Vol.11.Issue.3.2023 (July-Sept.)

http://www.bomsr.com

Email:editorbomsr@gmail.com

RESEARCH ARTICLE

BULLETIN OF MATHEMATICS AND STATISTICS RESEARCH

A Peer Reviewed International Research Journal

INTERNATIONAL STANDARD SERIAL NUMBER 2348-0580

THE VALIDITY OF T-TEST, MANN-WHITNEY TEST AND Z TEST FOR TESTING SIGNIFICANT DIFFERENCES BETWEEN TWO SAMPLE MEANS WHEN THE SAMPLE SIZE IS BETWEEN 10 AND 30

RAMNATH TAKIAR

Apartment no. 11, Building 9, 3 rd Floor, Mangol Expoprt Town, 1st Khorro, Olympic Street
Sukhbaatar District, Ulaanbaatar, Mongolia.
Email: ramnathtakiar@gmail.com, ramnath_takiar@yahoo.co.in
&
Scientist G – (Retired)
National Centre for Disease Informatics and Research (NCDIR), Indian Council of Medical Research (1978-2013) Bangalore – 562110, Karnataka, India
DOI:10.33329/bomsr.11.2.1

Ramnath Takiar

ABSTRACT

For the sample size below 30, to compare two sample means, t-test is advocated. In my recent, study (Takiar 2023), for small samples (≤10), at 5% α level, it was shown that the t-test can pick up 11.1%, 18.6% and 31.1% of the expected significant differences, between two sample means for the sample size of 3,6 and 9, respectively. Further, it was shown that Z test with estimated variance attains relatively a higher negative validity than the t-test. In the present study, an attempt is made to evaluate the validity of t-test as compared to Mann Whitney test and Z-test when the sample size is above 10 and below 30? The four set of Normal population of size 200, with different means and standard deviations, are generated and termed as P1, P2, P3 and P4 and comparisons are made between the means of P1 with P2 on one hand and P3 and P4 on the other hand. From each population, 500 Random samples of size 12, 18 and 24 are generated. The comparison between two sample means is carried out using the t-test, Mann Whitney test, and Z-test with estimated sample variance. For the study purposes, three α levels are chosen namely 5%, 10% and 15%. At 5% α level, the negative validity of the t-test and Mann Whitney test is observed to be

around 38%, 53% and 64% for the sample size of 12, 18 and 24, respectively. At 10% α level, the corresponding negative validity figures are 51%, 65% and 75%. At 15% α level, the comparable negative validity figures are 60%, 72%, 81%. At 5% and 10% α level, the t-test and Mann Whitney test are not suitable when the sample size is below 20. The results suggest that Z-EV test scores 5-8% more negative validity as compared to other two tests. When the sample size is between 15 and 30, Z-EV test can be used dropping the t-test and Mann Whitney test. Further, it is advised that α level should be 10% or 15% not 5%, as traditionally used.

Keywords: t-test, Mann Whitney test, Z-EV test, a level, Negative validity,

INTRODUCTION

A Z test is used for comparisons of two sample means, if the samples are large enough (≥ 30) and they are drawn from the same normal population with the known variance. In case of small samples, with the sample size below 30, the t-test is advocated (Snedecor and Cochran 1967, Gupta and Kapoor 2001, Gupta 2012). In my recent, study (Takiar 2023), for small samples (≤10), at 5% α level, it was shown that in more than 90% of the cases, the t-test is good enough in picking up correctly the non-significant differences. However, it picks up only 11.4%, 18.6% and 31.1% of the expected significant differences, between two sample means for the sample size of 3,6 and 9, respectively, when they are known to have been drawn from the two different normal populations. Further, it was shown that for small samples (≤ 10), the Z-test with estimated variance (Z-EV) can pick up 10% to 20% relatively more expected significant differences than the t-test. It was concluded therefore that even for small samples, Z-EV test is relatively a better choice than the t-test and the Mann Whitney test. The problem with the t-test is that it tends to accept H0 very often then required. In the present study, an attempt is made to evaluate the validity of t-test as compared to Mann Whitney test and Z-EV test when the sample size is above 10 and below 30?

OBJECTIVES

- The objectives of the present study when the sample size is between 10 and 30, are:
- To evaluate the validity of t-test?
- The evaluate the performance of the t-test in comparison to Z-EV test and Mann Whitney test? and
- To evaluate and compare the effect of sample size on the validity of the selected Significant tests?

MATERIAL AND METHODS

GENERATION OF NORMAL POPULATIONS

The four sets of Normal populations of size 200, with different means and standard deviations, are generated with the help of the function key “Random number Generation” provided in StatPlus 7.6.5. For the study purposes, they are termed as P1, P2, P3 and P4 and comparisons are made between the means of P1 with P2 on one hand and P3 and P4 on the other hand. The major parameters of the above four populations are provided in Table 1.

Based on the mean comparisons, the distribution of the population P1 is found to be significantly different from that of the population P2 on one hand and the distribution of the Population P3 with that of P4, on the other hand.

