

# Bus 274: Further Statistics For Business

Harald Schmidbauer



# About These Slides

- The present slides are not self-contained; they need to be explained and discussed. This will be done in the lectures.
- Even though being a “work in progress” and subject to revision, the slides constitute copyrighted material.  
If you want to reproduce or copy anything from the slides, please ask:

Harald Schmidbauer    **harald** at **hs-stat** dot **com**  
Angi Rösch            **angi** at **angi-stat** dot **com**

- The slides were produced using  $\text{\LaTeX}$  ([www.latex-project.org](http://www.latex-project.org)) and R ([www.R-project.org](http://www.R-project.org)) on a GNU/Linux system, all of which are free and open source software (FOSS).
- R files used for this course are available upon request.



# Chapter 17:

## Analysis of Variance

### (PART — DRAFT VERSION)



# 17.1 Introduction

## The scope of ANOVA.

- ANOVA is a method to test the hypothesis that the means of several normally distributed populations are equal.

- The question

Are averages equal?

can also be asked as:

Where does the variability come from?



# 17.1 Introduction

## The scope of ANOVA.

- Without further information, we can only analyze the variability in a data set, but we cannot explain where it comes from:



- If additional information is given, part of the variability can be accounted for:



# 17.1 Introduction

Example: Used cars.

- Is there a price difference between cars of different colours?
- To find out, we have to compare the means of several (say,  $K$ ) populations:
  - blue cars
  - green cars
  - red cars
  - . . .



# 17.1 Introduction

Example: A course at university.

- A course at university is taught by three lecturers:  
Ayşe, Bürzel, Çağla.
- There are three populations of students.
- Is there a difference between the populations with respect to their average grades?



# 17.1 Introduction

Example: A bank.

- A bank has three branch offices.
- Are average service times of customers the same in the branch offices?



# 17.1 Introduction

Example: A bank.

- A bank has three branch offices.
- Managers have classified the bank customers into two types.
- Are there any differences between customer types / branch offices with respect to average service time?
- How many populations are there in this case?



# 17.2 One-Way ANOVA

## The model.

- Compare  $K$  populations (“groups”, “treatments”) with respect to their means.
- Random variables:

$$X_1, X_2, \dots, X_K, \quad X_i \sim N(\mu_i, \sigma^2), \quad i = 1, 2, \dots, K$$

- This can be written as:

$$X_i = \mu + G_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \quad i = 1, 2, \dots, K,$$

where  $\sum G_i = 0$ .



# 17.2 One-Way ANOVA

The hypotheses.

- Null hypothesis:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_K$$

$$H_0 : G_1 = G_2 = \dots = G_K = 0$$

- Alternative hypothesis:

$$H_1 : \text{There are } i_1, i_2, i_1 \neq i_2, \text{ with } \mu_{i_1} \neq \mu_{i_2}$$

$$H_1 : \text{There is an } i \text{ with } G_i \neq 0$$



# 17.2 One-Way ANOVA

The random sample.

- From each group  $i = 1, \dots, K$ , a random sample of size  $n_i$  is drawn:

$$X_{ij} \sim \text{N}(\mu_i, \sigma^2), \quad i = 1, \dots, K, \quad j = 1, \dots, n_i$$

- This can be written as a “fixed effects model”:

$$X_{ij} = \mu + G_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim \text{N}(0, \sigma^2)$$

- Taking expectations:

$$E(X_{ij}) = \mu + G_i$$



# 17.2 One-Way ANOVA

The random sample.

- Arranging the data in a table:

					group	
1	...	$i$	...	$K$		
$X_{11}$	...	$X_{i1}$	...	$X_{K1}$		
$X_{12}$	...	$X_{i2}$	...	$X_{K2}$		
$\vdots$		$\vdots$		$\vdots$		
$X_{1n_1}$	...	$X_{in_i}$	...	$X_{Kn_K}$		
$\bar{X}_1$	...	$\bar{X}_i$	...	$\bar{X}_K$	$\bar{X}$	

- Sample means:

$$\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} \text{ (group } i), \quad \bar{X} = \frac{1}{n} \sum_{i=1}^K \sum_{j=1}^{n_i} X_{ij} \text{ (overall)}$$



## 17.2 One-Way ANOVA

The “empirical group effect” .

- The empirical analogue to  $E(X_{ij}) = \mu + G_i$  is:

$$\bar{X}_i = \bar{X} + (\bar{X}_i - \bar{X})$$

- The term  $\bar{X}_i - \bar{X}$  measures the group effect.
- How important is the group effect? —  
We use sums of squares to measure variabilities within and between groups.



# 17.2 One-Way ANOVA

Sums of squares.

- within groups: 
$$SSW = \sum_{i=1}^K \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2, \quad \text{df: } n - K$$

- between groups: 
$$SSG = \sum_{i=1}^K n_i (\bar{X}_i - \bar{X})^2, \quad \text{df: } K - 1$$

- total: 
$$SST = \sum_{i=1}^K \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2, \quad \text{df: } n - 1$$



# 17.2 One-Way ANOVA

Decomposition of variance.

