SILESIAN
UNIVERSITY
SCHOOL OF BUSINESS
ADMINISTRATION IN KARVINA

ANOVA

Analysis of Variance – One Way

Ing. Elena Mielcová, Ph.D.
STATISTICAL DATA PROCESSING/NPSTZ

OUTLINE OF THE LECTURE

1. One-way ANOVA
2. Sources of Variability
3. ANOVA Hypothesis
4. ANOVA Test
5. Statistical Software: ANOVA Tests
6. Correlation Ratio

ANOVA

- ANOVA = Analysis of Variance
- ANOVA
  - one of the most frequently used statistical procedures in marketing as well as other areas of data analysis.
  - this method enables one to assess the potential influence of a qualitative or quantitative variable on another quantitative variable.
- Example of ANOVA use:
  - to evaluate effects of different forms of a promotional campaign on the sales of a product. In this case, different promotional campaigns resent different categories of the observed qualitative variable = promotional campaign. The sales are then the quantitative variable in question.

ANOVA

Effects of factors:

- potential effect can be expressed mathematically in such a way that the expression analyses whether a change in the level of the qualitative/quantitative variable changes the population mean of the other observed quantitative variable.
- In this sense, ANOVA tests if there are any differences among the population means of the quantitative variable.
- ANOVA is based on decomposition of what is called the **total variability of the observed variable**.
- Depending on how many main sources or **factors** appear in the decomposition, we talk about one-way ANOVA, two-way ANOVA and so on.

One – Way ANOVA

Factor level	Data sample for the factor level	Sample size	Sample mean	Sample variance
1	$y_{11}, y_{12}, \ldots, y_{1j}, \ldots, y_{1n_1}$	n_1	$\bar{y}_1$	s_1^2
2	$y_{21}, y_{22}, \ldots, y_{2j}, \ldots, y_{2n_2}$	n_2	$\bar{y}_2$	s_2^2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	$y_{i1}, y_{i2}, \ldots, y_{ij}, \ldots, y_{in_i}$	n_i	$\bar{y}_i$	s_i^2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	$y_{k1}, y_{k2}, \ldots, y_{kj}, \ldots, y_{kn_k}$	n_k	$\bar{y}_k$	s_k^2
Total		N	$\bar{y}$	s^2

One – Way ANOVA

- Statistical test
- The main principle of the analysis of variance is to decompose the total variability of the observed variable.
- The total variability, measured by the sum of squared deviations of the individual values of the variable from their average, is divided by the decomposition into a part that reflects a variability within the samples and a part which reflects a variability between the samples.

Sources of Variability

- Total sum of squares (total variation):

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

- Where

$$\bar{y} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}}{N}$$

Sources of Variability

- Within-group variation, also called residual variation:

$$S_{yv} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

- Where

$$\bar{y}_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$$

Sources of Variability

- The between group variation:

$$S_{ym} = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2$$

Sources of Variability

- Using algebraic operations, the following fundamental formula for the one-way analysis of variance can be derived:

$$S_y = S_{yv} + S_{ym}$$

ANOVA Hypothesis

- **Analysis of variance is a statistical test.**
  - Therefore, we work with a pair of hypotheses: a null hypothesis and an alternative hypothesis.
- ANOVA has its conditions under which it was derived:
  - The method assumes that each of the *k* random samples comes from a normal distribution, and the distributions have the same variance.
  - Also, the samples were drawn independently of each other. In analysis of variance, more than two samples are usually worked with, and

ANOVA Hypothesis

- ANOVA test is trying to find an answer to the question if all *k* samples came from the same populations. In other words, whether the populations the samples came from have the same means (in other approach if there is no factor effect present).

- Null hypothesis:
- H_0: $\mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$ (factor has no effect)

- Alternative hypothesis:
- H_1:negation of H_0 (effect of factor is significant)

ANOVA – Test Criterion

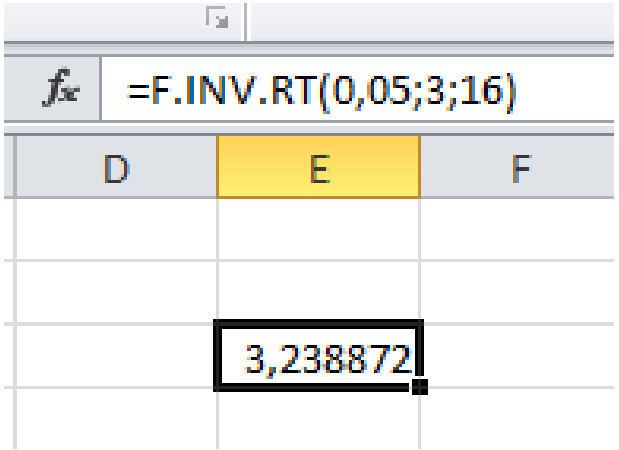
- Test criterion:

- $$T = \frac{\frac{S_{ym}}{k-1}}{\frac{S_{yv}}{N-k}}$$

- follows a **Fisher's distribution** with k-1 and N-k degrees of freedom.

