PSTAT 122 Final: Simulation Study of ANOVA

Alexander Bloyer

2024-12-10

Introduction

The **An**alysis **O**f **Va**riance (`ANOVA`) is a powerful statistical test that is utilized across a variety of fields and applications that deal with data. The fundamental ability that `ANOVA` possesses is to determine whether the mean, or average, of two or more groups are equal. While this is a simple statement, the repercussions due its mathematically justified output allows for conclusions to be made that are highly statistically significant.

`ANOVA` was developed by Sir Ronald Aylmer Fisher, who proposed the term variance and its rigorous analysis in the 1918 article *The Correlation Between Relatives on the Supposition of Mendelian Inheritance*. A subsequent application of the analysis of variance came three years later in the 1921 publication *Studies in Crop Variation I*. In these works, Fisher demonstrated that there was a permanent and significant advantage to using highly nitrogenous dressings in the Broadbalk Wheat fields over the course of 67 years by using the features of the comparison of mean yields across the crops and plots of land (Fisher, 1921). A consequential observation that Fisher drew from the data collected over 67 years across a variety of harvests was that the mean wheat yield from different fields or across different seasons could not approach, in accuracy, the comparison of the mean wheat yield across the same field and season (Fisher, 1921). From this observation, he formally recognized the limitations of `ANOVA` in practice. These theoretical limitations are the mathematical assumptions of the `ANVOA` test which are (1) that the data be normally distributed, (2) homoscedasticity (i.e. homogeneity of variances), and (3) independence. It should be noted that `ANOVA` is relatively robust to violations of homoscedasticity, in the case of a balanced design. If the sample sizes across treatments are not equal, this introduces varying degrees of distrust in the results. Most notably, if the factor levels of the experiment with the larger variances also have smaller sample sizes, the type I error rate, α, that we observe will be larger than anticipated. Conversely, a smaller than expected type I error rate will be observed if the factor levels with larger variances also have the larger sample sizes (Montgomery, 2007, p. 72). Considering these as the primary "weakpoints/blindspots" of `ANOVA`, it follows that discussion and experimentation on the degree of unreliability of the `ANOVA` under the violation of its assumptions is warranted.

I will investigate the implications of modeling assumptions not being met on the statistical performance of `ANOVA` by comparing its results (the statistical Power of the test, $1-\beta$, or the type I error rate, α) to the corresponding permutation test. A permutation test is a statistical test that relies on a proof by contradiction. The logical contradiction occurs when we assume that data points in a certain group can be swapped to another group if we randomize the points by sampling without replacement. This preserves the overall structure of the data and sample size of each group, but allows us to test the equality of group means across a multitude of iterations.

The following simulation study will be conducted under the circumstances of three distinct factors—equality and inequality of group means, sample sizes, and variances—each of which having three levels. These factor levels are chosen to explore the aforementioned cases in which `ANOVA` is weak to the violation of its assumptions. Specifically, there are two primary circumstances of interest: the factor levels with the larger variance having either the smaller or larger sample sizes. Furthermore, the data that is generated to conduct the `ANOVA` and permutation tests are from a gamma distribution using `rgamma`, so these situations will also explore the violation of the normality assumption. The skewness of the gamma distribution will heavily violate the normality assumption which makes it a perfect candidate for a non-normal distribution.

Methods

To conduct a simulation study of the nature described in the introduction, several considerations must be made to ensure a proper analysis of `ANOVA` and the permutation test. It is vital to the integrity and significance of the results that the proper factor levels are chosen with consideration to the data distribution, and the previously outlined theoretical scenarios. Each factor combination will be structured as three groups of data such that group i has a mean, sample size, and variance of μ_i, n_i, and σ_i^2 respectively. Thus, we can mathematically define the factor levels as follows for the three groups contained in each simulation loop.

For the group mean, μ_i, the three factor levels are:

- equality ($\mu_1 = \mu_2 = \mu_3 = 1$)
- one different ($\mu_1 = 3, \mu_2 = 1, \mu_3 = 1$)
- all slightly different ($\mu_1 = 3, \mu_2 = 2, \mu_3 = 1$)
- all very different ($\mu_1 = 7, \mu_2 = 4, \mu_3 = 2$)

For the sample size, n_i, the three factor levels are:

- equality ($n_1 = n_2 = n_3 = 15$)
- a slight imbalance ($n_1 = 10, n_2 = 15, n_3 = 20$)
- a large imbalance forward ($n_1 = 5, n_2 = 20, n_3 = 50$)
- a large imbalance backward ($n_1 = 50, n_2 = 20, n_3 = 5$)

For the group variance, σ_i^2, the three factor levels are:

- equality ($\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$)
- slight variation ($\sigma_1^2 = 1, \sigma_2^2 = 2, \sigma_3^2 = 4$)
- large variation ($\sigma_1^2 = 1, \sigma_2^2 = 3, \sigma_3^2 = 9$)

The process of choosing these specific levels began in the precursory stages of this simulation study. Originally, the generated data followed a normal distribution with generally larger parameters (e.g. $20 \leq \mu_i \leq 50$, $5 \leq \sigma_i^2 \leq 15$, and $10 \leq n_i \leq 30$). Through several attempts of varying parameter values, the most significant results were observed by conducting simulations with smaller overall values of the mean and variances and a moderate disparity in sample sizes. However, the differences in the results obtained from an `ANOVA` and permutation test were largely negligible across all combinations of factors by using the larger values due to the assumption of normality being met. Thus, it seemed that smaller parameter values and a violation of the normality assumption would be useful in gathering significant results.

