Open Access Scientific Reports

Open Access Scientific Reports

Research Article Open Access

W Statistic: Ties Adjusted Two Sample Test

Oyeka ICA[1] and Okeh UM[2]*

[1]Department of Applied Statistics, Nnamdi Azikiwe University, Awka, Nigeria
[2]Department of Industrial Mathematics and Applied Statistics, Ebonyi State University Abakaliki, Nigeria

Abstract

This paper presents a non-parametric statistical method for the analysis of two-sample data. The test statistic here termed 'W' is intrinsically adjusted for the possible presence of ties observations in the sampled populations. The populations themselves may be measurements on as low as the ordinal scale and need not be continuous. The proposed method is illustrated with some data and shown to compare favorably with existing methods.

Keywords: Mann-Whitney U–Test; Non-parametric test; Two-sample W test; Intrinsically

Introduction

When analyzing random samples drawn from two independent populations that satisfy the necessary assumptions of normality and homogeneity for the use of two-sample parametric 't' test cannot be properly used. Use of non-parametric alternatives is then indicated and preferable.

In these situations the non-parametric methods that readily suggest themselves are the median test and the Mann-Whitney U test [1,2]. However both of these tests are often encumbered by problems of tied observations in the data. If the ties are few, the problem of tied observations may be resolved by dropping these tied observations and reducing the sample sizes appropriately in subsequent analyses. The problem of ties, if they are not too many also be resolved by randomly assigning the tied observations to one of the two groups into which the data set has been dichotomized by the common median of the pooled sample in the case of the median test, or by assigning the tied observations their mean ranks in the case of the Mann-Whitney U test. If however there are too many tied observations in the data then these approaches many not are satisfactory in resolving the problem of ties. This is because too many ties in the data often seriously compromise the power of the median test and the Mann-Whitney U test leading to possible erroneous conclusions.

We here propose a non parametric test simply termed the 'two-sample W test' that intrinsically adjusts the statistic W for the possible presence of ties in the data. This approach consequently obviates the need to require the populations of interest to be continuous or even numeric. The sample populations may here be measurements on as low as the ordinal scale.

The Proposed Method

Let X and Y be two independent populations that are measurements on at least the ordinal scale. Let x_i be the i^{th} observation in a random sample of size m drawn from population X, for i=1,2,…,m and y_i be the j^{th} observation in a random sample of size n independently drawn from population Y for j=1,2,…,n. Let

$$u_{ij} = \begin{cases} 1, \text{ if } x_i > y_i \\ 0, \text{ if } x_i = y_i \\ -1, \text{ if } x_i < y_i \end{cases} \tag{1}$$

For i=1,2,…,m; j=1,2,….,n

$$\text{Let } \pi^+ = P(u_{ij}=1), \pi^0 = P(u_{ij}=0), \pi^- = P(u_{ij}=-1) \tag{2}$$

Where

$$\pi^+ + \pi^0 + \pi^{-1} = 1 \tag{3}$$

Define

$$W = \sum_{i=1}^{m}\sum_{j=1}^{n} u_{ij} \tag{4}$$

We then have that

$$E(u_{ij}) = \pi^+ - \pi^- \tag{5}$$

$$Var(u_{ij}) = E(u_{ij})^2 - \left(E(u_{ij})\right)^2$$

$$= (1)^2\pi^+ + (0)^0\pi^0 + (-1)^2\pi^- - (\pi^+ - \pi^-)^2 \text{ or}$$

$$Var(u_{ij}) = \pi^+ + \pi^- - (\pi^+ - \pi^-)^2 \tag{6}$$

Using Equation 5 in Equation 4 we have that

$$E(W) = \sum_{i=1}^{m}\sum_{j=1}^{n} E(u_{ij}) = mn(\pi^+ - \pi^-) \tag{7}$$

$$Var(W) = Var\left(\sum_{i=1}^{m}\sum_{j=1}^{n} u_{ij}\right)$$

$$= E\left(\sum_{i=1}^{m}\sum_{j=1}^{n} u_{ij}\right)^2 - \left(E\left(\sum_{i=1}^{m}\sum_{j=1}^{n} u_{ij}\right)\right)^2$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{n} var\ u_{ij} + 2\binom{mn}{2} Cov(u_{ij}, u_{kl})$$

$$Now\ Cov\ (u_{ij}, u_{kl}) = E(u_{ij}, u_{kl}) - E(u_{ij})E(u_{kl})$$

But u_{ij}, u_{kl} takes the values of -1, 0 and 1. The value -1 is obtained when either u_{ij}=1 and u_{kl}=-1

or

u_{ij}=1 and u_{kl}=-1 with probability

$$\pi^+\pi^- + \pi^-\pi^+ = 2\pi^+\pi^-$$

And u_{ij}, u_{kl} assumes the value zero whenever u_{ij} and u_{kl} are both equal to zero

or

u_{ij}=0 and u_{kl} is either -1 or 1

***Corresponding author:** Okeh UM, Department of Industrial Mathematics and Applied Statistics, Ebonyi State University Abakaliki, Nigeria, E-mail: uzomaokey@ymail.com

