Chapter 5

The Two-Sample Dispersion Problem and Other Two-Sample Problems

INTRODUCTION

In this chapter the data once again consist of two independent random samples, one sample from each of two underlying populations. This is the same as the data setting considered in Chapter 4, where we discussed procedures designed for statistical analyses in which the primary interest was on possible differences in the locations (medians) of the populations. In this chapter we deal with statistical procedures designed to make inferences about possible differences other than location between two populations.

In Section 5.1 we present a distribution-free rank test for the hypothesis of equal scale parameters when the two underlying populations have a common median. Section 5.2 is devoted to an asymptotically distribution-free test for equality of scale parameters when the assumption of common medians is not justified. In Section 5.3 we consider a distribution-free rank test for the dual hypothesis of equal location and equal scale parameters for the underlying populations. Section 5.4 contains a distribution-free test of the general hypothesis that two populations are identical in all respects. Some aspects of the asymptotic relative efficiencies of the procedures in this chapter with respect to their normal theory counterparts are discussed in Section 5.5.

Data. We obtain $N = m + n$ observations $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$.

Assumptions

A1. The observations $X_1, \ldots, X_m$ are a random sample from a continuous population 1; that is, the X's are mutually independent and identically distributed. The observations $Y_1, \ldots, Y_n$ are a random sample from a continuous population 2, so that the Y's are also mutually independent and identically distributed.

A2. The X's and Y's are mutually independent. Thus, in addition to assumptions of independence within each sample, we also assume independence between the two samples.

Nonparametric Statistical Methods, Third Edition. Myles Hollander, Douglas A. Wolfe, Eric Chicken.

5.1 A DISTRIBUTION-FREE RANK TEST FOR DISPERSION—MEDIANS EQUAL (ANSARI–BRADLEY)

Hypothesis

Let F and G be the distribution functions corresponding to populations 1 and 2, respectively. The null hypothesis of interest here is that the X and Y variables have the same probability distribution but that their common distribution is not specified. Formally stated, this null hypothesis is

$$H_0 : [F(t) = G(t), \text{ for every } t]. \tag{5.1}$$

The typical alternative hypothesis in a two-sample dispersion problem specifies that the Y population has greater (or less) variability associated with it than does the X population. One model that is often used to describe such alternatives is the location-scale parameter model. In our two-sample setting, this location-scale parameter model corresponds to taking

$$F(t) = H\left(\frac{t-\theta_1}{\eta_1}\right) \quad \text{and} \quad G(t) = H\left(\frac{t-\theta_2}{\eta_2}\right), \quad -\infty < t < \infty, \tag{5.2}$$

where $H(u)$ is the distribution function for a continuous distribution with median 0, so that $F(\theta_1) = G(\theta_2) = \frac{1}{2}$. Thus, θ_1 and θ_2 are the population medians for the X and Y distributions, respectively. Moreover, η_1 and η_2 are the scale parameters associated with the X and Y distributions, respectively. Model (5.2) states that the Y population has the same general form as the X population, but they could have different medians and scale parameters. Another way to express this is to write

$$\frac{X-\theta_1}{\eta_1} \stackrel{d}{=} \frac{Y-\theta_2}{\eta_2}, \tag{5.3}$$

where the symbol $\stackrel{d}{=}$ means "has the same distribution as."

This two-sample location-scale problem will be further discussed in this most general context in Sections 5.2 and 5.3. In this section, however, we impose the further restriction that $\theta_1 = \theta_2$; that is, we also assume

A3. The median (θ_1) of the X population is equal to the median (θ_2) of the Y population.

Under this additional assumption, A3, the equal-in-distribution statement in (5.3) simplifies to

$$\frac{X-\theta}{\eta_1} \stackrel{d}{=} \frac{Y-\theta}{\eta_2}, \tag{5.4}$$

where θ is the common median and the only possible difference between the X and Y populations is in their respective scale parameters, as illustrated in Figure 5.1. (If the medians θ_1 and θ_2 of the X and Y populations are not necessarily equal but are known, the shifted variables $X_1 - \theta_1, \ldots, X_m - \theta_1$ and $Y_1 - \theta_2, \ldots, Y_n - \theta_2$ will satisfy Assumptions A1, A2, and A3. In such a situation, the procedures of this section can be applied to the shifted $(X - \theta_1)$ and $(Y - \theta_2)$ sample observations. For more about this known medians setting, see Comment 1.)

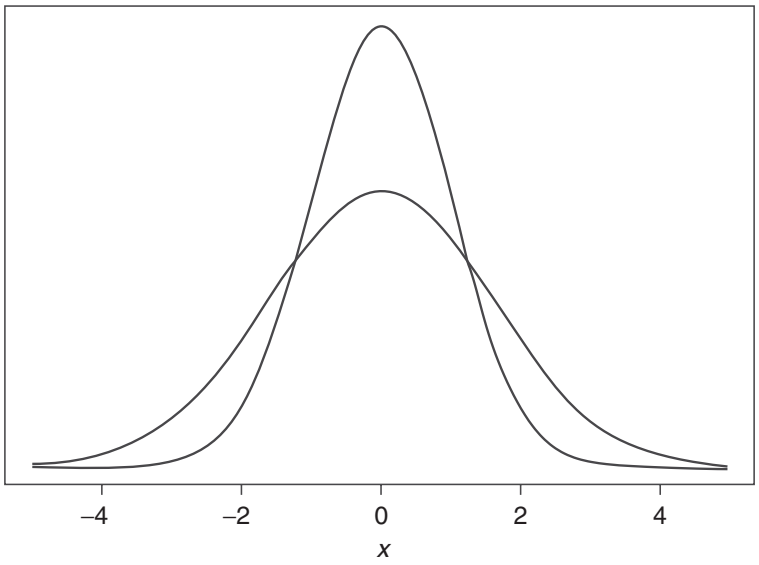

Figure 5.1 Probability distributions with the same general form and equal medians but different scale parameters.

Under Assumptions A1–A3, the parameter of interest in this section is the ratio of the scale parameters, $\gamma = (\eta_1/\eta_2)$ (see Comment 3). If the variance of population 1, var(X), exists and (5.4) is satisfied, then the variance of population 2, var(Y), also exists and

$$\gamma^2 = \left[\frac{\text{var}(X)}{\text{var}(Y)}\right], \tag{5.5}$$

the ratio of population variances (also see Comment 7). In terms of this location-scale parameter model with equal location parameters, as given in (5.4), the null hypothesis H_0 (5.1) reduces to $H_0 : \gamma^2 = 1$, corresponding to the assertion that the population scale parameters are equal.

Procedure

To compute the Ansari–Bradley two-sample scale statistic C, order the combined sample of $N = (m + n)$ X-values and Y-values from least to greatest. Assign the score 1 to both the smallest and largest observations in this combined sample, assign the score 2 to the second smallest and second largest, and continue in the manner. If N is an even integer, the array of assigned scores is $1, 2, 3, \ldots, N/2, N/2, \ldots, 3, 2, 1$. If N is an odd integer, the array of assigned scores is $1, 2, 3, \ldots, (N - 1)/2, (N + 1)/2, (N - 1)/2, \ldots, 3, 2, 1$. Let R_j denote the score assigned in this manner to Y_j, for $j = 1, \ldots, n$, and set

$$C = \sum_{j=1}^{n} R_j. \tag{5.6}$$

Thus the statistic C is the sum of the scores assigned via this scheme to the Y observations.

a. *One-Sided Upper-Tail Test*. To test

$$H_0 : \gamma^2 = 1$$

versus

$$H_1 : \gamma^2 > 1,$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } C \geq c_\alpha; \quad \text{otherwise do not reject,} \tag{5.7}$$

where the constant c_α is chosen to make the type I error probability equal to α. The constant c_α is the upper α percentile for the null ($\gamma^2 = 1$) distribution of C. Comment 4 explains how to obtain the critical value c_α for sample sizes m and n and available levels of α.

b. *One-Sided Lower-Tail Test*. To test

$$H_0 : \gamma^2 = 1$$

versus

$$H_2 : \gamma^2 < 1,$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } C \leq [c_{1-\alpha} - 1]; \quad \text{otherwise do not reject,} \tag{5.8}$$

where, as with the upper-tail test in (5.7), the appropriate value of $c_{1-\alpha}$ is obtained as stipulated in Comment 4.

c. *Two-Sided Test*. To test

$$H_0 : \gamma^2 = 1$$

versus

$$H_3 : \gamma^2 \neq 1,$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } C \geq c_{\alpha_1} \text{ or } C \leq [c_{1-\alpha_2} - 1]; \text{ otherwise do not reject,} \tag{5.9}$$

where $\alpha_1 + \alpha_2 = \alpha$ and the appropriate values of c_{α_1} and $c_{1-\alpha_2}$ are obtained as directed in Comment 4. We note that the null distribution of C is symmetric when $N = (m + n)$ is an even number (see Comment 5). In such a case, it is most natural to place an equal amount of probability in each tail of the null distribution of C, corresponding to setting $\alpha_1 = \alpha_2 = \alpha/2$. Thus, when N is even, the two-sided symmetric version of procedure (5.9) uses the critical values $c_{\alpha/2}$ and $[c_{1-(\alpha/2)} - 1]$.

Large-Sample Approximation

The large-sample approximation is based on the asymptotic normality of C, suitably standardized. For this purpose we first need to know the expected value and variance of C when the null hypothesis is true. Since the set of scores being assigned to the jointly ranked sample X and Y observations (see Procedure) depends on whether N is an even or odd integer, it is not surprising that the form of the mean and variance for C also depends on whether N is even or odd. When H_0 is true and $N = m + n$ is an even number, the expected value and variance of C are

$$E_0(C) = \frac{n(N+2)}{4} \tag{5.10}$$

and

$$\text{var}_0(C) = \frac{mn(N+2)(N-2)}{48(N-1)}, \tag{5.11}$$

respectively. When N is an odd integer, the null expected value and variance of C are

$$E_0(C) = \frac{n(N+1)^2}{4N} \tag{5.12}$$

and

$$\text{var}_0(C) = \frac{mn(N+1)(3+N^2)}{48N^2}, \tag{5.13}$$

respectively. These expressions for $E_0(C)$ and $\text{var}_0(C)$ are verified by direct calculations in Comment 8 for the special cases of $m = n = 2$ (where $N = 4$ is even) and $m = 3$, $n = 2$ (where $N = 5$ is odd). General derivations of the null expected value and variance expressions in (5.10), (5.11), (5.12), and (5.13) are presented in Comment 9.

For general N (even or odd), the standardized version of C is given by

$$C^* = \frac{C - E_0(C)}{\{\text{var}_0(C)\}^{1/2}}, \tag{5.14}$$

where $E_0(C)$ and $\text{var}_0(C)$ correspond to expressions (5.10) and (5.11), respectively, if N is even or to expressions (5.12) and (5.13), respectively, if N is odd. In either case, when H_0 is true, C^* has, as $\min(m, n)$ tends to infinity, an asymptotic $N(0,1)$ distribution (see Comment 9 for indications of the proof). The normal theory approximation for procedure (5.7) is

$$\text{Reject } H_0 \text{ if } C^* \geq z_\alpha; \quad \text{otherwise do not reject,} \tag{5.15}$$

the normal theory approximation for procedure (5.8) is

$$\text{Reject } H_0 \text{ if } C^* \leq -z_\alpha; \quad \text{otherwise do not reject,} \tag{5.16}$$

and the normal theory approximation for procedure (5.9) is

$$\text{Reject } H_0 \text{ if } |C^*| \geq z_{\alpha/2}; \quad \text{otherwise do not reject.} \tag{5.17}$$

Ties

If there are ties among the X and/or Y observations, assign each of the observations in a tied group the average of the integer scores that are associated with the tied group. After computing C with these average scores for tied observations, use the appropriate procedure (5.7), (5.8), or (5.9) with this tie-averaged value of C. Note, however, that this test associated with tied X and/or Y observations is only approximately, and not exactly, of significance level α. (To get an exact level α test even in this tied setting, see Comment 11.)

When applying the large-sample approximation, an additional factor must be taken into account. Although ties among the X and/or Y observations do not affect the null expected value of C, its null variance is reduced to

$$\text{var}_0(C) = \frac{mn\left[16\sum_{j=1}^{g} t_j r_j^2 - (N)(N+2)^2\right]}{16N(N-1)} \tag{5.18}$$

in the presence of ties when N is even and to

$$\text{var}_0(C) = \frac{mn\left[16N\sum_{j=1}^{g} t_j r_j^2 - (N+1)^4\right]}{16N^2(N-1)} \tag{5.19}$$

when N is odd, where in (5.18) and (5.19) g denotes the number of tied groups among the N sample observations, t_j is the size of tied group j, and r_j is the average score associated with the observations in tied group j. We note that an untied observation is considered to be a tied "group" of size 1. In particular, if there are no ties among the X's and/or Y's, then $g = N$ and $t_j = 1$ for $j = 1, \ldots, N$. In this case of no tied sample observations, we have

$$\sum_{j=1}^{g} t_j r_j^2 = 2\sum_{j=1}^{N/2} j^2 = \frac{2\left(\frac{N}{2}\right)\left(\frac{N}{2}+1\right)\left(2\left(\frac{N}{2}\right)+1\right)}{6} = \frac{N(N+1)(N+2)}{12},$$

when N is an even integer, and

$$\begin{aligned}\sum_{j=1}^{g} t_j r_j^2 &= 2\sum_{j=1}^{(N-1)/2} j^2 + \left(\frac{N+1}{2}\right)^2 \\ &= \frac{2}{6}\left(\frac{N-1}{2}\right)\left(\frac{N-1}{2}+1\right)\left(2\left(\frac{N-1}{2}\right)+1\right) + \left(\frac{N+1}{2}\right)^2 \\ &= \left(\frac{N+1}{12}\right)(N^2+2N+3),\end{aligned}$$

when N is an odd integer. Using these expressions for $\sum_{j=1}^{g} t_j r_j^2$, the associated ties-adjusted expressions for $\text{var}_0(C)$ given in (5.18) and (5.19) reduce to the corresponding untied null variances in (5.11) and (5.13), respectively, in the case of no tied observations.

As a consequence of the effect that ties have on the null variance of C, the following modification is needed to apply the large-sample approximation when there are ties among the X and/or Y observations. Compute C using average scores and set

$$C^* = \frac{C - E_0(C)}{\{\text{var}_0(C)\}^{1/2}}, \tag{5.20}$$

where $E_0(C)$ and $\text{var}_0(C)$ are now given by displays (5.10) and (5.18), respectively, if N is even or by displays (5.12) and (5.19), respectively, if N is odd. With this modified form of C^*, approximations (5.15), (5.16), or (5.17) can be applied.

EXAMPLE 5.1 ***Serum Iron Determination.***

The data in Table 5.1 are a portion of the data obtained by Jung and Parekh (1970) in a study concerned with techniques for direct determination of serum iron. In particular, they attempted to eliminate some of the problems associated with other commonly used methods, which often result in turbidity of the analyzed serum, as well as requiring large samples and slow, tedious analyses. To accomplish this, the authors proposed an improved method for serum iron determination based on a different detergent. One of the purposes of their investigation was to study the accuracy of their method for serum iron determination in comparison to a method due to Ramsay (1957). Twenty duplicate analyses were made, each by the proposed method and by the method of Ramsay, using Hyland control sera containing 105 μg of serum iron per 100 ml. Table 5.1 gives the serum iron detected (in μg/100 ml) for the 40 analyses in the study.

From the point of view of procedural technique, the Jung–Parekh method competes favorably with the Ramsay method for serum iron determination. An additional concern,

Table 5.1 Serum Iron (μg/100 ml) Determination Using Hyland Control Sera

Ramsay method	Jung–Parekh method
111	107
107	108
100	106
99	98
102	105
106	103
109	110
108	105
104	104
99	100
101	96
96	108
97	103
102	104
107	114
113	114
116	113
113	108
110	106
98	99

Source: D.H. Jung and A.C. Parekh (1970).

however, is whether there is a loss of accuracy when the Jung–Parekh procedure is used instead of the Ramsay procedure. As a result, the alternative of interest in this example is greater dispersion or variation for the Jung–Parekh method of serum iron determination than for the method of Ramsay. Hence, letting Y correspond to the Ramsay determinations and X to the Jung–Parekh determinations, we are interested in a one-sided test designed to detect the alternative $H_1 : \gamma^2 > 1$. Since there are ties among the X and Y sample observations and $N = m + n = 20 + 20 = 40$ is an even integer, we will apply the large-sample approximation (with ties), as detailed in (5.15) and (5.20), to procedure (5.7).

For the purpose of illustration, we consider the approximate level $\alpha = .05$. Hence, using the R command `pnorm(·)`, we set `1-pnorm(`$z_{.05}$`)` $= .05$ and obtain $z_{.05} = 1.645$, and the large-sample approximation to procedure (5.7) is given by

$$\text{Reject } H_0 \text{ if} \frac{C - E_0(C)}{\{\text{var}_0(C)\}^{1/2}} \geq 1.645,$$

where $E_0(C)$ and $\text{var}_0(C)$ are given by expressions (5.10) and (5.18), respectively.

To calculate C (5.6) we need the Ansari–Bradley ranks of the 20 Y (Ramsay) observations. In the following display we list in order (from least to greatest) the combined sample of 40 X (Jung–Parekh) and Y (Ramsay) values and assign the ranks according to the Ansari–Bradley scheme.