Since there are three factors of interest, two with four factor levels and one with three, it follows that $3 \cdot 4^2 = 48$ simulation loops are required to study each combination of factor levels. Throughout the combinations of factor levels, the combinations with equal group means will yield a result in the form of type I error rate because the null hypothesis is actually true. Conversely, the combinations with unequal group means (e.g., one different, all slightly different, all very different) will yield a result in the form of statistical power because the null hypothesis is actually false.

To provide context for the performance metrics of this simulation study, an explanation of the procedure of `ANOVA` and the permutation test are warranted. `ANOVA` can be broken down into several components, and to gain an intuition approach of the process, the null and alternative hypotheses must be established:

$$H_0 : \text{All group means are equal } (\mu_1 = \mu_2 = \mu_3).$$
$$H_A : \text{At least one of the group means differs.}$$

It should be noted that `ANOVA` only tests whether there is a difference across all the group means. Individual group differences can only be determined with post-hoc tests like Tukey's Test for pairwise comparisons of group means. This can be performed via `TukeyHSD` in `R`. Also, there are multiple types of `ANOVA`, but the simplest form is the One-Way `ANOVA` which examines one categorical variable with a certain number of variables. As mentioned beforehand, the key assumptions of `ANOVA` are:

- Independence: the observations within and across groups are independent.
- Normality: the distributions of residuals are approximately normal.
- Homoscedasticity: the variance across groups are approximately equal

Provided that these assumptions are met, `ANOVA` will return a p-value that can be used to draw conclusions about the equality of group means. The backbone of `ANOVA` is the ability to partition the total variability of the data into the variability of the residuals between and within groups. Mathematically, we can define the total sum of squares (SST), the sum of squares between groups (SSB), and the sum of squares within groups (SSW) as follows

$$SST = \sum_{i=1}^{N} (y_i - \bar{y})^2$$

$$SSB = \sum_{j=1}^{k} n_j (\bar{y}_j - \bar{y})^2$$

$$SSW = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$$

where N is the total sample size, y_i is the ith observation, k is the number of groups, n_j is the sample size of group j, $\bar{y}_j$ is the mean of group j, and $\bar{y}$ is the overall mean. The variability due to differences between group means is measured by SSB, and the variability within groups due to random noise is measured by SSW. It can be shown that the total sum of squares (SST) is equal to the sum of squares between groups (SSB) and the sum of squares within groups (SSW). In other words, $SST = SSB + SSW$. Using this result, the measure of variability due to differences between the groups can be quantified proportionally to the differences occuring by chance and used to extract a p-value. We begin by constructing a test statistic by using the ratio of the Mean Squared Error between groups and the Mean Squared Error within groups. These values are calculated by dividing the sum of squares terms by their respective degrees of freedom. The degrees of freedom for SST, SSB, and SSW are $n - 1$, $k - 1$, and $n - k$ respectively (note that $df_{SST} = df_{SSB} + df_{SSW}$). Thus,

$$MSB = \frac{SSB}{df_B} = \frac{SSB}{k-1}$$

$$MSW = \frac{SSW}{df_W} = \frac{SSW}{n-k}$$

It can be shown that these quantities will follow χ^2 distributions with their respective degrees of freedom (i.e., $MSB \sim \chi^2_{k-1}$ and $MSW \sim \chi^2_{n-k}$), and probability & statistics theory can be applied to show that we can construct a test statistic that follows an F distribution with:

$$F = \frac{MSB}{MSW} \sim F_{k-1,n-k}$$

Now that a test statistic with a known distribution has been obtained, the associated p-value can be calculated and compared to a significance level (typically $\alpha = 0.05$). Since this test statistic was built from the null hypothesis assumption that all of the group means are equal, the p-value will tell us whether this hypothesis should or should not be rejected.

While this procedure is effective at detecting a difference in group means, it is only statistically valid when the assumptions are met. In this simulation study, the primary objective is the examination of ANOVA under the violation of its assumptions to determine the impact on performance (as measured by power or type I error rate). Thus, we require a ulterior method that is “immune” to the same violation of assumptions as ANOVA. In simple terms, we seek to employ a test that will determine whether group means differ that does not require the same assumptions of ANOVA. Thus, a permutation test is desirable, and the specific function used in this simulation study was coded in R as the following:

```r
permutation.test <- function(df, reps){
  perm_F <- NA
  colnames(df) <- c("y", "x")

  for(i in 1:reps){
    df_perm <- df
    df_perm$y <- sample(df_perm$y)
    perm_F[i] <- summary(aov(y ~ x, data=df_perm))[[1]][1, 4]
  }

  F_0 <- summary(aov(y ~ x, data=df))[[1]][1, 4]

  return((sum(perm_F >= abs(F_0)) + sum(perm_F <= -abs(F_0))) / reps)
}
```

This function takes in two arguments: `df`, a data frame, and `reps`, the number of repetitions for the permutation test. The assumption that is required for a permutation test is the exchangeability of group labels, which allows for the randomization of data points across groups (Welch, 1987). The logic of the permutation test begins with a synonymous null hypothesis to that of ANOVA. By assuming that there is no difference in the group means across groups and thus that the individual values can be permuted across and among the groups randomly without changing the group means, a test statistic, F, can be obtained by running a standard ANOVA on the permuted data frames. By replicating this process many times and evaluating the proportion of F values that are as or more extreme than the value obtained from running an ANOVA on the original un-permuted data frame, we can calculate an associated p-value. Then, the process is repeated for a certain number of iterations, and the statistical power or type I error rate can be expressed as the proportion of p-values that are less than $\alpha = 0.05$ divided by the number of iterations.

The primary difference between ANOVA and the permutation test is that their required assumptions differ. The permutation test relies on the assumptions that the observations are exchangeable across groups, that the observations are independent, and that a sufficient number of repetitions are performed. Thus, the permutation test does not require the assumptions of normality and homoscedasticity that ANOVA requires.

