USE OF THE WILCOXON STATISTIC FOR A GENERALIZED BEHRENS-FISHER PROBLEM[1]

BY RICHARD F. POTTHOFF

University of North Carolina

0. Summary. Heretofore, the ordinary Wilcoxon statistic for the two-sample problem [9], [5] has been used only to test the null hypothesis that the two parent populations are identical. This paper presents a technique for utilizing the Wilcoxon statistic to test a broader type of null hypothesis, like that encountered in the Behrens-Fisher problem: we show that the usual Wilcoxon test, with $(m + n + 1)/12mn$ replaced by $1/[4 \min(m, n)]$, may be used to test the null hypothesis of the equality of the medians of two symmetrical (continuous) distributions which are of the same form but which have different (unknown) scale parameters; more generally, the test still works for testing the equality of the medians of any two symmetrical distributions.

1. Introduction and statement of results. We have a sample f m X_i's and a sample of n Y_j's from parent populations with c.d.f.'s $G(x)$ and $H(y)$ respectively. It is known that

$$G(x) = F(b_1x - b_1\theta_1) \quad \text{and} \quad H(y) = F(b_2y - b_2\theta_2), \tag{1.1}$$

where the c.d.f. F is continuous and is symmetric about the origin:

$$F(-w) + F(w) = 1, \qquad -\infty < w < \infty. \tag{1.2}$$

In (1.1), the scale parameters b_1 and b_2 and the medians θ_1 and θ_2 are assumed unknown; the function F may or may not be known. In order to test the null hypothesis

$$H_0 : \theta_1 = \theta_2 \tag{1.3}$$

against alternatives $H_a : \theta_1 \neq \theta_2$, we will utilize the basic Wilcoxon statistic [9], [5]

$$W_{m,n} = (1/mn) \sum_{i=1}^{m} \sum_{j=1}^{n} U(Y_j - X_i), \tag{1.4}$$

where the function $U(d)$ equals 0 if $d \leqq 0$ and 1 if $d > 0$. We may assume $m \leqq n$. We will show that, if we base a test of H_0 (1.3) on the critical region

$$2m^{\frac{1}{2}} \mid W_{m,n} - \tfrac{1}{2} \mid > z_{\alpha/2}, \tag{1.5}$$

then such a test (1.5) will be (approximately) of size α (where "size" means maximum probability of Type I. error) and will for practical purposes be con-

Received March 8, 1962; revised May 24, 1963.

[1] This research was supported partially by the Mathematics Division of the Air Force Office of Scientific Research and partially by Educational Testing Service.

sistent. [By $z_{\alpha/2}$ in (1.5) we are denoting the point above which lies $100(\alpha/2)\%$ of the $N(0, 1)$ distribution.]

More generally, the test (1.5) is still valid under a null hypothesis even broader than the set-up of the generalized Behrens-Fisher problem (1.1–1.3). If, instead of (1.1), we consider a model in which

$$G(x) = F_1(x - \theta_1) \quad \text{and} \quad H(y) = F_2(y - \theta_2), \tag{1.6}$$

where both $F_1(w)$ and $F_2(w)$ are symmetric about $w = 0$ [see (1.2)] and continuous, then the test (1.5) still works for testing H_0 (1.3). The test (1.5) is a conservative test.

2. Proof of results. To prove the results stated in the previous section, we will need to utilize the following five points:

(i) Regardless of what the two parent populations are, the distribution of $[W_{m,n} - E(W_{m,n})]/[\text{var}(W_{m,n})]^{\frac{1}{2}}$ is asymptotically $N(0, 1)$ under rather general conditions. This is well-known. (For proof and statement of conditions, refer to [4], Theorem 3.2 or [2], Theorem 6.1; alternatively, the proof follows easily from Theorem 8.1 of [3].)

(ii) Under the broad model associated with (1.6), the expectation of $W_{m,n}$ (1.4) will always be $\frac{1}{2}$ if H_0 (1.3) is true. In the course of proving this, let us establish a formula for $E(W_{m,n})$ for the case of general δ, where we define $\delta = \theta_2 - \theta_1$. We write

$$\begin{aligned} E(W_{m,n}) &= P\{X < Y\} = \int_{-\infty}^{\infty} F_1(y - \theta_1)\, dF_2(y - \theta_2) \\ &= \frac{1}{2} + \int_0^{\infty} [F_1(u + \delta) - F_1(u - \delta)]\, dF_2(u), \end{aligned} \tag{2.1}$$

the last step being reached through a series of manipulations some of which utilize the symmetry of F_1 and F_2. Now we plug $\delta = 0$ into (2.1) to get what we want.

(iii) We also see from (2.1) that, except for certain trivial types of odd-shaped distributions F_1 and F_2, $E(W_{m,n})$ will be $\neq \frac{1}{2}$ if H_0 (1.3) is not true (i.e., if $\delta \neq 0$). More exactly, $\delta > 0$ implies $E(W_{m,n}) \geqq \frac{1}{2}$ and $\delta < 0$ implies $E(W_{m,n}) \leqq \frac{1}{2}$, but in each case it will be an inequality rather than an equality unless F_1 and F_2 are of such peculiar form as to cause the integral in the last line of (2.1) to vanish. [It is not impossible for this integral to vanish: e. g., let $\delta = \frac{1}{4}$ and take $F_1(w) = F(w)$, $F_2(w) = F(10w)$, where F is defined by $F(w) = 0$, $w \leqq -1$; $F(w) = 1 + w$, $-1 \leqq w \leqq -\frac{1}{2}$; $F(w) = \frac{1}{2}$, $-\frac{1}{2} \leqq w \leqq \frac{1}{2}$; $F(w) = w$, $\frac{1}{2} \leqq w \leqq 1$; $F(w) = 1$, $w \geqq 1$.]

