Robustness of a Two-Stage Procedure When Comparing Population Means

By: Christino Lorenzo Barbosa [1] **| Mentor:** Dr. Rebecca Conley [2]
[1]STEM Division, Union County College - Cranford, NJ
[2]Saint Peter's University - Jersey City, NJ

Union County College
Transforming Our Community... One Student at a Time

Abstract

In this research I conducted a series of experiments designed to determine the accuracy of a two-stage procedure for comparing population means. To set up each experiment three groups of sample data were generated using random numbers from the normal, uniform, or mixed-normal populations with equal means. We applied the two-stage procedure, determined whether a type 1 error had been committed, and recorded the results. I found the two-stage procedure was robust for the samples I used.

Introduction

Through the course of this research, I tested the accuracy of a two-stage procedure when using hypotheses tests to compare the means of three populations. The analysis of variance (ANOVA) and the Welch test are used to compare the means of groups. This is referred to as preliminary testing or a two-stage procedure. We wish to determine whether the two-stage procedure changes the overall probability of a type 1 error.

A type 1 error is when the null hypothesis true but is rejected by the hypothesis test. When using a significance level, $\alpha = 0.05$, there is a 5% chance this will happen. We use β to represent the probability of a type 2 error, which is when the null hypothesis is false, but it was not rejected. Type 1 error and type 2 error can be used to tell us how accurate these methods are. A method is considered robust if the probability of a type one error is between 0.5α and 1.5α, where α is the significance level (Bradley 1978). In the case of this research, I am using $\alpha = 0.05$, so for a test to be robust, the probability of a type 1 error must be between 0.025 and 0.075.

Methods

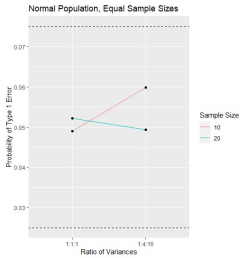
In practice, researchers often test their samples to determine if they are from populations with equal variances before deciding whether to use an ANOVA or a Welch test. I wrote a program in RStudio v1.4.1106 to use simulated data to calculate the probability of a type 1 error for the whole procedure. I looked at different population shapes (normal, mixed normal, and uniform), different ratios of variances, and different sample sizes. I conducted 16 experiments.

I created a series of tests with samples containing randomized values for the data. In each test we have three samples. I considered the sample sizes 10, 20, 30 and 40 and the variances of 1, 4, 16, and 64. I also considered samples drawn from the normal population, the uniform population, and the mixed-normal population. All the combinations are listed in the Results section. After creating the groups, the Levene test is used to determine whether to use Welch or ANOVA. If the p-value from the Levene test is less than or equal to 0.05 then the Welch test is used, otherwise I use ANOVA. The null hypothesis for the Levene test is all variances are equal and the alternative hypothesis is at least one of the pairs of variances is not equal (NIST/SEMATECH, 2021). The null hypothesis for ANOVA and Welch is that all means are equal, and the alternative hypothesis is at least one of the pairs of means is not equal (NIST/SEMATECH, 2021). I ran the procedure 10,000 and summarized the results by the number of times the conclusion is correct or incorrect (type I error).

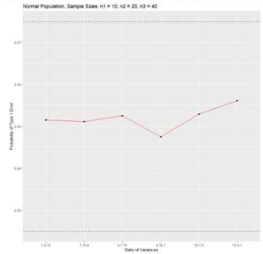
Results

The dashed black lines on the graphs are at 0.025 and 0.075, which are the bounds of where a test is considered robust according to the Bradley criteria (Bradley, 1978).

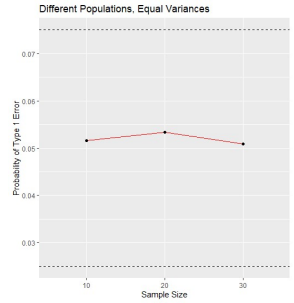
Experiment Number	Populations	Sample Sizes	Variance Ratio	P(Type I Error)
1	Normal	10, 10, 10	1:1:1	0.0490
2	Normal	20, 20, 20	1:1:1	0.0522
3	Normal	10, 10, 10	1:4:16	0.0598
4	Normal	20, 20, 20	1:4:16	0.0494



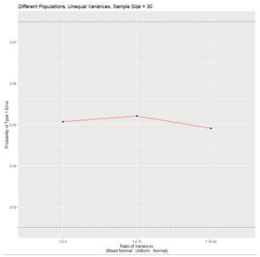
Experiment Number	Populations	Sample Sizes	Variance Ratio	P(Type I Error)
5	Normal	10, 20, 40	1:4:16	0.0515
6	Normal	10, 20, 40	1:16:4	0.0511
7	Normal	10, 20, 40	4:1:16	0.0525
8	Normal	10, 20, 40	4:16:1	0.0475
9	Normal	10, 20, 40	16:1:4	0.0529
10	Normal	10, 20, 40	16:4:1	0.0561



Experiment Number	Populations	Sample Sizes	Variance Ratio	P(Type I Error)
11	Different	10, 20, 30	1:1:1	0.0516
12	Different	10, 20, 30	1:1:1	0.0534
13	Different	10, 20, 30	1:1:1	0.0509



Experiment Number	Populations	Sample Sizes	Variance Ratio	P(Type I Error)
14	Different	30, 30, 30	1:2:4	0.0507
15	Different	30, 30, 30	1:4:16	0.0521
16	Different	30, 30, 30	1:16:64	0.0491



Conclusion

According to the results, we can conclude that the probability of committing a type 1 error is around 5% for about all the experiments. As shown through the p-value, all the values are within 0.04 to 0.06. This is expected because the significance level α was set equal to 0.05 at the beginning of the experiment. We can also conclude that all the tests are robust because the p-value falls within 0.025 and 0.075. Through this research I learned a process for testing methods of statistics. This research has helped me to think of ways to effectively communicate the process and results to people who may not be familiar with the vocabulary. I did this through clearly describing terms and making graphs and tables of the results that are easy to read.

Future Work

An infinite number of different scenarios could be tested, I merely covered a small fraction of them. Future work would include running tests with larger sample sizes, a different number of groups, different populations, and other combinations of variances. For even more accurate results I could repeat each test more than 10,000 times. I could consider preliminary tests other than Levene's test and see how it changes the results. I could use one step testing instead of the two-step used in this research. There are many possibilities in what this research could lead to in the future.

Acknowledgements

Grant Funding from the NSF IRAP Grant
Union County College STEM Division

Contact Information

Christino.Barbosa@owl.ucc.edi