CHAPTER 13

Nonparametric Statistical Inference

13.1 INTRODUCTION

In all the problems of statistical inference considered so far, we assumed that the distribution of the random variable being sampled is known except, perhaps, for some parameters. In practice, however, the functional form of the distribution is seldom, if ever, known. It is therefore desirable to devise methods that are free of this assumption concerning distribution. In this chapter we study some procedures that are commonly referred to as *distribution-free* or *nonparametric methods*. The term *distribution-free* refers to the fact that no assumptions are made about the underlying distribution except that the distribution function being sampled is absolutely continuous. The term *nonparametric* refers to the fact that there are no parameters involved in the traditional sense of the term *parameter* used thus far. To be sure, there is a parameter that indexes the family of absolutely continuous DFs, but it is not numerical, and hence the parameter set cannot be represented as a subset of $\mathcal{R}_n$, for any $n \geq 1$. The restriction to absolutely continuous distribution functions is a simplifying assumption that allows us to use the probability integral transformation (Theorem 5.3.1) and the fact that ties occur with probability 0.

Section 13.2 is devoted to the problem of unbiased (nonparametric) estimation. We develop the theory of U-statistics since many estimators and test statistics may be viewed as U-statistics. Sections 13.3 through 13.5 deal with some common hypothesis-testing problems. In Section 13.6 we investigate applications of order statistics in nonparametric methods. Section 13.7 considers underlying assumptions in some common parametric problems and the effect of relaxing these assumptions.

13.2 *U*-STATISTICS

In Chapter 7 we encountered several nonparametric estimators. For example, the empirical DF defined in Section 7.3 as an estimator of the population DF is distribution free, and so also are the sample moments as estimators of the population moments. These are examples of what are known as *U-statistics*, which lead to unbiased estimators of population characteristics. In this section we study the general theory of

U-statistics. Although the thrust of this investigation is unbiased estimation, many of the U-statistics defined in this section may be used as test statistics.

Let $X_1, X_2, \dots, X_n$ be iid RVs with common law $\mathcal{L}(X)$, and let $\mathcal{P}$ be the class of all possible distributions of X that consists of the absolutely continuous or discrete distributions, or subclasses of these.

Definition 1. A statistic $T(\mathbf{X})$ is *sufficient* for the family of distributions $\mathcal{P}$ if the conditional distribution of $\mathbf{X}$, given $T = t$, is the same whatever the true $F \in \mathcal{P}$.

Example 1. Let $X_1, X_2, \dots, X_n$ be a random sample from an absolutely continuous DF, and let $\mathbf{T} = (X_{(1)}, \dots, X_{(n)})$ be the order statistic. Then

$$f(\mathbf{x} \mid \mathbf{T} = \mathbf{t}) = (n!)^{-1},$$

and we see that $\mathbf{T}$ is sufficient for the family of absolutely continuous distributions on $\mathcal{R}$.

Definition 2. A family of distributions $\mathcal{P}$ is *complete* if the only unbiased estimator of 0 is the zero function itself, that is,

$$E_F h(\mathbf{X}) = 0 \qquad \text{for all } F \in \mathcal{P} \Rightarrow h(\mathbf{x}) = 0$$

for all $\mathbf{x}$ (except for a null set with respect to each $F \in \mathcal{P}$).

Definition 3. A statistic $T(\mathbf{X})$ is said to be *complete in relation to a class of distributions* $\mathcal{P}$ if the class of induced distributions of T is complete.

We have already encountered many examples of complete statistics or complete families of distributions in Chapter 8.

The following result is stated without proof. For the proof we refer to Fraser [29, pp. 27–30, 139–142].

Theorem 1. The order statistic $(X_{(1)}, X_{(2)}, \dots, X_{(n)})$ is a complete sufficient statistic provided that the iid RVs $X_1, X_2, \dots, X_n$ are of either the discrete or continuous type.

Definition 4. A real-valued parameter $g(F)$ is said to be *estimable* if it has an unbiased estimator, that is, if there exists a statistic $T(\mathbf{X})$ such that

$$E_F T(\mathbf{X}) = g(F) \qquad \text{for all } F \in \mathcal{P}. \tag{1}$$

Example 2. If $\mathcal{P}$ is the class of all distributions for which the second moment exists, $\overline{X}$ is an unbiased estimator of $\mu(F)$, the population mean. Similarly, $\mu_2(F) = \operatorname{var}_F(X)$ is also estimable, and an unbiased estimator is $S^2 = \sum_1^n (X_i - \overline{X})^2/(n-1)$. We would like to know whether $\overline{X}$ and S^2 are UMVUEs. Similarly, $F(x)$ and $P_F(X_1 + X_2 > 0)$ are estimable for $F \in \mathcal{P}$.

Definition 5. The *degree* m $(m \geq 1)$ of an estimable parameter $g(F)$ is the smallest sample size for which the parameter is estimable; that is, it is the smallest m such that there exists an unbiased estimator $T(X_1, X_2, \ldots, X_m)$ with

$$E_F T = g(F) \qquad \text{for all } F \in \mathcal{P}.$$

Example 3. The parameter $g(F) = P_F\{X > c\}$, where c is a known constant, has degree 1. Also, $\mu(F)$ is estimable with degree 1 [we assume that there is at least one $F \in \mathcal{P}$ such that $\mu(F) \neq 0$], and $\mu_2(F)$ is estimable with degree $m = 2$, since $\mu_2(F)$ cannot be estimated (unbiasedly) by one observation only. At least two observations are needed. Similarly, $\mu^2(F)$ has degree 2, and $P(X_1 + X_2 > 0)$ also is of degree 2.

Definition 6. An unbiased estimator of a parameter based on the smallest sample size (equal to degree m) is called a *kernel.*

Example 4. Clearly, X_i $1 \leq i \leq n$ is a kernel of $\mu(F)$; $T(X_i) = 1$, if $X_i > c$, and $= 0$ if $X_i \leq c$ is a kernal of $P(X > c)$. Similarly, $T(X_i, X_j) = 1$ if $X_i + X_j > 0$, and $= 0$ otherwise is a kernel of $P(X_i + X_j > 0)$, $X_i X_j$ is a kernel of $\mu^2(F)$ and $X_i^2 - X_i X_j$ is a kernel of $\mu_2(F)$.

Lemma 1. There exists a symmetric kernel for every estimable parameter.

Proof. If $T(X_1, X_2, \ldots, X_m)$ is a kernel of $g(F)$, so also is

$$T_s(X_1, X_2, \ldots, X_m) = \frac{1}{m!} \sum_P T(X_{i_1}, X_{i_2}, \ldots, X_{i_m}), \tag{2}$$

where the summation P is over all $m!$ permutations of $\{1, 2, \ldots, m\}$.

Example 5. A symmetric kernel for $\mu_2(F)$ is

$$\begin{aligned} T_s(X_i, X_j) &= \tfrac{1}{2}\{T(X_i, X_j) + T(X_j, X_i)\} \\ &= \tfrac{1}{2}(X_i - X_j)^2, \qquad i, j = 1, 2, \ldots, n \ (i \neq j). \end{aligned}$$

Definition 7. Let $g(F)$ be an estimable parameter of degree m, and let $X_1, X_2, \ldots, X_n$ be a sample of size n, $n \geq m$. Corresponding to any kernel $T(X_{i_1}, \ldots, X_{i_m})$ of $g(F)$, we define a *U-statistic* for the sample by

$$U(X_1, X_2, \ldots, X_n) = \binom{n}{m}^{-1} \sum_C T_s(X_{i_1}, \ldots, X_{i_m}), \tag{3}$$

where the summation C is over all $\binom{n}{m}$ combinations of m integers $(i_1, i_2, \ldots, i_m)$ chosen from $\{1, 2, \ldots, n\}$, and T_s is the symmetric kernel defined in (2).

Clearly, the U-statistic defined in (3) is symmetric in the X_i's, and

$$E_F U(\mathbf{X}) = g(F) \qquad \text{for all } F. \tag{4}$$

Moreover, $U(\mathbf{X})$ is a function of the complete sufficient statistic $X_{(1)}, X_{(2)}, \dots, X_{(n)}$. It follows from Theorem 8.4.6 that it is UMVUE of its expected value.

Example 6. For estimating $\mu(F)$, the U-statistic is $n^{-1}\sum_1^n X_i$. For estimating $\mu_2(F)$, a symmetric kernel is

$$T_s(X_{i_1}, X_{i_2}) = \tfrac{1}{2}(X_{i_1} - X_{i_2})^2, \qquad i_1 = 1, 2, \dots, n \ (i_1 \neq i_2),$$

so that the corresponding U-statistic is

$$\begin{aligned} U(\mathbf{X}) &= \binom{n}{2}^{-1} \sum_{i_1 < i_2} \frac{1}{2}(X_{i_1} - X_{i_2})^2 \\ &= \frac{1}{n-1}\sum_1^n (X_i - \overline{X})^2 \\ &= S^2. \end{aligned}$$

Similarly, for estimating $\mu^2(F)$, a symmetric kernel is $T_s(X_{i_1}, X_{i_2}) = X_{i_1}X_{i_2}$, and the corresponding U-statistic is

$$U(\mathbf{X}) = \frac{1}{\binom{n}{2}} \sum_{i<j} X_i X_j = \frac{1}{n(n-1)} \sum_{i \neq j} X_i X_j.$$

For estimating $\mu^3(F)$, a symmetric kernel is $T_s(X_{i_1}, X_{i_2}, X_{i_3}) = X_{i_1}X_{i_2}X_{i_3}$, so that the corresponding U-statistic is

$$U(\mathbf{X}) = \binom{n}{3}^{-1} \sum\sum\sum_{i<j<k} X_i X_j X_k = \frac{1}{n(n-1)(n-2)} \sum_{i \neq j \neq k} X_i X_j X_k.$$

For estimating $F(x)$ a symmetric kernel is $I_{[X_i \le x]}$, so the corresponding U-statistic is

$$U(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^n I_{[X_i \le x]} = F_n^*(x),$$

and for estimating $P(X > 0)$ the U-statistic is

$$U(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^n I_{[X_i > 0]} = 1 - F_n^*(0).$$

Finally, for estimating $P(X_1 + X_2 > 0)$ the U-statistic is

$$U(\mathbf{X}) = \frac{1}{\binom{n}{2}} \sum_{i<j} I_{[X_i+X_j>0]}.$$

Theorem 2. The variance of the U-statistic defined in (3) is given by

$$\text{(5)} \qquad \operatorname{var} U(\mathbf{X}) = \frac{1}{\binom{n}{m}} \sum_{c=1}^{m} \binom{m}{n} \binom{n-m}{m-c} \zeta_c,$$

where

$$\zeta_c = \operatorname{cov}_F \left\{ T_s\left(X_{i_1}, \dots, X_{i_m}\right), T_s\left(X_{j_i}, \dots, X_{j_m}\right) \right\}$$

with m, the degree of $g(F)$, and c is the common number of integers in the sets $\{i_1, \dots, i_m\}$ and $\{j_1, \dots, j_m\}$. [For $c = 0$, the two statistics $T(X_{i_1}, \dots, X_{i_m})$ and $T(X_{j_1}, \dots, X_{j_m})$ are independent and have zero covariance.]

Proof. Clearly,

$$\operatorname{var} U(\mathbf{X})$$
$$= \frac{1}{\left[\binom{n}{m}\right]^2} \sum \sum E_F \left[\left\{ T_s\left(X_{i_1}, \dots, X_{i_m}\right) - g(F) \right\} \left\{ T_s\left(X_{j_1}, \dots, X_{j_m}\right) - g(F) \right\} \right].$$

Let c be the number of common integers in $\{i_1, i_2, \dots, i_m\}$ and $\{j_2, j_2, \dots, j_m\}$. Then c takes values $0, 1, \dots, m$ and for $c = 0$, $T_s(X_{i_1}, \dots, X_{i_m})$ and $T_s(X_{j_1}, \dots, X_{j_m})$ are independent. It follows that

$$\text{(6)} \qquad \operatorname{var} U(\mathbf{X}) = \frac{1}{\left[\binom{n}{m}\right]^2} \sum_{c=1}^{m} \binom{n}{m} \binom{m}{c} \binom{n-m}{m-c} \zeta_c,$$

which is (5). The counting argument from (6) to (7) is as follows: First we select integers $\{i_1, \dots, i_m\}$ from $\{1, 2, \dots, n\}$ in $\binom{n}{m}$ ways. Next we select the integers in $\{j_1, \dots, j_m\}$. This is done by selecting first the c integers that will be in $\{i_1, \dots, i_m\}$ (hence common to both sets) and then the $m - c$ integers from $n - m$ integers which will not be $\{j_1, \dots, j_m\}$. Note that $\zeta_0 = 0$ from independence.

Example 7. Consider the U-statistic estimator $\overline{X}$ of $g(F) = \mu(F)$ in Example 6. Here $m = 1$, $T(x) = x$, and $\zeta_1 = \operatorname{var}(X_1) = \sigma^2$ so that $\operatorname{var}(\overline{X}) = \sigma^2/n$.

For the parameter $g(F) = \mu_2(F)$, $U(\mathbf{X}) = S^2$. In this case, $m = 2$, $T_s(X_{i_1}, X_{i_2}) = (X_{i_1} - X_{i_2})^2/2$, so

$$\operatorname{var} U(\mathbf{X}) = \frac{1}{\binom{n}{2}}\{2(n-2)\zeta_1 + \zeta_2\},$$

where

$$\zeta_2 = E_F\left[\frac{1}{4}\left(X_{i_1} - X_{i_2}\right)^4 - \sigma^4\right] = \frac{\mu_4 + \sigma^4}{2},$$

and

$$\zeta_1 = \operatorname{cov}\left[\tfrac{1}{2}\left(X_{i_1} - X_{i_2}\right)^2, \tfrac{1}{2}\left(X_{i_1} - X_{j_2}\right)^2\right],$$

where $i_2 \neq j_2$. Then

$$\zeta_1 = \frac{\mu_4 - \sigma^4}{4}$$

and

$$\begin{aligned}\operatorname{var} U(\mathbf{X}) = \operatorname{var}(S^2) &= \frac{2}{n(n-1)}\left[\frac{(n-2)(\mu_4 - \sigma^4)}{2} + \frac{\mu_4 + \sigma^4}{2}\right] \\ &= \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right),\end{aligned}$$

which agrees with Corollary 2 to Theorem 7.3.5.

For the parameter $g(F) = F(x)$, $\operatorname{var} U(\mathbf{X}) = F(x)(1 - F(x))/n$, and for $g(F) = P_F(X_1 + X_2 > 0)$,

$$\operatorname{var} U(\mathbf{X}) = \frac{1}{n(n-1)}[4(n-2)\zeta_1 + 2\zeta_2],$$

where

$$\zeta_1 = P_F(X_1 + X_2 > 0,\ X_1 + X_3 > 0) - P_F^2(X_1 + X_2 > 0)$$

and

$$\begin{aligned}\zeta_2 &= P_F(X_1 + X_2 > 0) - P_F^2(X_1 + X_2 > 0) \\ &= P_F(X_1 + X_2 > 0) P_F(X_1 + X_2 \leq 0).\end{aligned}$$

Corollary to Theorem 2. Let U be the U-statistic for a symmetric kernel $T_s(X_1, X_2, \ldots, X_m)$. Suppose that $E_F[T_s(X_1, \ldots, X_m)]^2 < \infty$. Then

$$\lim_{n\to\infty}\{n \operatorname{var} U(\mathbf{X})\} = m^2\zeta_1. \tag{7}$$

Proof. It is easily shown that $0 \le \zeta_c \le \zeta_m$ for $1 \le c \le m$. It follows from the hypothesis $\zeta_m = \text{var}[T_s(X_1, \dots, X_m)]^2 < \infty$ and (5) that $\text{var}\, U(\mathbf{X}) < \infty$. Now

$$n\frac{\binom{m}{c}\binom{n-m}{m-c}}{\binom{n}{m}}\zeta_c = \frac{(m!)^2 n}{c!\,[(m-c)!]^2}\frac{[(n-m)!]^2}{n!\,(n-2m+c)!}\zeta_c$$
$$= \frac{(m!)^2}{c!\,[(m-c)!]^2}n\frac{(n-m)(n-m-1)\cdots(n-2m+c+1)}{n(n-1)\cdots(n-m+1)}\zeta_c.$$

Note that the numerator has $m - c + 1$ factors involving n, while the denominator has m such factors so that for $c > 1$, the ratio involving n goes to zero as $n \to \infty$. For $c = 1$, this ratio $\to 1$ and

$$n\,\text{var}\, U(\mathbf{X}) \longrightarrow \frac{(m!)^2}{[(m-1)!]^2}\zeta_1 = m^2\zeta_1$$

as $n \to \infty$.

Example 8. In Example 7, $n\,\text{var}(\overline{X}) \equiv \sigma^2$ and

$$n\,\text{var}(S^2) \longrightarrow 2^2\zeta_1 = \mu_4 - \sigma^4$$

as $n \to \infty$.

Finally, we state, without proof, the following result due to Hoeffding [42], which establishes the asymptotic normality of a suitably centered and normed U-statistic. For proof we refer to Lehmann [59, pp. 364–365] or Randles and Wolfe [83, p. 82].

Theorem 3. Let $X_1, X_2, \dots, X_n$ be a random sample from a DF F and let $g(F)$ be an estimable parameter of degree m with symmetric kernel $T_s(X_1, X_2, \dots, X_m)$. If $E_F\,\{T_s(X_1, X_2, \dots, X_m)\}^2 < \infty$ and U is the U-statistic for g [as defined in (3)], then $\sqrt{n}(U(\mathbf{X}) - g(F)) \xrightarrow{L} \mathcal{N}(0, m^2\zeta_1)$, provided that

$$\zeta_1 = \text{cov}_F\left\{T_s\left(X_{i_1}, \dots, X_{i_m}\right), T_s\left(X_{j_1}, \dots, X_{j_m}\right)\right\} > 0.$$

In view of the corollary to Theorem 2, it follows that $[U - g(F)]/\sqrt{\text{var}(U)} \xrightarrow{L} \mathcal{N}(0, 1$ provided that $\zeta_1 > 0$.

Example 9. (*Example 7 continued*). Clearly, $\sqrt{n}(\overline{X} - \mu)/\sigma \xrightarrow{L} \mathcal{N}(0, 1)$ as $n \to \infty$ since $\zeta_1 = \sigma^2 > 0$.

For the parameter $g(F) = \mu_2(F)$,

$$\text{var}\, U(\mathbf{X}) = \text{var}(S^2) = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right), \qquad \zeta_1 = \frac{\mu_4 - \sigma^4}{4} > 0,$$

so it follows from Theorem 3 that

$$\sqrt{n}(S^2 - \sigma^2) \xrightarrow{L} \mathcal{N}(0, \mu^4 - \sigma^4).$$

The concept of U-statistic can be extended to multiple random samples. We will restrict ourselves to the case of two samples. Let $X_1, X_2, \ldots, X_{n_1}$ and $Y_1, Y_2, \ldots, Y_{n_2}$ be two independent random samples from DFs F and G, respectively.

Definition 8. A parameter $g(F, G)$ is estimable of degrees (m_1, m_2) if m_1 and m_2 are the smallest sample sizes for which there exists a statistic $T(X_1, \ldots, X_{m_1}; Y_1, \ldots, Y_{m_2})$ such that

$$E_{F,G}T\left(X_1, \ldots, X_{m_1}; Y_1, \ldots, Y_{m_2}\right) = g(F, G) \tag{8}$$

for all $F, G \in \mathcal{P}$.

The statistic T in Definition 8 is called a *kernel* of g and a symmetrized version of T, T_s, is called a *symmetric kernel* of g. Without loss of generality, therefore, we assume that the two-sample kernel T in (9) is a symmetric kernel.

Definition 9. Let $g(F, G)$, $F, G \in \mathcal{P}$ be an estimable parameter of degree (m_1, m_2). Then a (two-sample) U-statistic estimate of g is defined by

$$U(\mathbf{X}; \mathbf{Y}) = \binom{n_1}{m_1}^{-1} \binom{n_2}{m_2}^{-1} \sum_{\mathbf{i} \in A} \sum_{\mathbf{j} \in B} T\left(X_{i_1}, \ldots, X_{i_{m_1}}; Y_{j_1}, \ldots, Y_{j_{m_2}}\right), \tag{9}$$

where A and B are collections of all subsets of m_1 and m_2 integers chosen without replacement from the sets $\{1, 2, \ldots, n_1\}$ and $\{1, 2, \ldots, n_2\}$, respectively.