Ansari–Bradley Ranking Scheme for the Data of Table 5.1

Y	X	Y	Y	X	Y	Y	X	X	Y
96	96	97	98	98	99	99	99	100	100
1.5	1.5	3	4.5	4.5	7	7	7	9.5	9.5
Y	Y	Y	X	X	X	Y	X	X	X
101	102	102	103	103	104	104	104	105	105
11	12.5	12.5	14.5	14.5	17	17	17	19.5	19.5
X	X	Y	Y	X	Y	Y	X	X	X
106	106	106	107	107	107	108	108	108	108
19	19	19	16	16	16	12.5	12.5	12.5	12.5
Y	X	Y	Y	X	Y	Y	X	X	Y
109	110	110	111	113	113	113	114	114	116
10	8.5	8.5	7	5	5	5	2.5	2.5	1

Thus $C = \sum_{i=1}^{20} R_i = 185.5$. In order to calculate C^*, we need to evaluate expressions (5.10) and (5.18). We illustrate the calculation of $\sum_{j=1}^{g} t_j r_j^2$ in the following table, where for our data there are $g = 19$ tied groups.

Thus, we have $\sum_{j=1}^{19} t_j r_j^2 = 5721$, and from (5.10) and (5.18), we obtain

$$C^* = \frac{185.5 - [20(42)/4]}{\{(20)(20)[16(5721) - 40(42)^2]/[16(40)(39)]\}^{1/2}} = -1.34,$$

which tells us not to reject H_0 at the approximate $\alpha = .05$ level, since $C^* = -1.34 < 1.645 = z_{.05}$. Hence, there is not sufficient evidence to indicate loss of accuracy when the Jung–Parekh method is used instead of the Ramsay method.

Tied group	t_j	r_j^2	$t_j r_j^2$
1	2	2.25	4.5
2	1	9	9
3	2	20.25	40.5
4	3	49	147
5	2	90.25	180.5
6	1	121	121
7	2	156.25	312.5
8	2	210.25	420.5
9	3	289	867
10	2	380.25	760.5
11	3	361	1083
12	3	256	768
13	4	156.25	625
14	1	100	100
15	2	72.25	144.5
16	1	49	49
17	3	25	75
18	2	6.25	12.5
19	1	1	1

Since the one-sided P-value for these data is the lowest significance level at which we can reject H_0 in favor of $\gamma^2 > 1$ with the observed value of the test statistic C^*, we see, using the R command `pnorm(·)`, that the P-value for these data is approximately $P_0(C^* \geq -1.34) \approx 1 - \texttt{pnorm(-1.34)} = (1 - .0901) = .9099$. Thus, there is absolutely no evidence in the sample data to indicate any loss of accuracy with the Jung–Parekh method. In fact, the C^* value of -1.34 actually provides evidence pointing in the other direction, namely, $\gamma^2 < 1$, corresponding to improved accuracy with the Jung–Parekh method.

Comments

1. *Known Population Medians*. When the population median, θ_2, for the Y observations is known to be equal to $\theta_1 + \xi$, where θ_1 is the population median for the X observations and ξ is a known constant, we can create modified observations $X_i' = X_i + \xi, i = 1, \ldots, m$ and apply the Ansari–Bradley procedures of this section to the modified X' observations and the unchanged Y observations.
2. *Testing γ^2 Equal to Some Specified Value Other Than One*. To test the hypothesis $\gamma^2 = \gamma_0^2$, where γ_0^2 is some specified positive number different from 1, when the common median for the underlying X and Y populations has known value θ_0, we obtain the modified observations $X_i' = (X_i - \theta_0)/\gamma_0$, for $i = 1, \ldots, m$, and $Y_j' = (Y_j - \theta_0)$, for $j = 1, \ldots, n$, and compute C (5.6) using the X''s and Y''s (instead of the X's and Y's). Procedures (5.7), (5.8), or (5.9) or the corresponding large-sample approximations (5.15), (5.16), or (5.17) may then be applied as described.

3. *Motivation for the Test*. Under Assumptions A1– A3, the X and Y populations have the same median. Suppose, for example, that γ^2 is greater than 1. Then the X values would tend to be more spread out than the Y values. Thus, the Y's would tend to get larger scores than the X's from the scheme described in the Procedure and C (5.6) would tend to be larger. (Visualize an extreme sample where the sample values, when ordered, fall in the pattern *XXYYYXX*.) This serves as partial motivation for the one-sided upper-tail test procedure given in (5.7).
4. *Derivation of the Distribution of C under H_0 (No-Ties Case)*. Under H_0 (5.1), each of the $\binom{N}{n}$ possible "meshings" of the X's and Y's has probability $1/\binom{N}{n}$. This fact can be used to obtain the null distribution of C (5.6). We illustrate the steps involved in constructing this null distribution for the two cases $m = 3$, $n = 2$ (where $N = 5$ is odd) and $m = 2$, $n = 2$ (where $N = 4$ is even). First, for $m = 3$ and $n = 2$, we use the set of scores $\{1, 1, 2, 2, 3\}$. Let $R^{(1)} < R^{(2)}$ denote the ordered Y scores so that $C = R_1 + R_2 = R^{(1)} + R^{(2)}$. The $\binom{5}{2} = 10$ possible meshings and associated values of $(R^{(1)}, R^{(2)})$ and C are given in the following table.

Meshing	Probability	$(R^{(1)}, R^{(2)})$	$C = R^{(1)} + R^{(2)}$
YYXXX	$\frac{1}{10}$	(1, 2)	3
YXYXX	$\frac{1}{10}$	(1, 3)	4
YXXYX	$\frac{1}{10}$	(1, 2)	3
YXXXY	$\frac{1}{10}$	(1, 1)	2
XYYXX	$\frac{1}{10}$	(2, 3)	5
XYXYX	$\frac{1}{10}$	(2, 2)	4
XYXXY	$\frac{1}{10}$	(1, 2)	3
XXYYX	$\frac{1}{10}$	(2, 3)	5
XXYXY	$\frac{1}{10}$	(1, 3)	4
XXXYY	$\frac{1}{10}$	(1, 2)	3

Thus, for example, the probability is $\frac{3}{10}$ under H_0 that C is equal to 4, because $C = 4$ when either of the exclusive outcomes $(R^{(1)}, R^{(2)}) = (1, 3)$ or $(R^{(1)}, R^{(2)}) = (2, 2)$ occurs. These two outcomes for $(R^{(1)}, R^{(2)})$ are associated with three mutually exclusive meshings, each with null probability $\frac{1}{10}$. Hence, it follows that $P_0(C = 4) = 3(\frac{1}{10})$. Proceeding in the same manner for all possible values for C and simplifying, we obtain the null distribution.

Possible value of C	Probability under H_0
2	$\frac{1}{10}$
3	$\frac{4}{10}$
4	$\frac{3}{10}$
5	$\frac{2}{10}$

The probability, under H_0, that C is greater than or equal to 4, for example, is therefore

$$P_0(C \geq 4) = P_0(C = 4) + P_0(C = 5) = .3 + .2 = .5,$$

so that $c_{.5} = 4$. Note also that $c_{.2} = 5$.

For the case of $m = n = 2$ (where $N = 4$ is even), we use the set of scores $\{1, 1, 2, 2\}$. The $\binom{4}{2} = 6$ possible meshings, as well as the associated ordered Y-scores $(R^{(1)}, R^{(2)})$ and values of C are given in the following table.

Meshing	Probability	$(R^{(1)}, R^{(2)})$	$C = R^{(1)} + R^{(2)}$
XXYY	$\frac{1}{6}$	(1, 2)	3
XYXY	$\frac{1}{6}$	(1, 2)	3
YXXY	$\frac{1}{6}$	(1, 1)	2
XYYX	$\frac{1}{6}$	(2, 2)	4
YXYX	$\frac{1}{6}$	(1, 2)	3
YYXX	$\frac{1}{6}$	(1, 2)	3

Proceeding as for the previous case of $m = 3$, $n = 2$, we obtain the null distribution for C.

Possible value of C	Probability under H_0
2	$\frac{1}{6}$
3	$\frac{4}{6}$
4	$\frac{1}{6}$

Note that we have derived the null distribution of C without specifying the form of the common (under H_0) underlying X and Y populations beyond the point of requiring that they be continuous. This is why the test procedures based on C are called *distribution-free procedures*. From the null distribution of C, we can determine the critical value c_α and control the probability α of falsely rejecting H_0 when H_0 is true, and this error probability does not depend on the specific form of the common underlying continuous distribution for the X and Y observations.

For given sample sizes m and n, the R command `cAnsBrad`(α, m, n) can be used to find the available upper-tail critical values c_α for possible values of C. For a given available significance level α, the critical value c_α then corresponds to $P_0(C \geq c_\alpha) = \alpha$ and is given by `cAnsBrad`$(\alpha, m, n) = c_\alpha$. Thus, for example, for $m = 8$ and $n = 4$, we have $P_0(C \geq 20) = .0283$ so that $c_{.0283} = 20$ for $m = 8$ and $n = 4$.

5. *Symmetry of the Distribution of C under the Null Hypothesis When $N = m + n$ Is Even*. When H_0 is true and $N = m + n$ is an even integer, the distribution of

C is symmetric about its mean $n(N+2)/4$. (See Comment 4 for verification of this when $m = n = 2$.) This implies that when N is even

$$P_0(C \leq x) = P_0\left(C \geq \frac{n(N+2)}{2} - x\right), \tag{5.21}$$

for every possible value of x.

Equation (5.21) is directly used to convert upper-tail probabilities, as obtained from `cAnsBrad(`$\cdot$, *m, n*), to lower-tail probabilities when N is even. Thus, the lower-tail critical value $[c_{1-\alpha} - 1]$ used in test procedures (5.8) or (5.9) can be expressed in terms of the upper-tail critical value c_α by

$$[c_{1-\alpha} - 1] = \left[\frac{n(N+2)}{2} - c_\alpha\right], \tag{5.22}$$

when N is even.

6. *Equivalent Form*. The statistic C (5.6) is the sum of the scores assigned to the Y observations by the Ansari–Bradley scoring scheme described in the Procedure. Test procedures (5.7), (5.8), and (5.9) could equivalently be based on the statistic $C' = [\text{sum of the scores assigned by this scheme to the } X \text{ observations}]$, because $C' = [N(N+2)/4] - C$ when $N = m + n$ is even and $C' = [(N-1)^2/4] - C$ when N is odd (see Problem 6).

7. *Assumptions*. We can use the Ansari–Bradley test procedures in (5.7), (5.8), or (5.9) without even requiring that the variances for the X and Y populations exist. Indeed, our Assumptions A1–A3 for this section do not specify anything about the existence of even the first moments of the X and Y populations. However, when the first two moments (and, therefore, the variance) for the underlying distributional model $H(u)$ in (5.2) exist, we see from the equal-in-distribution statement in (5.3) that

$$\text{var}\left(\frac{X}{\eta_1}\right) = \text{var}\left(\frac{Y}{\eta_2}\right),$$

which, in turn, implies that

$$[\text{var}(X)]/\eta_1^2 = [\text{var}(Y)]/\eta_2^2.$$

Thus, when the variances exist, we see that $\gamma^2 = [\eta_1^2/\eta_2^2] = [\text{var}(X)/\text{var}(Y)]$.

Assumptions A1–A3 do imply that the only possible difference between the X and Y populations is the difference in scale parameters. In particular, these assumptions imply that the two populations do not differ in location, as they have a common median θ (see Comment 1 for a slight relaxation of this condition). While the requirement of equal medians is not necessary for the classical $\mathcal{F}$-test based on the ratio of the X and Y sample variances, this requirement is essential for the Ansari–Bradley test. For example, suppose that $m = 5$, $n = 4$, and the X and Y probability distributions are such that $P(X < Y) = 0$. Then, for *all* possible X and Y samples, the joint ordering of the five X observations and four Y observations would *always* result in a value of $C = 10$, regardless of the scale parameters for the two populations. That is, in such a setting, *no* information

about γ^2 can be obtained from the joint ranking and the Ansari–Bradley scoring scheme.

Moses (1963) has emphasized this bizarre behavior of tests for dispersion based on joint rankings of the sample X and Y observations and has shown that such tests are inadequate unless strong assumptions (such as equal or known medians) are made concerning the locations of the X and Y populations. For an asymptotically distribution-free test that does not require equal or known medians, see Section 5.2.

8. *Calculation of the Mean and Variance of C under the Null Hypothesis*, H_0. In (5.10) and (5.11), we presented formulas for the mean and variance of C when the null hypothesis is true and $N = (m + n)$ is an even number. The corresponding expressions for the null mean and variance of C when N is an odd number are given in (5.12) and (5.13). In this comment, we illustrate a direct calculation of $E_0(C)$ and $\text{var}_0(C)$ in the particular cases of $m = 3$, $n = 2$ (where $N = 5$ is odd) and $m = n = 2$ (where $N = 4$ is even), using the null distributions of C obtained in Comment 4. (Later, in Comment 9, we present general derivations of $E_0(C)$ and $\text{var}_0(C)$.) The null mean, $E_0(C)$, is obtained by multiplying each possible value of C by its probability under H_0. Thus, for $m = 3$, $n = 2$, we have

$$E_0(C) = 2(.1) + 3(.4) + 4(.3) + 5(.2) = 3.6.$$

This is in agreement with what we obtain using (5.12), namely.

$$E_0(C) = \frac{n(N+1)^2}{4N} = \frac{2(5+1)^2}{4(5)} = 3.6.$$

Similarly, for $m = n = 2$, we have by direct computation from the null distribution of C in Comment 4 that

$$E_0(C) = 2\left(\frac{1}{6}\right) + 3\left(\frac{4}{6}\right) + 4\left(\frac{1}{6}\right) = 3,$$

in agreement with the value obtained from (5.10), namely,

$$E_0(C) = \frac{n(N+2)}{4} = \frac{2(4+2)}{4} = 3.$$

Checks on the expressions for $\text{var}_0(C)$ are also easily performed, using the well-known fact that

$$\text{var}_0(C) = E_0(C^2) - \{E_0(C)\}^2.$$

The required values of $E_0(C^2)$, the second moment of the null distribution of C, are again obtained by multiplying the possible values of C^2 by the corresponding probabilities under H_0. For the case of $m = 3$, $n = 2$, we find that

$$E_0(C^2) = 2^2(.1) + 3^2(.4) + 4^2(.3) + 5^2(.2) = 13.8,$$

yielding

$$\text{var}_0(C) = 13.8 - (3.6)^2 = 13.8 - 12.96 = .84,$$

which is in agreement with the value obtained from (5.13), namely,

$$\begin{aligned}\text{var}_0(C) &= \frac{mn(N+1)(3+N^2)}{48N^2} \\ &= \frac{3(2)(5+1)(3+5^2)}{48(5)^2} = .84.\end{aligned}$$

Similarly, for $m = n = 2$, we have by direct computation from the null distribution in Comment 4 that

$$E_0(C^2) = 2^2\left(\frac{1}{6}\right) + 3^2\left(\frac{4}{6}\right) + 4^2\left(\frac{1}{6}\right) = \frac{56}{6},$$

yielding

$$\text{var}_0(C) = \tfrac{56}{6} - (3)^2 = \tfrac{1}{3},$$

which is in agreement with the value obtained from (5.11), namely,

$$\begin{aligned}\text{var}_0(C) &= \frac{mn(N+2)(N-2)}{48(N-1)} \\ &= \frac{2(2)(4+2)(4-2)}{48(4-1)} = \frac{1}{3}.\end{aligned}$$

9. *Large-Sample Approximation*. The statistic C/n is the average of the scores assigned to the Y observations. Since all $\binom{N}{n}$ possible distributions of the appropriate scores (depending on whether N is even or odd) to the X and Y observations are equally likely under H_0, the null distribution of C/n is the same as the distribution of the sample mean for a random sample of size n drawn without replacement from the finite population of scores S_N, where $S_N = \{1, 2, 3, \ldots, N/2, N/2, \ldots, 3, 2, 1\}$ if N is an even number and $S_N = \{1, 2, 3, \ldots, (N-1)/2, (N+1)/2, (N-1)/2, \ldots, 3, 2, 1\}$ if N is odd.

   From basic results for a random sample of size n drawn without replacement from a finite population of N elements, we know that

   (i) the expected value of the sample average is equal to the average, μ_{pop}, of the finite population,

   (ii) the variance of the sample average is equal to

$$\frac{\sigma^2_{\text{pop}}}{n}\left(\frac{N-n}{N-1}\right),$$

   where σ^2_{pop} is the variance of the finite population and the factor $(N-n)/(N-1)$ is known as the finite population correction factor.

