This research was supported in part by Grant No. Dot-HS-4-00897 (SY085)
from the National Highway Traffic Safety Administration

ON THE VARIANCE ESTIMATE OF A MANN-WHITNEY STATISTIC FOR ORDERED GROUPED DATA

by

Yosef Hochberg

Highway Safety Research Center
University of North Carolina

ISMS #1002

ON THE VARIANCE ESTIMATE OF A MANN-WHITNEY STATISTIC FOR ORDERED GROUPED DATA

by

Yosef Hochberg

Summary

Two consistent estimators are given for the variance of a Mann-Whitney type statistic used for comparing two populations based on samples grouped into ordered response categories.

1. Introduction. When comparing ordered grouped data (i.e. ordered response categories) one is urged to use methods generally more efficient than the conventional chi-square. Work in this direction is given by Bross (1958), where Ridit analysis is introduced, by Sen (1967a), where score statistics of some asymptotic optimality properties are proposed, and by Williams and Grizzle (1972), where the SGK approach (Grizzle, Starmer and Koch, 1969) is used.

The use of linear score statistics based on the theory of ranks for discrete distributions (see e.g. Vorlicova, 1970 and Conover, 1973) in the analysis of ordered categorical data has been advocated by many writers, for example, Grizzle et al (1969), Hobbs (1973), and Hobbs and Conover (1974). These writers were mainly concerned with hypotheses testing rather than with confidence interval estimation of certain functionals of interest to the experimenter.

Consider two populations with corresponding random variables X and Y. Flora (1974) uses the Mann-Whitney statistic for inference on $P(X<Y) - P(Y<X)$

in the case of ordered grouped data. However, he uses a conditional variance of the Mann-Whitney statistic under the null hypothesis Ho:P(X<Y) - P(Y<X) = 0 which disvalidates his confidence intervals. When the observations on X and Y are not in grouped form, the problem of inference on P(X<Y) received much attention in the statistical literature (mainly in the continuous case) the main problem being that of obtaining upper bounds on or estimates of the variance of the Mann-Whitney estimate of P(X<Y). (See, e.g. Sen, 1967b and Govindarajulu, 1968).

When the data is grouped into ordered categories, the functional P(X<Y) - P(Y<X) is usually more informative then P(X<Y) since P(X=Y) > 0, and often is quite large.

Here we give two consistent estimators of the variance of the Mann-Whitney estimator of P(X<Y) - P(Y<X) for grouped data which can be used (when sample sizes are sufficiently large) for obtaining a confidence interval for this functional.

In Sections 2 and 3 we give the two consistent estimators which are then exemplified in Section 4.

2. First estimator. Let $X_i, i=1,\ldots,n$ be i.i.d. observations from the distribution F(X) and $Y_i, i=1,\ldots,m$ be another sample of i.i.d. observations from the distribution G(Y). The data is presented in grouped form as in Table 2.1.

Table 2.1: Data Format

Ordered categories	C_1	C_2	. . .	C_k	Total
Observations (X) from F	f_1	f_2	. . .	f_k	n
Observations (Y) from G	g_1	g_2	. . .	g_k	m
Total	T_1	T_2	. . .	T_k	N=m+n

For example, F and G might correspond to the injury distribution of belted and unbelted drivers involved in accidents, respectively, and $C_1,\ldots,C_k$ are ordered injury categories ranging from least to most severe injury.

Let I(a) denote the group index of an observation with value a. Thus, $I(\cdot)$ is a random variable with the range of values: $1,2,\ldots,k$.

Let

$$A_{ij} = \begin{cases} 1 & \text{if } I(Y_i) > I(X_j) \\ 0 & \text{otherwise} \end{cases}$$

$$E_{ij} = \begin{cases} 1 & \text{if } I(Y_i) = I(X_j) \\ 0 & \text{otherwise} \end{cases}$$

$$B_{ij} = \begin{cases} 1 & \text{if } I(Y_i) < I(X_j) \\ 0 & \text{otherwise} \end{cases}$$

and define

$$\Pi^+ = \Pr\,[I(Y) > I(X)]$$

$$\Pi^\circ = \Pr\,[I(Y) = I(X)]$$

$$\Pi^- = \Pr\,[I(Y) < I(X)]\ .$$

We also let $D_{ij} = A_{ij} - B_{ij}$ which implies that

$$D_{ij} = \begin{cases} 1 & \text{if } I(Y_i) > I(K_j) \\ 0 & \text{if } I(Y_i) = I(X_j) \\ -1 & \text{if } I(Y_i) < I(X_j)\ . \end{cases}$$

