Chapter 4

The Two-Sample Location Problem

INTRODUCTION

In this chapter the data consist of two random samples, a sample from the control population and an independent sample from the treatment population. On the basis of these samples, we wish to investigate the presence of a treatment effect that results in a shift of location. The basic hypothesis is that of no treatment effect; that is, the samples can be thought of as a single sample from one population.

Section 4.1 presents a distribution-free rank sum test for the hypothesis of no treatment effect; Section 4.2, a point estimator associated with the rank sum statistic; and Section 4.3, a related distribution-free confidence interval that emanates from the rank sum test. The basic model for Sections 4.1, 4.2 and 4.3 assumes the populations differ only by a location shift. In Section 4.4 we present a test for location differences that allows the population dispersions to differ. Section 4.5 considers the asymptotic relative efficiencies for translation alternatives of the procedures based on the rank sum statistic with respect to their normal theory counterparts based on sample means.

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

A3. Populations 1 and 2 are continuous populations.

4.1 A DISTRIBUTION-FREE RANK SUM TEST (WILCOXON, MANN AND WHITNEY)

Hypothesis

Let F be the distribution function corresponding to population 1 and let G be the distribution function corresponding to population 2.

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

The null hypothesis is

$$H_0 : F(t) = G(t), \quad \text{for every } t. \tag{4.1}$$

The null hypothesis asserts that the X variable and the Y variable have the same probability distribution, but the common distribution is not specified.

The alternative hypothesis in a two-sample location problem typically specifies that Y tends to be larger (or smaller) than X. One model that is useful to describe such alternatives is the translation model—also called the *location-shift model*. The location-shift model is

$$G(t) = F(t - \Delta), \quad \text{for every } t. \tag{4.2}$$

Model (4.2) says that population 2 is the same as population 1 except it is shifted by the amount Δ. Another way of writing this is

$$Y \stackrel{d}{=} X + \Delta$$

where the symbol $\stackrel{d}{=}$ means "has the same distribution as." The parameter Δ is called the *location shift*. It is also known as the *treatment effect*. If X is a randomly selected value from population 1, the control population, and Y is a randomly selected value from population 2, the treatment population, then Δ is the expected effect due to the treatment. If Δ is positive, it is the expected increase due to the treatment, and if Δ is negative, it is the expected decrease due to the treatment. If the mean $E(X)$ of population 1 exists, then letting $E(Y)$ denote the mean of population 2,

$$\Delta = E(Y) - E(X),$$

the difference in population means. In terms of the location-shift model, the null hypothesis H_0 reduces to

$$H_0 : \Delta = 0,$$

the hypothesis that asserts the population means are equal or, equivalently, that the treatment has no effect.

We note that although we find it convenient to use the "treatment" and "control" terminology, many situations will arise in which we want to compare two random samples, neither one of which can be described as a sample from a control population. The procedures of this chapter are applicable even when there are no natural control or treatment designations.

Procedure

To compute the Wilcoxon two-sample rank sum statistic W, order the combined sample of $N = m + n$ X-values and Y-values from least to greatest. Let S_1 denote the rank of $Y_1, \ldots, S_n$ denote the rank of Y_n in this joint ordering. W is the sum of the ranks assigned to the Y-values. That is,

$$W = \sum_{j=1}^{n} S_j. \tag{4.3}$$

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

$$H_0 : \Delta = 0$$

versus

$$H_1 : \Delta > 0$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } W \geq w_\alpha; \quad \text{otherwise do not reject,} \tag{4.4}$$

where the constant w_α is chosen to make the type I error probability equal to α. Values of w_α can be obtained from the R functions `pwilcox` and `qwilcox` as illustrated in Example 4.1 and Comment 3.

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

$$H_0 : \Delta = 0$$

versus

$$H_2 : \Delta < 0$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } W \leq n(m+n+1) - w_\alpha; \quad \text{otherwise do not reject.} \tag{4.5}$$

c. *Two-Sided Test*. To test

$$H_0 : \Delta = 0$$

versus

$$H_3 : \Delta \neq 0$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } W \geq w_{\alpha/2} \text{ or if } W \leq n(m+n+1) - w_{\alpha/2}; \quad \text{otherwise do not reject.} \tag{4.6}$$

The two-sided procedure given by (4.6) is the two-sided symmetric test with the $\alpha/2$ probability in each tail of the distribution.

Large-Sample Approximation

The large-sample approximation is based on the asymptotic normality of W, suitably standardized. We first need to know the mean and variance of W when the null hypothesis is true. When H_0 is true, the mean and variance of W are, respectively,

$$E_0(W) = \frac{n(m+n+1)}{2} \tag{4.7}$$

$$\text{var}_0(W) = \frac{mn(m+n+1)}{12}. \tag{4.8}$$

Comment 4 gives direct calculations of $E_0(W)$ and $\text{var}_0(W)$ in the special case where $m = 3$, $n = 2$. Comment 6 gives general derivations.

The standardized version of W is

$$W^* = \frac{W - E_0(W)}{\{\text{var}_0(W)\}^{1/2}} = \frac{W - \{n(m+n+1)/2\}}{\{mn(m+n+1)/12\}^{1/2}}. \tag{4.9}$$

When H_0 is true, W^* has, as $\min(m,n)$ tends to infinity, an asymptotic $N(0,1)$ distribution.

The normal theory approximation to procedure (4.4) is

$$\text{Reject } H_0 \text{ if } W^* \geq z_\alpha; \quad \text{otherwise do not reject.} \tag{4.10}$$

The normal theory approximation to procedure (4.5) is

$$\text{Reject } H_0 \text{ if } W^* \leq -z_\alpha; \quad \text{otherwise do not reject.} \tag{4.11}$$

The normal theory approximation to procedure (4.6) is

$$\text{Reject } H_0 \text{ if } |W^*| \geq z_{\alpha/2}; \quad \text{otherwise do not reject.} \tag{4.12}$$

Ties

If there are ties, give tied observations the average of the ranks for which those observations are competing. After computing W using average ranks, use procedure (4.4), (4.5), or (4.6). Now, however, the test is approximate rather than exact. (To get an exact test, even in the tied case, see Comment 5.)

When applying the large-sample approximation, the following modification should be made. What there are ties, the null mean of W is unaffected, but the null variance is reduced to

$$\text{var}_0(W) = \frac{mn}{12}\left[m+n+1 - \frac{\sum_{j=1}^{g}(t_j-1)t_j(t_j+1)}{(m+n)(m+n-1)}\right], \tag{4.13}$$

or, equivalently,

$$\text{var}_0(W) = \frac{mn(N+1)}{12} - \left\{\frac{mn}{12N(N-1)} \cdot \sum_{j=1}^{g}(t_j-1)t_j(t_j+1)\right\}. \tag{4.14}$$

In displays (4.13) and (4.14) g denotes the number of tied groups and t_j is the size of tied group j. Furthermore, an untied observation is considered to be a tied "group" of size 1. In particular, if there are no tied observations, $g = N$, $t_j = 1$ for $j = 1, \ldots, N$, and thus each term of the form $(t_j - 1)(t_j)(t_j + 1)$ reduces to 0 and $\text{var}_0(W)$ reduces to $mn(m+n+1)/12$, the null variance of W when there are no ties. Note also that the term in curly braces on the right-hand side of display (4.14) measures the reductions in the null variance due to the presence of ties.

To apply the large-sample approximation when ties are present, compute W using average ranks and compute

$$W^* = \frac{W - [n(m+n+1)/2]}{\{\text{var}_0(W)\}^{1/2}},$$

where $\text{var}_0(W)$ is given by display (4.13). With this modified value of W^*, approximations (4.10), (4.11), and (4.12) can be applied.

The Mann–Whitney Statistic

Procedures (4.4), (4.5), and (4.6) based on the rank sum statistic can also be performed using the Mann–Whitney statistic. Let

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} \phi(X_i, Y_j), \tag{4.15}$$

where

$$\phi(X_i, Y_j) = \begin{cases} 1 & \text{if } X_i < Y_j, \\ 0 & \text{otherwise.} \end{cases}$$

The statistic U counts the number of "X before Y" predecessors. It is easy to show (see Comment 7) that

$$W = U + \frac{n(n+1)}{2}. \tag{4.16}$$

Thus tests based on W and U are equivalent. For example, the one-sided test given by (4.4) that rejects if $W \geq w_\alpha$ is equivalent to the one-sided test that rejects if $U \geq u_\alpha$ where u_α is the upper α percentile point of the null distribution of U. From (4.16) it follows that $w_\alpha = u_\alpha + (n(n+1)/2)$. Some textbooks and some software find it more convenient to use U rather than W. For example, the R functions `wilcox.test`, `pwilcox`, and `qwilcox`, illustrated in Comment 3 and Example 4.1, utilize

$$U' = U - mn, \tag{4.17}$$

the number of Y before X predecessors. The possible values of U and U' are 0, 1, . . . , mn. Furthermore, when H_0 is true, the mean and variance of U and U' are, respectively,

$$E_0(U) = E_0(U') = mn/2 \tag{4.18}$$

$$\text{Var}_0(U) = \text{Var}_0(U') = mn(m+n+1)/12. \tag{4.19}$$

The null distributions of U and U' are symmetric about the mean $mn/2$.

EXAMPLE 4.1 ***Water Transfer in Placental Membrane.***

The data in Table 4.1 are a portion of the data obtained by Lloyd et al. (1969). Among other things, these authors investigated whether there is a difference in the transfer of tritiated water (water containing tritium, a radioactive isotope of hydrogen) across the tissue layers in the term human chorioamnion (a placental membrane) and in the human chorioamnion between 3- and 6-months' gestational age. The objective measure used was the permeability constant *Pd* of the human chorioamnion to water. The tissues used for the study were obtained within 5 min of delivery from the placentas of healthy, uncomplicated pregnancies in the following two gestational age categories: (a) between 12 and 26 weeks following termination of pregnancy via abdominal hysterotomy (surgical incision of the uterus) for psychiatric indications and (b) term, uncomplicated vaginal deliveries. Tissues from 10 term pregnancies and five terminated pregnancies were used in the experiment. Table 4.1 gives the average permeability constant (in units of 10^{-4} cm/s) for six measurements on each of the 15 tissues in the study.

Table 4.1 Tritiated Water Diffusion Across Human Chorioamnion

$Pd(10^{-4}$ cm/s)	
At term	12–26 Weeks gestational age
0.80	1.15
0.83	0.88
1.89	0.90
1.04	0.74
1.45	1.21
1.38	
1.91	
1.64	
0.73	
1.46	

Source: S.J. Lloyd, K.D. Garlid, R.C. Reba and A.E. Seeds (1969).

In this example, the alternative of interest is greater permeability of the human chorioamnion for the term pregnancy. Thus, if we let X correspond to the Pd values of tissues from term pregnancies and Y to the Pd values of tissues from terminated pregnancies, we perform a one-sided test designed to detect the alternative $\Delta < 0$.

We list the combined sample in increasing order to facilitate the joint ranking. The ranks are given in parentheses

X	Y	X	X	Y	Y	X	Y
0.73	0.74	0.80	0.83	0.88	0.90	1.04	1.15
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Y	X	X	X	X	X	X	
1.21	1.38	1.45	1.46	1.64	1.89	1.91	
(9)	(10)	(11)	(12)	(13)	(14)	(15)	

We see that the Y-ranks are 2, 5, 6, 8, and 9 and thus

$$W = 2 + 5 + 6 + 8 + 9 = 30.$$

From (4.16) we find

$$U = W - n(n+1)/2 = 30 - 15 = 15.$$

The R function `wilcox.test` computes the value of $U' = U - mn$ and gives the P-value corresponding to U'. In the R output, U' is denoted by W, which is not to be confused with our use of W for the sum of the Y-ranks. Since $U = 15$, $U' = 50 - 15 = 35$, and that is the value (labeled W) provided by `wilcox.test`. If you let

```
at.term<−c (.80, .83, 1.89, 1.04, 1.45, 1.38, 1.91, 1.64, .73, 1.46),
gest.age<−c (1.15, .88, .90, .74, 1.21)
```

and perform `wilcox.test (at.term, gest.age, alternative="t", conf. int=T)` you get the two-sided P-value .2544 and the one-sided P-value for the test of $\Delta < 0$ is .127. This one-sided P-value is also obtained using `wilcox.test (x, y, alt="g")`.

