8-9-2019

Effect of Heterogeneity of Variance on the performance of ANOVA F-test and its Alternatives: Simulation Based Study

Jay Lopez
Portland State University

Effect of Heterogeneity of Variance on the performance of ANOVA F-test and its Alternatives: Simulation Based Study

By

Jay Lopez

An undergraduate honors thesis submitted in partial fulfillment
of the requirements for the degree of
Bachelor of Science
in
University Honors

Thesis Advisor Nadeeshani Jayasena, Ph. D.

Fariborz Maseeh Department of Mathematics and Statistics
Portland State University
2019

Abstract

Studies have shown that ANOVA F-test has a lower performance against heterogeneity of variances. It is important to provide more information on its alternatives and other methods that can prove useful. As a general guideline, Welch's ANOVA is a best alternative with low type 1 error rate in all cases of different population variances compared to other methods used in this study. In addition to Welch's ANOVA, Marascuilo's alternative to this test gives a less accurate result but provides simpler calculation methods. Similar to Moder K. (2010) , Kruskal Wallis test and Hotelling's T^2 test were taken into consideration. Kruskal Wallis test had higher type 1 error rate similar to the ANOVA F-test. Hotelling's T^2 test had significantly lower type 1 error rate in comparison to the ANOVA F-test and the Kruskal Wallis test. Depending on the amount of observations different studies may have, a multivariate analysis of variance using Hotelling's T^2 test is advisable. Otherwise Welch's ANOVA is a better choice for a test with lower type 1 error rate.

Introduction

This research analyzes the analysis of variance (ANOVA) F-test and its alternatives under heterogeneity of variances. Similar to previous studies, a simulation will be used in order to collect data. The goal is to clearly define each test and provide additional information based on the results. Research from other studies may contain population variances that differ in groups therefore, it is important to distinguish various tests that can be used under different parameters. It is important to continue this research in order to present beneficial information for conducting more accurate results.

The ANOVA F-test is the generic form of testing the equality of means between more than two different populations. The null hypothesis is that there is no difference in the group means and means should be equal. The alternative hypothesis is that at least one group mean is different. When using this test, it is assumed that the data from the populations are each normally distributed with equal variance and are independent from each other. F test statistic is computed using the following function,

$$F = \frac{MST}{MSE} = \frac{\sum_{j=1}^{k} n_j \left(x_{ij} - \bar{x}\right)^2 / (k-1)}{\sum_{i=1}^{n} \sum_{j=1}^{k} \left(x_{ij} - \bar{x}_j\right)^2 / (k(n-1))}$$

where n_j and k are the sample size of each group and number of groups, respectively while $\bar{x}$ and $\bar{x}_j$ are observed overall mean and group mean, respectively. P-value is computed using the F distribution with $(k-1)$ and $k(n-1)$ degrees of freedom. The assumptions of ANOVA F-test can cause limitations when the given data violates 1 or more of these assumptions. As a result, alternative tests may be employed to receive more accurate results.

The Welch's ANOVA test is an alternative of the ANOVA F-test. Welch's ANOVA performs better under conditions in which population variances are unequal. Similar to the ANOVA F-test, populations should have a normal distribution and hypotheses are the same. This test also uses the F statistic to calculate the p-value. This test has different sections in the function;

$$F = \frac{\frac{1}{k-1} \sum_{j=1}^{k} w_j (\overline{x}_j - \overline{x}')^2}{1 + \frac{2(k-2)}{k^2 - 1} \Lambda} \quad \text{is the F test statistic}$$

$$\text{where } w_j = \frac{n_j}{s_j^2}\ , \qquad w = \sum\nolimits_{j=1}^{k} w_j\ \ , \qquad \overline{x}' = \frac{\sum_{j=1}^{k} w_j \overline{x}_j}{w} \ \text{ and } \ \Lambda = \sum_{j=1}^{k} \left(\frac{1}{n_j - 1}\right)\left(1 - \frac{w_j}{w}\right)^2$$

$$\text{following } \ F \sim F(k - 1, df) \ \text{ where } df = \frac{(k^2-1)}{3\Lambda}$$

Marascuilo's alternative to Welch's ANOVA has simpler computation methods. Using this test is only advisable during time constraints. Assumptions and hypotheses are similar to the Welch's ANOVA. This will decrease accuracy, but processing time will be saved. The function difference is removing Λ and the its associated variables,

$$F = \frac{1}{k-1}\sum_{j=1}^{k} w_j\left(\overline{x}_j - \overline{x}'\right)^2$$

$$\text{following } \ F \sim F(k - 1, df) \ \text{ where } df = \frac{(k^2-1)}{3\Lambda}$$

