CHAPTER 7

Sample Moments and Their Distributions

7.1 INTRODUCTION

In the preceding chapters we discussed fundamental ideas and techniques of probability theory. In this development we created a mathematical model of a random experiment by associating with it a sample space in which random events correspond to sets of a certain σ-field. The notion of probability defined on this σ-field corresponds to the notion of uncertainty in the outcome on any performance of the random experiment.

In this chapter we begin the study of some problems of mathematical statistics. The methods of probability theory learned in preceding chapters are used extensively in this study. Suppose that we seek information about some numerical characteristics of a collection of elements, called a *population*. For reasons of time or cost we may not wish or be able to study each element of the population. Our object is to draw conclusions about the unknown population characteristics on the basis of information on some characteristics of a suitably selected *sample*. Formally, let X be a random variable that describes the population under investigation, and let F be the DF of X. There are two possibilities. Either X has a DF F_θ with a known functional form (except perhaps for the parameter θ, which may be a vector), or X has a DF F about which we know nothing (except perhaps that F is, say, absolutely continuous). In the former case let Θ be the set of possible values of the unknown parameter θ. Then the job of a statistician is to decide on the basis of a suitably selected sample which member or members of the family $\{F_\theta, \theta \in \Theta\}$ can represent the DF of X. Problems of this type, called problems of *parametric statistical inference*, are the subject of investigation in Chapters 8 through 12. The case in which nothing is known about the functional form of the DF F of X is clearly much more difficult. Inference problems of this type fall into the domain of *nonparametric statistics* and are discussed in Chapter 13.

To be sure, the scope of statistical methods is much wider than the statistical inference problems discussed in this book. Statisticians, for example, deal with problems

of planning and designing experiments, of collecting information, and of deciding how best the collected information should be used. However, here we concern ourselves only with the best methods of making inferences about probability distributions.

In Section 7.2 we introduce the notions of a (simple) *random sample* and *sample statistics*. In Section 7.3 we study sample moments and their exact distributions, and in Section 7.5 we consider their large-sample approximations. In Section 7.4 we consider some important distributions that arise in sampling from a normal population. Sections 7.6 and 7.7 are devoted to the study of sampling from univariate and bivariate normal distributions.

7.2 RANDOM SAMPLING

Consider a statistical experiment that culminates in outcomes x, which are the values assumed by an RV X. Let F be the DF of X. In practice, F will not be completely known; that is, one or more parameters associated with F will be unknown. The job of a statistician is to estimate these unknown parameters or to test the validity of certain statements about them. She can obtain n independent observations on X. This means that she observes n values $x_1, x_2, \ldots, x_n$ assumed by the RV X. Each x_i can be regarded as the value assumed by an RV X_i, $i = 1, 2, \ldots, n$, where $X_1, X_2, \ldots, X_n$ are independent RVs with common DF F. The observed values $(x_1, x_2, \ldots, x_n)$ are then values assumed by $(X_1, X_2, \ldots, X_n)$. The set $\{X_1, X_2, \ldots, X_n\}$ is then a *sample* of size n taken from a *population distribution* F. The set of n values $x_1, x_2, \ldots, x_n$ is called a *realization* of the sample. Note that the possible values of the RV $(X_1, X_2, \ldots, X_n)$ can be regarded as points in $\mathcal{R}_n$, which may be called the *sample space*. In practice one observes not $x_1, x_2, \ldots, x_n$ but some function $f(x_1, x_2, \ldots, x_n)$. Then $f(x_1, x_2, \ldots, x_n)$ are values assumed by the RV $f(X_1, X_2, \ldots, X_n)$.

Let us now formalize these concepts.

Definition 1. Let X be an RV with DF F, and let $X_1, X_2, \ldots, X_n$ be iid RVs with common DF F. Then the collection $X_1, X_2, \ldots, X_n$ is known as a *random sample* of size n from the DF F or simply as n independent observations on X.

If $X_1, X_2, \ldots, X_n$ is a random sample from F, their joint DF is given by

$$F^*(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} F(x_i). \tag{1}$$

Definition 2. Let $X_1, X_2, \ldots, X_n$ be n independent observations on an RV X, and let $f: \mathcal{R}_n \to \mathcal{R}_k$ be a Borel-measurable function. Then the RV $f(X_1, X_2, \ldots, X_n)$ is called a (*sample*) *statistic* provided that it is not a function of any unknown parameter(s).

Two of the most commonly used statistics are defined as follows.

Definition 3. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution function F. Then the statistic

$$\overline{X} = n^{-1}S_n = \sum_{i=1}^{n} \frac{X_i}{n} \tag{2}$$

is called the *sample mean*, and the statistic

$$S^2 = \sum_{1}^{n} \frac{(X_i - \overline{X})^2}{n-1} = \frac{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}{n-1} \tag{3}$$

is called the *sample variance*, and S is called the *sample standard deviation*.

Remark 1. Whenever the word *sample* is used subsequently, it will mean *random sample*.

Remark 2. Sampling from a probability distribution (Definition 1) is sometimes referred to as *sampling from an infinite population* since one can obtain samples of any size one desires even if the population is finite (by sampling *with replacement*).

Remark 3. In sampling *without replacement* from a finite population, the independence condition of Definition 1 is not satisfied. Suppose that a sample of size 2 is taken from a finite population $(a_1, a_2, \ldots, a_N)$ without replacement. Let X_i be the outcome on the ith draw. Then $P\{X_1 = a_1\} = 1/N$, $P\{X_2 = a_2 \mid X_1 = a_1\} = 1/(N-1)$, and $P\{X_2 = a_2 \mid X_1 = a_2\} = 0$. Thus the PMF of X_2 depends on the outcome of the first draw (that is, on the value of X_1), and X_1 and X_2 are not independent. Note, however, that

$$\begin{aligned} P\{X_2 = a_2\} &= \sum_{j=1}^{N} P\{X_1 = a_j\}P\{X_2 = a_2 \mid a_j\} \\ &= \sum_{j \neq 2} P\{X_1 = a_j\}P\{X_2 = a_2 \mid a_j\} = \frac{1}{N}, \end{aligned}$$

and $X_1 \overset{d}{=} X_2$. A similar argument can be used to show that $X_1, X_2, \ldots, X_n$ all have the same distribution but they are not independent. In fact, $X_1, X_2, \ldots, X_n$ are exchangeable RVs. Sampling without replacement from a finite population is often referred to as *simple random sampling*.

Remark 4. It should be remembered that sample statistics $\overline{X}$, S^2 (and others that we will define later) are random variables, while the population parameters μ, σ^2, and so on, are fixed constants that may be unknown.

Remark 5. In (3) we divide by $n - 1$ rather than n. The reason for this will become clear in the next section.

Remark 6. Other frequently occurring examples of statistics are sample order statistics $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ and their functions, as well as sample moments, which will be studied in the next section.

Example 1. Let $X \sim b(1, p)$, where p is possibly unknown. The DF of X is given by

$$F(x) = p\varepsilon(x - 1) + (1 - p)\varepsilon(x), \qquad x \in \mathcal{R}.$$

Suppose that five independent observations on X are 0, 1, 1, 1, 0. Then 0, 1, 1, 1, 0 is a realization of the sample $X_1, X_2, \ldots, X_5$. The sample mean is

$$\bar{x} = \frac{0 + 1 + 1 + 1 + 0}{5} = 0.6,$$

which is the value assumed by the RV $\overline{X}$. The sample variance is

$$s^2 = \sum_{i=1}^{5} \frac{(x_i - \bar{x})^2}{5 - 1} = \frac{2(0.6)^2 + 3(0.4)^2}{4} = 0.3,$$

which is the value assumed by the RV S^2. Also $s = \sqrt{0.3} = 0.55$.

Example 2. Let $X \sim \mathcal{N}(\mu, \sigma^2)$, where μ is known but σ^2 is unknown. Let $X_1, X_2, \ldots, X_n$ be a sample from $\mathcal{N}(\mu, \sigma^2)$. Then, according to our definition, $\sum_{i=1}^{n} X_i/\sigma^2$ is not a statistic.

Suppose that five observations on X are -0.864, 0.561, 2.355, 0.582, -0.774. Then the sample mean is 0.372, and the sample variance is 1.648.

PROBLEMS 7.2

1. Let X be a $b(1, \frac{1}{2})$ RV, and consider all possible random samples of size 3 on X. Compute $\overline{X}$ and S^2 for each of the eight samples, and also compute the PMFs of $\overline{X}$ and S^2.

2. A die is rolled. Let X be the face value that turns up, and X_1, X_2 be two independent observations on X. Compute the PMF of $\overline{X}$.

3. Let $X_1, X_2, \ldots, X_n$ be a sample from some population. Show that

$$\max_{1 \le i \le n} |X_i - \overline{X}| < \frac{(n - 1)S}{\sqrt{n}}$$

unless either all the n observations are equal or exactly $n-1$ of the X_j's are equal. (Samuelson [97])

4. Let $x_1, x_2, \ldots, x_n$ be real numbers, and let $x_{(n)} = \max\{x_1, x_2, \ldots, x_n\}$, $x_{(1)} = \min\{x_1, x_2, \ldots, x_n\}$. Show that for any set of real numbers $a_1, a_2, \ldots, a_n$ such that $\sum_{i=1}^n a_i = 0$, the following inequality holds:

$$\left|\sum_{i=1}^n a_i x_i\right| \leq \tfrac{1}{2}\left(x_{(n)} - x_{(1)}\right)\sum_{i=1}^n |a_i|.$$

5. For any set of real numbers $x_1, x_2, \ldots, x_n$, show that the fraction of $x_1, x_2, \ldots, x_n$ included in the interval $(\bar{x} - ks, \bar{x} + ks)$ for $k \geq 1$ is at least $1 - 1/k^2$. Here $\bar{x}$ is the mean and s the standard deviation of x's.

7.3 SAMPLE CHARACTERISTICS AND THEIR DISTRIBUTIONS

Let $X_1, X_2, \ldots, X_n$ be a sample from a population DF F. In this section we consider some commonly used sample characteristics and their distributions.

Definition 1. Let $F_n^*(x) = n^{-1}\sum_{j=1}^n \varepsilon(x - X_j)$. Then $nF_n^*(x)$ is the number of X_k's $(1 \leq k \leq n)$ that are $\leq x$. $F_n^*(x)$ is called the *sample* (or *empirical*) *distribution function*.

We note that $0 \leq F_n^*(x) \leq 1$ for all x, and moreover, that F_n^* is right continuous, nondecreasing, and $F_n^*(-\infty) = 0$, $F_n^*(\infty) = 1$. Thus F_n^* is a DF.

If $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ is the order statistic for $X_1, X_2, \ldots, X_n$, then clearly

$$(1) \qquad F_n^*(x) = \begin{cases} 0 & \text{if } x < X_{(1)} \\ \dfrac{k}{n} & \text{if } X_{(k)} \leq x < X_{(k+1)} \qquad (k = 1, 2, \ldots, n-1) \\ 1 & \text{if } x \geq X_{(n)}. \end{cases}$$

For fixed but otherwise arbitrary $x \in \mathcal{R}$, $F_n^*(x)$ itself is an RV of the discrete type. The following result is immediate.

Theorem 1. The RV $F_n^*(x)$ has the probability function

$$(2) \qquad P\left\{F_n^*(x) = \frac{j}{n}\right\} = \binom{n}{j}[F(x)]^j[1 - F(x)]^{n-j}, \qquad j = 0, 1, \ldots, n,$$

with mean

$$(3) \qquad EF_n^*(x) = F(x)$$

and variance

$$\text{(4)} \qquad \operatorname{var}(F_n^*(x)) = \frac{F(x)[1 - F(x)]}{n}.$$

Proof. Since $\varepsilon(x - X_j)$, $j = 1, 2, \ldots, n$, are iid RVs, each with PMF

$$P\{\varepsilon(x - X_j) = 1\} = P\{x - X_j \geq 0\} = F(x)$$

and

$$P\{\varepsilon(x - X_j) = 0\} = 1 - F(x),$$

their sum $nF_n^*(x)$ is a $b(n, p)$ RV, where $p = F(x)$. Relations (2), (3), and (4) follow immediately.

Corollary 1. For each $x \in R$,

$$F_n^*(x) \xrightarrow{P} F(x) \qquad \text{as } n \to \infty.$$

Corollary 2. For each $x \in R$,

$$\frac{\sqrt{n}\,[F_n^*(x) - F(x)]}{\sqrt{F(x)[1 - F(x)]}} \xrightarrow{L} Z \qquad \text{as } n \to \infty,$$

where Z is $\mathcal{N}(0, 1)$.

Corollary 1 follows from the WLLN and Corollary 2 from the CLT. The convergence in Corollary 1 is for each value of x. It is possible to make a probability statement simultaneously for all x. We state the result without proof.

Theorem 2 (Glivenko–Cantelli Theorem). $F_n^*(x)$ converges uniformly to $F(x)$, that is, for $\varepsilon > 0$,

$$\lim_{n \to \infty} P\left\{ \sup_{-\infty < x < \infty} |F_n^*(x) - F(x)| > \varepsilon \right\} = 0.$$

For a proof of Theorem 2, we refer to Fisz [28, p. 391].