   For the case of N even and the finite population $S_N = \{1, 2, 3, \ldots, N/2, N/2, \ldots, 3, 2, 1\}$, we see that

   (iii)
$$\mu_{\text{pop}} = \frac{2}{N}\sum_{i=1}^{N/2} i = \frac{(N/2)[(N/2)+1]}{2(N/2)} = \frac{N+2}{4}$$

   and

(iv)
$$\begin{aligned}\sigma^2_{\text{pop}} &= \left[\frac{2}{N}\left(\sum_{i=1}^{N/2} i^2\right) - \left(\frac{N+2}{4}\right)^2\right] \\ &= \left[\frac{(N/2)[(N/2)+1][2(N/2)+1]}{6(N/2)} - \left(\frac{N+2}{4}\right)^2\right] \\ &= \left[\frac{(N+2)(N+1)}{12} - \frac{(N+2)(N+2)}{16}\right] \\ &= \frac{(N+2)(N-2)}{48}.\end{aligned}$$

From (i), (ii), (iii), and (iv), it follows that

$$E_0\left(\frac{C}{n}\right) = \frac{N+2}{4}$$

and

$$\text{var}_0\left(\frac{C}{n}\right) = \left[\frac{(N+2)(N-2)}{48n}\right]\left[\frac{N-n}{N-1}\right] = \frac{m(N+2)(N-2)}{48n(N-1)}.$$

Thus,

$$E_0(C) = nE_0\left(\frac{C}{n}\right) = \frac{n(N+2)}{4}$$

and

$$\text{var}_0(C) = n^2\text{var}_0\left(\frac{C}{n}\right) = \frac{mn(N+2)(N-2)}{48(N-1)},$$

in agreement with the formulas in (5.10) and (5.11). The corresponding expressions for $E_0(C)$ and $\text{var}_0(C)$ when N is an odd integer, as given in (5.12) and (5.13), respectively, can be similarly obtained using the expressions in (i) and (ii) and the finite population

$$S_N = \left\{1, 2, 3, \ldots, \frac{N-1}{2}, \frac{N+1}{2}, \frac{N-1}{2}, \ldots, 3, 2, 1\right\}.$$

For any N (even or odd), the asymptotic normality under H_0 of the standardized

$$C^* = \frac{C - E_0(C)}{\sqrt{\text{var}_0(C)}}$$

follows from standard theory for the mean of a sample from a finite population (cf. Wilks, 1962, p. 268). Asymptotic normality results for C^* are also available under general alternatives to H_0 (see, for example, Ansari and Bradley (1960), Randles and Wolfe (1979), or Hájek and Šidák (1967)).

10. *Lower-Tail Critical Values*. In the expression for the one-sided lower-tail test in (5.9), the critical value is given to be $c_{1-\alpha} - 1$, where $c_{1-\alpha}$ is the upper $(1-\alpha)$th

percentile of the null distribution of C. This means that

$$P_0(C \leq c_{1-\alpha} - 1) = 1 - P_0(C > c_{1-\alpha} - 1) = 1 - P_0(C \geq c_{1-\alpha}),$$

where the last equality follows from the fact that C is a discrete random variable assuming only positive integer values. Since $c_{1-\alpha}$ is the upper $(1-\alpha)$th percentile for the null distribution of C, it follows that

$$P_0(C \leq c_{1-\alpha} - 1) = 1 - (1 - \alpha) = \alpha.$$

Hence, $c_{1-\alpha} - 1$ is, indeed, the *lower* αth percentile for the null distribution of C, as required for the level α one-sided lower-tail test procedure in expression (5.8).

When N is an even integer, we have already noted in Comment 5 that the null distribution of C is symmetric about its mean, $n(N+2)/4$. It follows that $[c_{1-\alpha} - 1] = [\{n(N+2)/2\} - c_\alpha]$ when N is even.

11. *Exact Conditional Distribution of C with Ties*. To have a test with exact significance level even in the presence of ties among the X's and/or Y's, we need to consider all $\binom{N}{n}$ possible assignments of the N observations with n observations serving as Y's and m observations serving as X's. As in Comment 4, it still follows that, under H_0 (5.1), each of the $\binom{N}{n}$ possible "meshings" of the X's and Y's has probability $1/\binom{N}{n}$. The only difference in the case of ties is that we now use average scores in the computation of C for each of these $\binom{N}{n}$ "meshings" leading to the tabulation of the null distribution. We illustrate this construction for N odd (a similar approach will work for N even) and the following $m = 3$, $n = 2$ data: $X_1 = 3.2$, $X_2 = 5.7$, $X_3 = 6.3$, $Y_1 = 1.9$, $Y_2 = 6.3$. The associated average scores assignments (taking into account the tie between X_3 and Y_2) are 2, 3, 1.5, 1, and 1.5, respectively, and the corresponding value of C, the sum of the scores for the Y observations, is $C = 1.5 + 1 = 2.5$. To assess the significance of this value of C, we obtain its conditional distribution by considering the $\binom{5}{2} = 10$ possible assignments of the observations 1.9, 3.2, 5.7, 6.3, and 6.3 to serve as three X observations and two Y observations, or, equivalently the 10 possible assignments of the average scores 1, 1.5, 1.5, 2, and 3 to serve as three X scores and two Y scores. These 10 assignments and the corresponding values of C are as follows.

Y scores	Probability under H_0	Value of C
1, 1.5	$\frac{1}{10}$	2.5
1, 1.5	$\frac{1}{10}$	2.5
1, 2	$\frac{1}{10}$	3
1, 3	$\frac{1}{10}$	4
1.5, 1.5	$\frac{1}{10}$	3
1.5, 2	$\frac{1}{10}$	3.5
1.5, 3	$\frac{1}{10}$	4.5
1.5, 2	$\frac{1}{10}$	3.5
1.5, 3	$\frac{1}{10}$	4.5
2, 3	$\frac{1}{10}$	5

This yields the null tail probabilities

$$P_0(C \geq 5) = \tfrac{1}{10},$$

$$P_0(C \geq 4.5) = \tfrac{3}{10},$$

$$P_0(C \geq 4) = \tfrac{4}{10},$$

$$P_0(C \geq 3.5) = \tfrac{6}{10},$$

$$P_0(C \geq 3) = \tfrac{8}{10},$$

$$P_0(C \geq 2.5) = 1.$$

This distribution is called the *conditional null distribution* or the *permutation null distribution* of C, given the set of tied scores $\{1, 1.5, 1.5, 2, 3\}$. For the particular observed value $C = 2.5$, we have $P_0(C \geq 2.5) = 1$, so that such a value does not indicate a deviation from H_0 in the direction of $\gamma^2 > 1$ (although it would provide marginal support for the alternative $\gamma^2 < 1$).

12. *Confidence Intervals, Confidence Bounds, and Point Estimators for* γ^2. The Ansari–Bradley statistic, C (5.6), is a member of a large class of rank statistics (referred to as *linear rank statistics* in the literature; see, for example, Section 9.3 of Randles and Wolfe (1979)) that can be used to test for equality of scale parameters under the strict assumption of equal or known medians for the X and Y populations. Bauer (1972) has shown how to invert some of these linear rank tests of $\gamma^2 = 1$, including the Ansari–Bradley procedure, to obtain point estimators and confidence intervals or bounds for γ^2 in such a setting.
13. *Unequal and Unknown Medians*. If the medians of the X and Y populations are not known and it is questionable whether or not they are equal, Ansari and Bradley (1960) suggested the following modification to their test procedures. Define the adjusted observations $X_i' = X_i - \tilde{X}, i = 1, \ldots, m$, and $Y_j' = Y_j - \tilde{Y}, j = 1, \ldots, n$, where $\tilde{X}$ and $\tilde{Y}$ are the sample medians for the X and Y observations, respectively. Let C' be C (5.6) calculated for these adjusted X' and Y' observations. Depending on the alternative to $H_0 : \gamma^2 = 1$ that is of interest, the appropriate procedure (5.7), (5.8), or (5.9), or the corresponding large-sample approximation, can then be applied directly to the modified statistic C' instead of C. Such tests based on C' are no longer strictly distribution-free. However, Gross (1966) has given sufficient conditions under which such procedures are asymptotically distribution-free. Under such conditions, the various tests based on C' maintain an approximate (both m and n large) significance level α over a large class of continuous underlying distributions.
14. *Consistency of the C-Tests*. Under Assumptions A1–A3, the consistency of the tests based on C depends on the parameter

$$\Delta^* = \left[P(X > Y > \theta) + P(X < Y < \theta) - \tfrac{1}{4}\right].$$

The test procedures defined by (5.7), (5.8), and (5.9) are consistent against the alternatives corresponding to $\Delta^* >, <$, and $\neq 0$, respectively.

15. *More General Alternatives*. In many two-sample situations, we are interested in simultaneously detecting either location or scale differences between the X and Y populations. One solution to this broader problem is to use a test procedure designed to detect quite general alternatives. One such test procedure based on the two-sample Kolmogorov (1933)–Smirnov (1939) statistic is discussed in Section 5.4. A second approach is to conduct simultaneously a test such as the Wilcoxon rank sum procedure based on W(4.3) for detecting differences in location and a second test such as the Ansari–Bradley procedure based on C (5.6) for detecting differences in scale. One such simultaneous testing approach, due to Lepage (1971, 1973), for dealing with general alternatives is the topic of Section 5.3. Randles and Hogg (1971) have shown that in such a situation, W and C are uncorrelated and, in fact, asymptotically independent when H_0 (5.1) is true. This implies, among other things, that if we conduct the Wilcoxon rank sum test at a significance level α_1, and the Ansari–Bradley test at a significance level α_2, then the probability of incorrectly rejecting with at least one of the two tests, given that H_0 (5.1) is true, is approximately $\alpha_1 + \alpha_2 - \alpha_1\alpha_2$.

Properties

1. *Consistency*. For our statement, we consider the more stringent location-scale parameter model described in (5.4). Then the tests defined by (5.7), (5.8), and (5.9) are consistent against the alternative $\gamma^2 >, <$, and $\neq 1$, respectively. See also Comment 14.
2. *Asymptotic Normality*. See Randles and Wolfe (1979, pp. 315–320).
3. *Efficiency*. See Section 5.5.

Problems

1. Consider the chorioamnion permeability data in Table 4.1. In Section 4.1 we saw that a test procedure based on the Wilcoxon rank sum statistic did not reject the null hypothesis that the human chorioamnion is as permeable to water transfer at 12–26 weeks gestational age as it is at term. With this in mind and using the same data, test the hypothesis of equal dispersions versus the alternative that the variation in tritiated water diffusion across human chorioamnion is different at term than at 12–26 weeks gestational age.

2. Find or construct an example in which there exists a level α and a constant d such that when C (5.6) is computed for the original data $X_1, \ldots, X_m$, $Y_1, \ldots, Y_n$, the level α procedure in (5.7) does not lead to rejection of $H_0 : \gamma^2 = 1$, but when C is computed for the values $X_1, \ldots, X_m, Y_1 + d, \ldots, Y_n + d$, the level α procedure in (5.7) does lead to rejection of H_0. Note that such an example exposes an undesirable aspect of the test procedures based on C. Let Y be a random member from a population II. Then, the population II* formed by adding the constant d to each member of II must, by any reasonable definition of dispersion, have the same dispersion as the II population. Thus, the difference in dispersions between the X and Y populations must be the same as the difference in dispersions between the X and $Y + d$ populations. Yet the tests based on C, applied to the data $X_1, \ldots, X_m, Y_1 + d, \ldots, Y_n + d$, can yield a decision that differs from the one that results from applying the same C-test to $X_1, \ldots, X_m, Y_1, \ldots, Y_n$. (For a related discussion, see Comment 7.)

3. Consider the television-viewing behavior data in Table 4.4. For these data, use the R command `pAnsBrad(x,y)` to find the P-value for an appropriate test of the hypothesis of equal dispersions versus the alternative that there is more variability in the time spent in the room after

witnessing the violent behavior for those children who had previously watched the *Karate Kid* than for those children who had previously watched parts of the 1984 Summer Olympic Games. Comment on the importance of the results of Problem 4.5 in relationship to this dispersion test.

4. Verify the expressions for $E_0(C)$ and $\text{var}_0(C)$ in (5.12) and (5.13), respectively, when $N = (m + n)$ is an odd integer. (See Comment 9 for guidance.)

5. Consider the following two-sample data for $m = 3$, $n = 3 : X_1 = -3.7$, $X_2 = 4.6$, $X_3 = 1.5$, $Y_1 = 1.5$, $Y_2 = 4.6$, $Y_3 = 1.5$. Here, $N = 3 + 3 = 6$ is an even integer. Using the approach discussed in Comment 11, find the exact conditional null distribution of the Ansari–Bradley statistic, C (5.6). Compare and contrast this conditional null distribution with the null distribution of C for $m = n = 3$ and no tied observations.

6. Let C' be the sum of the scores assigned to the X observations by the Ansari–Bradley scoring scheme described in the Procedure. Verify directly, or illustrate using the serum iron determination data in Table 5.1, that $C' = [N(N + 2)/4] - C$, when $N = (m + n)$ is even.

7. For an arbitrary total number of observations $N = (m + n)$, find an expression for the smallest and largest possible values of C. Consider the two cases of N even and N odd.

8. Let X and Y be independent, identically distributed continuous random variables with a common probability distribution with median θ. What is the value of Δ^* in Comment 14 for this setting?

9. Consider the alcoholic intake data in Table 4.2. In Example 4.2 the Wilcoxon rank sum test procedure led to the rejection of $H_0 : \Delta = 0$ in favor of $H_1 : \Delta < 0$. What does this result imply about the appropriateness of the Ansari–Bradley procedure in (5.9) or its approximate large-sample counterpart in (5.17) as a test of $H_0 : \gamma^2 = 1$ versus $H_1 : \gamma^2 \neq 1$ for these data? In view of this fact, find the approximate P-value for an appropriate modification of the large-sample procedure in (5.17) to test for possible differences in dispersions between the control and SST data. (See Comment 13.)

10. Consider the television-viewing behavior data in Table 4.4. *Without* the assumption of equal medians, find the approximate P-value for an appropriate test (see Comment 13) of the hypothesis of equal dispersions versus the alternative that there is more variability in the time spent in the room after witnessing the violent behavior for those children who had previously watched the *Karate Kid* than for those children who had previously watched parts of the 1984 Summer Olympic Games. Compare the P-value obtained here with the one found in Problem 3. Interpret the similarity or lack thereof between the two P-values.

11. Suppose $m = n = 10$. Compare the critical region for the exact level $\alpha = .056$ test of $H_0 : \gamma^2 = 1$ versus $H_1 : \gamma^2 > 1$ based on C with the critical region for the corresponding nominal level $\alpha = .056$ test based on the large-sample approximation. What is the exact significance level of this .056 nominal level test based on the large-sample approximation?

12. Generate the conditional permutation distribution of C (see Comments 4 and 11), given the set of tied values for the serum iron data in Example 5.1. From this conditional permutation distribution of C, obtain the exact conditional P-value, $P_0(C \geq 185.5)$, for the corresponding test of $H_0 : \gamma^2 = 1$ versus $H_1 : \gamma^2 > 1$. Compare this exact conditional P-value with the approximate P-value for the large-sample test procedure applied to these data in Example 5.1.

5.2 AN ASYMPTOTICALLY DISTRIBUTION-FREE TEST FOR DISPERSION BASED ON THE JACKKNIFE–MEDIANS NOT NECESSARILY EQUAL (MILLER)

Hypothesis

Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent random samples satisfying Assumptions A1 and A2 from continuous populations with distribution functions F and G, respectively,

satisfying the location-scale parameter model relationship in (5.2) and (5.3). In addition, we assume that the continuous distribution associated with the distribution function $H(\cdot)$ in (5.2) has finite fourth moment; that is, we assume

A4. If V is a continuous random variable with distribution function H, then $E(V^4) < \infty$.

Under Assumptions A1, A2, and A4, but without the equal median Assumption A3, we are once again interested in the ratio of scale parameters $\gamma = (\eta_1/\eta_2)$. In view of Assumption A4 (see Comment 7), we note that $\gamma^2 = [\text{var}(X)/\text{var}(Y)]$, the ratio of population variances. We are interested in testing the null hypothesis H_0 (5.1), which reduces to $H_0 : \gamma^2 = 1$, corresponding to the assertion that the population variances are equal, under the location-scale parameter model (5.2).