The code below depicts an example simulation loop where the group means are equal, the sample sizes have a slight imbalance, and there is large variation.

```r
reps <- 10000
iters <- 1000

aov_pvals <- NA
perm_pvals <- NA

for(i in 1:iters){
  group1 <- rgamma(n = 3, shape = 1, scale = 1)
  group2 <- rgamma(n = 5, shape = 0.5, scale = 2)
  group3 <- rgamma(n = 7, shape = 0.25, scale = 4)

  data <- data.frame(
    y = c(group1, group2, group3),
    x = c(rep("group1", length(group1)),
          rep("group2", length(group2)),
          rep("group3", length(group3)))
  )
  aov_pvals[i] <- summary(aov(y ~ x, data = data))[[1]][1, 5]
  perm_pvals[i] <- permutation.test(df = data, reps)
}

result_aov <- sum(aov_pvals < 0.05) / iters
result_perm <- sum(perm_pvals < 0.05) / iters
```

When using normally distributed data, the implementation of the simulation loop for any of the factor combinations is trivial. The `rnorm` function has the parameters of sample size, mean, and standard deviation. Therefore, the factor level values can be input directly. For gamma distributed data, the `rgamma` function takes in the parameters of sample size, shape, rate, and scale where shape is α and scale is β. A gamma distribution is characterized by the shape (α) and scale (β) parameters since scale is equal to $\frac{1}{rate}$, so only one of scale or rate needs to be specified. The mean and variance of a gamma distribution are $\alpha\beta$ and $\alpha\beta^2$, respectively, so we can still generate groups of gamma data with a specific mean and variance after some brief algebra. For an arbitrary group of gamma distributed data that is generated using `rgamma`, we can obtain a specific mean and variance by considering the following relation. If $X \sim Gamma(\alpha, \beta)$, then $E[X] = \alpha\beta$ and $V(X) = \alpha\beta^2$.

Suppose that we would like to construct a simulation for-loop with data that follows the previously defined factor combination (equal group means, a slight imbalance in the sample size, and large variation) with the aforementioned factor levels. For an arbitrary group of gamma data with mean μ and variance σ^2,

$$\mu = \alpha\beta \Rightarrow \beta = \frac{\mu}{\alpha} \Rightarrow \frac{1}{\beta} = \frac{\alpha}{\mu}$$

$$\sigma^2 = \alpha\beta^2 \Rightarrow \alpha = \frac{\sigma^2}{\beta^2} = \sigma^2(\frac{\alpha}{\mu})^2 = \frac{\sigma^2\alpha^2}{\mu^2} \Rightarrow \alpha = \frac{\mu^2}{\sigma^2}$$

$$\beta = \frac{\mu}{\alpha} = \mu(\frac{\sigma^2}{\mu^2}) = \frac{\sigma^2}{\mu}$$

Thus, we can construct a group of gamma distributed data with a sample size of n mean of μ and a variance of σ^2 with the following parameters values,

$$\text{rgamma}(n, \alpha = \frac{\mu^2}{\sigma^2}, \beta = \frac{\sigma^2}{\mu})$$

This formula can be applied to the example simulation loop where we require a group of gamma distributed data with a sample size of 7, a mean of 1 and a variance of 4. By plugging in the values, we observe that

$$\alpha = \frac{1^2}{4} = \frac{1}{4}$$

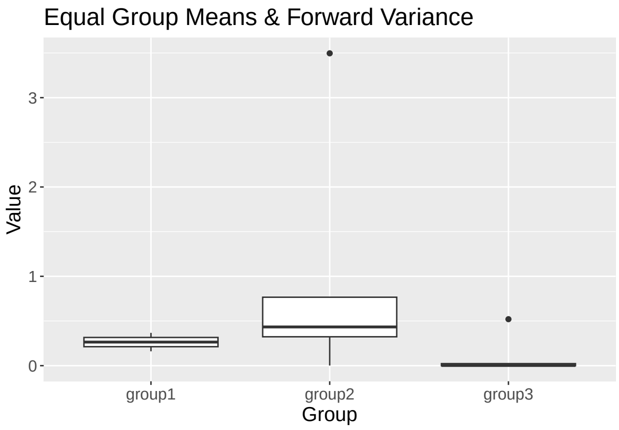
$$\beta = \frac{4}{1} = 4$$

This process can be repeated to construct the necessary groups of data for each factor combination across the simulation study loops.

Results

Before delving into the power and type I error rate values across the factor combinations, an examination of the simulated data is worthwhile. Since data was generated according to every possible combination of factor level, there are 48 uniquely defined circumstances, each of which generates three groups of data. Beyond this, 1000 iterations of each factor combination are executed to provide stability to the values of power and type I error rate. Therefore, this simulation study generates $48 \cdot 3 \cdot 1,000 = 144,000$ total groups of gamma distributed data with varying parameters. Considering the primary objective of the effects of assumption violation of `ANOVA`, it seems logical to explore examples of the data that are responsible for violating the assumptions.

The following two graphs visualize gamma data according to distributions with equal means and a large difference in variance. The primary reason for the difference in the two graphs, beyond the random nature of data generation, is the difference in sample sizes across the groups. The groups of the left graph have sample sizes that are approximately proportional to their respective variances (i.e., $n_1 = 2, n_2 = 5, n_3 = 8$ and $\sigma_1^2 = 1, \sigma_2^2 = 3, \sigma_3^2 = 9$). Conversely, the groups of the right graph have sample sizes are are approximately inversely proportional to their respective variances (i.e., $n_1 = 8, n_2 = 5, n_3 = 2$ and $\sigma_1^2 = 1, \sigma_2^2 = 3, \sigma_3^2 = 9$). Thus, these graphs show the difference in the smallest sample having the smallest and largest variance.