Example 10. Let $X_1, X_2, \ldots, X_{n_1}$ and $Y_1, Y_2, \ldots, Y_{n_2}$ be two independent samples from DFs F and G respectively. Let

$$g(F, G) = P(X < Y) = \int_{-\infty}^{\infty} F(x) g(x)\, dx = \int_{-\infty}^{\infty} P(Y > y) f(y)\, dy,$$

where f and g are the respective PDFs of F and G. Then

$$T(X_i; Y_j) = \begin{cases} 1 & \text{if } X_i < Y_j \\ 0 & \text{if } X_i \geq Y_j \end{cases}$$

is an unbiased estimator of g. Clearly, g has degree $(1, 1)$ and the two-sample U-statistic is given by

$$U(\mathbf{X}; \mathbf{Y}) = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} T(X_i; Y_j).$$

Theorem 4. The variance of the two-sample U-statistic defined in (9) is given by

$$\operatorname{var} U(\mathbf{X};\mathbf{Y}) = \frac{1}{\binom{n_1}{m_1}\binom{n_2}{m_2}} \sum_{c=0}^{m_1}\sum_{d=0}^{m_2} \binom{m_1}{c}\binom{n_1-m_1}{m_1-c}\binom{m_2}{d}\binom{n_2-m_2}{m_2-d}\zeta_{c,d},$$

(10)

where $\zeta_{c,d}$ is the covariance between $T(X_{i_1},\dots,X_{i_{m_1}};Y_{j_1},\dots,Y_{j_{m_1}})$ and $T\big(X_{k_1},\dots,X_{k_{m_1}};Y_{\ell_1},\dots,Y_{\ell_{m_2}}\big)$ with exactly c X's and d Y's in common.

Corollary. Suppose that $E_{F,G}T^2(X_1,\dots,X_{m_1};Y_1,\dots,Y_{m_2}) < \infty$ for all $F, G \in \mathcal{P}$. Let $N = n_1 + n_2$ and suppose that $n_1, n_2, N \to \infty$ such that $n_1/N \to \lambda$, $n_2/N \to 1-\lambda$. Then

$$\lim_{N\to\infty} N \operatorname{var} U(\mathbf{X};\mathbf{Y}) = \frac{m_1^2}{\lambda}\zeta_{1,0} + \frac{m_2^2}{1-\lambda}\zeta_{0,1}. \tag{11}$$

The proofs of Theorem 4 and its corollary parallel those of Theorem 2 and its corollary and are left to the reader.

Example 11. For the U-statistic in Example 10,

$$E_{F,G}U^2(\mathbf{X};\mathbf{Y}) = \frac{1}{n_1^2n_2^2}\sum\sum\sum\sum E_{F,G}\left\{T(X_i;Y_j)T(X_k;Y_\ell)\right\}.$$

Now

$$E_{F,G}\left\{T(X_i;Y_j)T(X_k;Y_\ell)\right\} = P(X_i < Y_j, X_k < Y_l)$$
$$= \begin{cases} \int_{-\infty}^{\infty} F(x)g(x)\,dx & \text{for } i=k,\ j=l,\\ \int_{-\infty}^{\infty}[1-G(x)]^2 f(x)\,dx & \text{for } i=k,\ j\neq l,\\ \int_{-\infty}^{\infty} F^2(x)g(x)\,dx & \text{for } i\neq k,\ j=l,\\ \left[\int_{-\infty}^{\infty} F(x)g(x)\,dx\right]^2 & \text{for } i\neq k,\ j\neq l, \end{cases}$$

where f and g are PDFs of F and G, respectively. Moreover,

$$\zeta_{1,0} = \int_{-\infty}^{\infty}[1-G(x)]^2 f(x)\,dx - [g(F,G)]^2$$

and

$$\zeta_{0,1} = \int_{-\infty}^{\infty} F^2(X)g(x)]\,dx - [g(F,G)]^2.$$

It follows that

$$\operatorname{var} U(\mathbf{X};\mathbf{Y}) = \frac{1}{n_1 n_2}\left\{g(F,G)[1-g(F,G)] + (n_1-1)\zeta_{1,0} + (n_2-1)\zeta_{0,1}\right\}.$$

In the special case when $F = G$, $g(F,G) = \frac{1}{2}$, $\zeta_{1,0} = \zeta_{0,1} = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$ and $\operatorname{var}(U) = (n_1 + n_2 + 1)/(12 n_1 n_2)$.

Finally, we state, without proof, the two-sample analog of Theorem 3, which establishes the asymptotic normality of the two-sample U-statistic defined in (9).

Theorem 5. Let $X_1, X_2, \ldots, X_{n_1}$ and $Y_1, Y_2, \ldots, Y_{n_2}$ be independent random samples from DFs F and G, respectively, and let $g(F, G)$ be an estimable parameter of degree (m_1, m_2). Let $T(X_1, \ldots, X_{m_1}; Y_1, \ldots, Y_{m_2})$ be a symmetric kernel for g such that $ET^2 < \infty$. Then

$$\sqrt{n_1 + n_2}\,\{U(\mathbf{X};\mathbf{Y}) - g(F,G)\} \xrightarrow{L} \mathcal{N}(0,\sigma^2),$$

where $\sigma^2 = m_1^2\zeta_{1,0}/\lambda + m_2^2\zeta_{0,1}/(1-\lambda)$, provided that $\sigma^2 > 0$, and $0 < \lambda = \lim_{N\to\infty}(m_1/N) = \lambda < 1$, $N = n_1 + n_2$.

In view of (12), we see that $(U - g)/\sqrt{\operatorname{var} U} \xrightarrow{L} \mathcal{N}(0,1)$, provided that $\sigma^2 > 0$.

For a proof of Theorem 5 we refer to Lehmann [59, p. 364], or Randles and Wolfe [83, p. 92].

Example 11. (*Continued*). In Example 11 we saw that in the special case when $F = G$, $\zeta_{1,0} = \zeta_{0,1} = \frac{1}{12}$, and $\operatorname{var} U = (n_1 + n_2 + 1)/(12 n_1 n_2)$. It follows from the remark following Theorem 5 that

$$\frac{U(\mathbf{X};\mathbf{Y}) - \frac{1}{2}}{\sqrt{(n_1+n_2+1)/(12n_1n_2)}} \xrightarrow{L} \mathcal{N}(0,1).$$

PROBLEMS 13.2

1. Let $(\mathcal{R}, \mathfrak{B}, P_\theta)$ be a probability space, and let $\mathcal{P} = \{P_\theta\colon \theta \in \Theta\}$. Let A be a Borel subset of $\mathcal{R}$, and consider the parameter $d(\theta) = P_\theta(A)$. Is d estimable? If so, what is the degree? Find the UMVUE for d, based on a sample of size n, assuming that $\mathcal{P}$ is the class of all continuous distributions.

2. Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be independent random samples from two absolutely continuous DFs. Find the UMVUEs of (a) $E(XY)$, and (b) $\operatorname{var}(X+Y)$.

3. Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be a random sample from an absolutely continuous distribution. Find the UMVUEs of (a) $E(XY)$ and (b) $\operatorname{var}(X+Y)$.

4. Let $T(X_1, X_2, \dots, X_n)$ be a statistic that is symmetric in the observations. Show that T can be written as a function of the order statistic. Conversely, if $T(X_1, X_2, \dots, X_n)$ can be written as a function of the order statistic, T is symmetric in the observations.

5. Let $X_1, X_2, \dots, X_n$ be a random sample from an absolutely continuous DF F, $F \in \mathcal{P}$. Find U-statistics for $g_1(F) = \mu^3(F)$ and $g_2(F) = \mu_3(F)$. Find the corresponding expressions for the variance of the U-statistic in each case.

6. In Example 3, show that $\mu_2(F)$ is not estimable with one observation. That is, show that the degree of $\mu_2(F)$ where $F \in \mathcal{P}$, the class of all distributions with finite second moment, is two.

7. Show that for $c = 1, 2, \dots, m$, $0 \le \zeta_c \le \zeta_m$.

8. Let $X_1, X_2, \dots, X_n$ be a random sample from an absolutely continuous DF F, $F \in \mathcal{P}$. Let

$$g(F) = E_F|X_1 - X_2|.$$

Find the U-statistic estimator of $g(F)$ and its variance.

13.3 SOME SINGLE-SAMPLE PROBLEMS

Let $X_1, X_2, \dots, X_n$ be a random sample from a DF F. In Section 13.2 we studied properties of U-statistics as nonparametric estimators of parameters $g(F)$. In this section we consider some nonparametric tests of hypotheses. Often, the test statistic may be viewed as a function of a U-statistic.

13.3.1 Goodness-of-Fit Problem

The problem of fit is to test the hypothesis that the sample comes from a specified DF F_0 against the alternative that it is from some other DF F, where $F(x) \ne F_0(x)$ for some $x \in \mathcal{R}$. In Section 10.3 we studied the chi-square test of goodness of fit for testing H_0: $X_i \sim F_0$. Here we consider the Kolmogorov–Smirnov test of H_0. Since H_0 concerns the underlying DF of the X's, it is natural to compare the U-statistic estimator of $g(F) = F(x)$ with the specified DF F_0 under H_0. The U-statistic for $g(F) = F(x)$ is the empirical DF $F_n^*(x)$.

Definition 1. Let $X_1, X_2, \dots, X_n$ be a sample from a DF F, and let F_n^* be a corresponding empirical DF. The statistic

$$D_n = \sup_x |F_n^*(x) - F(x)| \tag{1}$$

is called the (two-sided) *Kolmogorov–Smirnov statistic*. We write

$$D_n^+ = \sup_n [F_n^*(x) - F(x)] \tag{2}$$

and

$$D_n^- = \sup_x [F(x) - F_n^*(x)], \tag{3}$$

and call D_n^+, D_n^- the *one-sided Kolmogorov–Smirnov statistics.*

Theorem 1. The statistics D_n, D_n^-, D_n^+ are distribution-free for any continuous DF F.

Proof. Clearly, $D_n = \max(D_n^+, D_n^-)$. Let $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ be the order statistics of $X_1, X_2, \dots, X_n$, and define $X_{(0)} = -\infty$, $X_{(n+1)} = +\infty$. Then

$$F_n^*(x) = \frac{i}{n} \qquad \text{for } X_{(i)} \le x < X_{(i+1)}, \quad i = 0, 1, 2, \dots, n,$$

and we have

$$\begin{aligned} D_n^+ &= \max_{0 \le i \le n} \sup_{X_{(i)} \le x < X_{(i+1)}} \left\{ \frac{i}{n} - F(x) \right\} \\ &= \max_{0 \le i \le n} \left\{ \frac{i}{n} - \inf_{X_{(i)} \le x < X_{(i+1)}} F(x) \right\} \\ &= \max_{0 \le i \le n} \left\{ \frac{i}{n} - F(X_{(i)}) \right\} \\ &= \max \left\{ \max_{1 \le i \le n} \left[\frac{i}{n} - F(X_{(i)}) \right], 0 \right\}. \end{aligned}$$

Since $F(X_{(i)})$ is the ith-order statistic of a sample from $U(0, 1)$ irrespective of what F is, as long as it is continuous, we see that the distribution of D_n^+ is independent of F. Similarly,

$$D_n^- = \max \left\{ \max_{1 \le i \le n} \left[F(X_{(i)}) - \frac{i-1}{n} \right], 0 \right\},$$

and the result follows.

Without loss of generality, therefore, we assume that F is the DF of a $U(0, 1)$ RV.

Theorem 2. If F is continuous, then

$$(4)\quad P\left\{D_n \le v + \frac{1}{2n}\right\}$$

$$= \begin{cases} 0 & \text{if } v \le 0, \\ \displaystyle\int_{(1/2n)-v}^{v+(1/2n)} \int_{(3/2n)-v}^{v+(3/2n)} \cdots \\ \displaystyle\int_{[(2n-1)/2n]-v}^{v+[(2n-1)/2n]} f(u_1, u_2, \dots, u_n) \cdot \prod_{n}^{1} du_i & \text{if } 0 < v < \dfrac{2n-1}{2n}, \\ 1 & \text{if } v \ge \dfrac{2n-1}{2n}, \end{cases}$$

where

$$(5)\qquad f(u_1, u_2, \dots, u_n) = \begin{cases} n!, & 0 < u_1 < \cdots < u_n < 1, \\ 0, & \text{otherwise,} \end{cases}$$

is the joint PDF of the set of order statistics for a sample of size n from $U(0, 1)$.

We will not prove this result here. Let $D_{n,\alpha}$ be the upper α-percent point of the distribution of D_n, that is, $P\{D_n > D_{n,\alpha}\} \le \alpha$. The exact distribution of D_n for selected values of n and α has been tabulated by Miller [72], Owen [77], and Birnbaum [8]. The large-sample distribution of D_n was derived by Kolmogorov [51], and we state it without proof.

Theorem 3. Let F be any continuous DF. Then for every $z \ge 0$,

$$(6)\qquad \lim_{n\to\infty} P\{D_n \le zn^{-1/2}\} = L(z),$$

where

$$(7)\qquad L(z) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2z^2}.$$

Theorem 3 can be used to find d_α such that $\lim_{n\to\infty} P\{\sqrt{n}\, D_n \le d_\alpha\} = 1 - \alpha$. Tables of d_α for various values of α are also available in Owen [77].

The statistics D_n^+ and D_n^- have the same distribution because of symmetry, and their common distribution is given by the following theorem.

Theorem 4. Let F be a continuous DF. Then

$$(8)\quad P\{D_n^+ \le z\} = \begin{cases} 0 & \text{if } z \le 0, \\ \displaystyle\int_{1-z}^{1} \int_{[(n-1)/n]-z}^{u_n} \cdots \int_{(2/n)-z}^{u_3} \\ \displaystyle\times \int_{(1/n)-z}^{u_2} f(u_1, u_2, \ldots, u_n) \prod_{i=1}^{n} du_i & \text{if } 0 < z < 1, \\ 1 & \text{if } z \ge 1, \end{cases}$$

where f is given by (5).

We leave the reader to prove Theorem 4.

Tables for the critical values $D_{n,\alpha}^+$, where $P\{D_n^+ > D_{n,\alpha}^+\} \le \alpha$, are also available for selected values of n and α (see Birnbaum and Tingey [7]). Table ST7 gives $D_{n,\alpha}^+$ and $D_{n,\alpha}$ for some selected values of n and α. For large samples, Smirnov [106] showed that

$$(9)\qquad \lim_{n\to\infty} P\{\sqrt{n}\, D_n^+ \le z\} = 1 - e^{-2z^2}, \qquad z \ge 0.$$

In fact, in view of (9), the statistic $V_n = 4nD_n^{+2}$ has a limiting $\chi^2(2)$ distribution, for $4nD_n^{+2} \le 4z^2$ if and only if $\sqrt{n}\, D_n^+ \le z$, $z \ge 0$, and the result follows since

$$\lim_{n\to\infty} P\{V_n \le 4z^2\} = 1 - e^{-2z^2}, \qquad z \ge 0,$$

so that

$$\lim_{n\to\infty} P\{V_n \le x\} = 1 - e^{-x/2}, \qquad x \ge 0,$$

which is the DF of a $\chi^2(2)$ RV.

Example 1. Let $\alpha = 0.01$, and let us approximate $D_{n,\alpha}^+$. We have $\chi^2_{2,0.01} = 9.21$. Thus $V_n = 9.21$, yielding

$$D_{n,0.01}^+ = \sqrt{\frac{9.21}{4n}} = \frac{3.03}{2\sqrt{n}}.$$

If, for example, $n = 9$, then $D_{n,0.01}^+ = 3.03/6 = 0.50$. Of course, the approximation is better for large n.

The statistic D_n and its one-sided analogs can be used in testing $H_0\colon X \sim F_0$ against $H_1 : X \sim F$, where $F_0(x) \ne F(x)$ for some x.

Definition 2. To test $H_0\colon F(x) = F_0(x)$ for all x at level α, the *Kolmogorov–Smirnov test* rejects H_0 if $D_n > D_{n,\alpha}$. Similarly, it rejects $F(x) \ge F_0(x)$ for all x if $D_n^- > D_{n,\alpha}^+$ and rejects $F(x) \le F_0(x)$ for all x at level α if $D_n^+ > D_{n,\alpha}^+$.

For large samples we can approximate by using Theorem 3 or (9) to obtain an approximate α-level test.

Example 2. Let us consider the data in Example 10.3.3, and apply the Kolmogorov–Smirnov test to determine the goodness of the fit. Rearranging the data in increasing order of magnitude, we have the following result:

x	$F_0(x)$	$F^*_{20}(x)$	$i/20 - F_0(x_{(i)})$	$F_0(x_{(i)}) - (i-1)/20$
−1.787	0.0367	$\frac{1}{20}$	0.0133	0.0367
−1.229	0.1093	$\frac{2}{20}$	−0.0093	0.0593
−0.525	0.2998	$\frac{3}{20}$	−0.1498	0.1998
−0.513	0.3050	$\frac{4}{20}$	−0.1050	0.1550
−0.508	0.3050	$\frac{5}{20}$	−0.0550	0.1050
−0.486	0.3121	$\frac{6}{20}$	−0.0121	0.0621
−0.482	0.3156	$\frac{7}{20}$	0.0344	0.0156
−0.323	0.3745	$\frac{8}{20}$	0.0255	0.0245
−0.261	0.3974	$\frac{9}{20}$	0.0526	−0.0026
−0.068	0.4721	$\frac{10}{20}$	0.0279	0.0221
−0.057	0.4761	$\frac{11}{20}$	0.0739	−0.0239
0.137	0.5557	$\frac{12}{20}$	0.0443	0.0057
0.464	0.6772	$\frac{13}{20}$	−0.0272	0.0772
0.595	0.7257	$\frac{14}{20}$	−0.0257	0.0757
0.881	0.8106	$\frac{15}{20}$	−0.0606	0.1106
0.906	0.8186	$\frac{16}{20}$	−0.0186	0.0686
1.046	0.8531	$\frac{17}{20}$	−0.0031	0.0531
1.237	0.8925	$\frac{18}{20}$	0.0075	0.0425
1.678	0.9535	$\frac{19}{20}$	−0.0035	0.0535
2.455	0.9931	1	0.0069	0.0431

From Theorem 1,

$$D_{20}^- = 0.1998, \quad D_{20}^+ = 0.0739, \quad \text{and} \quad D_{20} = \max(D_{20}^+, D_{20}^-) = 0.1998.$$

Let us take $\alpha = 0.05$. Then $D_{20,0.05} = 0.294$. Since $0.1998 < 0.294$, we accept H_0 at the 0.05 level of significance.

It is worthwhile to compare the chi-square test of goodness of fit and the Kolmogorov–Smirnov test. The latter treats individual observations directly, whereas the former discretizes the data and sometimes loses information through grouping. Moreover, the Kolmogorov–Smirnov test is applicable even in the case of very small samples, but the chi-square test is essentially for large samples.

The chi-square test can be applied when the data are discrete or continuous, but the Kolmogorov–Smirnov test assumes continuity of the DF. This means that the latter test provides a more refined analysis of the data. If the distribution is actually discontinuous, the Kolmogorov–Smirnov test is conservative in that it favors H_0.

We next turn our attention to some other uses of the Kolmogorov–Smirnov statistic. Let $X_1, X_2, \ldots, X_n$ be a sample from a DF F, and let F_n^* be the sample DF. The estimate F_n^* of F for large n should be close to F. Indeed,

$$P\left\{|F_n^*(x) - F(x)| \le \frac{\lambda\sqrt{F(x)[1-F(x)]}}{\sqrt{n}}\right\} \ge 1 - \frac{1}{\lambda^2}, \tag{10}$$

and since $F(x)[1-F(x)] \le \frac{1}{4}$, we have

$$P\left\{|F_n^*(x) - F(x)| \le \frac{\lambda}{2\sqrt{n}}\right\} \ge 1 - \frac{1}{\lambda^2}. \tag{11}$$

Thus F_n^* can be made close to F with high probability by choosing λ and large enough n. The Kolmogorov–Smirnov statistic enables us to determine the smallest n such that the error in estimation never exceeds a fixed value ε with a large probability $1-\alpha$. Since

$$P\{D_n \le \varepsilon\} \ge 1 - \alpha, \tag{12}$$

$\varepsilon = D_{n,\alpha}$; and given ε and α, we can read n from the tables. For large n we can use the asymptotic distribution of D_n and solve $d_\alpha = \varepsilon\sqrt{n}$ for n.