Table 1: Description of Parameters of the Selected Populations with the result of Significance test

Parameter	NORMAL POPULATION			
	P1	P2	P3	P4
N	200	200	200	200
Mean	55.5	44.2	65.8	76.1
SD	16.01	11.7	16.79	17.94
Z Value	8.03		5.95	
P-value	**< 0.001**		**< 0.001**	
Critical t value at 0.001	3.09		3.09	

SAMPLE SELECTION

The Scheme of sample selection by different population and size is shown in Table 2. From each population, 500 Random samples of size 12, 18 and 24 are generated using the key "Random Sample" available with StatPlus 7.6.5. Thus, for each Normal Population, 1500 samples are drawn, as shown below:

Table 2: Scheme of Sample Selection according to Population, Sample size and Number of Samples drawn

Population	Sample size			**Total**
	12	18	24	
P1	500	500	500	**1500**
P2	500	500	500	**1500**
P3	500	500	500	**1500**
P4	500	500	500	**1500**
Total	**2000**	**2000**	**2000**	**6000**

SCHEME OF COMPARISONS BETWEEN THE MEANS OF SAME NORMAL POPULATIONS

The number of possible mean comparisons according to each sample size and Population is shown in Table 3. In total, the above scheme allows us to have 6000 mean comparisons.

Table 3: The Scheme of Mean Comparisons among Samples of same Population by the Varying Sample size

Sample Size	Samples for Comparisons	Number of Mean comparisons	Samples for Comparisons	Number of Mean comparisons
12	P1 with P1*	500	P3 with P3*	500
	P2 with P2*	500	P4 with P4*	500
18	P1 with P1*	500	P3 with P3*	500
	P2 with P2*	500	P4 with P4*	500
24	P1 with P1*	500	P3 with P3*	500
	P2 with P2*	500	P4 with P4*	500
Total		3000	Total	3000

* - Samples with Changed Sequence

SCHEME OF COMPARISONS BETWEEN THE MEANS OF DIFFERENT NORMAL POPULATIONS

The number of possible mean comparisons according to each sample size and Population is shown in Table 4.

Table 4: The Scheme of Mean Comparisons among the Samples of same Population by the Varying Sample size

Sample Size	Samples for Comparisons	Number of Mean comparisons	Samples for Comparisons	Number of Mean comparisons
12	P1 with P2	500	P3 with P3*	500
	P1 with P2*	500	P4 with P4*	500
18	P1 with P2	500	P3 with P3*	500
	P1 with P2*	500	P4 with P4*	500
24	P1 with P2	500	P3 with P3*	500
	P1 with P2*	500	P4 with P4*	500
Total		3000	Total	3000

* Samples with Changed Sequence

In above scheme, the sample means are compared between the populations of P1 and P2 on one hand and P3 and P4 on the other hand. Additionally, comparisons are made between sample means of P1 with P2* on one hand and P3 with P4* on other hand. In total, for three selected sample sizes, 6000 mean comparisons are made.

The formulae used are as follows: $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$; Z -EV= $\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

Where $S^2 = \frac{1}{(n-1)} \Sigma (x_i - \bar{x})^2$ $s^2 = \frac{1}{n} \Sigma (x_i - \bar{x})^2$

SIGNIFICANCE TEST FOR TWO SAMPLE MEAN COMPARISONS

In the present study, the comparison between two sample means is carried out using the following three tests namely t-test, Mann Whitney test, and Z test with estimated sample variance, termed as Z-EV test. For the study purposes, three α levels are chosen namely 5%, 10% and 15%.

VALIDITY OF T-TEST AND Z-TEST

WHEN SAMPLES ARE DRAWN FROM THE SAME NORMAL POPULATION

In this case, the Null Hypothesis is that "The sample means compared are not significantly different from each other." Since, the samples are known to have been drawn from the same normal population, it is logical not to reject the Null Hypothesis. Thus, the validity of the test under consideration can be defined as follows:

Positive Validity = [Number of non-significant differences found correctly /500] *100

Thus, the positive validity will range from 0 to 100.

WHEN SAMPLES ARE DRAWN FROM TWO DIFFERENT NORMAL POPULATIONS

In this case, as usual, the Null Hypothesis is that "The sample means compared are not significantly different from each other." Since, the samples are known to have been drawn from the different normal populations, it is logical to reject the Null Hypothesis. Thus, the validity of the test under consideration can be defined as follows:

Negative Validity = [Number of significant differences found correctly /500] *100

Like, the positive validity, the negative validity will also range from 0 to 100.