Received January 16, 2013; **Published** February 28, 2013

Citation: Oyeka ICA, Okeh UM (2013) W Statistic: Ties Adjusted Two Sample Test. 2: 639 doi:10.4172/scientificreports.639

or

u_{kl}=0 and u_{ij} is either -1 or 1

and this occurs with probability $(\pi^0)^2 + 2\pi^0(\pi^- + \pi^+)$. Finally u_{ij}, u_{kl} assumes the value +1 when either u_{ij}=-1 and u_{kl}=-1

or

u_{ij}=-1 and u_{kl}=+1 with probability

$$\pi^-\pi^- + \pi^+\pi^+ = (\pi^-)^2 + (\pi^+)^2$$

Therefore

$$E(u_{ij}, u_{kl}) = (-1)(2\pi^+\pi^-) + (0)\left((\pi^0)^2 + 2\pi^0(\pi^- + \pi^+)\right) + (1)\left((\pi^-)^2 + (\pi^+)^2\right)$$
$$= (\pi^+ - \pi^-)^2$$

Hence

$$Cov(u_{ij}, u_{kl}) = (\pi^+ - \pi^-)^2 - (\pi^+ - \pi^-)^2 = 0$$

Hence using equation 6 and the above results we have that

$$Var(W) = mn\left(\pi^+ + \pi^- - (\pi^+ - \pi^-)^2\right) \tag{8}$$

The second term in Equation (8) that is $mn(\pi^+ - \pi^-)^2 = \frac{w^2}{mn}$ is not affected by any possible ties between the observations from population X and observations from population Y.

The first term $mn(\pi^+ + \pi^-)$=mn(1-π^0) has by the specifications in Equations 1 and 2 been adjusted for any possible ties in the data. If these adjustments had not been made, the presence of any ties in the data would have been completely ignored and possibly erroneously assumed to be absent meaning that $mn(\pi^+ + \pi^-)$ would be equal to mn since from Equation 2, (π^+-π^-) would have erroneously automatically been set equal to 1. This would lead to an over estimation of the variance of W with an error that is equal to $\frac{\pi^0}{1-\pi^0}$ of its true value when provision have been made for the presence of ties resulting in the inflation of the variance to $\frac{1}{1-\pi^0}$ of its true value for fairly large m and n, $(m, n \geq 8)$, increasing for smaller m and n. This bias cannot become zero unless there are no ties whatsoever between observations from population X and observation from population Y, or the ties are so few in practice, it is reasonable to assume that their effect is negligibly small and can be ignored. This assumption is not necessary here because as we already pointed out, the model specifications in Equations 1 and 2 have already taken care of the possibility of ties. Hence the variance of W in Equation 8 is not affected by any ties that may exist in the data.

Estimating for π^+, π^0 and π^-

Now π^+, π^0 and π^- are respectively the probabilities that on the average a randomly selected subject from population X performs better, as well as, or worse than a randomly selected subject from population Y. These probabilities are estimated as respectively the relative frequencies of occurrence of 1s, 0s, and -1s in the frequency distribution of the mn values of these numbers in u_{ij}, i=1,2,....,m; j=1,2,...n. Thus if f_+,f_0 and f are the frequencies of occurrence of 1,0 and -1 respectively in the frequency distribution of the mn values of u_{ij} then

$$\hat{\pi}^+ = \frac{f^+}{mn}; \hat{\pi}^0 = \frac{f^0}{mn}; \hat{\pi}^- = \frac{f^-}{mn} \tag{9}$$

Not however that from Equation (7) that the difference between population X and Y in their probabilities of positive response, $\pi^+ - \pi^-$, is estimated as

$$\hat{\pi}^+ - \hat{\pi}^- = \hat{\pi}^+ - \hat{\pi}^- = \frac{w}{mn} = \frac{f^+ - f^-}{mn} \tag{10}$$

