Appendix B

ADDITIONAL TOPICS IN PROBABILITY AND ANALYSIS

In this appendix we give some results in probability theory, matrix algebra, and analysis that are essential in our treatment of statistics and that may not be treated in enough detail in more specialized texts. Some of the material in this appendix, as well as extensions, can be found in Anderson (1958), Billingsley (1995), Breiman (1968), Chung (1978), Dempster (1969), Feller (1971), Loeve (1977), and Rao (1973).

Measure theory will not be used. We make the blanket assumption that all sets and functions considered are measurable.

B.1 CONDITIONING BY A RANDOM VARIABLE OR VECTOR

The concept of conditioning is important in studying associations between random variables or vectors. In this section we present some results useful for prediction theory, estimation theory, and regression.

B.1.1 The Discrete Case

The reader is already familiar with the notion of the conditional probability of an event A given that another event B has occurred. If $\mathbf{Y}$ and $\mathbf{Z}$ are discrete random vectors possibly of different dimensions, we want to study the conditional probability structure of $\mathbf{Y}$ given that $\mathbf{Z}$ has taken on a particular value $\mathbf{z}$.

Define the *conditional frequency function* $p(\cdot \mid \mathbf{z})$ *of* $\mathbf{Y}$ *given* $\mathbf{Z} = \mathbf{z}$ by

$$p(\mathbf{y} \mid \mathbf{z}) = P[\mathbf{Y} = \mathbf{y} \mid \mathbf{Z} = \mathbf{z}] = \frac{p(\mathbf{y}, \mathbf{z})}{p_{\mathbf{Z}}(\mathbf{z})} \tag{B.1.1}$$

where p and $p_{\mathbf{Z}}$ are the frequency functions of $(\mathbf{Y}, \mathbf{Z})$ and $\mathbf{Z}$. The conditional frequency function p is defined only for values of $\mathbf{z}$ such that $p_{\mathbf{Z}}(\mathbf{z}) > 0$. With this definition it is

TABLE B.1

y \ z	0	10	20	$p_Y(y)$
0	0.25	0.05	0.05	0.35
1	0.05	0.15	0.05	0.25
2	0.05	0.10	0.25	0.40
$p_Z(z)$	0.35	0.30	0.35	1

clear that $p(\cdot \mid \mathbf{z})$ is the frequency of a probability distribution because

$$\Sigma_{\mathbf{y}} p(\mathbf{y} \mid \mathbf{z}) = \frac{\Sigma_{\mathbf{y}} p(\mathbf{y}, \mathbf{z})}{p_{\mathbf{Z}}(\mathbf{z})} = \frac{p_{\mathbf{Z}}(\mathbf{z})}{p_{\mathbf{Z}}(\mathbf{z})} = 1$$

by $(A.8.11)$. This probability distribution is called the *conditional distribution of* $\mathbf{Y}$ *given that* $\mathbf{Z} = \mathbf{z}$.

Example B.1.1 Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$, where the Y_i are the indicators of a set of n Bernoulli trials with success probability p. Let $Z = \Sigma_{i=1}^n Y_i$, the total number of successes. Then Z has a binomial, $\mathcal{B}(n, p)$, distribution and

$$p(\mathbf{y} \mid z) = \frac{P[\mathbf{Y} = \mathbf{y}, Z = z]}{\binom{n}{z} p^z (1-p)^{n-z}} = \frac{p^z (1-p)^{n-z}}{\binom{n}{z} p^z (1-p)^{n-z}} = \frac{1}{\binom{n}{z}} \tag{B.1.2}$$

if the y_i are all 0 or 1 and $\Sigma y_i = z$.

Thus, if we are told we obtained k successes in n binomial trials, then these successes are as likely to occur on one set of trials as on any other. □

Example B.1.2 Let Y and Z have the joint frequency function given by the table
For instance, suppose Z is the number of cigarettes that a person picked at random from a certain population smokes per day (to the nearest 10), and Y is a general health rating for the same person with 0 corresponding to good, 2 to poor, and 1 to neither. We find for $z = 20$

y	0	1	2
$p(y \mid 20)$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{5}{7}$

These figures would indicate an association between heavy smoking and poor health because $p(2 \mid 20)$ is almost twice as large as $p_Y(2)$. □

The conditional distribution of $\mathbf{Y}$ given $\mathbf{Z} = \mathbf{z}$ is easy to calculate in two special cases.

(i) If $\mathbf{Y}$ and $\mathbf{Z}$ are independent, then $p(\mathbf{y} \mid \mathbf{z}) = p_{\mathbf{Y}}(\mathbf{y})$ and the conditional distribution coincides with the marginal distribution.

(ii) If $\mathbf{Y}$ is a function of $\mathbf{Z}$, $h(\mathbf{Z})$, then the conditional distribution of $\mathbf{Y}$ is degenerate, $\mathbf{Y} = h(\mathbf{Z})$ with probability 1.

Both of these assertions follow immediately from Definition(B.1.1).

Two important formulae follow from (B.1.1) and (A.4.5). Let $q(\mathbf{z} \mid \mathbf{y})$ denote the conditional frequency function of $\mathbf{Z}$ given $\mathbf{Y} = \mathbf{y}$. Then

$$p(\mathbf{y}, \mathbf{z}) = p(\mathbf{y} \mid \mathbf{z}) p_{\mathbf{Z}}(\mathbf{z}) \tag{B.1.3}$$

$$p(\mathbf{y} \mid \mathbf{z}) = \frac{q(\mathbf{z} \mid \mathbf{y}) p_{\mathbf{Y}}(\mathbf{y})}{\Sigma_{\mathbf{y}} q(\mathbf{z} \mid \mathbf{y}) p_{\mathbf{Y}}(\mathbf{y})} \qquad \text{Bayes' Theorem} \tag{B.1.4}$$

whenever the denominator of the right-hand side is positive.

Equation (B.1.3) can be used for model construction. For instance, suppose that the number Z of defectives in a lot of N produced by a manufacturing process has a $\mathcal{B}(N, \theta)$ distribution. Suppose the lot is sampled n times without replacement and let Y be the number of defectives found in the sample. We know that given $Z = z$, Y has a hypergeometric, $\mathcal{H}(z, N, n)$, distribution. We can now use (B.1.3) to write down the joint distribution of Y and Z

$$P[Y = y, Z = z] = \binom{N}{z} \theta^z (1-\theta)^{N-z} \frac{\binom{z}{y}\binom{N-z}{n-y}}{\binom{N}{n}}$$

where the combinatorial coefficients $\binom{a}{b}$ vanish unless a, b are integers with $b \leq a$.

We can also use this model to illustrate (B.1.4). Because we would usually only observe Y, we may want to know what the conditional distribution of Z given $Y = y$ is. By (B.1.4) this is

$$P[Z = z \mid Y = y] = \binom{N}{z} \theta^z (1-\theta)^{N-z} \binom{z}{y}\binom{N-z}{n-y} \Big/ c(y) \tag{B.1.5}$$

where

$$c(y) = \Sigma_z \binom{N}{z} \theta^z (1-\theta)^{N-z} \binom{z}{y}\binom{N-z}{n-y}.$$

This formula simplifies to (see Problem B.1.11) the binomial probability,

$$P[Z = z \mid Y = y] = \binom{N-n}{z-y} \theta^{z-y} (1-\theta)^{N-n-(z-y)}. \tag{B.1.6}$$

B.1.2 Conditional Expectation for Discrete Variables

Suppose that Y is a random variable with $E(|Y|) < \infty$. Define the *conditional expectation of Y given* $\mathbf{Z} = \mathbf{z}$, written $E(Y \mid \mathbf{Z} = \mathbf{z})$, by

$$E(Y \mid \mathbf{Z} = \mathbf{z}) = \Sigma_y y p(y \mid \mathbf{z}). \tag{B.1.7}$$

Note that by (B.1.1), if $p_{\mathbf{Z}}(\mathbf{z}) > 0$,

$$E(|Y| \mid \mathbf{Z} = \mathbf{z}) = \Sigma_y |y| p(y \mid \mathbf{z}) \leq \Sigma_y |y| \frac{p_Y(y)}{p_{\mathbf{Z}}(\mathbf{z})} = \frac{E(|Y|)}{p_{\mathbf{Z}}(\mathbf{z})}. \tag{B.1.8}$$

Thus, when $p_{\mathbf{z}}(\mathbf{z}) > 0$, the conditional expected value of Y is finite whenever the expected value is finite.

Example B.1.3 Suppose Y and Z have the joint frequency function of Table B.1. We find

$$E(Y \mid Z = 20) = 0 \cdot \frac{1}{7} + 1 \cdot \frac{1}{7} + 2 \cdot \frac{5}{7} = \frac{11}{7} = 1.57.$$

Similarly, $E(Y \mid Z = 10) = \frac{7}{6} = 1.17$ and $E(Y \mid Z = 0) = \frac{3}{7} = 0.43$. Note that in the health versus smoking context, we can think of $E(Y \mid Z = z)$ as the mean health rating for people who smoke z cigarettes a day. □

Let $g(\mathbf{z}) = E(Y \mid \mathbf{Z} = \mathbf{z})$. The random variable $g(\mathbf{Z})$ is written $E(Y \mid \mathbf{Z})$ and is called the *conditional expectation of* Y *given* $\mathbf{Z}$.[(1)]

As an example we calculate $E(Y_1 \mid Z)$ where Y_1 and Z are given in Example B.1.1. We have

$$E(Y_1 \mid Z = i) = P[Y_1 = 1 \mid Z = i] = \frac{\binom{n-1}{i-1}}{\binom{n}{i}} = \frac{i}{n}. \tag{B.1.9}$$

The first of these equalities holds because Y_1 is an indicator. The second follows from (B.1.2) because $\binom{n-1}{i-1}$ is just the number of ways i successes can occur in n Bernoulli trials with the first trial being a success. Therefore,

$$E(Y_1 \mid Z) = \frac{Z}{n}. \tag{B.1.10}$$

B.1.3 Properties of Conditional Expected Values

In the context of Section A.4, the conditional distribution of a random vector $\mathbf{Y}$ given $\mathbf{Z} = \mathbf{z}$ corresponds to a single probability measure $P_{\mathbf{z}}$ on $(\Omega, \mathcal{A})$. Specifically, define for $A \in \mathcal{A}$,

$$P_{\mathbf{z}}(A) = P(A \mid [\mathbf{Z} = \mathbf{z}]) \text{ if } p_{\mathbf{Z}}(\mathbf{z}) > 0. \tag{B.1.11}$$

This $P_{\mathbf{z}}$ is just the conditional probability measure on $(\Omega, \mathcal{A})$ mentioned in (A.4.2). Now the conditional distribution of $\mathbf{Y}$ given $\mathbf{Z} = \mathbf{z}$ is the same as the distribution of $\mathbf{Y}$ if $P_{\mathbf{z}}$ is the probability measure on $(\Omega, \mathcal{A})$. Therefore, the conditional expectation *is* an ordinary expectation with respect to the probability measure $P_{\mathbf{z}}$. It follows that all the properties of the expectation given in (A.10.3)–(A.10.8) hold for the conditional expectation given $\mathbf{Z} = \mathbf{z}$. Thus, for any real-valued function $r(\mathbf{Y})$ with $E|r(\mathbf{Y})| < \infty$,

$$E(r(\mathbf{Y}) \mid \mathbf{Z} = \mathbf{z}) = \Sigma_{\mathbf{y}} r(\mathbf{y}) p(\mathbf{y} \mid \mathbf{z})$$

and

$$E(\alpha Y_1 + \beta Y_2 \mid \mathbf{Z} = \mathbf{z}) = \alpha E(Y_1 \mid \mathbf{Z} = \mathbf{z}) + \beta E(Y_2 \mid \mathbf{Z} = \mathbf{z}) \tag{B.1.12}$$

identically in $\mathbf{z}$ for any Y_1, Y_2 such that $E(|Y_1|)$, $E(|Y_2|)$ are finite. Because the identity holds for all $\mathbf{z}$, we have

$$E(\alpha Y_1 + \beta Y_2 \mid \mathbf{Z}) = \alpha E(Y_1 \mid \mathbf{Z}) + \beta E(Y_2 \mid \mathbf{Z}). \tag{B.1.13}$$

This process can be repeated for each of (A.10.3)–(A.10.8) to obtain analogous properties of the conditional expectations.

In two special cases we can calculate conditional expectations immediately. If Y and $\mathbf{Z}$ are independent and $E(|Y|) < \infty$, then

$$E(Y \mid \mathbf{Z}) = E(Y). \tag{B.1.14}$$

This is clear by (i).

On the other hand, by (ii)

$$E(h(\mathbf{Z}) \mid \mathbf{Z}) = h(\mathbf{Z}). \tag{B.1.15}$$

The notion implicit in (B.1.15) is that given $\mathbf{Z} = \mathbf{z}$, $\mathbf{Z}$ acts as a constant. If we carry this further, we have a relation that we shall call the *substitution theorem for conditional expectations*:

$$E(q(\mathbf{Y}, \mathbf{Z}) \mid \mathbf{Z} = \mathbf{z}) = E(q(\mathbf{Y}, \mathbf{z}) \mid \mathbf{Z} = \mathbf{z}). \tag{B.1.16}$$

This is valid for all $\mathbf{z}$ such that $p_{\mathbf{Z}}(\mathbf{z}) > 0$ if $E|q(\mathbf{Y}, \mathbf{Z})| < \infty$. This follows from definitions (B.1.11) and (B.1.7) because

$$P[q(\mathbf{Y}, \mathbf{Z}) = a \mid \mathbf{Z} = \mathbf{z}] = P[q(\mathbf{Y}, \mathbf{Z}) = a,\ \mathbf{Z} = \mathbf{z} \mid \mathbf{Z} = \mathbf{z}] = P[q(\mathbf{Y}, \mathbf{z}) = a \mid \mathbf{Z} = \mathbf{z}] \tag{B.1.17}$$

for any a.

If we put $q(\mathbf{Y}, \mathbf{Z}) = r(\mathbf{Y})h(\mathbf{Z})$, where $E|r(\mathbf{Y})h(\mathbf{Z})| < \infty$, we obtain by (B.1.16),

$$E(r(\mathbf{Y})h(\mathbf{Z}) \mid \mathbf{Z} = \mathbf{z}) = E(r(\mathbf{Y})h(\mathbf{z}) \mid \mathbf{Z} = \mathbf{z}) = h(\mathbf{z})E(r(\mathbf{Y}) \mid \mathbf{Z} = \mathbf{z}). \tag{B.1.18}$$

Therefore,

$$E(r(\mathbf{Y})h(\mathbf{Z}) \mid \mathbf{Z}) = h(\mathbf{Z})E(r(\mathbf{Y}) \mid \mathbf{Z}). \tag{B.1.19}$$

Another intuitively reasonable result is that the mean of the conditional means is the mean:

$$E(E(Y \mid \mathbf{Z})) = E(Y), \tag{B.1.20}$$

whenever Y has a finite expectation. We refer to this as the *double* or *iterated expectation theorem*.

To prove (B.1.20) we write, in view of (B.1.7) and (A.10.5),

$$E(E(Y \mid \mathbf{Z})) = \Sigma_{\mathbf{z}} p_{\mathbf{Z}}(\mathbf{z})[\Sigma_y y p(y \mid \mathbf{z})] = \Sigma_{y,\mathbf{z}} y p(y \mid \mathbf{z}) p_{\mathbf{Z}}(\mathbf{z}) = \Sigma_{y,\mathbf{z}} y p(y, \mathbf{z}) = E(Y). \tag{B.1.21}$$

The interchange of summation used is valid because the finiteness of $E(|Y|)$ implies that all sums converge absolutely.

As an illustration, we check (B.1.20) for $E(Y_1 \mid Z)$ given by (B.1.10). In this case,

$$E(E(Y_1 \mid Z)) = E\left(\frac{Z}{n}\right) = \frac{np}{n} = p = E(Y_1). \tag{B.1.22}$$

If we apply (B.1.20) to $Y = r(\mathbf{Y})h(\mathbf{Z})$ and use (B.1.19), we obtain the *product expectation formula*:

Theorem B.1.1 *If $E|r(\mathbf{Y})h(\mathbf{Z})| < \infty$, then*

$$E(r(\mathbf{Y})h(\mathbf{Z})) = E(h(\mathbf{Z})E(r(\mathbf{Y}) \mid \mathbf{Z})). \tag{B.1.23}$$

Note that we can express the conditional probability that $\mathbf{Y} \in A$ given $\mathbf{Z} = \mathbf{z}$ as

$$P[\mathbf{Y} \in A \mid \mathbf{Z} = \mathbf{z}] = E(1[\mathbf{Y} \in A] \mid \mathbf{Z} = \mathbf{z}) = \Sigma_{\mathbf{y} \in A} p(\mathbf{y} \mid \mathbf{z}).$$

Then by taking $r(\mathbf{Y}) = 1[\mathbf{Y} \in A]$, $h = 1$ in Theorem B.1.1 we can express the (unconditional) probability that $\mathbf{Y} \in A$ as

$$P[\mathbf{Y} \in A] = E(E(r(\mathbf{Y}) \mid \mathbf{Z})) = \Sigma_{\mathbf{z}} P[\mathbf{Y} \in A \mid \mathbf{Z} = \mathbf{z}] p_{\mathbf{Z}}(\mathbf{z}) = E[P(Y \in A \mid \mathbf{Z})]. \tag{B.1.24}$$

For example, if Y and Z are as in (B.1.5),

$$P[Y \le y] = \Sigma_z \begin{pmatrix} N \\ z \end{pmatrix} \theta^z (1-\theta)^{n-z} H_z(y)$$

where H_z is the distribution function of a hypergeometric distribution with parameters (z, N, n).

B.1.4 Continuous Variables

Suppose now that $(\mathbf{Y}, \mathbf{Z})$ is a continuous random vector having coordinates that are themselves vectors and having density function $p(\mathbf{y}, \mathbf{z})$. We define, following the analogy between frequency and density functions, the *conditional density* function of $\mathbf{Y}$ given $\mathbf{Z} = \mathbf{z}$ by

$$p(\mathbf{y} \mid \mathbf{z}) = \frac{p(\mathbf{y}, \mathbf{z})}{p_{\mathbf{Z}}(\mathbf{z})} \tag{B.1.25}$$

if $p_{\mathbf{Z}}(\mathbf{z}) > 0$.

Because the marginal density of $\mathbf{Z}$, $p_{\mathbf{Z}}(\mathbf{z})$, is given by (A.8.12), it is clear that $p(\cdot \mid \mathbf{z})$ is a density. Because (B.1.25) does not differ formally from (B.1.1), equations (B.1.3) and (B.1.6) go over verbatim. Expression (B.1.4) becomes

$$p(\mathbf{y} \mid \mathbf{z}) = \frac{p_{\mathbf{Y}}(\mathbf{y}) q(\mathbf{z} \mid \mathbf{y})}{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p_{\mathbf{Y}}(\mathbf{t}) q(\mathbf{z} \mid \mathbf{t}) dt_1 \ldots dt_n}, \tag{B.1.26}$$

where q is the conditional density of $\mathbf{Z}$ given $\mathbf{Y} = \mathbf{y}$. This is also called *Bayes' Theorem*.

If $\mathbf{Y}$ and $\mathbf{Z}$ are independent, the conditional distributions equal the marginals as in the discrete case.

Example B.1.4 Let Y_1 and Y_2 be independent and uniformly, $\mathcal{U}(0,1)$, distributed. Let $Z = \min(Y_1, Y_2)$, $Y = \max(Y_1, Y_2)$. The joint distribution of Z and Y is given by

$$\begin{aligned} F(z,y) &= 2P[Y_1 < Y_2, Y_1 < z, Y_2 < y] \\ &= 2\int_0^y \int_0^{\min(y_2,z)} dy_1 dy_2 = 2\int_0^y \min(y_2, z) dy_2 \end{aligned} \tag{B.1.27}$$

if $0 \le z, y \le 1$.

The joint density is, therefore,

$$\begin{aligned} p(z,y) &= 2 \text{ if } 0 < z \le y < 1 \\ &= 0 \text{ otherwise.} \end{aligned} \tag{B.1.28}$$

The marginal density of Z is given by

$$\begin{aligned} p_Z(z) &= \int_z^1 2dy = 2(1-z),\ 0 < z < 1 \\ &= 0 \text{ otherwise.} \end{aligned} \tag{B.1.29}$$

We conclude that the conditional density of Y given $Z = z$ is uniform on the interval $(z, 1)$. □

If $E(|Y|) < \infty$, we denote the *conditional expectation of* Y *given* $\mathbf{Z} = \mathbf{z}$ in analogy to the discrete case as the expected value of a random variable with density $p(y \mid \mathbf{z})$. More generally, if $E(|r(\mathbf{Y})|) < \infty$, (A.10.11) shows that the conditional expectation of $r(\mathbf{Y})$ given $\mathbf{Z} = \mathbf{z}$ can be obtained from

$$E(r(\mathbf{Y}) \mid \mathbf{Z} = \mathbf{z}) = \int_{-\infty}^{\infty} r(\mathbf{y}) p(\mathbf{y} \mid \mathbf{z}) d\mathbf{y}. \tag{B.1.30}$$

As before, if $g(\mathbf{z}) = E(r(\mathbf{Y}) \mid \mathbf{Z} = \mathbf{z})$, we write $g(\mathbf{Z})$ as $E(r(\mathbf{Y}) \mid \mathbf{Z})$, the conditional expectation of $r(\mathbf{Y})$ given $\mathbf{Z}$. With this definition we can show that formulas 12, 13, 14, 19, 20, 23, and 24 of this section hold in the continuous case also. As an illustration, we next derive B.1.23:

Let $g(\mathbf{z}) = E[r(\mathbf{Y}) \mid \mathbf{Z}]$, then, by (A.10.11),

$$\begin{aligned} E(h(\mathbf{Z})E(r(\mathbf{Y}) \mid \mathbf{Z})) = E(h(\mathbf{Z})g(\mathbf{Z})) &= \int_{-\infty}^{\infty} h(\mathbf{z}) g(\mathbf{z}) p_Z(\mathbf{z}) d\mathbf{z} \\ = \int_{-\infty}^{\infty} h(\mathbf{z}) p_Z(\mathbf{z}) \left[\int_{-\infty}^{\infty} r(\mathbf{y}) p(\mathbf{y} \mid \mathbf{z}) d\mathbf{y} \right] d\mathbf{z}. \end{aligned} \tag{B.1.31}$$

By a standard theorem on double integrals, we conclude that the right-hand side of (B.1.31) equals

$$\begin{aligned}&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} r(\mathbf{y})h(\mathbf{z})p_Z(\mathbf{z})p(\mathbf{y} \mid \mathbf{z})d\mathbf{y}d\mathbf{z}\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} r(\mathbf{y})h(\mathbf{z})p(\mathbf{y},\mathbf{z})d\mathbf{y}d\mathbf{z} = E(r(\mathbf{Y})h(\mathbf{Z}))\end{aligned} \tag{B.1.32}$$

by (A.10.11), and we have established B.1.23.

To illustrate these formulae, we calculate $E(Y \mid Z)$ in Example B.1.4. Here,

$$E(Y \mid Z = z) = \int_0^1 yp(y \mid z)dy = \frac{1}{(1-z)}\int_z^1 ydy = \frac{1+z}{2},\ 0 < z < 1,$$

and, hence,

$$E(Y \mid Z) = \frac{1+Z}{2}.$$

B.1.5 Comments on the General Case

Clearly the cases $(\mathbf{Y}, \mathbf{Z})$ discrete and $(\mathbf{Y}, \mathbf{Z})$ continuous do not cover the field. For example, if Y is uniform on $(0, 1)$ and $Z = Y^2$, then (Y, Z) neither has a joint frequency function nor a joint density. (The density would have to concentrate on $z = y^2$, but then it cannot satisfy $\int_0^1 \int_0^1 f(y, z)dydz = 1$.) Thus, (Y, Z) is neither discrete nor continuous in our sense. On the other hand, we should have a concept of conditional probability for which $P[Y = u \mid Z = \sqrt{u}] = 1$. To cover the general theory of conditioning is beyond the scope of this book. The interested student should refer to the books by Breiman (1968), Loève (1977), Chung (1974), or Billingsley (1995). We merely note that it is possible to define $E(\mathbf{Y} \mid \mathbf{Z} = \mathbf{z})$ and $E(\mathbf{Y} \mid \mathbf{Z})$ in such a way that they coincide with (B.1.7) and (B.1.30) in the discrete and continuous cases and moreover so that equations 15, 16, 20, and 23 of this section hold.

As an illustration, suppose that in Example B.1.4 we want to find the conditional expectation of $\sin(ZY)$ given $Z = z$. By our discussion we can calculate $E(\sin(ZY) \mid Z = z)$ as follows: First, apply (B.1.16) to get

$$E(\sin(ZY) \mid Z = z) = E(\sin(zY) \mid Z = z).$$

Because, given $Z = z$, Y has a $\mathcal{U}(z, 1)$ distribution, we can complete the computation by applying (A.10.11) to get

$$E(\sin(zY) \mid Z = z) = \frac{1}{(1-z)}\int_z^1 \sin(zy)dy = \frac{1}{z(1-z)}[\cos z^2 - \cos z].$$

B.2 DISTRIBUTION THEORY FOR TRANSFORMATIONS OF RANDOM VECTORS

B.2.1 The Basic Framework

In statistics we will need the distributions of functions of the random variables appearing in an experiment. Examples of such functions are sums, averages, differences, sums of squares, and so on. In this section we will develop a result that often is useful in finding the joint distribution of several functions of a continuous random vector. The result will generalize (A.8.9), which gives the density of a real-valued function of a continuous random variable.

Let $\mathbf{h} = (h_1, \ldots, h_k)^T$, where each h_i is a real-valued function on R^k. Thus, $\mathbf{h}$ is a transformation from R^k to R^k. Recall that the *Jacobian* $J_{\mathbf{h}}(\mathbf{t})$ of $\mathbf{h}$ evaluated at $\mathbf{t} = (t_1, \ldots, t_k)^T$ is by definition the determinant

$$J_{\mathbf{h}}(\mathbf{t}) = \begin{vmatrix} \dfrac{\partial}{\partial t_1} & h_1(\mathbf{t}) & \ldots & \dfrac{\partial}{\partial t_1} & h_k(\mathbf{t}) \\ \vdots & & & \vdots & \\ \dfrac{\partial}{\partial t_k} & h_1(\mathbf{t}) & \ldots & \dfrac{\partial}{\partial t_k} & h_k(\mathbf{t}) \end{vmatrix}.$$

The principal result of this section, Theorem B.2.2, rests on the change of variable theorem for multiple integrals from calculus. We now state this theorem without proof (see Apostol, 1974, p. 421).

Theorem B.2.1 *Let* $\mathbf{h} = (h_1, \ldots, h_k)^T$ *be a transformation defined on an open subset* B *of* R^k. *Suppose that:*(1)

(i) $\mathbf{h}$ *has continuous first partial derivatives in* B.

