# Simulation Study of the TukeyHSD Test

John Kimrey

## Introduction

The Tukey Honest Significant Difference (TukeyHSD) test is a widely used statistical procedure for performing pairwise comparisons among group means following an Analysis of Variance (ANOVA) (Fisher 1925; Tukey 1949). In many experimental and observational studies, researchers are interested not only in whether any group means differ (as tested by ANOVA), but also in which specific groups are different from each other. The TukeyHSD test addresses this need by controlling the family-wise error rate when making multiple comparisons, reducing the likelihood of false discoveries due to multiple testing (Tukey 1949; Westfall & Young 1993).

Despite its popularity, questions often arise regarding the true statistical properties of the TukeyHSD test, particularly its control of Type I error and its statistical power relative to the initial omnibus ANOVA test (Games & Howell 1976; Westfall & Young 1993). Standard practice is to first perform an ANOVA to detect any difference among groups and, if significant, follow up with the TukeyHSD test for pairwise group comparisons. This workflow is based on the premise that the TukeyHSD test, while providing strong control for the overall Type I error rate, may be more conservative and thus less powerful in detecting genuine differences between specific groups (Games & Howell 1976).

This report aims to systematically investigate the performance of the TukeyHSD test through simulation. Specifically, the study will address two key questions:

1. Type I Error Rate: Does the TukeyHSD test maintain an overall Type I error rate near the nominal level (e.g., alpha = 0.05) when ANOVA conditions are satisfied?
2. Statistical Power: How does the power of the TukeyHSD test compare to the overall ANOVA test in detecting true differences among groups?

By simulating data under a variety of conditions—including different numbers of groups, sample sizes, and effect sizes—this study seeks to provide an empirical understanding of when and how the TukeyHSD test meets its theoretical guarantees and where it may fall short compared to the omnibus ANOVA.

# Methods

This study uses Monte Carlo simulation to investigate the statistical properties of the Tukey Honest Significant Difference (TukeyHSD) test (Tukey 1949; Games & Howell 1976) in the context of one-way ANOVA (Fisher 1925). Specifically, we evaluate (1) the Type I error rate of TukeyHSD under the global null hypothesis and (2) its statistical power compared to the omnibus ANOVA F-test under alternatives with varying effect sizes (Westfall & Young 1993).

## Simulation Scenarios

For both aims, the following simulation scenarios were considered:

- **Number of groups:** 4, 5, or 6
- **Sample size per group:** 20
- **Group variance:** 1 (equal for all groups)

For Type I error simulations (Aim 1), all group means were set to 0 (i.e., no true group differences).
For power simulations (Aim 2), all but one group had a mean of 0, while the remaining group had a mean of 0.25, 0.5, or 1.0—representing small, medium, and large effect sizes, respectively.

## Simulation Procedure

For each scenario (combination of group count and effect size), the following steps were repeated for 10,000 iterations:

1. **Data Generation:**
   For each group, 20 independent samples were drawn from a normal distribution with the designated group mean and variance of 1.

2. **Statistical Testing:**

   - **ANOVA:** A one-way ANOVA was performed to test for any group differences (Fisher 1925).
   - **TukeyHSD:** The TukeyHSD test was applied post-hoc to all pairwise group comparisons (Tukey 1949; Games & Howell 1976).

3. **Outcome Recording:**

   - **Type I error (Aim 1):** A Type I error was recorded if at least one TukeyHSD pairwise p-value was less than 0.05 in the null case.
   - **Power (Aim 2):** Power was defined as the proportion of simulations in which a statistically significant result ($p < 0.05$) was observed in the ANOVA or in at least one TukeyHSD pairwise comparison under the alternative (Games & Howell 1976; Westfall & Young 1993).

## Data Analysis

For each scenario, the estimated Type I error rate or power was computed as the proportion of simulations in which the respective test yielded a significant result. Results were summarized in tables and plots to illustrate the dependence of Type I error and power on the number of groups and effect size (Westfall & Young 1993). Simulations took approximately 30 minutes per scenario on a standard laptop, and results were saved and loaded into RMarkdown to avoid computation time during knitting.