Procedure

Consider the X sample data with the first observation deleted and set

$$\bar{X}_1 = \sum_{s=2}^{m} \frac{X_s}{m-1} \quad \text{and} \quad D_1^2 = \sum_{s=2}^{m} \frac{(X_s - \bar{X}_1)^2}{m-2}. \tag{5.23}$$

Thus, $\bar{X}_1$ and D_1^2 are the sample average and sample variance for the data $X_2, \ldots, X_m$, corresponding to the X sample less X_1. Similarly, let

$$\bar{X}_i = \sum_{s \neq i}^{m} \frac{X_s}{m-1} \quad \text{and} \quad D_i^2 = \sum_{s \neq i}^{m} \frac{(X_s - \bar{X}_i)^2}{m-2} \tag{5.24}$$

be the sample average and sample variance, respectively, for the data $X_1, \ldots, X_{i-1}$, $X_{i+1}, \ldots, X_m$, corresponding to the X sample less X_i, for $i = 1, \ldots, m$. In the same fashion, let

$$\bar{Y}_j = \sum_{t \neq j}^{n} \frac{Y_t}{n-1} \quad \text{and} \quad E_j^2 = \sum_{t \neq j}^{n} \frac{(Y_t - \bar{Y}_j)^2}{n-2} \tag{5.25}$$

be the sample average and sample variance, respectively, for the data $Y_1, \ldots, Y_{j-1}$, $Y_{j+1}, \ldots, Y_n$, corresponding to the Y sample less Y_j, for $j = 1, \ldots, n$. Define $S_1, \ldots, S_m$ and $T_1, \ldots, T_n$ by

$$S_i = \ln D_i^2, \quad i = 1, \ldots, m, \tag{5.26}$$

and

$$T_j = \ln E_j^2, \quad j = 1, \ldots, n. \tag{5.27}$$

In addition, let

$$S_0 = \ln \left[\sum_{s=1}^{m} \frac{(X_s - \bar{X}_0)^2}{m-1} \right] \tag{5.28}$$

and

$$T_0 = \ln \left[\sum_{t=1}^{n} \frac{(Y_t - \bar{Y}_0)^2}{n-1} \right], \tag{5.29}$$

where $\bar{X}_0 = \sum_{s=1}^{m} X_s/m$ and $\bar{Y}_0 = \sum_{t=1}^{n} Y_t/n$, be the corresponding statistics for the complete samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$, respectively. Compute

$$A_i = mS_0 - (m-1)S_i, \quad \text{for } i = 1, \ldots, m, \tag{5.30}$$

and

$$B_j = nT_0 - (n-1)T_j, \quad \text{for } j = 1, \ldots, n. \tag{5.31}$$

(This is what is referred to as the *jackknifing process*, as applied to the sample variance.) Set

$$\bar{A} = \sum_{i=1}^{m} \frac{A_i}{m} \quad \text{and} \quad \bar{B} = \sum_{j=1}^{n} \frac{B_j}{n}, \tag{5.32}$$

and compute

$$V_1 = \sum_{i=1}^{m} \frac{(A_i - \bar{A})^2}{m(m-1)} \tag{5.33}$$

and

$$V_2 = \sum_{j=1}^{n} \frac{(B_j - \bar{B})^2}{n(n-1)}. \tag{5.34}$$

Finally, set

$$Q = \frac{\bar{A} - \bar{B}}{\sqrt{V_1 + V_2}}. \tag{5.35}$$

a. *One-Sided Upper-Tail Test*. To test

$$H_0 : \gamma^2 = 1$$

versus

$$H_1 : \gamma^2 > 1,$$

at the approximate α level of significance,

$$\text{Reject } H_0 \text{ if } Q \geq z_\alpha; \quad \text{otherwise do not reject,} \tag{5.36}$$

where, as previously, z_α is the upper αth percentile for the standard normal distribution.

b. *One-Sided Lower-Tail Test*. To test

$$H_0 : \gamma^2 = 1$$

versus

$$H_1 : \gamma^2 < 1,$$

at the approximate α level of significance,

$$\text{Reject } H_0 \text{ if } Q \leq -z_\alpha; \quad \text{otherwise do not reject.} \tag{5.37}$$

c. *Two-Sided Test*. To test

$$H_0 : \gamma^2 = 1$$

versus

$$H_1 : \gamma^2 \neq 1,$$

at the approximate α level of significance,

$$\text{Reject } H_0 \text{ if } |Q| \geq z_{\alpha/2}; \quad \text{otherwise do not reject.} \tag{5.38}$$

This two-sided procedure is the two-sided symmetric test with $\alpha/2$ probability in each tail of the approximating standard normal distribution.

When m and n are small and equal, the approximate level α test procedures given by (5.36), (5.37), and (5.38) can be improved slightly by replacing z_α and $z_{\alpha/2}$ by $t_{m+n-2,\alpha}$ and $t_{m+n-2,\alpha/2}$, respectively, where $t_{m+n-2,\alpha}$ is the upper α percentile point of the t distribution with $m + n - 2$ degrees of freedom. To find $t_{m+n-2,\alpha}$ for given sample sizes m and n, we use the R command `qt(1-α, m+n-2)`. For example, to find $t_{14,.05}$, we apply `qt(.95,14)` and obtain $t_{14,.05} = 1.761$.

Ties

The jackknife procedures are well defined when ties within or between the X's and Y's occur and further adjustments are not necessary.

EXAMPLE 5.2 ***Southern Armyworm and Pokeweed.***

Burnett and Jones (1973) investigated the idea of coevolution between the southern armyworm and pokeweed. They suspected that armyworms might have developed a greater resistance to the toxins from pokeweed populations that lie within their geographic range than to the toxins of pokeweeds found in other areas of the country. Pokeweed plants from Florida populations (within the range of southern armyworms) and Kentucky populations (well north of the range of southern armyworms) were raised under similar conditions in greenhouses for the study. Larval southern armyworms were then used in feeding experiments to determine whether they would eat less of the Kentucky pokeweed possessing toxins to which they are not resistant. Five samples of Kentucky pokeweed and five samples of Florida pokeweed were used, with each such sample being exposed to 10 separate southern armyworm larvae. (There were 100 *different* larvae used in the experiment.) Following an individual larva's 24-h feeding period (in darkness at $25 \pm 1\,^{\circ}$C) on a moist filter paper in a disposable petri dish, the fecal material of the larva was dried overnight in an oven and weighed the following day. This was then used as a measure of quantity of the plant material ingested by the armyworm larva during its feeding. The data in Table 5.2 are the average (over the 10 armyworm larvae replications) dry feces weights (in milligrams) for the five Kentucky pokeweed and five Florida pokeweed plant samples.

Table 5.2 Average Dry Feces Weight (mg)

Kentucky pokeweed	Florida pokeweed
6.2	9.5
5.9	9.8
8.9	9.5
6.5	9.6
8.6	10.3

Source: W. C. Burnett, Jr., and S. B. Jones, Jr. (1973).

It is clear from the data in Table 5.2 that the southern armyworm larvae had a tendency to eat more (on the average) of the Florida pokeweed than the Kentucky pokeweed. As a result, if we are interested in assessing whether there is any difference in the variability or dispersion of the southern armyworm's consumption of the two pokeweed varieties, it would not be appropriate to directly apply one of the Ansari–Bradley procedures discussed in Section 5.1, because they require equality of the respective population medians. However, the jackknifed variances procedure of this section makes no such assumption and can be applied directly to the sample data.

Letting X correspond to the Kentucky pokeweed observations and Y to the Florida pokeweed data, we consider testing the null hypothesis of no difference in dispersion against the alternative that the variability is greater for Kentucky pokeweed; that is, we want to use procedure (5.36) to test $H_0 : \gamma^2 = 1$ against the alternative $H_1 : \gamma^2 > 1$.

The five Kentucky pokeweed subgroups of four observations each, corresponding to the five different ways to delete a single measurement, are given by

$$G_1 = \{6.2, 5.9, 8.9, 6.5\}, \quad G_2 = \{6.2, 5.9, 8.9, 8.6\}, \quad G_3 = \{6.2, 5.9, 6.5, 8.6\},$$

$$G_4 = \{6.2, 8.9, 6.5, 8.6\} \quad \text{and} \quad G_5 = \{5.9, 8.9, 6.5, 8.6\}.$$

Following (5.23), the sample average and sample variance associated with subgroup G_1 are

$$\bar{X}_1 = \frac{6.2 + 5.9 + 8.9 + 6.5}{4} = 6.875 \tag{5.39}$$

and

$$D_1^2 = \frac{(6.2 - 6.875)^2 + (5.9 - 6.875)^2 + (8.9 - 6.875)^2 + (6.5 - 6.875)^2}{3} = 1.8825. \tag{5.40}$$

In a similar manner, it follows from (5.24) that the sample averages and sample variances for the other four Kentucky pokeweed subgroups are

Subgroup G_i	$\bar{X}_i$	D_i^2
G_2	7.4	2.46
G_3	6.8	1.50
G_4	7.55	1.95
G_5	7.475	2.2425

(5.41)

Proceeding in the same fashion with the Florida pokeweed data, we obtain the five deleted-observation subgroups

$$H_1 = \{9.5, 9.8, 9.5, 9.6\}, \quad H_2 = \{9.5, 9.8, 9.5, 10.3\}, \quad H_3 = \{9.5, 9.8, 9.6, 10.3\},$$

$$H_4 = \{9.5, 9.5, 9.6, 10.3\}, \quad \text{and} \quad H_5 = \{9.8, 9.5, 9.6, 10.3\}.$$

Using (5.25), we obtain the associated subgroup sample means and sample variances to be

Subgroup H_j	$\bar{Y}_j$	E_j^2
H_1	9.6	.02
H_2	9.775	.1425
H_3	9.8	.1267
H_4	9.725	.1492
H_5	9.8	.1267

(5.42)

Taking natural logarithms of the D_i^2's (in (5.40) and (5.41)) and the E_j^2's (in (5.42)), it follows from (5.26) and (5.27) that

$$S_1 = .6326, \quad S_2 = .9002, \quad S_3 = .4055, \quad S_4 = .6678, \quad S_5 = .8076 \tag{5.43}$$

and

$$T_1 = -3.9120, \quad T_2 = -1.9484, \quad T_3 = -2.0662, \quad T_4 = -1.9027, \quad T_5 = -2.0662. \tag{5.44}$$

Finally, using all five of the X sample observations, we see from (5.28) that

$$\bar{X}_0 = 7.22 \quad \text{and} \quad S_0 = \ln\left[\sum_{s=1}^{5} \frac{(X_s - 7.22)^2}{4}\right] = \ln 2.007 = .6966.$$

Similarly, using all five of the Y sample observations, it follows from (5.29) that

$$\bar{Y}_0 = 9.74 \quad \text{and} \quad T_0 = \ln\left[\sum_{t=1}^{5} \frac{(Y_t - 9.74)^2}{4}\right] = \ln .113 = -2.1804.$$

Combining these complete sample values with the subgroup calculations in (5.43) and (5.44) via the jackknifing process in (5.30) and (5.31), we obtain

$$A_1 = 5(.6966) - 4(.6326) = .9526, \quad A_2 = 5(.6966) - 4(.9002) = -.1178,$$

$$A_3 = 5(.6966) - 4(.4055) = 1.861, \tag{5.45}$$

$$A_4 = 5(.6966) - 4(.6678) = .8118, \quad A_5 = 5(.6966) - 4(.8076) = .2526$$

and

$$\begin{aligned} B_1 &= 5(-2.1804) - 4(-3.9120) = 4.746, \\ B_2 &= 5(-2.1804) - 4(-1.9484) = -3.1084, \\ B_3 &= 5(-2.1804) - 4(-2.0662) = -2.6372, \\ B_4 &= 5(-2.1804) - 4(-1.9027) = -3.2912, \\ B_5 &= 5(-2.1804) - 4(-2.0662) = -2.6372. \end{aligned} \tag{5.46}$$

From (5.32), (5.33), and (5.45), we see that

$$\bar{A} = .7520 \quad \text{and} \quad V_1 = \sum_{i=1}^{5} \frac{(A_i - \bar{A})^2}{5(4)} = .1140.$$

Similarly, from (5.32), (5.34), and (5.46), we have

$$\bar{B} = -1.3856 \quad \text{and} \quad V_2 = \sum_{j=1}^{5} \frac{(B_j - \bar{B})^2}{5(4)} = 2.3664.$$

It then follows from (5.35) that

$$Q = \frac{.7520 - (-1.3856)}{(.1140 + 2.3664)^{1/2}} = 1.36.$$

(We note that the R command `MillerJack(pokeweed$x, pokeweed$y)` can also be used to obtain the value $Q = 1.36$ for the (x, y) data in Table 5.2.)

Hence, from (5.36) and the R command `pnorm(·)`, we see that the lowest significance level at which we can reject H_0 in favor of $\gamma^2 > 1$ with the observed value of the test statistic Q (i.e., the one-sided P-value) is approximately $1 - \texttt{pnorm(1.36)} = .0869$. (Since the sample sizes $m = n = 5$ are small and equal, we can also use the t-distribution with $m + n - 2 = 5 + 5 - 2 = 8$ degrees of freedom to approximate the P-value. Using the R command `pt(·)`, we have P-value $\approx 1 - \texttt{pt(1.36, 8)} = \texttt{1 - .8945} = \texttt{.1055}$, in general agreement with the normal approximation.) Thus, there is only mild evidence in the sample data to indicate greater variability for the Kentucky pokeweed population.

Comments

16. *Assumptions*. Note that Assumptions A1, A2, and A4 do not impose the severe condition that the two underlying populations have equal medians. Although these assumptions do require that the two underlying populations have finite fourth moments, this is not a serious restriction for most common data collection settings. This means that the Miller procedures are applicable in more general settings than the Ansari–Bradley procedures based on C (5.6). (See Comment 7.)

17. *Testing γ^2 Equal to Some Specified Value Other Than One*. To test the hypothesis $\gamma^2 = \gamma_0^2$, where γ_0^2 is some specific positive number different from 1, we obtain the modified observations $Y_j' = \gamma_0 Y_j, j = 1, \ldots, n$, and compute Q(5.35) using the X's and the Y''s (instead of the X's and the Y's). The appropriate procedure (5.36), (5.37), or (5.38) may then he applied as described.

18. *Asymptotic Distribution-Freeness*. Asymptotically (i.e., for infinitely large samples) the true level of the tests defined by (5.36), (5.37), and (5.38) will agree with the nominal level α. Subject to Assumptions A1, A2, and A4, this asymptotic result does not depend on the underlying populations of the X's and Y's. More precisely, subject to Assumptions A1, A2, and A4, Q(5.35) has an asymptotic $N(0,1)$ distribution when H_0 is true. Since this asymptotic distribution does not depend on the underlying populations of the X's and Y's, we say that the tests based on Q are asymptotically distribution-free. Of course, in practice, we do not have the luxury of infinite samples. Thus, in any particular setting with m and n large, although the level of any of the tests based on Q is not necessarily exactly equal to the nominal level α, we hope it is close to it. The closeness of this approximation depends on m, n, α, and the underlying populations, but, for fixed α, the closeness generally improves as m and n increase, regardless of the underlying populations. In the case of the Q tests, the reader is cautioned that the question of how large m and n should be, in order for the normal approximation to be good, is relatively unanswered. Exact null distribution tables for Q cannot be provided for specified values of m and n, because the exact null distribution of Q depends on the underlying X and Y populations; thus, exact critical points would vary with the forms of the X and Y populations. The procedures (5.36), (5.37), and (5.38) based on Q, therefore, are not (strictly) distribution-free.

19. *Alternative Method of Calculation*. For $i = 1, \ldots, m, S_i$ is the natural log of the sample variance for the $(m-1)$ X observations $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_m$. Similarly, for $j = 1, \ldots, n, T_j$ is the natural log of the sample variance for the $(n-1)$ Y observations $Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_n$. The following equivalent formulas for D_i^2 (5.24) and E_j^2 (5.25), namely,

$$D_i^2 = \frac{\sum_{s \neq i}^{m} X_s^2 - \frac{\left(\sum_{s \neq i}^{m} X_s\right)^2}{m-1}}{m-2}$$

and

$$E_j^2 = \frac{\sum_{t \neq j}^{n} Y_t^2 - \frac{\left(\sum_{t \neq j}^{n} Y_t\right)^2}{n-1}}{n-2},$$

are computationally more convenient than the definitions given in (5.24) and (5.25), respectively.

20. *General Jackknife Technique*. The jackknife technique applied in this section to the problem of testing two-sample dispersion hypotheses is a tool that can be used successfully in certain statistical problems to accomplish two

goals: (a) reducing the bias of point estimators and (b) generating broadly applicable and reasonably powerful test procedures for problems where classical test procedures are sensitive to nonnormality of the underlying populations. Although the jackknife technique is not always effective in achieving these goals (see Miller (1964)), it performs well in the two-sample dispersion problem, providing us with asymptotically distribution-free test procedures for a problem where deviations from normality of the underlying populations can be disastrous for the classical $\mathcal{F}$-test for equal variances (cf. Box (1953), Shorack (1969), and Comment 26) and where distribution-free rank tests for the problem are limited in their applicability (see Comment 7 and Moses (1963)). Miller (1968) discussed in detail the advantages gained by applying the jackknife technique to this two-sample dispersion problem. (See also Comment 21.)

21. *Motivation*. The jackknife is an extension of an idea due to Quenouille (1949) and is designed to reduce the bias of an estimator. Suppose we have a sample of N independent observations, each from the same distribution that depends on an unknown parameter θ. Assume that we have a general method for estimating θ and let $\hat{\theta}$ denote this estimator based on all N observations. Divide the data into n groups of size k. Let $\hat{\theta}_{-i}, i = 1, \ldots, n$, denote the estimator of θ obtained by deleting the ith group and estimating θ from the remaining $(n-1)k$ observations. Define $\tilde{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_{-i}$. The jackknife estimator of θ is $\tilde{\theta} = \sum_{i=1}^{n} \tilde{\theta}_i / n$. In certain situations, the jackknife can be shown to be less biased than the estimator $\hat{\theta}$. Tukey (1958, 1962) extended the jackknife to construct approximate significance tests and confidence intervals for θ.

The traditional estimator of the variance of $\tilde{\theta}$, in the case of $k = 1$, is

$$\widehat{V}^2 = \frac{1}{N(N-1)} \sum_{i=1}^{N} (\widehat{\theta}_{-i} - \widehat{\theta})^2.$$

Asymptotic $100(1-\alpha)\%$ confidence intervals for θ are then

$$(\tilde{\theta} - z_{\alpha/2}\widehat{V}, \tilde{\theta} + z_{\alpha/2}\widehat{V}).$$

Under certain conditions (i.e., when $\widehat{\theta}$ is not "sufficiently smooth"), $\widehat{V}^2$ may be inconsistent. One such situation is when $\widehat{\theta}$ is a sample quantile (see, e.g., Miller (1974)). To overcome this difficulty, Shao (1988), Shao and Wu (1989), and Wu (1990) have studied a "delete-d" jackknife variance estimator. See also Maesono (1996) for further details.

