Chapter 10 - Comparison of several means

STAT 251

Analysis of Variance (ANOVA)

Dr. Lasantha Premarathna

Chapter 10 - Learning Outcomes

- Identify situations where one-way ANOVA is and is not appropriate.
- State the modeling assumptions underlying ANOVA.
- State the null and alternative hypothesis for the ANOVA test.
- Explain the partitioning of the total sum of squares into the within and between group components.
- Identify the degrees of freedom associated with each sum of squares.
- Interpret an ANOVA table.
- Perform the F test in ANOVA.
- Use the data to estimate the underlying within-group variance.
- Explain the output from a software package for an ANOVA study

Analysis of Variance (ANOVA)

- In previous lecture, we discussed the situation where we wanted to compare the distribution of two groups under the assumption that population variances are equal.
- For example,
  - Do patients who received treatment have a better recovery rate than patients who don't receive the treatment?
  - Does a company sells more online using Website A than using website B?

Analysis of Variance (ANOVA)

- ANOVA is a statistical method that tests the equality of three or more population means by analyzing sample variances or variation in the data.
- The simplest ANOVA problem is referred to as single-factor, single-classification or one-way ANOVA
- Examples
  - An experiment to study the effect of five different brands of gasoline on automobile engines operating efficiency (km per liter)
  - Is there a difference in sales if a company uses Website A, Website B, or Website C?
  - An experiment to study the effect of presence of three different sugar solutions (glucose, sucrose, fructose) on bacterial growth.
- Quantitative response and categorical explanatory variable (factor/treatments)

Case study

- A company manufactures rubber handles for axes
- They test 4 different machines for the manufacturing of the rubber
- Compare the tensile strength (kilograms per square millimetre) of 5 handles made by each machine

- Response: tensile strength (kg/mm^2)
- Factor: machine used
  - Levels: 4

One-way ANOVA (single-factor ANOVA)

One-way ANOVA focuses on a comparison of 3 or more population or treatment means. Let

k = the number of populations or treatments being compared

μ_1 = the mean of the population 1 or the true average response when treatment 1 is applied

$\vdots$

μ_k = the mean of the population k or the true average response when treatment k is applied

Analysis of Variance (ANOVA)

- Suppose we have k groups, and we want to test

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

(i.e., the null hypothesis states that all the means are the same) versus

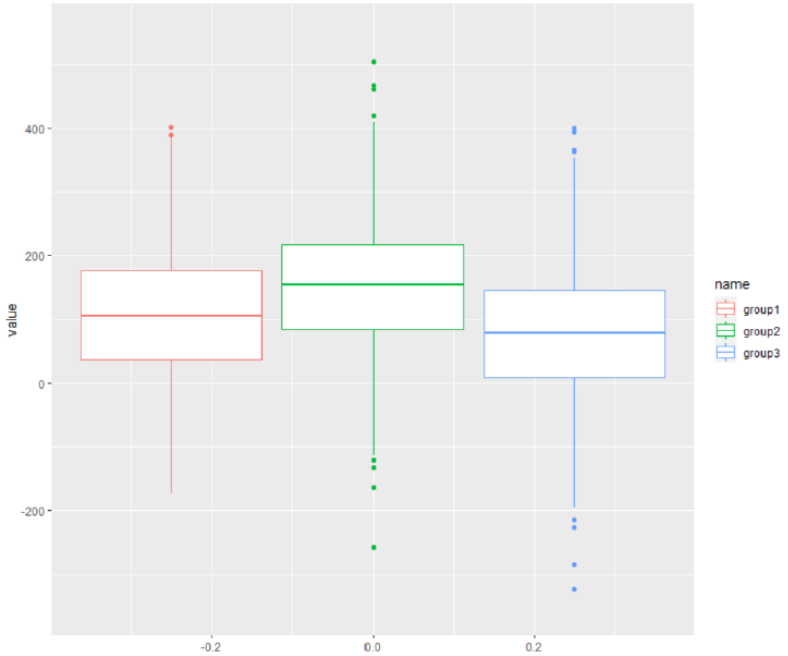
$$H_1 : \text{at least one of the means is different.}$$