# Results

## Example of a Simulated Dataset

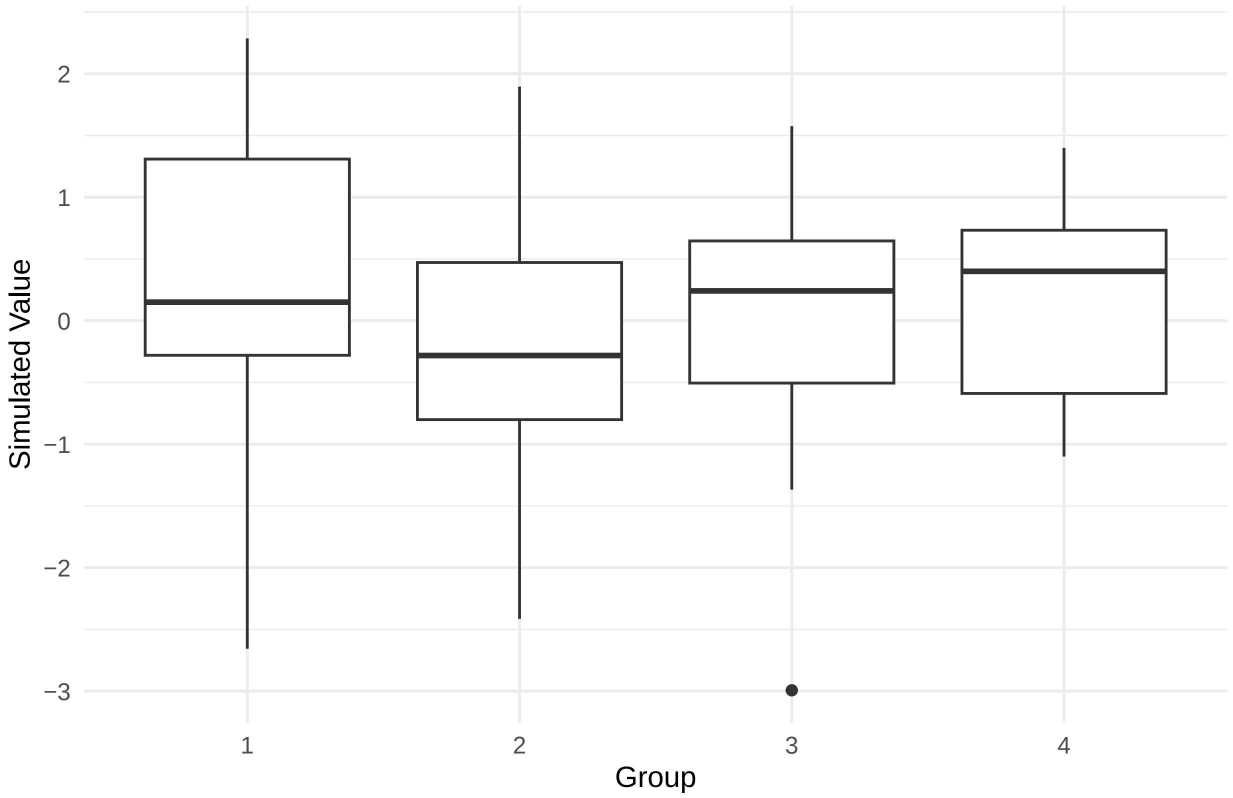
To provide context for the simulation results, Table 1 and Figure 1 below show a single simulated dataset generated for one scenario (e.g., 4 groups, sample size 20 per group, all means 0).

Table 1 below displays the first 10 observations from a single simulated dataset generated with 4 groups and 20 observations per group, with all group means set to zero. Figure 1 displays a boxplot of the entire simulated dataset, showing the distribution of values across groups. This provides a sense of the structure of the data underlying each simulation iteration.

Table 1: Sample of a single simulated dataset (first 10 observations)

| group | value |
|---|---:|
| 1 | 1.3709584 |
| 1 | -0.5646982 |
| 1 | 0.3631284 |
| 1 | 0.6328626 |
| 1 | 0.4042683 |
| 1 | -0.1061245 |
| 1 | 1.5115220 |
| 1 | -0.0946590 |
| 1 | 2.0184237 |
| 1 | -0.0627141 |

Figure 1: Boxplot of a Single Simulated Dataset (4 groups, n = 20 per group



### Model Checking for Example Simulated Dataset

To verify that the ANOVA model assumptions are satisfied for our simulated data, we examined diagnostic plots for a representative simulated dataset. Figure 2 shows a QQ plot of the ANOVA residuals, and Figure 3 shows a plot of residuals versus fitted values.

