STATISTICS IN USE:
Basic Statistical Methods and Tools

Methods, Tools and Applications

Petr V. Nazarov

Microarray Center,
CRP-Santé,
Luxembourg

petr.nazarov@crp-sante.lu

OUTLINE

I. Introduction

- Statistical questions
- Basic notation and numerical measures
- Random variables and their distributions

II. Methods and Examples

- Detection the outliers
- Interval estimation for the mean
- Comparison of the means of two samples
- Interval estimation for the variance
- Test of goodness of fit (model testing)
- Analysis of variance, ANOVA
- Principal component analysis, PCA
- Effect normalization

INTRODUCTION

Statistical Questions

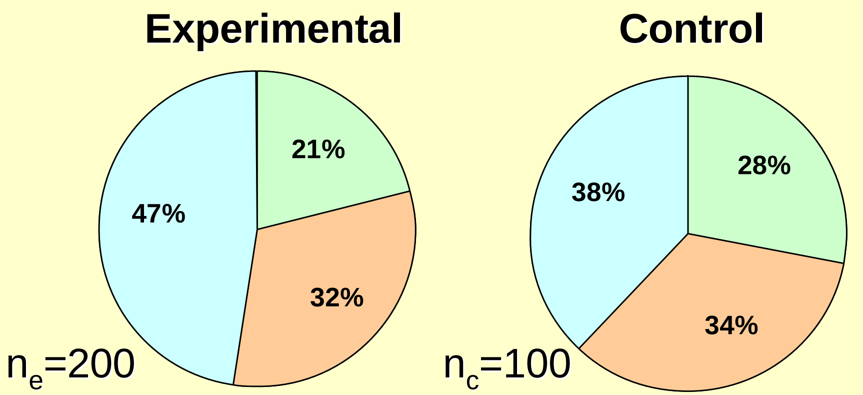
◆ The number of living cells measured in 5 independent experiments are 1520, 1231, 2102, 1867, 1625

What is the ***interval estimation*** for the real average number of living cells?

◆ The number of living cells measured in 3 independent experiments for 2 conditions are
A: 1520, 1231, 1425,
B: 2102, 1867, 1625

Are the average numbers of living cells ***significantly different*** for A and B?

◆ The proportions for 3 "classes" of patients with and without treatment are:



Are the proportions ***significantly different*** in control and experimental groups?

◆ The behaviour of a cell line is studied, being affected by several factors (e.g. concentration, time of treatment, temperature).

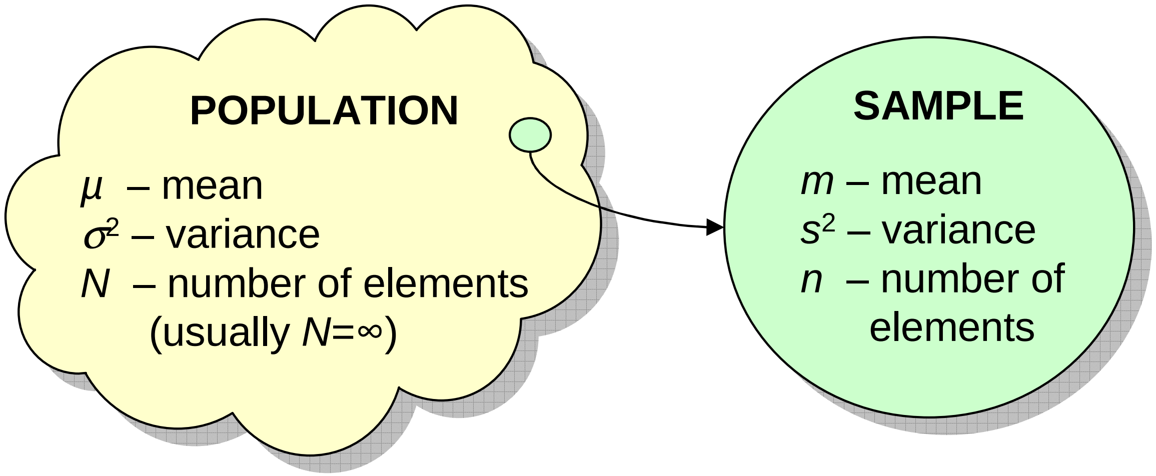
	Concentration				
Time	**0.1**	**0.2**	**0.5**	**1**	**2**
1	21.11	23.74	22.19	24.45	24.32
2	24.02	25.19	25.44	26.59	27.43
5	25.43	25.58	25.30	24.74	28.59
10	22.48	22.84	24.01	26.04	26.60
30	25.77	26.52	25.43	25.39	30.75
60	28.76	31.08	28.97	28.74	34.96

Which of the factors effect the behavior more and are more important?

