Stat 5101 Lecture Notes

Charles J. Geyer

January 16, 2001

Contents

Chapter 1

Random Variables and Change of Variables

1.1 Random Variables

1.1.1 Variables

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

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

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

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

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

1.1.2 Functions

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

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

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

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

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

Constant Functions

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

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

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

Identity Functions

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

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

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

Range and Codomain

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

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

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

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

$$A \xrightarrow{f} B$$

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

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

1.1.3 Random Variables: Informal Intuition

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

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

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

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

1.1.4 Random Variables: Formal Definition

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

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

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

$$S \xrightarrow{X} T$$

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

An Abuse of Notation

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

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

or

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

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

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

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

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

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

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

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

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

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

$$P(X \le 2)$$

or

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

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

The Incredible Disappearing Identity Random Variable

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

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

that is, X is the identity function on S.

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

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

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

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

Another Useful Notation

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

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

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

Examples

Example 1.1.1 (Exponential Random Variable).
Suppose

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

What is

$$P(X > x),$$

for $x > 0$?

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

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

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

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

The event A is

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

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

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

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

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

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

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

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

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

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

Note that in the second line

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

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

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

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

But know we want to know

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

where μ and a are positive real numbers.

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

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

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

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

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

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

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

1.1.5 Functions of Random Variables

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

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

and similarly for any other function of random variables.

1.2 Change of Variables

1.2.1 General Definition

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

Note that

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

1.2.2 Discrete Random Variables

For discrete random variables, probability measures are defined by sums

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

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

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

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

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

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

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

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

where

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

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

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

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

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

One-To-One Transformations

A transformation (change of variable)

$$S \xrightarrow{g} T$$

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

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

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

$$y = g(x).$$

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

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

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

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

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

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

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

Inverse Transformations

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

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

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

by

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

Every invertible function

$$S \xrightarrow{g} T$$

has an *inverse* function

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

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

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

and

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

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

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

The inverse function is *discovered* by trying to solve

$$y = g(x)$$

for x. For example, if

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

then

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

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

Change of Variable for Invertible Transformations

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

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

and the theorem can be stated as follows.

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

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

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

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

Some people like to start counting at one rather than zero (Lindgren among them) and prefer to call the distribution of the random variable $Y = X + 1$ the

"geometric distribution" (there is no standard, some people like one definition, some people like the other).

The transformation in question is quite simple

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

has inverse

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

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

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

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

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

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

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

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

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

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

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

1.2.3 Continuous Random Variables

For continuous random variables, probability measures are defined by integrals

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

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

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

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

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

There is, however, an analog of Theorem 1.3.

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

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

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

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

Example 1.2.5.
Suppose

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

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

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

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

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

The inverse transformation is, of course,

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

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

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

The density of X is

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

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

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

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

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

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

1.3 Random Vectors

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

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

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

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

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

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

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

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

1.3.1 Discrete Random Vectors

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

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

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

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

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

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

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

1.3.2 Continuous Random Vectors

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

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

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

The corresponding *probability measure* is defined by

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The inner integral is

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

So the outer integral is

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

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

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

1.4 The Support of a Random Variable

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

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

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

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

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

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

Simple examples show the idea.

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

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

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

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

and the other using indicator functions

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

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

1.5 Joint and Marginal Distributions

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

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

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

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

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

Theorem 1.1 applied to this case says that

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

where

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

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

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

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

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

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

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

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

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

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

In words we can state this result as follows

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

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

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

like *any* integral

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

Thus

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Of course the density of the uniform distribution is constant

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

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

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

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

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

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

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

1.6 Multivariable Change of Variables

1.6.1 The General and Discrete Cases

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

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

1.6.2 Continuous Random Vectors

Derivatives of Vector Functions

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

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

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

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

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

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

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

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

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

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

is the matrix with elements

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

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

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

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

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

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

This is sometimes written in calculus books as

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

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

Calculating these partial derivatives, we get

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

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

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

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

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

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

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

Invertible Transformations

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

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

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

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

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

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

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

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

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

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

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

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

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

Note that the univariate change-of-variable formula

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

is a special case.

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

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

where

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

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

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

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

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

Now applying Theorem 1.5 to the right hand side gives

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

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

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

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

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

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

Find the joint density of the variables

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

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

The inverse transformation is

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

This transformation has derivative

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

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

Thus the joint density of U and V is

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

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

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

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

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

Proof. Consider the change of variables

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

(this is the mapping **g** in the corollary, which gives the new variables in terms of the old) having inverse mapping

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

(this is the mapping **h** in the corollary, which gives the old variables in terms of the new). The Jacobian is

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

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

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

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

We find the marginal of U by integrating out V

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

which is the convolution formula. □

Noninvertible Transformations

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

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

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

defined by

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

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

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

and

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

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

$$S \xrightarrow{g} T$$

$$S \xleftarrow{h} T$$

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

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

and

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

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

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

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

and

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

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

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

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

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

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

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

where

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

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

have

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

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

Example 1.6.3.
Suppose X is a random variable with density

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

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

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

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

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

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

and

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

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

The sections have derivatives

$$h'_-(y) = -\frac{1}{2} y^{-1/2}$$

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

and applying the theorem gives

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

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

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

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

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

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

Moreover, if f_X is symmetric about zero, then

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

Chapter 2

Expectation

2.1 Introduction

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

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

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

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

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

2.2 The Law of Large Numbers

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

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

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

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

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

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

2.3 Basic Properties

2.3.1 Axioms for Expectation (Part I)

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

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

Expectation is in some respects much a much simpler concept that probability and in other respects a bit more complicated. The issue that makes

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Equation (2.3) is more commonly written

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

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

2.3.2 Derived Basic Properties

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

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

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

Then Axiom E1 implies

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

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

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

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

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

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

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

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

and reading from end to end gives

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

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

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

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

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

The corollary then implies

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

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

Theorem 2.3 (Linearity). *If X_1, ..., X_n are in L^1, and a_1, ..., a_n are real numbers then $a_1X_1 + \cdots a_nX_n$ is also in L^1, and*

$$E(a_1X_1 + \cdots + a_nX_n) = a_1E(X_1) + \cdots + a_nE(X_n).$$

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

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

$$E(a_1X_1 + \cdots + a_{k+1}X_{k+1}) = E(a_1X_1 + \cdots + a_kX_k) + E(a_{k+1}X_{k+1}).$$

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

$$E(a_1X_1 + \cdots + a_{k+1}X_{k+1}) = E(a_1X_1 + \cdots + a_kX_k) + a_{k+1}E(X_{k+1}).$$

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

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

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

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

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

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

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

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

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

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

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

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

$$E(a) = a.$$

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

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

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

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

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

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

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

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

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

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

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

2.3.3 Important Non-Properties

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

The Multiplicativity Non-Property

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

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

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

Taking a Function Outside an Expectation

Suppose g is a linear function defined by

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

where a and b are real numbers. Then

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

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

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

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

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

2.4 Moments

If k is a positive integer, then the real number

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

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

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

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

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

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

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

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

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

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

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

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

2.4.1 First Moments and Means

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

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

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

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

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

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

Symmetric Random Variables

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

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

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

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

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

$$P_X = P_Y$$
$$F_X = F_Y$$
$$f_X = f_Y$$

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

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

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

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

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

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

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

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

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

In notation, the two assertions of the theorem are

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

and

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

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

2.4.2 Second Moments and Variances

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

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

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

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

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

All second moments are related.

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

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

Proof. Using the fact

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

from algebra

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

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

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

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

The name of this theorem is rather strange. It is taken from an analogous theorem in physics about moments of inertia. So the name has nothing to do with probability in general and moments (as understood in probability theory rather than physics) in particular, and the theorem is not commonly called by that name. We will use it because Lindgren does, and perhaps because the theorem doesn't have any other widely used name. In fact, since it is so

simple, it is often not called a theorem but just a calculation formula or method. Sometimes it is called "completing the square" after the method of that name from high-school algebra, although that name isn't very appropriate either. It is a very simple theorem, just the algebraic identity (2.14), which is related to "completing the square" plus linearity of expectation, which isn't. Whatever it is called, the theorem is exceedingly important, and many important facts are derived from it. I sometimes call it "the most important formula in statistics."

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

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

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

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

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

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

and

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

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

The next theorem extends Theorem 2.2 from means to variances.

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

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

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

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

Sanity Check: *Variances are nonnegative.*

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

example, you can get a negative number as a result of the subtraction, if you have calculated one of the quantities being subtracted incorrectly.

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

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

2.4.3 Standard Deviations and Standardization

Standard Deviations

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

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

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

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

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

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

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

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

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

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

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

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

Standardization

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

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

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

has mean zero and standard deviation one.

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

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

has mean μ and standard deviation σ.

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

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

2.4.4 Mixed Moments and Covariances

When several random variables are involved in the discussion, there are several moments of each type, as we have already discussed. If we have two

random variables X and Y, then we also have two (ordinary) first moments μ_X and μ_Y and two second central moments σ_X^2 and σ_Y^2, but that is not the whole story. To see why, it is helpful to make a brief digression into the terminology of polynomials.

Polynomials and Monomials

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

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

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

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

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

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

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

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

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

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

Mixed Moments

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

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

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

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

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

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

just as (2.20) can be written

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

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

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

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

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

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

where

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

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

Covariance

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

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

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

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

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

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

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

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

Variance of a Linear Combination

A very important application of the covariance concept is the second-order analog of the linearity property given in Theorem 2.3. Expressions like the $a_1X_1 + \cdots + a_nX_n$ occurring in Theorem 2.3 arise so frequently that it is worth having a general term for them. An expression $a_1x_1 + \cdots a_nx_n$, where the a_i are constants and the x_i are variables is called a *linear combination* of these variables. The same terminology is used when the variables are random. With this terminology defined, the question of interest in this section can be stated: what can we say about variances and covariances of linear combinations?

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

In words,

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

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

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

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

Proof of Theorem 2.16. First define

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

Then note that by linearity of expectation

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

Then

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

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

2.4.5 Exchangeable Random Variables

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.4.6 Correlation

The Cauchy-Schwarz Inequality

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

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

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

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

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

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

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

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

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

is equal to zero, that is, at

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

the minimum being

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

And this is nonnegative if and only if

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

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

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

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

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

Correlation

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

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

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

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

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

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

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

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

Third, we have the correlation inequality.

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

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

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

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

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

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

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

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

Problems

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

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

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

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

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

2-3. Prove Corollary 2.5.

2-4. Prove Corollary 2.6.

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

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

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

2-6. Prove Corollary 2.7.

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

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

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

2-9. Prove Theorem 2.9.

2-10. Prove Theorem 2.10.

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

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

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

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

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

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

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

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

2-13. Prove Corollary 2.12.

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

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

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

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

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

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

2-17. Prove Theorem 2.15.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

helps.

2-25. Prove (2.33).

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

2.5 Probability Theory as Linear Algebra

This section has two objectives.

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

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

2.5.1 The Vector Space L^1

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

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

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

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

Vector Spaces

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

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

There are two things you can do with vectors.

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

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

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

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

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

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

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

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

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

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

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

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

Linear Transformations and Linear Functionals

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

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

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

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

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

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

2.5.2 Two Notions of Linear Functions

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

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

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

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

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

is an affine function.

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

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

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

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

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

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

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

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

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

2.5.3 Expectation on Finite Sample Spaces

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

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

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

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

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

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

where

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

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

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

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

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

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

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

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

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

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

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

Now we want to make another change of notation. If we want to talk about vectors that are elements of L^1 (and we do), we should use the usual notation,

denoting those elements (which are random variables) by X rather than $\mathbf{x}$ and their components by $X(s)$ giving

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

To summarize the argument of this section so far

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

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

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

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

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

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

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

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

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

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

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

and

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

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

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

A little section about *mathematics is invariant under changes of notation*. We often write (2.40) in different notation. If X is a random variable with density f_X having domain S (the range of possible values of X), then

$$E\{g(X)\} = \sum_{x \in S} g(x) f_X(x). \tag{2.42}$$

Note that (2.42) is exactly the same as (2.40) except for purely notational differences. The special case where g is the identity function

$$E(X) = \sum_{x \in S} x f_X(x) \tag{2.43}$$

is of some interest. Lindgren takes (2.43) as the *definition* of expectation. For us it is a trivial special case of the more general formula (2.42), which in turn is not a definition but a theorem (Theorem 2.27). For us the *definition* of expectation is "an operator satisfying the axioms."

Example 2.5.1 (The Binomial Distribution).
Recall the binomial distribution (Section B.1.2 of Appendix B) having density

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, \ldots, n.$$

We want to calculate $E(X)$. By the formulas in the preceding discussion

$$\begin{aligned}
E(X) &= \sum_{x=0}^{n} x f(x) \\
&= \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} \\
&= \sum_{k=0}^{n} k \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\
&= \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} p^k (1-p)^{n-k} \\
&= np \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} (1-p)^{n-k} \\
&= np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} \\
&= np \sum_{m=0}^{n-1} \binom{n-1}{m} p^m (1-p)^{n-1-m}
\end{aligned}$$

- Going from line 1 to line 2 we just plugged in the definition of $f(x)$ and changed the dummy variable of summation from x to k.

- Going from line 2 to line 3 we just plugged in the definition of the binomial coefficient.
- Going from line 3 to line 4 we just observed that the $k = 0$ term is zero and then canceled the k in the numerator with the k in the $k!$ in the denominator.
- Going from line 4 to line 5 we pulled an n out of the $n!$ and a p out of the p^k.
- Going from line 5 to line 6 we just used the definition of the binomial coefficient again.
- Going from line 6 to line 7 we changed the dummy variable of summation to $m = k - 1$.

Now the binomial theorem says the sum in the last line is equal to one. Alternatively, the sum in the last line is equal to one because the summand is the $\text{Bin}(n - 1, p)$ density, and *every* probability density sums to one. Hence

$$E(X) = np.$$

2.5.4 Axioms for Expectation (Part II)

Absolute Values

Axiom E5 (Absolute Values). *If X is in L^1, then so is $|X|$.*

Note that this axiom trivially applies to the probability models on a finite sample space discussed in the preceding section, because *every* real-valued function is in L^1. This axiom is only interesting when the sample space is infinite.

With this axiom, we can prove another basic property of expectation that is mostly used in theoretical arguments.

Theorem 2.28 (Absolute Values). *If X is in L^1, then*

$$|E(X)| \le E(|X|).$$

The name of this theorem is "taking an absolute value inside an expectation can only increase it." That's a long-winded name, but there is no widely used short name for the theorem.

Derivation of Property 2.28. First note that $X \le |X|$. Applying Property 2.8 to these two random variables gives

$$E(X) \le E(|X|),$$

which is what was to be proved in the case that $E(X)$ is nonnegative.

To prove the other case, we start with the fact that $-X \le |X|$. Another application of Property 2.8 along with Property 2.6 gives

$$-E(X) = E(-X) \le E(|X|).$$

But when $E(X)$ is negative $-E(X) = |E(X)|$, so that proves the other case. □

Note that there is no explicit mention of Axiom E5 in the proof. The *implicit* mention is that only the axiom allows us to talk about $E(|X|)$. None of the other axioms guarantee that $|X|$ has expectation.

Monotone Convergence

The last axiom for expectation analogous to the countable additivity axiom for probability (called Axiom 3a on p. 30 in Lindgren). This is the monotone convergence axiom. To understand it we need a preliminary definition. For a sequence of numbers $\{x_n\}$, the notation $x_n \uparrow x$ means $x_1 \le x_2 \le \ldots$ and $x_n \to x$. For a sequence of random variables $\{X_n\}$ on a sample space S, the notation $X_n \uparrow X$ means $X_n(s) \uparrow X(s)$ for all $s \in S$.

Axiom E6 (Monotone Convergence). *Suppose X_1, X_2, ... is a sequence of random variables in L^1 such that $X_n \uparrow X$. If*

$$\lim_{n\to\infty} E(X_n) < \infty,$$

then $X \in L^1$ and

$$E(X_n) \uparrow E(X).$$

Conversely, if

$$\lim_{n\to\infty} E(X_n) = \infty,$$

then $X \notin L^1$.

The monotone convergence axiom is a fairly difficult subject, so difficult that Lindgren omits it entirely from his book, although this makes no sense because the countable additivity axiom for probability is equally difficult and is included. So this is really more treating expectation is a second class concept, subsidiary to probability. Our insistence on including it is part and parcel of our notion that probability and expectation are equally important and deserve equal treatment.

That having been said, this axiom can be considered the dividing line between material at the level of this course and material over our heads. If a proof involves monotone convergence, it is too hard for us. We will state some results that can only be proved using the monotone convergence axiom, but we will leave the proofs for more advanced courses.

There is a "down arrow" concept defined in obvious analogy to the "up arrow" concept (the sequence converges down rather than up), and there is an analogous form of monotone convergence

Corollary 2.29 (Monotone Convergence). *Suppose X_1, X_2, ... is a sequence of random variables in L^1 such that $X_n \downarrow X$. If*

$$\lim_{n\to\infty} E(X_n) > -\infty,$$

then $X \in L^1$ and

$$E(X_n) \downarrow E(X).$$

Conversely, if

$$\lim_{n\to\infty} E(X_n) = -\infty,$$

then $X \notin L^1$.

2.5.5 General Discrete Probability Models

If the sample space S of a probability model is countably infinite, we would like to use the same formulas (2.40), (2.41a) and (2.41b), that we used for finite sample spaces, but we run into problems related to infinite series. The sum may not exist (the series may not converge), and if it does exist, its value may depend on the particular enumeration of the sample space that is used. Specifically, there are many ways to enumerate the sample space, writing it as a sequence $S = \{s_1, s_2, \ldots\}$, and when we write out the infinite sum explicitly as

$$E(X) = \sum_{i=1}^{\infty} X(s_i)p(s_i) = \lim_{n\to\infty} \sum_{i=1}^{n} X(s_i)p(s_i)$$

the limit may depend on the particular enumeration chosen. The axioms of expectation, however, solve both of these problems.

First, not all random variables have expectation, only those in L^1 so the fact that expectation may not be defined for some random variables should not bother us. For discrete probability models on a sample space S defined by a probability density p, we define L^1 to be the set of all functions $X : S \to \mathbb{R}$ satisfying

$$\sum_{s\in S} |X(s)|p(s) < \infty. \tag{2.44}$$

This definition trivially satisfies Axiom E5 and also satisfies the existence parts of Axioms E1, E2, and E4.

For $X \in L^1$ we define expectation by the same formula (2.40) as in the finite sample space case. Note that then the sum in (2.44) is $E(|X|)$. Thus our definition says that X has expectation if and only if $|X|$ also has expectation. Another way to say the same thing is that (2.40) defines an expectation if and only if the series is *absolutely summable*, which means the sum of the absolute values of the terms of the series exists.

Because of the rearrangement of series theorem from calculus, which says that if a series is absolutely summable then the sum of the series does not depend on the order in which the terms are summed, we can rearrange the terms in the sum as we please without changing the result. That is why we can write (2.40) as an *unordered* sum using notation that does not specify any particular ordering.

Theorem 2.30. *All probability models on a countable sample space S are defined by a function function $p : S \to \mathbb{R}$ satisfying*

$$p(s) \geq 0, \qquad s \in S, \tag{2.45a}$$

and

$$\sum_{s \in S} p(s) = 1. \tag{2.45b}$$

The corresponding expectation operator is $E : L^1 \to \mathbb{R}$*, where* L^1 *is the set of functions* $X : S \to \mathbb{R}$ *such that*

$$\sum_{s \in S} p(s)|X(s)| < \infty,$$

and

$$E(X) = \sum_{s \in S} p(s)X(s) \tag{2.46}$$

Following our policy that any proof that involves dominated convergence is beyond the scope of this course, we won't try to prove the theorem.

Note that the remarks about *mathematics is invariant under changes of notation* in the preceding section apply here too. In particular, (2.42) and (2.43) apply just as well in the case that S is countably infinite (so long as the expectation in question exists).

Example 2.5.2 (The Poisson Distribution).
The the Poisson distribution is the discrete distribution having density

$$f(x) = \frac{\mu^x}{x!} e^{-\mu}, \qquad x = 0, 1, \ldots.$$

(Section B.1.4 of Appendix B). If $X \sim \text{Poi}(\mu)$, then

$$\begin{aligned} E(X) &= \sum_{x=0}^{\infty} x f(x) \\ &= \sum_{k=0}^{\infty} k \frac{\mu^k}{k!} e^{-\mu} \\ &= \mu \sum_{k=1}^{\infty} \frac{\mu^(k-1)}{(k-1)!} e^{-\mu} \\ &= \mu \sum_{m=0}^{\infty} \frac{\mu^m}{m!} e^{-\mu} \end{aligned}$$

- Going from line 1 to line 2 we just plugged in the definition of $f(x)$ and changed the dummy variable of summation from x to k.
- Going from line 2 to line 3 we just observed that the $k = 0$ term is zero, then canceled the k in the numerator with the k in the $k!$ in the denominator, and then pulled a μ out of the μ^k.
- Going from line 3 to line 4 we changed the dummy variable of summation to $m = k - 1$.

The sum in the last line is equal to one because the summand is the Poi(μ) density, and *every* probability density sums to one. Hence

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

2.5.6 Continuous Probability Models

When the sample space is uncountable, like $\mathbb{R}$ or $\mathbb{R}^d$ we cannot use the formulas of Theorem 2.27 to define expectation. There is no notion of sums with an uncountably infinite number of terms.

There is, however, another concept that behaves much like summation, which is integration. We just replace the sums by integrals.

Theorem 2.31. *Probability models on having a subset S of $\mathbb{R}$ or $\mathbb{R}^d$ can be defined by a function function $f : S \to \mathbb{R}$ satisfying*

$$f(x) \ge 0, \qquad x \in S, \tag{2.47a}$$

and

$$\int_S f(x)\, dx = 1. \tag{2.47b}$$

The space L^1 of random variables having expectations is the set of real-valued functions $g : S \to \mathbb{R}$ such that

$$\int_S |g(x)| f(x)\, dx < \infty.$$

The corresponding expectation operator is $E : L^1 \to \mathbb{R}$ is defined by

$$E\{g(X)\} = \int_S g(x) f(x)\, dx. \tag{2.48}$$

As in the discrete case, we define expectation so that Y has expectation only if $|Y|$ also has expectation. Since we are using integrals rather than sums, we are now interested in absolute integrability rather than absolute summability, but there is a complete analogy between the two cases.

Similar formulas hold when the sample space is $\mathbb{R}^d$ or a subset S of $\mathbb{R}^d$. The general formula, written in vector notation and ordinary multiple-integral notation is

$$\begin{aligned} E\{g(\mathbf{X})\} &= \int_S g(\mathbf{x}) f(\mathbf{x})\, d\mathbf{x} \\ &= \idotsint_S g(x_1, x_2, \ldots, x_n) f(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n \end{aligned} \tag{2.49}$$

Now we take a time out for a comment that is "beyond the scope of this course." We just lied to you, sort of. Theorem 2.31 is not true if the integral signs indicate the kind of integral (the so-called *Riemann integral*) described in

elementary calculus courses. All the axioms except monotone convergence are satisfied, and monotone convergence

$$\lim_{n\to\infty} \int g_n(x) f(x)\, dx = \int g(x) f(x)\, dx, \qquad \text{if } g_n \uparrow g. \tag{2.50}$$

holds sometimes but not always.

The problem is that the limit of a sequence of Riemann integrable functions is not necessarily Riemann integrable. So even though (2.50) is true whenever all the functions involved are Riemann integrable, that isn't enough to satisfy the monotone convergence axiom. The way around this problem is a *tour de force* of higher mathematics. One just makes (2.50) hold *by definition*. First one shows that for two sequences $g_n \uparrow g$ and $h_n \uparrow g$ increasing to the same limit

$$\lim_{n\to\infty} \int g_n(x) f(x)\, dx = \lim_{n\to\infty} \int h_n(x) f(x)\, dx \tag{2.51}$$

Therefore if we just *define* the right hand side of (2.50) to be the left hand side, the equation is then true by definition. This definition is unambiguous because the value of the limit does not depend on the sequence chosen (2.51). This "extension by monotone convergence" of the definition of the integral is called the *Lebesgue integral* .

Note that the Riemann integral always agrees with the Lebesgue integral whenever both are defined, so this is not a totally new concept. Every function you already know how to integrate has the same integral in both senses. The only point of Lebesgue integration is that it allows the integration of some *really weird* functions, too weird to have Riemann integrals. Since no really weird functions are of any practical interest, the only point of the whole exercise is to prove theorems using the monotone convergence axiom. And since that is beyond the scope of this course, we won't worry about it.

Example 2.5.3 (The Gamma Distribution).
The the Gamma distribution is the continuous distribution having density

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \qquad x > 0$$

(Section B.2.3 of Appendix B). If $X \sim \text{Gam}(\alpha, \lambda)$, then

$$\begin{aligned} E(X) &= \int_0^\infty x f(x)\, dx \\ &= \int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} x^\alpha e^{-\lambda x}\, dx \\ &= \frac{\Gamma(\alpha+1)}{\lambda\Gamma(\alpha)} \int_0^\infty \frac{\lambda^{\alpha+1}}{\Gamma(\alpha+1)} x^\alpha e^{-\lambda x}\, dx \end{aligned}$$

- Going from line 1 to line 2 we just plugged in the definition of $f(x)$ and collected the x and $x^{\alpha-1}$ terms together.

- Going from line 2 to line 3 we just pulled some constants outside of the integral.

The integral in the last line is equal to one because the integrand is the density of the $\text{Gam}(\alpha + 1, \lambda)$ distribution, and *every* probability density integrates to one. Hence

$$E(X) = \frac{\Gamma(\alpha + 1)}{\lambda \Gamma(\alpha)} = \frac{\alpha}{\lambda}$$

the second equality being the recurrence relation for the gamma function (B.3) in Section B.3.1 of Appendix B.

2.5.7 The Trick of Recognizing a Probability Density

The astute reader may have recognized a pattern to Examples 2.5.1, 2.5.2, and 2.5.3. In each case the sum or integral was done by recognizing that by moving certain constants (terms not containing the variable of summation or integration) outside of the sum or integral leaving only the sum or integral of a *known probability density*, which is equal to one by definition.

Of course, you don't have to use the trick. There is more than one way to do it. In fact, we even mentioned that you could instead say that we used the binomial theorem to do the sum in Example 2.5.1. Similarly, you could say we use the Maclaurin series for the exponential function

$$e^x = 1 + x + \frac{x^2}{2} + \cdots + \frac{x^k}{k!} + \cdots$$

to do the sum in Example 2.5.2, and you could say we use the definition of the gamma function, (B.2) in Appendix B plus the change-of-variable formula to do the integral in Example 2.5.3. In fact, the argument we gave using the fact that densities sum or integrate to one as the case may be does use these *indirectly*, because those are the reasons why these densities sum or integrate to one.

The point we are making here is that in every problem involving an expectation in which you are doing a sum or integral, you already have a know sum or integral to work with. This is expecially important when there is a whole *parametric family* of densities to work with. In calculating the mean of a $\Gamma(\alpha, \lambda)$ distribution, we used the fact that a $\Gamma(\alpha + 1, \lambda)$ density, like all densities, integrates to one. This is a very common trick. One former student said that if you can't do an integral using this trick, then you can't do it at all, which is not quite true, but close. Most integrals and sums you will do to calculate expectations can be done using this trick.

2.5.8 Probability Zero

Events of probability zero are rather a nuisance, but they cannot be avoided in continuous probability models. First note that every outcome is an event of probability zero in a continuous probability model, because by definition

$$P(X = a) = \int_a^a f(x)\, dx,$$

and a definite integral over an interval of length zero is zero.

Often when we want to assert a fact, it turns out that the best we can get from probability is an assertion "with probability one" or "except for an event of probability zero." The most important of these is the following theorem, which is essentially the same as Theorem 5 of Chapter 4 in Lindgren.

Theorem 2.32. *If $Y = 0$ with probability one, then $E(Y) = 0$. Conversely, if $Y \geq 0$ and $E(Y) = 0$, then $Y = 0$ with probability one.*

The phrase "$Y = 0$ with probability one" means $P(Y = 0) = 1$. The proof of the theorem involves dominated convergence and is beyond the scope of this course.

Applying linearity of expectation to the first half of the theorem, we get an obvious corollary.

Corollary 2.33. *If $X = Y$ with probability one, then $E(X) = E(Y)$.*

If $X = Y$ with probability one, then the set

$$A = \{\, s : X(s) \neq Y(s) \,\}$$

has probability zero. Thus a colloquial way to rephrase the corollary is "what happens on a set of probability zero doesn't matter." Another rephrasing is "a random variable can be arbitrarily redefined on a set of probability zero without changing any expectations."

There are two more corollaries of this theorem that are important in statistics.

Corollary 2.34. $\operatorname{var}(X) = 0$ *if and only if X is constant with probability one.*

Proof. First, suppose $X = a$ with probability one. Then $E(X) = a = \mu$, and $(X - \mu)^2$ equals zero with probability one, hence by Theorem 2.32 its expectation, which is $\operatorname{var}(X)$, is zero.

Conversely, by the second part of Theorem 2.32, $\operatorname{var}(X) = E\{(X-\mu)^2\} = 0$ implies $(X-\mu)^2 = 0$ with probability one because $(X-\mu)^2$ is a random variable that is nonnegative and integrates to zero. Since $(X - \mu)^2$ is zero only when $X = \mu$, this implies $X = \mu$ with probability one. □

Corollary 2.35. $|\operatorname{cor}(X, Y)| = 1$ *if and only if there exist constants α and β such that $Y = \alpha + \beta X$ with probability one.*

Proof. First suppose $Y = \alpha + \beta X$ with probability one. Then by (2.33)

$$\operatorname{cor}(\alpha + \beta X, X) = \operatorname{sign}(\beta) \operatorname{cor}(X, X) = \pm 1.$$

That proves one direction of the "if and only if."

To prove the other direction, we assume $\rho_{X,Y} = \pm 1$ and have to prove that $Y = \alpha + \beta X$ with probability one, where α and β are constants we may choose. I claim that the appropriate choices are

$$\beta = \rho_{X,Y} \frac{\sigma_Y}{\sigma_X}$$
$$\alpha = \mu_Y - \beta \mu_X$$

(these are just pulled out of the air here, the choice will make sense after we have done best linear prediction).

We want to prove that $Y = \alpha + \beta X$ with probability one. We can do this by showing that $(Y - \alpha - \beta X)^2$ is zero with probability one, and this will follow from Theorem 2.32 if we can show that $(Y - \alpha - \beta X)^2$ has expectation zero. Hence let us calculate

$$\begin{aligned}
E\left\{(Y - \alpha - \beta X)^2\right\} &= E\left\{\left(Y - \mu_Y - \rho_{X,Y}\frac{\sigma_Y}{\sigma_X}(X - \mu_X)\right)^2\right\} \\
&= E\left\{(Y - \mu_Y)^2\right\} \\
&\qquad - 2E\left\{(Y - \mu_Y)\left(\rho_{X,Y}\frac{\sigma_Y}{\sigma_X}(X - \mu_X)\right)\right\} \\
&\qquad + E\left\{\left(\rho_{X,Y}\frac{\sigma_Y}{\sigma_X}(X - \mu_X)\right)^2\right\} \\
&= \operatorname{var}(Y) - 2\rho_{X,Y}\frac{\sigma_Y}{\sigma_X}\operatorname{cov}(X, Y) + \rho_{X,Y}^2\frac{\sigma_Y^2}{\sigma_X^2}\operatorname{var}(X) \\
&= \sigma_Y^2 - 2\rho_{X,Y}^2\sigma_Y^2 + \rho_{X,Y}^2\sigma_Y^2 \\
&= \sigma_Y^2(1 - \rho_{X,Y}^2)
\end{aligned}$$

which equals zero because of the assumption $|\rho_{X,Y}| = 1$. □

2.5.9 How to Tell When Expectations Exist

We say a random variable Y *dominates* a random variable X if $|X| \le |Y|$.

Theorem 2.36. *If Y dominates X and Y has expectation, then X also has expectation. Conversely if Y dominates X and the expectation of X does not exist, then the expectation of Y does not exist either.*

The proof involves monotone convergence and is hence beyond the scope of this this course.[1]

We say a random variable X is *bounded* if $|X| \le a$ for some constant a.

[1] Actually this theorem is way, way beyond the scope of this course, the one subject we will touch on that is really, really, really weird. Whether this theorem is true or false is a matter of taste. Its truth depends on an axiom of set theory (the so-called axiom of choice), which can be assumed or not without affecting anything of practical importance. If the theorem is false, that means there exists a random variable X dominated by another random variable Y such that Y is in L^1 and X isn't. However, the usual assumptions of advanced probability theory imply that every Riemann integrable random variable dominated by Y is in L^1, hence X cannot be written as the limit of a sequence $X_n \uparrow X$ for a sequence of Riemann integrable random variables X_n. This means that X is weird indeed. Any conceivable description of X (which like any random variable is a function on the sample space) would have not only infinite length but uncountably infinite length. That's weird! What is not widely known, even among experts, is that there is no need to assume such weird functions actually exist. The entirety of advanced probability theory can be carried through under the assumption that Theorem 2.36 is true (R. M. Solovay, "A Model of Set-Theory in Which Every Set of Reals is Lebesgue Measurable," *Annals of Mathematics*, 92:1-56, 1970).

Corollary 2.37. *Every bounded random variable is in L^1.*

Corollary 2.38. *In a probability model with a finite sample space, every random variable is in L^1.*

The corollaries take care of the trivial cases. Thus the question of existence or non-existence of expectations only applies to unbounded random variables in probability models on infinite sample spaces. Then Theorem 2.36 is used to determine whether expectations exist. An expectation is an infinite sum in the discrete case or an integral in the continuous case. The question is whether the integral or sum converges absolutely. That is, if we are interested in the expectation of the random variable $Y = g(X)$ where X has density f, we need to test the integral

$$E(|Y|) = \int |g(x)| f(x)\, dx$$

for finiteness in the continuous case, and we need to test the corresponding sum

$$E(|Y|) = \sum_{x \in S} |g(x)| f(x)$$

for finiteness in the discrete case. The fact that the integrand or summand has the particular product form $|g(x)|f(x)$ is irrelevant. What we need to know here are the rules for determining when an integral or infinite sum is finite.

We will cover the rules for integrals first. The rules for sums are very analogous. Since we are only interested in nonnegative integrands, we can always treat the integral as representing "area under the curve" where the curve in question is the graph of the integrand. Any part of the region under the curve that fits in a finite rectangle is, of course, finite. So the only way the area under the curve can be infinite is if part of the region does not fit in a finite rectangle: either the integrand has a singularity (a point where it goes to infinity), or the domain of integration is an unbounded interval. It helps if we focus on each problem separately: we test whether integrals over neighborhoods of singularities are finite, and we test whether integrals over unbounded intervals are finite. Integrals over bounded intervals not containing singularities do not need to be checked at all.

For example, suppose we want to test whether

$$\int_0^\infty h(x)\, dx$$

is finite, and suppose that the only singularity of h is at zero. For any numbers a and b such that $0 < a < b < \infty$ we can divide up this integral as

$$\int_0^\infty h(x)\, dx = \int_0^a h(x)\, dx + \int_a^b h(x)\, dx + \int_b^\infty h(x)\, dx$$

The first integral on the right hand side may be infinite because of the singularity. The third integral on the right hand side may be infinite because of the

unbounded domain of integration. The second integral on the right hand side must be finite: the integral of a bounded function over a bounded domain is always finite, we do not need to check.

It is rare that we can exactly evaluate the integrals. Usually we have to use Theorem 2.36 to settle the existence question by comparing with a simpler integral. The following lemmas give the most useful integrals for such comparisons. While we are at it, we give the analogous useful infinite sums. The proofs are all elementary calculus.

Lemma 2.39. *For any positive real number a or any positive integer m*

$$\int_a^\infty x^b \, dx \qquad \text{and} \qquad \sum_{n=m}^\infty n^b$$

exist if and only if $b < -1$.

Lemma 2.40. *For any positive real number a*

$$\int_0^a x^b \, dx$$

exists if and only if $b > -1$.

Lemma 2.41. *For any positive real number a or any positive integer m and any positive real number c and any real number b (positive or negative)*

$$\int_a^\infty x^b e^{-cx} \, dx \qquad \text{and} \qquad \sum_{n=m}^\infty n^b e^{-cn}$$

exist.

The following two lemmas give us more help using the domination theorem.

Lemma 2.42. *Suppose g and h are bounded, strictly positive functions on an interval $[a, \infty)$ and*

$$\lim_{x \to \infty} \frac{g(x)}{h(x)} = k, \tag{2.52}$$

where k is a strictly positive constant, then either both of the integrals

$$\int_a^\infty g(x) \, dx \qquad \text{and} \qquad \int_a^\infty h(x) \, dx \tag{2.53}$$

are finite, or neither is. Similarly, either both of the sums

$$\sum_{k=m}^\infty g(k) \qquad \text{and} \qquad \sum_{k=m}^\infty h(k) \tag{2.54}$$

are finite, or neither is, where m is any integer greater than a.

Example 2.5.4 (Exponentially Decreasing Tails).
The following densities

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad -\infty < x < \infty$$

and

$$f(x) = \frac{1}{2} e^{-|x|}, \qquad -\infty < x < \infty$$

have moments of all orders, that is, $E(|X|^p)$ exists for all $p > 0$.

Why? Because the densities are bounded (no singularities) and have exponentially decreasing tails, so Lemma 2.41 assures us that all moments exist.

Example 2.5.5 (Polynomially Decreasing Tails).
The following densities

$$f(x) = \frac{1}{\pi(1 + x^2)}, \qquad -\infty < x < \infty$$

and

$$f(x) = \frac{6}{\pi^2 x^2}, \qquad x = 1, 2, \ldots$$

do not have moments of all orders. In fact, for both $E(|X|^p)$ exists for $p > 0$ if and only if $p < 1$. Thus for these two distributions, neither the mean, nor the variance, nor any higher moment exists.

Why? In both cases, if we look at the integrand or summand $|x|^p f(x)$ in the integral or sum we need to check, we see that it behaves like $|x|^{p-2}$ at infinity. (More formally, the limit of the integrand or summand divided by $|x|^{p-2}$ converges to a constant as x goes to plus or minus infinity. Hence by Lemma 2.42, the expectation exists if and only if the integral or sum of $|x|^{p-2}$ exists.) By Lemma 2.39 the integral or sum exists if and only if $p - 2 < -1$, that is, $p < 1$.

To do problems involving singularities, we need another lemma analogous to Lemma 2.42. This lemma involves only integrals not sums because sequences cannot go to infinity except at infinity (all the terms are actually finite).

Lemma 2.43. *Suppose g and h are strictly positive functions on an interval (a, b) and both have singularities at a but are bounded elsewhere, and suppose*

$$\lim_{x \to a} \frac{g(x)}{h(x)} = k,$$

where k is a strictly positive constant, then either both of the integrals

$$\int_a^b g(x)\, dx \qquad \text{and} \qquad \int_a^b h(x)\, dx$$

are finite, or neither is.

Example 2.5.6 (The Gamma Distribution Again).
The the Gamma distribution is the continuous distribution having density

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \qquad x > 0$$

(Section B.2.3 of Appendix B). For $X \sim \text{Gam}(\alpha, \lambda)$, we consider here when X^p is in L^1 for any real number p, positive or negative. The integral that defines the expectation is

$$E(X^p) = \int_0^\infty x^p \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \, dx = \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha+p-1} e^{-\lambda x} \, dx$$

if the integral exists (which is the question we are examining).

From Lemma 2.41, the integral over (a, ∞) exists for for any $a > 0$ and any p positive or negative. The only issue is the possible singularity of the integrand at the origin. There is a singularity if $\alpha + p - 1 < 0$. Otherwise the integrand is bounded and the expectation exists.

Since $e^0 = 1$, the integrand behaves like $x^{\alpha+p-1}$ at zero and according to Lemma 2.43 this is integrable over a neighborhood of zero if and only if $\alpha+p-1 > -1$, that is, if and only if $p > -\alpha$.

2.5.10 L^p Spaces

We start with another consequence of the domination theorem and the methods for telling when expectations exist developed in the preceding section.

Theorem 2.44. *If X is a real-valued random variable and $|X - a|^p$ is in L^1 for some constant a and some $p \geq 1$, then*

$$|X - b|^q \in L^1,$$

for any constants b and any q such that $1 \leq q \leq p$.

Proof. First the case $q = p$. The ratio of the integrands defining the expectations of $|X - a|^p$ and $|X - b|^p$ converges, that is

$$\frac{|x - b|^p f(x)}{|x - a|^p f(x)} = \left| \frac{x - b}{x - a} \right|^p$$

goes to 1 as x goes to plus or minus infinity. Thus both integrals exist, and $|X - b|^p \in L^1$.

In the case $q < p$, the ratio of integrands

$$\frac{|x - b|^q f(x)}{|x - a|^p f(x)} = \frac{|x - b|^q}{|x - a|^p}$$

converges to zero as x goes to plus or minus infinity. Again this implies both integrals exist and $|X - b|^p \in L^1$. □

Definition 2.5.1 (L^p Spaces).
For any $p \geq 1$, the set of random variables X such that $|X|^p \in L^1$ is called L^p.

With this definition, we can rephrase the theorem. The condition of the theorem can now be stated concisely as $X \in L^p$, because if $|X - a|^p \in L^1$, then the theorem implies $|X|^p \in L^1$ too, which is the same as $X \in L^p$. The conclusion of the theorem can also be restated as $X \in L^q$. Hence $L^1 \supset L^q \supset L^p$ when $1 \leq q \leq p$.

The reason for the name "L^p space" is the following theorem, which we will not prove.

Theorem 2.45. *Each L^p is a vector space.*

What the theorem says is that L^p is closed under addition and scalar multiplication, that is,

$$X \in L^p \text{ and } Y \in L^p \text{ implies } X + Y \in L^p$$

and

$$X \in L^p \text{ and } a \in \mathbb{R} \text{ implies } aX \in L^p.$$

All of this having been said, I have to admit that the main use of the L^p concept at this level is purely as a shorthand. L^2 is the set of random variables having variances. By Theorem 2.44 and the following comment $L^1 \supset L^2$ so these random variables also have means. Thus we could have stated the condition "X is a random variable having first and second moments" in Corollary 2.12 and succeeding theorems about second moments much more concisely as "$X \in L^2$." Whether you like the shorthand or not is a matter of taste. One thing, though, that we did learn in this section is that the words "first and" could have been deleted from the condition of Corollary 2.12 and theorems with similar conditions. If second moments exist, then so do first moments by Theorem 2.44.

2.6 Probability is a Special Case of Expectation

A special kind of random variable is the *indicator function* (or indicator random variable) of an event A (a random variable is a function on the sample space, so an indicator function is a random variable). This is denoted I_A and defined by

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

The indicator function characterizes the set A. It is the set of points ω such that $I_A(\omega) = 1$. More importantly from our point of view, indicator functions connect probability and expectation. The relation

$$P(A) = E(I_A) \tag{2.55}$$

holds for all events A. Probability is just expectation of indicator functions. Thus probability is a dispensable concept. It is just a special case of expectation.

The proof of (2.55) for discrete probability models is trivial.

$$\begin{aligned} E(I_A) &= \sum_{\omega \in \Omega} I_A(\omega) p(\omega) \\ &= \sum_{\omega \in A} p(\omega) \\ &= P(A) \end{aligned}$$

The first equality is the definition (2.40), the third is the definition of probability (p. 30 in Lindgren), and the middle equality just uses the definition of indicator functions: terms for $\omega \in A$ have $I_A(\omega) = 1$ and terms for $\omega \notin A$ have $I_A(\omega) = 0$ and can be dropped from the sum. The proof of (2.55) for continuous probability models is the same except that we replace sums by integrals.

All of the probability axioms can be derived from the expectation axioms by just taking the special case when the random variables are indicator functions. Since indicator functions are nonnegative, Axiom E1 implies

$$E(I_A) = P(A) \ge 0$$

which is the first probability axiom. Axiom E2 implies

$$E(1) = E(I_\Omega) = P(\Omega) = 1$$

which is the second probability axiom. The sum of indicator functions is not necessarily an indicator function, in fact

$$I_A + I_B = I_{A \cup B} + I_{A \cap B}. \tag{2.56}$$

This is easily verified by checking the four possible cases, ω in or not in A and in or not in B. Applying Axiom E4 to both sides of (2.56) gives

$$\begin{aligned} P(A) + P(B) &= E(I_A) + E(I_B) \\ &= E(I_{A \cup B}) + E(I_{A \cap B}) \\ &= P(A \cup B) + P(A \cap B) \end{aligned}$$

which is the general addition rule for probabilities and implies the third probability axiom, which is the special case $A \cap B = \varnothing$.

The countable additivity axiom is applied by the monotone convergence. A nondecreasing sequence of indicator functions corresponds to a nondecreasing sequence of sets. Hence Axiom E5 implies

$$P(A_n) \uparrow P(A), \qquad \text{whenever } A_n \uparrow A$$

This statement, continuity of probability, implies countable additivity (just run the proof on p. 29 in Lindgren backwards).

2.7 Independence

2.7.1 Two Definitions

Lindgren (p. 79, equation (3)) gives the following as a definition of independent random variables.

Definition 2.7.1 (Independent Random Variables).
Random variables X and Y are independent *if*

$$P(X \in A \text{ and } Y \in B) = P(X \in A)P(Y \in B). \tag{2.57}$$

for every event A in the range of X and B in the range of Y.

We take a quite different statement as the definition.

Definition 2.7.2 (Independent Random Variables).
Random variables X and Y are independent *if*

$$E\{g(X)h(Y)\} = E\{g(X)\}E\{h(Y)\} \tag{2.58}$$

for all real-valued functions g and h such that these expectations exist.

These two definitions are equivalent—meaning they define the same concept. That means that we could take either statement as the definition and prove the other. Lindgren takes (2.57) as the definition and "proves" (2.58). This is Theorem 11 of Chapter 4 in Lindgren. But the "proof" contains a lot of hand waving. A correct proof is beyond the scope of this course.

That's one reason why we take Definition 2.7.2 as the definition of the concept. Then Definition 2.7.1 describes the trivial special case of Definition 2.7.2 in which the functions in question are indicator functions, that is, (2.57) says exactly the same thing as

$$E\{I_A(X)I_B(Y)\} = E\{I_A(X)\}E\{I_B(Y)\}. \tag{2.59}$$

only in different notation. Thus if we take Definition 2.7.2 as the definition, we easily (trivially) prove (2.57). But the other way around, the proof is beyond the scope of this course.

