CENTRE DE RECHERCHE PUBLIC

Microarray Center

APPLIED STATISTICS

Lecture 9

Analysis of Variance (ANOVA)

Petr Nazarov
petr.nazarov@crp-sante.lu

25-11-2009

- **Introduction to ANOVA**
  - why ANOVA
  - shoe experiment
  - assumptions with ANOVA
- **Single-factor ANOVA**
  - theory and application
  - ANOVA table
- **Multi-factor ANOVA**
  - theory and applications
  - factor effects
- **Experimental design**
  - randomized design
  - block design

Why ANOVA?

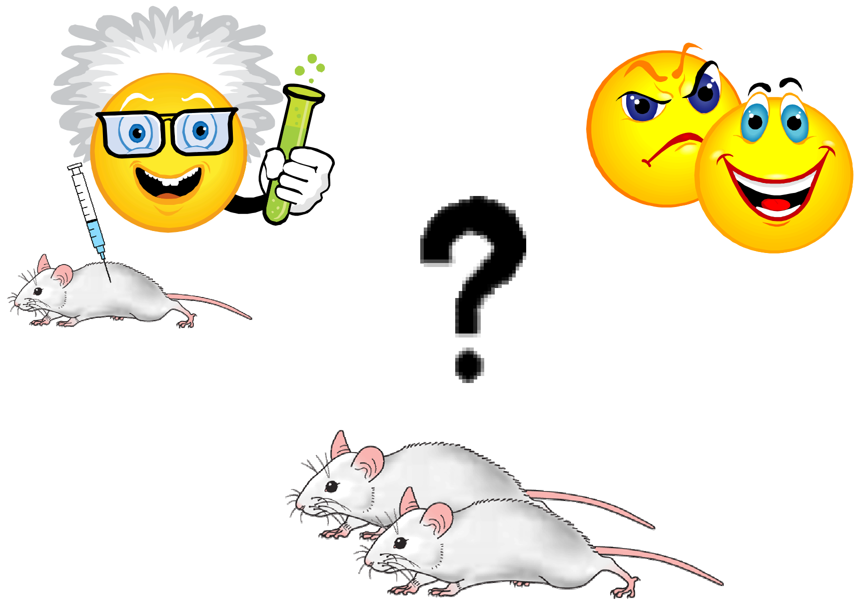
Means for more than 2 populations
We have measurements for 5 conditions. Are the means for these conditions equal?

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 more significant? Which can be neglected?



ANOVA example from Partek™

http://easylink.playstream.com/affymetrix/ambsymposium/partek_08.wvx

Example from Case Problem 3

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?

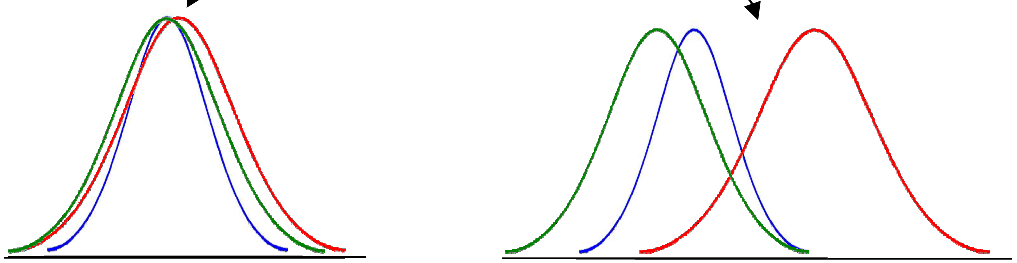
`depression.xls`

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
...	...	...