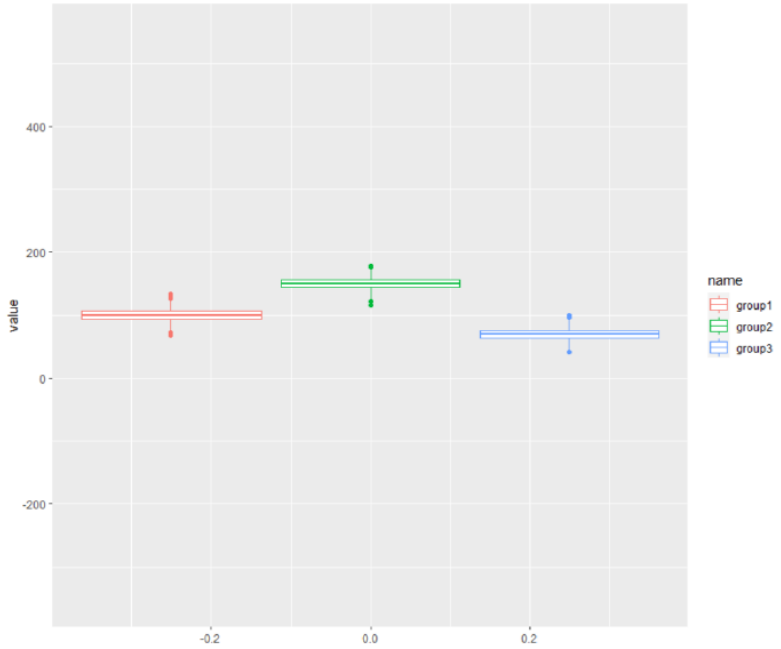
- Surprisingly, when we have multiple groups, we test the group's **means** by comparing the **variances**.

- Ok, I know, this seems weird! So, let's explore the idea.

Analysis of Variance (ANOVA)

Consider following two scenarios with respect to 3 groups in each case.





What do you think about above two scenarios?

1. It is very plausible that the centre of the three populations are the same and the difference observed is just due by random chance.
2. It is very unlikely the centre of all these three populations are the same.

