ON A USE OF THE MANN-WHITNEY STATISTIC

Z. W. BIRNBAUM
UNIVERSITY OF WASHINGTON

1. Introduction

Let X and Y be independent random variables with continuous cumulative probability functions F and G, respectively, and let

$$X_1, X_2, \cdots, X_m; \qquad Y_1, Y_2, \cdots, Y_n \tag{1.1}$$

be samples of X and of Y. Mann and Whitney [1] considered the statistic

$$U = \text{number of pairs } (X_j, Y_k) \text{ such that } Y_k < X_j . \tag{1.2}$$

For $m = n$ an equivalent statistic had been proposed and studied earlier by Wilcoxon [2]. The main aim of these studies was to develop a test of the hypothesis that X and Y have the same probability distribution: $F = G$. More about this test will be reported in section 2.

Independently, Haldane and Smith [3] investigated the following problem. In some hereditary conditions, the probability that a member of a sibship has the condition depends partly on his birth rank. Having records of sibships in the order of birth, stating for each individual whether it has or does not have the condition, how does one test for independence of the condition from birth rank? To answer this question, Haldane and Smith constructed a test statistic which is equivalent with the U statistic (1.2).

Without an attempt at completeness, we shall give in section 2 a brief survey of the known properties of the U statistic which are of importance for its use in testing hypotheses. The main purpose of the present paper, however, is to discuss another use of this statistic which, while not new, seems to have attracted less attention. Let

$$p = Pr\{Y < X\}. \tag{1.3}$$

If the samples (1.1) are available, then the statistic

$$\hat{p} = \frac{U}{mn} \tag{1.4}$$

can be used to estimate the parameter p. It is this particular use of U which will be explored in some detail in sections 3 and 4.

2. Properties of U useful in testing hypotheses

Under the hypothesis (H): $F = G$, Mann and Whitney [1] have tabulated the exact probability distribution of U for $m \leqq n \leqq 8$, and proved that

$$\frac{U - \frac{1}{2}mn}{\sqrt{mn(m+n+1)/12}} \tag{2.1}$$

This paper was prepared with the support of the Office of Ordnance Research, U.S. Army, under Contract DA-04-200-ORD-355.

has asymptotically the normal distribution $N(0, 1)$. This approximation is already very good for $m = n = 8$. They showed furthermore that, for alternatives (A): $F(s) > G(s)$ for all s, the one-sided test of (H) using the critical region $U \leqq mn/2 - t\sqrt{mn(m+n+1)/12}$ with $t > 0$ is consistent.

Lehmann [4] generalized one of these results by proving that the random variable $\sqrt{n}\,(\hat{p} - p)$ is asymptotically nondegenerate normal for any pair F, G, under the assumptions (a) $m = cn$, $n \to \infty$, (b) $0 < p < 1$.

The other result of Mann and Whitney was generalized by Van Dantzig [5] who showed that their one-sided test is consistent against any alternative such that $p < \frac{1}{2}$, and, for sufficiently small significance level, against no other alternative. Similar statements are true for a two-sided test, based on U, of the hypothesis $p = \frac{1}{2}$ against any alternative $p \neq \frac{1}{2}$.

3. Estimation of p

3.1. *An illustration.* If structural components of a mechanism are mass produced, the strength at failure Y of each single component (equals stress at which this component will fail) may be considered a random variable. The component is installed in an assembly and exposed to a stress which reaches its maximum value X, again a random variable. If $Y < X$, then the component will fail in use. In this situation, $p = Pr\{Y < X\}$ is the probability that failure will occur because, due to chance, a component with relatively low strength was paired off with a high stress. It clearly is of interest to estimate this probability, preferably from samples of X and of Y alone, since installing the components in complete assemblies and trying them out under conditions of actual use may involve nearly prohibitive expense and effort. It also will be important to be able to estimate p without knowing the distribution of the strengths of the components, or of the stresses, or of both.