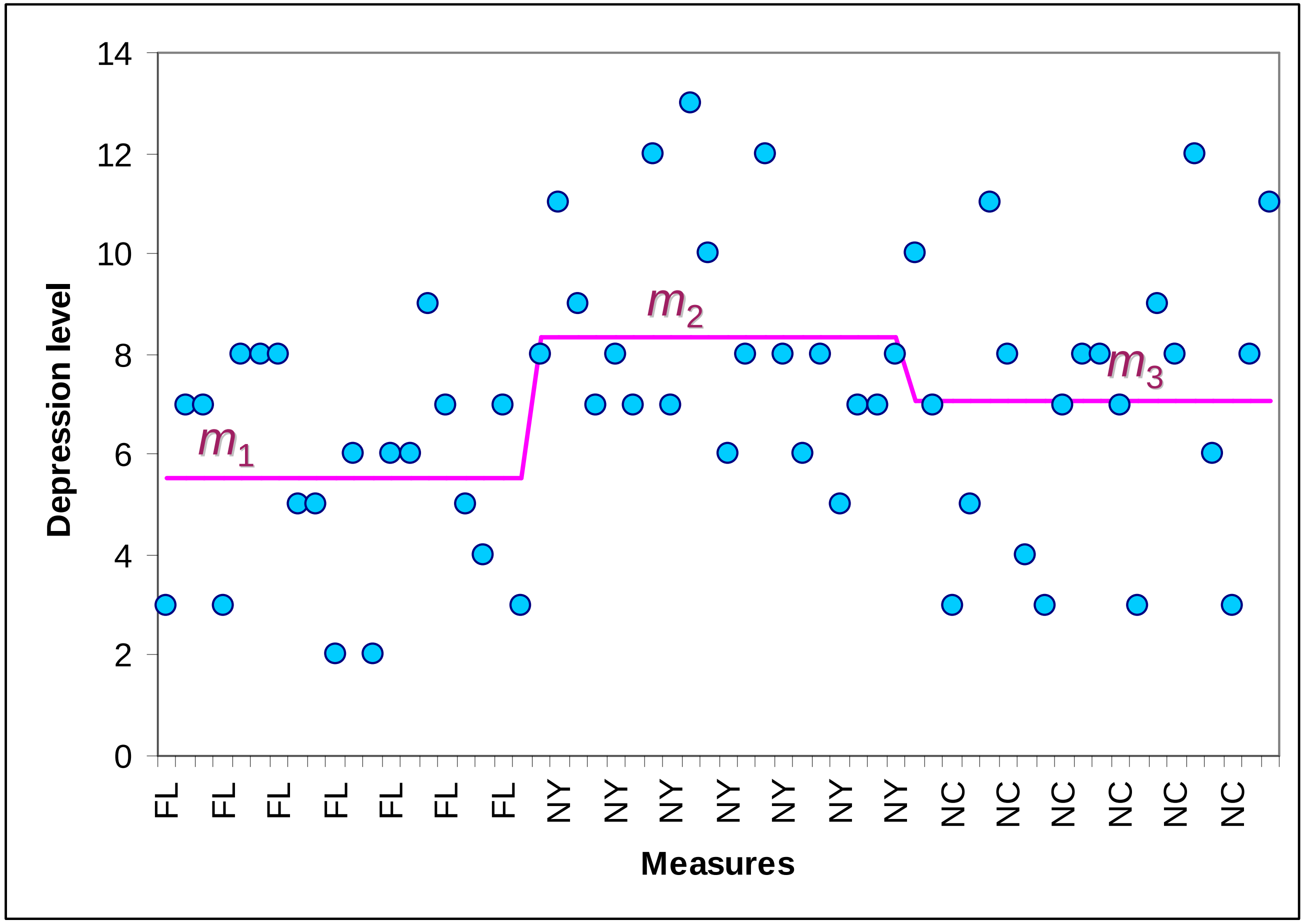
H_0: $\mu_1 = \mu_2 = \mu_3$

H_a: not all 3 means are equal



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

H_a: not all 3 means are equal

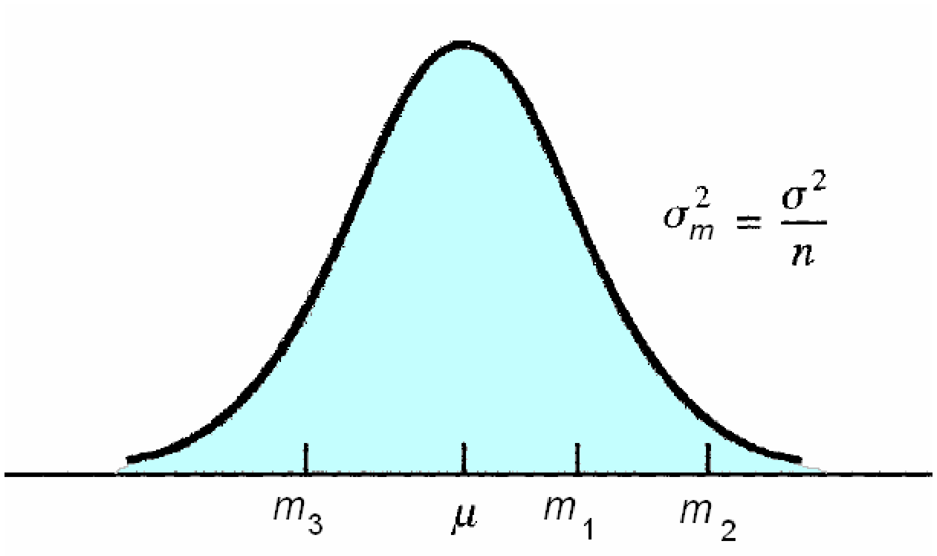


Assumptions for Analysis of Variance

1. For each population, the response variable is **normally distributed**

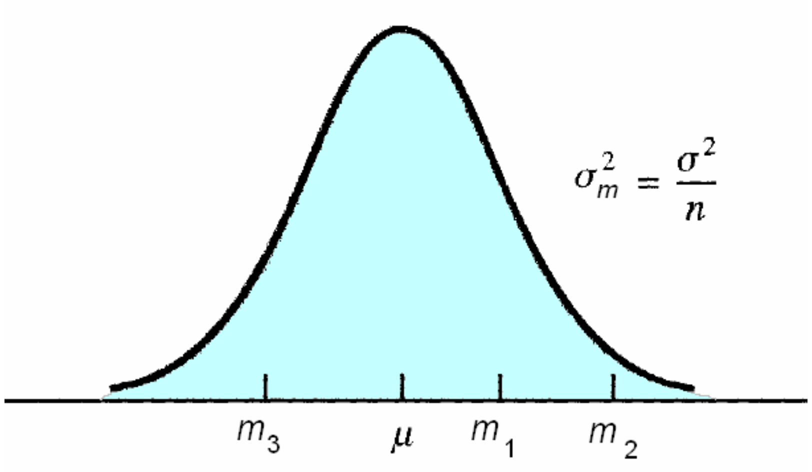
2. The variance of the respond variable, denoted as **σ^2 is the same** for all of the populations.

