Part of Results on Admissibility of the well-known Wilcoxon-Mann-Whitney test

Abstract

The admissibility of the Wilcoxon-Mann-Whitney test in the class of all two-samples tests has been a longstanding "difficult and unsolved problem" in statistics. The problem hasn't been solved completely, except for results under some special situation. In this note, under the assumption that the samples have no ties, we solve this open problem affirmatively for any sample sizes.

1. Definition and notation

Suppose a random variable X takes each of s possible values $a_1, \cdots, a_s$ with probability $P(X = a_i) = p_i, i = 1,2, \cdots, s$. And similarly, a random variable Y takes each of t possible values $b_1, \cdots, b_t$ with probability $P(Y = b_j) = q_j, j = 1,2, \cdots, t$. The expectations are then defined as

$$E(X) = \sum_{i=1}^{s} a_i\, p_i, \tag{1.1}$$

$$E(f(X,Y)) = \sum_{i=1}^{s} \sum_{j=1}^{t} f(a_i, b_j)\, p_i q_j, \tag{1.2}$$

Where $f(x,y)$ is a function of x and y. Particularly, let $f(x,y) = xy$, then we have

$$E(XY) = \sum_{i=1}^{s} \sum_{j=1}^{t} a_i b_j\, p_i q_j; \tag{1.3}$$

Also, if X has a Bernoulli distribution, that is, X takes either 0 or 1, the expectation of X is

$$E(X) = 0 \times P(X = 0) + 1 \times P(X = 1) = P(X = 1). \tag{1.4}$$

(4) is very important in that it is a bridge between two concepts: probability and expectation.

Next, we introduce the concept of the cumulative distribution function of a random variable$X$.

DEFINITION: The cumulative distribution function (cdf) of a random variable X, denoted by$F(x)$, is defined by

$$F(x) = P(X \le x) = \sum_{i=1}^{s} p_i \, I(a_i \le x), \qquad \text{for all } x,$$

Where $I(a_i \le x) = 1$, if $a_i \le x$, elsewhere $I(a_i \le x) = 0$.

DEFINITION: For any two distributions F and G, we say $F > G$ if $F(x) \ge G(x)$ for each x and $F(x) > G(x)$ for some x.

2. Formulation of the problem

In this section, we will present the problem we try to solve. We first introduce hypothesis test and the relative testing procedure.

2.1. Hypothesis test

In real life, we often run into some statements needed to determine which of the two complementary hypotheses is true. For example, (a) Whether does a new drug have effect or not? (b) Are some people more susceptible to a disease than the other? (c) whether is a particular product up to standard or not? and so on. If we were the "God", knowing all the information, we can make the judgment correctly and easily. However, we are definitely not. We can only infer based on incomplete information. Yet we might make a correct decision or make a mistake. Thus, a prerequisite to draw statistical inference is the quantization of the chance of making right or wrong decision. It is probability that provides a mean for this job. The aim of this part is not to give a thorough introduction to probability. Rather, we attempt to outline some relative concepts in hypothesis test by probability language.

In general, the hypothesis which tends to be refused in a hypothesis testing problem is called the null hypothesis and its complementary is called alternative hypothesis. They are denoted by H_0 and H_1, respectively. For example, in the statement (b) given above, the null hypothesis states that, on the average, the drug has no effect, and the alternative hypothesis states that there is some effect. If the result of the test does not correspond with the actual state of nature, an error has occurred. If, on the other hand, the result of the test corresponds with the actual state of nature, a correct decision has been made. These four different situations are depicted in Table 1.1.

		Decision	
		Accept H_0	Reject H_0
Truth	H_0	Correct decision $(1-\alpha)$	Type I (α)
	H_1	Type II (β)	Correct decision $(1-\beta)$

Table 1.1 the four different situations and their probabilities

As we can see in Table 1.1, a hypothesis test might make one of two types of errors. If the hypothesis test incorrectly decides to reject a null hypothesis H_0 when it is actually true, then the test has made a *Type I Error*. If, on the other hand, the test decides to accept H_0 when it is actually false, a *Type II Error* has been made. A test is a level α test if the maximum probability of its Type I Error equals α. The probability of a Type II Error is denoted by β. And the power of a hypothesis test is the probability of correctly rejecting a false null hypothesis. Therefore, the power is equal to $1-\beta$.