INTRODUCTION

Basic notation and numerical measures

Let the measured quantity be *x*. This *x* can be also referred as a *random variable*.



$$m = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - m)^2$$

$$r = \frac{s_{xy}}{s_x s_y}, \quad s_{xy}^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - m_x)(y_i - m_y)$$

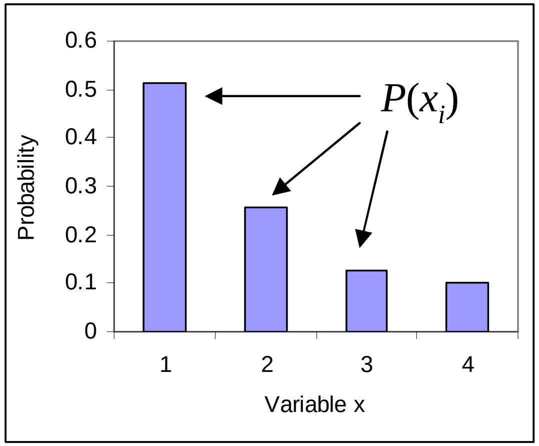
Numerical measures:

- Mean μ, m – characteristic of the position (unstable to outliers)
- Trimmed mean – characteristic of the position (stable to outliers)
- Median *med* – robust characteristic of the position (but less precise)
- Variance σ^2, s^2 – the characteristic of the scale (squared)
- Standard deviation σ, s – the characteristic of the scale (linear)
- Inter-quartile range IRQ – robust characteristic of the scale (but less precise)
- Correlation r – characteristic of linear dependency of 2 data sets

INTRODUCTION

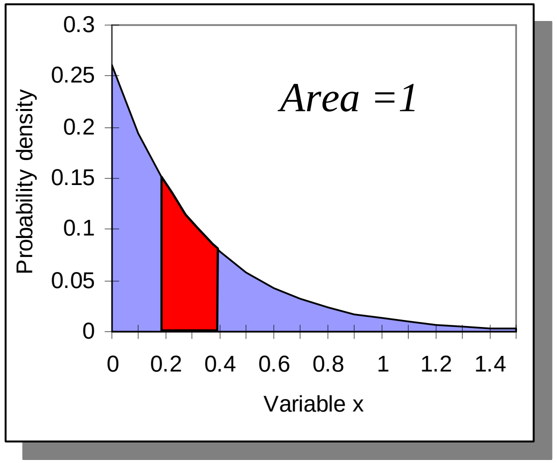
- Random variables can be *discrete* or *continues*.

Probability function:



$$\sum_{i=1}^{k} P(x_i) = 1$$

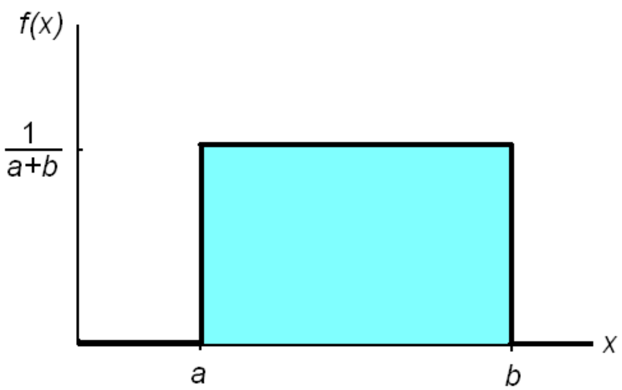
Probability density function:



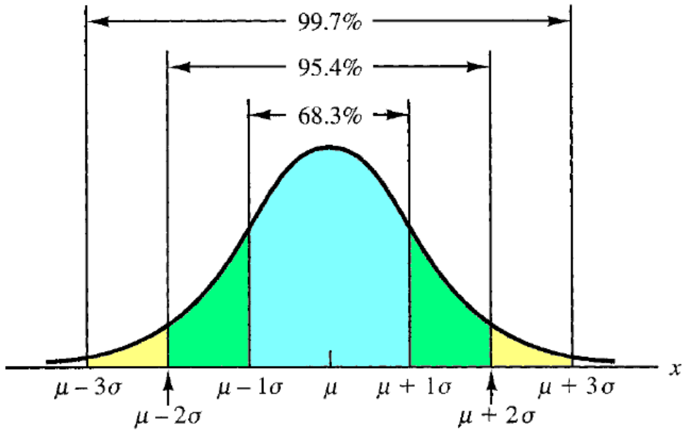
$$\int_{\min(x)}^{\max(x)} f(x)dx = 1$$

Examples of distributions for continues

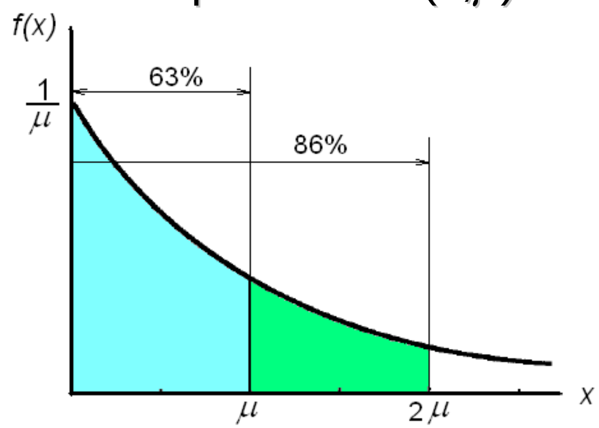
- Uniform: *f*(*x*,*a*,*b*)



- Normal (Gaussian): *f*(*x*,*μ*,*σ*)



- Exponential: *f*(*x*,*μ*)



METHODS AND APPLICATIONS

Detection of outliers

Chebyshev's theorem

- For any kind of distribution at least $1-z^{-2}$ of the data values must be within z standard deviations from the mean ($\mu \pm z\sigma$), where *z* is any number > 1.

- At least 75% of data have z-score < 2
- At least 89% of data have z-score < 3
- At least 94% of data have z-score < 4
- At least 96% of data have z-score < 5

$$z_i = \frac{x_i - \mu}{\sigma}$$

"Rule of thumb":

If $|z_i|$ **> 3** (for symmetrical distr.) or $|z_i|$ **> 5** (for skewed distr.) then x_i **is an outlier**.

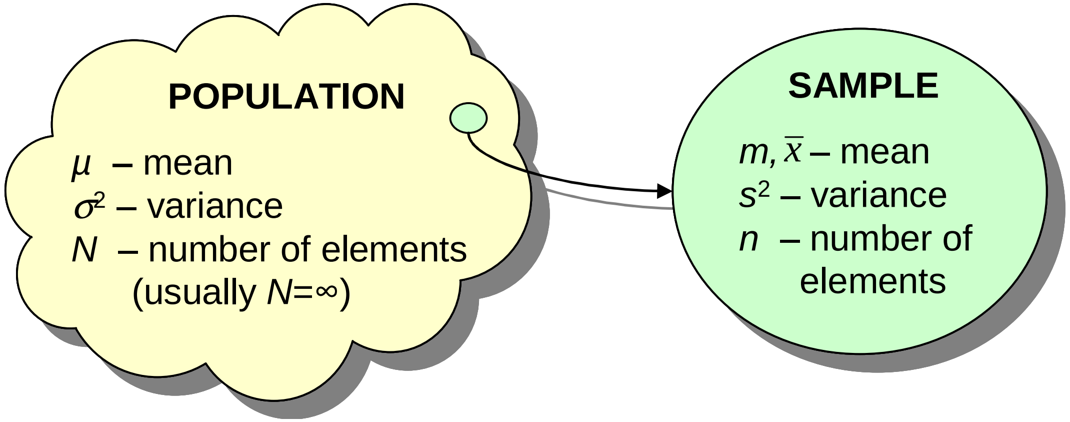
Example

Number of cells				
503	516	529	529	507
589	547	515	490	484
491	154	215	536	508
532	546	572	517	499
455	558	552	462	554
469	500	588	516	485
506	507	523	567	533
512	529	534	523	581
543	577	573	526	471
478	495	517	473	548

z-score				
-0.08	0.10	0.27	0.27	-0.02
1.07	0.51	0.08	-0.25	-0.33
-0.24	**-4.73**	**-3.92**	0.36	-0.01
0.31	0.49	0.84	0.11	-0.12
-0.72	0.66	0.58	-0.62	0.61
-0.53	-0.11	1.05	0.09	-0.32
-0.04	-0.02	0.19	0.78	0.32
0.04	0.27	0.34	0.19	0.97
0.46	0.91	0.86	0.24	-0.51
-0.41	-0.18	0.12	-0.48	0.53

INTERVAL ESTIMATIONS

Interval estimations for mean and proportion



If x is a random variable, then m and s are random variables too.