We can also form confidence bounds for F. Given α and n, we first find $D_{n,\alpha}$ such that

$$P\{D_n > D_{n,\alpha}\} \le \alpha, \tag{13}$$

which is the same as

$$P\left\{\sup_x |F_n^*(x) - F(x)| \le D_{n,\alpha}\right\} \ge 1 - \alpha.$$

Thus

$$P\{|F_n^*(x) - F(x)| \le D_{n,\alpha} \quad \text{for all } x\} \ge 1 - \alpha. \tag{14}$$

Define

$$L_n(x) = \max\{F_n^*(x) - D_{n,\alpha}, 0\} \tag{15}$$

and

$$U_n(x) = \min\{F_n^*(x) + D_{n,\alpha}, 1\}. \tag{16}$$

Then the region between $L_n(x)$ and $U_n(x)$ can be used as a confidence band for $F(x)$ with associated confidence coefficient $1-\alpha$.

Example 3. For the data on the standard normal distribution of Example 2, let us form a 0.90 confidence band for the DF. We have $D_{20,0.10} = 0.265$. The confidence band is, therefore, $F_{20}^*(x) \pm 0.265$ as long as the band is between 0 and 1.

13.3.2 Problem of Location

Let $X_1, X_2, \ldots, X_n$ be a sample of size n from some unknown DF F. Let p be a positive real number, $0 < p < 1$, and let $\mathfrak{z}_p(F)$ denote the quantile of order p for the DF F. In the following analysis we assume that F is absolutely continuous. The problem of location is to test $H_0\colon \mathfrak{z}_p(F) = \mathfrak{z}_0$, $\mathfrak{z}_0$ a given number, against one of the alternatives $\mathfrak{z}_p(F) > \mathfrak{z}_0$, $\mathfrak{z}_p < \mathfrak{z}_0$, and $\mathfrak{z}_p \neq \mathfrak{z}_0$. The problem of location and symmetry is to test $H_0'\colon \mathfrak{z}_{0.5}(F) = \mathfrak{z}_0$, and F is symmetric against $H_1'\colon \mathfrak{z}_{0.5}(F) \neq \mathfrak{z}_0$ or F is not symmetric.

We consider two tests of location. First, we describe the sign test.

Sign Test

Let $X_1, X_2, \ldots, X_n$ be iid RVs with common PDF f. Consider the hypothesis-testing problem

$$(17) \qquad H_0\colon \mathfrak{z}_p(f) = \mathfrak{z}_0 \qquad \text{against } H_1\colon \mathfrak{z}_p(f) > \mathfrak{z}_0,$$

where $\mathfrak{z}_p(f)$ is the quantile of order p of PDF f, $0 < p < 1$. Let $g(F) = P(X_i > \mathfrak{z}_0) = P(X_i - \mathfrak{z}_0 > 0)$. Then the corresponding U-statistic is given by

$$nU(\mathbf{X}) = R^+(\mathbf{X}),$$

the number of positive elements in $X_1 - \mathfrak{z}_0, X_2 - \mathfrak{z}_0, \ldots, X_n - \mathfrak{z}_0$. Clearly, $P(X_i = \mathfrak{z}_0) = 0$. Fraser [29, pp. 167–170] has shown that a UMP test of H_0 against H_1 is given by

$$(18) \qquad \varphi(\mathbf{x}) = \begin{cases} 1, & R^+(\mathbf{x}) > c; \\ \gamma, & R^+(\mathbf{x}) = c, \\ 0, & R^+(\mathbf{x}) < c, \end{cases}$$

where c and γ are chosen from the size restriction

$$(19) \qquad \alpha = \sum_{i=c+1}^{n} \binom{n}{R^+(\mathbf{x})} (1-p)^{R^+(\mathbf{x})} p^{n-R^+(\mathbf{x})} + \gamma \binom{n}{c} (1-p)^c p^{n-c}.$$

Note that under H_0, $\mathfrak{z}_p(f) = \mathfrak{z}_0$, so that $P_{H_0}(X \le \mathfrak{z}_0) = p$, and $R^+(\mathbf{X}) \sim b(n, 1-p)$. The same test is UMP for $H_0\colon \mathfrak{z}_p(f) \le \mathfrak{z}_0$ against $H_1\colon \mathfrak{z}_p(f) > \mathfrak{z}_0$. For the two-sided case, Fraser [29, p. 171] shows that the two-sided sign test is UMP unbiased.

If, in particular, $\mathfrak{z}_0$ is the median of f, then $p = \frac{1}{2}$ under H_0. In this case one can also use the sign test to test $H_0\colon \text{med}(X) = \mathfrak{z}_0$, F is symmetric.

For large n one can use the normal approximation to binomial to find c and γ in (19).

Example 4. Entering college freshmen have taken a particular high school achievement test for many years, and the upper quartile ($p = 0.75$) is well established at a score of 195. A particular high school sent 12 of its graduates to college, where they took the examination and obtained scores of 203, 168, 187, 235, 197, 163, 214, 233, 179, 185, 197, 216. Let us test the null hypothesis H_0 that $\mathfrak{z}_{0.75} \le 195$ against $H_1 : \mathfrak{z}_{0.75} > 195$ at the $\alpha = 0.05$ level.

We have to find c and γ such that

$$\sum_{i=c+1}^{12} \binom{12}{i} \left(\frac{1}{4}\right)^i \left(\frac{3}{4}\right)^{12-i} + \gamma \binom{12}{c} \left(\frac{1}{4}\right)^c \left(\frac{3}{4}\right)^{12-c} = 0.05.$$

From the table of cumulative binomial distribution (Table ST1) for $n = 12$, $p = \frac{1}{4}$, we see that $c = 6$. Then γ is given by

$$0.0142 + \gamma \binom{12}{6} \left(\frac{1}{4}\right)^6 \left(\frac{3}{4}\right)^6 = 0.05.$$

Thus

$$\gamma = \frac{0.0358}{0.0402} = 0.89.$$

In our case the number of positive signs, $x_i - 195$, $i = 1, 2, \ldots, 12$, is 7, so we reject H_0 that the upper quartile is ≤ 195.

Example 5. A random sample of size 8 is taken from a normal population with mean 0 and variance 1. The sample values are -0.465, 0.120, -0.238, -0.869, -1.016, 0.417, 0.056, 0.561. Let us test hypothesis $H_0 : \mu = -1.0$ against $H_1 : \mu > -1.0$. We should expect to reject H_0 since we know that it is false. The number of observations, $x_i - \mu_0 = x_i + 1.0$, that are ≥ 0 is 7. We have to find c and γ such that

$$\sum_{i=c+1}^{8} \binom{8}{i} \left(\frac{1}{2}\right)^8 + \gamma \binom{8}{c} \left(\frac{1}{2}\right)^8 = 0.05, \text{ say},$$

that is,

$$\sum_{i=c+1}^{8} \binom{8}{i} + \gamma \binom{8}{c} = 12.8.$$

We see that $c = 6$ and $\gamma = 0.13$. Since the number of positive $x_i - \mu_0$ is > 6, we reject H_0.

Let us now apply the parametric test here. We have

$$\bar{x} = -\frac{1.434}{8} = -0.179.$$

Since $\sigma = 1$, we reject H_0 if

$$\bar{x} > \mu_0 + \frac{1}{\sqrt{n}} z_\alpha = -1.0 + \frac{1}{\sqrt{8}} 1.64$$
$$= -0.42.$$

Since $-0.179 > -0.42$, we reject H_0.

The single-sample sign test described above can easily be modified to apply to sampling from a bivariate population. Let $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ be a random sample from a bivariate population. Let $Z_i = X_i - Y_i$, $i = 1, 2, \dots, n$, and assume that Z_i has an absolutely continuous DF. Then one can test hypotheses concerning the order parameters of Z by using the sign test. A hypothesis of interest here is that Z has a given median $\mathfrak{z}_0$. Without loss of generality, let $\mathfrak{z}_0 = 0$. Then H_0: $\text{med}(Z) = 0$; that is, $P\{Z > 0\} = P\{Z < 0\} = \frac{1}{2}$. Note that med($Z$) is not necessarily equal to med(X) − med(Y), so that H_0 is not that med(X) = med(Y) but that med(Z) = 0. The sign test is UMP against one-sided alternatives and UMP unbiased against two-sided alternatives.

Example 6. We consider an example due to Hahn and Nelson [37], in which two measuring devices take readings on each of 10 test units. Let X and Y, respectively, be the readings on a test unit by the first and second measuring devices. Let $X = A + \varepsilon_1$, $Y = A + \varepsilon_2$, where A, ε_1, ε_2, respectively, are the contributions to the readings due to the test unit and to the first and second measuring devices. Let A, ε_1, ε_2 be independent with $EA = \mu$, $\text{var}(A) = \sigma_a^2$, $E\varepsilon_1 = E\varepsilon_2 = 0$, $\text{var}(\varepsilon_1) = \sigma_1^2$, $\text{var}(\varepsilon_2) = \sigma_2^2$, so that X and Y have common mean μ and variances $\sigma_1^2 + \sigma_a^2$ and $\sigma_2^2 + \sigma_a^2$, respectively. Also, the covariance between X and Y is σ_a^2. The data are as follows:

	Test Unit									
	1	2	3	4	5	6	7	8	9	10
First device, X	71	108	72	140	61	97	90	127	101	114
Second device, Y	77	105	71	152	88	117	93	130	112	105
$Z = X - Y$	−6	3	1	−8	−17	−20	−3	−3	−11	9

Let us test the hypothesis H_0: med(Z) = 0. The number of Z_i's > 0 is 3. We have

$$P\{\text{number of } Z_i\text{'s} > 0 \text{ is } \le 3 \mid H_0\} = \sum_{k=0}^{3} \binom{10}{k} \left(\frac{1}{2}\right)^{10}$$
$$= 0.172.$$

Using the two-sided sign test, we cannot reject H_0 at level $\alpha = 0.05$, since $0.172 > 0.025$. The RVs Z_i can be considered to be distributed normally, so that under H_0

the common mean of Z_i's is 0. Using a paired comparison t-test on the data, we can show that $t = -0.88$ for 9 d.f., so we cannot reject the hypothesis of equality of means of X and Y at level $\alpha = 0.05$.

Finally, we consider the Wilcoxon signed-ranks test.

Wilcoxon Signed-Ranks Test

The sign test for median and symmetry loses information since it ignores the magnitude of the difference between the observations and the hypothesized median. The Wilcoxon signed-ranks test provides an alternative test of location (and symmetry) that also takes into account the magnitudes of these differences.

Let $X_1, X_2, \ldots, X_n$ be iid RVs with common absolutely continuous DF F, which is symmetric about the median $\mathfrak{z}_{1/2}$. The problem is to test H_0: $\mathfrak{z}_{1/2} = \mathfrak{z}_0$ against the usual one- or two-sided alternatives. Without loss of generality, we assume that $\mathfrak{z}_0 = 0$. Then $F(-x) = 1 - F(x)$ for all $x \in \mathcal{R}$. To test H_0: $F(0) = \frac{1}{2}$ or $\mathfrak{z}_{1/2} = 0$, we first arrange $|X_1|, |X_2|, \ldots, |X_n|$ in increasing order of magnitude and assign ranks $1, 2, \ldots, n$, keeping track of the original signs of X_i. For example, if $n = 4$ and $|X_2| < |X_4| < |X_1| < |X_3|$, the rank of $|X_1|$ is 3, of $|X_2|$ is 1, of $|X_3|$ is 4, and of $|X_4|$ is 2.

Let

$$\begin{cases} T^+ = \text{sum of the ranks of positive } X_i\text{'s}, \\ T^- = \text{sum of the ranks of negative } X_i\text{'s}. \end{cases} \tag{20}$$

Then, under H_0, we expect T^+ and T^- to be the same. Note that

$$T^+ + T^- = \sum_1^n i = \frac{n(n+1)}{2}, \tag{21}$$

so that T^+ and T^- are linearly related and offer equivalent criteria. Let us define

$$Z_i = \begin{cases} 1 & \text{if } X_i > 0 \\ 0 & \text{if } X_i < 0 \end{cases}, \quad i = 1, 2, \ldots, n, \tag{22}$$

and write $R(|X_i|) = R_i^+$ for the rank of $|X_i|$. Then $T^+ = \sum_{i=1}^n R_i^+ Z_i$ and $T^- = \sum_{i=1}^n (1 - Z_i) R_i^+$. Also,

$$\begin{aligned} T^+ - T^- &= -\sum_{i=1}^n R_i^+ + 2\sum_{i=1}^n Z_i R_i^+ \\ &= 2\sum_{i=1}^n R_i^+ Z_i - \frac{n(n+1)}{2}. \end{aligned} \tag{23}$$

The statistic T^+ (or T^-) is known as the *Wilcoxon statistic*. A large value of T^+ (or, equivalently, a small value of T^-) means that most of the large deviations from 0 are positive, and therefore we reject H_0 in favor of the alternative, H_1 : $\mathfrak{z}_{1/2} > 0$.

A similar analysis applies to the other two alternatives. We record the results as follows:

H_0	H_1	Reject H_0 if:
$\mathfrak{z}_{1/2} = 0$	$\mathfrak{z}_{1/2} > 0$	$T^+ > c_1$
$\mathfrak{z}_{1/2} = 0$	$\mathfrak{z}_{1/2} < 0$	$T^+ < c_2$
$\mathfrak{z}_{1/2} = 0$	$\mathfrak{z}_{1/2} \neq 0$	$T^+ < c_3$ or $T^+ > c_4$

We now show how the Wilcoxon signed-ranks test statistic is related to the U-statistic estimate of $g_2(F) = P_F(X_1 + X_2 > 0)$. Recall from Example 13.2.6 that the corresponding U-statistic is

$$U_2(\mathbf{X}) = \binom{n}{2}^{-1} \sum_{1\le i<j\le n} I_{[X_i+X_j>0]}. \tag{24}$$

First note that

$$\sum_{1\le i\le j\le n} I_{[X_i+X_j>0]} = \sum_{j=1}^{n} I_{[X_j>0]} + \sum_{1\le i<j\le n} I_{[X_i+X_j>0]}. \tag{25}$$

Next note that for $i < j$, $X_{(i)} + X_{(j)} > 0$ if and only if $X_{(j)} > 0$ and $|X_{(i)}| < |X_{(j)}|$. It follows that $\sum_{i=1}^{j} I_{[X_{(i)}+X_{(j)}>0]}$ is the signed rank of $X_{(j)}$. Consequently,

$$\begin{aligned} T^+ &= \sum_{j=1}^{n}\sum_{i=1}^{j} I_{[X_{(i)}+X_{(j)}>0]} = \sum_{1\le i\le j\le n} I_{[X_i+X_j>0]} \\ &= \sum_{j=1}^{n} I_{[X_j>0]} + \sum_{1\le i<j\le n} I_{[X_i+X_j>0]} \\ &= nU_1(\mathbf{X}) + \binom{n}{2} U_2(\mathbf{X}), \end{aligned} \tag{26}$$

where U_1 is the U-statistic for $g_1(F) = P_F(X_1 > 0)$.

We next compute the distribution of T^+ for small samples. The distribution of T^+ is tabulated by Kraft and Van Eeden [53, pp. 221–223].

Let

$$Z_{(i)} = \begin{cases} 1 & \text{if the } |X_j| \text{ that has rank } i \text{ is } > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Note that $T^+ = 0$ if all differences have negative signs, and $T^+ = n(n+1)/2$ if all differences have positive signs. Here a difference means a difference between the observations and the postulated value of the median. T^+ is completely determined by the indicators $Z_{(i)}$, so that the sample space can be considered as a set of 2^n n-tuples

$(z_1, z_2, \ldots, z_n)$, where each z_i is 0 or 1. Under H_0, $\mathfrak{z}_{1/2} = \mathfrak{z}_0$ and each arrangement is equally likely. Thus

$$\text{(27)} \qquad P_{H_0}\{T^+ = t\} = \frac{\begin{array}{c}\{\text{number of ways to assign } + \text{ or } - \text{ signs to} \\ \text{integers } 1, 2, \ldots, n \text{ so that the sum is } t\}\end{array}}{2^n}$$

$$= \frac{n(t)}{2^n}, \text{ say.}$$

Note that every assignment has a conjugate assignment with plus and minus signs interchanged so that for this conjugate, T^+ is given by

$$\text{(28)} \qquad \sum_1^n i(1 - Z_{(i)}) = \frac{n(n+1)}{2} - \sum_1^n iZ_{(i)}.$$

Thus under H_0 the distribution of T^+ is symmetric about the mean $n(n+1)/4$.

Example 7. Let us compute the null distribution for $n = 3$. $E_{H_0}T^+ = n(n+1)/4 = 3$, and T^+ takes values from 0 to $n(n+1)/2 = 6$:

Value of T^+	Ranks Associated with Positive Differences	$n(t)$
6	1, 2, 3	1
5	2, 3	1
4	1, 3	1
3	1, 2; 3	2

so that

$$\text{(29)} \qquad P_{H_0}\{T^+ = t\} = \begin{cases} \frac{1}{8}, & t = 4, 5, 6, 0, 1, 2, \\ \frac{2}{8}, & t = 3, \\ 0, & \text{otherwise.} \end{cases}$$

Similarly, for $n = 4$, one can show that

$$\text{(30)} \qquad P_{H_0}\{T^+ = t\} = \begin{cases} \frac{1}{16}, & t = 0, 1, 2, 8, 9, 10, \\ \frac{2}{16}, & t = 3, 4, 5, 6, 7, \\ 0, & \text{otherwise.} \end{cases}$$

An alternative procedure would be to use the MGF technique. Under H_0, the RVs $iZ_{(i)}$ are independent and have the PMF

$$P\{iZ_{(i)} = i\} = P\{iZ_{(i)} = 0\} = \tfrac{1}{2}.$$

Thus

$$
\begin{aligned}
(31) \qquad M(t) &= Ee^{tT^+} \\
&= \prod_{i=1}^{n} \frac{e^{it}+1}{2}.
\end{aligned}
$$

We express $M(t)$ as a sum of terms of the form $\alpha_j e^{jt}/2^n$. The PMF of T^+ can then be determined by inspection. For example, in the case $n = 4$, we have

$$
\begin{aligned}
M(t) &= \prod_{i=1}^{4} \frac{e^{it}+1}{2} = \frac{e^t+1}{2}\frac{e^{2t}+1}{2}\frac{e^{3t}+1}{2}\frac{e^{4t}+1}{2} \\
(32) \quad &= \frac{1}{4}(e^{3t}+e^{2t}+e^t+1)\frac{e^{3t}+1}{2}\frac{e^{4t}+1}{2} \\
(33) \quad &= \frac{1}{8}(e^{6t}+e^{5t}+e^{4t}+2e^{3t}+e^{2t}+e^t+1)\frac{e^{4t}+1}{2} \\
(34) \quad &= \frac{1}{16}(e^{10t}+e^{9t}+e^{8t}+2e^{7t}+2e^{6t}+2e^{5t}+2e^{4t}+2e^{3t}+e^{2t}+e^t+1).
\end{aligned}
$$

This method gives us the PMF of T^+ for $n = 2$, $n = 3$, and $n = 4$ immediately. Quite simply,

(35) $P_{H_0}\{T^+ = j\}$ = coefficient of e^{jt} in the expansion of $M(t)$, $j = 0, 1, \ldots, n(n+1)/2$.

See Problem 3.3.12 for the PGF of T^+.

Example 8. Let us return to the data of Example 5 and test $H_0\colon \mathfrak{z}_{1/2} = \mu = -1.0$ against $H_1\colon \mathfrak{z}_{1/2} > -1.0$. Ranking $|x_i - \mathfrak{z}_{1/2}|$ in increasing order of magnitude, we have

$$
\begin{array}{ccccccccccccccc}
0.016 & < & 0.131 & < & 0.535 & < & 0.762 & < & 1.056 & < & 1.120 & < & 1.417 & < & 1.561 \\
5 & & 4 & & 1 & & 3 & & 7 & & 2 & & 6 & & 8
\end{array}
$$

Thus

$$
\begin{array}{llll}
r_1 = 3, & r_2 = 6, & r_3 = 4, & r_4 = 2, \\
r_5 = 1, & r_6 = 7, & r_7 = 5, & r_8 = 8
\end{array}
$$

and

$$T^+ = 3 + 6 + 4 + 2 + 7 + 5 + 8 = 35.$$

From Table ST10, H_0 is rejected at level $\alpha = 0.05$ if $T^+ \geq 31$. Since $35 > 31$, we reject H_0.

