions.

□

Remarks 2.3.1 Here we develop some of the elementary properties of exponential families.

(a) Suppose **P** is an exponential family dominated by μ σ–finite with densities as given in (2.41). Let X_1, X_2, ..., X_n be i.i.d. with common distribution from **P**. Then $\underline{X} = (X_1,\ X_2,\ ...,\ X_n)$ has a distribution from an exponential family dominated by μ^n and with densities

$$f_\theta(\underline{x}) \ = \ \exp\left[\underline{\eta}(\theta)^t\underline{\tilde{T}}(\underline{x}) \ - \ \tilde{B}(\theta)\right]\tilde{h}(\underline{x})$$

where

$$\underline{\tilde{T}}(\underline{x}) \ = \ \sum_{i=1}^{n} T(x_i)$$

$$\tilde{B}(\theta) \ = \ nB(\theta)$$

$$\tilde{h}(\underline{x}) \ = \ \prod_{i=1}^{n} h(x_i)$$

which is easily verified from the product formula for the joint density.

(b) By changing dominating measures, we can take $h(x) \equiv 1$, i.e. eliminate the factor $h(x)$ in (2.41). Define a new dominating measure ν by

$$\frac{d\nu}{d\mu} \ = \ h \ . \tag{2.44}$$

We claim the ν is σ-finite. Fix $\theta \in \Theta$. Define

$$B_m \ = \ \{x : \exp[\underline{\eta}(\theta)^t\underline{T}(x) - B(\theta)] \geq 1/m\},$$

for $m = 1, 2, \ldots$. As the exponential function is always positive, we have $\bigcup_m B_m = \Xi$, so

$$\begin{aligned} \nu(B_m) &= \int_{B_m} h(x)\, d\mu(x) \\ &\leq m \cdot \int_{B_m} \exp[\underline{\eta}(\theta)^t \underline{T}(x) - B(\theta)] h(x)\, d\mu(x) \\ &\leq mP_\theta(B) \\ &\leq m. \end{aligned}$$

Now, for all θ and A, measurable,

$$\begin{aligned} P_\theta(A) &= \int_A \exp[\underline{\eta}(\theta)^t \underline{T}(x) - B(\theta)] h(x)\, d\mu(x) \\ &= \int_A \exp[\underline{\eta}(\theta)^t \underline{T}(x) - B(\theta)]\, d\nu(x) \end{aligned}$$

by Proposition 1.4.2 (a), so $P_\theta \ll \nu$ we have

$$\frac{dP_\theta}{d\nu}(x) = \exp[\underline{\eta}(\theta)^t \underline{T}(x) - B(\theta)], \tag{2.45}$$

which is an exponential family with $h \equiv 1$. For convenience, we will often delete the factor $h(x)$ when writing the density in an exponential family, although the student should recall that it may be present when initially writing a density in exponential family form. The student may wish to determine the new dominating measure ν in the previous examples which causes h to disappear.

(c) Note that the density in (2.45) is strictly positive, so we conclude that in general, the region where the density is positive is not dependent on the parameter θ. A family we consider below which is related to the $Exp(\beta)$ family is the $Exp[a, b]$ family. An $Exp[a, b]$ distribution has Lebesgue density given by

$$f_{ab}(x) = a^{-1} e^{-(x-b)/a} \quad , \quad x \geq b.$$

Here the parameter a is required to be positive and b is an arbitrary real number. Since the density is positive exactly on the set $[b, \infty)$, it follows that this family is not an exponential family.

(d) Remark (b) indicates one simplification of formula (2.41), and we now investigate a more subtle simplification via reparameterization. Let $\Lambda = \underline{\eta}(\Theta)$, then $\Lambda \subset I\!R^p$ and $\underline{\eta} \in \Lambda$ may be used as parameter rather than θ since the actual probability measure only depends on $\underline{\eta}^t \underline{T}(x)$. Thus, we may write

$$f_{\underline{\eta}}(x) = \exp\left[\underline{\eta}^t \underline{T}(x) - A(\underline{\eta}) \right]. \tag{2.46}$$

Keep in mind that $A(\underline{\eta})$ is just a normalizing constant so it can be computed as

$$A(\underline{\eta}) = \log\left(\int_\Xi \exp[\underline{\eta}^t \underline{T}(x)]\, d\mu(x) \right). \tag{2.47}$$

We have implicitly assumed that $h \equiv 1$ as in remark (b) above. The form of the exponential family given in (2.46) is called the *canonical form.* The new parameter $\underline{\eta}$ is called the *natural parameter.* If the natural parameterization is used, then we call the family a *natural parameter exponential family* or a *canonical form exponential family.* It is of course required that $A(\underline{\eta})$ be finite for (2.46) to define a probability density, but it is obviously also sufficient. The *natural parameter space* Λ_0 is the largest possible parameter space for the natural parameter, viz.

$$\Lambda_0 \;=\; \{\, \underline{\eta} \in \mathbb{R}^p \;:\; 0 \;<\; \int_{\Xi} \exp[\underline{\eta}^t \underline{T}(x)] h(x)\, d\mu(x) \;<\; \infty \,\} \tag{2.48}$$

$$= \; \{\, \underline{\eta} \;:\; -\infty \;<\; A(\underline{\eta}) \;<\; \infty \,\}$$

An exponential family in canonical form with the natural parameter space is called a *natural exponential family.*

(e) The canonical form representation is not unique. Indeed, let D be any nonsingular $p \times p$ matrix and put

$$\tilde{\underline{\eta}} \;=\; (D^{-1})^t \underline{\eta} \quad , \quad \tilde{\underline{T}}(x) \;=\; D\underline{T}(x) \;,$$

then

$$\underline{\eta}^t \underline{T} \;=\; \underline{\eta}^t D^{-1} D \underline{T} \;=\; [(D^{-1})^t \underline{\eta}]^t (D\underline{T}) \;=\; \tilde{\underline{\eta}}^t \tilde{\underline{T}}$$

and we may use the new parameter $\tilde{\underline{\eta}}$ in place of $\underline{\eta}$ provided we switch from $\underline{T}$ to $\tilde{\underline{T}}(x)$.

(f) If the parameter space $\Lambda \subset M$ where M is a linear manifold in $\mathbb{R}^p$ with $\dim(M) = q < p$, then the natural parameter satisfies $p - q$ independent linear constraints. (To wit, $C^t \underline{\eta} = C^t \underline{\zeta}$ where C is a $p \times (p-q)$ matrix with columns orthogonal to $M - \underline{\zeta}$ and $\underline{\zeta}$ is any element of M. Note that $M - \underline{\zeta}$ is a q-dimensional linear subspace since it contains $\underline{0}$, and is the unique such subspace parallel to M.) Then there is a $p \times q$ matrix B such that for any $\underline{\zeta} \in M$, there is for each $\underline{\eta} \in \Lambda$ a unique $\tilde{\underline{\eta}} \in \mathbb{R}^q$ such that $\underline{\eta} = B\tilde{\underline{\eta}} + \zeta$, and we will denote by $\tilde{\Lambda}$ the set of all such $\tilde{\underline{\eta}}$. (Here, B may be taken as any matrix whose columns span $M - \zeta$, and then the entries in $\tilde{\underline{\eta}}$ are just the coefficients in the expansion of $\underline{\eta} - \underline{\zeta}$ using the basis consisting of the columns of M.) Then $\underline{\eta}^t \underline{T} = \tilde{\underline{\eta}}^t (B^t \underline{T}) + \underline{\zeta}^t \underline{T}$, so

$$f_{\underline{\eta}}(x) \;=\; \exp[\underline{\eta}^t \underline{T}(x) - A(\underline{\eta})]$$

$$= \; \exp[\tilde{\underline{\eta}}^t \tilde{\underline{T}}(x) - \tilde{A}(\tilde{\underline{\eta}})]\, \tilde{h}(x)$$

where $\tilde{\underline{T}}(x) = B^t \underline{T}(x) \in \mathbb{R}^q$, $\tilde{A}(\tilde{\underline{\eta}}) = A(B\tilde{\underline{\eta}} + \underline{\zeta})$, and $\tilde{h}(x) = \exp[\underline{\zeta}^t \underline{T}(x)]$. Note that $\tilde{\underline{\eta}}$ does not appear in $\tilde{h}(x)$. Thus, we may reparameterize and reduce the dimension of $\underline{\eta}$ and $\underline{T}$ so that Λ does not belong to any proper linear manifold.

Similarly, suppose $\underline{T}$ satisfies some linear constraints, i.e. if there is a linear manifold M of dimension $q < p$,

$$P_\eta\{x : \underline{T}(x) \in M\} \;=\; 1$$

or, what is the same, there is a $p \times (p-q)$ matrix C and a $\underline{\zeta} \in I\!R^p$ such that

$$P_\eta\{x : C^t \underline{T}(x) = C^t \underline{\zeta}\} \;=\; 1 \;.$$

Note that if this happens for one $\underline{\eta}$ then it happens for all $\underline{\eta}$ since the set where $f_\eta > 0$ doesn't depend on η. Now let B be a $p \times q$ matrix with columns spanning M and $\underline{\tau} \in M$, then $T = B\tilde{T} + \tau$ for a unique $\tilde{T} \in I\!R^q$, and we may reparameterize with $\tilde{\eta} = B^t\eta$ and reduce dimensionality again and $\tilde{T}$ will not satisfy any linear constraints (i.e. not be confined to a proper linear manifold in $I\!R^q$). Note that even though η was not constrained here, we have lost nothing since if $(\eta_1 - \eta_2)$ is orthogonal to $M - \tau$, we have $\eta_1^t T = \eta_2^t T$ μ–a.e. where μ is the dominating measure, i.e. the original parameterization was not identifiable.

In conclusion, we can always reduce an exponential family in canonical form so that neither the parameter η nor the T satisfies any linear constraints. When a canonical form exponential family is such that neither η nor T satisfies any linear constraints, we say the family is *minimal*. If the parameter space of a minimal exponential family (in canonical form) has nonempty interior (i.e. the parameter space contains a nonempty open set, such as an open rectangle $(a_1, b_1) \times (a_2, b_2) \times ... \times (a_p, b_p)$ where $a_i < b_i$ for $1 \le i \le p$), then the family is said to be of *full rank.*

□

The following definition and result will be useful several times in the course of our study.

Definition 2.3.6 *Let A be an $n \times m$ matrix with entries A_{ij}, $1 \le i \le n$, $1 \le j \le m$. The* Frobenius norm *of A is*

$$\|A\| \;=\; \left(\sum_{i=1}^{n} \sum_{j=1}^{m} A_{ij}^2 \right)^{1/2} .$$

Proposition 2.3.1 *For any matrices A and B, $\|AB\| \le \|A\|\|B\|$, provided AB is defined. In particular, if x is a vector of appropriate dimension, $\|Ax\| \le \|A\|\|x\|$.*

Proof: Exercise 2.3.2.

Proposition 2.3.2 *Suppose $\{f_\eta : \eta \in \Lambda_0\}$ is a natural exponential family which is minimal.*

(a) The natural parameter space Λ_0 is a convex subset of $I\!R^p$ and the family is full rank.

(b) If η_0 is an interior point of Λ_0 (i.e. there is some open ball $B(\eta_0, \epsilon) \subset \Lambda_0$, where the radius $\epsilon > 0$), then the m.g.f. ψ_{η_0} of $Law_{\eta_0}[T(X)]$ is finite in a neighborhood of 0 *and is given by*

$$\psi_{\eta_0}(u) \;=\; \exp\left[\, A(\eta_0 + u) \;-\; A(\eta_0) \,\right] \tag{2.49}$$

and the cumulant generating function (defined in (2.36)) is

$$\kappa_{\eta_0}(u) \;=\; A(\eta_0 + u) \;-\; A(\eta_0)$$

In particular

$$E_{\eta_0}[T(X)] \;=\; \nabla A(\eta_0) \tag{2.50}$$

$$Cov_{\eta_0}[T(X)] \;=\; D^2 A(\eta_0) \tag{2.51}$$

Furthermore, $A(\eta)$ is a strictly convex function on the interior of Λ_0.

(c) Under the same hypotheses as (b), if $\phi : \Xi \longrightarrow I\!\!R$ is such that $E_{\eta_0}|\phi(X)| < \infty$, then the function

$$h(\eta) \;=\; E_\eta[\phi(X)]$$

is finite in a neighborhood of η_0. Furthermore, h is infinitely differentiable and the derivatives may be computed by interchange of differentiation and integration.

Proof. For (a), assume η_1, $\eta_2 \in \Lambda_0$ and put $\eta = \alpha\eta_1 + (1-\alpha)\eta_2$ for some $\alpha \in [0, 1]$. The exponential function is convex, so

$$\exp[\alpha\eta_1^t T + (1-\alpha)\eta_2^t T] \;\leq\; \alpha \exp[\eta_1^t T] \;+\; (1-\alpha)\exp[\eta_2^t T]\ .$$

Taking integrals w.r.t. the dominating measure (and noting that the integrands are positive, so the integrals exist) gives

$$\int_\Xi \exp[\eta^t T(x)]\, d\mu(x) \;\leq\; \alpha \int_\Xi \exp[\eta_1^t T]\, d\mu(x) \;+\; (1-\alpha)\int_\Xi \exp[\eta_2^t T]\, d\mu(x)$$

Thus, finiteness of the two integrals on the r.h.s. implies finiteness of the integral on the l.h.s., i.e. that η is in Λ_0 and hence that Λ_0 is convex.

To show that the family is full rank, it is only necessary to show that the natural parameter space has nonempty interior (since we know by minimality that T does not satisfy any linear constraint). Since the canonical parameter η does not satisfy any linear constraints, we know that Λ_0 does not lie in a lower dimensional linear manifold. Thus, we can find $p+1$ vectors $\underline{\eta}_0$, $\underline{\eta}_1$, ..., $\underline{\eta}_p$ such that $\{\underline{\eta}_1 - \underline{\eta}_0, \underline{\eta}_2 - \underline{\eta}_0, \ldots, \underline{\eta}_p - \underline{\eta}_0\}$ forms a linear independent set of p-dimensional vectors. We will assume without loss of generality that $\underline{\eta}_0 = \underline{0}$ (by subtracting $\underline{\eta}_0$ from every other $\underline{\eta}$). Put

$$\begin{aligned} K &= \left\{ \sum_{j=0}^{p} a_j \underline{\eta}_j : a_j \geq 0, 0 \leq j \leq p,\ \&\sum_{j=0}^{p} a_j = 1 \right\} \\ \underline{\bar{\eta}} &= (p+1)\sum_{j=0}^{p} a_j \underline{\eta}_j . \end{aligned}$$

Of course, $\underline{\bar{\eta}} \in K$, and our goal is to show that

for some $\epsilon > 0$, $\|\underline{\eta} - \underline{\bar{\eta}}\| < \epsilon$ implies $\underline{\eta} \in K$, i.e. that $\bar{\underline{\eta}}$ has a neighborhood contained in K, so K has nonempty interior. Now any $\underline{\eta} \in \mathbb{R}^p$ can be written as

$$\underline{\eta} \;=\; \bar{\underline{\eta}} + \sum_{j=1}^{p} b_j \underline{\eta}_j,$$

where the $\underline{b} = (b_1, \ldots, b_p)$ can be found by solving

$$A\underline{b} \;=\; \underline{\eta} - \bar{\underline{\eta}},$$

where A is the $p \times p$ matrix with j'th column equal to $\underline{\eta}_j$. We know that A is invertible, so by Proposition 2.3.1 ,

$$\|b\| \;\leq\; \|A^{-1}\|\|\underline{\eta} - \bar{\underline{\eta}}\|.$$

Now in order to guarantee that $\underline{\eta}$ is in K we need that

$$\begin{aligned} a_j &= (p+1)^{-1} + b_j, \quad 1 \leq j \leq p, \\ a_0 &= (p+1)^{-1} - \sum_{j=1}^{p} b_j, \end{aligned}$$

are nonnegative since they already sum to 1. For this it suffices that

$$\begin{aligned} \max_{1 \leq j \leq p} |b_j| &\leq (p+1)^{-1} \\ \sum_{j=1}^{p} |b_j| &\leq (p+1)^{-1}. \end{aligned}$$

Now $\max_{1 \leq j \leq p} |b_j| \leq \|\underline{b}\|$ and using Cauchy-Schwartz, it is easy to see that

$$\sum_{j=1}^{p} |b_j| \;\leq\; p^{1/2}\|\underline{b}\|.$$

Hence, as long as we make

$$\|\underline{b}\| \;<\; \min\{p^{-1/2}(p+1)^{-1}, (p+1)^{-1}\} \;=\; p^{-1/2}(p+1)^{-1},$$

then we will satisfy our requirements on $\underline{b}$ so that $\underline{\eta} \in K$. Thus, it suffices to take

$$\epsilon \;=\; \|A^{-1}\|^{-1} p^{-1/2}(p+1)^{-1}.$$

(b) For the m.g.f. calculation, we have

$$\begin{aligned} \psi_{\eta_0}(u) &= E_{\eta_0}[\exp(u^t T(X))] \\ &= \int_{\Xi} \exp[u^t T(x) + \eta_0^t T(x) - A(\eta_0)]\, d\mu(x) \\ &= \int_{\Xi} \exp[(u+\eta_0)^t T(x) - A(u+\eta_0)\, d\mu(x) \exp[A(u+\eta_0) - A(\eta_0)] \\ &= \exp[A(u+\eta_0) - A(\eta_0)] \end{aligned}$$

where this is valid provided $\eta_0 + u$ is in Λ_0. Since there is a neighborhood of η_0 contained in Λ_0, it follows that there is a neighborhood of 0 such that if u is in this neighborhood of 0, then $\eta_0 + u$ is in the neighborhood of η_0 contained in Λ_0, and everything in the last displayed calculation is finite, i.e. ψ_{η_0} is finite in a neighborhood of 0.

