Final Project - Simulation Study of RCBD

Matthew Keeley

2025-02-25

1 Introduction

Everyday we make assumptions about the world around us, and everyday we either assume correctly or not. Sometimes our assumptions are small, while other times they are more serious. Performing statistical tests is no different – we make assumptions about the shape of our data and the process of data collection based on the type of test we choose to perform, and the validity of our assumptions can have significant impacts in the legitimacy of our outcomes (Garson 2012). Knowing what to do when these assumptions are violated is thus critical in legitimizing our results.

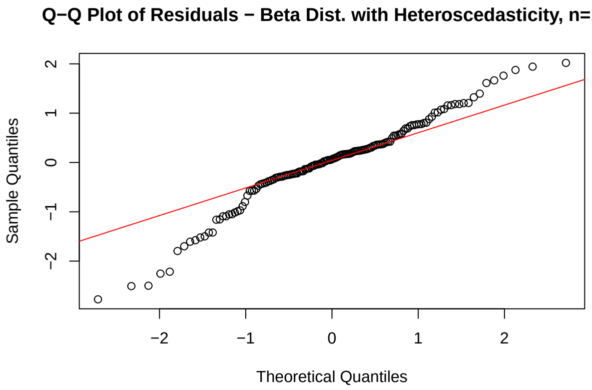
One type of statistical test that makes hypothesis testing more straightforward is ANOVA, or analysis of variance, which makes a variety of assumptions about the data (St, Wold, et al. 1989). Some key assumptions are normality, no structure to the data, and equal variances – when these assumptions are violated, the efficacy of ANOVA testing begins to fade, and other tests may begin to take precedence (Keselman et al. 1998). The conditions under which ANOVA experiences performance issues has seen extensive research over the years, and the consequences of violating assumptions is well-known within the field of statistical testing (Sheng 2008). However, the extent to which assumptions are violated is vital in understanding the significance of these violations, and testing assumptions is often overlooked in empirical studies, in large part because small deviations in assumptions can often yield acceptable results under parametric tests (Glass, Peckham, and Sanders 1972).

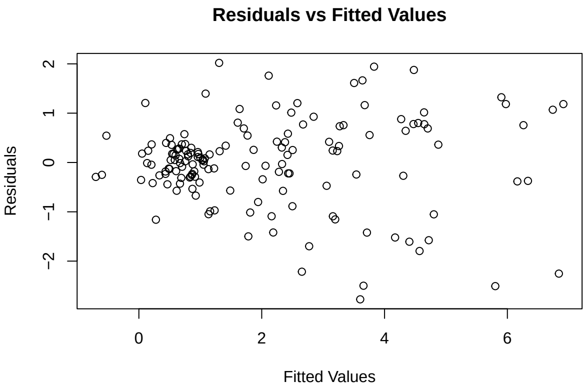
Performing statistical testing when the assumptions of ANOVA are violated is an important empirical question to help perform accurate statistical testing under circumstances where our assumptions do not hold, and permutation testing is a robust approach to getting results with higher power under certain circumstances. Knowing the specifics of these circumstances can help researchers decide which statistical test to perform for their precise use case, and get more powerful results from the same data (Hoekstra, Kiers, and Johnson 2012).

In this paper, the question that I aim to answer is under which scenarios permutation testing outperforms ANOVA when the assumptions are broken, and to what extent these violations must occur. Since permutation tests are non-parametric and make fewer assumptions about the shape of data, it is expected be more robust than ANOVA when the assumptions of ANOVA are violated – as long as the independence and exchangeability assumptions hold, permutation testing should perform fairly well, which will be explored in more detail in this paper (Pesarin and Salmaso 2010).