We next consider some typical values of the DF $F_n^*(x)$, called *sample statistics*. Since $F_n^*(x)$ has jump points X_j, $j = 1, 2, \ldots, n$, it is clear that all moments of $F_n^*(x)$ exist. Let us write

$$\text{(5)} \qquad a_k = n^{-1} \sum_{j=1}^{n} X_j^k$$

for the moment of order k about 0. Here a_k will be called the *sample moment of order k*. In this notation

$$a_1 = n^{-1} \sum_{j=1}^{n} X_j = \overline{X}. \tag{6}$$

The *sample central moment* is defined by

$$b_k = n^{-1} \sum_{j=1}^{n} (X_j - a_1)^k = n^{-1} \sum_{j=1}^{n} (X_j - \overline{X})^k. \tag{7}$$

Clearly,

$$b_1 = 0 \quad \text{and} \quad b_2 = \frac{n-1}{n} S^2.$$

As mentioned earlier, we do not call b_2 the sample variance. S^2 will be referred to as the *sample variance*, for reasons that will subsequently become clear. We have

$$b_2 = a_2 - a_1^2. \tag{8}$$

For the MGF of DF $F_n^*(x)$, we have

$$M^*(t) = n^{-1} \sum_{j=1}^{n} e^{tX_j}. \tag{9}$$

Similar definitions are made for sample moments of bivariate and multivariate distributions. For example, if $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ is a sample from a bivariate distribution, we write

$$\overline{X} = n^{-1} \sum_{j=1}^{n} X_j, \quad \text{and} \quad \overline{Y} = n^{-1} \sum_{j=1}^{n} Y_j \tag{10}$$

for the two sample means, and for the second-order sample central moments we write

$$b_{20} = n^{-1} \sum_{j=1}^{n} (X_j - \overline{X})^2, \quad b_{02} = n^{-1} \sum_{j=1}^{n} (Y_j - \overline{Y})^2, \quad \text{and} \tag{11}$$

$$b_{11} = n^{-1} \sum_{j=1}^{n} (X_j - \overline{X})(Y_j - \overline{Y}).$$

Once again we write

$$S_1^2 = (n-1)^{-1} \sum_{j=1}^{n} (X_j - \overline{X})^2 \quad \text{and} \quad S_2^2 = (n-1)^{-1} \sum_{j=1}^{n} (Y_j - \overline{Y})^2 \tag{12}$$

for the two *sample variances*, and for the *sample covariances* we use the quantity

$$S_{11} = (n-1)^{-1} \sum_{j=1}^{n} (X_j - \overline{X})(Y_j - \overline{Y}). \tag{13}$$

In particular, the *sample correlation coefficient* is defined by

$$R = \frac{b_{11}}{\sqrt{b_{20}b_{02}}} = \frac{S_{11}}{S_1 S_2}. \tag{14}$$

It can be shown (Problem 4) that $|R| \leq 1$; the extreme values ± 1 can occur only when all sample points $(X_1, Y_1), \ldots, (X_n, Y_n)$ lie on a straight line.

The sample quantiles are defined in a similar manner. Thus, if $0 < p < 1$, the *sample quantile of order* p, denoted by Z_p, is the order statistic $X_{(r)}$, where

$$r = \begin{cases} np & \text{if } np \text{ is an integer,} \\ [np + 1] & \text{if } np \text{ is not an integer.} \end{cases}$$

As usual, $[x]$ is the largest integer $\leq x$. Note that if np is an integer, we can take any value between $X_{(np)}$ and $X_{(np)+1}$ as the pth sample quantile. Thus, if $p = \frac{1}{2}$ and n is even, we can take any value between $X_{(n/2)}$ and $X_{(n/2)+1}$, the two middle values, as the median. It is customary to take the average. Thus the sample median is defined as

$$Z_{1/2} = \begin{cases} X_{((n+1)/2)} & \text{if } n \text{ is odd,} \\ \dfrac{X_{(n/2)} + X_{((n/2)+1)}}{2} & \text{if } n \text{ is even.} \end{cases} \tag{15}$$

Note that

$$\left[\frac{n}{2} + 1\right] = \frac{n+1}{2}$$

if n is odd.

Example 1. A random sample of 25 observations is taken from the interval (0,1):

0.50	0.24	0.89	0.54	0.34	0.89	0.92	0.17	0.32	0.80
0.06	0.21	0.58	0.07	0.56	0.20	0.31	0.17	0.41	0.38
0.88	0.61	0.35	0.06	0.90					

In order to compute F_{25}^*, the first step is to order the observations from smallest to largest. The ordered sample is

0.06,	0.06,	0.07,	0.17,	0.17,	0.20,	0.21,	0.24,	0.31,	0.32,
0.34,	0.35,	0.38,	0.41,	0.50,	0.54,	0.56,	0.58,	0.61,	0.80,
0.88,	0.89,	0.89,	0.90,	0.92					

Then the empirical DF is given by

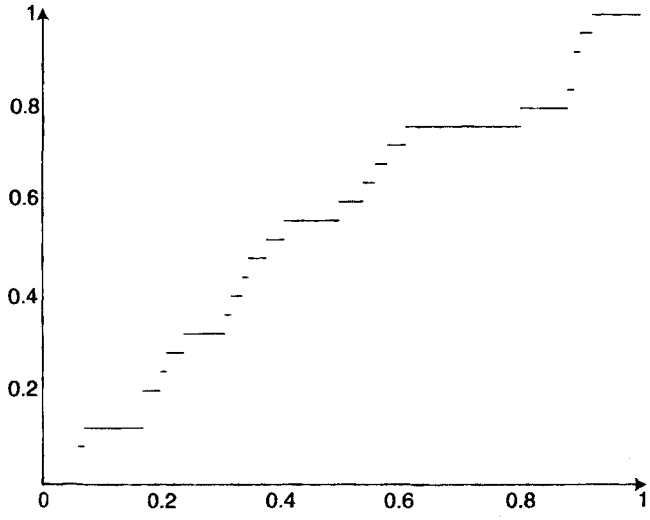

Fig. 1. Empirical DF for data of Example 1.

$$F_{25}^*(x) = \begin{cases} 0, & x < 0.06 \\ 2/25, & 0.06 \le x < 0.07 \\ 3/25, & 0.07 \le x < 0.17 \\ 5/25, & 0.17 \le x < 0.20 \\ \vdots & \\ 24/25, & 0.90 \le x < 0.92 \\ 1, & x \ge 0.92 \end{cases}$$

A plot of F_{25}^* is shown in Fig. 1. The sample mean and variance are

$$\bar{x} = 0.45, \quad s^2 = 0.084, \quad \text{and} \quad s = 0.29.$$

Also, sample median is the 13th observation in the ordered sample, namely, $z_{1/2} = 0.38$, and if $p = 0.2$, then $np = 5$ and $z_{.2} = 0.17$.

Next we consider the moments of sample characteristics. In the following we write $EX^k = m_k$ and $E(X - \mu)^k = \mu_k$ for the kth-order population moments. Whenever we use m_k (or μ_k), it will be assumed to exist. Also, σ^2 represents the population variance.

Theorem 3. Let $X_1, X_2, \ldots, X_n$ be a sample from a population with DF F. Then

$$E\overline{X} = \mu, \tag{16}$$

$$\operatorname{var}(\overline{X}) = \frac{\sigma^2}{n}, \tag{17}$$

$$E(\overline{X})^3 = \frac{m_3 + 3(n-1)m_2\mu + (n-1)(n-2)\mu^3}{n^2}, \tag{18}$$

and

$$E(\overline{X})^4 = \frac{m_4 + 4(n-1)m_3\mu + 6(n-1)(n-2)m_2\mu^2 + 3(n-1)m_2^2}{n^3} + \frac{(n-1)(n-2)(n-3)\mu^4}{n^3}. \tag{19}$$

Proof. In view of Theorems 4.5.3 and 4.5.7, it suffices to prove (18) and (19). We have

$$\left(\sum_{j=1}^{n} X_j\right)^3 = \sum_{j=1}^{n} X_j^3 + 3\sum_{j\neq k} X_j^2 X_k + \sum_{j\neq k\neq l} X_j X_k X_l,$$

and (18) follows. Similarly,

$$\begin{aligned}\left(\sum_{i=1}^{n} X_i\right)^4 &= \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{j=1}^{n} X_j^3 + 3\sum_{j\neq k} X_j^2 X_k + \sum_{j\neq k\neq l} X_j X_k X_l\right) \\ &= \sum_{i=1}^{n} X_i^4 + 4\sum_{j\neq k} X_j X_k^3 + 3\sum_{j\neq k} X_j^2 X_k^2 + 6\sum_{i\neq j\neq k} X_i^2 X_j X_k \\ &\quad + \sum_{i\neq j\neq k\neq l} X_i X_j X_k X_l,\end{aligned}$$

and (19) follows.

Theorem 4. For the third and fourth central moments of $\overline{X}$, we have

$$\mu_3(\overline{X}) = \frac{\mu_3}{n^2}, \tag{20}$$

and

$$\mu_4(\overline{X}) = \frac{\mu_4}{n^3} + 3\frac{(n-1)\mu_2^2}{n^3}. \tag{21}$$

Proof. We have

$$\mu_3(\bar{X}) = E(\bar{X} - \mu)^3 = \frac{1}{n^3} E\left[\sum_{i=1}^{n}(X_i - \mu)\right]^3$$

$$= \frac{1}{n^3}\sum_{i=1}^{n} E(X_i - \mu)^3 = \frac{\mu_3}{n^2},$$

and

$$\mu_4(\bar{X}) = E(\bar{X} - \mu)^4 = \frac{1}{n^4} E\left[\sum_{i=1}^{n}(X_i - \mu)\right]^4$$

$$= \frac{1}{n^4}\sum_{i=1}^{n} E(X_i - \mu)^4 + \binom{4}{2}\frac{1}{n^4}\sum_{i<j} E[(X_i - \mu)^2(X_j - \mu)^2]$$

$$= \frac{\mu_4}{n^3} + \frac{3(n-1)}{n^3}\mu_2^2.$$

Theorem 5. For the moments of b_2, we have

$$E(b_2) = \frac{(n-1)\sigma^2}{n}, \tag{22}$$

$$\operatorname{var}(b_2) = \frac{\mu_4 - \mu_2^2}{n} - \frac{2(\mu_4 - 2\mu_2^2)}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3}, \tag{23}$$

$$E(b_3) = \frac{(n-1)(n-2)}{n^2}\mu_3, \tag{24}$$

and

$$E(b_4) = \frac{(n-1)(n^2 - 3n + 3)}{n^3}\mu_4 + \frac{3(n-1)(2n-3)}{n^3}\mu_2^2. \tag{25}$$

Proof. We have

$$Eb_2 = \frac{1}{n} E\left[\sum_{1}^{n}(X_i - \mu + \mu - \bar{X})^2\right]$$

$$= \frac{1}{n} E\left[\sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2\right]$$

$$= \frac{1}{n}(n\sigma^2 - \sigma^2) = \frac{n-1}{n}\sigma^2.$$

Now

$$n^2 b_2^2 = \left[\sum_{i=1}^{n}(X_i - \mu)^2 - n(\overline{X} - \mu)^2\right]^2.$$

Writing $Y_i = X_i - \mu$, we see that $EY_i = 0$, $\operatorname{var}(Y_i) = \sigma^2$, and $EY_i^4 = \mu_4$. We have

$$\begin{aligned} n^2 E b_2^2 &= E\left(\sum_1^n Y_i^2 - n\overline{Y}^2\right)^2 \\ &= E\left[\sum_{i=1}^n Y_i^4 + \sum_{i\neq j} Y_i^2 Y_j^2 - \frac{2}{n}\left(\sum_{i\neq j} Y_i^2 Y_j^2 + \sum_{j=1}^n Y_j^4\right)\right. \\ &\qquad \left. + \frac{1}{n^2}\left(3\sum_{i\neq j} Y_i^2 Y_j^2 + \sum_1^n Y_j^4\right)\right]. \end{aligned}$$

It follows that

$$\begin{aligned} n^2 E b_2^2 &= n\mu_4 + n(n-1)\sigma^4 - \frac{2}{n}[n(n-1)\sigma^4 + n\mu_4] + \frac{1}{n^2}[3n(n-1)\sigma^4 + n\mu_4] \\ &= \left(n - 2 + \frac{1}{n}\right)\mu_4 + \left(n - 2 + \frac{3}{n}\right)(n-1)\mu_2^2 \qquad (\mu_2 = \sigma^2). \end{aligned}$$

Therefore,

$$\begin{aligned} \operatorname{var}(b_2) &= E b_2^2 - (E b_2)^2 \\ &= \left(n - 2 + \frac{1}{n}\right)\frac{\mu_4}{n^2} + (n-1)\left(n - 2 + \frac{3}{n}\right)\frac{\mu_2^2}{n^2} - \left(\frac{n-1}{n}\right)^2 \mu_2^2 \\ &= \left(n - 2 + \frac{1}{n}\right)\frac{\mu_4}{n^2} + (n-1)(3-n)\frac{\mu_2^2}{n^3}, \end{aligned}$$

as asserted.

Relations (24) and (25) can be proved similarly.

Corollary 1. $ES^2 = \sigma^2$.

This is precisely the reason why we call S^2, and not b_2, the sample variance.

Corollary 2. $\operatorname{var}(S^2) = \dfrac{\mu_4}{n} + \dfrac{3-n}{n(n-1)}\mu_2^2$.

Remark 1. The results of Theorems 3 to 5 can easily be modified and stated for the case when the X_i's are exchangeable RVs. Thus (16) holds and (17) has to be

modified to

$$\text{(17')} \qquad \operatorname{var}(\bar{X}) = \frac{\sigma^2}{n} + \frac{n-1}{n}\rho\sigma^2$$

where ρ is the correlation coefficient between X_i and X_j. The expressions for $(\Sigma X_j)^3$ and $(\Sigma X_j)^4$ in the proof of Theorem 3 still hold, but both (18) and (19) need appropriate modification. For example, (18) changes to

$$\text{(18')} \qquad E\bar{X}^3 = \frac{m_3 + 3(n-1)E(X_j^2 X_k) + (n-1)(n-2)E(X_j X_k X_l)}{n^2}.$$

Let us show how Corollary 1 changes for exchangeable RVs. Clearly,

$$(n-1)S^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2$$

so that

$$\begin{aligned}(n-1)ES^2 &= n\sigma^2 - nE(\bar{X} - \mu)^2 \\ &= n\sigma^2 - \left[\sigma^2 + (n-1)\rho\sigma^2\right].\end{aligned}$$

in view of (17'). It follows that

$$ES^2 = \sigma^2(1-\rho).$$

We note that $E(S^2 - \sigma^2) = -\rho\sigma^2$, and moreover, from Problem 4.5.19 [or from (17')] we note that $\rho \geq -1/(n-1)$, so that $1 - \rho \leq n/(n-1)$, and hence

$$0 \leq ES^2 \leq \frac{n}{n-1}\sigma^2.$$

Remark 2. In simple random sampling from a (finite) population of size N, we note that when $n = N$, $\bar{X} = \mu$, which is a constant, so that (17') reduces to

$$0 = \frac{\sigma^2}{N} + \frac{N-1}{N}\rho\sigma^2,$$

so that $\rho = -1/(N-1)$. It follows that

$$\text{(17'')} \qquad \operatorname{var}(\bar{X}) = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right) = \frac{N-n}{N-1}\frac{\sigma^2}{n}.$$

The factor $(N-n)/(N-1)$ in (17'') is called the *finite population correction factor*. As $N \to \infty$, with n fixed, $(N-n)/(N-1) \to 1$, so that the expression from $\operatorname{var}(\bar{X})$ in (17'') approaches that in (17).

