UNIVERSITY OF NORTH CAROLINA
Department of Statistics
Chapel Hill, N. C.

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

by

Richard F. Potthoff

February, 1962

Contract No. AF 49(638)-213

The ordinary Wilcoxon test is used to test the null hypothesis that two populations are identical. This paper presents a technique for utilizing the Wilcoxon statistic to test a null hypothesis of a broader type, such as is encountered in the Behrens-Fisher problem where it is required to test the equality of the means of two normal populations having (possibly) unequal variances. The discussion in this paper is on a rather technical level; for a less technical discussion, see Mimeograph Series No. 316.

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

Institute of Statistics
Mimeo Series No. 315

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

by

Richard F. Potthoff
University of North Carolina

1. Introduction and summary. Let $F(w)$ be any continuous c.d.f. (cumulative distribution function) which is symmetrical about the origin:

$$(1.1) \qquad F(-w) + F(w) = 1, \qquad -\infty < w < \infty \quad .$$

Suppose that we have a sample of m independent observations $X_1, X_2, \ldots, X_m$ from a population with c.d.f. $G(x)$, and a sample of n independent observations $Y_1, Y_2, \ldots, Y_n$ from a population with c.d.f. $H(y)$, where

$$(1.2a) \qquad G(x) = F(b_1 x - b_1 \theta_1)$$

and

$$(1.2b) \qquad H(y) = F(b_2 y - b_2 \theta_2) \quad .$$

The scale parameters b_1 and $b_2 (b_1 > 0, b_2 > 0)$ and the location parameters θ_1 and θ_2 are assumed unknown; the function F may or may not be known. With no loss of generality we may assume that $m \leq n$.

We suppose that it is desired to test the null hypothesis

$$(1.3) \qquad H_o: \theta_1 = \theta_2$$

against alternatives $H_a: \theta_1 \neq \theta_2$. Let us define

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

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

where the function $U(d)$ is 0 or 1 according to whether $d \leq 0$ or $d > 0$ respectively. In Section 3 it will be shown that

$$(1.5) \qquad E(W_{m,n}) = \frac{1}{2} \qquad \text{if and only if } \Theta_1 = \Theta_2$$

(except that the "only if" part is subject to a trivial further restriction on F), and that

$$(1.6) \qquad \sup_{0<b_1,b_2<\infty} \operatorname{var}(W_{m,n}) = 1/4m \qquad \text{whenever } \Theta_1 = \Theta_2 .$$

Since $W_{m,n}$ is (approximately) normal for sufficiently large samples, we will have (approximately) a size $-\alpha$ test of H_o if we use as the critical region

$$(1.7) \qquad 2m^{\frac{1}{2}} \left| W_{m,n} - \frac{1}{2} \right| > z_{\alpha/2} \qquad ,$$

where $z_{\alpha/2}$ is defined by

$$(2\pi)^{-\frac{1}{2}} \int_{-\infty}^{z_{\alpha/2}} e^{-z^2/2} \, dz = 1 - (\alpha/2) \quad .$$

This test (1.7) will actually be a conservative test; that is, the probability of rejecting H_o when H_o is true will generally be somewhat less than α, rather than exactly equal to α.

The statistic $W_{m,n}$ (1.4) is of course the basic element of the well-known Wilcoxon test [5, 9]. Heretofore, however, this Wilcoxon statistic $W_{m,n}$ has been used only for testing the null hypothesis $G = H$ (G being the c.d.f. of the X_i's and H being the c.d.f. of the Y_j's) against alternatives of the form $H(y) = G(y + \delta)$ where $\delta \neq 0$ (or, more generally, against alternatives of a form such that $p \neq \frac{1}{2}$, where we are defining $p = P\{X < Y\}$), and has not been used for testing a hypothesis of the form specified by (1.1, 1.2, 1.3). The variance of $W_{m,n}$ when $G = H$ is $(m + n + 1)/12mn$, which is of course smaller than (1.6). The asymptotic normality of $W_{m,n}$, alluded to in the preceding paragraph, has already been proved; regardless of whether $G = H$ or $G \neq H$, and regardless of what G and H are (so long as $0 < p < 1$), the distribution of