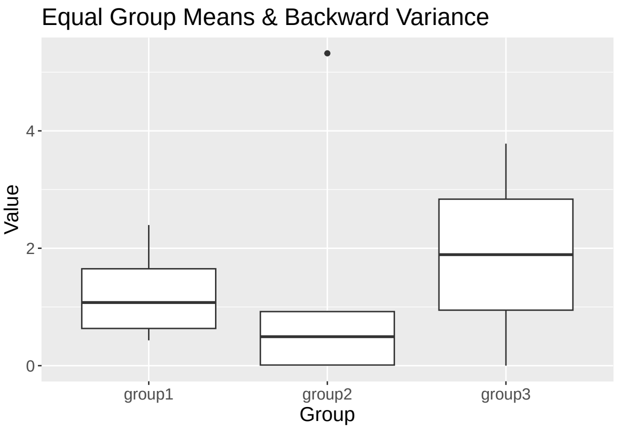


Table 1: Left Graph Summary Statistics

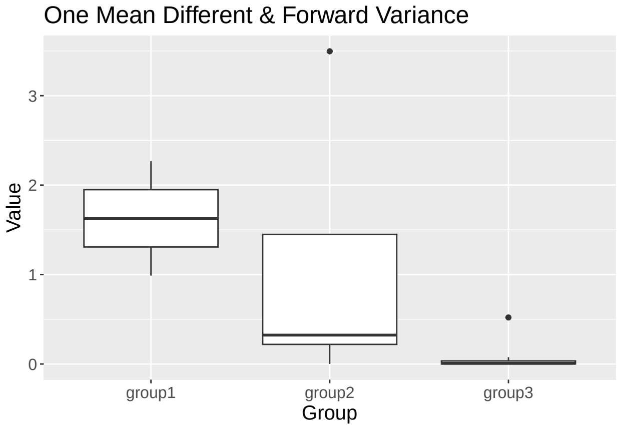
Group	Sample Size	Mean	Variance
group1	2	0.2639138	0.0213097
group2	5	1.0040551	2.0157127
group3	8	0.0717564	0.0329582

Table 2: Right Graph Summary Statistics

Group	Sample Size	Mean	Variance
group1	8	1.194447	0.4539365
group2	5	1.350634	5.0715391
group3	2	1.891094	7.1524707

While the above graphs violate the assumptions of normality and homoscedasticity required for `ANOVA`, the theoretical group means do not differ, so the null hypothesis is in fact true. Therefore, the performance metric of interest that we will measure from this data and other iterations is the type I error rate, α.

The following graphs explore data where one group mean differs and there is a large difference in variance across groups. The left and right graphs have the same samples size and variance combinations as the preceding graphs, but the theoretical means have been adjusted so that one differs (i.e., $\mu_1 = 3, \mu_2 = \mu_3 = 1$). Since the means differ, the null hypothesis is in fact false, so this data will ultimately measure statistical power. However, since only one group mean differs, the measurement of power will likely be less significant than if the difference in means were more dramatic.


One Mean Different & Forward Variance

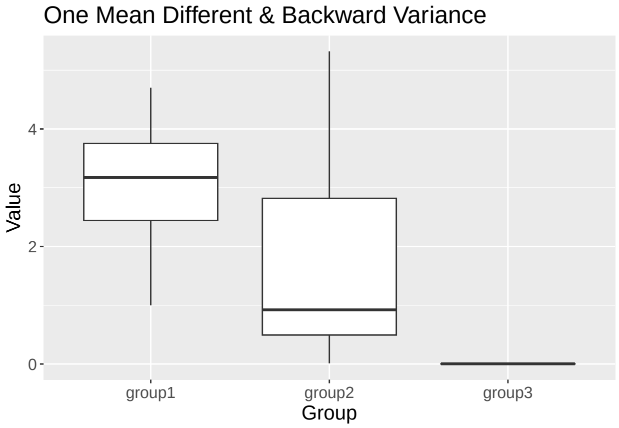

One Mean Different & Backward Variance

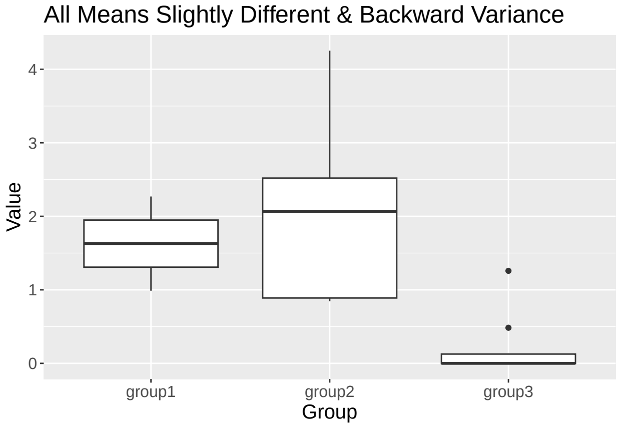
Table 3: Left Graph Summary Statistics

Group	Sample Size	Mean	Variance
group1	2	1.6287193	0.8208527
group2	5	1.0979155	2.1125487
group3	8	0.0793649	0.0323758

Table 4: Right Graph Summary Statistics

Group	Sample Size	Mean	Variance
group1	8	3.0063604	1.5309445
group2	5	1.9130697	4.7640137
group3	2	0.0029437	0.0000173

The following graphs explore the same situation with variance and sample sizes but with slightly different group means across each group. Specifically, the data is generated such that $\mu_1 = 3$, $\mu_2 = 2$, and $\mu_3 = 1$. The slightly more apparent difference in group means is likely to yield higher values of power for the `ANOVA` and permutation test.