2.7.2 The Factorization Criterion

Theorem 2.46 (Factorization Criterion). *A finite set of real-valued random variables is independent if and only if their joint distribution is the product of the marginals.*

What this says is that X_1, ..., X_n are independent if and only if

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i) \tag{2.60}$$

One direction of the theorem is easy to establish. If (2.60) holds

$$\begin{aligned} E\left\{\prod_{i=1}^n g_i(X_i)\right\} &= \int \cdots \int \left(\prod_{i=1}^n g_i(x_i) f_{X_i}(x_i)\right) dx_1 \cdots dx_n \\ &= \prod_{i=1}^n \int g_i(x_i) f_{X_i}(x_i) dx_i \\ &= \prod_{i=1}^n E\left\{g_i(X_i)\right\} \end{aligned}$$

So the X_i are independent. The proof of the other direction of the theorem is beyond the scope of this course.

The simple statement of Theorem 2.46 assumes the marginal densities are defined on the whole real line If necessary, they are extended by zero off the supports of the variables.

> *It is not enough to look only at the formulas defining the densities. You must also look at the domains of definition.*

The following example shows why.

Example 2.7.1 (A Cautionary Example).
The random variables X and Y having joint density

$$f(x, y) = 4xy, \qquad 0 < x < 1 \text{ and } 0 < y < 1 \tag{2.61}$$

are independent, but the random variables X and Y having joint density

$$f(x, y) = 8xy, \qquad 0 < x < y < 1 \tag{2.62}$$

are not! For more on this, see Problem 2-35.

The difference is easy to miss. The formulas defining the densities are very similar, both factor as a function of x times a function of y. The difference is in the domains of definition. The one for which the factorization criterion holds is a rectangle with sides parallel to the axes. The other isn't.

2.7.3 Independence and Correlation

Theorem 2.47. *Independent random variables are uncorrelated.*

The converse is false!

Example 2.7.2.
Suppose X is a nonconstant random variable having a distribution symmetric about zero, and suppose $Y = X^2$ is also nonconstant. For example, we could take $X \sim \mathcal{U}(-1, 1)$, but the details of the distribution do not matter, only that it is symmetric about zero and nonconstant and that X^2 also has a nonconstant distribution.

Then X and Y are uncorrelated (Problem 2-37) but not independent. Independence would require that

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

hold for *all* functions g and h. But it obviously does not hold when, to pick just one example, g is the squaring function and h is the identity function so $g(X) = Y$ and $h(Y) = Y$, because no nonconstant random variable is independent of itself.[2]

Problems

2-27. Suppose $X \sim \text{Bin}(n, p)$.

(a) Show that

$$E\{X(X-1)\} = n(n-1)p^2$$

Hint: Follow the pattern of Example 2.5.1.

(b) Show that

$$\text{var}(X) = np(1-p).$$

Hint: Use part (a).

2-28. Suppose $X \sim \text{Poi}(\mu)$.

(a) Show that

$$E\{X(X-1)\} = \mu^2$$

Hint: Follow the pattern of Example 2.5.2.

(b) Show that

$$\text{var}(X) = \mu.$$

Hint: Use part (a).

2-29. Verify the moments of the $\mathcal{DU}(1, n)$ distribution given in Section B.1.1 of Appendix B.
Hint: First establish

$$\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$$

by mathematical induction.

[2]Bizarrely, constant random variables are independent of all random variables, including themselves. This is just the homogeneity axiom and the "expectation of a constant is the constant" property:

$$E\{g(a)h(X)\} = g(a)E\{h(X)\} = E\{g(a)\}E\{h(X)\}$$

for any constant a and random variable X.

2-30. Verify the moments of the $\mathcal{U}(a, b)$ distribution given in Section B.2.1 of Appendix B.

2-31. The proof of Corollary 2.35 used $\operatorname{cor}(X, X) = 1$ without comment. Prove this.

2-32. Suppose $X \sim \operatorname{Gam}(\alpha, \lambda)$.

(a) For any real number $p > -\alpha$, the p-th ordinary moment

$$\alpha_p = E(X^p)$$

exists. Calculate it.

Hint: Follow the pattern of Example 2.5.3. Your answer will involve gamma functions that cannot be simplified using the recurrence relation if p is not an integer (which we didn't say it was).

(b) Show that

$$\operatorname{var}(X) = \frac{\alpha}{\lambda^2}$$

Hint: Use part (a) and the recurrence relation for gamma functions, (B.3) in Appendix B.

2-33. Suppose X has probability density

$$f(x) = \frac{3}{x^4}, \qquad x > 1$$

(note the domain).

(a) For what positive integers k is X^k in L^1?

(b) Calculate $E(X^k)$ for the positive integers k such that the expectation exists.

2-34. Suppose X has probability density

$$f(x) = \frac{1}{2\sqrt{x}}, \qquad 0 < x < 1$$

(note the domain).

(a) For what positive integers k is X^k in L^1?

(b) Calculate $E(X^k)$ for the positive integers k such that the expectation exists.

2-35. Calculate the marginal distributions for

(a) the density (2.61) and

(b) the density (2.62).

Show that the factorization criterion

(c) holds for the density (2.61) and

(d) fails for the density (2.62).

2-36. Prove Theorem 2.47.

2-37. This fills in some details left unsaid in Example 2.7.2.

(a) Prove that X and Y defined in Example 2.7.2 are uncorrelated.

Hint: Use Theorem 2.10.

(b) Prove that no nonconstant random variable is independent of itself.

Hint: If all we know is that X is nonconstant, then all we know is that there exists an event A such that $0 < P(X \in A) < 1$. Now use Definition 2.7.1.

2-38. Prove the following identities. For any $n \geq 1$

$$\mu_n = \sum_{k=0}^{n} \binom{n}{k} (-1)^k \alpha_1^k \alpha_{n-k}$$

and

$$\alpha_n = \sum_{k=0}^{n} \binom{n}{k} \alpha_1^k \mu_{n-k}$$

where, as defined in Section 2.4, μ_k is the k-th central moment and α_k is the k-th ordinary moment.
Hint: Use the binomial theorem (Problem 1-14 on p. 7 of Lindgren).

Chapter 3

Conditional Probability and Expectation

3.1 Parametric Families of Distributions

Scalar Variable and Parameter

Sometimes, like in the "brand name distributions" in Appendix B of these notes, we consider probability models having an adjustable constant in the formula for the density. Generically, we refer to such a constant as a *parameter* of the distribution. Usually, though not always, we use Greek letters for parameters to distinguish them from random variables (large Roman letters) and possible values of random variables (small Roman letters). A lot of different Greek letters are used for parameters (check out Appendix B), the Greek letter used for a "generic" parameter (when we are talking generally, not about any particular distribution) is θ (lower case theta, see Appendix A).

When we want to emphasize the dependence of the density on the parameter, we write f_θ or $f(\cdot \mid \theta)$ rather than just f for the density function and $f_\theta(x)$ or $f(x \mid \theta)$ for the value of the density function at the point x. Why two notations? The former is simpler and a good deal less clumsy in certain situations, but the latter shows explicitly the close connection between conditional probability and parametric families, which is the subject of this section and the following section.

Thus we say: let X be a random variable having density f_θ on a sample space S. This means that for each particular value of the parameter θ the function f_θ is a density, that is,

$$f_\theta(x) \ge 0, \qquad x \in S \tag{3.1a}$$

and

$$\int f_\theta(x)\, dx = 1 \tag{3.1b}$$

(with, as usual, the integral replaced by a sum in the discrete case). Note that this is exactly the usual condition for a function to be a probability density, just

like (2.47a) and (2.47b). The *only* novelty is writing f_θ in place of f. If you prefer the other notation, this condition would become

$$f(x \mid \theta) \ge 0, \qquad x \in S \tag{3.2a}$$

and

$$\int f(x \mid \theta)\, dx = 1 \tag{3.2b}$$

Again, there is no novelty here except for the purely notational novelty of writing $f(x \mid \theta)$ instead of $f_\theta(x)$ or $f(x)$.

Example 3.1.1 (The Exponential Distribution).
We want to write the exponential distribution (Section B.2.2 in Appendix B) in the notation of parametric families. The parameter is λ. We write the density as

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

or as

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

the only difference between either of these or the definition in Section B.2.2 being the notation on the left hand side: $f(x)$ or $f_\lambda(x)$ or $f(x \mid \lambda)$.

Each different value of the parameter θ gives a different probability distribution. As θ ranges over its possible values, which we call the *parameter space*, often denoted Θ when the parameter is denoted θ, we get a *parametric family* of densities

$$\{\, f_\theta : \theta \in \Theta \,\}$$

although we won't see this notation much until we get to statistics next semester.

Vector Variable or Parameter

Vector Variable

Another purely notational variant involves random vectors. We typically indicate vector variables with boldface type, as discussed in Section 1.3 of these notes, that is, we would write $f(\mathbf{x})$ or $f_\theta(\mathbf{x})$ or $f(\mathbf{x} \mid \theta)$. As usual we are sloppy about whether these are functions of a single vector variable $\mathbf{x} = (x_1, \ldots, x_n)$ or of many scalar variables x_1, ..., x_n. When we are thinking in the latter mode, we write $f(x_1, \ldots, x_n)$ or $f_\theta(x_1, \ldots, x_n)$ or $f(x_1, \ldots, x_n \mid \theta)$.

Example 3.1.2 (The Exponential Distribution).
Suppose X_1, ..., X_n are independent and identically distributed Exp(λ) random

variables. We write the density of the random vector $\mathbf{X} = (X_1, \ldots, X_n)$ as

$$\begin{aligned} f_\lambda(\mathbf{x}) &= \prod_{i=1}^n \lambda e^{-\lambda x_i} \\ &= \lambda^n \exp\left(-\lambda \sum_{i=1}^n x_i \right), \qquad x_i > 0,\ i = 1, \ldots, n. \end{aligned}$$

or, according to taste, we might write the left hand side as $f_\lambda(x_1, \ldots, x_n)$ or $f(\mathbf{x} \mid \lambda)$ or $f(x_1, \ldots, x_n \mid \lambda)$.

Vector Parameter

Similarly, when we have a vector parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$, we write the density as $f_{\boldsymbol{\theta}}(x)$ or $f(x \mid \boldsymbol{\theta})$. And, as usual, we are sloppy about whether there is really one vector parameter or several scalar parameters θ_1, ..., θ_m. When we are thinking in the latter mode, we write $f_{\theta_1, \ldots, \theta_m}(x)$ or $f(x \mid \theta_1, \ldots, \theta_m)$.

Example 3.1.3 (The Gamma Distribution).
We want to write the gamma distribution (Section B.2.3 in Appendix B) in the notation of parametric families. The parameter is $\boldsymbol{\theta} = (\alpha, \lambda)$. We write the density as

$$f_{\boldsymbol{\theta}}(x) = f_{\alpha,\lambda}(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \qquad x > 0$$

or if we prefer the other notation we write the left hand side as $f(x \mid \boldsymbol{\theta})$ or $f(x \mid \alpha, \lambda)$.

The parameter space of this probability model is

$$\Theta = \{\, (\alpha, \lambda) \in \mathbb{R}^2 : \alpha > 0,\ \lambda > 0 \,\}$$

that is, the first quadrant with boundary points excluded.

Vector Variable and Vector Parameter

And, of course, the two preceeding cases can be combined. If we have a vector random variable $\mathbf{X} = (X_1, \ldots, X_n)$ and a vector parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$, we can write write the density as any of

$$\begin{gathered} f_{\boldsymbol{\theta}}(\mathbf{x}) \\ f(\mathbf{x} \mid \boldsymbol{\theta}) \\ f_{\theta_1, \ldots, \theta_m}(x_1, \ldots, x_n) \\ f(x_1, \ldots, x_n \mid \theta_1, \ldots, \theta_m) \end{gathered}$$

according to taste.

3.2 Conditional Probability Distributions

Scalar Variables

The conditional probability distribution of one random variable Y given another X is the probability model you are supposed to use in the situation when you have seen X and know its value but have not yet seen Y and don't know its value. The point is that X is no longer random. Once you know its value x, it's a constant not a random variable.

We write the density of this probability model, the *conditional distribution of Y given X* as $f(y \mid x)$. We write expectations with respect to this model as $E(Y \mid x)$, and we write probabilities as

$$P(Y \in A \mid x) = E\{I_A(Y) \mid x\}$$

(couldn't resist an opportunity to reiterate the lesson of Section 2.6 that probability is a special case of expectation).

We calculate probabilities or expectations from the density in the usual way with integrals in the continuous case

$$E\{g(Y) \mid x\} = \int g(y) f(y \mid x)\, dy \tag{3.3}$$

$$P\{Y \in A \mid x\} = \int_A f(y \mid x)\, dy \tag{3.4}$$

and with the integrals replaced by sums in the discrete case.

Note that

> *A conditional probability density is just an ordinary probability density, when considered as a function of the variable(s) in front of the bar alone with the variable(s) behind the bar considered fixed.*

This means that in calculating a conditional probability or expectation from a conditional density

> *always integrate with respect to the variable(s) in front of the bar*

(with, of course, "integrate" replaced by "sum" in the discrete case).

Example 3.2.1 (Exponential Distribution).
Of course, one doesn't always have to do an integral or sum, expecially when a "brand name" distribution is involved. Suppose the conditional distribution of Y given X is $\text{Exp}(X)$, denoted

$$Y \mid X \sim \text{Exp}(X)$$

for short. This means, of course, that the conditional density is

$$f(y \mid x) = xe^{-xy}, \qquad y > 0$$

(just plug in x for λ in the formula in Section B.2.2 in Appendix B), but we don't need to use the density to calculate the conditional expectation, because we know that the mean of the $\text{Exp}(\lambda)$ distribution is $1/\lambda$, hence (again just plugging in x for λ

$$E(Y \mid x) = \frac{1}{x}$$

or

$$E(Y \mid X) = \frac{1}{X}$$

depending on whether we are thinking of the variable behind the bar as random (big X) or fixed (little x) As we shall see, both viewpoints are useful and we shall use both in different situations.

If the known formulas for a "brand name" distribution don't answer the question, then we do need an integral

$$\begin{aligned} P(a < Y < b \mid x) &= \int_a^b f(y \mid x)\, dy \\ &= \int_a^b x e^{-xy}\, dy \\ &= -e^{-xy} \Big|_a^b \\ &= e^{-xa} - e^{-xb} \end{aligned}$$

and, of course, if we are thinking of X as being random too, we would write

$$P(a < Y < b \mid X) = e^{-aX} - e^{-bX}$$

just the same except for big X instead of little x.

The astute reader will by now have understood from the hint given by the notation why this chapter started with a section on the seemingly unrelated topic of parametric families of distributions.

> *Conditional probability distributions are no different from parametric families of distributions.*

For each fixed value of x, the conditional density $f(y \mid x)$, considered as a function of y alone, is just an ordinary probability density. Hence it satisfies the two properties

$$f(y \mid x) \ge 0, \qquad \text{for all } y \tag{3.5a}$$

and

$$\int f(y \mid x)\, dy = 1 \tag{3.5b}$$

(with the integral replaced by a sum in the discrete case). Notice that there is no difference, except a purely notational one, between the pair of conditions (3.5a) and (3.5b) and the pair of conditions (3.2a) and (3.2b). Here we have a

Roman letter behind the bar; there we had a Greek letter behind the bar, but (mathematics is invariant under changes of notation) that makes no *conceptual* difference whatsoever.

The fact that conditional probability is a special case of ordinary probability (when we consider the variable or variables behind the bar fixed) means that we already know a lot about conditional probability. Every fact we have learned so far in the course about *ordinary* probability and expectation applies to its special case *conditional* probability and expectation. **Caution:** What we just said applies *only* when the variable(s) behind the bar are considered fixed. As we shall see, things become more complicated when both are treated as random variables.

Vector Variables

Of course, either of the variables involved in a conditional probability distribution can be vectors. Then we write either of

$$f(\mathbf{y} \mid \mathbf{x})$$
$$f(y_1, \ldots y_n \mid x_1, \ldots x_m)$$

according to taste, and similarly either of

$$E(\mathbf{Y} \mid \mathbf{x})$$
$$E(Y_1, \ldots Y_n \mid x_1, \ldots x_m)$$

Since we've already made this point in the context of parametric families of distributions, and conditional probability distributions are no different, we will leave it at that.

3.3 Axioms for Conditional Expectation

The conditional expectation $E(Y \mid x)$ is just another expectation operator, obeying all the axioms for expectation. This follows from the view explained in the preceeding section that conditional expectation is a special case of ordinary unconditional expectation (at least when we are considering the variable or variables behind the bar fixed). If we just replace unconditional expectations with conditional expectations everywhere in the axioms for unconditional expectation, they are still true.

There are, however, a couple of additional axioms for conditional expectation. Axiom E2 can be strengthened (as described in the next section), and an entirely new axiom (described in the two sections following the next) can be added to the set of axioms.

3.3.1 Functions of Conditioning Variables

Any function of the variable or variables behind the bar (the *conditioning* variables) behaves like a constant in conditional expectations.

Axiom CE1. *If Y is in L^1 and a is any function, then*

$$E\{a(X)Y \mid X\} = a(X)E(Y \mid X).$$

We don't have to verify that conditional expectation obeys the axioms of ordinary unconditional expectation, because conditional expectation is a special case of unconditional expectation (when thought about the right way), but this axiom isn't a property of unconditional expectation, so we do need to verify that it holds for conditional expectation as we have already defined it. But the verification is easy.

$$\begin{aligned} E\{a(X)Y \mid X\} &= \int a(X)yf(y \mid X)\,dy \\ &= a(X)\int yf(y \mid X)\,dy \\ &= a(X)E(Y \mid X) \end{aligned}$$

because any term that is not a function of the variable of integration can be pulled outside the integral (or sum in the discrete case).

Two comments:

- We could replace big X by little x if we want

  $$E\{a(x)Y \mid x\} = a(x)E(Y \mid x)$$

  though, of course, this now follows from Axiom E2 of ordinary expectation because $a(x)$ is a constant when x is a constant.

- We could replace big Y by any random variable, for example, $g(Y)$ for any function g, obtaining

  $$E\{a(X)g(Y) \mid X\} = a(X)E\{g(Y) \mid X\}.$$

3.3.2 The Regression Function

It is now time to confront squarely an issue we have been tiptoeing around with comments about writing $E(Y \mid x)$ or $E(Y \mid X)$ "according to taste." In order to clearly see the contrast with unconditional expectation, let first review something about ordinary unconditional expectation.

$E(X)$ is not a function of X. It's a constant, not a random variable.

This doesn't conflict with the fact that an expectation operator is a function $E : L^1 \to \mathbb{R}$ when considered abstractly. This is the usual distinction between a function and it's values: E is indeed a function (from L^1 to $\mathbb{R}$), but $E(X)$ isn't a function, it's the value that the expectation operator assigns to the random variable X, and that value is a real number, a constant, not a random variable (not a function on the sample space).

So $E(X)$ is very different from $g(X)$, where g is an ordinary function. The latter is a random variable (any function of a random variable is a random variable).

So what's the corresponding fact about conditional expectation?

> *$E(Y \mid X)$ is not a function of Y, but it is a function of X, hence a random variable.*

We saw this in Example 3.2.1

$$Y \mid X \sim \text{Exp}(X)$$

implies

$$E(Y \mid X) = \frac{1}{X}$$

which is, apparently, a function of X and not a function of Y.

In a way, there is nothing surprising here. If we consider the conditioning variable fixed, then $E(Y \mid x)$ is just a special case of ordinary expectation. Hence $E(Y \mid x)$ is not a function of Y any more than $E(Y)$ is. Furthermore, $E(Y \mid x)$ is not a random variable because x isn't a random variable (little x).

In another way, this is surprising. If we consider the conditioning variable to be random, then it no longer looks like conditional expectation is a special case of ordinary expectation, because the former is a random variable and the latter isn't! What happens is that which is a special case of which gets turned around.

> *Unconditional expectation is the special case of conditional expectation obtained by conditioning on an empty set of variables.*

This accords with the naive view that a conditional probability model for Y given X is what you use when you have seen X but not yet seen Y. Clearly, what you use when you have seen (nothing) but not yet seen Y is the the ordinary unconditional models we have been using all along. It says that $E(Y)$ can be thought of as $E(Y \mid \quad)$ with nothing behind the bar. Applying our other slogan to this special case we see that

> *$E(Y) = E(Y \mid \quad)$ is not a function of Y, but it is a function of (nothing), hence a constant random variable.*

Thus when we think of unconditional expectation as a special case of conditional expectation $E(Y)$ isn't a constant but a constant random variable, which is almost the same thing—only a mathematician and a rather pedantic one could care about the difference.

So we have two somewhat conflicting views of conditional probability and expectation.

- When we consider the conditioning variables (the variables behind the bar) fixed, conditional expectation is just a special case of ordinary unconditional expectation. The conditioning variables behave like parameters of the probability model.

- When we consider the conditioning variables (the variables behind the bar) random, unconditional expectation is just a special case of conditional expectation, what happens when we condition on an empty set of variables.

What's to blame for the confusion is partly just the notation, it's not clear from the notation that $E(Y \mid X)$ is a function of X but not a function of Y, and partly the real conflict between seeing the conditioning variable sometimes as random and sometimes as constant. There's nothing to be done about the second problem except to be very careful to always understand which situation you are in. For the first, we can change terminology and notation.

If $E(Y \mid X)$ is a function of X, we can write it as a function of X, say $g(X)$. In Example 3.2.1 we had

$$E(Y \mid X) = g(X) = \frac{1}{X}$$

which means that g is the function defined by

$$g(x) = \frac{1}{x}, \qquad x > 0$$

just an ordinary function of an ordinary variable, that is, g is an ordinary function, and $g(x)$ is an ordinary number, but, of course, $g(X)$ is a random variable (because of the big X).

Another name for this function g is the *regression function of Y on X*. When it's clear from the context which is the conditioning variable and which is the other variable, we can say just *regression function*. But when any confusion might arise, the longer form is essential. The regression function of Y on X, that is, $E(Y \mid X)$ is quite different from the regression function of X on Y, that is, $E(X \mid Y)$. For one thing, the former is a function of X and the latter is a function of Y. But not only that, they are in general quite different and unrelated functions.

3.3.3 Iterated Expectations

We saw in the preceding section that $E(Y \mid X)$ is a random variable, a function of X, say $g(X)$. This means we can take its expectation

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

The left hand side is nothing unusual, just an expectation like any other. The right hand side looks like something new. We call it an "iterated expectation" (an unconditional expectation of a conditional expectation). Iterated expectation has a very important property which is the last axiom for conditional probability.

Axiom CE2. *If $Y \in L^1$, then*

$$E\{E(Y \mid X)\} = E(Y). \tag{3.6}$$

A proof that the notion of conditional expectation we have so far developed satisfies this axiom will have to wait until the next section. First we give some examples and consequences.

Example 3.3.1 (Random Sum of Random Variables).
Suppose X_0, X_1, ... is an infinite sequence of identically distributed random variables, having mean $E(X_i) = \mu_X$, and suppose N is a nonnegative integer-valued random variable independent of the X_i and having mean $E(N) = \mu_N$. It is getting a bit ahead of ourselves, but we shall see in the next section that this implies

$$E(X_i \mid N) = E(X_i) = \mu_X. \tag{3.7}$$

Question: What is the expectation of

$$S_N = X_1 + \cdots + X_N$$

(a sum with a random number N of terms and each term X_i a random variable) where the sum with zero terms when $N = 0$ is defined to be zero?

Linearity of expectation, which applies to conditional as well as unconditional probability, implies

$$\begin{aligned} E(S_N \mid N) &= E(X_1 + \cdots X_n \mid N) \\ &= E(X_1 \mid N) + \cdots + E(X_n \mid N) \\ &= E(X_1) + \cdots + E(X_N) \\ &= N\mu_X \end{aligned}$$

the next to last equality being (3.7). Hence by the iterated expectation axiom

$$E(S_N) = E\{E(S_N \mid N)\} = E(N\mu_X) = E(N)\mu_X = \mu_N\mu_X.$$

Note that this example is impossible to do any other way than using the iterated expectation formula. Since no formulas were given for any of the densities, you can't use any formula involving explicit integrals.

If we combine the two conditional probability axioms, we get the following.

Theorem 3.1. *If X and Y are random variables and g and h are functions such that $g(X)$ and $h(Y)$ are in L^1, then*

$$E\{g(X)E[h(Y) \mid X]\} = E\{g(X)h(Y)\}. \tag{3.8}$$

Proof. Replace Y by $g(X)h(Y)$ in Axiom CE2 obtaining

$$E\{E[g(X)h(Y) \mid X]\} = E\{g(X)h(Y)\}.$$

then apply Axiom CE1 to pull $g(X)$ out of the inner conditional expectation obtaining (3.8). □

The reader should be advised that our treatment of conditional expectation is a bit unusual. Rather than state two axioms for conditional expectation, standard treatments in advanced probability textbooks give just one, which is essentially the statement of this theorem. As we have just seen, our two axioms imply this one, and conversely our two axioms are special cases of this

one: taking $g = a$ and h the identity function in (3.8) gives our Axiom CE1, and taking $g = 1$ and h the identity function in (3.8) gives our Axiom CE2. Thus our treatment characterizes the same notion of conditional probability as standard treatments.

Another aspect of advanced treatments of conditional probability is that standard treatments usually take the statement Theorem 3.1 as a *definition* rather than an *axiom*. The subtle difference is the following uniqueness assertion.

Theorem 3.2. *If X and Y are random variables and h is a function such that $h(Y) \in L^1$, then then there exists a function f such that $f(X) \in L^1$ and*

$$E\{g(X)f(X)\} = E\{g(X)h(Y)\} \tag{3.9}$$

for every function g such that $g(X)h(Y) \in L^1$. The function f is unique up to redefinition on sets of probability zero.

The proof of this theorem is far beyond the scope of this course. Having proved this theorem, advanced treatments take it as a definition of conditional expectation. The unique function f whose existence is guaranteed by the theorem is defined to be the conditional expectation, that is,

$$E\{h(Y) \mid X\} = f(X).$$

The theorem makes it clear that (as everywhere else in probability theory) redefinition on a set (event) of probability zero makes no difference.

Although we cannot prove Theorem 3.2, we can use it to prove a fancy version of the iterated expectation formula.

Theorem 3.3. *If $Y \in L^1$, then*

$$E\bigl\{E(Z \mid X, Y) \,\big|\, X\bigr\} = E(Z \mid X). \tag{3.10}$$

Of course, the theorem also holds when the conditioning variables are vectors, that is, if $m < n$

$$E\bigl\{E(Z \mid X_1, \ldots, X_n) \,\big|\, X_1, \ldots X_m\bigr\} = E(Z \mid X_1, \ldots, X_m).$$

In words, an iterated conditional expectation (a conditional expectation inside another conditional expectation) is just the conditional expectation conditioning on the set of variables of the outer conditional expectation, if the set of conditioning variables in the outer expectation is a *subset* of the conditioning variables in the inner expectation. That's a mouthful. The formula (3.10) is simpler.

Proof of Theorem 3.3. By Theorem 3.2 and the following comment,

- $E(Z \mid X, Y)$ is the unique (up to redefinition on sets of probability zero) function $f_1(X, Y)$ such that

  $$E\{g_1(X, Y)f_1(X, Y)\} = E\{g_1(X, Y)Z\} \tag{3.11a}$$

  for all functions g_1 such that $g_1(X, Y)Z \in L^1$.

- The iterated expectation on the left hand side of (3.10) is the unique (up to redefinition on sets of probability zero) function $f_2(X)$ such that

$$E\{g_2(X)f_2(X)\} = E\{g_2(X)f_1(X,Y)\} \tag{3.11b}$$

  for all functions g_2 such that $g_2(X)f_1(X,Y) \in L^1$.

- $E(Z \mid X)$ is the unique (up to redefinition on sets of probability zero) function $f_3(X)$ such that

$$E\{g_3(X)f_3(X)\} = E\{g_3(X)Z\} \tag{3.11c}$$

  for all functions g_3 such that $g_3(X)Z \in L^1$.

Since (3.11a) holds for any function g_1, it holds when $g_1(X,Y) = g_3(X)$, from which, combining (3.11a) and (3.11c), we get

$$E\{g_3(X)f_3(X)\} = E\{g_3(X)Z\} = E\{g_3(X)f_1(X,Y)\} \tag{3.11d}$$

Reading (3.11d) from end to end, we see it is the same as (3.11b), because (3.11d) must hold for any function g_3 and (3.11b) must hold for any function g_2. Thus by the uniqueness assertion of Theorem 3.2 we must have $f_2(X) = f_3(X)$, except perhaps on a set of probability zero (which does not matter). Since $f_2(X)$ is the left hand side of (3.10) and $f_3(X)$ is the right hand side, that is what was to be proved. □

Theorem 3.2 can also be used to prove a very important fact about independence and conditioning.

Theorem 3.4. *If X and Y are independent random variables and h is a function such that $h(Y) \in L^1$, then*

$$E\{h(Y) \mid X\} = E\{h(Y)\}.$$

In short, conditioning on an *independent* variable or variables is the same as conditioning on *no* variables, making conditional expectation the same as unconditional expectation.

Proof. If X and Y are independent, the right hand side of (3.9) becomes $E\{g(X)\}E\{h(Y)\}$ by Definition 2.7.2. Hence, in this special case, Theorem 3.2 asserts that $E\{h(Y) \mid X\}$ is the unique function $f(X)$ such that

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

whenever $g(X) \in L^1$. Certainly the constant $f(X) = a$, where $a = E\{h(Y)\}$ is one such function, because

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

so by the uniqueness part of Theorem 3.2 this is the conditional expectation, as was to be proved. □

3.4 Joint, Conditional, and Marginal

As was the case with unconditional expectation, our "axioms first" treatment of conditional expectation has been a bit abstract. When the problem is solved by pulling a function of the conditioning variables outside of a conditional expectation or by the iterated expectation formula, either the special case in Axiom CE2 with the outside expectation an unconditional one or the general case in Theorem 3.3 in which both expectations are conditional, then the axioms are just what you need. But for other problems you need to be able to calculate conditional probability densities and expectations by doing sums and integrals, and that is the subject to which we now turn.

3.4.1 Joint Equals Conditional Times Marginal

Note that the iterated expectation axiom (Axiom CE2), when we write out the expectations as integrals, equates

$$\begin{aligned} E\{E(Y \mid X)\} &= \int \left(\int y f(y \mid x) \, dy \right) f_X(x) \, dx \\ &= \iint y f(y \mid x) f_X(x) \, dx \, dy \end{aligned} \tag{3.12a}$$

and

$$E(Y) = \iint y f(x, y) \, dx \, dy. \tag{3.12b}$$

Equation (3.12b) is correct, because of the general definition of expectation of a function of two variables:

$$E\{g(X, Y)\} = \iint g(x, y) f(x, y) \, dx \, dy$$

whenever the expectation exists. Now just take $g(x, y) = y$.

One way that the right hand sides of (3.12a) and (3.12b) can be equal is if

$$f(x, y) = f(y \mid x) f_X(x) \tag{3.13}$$

or in words,

$$\text{joint} = \text{conditional} \times \text{marginal}$$

In fact, by the uniqueness theorem (Theorem 3.2), this is the only way the iterated expectation axiom can hold, except, as usual, for possible redefinition on sets of probability zero.

This gives a formula for calculating a conditional probability density from the joint

$$f(y \mid x) = \frac{f(x, y)}{f_X(x)} \tag{3.14}$$

or in words,

$$\text{conditional} = \frac{\text{joint}}{\text{marginal}}$$

Of course, there is a slight problem with (3.14) when the denominator is zero, but since the set of x such that $f_X(x) = 0$ is a set of probability zero, this does not matter, and $f(y \mid x)$ can be defined arbitrarily for all such x.

Example 3.4.1 (Uniform Distribution on a Triangle).
This continues Example 1.5.2. Recall from that example that if X and Y have joint density

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

that the marginal of X is

$$f_X(x) = 2(1 - x), \qquad 0 < x < 1.$$

Thus the conditional is

$$f(y \mid x) = \frac{2}{2(1 - x)} = \frac{1}{1 - x}$$

Or we should say this is the conditional for *some* values of x and y. As usual, we have to be careful about domains of definition or we get nonsense. First, the marginal only has the formula we used when $0 < x < 1$, so that is one requirement. Then for x in that range, the joint is only defined by the formula we used when $0 < y$ and $x + y < 1$, that is, when $0 < y < 1 - x$. Thus to be precise, we must say

$$f(y \mid x) = \frac{1}{1 - x}, \qquad 0 < y < 1 - x \text{ and } 0 < x < 1. \tag{3.15}$$

What about other values of x and y? What if we want the definition for all real x and y? First, for $f(y \mid x)$ to be a probability density (considered as a function of y for fixed x) it must integrate to 1 (integrating with respect to y). Since our formula already does integrate to one over its domain of definition $0 < y < 1 - x$, it must be zero elsewhere. Thus when $0 < x < 1$

$$f(y \mid x) = \begin{cases} \frac{1}{1-x}, & 0 < y < 1 - x \\ 0, & \text{elsewhere} \end{cases}$$

or, if you prefer a definition using an indicator function,

$$f(y \mid x) = \frac{1}{1 - x} I_{(0,1-x)}(y), \qquad y \in \mathbb{R}.$$

What about x outside $(0, 1)$? Those are x such that the marginal is zero, so the formula "joint over marginal" is undefined. As we have already said, the definition is then arbitrary, so we may say

$$f(y \mid x) = 42$$

or whatever we please when $x \leq 0$ or $1 \leq x$. (It doesn't even matter that this function doesn't integrate to one!) Mostly we will ignore such nonsense and

only define conditional densities where the values are not arbitrary and actually matter. The only reason we mention this issue at all is so that you won't think $f(y \mid x)$ has to have a sensible definition for all possible x.

So how about conditional expectations? Given the formula (3.15) for the conditional density, we just plug and chug

$$E(Y \mid x) = \int y f(y \mid x)\, dy = \frac{1 - x}{2} \tag{3.16a}$$

$$E(Y^2 \mid x) = \int y^2 f(y \mid x)\, dy = \frac{(1 - x)^2}{3} \tag{3.16b}$$

$$\operatorname{var}(Y \mid x) = E(Y^2 \mid x) - E(Y \mid x)^2 = \frac{(1 - x)^2}{12} \tag{3.16c}$$

and so forth, (3.16c) holding because of Corollary 2.12, which like every other fact about unconditional expectation, also holds for conditional expectation so long as we are considering the conditioning variables fixed.

We could end Section 3.4 right here. Formulas (3.13) and (3.14) tell us how to calculate conditionals from joints and joints from conditionals and marginals. And the fact that "conditional expectation is a special case of ordinary expectation" (so long as we are considering the conditioning variables fixed) tells how to compute expectations. So what else is there to know? Well, nothing, but a lot more can be said on the subject. The rest of Section 3.4 should give you a much better feel for the subject and allow you to calculate conditional densities and expectations more easily.

3.4.2 Normalization

A standard homework problem for courses like this specifies some nonnegative function $h(x)$ and then asks for what real number k is $f(x) = kh(x)$ a probability density.

Clearly we must have $k > 0$, because $k < 0$ would entail negative probabilities and $k = 0$ would make the density integrate (or sum in the discrete case) to zero. Either violates the defining properties for a probability density, which are (1.20a) and (1.20b) in the discrete case and (1.21a) and (1.21b) in the continuous case.

For reasons that will soon become apparent, we prefer to use $c = 1/k$. This is allowed because $k \neq 0$. Thus the problem becomes: for what real number c is

$$f(x) = \frac{1}{c} h(x)$$

a density function? The process of determining c is called *normalization* and c is called the *normalizing constant* for the *unnormalized density* $h(x)$.

To determine c we use the second defining property for a probability density (1.20b) or (1.21b) as the case may be, which implies

$$c = \int h(x)\, dx \tag{3.17}$$

(with integration replaced by summation if the probability model is discrete). In order for c to be a positive number, the integral (or sum in the discrete case) must exist and be nonzero. This gives us two conditions on unnormalized densities. A real-valued function $h(x)$ is an *unnormalized density* provided the following two conditions hold.

- It is nonnegative: $h(x) \ge 0$, for all x.
- It is integrable in the continuous case or summable in the discrete case and the integral or sum is nonzero.

Then

$$f(x) = \frac{1}{c} h(x)$$

is a normalized probability density, where c is given by (3.17) in the continuous case and by (3.17) with the integral replaced by a sum in the discrete case.

Example 3.4.2.
Consider the function

$$h(x) = x^{\alpha - 1} e^{-x}, \qquad x > 0,$$

where $\alpha > 0$. How do we normalize it to make a probability density?

The normalizing constant is

$$c = \int_0^\infty x^{\alpha - 1} e^{-x} \, dx = \Gamma(\alpha)$$

by (B.2) in Appendix B. Thus we obtain a gamma distribution density

$$f(x) = \frac{1}{\Gamma(\alpha)} x^{\alpha - 1} e^{-x}.$$

So what's the big deal? We already knew that! Is "normalization" just a fancy name for something trivial? Well, yes and no. You can form your own opinion, but not until the end of Section 3.4.

3.4.3 Renormalization

We start with a slogan

Conditional probability is renormalization.

What this means will become apparent presently.

First, $f(y \mid x)$ is just an ordinary probability density when considered as a function of y for fixed x. We maintain this view, y is the variable and x is fixed, throughout this subsection.

Second, since x is fixed, the denominator in

$$f(y \mid x) = \frac{f(x, y)}{f_X(x)} \tag{3.18}$$

is constant (not a function of y). Thus we can also write

$$f(y \mid x) \propto f(x, y) \tag{3.19}$$

the symbol $\propto$ meaning "proportional to" (still thinking of y as the only variable, the proportionality does not hold if we vary x). This says the joint is just like the conditional, at least proportional to it, the only thing wrong is that it doesn't integrate to one (still thinking of y as the only variable, the joint does, of course, integrate to one if we integrate with respect to x and y). Formula (3.19) says that if we graph the conditional and the joint (as functions of y!) we get the same picture, they are the same shape, the only difference is the scale on the vertical axis (the constant of proportionality). So if we put in the constant of proportionality, we get

$$f(y \mid x) = \frac{1}{c(x)} f(x, y). \tag{3.20}$$

We have written the "constant" as $c(x)$ because it is a function of x, in fact, comparing with (3.18) we see that

$$c(x) = f_X(x).$$

We call it a "constant" because we are considering x fixed.

All of this can be summarized in the following slogan.

> *A joint density is an unnormalized conditional density. Its normalizing constant is a marginal density.*

Spelled out in more detail, the joint density $f(x, y)$ considered as a function of y alone is an unnormalized probability density, in fact, is it proportional to the conditional density (3.19). In order to calculate the conditional density, we need to calculate the normalizing constant, which just happens to turn out to be the marginal $f_X(x)$, and divide by it (3.18).

If we take this argument a bit further and plug the definition of the marginal into (3.18), we get

$$f(y \mid x) = \frac{f(x, y)}{\int f(x, y)\, dy} \tag{3.21}$$

This shows more explicitly how "conditional probability is renormalization." You find a conditional probability density by dividing the joint density by what it integrates to. How do we remember which variable is the variable of integration here? That's easy. In this whole subsection y is the only variable; x is fixed. In general, a conditional density is an ordinary density (integrates to one, etc.) when considered a function of the variable "in front of the bar" with the conditioning variable, the variable "behind the bar" *fixed*. That's what we are doing here. Hence we divide by the integral of the joint density with respect to the variable "in front of the bar."

It is occasionally useful that (3.21) holds whether or not the joint density is normalized. Suppose we are given an unnormalized joint density $h(x, y)$ so that

$$f(x, y) = \frac{1}{c} h(x, y)$$

for some normalizing constant c. Plugging this into (3.21) gives

$$f(y \mid x) = \frac{h(x, y)}{\int h(x, y)\, dy} \tag{3.22}$$

The c's cancel in the numerator and denominator.

Our slogan about conditional probability and renormalization helps us remember which marginal is meant in

$$\text{conditional} = \frac{\text{joint}}{\text{marginal}}$$

- If the conditional in question is $f(y \mid x)$, then we are considering y the variable (x is fixed).
- Thus the marginal in question is the one obtained by integrating with respect to y (that's what we are considering variable).
- The marginal obtained by integrating out y is the marginal of the *other* variable (slogan on p. 19 in these notes). Hence the marginal is $f_X(x)$.

But even if you are confused about how to calculate marginals or which marginal you need to divide by, you should still be able to calculate conditionals using (3.21) and (3.22), which *contain no marginals* and are in fact derivable on the spot. Both are obvious consequences of the facts that

- Conditional densities are proportional to joint densities considered as functions of the variable(s) in front of the bar.
- Conditional densities integrate to one considered as functions of the variable(s) in front of the bar.

Example 3.4.3.
Consider the function

$$h(x, y) = (x + y^2)e^{-x-y}, \qquad x > 0,\ y > 0.$$

If we take this to be an unnormalized joint density, what are the two conditional densities $f(x \mid y)$ and $f(y \mid x)$?

Integrating with respect to x gives

$$\begin{aligned} \int_0^\infty h(x, y)\, dx &= e^{-y} \int_0^\infty x e^{-x}\, dx + y^2 e^{-y} \int_0^\infty e^{-x}\, dx \\ &= (1 + y^2)e^{-y} \end{aligned}$$

We used the formula

$$\int_0^\infty x^n e^{-x}\, dx = \Gamma(n + 1) = n! \tag{3.23}$$

to evaluate the integrals. Hence

$$f(x \mid y) = \frac{f(x, y)}{\int f(x, y)\, dx} = \frac{x + y^2}{1 + y^2} e^{-x}$$

Similarly

$$\begin{aligned} \int_0^\infty h(x, y)\, dy &= x e^{-x} \int_0^\infty e^{-y}\, dy + e^{-x} \int_0^\infty y^2 e^{-y}\, dy \\ &= (x + 2) e^{-x} \end{aligned}$$

Again, we used (3.23) to evaluate the integrals. So

$$f(y \mid x) = \frac{f(x, y)}{\int f(x, y)\, dy} = \frac{x + y^2}{x + 2} e^{-y}$$

Things become considerably more complicated when the support of the joint density is not a rectangle with sides parallel to the axes. Then the domains of integration depend on the values of the conditioning variable.

Example 3.4.4 (A Density with Weird Support).
Consider the function

$$h(x, y) = \begin{cases} x + y^2, & x > 0,\ y > 0,\ x + y < 1 \\ 0, & \text{otherwise} \end{cases}$$

If we take this to be an unnormalized joint density, what is the conditional density $f(x \mid y)$?

Integrating with respect to x gives

$$\int_{-\infty}^\infty h(x, y)\, dx = \int_0^{1-y} (x + y^2)\, dx = \left. \frac{x^2}{2} + x y^2 \right|_0^{1-y} = \tfrac{1}{2}(1 - y)(1 - y + 2y^2)$$

What is tricky is that the formula $x + y^2$ for $h(x, y)$ is valid only when $x > 0$ and $y > 0$ and $x + y < 1$. This means $0 < x < 1 - y$. For other values of x, the integrand is zero. Hence the domain of integration in the second integral must be $0 < x < 1 - y$. If you miss this point about the domain of integration, you make a complete mess of the problem. If you get this point, the rest is easy

$$f(x \mid y) = \frac{f(x, y)}{\int f(x, y)\, dx} = \frac{2(x + y^2)}{(1 - y)(1 - y + 2y^2)}$$

3.4.4 Renormalization, Part II

This subsection drops the other shoe in regard to "conditional probability is renormalization." So is conditional expectation. Plugging the definition (3.21) of conditional densities into (3.3) gives

$$E\{g(Y) \mid x\} = \frac{\int g(y) f(x, y)\, dy}{\int f(x, y)\, dy} \tag{3.24}$$

(and, of course, the discrete case is analogous with the integrals replaced by sums). It is a useful mnemonic device to write (3.24) lining up the analogous bits in the numerator and denominator

$$E\{g(Y) \mid x\} = \frac{\int g(y) f(x, y)\, dy}{\int \qquad f(x, y)\, dy}.$$

This looks a little funny, but it reminds us that the density in the numerator and denominator is the same, and the variable of integration is the same. The only difference between the numerator and denominator is the function $g(y)$ appearing in the numerator.

If we plug in (3.22) instead of (3.21) for $f(y \mid x)$ we get

$$E\{g(Y) \mid x\} = \frac{\int g(y) h(x, y)\, dy}{\int \qquad h(x, y)\, dy} \tag{3.25}$$

where $h(x, y)$ is an *unnormalized* joint density.

These formulas make it clear that we are choosing the denominator so that $E(1 \mid x) = 1$, which is the form the norm axiom takes when applied to conditional probability. That is, when we take the special case in which the function $g(y)$ is equal to one for all y, the numerator and denominator are the same.

Example 3.4.5.
Suppose X and Y have the unnormalized joint density

$$h(x, y) = (x + y)e^{-x-y}, \qquad x > 0,\ y > 0,$$

what is $E(X \mid y)$?

Using (3.25) with the roles of X and Y interchanged and g the identity function we get

$$\begin{aligned} E(X \mid y) &= \frac{\int x h(x, y)\, dx}{\int \quad h(x, y)\, dx} \\ &= \frac{\int x(x + y)e^{-x-y}\, dx}{\int \quad (x + y)e^{-x-y}\, dx} \end{aligned}$$

Using (3.23) the denominator is

$$\begin{aligned} \int_0^\infty (x + y)e^{-x-y}\, dx &= e^{-y} \int_0^\infty x e^{-x}\, dx + y e^{-y} \int_0^\infty e^{-x}\, dx \\ &= (1 + y)e^{-y} \end{aligned}$$

and the numerator is

$$\begin{aligned} \int_0^\infty x(x + y)e^{-x-y}\, dx &= e^{-y} \int_0^\infty x^2 e^{-x}\, dx + y e^{-y} \int_0^\infty x e^{-x}\, dx \\ &= (2 + y)e^{-y} \end{aligned}$$

Hence

$$E(X \mid y) = \frac{2 + y}{1 + y}, \qquad y > 0.$$

Recall from p. 90 in these notes

> **Sanity Check:** *$E(X \mid Y)$ is a function of Y and is not a function of X.*

Good. We did get a function of y. If you get confused about which variable to integrate with respect to, this sanity check will straighten you out. If you through some mistake get a function of both variables, this sanity check will at least tell you that you messed up somewhere.

3.4.5 Bayes Rule

Now we want to study the consequences of

$$\text{joint} = \text{conditional} \times \text{marginal} \tag{3.26}$$

Again we have the problem of remembering which marginal. If we recall our analysis of

$$\text{conditional} = \frac{\text{joint}}{\text{marginal}}$$

on p. 100 in these notes, we recall that it is the marginal of the variable "behind the bar."

