Simulation Study of ANOVA

Cyrus Navasca

December 2024

Introduction

One of the first questions that aspiring statisticians learn to answer is whether or not two group means have statistically significant differences. It turns out that this is one of the most important questions in statistical analysis, with diverse applications in fields such psychology, economics and healthcare. The test used to answer this question is most commonly referred to as a "two-sample t-test" and is relatively simple to implement. However, for comparing the means of more than two groups, statisticians utilize a slightly more rigorous test called the "Analysis of Variance", or ANOVA test for short.

The ANOVA test is "parametric", meaning there are assumptions that need to be met in order to ensure the validity of its results (Vimal, Venugopal, and Anandabaskar, 877-888). Specifically, these assumptions are: 1) Normality of data 2) Independence of observations 3) Equal variances between groups (Homoscedasticity). However, data from the real-world often fails to meet all three of these conditions. Thus, statisticians have posed the question: "how valid are the results of an ANOVA test when its assumptions are not met?"

Research suggests that the ANOVA test is robust to violations of the normality assumption, as well as violations of homoscedasticity when the experiment is balanced, meaning the sample sizes between groups is equal. However, in unbalanced designs, it follows that the ANOVA test becomes less robust with differences in variance, leading to inflated Type I (false positive) and Type II (false negative) error rates (Blanca et al. 937). One solution is to transform the underlying data to meet the assumptions, but this leads to difficulty in the interpretation of results and lack of ability to make conclusions on the original data (Sheng, 325).

However, a more promising solution is the use of the permutation ANOVA test, which unlike the standard ANOVA test is "non-parametric", meaning that it requires no assumptions about the underlying data. This makes it a good choice for comparing the means of more than two groups, especially in real-world applications where it can be difficult to come across data that meets all of the ANOVA test assumptions. The permutation test is conducted by first running the standard ANOVA test on the original data, then repeatedly shuffling the data and re-running the ANOVA test. The p-values obtained from these shuffled, or "permuted", ANOVA tests are then compared to the p-value obtained by the original ANOVA test. The proportion of permuted p-values that make the same decision as the original p-value is used as the test's final p-value.

It is recommended that the data is permuted at least 10,000 times to ensure the validity of the permutation test. However, while more permutations may lead to more accurate results, it also creates a downside of the test where it becomes more exhaustive and computationally extensive to run permutation tests compared to the standard ANOVA. This highlights the importance of balancing computational resources with accuracy of the statistical analysis.

The aim of this study is to examine the robustness of both the standard ANOVA test and the permutation ANOVA test when the assumptions of ANOVA are violated. A "robust test" in this study will be denoted by low Type I (false positive) Error rate and a high statistical power (true positive rate). We will examine how both of these tests perform under different combinations of groups means, group variances and sample sizes. Understanding when these tests perform well and when they don't should hopefully give a good idea on when to use each test or if we should use them at all.

Methods

Study Description

In this study, we will conduct 28 separate standard ANOVA and ANOVA permutation tests, where each will have a varying degree of difference in the following factors: group means, group variances and sample sizes.

When varying group variances, we will use group standard deviations of 7.5, 10 and 12.5 to represent a "low difference in variance" with standard deviations of 5, 10 and 15 to represent a "high difference in variance". For the case of equal group variances, we will use standard deviations of 1 for simplicity.

The group means for each test will be dependent on the group variance. For a "low difference in means", the difference between groups will be under 1 standard deviation and for a "high difference in means" the difference between groups will be 1 standard deviation or greater.

For sample sizes we will keep a consistent total of 90 samples with a median of 30. Group sample sizes will be set at 20, 30, and 40 to represent a "low difference in sample size", at 10, 30, and 50 to represent a "high difference in sample size" and at 30 each for equal sample sizes.

Additionally, we will randomly sample from the Normal distribution to easily manipulate group means and variances for the purpose of this study, as well as allow us to hone in on the effects of variance and sample sizes on these tests without worry of the Normality assumption.

Understanding ANOVA

Let us examine how the standard ANOVA test works. To restate the introduction page, the Analysis of Variance, or ANOVA test, is used to compare the means of three or more groups. In statistical terms, the ANOVA test allows us to make a decision on the following hypothesis:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$
$$H_A : \text{at least one mean is different}$$

where k is the total number of groups and μ_i is the mean of the i_{th} group.

Similar to a t-test, the ANOVA test outputs a test statistic which is compared to a critical value to give a decision on the above hypothesis. To understand where this test statistic comes from, let us define some of its key components.

First, let us define the "Sum of Squares due to Treatments" by:

$$SS_{Treatment} = \sum_{i=1}^{a} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})$$

where a is the total number of groups, n_i is the sample size of the i_{th} group, $\bar{y}_{i\cdot}$ is the mean of the i_{th} group and $\bar{y}_{\cdot\cdot}$ is the total mean of all observations.

Utilizing this result allows us to define the "Mean Squares due to Treatment", which is a representation of the variability across groups. Let us write it as:

$$MS_{Treatment} = \frac{SS_{Treatment}}{a-1}$$

Next, let us define the "Sum of Squares due to Error" as:

$$SS_E = \sum_{i=1}^{a} \sum_{j=1}^{n} (\bar{y}_{ij} - \bar{y}_{\cdot\cdot})$$

where $\bar{y}_{ij}$ is the value of the j_{th} observation in the i_{th} group.

This result allows us to define the "Mean Squares due to Error", which quantifies the average variability of observations within each group. It is written as:

$$MS_E = \frac{SS_E}{N-a}$$

where N is the total sample size including all groups.