The Kruskal-Wallis Test is used when the assumption of normality is not met for the ANOVA test. Assumptions for this test are similar to the ANOVA F test except that data do not need to be normally distributed. It is necessary for data to be sorted into ranks to compute this test statistic. Smallest value in a data has a rank of 1, second smallest has a rank 2, and so forth. Results explain the differences in rank means and not about the mean of the true variable. This is the only nonparametric test whereas all other tests in this study rely on data coming from a normal distribution. P-value is calculated from a chi-square distribution. This test function is,

$$H = \frac{12}{n(n+1)}\sum\nolimits_{j=1}^{k} \frac{R_j^2}{n_j} - 3(n+1)$$

$$\text{where } R_j = \text{the rank sum for the } j^{th} \text{ group,} \qquad n = \sum\nolimits_{j=1}^{k} n_j$$

$$\text{following } \ H \sim \chi^2(k - 1)$$

Hotelling's T^2 test is used for multivariate analysis of variance. Like the other tests, observations should be independent. This test requires vectors rather than different samples. The F statistic is given by the following function,

$$T^2 = n\left(\overline{\boldsymbol{X}} - \boldsymbol{\mu_0}\right)' \boldsymbol{S}^{-1}\left(\overline{\boldsymbol{X}} - \boldsymbol{\mu_0}\right)$$

$$\text{and } F = \frac{n-k}{k(n-1)}T^2 \ \sim \ F(k, n-k)$$

where $\overline{\boldsymbol{X}}$ and $\boldsymbol{\mu_0}$ are vectors of observed group means and true group mean, respectively while $\boldsymbol{S}$ is the observed variance covariance matrix of k groups.

This paper aims to determine how heterogeneity of variance affects the performance of these tests. The F-test and its alternatives will be run through a simulation, and the estimated type I error rate will be recorded. Following this, a discussion of findings and further actions will be discussed.

Literature Review

Much of the research around the ANOVA F-test and its alternatives employ Monte Carlo methods to test robustness. A few of these papers are listed below to establish the scholarship surrounding this topic. This study uses these papers to establish a framework for providing contemporary commentary on this subject.

Levy (1978) used the Monte Carlo simulation to check robustness against heterogeneity of variance and non-normality. Tests used in this study include the ANOVA F-test, Welch's v-test (now known as Welch's ANOVA), and Marascuilo's slight variant of Welch's v-test. Sample sizes used in this study are appropriate for other studies. This study looked at both varying and equal sample sizes under different variance levels. Through this study, Levy determined that Welch's v-test is a reasonable alternative to the ANOVA F-test.

Wilcox et al (1986) considered the parametric modifications of ANOVA F using F and W statistics described in a study done by Brown and Forsythe (1974). Wilcox et al reported situations in which F and W statistics do not provide ideal results such as having inadequate control over Type I errors (1986). Although these alternatives to the ANOVA F-test will not be observed, this paper is useful in understanding the use of Monte Carlo methods to check robustness of alternatives in statistical testing.

Simulation studies using SAS and R were used in a study by Moder in 2010. This study focuses on heteroscedasticity and its effects on Type I error rates for the ANOVA F-test, Welch-test, Kruskal Wallis test, Permutation test, and Hotelling's T^2 test. Simulations were based on balanced and unbalanced designs. Balanced designs contained equal sample sizes while unbalanced designs contained unequal sample sizes. Moder (2010) found that ANOVA F-test is unsuitable for analysis in cases of heteroscedasticity while Hotelling's T^2-test would be the best alternative because of its ability to maintain a small Type I error rate in all observed balanced designs in their study.

Blanca et al (2017) conducted a study of the ANOVA F-test and its ability to perform under differences in variance ratios by employing Monte Carlo methods. Variables in this study include unequal/equal group sample sizes and number of groups, total sample size, coefficient of sample size variation (amount of inequality in group sizes), ratio of largest to smallest variance, and patterns of variance and pairings of variance to group sample size (Blanca et al 2017). This study found that the ANOVA F-test was robust when group sample sizes were equal. When sample sizes were unequal, "robustness depends on the variance ratio, the pairing of variance with group size, and the coefficient of sample size variation, with the procedure being more robust when variance ratios are small, the pairing of variance is either zero or positive, and the coefficient of sample size variation is smaller" (Blanca et al 2017). Overall, this study concluded that 1.5 would be the highest variance ratio in order to avoid problems with Type I error. Blanca et al (2017) also considers the use of Kruskal Wallis test as an alternative, however it is noted that Monte Carlo methods in previous studies have shown that its Type I error rates are also affected by heterogeneity of variances.