3. The observations must be **independent.**



Example from Case Problem 3

Parameter	Florida	New York	N. Carolina
m=	5.55	8.35	7.05
overall mean=	6.98333		
var=	4.5763	4.7658	8.0500



Let's estimate the variance of sampling distribution. If H_0 is true, then all m_i belong to the same distribution

$$\sigma_m^2 = \frac{\sum_{i=1}^{k}(m_i - \overline{m})^2}{k-1} = \frac{(5.55-6.98)^2 + (8.35-6.98)^2 + (7.05-6.98)^2}{3-1} = 1.96$$

$$\boxed{\sigma^2 = n\sigma_m^2 = 20 \times 1.96 = 39.27}$$

– this is called between-treatment estimate, works only at H_0

At the same time, we can estimate the variance just by averaging out variances for each populations:

$$\boxed{\sigma^2 = \frac{\sum_{i=1}^{k}\sigma_i}{k} = \frac{4.58+4.77+8.05}{3} = 5.8}$$

– this is called within-treatment estimate

Does between-treatment estimate and within-treatment estimate give variances of the same "population"?

H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$

H_a: not all k means are equal

Means for treatments

$$m_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j}$$

Variances treatments

$$s_j^2 = \frac{\sum_{i=1}^{n_j} \left(x_{ij} - m_j\right)^2}{n_j - 1}$$

Total mean

$$\overline{m} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}}{n_T}$$

$$n_T = n_1 + n_2 + \cdots + n_k$$

due to treatment

Sum squares

$$SSTR = \sum_{j=1}^{k} n_j \left(m_j - \overline{m}\right)^2$$

Mean squares, $\sigma_{beetween}^2$

$$MSTR = \frac{SSTR}{k-1}$$

due to error

Sum squares

$$SSE = \sum_{j=1}^{k} \left(n_j - 1\right) s_j^2$$

Mean squares, σ_{within}^2

$$MSE = \frac{SSE}{n_r - k}$$

Test of variance equality

$$F = \frac{MSE}{SSTR}$$

p-value for the treatment effect

$$p - value$$

The Main Equation

SS due to treatment

$$SSTR = \sum_{j=1}^{k} n_j \left(m_j - \overline{m}\right)^2$$

Total sum squares

$$SST = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(x_{ij} - \overline{m}\right)^2$$

$$SST = SSTR + SSE$$

SS due to error

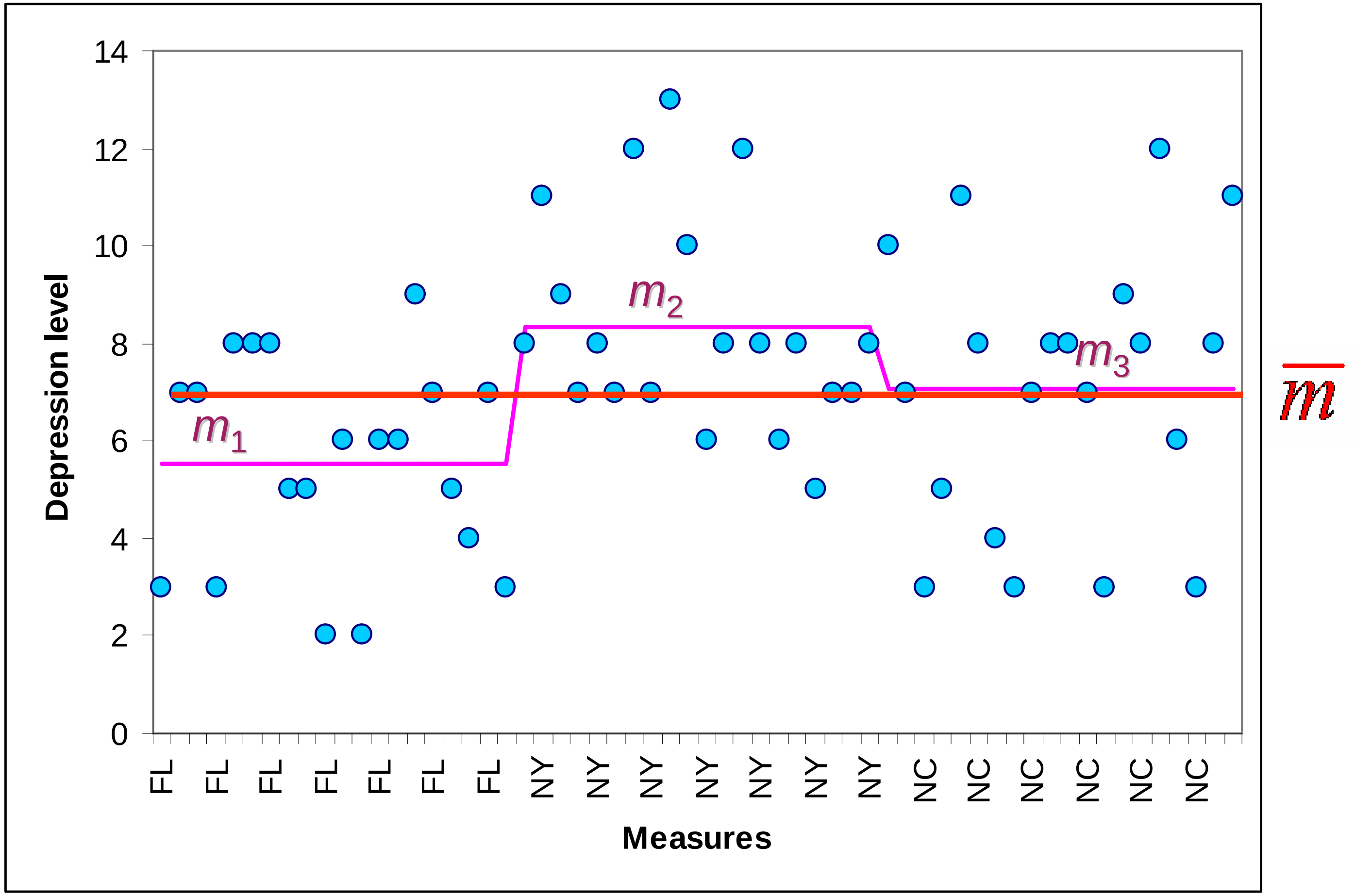
$$SSE = \sum_{j=1}^{k} \left(n_j - 1\right) s_j^2$$