(ii) $\mathbf{h}$ *is one-to-one on* B.

(iii) *The Jacobian of* $\mathbf{h}$ *does not vanish on* B.

Let f *be a real-valued function (defined and measurable) on the range* $\mathbf{h}(B) = \{(h_1(\mathbf{t}), \ldots, h_k(\mathbf{t})) : \mathbf{t} \in B\}$ *of* $\mathbf{h}$ *and suppose* f *satisfies*

$$\int_{\mathbf{h}(B)} |f(\mathbf{x})| d\mathbf{x} < \infty.$$

Then for every (measurable) subset K *of* $\mathbf{h}(B)$ *we have*

$$\int_K f(\mathbf{x}) d\mathbf{x} = \int_{\mathbf{h}^{-1}(K)} f(\mathbf{h}(\mathbf{t})) |J_{\mathbf{h}}(\mathbf{t})| d\mathbf{t}. \tag{B.2.1}$$

In these expressions we write $d\mathbf{x}$ for $dx_1 \dots dx_k$. Moreover, $\mathbf{h}^{-1}$ denotes the inverse of the transformation $\mathbf{h}$; that is, $\mathbf{h}^{-1}(\mathbf{x}) = \mathbf{t}$ if, and only if, $\mathbf{x} = \mathbf{h}(\mathbf{t})$. We also need another result from the calculus (see Apostol, 1974, p. 417),

$$J_{\mathbf{h}^{-1}}(\mathbf{t}) = \frac{1}{J_{\mathbf{h}}(\mathbf{h}^{-1}(\mathbf{t}))}. \tag{B.2.2}$$

It follows that a transformation $\mathbf{h}$ satisfies the conditions of Theorem B.2.1 if, and only if, $\mathbf{h}^{-1}$ does.

We can now derive the density of $\mathbf{Y} = \mathbf{g}(\mathbf{X}) = (g_1(\mathbf{X}), \dots, g_k(\mathbf{X}))^T$ when $\mathbf{g}$ satisfies the conditions of Theorem B.2.1 and $\mathbf{X} = (X_1, \dots, X_k)^T$ is a continuous random vector.

Theorem B.2.2 *Let $\mathbf{X}$ be continuous and let S be an open subset of R^k such that $P(\mathbf{X} \in S) = 1$. If $\mathbf{g} = (g_1, \dots, g_k)^T$ is a transformation from S to R^k such that $\mathbf{g}$ and S satisfy the conditions of Theorem B.2.1, then the density of $\mathbf{Y} = \mathbf{g}(\mathbf{X})$ is given by*

$$p_{\mathbf{Y}}(\mathbf{y}) = p_{\mathbf{X}}(\mathbf{g}^{-1}(\mathbf{y}))|J_{\mathbf{g}^{-1}}(\mathbf{y})| \tag{B.2.3}$$

for $\mathbf{y} \in \mathbf{g}(S)$.

Proof. The distribution function of $\mathbf{Y}$ is (see (A.7.8))

$$F_{\mathbf{Y}}(\mathbf{y}) = \int \cdots \int_{A_k} p_{\mathbf{X}}(x_1, \dots, x_k) dx_1 \dots dx_k$$

where $A_k = \{\mathbf{x} \in R^k : g_i(\mathbf{x}) \le y_i,\ i = 1, \dots, k\}$. Next we apply Theorem B.2.1 with $\mathbf{h} = \mathbf{g}^{-1}$ and $f = p_{\mathbf{x}}$. Because $\mathbf{h}^{-1}(A_k) = \mathbf{g}(A_k) = \{\mathbf{g}(\mathbf{x}) : g_i(\mathbf{x}) \le y_i,\ i = 1, \dots, k\} = \{\mathbf{t} : \mathbf{t} \le y_i,\ i = 1, \dots, k\}$, we obtain

$$F_{\mathbf{Y}}(\mathbf{y}) = \int_{-\infty}^{y_k} \cdots \int_{-\infty}^{y_1} p_{\mathbf{X}}(\mathbf{g}^{-1}(\mathbf{t}))|J_{\mathbf{g}^{-1}}(\mathbf{t})| dt_1 \dots dt_k.$$

The result now follows if we recall from Section A.7 that whenever $F_{\mathbf{Y}}(\mathbf{y}) = \int_{-\infty}^{y_k} \cdots \int_{-\infty}^{y_1} q(t_1, \dots, t_k) dt_1 \dots dt_k$ for some nonnegative function q, then q must be the density of $\mathbf{Y}$. □

Example B.2.1 Suppose $\mathbf{X} = (X_1, X_2)^T$ where X_1 and X_2 are independent with $\mathcal{N}(0, 1)$ and $\mathcal{N}(0, 4)$ distributions, respectively. What is the joint distribution of $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$? Here (see (A.13.17)),

$$p_{\mathbf{X}}(x_1, x_2) = \frac{1}{4\pi} \exp -\frac{1}{2}\left[x_1^2 + \frac{1}{4}x_2^2\right].$$

In this case, $S = R^2$. Also note that $g_1(\mathbf{x}) = x_1 + x_2$, $g_2(\mathbf{x}) = x_1 - x_2$, $g_1^{-1}(\mathbf{y}) = \frac{1}{2}(y_1 + y_2)$, $g_2^{-1}(\mathbf{y}) = \frac{1}{2}(y_1 - y_2)$, that the range $\mathbf{g}(S)$ is R^2 and that

$$J_{\mathbf{g}^{-1}}(\mathbf{y}) = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = -\frac{1}{2}.$$

Upon substituting these quantities in (B.2.3), we obtain

$$\begin{aligned} p_{\mathbf{Y}}(y_1, y_2) &= \frac{1}{2} p_{\mathbf{x}}\left(\frac{1}{2}(y_1 + y_2), \frac{1}{2}(y_1 - y_2)\right) \\ &= \frac{1}{8\pi} \exp -\frac{1}{2}\left[\frac{1}{4}(y_1 + y_2)^2 + \frac{1}{16}(y_1 - y_2)^2\right] \\ &= \frac{1}{8\pi} \exp -\frac{1}{32}[5y_1^2 + 5y_2^2 + 6y_1 y_2]. \end{aligned}$$

This is an example of bivariate normal density. Such densities will be considered further in Section B.4. □

Upon combining (B.2.2) and (B.2.3) we see that for $\mathbf{y} \in \mathbf{g}(S)$,

$$p_{\mathbf{Y}}(\mathbf{y}) = \frac{p_{\mathbf{X}}(\mathbf{g}^{-1}(\mathbf{y}))}{|J_{\mathbf{g}}(\mathbf{g}^{-1}(\mathbf{y}))|}. \tag{B.2.4}$$

If X is a random variable ($k = 1$), the Jacobian of g is just its derivative and the requirements (i) and (iii) that g' be continuous and nonvanishing imply that g is strictly monotone and, hence, satisfies (ii). In this case (B.2.4) reduces to the familiar formula (A.8.9).

It is possible to give useful generalizations of Theorem B.2.2 to situations where $\mathbf{g}$ is not one-to-one (Problem B.2.7).

Theorem B.2.2 provides one of the instances in which frequency and density functions differ. If $\mathbf{X}$ is discrete, $\mathbf{g}$ is one-to-one, and $\mathbf{Y} = \mathbf{g}(\mathbf{X})$, then $p_{\mathbf{Y}}(\mathbf{y}) = p_{\mathbf{x}}(\mathbf{g}^{-1}(\mathbf{y}))$. The extra factor in the continuous case appears roughly as follows. If $A(\mathbf{y})$ is a "small" cube surrounding $\mathbf{y}$ and we let $V(B)$ denote the volume of a set B, then

$$\begin{aligned} p_{\mathbf{Y}}(\mathbf{y}) &\approx \frac{P[\mathbf{g}(\mathbf{X}) \in A(\mathbf{y})]}{V(A(\mathbf{y}))} = \frac{P[\mathbf{X} \in \mathbf{g}^{-1}(A(\mathbf{y}))]}{V(\mathbf{g}^{-1}(A(\mathbf{y})))} \cdot \frac{V(\mathbf{g}^{-1}(A(\mathbf{y})))}{V(A(\mathbf{y}))} \\ &\approx p_{\mathbf{x}}(\mathbf{g}^{-1}(\mathbf{y})) \cdot \frac{V(\mathbf{g}^{-1}(A(\mathbf{y})))}{V(A(\mathbf{y}))}. \end{aligned}$$

Using the fact that $\mathbf{g}^{-1}$ is approximately linear on $A(\mathbf{y})$, it is not hard to show that

$$\frac{V(g^{-1}(A(\mathbf{y})))}{V(A(\mathbf{y}))} \approx |J_{\mathbf{g}^{-1}}(\mathbf{y})|.$$

The justification of these approximations is the content of Theorem B.2.2.

The following generalization of (A.8.10) is very important. For a review of the elementary properties of matrices needed in its formulation, we refer the reader to Section B.10.

Recall that $\mathbf{g}$ is called an *affine transformation* of R^k if there exists a $k \times k$ matrix $\mathbf{A}$ and a $k \times 1$ vector $\mathbf{c}$ such that $\mathbf{g}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{c}$. If $\mathbf{c} = \mathbf{0}$, $\mathbf{g}$ is called a *linear transformation*. The function $\mathbf{g}$ is one-to-one if, and only if, $\mathbf{A}$ is nonsingular and then

$$\mathbf{g}^{-1}(\mathbf{y}) = \mathbf{A}^{-1}(\mathbf{y} - \mathbf{c}), \tag{B.2.5}$$

$\mathbf{y} \in R^k$, where $\mathbf{A}^{-1}$ is the inverse of $\mathbf{A}$.

Corollary B.2.1 *Suppose* $\mathbf{X}$ *is continuous and* S *is such that* $P(\mathbf{X} \in S) = 1$. *If* $\mathbf{g}$ *is a one-to-one affine transformation as defined earlier, then* $\mathbf{Y} = \mathbf{g}(\mathbf{X})$ *has density*

$$p_{\mathbf{Y}}(\mathbf{y}) = |\det \mathbf{A}|^{-1} p_{\mathbf{X}}(\mathbf{A}^{-1}(\mathbf{y} - \mathbf{c})) \tag{B.2.6}$$

for $\mathbf{y} \in \mathbf{g}(S)$, *where* det $\mathbf{A}$ *is the determinant of* $\mathbf{A}$.

The corollary follows from (B.2.4), (B.2.5), and the relation,

$$J_{\mathbf{g}}(\mathbf{g}^{-1}(\mathbf{y})) \equiv \det \mathbf{A}. \tag{B.2.7}$$

Example B.2.1 is a special case of the corollary. Further applications appear in the next section. □

B.2.2 The Gamma and Beta Distributions

As a consequence of the transformation theorem we obtain basic properties of two important families of distributions, which will also figure in the next section. The first family has densities given by

$$g_{p,\lambda}(x) = \frac{\lambda^p x^{p-1} e^{-\lambda x}}{\Gamma(p)} \tag{B.2.8}$$

for $x > 0$, where the parameters p and λ are taken to be positive and $\Gamma(p)$ denotes the *Euler gamma function* defined by

$$\Gamma(p) = \int_0^\infty t^{p-1} e^{-t} dt. \tag{B.2.9}$$

A useful fact is that $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$. It follows by integration by parts that, for all $p > 0$,

$$\Gamma(p+1) = p\Gamma(p) \text{ and that } \Gamma(k) = (k-1)! \text{ for positive integers } k. \tag{B.2.10}$$

The family of distributions with densities given by (B.2.8) is referred to as the *gamma* family of distributions and we shall write $\Gamma(p, \lambda)$ for the distribution corresponding to $g_{p,\lambda}$. The special case $p = 1$ corresponds to the familiar exponential distribution $\mathcal{E}(\lambda)$ of (A.13.24). By (A.8.10), X is distributed $\Gamma(p, \lambda)$ if, and only if, λX is distributed $\Gamma(p, 1)$. Thus, $1/\lambda$ is a scale parameter for the $\Gamma(p, \lambda)$ family.

Let k be a positive integer. In statistics, the gamma density $g_{p,\lambda}$ with $p = \frac{1}{2}k$ and $\lambda = \frac{1}{2}$ is referred to as the *chi squared density with* k *degrees of freedom* and is denoted by χ^2_k.

The other family of distributions we wish to consider is the *beta* family, which is indexed by the positive parameters r and s. Its densities are given by

$$b_{r,s}(x) = \frac{x^{r-1}(1-x)^{s-1}}{B(r,s)} \tag{B.2.11}$$

for $0 < x < 1$, where $B(r, s) = [\Gamma(r)\Gamma(s)]/[\Gamma(r+s)]$ is the *beta function*. The distribution corresponding to $b_{r,s}$ will be written $\beta(r, s)$. Figures B.2.1 and B.2.2 show some typical members of the two families.

Theorem B.2.3 *If X_1 and X_2 are independent random variables with $\Gamma(p, \lambda)$ and $\Gamma(q, \lambda)$ distributions, respectively, then $Y_1 = X_1 + X_2$ and $Y_2 = X_1/(X_1 + X_2)$ are independent and have, respectively, $\Gamma(p + q, \lambda)$ and $\beta(p, q)$ distributions.*

Proof. If $\lambda = 1$, the joint density of X_1 and X_2 is

$$p(x_1, x_2) = [\Gamma(p)\Gamma(q)]^{-1} e^{-(x_1+x_2)} x_1^{p-1} x_2^{q-1} \tag{B.2.12}$$

for $x_1 > 0$, $x_2 > 0$. Let

$$(y_1, y_2)^T = \mathbf{g}(x_1, x_2) = \left(x_1 + x_2, \frac{x_1}{x_1 + x_2} \right)^T.$$

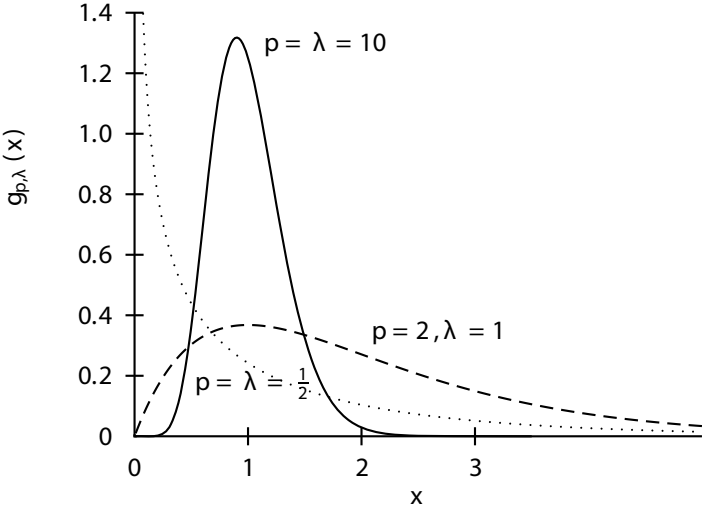

Figure B.2.1 The gamma density, $g_{p,\lambda}(x)$, for selected p, λ.

Then $\mathbf{g}$ is one-to-one on $S = \{(x_1, x_2)^T : x_1 > 0,\ x_2 > 0\}$ and its range is $S_1 = \{(y_1, y_2)^T : y_1 > 0,\ 0 < y_2 < 1\}$. We note that on S_1

$$\mathbf{g}^{-1}(y_1, y_2) = (y_1 y_2, y_1 - y_1 y_2)^T. \tag{B.2.13}$$

Therefore,

$$J_{\mathbf{g}^{-1}}(y_1, y_2) = \begin{vmatrix} y_2 & 1 - y_2 \\ y_1 & -y_1 \end{vmatrix} = -y_1. \tag{B.2.14}$$

If we now substitute (B.2.13) and (B.2.14) in (B.2.4) we get for the density of $(Y_1, Y_2)^T = \mathbf{g}(X_1, X_2)$,

$$p_{\mathbf{Y}}(y_1, y_2) = \frac{e^{-y_1}(y_1 y_2)^{p-1}(y_1 - y_1 y_2)^{q-1} y_1}{\Gamma(p)\Gamma(q)} \tag{B.2.15}$$

for $y_1 > 0$, $0 < y_2 < 1$. Simplifying (B.2.15) leads to

$$p_{\mathbf{Y}}(y_1, y_2) = g_{p+q,1}(y_1) b_{p,q}(y_2). \tag{B.2.16}$$

The result is proved for $\lambda = 1$. If $\lambda \neq 1$ define $X_1' = \lambda X_1$ and $X_2' = \lambda X_2$. Now X_1' and X_2' are independent $\Gamma(p, 1)$, $\Gamma(q, 1)$ variables respectively. Because $X_1' + X_2' = \lambda(X_1 + X_2)$ and $X_1'(X_1' + X_2')^{-1} = X_1(X_1 + X_2)^{-1}$ the theorem follows. □

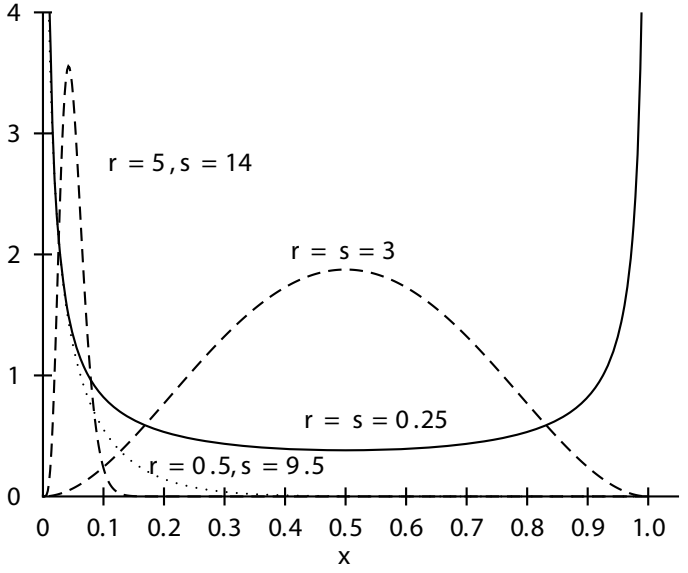

Figure B.2.2 The beta density, $b_{r,s}(x)$, for selected r, s.

By iterating the argument of Theorem B.2.3, we obtain the following general result.

Corollary B.2.2 *If $X_1, \ldots, X_n$ are independent random variables such that X_i has a $\Gamma(p_i, \lambda)$ distribution, $i = 1, \ldots, n$, then $\Sigma_{i=1}^n X_i$ has a $\Gamma(\Sigma_{i=1}^n p_i, \lambda)$ distribution.*

Some other properties of the gamma and beta families are given in the problems and in the next section.

B.3 DISTRIBUTION THEORY FOR SAMPLES FROM A NORMAL POPULATION

In this section we introduce some distributions that appear throughout modern statistics. We derive their densities as an illustration of the theory of Section B.2. However, these distributions should be remembered in terms of their definitions and qualitative properties rather than density formulas.

B.3.1 The χ^2, F, and t Distributions

Throughout this section we shall suppose that $\mathbf{X} = (X_1, \ldots, X_n)^T$ where the X_i form a sample from a $\mathcal{N}(0, \sigma^2)$ population. Some results for normal populations, whose mean differs from 0, are given in the problems. We begin by investigating the distribution of the $\Sigma_{i=1}^n X_i^2$, the squared distance of $\mathbf{X}$ from the origin.

Theorem B.3.1 *The random variable $V = \Sigma_{i=1}^n X_i^2/\sigma^2$ has a χ_n^2 distribution. That is, V has density*

$$p_V(v) = \frac{v^{\frac{1}{2}(n-2)} e^{-\frac{1}{2}v}}{2^{n/2}\Gamma(n/2)} \tag{B.3.1}$$

for $v > 0$.

Proof. Let $Z_i = X_i/\sigma$, $i = 1, \ldots, n$. Then $Z_i \sim \mathcal{N}(0, 1)$. Because the Z_i^2 are independent, it is enough to prove the theorem for $n = 1$ and then apply Corollary B.2.2. If $T = Z_1^2$, then the distribution function of T is

$$P[Z_1^2 \leq t] = P[-\sqrt{t} \leq Z_1 \leq \sqrt{t}] \tag{B.3.2}$$

and, thus,

$$F_T(t) = \Phi(\sqrt{t}) - \Phi(-\sqrt{t}). \tag{B.3.3}$$

Differentiating both sides we get the density of T

$$p_T(t) = t^{-\frac{1}{2}}\varphi(\sqrt{t}) = \frac{1}{\sqrt{2\pi}} t^{-\frac{1}{2}} e^{-t/2} \tag{B.3.4}$$

for $t > 0$, which agrees with $g_{\frac{1}{2},\frac{1}{2}}$ up to a multiplicative constant. Because the constant is determined by the requirement that p_T and $g_{\frac{1}{2},\frac{1}{2}}$ are densities, we must have $p_T = g_{\frac{1}{2},\frac{1}{2}}$ and the result follows. □

Let V and W be independent and have χ_k^2 and χ_m^2 distributions, respectively, and let $S = (V/k)/(W/m)$. The distribution of S is called the *F distribution with k and m degrees of freedom*. We shall denote it by $\mathcal{F}_{k,m}$.

Next, we introduce the *t distribution with k degrees of freedom*, which we shall denote by $\mathcal{T}_k$. By definition $\mathcal{T}_k$ is the distribution of $Q = Z/\sqrt{V/k}$, where Z and V are independent with $\mathcal{N}(0, 1)$ and χ_k^2 distributions, respectively. We can now state the following elementary consequence of Theorem B.3.1.

Corollary B.3.1 *The random variable* $(m/k)\Sigma_{i=1}^{k}X_i^2/\Sigma_{i=k+1}^{k+m}X_i^2$ *has an* $\mathcal{F}_{k,m}$ *distribution. The random variable* $X_1/\sqrt{(1/k)\Sigma_{i=2}^{k+1}X_i^2}$ *has a* $\mathcal{T}_k$ *distribution.*

Proof. For the first assertion we need only note that

$$\sum_{i=1}^{k}X_i^2\Bigg/\sum_{i=k+1}^{k+m}X_i^2=\frac{1}{\sigma^2}\sum_{i=1}^{k}X_i^2\Bigg/\frac{1}{\sigma^2}\sum_{i=k+1}^{k+m}X_i^2 \tag{B.3.5}$$

and apply the theorem and the definition of $\mathcal{F}_{k,m}$. The second assertion follows in the same way. □

To make the definitions of the $\mathcal{F}_{k,m}$ and $\mathcal{T}_k$ distributions useful for computation, we need their densities. We assume the S, Q, V, W are as in the definitions of these distributions.

To derive the density of S note that, if $U=V/(V+W)$, then

$$S=\frac{V/k}{W/m}=\frac{m}{k}\frac{U}{1-U}. \tag{B.3.6}$$

Because $V\sim\Gamma\left(\frac{1}{2}k,\frac{1}{2}\right)$, $W\sim\Gamma\left(\frac{1}{2}m,\frac{1}{2}\right)$ and V and W are independent, then by Theorem B.2.3, U has a beta distribution with parameters $\frac{1}{2}k$ and $\frac{1}{2}m$. To obtain the density of S we need only apply the change of variable formula (A.8.9) to U with $g(u)=(m/k)u/(1-u)$. After some calculation we arrive at the $\mathcal{F}_{k,m}$ density (see Figure B.3.1)

$$p_S(s)=\frac{(k/m)^{\frac{1}{2}k}s^{\frac{1}{2}(k-2)}(1+(k/m)s)^{-\frac{1}{2}(k+m)}}{B\left(\frac{1}{2}k,\frac{1}{2}m\right)} \tag{B.3.7}$$

for $s>0$.

To get the density of Q we argue as follows. Because $-Z$ has the same distribution as Z, we may conclude that Q and $-Q$ are identically distributed. It follows that

$$\begin{aligned}P[0<Q<q] &= P[0<-Q<q]\\ &= P[-q<Q<0]=\frac{1}{2}P[0<Q^2<q^2].\end{aligned} \tag{B.3.8}$$

Differentiating $P[0<Q<q]$, $P[-q<Q<0]$ and $\frac{1}{2}P[0<Q^2<q^2]$ we get

$$p_Q(q)=p_Q(-q)=qp_{Q^2}(q^2)\text{ if }q>0. \tag{B.3.9}$$

Now Q^2 has by Corollary B.3.1 an $\mathcal{F}_{1,k}$ distribution. We can, therefore, use (B.3.7) and (B.3.9) to conclude

$$p_Q(q)=\frac{\Gamma\left(\frac{1}{2}(k+1)\right)(1+(q^2/k))^{-\frac{1}{2}(k+1)}}{\sqrt{\pi k}\Gamma\left(\frac{1}{2}k\right)} \tag{B.3.10}$$

for $-\infty<q<\infty$.

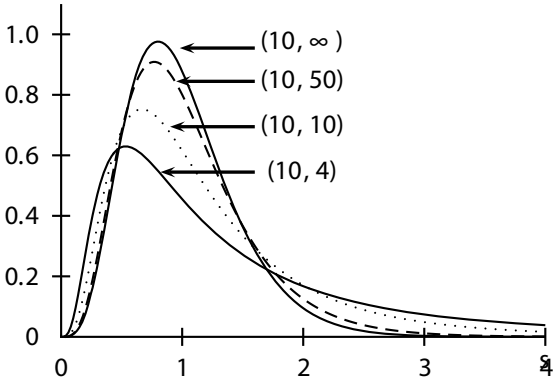

Figure B.3.1 The $\mathcal{F}_{k,m}$ density for selected (k, m).

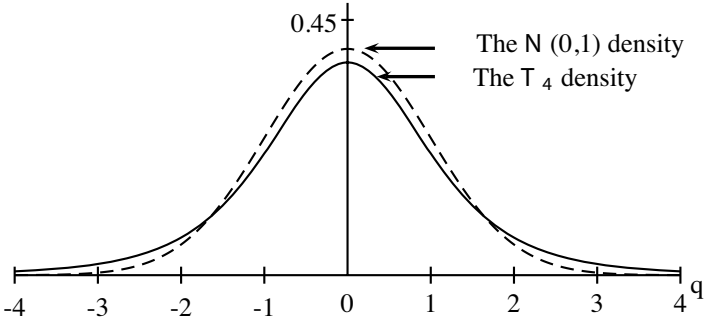

Figure B.3.2 The Student t and standard normal densities.

The χ^2, $\mathcal{T}$, and $\mathcal{F}$ cumulative distribution functions are given in Tables II, III, and IV, respectively. More precisely, these tables give the inverses or quantiles of these distributions. For $\alpha \in (0, 1)$, an αth *quantile* or 100 αth *percentile* of the continuous distribution F is by definition any number $x(\alpha)$ such that $F(x(\alpha)) = \alpha$.

Continuity of F guarantees the existence of $x(\alpha)$ for all α. If F is strictly increasing, $x(\alpha)$ is unique for each α. As an illustration, we read from Table III that the (0.95)th quantile or 95th percentile of the $\mathcal{T}_{20}$ distribution is $t(0.95) = 1.725$.