2 Methods

In order to explore the circumstances under which permutation tests and ANOVA tests performed and failed to perform, data was simulated to violate the normality assumption of ANOVA with 3 types of distributions – exponential, uniform, and beta distributions. Under the beta distribution, homoscedasticity is also violated through adding unequal standard deviations by scaling up observations with different multiples relative to their blocking group, which led to clustering around certain variances based on the scaling factors. This violation can be seen graphically below. I chose these methods by performing smaller models with fewer iterations in order to detect which factors appeared to have the most significant effects with regard to performance, and which appeared to be less close in my small simulations. These simulations are excluded from my results, but they all appeared to have fairly small impacts on the performance of the permutation test relative to the ANOVA test.





The QQ-plot and residual vs fitted values plots above help interpret the shape of the beta distribution with heteroscedasticity for n=150. As can be seen clearly in the QQ-plot, there are noticeable and significant deviations from normal, as expected. This helps for violating the normality assumption and analyzing the performance of ANOVA under such a violation. Furthermore, the residuals vs fitted values plot shows clustering, suggesting that the homoscedasticity assumption was successfully violated, allowing us to analyze the performance of ANOVA under both normality and homoscedasticity violations.

When model testing was performed on the data for each distribution, namely QQ-plot testing and the Shapiro-Wilk test, the assumption of normality for ANOVA appeared to be met – the P-values for all the Shapiro-Wilk tests were not significant, suggesting that there was insufficient evidence to reject the null hypothesis that the distribution of the data was not normal (even though none of the distributions were actually normal). This is likely due to each sample of data being generated fairly few times, with n=15 in each case as per the guidelines. I noticed that when I increased n to 150 and 1500, the Shapiro-Wilk test approached and surpassed the threshold of alpha = 0.05 for statistical significance, suggesting that more data was needed to confirm that the original data was not in fact normal. As might be expected, with few iterations, the ANOVA and permutation tests performed similarly to one another – it was only once the number of iterations was increased that some patterns started to surface, which helped determine the choice of distributions and situations to explore.

One curious situation was under the uniform distribution, where the permutation test seemed to perform worse than ANOVA – this distribution was thus chosen to explore whether this was due to too few iterations, or some other underlying factor. Effect sizes were chosen through similar techniques as the distributions, since the specific distribution was important for the impact of the effect size – the ranges and means of distributions were designed to be somewhat close (but still different) to one another so that effect sizes could have similar effects on each of them and to glean whether slight differences had large impacts on test performance. Effect sizes that were too large seemed to result in the ANOVA and permutation tests having indiscernible differences, likely because the correct answer was easier to detect for both, and so effect sizes were restrained from 0 to .75, at .25 increments. Effect sizes were implemented by creating an implicit meaning between treatment levels of 1, 2, and 3 as per guidelines, and then multiplying that treatment level by the effect size to generate the desired result.