Because "mathematics is invariant under changes of notation" (3.26) is also true when we interchange the roles of the variables Hence we can "factor" a joint density into marginal and conditional two different ways

$$f(x, y) = f(x \mid y) f_Y(y) \tag{3.27}$$
$$f(x, y) = f(y \mid x) f_X(x) \tag{3.28}$$

Plugging (3.27) into (3.21) gives

$$f(y \mid x) = \frac{f(x \mid y) f_Y(y)}{\int f(x \mid y) f_Y(y)\, dy} \tag{3.29}$$

This equation is called *Bayes rule*. It allows us to "turn around" conditional probabilities. That is, it is useful for problems that say: given $f(x \mid y)$, find $f(y \mid x)$. Or vice versa. Of course, because "mathematics is invariant under changes of notation" (3.29) is also true with all the x's and y's interchanged.

Example 3.4.6.
Suppose that X and Y are positive real-valued random variables and

$$f(x \mid y) = \tfrac{1}{2} x^2 y^3 e^{-xy}$$
$$f_Y(y) = e^{-y}$$

what is $f(y \mid x)$?

Note that this is slightly tricky in that the conditional wanted is not the one given by the Bayes rule formula (3.29). You need to interchange x's and y's in (3.29) to get the formula needed to do this problem

$$f(y \mid x) = \frac{f(x \mid y) f_Y(y)}{\int f(x \mid y) f_Y(y) \, dy}$$

The denominator is

$$\int_0^\infty x^2 y^3 e^{-xy-y} \, dy = x^2 \int_0^\infty y^3 e^{-(1+x)y} \, dy$$

The change of variable $y = u/(1+x)$ makes the right hand side

$$\frac{x^2}{(1+x)^4} \int_0^\infty u^3 e^{-u} \, du = \frac{6x^2}{(1+x)^4}$$

Thus

$$f(y \mid x) = \tfrac{1}{6}(1+x)^4 y^3 e^{-(1+x)y}, \qquad y > 0$$

Example 3.4.7 (Bayes and Brand Name Distributions).
Suppose

$$\begin{aligned} X &\sim \text{Exp}(\lambda) \\ Y \mid X &\sim \text{Exp}(X) \end{aligned}$$

meaning the marginal distribution of X is $\text{Exp}(\lambda)$ and the conditional distribution of Y given X is $\text{Exp}(X)$, that is,

$$f(y \mid x) = x e^{-xy}, \qquad y > 0. \tag{3.30}$$

This is a bit tricky, so let's go through it slowly. The formula for the density of the exponential distribution given in Section B.2.2 in Appendix B is

$$f(x \mid \lambda) = \lambda e^{-\lambda x}, \qquad x > 0. \tag{3.31}$$

We want to change x to y and λ to x. Note that it matters which order we do the substitution. If we change λ to x first, we get

$$f(x \mid x) = \lambda e^{-x^2}, \qquad x > 0.$$

but that's nonsense. First, the right hand side isn't a density. Second, the left hand side is the density of X given X, but this distribution is concentrated at X (if we know X, then we know X) and so isn't even continuous. So change x in (3.31) to y obtaining

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

and then change λ to x obtaining (3.30).

Of course, the joint is conditional times marginal

$$f(x, y) = f(y \mid x) f_X(x) = x e^{-xy} \cdot \lambda e^{-\lambda x} = \lambda x e^{-(\lambda + y)x} \tag{3.32}$$

Question: What is the other marginal (of Y) and the other conditional (of X given Y)? Note that these two problems are related. If we answer one, the answer to the other is easy, just a division

$$f(x \mid y) = \frac{f(x, y)}{f_Y(y)}$$

or

$$f_Y(y) = \frac{f(x, y)}{f(x \mid y)}$$

I find it a bit easier to get the conditional first. Note that the joint (3.32) is an unnormalized conditional when thought of as a function of x alone. Checking our inventory of "brand name" distributions, we see that the only one like (3.32) in having both a power and an exponential of the variable is the gamma distribution with density

$$f(x \mid \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha - 1} e^{-\lambda x}, \qquad x > 0. \tag{3.33}$$

Comparing the analogous parts of (3.32) and (3.33), we see that we must match up x with $x^{\alpha - 1}$, which tells us we need $\alpha = 2$, and we must match up $e^{-(\lambda + y)x}$ with $e^{-\lambda x}$ which tells us we need $\lambda + y$ in (3.32) to be the λ in (3.33), which is the second parameter of the gamma distribution. Thus (3.32) must be an unnormalized $\Gamma(2, \lambda + y)$ density, and the properly normalized density is

$$f(x \mid y) = (\lambda + y)^2 x e^{-(\lambda + y)x}, \qquad x > 0 \tag{3.34}$$

Again this is a bit tricky, so let's go through it slowly. We want to change α to 2 and λ to $\lambda + y$ in (3.33). That gives

$$f(x \mid y) = \frac{(\lambda + y)^2}{\Gamma(2)} x^{2 - 1} e^{-(\lambda + y)x}, \qquad x > 0.$$

and this cleans up to give (3.34).

3.5 Conditional Expectation and Prediction

The parallel axis theorem (Theorem 2.11 in these notes)

$$E[(X - a)^2] = \operatorname{var}(X) + [a - E(X)]^2$$

has an analog for conditional expectation. Just replace expectations by conditional expectations (and variances by conditional variances) and, because functions of the conditioning variable behave like constants, replace the constant by a function of the conditioning variable.

Theorem 3.5 (Conditional Parallel Axis Theorem). *If $Y \in L^1$*

$$E\{[Y - a(X)]^2 \mid X\} = \operatorname{var}(Y \mid X) + [a(X) - E(Y \mid X)]^2 \tag{3.35}$$

The argument is exactly the same as that given for the unconditional version, except for the need to use Axiom CE1 instead of Axiom E2 to pull a function of the conditioning variable out of the conditional expectation. Otherwise, only the notation changes.

If we take the unconditional expectation of both sides of (3.35), we get

$$E\big(E\{[Y - a(X)]^2 \mid X\}\big) = E\{\operatorname{var}(Y \mid X)\} + E\{[a(X) - E(Y \mid X)]^2\}$$

and by the iterated expectation axiom, the left hand side is the the unconditional expectation, that is,

$$E\{[Y - a(X)]^2\} = E\{\operatorname{var}(Y \mid X)\} + E\{[a(X) - E(Y \mid X)]^2\} \tag{3.36}$$

This relation has no special name, but it has two very important special cases. The first is the prediction theorem.

Theorem 3.6. *For predicting a random variable Y given the value of another random variable X, the predictor function $a(X)$ that minimizes the expected squared prediction error*

$$E\{[Y - a(X)]^2\}$$

is the conditional expectation $a(X) = E(Y \mid X)$.

The proof is extremely simple. The expected squared prediction error is the left hand side of (3.36). On the right hand side of (3.36), the first term does not contain $a(X)$. The second term is the expectation of the square of $a(X) - E(Y \mid X)$. Since a square is nonnegative and the expectation of a nonnegative random variable is nonnegative (Axiom E1), the second term is always nonnegative and hence is minimized when it is zero. By Theorem 2.32, that happens if and only if $a(X) = E(Y \mid X)$ with probability one. (Yet another place where redefinition on a set of probability zero changes nothing of importance).

Example 3.5.1 (Best Prediction).
Suppose X and Y have the unnormalized joint density

$$h(x, y) = (x + y)e^{-x-y}, \qquad x > 0,\ y > 0,$$

what function of Y is the best predictor of X in the sense of minimizing expected squared prediction error?

The predictor that minimizes expected squared prediction error is the regression function

$$a(Y) = E(X \mid Y) = \frac{2 + Y}{1 + Y}$$

found in Example 3.4.5.

The other important consequence of (3.36) is obtained by taking $a(X) = E(Y) = \mu_Y$ (that is, a is the constant function equal to μ_Y). This gives

$$E\{[Y - \mu_Y]^2\} = E\{\operatorname{var}(Y \mid X)\} + E\{[\mu_Y - E(Y \mid X)]^2\} \tag{3.37}$$

The left hand side of (3.37) is, by definition $\operatorname{var}(Y)$. By the iterated expectation axiom, $E\{E(Y \mid X)\} = E(Y) = \mu_Y$, so the second term on the right hand side is the expected squared deviation of $E(Y \mid X)$ from its expectation, which is, by definition, its variance. Thus we have obtained the following theorem.

Theorem 3.7 (Iterated Variance Formula). *If $Y \in L^2$,*

$$\operatorname{var}(Y) = E\{\operatorname{var}(Y \mid X)\} + \operatorname{var}\{E(Y \mid X)\}.$$

Example 3.5.2 (Example 3.3.1 Continued).
Suppose X_0, X_1, ... is an infinite sequence of identically distributed random variables, having mean $E(X_i) = \mu_X$ and variance $\operatorname{var}(X_i) = \sigma_X^2$, and suppose N is a nonnegative integer-valued random variable independent of the X_i having mean $E(N) = \mu_N$ and variance $\operatorname{var}(N) = \sigma_N^2$. Note that we have now tied up the loose end in Example 3.3.1. We now know from Theorem 3.4 that independence of the X_i and N implies

$$E(X_i \mid N) = E(X_i) = \mu_X.$$

and similarly

$$\operatorname{var}(X_i \mid N) = \operatorname{var}(X_i) = \sigma_X^2.$$

Question: What is the variance of

$$S_N = X_1 + \cdots + X_N$$

expressed in terms of the means and variances of the X_i and N?

This is easy using the iterated variance formula. First, as we found in Example 3.3.1,

$$E(S_N \mid N) = NE(X_i \mid N) = N\mu_X.$$

A similar calculation gives

$$\operatorname{var}(S_N \mid N) = N \operatorname{var}(X_i \mid N) = N\sigma_X^2$$

(because of the assumed independence of the X_i and N). Hence

$$\begin{aligned} \operatorname{var}(S_N) &= E\{\operatorname{var}(S_N \mid N)\} + \operatorname{var}\{E(S_N \mid N)\} \\ &= E(N\sigma_X^2) + \operatorname{var}(N\mu_X) \\ &= \sigma_X^2 E(N) + \mu_X^2 \operatorname{var}(N) \\ &= \sigma_X^2 \mu_N + \mu_X^2 \sigma_N^2 \end{aligned}$$

Again notice that it is impossible to do this problem any other way. There is not enough information given to use any other approach.

Also notice that the answer is not exactly obvious. You might just guess, using your intuition, the answer to Example 3.3.1. But you wouldn't guess this. You need the theory.

Problems

3-1. In class we found the moment generating function of the geometric distribution (Section B.1.3 in Appendix B) is defined by

$$\psi(t) = \frac{1 - p}{1 - pe^t}$$

on some neighborhood of zero. Find the variance of this random variable.

3-2. Verify the details in (3.16a), (3.16b), and (3.16c).

3-3. Suppose X is a positive random variable and the density of Y given X is

$$f(y \mid x) = \frac{2y}{x^2}, \qquad 0 < y < x.$$

(a) Find $E(Y \mid X)$.

(b) Find $\operatorname{var}(Y \mid X)$.

3-4. For what real values of θ is

$$f_\theta(x) = \frac{1}{c(\theta)} x^\theta, \qquad 0 < x < 1$$

a probability density, and what is the function $c(\theta)$?

3-5. Suppose X, Y, and Z are random variables such that

$$E(X \mid Y, Z) = Y \quad \text{and} \quad \operatorname{var}(X \mid Y, Z) = Z.$$

Find the (unconditional) mean and variance of X in terms of the means, variances, and covariance of Y and Z.

3-6. Suppose the random vector (X, Y) is uniformly distributed on the disk

$$S = \{\, (x, y) \in \mathbb{R}^2 : x^2 + y^2 < 4 \,\}$$

that is, (X, Y) has the $\mathcal{U}(S)$ distribution in the notation of Section B.2.1 of Appendix B.

(a) Find the conditional distributions of X given Y and of Y given X.

(b) Find the marginal distributions of X and Y.

(c) Find $E(Y \mid x)$.

(d) Find $P(|Y| < 1 \mid x)$.

3-7. Suppose the conditional distribution of Y given X is $\mathcal{N}(0, 1/X)$ and the marginal distribution of X is $\operatorname{Gam}(\alpha, \lambda)$.

(a) What is the conditional density of X given Y?

(b) What is the marginal density of Y?

3-8. Suppose X and Z are independent random variables and $E(Z) = 0$. Define $Y = X + X^2 + Z$.

(a) Find $E(Y \mid X)$.

(b) Find $\text{var}(Y \mid X)$.

(c) What function of X is the best predictor of Y in the sense of minimizing expected squared prediction error?

(d) What is the expected squared prediction error of this predictor?

Note: Any of the answers may involve moments of X and Z.

Chapter 4

Parametric Families of Distributions

The first thing the reader should do before reading the rest of this chapter is go back and review Section 3.1, since that establishes the basic notation for parametric families of distributions.

4.1 Location-Scale Families

Consider a probability density f of a real-valued random variable X. By the theorem on linear changes of variables (Theorem 7 of Chapter 3 in Lindgren), for any real number μ and any positive real number σ, the random variable $Y = \mu + \sigma X$ has the density

$$f_{\mu,\sigma}(y) = \frac{1}{\sigma} f\left(\frac{y-\mu}{\sigma}\right).$$

This generates a two-parameter family of densities called the *location-scale* family generated by the *reference density* f. The parameter μ is called the *location* parameter, and the parameter σ is called the *scale* parameter.

We could choose any distribution in the family as the reference distribution with density f. This gives a different *parameterization* of the family, but the *same* family. Suppose we choose $f_{\alpha,\beta}$ as the reference density. The family it generates has densities

$$\begin{aligned} f_{\mu,\sigma}(y) &= \frac{1}{\sigma} f_{\alpha,\beta}\left(\frac{y-\mu}{\sigma}\right). \\ &= \frac{1}{\sigma\beta} f\left(\frac{1}{\beta}\left[\frac{y-\mu}{\sigma} - \alpha\right]\right) \\ &= \frac{1}{\sigma\beta} f\left(\frac{y-\mu-\sigma\alpha}{\sigma\beta}\right) \end{aligned}$$

It is clear that as μ and σ run over all possible values we get the same *family* of distributions as before. The parameter values that go with each particular distribution have changed, but each density that appears in one family also appears in the other. The correspondence between the parameters in the two parameterizations is

$$\begin{aligned} \mu &\longleftrightarrow \mu + \sigma\alpha \\ \sigma &\longleftrightarrow \sigma\beta \end{aligned}$$

If the reference random variable X has a variance, then every distribution in the family has a variance (by Theorem 2.44 in these notes), and the distributions of the family have every possible mean and variance. Since we are free to choose the reference distribution as any distribution in the family, we may as well choose so that $E(X) = 0$ and $\operatorname{var}(X) = 1$, then μ is the mean and σ the standard deviation of the variable Y with density $f_{\mu,\sigma}$.

But the distributions of the family do not have to have either means or variances. In that case we cannot call μ the mean or σ the standard deviation. That is the reason why in general we call μ and σ the location and scale parameters.

Example 4.1.1 (Uniform Distributions).
The $\mathcal{U}(a, b)$ family of distribution defined in Section B.2.1 of Appendix B has densities

$$f(x \mid a, b) = \frac{1}{b - a}, \qquad a < x < b \tag{4.1}$$

and moments

$$\begin{aligned} E(X \mid a, b) &= \frac{a + b}{2} \\ \operatorname{var}(X \mid a, b) &= \frac{(b - a)^2}{12} \end{aligned}$$

Therefore the parameters a and b of the distribution having mean zero and standard deviation one is found by solving

$$\frac{a + b}{2} = 0$$

(from which we see that $b = -a$) and

$$\frac{(b - a)^2}{12} = 1$$

which becomes, plugging in $b = -a$,

$$\frac{(2 \cdot b)^2}{12} = 1$$

Hence $b = \sqrt{3}$. Giving the density

$$f(x) = \frac{1}{2\sqrt{3}}, \qquad -\sqrt{3} < x < +\sqrt{3}$$

Then use the formula for a general location-scale family, obtaining

$$f(y \mid \mu, \sigma) = \frac{1}{\sigma} f\left(\frac{y - \mu}{\sigma}\right) = \frac{1}{2\sigma\sqrt{3}}$$

on the domain of definition, whatever that is. The change of variable is $y = \mu + \sigma x$, so $x = \pm\sqrt{3}$ maps to $\mu \pm \sigma\sqrt{3}$, and those are the endpoints of the domain of definition. So

$$f(y \mid \mu, \sigma) = \frac{1}{2\sigma\sqrt{3}}, \qquad \mu - \sigma\sqrt{3} < y < \mu + \sigma\sqrt{3} \tag{4.2}$$

The reader may have lost track in all the formula smearing of how simple this all is. We have another description of the *same* family of densities. The correspondence between the two parameterizations is

$$a \longleftrightarrow \mu - \sigma\sqrt{3}$$
$$b \longleftrightarrow \mu + \sigma\sqrt{3}$$

It should be clear that (4.2) defines a density that is constant on an interval, just like (4.1) does. Furthermore, it should also be clear that as μ and σ range over all possible values we get distributions on all possible intervals. This is not so obvious from the range specification in (4.2), but is clear from the definition of μ and σ in terms of a and b

$$\mu = \frac{a + b}{2}$$
$$\sigma = \sqrt{\frac{(b - a)^2}{12}}$$

The only virtue of the new parameterization (4.2) over the old one (4.1) is that it explicitly describes the density in terms of the mean and standard deviation (μ is the mean and σ is the standard deviation, as explained in the comments immediately preceding the example). But for most people that is not a good enough reason to use the more complicated parameterization. Hence (4.1) is much more widely used.

Example 4.1.2 (Cauchy Distributions).
The function

$$f(x) = \frac{1}{\pi(1 + x^2)}, \qquad -\infty < x < +\infty$$

is a probability density, because

$$\int_{-\infty}^{\infty} \frac{1}{1 + x^2}\, dx = \tan^{-1} x \Big|_{-\infty}^{\infty} = \pi$$

This density is called the standard Cauchy density (Section 6.12 in Lindgren). This distribution has no mean or variance. If we try to calculate

$$E(|X|) = \int_{-\infty}^{\infty} \frac{|x|}{1 + x^2}\, dx = 2 \int_{0}^{\infty} \frac{x}{1 + x^2}\, dx$$

we see that, because the integrand is bounded, only the behavior of the integrand near infinity is important. And for large x

$$\frac{x}{1 + x^2} \approx \frac{1}{x}$$

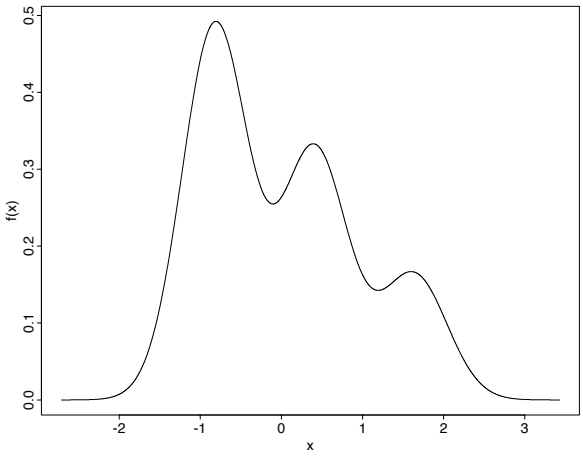
and so by Lemma 2.39 the integral does not exist. Hence by Theorem 2.44 neither does any moment of first or higher order. That is, no moments exist.

The Cauchy location-scale family has densities

$$f_{\mu,\sigma}(x) = \frac{\sigma}{\pi(\sigma^2 + [x - \mu]^2)}, \qquad -\infty < x < +\infty \tag{4.3}$$

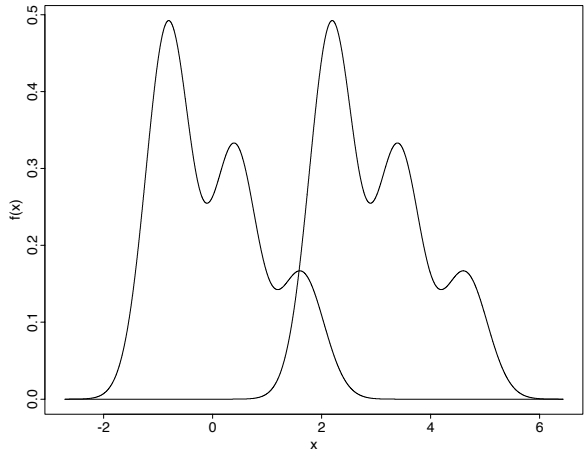
Here μ is not the mean, because Cauchy distributions do not have means. It is, however, the median because this distribution is symmetric with center of symmetry μ. Neither is σ the standard deviation, because Cauchy distributions do not have variances.

Example 4.1.3 (Blurfle Distributions).
All of the distributions in a location-scale family have the same *shape*. In fact we could use the same curve as the graph of every density in the family. Changing μ and σ only changes the scales on the axes, not the shape of the curve. Consider the distribution with the density shown below, which is of no particular interest, just an arbitrary p. d. f. Call it the "blurfle" distribution. It has been chosen so to have mean zero and variance one, so we can refer to it as the *standard* blurfle distribution.

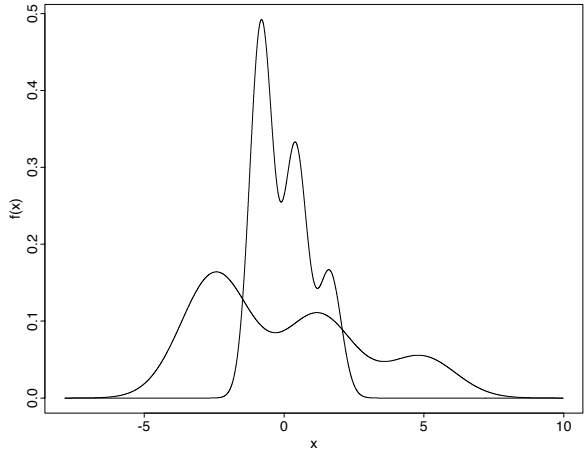


Like any other distribution, it generates a location-scale family, which we can call the blurfle family. Different blurfle distributions have the same shape, just different location and scale parameters. Changing the location parameter, but leaving the scale parameter unchanged just shifts the curve to the right or left along the number line.

Shown below are two different blurfle densities with same scale parameter but different location parameters.



And shown below are two different blurfle densities with same location parameter but different scale parameters.



4.2 The Gamma Distribution

The gamma function is defined for all real $\alpha > 0$ by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx. \tag{4.4}$$

Theorem 4.1 (Gamma Function Recursion Relation).

$$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha) \tag{4.5}$$

holds for all $\alpha > 0$.

Proof. This can be proved using the integration by parts formula: $\int u\,dv = uv - \int v\,du$. Let $u = x^\alpha$ and $dv = e^{-x}\,dx$, so $du = \alpha x^{\alpha-1}\,du$ and $v = -e^{-x}$, and

$$\begin{aligned}\Gamma(\alpha+1) &= \int_0^\infty x^\alpha e^{-x}\,dx \\ &= -x^\alpha e^{-x}\Big|_0^\infty - \int_0^\infty \alpha x^{\alpha-1} e^{-x}\,dx \\ &= \alpha\Gamma(\alpha)\end{aligned}$$

The uv term in the integration by parts is zero, because $x^\alpha e^{-x}$ goes to zero as x goes to either zero or infinity. □

Since

$$\Gamma(1) = \int_0^\infty e^{-x}\,dx = -e^{-x}\Big|_0^\infty = 1,$$

the gamma function interpolates the factorials

$$\begin{aligned}\Gamma(2) &= 1 \cdot \Gamma(1) = 1! \\ \Gamma(3) &= 2 \cdot \Gamma(2) = 2! \\ &\vdots \\ \Gamma(n+1) &= n \cdot \Gamma(n) = n!\end{aligned}$$

In a later section, we will find out that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, which can be used with the recursion relation (4.5) to find $\Gamma(\frac{n}{2})$ for odd positive integers n.

The integrand in the integral defining the gamma function (4.4) is nonnegative and integrates to a finite, nonzero constant. Hence, as we saw in Example 3.4.2, dividing it by what it integrates to makes a probability density

$$f(x \mid \alpha) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}, \qquad x > 0. \tag{4.6}$$

The parameter α of the family is neither a location nor a scale parameter. Each of these densities has a different shape. Hence we call it a *shape parameter*.

It is useful to enlarge the family of densities by adding a scale parameter. If X has the density (4.6), then σX has the density

$$f(x \mid \alpha, \sigma) = \frac{1}{\sigma} f\Big(\frac{x}{\sigma} \;\Big|\; \alpha\Big) = \frac{1}{\sigma^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\sigma}. \tag{4.7}$$

For reasons that will become apparent later Lindgren prefers to use the reciprocal scale parameter $\lambda = 1/\sigma$. If the units of X are feet, then so are the units of σ. The units of λ are reciprocal feet (ft^{-1}). In this parameterization the densities are

$$f(x \mid \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}. \tag{4.8}$$

You should be warned that there is no generally accepted parameterization of the gamma family of densities. Some books prefer one, some the other. In this

course we will always use (4.8), and following Lindgren we will use the notation $\text{Gam}(\alpha, \lambda)$ to denote the distribution with density (4.8). We will call λ the *inverse scale parameter* or, for reasons to be explained later (Section 4.4.3), the *rate parameter*. The fact that (4.8) must integrate to one tells us

$$\int_0^\infty x^{\alpha-1} e^{-\lambda x}\, dx = \frac{\Gamma(\alpha)}{\lambda^\alpha}.$$

We can find the mean and variance of the gamma using the trick of recognizing a probability density (Section 2.5.7).

$$\begin{aligned} E(X) &= \int_0^\infty x f(x \mid \alpha, \lambda)\, dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^\alpha e^{-\lambda x}\, dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+1)}{\lambda^{\alpha+1}} \\ &= \frac{\alpha}{\lambda} \end{aligned}$$

(we used the recursion (4.5) to simplify the ratio of gamma functions). Similarly

$$\begin{aligned} E(X^2) &= \int_0^\infty x^2 f(x \mid \alpha, \lambda)\, dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha+1} e^{-\lambda x}\, dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+2)}{\lambda^{\alpha+2}} \\ &= \frac{(\alpha+1)\alpha}{\lambda^2} \end{aligned}$$

(we used the recursion (4.5) twice). Hence

$$\text{var}(X) = E(X^2) - E(X)^2 = \frac{(\alpha+1)\alpha}{\lambda^2} - \left(\frac{\alpha}{\lambda}\right)^2 = \frac{\alpha}{\lambda^2}$$

The sum of independent gamma random variables with the same scale parameter is also gamma. If $X_1, \ldots, X_k$ are independent with $X_i \sim \text{Gam}(\alpha_i, \lambda)$, then

$$X_1 + \cdots + X_k \sim \text{Gam}(\alpha_1 + \cdots + \alpha_k, \lambda).$$

This will be proved in the following section (Theorem 4.2).

4.3 The Beta Distribution

For any real numbers $s > 0$ and $t > 0$, the function

$$h(x) = x^{s-1}(1-x)^{t-1}, \qquad 0 < x < 1$$

is an unnormalized probability density. This is clear when $s \geq 1$ and $t \geq 1$, because then it is bounded. When $s < 1$, it is unbounded near zero. When $t < 1$, it is unbounded near one. But even when unbounded it is integrable. For x near zero

$$h(x) \approx x^{s-1}$$

Hence h is integrable on $(0, \epsilon)$ for any $\epsilon > 0$ by Lemmas 2.40 and 2.43 because the exponent $s-1$ is greater than -1. The same argument (or just changing the variable from x to $1 - x$) shows that the unnormalized density h is integrable near one.

The normalizing constant for h depends on s and t and is called the beta function

$$B(s, t) = \int_0^1 x^{s-1}(1 - x)^{t-1} \, dx.$$

Dividing by the normalizing constant gives normalized densities

$$f(x \mid s, t) = \frac{1}{B(s, t)} x^{s-1}(1 - x)^{t-1}, \qquad 0 < x < 1.$$

The probability distributions having these densities are called *beta distributions* and are denoted $\text{Beta}(s, t)$.

The next theorem gives the "addition rule" for gamma distributions mentioned in the preceding section and a connection between the gamma and beta distributions.

Theorem 4.2. *If X and Y are independent random variables*

$$X \sim \text{Gam}(s, \lambda)$$
$$Y \sim \text{Gam}(t, \lambda)$$

Then

$$U = X + Y$$
$$V = \frac{X}{X + Y}$$

are also independent random variables, and

$$U \sim \text{Gam}(s + t, \lambda)$$
$$V \sim \text{Beta}(s, t)$$

Proof. To use the multivariate change of variable formula, we first solve for the old variables x and y in terms of the new

$$x = uv$$
$$y = u(1 - v)$$

Hence the Jacobian is

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

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

$$\begin{aligned} f_{U,V}(u,v) &= f_{X,Y}[uv, u(1-v)]|J(u,v)| \\ &= f_X(uv)f_Y[u(1-v)]u \\ &= \frac{\lambda^s}{\Gamma(s)}(uv)^{s-1}e^{-\lambda uv}\frac{\lambda^t}{\Gamma(t)}[u(1-v)]^{t-1}e^{-\lambda u(1-v)}u \\ &= \frac{\lambda^{s+t}}{\Gamma(s)\Gamma(t)}u^{s+t-1}e^{-\lambda u}v^{s-1}(1-v)^{t-1} \end{aligned}$$

Since the joint density factors into a function of u times a function of v, the variables U and V are independent. Since these functions are proportional to the gamma and beta densities asserted by the theorem, U and V must actually have these distributions. □

Corollary 4.3.

$$B(s,t) = \frac{\Gamma(s)\Gamma(t)}{\Gamma(s+t)}$$

Proof. The constant in the joint density found in the proof of the theorem must be the product of the constants for the beta and gamma densities. Hence

$$\frac{\lambda^{s+t}}{\Gamma(s)\Gamma(t)} = \frac{\lambda^{s+t}}{\Gamma(s+t)}\frac{1}{B(s,t)}$$

Solving for $B(s,t)$ gives the corollary. □

For moments of the beta distribution, see Lindgren pp. 176–177.

4.4 The Poisson Process

4.4.1 Spatial Point Processes

A *spatial point process* is a random pattern of points in a region of space. The space can be any dimension.

A point process is *simple* if it never has points on top of each other so that each point of the process is at a different location in space. A point process is *boundedly finite* if with probability one it has only a finite number of points in any bounded set.

Let N_A denote the number of points in a region A. Since the point pattern is random, N_A is a random variable. Since it counts points, N_A is a discrete

random variable taking values 0, 1, 2, ... If A is a bounded set and the point process is boundedly finite, then the event $N_A = \infty$ has probability zero.

A point x is a *fixed atom* if $P(N_{\{x\}} > 0) > 0$, that is, if there is positive probability of seeing a point at the particular location x in every random pattern. We are interested in point processes in which the locations of the points are continuous random variables, in which case the probability of seeing a point at any particular location is zero, so there are no fixed atoms.

For a general spatial point process, the joint distribution of the variables N_A for various sets A is very complicated. There is one process for which it is not complicated. This is the Poisson process, which is a model for a "completely random" pattern of points. One example of this process is given in Figure 4.1.

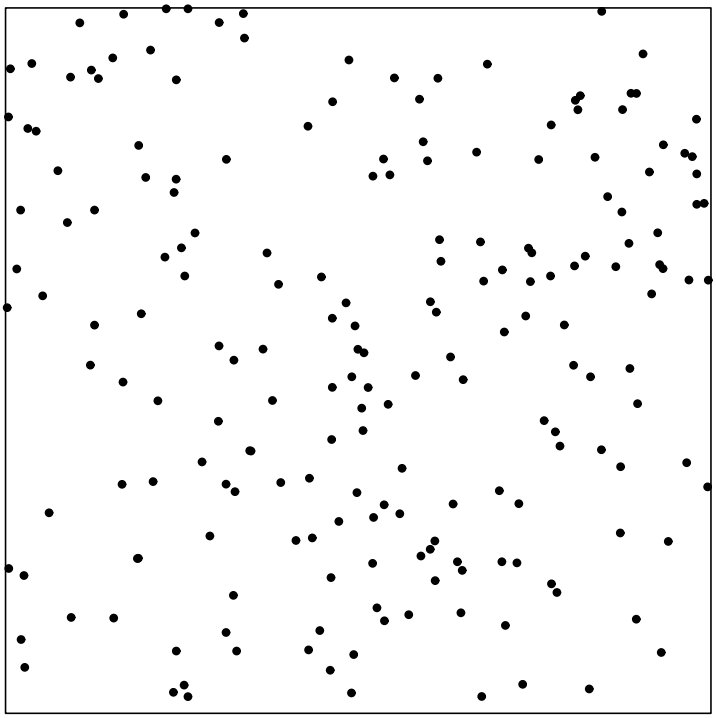


Figure 4.1: A single realization of a homogeneous Poisson process.

4.4.2 The Poisson Process

A Poisson process is a spatial point process characterized by a simple independence property.

Definition 4.4.1.
*A **Poisson process** is a simple, boundedly finite spatial point process with no fixed atoms having the property that N_{A_1}, N_{A_2}, ..., N_{A_k} are independent random variables, whenever A_1, A_2, ..., A_k are disjoint bounded sets.*

In short, counts of points in disjoint regions are independent random variables. It is a remarkable fact that the independence property alone determines the distribution of the counts.

Theorem 4.4. *For a Poisson process, N_A has a Poisson distribution for every bounded set A. Conversely, a simple point process with no fixed atoms such that N_A has a Poisson distribution for every bounded set A is a Poisson process.*

Write $\Lambda(A) = E(N_A)$. Since the parameter of the Poisson distribution is the mean, the theorem says N_A has the Poisson distribution with parameter $\Lambda(A)$. The function $\Lambda(A)$ is called the *intensity measure* of the process.

An important special case of the Poisson process occurs when the intensity measure is proportional to ordinary measure (length in one dimension, area in two, volume in three, and so forth): if we denote the ordinary measure of a region A by $m(A)$, then

$$\Lambda(A) = \lambda m(A) \tag{4.9}$$

for some $\lambda > 0$. The parameter λ is called the *rate parameter* of the process. A Poisson process for which (4.9) holds, the process is said to be a *homogeneous Poisson process*. Otherwise it is *inhomogeneous*.

The space could be the three-dimensional space of our ordinary experience. For example, the points could be the locations of raisins in a carrot cake. If the process is homogeneous, that models the situation where regions of equal volume have an equal number of raisins on average, as would happen if the batter was stirred well and the raisins didn't settle to the bottom of the cake pan before baking. If the process is inhomogeneous, that models the situation where some regions get more raisins per unit volume than others on average. Either the batter wasn't stirred well or the raisins settled or something of the sort.

There are two important corollaries of the characterization theorem.

Corollary 4.5. *The sum of independent Poisson random variables is a Poisson random variable.*

If $X_i \sim \text{Poi}(\mu_i)$ then the X_i could be the counts N_{A_i} in disjoint regions A_i having measures $m(A_i) = \mu_i$ in a homogeneous Poisson process with unit rate parameter. The sum is the count in the combined region

$$X_1 + \cdots + X_n = N_{A_1 \cup \cdots \cup A_n}$$

which has a Poisson distribution with mean

$$m(A_1 \cup \cdots \cup A_n) = m(A_1) + \cdots m(A_n)$$

because the measure of the union of disjoint regions is the sum of the measures. This is also obvious from linearity of expectation. We must have

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

Corollary 4.6. *The conditional distribution of a Poisson process in a region A^c given the process in A is the same as the unconditional distribution of the process in A^c.*

In other words, finding the point pattern in A tells you nothing whatsoever about the pattern in A^c. The pattern in A^c has the same distribution conditionally or unconditionally.

Proof. By Definition 4.4.1 and Theorem 4.4 N_B is independent of N_C when $B \subset A^c$ and $C \subset A$. Since this is true for all such C, the random variable N_B is independent of the whole pattern in A, and its conditional distribution given the pattern in A is the same as its unconditional distribution. Theorem 4.4 says Poisson distributions of the N_B for all subsets B of A^c imply that the process in A^c is a Poisson process. □

4.4.3 One-Dimensional Poisson Processes

In this section we consider Poisson processes in one-dimensional space, that is, on the real line. So a realization of the process is a pattern of points on the line. For specificity, we will call the dimension along the line "time" because for many applications it is time. For example, the calls arriving at a telephone exchange are often modeled by a Poisson process. So are the arrivals of customers at a bank teller's window, or at a toll plaza on an toll road. But you should remember that there is nothing in the theory specific to time. The theory is the same for all one-dimensional Poisson processes.

Continuing the time metaphor, the points of the process will always in the rest of this section be called *arrivals*. The time from a fixed point to the next arrival is called the *waiting time* until the arrival.

The special case of the gamma distribution with shape parameter one is called the exponential distribution, denoted $\text{Exp}(\lambda)$. Its density is

$$f(x) = \lambda e^{-\lambda x}, \qquad x > 0. \tag{4.10}$$

Theorem 4.7. *The distribution of the waiting time in a homogeneous Poisson process with rate parameter λ is $\text{Exp}(\lambda)$. The distribution is the same unconditionally, or conditional on the past history up to and including the time we start waiting.*

Call the waiting time X and the point where we start waiting a. Fix an $x > 0$, let $A = (a, a + x)$, and let $Y = N_{(a,a+x)}$ be the number of arrivals in the interval A. Then Y has a Poisson distribution with mean $\lambda m(A) = \lambda x$, since

measure in one dimension is length. Then the c. d. f. of X is given by

$$\begin{aligned} F(x) &= P(X \le x) \\ &= P(\text{there is at least one arrival in } (a, a + x)) \\ &= P(Y \ge 1) \\ &= 1 - P(Y = 0) \\ &= 1 - e^{-\lambda x} \end{aligned}$$

Differentiating gives the density (4.10).

The assertion about the conditional and unconditional distributions being the same is just the fact that the process on $(-\infty, a]$ is independent of the process on $(a, +\infty)$. Hence the waiting time distribution is the same whether or not we condition on the point pattern in $(-\infty, a]$.

The length of time between two consecutive arrivals is called the *interarrival time*. Theorem 4.7 also gives the distribution of the interarrival times, because it says the distribution is the same whether or not we condition on there being an arrival at the time we start waiting. Finally, the theorem says an interarrival time is independent of any past interarrival times. Since independence is a symmetric property (X is independent of Y if and only if Y is independent of X), this means all interarrival times are independent.

This means we can think of a one-dimensional Poisson process two different ways.

- The number of arrivals in disjoint intervals are independent Poisson random variables. The number of arrivals in an interval of length t is $\text{Poi}(\lambda t)$.

- Starting at an arbitrary point (say time zero), the waiting time to the first arrival is $\text{Exp}(\lambda)$. Then all the successive interarrival times are also $\text{Exp}(\lambda)$. And all the interarrival times are independent of each other and the waiting time to the first arrival.

  Thus if X_1, X_2, ... are i. i. d. $\text{Exp}(\lambda)$ random variables, the times T_1, T_2, ... defined by

  $$T_n = \sum_{i=1}^{n} X_i \tag{4.11}$$

  form a Poisson process on $(0, \infty)$.

Note that by the addition rule for the gamma distribution, the time of the nth arrival is the sum of n i. i. d. $\text{Gam}(1, \lambda)$ random variables and hence has a $\text{Gam}(n, \lambda)$ distribution.

These two ways of thinking give us a c. d. f. for the $\text{Gam}(n, \lambda)$ distribution

of T_n.

$$\begin{aligned} F(x) &= P(T_n \le x) \\ &= P(\text{there are at least n arrivals in } (0, x)) \\ &= 1 - P(\text{there are no more than } n - 1 \text{ arrivals in } (0, x)) \\ &= 1 - \sum_{k=0}^{n-1} \frac{(\lambda x)^k}{k!} e^{-\lambda x} \end{aligned}$$

Unfortunately, this trick does not work for gamma distributions with noninteger shape parameters. There is no closed form expression for the c. d. f. of a general gamma distribution.

In problems, it is best to use the way of thinking that makes the problem easiest.

Example 4.4.1.
Assume the service times for a bank teller form a homogeneous Poisson process with rate parameter λ. I arrive at the window, and am fifth in line with four people in front of me. What is the expected time until I leave?

There are four interarrival times and the waiting time until the first person in line is finished. All five times are i. i. d. Exp(λ) by the Poisson process assumption. The times have mean $1/\lambda$. The expectation of the sum is the sum of the expectations $5/\lambda$.

Alternatively, the distribution of the time I leave is the sum of the five interarrival and waiting times, which is Gam$(5, \lambda)$, which has mean $5/\lambda$.

Example 4.4.2.
With the same assumptions in the preceding example, suppose $\lambda = 10$ per hour. What is the probability that I get out in less than a half hour.

This is the probability that there are at least five points of the Poisson process in the interval $(0, 0.5)$, measuring time in hours (the time I leave is the fifth point in the process). The number of points Y has a Poi(λt) distribution with $t = 0.5$, hence $\lambda t = 5$. From Table II in the back of Lindgren $P(Y \ge 5) = 1 - P(Y \le 4) = 1 - .44 = .56$.

Problems

4-1. Prove Corollary 4.5 for the case of two Poisson random variables directly using the convolution formula Theorem 1.7 from Chapter 1 of these notes. Note that the two Poisson variables are allowed to have different means.
Hint: Use the binomial theorem (Problem 1-14 on p. 7 of Lindgren).

4-2. Suppose X_1, X_2, ... are i. i. d. random variables with mean μ and variance σ^2, and N is a Geo(p) random variable independent of the X_i. What is the mean and variance of

$$Y = X_1 + X_2 + \cdots + X_N$$

(note N is random).

4-3. A brand of raisin bran averages 84.2 raisins per box. The boxes are filled from large bins of well mixed raisin bran. What is the standard deviation of the number of raisins per box.

4-4. Let X be the number of winners of a lottery. If we assume that players pick their lottery numbers at random, then their choices are i. i. d. random variables and X is binomially distributed. Since the mean number of winners is small, the Poisson approximation is very good. Hence we may assume that $X \sim \text{Poi}(\mu)$ where μ is a constant that depends on the rules of the lottery and the number of tickets sold.

Because of our independence assumption, what other players do is independent of what you do. Hence the conditional distribution of the number of other winners given that you win is also $\text{Poi}(\mu)$. If you are lucky enough to win, you must split the prize with X other winners. You win $A/(X+1)$ where A is the total prize money. Thus

$$E\left(\frac{A}{X+1}\right)$$

is your expected winnings given that you win. Calculate this expectation.

4-5. Suppose X and Y are independent, but not necessarily identically distributed Poisson random variables, and define $N = X + Y$.

(a) Show that

$$X \mid N \sim \text{Bin}(N, p),$$

where p is some function of the parameters of the distributions of X, Y. Specify the function.

(b) Assume

$$Z \mid N \sim \text{Bin}(N, q),$$

where $0 < q < 1$. Show that

$$Z \sim \text{Poi}(\mu),$$

where μ is some function of q and the parameters of the distribution of X, Y. Specify the function.

4-6. Suppose $X \sim \text{Gam}(\alpha, \lambda)$. Let $Y = 1/X$.

(a) For which values of α and λ does $E(Y)$ exist?

(b) What is $E(Y)$ when it exists?

4-7. Suppose that X, Y, and Z are independent $\mathcal{N}(2, 2)$ random variables. What is $P(X > Y + Z)$? **Hint:** What is the distribution of $X - Y - Z$?

Chapter 5

Multivariate Distribution Theory

5.1 Random Vectors

5.1.1 Vectors, Scalars, and Matrices

It is common in linear algebra to refer to single numbers as *scalars* (in contrast to vectors and matrices). So in this chapter a real variable x or a real-valued random variable X will also be referred to as a *scalar variable* or a *scalar random variable*, respectively.

A *matrix* (plural *matrices*) is a rectangular array of scalars, called called the *elements* or *components* of the matrix, considered as a single mathematical object. We use the convention that matrices are denoted by boldface capital letters. The elements of a matrix are indicated by double subscripts, for example the elements of a matrix $\mathbf{A}$ may be denoted a_{ij}. Conventionally, the array is displayed as follows

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ & \vdots & & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{pmatrix} \tag{5.1}$$

The first index indicates the element's row, and the second index indicates the column. The matrix (5.1) has *row dimension* m and *column dimension* n, which is indicated by saying it is an $m \times n$ matrix.

The *transpose* of a matrix $\mathbf{A}$ with elements a_{ij} is the matrix $\mathbf{A}'$ with elements a_{ji}, that is, $\mathbf{A}'$ is obtained from $\mathbf{A}$ by making the rows columns and vice versa.

There are several ways to think of vectors. In the preceeding chapters of these notes we wrote vectors as tuples $\mathbf{x} = (x_1, \ldots, x_n)$. Now we will also

consider vectors as special cases of matrices. A *column vector* is an $n \times 1$ matrix

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \tag{5.2}$$

and a *row vector* is a $1 \times n$ matrix

$$\mathbf{x}' = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix} \tag{5.3}$$

Note that (5.2) is indeed the transpose of (5.3) as the notation $\mathbf{x}$ and $\mathbf{x}'$ indicates. Note that even when we consider vectors as special matrices we still use boldface lower case letters for nonrandom vectors, as we always have, rather than the boldface capital letters we use for matrices.

5.1.2 Random Vectors

A *random vector* is just a vector whose components are random scalars. We have always denoted random vectors using boldface capital letters $\mathbf{X} = (X_1, \ldots, X_n)$, which conflicts with the new convention that matrices are boldface capital letters. So when you see a boldface capital letter, you must decide whether this indicates a random vector or a constant (nonrandom) matrix. One hint is that we usually use letters like $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ for random vectors, and we will usually use letters earlier in the alphabet for matrices. If you are not sure what is meant by this notation (or any notation), look at the context, it should be defined nearby.

The *expectation* or *mean* of a random vector $\mathbf{X} = (X_1, \ldots, X_n)$ is defined componentwise. The mean of $\mathbf{X}$ is the vector

$$\boldsymbol{\mu}_{\mathbf{X}} = E(\mathbf{X}) = \big(E(X_1), \ldots, E(X_n)\big)$$

having components that are the expectations of the corresponding components of $\mathbf{X}$.

