Statistical Methodology for Software Engineering

Hadas Lapid, PhD

Contents

Analysis of Variance

Reference:
https://en.wikipedia.org/wiki/Analysis_of_variance

ANOVA: Analysis of Variance

Definition

Comparison of k population means in normal variable Y

Why isn’t t-test enough?

Multiple hypothesis testing problem

ANOVA

Model and definitions

- Normal variable Y is measured in k independent samples
- $\{\mu_i, \sigma^2\}$ are the i^{th} population's mean and variance
- $n_i = i^{th}$ population sample size
- Y_{ij} = observation j in the i^{th} sample
- $N = n_1 + n_2 + \ldots + n_k$ = total sample size

Assumptions:

- Normality of mutual distribution
- Equality of variances

$$\boxed{Y_{ij} \sim N\left(\mu_j, \sigma^2\right)}$$

Hypothesis testing:

$$\boxed{\begin{array}{l} H_0: \mu_1 = \mu_2 = \cdots = \mu_k \\ H_A: \text{otherwise} \end{array}}$$

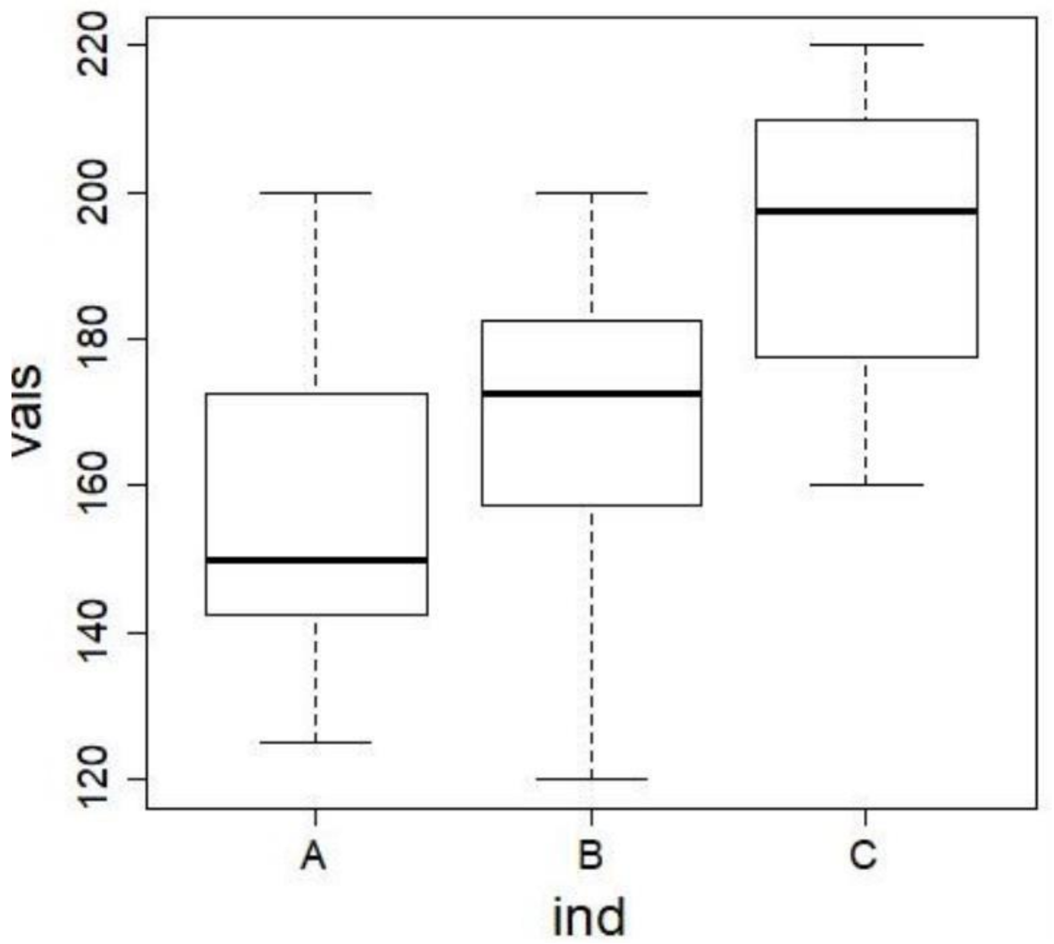
ANOVA

Example: Comparison between means of three treatment groups

Diet	**A**	**B**	**C**
Sample size, n_i	7	8	8
y_{ij}	165,200,140,125, 180,150,145	180,200,185,170, 120,175,165,150	195,215,175,220, 180,160,205,200
Sample mean, $\overline{Y_i}$	157.86	168.13	193.75
Sample variance, s_i^2	657.14	592.41	426.79
Sample standard deviation, s_i	25.63	24.34	20.66

Comparison between means of three treatment groups

Step 1: Boxplot visualization, targeting of outliers



Measuring the differences between groups

Variability between groups:

$\overline{Y_i}$ ≡ sample i mean, i:1,…k
$\overline{\overline{Y}}$ ≡ overall mean of all the data

$$\text{Sum of squares between groups} = SS_{\text{between groups}} = \sum_{i=1}^{k} n_i\left(\overline{Y}_i - \overline{\overline{Y}}\right)^2$$

- If one of more group means deviate from overall mean, SS_{bw} will be large
- How do we judge what is large?