Examining $MS_{Treatment}$ and MS_E provide insight onto the decision of our hypothesis. Observe that if our null hypothesis H_0 was true, meaning that there is no significant difference between group means, then each group mean should be roughly equal to the total mean as grouping would not effect the value of an observation. In other words, $\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}$ would be close to 0. Thus, a small $SS_{Treatment}$ and consequently a small $MS_{Treatment}$ would indicate that H_0 is true. This provides an intuition of the test statistic of the ANOVA test, called the F-statistic, which is defined as:

$$F_0 = \frac{MS_{Treatment}}{MS_E}$$

The above description allows one to assume that when $|F_0|$ is small, we have a greater chance of the null hypothesis H_0 being true. This is indeed the case, and the reason for this is because F_0 will follow an F-distribution when the null hypothesis is true. Thus, the critical value of the ANOVA test is

$$F_{a-1,N-a,\frac{\alpha}{2}}$$

where $a-1$ and $N-a$ are degrees of freedom, and α is the significance level which is typically equal to 0.05 for a 95% confidence interval.

Each of the above results lead to the rejection region of the test, which is defined as

$$|F_0| > F_{a-1,N-a,\frac{\alpha}{2}}$$

When the rejection region is satisfied, we reject the null hypothesis H_0 and conclude that there is a statistically significant difference between group means. Consequently, when the rejection region is not satisfied, we fail to reject the null hypothesis and do not have evidence to suggest a statistically significant difference between group means.

Understanding Permutation Tests

A permutation test is widely used for a variety of statistical tests, with one of the most prominent being the ANOVA test. The main benefit of the permutation ANOVA test is that it is non-parametric, meaning it does not require any assumptions about the underlying data unlike the standard ANOVA test. It is used in conjunction with the same hypothesis, but approaches it with a different method. The following steps are repeated at least 10,000 times as the bulk of the ANOVA permutation test:

1) Calculate the F-statistic of the original data
2) Randomly shuffle the values of the data
3) Compute the F-statistic of the permuted data

After this computation, the test's p-value is equal to the proportion of permuted F-statistics which are at least as extreme as the original F-statistic. If the tests p-value is less than or equal to the significance level α, then we reject the null hypothesis. Conversely, when the tests p-value is greater than α, we fail to reject the null hypothesis.

Technical Issues of the Study

As mentioned in the introduction, and as one may assume, there is a high runtime associated with running a permutation test, especially at 10,000 repetitions. Additionally, the estimation of Type I/Type II errors in this study were run with 1,000 repetitions. This means that for each of the 28 combinations of factors, 1,000 standard ANOVA and permutation tests are run each. Since each permutation test is composed of 10,000 permuted ANOVA tests, it follows that each simulation is made up of 10,001,000 ANOVA tests for a total of 28,028,000 ANOVA tests for this study (not including the calculation of the original F-statistic in each permutation test, which is negligible).

With that being said, this study was computationally exhaustive and the laptop that was tasked with running this simulation often experienced overheating. More importantly, the runtime of this study was extremely long. The simulations were split into two, with the first run of 11 simulations taking about 18 hours and the final run taking a similar amount of time.

Results

```
# Installing Dependencies
library(tidyverse)
library(knitr)
```

```
source("PSTAT-122-Simulations.R")
```

```
source("PSTAT-122-Simulations (2).R")
```

Examination of Type I Error Rate

In hypothesis testing, a Type I Error is classified as incorrectly rejecting the null hypothesis. Here, a Type I Error specifically means that we found evidence to suggest a statistically significant difference in means, when there is no significant difference.

In this section, we will examine the effect of different combinations of the aforementioned factors on Type I Error Rates of both standard ANOVA and permutation ANOVA tests.

```
# Congregating Type I Errors into vectors
aov_typeI_errors = c(
  meanEQ_sdEQ_nEQ$aov_typeI_error, meanEQ_sdEQ_nLOW$aov_typeI_error,
  meanEQ_sdEQ_nHIGH$aov_typeI_error, meanEQ_sdLOW_nEQ$aov_typeI_error,
  meanEQ_sdLOW_nLOW$aov_typeI_error, meanEQ_sdLOW_nHIGH$aov_typeI_error,
  meanEQ_sdHIGH_nEQ$aov_typeI_error, meanEQ_sdHIGH_nLOW$aov_typeI_error,
  meanEQ_sdHIGH_nHIGH$aov_typeI_error)

perm_typeI_errors = c(
  meanEQ_sdEQ_nEQ$perm_typeI_error, meanEQ_sdEQ_nLOW$perm_typeI_error,
  meanEQ_sdEQ_nHIGH$perm_typeI_error, meanEQ_sdLOW_nEQ$perm_typeI_error,
  meanEQ_sdLOW_nLOW$perm_typeI_error, meanEQ_sdLOW_nHIGH$perm_typeI_error,
  meanEQ_sdHIGH_nEQ$perm_typeI_error, meanEQ_sdHIGH_nLOW$perm_typeI_error,
  meanEQ_sdHIGH_nHIGH$perm_typeI_error
  )
```

```r
# Congregating data for simulations with equal means into a data frame
typeI_errors_df <- data.frame(
  mean = rep("Equal (=)", 9),
  sd = rep(c("Equal (=)", "Low (-)", "High (+)"), each = 3),
  n = rep(c("Equal (=)", "Low (-)", "High (+)"), times = 3),
  aov_error = aov_typeI_errors,
  perm_error = perm_typeI_errors
)

# Rounding Type I Errors
typeI_errors_rounded <- typeI_errors_df
typeI_errors_rounded[] <- lapply(typeI_errors_rounded, function(x) if(is.numeric(x))
  round(x, 2) else x)

# Renaming columns
colnames(typeI_errors_rounded) <- c("Difference in Group Mean",
                                    "Difference in Group Variance",
                                    "Difference in Sample Size",
                                    "Standard ANOVA Type I Error Rate",
                                    "Permutation Test Type I Error Rate")

# Output
kable(typeI_errors_rounded, caption="Simulated Type I Error Rates")
```