B.3.2 Orthogonal Transformations

We turn now to orthogonal transformations of normal samples. Let us begin by recalling some classical facts and definitions involving matrices, which may be found in standard texts, for example, Birkhoff and MacLane (1965).

Suppose that $\mathbf{A}$ is a $k \times m$ matrix with entry a_{ij} in the ith row and jth column, $i = 1, \ldots, k$; $j = 1, \ldots, m$. Then the *transpose* of $\mathbf{A}$, written $\mathbf{A}^T$, is the $m \times k$ matrix, whose entry in the ith row and jth column is a_{ji}. Thus, the transpose of a row vector is a column vector and the transpose of a square matrix is a square matrix.

An $n \times n$ matrix $\mathbf{A}$ is said to be *orthogonal* if, and only if,

$$\mathbf{A}^T = \mathbf{A}^{-1} \tag{B.3.11}$$

or equivalently if, and only if, either one of the following two matrix equations is satisfied,

$$\mathbf{A}^T\mathbf{A} = \mathbf{I} \tag{B.3.12}$$

$$\mathbf{A}\mathbf{A}^T = \mathbf{I} \tag{B.3.13}$$

where $\mathbf{I}$ is the $n \times n$ identity matrix. Equation (B.3.12) requires that the column vectors of $\mathbf{A}$ be of length 1 and mutually perpendicular, whereas (B.3.13) imposes the same requirement on the rows. Clearly, (B.3.12) and (B.3.13) are equivalent. Considered as transformations on R^n orthogonal matrices are rigid motions, which preserve the distance between points. That is, if $\mathbf{a} = (a_1, \ldots, a_n)^T$, $\mathbf{b} = (b_1, \ldots, b_n)^T$, $|\mathbf{a} - \mathbf{b}| = \sqrt{\Sigma_{i=1}^n (a_i - b_i)^2}$ is the Euclidean distance between $\mathbf{a}$ and $\mathbf{b}$, and $\mathbf{A}$ is orthogonal, then

$$|\mathbf{a} - \mathbf{b}| = |\mathbf{A}\mathbf{a} - \mathbf{A}\mathbf{b}|. \tag{B.3.14}$$

To see this, note that

$$\begin{aligned} |\mathbf{a} - \mathbf{b}|^2 &= (\mathbf{a} - \mathbf{b})^T(\mathbf{a} - \mathbf{b}) = (\mathbf{a} - \mathbf{b})^T\mathbf{A}^T\mathbf{A}(\mathbf{a} - \mathbf{b}) \\ &= [\mathbf{A}(\mathbf{a} - \mathbf{b})]^T[\mathbf{A}(\mathbf{a} - \mathbf{b})] = |\mathbf{A}\mathbf{a} - \mathbf{A}\mathbf{b}|^2. \end{aligned} \tag{B.3.15}$$

Finally, we shall use the fact that if $\mathbf{A}$ is orthogonal,

$$|\det \mathbf{A}| = 1. \tag{B.3.16}$$

This follows from[(1)]

$$[\det \mathbf{A}]^2 = [\det \mathbf{A}][\det \mathbf{A}^T] = \det[\mathbf{A}\mathbf{A}^T] = \det \mathbf{I} = 1. \tag{B.3.17}$$

Because $\mathbf{X} = (X_1, \ldots, X_n)^T$ is a vector of independent identically distributed $\mathcal{N}(0, \sigma^2)$ random variables we can write the density of $\mathbf{X}$ as

$$\begin{aligned} p_{\mathbf{X}}(x_1, \ldots, x_n) &= \frac{1}{[\sqrt{2\pi}\sigma]^n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2\right] \\ &= \frac{1}{[\sqrt{2\pi}\sigma]^n} \exp\left[-\frac{1}{2\sigma^2}|\mathbf{x}|^2\right]. \end{aligned} \tag{B.3.18}$$

We seek the density of $\mathbf{Y} = \mathbf{g}(\mathbf{X}) = \mathbf{A}\mathbf{X} + \mathbf{c}$, where $\mathbf{A}$ is orthogonal, $n \times n$, and $\mathbf{c} = (c_1, \ldots, c_n)$. By Corollary B.2.1, $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ has density

$$\begin{aligned} p_{\mathbf{Y}}(\mathbf{y}) &= \frac{1}{|\det \mathbf{A}|} p_{\mathbf{X}}(\mathbf{A}^{-1}(\mathbf{y} - \mathbf{c})) \\ &= p_{\mathbf{X}}(\mathbf{A}^T(\mathbf{y} - \mathbf{c})) \end{aligned} \tag{B.3.19}$$

by (B.3.11) and (B.3.16). If we substitute $\mathbf{A}^T(\mathbf{y} - \mathbf{c})$ for $\mathbf{x}$ in (B.3.18) and apply (B.3.15), we get

$$p_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{[\sqrt{2\pi}\sigma]^n} \exp\left[-\frac{1}{2\sigma^2}|\mathbf{y} - \mathbf{c}|^2\right] = p_{\mathbf{X}}(\mathbf{y} - \mathbf{c}). \tag{B.3.20}$$

Because

$$p_{\mathbf{X}}(\mathbf{y} - \mathbf{c}) = \prod_{i=1}^{n} \left\{ \frac{1}{\sigma} \varphi\left(\frac{y_i - c_i}{\sigma}\right) \right\} \tag{B.3.21}$$

we see that the Y_i are independent normal random variables with $E(Y_i) = c_i$ and common variance σ^2. If, in particular, $\mathbf{c} = \mathbf{0}$, then $Y_1, \ldots, Y_n$ are again a sample from a $\mathcal{N}(0, \sigma^2)$ population.

More generally it follows that if $\mathbf{Z} = \mathbf{X} + \mathbf{d}$, then $\mathbf{Y} = \mathbf{g}(\mathbf{Z}) = \mathbf{A}(\mathbf{X} + \mathbf{d}) + \mathbf{c} = \mathbf{A}\mathbf{X} + (\mathbf{A}\mathbf{d} + \mathbf{c})$ has density

$$p_{\mathbf{Y}}(\mathbf{y}) = p_{\mathbf{X}}(\mathbf{y} - (\mathbf{A}\mathbf{d} + \mathbf{c})). \tag{B.3.22}$$

Because $\mathbf{d} = (d_1, \ldots, d_n)^T$ is arbitrary and, by definition, $E(Z_i) = E(X_i + d_i) = d_i$, $i = 1, \ldots, n$, we see that we have proved the following theorem.

Theorem B.3.2 *If $\mathbf{Z} = (Z_1, \ldots, Z_n)^T$ has independent normally distributed components with the same variance σ^2 and $\mathbf{g}$ is an affine transformation defined by the orthogonal matrix $\mathbf{A}$ and vector $\mathbf{c} = (c_1, \ldots, c_n)^T$, then $\mathbf{Y} = \mathbf{g}(\mathbf{Z}) = (Y_1, \ldots, Y_n)^T$ has independent normally distributed components with variance σ^2. Furthermore, if $\mathbf{A} = (a_{ij})$*

$$E(Y_i) = c_i + \sum_{j=1}^{n} a_{ij} E(Z_j) \tag{B.3.23}$$

for $i = 1, \ldots, n$.

This fundamental result will be used repeatedly in the sequel. As an application, we shall derive another classical property of samples from a normal population.

Theorem B.3.3 *Let $(Z_1, \ldots, Z_n)^T$ be a sample from a $\mathcal{N}(\mu, \sigma^2)$ population. Define*

$$\bar{Z} = \frac{1}{n} \sum_{i=1}^{n} Z_i. \tag{B.3.24}$$

Then $\bar{Z}$ and $\Sigma_{k=1}^{n}(Z_i - \bar{Z})^2$ are independent. Furthermore, $\bar{Z}$ has a $\mathcal{N}(\mu, \sigma^2/n)$ distribution while $(1/\sigma^2)\Sigma_{i=1}^{n}(Z_i - \bar{Z})^2$ is distributed as χ^2_{n-1}.

Proof. Construct an orthogonal matrix $\mathbf{A} = (a_{ij})$ whose first row is

$$\mathbf{a}_1 = \left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right).$$

This is equivalent to finding one of the many orthogonal bases in R^n whose first member is $\mathbf{a}_1$ and may, for instance, be done by the Gram–Schmidt process (Birkhoff and MacLane, 1965, p. 180). An example of such an $\mathbf{A}$ is given in Problem B.3.15. Let $\mathbf{AZ} = (Y_1, \ldots, Y_n)^T$. By Theorem B.3.2, the Y_i are independent and normally distributed with variance σ^2 and means

$$E(Y_i) = \sum_{j=1}^{n} a_{ij} E(Z_j) = \mu \sum_{j=1}^{n} a_{ij}. \tag{B.3.25}$$

Because $a_{1j} = 1/\sqrt{n}$, $1 \le j \le n$, and $\mathbf{A}$ is orthogonal we see that

$$\sum_{j=1}^{n} a_{kj} = \sqrt{n} \sum_{j=1}^{n} a_{1j} a_{kj} = 0,\ k = 2, \ldots, n. \tag{B.3.26}$$

Therefore,

$$E(Y_1) = \mu\sqrt{n},\ E(Y_k) = 0,\ k = 2, \ldots, n. \tag{B.3.27}$$

By Theorem B.3.1, $(1/\sigma^2)\Sigma_{k=2}^{n} Y_k^2$ has a χ_{n-1}^2 distribution. Because by the definition of $\mathbf{A}$,

$$\bar{Z} = \frac{Y_1}{\sqrt{n}}, \tag{B.3.28}$$

the theorem will be proved once we establish the identity

$$\sum_{k=2}^{n} Y_k^2 = \sum_{k=1}^{n} (Z_k - \bar{Z})^2. \tag{B.3.29}$$

Now

$$\sum_{k=1}^{n} (Z_k - \bar{Z})^2 = \sum_{k=1}^{n} Z_k^2 - 2\bar{Z} \sum_{k=1}^{n} Z_k + n\bar{Z}^2 = \sum_{k=1}^{n} Z_k^2 - n\bar{Z}^2. \tag{B.3.30}$$

Therefore, by (B.3.28),

$$\sum_{k=1}^{n} (Z_k - \bar{Z})^2 = \sum_{k=1}^{n} Z_k^2 - Y_1^2. \tag{B.3.31}$$

Finally,

$$\sum_{k=1}^{n} Y_k^2 = |\mathbf{Y}|^2 = |\mathbf{AZ} - \mathbf{A0}|^2 = |\mathbf{Z}|^2 = \sum_{k=1}^{n} Z_k^2. \tag{B.3.32}$$

Assertion (B.3.29) follows. □

B.4 THE BIVARIATE NORMAL DISTRIBUTION

The normal distribution is the most ubiquitous object in statistics. It appears in theory as an approximation to the distribution of sums of independent random variables, of order statistics, of maximum likelihood estimates, and so on. In practice it turns out that variables arising in all sorts of situations, such as errors of measurement, height, weight, yields of chemical and biological processes, and so on, are approximately normally distributed.

In the same way, the family of k-variate normal distributions arises on theoretical grounds when we consider the limiting behavior of sums of independent k-vectors of random variables and in practice as an approximation to the joint distribution of k-variables. Examples are given in Section 6.2. In this section we focus on the important case $k = 2$ where all properties can be derived relatively easily without matrix calculus and we can draw pictures. The general k-variate distribution is presented in Section B.6 following a more thorough introduction to moments of random vectors.

Recall that if Z has a standard normal distribution, we obtain the $\mathcal{N}(\mu, \sigma^2)$ distribution as the distribution of $g(Z) = \sigma Z + \mu$. Thus, Z generates the location-scale family of $\mathcal{N}(\mu, \sigma^2)$ distributions. The analogue of the standard normal distribution in two dimensions is the distribution of a random pair with two independent standard normal components, whereas the generalization of the family of maps $g(z) = \sigma z + \mu$ is the group of affine transformations. This suggests the following definition, in which we let the independent $\mathcal{N}(0, 1)$ random variables Z_1 and Z_2 generate our family of bivariate distributions.

A planar vector (X, Y) has a *bivariate normal distribution* if, and only if, there exist constants a_{ij}, $1 \leq i, j \leq 2, \mu_1, \mu_2$, and independent standard normal random variables Z_1, Z_2 such that

$$\begin{aligned} X &= \mu_1 + a_{11}Z_1 + a_{12}Z_2 \\ Y &= \mu_2 + a_{21}Z_1 + a_{22}Z_2. \end{aligned} \tag{B.4.1}$$

In matrix notation, if $\mathbf{A} = (a_{ij})$, $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$, $\mathbf{X} = (X, Y)^T$, $\mathbf{Z} = (Z_1, Z_2)^T$, the definition is equivalent to

$$\mathbf{X} = \mathbf{AZ} + \boldsymbol{\mu}. \tag{B.4.2}$$

Two important properties follow from the Definition (B.4.1).

Proposition B.4.1 *The marginal distributions of the components of a bivariate normal random vector are (univariate) normal or degenerate (concentrate on one point).*

This is a consequence of (A.13.23). The converse is not true. See Problem B.4.10. Note that

$$E(X) = \mu_1 + a_{11}E(Z_1) + a_{12}E(Z_2) = \mu_1,\ E(Y) = \mu_2 \tag{B.4.3}$$

and define

$$\sigma_1 = \sqrt{\text{Var } X},\ \sigma_2 = \sqrt{\text{Var } Y}. \tag{B.4.4}$$

Then X has a $\mathcal{N}(\mu_1, \sigma_1^2)$ and Y a $\mathcal{N}(\mu_2, \sigma_2^2)$ distribution.

Proposition B.4.2 *If we apply an affine transformation* $\mathbf{g}(\mathbf{x}) = \mathbf{Cx} + \mathbf{d}$ *to a vector* $\mathbf{X}$*, which has a bivariate normal distribution, then* $\mathbf{g}(\mathbf{X})$ *also has such a distribution.*

This is clear because

$$\mathbf{CX} + \mathbf{d} = \mathbf{C}(\mathbf{AZ} + \boldsymbol{\mu}) + \mathbf{d} = (\mathbf{CA})\mathbf{Z} + (\mathbf{C}\boldsymbol{\mu} + \mathbf{d}). \tag{B.4.5}$$

We now show that the bivariate normal distribution can be characterized in terms of first- and second-order moments and derive its density. As in Section A.11, let

$$\rho = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_1 \sigma_2} \tag{B.4.6}$$

if $\sigma_1\sigma_2 \neq 0$. If $\sigma_1\sigma_2 = 0$, it will be convenient to let $\rho = 0$. We define the *variance-covariance matrix* of (X, Y) (or of the distribution of (X, Y)) as the matrix of central second moments

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}. \tag{B.4.7}$$

This symmetric matrix is in many ways the right generalization of the variance to two dimensions. We see this in Theorem B.4.1 and (B.4.21). A general definition and its properties (for k dimensions) are given in Section B.5.

Theorem B.4.1 *Suppose that $\sigma_1\sigma_2 \neq 0$ and $|\rho| < 1$. Then*

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{2\pi\sqrt{\det \boldsymbol{\Sigma}}} \exp\left[-\frac{1}{2}((\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}))\right] \tag{B.4.8}$$

$$\begin{aligned} &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right.\right. \\ &\qquad \left.\left. -2\rho\frac{(x-\mu_1)}{\sigma_1}\frac{(y-\mu_2)}{\sigma_2} + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right\}\right]. \end{aligned} \tag{B.4.9}$$

Proof. Because (X, Y) is an affine transformation of (Z_1, Z_2), we can use Corollary B.2.1 to obtain the joint density of (X, Y) provided $\mathbf{A}$ is nonsingular. We start by showing that $\mathbf{AA}^T = \boldsymbol{\Sigma}$. Note that

$$\mathbf{AA}^T = \begin{pmatrix} a_{11}^2 + a_{12}^2 & a_{11}a_{21} + a_{12}a_{22} \\ a_{11}a_{21} + a_{12}a_{22} & a_{21}^2 + a_{22}^2 \end{pmatrix}$$

while

$$\begin{aligned} \sigma_1^2 &= \text{Var}(a_{11}Z_1) + \text{Var}(a_{12}Z_2) = a_{11}^2 \text{ Var } Z_1 + a_{12}^2 \text{ Var } Z_2 \\ &= a_{11}^2 + a_{12}^2,\ \sigma_2^2 = a_{21}^2 + a_{22}^2 \end{aligned} \tag{B.4.10}$$

and

$$\begin{aligned} \rho\sigma_1\sigma_2 &= \text{Cov}(a_{11}Z_1 + a_{12}Z_2,\ a_{21}Z_1 + a_{22}Z_2) \\ &= a_{11}a_{21} \text{ Cov}(Z_1, Z_1) \\ &\quad +(a_{12}a_{21} + a_{11}a_{22}) \text{ Cov}(Z_1, Z_2) \\ &\quad +a_{12}a_{22} \text{ Cov}(Z_2, Z_2) \\ &= a_{11}a_{21} + a_{12}a_{22}. \end{aligned} \tag{B.4.11}$$

Therefore, $\mathbf{A}\mathbf{A}^T = \boldsymbol{\Sigma}$ and by using elementary properties of determinants we obtain

$$\begin{aligned} |\det \mathbf{A}| &= \sqrt{[\det \mathbf{A}]^2} = \sqrt{\det \mathbf{A} \det \mathbf{A}^T} = \sqrt{\det \mathbf{A}\mathbf{A}^T} \\ &= \sqrt{\det \boldsymbol{\Sigma}} = \sigma_1\sigma_2\sqrt{1-\rho^2}. \end{aligned} \tag{B.4.12}$$

Because $|\rho| < 1$ and $\sigma_1\sigma_2 \neq 0$, we see that $\mathbf{A}$ is nonsingular and can apply Corollary B.2.1 to obtain the density of $\mathbf{X}$. The density of $\mathbf{Z}$ can be written

$$p_{\mathbf{Z}}(\mathbf{z}) = \frac{1}{2\pi}\exp\left(-\frac{1}{2}\mathbf{z}^T\mathbf{z}\right). \tag{B.4.13}$$

As in (B.3.19),

$$\begin{aligned} p_{\mathbf{X}}(\mathbf{x}) &= \frac{1}{2\pi|\det \mathbf{A}|}\left\{\exp -\frac{1}{2}[\mathbf{A}^{-1}(\mathbf{x}-\boldsymbol{\mu})]^T[\mathbf{A}^{-1}(\mathbf{x}-\boldsymbol{\mu})]\right\} \\ &= \frac{1}{2\pi|\det \mathbf{A}|}\left\{\exp -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T[\mathbf{A}^{-1}]^T\mathbf{A}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}. \end{aligned} \tag{B.4.14}$$

Because

$$[\mathbf{A}^{-1}]^T \cdot \mathbf{A}^{-1} = [\mathbf{A}\mathbf{A}^T]^{-1} = \boldsymbol{\Sigma}^{-1} \tag{B.4.15}$$

we arrive at (B.4.8). Finally (B.4.9) follows because by the formulas for an inverse

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \frac{1}{\sigma_1^2(1-\rho^2)} & \frac{-\rho}{\sigma_1\sigma_2(1-\rho^2)} \\ \frac{-\rho}{\sigma_1\sigma_2(1-\rho^2)} & \frac{1}{\sigma_2^2(1-\rho^2)} \end{pmatrix}. \tag{B.4.16}$$

□

From (B.4.7) it is clear that $\boldsymbol{\Sigma}$ is nonsingular if, and only if, $\sigma_1\sigma_2 \neq 0$ and $|\rho| < 1$. Bivariate normal distributions with $\sigma_1\sigma_2 \neq 0$ and $|\rho| < 1$ are referred to as *nondegenerate*, whereas others are *degenerate*. If $\sigma_1^2 = a_{11}^2 + a_{12}^2 = 0$, then $X \equiv \mu_1$, Y is necessarily distributed as $\mathcal{N}(\mu_2, \sigma_2^2)$, while $\sigma_2^2 = 0$ implies that $Y \equiv \mu_2$ and X has a $\mathcal{N}(\mu_1, \sigma_1^2)$ distribution. Finally, $\sigma_1\sigma_2 \neq 0$ and $|\rho| = 1$ implies by (A.11.16) that

$$\frac{(Y-\mu_2)}{\sigma_2} = \rho\frac{(X-\mu_1)}{\sigma_1}. \tag{B.4.17}$$

Because the marginal distributions of X and Y are, as we have already noted, $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$ respectively, relation (B.4.17) specifies the joint distribution of (X, Y) completely. Degenerate distributions do not have densities but correspond to random vectors whose marginal distributions are normal or degenerate and are such that (X, Y) falls on a fixed line or a point with probability 1.

Note that when $\rho = 0$, $p_{\mathbf{X}}(\mathbf{x})$ becomes the joint density of two independent normal variables. Thus, in the bivariate normal case, independence is equivalent to correlation zero. This is not true in general. An example is given in Problem B.4.11.

Now, suppose that we are given nonnegative constants σ_1, σ_2, a number ρ such that $|\rho| \leq 1$ and numbers μ_1, μ_2. Then we can construct a random vector (X, Y) having a

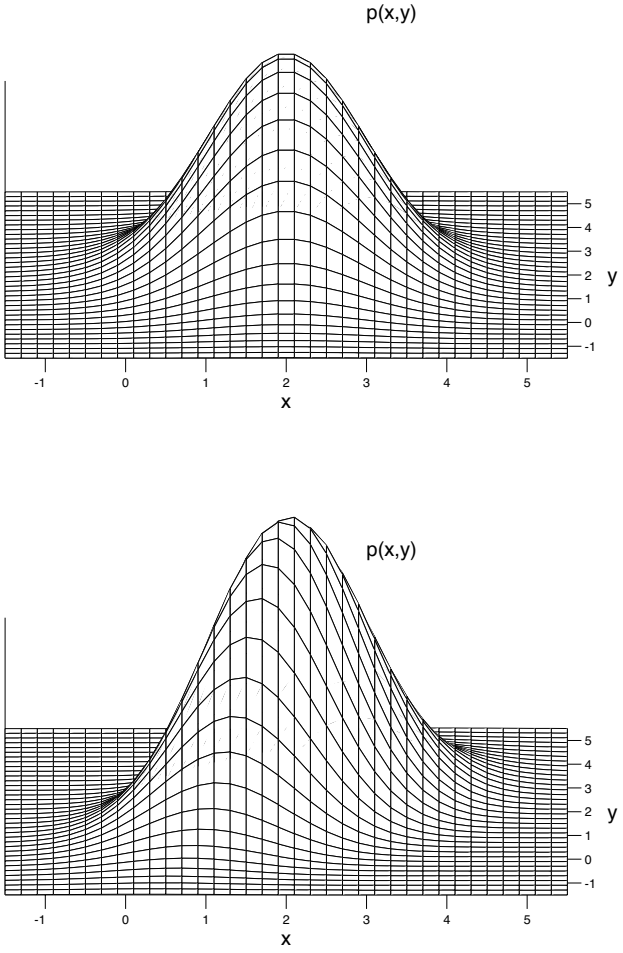

Figure B.4.1 A plot of the bivariate normal density $p(x, y)$ for $\mu_1 = \mu_2 = 2$, $\sigma_1 = \sigma_2 = 1$, $\rho = 0$ (top) and $\rho = 0.5$ (bottom).

bivariate normal distribution with vector of means (μ_1, μ_2) and variance-covariance matrix $\mathbf{\Sigma}$ given by (B.4.7). For example, take

$$X = \mu_1 + \sigma_1 Z_1,\ Y = \mu_2 + \sigma_2(\rho Z_1 + \sqrt{1-\rho^2} Z_2) \tag{B.4.18}$$

and apply (B.4.10) and (B.4.11). A bivariate normal distribution with this moment structure will be referred to as $\mathcal{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ or $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$.

Now suppose that $\mathbf{U} = (U_1, U_2)^T$ is obtained by an affine transformation,

$$\begin{aligned} U_1 &= c_{11}X + c_{12}Y + \nu_1 \\ U_2 &= c_{21}X + c_{22}Y + \nu_2 \end{aligned} \tag{B.4.19}$$

from a vector $(X, Y)^T$ having a $\mathcal{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ distribution. By Proposition B.4.2, $\mathbf{U}$ has a bivariate normal distribution. In view of our discussion, this distribution is completely determined by the means, variances, and covariances of U_1 and U_2, which in turn may be expressed in terms of the μ_i, σ_i^2, ρ, c_{ij}, and ν_i. Explicitly,

$$\begin{aligned} E(U_1) &= \nu_1 + c_{11}\mu_1 + c_{12}\mu_2,\ E(U_2) = \nu_2 + c_{21}\mu_1 + c_{22}\mu_2 \\ \text{Var } U_1 &= c_{11}^2\sigma_1^2 + c_{12}^2\sigma_2^2 + 2c_{12}c_{11}\rho\sigma_1\sigma_2 \\ \text{Var } U_2 &= c_{21}^2\sigma_1^2 + c_{22}^2\sigma_2^2 + 2c_{21}c_{22}\rho\sigma_1\sigma_2 \\ \text{Cov}(U_1, U_2) &= c_{11}c_{21}\sigma_1^2 + c_{12}c_{22}\sigma_2^2 + (c_{11}c_{22} + c_{12}c_{21})\rho\sigma_1\sigma_2. \end{aligned} \tag{B.4.20}$$

In matrix notation we can write compactly

$$\begin{aligned} (E(U_1), E(U_2)) &= (\nu_1, \nu_2)^T + \mathbf{C}(\mu_1, \mu_2)^T = \boldsymbol{\nu} + \mathbf{C}\boldsymbol{\mu}, \\ \boldsymbol{\Sigma}(\mathbf{U}) &= \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^T \end{aligned} \tag{B.4.21}$$

where $\boldsymbol{\Sigma}(\mathbf{U})$ denotes the covariance matrix of $\mathbf{U}$.

If the distribution of (X, Y) is nondegenerate and we take

$$\begin{aligned} \mathbf{C}^T &= \begin{pmatrix} \frac{1}{\sigma_1} & \frac{-\rho}{\sigma_1\sqrt{1-\rho^2}} \\ 0 & \frac{1}{\sigma_2\sqrt{1-\rho^2}} \end{pmatrix} \\ (\nu_1, \nu_2)' &= -\mathbf{C}(\mu_1, \mu_2)^T \end{aligned} \tag{B.4.22}$$

then U_1 and U_2 are independent and identically distributed standard normal random variables. Therefore, starting with any nondegenerate bivariate normal distribution we may by an affine transformation of the vector obtain any other bivariate normal distribution.