5.1.3 Random Matrices

Similarly, we define *random matrix* to be a matrix whose components are random scalars. Let $\mathbf{X}$ denote a random matrix with elements X_{ij}. We can see that the boldface and capital letter conventions have now pooped out. There is no "double bold" or "double capital" type face to indicate the difference between a random vector and a random matrix.[1] The reader will just have to remember in this section $\mathbf{X}$ is a matrix not a vector.

[1] This is one reason to avoid the "vectors are bold" and "random objects are capitals" conventions. They violate "mathematics is invariant under changes of notation." The type face conventions work in simple situations, but in complicated situations they are part of the problem rather than part of the solution. That's why modern advanced mathematics doesn't use the "vectors are bold" convention. It's nineteenth century notation still surviving in statistics.

Again like random vectors, the *expectation* or *mean* of a random matrix is a nonrandom matrix. If $\mathbf{X}$ is a random $m \times n$ matrix with elements X_{ij}, then the mean of $\mathbf{X}$ is the matrix $\mathbf{M}$ with elements

$$\mu_{ij} = E(X_{ij}), \tag{5.4}$$

and we also write $E(\mathbf{X}) = \mathbf{M}$ to indicate all of the mn equations (5.4) with one matrix equation.

5.1.4 Variance Matrices

In the preceding two sections we defined random vectors and random matrices and their expectations. The next topic is variances. One might think that the variance of a random vector should be similar to the mean, a vector having components that are the variances of the corresponding components of $\mathbf{X}$, but it turns out that this notion is not useful. The reason is that variances and covariances are inextricably entangled. We see this in the fact that the variance of a sum involves both variances and covariances (Corollary 2.19 of these notes and the following comments). Thus the following definition.

The *variance matrix* of an n-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_n)$ is the nonrandom $n \times n$ matrix $\mathbf{M}$ having elements

$$m_{ij} = \operatorname{cov}(X_i, X_j). \tag{5.5}$$

As with variances of random scalars, we also use the notation $\operatorname{var}(\mathbf{X})$ for the variance matrix. Note that the diagonal elements of $\mathbf{M}$ are variances because the covariance of a random scalar with itself is the variance, that is,

$$m_{ii} = \operatorname{cov}(X_i, X_i) = \operatorname{var}(X_i).$$

This concept is well established, but the name is not. Lindgren calls $\mathbf{M}$ the *covariance matrix* of $\mathbf{X}$, presumably because its elements are covariances. Other authors call it the *variance-covariance* matrix, because some of its elements are variances too. Some authors, to avoid the confusion about variance, covariance, or variance-covariance, call it the *dispersion* matrix. In my humble opinion, "variance matrix" is the right name because it is the generalization of the variance of a scalar random variable. But you're entitled to call it what you like. There is no standard terminology.

Example 5.1.1.
What are the mean vector and variance matrix of the random vector (X, X^2), where X is some random scalar? Let

$$\alpha_k = E(X^k)$$

denote the ordinary moments of X. Then, of course, the mean and variance of X are $\mu = \alpha_1$ and

$$\sigma^2 = E(X^2) - E(X)^2 = \alpha_2 - \alpha_1^2,$$

but it will be simpler if we stick to the notation using the α's. The mean vector is

$$\boldsymbol{\mu} = \begin{pmatrix} E(X) \\ E(X^2) \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \tag{5.6}$$

The moment matrix is the 2×2 matrix $\mathbf{M}$ with elements

$$\begin{aligned} m_{11} &= \operatorname{var}(X) \\ &= \alpha_2 - \alpha_1^2 \\ m_{22} &= \operatorname{var}(X^2) \\ &= E(X^4) - E(X^2)^2 \\ &= \alpha_4 - \alpha_2^2 \\ m_{12} &= \operatorname{cov}(X, X^2) \\ &= E(X^3) - E(X)E(X^2) \\ &= \alpha_3 - \alpha_1\alpha_2 \end{aligned}$$

Putting this all together we get

$$\mathbf{M} = \begin{pmatrix} \alpha_2 - \alpha_1^2 & \alpha_3 - \alpha_1\alpha_2 \\ \alpha_3 - \alpha_1\alpha_2 & \alpha_4 - \alpha_2^2 \end{pmatrix} \tag{5.7}$$

5.1.5 What is the Variance of a Random Matrix?

By analogy with random vectors, the variance of $\mathbf{X}$ should be a mathematical object with four indexes, the elements being

$$v_{ijkl} = \operatorname{cov}(X_{ij}, X_{kl}).$$

Even naming such an object takes outside the realm of linear algebra. One terminology for objects with more than two indices is *tensors*. So we can say that the variance of a random matrix is a nonrandom tensor. But this doesn't get us anywhere because we don't know anything about operations that apply to tensors.

Thus we see that random matrices present no problem so long as we only are interested in their means, but their variances are problematical. Fortunately, we can avoid random matrices except when we are interested only in their means, not their variances.[2]

[2] A solution to the problem of defining the variance of a random matrix that avoids tensors is to change notation and consider the random matrix a random vector. For example, a random $m \times n$ matrix $\mathbf{X}$ can be written as a vector

$$\mathbf{Y} = (X_{11}, X_{12}, \ldots X_{1n}, X_{21}, X_{22}, \ldots, X_{2n}, \ldots X_{m1}, X_{m2}, \ldots X_{mn})$$

So $Y_1 = X_{11}$, $Y_2 = X_{12}$, ..., $Y_{n+1} = X_{2n}$, and so forth. Now there is no problem defining the variance matrix of $\mathbf{Y}$, but this is unnatural and clumsy notation that will in most problems make things exceedingly messy.

5.1.6 Covariance Matrices

The *covariance matrix* of an m-dimensional random vector $\mathbf{X}$ and an n-dimensional random vector $\mathbf{Y}$ is the nonrandom matrix $\mathbf{C}$ with elements

$$c_{ij} = \operatorname{cov}(X_i, Y_j), \tag{5.8}$$

(where, as usual X_i is an element of $\mathbf{X}$ and Y_j an element of $\mathbf{Y}$). Note that if $\mathbf{X}$ is an m-dimensional vector and $\mathbf{Y}$ is an n-dimensional vector, then $\mathbf{C} = \operatorname{cov}(\mathbf{X}, \mathbf{Y})$ is an $m \times n$ matrix. Swapping the roles of $\mathbf{X}$ and $\mathbf{Y}$ we see that $\operatorname{cov}(\mathbf{Y}, \mathbf{X})$ is an $n \times m$ matrix. Thus it is obvious that the property $\operatorname{cov}(X, Y) = \operatorname{cov}(Y, X)$ that holds for covariances of scalar random variables, does not hold for covariances of random vectors. In fact, if we write

$$\begin{aligned} \mathbf{C} &= \operatorname{cov}(\mathbf{X}, \mathbf{Y}) \\ \mathbf{D} &= \operatorname{cov}(\mathbf{Y}, \mathbf{X}), \end{aligned}$$

then the elements of $\mathbf{C}$ are given by (5.8) and the elements of $\mathbf{D}$ are

$$d_{ij} = \operatorname{cov}(Y_i, X_j) = c_{ji}$$

Thus the two matrices are transposes of each other: $\mathbf{D} = \mathbf{C}'$.

With these definitions, we can easily generalize most of the formulas about variances and covariances of scalar random variables to vector random variables. We won't bother to go through all of them. The most important one is the formula for the variance of a sum of random vectors.

$$\operatorname{var}\left(\sum_{i=1}^{n} \mathbf{X}_i\right) = \sum_{i=1}^{n} \sum_{j=1}^{n} \operatorname{cov}(\mathbf{X}_i, \mathbf{X}_j) \tag{5.9}$$

which is the same as Corollary 2.19, except that it applies to vector random variables in place of scalar ones. The special case in which $\mathbf{X}_1$, ..., $\mathbf{X}_n$ are *uncorrelated* random vectors, meaning $\operatorname{cov}(\mathbf{X}_i, \mathbf{X}_j) = 0$ when $i \neq j$, gives

$$\operatorname{var}\left(\sum_{i=1}^{n} \mathbf{X}_i\right) = \sum_{i=1}^{n} \operatorname{var}(\mathbf{X}_i) \tag{5.10}$$

that is, the variance of the sum is the sum of the variances, which is the same as Corollary 2.21, except that it applies to vector random variables in place of scalar ones.

As with random scalars, independence implies lack of correlation, because $\mathbf{C} = \operatorname{cov}(\mathbf{X}, \mathbf{Y})$ has elements $c_{ij} = \operatorname{cov}(X_i, Y_j)$ which are all zero by this property for random scalars (Theorem 2.47). Hence (5.10) also holds when $\mathbf{X}_1$, ..., $\mathbf{X}_n$ are *independent* random vectors. This is by far the most important application of (5.10). As in the scalar case, you should remember

> *Independent implies uncorrelated, but uncorrelated does* ***not*** *imply independent.*

Thus independence is a sufficient but not necessary condition for (5.10) to hold. It is enough that the variables be uncorrelated.

In statistics, our main interest is not in sums per se but rather in averages

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i. \tag{5.11a}$$

The analogous formula for random vectors is just the same formula with boldface

$$\overline{\mathbf{X}}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i. \tag{5.11b}$$

Warning: the subscripts on the right hand side in (5.11b) do not indicate components of a vector, rather $\mathbf{X}_1$, $\mathbf{X}_2$, ... is simply a sequence of random vectors just as in (5.11a) X_1, X_2, ... is a sequence of random scalars. The formulas for the mean and variance of a sum also give us the mean and variance of an average.

Theorem 5.1. *If $\mathbf{X}_1$, $\mathbf{X}_2$, ... are random vectors having the same mean vector $\boldsymbol{\mu}$, then*

$$E(\overline{\mathbf{X}}_n) = \boldsymbol{\mu}. \tag{5.12a}$$

If $\mathbf{X}_1$, $\mathbf{X}_2$, ... also have the same variance matrix $\mathbf{M}$ and are uncorrelated, then

$$\operatorname{var}(\overline{\mathbf{X}}_n) = \frac{1}{n}\mathbf{M}. \tag{5.12b}$$

This is exactly analogous to the scalar case

$$E(\overline{X}_n) = \mu \tag{5.13a}$$

and

$$\operatorname{var}(\overline{X}_n) = \frac{\sigma^2}{n} \tag{5.13b}$$

Theorem 5.2 (Alternate Variance and Covariance Formulas). *If $\mathbf{X}$ and $\mathbf{Y}$ are random vectors with means $\boldsymbol{\mu}_{\mathbf{X}}$ and $\boldsymbol{\mu}_{\mathbf{Y}}$, then*

$$\operatorname{cov}(\mathbf{X}, \mathbf{Y}) = E\{(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}})'\} \tag{5.14a}$$

$$\operatorname{var}(\mathbf{X}) = E\{(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}})'\} \tag{5.14b}$$

This hardly deserves the name "theorem" since it is obvious once one interprets the matrix notation. If $\mathbf{X}$ is m-dimensional and $\mathbf{Y}$ is n-dimensional, then when we consider the vectors as matrices ("column vectors") we see that the dimensions are

$$\underset{m \times 1}{(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})} \quad \underset{1 \times n}{(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}})'}$$

so the "sum" implicit in the matrix multiplication has only one term. Thus (5.14a) is the $m \times n$ matrix with i, j element

$$E\{(X_i - \mu_{X_i})(Y_j - \mu_{Y_j})\} = \operatorname{cov}(X_i, Y_j)$$

and hence is the covariance matrix $\operatorname{cov}(\mathbf{X}, \mathbf{Y})$. Then we see that (5.14b) is just the special case where $\mathbf{Y} = \mathbf{X}$.

5.1.7 Linear Transformations

In this section, we derive the analogs of the formulas

$$E(a + bX) = a + bE(X) \tag{5.15a}$$
$$\operatorname{var}(a + bX) = b^2 \operatorname{var}(X) \tag{5.15b}$$

(Corollary 2.2 and Theorem 2.13 in Chapter 2 of these notes) that describe the moments of a linear transformation of a random variable. A general linear transformation has the form

$$\mathbf{y} = \mathbf{a} + \mathbf{B}\mathbf{x}$$

where $\mathbf{y}$ and $\mathbf{a}$ are m-dimensional vectors, $\mathbf{B}$ is an $m \times n$ matrix, and $\mathbf{x}$ is an n-dimensional vector. The dimensions of each object, considering the vectors as column vectors (that is, as matrices with just a single column), are

$$\begin{array}{ccccccc} \mathbf{y} & = & \mathbf{a} & + & \mathbf{B} & \mathbf{a} \\ m \times 1 & & m \times 1 & & m \times n & n \times 1 \end{array} \tag{5.16}$$

Note that the column dimension of $\mathbf{B}$ and the row dimension of $\mathbf{x}$ must agree, as in any matrix multiplication. Also note that the dimensions of $\mathbf{x}$ and $\mathbf{y}$ are not the same. We are mapping n-dimensional vectors to m-dimensional vectors.

Theorem 5.3. *If $\mathbf{Y} = \mathbf{a} + \mathbf{B}\mathbf{X}$, where $\mathbf{a}$ is a constant vector, $\mathbf{B}$ is a constant matrix, and $\mathbf{X}$ is a random vector, then*

$$E(\mathbf{Y}) = \mathbf{a} + \mathbf{B}E(\mathbf{X}) \tag{5.17a}$$
$$\operatorname{var}(\mathbf{Y}) = \mathbf{B} \operatorname{var}(\mathbf{X}) \mathbf{B}' \tag{5.17b}$$

If we write $\boldsymbol{\mu}_{\mathbf{X}}$ and $\mathbf{M}_{\mathbf{X}}$ for the mean and variance of $\mathbf{X}$ and similarly for $\mathbf{Y}$, then (5.17a) and (5.17b) become

$$\boldsymbol{\mu}_{\mathbf{Y}} = \mathbf{a} + \mathbf{B}\boldsymbol{\mu}_{\mathbf{X}} \tag{5.18a}$$
$$\mathbf{M}_{\mathbf{Y}} = \mathbf{B}\mathbf{M}_{\mathbf{X}}\mathbf{B}' \tag{5.18b}$$

If we were to add dimension information to (5.18a), it would look much like (5.16). If we add such information to (5.18b) it becomes

$$\begin{array}{ccccccc} \mathbf{M}_{\mathbf{Y}} & = & \mathbf{B} & \mathbf{M}_{\mathbf{X}} & \mathbf{B}' \\ m \times m & & m \times n & n \times n & n \times m \end{array}$$

Note again that, as in any matrix multiplication, the column dimension of the left hand factor agrees with row dimension of the right hand factor. In particular, the column dimension of $\mathbf{B}$ is the row dimension of $\mathbf{M}_{\mathbf{X}}$, and the column dimension of $\mathbf{M}_{\mathbf{X}}$ is the row dimension of $\mathbf{B}'$. Indeed, this is the only way these matrices can be multiplied together to get a result of the appropriate dimension. So merely getting the dimensions right tells you what the formula has to be.

Proof of Theorem 5.3. Since our only definition of the mean of a random vector involves components, we will have to prove this componentwise. The component equations of $\mathbf{Y} = \mathbf{a} + \mathbf{B}\mathbf{X}$ are

$$Y_i = a_i + \sum_{j=1}^{n} b_{ij} X_j$$

(where, as usual, the a_i are the components of $\mathbf{a}$, the b_{ij} are the components of $\mathbf{B}$, and so forth). Applying linearity of expectation for scalars gives

$$E(Y_i) = a_i + \sum_{j=1}^{n} b_{ij} E(X_j),$$

which are the component equations of (5.18a).

Now we can be a bit slicker about the second half of the proof using the alternate variance formula (5.14b).

$$\begin{aligned} \operatorname{var}(\mathbf{a} + \mathbf{B}\mathbf{X}) &= E\{(\mathbf{a} + \mathbf{B}\mathbf{X} - \boldsymbol{\mu}_{\mathbf{a}+\mathbf{B}\mathbf{X}})(\mathbf{a} + \mathbf{B}\mathbf{X} - \boldsymbol{\mu}_{\mathbf{a}+\mathbf{B}\mathbf{X}})'\} \\ &= E\{(\mathbf{B}\mathbf{X} - \mathbf{B}\boldsymbol{\mu}_{\mathbf{X}})(\mathbf{B}\mathbf{X} - \mathbf{B}\boldsymbol{\mu}_{\mathbf{X}})'\} \\ &= E\{\mathbf{B}(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})'\mathbf{B}'\} \\ &= \mathbf{B} E\{(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})'\}\mathbf{B}' \end{aligned}$$

Going from the first line to the second is just (5.18a). Going from the second line to the third uses the fact that the transpose of a matrix product is the product of the transposes in reverse order, that is, $(\mathbf{B}\mathbf{C})' = \mathbf{C}'\mathbf{B}$. And going from the third line to the forth uses (5.18a) again to pull the constant matrices outside the expectation. □

Of particular interest is the special case in which the linear transformation is scalar-valued, that is, $m = 1$ in (5.16). Then the matrix $\mathbf{B}$ must be $1 \times n$, hence a row vector. We usually write row vectors as transposes, say $\mathbf{c}'$, because convention requires unadorned vectors like $\mathbf{c}$ to be column vectors. Thus we write $\mathbf{B} = \mathbf{c}'$ and obtain

Corollary 5.4. *If $Y = a + \mathbf{c}'\mathbf{X}$, where a is a constant scalar, $\mathbf{c}$ is a constant vector, and $\mathbf{X}$ is a random vector, then*

$$E(Y) = a + \mathbf{c}' E(\mathbf{X}) \tag{5.19a}$$

$$\operatorname{var}(Y) = \mathbf{c}' \operatorname{var}(\mathbf{X}) \mathbf{c} \tag{5.19b}$$

Or, if you prefer the other notation

$$\mu_Y = a + \mathbf{c}'\boldsymbol{\mu}_{\mathbf{X}} \tag{5.20a}$$

$$\sigma_Y^2 = \mathbf{c}'\mathbf{M}_{\mathbf{X}}\mathbf{c} \tag{5.20b}$$

Note that, since $m = 1$, both Y and a are scalars (1×1 matrices), so we have written them in normal (not boldface) type and used the usual notation σ_Y^2 for the variance of a scalar. Also note that because $\mathbf{B} = \mathbf{c}'$ the transposes have switched sides in going from (5.18b) to (5.20b).

Example 5.1.2.
(This continues Example 5.1.1.) What are the mean and variance of $X + X^2$, where X is some random scalar? We don't have to use an multivariate theory to answer this question. We could just use the formulas for the mean and variance of a sum of random variables from Chapter 2 of these notes. But here we want to use the multivariate theory to illustrate how it works.

Write $Y = X + X^2$ and let

$$\mathbf{Z} = \begin{pmatrix} X \\ X^2 \end{pmatrix}$$

be the random vector whose mean vector and variance matrix were found in Example 5.1.1. Then $Y = \mathbf{u}'\mathbf{Z}$, where

$$\mathbf{u} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Thus by (5.20a) and (5.6)

$$E(Y) = \mathbf{u}'\boldsymbol{\mu}_{\mathbf{Z}} = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \alpha_1 + \alpha_2$$

And by (5.20b) and (5.7)

$$\begin{aligned} \operatorname{var}(Y) &= \mathbf{u}'\mathbf{M}_{\mathbf{Z}}\mathbf{u} \\ &= \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} \alpha_2 - \alpha_1^2 & \alpha_3 - \alpha_1\alpha_2 \\ \alpha_3 - \alpha_1\alpha_2 & \alpha_4 - \alpha_2^2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ &= \alpha_2 - \alpha_1^2 + 2(\alpha_3 - \alpha_1\alpha_2) + \alpha_4 - \alpha_2^2 \end{aligned}$$

Alternate Solution We could also do this problem the "old fashioned way" (without matrices)

$$\begin{aligned} \operatorname{var}(X + X^2) &= \operatorname{var}(X) + 2\operatorname{cov}(X, X^2) + \operatorname{var}(X^2) \\ &= (\alpha_2 - \alpha_1^2) + 2(\alpha_3 - \alpha_1\alpha_2) + (\alpha_4 - \alpha_2^2) \end{aligned}$$

Of course, both ways must give the same answer. We're just using matrices here to illustrate the use of matrices.

5.1.8 Characterization of Variance Matrices

A matrix $\mathbf{A}$ is said to be *positive semi-definite* if

$$\mathbf{c}'\mathbf{A}\mathbf{c} \geq 0, \qquad \text{for every vector } \mathbf{c} \tag{5.21}$$

and *positive definite* if

$$\mathbf{c}'\mathbf{A}\mathbf{c} > 0, \qquad \text{for every nonzero vector } \mathbf{c}.$$

Corollary 5.5. *The variance matrix of any random vector is symmetric and positive semi-definite.*

Proof. Symmetry follows from the symmetry property of covariances of random scalars: $\operatorname{cov}(X_i, X_j) = \operatorname{cov}(X_j, X_i)$.

The random scalar Y in Corollary 5.4 must have nonnegative variance. Thus (5.19b) implies $\mathbf{c}' \operatorname{var}(\mathbf{X})\mathbf{c} \ge 0$. Since $\mathbf{c}$ was an arbitrary vector, this proves $\operatorname{var}(\mathbf{X})$ is positive semi-definite. □

The corollary says that a *necessary condition* for a matrix $\mathbf{M}$ to be the variance matrix of some random vector $\mathbf{X}$ is that $\mathbf{M}$ be symmetric and positive semi-definite. This raises the obvious question: is this a *sufficient condition*, that is, for any symmetric and positive semi-definite matrix $\mathbf{M}$ does there exist a random vector $\mathbf{X}$ such that $\mathbf{M} = \operatorname{var}(\mathbf{X})$? We can't address this question now, because we don't have enough examples of random vectors for which we know the distributions. It will turn out that the answer to the sufficiency question is "yes." When we come to the multivariate normal distribution (Section 5.2) we will see that for any symmetric and positive semi-definite matrix $\mathbf{M}$ there is a multivariate normal random vector $\mathbf{X}$ such that $\mathbf{M} = \operatorname{var}(\mathbf{X})$.

A *hyperplane* in n-dimensional space $\mathbb{R}^n$ is a set of the form

$$H = \{\, \mathbf{x} \in \mathbb{R}^n : \mathbf{c}'\mathbf{x} = a \,\} \tag{5.22}$$

for some nonzero vector $\mathbf{c}$ and some scalar a. We say a random vector $\mathbf{X}$ is *concentrated* on the hyperplane H if $P(\mathbf{X} \in H) = 1$. Another way of describing the same phenomenon is to say that that H is a *support* of $\mathbf{X}$.

Corollary 5.6. *The variance matrix of a random vector $\mathbf{X}$ is positive definite if and only if $\mathbf{X}$ is not concentrated on any hyperplane.*

Proof. We will prove the equivalent statement that the variance matrix is *not* positive definite if and only if is *is* concentrated on some hyperplane.

First, suppose that $\mathbf{M} = \operatorname{var}(\mathbf{X})$ is not positive definite. Then there is some nonzero vector $\mathbf{c}$ such that $\mathbf{c}'\mathbf{M}\mathbf{c} = 0$. Consider the random scalar $Y = \mathbf{c}'\mathbf{X}$. By Corollary 5.4 $\operatorname{var}(Y) = \mathbf{c}'\mathbf{M}\mathbf{c} = 0$. Now by Corollary 2.34 of these notes $Y = \mu_Y$ with probability one. Since $E(Y) = \mathbf{c}'\mu_{\mathbf{X}}$ by (5.19a), this says that $\mathbf{X}$ is concentrated on the hyperplane (5.22) where $a = \mathbf{c}'\mu_{\mathbf{X}}$.

Conversely, suppose that $\mathbf{X}$ is concentrated on the hyperplane (5.22). Then the random scalar $Y = \mathbf{c}'\mathbf{x}$ is concentrated at the point a, and hence has variance zero, which is $\mathbf{c}'\mathbf{M}\mathbf{c}$ by Corollary 5.4. Thus $\mathbf{M}$ is not positive definite. □

5.1.9 Degenerate Random Vectors

Random vectors are sometimes called *degenerate* by those who believe in the kindergarten principle of calling things we don't like bad names. And why wouldn't we like a random vector concentrated on a hyperplane? Because it doesn't have a density. A hyperplane is a set of measure zero, hence any integral over the hyperplane is zero and cannot be used to define probabilities and expectations.

Example 5.1.3 (A Degenerate Random Vector).
Suppose U, V, and W are independent and identically distributed random variables having a distribution not concentrated at one point, so $\sigma^2 = \operatorname{var}(U) = \operatorname{var}(V) = \operatorname{var}(W)$ is strictly positive. Consider the random vector

$$\mathbf{X} = \begin{pmatrix} U - V \\ V - W \\ W - U \end{pmatrix} \tag{5.23}$$

Because of the assumed independence of U, V, and W, the diagonal elements of $\operatorname{var}(\mathbf{X})$ are all equal to

$$\operatorname{var}(U - V) = \operatorname{var}(U) + \operatorname{var}(V) = 2\sigma^2$$

and the off-diagonal elements are all equal to

$$\operatorname{cov}(U - V, V - W) = \operatorname{cov}(U, V) - \operatorname{cov}(U, W) - \operatorname{var}(V) + \operatorname{cov}(V, W) = -\sigma^2$$

Thus

$$\operatorname{var}(\mathbf{X}) = \sigma^2 \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}$$

Question Is $\mathbf{X}$ degenerate or non-degenerate? If degenerate, what hyperplane or hyperplanes is it concentrated on?

Answer We give two different ways of finding this out. The first uses some mathematical cleverness, the second brute force and ignorance (also called plug and chug).

The first way starts with the observation that each of the variables U, V, and W occurs twice in the components of $\mathbf{X}$, once with each sign, so the sum of the components of $\mathbf{X}$ is zero, that is $X_1 + X_2 + X_3 = 0$ with probability one. But if we introduce the vector

$$\mathbf{u} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

we see that $X_1 + X_2 + X_3 = \mathbf{u}'\mathbf{X}$. Hence $\mathbf{X}$ is concentrated on the hyperplane defined by

$$H = \{\, \mathbf{x} \in \mathbb{R}^3 : \mathbf{u}'\mathbf{x} = 0 \,\}$$

or if you prefer

$$H = \{\, (x_1, x_2, x_3) \in \mathbb{R}^3 : x_1 + x_2 + x_3 = 0 \,\}.$$

Thus we see that $\mathbf{X}$ is indeed degenerate (concentrated on H). Is is concentrated on any other hyperplanes? The answer is no, but our cleverness has run out. It's hard so show that there are no more except by the brute force approach.

The brute force approach is to find the eigenvalues and eigenvectors of the variance matrix. The random vector in question is concentrated on hyperplanes defined by eigenvectors corresponding to zero eigenvalues (Lemma 5.7 below). Eigenvalues and eigenvectors can be found by many numerical math packages. Here we will just demonstrate doing it in R.

```
> M <- matrix(c(2, -1, -1, -1, 2, -1, -1, -1, 2), nrow=3)
> M
     [,1] [,2] [,3]
[1,]    2   -1   -1
[2,]   -1    2   -1
[3,]   -1   -1    2
> eigen(M)
$values
[1]  3.000000e+00  3.000000e+00 -8.881784e-16

$vectors
           [,1]       [,2]      [,3]
[1,]  0.8156595  0.0369637 0.5773503
[2,] -0.3758182 -0.7248637 0.5773503
[3,] -0.4398412  0.6879000 0.5773503
```

Each eigenvector corresponding to a zero eigenvalue is a vector $\mathbf{c}$ defining a hyperplane by (5.22) on which the random vector is concentrated. There is just one zero eigenvalue. The corresponding eigenvector is

$$\mathbf{c} = \begin{pmatrix} 0.5773503 \\ 0.5773503 \\ 0.5773503 \end{pmatrix}$$

(the eigenvectors are the columns of the `$vectors` matrix returned by the `eigen` function). Since $\mathbf{c}$ is a multiple of $\mathbf{u}$ in the first answer, they define the same hyperplane. Since there is only one zero eigenvalue, there is only one hyperplane supporting the random vector.

Lemma 5.7. *A random vector* $\mathbf{X}$ *is concentrated on a hyperplane* (5.22) *if and only if the vector* $\mathbf{c}$ *in* (5.22) *is an eigenvector of* $\operatorname{var}(\mathbf{X})$ *corresponding to a zero eigenvalue.*

Proof. First suppose $\mathbf{c}$ is an eigenvector of $\mathbf{M} = \operatorname{var}(\mathbf{X})$ corresponding to a zero eigenvalue. This means $\mathbf{Mc} = 0$, which implies $\mathbf{c}'\mathbf{Mc} = 0$, which, as in the proof of Corollary 5.6, implies that $\mathbf{X}$ is concentrated on the hyperplane defined by (5.22).

Conversely, suppose $\mathbf{X}$ is concentrated on the hyperplane defined by (5.22), which, as in the proof of Corollary 5.6, implies $\mathbf{c}'\mathbf{Mc} = 0$. Write, using the spectral decomposition (Theorem E.4 in Appendix E) $\mathbf{M} = \mathbf{ODO}'$, where $\mathbf{D}$ is diagonal and $\mathbf{O}$ is orthogonal. Then

$$0 = \mathbf{c}'\mathbf{Mc} = \mathbf{c}'\mathbf{ODO}'\mathbf{c} = \mathbf{w}'\mathbf{Dw}$$

where we have written $\mathbf{w} = \mathbf{O}'\mathbf{c}$. Writing out the matrix multiplications with subscripts

$$\mathbf{w}'\mathbf{D}\mathbf{w} = \sum_i d_{ii} w_i^2 = 0$$

which implies, since $d_{ii} \geq 0$ for all i that

$$d_{ii} = 0 \quad \text{or} \quad w_i = 0, \qquad \text{for all } i$$

and this implies that actually $\mathbf{D}\mathbf{w} = 0$. Hence, plugging back in the definition of $\mathbf{w}$, that $\mathbf{D}\mathbf{O}'\mathbf{c} = 0$, and, multiplying on the left by $\mathbf{O}$, that

$$\mathbf{M}\mathbf{c} = \mathbf{O}\mathbf{D}\mathbf{O}'\mathbf{c} = 0$$

which says that $\mathbf{c}$ is an eigenvector of $\mathbf{M}$ corresponding to a zero eigenvalue, which is what we were proving. □

Degeneracy is not solely a phenomenon of concentration on hyperplanes. We say a random vector is degenerate if it is concentrated on any set of measure zero.

Example 5.1.4.
In Example 2.7.2 we considered the random vector $\mathbf{Z} = (X, Y)$, where $Y = X^2$ and X was any nonconstant random variable having a distribution symmetric about zero. It served there as an example of random variables X and Y that were uncorrelated but not independent.

Here we merely point out that the random vector $\mathbf{Z}$ is degenerate, because it is clearly concentrated on the parabola

$$S = \{\, (x, y) \in \mathbb{R}^2 : y = x^2 \,\}$$

which is, being a one-dimensional curve in $\mathbb{R}^2$, a set of measure zero.

So how does one handle degenerate random vectors? If they don't have densities, and most of the methods we know involve densities, what do we do? First let me remind you that we do know some useful methods that don't involve densities.

- The first part of Chapter 2 of these notes, through Section 2.4 never mentions densities. The same goes for Sections 3.3 and 3.5 in Chapter 3.

- In order to calculate $E(\mathbf{Y})$ where $\mathbf{Y} = \mathbf{g}(\mathbf{X})$, you don't need the density of $\mathbf{Y}$. You can use

  $$E(\mathbf{Y}) = \int \mathbf{g}(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}$$

  instead. Thus even if $\mathbf{Y}$ is degenerate, but is a known function of some non-degenerate random vector $\mathbf{X}$, we are still in business.

When a random vector $\mathbf{X}$ is degenerate, it is always possible in theory (not necessarily in practice) to eliminate one of the variables. For example, if $\mathbf{X}$ is concentrated on the hyperplane H defined by (5.22), then, since $\mathbf{c}$ is nonzero, it has at least one nonzero component, say c_j. Then rewriting $\mathbf{c}'\mathbf{x} = a$ with an explicit sum we get

$$\sum_{i=1}^{n} c_i X_i = a,$$

which can be solved for X_j

$$X_j = \frac{1}{c_j} \left(a - \sum_{\substack{i=1 \\ i \neq j}}^{n} c_i X_i \right)$$

Thus we can eliminate X_j and work with the remaining variables. If the random vector

$$\mathbf{X}' = (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n)$$

of the remaining variables is non-degenerate, then it has a density. If $\mathbf{X}'$ is still degenerate, then there is another variable we can eliminate. Eventually, unless $\mathbf{X}$ is a constant random vector, we get to some subset of variables that have a non-degenerate joint distribution and hence a density. Since the rest of the variables are a function of this subset, that indirectly describes all the variables.

Example 5.1.5 (Example 5.1.3 Continued).
In Example 5.1.3 we considered the random vector

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} U - V \\ V - W \\ W - U \end{pmatrix}$$

where U, V, and W are independent and identically distributed random variables. Now suppose they are independent standard normal.

In Example 5.1.3 we saw that $\mathbf{X}$ was degenerate because $X_1 + X_2 + X_3 = 0$ with probability one. We can eliminate X_3, since

$$X_3 = -(X_1 + X_2)$$

and consider the distribution of the vector (X_1, X_2), which we will see (in Section 5.2 below) has a non-degenerate multivariate normal distribution.

5.1.10 Correlation Matrices

If $\mathbf{X} = (X_1, \ldots, X_n)$ is a random vector having no constant components, that is, $\operatorname{var}(X_i) > 0$ for all i, the *correlation matrix* of $\mathbf{X}$ is the $n \times n$ matrix $\mathbf{C}$ with elements

$$c_{ij} = \frac{\operatorname{cov}(X_i, X_j)}{\sqrt{\operatorname{var}(X_i) \operatorname{var}(X_i)}} = \operatorname{cor}(X_i, X_j)$$

If $\mathbf{M} = \text{var}(\mathbf{X})$ has elements m_{ij}, then

$$c_{ij} = \frac{m_{ij}}{\sqrt{m_{ii} m_{jj}}}$$

Note that the diagonal elements c_{ii} of a correlation matrix are all equal to one, because the correlation of any random variable with itself is one.

Theorem 5.8. *Every correlation matrix is positive semi-definite. The correlation matrix of a random vector* $\mathbf{X}$ *is positive definite if and only the variance matrix of* $\mathbf{X}$ *is positive definite.*

Proof. This follows from the analogous facts about variance matrices. □

It is important to understand that the requirement that a variance matrix (or correlation matrix) be positive semi-definite is a much stronger requirement than the correlation inequality (correlations must be between -1 and $+1$). The two requirements are related: positive semi-definiteness implies the correlation inequality, but not vice versa. That positive semi-definiteness implies the correlation inequality is left as an exercise (Problem 5-4). That the two conditions are not equivalent is shown by the following example.

Example 5.1.6 (All Correlations the Same).
Suppose $\mathbf{X} = (X_1, \ldots, X_n)$ is a random vector and all the components have the same correlation, as would be the case if the components are *exchangeable* random variables, that is, $\text{cor}(X_i, X_j) = \rho$ for all i and j with $i \neq j$. Then the correlation matrix of $\mathbf{X}$ has one for all diagonal elements and ρ for all off-diagonal elements. In Problem 2-22 it is shown that positive definiteness of the correlation matrix requires

$$-\frac{1}{n-1} \leq \rho.$$

This is an additional inequality not implied by the correlation inequality.

The example says there is a limit to how negatively correlated a sequence of exchangeable random variables can be. But even more important than this specific discovery, is the general message that there is more to know about correlations than that they are always between -1 and $+1$. The requirement that a correlation matrix (or a variance matrix) be positive semi-definite is much stronger. It implies a lot of other inequalities. In fact it implies an infinite family of inequalities: $\mathbf{M}$ is positive semi-definite only if $\mathbf{c}'\mathbf{M}\mathbf{c} \geq 0$ for every vector $\mathbf{c}$. That's a different inequality for every vector $\mathbf{c}$ and there are infinitely many such vectors.

5.2 The Multivariate Normal Distribution

The *standard multivariate normal distribution* is the distribution of the random vector $\mathbf{Z} = (Z_1, \ldots, Z_n)$ having independent and identically $\mathcal{N}(0, 1)$ dis-

tributed components. Its density is, of course,

$$f_{\mathbf{Z}}(\mathbf{z}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-z_i^2/2} = \frac{1}{(2\pi)^{n/2}} e^{-\mathbf{z}'\mathbf{z}/2}, \qquad \mathbf{z} \in \mathbb{R}^n$$

Note for future reference that

$$\begin{aligned} E(\mathbf{Z}) &= 0 \\ \operatorname{var}(\mathbf{Z}) &= \mathbf{I} \end{aligned}$$

where $\mathbf{I}$ denotes an identity matrix. These are obvious from the definition of $\mathbf{Z}$. Its components are independent and standard normal, hence have mean zero, variance one, and covariances zero. Thus the variance matrix has ones on the diagonal and zeroes off the diagonal, which makes it an identity matrix.

As in the univariate case, we call a linear transformation of a standard normal random vector a (general) normal random vector. If we define $\mathbf{X} = \mathbf{a} + \mathbf{B}\mathbf{Z}$, then by Theorem 5.3

$$\begin{aligned} E(\mathbf{X}) &= \mathbf{a} + \mathbf{B}E(\mathbf{Z}) \\ &= \mathbf{a} \\ \operatorname{var}(\mathbf{X}) &= \mathbf{B} \operatorname{var}(\mathbf{Z}) \mathbf{B}' \\ &= \mathbf{B}\mathbf{B}' \end{aligned}$$

We say that $\mathbf{X}$ has the multivariate normal distribution with mean (vector) $\mathbf{a}$ and variance (matrix) $\mathbf{M} = \mathbf{B}\mathbf{B}'$, and abbreviate it as $\mathcal{N}_n(\mathbf{a}, \mathbf{M})$ if we want to emphasize the dimension n of the random vector, or just as $\mathcal{N}(\mathbf{a}, \mathbf{M})$ if we don't want to explicitly note the dimension. No confusion should arise between the univariate and multivariate case, because the parameters are scalars in the univariate case and a vector and a matrix in the multivariate case and are clearly distinguishable by capitalization and type face.

Lemma 5.9. *If $\mathbf{M}$ is a positive semi-definite matrix, then there exists a normal random vector $\mathbf{X}$ such that $E(\mathbf{X}) = \mu$ and $\operatorname{var}(\mathbf{X}) = \mathbf{M}$.*

Proof. In Corollary E.7 in Appendix E the symmetric square root $\mathbf{M}^{1/2}$ of $\mathbf{M}$ is defined. Now define $\mathbf{X} = \boldsymbol{\mu} + \mathbf{M}^{1/2}\mathbf{Z}$, where $\mathbf{Z}$ is multivariate standard normal. Then by Theorem 5.3

$$E(\mathbf{X}) = \boldsymbol{\mu} + \mathbf{M}^{1/2} E(\mathbf{Z}) = \boldsymbol{\mu}$$

and

$$\operatorname{var}(\mathbf{X}) = \mathbf{M}^{1/2} \operatorname{var}(\mathbf{Z}) \mathbf{M}^{1/2} = \mathbf{M}^{1/2} \mathbf{I} \mathbf{M}^{1/2} = \mathbf{M}^{1/2} \mathbf{M}^{1/2} = \mathbf{M}$$

□

5.2.1 The Density of a Non-Degenerate Normal Random Vector

How about the density of the multivariate normal distribution? First we have to say that it may not have a density. If the variance parameter $\mathbf{M}$ is not positive definite, then the random vector will be concentrated on a hyperplane (will be degenerate) by Theorem 5.6, in which case it won't have a density. Otherwise, it will.

Another approach to the same issue is to consider that $\mathbf{X}$ will have support on the whole of $\mathbb{R}^n$ only if the transformation

$$\mathbf{g}(\mathbf{z}) = \mathbf{a} + \mathbf{B}\mathbf{z}$$

is invertible, in which case its inverse is

$$\mathbf{h}(\mathbf{x}) = \mathbf{B}^{-1}(\mathbf{x} - \mathbf{a})$$

and has derivative matrix

$$\nabla \mathbf{h}(\mathbf{x}) = \mathbf{B}^{-1}$$

Thus we find the density of $\mathbf{X}$ by the multivariate change of variable theorem (Corollary 1.6 of Chapter 1 of these notes)

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= f_{\mathbf{Z}}[\mathbf{h}(\mathbf{x})] \cdot \left|\det\bigl(\nabla \mathbf{h}(\mathbf{x})\bigr)\right|. \\ &= f_{\mathbf{Z}}\bigl(\mathbf{B}^{-1}(\mathbf{x} - \mathbf{a})\bigr) \cdot \left|\det\bigl(\mathbf{B}^{-1}\bigr)\right|. \\ &= \frac{\left|\det\bigl(\mathbf{B}^{-1}\bigr)\right|}{(2\pi)^{n/2}} \exp\left(-\tfrac{1}{2}[\mathbf{B}^{-1}(\mathbf{x} - \mathbf{a})]'\mathbf{B}^{-1}(\mathbf{x} - \mathbf{a})\right) \\ &= \frac{\left|\det\bigl(\mathbf{B}^{-1}\bigr)\right|}{(2\pi)^{n/2}} \exp\left(-\tfrac{1}{2}(\mathbf{x} - \mathbf{a})'(\mathbf{B}^{-1})'\mathbf{B}^{-1}(\mathbf{x} - \mathbf{a})\right) \end{aligned}$$

Now we need several facts about matrices and determinants to clean this up. First, $(\mathbf{B}^{-1})'\mathbf{B}^{-1} = \mathbf{M}^{-1}$, where, as above, $\mathbf{M} = \operatorname{var}(\mathbf{X})$ because of two facts about inverses, transposes, and products.

- The inverse of a transpose is the transpose of the inverse.

  Hence $(\mathbf{B}^{-1})' = (\mathbf{B}')^{-1}$

- The inverse of a product is the product of the inverses in reverse order, that is, $(\mathbf{C}\mathbf{D})^{-1} = \mathbf{D}^{-1}\mathbf{C}^{-1}$ for any invertible matrices $\mathbf{C}$ and $\mathbf{D}$.

  Hence $(\mathbf{B}')^{-1}\mathbf{B}^{-1} = (\mathbf{B}\mathbf{B}')^{-1} = \mathbf{M}^{-1}$.

Second, $|\det\bigl(\mathbf{B}^{-1}\bigr)| = \det(\mathbf{M})^{-1/2}$ because of two facts about determinants, inverses, and products.

- The determinant of an inverse is the multiplicative inverse (reciprocal) of the determinant.

  Hence $\det\bigl(\mathbf{B}^{-1}\bigr) = \det(\mathbf{B})^{-1}$.

- The determinant of a matrix and its transpose are the same.

  Hence $\det(\mathbf{B}) = \det(\mathbf{B}')$.

- The determinant of a product is the product of the determinants, that is, $\det(\mathbf{CD}) = \det(\mathbf{C})\det(\mathbf{D})$ for any matrices $\mathbf{C}$ and $\mathbf{D}$.

  Hence $\det(\mathbf{M}) = \det(\mathbf{BB}') = \det(\mathbf{B})^2$.

Putting this all together, we get

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\det(\mathbf{M})^{1/2}} \exp\left(-\tfrac{1}{2}(\mathbf{x}-\mathbf{a})'\mathbf{M}^{-1}(\mathbf{x}-\mathbf{a})\right), \qquad \mathbf{x} \in \mathbb{R}^n \quad (5.24)$$

Note that the formula does not involve $\mathbf{B}$. The distribution does indeed only depend on the parameters $\mathbf{a}$ and $\mathbf{M}$ as the notation $\mathcal{N}_n(\mathbf{a}, \mathbf{M})$ implies.

Recall from Lemma 5.9 that there exists a $\mathcal{N}(\mathbf{a}, \mathbf{M})$ for every vector $\mathbf{a}$ and every symmetric positive semi-definite matrix $\mathbf{M}$. If $\mathbf{M}$ is not positive definite, then the distribution is degenerate and has no density. Otherwise, it has the density (5.24).

While we are on the subject, we want to point out that every density that looks like even vaguely like (5.24) is multivariate normal. Of course, we will have to be a bit more precise than "even vaguely like" to get a theorem. A general *quadratic form* is a function $q : \mathbb{R}^n \to \mathbb{R}$ defined by

$$q(\mathbf{x}) = \tfrac{1}{2}\mathbf{x}'\mathbf{A}\mathbf{x} + \mathbf{b}'\mathbf{x} + c \quad (5.25)$$

where $\mathbf{A}$ is an $n \times n$ matrix, $\mathbf{b}$ is an n vector, and c is a scalar. There is no loss of generality in assuming $\mathbf{A}$ is symmetric, because

$$\tfrac{1}{2}\mathbf{x}'\mathbf{A}\mathbf{x} = \tfrac{1}{2}\mathbf{x}'\mathbf{A}'\mathbf{x} = \mathbf{x}'(\mathbf{A} + \mathbf{A}')\mathbf{x},$$

the first equality following from the rule for the transpose of a product, and the second equality coming from averaging the two sides of the first equality. The matrix in the middle of the expression on the right hand side is symmetric. If we replaced $\mathbf{A}$ in the definition of q by the symmetric matrix $\frac{1}{2}(\mathbf{A} + \mathbf{A}')$, we would still be defining the same function. Thus we assume from here on that the matrix in the definition of any quadratic form is symmetric.

Theorem 5.10. *If q is a quadratic form defined by* (5.25) *and*

$$f(\mathbf{x}) = e^{-q(\mathbf{x})}, \qquad \mathbf{x} \in \mathbb{R}^n$$

is the probability density of a random variable $\mathbf{X}$, then

(a) $\mathbf{A}$ *is positive definite,*

(b) $\mathbf{X}$ *has a non-degenerate multivariate normal distribution,*

(c) $\operatorname{var}(\mathbf{X}) = \mathbf{A}^{-1}$*, and*

(d) $E(\mathbf{X}) = -\mathbf{A}^{-1}\mathbf{b}$.

Proof. The proof that $\mathbf{A}$ is positive definite has to do with the existence of the integral $\int f(\mathbf{x})\,d\mathbf{x} = 1$. We claim that unless $\mathbf{A}$ is positive definite the integral does not exist and cannot define a probability density.

