THE TREATMENT OF TIES IN THE WILCOXON TEST[1]

By Wolfgang J. Bühler

University of California, Berkeley

1. Introduction. Let $(X_1, \cdots, X_n)$ be a sample of n independent observations from a distribution F, and $(Y_1, \cdots, Y_m)$ be a sample of independent observations from G. Then, if all $m + n$ observations are different, the Wilcoxon test will reject the hypothesis $F = G$, when the sum S_{nm} of the ranks R_i of the X_i is too small or too large.

For the case with a positive probability of ties two procedures have been proposed. One is to order the tied observations randomly, the other is to replace S_{nm} by $S'_{nm} = \sum_{i=1}^{n} R_i'$. Here $R_i' = \text{midrank } (X_i) = \frac{1}{2}[N_1(i) + N_2(i) + 1]$. $N_1(\mathrm{i})$ is the number of observations smaller than X_i and $N_2(\mathrm{i})$ is the number of observations (including X_i) not larger than X_i.

If there are only finitely many values ξ_k at which ties may occur and if $p_k = P\{X_1 = \xi_k\}$, then as shown by Putter [3] under certain regularity conditions the asymptotic relative efficiency of the "randomized" with respect to the midrank test is $1 - \sum_{k=1}^{n} p_k^3$. Using a slight modification of Putter's argument this note will show that this conclusion is still true if $p_k = P\{X_1 = \xi_k\} > 0$ and $q_k = P\{Y_1 = \xi_k\} > 0$ for infinitely many values ξ_k. The result is illustrated by applying it to certain parametric families of distributions, for which the efficiency of the midrank test has been investigated by Chanda [1]. Putter's notation will be used throughout the paper.

2. The basic theorem. Following Putter, let for

$$k = 1, 2, \cdots, p_k = P\{X_1 = \xi_k\} > 0, \qquad q_k = P\{Y_1 = \xi_k\} > 0;$$

U_k = number of X's equal to ξ_k, V_k = number of Y's equal to ξ_k; $U = (U_1, U_2, \cdots), V = (V_1, V_2, \cdots), W = U + V; S^0_{nm}$ = any statistic whose distribution is that of S_{nm} under $F = G$; $\mu_{nm} = ES^0_{nm} = n(n + m + 1)/2$, $\sigma^2_{nm} = \operatorname{Var} S^0_{nm} = nm(n + m + 1)/12$; $T^0_{nm} = (S^0_{nm} - \mu_{nm})/\sigma_{nm}$.

Then the following theorem connects the asymptotic distributions of S_{nm} and of S'_{nm}.

Theorem 1. *If m/n converges to a positive number c as $m, n \to \infty$, then we have for any pair* $(F, G,$ *of distributions with common discontinuities* ξ_k, $k = 1, 2, \cdots$

$$\sigma^2_{U_kV_k}/\sigma^2_{nm} = a_k^2 \to_P b_k^2 = (1 + c)^{-1} p_k q_k(\theta)[p_k + c q_k(\theta)] \tag{2.1}$$

$$(S_{nm} - ES_{nm})/\sigma_{nm} = T_{nm} \to_{\mathfrak{L}} N(0, b^2) \tag{2.2}$$

$$(S'_{nm} - E\mathrm{S}_{nm})/\sigma_{nm} = T'_{nm} \to_{\mathfrak{L}} N(0, \bar{b}^2), \tag{2.3}$$

Received 25 July 1966; revised 3 September 1966.

[1] Prepared with the partial support of U. S. Public Health Grant GM-10525-03.

where the variances b^2 and $\bar{b}^2$ satisfy the relation

$$(2.4) \qquad \bar{b}^2 = b^2 - \textstyle\sum_{k=1}^{\infty} b_k^{\,2}.$$

PROOF.[2] Let $c(u) = \frac{1}{2} + \frac{1}{2}\operatorname{sgn}(u)$ and $d(u) = 1$ or 0 according to $u = 0$ or $u \neq 0$ respectively. Define $X_{n+j} = Y_j\,(j = 1, 2, \cdots, m)$ and let $Z_1, Z_2, \cdots, Z_{n+m}$ be mutually independent and independent of $X_1, X_2, \cdots, X_{n+m}$ and let the Z_i, have a common continuous (otherwise arbitrary) distribution. Then