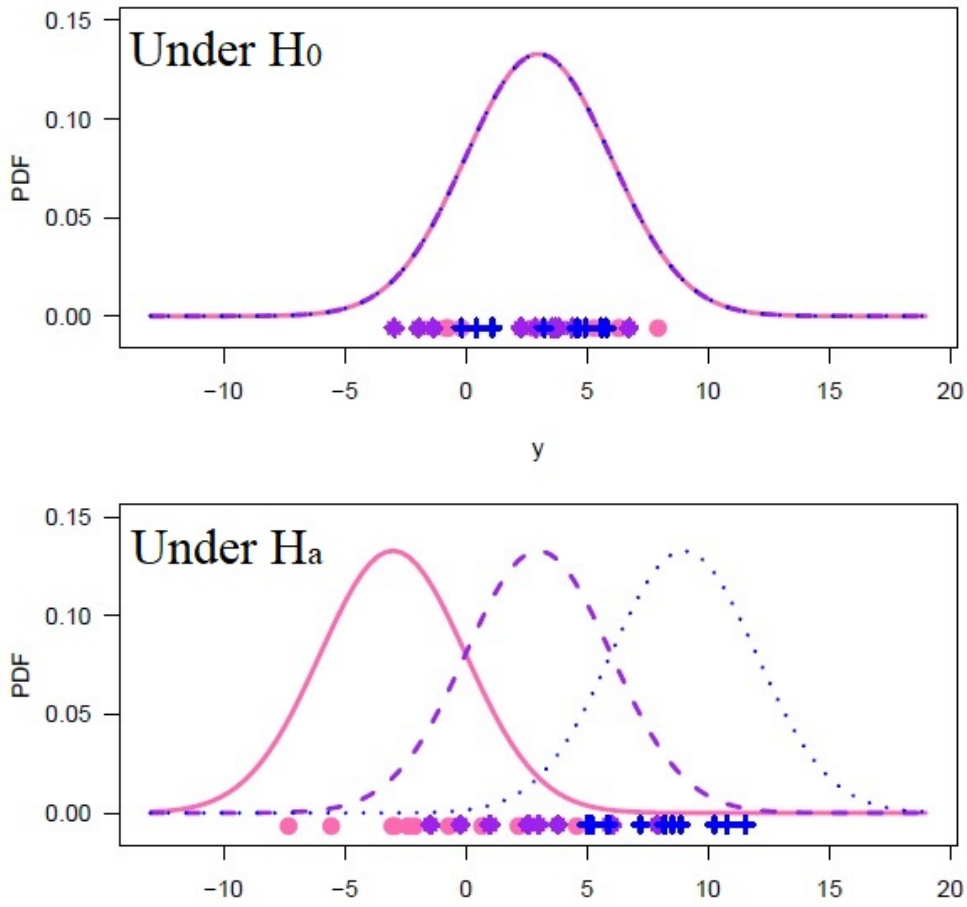
Hypotheses

- H_0: $\mu_1 = \mu_2 = ... = \mu_k$
- H_a: at least one of the means is different from the others (i.e. $\mu_i \neq \mu_j$ for $i \neq j$)

$\Rightarrow$ Reject H_0 means that at least two population means have different values

Hypotheses

Example: Consider $k = 3$. H_0: $\mu_1 = \mu_2 = \mu_3$. Under H_1 we have at least one of the means is different from the others. Here I have considered one possibility under H_1 that is all means are different.



Assumptions for ANOVA

- For each population, the response variable is normally distributed
- The variance (σ^2) of the response variable is same for all the populations
- data are independent

Notation

- k random samples are observed
- $y_{ij} = j^{th}$ observed value from the i^{th} population (treatment)

Treatment:	1	2	$\cdots$	i	$\cdots$	k
	y_{11}	y_{21}		y_{i1}		y_{k1}
	y_{12}	y_{22}		y_{i2}		y_{k2}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	y_{1n_1}	y_{2n_2}		y_{in_i}		y_{kn_k}
Total:	$y_{1\cdot}$	$y_{2\cdot}$		$y_{i\cdot}$		$y_{k\cdot}$
Mean:	$\bar{y}_{1\cdot}$	$\bar{y}_{2\cdot}$		$\bar{y}_{i\cdot}$		$\bar{y}_{k\cdot}$

$$\text{where } \bar{y}_{i\cdot} = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i} = \frac{y_{i\cdot}}{n_i}$$

Notation

- Total sample size $= n = n_1 + n_2 + \cdots + n_k$
- Grad Total (Overall Total) $= y_{\cdot\cdot} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}$
- Grad mean (Overall mean) $= \bar{y}_{\cdot\cdot} = \dfrac{y_{\cdot\cdot}}{n} = \dfrac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}}{n}$

Notation

- let Y_{ij} be the random variable that denote the j^{th} measurement taken from the i^{th} treatment
- then y_{ij} = observed value of Y_{ij}
- We can easily verify that (for $i = 1, 2, \cdots, k$)

$$E(\bar{Y}_{i\cdot}) = \mu_i$$

$\Rightarrow \bar{y}_{i\cdot}$ is an unbiased estimate for the unknown parameter μ_i

$$Var(\bar{Y}_{i\cdot}) = \frac{\sigma^2}{n_i}$$

Notation

- For k random samples, we can calculate sample variances

$$s_1^2, s_2^2, \cdots, s_k^2 \qquad \text{where } s_i^2 = \frac{\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2}{n_i - 1}$$

- $s_1^2, s_2^2, \cdots, s_k^2$ are k different unbiased estimates for the common variance σ^2