The following result provides a justification for our definition of sample covariance.

Theorem 6. Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be a sample from a bivariate population with variances σ_1^2, σ_2^2 and covariance $\rho\sigma_1\sigma_2$. Then

$$ES_1^2 = \sigma_1^2, \quad ES_2^2 = \sigma_2^2, \quad \text{and} \quad ES_{11} = \rho\sigma_1\sigma_2, \tag{26}$$

where S_1^2, S_2^2, and S_{11} are defined in (12) and (13).

Proof. It follows from Corollary 1 to Theorem 5 that $ES_1^2 = \sigma_1^2$ and $ES_2^2 = \sigma_2^2$. To prove that $ES_{11} = \rho\sigma_1\sigma_2$, we note that X_i is independent of $X_j (i \neq j)$ and Y_j $(i \neq j)$. We have

$$(n-1)ES_{11} = E\left[\sum_{j=1}^{n}(X_j - \overline{X})(Y_j - \overline{Y})\right].$$

Now

$$E[(X_j - \overline{X})(Y_j - \overline{Y})] = E\left[X_jY_j - X_j\frac{\sum_1^n Y_j}{n} - Y_j\frac{\sum_1^n X_j}{n} + \frac{\sum X_j \sum Y_j}{n^2}\right]$$

$$= E(XY) - \frac{1}{n}[E(XY) + (n-1)EXEY] - \frac{1}{n}[E(XY) + (n-1)EXEY]$$

$$+ \frac{1}{n^2}[nE(XY) + n(n-1)EXEY]$$

$$= \frac{n-1}{n}[E(XY) - EXEY]$$

and it follows that

$$(n-1)ES_{11} = n\frac{n-1}{n}[E(XY) - EXEY],$$

that is,

$$ES_{11} = E(XY) - EXEY = \operatorname{cov}(X, Y) = \rho\sigma_1\sigma_2,$$

as asserted.

We next turn our attention to the distributions of sample characteristics. Several possibilities exist. If the exact sampling distribution is required, the method of transformation described in Section 4.4 can be used. Sometimes the technique of MGF or CF can be applied. Thus, if $X_1, X_2, \ldots, X_n$ is a random sample from a population distribution for which the MGF exists, the MGF of the sample mean $\overline{X}$ is given by

$$M_{\overline{X}}(t) = \prod_{i=1}^{n} Ee^{tX_i/n} = \left[M\left(\frac{t}{n}\right)\right]^n, \tag{27}$$

where M is the MGF of the population distribution. If $M_{\overline{X}}(t)$ has one of the known forms, it is possible to write the PDF of $\overline{X}$. Although this method has the obvious drawback that it applies only to distributions for which all moments exist, we will see in Section 7.6 its effectiveness in the important case of sampling from a normal population where this condition is satisfied. An analog of (27) holds for CFs without any condition on existence of moments. Indeed,

$$\phi_{\overline{X}}(t) = \sum_{j=1}^{n} Ee^{itX_j/n} = \left[\phi\left(\frac{t}{n}\right)\right]^n, \tag{28}$$

where ϕ is the CF of X_j.

Example 2. Let $X_1, X_2, \ldots, X_n$ be a sample from a $G(\alpha, 1)$ distribution. We will compute the PDF of $\overline{X}$. We have

$$M_{\overline{X}}(t) = \left[M\left(\frac{t}{n}\right)\right]^n = \frac{1}{(1 - t/n)^{\alpha n}}, \qquad \frac{t}{n} < 1,$$

so that $\overline{X}$ is a $G(\alpha n, 1/n)$ variate.

Example 3. Let $X_1, X_2, \ldots, X_n$ be a random sample from a uniform distribution on (0, 1). Consider the geometric mean

$$Y_n = \left(\prod_{i=1}^{n} X_i\right)^{1/n}.$$

We have $\log Y_n = (1/n)\sum_{i=1}^{n} \log X_i$, so that $\log Y_n$ is the mean of $\log X_1, \ldots, \log X_n$.

The common PDF of $\log X_1, \ldots, \log X_n$ is

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

which is the negative exponential distribution with parameter $\beta = 1$. We see that the MGF of $\log Y_n$ is given by

$$M(t) = \prod_{i=1}^{n} Ee^{t \log X_i/n} = \frac{1}{(1 + t/n)^n},$$

and the PDF of $\log Y_n$ is given by

$$f^*(x) = \begin{cases} \dfrac{n^n}{\Gamma(n)}(-x)^{n-1}e^{nx}, & -\infty < x < 0, \\ 0, & \text{otherwise.} \end{cases}$$

It follows that Y_n has PDF

$$f_{Y_n}(y) = \begin{cases} \dfrac{n^n}{\Gamma(n)}y^{n-1}(-\log y)^{n-1}, & 0 < y < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Example 4 (Hogben [43]). Let $X_1, X_2, \ldots, X_n$ be a random sample from a Bernoulli distribution with parameter p, $0 < p < 1$. Let $\overline{X}$ be the sample mean and S^2 the sample variance. We will find the PMF of S^2. Note that $S_n = \sum_{i=1}^n X_i = \sum_{i=1}^n X_i^2$ and that S_n is $b(n, p)$. Since

$$\begin{aligned}(n-1)S^2 &= \sum_{i=1}^n X_i^2 - n(\overline{X})^2 \\ &= \frac{S_n(n-S_n)}{n},\end{aligned}$$

S^2 only assumes values of the form

$$t = \frac{i(n-i)}{n(n-1)}, \qquad i = 0, 1, 2, \ldots, \left[\frac{n}{2}\right],$$

where $[x]$ is the largest integer $\leq x$. Thus

$$\begin{aligned}P\{S^2 = t\} &= P\{nS_n - S_n^2 = i(n-i)\} = P\left\{\left(S_n - \frac{n}{2}\right)^2 - \left(i - \frac{n}{2}\right)^2\right\} \\ &= P\{S_n = i \text{ or } S_n = n-i\} \\ &= \binom{n}{i}p^i(1-p)^{n-i} + \binom{n}{i}p^{n-i}(1-p)^i \\ &= \binom{n}{i}p^i(1-p)^i\{(1-p)^{n-2i} + p^{n-2i}\}, \qquad i < \left[\frac{n}{2}\right].\end{aligned}$$

If n is even, $n = 2m$, say, where $m \geq 0$ is an integer, and $i = m$, then

$$P\left\{S^2 = \frac{m}{2(2m-1)}\right\} = 2\binom{2m}{m}p^m(1-p)^m$$

In particular, if $n = 7$, $S^2 = 0$, $\frac{1}{7}$, $\frac{5}{21}$, and $\frac{2}{7}$ with probabilities $\{p^7 + (1-p)^7\}$, $7p(1-p)\{p^5+(1-p)^5\}$, $21p^2(1-p)^2\{p^3+(1-p)^3\}$, and $35p^3(1-p)^3$, respectively.

If $n = 6$, then $S^2 = 0, \frac{1}{6}, \frac{4}{15}$, and $\frac{3}{10}$ with probabilities $\{p^6 + (1-p)^6\}$, $6p(1-p)\{p^4 + (1-p)^4\}$, $15p^2(1-p)^2\{p^2 + (1-p)^2\}$, and $40p^3(1-p)^3$, respectively.

We have already considered the distribution of the sample quantiles in Section 4.7 and the distribution of range $X_{(n)} - X_{(1)}$ in Example 4.7.4. It can be shown without much difficulty that the distribution of the sample median is given by

$$(29) \quad f_r(y) = \frac{n!}{(r-1)!\,(n-r)!}[F(y)]^{r-1}[1-F(y)]^{n-r} f(y) \qquad \text{if } r = \frac{n+1}{2},$$

where F and f are the population DF and PDF, respectively. If $n = 2m$ and the median is taken as the average of $X_{(m)}$ and $X_{(m+1)}$, then

$$(30) \quad f_r(y) = \frac{2(2m)!}{[(m-1)!]^2} \int_y^\infty [F(2y-v)]^{m-1}[1-F(v)]^{m-1} f(2y-v) f(v)\, dv.$$

Example 5. Let $X_1, X_2, \ldots, X_n$ be a random sample from $U(0, 1)$. Then the integrand in (30) is positive for the intersection of the regions $0 < 2y - v < 1$ and $0 < v < 1$. This gives $v/2 < y < (v+1)/2$, $y < v$, and $0 < v < 1$. The shaded area in Fig. 2 gives the limits on the integral as

$$y < v < 2y \qquad \text{if } 0 < y \leq \tfrac{1}{2},$$

and

$$y < v < 1 \qquad \text{if } \tfrac{1}{2} < y < 1.$$

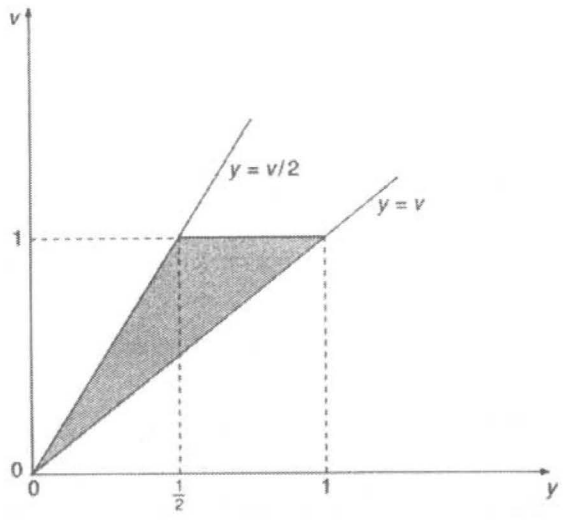

Fig. 2. $\{y < v \leq 2y, 0 < y < \frac{1}{2}$, and $y < v < 1, \frac{1}{2} < y \leq 1\}$.

In particular, if $m = 2$, the PDF of the median, $(X_{(2)} + X_{(3)})/2$, is given by

$$f_r(y) = \begin{cases} 8y^2(3 - 4y) & \text{if } 0 < y < \frac{1}{2}, \\ 8(4y^3 - 9y^2 + 6y - 1) & \text{if } \frac{1}{2} < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$

In Section 7.5 we study large-sample theory techniques to approximate distributions of sample statistics when n is large.

PROBLEMS 7.3

1. Let $X_1, X_2, \ldots, X_n$ be random sample from a DF F, and let $F_n^*(x)$ be the sample distribution function. Find $\text{cov}(F_n^*(x), F_n^*(y))$ for fixed real numbers x, y.

2. Let F_n^* be the empirical DF of a random sample from DF F. Show that

$$P\left\{|F_n^*(x) - F(x)| \geq \frac{\varepsilon}{2\sqrt{n}}\right\} \leq \frac{1}{\varepsilon^2} \qquad \text{for all } \varepsilon > 0.$$

3. For the data of Example 7.2.2, compute the sample distribution function.

4. (a) Show that the sample correlation coefficient R satisfies $|R| \leq 1$ with equality if and only if all sample points lie on a straight line.

   (b) If we write $U_i = aX_i + b$ $(a \neq 0)$ and $V_i = cY_i + d$ $(c \neq 0)$, what is the sample correlation coefficient between the U's and the V's?

5. (a) A sample of size 2 is taken from the PDF $f(x) = 1$, $0 \leq x \leq 1$, and $= 0$ otherwise. Find $P(\bar{X} \geq 0.9)$.

   (b) A sample of size 2 is taken from $b(1, p)$. Find (i) $P(\bar{X} \leq p)$, and (ii) $P(S^2 \geq 0.5)$.

6. Let $X_1, X_2, \ldots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$. Compute the first four sample moments of $\overline{X}$ about the origin and about the mean. Also compute the first four sample moments of S^2 about the mean.

7. Derive the PDF of the median given in (29) and (30).

8. Let $U_{(1)}, U_{(2)}, \ldots, U_{(n)}$ be the order statistics of a sample size n from $U(0, 1)$. Compute $EU_{(r)}^k$ for any $1 \leq r \leq n$ and integer k (> 0). In particular, show that

$$EU_{(r)} = \frac{r}{n+1} \quad \text{and} \quad \text{var}(U_{(r)}) = \frac{r(n-r+1)}{(n+1)^2(n+2)}.$$

   Show also that the correlation coefficient between $U_{(r)}$ and $U_{(s)}$ for $1 \leq r < s \leq n$ is given by $[r(n-s+1)/s(n-r+1)]^{1/2}$.

9. Let $X_1, X_2, \ldots, X_n$ be n independent observations on X. Find the sampling distribution of $\overline{X}$, the sample mean, if (a) $X \sim P(\lambda)$, (b) $X \sim C(1, 0)$, and (c) $X \sim \chi^2(m)$.

10. Let $X_1, X_2, \ldots, X_n$ be a random sample from $G(\alpha, \beta)$. Let us write $Y_n = (\overline{X} - \alpha\beta)/\beta\sqrt{\alpha/n}$, $n = 1, 2, \ldots$.

(a) Compute the first four moments of Y_n, and compare them with the first four moments of the standard normal distribution.

(b) Compute the coefficients of skewness α_3 and of kurtosis α_4 for the RVs Y_n. (For definitions of α_3, α_4, see Problem 3.2.10.)

11. Let $X_1, X_2, \ldots, X_n$ be a random sample from $U[0, 1]$. Also let $Z_n = (\overline{X} - 0.5)/\sqrt{1/12n}$. Repeat Problem 10 for the sequence Z_n.

12. Let $X_1, X_2, \ldots, X_n$ be a random sample from $P(\lambda)$. Find $\text{var}(S^2)$, and compare it with $\text{var}(\overline{X})$. Note that $E\overline{X} = \lambda = ES^2$. (*Hint:* Use Problem 3.2.9.)

13. Prove (24) and (25).

14. Multiple RVs $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ are exchangeable if the $n!$ permutations $(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}, \ldots, \mathbf{X}_{i_n})$ have the same multidimensional distribution. Consider the special case when $\mathbf{X}$'s are two-dimensional. Find an analog of Theorem 6 for exchangeable bivariate RVs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$.