Methodology and Results

Using information from previous studies, a similar simulation was created with different variances to better understand the changes between each test. Marascuilo's alternative to Welch's ANOVA was an uncommon test that was examined in comparison to the original. Simulation was conducted using R program. Three independent groups with equal sample size from normal distribution were the extent of this research. Data were randomly generated from the normal distribution with mean of 0 for all tests. Several different variance levels were chosen and four different sample sizes were examined.

To better explain the variance levels, this research chose a variance for each group. For instance, the first group would have a variance of 1, second group variance of 1, and finally third group had a variance of 5. This would be denoted as (1, 1, 5). In total all the variances including (1, 1, 5) are (1, 5, 5), (1, 1, 15), (1, 15, 15), (5, 15, 15), (5, 5, 15), (1, 1, 50), (1, 5, 15), (5, 5, 50), (1, 5, 50), (1, 50, 50), (5, 50, 50), (15, 50, 50), (1, 1, 1), (5, 5, 5), (15, 15, 15), and (50, 50, 50). Each variance level was used separately under different sample sizes. Sample sizes 5, 10, 25, 50 were used in order to understand how this might change the results of the tests. The true mean chosen was 0 in order to focus on heterogeneity of variance. These tests were analyzed under 5000 iterations for each sample size and variance level. These tests returned a p value that was analyzed with a significance level of $\alpha = 0.05$ at each iteration. And then the type 1 error rate was calculated as the proportion of time the p value was less than or equal to significance level (0.05) in 5000 iterations.

Table 1. Estimated Type 1 Error rate when sample size is 5

	ANOVA F	Welch's ANOVA	Marascuilo's Alternative	Kruskal Wallis	Hotelling's T2
Variances					
1, 1, 5	0.0974	0.0464	0.0576	0.065	0.0526
1,5, 5	0.0592	0.0548	0.0644	0.0594	0.053
1, 1, 15	0.1234	0.056	0.067	0.0746	0.0502
1, 15, 15	0.0698	0.059	0.0718	0.0822	0.052
5, 15, 15	0.0658	0.0526	0.0602	0.0572	0.0504
5, 5, 15	0.0824	0.0436	0.0536	0.0578	0.045
1, 1, 50	0.1222	0.057	0.0648	0.0696	0.048
1, 5, 15	0.1018	0.0519	0.065	0.0698	0.0488
5, 5, 50	0.1176	0.0496	0.0566	0.0692	0.0506
1, 50, 50	0.0648	0.0546	0.0684	0.0748	0.0494
1, 5, 50	0.1128	0.0596	0.0712	0.078	0.0456
5, 50, 50	0.0772	0.064	0.076	0.0786	0.0534
15, 50, 50	0.0612	0.055	0.0628	0.055	0.0512
1, 1, 1	0.0468	0.0414	0.0514	0.0386	0.0462
5, 5, 5	0.0484	0.0416	0.0512	0.042	0.0484
15, 15, 15	0.0464	0.0426	0.0518	0.0398	0.0492
50, 50, 50	0.048	0.0402	0.0522	0.0434	0.0476

