Selcuk Journal of Agriculture and Food Sciences

https://dergipark.org.tr/tr/pub/selcukjafsci

Research Article

(2023) 37 (2), 223-231 DOI:10.15316/SJAFS.2023.022 e-ISSN: 2458-8377 Selcuk J Agr Food Sci

Comparison of Student – t, Welch's t, and Mann – Whitney U Tests in Terms of Type I Error Rate and Test Power

Malik Ergin[1,*], Ozgur Koskan[1]

[1] University of Applied Sciences of Isparta, Faculty of Agriculture, Department of Animal Science, Isparta, Türkiye

HIGHLIGHTS

- Student – t, Welch's – t, and Mann – Whitney U tests were examined by various simulation combinations.
- It is indicated that Welch's t-test is robust for preserving the type I error rate when the distribution is normal.
- In real-life applications, using Welch t-test as an alternative to these two methods is recommended.

Abstract

In this study, we compared the Student's t-test, Welch's t-test, and Mann-Whitney U test, in terms of their type I error rate and statistical power when the assumptions of parametric tests are violated in different situations. Materials used in this study, consisted of random numbers generated using the Numpy library in the Python programming language. All random numbers were generated from a normal distribution with N (0, 1) parameters. Balanced and unbalanced experimental conditions were simulated 50 000 times for each combination. The study revealed that, in comparison to other tests, Welch's t - test was particularly more conservative in terms of type I error rate. It was discovered that the Student-t test had higher power values than the Mann-Whitney U test, mainly when only a small sample size of observations was used for the analysis. This simulation study indicated that Welch's t - test is robust for preserving type I error rate when the distribution is normal. Therefore, in practice, the use of Welch t-test is recommended based on the findings of this study. One of the recommendations of this study is that the tests in question should also be evaluated in cases where observations have different distributions.

Keywords: Type I error rate; Test power; Student – t test; Mann – Whitney U test; Welch's – t test; Simulation

1. Introduction

Depending on the situation, parametric or non–parametric statistical approaches are preferred in studies where the means or median values of two groups are compared. Parametric tests, as it is well–known, act by some parameters in the probability distribution to which the observation values belong. The Student – t-test, which compares the means of two independent groups, is also a parametric method that needs the normal distribution of observations and the homogeneity of group variances. Nonparametric methods, on the other

Citation: Ergin M, Koskan O (2023). Comparison of Student – t, Welch's t, and Mann – Whitney U Tests in Terms of Type I Error Rate and Test Power. *Selcuk Journal of Agriculture and Food Sciences*, 37(2),223-231. https://doi.org/10. 15316/SJAFS.2023.022
***Correspondence:** malikergin@isparta.edu.tr
Received date: 19/10/2022
Accepted date: 05/03/2023

hand, assume that the observation values are not obtained from a certain probability distribution. Therefore, it is represented in the literature as distribution – free. It is widely known that parametric tests are more robust than nonparametric tests when assumptions are provided, even with a small sample size. Furthermore, it has been reported that when the number of observations in both groups is equal (balanced design), the Student – t test is a powerful test even if the homogeneity of the variances, which is considered to be the most important assumption of the parametric tests, is violated. Moreover, it is reported that the heterogeneity of variances in experiments where the number of observations is not equal (unbalanced design) causes the probability of type I error determined at the beginning of the experiment to not be maintained at 5% (Zimmerman and Zumbo 1993).

In general, the options that researchers may use when the assumptions are not met can be summarized as (i) the Welch – t-test, which is one of the parametric alternative approaches, or (ii) the Mann – Whitney U test, which is one of the non–parametric approaches. The parametric alternative to the Student – t test is the Welch – t-test, which was developed by correcting the degrees of freedom of the independent two–group t–test in experiments where group variances were not homogeneous (Derrick et al. 2016). As a result of previous simulation studies, it has been reported that the Welch – t-test is more powerful when the assumption of homogeneity of group variances is not met (Oshima et al. 1991; Keselman et al. 1991). Welch–t-test has been reported to better preserve the type I error probability in unbalanced designs with unequal sample sizes (Zimmerman 2004). Keselman et al. (2004) have claimed that Welch – t-test is not affected by the heterogeneity of variances but is influenced by non–normality of the observations. In addition, Winter (2013) supported through a simulation study that applying the Welch–t-test on experiments with very small sample sizes is problematic.