**Figure 2: QQ Plot of Residuals**

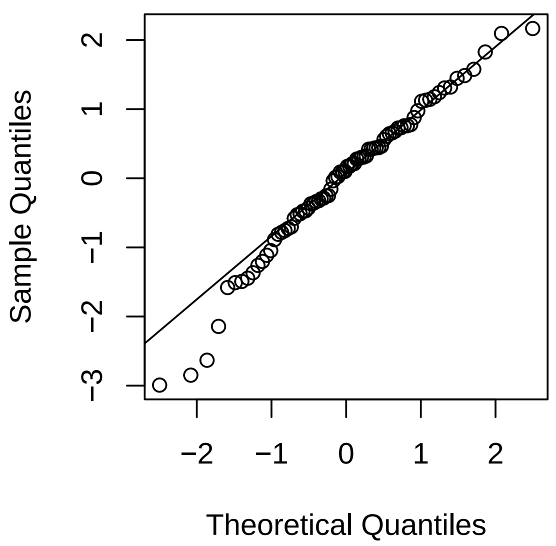
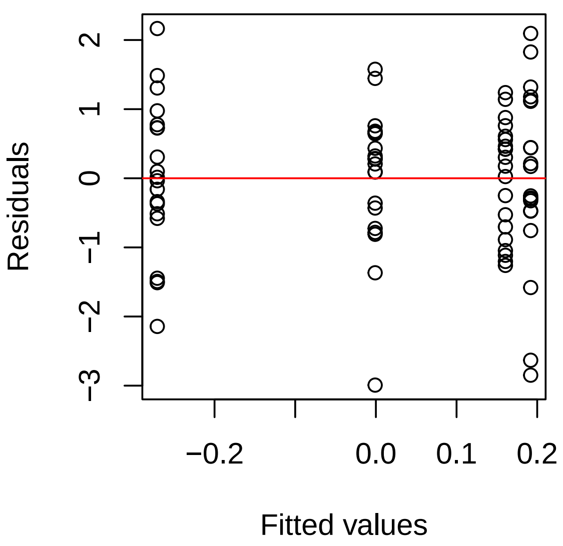


**Figure 3: Residuals vs Fitted Value**



As expected for data generated from a normal distribution with equal variance, the QQ plot indicates approximate normality of residuals, and the residuals vs fitted plot shows no systematic patterns. Thus, ANOVA assumptions are well-satisfied in these simulations, and no further adjustments or permutation testing were necessary.

### Simulation for Type I Error and Power

The code below loads the results of a Monte Carlo simulation to estimate the Type I error rate (when group means are equal) and power (when one group mean differs) for the TukeyHSD and ANOVA tests, varying the number of groups and effect sizes.

```
load("tukey_sim_results.RData")

# Display results table
kable(results, digits = 3, caption = "Estimated Type I Error and Power for TukeyHSD and ANOVA")
```