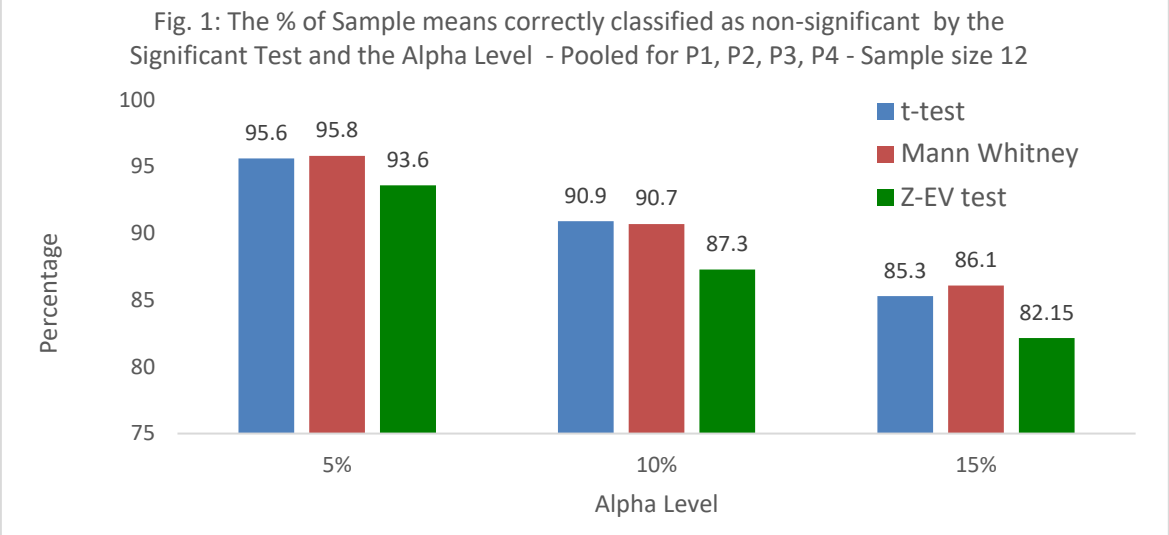
ANALYSIS OF THE DATA

For simultaneous comparisons of means of all the 500 samples, a program developed on Excel 2019, is utilized. In Z-EV test, the estimated sample variance is utilized. The function keys available on Excel 2019, are utilized to arrive at the probability of the calculated Z and t-statistic. To obtain the result by the Mann Whitney test, SPSS program, 2023 version, is utilized. To assess which test is better in picking up either the significant or the non-significant differences, correctly, the results obtained by the t-test, Mann Whitney and Z-EV test, at the given α level, are obtained and compared.

RESULTS

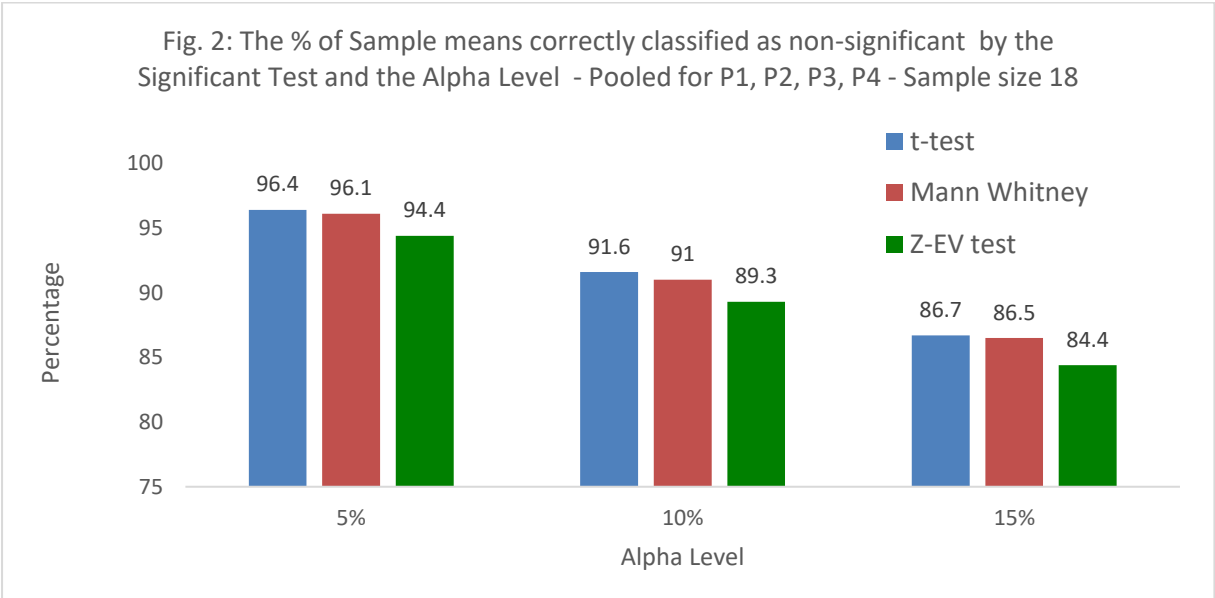
COMPARISON OF TWO SAMPLE MEANS WHEN SAMPLES ARE DRAWN FROM THE SAME NORMAL POPULATION

The Results of t-test, Mann Whitney test and Z-EV test, for testing the expected non-significant differences between sample means by the varying α levels, for the sample size of 12, are shown in Fig.1.