The formula for κ is immediate and the formulae for the first two moments of T under η_0 follows by an elementary calculation. Since the family is minimal, T is not almost surely confined to some proper linear manifold of $I\!R^p$, so by Proposition 2.1.7, the covariance in (2.51) is full rank, i.e. positive definite. This shows that $A(\eta)$ is strictly convex by the second derivative test.

(c) We apply Theorem 2.2.2. For the bounded function $f(x)$ take

$$f(x) \;=\; I_{[0,\infty)}(\phi(x)) \;-\; I_{(-\infty,0)}(\phi(x))$$

and for the measure $\tilde{\mu}$ in Theorem 2.2.2, use

$$d\tilde{\mu}(x) \;=\; |\phi(x)| d\mu(x) \; .$$

Note that $f(x)|\phi(x)| = \phi(x)$. Then apply that theorem to

$$B(\eta) \;=\; \int_{\Xi} f(x) \exp[\eta^t T(x)] \, d\tilde{\mu}(x) \; .$$

Infinite differentiability of B at an interior point of Λ_0 implies the same for h, and the interchangeability of the differentiation and integration operators follows as well.

□

Example 2.3.3 (Multinomial Family) Suppose Ω is partitioned into k events, say A_1, A_2, ..., A_k where the A_i are mutually exclusive and their union is Ω. Let p_i be the probability of the A_i, so $\underline{p} = (p_1, \, p_2, \, ..., \, p_k)$ is a *probability vector*, i.e. has nonnegative entries which sum to 1. Let $\underline{X}$ be the random k-vector which indicates which event in the partition of Ω occurs, i.e. $\underline{X} = (I_{A_1}, \, I_{A_2}, \, ..., \, I_{A_k})$. Then the i'th entry of $\underline{X}$ is 1 if the outcome is in A_i, and the other entries of $\underline{X}$ are 0. Now let $\underline{X}_1$, $\underline{X}_2$, ..., $\underline{X}_n$ be i.i.d. with the same distribution as $\underline{X}$, and let $\underline{Y} = \sum \underline{X}_i$. Thus, Y_i is the number of times A_i occurs in the n trials. Then $\underline{Y}$ has a *multinomial distribution* with parameters $(n, \underline{p})$, written $Mult(n, \underline{p})$. As the parameter n is always known (since $\sum Y_i = n$), we only show $\underline{p}$ in the probabilities, etc. One can check (Exercise 2.3.15) that the distribution of $\underline{Y}$ is given by

$$P_{\underline{p}}[\underline{Y} = \underline{y}] \;=\; \left(\begin{array}{c} n \\ \underline{y} \end{array} \right) \underline{p}^{\underline{y}} \tag{2.52}$$

where $\underline{y}$ is a k-multi-index (a k-vector with nonnegative integer entries) satisfying

$$\sum_{i=1}^{k} y_i \;=\; n \; . \tag{2.53}$$

Also,

$$\binom{n}{\underline{y}} = \frac{n!}{\underline{y}!} = \frac{n!}{\prod_{i=1}^{k} y_i!} \tag{2.54}$$

is a *multinomial coefficient*, and

$$\underline{p}^{\underline{y}} = \prod_{i=1}^{k} p_i^{y_i} \tag{2.55}$$

is the monomial defined in Remark 2.2.1 (b). It is convenient to define the multinomial coefficient to be 0 if (2.53) fails. Now if we take as dominating measure the discrete measure

$$\mu = \sum_{\underline{y}} \binom{n}{\underline{y}} \delta_{\underline{y}}$$

then the density of $Mult(n, \underline{p})$ is

$$\begin{aligned} f_{\underline{p}}(\underline{y}) &= \underline{p}^{\underline{y}} \qquad (2.56) \\ &= \exp\left[\sum_{i=1}^{k} y_i \log(p_i)\right] , \end{aligned}$$

provided it is known $p_i \neq 0$ for all i. This is an exponential family with natural parameters $\eta_i = \log(p_i)$ and $\underline{T} = \underline{y}$, but $\underline{T}$ satisfies the linear constraint in (2.53) and the η_i satisfy the nonlinear constraint $\sum \exp[\eta_i] = 1$. There are many ways of eliminating this indeterminacy, but the most common is to use $\underline{T} = (y_1, y_2, ..., y_{k-1})$ (i.e. leave off the last component which is determinable from the other components and (2.54)), and form the *multinomial logit*

$$\eta_i = \log\left(\frac{p_i}{1 - \sum_{j=1}^{k-1} p_j}\right) \quad , \quad 1 \le i \le (k-1) \tag{2.57}$$

Note that given any probability vector $\underline{p}$ one can obtain a $(k-1)$ vector $\underline{\eta}$ from (2.57), and conversely given any $\underline{\eta} \in I\!R^{(k-1)}$, one can obtain the corresponding probability vector through

$$p_k = \frac{1}{1 + \sum_{j=1}^{k-1} \exp[\eta_j]} , \tag{2.58}$$

$$p_i = p_k \exp[\eta_i] \quad , \quad 1 \le i \le k-1 . \tag{2.59}$$

Note that the multinomial logit $\underline{\eta}$ is an unconstrained vector in $I\!R^{(k-1)}$ whereas the probability vector $\underline{p}$ is a k vector which satisfies the constraints of nonnegativity and $\sum p_i = 1$. The density in (2.56) can be written as

$$f_{\underline{p}}(\underline{y}) = \exp\left[\sum_{i=1}^{k-1} y_i \log(p_i) + (n - \sum_{i=1}^{k-1} y_i) \log(1 - \sum_{j=1}^{k-1} p_j)\right] \tag{2.60}$$

$$= \exp\left[\sum_{i=1}^{k-1} y_i(\log(p_i) - \log(1 - \sum_{j=1}^{k-1} p_j)) \;+\; n\log(1 - \sum_{j=1}^{k-1} p_j) \right]$$

$$= \exp\left[\sum_{i=1}^{k-1} y_i\eta_i \;-\; n\log(1 + \sum_{j=1}^{k-1} \exp[\eta_j]) \right]$$

which is an exponential family in canonical form with

$$A(\underline{\eta}) \;=\; n\log\left(1 + \sum_{j=1}^{k-1} \exp[\eta_j] \right) \;. \tag{2.61}$$

From this in conjunction with Proposition 2.3.2 (b), we have for $1 \leq i < k$,

$$\begin{aligned} E_{\underline{p}}[Y_i] \;&=\; \frac{n\exp[\eta_i]}{1 + \sum_{j=1}^{k-1}\exp[\eta_j]} \\ &=\; np_i \end{aligned} \tag{2.62}$$

and since $Y_k = n - \sum_{j=1}^{k-1} Y_j$,

$$E_{\underline{p}}[Y_k] \;=\; n \;-\; \sum_{j=1}^{k-1} E_{\underline{p}}[Y_j]$$

$$= n \;-\; \sum_{j=1}^{k-1} np_j \;=\; n\left[1 - \sum_{j=1}^{k-1} p_j\right] \;=\; np_k \;.$$

Also, if $1 \leq i < j \leq k$, then

$$\begin{aligned} \mathrm{Cov}_{\underline{p}}[Y_i, Y_j] \;&=\; \frac{\partial^2}{\partial\eta_i\partial\eta_j} A(\underline{\eta}) \\ &=\; \frac{-n\exp[\eta_i]\exp[\eta_j]}{\left[1 + \sum_{m=1}^{k-1}\exp[\eta_m]\right]^2} \\ &=\; np_ip_j \end{aligned} \tag{2.63}$$

and for $1 \leq i < k$

$$\begin{aligned} \mathrm{Var}_{\underline{p}}[Y_i] \;&=\; \frac{\partial^2}{\partial^2\eta_i} A(\underline{\eta}) \\ &=\; \frac{n\left[\exp[\eta_i]\left(1 + \sum_{j=1}^{k-1}\exp[\eta_j]\right) \;-\; \exp[2\eta_i]\right]}{\left[1 + \sum_{m=1}^{k-1}\exp[\eta_m]\right]^2} \\ &=\; n\left[p_i - p_i^2\right] \;=\; np_i(1 - p_i) \;. \end{aligned} \tag{2.64}$$

One can check as before that (2.63) and (2.64) hold if one of the indices is equal to k. Also, (2.62) and (2.64) are easy to see directly since Y_i is $B(n, p_i)$. One can verify (2.63) by computing the covariance of X_i and X_j where $\underline{X}$ is $Mult(1, \underline{p})$. See Exercise 2.3.16.

□

2.3.2 Location–Scale Families.

We now build ourselves up by special cases to Definition 2.3.8 below.

Definition 2.3.7 *Let P be a Borel p.m. on $I\!R$.*

(a) The location family *generated by P is $\{P_b : b \in I\!R\}$ where $P_b(A) = P(A - b)$ and $A - b = \{x - b : x \in A\}$. Note that if $\tau_b : I\!R \longrightarrow I\!R$ is translation by b, i.e. $\tau_b(x) = x + b$, then $P_b = P \circ \tau_b^{-1}$, i.e. if $Z \sim P$ then $\tau_b(Z) = Z + b \sim P_b$.*

(b) The scale family *generated by P is $\{P_a : a > 0\}$ where $P_a(A) = P(a^{-1}A)$ and $a^{-1}A = \{a^{-1}x : x \in A\}$. (Note that if $\varsigma_a : I\!R \longrightarrow I\!R$ is multiplication by a, i.e. $\varsigma_a(x) = ax$, then $P_a = P \circ \varsigma_a^{-1}$, i.e. if $Z \sim P$ then $\varsigma_a(Z) = aZ \sim P_a$.)*

(c) The location–scale family *generated by P is $\{P_{ab} : a > 0$ and $b \in I\!R\}$ where $P_{ab}(A) = P(a^{-1}(A - b))$. (Note that if $Z \sim P$ then $\tau_b(\varsigma_a(Z)) \sim P_{ab}$, i.e. $P_{ab} \sim P \circ \varsigma_a^{-1} \circ \tau_b^{-1}$.)*

□

Remarks 2.3.2 (a) Suppose P has Lebesgue density $f(x)$ and $P_b = P \circ \tau_b^{-1}$. Then the Lebesgue density of P_b is $f(x - b)$. This follows from a simple argument with c.d.f.'s. If Z has c.d.f. F_0, then the c.d.f. of $P_b = \mathrm{Law}[Z + b]$ is

$$F_b(x) \;=\; P[Z + b \leq x] \;=\; P[Z \leq x - b] \;=\; F_0(x - b).$$

Of course $f = dF/dz$ and

$$\frac{d}{dx}F_b(x) \;=\; \frac{d}{dx}F_0(x - b) \;=\; f(x - b),$$

which proves the claim.

(b) In a similar fashion, we can derive the c.d.f. for a scale family: if $a > 0$, then

$$\begin{aligned} P \circ \varsigma_a^{-1}((-\infty, x]) &= P[\varsigma_a^{-1}(-\infty, x]] \\ &= P(\{y : ay \leq x\}) \\ &= P(\{y : y \leq x/a\}) \\ &= P((-\infty, x/a]) \end{aligned}$$

and if F_1 denotes the c.d.f. for P and F_a the c.d.f. for $P_a = P \circ \varsigma_a^{-1}$, we have

$$F_a(x) \;=\; F_1(x/a).$$

So if P has Lebesgue density $f_1(z)$ then the Lebesgue density of P_a is

$$f_a(x) \;=\; \frac{d}{dx}F_a(x) \;=\; \frac{d}{dx}F_1(x/a) \;=\; \frac{1}{a}f(x/a).$$

(c) For a location-scale family, if P has Lebesgue density $f(x)$, the Lebesgue density of $P_{ab} = P \circ \varsigma_a^{-1} \circ \tau_b^{-1}$ is $a^{-1}f(a^{-1}(x - b))$.

Example 2.3.4 Let $P = U[0, 1]$ be the uniform distribution on $[0, 1]$. Then the location–scale family generated by P is denote **Unif** and contains all p.m.'s of the form $P_{ab} = U[b, a + b]$, i.e. if $U \sim U[0, 1]$ then $aU + b \sim U[b, a + b]$. It is more convenient to parameterize the uniform density as $U[\alpha, \beta]$ where $\alpha < \beta$, i.e. to give the endpoints. In general, one can reparameterize a family of probability measures to whatever is convenient since the parameter in some sense is only a label for the distribution.

□

The above example is also an example of what is known as a *truncation family.* To define such a family, let $g : I\!\!R \longrightarrow [0, \infty)$ be a Borel function satisfying

$$0 \ < \ \int_a^b g(x)\, dx \ < \ \infty$$

for all $-\infty < a < b < \infty$. Then we put

$$f_{ab}(x) \ = \ \frac{g(x) I_{[a,b]}(x)}{\int_a^b g(y)\, dy}$$

for $a < b$. Clearly the uniform family is a truncation family with constant g. Such truncation families have little if any application in practice, although they seem to play an important role in mathematical statistics textbooks.

Example 2.3.5 Let $P = Exp(1)$, the exponential distribution with Lebesgue density

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

The location–scale family generated by P is called the shifted exponentials and will be denoted **ShExp**, and a member thereof will be denoted $Exp[a, b]$ and has Lebesgue density

$$f_{ab}(x) \ = \ a^{-1} \exp[-(x - b)/a] I_{[b,\infty)}(x) \ .$$

Note that the support $[b, \infty)$ depends on the parameter. The scale family **Exp** of distributions $Exp[a, 0]$, $a > 0$ (called the family of exponential distributions and not the exponential family) is perhaps more fundamental and is frequently used as a model for observations which must be positive, such as lifetimes or masses. Note that the shifted exponential family $\{Exp[\beta, b] : \beta > 0 \text{ and } b \in I\!\!R \ \}$ is not a subfamily of **Gamma**, and it is also not an exponential family since the support depends on the parameter b. See Remark 2.3.1 (c) above.

The shifted exponential family arises in a natural way from the ordinary family of exponential distributions as follows. Suppose the observable is the mass of tumors in a mouse liver. If a tumor is too small (say less than b where $b > 0$) then it is not observed at all. One can see then that the observed tumor masses are conditional on being larger than b. If the exponential distribution $Exp[a,0]$ applies for all tumor masses (observed and unobserved) then because of elementary properties of the exponential distribution the observed tumor masses will be $Exp[a,b]$. See Exercise 2.3.7. Note however that we will not obtain a true location–scale family since we will require that $b > 0$.

The location family of distributions $Exp[1,b]$ where b is an arbitrary real number has little application in practice. It is however an example of a *left truncation family*, which we leave as an exercise to define.

□

The next example illustrates how Definition 2.3.7 can be extended to scale families generated by more than one probability measure. In a similar fashion, on can generate location families or location-scale families beginning with a family of probability measures.

Example 2.3.6 Let $\mathbf{P_0} = \{Gamma(\alpha,1) : \alpha > 0\}$ and denote the elements of $\mathbf{P_0}$ by $P_{\alpha,1}$ with corresponding Lebesgue densities $f_{\alpha,1}$. For each $\beta > 0$ and each $f_{\alpha,1}$ we obtain a "rescaled density" as

$$f_{\alpha,\beta}(x) \;=\; \frac{1}{\beta} f_{\alpha,1}\left(\frac{x}{\beta}\right) \quad ,$$

and of course $\{f_{\alpha,\beta} : \alpha > 0 \text{ and } \beta > 0\ \}$ gives the entire **Gamma** family. We say that **Gamma** is the scale family generated by $\{Gamma(\alpha,1) : \alpha > 0\}$, and for this reason β is often called the *scale parameter*.

□

2.3.3 Group Families.

In this subsection we define a more general class of families of distribution which includes the location, scale, and location–scale families of the previous subsection.

Definition 2.3.8 *(a) A class of transformations* $\mathbf{T}$ *on* $(\Xi, \mathcal{G})$ *is called a* transformation group *iff the following hold:*

(i) *Every* $g \in \mathbf{T}$ *is measurable* $g : (\Xi, \mathcal{G}) \longrightarrow (\Xi, \mathcal{G})$.

(ii) $\mathbf{T}$ *is closed under composition, i.e. if* g_1 *and* g_2 *are in* $\mathbf{T}$ *the so is* $g_1 \circ g_2$.

(iii) $\mathbf{T}$ *is closed under taking inverses, i.e. if* $g \in \mathbf{T}$ *then* $g^{-1} \in \mathbf{T}$.

