

STATISTICAL DATA ANALYSIS IN EXCEL

Lecture 4

Analysis of Variance (ANOVA)

dr. Petr Nazarov

petr.nazarov@crp-sante.lu

31-10-2011

Means for more than 2 populations

We have measurements for 5 conditions. Are the means for these conditions equal?

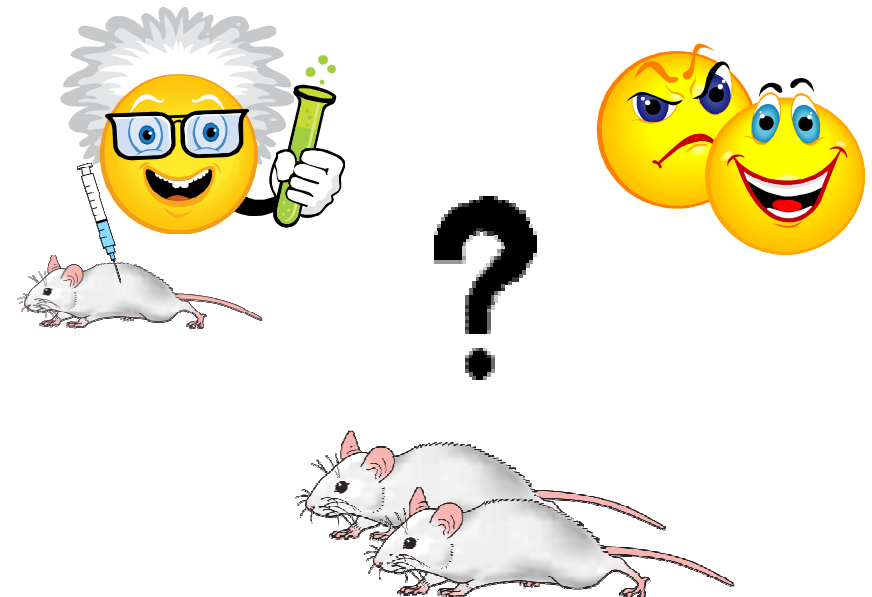
If we would use pairwise comparisons, what will be the probability of getting error?

Number of comparisons: $C_2^5 = \frac{5!}{2!3!} = 10$

Probability of an error: $1 - (0.95)^{10} = 0.4$

Validation of the effects

We assume that we have several factors affecting our data. Which factors are most significant? Which can be neglected?



ANOVA
example from Partek™

As part of a long-term study of individuals 65 years of age or older, sociologists and physicians at the Wentworth Medical Center in upstate New York investigated the relationship between geographic location and depression. A sample of 60 individuals, all in reasonably good health, was selected; 20 individuals were residents of Florida, 20 were residents of New York, and 20 were residents of North Carolina. Each of the individuals sampled was given a standardized test to measure depression. The data collected follow; higher test scores indicate higher levels of depression.

Q: Is the depression level same in all 3 locations?

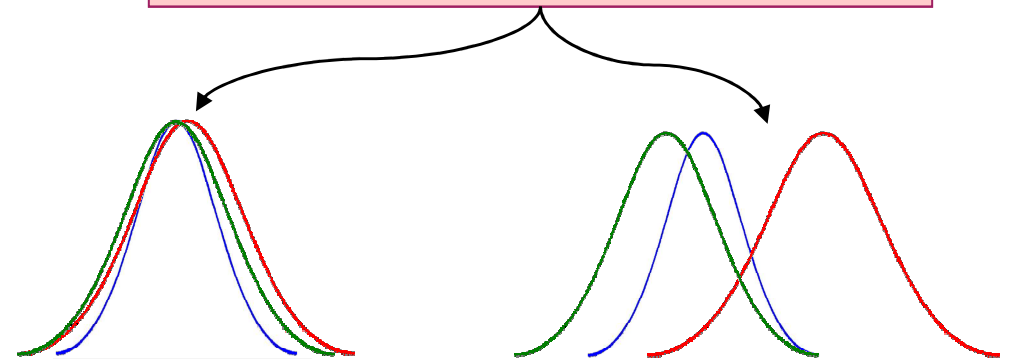
depression.xls

1. Good health respondents

Florida	New York	N. Carolina
3	8	10
7	11	7
7	9	3
3	7	5
8	8	11
8	7	8
...