$$(1.8) \qquad \frac{W_{m,n} - E(W_{m,n})}{[\mathrm{var}\,(W_{m,n})]^{\frac{1}{2}}}$$

approaches a normal distribution with zero mean and unit variance as $m, n \to \infty$, provided that $m, n \to \infty$ in such a way that the ratio $m/n \to$ some constant c. (For proof, refer to [4, Theorem 3.2] or [2, Theorem 6.1]; alternatively, the proof follows easily from Theorem (8.1) of [3].)

In the special case where F in (1.2) is known to be the normal c.d.f., the problem we are considering reduces to the Behrens-Fisher problem. Since originally it was the Behrens-Fisher problem which motivated this paper, and since this problem is of special interest, we will first give, in Section 2, some brief remarks pertaining to it.

2. <u>The special case of the Behrens-Fisher problem.</u> The m X_i's are $N(\Theta_1, \sigma_1^2)$, and the n Y_j's are $N(\Theta_2, \sigma_2^2)$, where $\sigma_1 = 1/b_1$, $\sigma_2 = 1/b_2$. F is the $N(0, 1)$ c.d.f. Note first that

$$(2.1) \quad t = \frac{(m+n-2)^{\frac{1}{2}}\,(\overline{X} - \overline{Y})}{\left[\left(\frac{1}{m} + \frac{b^2}{n}\right) \Sigma (X_i - \overline{X})^2 + \left(\frac{1}{n} + \frac{1}{mb^2}\right) \Sigma\, (Y_j - \overline{Y})^2\right]^{1/2}},$$

where $\overline{X}$ and $\overline{Y}$ are the means of the X_i's and Y_j's respectively and where $b = \sigma_2/\sigma_1 = b_1/b_2$, follows the t-distribution (under H_o) with $(m+n-2)$ degrees of freedom; but we cannot use (2.1) as a statistic for testing H_o (1.3) due to the fact that (2.1) is not free of the nuisance parameter b. Now we might be inclined to try to approach the Behrens-Fisher problem by maximizing the denominator of (2.1) with respect to b, since this would lead us to the minimum possible (absolute) value of t and would give us a conservative test valid for any value of the unknown parameter b. It quickly becomes apparent, however, that this approach doesn't work, due to the fact that the denominator of (2.1) becomes infinite as $b \rightarrow 0$ or as $b \rightarrow \infty$.

Consider what happens, though, if we use this same approach with respect to the statistic (1.8). Under H_o the denominator of (1.8), like the denominator of (2.1), depends on b [see (3.10)]. But when we try to maximize this denominator of (1.8) with respect to b, the maximum turns out to be finite [see (1.6)].

Thus, under H_o, the denominator of (2.1) is unbounded with respect to b, whereas the maximum value of the denominator of (1.8) with respect to b is finite. It is this fact which provided the key to the basic approach used in this paper.

The general proofs of both (1.5) and (1.6) given in Section 3 are of course fully adequate to cover the special case in which F is normal. As for (1.6), we rely strictly on the proof of Section 3. We may note, though, that (1.5) has a quick proof when F is normal. By (1.4), $E(W_{m,n}) = p$ (recall that we defined $p = P\{X < Y\}$). Now the variate $(Y_j - X_i)$ is normally distributed with mean $(\Theta_2 - \Theta_1)$, and hence has median $(\Theta_2 - \Theta_1)$. Thus $p = \frac{1}{2}$ if and only if $\Theta_1 = \Theta_2$, which establishes (1.5).

It appears from rough preliminary investigation that Scheffé's test [7] for the Behrens-Fisher problem (which is a t-test with $m-1$ degrees of freedom) generally has better power than the test proposed here (1.7). However, the advantage of the test (1.7) lies in the fact that it is still valid for testing the equality of the Θ's even when F is non-normal (our test of course even works if F is the Cauchy distribution), whereas the validity of Scheffé's test would be questionable if F were not normal. Thus, if data have been obtained under a model of the form (1.1 - 1.2), then Scheffé's test appears to be better if the experimenter is sure that F is normal; but if the experimenter is not certain about F being normal, then it might be safer to use the test (1.7).