If $g_1 \circ g_2 = g_2 \circ g_1$ *for all* g_1 *and* g_2 *in* $\mathbf{T}$*, then* $\mathbf{T}$ *is called* commutative.

(b) If $\mathbf{T}$ *is a transformation group and* $\mathbf{P}_0$ *is a family of probability measures on* $(\Xi, \mathcal{G})$*, then the* group family generated by $\mathbf{P}_0$ under $\mathbf{T}$ *is*

$$\mathbf{P}_0 \circ \mathbf{T}^{-1} \;=\; \{P \circ g^{-1} : P \in \mathbf{P}_0 \textit{ and } g \in \mathbf{T}\} \,.$$

□

Note that if $Z \sim P$, then $g(Z) \sim P \circ g^{-1}$.

Example 2.3.7 Consider the observation space $(\mathbb{R}^n, \mathcal{B}_n)$. For an $n \times n$ nonsingular matrix A and $\underline{b} \in \mathbb{R}^n$, define the transformation $g_{A,\underline{b}}(\underline{x}) = A\underline{x} + \underline{b}$. The family of transformations $\mathbf{T} = \{g_{A,\underline{b}} : A$ is an $n \times n$ nonsingular matrix and $\underline{b} \in \mathbb{R}^n\}$ is called the *affine group*. We verify that $\mathbf{T}$ is indeed a transformation group by checking the three defining properties.

(i) $g_{A,\underline{b}} : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is Borel measurable since it is continuous.

(ii) Given $g_{A_1,\underline{b}_1}$ and $g_{A_2,\underline{b}_2}$, we have by some simple algebra

$$(g_{A_1,\underline{b}_1} \circ g_{A_2,\underline{b}_2})(\underline{x}) \;=\; (A_1A_2)\underline{x} \,+\, (A_1\underline{b}_2 + \underline{b}_1) \;,$$

i.e. $g_{A_1,\underline{b}_1} \circ g_{A_2,\underline{b}_2} = g_{A,\underline{b}}$ where $A = A_1A_2$ and $\underline{b} = A_1\underline{b}_2 + \underline{b}_1$. This shows $\mathbf{T}$ is closed under composition.

(iii) Given $g_{A,\underline{b}}$ and $\underline{x} \in \mathbb{R}^n$, consider solving for $\underline{y}$ in $g_{A,\underline{b}}(\underline{y}) = \underline{x}$, which gives $\underline{y} = A^{-1}\underline{x} + (-A^{-1}\underline{b})$, i.e. $(g_{A,\underline{b}}^{-1}$ is an affine transformation with matrix A^{-1} and shift $-A^{-1}\underline{b}$. This shows $\mathbf{T}$ is closed under taking inverses.

We note that $\mathbf{T}$ is not commutative, even when $n = 1$. There are two interesting transformation *subgroups*, i.e. subsets of the affine group which are also closed under composition and taking inverses. One is the *general linear group* $\{g_{A,\underline{0}} : A$ is $n \times n$ and nonsingular $\}$, which is simply the group of all nonsingular linear transformations on $\mathbb{R}^n$. It is sometimes denoted $\mathbf{GL}(n)$. The other subgroup of interest is the translation subgroup $\{g_{I,\underline{b}} : \underline{b} \in \mathbb{R}^n\}$, where I is an $n \times n$ identity matrix.

We now generate a group family under the affine group. Let $\mathbf{P}_0$ consist of the single p.m. $N(\underline{0}, I)$, i.e. the standard normal distribution on $\mathbb{R}^n$. If $\underline{Z} \sim N(\underline{0}, I)$, then $A\underline{Z} + \underline{b} \sim N(\underline{b}, AA^t)$ (Exercise 2.3.3, (a)). Further, any nonsingular normal distribution on $\mathbb{R}^n$ can be so generated. We do have the following problem if we use the parameterization $(\underline{b}, A)$: two different A's can give rise to the same normal distribution (i.e. if $AA^t = A_1A_1^t$). Thus, the parameter does not uniquely define the distribution, i.e. the parameter is not identifiable in the terminology of Definition 2.3.2. To avoid this problem, we will use instead the parameters $(\underline{b}, V)$ where $V = AA^t$ is the covariance.

□

Definition 2.3.9 *Let* $\mathbf{T}$ *be a transformation group and let* $\mathbf{P}$ *be a family of probability measures on* $(\Xi, \mathcal{G})$*. We say that* $\mathbf{P}$ *is* $\mathbf{T}$–invariant *iff* $\mathbf{P} \circ \mathbf{T}^{-1} = \mathbf{P}$*.*

□

Proposition 2.3.3 *If* $\mathbf{P}$ *is a group family (generated by some* $\mathbf{P}_0$*) under* $\mathbf{T}$*, then* $\mathbf{P}$ *is* $\mathbf{T}$*-invariant.*

□

Example 2.3.8 Let **P** be the multivariate location family generated by the $N(\underline{0}, I)$ on $I\!\!R^n$, i.e. $\mathbf{P} = \{N(\underline{\mu}, I) \ : \ \underline{\mu} I\!\!R^n\}$. Then of course **P** is translation invariant by the last Proposition. It turns out that the family is also *spherically invariant*, whereby we mean that it is invariant under the group of orthogonal transformations $\mathbf{O}(n) := \{U : U$ is an $n \times n$ orthogonal matrix $\}$. To see this, note that if $X \sim N(\underline{\mu}, I)$ and U is orthogonal then $UX \sim N(U\underline{\mu}, UIU^t)$ and $UIU^t = UU^t = I$, so $UX \sim N(U\underline{\mu}, I)$.

□

Exercises for Section 2.3.

2.3.1 (a) Put the normal family of Example 2.3.1 in canonical form, determine the natural parameter space, and determine whether or not the family is of full rank.

(b) Same as (a) but for Example 2.3.2.

2.3.2 Prove Proposition 2.3.1.

2.3.3 Let $\underline{Z}$ be a random n-vector with independent $N(0,1)$ components, i.e. Z_1, Z_2, ..., Z_n are mutually independent r.v.'s with the same distribution which has Lebesgue density

$$f(z) \;=\; \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad , \quad -\infty < z < \infty \quad . \tag{2.65}$$

Let V be a given nonnegative definite matrix and $\underline{\mu}$ a given n-vector.

(a) Suppose A is any $n \times n$ matrix with $AA^t = V$. Show that $\underline{X} = A\underline{Z} + \underline{\mu}$ has a $N(\underline{\mu}, V)$ distribution, i.e. Law$[\underline{X}]$ has the m.g.f. given in Definition 2.3.5. (Note that such matrices A exist. Exercises 2.1.18 and x2.1.18 give two such examples.) Hint: Derive $\psi_{\underline{Z}}$ and use that to derive $\psi_{\underline{X}}$.

(b) Assume that if V is nonsingular then the $N(\underline{\mu}, V)$ distribution is dominated by m^n and the Lebesgue density is given by (2.43). (This is shown in Exercise 2.4.4.) Show that the family of multivariate normal distributions $\{N(\underline{\mu}, V) : \ \underline{\mu} \in I\!R^n$ and V is $n \times n$ strictly positive definite $\}$ is an exponential family.

(c) Following up on part (b), put the family in canonical form, determine the natural parameter space, and determine whether or not the family is of full rank.

2.3.4 Verify the form of the Lebesgue density for a location-scale family claimed in Remark 2.3.2 (c).

2.3.5 For $0 \le \alpha$ and $0 \le \phi < 2\pi$, the Fisher-von Mises density with parameters α and ϕ is given by

$$f_{\alpha,\phi}(x) \;=\; \frac{1}{2\pi I_0(\alpha)} \exp[\alpha \cos(x + \phi)] \quad , \quad 0 \le x < 2\pi \; ,$$

where

$$I_0(\alpha) \;=\; \frac{1}{2\pi} \int_0^{2\pi} e^{\alpha \cos(x)} \, dx$$

is a so-called modified Bessel function. We will denote the corresponding distribution by $FM(\alpha, \phi)$.

(a) Show that $f_{\alpha,\phi}$ is a Lebesgue density.

(b) Is the parameterization identifiable? What if we restrict to $\alpha > 0$?

(c) For $\theta \in [0, 2\pi)$, define modulo 2π translation by θ as the transformation $g_\theta : [0, 2\pi) \longrightarrow [0, 2\pi)$ given by

$$g_\theta(x) \;=\; \begin{cases} x + \theta & \text{if } x + \theta < 2\pi, \\ x + \theta - 2\pi & \text{if } x + \theta \geq 2\pi. \end{cases}$$

Show that $\mathbf{T} = \{g_\theta : \theta \in [0, 2\pi)\}$ is a transformation group. Is $\mathbf{T}$ commutative?

(d) Show that $\mathbf{FM} = \{f_{\alpha,\phi} : \alpha > 0 \text{ and } 0 \leq \phi < 2\pi\}$ is the group family generated by $FM_0 = \{f_{\alpha,0} : \alpha > 0\}$ under $\mathbf{T}$.

(e) Show directly (without recourse to Proposition 2.3.3) that **FM** is invariant under $\mathbf{T}$.

(f) Show that **FM** is an exponential family.

2.3.6 (a) Write **FM** from Exercise 2.3.5 in canonical form. Determine the natural parameter space, and show that the natural exponential family is of full rank.

(b) Consider the subfamily of **FM** given by $\{FM(\alpha, \phi_0) : 0 \leq \alpha < \infty\}$ where ϕ_0 is any fixed constant in $[0, 2\pi)$. Is this an exponential subfamily? Express it in canonical form, determine the natural parameter space, and determine whether or not the family is of full rank.

(c) Consider the subfamily of **FM** given by $\{FM(1, \phi) : 0 \leq \phi < 2\pi\}$. Is this family of full rank?

2.3.7 Suppose X is a r.v. with $Exp(a)$ distribution for some $a > 0$. However, you only get to observe the X if $X > b$ where $b > 0$ is some constant threshhold. Show that $P[X > x|X > b] = \exp[(x - b)/a]$ for $x > b$, and conclude that if we observe X, then it has a $Exp[a, b]$ distribution.

2.3.8 Let P be a Borel p.m. on $I\!R$. Show that the group family generated by P under the affine group (Example 2.3.7) is a location-scale family if and only if P is symmetric about some point b_0, i.e. $P(A - b_0) = P(b_0 - A)$ for all Borel sets A.

2.3.9 Determine which of the following transformation groups is commutative.

(a) The translation group on $I\!R^n$.

(b) The scale group on $I\!R$ given by $\{m_a : a > 0\}$ as in Definition 2.3.7 (b).

(c) The translation-scale group on $I\!R$ given by $\{t_b \circ m_a : b \in I\!R \text{ and } a > 0\}$ as in Definition 2.3.7 (c).

(d) The general linear group on $I\!R^n$ given in Example 2.3.7.

2.3.10 For each of the following dominated families, show that they are exponential families, put them in canonical form, determine the natural parameter space, and determine whether or not the family is of full rank.

(a) The Poisson family **Poisson** which has densities w.r.t. counting measure on $I\!\!N$ given by

$$f_\lambda(x) \;=\; \frac{e^{-\lambda}\lambda^x}{x!}$$

We denote the corresponding Poisson distribution by $Poisson(\lambda)$, where $\lambda > 0$.

(b) The Binomial family **Bin** of $B(n,p)$ distributions with parameter p, $0 < p < 1$.

(c) The Beta family **Beta** of $Beta(\alpha,\beta)$ distributions which have Lebesgue densities

$$f_{\alpha,\beta}(x) \;=\; \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \quad , \quad 0 < x < 1 \; .$$

Here, $0 < \alpha$ and $0 < \beta$.

(d) The negative binomial family **NBin** of distributions $NB(m,p)$ where m is a positive integer and $0 < p < 1$, which has density w.r.t. counting measure on $I\!\!N$ given by

$$f_p(x) \;=\; \left(\begin{array}{c} m+x-1 \\ m-1 \end{array} \right) p^m (1-p)^x \; .$$

2.3.11 (a) Suppose X_1, X_2, …, X_n are independent random observables where each Law$_\theta(X_i)$ is an exponential family of the form

$$f_\theta(x_i) \;=\; \exp[\underline{\eta}_i(\theta)^t \underline{T}_i(x) - B_i(\theta)]h_i(x),$$

w.r.t. the σ-finite dominating measure μ_i. Note that the parameter θ is common to each of the densities. Let $X = (X_1, \ldots, X_n)$. Show Law$_\theta(X)$ is from an exponential family.

(b) Suppose we apply the result of part (a) to the i.i.d. setting of Remark Rmk3.2.13(a). Explain what would be undesirable about the result and how it could be put in a desirable form.

2.3.12 Suppose X is a random n-vector so that each P_θ is a Borel measure, and assume $\{P_\theta : \theta \in \Theta\}$ is an exponential family. Show that $A = \mathrm{supp}(P_\theta)$ doesn't depend on θ. See Remark Rmk3.2.13(c).

2.3.13 Verify equations (2.50) and (2.51).

2.3.14 For each of the examples in Exercise 2.3.10, compute the mean and variance using Proposition 2.3.2 (b). You should be able to check your formulae by computing the moments directly as well.

2.3.15 In Example 2.3.3, verify (2.52) is valid for $n = 1$. Compute the m.g.f. for the distribution given by (2.52) and verify it is the m.g.f. for the distribution of $\underline{Y} = \sum \underline{X}_i$ where the $\underline{X}_i$ are i.i.d. $Mult(1, \underline{p})$.

2.3.16 In Example 2.3.3, verify (2.62) through (2.64) by direct calculation by computing the mean and covariance of a $Mult(1, \underline{p})$ random vector and then using the fact that a $Mult(n, \underline{p})$ random vector is the sum of n i.i.d. $Mult(1, \underline{p})$ random vectors.

2.3.17 Suppose Y_1, Y_2, ..., Y_n are independent r.v.'s with $Y_i \sim Poisson(\lambda t_i)$ where t_1, t_2, ..., t_n are known positive numbers, and $\lambda > 0$ is unknown.

(a) Show that the model for $\underline{Y} = (Y_1, Y_2, ..., Y_n)$ is an exponential family.

(b) Put the family in minimal canonical form and identify the natural parameter space.

(c) Find the m.g.f. of $\sum_{i=1}^n Y_i$ using exponential family theory. What is the distribution of this sum?

2.3.18 Let $\{\text{Law}_\theta[X] : \theta \in \Theta\}$ be an exponential family with σ-finite dominating measure μ and densities as in (2.41). Let $T(X)$ be the random p-vector where T is as given in (2.41). Show that there is a σ-finite dominating measure on $I\!R^p$ for $\text{Law}_\theta[T(X)]$ such that $\text{Law}_\theta[T(X)]$ has a density of the form

$$f_\theta(t) \;=\; \exp[\eta(\theta)^t t - C(\theta)]$$

2.3.19 Prove Proposition 2.3.3.

2.4 Distributional Calculations.

Many problems in theoretical statistics require the calculation of distributions. W. S. Gosset (writing under the pseudonym of Student) made history in 1908 with his derivation of the density for the t distribution. This is but one example of the importance of this general issue. Unfortunately, there is no single, general method one can apply, but rather a whole toolbox of techniques is necessary. One then attempts to use various tools from the toolbox until one works. We will use four general classes of methods: methods based on the c.d.f., methods based on transformations, methods based on conditioning, and methods based on moment generating functions. We have already seen the use of c.d.f.'s in computing distributions in Remarks 2.3.2. Here, we will concentrate on the other methods.

2.4.1 Lebesgue Densities and Transformations.

In conjunction with the change of variables theorem (Theorem 1.2.10), it was mentioned that one often encounters a Jacobian in actually computing the induced measure, which we now explain. First, some more review of advanced calculus on $I\!R^n$. Let U be an open subset of $I\!R^n$ and $h : U \longrightarrow I\!R^k$ have continuous partial derivatives $\partial h_i/\partial x_j$ of all component functions, $1 \leq i \leq k$, $1 \leq j \leq n$. The *derivative* $Dh(x)$ is the $k \times n$ matrix with (i, j) entry $[\partial h_i/\partial x_j](x)$. $Dh(x)$ is sometimes called the *Jacobian matrix*. It is a matrix valued function of x. Also, $Dh(x)$ may be used for local linear approximation of h in the sense that

$$h(y) \;=\; h(x) \,+\, Dh(x)(y-x) \,+\, \mathrm{Rem}(x,y) \quad , \tag{2.66}$$

where the remainder term satisfies

$$\lim_{y \to x} \frac{\|\mathrm{Rem}(x,y)\|}{\|y-x\|} \;=\; 0 \quad .$$

This last equation states that $\|\mathrm{Rem}(x, y)\|$ tends to be much smaller than $\|y-x\|$ if y is close to x, and so the "linear" function $h(x) + Dh(x)(y - x)$ as a function of y tends to be a good approximation to $h(y)$ for y close to x. If $U \subset I\!R^n$ and $h : U \longrightarrow I\!R^n$, then $Dh(x)$ is a square $n \times n$ matrix, so its determinant

$$\det\, Dh(x) \;=\; J(x) \quad ,$$

is defined, and is sometimes called the *Jacobian (determinant)*. The Inverse Function Theorem (p. 221 of Rudin, *Principles of Mathematical Analysis*) states that under these conditions, if $J(a) \neq 0$ at some $a \in U$, then h is invertible in a neighborhood of a and h^{-1} has derivative $[D(h^{-1})](y) = [(Dh)(h^{-1}(y))]^{-1}$ $= [((Dh) \circ h^{-1})(y)]^{-1}$ at a point y in this neighborhood of $h(a)$. Part of the conclusion is that this inverse matrix exists in the neighborhood of $h(a)$. Also, if

$J(x) \neq 0$ for all $x \in U$, then $h(V)$ is an open set for any open set $V \subset U$. This latter fact (V open implies $h(V)$ open) implies that h^{-1} is measurable, if it exists on all of $h(U)$ (Exercise 1.4.20).