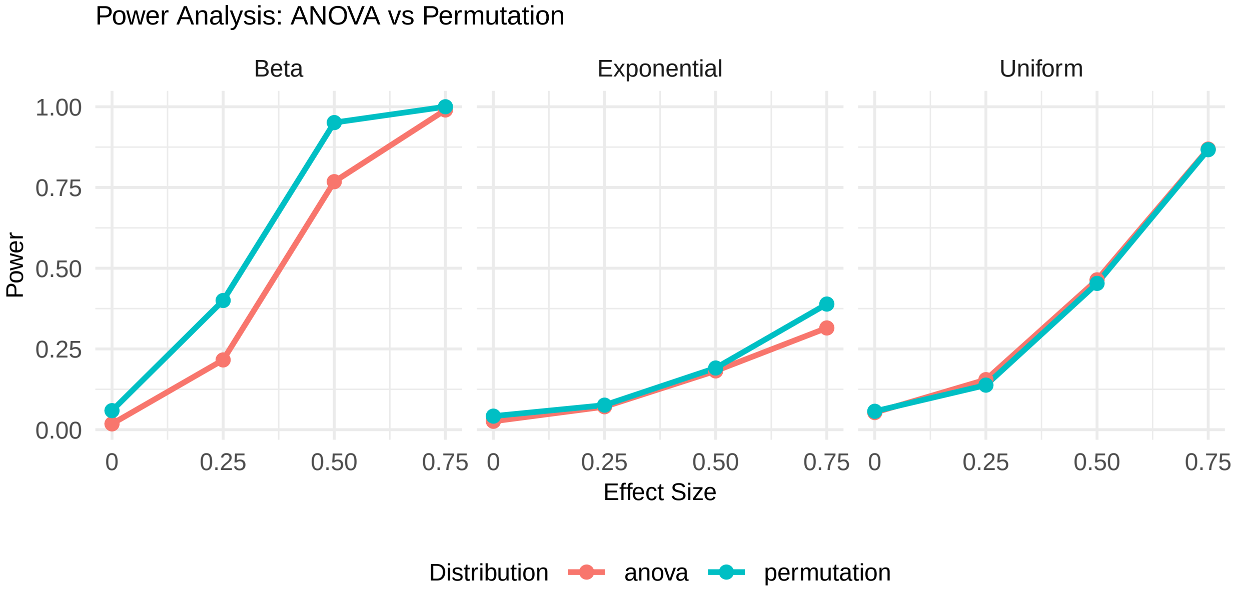
The RCBD controls for variation between blocks by making sure each block is treated separately, and assigning each treatment to each block exactly once. The associated ANOVA then tests the effects of treatment by accounting for block-to-block variation, isolating the treatment effect. The permutation test works by randomly re-arranging values within blocks to maintain the blocking, but changing observation values between treatments to test the null hypothesis. This works because if the null hypothesis is true, which is that the treatment has no effect on the mean value of the observation, then changing every value is just as likely to have occurred under each treatment. As such, if the null hypothesis is true, the permutation test should theoretically not change the outcome. Then, the permutation test is iterated many times, and each iteration's F-value is compared to the original data's F-value – every time the permutation test results in an F-value at least as extreme as the original F-value is summed, and the total number of instances is divided by the total number of iterations to get the permutation test P-value.

The assumptions of a RCBD are that observations within blocks are homogeneous, treatments are randomly assigned, and error terms are independent to one another. For the simulated data, this should be the case, since data within blocks was not manipulated to vary heterogeneously, and all treatments were randomly assigned. Errors between observations were also randomized and thus independent. For ANOVA, the key assumptions are independence between observations in the structure of data, normality of the data, and equal variances. Violations to these assumptions is the key component of this paper, as detailed above - normality is violated in all 3 distributions to varying extents, and homoscedasticity is violated in the beta distribution. Permutation tests are non-parametric, which requires fewer assumptions compared to parametric tests such as ANOVA - as long as observations are exchangeable under the null hypothesis, meaning if the null hypothesis is true, then any permutation should be just as likely to have occurred as the original, and the permutation test should be robust to the violations in ANOVA.

Some technical difficulties encountered while performing this simulation relate to the simulation taking nearly 48 hours to run (continuously), which meant keeping my laptop running continuously. This involved a lot of charging and running overnight, all while I was biking to the library to study for other classes and moving around a lot. My first obstacle was keeping the code running while my laptop was closed, so that I could safely bike to and from the library without having to carry my computer. To solve this, I changed a variety of settings – my computer will not sleep while closed, will not sleep while plugged in, and will not sleep while open and not plugged in. R will continue to run in the background, I will not allocate more memory to other programs to prevent R from terminating, to name the most glaring issues. I also made my simulation save every so often so that if somehow my code were to stop, I could at least resume from some recent iteration in order to maintain most of my progress without having to restart the 48 hour simulation over from scratch. Fortunately, I never had to implement this, as my other approaches proved sufficient. I also encountered obstacles with the code running slowly while my laptop was closed, so I kept it open as much as possible to speed up the process – overnight, at the library while I was taking a practice test for another class, and at home while eating breakfast. Thanks to all these efforts, 48 hours later my code finished – "Progress: 100%", RStudio chimed out. I could finally rest easy, knowing that my simulation was complete and I could finally begin my analysis.