$$E(s_i^2) = \sigma^2 \,, \quad \text{for } i = 1, 2, \cdots, k$$

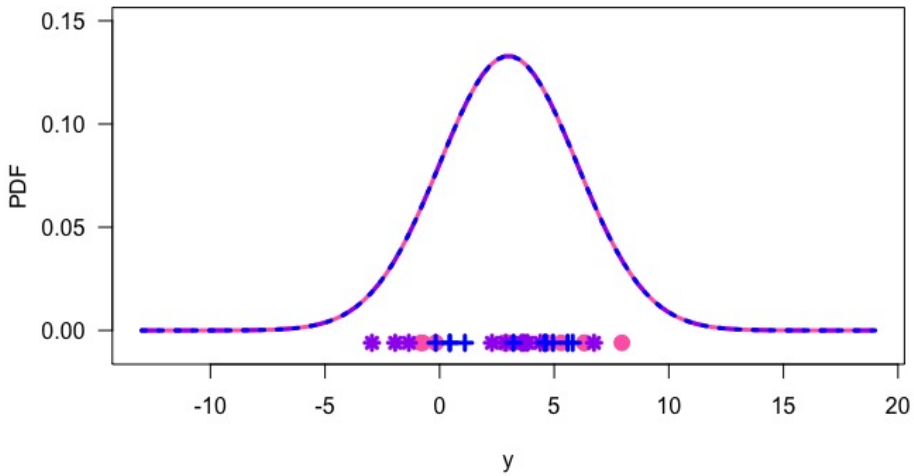
- These k estimates can be combined to obtain unbiased estimate for σ^2

$$s^2 = \frac{\sum_{i=1}^{k} (n_i - 1) s_i^2}{n - k} = MSE \text{ (explain MSE later)}$$

Motivation for ANOVA

- When H_0 true (i.e. $H_0 : \mu_1 = \mu_2 = ... = \mu_k$)
  Sample means are close together because there is only one sampling distribution.

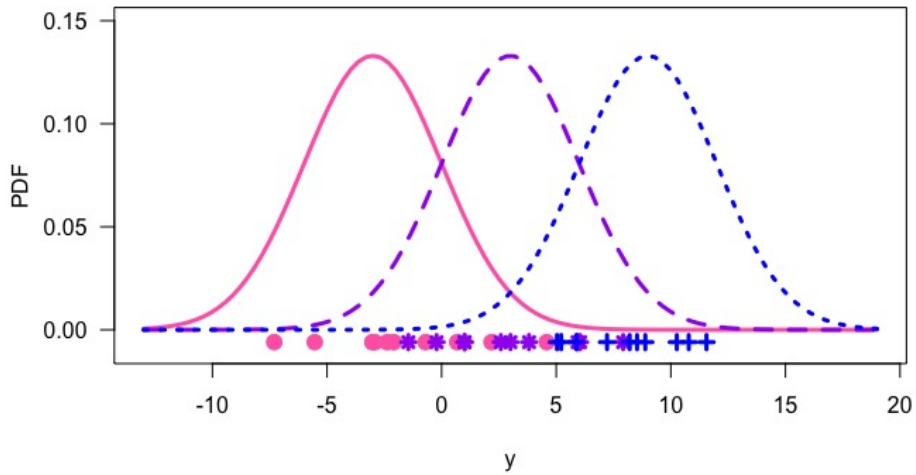
  Under H_0



Motivation for ANOVA

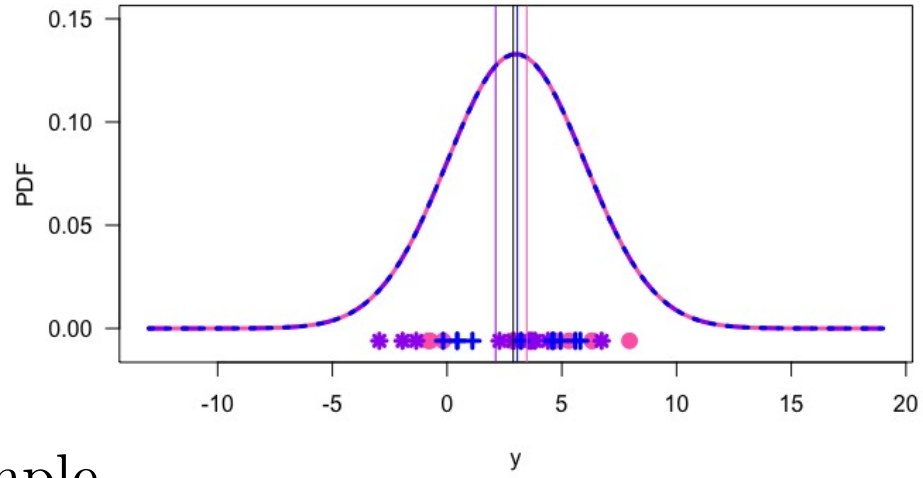
- When H_0 false (at least one of the means is different from the others)
  Sample means come from different sampling distributions and are not close together.