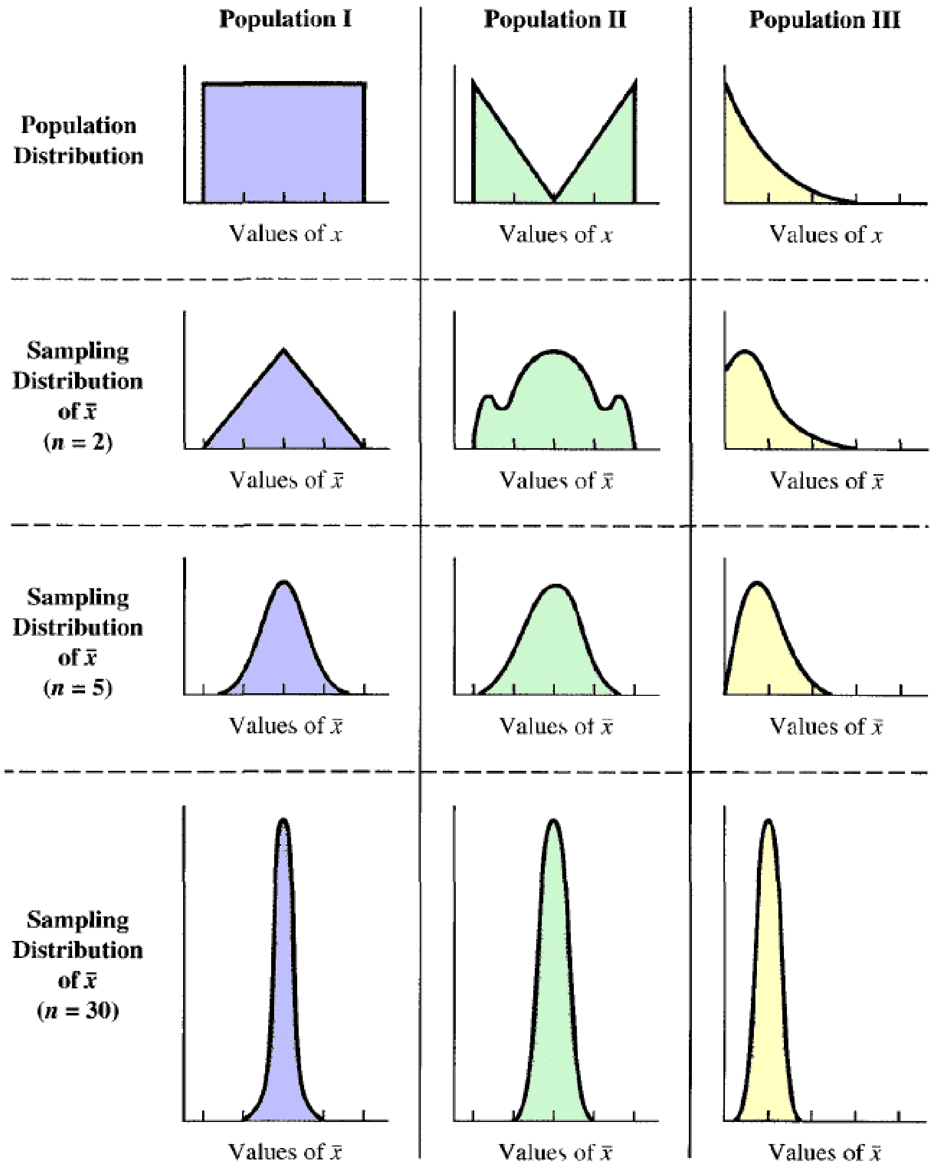
Central Limit Theorem:

The distribution of the sample mean tends to the normal distribution, when the sample size *n* increases.

In practice if the sample size is **>30**, the normal distribution is a good approximation for the sample mean for any initial distribution.

NOTE: here and below $\overline{X}$ will be used together with m as a sample mean.