Chart 1. Estimated Type 1 Error rate when sample size is 5

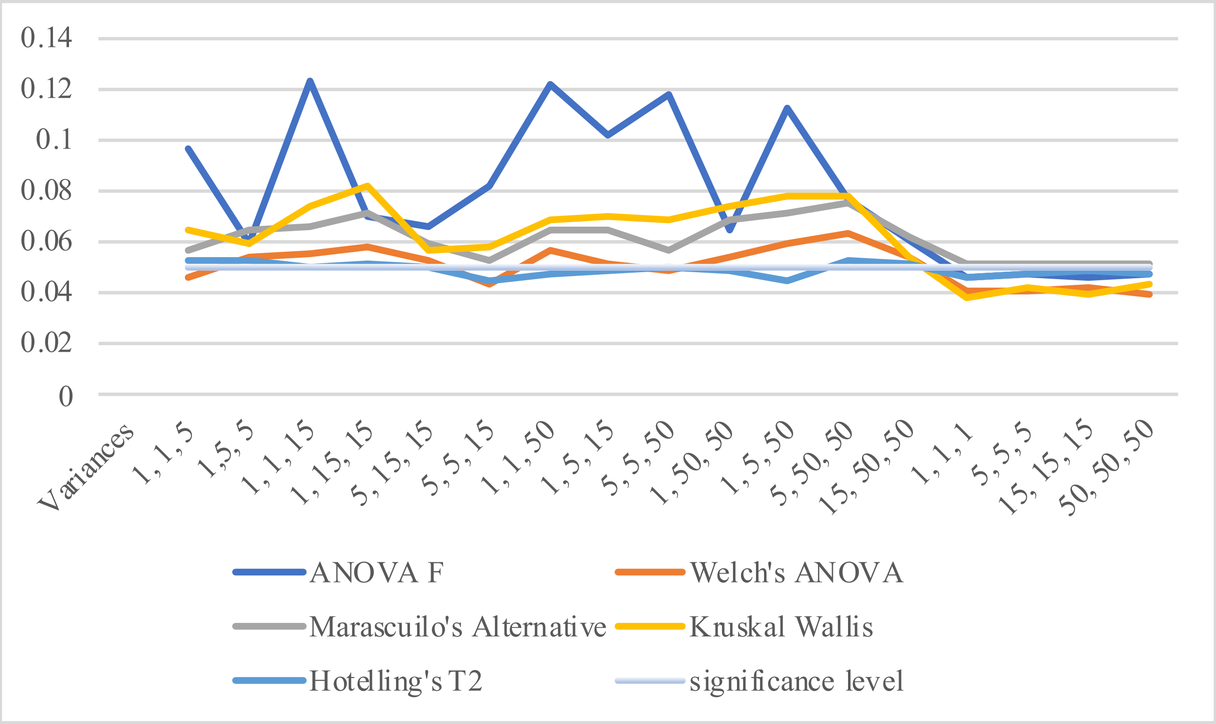


Table 2. Estimated Type 1 Error rate when sample size is 10

	ANOVA F	Welch's ANOVA	Marascuilo's Alternative	Kruskal Wallis	Hotelling's T2
Variances					
1, 1, 5	0.1018	0.0508	0.058	0.0812	0.0476
1,5, 5	0.0596	0.0488	0.0544	0.0618	0.0532
1, 1, 15	0.0876	0.0518	0.0582	0.1016	0.0492
1, 15, 15	0.0634	0.0504	0.0564	0.0728	0.054
5, 15, 15	0.0568	0.0464	0.0514	0.0512	0.044
5, 5, 15	0.0768	0.0452	0.0502	0.0596	0.0444
1, 1, 50	0.0934	0.0484	0.0544	0.0966	0.0524
1, 5, 15	0.0858	0.0574	0.064	0.0802	0.0502
5, 5, 50	0.0964	0.0552	0.0608	0.092	0.0544
1, 50, 50	0.065	0.0534	0.0608	0.0736	0.0502
1, 5, 50	0.0978	0.05	0.0594	0.1076	0.0464
5, 50, 50	0.063	0.053	0.0584	0.0716	0.0442
15, 50, 50	0.058	0.0506	0.0594	0.0586	0.049
1, 1, 1	0.047	0.0454	0.0494	0.0442	0.0406
5, 5, 5	0.0516	0.0468	0.0524	0.048	0.0516
15, 15, 15	0.0478	0.0446	0.0512	0.0432	0.0462
50, 50, 50	0.0556	0.0538	0.059	0.0512	0.0516

