Received: April 25, 2023 | Revised: July 25, 2023 | Accepted: July 31, 2023

Mixed Bootstrap-Marascuilo test for testing equality of means under unequal variances in a completely randomized design

Phannipa Worapun[1] and Tammarat Kleebmek[2*]

[1]Department of Applied Mathematics and Statistics, Faculty of Sciences and Liberal Arts, Rajamangala University of Technology Isan, Nakhon Ratchasima 30000, THAILAND

[2]Department of Statistics, Faculty of Science, Khon Kaen University, Khon Kaen 40002, THAILAND

[*]Corresponding author: tammarat@kku.ac.th

ABSTRACT

This study introduces a pioneering statistical methodology, the Mixed Bootstrap-Marascuilo test, which emerges from the fusion of the Bootstrap method and the Marascuilo test. The test examines the means of three or more populations with unequal variances in a completely randomized design. The research compares the Mixed Bootstrap-Marascuilo test's ability to control the probability of type I error using Cochran's criteria and its test power using penalized power with those of One-Way ANOVA and the Marascuilo test. The research design thoughtfully encompasses three populations: small, medium, and large, each exhibiting variations in size and equality. It systematically manipulates variances, encompassing a range from equal to slightly different, moderately different, and significantly different. Furthermore, the error distribution is specified to be normal with a designated mean. The research methodology embraces the utilization of R ver.4.2.2 and the Monte Carlo technique, leveraging 5,000 simulations per case to ensure the robustness and reliability of the study's findings. The outcomes of the comprehensive analysis yield intriguing insights. The results indicate that the Mixed Bootstrap-Marascuilo test exhibits superior testing ability in scenarios involving equal small sample sizes with equal and moderate variances, as well as medium sample sizes with equal and small variances. Furthermore, it demonstrates effectiveness even when dealing with large, unequal sample sizes with equal and large variances. In essence, this research advances the realm of statistical hypothesis testing through the introduction and meticulous evaluation of the Mixed Bootstrap-Marascuilo test. Its demonstrated ability to navigate complexities in mean examination across diverse populations, coupled with its versatile applicability to scenarios of unequal variances and varying sample sizes, underscores its potential as a valuable tool for researchers across disciplines.

Keywords: Unequal variances, ANOVA, Bootstrap, Marascuilo test, Mixed Bootstrap-Marascuilo test

INTRODUCTION

Statistical analysis is essential to most research studies and is crucial to data analysis. In experimental research, where experimental design is a critical step, statistical methods are necessary to ensure reliable results. Researchers need to have statistical knowledge to analyze data and select appropriate statistical methods based on the characteristics of the data, in addition to subject matter expertise and the ability to design experiments. The Completely Randomized Design (CRD) is a data analysis technique used for a single factor with several levels, requiring the experimental units to be similar. The One-Way Analysis of Variance (One-Way ANOVA) is the analysis method employed. However, in practice, the data may not adhere to the initial assumption that the error follows a normal distribution and homogeneity of variance, particularly when the variances of each population are different. Using the ANOVA technique under such conditions may result in inaccurate conclusions. To address this, statisticians have developed and refined tests to evaluate the difference in means among multiple population groups when analyzing data with unequal variances. Examples of such tests include Welch's Test, Marascuilo Test, Brown-Forsythe's Test, Modified Brown-Forsythe's Test, Generalized F Test, and others. Komduan [1] conducted a comparative study of the Welch test, Brown-Forsythe test, Marascuilo test, and Parametric Bootstrap approach. The study revealed that the Welch test and Marascuilo test effectively controlled the probability of Type I error

DOI: 10.60101/jarst.2023.251663

and had the highest testing power when the error followed a normal distribution and the variance of the error differed. Hananurak [2] also found that the Marascuilo test could control Type I error probability when the population followed a t-distribution and had the highest testing capacity when the population was normally distributed. Additionally, Efron and Tibshirani [3] and Sangthong [4] proposed a parameter estimation method that includes estimating the variance of various estimators, known as the Bootstrap technique. Reiczigel et al. [5] compared the Bootstrap Rank-Welch (BRW) test with the Wilcoxon-Mann-Whitney (WMW) test, the Rank Welch (RW) test, and the Brunner-Mansel (BM) test. They found that the BM test could control the probability of error well in a large sample, and the Bootstrap Rank-Welch test could control the probability of error as well as the Wilcoxon-Man-Whitney test and the Rank-Welch test in a small sample. Dag et.al. [6] compared seven one-way tests in the oneway tests package. The Kruskal-Wallis test performs best with small and medium effect sizes. Cavus and Yazaci [7] present the doex package, which contains the tests for equality of normal distributed and independent group means under unequal variances. Although new test statistics are continually being proposed, it remains unclear which test statistics are most effective under the same circumstances. Thus, this study aims to compare the statistical efficiency of three tests: One-Way ANOVA, Marascuilo procedure, and a novel approach that combines the Marascuilo test procedure with the Bootstrap technique called the Mixed Bootstrap-Marascuilo (MBM) test. The study examines scenarios involving normal distributions, varying variances, and equal or unequal sample sizes across three populations. The primary criterion for assessing performance is controlling Type I error rates and achieving the highest testing power.