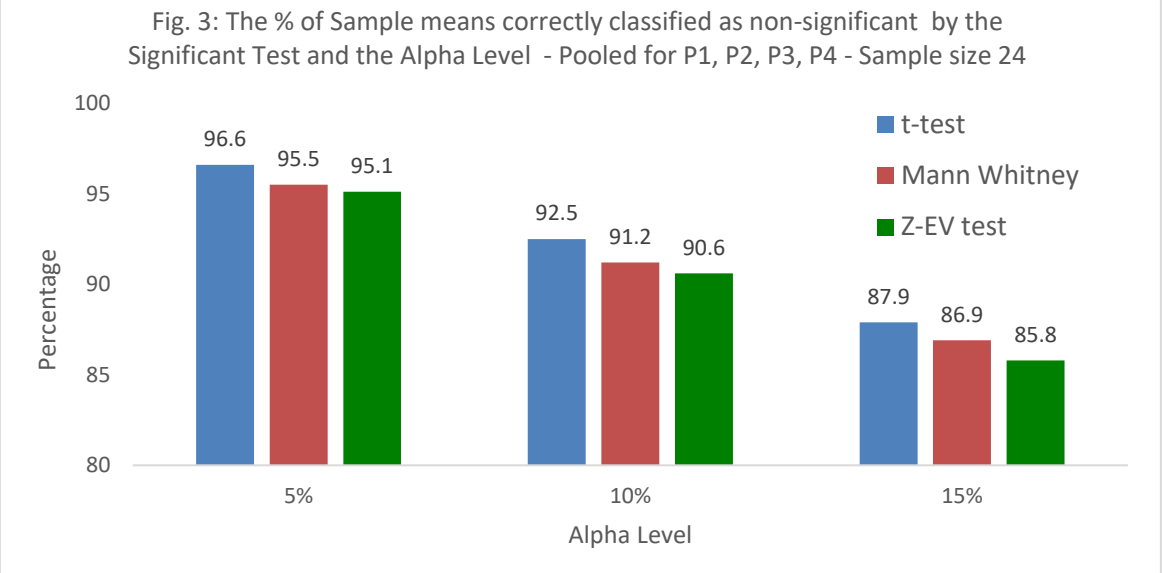


Assuming 80% as the cut-off level, the positive validity is acceptable at all the three α levels. The difference in the validity is not much between the t-test, Mann Whitney test and the Z-EV test.

The Results of t-test, Mann Whitney test and Z-EV test, for testing the expected non-significant differences between sample means by the varying α levels, for the sample size of 18, are shown in Fig.2. Again, the positive validity is acceptable at all the three α levels. The difference in the validity is not much between the t-test, Mann Whitney test and the Z-EV test.



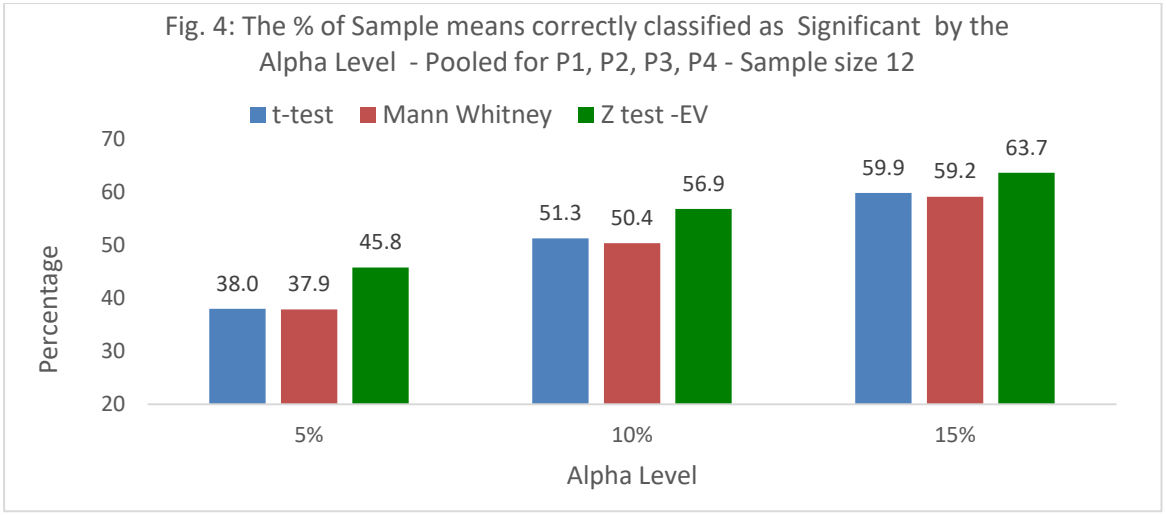
The Results of t-test and Z-EV test, for testing the expected non-significant differences between the sample means by the varying α levels, for the sample size of 24, are shown in Fig. 3. Here, also, the positive validity is quite high and acceptable at all the three α levels. Like before, the validity, is almost similar in all the three tests.