Remarks 2.4.1 If $h : I\!\!R^n \longrightarrow I\!\!R$, then the derivative Dh as defined above is an $n \times 1$ matrix, i.e. a "row vector," whereas the gradient ∇h is a $1 \times n$ "column vector." Note that $Dh = (\nabla h)^t$. The difference is not really important, but one should remember that formulae such as (2.66) will involve a transpose when expressed in terms of the gradient, i.e.

$$h(y) \ = \ h(x) \ + \ (\nabla h(x))^t (y - x) \ + \ \mathrm{Rem}(x, y) \quad ,$$

when h is real valued.

□

Theorem 2.4.1 *Suppose $\Omega \subset I\!\!R^n$ is open and $h : \Omega \longrightarrow I\!\!R^n$ is a one to one mapping with nonvanishing Jacobian (i.e. $J(x) \neq 0$ for all $x \in \Omega$). Let $\Lambda = h(\Omega)$, and let ν be Lebesgue measure restricted to Ω. Then $\nu \circ h^{-1}$ is a Borel measure on Λ, $\nu \circ h^{-1} \ll m$, and*

$$\frac{d(\nu \circ h^{-1})}{dm}(y) \ = \ \left\{ \begin{array}{cc} |\det D(h^{-1})(y)| & \textit{if } y \in \Lambda, \\ & \\ 0 & \textit{otherwise.} \end{array} \right\} \qquad m - a.e.$$

□

This result is Theorem 17.2, p. 229 of Billingsley. See also Theorem 10.9, page 252 of Rudin. To check that det $Dh(x) \neq 0$ for all x, it suffices to show that det $D(h^{-1})(y) \neq 0$ for all y by the Inverse Function Theorem applied to h^{-1}. A relation between the Jacobian of h^{-1} and h is given by

$$D(h^{-1})(y) \ = \ \left[Dh(h^{-1}(y)) \right]^{-1} . \tag{2.67}$$

This follows from the chain rule (see Exercise 2.4.2). Now we show the usefulness of Theorem 2.4.1 in probability theory.

Proposition 2.4.2 *Suppose P is a Borel p.m. on $I\!\!R^n$ which has Lebesgue density f. Let $h : \Omega \longrightarrow \Lambda$ be as in Theorem 2.4.1 where $\Lambda = h(\Omega)$ and suppose $P(\Omega) = 1$. Then $P \circ h^{-1}$ has Lebesgue density g given by*

$$g(y) \ = \ f(h^{-1}(y)) |\det D(h^{-1})(y)| \quad , \textit{ for all } y \in \Lambda \quad .$$

Put otherwise, if Law$[X] = P$ and $Y = h(X)$, then Law$[Y]$ has Lebesgue density given by g above.

Proof. Let $J(y) = \det\, D(h^{-1})(y)$, and let $B \subset \Lambda$ be a Borel set. Then

$$(P \circ h^{-1})(B) \;=\; P(h^{-1}(B)) \;=\; \int_{h^{-1}(B)} f(x)\,dx$$

$$= \int_\Omega I_{h^{-1}(B)}(x) f(x)\,dx \;=\; \int_\Omega I_B(h(x)) f(x)\,dx \quad ,$$

where the last equality follows since $x \in h^{-1}(B)$ iff $h(x) \in B$. Now put

$$\beta(y) \;=\; I_B(y)(f \circ h^{-1})(y) \quad ,$$

and since $(f \circ h^{-1})(h(x)) = f(x)$, we have

$$\begin{aligned}(P \circ h^{-1})(B) \;&=\; \int_\Omega \beta((h(x))\,dm^n(x) \\ &=\; \int_\Lambda \beta(y)\,d(m^n \circ h^{-1})(y) \quad ,\end{aligned}$$

where the last equation follows from the change of variables theorem (Theorem 1.1.2.10). By Proposition 1.4.2 (a) and the previous theorem,

$$\begin{aligned}(P \circ h^{-1})(B) \;&=\; \int_\Lambda \beta(y)|J(y)|\,dm^n(y) \\ &=\; \int_\Lambda I_B(y)(f \circ h^{-1})(y)|J(y)|\,dy \\ &=\; \int_B (f \circ h^{-1})(y)|J(y)|\,dy \\ &=\; \int_B g(y)\,dy \quad .\end{aligned}$$

Since the above result holds for arbitrary Borel $B \subset \Lambda$, it follows that $d(P \circ h^{-1})/dm^n$ exists and equals g, by the uniqueness part of the Radon-Nikodym Theorem.

□

Example 2.4.1 (Log-Normal Distribution) Suppose $X \sim N(\mu, \sigma^2)$ and $Y = \exp[X]$. Then Y is said to have a *log-normal distribution with parameters μ and σ^2*. Perhaps we should say Y has an "exponential-normal distribution" as it is the exponential of a normal r.v., but the terminology "log-normal" is standard. It presumably arose from something like the statement, "The logarithm is normally distributed."

Now we derive the Lebesgue density using the previous theorem. Now $\Omega = I\!\!R$ and $\Lambda = (0, \infty)$. Of course, $h(x) = \exp[x]$ and $h^{-1}(y) = \log y$, so $D(h^{-1})(y) = 1/y$. Hence, letting f be the $N(\mu, \sigma^2)$ density we have

$$\begin{aligned}g(y) \;&=\; f(\log y)\frac{1}{y} \quad , \quad y > 0, \\ &=\; \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(\log y - \mu)^2\right] \quad , \quad y > 0.\end{aligned}$$

Next we consider the problem of computing the mean and variance of Y. One approach would be to compute $\int_0^\infty y^m\, dy$ for $y = 1, 2$. However, one should always consider all options in computing expectations via the law of the unconscious statistician. Now

$$E[Y^m] \;=\; E[\exp(mX)]$$

which is the m.g.f. of X evaluated at m. Recalling the m.g.f. of a univariate normal distribution

$$\psi_{N(\mu,\sigma^2)}(t) \;=\; \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right]$$

we have

$$E[Y^m] \;=\; \exp\left[\mu m + \frac{1}{2}\sigma^2 m^2\right]$$

and so

$$E[Y] \;=\; e^{\mu+\sigma^2}$$

$$\begin{aligned} \mathrm{Var}[Y] &= E[Y^2] - E[Y]^2 \\ &= e^{2\mu+4\sigma^2} \;-\; e^{2\mu+2\sigma^2} \\ &= e^{2\mu+2\sigma^2}\left[e^{2\sigma^2} - 1\right]. \end{aligned}$$

□

Example 2.4.2 (Student's t-distribution) Suppose X and Y are independent r.v.'s with the following distributions:

$$\mathrm{Law}[X] \;=\; N(0,1) \quad , \quad \mathrm{Law}[Y] \;=\; \chi^2_n \,,$$

i.e. the Lebesgue densities are given by

$$f_X(x) \;=\; \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \,,$$

$$f_Y(y) \;=\; \frac{y^{n/2-1} e^{-y/2}}{\Gamma(n/2) 2^{n/2}} I_{(0,\infty)}(y) \,.$$

Note that X has the standard normal distribution and Y has a chi-squared distribution with n degrees of freedom. Let

$$T \;=\; \frac{X}{\sqrt{Y/n}} \,.$$

Then T is said to have *Student's t-distribution with* n *degrees of freedom.* We will derive the Lebesgue density for T. By Proposition 1.4.3, the joint density for X

and Y is $f_{XY}(x,y) = f_X(x)f_Y(y)$. Letting $\Omega = I\!\!R \times [0,\infty) = \mathrm{supp}(\mathrm{Law}[X,Y])$ (this last equality follows from Exercise 1.4.19) and

$$h(x,y) \;=\; (\; x/\sqrt{y/n}\;,\; y\;) \quad , \quad (x,y) \in \Omega \; ,$$

then $h(\Omega) = \Omega$ and h is one to one on Ω since

$$h(x,y) \;=\; (t,u) \text{ iff } x = t\sqrt{u/n} \text{ and } y = u \; ,$$

and this gives the inverse function

$$h^{-1}(t,u) \;=\; (\; t\sqrt{u/n}\;,\; u\;) \; .$$

Now the Jacobian matrix for h^{-1} is

$$Dh^{-1}(t,u) \;=\; \left[\begin{array}{cc} \sqrt{u/n} & t/(2\sqrt{un}) \\ 0 & 1 \end{array} \right] \quad , \qquad (2.68)$$

with Jacobian

$$\det\, Dh^{-1}(t,u) \;=\; \sqrt{u/n} \quad , \qquad (2.69)$$

which is nonvanishing for all $(t,u) \in \Omega$. Hence, the joint density of (T,U) is by Theorem 2.4.1

$$\begin{aligned} f_{TU}(t,u) \;&=\; f_{XY}(h^{-1}(t,u))|\det\, Dh^{-1}(t,u)| \\ &=\; \left[\frac{1}{\sqrt{2\pi}} e^{-t^2u/(2n)} \right] \left[\frac{u^{n/2-1}e^{-u/2}}{\Gamma(n/2)2^{n/2}}\, I_{(0,\infty)}(u) \right] \sqrt{u/n} \\ &=\; \frac{1}{\pi^1/2\Gamma(n/2)2^{(n+1)/2}n^{1/2}}\, u^{(n-1)/2} e^{-(1+t^2/n)u/2} I_{(0,\infty)}(u) \quad . \end{aligned}$$

To get the marginal density for T, we apply Proposition 1.4.4 (or Exercise 1.4.12 (a)) to obtain

$$\begin{aligned} f_T(t) \;&=\; \int f_{TU}(t,u)\, du \\ &=\; \frac{1}{\pi^{1/2}\Gamma(n/2)2^{(n+1)/2}n^{1/2}} \int_0^\infty u^{(n-1)/2} e^{-(1+t^2/n)u/2}\, du \end{aligned}$$

In the last integral make the change of variables

$$v \;=\; (1 + t^2/n)u \; , \text{ so that } du \;=\; \frac{dv}{1 + t^2/n} \quad .$$

This gives

$$\begin{aligned} f_T(t) \;&=\; \frac{1}{\pi^{1/2}\Gamma(n/2)2^{(n+1)/2}n^{1/2}} (1 + t^2/n)^{-(n+1)/2} \int_0^\infty v^{(n+1)/2-1} e^{-v/2}\, dv \\ &=\; \frac{1}{\pi^{1/2}\Gamma(n/2)2^{(n+1)/2}n^{1/2}} (1 + t^2/n)^{-(n+1)/2}\, \Gamma((n+1)/2)2^{(n+1)/2} \end{aligned}$$

where the last line follows since the integrand in the previous line is the $\chi^2_{(n+1)}$ density without the normalizing constant. In summary,

$$f_T(t) \;=\; \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)}\,(1+t^2/n)^{-(n+1)/2} \quad . \tag{2.70}$$

This is the (Lebesgue) density of Student's t–distribution with n degrees of freedom.

□

The preceding example is typical of how the method gets used when one wishes to obtain the Lebesgue density for a real valued random variable Y that is a function of a random vector $\underline{X}$: one must extend Y to a vector $\underline{Y}$ of the same dimension as $\underline{X}$ to obtain a one to one trasformation with nonsingular Jacobian and then apply mariginalization to get the desired density. Sometimes, it is not possible to compute the marginal density in a neat closed form and one must be satisified with an integral expression or something similar, as in Exercises 2.4.5 and 2.4.7.

2.4.2 Applications of Conditional Distributions.

Conditional distributions can be very useful for proving results about conditional expectations. For instance, suppose X and Y are r.v.'s and let $h(X)$ and $g(X)$ be two functions of X such that $h(X) \leq g(X)$ a.s. Then of course $E[h(X)|Y] \leq E[g(X)|Y]$ by Theorem 1.5.7 (b), assuming that $E[|h(X)|] < \infty$ and $E[|g(X)|] < \infty$. But we can prove it with conditinal distributions using the elementary properties of integrals, viz.

$$\begin{aligned} E[h(X)|Y=y] \;&=\; \int h(x)\, d\mathrm{Law}[X|Y=y](x) \\ &\leq\; \int g(x)\, d\mathrm{Law}[X|Y=y](x) \\ &=\; E[g(X)|Y=y] \end{aligned}$$

where the inequality follows for each y since we are simply integrating w.r.t. the measure Law$[X|Y=y]$.

The following conditional moment inequality is extremely useful in statistics.

Theorem 2.4.3 (Jensen's Inequality for Conditional Expectation.) *Let* $Y : (\Omega, \mathcal{F}, P) \longrightarrow (\Lambda, \mathcal{G})$ *be a random element and* $\underline{X}$ *a random* n*–vector defined on the same probability space. Assume there is a convex Borel set* $K \subset I\!R^n$ *such that* $P[\underline{X} \in K] = 1$. *Let* $g : K \times \Lambda \longrightarrow I\!R$ *be a measurable function on* $(K \times \Lambda, \mathcal{B}_K \times \mathcal{G})$

where $\mathcal{B}_K$ denotes the Borel subsets of K. Assume that $g(\cdot, y)$ is a convex function on K for each fixed $y \in \Lambda$ and that $E|g(\underline{X}, Y)| < \infty$. Then

$$E[g(\underline{X}, Y)|Y = y] \;\geq\; g(E[\underline{X}|Y = y]) \quad , \quad Law[Y] - a.s. \tag{2.71}$$

Furthermore, if for $Law[Y]$ almost all $y \in \Lambda$, $Law[\underline{X}|Y = y]$ is nondegenerate, and if $g(\cdot, y)$ is strictly convex, then strict inequality holds in (2.71).

Proof. By equation (1.71),

$$E[g(\underline{X}, Y)|Y = y] \;=\; \int_K g(\underline{x}, y)\, dP_{\underline{X}|Y}(\underline{x}|y) \quad , \quad \text{Law}[Y] - a.s.$$

where the integral may be taken over K since $I_K(\underline{X}) = 1$ a.s. by assumption. Applying the ordinary Jensen's inequality (Theorem 2.1.4) to the p.m. $P_{\underline{X}|Y}(\cdot|Y = y)$ and the convex function $g(\cdot, y)$ on the r.h.s. of the last displayed equation we have

$$E[g(\underline{X}, Y)|Y = y] \;\geq\; g\left(\int_K \underline{x}\, dP_{\underline{X}|Y}(\underline{x}|y)\right) \quad , \quad \text{Law}[Y] - a.s.$$

which is the desired result. The claim involving strict inequality follows from the analogous claim in Theorem 2.1.4.

□

An alternative (and more general) proof to the previous result may be found in Billingsley, p. 470.

Examples of Conditional Distributions.

We will find many uses of the notions of this section in the remainder of the text. Now we will introduce some applications for the purposes of illustration.

One of the most useful results for deriving conditional distributions is Proposition 1.5.5, derived from Proposition 1.5.4 (see also Remark 1.5.5). Propositions 1.5.4 and 1.5.5 tell us how to obtain a density for a conditional distribution when the joint distribution is dominated by a product measure. To summarize, let $(\underline{X}, \underline{Y})$ have a (joint) density $f_{(\underline{X},\underline{Y})}(\underline{x}, \underline{y})$ w.r.t. $\mu_1 \times \mu_2$. Letting $f_{\underline{Y}}(\underline{y}) = \int f_{(\underline{X},\underline{Y})}(\underline{x}, \underline{y})\, d\mu_1(\underline{x})$ denote the marginal density of $\underline{Y}$ w.r.t. μ_2, we can write the conditional density of $\underline{X}$ given $\underline{Y} = \underline{y}$ as

$$f_{\underline{X}|\underline{Y}}(\underline{x}|\underline{y}) \;=\; \frac{f_{(\underline{X},\underline{Y})}(\underline{x}, \underline{y})}{f_{\underline{Y}}(\underline{y})} \quad .$$

Too often inexperienced students will laboriously compute $f_{\underline{Y}}(\underline{y})$ and divide it into $f_{(\underline{X},\underline{Y})}(\underline{x}, \underline{y})$ to obtain the conditional density. In fact, one can often recognize $f_{\underline{X}|\underline{Y}}(\underline{x}|\underline{y})$ by inspection of $f_{(\underline{X},\underline{Y})}(\underline{x}, \underline{y})$, possibly after a little algebra. If we look at $f_{(\underline{X},\underline{Y})}(\underline{x}, \underline{y})$ as a function of $\underline{x}$ for fixed $\underline{y}$, then it is already the conditional density except for the "normalizing constant" $1/f_{\underline{Y}}(\underline{y})$.

