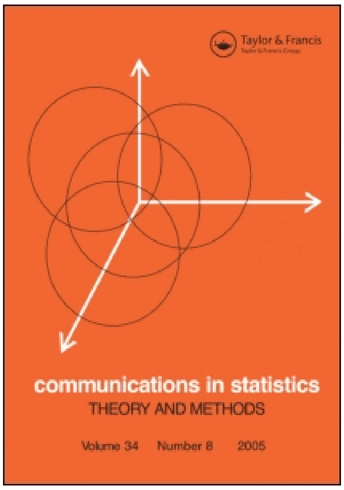

Communications in Statistics - Theory and Methods
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597238

A generalized wilcoxon-mann-whitney statistic

Carey E. Priebe[a]; Lenore J. Cowen[a]
[a] Department of Mathematical Sciences, The Johns Hopkins University, Baltimore, MD, USA

To cite this Article Priebe, Carey E. and Cowen, Lenore J.(1999) 'A generalized wilcoxon-mann-whitney statistic', Communications in Statistics - Theory and Methods, 28: 12, 2871 — 2878
To link to this Article: DOI: 10.1080/03610929908832454
URL: http://dx.doi.org/10.1080/03610929908832454

COMMUN. STATIST.—THEORY METH., 28(12), 2871–2878 (1999)

A GENERALIZED WILCOXON-MANN-WHITNEY STATISTIC

Carey E. Priebe and Lenore J. Cowen

Department of Mathematical Sciences
The Johns Hopkins University
Baltimore, MD 21218-2682 USA

Key words: U-statistic; rank statistic; stochastic ordering; permutation test; subsample test; distribution-free test; Pitman's asymptotic relative efficiency; exact distribution; recurrence; generating function; discriminant analysis; classification.

ABSTRACT

We develop a simple but useful generalization of the classical Wilcoxon-Mann-Whitney statistic. A normal approximation and a recurrence for the exact distribution of this generalization are available. The statistic has potential application in nonparametric discriminant analysis.

1. INTRODUCTION

Consider a nonparametric rank-based test for location in the two-sample case. Let $\chi_1 = \{X_1^1, \ldots, X_{n_1}^1\}$ be i.i.d. F_1 and $\chi_2 = \{X_1^2, \ldots, X_{n_2}^2\}$ be i.i.d. F_2 with χ_1, χ_2 mutually independent. We wish to test $H_0{:}F_1 = F_2$ against the alternative of stochastic ordering. We will assume for simplicity that the distributions F_j are continuous, implying that rank ties occur with probability zero.

The Wilcoxon-Mann-Whitney (WMW) statistic (Wilcoxon 1945; Mann and Whitney 1947) is based on pairwise comparisons; $W = (n_1 n_2)^{-1} \sum_i \sum_j I_{\{X_i^1 < X_j^2\}}$. The statistic W is an estimator of $P\left[X_1^1 < X_1^2\right]$; $E_0[W] = 1/2$ and the associated test rejects for large values of $|W - 1/2|$.

We present a generalized WMW statistic based on the comparison of subsample minima. Let $1 \le k_1 \le n_1$ and $1 \le k_2 \le n_2$, and define

$$W_{k_1,k_2} = \left|\Delta_{k_1,k_2}\right|^{-1} \sum_{(C_1,C_2) \in \Delta_{k_1,k_2}} I_{\{min(C_1) < min(C_2)\}} \tag{1}$$

where $|S|$ represents the cardinality of the set S. The summation is over elements of

$$\Delta_{k_1,k_2} = \{ (C_1, C_2) : C_1 \subset \chi_1, |C_1| = k_1, C_2 \subset \chi_2, |C_2| = k_2 \} ;$$

thus $\left|\Delta_{k_1,k_2}\right| = \binom{n_1}{k_1}\binom{n_2}{k_2}$.

As is the case for the conventional WMW statistic W, the W_{k_1,k_2} are U-statistics and are distribution-free under H_0. A normal approximation is available (see Xie and Priebe 1999 for details).