Remark 1. The Wilcoxon test statistic can also be used to test for symmetry. Let $X_1, X_2, \ldots, X_n$ be iid observations on an RV with absolutely continuous DF F. We set the null hypothesis as

$$H_0\colon \mathfrak{z}_{1/2} = \mathfrak{z}_0, \text{ and DF } F \text{ is symmetric about } \mathfrak{z}_0.$$

The alternative is

$$H_1 : \mathfrak{z}_{1/2} \neq \mathfrak{z}_0 \text{ and } F \text{ symmetric, or } F \text{ asymmetric.}$$

The test is the same since the null distribution of T^+ is the same.

Remark 2. If we have n independent pairs of observations $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ from a bivariate DF, we form the differences $Z_i = X_i - Y_i$, $i = 1, 2, \ldots, n$. Assuming that $Z_1, Z_2, \ldots, Z_n$ are (independent) observations from a population of differences with absolutely continuous DF F that is symmetric with median $\mathfrak{z}_{1/2}$, we can use the Wilcoxon statistic to test $H_0\colon \mathfrak{z}_{1/2} = \mathfrak{z}_0$.

We present some examples.

Example 9. For the data of Example 10.3.3, let us apply the Wilcoxon statistic to test $H_0\colon \mathfrak{z}_{1/2} = 0$ and F is symmetric against $H_1 : \mathfrak{z}_{1/2} \neq 0$ and F symmetric or F not symmetric.

The absolute values, when arranged in increasing order of magnitude, are as follows:

$$\begin{array}{cccccccccccccccc}
 & 0.057 & < & 0.068 & < & 0.137 & < & 0.261 & < & 0.323 & < & 0.464 & < & 0.482 & < & 0.486 \\
 & 13 & & 5 & & 2 & & 17 & & 4 & & 1 & & 11 & & 15 \\
< & 0.508 & < & 0.513 & < & 0.525 & < & 0.595 & < & 0.881 & < & 0.906 & < & 1.046 & & \\
 & 20 & & 7 & & 8 & & 9 & & 10 & & 6 & & 19 & & \\
< & 1.229 & < & 1.237 & < & 1.678 & < & 1.787 & < & 2.455 & & & & & & \\
 & 14 & & 18 & & 12 & & 16 & & 3 & & & & & &
\end{array}$$

Thus

$$\begin{array}{llllll}
r_1 = 6, & r_2 = 3, & r_3 = 20, & r_4 = 5, & r_5 = 2, & r_6 = 14, \\
r_7 = 10, & r_8 = 11, & r_9 = 12, & r_{10} = 13, & r_{11} = 7, & r_{12} = 18, \\
r_{13} = 1, & r_{14} = 16, & r_{15} = 8, & r_{16} = 19, & r_{17} = 4, & r_{18} = 17, \\
r_{19} = 15, & r_{20} = 9, & & & &
\end{array}$$

and

$$T^+ = 6 + 3 + 20 + 14 + 12 + 13 + 18 + 17 + 15 = 118.$$

From Table ST10 we see that H_0 cannot be rejected even at level $\alpha = 0.20$.

Example 10. Returning to the data of Example 6, we apply the Wilcoxon test to the differences $Z_i = X_i - Y_i$. The differences are $-6, 3, 1, -8, -17, -20, -3, -3, -11, 9$. To test $H_0\colon \mathfrak{z}_{1/2} = 0$ against $H_1\colon \mathfrak{z}_{1/2} \neq 0$, we rank the absolute values of z_i in increasing order to get

$$1 < 3 = 3 = 3 < 6 < 8 < 9 < 11 < 17 < 20$$

and

$$T^+ = 1 + 2 + 7 = 10.$$

Here we have assigned ranks 2, 3, 4 to observations $+3, -3, -3$. (If we assign rank 4 to observation 3, then $T^+ = 12$ without appreciably changing the result.)

From Table ST10 we reject H_0 at $\alpha = 0.05$ if either $T^+ > 46$ or $T^+ < 9$. Since $T^+ > 9$ and < 46, we accept H_0. Note that hypothesis H_0 was also accepted by the sign test.

For large samples we use the normal approximation. In fact, from (26) we see that

$$\frac{\sqrt{n}(T^+ - ET^+)}{\binom{n}{2}} = \frac{n^{3/2}}{\binom{n}{2}}(U_1 - EU_1) + \sqrt{n}(U_2 - EU_2).$$

Clearly $U_1 - EU_1 \xrightarrow{P} 0$ and since $n^{3/2}\Big/\binom{n}{2} \to 0$, the first term $\to 0$ in probability as $n \to \infty$. By Slutsky's theorem (Theorem 6.2.15) it follows that

$$\frac{\sqrt{n}}{\binom{n}{2}}(T^+ - ET^+) \quad \text{and} \quad \sqrt{n}(U_2 - EU_2)$$

have the same limiting distribution. From Theorem 13.2.3 and Example 13.2.7 it follows that $\sqrt{n}(U_2 - EU_2)$, and hence $(T^+ - ET^+)\sqrt{n}\Big/\binom{n}{2}$, has a limiting normal distribution with mean 0 and variance

$$4\zeta_1 = 4P_F(X_1 + X_2 > 0, X_1 + X_3 > 0) - 4P_F^2(X_1 + X_2 > 0).$$

Under H_0, the RVs $iZ_{(i)}$ are independent $b(1, \frac{1}{2})$ so

$$E_{H_0}T^+ = \frac{n(n+1)}{4} \quad \text{and} \quad \text{var}_{H_0} T^+ = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{24}.$$

Also, under H_0, F is continuous and symmetric, so

$$P_F(X_1 + X_2 > 0) = \int_{-\infty}^{\infty} P_F(X_1 > -x) f(x)\, dx = \tfrac{1}{2},$$

and

$$P_F(X_1 + X_2 > 0, X_1 + X_3 > 0) = \int_{-\infty}^{\infty} [P_F(X_1 > -x)]^2 f(x)\,dx = \tfrac{1}{3},$$

Thus $4\zeta_1 = \frac{4}{3} - \frac{4}{4} = \frac{1}{3}$, so that

$$\frac{T^+ - E_{H_0}T^+}{\binom{n}{2}\sqrt{\dfrac{1}{3n}}} \xrightarrow{L} \mathcal{N}(0, 1).$$

However,

$$\frac{(\text{var}_{H_0} T^+)^{1/2}}{\binom{n}{2}\sqrt{\dfrac{1}{3n}}} = \frac{[n(n+1)(2n+4)/24]^{1/2}}{\dfrac{n(n-1)}{2}\sqrt{\dfrac{1}{3n}}} \to 1$$

as $n \to \infty$. Consequently, under H_0,

$$T^+ \sim \text{AN}\left(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\right).$$

Thus, for large enough n we can determine the critical values for a test based on T^+ by using normal approximation.

As an example, take $n = 20$. From Table ST10 the P-value associated with $t^+ = 140$ is 0.10. Using normal approximation yields

$$P_{H_0}(T^+ > 140) \approx P\left(Z > \frac{140 - 105}{27.45}\right) = P(Z > 1.28) = 0.10003$$

PROBLEMS 13.3

1. Prove Theorem 4.
2. A random sample of size 16 from a continuous DF on [0, 1] yields the following data: 0.59, 0.72, 0.47, 0.43, 0.31, 0.56, 0.22, 0.90, 0.96, 0.78, 0.66, 0.18, 0.73, 0.43, 0.58, 0.11. Test the hypothesis that the sample comes from $U[0, 1]$.
3. Test the goodness of fit of normality for the data of Problem 10.3.6 using the Kolmogorov–Smirnov test.
4. For the data of Problem 10.3.6, find a 0.95 level confidence band for the distribution function.
5. The following data represent a sample of size 20 from $U[0, 1]$: 0.277, 0.435, 0.130, 0.143, 0.853, 0.889, 0.294, 0.697, 0.940, 0.648, 0.324, 0.482, 0.540, 0.152, 0.477, 0.667, 0.741, 0.882, 0.885, 0.740. Construct a 0.90 level confidence band for $F(x)$.

6. In Problem 5, test the hypothesis that the distribution is $U[0, 1]$. Take $\alpha = 0.05$.

7. For the data of Example 2, test, by means of the sign test, the null hypothesis $H_0: \mu = 1.5$ against $H_1: \mu \neq 1.5$.

8. For the data of Problem 5, test the hypothesis that the quantile of order $p = 0.20$ is 0.20.

9. For the data of Problem 10.4.8, use the sign test to test the hypothesis of no difference between the two averages.

10. Use the sign test for the data of Problem 10.4.9 to test the hypothesis of no difference in grade-point averages.

11. For the data of Problem 5, apply the signed-rank test to test $H_0: \mathfrak{z}_{1/2} = 0.5$ against $H_1: \mathfrak{z}_{1/2} \neq 0.5$.

12. For the data of Problems 10.4.8 and 10.4.9, apply the signed-rank test to the differences to test $H_0: \mathfrak{z}_{1/2} = 0$ against $H_1: \mathfrak{z}_{1/2} \neq 0$.

13.4 SOME TWO-SAMPLE PROBLEMS

In this section we consider some two-sample tests. Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be independent samples from two absolutely continuous distribution functions F_X and F_Y, respectively. The problem is to test the null hypothesis $H_0: F_X(x) = F_Y(x)$ for all $x \in \mathcal{R}$ against the usual one- and two-sided alternatives.

Tests of H_0 depend on the type of alternative specified. We state some of the alternatives of interest even though we do not consider all of these in this book.

I Location alternative: $F_Y(x) = F_X(x - \theta)$, $\theta \neq 0$.

II Scale alternative: $F_Y(x) = F_X(x/\sigma)$, $\sigma > 0$.

III Lehmann alternative: $F_Y(x) = 1 - [1 - F_X(x)]^{\theta+1}$, $\theta + 1 > 0$.

IV Stochastic alternative: $F_Y(x) \geq F_X(x)$ for all x, and $F_Y(x) > F_X(x)$ for at least one x.

V General alternative: $F_Y(x) \neq F_X(x)$ for some x.

Some comments are in order. Clearly, I through IV are special cases of V. Alternatives I and II show differences in F_X and F_Y in location and scale, respectively. Alternative III states that $P(Y > x) = [P(X > x)]^{\theta+1}$. In the special case when θ is an integer, it states that Y has the same distribution as the smallest of the $\theta + 1$ of X-variables. A similar alternative to test that is sometimes used is $F_Y(x) = [F_X(x)]^{\alpha}$ for some $\alpha > 0$ and all x. When α is an integer, this states that Y is distributed as the largest of the α X-variables. Alternative IV refers to the relative magnitudes of X's and Y's. It states that

$$P(Y \leq x) \geq P(X \leq x) \qquad \text{for all } x,$$

so that

$$(1) \qquad P(Y > x) \le P(X > x),$$

for all x. In other words, X's tend to be larger than the Y's.

Definition 1. We say that a continuous RV X is *stochastically larger* than a continuous RV Y if inequality (1) is satisfied for all x with strict inequality for some x.

A similar interpretation may be given to the one-sided alternative $F_X > F_Y$. In the special case where both X and Y are normal RVs with means μ_1, μ_2 and common variance σ^2, $F_X = F_Y$ corresponds to $\mu_1 = \mu_2$ and $F_X > F_Y$ corresponds to $\mu_1 < \mu_2$.

In this section we consider some common two-sample tests for location (case I) and stochastic ordering (case IV) alternatives. First, note that a test of stochastic ordering may also be used as a test of less restrictive location alternatives since, for example, $F_X > F_Y$ corresponds to larger Y's and hence larger location for Y. Second, we note that the chi-square test of homogeneity described in Section 10.3 can be used to test general alternatives (case V) H_1: $F(x) \ne G(x)$ for some x. Briefly, one partitions the real line into Borel sets $A_1, A_2, \ldots, A_k$. Let

$$p_{i1} = P(X_j \in A_i) \quad \text{and} \quad p_{i2} = P(Y_j \in A_i),$$

$i = 1, 2, \ldots, k$. Under H_0: $F = G$, $p_{i1} = p_{i2}$, $i = 1, 2, \ldots, k$, which is the problem of testing equality of two independent multinomial distributions discussed in Section 10.3.

We first consider a simple test of location. This test, based on the sample median of the combined sample, is a test of the equality of medians of the two DFs. It will tend to accept H_0: $F = G$ even if the shapes of F and G are different as long as their medians are equal.

13.4.1 Median Test

The combined sample $X_1, X_2, \ldots, X_m, Y_1, Y_2, \ldots, Y_n$ is ordered and a sample median is found. If $m+n$ is odd, the median is the $[(m+n+1)/2]$th value in the ordered arrangement. *If* $m + n$ is even, the median is any number between the two middle values. Let V be the number of observed values of X that are $\le$ the sample median for the combined sample. If V is large, it is reasonable to conclude that the actual median of X is smaller than the median of Y. One therefore rejects H_0: $F = G$ in favor of H_1: $F(x) \ge G(x)$ for all x and $F(x) > G(x)$ for some x if V is too large, that is, if $V \ge c$. If, however, the alternative is $F(x) \le G(x)$ for all x and $F(x) < G(x)$ for some x, the median test rejects H_0 if $V \le c$. For the two-sided alternative that $F(x) \ne G(x)$ for some x, we use the two-sided test.

We next compute the null distribution of the RV V. If $m + n = 2p$, p a positive integer, then

$$(2)\qquad P_{H_0}\{V=v\}=P_{H_0}\{\text{exactly } v \text{ of the } X_i\text{'s are} \le \text{combined median}\}$$

$$=\begin{cases}\dfrac{\dbinom{m}{v}\dbinom{n}{p-v}}{\dbinom{m+n}{p}}, & v=0,1,2,\ldots,m,\\[2ex] 0, & \text{otherwise.}\end{cases}$$

Here $0 \le V \le \min(m, p)$. If $m+n=2p+1$, $p>0$, is an integer, the $[(m+n+1)/2]$th value is the median in the combined sample, and

$$(3)\qquad P_{H_0}\{V=v\}=P\{\text{exactly } v \text{ of the } X_i\text{'s are below the } (p+1)\text{th value in the ordered arrangement}\}$$

$$=\begin{cases}\dfrac{\dbinom{m}{v}\dbinom{n}{p-v}}{\dbinom{m+n}{p}}, & v=0,1,\ldots,\min(m,p),\\[2ex] 0, & \text{otherwise.}\end{cases}$$

Remark 1. Under H_0 we expect $(m+n)/2$ observations above the median and $(m+n)/2$ below the median. One can therefore apply the chi-square test with 1 d.f. to test H_0 against the two-sided alternative.

Example 1. The following data represent lifetimes (hours) of batteries for two different brands:

Brand A:	40	30	40	45	55	30
Brand B:	50	50	45	55	60	40

The combined ordered sample is 30, 30, 40, 40, 40, 45, 45, 50, 50, 55, 55, 60. Since $m+n=12$ is even, the median is 45. Thus

$$\begin{aligned} v &= \text{number of observed values of } X \text{ that are less than or equal to } 45\\ &= 5.\end{aligned}$$

Now

$$P_{H_0}\{V \ge 5\}=\frac{\dbinom{6}{5}\dbinom{6}{1}}{\dbinom{12}{6}}+\frac{\dbinom{6}{6}\dbinom{6}{0}}{\dbinom{12}{6}}\approx 0.04.$$

Since $P_{H_0}\{V \ge 5\} > 0.025$, we cannot reject H_0, that the two samples come from the same population.

We now consider two tests of the stochastic alternatives. As mentioned earlier, they may also be used as tests of location.

13.4.2 Kolmogorov–Smirnov Test

Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be independent random samples from continuous DFs F and G, respectively. Let F_m^* and G_n^*, respectively, be the empirical DFs of the X's and Y's. Recall that F_m^* is the U-statistic for F, and G_n^*, that for G. Under H_0: $F(x) = G(x)$ for all x, we expect a reasonable agreement between the two sample DFs. We define

$$D_{m,n} = \sup_x |F_m^*(x) - G_n^*(x)|. \tag{4}$$

Then $D_{m,n}$ may be used to test H_0 against the two-sided alternative H_1: $F(x) \neq G(x)$ for some x. The test rejects H_0 at level α if

$$D_{m,n} \geq D_{m,n,\alpha}, \tag{5}$$

where $P_{H_0}\{D_{m,n} \geq D_{m,n,\alpha}\} \leq \alpha$.

Similarly, one can define the one-sided statistics

$$D_{m,n}^+ = \sup_x [F_m^*(x) - G_n^*(x)] \tag{6}$$

and

$$D_{m,n}^- = \sup_x [G_n^*(x) - F_m^*(x)], \tag{7}$$

to be used against the one-sided alternatives

$$\begin{aligned} &G(x) \leq F(x) \quad \text{for all } x \quad \text{and} \quad G(x) < F(x) \quad \text{for some } x \\ &\text{with rejection region } D_{m,n}^+ \geq D_{m,n,\alpha}^+ \end{aligned} \tag{8}$$

and

$$\begin{aligned} &F(x) \leq G(x) \quad \text{for all } x \quad \text{and} \quad F(x) < G(x) \quad \text{for some } x \\ &\text{with rejection region } D_{m,n}^- \geq D_{m,n,\alpha}^-, \end{aligned} \tag{9}$$

respectively.

For small samples, tables due to Massey [70] are available. In Table ST9 we give the values of $D_{m,n,\alpha}$ and $D_{m,n,\alpha}^+$ for some selected values of m, n, and α. Table ST8 gives the corresponding values for the $m = n$ case.

For large samples we use the limiting result due to Smirnov [105]. Let $N = mn/(m + n)$. Then

$$\lim_{m,n\to\infty} P\{\sqrt{N}\,D^+_{m,n} \le \lambda\} = \begin{cases} 1 - e^{-2\lambda^2}, & \lambda > 0, \\ 0, & \lambda \le 0, \end{cases} \tag{10}$$

and

$$\lim_{m,n\to\infty} P\{\sqrt{N}\,D_{m,n} \le \lambda\} = \begin{cases} \displaystyle\sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2\lambda^2}, & \lambda > 0, \\ 0, & \lambda \le 0. \end{cases} \tag{11}$$

Relations (10) and (11) give the distribution of $D^+_{m,n}$ and $D_{m,n}$, respectively, under H_0: $F(x) = G(x)$ for all $x \in \mathcal{R}$.

Example 2. Let us apply the test to data from Example 10. Do the two brands differ with respect to average life?

Let us first apply the Kolmogorov–Smirnov test to test H_0 that the population distribution of length of life for the two brands is the same.

x	$F^*_6(x)$	$G^*_6(x)$	$\lvert F^*_6(x) - G^*_6(x)\rvert$
30	$\frac{2}{6}$	0	$\frac{2}{6}$
40	$\frac{4}{6}$	$\frac{1}{6}$	$\frac{3}{6}$
45	$\frac{5}{6}$	$\frac{2}{6}$	$\frac{3}{6}$
50	$\frac{5}{6}$	$\frac{4}{6}$	$\frac{1}{6}$
55	1	$\frac{5}{6}$	$\frac{1}{6}$
60	1	1	0

$$D_{6,6} = \sup_x \lvert F^*_6(x) - G^*_6(x)\rvert = \frac{3}{6}.$$

From Table ST8 the critical value for $m = n = 6$ at level $\alpha = 0.05$ is $D_{6,6,0.05} = \frac{4}{6}$. Since $D_{6,6} \not> D_{6,6,0.05}$, we accept H_0 that the population distribution for the length of life for the two brands is the same.

Let us next apply the two-sample t-test. We have $\bar{x} = 40$, $\bar{y} = 50$, $s_1^2 = 90$, $s_2^2 = 50$, $s_p^2 = 70$. Thus

$$t = \frac{40 - 50}{\sqrt{70}\sqrt{\frac{1}{6} + \frac{1}{6}}} = -2.08.$$

Since $t_{10,0.025} = 2.2281$, we accept the hypothesis that the two samples come from the same (normal) population.

The second test of stochastic ordering alternatives we consider is the Mann–Whitney–Wilcoxon test, which can be viewed as a test based on a U-statistic.

