

Analysis of Variance

Mahbub Latif, PhD

July 2025

Plan

- One-way analysis of variance

A motivating example

- A large company is considering purchasing one of four computer packages designed to teach a new programming language
- Some experts of the company have claimed that these four packages are interchangeable
- To test whether these packages are interchangeable, the company decided to select 160 of its engineers and randomly divide them into four groups of size 40
- Each member of group i will be given teaching package to learn a new package ($i = 1, 2, 3, 4$) and when all the engineers will complete their study, a comprehensive exam will be taken
- Based on these exam results, the company wants to decide which computer package they should buy?

How can they do it?

A motivating example

- Let Y_{ij} be test score of the j th engineer of the i th group ($i = 1, 2, 3, 4, j = 1, \dots, 40$)
- Assume test scores of a group are independent and follow a normal distribution

$$Y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$$

- The null hypothesis to test the computer packages are interchangeable

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$H_1 : \text{not all means are equal}$$

- The method of comparing means of several (more than 2) populations is known as *analysis of variance* (ANOVA)

Some theoretical results

- Let Y_1, \dots, Y_n be a random sample from a normal distribution with parameters μ_i and σ^2

$$Z_i = \frac{Y_i - E(Y_i)}{\sigma} = \frac{Y_i - \mu_i}{\sigma} \sim \mathcal{N}(0, 1)$$

- Relationship between standard normal and chi-square distributions

$$Z_i^2 = \left[\frac{Y_i - \mu_i}{\sigma} \right]^2 \sim \chi_1^2 \text{ and } \sum_{i=1}^n Z_i^2 = \sum_{i=1}^n \left[\frac{Y_i - \mu_i}{\sigma} \right]^2 \sim \chi_n^2$$

- Suppose $\hat{\mu}_i$ depends on k parameters, e.g. $\mu_i = \beta_0 + \beta_1 x_1 + \dots + \beta_{k-1} x_{k-1}$

$$\sum_{i=1}^n \left[\frac{Y_i - \hat{\mu}_i}{\sigma} \right]^2 \sim \chi_{n-k}^2$$

One-way analysis of variance

- Let Y_{ij} be a response obtained from the j th subject of the i th group and assume

$$Y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2) \quad (i = 1, \dots, m, j = 1, \dots, n)$$

- We are interested to test

$$H_0 : \mu_1 = \dots = \mu_m = \mu \text{ (say)}$$

$$H_1 : \text{not all means are equal}$$

- It can be shown that

$$\frac{Y_{ij} - \mu_i}{\sigma} \sim \mathcal{N}(0, 1) \Rightarrow \sum_{i=1}^m \sum_{j=1}^n \left[\frac{Y_{ij} - \mu_i}{\sigma} \right]^2 \sim \chi_{nm}^2$$

- Overall and group-specific mean

$$\hat{\mu} = \bar{Y}_{..} = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n Y_{ij} \quad \text{and} \quad \hat{\mu}_i = \bar{Y}_{i.} = \frac{1}{n} \sum_{j=1}^n Y_{ij}$$

- Total sum of squares

$$SS_T = \sum_{i=1}^m \sum_{j=1}^n (Y_{ij} - \hat{\mu})^2$$

- Within-group sum of squares

$$SS_W = \sum_{i=1}^m \sum_{j=1}^n (Y_{ij} - \hat{\mu}_i)^2$$

- Between-group sum of squares

$$SS_B = \sum_{i=1}^m \sum_{j=1}^n (\hat{\mu}_i - \hat{\mu})^2 = n \sum_{i=1}^m (\hat{\mu}_i - \hat{\mu})^2$$

- Within-group sum of squares

$$\sum_{i=1}^m \sum_{j=1}^n \left[\frac{Y_{ij} - \mu_i}{\sigma} \right]^2 \sim \chi_{nm}^2 \Rightarrow \sum_{i=1}^m \sum_{j=1}^n \left[\frac{Y_{ij} - \hat{\mu}_i}{\sigma} \right]^2 = \frac{SS_W}{\sigma^2} \sim \chi_{mn-m}^2$$

- If $X \sim \chi_n^2$ then $E(X) = n$, so

$$E\left[\frac{SS_W}{\sigma^2}\right] = mn - m \Rightarrow E\left[\frac{SS_W}{m(n-1)}\right] = \sigma^2$$