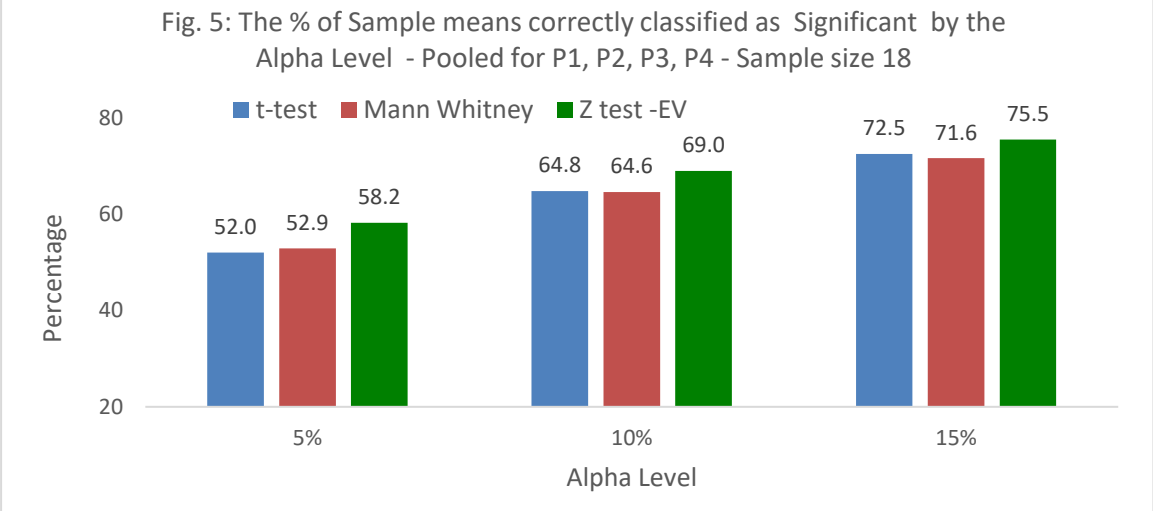
COMPARISON OF TWO SAMPLE MEANS WHEN SAMPLES ARE DRAWN FROM TWO DIFFERENT NORMAL POPULATIONS

The Results of t-test, Mann Whitney test and Z-EV test, for the expected significant differences between two sample means by the varying α levels, for the sample size of 12, are shown in Fig.4.

The negative validity, that is the ability to pick up correctly, the significant differences between two sample means is quite low. At 5% α level, the validity is around 38% for the t-test and the Mann Whitney test and 45.8% for the Z-EV test. However, it improves with the rise in α levels. For the t-test, for α level of 10% and 15%, the negative validity is found to be 51.3% and 59.9%, respectively. The validity of the Mann Whitney test is found to be comparable with the of t-test. In case of Z-EV test, as compared to t-test and Mann Whitney test, the negative validity is found to be relatively higher by approximately 5%.



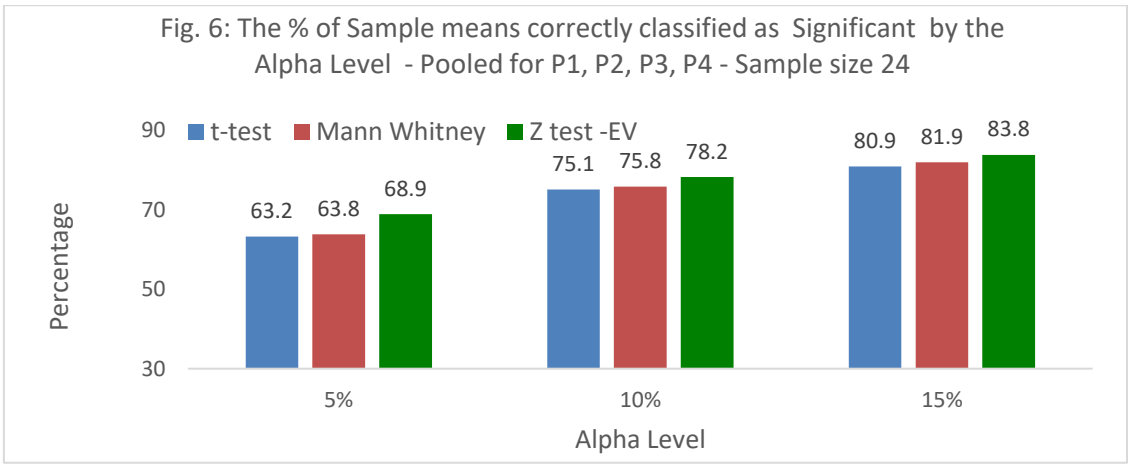
The Results of the t-test, Mann Whitney test and the Z-EV test, for the expected significant differences between two sample means by the varying α levels, for the sample size of 18, are shown in Fig.5.