13.4.3 Mann–Whitney–Wilcoxon Test

Let $X_1, X_2, \dots, X_m$ and $Y_1, Y_2, \dots, Y_n$ be independent samples from two continuous DFs, F and G, respectively. As in Example 13.2.9, let

$$T(X_i; Y_j) = \begin{cases} 1 & \text{if } X_i < Y_j, \\ 0 & \text{if } X_i \ge Y_j, \end{cases}$$

for $i = 1, 2, \dots, m$, $j = 1, 2, \dots, n$. Recall that $T(X_i; Y_j)$ is an unbiased estimator of $g(F, G) = P_{F,G}(X < Y)$ and the two-sample U-statistic for g is given by $U_1(\mathbf{X}; \mathbf{Y}) = (m, n)^{-1} \sum_{i=1}^m \sum_{j=1}^n T(X_i; Y_j)$. For notational convenience, let us write

$$U = mnU_1(\mathbf{X}; \mathbf{Y}) = \sum_{i=1}^{m} \sum_{j=1}^{n} T(X_i; Y_j). \tag{12}$$

Then U is the number of values of $X_1, X_2, \dots, X_m$ that are smaller than each of $Y_1, Y_2, \dots, Y_n$. The statistic U is called the *Mann–Whitney statistic*. An alternative equivalent form using Wilcoxon scores is the linear rank statistic given by

$$W = \sum_{j=1}^{n} Q_j, \tag{13}$$

where Q_j = rank of Y_j among the combined $m + n$ observations. Indeed,

$$Q_j = \text{rank of } Y_j = (\text{no. of } X_i\text{'s} < Y_j) + \text{ rank of } Y_j \text{ in } Y\text{'s}.$$

Thus

$$W = \sum_{j=1}^{n} Q_j = U + \sum_{j=1}^{n} j = U + \frac{n(n+1)}{2}, \tag{14}$$

so that U and W are equivalent test statistics—hence the name *Mann–Whitney–Wilcoxon test*. We restrict our attention to U as the test statistic.

Example 3. Let $m = 4$, $n = 3$, and suppose that the combined sample when ordered is as follows:

$$x_2 < x_1 < y_3 < y_2 < x_4 < y_1 < x_3.$$

Then $U = 7$, since there are three values of $x < y_1$, two values of $x < y_2$, and two values of $x < y_3$. Also, $W = 13$, so $U = 13 - 3(4)/2 = 7$.

Note that $U = 0$ if all the X_i's are larger than all the Y_j's and $U = mn$ if all the X_i's are smaller than all the Y_j's, because then there are m X's $< Y_1$, m X's $< Y_2$, and so on. Thus $0 \le U \le mn$. If U is large, the values of Y tend to be larger than the values of X (Y is stochastically larger than X), and this supports the alternative $F(x) \ge G(x)$ for all x and $F(x) > G(x)$ for some x. Similarly, if U is small, the Y values tend to be smaller than the X values, and this supports the alternative $F(x) \le G(x)$ for all x and $F(x) < G(x)$ for some x. We summarize these results as follows:

H_0	H_1	Reject H_0 if:
$F = G$	$F \ge G$	$U \ge c_1$
$F = G$	$F \le G$	$U \le c_2$
$F = G$	$F \ne G$	$U \ge c_3$ or $U \le c_4$

To compute the critical values we need the null distribution of U. Let

$$p_{m,n}(u) = P_{H_0}\{U = u\}. \tag{15}$$

We will set up a difference equation relating $p_{m,n}$ to $p_{m-1,n}$ and $p_{m,n-1}$. If the observations are arranged in increasing order of magnitude, the largest value can be either an x value or a y value. Under H_0, all $m + n$ values are equally likely, so the probability that the largest value will be an x value is $m/(m+n)$ and that it will be a y value is $n/(m+n)$.

Now, if the largest value is an x, it does not contribute to U, and the remaining $m - 1$ values of x and n values of y can be arranged to give the observed value $U = u$ with probability $p_{m-1,n}(u)$. If the largest value is a Y, this value is larger than all the m x's. Thus, to get $U = u$, the remaining $n - 1$ values of Y and m values of x contribute $U = u - m$. It follows that

$$p_{m,n}(u) = \frac{m}{m+n} p_{m-1,n}(u) + \frac{n}{m+n} p_{m,n-1}(u - m). \tag{16}$$

If $m = 0$, then for $n \ge 1$,

$$p_{0,n}(u) = \begin{cases} 1 & \text{if } u = 0, \\ 0 & \text{if } u > 0. \end{cases} \tag{17}$$

If $n = 0$, $m \ge 1$, then

$$p_{m,0}(u) = \begin{cases} 1 & \text{if } u = 0, \\ 0 & \text{if } u > 0, \end{cases} \tag{18}$$

and

$$p_{m,n}(u) = 0 \qquad \text{if } u < 0, \quad m \ge 0, \quad n \ge 0. \tag{19}$$

For small values of m and n, one can easily compute the null PMF of U. Thus, if $m = n = 1$, then

$$p_{1,1}(0) = \tfrac{1}{2} \quad \text{and} \quad p_{1,1}(1) = \tfrac{1}{2}.$$

If $m = 1, n = 2$, then

$$p_{1,2}(0) = p_{1,2}(1) = p_{1,2}(2) = \tfrac{1}{3}.$$

Tables for critical values are available for small values of m and n, $m \leq n$ (see, e.g., Auble [2] or Mann and Whitney [69]). Table ST11 gives the values of u_α for which $P_{H_0}\{U > u_\alpha\} \leq \alpha$ for selected values of m, n, and α.

If m, n are large, we can use the asymptotic normality of U. In Example 13.2.10 we showed that under H_0,

$$\frac{U/(mn) - \frac{1}{2}}{\sqrt{(m+n+1)/12mn}} \xrightarrow{L} \mathcal{N}(0,1)$$

as $m, n \to \infty$ such that $m/(m+n) \to$ constant. The approximation is fairly good for $m, n \geq 8$.

Example 4. Two samples are as follows:

Values of X_i:	1, 2, 3, 5, 7, 9, 11, 18
Values of Y_i:	4, 6, 8, 10, 12, 13, 14, 15, 19

Thus $m = 8, n = 9$, and $U = 3 + 4 + 5 + 6 + 7 + 7 + 7 + 7 + 8 = 54$. The (exact) P-value is $P_{H_0}(U \geq 54) = 0.046$, so we reject H_0 at (two-tailed) level $\alpha = 0.1$. Let us apply the normal approximation. We have

$$E_{H_0}U = \frac{8 \cdot 9}{2} = 36, \qquad \text{var}_{H_0}(U) = \frac{8 \cdot 9}{12}(8 + 9 + 1) = 108,$$

and

$$Z = \frac{54 - 36}{\sqrt{108}} = \frac{18}{6\sqrt{3}} = \sqrt{3} = 1.732.$$

We note that $P(Z > 1.73) = 0.042$.

PROBLEMS 13.4

1. For the data of Example 4, apply the median test.

2. Twelve 4-year-old boys and twelve 4-year-old girls were observed during two 15-minute play sessions, and each child's play during these two periods was scored as follows for incidence and degree of aggression:

Boys: 86, 69, 72, 65, 113, 65, 118, 45, 141, 104, 41, 50

Girls: 55, 40, 22, 58, 16, 7, 9, 16, 26, 36, 20, 15

Test the hypothesis that there were gender differences in the amount of aggression shown, using (a) the median test, and (b) the Mann–Whitney–Wilcoxon test. (Siegel [103])

3. To compare the variability of two brands of tires, the following mileages (1000 miles) were obtained for eight tires of each kind:

   Brand A: 32.1, 2.6, 17.8, 28.4, 19.6, 21.4, 19.9, 3.1

   Brand B: 19.8, 27.6, 3.8, 27.6, 34.1, 18.7, 16.9, 17.9

   Test the null hypothesis that the two samples come from the same population, using the Mann–Whitney–Wilcoxon test.

4. Use the data of Problem 2 to apply the Kolmogorov–Smirnov test.

5. Apply the Kolmogorov–Smirnov test to the data of Problem 3.

6. Yet another test for testing $H_0: F = G$ against general alternatives is the *runs test.* A *run* is a succession of one or more identical symbols which are preceded and followed by a different symbol (or no symbol). The *length* of a run is the number of like symbols in a run. The total number of runs, R, in the combined sample of X's and Y's when arranged in increasing order can be used as a test of H_0. Under H_0 the X and Y symbols are expected to be well mixed. A small value of R supports $H_1: F \neq G$. A test based on R is appropriate only for two-sided (general) alternatives. Tables of critical values are available. For large samples, one uses normal approximation:

$$R \sim \text{AN}\left(1 + \frac{2mn}{m+n}, \frac{2mn(2mn - m - n)}{(m+n-1)(m+n)^2}\right).$$

   (a) Let R_1 = number of X-runs, R_2 = number of Y-runs, and $R = R_1 + R_2$. Under H_0, show that

$$P(R_1 = r_1, R_2 = r_2) = k\frac{\binom{m-1}{r_1-1}\binom{n-1}{r_2-1}}{\binom{m+n}{m}},$$

   where $k = 2$ if $r_1 = r_2$, $= 1$ if $|r_1 - r_2| = 1$, $r_2 = 1, 2, \dots, m$ and $r_2 = 1, 2, \dots, n$.

   (b) Show that

$$P_{H_0}(R_1 = r_1) = \binom{m-1}{r_1-1}\binom{n+1}{r_1}\Big/\binom{m+n}{m}, \qquad 0 \le r_1 \le m.$$

7. Fifteen 3-year-old boys and fifteen 3-year-old girls were observed during two sessions of recess in a nursery school. Each child's play was scored for incidence and degree of aggression as follows:

Boys:	96,	65,	74,	78,	82,	121,	68,	79,	111,	48,	53,	92,	81,	31,	40
Girls:	12,	47,	32,	59,	83,	14,	32,	15,	17,	82,	21,	34,	9,	15,	51

Is there evidence to suggest that there are gender differences in the incidence and amount of aggression? Use both Mann–Whitney–Wilcoxon and runs tests.

13.5 TESTS OF INDEPENDENCE

Let X and Y be two RVs with joint DF $F(x, y)$, and let F_1 and F_2, respectively, be the marginal DFs of X and Y. In this section we study some tests of the hypothesis of independence, namely,

$$H_0\colon\ F(x, y) = F_1(x)F_2(y) \qquad \text{for all } (x, y) \in \mathcal{R}_2$$

against the alternative

$$H_1\colon\ F(x, y) \neq F_1(x)F_2(y) \qquad \text{for some } (x, y).$$

If the joint distribution function F is bivariate normal, we know that X and Y are independent if and only if the correlation coefficient $\rho = 0$. In this case, the test of independence is to test $H_0\colon \rho = 0$.

In the nonparametric situation the most commonly used test of independence is the chi-square test, which we now study.

13.5.1 Chi-Square Test of Independence (Contingency Tables)

Let X and Y be two RVs, and suppose that we have n observations on (X, Y). Let us divide the space of values assumed by X (the real line) into r mutually exclusive intervals $A_1, A_2, \dots, A_r$. Similarly, the space of values of Y is divided into c disjoint intervals $B_1, B_2, \dots, B_c$. As a rule of thumb, we choose the length of each interval in such a way that the probability that $X(Y)$ lies in an interval is approximately $(1/r)(1/c)$. Moreover, it is desirable to have n/r and n/c at least equal to 5. Let X_{ij} denote the number of pairs (X_k, Y_k), $k = 1, 2, \dots, n$, that lie in $A_i \times B_j$, and let

$$(1) \qquad p_{ij} = P\{(X, Y) \in A_i \times B_j\} = P\{X \in A_i \text{ and } Y \in B_j\},$$

where $i = 1, 2, \dots, r$, $j = 1, 2, \dots, c$. If each p_{ij} is known, the quantity

$$(2) \qquad \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(X_{ij} - np_{ij})^2}{np_{ij}}$$

has approximately a chi-square distribution with $rc - 1$ d.f., provided that n is large (see Theorem 10.3.2). If X and Y are independent, $P\{(X, Y) \in A_i \times B_j\} = P\{X \in A_i\}P\{Y \in B_j\}$. Let us write $p_{i\cdot} = P\{X \in A_i\}$ and $p_{\cdot j} = P\{Y \in B_j\}$. Then under H_0: $p_{ij} = p_{i\cdot}p_{\cdot j}$, $i = 1, 2, \ldots, r$, $j = 1, 2, \ldots, c$. In practice, p_{ij} will not be known. We replace p_{ij} by their estimates. Under H_0, we estimate $p_{i\cdot}$ by

$$(3) \qquad \hat{p}_{i\cdot} = \frac{\sum_{j=1}^{c} X_{ij}}{n}, \qquad i = 1, 2, \ldots, r,$$

and $p_{\cdot j}$ by

$$(4) \qquad \hat{p}_{\cdot j} = \sum_{i=1}^{r} \frac{X_{ij}}{n}, \qquad j = 1, 2, \ldots, c.$$

Since $\sum_{j=1}^{c} \hat{p}_{\cdot j} = 1 = \sum_{1}^{r} \hat{p}_{i\cdot}$, we have estimated only $r - 1 + c - 1 = r + c - 2$ parameters. It follows (see Theorem 1.3.4) that the RV

$$(5) \qquad U = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(X_{ij} - n\hat{p}_{i\cdot}\hat{p}_{\cdot j})^2}{n\hat{p}_{i\cdot}\hat{p}_{\cdot j}}$$

is asymptotically distributed as χ^2 with $rc - 1 - (r + c - 2) = (r - 1)(c - 1)$ d.f., under H_0. The null hypothesis is rejected if the computed value of U exceeds $\chi^2_{(r-1)(c-1),\alpha}$.

It is frequently convenient to list the observed and expected frequencies of the rc events $A_i \times B_j$ in an $r \times c$ table, called a *contingency table*, as follows:

	Observed Frequency O_{ij}			Expected Frequency E_{ij}		
	B_1	$B_2 \cdots B_c$		B_1	$B_2 \cdots B_c$	
A_1	X_{11}	$X_{12} \cdots X_{1c}$	$\sum X_{1j}$	$np_{1\cdot}p_{\cdot 1}$	$np_{1\cdot}p_{\cdot 2} \cdots np_{1\cdot}p_{\cdot c}$	$np_{1\cdot}$
A_2	X_{21}	$X_{22} \cdots X_{2c}$	$\sum X_{2j}$	$np_{2\cdot}p_{\cdot 1}$	$np_{2\cdot}p_{\cdot 2} \cdots np_{2\cdot}p_{\cdot c}$	$np_{2\cdot}$
$\cdot$	$\cdot$	$\cdot\cdot\cdot\cdot\cdot$	$\cdot$	$\cdot$	$\cdot \quad \cdots$	
$\cdot$	$\cdot$	$\cdot\cdot\cdot\cdot\cdot$	$\cdot$	$\cdot$	$\cdot \quad \cdots$	
$\cdot$	$\cdot$	$\cdot\cdot\cdot\cdot\cdot$	$\cdot$	$\cdot$	$\cdot \quad \cdots$	
A_r	X_{r1}	$X_{r2} \cdots X_{rc}$	$\sum X_{rj}$	$np_{r\cdot}p_{\cdot 1}$	$np_{r\cdot}p_{\cdot 2} \cdots np_{r\cdot}p_{\cdot c}$	$np_{r\cdot}$
	$\sum X_{i1}$	$\sum X_{i2}$ $\sum X_{ic}$	n	$np_{\cdot 1}$	$np_{\cdot 2}$ $np_{\cdot c}$	n

Note that the X_{ij}'s in the table are frequencies. Once the category $A_i \times B_j$ is determined for an observation (X, Y), numerical values of X and Y are irrelevant. Next, we need to compute the expected frequency table. This is done quite simply by multiplying the row and column totals for each pair (i, j) and dividing the product by n. Then we compute the quantity

$$\sum_i \sum_j \frac{(E_{ij} - O_{ij})^2}{E_{ij}}$$

and compare it with the tabulated χ^2 value. In this form the test can be applied even to qualitative data. $A_1, A_2, \ldots, A_r$ and $B_1, B_2, \ldots, B_c$ represent the two attributes, and the null hypothesis to be tested is that the attributes A and B are independent.

Example 1. Following are the results for a random sample of 400 employees:

	Annual Income (dollars)			
Time (years) with the Same Company	Less Than 40,000	40,000–75,000	More Than 75,000	Total
< 5	50	75	25	150
5–10	25	50	25	100
10 or more	25	75	50	150
Total	100	200	100	400

If X denotes the length of service with the same company, and Y, the annual income, we wish to test the hypothesis that X and Y are independent. The expected frequencies are as follows:

	Expected Frequency for Income of:			
Time (years) with the Same Company	< 40,000	40,000–75,000	≥ 75,000	Total
< 5	37.5	75	37.5	150
5–10	25	50	25	100
≥ 10	37.5	75	37.5	150
Total	100	200	100	400

Thus

$$\begin{aligned} U &= \frac{(12.5)^2}{37.5} + \frac{0}{25} + \frac{(12.5)^2}{37.5} + 0 + 0 + 0 + \frac{(12.5)^2}{37.5} + 0 + \frac{(12.5)^2}{37.5} \\ &= 16.66. \end{aligned}$$

The number of degrees of freedom is $(3-1)(3-1) = 4$, and $\chi^2_{4,0.05} = 9.488$. Since $16.66 > 9.488$, we reject H_0 at level 0.05 and conclude that length of service with a company is not independent of annual income.

13.5.2 Kendall's Tau

Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be a sample from a bivariate population.

Definition 1. For any two pairs (X_i, Y_i) and (X_j, Y_j) we say that the relation is *perfect concordance* (or *agreement*) if

(6) $\qquad X_i < X_j$ whenever $Y_i < Y_j$ or $X_i > X_j$ whenever $Y_i > Y_j$

and that the relation is *perfect discordance* (disagreement) if

$$(7)\qquad X_i > X_j \text{ whenever } Y_i < Y_j \quad \text{or} \quad X_i < X_j \text{ whenever } Y_i > Y_j.$$

Writing π_c and π_d for the probability of perfect concordance and of perfect discordance, respectively, we have

$$(8)\qquad \pi_c = P\{(X_j - X_i)(Y_j - Y_i) > 0\}$$

and

$$(9)\qquad \pi_d = P\{(X_j - X_i)(Y_j - Y_i) < 0\},$$

and if the marginal distributions of X and Y are continuous,

$$(10)\qquad \begin{aligned}\pi_c &= [P\{Y_i < Y_j\} - P\{X_i > X_j \text{ and } Y_i < Y_j\}]\\ &\quad + [P\{Y_i > Y_j\} - P\{X_i < X_j \text{ and } Y_i > Y_j\}] = 1 - \pi_d.\end{aligned}$$

Definition 2. The measure of association between the RVs X and Y defined by

$$(11)\qquad \tau = \pi_c - \pi_d$$

is known as *Kendall's tau*.

If the marginal distributions of X and Y are continuous, we may rewrite (11), in view of (10), as follows:

$$(12)\qquad \tau = 1 - 2\pi_d = 2\pi_c - 1.$$

In particular, if X and Y are independent and continuous RVs, then

$$P\{X_i < X_j\} = P\{X_i > X_j\} = \tfrac{1}{2},$$

since then $X_i - X_j$ is a symmetric RV. Then

$$\begin{aligned}\pi_c &= P\{X_i < X_j\}P\{Y_i < Y_j\} + P\{X_i > X_j\}P\{Y_i > Y_j\}\\ &= P\{X_i > X_j\}P\{Y_i < Y_j\} + P\{X_i < X_j\}P\{Y_i > Y_j = \pi_d,\end{aligned}$$

and it follows that $\tau = 0$ for independent continuous RVs.

Note that, in general, $\tau = 0$ does not imply independence. However, for the bivariate normal distribution, $\tau = 0$ if and only if the correlation coefficient ρ between X and Y is 0, so that $\tau = 0$ if and only if X and Y are independent (Problem 6).

Let

$$(13)\qquad \psi((x_1, y_1), (x_2, y_2)) = \begin{cases} 1, & (y_2 - y_1)(x_2 - x_1) > 0,\\ 0, & \text{otherwise.}\end{cases}$$

Then $E\psi((X_1, Y_1), (X_2, Y_2)) = \tau_c = (1+\tau)/2$, and we see that τ_c is estimable of degree 2, with symmetric kernel ψ defined in (13). The corresponding one-sample U-statistic is given by

$$(14)\quad U((X_1, Y_1), \ldots, (X_n, Y_n)) = \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} \psi\left((X_i, Y_i), (X_j, Y_j)\right).$$

Then the corresponding estimator of Kendall's tau is

$$(15)\qquad T = 2U - 1$$

and is called *Kendall's sample correlation coefficient.*

Note that $-1 \le T \le 1$. To test H_0 that X and Y are independent against $H_1: X$ and Y are dependent, we reject H_0 if $|T|$ is large. Under H_0, $\tau = 0$, so that the null distribution of T is symmetric about 0. Thus we reject H_0 at level α if the observed value of T, t, satisfies $|t| > t_{\alpha/2}$, where $P\{|T| \ge t_{\alpha/2} \mid H_0\} = \alpha$.

For small values of n the null distribution can be evaluated directly. Values for $4 \le n \le 10$ are tabulated by Kendall [49]. Table ST12 gives the values of S_α for which $P\{S > S_\alpha\} \le \alpha$, where $S = \binom{n}{2}T$ for selected values of n and α.