MATERIALS AND METHODS

The Completely Randomized Design (CRD) is a trial scheme randomly assigning treatments to trial units without specific conditions or restrictions. This design ensures that each trial unit has an equal chance of receiving any treatments, making it most like the trial units themselves. The primary goal of CRD is to compare the means of k populations.

$$y_{ij}=\mu+\tau_i+\varepsilon_{ij} \quad (1)$$

$i = 1,2,\ldots..,k, j = 1,2,\ldots..,n_i$

where
n_i is the number of the i[th] treatment,
y_{ij} is the j[th] observation of the i[th] treatment,
μ is the population mean,
τ_i is the treatment effect of the i[th] treatment,
and ε_{ij} is the random error.

The null hypothesis is simply that all the group population means are the same. The alternative hypothesis is that at least one pair of means is different. For example, if there are k groups:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$$

H_1: At least one of the group's means is different.

1. The method of test statistics

This study is experimental research that aims to compare the efficiency of test statistics for comparing population means across three or more groups, specifically the One-Way ANOVA, Marascuilo Test, and the Mixed Bootstrap-Marascuilo Test obtained by applying the bootstrap technique to the Marascuilo test statistics. The RStudio program was used to simulate data and test hypotheses. The research methodology is described as follows:

1.1 One-Way ANOVA

The One-Way ANOVA is applied to assess whether there are statistically significant distinctions among the means of three or more independent (unrelated) groups. There are two primary assumptions: the responses for each factor level have a normal population distribution. These distributions have the same variance. Examining the formula for One-Way ANOVA as shown in Table 1.

Table 1 One-way ANOVA Table.

SV	***DF***	***SS***	***MS***	F_{cal}
Treatment	$k-1$	$SSTr$	$MSTr$	$\frac{MSTr}{MSE}$
Error	$\sum n_i - k$	SSE	MSE	
Total	$\sum n_i - 1$	SST		

where SV : Source of Variation,
DF : Degree of Freedom,
SS : Sum of Square,
MS : Mean Square.

We have

$$F_{cal} = \frac{MSTr}{MSE}$$

Where $MSE = \frac{SSE}{\sum n_i - k}, MSTr = \frac{SSTr}{k-1},$

$SSTr$: treatment sum of square

$$SSTr = \sum_{i=1}^{k} \frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{\sum n_i - k},$$

SSE : Error sum of square

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^2 - \sum_{i=1}^{k} \frac{y_{i.}^2}{n_i}$$

DOI: 10.60101/jarst.2023.251663

JOURNAL OF APPLIED RESEARCH ON SCIENCE AND TECHNOLOGY

and

$$SSE = SST - SSTr,$$

to reject H_0 when $F_{cal} > F_{\alpha,k-1,\sum n_i - k}$.

1.2 Marascuilo test

Marascuilo (M) test [8] is a modified version of the Welch test, simplifying the calculation process. It is a method for testing the equality of means in more than two populations in cases where the variances are unequal. This test is suitable for large sample sizes and for groups with different population variances. Let's delve into the formula for the Marascuilo test:

$$M = \sum_{i=1}^{k} w_i[\bar{y}_i - \tilde{y})^2/(k-1)]$$

Where $w_i = \frac{n_i}{S_i^2}, u = \sum_{i=1}^{k} w_i, \tilde{x} = \sum_{i=1}^{k} \frac{w_i \bar{y}_i}{u}$.

$\bar{y}_i$ is mean of the i[th] treatment,
s_i^2 is variance of the i[th] treatment
To reject H_0 when $M > F_{\alpha,k-1,f}$

$$f = \left[\frac{3}{(k^2-1)}\left(\sum_{i=1}^{k}\frac{\left(1-\frac{w_i}{u}\right)^2}{n_i - 1}\right)\right]^{-1}$$

At the level of significant α

1.3 The Mixed Bootstrap-Marascuilo test (MBM)

The test is obtained by applying the bootstrap technique to the Marascuilo test statistics, which are described as follows:

1.3.1 Let $Y_i = y_{i1}, y_{i2}, \ldots, y_{in_i}$ is an observation from sample i, sample size n_i and have a normal distribution.