In this simulation study, Student – t, Welch – t, and Mann – Whitney U tests which compare the mean or median values of two groups, were examined in terms of type I error rate and test power by designing various simulation combinations.

2. Materials and Methods

In the present study, random numbers generated from normal distribution by using the "random" function in the Numpy library of the Python programming language were used. Student – t, Welch – t, and Mann – Whitney U tests were compared in terms of both type I error rate and test power values by constituting various situations. Detailed information on the simulation design for type I error rate and test power is presented in Table 1. The simulation scenario used to calculate Type I error rates is as follows:

1. Define a variable as count = 0.
2. Generate two samples of data from normal (0,1) distribution using the "random" function to consider the null hypothesis.
3. Perform all tests at the predetermined significance level (α = 5%).
4. Compute and store all p – values in a list variable.
5. If p – value ≤ 0.05, increase count variable by one.
6. Perform these procedures 50 000 times.
7. Calculate the type I error rates as follow:

$$\frac{(\text{Number of Rejected } H_0 \text{ Hypothesis})}{(\text{Number of Total Simulations})}$$

The simulation scenario for calculating power values is as follows:

1. Define a variable as count = 0.

2. Generate two samples of data from normal (0,1) distribution using the "random" function to consider the null hypothesis.
3. Add constant value to the mean of the first group to create the desired standard deviation differences (Repeat this step for all standard deviation differences such as 0.75, 1, and 1.5).
4. Perform all tests at the predetermined significance level (α = 5%).
5. Compute and store all p – values in a list variable.
6. If p – value ≤ 0.05, increase count variable by one.
7. Perform these procedures 50 000 times.
8. Calculate the type I error rates as follow:

$$\frac{(\text{Number of Rejected } H_0 \text{ Hypothesis})}{(\text{Number of Total Simulations})}$$

Table 1. Simulation design.

Type I Error Rate		**Test Power**		
n	$\boldsymbol{\sigma^2:\sigma^2}$	**n**	$\boldsymbol{\sigma^2:\sigma^2}$	**Δ**
8,8 15,15 25,25 35,35 45,45 65,65	1:1 1:3 1:5 1:10	8,8 10,10 15,15 25,25 30,30 45,45	1:1	0.75 1 1.5
8,15 15,20 20,27 35,45 50,60	1:3 3:1 1:7 7:1			

Student – t-test

In general, population variance is unknown and therefore the variance estimated from the sample should be used. The t-test that is developed by William S. Gosset as the "Student" nickname, is suitable where the population variance is unknown. The difference between the two sample variances should be equal to the variances calculated from samples drawn from the same population, indicating that the variances are homogeneous. When considering whether the difference between the means of the samples is part of the distribution, there are two estimates of the population variance, as each sample variance is an estimate of the population variance for the calculated t-value. Therefore, the weighted average of these variance estimates, based on their degrees of freedom, will provide a more reliable estimate of the population variance. Student – t test value calculated by Equation (1) and Equation (2).

$$t = \frac{\overline{A} - \overline{B}}{S_D} \tag{1}$$

$$S_D = \sqrt{\frac{\sum d_A^2 + \sum d_B^2}{(n_A - 1) + (n_B - 1)} * \frac{(n_A + n_B)}{n_A * n_B}} \tag{2}$$

It shows the theoretical t distribution with $(n_A - 1) + (n_B - 1)$ degrees of freedom.