3. The general case. In this section we prove the relations (1.5) and (1.6) for general F. To establish (1.5), we first write

$$(3.1)\quad E(W_{m,n}) = p = \int_{-\infty}^{\infty} G(y)\, dH(y)$$

$$= \int_{-\infty}^{\infty} F(b_1 y - b_1\,\theta_1)\, dF(b_2 y - b_2\,\theta_2)$$

$$= \int_{-\infty}^{\infty} F(bu + b_1\delta)\, dF(u)$$

where $b = b_1/b_2$, $\delta = \theta_2 - \theta_1$, $u = b_2 y - b_2\,\theta_2$. Next we have

$$(3.2)\quad \int_{-\infty}^{0} F(bu + b_1\,\delta)\, dF(u) = \int_{\infty}^{0} F(-bv + b_1\delta)\, dF(-v)$$

$$= \frac{1}{2} - \int_{0}^{\infty} F(bv - b_1\delta)\, dF(v)\ ,$$

where we first set $u = -v$ and then applied (1.1) thrice. Substituting (3.2) into (3.1), we obtain

$$(3.3)\qquad p = \frac{1}{2} + \int_{0}^{\infty} \left[F(bu + b_1\delta) - F(bu - b_1\delta) \right] dF(u)\quad .$$

If $\delta = 0$, (3.3) becomes $p = \frac{1}{2}$. If $\delta > 0$, (3.3) tells us that $p \geq \frac{1}{2}$, and if

$\delta < 0$, then $p \leq \frac{1}{2}$; in each case, the equality sign holds only if the integral on the right-hand side of (3.3) is 0. Thus, for $\delta \neq 0$ to imply $p \neq \frac{1}{2}$, we must impose the condition that this integral be non-vanishing. This is a somewhat trivial condition, however; only a rather odd-shaped F would fail to satisfy it. [E.g., the distribution $F(w) = 0$, $w \leq -1$; $F(w) = 1 + w$, $-1 \leq w \leq -\frac{1}{2}$; $F(w) = \frac{1}{2}$, $-\frac{1}{2} \leq w \leq \frac{1}{2}$; $F(w) = w$, $\frac{1}{2} \leq w \leq 1$; $F(w) = 1$, $w \geq 1$ fails to satisfy the condition if $b = 10$, $b_1 = 1$, $\delta = 1$.]

Having proved (1.5), we now turn to (1.6). We will start by establishing a result somewhat more general than what we actually need for proving (1.6). For the time being we are going to assume that G and H may be any c.d.f.'s [not necessarily satisfying (1.1, 1.2, 1.3)]. Then we have the well-known formula (see [8] or [6])

$$(3.4) \quad \mathrm{var}(W_{m,n}) = (1/mn)[p+(1-m-n)p^2 + (m-1)P\{X_\alpha < Y, X_\beta < Y\}+(n-1)P\{X < Y_\gamma, X < Y_\nu\}].$$

Now let E_{ij} denote the event $X_i < Y_j$, and consider two observations X_1, X_2 from G and two observations Y_1, Y_2 from H. Then

$$(3.5) \quad \begin{aligned} P\{X_\alpha < Y, X_\beta < Y\}+P\{X < Y_\gamma, X < Y_\nu\} &= P\{E_{11}E_{21}\}+P\{E_{11}E_{12}\} \\ &= P\{(E_{11}E_{21})\cup(E_{11}E_{12})\}+P\{(E_{11}E_{21})(E_{11}E_{12})\} \\ &= P\{(E_{11})(E_{21}\cup E_{12})\}+P\{(E_{11})(E_{21}E_{12})\} \\ &\leq P\{E_{11}\} + P\{E_{21}E_{12}\} = p + p^2 . \end{aligned}$$