First note that, since the density is continuous, it is bounded on bounded sets. We only need to worry about the behavior of the integrand near infinity. Second, since

$$\frac{f(\mathbf{x})}{e^{-\mathbf{x}'\mathbf{A}\mathbf{x}/2}} \to 1, \qquad \text{as } \mathbf{x} \to \infty,$$

we may in determining when the integral exists consider only the quadratic part in the definition of q. Let $\mathbf{A} = \mathbf{O}\mathbf{D}\mathbf{O}'$ be the spectral decomposition (Theorem E.4 in Appendix E) of $\mathbf{A}$, and consider the change of variables $\mathbf{y} = \mathbf{O}'\mathbf{x}$, which has inverse transformation $\mathbf{x} = \mathbf{O}\mathbf{y}$ and Jacobian one. Using this change of variables we see

$$\begin{aligned}
\int e^{-\mathbf{x}'\mathbf{A}\mathbf{x}/2}\,d\mathbf{x} &= \int e^{-\mathbf{y}'\mathbf{D}\mathbf{y}/2}\,d\mathbf{y} \\
&= \iint \cdots \int \exp\left(-\frac{1}{2}\sum_{i=1}^{n} d_{ii}y_i^2\right) dy_1\,dy_2 \cdots dy_n \\
&= \prod_{i=1}^{n}\left(\int_{-\infty}^{\infty} e^{-d_{ii}y_i^2/2}\,dy_i\right)
\end{aligned}$$

It is obvious that all the integrals in the last line exist if and only if each d_{ii} is strictly positive, which happens if and only if $\mathbf{A}$ is positive definite. That proves (a).

Now we just "complete the square." We want to put $q(\mathbf{x})$ in the same form as the quadratic form

$$\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\mathbf{M}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \tag{5.26}$$

in the exponent of the usual expression for the normal distribution. Expand (5.26)

$$\begin{aligned}
\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\mathbf{M}^{-1}(\mathbf{x} - \boldsymbol{\mu}) &= \tfrac{1}{2}\mathbf{x}'\mathbf{M}^{-1}\mathbf{x} - \tfrac{1}{2}\mathbf{x}'\mathbf{M}^{-1}\boldsymbol{\mu} - \tfrac{1}{2}\boldsymbol{\mu}'\mathbf{M}^{-1}\mathbf{x} + \tfrac{1}{2}\boldsymbol{\mu}'\mathbf{M}^{-1}\boldsymbol{\mu} \\
&= \tfrac{1}{2}\mathbf{x}'\mathbf{M}^{-1}\mathbf{x} - \boldsymbol{\mu}'\mathbf{M}^{-1}\mathbf{x} + \tfrac{1}{2}\boldsymbol{\mu}'\mathbf{M}^{-1}\boldsymbol{\mu}
\end{aligned}$$

(the second equality holding because of the rule for the transpose of a product). Now the only way $q(\mathbf{x})$ can match up with this is if the constants in the quadratic and linear terms both match, that is,

$$\mathbf{A} = \mathbf{M}^{-1}$$

and

$$\mathbf{b}' = -\boldsymbol{\mu}'\mathbf{M}^{-1},$$

and these in turn imply

$$\boldsymbol{\mu} = -\mathbf{A}^{-1}\mathbf{b} \tag{5.27}$$

$$\mathbf{M} = \mathbf{A}^{-1} \tag{5.28}$$

which in turn are (c) and (d) if (b) is true. So all that remains is to prove (b).

We have now shown that the quadratic and linear terms of $q(\mathbf{x})$ and (5.26) match when we define $\boldsymbol{\mu}$ and $\mathbf{M}$ by (5.27) and (5.28). Hence

$$q(\mathbf{x}) = \tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\mathbf{M}^{-1}(\mathbf{x} - \boldsymbol{\mu}) + c - \tfrac{1}{2}\boldsymbol{\mu}'\mathbf{M}^{-1}\boldsymbol{\mu}$$

and

$$f(\mathbf{x}) = \exp\left(-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\mathbf{M}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) \exp\left(c - \tfrac{1}{2}\boldsymbol{\mu}'\mathbf{M}^{-1}\boldsymbol{\mu}\right)$$

Since the first term on the right hand side is an unnormalized density of the $\mathcal{N}(\boldsymbol{\mu}, \mathbf{M})$ distribution, the second term must be the reciprocal of the normalizing constant so that $f(\mathbf{x})$ integrates to one. That proves (b), and we are done. □

I call this the "e to a quadratic" theorem. If the density is the exponential of a quadratic form, then the distribution must be non-degenerate multivariate normal, and the mean and variance can be read off the density.

5.2.2 Marginals

Lemma 5.11. *Every linear transformation of a multivariate normal random vector is (multivariate or univariate) normal.*

This obvious because a linear transformation of a linear transformation is linear. If $\mathbf{X}$ is multivariate normal, then, by definition, it has the form $\mathbf{X} = \mathbf{a} + \mathbf{B}\mathbf{Z}$, where $\mathbf{Z}$ is standard normal, $\mathbf{a}$ is a constant vector, and $\mathbf{B}$ is a constant matrix. So if $\mathbf{Y} = \mathbf{c} + \mathbf{D}\mathbf{X}$, where $\mathbf{c}$ is a constant vector and $\mathbf{D}$ is a constant matrix, then

$$\begin{aligned} \mathbf{Y} &= \mathbf{c} + \mathbf{D}\mathbf{X} \\ &= \mathbf{c} + \mathbf{D}(\mathbf{a} + \mathbf{B}\mathbf{Z}) \\ &= (\mathbf{c} + \mathbf{D}\mathbf{a}) + (\mathbf{D}\mathbf{B})\mathbf{Z}, \end{aligned}$$

which is clearly a linear transformation of $\mathbf{Z}$, hence normal.

Corollary 5.12. *Every marginal distribution of a multivariate normal distribution is (multivariate or univariate) normal.*

This is an obvious consequence of the lemma, because the operation of finding a marginal defines a linear transformation, simply because of the definitions of vector addition and scalar multiplication, that is, because the i-th component of $a\mathbf{X} + b\mathbf{Y}$ is $aX_i + bY_i$.

5.2.3 Partitioned Matrices

This section has no probability theory, just an odd bit of matrix algebra. The notation

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix} \tag{5.29}$$

indicates a *partitioned matrix*. Here each of the $\mathbf{B}_{ij}$ is itself a matrix. $\mathbf{B}$ is just the matrix having the elements of $\mathbf{B}_{11}$ in its upper left corner, with the elements of $\mathbf{B}_{12}$ to their right, and so forth. Of course the dimensions of the $\mathbf{B}_{ij}$ must fit together the right way.

One thing about partitioned matrices that makes them very useful is that matrix multiplication looks "just like" matrix multiplication of non-partitioned matrices. You just treat the matrices like scalar elements of an ordinary array

$$\begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{B}_{11}\mathbf{C}_{11} + \mathbf{B}_{12}\mathbf{C}_{21} & \mathbf{B}_{11}\mathbf{C}_{12} + \mathbf{B}_{12}\mathbf{C}_{22} \\ \mathbf{B}_{21}\mathbf{C}_{11} + \mathbf{B}_{22}\mathbf{C}_{21} & \mathbf{B}_{21}\mathbf{C}_{12} + \mathbf{B}_{22}\mathbf{C}_{22} \end{pmatrix}$$

If one of the matrixes is a partitioned column vector, it looks like the multiplication of a vector by a matrix

$$\begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{B}_{11}\mathbf{x}_1 + \mathbf{B}_{12}\mathbf{x}_2 \\ \mathbf{B}_{21}\mathbf{x}_1 + \mathbf{B}_{22}\mathbf{x}_2 \end{pmatrix}$$

and similarly for

$$\begin{aligned} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}' \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} &= \begin{pmatrix} \mathbf{x}_1' & \mathbf{x}_2' \end{pmatrix} \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{x}_1' & \mathbf{x}_2' \end{pmatrix} \begin{pmatrix} \mathbf{B}_{11}\mathbf{x}_1 + \mathbf{B}_{12}\mathbf{x}_2 \\ \mathbf{B}_{21}\mathbf{x}_1 + \mathbf{B}_{22}\mathbf{x}_2 \end{pmatrix} \\ &= \mathbf{x}_1'\mathbf{B}_{11}\mathbf{x}_1 + \mathbf{x}_1'\mathbf{B}_{12}\mathbf{x}_2 + \mathbf{x}_2'\mathbf{B}_{21}\mathbf{x}_1 + \mathbf{x}_2'\mathbf{B}_{22}\mathbf{x}_2 \end{aligned}$$

Of course, in all of these, the dimensions have be such that the matrix multiplications make sense.

If $\mathbf{X}$ is a partitioned random vector

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}, \tag{5.30a}$$

then its mean mean vector is

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \tag{5.30b}$$

where

$$\boldsymbol{\mu}_i = E(\mathbf{X}_i),$$

and its variance matrix is

$$\mathbf{M} = \begin{pmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{pmatrix}, \tag{5.30c}$$

where

$$\mathbf{M}_{ij} = \operatorname{cov}(\mathbf{X}_i, \mathbf{X}_j).$$

Again, every thing looks very analogous to the situation with scalar rather than vector or matrix components.

A partitioned matrix is called *block diagonal* if the "off-diagonal" matrices are all zero. The partitioned matrix (5.29) is block diagonal if $\mathbf{B}_{12} = 0$ and $\mathbf{B}_{21} = 0$. The partitioned matrix (5.30c) is block diagonal if $\mathbf{X}_1$ and $\mathbf{X}_2$ are uncorrelated, that is, $\operatorname{cov}(\mathbf{X}_1, \mathbf{X}_2) = 0$.

A block diagonal matrix with square blocks on the diagonal, is easy to invert, just invert each block. For example, if (5.30c) is block diagonal, then

$$\mathbf{M}^{-1} = \begin{pmatrix} \mathbf{M}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{M}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{22}^{-1} \end{pmatrix} \tag{5.31}$$

5.2.4 Conditionals and Independence

In this section we consider a normal random vector $\mathbf{X}$ partitioned as in (5.30a) with variance matrix $\mathbf{M}$, which must be partitioned as in (5.30c). We will need a notation for the inverse variance matrix: we adopt $\mathbf{W} = \mathbf{M}^{-1}$. Of course, it can be partitioned in the same way

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{21} & \mathbf{W}_{22} \end{pmatrix} \tag{5.32}$$

Note from (5.31) that if $\mathbf{M}$ is block diagonal and invertible, then so is $\mathbf{W}$ and $\mathbf{W}_{ii} = \mathbf{M}_{ii}^{-1}$. When $\mathbf{M}$ is not block diagonal, then neither is $\mathbf{W}$ and the relation between the two is complicated.

Theorem 5.13. *Random vectors that are jointly multivariate normal and uncorrelated are independent.*

In notation, what the theorem says is that if $\mathbf{X}$ is multivariate normal and partitioned as in (5.30a) with variance matrix (5.30c), then

$$\mathbf{M}_{12} = \operatorname{cov}(\mathbf{X}_1, \mathbf{X}_2) = 0$$

implies that $\mathbf{X}_1$ and $\mathbf{X}_2$ are actually independent random vectors.

Please note the contrast with the general case.

> *In general independent implies uncorrelated, but uncorrelated does* ***not*** *imply independent.*
>
> *Only when the random variables are* ***jointly multivariate normal*** *does uncorrelated imply independent.*

Proof. Without loss of generality, we may assume the means are zero, because $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent if and only if $\mathbf{X}_1 - \boldsymbol{\mu}_1$ and $\mathbf{X}_2 - \boldsymbol{\mu}_2$ are independent.

We first prove the special case in which $\mathbf{X}$ has a non-degenerate distribution. Then the unnormalized density (ignoring constants) is

$$\exp\left(-\tfrac{1}{2}\mathbf{x}'\mathbf{W}\mathbf{x}\right) = \exp\left(-\tfrac{1}{2}\mathbf{x}_1'\mathbf{W}_{11}\mathbf{x}_1\right) \exp\left(-\tfrac{1}{2}\mathbf{x}_2'\mathbf{W}_{22}\mathbf{x}_2\right)$$

In general, there is also a $\mathbf{x}_1'\mathbf{W}_{12}\mathbf{x}_2$ term in the exponent, but it vanishes here because $\mathbf{W}$ is block diagonal because of (5.31). Since the density factors, the random vectors are independent.

We now prove the general case by expressing some variables in terms of the others. If $\mathbf{X}$ is concentrated on a hyperplane, then we can express one variable as a linear combination of the remaining $n - 1$ variables. If these are still concentrated on a hyperplane, then we can express another variable as a linear combination of the remaining $n - 2$ and so forth. We stop when we have expressed some variables as linear combinations of a set of k variables which have a non-degenerate multivariate normal distribution. We can now partition $\mathbf{X}$ as

$$\mathbf{X} = \begin{pmatrix} \mathbf{U}_1 \\ \mathbf{V}_1 \\ \mathbf{U}_2 \\ \mathbf{V}_2 \end{pmatrix}$$

where $(\mathbf{U}_1, \mathbf{U}_2)$ has a non-degenerate multivariate normal distribution and

$$\begin{aligned} \mathbf{V}_1 &= \mathbf{B}_{11}\mathbf{U}_1 + \mathbf{B}_{12}\mathbf{U}_2 \\ \mathbf{V}_2 &= \mathbf{B}_{21}\mathbf{U}_1 + \mathbf{B}_{22}\mathbf{U}_2 \end{aligned}$$

for some matrix $\mathbf{B}$ partitioned as in (5.29), and $\mathbf{X}_i = (\mathbf{U}_i, \mathbf{V}_i)$. Note that the assumption that $\mathbf{X}_1$ and $\mathbf{X}_2$ are uncorrelated implies that $\mathbf{U}_1$ and $\mathbf{U}_2$ are also uncorrelated and hence, by what has already been proved independent (since they are jointly non-degenerate multivariate normal).

Then, using the additional notation

$$\begin{aligned} \operatorname{var}(\mathbf{U}_1) &= \mathbf{S}_{11} \\ \operatorname{var}(\mathbf{U}_2) &= \mathbf{S}_{22} \end{aligned}$$

we calculate that $\operatorname{var}(\mathbf{X})$ is

$$\begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{11}\mathbf{B}_{11}' & \mathbf{0} & \mathbf{S}_{11}\mathbf{B}_{21}' \\ \mathbf{B}_{11}\mathbf{S}_{11} & \mathbf{B}_{11}\mathbf{S}_{11}\mathbf{B}_{11}' + \mathbf{B}_{12}\mathbf{S}_{22}\mathbf{B}_{12}' & \mathbf{B}_{12}\mathbf{S}_{22} & \mathbf{B}_{11}\mathbf{S}_{11}\mathbf{B}_{21}' + \mathbf{B}_{12}\mathbf{S}_{22}\mathbf{B}_{22}' \\ \mathbf{0} & \mathbf{S}_{22}\mathbf{B}_{12}' & \mathbf{S}_{22} & \mathbf{S}_{22}\mathbf{B}_{22}' \\ \mathbf{B}_{21}\mathbf{S}_{11} & \mathbf{B}_{21}\mathbf{S}_{11}\mathbf{B}_{11}' + \mathbf{B}_{22}\mathbf{S}_{22}\mathbf{B}_{12}' & \mathbf{B}_{22}\mathbf{S}_{22} & \mathbf{B}_{21}\mathbf{S}_{11}\mathbf{B}_{21}' + \mathbf{B}_{22}\mathbf{S}_{22}\mathbf{B}_{22}' \end{pmatrix}$$

Now the assumption of the theorem is that this matrix is block diagonal, with the blocks now 2×2. Since $\mathbf{U}_1$ and $\mathbf{U}_2$ are nondegenerate, their variance matrices are invertible, thus the only way we can have $\mathbf{B}_{21}\mathbf{S}_{11} = 0$ and $\mathbf{B}_{12}\mathbf{S}_{22} = 0$ is if $\mathbf{B}_{21} = 0$ and $\mathbf{B}_{12} = 0$. But this implies

$$\mathbf{X}_i = \begin{pmatrix} \mathbf{U}_i \\ \mathbf{B}_{ii}\mathbf{U}_i \end{pmatrix}$$

for $i = 1$, 2, and since these are functions of the independent random vectors $\mathbf{U}_1$ and $\mathbf{U}_2$, they are independent. □

Every conditional of a normal random vector is normal too, but it is hard for us to give an explicit expression for the degenerate case. This is not surprising, because all our methods for finding conditional distributions involve densities and degenerate normal distributions don't have densities.

First a lemma.

Lemma 5.14. *Suppose* $\mathbf{X}$ *is partitioned as in* (5.30a) *and has variance matrix* (5.30c)*, and suppose that* $\mathbf{M}_{22}$ *is positive definite. Then*

$$\mathbf{X}_1 - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{X}_2 \quad \textit{and} \quad \mathbf{X}_2$$

are uncorrelated.

And, we should note, by Theorem 5.13, if $\mathbf{X}$ is multivariate normal, then $\mathbf{X}_1 - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{X}_2$ is independent of $\mathbf{X}_2$.

Proof. Obvious, just calculate the covariance

$$\begin{aligned}
\operatorname{cov}(\mathbf{X}_1 - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{X}_2, \mathbf{X}_2) &= \operatorname{cov}(\mathbf{X}_1, \mathbf{X}_2) - \mathbf{M}_{12}\mathbf{M}_{22}^{-1} \operatorname{cov}(\mathbf{X}_2, \mathbf{X}_2) \\
&= \mathbf{M}_{12} - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{22} \\
&= 0
\end{aligned}$$

□

Every conditional of a normal random vector is also normal, but it is hard for us to give an explicit expression for the degenerate case. This is not surprising, because all our methods for finding conditional densities and degenerate normal distributions don't have densities. So here we will be satisfied with describing the non-degenerate case.

Theorem 5.15. *Every condition distribution of a non-degenerate multivariate normal distribution is non-degenerate (multivariate or univariate) normal.*

In particular, if $\mathbf{X}$ *is partitioned as in* (5.30a)*, has the multivariate normal distribution with mean vector* (5.30b) *and variance matrix* (5.30c)*, then*

$$\mathbf{X}_1 \mid \mathbf{X}_2 \sim \mathcal{N}(\boldsymbol{\mu}_1 + \mathbf{M}_{12}\mathbf{M}_{22}^{-1}[\mathbf{X}_2 - \boldsymbol{\mu}_2], \mathbf{M}_{11} - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{21}). \tag{5.33}$$

Proof. First note that the conditional distribution is multivariate normal by Lemma 5.10, because the joint density is the exponential of a quadratic, hence so is the conditional, which is just the joint density considered as a function of x_1 with x_2 fixed renormalized.

So all that remains to be done is figuring out the conditional mean and variance. For the conditional mean, we use Lemma 5.14 and the comment following it. Because of the independence of $\mathbf{X}_1 - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{X}_2$ and $\mathbf{X}_2$,

$$E(\mathbf{X}_1 - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{X}_2 \mid \mathbf{X}_2) = E(\mathbf{X}_1 - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{X}_2)$$

but

$$E(\mathbf{X}_1 - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{X}_2 \mid \mathbf{X}_2) = E(\mathbf{X}_1 \mid \mathbf{X}_2) - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{X}_2$$

by linearity of expectations and functions of the conditioning variable behaving like constants, and

$$E(\mathbf{X}_1 - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{X}_2) = \boldsymbol{\mu}_1 - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\boldsymbol{\mu}_2.$$

Thus

$$E(\mathbf{X}_1 \mid \mathbf{X}_2) - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{X}_2 = \boldsymbol{\mu}_1 - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\boldsymbol{\mu}_2,$$

which establishes the conditional expectation given in (5.33).

To calculate the variance, we first observe that

$$\operatorname{var}(\mathbf{X}_1 \mid \mathbf{X}_2) = W_{11}^{-1} \tag{5.34}$$

where $\mathbf{W} = \mathbf{M}^{-1}$ is partitioned as in (5.32), because the quadratic form in the exponent of the density has quadratic term $\mathbf{x}_1 W_{11} \mathbf{x}_1$ and Theorem 5.10 says that is the inverse variance matrix of the vector in question, which in this case is x_1 given x_2. We don't know what the form of W_{11} or it's inverse it, but we do know it is a constant matrix, which is all we need. The rest of the job can be done by the vector version of the iterated variance formula (Theorem 3.7)

$$\operatorname{var}(\mathbf{X}_1) = \operatorname{var}\{E(\mathbf{X}_1 \mid \mathbf{X}_2)\} + E\{\operatorname{var}(\mathbf{X}_1 \mid \mathbf{X}_2)\} \tag{5.35}$$

(which we haven't actually proved but is proved in exactly the same way as the scalar formula). We know

$$\operatorname{var}(\mathbf{X}_1) = \mathbf{M}_{11}$$

but

$$\begin{aligned}
&\operatorname{var}\{E(\mathbf{X}_1 \mid \mathbf{X}_2)\} + E\{\operatorname{var}(\mathbf{X}_1 \mid \mathbf{X}_2)\} \\
&\qquad = \operatorname{var}\{\boldsymbol{\mu}_1 + \mathbf{M}_{12}\mathbf{M}_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2)\} + E\{\mathbf{W}_{11}^{-1}\} \\
&\qquad = \operatorname{var}(\mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{X}_2) + \mathbf{W}_{11}^{-1} \\
&\qquad = \mathbf{M}_{12}\mathbf{M}_{22}^{-1} \operatorname{var}(\mathbf{X}_2)\mathbf{M}_{22}^{-1}\mathbf{M}_{12}' + \mathbf{W}_{11}^{-1} \\
&\qquad = \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{22}\mathbf{M}_{22}^{-1}\mathbf{M}_{21} + \mathbf{W}_{11}^{-1} \\
&\qquad = \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{21} + \mathbf{W}_{11}^{-1}
\end{aligned}$$

Equating the two gives

$$\mathbf{M}_{11} = \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{21} + \mathbf{W}_{11}^{-1}$$

which along with (5.34) establishes the conditional variance given in (5.33). □

5.3 Bernoulli Random Vectors

To start we generalize the notion of a Bernoulli random variables. One might think that should be a vector with i. i. d. Bernoulli components, but something quite different is in order. A (univariate) Bernoulli random variable is really an indicator function. *All* zero-or-one valued random variables are indicator functions: they indicate the set on which they are one. How do we generalize the notion of an indicator function to the multivariate case? We consider a vector of indicator functions.

We give three closely related definitions.

Definition 5.3.1 (Bernoulli Random Vector).
A random vector $\mathbf{X} = (X_1, \ldots, X_k)$ *is* ***Bernoulli*** *if the* X_i *are the indicators of a partition of the sample space, that is,*

$$X_i = I_{A_i}$$

where

$$A_i \cap A_j = \varnothing, \qquad i \neq j$$

and

$$\bigcup_{i=1}^{k} A_i$$

is the whole sample space.

Definition 5.3.2 (Bernoulli Random Vector).
A random vector $\mathbf{X} = (X_1, \ldots, X_k)$ *is* ***Bernoulli*** *if the* X_i *are zero-or-one-valued random variables and*

$$X_1 + \cdots + X_k = 1.$$

with probability one.

Definition 5.3.3 (Bernoulli Random Vector).
A random vector $\mathbf{X} = (X_1, \ldots, X_k)$ *is* ***Bernoulli*** *if the* X_i *are zero-or-one-valued random variables and with probability one exactly one of* X_1, ..., X_k *is one and the rest are zero.*

The equivalence of Definitions 5.3.2 and 5.3.3 is obvious. The only way a bunch of zeros and ones can add to one is if there is exactly one one.

The equivalence of Definitions 5.3.1 and 5.3.3 is also obvious. If the A_i form a partition, then exactly one of the

$$X_i(\omega) = I_{A_i}(\omega)$$

is equal to one for any outcome ω, the one for which $\omega \in A_i$. There is, of course, exactly one i such that $\omega \in A_i$ just by definition of "partition."

5.3.1 Categorical Random Variables

Bernoulli random vectors are closely related to *categorical random variables* taking values in an arbitrary finite set. You may have gotten the impression up to know that probability theorists have a heavy preference for numerical random variables. That's so. Our only "brand name" distribution that is not necessarily numerical valued is the discrete uniform distribution. In principle, though a random variable can take values in *any* set. So although we haven't done much with such variables so far, we haven't ruled them out either. Of course, if Y is a random variable taking values in the set

$$S = \{\text{strongly agree}, \text{agree}, \text{neutral}, \text{disagree}, \text{strongly disagree}\} \tag{5.36}$$

you can't talk about expectations or moments, $E(Y)$ is defined only for numerical (or numerical vector) random variables, not for categorical random variables.

However, if we number the categories

$$S = \{s_1, s_2, \ldots, s_5\}$$

with s_1 = strongly agree, and so forth, then we can identify the categorical random variable Y with a Bernoulli random vector $\mathbf{X}$

$$X_i = I_{\{s_i\}}(Y)$$

that is

$$X_i = 1 \quad \text{if and only if} \quad Y = s_i.$$

Thus Bernoulli random variables are an artifice. They are introduced to inject some numbers into categorical problems. We can't talk about $E(Y)$, but we can talk about $E(\mathbf{X})$. A thorough analysis of the properties of the distribution of the random vector $\mathbf{X}$ will also tell us everything we want to know about the categorical random variable Y, and it will do so allowing us to use the tools (moments, etc.) that we already know.

5.3.2 Moments

Each of the X_i is, of course, univariate Bernoulli, write

$$X_i \sim \text{Ber}(p_i)$$

and collect these parameters into a vector

$$\mathbf{p} = (\mathbf{p}_1, \ldots, \mathbf{p}_k)$$

Then we abbreviate the distribution of $\mathbf{X}$ as

$$\mathbf{X} \sim \text{Ber}_k(\mathbf{p})$$

if we want to indicate the dimension k or just as $\mathbf{X} \sim \text{Ber}(\mathbf{p})$ if the dimension is clear from the context (the boldface type indicating a vector parameter makes it clear this is not the univariate Bernoulli).

Since each X_i is univariate Bernoulli,

$$\begin{aligned} E(X_i) &= p_i \\ \text{var}(X_i) &= p_i(1 - p_i) \end{aligned}$$

That tells us

$$E(\mathbf{X}) = \mathbf{p}.$$

To find the variance matrix we need to calculate covariances. For $i \neq j$,

$$\text{cov}(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j) = -p_i p_j,$$

because $X_i X_j = 0$ with probability one.

Hence $\operatorname{var}(\mathbf{X}) = \mathbf{M}$ has components

$$m_{ij} = \begin{cases} p_i(1 - p_i), & i = j \\ -p_i p_j & i \neq j \end{cases} \tag{5.37}$$

We can also write this using more matrixy notation by introducing the diagonal matrix $\mathbf{P}$ having diagonal elements p_i and noting that the "outer product" $\mathbf{p}\mathbf{p}'$ has elements $p_i p_j$, hence

$$\operatorname{var}(\mathbf{X}) = \mathbf{P} - \mathbf{p}\mathbf{p}'$$

Question: Is $\operatorname{var}(\mathbf{X})$ positive definite? This is of course related to the question of whether $\mathbf{X}$ is degenerate. We haven't said anything explicit about either, but the information needed to answer these questions is in the text above. It should be obvious if you know what to look for (a good exercise testing your understanding of degenerate random vectors).

5.4 The Multinomial Distribution

The multinomial distribution is the multivariate analog of the binomial distribution. It is sort of, but not quite, the multivariate generalization, that is, the binomial distribution is sort of, but not precisely, a special case of the multinomial distribution. Thus is unlike the normal, where the univariate normal distribution is precisely the one-dimensional case of the multivariate normal.

Suppose $\mathbf{X}_1$, $\mathbf{X}_2$ are an i. i. d. sequence of $\operatorname{Ber}_k(\mathbf{p})$ random vectors (caution: the subscripts on the $\mathbf{X}_i$ indicate elements of an infinite sequence of i. i. d. random vectors, not components of one vector). Then

$$\mathbf{Y} = \mathbf{X}_1 + \cdots + \mathbf{X}_n$$

has the *multinomial distribution* with *sample size* n and dimension k, abbreviated

$$\mathbf{Y} \sim \operatorname{Multi}_k(n, \mathbf{p})$$

if we want to indicate the dimension in the notation or just $\mathbf{Y} \sim \operatorname{Multi}(n, \mathbf{p})$ if the dimension is clear from the context.

Note the dimension is k, not n, that is, both $\mathbf{Y}$ and $\mathbf{p}$ are vectors of dimension k.

5.4.1 Categorical Random Variables

Recall that a multinomial random vector is the sum of i. i. d. Bernoullis

$$\mathbf{Y} = \mathbf{X}_1 + \cdot + \mathbf{X}_n$$

and that each Bernoulli is related to a categorical random variable: $X_{i,j} = 1$ if and only if the i-th observation fell in the j-th category. Thus $Y_j = \sum_i X_{i,j}$ is the number of individuals that fell in the j-th category.

This gives us another distribution of multinomial random vectors. A random vector $\mathbf{Y} = \text{Multi}(n, p)$ arises by observing a sequence of n independent random variables (taking values in any set) and letting Y_j be the number of observations that fall in the j-th category. The parameter p_j is the probability of any one individual observation falling in the j-th category.

5.4.2 Moments

Obvious, just n times the moments of $\text{Ber}(\mathbf{p})$

$$\begin{aligned} E(\mathbf{X}) &= n\mathbf{p} \\ \text{var}(\mathbf{X}) &= n(\mathbf{P} - \mathbf{p}\mathbf{p}') \end{aligned}$$

5.4.3 Degeneracy

Since the components of a $\text{Ber}(\mathbf{p})$ random vector sum to one, the components of a $\text{Multi}(n, \mathbf{p})$ random vector sum to n. That is, if $\mathbf{Y} \sim \text{Multi}(n, \mathbf{p})$, then

$$Y_1 + \cdots Y_k = n$$

with probability one. This can be written $\mathbf{u}'\mathbf{Y} = n$ with probability one, where $\mathbf{u} = (1, 1, \ldots, 1)$. Thus $\mathbf{Y}$ is concentrated on the hyperplane

$$H = \{\, \mathbf{y} \in \mathbb{R}^k : \mathbf{u}'\mathbf{y} = n \,\}$$

Is $\mathbf{Y}$ concentrated on any other hyperplanes? Since the $\text{Ber}_k(\mathbf{p})$ distribution and the $\text{Multi}_k(n, \mathbf{p})$ distribution have the same variance matrices except for a constant of proportionality ($\mathbf{M}$ and $n\mathbf{M}$, respectively), they both are supported on the same hyperplanes. We might as well drop the n and ask the question about the Bernoulli.

Let $\mathbf{c} = (c_1, \ldots, c_k)$ be an arbitrary vector. Such a vector is associated with a hyperplane supporting the distribution if

$$\begin{aligned} \mathbf{c}'\mathbf{M}\mathbf{c} &= \sum_{i=1}^{k} \sum_{j=1}^{k} m_{ij} c_i c_j \\ &= \sum_{i=1}^{k} p_i c_i^2 - \sum_{i=1}^{k} \sum_{j=1}^{k} p_i p_j c_i c_j \\ &= \sum_{i=1}^{k} p_i c_i^2 - \left(\sum_{j=1}^{k} p_j c_j \right)^2 \end{aligned}$$

is zero. Thinking of this as a function of $\mathbf{c}$ for fixed $\mathbf{p}$, write it as $q(\mathbf{c})$. Being a variance, it is nonnegative, hence it is zero only where it is achieving its

minimum value, and where, since it is a smooth function, its derivative must be zero, that is,

$$\frac{\partial q(\mathbf{c})}{\partial c_i} = 2p_i c_i - 2p_i \left(\sum_{j=1}^{k} p_j c_j \right) = 0$$

Now we do not know what the quantity in parentheses is, but it does not depend on i or j, so we can write it as a single letter d with no subscripts. Thus we have to solve

$$2p_i c_i - 2dp_i = 0 \tag{5.38}$$

for c_i. This splits into two cases.

Case I. If none of the p_i are zero, the only solution is $c_i = d$. Thus the only null eigenvectors are proportional to the vector $\mathbf{u} = (1, 1, \ldots, 1)$. And all such vectors determine the same hyperplane.

Case II. If any of the p_i are zero, we get more solutions. Equation (5.38) becomes $0 = 0$ when $p_i = 0$, and since this is the only equation containing c_i, the equations say nothing about c_i, thus the solution is

$$\begin{aligned} c_i &= d, && p_i > 0 \\ c_i &= \text{arbitrary}, && p_i = 0 \end{aligned}$$

In hindsight, case II was rather obvious too. If $p_i = 0$ then $X_i = 0$ with probability one, and that is another degeneracy. But our real interest is in case I. If none of the success probabilities are zero, then the only degeneracy is $Y_1 + \cdots + Y_k = n$ with probability one.

5.4.4 Density

Density? Don't degenerate distribution have no densities? In the continuous case, yes. Degenerate continuous random vectors have no densities. But discrete random vectors always have densities, as always, $f(\mathbf{x}) = P(\mathbf{X} = \mathbf{x})$.

The derivation of the density is exactly like the derivation of the binomial density. First we look at one particular outcome, then collect the outcomes that lead to the same $\mathbf{Y}$ values. Write $X_{i,j}$ for the components of $\mathbf{X}_i$, and note that if we know $X_{i,m} = 1$, then we also know $X_{i,j} = 0$ for $j \neq m$, so it is enough to determine the probability of an outcome if we simply record the X_{ij} that are equal to one. Then by the multiplication rule

$$\begin{aligned} P(X_{1,j_1} = 1 \text{ and} \cdots \text{and } X_{n,j_n} = 1) &= \prod_{i=1}^{n} P(X_{i,j_i} = 1) \\ &= \prod_{i=1}^{n} p_{j_i} \\ &= \prod_{j=1}^{k} p_j^{y_j} \end{aligned}$$

The last equality records the same kind of simplification we saw in deriving the binomial density. The product from 1 to n in the next to last line may repeat some of the p's. How often are they repeated? There is one p_j for each X_{ij} that is equal to one, and there are $Y_j = \sum_i X_{ij}$ of them.

We are not done, however, because more than one outcome can lead to the same right hand side here. How many ways are there to get exactly y_j of the X_{ij} equal to one? This is the same as asking how many ways there are to assign the numbers $i = 1, \ldots, n$ to one of k categories, so that there are y_i in the i-th category, and the answer is the multinomial coefficient

$$\binom{n}{y_1, \ldots, y_k} = \frac{n!}{y_1! \cdots y_k!}$$

Thus the density is

$$f(\mathbf{y}) = \binom{n}{y_1, \ldots, y_k} \prod_{j=1}^{k} p_j^{y_j}, \qquad \mathbf{y} \in S$$

where the sample space S is defined by

$$S = \{\, \mathbf{y} \in \mathbb{N}^k : y_1 + \cdots y_k = n \,\}$$

where $\mathbb{N}$ denotes the "natural numbers" 0, 1, 2, ... In other words, the sample space S consists of vectors $\mathbf{y}$ having nonnegative integer coordinates that sum to n.

5.4.5 Marginals and "Sort Of" Marginals

The univariate marginals are obvious. Since the univariate marginals of $\text{Ber}(\mathbf{p})$ are $\text{Ber}(p_i)$, the univariate marginals of $\text{Multi}(n, \mathbf{p})$ are $\text{Bin}(n, p_i)$.

Strictly speaking, the multivariate marginals do not have a brand name distribution. Lindgren (Theorem 8 of Chapter 6) says the marginals of a multinomial are multinomial, but this is, strictly speaking, complete rubbish, given the way he (and we) defined "marginal" and "multinomial." It is obviously wrong. If $\mathbf{X} = (X_1, \ldots, X_k)$ is multinomial, then it is degenerate. But $(X_1, \ldots, X_{k-1})$ is not degenerate, hence not multinomial (all multinomial distributions are degenerate). The same goes for any other subvector, (X_2, X_5, X_{10}), for example.

Of course, Lindgren knows this as well as I do. He is just being sloppy about terminology. What he means is clear from his discussion leading up to the "theorem" (really a non-theorem). Here's the correct statement.

Theorem 5.16. *Suppose* $\mathbf{Y} = \text{Multi}_k(n, \mathbf{p})$ *and* $\mathbf{Z}$ *is a random vector formed by collapsing some of the categories for* $\mathbf{Y}$*, that is, each component of* $\mathbf{Z}$ *has the form*

$$Z_j = Y_{i_1} + \cdots + Y_{i_{m_j}}$$

where each Y_i *contributes to exactly one* Z_j *so that*

$$Z_1 + \cdots + Z_l = Y_1 + \cdots + Y_k = n,$$

then

$$\mathbf{Z} \sim \text{Multi}_l(n, \mathbf{q})$$

where the parameter vector $\mathbf{q}$ *has components*

$$q_j = p_{i_1} + \cdots + p_{i_{m_j}}$$

is formed by collapsing the categories in the same way as in forming $\mathbf{Z}$ *from* $\mathbf{Y}$*.*

No wonder Lindgren felt the urge to sloppiness here. The correct statement is a really obnoxious mess of notation. But the idea is simple and obvious. If we collapse some categories, that gives a different (coarser) partition of the sample space and a multinomial distribution with fewer categories.

Example 5.4.1.
Consider the multinomial random vector $\mathbf{Y}$ associated with i. i. d. sampling of a categorical random variable taking values in the set (5.36). Let $\mathbf{Z}$ be the multinomial random vector associated with the categorical random variable obtained by collapsing the categories on the ends, that is, we collapse the categories "strongly agree" and "agree" and we collapse the categories "strongly disagree" and "disagree." Thus

$$\begin{aligned} \mathbf{Y} &\sim \text{Multi}_5(n, \mathbf{p}) \\ \mathbf{Z} &\sim \text{Multi}_3(n, \mathbf{q}) \end{aligned}$$

where

$$\begin{aligned} Z_1 &= Y_1 + Y_2 \\ Z_2 &= Y_3 \\ Z_3 &= Y_4 + Y_5 \end{aligned}$$

and

$$\begin{aligned} q_1 &= p_1 + p_2 \\ q_2 &= p_3 \\ q_3 &= p_4 + p_5 \end{aligned}$$

The notation is simpler than in the theorem, but still messy, obscuring the simple idea of collapsing categories. Maybe Lindgren has the right idea. Slop is good here. The marginals of a multinomial are sort of, but not precisely, multinomial. Or should that be the sort-of-but-not-precisely marginals of a multinomial are multinomial?

Recall that we started this section with the observation that one-dimensional marginal distributions of a multinomial are binomial (with no "sort of"). But two-dimensional multinomial distributions must also be somehow related to the binomial distribution. The $k = 2$ multinomial coefficients *are* binomial coefficients, that is,

$$\binom{n}{y_1, y_2} = \frac{n!}{y_1! y_2!} = \binom{n}{y_1} = \binom{n}{y_2}$$

because the multinomial coefficient is only defined when the numbers in the second row add up to number in the first row, that is, here $y_1 + y_2 = n$.

And the relation between distributions is obvious too, just because the marginals are binomial. If

$$\mathbf{Y} = \text{Multi}_2(n, \mathbf{p}),$$

then

$$Y_i = \text{Bin}(n, p_i)$$

and

$$Y_2 = n - Y_1.$$

Conversely, if

$$X \sim \text{Bin}(n, p),$$

then

$$(X, n - X) \sim \text{Multi}_2\big(n, (p, 1 - p)\big)$$

So the two-dimensional multinomial is the distribution of $(X, n - X)$ when X is binomial. Recall the conventional terminology that X is the number of "successes" in n Bernoulli "trials" and $n - X$ is the number of "failures." Either of the successes or the failures considered by themselves are binomial. When we paste them together in a two-dimensional vector, the vector is degenerate because the successes and failures sum to the number of trials, and that degenerate random vector is the two-dimensional multinomial.

5.4.6 Conditionals

Theorem 5.17. *Every conditional of a multinomial is multinomial. Suppose* $\mathbf{Y} \sim \text{Multi}_k(n, \mathbf{p})$*, then*

$$(Y_1, \ldots, Y_j) \mid (Y_{j+1}, \ldots, Y_k) \sim \text{Multi}_j(n - Y_{j+1} - \cdots - Y_k, \mathbf{q}), \tag{5.39a}$$

where

$$q_i = \frac{p_i}{p_1 + \cdots + p_j}, \qquad i = 1, \ldots, j. \tag{5.39b}$$

In words, the variables that are still random (the ones "in front of the bar") are multinomial. The number of categories is the number (here j) of such variables. The sample size is the number of observations still random, which is the original sample size minus the observations in the variables now known (the ones "behind the bar"). And the parameter vector $\mathbf{q}$ is the part of the original parameter vector corresponding to the variables in front of the bar renormalized.

Renormalized? Why are we renormalizing parameters? The parameter vector for a multinomial distribution can be thought of as a probability density (it's numbers that are nonnegative and sum to one). When we take a subvector, we need to renormalize to get another multinomial parameter vector (do what it takes to make the numbers sum to one). That's what's going on in (5.39b).

Proof of Theorem 5.17. Just calculate. The relevant marginal is the distribution of $(Y_{j+1}, \ldots, Y_k)$ but that isn't a brand name distribution. Almost as good is the marginal of

$$\mathbf{Z} = (Y_1 + \cdots + Y_j, Y_{j+1}, \ldots, Y_k) = (n - Y_{j+1} - \cdots - Y_k, Y_{j+1}, \ldots, Y_k) \quad (5.40)$$

which is $\text{Multi}_{k-j+1}(n, \mathbf{q})$ with

$$\mathbf{q} = (p_1 + \cdots + p_j, p_{j+1}, \ldots, p_k) = (n - p_{j+1} - \cdots - p_k, p_{j+1}, \ldots, p_k)$$

It's almost the same thing really, because the right hand side of (5.40) is a function of Y_{j+1}, ..., Y_k alone, hence

$$\begin{aligned} &P(Y_i = y_i, \ i = j+1, \ldots, k) \\ &\qquad = \binom{n}{n - y_{j+1} - \cdots - y_k, y_{j+1}, \ldots, y_k} \\ &\qquad\qquad \times (1 - p_{j+1} - \cdots - p_k)^{n - y_{j+1} - \cdots - y_k} p_{j+1}^{y_{j+1}} \cdots p_k^{y_k} \end{aligned}$$

And, of course, conditional equals joint over marginal

$$\begin{aligned} &\frac{\binom{n}{y_1, \ldots, y_k} p_1^{y_1} \cdots p_k^{y_k}}{\binom{n}{n - y_{j+1} - \cdots - y_k, y_{j+1}, \ldots, y_k} (1 - p_{j+1} - \cdots - p_k)^{n - y_{j+1} - \cdots - y_k} p_{j+1}^{y_{j+1}} \cdots p_k^{y_k}} \\ &\qquad = \frac{n!}{y_1! \cdots y_k!} \cdot \frac{(n - y_{j+1} - \cdots - y_k)! y_{j+1}! \cdots y_k!}{n!} \\ &\qquad\qquad \times \frac{p_1^{y_1} \cdots p_j^{y_j}}{(1 - p_{j+1} - \cdots - p_k)^{n - y_{j+1} - \cdots y_k}} \\ &\qquad = \frac{(n - y_{j+1} - \cdots - y_k)!}{y_1! \cdots y_j!} \prod_{i=1}^{j} \left(\frac{p_i}{1 - p_{j+1} - \cdots p_k} \right)^{y_j} \\ &\qquad\qquad = \binom{n - y_{j+1} - \cdots - y_k}{y_1, \ldots, y_j} \prod_{i=1}^{j} \left(\frac{p_i}{p_1 + \cdots + p_j} \right)^{y_j} \end{aligned}$$

and that's the conditional density asserted by the theorem. □

Problems

5-1. Is

$$\begin{pmatrix} 3 & 2 & -1 \\ 2 & 3 & 2 \\ -1 & 2 & 3 \end{pmatrix}$$

a covariance matrix? If not, why not?

5-2. Is

$$\begin{pmatrix} 3 & 2 & -1/3 \\ 2 & 3 & 2 \\ -1/3 & 2 & 3 \end{pmatrix}$$

a covariance matrix? If not, why not? If it is a covariance matrix, is a random vector having this covariance matrix degenerate or non-degenerate?

5-3. Consider the degenerate random vector (X, Y) in $\mathbb{R}^2$ defined by

$$\begin{aligned} X &= \sin(U) \\ Y &= \cos(U) \end{aligned}$$

where $U \sim \mathcal{U}(0, 2\pi)$. We say that (X, Y) has the uniform distribution on the unit circle. Find the mean vector and covariance matrix of (X, Y).

5-4. Let $\mathbf{M}$ be any symmetric positive semi-definite matrix, and denote its elements m_{ij}. Show that for any i and j

$$-1 \le \frac{m_{ij}}{\sqrt{m_{ii} m_{jj}}} \le 1 \tag{5.41}$$

Hint: Consider $\mathbf{w}'\mathbf{M}\mathbf{w}$ for vectors $\mathbf{w}$ having all elements zero except the i-th and j-th.

The point of the problem (this isn't part of the problem, just the explanation of why it is interesting) is that if $\mathbf{M}$ is a variance, then the fraction in (5.41) is $\operatorname{cor}(X_i, X_j)$. Thus positive semi-definiteness is a stronger requirement than the correlation inequality, as claimed in Section 5.1.4.

5-5. Show that the usual formula for the univariate normal distribution is the one-dimensional case of the formula for the multivariate normal distribution.

5-6. Show that a constant random vector (a random vector having a distribution concentrated at one point) is a (degenerate) special case of the multivariate normal distribution.

5-7. Suppose $\mathbf{X} = (X_1, \ldots, X_k)$ has the multinomial distribution with sample size n and parameter vector $\mathbf{p} = (p_1, \ldots, p_k)$, show that for $i \neq j$

$$\frac{\operatorname{var}(X_i - X_j)}{n} = p_i + p_j - (p_i - p_j)^2$$

5-8. If $\mathbf{X} \sim \mathcal{N}(0, \mathbf{M})$ is a non-degenerate normal random vector, what is the distribution of $\mathbf{Y} = \mathbf{M}^{-1}\mathbf{X}$?

5-9. Prove (5.35).
Hint: Write

$$\mathbf{X}_1 - \boldsymbol{\mu}_1 = [\mathbf{X}_1 - E(\mathbf{X}_1 \mid \mathbf{X}_2)] + [E(\mathbf{X}_1 \mid \mathbf{X}_2) - \boldsymbol{\mu}_1]$$

then use the alternate variance and covariance expressions in Theorem 5.2 and linearity of expectation.

5-10. Specialize the formula (5.24) for the non-degenerate multivariate normal density to the two-dimensional case, obtaining

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \times \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right)$$

Hint: To do this you need to know how to invert a 2×2 matrix and calculate its determinant. If

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

then

$$\det(\mathbf{A}) = a_{11}a_{22} - a_{12}a_{21}$$

and

$$\mathbf{A}^{-1} = \frac{\begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}}{\det(\mathbf{A})}$$

(This is a special case of Cramer's rule. It can also be verified by just doing the matrix multiplication. Verification of the formulas in the hint is *not* part of the problem.)