In practice, we wish to find a good test that has both types of error probability that as small as possible. However, for a fixed sample size, it is usually impossible to achieve this goal. Although the goal has a highly tight connection with sample size, we are not going to give further discussion on the connection. For fixed sample size, *the level of significance* of a test decreases at the cost of raising the probability of its Type II Error (As the saying goes, you can't have your cake and eat it too). In searching for a good test, it is common to restrict consideration to tests at a specified level of significance, with typical choices being α=0.05 (significant) and 0.01 (highly significant). Within this class of tests we then search for tests that have Type II Error probability that as small as possible.

In statistics, a hypothesis test is specified in terms of a *test statistic*, a function of the sample. It is worth to mention here, even for the same problem, we can have different test statistics.

2.2. WMW(Wilcoxon-Mann-Whitney) test statistic

Comparison between two samples of observations is one of the problems we are faced with widely in fields such as agricultural production, clinical trial and financial analysis. A typical example is the comparison of a treatment with a control, where the null hypothesis of no treatment effect H_0 is test against the alternative of a beneficial effect H_1. The WMW test is one of the best-known tests for assessing whether the two independent samples come from the same distribution.

Suppose we have two independent samples without ties $\{x_1, \cdots, x_m\}$ and $\{y, \cdots, y_n\}$ from two unknown distribution F and G which need not to be discrete distributions. Here we only consider a one-tailed test

$H_0: F = G$ VS $H_1: F > G$

The WMW test is based upon comparing each observation of the two samples and counting the number of $x_i < y_j$. Obviously, if it is big, we are likely to accept H_1. The number can be treated as a test statistic called by WMW test statistic.

$$W(x,y) = \sum_{i=1}^{s}\sum_{j=1}^{t} I(x_i < y_j), \qquad (2.1)$$

Where $I(x_i < y_j) = 1$, if $x_i < y_j$, elsewhere $I(x_i < y_j) = 0$. Let $x = (x_1, \cdots, x_m)$ and $y = (y, \cdots, y_n)$, WMW test can be defined as

$$\phi = I(W(x,y) > k_0) + \gamma I(W(x,y) = k_0), \qquad (2.2)$$

where k_0 is a non-negative integer and $0 \le \gamma < 1$.

In fact, the rejection region of the WMW test is presented (Rejection region is the subsets of the sample space for which H_0 will be rejected). If $W(x,y) > k_0$, we reject H_0 and determine $F > G$. If, on the other hand, $W(x,y) < k_0$, we accept H_0 and say that $F = G$. Otherwise, if $W(x,y) = k_0$, namely the observations fall on the boundary of the rejection region, we reject H_0 with probability γ.

2.3. Admissibility

Suppose there are two tests ϕ_1 and ϕ_2 for the same hypothesis testing problem, if the two types error probabilities of ϕ_1 are less than those of ϕ_2,then test ϕ_1 is considered to be better than test ϕ_2. The formal definition is given below.

DEFINITION: For any two tests ϕ_1 and ϕ_2, we say ϕ_1 is better than ϕ_2 if

$$E\phi_1 \le E\phi_2, \qquad \forall F = G,$$

$$E\phi_1 \ge E\phi_2, \qquad \forall F > G,$$

and at least one inequality is strict for some F and G.

The definition posed above tells us whether one test is better than another one. After choosing a test statistic to handle problems in the practical life, we are definitely not willing to see that there is another test better than the one we choose. In that case, we will choose the “better” test naturally. Therefore, when searching for a test, we should follow the principle that there isn’t any other test that is better than ours. This principle will lead to many other tests that is well-matched with ours. Of course, these tests are “admissible”. The important concept of admissibility can be drawn from here.

