# Week 4 (LAB2): Analysis of Variance

## PSTAT122: Design and Analysis of Experiments

Delaney Thau

**Submission Instructions**

- This LAB must be completed and submitted **individually**. Collaboration is allowed for discussion, but each student must submit their own work.
- Ensure that all `R` code are presented clearly and appropriately.
- All figures should be numbered, and axes must be labeled.
- Please use the provided `LAB 2 M.qmd` file to type your solutions and submit the completed LAB as a PDF file.

You can utilize `RStudio` for this purpose. For guidance, refer to the Tutorial: Hello, Quarto).

- – Submit your solution via **Gradescope** as **a single PDF file**.

**Due Date**

**Due Date:** Wednesday, January 28, 2026, 11:59 PM

**Overview**

In this assignment, you will implement the ANOVA (Analysis of Variance) procedure from scratch, using basic operations in `R`. You'll compare your custom function's output with the built-in `aov` function to ensure correctness, and apply it to real datasets from class examples and homework.

# 1 Objectives:

1. Develop a custom ANOVA function that works with unbalanced datasets.
2. Validate your function by comparing its output to `R`'s built-in `aov` function.
3. Apply the function to multiple datasets to demonstrate its accuracy and versatility.

**!** Implementing ANOVA from Scratch (15 Points)

Write a function to perform ANOVA from scratch on any arbitrary dataframe.

- Your function needs to
- take a `dataframe` as its input.
- expect that the dataframe will have two columns:
- the outcome variable in the first column
- the factor variable in the second column
- It should do the version of the calculations from the slides that do not rely on having a balanced dataset (i.e consider different values of $n_i$ as in slide 51 of Chapter 3). Include a code comment that indicates exactly where in your calculations you are doing this.
- It should output the same exact values that a `summary` of the `aov` function does, in a matrix with the proper column and row names (elements of the matrix that have nothing in it should be filled in with `NA`).

**ANSWERS TO Implementing ANOVA from Scratch:**

```r
library("dplyr")
anova_from_scratch <- function(dataframe) {
  #Rename columns for consistency
  colnames(dataframe) <- c("y", "group")

  # Group by the factor and summarize: mean, variance, and count per group
  summary_stats <- dataframe %>%
    group_by(group) %>%
    summarize(
      y_bar_i = mean(y),
      S_i = var(y),
```

```r
    n_i = n(),
    .groups = 'drop'
  )

# Grand mean
grand_mean <- mean(dataframe$y)

# Number of groups (a) and total sample size (N)
a <- nrow(summary_stats) # number of groups
N <- nrow(dataframe) #total sample size

# Sum of Squares Between (SSB)
# Taking account for unequal group sizes using n_i (slide 51 from Ch 3)
SSB <- sum(summary_stats$n_i * (summary_stats$y_bar_i - grand_mean)^2)

# Sum of Squares Within/Error (SSE)
SSE <- sum(summary_stats$S_i * (summary_stats$n_i - 1))

# Degrees of Freedom
df_between <- a - 1
df_within <- N - a

# Mean Squares
MSB <- SSB / df_between
MSE <- SSE / df_within

# F Statistic
F_value <- MSB / MSE

# p-value
p_value <- pf(F_value, df_between, df_within, lower.tail = FALSE)

# Construct ANOVA table matrix
anova_matrix <- matrix(NA, nrow = 2, ncol = 5)
rownames(anova_matrix) <- c("group", "Residuals")
colnames(anova_matrix) <- c("DF", "Sum Sq", "Mean Sq", "F value", "Pr(>F)")

# Fill in values
anova_matrix[1, ] <- c(df_between, SSB, MSB, F_value, p_value)
anova_matrix[2, 1:3] <- c(df_within, SSE, MSE)

```

```r
  return(anova_matrix)
}
```

**!** Testing on Class Example Datasets (10 Points)

Test this function on the navigation data from Chapter 3 (Soft Drink) and (Paper Strength). Also re-run the `aov` function with its `summary` here to show that you get the same values either way.

**ANSWERS TO Testing on Class Example Datasets:**

```r
softdrink<- c(5.43, 5.71, 6.22, 6.01, 5.29, 6.24, 6.71, 5.98,
              5.66, 6.60, 8.79, 9.20, 7.90, 8.15, 7.55)
y <- as.vector(t(softdrink))
n = rep(5,3)
group <- rep(1:3, n)
data = data.frame(y = y, group = factor(group))

anova_from_scratch(data)
```

```
          DF   Sum Sq   Mean Sq  F value       Pr(>F)
group      2 18.78305 9.3915267 35.77225 8.781737e-06
Residuals 12  3.15044 0.2625367       NA           NA
```

```r
anova_real <- aov(y ~ group, data = data)
summary(anova_real)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
group        2  18.78   9.392   35.77 8.78e-06 ***
Residuals   12   3.15   0.263
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
paperstrength <- c(14.7, 15.1, 16.4, 14.1, 14.5, 12.7, 9.88, 10.7,
                   13, 13.3, 15.9, 12.8, 15, 18.4, 18.7, 19.3, 20,
                   18, 17.7, 16.3, 15.6, 18.6, 17.6, 15.1)
y <- as.vector(t(paperstrength))
n = rep(6,4)
group <- rep(1:4, n)
data = data.frame(y = y, group = factor(group))

```

```r
anova_from_scratch(data)
```

```
          DF    Sum Sq   Mean Sq  F value       Pr(>F)
group      3 110.76672 36.922239 13.61481 4.561864e-05
Residuals 20  54.23833  2.711917       NA           NA
```

```r
anova_real <- aov(y ~ group, data = data)
summary(anova_real)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
group        3 110.77   36.92   13.62 4.56e-05 ***
Residuals   20  54.24    2.71
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**!** Testing on Homework Dataset (5 Points)

Test your function on the data from Question 5 in the Homework 2. Run the `aov` function with its `summary` here as well to show that you get the same values either way.

**ANSWERS TO Testing on Homework Dataset:**

```r
curlingmethods <- c(42.5, 40.8, 39.6, 41.2, 45.1, 46.3, 44, 45.7, 38.9,
                    40.1, 41.5, 42, 36.2, 37.4, 35.8, 38)
y <- as.vector(t(curlingmethods))
n = rep(4,4)
group <- rep(1:4, n)
data = data.frame(y = y, group = factor(group))

anova_from_scratch(data)
```

```
          DF   Sum Sq   Mean Sq  F value       Pr(>F)
group      3 142.5069 47.502292 35.11643 3.202802e-06
Residuals 12  16.2325  1.352708       NA           NA
```

```r
anova_real <- aov(y ~ group, data = data)
summary(anova_real)
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
group        3 142.51   47.50   35.12 3.2e-06 ***
Residuals   12  16.23    1.35
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```