5-11. Specialize the conditional mean and variance in Theorem 5.15 to the two-dimensional case, obtaining

$$\begin{aligned} E(X \mid Y) &= \mu_X + \rho\frac{\sigma_X}{\sigma_Y}(Y - \mu_Y) \\ \operatorname{var}(X \mid Y) &= \sigma_X^2(1 - \rho^2) \end{aligned}$$

5-12 (Ellipsoids of Concentration). Suppose $\mathbf{X}$ is a non-degenerate normal random variable with density (5.24), which we rewrite as

$$f(\mathbf{x}) = \frac{e^{-q(\mathbf{x})/2}}{(2\pi)^{n/2}\det(\mathbf{M})^{1/2}}$$

A *level set* of the density, also called a *highest density region* is a set of the form

$$S = \{\, \mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) > c \,\}$$

for some constant c. Show that this can also be written

$$S = \{\, \mathbf{x} \in \mathbb{R}^n : q(\mathbf{x}) < d \,\}$$

for some other constant d. (A set like this, a level set of a positive definite quadratic form, is called an ellipsoid.) Give a formula for $P(\mathbf{X} \in S)$ as a function of the constant d in terms of the probabilities for a univariate brand name distribution. (**Hint:** Use Problem 12-32 in Lindgren.)

5-13. For the random vector $\mathbf{X}$ defined by (5.23) in Example 5.1.3 suppose U, V, and W are i. i. d. standard normal random variables.

(a) What is the joint distribution of the two-dimensional random vector whose components are the first two components of $\mathbf{X}$?

(b) What is the conditional distribution of the first component of $\mathbf{X}$ given the second?

5-14. Suppose Z_1, Z_2, ... are i. i. d. $\mathcal{N}(0, \tau^2)$ random variables and X_1, X_2, ... are defined recursively as follows.

- X_1 is a $\mathcal{N}(0, \sigma^2)$ random variable that is independent of all the Z_i.
- for $i > 1$
$$X_{i+1} = \rho X_i + Z_i.$$

There are three unknown parameters, ρ, σ^2, and τ^2, in this model. Because they are variances, we must have $\sigma^2 > 0$ and $\tau^2 > 0$. The model is called an autoregressive time series of order one or AR(1) for short. The model is said to be *stationary* if X_i has the same marginal distribution for all i.

(a) Show that the joint distribution of X_1, X_2, ..., X_n is multivariate normal.

(b) Show that $E(X_i) = 0$ for all i.

(c) Show that the model is stationary only if $\rho^2 < 1$ and
$$\sigma^2 = \frac{\tau^2}{1 - \rho^2}$$

Hint: Consider $\operatorname{var}(X_i)$.

(d) Show that
$$\operatorname{cov}(X_i, X_{i+k}) = \rho^k \sigma^2, \qquad k \geq 0$$
in the stationary model.

Chapter 6

Convergence Concepts

6.1 Univariate Theory

Chapter 5 in Lindgren is a jumble of convergence theory. Here we will follow one thread through the jumble, ignoring many of the convergence concepts discussed by Lindgren. The only ones widely used in statistics are *convergence in distribution* and its special case *convergence in probability to a constant*. We will concentrate on them.

6.1.1 Convergence in Distribution

Definition 6.1.1 (Convergence in Distribution).
A sequence of random variables X_1, X_2, ... with X_n having distribution function F_n converges in distribution *to a random variable X with distribution function F if*

$$F_n(x) \to F(x), \qquad \text{as } n \to \infty$$

for every real number x that is a continuity point of F. We indicate this by writing

$$X_n \xrightarrow{\mathcal{D}} X, \qquad \text{as } n \to \infty.$$

"Continuity point" means a point x such that F is continuous at x (a point where F does not jump). If the limiting random variable X is continuous, then every point is a continuity point. If X is discrete or of mixed type, then $F_n(x) \to F(x)$ must hold at points x where F does not jump but it does not have to hold at the jumps.

From the definition it is clear that convergence in distribution is a statement about distributions not variables. Though we write $X_n \xrightarrow{\mathcal{D}} X$, what this means is that the *distribution of* X_n converges to the *distribution of* X. We could dispense with the notion of convergence in distribution and always write $F_{X_n}(x) \to F_X(x)$ for all continuity points x of F_X in place of $X_n \xrightarrow{\mathcal{D}} X$, but that would be terribly cumbersome.

There is a much more general notion of convergence in distribution (also called *convergence in law* or *weak convergence*) that is equivalent to the concept defined in Definition 6.1.1.

Theorem 6.1 (Helly-Bray). *A sequence of random variables X_1, X_2, ... converges in distribution to a random variable X if and only if*

$$E\{g(X_n)\} \to E\{g(X)\}$$

for every bounded continuous function $g : \mathbb{R} \to \mathbb{R}$.

For comparison, Definition 6.1.1 says, when rewritten in analogous notation

$$E\{I_{(-\infty,x]}(X_n)\} \to E\{I_{(-\infty,x]}(X)\}, \qquad \text{whenever } P(X = x) = 0. \tag{6.1}$$

Theorem 6.1 doesn't explicitly mention continuity points, but the continuity issue is there implicitly. Note that

$$E\{I_A(X_n)\} = P(X_n \in A)$$

may fail to converge to

$$E\{I_A(X)\} = P(X \in A)$$

because indicator functions, though bounded, are not continuous. And (6.1) says that expectations of some indicator functions converge and others don't (at least not necessarily).

Also note that $E(X_n)$ may fail to converge to $E(X)$ because the identity function, though continuous, is unbounded. Nevertheless, the Theorem 6.1 does imply convergence of expectations of many interesting functions.

How does one establish that a sequence of random variables converges in distribution? By writing down the distribution functions and showing that they converge? No. In the common applications of convergence in distribution in statistics, convergence in distribution is a consequence of the central limit theorem or the law of large numbers.

6.1.2 The Central Limit Theorem

Theorem 6.2 (The Central Limit Theorem (CLT)). *If X_1, X_2, ... is a sequence of independent, identically distributed random variables having mean μ and variance σ^2 and*

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{6.2}$$

is the sample mean for sample size n, then

$$\sqrt{n} \left(\overline{X}_n - \mu \right) \xrightarrow{\mathcal{D}} Y, \qquad \text{as } n \to \infty, \tag{6.3}$$

where $Y \sim \mathcal{N}(0, \sigma^2)$.

It simplifies notation if we are allowed to write a distribution on the right hand side of a statement about convergence in distribution, simplifying (6.3) and the rest of the sentence following it to

$$\sqrt{n}\,(\overline{X}_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2), \qquad \text{as } n \to \infty. \tag{6.4}$$

There's nothing wrong with this mixed notation because (to repeat something said earlier) convergence in distribution is a statement about distributions of random variables, not about the random variables themselves. So when we replace a random variable with its distribution, the meaning is still clear.

The only requirement for the CLT to hold is that the variance σ^2 exist (this implies that the mean μ also exists by Theorem 2.44 of Chapter 2 of these notes. No other property of the distribution of the X_i matters.

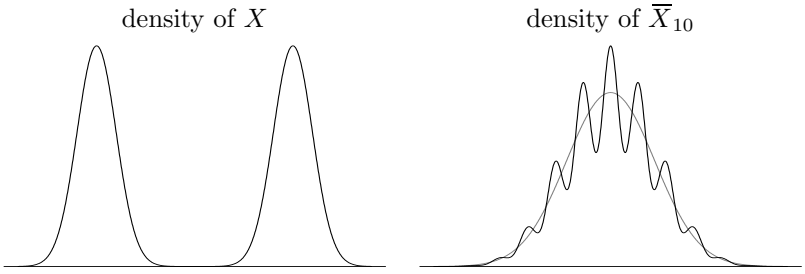
The left hand side of (6.3) always has mean zero and variance σ^2 for all n, regardless of the distribution of the X_i so long as the variance exists. Thus the central limit theorem doesn't say anything about means and variances, rather it says that the *shape* of the distribution of $\overline{X}_n$ approaches the bell-shaped curve of the normal distribution as $n \to \infty$.

A sloppy way of rephrasing (6.3) is

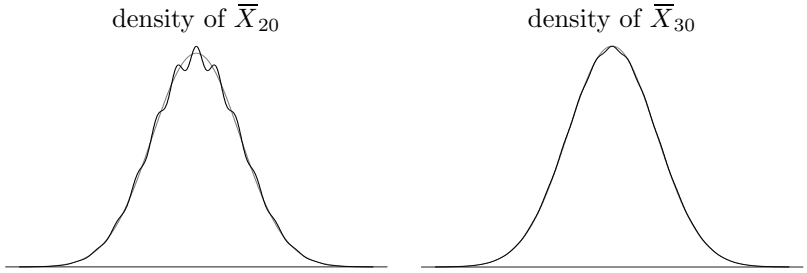
$$\overline{X}_n \approx \mathcal{N}\left(\mu, \tfrac{\sigma^2}{n}\right)$$

for "large n." Most of the time the sloppiness causes no harm and no one is confused. The mean and variance of $\overline{X}_n$ are indeed μ and σ^2/n and the shape of the distribution is approximately normal if n is large. What one cannot do is say $\overline{X}_n$ converges in distribution to Z where Z is $\mathcal{N}(\mu, \sigma^2/n)$. Having an n in the supposed limit of a sequence is mathematical nonsense.

Example 6.1.1 (A Symmetric Bimodal Distribution).
Let us take a look at how the CLT works in practice. How large does n have to be before the distribution of $\overline{X}_n$ is approximately normal?



On the left is a severely bimodal probability density function. On the right is the density of (6.2), where $n = 10$ and the X_i are i. i. d. with the density on the left. The wiggly curve is the density of $\overline{X}_{10}$ and the smooth curve is the normal density with the same mean and variance. The two densities on the right are not very close. The CLT doesn't provide a good approximation at $n = 10$.



At $n = 20$ and $n = 30$ we have much better results. The density of $\overline{X}_{30}$ is almost indistinguishable from the normal density with the same mean and variance. There is a bit of wiggle at the top of the curve, but everywhere else the fit is terrific. It is this kind of behavior that leads to the rule of thumb propounded in elementary statistics texts that $n > 30$ is "large sample" territory, thirty is practically infinity.

The symmetric bimodal density we started with in this example is of no practical importance. Its only virtue giving rise to a density for $\overline{X}_n$ that is easy to calculate. If you are not interested in the details of this example, skip to the next example. If you wish to play around with this example, varying different aspects to see what happens, go to the web page

`http://www.stat.umn.edu/geyer/5101/clt.html#bi`

The symmetric bimodal density here is the density of $X = Y + Z$, where $Y \sim \text{Ber}(p)$ and $Z \sim \mathcal{N}(0, \sigma^2)$, where $p = \frac{1}{2}$ and $\sigma = 0.1$. If Y_i and Z_i are i. i. d. sequences, then, of course

$$\sum_{i=1}^{n} Y_i \sim \text{Bin}\,(n, p)$$
$$\sum_{i=1}^{n} Z_i \sim \mathcal{N}\left(0, n\sigma^2\right)$$

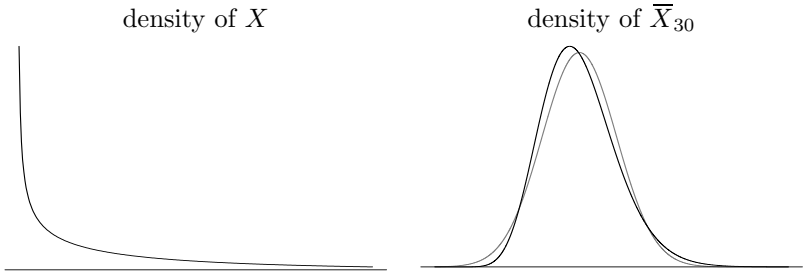
So by the convolution theorem the density of their sum is

$$f_{X_1+\cdots+X_n}(s) = \sum_{k=0}^{n} f(k \mid n, p)\phi(s - k \mid 0, n\sigma^2)$$

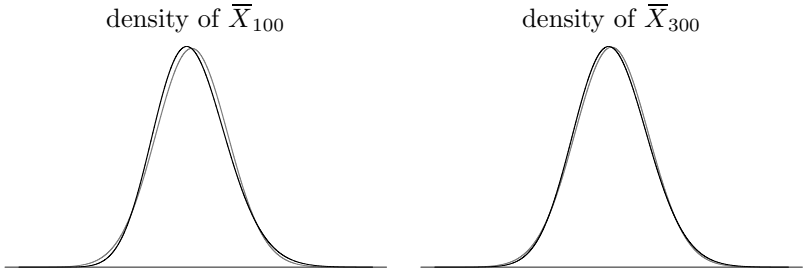
where $f(k \mid n, p)$ is the the $\text{Bin}(n, p)$ density and $\phi(z \mid \mu, \sigma^2)$ is the $\mathcal{N}(\mu, \sigma^2)$ density. The the distribution of $\overline{X}_n$ is given by

$$f_{\overline{X}_n}(w) = n f_{X_1+\cdots+X_n}(nw) = n \sum_{k=0}^{n} f(k \mid n, p)\phi(nw - k \mid 0, n\sigma^2) \tag{6.5}$$

Example 6.1.2 (A Skewed Distribution).
The $30 = \infty$ "rule" promulgated in introductory statistics texts does not hold for skewed distributions. Consider X having the chi-square distribution with one degree of freedom.



The density of X is shown on the left. It is extremely skewed going to infinity at zero. On the right is the density of $\overline{X}_{30}$ and the normal density with the same mean and variance. The fit is not good. The density of $\overline{X}_{30}$, a rescaled chi^2(30) density, is still rather skewed and so cannot be close to a normal density, which of course is symmetric.



The fit is better at $n = 100$ and $n = 300$, but still not as good as our bimodal example at $n = 30$. The moral of the story is that skewness slows convergence in the central limit theorem.

If you wish to play around with this example, varying different aspects to see what happens, go to the web page

`http://www.stat.umn.edu/geyer/5101/clt.html#expo`

6.1.3 Convergence in Probability

A special case of convergence in distribution is convergence in distribution to a degenerate random variable concentrated at one point, $X_n \xrightarrow{\mathcal{D}} a$ where a is a constant. Theorem 2 of Chapter 5 in Lindgren says that this is equivalent to the following notion.

Definition 6.1.2 (Convergence in Probability to a Constant).
A sequence of random variables X_1, X_2, ... converges in probability *to a constant a if for every $\epsilon > 0$*

$$P(|X_n - a| > \epsilon) \to 0, \qquad \text{as } n \to \infty.$$

We indicate X_n converging in probability to a by writing

$$X_n \xrightarrow{P} a, \qquad \text{as } n \to \infty.$$

Convergence in probability to a constant and convergence in distribution to a constant are the same thing, so we could write $X_n \xrightarrow{\mathcal{D}} a$ instead of $X_n \xrightarrow{P} a$, but the latter is traditional. There is also a more general notion of convergence in probability to a *random variable*, but it has no application in statistics and we shall ignore it.

6.1.4 The Law of Large Numbers

One place convergence in probability appears is in the law of large numbers.

Theorem 6.3 (Law of Large Numbers (LLN)). *If X_1, X_2, ... is a sequence of independent, identically distributed random variables having mean μ, and*

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

is the sample mean for sample size n, then

$$\overline{X}_n \xrightarrow{P} \mu, \qquad \text{as } n \to \infty. \tag{6.6}$$

The only requirement is that the mean μ exist. No other property of the distribution of the X_i matters.

6.1.5 The Continuous Mapping Theorem

Theorem 6.4 (Continuous Mapping). *If g is a function continuous at all points of a set A, if $X_n \xrightarrow{\mathcal{D}} X$, and if $P(X \in A) = 1$, then $g(X_n) \xrightarrow{\mathcal{D}} g(X)$.*

The main point of the theorem is the following two corollaries.

Corollary 6.5. *If g is an everywhere continuous function and $X_n \xrightarrow{\mathcal{D}} X$, then $g(X_n) \xrightarrow{\mathcal{D}} g(X)$.*

Here the set A in the theorem is the whole real line. Hence the condition $P(X \in A) = 1$ is trivial.

Corollary 6.6. *If g is a function continuous at the point a and $X_n \xrightarrow{P} a$, then $g(X_n) \xrightarrow{P} g(a)$.*

Here the set A in the theorem is just the singleton set $\{a\}$, but the limit variable in question is the constant random variable satisfying $P(X = a) = 1$.

These theorems say that convergence in distribution and convergence in probability to a constant behave well under a continuous change of variable.

Rewriting the CLT

The CLT can be written in a variety of slightly different forms. To start, let us rewrite (6.3) as

$$\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow{\mathcal{D}} \sigma Z, \qquad \text{as } n \to \infty,$$

where now Z is a standard normal random variable. If $\sigma > 0$, then we can divide both sides by σ. This is a simple application of the continuous mapping theorem, the function defined by $g(x) = x/\sigma$ being continuous. It gives

$$\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \xrightarrow{\mathcal{D}} Z$$

Moving the $\sqrt{n}$ from the numerator to the denominator of the denominator gives

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{D}} Z \tag{6.7}$$

6.1.6 Slutsky's Theorem

Theorem 6.7 (Slutsky). *If $g(x, y)$ is a function jointly continuous at every point of the form (x, a) for some fixed a, and if $X_n \xrightarrow{\mathcal{D}} X$ and $Y_n \xrightarrow{P} a$, then*

$$g(X_n, Y_n) \xrightarrow{\mathcal{D}} g(X, a).$$

Corollary 6.8. *If $X_n \xrightarrow{\mathcal{D}} X$ and $Y_n \xrightarrow{P} a$, then*

$$X_n + Y_n \xrightarrow{\mathcal{D}} X + a,$$
$$Y_n X_n \xrightarrow{\mathcal{D}} aX,$$

and if $a \neq 0$

$$X_n / Y_n \xrightarrow{\mathcal{D}} X/a.$$

In other words, we have all the nice properties we expect of limits, the limit of a sum is the sum of the limits, and so forth. The point of the theorem is this is *not true* unless one of the limits is a constant. If we only had $X_n \xrightarrow{\mathcal{D}} X$ and $Y_n \xrightarrow{\mathcal{D}} Y$, we couldn't say anything about the limit of $X_n + Y_n$ without knowing about the *joint* distribution of X_n and Y_n. When Y_n converges to a constant, Slutsky's theorem tells us that we don't need to know anything about joint distributions.

A special case of Slutsky's theorem involves two sequences converging in probability. If $X_n \xrightarrow{P} a$ and $Y_n \xrightarrow{P} b$, then $X_n + Y_n \xrightarrow{P} a + b$, and so forth. This is a special case of Slutsky's theorem because convergence in probability to a constant is the same as convergence in distribution to a constant.

6.1.7 Comparison of the LLN and the CLT

When X_1, X_2, ... is an i. i. d. sequence of random variables having a variance, both the law of large numbers and the central limit theorem apply, but the CLT tells us much more than the LLN.

It could not tell us less, because the CLT implies the LLN. By Slutsky's theorem, the CLT (6.3) implies

$$\overline{X}_n - \mu = \frac{1}{\sqrt{n}} \cdot \sqrt{n} \left(\overline{X}_n - \mu \right) \xrightarrow{\mathcal{D}} 0 \cdot Y = 0$$

where $Y \sim \mathcal{N}(0, \sigma^2)$. Because convergence in distribution to a constant and convergence in probability to a constant are the same thing, this implies the LLN.

But the CLT gives much more information than the LLN. It says that the size of the estimation error $\overline{X}_n - \mu$ is about $\sigma/\sqrt{n}$ and also gives us the *shape* of the error distribution (i. e., normal).

So why do we even care about the law of large numbers? Is it because there are lots of important probability models having a mean but no variance (so the LLN holds but the CLT does not)? No, not any used for real data. The point is that sometimes we don't care about the information obtained from the central limit theorem. When the only fact we want to use is $\overline{X}_n \xrightarrow{P} \mu$, we refer to the law of large numbers as our authority. Its statement is simpler, and there is no point in dragging an unnecessary assumption about variance in where it's not needed.

6.1.8 Applying the CLT to Addition Rules

The central limit theorem says that the sum of i. i. d. random variables with a variance is approximately normally distributed if the number of variables in the sum is "large." Applying this to the addition rules above gives several normal approximations.

Binomial The $\text{Bin}(n, p)$ distribution is approximately normal with mean np and variance $np(1 - p)$ if n is large.

Negative Binomial The $\text{NegBin}(n, p)$ distribution is approximately normal with mean n/p and variance $n(1 - p)/p^2$ if n is large.

Poisson The $\text{Poi}(\mu)$ distribution is approximately normal with mean μ and variance μ if μ is large.

Gamma The $\text{Gam}(\alpha, \lambda)$ distribution is approximately normal with mean α/λ and variance α/λ^2 if α is large.

Chi-Square The $\text{chi}^2(n)$ distribution is approximately normal with mean n and variance $2n$ if n is large.

Comment The rules containing n are obvious combinations of the relevant addition rule and the CLT. The rules for the Poisson and gamma distributions are a bit weird in that there is no n. To understand them we need the notion of an infinitely divisible distribution.

Definition 6.1.3.
*A probability distribution P is **infinitely divisible** if for every positive integer n there exist independent and identically distributed random variables X_1, ..., X_n such that $X_1 + \cdots + X_n$ has the distribution P.*

Example 6.1.3 (Infinite Divisibility of the Poisson).
By the addition rule for Poisson random variables, $X_1 + \cdots + X_n \sim \text{Poi}(\mu)$ when the X_i are i. i. d. $\text{Poi}(\mu/n)$. Thus the $\text{Poi}(\mu)$ distribution is infinitely divisible for any $\mu > 0$.

Example 6.1.4 (Infinite Divisibility of the Gamma).
By the addition rule for gamma random variables, $X_1 + \cdots + X_n \sim \text{Gam}(\alpha, \lambda)$ when the X_i are i. i. d. $\text{Gam}(\alpha/n, \lambda)$. Thus the $\text{Gam}(\alpha, \lambda)$ distribution is infinitely divisible for any $\alpha > 0$ and $\lambda > 0$.

The infinite divisibility of the Poisson and gamma distributions explains the applicability of the CLT. But we have to be careful. Things are not quite as simple as they look.

A Bogus Proof that Poisson is Normal Every Poisson random variable is the sum of n i. i. d. random variables and n can be chosen as large as we please. Thus by the CLT the Poisson distribution is arbitrarily close to normal. Therefore it is normal.

Critique of the Bogus Proof For one thing, it is obviously wrong. The Poisson discrete is discrete. The Normal distribution is continuous. They can't be equal. But what's wrong with the proof?

The problem is in sloppy application of the CLT. It is often taken to say what the bogus proof uses, and the sloppy notation (6.4) encourages this sloppy use, which usually does no harm, but is the problem here.

A more careful statement of the CLT says that for any fixed μ and large enough n the $\text{Poi}(n\mu)$ distribution is approximately normal. The n that is required to get close to normal depends on μ. This does tell us that for sufficient large values of the parameter, the Poisson distribution is approximately normal. It does *not* tell us the Poisson distribution is approximately normal for *any* value of the parameter, which the sloppy version seems to imply.

The argument for the gamma distribution is exactly analogous to the argument for the Poisson. For large enough values of the parameter α involved in the infinite divisibility argument, the distribution is approximately normal. The statement about the chi-square distribution is a special case of the statement for the gamma distribution.

6.1.9 The Cauchy Distribution

The Cauchy location-scale family, abbreviated Cauchy(μ, σ) is described in Section B.2.7 of Appendix B an addition rule given by (C.11) in Appendix C, which we repeat here

$$X_1 + \cdots + X_n \sim \text{Cauchy}(n\mu, n\sigma) \tag{6.8}$$

from which we can derive the distribution of the sample mean

$$\overline{X}_n \sim \text{Cauchy}(\mu, \sigma) \tag{6.9}$$

(Problem 6-1).

The Cauchy family is not a useful model for real data, but it is theoretically important as a source of counterexamples. A Cauchy(μ, σ) distribution has center of symmetry μ. Hence μ is the median, but μ is not the mean because the mean does not exist.

The rule for the mean (6.9) can be trivially restated as a convergence in distribution result

$$\overline{X}_n \xrightarrow{\mathcal{D}} \text{Cauchy}(\mu, \sigma), \qquad \text{as } n \to \infty \tag{6.10}$$

a "trivial" result because $\overline{X}_n$ actually has exactly the Cauchy(μ, σ) distribution for all n, so the assertion that is gets close to that distribution for large n is trivial (exactly correct is indeed a special case of "close").

The reason for stating (6.10) is for contrast with the law of large numbers (LLN), which can be stated as follows: if X_1, X_2, ... are i. i. d. from a distribution with mean μ, then

$$\overline{X}_n \xrightarrow{P} \mu \qquad \text{as } n \to \infty \tag{6.11}$$

The condition for the LLN, that the mean exist, does not hold for the Cauchy. Furthermore, since μ does not exist, $\overline{X}_n$ cannot converge to it. But it is conceivable that

$$\overline{X}_n \xrightarrow{P} c \qquad \text{as } n \to \infty \tag{6.12}$$

for some constant c, even though this does not follow from the LLN. The result (6.10) for the Cauchy rules this out. Convergence in probability to a constant is the same as convergence in distribution to a constant (Theorem 2 of Chapter 5 in Lindgren). Thus (6.12) and (6.10) are contradictory. Since (6.10) is correct, (6.12) must be wrong. For the Cauchy distribution $\overline{X}_n$ does not converge in probability to anything.

Of course, the CLT also fails for the Cauchy distribution. The CLT implies the LLN. Hence if the CLT held, the LLN would also hold. Since the LLN doesn't hold for the Cauchy, the CLT can't hold either.

Problems

6-1. Derive (6.9) from (6.8) using the change of variable theorem.

6-2. Suppose that S_1, S_2, ... is any sequence of random variables such that $S_n \xrightarrow{P} \sigma$, and X_1, X_2, ... are independent and identically distributed with mean μ and variance σ^2 and $\sigma > 0$. Show that

$$\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1), \qquad \text{as } n \to \infty,$$

where, as usual,

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

6-3. Suppose X_1, X_2, ... are i. i. d. with common probability measure P, and define $Y_n = I_A(X_n)$ for some event A, that is,

$$Y_n = \begin{cases} 1, & X_n \in A \\ 0, & X_n \notin A \end{cases}$$

Show that $\overline{Y}_n \xrightarrow{P} P(A)$.

6-4. Suppose the sequences X_1, X_2, ... and Y_1, Y_2, ... are defined as in Problem 6-3, and write $P(A) = p$. Show that

$$\sqrt{n}(\overline{Y}_n - p) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, p(1-p)\big)$$

and also show that

$$\frac{\overline{Y}_n - p}{\sqrt{\overline{Y}_n(1-\overline{Y}_n)/n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

Hint: What is the distribution of $\sum_i Y_i$? Also use Problem 6-2.

Chapter 7

Sampling Theory

7.1 Empirical Distributions

In statistics, we often deal with complicated data, but for learning it is best to start simple. The simplest sort of data is just a set of numbers that are measurements of one variable on a set of individuals. In the next section we will see that it is important that these individuals be a random sample from some population of interest. For now we will just treat the data as a set of numbers.

Example 7.1.1 (A Data Set).
The numbers below were generated by computer and are a random sample from an Exp(1) distribution rounded to one significant figure. Because of the rounding, there are duplicate values. If not rounded the values would all be different, as would be the case for any sample from any continuous distribution.

$$0.12 \quad 3.15 \quad 0.77 \quad 1.02 \quad 0.08 \quad 0.35 \quad 0.29 \quad 1.05 \quad 0.49 \quad 0.81$$

A vector

$$\mathbf{x} = (x_1, \ldots, x_n) \tag{7.1}$$

can be thought of as a function of the index variable i. To indicate this we can write the components as $x(i)$ instead of x_i. Then x is a function on the index set $\{1, \ldots, n\}$. Sometimes we don't even bother to change the notation but still think of the vector as being the function $i \mapsto x_i$.

This idea is useful in probability theory because of the dogma "a random variable is a function on the sample space." So let us think of the index set $S = \{1, \ldots, n\}$ as the sample space, and X as a random variable having values $X(i)$, also written x_i. When we consider a uniform distribution on the sample space, which means each point gets probability $1/n$ since there are n points, then the distribution of X is called the *empirical distribution* associated with the vector (7.1).

By definition, the probability function of X is

$$f(x) = P(X = x) = \sum_{\substack{i \in S \\ x_i = x}} \frac{1}{n} = \frac{\text{card}(\{\, i \in S : x_i = x \,\})}{n}$$

where, as usual, $\text{card}(A)$ denotes the *cardinality* of the set A. If all of the x_i are distinct, then the distribution of X is also uniform. Otherwise, it is not. If the point x occurs m times among the x_i, then $f(x) = m/n$. This makes the definition of the empirical distribution in terms of its probability function rather messy. So we won't use it.

The description in terms of expectation is much simpler.

Definition 7.1.1 (Empirical Expectation).
The empirical expectation operator associated with the vector $(x_1, \ldots, x_n)$ is denoted E_n and defined by

$$E_n\{g(X)\} = \frac{1}{n} \sum_{i=1}^{n} g(x_i). \tag{7.2}$$

Example 7.1.2.
For the data in Example 7.1.1 we have for the function $g(x) = x$

$$E_n(X) = \frac{1}{n} \sum_{i=1}^{n} x_i = 0.813$$

and for the function $g(x) = x^2$

$$E_n(X^2) = \frac{1}{n} \sum_{i=1}^{n} x_i^2 = 1.37819$$

The corresponding probability measure P_n is found by using "probability is just expectation of indicator functions."

Definition 7.1.2 (Empirical Probability Measure).
The empirical probability measure associated with the vector $(x_1, \ldots, x_n)$ is denoted P_n and defined by

$$P_n(A) = \frac{1}{n} \sum_{i=1}^{n} I_A(x_i). \tag{7.3}$$

Example 7.1.3.
For the data in Example 7.1.1 we have for the event $X > 2$

$$P_n(X > 2) = \frac{1}{n} \sum_{i=1}^{n} I_{(2,\infty)}(x_i) = \frac{\text{number of } x_i \text{ greater than } 2}{n} = 0.1$$

and for the event $1 < X < 2$

$$P_n(1 < X < 2) = \frac{1}{n} \sum_{i=1}^{n} I_{(1,2)}(x_i) = \frac{\text{number of } x_i \text{ between 1 and 2}}{n} = 0.2$$

7.1.1 The Mean of the Empirical Distribution

For the rest of this section we consider the special case in which all of the x_i are real numbers.

The *mean* of the empirical distribution is conventionally denoted by $\bar{x}_n$ and is obtained by taking the case $g(x) = x$ in (7.2)

$$\bar{x}_n = E_n(X) = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

7.1.2 The Variance of the Empirical Distribution

The variance of the empirical distribution has no conventional notation, but we will use both $\operatorname{var}_n(X)$ and v_n. Just like any other variance, it is the expected squared deviation from the mean. The mean is $\bar{x}_n$, so

$$v_n = \operatorname{var}_n(X) = E_n\{(X - \bar{x}_n)^2\} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \tag{7.4}$$

It is important that you think of the empirical distribution as a probability distribution just like any other. This gives us many facts about empirical distributions, that are derived from general facts about probability and expectation. For example, the parallel axis theorem holds, just as it does for any probability distribution. For ease of comparison, we repeat the general parallel axis theorem (Theorem 2.11 of Chapter 2.27 of these notes).

If X is a real-valued random variable having finite variance and a is any real number, then

$$E\{(X - a)^2\} = \operatorname{var}(X) + [a - E(X)]^2 \tag{7.5}$$

Corollary 7.1 (Empirical Parallel Axis Theorem).

$$E_n\{(X - a)^2\} = \operatorname{var}_n(X) + [a - E_n(X)]^2$$

or, in other notation,

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - a)^2 = v_n + (a - \bar{x}_n)^2 \tag{7.6}$$

In particular, the case $a = 0$ gives the empirical version of

$$\operatorname{var}(X) = E(X^2) - E(X)^2$$

which is

$$\operatorname{var}_n(X) = E_n(X^2) - E_n(X)^2$$

or, in other notation,

$$v_n = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}_n^2. \tag{7.7}$$

Example 7.1.4.
In Example 7.1.2 we found for the data in Example 7.1.1

$$\bar{x}_n = E_n(X) = 0.813$$

and

$$E_n(X^2) = 1.37819.$$

Although we could use the definition (7.4) directly, we can also use the empirical parallel axis theorem in the form (7.7)

$$v_n = 1.37819 - 0.813^2 = 0.717221.$$

7.1.3 Characterization of the Mean

Considering a as a variable in (7.5) or (7.6) gives the following pair of theorems. The first one is just the corollary to the parallel axis theorem in Lindgren (p. 107) in different language. It is also the special case of the characterization of conditional expectation as best prediction (Theorem 3.6 in Chapter 3 of these notes) when the conditional expectation is actually unconditional.

Corollary 7.2 (Characterization of the Mean). *The mean of a real-valued random variable X having finite variance is the value of a that minimizes the function*

$$g(a) = E\{(X - a)^2\}$$

which is the expected squared deviation from a.

Corollary 7.3 (Characterization of the Empirical Mean). *The mean of the empirical distribution is the value of a that minimizes the function*

$$g(a) = E_n\{(X - a)^2\} = \frac{1}{n} \sum_{i=1}^{n} (x_i - a)^2$$

which is the average squared deviation from a.

The point of these two corollaries is that they describe the sense in which the mean is the "center" of a distribution. It is the point to which all other points are closest on average, when "close" is defined in terms of squared differences. The mean is the point from which the average squared deviation is the smallest. We will contrast this characterization with an analogous characterization of the median in Section 7.1.7.

7.1.4 Review of Quantiles

Recall from Section 3.2 in Lindgren the definition of a quantile of a probability distribution.

Definition 7.1.3 (Quantile).
For $0 < p < 1$, a point x is a p-th quantile *of the distribution of a real-valued random variable X if*

$$P(X \le x) \ge p \quad \text{and} \quad P(X \ge x) \ge 1 - p$$

If the c. d. f. of X is invertible, then there is a much simpler characterization of quantiles. For $0 < p < 1$, the p-th quantile is the unique solution x of the equation

$$F(x) = p, \tag{7.8a}$$

or in other notation

$$x = F^{-1}(p). \tag{7.8b}$$

The following lemma tells us we are usually in this situation when dealing with continuous random variables

Lemma 7.4. *A continuous random variable having a strictly positive p. d. f. has an invertible c. d. f.*

Proof. There exists a solution to (7.8a), by the intermediate value theorem from calculus, because F is continuous and goes from zero to one as x goes from $-\infty$ to $+\infty$. The solution is unique because

$$F(x + h) = F(x) + \int_x^{x+h} f(x)\, dx$$

and the integral is not zero unless $h = 0$, because the integral of a strictly positive function cannot be zero. □

In general, the p-th quantile need not be unique and it need not be a point satisfying $F(x) = p$ (see Figure 3.3 in Lindgren for examples of each of these phenomena). Hence the technical fussiness of Definition 7.1.3. That definition can be rephrased in terms of c. d. f.'s as follows. A point x is a p-th quantile of a random variable with c. d. f. F if

$$F(x) \ge p \qquad \text{and} \qquad F(y) \le p, \quad \text{for all } y < x$$

Here the asymmetry of the definition of c. d. f.'s (right continuous but not necessarily left continuous) makes the two conditions asymmetric. Definition 7.1.3 makes the symmetry between left and right clear. If x is a p-th quantile of X, then $-x$ is also a q-th quantile of $-X$, where $q = 1 - p$.

7.1.5 Quantiles of the Empirical Distribution

Now we want to look at the quantiles of the empirical distribution associated with a vector $\mathbf{x}$. In order to discuss this, it helps to establish the following notation. We denote the sorted values of the components of $\mathbf{x}$ by

$$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}.$$

That is, when we put parentheses around the subscripts, that means we have put the values in ascending order. For any real number x, the notation $\lceil x \rceil$ denotes the smallest integer greater than or equal to x, which is called the *ceiling* of x, and the notation $\lfloor x \rfloor$ denotes the largest integer less than or equal to x, which is called the *floor* of x,

Theorem 7.5. *If np is not an integer, then the p-th quantile of the empirical distribution associated with the vector $\mathbf{x}$ is unique and is equal to $x_{(\lceil np \rceil)}$.*

When np is an integer, then any point x such that

$$x_{(np)} \le x \le x_{(np+1)} \tag{7.9}$$

is a p-th quantile.

Proof. The p-th quantile must be a point x such that there are at least np of the x_i at or below x and at least $n(1-p)$ at or above x.

In the case that np is not an integer, let $k = \lceil np \rceil$. Since np is not an integer, and $\lceil np \rceil$ is the least integer greater than k, we have $k > np > k - 1$. What we must show is that $x_{(k)}$ is the unique p-th quantile.

There are at least $k > np$ data points

$$x_{(1)} \le \cdots \le x_{(k)}$$

at or below $x_{(k)}$. Furthermore, if $i < k$, then $i \le k - 1 < np$ so there are fewer than np data points at or below $x_{(i)}$ unless $x_{(i)}$ happens to be equal to $x_{(k)}$.

Similarly, there are at least $n - k + 1 > n(1-p)$ data points

$$x_{(k)} \le \cdots \le x_{(n)}$$

at or above $x_{(k)}$. Furthermore, if $i > k$, then $n - i + 1 \le n - k < n(1-p)$ so there are fewer than $n(1-p)$ data points at or above $x_{(i)}$ unless $x_{(i)}$ happens to be equal to $x_{(k)}$.

In the case $np = k$, let x be any point satisfying (7.9). Then there are at least $k = np$ data points

$$x_{(1)} \le \cdots \le x_{(k)} \le x$$

at or below x, and there are at least $n - k = n(1-p)$ data points

$$x \le x_{(k+1)} \le \cdots \le x_{(n)}$$

at or above x. Hence x is a p-th quantile. □

Example 7.1.5.
The data in Example 7.1.1 have 10 data points. Thus by the theorem, the empirical quantiles are uniquely defined when np is not an integer, that is, when p is not a multiple of one-tenth.

The first step in figuring out empirical quantiles is always to *sort the data*. Don't forget this step. The sorted data are

0.08 0.12 0.29 0.35 0.49 0.77 0.81 1.02 1.05 3.15

To find the 0.25 quantile, also called the 25-th percentile, the theorem says we find $\lceil np \rceil$, which is the integer above $np = 2.5$, which is 3, and then the empirical quantile is the corresponding order statistic, that is $x_{(3)} = 0.29$.

We remark in passing that if the 25-th percentile is 3 in from the left end of the data in *sorted order*, then the 75-th percentile is 3 in from the right end, so the definition behaves as we expect. Let's check this. First $np = 7.5$. Rounding up gives 8. And $x_{(8)} = 1.02$ is indeed the third from the right.

The definition gets tricky is when np is an integer. If we want the 40-th percentile, $np = 4$. Then the theorem says that any point x between $x_{(4)} = 0.35$ and $x_{(5)} = 0.49$ is a 40-th percentile (0.4 quantile) of these data. For example, 0.35, 0.39, 0.43, and 0.49 are all 40-th percentiles. A bit weird, but that's how the definition works.

7.1.6 The Empirical Median

The median of the empirical distribution we denote by $\tilde{x}_n$. It is the p-th quantile for $p = 1/2$. By the theorem, the median is unique when np is not an integer, which happens whenever n is an odd number. When n is an even number, the empirical median is not unique. It is any point x satisfying (7.9), where $k = n/2$. This nonuniqueness is unsettling to ordinary users of statistics, so a convention has grown up of taking the empirical median to be the midpoint of the interval given by (7.9).

Definition 7.1.4 (Empirical Median).
The median *of the values $x_1, \ldots, x_n$ is the middle value in sorted order when n is odd*

$$\tilde{x}_n = x_{(\lceil n/2 \rceil)}$$

and the average of the two middle values when n is even

$$\tilde{x}_n = \frac{x_{(n/2)} + x_{(n/2+1)}}{2}$$

Example 7.1.6.
The data in Example 7.1.1 have 10 data points. So we are in the "n even" case, and the empirical median is the average of the two middle values of the data *in sorted order*, that is,

$$\tilde{x}_n = \frac{x_{(5)} + x_{(6)}}{2} = \frac{0.49 + 0.77}{2} = 0.63$$

7.1.7 Characterization of the Median

Corollary 7.6 (Characterization of the Median). *If X is a real-valued random variable having finite expectation, then a median of X is any value of a that minimizes the function*

$$g(a) = E\{|X - a|\}$$

which is the expected absolute deviation from a.

Proof. What we need to show is that if m is a median, that is, if

$$P(X \le m) \ge \tfrac{1}{2} \qquad \text{and} \qquad P(X \ge m) \ge \tfrac{1}{2}$$

and a is any real number, then

$$E(|X - a|) \ge E(|X - m|).$$

Without loss of generality, we may suppose $a > m$. (The case $a = m$ is trivial. The case $a < m$ follows from the other case by considering the distribution of $-X$.)

Define

$$g(x) = |x - a| - |x - m|$$

so, by linearity of expectation,

$$E(|X - a|) - E(|X - m|) = E(|X - a| - |X - m|) = E\{g(X)\}$$

So what must be shown is that $E\{g(X)\} \ge 0$.

When $x \le m < a$,

$$g(x) = (a - x) - (m - x) = a - m.$$

Similarly, when $m < a \le x$,

$$g(x) = -(a - m).$$

When $m < x < a$,

$$g(x) = (x - a) - (m - x) = 2(x - m) - (a - m) \ge -(a - m).$$

Thus $g(x) \ge h(x)$ for all x, where

$$h(x) = \begin{cases} a - m, & x \le m \\ -(a - m), & x > m \end{cases}$$

The point is that h can be written in terms of indicator functions

$$h(x) = (a - m)\big[I_{(-\infty, m]}(x) - I_{(m, +\infty)}(x)\big]$$

so by monotonicity of expectation, linearity of expectation, and "probability is expectation of indicator functions"

$$E\{g(X)\} \ge E\{h(X)\} = (a - m)\big[P(X \le m) - P(X > m)\big]$$

Because m is a median, the quantity in the square brackets is nonnegative. □

Corollary 7.7 (Characterization of the Empirical Median). *A median of the empirical distribution is a value of a that minimizes the function*

$$g(a) = E_n\{|X - a|\} = \frac{1}{n} \sum_{i=1}^{n} |x_i - a| \tag{7.10}$$

which is the average absolute deviation from a.

There is no end to this game. Every notion that is defined for general probability models, we can specialize to empirical distributions. We can define empirical moments and central moments of all orders, and so forth and so on. But we won't do that in gory detail. What we've done so far is enough for now.

7.2 Samples and Populations

7.2.1 Finite Population Sampling

It is common to apply statistics to a *sample* from a *population*. The population can be any finite set of individuals. Examples are the population of Minnesota today, the set of registered voters in Minneapolis on election day, the set of wolves in Minnesota. A sample is any subset of the population. Examples are the set of voters called by an opinion poll and asked how they intend to vote, the set of wolves fitted with radio collars for a biological experiment. By convention we denote the population size by N and the sample size by n. Typically n is much less than N. For an opinion poll, n is typically about a thousand, and N is in the millions.

Random Sampling

A *random sample* is one drawn so that every individual in the population is equally likely to be in the sample. There are two kinds.

Sampling without Replacement The model for sampling without replacement is dealing from a well-shuffled deck of cards. If we deal n cards from a deck of N cards, there are $(N)_n$ possible outcomes, all equally likely (here we are considering that the order in which the cards are dealt matters). Similarly there are $(N)_n$ possible samples without replacement of size n from a population of size N. If the samples are drawn in such a way that all are equally likely we say we have a *random sample without replacement* from the population.

Sampling with Replacement The model for sampling with replacement is spinning a roulette wheel. If we do n spins and the wheel has N pockets, there are N^n possible outcomes, all equally likely. Similarly there are N^n possible samples with replacement of size n from a population of size N. If the samples are drawn in such a way that all are equally likely we say we have a *random sample with replacement* from the population.

Lindgren calls this a *simple random sample*, although there is no standard meaning of the word "simple" here. Many statisticians would apply "simple" to sampling either with or without replacement using it to mean that all samples are equally likely in contrast to more complicated sampling schemes in which the samples are not all equally likely.

Random Variables

Suppose we are interested in a particular variable, which in principle could be measured for each individual in the population. Write the vector of population values

$$\mathbf{x} = (x_1, \ldots, x_N).$$

Sometimes when x is the only variable of interest we think of this collection of x values as being the population (as opposed to the population being the collection of individuals on whom these measurements could be made).

The vector of population values is *not* a random vector.[1] The population is what it is, and the value x_i for the i-th individual of the population is what it is. Because $\mathbf{x}$ is not random, we use a lower case letter, following the "big X" for random and "little x" for nonrandom convention.

When we take a random sample of size n from the population we obtain a sequence X_1, ..., X_n of values of the variable. Each sample value X_i is one of the population values x_j, but which one is random. That is why we use capital letters for the sample values. When we think of the sample as one thing rather than n things, it is a vector

$$\mathbf{X} = (X_1, \ldots, X_n).$$

Thus we can talk about the probability distributions of each X_i and the joint distribution of all the X_i, which is the same thing as the distribution of the random vector $\mathbf{X}$.

Theorem 7.8 (Sampling Distributions). *If X_1, ..., X_n are a random sample from a population of size n, then the marginal distribution of each X_i is the empirical distribution associated with the population values x_1, ..., x_N.*

If the sampling is with replacement, then the X_i are independent and identically distributed. If the sampling is without replacement, then the X_i are exchangeable but not independent.

Proof. The X_i are exchangeable by definition: every permutation of the sample is equally likely. Hence they are identically distributed, and the marginal distribution of the X_i is the marginal distribution of X_1. Since every individual is equally likely to be the first one drawn, X_1 has the empirical distribution.

Under sampling with replacement, every sample has probability $1/N^n$, which is the product of the marginals. Hence the X_i are independent random variables. Under sampling without replacement, every sample has probability $1/(N)_n$, which is not the product of the marginals. Hence the X_i are dependent random variables. □

Thus, when we have sampling with replacement, we can use formulas that require independence, the most important of these being

[1] When we get to the chapter on Bayesian inference we will see that this sentence carries unexamined philosophical baggage. A Bayesian would say the population values are random too. But we won't worry about that for now.

- the variance of a sum is the sum of the variances

$$\operatorname{var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{var}(X_i) = n\sigma^2 \tag{7.11}$$

  where we have written σ^2 for the variance of all of the X_i (they must have the same variance because they are identically distributed), and

- the joint density is the product of the marginals

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} f_{X_i}(x_i) = \prod_{i=1}^{n} f(x_i) \tag{7.12}$$

  where we have written f for the marginal density of all of the X_i (they must have the same density because they are identically distributed).