In the dispersion problem, Miller jackknifed the natural logs of the sample variances rather than the sample variances themselves because the natural log transformation tends to stabilize the variance and create a distribution that is "closer" to the normal distribution. The statistic $\bar{B}$ (5.32) is an estimator of $\ln\{\text{var}(Y)\}$, the statistic $\bar{A}$ (5.32) is an estimator of $\ln\{\text{var}(X)\}$, and $\bar{A} - \bar{B}$ estimates $\ln\{\text{var}(X)/\text{var}(Y)\} = \ln \gamma^2$. The quantity $(V_1 + V_2)^{1/2}$ in the denominator of Q (5.35) is an estimator of the standard deviation of $\bar{A} - \bar{B}$. If, for example, the X's are more disperse than the Y's, $\bar{A} - \bar{B}$ would tend to be large, and this is partial motivation for procedure (5.36).

22. *Generalization*. In its most general formulation (see Comment 21), the jackknife process can be applied to any randomly selected partition of the data set into subsets of size k each, where k can be any positive integer that is a factor of the number of observations in the data set. In fact, Miller (1968) discussed the test for dispersion based on jackknifing the natural logs of the sample variances in the context of this most general formulation. However, for any integer $k > 1$, the associated Miller jackknifed variances procedures have the rather severe deficiency that it is possible for two different people to arrive at different conclusions when analyzing the same data set with the same test and at the same significance level. This possibility arises because of the variety of ways that the data set could be randomly partitioned into subsets of size k each. To avoid this undesirable feature, we have chosen to discuss the jackknifed variances procedure only for $k = 1$. In this case, there is no flexibility in partitioning a data set and the associated Miller test procedures are unambiguous in their conclusions.
23. *t Distribution Approximation*. The standard normal percentiles used in (5.36) to (5.38) should be replaced by the corresponding percentile points for a t distribution with $m + n - 2$ degrees of freedom only when m and n are small and equal. For other situations, the matter of which t distribution (i.e., what degrees of freedom) should be used to find the approximating percentile is somewhat ambiguous.
24. *Point Estimators and Confidence Intervals and Bounds for* γ^2. Point estimators of and approximate confidence intervals and bounds for γ^2 can be readily obtained from the jackknife procedures. In particular, the estimator for γ^2 associated with the jackknifed variances procedures is

$$\tilde{\gamma}^2 = e^{\bar{A}-\bar{B}}. \tag{5.47}$$

Moreover, an asymptotically distribution-free confidence interval for γ^2, with approximate confidence coefficient $1 - \alpha$, based on the jackknifed variances procedures is given by

$$(\gamma_L^2, \gamma_U^2), \tag{5.48}$$

where

$$\gamma_L^2 = e^{[(\bar{A}-\bar{B})-z_{\alpha/2}(V_1+V_2)^{1/2}]} \tag{5.49}$$

and

$$\gamma_U^2 = e^{[(\bar{A}-\bar{B})+z_{\alpha/2}(V_1+V_2)^{1/2}]}. \tag{5.50}$$

With γ_L^2 and γ_U^2 given by (5.49) and (5.50), we have

$$P_{\gamma^2}\{\gamma_L^2 < \gamma^2 < \gamma_U^2\} \approx 1 - \alpha. \tag{5.51}$$

The corresponding asymptotically distribution-free approximate $100\ (1 - \alpha)\%$ *lower* and *upper* confidence bounds for γ^2 based on the jackknifed variances procedures are

$$\gamma_L^{*2} = e^{[(\bar{A}-\bar{B})-z_{\alpha}(V_1+V_2)^{1/2}]} \tag{5.52}$$

and

$$\gamma_U^{*2} = e^{[(\bar{A}-\bar{B})+z_\alpha(V_1+V_2)^{1/2}]}, \tag{5.53}$$

respectively, satisfying

$$P_{\gamma^2}\{\gamma_L^{*2} < \gamma^2\} \approx 1-\alpha \quad \text{and} \quad P_{\gamma^2}\{\gamma^2 < \gamma_U^{*2}\} \approx 1-\alpha. \tag{5.54}$$

For the armyworm/pokeweed data in Example 5.2, the point estimate of γ^2 is $\tilde{\gamma}^2 = e^{[.7520-(-1.3856)]} = e^{2.1376} = 8.479$ and, with $\alpha = .0548$, the approximate 94.52% lower confidence bound for γ^2 is

$$\begin{aligned}\gamma_L^{*2} &= e^{[(.7520-(-1.3856))-1.6(.1140+2.3664)^{1/2}]} \\ &= e^{-.3823} = .6823.\end{aligned}$$

25. *Asymptotic Coverage Probability*. Asymptotically (i.e., for infinitely large samples), the true coverage probabilities of the confidence interval defined by (5.48) and the confidence bounds defined by (5.52) and (5.53) will agree with the nominal confidence coefficient $1-\alpha$. Subject to Assumptions A1, A2, and A4, this asymptotic result does not depend on the form of the distribution function $H(\cdot)$ in (5.2). Hence, we say that the interval given by (5.48) and the bounds given by (5.52) and (5.53) are an asymptotically distribution-free confidence interval and asymptotically distribution-free confidence bounds, respectively, for γ^2.

    The interval (5.48) has also been defined so that it is "asymptotically symmetric." Here, the word symmetric refers to the equal-tail probabilities of $\alpha/2$. The $1-\alpha$ confidence interval for γ^2 defined by (5.48) can be called asymptotically symmetric, because it is constructed so that $P_{\gamma^2}(\gamma_U^2 \le \gamma^2) \approx P_{\gamma^2}(\gamma_L^2 \ge \gamma^2) \approx \alpha/2$. The approximation is a result of approximating the true distribution of the statistic $\bar{A}-\bar{B}$ by its asymptotic normal distribution.

26. *Lack of Robustness of the Classical $\mathcal{F}$-Test for Equal Variances*. The classical normal theory $\mathcal{F}$-test for equality of variances is not robust with respect to the assumption of normality in the sense that when the underlying populations are not normal, the true level of an $\mathcal{F}$-test that is supposed to be of size α may be quite far from α. Box (1953) gave examples in which the level of the $\mathcal{F}$-test is specified to be .05, although the actual level is as large as .166 or as small as .0056. Furthermore, there exist nonnormal populations in which, even with large samples, the level of the $\mathcal{F}$-test will not be what it is supposed to be. This nonrobustness, which was pointed out as early as 1931 by Pearson (1931), has been emphasized by Box (1953) and more recently by Miller (1968) and Shorack (1969).

Properties

1. *Consistency*. The tests given by (5.36), (5.37), and (5.38) are consistent against the alternatives $\gamma^2 >$, $<$, and $\neq 1$, respectively.
2. *Asymptotic Normality.* See Miller (1968).
3. *Efficiency.* See Miller (1968) and Section 5.5.

Problems

13. The data in Table 5.3 are a portion of those collected by Bugyi et al. (1969) in a study concerned with ascertaining sodium ion content in erythrocytes (red blood cells). Such determinations are helpful in the diagnoses of certain diseases, where merely knowing the sodium ion content in plasma does not provide sufficient information. However, erythrocyte sodium ion determination is extremely variable and subject to error. This prompted the authors to propose using the flame photometric method to determine sodium ion content in erythrocytes, with the hope of providing better accuracy than can be obtained with the inefficient procedures commonly used at that time.

One of the ways to assess the accuracy of the proposed method is to compare the variation in erythrocyte sodium ion measurements with the variation in plasma sodium ion determinations, where it is known that the measurement variation for the flame photometric method is acceptably low. Sodium ion determinations were obtained by the flame photometric method on each of 10 plasma and 10 erythrocyte samples. Table 5.3 gives the sodium ion content in mequiv/l for the 20 samples.

Use a Miller jackknife procedure to test the hypothesis of equal dispersions for the plasma and erythrocyte sodium ion measurements against the alternative of interest in the study. Find the approximate P-value for the test.

14. Consider the chorioamnion permeability data given in Table 4.1. Find the approximate P-value for the Miller jackknife test of the hypothesis of equal dispersions versus the alternative that the variation in tritiated water diffusion across human chorioamnion is different at term than at 12–26 weeks gestational age. Compare your findings with those obtained in Problem 1 using an Ansari–Bradley procedure to analyze the data.

15. Consider the television-viewing behavior data in Table 4.4. Find the approximate P-value for the Miller jackknife test of equal dispersions versus the alternative that there is more variability in the time spent in the room after witnessing the violent behavior for those children who had previously watched the *Karate Kid* than for those children who had previously watched parts of the 1984 Summer Olympic Games. Comment on your analysis in conjunction with the findings of Problems 3 and 4.5.

16. Consider the alcoholic intake data in Table 4.2. Find the approximate P-value for the Miller jackknife test of whether there is any difference in dispersions for the control and SST data. Comment on your analysis relative to the findings in Problem 9 and Example 4.2.

17. For the sodium ion determination data of Table 5.3, compute the value of the estimator $\tilde{\gamma}^2$ defined in expression (5.47).

18. For the chorioamnion permeability data given in Table 4.1, obtain the value of the estimator $\tilde{\gamma}^2$ defined in expression (5.47).

Table 5.3 Sodium Ion Content (mequiv/l)

Plasma	Erythrocytes
147.0	10.3
147.0	12.2
146.0	16.5
145.0	19.3
146.5	8.3
161.0	15.2
141.0	27.0
146.5	26.3
145.0	17.5
153.5	21.7

Source: H. I. Bugyi, E. Magnier, W. Joseph, and G. Frank (1969).

19. Obtain the value of the estimator $\tilde{\gamma}^2$ defined in expression (5.47) for the television-viewing behavior data in Table 4.4.

20. Compute the value of the estimator $\tilde{\gamma}^2$ defined in expression (5.47) for the alcoholic intake data in Table 4.2.

21. With respect to the chorioamnion permeability data given in Table 4.1, find an approximate 96.6% confidence interval for γ^2 utilizing the procedure discussed in Comment 24.

22. Consider the alcoholic intake data in Table 4.2. Using the procedure discussed in Comment 24, find an approximate 93.72% confidence interval for γ^2.

23. Consider the television-viewing behavior data in Table 4.4. Labeling the Olympic watchers data as the X sample, use the procedure discussed in Comment 24 to find an approximate 98.96% upper confidence bound for γ^2.

24. Consider the sodium ion determination data of Table 5.3. Labeling the erythrocyte sodium ion measurements as the X sample, use the procedure discussed in Comment 24 to find an approximate 91.92% lower confidence bound for γ^2.

5.3 A DISTRIBUTION-FREE RANK TEST FOR EITHER LOCATION OR DISPERSION (LEPAGE)

Hypothesis

Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent random samples satisfying Assumptions Al and A2 from continuous populations with distribution functions F and G, respectively, satisfying the location-scale parameter model relationships in (5.2) and (5.3).

Under these assumptions, we are interested in assessing whether there are differences in *either* the location parameters (i.e., medians) θ_1 and θ_2 *or* the scale parameters η_1 and η_2 for the X and Y populations. Thus, we are interested in testing the null hypothesis H_0 (5.1) versus the general alternative H_1: $[\theta_1 \neq \theta_2$ and/or $\eta_1 \neq \eta_2]$. Note that under the location-scale parameter model, as stated in (5.2) and (5.3), the null hypothesis H_0 (5.1) reduces to H_0: $[\theta_1 = \theta_2$ and $\eta_1 = \eta_2]$, corresponding to the assertion that both the population location parameters and the population scale parameters are equal.

Procedure

To compute the Lepage two-sample location-scale statistic D, order the combined sample of $N = (m + n)$ X-values and Y-values from least to greatest. Let S_j denote the combined samples rank of Y_j, for $j = 1, \ldots, n$, and let $W = \sum_{j=1}^{n} S_j$ be the Wilcoxon rank sum statistic defined in (4.3). In addition, for $j = 1, \ldots, n$, let R_j be the score assigned to Y_j by the Ansari–Bradley scoring scheme discussed in the Procedure of Section 5.2 and let $C = \sum_{j=1}^{n} R_j$ be the Ansari–Bradley scale statistic defined in (5.6). The Lepage rank statistic is then defined by

$$D = \frac{[W - E_0(W)]^2}{\text{var}_0(W)} + \frac{[C - E_0(C)]^2}{\text{var}_0(C)}, \tag{5.55}$$

where $E_0(W)$ and $\text{var}_0(W)$ are the expected value and variance of W under H_0 (5.1), as given in (4.7) and (4.8), respectively, and $E_0(C)$ and $\text{var}_0(C)$ are the corresponding expected value and variance of C under H_0 (5.1), as stated in (5.10) and

(5.11), respectively, when $N = (m + n)$ is an even number, or in (5.12) and (5.13), respectively, when N is odd. Thus, if we let W^* (4.9) and C^* (5.20) represent the standardized forms for the Wilcoxon rank sum statistic and Ansari–Bradley scale statistic, respectively, then the Lepage statistic D can be written as

$$D = (W^*)^2 + (C^*)^2. \tag{5.56}$$

To test H_0 (5.1), corresponding to the equality of both the location and the scale parameters for the X and Y populations, versus the general alternatives that the location parameters are different or the scale parameters are different or both, corresponding to

$$H_1 : [\theta_1 \neq \theta_2 \text{ and/or } \eta_1 \neq \eta_2], \tag{5.57}$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } D \geq d_\alpha; \text{ otherwise do not reject,} \tag{5.58}$$

where the constant d_α is chosen to make the type I error probability equal to α. The constant d_α is the upper α percentile for the null H_0(5.1) distribution of D. Comment 28 explains how to obtain the critical value d_α for sample sizes m and n and available values of α.

Large-Sample Approximation

The large-sample approximation is based on the fact that when H_0 (5.1) is true, the statistic D has, as $\min(m, n)$ tends to infinity, a chi-square distribution with 2 degrees of freedom (see Comment 31 for indications of the proof). The large-sample approximation for the exact level α procedure in (5.58) is

$$\text{Reject } H_0 \text{ if } D \geq \chi^2_{2,\alpha}; \text{ otherwise do not reject,} \tag{5.59}$$

where $\chi^2_{2,\alpha}$ is the upper α percentile point of the chi-square distribution with 2 degrees of freedom. To find $\chi^2_{2,\alpha}$, we use the R command `qchisq(1 - α, 2)`. For example, to find $\chi^2_{2,.05}$ we apply `qchisq(.95, 2)` and obtain $\chi^2_{2,.05} = 5.991$.

Ties

If there are ties among the X and/or Y observations, we modify the standardized Wilcoxon rank sum statistic W^* and the standardized Ansari–Bradley scale statistic C^* in the manners prescribed for the large-sample approximations in the Ties portions of Sections 4.1 and 5.1, respectively. When applying either the small-sample procedure in (5.58) or the large-sample approximation in (5.59), the Lepage statistic D should be computed using these ties-modified versions of W^* and C^*. The corresponding modified version of procedure (5.58) in the case of ties among the X and/or Y observations is only approximately, and not exactly, of significance level α. (To get an exact level α test even in this tied setting, see Comment 32.)

EXAMPLE 5.3 ***Effect of Maternal Steroid Therapy on Platelet Counts of Newborn Infants.***

Autoimmune thrombocytopenic purpura (ATP) is a disease in which the patient produces antibodies to his/her own platelets. Due to transplacental passage of antiplatelet antibodies during pregnancy, children of women with ATP are often born with low platelet counts. For this reason, there is medical concern that a vaginal delivery for a mother with ATP could result in intracranial hemorrhage for the infant. However, the proper obstetrical management of pregnant women with ATP is controversial. Most doctors have advocated cesarean section as the preferable method of delivery for mothers with ATP. Others suggest that cesarean section, with its obvious complications for both mother and infant, be avoided unless there is some additional obstetrical reason for it. Karpatkin, Porges, and Karpatkin (1981) studied the effect of administering the corticosteroid prednisone to pregnant women with ATP with the intent of raising the infants' platelet counts to safe levels during their deliveries. The rationale for this treatment is the fact that steroids, in general, increase the platelet counts in patients with ATP by blocking splenic destruction of antibody-coated platelets. In theory, then, the corticosteroid prednisone should cross the placenta, enter the infant's circulation, and prevent splenic removal of those infant's platelets that are coated by the mother's antibodies.

The data in Table 5.4 are a subset of the data obtained by Karpatkin et al. in their study of the effect that administration of prednisone to pregnant women with ATP had on their infants' platelet counts. All the infants included in this example were delivered vaginally. Table 5.4 gives the platelet counts (per cubic millimeter) of 10 infants whose mothers received the steroid prednisone prior to delivery and 6 infants whose mothers were not treated with prednisone prior to delivery. All 16 mothers in the study were diagnosed with ATP.