Hence if we replace π^+ and π^- in Equation (8) with their sample estimates in Equation 9 and then solve Equation 8 and 10 simultaneously for $\hat{\pi}^+ and\ \hat{\pi}^-$ we obtain

$$\left.\begin{aligned} \hat{\pi}^+ &= \frac{1}{2mn}\left(Var(w) + w\left(\frac{w}{mn} + 1\right)\right) \\ \hat{\pi}^- &= \frac{1}{2mn}\left(Var(w) + w\left(\frac{w}{mn} - 1\right)\right) \\ \hat{\pi}^0 &= 1 - \frac{1}{mn}\left(Var(w) + \frac{w^2}{mn}\right) \end{aligned}\right\} \tag{11}$$

Where

$$w = f^+ - f^-; Var(w) = f^+ + f^- - \frac{w^2}{mn}$$

Testing of Hypothesis

Interest may be in testing the null hypothesis that subjects in population X on the average perform better (or worse) than subjects in population Y. This is equivalent to testing the null hypothesis

$$H_0: \pi^+ - \pi^- \geq \delta_o \text{ versus } H_1: \pi^+ - \pi^- < \delta_o \text{, say, } (-1 < \delta_o < 1) \tag{13}$$

Now for sufficiently large m and n $(m, n \geq \delta_o)$, the test statistic

$$\chi^2 = \frac{\left(w - mn(\pi^+ - \pi^-)\right)^2}{Var(w)} = \frac{\left(w - mn\delta_o\right)^2}{mn\left(\pi^+ + \pi^-\right) - \frac{w^2}{mn}} \tag{14}$$

has approximately a chi-square distribution with 1 degree of freedom and may be used to test the null hypothesis of Equation 13 H_0 is rejected at the a level of significance if

$$\chi^2 \geq \chi^2_{1-\alpha,1} \tag{15}$$

Otherwise H_0 is accepted

If the null hypothesis of Equation 13 is that subjects in population X and Y perform equally well

$(H_0: \pi^+ - \pi^- = \delta_o = 0)$, then the test statistic of Equation 14 reduces to

$$\chi^2 = \frac{w^2}{mn\left(\pi^+ + \pi^-\right) - \frac{w^2}{mn}} \tag{16}$$

In practical applications, π^+ and π^- in Equations 14 and 16 are replaced with their sample estimates $\hat{\pi}^+$ and $\hat{\pi}^-$ respectively in Equation 9.

Modified Sign Test

If m=n and i=j is put in Equation (1). That is if

$$u_{ij} = u_{ii} = u_i = \begin{cases} 1, if\ x_i > y_i \\ 1, if\ x_i = y_i \\ -1, if\ x_i < y_i \end{cases} \tag{17}$$

for i=1,2,…,n

Then the above method can be used as an alternative approach to the sign test, adjusted for the possible presence of tied observations in the data, for either paired samples or two independent samples of equal size n. Using Equation 2 in Equation 13 we have

$$E(u_i) = \pi^+ - \pi^- and\ Var(u_i) = \pi^+ + \pi^- - (\pi^+ - \pi^-)^2 \tag{18}$$

Then equation 4 becomes

$$w = \sum_{i=1}^{n} u_i,$$

So that

$$E(w) = n(\pi^+ - \pi^-) \tag{19}$$

and

$$Var(w) = n\left((\pi^+ + \pi^-) - (\pi^+ - \pi^-)^2\right) = n(\pi^+ + \pi^-) - \frac{w^2}{n} \quad (20)$$

Estimates of π^+, π^0 and π^- for the sign test, similar to Equation 9 then becomes respectively

$$\hat{\pi}^+ = \frac{f^+}{n}; \hat{\pi}^0 = \frac{f^0}{n}; \hat{\pi}^- = \frac{f^-}{n} \quad (21)$$

Where f^+, f^0 *and* f^- are respectively the number of 1s, 0s and -1s in the frequency distribution of the 'n' values of these numbers in u_i, i=1,2,....n. To test a null hypothesis often using the sign test $(H_0 : \pi^+ - \pi^- = \delta_0 = 0)$, we may use the test statistic

$$\chi^2 = \frac{w^2}{n\left((\pi^+ - \pi^-) - (\pi^+ - \pi^-)^2\right)} = \frac{w^2}{n(\pi^+ + \pi^-) - \frac{w^2}{n}}$$

Which for sufficiently large $n(n \geq 8)$ has approximately a chi-square distribution with 1 degree of freedom. H_0 is rejected at the α level of significance if Equation 15 is satisfied, otherwise H_0 is accepted.