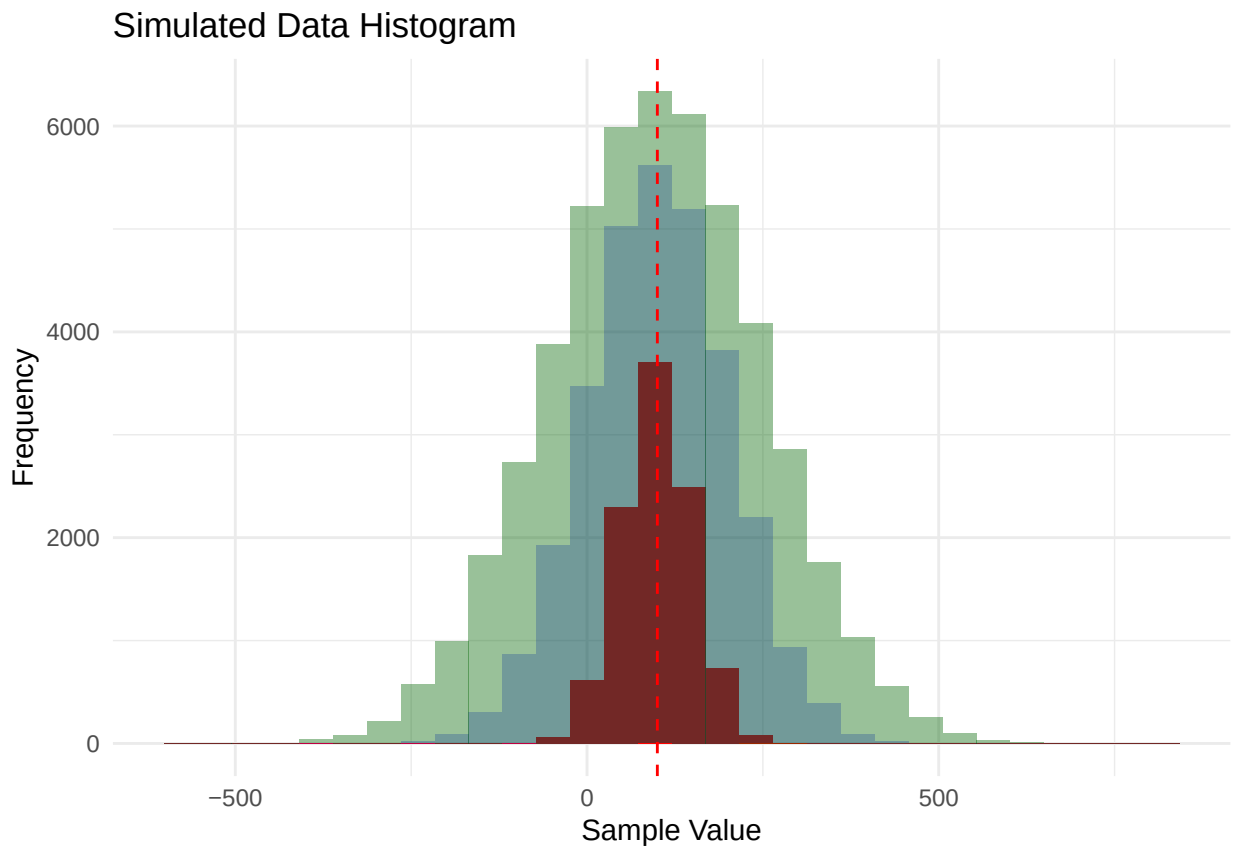
Table 1: Simulated Type I Error Rates

Difference in Group Mean	Difference in Group Variance	Difference in Sample Size	Standard ANOVA Type I Error Rate	Permutation Test Type I Error Rate
Equal (=)	Equal (=)	Equal (=)	0.06	0.08
Equal (=)	Equal (=)	Low (-)	0.04	0.07
Equal (=)	Equal (=)	High (+)	0.07	0.06
Equal (=)	Low (-)	Equal (=)	0.05	0.04
Equal (=)	Low (-)	Low (-)	0.03	0.03
Equal (=)	Low (-)	High (+)	0.01	0.01
Equal (=)	High (+)	Equal (=)	0.06	0.06
Equal (=)	High (+)	Low (-)	0.03	0.05
Equal (=)	High (+)	High (+)	0.02	0.02

By examining the table above, we find that the greatest Type I Error Rate for the standard ANOVA test is 0.07 and 0.05 for the permutation ANOVA test. Furthermore, although we may expect Type I Error Rate to greatly increase as differences in group variances and sample sizes increase, we actually see the opposite. To further this point, let us add another row to this table where the difference in sample size is high and difference in group variance is very high. Specifically, we will continue to use sample sizes of 10, 30, 50 but now in conjunction with group standard deviations of 50, 100 and 150. Let us visualize the randomly sampled groups of this experiment before analyzing the results.

```r
ggplot() +
  geom_histogram(aes(x = meanEQ_sdEXTREME_nHIGH$sample1), fill="red") +
  geom_histogram(aes(x = meanEQ_sdEXTREME_nHIGH$sample2),
                 fill="blue", alpha=0.25) +
  geom_histogram(aes(x = meanEQ_sdEXTREME_nHIGH$sample3),
                 fill="darkgreen", alpha=0.4) +
```

```
  geom_vline(xintercept=100, color="red", linetype="dashed") +
  labs(x="Sample Value", y="Frequency",
       title="Simulated Data Histogram") +
  theme_minimal()
```