  An H_1 example (consider all means are different)

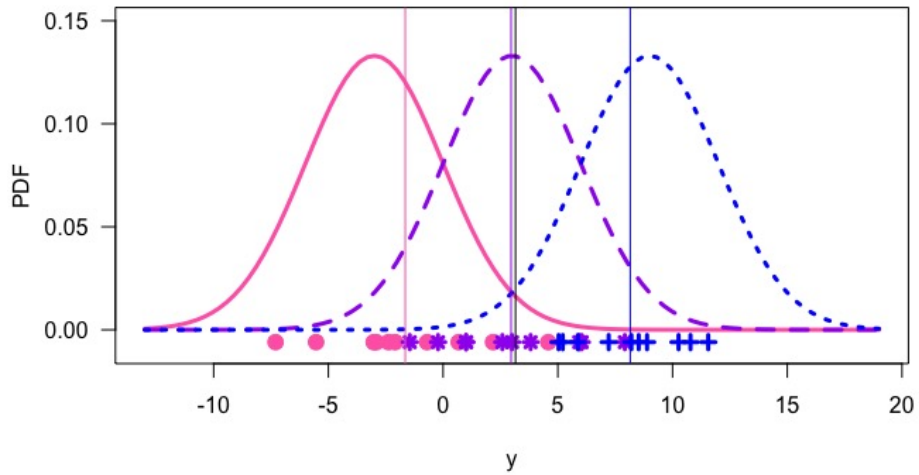


Motivation for ANOVA

Under H_0



An H_a example



Total Variation

- In ANOVA, variations are measured by sums of squares (SS)
- Total Variation in the data (SST - Total sum of squares) comes from two sources.
  1. Variation between groups/treatments
     (SSTr - treatment sum of squares)
  2. Variation withing groups/treatments
     (SSE - Error sum of squares)