Another very important property of bivariate normal distributions is that normality is preserved under conditioning. That is,

Theorem B.4.2 *If (X, Y) has a nondegenerate $\mathcal{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ distribution, then the conditional distribution of Y given $X = x$ is*

$$\mathcal{N}\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1),\ \sigma_2^2(1 - \rho^2)\right).$$

Proof. Because X has a $\mathcal{N}(\mu_1, \sigma_1^2)$ distribution and (Y, X) is nondegenerate, we need only

calculate

$$\begin{aligned} p(y \mid x) &= \frac{p_{(\mathbf{Y},\mathbf{X})}(y,x)}{p_{\mathbf{X}}(x)} \\ &= \frac{1}{\sigma_2\sqrt{2\pi(1-\rho^2)}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(y-\mu_2)^2}{\sigma_2^2} - \frac{2\rho}{\sigma_2\sigma_1}(y-\mu_2)(x-\mu_1)\right.\right. \\ &\quad \left.\left.+[1-(1-\rho^2)]\frac{(x-\mu_1)^2}{\sigma_1^2}\right]\right\} \\ &= \frac{1}{\sigma_2\sqrt{2\pi(1-\rho^2)}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(y-\mu_2)}{\sigma_2} - \rho\frac{(x-\mu_1)}{\sigma_1}\right]^2\right\}. \end{aligned} \tag{B.4.23}$$

This is the density we want. □

Theorem B.4.2 shows that the conditional mean of Y given $X = x$ falls on the line $y = \mu_2 + \rho(\sigma_2/\sigma_1)(x - \mu_1)$. This line is called the *regression line*. See Figure B.4.2, which also gives the *contour plot* $S_c = \{(x, y) : f(x, y) = c\}$ where c is selected so that $P((X, Y) \in S_c) = \gamma$, $\gamma = 0.25$, 0.50, and 0.75. Such a contour plot is also called a $100\gamma\%$ *probability level curve*. See also Problem B.4.6.

By interchanging the roles of Y and X, we see that the conditional distribution of X given $Y = y$ is $\mathcal{N}(\mu_1 + (\sigma_1/\sigma_2)\rho(y - \mu_2),\ \sigma_1^2(1 - \rho^2))$.

With the convention that $0/0 = 0$ the theorem holds in the degenerate case as well. More generally, the conditional distribution of any linear combination of X and Y given any other linear combination of X and Y is normal (Problem B.4.9).

As we indicated at the beginning of this section the bivariate normal family of distributions arises naturally in limit theorems for sums of independent random vectors. The main result of this type is the bivariate central limit theorem. We postpone its statement and proof to Section B.6 where we can state it for the k-variate normal.

B.5 MOMENTS OF RANDOM VECTORS AND MATRICES

We generalize univariate notions from Sections A.10 to A.12 in this section. Let $\mathbf{U}$, respectively $\mathbf{V}$, denote a random k, respectively l, vector or more generally $\mathbf{U} = \|U_{ij}\|_{k\times l}$, $\mathbf{V} = \|V_{ij}\|_{h\times l}$, matrices of random variables. Suppose $E|U_{ij}| < \infty$ for all i, j. Define the expectation of $\mathbf{U}$ by

$$E(\mathbf{U}) = \|E(U_{ij})\|_{k\times l}.$$

B.5.1 Basic Properties of Expectations

If $\mathbf{A}_{m\times k}$, $\mathbf{B}_{m\times h}$ are nonrandom and $E\mathbf{U}$, $E\mathbf{V}$ are defined, then

$$E(\mathbf{AU} + \mathbf{BV}) = \mathbf{A}E(\mathbf{U}) + \mathbf{B}E(\mathbf{V}). \tag{B.5.1}$$

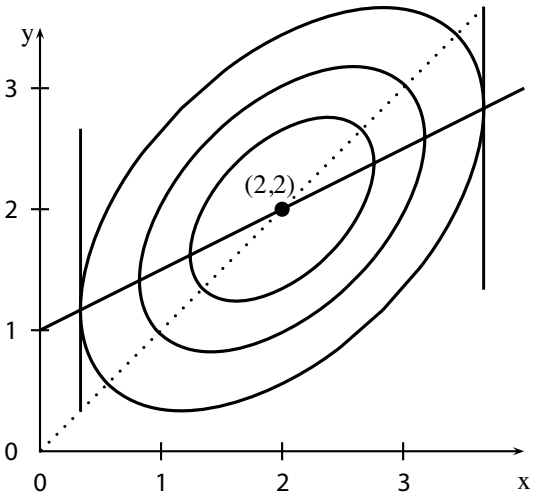

Figure B.4.2 25%, 50%, and 75% probability level-curves, the regression line (solid line), and major axis (dotted line) for the $\mathcal{N}(2, 2, 1, 1, 0.5)$ density.

This is an immediate consequence of the linearity of expectation for random variables and the definitions of matrix multiplication.

If $\mathbf{U} = \mathbf{c}$ with probability 1,

$$E(\mathbf{U}) = \mathbf{c}.$$

For a random vector $\mathbf{U}$, suppose $EU_i^2 < \infty$ for $i = 1, \ldots, k$ or equivalently $E(|\mathbf{U}|^2) < \infty$, where $|\cdot|$ denotes Euclidean distance. Define the *variance* of $\mathbf{U}$, often called the *variance-covariance* matrix, by

$$\begin{aligned} \mathrm{Var}(\mathbf{U}) &= E(\mathbf{U} - E(\mathbf{U}))(\mathbf{U} - E(\mathbf{U}))^T \\ &= \|\mathrm{Cov}(U_i, U_j)\|_{k\times k}, \end{aligned} \tag{B.5.2}$$

a symmetric matrix.

B.5.2 Properties of Variance

If $\mathbf{A}$ is $m \times k$ as before,

$$\mathrm{Var}(\mathbf{AU}) = \mathbf{A}\mathrm{Var}(\mathbf{U})\mathbf{A}^T. \tag{B.5.3}$$

Note that $\mathrm{Var}(\mathbf{U})$ is $k \times k$, $\mathrm{Var}(\mathbf{AU})$ is $m \times m$.

Let $\mathbf{c}_{k\times 1}$ denote a constant vector. Then

$$\text{Var}(\mathbf{U}+\mathbf{c}) = \text{Var}(\mathbf{U}). \tag{B.5.4}$$

$$\text{Var}(\mathbf{c}) = \|0\|_{k\times k}. \tag{B.5.5}$$

If $\mathbf{a}_{k\times 1}$ is constant we can apply (B.5.3) to obtain

$$\begin{aligned} \text{Var}(\mathbf{a}^T\mathbf{U}) &= \text{Var}(\Sigma_{j=1}^k a_j U_j) \\ &= \mathbf{a}^T\text{Var}(\mathbf{U})\mathbf{a} = \textstyle\sum_{i,j} a_i a_j \,\text{Cov}(U_i, U_j). \end{aligned} \tag{B.5.6}$$

Because the variance of any random variable is nonnegative and $\mathbf{a}$ is arbitrary, we conclude from (B.5.6) that $\text{Var}(\mathbf{U})$ is a *nonnegative definite symmetric matrix*.

The following proposition is important.

Proposition B.5.1 *If $E|\mathbf{U}|^2 < \infty$, then Var$(\mathbf{U})$ is positive definite if and only if, for every $\mathbf{a} \neq \mathbf{0}$, b,*

$$P[\mathbf{a}^T\mathbf{U} + b = 0] < 1. \tag{B.5.7}$$

Proof. By the definition of positive definite, (B.10.1), $\text{Var}(\mathbf{U})$ is not positive definite iff $\mathbf{a}^T\text{Var}(\mathbf{U})\mathbf{a} = 0$ for some $\mathbf{a} \neq 0$. By (B.5.6) that is equivalent to $\text{Var}(\mathbf{a}^T\mathbf{U}) = 0$, which is equivalent to (B.5.7) by (A.11.9). □

If $\mathbf{U}_{k\times 1}$ and $\mathbf{W}_{k\times 1}$ are independent random vectors with $E|\mathbf{U}|^2 < \infty$, $E|\mathbf{W}|^2 < \infty$, then

$$\text{Var}(\mathbf{U}+\mathbf{W}) = \text{Var}(\mathbf{U}) + \text{Var}(\mathbf{W}). \tag{B.5.8}$$

This follows by checking the identity element by element.

More generally, if $E|\mathbf{U}|^2 < \infty$, $E|\mathbf{V}|^2 < \infty$, define the covariance of $\mathbf{U}_{k\times 1}$, $\mathbf{V}_{l\times 1}$ by

$$\begin{aligned} \text{Cov}(\mathbf{U},\mathbf{V}) &= E(\mathbf{U}-E(\mathbf{U}))(\mathbf{V}-E(\mathbf{V}))^T \\ &= \|\text{Cov}(U_i, V_j)\|_{k\times l}. \end{aligned}$$

Then, if $\mathbf{U}, \mathbf{V}$ are independent

$$\text{Cov}(\mathbf{U},\mathbf{V}) = \mathbf{0}. \tag{B.5.9}$$

In general

$$\text{Cov}(\mathbf{A}\mathbf{U}+\mathbf{a},\ \mathbf{B}\mathbf{V}+\mathbf{b}) = \mathbf{A}\text{Cov}(\mathbf{U},\mathbf{V})\mathbf{B}^T \tag{B.5.10}$$

for nonrandom $\mathbf{A}$, $\mathbf{a}$, $\mathbf{B}$, $\mathbf{b}$, and

$$\text{Var}(\mathbf{U}+\mathbf{V}) = \text{Var}(\mathbf{U}) + 2\,\text{Cov}(\mathbf{U},\mathbf{V}) + \text{Var}(\mathbf{V}). \tag{B.5.11}$$

We leave (B.5.10) and (B.5.11) to the problems.

Define the *moment generating function* (m.g.f.) of $\mathbf{U}_{k\times 1}$ for $\mathbf{t} \in R^k$ by

$$M(\mathbf{t}) = M_{\mathbf{U}}(\mathbf{t}) = E(e^{\mathbf{t}^T\mathbf{U}}) = E(e^{\Sigma_{j=1}^k t_j U_j}).$$

Note that $M(\mathbf{t})$ can be ∞ for all $\mathbf{t} \neq \mathbf{0}$. In parallel[(1)] define the *characteristic function* (c.f.) of $\mathbf{U}$ by,

$$\varphi(\mathbf{t}) = E(e^{i\mathbf{t}^T\mathbf{U}}) = E(\cos(\mathbf{t}^T U)) + iE(\sin(\mathbf{t}^T\mathbf{U}))$$

where $i = \sqrt{-1}$. Note that φ is defined for all $\mathbf{t} \in R^k$, all $\mathbf{U}$. The proofs of the following theorems are beyond the scope of this book.

Theorem B.5.1 *Let $S = \{\mathbf{t} : M(\mathbf{t}) < \infty\}$. Then,*

(a) *S is convex.* (*See* **B.9**).

(b) *If S has a nonempty interior S^0, (contains a sphere $S(\mathbf{0}, \epsilon)$, $\epsilon > 0$), then M is analytic on S^0. In that case $E|\mathbf{U}|^p < \infty$ for all p. Then, if $i_1 + \cdots + i_k = p$,*

$$\frac{\partial^p M(\mathbf{0})}{\partial t_1^{i_1} \ldots \partial t_k^{i_k}} = E(U_1^{i_1} \ldots U_k^{i_k}). \tag{B.5.12}$$

In particular,

$$\left\| \frac{\partial M}{\partial t_j}(\mathbf{0}) \right\|_{k\times 1} = E(\mathbf{U}) \tag{B.5.13}$$

and

$$\left\| \frac{\partial^2 M(\mathbf{0})}{\partial t_i \partial t_j} \right\|_{k\times k} = E(\mathbf{U}\mathbf{U}^T). \tag{B.5.14}$$

(c) *If S^0 is nonempty, M determines the distribution of $\mathbf{U}$ uniquely.*

Expressions (B.5.12)–(B.5.14) are valid with φ replacing M if $E|\mathbf{U}|^{p/2} < \infty$, $E|\mathbf{U}| < \infty$, $E|\mathbf{U}|^2 < \infty$, respectively. The characteristic function always determines the distribution of $\mathbf{U}$ uniquely.

Proof. See Billingsley (1995).

The *cumulant generating function* of $\mathbf{U}$ is defined by $K(\mathbf{t}) = K_{\mathbf{U}}(\mathbf{t}) = \log M(\mathbf{t})$. If $S(t) = \{\mathbf{t} : M(\mathbf{t}) < \infty\}$ has a nonempty interior, then we define the cumulants as

$$c_{i_1 \ldots i_k} = c_{i_1 \ldots i_k}(\mathbf{U}) = \frac{\partial^p}{\partial t_1^{i_1} \ldots \partial t_k^{i_k}} K(\mathbf{t})\Big|_{\mathbf{t}=\mathbf{0}}, \; i_1 + \cdots + i_k = p.$$

An important consequence of the definitions and (A.9.3) is that if $\mathbf{U}_{k\times 1}$, $\mathbf{V}_{k\times 1}$ are independent then

$$M_{\mathbf{U}+\mathbf{V}}(\mathbf{t}) = M_{\mathbf{U}}(\mathbf{t}) M_{\mathbf{V}}(\mathbf{t}), \; K_{\mathbf{U}+\mathbf{V}}(\mathbf{t}) = K_{\mathbf{U}}(\mathbf{t}) + K_{\mathbf{V}}(\mathbf{t}) \tag{B.5.15}$$

where we use subscripts to indicate the vector to which the m.g.f. belongs. The same type of identity holds for c.f.'s. Other properties of cumulants are explored in the problems. See also Barndorff–Nielsen and Cox (1989).

Example B.5.1 *The Bivariate Normal Distribution.* If $(U_1, U_2)^T$ have a $\mathcal{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ distribution then, it is easy to check that

$$E(\mathbf{U}) = \boldsymbol{\mu} \tag{B.5.16}$$

$$\text{Var}(\mathbf{U}) = \boldsymbol{\Sigma} \tag{B.5.17}$$

$$M_{\mathbf{U}}(\mathbf{t}) = \exp\left\{\mathbf{t}^T\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^T\boldsymbol{\Sigma}\mathbf{t}\right\} \tag{B.5.18}$$

where $\mathbf{t} = (t_1, t_2)^T$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ are defined as in (B.4.7), (B.4.8). Similarly

$$\varphi_{\mathbf{U}}(\mathbf{t}) = \exp\left\{i\mathbf{t}^T\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^T\boldsymbol{\Sigma}\mathbf{t}\right\} \tag{B.5.19}$$

obtained by substituting it_j for t_j, $1 \le j \le k$, in (B.5.18). The result follows directly from (A.13.20) because

$$E(\exp(\mathbf{t}^T\mathbf{U})) = E(\exp(t_1U_1 + t_2U_2)) \tag{B.5.20}$$

and by (B.4.20), $t_1U_1 + t_2U_2$ has a $\mathcal{N}(\mathbf{t}^T\boldsymbol{\mu}, \mathbf{t}^T\boldsymbol{\Sigma}\mathbf{t})$ distribution.

By taking the log in (B.5.18) and differentiating we find the first five cumulants

$$(c_{10}, c_{01}, c_{20}, c_{02}, c_{11}) = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \sigma_1\sigma_2\rho).$$

All other cumulants are zero. □

B.6 THE MULTIVARIATE NORMAL DISTRIBUTION

B.6.1 Definition and Density

We define the multivariate normal distribution in two ways and show they are equivalent. From the equivalence we are able to derive the basic properties of this family of distributions rapidly.

Definition B.6.1 $\mathbf{U}_{k\times 1}$ has a *multivariate (k-variate) normal distribution* iff $\mathbf{U}$ can be written as

$$\mathbf{U} = \boldsymbol{\mu} + \mathbf{A}\mathbf{Z}$$

when $\boldsymbol{\mu}_{k\times 1}$, $\mathbf{A}_{k\times k}$ are constant and $\mathbf{Z} = (Z_1, \ldots, Z_k)^T$ where the Z_j are independent standard normal variables. This is the immediate generalization of our definition of the bivariate normal. We shall show that as in the bivariate case the distribution of $\mathbf{U}$ depends on $\boldsymbol{\mu} = E(\mathbf{U})$ and $\boldsymbol{\Sigma} \equiv \text{Var}(\mathbf{U})$, only.

Definition B.6.2 $\mathbf{U}_{k\times 1}$ has a multivariate normal distribution iff for every $\mathbf{a}_{k\times 1}$ nonrandom, $\mathbf{a}^T\mathbf{U} = \Sigma_{j=1}^k a_jU_j$ has a univariate normal distribution.

Theorem B.6.1 *Definitions* B.6.1 *and* B.6.2 *define the same family of distributions.*

Proof. If $\mathbf{U}$ is given by Definition B.6.1, then $\mathbf{a}^T\mathbf{U} = \mathbf{a}^T(\mathbf{AZ} + \boldsymbol{\mu}) = [\mathbf{A}^T\mathbf{a}]^T\mathbf{Z} + \mathbf{a}^T\boldsymbol{\mu}$, a linear combination of independent normal variables and, hence, normal. Conversely, if $X = \mathbf{a}^T\mathbf{U}$ has a univariate normal distribution, necessarily this is $\mathcal{N}(E(X), \text{Var}(X))$. But

$$E(X) = \mathbf{a}^T E(\mathbf{U}) \tag{B.6.1}$$

$$\text{Var}(X) = \mathbf{a}^T \text{ Var}(\mathbf{U})\mathbf{a} \tag{B.6.2}$$

from (B.5.1) and (B.5.4). Note that the finiteness of $E(|\mathbf{U}|)$ and $\text{Var}(\mathbf{U})$ is guaranteed by applying Definition B.6.2 to $\mathbf{e}_j^T\mathbf{U}$, where $\mathbf{e}_j$ denotes the $k \times 1$ coordinate vector with 1 in the jth coordinate and 0 elsewhere. Now, by definition,

$$\begin{aligned} M_{\mathbf{U}}(\mathbf{a}) &= E(\exp(\mathbf{a}^T\mathbf{U})) = E(e^X) \\ &= \exp\left\{\mathbf{a}^T E(\mathbf{U}) + \tfrac{1}{2}\mathbf{a}^T \text{ Var}(\mathbf{U})\mathbf{a}\right\} \end{aligned} \tag{B.6.3}$$

from (A.13.20), for all $\mathbf{a}$. Thus, by Theorem B.5.1 the distribution of $\mathbf{U}$ under Definition B.6.2 is completely determined by $E(\mathbf{U})$, $\text{Var}(\mathbf{U})$. We now appeal to the principal axis theorem (B.10.1.1). If $\boldsymbol{\Sigma}$ is nonnegative definite symmetric, there exists $\mathbf{A}_{k\times k}$ such that

$$\boldsymbol{\Sigma} = \mathbf{AA}^T, \tag{B.6.4}$$

where $\mathbf{A}$ is nonsingular iff $\boldsymbol{\Sigma}$ is positive definite. Now, given $\mathbf{U}$ defined by B.6.2 with $E(\mathbf{U}) = \boldsymbol{\mu}$, $\text{Var}(\mathbf{U}) = \boldsymbol{\Sigma}$, consider

$$\mathbf{V}_{k\times 1} = \boldsymbol{\mu} + \mathbf{AZ}$$

where $\mathbf{A}$ and $\mathbf{Z}$ are as in Definition B.6.1. Then

$$E(\mathbf{V}) = \boldsymbol{\mu}, \ \text{Var}(\mathbf{V}) = \mathbf{A}\text{Var}(\mathbf{Z})\mathbf{A}^T = \mathbf{AA}^T$$

because $\text{Var}(\mathbf{Z}) = \mathbf{J}_{k\times k}$, the identity matrix.

Then, by definition, $\mathbf{V}$ satisfies Definition B.6.1 and, hence, B.6.2 and has the same first and second moments as $\mathbf{U}$. Since first and second moments determine the k-variate normal distribution uniquely, $\mathbf{U}$ and $\mathbf{V}$ have the same distribution and the theorem follows. □

Notice that we have also proved:

Corollary B.6.1. *Given arbitrary $\boldsymbol{\mu}_{k\times 1}$ and $\boldsymbol{\Sigma}$ nonnegative definite symmetric, there is a unique k-variate normal distribution with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$.*

We use $\mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote the k-variate normal distribution of Corolary B.6.1. Arguing from Corollary B.2.1 we see the following.

Theorem B.6.2 *If $\boldsymbol{\Sigma}$ is positive definite or equivalently nonsingular, then if $\mathbf{U} \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{U}$ has a density given by*

$$p_{\mathbf{U}}(\mathbf{x}) = \frac{1}{(2\pi)^{k/2}[\det(\boldsymbol{\Sigma})]^{1/2}} \times \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\}. \tag{B.6.5}$$

Proof. Apply Definition B.6.1 with $\mathbf{A}$ such that $\mathbf{\Sigma} = \mathbf{A}\mathbf{A}^T$.

The converse that $\mathbf{U}$ has a density only if $\mathbf{\Sigma}$ is positive definite and similar more refined results are left to Problem B.6.2.

There is another important result that follows from the spectral decomposition theorem (B.10.1.2).

Theorem B.6.3 *If $\mathbf{U}_{k\times 1}$ has $\mathcal{N}_k(\boldsymbol{\mu}, \mathbf{\Sigma})$ distribution, there exists an orthogonal matrix $\mathbf{P}_{k\times k}$ such that $\mathbf{P}^T\mathbf{U}$ has an $\mathcal{N}_k(\boldsymbol{\nu}, \mathbf{D}_{k\times k})$ distribution where $\boldsymbol{\nu} = \mathbf{P}^T\boldsymbol{\mu}$ and $\mathbf{D}_{k\times k}$ is the diagonal matrix whose diagonal entries are the necessarily nonnegative eigenvalues of $\mathbf{\Sigma}$. If $\mathbf{\Sigma}$ is of rank $l < k$, necessarily only l eigenvalues are positive and conversely.*

Proof. By the spectral decomposition theorem there exists $\mathbf{P}$ orthogonal such that

$$\mathbf{\Sigma} = \mathbf{P}\mathbf{D}\mathbf{P}^T.$$

Then $\mathbf{P}^T\mathbf{U}$ has a $\mathcal{N}_k(\boldsymbol{\nu}, \mathbf{D})$ distribution since $\mathrm{Var}(\mathbf{P}^T\mathbf{U}) = \mathbf{P}^T\mathbf{\Sigma}\mathbf{P} = \mathbf{D}$ by orthogonality of $\mathbf{P}$. □

This result shows that an arbitrary normal random vector can be linearly transformed to a normal random vector with independent coordinates, some possibly degenerate. In the bivariate normal case, (B.4.19) and (B.4.22) transformed an arbitrary nondegenerate bivariate normal pair to an i.i.d. $\mathcal{N}(0,1)$ pair.

Note that if rank $\mathbf{\Sigma} = k$ and we set

$$\mathbf{\Sigma}^{\frac{1}{2}} = \mathbf{P}\mathbf{D}^{\frac{1}{2}}\mathbf{P}^T,\ \mathbf{\Sigma}^{-\frac{1}{2}} = \left(\mathbf{\Sigma}^{\frac{1}{2}}\right)^{-1} = \mathbf{P}\left(\mathbf{D}^{\frac{1}{2}}\right)^{-1}\mathbf{P}^T$$

where $\mathbf{D}^{\frac{1}{2}}$ is the diagonal matrix with diagonal entries equal to the square root of the eigenvalues of $\mathbf{\Sigma}$, then

$$\mathbf{Z} = \mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{U} - \boldsymbol{\mu}) \tag{B.6.6}$$

has a $\mathcal{N}(\mathbf{0}, \mathbf{J})$ distribution, where $\mathbf{J}$ is the $k \times k$ identity matrix.

Corollary B.6.2 *If $\mathbf{U}$ has an $\mathcal{N}_k(\mathbf{0}, \mathbf{\Sigma})$ distribution and $\mathbf{\Sigma}$ is of rank k, then $\mathbf{U}^T\mathbf{\Sigma}^{-1}\mathbf{U}$ has a χ^2_k distribution.*

Proof. By (B.6.6), $\mathbf{U}^T\mathbf{\Sigma}^{-1}\mathbf{U} = \mathbf{Z}^T\mathbf{Z}$, where $\mathbf{Z}$ is given by (B.6.6). But $\mathbf{Z}^T\mathbf{Z} = \sum_{i=1}^k Z_i^2$ where Z_i are i.i.d. $\mathcal{N}(0,1)$. The result follows from (B.3.1). □

B.6.2 Basic Properties. Conditional Distributions

If $\mathbf{U}$ is $\mathcal{N}_k(\boldsymbol{\mu}, \mathbf{\Sigma})$ and $\mathbf{A}_{l\times k}$, $\mathbf{b}_{l\times 1}$ are nonrandom, then $\mathbf{A}\mathbf{U} + \mathbf{b}$ is $\mathcal{N}_l(\mathbf{A}\boldsymbol{\mu} + \mathbf{b},\ \mathbf{A}\mathbf{\Sigma}\mathbf{A}^T)$. This follows immediately from Definition B.6.2. In particular, marginal distributions of blocks of coordinates of $\mathbf{U}$ are normal. For the next statement we need the following block matrix notation. Given $\mathbf{\Sigma}_{k\times k}$ positive definite, write

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix} \tag{B.6.7}$$

where $\mathbf{\Sigma}_{11}$ is the $l \times l$ variance of $(U_1, \ldots, U_l)^T$, which we denote by $\mathbf{U}^{(1)}$, $\mathbf{\Sigma}_{22}$ the $(k-l)\times(k-l)$ variance of $(U_{l+1}, \ldots, U_k)^T$ denoted by $\mathbf{U}^{(2)}$, and $\mathbf{\Sigma}_{12} = \mathrm{Cov}(\mathbf{U}^{(1)}, \mathbf{U}^{(2)})_{l\times k-l}$, $\mathbf{\Sigma}_{21} = \mathbf{\Sigma}_{12}^T$. Similarly write $\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{pmatrix}$, where $\boldsymbol{\mu}^{(1)}$ and $\boldsymbol{\mu}^{(2)}$ are the mean vectors of $\mathbf{U}^{(1)}$ and $\mathbf{U}^{(2)}$.