- An estimate of σ^2

$$\hat{\sigma}^2 = \frac{SS_W}{mn - m} = MS_W$$

- If $H_0 : \mu_1 = \cdots = \mu_m = \mu$ is true, then it can be shown

$$\hat{\mu}_i = \bar{Y}_{i\cdot} = \frac{1}{n} \sum_{j=1}^n Y_{ij} \sim \mathcal{N}(\mu, \sigma^2/n)$$

- Between-group sum of squares

$$\sum_{i=1}^m \left[\frac{\hat{\mu}_i - \mu}{\sigma/\sqrt{n}} \right]^2 \sim \chi_m^2 \Rightarrow \sum_{i=1}^m \left[\frac{\hat{\mu}_i - \bar{\mu}}{\sigma/\sqrt{n}} \right]^2 = \frac{SS_B}{\sigma^2} \sim \chi_{m-1}^2 \Rightarrow E \left[\frac{SS_B}{\sigma^2} \right] = m - 1$$

- Under H_0 , an estimate of σ^2

$$\hat{\sigma}^2 = \frac{SS_B}{m-1} = MS_B$$

- Estimates of σ^2

$$\hat{\sigma}^2 = \begin{cases} SS_W/(mn - m) & \text{does not depend on whether } H_0 \text{ is true or not} \\ SS_B/(m - 1) & \text{if } H_0 \text{ is true} \end{cases}$$

- The following statistic will be large if H_0 is not true, i.e., if we reject H_0

$$F = \frac{SS_B/(mn - m)}{SS_W/(m - 1)}$$

- Since $SS_W/\sigma^2 \sim \chi_{mn-m}^2$ and $SS_B/\sigma^2 \sim \chi_{m-1}^2$, and they are independent

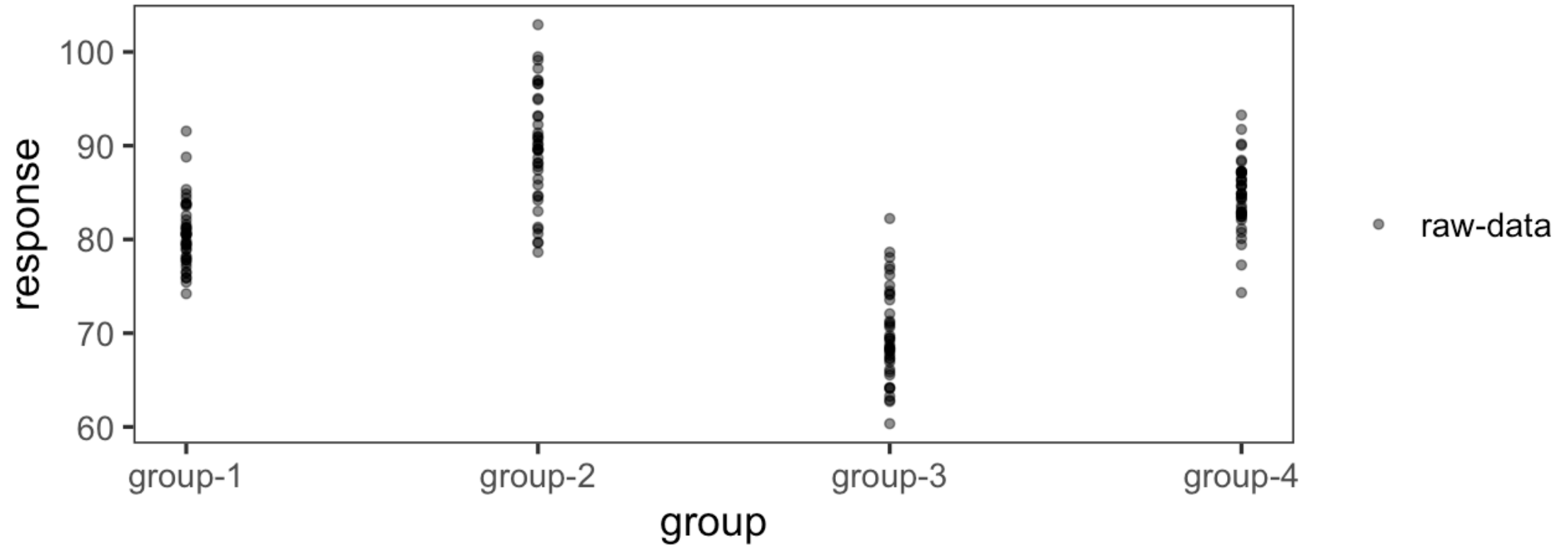
$$F = \frac{SS_B/(mn - m)}{SS_W/(m - 1)} \sim F_{mn-m, m-1}$$