3.2. *Properties of U useful for estimating p.* It is easily seen that

$$E(\hat{p}) = E\left(\frac{U}{mn}\right) = p \tag{3.2.1}$$

so that $\hat{p}$ is an unbiased estimate of p. Van Dantzig [5] obtained the sharp upper bound for the variance of $\hat{p}$

$$\sigma^2(\hat{p}) \leqq \frac{p(1-p)}{\min(m, n)} \tag{3.2.2}$$

and thereby showed that for $m \to \infty$, $n \to \infty$ the statistic $\hat{p}$ is a consistent estimate of p. Lehmann (see pp. 3-23–3-24 in [6]) also proved that $\hat{p}$ is the UMV unbiased estimate of p.

These properties as well as those enumerated in section 2 make $\hat{p}$ a very good point estimate of p. If a confidence interval for p is wanted, one can make use of Lehmann's theorem on the asymptotic normality of $\sqrt{n}\,(\hat{p} - p)$, together with Van Dantzig's inequality (3.2.2), to obtain the sample size and the confidence interval for a given confidence level.

Difficulties will arise when a confidence interval for p is desired and the assumptions of Lehmann's theorem are not fulfilled. These assumptions are, intuitively speaking, (a) that the sample sizes m and n are of the same order of magnitude, and (b) that p is sufficiently far from 0 and from 1. If one or both of these assumptions are not clearly satisfied, that is, if one sample size is much larger than the other, or p is close to 0 or close

to 1, then there is no assurance that the normal approximation is still good enough to be used.

In practical situations, one or both of these assumptions are often not fulfilled. In the problem described in 3.1, for example, p will most likely be close to 0. It can also happen that the cost of obtaining observations on X is so different from that for Y that it will be desirable to choose m and n very differently. In neither case will it be safe to rely on the normal approximation. This suggests the following concrete problem: to obtain a statistic ψ and, for any ϵ, $\alpha > 0$, a pair of numbers $M_{\epsilon,\alpha}$, $N_{\epsilon,\alpha}$ so that

$$Pr\{p \leqq \psi + \epsilon\} \geqq 1 - \alpha\,, \qquad \text{if } m \geqq M_{\epsilon,\alpha},\ n \geqq N_{\epsilon,\alpha}. \tag{3.2.3}$$

From here on we shall be concerned mainly with this problem.

3.3. *A one-sided confidence interval for p, not depending on normal approximation.* Let us first consider the case when F is known and G not known. This situation arises, for example, when it is easy to obtain a practically unlimited number of observations of X, and hence to reconstruct F as accurately as desired, but only a finite sample $Y_1, \cdots, Y_n$ of Y can be obtained [this would correspond to $\lim (n/m) = 0$ in the general case]. Let $Y_1^* < Y_2^* < \cdots < Y_n^*$ be the ordered sample of Y and

$$G_n(s) = \begin{cases} 0 & \text{for } s < Y_1^* \\ \dfrac{k}{n} & \text{for } Y_k^* \leqq s < Y_{k+1}^* \\ 1 & \text{for } Y_n^* \leqq s \end{cases} \tag{3.3.1}$$

the empirical distribution function (e.d.f.). We consider the statistic

$$\hat{p}_1 = \int_{-\infty}^{+\infty} G_n(s)\, dF(s) = 1 - \frac{1}{n}\sum_{k=1}^{n} F(Y_k^*) \tag{3.3.2}$$

$$= 1 - \frac{1}{n}\sum_{k=1}^{n} F(Y_k)$$

where, by definition, $F(Y_0^*) = 0$, $F(Y_{n+1}^*) = 1$. Since

$$E(\hat{p}_1) = 1 - EF(Y) = 1 - \int_{-\infty}^{+\infty} F(s)\, dG(s) \tag{3.3.3}$$

$$= \int_{-\infty}^{+\infty} G(s)\, dF(s) = p\,,$$

$\hat{p}_1$ is an unbiased estimate of p. To obtain a one-sided (upper) confidence interval for p we observe that

$$p - \hat{p}_1 = \int_{-\infty}^{+\infty} (G - G_n)\, dF \leqq \sup_{(s)} \{G(s) - G_n(s)\} = D_n^+ \tag{3.3.4}$$

hence

$$Pr\{p - \hat{p}_1 < \epsilon\} \geqq Pr\{D_n^+ < \epsilon\} = P_n(\epsilon)\,. \tag{3.3.5}$$