Example 2.4.3 Let # be counting measure on $I\!N = \{0, 1, 2, \ldots\}$ and let m be Lebesgue measure on $I\!R$. Suppose (X, Y) is a random 2–vector having joint density w.r.t. $\# \times m$

$$\begin{aligned} f(x,y) &= \frac{e^{-2y} y^x}{x!} I_{(0,\infty)}(y) \\ &= C(y) \frac{y^x}{x!}, \end{aligned}$$

where $C(y)$ doesn't depend on x. Looking at the factor $y^x/x!$ in $f(x, y)$ which does depend on x, we see that it is the density (w.r.t. #) of a Poisson r.v. with mean y. Hence, $\text{Law}[X|Y = y] = Poisson(y)$. Note that we have not computed the marginal density for Y, but since $\sum_{x=0}^{\infty}(y^x/x!) = e^y$, we see $f_Y(y) = e^{-y} I_{(0,\infty)}(y)$ is an exponential distribution with mean 1. Thus, we see immediately that $E[X|Y]$ = 1 a.s. and $\text{Var}[X|Y] = 1$ a.s.

Similarly, the functional dependence of $f(x, y)$ on y can be concentrated in a factor $e^{-2y} y^x I_{(0,\infty)}(y)$, which is a $Gamma(x+1, 1/2)$ density except for a normalizing constant (namely $1/(\Gamma(x+1)(1/2)^{(x+1)})$), so $Law[Y|X] = Gamma(x+1, 1/2)$. Notice that we did not compute the marginal density of X w.r.t. # to obtain this conditional distribution. (See Exercise 2.4.9.)

□

Example 2.4.4 Suppose $\underline{Z} = (\underline{X}, \underline{Y})$ has a (joint) multivariate normal distribution which is nonsingular on $I\!R^{(m+n)}$ where $\underline{X}$ is m–dimensional and $\underline{Y}$ is n–dimensional. We wish to obtain the conditional distribution of $\underline{X}$ given $\underline{Y} = \underline{y}$. In order to appropriately split up the mean vector and covariance matrix, we will write them in "partitioned" form as

$$\underline{\mu}_Z = E\left[\begin{array}{c} \underline{X} \\ \underline{Y} \end{array}\right] = \left[\begin{array}{c} \underline{\mu}_X \\ \underline{\mu}_Y \end{array}\right], \tag{2.72}$$

where $\underline{\mu}_X = E[\underline{X}] \in I\!R^m$ and $\underline{\mu}_Y = E[\underline{Y}] \in I\!R^n$.

$$\Sigma_Z = \text{Cov}\left[\begin{array}{c} \underline{X} \\ \underline{Y} \end{array}\right] = \left[\begin{array}{cc} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{array}\right] \tag{2.73}$$

where

$$\begin{array}{rcl} \Sigma_{XX} = \text{Cov}[\underline{X}] & \text{is} & m \times m, \\ \Sigma_{XY} = \text{Cov}[\underline{X}, \underline{Y}] & \text{is} & m \times n, \\ \Sigma_{YX} = \Sigma_{XY}^t & \text{is} & n \times m, \\ \Sigma_{YY} = \text{Cov}[\underline{Y}] & \text{is} & n \times n. \end{array}$$

The conditional density will involve the inverse of the covariance matrix.

Lemma 2.4.4 (Inverse of a symmetric 2×2 partitioned matrix) *Let M be an $(n+m) \times (n+m)$ symmetric nonsingular matrix partitioned as*

$$M = \begin{bmatrix} A & B \\ B^t & C \end{bmatrix} \tag{2.74}$$

where

$$\begin{array}{ccc} A & \text{is } n \times n & \text{and symmetric} \\ C & \text{is } m \times m & \text{and symmetric} \\ B & \text{is } n \times m & \end{array}$$

Then A and C are nonsingular and

$$M^{-1} = \begin{bmatrix} D & E \\ E^t & F \end{bmatrix} \tag{2.75}$$

where

$$D = \left(A - BC^{-1}B^t\right)^{-1} \tag{2.76}$$

$$F = \left(C - B^tA^{-1}B\right)^{-1} \tag{2.77}$$

$$\begin{aligned} E &= -A^{-1}B\left(C - B^tA^{-1}B\right)^{-1} && (2.78) \\ &= -C^{-1}B^t\left(A - BC^{-1}B^t\right)^{-1} && (2.79) \end{aligned}$$

Proof. Invertibility of M implies invertibility of A and C. M^{-1} is symmetric, so it has the form (2.75) for some D, E, and F. We obtain matrix equations these must satisfy from

$$\begin{aligned} I_{m+n} = \begin{bmatrix} I_n & 0 \\ 0 & I_m \end{bmatrix} = MM^{-1} &= \begin{bmatrix} A & B \\ B^t & C \end{bmatrix} \begin{bmatrix} D & E \\ E^t & F \end{bmatrix} \\ &= \begin{bmatrix} AD + BE^t & AE + BF \\ B^tD + CE^t & B^tE + CF \end{bmatrix} \end{aligned}$$

where I_k denotes a $k \times k$ identity matrix. Note that in the last equality we have multiplied the partitioned matrices together (almost) as if the entries were scalars (but we have kept track of the order of the multiplications, as matrix multiplication is not commutative). The reader should verify that this formula is correct by checking individual matrix entries, if necessary. This leads to the system of equations

$$\begin{aligned} AD + BE^t &= I \\ AE + BF &= 0 \\ B^tD + CE^t &= 0 \\ B^tE + CF &= I \end{aligned}$$

At this point, one can try various algebraic steps, but remember that only A and C (not B) are invertible. Anyway, one thing that works is to solve the third equation for E^t and plug this into the first:

$$\begin{array}{rclcrcl} E^t & = & -C^{-1}B^tD & \Longrightarrow & AD - BC^{-1}B^tD & = & I \quad \Longrightarrow \\ (A - BC^{-1}B^t)D & = & I & \Longrightarrow & D & = & \left(A - BC^{-1}B^t\right)^{-1} \end{array}$$

which is (2.76). Plugging the formula for D back into the first equation and transposing gives (2.79). The other two formulae can be obtained in a similar manner, or by an appeal to symmetry.

□

Applying this result to obtain the form of the joint Lebesgue density for $(\underline{X}, \underline{Y})$, we obtain

$$\begin{aligned} f(\underline{x}, \underline{y}) &= (2\pi)^{-(n+m)/2} \left(\det \Sigma_Z\right)^{-1/2} \exp\left[-\frac{1}{2}\left\{\left[\begin{array}{c} \underline{x} - \underline{\mu}_X \\ \underline{y} - \underline{\mu}_Y \end{array}\right]^t \Sigma_Z^{-1} \left[\begin{array}{c} \underline{x} - \underline{\mu}_X \\ \underline{y} - \underline{\mu}_Y \end{array}\right]\right\}\right] \qquad (2.80) \\ &= (2\pi)^{-(n+m)/2} \left(\det \Sigma_Z\right)^{-1/2} \exp\left[-\frac{1}{2}\left\{(\underline{x} - \underline{\mu}_X)^t D (\underline{x} - \underline{\mu}_X)\right.\right. \\ &\qquad \left.\left. + 2(\underline{x} - \underline{\mu}_X)^t E (\underline{y} - \underline{\mu}_Y) + (\underline{y} - \underline{\mu}_Y)^t F (\underline{y} - \underline{\mu}_Y)\right\}\right] \end{aligned}$$

where

$$\begin{aligned} D &= \left(\Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{XY}^t\right)^{-1} \\ F &= \left(\Sigma_{YY} - \Sigma_{XY}^t\Sigma_{XX}^{-1}\Sigma_{XY}\right)^{-1} \\ E &= -\Sigma_{XX}^{-1}\Sigma_{XY}\left(\Sigma_{YY} - \Sigma_{XY}^t\Sigma_{XX}^{-1}\Sigma_{XY}\right)^{-1} \\ &= -\Sigma_{YY}^{-1}\Sigma_{XY}^t\left(\Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{XY}^t\right)^{-1} \end{aligned}$$

To derive the conditional density, we really only care about the joint density as a function of $\underline{x}$, as the factor that depends on $\underline{y}$ or parameters only will cancel out when we divide the joint density by the marginal $f_{\underline{Y}}$. Letting "C" denote a quantity which doesn't depend on $\underline{x}$, and is not necessarily the same in each appearance, we have

$$f(\underline{x}, \underline{y}) = C \exp\left[-\frac{1}{2}\left\{(\underline{x} - \underline{\mu}_X)^t D (\underline{x} - \underline{\mu}_X) + 2(\underline{x} - \underline{\mu}_X)^t E (\underline{y} - \underline{\mu}_Y)\right\}\right] \qquad (2.81)$$

Now we will show that this can be put in the form

$$C\exp\left[-\frac{1}{2}\left\{(\underline{x}-\underline{\mu}_{X|Y})^t\Sigma_{X|Y}^{-1}(\underline{x}-\underline{\mu}_{X|Y})\right\}\right] \tag{2.82}$$

for some $\underline{\mu}_{X|Y} \in I\!\!R^n$ (which depends on $\underline{y}$) and $\Sigma_{X|Y}$ which is $n \times n$. When we have done this, it will follow that the conditional distribution of $\underline{X}$ given $\underline{Y}$ is $N(\underline{\mu}_{X|Y}, \Sigma_{X|Y})$. Expanding out the form in the exponential in (2.81), we have

$$\begin{aligned}
&(\underline{x}-\underline{\mu}_X)^t D(\underline{x}-\underline{\mu}_X) + 2(\underline{x}-\underline{\mu}_X)^t E(\underline{y}-\underline{\mu}_Y) \\
&= \ \underline{x}^t D\underline{x} \ - \ 2\underline{x}^t D\underline{\mu}_X \ + \ 2\underline{x}^t E(\underline{y}-\underline{\mu}_Y) \ + \ C \\
&= \ \underline{x}^t D\underline{x} \ - \ 2\underline{x}^t D[\underline{\mu}_X - D^{-1}E(\underline{y}-\underline{\mu}_Y] \ + \ C \\
&= \ \{\underline{x}^t - [\underline{\mu}_X - D^{-1}E(\underline{y}-\underline{\mu}_Y]\}^t D\{\underline{x}^t - [\underline{\mu}_X - D^{-1}E(\underline{y}-\underline{\mu}_Y]\}^t \ + \ C
\end{aligned}$$

So, it is clear that

$$\begin{aligned}
\Sigma_{X|Y} &= D^{-1} \\
&= \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{XY}^t
\end{aligned} \tag{2.83}$$

$$\begin{aligned}
\underline{\mu}_{X|Y} &= \underline{\mu}_X - D^{-1}E(\underline{y}-\underline{\mu}_Y) \\
&= \underline{\mu}_X + \Sigma_{XY}\Sigma_{YY}^{-1}(\underline{y}-\underline{\mu}_Y)
\end{aligned} \tag{2.84}$$

Now we have shown that the joint density considered as a function of $\underline{x}$ alone can be put in the form (2.82) where C does not depend on $\underline{x}$. When this is normalized to be a Lebesgue probability density function in $\underline{x}$, we will obtain

$$f_{\underline{X}|\underline{Y}}(\underline{x}|\underline{y}) \ = \ (2\pi)^{-m/2}\left(\det\Sigma_{X|Y}\right)^{-1/2}\exp\left[-\frac{1}{2}\left\{(\underline{x}-\underline{\mu}_{X|Y})^t\Sigma_{X|Y}^{-1}(\underline{x}-\underline{\mu}_{X|Y})\right\}\right] \quad .$$

This proves our claim about the conditional distribution. Note that only the conditional mean and not the conditional variance depends on $\underline{y}$, and the conditional mean is a linear transformation of $\underline{y}$.

□

Example 2.4.5 Next we will illustrate the use of the Two Stage Experiment Theorem. Other authors refer to the technique as "heirarchical modelling." Consider the following data set of the numbers of moths caught in a trap on 24 consecutive nights at a site in North Buckinghamshire, England (taken from *A Handbook of Small Data Sets* by Hand, et.al.)

$$47, 21, 16, 39, 24, 34, 21, 34, 49, 20, 37, 65, 67, 21, 37, 46, 29, 41, 47, 24, 22, 19, 54, 71$$

We might expect this data to follow a Poisson distribution as there are presumably numerous moths in the area but a small probability of catching any individual moth. (Why does this motivate the use of a Poisson distribution for the model?). However, the sample mean is 36.9 while the sample variance is 251.7, suggesting the Poisson model is probably not valid since the mean and variance of a Poisson random variable are equal. However, we need not abandon the Poisson model entirely. It would only be valid if the number of moths and the probability of catching any individual moth were the same each night, but in fact we expect that these may vary from night to night in a random way. For instance, the number of moths may vary considerably and their propensity for being trapped may depend heavily on meteoroligical conditions which can change markedly from night to night.

Let us suppose that on a given night, the number of moths caught in the trap is a realization of a random variable $X \in I\!N$. Assume there is an unobservable $Y \in (0, \infty)$ such that $\mathrm{Law}[X|Y = y] = Poisson(y)$. Then once a distribution for Y is specified, we have a joint distribution by the Two Stage Experiment Theorem, and hence also a marginal distribution for X. Note that $E[X] = E[E[X|Y]] = E[Y]$, and by Exercise 1.5.7(b),

$$Var[X] \;=\; E[Var[X|Y]] + Var[E[X|Y]] \;=\; E[Y] + \mathrm{Var}[Y].$$

Using our sample values, we would estimate $E[Y]$ as about 37 and $\mathrm{Var}[Y]$ as about $251.7 - 36.9 \doteq 215$, which gives a standard deviation of about 15. Thus, we can retain our Poisson model for the data, but explain the "extra" variability by variability in the underlying Poisson parameter.

□

2.4.3 Moment Generating Functions.

Moment generating functions and characteristic functions are primarily useful for one type of distributional calculation: computing the distribution of the sum of independent random variables. If X and Y are independent random variables with m.g.f's ψ_X and ψ_Y, then the m.g.f. of their sum is

$$\begin{aligned}\psi_{X+Y}(u) &= E\left[\exp\left\{u(X+Y)\right\}\right] \\ &= E\left[\exp\{uX\}\exp\{uY\}\right] \\ &= E\left[\exp\{uX\}\right] E\left[\exp\{uY\}\right] \\ &= \psi_X(u)\psi_Y(u)\end{aligned}$$

where independence was used at the third equality to write the expectation of the product as the product of the expectations. The results extends to random vectors, characteristic functions, and more than two independent summands. Of course, for the m.g.f. to be useful, we must have the conditions wherein uniqueness holds (Theorem 2.2.1(d)).

Exercises for Section 2.4.

2.4.1 (a) Assume that X is a r.v. with c.d.f. F and Lebesgue density $f = dF/dx$. Assume there is an open interval (a, b) such that $P[X \in (a, b)] = 1$. Let $h(x)$ be a transformation which is differentiable and on (a, b) and satisfies $h'(x) > 0$ for all $x \in (a, b)$. Using an argument based on c.d.f.'s show that the r.v. $Y = h(X)$ has a Lebesgue density

$$g(y) \; = \; f(h^{-1}(y)) \frac{1}{h'(h^{-1}(y))} .$$

(b) Same as part (a) but suppose $h'(x) < 0$.

(c) Derive the results in parts (a) and (b) using Proposition 2.4.2.

2.4.2 The *chain rule* for vector valued functions of several variables states that if $h : U \longrightarrow V$ and $g : V \longrightarrow I\!\!R^k$ with $U \subset I\!\!R^n$ and $V \subset I\!\!R^m$ both open sets and h and g differentiable, then the composite function $g \circ h$ is differentiable and

$$D(g \circ h)(x) \; = \; Dg(h(x))Dh(x) \quad .$$

(a) Verify that the matrices on the right hand side of the equation above have appropriate dimensions so the product is defined, and that the dimensions of the product equal the dimensions of the matrix on the left hand side.

(b) Use the chain rule applied to the equation $h \circ (h^{-1}) = \iota$ with ι the identity to verify equation (2.67).

(c) With the notation and hypotheses of Proposition 2.4.2, show that

$$g(y) \; = \; f(h^{-1}(y)) |\det Dh(h^{-1}(y))|^{-1} \quad , \text{ for all } y \in \Lambda \quad .$$

2.4.3 Verify equations (2.68) and (2.69).

2.4.4 As in Exercise 2.3.3, let $\underline{Z}$ be a random n-vector with independent $N(0, 1)$ components, V a given nonnegative definite matrix, and $\underline{\mu}$ a given n-vector.

(a) Suppose A is an $n \times n$ matrix with $AA^t = V$. Show that $\det(A) = [\det(V)]^{1/2}$. Assuming V is nonsingular, show that any such square matrix A satisfying $AA^t = V$ must also be nonsingular.

(b) Show that if $\det(V) \neq 0$, then the $N(\underline{\mu}, V)$ distribution is dominated by m^n and the Lebesgue density is given by (2.43).