Chart 2. Estimated Type 1 Error rate when sample size is 10

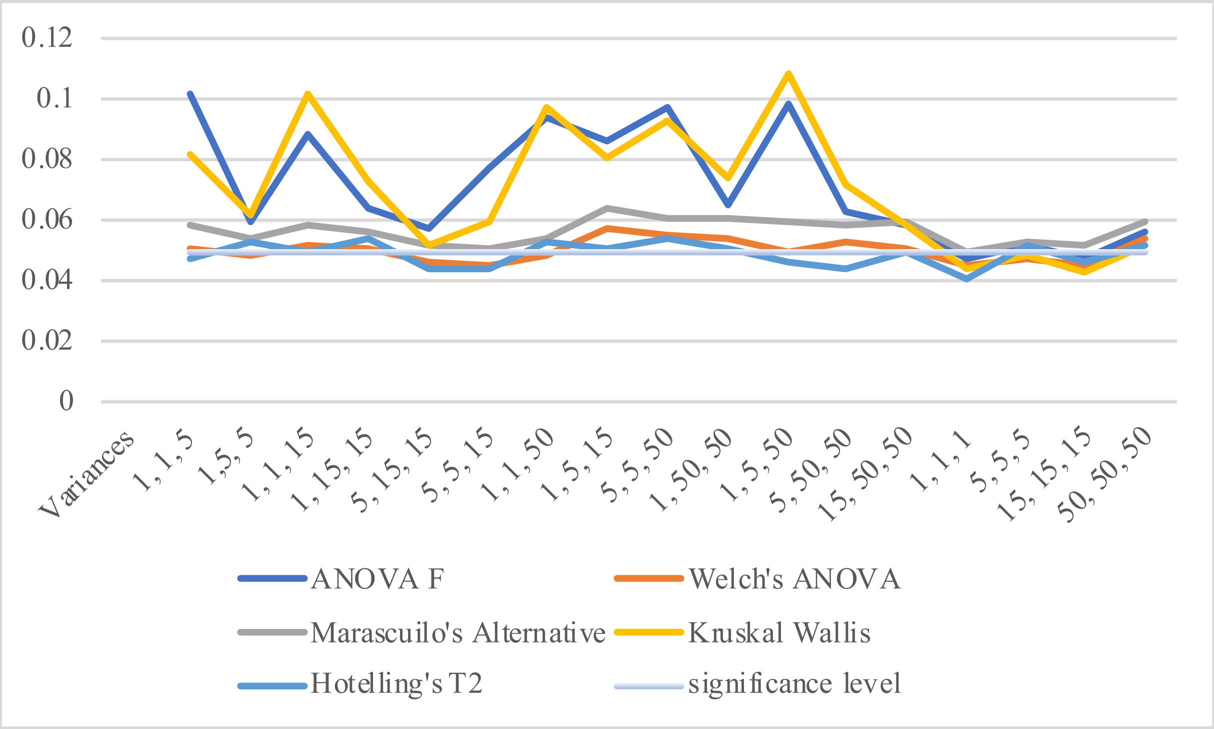


Table 3. Estimated Type 1 Error rate when sample size is 25

	ANOVA F	Welch's ANOVA	Marascuilo's Alternative	Kruskal Wallis	Hotelling's T2
Variances					
1, 1, 5	0.0804	0.0474	0.0494	0.0724	0.0458
1,5, 5	0.0554	0.0468	0.0498	0.0668	0.0502
1, 1, 15	0.088	0.048	0.0508	0.0928	0.0516
1, 15, 15	0.0644	0.05	0.053	0.0828	0.0576
5, 15, 15	0.0584	0.0462	0.0482	0.0576	0.0508
5, 5, 15	0.0716	0.0564	0.06	0.073	0.0556
1, 1, 50	0.0834	0.0486	0.0502	0.1062	0.0508
1, 5, 15	0.0696	0.045	0.0474	0.073	0.0456
5, 5, 50	0.0948	0.0464	0.0482	0.0846	0.0452
1, 50, 50	0.062	0.0484	0.0502	0.0804	0.0514
1, 5, 50	0.092	0.0522	0.055	0.1002	0.0528
5, 50, 50	0.0524	0.0476	0.05	0.07	0.0472
15, 50, 50	0.0602	0.0526	0.0542	0.0618	0.052
1, 1, 1	0.0524	0.0544	0.0554	0.0516	0.0536
5, 5, 5	0.0498	0.0468	0.0488	0.045	0.0484
15, 15, 15	0.0462	0.0472	0.0486	0.0472	0.0496
50, 50, 50	0.049	0.05	0.0516	0.0482	0.0506