Total variability of the data include variability due to treatment and variability due to error

$$d.f.(SST) = d.f.(SSTR) + d.f.(SSE)$$
$$n_T - 1 = (k - 1) + (n_T - k)$$

Partitioning
The process of allocating the total sum of squares and degrees of freedom to the various components.

Example



$$SST = SSTR + SSE$$

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

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	78.53333	2	39.26667	6.773188	$\boxed{0.002296}$	3.158843
Within Groups	330.45	57	5.797368			
Total	408.9833	59				

SSE

Factors and Treatments

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.

ANOVA example from Partek™

2-factor ANOVA with *r* Replicates

Replications
The number of times each experimental condition is repeated in an experiment.

a = number of levels of factor A
b = number of levels of factor B
r = number of replications
n_T = total number of observations taken in the experiment; $n_T = abr$

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Factor A	SSA	$a - 1$	$\text{MSA} = \dfrac{\text{SSA}}{a - 1}$	$\dfrac{\text{MSA}}{\text{MSE}}$
Factor B	SSB	$b - 1$	$\text{MSB} = \dfrac{\text{SSB}}{b - 1}$	$\dfrac{\text{MSB}}{\text{MSE}}$
Interaction	SSAB	$(a - 1)(b - 1)$	$\text{MSAB} = \dfrac{\text{SSAB}}{(a - 1)(b - 1)}$	$\dfrac{\text{MSAB}}{\text{MSE}}$
Error	SSE	$ab(r - 1)$	$\text{MSE} = \dfrac{\text{SSE}}{ab(r - 1)}$	
Total	SST	$n_T - 1$		

2-factor ANOVA with *r* Replicates: Example

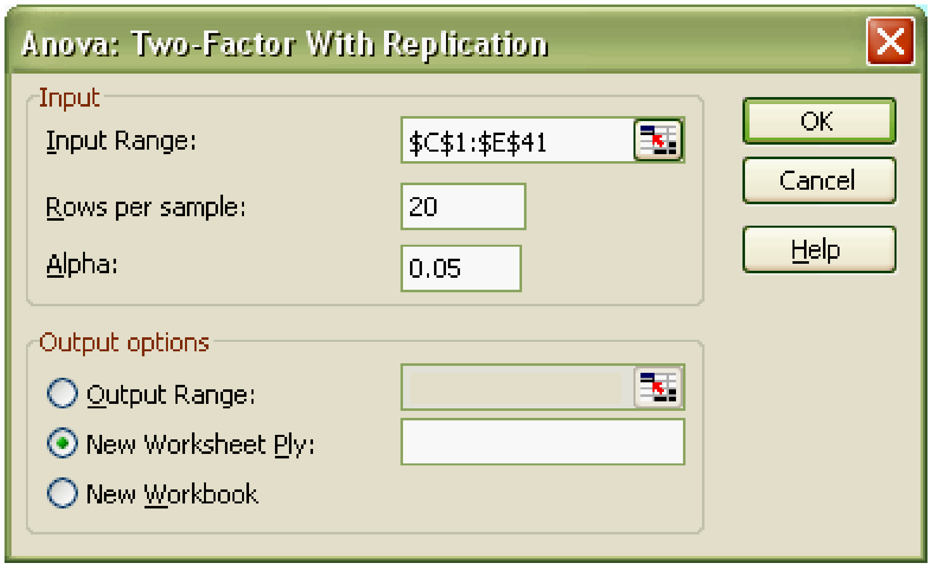
depression.xls

- **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**

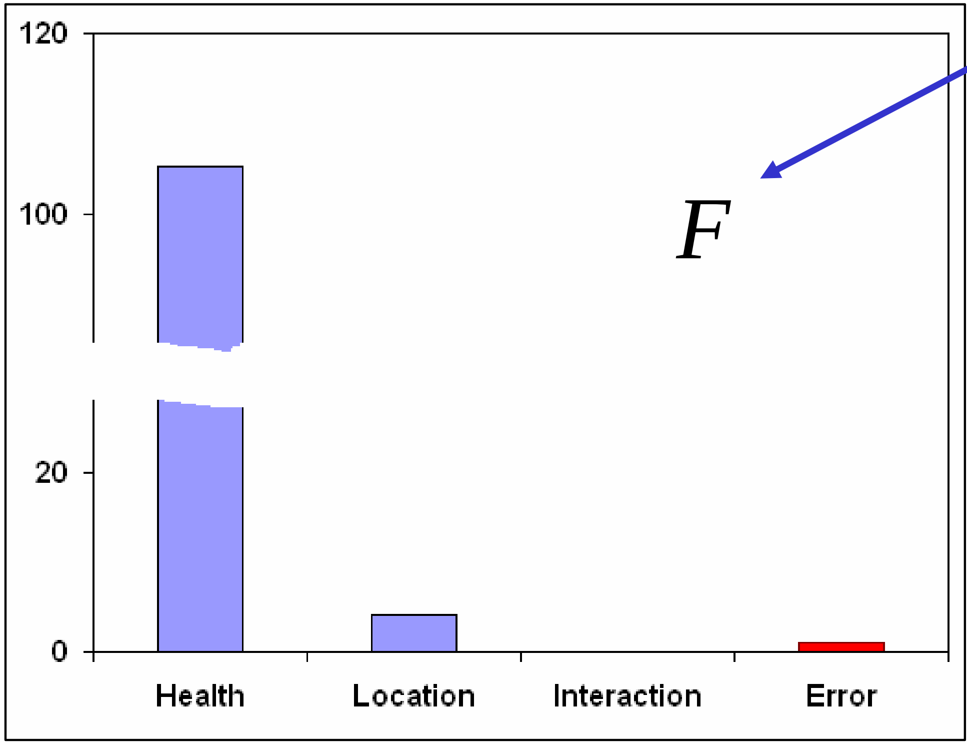
2. Use Tools → Data Analysis → ANOVA: Two-factor with replicates