- Total variability can be decomposed as follows:

$$\begin{array}{rclcl} \text{SST} & = & \text{SSW} & + & \text{SSG} \\ n - 1 & = & (n - K) & + & (K - 1) \end{array}$$

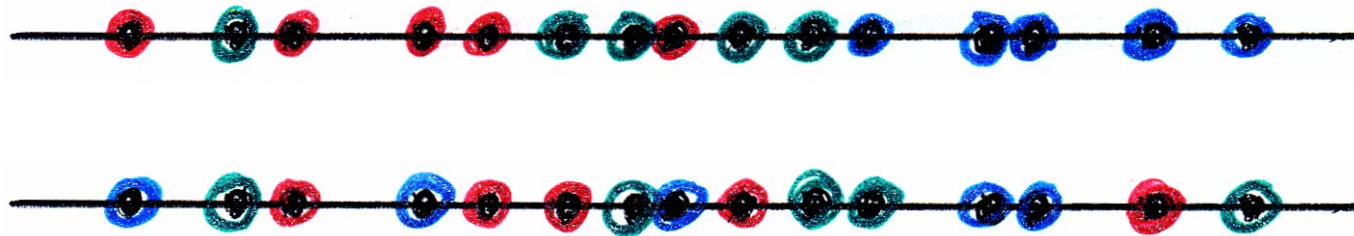
- How can this contribute to testing our hypothesis?



# 17.2 One-Way ANOVA

Decomposition of variance.

- Consider two situations:



- SST is the same in both cases, but it is decomposed into different components.
- $H_0$  might be rejected in the first case.



# 17.2 One-Way ANOVA

Mean sums of squares.

- $MSW = \frac{SSW}{n - K} = \frac{1}{n - K} \sum_{i=1}^K \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2, \quad \text{df: } n - K$

- $MSG = \frac{SSG}{K - 1} = \frac{1}{K - 1} \sum_{i=1}^K n_i (\bar{X}_i - \bar{X})^2, \quad \text{df: } K - 1$

- $MST = \frac{SST}{n - 1} = \frac{1}{n - 1} \sum_{i=1}^K \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2, \quad \text{df: } n - 1$



# 17.2 One-Way ANOVA

Testing the null hypothesis  $H_0 : \mu_1 = \dots = \mu_K$ .

- If  $H_0$  is true,

$$F = \frac{MSG}{MSW} \sim F_{K-1, n-K}$$

- Critical for  $H_0$ : too large values of  $F$ .



# 17.2 One-Way ANOVA

The ANOVA table.

- It contains all relevant measures.
- One-way ANOVA table:

Source of variation	SS	df	MS	$F_{\text{calc}}$	$F_{\text{crit}}$
between groups	SSG	$K - 1$	MSG	MSG/MSW	
within groups	SSW	$n - K$	MSW		
total	SST	$n - 1$			



# 17.3 Two-Way ANOVA

## Introduction.

- Two-way ANOVA assumes that the population is stratified (cross-classified) in two ways:
  - with respect to  $K$  groups,
  - with respect to  $H$  blocks.
- Each combination of group and block is represented by a random variable:

$$X_{ij} \sim \text{N}(\mu_{ij}, \sigma^2), \quad i = 1, \dots, K, \quad j = 1, \dots, H$$

- Example: See handout.



# 17.3 Two-Way ANOVA

The model.

- Random variables:

$$X_{ij} \sim \text{N}(\mu_{ij}, \sigma^2), \quad i = 1, \dots, K, \quad j = 1, \dots, H$$

- This can be written as:

$$X_{ij} = \mu + G_i + B_j + I_{ij} + \epsilon_{ij}, \quad \epsilon_{ij} \sim \text{N}(0, \sigma^2)$$

with  $\sum_i G_i = \sum_j B_j = \sum_{i,j} I_{ij} = 0$ .

- Taking expectations:

$$\mu_{ij} = \mu + G_i + B_j + I_{ij}$$



# 17.3 Two-Way ANOVA

The meaning of  $\mu$ ,  $G_i$ ,  $B_j$ , and  $I_{ij}$ .

- $\mu$  is the overall mean.
- $G_i$  reflects differences between groups.
- $B_j$  reflects differences between blocks.
- $I_{ij}$  reflects interactions between groups and blocks.



# 17.3 Two-Way ANOVA

## The null hypotheses.

The goal of two-way ANOVA is to test three hypotheses, using the same data set:

- $H_0$  : There is no group effect, that is:  $G_i = 0$  for all  $i$ .
- $H_0$  : There is no block effect, that is:  $B_j = 0$  for all  $j$ .
- $H_0$  : There are no interactions, that is:  $I_{ij} = 0$  for all  $i, j$ .



# 17.3 Two-Way ANOVA

Example: Fuel consumption of cars.

- How do driver/car combinations of
  - cars from three car types (the groups) and
  - drivers from three age classes (the blocks)perform w.r.t. fuel consumption?
- The following tables show *assumed* values of  $\mu_{ij}$ ,  $G_i$ ,  $B_j$ , and  $I_{ij}$ .
- We shall see two situations:  
one with and one without interactions.