7.4 CHI-SQUARE, *t*-, AND *F*-DISTRIBUTIONS: EXACT SAMPLING DISTRIBUTIONS

In this section we investigate certain distributions that arise in sampling from a normal population. Let $X_1, X_2, \ldots, X_n$ be a sample from $\mathcal{N}(\mu, \sigma^2)$. Then we know that $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$. Also, $\{\sqrt{n}(\overline{X} - \mu)/\sigma\}^2$ is $\chi^2(1)$. We determine the distribution of S^2 in the next section. Here we define mainly chi-square, t-, and F-distributions and study their properties. Their importance will become evident in the next section and later in the testing of statistical hypotheses (Chapter 10).

The first distribution of interest is the *chi-square distribution*, defined in Chapter 5 as a special case of the gamma distribution. Let $n > 0$ be an integer. Then $G(n/2, 2)$ is a $\chi^2(n)$ RV. In view of Theorem 5.3.29 and Corollary 2 to Theorem 5.3.4, the following result holds.

Theorem 1. Let $X_1, X_2, \ldots, X_n$ be iid RVs, and let $S_n = \sum_{k=1}^n X_k$. Then

(a) $S_n \sim \chi^2(n) \Leftrightarrow X_1 \sim \chi^2(1)$, and

(b) $X_1 \sim \mathcal{N}(0, 1) \Rightarrow \sum_{k=1}^n X_k^2 \sim \chi^2(n)$.

If X has a chi-square distribution with n d.f., we write $X \sim \chi^2(n)$. We recall that if $X \sim \chi^2(n)$, its PDF is given by

$$f(x) = \begin{cases} \dfrac{x^{n/2-1}e^{-x/2}}{2^{n/2}\Gamma(n/2)} & \text{if } x \geq 0, \\ 0 & \text{if } x < 0, \end{cases} \tag{1}$$

the MGF by

$$M(t) = (1 - 2t)^{-n/2} \qquad \text{for } t < \tfrac{1}{2}, \tag{2}$$

and the mean and the variance by

$$EX = n, \quad \text{and} \quad \operatorname{var}(X) = 2n. \tag{3}$$

The $\chi^2(n)$ distribution is tabulated for values of $n = 1, 2, \ldots$. Tables usually go up to $n = 30$, since for $n > 30$ it is possible to use normal approximation. In Fig. 1 we plot the PDF (1) for selected values of n.

We will write $\chi^2_{n,\alpha}$ for the upper α percent point of the $\chi^2(n)$ distribution, that is,

$$P\{\chi^2(n) > \chi^2_{n,\alpha}\} = \alpha. \tag{4}$$

Table ST3 at the end of the book gives the values of $\chi^2_{n,\alpha}$ for some selected values of n and α.

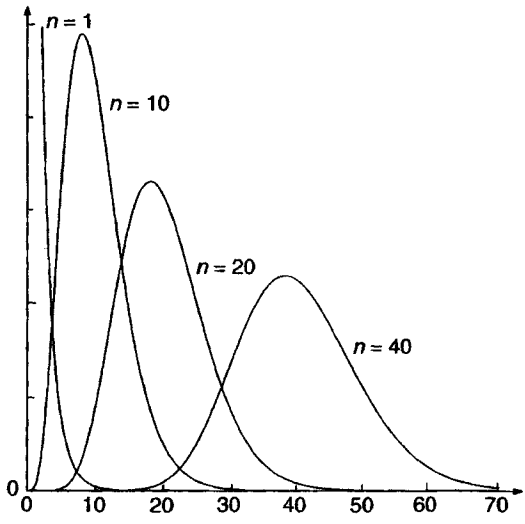

Fig. 1. Chi-square densities.

Example 1. Let $n = 25$. Then, from Table ST3,

$$P\{\chi^2(25) \leq 34.382\} = 0.90.$$

Let us approximate this probability using CLT. We see that $E\chi^2(25) = 25$, $\operatorname{var} \chi^2(25) = 50$, so that

$$\begin{aligned} P\{\chi^2(25) \leq 34.382\} &= P\left\{\frac{\chi^2(25) - 25}{\sqrt{50}} \leq \frac{34.382 - 25}{5\sqrt{2}}\right\} \\ &\approx P\{Z \leq 1.32\} \\ &= 0.9066. \end{aligned}$$

Definition 1. Let $X_1, X_2, \ldots, X_n$ be independent normal RVs with $EX_i = \mu_i$ and $\operatorname{var}(X_i) = \sigma^2$, $i = 1, 2, \ldots, n$. Also, let $Y = \sum_{i=1}^n X_i^2/\sigma^2$. The RV Y is said to be a *noncentral chi-square* RV with *noncentrality parameter* $\sum_{i=1}^n \mu_i^2/\sigma^2$ and n d.f. We will write $Y \sim \chi^2(n, \delta)$, where $\delta = \sum_{i=1}^n \mu_i^2/\sigma^2$.

Although the PDF of a $\chi^2(n, \delta)$ RV is hard to compute (see Problem 16), its MGF is easily evaluated. We have

$$M(t) = Ee^{t\sum_1^n X_i^2/\sigma^2} = \prod_1^n Ee^{tX_i^2/\sigma^2},$$

where $X_i \sim \mathcal{N}(\mu_i, \sigma^2)$. Thus

$$Ee^{tX_i^2/\sigma^2} = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{tx_i^2}{\sigma^2} - \frac{(x_i - \mu_i)^2}{2\sigma^2}\right] dx_i,$$

where the integral exists for $t < \frac{1}{2}$. In the integrand we complete squares, and after some simple algebra we obtain

$$Ee^{tX_i^2/\sigma^2} = \frac{1}{\sqrt{1-2t}} \exp\left[\frac{t\mu_i^2}{\sigma^2(1-2t)}\right], \qquad t < \frac{1}{2}.$$

It follows that

$$M(t) = (1-2t)^{-n/2} \exp\left(\frac{t}{1-2t}\frac{\sum \mu_i^2}{\sigma^2}\right), \qquad t < \frac{1}{2}, \tag{5}$$

and the MGF of a $\chi^2(n, \delta)$ RV is therefore

$$M(t) = (1-2t)^{-n/2} \exp\left(\frac{t}{1-2t}\delta\right), \qquad t < \frac{1}{2}. \tag{6}$$

It is immediate that if $Y_1, Y_2, \ldots, Y_k$ are independent, $Y_i \sim \chi^2(n_i, \delta_i)$, $i = 1, 2, \ldots, k$, then $\sum_{i=1}^k Y_i$ is $\chi^2(\sum_{i=1}^k n_i, \sum_{i=1}^k \delta_i)$.

The mean and variance of $\chi^2(n, \delta)$ are easy to calculate. We have

$$EY = \frac{\sum_1^n EX_i^2}{\sigma^2} = \frac{\sum_1^n [\text{var}(X_i) + (EX_i)^2]}{\sigma^2}$$
$$= \frac{n\sigma^2 + \sum_1^n \mu_i^2}{\sigma^2} = n + \delta$$

and

$$\text{var}(Y) = \text{var}\left(\frac{\sum_1^n X_i^2}{\sigma^2}\right) = \frac{1}{\sigma^4}\left[\sum_{i=1}^n \text{var}(X_i^2)\right]$$
$$= \frac{1}{\sigma^4}\left[\sum_{i=1}^n EX_i^4 - \sum_{i=1}^n [E(X_i^2)]^2\right]$$
$$= \frac{1}{\sigma^4}\left[\sum_{i=1}^n (3\sigma^4 + 6\sigma^2\mu_i^2 + \mu_i^4) - \sum_{i=1}^n (\sigma^2 + \mu_i^2)^2\right]$$
$$= \frac{1}{\sigma^4}\left(2n\sigma^4 + 4\sigma^2 \sum \mu_i^2\right) = 2n + 4\delta.$$

We next turn our attention to *Student's t-statistic*, which arises quite naturally in sampling from a normal population.

Definition 2. Let $X \sim \mathcal{N}(0, 1)$ and $Y \sim \chi^2(n)$, and let X and Y be independent. Then the statistic

$$T = \frac{X}{\sqrt{Y/n}} \tag{7}$$

is said to have a *t-distribution* with n d.f. and we write $T \sim t(n)$.

Theorem 2. The PDF of T defined in (7) is given by

$$f_n(t) = \frac{\Gamma[(n+1)/2]}{\Gamma(n/2)\sqrt{n\pi}}(1 + t^2/n)^{-(n+1)/2}, \qquad -\infty < t < \infty. \tag{8}$$

The proof is left as an exercise.

Remark 1. For $n = 1$, T is a Cauchy RV. We will therefore assume that $n > 1$. For each n, we have a different PDF. In Fig. 2 we plot $f_n(t)$ for some selected values of n. Like the normal distribution, the t-distribution is important in the theory of statistics and hence is tabulated (Table ST4).

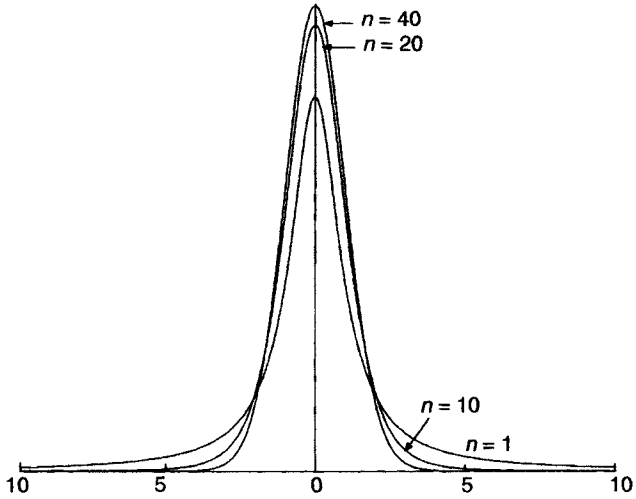

Fig. 2. Student's t-densities.

Remark 2. The PDF $f_n(t)$ is symmetric in t, and $f_n(t) \to 0$ as $t \to +\infty$. For large n, the t-distribution is close to the normal distribution. Indeed, $(1 + t^2/n)^{-(n+1)/2} \to e^{-t^2/2}$ as $n \to \infty$. Moreover, as $t \to \infty$ or $t \to -\infty$, the tails of $f_n(t) \to 0$ much more slowly than do the tails of the $\mathcal{N}(0, 1)$ PDF. Thus for small n and large t_0,

$$P\{|T| > t_0\} \geq P\{|Z| > t_0\}, \qquad Z \sim \mathcal{N}(0, 1);$$

that is, there is more probability in the tail of the t-distribution than in the tail of the standard normal. In what follows we write $t_{n,\alpha/2}$ for the value (Fig. 3) of T for which

$$P\{|T| > t_{n,\alpha/2}\} = \alpha. \tag{9}$$

In Table ST4 positive values of $t_{n,\alpha}$ are tabulated for selected values of n and α. Negative values may be obtained from symmetry, $t_{n,1-\alpha} = -t_{n,\alpha}$.

Example 2. Let $n = 5$. Then from Table ST4, we get $t_{5,0.025} = 2.571$ and $t_{5,0.05} = 2.015$. The corresponding values under the $\mathcal{N}(0, 1)$ distribution are $z_{0.025} = 1.96$ and $z_{0.05} = 1.65$. For $n = 30$,

$$t_{30,0.05} = 1.697 \qquad \text{and} \qquad z_{0.05} = 1.65.$$

Theorem 3. Let $X \sim t(n)$, $n > 1$. Then EX^r exists for $r < n$. In particular, if $r < n$ is odd,

$$EX^r = 0, \tag{10}$$

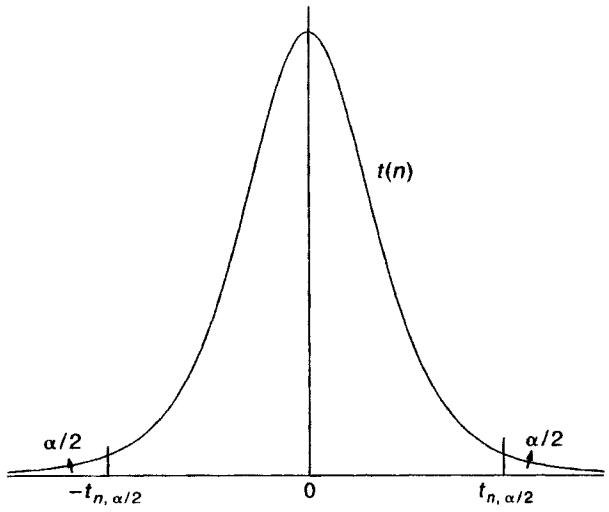

Fig. 3.

and if $r < n$ is even,

$$EX^r = n^{r/2}\frac{\Gamma[(r+1)/2]\Gamma[(n-r)/2]}{\Gamma(1/2)\Gamma(n/2)}. \tag{11}$$

Corollary. If $n > 2$, $EX = 0$ and $EX^2 = \text{var}(X) = n/(n-2)$.

Remark 3. If in Definition 2 we take $X \sim \mathcal{N}(\mu, \sigma^2)$, $Y/\sigma^2 \sim \chi^2(n)$, and X and Y independent,

$$T = \frac{X}{\sqrt{Y/n}}$$

is said to have a *noncentral t-distribution with parameter* (also called *noncentrality parameter*) $\delta = \mu/\sigma$ and d.f. n. Various moments of noncentral t-distribution may be computed by using the fact that expectation of a product of independent RVs is the product of their expectations.

We leave the reader to show (Problem 3) that if T has a noncentral t-distribution with n d.f. and noncentrality parameter δ, then

$$ET = \delta\frac{\Gamma[(n-1)/2]}{\Gamma(n/2)}\sqrt{\frac{n}{2}}, \qquad n > 1, \tag{12}$$

and

$$\text{var}(T) = \frac{n(1+\delta^2)}{n-2} - \frac{\delta^2 n}{2}\left(\frac{\Gamma[(n-1)/2]}{\Gamma(n/2)}\right)^2, \qquad n > 2. \tag{13}$$

Definition 3. Let X and Y be independent χ^2 RVs with m and n d.f., respectively. The RV

$$F = \frac{X/m}{Y/n} \tag{14}$$

is said to have an *F-distribution* with (m, n) d.f., and we write $F \sim F(m, n)$.