The function `wilcox.test` not only performs the test but also provides the Hodges–Lehmann estimator of Section 4.2 and the confidence interval of Section 4.3.

There is no need to perform the large-sample approximation because we have the result for the exact test. Nevertheless, it is informative to see how close the P-value given by the large-sample approximation is to the exact P-value. From (4.9) we find $W^* = -1.225$ and the R function `pnorm` gives `pnorm` $(-1.225) = .110$. Thus, the one-sided P-value based on the large-sample approximation is .110 compared to the exact one-sided P-value of .127.

Both the exact test and the large-sample approximation indicate that there is no sufficiently strong evidence to support the hypothesis that human chorioamnion is more permeable to water transfer at term than at 12–26 weeks' gestational age.

EXAMPLE 4.2 ***Alcohol Intakes.***

Eriksen, Björnstad, and Götestam (1986) studied a social skills training program for alcoholics. Twenty-four "alcohol-dependent" male inpatients at an alcohol treatment center were randomly assigned to two groups. The control group patients were given a traditional treatment program. The treatment group patients were given the traditional treatment program plus a class in social skills training (SST). After being discharged from the program, each patient reported—in 2-week intervals—the quantity of alcohol consumed, the number of days prior to his first drink, the number of sober days, the days worked, the times admitted to an institution, and the nights slept at home. Reports were verified by other sources (wives or family members). (Such data can be unreliable!) One patient in the SST group, discovered to be an opiate addict, disappeared after discharge and submitted no reports. The remaining 23 patients reported faithfully for a year. The results for alcohol intake are given in Table 4.2. The ranks in the joint ranking of the 23 observations are given in parentheses in Table 4.2 and we find that the sum of the SST ranks is $W = 81$.

To test H_0 versus the alternative that the SST group tends to have lower alcohol intakes, we need to test $H_0 : \Delta = 0$ versus $H_2 : \Delta < 0$. We will use the R function `wilcox.test`. Let

```
x<−c(1042, 1617, 1180, 973, 1552, 1251, 1151, 1511, 728, 1079, 951, 1319)
y<−c(874, 389, 612, 798, 1152, 893, 541, 741, 1064, 862, 213)
```

Table 4.2 Alcohol Intake for 1 Year (Centiliter of Pure Alcohol)

Control		SST	
1042	(13)	874	(9)
1617	(23)	389	(2)
1180	(18)	612	(4)
973	(12)	798	(7)
1552	(22)	1152	(17)
1251	(19)	893	(10)
1151	(16)	541	(3)
1511	(21)	741	(6)
728	(5)	1064	(14)
1079	(15)	862	(8)
951	(11)	213	(1)
1319	(20)		

Source: L. Eriksen, S. Björnstad, and K. G. Götestam (1986).

The alternative that the SST groups tend to have lower alcohol intakes would be reflected in small W values or, in terms of the function `wilcox.test`, which uses $U' = U\text{-}mn$, large values of U. Thus we use `wilcox.test (x, y, alt="g")`. This yields a one-sided P-value of .00049. Thus there is strong evidence that the SST class in combination with the traditional treatment program tends to lower alcohol intake in alcoholics.

Comments

1. *Motivation for the Test*. When Δ is greater than 0, the Y-values will tend to be larger than the X-values, and thus the Y-ranks will tend to be larger than the X-ranks. Hence the value of W will tend to be large. This suggests rejecting H_0 in favor of $\Delta > 0$ for large values of W and motivates procedure (4.4). An analogous motivation leads to procedure (4.5).

   The test based on W was introduced by Wilcoxon in 1945. An equivalent test based on the number of X before Y occurrences in the jointly ordered sample (see Comment 7) was proposed by Mann and Whitney (1947). Kruskal (1957) gives a detailed history of the Wilcoxon statistic dating back to 1914.

2. *Testing Δ is Equal to Some Specified Nonzero Value*. Procedures (4.4), (4.5), and (4.6) and the corresponding large-sample approximations given by procedures (4.10), (4.11), and (4.12) are for testing if Δ is equal to zero. To test $\Delta = \Delta_0$, where Δ_0 is some specified nonzero number, subtract Δ_0 from each Y-value to form a pseudosample, namely, $Y_1' = Y_1 - \Delta_0$, $Y_2' = Y_2 - \Delta_0, \ldots, Y_n' = Y_n - \Delta_0$. Then compute W as the sum of the Y'-ranks in the joint ranking of the m X-values and the n Y'-values. Then procedures (4.4), (4.5), and (4.6), and their corresponding large-sample approximations given by displays (4.10), (4.11), and (4.12), can be applied as described earlier.

3. *Derivation of the Distribution of W under H_0 (No-Ties Case)*. Assume that the underlying distribution under H_0 is continuous so that ties have probability zero of occurring. Then under H_0, all $\binom{N}{n}$ possible assignments for the Y-ranks are equally likely, each having probability $1/\binom{N}{n}$. For example, in the case of $m = 3$, $n = 2$, the $\binom{5}{2} = 10$ possible outcomes for the ranks attained by the two Y observations and the corresponding values of W are given in the following table.

Y-ranks	Probability	W
1, 2	$\frac{1}{10}$	3
1, 3	$\frac{1}{10}$	4
1, 4	$\frac{1}{10}$	5
1, 5	$\frac{1}{10}$	6
2, 3	$\frac{1}{10}$	5
2, 4	$\frac{1}{10}$	6
2, 5	$\frac{1}{10}$	7
3, 4	$\frac{1}{10}$	7
3, 5	$\frac{1}{10}$	8
4, 5	$\frac{1}{10}$	9

Thus, for example, under H_0, the probability is $\frac{2}{10}$ that W is equal to 5, because $W = 5$ when either Y-rank configuration {1, 4} or Y-rank configuration {2, 3} occurs, each has a $\frac{1}{10}$ chance of occurring (and, of course, they cannot both occur simultaneously). Simplifying, we obtain the null distribution.

Possible value of W	Probability of value
3	.1
4	.1
5	.2
6	.2
7	.2
8	.1
9	.1

Thus, for example, under H_0, the probability that W is greater than or equal to 7 is

$$P_0(W \geq 7) = P_0(W = 7) + P_0(W = 8) + P_0(W = 9)$$
$$= .2 + .1 + .1 = .4.$$

The R command `pwilcox` enumerates the null distribution of U', which is the same as the null distribution of U. The command `pwilcox(0:6,2,3, lower.tail=T)` gives the lower tail probabilities, that is, the cumulative distribution, corresponding to the six possible values of U. The output is .1, .2, .4, .6, .8, .9, and 1.0; that is,

$$P(U < 0) = .1,\ P(U < 1) = .2,\ P(U < 2) = .4,$$
$$P(U < 3) = .6,\ P(U < 4) = .8,\ P(U < 5) = .9,\ P(U < 6) = 1.0.$$

Recall that $W = U + n(n+1)/2 = U + 3$ to verify that this output agrees with results for W.

Observe that we have derived the null distribution of W (and equivalently U) without specifying the common underlying continuous distribution of the two populations. This is why the procedures based on W are called *distribution-free procedures*. From the null distribution of W, we can determine the critical values w_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 common underlying distribution.

4. *Calculation of the Mean and Variance of W under the Null Hypothesis*. In displays (4.7) and (4.8), we presented formulas for the mean and variance of W when the null hypothesis is true. In this comment, we illustrate a direct calculation of $E_0(W)$ and $\text{var}_0(W)$ in a particular case. We use the null distribution of W obtained in Comment 3. (Later, in Comment 6, we present general derivations of $E_0(W)$ and $\text{var}_0(W)$.) Comment 3 treated the case where $m = 3, n = 2$. The null mean of W, $E_0(W)$, is obtained by multiplying each possible value of W with its probability under H_0. Thus

$$E_0(W) = 3(.1) + 4(.1) + 5(.2) + 6(.2) + 7(.2) + 8(.1) + 9(.1) = 6.$$

This is in agreement with what we obtain using (4.7), namely,

$$E_0(W) = \frac{n(m+n+1)}{2} = \frac{2(3+2+1)}{2} = 6.$$

A check on the expression for $\text{var}_0(W)$ is also easily performed. Recall

$$\text{var}_0(W) = E_0(W^2) - \{E_0(W)\}^2,$$

where $E_0(W^2)$, the second moment of the distribution of W, is again obtained by multiplying possible values (in this case, values of W^2) by the corresponding probabilities under H_0. We find

$$E_0(W^2) = 9(.1) + 16(.1) + 25(.2) + 36(.2) + 49(.2) + 64(.1) + 81(.1) = 39.$$

Thus

$$\text{var}_0(W) = 39 - (6)^2 = 3.$$

This agrees with what we obtain using (4.8) directly, namely,

$$\text{var}_0(W) = \frac{3(2)(3+2+1)}{12} = 3.$$

5. *Exact Conditional Distribution of W with Ties*. To get an exact test in the presence of ties, we 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. For each such assignment, we compute a value of W. Then we see how extreme our observed value of W is in this "built-up" conditional distribution. To keep computations simple, we illustrate for the $n = 2$, $m = 3$ data

Y	X	X	Y	X
.7	1.2	1.7	1.7	2.8
(1)	(2)	(3.5)	(3.5)	(5)

Note that the two tied 1.7 values get the average ranks 3.5. We then find that W, the sum of the Y-ranks, is

$$W = 1 + 3.5 = 4.5.$$

To assess the significance of W, we obtain a conditional distribution by considering the $\binom{5}{2} = 10$ possible assignments of the observations

$$.7, \quad 1.2, \quad 1.7, \quad 1.7, \quad 2.8,$$

to serve as three X-values and two Y-values, or, equivalently, the 10 possible assignments of the ranks

$$1, \quad 2, \quad 3.5, \quad 3.5, \quad 5,$$

to serve as three X-ranks and two Y-ranks. These 10 assignments and the corresponding values of W are shown in the following table.

Y-ranks	Probability	W
1, 2	$\frac{1}{10}$	3
1, 3.5	$\frac{1}{10}$	4.5
1, 3.5	$\frac{1}{10}$	4.5
1, 5	$\frac{1}{10}$	6
2, 3.5	$\frac{1}{10}$	5.5
2, 3.5	$\frac{1}{10}$	5.5
2, 5	$\frac{1}{10}$	7
3.5, 3.5	$\frac{1}{10}$	7
3.5, 5	$\frac{1}{10}$	8.5
3.5, 5	$\frac{1}{10}$	8.5

Then, for the tail probabilities, we obtain

$$P_0(W \geq 8.5) = \tfrac{2}{10},$$
$$P_0(W \geq 7) = \tfrac{4}{10},$$
$$P_0(W \geq 6) = \tfrac{5}{10},$$
$$P_0(W \geq 5.5) = \tfrac{7}{10},$$
$$P_0(W \geq 4.5) = \tfrac{9}{10},$$
$$P_0(W \geq 3) = 1.$$

This distribution is called the *conditional distribution* or the permutation distribution of W. For the particular observed value $W = 4.5$, we see $P_0(W \leq 4.5) = 1 - P_0(W \geq 5.5) = \frac{3}{10}$, and such a value would not indicate a deviation from H_0.

