IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728,p-ISSN: 2319-765X, Volume 7, Issue 4 (Jul. - Aug. 2013), PP 52-56
www.iosrjournals.org

Modified Intrinsically Ties Adjusted Mann-Whitney U Test

[1.] Oyeka I. C. A and [2]Okeh U. M

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

Abstract: *This paper proposes to develop a ties adjusted or corrected approach to the usual Mann-Whitney U test that structurally makes provision for the correction of the U statistic and its variance for the possible presence of ties observation between the two sampled populations. The modification makes it unnecessary as is the case with the ordinary Mann-Whitney U test for the populations to be continuous with this method data analysis may precede without any problem when the populations are measurements on as low as the ordinal scale and may not be numerical. The method which is based on both magnitude and direction enables the estimation of the probabilities that a randomly selected subject from one population performs or scores higher, as well as or lower than a randomly selected subject from the other population. The procedure is illustrated with some sample data and shown to be at least as powerful as the modified median test and the usual Mann-Whitney U test.*

I. 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 (Oyeka,2009), then this method would be properly used for this purpose. However if these assumptions are not satisfied or if the populations are measurements on as low as the ordinal scale, then the parametric alternatives is then indicated and preferable (Spiegel,1998). In these situations, non-parametric methods that readily suggest themselves are the median test and the Mann-Whitney U test (Gibbons, 1971;Oyeka,2009).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 analysis. The problem of ties, if they are not two many may 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(Oyeka,2009;Gibbons,1971Munzel and Brunner,2002). If however there are two many tied observations in the data then these approaches may not be satisfactory in resolving the problem of ties. This is because too many ties in the data often seriously compromise the power of thee median test and the Mann-Whitney U test, leading to possible erroneous conclusions. Fortunately some workers have recently developed procedures for intrinsically adjusting or correcting test statistics in two sample problems for possible presence of ties in the data (see for example, Oyeka et al,2009;Munzel and Brunner,2002;Oyeka and Okeh,2012;Okeh,2009).An expression also exists for correcting the variance of the Mann-Whitney U test statistic for the presence of tied observations in the sampled populations (Hays,1973).However evaluating this expression in practical applications is often rather tedious and cumbersome. We here propose to develop a test statistic that intrinsically and structurally adjusts or corrects the Mann-Whitney U test and makes provision for the possible presence of tied observations between sampled populations. The procedures therefore obviates the need to require the populations to be continuous or even numeric. The sampled populations may be measurements on as low as the ordinal scale and need not be continuous.

II. The Proposed Method

Let x_{ij} be the ith observation in a random sample of size n_j drawn from population j,for $i=1,2,\ldots,n_j;j=1,2$.The populations may be measurements on as low as the ordinal scale and need not be continuous or numeric(Oyeka et al,2012). To develop the proposed method we would first pool the two samples into one combined sample of size $n=n_1+n_2$.let x_l be the lth observation in such a pooled sample for $l=1,2,..,n$.Suppose these pooled 'n' observations are now ranked from the smallest to the largest or largest to smallest specifically let r_{ij} be the rank assigned to the ith observation from population 'j' in the combined ranking of the 'n' observations, for $i=1,2,\ldots,n_j;j=1,2$.
Let

$$u_{gh} = \begin{cases} 1, if \ x_{g1} > x_{h2} \\ 0, if \ x_{g1} = x_{h2} \\ -1, if \ x_{g1} < x_{h2} \end{cases} \qquad 1$$

$$for \ g = 1, 2, \ldots, n_1 ; h = 1, 2, \ldots, n_2$$

In other words u_{gh} assume the values 1,0 and -1 respectively if the gth observation from population 1 is greater than equal to,or less than the hth observation from population 2.

Let

$$\pi^{+} = P(u_{gh} = 1); \pi^{0} = P(u_{gh} = 0); \pi^{-} = P(u_{gh} = -1) \qquad 2$$

$$where$$

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

$$Also \ let$$

$$\dot{r}_{gh} = r_{g1} - r_{h2} \qquad 4$$

be the difference between the ranks assigned to the gth and hth observations from population 1 and 2,for g=1,2,..,n_1;h=1,2,…,n_2.

Also let

$$W = \sum_{g=1}^{n_1}\sum_{h=1}^{n_2} |r_{gh}|.u_{gh} = \sum_{g=1}^{n_1}\sum_{h=1}^{n_2} |r_{g1} - r_{h2}|.u_{gh}$$

Or

$$W = n_{2.}R_{.1} - n_{2.}R_{.2} \qquad 5$$

$$where \ R_{.1} = \sum_{g=1}^{n_1} r_{g1} ; R_{.2} = \sum_{h=1}^{n_2} r_{h2} \qquad 6$$

Are respectively the sums of the ranks assigned to observations from populations x_1 and x_2 in the combined ranking of these observations from the two populations? We have here for ease of presentation but without loss of generality assumed that $n_{2.}R_{.1}$ is greater than $n_{1.}R_{.2}$ .This assumption does not affect the final results of the analysis because its only effect is a possible change in the sign of W while all statistical tests use the square of this statistic.