Theorem 4. The PDF of the F-statistic defined in (14) is given by

$$g(f) = \begin{cases} \dfrac{\Gamma[(m+n)/2]}{\Gamma(m/2)\Gamma(n/2)} \left(\dfrac{m}{n}\right) \left(\dfrac{m}{n} f\right)^{(m/2)-1} \\ \quad \cdot \left(1 + \dfrac{m}{n} f\right)^{-(m+n)/2}, & f > 0, \\ 0, & f \le 0. \end{cases} \tag{15}$$

The proof is left as an exercise.

Remark 4. If $X \sim F(m, n)$, then $1/X \sim F(n, m)$. If we take $m = 1$, then $F = [t(n)]^2$, so that $F(1, n)$ and $t^2(n)$ have the same distribution. It also follows that if Z is $C(1, 0)$ [which is the same as $t(1)$], Z^2 is $F(1, 1)$.

Remark 5. As usual, we write $F_{m,n,\alpha}$ for the upper α percent point of the $F(m, n)$ distribution, that is,

$$P\{F(m, n) > F_{m,n,\alpha}\} = \alpha. \tag{16}$$

From Remark 4, we have the following relation:

$$F_{m,n,1-\alpha} = \frac{1}{F_{n,m,\alpha}}. \tag{17}$$

It therefore suffices to tabulate values of F that are ≥ 1. This is done in Table ST5, where values of $F_{m,n,\alpha}$ are listed for selected values of m, n, and α. See Fig. 4 for a plot of $g(f)$.

Theorem 5. Let $X \sim F(m, n)$. Then, for $k > 0$, integral,

$$EX^k = \left(\frac{n}{m}\right)^k \frac{\Gamma[k + (m/2)]\Gamma[(n/2) - k]}{\Gamma[(m/2)\Gamma(n/2)]} \qquad \text{for } n > 2k. \tag{18}$$

In particular,

$$EX = \frac{n}{n-2}, \qquad n > 2, \tag{19}$$

and

$$\text{var}(X) = \frac{n^2(2m + 2n - 4)}{m(n-2)^2(n-4)}, \qquad n > 4. \tag{20}$$

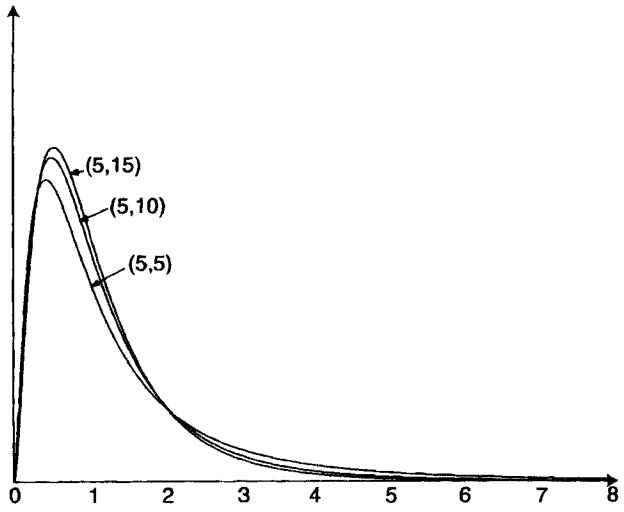

Fig. 4. F densities.

Proof. We have for a positive integer k,

$$\int_0^\infty f^k f^{m/2-1}\left(1+\frac{m}{n}f\right)^{-(m+n)/2} DF \tag{21}$$
$$= \left(\frac{m}{n}\right)^{k+(m/2)} \int_0^1 x^{k+(m/2)-1}(1-x)^{(n/2)-k-1}\,dx,$$

where we have changed the variable to $x = (m/n)f[1 + (m/n)f]^{-1}$. The integral in the right side of (21) converges for $(n/2) - k > 0$ and diverges for $(n/2) - k \leq 0$. We have

$$EX^k = \frac{\Gamma[(m+n)/2]}{\Gamma(m/2)\Gamma(n/2)}\left(\frac{m}{n}\right)^{m/2}\left(\frac{n}{m}\right)^{k+(m/2)} B\left(k+\frac{m}{2}, \frac{n}{2}-k\right),$$

as asserted.

For $k = 1$ we get

$$EX = \frac{n}{m}\frac{m/2}{(n/2)-1} = \frac{n}{n-2}, \qquad n > 2.$$

Also,

$$EX^2 = \left(\frac{n}{m}\right)^2 \frac{(m/2)[(m/2)+1]}{[(n/2)-1][(n/2)-2]}, \qquad n > 4,$$
$$= \left(\frac{n}{m}\right)^2 \frac{m(m+2)}{(n-2)(n-4)},$$

and

$$\operatorname{var}(X) = \left(\frac{n}{m}\right)^2 \frac{m(m+2)}{(n-2)(n-4)} - \left(\frac{n}{n-2}\right)^2$$
$$= \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}, \qquad n > 4.$$

Theorem 6. If $X \sim F(m, n)$, then $Y = 1/[1 + (m/n)X]$ is $B(n/2, m/2)$. Consequently, for each $x > 0$,

$$F_X(x) = 1 - F_Y\left[\frac{1}{1 + (m/n)x}\right].$$

If in Definition 3 we take X to be a noncentral χ^2 RV with n d.f. and noncentrality parameter δ, we get a *noncentral F* RV.

Definition 4. Let $X \sim \chi^2(m, \delta)$ and $Y \sim \chi^2(n)$, and let X and Y be independent. Then the RV

$$F = \frac{X/m}{Y/n} \tag{22}$$

is said to have a *noncentral F-distribution* with (m, n) d.f. and *noncentrality parameter* δ.

It is shown in Problem 2 that if F has a noncentral F-distribution with (m, n) d.f. and noncentrality parameter δ,

$$EF = \frac{n(m+\delta)}{m(n-2)}, \qquad n > 2,$$

and

$$\operatorname{var}(F) = \frac{2n^2}{m^2(n-4)(n-2)^2}[(m+\delta)^2 + (n-2)(m+2\delta)], \qquad n > 4.$$

PROBLEMS 7.4

1. Let

$$P_x = \left\{\Gamma\left(\frac{n}{2}\right)2^{n/2}\right\}^{-1} \int_0^x \omega^{(n-2)/2} e^{-\omega/2}\, d\omega, \qquad x > 0.$$

Show that

$$x < \frac{n}{1 - P_x}.$$

2. Let $X \sim F(m, n, \delta)$. Find EX and $\text{var}(X)$.

3. Let T be a noncentral t-statistic with n d.f. and noncentrality parameter δ. Find ET and $\text{var}(T)$.

4. Let $F \sim F(m, n)$. Then

$$Y = \left(1 + \frac{m}{n}F\right)^{-1} \sim B\left(\frac{n}{2}, \frac{m}{2}\right).$$

Deduce that for $x > 0$,

$$P\{F \leq x\} = 1 - P\left\{Y \leq \left(1 + \frac{m}{n}x\right)^{-1}\right\}.$$

5. Derive the PDF of an F-statistic with (m, n) d.f.

6. Show that the square of a noncentral t-statistic is a noncentral F-statistic.

7. A sample of size 16 showed a variance of 5.76. Find c such that $P\{|\overline{X} - \mu| < c\} = 0.95$, where $\overline{X}$ is the sample mean and μ is the population mean. Assume that the sample comes from a normal population.

8. A sample from a normal population produced variance 4.0. Find the size of the sample if the sample mean deviates from the population mean by no more than 2.0 with a probability of at least 0.95.

9. Let X_1, X_2, X_3, X_4, X_5 be a sample from $\mathcal{N}(0, 4)$. Find $P\{\sum_{i=1}^{5} X_i^2 \geq 5.75\}$.

10. Let $X \sim \chi^2(61)$. Find $P\{X > 50\}$.

11. Let $F \sim F(m, n)$. The random variable $Z = \frac{1}{2}\log F$ is known as *Fisher's Z-statistic*. Find the PDF of Z.

12. Prove Theorem 1.

13. Prove Theorem 2.

14. Prove Theorem 3.

15. Prove Theorem 4.

16. (a) Let $f_1, f_2, \ldots$ be PDFs with corresponding MGFs $M_1, M_2, \ldots$, respectively. Let α_j $(0 < \alpha_j < 1)$ be constants such that $\sum_{j=1}^{\infty} \alpha_j = 1$. Then $f = \sum_1^{\infty} \alpha_j f_j$ is a PDF with MGF $M = \sum_{j=1}^{\infty} \alpha_j M_j$.

(b) Write the MGF of a $\chi^2(n, \delta)$ RV in (6) as

$$M(t) = \sum_{j=0}^{\infty} \alpha_j M_j(t)$$

where $M_j(t) = (1 - 2t)^{-(2j+n)/2}$ is the MGF of a $\chi^2(2j + n)$ RV and $\alpha_j = e^{-\delta/2}(\delta/2)^j/j!$ is the PMF of a $P(\delta/2)$ RV. Conclude that PDF of $Y \sim$

$\chi^2(n, \delta)$ is the weighted sum of PDFs of $\chi^2(2j + n)$ RVs, $j = 0, 1, 2, \ldots$ with Poisson weights and hence

$$f_Y(y) = \sum_{j=0}^{\infty} \frac{e^{-\delta/2}(\delta/2)^j}{j!} \frac{y^{(2j+n)/2-1} \exp(-y/2)}{2^{(2j+n)/2}\,\Gamma[(2j+n)/2]}.$$

7.5 LARGE-SAMPLE THEORY

In many applications of probability one needs the distribution of a statistic or some function of it. The methods of Section 7.3 when applicable lead to the exact distribution of the statistic under consideration. If not, it may be sufficient to approximate this distribution provided that the sample size is large enough.

Let $\{X_n\}$ be a sequence of RVs that converges in law to $N(\mu, \sigma^2)$. Then $\{(X_n - \mu)/\sigma)\}$ converges in law to $N(0, 1)$, and conversely. We will say alternatively and equivalently that $\{X_n\}$ is *asymptotically normal* with mean μ and variance σ^2. More generally, we say that X_n is *asymptotically normal* with "mean" μ_n and "variance" σ_n^2, and write X_n is $\mathrm{AN}(\mu_n, \sigma_n^2)$, if $\sigma_n > 0$ and as $n \to \infty$,

$$\frac{X_n - \mu_n}{\sigma_n} \xrightarrow{L} \mathcal{N}(0, 1). \tag{1}$$

Here μ_n is not necessarily the mean of X_n and σ_n^2, not necessarily its variance. In this case we can approximate, for sufficiently large n, $P\{X_n \le t\}$ by $P\{Z \le (t-\mu_n)/\sigma_n\}$ where Z is $\mathcal{N}(0, 1)$.

The most common method to show that X_n is $\mathrm{AN}(\mu_n, \sigma_n^2)$ is the central limit theorem of Section 6.6. Thus, according to Theorem 6.6.1, $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{L} \mathcal{N}(0, \sigma^2)$ as $n \to \infty$, where $\overline{X}_n$ is the sample mean of n iid RVs with mean μ and variance σ^2. The same result applies to the kth sample moment, provided that $E|X|^{2k} < \infty$. Thus

$$\sum_{j=1}^{n} \frac{X_n^k}{n} \text{ is AN}\left(EX^k, \frac{\mathrm{var}(X^k)}{n}\right).$$

In many large-sample approximations an application of the CLT along with Slutsky's theorem suffices.

Example 1. Let $X_1, X_2, \ldots$ be iid $\mathcal{N}(\mu, \sigma^2)$. Consider the RV

$$T_n = \frac{\sqrt{n}(\overline{X} - \mu)}{S}.$$

The statistic T_n is well known for its applications in statistics and in Section 7.6 we determine its exact distribution. From Example 6.3.4, $(n-1)S^2/n \xrightarrow{P} \sigma^2$ and hence $S/\sigma \xrightarrow{P} 1$. Since $\sqrt{n}(\overline{X} - \mu)/\sigma \xrightarrow{L} Z \sim \mathcal{N}(0, 1)$, it follows from Slutsky's

theorem that $T_n \xrightarrow{L} Z$. Thus for sufficiently large n ($n \geq 30$), we can approximate $P\{T_n \leq t\}$ by $P\{Z \leq t\}$.

Actually, we do not need X's to be normally distributed (see Problem 6.6.5).

Often, we need to approximate the distribution of $g(Y_n)$ given that Y_n is AN(μ, σ^2).

Theorem 1. Suppose that Y_n is AN(μ, σ_n^2), with $\sigma_n \to 0$ and μ a fixed real number. Let g be a real-valued function that is differentiable at $x = \mu$, with $g'(\mu) \neq 0$. Then

$$g(Y_n) \text{ is AN}\left(g(\mu), [g'(\mu)]^2\sigma_n^2\right). \tag{2}$$

Proof. We first show that

$$\frac{g(Y_n) - g(\mu)}{g'(\mu)\sigma_n} - \frac{Y_n - \mu}{\sigma_n} \xrightarrow{P} 0. \tag{3}$$

Set

$$h(x) = \begin{cases} \dfrac{g(x) - g(\mu)}{x - \mu} - g'(\mu), & x \neq \mu \\ 0, & x = \mu. \end{cases}$$

Then h is continuous at $x = \mu$. Since

$$Y_n - \mu = \sigma_n \frac{Y_n - \mu}{\sigma_n} \xrightarrow{L} 0$$

by Problem 6.2.7, $Y_n - \mu \xrightarrow{P} 0$, and it follows from Theorem 6.2.4 that $h(Y_n) \xrightarrow{P} h(\mu) = 0$. By Slutsky's theorem, therefore,

$$h(Y_n)\frac{Y_n - \mu}{\sigma_n} \xrightarrow{P} 0.$$

That is,

$$\frac{g(Y_n) - g(\mu)}{\sigma_n g'(\mu)} - \frac{Y_n - \mu}{\sigma_n} \xrightarrow{P} 0.$$

It follows again by Slutsky's theorem that $[g(Y_n) - g(\mu)]/[g'(\mu)\sigma_n]$ has the same limit law as $(Y_n - \mu)/\sigma_n$.