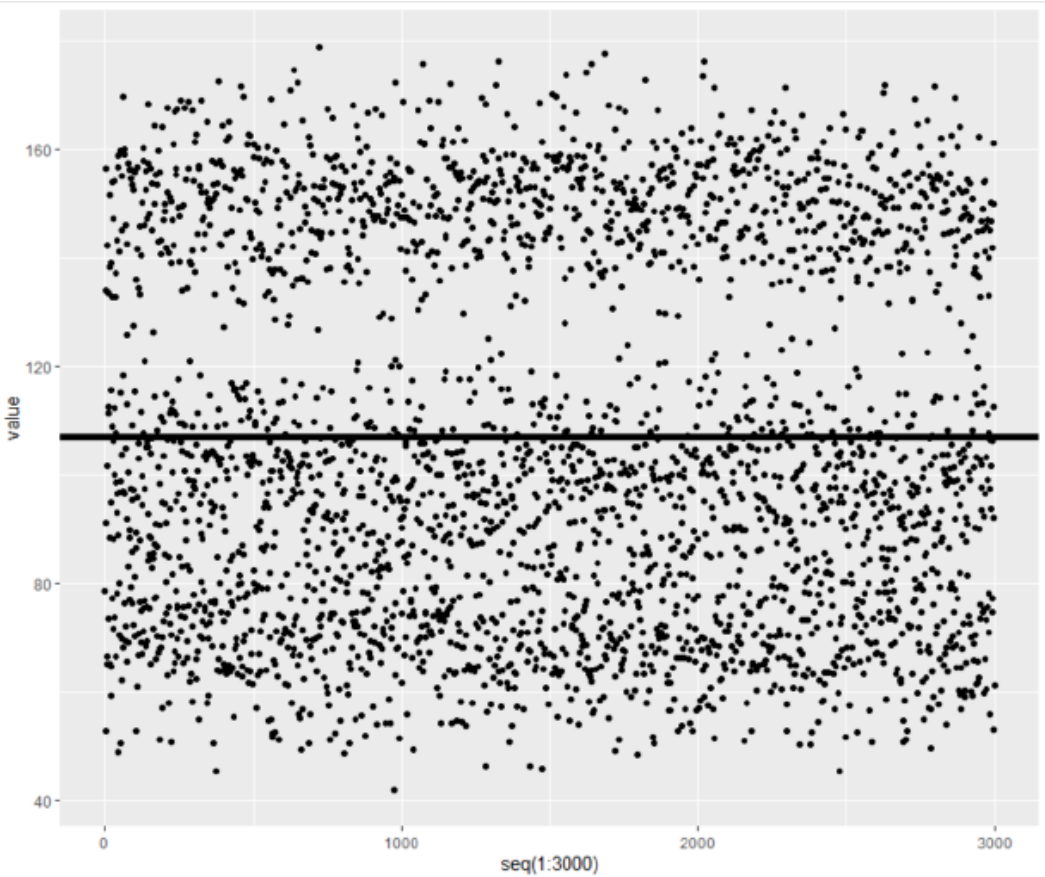
$$SST = SSTr + SSE$$

In other words this can be expressed

Overall variability = Between group SS + Within group SS

Analysis of Variance (ANOVA)

In ANOVA, Variations are measured by sums of squares (SS). Imagine a problem where we have three groups



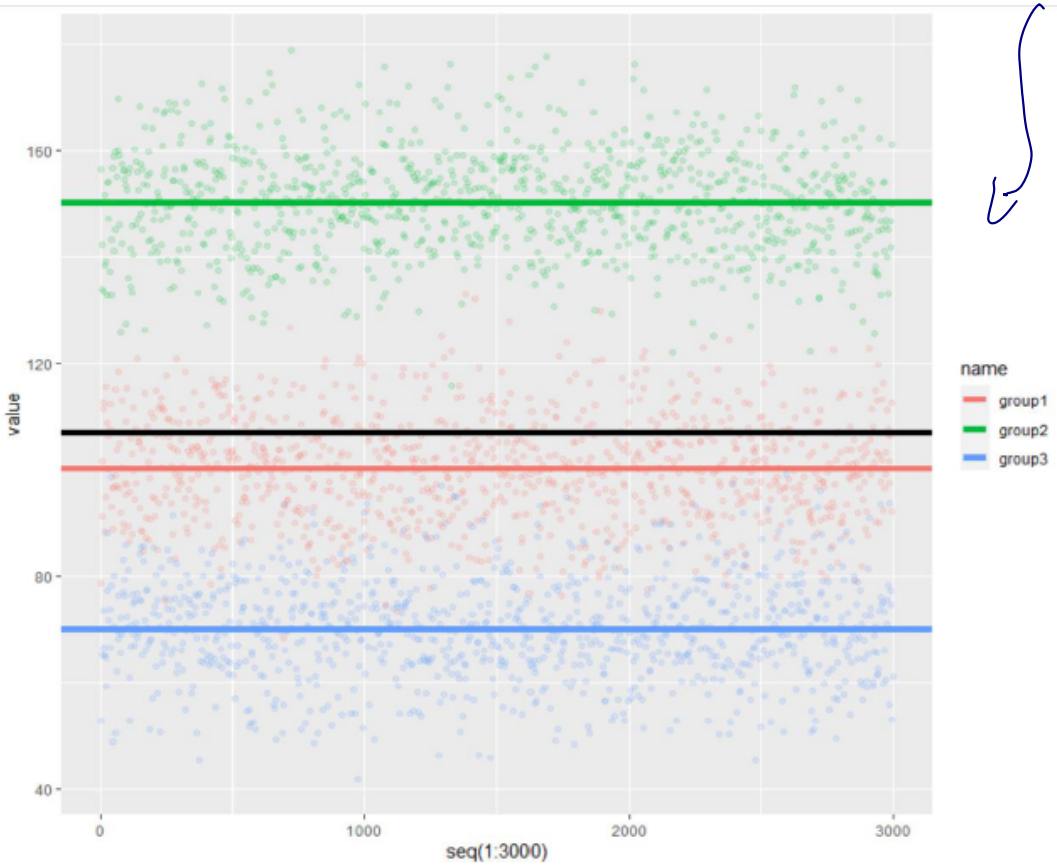
Overall mean

Overall variance:

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$$

Analysis of Variance (ANOVA)

In ANOVA, Variations are measured by sums of squares (SS). Imagine a problem where we have three groups



Means.