Theorem 1. As $n_1 \to \infty$ and $n_2 \to \infty$ such that $n_1/(n_1+n_2) \to \lambda \in (0, 1)$, $\sqrt{n_1+n_2}\,(W_{k_1,k_2} - E[W_{k_1,k_2}])$ is asymptotically normal with mean 0 and variance $VAR[W_{k_1,k_2}]$. Under H_0, $E_0[W_{k_1,k_2}] = k_1/(k_1+k_2)$ and $VAR_0[W_{k_1,k_2}] = k_1^2k_2^2/\left(\lambda(1-\lambda)\,(k_1+k_2)^2\,(2k_1+2k_2-1)\right)$.

Consideration of subsample maxima — replacing '*min*' with '*max*' in equation (1) — results in the generalization of WMW proposed by Kochar (1978), Deshpande and Kochar (1980), Stephenson and Ghosh (1985), and Ahmad (1996); W_{k_1,k_2} can be obtained from the analogous subsample maxima statistic by replacing $I_{\{min(C_1) < min(C_2)\}}$ with $I_{\{max(-C_2) < max(-C_1)\}}$ in (1). Shetty and Govindarajulu (1988) and Kumar (1997) propose using '*median*'. Xie and Priebe (1999) consider general order statistics.

In Section 2 we present a recurrence for the exact distribution $F_{W_{k_1,k_2}}$. Section 3 presents an example, motivated by a class of nonparametric discriminant analysis applications, indicating the utility of the generalization. In Section 4 a recurrence is presented for the generalization of statistic (1) to the $J \ge 3$ sample case.

2. A RECURRENCE FOR $F_{W_{k_1,k_2}}$

For the classical WMW statistic, the inadequacy of the normal approximation for small samples and the resultant desire for exact inference has spurred continuing interest in recurrences for the exact distribution (Lehmann 1975; Brus 1989; Chang 1992; Di Bucchianico 1997; Cheung and Klotz 1997). These same considerations motivate the derivation of a recurrence for the exact distribution of W_{k_1,k_2}. Classical combinatorics provides the required result. (The analogous recurrence for the variant of (1) employing subsample maxima is due to Kochar (1978) and Deshpande and Kochar (1980).)

Let $Z_{(1)}, \ldots, Z_{(n_1+n_2)}$ represent the ordered combined sample $\chi_1 \cup \chi_2$ and let $S = S_1 S_2 \ldots S_{n_1+n_2}$ be the sequence representing the sample labels for the ordered combined sample; $S_i = 2 - I_{\{Z_{(i)} \in \chi_1\}}$. Given n_1, n_2 and k_1, k_2, the random variable W_{k_1,k_2} is a function of only the probability measure on S. All $\binom{n_1+n_2}{n_1}$ sequences S of n_1 1's and n_2 2's are equally likely under H_0: $F_1 = F_2$ regardless of this common distribution; W_{k_1,k_2} is distribution-free.

Define $\gamma(i;k_1,k_2,n_1,n_2)$ to be the number of sequences S of n_1 1's and n_2 2's such that there are exactly i subset pair selections $(C_1, C_2) \in \Delta_{k_1,k_2}$ of k_1 1's and k_2 2's for which $min(C_1) < min(C_2)$. That is,

$$\gamma(i;k_1,k_2,n_1,n_2) = |\{S : |\Delta_{k_1,k_2}| W_{k_1,k_2}(S) = i\}|.$$

Then

$$f_{W_{k_1,k_2}}(i) = P(|\Delta_{k_1,k_2}| W_{k_1,k_2} = i) = \gamma(i;k_1,k_2,n_1,n_2) \Big/ \binom{n_1+n_2}{n_1}$$

is the probability mass function for the generalized WMW statistic (1) under H_0, and the desired probability distribution function $F_{W_{k_1,k_2}}$ is available therefrom.