ANOVA – Critical Value

- The critical value of the test:
- $F_{critical} = F_{k-1,N-k}(\alpha)$
- The critical value is tabulated in statistical tables, exact values can be calculated via statistical programs, for example $F_{k-1,N-k}(\alpha)$ can be obtained with the Excel function F.INV.RT(alpha , *k*-1, *N*-*k*) or in older versions of Excel by function FINV(alpha , *k*-1, *N*-*k*):



ANOVA - Results

- Results are taken from comparison of the test value T with the critical value $F_{critical}$.

- If $T < F_{critical}$, then the null hypothesis H_0 is accepted and the factor X can be pronounced not influential in relation to
- the variable Y.

- If $T \geq F_{critical}$, then the null hypothesis H_0 is rejected, meaning the factor X has a statistically significant influence on the variable Y.

ANOVA - Results

- If the test confirms that the factor *X* affects *Y*, we may ask which population means are different.

- It can be the case that only population means are different, while all the other population means are the same.

- There are methods that try to answer this question, one of them being devised by Scheffé and one by Tukey.

ANOVA – Alternative way of calculation of SS

•These formulas are more convenient if the variabilities are to be calculated on a calculator:

$$S_y = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{1}{N}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}\right)^2,$$

$$S_{y,m} = \sum_{i=1}^{k} n_i \bar{y}_i^2 - \frac{1}{N}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}\right)^2,$$

$$S_{y,v} = S_y - S_{y,m}.$$

ANOVA: Typical computer outcome:

Typical computer outcome of one-way ANOVA analysis consists of an ANOVA table with these components:

Source of Variation	Sum of Squares	Degree of Freedom	Mean Squares	Test Criterion	p-value	Critical Value
Between groups	S_{ym}	$k-1$	$\frac{S_{ym}}{k-1}$	$\frac{\frac{S_{ym}}{k-1}}{\frac{S_{yv}}{N-k}}$	p-value	$F_{k-1,N-k}(\alpha)$
Within groups	S_{yv}	$N-k$	$\frac{S_{yv}}{N-k}$			
Total	S_y	$N-1$				

Result can be seen from p-value or from comparison of calculated test criterion with provided critical value.

Example

•The following table contains data obtained through several independent random samplings. The observed factor is the number of octanes used to describe the quality of car fuel (90, 91, 95, 98 octanes are usually available). Thus, the factor is monitored at four possible levels. For each of the levels, five car drivers using the fuel of the corresponding quality were randomly selected. In this case, all samples have the same size, which is not required for one-way ANOVA. We want to know whether the quality of the fuel affects fuel consumption (car mileage).

Example

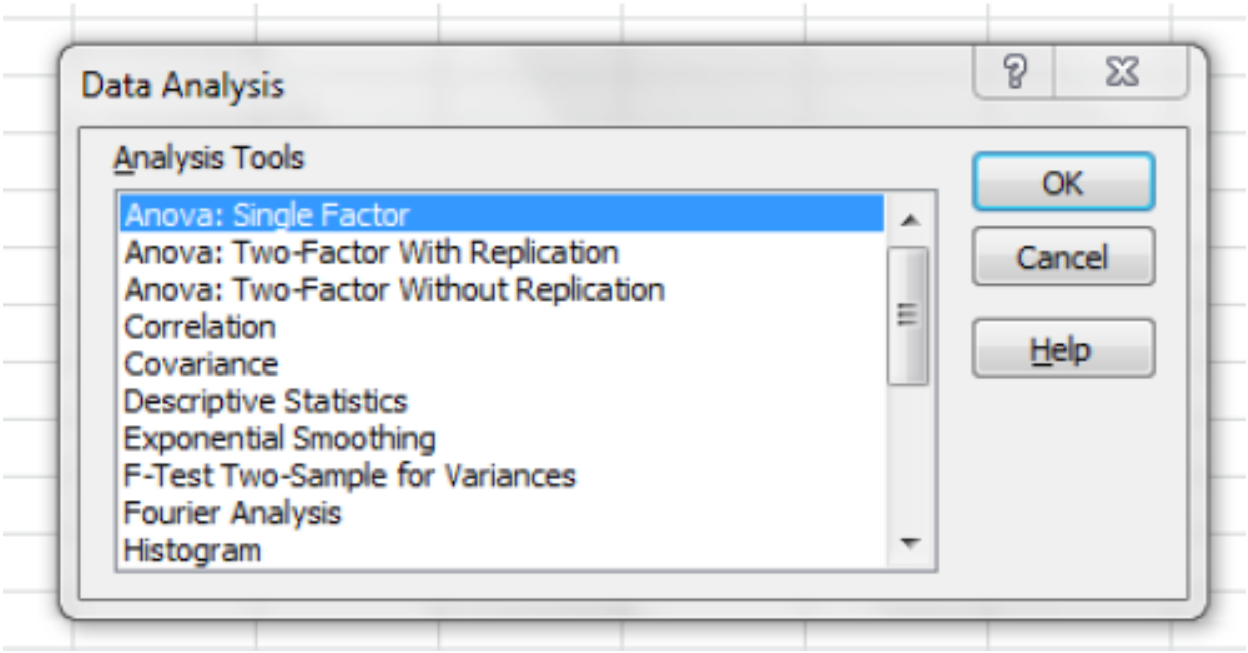
- Data:

Car mileage for different types of fuel

Factor levels	90	91	95	98
	8,1	7,7	7,6	7,5
	8	7,8	7,6	7,8
Samples	7,9	7,9	7,5	7,6
	7,8	7,6	7,6	7,5
	8,2	7,8	7,6	7,5

Example - Solution

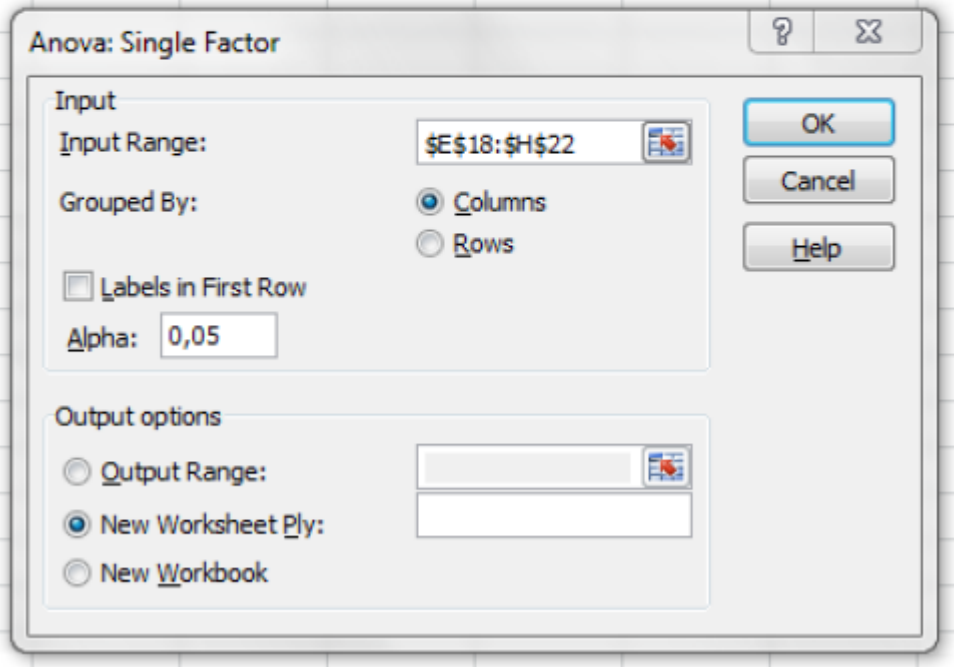
- Calculation via Excel – Data Analysis tool:



Example - Solution

•In the dialogue window, it is necessary to insert as the Input Range a reference to the area of the Excel spreadsheet that contains the data samples to be worked with in ANOVA:

	Factor levels			
	8,1	7,7	7,6	7,5
	8	7,8	7,6	7,8
Samples	7,9	7,9	7,5	7,6
	7,8	7,6	7,6	7,5
	8,2	7,8	7,6	7,5



Example - Solution

- Excel results are given in a form of table:

ANOVA

Source of variability	*SS*	*Difference*	*MS*	*F*	*P-value*	*F krit*
Among-groups	0,594	3	0,198	13,89474	0,0001015	3,238872
Within-groups	0,228	16	0,01425			
Total	0,822	19				

- Since the test criterion is greater than critical value, we reject the null hypothesis that fuel quality has no effect on car mileage.
- In other words, it seems the factor does have an influence on car mileage.

A Measure of Dependence

- Variability of $\bar{y}_t$ around $\bar{y}$ is caused by a dependence of *Y* on *X*. We described such variability with the among-group sum of squares S_{ym}.
- The within-group variability $S_{y,v}$ is induced by factors other than *X*.
- Higher S_{ym} implies a stronger dependence of *Y* on *X*.
- This type of dependence can be measured using the *determination ratio*, denoted P^2:

$$P^2 = \frac{S_{ym}}{S_y}$$

- The square root of P^2 is called the correlation ratio:

$$P = \sqrt{P^2} = \sqrt{\frac{S_{ym}}{S_y}}$$

A Measure of Dependence

- Determination ratio P^2 can take on any value from interval [0,1].

- The stronger the dependence of *Y* on *X*, the closer the characteristic is to one, and the closer the among-group sum of squares is to the total sum of squares (total variability). Under such condition the within-group variability approaches zero.

- The closer the determination ratio is to zero, the smaller the part of the total variability which is accounted for by the among-group variability. In this case, the dependence of *Y* on *X* is weak.

Reading List

- GUJARATI, D., 2009 . *Essentials of Econometrics – Appendix C*
- TOŠENOVSKÝ, F., 2014. *Statistical Methods for Economists – Chapter 6*

Next Lecture:

- Two-Way ANOVA and Latin Squares

THANK YOU