The R package coin contains the program `wilcox_test` that computes the P-value attained by referring W to its conditional distribution. For our $n = 2$, $m = 3$ data, let x<-c(1.2, 1.7, 2.8) and y<-c(.7, 1.7). Then the command `wilcox_test(c(x,y)~factor(c(0, 0, 0, 1, 1)), distribution = "exact", alt = "g")` yields the P-value .3 agreeing with what we obtained by enumeration.

6. *Large-Sample Approximation*. The statistic W/n is the average of the Y-ranks. All $\binom{N}{n}$ possible outcomes of the Y-ranks are equally likely under H_0. It follows that the null distribution of W/n is the same as the distribution of the sample mean of a random sample of size n drawn without replacement from the finite population $\{1, 2, \ldots, N\}$ of the first N integers. Next, we use results (i) and (ii), which are basic results from finite population theory concerning the mean and variance of the distribution of the sample mean of a sample of size n drawn without replacement from a finite population of N elements:

   (i) The mean is equal to the mean μ_{pop} of the finite population.

   (ii) The variance is equal to

$$\frac{\sigma^2_{\text{pop}}}{n} \times \frac{N-n}{N-1},$$

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

For the finite population $\{1, 2, \ldots, N\}$, direct calculations establish

(iii) $\mu_{\text{pop}} = \dfrac{1 + 2 + \cdots + N}{N} = \dfrac{N + 1}{2}$,

(iv) $\sigma^2_{\text{pop}} = \dfrac{1}{N}\{1^2 + 2^2 + \cdots + N^2\} - \left(\dfrac{N + 1}{2}\right)^2 = \dfrac{(N - 1)(N + 1)}{12}$.

From (i), (ii), (iii), and (iv), we then obtain

$$E_0\left(\frac{W}{n}\right) = \frac{N + 1}{2},$$

$$\text{var}_0\left(\frac{W}{n}\right) = \frac{(N - 1)(N + 1)}{12n} \times \frac{N - n}{N - 1} = \frac{m(N + 1)}{12n},$$

and it follows that

$$\text{var}_0(W) = \frac{mn(N + 1)}{12}.$$

Asymptotic normality of

$$W^* = \frac{W - \frac{n(N+1)}{2}}{\sqrt{\frac{mn(N+1)}{12}}} = \frac{W - E_0(W)}{\sigma_0(W)}$$

follows from standard theory for the mean of a sample from a finite population (cf. Wilks, 1962, p. 268).

Asymptotic normality results are also obtainable under general alternatives. See, for example, Lehmann's (1951) extension of Hoeffding's (1948a) U-statistic theorem as stated and applied to the Wilcoxon statistic on pages 92–94 of Randles and Wolfe (1979).

7. *The Mann–Whitney U Statistic*. For testing the hypothesis $H_0 : \Delta = 0$, Mann and Whitney (1947) proposed the statistic U given by (4.15),

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} \phi(X_i, Y_j),$$

where

$$\phi(X_i, Y_j) = \begin{cases} 1, & \text{if } X_i < Y_j, \\ 0, & \text{otherwise.} \end{cases}$$

The statistic U can be computed as follows. For each pair of values X_i and Y_j, observe which is smaller. If the X_i value is smaller, score 1 for that pair; if the Y_j value is smaller, score 0 for that pair. Add up the 0's and 1's and call the sum U. Mann and Whitney showed that, in the case of no ties,

$$W = U + \frac{n(n + 1)}{2}. \tag{4.20}$$

This implies that tests based on U are equivalent to tests based on W.

To establish (4.20), write

$$W = \sum_{j=1}^{n} R(Y_j), \tag{4.21}$$

where $R(Y_j)$ denotes the rank of Y_j in the joint ranking of the $m + n$ X's and Y's. Since the rank of Y_j is equal to the number of X's less than Y_j plus the number of Y's less than Y_j plus 1, write

$$R(Y_j) = \sum_{i=1}^{m} \phi(X_i, Y_j) + \sum_{j'=1}^{n} \phi(Y_{j'}, Y_j) + 1. \tag{4.22}$$

Substituting (4.22) into (4.21) yields

$$W = \sum_{j=1}^{n} \sum_{i=1}^{m} \phi(X_i, Y_j) + \sum_{j=1}^{n} \sum_{j'=1}^{n} \phi(Y_{j'}, Y_j) + n. \tag{4.23}$$

In (4.23), the first term on the right is U. The second term on the right is equal to the number of Y's less than the smallest Y plus the number of Y's less than the second smallest Y plus ... plus the number of Y's less than the largest Y, that is $0 + 1 + \cdots + n - 1$. Thus

$$W = U + \{1 + 2 + \cdots + n - 1\} + n = U + \frac{n(n+1)}{2},$$

recalling that the sum of the first n integers is equal to $n(n+1)/2$.

We illustrate the computation of U for the diffusion data of Table 4.1. The first row below counts the number of X-values less than 1.15, the second row counts the number of X-values less than .88, the third row the number of X-values less than .90, the fourth row the number of X-values less than .74, and the fifth row the number of X-values less than 1.21. The counts are given in brackets; each count is simply the number of 1s in the particular row. U is then obtained by summing the counts, which is equivalent to summing all the 1s.

$$\begin{aligned} U = {} & 1+1+0+1+0+0+0+0+1+0 && [4] \\ & +1+1+0+0+0+0+0+0+1+0 && [3] \\ & +1+1+0+0+0+0+0+0+1+0 && [3] \\ & +0+0+0+0+0+0+0+0+1+0 && [1] \\ & +1+1+0+1+0+0+0+0+1+0 && [4] \\ = {} & 15. \end{aligned}$$

Recall that in Example 4.1 we found $W = 30$. We could alternatively obtain W by computing U as before and then using (4.16) to find

$$W = 15 + \frac{5(6)}{2} = 30.$$

There is a generalization of (4.16) that holds when there are ties. If W is computed using average ranks and U is computed via

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} \phi^*(X_i, Y_j),$$

where

$$\phi^*(X_i, Y_j) = \begin{cases} 1, & \text{if } X_i < Y_j \\ \frac{1}{2}, & \text{if } X_i = Y_j \\ 0, & \text{if } X_i > Y_j, \end{cases}$$

then we still have $W = U + n(n+1)/2$. In other words, when there are ties, instead of scoring 1 if X is less than Y and 0 otherwise, compute U by scoring 1 if X is less than Y, $\frac{1}{2}$ if X equals Y, and 0 if X is greater than Y.

Bohn and Wolfe (1992, 1994) developed statistical procedures based on an analog of the Mann–Whitney statistic U for data obtained under the structure of ranked-set sampling. This form of data collection is a preferable alternative to simple random sampling when the actual sample measurements are costly and/or difficult to obtain, but ranking a small set of items is relatively easy and inexpensive.

Bohn (1996) provides a nice review of the general concept of ranked-set sampling, as well as an overview of the related nonparametric literature in this area of research. See Chapter 15 for more on ranked-set sampling.

8. *Symmetry of the Distribution of W under the Null Hypothesis*. When H_0 is true, the distribution of W is symmetric about its mean. This implies that when H_0 is true,

$$P(W \leq x) = P(W \geq n(m+n+1) - x), \tag{4.24}$$

for $x = n(n+1)/2, \ldots, n(2m+n+1)/2$.

Equation (4.24) is useful for converting upper-tail probabilities to lower-tail probabilities.

9. *Some Power Results for the Wilcoxon Test*. We consider the upper-tail α-level test of $H_0 : \Delta = 0$ versus $H_1 : \Delta > 0$ given by procedure (4.4). Suppose that the Y-population is the X-population shifted by an amount Δ, so that model (4.2) holds. Recall that

$$\text{Power} = \text{probability of rejecting } H_0, \text{ given that } H_0 \text{ is false.}$$

Then for Δ values "near" the null hypothesis value of 0, the power can be approximated as

$$\text{Power} \doteq \Phi(A_F), \tag{4.25}$$

where $\Phi(A_F)$ is the area under a standard normal density to the left of the point

$$A_F = \left[\left(\frac{12mn}{N+1} \right)^{1/2} \cdot f^*(0) \cdot \Delta \right] - z_\alpha, \tag{4.26}$$

where $f^*(0)$ is the density function, evaluated at 0, of the difference between two independent values drawn from the X-population having distribution F (cf. Lehmann (1975, p. 72, 403)).

When F is normal with standard deviation σ, $f^*(0) = 1/\{2\sigma(\pi)^{1/2}\}$ and A_F reduces to

$$A_{\text{normal}} = \left(\sqrt{\frac{3mn}{(N+1)\pi}} \cdot \frac{\Delta}{\sigma}\right) - z_\alpha. \tag{4.27}$$

Equation (4.27) shows that when F is normal, the approximate power depends on Δ and σ only through their ratio Δ/σ. (This is also true of the exact power.) Thus, for example, the power for the pair ($\Delta = 1$, $\sigma = 2$) is the same as the power for the pair ($\Delta = .5$, $\sigma = 1$).

Exact power values for the one-sided Wilcoxon test for model (4.2) when F is normal are given in Table B-1 of Milton (1970). Exact power values for the two-sided Wilcoxon test when F is normal are given in Table B-2 of Milton (1970). Milton's tables give power values for all sample sizes $2 \le n \le m \le 7$ that yield nontrivial results. If the sample of size m (or n) is from a normal population with mean μ_1 (or μ_2), $\mu_2 > \mu_1$, and variance σ^2, the location-shift alternative is defined in terms of $d = \{(\mu_2 - \mu_1)/\sigma\} = \Delta/\sigma$. Values are given for $d = .2(.2)1.0, 1.5, 2.0, 3.0$. Entries in the tables are ordered according to increasing values of $m + n$, from $2 \le m + n \le 14$. In Tables B-1 and B-2, the nominal levels of α are $\alpha = .25, .10, .05, .025, .01, .005$. The α's appearing in the tables are the attainable levels of significance nearest to but less than the nominal α's.

We suppose, for purposes of illustration, that model (4.2) holds with the underlying population F taken to be normal with variance $\sigma^2 = 16$ and the treatment effect $\Delta = 4$. Suppose further that we wish to determine, in a case where $m = 7$ and $n = 7$, the power of the $\alpha = .082$ test that rejects H_0 if $W \ge 64$ and accepts H_0 if $W < 64$. Substituting into (4.26) yields

$$A_{\text{normal}} = \left\{\left(\sqrt{\frac{3(7)(7)}{(15)\pi}} \cdot \left(\frac{4}{4}\right) - 1.39\right)\right\} = .376$$

and thus the power is approximately

$$\text{Power} \doteq \Phi(.376) = 1 - .35 = .65.$$

The exact power in this case is found from Table B-1 of Milton (1970) to be .635.

10. *Sample-Size Determination*. The Wilcoxon rank sum test detects a more general class of alternatives than the location-shift alternatives described by model (4.2). The one-sided upper-tail test defined by procedure (4.4) is consistent (i.e., has power tending to 1 as m, n tend to infinity) against those (F, G) populations for which $\delta > \frac{1}{2}$, where

$$\delta = P(X < Y). \tag{4.28}$$

The parameter δ defined by (4.28) is the probability that an X randomly selected from the distribution F will be less than an independent Y randomly selected from the distribution G. We say more about δ in Comment 18.

Noether (1987) shows how to determine an approximate total sample size N so that the α-level one-sided test given by procedure (4.4) will have an approximate power $1-\beta$ against an alternative value δ, where δ is greater than $\frac{1}{2}$. With $m = cN$, the approximate value of N is

$$N \doteq \frac{(z_\alpha + z_\beta)^2}{12c(1-c)(\delta - \frac{1}{2})^2}. \tag{4.29}$$

We illustrate the use of (4.29). Suppose we are testing H_0 and we desire to use an upper-tail $\alpha = .05$ test with power $= 1-\beta$ at least .90 against an alternative where $\delta = P(X < Y) = .7$ (recall that under H_0, $\delta = .5$). For simplicity, we take $m = n$ so that $c = .5$. From (4.29) with $z_\alpha = z_{.05} = 1.65$, $z_\beta = z_{.10} = 1.28$, and $\delta = .7$, we find

$$N \doteq \frac{(1.65 + 1.28)^2}{12(.5)(.5)(.7 - .5)^2} = 71.54, \qquad m = n = \frac{N}{2} = 35.8.$$

To be conservative take $m = n = 36$ rather than 35.