The primary interest in the study is in whether or not the predelivery administration of prednisone typically leads to an increased newborn platelet count. Thus, the principal statistical issue in the study is that of a possible difference in locations for the prednisone and nonprednisone populations. However, there is some concern that the administration of predelivery prednisone could also lead to a rather large increase in variability in the newborn platelet counts. (Such a finding would certainly affect our interpretation of any possible increase in typical platelet count resulting from the prednisone.) As a result, we will apply the Lepage test procedure to test H_0 (5.1) versus the general alternative H_1 5.57). For the purposes of illustration, we consider the exact procedure (5.58) with level of

Table 5.4 Platelet Counts of Newborn Infants (per Millimeter^3)

Mothers given prednisone	Mothers not given prednisone
120,000	12,000
124,000	20,000
215,000	112,000
90,000	32,000
67,000	60,000
95,000	40,000
190,000	
180,000	
135,000	
399,000	

Source: M. Karpatkin, R. F. Porges, and S. Karpatkin (1981).

significance $\alpha = .02$. For convenience, we take the infant platelet count data for mothers given prednisone to be the Y sample ($n = 10$) and the corresponding control (nonprednisone) data to be the X sample ($m = 6$). Using the R command `cLepage(.02, 6, 10)`, we obtain $d_{.02} = 6.903$ so that procedure (5.58) is

$$\text{Reject } H_0 \text{ if } D \geq 6.903. \tag{5.60}$$

To calculate D, we need first to calculate the standardized versions of the Wilcoxon rank sum and Ansari–Bradley statistics. Proceeding as in (4.3), we note that the combined samples ranks for the 10 Y observations are 10, 11, 16, 7, 6, 8, 14, 13, 12, and 15, yielding a value of

$$W = 10 + 11 + 16 + 7 + 6 + 8 + 14 + 13 + 12 + 15 = 112$$

for the Wilcoxon rank sum statistic. Since there are no ties among the 16 X and Y observations, it follows from (4.7), (4.8), and (4.9) that the standardized form of W for the data in Table 5.4 is

$$W^* = \frac{112 - \{10(6 + 10 + 1)/2\}}{\{6(10)(6 + 10 + 1)/12\}^{1/2}} = 2.929. \tag{5.61}$$

For the calculation of the standardized Ansari–Bradley statistic C^*, we observe that the Ansari–Bradley scores (as defined in the Procedure of Section 5.1) for the 10 Y observations are 7, 6, 2, 7, 6, 8, 3, 4, 5, and 1. From (5.6), this produces a value of

$$C = 7 + 6 + 2 + 7 + 6 + 8 + 3 + 4 + 5 + 1 = 49.$$

Since $N = 6 + 10 = 16$ is an even number and there are no ties among the 16 X and Y observations, it follows from (5.10), (5.11), and (5.14) that the standardized form of C for the data in Table 5.4 is

$$C^* = \frac{49 - \{10(16 + 2)/4\}}{\left\{\dfrac{10(6)(16 + 2)(16 - 2)}{48(16 - 1)}\right\}^{1/2}} = .873. \tag{5.62}$$

Using these values of W^* (5.61) and C^* (5.62) in (5.56) yields

$$D = (2.929)^2 + (.873)^2 = 9.34, \tag{5.63}$$

which, in view of expression (5.60), tells us to reject H_0 at the $\alpha = .02$ level, because $D = 9.34 > d_{.02} = 6.903$. Hence, there is rather strong evidence that there are differences in locations or scales (or both) between the prednisone and control infant platelet count populations. In fact, using the R command `pLepage(`**platelet.counts$x**, **platelet.counts$y**`)`, the P-value for these data with observed value $D = 9.34$ is given by P-value = `pLepage(`**platelet.counts$x**, **platelet.counts$y**`)` = .0035, providing an even stronger statement in favor of the alternative H_1 (5.57).

For the large-sample approximation, we see from (5.59) that the approximate P-value for these data is

$$P\text{-value} \approx P(Q \geq 9.34),$$

where Q has a chi-square distribution with 2 degrees of freedom. This approximate P-value is then given by $1 - \texttt{pchisq(9.34, 2)} = 1 - .9906 = .0094$, in general agreement with the exact P-value of .0035 previously obtained.

We conclude this example by noting that the large value of D is due primarily to a large value of W^*. This would suggest intuitively that the rejection of H_0 is due primarily to a difference in locations between the infant platelet counts for the prednisone and control populations. However, we emphasize that such a conclusion is not statistically justified through the application of the general Lepage procedure (5.58). The only valid conclusion based on the Lepage procedure is that of the general alternative H_1 (5.57). (If you do, however, apply the Wilcoxon rank sum procedure of Section 4.1 to the data in Table 5.4, you would be able to conclude that there is, indeed, a difference in locations between the infant platelet counts for the prednisone and control populations. In view of this fact, would it be legitimate to then apply the Ansari–Bradley procedure of Section 5.1 directly to the data in Table 5.4 to test for possible scale differences in the two populations?)

Comments

27. *Motivation for the Test*. From Section 4.1 we know that a large value of $(W^*)^2$ is indicative of a possible difference in locations for the X and Y populations. We also know from Section 5.1 that a large value of $(C^*)^2$ is indicative of a possible difference in dispersions for the X and Y populations. Since D (5.56) will be large if and only if $(W^*)^2$ is large or $(C^*)^2$ is large or both, then such a large value of D is indicative of $\theta_1 \neq \theta_2$ or $\eta_1 \neq \eta_2$ or both. This serves as partial motivation for the test procedure given by (5.58).

28. *Derivation of the Distribution of D under H_0 (No-Ties Case)*. Under H_0 (5.1), each of the $\binom{N}{n}$ possible "meshings" of the X's and Y's has probability $1/\binom{N}{n}$. This fact can be used to obtain the null distribution of D (5.56). We illustrate the steps involved in constructing this null distribution for the simple case $m = 2$, $n = 2$. Since $N = 4$, we must consider $\binom{4}{2} = 6$ possible meshings of the X and Y observations. For this setting, it follows from (4.7) and (4.8) that $E_0(W) = 2(2 + 2 + 1)/2 = 5$ and $\text{var}_0(W) = 2(2)(2 + 2 + 1)/12 = \frac{5}{3}$. Similarly, from (5.10) and (5.11), we have $E_0(C) = 2(4 + 2)/4 = 3$ and $\text{var}_0(C) = [2(2)(4 + 2)(4 - 2)]/48(4 - 1) = \frac{1}{3}$. Thus, for $m = n = 2$, we have $(W^*)^2 = 3(W - 5)^2/5$ and $(C^*)^2 = 3(C - 3)^2$. Using these facts and the same approach taken in Comments 4.3 and 5.4 for the calculations of W and C, respectively, the values of D for these six meshings are given in the following table.

Meshing	Probability	$D = (W^*)^2 + (C^*)^2$
XXYY	$\frac{1}{6}$	2.4
XYXY	$\frac{1}{6}$	.6
YXXY	$\frac{1}{6}$	3.0
XYYX	$\frac{1}{6}$	3.0
YXYX	$\frac{1}{6}$	.6
YYXX	$\frac{1}{6}$	2.4

Thus, for example, the probability is $\frac{1}{3}$ under H_0 that D is equal to .6, since $D = .6$ when either of the exclusive meshings *XYXY* or *YXYX* occurs and each of these meshings has null probability $\frac{1}{6}$. Proceeding in the same manner for all possible values for D and simplifying, we obtain the null distribution.

Value of D	Probability under H_0
0.6	$\frac{1}{3}$
2.4	$\frac{1}{3}$
3.0	$\frac{1}{3}$

Thus, for example, the probability under H_0 that D is greater than or equal to 3 is, therefore, $P_0(D \geq 3) = \frac{1}{3}$, which implies that $d_{1/3} = 3$ for the setting $m = n = 2$.

Note that we have derived the null distribution of D without specifying the form of the common (under H_0) underlying X and Y populations beyond the point of requiring that they be continuous. This is why the test procedure based on D is called a *distribution-free procedure*. From the null distribution of D we can determine the critical value d_α and control the probability α of falsely rejecting H_0 when H_0 is true, and this error probability does not depend on the specific form of the common underlying distribution for the X and Y populations.

For given sample sizes m and n, the R command `cLepage(`α`,m,n)` can be used to find the available upper-tail critical values d_α for possible values of D. For a given available significance level α, the critical value d_α then corresponds to $P_0(D \geq d_\alpha) = \alpha$ and is given by `cLepage(`$\alpha, m, n) = d_\alpha$. Thus, for example, for $m = 5$ and $n = 8$, we have $P_0(D \geq 6.875) = .0194$ so that $d_{.0194} =$ `cLepage(.0194,5,8)` $= 6.875$ for $m = 5$ and $n = 8$.

29. *Equivalent Form.* In computing the Wilcoxon rank sum statistic W (4.3), we use the combined samples ranks of the Y observations. In computing the Ansari–Bradley statistic C (5.6), we use the scores assigned to the Y observations by the Ansari–Bradley outside-in scoring scheme. However, both W and C, and therefore D (5.55), can be computed solely from knowledge of the combined samples ranks of the Y observations. This follows directly from the fact that the Ansari–Bradley statistic can also be represented (in the case of no tied X and/or Y observations) as

$$C = \frac{n(N+1)}{2} - \sum_{j=1}^{n} \left| S_j - \frac{N+1}{2} \right|, \tag{5.64}$$

where, as in the calculation of W, S_j is the combined samples rank of Y_j, for $j = 1, \ldots, n$. In fact, both W and C are members of a very large class of statistics based solely on the combined samples ranks in a special way. This collection is referred to as the class of two-sample linear rank statistics, and they have been extensively studied in the literature (see, e.g., Randles and Wolfe (1979)).

Thus, although the Ansari–Bradley scoring scheme is useful in helping to motivate the statistic C as one appropriate for assessing possible scale differences in the X and Y populations, we could, in view of (5.64), just as easily have initially defined C in terms of the combined samples ranks as we

did for W (4.3). This means, of course, that D (5.55) is also a function of the X and Y observations only through their combined samples ranks.

30. *Assumptions.* We can use the Lepage test procedure in (5.58) without even requiring that the variances for the X and Y populations exist. Indeed, neither Assumptions A1 and A2 nor the location-scale parameter model in (5.2) and (5.3) specify anything about the existence of even the first moments of the X and Y populations. However, when the first two moments (and, therefore, the variance) for the underlying distributional model $H(u)$ in (5.2) exist, we see from the equal-in-distribution statement in (5.3) that

$$\text{var}\left(\frac{X}{\eta_1}\right) = \text{var}\left(\frac{Y}{\eta_2}\right),$$

which, in turn, implies that

$$\frac{\text{var}(X)}{\eta_1^2} = \frac{\text{var}(Y)}{\eta_2^2}.$$

Thus, when the variances exist, we see that $\gamma^2 = [\eta_1^2/\eta_2^2] = [\text{var}(X)/\text{var}(Y)]$.

It also follows from (5.3) that

$$E\left[\frac{X-\theta_1}{\eta_1}\right] = E\left[\frac{Y-\theta_2}{\eta_2}\right],$$

provided only that the first moment exists for $H(u)$ in (5.2). Thus, if this first moment exists, we have the relationship

$$E[Y] - \theta_2 = \frac{\eta_2}{\eta_1}\{E[X] - \theta_1\}.$$

As a result, if $\gamma^2 = \eta_1^2/\eta_2^2 = 1$, then $E[Y] - E[X] = \theta_2 - \theta_1$, corresponding to the standard interpretation of a location-only difference between two populations. This is the setting previously considered in Chapter 4 with the identification $\Delta = \theta_2 - \theta_1$. (We emphasize, however, that the existence of the first moment is not a necessary assumption for any of the statistical procedures developed in Chapter 4.)

31. *Large-Sample Approximation.* We have previously seen that both W^* (see Comment 4.6) and C^* (see Comment 9) have asymptotic standard normal distributions under H_0 (5.1) as $\min(m, n)$ becomes infinite. Moreover, it can be shown (see, e.g., Lepage (1971)) that W^* and C^* are asymptotically independent under H_0 (5.1) as $\min(m, n)$ becomes infinite. The conclusion that the statistic D (5.56) has, as $\min(m, n)$ tends to infinity, an asymptotic distribution under H_0 (5.1) that is chi-square with 2 degrees of freedom then follows from the properties (i) the square of a standard normal variable has a chi-square distribution with 1 degree of freedom and (ii) the sum of independent chi-square variables with degrees of freedom f_1 and f_2 has a chi-square distribution with $f_1 + f_2$ degrees of freedom.

32. *Exact Conditional Distribution of D with Ties.* To have a test with exact significance level even in the presence of ties among the X's and/or Y's, we need to consider all $\binom{N}{n}$ possible assignments of the N observations,

with n observations serving as Y's and m observations serving as X's. As in Comment 28, it still follows that under H_0 (5.1), each of the $\binom{N}{n}$ possible "meshings" of the X's and Y's has probability $1/\binom{N}{n}$. The only difference in the case of ties (see Ties in this section) is that we now use average scores and the appropriately modified $\text{var}_0(C)$ in the computation of C^* and average ranks and the appropriately modified $\text{var}_0(W)$ in the computation of W^* to calculate the value of D for each of these $\binom{N}{n}$ meshings leading to the tabulation of the exact conditional null distribution of D.

An example illustrating how to obtain such a conditional null distribution of D for a specific case of tied observations is not included here, because the details are much the same as those provided in Comments 4.5 and 11 for the conditional null distributions of W and C, respectively, in the presence of tied observations.

33. *More General Alternatives.* In his original discussion of the test procedure based on D, Lepage (1971) considered a slightly more general setting than that dictated by the location-scale parameter model in (5.2). In addition to Assumptions A1 and A2, he required that the X distribution function F and the Y distribution function G be related by the equation

$$G(t) = F(at + b), \quad \text{for every } t, \tag{5.65}$$

for some constants $a > 0$ and $-\infty < b < \infty$. He then considered tests of $H_0^*: [a = 1, b = 0]$ versus H_1^*: [$a \neq 1$ or $b \neq 0$ or both]. His null hypothesis H_0^* is, of course, identical to H_0 (5.1) considered in this section. However, his alternative H_1^* is more general than the location-scale parameter alternative H_1 (5.57) discussed here. The alternative H_1 (5.57) represents a slightly reduced subset of H_1^* corresponding to the identifications $a = \eta_1/\eta_2$ and $b = (\eta_2\theta_1 - \eta_1\theta_2)/\eta_2$.

34. *Consistency of the D Tests.* Let $\delta^* = [P(X < Y) - \frac{1}{2}]$ and $\Delta_\theta^* = [P(X > Y > \theta) + P(X < Y < \theta) - \frac{1}{4}]$. Under the minimal Assumptions A1 and A2 only, the test procedure (5.58) based on D is consistent if either $\delta^* \neq 0$ or $\theta_1 = \theta_2 = \theta$ and $\Delta_\theta^* \neq 0$. (See Comments 4.14 and 14.)

Under Assumptions A1, A2, and the additional general distributional relationship given by (5.65), the test procedure (5.58) based on D is consistent against any alternative for which either $a \neq 1$ or $b \neq 0$.

Properties

1. *Consistency.* For our statement we consider the more stringent location-scale parameter model described in (5.2). Then the test defined by (5.58) is consistent against alternatives for which either $\eta_1 \neq \eta_2$ or $\theta_1 \neq \theta_2$. (See also Comment 34.)
2. *Asymptotic Chi-Squareness.* See Lepage (1971) and Comment 31.
3. *Efficiency.* See Section 5.5.

Problems

25. It has long been generally accepted by medical doctors that exercise tends to stimulate the release of growth hormones in adolescents. However, little previous research had been directed toward assessment of possible effects that various medications might have on this phenomenon.

This fact led Falkner et al. (1981) to investigate whether the use of the drug clonidine to treat hypertension in adolescents has any effect on this exercise-induced release of growth hormones. Two groups of adolescents were involved in the study. The first was a control group consisting of 10 teenagers who had been diagnosed as hypertensive but were not being treated with clonidine. (Note that the "control" group considered by Falkner et al. included an additional seven nonhypertensive teenagers. In order not to possibly confound the effects of hypertension itself and the treatment clonidine on the release of the growth hormone during exercise, these seven subjects are not included in the control group presented in the problem. In addition, two subjects studied both as controls and again later after clonidine treatment are included here only in the control sample.) The second treatment group consisted of 13 hypertensive teenagers who were being treated with clonidine.

The experiment proceeded as follows. First, the basal level of growth hormone in the blood was measured for each of the subjects prior to exercising. Then each subject exercised on a treadmill until attaining a heart rate of 180–200 beats/min, at which time the blood level of growth hormone was once again obtained. The data in Table 5.5 represent these pre- and postexercise growth hormone blood levels (ng/ml) for the 23 subjects in the study.

Use an appropriate nonparametric test procedure to assess whether there are significant location or dispersion differences between the control hypertension population and the clonidine-treated population in their increases in growth hormone levels following exercise. Find the approximate P-value for the test.

26. In Example 5.3 we used a Lepage test procedure to assess whether or not the administration of the corticosteroid prednisone to pregnant women with ATP resulted in any location or dispersion changes in the platelet counts of their newborn infants. It would also be of interest to know whether there were any baseline (predelivery) differences in the platelet counts of those mothers in the study who were given the prednisone and those who served as the

Table 5.5 Growth Hormone Level (ng/ml)

	Preexercise	Postexercise
Control		
1	1.3	19.0
2	1.3	40.0
3	5.8	3.8
4	2.0	6.5
5	2.7	16.0
6	1.7	13.0
7	1.8	18.0
8	1.7	2.6
9	1.8	18.0
10	4.7	5.8
Clonidine-treated		
1	1.2	5.1
2	1.2	7.2
3	5.8	14.0
4	.3	4.0
5	3.3	25.0
6	2.2	15.0
7	4.1	10.0
8	1.2	7.6
9	6.4	10.0
10	1.8	10.0
11	1.8	8.0
12	5.2	40.0
13	1.3	21.0

Source: B. Falkner, G. Onesti, T. Moshang, Jr., and D. T. Lowenthal (1981).