It is well known [7] that $P_n(\epsilon)$ is independent of G. Smirnov [8] has shown that

$$\lim_{n\to\infty} P_n\left(\frac{z}{\sqrt{n}}\right) = 1 - e^{-2z^2}. \tag{3.3.6}$$

In [9] a closed expression is given for $P_n(\epsilon)$, as well as a tabulation showing that the solutions $\epsilon_{n,\alpha}$ of the equation

$$P_n(\epsilon) = 1 - \alpha \tag{3.3.7}$$

for $\alpha = .10, .05, .01, .001$ differ from those obtained by using approximation (3.3.6) by less than .005 as soon as $n > 50$. For practical purposes, therefore, (3.3.5) may be rewritten

$$Pr\{p < \hat{p}_1 + \epsilon\} > 1 - e^{-2n\epsilon^2}. \tag{3.3.8}$$

This shows that $\hat{p}_1 + \epsilon$ is an upper confidence bound for p on a confidence level greater than or equal to $1 - \exp(-2n\epsilon^2)$, and that $\hat{p}_1$ is a statistic which answers the problem stated at the end of section 3.2.

A numerical example may be of interest. If we wish to state that $Pr\{p < \hat{p}_1 + .05\} \geqq .99$, the required sample size obtained from (3.3.8) is $n = 921$. Using Chebyshev's inequality and the bound $\sigma^2(\hat{p}) < p(1-p)/n \leqq (4n)^{-1}$, we would obtain $n = 10{,}000$. If we were to use the normal approximation, not knowing whether this is justified, and the same upper bound for $\sigma^2(\hat{p})$, the result would be $n = 541$.

3.4. *Case of F and G unknown.* Let $F_m(s)$ and $G_n(s)$ be e.d.f.'s corresponding to the samples (1.1). It is easily verified that

$$\hat{p} = \int_{-\infty}^{+\infty} G_n(s)\, dF_m(s) \tag{3.4.1}$$

and this together with (1.3) yields

$$\begin{aligned} p - \hat{p} &= \int_{-\infty}^{+\infty} G\, d(F - F_m) + \int_{-\infty}^{+\infty} (G - G_n)\, dF_m \\ &= \int_{-\infty}^{+\infty} (F_m - F)\, dG + \int_{-\infty}^{+\infty} (G - G_n)\, dF_m \end{aligned} \tag{3.4.2}$$

so that

$$p - \hat{p} \leqq \sup_{(s)} [F_m(s) - F(s)] + \sup_{(s)} [G(s) - G_n(s)] = D_m^+ + D_n^+ . \tag{3.4.3}$$

We have, therefore,

$$Pr\{p \leqq \hat{p} + \epsilon\} \geqq Pr\{D_m^+ + D_n^+ \leqq \epsilon\} = P_{m,n}(\epsilon) \tag{3.4.4}$$

which shows that $\hat{p}$ is a statistic of the kind asked at the end of section 3.2, provided we can determine $M_{\epsilon,\alpha}$, $N_{\epsilon,\alpha}$.