The following choice of factors results in the following:

```
# Creating of data for simulation with very high difference of variance
typeI_errors_df2 <- data.frame(
  mean = "Equal (=)",
  sd = "Very High (++)",
  n = "High (+)",
  aov_error = meanEQ_sdEXTREME_nHIGH$aov_typeI_error,
  perm_error = meanEQ_sdEXTREME_nHIGH$perm_typeI_error
)

# Rounding Type I Errors
typeI_errors_rounded2 <- typeI_errors_df2
typeI_errors_rounded2[] <- lapply(typeI_errors_rounded2, function(x) if(is.numeric(x))
  round(x, 2) else x)

# Renaming columns
colnames(typeI_errors_rounded2) <- c("Difference in Group Mean",
                                     "Difference in Group Variance",
                                     "Difference in Sample Size",
                                     "Standard ANOVA Type I Error Rate",
```

```
                                    "Permutation Test Type I Error Rate")

# Output
kable(typeI_errors_rounded2, caption="Simulated Type I Error Rate with Very
      High Difference in Variance and High Difference in Sample Size")
```

Table 2: Simulated Type I Error Rate with Very High Difference in Variance and High Difference in Sample Size

Difference in Group Mean	Difference in Group Variance	Difference in Sample Size	Standard ANOVA Type I Error Rate	Permutation Test Type I Error Rate
Equal (=)	Very High (++)	High (+)	0.02	0.02

The above simulation results in Type I Error Rates of 0.02 for both standard ANOVA and permutation ANOVA tests. This means that in this scenario, we have an estimated 2% probability to incorrectly reject the null hypothesis. This is is a very optimal result, as indicates that both tests are robust.

This solidifies the observation that both standard ANOVA and permutation ANOVA tests perform well when group means are equal and result in low Type I Error Rates regardless of differences in group variances and sample sizes.

While this may seem contradictory to claims made by researchers in the introduction, this is only one side of the coin. In the next section, we will examine the performance of these statistical tests when group means are not equal.

Examination of Statistical Power

Statistical power is defined as the probability of correctly rejecting the null hypothesis. This means that our statistical analysis appropriately found evidence suggesting a significant difference in group means. It can be simply calculated by 1 - Type II Error Rate.

In this section, we will examine the effectiveness of both standard ANOVA and permutation ANOVA tests in terms of statistical power,

```
# Congregating statistical powers into vectors
aov_powers <- c(
    1 - meanLOW_sdEQ_nEQ$aov_typeII_error, 1 - meanLOW_sdEQ_nLOW$aov_typeII_error,
    1 - meanLOW_sdEQ_nHIGH$aov_typeII_error, 1 - meanLOW_sdLOW_nEQ$aov_typeII_error,
    1 - meanLOW_sdLOW_nLOW$aov_typeII_error, 1 - meanLOW_sdLOW_nHIGH$aov_typeII_error,
    1 - meanLOW_sdHIGH_nEQ$aov_typeII_error, 1 - meanLOW_sdHIGH_nLOW$aov_typeII_error,
    1 - meanLOW_sdHIGH_nHIGH$aov_typeII_error, 1 - meanHIGH_sdEQ_nEQ$aov_typeII_error,
    1 - meanHIGH_sdEQ_nLOW$aov_typeII_error, 1 - meanHIGH_sdEQ_nHIGH$aov_typeII_error,
    1 - meanHIGH_sdLOW_nEQ$aov_typeII_error, 1 - meanHIGH_sdLOW_nLOW$aov_typeII_error,
    1 - meanHIGH_sdLOW_nHIGH$aov_typeII_error, 1 - meanHIGH_sdHIGH_nEQ$aov_typeII_error,
    1 - meanHIGH_sdHIGH_nLOW$aov_typeII_error, 1 - meanHIGH_sdHIGH_nHIGH$aov_typeII_error
)

perm_powers <- c(
    1 - meanLOW_sdEQ_nEQ$perm_typeII_error, 1 - meanLOW_sdEQ_nLOW$perm_typeII_error,
    1 - meanLOW_sdEQ_nHIGH$perm_typeII_error, 1 - meanLOW_sdLOW_nEQ$perm_typeII_error,
    1 - meanLOW_sdLOW_nLOW$perm_typeII_error, 1 - meanLOW_sdLOW_nHIGH$perm_typeII_error,
    1 - meanLOW_sdHIGH_nEQ$perm_typeII_error, 1 - meanLOW_sdHIGH_nLOW$perm_typeII_error,
```

```
    1 - meanLOW_sdHIGH_nHIGH$perm_typeII_error, 1 - meanHIGH_sdEQ_nEQ$perm_typeII_error,
    1 - meanHIGH_sdEQ_nLOW$perm_typeII_error, 1 - meanHIGH_sdEQ_nHIGH$perm_typeII_error,
    1 - meanHIGH_sdLOW_nEQ$perm_typeII_error, 1 - meanHIGH_sdLOW_nLOW$perm_typeII_error,
    1 - meanHIGH_sdLOW_nHIGH$perm_typeII_error, 1 - meanHIGH_sdHIGH_nEQ$perm_typeII_error,
    1 - meanHIGH_sdHIGH_nLOW$perm_typeII_error, 1 - meanHIGH_sdHIGH_nHIGH$perm_typeII_error
)
```

```
# Congregating data for simulations with unequal means into a data frame
powers_df <- data.frame(
  mean = rep(c("Low (-)", "High (+)"), each = 9),
  sd = rep(c("Equal (=)", "Low (-)", "High (+)"), each = 3, times = 2),
  n = rep(c("Equal (=)", "Low (-)", "High (+)"), times = 6),
  aov_power = aov_powers,
  perm_power = perm_powers
)

# Rounding power values
powers_rounded <- powers_df
powers_rounded[] <- lapply(powers_rounded, function(x) if(is.numeric(x))
  round(x, 2) else x)

# Renaming columns
colnames(powers_rounded) <- c("Difference in Group Mean",
                              "Difference in Group Variance",
                              "Difference in Sample Size",
                              "Standard ANOVA Power",
                              "Permutation Test Power")

# Output
kable(powers_rounded, caption="Simulated Statistical Powers")
```