We next show that independence and uncorrelatedness are the same for the k variate normal. Specifically

Theorem B.6.4 *If* $\mathbf{U}_{(k+l)\times 1} = \begin{pmatrix} \mathbf{U}^{(1)} \\ \mathbf{U}^{(2)} \end{pmatrix}$, *where* $\mathbf{U}^{(1)}$ *and* $\mathbf{U}^{(2)}$ *are* k *and* l *vectors, respectively, has a* $k + l$ *variate normal distribution, then* $\mathbf{U}^{(1)}$ *and* $\mathbf{U}^{(2)}$ *are independent iff*

$$\mathrm{Cov}(\mathbf{U}^{(1)}, \mathbf{U}^{(2)})_{k\times l} = \|\mathbf{0}\|_{k\times l}. \tag{B.6.8}$$

Proof. If $\mathbf{U}^{(1)}$ and $\mathbf{U}^{(2)}$ are independent, (B.6.8) follows from (A.11.22). Let $\mathbf{U}^{(1)*}$ and $\mathbf{U}^{(2)*}$ be independent $\mathcal{N}_k(E(\mathbf{U}^{(1)}), \mathrm{Var}(\mathbf{U}^{(1)}))$, $\mathcal{N}_l(E(\mathbf{U}^{(2)}), \mathrm{Var}(\mathbf{U}^{(2)}))$. Then we show below that $\mathbf{U}^* \equiv \begin{pmatrix} \mathbf{U}^{(1)*} \\ \mathbf{U}^{(2)*} \end{pmatrix}$ has the same distribution as $\mathbf{U}$ and, hence, $\mathbf{U}^{(1)}, \mathbf{U}^{(2)}$ are independent. To see that $\mathbf{U}^*$ has the same distribution as $\mathbf{U}$ note that

$$E\mathbf{U}^* = E\mathbf{U} \tag{B.6.9}$$

by definition, and because $\mathbf{U}^{(1)}$ and $\mathbf{U}^{(2)}$ have $\Sigma_{12} = \mathbf{0}$ and by construction $\mathrm{Var}(\mathbf{U}^{(j)*}) = \Sigma_{jj}$, $j = 1, 2$, then

$$\mathrm{Var}(\mathbf{U}) = \mathrm{Var}(\mathbf{U}^*) \tag{B.6.10}$$

by (B.6.7). Therefore, $\mathbf{U}$ and $\mathbf{U}^*$ must have the same distribution by the determination of the k-variate normal by first and second moments. □

Theorem B.6.5 *If* $\mathbf{U}$ *is distributed as* $\mathcal{N}_k(\boldsymbol{\mu}, \mathbf{\Sigma})$, *with* $\mathbf{\Sigma}$ *positive definite as previously, then the conditional distribution of* $\mathbf{U}^{(1)}$ *given* $\mathbf{U}^{(2)} = \mathbf{u}^{(2)}$ *is* $\mathcal{N}_l(\boldsymbol{\mu}^{(1)} + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}(\mathbf{u}^{(2)} - \boldsymbol{\mu}^{(2)}), \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$. *Moreover* $\mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$ *is positive definite so there is a conditional density given by (B.6.5) with* $(\mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21})$ *substituted for* $\mathbf{\Sigma}$ *and* $\boldsymbol{\mu}^{(1)} + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}(\mathbf{u}^{(2)} - \boldsymbol{\mu}^{(2)})$ *for* $\boldsymbol{\mu}$.

Proof. Identification of the conditional density as normal can be done by direct computation from the formula

$$p(\mathbf{u}^{(1)} \mid \mathbf{u}^{(2)}) = p_{\mathbf{U}}(\mathbf{u})/p_{\mathbf{U}^{(1)}}(\mathbf{u}^{(2)}) \tag{B.6.11}$$

after noting that $\mathbf{\Sigma}_{11}$ is positive definite because the marginal density must exist.

To derive this and also obtain the required formula for conditional expectation and variance we proceed as follows. That $\mathbf{\Sigma}_{11}, \mathbf{\Sigma}_{22}$ are positive definite follows by using $\mathbf{a}^T\mathbf{\Sigma}\mathbf{a} > 0$ with $\mathbf{a}$ whose last $k - l$ or first l coordinates are 0. Next note that

$$\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21} = \mathrm{Var}(\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{U}^{(2)}) \tag{B.6.12}$$

because

$$\begin{aligned} \text{Var}(\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{U}^{(2)}) &= \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\,\text{Var}(\mathbf{U}^{(2)})\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \\ &= \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{22}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \end{aligned} \tag{B.6.13}$$

by (B.5.3). Furthermore, we claim

$$\text{Cov}(\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{U}^{(2)},\ \mathbf{U}^{(1)} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{U}^{(2)}) = \mathbf{0} \tag{B.6.14}$$

(Problem B.6.4) and, hence, by Theorem B.6.4, $\mathbf{U}^{(1)} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{U}^{(2)}$ and $\mathbf{U}^{(2)}$ are independent. Thus, the conditional distribution of $\mathbf{U}^{(1)} - \Sigma_{12}\Sigma_{22}^{-1}\mathbf{U}^{(2)}$ given $\mathbf{U}^{(2)} = \mathbf{u}^{(2)}$ is the same as its marginal distribution. By the substitution property of the conditional distribution this is the same as the conditional distribution of $\mathbf{U}^{(1)} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{u}^{(2)}$ given $\mathbf{U}^{(2)} = \mathbf{u}^{(2)}$. The result now follows by adding $\Sigma_{12}\Sigma_{22}^{-1}\mathbf{U}^{(2)}$ and noting that

$$\text{Var}(\mathbf{U}^{(1)} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{U}^{(2)}) = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \tag{B.6.15}$$

and

$$\begin{aligned} E(\mathbf{U}_1 \mid \mathbf{U}_2 = \mathbf{u}_2) &= E(\mathbf{U}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{U}_2 \mid \mathbf{U}_2 = \mathbf{u}_2) + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{u}_2 \\ &= E(\mathbf{U}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{U}_2) + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{u}_2 \\ &= \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{u}_2 \\ &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{u}_2 - \boldsymbol{\mu}_2). \end{aligned}$$

□

Theorem B.6.6 The Multivariate Central Limit Theorem. *Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be independent and identically distributed random k vectors with $E|\mathbf{X}_1|^2 < \infty$. Let $E(\mathbf{X}_1) = \boldsymbol{\mu}$, $\text{Var}(\mathbf{X}_1) = \boldsymbol{\Sigma}$, and let $\mathbf{S}_n = \Sigma_{i=1}^n \mathbf{X}_i$. Then, for every continuous function $g : R^k \to R$,*

$$g\left(\frac{\mathbf{S}_n - n\boldsymbol{\mu}}{\sqrt{n}}\right) \xrightarrow{\mathcal{L}} g(\mathbf{Z}) \tag{B.6.16}$$

where $\mathbf{Z} \sim \mathcal{N}_k(\mathbf{0}, \boldsymbol{\Sigma})$.

As a consequence, if $\boldsymbol{\Sigma}$ is positive definite, we can use Theorem B.7.1 to conclude that

$$P\left[\frac{\mathbf{S}_n - n\boldsymbol{\mu}}{\sqrt{n}} \le \mathbf{z}\right] \to P[\mathbf{Z} \le \mathbf{z}] \tag{B.6.17}$$

for all $\mathbf{z} \in R^k$. Here $\{\mathbf{x} : \mathbf{x} \le \mathbf{z}\} = \{\mathbf{x} : x_i \le z_i,\ i = 1, \ldots, k\}$ where as usual subscripts indicate coordinate labels.

A proof of this result may be found in more advanced texts in probability, for instance, Billingsley (1995) and in Chung (1974).

An important corollary follows.

Corollary B.6.7 *If the $\mathbf{X}_i$ are as in the statement of Theorem* B.6.6, *if $\boldsymbol{\Sigma}$ is positive definite and if $\bar{\mathbf{X}} = \frac{1}{n}\Sigma_{i=1}^n \mathbf{X}_i$, then*

$$n(\bar{\mathbf{X}} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \xrightarrow{\mathcal{L}} \chi_k^2. \tag{B.6.18}$$

Proof. $\sqrt{n}(\bar{\mathbf{X}} - \boldsymbol{\mu}) = (\mathbf{S}_n - n\mu)/\sqrt{n}$. Thus, we need only note that the function $g(\mathbf{x}) \equiv \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}$ from R^k to R is continuous and that if $\mathbf{Z} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, then $\mathbf{Z}^T\boldsymbol{\Sigma}^{-1}\mathbf{Z} \sim \chi^2_k$ (Corollary B.6.2). □

B.7 CONVERGENCE FOR RANDOM VECTORS: O_P AND o_P NOTATION

The notion of convergence in probability and convergence in law for random variables discussed in section A.1.5 generalizes to random vectors and even abstract valued random elements taking their values in metric spaces. We give the required generalizations for random vectors and, hence, random matrices here. We shall also introduce a unique notation that makes many computations easier. In the following, $|\cdot|$ denotes Euclidean distance.

B.7.1 A sequence of random vectors $\mathbf{Z}_n \equiv (Z_{n1}, \ldots, Z_{nd})^T$ converges *in probability* to $\mathbf{Z} \equiv (Z_1, \ldots, Z_d)^T$ iff

$$|\mathbf{Z}_n - \mathbf{Z}| \xrightarrow{P} 0$$

or equivalently $Z_{nj} \xrightarrow{P} Z_j$ for $1 \le j \le d$.

Note that this definition also makes sense if the $\mathbf{Z}_n$ are considered under probabilities P_n that depend on n. Thus, $\mathbf{Z}_n \xrightarrow{P_n} \mathbf{Z}$ iff

$$P_n[|\mathbf{Z}_n - \mathbf{Z}| \ge \epsilon] \to 0 \qquad \text{for every } \epsilon > 0.$$

WLLN (the weak law of large numbers). *Let* $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ *be i.i.d. as* $\mathbf{Z}$ *and let* $\bar{\mathbf{Z}}_n = n^{-1}\sum_{i=1}^n \mathbf{Z}_i$. *If* $E|\mathbf{Z}| < \infty$, *then* $\bar{\mathbf{Z}}_n \xrightarrow{P} \boldsymbol{\mu} = E\mathbf{Z}$.

When $E|\mathbf{Z}|^2 < \infty$, the result follows from Chebychev's inequality as in Appendix A. For a proof in the $E|\mathbf{Z}| < \infty$ case, see Billingsley (1995).

The following definition is subtler.

B.7.2 A sequence $\{\mathbf{Z}_n\}$ of random vectors *converges in law* to $\mathbf{Z}$, written $\mathbf{Z}_n \xrightarrow{\mathcal{L}} \mathbf{Z}$ or $\mathcal{L}(\mathbf{Z}_n) \to \mathcal{L}(\mathbf{Z})$, iff

$$h(\mathbf{Z}_n) \xrightarrow{\mathcal{L}} h(\mathbf{Z})$$

for all functions $h : R^d \to R$, h continuous.

We saw this type of convergence in the central limit theorem (B.6.6).

Note that in the definition of convergence in law, the random vectors $\mathbf{Z}_n, \mathbf{Z}$ only play the role of defining marginal distributions. No requirement is put on joint distributions of $\{\mathbf{Z}_n\}, \mathbf{Z}$. Thus, if $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ are i.i.d., $\mathbf{Z}_n \xrightarrow{\mathcal{L}} \mathbf{Z}_1$, but $\mathbf{Z}_n \not\xrightarrow{P} \mathbf{Z}_1$.

An equivalent statement to (B.7.2) is

$$Eg(\mathbf{Z}_n) \to Eg(\mathbf{Z}) \tag{B.7.3}$$

for all $\mathbf{g} : R^d \to R$ continuous and bounded. Note that (B.7.3) implies (A.14.6). The following stronger statement can be established.

Theorem B.7.1 $\mathbf{Z}_n \xrightarrow{\mathcal{L}} \mathbf{Z}$ *iff (B.7.3) holds for every* $\mathbf{g} : R^d \to R^p$ *such that* $\mathbf{g}$ *is bounded and if* $A_{\mathbf{g}} \equiv \{\mathbf{z} : \mathbf{g}$ *is continuous at* $\mathbf{z}\}$ *then* $P[\mathbf{Z} \in A_{\mathbf{g}}] = 1$.

Here are some further properties.

Proposition B.7.1

(a) *If* $\mathbf{Z}_n \xrightarrow{P} \mathbf{Z}$ *and* $\mathbf{g}$ *is continuous from* R^d *to* R^p, *then* $\mathbf{g}(\mathbf{Z}_n) \xrightarrow{P} \mathbf{g}(\mathbf{Z})$.

(b) *The implication in* (a) *continues to hold if "*P*" is replaced by "*$\mathcal{L}$*" in premise and conclusion above*.

(c) *The conclusion of* (a) *and* (b) *continues to hold if continuity of* $\mathbf{g}$ *is replaced by* $P[\mathbf{Z} \in A_{\mathbf{g}}] = 1$ *where* $A_{\mathbf{g}} \equiv \{\mathbf{z} : \mathbf{g}$ *is continuous at* $\mathbf{z}\}$.

B.7.4 If $\mathbf{Z}_n \xrightarrow{P} \mathbf{Z}$ then $\mathbf{Z}_n \xrightarrow{\mathcal{L}} \mathbf{Z}$.

A partial converse follows.

B.7.5 If $\mathbf{Z}_n \xrightarrow{\mathcal{L}} \mathbf{z}_0$ (a constant), then $\mathbf{Z}_n \xrightarrow{P} \mathbf{z}_0$.

Note that (B.7.4) and (B.7.5) generalize (A.14.3), (A.14.4).

Theorem B.7.2 Slutsky's Theorem. *Suppose* $\boldsymbol{Z}^{\boldsymbol{T}}{}_n = (\boldsymbol{U}^{\boldsymbol{T}}{}_n, \boldsymbol{V}^{\boldsymbol{T}}{}_n)$ *where* $\mathbf{Z}_n$ *is a* d *vector,* $\boldsymbol{U}_n$ *is* b*-dimensional,* $\boldsymbol{V}_n$ *is* $c = d - b$*-dimensional and*

(a) $\boldsymbol{U}_n \xrightarrow{\mathcal{L}} \boldsymbol{U}$

(b) $\boldsymbol{V}_n \xrightarrow{\mathcal{L}} \boldsymbol{v}$ *where* $\boldsymbol{v}$ *is a constant vector.*

(c) $\mathbf{g}$ *is a continuous function from* R^d *to* R^b.

Then

$$\mathbf{g}(\boldsymbol{U}_{\boldsymbol{n}}^{\boldsymbol{T}}, \boldsymbol{V}_{\boldsymbol{n}}^{\boldsymbol{T}}) \xrightarrow{\mathcal{L}} \mathbf{g}(\boldsymbol{U}^{\boldsymbol{T}}, \boldsymbol{v}^{\boldsymbol{T}}).$$

Again continuity of $\mathbf{g}$ can be weakened to $P[(\boldsymbol{U}^{\boldsymbol{T}}, \boldsymbol{v}^{\boldsymbol{T}})^T \in A_{\mathbf{g}}] = 1$.
We next give important special cases of Slutsky's theorem:

Example B.7.1

(a) $d = 2, b = c = 1, g(u, v) = \alpha u + \beta v$, $g(u, v) = uv$ or $g(u, v) = \frac{u}{v}$ and $v \neq 0$. This covers (A.14.9)

(b) $\boldsymbol{V}_n = ||V_{nij}||_{b \times b}, c = b^2, g(\boldsymbol{u}^{\boldsymbol{T}}, \boldsymbol{v}^{\boldsymbol{T}}) = \boldsymbol{vu}$ where $\boldsymbol{v}$ is a $b \times b$ matrix. To apply Theorem B.7.2, rearrange $\mathbf{V}_n$ and $\mathbf{v}$ as $c \times 1$ vectors with $c = b^2$.

Combining this with $b = c = d/2, g(\boldsymbol{u}^T, \boldsymbol{v}^T) = \boldsymbol{u} + \boldsymbol{v}$, we obtain, that if the $b \times b$ matrix $||\boldsymbol{V}_n|| \xrightarrow{P} ||\boldsymbol{v}||$ and $\boldsymbol{W}_n$, $b \times 1$, tends in probability to $\boldsymbol{w}$, a constant vector, and $\boldsymbol{U}_n \xrightarrow{\mathcal{L}} \boldsymbol{U}$, then

$$\boldsymbol{V}_n\boldsymbol{U}_n + \boldsymbol{W}_n \xrightarrow{\mathcal{L}} \boldsymbol{v}\boldsymbol{U} + \boldsymbol{w}. \tag{B.7.6}$$

The proof of Theorem B.7.1 and other preceding results comes from the following theorem due to Hammersley (1952), which relates the two modes of convergence. Skorokhod (1956) extended the result to function spaces.

Theorem B.7.3 Hammersley. *Suppose vectors* $\mathbf{Z}_n \xrightarrow{\mathcal{L}} \mathbf{Z}$ *in the sense of Definition B.7.2. There exists (on a suitable probability space) a sequence of random vectors* $\{\boldsymbol{Z}_n^*\}$ *and a vector* $\boldsymbol{Z}^*$ *such that*

(i) $\mathcal{L}(\boldsymbol{Z}_n^*) = \mathcal{L}(\mathbf{Z}_n)$ for all n, $\mathcal{L}(\boldsymbol{Z}^*) = \mathcal{L}(\mathbf{Z})$

(ii) $\boldsymbol{Z}_n^* \xrightarrow{P} \boldsymbol{Z}^*$.

A somewhat stronger statement can also be made, namely, that

$$\boldsymbol{Z}_n^* \xrightarrow{a.s.} \boldsymbol{Z}^*$$

where $\xrightarrow{a.s.}$ refers to almost sure convergence defined by

$$\mathbf{Z}_n \xrightarrow{a.s.} \mathbf{Z} \text{ if } P\left(\lim_{n\to\infty} \mathbf{Z}_n = \mathbf{Z}\right) = 1.$$

This type of convergence also appears in the following famous law.

SLLN (the strong law of large numbers). *Let* $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ *be i.i.d. as* $\mathbf{Z}$ *and let* $\bar{\mathbf{Z}}_n = n^{-1}\sum_{i=1}^n \mathbf{Z}_i$, *then* $\bar{\mathbf{Z}}_n \xrightarrow{a.s.} \mu = E\mathbf{Z}$ *iff* $E|\mathbf{Z}| < \infty$.

For a proof, see Billingsley (1995).

The proof of Theorem B.7.3 is easy for $d = 1$. (Problem B.7.1) For the general case refer to Skorokhod (1956). Here are the proofs of some of the preceding assertions using Hammersley's theorem and the following.

Theorem B.7.4. Bounded convergence theorem. *If the vector* $\mathbf{U}_n$ *converges in probability to* $\mathbf{U}$ *and* g *is bounded and* $P[\mathbf{U} \in A_g] = 1$, *then*

$$Eg(\mathbf{U}_n) \to Eg(\mathbf{U}).$$

For a proof see Billingsley (1995, p. 209). Evidently, Theorem B.7.4 gives the equivalence between (B.7.2) and (B.7.3) and establishes Theorem B.7.1.

The proof of Proposition B.7.1(a) is easy if g is uniformly continuous; that is, for every $\epsilon > 0$ there exists $\delta > 0$ such that

$$\{(\mathbf{z}_1, \mathbf{z}_2) : |g(\mathbf{z}_1) - g(\mathbf{z}_2)| \geq \epsilon\} \subset \{(\mathbf{z}_1, \mathbf{z}_2) : |\mathbf{z}_1 - \mathbf{z}_2| \geq \delta\}.$$

A stronger result (in view of Theorem B.7.4) is as follows.

Theorem B.7.5 Dominated Convergence Theorem. *If $\{W_n\}, W$ and V are random variables with $W_n \xrightarrow{P} W$, $P[|W_n| \le |V|] = 1$, and $E|V| < \infty$, then $EW_n \to EW$.*

Proposition B.7.1(b) and (c) follow from the (a) part and Hammersley's theorem. Then (B.7.3) follows from the dominated convergence because if g is bounded by M and uniformly continuous, then for $\delta > 0$

$$|E_P g(\mathbf{Z}_n) - E_P g(\mathbf{Z})| \le \sup\{|g(\boldsymbol{z}) - g(\boldsymbol{z'})| : |\boldsymbol{z} - \boldsymbol{z'}| \le \delta\} + MP[|\mathbf{Z}_n - \mathbf{Z}| \ge \delta] \tag{B.7.7}$$

Let $n \to \infty$ to obtain that

$$\limsup_n |E_P g(\mathbf{Z}_n) - E_P g(\mathbf{Z})| \le \sup\{|g(\mathbf{z}) - g(\mathbf{z}')| : |\mathbf{z} - \mathbf{z}'| \le \delta\} \tag{B.7.8}$$

and let $\delta \to 0$. The general argument is sketched in Problem B.7.3.

For B.7.5 let $h_\epsilon(\boldsymbol{z}) = 1(|\boldsymbol{z} - \boldsymbol{z_0}| \ge \epsilon)$. Note that $A_{h_\epsilon} = \{\boldsymbol{z} : |\boldsymbol{z} - \boldsymbol{z_0}| \ne \epsilon\}$. Evidently if $P[\mathbf{Z} = \boldsymbol{z_0}] = 1$, $P[Z \in A_{h_\epsilon}] = 1$ for all $\epsilon > 0$. Therefore, by Problem B.7.4, $P[|\mathbf{Z}_n - \boldsymbol{z_0}| \ge \epsilon] \to P[|\mathbf{Z} - \boldsymbol{z_0}| \ge \epsilon] = 0$ because $P[\mathbf{Z} = \boldsymbol{z_0}] = 1$ and the result follows.

Finally Slutsky's theorem is easy because by Hammersley's theorem there exist $\boldsymbol{V^*}_n$, $\boldsymbol{U^*}_n$ with the same marginal distributions as $\boldsymbol{V}_n, \boldsymbol{U}_n$ and $\boldsymbol{U^*}_n \xrightarrow{P} \boldsymbol{U^*}, \boldsymbol{V^*}_n \xrightarrow{P} \boldsymbol{v}$. Then $(\boldsymbol{U^*}_n, \boldsymbol{V^*}_n) \xrightarrow{P} (\boldsymbol{U^*}, \boldsymbol{v})$, which by Proposition B.7.1 implies that $(\boldsymbol{U}_n, \boldsymbol{V}_n) \xrightarrow{\mathcal{L}} (\boldsymbol{U}, \boldsymbol{v})$, which by Theorem B.7.1 implies Slutsky's theorem.

In deriving asymptotic properties of some statistical methods, it will be convenient to use convergence of densities. We will use the following.

Theorem B.7.6 Scheffé's Theorem. *Suppose $p_n(\mathbf{z})$ and $p(\mathbf{z})$ are densities or frequency functions on R^d such that $p_n(\mathbf{z}) \to p(\mathbf{z})$ as $n \to \infty$ for all $\mathbf{z} \in R^d$. Then*

$$\int |p_n(\mathbf{z}) - p(\mathbf{z})| d\mathbf{z} \to 0 \text{ as } n \to \infty$$

in the continuous case with a sum replacing the integral in the discrete case.

Proof. We give the proof in the continuous case. Note that

$$|p_n(\mathbf{z}) - p(\mathbf{z})| = p_n(\mathbf{z}) - p(\mathbf{z}) + 2[p(\mathbf{z}) - p_n(\mathbf{z})]^+$$

where $x^+ = \max\{0, x\}$. Thus,

$$\int |p_n(\mathbf{z}) - p(\mathbf{z})| dx = \int [p_n(\mathbf{z}) - p(\mathbf{z})] d\mathbf{z} + 2 \int [p(\mathbf{z}) - p_n(\mathbf{z})]^+ d\mathbf{z}.$$

The first term on the right is zero. The second term tends to zero by applying the dominated convergence theorem to $U_n = [p(\mathbf{Z}) - p_n(\mathbf{Z})]^+ / p(\mathbf{Z})$ and $g(u) = u$, $u \in [0, 1]$, because $[p(\mathbf{z}) - p_n(\mathbf{z})]^+ \le p(\mathbf{z})$. □

Proposition B.7.2 *If $\mathbf{Z}_n$ and $\mathbf{Z}$ have densities or frequency functions $p_n(\mathbf{z})$ and $p(\mathbf{z})$ with $p_n(\mathbf{z}) \to p(\mathbf{z})$ as $n \to \infty$ for all $\mathbf{z} \in R^d$, then $\mathbf{Z}_n \xrightarrow{\mathcal{L}} \mathbf{Z}$.*

Proof. We give the proof in the continuous case. Let $g : R^d \to R$ be continuous and bounded, say $|g| \leq M < \infty$. Then

$$|Eg(\mathbf{Z}_n) - Eg(\mathbf{Z})| = \left| \int g(\mathbf{z})[p_n(\mathbf{z}) - p(\mathbf{z})]d\mathbf{z} \right| \leq M \int |p_n(\mathbf{z}) - p(\mathbf{z})|d\mathbf{z}$$

and the result follows from (B.7.3) and Theorem B.7.5. □

Remark B.7.1 Theorem B.7.5 can be strengthened considerably with a suitable background in measure theory. Specifically, suppose μ is a sigma finite measure on $\mathcal{X}$. If g_n and g are measurable functions from $\mathcal{X}$ to R such that

(1) $g_n \to g$ in measure, i.e., $\mu\{x : |g_n(x) - g(x)| \geq \epsilon\} \to 0$ as $n \to \infty$ for all $\epsilon > 0$

and

(2) $\int |g_n|^r d\mu \to \int |g|^r d\mu$ as $n \to \infty$ for some $r \geq 1$, then $\int |g_n - g| d\mu \to 0$ as $n \to \infty$. A proof of this result can be found in Billingsley (1979, p. 184). □

Theorem B.7.7 Polya's Theorem. *Suppose real-valued $X_n \xrightarrow{\mathcal{L}} X$. Let F_n, F be the distribution functions of X_n, X, respectively. Suppose F is continuous. Then*

$$\sup_x |F_n(x) - F(x)| \to 0.$$

Outline of Proof. By Proposition B.7.1, $F_n(x) \to F(x)$ and $F_n(x - 0) \to F(x)$ for all x. Given $\epsilon > 0$, choose $\underline{x}$, $\bar{x}$ such that $F(\underline{x}) \leq \epsilon$, $1 - F(\bar{x}) \leq \epsilon$. Because F is uniformly continuous on $[\underline{x}, \bar{x}]$, there exists $\delta(\epsilon) > 0$ such that for all $\underline{x} \leq x_1, x_2 \leq \bar{x}$, $|x_1 - x_2| \leq \delta(\epsilon) \Rightarrow |F(x_1) - F(x_2)| \leq \epsilon$. Let $\underline{x} = x_0 < x_1 \cdots < x_K = \bar{x}$ be such that $|x_j - x_{j-1}| \leq \delta(\epsilon)$ for all j.

Then

$$\begin{aligned}\sup_{x_j \leq x \leq x_{j+1}} |F_n(x) - F(x)| &\leq \max\{|F_n(x_j) - F(x_j)|, |F_n(x_{j+1}) - F(x_{j+1})|\} \\ &+ \sup_{x_j \leq x \leq x_{j+1}} \{\max\{(F_n(x) - F_n(x_j)), F_n(x_{j+1}) - F_n(x)\} \\ &+ \max\{(F(x) - F(x_j)), F(x_{j+1}) - F(x)\}\}.\end{aligned}$$

The second term equals $(F_n(x_{j+1}) - F_n(x_j)) + (F(x_{j+1}) - F(x_j))$. Similarly,

$$\begin{aligned}\sup_{x \leq \underline{x}} |F_n(x) - F(x)| &\leq F_n(\underline{x}) + F(\underline{x}) \\ \sup_{x \geq \bar{x}} |F_n(x) - F(x)| &\leq (1 - F_n(\bar{x})) + (1 - F(\bar{x})).\end{aligned}$$

Conclude that, $\overline{\lim}_n \sup_x |F_n(x) - F(x)| \leq 3\epsilon$ and the theorem follows. □

We end this section with some useful notation.