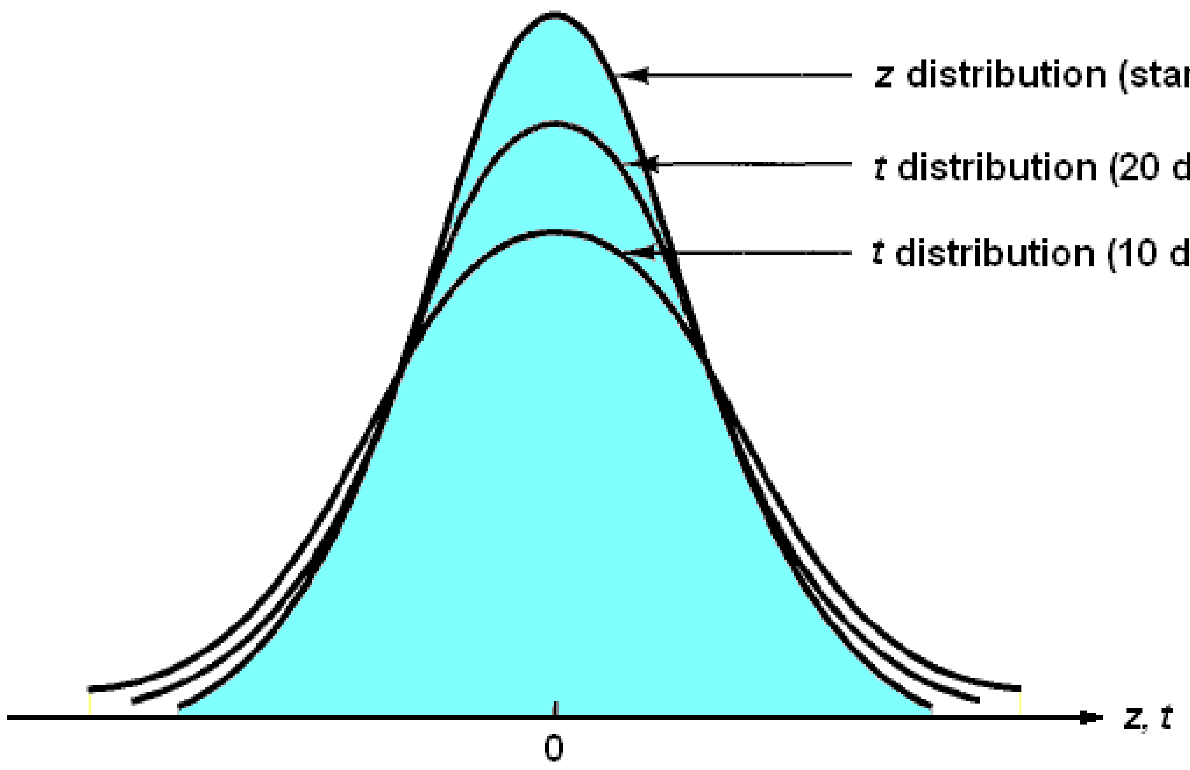
- Illustration of the central limit theorem.



INTERVAL ESTIMATIONS

Statistics used for means and proportions

- In the case of known population variance σ^2 (rare!): z-statistics (Gaussian)
- In the case of unknown population variance: t-statistics (Student's)
- Population proportion: z-statistics



INTERVAL ESTIMATIONS

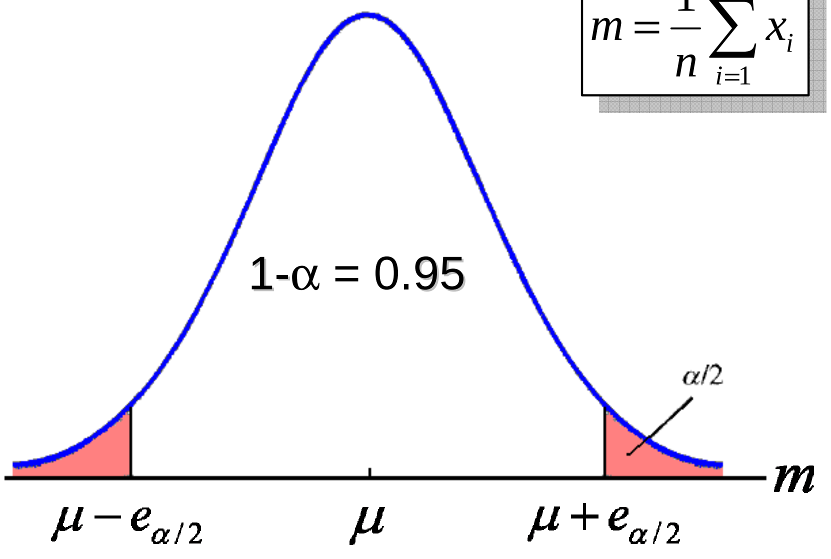
Population mean

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i$$

◆ Interval estimation for the population mean

Let us define α as "error probability", then 1-α is called *confidence interval*. For example let α=0.05

