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 $R_i' = \text{midrank}(X_i) = \frac{1}{2}[N_1(i) + N_2(i) + 1]$. $N_1(\text{i})$ is the number of observations smaller than X_i and $N_2(\text{i})$ 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 $F = G$; $\mu_{nm} = ES^0_{nm} = n(n + m + 1)/2$, $\sigma^2_{nm} = \text{Var}\, S^0_{nm} = nm(n + m + 1)/12$; $T^0_{nm} = (S^0_{nm} - \mu_{nm})/\sigma_{nm}$.

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, $k = 1, 2, \cdots$

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

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

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

Received 25 July 1966; revised 3 September 1966.

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