11. *Robustness of Level*. The significance level of the rank sum test is not preserved if the two populations differ in dispersion or shape. This is also the case for the normal theory two-sample t-test. For the effect of shape differences between the populations on the level of the rank sum test and other two-sample location procedures, see Pratt (1964). For a test of location differences that does not assume equal dispersions, see Fligner and Policello (1981) and Section 4.4.

    The level of the rank sum test is not preserved if dependencies exist among the X's or among the Y's, or if the X's are not independent of the Y's. Recall we have assumed that the NX's and Y's are mutually independent. For the effect on the level when this assumption is relaxed so that dependencies are allowed, see Serfling (1968); Hollander, Pledger, and Lin (1974) and Pettitt and Siskind (1981).

    There are other situations and designs in which the exact conditional randomization distribution of the Wilcoxon statistic is different than the usual Wilcoxon null distribution, and different approaches need to be used to obtain a P-value for comparing two treatments. See, for example, Efron's (1971) biased coin design and other restricted randomization designs considered by Hollander and Peña (1988) and Mehta, Patel, and Wei (1988).

12. *van der Waerden's Test*. The van der Waerden's rank statistic is

$$c = \sum_{j=1}^{n} \Phi^{-1}\left(\frac{S_j}{N+1}\right), \tag{4.30}$$

where, as before, $S_1, \ldots, S_n$ are the Y-ranks and $\Phi^{-1}(t)$ is the tth percentile of the $N(0,1)$ distribution. That is, $\Phi^{-1}(t)$ is the point such that the area under a $N(0,1)$ curve to the left of $\Phi^{-1}(t)$ is equal to t. The test of H_0 based on c has competitive efficiency properties versus the test based on W (see Section 4.5) and therefore is a popular competitor of W. To test $H_0 : \Delta = 0$ versus $H_1 : \Delta > 0$, reject H_0 for significantly large values of c. To test $H_0 : \Delta = 0$ versus $H_1 : \Delta < 0$, reject H_0 for significantly small values of c. To test $H_0 : \Delta = 0$ versus $H_3 : \Delta \neq 0$, reject

H_0 for significantly large values of $|c|$. Under H_0, the distribution of c is symmetric about 0. Tables of critical values are given by van der Waerden and Nievergelt (1956). The R package agricolae contains the program `waerden.test`, which gives P-values for van der Waerden's test. We illustrate using the water transfer data of Table 4.1. Let `at.term <-c(.80, .83, 1.89, 1.04, 1.45, 1.38, 1.91, 1.64, .73, 1.46)` and `gest.age <-c(1.15, .88, .90, .74, 1.21)`. Then, the R command `waerden.test (c(at.term, gest.age), factor (c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1)), group = T)` yields the two-sided P-value .26, which is in agreement with the one-sided P-value .13, which we illustrate in the following text.

The large-sample approximation is easy to perform. Under H_0, c has mean 0 and variance

$$\text{var}_0(c) = \frac{mn\left[\sum_{i=1}^{N}\{\Phi^{-1}(i/(N+1))\}^2\right]}{N(N-1)}. \tag{4.31}$$

The normal theory approximation to the distribution of

$$c^* = \frac{c}{\sqrt{\text{var}_0(c)}}, \tag{4.32}$$

treats c^* as an approximate $N(0,1)$ random variable for large m, n.

We illustrate the large-sample test based on c^* using the chorioamnion permeability data of Table 4.1 for which $m = 10$, $n = 5$, and $N = 15$. From the symmetry of the normal distribution, we note $\Phi^{-1}(i/16) = -\Phi^{-1}((16-i)/16)$ for $i = 1, \ldots, 7$, and $\Phi^{-1}\left(\frac{8}{16}\right) = \Phi^{-1}\left(\frac{1}{2}\right) = 0$. The values of $\Phi^{-1}(i/16)$ are found by using the R command `qnorm(x, 0, 1)`. For example, `qnorm(1/16, 0, 1) = -1.534` and so forth.

i:	1	2	3	4	5	6	7	8
$\Phi^{-1}(i/16)$:	−1.534	−1.150	−.887	−.674	−.489	−.319	−.157	0
i:	9	10	11	12	13	14	15	
$\Phi^{-1}(i/16)$:	.157	.319	.489	.674	.887	1.150	1.534.	

Recall that the Y-ranks for the data of Table 4.1 are 2, 5, 6, 8, and 9. From (4.30) we obtain

$$\begin{aligned} c &= \Phi^{-1}\left(\tfrac{2}{16}\right) + \Phi^{-1}\left(\tfrac{5}{16}\right) + \Phi^{-1}\left(\tfrac{6}{16}\right) + \Phi^{-1}(\tfrac{8}{16}) + \Phi^{-1}\left(\tfrac{9}{16}\right) \\ &= -1.150 - .489 - .319 + 0 + .157 = -1.80. \end{aligned}$$

From (4.31) we obtain

$$\begin{aligned} \text{var}_0(c) = \frac{10(5)}{15(14)}\{&(-1.534)^2 + (-1.150)^2 + (-.887)^2 + (-.674)^2 \\ &+ (-.489)^2 + (-.319)^2 + (-.157)^2 + (.157)^2 + (.319)^2 + (.489)^2 \\ &+ (.674)^2 + (.887)^2 + (1.150)^2 + (1.534)^2\} = 2.52. \end{aligned}$$

Then from (4.28) we find

$$c^* = \frac{-1.80}{\sqrt{2.52}} = -1.14,$$

with a one-sided P-value of .13.

Note that the results based on c are very close to those based on W that we found in Example 4.1. The large-sample approximation based on W gave a one-sided P-value of .11 and the exact P-value from W is $P = P_0(W \leq 30) = .127$.

A test that is asymptotically equivalent to the test based on c is the Fisher–Yates–Terry–Hoeffding (cf. Terry (1952), Hoeffding (1951)) test based on

$$c_1 = \sum_{j=1}^{n} E(V^{(S_j)}),$$

where $V^{(1)} < V^{(2)} < \cdots < V^{(N)}$ are the order statistics of a sample of size N from a $N(0,1)$ distribution and $S_1, \ldots, S_n$ are the Y-ranks. Values of $E(V^{(i)})$, $i = 1, \ldots, N$ for $N \leq 100$ and some larger sizes are given in Harter (1961). Exact tables can be found in Terry (1952) and Klotz (1964). Under H_0, the distribution of c_1 is symmetric about 0. The large-sample normal theory approximation treats

$$c_1^* = \frac{c_1}{\sqrt{\text{var}_0(c_1)}},$$

as a $N(0,1)$ random variable under H_0, where

$$\text{var}_0(c_1) = \frac{mn \sum_{i=1}^{N} \{E(V^{(i)})\}^2}{N(N-1)}.$$

Because $E(V^{(i)}) \doteq \Phi^{-1}(i/(N+1))$, it can be shown that tests based on c and c_1 are asymptotically equivalent. Both tests are often referred to as the *normal scores test*. Exact power values for the one-sided and two-sided tests based on c_1 for model (4.2) when F is normal are given in Tables B-3 and B-4 of Milton (1970).

13. *The Location-Shift Function*. Model (4.2) implies that the treatment effect is the same constant value Δ for each possible value of X. In some instances, it will be more appropriate to use a model that allows the treatment effects to be a function $\Delta(X)$ that is allowed to vary with X. For example, the treatment effect may be the expected increase (decrease) in systolic blood pressure due to taking a tranquilizer. In such a case, $\Delta(X)$ would depend on the patient's pretranquilizer blood pressure level X. This suggests the model

$$Y \stackrel{d}{=} X + \Delta(X), \tag{4.33}$$

where Y is systolic blood pressure after taking the tranquilizer. Model (4.33) was introduced by Lehmann (1975, p. 68). The function $\Delta(X)$ is called the *location-shift function*. Properties of $\Delta(X)$ were developed by Doksum (1974) and Switzer (1976). Doksum and Sievers (1976) derived simultaneous confidence bands for $\Delta(X)$. Hollander and Korwar (1982) and Wells and Tiwari (1989) extended the results of Doksum (1974) and Switzer (1976) to a nonparametric Bayesian framework. Lu, Wells, and Tiwari (1994) studied the location-shift function when the two samples are censored.

14. *Consistency of the W Test*. Under Assumptions A1–A3, the consistency of the tests based on W depends on the parameter

$$\delta^* = P(X < Y) - \tfrac{1}{2}.$$

The test procedures defined by (4.4), (4.5), and (4.6) are consistent against the alternatives for which $\delta^* >$, $<$, and $\neq 0$, respectively.

Properties

1. *Consistency*. For the location-shift model defined by (4.2), the tests defined by (4.4), (4.5), and (4.6) are consistent against the alternatives $\Delta >$, $<$, and $\neq 0$, respectively. Also, see Comment 14.
2. *Asymptotic Normality*. See Lehmann (1975, pp. 365–366).
3. *Efficiency*. See Section 4.5.

Problems

1. The data in Table 4.3 are a subset of the data obtained by Thomas and Simmons (1969), who investigated the relation of sputum histamine levels to inhaled irritants or allergens. The histamine content was reported in micrograms per gram of dry weight of sputum. The subjects for this portion of the study consisted of 22 smokers; 9 of them were allergics and the remaining 13 were asymptomatic (nonallergic) individuals. Care was taken to avoid people who carried out part of their daily work in an atmosphere of noxious gases or other respiratory toxicants. Table 4.3 gives the ordered sputum histamine levels for the 22 individuals in the study.

Test the hypothesis of equal levels versus the alternative that allergic smokers have higher sputum histamine levels than nonallergic smokers. Use the large-sample approximation.

2. Let W' be the sum of the ranks of the X observations. Verify directly, or illustrate using the chorioamnion permeability data of Table 4.1, the equation $W + W' = (m + n)(m + n + 1)/2$.

Table 4.3 Sputum Histamine Levels ($\mu g/g$ Dry Weight Sputum)

Allergics	Nonallergics
1651.0	48.1
1112.0	48.0
102.4	45.5
100.0	41.7
67.6	35.4
65.9	34.3
64.7	32.4
39.6	29.1
31.0	27.3
	18.9
	6.6
	5.2
	4.7

Source: H. V. Thomas and E. Simmons (1969).

3. Suppose a sixth Y observation is added to the five Y's of Table 4.1, and assume that the value $W = 30$ based on the original 10 X's and five Y's has already been calculated. How would you calculate the new value of W? Compare the method of reranking (to obtain new Y ranks) with a method based on using the Mann–Whitney statistic U in conjunction with the equation relating U and W (see Comment 7). Generalize the problem to different m and n values and make the same comparison.

4. Let U' denote the number of (X_i, Y_j) pairs for which $X_i > Y_j$. Assume that there are no $X = Y$ ties, and either establish directly or illustrate with the chorioamnion permeability data of Table 4.1, the relation $U' + U = mn$.

5. Molitor (1989) conducted a study to see if children who watched TV or film violence were significantly more tolerant of "real-life" violent behavior than children who instead watched a nonviolent TV show or film. Half of the 42 children in the study were shown violent TV (an edited version of *The Karate Kid*), whereas the other half watched exciting but nonviolent sports (highlights from the 1984 Summer Olympic Games). Each child was asked to "watch over" two younger children, supposedly in the next room, via a television monitor. Each child was instructed to go and get the research assistant (who stated she had to leave for an emergency) if the younger children "got into trouble." What each child witnessed, while alone, was actually a videotaped sequence depicting two small children first play with blocks and then progressively get more violent. That is, they called each other names, then pushed each other, chased each other, fought, and then supposedly broke a video camera while fighting.