Let $G(q;k_1,k_2,n_1,n_2) = \sum_i \gamma(i;k_1,k_2,n_1,n_2) q^i$ be the generating function for which the coefficient of q^i is precisely the required value $\gamma(i;k_1,k_2,n_1,n_2)$. Theorem 2 provides a recurrence for G.

Theorem 2. G satisfies the recurrence

$$G(q;k_1,k_2,n_1,n_2) = G(q;k_1,k_2,n_1,n_2-1) + G(q;k_1,k_2,n_1-1,n_2)\,q^{\binom{n_1-1}{k_1-1}\binom{n_2}{k_2}}$$

where the base cases of the recurrence are given by

$$G(q;k_1,k_2,m,k_2) = \sum_{l=0}^{m}\binom{l+k_2-1}{l}q^{\binom{m}{k_1}-\binom{l}{k_1}} \quad for\ k_1 \le m \le n_1$$

and

$$G(q;k_1,k_2,k_1,m) = \sum_{l=0}^{m}\binom{l+k_1-1}{l}q^{\binom{l}{k_2}} \quad for\ k_2 \le m \le n_2.$$

3. EXAMPLE

For the class of nonparametric discriminant analysis applications — especially in high dimensions — it is common practice to reduce the problem of comparing high-dimensional samples to that of the one-dimensional comparison of interpoint distances $Y_i^j(Z) = d(Z, X_i^j)$. See the discussion in Maa, Pearl and Bartoszynski (1996) and Bartoszynski, Pearl, and Lawrence (1997). In particular, the classical nearest neighbor classifiers (Fix and Hodges 1951; Cover and Hart 1967; Duda and Hart 1973; Devroye, Gyorfi and Lugosi 1996) are based on the ranks of these interpoint distances. The class-conditional distributions F_{Y^1} and F_{Y^2} are typically skewed to the right. For such distributions the statistic W_{k_1,k_2}, an estimator of $P\left[min(Y_1^1, \ldots, Y_{k_1}^1) < min(Y_1^2, \ldots, Y_{k_2}^2)\right]$, can be superior to the classical WMW or any order statistic-based generalization thereof.

For example, if the F_j are normal and $Z \sim F_1$, an analysis of W_{k_1,k_2} in terms of Pitman's asymptotic relative efficiency ρ (Pitman 1949; Lehmann 1975) yields $\rho(W_{10,10}, W) = 7.15$ and $\rho(W_{10,10}, W') > 2.38$ for all order statistic-based generalizations W' of (1) with subsamples of size 10. (Employing subsample maxima yields a relative efficiency of 53.8; medians yield 11.7.) The statistic W_{k_1,k_2} is admissible, and is preferred for this class of applications.

A more elaborate investigation of the class of order statistic-based generalizations of (1) in terms of Pitman efficiency is presented in Xie and Priebe (1999). A comparison of classifiers based on W_{k_1,k_2} to nearest neighbor classifiers for a specific discriminant analysis application is presented in Priebe (1998).

4. THE $J \geq 3$ SAMPLE CASE

Motivated by polychotomous nonparametric discriminant analysis based on the one-dimensional comparison of interpoint distances $d(Z, X_i^j)$, we consider now the $J \geq 3$ sample case, with mutually independent i.i.d. samples $\chi_1, \ldots, \chi_J$. Let $\boldsymbol{n} = [n_1, \ldots, n_J]^T$ and $\boldsymbol{k} = [k_1, \ldots, k_J]^T$ with $1 \leq k_j \leq n_j$. For $j = 1, \ldots, J$ define the statistics

$$W_{\boldsymbol{k}}^j = |\Delta_{\boldsymbol{k}}|^{-1} \sum_{(C_1, \ldots, C_J) \in \Delta_{\boldsymbol{k}}} I_{\{min(C_j) < min(\bigcup_{l \neq j} C_l)\}} .$$

The summation is over elements of

$$\Delta_{\boldsymbol{k}} = \{ (C_1, \ldots, C_J) : C_j \subset \chi_j, |C_j| = k_j \text{ for all } j \} ;$$

thus $|\Delta_{\boldsymbol{k}}| = \prod_{j=1}^{J} \binom{n_j}{k_j}$.

