Biomedical Engineering Degree

7. ANalysis Of VAriance ANOVA

Felipe Alonso Atienza

✉felipe.alonso@urjc.es

@FelipeURJC

Escuela Técnica Superior de Ingeniería de Telecomunicación
Universidad Rey Juan Carlos

References

1. R. Bernard. *Fundamentals of Biostatistics*. Ed.: Thompson. Chapter 12
2. D. Díez, M Cetinkaya-Rundel and CD Barr. *OpenIntro Statistics*. Chapter 7.

Introduction

- **Generalization** of the $t-$test for $K > 2$ groups

- ANOVA uses a single hypothesis test to check whether the **means across groups are equal**:

  H_0: The mean outcome is the same across all groups $\mu_1 = \mu_2 = \ldots = \mu_K$, where μ_k represents the mean of the outcome for observations in category k.

  H_1: At least one mean is different.

  - Examples
    1. We might like to determine whether there are statistically significant differences in **exam scores** for different lectures of the same course
    2. We might like to determine whether there are statistically significant differences in **time spent on a website** for different customer categories: promoters / neutrals / detractors
    3. We might like to determine whether there are statistically significant differences in **house prices** for different `grade` conditions:

Introduction

- ANOVA Assumptions:

    1. The observations are **independent** within and across groups
    2. The data within each group are **nearly normal**
    3. The **variability** across the groups is about equal

Outline

One-way ANOVA: an example

- When it applies to a **single variable**, as previous examples

Plant Growth

The following table[a] shows the result from an experiment to compare yields (as measured by dried **weight** of plants) obtained under a control and two different treatment conditions:

control	4.17	5.58	5.18	6.11	4.5	4.61	5.17	4.53	5.33	5.14
treatment 1	4.81	4.17	4.41	3.59	5.87	3.83	6.03	4.89	4.32	4.69
treatment 2	6.31	5.12	5.54	5.5	5.37	5.29	4.92	6.15	5.8	5.26

- We might like to determine whether there are statistically significant differences in **mean weight** for the control and the different treatments

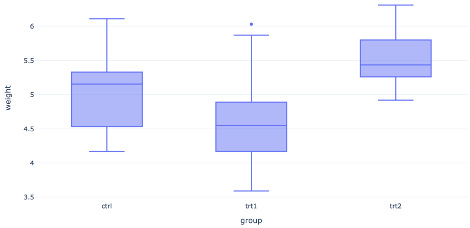
[a] This dataset, among others, can be found *here*

One-way ANOVA: example cont.

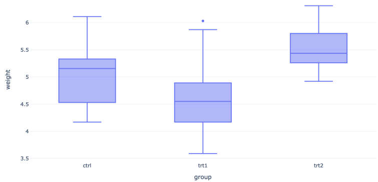
control	4.17	5.58	5.18	6.11	4.5	4.61	5.17	4.53	5.33	5.14
treatment 1	4.81	4.17	4.41	3.59	5.87	3.83	6.03	4.89	4.32	4.69
treatment 2	6.31	5.12	5.54	5.5	5.37	5.29	4.92	6.15	5.8	5.26

- First, let's calculate some summary statistics

	Sample size (n_k)	sample mean ($\bar{x}_k$)	sample (unbiased) SD (s_k)
control	10	5.03	0.58
treatment 1	10	4.66	0.79
treatment 2	10	5.53	0.44



One-way ANOVA: intuition



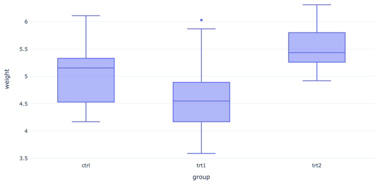
- It tries to answer to the following question: is the variability in the sample means so large that it seems unlikely to be from chance alone?
- For doing this, ANOVA analyzes the **Sum of Squares Total**[1] (SST)

$$\text{SST} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

which is further split into two terms $\boxed{\text{SST} = \text{SSW} + \text{SSB}}$

[1]Recall that the sample (unbiased) variance is calculated as $\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

One-way ANOVA: intuition



$$\boxed{\text{SST} = \text{SSW} + \text{SSB}}$$

1. **Sum of Squares within groups** (SSW):

$$\text{SSW} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_{i,k} - \bar{x}_k)^2$$

2. **Sum of Squares between groups** (SSB):

$$\text{SSB} = \sum_{k=1}^{K} n_k (\bar{x}_k - \bar{x})^2$$

One-way ANOVA: calculations

Plant Growth

$$x = [4.17, 5.58, 5.18, 6.11, 4.5, 4.61, 5.17, 4.53, 5.33, 5.14, \ldots \\ 4.81, 4.17, 4.41, 3.59, 5.87, 3.83, 6.03, 4.89, 4.32, 4.69, \ldots \\ 6.31, 5.12, 5.54, 5.5, 5.37, 5.29, 4.92, 6.15, 5.8, 5.26]$$

where $\bar{x} = 5.073$, so that

$$\text{SST} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{30}(x_i - 5.073)^2 = 14.26$$

	control ($k=1$) $\bar{x}_1 = 5.03$ $x_{i,1} = [4.17, \ldots, 5.14]$	treatment 1 ($k=2$) $\bar{x}_2 = 4.66$ $x_{i,2} = [4.81, \ldots, 4.69]$	treatment 2 ($k=3$) $\bar{x}_3 = 5.53$ $x_{i,3} = [6.31, \ldots, 5.26]$	TOTAL
SSW_k	3.06	5.67	1.76	SSW = **10.49**
SSB_k	0.02	1.7	2.05	SSB = **3.77**
				SST = 14.26