$$m = \mu \pm e_{a/2} \quad \Leftrightarrow \quad \mu = m \pm e_{a/2}$$



In the case of unknown σ^2 the interval is defined as:

$$\mu = m \pm t_{\alpha/2}^{df=n-1} \frac{s}{\sqrt{n}}$$

An example in Excel

x	mean(x)	e
1.421233	1.463722	0.371382
1.748418		
1.081124		
1.112433		
1.985844		
1.433279		

m = AVERAGE (A2:A7)

e = **TINV**(0.05,6-1)*STDEV(A2:A7)/SQRT(n)

α

degree of freedom = *n* -1

NOTE: there is α value in TINV instead α/2.

INTERVAL ESTIMATIONS

Population proportion

Interval estimation for the population proportion (π)

Again α is "error probability", 1-α is *confidence interval*. Let α=0.05

$$P = \frac{n_{good}}{n}$$

$$\Pi = P \pm e_{\alpha/2}$$

Usually *z*-statistics is used. But the requirement must be obeyed →

$$\begin{cases} nP > 5 \\ n(1-P) > 5 \end{cases}$$

$$\Pi = P \pm z_{\alpha/2}\sqrt{\frac{P(1-P)}{n}}$$

Example in Excel

#	DATA		
1	1	1	0
2	0	0	0
3	1	1	0
4	0	1	0
5	0	1	1
6	1	0	1
7	1	0	1
8	1	1	0
9	1	0	1
10	0	0	1
11	1	1	1
12	1	0	0
13	0	0	0
14	0	1	0
15	1	1	1
16	1	1	1

n	48
p(1)	0.5625
e	-0.14

data

P = COUNTIF(F6:H21,"=1")/n

e =**NORMINV**(0.025,0,1)*SQRT(P*(1-P)/n)

$\alpha/2$

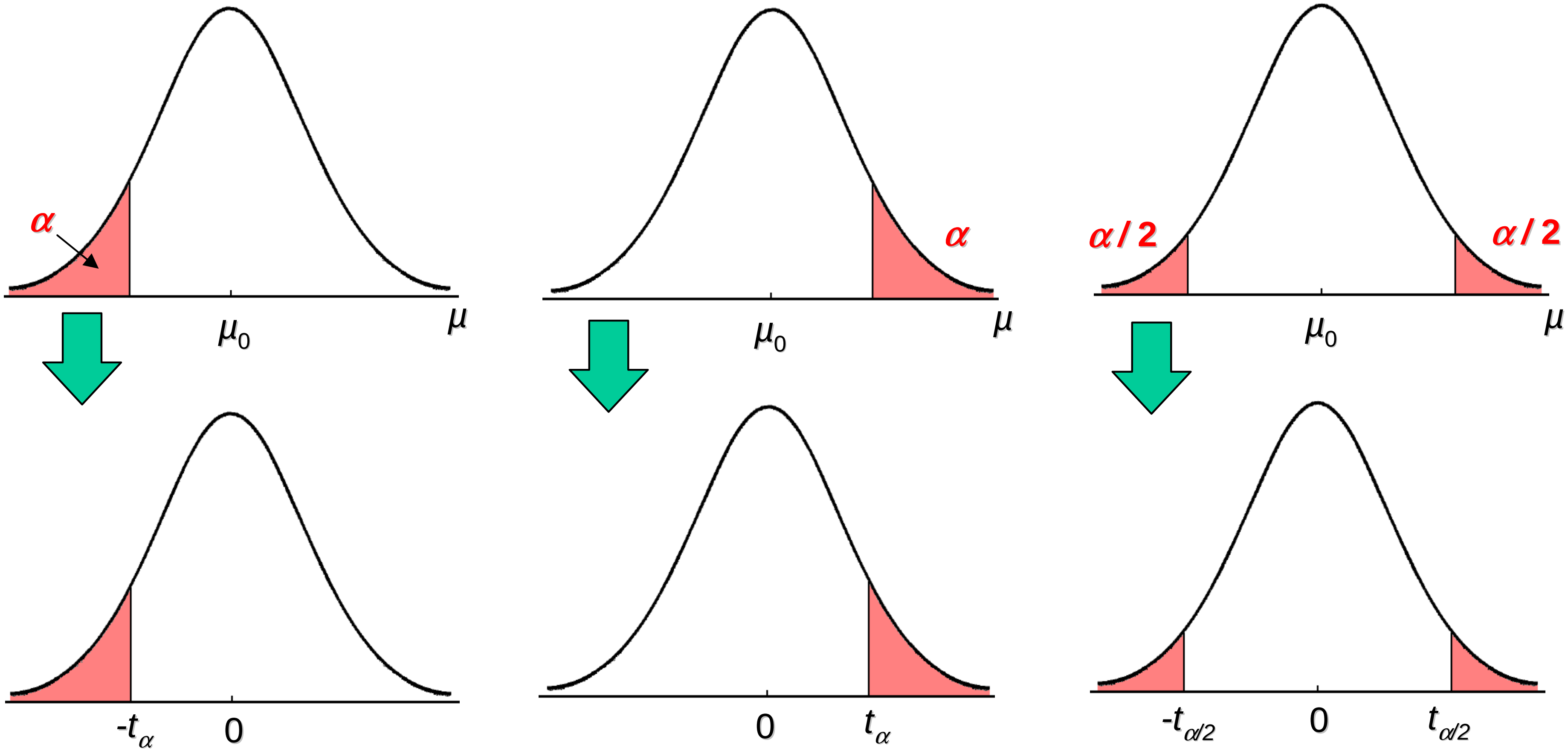
NOTE: there is α/2 value in NORMINV !!!