# 17.3 Two-Way ANOVA

Example: Fuel consumption of cars. Situation I:

	car type			block mean	$B_j$		car type		
	$\alpha$	$\beta$	$\gamma$				$\alpha$	$\beta$	$\gamma$
young	9	10	11	10	1	young	0	0	0
middle	7	8	9	8	-1	middle	0	0	0
old	8	9	10	9	0	old	0	0	0
group mean	8	9	10	9					
$G_i$	-1	0	1						

- There are no interactions in this case.
- For example,

$$\begin{aligned} \mu_{\gamma, \text{old}} &= \mu + G_{\gamma} + B_{\text{old}} + I_{\gamma, \text{old}} \\ 10 &= 9 + 1 + 0 + 0 \end{aligned}$$



# 17.3 Two-Way ANOVA

Example: Fuel consumption of cars. Situation II:

	car type			block mean	$B_j$		car type		
	$\alpha$	$\beta$	$\gamma$				$\alpha$	$\beta$	$\gamma$
young	9	10	11	10	1	young	0	0	0
middle	9	8	7	8	-1	middle	2	0	-2
old	6	9	12	9	0	old	-2	0	2
group mean	8	9	10	9					
$G_i$	-1	0	1						

- Now, interactions exist.
- For example,

$$\begin{aligned} \mu_{\gamma, \text{old}} &= \mu + G_{\gamma} + B_{\text{old}} + I_{\gamma, \text{old}} \\ 12 &= 9 + 1 + 0 + 2 \end{aligned}$$



# 17.3 Two-Way ANOVA

Sample and sample means.

Sample:  $X_{ijl}$ ,  $i = 1, \dots, K$ ,  $j = 1, \dots, H$ ,  $l = 1, \dots, L$

	1	...	$i$	...	$K$	
1	$X_{111}, \dots, X_{11L}$	...	...	...	...	$\bar{X}_{.1.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$j$	...	...	$X_{ij1}, \dots, X_{ijL}$	...	...	$\bar{X}_{.j.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$H$	...	...	...	...	$X_{KH1}, \dots, X_{KHL}$	$\bar{X}_{.H.}$
	$\bar{X}_{1..}$	...	$\bar{X}_{i..}$	...	$\bar{X}_{K..}$	$\bar{X}$

Here,

$$\bar{X}_{i..} = \frac{1}{HL} \sum_{j,l} X_{ijl}, \quad \bar{X}_{.j.} = \frac{1}{KL} \sum_{i,l} X_{ijl}, \quad \bar{X} = \frac{1}{KHL} \sum_{i,j,l} X_{ijl}.$$



# 17.3 Two-Way ANOVA

Parameters and their empirical analogues.

parameter	empirical analogue
$\mu$	$\bar{X}$
$G_i$	$\bar{X}_{i..} - \bar{X}$
$B_j$	$\bar{X}_{.j.} - \bar{X}$
$\mu_{ij}$	$\bar{X}_{ij.}$
$I_{ij}$	$\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}$

To understand the last relationship, observe that

$$I_{ij} = \mu_{ij} - \mu - G_i - B_j$$

$$\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X} = \bar{X}_{ij.} - \bar{X} - (\bar{X}_{i..} - \bar{X}) - (\bar{X}_{.j.} - \bar{X}),$$

because of  $\mu_{ij} = \mu + G_i + B_j + I_{ij}$ .



# 17.3 Two-Way ANOVA

Sums of squares.

between groups:  $SSG = HL \sum_{i=1}^K (\bar{X}_{i..} - \bar{X})^2$

between blocks:  $SSB = KL \sum_{j=1}^H (\bar{X}_{.j.} - \bar{X})^2$

interaction:  $SSI = L \sum_{i=1}^K \sum_{j=1}^H (\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X})^2$

error:  $SSE = \sum_{i=1}^K \sum_{j=1}^H \sum_{l=1}^L (X_{ijl} - \bar{X}_{ij.})^2$

total:  $SST = \sum_{i=1}^K \sum_{j=1}^H \sum_{l=1}^L (X_{ijl} - \bar{X})^2$



# 17.3 Two-Way ANOVA

The ANOVA table.

Two-way ANOVA table:

source of variation	sum of squares	df	mean squares	$F_{\text{obs}}$	$F_{\text{crit}}$
groups	SSG	$K - 1$	MSG	$\frac{\text{MSG}}{\text{MSE}}$	$F_{1-\alpha; K-1, KH(L-1)}$
blocks	SSB	$H - 1$	MSB	$\frac{\text{MSB}}{\text{MSE}}$	$F_{1-\alpha; H-1, KH(L-1)}$
interaction	SSI	$(K - 1)(H - 1)$	MSI	$\frac{\text{MSI}}{\text{MSE}}$	$F_{1-\alpha; (K-1)(H-1), KH(L-1)}$
error	SSE	$KH(L - 1)$	MSE		
total	SST	$KHL - 1$			