$$R_i = \tfrac{1}{2} + \textstyle\sum_{j=1}^{m+n} \{[1 - d(X_i - X_j)]c(X_i - X_j) + d(X_i - X_j)c(Z_i - Z_j)\},$$

$$R_i' = \tfrac{1}{2} + \textstyle\sum_{j=1}^{m+n} c(X_i - X_j), \quad \text{and therefore}$$

$$\begin{aligned} S'_{nm} &= \textstyle\sum_{i=1}^{n} R_i' = \sum_{i=1}^{n} \{\frac{1}{2} + \sum_{j=1}^{n} c(X_i - X_j) + \sum_{j=n+1}^{n+m} c(X_i - X_j)\} \\ &= n(n+1)/2 + \textstyle\sum_{i=1}^{n} \sum_{j=1}^{m} c(X_i - X_{n+j}) \\ &\qquad\qquad = \textstyle\sum_{i=1}^{n} \sum_{j=1}^{m} \{(n+1)/2m + c(X_i - Y_j)\}, \end{aligned}$$

and similarly S_{nm}, is a two-sample U statistic. Thus (2.2) and (2.3) follow immediately (Lehmann [2]). To establish (2.4) we note that

$$T_{nm} = T'_{nm} + \textstyle\sum_{k=1}^{\infty} a_k T^0_{U_k V_k},$$

where all the summands on the right hand side are conditionally independent given (U, V). This implies $\operatorname{Var}(T_{nm} \mid U, V) = \operatorname{Var}(T'_{nm} \mid U, V) + \sum_{k=1}^{\infty} a_k^{\,2}$. Using, that $ES_{nm} = ES'_{nm}$ and even $E(S_{nm} \mid U, V) = E(S'_{nm} \mid U, V) = S'_{nm}$, it can be seen that

$$\begin{aligned} (2.5) \quad & \operatorname{Var}(T_{nm}) - \operatorname{Var}(T'_{nm}) \\ &\qquad = E\{\operatorname{Var}(T_{nm} \mid U, V) - \operatorname{Var}(T'_{nm} \mid U, V)\} = E\{\textstyle\sum_{k=1}^{\infty} a_k^{\,2}\}. \end{aligned}$$

Finally we let n tend to infinity in (2.5) to prove the relation (2.4).

3. The conclusions. As in Putter [3] we obtain the following immediate consequence of Theorem 1:

THEOREM 2. *If $F = G$, then $S'_{nm} - \mu_{nm}/\sigma_{nm} \to_{\mathfrak{L}} N(0, 1 - \sum p_k^{\,3})$ as $n, m \to \infty$. Therefore, if $s_{nm}(U, V)$ is any sequence of statistics with $s^2_{nm}(U, V)/\sigma^2_{nm} \to_P 1 - \sum_{k=1}^{\infty} p_k^{\,3}$, then $S'_{nm} - \mu_{nm}/s_{nm}(U, V) \to_{\mathfrak{L}} N(0, 1)$ as $n, m \to \infty$.*

Now we can state Putter's result about the asymptotic relative efficiency for the case of infinitely many points ξ_k where ties may occur.

THEOREM 3. *Let m/n converge to a fixed number $c > 0$ and let $\{G_\theta, 0 \leqq \theta \leqq \theta_1\}$ be a family of purely discontinuous distributions all having the same discontinuities $\xi_1 < \xi_2 < \cdots$ with jumps $q_k(\theta)$. Let $s_{nm}(U, V)$ be functions of U and V having, under each G_θ, finite variances, and let the following conditions be satisfied:*

(1) *$q_k(\theta) \geqq q_k > 0$, $q_k(0) = p_k$, $k = 1, 2, \cdots$;*

(2) *if X has distribution $F = G_0$, Y has distribution G_θ, then $\theta = P(X > Y) + \frac{1}{2}P(X = Y) - \frac{1}{2}$;*

[2] I am indebted to Professor W. Hoeffding for pointing out this proof which is much simpler than my original one.