When we have sampling without replacement *neither* (7.11) nor (7.12) holds. The analog of (7.11) is derived as follows.

Theorem 7.9 (Finite Population Correction). *If X_1, X_2, ..., X_n are a random sample without replacement from a finite population of size N, then all the X_i have the same variance σ^2 and*

$$\operatorname{var}\left(\sum_{i=1}^{n} X_i\right) = n\sigma^2 \cdot \frac{N-n}{N-1} \tag{7.13}$$

The factor $(N-n)/(N-1)$ by which (7.13) differs from (7.11) is called the *finite population correction*.

Proof. Since the X_i are exchangeable, each X_i has the same variance σ^2 and each pair X_i and X_j has the same correlation ρ. Thus

$$\begin{aligned} \operatorname{var}\left(\sum_{i=1}^{n} X_i\right) &= \sum_{i=1}^{n} \sum_{j=1}^{n} \operatorname{cov}(X_i, X_j) \\ &= n\sigma^2 + n(n-1)\sigma^2\rho \\ &= n\sigma^2 \left[1 + (n-1)\rho\right] \end{aligned} \tag{7.14}$$

The correlation ρ does not depend on the sample size, because by exchangeability it is the correlation of X_1 and X_2, and the marginal distribution of these two individuals does not depend on what happens after they are drawn. Therefore (a trick!) we can determine ρ by looking at the special case when $N = n$, when the sample is the whole population and

$$\sum_{i=1}^{n} X_i = \sum_{i=1}^{N} x_i$$

is not random (as is clear from the "little x" notation on the right hand side). Hence when $N = n$ the variance is zero, and we must have

$$1 + (N - 1)\rho = 0$$

which, solving for ρ, implies

$$\rho = -\frac{1}{N - 1}$$

Plugging this into (7.14) gives (7.13). □

7.2.2 Repeated Experiments

If X_1, ..., X_n are the outcomes of a series of random experiments which are absolutely identical and have nothing to do with each other, then they are independent and identically distributed, a phrase so widely used in statistics that its abbreviation i. i. d. is universally recognized.

This situation is analogous to sampling with replacement in that the variables of interest are i. i. d. and all the consequences of the i. i. d. property, such as (7.11) and (7.12), hold. The situation is so analogous that many people use the language of random sampling to describe this situation too. Saying that X_1, ..., X_n are a random sample from a hypothetical infinite population. There is nothing wrong with this so long as everyone understands it is only an analogy. There is no sense in which i. i. d. random variables actually are a random sample from some population.

We will use the same language. It lends color to otherwise dry and dusty discussions if you imagine we are sampling a population to answer some interesting question. That may lead us into some language a pedant would call sloppy, such as, "suppose we have a sample of size n from a population with finite variance." If the population is finite, then it automatically has a finite variance. If the population is infinite, then the variance is not really defined, since infinite populations don't exist except as a vague analogy. What is meant, of course, is "suppose X_1, ..., X_n are i. i. d. and have finite variance." That's well defined.

7.3 Sampling Distributions of Sample Moments

7.3.1 Sample Moments

If X_1, ..., X_n are a random sample, the *sample moments* are the moments of the empirical distribution associated with the vector $\mathbf{X} = (X_1, \ldots, X_n)$. The first moment is the *sample mean*

$$\overline{X}_n = E_n(X) = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{7.15}$$

The k-th moment is

$$A_{k,n} = E_n(X^k) = \frac{1}{n} \sum_{i=1}^{n} X_i^k.$$

The central moments of this empirical distribution are

$$M_{k,n} = E_n\big\{[X - E_n(X)]^k\big\} = E_n\big\{(X - \overline{X}_n)^k\big\} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^k$$

As with any distribution, the first central moment is zero, and the second is

$$V_n = \operatorname{var}_n(X) = E_n\big\{(X - \overline{X}_n)^2\big\} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2. \tag{7.16}$$

If there were any logic to statistics V_n would be called the "sample variance," but Lindgren, agreeing with most other textbooks, uses that term for something slightly different

$$S_n^2 = \frac{n}{n-1} V_n = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2. \tag{7.17}$$

The $n-1$ rather than n in the definition makes all of the formulas involving S_n^2 ugly, and makes S_n^2 not satisfy any of the usual rules involving variances. So be warned, and be careful! For example, V_n obeys the parallel axis theorem, hence

$$V_n = E_n(X^2) - E_n(X)^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \overline{X}_n^2.$$

Clearly S_n^2 cannot satisfy the same rule or it would be V_n. The only way to figure out the analogous rule for S_n^2 is to remember the rule for V_n (which makes sense) and derive the one for S_n^2.

$$\begin{aligned} S_n^2 &= \frac{n}{n-1} V_n \\ &= \frac{n}{n-1} \left[\frac{1}{n} \sum_{i=1}^{n} X_i^2 - \overline{X}_n^2 \right] \\ &= \frac{1}{n-1} \sum_{i=1}^{n} X_i^2 - \frac{n}{n-1} \overline{X}_n^2 \end{aligned}$$

No matter how you try to write it, it involves both n and $n-1$, and makes no sense.

Since S_n^2 is so ugly, why does anyone use it? The answer, as with so many other things, is circular. Almost everyone uses it because it's the standard, and it's the standard because almost everyone uses it. And "almost everyone" includes a lot of people, because S_n^2 is a topic in most introductory statistics courses.

Our position is that it simply does not matter whether you use V_n or S_n^2. Since one is a constant times the other, any place you could use one, you could use the other, so long as you make the appropriate changes in formulas. So the only reason for using S_n^2 is to avoid fighting tradition. Sometimes it's easier to follow the herd.

7.3.2 Sampling Distributions

Since a sample moment is a random variable, it has a probability distribution. We may not be able to give a formula for the density or distribution function, but it does have a distribution. So we can talk about its distribution and investigate its properties.

In a few specific cases we know the distribution of $\overline{X}_n$. It is given implicitly by what we call "addition rules" and which are summarized in Appendix C of these notes. They give the distribution of $Y = \sum_i X_i$ when the X_i are i. i. d.

- Binomial (including Bernoulli)
- Negative Binomial (including Geometric)
- Poisson
- Gamma (including Exponential and Chi-Square)
- Normal
- Cauchy

Given the distribution of Y, the distribution of $\overline{X}_n$ is found by a simple change of scale. If the X_i are continuous random variables, then

$$f_{\overline{X}_n}(z) = n f_Y(nz). \tag{7.18}$$

Example 7.3.1 (I. I. D. Exponential).
Let $X_1, \ldots, X_n$ be i. i. d. Exp(λ). Then the distribution of $Y = X_1 + \cdots + X_n$ is Gam(n, λ) by the addition rule for Gamma distributions (Appendix C) and the fact that the Exp(λ) is Gam($1, \lambda$). Hence by Problem 7-10

$$\overline{X}_n \sim \text{Gam}(n, n\lambda).$$

Many statistics textbooks, including Lindgren, have no tables of the gamma distribution. Thus we have to use the fact that gamma random variables having integer and half-integer values of their shape parameters are proportional to chi-square random variables, because $\text{chi}^2(n) = \text{Gam}(\frac{n}{2}, \frac{1}{2})$ and the second parameter of the gamma distribution is a shape parameter (Problem 7-10).

Lemma 7.10. *Suppose*

$$X \sim \text{Gam}(n, \lambda)$$

where n is an integer, then

$$2\lambda X \sim \text{chi}^2(2n).$$

The proof is Exercise 7-2.

Example 7.3.2 (Table Look-Up).
(Continues Example 7.3.1). Using the lemma, we can calculate probabilities for the sampling distribution of the sample mean of i. i. d. Exp(λ) data. Suppose $\lambda = 6.25$ so $\mu = 1/\lambda = 0.16$, and $n = 9$. What is $P(\overline{X}_n > 0.24)$.

In Example 7.3.1 we figured out that

$$\overline{X}_n \sim \text{Gam}(n, n\lambda)$$

so in this case

$$\overline{X}_n \sim \text{Gam}(9, 56.25) \tag{7.19}$$

($n\lambda = 9 \times 6.25 = 56.25$).

But to use the tables in Lindgren, we must use the lemma, which says

$$2n\lambda\overline{X}_n \sim \text{chi}^2(2n).$$

(there is an n on the left hand side, because the scale parameter of the gamma distribution is $n\lambda$ here rather than λ).

If $\overline{X}_n = 0.24$, then $2n\lambda\overline{X}_n = 2 \cdot 9 \cdot 6.25 \cdot 0.24 = 27.0$, and the answer to our problem is $P(Y > 27.0)$, where $Y \sim \text{chi}^2(18)$. Looking this up in Table Va in Lindgren, we get 0.079 for the answer.

Example 7.3.3 (Table Look-Up using Computers).
(Continues Example 7.3.1). The tables in Lindgren, or in other statistics books are not adequate for many problems. For many problems you need either a huge book of tables, commonly found in the reference section of a math library, or a computer.

Many mathematics and statistics computer software packages do calculations about probability distributions. In this course, we will only describe two of them: the statistical computing language R and the symbolic mathematics language Mathematica.

R In R the lookup is very simple. It uses the function `pgamma` which evaluates the gamma c. d. f.

```
> 1 - pgamma(0.24, 9, 1 / 56.25)
[1] 0.07899549
```

This statement evaluates $P(X \le x)$ when $X \sim \text{Gam}(9, 56.25)$ and $x = 0.24$, as (7.19) requires. We don't have to use the property that this gamma distribution is also a chi-square distribution. One caution: both R and Mathematica use a different parameterization of the gamma distribution than Lindgren. The shape parameter is the same, but the scale parameter is the reciprocal of Lindgren's scale parameter (See Problem 7-10). That's why the third argument of the `pgamma` function is 1/56.25 rather than 56.25.

Mathematica Mathematica makes things a bit more complicated. First you have to load a special package for probability distributions (always available, but not loaded by default), then you have to tell Mathematica which distribution you want, then you do the calculation

```
In[1]:= <<Statistics`ContinuousDistributions`

In[2]:= dist = GammaDistribution[9, 1 / 56.25]

Out[2]= GammaDistribution[9, 0.0177778]

In[3]:= F[x_] = CDF[dist, x]

Out[3]= GammaRegularized[9, 0, 56.25 x]

In[4]:= 1 - F[0.24]

Out[4]= 0.0789955
```

of course, the last three statements can be combined into one but just plugging in definitions

```
In[5]:= 1 - CDF[GammaDistribution[9, 1 / 56.25], 0.24]

Out[5]= 0.0789955
```

but that's a cluttered and obscure. For more on computing in general see the course computing web page

```
http://www.stat.umn.edu/geyer/5101/compute
```

and the pages on *Probability Distributions* in *R* and *Mathematica* in particular (follow the links from the main computing page).

Example 7.3.4 (I. I. D. Bernoulli).
If $X_1, \ldots, X_n$ are i. i. d. Ber(p) random variables, then $Y = \sum_i X_i$ is a Bin(n, p) random variable, and since $\overline{X}_n = Y/n$, we also have

$$n\overline{X}_n \sim \text{Bin}(n, p).$$

Example 7.3.5 (Another Computer Table Look-Up).
(Continues Example 7.3.4). Suppose $\overline{X}_n$ is the sample mean of 10 i. i. d. Ber(0.2) random variables. What is the probability $P(\overline{X}_n \le 0.1)$?

By the preceding example, $n\overline{X}_n \sim \text{Bin}(10, 0.2)$ and here $n\overline{X}_n = 10 \cdot 0.1 = 1$. So we need to look up $P(Y \le 1)$ when $Y \sim \text{Bin}(10, 0.2)$. In R this is

```
> pbinom(1, 10, 0.2)
[1] 0.3758096
```

In Mathematica it is

```
In[1]:= <<Statistics`DiscreteDistributions`

In[2]:= dist = BinomialDistribution[10, 0.2]

Out[2]= BinomialDistribution[10, 0.2]

In[3]:= F[x_] = CDF[dist, x]

Out[3]= BetaRegularized[0.8, 10 - Floor[x], 1 + Floor[x]]

In[4]:= F[1]

Out[4]= 0.37581
```

Our textbook has no tables of the binomial distribution, so there is no way to do this problem with pencil and paper except by evaluating the terms

$$\binom{n}{0} p^0 q^n + \binom{n}{1} p^1 q^{n-1}$$

(not so hard here, but very messy if there are many terms). You can't use the normal approximation because n is not large enough. Anyway, why use an approximation when the computer gives you the exact answer?

We can calculate the density using the convolution theorem. Mathematical induction applied to the convolution formula (Theorem 23 of Chapter 4 in Lindgren) gives the following result.

Theorem 7.11. *If X_1, ..., X_n are i. i. d. continuous random variables with common marginal density f_X, then $Y = X_1 + \cdots + X_n$ has density*

$$f_Y(y) = \int \cdots \iint f_X(y - x_2 - \cdots - x_n) f_X(x_2) \cdots f_X(x_n) \, dx_2 \cdots dx_n \quad (7.20)$$

Then (7.18) gives the density of $\overline{X}_n$. But this is no help if we can't do the integrals, which we usually can't, with the notable exceptions of the "brand name" distributions with "addition rules" (Appendix C).

Higher Moments So far we haven't considered any sample moment except $\overline{X}_n$. For other sample moments, the situation is even more complicated.

It is a sad fact is that the methods discussed in this section don't always work. In fact they usually don't work. Usually, nothing works, and you just can't find a closed form expression for the sampling distribution of a particular sample moment.

What is important to understand, though, and understand clearly, is that every sample moment does *have* a sampling distribution. Hence we can talk about properties of that distribution. The properties exist in principle, so we can talk about them whether or not we can calculate them.

7.3.3 Moments

In this section we calculate moments of sample moments. At first this sounds confusing, even bizarre, but sample moments are random variables and like any random variables they have moments.

Theorem 7.12. *If X_1, ..., X_n are identically distributed random variables with mean μ and variance σ^2, then*

$$E(\overline{X}_n) = \mu. \tag{7.21a}$$

If in addition, they are uncorrelated, then

$$\operatorname{var}(\overline{X}_n) = \frac{\sigma^2}{n}. \tag{7.21b}$$

If instead they are samples without replacement from a population of size N, then

$$\operatorname{var}(\overline{X}_n) = \frac{\sigma^2}{n} \cdot \frac{N - n}{N - 1}. \tag{7.21c}$$

Note in particular, that because independence implies lack of correlation, (7.21a) and (7.21b) hold in the i. i. d. case.

Proof. By the usual rules for linear transformations, $E(a + bX) = a + bE(X)$ and $\operatorname{var}(a + bX) = b^2 \operatorname{var}(X)$

$$E(\overline{X}_n) = \frac{1}{n} E\left(\sum_{i=1}^n X_i \right)$$

and

$$\operatorname{var}(\overline{X}_n) = \frac{1}{n^2} \operatorname{var}\left(\sum_{i=1}^n X_i \right)$$

Now apply Corollary 1 of Theorem 9 of Chapter 4 in Lindgren and (7.11) and (7.13). □

Theorem 7.13. *If X_1, ..., X_n are uncorrelated, identically distributed random variables with variance σ^2, then*

$$E(V_n) = \frac{n - 1}{n} \sigma^2, \tag{7.22a}$$

and

$$E(S_n^2) = \sigma^2. \tag{7.22b}$$

Proof. The reason why (7.22a) doesn't work out simply is that V_n involves deviations from the sample mean $\overline{X}_n$ and σ^2 involves deviations from the population mean μ. So use the empirical parallel axis theorem to rewrite V_n in terms of deviations from μ

$$E_n\{(X - \mu)^2\} = V_n + (\overline{X}_n - \mu)^2. \tag{7.23}$$

The left hand side is just $\overline{Y}_n$, where $Y_i = (X_i - \mu)^2$. Taking expectations of both sides of (7.23) gives

$$E(\overline{Y}_n) = E(V_n) + E\{(\overline{X}_n - \mu)^2\}$$

On the left hand side we have

$$E(\overline{Y}_n) = E(Y_i) = \operatorname{var}(X_i) = \sigma^2$$

And the second term on the right hand side is

$$\operatorname{var}(\overline{X}_n) = \frac{\sigma^2}{n}.$$

Collecting terms gives (7.22a). Then linearity of expectation gives (7.22b). □

The assertions (7.22a) and (7.22b) of this theorem are one place where S_n^2 seems simpler than V_n. It's why S_n^2 was invented, to make (7.22b) simple.

The sample moment formulas (7.21a), (7.21b), and (7.22b) are the ones most commonly used in everyday statistics. Moments of other sample moments exist but are mostly of theoretical interest.

Theorem 7.14. *If X_1, ..., X_n are i. i. d. random variables having moments of order k, then all sample moments of order k have expectation. If the X_i have moments of order $2k$, then sample moments of order k have finite variance. In particular,*

$$E(A_{k,n}) = \alpha_k$$

and

$$\operatorname{var}(A_{k,n}) = \frac{\alpha_{2k} - \alpha_k^2}{n},$$

where α_k is the k-th population moment.

We do not give formulas for the central moments because they are a mess. Even the formula for the variance of the sample variance given (though not proved) in Theorem 7 of Chapter 7 in Lindgren is already a mess. The formulas for higher moments are worse. They are, however, a straightforward mess. The proof below shows how the calculation would start. Continuing the calculation without making any mistakes would produce an explicit formula (a symbolic mathematics computer package like Maple or Mathematica would help a lot).

Proof. The k-th sample moment $A_{k,n}$ is the sample average of the random variables $Y_i = X_i^k$. Since

$$E(Y_i) = E(X_i^k) = \alpha_k \tag{7.24a}$$

and

$$\begin{aligned} \operatorname{var}(Y_i) &= E(Y_i^2) - E(Y_i)^2 \\ &= E(X_i^{2k}) - E(X_i^k)^2 \\ &= \alpha_{2k} - \alpha_k^2 \end{aligned} \tag{7.24b}$$

the formulas in the theorem follow by the usual rules for the moments of a sample mean.

The k-th central sample moment

$$\begin{aligned} M_{k,n} &= \frac{1}{n} \sum_{i=1}^n \bigl(X_i - \overline{X}_n \bigr)^k \\ &= \frac{1}{n} \sum_{i=1}^n \left(\frac{n-1}{n} X_i - \sum_{j \neq i} \frac{1}{n} X_j \right)^k \end{aligned}$$

is a k-th degree polynomial in the X_i. A single term of such a polynomial has the form

$$a \prod_{i=1}^n X_i^{m_i}$$

where the m_i are nonnegative integers such that $m_1 + \cdots + m_n = k$, and a is some constant (a different constant for each term of the polynomial, although the notation doesn't indicate that). By independence

$$E\left(a \prod_{i=1}^n X_i^{m_i} \right) = a \prod_{i=1}^n E(X_i^{m_i}) = a \prod_{i=1}^n \alpha_{m_i}. \tag{7.25}$$

If k-th moments exist, then all of the moments α_{m_i} in (7.25) exist because $m_i \le k$.

Similarly, $M_{k,n}^2$ is a polynomial of degree $2k$ in the X_i and hence has expectation if population moments of order $2k$ exist. Then $\operatorname{var}(M_{k,n}) = E(M_{k,n}^2) - E(M_{k,n})^2$ also exists. □

7.3.4 Asymptotic Distributions

Often we cannot calculate the exact sampling distribution of a sample moment, but we can always get large sample properties of the distribution from law of large numbers, the central limit theorem, and Slutsky's theorem.

Theorem 7.15. *Under i. i. d. sampling every sample moment converges in probability to the corresponding population moment provided the population moment exists.*

Proof. For ordinary moments, this was done as a homework problem (Problem 5-3 in Lindgren). If we let α_k be the k-th ordinary population moment and $A_{k,n}$ be the corresponding ordinary sample moment for sample size n, then

$$A_{k,n} = \frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{P} E(X_1^k) = \alpha_k.$$

Let μ_k be the k-th population central moment and $M_{k,n}$ be the corresponding sample central moment, then

$$\begin{aligned} M_{k,n} &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^k && \text{(7.26a)} \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{k}\binom{k}{j}(-1)^j(\overline{X}_n - \mu)^j(X_i - \mu)^{k-j} \\ &= \sum_{j=0}^{k}\binom{k}{j}(-1)^j(\overline{X}_n - \mu)^j\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^{k-j} \\ &= \sum_{j=0}^{k}\binom{k}{j}(-1)^j(\overline{X}_n - \mu)^j M'_{k-j,n} && \text{(7.26b)} \end{aligned}$$

where we have introduced the notation

$$M'_{k,n} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^k.$$

This is almost the same as (7.26a), the only difference being the replacement of $\overline{X}_n$ by μ. The asymptotics of $M'_{k,n}$ are much simpler than those for $M_{k,n}$ because $M'_{k,n}$ is the sum of i. i. d. terms so the LLN and CLT apply directly to it. In particular

$$M'_{k,n} \xrightarrow{P} E\{(X_i - \mu)^k\} = \mu_k \tag{7.27}$$

also

$$\overline{X}_n - \mu \xrightarrow{P} 0 \tag{7.28}$$

by the LLN and the continuous mapping theorem. Then (7.28) and Slutsky's theorem imply that every term of (7.26b) converges in probability to zero except the $j = 0$ term, which is $M'_{k,n}$. Thus (7.27) establishes

$$M_{k,n} \xrightarrow{P} \mu_k \tag{7.29}$$

which is what was to be proved. □

Theorem 7.16. *Under i. i. d. sampling every sample k-th moment is asymptotically normal if population moments of order $2k$ exist. In particular,*

$$\sqrt{n}(A_{k,n} - \alpha_k) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \alpha_{2k} - \alpha_k^2) \tag{7.30}$$

and

$$\sqrt{n}(M_{k,n} - \mu_k) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mu_{2k} - \mu_k^2 - 2k\mu_{k-1}\mu_{k+1} + k^2\mu_2\mu_{k-1}^2) \tag{7.31}$$

For ordinary moments, this is a homework problem (Problem 7-17 in Lindgren). For central moments, the proof will have to wait until we have developed multivariate convergence in distribution in the following chapter.

The special case $k = 2$ is worth noting.

Corollary 7.17. *Suppose X_1, X_2, ... are i. i. d. and have fourth moments. Then*

$$\sqrt{n}(V_n - \sigma^2) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mu_4 - \mu_2^2)$$

where V_n is defined by (7.16).

This is the case $V_n = M_{2,n}$ of the theorem. The third and forth terms of the asymptotic variance formula are zero because $\mu_1 = 0$ (Theorem 2.9 in Chapter 2 of these notes).

Example 7.3.6 (I. I. D. Normal).
Suppose X_1, ..., X_n are i. i. d. $\mathcal{N}(\mu, \sigma^2)$. What is the asymptotic distribution of $\overline{X}_n$, of V_n, of $M_{3,n}$?

The CLT, of course, tells us the asymptotic distribution of $\overline{X}_n$. Here we just want to check that the $k = 1$ case of (7.30) agrees with the CLT. Note that $A_{1,n} = \overline{X}_n$ and $\alpha_1 = \mu$, so the left hand side of (7.30) is the same as the left hand side of the CLT (6.7). Also $\alpha_2 - \alpha_1^2 = \sigma^2$ because this is just $\operatorname{var}(X) = E(X^2) - E(X)^2$ in different notation. So the $k = 1$ case of (7.30) does agree with the CLT.

The asymptotic distribution of $V_n = M_{2,n}$ is given by the $k = 2$ case of (7.31) or by Theorem 7.17. All we need to do is calculate the asymptotic variance $\mu_4 - \mu_2^2$. The fourth central moment of the standard normal distribution is given by the $k = 2$ case of equation (5) on p. 178 in Lindgren to be $\mu_4 = 3$. A general normal random variable has the form $X = \mu + \sigma Z$, where Z is standard normal, and this has fourth central moment $3\sigma^4$ by Problem 7-11. Thus $\mu_4 - \mu_2^2 = 3\sigma^4 - \sigma^4 = 2\sigma^4$, and finally we get

$$V_n \approx \mathcal{N}\left(\sigma^2, \frac{2\sigma^4}{n}\right)$$

Note this formula holds for i. i. d. normal data *only*. Other statistical models can have rather different distributions (Problem 7-12).

The asymptotic distribution of $M_{3,n}$ is given by the $k = 3$ case of (7.31)

$$\begin{aligned} \mu_6 - \mu_3^2 - 2 \cdot 3\mu_2\mu_4 + 3^2\mu_2 \cdot \mu_2^2 &= \mu_6 - \mu_3^2 - 6\mu_2\mu_4 + 9\mu_2^3 \\ &= \mu_6 - 6\mu_2\mu_4 + 9\mu_2^3 \end{aligned}$$

because odd central moments are zero (Theorem 2.10 of Chapter 2 of these notes). We already know $\mu_2 = \sigma^2$ and $\mu_4 = 3\sigma^2$. Now we need to use the

$k = 3$ case of equation (5) on p. 178 in Lindgren and Problem 7-11 to get to be $\mu_6 = 15\sigma^2$. Hence the asymptotic variance is

$$\mu_6 - 6\mu_2\mu_4 + 9\mu_2^3 = (15 - 6 \cdot 1 \cdot 3 + 9)\sigma^6 = 6\sigma^6$$

and

$$M_{3,n} \approx \mathcal{N}\left(0, \frac{6\sigma^6}{n}\right)$$

(the asymptotic mean is $\mu_3 = 0$).

7.3.5 The t Distribution

We now derive two other "brand name" distributions that arise as exact sampling distributions of statistics derived from sampling normal populations. The distributions are called the t and F distributions (whoever thought up those names must have had a real imagination!)

Before we get to them, we want to generalize the notion of degrees of freedom to noninteger values. This will be useful when we come to Bayesian inference.

Definition 7.3.1 (Chi-Square Distribution).
The chi-square with noninteger degrees of freedom $\nu > 0$ is the $\text{Gam}(\frac{\nu}{2}, \frac{1}{2})$ *distribution.*

This agrees with our previous definition when ν is an integer.

Definition 7.3.2 (Student's t Distribution).
If Z and Y are independent random variables, Z is standard normal and Y is $\text{chi}^2(\nu)$*, then the random variable*

$$T = \frac{Z}{\sqrt{Y/\nu}}$$

is said to have a t-distribution with ν degrees of freedom*, abbreviated $t(\nu)$. The parameter ν can be any strictly positive real number.*

The reason for the "Student" sometimes attached to the name of the distribution is that the distribution was discovered and published by W. S. Gosset, the chief statistician for the Guiness brewery in Ireland. The brewery had a company policy that employees were not allowed to publish under their own names, so Gosset used the pseudonym "Student" and this pseudonym is still attached to the distribution by those who like eponyms.

Theorem 7.18. *The p. d. f. of the $t(\nu)$ distribution is*

$$f_\nu(x) = \frac{1}{\sqrt{\nu\pi}} \cdot \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \cdot \frac{1}{\left(1 + \frac{x^2}{\nu}\right)^{(\nu+1)/2}}, \qquad -\infty < x < +\infty \tag{7.32}$$

The normalizing constant can also be written using a beta function because $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. Thus

$$\frac{1}{\sqrt{\nu\pi}} \cdot \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} = \frac{1}{\sqrt{\nu}} \cdot \frac{1}{B(\frac{\nu}{2}, \frac{1}{2})}$$

The connection with the beta distribution is obscure but will be clear after we finish this section and do Problem 7-3.

Proof. The joint distribution of Z and Y in the definition is

$$f(z, y) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \frac{\left(\frac{1}{2}\right)^{\nu/2}}{\Gamma(\nu/2)} y^{\nu/2-1} e^{-y/2}$$

Make the change of variables $t = z/\sqrt{y/\nu}$ and $u = y$, which has inverse transformation

$$\begin{aligned} z &= t\sqrt{u/\nu} \\ y &= u \end{aligned}$$

and Jacobian

$$\begin{vmatrix} \sqrt{u/\nu} & t/2\sqrt{u\nu} \\ 0 & 1 \end{vmatrix} = \sqrt{u/\nu}$$

Thus the joint distribution of T and U given by the multivariate change of variable formula is

$$\begin{aligned} f(t, u) &= \frac{1}{\sqrt{2\pi}} e^{-(t\sqrt{u/\nu})^2/2} \frac{\left(\frac{1}{2}\right)^{\nu/2}}{\Gamma(\nu/2)} u^{\nu/2-1} e^{-u/2} \cdot \sqrt{u/\nu} \\ &= \frac{1}{\sqrt{2\pi}} \frac{\left(\frac{1}{2}\right)^{\nu/2}}{\Gamma(\nu/2)} \frac{1}{\sqrt{\nu}} u^{\nu/2-1/2} \exp\left\{ -\left(1 + \frac{t^2}{\nu}\right) \frac{u}{2} \right\} \end{aligned}$$

Thought of as a function of u for fixed t, this is proportional to a gamma density with shape parameter $(\nu + 1)/2$ and inverse scale parameter $\frac{1}{2}(1 + \frac{t^2}{\nu})$. Hence we can use the "recognize the unnormalized density trick" (Section 2.5.7 in Chapter 2 of these notes) to integrate out u getting the marginal of t

$$f(t) = \frac{1}{\sqrt{2\pi}} \cdot \frac{\left(\frac{1}{2}\right)^{\nu/2}}{\Gamma(\nu/2)} \cdot \frac{1}{\sqrt{\nu}} \cdot \frac{\Gamma(\frac{\nu+1}{2})}{[\frac{1}{2}(1 + \frac{t^2}{\nu})]^{(\nu+1)/2}}$$

which, after changing t to x, simplifies to (7.32). □

The formula for the density of the t distribution shows that it is symmetric about zero. Hence the median is zero, and the mean is also zero when it exists. In fact, all odd central moments are zero when they exist, because this is true of any symmetric random variable (Theorem 2.10 of Chapter 2 of these notes).

The question of when moments exist is settled by the following theorem.

Theorem 7.19. *If X has a Student t distribution with ν degrees of freedom, then moments of order k exist if and only if $k < \nu$.*

Proof. The density (7.32) is clearly bounded. Hence we only need to check whether $|x|^k f(x)$ is integrable near infinity. Since the density is symmetric, we only need to check one tail. For x near $+\infty$

$$|x|^k f(x) \approx k x^{k-(\nu+1)}$$

for some constant k. From Lemma 2.39 of Chapter 2 of these notes the integral is finite if and only if $k - (\nu + 1) < -1$, which the same as $\nu > k$. □

We also want to know the variance of the t distribution.

Theorem 7.20. *If $\nu > 2$ and $X \sim t(\nu)$, then*

$$\operatorname{var}(X) = \frac{\nu}{\nu - 2}.$$

The proof is a homework problem (7-5).

Another important property of the t distribution is given in the following theorem, which we state without proof since it involves the Stirling approximation for the gamma function, which we have not developed, although we will prove a weaker form of the second statement of the theorem in the next chapter after we have developed some more tools.

Theorem 7.21. *For every $x \in \mathbb{R}$*

$$f_\nu(x) \to \phi(x), \qquad \text{as } \nu \to \infty,$$

where ϕ is the standard normal density, and

$$t(\nu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1), \qquad \text{as } \nu \to \infty.$$

Comparison of the $t(1)$ density to the standard Cauchy density given by equation (1) on p. 191 in Lindgren shows they are the same (it is obvious that the part depending on x is the same, hence the normalizing constants must be the same if both integrate to one, but in fact we already know that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ also shows the normalizing constants are equal). Thus $t(1)$ is another name for the standard Cauchy distribution. The theorem above says we can think of $t(\infty)$ as another name for the standard normal distribution. Tables of the t distribution, including Tables IIIa and IIIb in the Appendix of Lindgren include the normal distribution labeled as ∞ degrees of freedom. Thus the t family of distributions provides lots of examples between the best behaved distribution of those we've studied, which is the normal, and the worst behaved, which is the Cauchy. In particular, the $t(2)$ distribution has a mean but no variance, hence the sample mean of i. i. d. $t(2)$ random variables obeys the LLN but not the CLT. For $\nu > 2$, The $t(\nu)$ distribution has both mean and variance, hence the sample mean of i. i. d. $t(\nu)$ random variables obeys both LLN and CLT, but the $t(\nu)$ distribution is much more heavy-tailed than other distributions we have previously considered.

7.3.6 The F Distribution

The letter F for the random variable having the "F distribution" was chosen by Snedecor in honor of R. A. Fisher who more or less invented the F distribution. Actually, he proposed a monotone transformation of this variable $Z = \frac{1}{2} \log F$, which has a better normal approximation.

Definition 7.3.3 (The F Distribution).
If Y_1 and Y_2 are independent random variables, and $Y_i \sim \text{chi}^2(\nu_i)$, then the random variable

$$U = \frac{Y_1/\nu_1}{Y_2/\nu_2}$$

has an F distribution with ν_1 numerator degrees of freedom and ν_2 denominator degrees of freedom, abbreviated $F(\nu_1, \nu_2)$.

Theorem 7.22. *If Y_1 and Y_2 are independent random variables, and $Y_i \sim \text{chi}^2(\nu_i)$, then the random variable*

$$W = \frac{Y_1}{Y_1 + Y_2}$$

has a $\text{Beta}(\frac{\nu_1}{2}, \frac{\nu_2}{2})$ distribution.

Proof. Since we know that the chi-square distribution is a special case of the gamma distribution $\text{chi}^2(k) = \text{Gam}(\frac{k}{2}, \frac{1}{2})$, this is one of the conclusions of Theorem 4.2 of Chapter 4 of these notes. □

Corollary 7.23. *If $U \sim F(\nu_1, \nu_2)$, then*

$$W = \frac{\frac{\nu_1}{\nu_2} U}{1 + \frac{\nu_1}{\nu_2} U}$$

has a $\text{Beta}(\frac{\nu_1}{2}, \frac{\nu_2}{2})$ distribution.

Hence the F distribution is not really new, it is just a transformed beta distribution. The only reason for defining the F distribution is convention. Tables of the F distribution are common. There is one in the appendix of Lindgren. Tables of the beta distribution are rare. So we mostly use F tables rather than beta tables. When using a computer, the distinction doesn't matter. Mathematica and R have functions that evaluate either F or beta probabilities.

7.3.7 Sampling Distributions Related to the Normal

When the data are i. i. d. normal, the exact (not asymptotic) sampling distributions are known for many quantities of interest.

Theorem 7.24. *If $X_1, \ldots, X_n$ are i. i. d. $\mathcal{N}(\mu, \sigma^2)$, then $\overline{X}_n$ and S_n^2 given by* (7.15) *and* (7.17) *are independent random variables and*

$$\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \tag{7.33a}$$

$$(n-1)S_n^2/\sigma^2 \sim \text{chi}^2(n-1) \tag{7.33b}$$

This is a combination of Theorems 9, 10, and 11 and the Corollary to Theorem 10 in Section 7.5 of Lindgren.

Note that the theorem implicitly gives the distribution of S_n^2, since $\text{chi}^2(n-1)$ is just another name for $\text{Gam}(\frac{n-1}{2}, \frac{1}{2})$ and the second parameter of the gamma is an upside down scale parameter, which implies

$$S_n^2 \sim \text{Gam}\left(\frac{n-1}{2}, \frac{n-1}{2\sigma^2}\right) \tag{7.34}$$

The theorem is stated the way it is because chi-square tables are widely available (including in the Appendix of Lindgren) and gamma tables are not. Hence (7.33b) is a more useful description of the sampling distribution of S_n^2 than is (7.34) when you are using tables (if you are using a computer, either works).

The main importance of the t distribution in statistics comes from the following corollary.

Corollary 7.25. *If X_1, ..., X_n are i. i. d. $\mathcal{N}(\mu, \sigma^2)$, then*

$$T = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}}$$

has a $t(n-1)$ distribution.

Proof.

$$Z = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

is standard normal, and independent of $Y = (n-1)S_n^2/\sigma^2$ which is $\text{chi}^2(n-1)$ by Theorem 7.24. Then $Z/\sqrt{Y/(n-1)}$ is T. □

One use of the F distribution in statistics (not the most important) comes from the following corollary.

Corollary 7.26. *If X_1, ..., X_m are i. i. d. $\mathcal{N}(\mu_X, \sigma_X^2)$ and Y_1, ..., Y_n are i. i. d. $\mathcal{N}(\mu_Y, \sigma_Y^2)$, and all of the X_i are independent of all of the Y_j, then*

$$F = \frac{S_{m,X}^2}{S_{n,Y}^2} \cdot \frac{\sigma_Y^2}{\sigma_X^2}$$

has an $F(m-1, n-1)$ distribution, where $S_{m,X}^2$ is the sample variance of the X_i and $S_{n,Y}^2$ is the sample variance of the Y_i.

The proof is obvious from Theorem 7.24 and the definition of the F distribution.

Example 7.3.7 (T Distribution).
Suppose X_1, ..., X_{20} are i. i. d. standard normal. Compare $P(\overline{X}_n > \sigma/\sqrt{n})$

and $P(\overline{X}_n > S_n/\sqrt{n})$. We know that

$$\frac{\overline{X}_n}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$
$$\frac{\overline{X}_n}{S_n/\sqrt{n}} \sim t(19)$$

So we need to compare $P(Z > 1)$ where Z is standard normal and $P(T > 1)$ where $T \sim t(19)$.

From Tables I and IIIa in Lindgren, these probabilities are .1587 and .165, respectively. The following R commands do the same lookup

```
> 1 - pnorm(1)
[1] 0.1586553
> 1 - pt(1, 19)
[1] 0.1649384
```

Example 7.3.8 (F Distribution).
Suppose S_1^2 and S_2^2 are sample variances of two independent samples from two normal populations with equal variances, and the sample sizes are $n_1 = 10$ and $n_2 = 20$, respectively. What is $P(S_1^2 > 2S_2^2)$? We know that

$$\frac{S_1^2}{S_2^2} \sim F(9, 19)$$

So the answer is $P(Y > 2)$ where $Y \sim F(9, 19)$. Tables IVa and IVb in Lindgren (his only tables of the F distribution) are useless for this problem. We must use the computer. In R it's simple

```
> 1 - pf(2, 9, 19)
[1] 0.0974132
```

For this example, we also show how to do it in Mathematica

```
In[1]:= <<Statistics`ContinuousDistributions`

In[2]:= dist = FRatioDistribution[9, 19]

Out[2]= FRatioDistribution[9, 19]

In[3]:= F[x_] = CDF[dist, x]

                             19        19  9
Out[3]= BetaRegularized[--------, 1, --, -]
                       19 + 9 x     2   2

In[4]:= 1 - F[2]
```

```
                              19     19  9
Out[4]= 1 - BetaRegularized[--, 1, --, -]
                              37     2   2

In[5]:= N[%]

Out[5]= 0.0974132
```

(The last command tells Mathematica to evaluate the immediately preceding expression giving a numerical result). This can be done more concisely if less intelligibly as

```
In[6]:= N[1 - CDF[FRatioDistribution[9, 19], 2]]

Out[6]= 0.0974132
```

7.4 Sampling Distributions of Sample Quantiles

The *sample quantiles* are the quantiles of the empirical distribution associated with the data vector $\mathbf{X} = (X_1, \ldots, X_n)$. They are mostly of interest only for continuous population distributions. A sample quantile can always be taken to be an order statistic by Theorem 7.5. Hence the exact sampling distributions of the empirical quantiles are given by the exact sampling distributions for order statistics, which are given by equation (5) on p. 217 of Lindgren

$$f_{X_{(k)}}(y) = \frac{n!}{(k-1)!(n-k)!} F(y)^{k-1} [1 - F(y)]^{n-k} f(y) \tag{7.35}$$

when the population distribution is continuous, (where, as usual, F is the c. d. f. of the X_i and f is their p. d. f.). Although this is a nice formula, it is fairly useless. We can't calculate any moments or other useful quantities, except in the special case where the X_i have a $\mathcal{U}(0, 1)$ distribution, so $F(y) = y$ and $f(y) = 1$ for all y and we recognize

$$f_{X_{(k)}}(y) = \frac{n!}{(k-1)!(n-k)!} y^{k-1} (1 - y)^{n-k} \tag{7.36}$$

as a $\text{Beta}(k, n - k + 1)$ distribution.

Much more useful is the asymptotic distribution of the sample quantiles given by the following. We will delay the proof of the theorem until the following chapter, where we will develop the tools of multivariate convergence in distribution used in the proof.

Theorem 7.27. *Suppose X_1, X_2, ... are continuous random variables that are independent and identically distributed with density f that is nonzero at the p-th quantile x_p, and suppose*

$$\sqrt{n} \left(\frac{k_n}{n} - p \right) \to 0, \qquad \text{as } n \to \infty, \tag{7.37}$$

then

$$\sqrt{n}\big(X_{(k_n)} - x_p\big) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{p(1-p)}{f(x_p)^2}\right), \qquad \text{as } n \to \infty. \tag{7.38}$$

Or the sloppy version

$$X_{(k_n)} \approx \mathcal{N}\left(x_p, \frac{p(1-p)}{nf(x_p)^2}\right).$$

In particular, if we define $k_n = \lceil np \rceil$, then $X_{(k_n)}$ is a sample p-th quantile by Theorem 7.5. The reason for the extra generality, is that the theorem makes it clear that $X_{(k_n+1)}$ also has the same asymptotic distribution. Since $X_{(k_n)} \le X_{(k_n+1)}$ always holds by definition of order statistics, this can only happen if

$$\sqrt{n}\big(X_{(k_n+1)} - X_{(k_n)}\big) \xrightarrow{P} 0.$$

Hence the average

$$\widetilde{X}_n = \frac{X_{(k_n)} + X_{(k_n+1)}}{2}$$

which is the conventional definition of the sample median, has the same asymptotic normal distribution as either $X_{(k_n)}$ or $X_{(k_n+1)}$.

Corollary 7.28. *Suppose X_1, X_2, ... are continuous random variables that are independent and identically distributed with density f that is nonzero the population median m, then*

$$\sqrt{n}\big(\widetilde{X}_n - m\big) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{1}{4f(x_p)^2}\right), \qquad \text{as } n \to \infty.$$

This is just the theorem with $x_p = m$ and $p = 1/2$. The sloppy version is

$$\widetilde{X}_n \approx \mathcal{N}\left(m, \frac{1}{4nf(m)^2}\right).$$

Example 7.4.1 (Median, Normal Population).
If X_1, X_2, ... are i. i. d. $\mathcal{N}(\mu, \sigma^2)$, then the population median is μ by symmetry and the p. d. f. at the median is

$$f(\mu) = \frac{1}{\sigma\sqrt{2\pi}}$$

Hence

$$\widetilde{X}_n \approx \mathcal{N}\left(\mu, \frac{\pi\sigma^2}{2n}\right).$$

or, more precisely,

$$\sqrt{n}(\widetilde{X}_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{\pi\sigma^2}{2}\right)$$

Problems

7-1. The *median absolute deviation from the median* (MAD) of a random variable X with unique median m is the median of the random variable $Y = |X - m|$. The MAD of the values $x_1, \ldots, x_n$ is the median of the values $x_i - \tilde{x}_n$, where $\tilde{x}_n$ is the empirical median defined in Definition 7.1.4. This is much more widely used than the "other MAD," mean absolute deviation from the mean, discussed in Lindgren.

(a) Show that for a symmetric continuous random variable with strictly positive p. d. f. the MAD is half the interquartile range. (The point of requiring a strictly positive p. d. f. is that this makes all the quantiles unique and distinct. The phenomena illustrated in the middle and right panels of Figure 3-3 in Lindgren cannot occur.)

(b) Calculate the MAD for the standard normal distribution.

(c) Calculate the MAD for the data in Problem 7-4 in Lindgren.

7-2. Prove Lemma 7.10.

7-3. Show that if $T \sim t(\nu)$, then $T^2 \sim F(1, \nu)$.

7-4. Show that if $X \sim F(\mu, \nu)$ and $\nu > 2$, then

$$E(X) = \frac{\nu}{\nu - 2}$$

7-5. Prove Theorem 7.20.

7-6. Find the asymptotic distribution of the sample median of an i. i. d. sample from the following distributions:

(a) Cauchy(μ, σ) with density $f_{\mu,\sigma}$ given by

$$f_{\mu,\sigma}(x) = \frac{\sigma}{\pi(\sigma^2 + [x - \mu]^2)}, \qquad -\infty < x < +\infty$$

(b) The double exponential distribution (also called Laplace distribution) having density

$$f_{\mu,\sigma}(x) = \frac{1}{2\sigma} e^{-|x-\mu|/\sigma}, \qquad -\infty < x < +\infty$$

7-7. Suppose X_1, X_2, ... are i. i. d. $\mathcal{U}(0, \theta)$. As usual $X_{(n)}$ denotes the n-th order statistic, which is the maximum of the X_i.

(a) Show that

$$X_{(n)} \xrightarrow{P} \theta, \qquad \text{as } n \to \infty.$$

(b) Show that

$$n\big(\theta - X_{(n)}\big) \xrightarrow{\mathcal{D}} \text{Exp}(1/\theta), \qquad \text{as } n \to \infty.$$

Hints This is a rare problem (the only one of the kind we will meet in this course) when we can't use the LLN or the CLT to get convergence in probability and convergence in distribution results (obvious because the problem is not about $\overline{X}_n$ and the asymptotic distribution we seek isn't normal). Thus we need to derive convergence in distribution directly from the definition (Definition 6.1.1 in these notes or the definition on p. 135 in Lindgren).
Hint for Part (a): Show that the c. d. f. of $X_{(n)}$ converges to the c. d. f. of the constant random variable θ. (Why does this do the job?)
Hint for Part (b): Define

$$Y_n = n\big(\theta - X_{(n)}\big)$$

(the random variable we're trying to get an asymptotic distribution for). Derive its c. d. f. $F_{Y_n}(y)$. What you need to show is that

$$F_{Y_n}(y) \to F(y), \qquad \text{for all } y$$

where F is the c. d. f. of the $\text{Exp}(1/\theta)$ distribution. The fact from calculus

$$\lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n = e^x$$

is useful in this.

You can derive the c. d. f. of Y_n from the c. d. f. of $X_{(n)}$, which is given in the first displayed equation (unnumbered) of Section 7.6 in Lindgren.

7-8. Suppose X_1, ..., X_n are i. i. d. $\mathcal{N}(\mu, \sigma^2)$. What is the probability that $|\overline{X}_n - \mu| > 2S_n/\sqrt{n}$ if $n = 10$?

7-9. Suppose X_1, ..., X_n are i. i. d. $\mathcal{N}(\mu, \sigma^2)$. What is the probability that $S_n^2 > 2\sigma^2$ if $n = 10$?