Table 3: Simulated Statistical Powers

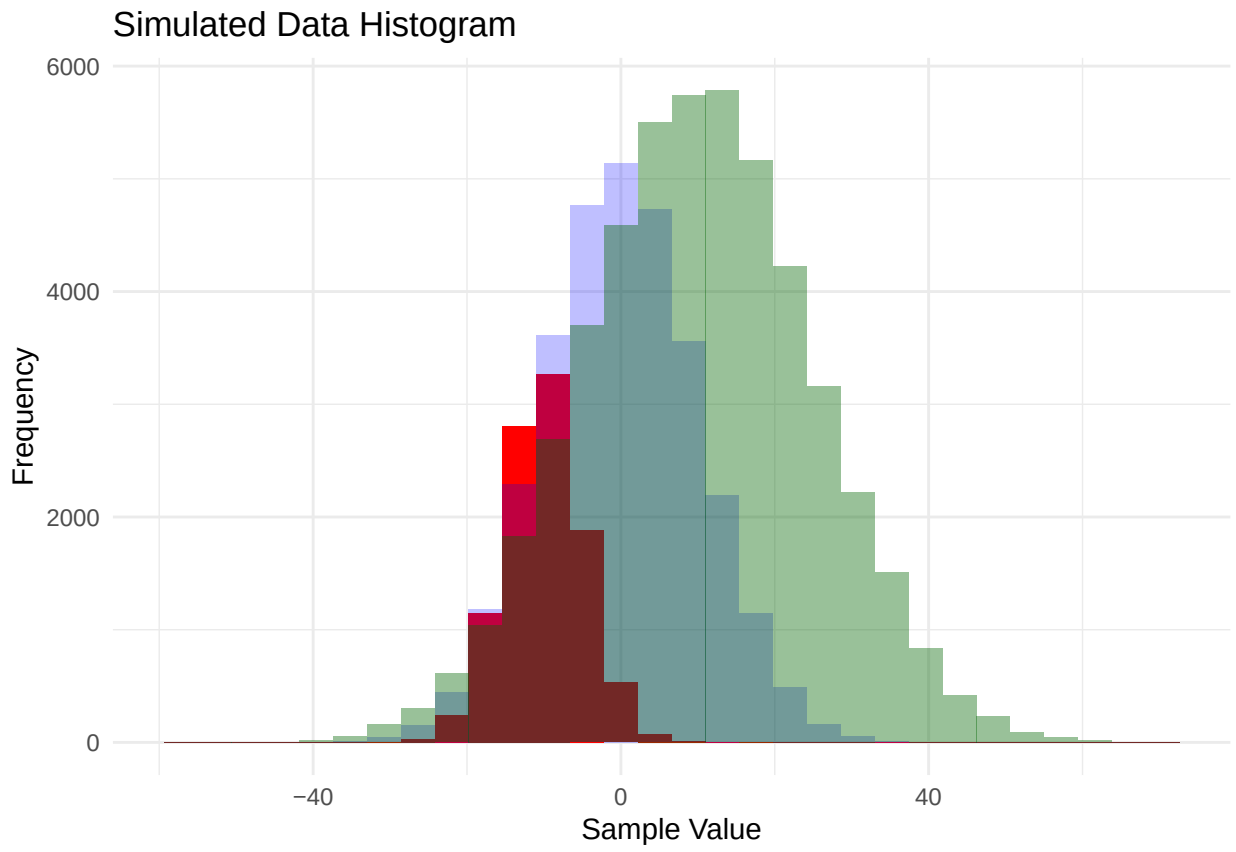
Difference in Group Mean	Difference in Group Variance	Difference in Sample Size	Standard ANOVA Power	Permutation Test Power
Low (-)	Equal (=)	Equal (=)	0.96	0.96
Low (-)	Equal (=)	Low (-)	0.93	0.93
Low (-)	Equal (=)	High (+)	0.83	0.83
Low (-)	Low (-)	Equal (=)	0.93	0.93
Low (-)	Low (-)	Low (-)	1.00	1.00
Low (-)	Low (-)	High (+)	0.95	0.95
Low (-)	High (+)	Equal (=)	0.89	0.89
Low (-)	High (+)	Low (-)	0.81	0.81
Low (-)	High (+)	High (+)	0.54	0.54
High (+)	Equal (=)	Equal (=)	1.00	1.00
High (+)	Equal (=)	Low (-)	1.00	1.00
High (+)	Equal (=)	High (+)	1.00	1.00
High (+)	Low (-)	Equal (=)	0.92	0.93
High (+)	Low (-)	Low (-)	0.91	0.89
High (+)	Low (-)	High (+)	0.69	0.65
High (+)	High (+)	Equal (=)	0.87	0.86
High (+)	High (+)	Low (-)	0.83	0.84

Difference in Group Mean	Difference in Group Variance	Difference in Sample Size	Standard ANOVA Power	Permutation Test Power
High (+)	High (+)	High (+)	1.00	1.00

By analyzing the results above, we approach the observations made by researchers as cited in the introduction. When the difference in group means are low along with high differences in variance and sample size, both the standard ANOVA test and permutation ANOVA test have powers of 0.54. This interprets as a 54% probability of correctly rejecting the null hypothesis, which is slighly more accurate than a coin flip. This is unideal for a statistical analysis and gives evidence to support the claim that the ANOVA test is not robust to violations of homoscedasticity in unbalanced designs.