For a direct evaluation of the null distribution we note that the numerical value of T is clearly invariant under all order-preserving transformations. It is therefore convenient to order X and Y values and assign them ranks. If we write the pairs from the smallest to the largest according to, say, X values, the number of pairs of values of $1 \le i < j \le n$ for which $Y_j - Y_i > 0$ is the number of concordant pairs, P.

Example 2. Let $n = 4$, and let us find the null distribution of T. There are 4! different permutations of ranks of Y:

Ranks of X values:	1,	2,	3,	4
Ranks of Y values:	a_1,	a_2,	a_3,	a_4

where (a_1, a_2, a_3, a_4) is one of the 24 permutations of 1, 2, 3, 4. Since the distribution is symmetric about 0, we need only compute one-half of the distribution.

P	T	Number of Permutations	$P_{H_0}\{T = t\}$
0	-1.00	1	$\frac{1}{24}$
1	-0.67	3	$\frac{3}{24}$
2	-0.33	5	$\frac{5}{24}$
3	0.00	6	$\frac{6}{24}$

Similarly, for $n = 3$, the distribution of T under H_0 is as follows:

P	T	Number of Permutations	$P_{H_0}\{T = t\}$
0	-1.00	1: (3, 2, 1)	$\frac{1}{6}$
1	-0.33	2: (2, 3, 1), (3, 1, 2)	$\frac{2}{6}$

Example 3. Two judges rank four essays as follows:

	Essay			
Judge	1	2	3	4
1, X	3	4	2	1
2, Y	3	1	4	2

To test H_0: rankings of the two judges are independent, let us arrange the rankings of the first judge from 1 to 4. Then we have:

Judge 1, X:	1,	2,	3,	4
Judge 2, Y:	2,	4,	3,	1

P = number of pairs of rankings for judge 2 such that for $j > i$, $Y_j - Y_i > 0 = 2$ [the pairs (2, 4) and (2, 3)], and

$$t = \frac{2 \cdot 2}{\binom{4}{2}} - 1 = -0.33.$$

Since

$$P_{H_0}\{|T| \geq 0.33\} = \frac{18}{24} = 0.75,$$

we cannot reject H_0.

For large n we can use an extension of Theorem 13.3.3 to bivariate case to conclude that $\sqrt{n}(U - \tau_c) \xrightarrow{L} \mathcal{N}(0, 4\zeta_1)$, where

$$\zeta_1 = \operatorname{cov}\{\psi((X_1, Y_1), (X_2, Y_2)), \psi((X_1, Y_1), (X_3, Y_3))\}.$$

Under H_0 it can be shown that

$$\frac{3\sqrt{n(n-1)}}{\sqrt{2(2n+5)}} T \xrightarrow{L} \mathcal{N}(0, 1).$$

See, for example, Kendall [49], Randles and Wolfe [83] or Gibbons [32]. Approximation is good for $n \geq 8$.

13.5.3 Spearman's Rank Correlation Coefficient

Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be a sample from a bivariate population. In Section 7.3 we defined the sample correlation coefficient by

$$R = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\left[\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2\right]^{1/2}}, \tag{16}$$

where

$$\overline{X} = n^{-1}\sum_{i=1}^{n} X_i \quad \text{and} \quad \overline{Y} = n^{-1}\sum_{i=1}^{n} Y_i.$$

If the sample values $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ are each ranked from 1 to n in increasing order of magnitude separately, and if the X's and Y's have continuous DFs, we get a unique set of rankings. The data will then reduce to n pairs of rankings. Let us write

$$R_i = \text{rank}(X_i) \quad \text{and} \quad S_i = \text{rank}(Y_i);$$

then R_i and $S_i \in \{1, 2, \ldots, n\}$. Also,

$$\sum_{1}^{n} R_i = \sum_{1}^{n} S_i = \frac{n(n+1)}{2}, \tag{17}$$

$$\overline{R} = n^{-1}\sum_{1}^{n} R_i = \frac{n+1}{2}, \quad \overline{S} = n^{-1}\sum_{1}^{n} S_i = \frac{n+1}{2}, \tag{18}$$

and

$$\sum_{1}^{n}(R_i - \overline{R})^2 = \sum_{1}^{n}(S_i - \overline{S})^2 = \frac{n(n^2-1)}{12}. \tag{19}$$

Substituting in (16), we obtain

$$R = \frac{12\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{n^3 - n} = \frac{12\sum_{1}^{n} R_i S_i}{n(n^2-1)} - \frac{3(n+1)}{n-1}. \tag{20}$$

Writing $D_i = R_i - S_i = (R_i - \overline{R}) - (S_i - \overline{S})$, we have

$$\begin{aligned}\sum_{i=1}^{n} D_i^2 &= \sum_{i=1}^{n}(R_i - \overline{R})^2 + \sum_{i=1}^{n}(S_i - \overline{S})^2 - 2\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S}) \\ &= \frac{1}{6}n(n^2-1) - 2\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S}),\end{aligned}$$

and it follows that

$$R = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)}. \tag{21}$$

The statistic R defined in (20) and (21) is called *Spearman's rank correlation coefficient* (see also Example 4.5.2).

From (20) we see that

$$\begin{aligned} ER &= \frac{12}{n(n^2-1)} E\left(\sum_{i=1}^{n} R_i S_i\right) - \frac{3(n+1)}{n-1} \\ &= \frac{12}{n^2-1} E(R_i S_i) - \frac{3(n+1)}{n-1}. \end{aligned} \tag{22}$$

Under H_0, the RVs X and Y are independent, so that the ranks R_i and S_i are also independent. It follows that

$$E_{H_0}(R_i S_i) = ER_i ES_i = \left(\frac{n+1}{2}\right)^2$$

and

$$E_{H_0} R = \frac{12}{n^2-1}\left(\frac{n+1}{2}\right)^2 - \frac{3(n+1)}{n-1} = 0. \tag{23}$$

Thus we should reject H_0 if the absolute value of R is large, that is, reject H_0 if

$$|R| > R_\alpha, \tag{24}$$

where $P_{H_0}\{|R| > R_\alpha\} \le \alpha$. To compute R_α we need the null distribution of R. For this purpose it is convenient to assume, without loss of generality, that $R_i = i$, $i = 1, 2, \dots, n$. Then $D_i = i - S_i$, $i = 1, 2, \dots, n$. Under H_0, X and Y being independent, the $n!$ pairs (i, S_i) of ranks are equally likely. It follows that

$$\begin{aligned} P_{H_0}\{R = r\} &= (n!)^{-1} \times (\text{number of pairs for which } R = r) \\ &= \frac{n_r}{n!}, \text{ say.} \end{aligned} \tag{25}$$

Note that $-1 \le R \le 1$, and the extreme values can occur only when either the rankings match, that is, $R_i = S_i$, in which case $R = 1$, or $R_i = n + 1 - S_i$, in which case $R = -1$. Moreover, one need compute only one-half of the distribution, since it is symmetric about 0 (Problem 7).

In the following example we compute the distribution of R for $n = 3$ and 4. The exact complete distribution of $\sum_{i=1}^{n} D_i^2$, and hence R, for $n \le 10$ has been tabulated by Kendall [49]. Table ST13 gives values of R_α for selected values of n and α.

Example 4. Let us first enumerate the null distribution of R for $n = 3$. This is done in the following table:

(s_1, s_2, s_3)	$\sum_{i=1}^{n} is_i$	$r = \dfrac{12\sum_1^n is_i}{n(n^2-1)} - \dfrac{3(n+1)}{n-1}$
(1, 2, 3)	14	1.0
(1, 3, 2)	13	0.5
(2, 1, 3)	13	0.5

Thus

$$P_{H_0}\{R = r\} = \begin{cases} \frac{1}{6}, & r = 1.0, \\ \frac{2}{6}, & r = 0.5, \\ \frac{2}{6}, & r = -0.5, \\ \frac{1}{6}, & r = -1.0. \end{cases}$$

Similarly, for $n = 4$ we have the following:

(s_1, s_2, s_3, s_4)	$\sum_1^n is_i$	r	n_r	$P_{H_0}\{R = r\}$
(1, 2, 3, 4)	30	1	1	$\frac{1}{24}$
(1, 3, 2, 4), (2, 1, 3, 4), (1, 2, 4, 3)	29	0.8	3	$\frac{3}{24}$
(2, 1, 4, 3)	28	0.6	1	$\frac{1}{24}$
(1, 3, 4, 2), (1, 4, 2, 3), (2, 3, 1, 4), (3, 1, 2, 4)	27	0.4	4	$\frac{4}{24}$
(1, 4, 3, 2), (3, 2, 1, 4)	26	0.2	2	$\frac{2}{24}$
	25	0.0	2	$\frac{2}{24}$

The last value is obtained from symmetry.

Example 5. In Example 3 we see that

$$r = \frac{12 \times 23}{4 \times 15} - \frac{3 \times 5}{3} = -0.4.$$

Since $P_{H_0}\{|R| \geq 0.4\} = 18/24 = 0.75$, we cannot reject H_0 at $\alpha = 0.05$ or $\alpha = 0.10$.

For large samples it is possible to use a normal approximation. It can be shown (see, for example, Fraser [29, pp. 247–248]) that under H_0 the RV

$$Z = \left(12\sum_{i=1}^{n} R_i S_i - 3n^3\right) n^{-5/2}$$

or, equivalently,

$$Z = R\sqrt{n-1}$$

has approximately a standard normal distribution. The approximation is good for $n \geq 10$.

PROBLEMS 13.5

1. A sample of 240 men was classified according to characteristics A and B. Characteristic A was subdivided into four classes, A_1, A_2, A_3, and A_4, while B was subdivided into three classes, B_1, B_2, and B_3, with the following result:

	A_1	A_2	A_3	A_4	
B_1	12	25	32	11	80
B_2	17	18	22	23	80
B_3	21	17	16	26	80
	50	60	70	60	240

Is there evidence to support the theory that A and B are independent?

2. The following data represent the blood types and ethnic groups of a sample of Iraqi citizens:

	Blood Type			
Ethnic Group	O	A	B	AB
Kurd	531	450	293	226
Arab	174	150	133	36
Jew	42	26	26	8
Turkoman	47	49	22	10
Ossetian	50	59	26	15

Is there evidence to conclude that blood type is independent of ethnic group?

3. In a public opinion poll, a random sample of 500 American adults across the country was asked the following question: "Do you believe that there was a concerted effort to cover up the Watergate scandal? Answer yes, no, or no opinion." The responses according to political beliefs were as follows:

Political Affiliation	Response Yes	No	No Opinion	Total
Republican	45	75	30	150
Independent	85	45	20	150
Democrat	140	30	30	200
Total	270	150	80	500

Test the hypothesis that attitude toward the Watergate cover-up is independent of political party affiliation.

4. A random sample of 100 families in Bowling Green, Ohio, showed the following distribution of home ownership by family income:

	Annual Income (dollars)		
Residential Status	Less Than 30,000	30,000–50,000	50,000 or Above
Homeowner	10	15	30
Renter	8	17	20

Is home ownership in Bowling Green independent of family income?

5. In a flower show the judges agreed that five exhibits were outstanding, and these were numbered arbitrarily from 1 to 5. Three judges each arranged these five exhibits in order of merit, giving the following rankings:

Judge A:	5,	3,	1,	2,	4
Judge B:	3,	1,	5,	4,	2
Judge C:	5,	2,	3,	1,	4

Compute the average values of Spearman's rank correlation coefficient R and Kendall's sample tau coefficient T from the three possible pairs of rankings.

6. For the bivariate normally distributed RV (X, Y), show that $\tau = 0$ if and only if X and Y are independent. [*Hint:* Show that $\tau = (2/\pi)\sin^{-1}\rho$, where ρ is the correlation coefficient between X and Y.]

7. Show that the distribution of Spearman's rank correlation coefficient R is symmetric about 0 under H_0.

8. In Problem 5, test the null hypothesis that rankings of judge A and judge C are independent. Use both Kendall's tau and Spearman's rank correlation tests.

9. A random sample of 12 couples showed the following distribution of heights:

	Height (in.)			Height (in.)	
Couple	Husband	Wife	Couple	Husband	Wife
1	80	72	7	74	68
2	70	60	8	71	71
3	73	76	9	63	61
4	72	62	10	64	65
5	62	63	11	68	66
6	65	46	12	67	67

(a) Compute T.

(b) Compute R.

(c) Test the hypothesis that the heights of husband and wife are independent, using T as well as R. In each case use the normal approximation.

13.6 SOME APPLICATIONS OF ORDER STATISTICS

In this section we consider some applications of order statistics. We are mainly interested in three applications: tolerance intervals for distributions, coverages, and confidence interval estimates for quantiles and location parameters.

Definition 1. Let F be a continuous DF. A *tolerance interval* for F with tolerance coefficient γ is a random interval such that the probability is γ that this random interval covers at least a specific percentage $(100p)$ of the distribution.

Let $X_1, X_2, \ldots, X_n$ be a sample of size n from F, and let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ be the corresponding set of order statistics. If the endpoints of the tolerance interval are two order statistics $X_{(r)}, X_{(s)}, r < s$, we have

$$(1) \qquad P\{P\{X_{(r)} < X < X_{(s)}\} \geq p\} = \gamma.$$

Since F is continuous, $F(X)$ is $U(0, 1)$, and we have

$$(2) \qquad \begin{aligned} P\{X_{(r)} < X < X_{(s)}\} &= P\{X < X_{(s)}\} - P\{X \leq X_{(r)}\} \\ &= F(X_{(s)}) - F(X_{(r)}) \\ &= U_{(s)} - U_{(r)}, \end{aligned}$$

where $U_{(r)}, U_{(s)}$ are the order statistics from $U(0, 1)$. Thus (1) reduces to

$$(3) \qquad P\{U_{(s)} - U_{(r)} \geq p\} = \gamma.$$

The statistic $V = U_{(s)} - U_{(r)}$, $1 \leq r < s \leq n$, is called the *coverage of the interval* $(X_{(r)}, X_{(s)})$. More precisely, the differences $V_k = F(X_{(k)}) - F(X_{(k-1)}) =$

$U_{(k)} - U_{(k-1)}$, for $k = 1, 2, \ldots, n+1$, where $U_{(0)} = -\infty$ and $U_{(n+1)} = 1$, are called *elementary coverages*.

Since the joint PDF of $U_{(1)}, U_{(2)}, \ldots, U_{(n)}$ is given by

$$f(u_1, u_2, \ldots, u_n) = \begin{cases} n!, & 0 < u_1 < u_2 < \cdots < u_n, \\ 0, & \text{otherwise,} \end{cases}$$

the joint PDF of $V_1, V_2, \ldots, V_n$ is easily seen to be

$$(4) \qquad h(v_1, v_2, \ldots, v_n) = \begin{cases} n!, & v_i \geq 0,\ i = 1, 2, \ldots, n,\ \sum_1^n v_i \leq 1 \\ 0, & \text{otherwise.} \end{cases}$$

Note that h is symmetric in its arguments. Consequently, V_i's are exchangeable RVs and the distribution of every sum of r, $r < n$, of these coverages is the same, and in particular, it is the distribution of $U_{(r)} = \sum_{j=1}^r V_j$, namely,

$$(5) \qquad g_r(u) = \begin{cases} n\binom{n-1}{r-1} u^{r-1}(1-u)^{n-r}, & 0 < u < 1 \\ 0, & \text{otherwise.} \end{cases}$$

The common distribution of elementary coverages is

$$g_1(u) = n(1-u)^{n-1}, \qquad 0 < u < 1, \qquad = 0, \text{ otherwise.}$$

Thus $EV_i = 1/(n+1)$ and $\sum_{i=1}^r EV_i = r/(n+1)$. This may be interpreted as follows: The order statistics $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ partition the area under the PDF in $n+1$ parts such that each part has the same average (expected) area.

The sum of any r successive elementary coverages $V_{i+1}, V_{i+1}, \ldots, V_{i+r}$ is called an *r-coverage*. Clearly,

$$(6) \qquad \sum_{j=1}^{r} V_{i+j} = U_{(i+r)} - U_{(i)}, \qquad i + r \leq n,$$

and, in particular, $U_{(s)} - U_{(r)} = \sum_{j=r+1}^{s} V_j$. Since V's are exchangeable, it follows that

$$(7) \qquad U_{(s)} - U_{(r)} \overset{d}{=} U_{(s-r)}$$

with PDF

$$g_{s-r}(u) = n\binom{n-1}{s-r-1} u^{s-r-1}(1-u)^{n-s+r}, \qquad 0 < u < 1.$$

From (3), therefore,

(8) $$\gamma = \int_p^1 g_{s-r}(u)\,du = \sum_{i=0}^{s-r-1} \binom{n}{i} p^i (1-p)^{n-i}$$

where the last equality follows from (5.3.48). Given $n,\ p,\ \gamma$ it may not always be possible to find $s-r$ to satisfy (8).

Example 1. Let $s = n$ and $r = 1$. Then

$$\gamma = \sum_{i=0}^{n-2} \binom{n}{i} p^i (1-p)^{n-i} = 1 - p^n - np^{n-1}(1-p).$$

If $p = 0.8, n = 5, r = 1$, then

$$\gamma = 1 - (0.8)^5 - 5(0.8)^4(0.2) = 0.263.$$

Thus the interval $(X_{(1)}, X_{(5)})$ in this case defines a 26 percent tolerance interval for 0.80 probability under the distribution (of X).

Example 2. Let X_1, X_2, X_3, X_4, X_5 be a sample from a continuous DF F. Let us find r and $s, r < s$, such that $(X_{(r)}, X_{(s)})$ is a 90 percent tolerance interval for 0.50 probability under F. We have

$$0.90 = P\left\{U \geq \frac{1}{2}\right\} = \sum_{i=0}^{s-r-1} \binom{5}{i} \left(\frac{1}{2}\right)^5.$$

It follows that if we choose $s - r = 4$, then $\gamma = 0.81$; and if we choose $s - r = 5$, then $\gamma = 0.969$. In this case we must settle for an interval with tolerance coefficient 0.969, exceeding the desired value 0.90.

In general, given $p, 0 < p < 1$, it is possible to choose a sufficiently large sample size n and a corresponding value of $s - r$ such that with probability $\geq \gamma$ an interval of the form $(X_{(r)}, X_{(s)})$ covers at least $100p$ percent of the distribution. If $s - r$ is specified as a function of n, one chooses the smallest sample size n.

Example 3. Let $p = \frac{3}{4}$ and $\gamma = 0.75$. Suppose that we want to choose the smallest sample size required such that $(X_{(2)}, X_{(n)})$ covers at least 75 percent of the distribution. Thus we want the smallest n to satisfy

$$0.75 \leq \sum_{i=0}^{n-3} \binom{n}{i} \left(\frac{3}{4}\right)^i \left(\frac{1}{4}\right)^{n-i}.$$

From Table ST1 of binomial distributions we see that $n = 14$.

We next consider the use of order statistics in constructing confidence intervals for population quantiles. Let X be an RV with a continuous DF F, $0 < p < 1$. Then the quantile of order p satisfies

$$(9) \qquad F(\mathfrak{z}_p) = p.$$

Let $X_1, X_2, \dots, X_n$ be n independent observations on X. Then the number of X_i's $< \mathfrak{z}_p$ is an RV that has a binomial distribution with parameters n and p. Similarly, the number of X_i's that are at least $\mathfrak{z}_p$ has a binomial distribution with parameters n and $1 - p$.

Let $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ be the set of order statistics for the sample. Then

$$(10) \qquad \begin{aligned} P\{X_{(r)} \le \mathfrak{z}_p\} &= P\{\text{at least } r \text{ of the } X_i\text{'s} \le \mathfrak{z}_p\} \\ &= \sum_{i=r}^{n} \binom{n}{i} p^i (1-p)^{n-i}. \end{aligned}$$

Similarly,

$$(11) \qquad \begin{aligned} P\{X_{(s)} \ge \mathfrak{z}_p\} &= P\{\text{at least } n-s+1 \text{ of the } X_i\text{'s} \ge \mathfrak{z}_p\} \\ &= P\{\text{at most } s-1 \text{ of the } X_i\text{'s} < \mathfrak{z}_p\} \\ &= \sum_{i=0}^{s-1} \binom{n}{i} p^i (1-p)^{n-i}. \end{aligned}$$

It follows from (10) and (11) that

$$(12) \quad \begin{aligned} P\{X_{(r)} \le \mathfrak{z}_p \le X_{(s)}\} &= P\{X_{(s)} \ge \mathfrak{z}_p\} - P\{X_{(r)} > \mathfrak{z}_p\} \\ &= P\{X_{(r)} \le \mathfrak{z}_p\} - 1 + P\{X_{(s)} \ge \mathfrak{z}_p\} \\ &= \sum_{i=r}^{n} \binom{n}{i} p^i (1-p)^{n-i} + \sum_{i=0}^{s-1} \binom{n}{i} p^i (1-p)^{n-i} - 1 \\ &= \sum_{i=r}^{s-1} \binom{n}{i} p^i (1-p)^{n-i}. \end{aligned}$$

It is easy to determine a confidence interval for $\mathfrak{z}_p$ from (12) once the confidence level is given. In practice, one determines r and s such that $s - r$ is as small as possible, subject to the condition that the level is $1 - \alpha$.