Example 2. We know by the CLT theorem that $Y_n = \overline{X}$ is AN$(\mu, \sigma^2/n)$. Suppose that $g(\overline{X}) = \overline{X}(1 - \overline{X})$, where $\overline{X}$ is the sample mean in random sampling from a population with mean μ and variance σ^2. Since $g'(\mu) = 1 - 2\mu \neq 0$ for $\mu \neq \frac{1}{2}$, it follows that for $\mu \neq \frac{1}{2}$, $\sigma^2 < \infty$, $\overline{X}(1 - \overline{X})$ is AN$(\mu(1 - \mu), (1 - 2\mu)^2\sigma^2/n)$.

Thus

$$P\{\overline{X}(1-\overline{X}) \le y\} = P\left\{\frac{\overline{X}(1-\overline{X}) - \mu(1-\mu)}{|1-2\mu|\sigma/\sqrt{n}} \le \frac{y-\mu(1-\mu)}{|1-2\mu|\sigma/\sqrt{n}}\right\}$$
$$\approx \Phi\left(\frac{y-\mu(1-\mu)}{|1-2\mu|\sigma/\sqrt{n}}\right)$$

for large n.

Remark 1. Suppose that g in Theorem 1 is differentiable k times, $k \ge 1$, at $x = \mu$ and $g^{(i)}(\mu) = 0$ for $1 \le i \le k-1$, $g^{(k)}(\mu) \ne 0$. Then a similar argument using Taylor's theorem shows that

$$[g(Y_n) - g(\mu)] \Big/ \left[\frac{1}{k!} g^{(k)}(\mu)\sigma_n^k\right] \xrightarrow{L} Z^k \tag{4}$$

where Z is a $\mathcal{N}(0, 1)$ RV. Thus in Example 2, when $\mu = \frac{1}{2}$, $g'(\frac{1}{2}) = 0$ and $g''(\frac{1}{2}) = -2 \ne 0$. It follows that

$$n[\overline{X}(1-\overline{X}) - \tfrac{1}{4}] \xrightarrow{L} -\sigma^2\chi^2(1)$$

since $Z^2 \stackrel{d}{=} \chi^2(1)$.

Remark 2. Theorem 1 can be extended to the multivariate case, but we will not pursue the development. We refer the reader to Ferguson [26] or Serfling [100].

Remark 3. In general, the asymptotic variance $[g'(\mu)]^2\sigma_n^2$ of $g(Y_n)$ will depend on the parameter μ. In problems of inference it will often be desirable to use transformation g such that the approximate variance $\operatorname{var} g(Y_n)$ is free of the parameter. Such transformations are called *variance stabilizing transformations*. Let us write $\sigma_n^2 = \sigma^2(\mu)/n$. Then finding a g such that $\operatorname{var}\ g(Y_n)$ is free of μ is equivalent to finding a g such that

$$g'(\mu) = \frac{c}{\sigma(\mu)}$$

for all μ, where c is a constant independent of μ. It follows that

$$g(x) = c\int \frac{dx}{\sigma(x)}. \tag{5}$$

Example 3. In Example 2, $\sigma^2(\mu) = \mu(1-\mu)$. Suppose that $X_1, \ldots, X_n$ are iid $b(1, p)$. Then $\sigma^2(p) = p(1-p)$ and (5) reduces to

$$g(x) = c \int \frac{dx}{\sqrt{x(1-x)}} = 2 \arcsin \sqrt{x}.$$

Since $g(0) = 0$, $g(1) = 1$, $c = 2/\pi$ and $g(x) = (2/\pi) \arcsin \sqrt{x}$.

Remark 4. In Section 7.3 we computed exact moments of some statistics in terms of population parameters. Approximations for moments of $g(\overline{X})$ can also be obtained from series expansions of g. Suppose that g is twice differentiable at $x = \mu$. Then

$$Eg(X) \approx g(\mu) + E(X-\mu)g'(\mu) + \tfrac{1}{2}g''(\mu)E(X-\mu)^2 \tag{6}$$

and

$$E[g(X) - g(\mu)]^2 \approx [g'(\mu)]^2 E(X-\mu)^2, \tag{7}$$

by dropping remainder terms. The case of most interest is to approximate $Eg(\overline{X})$ and $\operatorname{var} g(\overline{X})$. In this case, under suitable conditions, one can show that

$$Eg(\overline{X}) \approx g(\mu) + \frac{\sigma^2}{2n} g''(\mu) \tag{8}$$

and

$$\operatorname{var} g(\overline{X}) \approx \frac{\sigma^2}{n} [g'(\mu)]^2 \tag{9}$$

where $EX = \mu$ and $\operatorname{var}(X) = \sigma^2$.

In Example 2, when X_i's are iid $b(1, p)$, $g(x) = x(1-x)$, $g'(x) = 1 - 2x$, $g''(x) = -2$, so that

$$\begin{aligned} Eg(\overline{X}) &\approx E[\overline{X}(1-\overline{X})] \approx p(1-p) + \frac{\sigma^2}{2n}(-2) \\ &= p(1-p)\frac{n-1}{n} \end{aligned}$$

and

$$\operatorname{var} g(\overline{X}) \approx \frac{p(1-p)}{n}(1-2p)^2.$$

In this case we can compute $Eg(\overline{X})$ and $\operatorname{var} g(\overline{X})$ exactly. We have

$$Eg(\overline{X}) = E\overline{X} - E\overline{X}^2 = p - \left[\frac{p(1-p)}{n} + p^2\right] = p(1-p)\frac{n-1}{n},$$

so that (8) is exact. Also, since $X_i^k = X_i$, using Theorem 7.3.4 we have

$$\begin{aligned}\operatorname{var} g(\overline{X}) &= \operatorname{var}(\overline{X} - \overline{X}^2) \\ &= \operatorname{var}(\overline{X}) - 2\operatorname{cov}(\overline{X}, \overline{X}^2) + E\overline{X}^4 - (E\overline{X}^2)^2 \\ &= \frac{p(1-p)}{n}\left[(1-2p)^2 + \frac{2p(1-p)}{n-1}\right]\left(\frac{n-1}{n}\right)^2.\end{aligned}$$

Thus the error in approximation (9) is

$$\text{error} = \frac{2p^2(1-p)^2}{n^3}(n-1).$$

Remark 5. Approximations (6) through (9) do not assert the existence of $Eg(X)$ or $Eg(\overline{X})$, or $\operatorname{var} g(X)$ or $\operatorname{var} g(\overline{X})$.

Remark 6. It is possible to extend (6) through (9) to two (or more) variables by using Taylor series expansion in two (or more) variables.

Finally, we state the following result, which gives the asymptotic distribution of the rth order statistic, $1 \leq r \leq n$, in sampling from a population with an absolutely continuous DF F with PDF f. For a proof, see Problem 4.

Theorem 2. If $X_{(r)}$ denotes the rth-order statistic of a sample $X_1, X_2, \ldots, X_n$ from an absolutely continuous DF F with PDF f, then

$$\left[\frac{n}{p(1-p)}\right]^{1/2} f(\xi_p)\{X_{(r)} - \xi_p\} \xrightarrow{L} Z \qquad \text{as } n \to \infty, \tag{10}$$

so that r/n remains fixed, $r/n = p$, where Z is $\mathcal{N}(0, 1)$, and ξ_p is the unique solution of $F(\xi_p) = p$ (that is, ξ_p is the population quantile of order p assumed unique).

Remark 7. The sample quantile of order p, Z_p, is

$$\text{AN}\left(\xi_p, \frac{1}{[f(\xi_p)]^2}\frac{p(1-p)}{n}\right),$$

where ξ_p is the corresponding population quantile and f is the PDF of the population distribution function. It also follows that $Z_p \xrightarrow{P} \xi_p$.

PROBLEMS 7.5

1. In sampling from a distribution with mean μ and variance σ^2, find the asymptotic distribution of (a) $\overline{X}^2$, (b) $1/\overline{X}$, (c) $\ln|\overline{X}|^2$, and (d) $\exp(\overline{X})$, both when $\mu \neq 0$ and when $\mu = 0$.

2. Let $X \sim P(\lambda)$. Then $(X - \lambda)/\sqrt{\lambda} \xrightarrow{L} \mathcal{N}(0, 1)$. Find a transformation g such that $(g(X) - g(\lambda))$ has an asymptotic $\mathcal{N}(0, c)$ distribution for large μ, where c is a suitable constant.

3. Let $X_1, X_2, \ldots, X_n$ be a sample from an absolutely continuous DF F with PDF f. Show that

$$EX_{(r)} \approx F^{-1}\left(\frac{r}{n+1}\right)$$

and

$$\text{var}(X_{(r)}) \approx \frac{r(n-r+1)}{(n+1)^2(n+2)} \frac{1}{\{f[F^{-1}(r/n+1)]\}^2}.$$

[*Hint*: Let Y be an RV with mean μ, and ϕ be a Borel function such that $E\phi(Y)$ exists. Expand $\phi(Y)$ about the point μ by a Taylor series expansion, and use the fact that $F(X_{(r)}) = U_{(r)}$.]

4. Prove Theorem 7. [*Hint:* For any real μ and σ (> 0), compute the PDF of $(U_{(r)} - \mu)/\sigma$ and show that the standardized $U_{(r)}$, $(U_{(r)} - \mu)/\sigma$, is asymptotically $\mathcal{N}(0, 1)$ under the conditions of the theorem.]

5. Let $X \sim \chi^2(n)$. Then $(X - n)/\sqrt{2n}$ is AN(0, 1) and X/n is AN(1, $2/n$). Find a transformation g such that the distribution of $g(X) - g(n)$ is AN(0, c).

6. Suppose that X is $G(1, \theta)$. Find g such that $g(\overline{X}) - g(\theta)$ is AN(0, c).

7. Let $X_1, X_2, \ldots, X_n$ be iid RVs with $E|X_1|^4 < \infty$. Let $\text{var}(X) = \sigma^2$ and $\beta_2 = \mu_4/\sigma^4$.

   (a) Using the CLT for iid RVs, show that $\sqrt{n}(S^2 - \sigma^2) \xrightarrow{L} \mathcal{N}(0, \mu_4 - \sigma^4)$.

   (b) Find a transformation g such that $g(S^2)$ has an asymptotic distribution that depends on β_2 alone, not on σ^2.

7.6 DISTRIBUTION OF $(\overline{X}, S^2)$ IN SAMPLING FROM A NORMAL POPULATION

Let $X_1, X_2, \ldots, X_n$ be a sample from $\mathcal{N}(\mu, \sigma^2)$, and write $\overline{X} = n^{-1}\sum_{i=1}^n X_i$ and $S^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \overline{X})^2$. In this section we show that $\overline{X}$ and S^2 are independent and derive the distribution of S^2. More precisely, we prove the following important result.

Theorem 1. Let $X_1, X_2, \ldots, X_n$ be iid $\mathcal{N}(\mu, \sigma^2)$ RVs. Then $\overline{X}$ and $(X_1 - \overline{X}, X_2 - \overline{X}, \ldots, X_n - \overline{X})$ are independent.

Proof. We compute the MGF of $\overline{X}$ and $X_1 - \overline{X}, X_2 - \overline{X}, \ldots, X_n - \overline{X}$ as follows:

$$
\begin{aligned}
M(t, t_1, t_2, \ldots, t_n) &= E \exp\{t\overline{X} + t_1(X_1 - \overline{X}) + t_2(X_2 - \overline{X}) + \cdots + t_n(X_n - \overline{X})\} \\
&= E \exp\left[\sum_{i=1}^{n} t_i X_i - \left(\sum_{i=1}^{n} t_i - t\right)\overline{X}\right] \\
&= E \exp\left[\sum_{i=1}^{n} X_i \left(t_i - \frac{t_1 + t_2 + \cdots + t_n - t}{n}\right)\right] \\
&= E\left\{\prod_{i=1}^{n} \exp\left[\frac{X_i(nt_i - n\bar{t} + t)}{n}\right]\right\} \quad \left(\text{where } \bar{t} = n^{-1}\sum_{i=1}^{n} t_i\right) \\
&= \prod_{i=1}^{n} E \exp\left\{\frac{X_i[t + n(t_i - \bar{t})]}{n}\right\} \\
&= \prod_{1}^{n} \exp\left\{\frac{\mu[t + n(t_i - \bar{t})]}{n} + \frac{\sigma^2}{2}\frac{1}{n^2}[t + n(t_i - \bar{t})]^2\right\} \\
&= \exp\left\{\frac{\mu}{n}[nt + n\sum_{i=1}^{n}(t_i - \bar{t})] + \frac{\sigma^2}{2n^2}\sum_{i=1}^{n}[t + n(t_i - \bar{t})]^2\right\} \\
&= \exp(\mu t) \exp\left[\frac{\sigma^2}{2n^2}\left(nt^2 + n^2\sum_{i=1}^{n}(t_i - \bar{t})^2\right)\right] \\
&= \exp\left(\mu t + \frac{\sigma^2}{2n}t^2\right) \exp\left[\frac{\sigma^2}{2}\sum_{i=1}^{n}(t_i - \bar{t})^2\right] \\
&= M_{\overline{X}}(t) M_{X_1 - \overline{X}, \ldots, X_n - \overline{X}}(t_1, t_2, \ldots, t_n) \\
&= M(t, 0, 0, \ldots, 0) M(0, t_1, t_2, \ldots, t_n).
\end{aligned}
$$

Corollary 1. $\overline{X}$ and S^2 are independent.

Corollary 2. $(n-1)S^2/\sigma^2$ is $\chi^2(n-1)$.

Since

$$
\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi^2(n), \qquad n\left(\frac{\overline{X} - \mu}{\sigma}\right)^2 \sim \chi^2(1),
$$

and $\overline{X}$ and S^2 are independent, it follows from

$$\frac{\sum_1^n (X_i - \mu)^2}{\sigma^2} = n\left(\frac{\overline{X} - \mu}{\sigma}\right)^2 + (n-1)\frac{S^2}{\sigma^2}$$

that

$$E\left\{\exp\left[t\sum_1^n \frac{(X_i - \mu)^2}{\sigma^2}\right]\right\} = E\left\{\exp\left[tn\left(\frac{\overline{X} - \mu}{\sigma}\right)^2 + (n-1)\frac{S^2}{\sigma^2}t\right]\right\}$$

$$= E\exp\left[n\left(\frac{\overline{X} - \mu}{\sigma}\right)^2 t\right] E\exp\left[(n-1)\frac{S^2}{\sigma^2}t\right],$$

that is,

$$(1-2t)^{-n/2} = (1-2t)^{-1/2} E\exp\left[(n-1)\frac{S^2}{\sigma^2}t\right], \qquad t < \frac{1}{2},$$

and we see that

$$E\exp\left[(n-1)\frac{S^2}{\sigma^2}t\right] = (1-2t)^{-(n-1)/2}, \qquad t < \frac{1}{2}.$$

By the uniqueness of the MGF it follows that $(n-1)S^2/\sigma^2$ is $\chi^2(n-1)$.