Table 5.6 Maternal Platelet Counts (per mm^3)

Mothers given prednisone	Mothers not given prednisone
12,000	15,000
25,000	44,000
30,000	52,000
38,000	64,000
50,000	65,000
80,000	80,000
85,000	
126,000	
130,000	
180,000	

Source: M. Karpatkin, R. F. Porges, and S. Karpatkin (1981).

no-prednisone control group. The platelet count (per cubic millimeter) data for the mothers are given in Table 5.6.

Find the P-value for an appropriate nonparametric test procedure to assess whether are any significant location or dispersion differences in the predelivery maternal platelet counts for the control and prednisone-treated groups.

27. When there are no tied X and/or Y observations, show that the representation for C given in (5.64) in Comment 29 is indeed equivalent to the original definition of C in (5.6).

28. Generate the exact null distribution of D for the setting $m = 2$, $n = 3$. (See Comment 28.)

29. Consider the general relationship between the distribution functions for the X and Y populations prescribed in (5.65) of Comment 33. Verify that the location-scale parameter model relationship given in (5.2) corresponds to the special case of (5.65) with $a = \eta_1/\eta_2$ and $b = (\eta_2\theta_1 - \eta_1\theta_2)/\eta_2$.

30. Consider the television-viewing behavior data in Table 4.4. For these data, find the approximate P-value for an appropriate test of whether there are either location or dispersion differences in the time spent in the room after witnessing the violent behavior of those children who had previously watched the *Karate Kid* versus those children who had previously watched parts of the 1984 Summer Olympic Games. Comment on your finding in view of the results of Problems 3 and 4.5.

31. Consider the alcoholic intake data in Table 4.2. For these data, find the P-value for an appropriate test of whether there are either location or dispersion differences between the control and SST data. Discuss the result in conjunction with the previous findings in Example 4.2 and Problem 9.

32. Consider the following two-sample data for $m = 3$, $n = 3$: $X_1 = -3.7$, $X_2 = 4.6$, $X_3 = 1.5$, $Y_1 = 1.5$, $Y_2 = 4.6$, $Y_3 = 1.5$. Using the approach discussed in Comment 32, find the exact conditional null distribution of the Lepage statistic D (5.55). Compare and contrast the upper $\alpha = .10$ percentile for this exact conditional null distribution with the corresponding upper $\alpha = .10$ percentile for the null distribution of D for $m = n = 3$ and no tied observations.

5.4 A DISTRIBUTION-FREE TEST FOR GENERAL DIFFERENCES IN TWO POPULATIONS (KOLMOGOROV–SMIRNOV)

Hypothesis

Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent random samples satisfying Assumptions A1 and A2 from continuous populations with distribution functions F and G, respectively.

Under these assumptions we are interested in assessing whether there are *any* differences whatsoever between the X and Y probability distributions. Thus, we are interested in testing the null hypothesis H_0 (5.1) against the most general alternative possible, namely,

$$H_1 : [F(t) \neq G(t) \text{ for at least one } t]. \tag{5.66}$$

Procedure

To compute the two-sided two-sample Kolmogorov–Smirnov general alternative statistic J, we first need to obtain the empirical distribution functions for the X and Y samples. For every real number t, let

$$F_m(t) = \frac{\text{number of sample } X\text{'s} \leq t}{m} \tag{5.67}$$

and

$$G_n(t) = \frac{\text{number of sample } Y\text{'s} \leq t}{n}. \tag{5.68}$$

(The functions $F_m(t)$ and $G_n(t)$ are called the *empirical distribution functions* for the X and Y samples, respectively.) Let

$$d = \text{greatest common divisor of } m \text{ and } n \tag{5.69}$$

and set

$$J = \frac{mn}{d} \max_{(-\infty<t<\infty)} \{|F_m(t) - G_n(t)|\}. \tag{5.70}$$

The statistic J is the two-sided two-sample Kolmogorov–Smirnov statistic. To actually calculate J for the given X and Y samples, we use the fact that $F_m(t)$ and $G_n(t)$ are step functions changing functional values only at the observed X and Y sample observations, respectively. Thus, if we let $Z_{(1)} \leq \cdots \leq Z_{(N)}$ denote the $N = (m + n)$ ordered values for the combined sample of $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$, then we can rewrite J (5.70) in the computational form

$$J = \frac{mn}{d} \max_{i=1,\ldots,N} \{|F_m(Z_{(i)}) - G_n(Z_{(i)})|\}. \tag{5.71}$$

To test H_0 (5.1), corresponding to identical X and Y probability distributions, versus the general alternative H_1 (5.66), corresponding to *any* possible difference between the X and Y probability distributions, at the α level of significance,

$$\text{Reject } H_0 \text{ if } J \geq j_\alpha; \quad \text{otherwise do not reject,} \tag{5.72}$$

where the constant j_α is chosen to make the type I error probability equal to α. The constant j_α is the upper α percentile for the null H_0(5.1) distribution of J. Comment 38 explains how to obtain the critical value j_α for sample sizes m and n and available values of α.

Large-Sample Approximation

The large-sample approximation is based on the asymptotic distribution of J, suitably normalized, as $\min(m, n)$ tends to infinity. Set

$$J^* = \left(\frac{mn}{N}\right)^{1/2} \max_{i=1,\ldots,N} \{|F_m(Z_{(i)}) - G_n(Z_{(i)})|\} = \frac{d}{(mnN)^{1/2}} J. \tag{5.73}$$

As $\min(m, n)$ tends to infinity,

$$P_0(J^* < s) \longrightarrow \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 s^2}, 0 \quad \text{for} \quad s >, \leq 0. \tag{5.74}$$

Defining the function $Q(\mathrm{s})$ by

$$Q(s) = 1 - \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 s^2}, \quad s > 0, \tag{5.75}$$

the large-sample approximation to procedure (5.72) based on (5.74) and (5.75) is

$$\text{Reject } H_0 \text{ if } J^* \geq q_\alpha^*; \quad \text{otherwise do not reject,} \tag{5.76}$$

where q_α^* is defined by

$$Q(q_\alpha^*) = \alpha. \tag{5.77}$$

To find q_α^*, we use the `R` command `qKolSmirnLSA`(α). For example, to find $q_{.05}^*$, we apply `qKolSmirnLSA(.05)` and obtain $q_{.05}^* = 1.358$.

Ties

The empirical distribution functions $F_m(t)$ and $G_n(t)$, given by (5.67) and (5.68), respectively, are well defined in the case of ties and no adjustments are necessary in the calculation of J (5.70). (See Comment 39.) The test is then conducted using the same critical point j_α (5.72) as specified for the untied case. This approach is conservative; it yields a test with a significance level that does not exceed the nominal level α (see Hájek and Šidák (1967, p. 123), Noether (1963), and Walsh (1963)). For different methods of treating ties when using the Kolmogorov–Smirnov statistic J, see Hájek (1969, p. 134, 145).

EXAMPLE 5.4 ***Effect of Feedback on Salivation Rate.***

The effect of enabling a subject to hear himself salivate while trying to increase or decrease his salivary rate has been studied by Delse and Feather (1968). Two groups of subjects were told to attempt to increase their salivary rates upon observing a light to the left and decrease their salivary rates upon observing a light to the right. The apparatus for collecting and recording the amounts of saliva was described by Delse and Feather (1968) and also Feather and Wells (1966). Members of the feedback group received a 0.2-s, 1000-cps tone for each drop collected, whereas members of the no-feedback group did not receive any indication of their salivary rates. Table 5.7 gives differences of the form mean number of drops over 13 increase signals minus mean number of drops

Table 5.7 Mean Drop Differences

Feedback group	No-Feedback group
−.15	2.55
8.60	12.07
5.00	.46
3.71	.35
4.29	2.69
7.74	−.94
2.48	1.73
3.25	.73
−1.15	−.35
8.38	−.37

Source: F. C. Delse and B. W. Feather (1968).

over 13 decrease signals for the feedback group and the no-feedback group, each group consisting of 10 subjects.

Since both sample sizes are equal to 10, we arbitrarily choose to label the feedback group data as the X sample and the no-feedback group data as the Y sample. Thus, we have $m = n = 10$, $N = (10 + 10) = 20$, and $d = 10$. We simultaneously illustrate the calculation of the values of the empirical distribution functions $F_{10}(t)$ and $G_{10}(t)$ at the ordered combined sample values $Z_{(1)} \leq \cdots \leq Z_{(20)}$ from Table 5.7, as well as the absolute differences $|F_{10}(Z_{(i)}) - G_{10}(Z_{(i)})|$, in the following display.

i	$Z_{(i)}$	$F_{10}(Z_{(i)})$	$G_{10}(Z_{(i)})$	$\lvert F_{10}(Z_{(i)}) - G_{10}(Z_{(i)})\rvert$
1	−1.15	$\frac{1}{10}$	$\frac{0}{10}$	$\frac{1}{10}$
2	−.94	$\frac{1}{10}$	$\frac{1}{10}$	0
3	−.37	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{1}{10}$
4	−.35	$\frac{1}{10}$	$\frac{3}{10}$	$\frac{2}{10}$
5	−.15	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{1}{10}$
6	.35	$\frac{2}{10}$	$\frac{4}{10}$	$\frac{2}{10}$
7	.46	$\frac{2}{10}$	$\frac{5}{10}$	$\frac{3}{10}$
8	.73	$\frac{2}{10}$	$\frac{6}{10}$	$\frac{4}{10}$
9	1.73	$\frac{2}{10}$	$\frac{7}{10}$	$\frac{5}{10}$
10	2.48	$\frac{3}{10}$	$\frac{7}{10}$	$\frac{4}{10}$
11	2.55	$\frac{3}{10}$	$\frac{8}{10}$	$\frac{5}{10}$
12	2.69	$\frac{3}{10}$	$\frac{9}{10}$	$\frac{6}{10}$
13	3.25	$\frac{4}{10}$	$\frac{9}{10}$	$\frac{5}{10}$
14	3.71	$\frac{5}{10}$	$\frac{9}{10}$	$\frac{4}{10}$
15	4.29	$\frac{6}{10}$	$\frac{9}{10}$	$\frac{3}{10}$
16	5.00	$\frac{7}{10}$	$\frac{9}{10}$	$\frac{2}{10}$
17	7.74	$\frac{8}{10}$	$\frac{9}{10}$	$\frac{1}{10}$
18	8.38	$\frac{9}{10}$	$\frac{9}{10}$	0
19	8.60	$\frac{10}{10}$	$\frac{9}{10}$	$\frac{1}{10}$
20	12.07	$\frac{10}{10}$	$\frac{10}{10}$	0

For example, consider the evaluation of $F_{10}(Z_{(4)})$. We must count the number of X's less than or equal to $Z_{(4)} = -.35$, and divide this count by 10. From Table 5.7, we find that only one of the X values (-1.15) is less than $-.35$, none is equal to $-.35$, and thus $F_{10}(Z_{(4)}) = \frac{1}{10}$. Similarly, $G_{10}(Z_{(4)})$ is equal to {the number of Y's that are less than or equal to $-.35$}/10. From Table 5.7, we find two Y-values ($-.94$ and $-.37$) that are less than $-.35$ and one Y-value that is equal to $-.35$; thus, $G_{10}(Z_{(4)}) = \frac{3}{10}$. From this computational Table for the $|F_{10}(Z_{(i)}) - G_{10}(Z_{(i)})|$ values, we find

$$\max_{i=1,\ldots,20} \{|F_{10}(Z_{(i)}) - G_{10}(Z_{(i)})|\} = \frac{6}{10},$$

corresponding to $Z_{(12)}$. It follows from (5.71) that $J = [(10)(10)/10](6/10) = 6$.

Applying the R command `pKolSmirn(mean.drop$x,mean.drop$y)`, we find that `pKolSmirn(mean.drop$x,mean.drop$y)` $= P_0(J \geq 6) = .0524$. That is, in the notation of (5.72) with $m = n = 10$, we have $j_{.0524} = 6$. Thus, the lowest level at which we can reject H_0 (5.1) with our observed value of $J = 6$ (i.e., the P-value for the data) using procedure (5.72) is .0524, indicating some marginal evidence in the samples that feedback might have an effect on salivation rate.

To perform the large-sample approximation, we compute J^* (5.73). We find that $J^* = \{10/[10(10)(20)]^{1/2}\}(6) = 1.34$. Since `qKolSmirnLSA(.0551)` $= J^* = 1.34$, the smallest significance level at which we reject H_0, using the large-sample approximation to the Kolmogorov–Smirnov test, is approximately .0551.

Comments

35. *Motivation for the Test.* The empirical distribution functions $F_m(t)$ (5.67) and $G_n(t)$ (5.68) are estimators of the underlying distribution functions $F(t) = P\{X \leq t\}$ and $G(t) = P\{Y \leq t\}$, respectively. Thus, dJ/mn may be viewed as an estimator of $\max_{-\infty<t<\infty} |F(t) - G(t)| = \max_{-\infty<t<\infty} |P\{X \leq t\} - P\{Y \leq t\}|$, and this parameter is zero when H_0 (5.1) is true. Hence, large J values indicate a deviation from H_0 in the direction of the general alternative specified by (5.66).

36. *Equivalent Form.* In the case of no ties among the N combined $Z_{(i)}$ values, there is an alternative counting formulation for the test statistic J (5.70). Define the variables $\delta_i, i = 1, \ldots, N$, by

$$\delta_i = \begin{cases} 1, & \text{if } Z_{(i)} \text{ is an } X \text{ observation,} \\ 0, & \text{if } Z_{(i)} \text{ is a } Y \text{ observation.} \end{cases} \tag{5.78}$$

Set

$$s_j = \left[\frac{jm}{N} - \delta_1 - \cdots - \delta_j\right], \quad j = 1, \ldots, N. \tag{5.79}$$

Then the Kolmogorov–Smirnov statistic J (5.70) can also be expressed as

$$J = (N/d) \max\{|s_1|, \ldots, |s_N|\}. \tag{5.80}$$

(We note that, unlike expression (5.70), the formulation in (5.80) is not well defined in the case of ties among the $Z_{(i)}$'s.)

37. *Equal Sample Sizes.* In settings where $m = n$, the computational expression for J (5.71) can be simplified to

$$J = \max_{i=1,\ldots,N} |Q(Z_{(i)}) - S(Z_{(i)})|, \tag{5.81}$$

where, for every real number t,

$$Q(t) = mF_m(t) = [\text{number of sample } X\text{'s} \leq t] \tag{5.82}$$

and

$$S(t) = nG_n(t) = [\text{number of sample } Y\text{'s} \leq t]. \tag{5.83}$$

Thus, for the salivation data in Example 5.4, we have

$$J = |Q(Z_{(12)}) - S(Z_{(12)})| = |3 - 9| = 6,$$

in agreement with the value obtained via (5.71).

38. *Derivation of the Distribution of J under H_0 (No-Ties Case).* The null (H_0) distribution of J in the case of no ties can be obtained by using the fact that under H_0 (5.1) all possible $\binom{N}{n}$ meshings of the X's and Y's are equally likely, each having probability $1/\binom{N}{n}$. In the ensuing illustration, we derive the null distribution of J (5.70) for the sample sizes $m = 1$, $n = 3$. Here, $N = 4, d = 1$, and thus $J = 3 \max_{i=1,\ldots,4} |F_1(Z_{(i)}) - G_3(Z_{(i)})|$. We now list the $\binom{4}{1} = 4$ possible meshings, and for each of these meshings we give the associated values of $(F_1(Z_{(1)}), \ldots, F_1(Z_{(4)}))$, the associated values of $(G_3(Z_{(1)}), \ldots, G_3(Z_{(4)}))$, and finally the values of J. Thus, $P_0\{J = 2\} = \binom{2}{4} = .5$ and $P_0\{J = 3\} = .5$.

Meshings	$(F_1(Z_{(1)}), \ldots, F_1(Z_{(4)}))$	$(G_3(Z_{(1)}), \ldots, G_3(Z_{(4)}))$	J
XYYY	(1,1,1,1)	$\left(0, \frac{1}{3}, \frac{2}{3}, 1\right)$	3
YXYY	(0,1,1,1)	$\left(\frac{1}{3}, \frac{1}{3}, \frac{2}{3}, 1\right)$	2
YYXY	(0,0,1,1)	$\left(\frac{1}{3}, \frac{2}{3}, \frac{2}{3}, 1\right)$	2
YYYX	(0,0,0,1)	$\left(\frac{1}{3}, \frac{2}{3}, 1, 1\right)$	3

For given sample sizes m and n, the R command `cKolSmirn`(α, m, n) can be used to find the available upper-tail critical values j_α for possible values of J. For a given available significance level α, the critical value j_α then corresponds to $P_0(J \geq j_\alpha) = \alpha$ and is given by `cKolSmirn`$(\alpha, m, n) = j_\alpha$. Thus, for example, for $m = 4$ and $n = 6$, we have `cKolSmirn(.04762,4,6)` $= 10$ so that $P_0(J \geq 10) = .04762$ and $j_{.04762} = 10$ for $m = 4$ and $n = 6$.