Unlike the case $J = 2$, here we need the joint distribution $F_{W_{\boldsymbol{k}}^1, \ldots, W_{\boldsymbol{k}}^J}(i_1, \ldots, i_J)$ since the value of the largest of the $W_{\boldsymbol{k}}^j$ no longer completely determines the values for the remaining $J-1$. (For $J = 2$, $W_{\boldsymbol{k}}^2 = 1 - W_{\boldsymbol{k}}^1$.) Fortunately, a general recurrence is available.

Let $\gamma_j(i;\boldsymbol{k}, \boldsymbol{n})$ be the number of sequences S of n_1 1's, ..., n_J J's such that there are exactly i subset selections $(C_1, \ldots, C_J) \in \Delta_{\boldsymbol{k}}$ of k_1 1's, ..., k_J J's for which $min(C_j) < min(\bigcup_{l \neq j} C_l)$. The calculation required for the joint distribution $F_{W_{\boldsymbol{k}}^1, \ldots, W_{\boldsymbol{k}}^J}(i_1, \ldots, i_J)$ is that of $P[\gamma_1 = i_1, \ldots, \gamma_J = i_J]$. That is, a recurrence is available which yields the necessary values $\gamma(i_1, \ldots, i_J; \boldsymbol{k}, \boldsymbol{n})$, the number of sequences S of n_1 1's, ..., n_J J's such that, simultaneously for each j, there are exactly i_j subset selections $(C_1, \ldots, C_J) \in \Delta_{\boldsymbol{k}}$ for which $min(C_j) < min(\bigcup_{l \neq j} C_l)$.

The generating function of interest, using the $\gamma(i_1, \ldots, i_J; \boldsymbol{k}, \boldsymbol{n})$ as coefficients, is

$$G(\boldsymbol{q};\boldsymbol{k},\boldsymbol{n}) = \sum_{j=1}^{J} \gamma(\boldsymbol{i};\boldsymbol{k},\boldsymbol{n})\, \boldsymbol{q}^{\boldsymbol{i}}$$

where $\boldsymbol{i} = [i_1, \ldots, i_J]^T$, $\boldsymbol{q} = [q_1, \ldots, q_J]^T$, and $\boldsymbol{q}^{\boldsymbol{i}} = \prod q_j^{i_j}$. For ease of notation we will write $\boldsymbol{n}_{(j)}$ for the indices $n_1, \ldots, n_J$ with n_j replaced by $n_j - 1$. By a combinatorical argument similar to that used for the two sample case, we have

Theorem 3. G satisfies the recurrence

$$G(\boldsymbol{q};\boldsymbol{k},\boldsymbol{n}) = \sum_{j=1}^{J} G(\boldsymbol{q};\boldsymbol{k},\boldsymbol{n}_{(j)}) \cdot q_j^{\binom{n_j-1}{k_j-1}\prod_{l \neq j}\binom{n_l}{k_l}}.$$

To write the base cases, the indices are reordered so that the first a n_j's have all been decreased to k_j $(n_1 = k_1, \ldots, n_a = k_a)$ and the remaining $J-a$ n_j's $n_{a+1}, \ldots, n_J$ are all still greater than k_j. The bases cases can then be written via the single formula

$$G(\boldsymbol{q};\boldsymbol{k},\boldsymbol{n}) =$$

$$\sum_{j=1}^{a} \frac{\left(\sum_{l=1}^{a} k_l + \left(\sum_{l=a+1}^{J} n_l\right) - 1\right)!}{(k_j-1)!\prod_{\substack{l=1 \\ l \neq j}}^{a} k_l! \prod_{l=a+1}^{J} n_l!}\, q_j^{\prod_{l=a+1}^{J}\binom{n_l}{k_l}}$$

$$+ \sum_{j=a+1}^{J} G(\boldsymbol{q};\boldsymbol{k},\boldsymbol{n}_{(j)}) \cdot q_j^{\binom{n_j-1}{k_j-1}\prod_{\substack{l=a+1 \\ l \neq j}}^{J}\binom{n_l}{k_l}}.$$