Welch – t-test

It has been reported that deterioration of variance homogeneity causes changes in the performance of the Student – t test, both in terms of type I error rate and test power (Delacre et al. 2017). For this reason, Welch 1947) developed an approach based on the separated variances and correction of degrees of freedom. When the assumptions of the Student's t-test are not met, the Welch t-test can be used as a parametric alternative. The t statistic in question may be calculated by Equation (2.3). Some statistical software calculates the Welch t-test value using a formula that differs from the generally accepted one. For instance, IBM SPSS software calculates degrees of freedom with Equation (2.4), while Minitab software calculates degrees of freedom according to Equation (2.5).

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{3}$$

$$S.D. = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}} \tag{4}$$

$$S.D. = \frac{(VAR_1 + VAR_2)^2}{\left(\frac{VAR_1{}^2}{n_1 - 1} + \frac{VAR_1{}^2}{n_2 - 1}\right)} \tag{5}$$

Mann – Whitney U test

In the comparison of two independent groups, in cases where the assumptions of the t-test, which is a parametric test, are not met, that is, the distribution is not normal or the distribution type of the feature in question is not known, the non-parametric Mann Whitney U test is used (Mendes 2012). The working principle of the Mann-Whitney U test is to test whether the two groups represent samples from the same population, according to the group medians.

Observations belonging to two groups, A and B, are brought combined into a single group (nA+ nB = N). After the observations are ranked with ordinal scores from 1 to N, U-test statistics are calculated through Equation (2.6) and Equation (2.7).

$$n_A > n_B : U = T_A - \frac{n_A(n_A + 1)}{2} \tag{6}$$

$$n_A < n_B : U = T_B - \frac{n_B(n_B + 1)}{2} \tag{7}$$

Since the U test statistic shows a uniform distribution, it provides the determination of the critical value necessary for testing the H_0 hypothesis. If the said U test statistic value is greater than the critical value, the H_0 hypothesis is rejected, meaning that these two groups do not represent the same population (McKnight and Najab 2009; Zar 1984).

To evaluate whether all tests are conservative to various simulation combinations, Bradley's criterion of robustness was considered. The results of Bradley's (1978) study indicate that when testing at the significance level of 5.00%, a robust test's actual Type I error rate should be between 4.50% and 5.50%. Furthermore, Murphy and Myors (2014) reported that a power level that reaches or exceeds 80% is typically considered to be adequate. Therefore, we will consider 80% is a standard for sufficiency in test power.

3. Results

Type I error rates in simulation scenarios where the number of observations in both groups is balanced ($n_1 = n_2$) 8, 15, 25, 35, 45, 65, and variance ratios are 1:1, 1:3, 1:5, 1:10, it is given in Table 2 as a percentage. When the variance ratios are equal in a 1:1 ratio, the type I error rate has consistently remained at 5% across all three tests in the simulation design with equal observations. It has been remarked due to the increase in the number of observations.

In the case of a 1:3 ratio where the variance ratio deviates from homogeneity, the type I error rates seemed robust according to Bradley's criterion of robustness. The Welch's t-test only satisfied Bradley's criterion for robustness in all sample sizes when the variance ratios were 1:1.5. When the variance ratio was 1:10, the type I error rates calculated after conducting the Student t-test and Mann-Whitney U test for all sample sizes exceeded the Bradley's criterion. However, the type I error rates seemed conservative after conducting Welch's t-test.

Table 2. The type I error rates at various variance ratios occur when the number of observations is equal.

n	$\sigma^2:\sigma^2$	S_t	W_t	MWU
(8,8)	1:1	5.0	4.9	5.0
	1:3	5.4	5.0	5.4
	1:5	5.8	5.2	6.1
	1:10	6.0	5.1	6.8
(15,15)	1:1	4.9	4.9	4.5
	1:3	5.2	4.7	5.3
	1:5	5.7	5.0	5.9
	1:10	6.1	5.1	6.6
(25,25)	1:1	5.0	5.0	4.7
	1:3	5.4	5.0	5.4
	1:5	5.6	4.9	5.8
	1:10	6.1	5.1	6.8
(35,35)	1:1	5.0	5.0	5.0
	1:3	5.5	5.1	5.5
	1:5	5.6	4.9	5.9
	1:10	6.2	5.2	6.8
(45,45)	1:1	4.9	4.9	4.9
	1:3	5.3	4.9	5.3
	1:5	5.7	5.0	5.8
	1:10	6.2	5.2	6.6
(65,65)	1:1	5.0	5.0	4.9
	1:3	5.5	5.1	5.5
	1:5	5.5	4.9	5.7
	1:10	6.2	5.2	6.6