However, an interesting result appears when examining the same situation but with a greater difference in means. Here, we obtain statistical powers of 1 for both the standard ANOVA and permutation ANOVA tests. Let us first examine the simulated data that was used in this test:

```
ggplot() +
  geom_histogram(aes(x = meanHIGH_sdHIGH_nHIGH$sample1), fill="red") +
  geom_histogram(aes(x = meanHIGH_sdHIGH_nHIGH$sample2),
                 fill="blue", alpha=0.25) +
  geom_histogram(aes(x = meanHIGH_sdHIGH_nHIGH$sample3),
                 fill="darkgreen", alpha=0.4) +
  labs(x="Sample Value", y="Frequency",
       title="Simulated Data Histogram") +
  theme_minimal()
```



The simulated data appears to match the parameters given to it, so why do we obtain such an unusual

result? One plausible explanation may be that this combination of factors make the greater difference in means easier detect in comparison to a lower difference in group means for example. However, this is still an anomaly in our study and statistical power may need to be estimated with a greater amount of repetitions.

Moving on, let us examine the powers of both tests when group variances are equal.

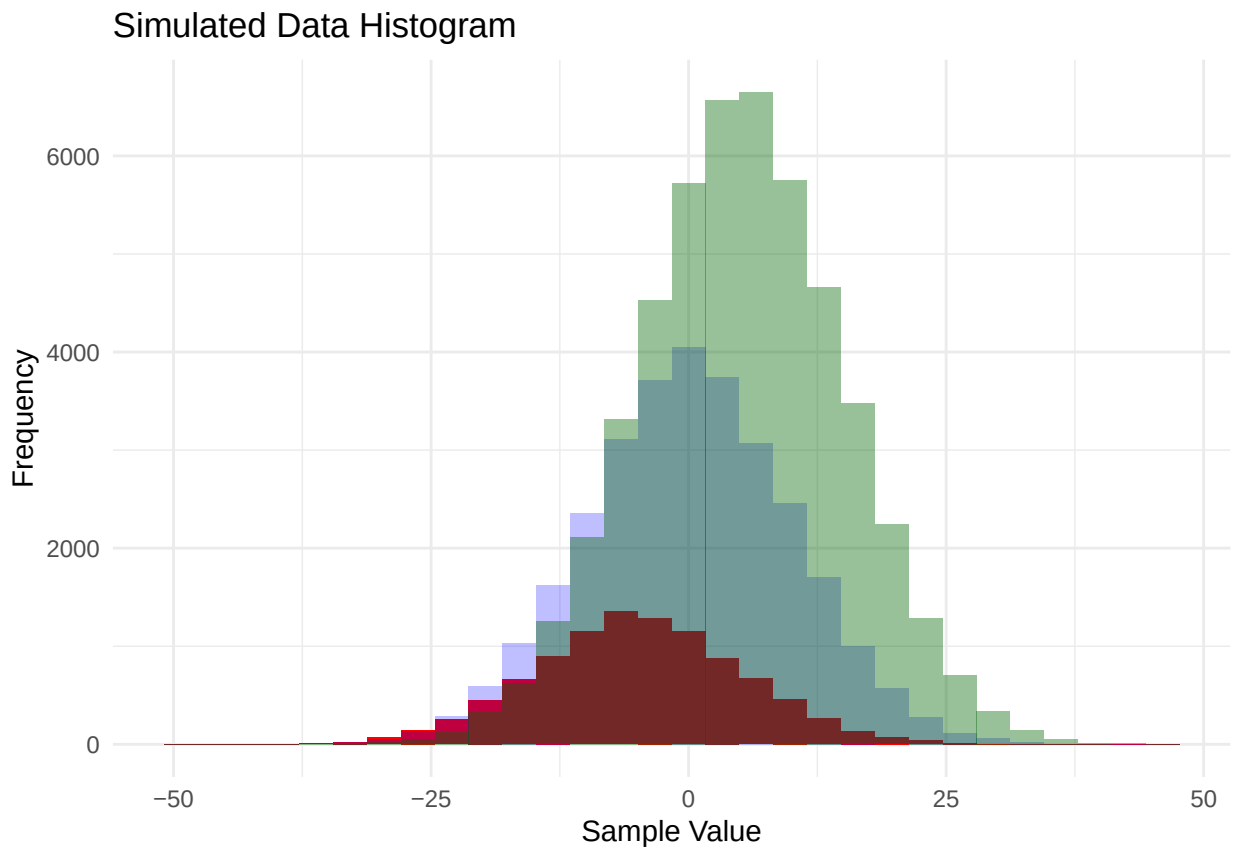
```
# Retrieving values where sample sizes are equal
powers_equalsd_df <- powers_df[powers_df$sd == "Equal (=)", ]

# Removing row indexing
rownames(powers_equalsd_df) <- NULL

# Renaming columns
colnames(powers_equalsd_df) <- c("Difference in Group Mean",
                                 "Difference in Group Variance",
                                 "Difference in Sample Size",
                                 "Standard ANOVA Power",
                                 "Permutation Test Power")

# Rounding power values
powers_equalsd_rounded <- powers_equalsd_df
powers_equalsd_rounded[] <- lapply(powers_equalsd_rounded, function(x) if(is.numeric(x))
  round(x, 2) else x)

# Output
kable(powers_equalsd_rounded, caption="Simulated Statistical Powers with Equal
     Group Variances")
```

Table 4: Simulated Statistical Powers with Equal Group Variances

Difference in Group Mean	Difference in Group Variance	Difference in Sample Size	Standard ANOVA Power	Permutation Test Power
Low (-)	Equal (=)	Equal (=)	0.96	0.96
Low (-)	Equal (=)	Low (-)	0.93	0.93
Low (-)	Equal (=)	High (+)	0.83	0.83
High (+)	Equal (=)	Equal (=)	1.00	1.00
High (+)	Equal (=)	Low (-)	1.00	1.00
High (+)	Equal (=)	High (+)	1.00	1.00