### Type I Error Rate (Aim 1)

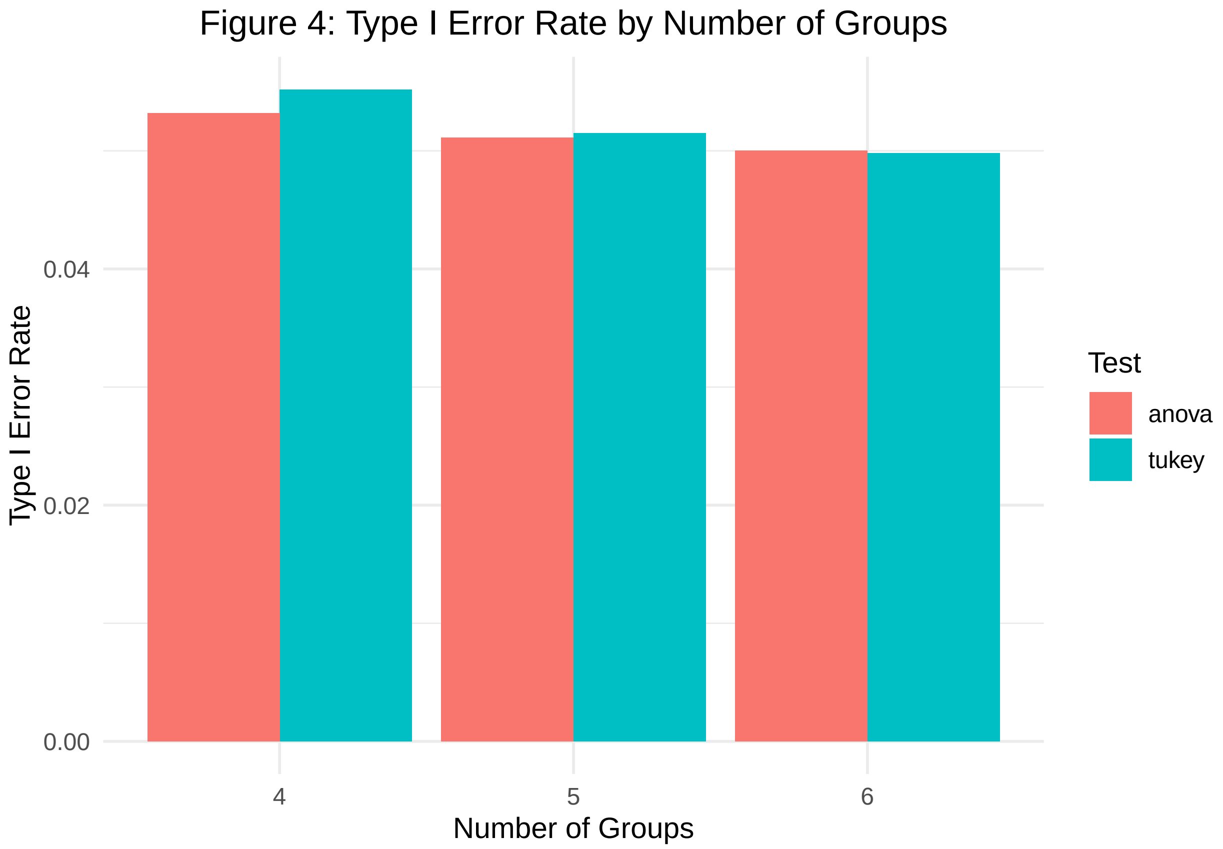
The estimated Type I error rates for the TukeyHSD and ANOVA tests, under the global null hypothesis (effect = 0), are shown below.

Table 2: Tukey and ANOVA Results at Effect = 0

| n_groups | effect | tukey | anova |
|---:|---:|---:|---:|
| 4 | 0 | 0.055 | 0.053 |
| 5 | 0 | 0.052 | 0.051 |
| 6 | 0 | 0.050 | 0.050 |

**Visualization**

Figure 4 displays the estimated Type I error rates for TukeyHSD and ANOVA across different numbers of groups, under the global null hypothesis.



**Interpretation:**

Both TukeyHSD and ANOVA maintain a Type I error rate very close to the nominal 0.05 level for all numbers of groups, confirming that the TukeyHSD test is appropriately controlling the family-wise error rate as expected under ANOVA assumptions.

## Statistical Power (Aim 2)

Estimated power for both tests as a function of effect size and number of groups:

Table 3: Tukey and ANOVA Results by Effect Size

| n_groups | effect | tukey | anova |
|---:|---:|---:|---:|
| 4 | 0.25 | 0.108 | 0.108 |
| 5 | 0.25 | 0.102 | 0.102 |
| 6 | 0.25 | 0.094 | 0.099 |
| 4 | 0.50 | 0.323 | 0.324 |
| 5 | 0.50 | 0.303 | 0.296 |
| 6 | 0.50 | 0.301 | 0.294 |
| 4 | 1.00 | 0.902 | 0.907 |
| 5 | 1.00 | 0.899 | 0.896 |
| 6 | 1.00 | 0.895 | 0.887 |

### Visualization

Figure 5 shows the estimated power for both TukeyHSD and ANOVA as a function of effect size, with separate curves for each number of groups.