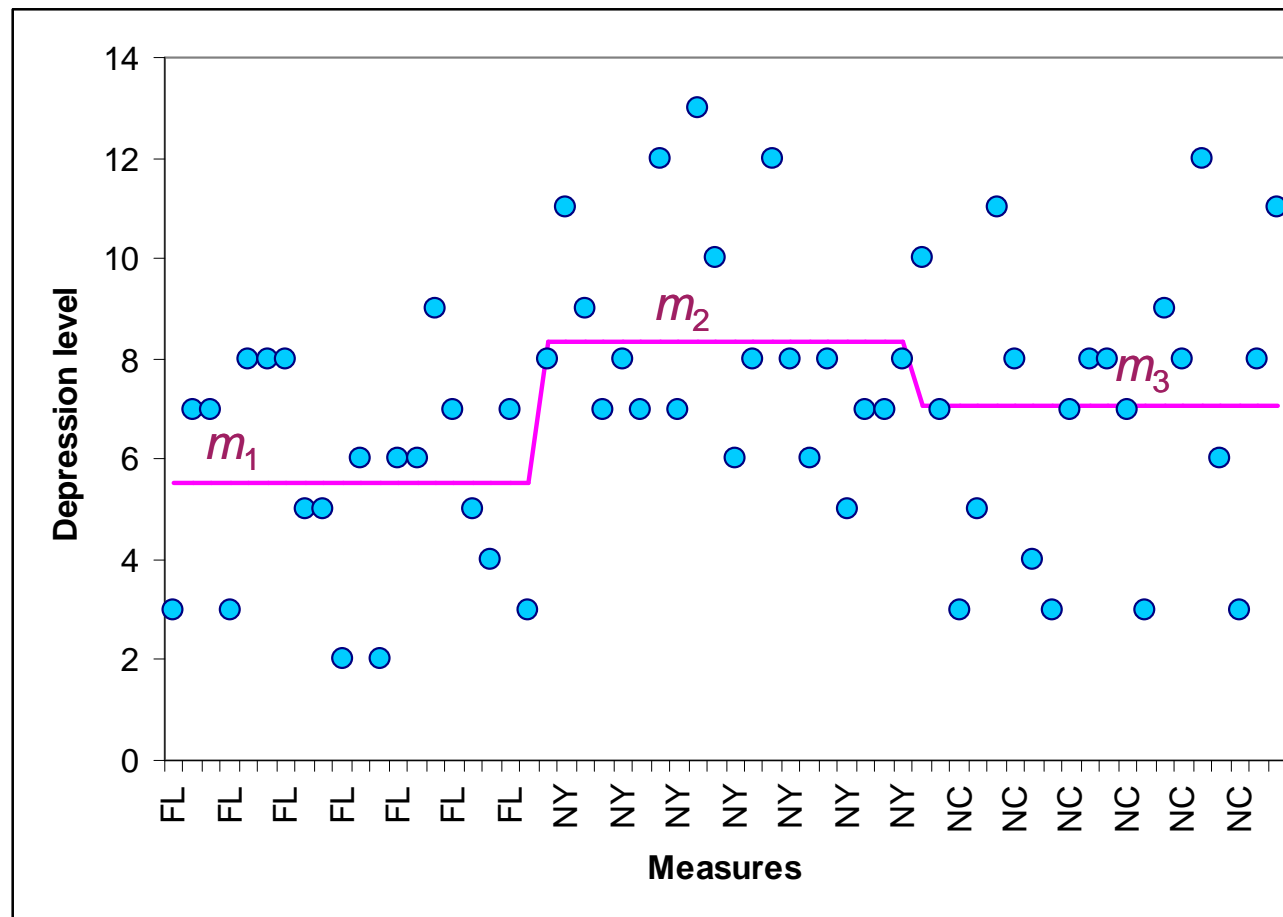
$$H_0: \mu_1 = \mu_2 = \mu_3$$

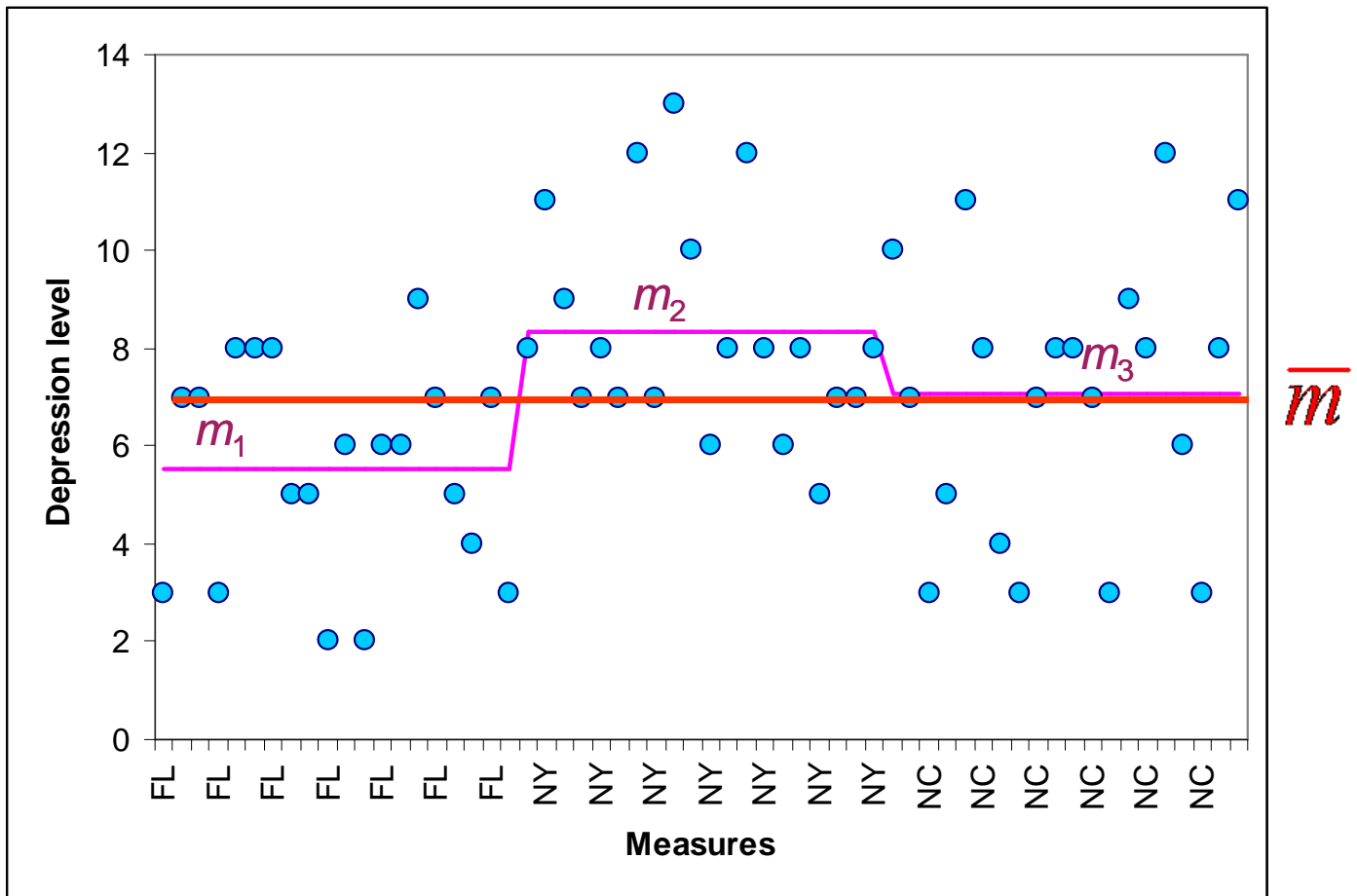
$$H_a: \text{not all 3 means are equal}$$



$$H_0: \mu_1 = \mu_2 = \mu_3$$

H_a : not all 3 means are equal





$$SST = SSTR + SSE$$

ANOVA table

A table used to summarize the analysis of variance computations and results. It contains columns showing the source of variation, the sum of squares, the degrees of freedom, the mean square, and the F value(s).