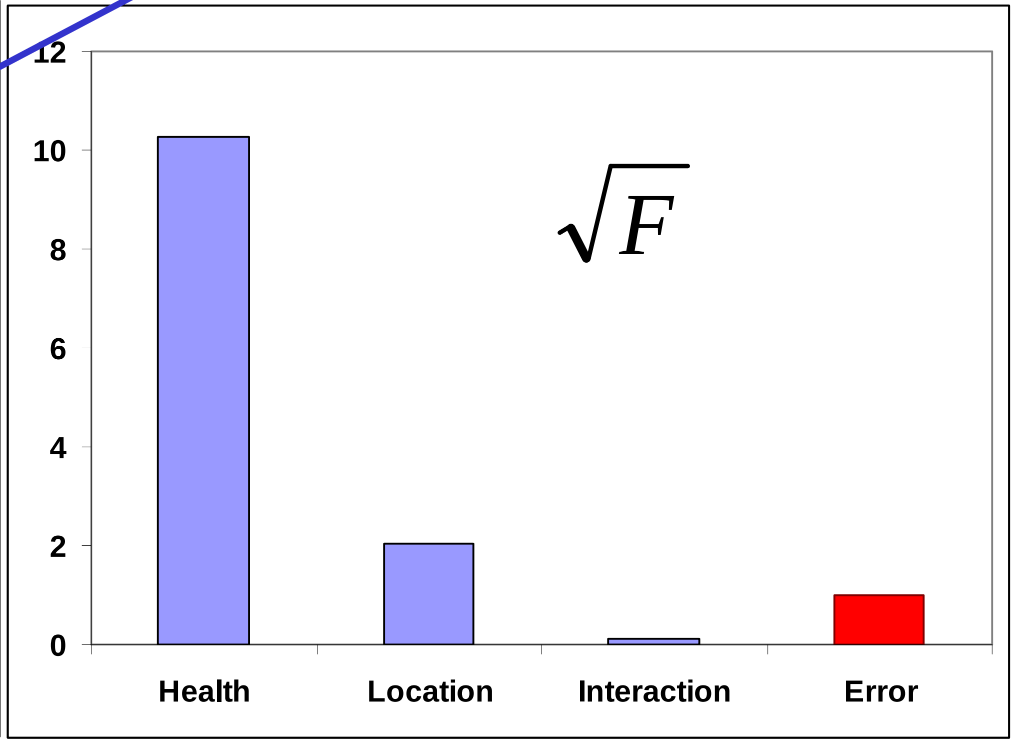


2-factor ANOVA with *r* Replicates: Example

ANOVA

	Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Health	Sample	973.0125	1	973.0125	105.1531	5.49E-16	3.96676
Location	Columns	37.8125	1	37.8125	4.086385	0.046751	3.96676
Interaction	Interaction	0.1125	1	0.1125	0.012158	0.912492	3.96676
Error	Within	703.25	76	9.253289			
	Total	1714.188	79				





Example 2

salaries.xls

Salary/week	Occupation	Gender
872	Financial Manager	Male
859	Financial Manager	Male
1028	Financial Manager	Male
1117	Financial Manager	Male
1019	Financial Manager	Male
519	Financial Manager	Female
702	Financial Manager	Female
805	Financial Manager	Female
558	Financial Manager	Female
591	Financial Manager	Female

	Ocupation		
Sex	Financial Manage	Computer Programme	Pharmacist
Male	872	747	1105
	859	766	1144
	1028	901	1085
	1117	690	903
	1019	881	998
Female	519	884	813
	702	765	985
	805	685	1006
	558	700	1034
	591	671	817