HYPOTHESIS TESTING

Hypothesis about population mean

- **Standard hypotheses look like:**

Lower Tail	***Upper Tail***	***Two Tail***
$H_0: \mu \geq \mu_0$	$H_0: \mu \leq \mu_0$	$H_0: \mu = \mu_0$
$H_a: \mu < \mu_0$	$H_a: \mu > \mu_0$	$H_a: \mu \neq \mu_0$



HYPOTHESIS TESTING

Hypothesis about population mean

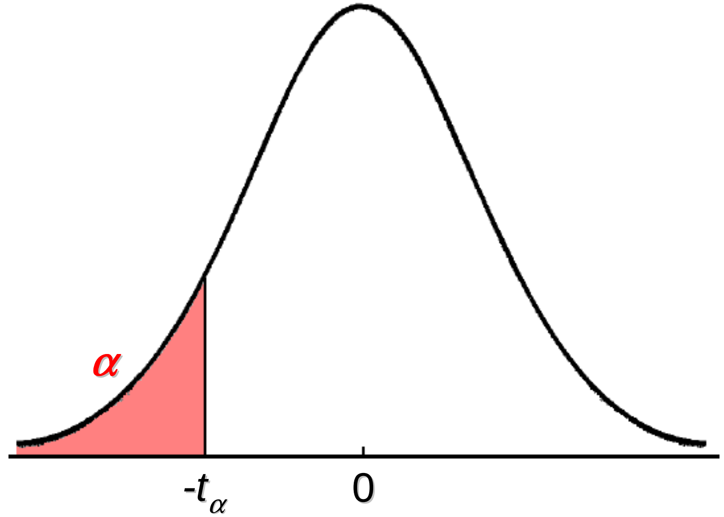
Lower Tail

H_0: $\mu \geq \mu_0$

H_a: $\mu < \mu_0$

- (1) Build a proper statistics

$$t, z = \frac{m - \mu_0}{s}$$



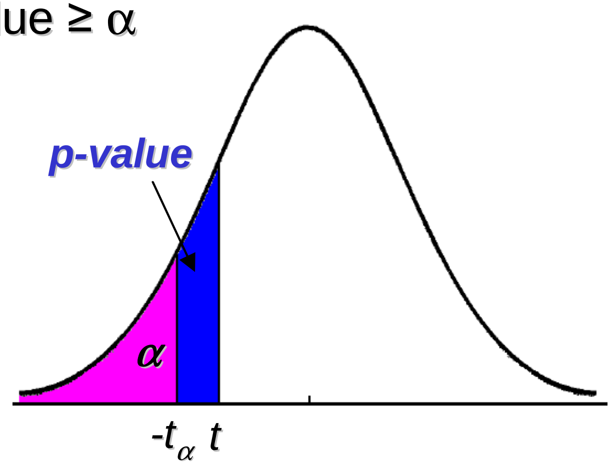
- (2) Check the position of t with respect to t_α