3 Results

The simulated data can be best understood from the graph below, where the power of the ANOVA and permutation tests are overlaid on each distribution, with the x-axis showing effect size and power on the y-axis. It is important to note that for each effect size of 0, the y-value is actually the probability of making at type 1 error and not the power of the test. This is because with an effect size of 0 there is no true difference in mean between treatments, and so any result rejecting the null is a Type 1 Error and not part of the power calculations.



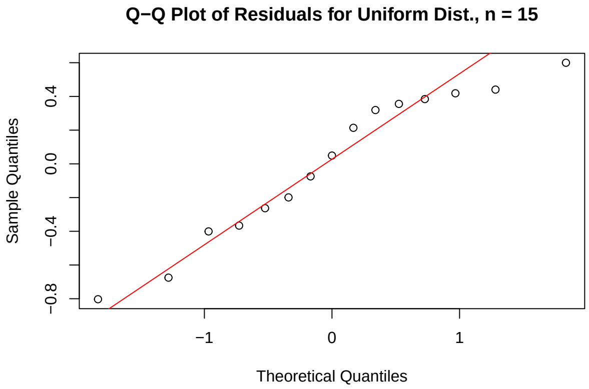
Graphically, the greatest improvement in the permutation test over ANOVA is in the beta distribution - whether this is solely due to the violation of homoscedasticity or a combination of that and the beta distribution being a greater violation of normality than the other distributions would require further testing, but it appears that the violation of homoscedasticity is a significant factor in the power of each test. Interestingly, the exponential distribution has an extremely low power at each effect size for both tests, though the permutation test appears to have an ever so slightly higher power graphically. Furthermore, the uniform distribution exhibits nearly identical powers for each test, though perhaps slightly higher power for ANOVA.

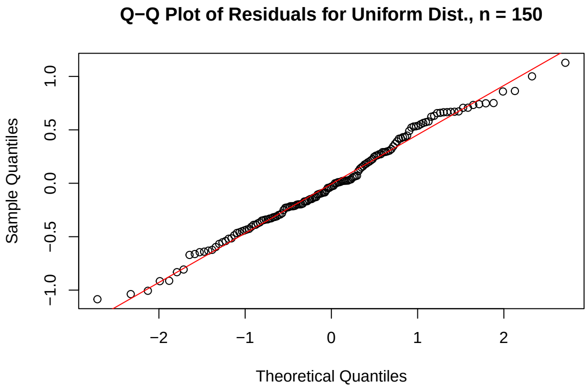
The figure shows the resulting powers and type 1 errors for each combination of factors. The power increases for each distribution with effect size, as expected due to there being a larger difference in means between treatments, and is thus easier to detect. Of note, the exponential distribution has less power than the other distributions. Furthermore, the difference between the ANOVA and permutation powers seems negligible in the uniform distribution at all effect sizes, even perhaps with a slight improvement in the ANOVA performance relative to the permutation test. It is important to note that for every distribution tested, ANOVA has a lower type 1 error rate.

Simulation Results

distribution	effect size	anova power	permutation power
Exponential	0.00	0.026	0.042
Uniform	0.00	0.053	0.057
Beta	0.00	0.018	0.059
Exponential	0.25	0.071	0.076
Uniform	0.25	0.155	0.138
Beta	0.25	0.216	0.400
Exponential	0.50	0.182	0.191
Uniform	0.50	0.464	0.453
Beta	0.50	0.768	0.951
Exponential	0.75	0.315	0.389
Uniform	0.75	0.869	0.867
Beta	0.75	0.990	1.000