The $O_P, \asymp_P$, and o_P Notation

The following asymptotic order in probability notation is useful.

$$\begin{aligned}
\boldsymbol{U}_n = o_P(1) \quad &\text{iff} \quad \boldsymbol{U}_n \xrightarrow{P} 0 \\
\boldsymbol{U}_n = O_P(1) \quad &\text{iff} \quad \forall \epsilon > 0, \exists M < \infty \text{ such that } \forall n \quad P[|\boldsymbol{U}_n| \geq M] \leq \epsilon \\
\boldsymbol{U}_n = o_P(\boldsymbol{V}_n) \quad &\text{iff} \quad \frac{|\boldsymbol{U}_n|}{|\boldsymbol{V}_n|} = o_P(1) \\
\boldsymbol{U}_n = O_P(\boldsymbol{V}_n) \quad &\text{iff} \quad \frac{|\boldsymbol{U}_n|}{|\boldsymbol{V}_n|} = O_P(1) \\
\boldsymbol{U}_n \asymp_P \boldsymbol{V}_n \quad &\text{iff} \quad \boldsymbol{U}_n = O_P(\boldsymbol{V}_n) \quad \text{and} \quad \boldsymbol{V}_n = O_P(\boldsymbol{U}_n).
\end{aligned}$$

Note that

$$O_P(1)o_P(1) = o_P(1),\ O_P(1) + o_P(1) = O_P(1), \tag{B.7.9}$$

and $\boldsymbol{U}_n \xrightarrow{\mathcal{L}} \boldsymbol{U} \Rightarrow \boldsymbol{U}_n = O_P(1)$.

Suppose $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ are i.i.d. as $\mathbf{Z}$ with $E|\mathbf{Z}| < \infty$. Set $\boldsymbol{\mu} = E(\mathbf{Z})$, then $\bar{\mathbf{Z}}_n = \boldsymbol{\mu} + o_P(1)$ by the WLLN. If $E|\mathbf{Z}|^2 < \infty$, then $\bar{\mathbf{Z}}_n = \boldsymbol{\mu} + O_P(n^{-\frac{1}{2}})$ by the central limit theorem.

B.8 MULTIVARIATE CALCULUS

B.8.1 A function $T : R^d \to R$ is *linear* iff

$$T(\alpha \mathbf{x}_1 + \beta \mathbf{x}_2) = \alpha T(\mathbf{x}_1) + \beta T(\mathbf{x}_2)$$

for all $\alpha, \beta \in R$, $\mathbf{x}_1, \mathbf{x}_2 \in R^d$. More generally, $T : \underbrace{R^d \times \cdots \times R^d}_{k} \to R$ is k *linear* iff $T(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k)$ is linear in each coordinate separately when the others are held fixed.

B.8.2 $\mathbf{T} \equiv (T_1, \ldots, T_p)$ mapping $\underbrace{R^d \times \cdots \times R^d}_{k} \to R^p$ is said to be k linear iff $T_1, \ldots, T_p$ are k linear as in B.8.1.

B.8.3 T is k linear as in B.8.1 iff there exists an array $\{a_{i_1,\ldots,i_k} : 1 \leq i_j \leq d,\ 1 \leq j \leq k\}$ such that if $\mathbf{x}_t \equiv (x_{t1}, \ldots, x_{td})$, $1 \leq t \leq k$, then

$$T(\mathbf{x}_1, \ldots, \mathbf{x}_k) = \sum_{i_k=1}^{d} \cdots \sum_{i_1=1}^{d} a_{i_1,\ldots,i_k} \prod_{j=1}^{k} x_{ji_j} \tag{B.8.4}$$

B.8.5 If $\mathbf{h} : \mathcal{O} \to R^p$, $\mathcal{O}$ open $\subset R^d$, $\mathbf{h} \equiv (h_1, \ldots, h_p)$, then $\mathbf{h}$ is *Fréchet differentiable* at $\mathbf{x} \in \mathcal{O}$ iff there exists a (necessarily unique) linear map $\mathbf{Dh}(\mathbf{x}) : R^d \to R^p$ such that

$$|\mathbf{h}(\mathbf{y}) - \mathbf{h}(\mathbf{x}) - \mathbf{Dh}(\mathbf{x})(\mathbf{y} - \mathbf{x})| = o(|\mathbf{y} - \mathbf{x}|) \tag{B.8.6}$$

where $|\cdot|$ is the Euclidean norm. If $p = 1$, $\mathbf{Dh}$ is the *total differential*.

More generally, $\mathbf{h}$ is m times *Fréchet differentiable* iff there exist l linear operators $\mathbf{D}^l\mathbf{h}(\mathbf{x}) : \underbrace{R^d \times \cdots \times R^d}_{l} \to R^p$, $1 \le l \le m$ such that

$$\left|\mathbf{h}(\mathbf{y}) - \mathbf{h}(\mathbf{x}) - \sum_{l=1}^{m} \frac{\mathbf{D}^l\mathbf{h}}{l!}(\mathbf{x})(\mathbf{y}-\mathbf{x},\ldots,\mathbf{y}-\mathbf{x})\right| = o(|\mathbf{y}-\mathbf{x}|^m). \tag{B.8.7}$$

B.8.8 If $\mathbf{h}$ is m times Fréchet differentiable, then for $1 \le j \le p$, h_j has partial derivatives of order $\le m$ at $\mathbf{x}$ and the jth component of $\mathbf{D}^l\mathbf{h}(\mathbf{x})$ is defined by the array $\left\{\frac{\partial^l h_j(\mathbf{x})}{\partial x_1^{\epsilon_1}\ldots\partial x_d^{\epsilon_d}} : \epsilon_1+\cdots+\epsilon_d = l,\ 0 \le \epsilon_i \le l,\ 1 \le i \le d\right\}$.

B.8.9 $\mathbf{h}$ is m times Fréchet differentiable at $\mathbf{x}$ if h_j has partial derivatives of order up to m on $\mathcal{O}$ that are continuous at $\mathbf{x}$.

B.8.10 Taylor's Formula

If h_j, $1 \le j \le p$ has continuous partial derivatives of order up to $m+1$ on $\mathcal{O}$, then, for all $\mathbf{x}, \mathbf{y} \in \mathcal{O}$,

$$\mathbf{h}(\mathbf{y}) = \mathbf{h}(\mathbf{x}) + \sum_{l=1}^{m} \frac{\mathbf{D}^l\mathbf{h}(\mathbf{x})}{l!}(\mathbf{y}-\mathbf{x},\ldots,\mathbf{y}-\mathbf{x}) + \frac{\mathbf{D}^{m+1}\mathbf{h}(\mathbf{x}^*)}{(m+1)!}(\mathbf{y}-\mathbf{x},\ldots,\mathbf{y}-\mathbf{x}) \tag{B.8.11}$$

for some $\mathbf{x}^* = \mathbf{x} + \alpha^*(\mathbf{y}-\mathbf{x})$, $0 \le \alpha^* \le 1$. These classical results may be found, for instance, in Dieudonné (1960) and Rudin (1991). As a consequence, we obtain the following.

B.8.12 Under the conditions of B.8.10,

$$\left|\mathbf{h}(\mathbf{y}) - \mathbf{h}(\mathbf{x}) - \sum_{l=1}^{m} \frac{\mathbf{D}^l\mathbf{h}(\mathbf{x})}{l!}(\mathbf{y}-\mathbf{x},\ldots,\mathbf{y}-\mathbf{x})\right|$$

$$\le ((m+1)!)^{-1} \sup\{|\mathbf{D}^{m+1}\mathbf{h}(\mathbf{x}')| : |\mathbf{x}'-\mathbf{x}| \le |\mathbf{y}-\mathbf{x}|\}|\mathbf{y}-\mathbf{x}|^{m+1}$$

for all $\mathbf{x}, \mathbf{y} \in \mathcal{O}$.

B.8.13 *Chain Rule.* Suppose $\mathbf{h} : \mathcal{O} \to R^p$ with derivative $\mathbf{Dh}$ and $\mathbf{g} : \mathbf{h}(\mathcal{O}) \to R^q$ with derivative $\mathbf{Dg}$. Then the composition $\mathbf{g} \circ \mathbf{h} : \mathcal{O} \to R^q$ is differentiable and

$$\mathbf{D}(\mathbf{g} \circ \mathbf{h})(\mathbf{x}) = (\mathbf{Dg}|_{\mathbf{h}(\mathbf{x})})(\mathbf{Dh}(x)).$$

As a consequence, we obtain the following.

B.8.14 Let $d = p$, $\mathbf{h}$ be $1-1$ and continuously Fréchet differentiable on a neighborhood of $\mathbf{x} \in \mathcal{O}$, and $\mathbf{Dh}(\mathbf{x}) = \left\|\frac{\partial h_i}{\partial x_j}(\mathbf{x})\right\|_{p\times p}$ be nonsingular. Then $\mathbf{h}^{-1} : \mathbf{h}(\mathcal{O}) \to \mathcal{O}$ is Fréchet differentiable at $\mathbf{y} = \mathbf{h}(\mathbf{x})$ and

$$\mathbf{Dh}^{-1}(\mathbf{h}(\mathbf{x})) = [\mathbf{Dh}(\mathbf{x})]^{-1}.$$

B.9 CONVEXITY AND INEQUALITIES

Convexity

A subset S of R^k is said to be *convex* if for every $\mathbf{x}, \mathbf{y} \in S$, and every $\alpha \in [0,1]$, $\alpha\mathbf{x}+(1-\alpha)\mathbf{y} \in S$. When $k=1$, convex sets are finite and infinite intervals. When $k>1$, spheres, rectangles, and hyperplanes are convex. The point $\mathbf{x}_0$ belongs to the interior S^0 of the convex set S iff for every $\mathbf{d} \neq 0$,

$$\{\mathbf{x} : \mathbf{d}^T\mathbf{x} > \mathbf{d}^T\mathbf{x}_0\} \cap S^0 \neq \emptyset \text{ and } \{\mathbf{x} : \mathbf{d}^T\mathbf{x} < \mathbf{d}^T\mathbf{x}_0\} \cap S^0 \neq \emptyset \tag{B.9.1}$$

where $\emptyset$ denotes the empty set.

A function g from a convex set S to R is said to be *convex* if

$$g(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha g(\mathbf{x}) + (1-\alpha)g(\mathbf{y}), \text{ all } \mathbf{x}, \mathbf{y} \in S, \ \alpha \in [0,1]. \tag{B.9.2}$$

g is said to be *strictly convex* if (B.9.2) holds with $\leq$ replaced by $<$ for all $\mathbf{x} \neq \mathbf{y}$, $\alpha \notin \{0,1\}$. Convex functions are continuous on S^0. When $k=1$, if g'' exists, convexity is equivalent to $g''(\mathbf{x}) \geq 0$, $\mathbf{x} \in S$; strict convexity holds if $g''(\mathbf{x}) > 0$, $\mathbf{x} \in S$. For g convex and fixed $\mathbf{x}, \mathbf{y} \in S$, $h(\alpha) = g(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}))$ is convex in $\alpha, \alpha \in [0,1]$. When $k>1$, if $\partial g^2(\mathbf{x})/\partial x_i \partial x_j$ exists, convexity is equivalent to

$$\sum_{i,j} u_i u_j \partial^2 g(\mathbf{x})/\partial x_i \partial x_j \geq 0, \text{ all } \mathbf{u} \in R^k \text{ and } \mathbf{x} \in S.$$

A function h from a convex set S to R is said to be *(strictly) concave* if $g = -h$ is (strictly) convex.

Jensen's Inequality. If $S \subset R^k$ is convex and closed, g is convex on S, $P[\mathbf{U} \in S] = 1$, and $E\mathbf{U}$ is finite, then $E\mathbf{U} \in S$, $Eg(\mathbf{U})$ exists and

$$Eg(\mathbf{U}) \geq g(E\mathbf{U}) \tag{B.9.3}$$

with equality if and only if there are a and $\mathbf{b}_{k\times 1}$ such that

$$P[g(\mathbf{U}) = \mathbf{a} + \mathbf{b}^T\mathbf{U}] = 1.$$

In particular, if g is strictly convex, equality holds in (B.9.3) if and only if $P[\mathbf{U} = \mathbf{c}] = 1$ for some $\mathbf{c}_{k\times 1}$.

For a proof see Rockafellar (1970). We next give a useful inequality relating product moments to marginal moments:

Hölder's Inequality. Let r and s be numbers with $r, s > 1$, $r^{-1} + s^{-1} = 1$. Then

$$E|XY| \leq \{E|X|^r\}^{\frac{1}{r}}\{E|Y|^s\}^{\frac{1}{s}}. \tag{B.9.4}$$

When $r = s = 2$, Hölder's inequality becomes the Cauchy–Schwartz inequality (A.11.17). For a proof of (B.9.4), see Billingsley (1995, p. 80) or Problem B.9.3.

We conclude with bounds for tails of distributions.

Bernstein Inequality for the Binomial Case. Let $S_n \sim \mathcal{B}(n,p)$, then

$$P(|S_n - np| \geq n\epsilon) \leq 2\exp\{-n\epsilon^2/2\} \text{ for } \epsilon > 0. \tag{B.9.5}$$

That is, the probability that S_n exceeds its expected value np by more than a multiple $n\epsilon$ of n tends to zero exponentially fast as $n \to \infty$. For a proof, see Problem B.9.1.

Hoeffding's Inequality. The exponential convergence rate (B.9.5) for the sum of independent Bernoulli variables extends to the sum $S_n = \sum_{i=1}^n X_i$ of i.i.d. bounded variables X_i, $|X_i - \mu| \leq c_i$, where $\mu = E(X_1)$

$$P[|S_n - n\mu| \geq x) \leq 2\exp\left\{-\tfrac{1}{2}x^2 / \sum_{i=1}^n c_i^2\right\}. \tag{B.9.6}$$

For a proof, see Grimmett and Stirzaker (1992, p. 449) or Hoeffding (1963).

B.10 TOPICS IN MATRIX THEORY AND ELEMENTARY HILBERT SPACE THEORY

B.10.1 Symmetric Matrices

We establish some of the results on symmetric nonnegative definite matrices used in the text and B.6. Recall $A_{p\times p}$ is *symmetric* iff $A = A^T$. A is *nonnegative definite* (nd) iff $\mathbf{x}^T A\mathbf{x} \geq 0$ for all $\mathbf{x}$, *positive definite* (pd) if the inequality is strict unless $\mathbf{x} = \mathbf{0}$.

B.10.1.1. The Principal Axis Theorem

(a) A is symmetric nonnegative definite (snd) iff there exist $C_{p\times p}$ such that

$$A = CC^T. \tag{B.10.1}$$

(b) A is symmetric positive definite (spd) iff C above is nonsingular.

The "if" part in (a) is trivial because then $\mathbf{x}^T A\mathbf{x} = \mathbf{x}^T CC^T\mathbf{x} = |C\mathbf{x}|^2$. The "only if" part in (b) follows because $|C\mathbf{x}|^2 > 0$ unless $\mathbf{x} = \mathbf{0}$ is equivalent to $C\mathbf{x} \neq \mathbf{0}$ unless $\mathbf{x} = \mathbf{0}$, which is nonsingularity. The "if" part in (b) follows by noting that C nonsingular iff $\det(C) \neq 0$ and $\det(CC^T) = \det^2(C)$. Parenthetically we note that if A is positive definite, A is nonsingular (Problem B.10.1). The "only if" part of (a) is deeper and follows from the spectral theorem.

B.10.1.2 Spectral Theorem

(a) $A_{p\times p}$ is symmetric iff there exists P orthogonal and $D = \text{diag}(\lambda_1, \ldots, \lambda_p)$ such that

$$A = PDP^T. \tag{B.10.2}$$

(b) The λ_j are real, unique up to labeling, and are the eigenvalues of A. That is, there exist vectors $\mathbf{e}_j$, $|\mathbf{e}_j| = 1$ such that

$$A\mathbf{e}_j = \lambda_j \mathbf{e}_j. \tag{B.10.3}$$

(c) If A is also snd, all the λ_j are nonnegative. The rank of A is the number of nonzero eigenvalues. Thus, A is positive definite iff all its eigenvalues are positive.

(d) In any case the vectors $\mathbf{e}_i$ can be chosen orthonormal and are then unique up to label.

Thus, Theorem B.10.1.2 may equivalently be written

$$A = \sum_{i=1}^{p} \mathbf{e}_i \mathbf{e}_i^T \lambda_i \tag{B.10.4}$$

where $\mathbf{e}_i \mathbf{e}_i^T$ can be interpreted as projection on the one-dimensional space spanned by $\mathbf{e}_i$ (Problem B.10.2).

(B.10.1) follows easily from B.10.3 by taking $C = P \operatorname{diag}(\lambda_1^{\frac{1}{2}}, \ldots, \lambda_p^{\frac{1}{2}})$ in (B.10.1).

The proof of the spectral theorem is somewhat beyond our scope—see Birkhoff and MacLane (1953, pp. 275–277, 314), for instance.

B.10.1.3 If A is spd, so is A^{-1}.

Proof. $A = P \operatorname{diag}(\lambda_1, \ldots, \lambda_p) P^T \Rightarrow A^{-1} = P \operatorname{diag}(\lambda_1^{-1}, \ldots, \lambda_p^{-1}) P^T$.

B.10.1.4 If A is spd, then $\max\{\mathbf{x}^T A \mathbf{x} : \mathbf{x}^T \mathbf{x} \leq 1\} = \max_j \lambda_j$.

B.10.2 Order on Symmetric Matrices

As we defined in the text for A, B symmetric $A \leq B$ iff $B - A$ is nonnegative definite. This is easily seen to be an ordering.

B.10.2.1 If A and B are symmetric and $A \leq B$, then for any C

$$CAC^T \leq CBC^T. \tag{B.10.5}$$

This follows from definition of snd or the principal axis theorem because $B - A$ snd means $B - A = EE^T$ and then $CBC^T - CAC^T = C(B-A)C^T = CEE^TC^T = (CE)(CE)^T$.

Furthermore, if A and B are spd and $A \leq B$, then

$$A^{-1} \geq B^{-1}. \tag{B.10.6}$$

Proof. After Bellman (1960, p. 92, Problems 13, 14). Note first that, if A is symmetric,

$$\mathbf{x}^T A^{-1} \mathbf{x} = \max\{\mathbf{y} : 2\mathbf{x}^T \mathbf{y} - \mathbf{y}^T A \mathbf{y}\} \tag{B.10.7}$$

because $\mathbf{y} = A^{-1}\mathbf{x}$ maximizes the quadratic form. Then, if $A \leq B$,

$$2\mathbf{x}^T\mathbf{y} - \mathbf{y}^T A\mathbf{y} \geq 2\mathbf{x}^T\mathbf{y} - \mathbf{y}^T B\mathbf{y}$$

for all $\mathbf{x}, \mathbf{y}$. By (B.10.7) we obtain $\mathbf{x}^T A^{-1}\mathbf{x} \geq \mathbf{x}^T B^{-1}\mathbf{x}$ for all $\mathbf{x}$ and the result follows. □

B.10.2.2 The Generalized Cauchy–Schwarz Inequality

Let $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ be spd, $(p+q)\times(p+q)$, with Σ_{11}, $p \times p$, Σ_{22}, $q \times q$. Then Σ_{11}, Σ_{22} are spd. Furthermore,

$$\Sigma_{11} \geq \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \tag{B.10.8}$$

Proof. From Section B.6 we have noted that there exist (Gaussian) random vectors $\mathbf{U}_{p\times 1}$, $\mathbf{V}_{q\times 1}$ such that $\Sigma = \text{Var}(\mathbf{U}^T, \mathbf{V}^T)^T$, $\Sigma_{11} = \text{Var}(\mathbf{U})$, $\Sigma_{22} = \text{Var}(\mathbf{V})$, $\Sigma_{12} = \text{cov}(\mathbf{U}, \mathbf{V})$. The argument given in B.6 establishes that

$$\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \text{Var}(\mathbf{U} - \Sigma_{12}\Sigma_{22}^{-1}\mathbf{V}) \tag{B.10.9}$$

and the result follows. □

B.10.2.3 We note also, although this is not strictly part of this section, that if $\mathbf{U}, \mathbf{V}$ are random vectors as previously (not necessarily Gaussian), then equality holds in (B.10.8) iff for some $\mathbf{b}$

$$\mathbf{U} = \mathbf{b} + \Sigma_{12}\Sigma_{22}^{-1}\mathbf{V} \tag{B.10.10}$$

with probability 1. This follows from (B.10.9) since $\mathbf{a}^T \text{Var}(\mathbf{U} - \Sigma_{12}\Sigma_{22}^{-1}\mathbf{V})\mathbf{a} = 0$ for all $\mathbf{a}$ iff

$$\mathbf{a}^T(\mathbf{U} - \Sigma_{12}\Sigma_{22}^{-1}\mathbf{V} - \mathbf{b}) = 0 \tag{B.10.11}$$

for all $\mathbf{a}$ where $\mathbf{b}$ is $E(\mathbf{U} - \Sigma_{12}\Sigma_{22}^{-1}\mathbf{V})$. But (B.10.11) for all $\mathbf{a}$ is equivalent to (B.10.10). □

B.10.3 Elementary Hilbert Space Theory

A linear space $\mathcal{H}$ over the reals is a Hilbert space iff

(i) It is endowed with an inner product $(\cdot,\cdot) : \mathcal{H} \times \mathcal{H} \to R$ such that $(\cdot,\cdot)$ is *bilinear*,

$$(ah_1 + bh_2, ch_3 + dh_4) = ab(h_1, h_2) + ac(h_1, h_3) + bc(h_2, h_3) + bd(h_2, h_4),$$

symmetric, $(h_1, h_2) = (h_2, h_1)$, and

$$(h, h) \geq 0$$

with equality iff $h = 0$.

It follows that if $\|h\|^2 \equiv (h,h)$, then $\|\cdot\|$ is a norm. That is,

(a) $\|h\| = 0$ iff $h = 0$
(b) $\|ah\| = |a|\|h\|$ for any scalar a
(c) $\|h_1 + h_2\| \leq \|h_1\| + \|h_2\|$. *Triangle inequality*

(ii) $\mathcal{H}$ is complete. That is, if $\{h_m\}_{m\geq 1}$ is such that $\|h_m - h_n\| \to 0$ as $m, n \to \infty$ then there exists $h \in \mathcal{H}$ such that $\|h_n - h\| \to 0$.

The prototypical example of a Hilbert space is Euclidean space R^p from which the abstraction is drawn. In this case if $\mathbf{x} = (x_1, \ldots, x_p)^T$, $\mathbf{y} = (y_1, \ldots, y_p)^T \in R^p$, $(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y} = \sum_{j=1}^p x_j y_j$, $\|\mathbf{x}\|^2 = \sum_{j=1}^p x_j^2$ is the squared length, and so on.

B.10.3.1 Orthogonality and Pythagoras's Theorem

h_1 is *orthogonal* to h_2 iff $(h_1, h_2) = 0$. This is written $h_1 \perp h_2$. This is the usual notion of orthogonality in Euclidean space. We then have

Pythagoras's Theorem. *If $h_1 \perp h_2$, then*

$$\|h_1 + h_2\|^2 = \|h_1\|^2 + \|h_2\|^2. \tag{B.10.12}$$

An interesting consequence is the inequality valid for all h_1, h_2,

$$|(h_1, h_2)| \leq \|h_1\|\|h_2\|. \tag{B.10.13}$$

In R^2 (B.10.12) is the familiar "square on the hypotenuse" theorem whereas (B.10.13) says that the cosine between $\mathbf{x}_1$ and $\mathbf{x}_2$ is ≤ 1 in absolute value.

B.10.3.2 Projections on Linear Spaces

We naturally define that a sequence $h_n \in \mathcal{H}$ converges to h iff $\|h_n - h\| \to 0$. A linear subspace $\mathcal{L}$ of $\mathcal{H}$ is *closed* iff $h_n \in \mathcal{L}$ for all n, $h_n \to h \Rightarrow h \in \mathcal{L}$. Given a closed linear subspace $\mathcal{L}$ of $\mathcal{H}$ we define the projection operator $\Pi(\cdot \mid \mathcal{L}) : \mathcal{H} \to \mathcal{L}$ by: $\Pi(h \mid \mathcal{L})$ is that $h' \in \mathcal{L}$ that achieves $\min\{\|h - h'\| : h' \in \mathcal{L}\}$. It may be shown that Π is characterized by the property

$$h - \Pi(h \mid \mathcal{L}) \perp h' \text{ for all } h' \in \mathcal{L}. \tag{B.10.14}$$

Furthermore,

(i) $\Pi(h \mid \mathcal{L})$ exists and is uniquely defined.

(ii) $\Pi(\cdot \mid \mathcal{L})$ is a linear operator

$$\Pi(\alpha h_1 + \beta h_2 \mid \mathcal{L}) = \alpha\Pi(h_1 \mid \mathcal{L}) + \beta\Pi(h_2 \mid \mathcal{L}).$$

(iii) Π is *idempotent*, $\Pi^2 = \Pi$.

(iv) Π is *norm reducing*

$$\|\Pi(h \mid \mathcal{L})\| \leq \|h\|. \tag{B.10.15}$$

In fact, and this follows from (B.10.12),

$$\|h\|^2 = \|\Pi(h \mid \mathcal{L})\|^2 + \|h - \Pi(h \mid \mathcal{L})\|^2. \tag{B.10.16}$$

Here $h - \Pi(h \mid \mathcal{L})$ may be interpreted as a projection on $\mathcal{L}^\perp \equiv \{h : (h, h') = 0 \text{ for all } h' \in \mathcal{L}\}$. Properties (i)–(iii) of Π above are immediate.

All of these correspond to geometric results in Euclidean space. If $\mathbf{x}$ is a vector in R^p, $\Pi(\mathbf{x} \mid \mathcal{L})$ is the point of $\mathcal{L}$ at which the perpendicular to $\mathcal{L}$ from $\mathbf{x}$ meets $\mathcal{L}$. (B.10.16) is Pythagoras's theorem again. If $\mathcal{L}$ is the column space of a matrix $A_{n\times p}$ of rank $p < n$, then

$$\Pi(\mathbf{x} \mid \mathcal{L}) = A[A^T A]^{-1} A^T \mathbf{x}. \tag{B.10.17}$$

This is the formula for obtaining the fitted value vector $\widehat{\mathbf{Y}} = (\widehat{Y}_1, \ldots, \widehat{Y}_n)^T$ by least squares in a linear regression $\mathbf{Y} = A\boldsymbol{\beta} + \boldsymbol{\epsilon}$ and (B.10.16) is the ANOVA identity.

The most important Hilbert space other than R^p is $L_2(P) \equiv \{\text{All random variables } X \text{ on a (separable) probability space such that } EX^2 < \infty\}$. In this case we define the inner product by

$$(X, Y) \equiv E(XY) \tag{B.10.18}$$

so that

$$\|X\| = E^{\frac{1}{2}}(X^2). \tag{B.10.19}$$

All properties needed for this to be a Hilbert space are immediate save for completeness, which is a theorem of F. Riesz. Maintaining our geometric intuition we see that, if $E(X) = E(Y) = 0$, orthogonality simply corresponds to uncorrelatedness and Pythagoras's theorem is just the familiar

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

if X and Y are uncorrelated.