In Excel use:

◆ Tools → Data Analysis → ANOVA Single Factor

depression.xls

Let's perform for dataset 1: "good health"

SSTR						
ANOVA						
<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	78.53333	2	39.26667	6.773188	0.002296	3.158843
Within Groups	330.45	57	5.797368			
Total	408.9833	59				

SSE

Factor

Another word for the independent variable of interest.

Factorial experiment

An experimental design that allows statistical conclusions about two or more factors.

Treatments

Different levels of a factor.

depression.xls

Factor 1: Health

good health

bad health

Factor 2: Location

Florida

New York

North Carolina

$$\text{Depression} = \mu + \text{Health} + \text{Location} + \text{Health} \times \text{Location} + \varepsilon$$

Interaction

The effect produced when the levels of one factor interact with the levels of another factor in influencing the response variable.

2-factor ANOVA with r Replicates: Example

depression.xls

Factor 1: Health

Factor 2: Location

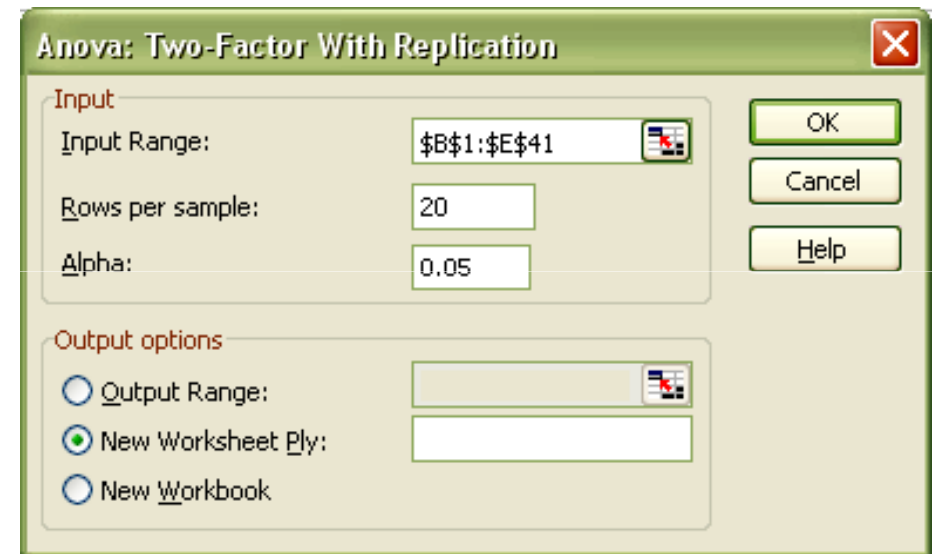
1. Reorder the data into format understandable for Excel