Now

$$E(u_{gh}) = \pi^{+} - \pi^{-} ; Var(u_{gh}) = \pi^{+} + \pi^{-} - (\pi^{+} - \pi^{-})^2 \qquad 7$$

Also

$$E(W) = \sum_{g=1}^{n_1}\sum_{h=1}^{n_2} |r_{gh}|.E(u_{gh})$$

$$= (\pi^{+} - \pi^{-})\sum_{g=1}^{n_1}\sum_{h=1}^{n_2} |r_{g1} - r_{h2}| = (\pi^{+} - \pi^{-})\left(n_2 \sum_{g=1}^{n_1} r_{g1} - n_1 \sum_{h=1}^{n_2} r_{h2} \right) \text{Also}$$

$$Or$$

$$E(W) = \left(n_{2.}R_{.1} - n_{1.}R_{.2}\right)\left(\pi^{+} - \pi^{-}\right) \qquad 8$$

$$Var(W) = \sum_{g=1}^{n_1}\sum_{h=1}^{n_2} r_{gh}^2 .Var(u_{gh}) = \left(\pi^{+} + \pi^{-} - (\pi^{+} - \pi^{-})^2\right).\sum_{g=1}^{n_1}\sum_{h=1}^{n_2}\left(r_{g1} - r_{h2}\right)^2$$

$$= \left(\pi^{+} + \pi^{-} - (\pi^{+} - \pi^{-})^2\right).\left(\sum_{g=1}^{n_1}\sum_{h=1}^{n_2} r_{g1}^2 + \sum_{g=1}^{n_1}\sum_{h=1}^{n_2} r_{h2}^2 - 2\sum_{g=1}^{n_1}\sum_{h=1}^{n_2} r_{g1}.r_{h2} \right)$$

$$or$$

$$Var(W) = \left(\pi^{+} + \pi^{-} - (\pi^{+} - \pi^{-})^2\right)\left(n_{2.}R_{.1}^{*2} + n_{1.}R_{.2}^{*2} - 2R_{.1}R_{.2}\right) \qquad 9$$

$$where$$

$$R_{.1}^{*2} = \sum_{g=1}^{n_1} r_{g1}^2 ; R_{.2}^{*2} = \sum_{h=1}^{n_2} r_{h2}^2 \qquad 10$$

Are respectively the sums of squares of the ranks assigned to observations from populations 1 and 2.Note that W in Eqn 5 and in particular its variance in Eqn 9 have by specifications been intrinsically adjusted for the possible presence of tied observations between the populations.

Now π^+, π^0 *and* π^- are respectively the probabilities that observations or scores by subjects from population X_1 are on the average greater than,equal to,or less than observations or scores by subjects from population X_2.Their sample estimates are respectively

$$\hat{\pi}^+ = \frac{f^+}{n_1.n_2}; \hat{\pi}^0 = \frac{f^0}{n_1.n_2}; \hat{\pi}^- = \frac{f^-}{n_1.n_2} \qquad 11$$

Where f^+, f^0 *and* f^- are respectively the number of times or number of subjects in the sample drawn from population X_1 whose scores are greater than ,or equal to or less than the scores by subjects drawn from population X_2.In other words f^+, f^0 *and* f^- are respectively the number of 1s,0s and -1 in the frequency distribution of the $n_{1.}n_2$ values of these numbers in u_{gh} ,g=1,2,…,n_1;h=1,2,..,n_2.

Now

$$W = n_{2.}R_{.1} - n_{1.}R_{.2} \qquad 12$$

Is the sample estimate of the expected sum of the ranks assigned to subjects from population X_1 less the sum of the ranks assigned to subjects from population X_2.Note that by the specifications of Eqns 1-3 and the introduction of π^0,Eqn 8 and hence 2 has been adjusted for the possibility of ties between populations X_1 and X_2,since $\pi^+ - \pi^-$ is not affected, that is has been adjusted for the possible presence of ties between the two populations. The corresponding sample variance of W is from Eqn 9

$$Var(W) = \left(n_{2.}R_{.1}^{*2} + n_{1.}R_{.2}^{*2} - 2R_{.1}R_{.2}\right)\left(\pi^+ + \pi^- - (\pi^+ - \pi^-)^2\right) \qquad 13$$