2.4.5 Let X and Y be independent r.v.'s with the χ^2_n and χ^2_m distributions, respectively. Let $W = (X/n)/(Y/m)$. Show that W has Lebesgue density given by

$$f_W(w) \; = \; \frac{\Gamma((n+m)/2)(n/m)^{n/2}}{\Gamma(n/2)\Gamma(m/2)} \frac{w^{n/2-1}}{(1 + nw/m)^{(n+m)/2}} \quad , \quad w > 0 \, .$$

The probability measure with this Lebesgue density is known as the *F–distribution with n and m degrees of freedom* (or *n degrees of freedom for the numerator and m degrees of freedom for the denominator.* The "F" comes from the first letter of the last name of Sir Ronald A. Fisher who is perhaps the greatest statistician of all times and the first to recognize the importance of the F–distribution.

2.4.6 Suppose T has Student's t-distribution with ν degrees of freedom. For what values of m is the moment $E[T^m]$ defined? Show that for such values,

$$E[T^m] \;=\; \nu^{-m/2} E[Z^m] E[V^{-m/2}]$$

where $Z \sim N(0,1)$ and $V \sim \chi^2_\nu$ are independent.

2.4.7 Here we derive some formulae for the Lebesgue density of the so-called *noncentral t-distribution.* Let X and Y be independent random variables with

$$\text{Law}[X] \;=\; N(\delta, 1) \quad , \quad \text{Law}[Y] \;=\; \chi^2_n \; .$$

As in Example 2.4.2, define

$$T \;=\; \frac{X}{\sqrt{Y/n}} \; .$$

Then we say T has the noncentral t-distribution with n *degrees of freedom* and *noncentrality parameter δ.* Derive the following formulae for the Lebesgue density of this distribution:

$$\begin{aligned}
& f(t|\delta) \\
&= \frac{1}{\sqrt{\pi\nu}\,\Gamma(\nu/2) 2^{(\nu+1)/2}} \int_0^\infty s^{(\nu-1)/2} \exp\left[-(t\sqrt{s/\nu} - \delta)^2/2 \;-\; s/2 \right] ds \\
&= \frac{1}{\sqrt{\pi\nu}\,\Gamma(\nu/2) 2^{(\nu+1)/2}} e^{-\delta^2/2} \int_0^\infty s^{(\nu-1)/2} \exp\left[-s(1+t^2/\nu)/2 \;+\; \sqrt{s}\delta t/\sqrt{\nu} \right] ds \\
&= \frac{1}{\sqrt{\pi\nu}\,\Gamma(\nu/2) 2^{(\nu+1)/2}} e^{-\delta^2/2} \int_0^\infty s^{(\nu-1)/2} \exp\left[-s(1+t^2/\nu)/2 \right] \sum_{k=0}^\infty \frac{1}{k!} (\sqrt{s}\delta t/\sqrt{\nu})^k \, ds \\
&= \frac{1}{\sqrt{\pi\nu}\,\Gamma(\nu/2) 2^{(\nu+1)/2}} e^{-\delta^2/2} \sum_{k=0}^\infty \frac{1}{k!} \left(\frac{\delta t}{\sqrt{\nu}} \right)^k \int_0^\infty s^{(k+\nu-1)/2} \exp\left[-s(1+t^2/\nu)/2 \right] ds \\
&= \frac{e^{-\delta^2/2}}{\Gamma(1/2)\Gamma(\nu/2)\sqrt{\nu}} \sum_{k=0}^\infty \frac{(2/\nu)^{k/2}(\delta t)^k}{k!} \frac{\Gamma([\nu+k+1]/2)}{(1+t^2/\nu)^{(\nu+k+1)/2}},
\end{aligned}$$

2.4.8 Show that Theorem 2.4.1 follows from Theorem 10.9 of Rudin.

2.4.9 Show that the marginal distribution of X in Example 2.4.3 is $NB(1, 1/2)$, where the negative binomial family $NB(m, p)$ is defined in Exercise 2.3.10 (d).

2.5 Order Statistics.

In this section, we investigate some ideas which are very useful in statistics. One of the unifying concepts of statistics is empirical distributions, which was introduced in Chapter 1 (see equation 1.4). For definiteness, let X be a random variable. It will usually be the case that X is the outcome of a measurement in an experiment, and that the experiment is repeatable, so that we may obtain further "replications" of X. For instance, consider the experiment of selecting a person at random (so the underlying probability space is the set of all people), and then measuring height, so that X is the height of the randomly selected person in some units (e.g. inches). Of course, we may obtain more replications of X by selecting more people. It may be reasonable to assume that the different replications or *trials* of the experiment are independent and identically distributed. Thus, using a subscript to denote the outcome of the i'th trial, we would obtain in n trials X_1, X_2, ..., X_n which are *independent and identically distributed* (abbreviated *i.i.d.*) where the common distribution is Law$[X] = P_X$. Note that Law$[X]$ is a probability measure on $(I\!\!R, \mathcal{B})$. As discussed back in Chapter 2, section 2, we need not be too concerned with the rather "messy" underlying probability space of people, but can focus on the real numbers (where are measurements lie) and distributions thereon. Sometimes, X_1, X_2, ..., X_n are referred to as a *random sample from* P_X. Here, n is the *sample size* or the number of trials. We can construct an "estimator" for P_X given by

$$\hat{P}_n \;=\; \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i} \;\; .$$

Now for a fixed ω (which gives observed values $X_1(\omega)$, ..., $X_n(\omega)$ which are fixed real numbers), $\hat{P}_n$ is a probability measure. Indeed, $\hat{P}_n$ with the random X_i's is a random probability measure on $I\!\!R$. From (1.30), if h is a real valued function on $I\!\!R$ then

$$\int h(x)\, d\hat{P}_n(x) \;=\; \frac{1}{n}\sum_{i=1}^{n} h(X_i) \;\; , \tag{2.85}$$

i.e. integration w.r.t. $\hat{P}_n$ amounts to averaging the function h over the sample. This is also a random variable (with the X_i's appearing as in (2.85)). Thus,

$$E\left[\int h(x)\, d\hat{P}_n(x)]\right] \;=\; E[h(X)] \;\; . \tag{2.86}$$

This latter equation says that $(1/n)\sum_{i=1}^{n} h(X_i)$ is an *unbiased estimator* of $E[h(X)]$.

In Chapter 1, Section 1, we introduced the following simple rule: To estimate a functional of an unknown probability distribution, simply replace the unknown probability distribution by the empirical distribution. Here, a "functional" of a distribution is a function which assigns a real number to each probability distribution (in a certain class). Thus, if h is a given Borel function of a real variable, then the map $H(P_X) = \int h(x)dP_X(x)$ is a functional defined on all Borel p.m.'s P_X for which the integral exists and is finite (e.g. $h(x) = x$ gives the mean functional). Another functional we may wish to estimate is the minimal α quantile $(0 \leq \alpha \leq 1)$ $F^-(\alpha) = \inf\{x : F(x) \geq \alpha\}$, where F is the c.d.f. of X. For fixed α, $F \mapsto F^-(\alpha)$ is a functional on all distributions on $I\!R$. Replacing F in the definition by the empirical distribution function $\hat{F}_n$ gives the minimal α sample quantile $\hat{F}_n^-(\alpha)$. In general, we don't have the relation that $E[\hat{F}^-(\alpha)] = F^-(\alpha)$ as in (2.86). That is, the sample quantile is generally a biased estimator of the true quantile. See Example 2.5.1 below. But the estimate is still a very natural one, and the bias is generally quite small. See Exercise 1.1.17 for more on sample quantiles.

Above we spoke of $\hat{P}_n$ as being a "random probability measure". Such a random object is not well defined at this point because we have not introduced a σ–field on the set of probability measures on $I\!R$. Also, we don't know what it means for $\hat{F}_n$ to be a "random distribution function" since we haven't introduced a σ–field on the set of cumulative distribution functions. However, $\hat{P}_n$ has a very special form since it is discrete, supp$[\hat{P}_n]$ has at most n points (exactly n points if all values in the sample are distinct), and the amount of probability mass at each point is a positive integer times $1/n$ (exactly $1/n$ if the points are distinct). Thus, we can think of the subset of such probability measures, which is "isomorphic" with a Euclidean space. Put less technically, we only need a finite number of numbers to determine $\hat{P}_n$, e.g. n numbers where n is the sample size, since if we know all n observed values then we know $\hat{P}_n$. Similar remarks hold for $\hat{F}_n$. However, the mapping from $I\!R^n$ to discrete probability measures given by

$$p(x_1, x_2, ..., x_n) \;=\; \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i} \quad , \quad x_i \in I\!R \text{ for } 1 \leq i \leq n \tag{2.87}$$

is not one to one since we can't reconstruct the order of the observations from $\hat{P}_n$. For instance, if π is a permutation of $\{1, 2, ..., n\}$, then

$$p(x_{\pi(1)}, ..., x_{\pi(n)}) \;=\; p(x_1, ..., x_n) \ .$$

Recall that a *permutation* of $\{1, 2, ..., n\}$ is a one to one correspondence (bijective map) of the finite set with (into) itself. Thus, if π is a permutation of $\{1, 2, ..., n\}$, then $\{\pi(1), \pi(2), ..., \pi(n)\}$ is simpy a reordering of $\{1, 2, ..., n\}$. This last displayed equation merely states the obvious fact that if we reorder the observations, then we get the same empirical probability.

2.5.1 Basic Results.

What we have said above can easily be extended to observations which are random vectors, but now we will use the order properties of real numbers. Let $\underline{X} = (X_1, X_2, ..., X_n)$ denote the vector of all observations. Consider the subset of $I\!R^n$ given by

$$\mathbb{P}^n = \{\underline{x} \in \mathbb{R}^n : x_1 \leq x_2 \leq ... \leq x_n\} .$$

On $\mathbb{P}^n$, the mapping p above is a one to one correspondence, so we can identify the set of possible empirical probability distributions (or empirical c.d.f.'s) with $\mathbb{P}^n$, since given an element of $\mathbb{P}^n$, we can associate a unique empirical probability distribution, and vice versa.

The mapping which "orders" our sample $\underline{X} = (X_1, X_2, ..., X_n)$ so that it becomes a random vector taking values in $\mathbb{P}^n$ will be denoted **Sort**, i.e. $\underline{Y} = \mathbf{Sort}(\underline{X})$ means $\underline{Y} \in \mathbb{P}^n$ and there is a permutation π of $\{1, 2, ..., n\}$ such that $Y_i = X_{\pi(i)}$ for all i. That is, the components of $\underline{Y}$ are obtained by rearranging the components of $\underline{X}$ in ascending order. $\underline{Y}$ is known as the vector of *order statistics*. Two notations for the i'th component Y_i of $\underline{Y}$ that are frequently used are $X_{(i)}$ and $X_{i:n}$. Intuitively, if we believe the components of $\underline{X}$ are i.i.d., then $\mathbf{Sort}(\underline{X})$ contains "as much information" about the unknown probability distribution as the original vector of observations $\underline{X}$. We will show below in fact that given $\mathbf{Sort}(\underline{X})$, it is possible to "reconstruct" $\underline{X}$ in the sense that we can obtain a random vector with the same distribution.

Let $\mathbf{Perm}(n)$ denote the set of all permutations of $\{1, 2, ..., n\}$, then $\#\mathbf{Perm}(n) = n!$, of course. Note that $\mathbf{Perm}(n)$ has the following properties:

(i) If π_1 and π_2 are in $\mathbf{Perm}(n)$, then so is $\pi_1 \circ \pi_2$.

(ii) There is an element $\iota \in \mathbf{Perm}(n)$ such that $\iota \circ \pi = \pi \circ \iota = \pi$ for every $\pi \in \mathbf{Perm}(n)$.

(iii) For every $\pi \in \mathbf{Perm}(n)$, there is an element $\pi^{-1} \in \mathbf{Perm}(n)$ such that $\pi \circ \pi^{-1} = \pi^{-1} \circ \pi = \iota$.

These three properties make $\mathbf{Perm}(n)$ into a *group* under the *(group) operation* of composition (i.e. $\circ$). Note that $\mathbb{R}$ is a group under $+$, and both $\mathbb{R} - \{0\}$ and $(0, \infty)$ are groups under multiplication. Now define

$$\zeta \circ \mathbf{Perm}(n) = \{\zeta \circ \pi : \pi \in \mathbf{Perm}(n)\}.$$

Using the properties just discussed, one can show that for all $\zeta \in \mathbf{Perm}(n)$,

$$\zeta \circ \mathbf{Perm}(n) = \mathbf{Perm}(n) . \qquad (2.88)$$

See Exercise 2.5.1.

We will write **Perm** when n is clear from context. To each permutation $\pi \in \mathbf{Perm}(n)$ there corresponds a unique linear transformation $\tilde{\pi}$ on $\mathbb{R}^n$ which reorders the components of a vector, viz.

$$\tilde{\pi}(y_1, y_2, ...y_n) = (y_{\pi(1)}, y_{\pi(2)}, ..., y_{\pi(n)}) .$$

One can easily see that the $n \times n$ matrix corresponding to $\tilde{\pi}$ is A where $A_{ij} = 1$ if $\pi(j) = i$ and otherwise $A_{ij} = 0$. Note that there is a single 1 in every row and in every column of A, and the remaining entries are 0. Such a matrix is called a *permutation matrix*. Also, one can show that $A^{-1} = A'$, i.e. A is an orthogonal matrix (Exercise 2.5.2).

If π is any permutation, then clearly $\mathbf{Sort}(\tilde{\pi}\underline{x}) = \mathbf{Sort}(\underline{x})$, i.e. if we permute the components of $\underline{x}$ and then rearrange permuted components into ascending order, we obtain the same result as if we didn't permute the components before ordering them. Thus, we say **Sort** is *invariant under coordinate permutations*, or simply *permutation invariant*. Now, we characterize some measurability properties of the mapping $\mathbf{Sort} : I\!\!R^n \longrightarrow I\!\!P^n$.

Theorem 2.5.1 *(a) A Borel set B is $\sigma(\mathbf{Sort})$ measurable iff it satisfies the following symmetry property: $\underline{x} \in B$ imples $\tilde{\pi}\underline{x} \in B$ for all $\pi \in \mathbf{Perm}$.*

A function $h : I\!\!R^n \longrightarrow I\!\!R$ is $\sigma(\mathbf{Sort})$ measurable iff it is invariant under permutations of the variables, i.e. $h \circ \tilde{\pi} = h$ for all $\pi \in \mathbf{Perm}$.

(b) Suppose $\underline{X}$ is a random n-vector with i.i.d. components and continuous one dimensional marginal c.d.f. Then

$$P[\mathbf{Sort}(\underline{X}) \in D] \;=\; n!P[\underline{X} \in D] \quad , \quad \text{for } D \subset I\!\!P^n \; . \tag{2.89}$$

In particular, if X_1 has a Lebesgue density f, then under the i.i.d. assumption, $\underline{Y} = \mathbf{Sort}(\underline{X})$ has a Lebesgue density (on $I\!\!R^n$) given by

$$f_{\underline{Y}}(\underline{y}) \;=\; \begin{cases} n! \prod_{i=1}^n f(y_i) & \text{if } \underline{y} \in I\!\!P^n, \\ \\ 0 & \text{if } \underline{y} \notin I\!\!P^n \; . \end{cases} \tag{2.90}$$

(c) If $\underline{X}$ has i.i.d. components with continuous c.d.f., as in part (b), then

$$Law[\underline{X}|\mathbf{Sort}(\underline{X}) = \underline{y}] \;=\; \frac{1}{n!} \sum_{\pi \in \mathbf{Perm}} \delta_{\tilde{\pi}\underline{y}} \; . \tag{2.91}$$

Hence,

$$E[h(\underline{X})|\mathbf{Sort}(\underline{X})] \;=\; \frac{1}{n!} \sum_{\pi \in \mathbf{Perm}} h(\tilde{\pi}\underline{X}) \; . \tag{2.92}$$

Remarks 2.5.1 Note that for each fixed $\underline{y} \in I\!\!P^n$, Law$[\underline{X}|\mathbf{Sort}(\underline{X}) = \underline{y}]$ is a p.m. on $I\!\!R^n$. To paraphrase the result in (2.91), given the order statistics, each of the $n!$ possible permutations of the data is equally likely.

Proof. (a) We claim that for $A \subset I\!\!P^n$,

$$\mathbf{Sort}^{-1}(A) \;=\; \bigcup_{\pi \in \mathbf{Perm}} \tilde{\pi}^{-1}(A) \; . \tag{2.93}$$

Let $\underline{x} \in I\!R^n$ and let $\underline{y} = \mathbf{Sort}(\underline{x})$. Then $\underline{y}$ is obtained by rearranging the components of $\underline{x}$ into ascending order. So if $\mathbf{Sort}(\underline{x}) \in A$, then $\tilde{\pi}\underline{x} \in A$ for some permutation π, and hence $\mathbf{Sort}^{-1}(A) \subset \bigcup_\pi \{\underline{x} : \tilde{\pi}\underline{x} \in A\}$. On the other hand, if $\tilde{\pi}\underline{x} \in A$ for some π, then $\tilde{\pi}\underline{x} = \mathbf{Sort}(\underline{x})$ since $A \subset I\!P^n$ and $\mathbf{Sort}(\underline{x})$ is the unique element of $I\!P^n$ that can be obtained by permuting the components of $\underline{x}$, so $\mathbf{Sort}(\underline{x}) \in A$, and we have shown that $\bigcup_\pi \{\underline{x} : \tilde{\pi}\underline{x} \in A\} \subset \mathbf{Sort}^{-1}(A)$. This completes the proof of (2.93).