7-10. R and Mathematica and many textbooks use a different parameterization of the gamma distribution. They write

$$f(x \mid \alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} \tag{7.39}$$

rather than

$$f(x \mid \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \tag{7.40}$$

Clearly the two parameterizations have the same first parameter α, as the notation suggests, and second parameters related by $\lambda = 1/\beta$.

(a) Show that β is the usual kind of scale parameter, that if X has p. d. f. (7.39), then σX has p. d. f. $f(x \mid \alpha, \sigma\beta)$, where again the p. d. f. is defined by (7.39).

(b) Show that λ is an "upside down" scale parameter, that if X has p. d. f. (7.40), then σX has p. d. f. $f(x \mid \alpha, \lambda/\sigma)$, where now the p. d. f. is defined by (7.40).

7-11. Show if X has k-th central moment

$$\mu_k = E\{(X - \mu)^k\}$$

where, as usual, $\mu = E(X)$, then $Y = a + bX$ has k-th central moment $b^k \mu_k$.

7-12. What is the asymptotic distribution of the variance V_n of the empirical distribution for an i. i. d. Exp(λ) sample?

7-13. Suppose X is standard normal (so $\mu_X = 0$ and $\sigma_X = 1$).

(a) What is $P(|X| > 2\sigma_X)$?

In contrast, suppose X has a $t(3)$ distribution (so $\mu_X = 0$ and the variance σ_X^2 is given by Problem 7-5)

(b) Now what is $P(|X| > 2\sigma_X)$?

7-14. With all the same assumptions as in Example 7.3.8, what are

(a) $P(S_2^2 > S_1^2)$?

(b) $P(S_2^2 > 2S_1^2)$?

7-15. Suppose X_1, X_2, X_3, ... is an i. i. d. sequence of random variables with mean μ and variance σ^2, and $\overline{X}_n$ is the sample mean. Show that

$$\sqrt{n}\left(\overline{X}_n - \mu\right)^k \xrightarrow{P} 0$$

for any integer $k > 1$. (**Hint:** Use the CLT, the continuous mapping theorem for convergence in distribution, and Slutsky's theorem.)

Appendix A

Greek Letters

Table A.1: Table of Greek Letters (Continued on following page.)

name	capital letter	small letter	pronunciation	sound
alpha	A	α	AL-fah	short a
beta	B	β	BAY-tah	b
gamma	Γ	γ	GAM-ah	g
delta	Δ	δ	DEL-tah	d
epsilon	E	ϵ	EP-si-lon	e
zeta	Z	ζ	ZAY-tah	z
eta	H	η	AY-tah	long a
theta	Θ	θ or ϑ	THAY-thah	soft th (as in thin)
iota	I	ι	EYE-oh-tah	i
kappa	K	κ	KAP-ah	k
lambda	Λ	λ	LAM-dah	l
mu	M	μ	MYOO	m
nu	N	ν	NOO	n
xi	Ξ	ξ	KSEE	x (as in box)
omicron	O	o	OH-mi-kron	o
pi	Π	π	PIE	p
rho	R	ρ	RHOH	rh[1]
sigma	Σ	σ	SIG-mah	s
tau	T	τ	TAOW	t
upsilon	Υ	υ	UP-si-lon	u

[1]The sound of the Greek letter ρ is not used in English. English words, like *rhetoric* and *rhinoceros* that are descended from Greek words beginning with ρ have English pronunciations beginning with an "r" sound rather than "rh" (though the spelling reminds us of the Greek origin).

Table A.2: Table of Greek Letters (Continued.)

name	capital letter	small letter	pronunciation	sound
phi	Φ	ϕ or φ	FIE	f
chi	X	χ	KIE	guttural ch[2]
psi	Ψ	ψ	PSY	ps (as in stops)[3]
omega	Ω	ω	oh-MEG-ah	o

[2]The sound of the Greek letter χ is not used in English. It is heard in the German *Buch* or Scottish *loch*. English words, like *chemistry* and *chorus* that are descended from Greek words beginning with χ have English pronunciations beginning with a "k" sound rather than "guttural ch" (though the spelling reminds us of the Greek origin).

[3]English words, like *pseudonym* and *psychology* that are descended from Greek words beginning with ψ have English pronunciations beginning with an "s" sound rather than "ps" (though the spelling reminds us of the Greek origin).

Appendix B

Summary of Brand-Name Distributions

B.1 Discrete Distributions

B.1.1 The Discrete Uniform Distribution

The Abbreviation $\mathcal{DU}(S)$.

The Sample Space Any finite set S.

The Density

$$f(x) = \frac{1}{n}, \qquad x \in S,$$

where $n = \text{card}(S)$.

Specialization The case in which the sample space consists of consecutive integers $S = \{m, m + 1, \ldots, n\}$ is denoted $\mathcal{DU}(m, n)$.

Moments If $X \sim \mathcal{DU}(1, n)$, then

$$E(X) = \frac{n + 1}{2}$$
$$\text{var}(X) = \frac{n^2 - 1}{12}$$

B.1.2 The Binomial Distribution

The Abbreviation $\text{Bin}(n, p)$

The Sample Space The integers 0, ..., n.

The Parameter p such that $0 < p < 1$.

The Density

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, \ldots, n.$$

Moments

$$\begin{aligned} E(X) &= np \\ \operatorname{var}(X) &= np(1-p) \end{aligned}$$

Specialization

$$\operatorname{Ber}(p) = \operatorname{Bin}(1, p)$$

B.1.3 The Geometric Distribution, Type II

Note This section has changed. The roles of p and $1-p$ have been reversed, and the abbreviation $\operatorname{Geo}(p)$ is no longer used to refer to this distribution but the distribution defined in Section B.1.8. All of the changes are to match up with Chapter 6 in Lindgren.

The Abbreviation No abbreviation to avoid confusion with the other type defined in Section B.1.8.

Relation Between the Types If $X \sim \operatorname{Geo}(p)$, then $Y = X - 1$ has the distribution defined in this section.

X is the number of *trials* before the first success in an i. i. d. sequence of $\operatorname{Ber}(p)$ random variables. Y is the number of *failures* before the first success.

The Sample Space The integers 0, 1, ...

The Parameter p such that $0 < p < 1$.

The Density

$$f(x) = p(1-p)^x, \qquad x = 0, 1, \ldots.$$

Moments

$$\begin{aligned} E(X) &= \frac{1}{p} - 1 = \frac{1-p}{p} \\ \operatorname{var}(X) &= \frac{1-p}{p^2} \end{aligned}$$

B.1.4 The Poisson Distribution

The Abbreviation $\text{Poi}(\mu)$

The Sample Space The integers 0, 1, ...

The Parameter μ such that $\mu > 0$.

The Density

$$f(x) = \frac{\mu^x}{x!} e^{-\mu}, \qquad x = 0, 1, \ldots .$$

Moments

$$\begin{aligned} E(X) &= \mu \\ \text{var}(X) &= \mu \end{aligned}$$

B.1.5 The Bernoulli Distribution

The Abbreviation $\text{Ber}(p)$

The Sample Space The integers 0 and 1.

The Parameter p such that $0 < p < 1$.

The Density

$$f(x) = \begin{cases} p, & x = 1 \\ 1 - p & x = 0 \end{cases}$$

Moments

$$\begin{aligned} E(X) &= p \\ \text{var}(X) &= p(1 - p) \end{aligned}$$

Generalization

$$\text{Ber}(p) = \text{Bin}(1, p)$$

B.1.6 The Negative Binomial Distribution, Type I

The Abbreviation $\text{NegBin}(k, p)$

The Sample Space The integers k, $k + 1$, ...

The Parameter p such that $0 < p < 1$.

The Density

$$f(x) = \binom{x-1}{k-1} p^k (1-p)^{x-k}, \qquad x = k, k+1, \ldots.$$

Moments

$$E(X) = \frac{k}{p}$$
$$\operatorname{var}(X) = \frac{k(1-p)}{p^2}$$

Specialization

$$\operatorname{Geo}(p) = \operatorname{NegBin}(1, p)$$

B.1.7 The Negative Binomial Distribution, Type II

The Abbreviation No abbreviation to avoid confusion with the other type defined in Section B.1.6.

Relation Between the Types If $X \sim \operatorname{NegBin}(k, p)$, then $Y = X - k$ has the distribution defined in this section.

X is the number of *trials* before the k-th success in an i. i. d. sequence of $\operatorname{Ber}(p)$ random variables. Y is the number of *failures* before the k-th success.

The Sample Space The integers 0, 1, ...

The Parameter p such that $0 < p < 1$.

The Density

$$f(x) = \binom{x-1}{k-1} p^k (1-p)^{x}, \qquad x = 0, 1, \ldots.$$

Moments

$$E(X) = \frac{k}{p} - k = \frac{k(1-p)}{p}$$
$$\operatorname{var}(X) = \frac{k(1-p)}{p^2}$$

B.1.8 The Geometric Distribution, Type I

The Abbreviation $\operatorname{Geo}(p)$

The Sample Space The integers 1, 2, ...

The Parameter p such that $0 < p < 1$.

The Density

$$f(x) = p(1-p)^{x-1}, \qquad x = 1, 2, \ldots.$$

Moments

$$E(X) = \frac{1}{p}$$
$$\text{var}(X) = \frac{1-p}{p^2}$$

Generalization

$$\text{Geo}(p) = \text{NegBin}(1, p)$$

B.2 Continuous Distributions

B.2.1 The Uniform Distribution

The Abbreviation $\mathcal{U}(S)$.

The Sample Space Any subset S of $\mathbb{R}^d$.

The Density

$$f(x) = \frac{1}{c}, \qquad x \in S,$$

where

$$c = m(S) = \int_S dx$$

is the measure of S (length in $\mathbb{R}^1$, area in $\mathbb{R}^2$, volume in $\mathbb{R}^3$, and so forth).

Specialization The case having $S = (a, b)$ in $\mathbb{R}^1$ and density

$$f(x) = \frac{1}{b-a}, \qquad a < x < b$$

is denoted $\mathcal{U}(a, b)$.

Moments If $X \sim \mathcal{U}(a, b)$, then

$$E(X) = \frac{a+b}{2}$$
$$\text{var}(X) = \frac{(b-a)^2}{12}$$

B.2.2 The Exponential Distribution

The Abbreviation $\text{Exp}(\lambda)$.

The Sample Space The interval $(0, \infty)$ of the real numbers.

The Parameter λ such that $\lambda > 0$.

The Density

$$f(x) = \lambda e^{-\lambda x}, \qquad x > 0.$$

Moments

$$E(X) = \frac{1}{\lambda}$$
$$\text{var}(X) = \frac{1}{\lambda^2}$$

Generalization

$$\text{Exp}(\lambda) = \text{Gam}(1, \lambda)$$

B.2.3 The Gamma Distribution

The Abbreviation $\text{Gam}(\alpha, \lambda)$.

The Sample Space The interval $(0, \infty)$ of the real numbers.

The Parameters α and λ such that $\alpha > 0$ and $\lambda > 0$.

The Density

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha - 1} e^{-\lambda x}, \qquad x > 0.$$

where $\Gamma(\alpha)$ is the gamma function (Section B.3.1 below).

Moments

$$E(X) = \frac{\alpha}{\lambda}$$
$$\text{var}(X) = \frac{\alpha}{\lambda^2}$$

Specialization

$$\text{Exp}(\lambda) = \text{Gam}(1, \lambda)$$
$$\text{chi}^2(k) = \text{Gam}\left(\tfrac{k}{2}, \tfrac{1}{2}\right)$$

B.2.4 The Beta Distribution

The Abbreviation Beta(s, t).

The Sample Space The interval $(0, 1)$ of the real numbers.

The Parameters s and t such that $s > 0$ and $t > 0$.

The Density

$$f(x) = \frac{1}{B(s,t)} x^{s-1}(1-x)^{t-1} \qquad 0 < x < 1.$$

where $B(s, t)$ is the *beta function* defined by

$$B(s,t) = \frac{\Gamma(s)\Gamma(t)}{\Gamma(s+t)} \tag{B.1}$$

Moments

$$\begin{aligned} E(X) &= \frac{s}{s+t} \\ \operatorname{var}(X) &= \frac{st}{(s+t)^2(s+t+1)} \end{aligned}$$

B.2.5 The Normal Distribution

The Abbreviation $\mathcal{N}(\mu, \sigma^2)$.

The Sample Space The real line $\mathbb{R}$.

The Parameters μ and σ^2 such that $\sigma^2 > 0$.

The Density

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad x \in \mathbb{R}.$$

Moments

$$\begin{aligned} E(X) &= \mu \\ \operatorname{var}(X) &= \sigma^2 \\ \mu_4 &= 3\sigma^4 \end{aligned}$$

B.2.6 The Chi-Square Distribution

The Abbreviation $\text{chi}^2(k)$.

The Sample Space The interval $(0, \infty)$ of the real numbers.

The Parameter A positive integer k.

The Density

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}, \qquad x > 0.$$

Moments

$$E(X) = k$$
$$\operatorname{var}(X) = 2k$$

Generalization

$$\text{chi}^2(k) = \text{Gam}\left(\tfrac{k}{2}, \tfrac{1}{2}\right)$$

B.2.7 The Cauchy Distribution

The Abbreviation Cauchy(μ, σ).

The Sample Space The real line $\mathbb{R}$.

The Parameters μ and σ such that $\sigma > 0$.

The Density

$$f(x) = \frac{1}{\pi} \cdot \frac{\sigma}{\sigma^2 + (x - \mu)^2}, \qquad x \in \mathbb{R}.$$

Moments None: $E(|X|) = \infty$.

B.3 Special Functions

B.3.1 The Gamma Function

The Definition

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx, \qquad \alpha > 0 \tag{B.2}$$

The Recursion Relation

$$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha) \tag{B.3}$$

Known Values

$$\Gamma(1) = 1$$

and hence using the recursion relation

$$\Gamma(n+1) = n!$$

for any nonnegative integer n.

Also

$$\Gamma(\tfrac{1}{2}) = \sqrt{\pi}$$

and hence using the recursion relation

$$\begin{aligned}
\Gamma(\tfrac{3}{2}) &= \tfrac{1}{2}\sqrt{\pi} \\
\Gamma(\tfrac{5}{2}) &= \tfrac{3}{2} \cdot \tfrac{1}{2}\sqrt{\pi} \\
\Gamma(\tfrac{7}{2}) &= \tfrac{5}{2} \cdot \tfrac{3}{2} \cdot \tfrac{1}{2}\sqrt{\pi}
\end{aligned}$$

and so forth.

B.3.2 The Beta Function

The function $B(s, t)$ defined by (B.1).

B.4 Discrete Multivariate Distributions

B.4.1 The Multinomial Distribution

The Abbreviation $\text{Multi}_k(n, \mathbf{p})$ or $\text{Multi}(n, \mathbf{p})$ if the dimension k is clear from context.

The Sample Space

$$S = \{\, \mathbf{y} \in \mathbb{N}^k : y_1 + \cdots y_k = n \,\}$$

where $\mathbb{N}$ denotes the "natural numbers" 0, 1, 2, ...

The Parameter $\mathbf{p} = (p_1, \ldots, p_k)$ such that $p_i \geq 0$ for all i and $\sum_i p_i = 1$.

The Density

$$f(\mathbf{y}) = \binom{n}{y_1, \ldots, y_k} \prod_{j=1}^{k} p_j^{y_j}, \qquad \mathbf{y} \in S$$

Moments

$$E(\mathbf{Y}) = n\mathbf{p}$$
$$\operatorname{var}(\mathbf{Y}) = \mathbf{M}$$

where $\mathbf{M}$ is the $k \times k$ matrix with elements

$$m_{ij} = \begin{cases} np_i(1 - p_i), & i = j \\ -np_ip_j & i \neq j \end{cases}$$

Specialization The special case $n = 1$ is called the multivariate Bernoulli distribution

$$\operatorname{Ber}_k(\mathbf{p}) = \operatorname{Bin}_k(1, \mathbf{p})$$

but for once we will not spell out the details with a special section for the multivariate Bernoulli. Just take $n = 1$ in this section.

Marginal Distributions Distributions obtained by collapsing categories are again multinomial (Section 5.4.5 in these notes).

In particular, if $\mathbf{Y} \sim \operatorname{Multi}_k(n, \mathbf{p})$, then

$$(Y_1, \ldots, Y_j, Y_{j+1} + \cdots + Y_k) \sim \operatorname{Multi}_{j+1}(n, \mathbf{q}) \tag{B.4}$$

where

$$\begin{aligned} q_i &= p_i, && i \leq j \\ q_{j+1} &= p_{j+1} + \cdots p_k \end{aligned}$$

Because the random vector in (B.4) is degenerate, this equation also gives implicitly the marginal distribution of Y_1, ..., Y_j

$$\begin{aligned} &f(y_1, \ldots, y_j) \\ &\quad = \binom{n}{y_1, \ldots, y_j, n - y_1 - \cdots - y_j} p_1^{y_1} \cdots p_j^{y_j} (1 - p_1 - \cdots - p_j)^{n - y_1 - \cdots - y_j} \end{aligned}$$

Univariate Marginal Distributions If $\mathbf{Y} \sim \operatorname{Multi}(n, \mathbf{p})$, then

$$Y_i \sim \operatorname{Bin}(n, p_i).$$

Conditional Distributions If $\mathbf{Y} \sim \operatorname{Multi}_k(n, \mathbf{p})$, then

$$(Y_1, \ldots, Y_j) \mid (Y_{j+1}, \ldots, Y_k) \sim \operatorname{Multi}_j(n - Y_{j+1} - \cdots - Y_k, \mathbf{q}),$$

where

$$q_i = \frac{p_i}{p_1 + \cdots + p_j}, \qquad i = 1, \ldots, j.$$

B.5 Continuous Multivariate Distributions

B.5.1 The Uniform Distribution

The uniform distribution defined in Section B.2.1 actually made no mention of dimension. If the set S on which the distribution is defined lies in $\mathbb{R}^n$, then this is a multivariate distribution.

Conditional Distributions Every conditional distribution of a multivariate uniform distribution is uniform.

Marginal Distributions No regularity. Depends on the particular distribution. Marginals of the uniform distribution on a rectangle with sides parallel to the coordinate axes are uniform. Marginals of the uniform distribution on a disk or triangle are not uniform.

B.5.2 The Standard Normal Distribution

The distribution of a random vector $\mathbf{Z} = (Z_1, \ldots, Z_k)$ with the Z_i i. i. d. standard normal.

Moments

$$\begin{aligned} E(\mathbf{Z}) &= 0 \\ \operatorname{var}(\mathbf{Z}) &= \mathbf{I}, \end{aligned}$$

where $\mathbf{I}$ denotes the $k \times k$ identity matrix.

B.5.3 The Multivariate Normal Distribution

The distribution of a random vector $\mathbf{X} = \mathbf{a} + \mathbf{B}\mathbf{Z}$, where $\mathbf{Z}$ is multivariate standard normal.

Moments

$$\begin{aligned} E(\mathbf{X}) &= \boldsymbol{\mu} = \mathbf{a} \\ \operatorname{var}(\mathbf{X}) &= \mathbf{M} = \mathbf{B}\mathbf{B}' \end{aligned}$$

The Abbreviation $\mathcal{N}_k(\boldsymbol{\mu}, \mathbf{M})$ or $\mathcal{N}(\boldsymbol{\mu}, \mathbf{M})$ if the dimension k is clear from context.

The Sample Space If $\mathbf{M}$ is positive definite, the sample space is $\mathbb{R}^k$.

Otherwise, X is concentrated on the intersection of hyperplanes determined by null eigenvectors of $\mathbf{M}$

$$S = \{\, \mathbf{x} \in \mathbb{R}^k : \mathbf{z}'\mathbf{x} = \mathbf{z}'\boldsymbol{\mu} \text{ whenever } \mathbf{M}\mathbf{z} = 0 \,\}$$

The Parameters The mean vector $\boldsymbol{\mu}$ and variance matrix $\mathbf{M}$.

The Density Only exists if the distribution is nondegenerate ($\mathbf{M}$ is positive definite). Then

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} \det(\mathbf{M})^{1/2}} \exp\left(-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu})' \mathbf{M}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right), \qquad \mathbf{x} \in \mathbb{R}^k$$

Marginal Distributions All are normal. If

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$$

is a partitioned random vector with (partitioned) mean vector

$$E(\mathbf{X}) = \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}$$

and (partitioned) variance matrix

$$\operatorname{var}(\mathbf{X}) = \mathbf{M} = \begin{pmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{pmatrix}$$

and $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{M})$, then

$$\mathbf{X}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{M}_{11}).$$

Conditional Distributions All are normal. If $\mathbf{X}$ is as in the preceding section and $\mathbf{X}_2$ is nondegenerate, then the conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_2$ is normal with

$$\begin{aligned} E(\mathbf{X}_1 \mid \mathbf{X}_2) &= \boldsymbol{\mu}_1 + \mathbf{M}_{12} \mathbf{M}_{22}^{-1} (\mathbf{X}_2 - \boldsymbol{\mu}_2) \\ \operatorname{var}(\mathbf{X}_1 \mid \mathbf{X}_2) &= \mathbf{M}_{11} - \mathbf{M}_{12} \mathbf{M}_{22}^{-1} \mathbf{M}_{21} \end{aligned}$$

If $\mathbf{X}_2$ is degenerate so $\mathbf{M}_{22}$ is not invertible, then the conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_2$ is still normal and the same formulas work if $\mathbf{M}_{22}^{-1}$ is replaced by a generalized inverse.

B.5.4 The Bivariate Normal Distribution

The special case $k = 2$ of the preceeding section.

The Density

$$\begin{aligned} f(x, y) = {} & \frac{1}{2\pi \sigma_X \sigma_Y \sqrt{1 - \rho^2}} \times \\ & \exp\left(-\frac{1}{2(1 - \rho^2)} \left[\frac{(x - \mu_X)^2}{\sigma_X^2} - \frac{2\rho (x - \mu_X)(y - \mu_Y)}{\sigma_X \sigma_Y} + \frac{(y - \mu_Y)^2}{\sigma_Y^2} \right] \right) \end{aligned}$$

Marginal Distributions

$$Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$$

Conditional Distributions The conditional distribution of X given Y is normal with

$$E(X \mid Y) = \mu_X + \rho \frac{\sigma_X}{\sigma_Y}(Y - \mu_Y)$$
$$\text{var}(X \mid Y) = \sigma_X^2(1 - \rho^2)$$

where $\rho = \text{cor}(X, Y)$.

Appendix C

Addition Rules for Distributions

"Addition rules" for distributions are rules of the form: if $X_1, \ldots, X_k$ are independent with some specified distributions, then $X_1 + \cdots + X_k$ has some other specified distribution.

Bernoulli If $X_1, \ldots, X_k$ are i. i. d. $\text{Ber}(p)$, then

$$X_1 + \cdots + X_k \sim \text{Bin}(k, p). \tag{C.1}$$

- All the Bernoulli distributions must have the *same* success probability p.

Binomial If $X_1, \ldots, X_k$ are independent with $X_i \sim \text{Bin}(n_i, p)$, then

$$X_1 + \cdots + X_k \sim \text{Bin}(n_1 + \cdots + n_k, p). \tag{C.2}$$

- All the binomial distributions must have the *same* success probability p.
- (C.1) is the special case of (C.2) obtained by setting $n_1 = \cdots = n_k = 1$.

Geometric If $X_1, \ldots, X_k$ are i. i. d. $\text{Geo}(p)$, then

$$X_1 + \cdots + X_k \sim \text{NegBin}(k, p). \tag{C.3}$$

- All the geometric distributions must have the *same* success probability p.

Negative Binomial If $X_1, \ldots, X_k$ are independent with $X_i \sim \text{NegBin}(n_i, p)$, then

$$X_1 + \cdots + X_k \sim \text{NegBin}(n_1 + \cdots + n_k, p). \tag{C.4}$$

- All the negative binomial distributions must have the *same* success probability p.
- (C.3) is the special case of (C.4) obtained by setting $n_1 = \cdots = n_k = 1$.

Poisson If X_1, ..., X_k are independent with $X_i \sim \text{Poi}(\mu_i)$, then

$$X_1 + \cdots + X_k \sim \text{Poi}(\mu_1 + \cdots + \mu_k). \tag{C.5}$$

Exponential If X_1, ..., X_k are i. i. d. $\text{Exp}(\lambda)$, then

$$X_1 + \cdots + X_k \sim \text{Gam}(n, \lambda). \tag{C.6}$$

- All the exponential distributions must have the *same* rate parameter λ.

Gamma If X_1, ..., X_k are independent with $X_i \sim \text{Gam}(\alpha_i, \lambda)$, then

$$X_1 + \cdots + X_k \sim \text{Gam}(\alpha_1 + \cdots + \alpha_k, \lambda). \tag{C.7}$$

- All the gamma distributions must have the *same* rate parameter λ.
- (C.6) is the special case of (C.7) obtained by setting $\alpha_1 = \cdots = \alpha_k = 1$.

Chi-Square If X_1, ..., X_k are independent with $X_i \sim \text{chi}^2(n_i)$, then

$$X_1 + \cdots + X_k \sim \text{chi}^2(n_1 + \cdots + n_k). \tag{C.8}$$

- (C.8) is the special case of (C.7) obtained by setting

$$\alpha_i = n_i/2 \quad \text{and} \quad \lambda_i = 1/2, \qquad i = 1, \ldots, k.$$

Normal If X_1, ..., X_k are independent with $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, then

$$X_1 + \cdots + X_k \sim \mathcal{N}(\mu_1 + \cdots + \mu_k, \sigma_1^2 + \cdots + \sigma_k^2). \tag{C.9}$$

Linear Combination of Normals If X_1, ..., X_k are independent with $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ and a_1, ..., a_k are constants, then

$$\sum_{i=1}^{k} a_i X_i \sim \mathcal{N}\left(\sum_{i=1}^{k} a_i \mu_i, \sum_{i=1}^{k} a_i^2 \sigma_i^2 \right). \tag{C.10}$$

- (C.9) is the special case of (C.10) obtained by setting $a_1 = \cdots = a_k = 1$.

Cauchy If X_1, ..., X_k are independent with $X_i \sim \text{Cauchy}(\mu, \sigma)$, then

$$X_1 + \cdots + X_k \sim \text{Cauchy}(n\mu, n\sigma). \tag{C.11}$$

Appendix D

Relations Among Brand Name Distributions

D.1 Special Cases

First there are the special cases, which were also noted in Appendix B.

$$\begin{aligned}
\text{Ber}(p) &= \text{Bin}(1, p) \\
\text{Geo}(p) &= \text{NegBin}(1, p) \\
\text{Exp}(\lambda) &= \text{Gam}(1, \lambda) \\
\text{chi}^2(k) &= \text{Gam}\left(\tfrac{k}{2}, \tfrac{1}{2}\right)
\end{aligned}$$

The main point of this appendix are the relationships that involve more theoretical issues.

D.2 Relations Involving Bernoulli Sequences

Suppose X_1, X_2, ... are i. i. d. Ber(p) random variables.
If n is a positive integer and

$$Y = X_1 + \cdots + X_n$$

is the number of "successes" in the n Bernoulli trials, then

$$Y \sim \text{Bin}(n, p).$$

On the other hand, if y is positive integer and N is the trial at which the y-th success occurs, that is the random number N such that

$$\begin{aligned}
X_1 + \cdots + X_N &= y \\
X_1 + \cdots + X_k &< y, \qquad k < N,
\end{aligned}$$

then

$$N \sim \text{NegBin}(y, p).$$

D.3 Relations Involving Poisson Processes

In a one-dimensional homogeneous Poisson process with rate parameter λ, the counts are Poisson and the waiting and interarrival times are exponential. Specifically, the number of points (arrivals) in an interval of length t has the $\text{Poi}(\lambda t)$ distribution, and the waiting times and interarrival times are independent and indentically $\text{Exp}(\lambda)$ distributed.

Even more specifically, let X_1, X_2, ... be i. i. d. $\text{Exp}(\lambda)$ random variables. Take these to be the waiting and interarrival times of a Poisson process. This means the arrival times themselves are

$$T_k = \sum_{i=1}^{k} X_i$$

Note that

$$0 < T_1 < T_2 < \cdots$$

and

$$X_i = T_i - T_{i-1}, \qquad i > 1$$

so these are the interarrival times and $X_1 = T_1$ is the waiting time until the first arrival.

The characteristic property of the Poisson process, that counts have the Poisson distribution, says the number of points in the interval $(0, t)$, that is, the number of T_i such that $T_i < t$, has the $\text{Poi}(\lambda t)$ distribution.

D.4 Normal and Chi-Square

If Z_1, Z_2, ... are i. i. d. $\mathcal{N}(0, 1)$, then

$$Z_1^2 + \ldots Z_n^2 \sim \text{chi}^2(n).$$

Appendix E

Eigenvalues and Eigenvectors

E.1 Orthogonal and Orthonormal Vectors

If $\mathbf{x}$ and $\mathbf{y}$ are vectors of the same dimension, we say they are *orthogonal* if $\mathbf{x}'\mathbf{y} = 0$. Since the transpose of a matrix product is the product of the transposes in reverse order, an equivalent condition is $\mathbf{y}'\mathbf{x} = 0$. Orthogonality is the n-dimensional generalization of perpendicularity. In a sense, it says that two vectors make a right angle.

The *length* or *norm* of a vector $\mathbf{x} = (x_1, \ldots, x_n)$ is defined to be

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}'\mathbf{x}} = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

Squaring both sides gives

$$\|\mathbf{x}\|^2 = \sum_{i=1}^{n} x_i^2,$$

which is one version of the Pythagorean theorem, as it appears in analytic geometry.

Orthogonal vectors give another generalization of the Pythagorean theorem. We say a set of vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ is *orthogonal* if

$$\mathbf{x}_i'\mathbf{x}_j = 0, \qquad i \neq j. \tag{E.1}$$

Then

$$\begin{aligned}\|\mathbf{x}_1 + \cdots + \mathbf{x}_k\|^2 &= (\mathbf{x}_1 + \cdots + \mathbf{x}_k)'(\mathbf{x}_1 + \cdots + \mathbf{x}_k) \\ &= \sum_{i=1}^{k} \sum_{j=1}^{k} \mathbf{x}_i' \mathbf{x}_j \\ &= \sum_{i=1}^{k} \mathbf{x}_i' \mathbf{x}_i \\ &= \sum_{i=1}^{k} \|\mathbf{x}_i\|^2\end{aligned}$$

because, by definition of orthogonality, all terms in the second line with $i \neq j$ are zero.

We say an orthogonal set of vectors is *orthonormal* if

$$\mathbf{x}_i' \mathbf{x}_i = 1. \tag{E.2}$$

That is, a set of vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ is orthonormal if it satisfies both (E.1) and (E.2).

An orthonormal set is automatically linearly independent because if

$$\sum_{i=1}^{k} c_i \mathbf{x}_i = 0,$$

then

$$0 = \mathbf{x}_j' \left(\sum_{i=1}^{k} c_i \mathbf{x}_i \right) = c_j \mathbf{x}_j' \mathbf{x}_j = c_j$$

holds for all j. Hence the only linear combination that is zero is the one with all coefficients zero, which is the definition of linear independence.

Being linearly independent, an orthonormal set is always a *basis* for whatever subspace it spans. If we are working in n-dimensional space, and there are n vectors in the orthonormal set, then they make up a basis for the whole space. If there are $k < n$ vectors in the set, then they make up a basis for some proper subspace.

It is always possible to choose an orthogonal basis for any vector space or subspace. One way to do this is the Gram-Schmidt orthogonalization procedure, which converts an arbitrary basis $\mathbf{y}_1$, ..., $\mathbf{y}_n$ to an orthonormal basis $\mathbf{x}_1$, ..., $\mathbf{x}_n$ as follows. First let

$$\mathbf{x}_1 = \frac{\mathbf{y}_1}{\|\mathbf{y}_1\|}.$$

Then define the $\mathbf{x}_i$ in order. After $\mathbf{x}_1$, ..., $\mathbf{x}_{k-1}$ have been defined, let

$$\mathbf{z}_k = \mathbf{y}_k - \sum_{i=1}^{k-1} \mathbf{x}_i \mathbf{x}_i' \mathbf{y}$$

and

$$\mathbf{x}_k = \frac{\mathbf{z}_k}{\|\mathbf{z}_k\|}.$$

It is easily verified that this does produce an orthonormal set, and it is only slightly harder to prove that none of the $\mathbf{x}_i$ are zero because that would imply linear dependence of the $\mathbf{y}_i$.

E.2 Eigenvalues and Eigenvectors

If $\mathbf{A}$ is any matrix, we say that λ is a *right eigenvalue* corresponding to a *right eigenvector* $\mathbf{x}$ if

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$$

Left eigenvalues and eigenvectors are defined analogously with "left multiplication" $\mathbf{x}'\mathbf{A} = \lambda\mathbf{x}'$, which is equivalent to $\mathbf{A}'\mathbf{x} = \lambda\mathbf{x}$. So the right eigenvalues and eigenvectors of $\mathbf{A}'$ are the left eigenvalues and eigenvectors of $\mathbf{A}$. When $\mathbf{A}$ is symmetric ($\mathbf{A}' = \mathbf{A}$), the "left" and "right" concepts are the same and the adjectives "left" and "right" are unnecessary. Fortunately, this is the most interesting case, and the only one in which we will be interested. From now on we discuss only eigenvalues and eigenvectors of *symmetric* matrices.

There are three important facts about eigenvalues and eigenvectors. Two elementary and one very deep. Here's the first (one of the elementary facts).

Lemma E.1. *Eigenvectors corresponding to distinct eigenvalues are orthogonal.*

This means that if

$$\mathbf{A}\mathbf{x}_i = \lambda_i\mathbf{x}_i \tag{E.3}$$

then

$$\lambda_i \neq \lambda_j \qquad \text{implies} \qquad \mathbf{x}_i'\mathbf{x}_j = 0.$$

Proof. Suppose $\lambda_i \neq \lambda_j$, then at least one of the two is not zero, say λ_j. Then

$$\mathbf{x}_i'\mathbf{x}_j = \frac{\mathbf{x}_i'\mathbf{A}\mathbf{x}_j}{\lambda_j} = \frac{(\mathbf{A}\mathbf{x}_i)'\mathbf{x}_j}{\lambda_j} = \frac{\lambda_i\mathbf{x}_i'\mathbf{x}_j}{\lambda_j} = \frac{\lambda_i}{\lambda_j} \cdot \mathbf{x}_i'\mathbf{x}_j$$

and since $\lambda_i \neq \lambda_j$ the only way this can happen is if $\mathbf{x}_i'\mathbf{x}_j = 0$. □

Here's the second important fact (also elementary).

Lemma E.2. *Every linear combination of eigenvectors corresponding to the same eigenvalue is another eigenvector corresponding to that eigenvalue.*

This means that if

$$\mathbf{A}\mathbf{x}_i = \lambda\mathbf{x}_i$$

then

$$\mathbf{A}\left(\sum_{i=1}^{k} c_i\mathbf{x}_i\right) = \lambda\left(\sum_{i=1}^{k} c_i\mathbf{x}_i\right)$$

Proof. This is just linearity of matrix multiplication. □

The second property means that all the eigenvectors corresponding to one eigenvalue constitute a subspace. If the dimension of that subspace is k, then it is possible to choose an orthonormal basis of k vectors that span the subspace. Since the first property of eigenvalues and eigenvectors says that (E.1) is also satisfied by eigenvectors corresponding to different eigenvalues, all of the eigenvectors chosen this way form an orthonormal set.

Thus our orthonormal set of eigenvectors spans a subspace of dimension m which contains all eigenvectors of the matrix in question. The question then arises whether this set is *complete*, that is, whether it is a basis for the whole space, or in symbols whether $m = n$, where n is the dimension of the whole space ($\mathbf{A}$ is an $n \times n$ matrix and the $\mathbf{x}_i$ are vectors of dimension n). It turns out that the set *is* always complete, and this is the third important fact about eigenvalues and eigenvectors.

Lemma E.3. *Every real symmetric matrix has an orthonormal set of eigenvectors that form a basis for the space.*

In contrast to the first two facts, this is deep, and we shall not say anything about its proof, other than that about half of the typical linear algebra book is given over to building up to the proof of this one fact.

The "third important fact" says that *any* vector can be written as a linear combination of eigenvectors

$$\mathbf{y} = \sum_{i=1}^{n} c_i \mathbf{x}_i$$

and this allows a very simple description of the action of the linear operator described by the matrix

$$\mathbf{A}\mathbf{y} = \sum_{i=1}^{n} c_i \mathbf{A}\mathbf{x}_i = \sum_{i=1}^{n} c_i \lambda_i \mathbf{x}_i \tag{E.4}$$

So this says that *when we use an orthonormal eigenvector basis*, if $\mathbf{y}$ has the representation $(c_1, \ldots, c_n)$, then $\mathbf{A}y$ has the representation $(c_1\lambda_1, \ldots, c_n\lambda_n)$. Let $\mathbf{D}$ be the representation in the orthonormal eigenvector basis of the linear operator represented by $\mathbf{A}$ in the standard basis. Then our analysis above says the i-the element of $\mathbf{D}\mathbf{c}$ is $c_i\lambda_i$, that is,

$$\sum_{j=1}^{n} d_{ij} c_j = \lambda_i c_i.$$

In order for this to hold for all real numbers c_i, it must be that $\mathbf{D}$ is diagonal

$$\begin{aligned} d_{ii} &= \lambda_i \\ d_{ij} &= 0, \qquad i \neq j \end{aligned}$$

In short, using the orthonormal eigenvector basis *diagonalizes* the linear operator represented by the matrix in question.

There is another way to describe this same fact without mentioning bases. Many people find it a simpler description, though its relation to eigenvalues and eigenvectors is hidden in the notation, no longer immediately apparent. Let $\mathbf{O}$ denote the matrix whose columns are the orthonormal eigenvector basis ($\mathbf{x}_1$, ..., $\mathbf{x}_n$), that is, if o_{ij} are the elements of $\mathbf{O}$, then

$$\mathbf{x}_i = (o_{1i}, \ldots, o_{ni}).$$

Now (E.1) and (E.2) can be combined as one matrix equation

$$\mathbf{O}'\mathbf{O} = \mathbf{I} \tag{E.5}$$

(where, as usual, $\mathbf{I}$ is the $n \times n$ identity matrix). A matrix $\mathbf{O}$ satisfying this property is said to be *orthogonal*. Another way to read (E.5) is that it says $\mathbf{O}' = \mathbf{O}^{-1}$ (an orthogonal matrix is one whose inverse is its transpose). The fact that inverses are two-sided ($\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ for any invertible matrix $\mathbf{A}$) implies that $\mathbf{O}\mathbf{O}' = \mathbf{I}$ as well.

Furthermore, the eigenvalue-eigenvector equation (E.3) can be written out with explicit subscripts and summations as

$$\sum_{j=1}^{n} a_{ij} o_{jk} = \lambda_k o_{ik} = o_{ik} d_{kk} = \sum_{j=1}^{n} o_{ij} d_{jk}$$

(where $\mathbf{D}$ is the the diagonal matrix with eigenvalues on the diagonal defined above). Going back to matrix notation gives

$$\mathbf{A}\mathbf{O} = \mathbf{O}\mathbf{D} \tag{E.6}$$

The two equations (E.3) and (E.6) may not look much alike, but as we have just seen, they say exactly the same thing in different notation. Using the orthogonality property ($\mathbf{O}' = \mathbf{O}^{-1}$) we can rewrite (E.6) in two different ways.

Theorem E.4 (Spectral Decomposition). *Any real symmetric matrix* $\mathbf{A}$ *can be written*

$$\mathbf{A} = \mathbf{O}\mathbf{D}\mathbf{O}' \tag{E.7}$$

where $\mathbf{D}$ *is diagonal and* $\mathbf{O}$ *is orthogonal.*

Conversely, for any real symmetric matrix $\mathbf{A}$ *there exists an orthogonal matrix* $\mathbf{O}$ *such that*

$$\mathbf{D} = \mathbf{O}'\mathbf{A}\mathbf{O}$$

is diagonal.

(The reason for the name of the theorem is that the set of eigenvalues is sometimes called the *spectrum* of $\mathbf{A}$). The spectral decomposition theorem says nothing about eigenvalues and eigenvectors, but we know from the discussion above that the diagonal elements of $\mathbf{D}$ are the eigenvalues of $\mathbf{A}$, and the columns of $\mathbf{O}$ are the corresponding eigenvectors.

E.3 Positive Definite Matrices

Using the spectral theorem, we can prove several interesting things about positive definite matrices.

Corollary E.5. *A real symmetric matrix* $\mathbf{A}$ *is positive semi-definite if and only if its spectrum is nonnegative. A real symmetric matrix* $\mathbf{A}$ *is positive definite if and only if its spectrum is strictly positive.*

Proof. First suppose that $\mathbf{A}$ is positive semi-definite with spectral decomposition (E.7). Let $\mathbf{e}_i$ denote the vector having elements that are all zero except the i-th, which is one, and define $\mathbf{w} = \mathbf{O}\mathbf{e}_i$, so

$$0 \le \mathbf{w}'\mathbf{A}\mathbf{w} = \mathbf{e}_i'\mathbf{O}'\mathbf{O}\mathbf{D}\mathbf{O}'\mathbf{O}\mathbf{e}_i = \mathbf{e}_i'\mathbf{D}\mathbf{e}_i = d_{ii} \tag{E.8}$$

using $\mathbf{O}'\mathbf{O} = I$. Hence the spectrum is nonnegative.

Conversely, suppose the d_{ii} are nonnegative. Then for any vector $\mathbf{w}$ define $\mathbf{z} = \mathbf{O}'\mathbf{w}$, so

$$\mathbf{w}'\mathbf{A}\mathbf{w} = \mathbf{w}'\mathbf{O}\mathbf{D}\mathbf{O}'\mathbf{w} = \mathbf{z}'\mathbf{D}\mathbf{z} = \sum\nolimits_i d_{ii} z_i^2 \ge 0$$

Hence $\mathbf{A}$ is positive semi-definite.

The assertions about positive definiteness are proved in almost the same way. Suppose that $\mathbf{A}$ is positive definite. Since $\mathbf{e}_i$ is nonzero, $\mathbf{w}$ in (E.8) is also nonzero because $\mathbf{e}_i = \mathbf{O}'\mathbf{w}$ would be zero (and it isn't) if $\mathbf{w}$ were zero. Thus the inequality in (E.8) is actually strict. Hence the spectrum of is strictly positive.

Conversely, suppose the d_{ii} are strictly positive. Then for any nonzero vector $\mathbf{w}$ define $\mathbf{z} = \mathbf{O}'\mathbf{w}$ as before, and again note that $\mathbf{z}$ is nonzero because $\mathbf{w} = \mathbf{O}\mathbf{z}$ and $\mathbf{w}$ is nonzero. Thus $\mathbf{w}'\mathbf{A}\mathbf{w} = \mathbf{z}'\mathbf{D}\mathbf{z} > 0$, and hence $\mathbf{A}$ is positive definite. □

Corollary E.6. *A positive semi-definite matrix is invertible if and only if it is positive definite.*

Proof. It is easily verified that the product of diagonal matrices is diagonal and the diagonal elements of the product are the products of the diagonal elements of the multiplicands. Thus a diagonal matrix $\mathbf{D}$ is invertible if and only if all its diagonal elements d_{ii} are nonzero, in which case $\mathbf{D}^{-1}$ is diagonal with diagonal elements $1/d_{ii}$.

Since $\mathbf{O}$ and $\mathbf{O}'$ in the spectral decomposition (E.7) are invertible, $\mathbf{A}$ is invertible if and only if $\mathbf{D}$ is, hence if and only if its spectrum is nonzero, in which case

$$\mathbf{A}^{-1} = \mathbf{O}\mathbf{D}^{-1}\mathbf{O}'.$$

By the preceding corollary the spectrum of a positive semi-definite matrix is nonnegative, hence nonzero if and only if strictly positive, which (again by the preceding corollary) occurs if and only if the matrix is positive definite. □

Corollary E.7. *Every real symmetric positive semi-definite matrix* $\mathbf{A}$ *has a symmetric square root*

$$\mathbf{A}^{1/2} = \mathbf{O}\mathbf{D}^{1/2}\mathbf{O}' \tag{E.9}$$

where (E.7) *is the spectral decomposition of* $\mathbf{A}$ *and where* $\mathbf{D}^{1/2}$ *is defined to be the diagonal matrix whose diagonal elements are* $\sqrt{d_{ii}}$*, where* d_{ii} *are the diagonal elements of* $\mathbf{D}$*.*

Moreover, $\mathbf{A}^{1/2}$ *is positive definite if and only if* $\mathbf{A}$ *is positive definite.*

Note that by Corollary E.5 all of the diagonal elements of $\mathbf{D}$ are nonnegative and hence have real square roots.

Proof.

$$\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{O}\mathbf{D}^{1/2}\mathbf{O}'\mathbf{O}\mathbf{D}^{1/2}\mathbf{O}' = \mathbf{O}\mathbf{D}^{1/2}\mathbf{D}^{1/2}\mathbf{O}' = \mathbf{O}\mathbf{D}\mathbf{O}' = \mathbf{A}$$

because $\mathbf{O}'\mathbf{O} = \mathbf{I}$ and $\mathbf{D}^{1/2}\mathbf{D}^{1/2} = \mathbf{D}$.

From Corollary E.5 we know that $\mathbf{A}$ is positive definite if and only if all the d_{ii} are strictly positive. Since (E.9) is the spectral decomposition of $\mathbf{A}^{1/2}$, we see that $\mathbf{A}^{1/2}$ is positive definite if and only if all the $\sqrt{d_{ii}}$ are strictly positive. Clearly $d_{ii} > 0$ if and only if $\sqrt{d_{ii}} > 0$. □

Appendix F

Normal Approximations for Distributions

F.1 Binomial Distribution

The $\text{Bin}(n, p)$ distribution is approximately normal with mean np and variance $np(1-p)$ if n is large.

F.2 Negative Binomial Distribution

The $\text{NegBin}(n, p)$ distribution is approximately normal with mean n/p and variance $n(1-p)/p^2$ if n is large.

F.3 Poisson Distribution

The $\text{Poi}(\mu)$ distribution is approximately normal with mean μ and variance μ if μ is large.

F.4 Gamma Distribution

The $\text{Gam}(\alpha, \lambda)$ distribution is approximately normal with mean α/λ and variance α/λ^2 if α is large.

F.5 Chi-Square Distribution

The $\text{chi}^2(n)$ distribution is approximately normal with mean n and variance $2n$ if n is large.