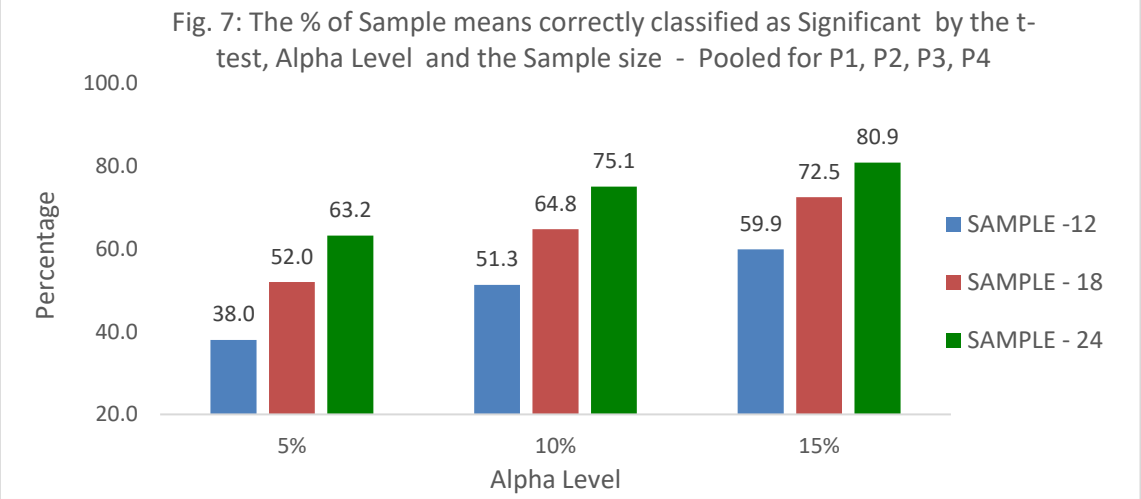
At α = 5%, the negative validity is observed to be around 52.0% for the t-test and the Mann Whitney test and 58.2% for the Z-EV test. At α=10%, the negative validity is around 64.0% for the t-test and the Mann Whitney test as against 69.0% seen in the case of Z-EV test. At α = 15%, the negative validity crosses the level of 70% for all the three tests. In general, the negative validity is observed to be higher for the Z-EV test as compared to other two tests.

The Results of the t-test, Mann Whitney test and the Z-test, for the expected significant differences between two sample means by the varying α levels, for the sample size of 24, are shown in Fig.6.

At α = 5%, the negative validity is around 63.0% for the t-test and the Mann Whitney test. It is 68.9% for the Z-EV test. At α = 15%, the negative validity crosses 80% for all the three tests considered.



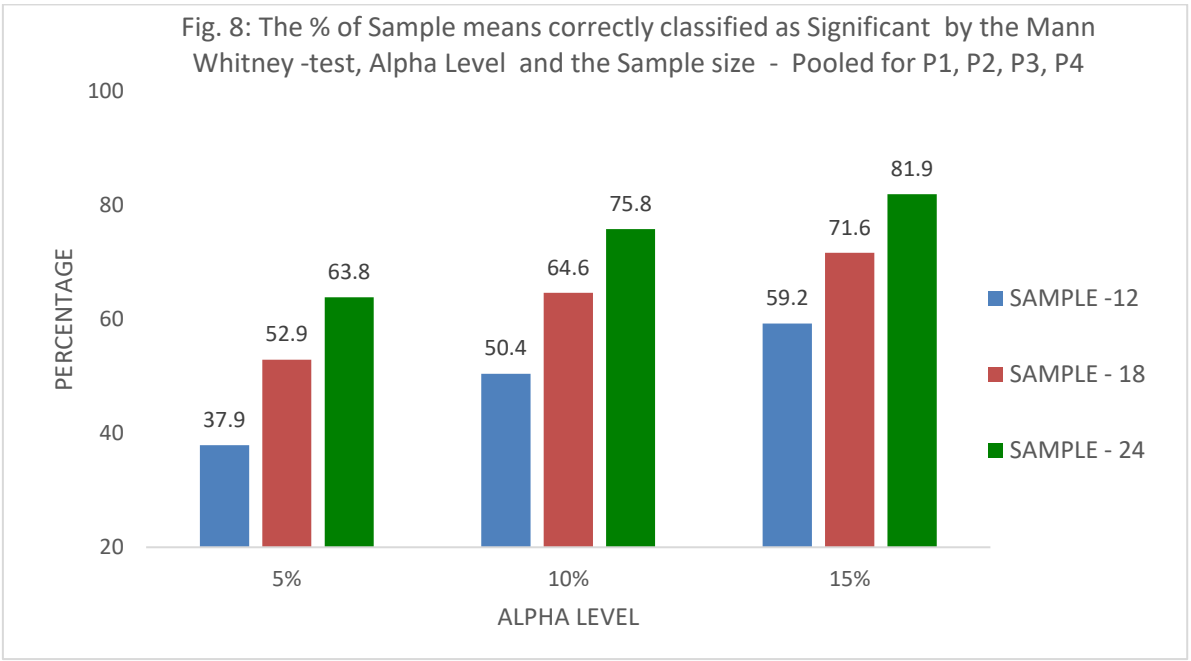
An attempt is also made in the present paper to find out the effect of sample size on the negative validity of the considered test. The variation in the % significance by the sample size and Alpha level for the t-test is shown in Fig. 7.