The projection formulation now reveals that what we obtained in Section 1.4 are formulas for projection operators in two situations,

(a) $\mathcal{L}$ is the linear span of $1, Z_1, \ldots, Z_d$. Here

$$\Pi(Y \mid \mathcal{L}) = E(Y) + (\Sigma_{\mathbf{ZZ}}^{-1}\Sigma_{\mathbf{Z}Y})^T(\mathbf{Z} - E(\mathbf{Z})). \tag{B.10.20}$$

This is just (1.4.14).

(b) $\mathcal{L}$ is the space of all $X = g(\mathbf{Z})$ for some g (measurable). This is evidently a linear space that can be shown to be closed. Here,

$$\Pi(Y \mid \mathcal{L}) = E(Y \mid \mathbf{Z}). \tag{B.10.21}$$

That is what (1.4.4) tells us.

The identities and inequalities of Section 1.4 can readily be seen to be special cases of (B.10.16) and (B.10.15).

For a fuller treatment of these introductory aspects of Hilbert space theory, see Halmos (1951), Royden (1968), Rudin (1991), or more extensive works on functional analysis such as Dunford and Schwartz (1964).

B.11 PROBLEMS AND COMPLEMENTS

Problems for Section B.1

1. An urn contains four red and four black balls. Four balls are drawn at random without replacement. Let Z be the number of red balls obtained in the first two draws and Y the total number of red balls drawn.

(a) Find the joint distribution of Z and Y and the conditional distribution of Y given Z and Z given Y.

(b) Find $E(Y \mid Z = z)$ for $z = 0, 1, 2$.

2. Suppose Y and Z have the joint density $p(z, y) = k(k-1)(z-y)^{k-2}$ for $0 < y \leq z < 1$, where $k \geq 2$ is an integer.

(a) Find $E(Y \mid Z)$.

(b) Compute $EY = E(E(Y \mid Z))$ using (a).

3. Suppose Z_1 and Z_2 are independent with exponential $\mathcal{E}(\lambda)$ distributions. Find $E(X \mid Y)$ when $X = Z_1$ and $Y = Z_1 + Z_2$.
Hint: $E(Z_1 + Z_2 \mid Y) = Y$.

4. Suppose Y and Z have joint density function $p(z, y) = z + y$ for $0 < z < 1, 0 < y < 1$.

(a) Find $E(Y \mid Z = z)$.

(b) Find $E(Ye^{[Z+(1/Z)]} \mid Z = z)$.

5. Let $(X_1, \ldots, X_n)$ be a sample from a Poisson $\mathcal{P}(\lambda)$ distribution and let $S_m = \sum_{i=1}^{m} X_i$, $m \leq n$.

(a) Show that the conditional distribution of $\mathbf{X}$ given $S_n = k$ is multinomial $\mathcal{M}(k, 1/n, \ldots, 1/n)$.

(b) Show that $E(S_m \mid S_n) = (m/n)S_n$.

6. A random variable X has a $\mathcal{P}(\lambda)$ distribution. Given $X = k$, Y has a binomial $\mathcal{B}(k, p)$ distribution.

(a) Using the relation $E(e^{tY}) = E(E(e^{tY} \mid X))$ and the uniqueness of moment generating functions show that Y has a $\mathcal{P}(\lambda p)$ distribution.

(b) Show that Y and $X - Y$ are independent and find the conditional distribution of X given $Y = y$.

7. Suppose that X has a normal $\mathcal{N}(\mu, \sigma^2)$ distribution and that $Y = X + Z$, where Z is independent of X and has a $\mathcal{N}(\gamma, \tau^2)$ distribution.

(a) What is the conditional distribution of Y given $X = x$?

(b) Using Bayes Theorem find the conditional distribution of X given $Y = y$.

8. In each of the following examples:

(a) State whether the conditional distribution of Y given $Z = z$ is discrete, continuous, or of neither type.

(b) Give the conditional frequency, density, or distribution function in each case.

(c) Check the identity $E[E(Y \mid Z)] = E(Y)$

(i)

$$\begin{aligned} p_{(Z,Y)}(z, y) &= \frac{1}{\pi}, z^2 + y^2 < 1 \\ &= 0 \text{ otherwise.} \end{aligned}$$

(ii)

$$\begin{aligned} p_{(Z,Y)}(z, y) &= 4zy,\ 0 < z < 1,\ 0 < y < 1 \\ &= 0 \text{ otherwise.} \end{aligned}$$

(iii) Z has a uniform $\mathcal{U}(0, 1)$ distribution, $Y = Z^2$.

(iv) Y has a $\mathcal{U}(-1, 1)$ distribution, $Z = Y^2$.

(v) Y has a $\mathcal{U}(-1, 1)$ distribution, $Z = Y^2$ if $Y^2 < \frac{1}{4}$ and $Z = \frac{1}{4}$ if $Y^2 \geq \frac{1}{4}$.

9. (a) Show that if $E(X^2)$ and $E(Y^2)$ are finite then

$$\text{Cov}(X, Y) = \text{Cov}(X, E(Y \mid X)).$$

(b) Deduce that the random variables Z and Y in Problem B.1.8(i) have correlation 0 although they are not independent.

10. (a) If $X_1, \ldots, X_n$ is a sample from *any* population and $S_m = \sum_{i=1}^m X_i$, $m \leq n$, show that the joint distribution of (X_i, S_m) does not depend on $i, i \leq m$.

Hint: Show that the joint distribution of $(X_1, \ldots, X_n)$ is the same as that of $(X_{i_1}, \ldots, X_{i_n})$ where $(i_1, \ldots, i_n)$ is any permutation of $(1, \ldots, n)$.

(b) Assume that if X and Y are any two random variables, then the family of conditional distributions of X given Y depends only on the joint distribution of (X, Y). Deduce from (a) that $E(X_1 \mid S_n) = \cdots = E(X_n \mid S_n)$ and, hence, that $E(S_m \mid S_n) = (m/n)S_n$.

11. Suppose that Z has a binomial, $\mathcal{B}(N, \theta)$, distribution and that given $Z = z$, Y has a hypergeometric, $\mathcal{H}(z, N, n)$, distribution. Show that

$$P[Z = z \mid Y = y] = \left(\begin{array}{c} N-n \\ z-y \end{array} \right) \theta^{z-y}(1-\theta)^{N-n-(z-y)}$$

(i.e., the binomial probability of $z - y$ successes in $N - n$ trials).

Hint: Divide the numerator and denominator of (B.1.5) by

$$\left(\begin{array}{c} N \\ n \end{array} \right) \left(\begin{array}{c} n \\ y \end{array} \right) \theta^y (1-\theta)^{n-y}.$$

This gives the required binomial probability in the numerator. Because the binomial probabilities add to one, the denominator must be one.

Problems for Section B.2

1. If θ is uniformly distributed on $(-\pi/2, \pi/2)$ show that $Y = \tan\theta$ has a *Cauchy* distribution whose density is given by $p(y) = 1/[\pi(1 + y^2)]$, $-\infty < y < \infty$. Note that this density coincides with the Student t density with one degree of freedom obtainable from (B.3.10).

2. Suppose X_1 and X_2 are independent exponential $\mathcal{E}(\lambda)$ random variables. Let $Y_1 = X_1 - X_2$ and $Y_2 = X_2$.

(a) Find the joint density of Y_1 and Y_2.

(b) Show that Y_1 has density $p(y) = \frac{1}{2}\lambda e^{-\lambda|y|}$, $-\infty < y < \infty$. This is known as the *double exponential* or *Laplace density*.

3. Let X_1 and X_2 be independent with $\beta(r_1, s_1)$ and $\beta(r_2, s_2)$ distributions, respectively. Find the joint density of $Y_1 = X_1$ and $Y_2 = X_2(1 - X_1)$.

4. Show that if X has a gamma $\Gamma(p, \lambda)$ distribution, then

(a) $M_X(t) = E(e^{tX}) = \left(\frac{\lambda}{\lambda - t}\right)^p$, $t < \lambda$.

(b) $E(X^r) = \frac{\Gamma(r+p)}{\lambda^r \Gamma(p)}$, $r > -p$.

(c) $E(X) = p/\lambda$, $\mathrm{Var}(X) = p/\lambda^2$.

5. Show that if X has a beta $\beta(r, s)$ distribution, then

(a) $E(X^k) = \frac{r \ldots (r+(k-1))}{(r+s)\ldots(r+s+(k-1))}$, $k = 1, 2, \ldots$.

(b) $\mathrm{Var}\, X = \frac{rs}{(r+s)^2(r+s+1)}$.

6. Let $V_1, \ldots, V_{n+1}$ be a sample from a population with an exponential $\mathcal{E}(1)$ distribution (see (A.13.24)) and let $S_m = \sum_{i=1}^m V_i$, $m \leq n + 1$.

(a) Show that $\mathbf{T} = \left(\frac{V_1}{S_{n+1}}, \ldots, \frac{V_n}{S_{n+1}}\right)^T$ has a density given by

$$\begin{aligned} p_{\mathbf{T}}(t_1, \ldots, t_n) &= n!,\ t_i > 0,\ 1 \le i \le n,\ \textstyle\sum_{i=1}^n t_i < 1, \\ &= 0 \text{ otherwise.} \end{aligned}$$

Hint: Derive first the joint distribution of $\left(\frac{V_1}{S_{n+1}}, \ldots, \frac{V_n}{S_{n+1}},\ S_{n+1}\right)^T$.

(b) Show that $\mathbf{U} = \left(\frac{S_1}{S_{n+1}}, \ldots, \frac{S_n}{S_{n+1}}\right)^T$ has a density given by

$$\begin{aligned} p_{\mathbf{U}}(u_1, \ldots, u_n) &= n!,\ 0 < u_1 < u_2 < \cdots < u_n < 1, \\ &= 0 \text{ otherwise.} \end{aligned}$$

7. Let $S_1, \ldots, S_r$ be r disjoint open subsets of R^n such that $P[\mathbf{X} \in \cup_{i=1}^r S_i] = 1$. Suppose that $\mathbf{g}$ is a transformation from $\cup_{i=1}^r S_i$ to R^n such that

(i) $\mathbf{g}$ has continuous first partial derivatives in S_i for each i.

(ii) $\mathbf{g}$ is one to one on each S_i.

(iii) The Jacobian of $\mathbf{g}$ does not vanish on each S_i.

Show that if $\mathbf{X}$ has density $p_{\mathbf{X}}$, $\mathbf{Y} = \mathbf{g}(\mathbf{X})$ has density given by

$$p_{\mathbf{Y}}(\mathbf{y}) = \sum_{i=1}^r p_{\mathbf{X}}(\mathbf{g}_i^{-1}(\mathbf{y}))|J_{\mathbf{g}_i}(\mathbf{g}_i^{-1}(\mathbf{y}))|^{-1} I_i(\mathbf{y}) \text{ for } \mathbf{y} \in \mathbf{g}(\cup_{i=1}^r S_i)$$

where $\mathbf{g}_i$ is the restriction of $\mathbf{g}$ to S_i and $I_i(\mathbf{y})$ is 1 if $\mathbf{y} \in \mathbf{g}(S_i)$ and 0 otherwise. (If $I_i(\mathbf{y}) = 0$, the whole summand is taken to be 0 even though $\mathbf{g}_i^{-1}$ is in fact undefined.)

Hint: $P[\mathbf{g}(\mathbf{X}) \in B] = \sum_{i=1}^r P[\mathbf{g}(\mathbf{X}) \in B, \mathbf{X} \in S_i]$.

8. Suppose that $X_1, \ldots, X_n$ is a sample from a population with density f. The X_i arranged in order from smallest to largest are called the *order statistics* and are denoted by $X_{(1)}, \ldots, X_{(n)}$. Show that $\mathbf{Y} = \mathbf{g}(\mathbf{X}) = (X_{(1)}, \ldots, X_{(n)})^T$ has density

$$p_{\mathbf{Y}}(\mathbf{y}) = n! \prod_{i=1}^n f(y_i) \text{ for } y_1 < y_2 < \cdots < y_n$$

Hint: Let

$$\begin{aligned} S_1 &= \{(x_1, \ldots, x_n) : x_1 < \cdots < x_n\}, \\ S_2 &= \{(x_1, \ldots, x_n) : x_2 < x_1 < \cdots < x_n\} \end{aligned}$$

and so on up to $S_{n!}$. Apply the previous problem.

9. Let $X_1, \ldots, X_n$ be a sample from a uniform $\mathcal{U}(0,1)$ distribution (cf. (A.13.29)).

(a) Show that the order statistics of $\mathbf{X} = (X_1, \ldots, X_n)$ have the distribution whose density is given in Problem B.2.6(b).

(b) Deduce that $X_{(k)}$ has a $\beta(k, n-k+1)$ distribution and that $X_{(l)} - X_{(k)} \sim \beta(l-k, n-l+k+1)$.

(c) Show that $EX_{(k)} = k/(n+1)$ and Var $X_{(k)} = k(n-k+1)/(n+1)^2(n+2)$.
Hint: Use Problem B.2.5.

10. Let $X_1, \ldots, X_n$ be a sample from a population with density f and d.f. F.

(a) Show that the conditional density of $(X_{(1)}, \ldots, X_{(r)})^T$ given $(X_{(r+1)}, \ldots, X_{(n)})^T$ is

$$p(x_{(1)}, \ldots, x_{(r)} \mid x_{(r+1)}, \ldots, x_{(n)}) = \frac{r! \prod_{i=1}^{r} f(x_{(i)})}{F^r(x_{(r+1)})}$$

if $x_{(1)} < \cdots < x_{(r)} < x_{(r+1)}$.

(b) Interpret this result.

11. (a) Show that if the population in Problem B.2.10 is $\mathcal{U}(0,1)$, then

$$\left(\frac{X_{(1)}}{X_{(r+1)}}, \ldots, \frac{X_{(r)}}{X_{(r+1)}}\right)^T \text{ and } (X_{(r+1)}, \ldots, X_{(n)})^T \text{ are independent.}$$

(b) Deduce that $X_{(n)}, \ldots, \frac{X_{(n)}}{X_{(n-1)}}, \frac{X_{(n-1)}}{X_{(n-2)}}, \ldots, \frac{X_{(2)}}{X_{(1)}}$ are independent in this case.

12. Let the d.f. F have a density f that is continuous and positive on an interval (a,b) such that $F(b) - F(a) = 1$, $-\infty \le a < b \le \infty$. (The results are in fact valid if we only suppose that F is continuous.)

(a) Show that if X has density f, then $Y = F(X)$ is uniformly distributed on $(0,1)$.

(b) Show that if $U \sim \mathcal{U}(0,1)$, then $F^{-1}(U)$ has density f.

(c) Let $U_{(1)} < \cdots < U_{(n)}$ be the order statistics of a sample of size n from a $\mathcal{U}(0,1)$ population. Show that then $F^{-1}(U_{(1)}) < \cdots < F^{-1}(U_{(n)})$ are distributed as the order statistics of a sample of size n from a population with density f.

13. Using Problems B.2.9(b) and B.2.12 show that if $X_{(k)}$ is the kth order statistic of a sample of size n from a population with density f, then

$$p_{X_{(k)}}(t) = \frac{n!}{(k-1)!(n-k)!} F^{k-1}(t)(1 - F(t))^{n-k} f(t).$$

14. Let $X_{(1)}, \ldots, X_{(n)}$ be the order statistics of a sample of size n from an $\mathcal{E}(1)$ population. Show that $nX_{(1)}$, $(n-1)(X_{(2)} - X_{(1)})$, $(n-2)(X_{(3)} - X_{(2)}), \ldots, (X_{(n)} - X_{(n-1)})$ are independent and identically distributed according to $\mathcal{E}(1)$.
Hint: Apply Theorem B.2.2 directly to the density given by Problem B.2.8.

15. Let T_k be the time of the kth occurrence of an event in a Poisson process as in (A.16.4).

(a) Show that T_k has a $\Gamma(k, \lambda)$ distribution.

(b) From the identity of the events, $[N(1) \leq k-1] = [T_k > 1]$, deduce the identity

$$\int_\lambda^\infty g_{k,1}(s)ds = \sum_{j=0}^{k-1} \frac{\lambda^j}{j!} e^{-\lambda}.$$

Problems for Section B.3

1. Let X and Y be independent and identically distributed $\mathcal{N}(0, \sigma^2)$ random variables.

(a) Show that $X^2 + Y^2$ and $\frac{X}{\sqrt{X^2+Y^2}}$ are independent.

(b) Let $\theta = \sin^{-1} \frac{X}{\sqrt{X^2+Y^2}}$. Show that θ is uniformly distributed on $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$.

(c) Show that X/Y has a Cauchy distribution.
Hint: Use Problem B.2.1.

2. Suppose that $Z \sim \Gamma\left(\frac{1}{2}k, \frac{1}{2}k\right)$, $k > 0$, and that given $Z = z$, the conditional distribution of Y is $\mathcal{N}(0, z^{-1})$. Show that Y has a $\mathcal{T}_k$ distribution. When $k = 1$, this is an example where $E(E(Y \mid Z)) = 0$, while $E(Y)$ does not exist.

3. Show that if $Z_1, \ldots, Z_n$ are as in the statement of Theorem B.3.3, then

$$\sqrt{n}(\bar{Z} - \mu) / \sqrt{\sum_{i=1}^n (Z_i - \bar{Z})^2/(n-1)}$$

has a $\mathcal{T}_{n-1}$ distribution.

4. Show that if $X_1, \ldots, X_n$ are independent $\mathcal{E}(\lambda)$ random variables, then $T = 2\lambda \sum_{i=1}^n X_i$ has a χ^2_{2n} distribution.
Hint: First show that $2\lambda X_i$ has a $\Gamma\left(1, \frac{1}{2}\right) = \chi^2_2$ distribution.

5. Show that if $X_1, \ldots, X_m$; $Y_1, \ldots, Y_n$ are independent $\mathcal{E}(\lambda)$ random variables, then $S = (n/m)\left(\sum_{i=1}^m X_i\right) / \left(\sum_{j=1}^n Y_j\right)$ has a $\mathcal{F}_{2m,2n}$ distribution.

6. Suppose that X_1 and X_2 are independent with $\Gamma(p, 1)$ and $\Gamma\left(p + \frac{1}{2}, 1\right)$ distributions. Show that $Y = 2\sqrt{X_1 X_2}$ has a $\Gamma(2p, 1)$ distribution.

7. Suppose X has density p that is symmetric about 0; that is, $p(x) = p(-x)$ for all x. Show that $E(X^k) = 0$ if k is odd and the kth moment is finite.

8. Let $X \sim \mathcal{N}(\mu, \sigma^2)$.

(a) Show that the rth central moment of X is

$$\begin{aligned} E(X-\mu)^r &= \frac{r!\sigma^r}{2^{\frac{1}{2}r}(r/2)!}, \quad r \text{ even} \\ &= 0, \qquad\qquad\quad r \text{ odd}. \end{aligned}$$

(b) Show the rth cumulant c_r is xero for $r \geq 3$.

Hint: Use Problem B.3.7 for r odd. For r even set $m = r/2$ and note that because $Y = [(X-\mu)/\sigma]^2$ has a χ_1^2 distribution, we can find $E(Y^m)$ from Problem B.2.4. Now use $E(X-\mu)^r = \sigma^r E(Y^m)$.

9. Show that if $X \sim \mathcal{T}_k$, then

$$E(X^r) = \frac{k^{\frac{1}{2}r}\Gamma\left(\frac{1}{2}(1+r)\right)\Gamma\left(\frac{1}{2}(k-r)\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{1}{2}k\right)}$$

for r even and $r < k$. The moments do not exist for $r \geq k$, the odd moments are zero when $r < k$. The mean of X is 0, for $k > 1$, and Var $X = k/(k-2)$ for $k > 2$.

Hint: Using the notation of Section B.3, for r even $E(X^r) = E(Q^r) = k^{\frac{1}{2}r}E(Z^r)E(V^{-\frac{1}{2}r})$, where $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi_k^2$. Now use Problems B.2.4 and B.3.7.

10. Let $X \sim \mathcal{F}_{k,m}$, then

$$E(X^r) = \frac{m^r\Gamma\left(\frac{1}{2}k+r\right)\Gamma\left(\frac{1}{2}m-r\right)}{k^r\Gamma\left(\frac{1}{2}k\right)\Gamma\left(\frac{1}{2}m\right)}$$

provided $-\frac{1}{2}k < r < \frac{1}{2}m$. For other r, $E(X^r)$ does not exist. When $m > 2$, $E(X) = m/(m-2)$, and when $m > 4$,

$$\text{Var } X = \frac{2m^2(k+m-2)}{k(m-2)^2(m-4)}.$$

Hint: Using the notation of Section B.3, $E(X^r) = E(Q^r) = (m/k)^r E(V^r)E(W^{-r})$, where $V \sim \chi_k^2$ and $W \sim \chi_m^2$. Now use Problem B.2.4.

11. Let X have a $\mathcal{N}(\theta, 1)$ distribution.

(a) Show that $Y = X^2$ has density

$$p_Y(y) = \frac{1}{2\sqrt{2\pi y}}e^{-\frac{1}{2}(y+\theta^2)}(e^{\theta\sqrt{y}} + e^{-\theta\sqrt{y}}),\ y > 0.$$

This density corresponds to the distribution known as *noncentral* χ^2 *with* 1 *degree of freedom and noncentrality parameter* θ^2.

(b) Show that we can write

$$p_Y(y) = \sum_{i=0}^{\infty} P(R=i)f_{2i+1}(y)$$

where $R \sim \mathcal{P}\left(\frac{1}{2}\theta^2\right)$ and f_m is the χ_m^2 density. Give a probabilistic interpretation of this formula.

Hint: Use the Taylor expansions for $e^{\theta\sqrt{y}}$ and $e^{-\theta\sqrt{y}}$ in powers of $\sqrt{y}$.

12. Let $X_1, \ldots, X_n$ be independent normal random variables each having variance 1 and $E(X_i) = \theta_i$, $i = 1, \ldots, n$, and let $\theta^2 = \sum_{i=1}^n \theta_i^2$. Show that the density of $V = \sum_{i=1}^n X_i^2$ is given by

$$p_V(v) = \sum_{i=0}^{\infty} P(R = i) f_{2i+n}(v),\ v > 0$$

where $R \sim \mathcal{P}\left(\frac{1}{2}\theta^2\right)$ and f_m is the χ^2_m density. The distribution of V is known as the *noncentral χ^2 with n degrees of freedom and (noncentrality) parameter θ^2. Hint:* Use an orthogonal transformation $\mathbf{Y} = \mathbf{AX}$ such that $Y_1 = \sum_{i=1}^n (\theta_i X_i/\theta)$. Now V has the same distribution as $\sum_{i=1}^n Y_i^2$ where $Y_1, \ldots, Y_n$ are independent with variances 1 and $E(Y_1) = \theta$, $E(Y_i) = 0$, $i = 2, \ldots, n$. Next use Problem B.3.11 and

$$p_V(v) = \int_0^\infty \left[\sum_{i=0}^{\infty} P(R = i) f_{2i+1}(v - s)\right] f_{n-1}(s) ds.$$

13. Let $X_1, \ldots, X_n$ be independent $\mathcal{N}(0, 1)$ random variables and let $V = (X_1 + \theta)^2 + \sum_{i=2}^n X_i^2$. Show that for fixed v and n, $P(V \geq v)$ is a strictly increasing function of θ^2. Note that V has a noncentral χ^2_n distribution with parameter θ^2.

14. Let V and W be independent with $W \sim \chi^2_m$ and V having a noncentral χ^2_k distribution with noncentrality parameter θ^2. Show that $S = (V/k)/(W/m)$ has density

$$p_S(s) = \sum_{i=0}^{\infty} P(R = i) f_{k+2i,m}(s)$$

where $R \sim \mathcal{P}\left(\frac{1}{2}\theta^2\right)$ and $f_{j,m}$ is the density of $\mathcal{F}_{j,m}$. The distribution of S is known as the *noncentral $\mathcal{F}_{k,m}$ distribution with (noncentrality) parameter θ^2.*

15. Let $X_1, \ldots, X_n$ be independent normal random variables with common mean and variance. Define $\bar{X}_{(m)} = (1/m)\sum_{i=1}^m X_i$, and $S_m^2 = \sum_{i=1}^m (X_i - \bar{X}_{(m)})^2$.

(a) Show that

$$S_m^2 = S_{m-1}^2 + \frac{(m-1)}{m}(X_m - \bar{X}_{(m-1)})^2.$$

(b) Let

$$Y_1 = \sqrt{n}\bar{X}_{(n)}, Y_2 = (X_2 - \bar{X}_{(1)})\sqrt{\frac{1}{2}}, Y_3 = (X_3 - \bar{X}_{(2)})\sqrt{\frac{2}{3}}, \ldots,$$
$$Y_n = (X_n - \bar{X}_{(n-1)})\sqrt{\frac{n-1}{n}}.$$

Show that the matrix $\mathbf{A}$ defined by $\mathbf{Y} = \mathbf{AX}$ is orthogonal and, thus, satisfies the requirements of Theorem B.3.2.

(c) Give the joint density of $(\bar{X}_{(n)}, S_2^2, \ldots, S_n^2)^T$.

16. Show that under the assumptions of Theorem B.3.3, $\bar{Z}$ and $(Z_1 - \bar{Z}, \ldots, Z_n - \bar{Z})$ are independent.

Hint: It suffices to show that $\bar{Z}$ is independent of $(Z_2 - \bar{Z}, \ldots, Z_n - \bar{Z})$. This provides another proof that $\bar{Z}$ and $\sum_{i=1}^n (Z_i - \bar{Z})^2$ are independent.

Problems for Section B.4

1. Let $(X, Y) \sim \mathcal{N}\left(1, 1, 4, 1, \frac{1}{2}\right)$. Find

(a) $P(X + 2Y \leq 4)$.

(b) $P(X \leq 2 \mid Y = 1)$.

(c) The joint distribution of $X + 2Y$ and $3Y - 2X$.

Let (X, Y) have a $\mathcal{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ distribution in the problems 2–6, 9 that follow.

2. Let $F(\cdot, \cdot, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ denote the d.f. of (X, Y). Show that

$$\left(\frac{X - \mu_1}{\sigma_1}, \frac{Y - \mu_2}{\sigma_2}\right)$$

has a $\mathcal{N}(0, 0, 1, 1, \rho)$ distribution and, hence, express $F(\cdot, \cdot, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ in terms of $F(\cdot, \cdot, 0, 0, 1, 1, \rho)$.

3. Show that $X + Y$ and $X - Y$ are independent, if and only if, $\sigma_1^2 = \sigma_2^2$.

4. Show that if $\sigma_1 \sigma_2 > 0$, $|\rho| < 1$, then the following expression has a χ_2^2 distribution.

$$\frac{1}{(1 - \rho^2)} \left\{ \frac{(X - \mu_1)^2}{\sigma_1^2} - 2\rho \frac{(X - \mu_1)(Y - \mu_2)}{\sigma_1 \sigma_2} + \frac{(Y - \mu_2)^2}{\sigma_2^2} \right\}$$

Hint: Consider (U_1, U_2) defined by (B.4.19) and (B.4.22).