Also,

$$(3.6) \quad P\{X < Y_\gamma, X < Y_\nu\} = P\{E_{11}E_{12}\} \leq P\{E_{11}\} = p .$$

Recalling that $n-m \geq 0$ and substituting (3.5) and (3.6) into (3.4), we obtain

$$\mathrm{var}(W_{m,n}) \leq (1/mn) [p + (1-m-n)\, p^2 + (m-1)(p + p^2) + (n-m)\, p] ,$$

or

$$(3.7) \quad \mathrm{var}(W_{m,n}) \leq p(1 - p)/m \quad .$$

We return now to the model specified by (1.1, 1.2) with null hypothesis (1.3). Under H_o (1.3), we have $p = \frac{1}{2}$, and so (3.7) tells us that

$$(3.8) \quad \mathrm{var}(W_{m,n}) \leq 1/4m \qquad \text{whenever } \Theta_1 = \Theta_2 .$$

[Obviously $\mathrm{var}(W_{m,n})$ cannot exceed $1/4m$ if H_o is false, either.] We will next prove that, regardless of what F in (1.2) is,

$$(3.9) \quad \lim_{b \to o} \mathrm{var}(W_{m,n}) = 1/4m \qquad \text{whenever } \Theta_1 = \Theta_2 .$$

Since we have no knowledge of what b is, (3.9) tells us that the bound $1/4m$ in (3.8) cannot be improved upon (i.e., reduced) for any F. The relations (3.8) and (3.9) together will be sufficient to establish (1.6).

To prove (3.9), we first use (3.4) and (1.2) (setting $p = \frac{1}{2}$ and $\Theta_1 = \Theta_2$) to obtain

$$(3.10) \quad \mathrm{var}(W_{m,n}) = \frac{1}{mn} \Big[\frac{3-m-n}{4} + (m-1)\int G^2 dH + (n-1)\Big(\int dG - 2\int H dG + \int H^2 dG\Big)\Big]$$

$$= \frac{1}{mn} \Big[\frac{3-m-n}{4} + (m-1) \int_{-\infty}^{\infty} F^2(b_1 y - b_1\Theta_1)\, dF(b_2 y - b_2\Theta_1)$$

$$+ (n-1) \int_{-\infty}^{\infty} F^2(b_2 x - b_2\Theta_1)\, dF(b_1 x - b_1\Theta_1)\Big]$$

$$= \frac{1}{mn} \Big[\frac{3-m-n}{4} + (m-1) \int_{-\infty}^{\infty} F^2(bv)\, dF(v) + (n-1)\Big\{\int_{-\infty}^{0} F^2(u/b)\, dF(u) +$$

$$+ \int_0^\infty F^2(u/b)dF(u)\}\_7,$$

where $b_2y - b_2\theta_1 = v$, $b_1x - b_1\theta_1 = u$. Recalling that F is continuous and referring to /‾1, p. 67, (I)_7, we obtain from (3.10)

$$\lim_{b \to 0} \text{var}(W_{m,n}) = \frac{1}{mn} \left[\frac{3-m-n}{4} + (m-1)(\tfrac{1}{4}) + (n-1)\{0 + \tfrac{1}{2}\} \right] .$$

This proves (3.9).

Now (1.5) and (1.6) are sufficient to establish that the test (1.7) is (approximately) of size α. For, since $W_{m,n}$ is (approximately) normal, and has expectation $\frac{1}{2}$ under H_o, we have (approximately)

$$P\{|W_{m,n} - \tfrac{1}{2}| > [\text{var}(W_{m,n})]^{\frac{1}{2}} z_{\alpha/2}\} = \alpha ,$$

i.e.,

$$(3.11)\quad P\{2m^{\frac{1}{2}} |W_{m,n} - \tfrac{1}{2}| > \left[\frac{\text{var}(W_{m,n})}{1/4m} \right]^{\frac{1}{2}} z_{\alpha/2}\} = \alpha ,$$

under H_o. Comparing (1.7) with (3.11), and noting that /‾by (1.6)_7 the supremum of the coefficient of $z_{\alpha/2}$ in (3.11) is 1, we conclude that (under H_o) the supremum of the probability of the event (1.7) is (approximately) α.