S_t: Student – t-test, W_t: Welch – t-test, MWU: Mann – Whitney U test, n: Number of observations

The type I error rates in simulation designs where the number of observations per group is unbalanced ($n_1 \neq n_2$) and variance ratios are 1:3, 1:7, 3:1, and 7:1, it is given in Table 3 as a percentage. Across all the combinations of variance ratios and sample sizes we tested, Welch's t-test yielded results that satisfied the Bradley's criterion of robustness. Similar to Bindak's (2014) study, cases such as positive and negative associations were also tried. Based on the results of the large variances and large sample sizes simulations (positive association), Student – t test did not satisfy Bradley's criterion. Yet, Mann – Whitney U test seemed robust with increasing number of observations. In the case of the large variance in small sample sizes (negative association), the Student – t and Mann – Whitney U tests exceeded the Bradley's criterion and seemed non-robust. Welch's t-test produced conservative results for all sample size and associations of variances.

Table 3. The type I error rates at various variance ratios occur when the number of observations is not equal.

n	σ^2: σ^2	S_t	Wt	MWU
(8,15)		2.5	4.8	3.2
		1.7	5.0	3.3
		9.5	5.2	7.3
		13.3	5.3	9.4
(15,20)		3.7	4.9	4.1
		3.3	5.1	4.7
		6.9	5.0	6.1
		8.5	5.2	7.7
(20,27)	1:3	3.7	5.1	4.6
	1:7	3.1	5.0	4.8
	3:1	7.2	5.1	6.6
	7:1	8.5	4.9	7.9
(35,45)		3.8	5.1	4.8
		3.3	5.0	5.2
		6.5	4.9	6.3
		7.7	5.0	7.7
(50,60)		4.0	5.0	5.1
		3.7	5.0	5.7
		6.2	5.0	6.1
		6.8	5.0	7.6

S_t: Student – t-test, W_t: Welch – t-test, MWU: Mann – Whitney U test, n: Number of observations

Table 4 presents the rejection rates of the null hypothesis (i.e., the results related to the test power) as a percentage when there are certain differences between group means. Assuming homogeneous variances and a standard deviation difference of 0.75, the power values for all tests remained below 80% until the sample sizes of the groups reached 25. The test power exceeded the desired power of 80% except for the Mann – Whitney U test when the sample sizes are 30. Furthermore, the power values for all three tests exceed 90%, when the difference between population means in standard deviation 1 and sample size is 25. The test power values exceeded 90% in small sample sizes due to the increase in standard deviation difference. It is more pronounced when the standard deviation difference is 1.5. For instance, while the sample size is 10 in both groups, the test powers calculated after all three tests is exceeded the desired level of 80%. When the sample size is 45, the power values calculated could reach 100%.

Table 4. The test powers when the number of the observations in the groups is equal and variances are homogenous.

Δ	0.75			1			1.5		
n	S_t	W_t	MWU	S_t	W_t	MWU	S_t	W_t	MWU
(8,8)	28.59	27.89	27.2	46.19	45.29	43.95	79.87	79.14	77.22
(10,10)	35.51	34.93	31.53	55.92	55.41	50.94	88.41	88.10	84.96
(15,15)	50.6	50.3	46.98	75.51	75.32	71.63	97.71	97.69	96.71
(25,25)	73.7	73.64	70.63	93.38	93.36	91.74	99.93	99.93	99.88
(30,30)	81.44	81.41	79.04	96.91	96.90	95.94	99.99	99.99	99.99
(45,45)	94.11	94.11	92.93	99.72	99.72	99.58	100	100	100

