ANOVA: Analysis of variance

Peter Tea

University of Ottawa

July 2017

uOttawa

Statistics

- Extract sensible information
- Investigate trends or patterns
- Find associations between variables

Statistics

- Extract sensible information
- Investigate trends or patterns
- Find associations between variables

Health Data:

Daily Exercise	Plasma Cholesterol (mmol/L)
A) 60 + minutes	
B) 31 - 60 minutes	
C) 15 - 30 minutes	
D) $<$ 15 minutes	

Analysis of Variance

Are the groups (A, B, C, D) actually different from one another in terms of the measured plasma cholesterol levels?

Analysis of Variance

Are the groups (A, B, C, D) actually different from one another in terms of the measured plasma cholesterol levels?

$$H_0: \quad \mu_A = \mu_B = \mu_C = \mu_D \tag{1}$$

$$H_1: \quad \mu_i \neq \mu_j \textit{ for at least one pair } (i, j) \tag{2}$$

ANOVA: How it Works

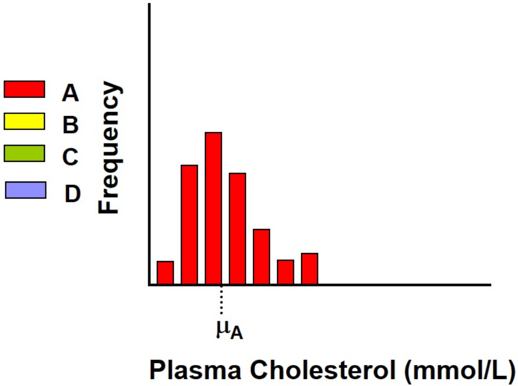
1. Split the data into groups corresponding to the different levels of the variable *Exercise*.
2. Analyse the variances among groups and compare to variances within groups.

Analysis of Variance

1. Split the data into groups corresponding to the different levels of the variable *Exercise*.

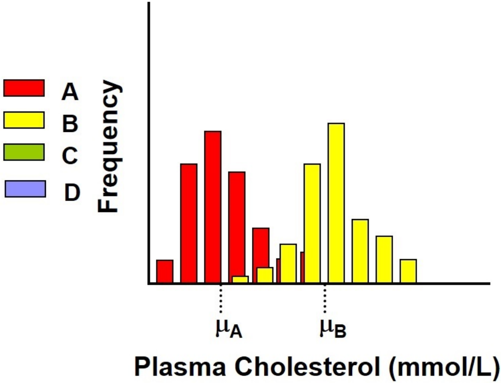
Analysis of Variance

1. Split the data into groups corresponding to the different levels of the variable *Exercise*.



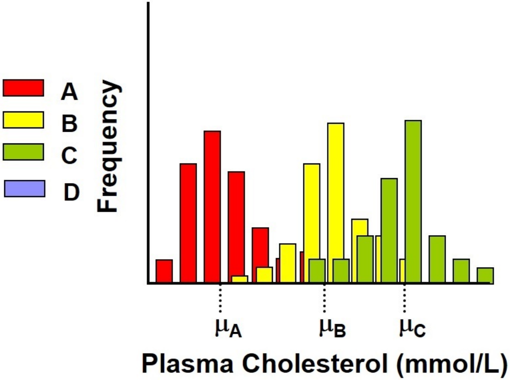
Analysis of Variance

1. Split the data into groups corresponding to the different levels of the variable *Exercise*.



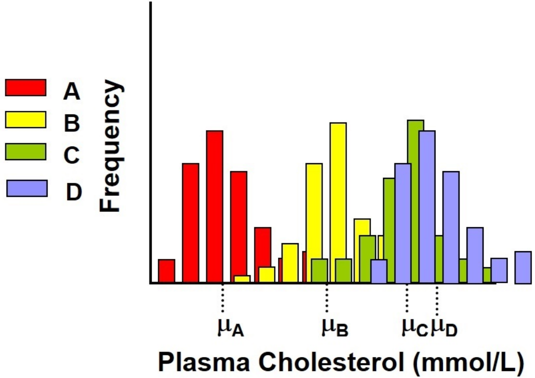
Analysis of Variance

1. Split the data into groups corresponding to the different levels of the variable *Exercise*.



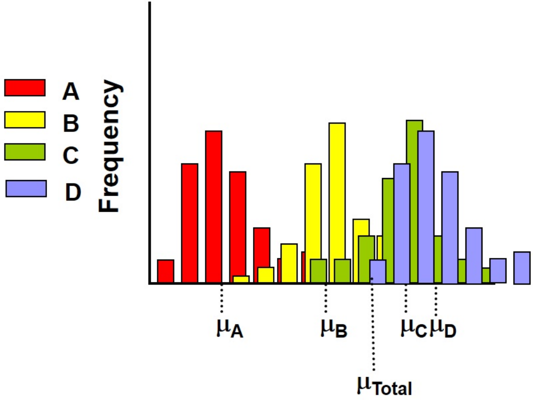
Analysis of Variance

1. Split the data into groups corresponding to the different levels of the variable *Exercise*.



Analysis of Variance

1. Split the data into groups corresponding to the different levels of the variable *Exercise*.



Analysis of Variance

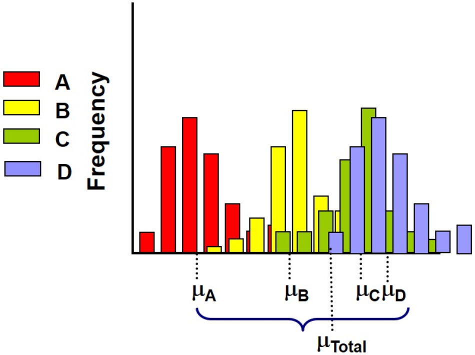
2. Analyse the variances among groups and compare to variances within groups.