Here, we see that both tests are relatively robust to almost all combinations of differences in group means and sample size with equal group variances. Moreover, differences in group means are high, we obtain statistical powers of 1 for each combination for both tests. The one exception is when differences in group means is low and differences in sample size is high. Let us visualize the sampled data in this scenario:

```
ggplot() +
  geom_histogram(aes(x = meanLOW_sdEQ_nHIGH$sample1), fill="red") +
  geom_histogram(aes(x = meanLOW_sdEQ_nHIGH$sample2),
                 fill="blue", alpha=0.25) +
  geom_histogram(aes(x = meanLOW_sdEQ_nHIGH$sample3),
                 fill="darkgreen", alpha=0.4) +
  labs(x="Sample Value", y="Frequency",
       title="Simulated Data Histogram") +
  theme_minimal()
```



Note that in this visualization, the red chart is the sample with the lowest sample size. This visualization reveals that as the sample size decreases, the variance of the sample seems to alter, with the data becoming more spread out. This is a likely explanation to the relatively lower powers of 0.83 that both tests obtain in this combination of factors. But even with this result, we find that both standard ANOVA and permutation ANOVA tests are relatively robust when group variances are equal.

Next, let us examine the performances of these tests when samples sizes are equal:

```
# Retrieving values where sample sizes are equal
powers_equaln_df <- powers_df[powers_df$n == "Equal (=)", ]

# Removing row indexing
rownames(powers_equaln_df) <- NULL

# Renaming columns
colnames(powers_equaln_df) <- c("Difference in Group Mean",
                                "Difference in Group Variance",
                                "Difference in Sample Size",
                                "Standard ANOVA Power",
                                "Permutation Test Power")

# Rounding power values
powers_equaln_rounded <- powers_equaln_df
powers_equaln_rounded[] <- lapply(powers_equaln_rounded, function(x) if(is.numeric(x))
  round(x, 2) else x)

# Output
```

```
kable(powers_equaln_rounded, caption="Simulated Statistical Powers with Equal
      Sample Sizes")
```

Table 5: Simulated Statistical Powers with Equal Sample Sizes

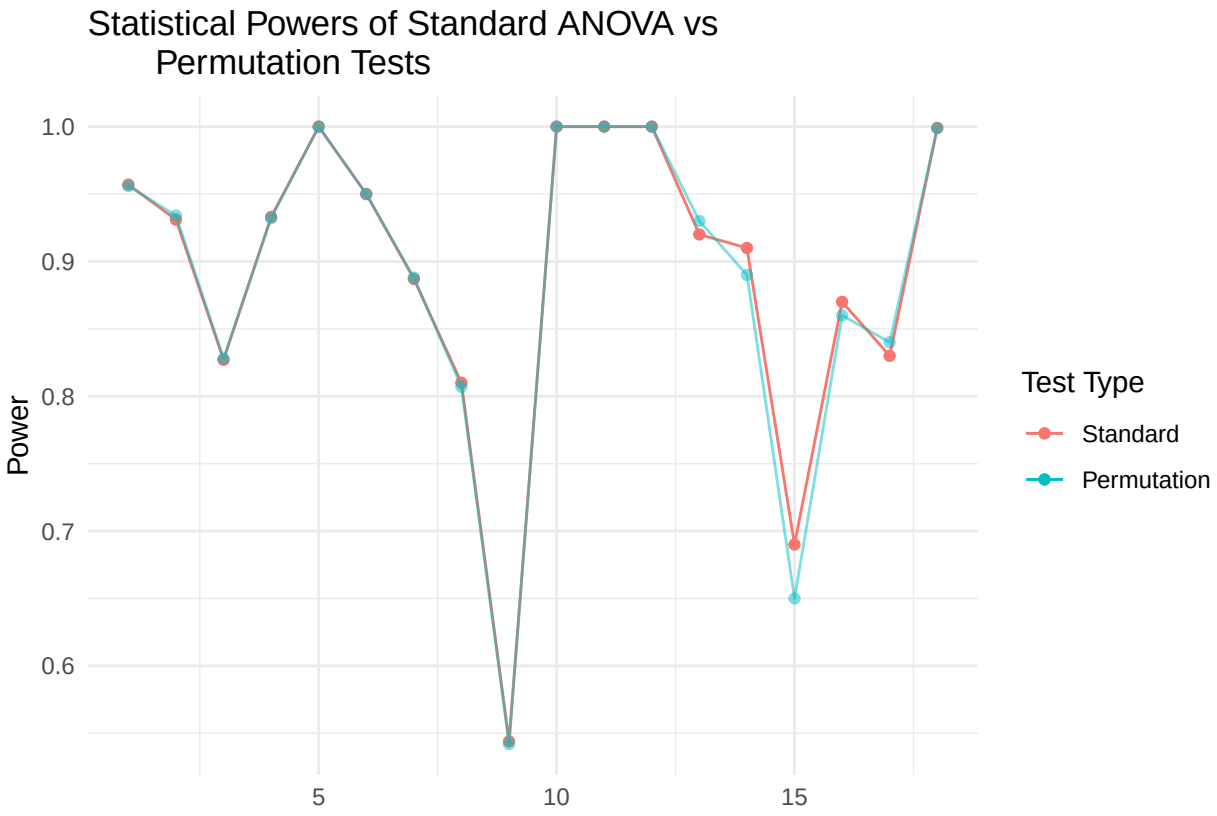
Difference in Group Mean	Difference in Group Variance	Difference in Sample Size	Standard ANOVA Power	Permutation Test Power
Low (-)	Equal (=)	Equal (=)	0.96	0.96
Low (-)	Low (-)	Equal (=)	0.93	0.93
Low (-)	High (+)	Equal (=)	0.89	0.89
High (+)	Equal (=)	Equal (=)	1.00	1.00
High (+)	Low (-)	Equal (=)	0.92	0.93
High (+)	High (+)	Equal (=)	0.87	0.86

The results above reveal that both tests perform well with any combination of both differences in group mean and group variance. It is worth noting that when differences in group variance is high, the statistical powers of both tests drop to a range of 0.86-0.89, or probabilities of 86-89% chance to correctly reject the null hypothesis. This is still optimal and indicates good performance, but may indicate that ANOVA may be still be sensitive to differences in group variance in balanced designed, contrary to claims in the introduction.