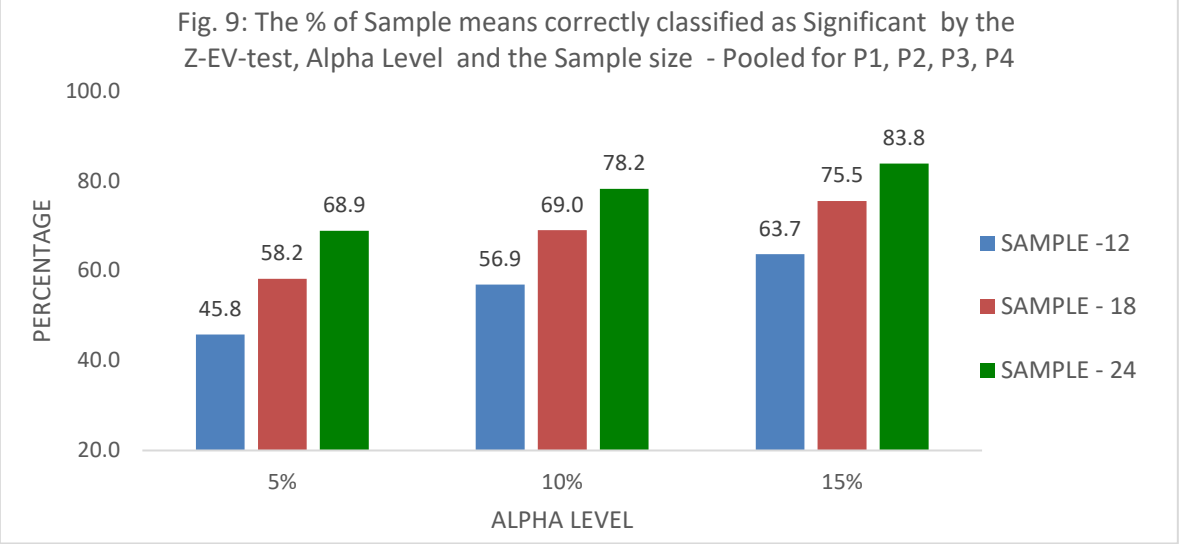
It is to be noted that for the t-test, the sample size of 12, at α = 5%, the validity is only 38% which is not acceptable. It changes to 52.0% and 63.2% for the sample size of 18 and 24, respectively. This raises the doubt of continuing the use of t-test even when the sample size is 24 and α is 5%. However, with the rise in α levels, the scenario changes favorably. At α = 10%, for the sample size of 18, the validity registered the level of 64.8% which raises to 75.1% for the sample size of 24. At α = 15%, for the sample size of 18, the validity is 72.5% which raises to 80.9% for the sample size of 24.

The variation in the % significance by the sample size and Alpha level for the Mann Whitney-test is shown in Fig. 8.

Surprisingly, the Mann Whitney test appears to be comparable in negative validity to the t-test at all the three α levels.



The variation in the significance by the sample size and Alpha level for the Z-EV-test is shown in Fig. 9.



%

At α = 5%, for the sample size of 12, for Z-EV test, the negative validity is only 45.8%. It changes to 58.2% and 68.9% for the sample size of 18 and 24, respectively which is not acceptable. At α =10%, for the sample size below 20, the negative validity remains below 70% while for the sample size of 24, the negative validity attains the level of 78.2%. At α = 15%, for the sample size of 18, the validity is 75.5% which raises to 83.8% for the sample size of 24.

DISCUSSION

The present study is in continuation to the study reported earlier, comparing, t-test, Mann Whitney test and Z EV test (Takiar 2023). In our previous study we dealt with the sample size of 9,6 and 3. In the present study, the sample sizes considered are 12,18 and 24. In previous study, it was reported that there was no problem in picking up the expected non-significant differences. In the present study also, no problem is found in picking up the expected non-significant differences correctly. Whether α is 5%, or 10%, the ability of all the three tests remained more than 90% suggesting that the three tests considered are associated with the higher positive validity. However, the situation does not remain the same when we wish to consider the negative validity.

The findings of study, strongly points out that when $\alpha = 5\%$, whether the sample size is 12, 18 or 24, the negative validity remains low and is observed to be below 65%. With $\alpha = 5\%$, it is expected that if we carry out 100 mean comparisons, when samples are known to have been drawn from two different normal populations, only 5% comparisons should be rejected by chance but based on the findings of the current study, the t-test is rejecting around 38%. This amounts to failure of the theory and belief that the t-test should be employed for small samples. It can be concluded therefore that the t-test cannot be accepted as a suitable test for picking up the expected significant differences between the sample means.