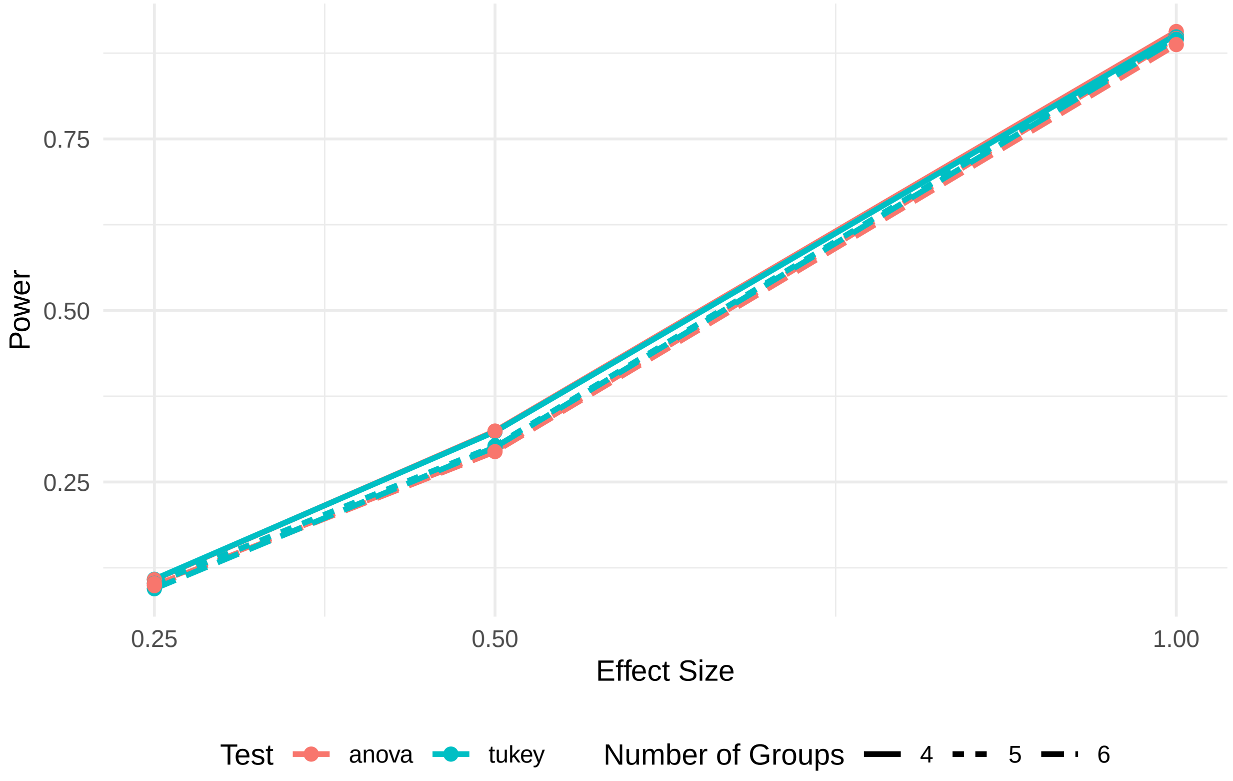


Figure 5: Power Curve for TukeyHSD and ANOVA

### Interpretation:

Power increases with effect size for both tests, as expected. Both TukeyHSD and ANOVA tests have very similar power in these scenarios. Power slightly decreases as the number of groups increases for small effect sizes, which is consistent with the increased multiple comparisons burden. For large effects (effect = 1), both tests have high power (>0.89) regardless of group count.

# Discussion

This simulation study provides empirical evidence that the TukeyHSD test, as a post-hoc procedure following ANOVA, maintains control of the overall Type I error rate across a range of group counts, in close agreement with the theoretical expectations and previous literature (Tukey 1949; Westfall & Young 1993). Both TukeyHSD and the omnibus ANOVA F-test demonstrate Type I error rates near the nominal 0.05 level under the null hypothesis, even as the number of groups increases (Fisher 1925; Games & Howell 1976). This result is reassuring for practitioners, confirming that TukeyHSD remains a reliable option for multiple comparisons when standard ANOVA assumptions are met (Westfall & Young 1993).