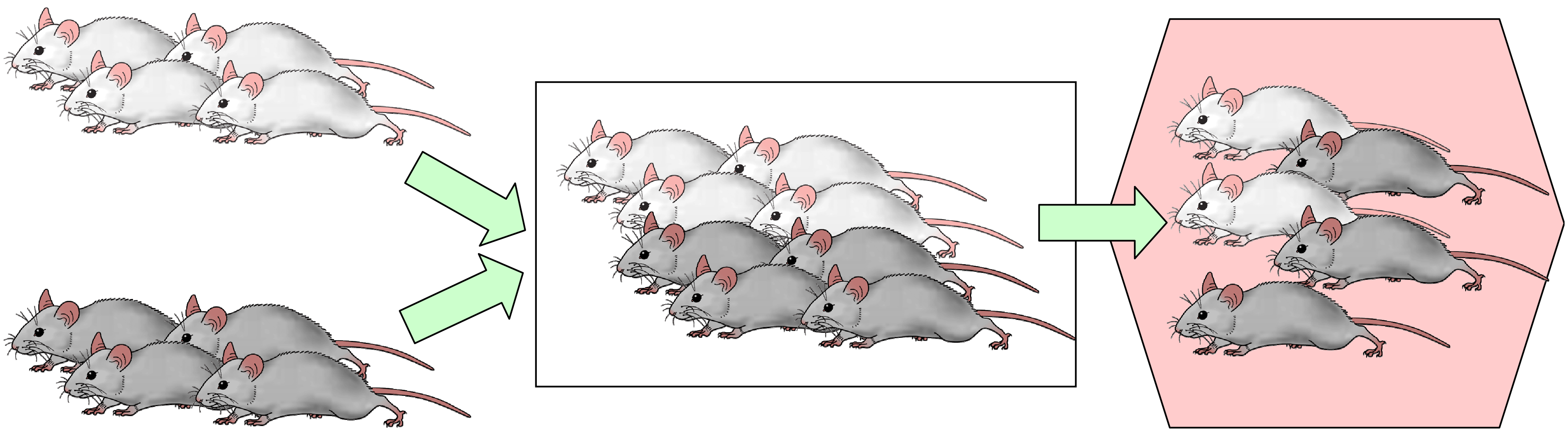
Q: Which factors have significant effect on the salary

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	36980	1	36980	4.0265	0.062	4.494
Columns	242000	1	242000	26.349	0.0001	4.494
Interaction	4500	1	4500	0.49	0.49399	4.494
Within	146948	16	9184.25			
Total	430428	19				

Tools → Data Analysis → ANOVA: Two-factor with replicates

Completely randomized design
An experimental design in which the treatments are randomly assigned to the experimental units.



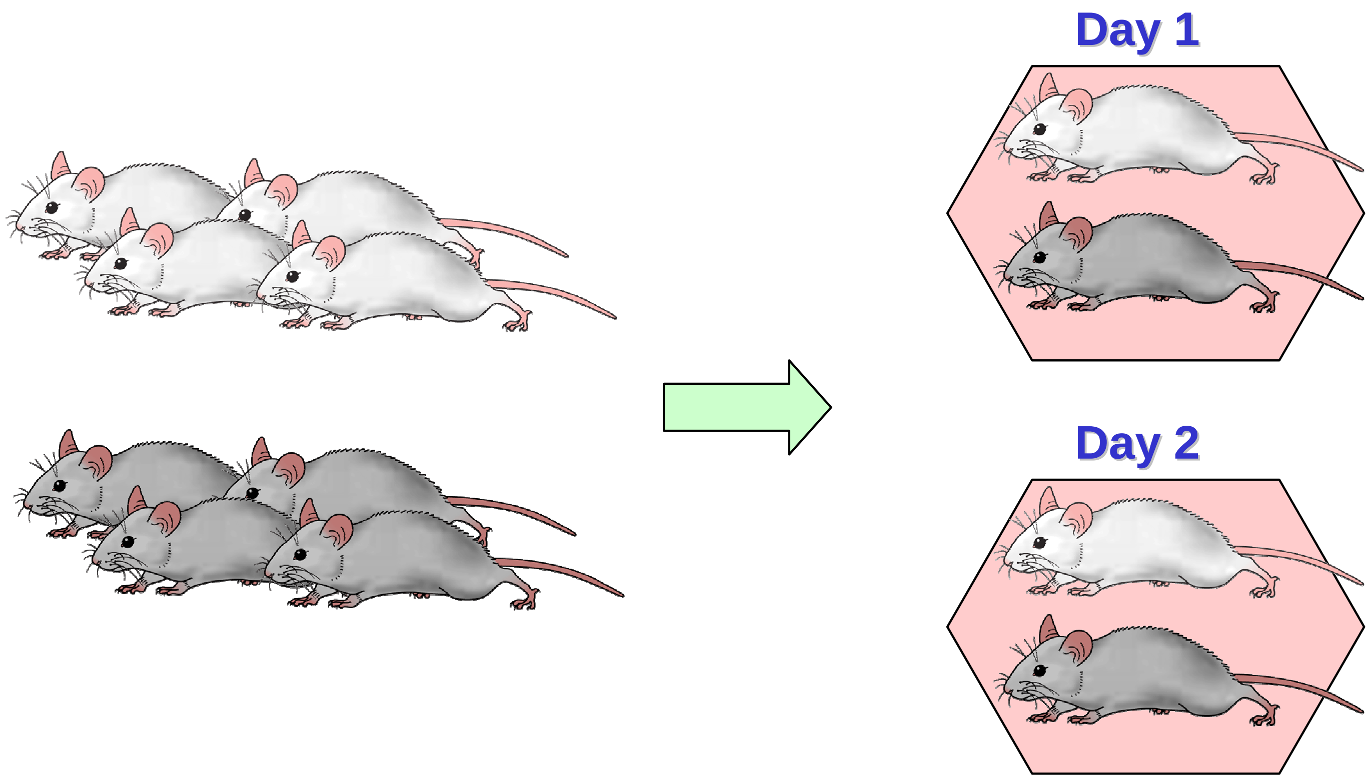
We can nicely randomize:

Day effect

Batch effect

Blocking

The process of using the same or similar experimental units for all treatments. The purpose of blocking is to remove a source of variation from the error term and hence provide a more powerful test for a difference in population or treatment means.