Example 4. Suppose that we want a confidence interval for the median ($p = \frac{1}{2}$), based on a sample of size 7 with confidence level 0.90. It suffices to find r and s, $r < s$, such that

$$\sum_{i=r}^{s-1} \binom{7}{i} \left(\frac{1}{2}\right)^7 \ge 0.90.$$

By trial and error, using the probability distribution $b(7, \frac{1}{2})$ we see that we can choose $s = 7$, $r = 2$ or $r = 1$, $s = 6$; in either case $s - r$ is minimum ($= 5$), and the confidence level is at least 0.92.

Example 5. Let us compute the number of observations required for $(X_{(1)}, X_{(n)})$ to be a 0.95 level confidence interval for the median, that is, we want to find n such that

$$P\{X_{(1)} \le \mathfrak{z}_{1/2} \le X_{(n)}\} \ge 0.95.$$

It suffices to find n such that

$$\sum_{i=1}^{n-1} \binom{n}{i} \left(\frac{1}{2}\right)^n \ge 0.95.$$

It follows from Table ST1 that $n = 6$.

Finally, we consider applications of order statistics to constructing confidence intervals for a location parameter. For this purpose we use the method of test inversion discussed in Chapter 11. We first consider confidence estimation based on the sign test of location.

Let $X_1, X_2, \ldots, X_n$ be a random sample from a symmetric, continuous DF $F(x - \theta)$ and suppose that we wish to find a confidence interval for θ. Let $R^+(\mathbf{X} - \theta_0) =$ number of X_i's $> \theta_0$ be the sign-test statistic for testing $H_0\colon \theta = \theta_0$ against $H_1 : \theta \ne \theta_0$. Clearly, $R^+(\mathbf{X} - \theta_0) \sim b(n, \frac{1}{2})$ under H_0. The sign-test rejects H_0 if

$$\min\{R^+(\mathbf{X} - \theta_0),\ R^+(\theta_0 - \mathbf{X})\} \le c \tag{13}$$

for some integer c to be determined from the level of the test. Let $r = c + 1$. Then any value of θ is acceptable provided that it is greater than the rth smallest observation and smaller than the rth largest observation, giving as the confidence interval

$$X_{(r)} < \theta < X_{(n+1-r)}. \tag{14}$$

If we want level $1 - \alpha$ to be associated with (14), we choose c so that the level of test (13) is α.

Example 6. The following 12 observations come from a symmetric, continuous DF $F(x - \theta)$:

$$-223, -380, -94, -179, 194, 25, -177, -274, -496, -507, -20, 122.$$

We wish to obtain a 95 percent confidence interval for θ. Sign test rejects H_0 if $R^+(\mathbf{X}) \ge 9$ or ≤ 2 at level 0.05. Thus

$$P\{3 < R^+(\mathbf{X} - \theta) < 10\} = 1 - 2(0.0193) = 0.9614 \ge 0.95.$$

It follows that a 95 percent confidence interval for θ is given by $(X_{(3)}, X_{(10)})$ or $(-380, 25)$.

We next consider the Wilcoxon signed-ranks test of $H_0\colon \theta = \theta_0$ to construct a confidence interval for θ. The test statistic in this case is T^+ = sum of ranks of positive $(X_i - \theta_0)$'s in the ordered $|X_i - \theta_0|$'s. From (13.3.4),

$$\begin{aligned} T^+ &= \sum_{1 \le i \le j \le n} I_{[X_i + X_j > 2\theta_0]} \\ &= \text{number of } \frac{(X_i + X_j)}{2} > \theta_0. \end{aligned}$$

Let $T_{ij} = (X_i + X_j)/2$, $1 \le i \le j \le n$ and order the $N = \binom{n+1}{2}$ T_{ij}'s in increasing order of magnitude

$$T_{(1)} < T_{(2)} < \cdots < T_{(N)}.$$

Then using the argument that converts (13) to (14), we see that a confidence interval for θ is given by

$$T_{(r)} < \theta < T_{(N+1-r)}. \tag{15}$$

Critical values c are taken from Table ST10.

***Example* 7.** For the data in Example 6, the Wilcoxon signed-rank test rejects $H_0\colon \theta = \theta_0$ at level 0.05 if $T^+ > 64$ or $T^+ < 14$. Thus

$$P\{14 \le T^+(\mathbf{X} - \theta_0) \le 64\} \ge 0.95.$$

It follows that a 95% confidence interval for θ is given by $[T_{(14)}, T_{(64)}] = [-336.5, -20]$.

PROBLEMS 13.6

1. Find the smallest values of n such that the intervals (a) $(X_{(1)}, X_{(n)})$, and (b) $(X_{(2)}, X_{(n-1)})$ contain the median with probability ≥ 0.90.

2. Find the smallest sample size required such that $(X_{(1)}, X_{(n)})$ covers at least 90 percent of the distribution with probability ≥ 0.98.

3. Find the relation between n and p such that $(X_{(1)}, X_{(n)})$ covers at least $100p$ percent of the distribution with probability $\ge 1 - p$.

4. Given γ, δ, p_0, p_1 with $p_1 > p_0$, find the smallest n such that

$$P\{F(X_{(s)}) - F(X_{(r)}) \ge p_0\} \ge \gamma$$

and

$$P\{F(X_{(s)}) - F(X_{(r)}) \geq p_1\} \leq \delta.$$

Find also $s - r$. [*Hint:* Use normal approximation to the binomial distribution.]

5. In Problem 4, find the smallest n and the associated value of $s - r$ if $\gamma = 0.95$, $\delta = 0.10$, $p_1 = 0.75$, $p_0 = 0.50$.

6. Let $X_1, X_2, \dots, X_7$ be a random sample from a continuous DF F. Compute:
   (a) $P(X_{(1)} < \mathfrak{z}_{0.5} < X_{(7)})$.
   (b) $P(X_{(2)} < \mathfrak{z}_{0.3} < X_{(5)})$.
   (c) $P(X_{(3)} < \mathfrak{z}_{0.8} < X_{(6)})$.

7. Let $X_1, X_2, \dots, X_n$ be iid with common continuous DF F.
   (a) What is the distribution of

   $$F(X_{(n-1)}) - F(X_{(j)}) + F(X_{(i)}) - F(X_{(2)})$$

   for $2 \leq i < j \leq n - 1$?
   (b) What is the distribution of $[F(X_{(n)}) - F(X_{(2)})]/[F(X_{(n)}) - F(X_{(1)})]$?

13.7 ROBUSTNESS

Most of the statistical inference problems treated in this book are parametric in nature. We have assumed that the functional form of the distribution being sampled is known except for a finite number of parameters. It is to be expected that any estimator or test of hypothesis concerning the unknown parameter constructed on this assumption will perform better than the corresponding nonparametric procedure, provided that the underlying assumptions are satisfied. It is therefore of interest to know how well the parametric optimal tests or estimators constructed for one population perform when the basic assumptions are modified. If we can construct tests or estimators that perform well for a variety of distributions, for example, there would be little point in using the corresponding nonparametric method unless the assumptions are seriously violated.

In practice, one makes many assumptions in parametric inference, and any one or all of these may be violated. Thus one seldom has accurate knowledge about the true underlying distribution. Similarly, the assumption of mutual independence or even identical distribution may not hold. Any test or estimator that performs well under modifications of underlying assumptions is usually referred to as *robust.*

In this section we first consider the effect that slight variation in model assumptions have on some common parametric estimators and tests of hypotheses. Next we consider some corresponding nonparametric competitors and show that they are quite robust.

13.7.1 Effect of Deviations from Model Assumptions on Some Parametric Procedures

Let us first consider the effect of contamination on sample mean as an estimator of the population mean. The most commonly used estimator of the population mean μ is the sample mean $\overline{X}$. It has the property of unbiasedness for all populations with finite mean. For many parent populations (normal, Poisson, Bernoulli, gamma, etc.) it is a complete sufficient statistic and hence a UMVUE. Moreover, it is consistent and has asymptotic normal distribution whenever the conditions of the central limit theorem are satisfied. Nevertheless, the sample mean is affected by extreme observations, and a single observation that is either too large or too small may make $\overline{X}$ worthless as an estimator of μ. Suppose, for example, that $X_1, X_2, \ldots, X_n$ is a sample from some normal population. Occasionally, something happens to the system, and a wild observation is obtained; that is, suppose that one is sampling from $\mathcal{N}(\mu, \sigma^2)$, say, 100α percent of the time and from $\mathcal{N}(\mu, k\sigma^2)$, where $k > 1$, $(1-\alpha)100$ percent of the time. Here both μ and σ^2 are unknown, and one wishes to estimate μ. In this case one is really sampling from the density function

$$f(x) = \alpha f_0(x) + (1-\alpha) f_1(x), \tag{1}$$

where f_0 is the PDF of $\mathcal{N}(\mu, \sigma^2)$ and f_1 is the PDF of $\mathcal{N}(\mu, k\sigma^2)$. Clearly,

$$\overline{X} = \frac{\sum_1^n X_i}{n} \tag{2}$$

is still unbiased for μ. If α is nearly 1, there is no problem since the underlying distribution is nearly $\mathcal{N}(\mu, \sigma^2)$, and $\overline{X}$ is nearly the UMVUE of μ with variance σ^2/n. If $1-\alpha$ is large (i.e., not nearly 0), then, since one is sampling from f, the variance of X_1 is σ^2 with probability α and is $k\sigma^2$ with probability $1-\alpha$, and we have

$$\text{var}_\sigma(\overline{X}) = \frac{1}{n}\,\text{var}(X_1) = \frac{\sigma^2}{n}[\alpha + (1-\alpha)k]. \tag{3}$$

If $k(1-\alpha)$ is large, $\text{var}_\sigma(\overline{X})$ is large and we see that even an occasional wild observation makes $\overline{X}$ subject to a sizable error. The presence of an occasional observation from $\mathcal{N}(\mu, k\sigma^2)$ is frequently referred to as *contamination*. The problem is that we do not know, in practice, the distribution of the wild observations, and hence we do not know the PDF f. It is known that the sample median is a much better estimator than the mean in the presence of extreme values. In the contamination model discussed above, if we use $Z_{1/2}$, the sample median of the X_i's, as an estimator of μ (which is the population median), then for large n

$$E(Z_{1/2} - \mu)^2 = \text{var}(Z_{1/2}) \approx \frac{1}{4n}\frac{1}{[f(\mu)]^2} \tag{4}$$

(see Theorem 7.5.2 and Remark 7.5.7). Since

$$\begin{aligned} f(\mu) &= \alpha f_0(\mu) + (1-\alpha) f_1(\mu) \\ &= \frac{\alpha}{\sigma\sqrt{2\pi}} + (1-\alpha)\frac{1}{\sigma\sqrt{2\pi}k} = \left(\alpha + \frac{1-\alpha}{\sqrt{k}}\right)\frac{1}{\sigma\sqrt{2\pi}}, \end{aligned}$$

we have

(5) $$\operatorname{var}(Z_{1/2}) \approx \frac{\pi\sigma^2}{2n}\frac{1}{\{\alpha + [(1-\alpha)/\sqrt{k}]\}^2}.$$

As $k \to \infty$, $\operatorname{var}(Z_{1/2}) \approx \pi\sigma^2/2n\alpha^2$. If there is no contamination, $\alpha = 1$ and $\operatorname{var}(Z_{1/2}) \approx \pi\sigma^2/2n$. Also,

$$\frac{\pi\sigma^2/2n\alpha^2}{\pi\sigma^2/2n} = \frac{1}{\alpha^2},$$

which will be close to 1 if α is close to 1. Thus the estimator $Z_{1/2}$ will not be greatly affected by how large k is, that is, how wild the observations are. We have

$$\frac{\operatorname{var}(\overline{X})}{\operatorname{var}(Z_{1/2})} = \frac{2}{\pi}[\alpha + (1-\alpha)k]\left(\alpha + \frac{(1-\alpha)}{\sqrt{k}}\right)^2 \to \infty \qquad \text{as } k \to \infty.$$

Indeed, $\operatorname{var}(\overline{X}) \to \infty$ as $k \to \infty$, whereas $\operatorname{var}(Z_{1/2}) \to \pi\sigma^2/2n\alpha^2$ as $k \to \infty$. One can check that when $k = 9$ and $\alpha \approx 0.915$, the two variances are (approximately) equal. As k becomes larger than 9 or α smaller than 0.915, $Z_{1/2}$ becomes a better estimator of μ than $\overline{X}$.

There are other flaws as well. Suppose, for example, that $X_1, X_2, \dots, X_n$ is a sample from $U(0,\theta)$, $\theta > 0$. Then both $\overline{X}$ and $T(\mathbf{X}) = (X_{(1)} + X_{(n)})/2$, where $X_{(1)} = \min(X_1, \dots, X_n)$, $X_{(n)} = \max(X_1, \dots, X_n)$, are unbiased for $EX = \theta/2$. Also, $\operatorname{var}_\theta(\overline{X}) = \operatorname{var}(X)/n = \theta^2/[12n]$, and one can show that $\operatorname{var}(T) = \theta^2/[2(n+1)(n+2)]$. It follows that the efficiency of $\overline{X}$ relative to that of T is

$$\operatorname{eff}_\theta(\overline{X} \mid T) = \frac{\operatorname{var}_\theta(T)}{\operatorname{var}_\theta(\overline{X})} = \frac{6n}{(n+1)(n+2)} < 1 \qquad \text{if } n > 2.$$

In fact, $\operatorname{eff}_\theta(\overline{X} \mid T) \to 0$ as $n \to \infty$, so that in sampling from a uniform parent $\overline{X}$ is much worse than T, even for moderately large values of n.

Let us next turn our attention to the estimation of standard deviation. Let $X_1, X_2, \dots, X_n$ be a sample from $\mathcal{N}(\mu, \sigma^2)$. Then the MLE of σ is

(6) $$\hat{\sigma} = \left[\sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n}\right]^{1/2} = \left(\frac{n-1}{n}\right)^{1/2} S.$$

Note that the lower bound for the variance of any unbiased estimator for σ is $\sigma^2/2n$. Although $\hat{\sigma}$ is not unbiased, the estimator

$$S_1 = \sqrt{\frac{n}{2}}\frac{\Gamma[(n-1)/2]}{\Gamma(n/2)}\hat{\sigma} = \sqrt{\frac{n-1}{2}}\frac{\Gamma[(n-1)/2]}{\Gamma(n/2)}S \tag{7}$$

is unbiased for σ. Also,

$$\begin{aligned} \operatorname{var}(S_1) &= \sigma^2\left\{\frac{n-1}{2}\left[\frac{\Gamma[(n-1)/2]}{\Gamma(n/2)}\right]^2 - 1\right\} \\ &= \frac{\sigma^2}{2n} + O\left(\frac{1}{n^2}\right). \end{aligned} \tag{8}$$

Thus the efficiency of S_1 (relative to the estimator with least variance $= \sigma^2/2n$) is

$$\frac{\sigma^2/2n}{\operatorname{var}(S_1)} = \frac{1}{1+\sigma^2 O(2/n)} < 1$$

and → 1 as $n \to \infty$. For small n, the efficiency of S_1 is considerably smaller than 1. Thus, for $n = 2$, $\operatorname{eff}(X_1) = 1/[2(\pi - 2)] = 0.438$, and for $n = 3$, $\operatorname{eff}(S_1) = \pi/[6(4-\pi)] = 0.61$.

Yet another estimator of σ is the sample mean deviation

$$S_2 = \frac{1}{n}\sum_{i=1}^{n}|X_i - \overline{X}|. \tag{9}$$

Note that

$$E\left(\sqrt{\frac{\pi}{2}}\frac{1}{n}\sum_{i=1}^{n}|X_i - \mu|\right) = \sqrt{\frac{\pi}{2}}E|X_i - \mu| = \sigma$$

and

$$\operatorname{var}\left(\sqrt{\frac{\pi}{2}}\frac{1}{n}\sum_{i=1}^{n}|X_i - \mu|\right) = \frac{\pi-2}{2n}\sigma^2. \tag{10}$$

If n is large enough so that $\overline{X} \approx \mu$, we see that $S_3 = \sqrt{(\pi/2)}\,S_2$ is nearly unbiased for σ with variance $[(\pi-2)/2n]\sigma^2$. The efficiency of S_3 is

$$\frac{\sigma^2(2n)}{\sigma^2[(\pi-2)/2n]} = \frac{1}{\pi - 2} < 1.$$

For large n, the efficiency of S_1 relative to S_3 is

$$\frac{\operatorname{var}(S_3)}{\operatorname{var}(S_1)} = \frac{[(\pi-2)/2n]\sigma^2}{\sigma^2/2n + O(1/n^2)} = \pi - 2 + \frac{\pi - 2}{O(2/n)} > 1.$$

Now suppose that there is some contamination. As before, let us suppose that for a proportion α of the time we sample from $\mathcal{N}(\mu, \sigma^2)$, and for a proportion $1 - \alpha$ of the time we get a wild observation from $\mathcal{N}(\mu, k\sigma^2)$, $k > 1$. Assuming that both μ and σ^2 are unknown, suppose that we wish to estimate σ. In the notation used above, let

$$f(x) = \alpha f_0(x) + (1 - \alpha) f_1(x),$$

where f_0 is the PDF of $\mathcal{N}(\mu, \sigma^2)$ and f_1 is the PDF of $\mathcal{N}(\mu, k\sigma^2)$. Let us see how even small contamination can make the maximum likelihood estimator $\hat{\sigma}$ of σ quite useless.

If $\hat{\theta}$ is the MLE of θ, and φ is a function of θ, then $\varphi(\hat{\theta})$ is the MLE of $\varphi(\theta)$. In view of (7.5.7) we get

$$(11) \qquad E(\hat{\sigma} - \sigma)^2 \approx \frac{1}{4\sigma^2} E(\hat{\sigma}^2 - \sigma^2)^2.$$

Using Theorem 7.3.5, we see that

$$(12) \qquad E(\hat{\sigma}^2 - \sigma^2)^2 \approx \frac{\mu_4 - \mu_2^2}{n}$$

(dropping the other two terms with n^2 and n^3 in the denominator), so that

$$(13) \qquad E(\hat{\sigma} - \sigma)^2 \approx \frac{1}{4\sigma^2 n}(\mu_4 - \mu_2^2).$$

For the density f, we see that

$$(14) \qquad \mu_4 = 3\sigma^4[\alpha + k^2(1 - \alpha)]$$

and

$$(15) \qquad \mu_2 = \sigma^2[\alpha + k(1 - \alpha)].$$

It follows that

$$(16) \qquad E\{\hat{\sigma} - \sigma\}^2 \approx \frac{\sigma^2}{4n}\left\{3[\alpha + k^2(1 - \alpha)] - [\alpha + k(1 - \alpha)]^2\right\}.$$

If we are interested in the effect of very small contamination, $\alpha \approx 1$ and $1 - \alpha \approx 0$. Assuming that $k(1 - \alpha) \approx 0$, we see that

$$(17) \qquad \begin{aligned} E\{\hat{\sigma} - \sigma\}^2 &\approx \frac{\sigma^2}{4n}\{3[1 + k^2(1 - \alpha)] - 1\} \\ &= \frac{\sigma^2}{2n}\left[1 + \tfrac{3}{2}k^2(1 - \alpha)\right]. \end{aligned}$$

In the normal case, $\mu_4 = 3\sigma^4$ and $\mu_2^2 = \sigma^4$, so that from (11)

$$E\{\hat{\sigma} - \sigma\}^2 \approx \frac{\sigma^2}{2n}.$$

Thus we see that the mean square error due to a small contamination is now multiplied by a factor $[1 + \frac{3}{2}k^2(1 - \alpha)]$. If, for example, $k = 10$, $\alpha = 0.99$, then $1 + \frac{3}{2}k^2(1 - \alpha) = \frac{5}{2}$. If $k = 10$, $\alpha = 0.98$, then $1 + \frac{3}{2}k^2(1 - \alpha) = 4$, and so on.