- Reject H_0 at α level of significance if

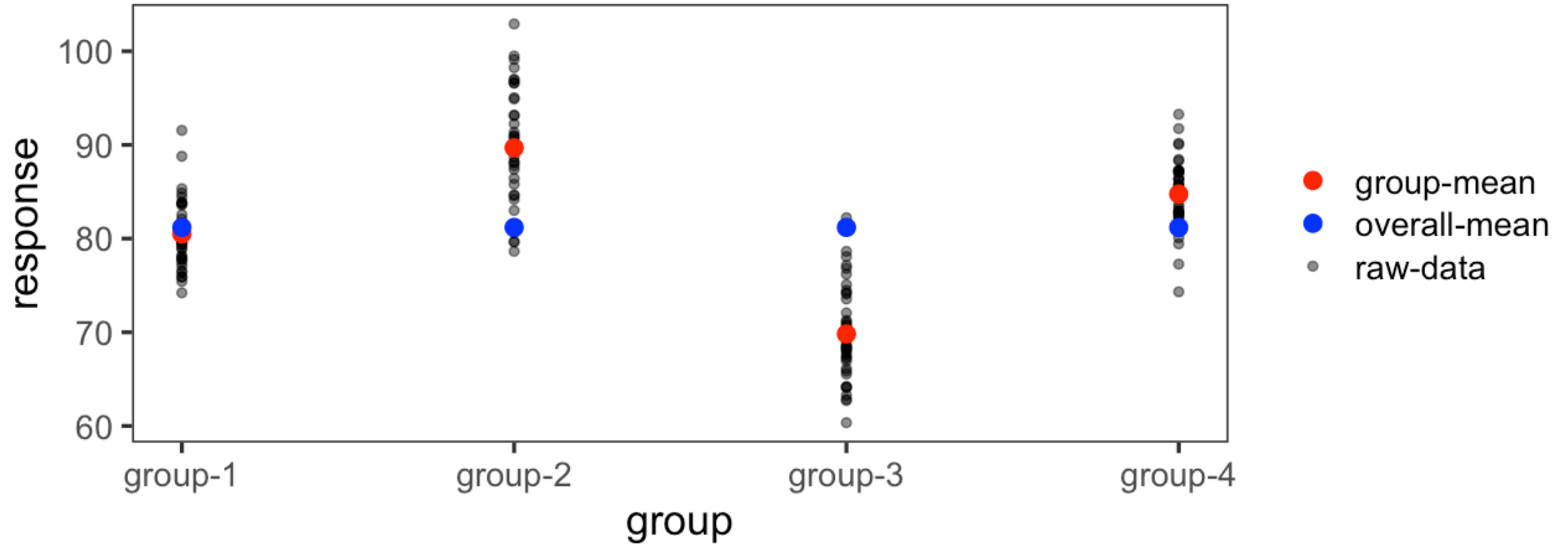
$$F > F_{mn-m, m-1, \alpha}$$

Data for comparing four group means

- Each group has 40 subjects

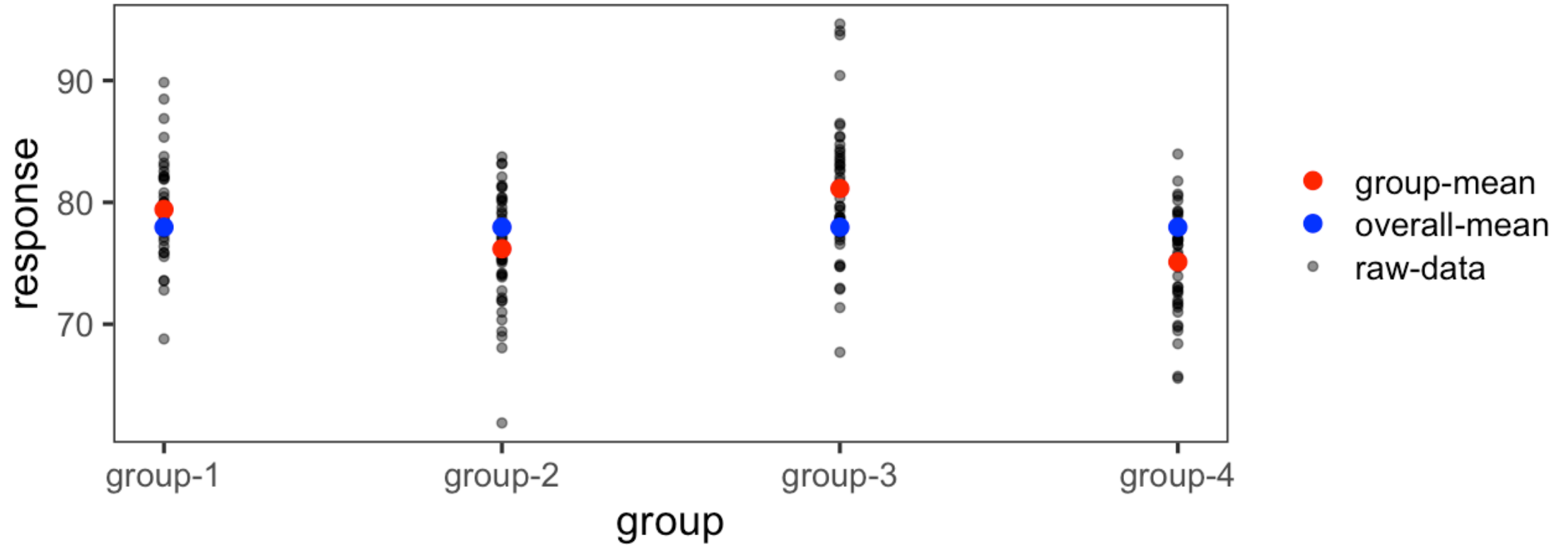


Case I



- $SS_B \rightarrow$ comparison between red and blue points
- $SS_W \rightarrow$ comparison between black and red points

Case II



- SS_B → comparison between red and blue points
- SS_W → comparison between black and red points

ANOVA table

sources	df	SS	MS	statistic
Between	$m - 1$	SS_B	$\frac{SS_B}{m-1}$	$F = \frac{MS_B}{MS_W}$
Within	$mn - m$	SS_W	$\frac{SS_W}{mn-m}$	

- The expressions of sum of squares

$$SS_B = n \sum_{i=1}^m (\hat{\mu}_i - \hat{\mu})^2 = n \left(\sum_{i=1}^m \hat{\mu}_i^2 - m\hat{\mu}^2 \right)$$

$$SS_W = \sum_{i=1}^m \sum_{j=1}^n (Y_{ij} - \hat{\mu})^2 - SS_B = \sum_{i=1}^m \sum_{j=1}^n Y_{ij}^2 - mn\hat{\mu}^2 - SS_B$$

Example 10.3a

- A auto rental firm is using 15 identical motors that are adjusted to run at a fixed speed to test 3 different brands of gasoline. Each brand is assigned to exactly 5 of the motors. Each motors run on 10 gallons of gasoline until it is out of fuel.
- The following represents the total mileages obtained by the different motors:
 - Gas 1: 220 251 226 246 260
 - Gas 2: 244 235 232 242 225
 - Gas 3: 252 272 250 238 256
- Test the hypothesis that the average mileage obtained is not affected by the type of gas used.

Example 10.3a

- The ANOVA table

sources	df	SS	MS	statistic
Between	2	863.33	431.67	2.6
Within	12	1991.60	165.97	NA

- Critical value $F_{2,12,.05} = 0.052$

Pair-wise comparison

- After rejecting the null hypothesis of equality of several means, the next step of the analysis is to compare pair-wise means, e.g. $H_0 : \mu_1 = \mu_2$
- Two means can be compared using the confidence interval of the difference of two means
 - $100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$

$$(\hat{\mu}_1 - \hat{\mu}_2) \pm t_{df, \alpha/2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{1}{n}}$$

- df is the residual degrees of freedom
- $\hat{\sigma}^2 = MS_W$

Pair-wise comparison

Gas	Mean
1	240.6
2	235.6
3	253.6

- $\hat{\sigma}^2 = MS_W = 165.97$

- 95% CI for $\mu_1 - \mu_2$

$$(\hat{\mu}_1 - \hat{\mu}_2) \pm t_{12,.025} \hat{\sigma} \sqrt{2/n} = 5 \pm (2.18)(\sqrt{165.97})(\sqrt{2/5}) = (-12.8, 22.8)$$

- $\mu_1 - \mu_3 : (-30.8, 4.8)$
- $\mu_2 - \mu_3 : (-35.8, -0.2)$

Problem (page 472-)

- 1, 2, 4, 5, 11, 12, 13,