Suppose $B \in \sigma(\mathbf{Sort})$ so by (2.93), $B = \bigcup_\pi \tilde{\pi}^{-1}A$ for some Borel set $A \subset I\!P^n$. If $\zeta \in \mathbf{Perm}$ then

$$\begin{aligned} \tilde{\zeta}^{-1}B \;&=\; \bigcup_{\pi\in\mathbf{Perm}} \tilde{\zeta}^{-1}\tilde{\pi}^{-1}A \\ &= \bigcup_\pi (\tilde{\pi}\circ\tilde{\zeta})^{-1}A \;=\; \bigcup_\pi \tilde{\pi}^{-1}A \;. \end{aligned} \tag{2.94}$$

The last equality follows from (2.88). This shows that B is symmetric. Conversely, if B is symmetric, then it is easy to see that $B = \bigcup_\pi \tilde{\pi}^{-1}A$ with $A = B \cap I\!P^n$, and hence B is $\sigma(\mathbf{Sort})$ measurable.

By Theorem 1.5.1, h is $\sigma(\mathbf{Sort})$ measurable iff there is a $g : I\!P^n \longrightarrow I\!R$ such that $h = g \circ (\mathbf{Sort})$. It follows that if h is $\sigma(\mathbf{Sort})$ measurable then $h \circ \tilde{\pi} = g \circ (\mathbf{Sort}) \circ \tilde{\pi} = g \circ (\mathbf{Sort}) = h$ since $(\mathbf{Sort}) \circ \tilde{\pi} = \mathbf{Sort}$ for any $\pi \in \mathbf{Perm}$. Conversely, suppose $h = h \circ \tilde{\pi}$ for all $\pi \in \mathbf{Perm}$. Now for every $\underline{x} \in I\!R^n$, $\mathbf{Sort}(\underline{x})$ is obtained by a permutation of the components of $\underline{x}$, so $h(\underline{x}) = h(\mathbf{Sort}(\underline{x}))$ so h is $\sigma(\mathbf{Sort})$ measurable by Proposition 1.2.3.

(b) From (2.93), if $D \subset I\!P^n$ then

$$\begin{aligned} P[\mathbf{Sort}(\underline{X}) \in D] \;&=\; P\left[\underline{X} \in \mathbf{Sort}^{-1}(D)\right] \\ &=\; P\left[\,\underline{X} \in \bigcup_\pi \tilde{\pi}^{-1}(D)\,\right], \end{aligned}$$

where the union is over all $\pi \in \mathbf{Perm}$. Now we claim that if $\pi \neq \zeta$, then

$$P\left[\underline{X} \in \tilde{\pi}^{-1}(D) \cap \tilde{\zeta}^{-1}(D)\right] \;=\; 0.$$

Assuming the claim is true, it follows that the sets in the union $\bigcup_\pi \tilde{\pi}^{-1}(D)$ are "essentially disjoint" and hence

$$P\left[\underline{X} \in \bigcup_\pi \tilde{\pi}^{-1}(D)\right] \;=\; \sum_\pi P[\underline{X} \in \tilde{\pi}^{-1}D]. \tag{2.95}$$

Here, by "essentially disjoint" we mean that the intersection has probability (measure) 0. This will complete the proof of this part of the theorem.

We hope that these claims are fairly obvious, but for the sake of mathematical formalism, we will show that

$$I_{\bigcup_\pi \tilde{\pi}^{-1}D}(\underline{x}) \;=\; \sum_{\pi\in\mathbf{Perm}} I_{\tilde{\pi}^{-1}D}(\underline{x})\;, \quad \text{for Law}[\underline{X}] \text{ almost all } \underline{x} \qquad , \tag{2.96}$$

Taking expectations (i.e. integrating w.r.t. the distribution of $\underline{X}$) of both sides gives (2.95). Now for given $\underline{x}$ the sum on the l.h.s. of (2.96) is the number of sets $\tilde{\pi}^{-1}D$ to which $\underline{x}$ belongs, and $\underline{x}$ belongs to two or more $\tilde{\pi}^{-1}D$ iff there is a pair of distinct permutations π and ζ such that $\underline{x} \in (\tilde{\pi}^{-1}D) \cap (\tilde{\zeta}^{-1}D)$. However, this means that $\underline{x} = \tilde{\pi}\underline{y} = \tilde{\zeta}\underline{y}$ for some $\underline{y} \in D$, where π and ζ are distinct permutations. However, when π and ζ are distinct permutations it is true that

$$(\tilde{\pi}^{-1}D) \cap (\tilde{\zeta}^{-1}D) \subset \{\underline{x} : x_i = x_j \text{ for some } i \neq j\} \ . \tag{2.97}$$

To see this, note that π and ζ being distinct permutations implies $\pi(k) \neq \zeta(k)$ for some $k \in \{1, \ldots, n\}$. Now suppose $\underline{x} \in (\tilde{\pi}^{-1}D) \cap (\tilde{\zeta}^{-1}D)$, where π and ζ are distinct permutations. This means $\underline{x} = \tilde{\pi}\underline{y} = \tilde{\zeta}\underline{y}$ for some $\underline{y} \in D$. But because π and ζ are distinct permutations, it follows that $y_{\pi(k)} = y_{\zeta(k)}$, and taking $i = \pi(k)$ and $j = \zeta(k)$, we have $i \neq j$ but $x_i = y_{\pi(k)} = x_j = y_{\zeta(k)}$, and hence $x \in \{\underline{x} : x_i = x_j \text{ for some } i \neq j\}$. This establishes (2.97). Now (2.97) implies the inequality

$$P_{\underline{X}}[(\tilde{\pi}^{-1}D) \cap (\tilde{\zeta}^{-1}D)] \ \leq \ P_{\underline{X}}(\{\underline{x} : x_i = x_j \text{ for some } i \neq j\} \ = \ 0 \ . \tag{2.98}$$

The equality in (2.98) follows by the assumption that the common c.d.f. of the X_i is continuous. This implies that $P[X_i = x] = 0$ for every $x \in I\!\!R$. One can then apply the argument of Exercise 1.3.18 with Lebesgue measure replaced by P_X, the common one dimensional marginal, to obtain $P_{\underline{X}}\{\underline{x} : x_i = x_j$ for some $i \neq j\} = 0$.

(c) For $B \in \mathcal{B}_n$ and $\underline{y} \in I\!\!P^n$, let

$$p(B, \underline{y}) \ = \ \frac{1}{n!} \sum_{\pi \in \mathbf{Perm}} \delta_{\tilde{\pi}\underline{y}}(B) \ .$$

For fixed $\underline{y}$, $p(\cdot, \underline{y})$ is clearly a p.m. Thus, we need to show

$$P[\underline{X} \in B | \mathbf{Sort}(\underline{X}) = \underline{y}] \ = \ \frac{1}{n!} \sum_{\pi \in \mathbf{Perm}} I_B(\tilde{\pi}\underline{y}) \ , \tag{2.99}$$

$$\text{for Law}[\mathbf{Sort}(\underline{X})] \text{ almost all } \underline{y} \ ,$$

i.e. that $p(B, \underline{y})$ is a version of $P[\underline{X} \in B | \mathbf{Sort}(\underline{X}) = \underline{y}]$. Clearly the r.h.s. of (2.99) is a Borel function of $\underline{y}$. Thus, we need to check that if $A \subset \Omega$ which is $\sigma(\mathbf{Sort}(\underline{X}))$ measurable, then

$$\int_A I_B(\underline{X}) \, dP \ = \ \int_A p(B, \mathbf{Sort}(\underline{X})) \, dP \ . \tag{2.100}$$

Since A is in $\sigma(\mathbf{Sort}(\underline{X}))$ it follows that $A = \underline{X}^{-1}(C)$ for some $C \in \sigma(\mathbf{Sort})$, and by (2.93), $C = \bigcup_\pi \tilde{\pi}^{-1}D$ for some Borel $D \subset I\!\!P^n$. Hence, by the change of variables

$$\int_A p(B, \mathbf{Sort}(\underline{X})) \, dP \ = \ \int_C p(B, \mathbf{Sort}(\underline{x})) \, dP_{\underline{X}}(\underline{x})$$

$$= \int_{\bigcup_\pi \tilde{\pi}^{-1}D} p(B, \mathbf{Sort}(\underline{x}))\, dP_{\underline{X}}(\underline{x}) \ . \tag{2.101}$$

Now (2.96) allows us to break the integral in (2.101) up into a sum over each of the $\tilde{\pi}^{-1}D$ and hence

$$\int_A p(B, \mathbf{Sort}(\underline{X}))\, dP \ = \ \sum_{\pi \in \mathbf{Perm}} \int_{\tilde{\pi}^{-1}D} p(B, \mathbf{Sort}(\underline{x}))\, dP_{\underline{X}}(\underline{x})$$

$$= \sum_{\pi \in \mathbf{Perm}} \int_D p(B, \mathbf{Sort}(\tilde{\pi}^{-1}\underline{w}))\, dP_{\tilde{\pi}\underline{X}}(\underline{w}) \ . \tag{2.102}$$

In the latter equality, we made the change of variables within each integral that $\underline{w} = \tilde{\pi}\underline{x}$. Note that if $\underline{W} = \tilde{\pi}\underline{X}$ then Law[$\underline{W}$] = Law[$\tilde{\pi}\underline{X}$]. However, Law[$\tilde{\pi}\underline{X}$] = Law[$\underline{X}$] since

$$P_{\tilde{\pi}\underline{X}} \ = \ \prod_{i=1}^n P_{X_{\pi(i)}} \ = \ \prod_{i=1}^n P_{X_1} \tag{2.103}$$

because X_1, ..., X_n all have the same marginal distribution. This shows that

$$P_{\tilde{\pi}\underline{X}} \ = \ P_{\underline{X}} \ .$$

Hence, plugging this back into (2.102) and using the fact that **Sort** is permutation invariant gives

$$\int_A p(B, \mathbf{Sort}(\underline{X}))\, dP \ = \ \sum_{\pi \in \mathbf{Perm}} \int_D p(B, \mathbf{Sort}(\underline{w}))\, dP_{\underline{X}}(\underline{w})$$

$$= n! \int_D p(B, \mathbf{Sort}(\underline{x}))\, dP_{\underline{X}}(\underline{x}) \ . \tag{2.104}$$

Now we substitute the form of $p(B, \mathbf{Sort}(\underline{x}))$ into this last expression to obtain

$$\int_A p(B, \mathbf{Sort}(\underline{X}))\, dP \ = \ n! \int_D \left[\frac{1}{n!} \sum_{\pi \in \mathbf{Perm}} I_B(\tilde{\pi}\mathbf{Sort}(\underline{x})) \right] dP_{\underline{X}}(\underline{x}) \tag{2.105}$$

$$= \sum_{\pi \in \mathbf{Perm}} \int_D I_B(\tilde{\pi}\mathbf{Sort}(\underline{x}))\, dP_{\underline{X}}(\underline{x}) \ = \ \int_D \left[\sum_{\pi \in \mathbf{Perm}} I_B(\tilde{\pi}\mathbf{Sort}(\underline{x})) \right] dP_{\underline{X}}(\underline{x}) \ .$$

Note that for a given $\underline{x}$, $\tilde{\pi}\mathbf{Sort}(\underline{x})$ ranges over the same collection of values as $\tilde{\pi}\underline{x}$ when π ranges over all of **Perm**, so the last expression in (2.105)

$$= \int_D \left[\sum_{\pi \in \mathbf{Perm}} I_B(\tilde{\pi}\underline{x}) \right] dP_{\underline{X}}(\underline{x}) \ = \ \sum_{\pi \in \mathbf{Perm}} \int_D I_B(\tilde{\pi}\underline{x})\, dP_{\underline{X}}(\underline{x})$$

$$= \sum_{\pi \in \mathbf{Perm}} \int_{\tilde{\pi}^{-1}D} I_B(\underline{w})\, dP_{\tilde{\pi}\underline{X}}(\underline{w})$$

where the latter follows as in (2.102), and by (2.103) this in turn

$$= \sum_{\pi \in \mathbf{Perm}} \int_{\tilde{\pi}^{-1}D} I_B(\underline{w})\, dP_{\underline{X}}(\underline{w}) \; = \int_{\bigcup_\pi \tilde{\pi}^{-1}D} I_B(\underline{x})\, dP_{\underline{X}}(\underline{x})$$

by the same argument as in (2.96), which

$$= \int_C I_B(\underline{x})\, dP_{\underline{X}}(\underline{x}) \; = \int_A I_B(\underline{X})\, dP \; .$$

which completes the proof.

□

Part (c) of the above theorem in combination with the Two Stage Experiment Theorem tells us that if we are given the order statistics **Sort**$(\underline{X})$, then we may obtain a "probabilistic replica" of the original sample $\underline{X}$ by choosing a permutation from **Perm**(n) at random (i.e. using the uniform distribution) and applying that permutation to reorder the **Sort**$(\underline{X})$ in a random fashion. Of course, it would in general be silly to do this, but if someone insisted that the data be made to look "realistic" it could be done. In general, one would think that statistical methods for use on i.i.d. samples should not depend on the order of the observations. We shall investigate this in a subsequent chapter.

As an example, we consider the following. The ages of 10 persons in a study (this was a sample from a total of 200) are

61 63 50 83 50 60 60 54 53 54

These data are given above in their original order, but suppose they were sorted before being given to the statistician who is working on the project. The following Splus code converts the data in ascending order to a random order that "looks" more realistic as a possible order in which the data were taken.

```
> ages_c(50,50,53,54,54,60,60,61,63,83)
> #these data are already sorted in ascending order!?
> #Rearrange in a random order:
> ages_sample(ages)
> ages
 [1] 53 83 60 63 54 54 61 60 50 50
> #Now this looks "realistic".
> #Of course, we can always recover the order statistics:
> sort(ages)
 [1] 50 50 53 54 54 60 60 61 63 83
```

Note that comments in the Splus code are on lines beginning with a #. The functions "`sample`" and "`sort`" are supplied functions in Splus. The function "`c`" simply creates a vector from a list. The underline character "`_`" denotes assignment. If an object (in this case, a vector or 1 dimensional array of numbers) is simply listed without any assignment, then it is printed out (as in the line where `ages` appears, or the last line where `sort(ages)` appears).

2.5.2 Some Applications.

One can use (2.90) to derive the marginal distribution of individual order statistics, but the following approach is easier. Assume $\underline{X}$ has i.i.d. components and let $X_{(i)}$ denote the i'th order statistic (i.e. the i'th component of $\mathbf{Sort}(\underline{X})$). Then we can derive the c.d.f. of $X_{(i)}$ directly by

$$\begin{aligned} P[X_{(i)} \leq x] &= P[\text{ at least } i \text{ components of } \underline{X} \text{ are } \leq x] \\ &= P[\sum_{j=1}^{n} I_{(-\infty,x]}(X_j) \;\geq\; i\,] = \sum_{j=i}^{n} \binom{n}{j} F(x)^j [1-F(x)]^{n-j} \qquad (2.106) \end{aligned}$$

where the latter results from the fact that the $I_{(-\infty,x]}(X_j)$ are independent Bernoulli r.v.'s with "success" probability $F(x)$, so the summation in (2.106) is a binomial r.v. with parameters n and $p = F(x)$, written as $B(n, F(x))$. Assuming the X_i's have Lebesgue density f, then differentiating the c.d.f. in (2.106) gives the Lebesgue density function of the i'th order statistic as

$$\begin{aligned} f_{X_{(i)}}(x) &= i \binom{n}{i} F(x)^{i-1} [1-F(x)]^{n-i} f(x) \qquad (2.107) \\ &= n \binom{n-1}{i-1} F(x)^{i-1} [1-F(x)]^{n-i} f(x) \;. \qquad (2.108) \end{aligned}$$

See Exercise 2.5.5.

Example 2.5.1 Let X_1, X_2, …, X_n be i.i.d. $Unif(0,1)$. Then the Lebesgue density for the i'th order statistic is

$$\begin{aligned} f_i(x) &= \frac{n!}{(i-1)!(n-i)!} x^{i-1}(1-x)^{n-i} \;, \quad 0 < x < 1 \\ &= \frac{\Gamma(n+1)}{\Gamma(i)\Gamma(n+1-i)} x^{i-1}(1-x)^{n-i} \;, \quad 0 < x < 1. \end{aligned}$$

The latter form shows that this is a $Beta(n+1, i)$ distribution. It is straightforward to show that

$$E[X_{(i)}] = \frac{\Gamma(i+1)\Gamma(n-i)}{\Gamma(i)\Gamma(n+1-i)} = \frac{i}{n+1}. \qquad (2.109)$$

In particular, we know that $X_{(i)} = \hat{F}_n^-(\alpha)$ for $\alpha \in ((i-1)/n, i/n]$, so we see that $\hat{F}_n^-(\alpha)$ is an unbiased estimator of α only for the particular value $\alpha = i/(n+1)$.