With respect to statistical power, the results show that both tests have nearly identical power profiles across a range of effect sizes and group counts, consistent with findings from prior simulation studies (Games & Howell 1976). Power increases with effect size, as expected, and for large effects, both methods detect true differences with high probability. Interestingly, while increasing the number of groups introduces more pairwise comparisons and a greater burden for family-wise error control, the reduction in power is minimal when using TukeyHSD compared to the omnibus ANOVA. This suggests that concerns about the conservativeness of TukeyHSD may be less severe than sometimes assumed, at least in balanced and ideal conditions (Tukey 1949; Westfall & Young 1993).

These findings support common practice: it is appropriate to use TukeyHSD as a follow-up to a significant ANOVA, with little concern for excessive loss of power relative to the overall F-test. The method's strong error control makes it especially valuable when identifying specific group differences is of scientific interest (Westfall & Young 1993).

A limitation of this study is the focus on idealized settings—balanced group sizes, equal variances, and data generated from normal distributions. In practice, violations of these assumptions are possible and may affect both Type I error rates and statistical power (Games & Howell 1976). Additionally, only scenarios with a single nonzero group mean were simulated; more complex patterns of group differences could yield different results. Exploring situations with unequal variances, non-normal distributions, or unbalanced group sizes would provide a fuller understanding of TukeyHSD's robustness in applied settings. Alternative post-hoc procedures, such as the Games-Howell test or resampling-based methods, could also be investigated (Games & Howell 1976; Westfall & Young 1993).

In summary, this study confirms the reliability of the TukeyHSD test for controlling family-wise error rate and retaining competitive power compared to ANOVA in the context of balanced, normally-distributed data. These results support the use of TukeyHSD in standard one-way ANOVA settings, while highlighting the need for further research in more challenging data scenarios.

# References

Fisher, R.A. (1925). Statistical methods for research workers. *Oliver and Boyd*, Edinburgh.

Games, P.A., & Howell, J.F. (1976). Pairwise multiple comparison procedures with unequal n's and/or variances: A Monte Carlo study. *Journal of Educational Statistics*, 1(2), 113-125.

Tukey, J.W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, 5(2), 99-114.

Westfall, P. H., & Young, S. S. (1993). Resampling-based multiple testing: Examples and methods for p-value adjustment. *Journal of the American Statistical Association*, 88(423), 881-889.

# Appendix

## Full Simulation Code

The following R script was used to generate all simulation results reported in this study. The code implements a Monte Carlo simulation for both Type I error and power, across all required combinations of group counts and effect sizes. This script was run outside the RMarkdown and the results were loaded into the Rmd via load(“tukey_sim_results.RData”). All further analysis, tables, and plots were created using the loaded results.

```r
# Set up scenarios
n_iter <- 10000
n_groups_set <- c(4, 5, 6)
effect_sizes <- c(0, 0.25, 0.5, 1)  # Need to include 0 for Type I error

# Pre-allocate results
results <- expand.grid(n_groups = n_groups_set, effect = effect_sizes)
results$tukey <- NA
results$anova <- NA

# Simulation function
run_sim <- function(n_groups, n_per_group, group_means, n_iter) {
  tukey_sig <- numeric(n_iter)
  anova_sig <- numeric(n_iter)
  for (i in 1:n_iter) {
    data <- data.frame(
      group = factor(rep(1:n_groups, each = n_per_group)),
      value = unlist(lapply(group_means, function(mu) rnorm(n_per_group, mu, 1)))
    )
    fit <- aov(value ~ group, data = data)
    anova_p <- summary(fit)[[1]][["Pr(>F)"]][1]
    anova_sig[i] <- (anova_p < 0.05)
    tukey_p <- TukeyHSD(fit)$group[,4]
    tukey_sig[i] <- any(tukey_p < 0.05)
  }
  c(tukey = mean(tukey_sig), anova = mean(anova_sig))
}

# Run all scenarios
for (i in 1:nrow(results)) {
  n_groups <- results$n_groups[i]
  effect <- results$effect[i]
  group_means <- c(rep(0, n_groups-1), effect)
  stats <- run_sim(n_groups, 20, group_means, n_iter)
  results$tukey[i] <- stats["tukey"]
  results$anova[i] <- stats["anova"]
  print(paste("Done:", n_groups, "groups; effect", effect))
}

# Save results
save(results, file = "tukey_sim_results.RData")
```