Corollary 3. The distribution of $\sqrt{n}(\overline{X} - \mu)/S$ is $t(n-1)$.

Proof. Since $\sqrt{n}(\overline{X} - \mu)/\sigma$ is $\mathcal{N}(0, 1)$ and $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$, and since $\overline{X}$ and S^2 are independent,

$$\frac{\sqrt{n}\,(\overline{X} - \mu)/\sigma}{\sqrt{[(n-1)S^2/\sigma^2]/(n-1)}} = \frac{\sqrt{n}\,(\overline{X} - \mu)}{S}$$

is $t(n-1)$.

Corollary 4. If $X_1, X_2, \ldots, X_m$ are iid $\mathcal{N}(\mu_1, \sigma_1^2)$ RVs, $Y_1, Y_2, \ldots, Y_n$ are iid $\mathcal{N}(\mu_2, \sigma_2^2)$ RVs, and the two samples are taken independently, $(S_1^2/\sigma_1^2)/(S_2^2/\sigma_2^2)$ is $F(m-1, n-1)$. If, in particular, $\sigma_1 = \sigma_2$, then S_1^2/S_2^2 is $F(m-1, n-1)$.

Corollary 5. Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$, respectively, be independent samples from $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$. Then

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\{[(m-1)S_1^2/\sigma_1^2] + [(n-1)S_2^2/\sigma_2^2]\}^{1/2}} \sqrt{\frac{m+n-2}{\sigma_1^2/m + \sigma_2^2/n}} \sim t(m+n-2).$$

In particular, if $\sigma_1 = \sigma_2$, then

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{[(m-1)S_1^2 + (n-1)S_2^2]}} \sqrt{\frac{mn(m+n-2)}{m+n}} \sim t(m+n-2).$$

Corollary 5 follows since

$$\overline{X} - \overline{Y} \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)$$

and

$$\frac{(m-1)S_1^2}{\sigma_1^2} + \frac{(n-1)S_2^2}{\sigma_2^2} \sim \chi^2(m+n-2)$$

and the two statistics are independent.

Remark 1. The converse of Corollary 1 also holds (see Theorem 5.3.28).

Remark 2. In sampling from a symmetric distribution, $\overline{X}$ and S^2 are uncorrelated (see Problem 4.5.14).

Remark 3. Alternatively, Corollary 1 could have been derived from Corollary 2 to Theorem 5.4.6 by using the Helmert orthogonal matrix:

$$\mathbf{A} = \begin{bmatrix} 1/\sqrt{n} & 1/\sqrt{n} & 1/\sqrt{n} & \cdots & 1/\sqrt{n} \\ -1/\sqrt{2} & 1/\sqrt{2} & 0 & \cdots & 0 \\ -1/\sqrt{6} & -1/\sqrt{6} & 2/\sqrt{6} & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & 0 \\ -1/\sqrt{n(n-1)} & -1/\sqrt{n(n-1)} & -1/\sqrt{n(n-1)} & \cdots & (n-1)/\sqrt{n(n-1)} \end{bmatrix}$$

For the case of $n = 3$ this was done in Example 4.4.6. In Problem 7 the reader is asked to work out the details in the general case.

Remark 4. An analytic approach to the development of the distribution of $\overline{X}$ and S^2 is as follows. Assuming without loss of generality that X_i is $\mathcal{N}(0, 1)$, we have as the joint PDF of $(X_1, X_2, \ldots, X_n)$

$$f(x_1, x_2, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{j=1}^{n} x_j^2\right)$$

$$= \frac{1}{(2\pi)^{n/2}} \exp\left[-\frac{(n-1)s^2 + n\overline{x}^2}{2}\right].$$

Changing the variables to $y_1, y_2, \ldots, y_n$ by using the transformation $y_k = (x_k - \overline{x})/s$, we see that

$$\sum_{k=1}^{n} y_k = 0 \quad \text{and} \quad \sum_{k=1}^{n} y_k^2 = n - 1.$$

It follows that two of the y_k's, say y_{n-1} and y_n are functions of the remaining y_k. Thus either

$$y_{n-1} = \frac{\alpha + \beta}{2} \quad \text{and} \quad y_n = \frac{\alpha - \beta}{2},$$

or

$$y_{n-1} = \frac{\alpha - \beta}{2} \quad \text{and} \quad y_n = \frac{\alpha + \beta}{2},$$

where

$$\alpha = -\sum_{k=1}^{n-2} y_k \quad \text{and} \quad \beta = \sqrt{2(n-1) - 2\sum_{k=1}^{n-2} y_k^2 - \left(\sum_{k=1}^{n-2} y_k\right)^2}.$$

We leave the reader to derive the joint PDF of $(Y_1, Y_2, \ldots, Y_{n-2}, \overline{X}, S^2)$, using the result described in Remark 4.4.2 and to show that the RVs $\overline{X}$, S^2 and $(Y_1, Y_2, \ldots, Y_{n-2})$ are independent.

PROBLEMS 7.6

1. Let $X_1, X_2, \ldots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$ and $\overline{X}$ and S^2, respectively, be the sample mean and the sample variance. Let $X_{n+1} \sim \mathcal{N}(\mu, \sigma^2)$, and assume that $X_1, X_2, \ldots, X_n, X_{n+1}$ are independent. Find the sampling distribution of $[(X_{n+1} - \overline{X})/S]\sqrt{n/(n+1)}$.

2. Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be independent random samples from $\mathcal{N}(\mu_1, \sigma^2)$ and $\mathcal{N}(\mu_2, \sigma^2)$, respectively. Also, let α, β be two fixed real numbers. If $\overline{X}, \overline{Y}$ denote the corresponding sample means, what is the sampling distribution of

$$\frac{\alpha(\overline{X} - \mu_1) + \beta(\overline{Y} - \mu_2)}{\sqrt{\dfrac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2}}\sqrt{\dfrac{\alpha^2}{m} + \dfrac{\beta^2}{n}}},$$

where S_1^2 and S_2^2, respectively, denote the sample variances of the X's and the Y's?

3. Let $X_1, X_2, \ldots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$ and k be a positive integer. Find $E(S^{2k})$. In particular, find $E(S^2)$ and $\text{var}(S^2)$.

4. A random sample of 5 is taken from a normal population with mean 2.5 and variance $\sigma^2 = 36$.
   (a) Find the probability that the sample variance lies between 30 and 44.
   (b) Find the probability that the sample mean lies between 1.3 and 3.5, while the sample variance lies between 30 and 44.

5. The mean life of a sample of 10 light bulbs was observed to be 1327 hours with a standard deviation of 425 hours. A second sample of 6 bulbs chosen from a different batch showed a mean life of 1215 hours with a standard deviation of 375 hours. If the means of the two batches are assumed to be same, how probable is the observed difference between the two sample means?

6. Let S_1^2 and S_2^2 be the sample variances from two independent samples of sizes $n_1 = 5$ and $n_2 = 4$ from two populations having the same unknown variance σ^2. Find (approximately) the probability that $S_1^2/S_2^2 < 1/5.2$ or > 6.25.

7. Let $X_1, X_2, \ldots, X_n$ be a sample from $\mathcal{N}(\mu, \sigma^2)$. By using the Helmert orthogonal transformation defined in Remark 3, show that $\overline{X}$ and S^2 are independent.

8. Derive the joint PDF of $\overline{X}$ and S^2 by using the transformation described in Remark 4.

7.7 SAMPLING FROM A BIVARIATE NORMAL DISTRIBUTION

Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be a sample from a bivariate normal population with parameters $\mu_1, \mu_2, \rho, \sigma_1^2, \sigma_2^2$. Let us write

$$\overline{X} = n^{-1}\sum_{i=1}^{n} X_i, \qquad \overline{Y} = n^{-1}\sum_{i=1}^{n} Y_i,$$

$$S_1^2 = (n-1)^{-1}\sum_{i=1}^{n}(X_i - \overline{X})^2, \qquad S_2^2 = (n-1)^{-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2,$$

and

$$S_{11} = (n-1)^{-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}).$$

In this section we show that $(\overline{X}, \overline{Y})$ is independent of (S_1^2, S_{11}, S_2^2) and obtain the distribution of the sample correlation coefficient and regression coefficients (at least in the special case where $\rho = 0$).

Theorem 1. The random vectors $(\overline{X}, \overline{Y})$ and $(X_1 - \overline{X}, X_2 - \overline{X}, \ldots, X_n - \overline{X}, Y_1 - \overline{Y}, Y_2 - \overline{Y}, \ldots, Y_n - \overline{Y})$ are independent. The joint distribution of $(\overline{X}, \overline{Y})$ is bivariate normal with parameters $\mu_1, \mu_2, \rho, \sigma_1^2/n, \sigma_2^2/n$.

Proof. The proof follows along the lines of the proof of Theorem 7.6.1. The MGF of $(\overline{X}, \overline{Y}, X_1 - \overline{X}, \ldots, X_n - \overline{X}, Y_1 - \overline{Y}, \ldots, Y_n - \overline{Y})$ is given by

$$\begin{aligned} M^* &= M(u, v, t_1, t_2, \ldots, t_n, s_1, s_2, \ldots, s_n) \\ &= E \exp\left[u\overline{X} + v\overline{Y} + \sum_{i=1}^{n} t_i(X_i - \overline{X}) + \sum_{i=1}^{n} s_i(Y_i - \overline{Y})\right] \\ &= E \exp\left[\sum_{i=1}^{n} X_i\left(\frac{u}{n} + t_i - \bar{t}\right) + \sum_{i=1}^{n} Y_i\left(\frac{v}{n} + s_i - \bar{s}\right)\right], \end{aligned}$$

where $\bar{t} = n^{-1}\sum_{i=1}^{n} t_i$, $\bar{s} = n^{-1}\sum_{i=1}^{n} s_i$. Therefore,

$$\begin{aligned} M^* &= \prod_{i=1}^{n} E \exp\left[\left(\frac{u}{n} + t_i - \bar{t}\right) X_i + \left(\frac{v}{n} + s_i - \bar{s}\right) Y_i\right] \\ &= \prod_{i=1}^{n} \exp\left\{\left(\frac{u}{n} + t_i - \bar{t}\right)\mu_1 + \left(\frac{v}{n} + s_i - \bar{s}\right)\mu_2 \right. \\ &\quad + \frac{\sigma_1^2[(u/n) + t_i - \bar{t}]^2 + 2\rho\sigma_1\sigma_2[(u/n) + t_i - \bar{t}][(v/n) + s_i - \bar{s}]}{2} \\ &\quad \left. \frac{+ \sigma_2^2[(v/n) + s_i - \bar{s}]^2}{2}\right\} \\ &= \exp\left(\mu_1 u + \mu_2 v + \frac{u^2\sigma_1^2 + 2\rho\sigma_1\sigma_2 uv + v^2\sigma_2^2}{2n}\right) \\ &\quad \cdot \exp\left[\frac{1}{2}\sigma_1^2\sum_{i=1}^{n}(t_i - \bar{t})^2 + \rho\sigma_1\sigma_2\sum_{i=1}^{n}(t_i - \bar{t})(s_i - \bar{s}) + \frac{1}{2}\sigma_2^2\sum_{i=1}^{2}(s_i - \bar{s})^2\right] \\ &= M_1(u, v)M_2(t_1, t_2, \ldots, t_n, s_1, s_2, \ldots, s_n) \end{aligned}$$

for all real $u, v, t_1, t_2, \ldots, t_n, s_1, s_2, \ldots, s_n$, where M_1 is the MGF of $(\overline{X}, \overline{Y})$ and M_2 is the MGF of $(X_1 - \overline{X}, \ldots, X_n - \overline{X}, Y_1 - \overline{Y}, \ldots, Y_n - \overline{Y})$. Also, M_1 is the MGF of a bivariate normal distribution. This completes the proof.

Corollary. The sample mean vector $(\overline{X}, \overline{Y})$ is independent of the sample variance–covariance matrix $\begin{pmatrix} s_1^2 & s_{11} \\ s_{11} & s_2^2 \end{pmatrix}$ in sampling from a bivariate normal population.

Remark 1. The result of Theorem 1 can be generalized to the case of sampling from a k-variate normal population. We do not propose to do so here.

Remark 2. Unfortunately, the method of proof of Theorem 1 does not lead to the distribution of the variance–covariance matrix. The distribution of $(\overline{X}, \overline{Y}, S_1^2, S_{11}, S_2^2)$ was found by Fisher [27] and Romanovsky [90]. The general case is due to Wishart [118], who determined the distribution of the sample variance–covariance matrix in sampling from a k-dimensional normal distribution. The distribution is named after him.

We will next compute the distribution of the sample correlation coefficient:

$$R = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\left[\sum_{i=1}^n (X_i - \overline{X})^2 \sum_{i=1}^n (Y_i - \overline{Y})^2\right]^{1/2}} = \frac{S_{11}}{S_1 S_2}. \tag{1}$$

It is convenient to introduce the sample regression coefficient of Y on X

$$B_{Y|X} = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} = \frac{S_{11}}{S_1^2} = R\frac{S_2}{S_1}. \tag{2}$$

Since we will need only the distribution of R and $B_{Y|X}$ whenever $\rho = 0$, we make this simplifying assumption in what follows. The general case is computationally quite complicated. We refer the reader to Cramér [16] for details.

We note that

$$R = \frac{\sum_{i=1}^n Y_i (X_i - \overline{X})}{(n-1) S_1 S_2} \tag{3}$$

and

$$B_{Y|X} = \frac{\sum_{i=1}^n Y_i (X_i - \overline{X})}{(n-1) S_1^2}. \tag{4}$$

Moreover,

$$R^2 = \frac{B_{Y|X}^2 S_1^2}{S_2^2}. \tag{5}$$

In the following we write $B = B_{Y|X}$.