A quick comparison with S_3 shows that although S_1 (or even $\hat{\sigma}$) is a better estimator of σ than S_3 if there is no contamination, S_3 becomes a much better estimator in the presence of contamination as k becomes large.

Next we consider the effect of deviation from model assumptions on tests of hypotheses. One of the most commonly used tests in statistics is Student's t-test for testing the mean of a normal population when the variance is unknown. Let $X_1, X_2, \dots, X_n$ be a sample from some population with mean μ and finite variance σ^2. As usual, let $\overline{X}$ denote the sample mean, and S^2, the sample variance. If the population being sampled is normal, the t-test rejects $H_0\colon \mu = \mu_0$ against $H_1\colon \mu \neq \mu_0$ at level α if $|\overline{x} - \mu_0| > t_{n-1,\alpha/2}(s/\sqrt{n})$. If n is large, we replace $t_{n-1,\alpha/2}$ by the corresponding critical value, $z_{\alpha/2}$, under the standard normal law. If the sample does not come from a normal population, the statistic $T = [(\overline{X} - \mu_0)/S]\sqrt{n}$ is no longer distributed as a $t(n-1)$ statistic. If, however, n is sufficiently large, we know that T has an asymptotic normal distribution irrespective of the population being sampled, as long as it has a finite variance. Thus, for large n, the distribution of T is independent of the form of the population, and the t-test is stable. The same considerations apply to testing the difference between two means when the two variances are equal. Although we assumed that n is sufficiently large for Slutsky's result (Theorem 6.2.15) to hold, empirical investigations have shown that the test based on Student's statistic is robust. Thus a significant value of t may not be interpreted to mean a departure from normality of the observations. Let us next consider the effect of departure from independence on the t-distribution. Suppose that the observations $X_1, X_2, \dots, X_n$ have a multivariate normal distribution with $EX_i = \mu$, $\text{var}(X_i) = \sigma^2$, and ρ as the common correlation coefficient between any X_i and X_j, $i \neq j$. Then

$$(18) \qquad E\overline{X} = \mu, \quad \text{and} \quad \text{var}(\overline{X}) = \frac{\sigma^2}{n}[1 + (n-1)\rho],$$

and since X_i's are exchangeable it follows from Remark 7.3.1 that

$$(19) \qquad ES^2 = \sigma^2(1 - \rho).$$

For large n, the statistic $\sqrt{n}(\overline{X} - \mu_0)/S$ will be asymptotically distributed as $\mathcal{N}(0, 1+n\rho/(1-\rho))$, instead of $\mathcal{N}(0, 1)$. Under H_0, $\rho = 0$ and $T^2 = n(\overline{X} - \mu_0)^2/S^2$ is distributed as $F(1, n-1)$. Consider the ratio

$$(20) \qquad \frac{nE(\overline{X} - \mu_0)^2}{ES^2} = \frac{\sigma^2[1 + (n-1)\rho]}{\sigma^2(1-\rho)} = 1 + \frac{n\rho}{1-\rho}.$$

The ratio equals 1 if $\rho = 0$ but is > 0 for $\rho > 0$ and $\to \infty$ as $\rho \to 1$. It follows that a large value of T is likely to occur when $\rho > 0$ and is large, even though μ_0 is the true value of the mean. Thus a significant value of t may be due to departure from independence, and the effect can be serious.

Next, consider a test of the null hypothesis $H_0\colon \sigma = \sigma_0$ against $H_1\colon \sigma \neq \sigma_0$. Under the usual normality assumptions on the observations $X_1, X_2, \dots, X_n$, the test statistic used is

$$V = \frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2}, \tag{21}$$

which has a $\chi^2(n-1)$ distribution under H_0. The usual test is to reject H_0 if

$$V_0 = \frac{(n-1)S^2}{\sigma_0^2} > \chi^2_{n-1,\alpha/2} \quad \text{or} \quad V_0 < \chi^2_{n-1,1-\alpha/2}. \tag{22}$$

Let us suppose that $X_1, X_2, \dots, X_n$ are not normal. It follows from Corollary 2 of Theorem 7.3.5 that

$$\operatorname{var}(S^2) = \frac{\mu_4}{n} + \frac{3-n}{n(n-1)}\mu_2^2, \tag{23}$$

so that

$$\operatorname{var}\left(\frac{S^2}{\sigma^2}\right) = \frac{1}{n}\frac{\mu_4}{\sigma^4} + \frac{3-n}{n(n-1)}. \tag{24}$$

Writing $\gamma_2 = (\mu_4/\sigma^4) - 3$, we have

$$\operatorname{var}\left(\frac{S^2}{\sigma^2}\right) = \frac{\gamma_2}{n} + \frac{2}{n-1} \tag{25}$$

when the X_i's are not normal, and

$$\operatorname{var}\left(\frac{S^2}{\sigma^2}\right) = \frac{2}{n-1} \tag{26}$$

when the X_i's are normal ($\gamma_2 = 0$). Now $(n-1)S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2$ is the sum of n identically distributed but dependent RVs $(X_j - \overline{X})^2$, $j = 1, 2, \dots, n$. Using a version of the central limit theorem for dependent RVs (see, e.g., Cramér [16, p. 365]), it follows that

$$\left(\frac{n-1}{2}\right)^{-1/2}\left(\frac{S^2}{\sigma^2} - 1\right),$$

under H_0, is asymptotically $\mathcal{N}(0, 1 + (\gamma_2/2))$, and not $\mathcal{N}(0, 1)$ as under the normal theory. As a result, the size of the test based on the statistic V_0 will be different from the stated level of significance if γ_2 differs greatly from 0. It is clear that the effect of violation of the normality assumption can be quite serious on inferences about variances, and the chi-square test is not robust.

In the discussion above we have used somewhat crude calculations to investigate the behavior of the most commonly used estimators and test statistics when one or more of the underlying assumptions are violated. Our purpose here was to indicate that some tests or estimators are robust, whereas others are not. The moral is clear: One should check carefully to see that the underlying assumptions are satisfied before using parametric procedures.

13.7.2 Some Robust Procedures

Let $X_1, X_2, \dots, X_n$ be a random sample from a continuous PDF $f(x - \theta), \theta \in \mathcal{R}$, and assume that f is symmetric about θ. We shall be interested in estimation or tests of hypotheses concerning θ. Our objective is to find procedures that perform well for several different types of distributions but do not have to be optimal for any particular distribution. We will call such procedures *robust*. We first consider estimation of θ.

The estimators fall under one of the following three types:

1. Estimators that are functions of $\mathbf{R} = (R_1, R_2, \dots, R_n)$, where $\mathcal{R}_j$ is the rank of X_j, are known as *R-estimators*. Hodges and Lehmann [41] devised a method of deriving such estimators from rank tests. These include the sample median $\tilde{X}$ (based on the sign test), and $W = \{\text{med}\{(X_i + X_j)/2,\ 1 \le i \le j \le n\}$ based on the Wilcoxon signed-rank test.
2. Estimators of the form $\sum_{i=1}^{n} a_i X_{(i)}$ are called *L-estimators*, being linear combinations of order statistics. This class includes the median, the mean, and the trimmed mean obtained by dropping a prespecified proportion of extreme observations.
3. Maximum likelihood type estimators obtained as solutions to certain equations $\sum_{j=1}^{n} \psi(X_j - \theta) = 0$ are called *M-estimators*. The function $\psi(t) = -f'(t)/f(t)$ gives MLEs.

Definition 1. Let $k = [n\alpha]$ be the largest integer $\le n\alpha$, where $0 < \alpha < \frac{1}{2}$. Then the estimator

$$\overline{X}_\alpha = \sum_{j=k+1}^{n-k} \frac{X_{(j)}}{n - 2k} \tag{27}$$

is called a *trimmed mean*.

Two extreme examples of trimmed means are the sample mean $\overline{X}(\alpha = 0)$ and the median $\tilde{X}$ when all except the central (n odd) or the two central (n even) observations are excluded.

Example 1. Consider the following sample of size 15 taken from a symmetric distribution.

0.97 0.66 0.73 0.78 1.30 0.58 0.79 0.94
0.52 0.52 0.83 1.25 1.47 0.96 0.71

Suppose that $\alpha = 0.10$. Then $k = [n\alpha] = 1$ and

$$\bar{x}_{0.10} = \frac{\sum_{j=2}^{14} x_{(j)}}{15-2} = 0.85.$$

Here $\bar{x} = 0.867$, $\text{med}_{1\le j\le 15}\, x_j = x_{(8)} = 0.79$.

We limit this discussion to four estimators of location: the sample median, trimmed mean, sample mean, and Hodges–Lehmann estimator based on Wilcoxon signed-rank test. To compare the performance of two procedures A and B, we use a (large-sample) measure of relative efficiency due to Pitman. Pitman's *asymptotic relative efficiency* (ARE) of procedure B relative to procedure A is the limit of the ratio of sample sizes n_A/n_B, where n_A and n_B are sample sizes needed for procedures A and B to perform equivalently with respect to a specified criterion. For example, suppose that $\{T_{n(A)}\}$ and $\{T_{n(B)}\}$ are two sequences of estimators for $\psi(\theta)$ such that

$$T_{n(A)} \sim \text{AN}\left(\psi(\theta), \frac{\sigma_A^2(\theta)}{n(A)}\right)$$

and

$$T_{n(B)} \sim \text{AN}\left(\psi(\theta), \frac{\sigma_B^2(\theta)}{n(B)}\right).$$

Suppose further that A and B perform equivalently if their asymptotic variances are the same, that is,

$$\frac{\sigma_A^2(\theta)}{n(A)} \approx \frac{\sigma_B^2(\theta)}{n(B)}.$$

Then

$$\frac{n(A)}{n(B)} \longrightarrow \frac{\sigma_A^2(\theta)}{\sigma_B^2(\theta)}.$$

Clearly, different performance measures may lead to different measures of ARE.

Similarly, if procedures A and B lead to two sequences of tests, then ARE is the limiting ratio of the sample sizes needed by the tests to reach a certain power β_0 against the same alternative and at the same limiting level α.

Accordingly, let $e(B, A)$ denote the ARE of B relative to A. If $e(B, A) = \frac{1}{2}$, say, then procedure A requires (approximately) half as many observations as procedure B. We will write $e_F(B, A)$ whenever necessary to indicate the dependence of ARE on the underlying DF F.

For detailed discussion of Pitman efficiency we refer to Lehmann [59, pp. 371–380], Lehmann [61, Sec. 5.2], Randles and Wolfe [83, Chap. 5], Serfling [100, Chap. 10], and Zacks [120]. The expressions for AREs of median and the Hodges–Lehmann estimators of location parameter θ with respect to the sample mean $\overline{X}$ are

$$e_F(\tilde{X}, \overline{X}) = 4\sigma_F^2 f(0) \tag{28}$$

and

$$e_F(W, \overline{X}) = 12\sigma_F^2 \left[\int_{-\infty}^{\infty} f^2(x)\,dx\right]^2, \tag{29}$$

where f is the PDF corresponding to F. To get $e_F(\tilde{X}, W)$ we use the fact that

$$\begin{aligned} e_F(\tilde{X}, W) &= \frac{e_F(\tilde{X}, \overline{X})}{e_F(W, \overline{X})} \\ &= \frac{f(0)}{3\left[\int_{-\infty}^{\infty} f^2(x)\,dx\right]^2}. \end{aligned} \tag{30}$$

Bickel [4] showed that

$$e_F(\overline{X}_\alpha, \overline{X}) = \frac{\sigma_F^2}{\sigma_\alpha^2}, \tag{31}$$

where

$$\sigma_\alpha^2 = \frac{2}{(1-2\alpha)^2}\left[\int_0^{\mathfrak{z}_{1-\alpha}} t^2 f(t)\,dt + \alpha \mathfrak{z}_{1-\alpha}\right] \tag{32}$$

and $\mathfrak{z}_\alpha$ is the unique αth percentile of F. It is clear from (32) that no closed-form expression for $e_F(\overline{X}_\alpha, \overline{X})$ is possible for most DFs F.

In the following table we give the AREs for some selected F.

ARE Computations for Selected F

F	$e(\tilde{X}, \overline{X})$	$e(W, \overline{X})$	$e(\tilde{X}, W)$
$U(-\frac{1}{2}, \frac{1}{2})$	$\frac{1}{3}$	1	$\frac{1}{3}$
$\mathcal{N}(0, 1)$	$2/\pi = 0.637$	$3/\pi = 0.955$	$\frac{2}{3}$
Logistic, $f(x) = e^{-x}\left(1 + e^{-x}\right)^{-1}$	$\pi^2/12 = 0.822$	1.10	0.748
Double exponential, $f(x) = \frac{1}{2}\exp(-\lvert x\rvert)$	2	1.5	$\frac{4}{3}$
$C(0, 1)$	∞	∞	$\frac{4}{3}$

It can be shown that $e_F(\tilde{X}, \overline{X}) \geq \frac{1}{3}$ for all symmetric F, so $\tilde{X}$ is quite inefficient compared to $\overline{X}$ for $U(-\frac{1}{2}, \frac{1}{2})$. Even for normal f, $\tilde{X}$ would require 157 observations to achieve the same accuracy that $\overline{X}$ achieves with 100 observations. For heavier-tailed distributions, however, $\tilde{X}$ provides more protection that $\overline{X}$.

The values of $e(W, \overline{X})$, on the other hand, are quite high for most F and, in fact, $e_F(W, \overline{X}) \geq 0.864$ for all symmetric F. Even for normal F one loses little (4.5%) in using W instead of $\overline{X}$. Thus W is more robust as an estimator of θ.

A look at the values of $e(\tilde{X}, W)$ shows that $\tilde{X}$ is worse than W for distributions with light-tails but does slightly better than W for heavier-tailed F.

Let us now compare the AREs of $\overline{X}_\alpha$, $\overline{X}$, and W. The following AREs for selected α are due to Bickel [4].

ARE Comparisons

	$\alpha = 0.01$		$\alpha = 0.05$	
F	$e(\overline{X}_\alpha, \overline{X})$	$e(W, \overline{X}_\alpha)$	$e(\overline{X}_\alpha, \overline{X})$	$e(W, \overline{X}_\alpha)$
Uniform	0.96	1.04	0.83	1.20
Normal	0.995	0.96	0.97	0.985
Double exponential	1.06	1.41	1.21	1.24
Cauchy	∞	6.72	∞	2.67

We note that $\overline{X}_\alpha$ performs quite well compared to $\overline{X}$. In fact, for normal distribution the efficiency is quite close to 1, so there is little loss in using $\overline{X}_\alpha$. For heavier-tailed distributions, $\overline{X}_\alpha$ is preferable. For small values of α, it should be noted that $\overline{X}_\alpha$ does not differ much from $\overline{X}$. Nevertheless, $\overline{X}_\alpha$ is more robust; it cannot do much worse than $\overline{X}$ but can do much better. Compared to the Hodges–Lehmann estimator, $\overline{X}_\alpha$ does not perform as well. It (W) provides better protection against outliers (heavy tails) and gives up little in the normal case.

Finally, we consider testing $H_0\colon \theta = \theta_0$ against $H_1\colon \theta > \theta_0$. Recall that $X_1, X_2, \ldots, X_n$ are iid with common continuous symmetric DF $F(x - \theta)$, $\theta \in \mathcal{R}$ and PDF $f(x - \theta)$. Suppose that $\sigma_F^2 = \operatorname{var}(X_1) < \infty$. Let S denote the sign test based on the statistic $R^+(\mathbf{X}) = \sum_{i=1}^{n} I_{[X_i > \theta_0]}$, W denote the Wilcoxon signed-rank test based on the statistic $T^+(\mathbf{X}) = \sum_{1 \le i \le j \le n} I_{[X_i + X_j > 2\theta_0]}$, M denote the test based on the Z-statistic $Z = \sqrt{n}(\overline{X} - \theta_0)/\sigma_F$, and t denote Student's t-test based on the statistic $\sqrt{n}(\overline{X} - \theta_0)/S$, where S^2 is the sample variance.

First note that $e(T, M) = 1$. Next we note that $e_F(S, t) = e_F(\tilde{X}, \overline{X})$, $e_F(W, t) = e_F(W, \overline{X})$, so that AREs are the same as given in (28), (29), and (30), and values of ARE given in the table for various F remain the same for corresponding tests.

Similar remarks apply as for the case of estimation of θ. The sign test is not as efficient as the Wilcoxon signed-rank test. But for heavier-tailed distributions such as Cauchy and double exponential, the sign test does better than the Wilcoxon signed-rank test.

PROBLEMS 13.7

1. Let $(X_1, X_2, \ldots, X_n)$ be jointly normal with $EX_i = \mu$, $\operatorname{var}(X_i) = \sigma^2$, and $\operatorname{cov}(X_i, X_j) = \rho\sigma^2$ if $|i - j| = 1$, $i \ne j$, and $= 0$ otherwise.

   (a) Show that

$$\operatorname{var}(\overline{X}) = \frac{\sigma^2}{n}\left[1 + 2\rho\left(1 - \frac{1}{n}\right)\right]$$

   and

$$E(S^2) = \sigma^2\left(1 - \frac{2\rho}{n}\right).$$

   (b) Show that the t-statistic $\sqrt{n}(\overline{X} - \mu)/S$ is asymptotically normally distributed with mean 0 and variance $1 + 2\rho$. Conclude that the significance of t is overestimated for positive values of ρ and underestimated for $\rho < 0$ in large samples.

   (c) For finite n, consider the statistic

$$T^2 = \frac{n(\overline{X} - \mu)^2}{S^2}.$$

   Compare the expected values of the numerator and the denominator of T^2 and study the effect of $\rho \ne 0$ to interpret significant t values. (Scheffé [99, p. 338])

2. Let $X_1, X_2, \ldots, X_n$ be a random sample from $G(\alpha, \beta), \alpha > 0, \beta > 0$.

   (a) Show that

$$\mu_2 = \alpha\beta^2, \quad \text{and} \quad \mu_4 = \frac{3\alpha(\alpha + 2)}{\beta^4}.$$

   (b) Show that

$$\operatorname{var}\left\{(n-1)\frac{S^2}{\sigma^2}\right\} \approx (n-1)\left(2+\frac{6}{\alpha}\right).$$

(c) Show that the large sample distribution of $(n-1)S^2/\sigma^2$ is normal.

(d) Compare the large-sample test of $H_0\colon \sigma = \sigma_0$ based on the asymptotic normality of $(n-1)S^2/\sigma^2$ with the large-sample test based on the same statistic when the observations are taken from a normal population. In particular, take $\alpha = 2$.

3. Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be two independent random samples from populations with means μ_1 and μ_2, and variances σ_1^2 and σ_2^2, respectively. Let $\overline{X}, \overline{Y}$ be the two sample means, and S_1^2, S_2^2 be the two sample variances. Write $N = m + n$, $R = m/n$, and $\theta = \sigma_1^2/\sigma_2^2$. The usual normal theory test of $H_0\colon \mu_1 - \mu_2 = \delta_0$ is the t-test based on the statistic

$$T = \frac{\overline{X} - \overline{Y} - \delta_0}{S_p(1/m + 1/n)^{1/2}},$$

where

$$S_p^2 = \frac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2}.$$

Under H_0, the statistic T has a t-distribution with $N-2$ d.f., provided that $\sigma_1^2 = \sigma_2^2$. Show that the asymptotic distribution of T in the nonnormal case is $\mathcal{N}(0, (\theta + R)(1 + R\theta)^{-1})$ for large m and n. Thus if $R = 1$, T is asymptotically $\mathcal{N}(0, 1)$ as in the normal theory case assuming equal variances, even though the two samples come from nonnormal populations with unequal variances. Conclude that the test is robust in the case of large, equal sample sizes. (Scheffé [99, p. 339])

4. Verify the ARE computations for F in the table above using the expressions of ARE in (28), (29), and (30).

5. Suppose that F is a $G(\alpha, \beta)$ RV. Show that

$$e(W, \overline{X}) = \frac{3\alpha\Gamma^2(2\alpha)}{2^{4(\alpha-1)}(2\alpha-1)^2\{\Gamma(\alpha)\}^4}.$$

(Note that F is not symmetric.)

6. Suppose that F has PDF

$$f(x) = \frac{\Gamma(m)}{\Gamma(1/2)\Gamma((m-1)/2)(1+x^2)^m}, \qquad -\infty < x < \infty,$$

for $m \geq 1$. Compute $e(\tilde{X}, \overline{X})$, $e(W, \overline{X})$, and $e(\tilde{X}, W)$. (From Problem 3.2.3, $E|X|^k < \infty$ if $k < m - \frac{1}{2}$.)