(3) $(S_{nm} - ES_{nm})/\sigma_{nm} \to_{\mathcal{L}} N(0, b^2(\theta))$ *uniformly in* θ;

(4) *the functions* $q_k(\theta)$ *are continuous at* $\theta = 0$;

(5) $s_{nm}(U, V)/n^{\frac{3}{2}} = \sum_{k=1}^{\infty} \alpha_k(\theta)(U_k - np_k) + \sum_{k=1}^{\infty} \beta_k(\theta)(V_k - mq_k(\theta)) + \gamma(\theta)n + o_p(n^{\frac{1}{2}})$;

(6) $\gamma^2(0) = [c(1 + c)/12](1 - \sum_{k=1}^{\infty} p_k^3)$;

(7) $\gamma(\theta)$ *is differentiable,* $\gamma'(\theta)$ *is continuous at* 0;

(8) *At least one of the inequalities* $c\theta\alpha_k(\theta) \neq \gamma(\theta)\bar{\alpha}_k(\theta)$, $c\theta\beta_k(\theta) \neq \gamma(\theta)\bar{\beta}_k$ *holds where* $\bar{\alpha}_k(\theta) = 1 + c[\sum_{j<k} q_j(\theta) + \frac{1}{2}q_k(\theta)]$, $\bar{\beta}_k = \sum_{j<k} p_j + \frac{1}{2}p_k$.
Under these conditions the asymptotic relative efficiency of the randomized with respect to the nonrandomized test is $R = 1 - \sum_{k=1}^{\infty} p_k^3$.

The proof of Theorem 3 follows the lines of Putter's proof using that the convergence in the proof of Theorem 1 is uniform in θ. It can be seen that (3) holds whenever $b^2(\theta)$ is bounded away from zero. Also, modifying remark (i) of Putter's paper, one shows that conditions (5), (6) and (8) are satisfied e.g., when s_{nm} is given by

$$(3.1)\quad s_{nm}^2(U, V) = nm(n + m + 1)/12 - \sum_{k=1}^{\infty} U_k V_k(U_k + V_k + 1)/12.$$

In this case we have

$$\alpha_k(\theta) = -(c/24\gamma(\theta))q_k(\theta)[2p_k + cq_k(\theta)]$$
$$\beta_k(\theta) = -(c/24\gamma(\theta))p_k[p_k + 2cq_k(\theta)]$$
$$\gamma^2(\theta) = c[1 + c - \sum p_k q_k(\theta)(p_k + cq_k(\theta))]/12.$$

4. Illustrations. Using the above remarks it is easy to verify that the result of Theorem 3 can be applied to many parametric families of distributions (a reparametrization may be needed to satisfy (2)). In particular we shall apply it to the examples considered by Chanda [1]. For this purpose let us denote by e the asymptotic efficiency of the midrank test relative to the best parametric test and by E = Re the asymptotic efficiency of the randomized test. All values of e given in the following, in particular the numerical values in Table 1 are taken from Chanda [1].

EXAMPLE 1. *Poisson distribution with parameter* λ.

TABLE 1

	0	0.2	0.5	1.0	3.0	∞
e	1	0.92	0.91	0.92	0.94	0.95
E	0	0.42	0.68	0.82	0.91	0.95

EXAMPLE 2. Binomial distribution with parameter $p = P(X = 1)$. $e = 1$, $E = R = 1 - p^3 - (1 - p)^3 = 3p(1 - p)$. Thus E is zero at $p = 0$ and at $p = 1$ and takes its maximum $\frac{3}{4}$ at $p = \frac{1}{2}$.

EXAMPLE 3. *Geometric distribution with parameter* p. $e = (1 + p + p^2)/(1 + p)^2$, $R = 3p/(1 + p + p^2)$, $E = 3p/(1 + p)^2$. At $p = 0$ we have $e = 1$

and $E = 0$. As p increases to 1, e is monotone decreasing to $\frac{3}{4}$ whereas E is monotone increasing to the same value $\frac{3}{4}$. As should be expected these examples indicate that the loss of efficiency when using the randomized procedure is the more severe the more the distribution is concentrated in a few points.

5. Acknowledgment. The author wishes to thank Professor E. L. Lehmann for suggesting the subject of this paper and for his advice and encouragement during the course of the work.

REFERENCES

[1] CHANDA, K. C. (1963). On the efficiency of two-sample Mann-Whitney test for discrete populations. *Ann. Math. Statist.* **34** 612–617.

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

[3] PUTTER, J. (1955). The treatment of ties in some nonparametric tests. *Ann. Math. Statist.* **26** 368–386.