All Means Slightly Different & Backward Variance

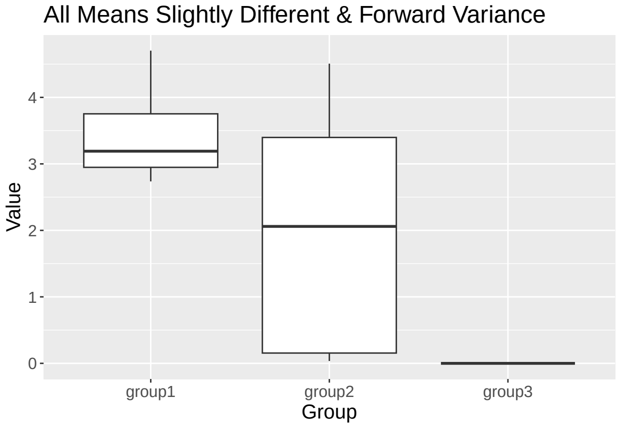

All Means Slightly Different & Forward Variance

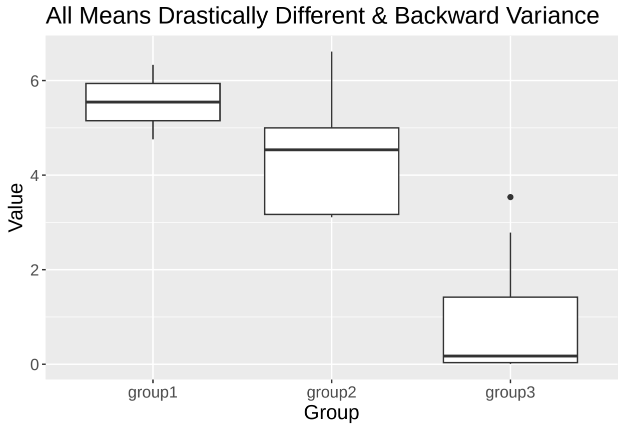
Table 5: Left Graph Summary Statistics

Group	Sample Size	Mean	Variance
group1	2	1.6287193	0.8208527
group2	5	2.1144291	1.9636900
group3	8	0.2187212	0.2051037

Table 6: Right Graph Summary Statistics

Group	Sample Size	Mean	Variance
group1	8	3.4200131	0.4645768
group2	5	2.0307351	3.8753609
group3	2	0.0000192	0.0000000

These two graphs deal with the most extreme case of data: the group means, sample sizes, and variances all differ drastically. The groups of data have theoretical means of $\mu_1 = 7$, $\mu_2 = 4$, and $\mu_3 = 2$ with the same configuration of the left graph depicting a group with the largest variance and sample size and the right graph depicting a group with the largest variance and smallest sample size. From this data, we would expect a high power from the permutation test and a lower paower from `ANOVA` because its assumptions are significantly violated.



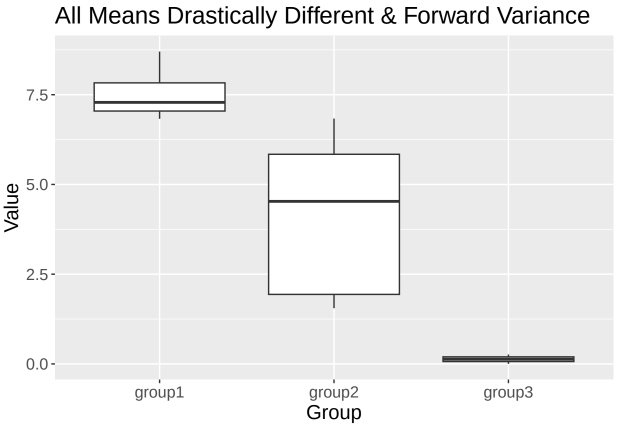


Table 7: Left Graph Summary Statistics

Group	Sample Size	Mean	Variance
group1	2	5.5447097	1.241590
group2	5	4.4850567	2.105775
group3	8	0.9616452	1.978037

Table 8: Right Graph Summary Statistics

Group	Sample Size	Mean	Variance
group1	8	7.4975468	0.4231988
group2	5	4.1388205	5.4651339
group3	2	0.1315701	0.0344480

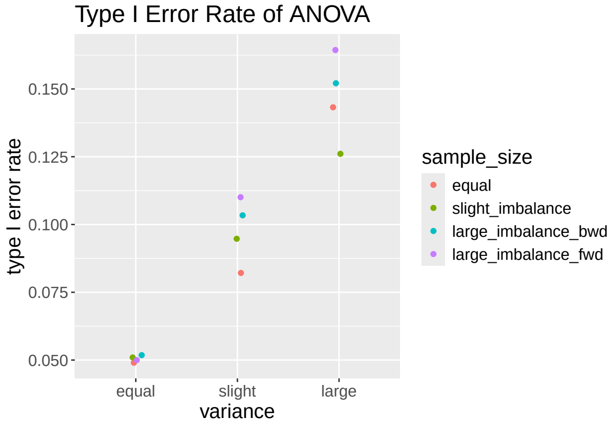
Table 9: Type I Error & Power Results Across Factor Combinations