The Π^+, Π° and Π^- are regular functionals on the class of all pairs of distribution functions (F,G), i.e., they admit unbiased estimators which are given by

$$\hat{\Pi}^{+} = \sum_{ij}\sum A_{ij}/mn = \sum_{i=1}^{k-1}\sum_{j=i+1}^{k} f_i g_j/mn$$

$$\hat{\Pi}^{\circ} = \sum_{ij}\sum E_{ij}/mn = \sum_{i=1}^{k} f_i g_i/mn$$

$$\hat{\Pi}^{-} = \sum_{ij}\sum B_{ij}/mn = \sum_{i=2}^{k}\sum_{j=1}^{i-1} f_i g_j/mn$$

A reasonable functional for comparing F with G is $\Pi^{+}-\Pi^{-}$ which admits the unbiased estimator

$$\frac{1}{mn}\sum_{ij}\sum D_{ij} \equiv \frac{1}{mn} W = \hat{\Pi}^{+}-\hat{\Pi}^{-}.$$

Regarding the variance of W, the situation is less simple. Flora (1974) gives the expression

$$V_0(W) = \frac{mn(N+1)}{3}\left[1 - \frac{\sum_i (T_i^3-T_i)}{N^3-N}\right].$$

He then uses (2.1) to obtain large sample test procedures of the hypothesis H_0: $\Pi^{+}-\Pi^{-} = 0$ and to construct a confidence interval for $\Pi^{+}-\Pi^{-}$.

However, (2.1) is the variance of W only under H_0 (and when conditioning on the T_i, see Conover, 1973). Thus, the large sample test procedure of H_0 based on $Z^* = W/V_0(W)$ being asymptotically a standard normal variate, is correct, but the confidence interval given by Flora, 1974 for $\Pi^{+}-\Pi^{-}$ as

$$\left[\frac{W}{mn} \pm Z^*_{\alpha/2} \left\{\frac{V_0(W)}{m^2n^2}\right\}^{\frac{1}{2}}\right]$$

(where $Z^*_{\alpha/2}$ is the $1-\frac{\alpha}{2}$ quantile of Z^*)

is in error, since $V_0(W)$ is not the variance of W under a non-null hypothesis. As an example consider the case when for some number h, $1<h<k$, we have

$$\Pr[I(X)<h] = 1$$

$$\Pr[I(Y)>h] = 1 \ .$$

In this case the variance of W is zero, and note that (2.1) does not give zero.

Next we denote the variance of W by $V_{F,G}(W)$ to distinguish it from (2.1). This is the non-conditional variance of W under general F,G. To express $V_{F,G}(W)$ we need some more notation which we now introduce. For independent X_j, X_ℓ and Y_i let

$$\Pi_{xxy} = \Pr[I(Y_i) > \max\{I(X_j), I(X_\ell)\}]$$

$$\Pi_{yxx} = \Pr[I(Y_i) < \min\{I(X_j), I(X_\ell)\}]$$

$$\Pi_{xyx} = \Pr[I(X_j) < I(Y_i) < I(X_\ell)]$$

Similarly define by symmetry Π_{yyx}, Π_{xyy} and Π_{yxy}. We can now obtain an explicit expression for $V_{F,G}(W)$. On writing

$$E(W) = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{s=1}^{n} \sum_{t=1}^{m} E(D_{ij}D_{st})$$

$$= \sum_i \sum_j \sum_s \sum_t \Big\{\Pr(D_{ij}=1;\ D_{st}=1) + \Pr(D_{ij}=-1;\ D_{st}=-1)$$

$$- \Pr(D_{ij}=1; D_{st}=-1) - \Pr(D_{ij}=-1; D_{st}=1)\Big\}$$

it is not difficult to verify that

$$E(W^2) = m(m-1)n(n-1)(\Pi^+-\Pi^-)^2 + mn(m-1)(\Pi_{xyy}+\Pi_{yyx}-2\Pi_{yxy})$$

$$+ mn(n-1)(\Pi_{xxy}+\Pi_{yxx}-2\Pi_{xyx}) + mn(\Pi^++\Pi^-).$$

We thus get (note that $E(W) = mn(\Pi^+ - \Pi^-)$)

$$V_{F,G}(W) = mn\Big[\Pi^+ + \Pi^- + (m-1)(\Pi_{xyy} + \Pi_{yyx} - 2\Pi_{yxy})$$

$$+ (n-1)(\Pi_{xxy} + \Pi_{yxx} - 2\Pi_{xyx}) - (n+m-1)(\Pi^+ - \Pi^-)^2\Big]$$

For constructing large sample confidence intervals, we need consistent estimators of the functionals involved in the expression for $V_{F,G}(W)$.