Define average variability between groups:

$$MS_{\text{between groups}} = SS_{\text{between groups}}/(k-1)$$

ANOVA

Numeric Example: **Variation between group means**

Overall mean of means, $\bar{\bar{Y}} = \frac{1}{N} \cdot \sum_{j=1}^{3} \sum_{i=1}^{n_i} y_{ij} = 173.91$

Diet	**A**	**B**	**C**
Sample size, n_i	7	8	8
Sample mean, $\overline{Y_i}$	157.86	168.13	193.75
$n_i\left(\bar{Y}_i - \bar{\bar{Y}}\right)^2$	7*(157.86-173.91)^2 = 1803.22	8*(168.13-173.91)^2 = 267.27	8*(193.75-173.91)^2 = 3149.00

$SS_{between}$ = 1803.22+267.27+3149.00 = 5219.52

$MS_{between}$ = 5219.52/(3-1) = 2609.76

Measuring the differences within groups

In order to estimate how large/small is the variation between groups, we will compare it to the variability within groups.

Assume equal variance in all groups, estimated as σ^2 - weighted mean of k variances

$$\boxed{\sigma^2 = \frac{\sum_{i=1}^{k}(n_i - 1)\cdot S_i^2}{\sum_{i=1}^{k}(n_i - 1)}}$$

- $SS_{error} = \sum_{i=1}^{k}(n_i - 1)\cdot S_i^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_i\right)^2$
- $\sum_{i=1}^{k}(n_i - 1)$ = n-k

$$\longrightarrow \boxed{MS_{error} = \sigma^2 = \frac{SS_{error}}{n-k}}$$

The F Distribution

Estimated population variance in each group

$$\boxed{MS_{error} = \sigma^2 = \frac{\sum_{i=1}^{k}(n_i - 1)\cdot S_i^2}{\sum_{i=1}^{k}(n_i - 1)}}$$

$$MS_{error} = \sigma^2 = \frac{6*657.14+7*592.41+7*426.79}{6+7+7} = 553.86$$

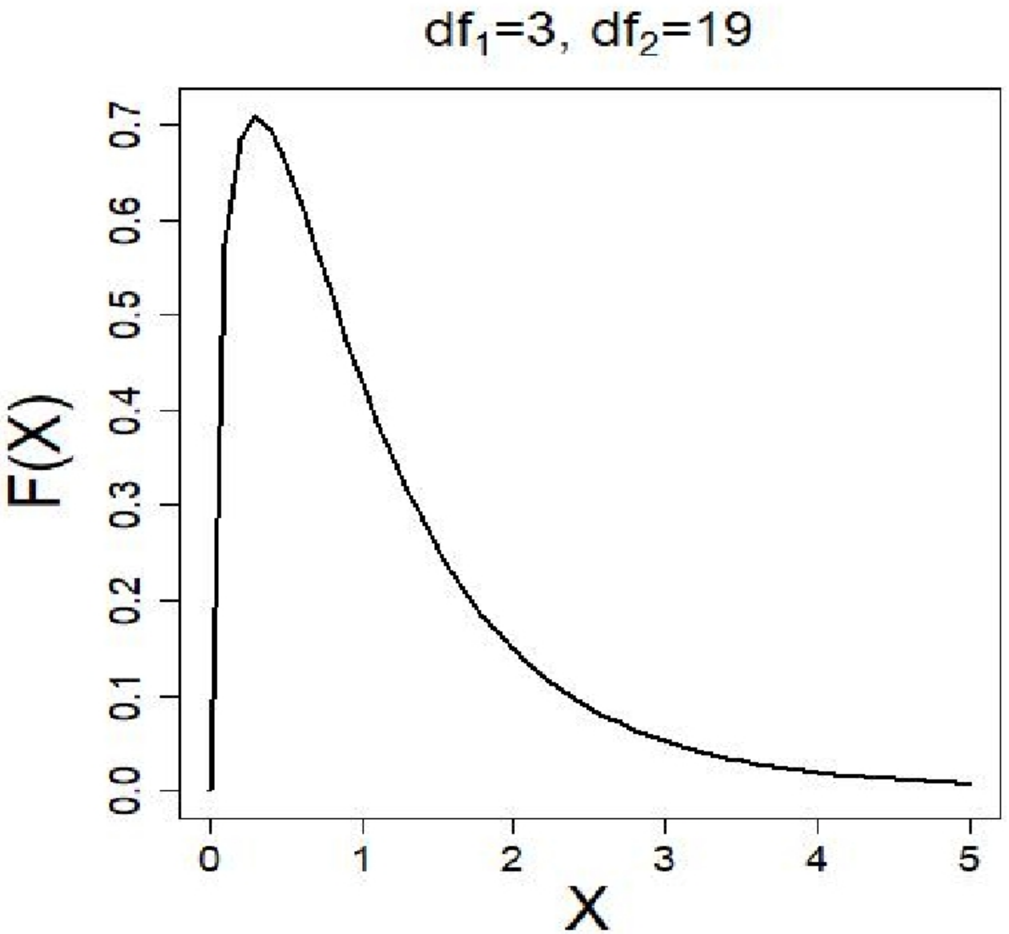
Statistical estimate of hypothesis testing:

$$\boldsymbol{F} = \frac{\boldsymbol{MS_{between\ groups}}}{\boldsymbol{MS_{error}}} = \frac{2609.76}{553.86} = 4.71$$

H_0 is rejected for $F>F_c$

F – the Statistical Estimate of ANOVA

$F(df_1,df_2)$ is χ^2-like distribution
df_1 = k-1 (k number of groups)
df_2 = n-k (n overall sample size)



Standard layout of ANOVA table

$SS_{between}$

F statistic

P_v **(Prob>F)**

$MS_{between}$

df1 = K-1 groups

```
              Df Sum Sq Mean Sq F value Pr(>F)
ind            2   5221  2610.3   4.713  0.021 *
Residuals     20  11077   553.9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

df2=n-k samples

MS_{error} (σ^2)