4. Concluding remarks. We make some final observations:

(i) Sundrum /‾8_7 presents a result which is of the same form as (3.5) /‾and also gives a relation of the form (3.7) for the special case $m = n$_7, but his rather unusual proof of (3.5) seems to assume a special type of discrete distribution. Although the idea behind his proof would appear intuitively to apply to other distributions also, it would not seem that such an extension, to a class of distributions wider than the class for which his proof was originally intended, would be completely rigorous. Thus the proof in (3.5), in addition to being short, makes no restrictions on G or H.

(ii) If $m,n \rightarrow \infty$ in such a way that $m/n \rightarrow c$, the test (1.7) will be consistent against all alternatives such that $p \neq \frac{1}{2}$ [recall that (3.3) showed that, with certain trivial exceptions, $H_a: \theta_1 \neq \theta_2$ implies $p \neq \frac{1}{2}$]. The consistency property can be proved (e.g.) by using an argument similar to that given in [5, pp. 58-59].

(iii) It does not appear that the test (1.7) is unbiased. This of course is (indirectly) a consequence of the fact that the probability of rejection when H_o is true will not be constant, but rather will vary with b.

(iv) Confidence bounds on δ $(\delta = \theta_2 - \theta_1)$ associated with the test (1.7) are easily obtained, using exactly the same technique as is commonly used with the ordinary Wilcoxon test. [The technique consists of finding that value of δ which, when subtracted from each $(Y_j - X_i)$ in (1.4), will cause the resulting new $W_{m,n}$ to be on the threshhold of being significantly small or large.]

(v) The discussion in this paper was in terms of a two-tailed test. Obviously, it could just as easily have been in terms of a one-tailed test.

(vi) The ordinary Wilcoxon test (for testing G = H) is the same as the test (1.7) except that the factor $2m^{\frac{1}{2}}$ $[= (1/4m)^{-\frac{1}{2}}]$ in (1.7) is replaced by $[(m + n + 1)/12\, mn]^{-\frac{1}{2}}$. By comparing these two factors, we can get some inkling as to the price we have to pay for eliminating the nuisance parameter b when we use the test (1.7). For example, if $m = n$, $1/4m$ is not quite $50^o/o$ larger than $(m + n + 1)/12mn$; if $m < n$, the comparison of course is worse (i.e., $\geq 50^o/o$).

(vii) We mention now the possibility of using the test (1.7) for testing null hypotheses broader than the null hypothesis indicated by (1.1, 1.2, 1.3).

For example, suppose that the specifications (1.2) are replaced by the less restrictive assumptions

(4.1a) $$G(x) = F_1(x - \theta_1)$$

and

(4.1b) $$H(y) = F_2(y - \theta_2) \quad ,$$

where both $F_1(w)$ and $F_2(w)$ are assumed to be symmetrical about $w = 0$ [see (1.1)]. Then the proof given in the first paragraph of Section 3 goes through much the same as before, and in place of (3.3) we end up with

(4.2) $$p = \frac{1}{2} + \int_0^\infty [F_1(u + \delta) - F_1(u - \delta)] \, d\, F_2(u) \quad .$$

From (4.2) we see that $\delta = 0$ implies $p = \frac{1}{2}$, and (with trivial exceptions) $\delta \neq 0$ implies $p \neq \frac{1}{2}$.

Now obviously [see (3.7)] $\mathrm{var}(W_{m,n})$ cannot exceed $1/4m$ even under the broader null hypothesis that we are now considering. Hence (1.7) will be a satisfactory test of $\theta_1 = \theta_2$ even under the broader model associated with the modification (4.1), in the sense that (a) the probability of rejecting when $\theta_1 = \theta_2$ will never exceed α (we are disregarding inaccuracies due to the normal approximation), and (b) the test will be consistent against all alternatives $\theta_1 \neq \theta_2$ (with certain trivial exceptions).