DEFINITION: A test ϕ_0 is said to be admissible if there is no other test ϕ such that

$$\mathrm{E}\phi \le \mathrm{E}\phi_0, \qquad \forall \mathrm{F} = \mathrm{G}, \tag{2.3}$$

$\mathrm{E}\phi \ge \mathrm{E}\phi_0, \qquad \forall \mathrm{F} > \mathrm{G}.$

3. Main results and Proofs

THEOREM 3.1 Suppose the two samples have no ties, for testing the two sample hypothesis $H_0: F = G \; vs \; H_1: F > G$, WMW test is admissible in the class of all tests for arbitrary sample sizes m and n.

Note: the theorem is under the assumption of the samples without ties. And our proof can be applied to a two-tailed test $H_0: F = G \; vs \; H_1: F \neq G$ with a few changes. The solution of the admissibility of the Wilcoxon test is still a long way off, yet this theorem is the best one for the present. We believe that our proof offers a new way for the final solution of the admissibility of the Wilcoxon test.

Proof we will prove the theorem by contradiction. The basic idea is converting the expectation to a polynomial by constructing a discrete distribution, and completing the proof with the relation between the test statistic $\mathrm{W(x,y)}$ and the power of the polynomial.

Suppose that there is a partially symmetric test ϕ which is the same for any permutation of x or y such that as good as the WMW test ϕ_0. Thus ϕ satisfies

Equation (2.3). It is suffice to prove that $\phi = \phi_0$, that is, $\phi(x_0, y_0) = \phi_0(x_0, y_0)$, $\forall x_0, y_0$. Without loss of general, we assume there are no ties in x_0 and y_0.

Denote non-zero vector $a = (a_0, a_1)$ satisfied $a_0(m+n) + a_1 W(x_0, y_0) = 1$. And define $(m+n)$ tuples $b(x_0, y_0) = \big(b(x_1), \cdots, b(x_m), b(y_1), \cdots, b(y_n)\big)$ and $c(x_0, y_0) = \big(c(x_1), \cdots, c(x_m), c(y_1), \cdots, c(y_n)\big)$ as follows

$$b(x_i) = a_0 + a_1 \sum_{j=1}^{n} I(x_i < y_j), \qquad c(x_i) = a_0, \qquad i = 1, \cdots, m, \qquad (3.1a)$$

$$b(y_j) = a_0, \qquad c(y_j) = a_0 + a_1 \sum_{i=1}^{m} I(x_i < y_j), \qquad j = 1, \cdots, n. \qquad (3.1b)$$

Then we can define two discrete distribution functions as

$$F_0(x) = \sum_{i=1}^{m} b(x_i) I(x_i \le x) + \sum_{j=1}^{n} b(y_j) I(y_j \le x), \qquad (3.2a)$$

and

$$G_0(x) = \sum_{i=1}^{m} c(x_i) I(x_i \le x) + \sum_{j=1}^{n} c(y_j) I(y_j \le x). \qquad (3.2b)$$

It is easy to show that

$$F_0(x) - G_0(x) = a_1 \sum_{i=1}^{m} \sum_{j=1}^{n} [I(x_i \le x) - I(y_j \le x)] I(x_i < y_j).$$

Note that $I(x_i \le x) > I(y_j \le x)$ if and only if $x_i < y_j$, and $U(x_0, y_0) > 0$ implies that at least a pair say (x_{i_0}, y_{j_0}) such that $x_{i_0} < y_{j_0}$. We get

$$F_0(x) - G_0(x) \geq [I(x_{i_0} \leq x) - I(y_{j_0} \leq x)]I(x_{i_0} < y_{j_0}) > 0 . \qquad (3.3)$$

That is, $F_0 > G_0$.