ACKNOWLEDGEMENTS

Research partially supported by Office of Naval Research Grants N00014-95-1-0777 and N00014-96-1-0829. The authors thank Jingdong Xie for many helpful comments. An anonymous reviewer's report improved the presentation of this material.

BIBLIOGRAPHY

Ahmad, I.A. (1996). "A class of Mann-Whitney-Wilcoxon type statistics," *Amer. Statist.* 50 324-327.

Bartoszynski, R., Pearl, D.K. and Lawrence, J. (1997). "A multidimensional goodness-of-fit test based on interpoint distances," *J. Amer. Statist. Assoc.* 92 577-586.

Brus, T. (1989). "A recurrence formula for the distribution of the Wilcoxon rank sum statistic," *Statist. Probab. Lett.* 7 161-165.

Chang, D.K. (1992). "A note on the distribution of the Wilcoxon rank sum statistic," *Statist. Probab. Lett.* 13 343-349.

Cheung, Y.K. and Klotz, J.H. (1997). "The Mann Whitney Wilcoxon distribution using linked lists," *Statist. Sinica* 7 805-813.

Cover, T.M. and Hart, P.E. (1967). "Nearest neighbor pattern classification," *IEEE Trans. Info. Theory* 13 21-27.

Deshpande, J.V. and Kochar, S.C. (1980). "Some competitors of tests based on powers of ranks for the two-sample problem," *Sankhya* 42 236-241.

Devroye, L., Gyorfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer, New York.

Di Bucchianico, A. (1997). "Computer algebra, combinatorics, and the Wilcoxon-Mann-Whitney statistic," preprint (submitted for publication).

Duda, R.O. and Hart, P.E. (1973). *Pattern Classification and Scene Analysis*. Wiley, New York.

Fix, E. and Hodges, J.L. (1951). "Discriminatory analysis: nonparametric discrimination: consistency properties," Report No. 4, USAF School of Aviation Medicine, Randolph Field, TX.

Kochar, S.C. (1978). "A class of distribution-free tests for the two-sample slippage problem," *Comm. in Statist. - Theo. and Meth.* A7 1243-1252.

Kumar, N. (1997). "A class of two-sample tests for location based on sub-sample medians," *Comm. in Statist. - Theo. and Meth.* 26 943-951.

Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods Based on Ranks.* Holden-Day, San Francisco.

Maa, J.-F., Pearl, D.K. and Bartoszynski, R. (1996). "Reducing multidimensional two-sample data to one-dimensional interpoint comparisons," *Ann. Statist.* 24 1069-1074.

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

Pitman, E.J.G. (1949). "Lecture Notes on Nonparametric Statistical Inference," Columbia University.

Priebe, C.E. (1998). "Olfactory classification via randomly weighted nearest neighbors," Johns Hopkins University Department of Mathematical Sciences Technical Report #585, Baltimore, MD.

Shetty, I.D. and Govindarajulu, Z. (1988). "A two-sample test for location," *Comm. in Statist. - Theo. and Meth.* 17 2389-2401.

Stephenson, R.W. and Ghosh, M. (1985). "Two-sample nonparametric tests based on subsamples," *Comm. in Statist. - Theo. and Meth.* 14 1669-1684.

Wilcoxon, F. (1945). "Individual comparisons by ranking methods," *Biometrics* 1 80-83.

Xie, J. and Priebe, C.E. (1999). "Generalizing the Mann-Whitney-Wilcoxon statistic," *J. Nonpar. Statist.* (to appear).

Received November, 1998; Revised July, 1999.