□

Example 2.5.2 Let X_1, X_2, …, X_n be i.i.d. *Expo*(1) and $X_{(1)}$, $X_{(2)}$, …, $X_{(n)}$ the corresponding order statistics. One can of course derive the joint and marginal distributions of the order statistics, but in this setting one achieves a particularly nice result for the joint distribution of the *spacings* defined by

$$\begin{aligned} Y_1 &= X_{(1)}, \\ Y_i &= X_{(i)} - X_{(i-1)}, \quad 1 < i \leq n. \end{aligned}$$

We will employ the transformation theory based on Jacobians. The inverse transformation is

$$x_{(i)} = \sum_{j=1}^{i} y_j,$$

so the matrix of partial derivatives is

$$\frac{d\mathbf{Sort}(\underline{x})}{d\underline{y}} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 1 & 0 & \cdots & 0 & 0 \\ 1 & 1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 1 & 1 & \cdots & 1 & 0 \\ 1 & 1 & 1 & \cdots & 1 & 1 \end{bmatrix}.$$

Hence, the absolute determinant of the Jacobian is 1. Now by (2.90) the joint Lebesgue density of $X_{(1)}$, $X_{(2)}$, …, $X_{(n)}$ is

$$f(x_{(1)}, \ldots, x_{(n)}) = n! \exp\left[-\sum_{i=1}^{n} x_{(i)}\right], \quad x_{(1)} < x_{(2)} < \cdots < x_{(n)}.$$

Note that

$$\begin{aligned} \sum_{i=1}^{n} x_{(i)} &= \sum_{i=1}^{n}\sum_{j=1}^{i} y_j \\ &= \sum_{j=1}^{n}\sum_{i=j}^{n} y_j \\ &= \sum_{j=1}^{n} (n-j+1) y_j \end{aligned}$$

so we get for the Lebesgue density of the Y_i's

$$\begin{aligned} f(\underline{y}) &= n! \exp\left[-\sum_{j=1}^{n}(n-j+1)y_j\right], \quad y_j > 0,\ 1 \leq j \leq n, \\ &= \prod_{j=1}^{n} (n-j+1) e^{-(n-j+1)y_j}, \quad y_j > 0,\ 1 \leq j \leq n. \end{aligned}$$

Note that this is the Lebesgue density of independent random variables Y_i with $\text{Law}[Y_i] = Expo[1/(n-i+1)]$.

□

Example 2.5.3 Let X_1, X_2, …, X_n be i.i.d. $N(\mu, \sigma^2)$. Consider the *sample range*, given by

$$\text{Range}(\underline{X}) \;=\; \max(\underline{X}) \;-\; \min(\underline{X}) \;=\; X_{(n)} \;-\; X_{(1)}.$$

In quality control, it is common to estimate the standard deviation σ by a multiple of the sample range, i.e. to use the estimator

$$\hat{\sigma}_R \;=\; C_n \text{Range}(\underline{X}),$$

where C_n is chosen so that the estimator is unbiased, i.e.

$$E\left[\hat{\sigma}_R\right] \;=\; \sigma.$$

Such an estimator is used in this context because it is easy to calculate for a person who has limited knowledge and computational resources. In fact, this is not a very good estimator from the point of view of accuracy, an issue we will take up in a later chapter. Values for C_n are tabulated in Table 3.1 of Thompson and Koronacki. We will derive a formula for C_n and indicate how such tables can be constructed.

Firstly, we will show that we can reduce the problem to one of i.i.d. $N(0,1)$ observations. To this end, note that if Z_1, Z_2, …, Z_n are i.i.d. $N(0,1)$, then $X_i = \sigma Z_i + \mu$, $1 \le i \le n$, are i.i.d. $N(\mu, \sigma^2)$. Further, $\max(\underline{X}) = \sigma \max(\underline{Z}) + \mu$ and similarly for $\min(\underline{X})$, so $\text{Range}(\underline{X}) = \sigma \text{Range}(\underline{Z})$. Thus

$$E\left[C_n \text{Range}(\underline{X})\right] \;=\; C_n \sigma E\left[\text{Range}(\underline{Z})\right],$$

so if we take

$$C_n^{-1} \;=\; E[\text{Range}(\underline{Z})],$$

then

$$E\left[C_n \text{Range}(\underline{X})\right] \;=\; \sigma.$$

Now, notice that $\underline{Z}$ and $-\underline{Z}$ have the same distribution, and of course $\max(-\underline{Z}) = -\min(\underline{Z})$. Hence,

$$\begin{aligned}
E\left[\text{Range}(\underline{Z})\right] &= E\left[\max(\underline{Z})\right] - E\left[\min(\underline{Z})\right] \\
&= E\left[\max(\underline{Z})\right] + E\left[\max(-\underline{Z})\right] \\
&= E\left[\max(\underline{Z})\right] + E\left[\max(\underline{Z})\right] \\
&= 2E\left[\max(\underline{Z})\right].
\end{aligned}$$

So, we need only compute the expectation of a single order statistic.

Now the Lebesgue density for $Z_{(n)}$ is

$$f(z) \;=\; n\Phi(z)^{n-1}\phi(z),$$

where ϕ denotes the $N(0,1)$ density and Φ the $N(0,1)$ c.d.f. Thus,

$$E\left[\max(\underline{Z})\right] \;=\; n\int_{-\infty}^{\infty} z\phi(z)\Phi(z)^{n-1}\,dz.$$

One can calculate the integral numerically for different values of n and then produce tables as indicated above. We will investigate useful approximations when n is large in a subsequent chapter.

□

To obtain higher order marginal densities it is probably just as easy to integrate (2.90). Some effort can be saved by specializing to the i.i.d. $Unif(0,1)$ case and then employing a trick. Let $\underline{U}$ be a random n-vector with i.i.d. components which are uniformly distributed on $[0,1]$. If $F(x)$ is a given c.d.f., then $X_i = F^-(U_i)$ gives a random vector $\underline{X}$ with i.i.d. components having marginal c.d.f. F (Proposition 1.2.4). Furthermore, if $\underline{V} = \textbf{Sort}(\underline{U})$ is the vector of order statistics for the uniform sample, then

$$\underline{Y} \;=\; \textbf{Sort}(\underline{X}) \;=\; \left(F^-(V_1), F^-(V_2), \ldots, F^-(V_n)\right). \qquad (2.110)$$

Assuming further that Law$[X_i]$ has a Lebesgue density $f(x)$, one can show that

$$\frac{dv_i}{dy_i} \;=\; f(y_i). \qquad (2.111)$$

Hence, in particular, if $i < j$, then

$$f_{Y_i,Y_j}(y_i,y_j) \;=\; f_{V_i,V_j}(F(y_i),F(y_j))f(y_i)f(y_j). \qquad (2.112)$$

Now to compute a bivariate marginal Lebesgue density for V_i and V_j with $i < j$, we will use the integration formulae

$$\int_0^{v_i}\cdots\int_0^{v_3}\int_0^{v_2} dv_1\,dv_2\cdots dv_{i-1} \;=\; \frac{1}{(i-1)!}v_i^{i-1}, \qquad (2.113)$$

$$\int_{v_j}^{1}\cdots\int_{v_{n-2}}^{1}\int_{v_{n-1}}^{1} dv_n\,dv_{n-1}\cdots dv_{j+1} \;=\; \frac{1}{(n-j)!}(1-v_j)^{n-j}, \qquad (2.114)$$

$$\int_{v_i}^{v_j}\cdots\int_{v_i}^{v_{i+3}}\int_{v_i}^{v_{i+2}} dv_{i+1}\,dv_{i+2}\cdots dv_{j-1} \;=\; \frac{1}{(j-i+1)!}(v_j-v_i)^{j-i+1}. \qquad (2.115)$$

From these it follows that

$$f_{V_i,V_j}(v_i,v_j) \;=\; \frac{n!}{(i-1)!(j-i+1)!(n-j)!} v_i^{i-1}(v_j-v_i)^{j-i+1}(1-v_j)^{n-j},$$
$$0 < v_i < v_j < 1, \qquad (2.116)$$

and hence that

$$f_{Y_i,Y_j}(y_i,y_j) \;=\;$$
$$\frac{n!}{(i-1)!(j-i+1)!(n-j)!} F(y_i)^{i-1}[F(y_j)-F(y_i)]^{j-i+1}[1-F(y_j)]^{n-j} f(y_i)f(y_j),$$
$$y_i < y_j. \qquad (2.117)$$

2.5.3 Further Results.

We shall also need some results on the conditional distributions of order statistics. This turns out to be especially easy if the components have a uniform distribution, and then we can extend the results to other distributions by the trick employed above. So let $\underline{U}_n$ be a random n-vector with i.i.d. components which are uniformly distributed on $[0,1]$ and put $\underline{V}_n = \mathbf{Sort}(\underline{U}_n)$. It will be convenient to include the sample size n in the notation here. For $1 \le i \le j \le n$, define $\underline{V}_n[i:j] = (V_{i,n},\, V_{i+1,n},\, ...,\, V_{j,n})$ to be the random $j-i+1$-vector obtained by selecting the indicated block of components of $\underline{V}_n$. If $i > j$ then $\underline{V}_n[i:j]$ is interpreted as being an "empty" vector with no components. Also, for completeness we define $V_{0,n} = 0$ and $V_{n+1,n} = 1$. We wish to determine $\mathrm{Law}[\,\underline{V}_n[i:j] \mid \underline{V}_n[1:(i-1)],\, \underline{V}_n[(j+1):n]\,]$. To this end note that by (2.90) the Lebesgue density of $\underline{V}_n$ is given by

$$f_{\underline{V}_n}(\underline{v}) \;=\; n! \quad , \quad 0 \le v_1 \le v_2 \le ... \le v_n \le 1 \; . \qquad (2.118)$$

Assuming that $\underline{v}$ satisfies the inequalities in (2.118) (for otherwise we don't care how the conditional distribution is defined), then the conditional density of $\underline{V}_n[i:j]$ given $\underline{V}_n[1:(i-1)]$ and $\underline{V}_n[(j+1):n]$ is

$$f_{\underline{V}_n[i:j]|(\underline{V}_n[1:(i-1)],\underline{V}_n[(j+1):n])}\big(\,\underline{v}[i:j] \mid (\underline{v}[1:(i-1)], \underline{v}[(j+1):n])\,\big) \qquad (2.119)$$
$$= \frac{f_{\underline{V}_n}(\underline{v})}{f_{(\underline{V}_n[1:(i-1)],\underline{V}_n[(j+1):n])}(\underline{v}[1:(i-1)], \underline{v}[(j+1):n])}$$

Note that the numerator is constant in the region where density of $\underline{V}_n$ is positive, so as a function of $\underline{v}[i:j]$, the conditional density is constant, i.e. it is a uniform density on the region of $\mathbb{R}^{j-i+1}$ where it is positive. Thus, it is only necessary to determine the region where it is positive, which clearly is

$$v_{i-1} \;\le\; v_i \;\le\; v_{i+1} \;\le\; ... \;\le\; v_j \;\le v_{j+1} \; . \qquad (2.120)$$

Note however that this is the Lebesgue density of the order statistics of $j-i+1$ i.i.d. random variables with the uniform distribution on $[v_{i-1}, v_{j+1}]$ (note that the conventions $v_0 = 0$ and $v_{n+1} = 1$ are in force). Thus,

$$\begin{aligned}\mathrm{Law}[\,\underline{V}_n[i:j]\,|\,\underline{V}_n[1:(i-1)] = v[1:(i-1)]\,\&\,\underline{V}_n[(j+1):n] = v[(j+1):n]\,] \\ = \mathrm{Law}[(v_{j+1} - v_{i-1})\underline{V}_{j-i+1} + v_{i-1}]\ . \end{aligned} \tag{2.121}$$

The latter follows since if U is uniform on $[0,1]$ then $aU+b$ is uniform on $[b, b+a]$, for $a > 0$.

Using (2.121) and the fact that nonuniform r.v.'s can be obtained by a transform of uniform r.v.'s as in (2.110), (2.111), (2.112), and (2.117), one can show that if $\underline{X}$ has i.i.d. components with Lebesgue density f, then denoting the order statistics by $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$, we have for instance Law$[X_{(2)}, ..., X_{(n-1)}\,|\, X_{(1)} = x_{(1)}, X_{(n)} = x_{(n)}\,]$ has a Lebesgue density on $I\!\!R^{n-2}$ given by

$$f(x_{(2)}, ..., x_{(n-1)}|x_{(1)}, x_{(n)}) \;=\; \frac{(n-2)!\prod_{i=2}^{n-1} f(x_{(i)})}{[F(x_{(n)}) - F(x_{(1)})]^{n-2}} \tag{2.122}$$

$$\text{for } x_{(1)} \;\leq\; x_{(2)} \;\leq\; ... \;\leq\; x_{(n-1)} \;\leq\; x_{(n)}\ .$$

Here, F is the common c.d.f. of the components of $\underline{X}$. See Exercise 2.5.10.

In the above discussion of order statistics we have already mentioned that for $\underline{x} \in I\!\!R^n$ there is a permutation π (which depends on $\underline{x}$) such that $\tilde{\pi}\underline{x} = \mathbf{Sort}(\underline{x})$, i.e. π is the permutation of the componont which rearranges them into ascending order. Furthermore, if the components of $\underline{x}$ are distinct (so there are not "ties" in $\mathbf{Sort}(\underline{x})$), then π is unique. For this case, define $\mathbf{Rank}(\underline{x}) = (\pi^{-1}(1), \pi^{-1}(2), ..., \pi^{-1}(n))$, i.e. the i'th component of $\mathbf{Rank}(\underline{x})$ is the index of the component of $\mathbf{Sort}(\underline{x})$ which equals x_i. For simplicity, we shall identify the vector of integers $\mathbf{Rank}(\underline{x})$ with the permutation π^{-1}. The student is asked to prove the following in Exercise 2.5.12.

Proposition 2.5.2 *Let* $\underline{X}$ *be as in Theorem 2.5.1 (b). Then* $\mathbf{Rank}(\underline{X})$ *has a uniform distribution on* $\mathbf{Perm}(n)$, *i.e.* $P[\mathbf{Rank}(\underline{X}) = \pi] = 1/n!$ *for all* $\pi \in \mathbf{Perm}(n)$.

□

Exercises for Section 2.5.

2.5.1 Verify (2.88).

2.5.2 Show that if A is a permutation matrix (i.e. there is a single 1 in each row and column, and 0 elsewhere), then $A^{-1} = A'$.

2.5.3 Verify (2.90) from (2.89).

2.5.4 Suppose $\underline{X}$ is a random n-vector satisfying the following two conditions:

(i) Law$[\underline{X}]$ is *exchangeable*, i.e. Law$[\tilde{\pi}\underline{X}]$ = Law$[\underline{X}]$ for all $\pi \in$ **Perm**;

(ii) $P[X_i = X_j \text{ for any } i \neq j] = 0$.

Show that the conclusion of Theorem 2.5.1 (b) still holds, i.e. that the same formula holds for Law$[\underline{X}|\mathbf{Sort}(\underline{X}) = \underline{y}]$.

2.5.5 Verify (2.107) and (2.108).

2.5.6 For each of the following, obtain (i) the Lebesgue density for the i'th order statistic $X_{(i)}$; (ii) $E[X_{(i)}]$; and (iii) the expected value of the sample range Range$(\underline{X})$. You can save yourself some work if you use results from Examples 2.5.1 and 2.5.2 and methods from Example 2.5.3.

(a) $X_1, X_2, \ldots, X_n$ be i.i.d. $Unif(a, b)$.

(b) $X_1, X_2, \ldots, X_n$ be i.i.d. $Expo(\mu)$.

2.5.7 For each of the settings in Exercise 2.5.6, determine a constant C_n so that C_nRange$(\underline{X})$ is an unbiased estimator of the standard deviation $\sigma = \sqrt{\text{Var}[X_i]}$. Compare with the value of C_n in Example 2.5.3.

2.5.8 Verify (2.110) through (2.117).

2.5.9 Find the Lebesgue density of the sample range in the setting of Example 2.5.3.

2.5.10 Verify (2.122).

2.5.11 Suppose $\underline{X}$ has i.i.d. components with Lebesgue density f, and let $X_{(1)} \leq ... \leq X_{(n)}$ denote the order statistics.

(a) What is the Lebesgue density of the conditional distribution of $X_{(n)}$ given $X_{(1)}, ..., X_{(n-1)}$?

(b) What is the Lebesgue density of the conditional distribution of $X_{(i)}$ given $X_{(1)}, ..., X_{(i-1)}, X_{(i+1)}, ... X_{(n)}$?

(c) Same as (b) but only given $X_{(i-1)}$ and $X_{(i+1)}$?

2.5.12 Prove Proposition 2.5.2.

2.5.13 Assuming $X_1, X_2, \ldots, X_n$ are i.i.d. with Lebesgue density $f(x)$, derive a formula for the Lebesgue density of the sample median. You will need different formulae depending on whether n is even or odd.