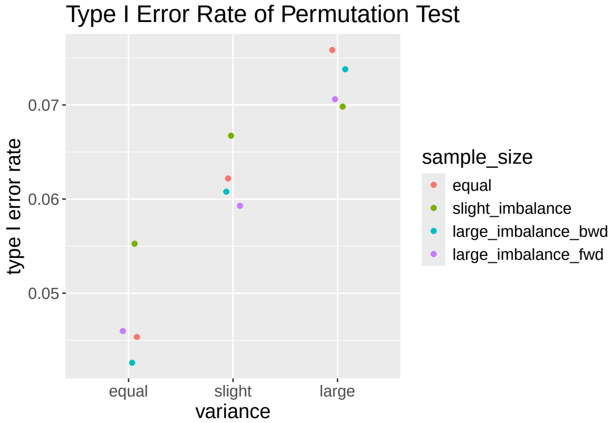
metric	aov	perm	mean	sample_size	variance
type_I_error	0.049	0.045	equal	equal	equal
type_I_error	0.082	0.062	equal	equal	slight_variation
type_I_error	0.143	0.076	equal	equal	large_variation
type_I_error	0.051	0.055	equal	slight_imbalance	equal
type_I_error	0.095	0.067	equal	slight_imbalance	slight_variation
type_I_error	0.126	0.070	equal	slight_imbalance	large_variation
type_I_error	0.050	0.046	equal	large_imbalance_fwd	equal
type_I_error	0.110	0.059	equal	large_imbalance_fwd	slight_variation
type_I_error	0.164	0.071	equal	large_imbalance_fwd	large_variation
type_I_error	0.052	0.043	equal	large_imbalance_bwd	equal
type_I_error	0.103	0.061	equal	large_imbalance_bwd	slight_variation
type_I_error	0.152	0.074	equal	large_imbalance_bwd	large_variation
power	0.611	0.596	one_different	equal	equal
power	0.580	0.493	one_different	equal	slight_variation
power	0.551	0.591	one_different	equal	large_variation
power	0.580	0.660	one_different	slight_imbalance	equal
power	0.451	0.520	one_different	slight_imbalance	slight_variation
power	0.427	0.640	one_different	slight_imbalance	large_variation
power	0.532	0.665	one_different	large_imbalance_fwd	equal
power	0.494	0.545	one_different	large_imbalance_fwd	slight_variation
power	0.480	0.540	one_different	large_imbalance_fwd	large_variation
power	0.416	0.613	one_different	large_imbalance_bwd	equal
power	0.401	0.565	one_different	large_imbalance_bwd	slight_variation
power	0.360	0.530	one_different	large_imbalance_bwd	large_variation
power	0.521	0.630	all_slightly_different	equal	equal
power	0.510	0.585	all_slightly_different	equal	slight_variation
power	0.540	0.615	all_slightly_different	equal	large_variation
power	0.565	0.650	all_slightly_different	slight_imbalance	equal
power	0.594	0.635	all_slightly_different	slight_imbalance	slight_variation
power	0.511	0.525	all_slightly_different	slight_imbalance	large_variation
power	0.566	0.765	all_slightly_different	large_imbalance_fwd	equal
power	0.431	0.746	all_slightly_different	large_imbalance_fwd	slight_variation
power	0.459	0.691	all_slightly_different	large_imbalance_fwd	large_variation
power	0.657	0.940	all_slightly_different	large_imbalance_bwd	equal
power	0.614	0.908	all_slightly_different	large_imbalance_bwd	slight_variation
power	0.528	0.895	all_slightly_different	large_imbalance_bwd	large_variation
power	0.897	1.000	all_very_different	equal	equal
power	0.795	0.999	all_very_different	equal	slight_variation
power	0.775	0.996	all_very_different	equal	large_variation
power	0.895	1.000	all_very_different	slight_imbalance	equal
power	0.830	0.973	all_very_different	slight_imbalance	slight_variation
power	0.630	0.962	all_very_different	slight_imbalance	large_variation
power	0.932	1.000	all_very_different	large_imbalance_fwd	equal
power	0.805	0.971	all_very_different	large_imbalance_fwd	slight_variation
power	0.695	0.979	all_very_different	large_imbalance_fwd	large_variation
power	0.956	1.000	all_very_different	large_imbalance_bwd	equal
power	0.925	0.975	all_very_different	large_imbalance_bwd	slight_variation
power	0.897	0.921	all_very_different	large_imbalance_bwd	large_variation

To draw conclusions from these results, it would be useful to visualize the power and type I error rates.

To begin, we will focus on the type I error rates, as they are a measure of the probability that the null hypothesis is rejected when it is in fact true. Since the null hypothesis is true for the factor combinations with equal group means, these combinations produce a type I error rate as a result of the simulation loops.

Type I Error rate: Group means are theoretically equal

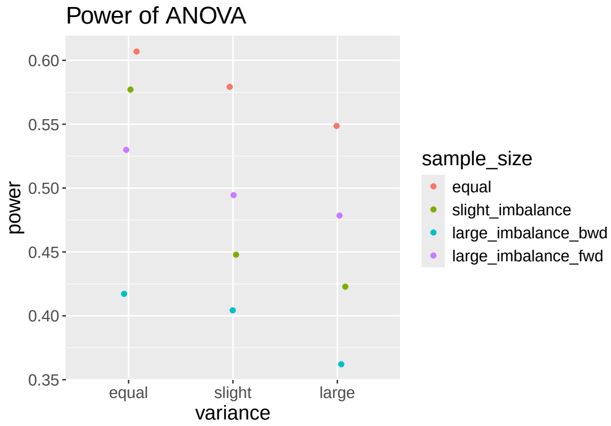


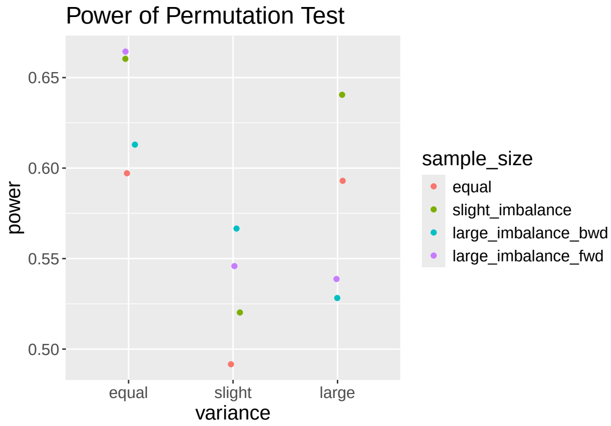


We observe that the values of type I error are relatively close to what is expected for the case of equal variance. In the `ANOVA` graph, the equal variance factor combinations has a type I error rate of approximately 0.05. As the variance increases, we observe both larger values and a larger spread of type I error rates. We observe that for a slight difference in variance, the type I error rate values range from 0.08-0.11 and for a large difference in variance the values range from 0.12-0.16. This aligns with our expectation because the assumption of homoscedasticity is becoming increasingly violated as the magnitude of difference in the variance increases. It is important to note that across the variance factor levels, the largest type I error values correspond to the largest imbalance in sample size.