Toleration of violence was measured by the time (in seconds) each child stayed in the room after he or she witnessed the two younger children's first act of violence. As soon as the subject child left the room, the timing clock was stopped. Each child was subsequently assured that an adult had entered the room where the two children were and that they were not hurt and the video camera was not damaged.

Do the data of Table 4.4 indicate that the children who viewed the violent TV tend to take longer to seek help (were more tolerant) than the children who viewed the nonviolent sports-action TV? Use Wilcoxon's W.

6. Assume that model (4.2) holds and that F is normal with variance 13. We have eight X observations and eight Y observations. If we use the $\alpha = .065$ test of $H_0 : \Delta = 0$ versus the alternative $\Delta > 0$, what is the approximate power of this test when the treatment effect is $\Delta = 2$?

7. For testing $H_0 : \Delta = 0$ versus the alternative $\Delta > 0$, you choose to use a type I error probability $\alpha = .10$. Using equal sample sizes, what should the common value of m, n be to have power at least .88 against an alternative where $\delta = .8$?

8. We observe $X_1 = 2.1, X_2 = 1.9, X_3 = 2.6, X_4 = 3.3, Y_1 = 1.9, Y_2 = 2.6$, and $Y_3 = 3.7$. What is the conditional distribution of W obtained by considering all $\binom{7}{3}$ possible choices of three data points to serve as the Y-values? How extreme is the observed value of W in this conditional distribution?

9. Apply van der Waerden's test based on c to the data of Table 4.4. Compare your result with that obtained in Problem 5 using Wilcoxon's W.

10. Apply the test based on W to the plasma glucose data of Table 4.6.

11. Apply the test based on c to the plasmas glucose data of Table 4.6. Compare with the results obtained in Problem 10.

12. Show directly, or illustrate via an example, that the maximum value of W is $n(2m + n + 1)/2$. What is the minimum value of W?

13. Suppose you reject H_0 if $W = n(2m + n + 1)/2$ or if $W = n(n + 1)/2$, and you accept H_0 otherwise. What is α for this test?

14. Suppose $m = n = 7$. Compare the exact $\alpha = .049$ test of $H_0 : \Delta = 0$ versus $H_1 : \Delta > 0$ based on W with its corresponding test based on large-sample approximation. What is the

Table 4.4 Seconds Spent in Room after Witnessing Violence

Olympics watchers	*Karate Kid* watchers
12	37
44	39
34	30
14	7
9	13
19	139
156	45
23	25
13	16
11	146
47	94
26	16
14	23
33	1
15	290
62	169
5	62
8	145
0	36
154	20
146	13

Source: F. T. Molitor (1989).

exact α value of the test based on the large-sample approximation whose nominal α value is .049?

15. Phadke et al. (2006) conducted a study to evaluate the soleus Hoffman reflex (H-reflex) for two different leg loading conditions on people who have not experienced spinal cord injuries (non-injured subjects) and people with incomplete spinal cord injuries (i-SCI subjects). The Phadke et al. (2006) paper was selected by Erin Easton (2006) for her term project in M. Hollander's 2006 Applied Nonparametric Statistics class. This problem is based on a portion of her analysis. Decreasing the load of weight on the leg is one way that patients with SCI undergo rehabilitation in order to relearn how to stand and walk. Leg loading is controlled through a body weight support (BWS) system that consists of a harness and a suspension system. The typical setting for BWS during rehabilitation for post-SCI patients is 60% leg loading (or 40% BWS). In the Phadke et al. study, 40% BWS was compared to 0% BWS (or 100% leg loading) for both i-SCI and noninjured subjects in order to determine whether a change in percent BWS changed the soleus H-reflex response for subjects in a standing position. Here, we focus on a portion of their data comparing noninjured to i-SCI subjects for 40% BWS.

The soleus muscle is one of the muscles that run from the just below the back of the knee down to the heel, and contraction of this muscle results in plantar flexion of the foot (pointing of the toes) and in maintenance of the body in a stable standing position. The H-reflex is an involuntary response (or flexion) in a muscle on electrical stimulation of the nerves that controls contraction and relaxation of the muscle. The tibialis anterior muscle is a muscle that runs along the front side of the tibia from below the knee to the top of the foot, and contraction of this muscle results in the dorsal flexion of the foot (rise of the foot toward the front of the leg). The tibial nerve runs along the entire back side of the leg, and it supplies electrical impulses to the muscles of the back of the leg, including the soleus. An electromyogram (EMG) is used to measure the electrical current in a muscle. The current is generally proportional to the activity level of the muscle, where an inactive muscle has no current. The H/M ratio is the ratio of the maximum soleus H-reflex to the maximum soleus muscle potential (or to a preset percentage

Table 4.5 H/M Ratios of Noninjured Subjects and i-SCI Subjects For 40% BWS

Noninjured H/M ratios	Ranks	i-SCI H/M ratios	Ranks
.19	4	.89	13
.14	3	.76	10
.02	1.5	.63	8
.44	6	.69	9
.37	5	.58	7
		.79	11.5
		.02	1.5
		.79	11.5

Source: C.P. Phadke, S.S. Wu, F.J. Thompson, and A.L. Behrman (2006).

of the maximum potential). Table 4.5 gives the H/M ratios for five noninjured subjects and eight i-SCI subjects for 40% BWS.

Is there evidence, for this 40% BWS situation, that the H/M ratios of the i-SCI subjects are significantly larger than the H/M ratios of the noninjured subjects? What is the approximate P-value achieved by your test.

16. Apply the exact conditional test based on W (see Comment 5) to the H/M ratios data of Table 4.5. Compare your result with that obtained in Problem 15.

17. Apply van der Waerden's test to the H/M ratios data of Table 4.5. Compare your result with the results of Problems 15 and 16.

4.2 AN ESTIMATOR ASSOCIATED WITH WILCOXON'S RANK SUM STATISTIC (HODGES–LEHMANN)

Procedure

To estimate Δ of model (4.2), form the mn differences $Y_j - X_i$, for $i = 1, \ldots, m$ and $j = 1, \ldots, n$. The estimator of Δ associated with the Wilcoxon rank sum statistic (see Comment 15) is

$$\widehat{\Delta} = \text{median}\{(Y_j - X_i), i = 1, \ldots, m; j = 1, \ldots, n\}. \tag{4.34}$$

Let $U^{(1)} \leq \cdots \leq U^{(mn)}$ denote the ordered values of $Y_j - X_i$. Then if mn is odd, say $mn = 2k + 1$, we have $k = (mn - 1)/2$ and

$$\widehat{\Delta} = U^{(k+1)}, \tag{4.35}$$

the value that occupies the position $k + 1$ in the list of the ordered $Y - X$ differences. If mn is even, say $mn = 2k$, then $k = mn/2$ and

$$\widehat{\Delta} = \frac{U^{(k)} + U^{(k+1)}}{2}. \tag{4.36}$$

That is, $\widehat{\Delta}$ is the average of the two $Y - X$ differences that occupy the positions k and $k + 1$ in the ordered list of the mn differences.

EXAMPLE 4.3 ***Continuation of Example 4.1.***

To estimate Δ for the chorioamnion permeability data of Table 4.1, we obtain, using R, the ordered values $U^{(1)} \leq \cdots \leq U^{(50)}$.

```
diff<-numeric(0); m<-10; n<-5
for (i in 1:m) for (j in 1:n) diff<-c(diff,
gest.age[j]-at.term[i])
diff<-sort(diff)
```

Then we obtain the 50 ordered $Y_j - X_i$ differences displayed in Table 4.6.
The value of $mn = 50$ is even and thus we use (4.36) with $k = \frac{50}{2} = 25$ to obtain

$$\widehat{\Delta} = \frac{U^{(25)} + U^{(26)}}{2} = \frac{-.31 - .30}{2} = -.305.$$

The estimate $\widehat{\Delta} = -.305$ is directly available from the R command `wilcox.test`. If you perform `wilcox.test(at.term, gest.age, alternative="t", conf.int=T)`, the output .305 is the estimate for the difference in location because `wilcox.test` computes the median of the X-Y differences. To get the median of the Y-X differences perform `wilcox.test(gest.age, at.term, alternative="t", conf.int=T)` and obtain $\widehat{\Delta} = -.305$.

Comments

15. *Motivation for the Hodges–Lehmann Estimator*. The Hodges–Lehmann (1963) estimator $\widehat{\Delta}$ defined by (4.34) is associated with the Wilcoxon rank sum test. When $\Delta = 0$, the distribution of the statistic W is symmetric about its mean $n(m + n + 1)/2$ (see Comment 8). A reasonable estimator of Δ is the amount $\widehat{\Delta}$ (say) that should be subtracted from each Y_j so that the value of W, when applied to the aligned samples $X_1, \ldots, X_m, Y_1 - \widehat{\Delta}, \ldots, Y_n - \widehat{\Delta}$, is $n(m + n + 1)/2$. Roughly speaking, we estimate Δ by the amount $(\widehat{\Delta})$ that the Y sample should be shifted in order that $X_1, \ldots, X_m$ and $Y_1 - \widehat{\Delta}, \ldots, Y_n - \widehat{\Delta}$ appear (when "viewed" by the rank sum statistic W) as two samples from the same population. (Under Assumptions A1–A3, the variables $X_1, \ldots, X_m$ and $Y_1 - \Delta, \ldots, Y_n - \Delta$ can be taken as a single sample of size $N = m + n$ from the underlying population.)

The Hodges–Lehmann method can be applied to large classes of statistics, which, for example, include van der Waerden's V. The forms of the resulting

Table 4.6 Ordered Y-X Differences for the Chorioamnion Permeability Data

$U^{(1)} \leq U^{(2)} \leq \cdots \leq U^{(50)}$											
−1.17	−1.15	−1.03	−1.01	−1.01	−0.99	−0.90	−0.76	−0.76	−0.74	−0.74	−0.72
−0.71	−0.70	−0.68	−0.64	−0.58	−0.57	−0.56	−0.55	−0.50	−0.49	−0.48	−0.43
−0.31	−0.30	−0.30	−0.25	−0.24	−0.23	−0.17	−0.16	−0.14	−0.09	−0.06	0.01
0.05	0.07	0.08	0.10	0.11	0.15	0.17	0.17	0.32	0.35	0.38	0.41
0.42	0.48										

estimators are not always as convenient for calculation as in the case of $\widehat{\Delta}$. See Hodges and Lehmann (1983) for an expository article on their method. See McKean and Ryan (1977) for an algorithm for computing $\widehat{\Delta}$.

16. *Sensitivity to Gross Errors*. The estimator $\widehat{\Delta}$ is less sensitive to gross errors than its normal theory analog $\overline{Y} - \overline{X}$, the difference of the sample averages.