Which has also by the specifications of Eqn 1-3 been adjusted for the possible presence of ties between the sampled populations. This is because the right hand side of Eqn 13 namely

$$\left(n_{2.}R_{.1}^{*2} + n_{1.}R_{.2}^{*2} - 2R_{.1}R_{.2}\right)\left(\pi^+ + \pi^-\right) = \left(n_{2.}R_{.1}^{*2} + n_{1.}R_{.2}^{*2} - 2R_{.1}R_{.2}\right)\left(1 - \hat{\pi}^0\right)$$

has also been adjusted for any ties between the two populations. If this adjustment had not been made, $\hat{\pi}^0$ would automatically have been set equal to zero, so that this later expression would have been erroneously calculated as $n_{2.}R_{.1}^{*2} + n_{1.}R_{.2}^{*2} - 2R_{.1}R_{.2}$,thereby inflating the variance of W as an increasing function of π^0.Now under the null hypothesis of equal population medians the expectation would be that any randomly selected observation or subject from population X_1 is as likely to be greater as less than any randomly selected subject from population X_2 and hence equally likely to be assigned the same rank as any randomly selected subject from population X_2 in the combined ranking of subjects from the two populations. In other words, the expected value of W,E(W),the expected difference between the sums of the ranks assigned to subjects from the two populations would be zero or equivalently $\pi^+ - \pi^- = 0$.Thus a null hypothesis that may be tested would be

$$H_0 : E(W) = 0 \; or \; H_0 : \pi^+ - \pi^- = 0 \qquad 14$$

Versus any desired one-sided or two sided alternative. The null hypothesis of Eqn 14 may be tested using the test statistic

$$\chi^2 = \frac{W^2}{Var(W)} = \frac{\left(n_{2.}R_{.1} - n_{1.}R_{.2}\right)^2}{\left(n_{2.}R_{.1}^{*2} + n_{1.}R_{.2}^{*2} - 2R_{.1}R_{.2}\right)\left(\pi^+ + \pi^- - (\pi^+ - \pi^-)^2\right)} \qquad 15$$

Which under H_0 has approximately the chi-square distribution with 1 degree of freedom for sufficiently large n_1 and n_2 $(n_1, n_2 \geq 8)$ .The null hypothesis of Eqn 14 is rejected at that level of significance if

$$\chi^2 \geq \chi^2_{1-\alpha;1} \qquad 16$$

Otherwise H_0 is accepted.

We have in this paper only presented a procedure that may enable one intrinsically adjust a test statistic in a two sample problem for possible presence of tied observations between the populations but not within the populations themselves(Okeh and Oyeka,2012. This is because although in tests concerning the equality of population medians, ties within these populations may still affect the results obtained but these effects may often be less serious than the effects of ties between the populations which are more likely to invalidate conclusions.

III. The Illustrative Example

We here illustrate the proposed method with two sample data on the lengths of hospitalization in days of patients admitted in a certain hospital for two types of illnesses namely Malaria and Hypertension (Table 1)

Table 1 Lengths of Hospitalization (in days)of Malaria and Hypertension patients

Data on Malaria				Data of hypertension			
Malaria Patients	No of days	r_{g1}	r_{g1}^2	Hypertension Patients	No of days	r_{h2}	r_{h2}^2
1	11	19	361	1	4	8	64
2	3	5	25	2	17	23.5	552.25
3	1	1	1	3	5	11	121
4	5	11	121	4	16	22	484
5	2	2.5	6.25	5	7	14	196
6	3	5	25	6	9	16.5	272.25
7	4	8	64	7	18	25	625
8	2	2.5	6.25	8	9	16.5	272.25
9	7	14	196	9	17	23.5	552.25
10	7	14	196	10	13	20.5	420.25
11	3	5	25	11	10	18	324
12				12	5	11	121
13				13	4	8	64
14				14	13	20.5	420.25
Total		87($R_{.1}$)	1026.5($R_{.1}^{*2}$)			238($R_{.2}$)	4488.5($R_{.2}^{*2}$)

Now to illustrate the proposed method we have here apply Eqn 1 to the data of Table 1 to obtain values of u_{gh} which yields $f^+ = 18; f^0 = 6; f^- = 130.$ Hence from Eqn 11 we have that

$$\hat{\pi}^+ = \frac{18}{(11)(14)} = \frac{18}{154} = 0.117; \hat{\pi}^0 = \frac{6}{154} = 0.039; and\ \hat{\pi}^- = \frac{130}{154} = 0.844.$$