Thus, if G and H are any two (continuous) symmetrical distributions, not necessarily of the same form, then (1.7) can be used to test the hypothesis that their medians are equal.

(viii) If we wish, we can broaden the null hypothesis still further: we can use (1.7) to test the null hypothesis $p = \frac{1}{2}$ against alternatives $p \neq \frac{1}{2}$. However, such a broad null hypothesis as $p = \frac{1}{2}$ might not be very meaningful in most practical situations.

(ix) An upper bound for $var(W_{m,n})$ is given by (3.7). We might note that an interesting lower bound for $var(W_{m,n})$ can be obtained for the particular case $m = n$, $p = \frac{1}{2}$. For this case, (3.4) becomes

$$(4.3) \quad var(W_{n,n}) = (1/n^2) \left[\frac{3-2n}{4} + (n-1) \left(\int G^2 \, d\, H + \int H^2 \, d\, G \right) \right] \quad .$$

(We assume that G and H are continuous.) We can write

$$(4.4) \int G^2 dH + \int H^2 dG = \frac{1}{2} \int G^2 dH + \frac{1}{2} \int H^2 dG + \frac{1}{2} \int (G^2 + H^2) \, d\, (G + H) - \frac{1}{3}$$

$$= 1 - \frac{1}{2} \int H dG^2 - \frac{1}{2} \int G dH^2 + \int (G^2 + H^2) \, d\, \left(\frac{G + H}{2}\right) - \frac{1}{3}$$

$$= \frac{2}{3} + \int (G - H)^2 \, d\, \left(\frac{G + H}{2}\right) \quad .$$

Lehmann [4, pp. 172-173] proves that the non-negative integral appearing in the last line of (4.4) is equal to 0 if and only if $G = H$. Hence, after combining (4.3) and (4.4), we conclude that

$$(4.5) \qquad var(W_{n,n}) \geq (2n + 1)/12n^2 \qquad \text{if} \quad p = \frac{1}{2} \; ,$$

with the equality sign holding if and only if $G = H$.

One implication of (4.5) is that, if an experimenter (erroneously) uses the ordinary Wilcoxon test rather than the test (1.7) for testing a null hypothesis of the form (1.1, 1.2, 1.3) [or of the form associated with the modification (4.1)] then (if $m = n$) his Type I. error will always exceed the intended value α, unless $G = H$.

(x) The line of thought on which this paper was based, expressed in the first three paragraphs of Section 2, has been extended to a problem involving comparisons of two regression lines having unequal variances. This problem will be discussed in other papers.

5. Acknowledgments. The idea as to the possibility of using the Wilcoxon statistic for the Behrens-Fisher problem arose in a conversation with Frederic M. Lord of Educational Testing Service. The author also wishes to acknowledge helpful discussions with Y. S. Sathe and with Professor Wassily Hoeffding.

REFERENCES

[1] Harald Cramér, Mathematical Methods of Statistics, Princeton University Press, Princeton, New Jersey, 1951.

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

[3] W. Hoeffding, "A class of statistics with asymptotically normal distributions," Ann. Math. Stat., vol.19 (1948), pp. 293-325.

[4] E. L. Lehmann, "Consistency and unbiasedness of certain nonparametric tests," Ann. Math. Stat., vol. 22 (1951), pp. 165-179.

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

[6] E. J. G. Pitman, "Lecture notes on non-parametric inference"(mimeographed), Institute of Statistics, University of North Carolina, 1948.

[7] Henry Scheffé, "On solutions of the Behrens-Fisher problem, based on the t-distribution," Ann. Math. Stat., vol. 14 (1943), pp. 35-44.

[8] R. M. Sundrum, "The power of Wilcoxon's 2-sample test," J. Roy. Stat. Soc., Ser. B, vol. 15 (1953), pp. 246-252.

[9] F. Wilcoxon, "Individual comparison by ranking methods," Biometrics, vol. 1 (1945), pp. 80-83.