17. *Competing Estimators*. Observe that the estimator $\widehat{\Delta}$ cannot be written as a difference of a statistic based on the Y observations only and a second statistic based on the X observations only. The classical estimator $\overline{\Delta} = \overline{Y} - \overline{X}$ can be written as such a difference. When the underlying population is symmetric, Lehmann (1963a) proposed to estimate Δ by

$$\widehat{\widehat{\Delta}} = \widehat{\theta}_2 - \widehat{\theta}_1,$$

where $\widehat{\theta}_1(\widehat{\theta}_2)$ is the estimator (3.10) associated with the signed rank statistic T^+ for estimating the location of the population corresponding to the $X(Y)$ observations. That is,

$$\widehat{\widehat{\Delta}} = \text{median}\left\{\frac{Y_i + Y_j}{2}, 1 \le i \le j \le n\right\} - \text{median}\left\{\frac{X_i + X_j}{2}, 1 \le i \le j \le m\right\}. \tag{4.37}$$

The standard deviation of $\widehat{\widehat{\Delta}}$ can be estimated by

$$\widehat{\sigma}_{\widehat{\widehat{\Delta}}} = \left\{\left(\frac{\theta_{2U} - \theta_{2L}}{2z_{\alpha_2/2}}\right)^2 + \left(\frac{\theta_{1U} - \theta_{1L}}{2z_{\alpha_1/2}}\right)^2\right\}^{1/2}, \tag{4.38}$$

where θ_{2U} and θ_{2L} are the upper and lower endpoints, respectively, of the $100(1-\alpha_2)\%$ confidence interval obtained from the method of Section 3.3 by replacing the Z's of Section 3.3 by the n Y's of sample 2. Similarly, θ_{1U} and θ_{1L} are the end points of the $100(1-\alpha_1)\%$ confidence interval obtained by the method of Section 3.3 by replacing the Z's of Section 3.3 by the m X's of sample 1.

An approximate confidence interval for Δ, with the confidence coefficient $1-\alpha$, is

$$\Delta_\ell = \widehat{\widehat{\Delta}} - z_{\alpha/2}\widehat{\sigma}_{\widehat{\widehat{\Delta}}}, \qquad \Delta_u = \widehat{\widehat{\Delta}} + z_{\alpha/2}\widehat{\sigma}_{\widehat{\widehat{\Delta}}}. \tag{4.39}$$

Lehmann (1963a), Høyland (1965), and Ramachandramurty (1966a) investigated the properties of $\overline{\Delta}$, $\widehat{\Delta}$ and $\widehat{\widehat{\Delta}}$ for various deviations from the assumptions, including asymmetry and non-location-differences between the populations.

Other competing estimators of $\widehat{\Delta}$ include those in classes initiated by Serfling (1984), Akritas (1986), and Serfling (1992).

18. *The Probability That X Is Less Than Y*. A quantity of interest in the two-sample location problem is the parameter $\delta = P(X_1 < Y_1)$, where X_1 is a random member from the X population, Y_1 is a random member from the Y population, and X_1 and Y_1 are independent; that is, δ is the probability that a single Y observation will be larger than a single X observation. Pitman (1948) and Birnbaum (1956) discussed a point estimator for δ given by $\widehat{\delta} = U/mn$, where U is the

Mann–Whitney form of the rank sum statistic (see Comment 7). Upper bounds for the variance of U, which are useful when using $\widehat{\delta}$ as a point estimator for δ, were obtained in terms of δ by van Dantzig (1951). See Birnbaum and Klose (1957) for lower bounds. Lehmann (1951) showed that $\widehat{\delta}$ is the uniform minimum variance unbiased estimator of δ over the class of continuous populations (also see Blyth (1950)). For the use of the sign statistic in obtaining a point estimator for δ, see Saxena (1969).

Many statisticians, including Wolfe and Hogg (1971), have emphasized the importance of natural parameters such as δ. Consider a medical application in which X represents the response to treatment A and Y is the response to treatment B. Let μ_1, μ_2 be the respective means of the X and Y populations and let σ denote the (assumed) common standard deviation. Then, $P(X < Y) = .76$ will usually make more sense to a doctor than the statement $\{(\mu_2 - \mu_1)/\sigma\} = 1$. (If X and Y are normal, and independent, with means μ_1, μ_2, and common standard deviation σ, then $\{(\mu_2 - \mu_1)/\sigma\} = 1$ implies $P(X < Y) = .76$.) Furthermore, we are often more interested in the probability that X is less than Y than, say, in the difference between the Y and X means. This is true in a good deal of biological research, where, for example, a large liver is a large liver, but how large it is makes little difference except possibly in comparison with other livers (rather than in comparison with scale measurements on a weighing machine). In situations such as these, the estimator $\widehat{\delta}$ may be more useful than the estimator $\widehat{\Delta}$.

Birnbaum (1956) and Birnbaum and McCarty (1958) considered a distribution-free upper confidence bound for $\delta = P(X_1 < Y_1)$ based on the Mann–Whitney U when the underlying populations are continuous. Owen, Craswell, and Hanson (1964) extended this to discrete populations, and Govindarajulu (1968) sharpened the Birnbaum–McCarty upper bound and provided corresponding two-sided distribution-free confidence intervals for δ. Sen (1967) and Govindarajulu (1968) considered asymptotically distribution-free confidence bounds for δ based on consistent estimators of the variance of the Mann–Whitney U. Saxena (1969) discussed distribution-free confidence bounds for δ based on the sign statistic.

The parameter δ also arises naturally in reliability. Let X be the stress on a component and let Y be the strength of the component. Then, $\delta = P(X < Y)$ is the probability that the component functions properly. Johnson (1988) surveys many of the methods referenced in this comment in the context of reliability. His focus is on getting estimators and confidence bounds on system reliability in reliability systems such as "k out of n" systems.

Sen's (1967) asymptotic nonparametric interval for δ is relatively easy to obtain. Sen's interval is based on the asymptotic normality of $\sqrt{n_0}(\widehat{\delta} - \delta)/s$, where $n_0 = mn/(m+n)$. Here, s is a consistent estimator of the standard deviation of $\sqrt{n_0}\widehat{\delta}$. Many estimators are available. A particularly convenient one defined by Sen is

$$s^2 = \frac{nS_{10}^2 + mS_{01}^2}{m+n},$$

where

$$S_{10}^2 = \frac{\sum_{i=1}^{m}(R_i - i)^2 - m(\overline{R} - (m+1)/2)^2}{(m-1)n^2},$$

and

$$S_{01}^2 = \frac{\sum_{j=1}^{n}(S_j - j)^2 - n(\overline{S} - (n+1)/2)^2}{(n-1)m^2}.$$

Here, R_i is the rank of $X_{(i)}$ in the joint ranking of the X's and Y's, S_j is the rank of $Y_{(j)}$ in the joint ranking of the X's and Y's, $\overline{R} = \sum_{i=1}^{m} R_i/m$, and $\overline{S} = \sum_{j=1}^{n} S_j/n$. Recall that $X_{(1)} \le \cdots \le X_{(m)}$ are the ordered X-values and $Y_{(1)} \le \cdots \le Y_{(n)}$ are the ordered Y-values. The lower and upper end points, δ_{L} and δ_{U}, respectively, of the asymptotic $1-\alpha$ confidence interval are

$$\delta_{\mathrm{L}}^S = \widehat{\delta} - z_{\alpha/2}\sqrt{\frac{nS_{10}^2 + mS_{01}^2}{mn}},$$

$$\delta_{\mathrm{U}}^S = \widehat{\delta} + z_{\alpha/2}\sqrt{\frac{nS_{10}^2 + mS_{01}^2}{mn}}. \tag{4.40}$$

A competing interval has been proposed by Halperin, Gilbert, and Lachin (1987). Their $1-\alpha$ confidence interval is

$$\delta_{\mathrm{L}}^H = \frac{A-B}{C}, \qquad \delta_{\mathrm{U}}^H = \frac{A+B}{C}, \tag{4.41}$$

where

$$A = \widehat{\delta} + \frac{\gamma z_{\alpha/2}^2}{2mn},$$

$$B = \left(\frac{(\widehat{\delta}(1-\widehat{\delta})\gamma z_{\alpha/2}^2 + \gamma^2 z_{\alpha/2}^4/4mn)}{mn}\right)^{1/2},$$

$$C = 1 + \frac{\gamma z_{\alpha/2}^2}{mn},$$

$$\gamma = \widehat{\theta}(m+n-2) + 1,$$

$$\widehat{\theta} = \frac{\dfrac{\widehat{K} + 2(n-1)\widehat{\delta}}{m+n-2} - \widehat{\delta}^2}{\widehat{\delta}(1-\widehat{\delta})},$$

$$\widehat{K} = \left\{\frac{\sum_{j=1}^{n} r_{1j}(r_{1j}-1)}{mn}\right\} + \left\{\frac{\sum_{i=1}^{m} s_{1i}(s_{1i}-1)}{mn}\right\} - (n-1),$$

where r_{1j} is the number of X-observations that are less than $Y_{(j)}$ and s_{1i} is the number of Y-observations that are less than $X_{(i)}$.

Halperin, Gilbert, and Lachin point out that δ_{U}^H is less than 1 and δ_{L}^U is greater than 0. Also, $\widehat{\theta} \le 1$ if $\widehat{\delta}$ is neither 0 nor 1, but for some samples, $\widehat{\theta}$ may be less than 0. If that happens, take $\widehat{\theta} = 0$ in the definition of γ. If $\widehat{\delta} = 0$ or 1, take $\widehat{\theta} = 1$.

Halperin, Gilbert, and Lachin did simulations that indicated their method generally yields coverage probabilities closer to the nominal $1-\alpha$ than does the Sen method.

For the chorioamnion permeability data of Table 4.1, the approximate 95% Sen confidence interval for δ and the approximate 95% Halperin–Gilbert–Lachin confidence interval for δ are as follows. Recall that for these data, we have found (see Comment 7) $U = 15$ and thus

$$\widehat{\delta} = \frac{15}{10(5)} = .3.$$

For the Sen interval,

$$S_{10}^2 = .171, \qquad S_{01}^2 = .015.$$

From display (4.40), we obtain, with $\alpha = .05$,

$$\delta_{\mathrm{L}}^S = .02, \qquad \delta_{\mathrm{U}}^S = .58.$$

For the Halperin–Gilbert–Lachin interval, with $\alpha = .05$, we find

$$\widehat{K} = -.76, \qquad \widehat{\theta} = .172, \qquad \gamma = 3.24$$
$$A = .424, \qquad B = .260, \qquad C = 1.25.$$

From display (4.41), we obtain

$$\delta_{\mathrm{L}}^H = .13, \qquad \delta_{\mathrm{U}}^H = .55.$$

Properties

1. *Standard Deviation of* $\widehat{\Delta}$. For the asymptotic standard deviation of $\widehat{\Delta}$, see Hodges and Lehmann (1963), Lehmann (1963c), and Comment 21.
2. *Asymptotic Normality*. See Hodges and Lehmann (1963) and Ramachandramurty (1966a).
3. *Efficiency*. See Hodges and Lehmann (1963), Høyland (1965), Ramachandramurty (1966a), and Section 4.5.

Problems

18. Consider the data of Table 4.3. Associate the Y's (X's) with the allergies (nonallergies) and estimate Δ of model (4.2) using $\widehat{\Delta}$.

19. Again consider the data of Table 4.3. Estimate Δ using $\widehat{\widehat{\Delta}}$ and compare your estimate with $\widehat{\Delta}$ obtained in Problem 15.

20. Consider the data of Table 4.3. Use display (4.35) to obtain an approximate 95% confidence interval for Δ.

21. Consider the data of Table 4.3. Estimate $\delta = P(X < Y)$ and determine an approximate 90% confidence interval for δ.

22. Consider the data of Table 4.4. Estimate Δ of model (4.2) using $\widehat{\Delta}$.

23. Consider the data of Table 4.4. Estimate Δ using $\widehat{\widehat{\Delta}}$ and compare your estimate with $\widehat{\Delta}$ obtained in Problem 22.

24. Consider the data of Table 4.4. Use Comment 17 to obtain an approximate 93% confidence interval for Δ.

25. Consider the data of Table 4.4. Estimate $\delta = P(X < Y)$ and determine (a) an approximate 93% confidence interval for δ using the Sen's interval and (b) an approximate 93% confidence interval for δ using the Halperin–Gilbert–Lachin interval.