(iv) No matter what the two parent populations are,

$$\text{var}(W_{m,n}) \leqq 1/4m. \tag{2.2}$$

This upper bound (2.2) follows trivially from [1], Formula (2.5), or from an equivalent inequality in [8], p. 3.

(v) For the model based on (1.1),

$$\sup_{0<b_1,b_2<\infty} \mathrm{var}(W_{m,n}) = 1/4m \quad \text{if } H_0 \text{ is true.} \tag{2.3}$$

To prove this, we take note of (2.2) and observe that (2.3) will be established if we show that

$$\lim_{b\to 0} \mathrm{var}(W_{m,n}) = 1/4m \quad \text{if } H_0 \text{ is true,} \tag{2.4}$$

where we define $b = b_1/b_2$. Now the basic formula

$$\begin{aligned} \mathrm{var}\,(W_{m,n}) &= (1/mn)[P\{X < Y\} + (1 - m - n)(P\{X < Y\})^2 \\ &+ (m - 1)P\{X_i < Y, X_I < Y\} + (n - 1)P\{X < Y_j, X < Y_J\} \end{aligned} \tag{2.5}$$

is well-known {see, e.g., [6], [7], Equation (3), or [1], Equation (2.4.2)}. Given (1.1 − 1.3), it is easily seen that $\lim_{b\to 0} P\{X_i < Y, X_I < Y\} = \frac{1}{4}$ and $\lim_{b\to 0} P\{X < Y_j, X < Y_J\} = \frac{1}{2}$; applying these two relations to (2.5), we obtain (2.4).

We now have the five points we need. From (i), (ii), and (v), it follows that, under the generalized Behrens-Fisher model (1.1 − 1.2), the test with critical region (1.5) is a size-α test of H_0 (1.3) (disregarding inaccuracies due to the normal approximation). More generally, it follows from (i), (ii), and (iv) that, even under the broad model associated with (1.6), the test (1.5) is such that its probability of falsely rejecting H_0 (1.3) can never exceed α (again disregarding the approximation). The consistency of the test (1.5) is of course established by utilizing (iii), and employing an argument similar to that given in [5], pp. 58–59.

3. Remarks. (a) Our discussion considered a model and null hypothesis based on (1.1–1.3), and also considered the more general set-up associated with (1.6). If we wish, we can consider a still more general null hypothesis: the test (1.5) can be used to test $P\{X < Y\} = \frac{1}{2}$ against $P\{X < Y\} \neq \frac{1}{2}$. But this latter set-up appears to be so general as to be almost useless.

(b) By way of comparison with (2.2), we may recall that, when $G = H$, $W_{m,n}$ has variance $(m + n + 1)/12mn$ (which is the value used by the ordinary Wilcoxon test [9], [5]).

(c) Although our entire discussion was in terms of a two-tailed test, the extension to one-tailed tests is immediate.

(d) Confidence bounds on δ associated with the test (1.5) are easily obtained.

(e) It does not appear that the test (1.5) is unbiased.

(f) For the case $m = n$, we find from [1], Formula (2.6), that

$$\mathrm{var}\,(W_{n,n}) \geqq (2n + 1)/12n^2 \quad \text{if} \quad P\{X < Y\} = \tfrac{1}{2}, \tag{3.1}$$

it being assumed only that the two parent c.d.f.'s G and H are continuous; furthermore, it can be shown that the equality sign in (3.1) holds if and only if $G = H$. The following implication of (3.1) is of some practical relevance to our present interests: if an experimenter (erroneously) uses the ordinary Wilcoxon test rather than the test (1.5) for testing a null hypothesis of either of the types specified in Section 1, then (if $m = n$) his probability of Type I. error will always exceed the intended value α, unless $G = H$.

4. Acknowledgments. The test presented here evolved subsequent to a conversation with Dr. Frederic M. Lord of Educational Testing Service in which the idea arose as to the possibility of using the Wilcoxon statistic for the Behrens-Fisher problem. The author also wishes to acknowledge helpful discussions with Dr. Y. S. Sathe and with Professor Wassily Hoeffding. Finally, the author wishes to thank the referee for some helpful comments.

REFERENCES

[1] Birnbaum, Z. W. and Klose, O. M. (1957). Bounds for the variance of the Mann-Whitney statistic. *Ann. Math. Statist.* **28** 933–945.

[2] Dwass, Meyer (1956). The large-sample power of rank-order tests in the two-sample problem. *Ann. Math. Statist.* **27** 352–374.

[3] Hoeffding, Wassily (1948). A class of statistics with asymptotically normal distributions. *Ann. Math. Statist.* **19** 293–235.

[4] Lehmann, E. L. (1951). Consistency and unbiasedness of certain nonparametric tests. *Ann. Math. Statist.* **22** 165–179.

[5] Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.* **18** 50–60.

[6] Pitman, E. J. G. (1948). *Lecture Notes on Non-parametric Inference* (mimeographed). Institute of Statistics, University of North Carolina.

[7] Sundrum, R. M. (1953). The power of Wilcoxon's 2-sample test. *J. Roy. Statist. Soc. Ser. B* **15** 246–252.

[8] Van Dantzig, D. (1951). On the consistency and the power of Wilcoxon's two sample test. *Nederl. Akad. Wetensch. Proc. Ser. A* **54** 1–9.

[9] Wilcoxon, Frank (1945). Individual comparison by ranking methods. *Biometrics Bull.* **1** 80–83.