1.3.2 A random sampling with replacement by using the technics Bootstrap, Define $Y_i^* = y_{i1}^*, y_{i2}^*, \ldots, y_{in_i}^*$ from n_i^* which is sample size equal n_i.

1.3.3 To compute the test of Marascuilo (M^*) from the sample Y_i^*, n_i^* as follows:

$$M^* = \sum_{i=1}^{k} w_i^*[(\bar{y}_i^* - \tilde{y}^*)^2/(k^* - 1)]$$

where

$$w_i^* = \frac{n_i^*}{S_i^{2*}}, u^* = \sum_{i=1}^{k} w_i^*, \tilde{y}^* = \sum_{i=1}^{k} \frac{w_i^* \bar{y}_i^*}{u^*}.$$

1.3.4 Repeat step 3 until 5,000 rounds have been completed.

1.3.5 Take the test statistics obtained from the 5,000 rounds of bootstrap sampling to find the mean, and compare the mean of the test statistics obtained with the critical.

2. The simulation and compare the test statistics

The study was conducted using the Monte Carlo data simulation method. The R programming language (version 4.2.2) was used to simulate data and test hypotheses. The operating procedures were designed to determine α_i is the probability of a type I error of the test (to reject the null hypothesis when it is true), in comparison to the Bradley criteria [9], and used penalized power introduced by Cavus et.al. [10] to compare the power of the test as follows:

$$\gamma = \frac{1-\beta}{\sqrt{1+\left|1-\frac{\alpha_i}{\alpha_0}\right|}}$$

where β is type II error and α_0 is the nominal level. The procedures are as follows:

1) Set the criteria for determining the effectiveness of controlling the probability of Type I errors from Bradley [9]. If the probability of a Type I error resulting from the experiment falls within the range specified for each level of significance, it means that the test statistic can control the Type I error probability. It can be classified according to the significance level, as shown in Table 2.

Table 2 The level of significant of Type 1 error.

The level of significant	Confidence Interval
0.25	[0.200, 0.300]
0.20	[0.160, 0.240]
0.15	[0.120, 0.180]
0.10	[0.080, 0.120]
0.05	[0.040, 0.060]
0.01	[0.007, 0.015]

2) Set the number of treatments and replicates of experimental units in each treatment to be the same for every treatment. There were two cases, as shown in Table 3.

Table 3 Sample size determination for various scenarios.

(n_1, n_2, n_3)	small	medium	large
equal	(10,10,10)	(30, 30, 30)	(100,100,100)
unequal	(5,10,15)	(25, 30, 35)	(95,100,105)

3) The treatment effect on the constant mean $\mu = 50$ in two cases by Hasalems and Jitthavech [11].

a) When all values have the same mean.
b) When the means are not equal.

DOI: 10.60101/jarst.2023.251663

4) Set the distribution of errors (ε_i) to be normal with a specified mean and equal variance.

a) When the variance is the same for all populations, the variance ratio is 1:1:1.

b) When the variance is different for each population, three cases were studied.

Table 4 Ratio of variance classification of size.

Variance	Ratio
Small	1.5:2.4:3
Medium	1.5:6:9
Large	1.5:18:36

5) Randomly sample from the simulated data for each treatment with an equal number of replicates.

6) Test for equality of variances of the variable at a specified significance level of 0.05.

7) To compare the mean of 3 treatments from the hypothesis.

8) Repeat the above steps 5,000 times for each situation.

9) Compare the calculated test statistic value with the critical value using a specified significance level. Repeat this process 5,000 times for each situation to calculate the probability of Type I error and the power of the test.

10) Consider α_i in confidence interval as the Bradley's criteria when the nominal level is 0.05; if the calculated α_i falls within the criteria, it can be concluded that the test is able to control the probability of Type I error, and the penalized power of the three testing statistics can be calculated in the case where the test is able to control the probability of Type 1 error.

RESULTS AND DISCUSSIONS

This research is an experimental study to investigate and compare the efficiency of three test statistics for testing the mean of three populations, including the analysis of variance, (one-way ANOVA), Marascuilo Test, and the Mixed Bootstrap-Marascuilo Test. The results of the research can be summarized into two parts: 1. The ability to control the Type I error rate of the first Type of error using the Cochran criterion, and 2. Comparing the power of the tests. The symbols representing the test statistics are defined as follows:

F : One-Way ANOVA,
M : Marascuilo Test,
MBM : Mixed Bootstrap-Marascuilo Test.