26. Change the value 102.4, appearing in Table 4.3, to 1024. How does this affect the estimate of Δ given by $\widehat{\Delta}$? How does this affect the estimate of Δ given by $\overline{\Delta} = \overline{Y} - \overline{X}$?

27. (a) What happens to $\widehat{\Delta}$ when we add a number b to each of the m X values and a number c to each of the n Y values? In particular, what happens when $b = c$?

(b) What happens to $\widehat{\Delta}$ when we multiply each of the X and Y values by the same number d?

28. Answer parts (a) and (b) of Problem 27 with $\widehat{\Delta}$ replaced by $\widehat{\widehat{\Delta}}$.

29. Do you need to calculate the values of all mn $Y - X$ differences in order to compute the value of $\widehat{\Delta}$? Explain.

4.3 A DISTRIBUTION-FREE CONFIDENCE INTERVAL BASED ON WILCOXON'S RANK SUM TEST (MOSES)

Procedure

For a symmetric two-sided confidence interval for Δ, with the confidence coefficient $1 - \alpha$, let $w_{\alpha/2}$ denote the upper $\alpha/2$ percentile point of the null distribution of W.

Then with

$$C_\alpha = \frac{n(2m + n + 1)}{2} + 1 - w_{\alpha/2}, \tag{4.42}$$

the $1 - \alpha$ confidence interval (Δ_L, Δ_U) is given by

$$\Delta_L = U^{(C_\alpha)}, \qquad \Delta_U = U^{(mn+1-C_\alpha)}. \tag{4.43}$$

That is, Δ_L is the $Y - X$ difference that occupies the position C_α in the list of the mn ordered $Y - X$ differences. The upper endpoint Δ_U is the $Y - X$ difference that occupies the position $mn + 1 - C_\alpha$ in the ordered list. With Δ_L and Δ_U given by display (4.43), we have, for all Δ,

$$P_\Delta(\Delta_L < \Delta < \Delta_U) = 1 - \alpha. \tag{4.44}$$

The confidence interval is found directly from the `R` command `wilcox.test`. We illustrate this in Example 4.4.

Large-Sample Approximation

For large m and n, the integer C_α may be approximated by

$$C_\alpha \approx \frac{mn}{2} - z_{\alpha/2}\left\{\frac{mn(m + n + 1)}{12}\right\}^{1/2}. \tag{4.45}$$

In general, the value of the right-hand side of (4.45) is not an integer. To be conservative, take C_α to be the largest integer that is less than or equal to the right-hand side of (4.45).

EXAMPLE 4.4 ***Continuation of Example 4.1.***

Consider the chorioamnion permeability data of Table 4.1. We will illustrate how to obtain the 96% confidence interval for Δ. Use the R command `wilcox.test (gest.age, at.term, conf.int = T, conf.level = .96)` to obtain the interval $\Delta_L = -.76$, $\Delta_U = .15$. Note $\Delta_L = U^{(9)}$ and $\Delta_U = U^{(42)}$ (see Table 4.6).

Applying the large-sample approximation, we find from approximation (4.45)

$$C_{.04} \approx \frac{10(5)}{2} - 2.05\left\{\frac{10(5)(10+5+1)}{12}\right\}^{1/2} = 8.3.$$

Thus, with the large-sample approximation, we set $C_{.04}$ equal to 8 and

$$\Delta_L = U^{(8)} = -.76, \qquad \Delta_U = U^{(43)} = .17.$$

Comments

19. *Relationship of Confidence Interval to Test*. The $1-\alpha$ confidence interval given by display (4.43) can be obtained from the two-sided rank sum test as follows. The confidence interval (Δ_L, Δ_U) consists of those Δ_0 values for which the two-sided α-level test of $\Delta = \Delta_0$ (see Comment 2) accepts the hypothesis $\Delta = \Delta_0$. The confidence interval given by display (4.43) was defined by way of a graphical procedure by Lincoln Moses in Chapter 18 of Walker and Lev (1953). See Lehmann (1986, p. 90) for a general result relating confidence intervals and acceptance regions of tests, and see Lehmann (1963c) for the specific result involving the rank sum test.
20. *Midpoint of Confidence Interval as an Estimator*. The midpoint of the interval (4.43), namely, $\{U^{(C_\alpha)} + U^{(mn+1-C_\alpha)}\}/2$, suggests itself as a reasonable estimator of Δ. (Note that this actually yields a class of estimators depending on the value of α.) In general this midpoint is not the same as $\widehat{\Delta}$. Lehmann (1963b) has also dealt with an asymptotically distribution-free confidence interval for Δ centered at $\widehat{\Delta}$, and Lehmann (1963c) has shown that the asymptotically distribution-free confidence interval has the same asymptotic behavior as the distribution-free confidence interval given by display (4.43).
21. *Estimating the Asymptotic Standard Deviation of* $\widehat{\Delta}$. The quantity $(\Delta_U - \Delta_L)/(2z_{\alpha/2})$, where (Δ_L, Δ_U) is the $1-\alpha$ confidence interval defined by display (4.43), provides us with a consistent estimator for the asymptotic standard deviation of the point estimator $\widehat{\Delta}$ (see Lehmann, (1963c)).
22. *Confidence Bounds*. To obtain a lower confidence bound for Δ, with the confidence coefficient $1-\alpha$, set

$$C_\alpha^* = \frac{n(2m+n+1)}{2} + 1 - w_\alpha, \tag{4.46}$$

where w_α, the upper α percentile point of the null distribution of W. The $100(1-\alpha)\%$ lower confidence bound Δ_L^* for Δ that is associated with the one-sided Wilcoxon rank sum test of $H_0 : \Delta = 0$ against the alternative $H_1 : \Delta > 0$ is given by

$$(\Delta_L^*, \infty) = (U^{(C_\alpha^*)}, \infty), \tag{4.47}$$

where $U^{(1)} \le \cdots \le U^{(mn)}$ are the ordered values of $Y_j - X_i$. With Δ_L^* defined by (4.47), we have, for all Δ,

$$P_\Delta(\Delta_L^* < \Delta < \infty) = 1 - \alpha. \tag{4.48}$$

The $100(1-\alpha)\%$ upper confidence bound Δ_U^* for Δ that is associated with the one-sided Wilcoxon rank sum test of $H_0 : \Delta = 0$ against the alternative $H_1 : \Delta < 0$ is given by

$$(-\infty, \Delta_U^*) = (-\infty, U^{(mn+1-C_\alpha^*)}), \tag{4.49}$$

where C_α^* is given by (4.46). With Δ_U^* defined by (4.49), we have, for all Δ,

$$P_\Delta(-\infty < \Delta < \Delta_U^*) = 1 - \alpha. \tag{4.50}$$

For large m, n, the integer C_α^* can be approximated by

$$C_\alpha^* \cong \frac{mn}{2} - z_\alpha \left\{ \frac{mn(m+n+1)}{12} \right\}^{1/2}. \tag{4.51}$$

Properties

1. Under Assumptions A1–A3 and model (4.2), (4.44) holds. Hence, we can control the coverage probability to be $1-\alpha$ without having more specific knowledge about the form of the underlying distribution. Thus (Δ_L, Δ_U) is a distribution-free confidence interval for Δ over a very large class of populations.
2. *Efficiency*. See Lehmann (1963c) and Section 4.5.

Problems

30. Refer to Problem 18 and obtain a confidence interval for Δ with approximate confidence coefficient .95.

31. For the chorioamnion permeability data of Table 4.1, compute an estimate of Δ utilizing the estimator defined in Comment 20. Compare with the value of $\widehat{\Delta}$ obtained in Example 4.3.

32. Use the results of Example 4.4 to obtain an estimate for the asymptotic standard deviation of $\widehat{\Delta}$ (see Comment 21).

33. Consider the $1-\alpha$ confidence interval defined by display (4.43). Show that when $\alpha = 2/\binom{N}{n}$,

$$\Delta_L = Y_{(1)} - X_{(m)}, \qquad \Delta_U = Y_{(n)} - X_{(1)},$$

where $X_{(1)} \le \ldots \le X_{(m)}$ are the ordered X's and $Y_{(1)} \le \cdots \le Y_{(n)}$ are the ordered Y's.

34. Consider the $1-\alpha$ confidence interval defined by display (4.43). Show that when $\alpha = 4/\binom{N}{n}$,

$$\Delta_L = \text{minimum}\{Y_{(2)} - X_{(m)}, Y_{(1)} - X_{(m-1)}\},$$
$$\Delta_U = \text{maximum}\{Y_{(n)} - X_{(2)}, Y_{(n-1)} - X_{(1)}\}.$$

35. Consider the data of Table 4.3. Obtain an approximate 95% confidence interval for Δ using the large-sample approximation of this section. Compare your result with the approximate 95% confidence interval obtained in Problem 20.

36. Consider the data of Table 4.2 and obtain an approximate 90% confidence interval for Δ using the large-sample approximation of this section.

37. Consider the data of Table 4.4 and obtain an approximation 99% confidence interval for Δ using the large-sample approximation of this section.

38. Consider the case $m = n = 8$ and compare the exact 91.8% confidence interval given by display (4.43) with that obtained by the large-sample approximation.

39. Consider the case $m = n = 10$ and compare the exact 91% confidence interval given by display (4.43) with that obtained by the large-sample approximation.

40. Consider the data of Table 4.5 and obtain a 95% confidence interval for Δ.

4.4 A ROBUST RANK TEST FOR THE BEHRENS–FISHER PROBLEM (FLIGNER–POLICELLO)

Hypothesis

In this section we introduce new assumptions. Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent random samples from continuous distributions that are symmetric about the population medians θ_x and θ_y, respectively. Note that we do not require the X and Y populations to have the same distributional form nor do we assume that the variances of the two populations are equal. We are interested in testing $H_0' : \theta_x = \theta_y$ versus $\theta_x < \theta_y$ [or $\theta_x > \theta_y$ or $\theta_x \neq \theta_y$]. This problem of testing $H_0' : \theta_x = \theta_y$ without assuming equal variances is often referred to as the *Behrens–Fisher problem*.

Procedure

Let

$$P_i = [\text{number of sample } Y \text{ observations less than } X_i], \tag{4.52}$$

for $i = 1, \ldots, m$. Similarly, set

$$Q_j = [\text{number of sample } X \text{ observations less than } Y_j], \tag{4.53}$$

for $j = 1, \ldots, n$. We call P_i and Q_j the *placements* of X_i and Y_j, respectively. Compute

$$\overline{P} = \frac{1}{m}\sum_{i=1}^{m} P_i = \text{average } X \text{ sample placement}, \tag{4.54}$$

and

$$\overline{Q} = \frac{1}{n}\sum_{j=1}^{n} Q_j = \text{average } Y \text{ sample placement}. \tag{4.55}$$

Let

$$V_1 = \sum_{i=1}^{m}(P_i - \overline{P})^2 \quad \text{and} \quad V_2 = \sum_{j=1}^{n}(Q_j - \overline{Q})^2, \tag{4.56}$$

and set

$$\widehat{U} = \frac{\sum_{j=1}^{n} Q_j - \sum_{i=1}^{m} P_i}{2(V_1 + V_2 + \overline{P}\,\overline{Q})^{1/2}}. \tag{4.57}$$

a. *One-Sided Upper-Tail Test*. For a one-sided test of $H_0' : \theta_x = \theta_y$ versus the one-sided alternative $H_1' : \theta_y > \theta_x$ at the approximate α level of significance,

$$\text{Reject } H_0' \text{ if } \widehat{U} \geq u_\alpha; \qquad \text{otherwise do not reject,} \tag{4.58}$$

where u_α is a constant satisfying $P_0(\widehat{U} \geq u_\alpha) \approx \alpha$.

b. *One-Sided Lower-Tail Test*. For a one-sided test of $H_0' : \theta_x = \theta_y$ versus the alternative $H_2' : \theta_y < \theta_x$ at the approximate α level of significance, we

$$\text{Reject } H_0' \text{ if } \widehat{U} \leq -u_\alpha; \qquad \text{otherwise do not reject.} \tag{4.59}$$

c. *Two-Sided Test*. For a two-sided test of $H_0' : \theta_x = \theta_y$ versus the alternative $H_3' : \theta_y \neq \theta_x$ at the approximate α level of significance, we

$$\text{Reject } H_0' \text{ if } |\widehat{U}| \geq u_{\alpha/2}; \qquad \text{otherwise do not reject.} \tag{4.60}$$

Any u_α can be computed exactly or estimated using Monte Carlo simulation for large m and n using the `R` command `cFligPoli`.