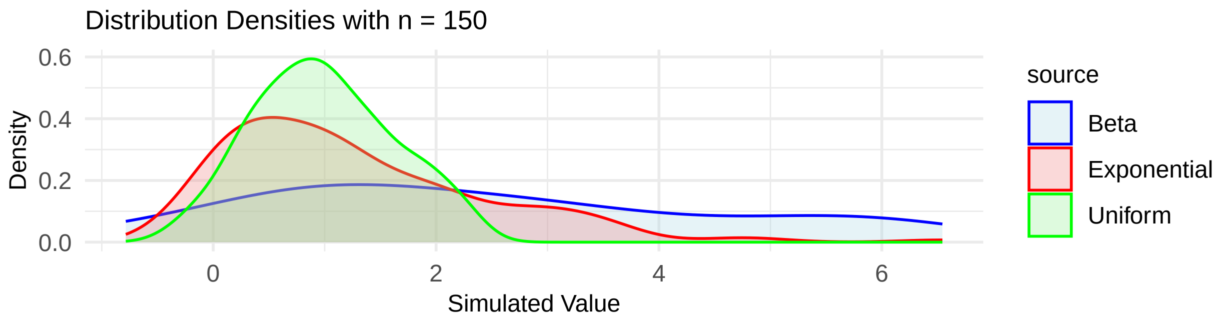
In order to investigate why the uniform distribution, while not normal, seems to work just as well with ANOVA as the permutation test (if not better), I perform an analysis of the assumptions of ANOVA on the uniform distribution as seen in the graphs below.





These Q-Q plots help us visualize the simulated data vs a normal distribution, shown as a red line. In both plots, there appears to be a little deviation at the tails, but the bulk of the data seems to follow the normal distribution quite closely. A simulation with 1000 iterations of the Shapiro-Wilk test is performed to test whether the difference between simulated data and normal data is statistically significant at n=15 and n=150. The results of the uniform distribution with n=15 show that the proportion of statistically significantly non-normal results is just 0.031, while n=150 has a proportion of 0.18 significantly non-normal results. For comparison, the same simulation under the beta distribution at n=150 has a proportion of 0.997 significantly non-normal iterations, a significant violation of normality. The subtlety of the violation in normality under the uniform distribution is likely a factor in why ANOVA performs well compared to the permutation test for the uniform distribution, since it is only a minor deviation from normality at the sample size that we have selected, n=15. For the sake of brevity, this explanation will suffice for explaining the poor performance of the permutation test relative to ANOVA for the uniform distribution. To test whether the exponential distribution's lower power level is for similar reasons, the Shapiro-Wilk simulation is run for the exponential distribution at n=150, and yields a high proportion of 0.966, suggesting that a violation in normality is strong in the exponential distribution at high sample sizes.

In order to further explore the low power of both tests on the exponential distribution, the shapes of each distribution are seen below, at n=150. The exponential distribution seems to follow a shape somewhat in the middle of the uniform and beta distributions, with a high concentration of observations between 0 and 2, but the exponential distribution has a right tail similar to the beta distribution that the uniform distribution lacks. This visualization suggests that the scale of the exponential distribution should not present an issue when performing statistical tests compared to the other distributions included. While further exploration for this would be appropriate in order to understand why there is lower power for the exponential distribution compared to other distributions, since there is minimal difference between the ANOVA and permutation test performance, it is not closely related to my research question and is left open to the reader to explore.



4 Discussion

In order to investigate the factors with the greatest impact on the performance of ANOVA, and to compare the results relative to the permutation test, a simulation of the beta, exponential, and uniform distributions was performed under a RCBD. The beta distribution was also sampled with unequal variances between blocks - in all, 5 blocks were each randomly sampled with 3 treatments. After data was simulated with n=15, various effect sizes were explored through early experimentation, and the effect sizes of 0, 0.25, 0.5, and 0.75 were chosen to explore further. This was done so because after an effect size of 0.75, most distributions already exhibited a very high power for both the ANOVA and permutation tests in the smaller simulations. Then, the permutation test was run with 10000 reps each, and the simulation for power was run with 1000 iterations for combination of factors, and the resulting powers for each combination were stored. Knowing when and why ANOVA fails to deliver on performance is critical to getting accurate conclusions from data, especially under small samples sizes like in this simulation, where n=15. Being able to work with small samples can help with preliminary experiments for many fields, and delivering on power while minimizing type 1 error is critical in yielding desirable results - the circumstances under which non-parametric tests such as the permutation test can rectify violations in assumption is thus critical to working with small samples that violate the assumptions of parametric counterparts, such as ANOVA.