$$S^2 = \sum \frac{(X - \bar{X})^2}{N - 1}$$

Analysis of Variance

2. Analyse the variances among groups and compare to variances within groups.

$$S^2 = \sum \frac{(X - \bar{X})^2}{N - 1}$$



Analysis of Variance

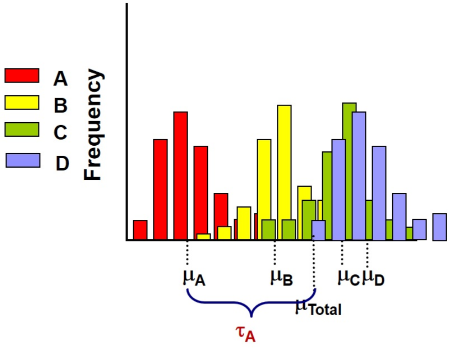
Data Modelling:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \begin{cases} i = 1, 2, 3, 4 \\ j = 1, 2, 3, ..., n \end{cases}$$

Analysis of Variance

Data Modelling:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \begin{cases} i = 1, 2, 3, 4 \\ j = 1, 2, 3, ..., n \end{cases}$$



Analysis of Variance

Data Modelling:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \begin{cases} i = 1, 2, 3, 4 \\ j = 1, 2, 3, ..., n \end{cases}$$

- y_{ij} represents the ijth observation
- μ represents the overall mean (i.e. the mean pooled across all levels)
- τ_i is a unique parameter to each group level and is referred to as the *treatment effect*. τ_i represents the deviation from the overall mean resulting from the ith group level.
- ϵ_{ij} is the random error of the experiment. The random error represents other sources of variability (eg. variability due to measurement errors or due to background noise.)

Analysis of Variance

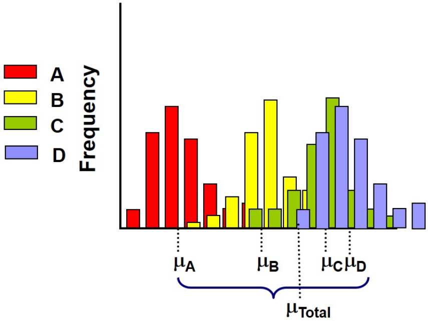
Compare variances among groups to variances within groups with the F-Test:

$$F_0 = \frac{\frac{SS_{Levels}}{a-1}}{\frac{SS_E}{N-a}} = \frac{MS_{levels}}{MS_E} = \frac{\text{Variation among groups}}{\text{Variation within groups}}$$

Analysis of Variance

Compare variances among groups to variances within groups with the F-Test:

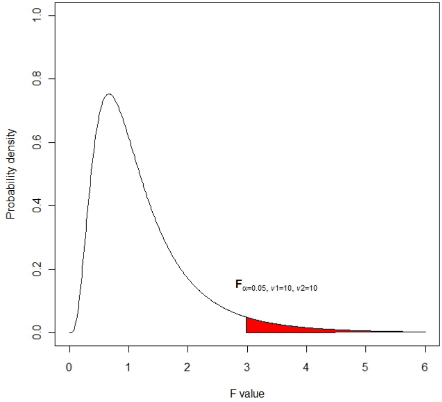
$$F_0 = \frac{\frac{SS_{Levels}}{a-1}}{\frac{SS_E}{N-a}} = \frac{MS_{levels}}{MS_E} = \frac{\text{Variation among groups}}{\text{Variation within groups}}$$



F-Test

The null hypothesis should be rejected if

$$F_0 > F_{\alpha,a-1,N-a}$$

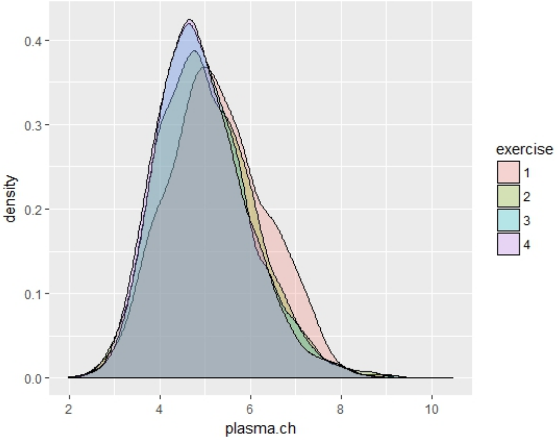


```
> setwd("C:/Users/Peter/Documents/Summer Research/
Topic 1 R")
> data <-read.csv("heartdata.csv", header = TRUE,
 sep = ",")

> model.plasma <- aov(plasma.ch~ exercise,
data = new.data) #Set up the One-Way ANOVA

> summary(model.plasma)
               Df Sum Sq Mean Sq F value Pr(>F)
exercise        3     86  28.723   27.33 <2e-16 ***
Residuals   11730  12329   1.051
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
```

Analysis of Variance



Problems?