Consistent estimators of the regular functionals Π_{xyy}, Π_{yxx}, etc., are obtained by constructing the corresponding U-statistics. (See, for example, Puri and Sen, 1971, Ch. 3). For example, the consistent estimators of Π_{xyy}, Π_{yxy} and of Π_{yyx} are given by

$$\hat{\Pi}_{xyy} = \frac{1}{mn(m-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} f_i g_j \left[\sum_{j=i+1}^{k} g_j - 1 \right]$$

$$\hat{\Pi}_{yxy} = \frac{2}{mn(m-1)} \sum_{i=2}^{k-1} f_i \left(\sum_{j=1}^{i-1} g_j \right) \left(\sum_{j=i+1}^{k} g_j \right) \quad \text{and}$$

$$\hat{\Pi}_{yyx} = \frac{1}{mn(m-1)} \sum_{i=2}^{k} \sum_{j=1}^{i-1} f_i g_j \left[\sum_{j=1}^{i-1} g_j - 1 \right] .$$

These estimates are then substituted into the formula for $V_{F,G}(W)$ to give $\hat{V}_{F,G}(W)$. The large sample $1-\alpha$ confidence interval for $\Pi^+ - \Pi^-$ is given by

$$\frac{1}{mn}\left[W + Z_{\alpha/2} \hat{V}^{\frac{1}{2}}_{F,G}(W)\right]$$

where $Z_{\alpha/2}$ is the $1 - \frac{\alpha}{2}$ quantile of $N(0,1)$.

3. A second consistent estimator. Our second estimator is based on the Delta method following the GSK approach and its generalization in Forthofer and Koch (1973). We let $n_1=n$, $n_2=m$, $\underset{\sim}{p}_i = (p_{i1},\ldots,p_{ik})'$ and $\underset{\sim}{\Pi}_i = (\Pi_{i1},\ldots,\Pi_{ik})'$, $i=1,2$. We have: $E(\underset{\sim}{p}_i) = \underset{\sim}{\Pi}_i$ and the dispersion matrix of $\underset{\sim}{p}_i$ is given by

$$\frac{1}{n_i}\{\mathrm{Diag}(\underset{\sim}{\Pi}_i) - \underset{\sim}{\Pi}_i\underset{\sim}{\Pi}_i'\} \equiv \underset{\sim}{V}(\underset{\sim}{\Pi}_i),\ i=1,2. \qquad (3.1)$$

On letting $\underset{\sim}{p}' = (\underset{\sim}{p}_1', \underset{\sim}{p}_2')$ and $\underset{\sim}{\Pi}' = (\underset{\sim}{\Pi}_1', \underset{\sim}{\Pi}_2')$ we have $E(\underset{\sim}{p}) = \underset{\sim}{\Pi}$ and the dispersion matrix of $\underset{\sim}{p}$ (which we shall denote by $\underset{\sim}{V}(\underset{\sim}{\Pi})$) is block diagonal with the $\underset{\sim}{V}(\underset{\sim}{\Pi}_i)$ on the main diagonal. The sample estimate of $\underset{\sim}{V}(\underset{\sim}{\Pi})$ is obtained by substituting p's instead of Π's in (3.1), this estimate is denoted by $\underset{\sim}{V}(\underset{\sim}{p})$.