Illustrative Example

We now illustrative the above procedures with the following data on the performance reported in later grades by random samples of male and female applicants in an interview for a certain employment.

Male: C A B$^-$ A$^+$ C B A C C$^-$ C$^-$ D E F

Female: E A B F F F E B B$^+$ C F A$^+$ E E C$^+$ E

If we now label male X and female Y then the values of u_{ij} of Equation are easily and more clearly presented in a tabular form (Table 1).

Results

From table 1 we have that

f^+=123, f^0=17 and f^-=68. Hence with m=13 and n=16, we have using Equation 9 that $\hat{\pi}^+ = \frac{123}{208} = 0.591; \hat{\pi}^0 = \frac{17}{208} = 0.082; \hat{\pi}^- = \frac{68}{208} = 0.327$ and from Equation 12, w=123-68=55 with variance (Eqn 8) estimated as $Var(w) = (13)(16)(0.591 + 0.327) - \frac{(53)^2}{(13)(16)} = 190.944 - 14.542 = 176.401$.

Hence the test statistic for the null hypothesis of Equation (13) with $\delta_0 = 0$ is from Equation (16)

$$\chi^2 = \frac{55^2}{176.461} = \frac{3025}{176.401} = 17.148$$

which with 1 degree of freedom is highly statistically significant.

It would be instructive to use the present data to compare the proposed method with the Mann-Whitney U-test. To apply the Mann-Whitney U-test, we rank the combined samples from the highest grade A$^+$ to the lowest grade F and then separate and sum the ranks assigned to each of the samples as follows (Table 2).

Now the Mann-Whitney U statistic based on the ranks of X is

$$U = mn + \frac{m(m+)}{2} - R_x$$

$$= (13)(16) + 13\frac{(13+1)}{2} - 167.5 = 299 - 167.5 = 131.5$$

The mean of U is estimated as $U = \frac{mn}{2} = \frac{(13)(16)}{2} = \frac{208}{2} = 104$

And standard deviation

$$se(U) = \sqrt{\frac{mn(m+n+1)}{12}} = \sqrt{\frac{(13)(16)(13+16+1)}{12}} = \sqrt{520} = 22.804$$

Hence the Mann-Whitney U test statistic is $z = \frac{u - u_u}{se(u)} = \frac{131.5 - 104}{22.804} = \frac{27.5}{22.804} = 1.206$

(P-value=0.1131), which is not statistically significance at the 5 percent level.

Hence using the Mann-Whitney U test with the data we would accept the null hypothesis that male and female candidates perform equally well in the job interview, while the same null hypothesis would be rejected using the proposed method. Notice that for the present data. The variance of the test statistic of the proposed method is only 176.401 compared with a variance of $(22.804)^2$=520.00.

Thus for the present data, at least, the Mann-Whitney U–test is seen to be less efficient than the proposed method and the test results suggest that the Mann-Whitney U–test is likely to lead to an acceptance of a false null hypothesis (Type II Error) more frequently than the proposed method, and hence is likely to be powerful.

Now to illustrate the application of the proposed method to a two-sample sign test problem, we use the following data on family size preferences.

Random samples of twelve newly married couples were asked to state their family size preferences. The results are shown in table 3.

We apply equation (17) to the sample data in table 3 to obtain the values of u_i shown in the last column of the table. We thus have

Col=Female Row=Male	E	A	B	F	F	F	E	B	B$^+$	C	F	A$^+$	E	E	C$^+$	E
C	1	–1	–1	1	1	1	1	–1	–1	0	1	–1	1	1	–1	1
A	1	0	–1	1	1	1	1	1	1	1	1	–1	1	1	1	1
B$^–$	1	–1	–1	1	1	1	1	–1	–1	1	1	–1	1	1	1	1
A$^+$	1	1	–1	1	1	1	1	1	1	1	1	0	1	1	1	1
C	1	–1	–1	1	1	1	1	–1	–1	0	1	–1	1	1	1	1
B	1	–1	0	1	1	1	1	0	–1	1	1	–1	1	1	1	1
A	1	0	1	1	1	1	1	1	1	1	1	–1	1	1	1	1
C	1	–1	–1	1	1	1	1	–1	–1	0	1	–1	1	1	–1	1
C$^–$	1	–1	–1	1	1	1	1	–1	–1	–1	1	–1	1	1	1	1
C$^–$	1	–1	–1	1	1	1	1	–1	–1	–1	1	–1	1	1	1	1
D	1	–1	–1	1	1	1	1	–1	–1	–1	1	–1	1	1	1	1
E	0	–1	–1	1	1	1	0	–1	–1	–1	1	–1	0	0	–1	0
F	–1	–1	–1	0	0	0	–1	–1	–1	–1	0	–1	–1	–1	1	1

Table 1: Values of u_{ij}.