Chart 3. Estimated Type 1 Error rate when sample size is 25

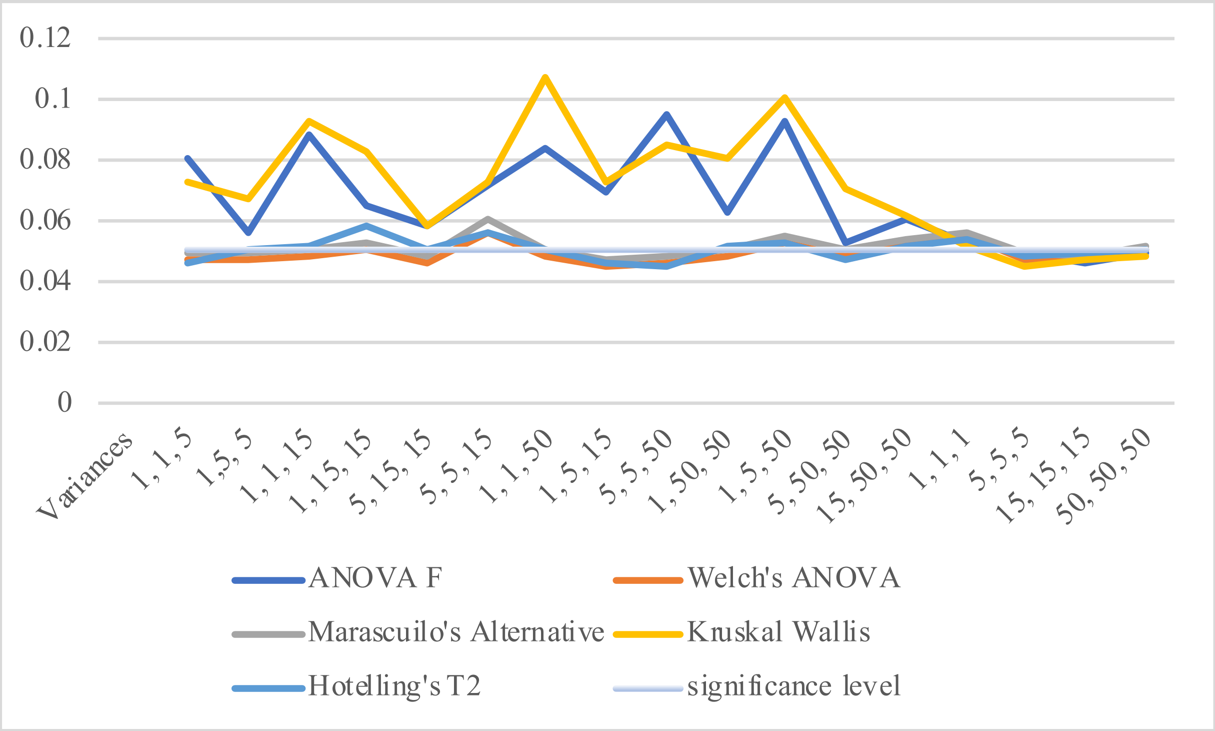
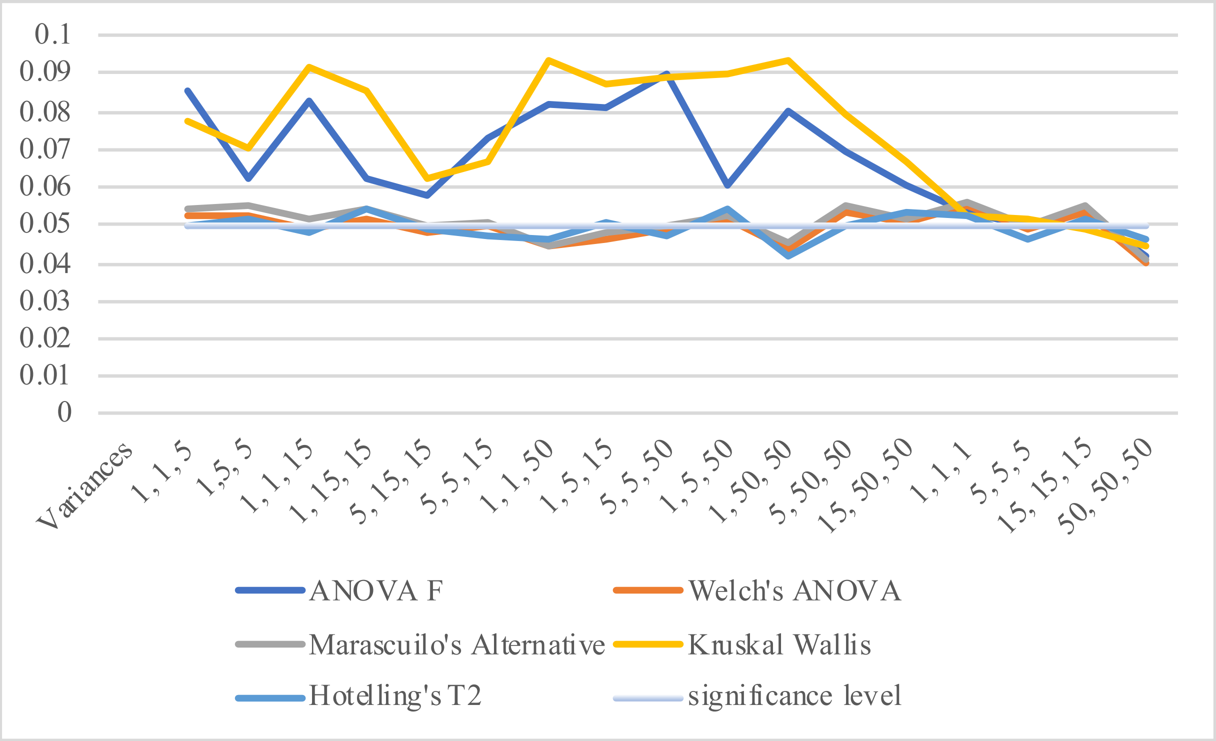