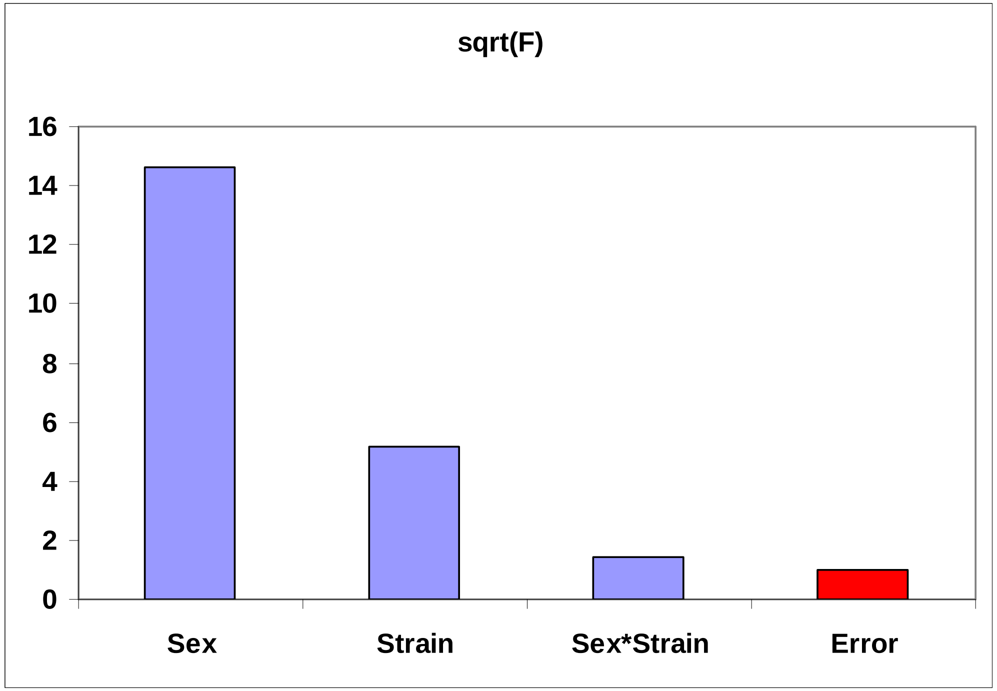
A good suggestion… ☺

Block what you can block, **randomize** what you cannot, and try to **avoid** unnecessary factors mice.xls

Q: Does mouse strain affect the weight? Show the effects of sex and strain using ANOVA

		129S1/SvImJ	A/J	AKR/J	BALB/cByJ	BTBR_T+_	BUB/BnJ	C3H/HeJ
1	Female	20.5	23.2	24.6	22.8	28	27.1	21.4
2		20.8	22.4	26	23.5	25.8	24.1	28.2
3		19.8	22.7	31	23.8	26	25.9	23.5
4		21	21.4	25.7	22.7	26.5	25.9	23.9
5		21.9	22.6	23.7	19.7	26.3	26	22.8
6		22.1	20	21.1	26.2	27	27.1	18.4
7		21.3	21.8	23.7	24.1	26	26.2	21.8
8		20.1	20.8	24.5	23.5	28.8	27.5	25
9		18.9	19.5	32.3	23.8	28	30.2	20.1
10	Male	24.7	25.8	42.8	29.3	34.1	36.2	31.2
11		27.2	27.7	32.6	32.2	33	36.9	28.2
12		23.9	29.9	34.8	29.7	38.7	34.4	26.7
13		26.3	24.8	32.8	30	39	34.3	29.3
14		26	22.9	34.8	27	31	31.7	33.1
15		23.3	24.5	32.8	30	32	33	28.2
16		26.5	24.6	33.6	33.1	33.7	33.2	31.2
17		27.4	21.6	30.7	30.6	33.1	34	27.7
18		27.5	26.9	36.5	28.7	32.5	31	27.5

mice.xls



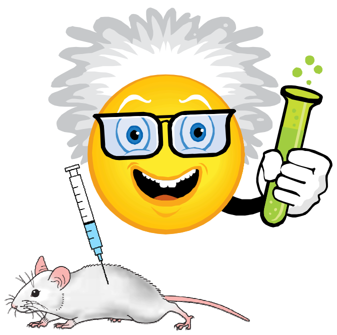
Factor	sqrt(F)
Sex	14.64136
Strain	5.193487
Sex*Strain	1.447993
Error	1

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	1206.676	1	1206.676	214.3693	3.36E-26	3.940163
Columns	759.13	5	151.826	26.97231	6.06E-17	2.309202
Interaction	59.01074	5	11.80215	2.096684	0.072376	2.309202
Within	540.38	96	5.628958			
Total	2565.197	107				