Conversely, the permutation test graph reveals significantly lower values for all factor levels of variance. Also, there tends to be less variability of the rates compared to `ANOVA`. This aligns with our expectation because the permutation test is immune to the assumptions that `ANOVA` requires. However, there is a clear trend showing an increase in the type I error rates for the permutation test as the variance increases. Even though the test is immune to the violation of homoscedasticity (which is apparent in the lower overall rates compared to `ANOVA`), the increase in variance still manages to affect the values as there seems to be a gradual increase in type I error.

Power: One group means differs

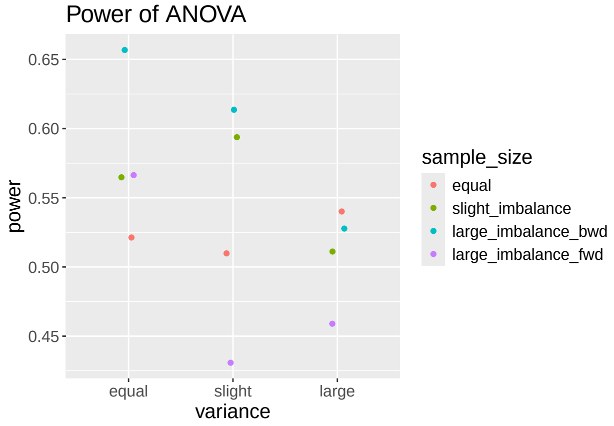


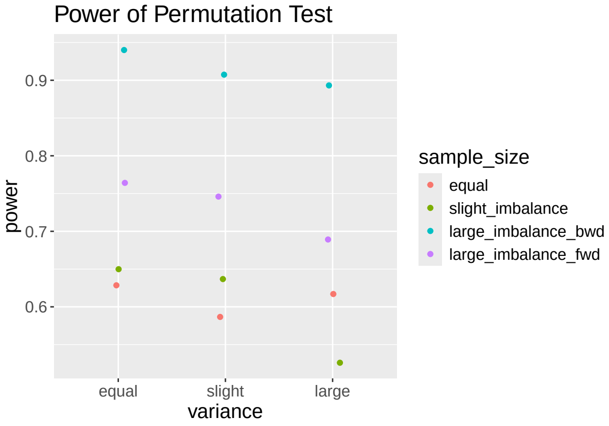


In the factor combinations with one different group mean, we observe that the power obtained from an `ANOVA` ranges from 0.4-0.6 with equal variance, 0.4-0.575 with a slight difference in group variances, and 0.35-0.55 with a large difference in group variances. These results track with our expectations of the power decreasing as the degree of violation of homoscedasticity increases. Beyond this, we observe a clear trend in the power based on the sample size factor level. Once again, we observe that in the case of an unbalanced design, `ANOVA` yields a lower power. There is a consistent spread of the values across each variance factor level. With equal sample sizes, we observe `ANOVA` perform "optimally" with the maximum observed power values.

The permutation test yields slightly more interesting results as there is less of a direct trend across all factors. Under equal variance, the power values are consistent and range from 0.6-0.67. However, we observe that a difference in variance does result in lower overall power values. For a slight difference in variance, the power ranges from 0.5-0.575, and for a large difference in variance, the power ranges from 0.375-0.525.

Power: All group means differ slightly

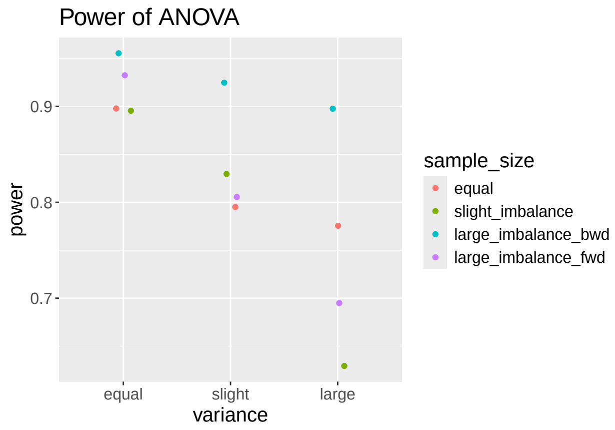


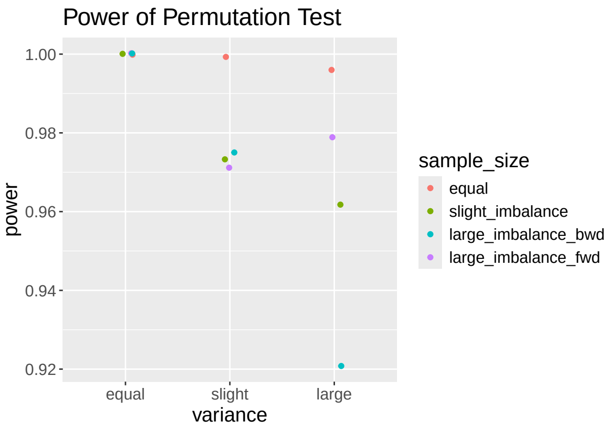


For the mean factor level of a slight difference across all groups, we observe a maximum power of 0.66 and a minimum power of 0.43. Additionally, the power observed with a large forward imbalance in sample size generally results in the largest values of power across the variance factor levels. We also observe a difference in the spread of the power values across the variance factor levels. For equal variance, there is a medium spread from 0.525-0.66, for a slight difference in variance, there is a large spread from 0.43-0.62, and for a large difference in variance, there is a small spread from 0.45-0.55. The violations in the assumptions of `ANOVA` clearly impact this factor combination due to the smaller power values as the variance increases.