Next we will evaluate $E_{(F_0,G_0)}[\phi - \phi_0]$. Note that F_0 and G_0 are discrete, the integral takes non-zero values only if the random vector (X_0, Y_0) takes all permutations of (x_0, y_0). Thus, it can be represented as a function of $b(x_0, y_0)$ and $c(x_0, y_0)$, indeed, a polynomial of a_0.

Without loss of general (W.L.O.G.), assume $x_1 < x_2 < \cdots < x_m$ and $y_1 < y_2 < \cdots < y_n$. Let π_1 and σ_1 be subsets of $[m]$ and π_2 and σ_2 be subsets of [n] such that

$$x_I = (x_{\pi_1}, y_{\pi_2}), y_I = (x_{\sigma_1}, y_{\sigma_2}).$$

Then $|\pi_1| + |\pi_2| = m, |\sigma_1| + |\sigma_2| = n$. Thus, we could recount all permutations in the integral in the following way:

$$0 \leq \frac{1}{m!\,n!} E_{F_0,G_0} (\phi - \phi_0)(X_0, Y_0)$$

$$= E_{F_0,G_0}(\phi - \phi_0)(X_{(0)}, Y_{(0)})$$

$$= \sum_{|\pi_1|+|\pi_2|=m,|\sigma_1|+|\sigma_2|=n} (\phi - \phi_0)(x_I, y_I)\, b(x_I)c(y_I)$$

$$= \sum_{|\pi_1|+|\pi_2|=m,|\sigma_1|+|\sigma_2|=n} (\phi - 1)(x_I, y_I) b(x_I) c(y_I) I(W(x_I, y_I) > k_0)$$

$$+ \sum_{|\pi_1|+|\pi_2|=m,|\sigma_1|+|\sigma_2|=n} (\phi - \gamma)(x_I, y_I) b(x_I) c(y_I) I(W(x_I, y_I) = k_0)$$

$$+ \sum_{|\pi_1|+|\pi_2|=m,|\sigma_1|+|\sigma_2|=n} \phi(x_I, y_I) b(x_I) c(y_I) I(W(x_I, y_I) < k_0)$$

where $b(x_I) = \prod_{i\in\pi_1} b(x_i) \prod_{j\in\pi_2} b(y_j)$, $c(y_I) = \prod_{i\in\sigma_1} c(x_i) \prod_{j\in\sigma_2} c(y_j)$.

In the case $W(x_0, y_0) = k > k_0 (> 0)$, the terms $(\phi - \phi_0)(x_I, y_I) b(x_I) c(y_I)$ are the polynomial of a_0. By lemma 3.1, the power of a_0 in $b(x_0)c(y_0)$, denoted by p, is the minimum, and $b(x_0)c(y_0)/a_1^p \to C(> 0)$ as $a_0 \to 0$. Particularly, if $W(x_I, y_I) < k$, then

$$\frac{b(x_I)c(y_I)}{a_0^p} \to 0, \qquad \text{as } a_0 \to 0 . \qquad (3.4)$$

In the inequality of the expectation given above, after dividing a_0^p, let $a_0 \to 0$, note that $0 \le \phi \le 1$, we get $C(\phi - 1)(x_0, y_0) = 0$. Hence $\phi(x_0, y_0) = 1$.

For $W(x_0, y_0) = k_0$, analogous augment leads to $\phi(x_0, y_0) \ge \gamma$.

Since $E_{(F,F)}[\phi - \phi_0] \le 0$ for all F by Equation (2.1) and

$$E_{(F,F)}[\phi - \phi_0]$$

$$= E_{(F,F)}[\phi - \phi_0]I(W > k_0) + E_{(F,F)}[\phi - \phi_0]I(W = k_0) + E_{(F,F)}[\phi - \phi_0]I(W < k_0)$$

$$= E_{(F,F)}[\phi - \phi_0]I(W = k_0) + E_{(F,F)}[\phi - \phi_0]I(W < k_0)\,,$$

$\phi = \gamma$ for $W = k_0$ and $\phi = 0$ for $W < k_0$. That is, $\phi = \phi_0$.