When we choose α = 10%, the negative validity for the t-test remains below 65% for the sample size of 12 and 18. However, it changes to 75.1%, when the sample size is 24. While the negative validity improves with α = 15%, many researchers may not like to go for such a higher level of α. The best possible solution is therefore to use t-test when the sample size is 20 or above and α is essentially equals to 10% or 15%. By choosing α = 15%, there is a gain in negative validity and which will increase to more than 75% which may be quite acceptable to many. Which level of negative validity should be acceptable, can be a point of discussion? If one wants to stick to the α level of 5%, then the t-test

and Mann Whitney test are found to be associated with very low negative validity and thus not suitable to carry out the mean differences and can be dropped from the future comparisons of sample means when n is below 30.

In the case of Mann Whitney test, its performance is observed to be comparable with that of t-test. It may make very little difference when you choose t-test or Mann Whitney test for comparing the significant differences between two sample means. Again, the best possible scenario for these tests is also that the sample size should be 20 or more and α is essentially equals to 10% or 15%.

In the case of Z-EV test, in comparisons to other two tests, the negative validity is found to be higher at all the three selected α levels. In view of Z-EV test scoring better in negative validity as compared to t-test, Z-EV test can be opted instead of t-test even for testing the significance differences between 2 sample means when n is 20 or above and α = 10% or 15%.

CONCLUSIONS

- At α = 5%, for the all the three selected sample sizes namely 12, 18 and 24, the negative validity remained below 65% for the t-test and the Mann Whitney test.
- Such a low negative validity observed in the case of t-test and Mann Whitney test raises a doubt about their continuing use when the sample size is below 30.
- Even at α = 10%, for the sample size of 12 and 18, the negative validity for the t-test and Mann Whitney test remained below 70%.
- It is concluded therefore that the t-test and Mann Whitney test are not suitable when the sample size is below 20.
- Relatively, at all the sample sizes selected, the Z-EV test performed better as compared to the t-test and the Mann Whitney test.
- Based on the results obtained in the present study, it is concluded that the use of Z-EV test is more appropriate as compared to the t-test and the Mann Whitney test even when the sample size is above 18 and below 30.

RECOMMENDATIONS

- Assuming the negative validity of 70% or above is acceptable, the minimum sample size should be above 18 and α to be 10% or 15%.
- For small samples, between 18 and 30, Z-EV test can be used instead of t-test or Mann Whitney test.
- In case of very small samples (below 15), to get more meaningful and valid results, it is advised to choose essentially the α level to be 15%.

REFERENCES

[1]. Gupta SC 2012: Fundamental of Statistics, Seventh Edition, Himalaya Publishing House; Page 19.1-19.2, 19.12-19.14.

[2]. Gupta SC, Kapoor VK 2001: Fundamentals of Mathematical Statistics, Sultan Chand & Sons; Tenth revised Edition, Page 14.1-14.3 and 14.24-14.26.

[3]. IBM Corp. (2015). IBM SPSS Statistics for Windows -Version 23.0, IBM Corp.

[4]. Microsoft Corporation, 2019. Microsoft Excel, Available at: https://office.microsoft.com/excel.

[5]. Snedecor GW, Cochran WG 1967; Statistical Methods, Sixth Edition; Oxford & IBH Publishing Co., Page 59-60, 100-105.

[6]. StatPlus 7.6.5.0 2021, AnalystSoft Inc.-Statistical analysis program. Version v7. See https://www.analystsoft.com/en/

[7]. Takiar R (2023): The Validity of t-test, Mann-Whitney test and Z test for Testing Significant differences between two Sample Means When Sample size is 10 or below, Bulletin of Mathematics and Statistics Research, Vol. 11(2), Page 1-15.

Biography of corresponding author: Dr. Ramnath Takiar

I am a Post graduate in Statistics from Osmania University, Hyderabad. I did my Ph.D. from Jai Narain Vyas University of Jodhpur, Jodhpur, while in service, as an external candidate. I worked as a research scientist (Statistician) for Indian Council of Medical Research from 1978 to 2013 and retired from the service as Scientist G (Director Grade Scientist). I am quite experienced in large scale data handling, data analysis and report writing. I have 65 research publications in national and International Journals related to various fields like Nutrition, Occupational Health, Fertility and Cancer epidemiology. During the tenure of my service, I attended three International conferences namely in Goiana (Brazil-2006), Sydney (Australia-2008) and Yokohoma (Japan-2010) and presented a paper in each. I also attended the Summer School related to Cancer Epidemiology (Modul I and Module II) conducted by International Agency for Research in Cancer (IARC), Lyon, France from 19th to 30th June 2007. After my retirement, I joined my son at Ulaanbaatar, Mongolia. I worked in Ulaanbaatar as a Professor and Consultant from 2013-2018 and was responsible for teaching and guiding Ph.D. students. I also taugth Mathematics to undergraduates and Econometrics to MBA students. During my service there, I also acted as the Executive Editor for the in-house Journal “International Journal of Management”. I am still active in research and have published 7 research papers during 2021-23.