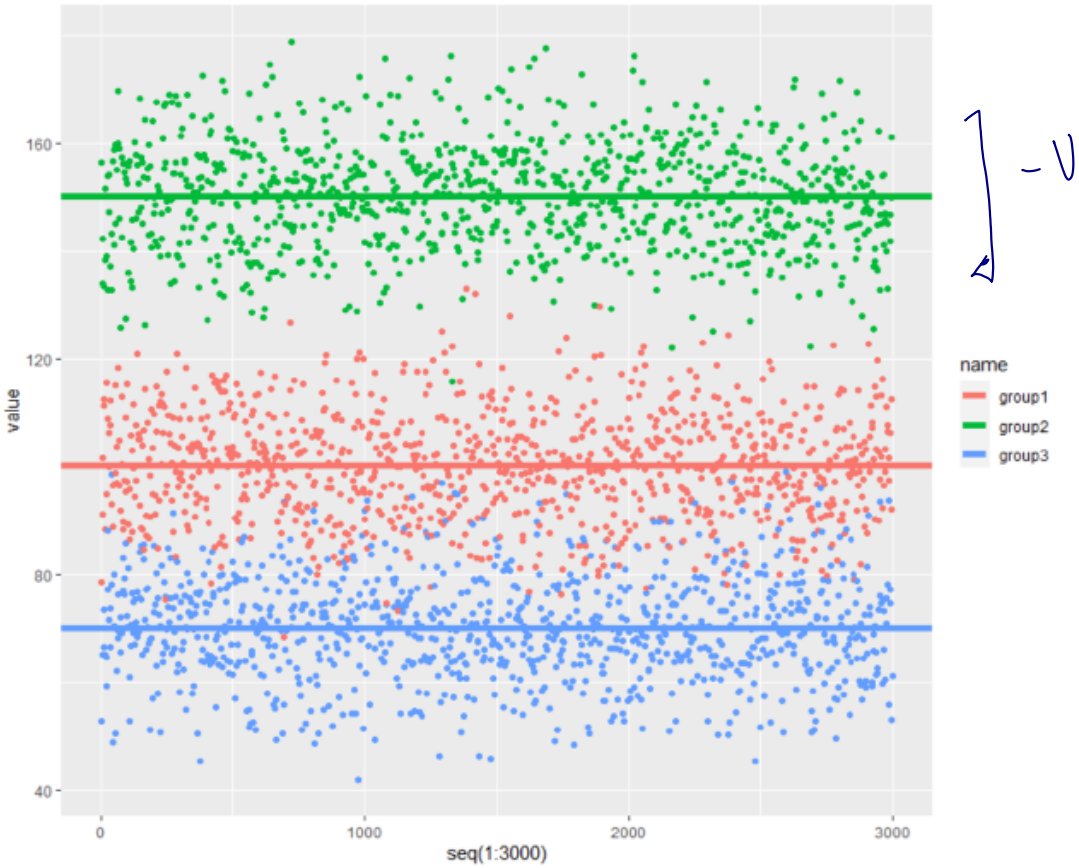
Variance **between** groups:

$$SSTr = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$$

Overall mean.

Analysis of Variance (ANOVA)

In ANOVA, Variations are measured by sums of squares (SS). Imagine a problem where we have three groups



- Variance

Variance **within** groups:

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2$$

SST = SSTr + SSE

$$SST = SSTr + SSE$$

Overall variability = Between group SS + Within group SS

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{1}{n}\, y_{..}^2$$

$$SSTr = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_{i.} - \bar{y}_{..})^2 = \sum_{i=1}^{k}\frac{1}{n_i}\, y_{i.}^2 - \frac{1}{n}\, y_{..}^2$$

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{k}\frac{y_{i.}^2}{n_i}$$

$$\text{also } SSE = \sum_{i=1}^{k}(n_i - 1)s_i^2 \qquad \text{where } s_i^2 = \frac{\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2}{n_i - 1}$$

Degrees of Freedom

.