Table 4. Estimated Type 1 Error rate when sample size is 50

	ANOVA F	Welch's ANOVA	Marascuilo's Alternative	Kruskal Wallis	Hotelling's T2
Variances					
1, 1, 5	0.0854	0.0526	0.054	0.0776	0.05
1,5, 5	0.0626	0.0528	0.0548	0.0704	0.0512
1, 1, 15	0.083	0.0488	0.0514	0.092	0.0478
1, 15, 15	0.0622	0.0516	0.0544	0.0852	0.0544
5, 15, 15	0.0576	0.0484	0.0494	0.0624	0.0488
5, 5, 15	0.0732	0.0496	0.0508	0.0664	0.0474
1, 1, 50	0.0822	0.0442	0.0446	0.0934	0.0462
1, 5, 15	0.0808	0.0464	0.0478	0.0874	0.0506
5, 5, 50	0.09	0.0492	0.0496	0.0888	0.0476
1, 5, 50	0.0602	0.0518	0.0528	0.0896	0.0544
1, 50, 50	0.0804	0.044	0.045	0.0934	0.0414
5, 50, 50	0.0696	0.053	0.0554	0.0788	0.0502
15, 50, 50	0.0608	0.0506	0.0516	0.0668	0.053
1, 1, 1	0.0532	0.0556	0.056	0.0522	0.0524
5, 5, 5	0.0502	0.0492	0.0496	0.0514	0.0466
15, 15, 15	0.0534	0.0538	0.055	0.049	0.0514
50, 50, 50	0.0414	0.0404	0.041	0.0448	0.046

Chart 4. Estimated Type 1 Error rate when sample size is 50



The ANOVA F-test had a higher type 1 error rate due to the violation of heterogeneity of variances. The ANOVA F-test resulted in having a higher type 1 error rate when two groups had a lower variance and one group had a higher variance. Cases where variances were equal reduced the type 1 error significantly. Smaller sample sizes had a slight effect on the type I error rate. In some cases where variances were equal, the ANOVA F-test had a lower type 1 error than Welch's ANOVA and Hotelling's T^2 test. With sample size 5, the ANOVA F-test was less robust than samples 10, 25, and 50. The ANOVA F-test and the Kruskal Wallis test had some of the highest type 1 error rates. From the charts, it is visible that the type 1 error rates are closer to the significance level as sample size increases. It isn't until sample size 50 where the ANOVA F-test reduces the type 1 error rate below 0.1, but still fails to be within a reasonable range of the significance level.

Welch's ANOVA resulted in a smaller type 1 error rate than the ANOVA F-test. There were variance levels in which the type 1 error rate was reduced below the significance level of 0.05 and never exceeded an error rate of 0.06. Heterogeneity of variance had no effect on this test which shows that the assumption of equal variances isn't required for Welch's ANOVA. This test, along with Hotelling's T^2 test, was the most robust with very low type 1 error rates. Cases where variances were equal didn't have any noticeable trend in Welch's ANOVA.

Marascuilo's alternative to Welch's ANOVA had lower type 1 error rates than the ANOVA F-test and Kruskal Wallis test when there was heterogeneity of variances. Similar to Welch's ANOVA, heterogeneity of variance had no effect on the type 1 error rate. In every chart, type 1 error rate of both Welch's ANOVA and Marascuilo's alternative fluctuated around the significance level (0.05). The difference in type 1 error rate between Marascuilo's alternative and Welch's ANOVA is minimal. Marascuilo's alternative performed just as well as Hotelling's T^2 test except when the sample size was 5.

Kruskal Wallis test performed poorly against heterogeneity of variance. Type 1 error rate was significantly large. A few variance levels resulted in type 1 error rates close to 0.05, but the type I error rates for majority of variance levels were much higher than the significance level of 0.05. The Kruskal Wallis test had better results than the ANOVA F-test when sample size was 5. Sample sizes of 10, 15, and 25 are where the Kruskal Wallis test began to perform more poorly than the ANOVA F-test. Cases where variances were equal reduced the type 1 error rate significantly. In some cases where variances were equal, the Kruskal Wallis test had a lower type

1 error rate than Welch's ANOVA and Hotelling's T^2 test. The Kruskal Wallis test did not have a trend in cases where one group has a larger variance or where one group has a smaller variance. Although the type 1 error rate of the Kruskal Wallis test did follow the ANOVA F-test, it did not have the same results.