From the permutation test, we observe clear trends across the variance and sample size factor. The spread of power for equal, slightly different, and largely different variance are 0.62-0.95, 0.58-0.9, and 0.53-0.9. Thus, we observe similar values of power across these levels with a slight drop in power as the variance increases. Also, we observe that the larger values of power are obtained with a large, forward imbalance in sample size. This is likely due to the fact that the data largest sample size with the largest variation has the largest impact on the overall results.

Power: All group means differ drastically





In the factor combination of drastically different group means, we observe stark trends in both the `ANOVA` and permutation test results. In the `ANOVA` graph, we observe high powers for equal variance ranging from 0.9-0.95. However, as the variance increases, we observe the same trend that was apparent in the previous graphs: the power decreases as variance among the groups increases. This can be observed with the factor level of a slight difference in the variance in which the power values range from 0.8-0.92. Also, for a large difference in variance, we observe a much larger range in power values from 0.63-0.9 with much more overall variability.

The permutation test graph reveals similar trends of increasing variability in the power values as the difference in variance increases. For equal variance, we observe values of power that are approximately 1, for a slight difference in variance, we observe values of power ranging from 0.97-1, and for a large difference in variance, we observe values of power ranging from 0.92-1. Thus, the permutation test performs with high power across all cases of the variance factor levels. It should be noted that the power obtained from groups with equal sample size are the highest, which is what is expected for the exchangeability logic of the permutation test.

Discussion

The previously graphed and tabulated values of type I error rate and power across factor levels effectively depict the impacts of the violation of assumptions on the ANOVA by using the permutation test as a frame of reference.

For the factor combinations that have theoretically equal means, we measure the performance of ANOVA and the permutation test through type I error rate. In the simulations with equal means, equal variances, and balanced sample sizes, we conclude that the ANOVA and permutation test yield equivalent results. From the previous graphs, we also conclude that as the sample size becomes unbalanced and as the difference in variance across the groups increases, the type I error rates from the ANOVA become inflated. Additionally, a similar, less intense trend can be observed with the permutation test type I error rates. Although the rates remain lower than the ANOVA values (due to the immunity to violations of normality and homoscedasticity), we do observe the permutation test values increase as variance and sample size increases. The largest type I error rate we observe from the permutation test is 0.076, while the maximum ANOVA type I error rates are nearly double, 0.164. The primary takeaway from the type I error rate data is that it inflates as the sample size becomes unbalanced and as the difference in variance increases.

For the factor combinations with theoretically unequal means, we measure the performance of the tests through statistical power. For the combinations with one different group mean, we observe power values of 0.35-0.6 from the ANOVA and 0.5-0.66 from the permutation test. For ANOVA, The largest power values are obtained with equal samples sizes across the groups, and the power decreases consistently as the sample sizes becomes slightly and significantly imbalanced. The same conclusion can be made of the effects of the increase in the difference in variance on the power. Thus, we observe an equivalent phenomenon of the performance metric worsening as the assumptions of ANOVA are violated. In the permutation test power values, we observe a similar trend with respect to the the variance, however, the presence of sample size imbalance seems to have less of an impact. This logically follows from the fact that the permutation test does not require normality of the data or homoscedasticity. For the combinations with all group means being slightly different, similar trends are observed for the decrease of power as the difference in variance increases. However, we interestingly observe a reversal of the impacts of the sample size imbalance. There is less of a distinct pattern the among the ANOVA values, and a stronger impact on the permutation test values. This can likely be attributed to the stronger overall difference in means, and the permutation test's ability to function reliably in circumstances of sample size imbalances. Finally, the factor combinations with drastically different group means portray the strongest impact of the violation of assumptions in the statistical power of ANOVA. The ANOVA power values range from 0.6-9.6, while the permutation test values range from 0.92-1. For the factor level of equal variance, we observe power values of 0.9-0.95 which is expected considering that the only assumption of ANOVA that is being violated in normality. However, a sharp decrease in power is apparent as the variance increases. Simultaneously, the power of the permutation test remains approximately consistent across the variance factor levels with equal sample size. In the circumstances of an imbalance in sample size, the power is slightly reduced but in the most extreme case, only to 0.92. Thus, we conclude that as the assumptions of ANOVA are systematically violated, its statistical power decreases steadily which when juxtaposed with the more consistent results of the permutation test demonstrate its ability to provide context for the detection of group mean difference.

The primary shortcoming of this simulation study is the lack of holistic assessment of the normality assumption. The results of an ANOVA have been shown to be impacted severely by skewness (Y. Sheng, 2008), which supports the choice to use gamma distributed data to explore the overall effects of the violation of assumptions. In future simulation studies, an examination of several other non-normal data with varying characteristics would be optimal to further explore the violation of the assumption of normality.

References

Fisher, R. A. (1919). XV.—The correlation between relatives on the supposition of Mendelian inheritance. Earth and Environmental Science Transactions of the Royal Society of Edinburgh, 52(2), 399-433.

Fisher, R. A. (1921). Studies in crop variation. I. An examination of the yield of dressed grain from Broadbalk. The Journal of Agricultural Science, 11(2), 107-135.

Montgomery, D. C. (2017). Design and analysis of experiments. John wiley & sons.

Welch, W. J. (1990). Construction of Permutation Tests. Journal of the American Statistical Association, 85(411), 693–698.

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