$$df(SST) = n - 1$$

$$df(SSTr) = k - 1$$

$$df(SSE) = n - k \qquad 20 - 4$$

$$SST = SSTr + SSE$$

$$\Rightarrow \quad df(SST) = df(SSTr) + df(SSE)$$

Mean Squares

- Mean Squares Treatment $= MSTr = \dfrac{SSTr}{k-1}$
- Mean Squares Error $= MSE = \dfrac{SSE}{n-k}$
  - MSE is a measure of within-samples variability

Note:

- If we increase sample size, SSE will be increased.
- If we consider more groups, $SSTr$ will be increased.
- We want a test statistic that works in general to test hypotheses, no matter the size of your sample or the number of groups we have. Therefore, we consider Mean Squares as above to calculate the test statistic.

ANOVA Test Procedure

- Hypotheses

  $H_0 : \mu_1 = \mu_2 = ... = \mu_k$

  H_a: at least one of the means is different from the others

- Test statistic

$$F_{obs} = \frac{\frac{SSTr}{k-1}}{\frac{SSE}{n-k}} = \frac{MSTr}{MSE} \sim F_{\nu_1, \nu_2}$$

Under $H_0, F_{obs} = \frac{MSTr}{MSE}$ follows the F**-distribution** with degrees of freedom ν_1 (numerator df) and ν_2 (denominator df)

$$\nu_1 = df(SSTr) = k - 1$$

$$\nu_2 = df(SSE) = n - k$$

ANOVA Test Procedure

- Hypotheses

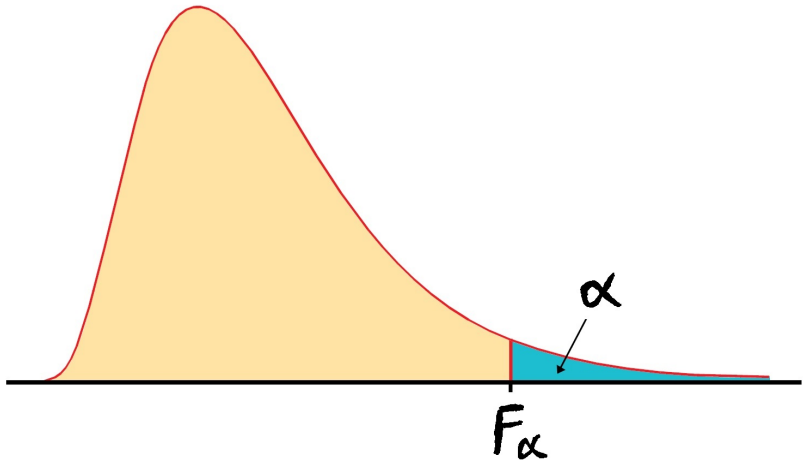
  $H_0 : \mu_1 = \mu_2 = ... = \mu_k$

  H_a: at least one of the means is different from the others

- Test statistic

$$F_{obs} = \frac{MSTr}{MSE} \sim F_{\nu_1,\, \nu_2}$$

- Reject H_0 if $F_{obs} \geq F_\alpha$ or p-value $\leq \alpha$



- **F-distribution** is an asymmetric distribution.

The ANOVA Table

Source of Variation	df	Sum of Squares (SS)	Mean Square (MS)	F-ratio
Treatment (Between group)	$k-1$	$SSTr$	$MSTr = \frac{SSTr}{k-1}$	$\frac{MSTr}{MSE}$
Error (Within group)	$n-k$	SSE	$MSE = \frac{SSE}{n-k}$	
Total	$n-1$	SST		

The ANOVA Model

- The assumptions of one-way ANOVA can be described succinctly by mean of the "model equation"

- Each measurement will be represented as the sum of two terms; unknown constant μ_i and a random variable ϵ_{ij}

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

$$i = 1, 2, \cdots, k \qquad \text{and} \qquad j = 1, 2, \cdots, n_i$$

Where ϵ_{ij} represents a random deviation from the population or true treatment mean μ_i

$\Rightarrow$ ϵ_{ij}'s are iid rvs such that $\epsilon_{ij} \sim N(0, \sigma^2)$

Before the next class ...

Visit the course website at canvas.ubc.ca

- Review Lecture 31 and related sections in the text book
- Topic of next class: **Chapter 10: ANOVA Examples**