Next we show that the Mann-Whitney statistic -- $\frac{1}{mn}W$ (see Section 2) can be expressed as an exponential-logaritmic-linear function of $\underset{\sim}{p}$ as in Forthofer and Koch (1973), and thus, the Delta method can be used for obtaining a consistent estimator for $V_{F,G}(W)$. To see this we let

$$\underset{4(k-1)\times 2k}{\underset{\sim}{A}} = \begin{bmatrix} \underset{\sim}{L} & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{T} \\ \underset{\sim}{T} & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{L} \end{bmatrix}$$

where $\underset{\sim}{L}:(k-1)\times k = [\underset{\sim}{I}:(k-1)\times(k-1),\ \underset{\sim}{0}:(k-1)\times 1]$ and $\underset{\sim}{T} = [\underset{\sim}{0}:(k-1)\times 1,\ \underset{\sim}{R}:(k-1)\times(k-1)]$ where $\underset{\sim}{R}$ is an upper triangular matrix with the common element -- 1 on and above the diagonal.

$$\underset{2(k-1)\times 4(k-1)}{\underset{\sim}{K}} = \begin{bmatrix} \underset{\sim}{I} & \underset{\sim}{I} & \underset{\sim}{0} & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{0} & \underset{\sim}{I} & \underset{\sim}{I} \end{bmatrix}$$

where both $\underset{\sim}{0}$ and $\underset{\sim}{I}$ are of dimensions $(k-1)\times(k-1)$ and

$$\underset{2\times 2(k-1)}{\underset{\sim}{Q}} = \begin{bmatrix} \underset{\sim}{1}' & \underset{\sim}{0}' \\ \underset{\sim}{0}' & \underset{\sim}{1}' \end{bmatrix}$$

where $\underset{\sim}{1}:(k-1)\times 1 = (1,\ldots,1)'$ and $\underset{\sim}{0}:(k-1)\times 1 = (0,\ldots,0)'$.

We have

$$\frac{1}{mn}W = \frac{1}{mn}\left[\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} f_i g_j - \sum_{i=2}^{k}\sum_{j=1}^{i-1} f_i g_j\right]$$

$$= \sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\left(p_{1i}p_{2j} - p_{1j}p_{2i}\right),$$

and it is not difficult to verify that on letting $\underset{\sim}{F}(\underset{\sim}{p}) = (F_1(\underset{\sim}{p}), F_2(\underset{\sim}{p}))' = \underset{\sim}{Q}(\exp\{\underset{\sim}{K}[\ln(\underset{\sim}{A}\underset{\sim}{p})]\})$, the Mann-Whitney statistic can be written as

$$\frac{1}{mn} W = F_1(\underset{\sim}{p}) - F_2(\underset{\sim}{p}).$$

By the Delta method, the consistent estimator of the dispersion of $\underset{\sim}{F}(\underset{\sim}{p})$ is given by $\underset{\sim}{S}$

$$\underset{\sim}{S} = \underset{\sim}{Q}\underset{\sim}{D}_{\underset{\sim}{y}}\underset{\sim}{K}\underset{\sim}{D}_{\underset{\sim}{a}}^{-1}\underset{\sim}{A}[\underset{\sim}{V}(\underset{\sim}{p})]\underset{\sim}{A}'\underset{\sim}{D}_{\underset{\sim}{a}}^{-1}\underset{\sim}{K}'\underset{\sim}{D}_{\underset{\sim}{y}}\underset{\sim}{Q}'$$

where

$$\underset{\sim}{a} = \underset{\sim}{A}\underset{\sim}{p}, \quad \underset{\sim}{y} = \exp\{\underset{\sim}{K}\ln(\underset{\sim}{a})\}$$

and $\underset{\sim}{D}_{\underset{\sim}{a}}(\underset{\sim}{D}_{\underset{\sim}{y}})$ is a diagonal matrix with $\underset{\sim}{a}(\underset{\sim}{y})$ on the diagonal.

This gives a consistent estimator for the variance of W/mn as $\underset{\sim}{c}'\underset{\sim}{S}\underset{\sim}{c}$, where $\underset{\sim}{c} = (1,-1)'$. The use of this method is quite convenient using the available programs for the implementation of the GSK method in the setup of Forthofer and Koch (1973).

4. Example. Consider the following data from Table 48 of McLean (1973).

Table 4.1: Injury severity by group drivers in left side impact.

Group	Injury Severity					Total
	None	C	B	A	K	
Door Beam	286	14	10	17	2	329
No Door Beam	731	41	28	72	9	881
Total	1017	55	38	89	11	1210

This data is analyzed by Flora (1974) twice, first omitting the "none" category and second including all categories. For our illustrative purposes, we go through the same two analyses.

The "No Door Beam" population here corresponds to G(Y) in section 2 and the "Door Beam" population corresponds to F(X).