S_t: Student – t-test, W_t: Welch – t-test, MWU: Mann – Whitney U test, n: Number of observations, Δ: Standard deviation differences

4. Discussion

In balanced simulation designs where the number of observations in the groups is equal and the variances are homogeneous, all methods met Bradley's criterion. However, when the assumed population variances of the groups deviated from the homogeneity, the Student t-test and Mann-Whitney U test increased the type I error rate and still seemed conservative. The Welch's t-test successfully maintained the type I error rate of 5% for all sample sizes and seemed conservative. In terms of these findings, our study's results are consistent with the studies of Derrick et al. (2016), Kasuya (2001), and Ruxton (2006).

In the unbalanced simulation designs where the number of observations in both groups was not equal, positive, and negative associations of variances were tested. For all combinations, Welch's t-test yielded results very close to the type I error rate determined at the beginning of the experiment 5% and may be definable as conservative. These findings are consistent with the studies of Ahad and Yahaya (2014), Bindak (2014), and Ruxton (2006).

When considering the power values, the Student t-test and Welch's t-test performed better than the Mann-Whitney U test, especially for small sample sizes. The reason for the Mann-Whitney U test having such a trend could be the assumption that the populations from which we took our groups were normally distributed. This result is similar to the study conducted by Bindak (2014). While the variances were homogeneous for the Student's t-test, a desired power value of 80% could be achieved with sample sizes of 30, 25, and 15 for all differences in standard deviations, respectively. Aslan et al. (2021) reported that to achieve a test power of 80-90% for the t-test with effect sizes of 0.75, 1, and 1.5, sample sizes of 33, 23, and 13 were required, respectively. In addition, Koskan et al. (2022) demonstrated that as the standard deviation differences between two treatments increase, the minimum sample sizes required to achieve a test power of 85-95% were decreased. These results are consistent with our findings.

In this study, the Student t-test, Welch's t-test, and Mann-Whitney U test were evaluated in terms of both Type I error rate and test power in two groups assumed to be normally distributed. Especially, it was noteworthy that Welch's t-test was more conservative in terms of type I error than the other tests. In terms of power values, it is seen that the Student t-test is more powerful than the Mann-Whitney U test, especially for small sample sizes. Therefore, the distributions of the populations from which the groups were assumed to be taken should be investigated by creating distributions with more skewed distributions.

5. Conclusions

In the present study, consistent with the literature, it was observed that the power values of all tests increased as expected when the number of observations in the groups and the differences in standard deviation between the group averages increased. All statistical tests produced similar results in terms of test power. Again, if the distribution is normal and the variances are homogeneous, it is seen in parallel with the literature that all tests preserve the type I error at the level of 5%. When the homogeneity of variances starts to be violated, it is observed that the type I error values of the Mann-Whitney U test are higher than 5%. In addition, Welch's t-test is better than other tests (Student – t, and Mann – Whitney U) in unbalanced designs for preserving the type I error rate at the level of 5%. The Student – t test is more powerful than Mann – Whitney U test, especially when studied with a small sample size. In conclusion, it should be investigated how this situation changes by constructing the distributions of the assumed populations from which the groups are taken to have more skewed distributions.

In real life, according to the literature, the most commonly used and reliable statistical method for comparing two groups is the Student t-test when the assumptions were met. Although the effectiveness of the t-test decreases relatively when the assumptions were not met, the Mann-Whitney U test is commonly used

in such cases. The Welch t-test, which is independent of the assumptions, has been shown to be an alternative to the other two methods in terms of both type I error rate and test power. It is observed to produce better results in some extreme cases. Therefore, in real-life applications, using Welch t-test as an alternative to these two methods is recommended.