	Florida	New York	North Carolina
Good health	3	8	10
	7	11	7
	7	9	3
	3	7	5

	7	7	8
	3	8	11
bad health	13	14	10
	12	9	12
	17	15	15
	17	12	18

	11	13	13
	17	11	11

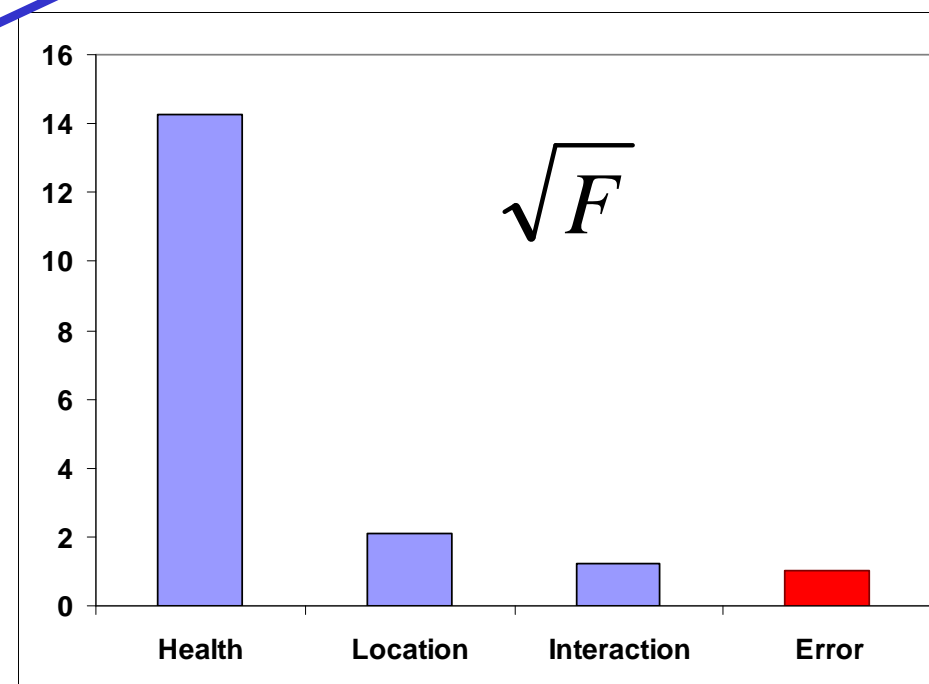
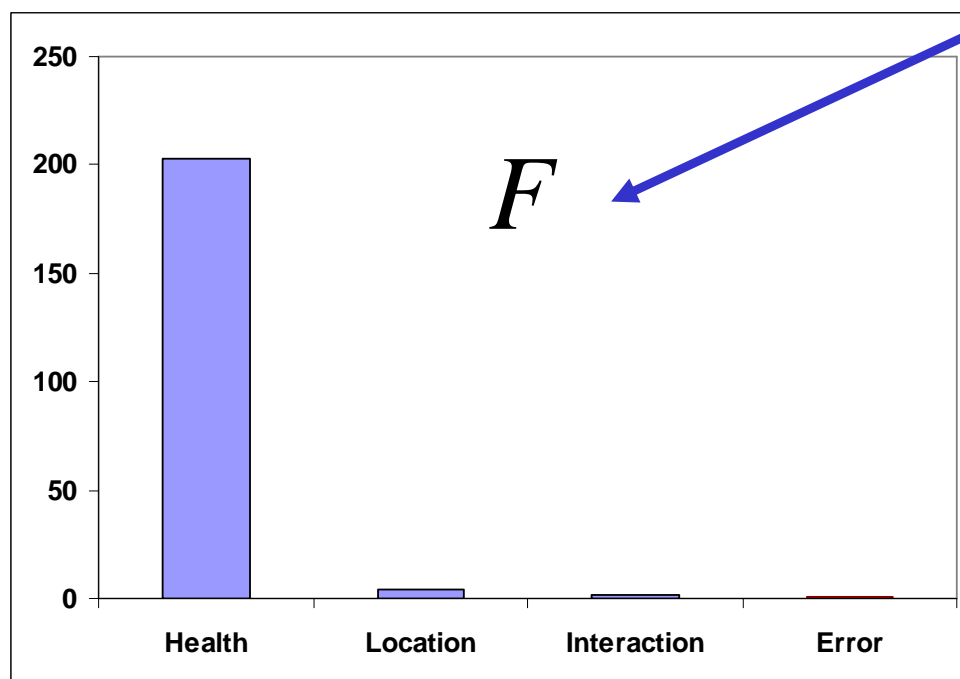
2. Use Tools → Data Analysis → ANOVA: Two-factor with replicates



2-factor ANOVA with r Replicates: Example

ANOVA

	Source of Variation	SS	df	MS	F	P-value	F crit
Health Location Interaction Error	Sample	1748.033	1	1748.033	203.094	4.4E-27	3.92433
	Columns	73.85	2	36.925	4.290104	0.015981	3.075853
	Interaction	26.11667	2	13.05833	1.517173	0.223726	3.075853
	Within	981.2	114	8.607018			
	Total	2829.2	119				



Thank you for your attention

to be continued...