Theorem 2. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$, $n \geq 2$, be a sample from a bivariate normal population with parameters $EX = \mu_1$, $EY = \mu_2$, $\operatorname{var}(X) = \sigma_1^2$, $\operatorname{var}(Y) = \sigma_2^2$, and $\operatorname{cov}(X, Y) = 0$. In other words, let $X_1, X_2, \ldots, X_n$ be iid $\mathcal{N}(\mu_1, \sigma_1^2)$ RVs, and $Y_1, Y_2, \ldots, Y_n$ be iid $\mathcal{N}(\mu_2, \sigma_2^2)$ RVs, and suppose that the X's and Y's are independent. Then the PDF of R is given by

$$(6)\qquad f_1(r) = \begin{cases} \dfrac{\Gamma[(n-1)/2]}{\Gamma(\frac{1}{2})\Gamma[(n-2)/2]}(1-r^2)^{(n-4)/2}, & -1 \le r \le 1, \\ 0, & \text{otherwise;} \end{cases}$$

and the PDF of B is given by

$$(7)\qquad h_1(b) = \frac{\Gamma(n/2)}{\Gamma(\frac{1}{2})\Gamma[(n-1)/2]} \frac{\sigma_1\sigma_2^{n-1}}{(\sigma_2^2 + \sigma_1^2 b^2)^{n/2}}, \qquad -\infty < b < \infty.$$

Proof. Without any loss of generality, we assume that $\mu_1 = \mu_2 = 0$ and $\sigma_1^2 = \sigma_2^2 = 1$, for we can always define

$$(8)\qquad X_i^* = \frac{X_i - \mu_1}{\sigma_1} \qquad \text{and} \qquad Y_i^* = \frac{Y_i - \mu_2}{\sigma_2}.$$

Now note that the conditional distribution of Y_i, given $X_1, X_2, \ldots, X_n$, is $\mathcal{N}(0, 1)$, and $Y_1, Y_2, \ldots, Y_n$, given $X_1, X_2, \ldots, X_n$, are mutually independent. Let us define the following orthogonal transformation:

$$(9)\qquad u_i = \sum_{j=1}^{n} c_{ij} y_j, \qquad i = 1, 2, \ldots, n,$$

where $((c_{ij}))_{i,j=1,2,\ldots,n}$ is an orthogonal matrix with the first two rows

$$(10)\qquad c_{1j} = \frac{1}{\sqrt{n}}, \qquad j = 1, 2, \ldots, n,$$

and

$$(11)\qquad c_{2j} = \frac{x_j - \bar{x}}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{1/2}}, \qquad j = 1, 2, \ldots, n.$$

It follows from orthogonality that for any $i \ge 2$,

$$(12)\qquad \sum_{j=1}^{n} c_{ij} = \sqrt{n} \sum_{j=1}^{n} c_{ij} \frac{1}{\sqrt{n}} = \sqrt{n} \sum_{j=1}^{n} c_{ij} c_{1j} = 0$$

and

$$(13)\qquad \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} c_{ij} y_j \sum_{j'=1}^{n} c_{ij'} y_{j'} \right)$$

$$= \sum_{j=1}^{n} \sum_{j'=1}^{n} \left(\sum_{i=1}^{n} c_{ij} c_{ij'} \right) y_j y_{j'} = \sum_{j=1}^{n} y_{j'}^2.$$

Moreover,

$$u_1 = \sqrt{n}\,\bar{y} \tag{14}$$

and

$$u_2 = b\sqrt{\sum (x_i - \bar{x})^2}, \tag{15}$$

where b is a value assumed by RV B. Also, $U_1, U_2, \ldots, U_n$, given $X_1, X_2, \ldots, X_n$, are normal RVs (being linear combinations of the Y's). Thus

$$\begin{aligned} E\{U_i \mid X_1, X_2, \ldots, X_n\} &= \sum_{j=1}^{n} c_{ij} E\{Y_j \mid X_1, X_2, \ldots, X_n\} \\ &= 0 \end{aligned} \tag{16}$$

and

$$\begin{aligned} \operatorname{cov}\{U_i, U_k \mid X_1, X_2, \ldots, X_n\} &= \operatorname{cov}\left[\sum_{j=1}^{n} c_{ij} Y_j, \sum_{p=1}^{n} c_{kp} Y_p \mid X_1, X_2, \ldots, X_n\right] \\ &= \sum_{j=1}^{n} \sum_{p=1}^{n} c_{ij} c_{kp} \operatorname{cov}\{Y_j, Y_p \mid X_1, X_2, \ldots, X_n\} \\ &= \sum_{j=1}^{n} c_{ij} c_{kj}. \end{aligned}$$

This last equality follows since

$$\operatorname{cov}\{Y_j, Y_p \mid X_1, X_2, \ldots, X_n\} = \begin{cases} 0, & j \neq p, \\ 1, & j = p. \end{cases}$$

From orthogonality, we have

$$\operatorname{cov}\{U_i, U_k \mid X_1, X_2, \ldots, X_n\} = \begin{cases} 0, & i \neq k, \\ 1, & i = k; \end{cases} \tag{17}$$

and it follows that the RVs $U_1, U_2, \ldots, U_n$, given $X_1, X_2, \ldots, X_n$, are mutually independent $\mathcal{N}(0, 1)$. Now

$$\sum_{j=1}^{n} (y_j - \bar{y}^2) = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 \tag{18}$$

$$= \sum_{j=1}^{n} u_j^2 - u_1^2$$

$$= \sum_{j=2}^{n} u_j^2.$$

Thus

$$R^2 = \frac{U_2^2}{\sum_{i=2}^{n} U_i^2} = \frac{U_2^2}{U_2^2 + \sum_{i=3}^{n} U_i^2}. \tag{19}$$

Writing $U = U_2^2$ and $W = \sum_{i=3}^{n} U_i^2$, we see that the conditional distribution of U, given $X_1, X_2, \ldots, X_n$, is $\chi^2(1)$, and that of W, given $X_1, X_2, \ldots, X_n$, is $\chi^2(n-2)$. Moreover, U and W are independent. Since these conditional distributions do not involve the X's, we see that U and W are unconditionally independent with $\chi^2(1)$ and $\chi^2(n-2)$ distributions, respectively. The joint PDF of U and W is

$$f(u, w) = \frac{1}{\Gamma(\frac{1}{2})\sqrt{2}} u^{1/2-1} e^{-u/2} \frac{1}{\Gamma[(n-2)/2] 2^{(n-2)/2}} w^{(n-2)/2-1} e^{-w/2}.$$

Let $u + w = z$; then $u = r^2 z$ and $w = z(1 - r^2)$. The Jacobian of this transformation is z, so that the joint PDF of R^2 and Z is given by

$$f^*(r^2, z) = \frac{1}{\Gamma(\frac{1}{2})\Gamma[(n-2)/2] 2^{(n-1)/2}} z^{n/2-3/2} e^{-z/2} (r^2)^{-1/2} (1 - r^2)^{n/2-2}.$$

The marginal PDF of R^2 is easily computed as

$$f_1^*(r^2) = \frac{\Gamma[(n-1)/2]}{\Gamma(\frac{1}{2})\Gamma[(n-2)/2]} (r^2)^{-1/2} (1 - r^2)^{n/2-2}, \qquad 0 \le r^2 \le 1. \tag{20}$$

Finally, using Theorem 2.5.4, we get the PDF of R as

$$f_1(r) = \frac{\Gamma[(n-1)/2]}{\Gamma(\frac{1}{2})\Gamma[(n-2)/2]} (1 - r^2)^{n/2-2}, \qquad -1 \le r \le 1.$$

As for the distribution of B, note that the conditional PDF of $U_2 = \sqrt{n-1}\, BS_1$, given $X_1, X_2, \ldots, X_n$, is $\mathcal{N}(0, 1)$, so that the conditional PDF of B, given $X_1, X_2, \ldots, X_n$, is $\mathcal{N}(0, 1/\sum(x_i - \bar{x})^2)$. Let us write $\Lambda = (n-1)S_1^2$. Then the PDF of RV Λ is that of a $\chi^2(n-1)$ RV. Thus the joint PDF of B and Λ is given by

$$h(b, \lambda) = g(b \mid \lambda) h_2(\lambda), \tag{21}$$

where $g(b \mid \lambda)$ is $\mathcal{N}(0, 1/\lambda)$, and $h_2(\lambda)$ is $\chi^2(n-1)$. We have

$$h_1(b) = \int_0^\infty h(b, \lambda)\, d\lambda \tag{22}$$

$$= \frac{1}{2^{n/2}\Gamma(\frac{1}{2})\Gamma[(n-1)/2]} \int_0^\infty \lambda^{n/2-1} e^{-\lambda/2(1+b^2)}\, d\lambda$$

$$= \frac{\Gamma(n/2)}{\Gamma(\frac{1}{2})\Gamma[(n-1)/2]} \frac{1}{(1+b^2)^{n/2}}, \qquad -\infty < b < \infty.$$

To complete the proof let us write

$$X_i = \mu_1 + X_i^* \sigma_1 \quad \text{and} \quad Y_i = \mu_2 + Y_i^* \sigma_2,$$

where $X_i^* \sim \mathcal{N}(0, 1)$ and $Y_i^* \sim \mathcal{N}(0, 1)$. Then $X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y_i \sim \mathcal{N}(\mu_2, \sigma_2^2)$, and

$$R = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2 \sum_{i=1}^n (Y_i - \overline{Y})^2}} \tag{23}$$

$$= R^*,$$

so that the PDF of R is the same as derived above. Also,

$$B = \frac{\sigma_1 \sigma_2 \sum_{i=1}^n (X_i^* - \overline{X}^*)(Y_i^* - \overline{Y}^*)}{\sigma_1^2 \sum_{i=1}^n (X_i^* - \overline{X}^*)^2} \tag{24}$$

$$= \frac{\sigma_2}{\sigma_1} B^*,$$

where the PDF of B^* is given by (22). Relations (23) and (24) are used to find the PDF of B. We leave the reader to carry out these simple details.

Remark 3. In view of (23), namely the invariance of R under translation and (positive) scale changes, we note that for fixed n the sampling distribution of R, under $\rho = 0$, does not depend on μ_1, μ_2, σ_1, and σ_2. In the general case when $\rho \neq 0$, one can show that for fixed n the distribution of R depends only on ρ but not on μ_1, μ_2, σ_1, and σ_2 (see, for example, Cramér [16, p. 398]).

Remark 4. Let us change the variable to

$$T = \frac{R}{\sqrt{1 - R^2}} \sqrt{n-2}. \tag{25}$$

Then

$$1 - R^2 = \left(1 + \frac{T^2}{n-2}\right)^{-1},$$

and the PDF of T is given by

$$p(t) = \frac{1}{\sqrt{n-2}} \frac{1}{B[(n-2)/2, \frac{1}{2}]} \frac{1}{[1 + t^2/(n-2)]^{(n-1)/2}}, \tag{26}$$

which is the PDF of a t-statistic with $n-2$ d.f. Thus T defined by (25) has a $t(n-2)$ distribution, provided that $\rho = 0$. This result facilitates the computation of probabilities under the PDF of R when $\rho = 0$.

Remark 5. To compute the PDF of $B_{X|Y} = R(S_1/S_2)$, the *sample regression coefficient* of X on Y, all we need to do is to interchange σ_1 and σ_2 in (7).

Remark 6. From (7) we can compute the mean and variance of B. For $n > 2$, clearly,

$$EB = 0,$$

and for $n > 3$, we can show that

$$EB^2 = \text{var}(B) = \frac{\sigma_2^2}{\sigma_1^2} \frac{1}{n-3}.$$

Similarly, we can use (6) to compute the mean and variance of R. We have, for $n > 4$, under $\rho = 0$,

$$ER = 0$$

and

$$ER^2 = \text{var}(R) = \frac{1}{n-1}.$$

PROBLEMS 7.7

1. Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be a random sample from a bivariate normal population with $EX = \mu_1$, $EY = \mu_2$, $\text{var}(X) = \text{var}(Y) = \sigma^2$, and $\text{cov}(X, Y) = \rho\sigma^2$. Let $\overline{X}, \overline{Y}$ denote the corresponding sample means, S_1^2, S_2^2, the corresponding sample variances, and S_{11}, the sample covariance. Write $R =$

$2S_{11}/(S_1^2 + S_2^2)$. Show that the PDF of R is given by

$$f(r) = \frac{\Gamma(n/2)}{\sqrt{\pi}\,\Gamma[(n-1)/2]}(1-\rho^2)^{(n-1)/2}(1-\rho r)^{-(n-1)}(1-r^2)^{(n-3)/2},$$
$$|r| < 1.$$

[*Hint:* Let $U = (X + Y)/2$, and $V = (X - Y)/2$, and observe that the random vector (U, V) is also bivariate normal. In fact, U and V are independent.] (Rastogi [87])

2. Let X and Y be independent normal RVs. A sample of $n = 11$ observations on (X, Y) produces sample correlation coefficient $r = 0.40$. Find the probability of obtaining a value of R that exceeds the observed value.

3. Let X_1, X_2 be jointly normally distributed with zero means, unit variances, and correlation coefficient ρ. Let S be a $\chi^2(n)$ RV that is independent of (X_1, X_2). Then the joint distribution of $Y_1 = X_1/\sqrt{S/n}$ and $Y_2 = X_2/\sqrt{S/n}$ is known as a *central bivariate t-distribution.* Find the joint PDF of (Y_1, Y_2) and the marginal PDFs of Y_1 and Y_2, respectively.

4. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample from a bivariate normal distribution with parameters $EX_1 = \mu_1$, $EY_i = \mu_2$, $\operatorname{var}(X_i) = \operatorname{var}(Y_i) = \sigma^2$, and $\operatorname{cov}(X_i, Y_i) = \rho\sigma^2$, $i = 1, 2, \ldots, n$. Find the distribution of the statistic

$$T(\mathbf{X}, \mathbf{Y}) = \sqrt{n}\frac{(\overline{X} - \mu_1) - (\overline{Y} - \mu_2)}{\sqrt{\sum_{i=1}^{n}(X_i - Y_i - \overline{X} + \overline{Y})^2}}.$$