One-way ANOVA: $F-$test

- We need an statistical test to assess if the SSB is big enough to say that there's an statistical difference among group means

$$F = \frac{\text{MSB}}{\text{MSW}} = \frac{\text{SSB}/(K-1)}{\text{SSW}/(n-K)} \sim F_{K-1,n-K} \quad \text{(Snedecor's F distribution)}$$

where

  - MSB stands for the mean square between groups
  - MSW is the mean square within groups
  - K is the number of groups
  - $K-1$ are the degrees of freedom associated with MSB
  - n is the number of observations in the whole dataset
  - $n-K$ are the the degrees of freedom associated with MSW

- Therefore, if the means are further apart, the F statistic is going to increase

Plant Growth

$$F = \frac{\text{SSB}/(K-1)}{\text{SSW}/(n-K)} = \frac{3.77/(3-1)}{10.49/(30-3)} = 4.85$$

One-way ANOVA: Hypothesis testing

- To test the hypothesis

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_K \quad \text{vs} \quad H_1 : \text{at least one mean is different}$$

with a significance level of α

Compute

$$F = \frac{\mathsf{SSB}/(K-1)}{\mathsf{SSW}/(n-K)} \sim F_{K-1,n-K}$$

- if $F > F_{K-1,n-K,1-\alpha}$, then we reject H_0
- if $F \leq F_{K-1,n-K,1-\alpha}$, then we fail to reject H_0

- p − value $= P(F_{K-1,n-K} > F)$

One-way ANOVA: Hypothesis testing

Plant Growth

$$F = \frac{\mathsf{SSB}/(K-1)}{\mathsf{SSW}/(n-K)} = \frac{3.77/(3-1)}{10.49/(30-3)} = 4.85 \sim F_{2,27}$$

- On the one hand, if $\alpha = 0.05$

$$F_{2,27,1-\alpha} = \texttt{f(2,27).ppf(0.95)} = 3.35$$

- On the other hand

$$p = P(F_{2,27} > 4.85) = \texttt{1-f(2,27).cdf(4.85)} = 0.016$$

Therefore

$$\boxed{\text{We reject } H_0 \text{ at a } 5\,\% \text{ confidence level}}$$

One-way ANOVA: Python

Plant Growth

```
f, p = f_oneway(x1, x2, x3)
```

where `x1, x2, x3` refer to the control, treatment 1 and treatment 2 weight values.

```
>>print(f,p)
```

```
4.846087862380136 0.0159099583256229
```

Outline

The ANOVA decomposition

- We can perform ANOVA on linear regression (we did it when defining R^2)
- When fitting the simple linear regression model $\hat{y}_i = \beta_0 + \beta_1 x_i$ to a data set (x_i, y_i) for $i = 1, \ldots, n$, we may identify three sources of variability in the responses

1. **Total variability in the responses**, represented by the Sum of Squares Total[2] (SST)

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

2. **Variability associated to the model**: measures the variation in the responses due to the regression model

$$\text{SSM} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

3. **Variability of the residuals** (previously named as RSS)

$$\text{SSR} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

[2] In Chapter 6 we called it Total Sum of Squares (TSS)

The ANOVA decomposition

- The ANOVA decomposition states[3] that

$$\boxed{\text{SST} = \text{SSM} + \text{SSR}}$$

and therefore

$$\text{R}^2 = \frac{\text{SSM}}{\text{SST}} = 1 - \frac{\text{SSR}}{\text{SST}}$$

which is the proportion of the variation in the responses that is explained by the regression model

[3]Notice that $y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$

The ANOVA table

Wheat production

For the Spanish wheat production data from the 80's with production (X) and price per kilo in pesetas (Y) we have the following table

production	30	28	32	25	25	25	22	24	35	40
price	25	30	27	40	42	40	50	45	30	25

- Fill-in the ANOVA table:

Source of variability	SS	DF	MS	F statistic
Model	SSM	1	SSM$/1$	MSM/MSR
Residual	SSR	$n-1-1$	SSR$/(n-2)$	
Total	SST	$n-1$		

The ANOVA table

Source of variability	SS	DF	MS	F statistic
Model	SSM	1	SSM$/1$	MSM$/$MSR
Residual	SSR	$n-1-1$	SSR$/(n-2)$	
Total	SST	$n-1$		

Source of variability	SS	DF	MS	F statistic
Model	528.47	1	528.47	20.33
Residual	207.93	$10-1-1$	25.99	
Total	736.4	$10-1$		

ANOVA hypothesis testing

- For **one predictor** (as in previous example):

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0$$

- The statistic

$$F = \frac{\mathsf{MSM}}{\mathsf{MSR}} = 20.33 \sim F_{1,n-2}$$

- At a significance level $\alpha = 0.05$,
  - $F_{1,n-2,1-\alpha} = 5.32$
  - $p = P(F_{1,n-2} > 20.33) =$ `1-f(1,8).cdf(20.33)` $= 0.0019$
- Therefore

$$\boxed{\text{We reject } H_0 \text{ at a } 5\,\% \text{ confidence level}}$$

- How does this result relate to the test based on the Student-t we saw in Chapter 6? **They are equivalent**

ANOVA hypothesis testing

- For **more than one** predictors:

$$\hat{y}_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_k x_{i,k}$$

- The ANOVA test would be

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \quad \text{vs} \quad H_1 : \beta_i \neq 0$$

  - If the null hypothesis is true, then **there is no (linear) relationship between the response and predictors**
  - This test evaluates the whole model, no individual coefficients
  - Could be used to evaluate a subset $p < k$ predictors

Outline

ANOVA advanced topics

- ANOVA could be applied also to two-way tables: **two-way ANOVA**
  - To simultaneously evaluate how `grade` and `n_bathrooms` affects the price of a house

- **Kruskall-Wallis** test: It is a non-parametric version of one-way ANOVA
  - Ordinal data
  - The underlying distribution is not normal