The results of comparing the probability of Type I error rates are shown in Table 5 and Table 6. The results of comparing the power of the test are shown in Table 7 and Table 8.

Table 5 Probability of Type I error rate in the case of an equal sample size.

n_1, n_2, n_3	$\sigma_1^2, \sigma_2^2, \sigma_3^2$	Test		
		F	M	MBM
10,10,10	1:1:1	0.0516	0.0522	0.0502
	1.5:2.4:3	0.0526	0.0540	0.0554
	1.5:6:9	0.0584	0.0576	0.0560
	1.5:18:36	0.0644	0.0534	0.0494
30,30,30	1:1:1	0.0464	0.0492	0.0514
	1.5:2.4:3	0.0496	0.0530	0.0500
	1.5:6:9	0.0524	0.0500	0.0512
	1.5:18:36	0.0584	0.0510	0.0486
100,100,100	1:1:1	0.0510	0.0500	0.0464
	1.5:2.4:3	0.0506	0.0514	0.0410
	1.5:6:9	0.0526	0.0490	0.0492
	1.5:18:36	0.0582	0.0496	0.0536

Note: The underline indicates that the Type I error rate cannot be controlled by the Cochran criterion.

Table 6 Probability of Type I error rate in the case of unequal sample size.

n_1, n_2, n_3	$\sigma_1^2, \sigma_2^2, \sigma_3^2$	Test		
		F	M	MB_M
5,10,15	1:1:1	0.0482	0.0574	0.0612
	1.5:2.4:3	0.0618	0.0566	0.0694
	1.5:6:9	0.0650	0.0544	0.0612
	1.5:18:36	0.0628	0.0558	0.0640
25,30,35	1:1:1	0.0500	0.0498	0.0484
	1.5:2.4:3	0.0450	0.0516	0.0500
	1.5:6:9	0.0454	0.0544	0.0496
	1.5:18:36	0.0452	0.0568	0.0532
95,100,105	1:1:1	0.0488	0.0474	0.0492
	1.5:2.4:3	0.0482	0.0486	0.0518
	1.5:6:9	0.0530	0.0526	0.0584
	1.5:18:36	0.0564	0.0520	0.0458

Note: The underline indicates that the Type I error rate cannot be controlled by the Cochran criterion.

Table 7 Power of the test in equal sample size.

n_1, n_2, n_3	$\sigma_1^2, \sigma_2^2, \sigma_3^2$	Test		
		F	M	MBM
10,10,10	1:1:1	0.13978	0.1391	0.1614
	1.5:2.4:3	0.09048	0.0904	0.0904
	1.5:6:9	0.06588	0.0695	0.0699
	1.5:18:36	-	0.0561	0.0512
30,30,30	1:1:1	0.3738	0.3849	0.4008
	1.5:2.4:3	0.1795	0.1882	0.1962
	1.5:6:9	0.0973	0.1108	0.1101
	1.5:18:36	0.0685	0.0743	0.0608
100,100, 100	1:1:1	0.8816	0.8876	0.8565
	1.5:2.4:3	0.5330	0.5503	0.4840
	1.5:6:9	0.2418	0.3054	0.2591
	1.5:18:36	0.0999	0.1219	0.1003

Note: The underline indicates the maximum power of the test.

DOI: 10.60101/jarst.2023.251663

JOURNAL OF APPLIED RESEARCH ON SCIENCE AND TECHNOLOGY

Table 8 Power of the test in the case of unequal sample sizes.

n_1, n_2, n_3	$\sigma_1^2, \sigma_2^2, \sigma_3^2$	Test		
		F	M	MBM
5,10,15	1:1:1	0.1116	0.1167	-
	1.5:2.4:3	-	0.0876	-
	1.5:6:9	-	0.0669	-
	1.5:18:36	-	0.0530	-
25,30,35	1:1:1	0.3728	0.3705	0.3737
	1.5:2.4:3	0.1547	0.1918	0.1918
	1.5:6:9	0.0852	0.1239	0.1010
	1.5:18:36	0.0537	0.0696	0.0630
95,100,105	1:1:1	0.8799	0.8675	0.8905
	1.5:2.4:3	0.5313	0.5626	0.5170
	1.5:6:9	0.2211	0.2952	0.2376
	1.5:18:36	0.0943	0.1257	0.0987

Note: The underline indicates the maximum Power of the test.

The analysis of variance for a single-factor design can control the Type I error rate in all cases when the sample sizes are the same. This finding is consistent with the research conducted by Hasalems and Jitthavech [11].