5. Establish the following relation due to Sheppard.

$$F(0, 0, 0, 0, 1, 1, \rho) = \frac{1}{4} + (1/2\pi) \sin^{-1} \rho.$$

Hint: Let U_1 and U_2 be as defined by (B.4.19) and B.4.22, then

$$\begin{aligned} P[X < 0, Y < 0] &= P[U_1 < 0, \rho U_1 + \sqrt{1 - \rho^2} U_2 < 0] \\ &= P\left[U_1 < 0, \frac{U_2}{U_1} > \frac{-\rho}{\sqrt{1 - \rho^2}}\right]. \end{aligned}$$

6. *The geometry of the bivariate normal surface.*

(a) Let $S_c = \{(x, y) : p_{(X,Y)}(x, y) = c\}$. Suppose that $\sigma_1^2 = \sigma_2^2$. Show that $\{S_c; c > 0\}$ is a family of ellipses centered at (μ_1, μ_2) with common major axis given by $(y - \mu_2) =$

$(x-\mu_1)$ if $\rho>0$, $(y-\mu_2)=-(x-\mu_1)$ if $\rho<0$. If $\rho=0$, $\{S_c\}$ is a family of concentric circles.

(b) If $x=c$, $p_{\mathbf{X}}(c,y)$ is proportional to a normal density as a function of y. That is, sections of the surface $z=p_{\mathbf{X}}(x,y)$ by planes parallel to the (y,z) plane are proportional to Gaussian (normal) densities. This is in fact true for sections by any plane perpendicular to the (x,y) plane.

(c) Show that the tangents to S_c at the two points where the line $y=\mu_2+\rho(\sigma_2/\sigma_1)(x-\mu_1)$ intersects S_c are vertical. See Figure B.4.2.

7. Let $(X_1,Y_1),\ldots,(X_n,Y_n)$ be a sample from a $\mathcal{N}(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)=\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ distribution. Let $\bar{X}=(1/n)\sum_{i=1}^n X_i$, $\bar{Y}=(1/n)\sum_{i=1}^n Y_i$, $S_1^2=\sum_{i=1}^n(X_i-\bar{X})^2$, $S_2^2=\sum_{i=1}^n(Y_i-\bar{Y})^2$, $S_{12}=\sum_{i=1}^n(X_i-\bar{X})(Y_i-\bar{Y})$.

(a) Show that $n(\bar{X}-\mu_1,\bar{Y}-\mu_2)^T\boldsymbol{\Sigma}^{-1}(\bar{X}-\mu_1,\bar{Y}-\mu_2)$ has a χ^2_2 distribution.

(b) Show that $(\bar{X},\bar{Y})$ and (S_1^2,S_2^2,S_{12}) are independent.

Hint: (a): See Problem B.4.4.

(b): Let $\mathbf{A}$ be an orthogonal matrix whose first row is $\left(n^{-\frac{1}{2}},\ldots,n^{-\frac{1}{2}}\right)$. Let $\mathbf{U}=\mathbf{AX}$ and $\mathbf{V}=\mathbf{AY}$, where $\mathbf{X}=(X_1,\ldots,X_n)^T$ and $\mathbf{Y}=(Y_1,\ldots,Y_n)^T$. Show that $(U_2,V_2),\ldots,(U_n,V_n)$ form a sample from a $\mathcal{N}(0,0,\sigma_1^2,\sigma_2^2,\rho)$ population. Note that $S_1^2=\sum_{i=2}^n U_i^2$, $S_2^2=\sum_{i=2}^n V_i^2$, $S_{12}=\sum_{i=2}^n U_iV_i$, while $\bar{X}=U_1/\sqrt{n}$, $\bar{Y}=V_1/\sqrt{n}$.

8. In the model of Problem B.4.7 let $R=S_{12}/S_1S_2$ and

$$T=\frac{\sqrt{(n-2)}R}{\sqrt{1-R^2}}.$$

(a) Show that when $\rho=0$, T has a $\mathcal{T}_{n-2}$ distribution.

(b) Find the density of R if $\rho=0$.

Hint: Without loss of generality, take $\sigma_1^2=\sigma_2^2=1$. Let $\mathbf{C}$ be an $(n-1)\times(n-1)$ orthogonal matrix whose first row is $(U_2,\ldots,U_n)/S_1$. Define $(W_2,\ldots,W_n)^T=\mathbf{C}(V_2,\ldots,V_n)^T$ and show that T can be written in the form $T=L/M$ where $L=S_{12}/S_1=W_2$ and $M^2=(S_1^2S_2^2-S_{12}^2)/(n-2)S_1^2=\sum_{i=3}^n W_i^2/(n-2)$. Argue that given $U_2=u_2,\ldots,U_n=u_n$, no matter what $u_2,\ldots,u_n$ are, T has a $\mathcal{T}_{n-2}$ distribution. Now use the continuous version of (B.1.24).

9. Show that the conditional distribution of $aX+bY$ given $cX+dY=t$ is normal.

Hint: Without loss of generality take $a=d=1$, $b=c=0$ because $(aX+bY,cX+dY)$ also has a bivariate normal distribution. Deal directly with the cases $\sigma_1\sigma_2=0$ and $|\rho|=1$.

10. Let p_1 denote the $\mathcal{N}(0,0,1,1,0)$ density and let p_2 be the $\mathcal{N}(0,0,1,1,\rho)$ density. Suppose that (X,Y) have the joint density

$$p(x,y)=\frac{1}{2}p_1(x,y)+\frac{1}{2}p_2(x,y).$$

Show that X and Y have normal marginal densities, but that the joint density is normal, if and only if, $\rho = 0$.

11. Use a construction similar to that of Problem B.4.10 to obtain a pair of random variables (X, Y) that

(i) have marginal normal distributions.

(ii) are uncorrelated.

(iii) are *not* independent.

Do these variables have a bivariate normal distribution?

Problems for Section B.5

1. Establish (B.5.10) and (B.5.11).

2. Let $\mathbf{a}_{k\times 1}$ and $\mathbf{B}_{k\times k}$ be nonrandom. Show that

$$M_{\mathbf{a}+\mathbf{BU}}(\mathbf{t}) = \exp\{\mathbf{a}^T\mathbf{t}\}M_U(\mathbf{B}^T\mathbf{t})$$

and

$$K_{\mathbf{a}+\mathbf{BU}}(t) = \mathbf{a}^T\mathbf{t} + K_{\mathbf{U}}(\mathbf{B}^T\mathbf{t}).$$

3. Show that if $M_{\mathbf{U}}(\mathbf{t})$ is well defined in a neighborhood of zero then

$$M_{\mathbf{U}}(\mathbf{t}) = 1 + \sum_{p=1}^{\infty} \frac{1}{p!}\mu_{i_1\ldots i_k} t_1^{i_1}\cdots t_k^{i_k}$$

where $\mu_{i_1\cdots i_k} = E(U_1^{i_1}\cdots U_k^{i_k})$ and the sum is over all $(i_1,\ldots,i_k)$ with $i_j \geq 0$, $\sum_{j=1}^k i_j = p$, $p = 1, 2, \ldots$. Moreover,

$$K_{\mathbf{U}}(\mathbf{t}) = \sum_{p=1}^{\infty} \frac{1}{p!} c_{i_1\cdots i_k} t_1^{i_1}\ldots t_k^{i_k}.$$

That is, the Taylor series for K_U converges in a neighborhood of zero.

4. Show that the second- and higher-degree cumulants (where $p = \sum_{j=1}^k i_j \geq 2$) are invariant under shift; thus, they depend only on the moments about the mean.

5. Establish (B.5.16)–(B.5.19).

6. In the bivariate case write $\boldsymbol{\mu} = E(\mathbf{U})$, $\sigma_{ij} = E(U_1-\mu_1)^i(U_2-\mu_2)^j$, $\sigma_1^2 = \sigma_{20}$, $\sigma_2^2 = \sigma_{02}$. Show that

$$(c_{10}, c_{01}, c_{20}, c_{02}, c_{11}, c_{30}, c_{03}, c_{21}, c_{12}) = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \sigma_{11}, \sigma_{30}, \sigma_{03}, \sigma_{21}, \sigma_{12})$$

and

$$(c_{40}, c_{04}, c_{22}, c_{31}, c_{13})$$
$$= (\sigma_{40} - 3\sigma_1^2, \sigma_{04} - 3\sigma_2^2, \sigma_{22} - \sigma_1^2\sigma_2^2 - 2\sigma_{11}^2, \sigma_{31} - 3\sigma_1^2\sigma_{11}, \sigma_{13} - 3\sigma_2^2\sigma_{11}).$$

7. Suppose V, W, and Z are independent and that $U_1 = Z + V$ and $U_2 = Z + W$. Show that

$$M_{\mathbf{U}}(\mathbf{t}) = M_V(t_1)M_W(t_2)M_Z(t_1 + t_2)$$
$$K_{\mathbf{U}}(\mathbf{t}) = K_V(t_1) + K_W(t_2) + K_Z(t_1 + t_2)$$

and show that $c_{ij}(\mathbf{U}) = c_{i+j}(Z)$ for $i \neq j$; $i, j > 0$.

8. (The bivariate log normal distribution). Suppose $\mathbf{U} = (U_1, U_2)^T$ has a bivariate $\mathcal{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ distribution. Then $\mathbf{Y} = (Y_1, Y_2)^T = (e^{U_1}, e^{U_2})^T$ is said to have a *bivariate log normal distribution*. Show that

$$E(Y_1^i Y_2^j) = \exp\left\{i\mu_1 + j\mu_2 + \frac{1}{2}i^2\sigma_1^2 + ij\sigma_{11} + \frac{1}{2}j^2\sigma_2^2\right\}$$

where $\sigma_{11} = \sigma_1\sigma_2\rho$.

9. (a) Suppose $\mathbf{Z}$ is $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Show that all cumulants of degree higher than 2 (where $p = \sum_{j=1}^k i_j > 2$) are zero.

(b) Suppose $\mathbf{U}_1, \ldots, \mathbf{U}_n$ are i.i.d. as $\mathbf{U}$. Let $\mathbf{Z}_n = n^{-\frac{1}{2}} \sum_{i=1}^n (\mathbf{U}_i - \boldsymbol{\mu})$. Show that $K_{\mathbf{Z}_n}(\mathbf{t}) = nK_{\mathbf{U}}(n^{-\frac{1}{2}}\mathbf{t}) - n^{\frac{1}{2}}\boldsymbol{\mu}^T\mathbf{t}$ and that all cumulants of degree higher than 2 tend to zero as $n \to \infty$.

10. Suppose $\mathbf{U}_{k\times 1}$ and $\mathbf{V}_{m\times 1}$ are independent and $\mathbf{Z}_{(k+m)\times 1} = (\mathbf{U}^T, \mathbf{V}^T)^T$. Let $C_{I,J}$ where $I = \{i_1, \ldots, i_k\}$ and $J = \{i_{k+1}, \ldots, i_{k+m}\}$ be a cumulant of $\mathbf{Z}$. Show that $C_{I,J} \neq 0$ unless either $I = \{0, \ldots, 0\}$ or $J = \{0, \ldots, 0\}$.

Problems for Section B.6

1. (a) Suppose $U_i = \mu + \alpha Z_i + \beta Z_{i-1}$, $i = 1, \ldots, k$, where $Z_0, \ldots, Z_k$ are independent $\mathcal{N}(0, \sigma^2)$ random variables. Compute the expectation and covariance matrix of $\mathbf{U} = (U_1, \ldots, U_k)$. Is $\mathbf{U}$ k-variate normal?

(b) Perform the same operation and answer the same question for $\bar{U}_i$ defined as follows:

$$\bar{U}_1 = Z_1, \bar{U}_2 = Z_2 + \alpha\bar{U}_1, \bar{U}_3 = Z_3 + \alpha\bar{U}_2, \ldots, \bar{U}_k = Z_k + \alpha\bar{U}_{k-1}.$$

2. Let $\mathbf{U}$ be as in Definition B.6.1. Show that if $\boldsymbol{\Sigma}$ is not positive definite, then $\mathbf{U}$ does not have a density.

3. Suppose $\mathbf{U}_{k\times 1}$ has positive definite variance $\boldsymbol{\Sigma}$. Let $U^{(1)}_{l\times 1}$ and $U^{(2)}_{(k-l)\times 1}$ be a partition of $\mathbf{U}$ with variances $\boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{22}$ and covariance $\boldsymbol{\Sigma}_{12} = \mathrm{Cov}(\mathbf{U}^{(1)}, \mathbf{U}^{(2)})_{l\times(k-l)}$. Show that

$$\mathrm{Cov}(\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{U}^{(2)}, \mathbf{U}^{(1)} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{U}^{(2)}) = 0.$$

Problems for Section B.7

1. Prove Theorem B.7.3 for $d = 1$ when Z an Z_n have continuous distribution functions F and F_n.

Hint: Let U denote a uniform, $\mathcal{U}(0,1)$, random variable. For any d.f. G define the left inverse by $G^{-1}(u) = \inf\{t : G(t) \geq u\}$. Now define $Z_n^* = F_n^{-1}(U)$ and $Z^* = F^{-1}(U)$.

2. Prove Proposition B.7.1(a).

3. Establish (B.7.8).

4. Show that if $\mathbf{Z}_n \xrightarrow{\mathcal{L}} \mathbf{z}_0$, then $P(|\mathbf{Z}_n - \mathbf{z}_0| \geq \epsilon) \to P(|\mathbf{Z} - \mathbf{z}_0| \geq \epsilon)$.
Hint: Extend (A.14.5).

5. The L_p *norm* of a random vector $\mathbf{X}$ is defined by $|\mathbf{X}|_p = \{E|\mathbf{X}|^p\}^{\frac{1}{p}}$, $p \geq 1$. The sequence of random variables $\{\mathbf{Z}_n\}$ is said to *converge to* $\mathbf{Z}$ *in* L_p *norm* if $|\mathbf{Z}_n - \mathbf{Z}|_p \to 0$ as $n \to \infty$. We write $\mathbf{Z}_n \xrightarrow{L_p} \mathbf{Z}$. Show that

(a) if $p < q$, then $\mathbf{Z}_n \xrightarrow{L_q} \mathbf{Z} \Rightarrow \mathbf{Z}_n \xrightarrow{L_p} \mathbf{Z}$.
Hint: Use Jensen's inequality B.9.3.

(b) if $\mathbf{Z}_n \xrightarrow{L_p} \mathbf{Z}$, then $\mathbf{Z}_n \xrightarrow{P} \mathbf{Z}$.
Hint:

$$E|\mathbf{Z}_n - \mathbf{Z}|^p \geq E[|\mathbf{Z}_n - \mathbf{Z}|^p 1\{|\mathbf{Z}_n - \mathbf{Z}| \geq \epsilon\}] \geq \epsilon^p P(|\mathbf{Z}_n - \mathbf{Z}| \geq \epsilon).$$

6. Show that $|\mathbf{Z}_n - \mathbf{Z}| \xrightarrow{P} 0$ is equivalent to $Z_{nj} \xrightarrow{P} Z_j$ for $1 \leq j \leq d$.
Hint: Use (B.7.3) and note that $|Z_{nj} - Z_j|^2 \leq |\mathbf{Z}_n - \mathbf{Z}|^2$.

7. Let $U \sim \mathcal{U}(0,1)$ and let $U_1 = 1$, $U_2 = 1\{U \in [0, \frac{1}{2})\}$, $U_3 = 1\{U \in [\frac{1}{2}, 1)\}$, $U_4 = 1\{U \in [0, \frac{1}{4})\}$, $U_5 = 1\{U \in [\frac{1}{4}, \frac{1}{2})\}, \ldots, U_n = 1\{U \in [m2^{-k}, (m+1)2^{-k})\}$, where $n = m + 2^k$, $0 \leq m \leq 2^k$ and $k \geq 0$. Show that $U_n \xrightarrow{P} 0$ but $U_n \overset{a.s.}{\not\to} 0$.

8. Let $U \sim \mathcal{U}(0,1)$ and set $U_n = 2^n 1\{U \in [0, \frac{1}{n})\}$. Show that $U_n \overset{a.s.}{\to} 0$, $U_n \xrightarrow{P} 0$, but $U_n \overset{L_p}{\not\to} 0$, $p \geq 1$, where L_p is defined in Problem B.7.5.

9. Establish B.7.9.

10. Show that Theorem B.7.5 implies Theorem B.7.4.

11. Suppose that as in Theorem B.7.7, $F_n(x) \to F(x)$ for all x, F is continuous, and strictly increasing so that $F^{-1}(\alpha)$ is unique for all $0 < \alpha < 1$. Show that

$$\sup\{|F_n^{-1}(\alpha) - F^{-1}(\alpha)| : \epsilon \leq \alpha \leq 1 - \epsilon\} \to 0$$

for all $\epsilon > 0$. Here $F_n^{-1}(\alpha) = \inf\{x : F_n(x) \geq \alpha\}$.
Hint: Argue by contradiction.

Problems for Section B.8

1. If $h : R^d \to R^p$ and $\dot{\mathbf{h}}(\mathbf{x}) = \mathbf{D}\mathbf{h}(\mathbf{x})$ is continuous in a sphere $\{\mathbf{x} : |\mathbf{x} - \mathbf{x}_0| < \delta\}$, then for $|\mathbf{z}| < \delta$

$$\mathbf{h}(\mathbf{x}_0 + \mathbf{z}) = \mathbf{h}(\mathbf{x}_0) + \left(\int_0^1 \dot{\mathbf{h}}(x_0 + u\mathbf{z})\mathbf{z}du \right) \mathbf{z}^T.$$

Here the integral is a $p \times d$ matrix of integrals.

Hint: Let $\mathbf{g}(u) = \mathbf{h}(\mathbf{x}_0 + u\mathbf{z})$. Then by the chain rule, $\dot{\mathbf{g}}(u) = \dot{\mathbf{h}}(\mathbf{x}_0 + u\mathbf{z})\mathbf{z}$ and

$$\int_0^1 \dot{\mathbf{h}}(\mathbf{x}_0 + u\mathbf{z})\mathbf{z}du = \int_0^1 \dot{\mathbf{g}}(u)du = \mathbf{g}(1) - \mathbf{g}(0) = \mathbf{h}(\mathbf{x}_0 + \mathbf{z}) - \mathbf{h}(\mathbf{x}_0).$$

2. If $h : R^d \to R$ and $\ddot{h}(\mathbf{x}) = D^2 h(\mathbf{x})$ is continuous in a sphere $\{\mathbf{x} : |\mathbf{x} - \mathbf{x}_0| < \delta\}$, then for $|\mathbf{z}| < \delta$,

$$h(\mathbf{x}_0 + \mathbf{z}) = h(\mathbf{x}_0) + \dot{h}(\mathbf{x}_0)\mathbf{z} + \mathbf{z}^T \left[\int_0^1 \int_0^1 \ddot{h}(\mathbf{x}_0 + uv\mathbf{z})vdudv \right] \mathbf{z}.$$

Hint: Apply Problem B.8.1 to $h(\mathbf{x}_0 + \mathbf{z}) - h(\mathbf{x}_0) - \dot{h}(\mathbf{x}_0)\mathbf{z}$.

3. Apply Problems B.8.1 and B.8.2 to obtain special cases of Taylor's Theorem B.8.12.

Problems for Section B.9

1. State and prove Jensen's inequality for conditional expectations.

2. Use Hoeffding's inequality (B.9.6) to establish Bernstein's inequality (B.9.5). Show that if $p = \frac{1}{2}$, the bound can be improved to $2\exp\{-2n/\epsilon^2\}$.

3. Derive Hölder's inequality from Jensen's inequality with $k = 2$.

Hint: For $(x, y) \in R^2$, consider $g(x, y) = \frac{|x|^r}{r} + \frac{|y|^s}{s}$, $\frac{1}{r} + \frac{1}{s} = 1$.

4. Show that if $k = 1$ and $g''(x)$ exists, then $g''(x) \geq 0$, all $\mathbf{x} \in S$, and convexity are equivalent.

5. Show that convexity is equivalent to the convexity of $g(\alpha\mathbf{x} + (1 - \alpha)\mathbf{y}))$ as function of $\alpha \in [0, 1]$ for all $\mathbf{x}$ and $\mathbf{y}$ in S.

6. Use Problem 5 above to generalize Problem 4 above to the case $k > 1$.

7. Show that if $\frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j}g^2(\mathbf{x})$ exists and the matrix $\left|\frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j}g^2(\mathbf{x})\right|$ is positive definite, then g is strictly convex.

8. Show that

$$P(X \geq a) \leq \inf\{e^{-ta}Ee^{tX} : t \geq 0\}.$$

Hint: Use inequality (A.15.4).

9. Use Problem 8 above to prove Bernstein's inequality.

10. Show that the sum of (strictly) convex functions is (strictly) convex.

Problems for Section B.10

1. Verify that if A is snd, then A is ppd iff A is nonsingular.

2. Show that if S is the one-dimensional space $S = \{a\mathbf{e} : a \in R\}$ for $\mathbf{e}$ orthonormal, then the projection matrix onto S (B.10.17) is just $\mathbf{e}\mathbf{e}^T$.

3. Establish (B.10.15) and (B.10.16).

4. Show that $h - \Pi(h \mid \mathcal{L}) = \Pi(h \mid \mathcal{L}^\perp)$ using (B.10.14).

5. Establish (B.10.17).

B.12 NOTES

Notes for Section B.1.2

(1) We shall follow the convention of also calling $E(Y \mid \mathbf{Z})$ any variable that is equal to $g(\mathbf{Z})$ with probability 1.

Notes for Section B.1.3

(1) The definition of the conditional density (B.1.25) can be motivated as follows: Suppose that $A(\mathbf{x})$, $A(\mathbf{y})$ are small "cubes" with centers $\mathbf{x}$ and $\mathbf{y}$ and volumes $d\mathbf{x}$, $d\mathbf{y}$ and $p(\mathbf{x}, \mathbf{y})$ is continuous. Then $P[\mathbf{X} \in A(\mathbf{x}) \mid \mathbf{Y} \in A(\mathbf{y})] = P[\mathbf{X} \in A(\mathbf{x}),\ \mathbf{Y} \in A(\mathbf{y})]/P[\mathbf{Y} \in A(\mathbf{y})]$. But $P[\mathbf{X} \in A(\mathbf{x}),\ \mathbf{Y} \in A(\mathbf{y})] \approx p(\mathbf{x}, \mathbf{y})d\mathbf{x}\, d\mathbf{y}$, $P[\mathbf{Y} \in A(\mathbf{y})] \approx p_Y(\mathbf{y})d\mathbf{y}$, and it is reasonable that we should have $p(\mathbf{x} \mid \mathbf{y}) \approx P[\mathbf{X} \in A(\mathbf{x}) \mid \mathbf{Y} \in A(\mathbf{y})]/d\mathbf{x} \approx p(\mathbf{x}, \mathbf{y})/p_{\mathbf{Y}}(\mathbf{y})$.

Notes for Section B.2

(1) We do not dwell on the stated conditions of the transformation Theorem B.2.1 because the conditions are too restrictive. It may, however, be shown that (B.2.1) continues to hold even if f is assumed only to be absolutely integrable in the sense of Lebesgue and K is any member of $\mathcal{B}^k$, the Borel σ-field on R^k. Thus, f can be any density function and K any set in R^k that one commonly encounters.

Notes for Section B.3.2

(1) In deriving (B.3.15) and (B.3.17) we are using the standard relations, $[\mathbf{AB}]^T = \mathbf{B}^T\mathbf{A}^T$, $\det[\mathbf{AB}] = \det \mathbf{A} \det \mathbf{B}$, and $\det \mathbf{A} = \det \mathbf{A}^T$.

Notes for Section B.5

(1) Both m.g.f.'s and c.f.'s are special cases of the Laplace transform ψ of the distribution of $\mathbf{U}$ defined by

$$\psi(\mathbf{z}) = E(e^{\mathbf{z}^T\mathbf{U}}),$$

where $\mathbf{z}$ is in the set of k tuples of complex numbers.

B.13 REFERENCES

ANDERSON, T. W., *An Introduction to Multivariate Statistical Analysis* New York: J. Wiley & Sons, 1958.

APOSTOL, T., *Mathematical Analysis*, 2nd ed. Reading, MA: Addison–Wesley,1974.

BARNDORFF–NIELSEN, O. E., AND D. P. COX, *Asymptotic Techniques for Use in Statistics* New York: Chapman and Hall, 1989.

BILLINGSLEY, P., *Probability and Measure*, 3rd ed. New York: J. Wiley & Sons, 1979, 1995.

BIRKHOFF, G., AND S. MacLANE, *A Survey of Modern Algebra*, rev. ed. New York: Macmillan, 1953.

BIRKHOFF, G., AND S. MacLANE, *A Survey of Modern Algebra*, 3rd ed. New York: MacMillan, 1965.

BREIMAN, L., *Probability* Reading, MA: Addison–Wesley, 1968.

CHUNG, K. L., *A Course in Probability Theory* New York: Academic Press, 1974.

DEMPSTER, A. P., *Elements of Continuous Multivariate Analysis* Reading, MA: Addison–Wesley, 1969.

DIEUDONNÉ, J., *Foundation of Modern Analysis*, v. 1, Pure and Applied Math. Series, Volume 10 New York: Academic Press, 1960.

DUNFORD, N., AND J. T. SCHWARTZ, *Linear Operators*, Volume 1, *Interscience* New York: J. Wiley & Sons, 1964.

FELLER, W., *An Introduction to Probability Theory and Its Applications*, Vol. II, 2nd ed. New York: J. Wiley & Sons, 1971.

GRIMMETT, G. R., AND D. R. STIRSAKER, *Probability and Random Processes* Oxford: Clarendon Press, 1992.

HALMOS, P. R., *An Introduction to Hilbert Space and the Theory of Spectral Multiplicity*, 2nd ed. New York: Chelsea, 1951.

HAMMERSLEY, J., "An extension of the Slutsky–Fréchet theorem," *Acta Mathematica, 87*, 243–247 (1952).

HOEFFDING, W., "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Assoc., 58*, 13–30 (1963).

LOÈVE, M., *Probability Theory*, Vol. 1, 4th ed. Berlin: Springer, 1977.

RAO, C. R., *Linear Statistical Inference and Its Applications*, 2nd ed. New York: J. Wiley & Sons, 1973.

ROCKAFELLAR, R. T., *Convex Analysis* Princeton, NJ: Princeton University Press, 1970.

ROYDEN, H. L., *Real Analysis*, 2nd ed. New York: MacMillan, 1968.

RUDIN, W., *Functional Analysis*, 2nd ed. New York: McGraw–Hill, 1991.

SKOROKHOD, A. V., "Limit theorems for stochastic processes," *Th. Prob. Applic., 1*, 261–290 (1956).

This page intentionally left blank