The findings suggested that the beta distribution (with heteroscedasticity), violating both the normality and homoscedasticity assumptions, had the most discernible difference between the power of the permutation test and ANOVA. Notably, it failed to have a lower type 1 error, which is found from the "power" under a zero effect size, which was perhaps as expected based on the findings by Glass et al in 1972 (Glass, Peckham, and Sanders 1972). While the exponential distribution showed signs that the permutation test may perform better than ANOVA for that distribution, the evidence is narrow and would require further analysis. Analysis of the exponential distribution suggested that although it was notably different from the normal distribution under a Shapiro-Wilk test at n=150, it is likely that the sample at n=15 combined with the low overall power of testing on the exponential distribution make it more difficult to find a noticeable impact to the performance between the ANOVA and permutation tests. That being said, under small sample sizes, the differences between ANOVA and permutation tests proved smaller than expected. Under a uniform distribution, the ANOVA test outperformed the permutation test under every effect size, and even exhibited a smaller type 1 error - the violation of normality in this instance was small, and in many cases not statistically significantly different from the normal distribution as revealed under the Shapiro-Wilk test. This corroborates findings that small deviations from normality may not significantly reduce the power of ANOVA (Morgan 2017).

Since the simulation involved lengthy calculations, it took a very long time to run, and required leaving the code running overnight. Additionally, as this was done during and before finals week, it required closing my laptop to bike to and from study spots and final exams - this meant the code had to stay active even while the computer was closed. In order to resolve this, I set my computer to not sleep while closed, or under any other circumstances. With my computer running around the clock, I was ready to run my dauntingly long simulation. That being said, I also saved the code around every 5%, so theoretically if I encountered an issue with that, I should have been prepared to address it, but fortunately I did not have to rely on that. 40 hours (and many searches for wall outlets to charge my computer) later, my simulation was complete, and the analysis found above was performed.

References

Garson, G David. 2012. "Testing Statistical Assumptions." Statistical associates publishing Asheboro, NC.

Glass, Gene V, Percy D Peckham, and James R Sanders. 1972. "Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance." *Review of Educational Research* 42 (3): 237–88.

Hoekstra, Rink, Henk AL Kiers, and Addie Johnson. 2012. "Are Assumptions of Well-Known Statistical Techniques Checked, and Why (Not)?" *Frontiers in Psychology* 3: 137.

Keselman, Harvey J, Carl J Huberty, Lisa M Lix, Stephen Olejnik, Robert A Cribbie, Barbara Donahue, Rhonda K Kowalchuk, et al. 1998. "Statistical Practices of Educational Researchers: An Analysis of Their ANOVA, MANOVA, and ANCOVA Analyses." *Review of Educational Research* 68 (3): 350–86.

Morgan, Charity J. 2017. "Use of Proper Statistical Techniques for Research Studies with Small Samples." *American Journal of Physiology-Lung Cellular and Molecular Physiology* 313 (5): L873–77.

Pesarin, Fortunato, and Luigi Salmaso. 2010. "The Permutation Testing Approach: A Review." *Statistica* 70 (4): 481–509.

Sheng, YANYAN. 2008. "Testing the Assumptions of Analysis of Variance." *Best Practices in Quantitative Methods*, 324–40.

St, Lars, Svante Wold, et al. 1989. "Analysis of Variance (ANOVA)." *Chemometrics and Intelligent Laboratory Systems* 6 (4): 259–72.