Since D_m^+ and D_n^+ are independent, and for $m \geqq 50$, $n \geqq 50$ Smirnov's approximation (3.3.6) is quite close, a good approximation to $P_{m,n}(\epsilon)$ could be obtained by convolution of two c.p.f.'s of the form (3.3.6). To make this expression practically useful, one would need numerical tabulations which are not available at this time. A somewhat crude but computationally easier procedure is the following. For any η such that $0 < \eta < \epsilon$ we have

$$P_{m,n}(\epsilon) = Pr\{D_m^+ + D_n^+ \leqq \epsilon\} \geqq Pr\{D_m^+ \leqq \eta\} \cdot Pr\{D_n^+ \leqq \epsilon - \eta\}, \tag{3.4.5}$$

and using the approximation (3.3.6) we obtain

$$P_{m,n}(\epsilon) \geqq (1 - e^{-2m\eta^2})(1 - e^{-2n(\epsilon-\eta)^2}). \tag{3.4.6}$$

The right-hand term can be maximized in η, which requires solving numerically a transcendental equation, and m and n can then be determined to make it equal to $1 - \alpha$. This procedure becomes quite simple if one wishes to have $m = n$, since then the maximum of the right term in (3.4.6) is attained for $\eta = \epsilon/2$ and one has

$$P_{n,n}(\epsilon) \geqq (1 - e^{-(n\epsilon^2/2)})^2 . \tag{3.4.7}$$

For example, to have $\epsilon = .05$, $\alpha = .01$, it is certainly sufficient to choose $M = N = 4238$. Using Chebyshev's inequality and the bound $\sigma^2(\hat{p}) \leqq p(1 - p)/n \leqq (4n)^{-1}$ one would again obtain $M, N = 10{,}000$.

4. Concluding remarks

The procedure outlined above for obtaining one-sided confidence bounds for p can clearly be used also for obtaining two-sided confidence intervals. The Kolmogorov statistic $\sup_{(x)} |F(x) - F_m(x)|$ would then take the place of $\sup_{(x)} \{F(x) - F_m(x)\}$.

This procedure, for the one-sided as well as the two-sided case, requires sample sizes which are most likely much too large. The obvious reason is the crudeness of inequalities such as (3.3.4) or (3.4.3). An improvement could be expected from a study of the asymptotic probability distribution of U if mnp is not large, although m and n are large. To our knowledge, no results are available in this direction. Another possibility of obtaining improved estimates for p would consist in making additional assumptions on F and G and arriving at bounds for $\sigma^2(\hat{p})$ which are better than (3.2.2). Some results of this kind have been obtained and will be published separately [10].

REFERENCES

[1] H. B. MANN and D. R. WHITNEY, "On a test of whether one of two random variables is stochastically larger than the other," *Annals of Math. Stat.*, Vol. 18 (1947), pp. 50–60.

[2] FRANK WILCOXON, "Individual comparisons by ranking methods," *Biometrics Bull.*, Vol. 1 (1945), pp. 80–83.

[3] J. B. S. HALDANE and CEDRIC A. B. SMITH, " A simple test for birth-order effect," *Annals of Eugenics*, Vol. 14 (1947), pp. 117–124.

[4] E. L. LEHMANN, "Consistency and unbiasedness of certain nonparametric tests," *Annals of Math. Stat.*, Vol. 22 (1951), pp. 165–179.

[5] D. VAN DANTZIG, "On the consistency and the power of Wilcoxon's two sample test," *Nederlandse Akad. Wetensch. Proc.*, Ser. A., Vol. 54 (1951), pp. 1–8.

[6] E. L. LEHMANN, "Notes on the theory of estimation," mimeographed lectures, University of California, 1950.

[7] A. WALD and J. WOLFOWITZ, "Confidence limits for continuous distribution functions," *Annals of Math. Stat.*, Vol. 10 (1939), pp. 105–118.

[8] N. SMIRNOV, "Sur les écarts de la courbe de distribution empirique," *Recueil Math. (Mat. Sbornik)*, N.S., Vol. 6 (1939), pp. 3–26.

[9] Z. W. BIRNBAUM and FRED H. TINGEY, "One-sided confidence contours for probability distribution functions," *Annals of Math. Stat.*, Vol. 22 (1951), pp. 592–596.

[10] Z. W. BIRNBAUM and ORVAL M. KLOSE, "On the variance of the Mann-Whitney statistic," to be published.