Male Grade	Rank of X	Female Grade	Rank of Y
X		Y	
C	13.5	E	21.5
A	4	A	4
B⁻	10	B	8
A⁺	1.5	F	27
C	13.5	F	27
B	8	F	27
A	4	E	21.5
C	13.5	B⁺	8
C⁻	16.5	B⁺	6
C⁻	16.5	C	13.5
D	18	F	27
E	21.5	A⁺	1.5
F	27	E	21.5
		E	21.5
		C⁺	11
		E	21.5
Sum of Rank for X(R_x)	167.5	Sum Rank for Y(R_y)	267.5

Table 2: Ranks of illustrative Data for use with the Mann-Whitney U-test.

Couple (i)	Husband	Wife	u_i
1	4	5	–1
2	2	5	–1
3	–6	5	1
4	2	6	–1
5	7	5	1
6	–2	9	–1
7	4	4	0
8	2	6	–1
9	–8	8	0
10	5	5	0
11	4	4	0
12	4	5	–1

Table 3: Family size preferences by a random sample of newly married couples.

$$f^+ = 2; f^0 = 4 \; and \; f^- = 6 \; so \; that \; \hat{\pi}^+ = \frac{2}{12} = 0.167;$$

$$\hat{\pi}^0 = \frac{4}{12} = 0.333, \hat{\pi}^- = \frac{6}{12} = 0.500$$

$$Also \; w = f^+ - f^- = 2 - 6 = -4 \; and \; from \; Equation \; 20$$

$$Var(w)12(0.167 + 0.500) - \frac{-4}{12} = 8.004 - 1.333 = 6.6771$$

From Equation 22 we have that the test statistic for testing the null hypothesis of Equation 13 with $\delta_0 = 0$ is $\chi^2 = \frac{(-4)^2}{6.671} = \frac{16}{6.671} = 2.398(P-value = 0.1131)$ which with 1 degree of freedom is not statistically significant at the 5 percent level. We may thus conclude that newly married husband and wives in the community of interest do not seem to differ in their family size preferences.

We now for comparative purposes re-analyze the data using ordinary sign test for two equal independent samples. We note from the last column of table 3 that there are x=2+signs (1s) and 4 zero's (ties). Hence the effective sample size to be used here is n=12-4=8. We therefore calculate the binomial probability for $P(x \le 2)$ with p=0.5 as $P(x \le 2) = \sum_{x=0}^{2} \binom{8}{x} (0.5)^8 = (1+8+28)(0.05)^8 = 37(0.5) = 0.1445$ which is greater than $\alpha/_2 = 0.05/_2 = 0.025$. We therefore do not also here reject the null hypothesis of no difference between newly married husbands and wives in their family size preferences.

Discussion

It is however seen from the above two results that even though the proposed method and the ordinary sign test both do not here reject the null hypothesis, the attained significance levels indicate that the sign test is nonetheless likely to accept a false null hypothesis (Type II Error) more frequently than the proposed method and is hence likely to be less powerful, at least for the present data.

Summary and Conclusion

We have here presented and discussed a non-parametric statistical method, termed the two-sample W test modified or adjusted for the possible presence of tied observation in the data. The sampled populations may be continuous. The proposed method may also be used as an alternative to and an improvement over the two-sample sign test for independent samples. The proposed method is illustrated with some data and shown to compare favorably with the Mann-Whitney U test.

References

1. Gibbons JD (1971) Nonparametric statistical inferences. Mc Graw Hill, New York, USA.
2. Oyeka CA (1996) An introduction to applied statistical methods. (3rdedn) Nobern avocation publishing Company Enugu, Nigeria.

3. Oyeka ICA, Utazi CE, Nwosu CR, Ebuh GU, Ikpegbu PA, et al. (2010) A Statistical Comparison of Test Scores for Non-parametric Approach. Journal of mathematics sciences 21: 77-87.
4. Siegel S, Castellan NJ (1988) Nonparametric statistics for the behavioral sciences. (2ndedn) Mc-Graw Hill, USA.