However, we still observe evidence to support the claim that ANOVA is robust to violations of homoscedasticity in balanced designs.

Finally, let us compare the statistical powers of standard ANOVA vs permutation ANOVA:

```
ggplot(powers_df) +
  geom_line(aes(x=1:length(aov_power), y=aov_power, color="blue")) +
  geom_point(aes(x=1:length(aov_power), y=aov_power, color="blue")) +
  geom_line(aes(x=1:length(perm_power), y=perm_power, color="red"), alpha=0.5) +
  geom_point(aes(x=1:length(perm_power), y=perm_power, color="red"), alpha=0.5) +
  labs(x="", y="Power", title="Statistical Powers of Standard ANOVA vs
       Permutation Tests", color="Test Type") +
  scale_color_discrete(labels = c("Standard", "Permutation")) +
  theme_minimal()
```



Note that the x-axis of the above chart has no ordinal value, but where the powers of each test line up correlate to a single simulation. Thus, the above line chart comparing the powers of both tests reveal that they are nearly identical for each experiment. This is contrary to what one might assume, as the permutation test is non-parametric and should be more robust when the assumptions of ANOVA are broken, as we have done in many of these simulations. There are two explanations for this which may be plausible:

1) The estimation of Type II Error was conducted with too many repetitions, eliminating any difference in performance

2) The Normality of data leads to similar performances in both tests

But if the results of this section are true, this would be a beneficial finding, as it would show that permutation ANOVA tests perform no better than standard ANOVA tests on Normal data. This would save a lot of time as the standard ANOVA test is a lot more computationally efficient, which would save lots of time in these statistical analyses.

Discussion

In conclusion, this study was conducted in order to examine the Type I Error Rates and statistical powers of both standard ANOVA and permutation ANOVA tests under different combinations of differences in group means, differences in group variances and differences in group sample sizes. Specifically, it has been found that standard ANOVA tests perform particularly poorly to violations of homoscedasticity when sample sizes are equal. However for permutation tests, since they are non-parametric, they should perform well regardless of these violations. By randomly simulating data and repeatedly conducting both of these tests, we were able to receive great insight on these claims.

Summary of Results

For each combination of differences in group variance and differences in sample size, we estimated the Type I Error Rates of both tests to be in the range of 0.01 to 0.08. This means that when group means were equal, either test falsely rejected the null hypothesis only 1-8% of the time. This revealed that both the standard ANOVA and permutation ANOVA tests performed extremely well when detecting equal means. While these findings seem contradictory to the claim that ANOVA tests do not perform well when there are violations to homoscedasticity in unbalanced designs, we were able to investigate more effectively by examining the statistical powers of these tests.

When conducting standard ANOVA and permutation ANOVA tests on data with true differences in means, it was first observed that when differences in group means were low but differences in both group variances and sample sizes were high, both tests had powers of 0.54, which were the lowest in this study. While this simulation confirmed the claim that ANOVA tests do not perform well when there are violations to homoscedasticity in unbalanced designs, we also obtained another unusual result. Under the same combination of factors but with a high difference in group means, both tests had powers of 1. This finding gave room for curiosity to the validity of this claim.

Otherwise, with exception to one simulation with skewed data, both tests had statistical powers in the range of 0.81-1, indicating that these tests are extremely robust. This includes group variances being equal and sample sizes being equal, proving the validity of these tests' performances

However, the final result of this study was particularly interesting. Here, we found that the statistical powers between standard ANOVA and permutation ANOVA tests were nearly identical throughout each simulation. While this may indicate a flaw in the study such as wrongly utilizing the Normal distribution or estimating power with too many simulations, this may also provide a beneficial result. If there is truly no difference in the standard ANOVA and permutation ANOVA tests, especially with Normal data, it would save a great deal of computational resources, no longer need to deal with the long runtimes of permutation tests.

Limitations

It is important to note that findings in this study are not causal and it may be wise to investigate further with more levels of each factor and with different distributions to solidify claims about ANOVA. For example, the choices of factors that represented "low" and "high" differences were extremely subjective and may not accurately represent what they are meant to represent. Additionally, three levels to each factor may not be enough to properly make conclusions, and it may follow that Error Rates may continually increase as differences increase, leading to more dramatic results than found in this study.

However, this study provides a baseline investigation into some of the claims made about ANOVA and hopefully challenges them in a valid way.

References

Vimal, Mourouguessine, Vinayagamoorthy Venugopal, and Nishanthi Anandabaskar. "Parametric Tests." Introduction to Basics of Pharmacology and Toxicology: Volume 3: Experimental Pharmacology: Research Methodology and Biostatistics. Singapore: Springer Nature Singapore, 2022. 877-888.

Blanca, María J., et al. "Effect of variance ratio on ANOVA robustness: Might 1.5 be the limit?." Behavior Research Methods 50 (2018): 937-962.

Sheng, Y. A. N. Y. A. N. "Testing the assumptions of analysis of variance." Best practices in quantitative methods (2008): 324-340.