Hotelling's T^2 test had the other lowest type 1 error rate similar to Welch's ANOVA and Marascuilo's alternative. A few cases resulted in Hotelling's T^2 test performing better than Welch's ANOVA and Marascuilo's alternative. The difference in type 1 error rates between these three tests is very small. Hotelling's T^2 test performed significantly better than the Kruskal Wallis test and the ANOVA F-test. Sample size did not affect Hotelling's T^2 test. Similar to Welch's ANOVA, heterogeneity of variance did not cause any noticeable change in the type 1 error rate.

Conclusion

Overall, Hotelling's T^2 test and Welch's ANOVA proved to be more robust against heterogeneity of variances. It is important to acknowledge that Marascuilo's alternative to Welch's ANOVA did significantly better than the Kruskal Wallis test and ANOVA F-test. In some cases, Welch's ANOVA had a lower type 1 error rate than Hotelling's T^2 test. Performance of the Hotelling's T^2 test was similar to that of Welch's ANOVA. Therefore, it is advised to use either Hotelling's T^2 test or Welch's ANOVA depending on heterogeneity of variances. Further, the ANOVA F-test and Kruskal Wallis test have smaller type 1 error rates (though still more than significance level) when more groups have higher variances. Choosing an exact test all depends on the assumptions provided from a research so it is difficult to determine an exact test to use. Having this flexibility of being able to choose multiple tests can be overwhelming so it's beneficial comparing statistical tests under parameters that would come up in real life. Not all groups will have same variances. There might be differences in means, data coming from different distributions, or even different group sizes. In order to find more accurate results, it would be helpful to continue this study.

In the future, it would be interesting to see cases how effective and quick calculations from Marascuilo's alternative are given in comparison to Welch's ANOVA for larger samples. Changes in the equation may be slight, but it is beneficial to see more of the theory behind each test. Other methods for multivariate analysis of variance should be taken into consideration. Future work can also consist of using different distributions and finding more results on how these methods can be applied to count data

Acknowledgement

Thank you Dr. Nadeeshani Jayasena for your assistance throughout my time spent at Portland State University. Through your guidance I was able to understand difficult topics. Your teachings have helped so much these past two years. It’s been a very beneficial experience examining ANOVA F-test and its alternatives. I am grateful for this opportunity that I had working with you. I would also like to thank the researchers that have explored many more parameters and for sharing their work online.

Bibliography

Blanca M., Alarcon R., Arnau J., Bono R., Bendayan R. (2017) Effect of variance ratio on ANOVA robustness: Might 1.5 be the limit? 50:937. https://doi.org/10.3758/s13428-017-0918-2

Brown, M. B., & Forsythe, A. B. (1974). The small sample behaviour of some statistics which test the equality of several means. *Technomectrics, 16,* 129–132.
doi: 10.1080/00401706.1974.10489158

"F-Test: Compare Two Variances in R." STHDA. Accessed July 24, 2019. http://www.sthda.com/english/wiki/f-test-compare-two-variances-in-r.

Hotelling, Harold. "The Generalization of Students Ratio." *The Annals of Mathematical Statistics*2, no. 3 (1931): 360-78. doi:10.1214/aoms/1177732979.

"Kruskal-Wallis Test in R." STHDA. Accessed July 24, 2019. http://www.sthda.com/english/wiki/kruskal-wallis-test-in-r.

Levy K. J. (1978) An empirical comparison of the ANOVA F-test with alternatives which are more robust against heterogeneity of variance, Journal of Statistical Computation and Simulation, 8:1, 49-57.

Moder, Karl. "How to Keep the Type I Error Rate in ANOVA If Variances Are Heteroscedastic." *Austrian Journal of Statistics*36, no. 3 (January 2007): 179-88. doi:10.17713/ajs.v36i3.329.

Moder Karl. (2010) Alternatives to F-Test in One Way ANOVA in case of heterogeneity of variances (a simulation study) , Psychological Test and Assessment Modeling, Volume 52, 2010 (4), 343-353

Spector, Phil. "Using T-tests in R." Department of Statistics. Accessed July 24, 2019. https://statistics.berkeley.edu/computing/r-t-tests.

"Unpaired Two-Samples T-test in R." STHDA. Accessed July 24, 2019. http://www.sthda.com/english/wiki/wiki.php?id_contents=7600.

Wilcox R. R., Ventura L., Char1in, L., Thompson, C. L. (1986) New monte carlo results on the robustness of the anova f, w and f statistics, Communications in Statistics - Simulation and Computation, 15:4, 933-943, DOI: 10.1080/03610918608812553