Large-Sample Approximation

When $H_0' : \theta_x = \theta_y$ is true, the statistic $\widehat{U}$ has an asymptotic ($\min(m, n)$ tending to infinity) $N(0, 1)$ distribution. Thus the normal theory approximations to procedures (4.58), (4.59), and (4.60) are obtained by replacing u_α and $u_{\alpha/2}$ by z_α and $z_{\alpha/2}$, respectively.

The `R` function `pFligPoli` (with `method="Monte carlo"`) performs a Monte Carlo approximation to the P-value of the statistic $\hat{U}$ and the `R` function `pFligPoli` (with `method="Asymptotic"`) performs the large-sample approximation.

Ties

If there are ties among the N sample observations, replace the placement formulas (4.52) and (4.53) by

$$P_i = \Big\{[\text{number of sample } Y \text{ observations less than } X_i] + \tfrac{1}{2}[\text{number of sample } Y \text{ observations equal to } X_i]\Big\} \tag{4.61}$$

and

$$Q_j = \Big\{[\text{number of sample } X \text{ observations less than } Y_j] + \tfrac{1}{2}[\text{number of sample } X \text{ observations equal to } Y_j]\Big\}, \tag{4.62}$$

respectively.

Table 4.7 Plasma Glucose Values

Healthy geese	Lead-poisoned geese
297	293
340	291
325	289
227	430
277	510
337	353
250	318
290	

Source: G. L. March, T. M. John, B. A. McKeown, L. Sileo and J. C. George (1976).

EXAMPLE 4.5 ***Plasma Glucose in Geese.***

March et al. (1976) were interested in, among other things, examining the differences between healthy (normal) and lead-poisoned Canadian geese. In particular, one of the measures examined was plasma glucose (mg/100 ml plasma). The data they obtained for eight healthy and seven lead-poisoned geese are given in Table 4.7.

Labeling the lead-poisoned geese as the Y-sample (because there are fewer lead-poisoned observations), the authors were interested in testing H_0': $\theta_x = \theta_y$ versus H_1': $\theta_y > \theta_x$; that is, do lead-poisoned Canadian geese tend to have larger plasma glucose values than healthy geese? Computing the placements for the X and Y observations, we obtain

$$P_1 = 3, \quad P_2 = 4, \quad P_3 = 4, \quad P_4 = 0, \quad P_5 = 0, \quad P_6 = 4, \quad P_7 = 0, \quad P_8 = 1$$

and

$$Q_1 = 4, \quad Q_2 = 4, \quad Q_3 = 3, \quad Q_4 = 8, \quad Q_5 = 8, \quad Q_6 = 8, \quad Q_7 = 5.$$

Thus,

$$\overline{P} = \frac{3+4+4+0+0+4+0+1}{8} = 2$$

and

$$\overline{Q} = \frac{4+4+3+8+8+8+5}{7} = \frac{40}{7}.$$

Using the values in (4.56), we have

$$V_1 = [(3-2)^2 + (4-2)^2 + (4-2)^2 + (0-2)^2 + (0-2)^2 + (4-2)^2 + (0-2)^2 + (1-2)^2] = 26$$

and

$$V_2 = \left[\left(4-\frac{40}{7}\right)^2+\left(4-\frac{40}{7}\right)^2+\left(3-\frac{40}{7}\right)^2+\left(8-\frac{40}{7}\right)^2 + \left(8-\frac{40}{7}\right)^2+\left(8-\frac{40}{7}\right)^2+\left(5-\frac{40}{7}\right)^2\right] = \frac{206}{7}.$$

Combining these quantities, we obtain

$$\widehat{U} = \frac{(40-16)}{2\left[26+\frac{206}{7}+2\left(\frac{40}{7}\right)\right]^{1/2}} = 1.468.$$

From the R commands `cFligPoli(alpha=0.05035),m=8,n=7` and `cFligPolio(alpha= 0.1001, m=8,n=7)`, we find $u_{.05035} = 1.807$ and $u_{.1001} = 1.310$. Thus for these data, the P-value obtained for testing $H_0' : \theta_x = \theta_y$ versus $H_1' : \theta_y = \theta_x$ is between .05035 and .1001. Also see Comment 27.

Comments

23. *Relationship of $\widehat{U}$ to U*. The statistic $\widehat{U}$ defined by (4.57) is of the form

$$\widehat{U} = \frac{n^{1/2}\left\{(U/mn)-\frac{1}{2}\right\}}{\widehat{\sigma}}, \tag{4.63}$$

where U is the Mann–Whitney statistic defined by (4.15) and

$$\widehat{\sigma}^2 = \frac{\sum_{j=1}^{n}(Q_j-\overline{Q})^2+\sum_{i=1}^{m}(P_i-\overline{P})^2+\overline{PQ}}{m^2n}.$$

Fligner and Policello (1981) point out that when written in the form (4.57), namely,

$$\widehat{U} = \frac{\sum_{j=1}^{n}Q_j-\sum_{i=1}^{m}P_i}{2\{V_1+V_2+\overline{P}\,\overline{Q}\}^{1/2}},$$

$\widehat{U}$ resembles Welch's t statistic (Welch 1937, 1947) for the normal theory Behrens–Fisher problem.

24. *Symmetry of the Distribution of $\widehat{U}$*. When H_0: [Identical X and Y distributions] is true, the distribution of $\widehat{U}$ is symmetric about its mean 0, which implies that

$$P_0(\widehat{U} \geq x) = P_0(\widehat{U} \leq -x)$$

for every x. From this, it follows that the lower αth percentile for the null H_0 distribution of $\widehat{U}$ is $-u_\alpha$; hence, its use in the test of H_0' versus H_2' defined by (4.59).

25. *Maintaining Levels*. The test procedures in (4.58), (4.59), and (4.60) have *exact* significance levels equal to α for testing H_0: [Identical X and Y distributions]. However, they also maintain *approximate* level α for the more general null hypothesis $H_0' : \theta_x = \theta_y$, without requiring equal variances or identical distributional forms for the two underlying population.

26. *Consistency of the Test Based on* $\widehat{U}$. Fligner and Policello (1981) consider the consistency of their test based on $\widehat{U}$. To test $H_0 : \theta_x = \theta_y$ versus $\theta_x < \theta_y$, it is necessary to impose conditions on F and G to ensure that whenever $\theta_x = \theta_y$, we have $P(X < Y) = \frac{1}{2}$ and whenever $\theta_x < \theta_y$, we have $P(X < Y) > \frac{1}{2}$. Fligner and Policello point out that a sufficient condition is that F and G be symmetric.
27. Exact Fligner-Policello test. The exact P-value for the Fligner-Policello test can be obtained from the `R` function `pFligPoli`. The `R` function `pFligPoli` computes the P-value of the statistic $\hat{U}$ based on the exact calculations, a Monte Carlo simulation, or the large-sample approximation. By default, the exact calculations are used when $\binom{m+n}{n} \leq 10{,}000$ and a Monte Carlo simulation otherwise. The user may specify which method to use by the `method=option` and, if applicable, the number of Monte Carlo samples to use by the `n.mc=option`. Applying `pFligPoli` to the plasma glucose data of Table 4.7 yields an exact P-value of .0808.

Properties

1. *Consistency*. Assuming F, G are symmetric, the tests defined by (4.58), (4.59), and (4.60) are consistent against the alternatives for which $\theta_x < \theta_y$, $\theta_x > \theta_y$, and $\theta_x \neq \theta_y$, respectively.
2. *Asymptotic Normality*. See Fligner and Policello (1981).
3. *Efficiency*. See Fligner and Policello (1981) and Section 4.5.

Problems

41. Apply the test based on $\widehat{U}$ to the data of Table 4.1. Compare your results with those of Example 4.1.

42. Apply the test based on $\widehat{U}$ to the data of Table 4.2. Compare your results with those of Example 4.2.

43. Apply the test based on $\widehat{U}$ to the data of Table 4.3. Compare your results with those of Problem 1.

44. Apply the test based on $\widehat{U}$ to the data of Table 4.4. Compare your results with those of Problem 5.

45. Apply the test based on $\widehat{U}$ to the data of Table 4.5. Compare your results with those of Problems 15, 16, and 17.

46. Establish (4.63) directly or illustrate it using an example.

47. Show that $\widehat{U}$ is a rank statistic. That is, show that you can compute $\widehat{U}$ from knowledge of $S_1, \ldots, S_n$, where S_j = rank of Y_j in the joint ranking of the N X's and Y's.

4.5 EFFICIENCIES OF TWO-SAMPLE LOCATION PROCEDURES

Recall the normal theory t-test based on

$$t = \frac{\overline{Y} - \overline{X}}{s_p \sqrt{\frac{m+n}{mn}}},$$

where

$$s_p^2 = \frac{\sum_{i=1}^{m}(X_i - \overline{X})^2 + \sum_{j=1}^{n}(Y_j - \overline{Y})^2}{m+n-2}$$

is the pooled variance. The Pitman asymptotic relative efficiency of the test based on W versus the test based on t is

$$E(W,t) = 12\sigma_F^2 \left\{ \int f^2 \right\}^2. \tag{4.64}$$

In (4.64), σ_F^2 is the variance of the population with distribution F and f is the probability density corresponding to F. The parameter $\int f^2$ is the area under the curve of f^2.

Equation (4.64) was derived by Pitman (1948) in the testing context and shown by Hodges and Lehmann (1963) to hold also for the asymptotic relative efficiency of the point estimator $\widehat{\Delta}$ with respect to $\overline{\Delta} = \overline{Y} - \overline{\Delta}$. Lehmann (1963c) showed that (4.64) also gives the asymptotic relative efficiency of the confidence interval derived from W to that of the confidence interval derived from $\overline{Y} - \overline{X}$.

Hodges and Lehmann (1956) showed that for all populations, $E(W,t)$ is at least .864. Thus the most efficiency one can lose when employing the Wilcoxon test instead of the t-test is about 14%. When F is the normal distribution (the home turf of the t-test), $E(W,t) = .955$. For many populations, $E(W,t)$ exceeds 1, and it can be infinite, as it is in the case when F is Cauchy. Some values of $E(W,t)$ are in the following table.

F	Normal	Uniform	Logistic	Double exponential	Cauchy	Exponential
$E(W,t)$	.955	1.000	1.097	1.500	∞	3.00

Some asymptotic relative efficiency values of van der Waerden's c test relative to the test based on W are in the following table.

F	Normal	Uniform	Logistic	Double exponential	Cauchy	Exponential
$E(c,W)$	1.047	∞	.955	.847	.708	∞

These values are also the values of the asymptotic relative efficiency $E(c_1, W)$, where c_1 is the Fisher–Yates–Terry–Hoeffding statistic.

Chernoff and Savage (1958) showed that for all populations, the asymptotic relative efficiency of the Fisher–Yates–Terry–Hoeffding test with respect to the t-test is always greater than or equal to 1. It equals 1 when F is normal.

For model (4.2), the asymptotic relative efficiency of the Fligner–Policello test based on $\widehat{U}$ with respect to W is 1 for all F.