First consider the case when the "None" category is omitted. We get $W = 624$ and

$$\hat{\Pi}^{+} = .3859 \qquad \hat{\Pi}^{\circ} = .3250 \qquad \hat{\Pi}^{-} = .2891$$

corresponding to probabilities of greater, equal, or less severity of injury, (given injury) respectively, without the side door beam.

One also obtains

$$\hat{\Pi}_{xyy} = .2402 \qquad \hat{\Pi}_{yyx} = .1411 \qquad \hat{\Pi}_{yxy} = .0455$$

$$\hat{\Pi}_{yxx} = .1588 \qquad \hat{\Pi}_{xxy} = .2200 \qquad \hat{\Pi}_{xyx} = .0805$$

and substituting in (2.2) gives

$$\hat{V}_{F,G}(W) = (517.2)^2 .$$

One may compare this value with the value $(602.74)^2$ reported by Flora (1974) based on (2.1).

Consider now the second type analysis where all categories are used. We have $W = 12{,}091$ and

$$\hat{\Pi}^{+} = .1565 \qquad \hat{\Pi}^{\circ} = .7286 \qquad \hat{\Pi}^{-} = .1149$$

corresponding to probabilities of greater, equal, or less severity of injury respectively, without the side door beam.

We also obtain

$$\hat{\Pi}_{xyy} = .0259 \qquad \hat{\Pi}_{yyx} = .1012 \qquad \hat{\Pi}_{yxy} = .0146$$

$$\hat{\Pi}_{yxx} = .0143 \qquad \hat{\Pi}_{xxy} = .1442 \qquad \hat{\Pi}_{xyx} = .0115 \ .$$

Substituting in (2.2) we get

$$\hat{V}_{F,G}(W) = (6116.3)^2$$

Again, this might be compared with the value $(6888.0)^2$ obtained by (2,1).

Finally, we note that based on the consistent estimator of Section 3 we get for the first type analysis $\hat{V}_{F,G}(W) = (597.3)^2$ and for the second type analysis -- $\hat{V}_{F,G}(W) = (6499.0)^2$.

References

Bross, I.D.J. [1958]. How To Use Ridit Analysis, Biometrics, 14, 18-38.

Conover, W.J. [1973]. Rank Tests For One Sample, Two Samples, and k Samples Without the Assumption of a Continuous Distribution Function, Annals of Statistics, 1, 1105-1125.

Flora, J.D. Jr. [1974]. A Note on Ridit Analysis, Memeo. Ser., No. 3, Department of Biostatics, School of Public Health, Univ. of Michigan, Ann Arbor.

Forthofer, R.N. and G.G. Koch. [1973]. An Analysis for Compounded Functions of Categorical Data, Biometrics, 29, 143-157.

Govindarajulu, Z. [1968]. Distribution-Free Confidence Bounds for $P(X<Y)$, Ann. Inst. Statist. Math. 20, 229-238.

Grizzle, J.E., Starmer, C.F., and Koch, G.G. [1969]. Analysis of Categorical Data by Linear Models, Biometrics, 25, 489-504.

Hobbs, G.R. [1973]. Some Results on the Theory of Rank Tests for Two or More Samples from Categorical Populations, Ph.D. dissertation, Kansas State University.

Hobbs, G.R. and Conover, W.J. [1974]. The Asymptotic Efficienty of the Kruskal-Wallis Test and the Median Test in Contingency Tables with Ordered Categories, Commun. in Stat., Vol 3, No 12. 1131-1138.

McLean, A.J. [1973]. Collection and Analysis fo Collision Data for Determining the Effectiveness of Some Vehicle Systems, HSRC, University of North Carolina, Technical Report No. 7301-C19.

Sen, P.K. [1967a]. Asymptotically Most Powerful Rank Order Tests for Grouped Data, Ann. Math. Statist., 38, 1229-1239.

Sen, P.K. [1967b]. A Note on Asymptotically Distribution-Free Confidence Bounds for $P(X<Y)$, Based on Two Independent Samples, Sankhyā A, 29, 351-372.

Vorlicova, D. [1970]. Asymptotic Properties of Rank Tests Under Discrete Distributions, Zeit. Wahrscheinlichkeitsth, 14, 275-289.

Williams, O.D. and Grizzle, J.E. [1972]. Analysis of Contingency Tables Having Ordered Response Categories, J. Amer. Statist. Assoc. 67, 55-63.