39. *Ties.* To illustrate how the computational formula for J given in expression (5.71) is well defined in the case of ties, we consider the following artificial set of tied data: $X_1 = 3$, $X_2 = 3$, $X_3 = 5$, $X_4 = 7$, $X_5 = 9$ and $Y_1 = 3$, $Y_2 = 4$, $Y_3 = 4$, $Y_4 = 6$, $Y_5 = 7$, $Y_6 = 8$, $Y_7 = 10$, $Y_8 = 10$, $Y_9 = 11$, $Y_{10} = 12$. Here we

have $m = 5$, $n = 10$, $N = 15$, and $d = 5$. Following the tabular approach of Example 5.4 for computation of J, we obtain

i	$Z_{(i)}$	$F_5(Z_{(i)})$	$G_{10}(Z_{(i)})$	$\lvert F_5(Z_{(i)}) - G_{10}(Z_{(i)})\rvert$
1	3	$\frac{2}{5}$	$\frac{1}{10}$	$\frac{3}{10}$
2	3	$\frac{2}{5}$	$\frac{1}{10}$	$\frac{3}{10}$
3	3	$\frac{2}{5}$	$\frac{1}{10}$	$\frac{3}{10}$
4	4	$\frac{2}{5}$	$\frac{3}{10}$	$\frac{1}{10}$
5	4	$\frac{2}{5}$	$\frac{3}{10}$	$\frac{1}{10}$
6	5	$\frac{3}{5}$	$\frac{3}{10}$	$\frac{3}{10}$
7	6	$\frac{3}{5}$	$\frac{4}{10}$	$\frac{3}{10}$
8	7	$\frac{4}{5}$	$\frac{5}{10}$	$\frac{3}{10}$
9	7	$\frac{4}{5}$	$\frac{5}{10}$	$\frac{3}{10}$
10	8	$\frac{4}{5}$	$\frac{6}{10}$	$\frac{2}{10}$
11	9	$\frac{5}{5}$	$\frac{6}{10}$	$\frac{4}{10}$
12	10	$\frac{5}{5}$	$\frac{8}{10}$	$\frac{2}{10}$
13	10	$\frac{5}{5}$	$\frac{8}{10}$	$\frac{2}{10}$
14	11	$\frac{5}{5}$	$\frac{9}{10}$	$\frac{1}{10}$
15	12	$\frac{5}{5}$	$\frac{10}{10}$	0

Thus, the empirical distribution function $F_5(t)$ for the X sample jumps from 0 to $\frac{2}{5}$ at $Z_{(1)} = Z_{(2)} = Z_{(3)} = 3$, since two of the five X values are 3's. Similarly, the empirical distribution function $G_{10}(t)$ for the Y sample jumps from $\frac{1}{10}$ to $\frac{3}{10}$ and $\frac{6}{10}$ to $\frac{8}{10}$ at $Z_{(4)} = Z_{(5)} = 4$ and $Z_{(12)} = Z_{(13)} = 10$, respectively, since there are two 4's and two 10's among the Y observations. For these tied data, we find

$$\max_{i=1,\ldots,15} |F_5(Z_{(i)}) - G_{10}(Z_{(i)})| = |F_5(Z_{(11)}) - G_{10}(Z_{(11)})| = \tfrac{4}{10}.$$

From (5.71) it then follows (as in the case of no ties) that $J = \left[\frac{5(10)}{5}\right]\left(\frac{4}{10}\right) = 4$.

40. *Exact Conditional Distribution of J with Ties.* To have a test with exact significance level even in the presence of ties among the X's and/or Y's, we need to consider all $\binom{N}{n}$ possible assignments of the N observations with n observations serving as Y's and m observations serving as X's. As in Comment 38, it still follows that, under H_0 (5.1), each of the $\binom{N}{n}$ possible meshings of the X's and Y's has probability $1/\binom{N}{n}$. The only difference in the case of ties is that now in the computation of J for each of these $\binom{N}{n}$ meshings, the jumps in the X and Y empirical distribution functions can occur at common observations and the sizes of these jumps can be greater than $1/m$ or $1/n$, respectively. We illustrate this construction for the following $m = 2, n = 3$ data: $X_1 = 3.2, X_2 = 6.3, Y_1 = 1.9, Y_2 = 1.9, Y_3 = 6.3$. The associated

ordered $Z_{(i)}$ values are $Z_{(1)} = Z_{(2)} = 1.9 < Z_{(3)} = 3.2 < Z_{(4)} = Z_{(5)} = 6.3$ and the corresponding value of J (5.71) is $\frac{2(3)}{1}|F_2(Z_{(1)}) - G_3(Z_{(1)})| = 6|F_2(Z_{(2)}) - G_3(Z_{(2)})| = 6|F_2(1.9) - G_3(1.9)| = 6|0 - \frac{2}{3}| = 4$. To assess the significance of this value of J, we obtain its conditional distribution by considering the $\binom{5}{3} = 10$ possible assignments of the observations 1.9, 1.9, 3.2, 6.3, and 6.3 to serve as two X observations and three Y observations. These 10 assignments and the corresponding values of J are

X observations	Y observations	Probability under H_0	Value of J
1.9, 1.9	3.2, 6.3, 6.3	$\frac{1}{10}$	6
1.9, 3.2	1.9, 6.3, 6.3	$\frac{1}{10}$	4
1.9, 3.2	1.9, 6.3, 6.3	$\frac{1}{10}$	4
1.9, 6.3	1.9, 3.2, 6.3	$\frac{1}{10}$	1
1.9, 6.3	1.9, 3.2, 6.3	$\frac{1}{10}$	1
1.9, 6.3	1.9, 3.2, 6.3	$\frac{1}{10}$	1
1.9, 6.3	1.9, 3.2, 6.3	$\frac{1}{10}$	1
3.2, 6.3	1.9, 1.9, 6.3	$\frac{1}{10}$	4
3.2, 6.3	1.9, 1.9, 6.3	$\frac{1}{10}$	4
6.3, 6.3	1.9, 1.9, 3.2	$\frac{1}{10}$	6

This yields the null tail probabilities

$$P_0(J \geq 6) = \tfrac{2}{10}, \quad P_0(J \geq 4) = \tfrac{6}{10}, \quad P_0(J \geq 1) = 1.$$

This distribution is called the *conditional null distribution* or the *permutation null distribution* of J, given the set of tied observations {1.9, 1.9, 3.2, 6.3, 6.3}. For the particular observed value $J = 4$, we have that $P_0(J \geq 4) = \frac{6}{10}$. (Note that the particular observed X and Y sample values are not important to the calculation of this conditional null distribution of J. It is critical only that the two smallest observations are tied in value, the middle ordered value is untied, and the two largest observations are tied. Thus, for example, the two sets of sample observations $\{X_1 = 3.2,\ X_2 = 6.3,\ Y_1 = 1.9,\ Y_2 = 1.9,\ Y_3 = 6.3\}$ and $\{X_1 = -12.1, X_2 = 13.7, Y_1 = -12.1, Y_2 = 0, Y_3 = 13.7\}$ yield the same exact conditional null distribution of J.)

41. *Large-Sample Approximation.* Smirnov (1939) derived the asymptotic (min (m, n) tending to infinity) distribution of the standardized Kolmogorov–Smirnov statistic J^* (5.73) using the work of Kolmogorov (1933) on the asymptotic (m tending to infinity) distribution of the one-sample statistic

$$J_0 = \sqrt{m} \max_{-\infty < a < \infty} |F_m(a) - F_0(a)|, \tag{5.84}$$

where $F_m(\cdot)$ is the empirical distribution function for a random sample of size m from the (assumed) continuous distribution with distribution function $F(a) =$

$P(X \leq a)$ and $F_0(a)$ is a completely specified distribution function. The statistic J_0 can be used to test the goodness-of-fit hypothesis that the random sample $X_1, \ldots, X_m$ has been drawn from a population with distribution function F_0, namely,

$$H_0' : [P(X \leq a) = F_0(a) \text{ for all } -\infty < a < \infty], \tag{5.85}$$

versus the broad alternative that the population from which the sample was drawn does not have distribution function F_0.

42. *Test Based on the One-Sample Limit of the Wilcoxon Rank Sum Statistic.* It is of interest to note that the two-sample Wilcoxon test discussed in Section 4.1 can be reduced to a test of H_0' (5.85) by allowing one of the sample sizes, say n, to become infinite. Moses (1964) showed how this leads to a test based on $W_0 = \sum_{j=1}^{m} F_0(X_j)$. (The normal approximation to W_0 treats $[W_0 - (m/2)]/(m/12)^{1/2}$ as an approximate $N(0, 1)$ random variable under H_0'.) Moses pointed out that a test based on W_0 is particularly convenient when F_0 is known but is specified by tabular data, such as demographic data on age of death distributions, rather than being given by a mathematical expression.
43. *Consistency of the J Tests.* Define the class $\mathcal{C}$ of pairs of distribution functions F and G by

$$\mathcal{C} = \{(F, G) : F(x) \neq G(x) \text{ for at least one } x\}. \tag{5.86}$$

Under the minimal Assumptions A1 and A2 only, the test procedure (5.72) is consistent for any $(F, G) \in \mathcal{C}$; that is, the test is consistent against *any* differences between the F and G distributions (i.e., *whenever* H_0 (5.1) is false). In gaining this extra protection against all differences, we do, however, sacrifice power against specific subclasses of alternatives (such as location shifts or differences in dispersions).

Properties

1. *Consistency.* See Comment 43.
2. *Asymptotic Distribution*. See Smirnov (1939) and Comment 41.
3. *Efficiency.* See Capon (1965), Ramachandramurty (1966b), Yu (1971), and Section 5.5.

Problems

33. The data in Table 5.8 are a subset of the data obtained by Friedman et al. (1971) in an experiment comparing the average concentrations of human plasma growth hormone both resting and after arginine hydrochloride infusion in relatively coronary-prone subjects (persons with type A behavior patterns) with the corresponding concentrations of relatively coronary-resistant individuals (subjects with type B behavior patterns). Type A behavior is characterized by an excessive sense of time urgency, drive, and competitiveness; type B denotes a converse type of behavior. Earlier studies (cf. Friedman and Rosenman (1959)) indicated that type A individuals may be more prone to coronary heart disease than type B individuals.

Table 5.8 Peak Levels of Human Plasma Growth Hormone after Arginine Hydrochloride Infusion (Initial Test, ng/ml)

Type A subjects	Type B subjects
3.6	16.2
2.6	17.4
4.7	8.5
8.0	15.6
3.1	5.4
8.8	9.8
4.6	14.9
5.8	16.6
4.0	15.9
4.6	5.3
	10.5

Source: M. Friedman, S. O. Byers, R. H. Rosenman, and R. Neuman (1971).

Find the P-value for an appropriate test of whether there is any difference between the probability distribution of peak level human plasma growth hormone (after arginine hydrochloride infusion) for type A subjects and that for type B subjects.

34. Consider the alcoholic intake data in Table 4.2. For these data, find the P-value for an appropriate test of whether there are *any* differences between the control and SST probability distributions. Discuss this result in conjunction with the previous findings in Example 4.2, Problem 9, and Problem 31.

35. Verify directly, or illustrate with a numerical example, that representations (5.71) and (5.80) for J are indeed equivalent.

36. When $m = n$, show that both representations (5.71) and (5.80) for J are equivalent to the expression

$$J = \max\{|t_1|, |t_2|, \ldots, |t_N|\}, \tag{5.87}$$

where

$$t_j = (1 - 2\delta_1) + (1 - 2\delta_2) + \cdots + (1 - 2\delta_j), \tag{5.88}$$

and the δ's are given by (5.78).

37. Calculate the value of J for the salivation data in Table 5.7 using the equivalent (when $m = n$) expression in (5.87).

38. Apply the two-sided Wilcoxon rank sum test procedure from Section 4.1 to the salivation data in Table 5.7 by finding the appropriate P-value. Compare the conclusion indicated by this Wilcoxon rank sum procedure with that indicated by the Kolmogorov–Smirnov procedure in Example 5.4. Comment on your findings.

39. Generate the exact null distribution of J (5.70) for the setting $m = 3$, $n = 3$. (See Comment 38.)

40. Consider the growth hormone level data found in Table 5.5. Use the Kolmogorov–Smirnov test procedure to assess whether there are significant differences of *any* kind between the control hypertension population and the clonidine-treated population in their increases in growth hormone levels following exercise. Find the appropriate P-value for the test and compare it with the P-value obtained in Problem 25.

41. Consider the following two-sample data for $m = 3$, $n = 3$: $X_1 = -3.7$, $X_2 = 4.6$, $X_3 = 1.5$, $Y_1 = 1.5$, $Y_2 = 4.6$, $Y_3 = 1.5$. Using the approach discussed in Comment 40, find the exact conditional null distribution of the Kolmogorov–Smirnov statistic J (5.70). Compare and

contrast this exact conditional null distribution with the corresponding null distribution of J for $m = n = 3$ and no tied observations, as obtained in Problem 39.

42. Consider the serum iron data in Table 5.1. Use the Kolmogorov–Smirnov test procedure to assess whether there are significant differences of *any* kind between the distribution of serum iron values obtained by the Ramsay method and the distribution of serum iron values obtained by the Jung–Parekh method. Find the P-value for the test and compare it with the results discussed in Example 5.1.

5.5 EFFICIENCIES OF TWO-SAMPLE DISPERSION AND BROAD ALTERNATIVES PROCEDURES

Recall the classical normal theory $\mathcal{F}$-test for equality of variances based on the statistic

$$D = \frac{S_x^2}{S_y^2}, \tag{5.89}$$

where $S_x^2 = \sum_{i=1}^m (X_i - \bar{X})^2/(m-1), S_y^2 = \sum_{j=1}^n (Y_j - \bar{Y})^2/(n-1), \bar{X} = \sum_{i=1}^m X_i/m$, and $\bar{Y} = \sum_{j=1}^n Y_j/n$. The significance level of this $\mathcal{F}$-test is extremely sensitive to nonnormality. (See Comment 26.) This is also true of the coverage probability of the confidence intervals for σ_2^2/σ_1^2 that are based on the ratio of sample variances and derived from the $\mathcal{F}$-test. The Box–Andersen (1955) test "adjusts" the $\mathcal{F}$-test to remedy this difficulty. Since this Box–Andersen approach has desirable properties, we report asymptotic efficiencies of the test procedures of Sections 5.1 and 5.2, as well as the point estimators and confidence intervals/bounds associated with the jackknife approach (see Comment 24), with respect to the corresponding Box–Andersen procedures. (The specific Box–Andersen procedures that we refer to are (a) the APF test of Shorack (1969), which is a slight variation of the test used by Box and Andersen for the case where the parameters θ_1 and θ_2 of model (5.2) are known; (b) an associated estimator given by Shorack (1965); and (c) the associated confidence interval and bounds discussed in Shorack (1969).)

The Pitman asymptotic relative efficiency for scale alternatives of the Ansari–Bradley test based on C (5.6) relative to the Box–Andersen adjusted $\mathcal{F}$-test based on D (5.89) is

$$e(C, D) = 12(\beta_G - 1)\left[\int_{-\infty}^{0} x g^2(x)dx - \int_{0}^{\infty} x g^2(x)dx\right]^2, \tag{5.90}$$

where

$$\beta_G = \frac{\int_{-\infty}^{\infty}(x-\mu)^4 g(x)dx}{\left\{\int_{-\infty}^{\infty}(x-\mu)^2 g(x)dx\right\}^2}$$

is the kurtosis and $\mu = \int_{-\infty}^{\infty} x g(x)dx$ is the mean of the population with distribution function $G(\cdot)$ and probability density function $g(\cdot)$.

The expression in (5.90) was obtained by Ansari and Bradley (1960). Some values of $e(C, D)$ for selected $G(\cdot)$ are

G	Normal	Uniform	Double exponential
$e(C, D)$	.61	.60	.94

Miller (1968) pointed out that the asymptotic relative efficiency of the jackknife procedures (tests, point estimators, and confidence intervals/bounds) with respect to the Box–Andersen procedures has the value 1 for *any* underlying distribution $F(\cdot)$; that is, $e(Q, D) \equiv 1$, where Q is given by (5.35) and D represents the Box–Andersen adjusted $\mathcal{F}$-test procedures.

We do not know of any results for the asymptotic efficiencies of the Lepage test for location or scale differences (Section 5.3).

The determination of asymptotic relative efficiencies for the Kolmogorov–Smirnov test based on J (5.70) is difficult, owing to the complicated form of the asymptotic distribution of the Kolmogorov–Smirnov statistic. Capon (1965) obtained lower bounds for the asymptotic relative efficiency of the Kolmogorov–Smirnov test. In particular, for normal translation alternatives, Capon derived the lower bound of .637 for the asymptotic relative efficiency of the Kolmogorov–Smirnov test with respect to the normal theory two-sample t test (see Section 4.5 and also Ramachandramurty (1966b) and Yu (1971)). For related efficiency results using different notions of asymptotic efficiency, see Klotz (1967), Hájek and Šidák (1967, p. 272), and Anděl (1967).