This completes the proof.

LEMMA 3.1 Assume $x_1 < x_2 < \cdots < x_m$ and $y_1 < y_2 < \cdots < y_n$. Given π_1 and σ_1 be subsets of $[m]$ and π_2 and σ_2 be subsets of [n] satisfied $|\pi_1| + |\pi_2| = m, |\sigma_1| + |\sigma_2| = n$, we have

$$b(x_I)c(y_I) \le b(x_0)c(y_0)\,. \tag{3.5}$$

With equality holds if and only if $W(x_I, y_I) \ge k$.

Proof By the definition of $b(x_0, y_0)$ and $c(x_0, y_0)$, we have $b(y_j) = c(x_i) \equiv a_0$. Furthermore, we can verify that

$$b(x_i) = a_0 + a_1 \sum_{j=1}^{n} I(x_i < y_j) \le a_0 + a_1 \sum_{j=1}^{n} I(x_{i-1} < y_j) = b(x_{i-1}),$$

$$c(y_j) = a_0 + a_1 \sum_{i=1}^{m} I(x_i < y_j) \le a_0 + a_1 \sum_{i=1}^{m} I(x_i < y_{j+1}) = c(y_{j+1}),$$

where $i = 2, \cdots, m$ and $j = 1, \cdots, n-1$. Thus,

$a_0 \le b(x_m) \le \cdots \le b(x_2) \le b(x_1), \qquad a_0 \le c(y_1) \le c(y_2) \le \cdots \le c(y_n).$

Therefore, for $|\pi_1| + |\pi_2| = m, |\sigma_1| + |\sigma_2| = n$

$$b(x_I)c(y_I) = b(x_{\pi_1})b(y_{\pi_2})c(x_{\sigma_1})c(y_{\sigma_2})$$

$$= b(x_{\pi_1})\, a_1^{|\pi_2|+|\sigma_1|} c(y_{\sigma_2})$$

$$\le b(x_{\pi_1})b(x_{\pi_1^c})c(y_{\sigma_2^c})c(y_{\sigma_2})$$

$$= b(x_0)c(y_0).$$

The equality holds if and only if $b(x_{\pi_1^c}) = a_1^{|\pi_2|} = a_1^{m-|\pi_1|}$,

$c(y_{\sigma_2^c}) = a_1^{|\sigma_1|} = a_1^{n-|\sigma_2|}$. In that case, $b(x_i) = c(y_j) = a_0$, $i \in \pi_1^c$, $j \in \sigma_2^c$. Further,

$x_i > y_j$ for $i \in \pi_1^c, j \in [n]$ or $i \in [m], j \in \sigma_2^c$. Hence

$$k = W(x_0, y_0)$$

$$= \Big(\sum_{i\in\pi_1, j\in\sigma_2} + \sum_{i\in\pi_1^c, j\in\sigma_2} + \sum_{i\in\pi_1, j\in\sigma_2^c} + \sum_{i\in\pi_1^c, j\in\sigma_2^c}\Big) I\{x_i < y_j\}$$

$$= \sum_{i\in\pi_1, j\in\sigma_2} I\{x_i < y_j\}$$

$$\le \sum_{i\in\pi_1, j\in\sigma_2} I\{x_i < y_j\} + |\pi_2| * |\sigma_1|$$

$$\leq \sum_{i\in\pi_1, j\in\sigma_2} I\{x_i < y_j\} + \sum_{i\in\pi_1, j\in\sigma_1} I\{x_i < x_j\} + \sum_{i\in\pi_2, j\in\sigma_2} I\{y_i < y_j\} + \sum_{i\in\pi_2, j\in\sigma_1} I\{y_i < x_j\}$$

$$= U\left(\left(x_{\pi_1}, y_{\pi_2}\right), \left(\left(x_{\sigma_1}, y_{\sigma_2}\right)\right)\right.$$

$$= U(x_I, y_I) .$$