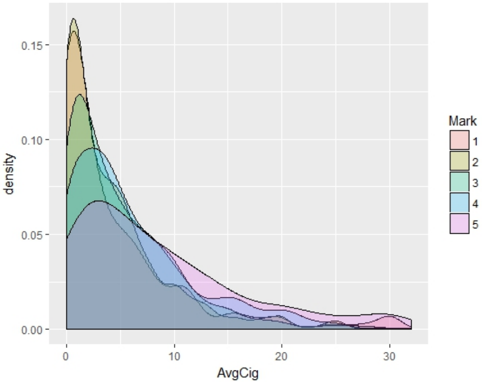
Assumption: The errors are normally and independently distributed random variables

$$\epsilon_{ij} \sim N(0, \sigma^2)$$

Problems?

Assumption: The errors are normally and independently distributed random variables

$$\epsilon_{ij} \sim N(0, \sigma^2)$$



Problems?

$$H_0: \quad \mu_A = \mu_B = \mu_C = \mu_D \tag{3}$$

$$H_1: \quad \mu_i \neq \mu_j \text{ for at least one pair } (i, j) \tag{4}$$

Multiple Comparisons

$A \leftrightarrow B$	$B \leftrightarrow C$	$C \leftrightarrow D$
$A \leftrightarrow C$	$B \leftrightarrow D$	
$A \leftrightarrow D$		

Table: All pairwise comparisons

Multiple Comparisons

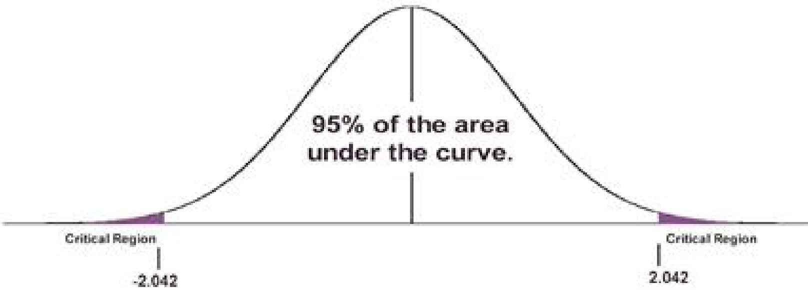
$A \leftrightarrow B$	$B \leftrightarrow C$	$C \leftrightarrow D$
$A \leftrightarrow C$	$B \leftrightarrow D$	
$A \leftrightarrow D$		

Table: All pairwise comparisons

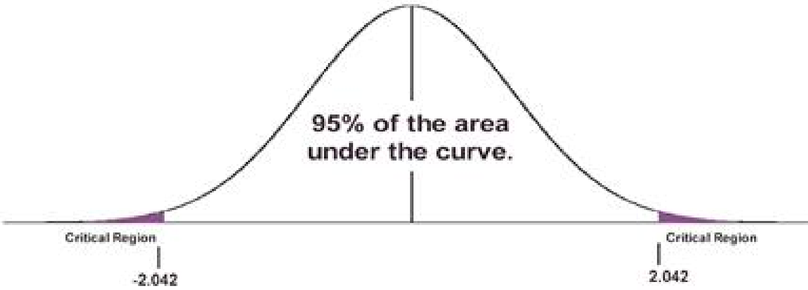
T-test: $\dfrac{\bar{X}_1 - \bar{X}_2}{Sp\sqrt{(1/N_1 + 1/N_2)}}$

Multiple Comparisons

T-test: $$\frac{\bar{X}_1 - \bar{X}_2}{Sp\sqrt{(1/N_1 + 1/N_2)}}$$



Multiple Comparisons



Type 1 error rate (α): Pr(Falsely rejecting H_0 | H_0 is true)

If $\alpha = 0.05$ then:

$$\begin{aligned} Pr(\text{not rejecting } H_0 | H_0 \text{ is true}) &= 1 - 0.05 \\ &= 0.95 \end{aligned}$$

Multiple Comparisons

Type 1 error rate (α): Pr(Falsely rejecting H_0 | H_0 is true)

If $\alpha = 0.05$ then:

$$\begin{aligned} Pr(\text{not rejecting } H_0 | H_0 \text{ is true}) &= 1 - 0.05 \\ &= 0.95 \end{aligned}$$

However, there are 6 total unique comparisons that can be made on the same data.
The probability of obtaining the correct decision in *all* comparisons made is:

$$\begin{aligned} (1-\alpha)^6 &= (1-0.05)^6 \\ &= 0.74 \end{aligned}$$

The type I error rate is inflated to:

$$1 - 0.74 = 0.26$$