Computer Assistance Computer programs containing the Mann-Whitney Test include *Minitab, S-Plus, SAS,* and *StatXact*.

Comment

The Mann-Whitney test is unbiased and consistent when testing the preceding hypotheses involving $P(X > Y)$. However, the same is not always true for the hypotheses involving $E(X)$ and $E(Y)$. To insure that the test remains consistent and unbiased for hypotheses involving $E(X)$ it is sufficient to add another assumption to the previous model.

Assumption 4. If there is a difference between population distribution functions, that difference is a difference in the location of the distributions. That is, if $F(x)$ is not identical with $G(x)$, then $F(x)$ is identical with $G(x + c)$, where c is some constant.

EXAMPLE 1

The senior class in a particular high school had 48 boys. Twelve boys lived on farms and the other 36 lived in town. A test was devised to see if farm boys in general were more physically fit than town boys. Each boy in the class was given a physical fitness test in which a low score indicates poor physical condition. The scores of the farm boys (X_i) and the town boys (Y_j) are as follows.

X_i: *Farm Boys*		Y_j: *Town Boys*					
14.8	10.6	12.7	16.9	7.6	2.4	6.2	9.9
7.3	12.5	14.2	7.9	11.3	6.4	6.1	10.6
5.6	12.9	12.6	16.0	8.3	9.1	15.3	14.8
6.3	16.1	2.1	10.6	6.7	6.7	10.6	5.0
9.0	11.4	17.7	5.6	3.6	18.6	1.8	2.6
4.2	2.7	11.8	5.6	1.0	3.2	5.9	4.0

Neither group of boys is a random sample from any population. However, it is reasonable to assume that these scores resemble hypothetical random samples from the populations of farm and town boys in that age group, at least for similar localities. The other assumptions of the model seem to be reasonable, such as independence between groups. Therefore the Mann-Whitney test is selected to test

H_0: Farm boys do not tend to be more fit, physically, than town boys

H_1: Farm boys tend to be more fit than town boys

These hypotheses suggest an upper-tailed test as stated in set C of hypotheses.

The scores are ranked as follows.

X	Y	*Rank*	X	Y	*Rank*	X	Y	*Rank*
	1.0	1		6.2	17		11.3	33
	1.8	2	6.3		18	11.4		34
	2.1	3		6.4	19		11.8	35
	2.4	4		6.7	20.5	12.5		36
	2.6	5		6.7	20.5		12.6	37
2.7		6	7.3		22		12.7	38
	3.2	7		7.6	23	12.9		39
	3.6	8		7.9	24		14.2	40
	4.0	9		8.3	25		14.8	41.5
4.2		10	9.0		26	14.8		41.5
	5.0	11		9.1	27		15.3	43
	5.6	13		9.9	28		16.0	44
	5.6	13		10.6	30.5	16.1		45
5.6		13		10.6	30.5		16.9	46
	5.9	15	10.6		30.5		17.7	47
	6.1	16		10.6	30.5		18.6	48

There are four groups of tied scores, as indicated by the square brackets. Within each group the ranks that should have been assigned are averaged, and the average rank is assigned instead, as illustrated.

The test is one tailed. The critical region corresponds to large values of T_1. Note that this is not a large number of ties, so it is probably acceptable to use T instead of T_1. Both methods will be compared later in the example. From Table A1 we see that a critical region of size $\alpha = 0.05$ corresponds to values of T_1 greater than 1.6449.

Here we have $n = 12$, $m = 36$, so $N = 12 + 36 = 48$. The sum of the ranks assigned to the Xs is

$$\begin{aligned} T &= \sum_{i=1}^{n} R(X_i) \\ &= 6 + 10 + 13 + 18 + 22 + 26 + 30.5 + 34 + 36 + 39 + 41.5 + 45 = 321 \end{aligned}$$

The sum of the squares of all 48 ranks is

$$\sum_{i=1}^{N} R_i^2 = 38{,}016$$

which is slightly less than the sum 38,024, of the squares of all the ranks from 1 to 48 if there had been no ties (using Lemma 1.4.2.). Now we can compute T_1.

$$
\begin{aligned}
T_1 &= \frac{T - n\dfrac{N+1}{2}}{\sqrt{\dfrac{nm}{N(N-1)}\displaystyle\sum_{i=1}^{N} R_i^2 - \dfrac{nm(N+1)^2}{4(N-1)}}} \\
&= \frac{321 - 12\dfrac{49}{2}}{\sqrt{\dfrac{(12)(36)}{(48)(47)}(38{,}016) - \dfrac{(12)(36)(49)^2}{4(47)}}} \\
&= 0.6431
\end{aligned}
$$

which is not in the critical region, so H_0 is accepted and we conclude that these data do not show that farm boys are more physically fit than town boys. A comparison of $T_1 = 0.6431$ with Table A1 shows that 0.6431 is close to the 0.74 quantile, so the null hypothesis could have been rejected at an α level of about $1 - 0.74 = 0.26$, and therefore the p-value is 0.26.

If we had ignored the few ties and used the large sample approximation in Equation 5 we would have obtained an approximate 0.95 quantile for T as

$$
\begin{aligned}
w_{0.95} &= n\frac{N+1}{2} + (1.6449)\sqrt{nm(N+1)/12} \\
&= 294 + (1.6449)(42) \\
&= 363.1
\end{aligned}
$$

and H_0 would have been accepted as before. ■

The next example illustrates a situation in which no random variables are defined explicitly. The pieces of flint are ranked according to hardness by direct comparison with each other. A random variable that assigns a measure of hardness to each piece of flint is conceivable but unnecessary in this case.

EXAMPLE 2

A simple experiment was designed to see if flint in area A tended to have the same degree of hardness as flint in area B. Four sample pieces of flint were collected in area A and five sample pieces of flint were collected in area B. To determine which of two pieces of flint was harder, the two pieces were rubbed against each other. The piece sustaining less damage was judged the harder of the two. In this manner all nine pieces of flint were ordered according to hardness. The rank 1 was assigned to the softest piece, rank 2 to the next softest, and so on.

these tests is examined by Mikulski (1963), Basu (1967a), Hollander (1967a), and Gibbons and Gastwirth (1970). Other related papers include Hollander, Pledger, and Lin (1974), Bickel and Lehmann (1975), Hettmansperger and Malin (1975), Doksum and Sievers (1976), and Fligner, Hogg, and Killeen (1976).

The method for finding confidence intervals is discussed by Noether (1967a), and a related graphical procedure is described by Moses in Walker and Lev (1953). An algorithm that may be useful when sample sizes are large is given by McKean and Ryan (1977). Other estimates of location differences are discussed by Hodges and Lehmann (1963), Høyland (1965), Rao, Schuster, and Littell (1975), and Switzer (1976). Related papers are by Moses (1965), Govindarajulu (1968), Bauer (1972), Ury (1972), and Kraft and van Eeden (1972).

EXERCISES

1. Test the following data to see if the mean high temperature in Des Moines is higher than the mean high temperature in Spokane, for randomly sampled days in the summer.

Des Moines	Spokane
83	78
91	82
94	81
89	77
89	79
96	81
91	80
92	81
90	

2. In a controlled environment laboratory, 10 men and 10 women were tested to determine the room temperature they found to be the most comfortable. The results were as follows.

Men	Women
74	75
72	77
77	78
76	79
76	77
73	73
75	78
73	79
74	78
75	80

Assuming that these temperatures resemble a random sample from their respective populations, is the average comfortable temperature the same for men and women?

3. Seven students were taught algebra using the present method, and six students learned