The Marascuilo Test can control the Type I error rate and has the highest power when the population variances differ greatly. This is in line with the research results reported by Komduan [1].

Table 9 Test statistics that can control the Type I error rate probability.

n_1, n_2, n_3	$\sigma_1^2, \sigma_2^2, \sigma_3^2$			
	1:1:1	1.5:2.4:3	1.5:6:9	1.5:18:36
5,10,15	F, M	M	M	M
10,10,10	F, M, MBM	F, M, MBM	F, M, MBM	M, MBM
25,30,35	F, M, MBM	F, M, MBM	F, M, MBM	F, M, MBM
30,30,30	F, M, MBM	F, M, MBM	F, M, MBM	F, M, MBM
95,100,105	F, M, MBM	F, M, MBM	F, M, MBM	F, M, MBM
100,100,100	F, M, MBM	F, M, MBM	F, M, MBM	F, M, MBM

Table 10 The comparison of the power of the tests.

n_1, n_2, n_3	$\sigma_1^2, \sigma_2^2, \sigma_3^2$			
	1:1:1	1.5:2.4:3	1.5:6:9	1.5:18:36
5,10,15	M	M	M	M
10,10,10	MBM	F	MBM	M
25,30,35	MBM	MBM	M	M
30,30,30	MBM	MBM	M	M
95,100,105	MBM	M	M	MBM
100,100,100	M	M	M	M

CONCLUSION

The analysis of variance with one-way can control the Type I error rate probability of Type I errors in almost all cases except when the sample size is the same and small (10, 10, 10) with significantly different variances (1.5:18:36) and when the sample size is different and small (5, 10, 15) with significantly different variances in all cases. The Marascuilo Test can control the Type I error rate probability of Type I errors in all cases, and the Mixed Bootstrap-Marascuilo Test can control the Type I error rate probability of Type I errors in almost all cases, except for the case of a small sample size (5, 10, 15) with different variances for all variances, as shown in Table 9. According to the results, the Mixed Bootstrap-Marascuilo test has the greatest testing ability in situations where there are equal small sample sizes with equal and moderate variances, as well as medium sample sizes with equal and small variances. Additionally, it shows an unequally large sample size with equal and large variances, as shown in Table10. This research can be used by organizations or researchers to select appropriate test statistics that are suitable for the characteristics of the data, both in cases where variances are equal and unequal and sample sizes are equal and unequal. In future research, the efficiency of test statistics for testing the equality of other means should be studied. This study used sample size, variance, and the influence of treatment factors that affect the ability to control Type I error and test power in only some cases, not all types of data. Therefore, in the future, the characteristics of data in different scenarios should be studied in more detail.

REFERENCES

1. Komduan J. Comparing the efficiency of mean testing statistics for three populations. [Master's thesis]. Bangkok: Faculty of Science, King Mongkut's Institute of Technology Ladkrabang; 2017.
2. Hananurak P. Comparing the Wilcoxon rank-sum test and the Mann-Whitney U test for testing population means with unequal variances. [Master's thesis]. Bangkok: Kasetsart University; 2006.
3. Efron B, Tibshirani RJ. An introduction to the bootstrap. 1st ed. New York: Chapman and Hall/ CRC;1993.
4. Sangthong M. Nonparametric Bootstrap Method for Location Testing between Two Populations under Combined Assumption Violations. Burapha Science Journal. 2020;25(3):864-79.
5. Reiczigel J, Zakarias I, Rozsa L. A Bootstrap test of stochastic equality of two populations. The American Statistician. 2005;59:156-61.

DOI: 10.60101/jarst.2023.251663

6. Dag O, Dolgun A, Konar NM. Onewaytests: An R Package for One-Way Tests in Independent Groups Designs. R Journal. 2018;10(1):175-99.
7. Cavus M, Yazici B. Testing the equality of normal distributed and independent groups' means under unequal variances by doex package. The R Journal. 2021;12(2):134-54.
8. Marascuilo LA. Statistical for behavioral science research. New York: McGraw-Hill Book Co; 1971.
9. Bradley JV. Robustness. Br J Math Stat Psychol. 1978;31(2):144-52.
10. Cavus M, Yazici B, Sezer A. Penalized power approach to compare the power of the tests when Type I error probabilities are different. Commun Stat - Simul Comput. 2021;50(7):1912-26.
11. Hasalems P, Jitthavech J. Comparing mean testing under variance heterogeneity in complete randomized experiments. Burapha Science Journal. 2018;23(1):135-45.

DOI: 10.60101/jarst.2023.251663