We next pool the two sample together and rank the observations from the smallest to the largest. The results are shown in Table 1.Using these ranks in Eqn 12 we have that

W=(14)(87)-(11)(238)=1218-2618=—1400

Whose variance is obtained using Eqn 13 as

$$Var(W) = \left((14)(1026.5) + (11)(4488.5) - 2(87)(238)\right)\left(0.117 + 0.844 + -(0.117 - 0.833)^2\right)$$
$$= \left(14371 + 49373.5 - 41412\right)\left(0.961 - 0.529\right)$$
$$= \left(63744.5 - 41412\right)\left(0.432\right) = \left(22332.5\right)\left(0.432\right) = 9647.64$$

Therefore to test the null hypothesis of equal population medians (Eqn 14) we have from Eqn 15 that

$$\chi^2 = \frac{(-1400)^2}{9647.64} = \frac{1960000}{9647.64} = 203.158(P - value = 0.0000)$$

Which is highly statistical significant, indicating that malaria and hypertension patients have different median lengths of hospitalization. If we had applied the Mann-Whitney U test to analyze. The sample data of Table 1 we would have using the ranks in Table 1 obtain after the ranking of the combined observations that the sums of the ranks assigned to observations from populations X_1 and X_2 are respectively $R_{.1}$=87 and $R_{.2}$ =238.Hence the Mann-Whitney u test statistic is

$$U = \frac{n_1.n_2 + n_1(n_1+1) - R_{.1}}{2} = (11)(14) + 11(6) - 87 = 154 + 66 - 87 = 133$$

With mean

$$E(u) = \frac{(11)(14)}{2} = 77$$

and var*iance*

$$Var(u) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} = \frac{11(14)(26)}{12} = \frac{4004}{12} = 333.667$$

The corresponding Mann-Whitney U test statistic is

$$\chi^2 = \frac{\left(u - E(u)\right)^2}{Var(u)} = \frac{\left(133 - 77\right)^2}{333.667} = \frac{3136}{333.667} = 9.399(P - value = 0.0000)$$

Which with 1 degree of freedom is highly statistically significant again indicating that malaria and hypertension patients have statistically different median lengths of hospitalization. However the calculated values of the corresponding chi-square statistic clearly show that the Mann-Whitney U Test is likely to lead to an acceptance of a false null hypothesis (Type II Error)more frequently and is hence likely to be less powerful than the proposed modified test that is intrinsically adjusted for ties between the populations(Oyeka and Okeh 2012. Statistically significant again supporting the findings using proposed method. Note however that the proposed method is still likely to be more powerful than the modified median test which is itself more powerful than the usual Mann-Whitney U Test that unlike the former two procedures is not modified or adjusted for possible ties in the sampled populations.

IV. Summary And Conclusion

We have in this paper proposed and presented the statistical procedure that enables the Mann-Whitney U statistic to be intrinsically and structurally adjusted for possible presence of tied observations between the two populations of research interest. The method makes it unnecessary to require as is the case with the usual Mann-Whitney U test, at least theoretically than the populations of interest to be continuous. The provision for the possible presence of ties in the proposed method makes this requirement unnecessary. The data of research interest may be measurements on as low as the ordinal scale. The proposed method which is based on both magnitude and direction enables the estimation of the probabilities that a randomly selected subject from one population performs or scores higher (better, more), as well as (the same as) or lower (worse, less) than a randomly selected subject from the other population. The proposed method is illustrated with some sample data and shown to be at least as powerful as the Mann-Whitney U Test and the modified median test.

Reference

[1]. Argesti A.(1992).Analysis of Ordinal Paired Comparison data. Appl Statist 41:287-297.
[2]. Gibbons, JD (1971). Nonparametric Statistical Inferences. McGraw Hill, New York.
[3]. Hay, William (1973).Statistics for the social sciences: Holt, Rinchart and Winston Inc New York, PP 778-780.
[4]. Munzel U, Brunner E.(2002).An exact paired rank test. Biometrical Journal 44:584-593.
[5]. *Oyeka I.C.A (2009)*: An Introduction to Applied Statistical Methods, Nobern Avocation Publishing Company, pp 530-533.
[6]. Okeh,UM.(2009).Statistical analysis of the application of Wilcoxon and Mann-Whitney U test in medical research studies. Biotechnology and Molecular Biology Reviews Vol. 3 (5), pp. 105-110.
[7]. *Oyeka I.C.A and Okeh U.M.(2012).* Intrinsically Ties Adjusted Sign Test by Ranks. J Biom Biostat 2012, 3:8
[8]. *Spiegel M.R (1998)*: Theory and Problems of Statistics, McGraw-Hill Book Company, New York. Pp 261-264.