Thank you for your attention

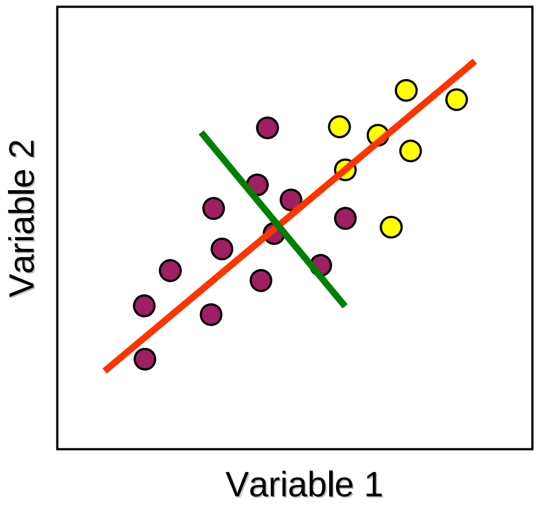
to be continued…



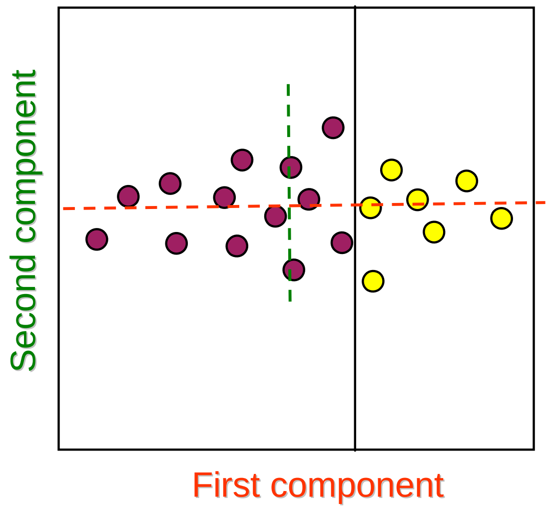
- Principal component analysis (PCA) is a vector space transform often used to reduce multidimensional data sets to lower dimensions for analysis. It selects the coordinates along which the variation of the data is bigger.

- Example for 2D case: for the simplicity let us consider 2 parametric situation both in terms of data and resulting PCA.

Scatter plot in "natural" coordinates



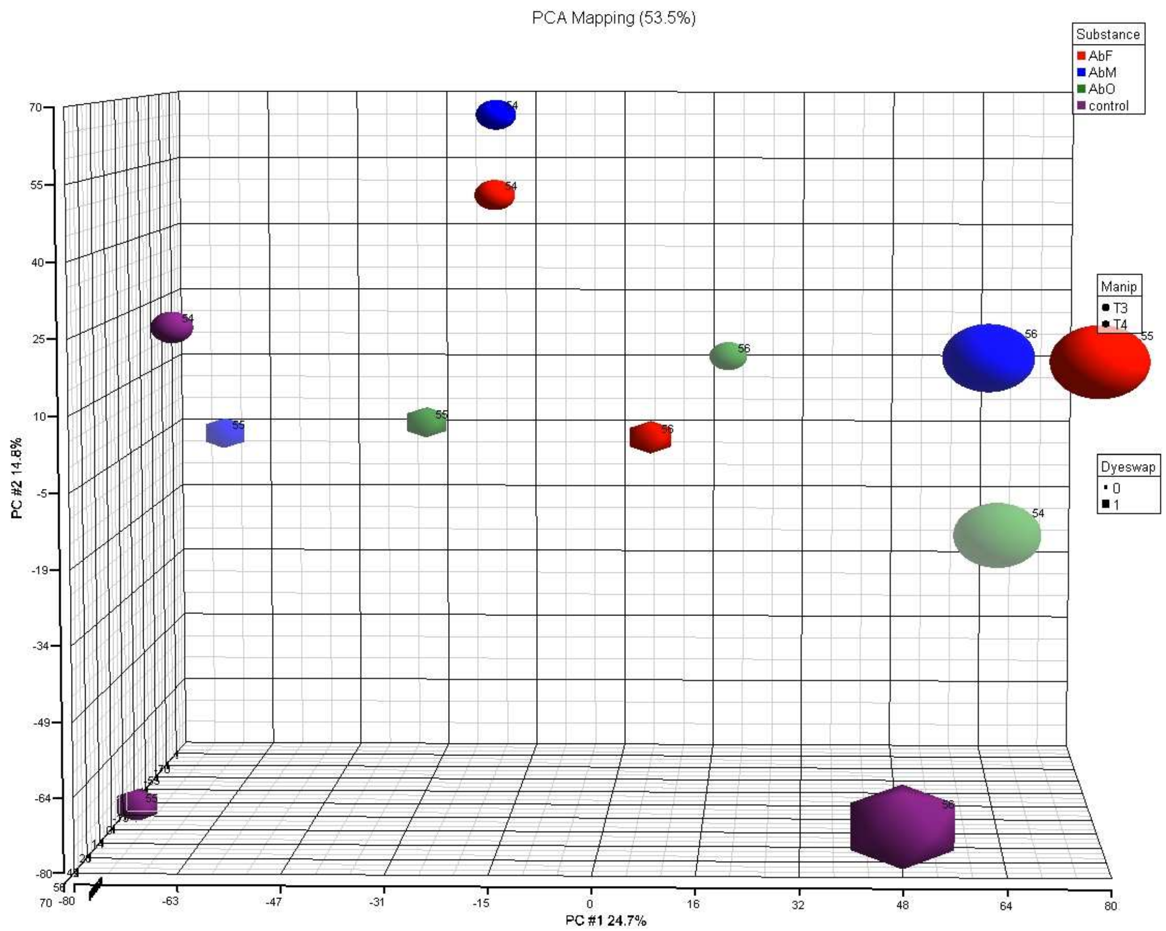
Scatter plot in PC



- Instead of using 2 "natural" parameters for the classification, we can use the first component!

PCA Example in Partek

- Transcriptomic profile of a sample contains thousands of genes, i.e. thousands of coordinates/parameters.
- PCA is extremely useful for initial data analysis in transcriptomics, as it allows to depict thousands of parameters just in 2 or 3 dimension space.



3 factors can influence the distribution of the variability:

- Substance
- Manip (bio replicate)
- Dye swap