Author Contributions: Conceptualization, M.E. and O.K.; methodology, M.E.; software, M.E.; validation, M.E., and O.K.; investigation, M.E.; writing—original draft preparation, M.E.; writing—review and editing, M.E. and O.K.; supervision, O.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest

References

Ahad NA, Yahaya SSS (2014). Sensitivity analysis of Welch's t -test. *AIP Conference Proceedings. American Institute of Physics Inc. C.*, 1605(1): 888-893. Doi:10.1063/1.4887707

Aslan E, Koşkan Ö, Altay Y (2021). Determination of the sample size on different independent K group comparisons by power analysis. *Türkiye Tarımsal Araştırmalar Dergisi*, 8(1): 34-41. Doi:10.19159/tutad.792694

Bindak R (2014). Comparision Mann-Whitney U Test and Students' t Test in terms of type 1 error rate and test power: a Monte Carlo simulation study. *Afyon Kocatepe University Journal of Sciences and Engineering*, 14(1): 5-11. Doi:10.5578/fmbd.7380

Bradley JV (1978). Robustness. *British Journal of Mathematical and Statistical Psychology*, 31(2):144-152. Doi:10.1111/j.2044-8317.1978.tb00581.x

Delacre M, Lakens D, Leys C (2017). Why psychologists should by default use welch's t-Test instead of student's t-Test. *International Review of Social Psychology*, 30(1): 92-101. Doi:10.5334/irsp.82

Derrick B, Toher D, White, P (2016). Why Welch's test is type 1 error robust. *The Quantitative Methods in Psychology*, 12(1). Doi:10.20982/tqmp.12.1.p030

Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Oliphant TE (2020). Array programming with NumPy. *Nature*, 585(7825): 357-362. Doi:10.1038/s41586-020-2649-2

Kasuya E (2001). Mann-Whitney U test when variances are unequal. *Animal Behaviour*, 61: 1247-1249. Doi:10.1006/anbe.2001.1691

Keselman HJ, Keselman JC, Shaffer JP (1991). Multiple pairwise comparisons of repeated measures means under violation of multisample sphericity. *Psychological Bulletin*, 110(1): 162. Doi:10.1037/0033-2909.110.1.162

Keselman HJ, Othman AR, Wilcox RR, Fradette K (2004). The new and improved two-sample t test. *Psychological Science*, 15(1): 47-51. Doi:10.1111/j.0963-7214.2004.01501008.x

Koskan O, Koknaroglu H, Altay Y (2022). Determination of minimum number of animals in comparing treatment means by power analysis. *MVZ Córdoba*, 27(2): 1-11. Doi:10.21897/rmvz.2572

McKnight PE, Najab J (2010). Mann-Whitney U Test. *The Corsini encyclopedia of psychology*, 1(1). Doi:10.1002/9780470479216.CORPSY0524

Murphy KR, Myors B, Wolach A (2014). Statistical Power Analysis: A Simple And General Model For Traditional And Modern Hypothesis Tests. *Routledge*, New York, USA. p. 244. Doi: 10.4324/9781315773155

Ruxton GD (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test. *Behavioral Ecology*, 17(4): 688-690. Doi:10.1093/beheco/ark016

Welch BL (1947). The generalization of "Student's" problem when several different population variances are involved. *Biometrika*, 34(1-2): 28-35. Doi:10.1093/biomet/34.1-2.28

Winter JCF (2013). Using the Student's t-test with extremely small sample sizes. *Practical Assessment, Research, and Evaluation Practical Assessment*, 18(1): 10. Doi:10.7275/e4r6-dj05

Zimmerman DW (2004). Conditional probabilities of rejecting h_0 by pooled and separate-variances t tests given heterogeneity of sample variances. *Communications in Statistics Part B: Simulation and Computation*, 33(1): 69-81. Doi:10.1081/SAC-120028434

Zimmerman DW, Zumbo BD (1993). Rank transformations and the power of the Student t test and Welch t test for non-normal populations with unequal variances. *Canadian Journal of Experimental Psychology*, 47(3): 523. Doi:10.1037/h0078850