OR

- (2) Calculate *p-value* (area) using inverse distribution

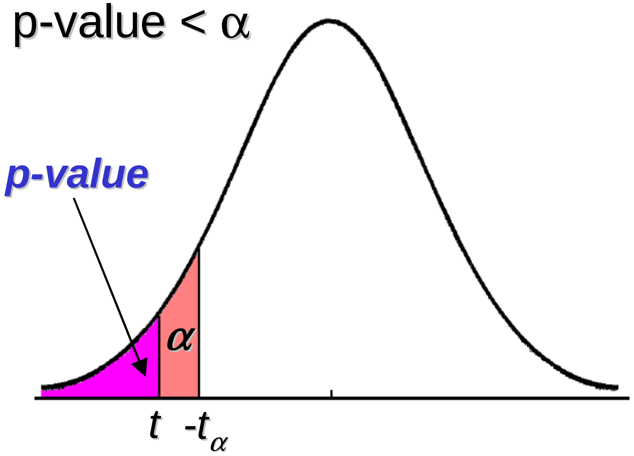
(3) **2 possible situations:**

- p-value ≥ α



The null-hypothesis H_0 cannot be rejected

- p-value < α



The null-hypothesis H_0 can be rejected with 1-α confidence

HYPOTHESIS TESTING

Excel example: hypothesis about population mean

- Number of living cells in 5 wells under some conditions are given in the table, with average value of 4705. In a reference literature source authors clamed a mean quantity of 5000 living cells under the same conditions.

# well	Living cells
1	5128
2	4806
3	5037
4	4231
5	4322

m= 4704.8
s= 409.49

- Question: is our experiment significantly different from the one performed in a reference article?

- **Solution**

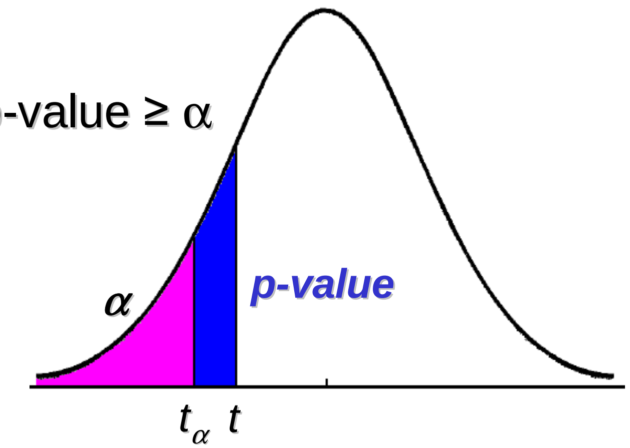
Lower Tail

H_0: $\mu \geq 5000$

H_a: $\mu < 5000$

$$t = \frac{m - \mu_0}{s} = \frac{4704.8 - 5000}{409.5} = -1.61$$

p-value ≥ α



x		
5128	m=	4704.8
4806	s=	409.4871
5037	mu0=	5000
4231	t=	-1.61199
4322	p-value=	0.091129

m = AVERAGE(A2:A6)

s = STDEV(A2:A6)

μ_0 = 5000

t = (m- μ_0)/s*SQRT(5)

p-value = **TDIST**(ABS(t);5-1;1)

The null-hypothesis H_0 cannot be rejected: no significant difference between reference and actual experiments

HYPOTHESIS TESTING

Testing hypothesis about means of two population

- **One way to compare means:**

Two Tail

$H_0: \mu_1 = \mu_2$

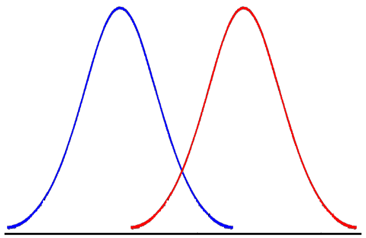
$H_a: \mu_1 \neq \mu_2$

Two Tail

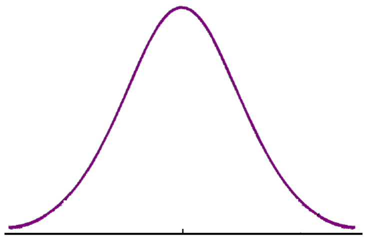
$H_0: \mu_1 - \mu_2 = 0$

$H_a: \mu_1 - \mu_2 \neq 0$

Distributions of m_1 and m_2



Distribution of m_1-m_2



$$\sigma^2_{m_1 - m_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

$$t = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

- **And another… :**

Excel → Tools → Data Analysis

Select for example t-Test for unequal variances

A	B
1520	2102
1231	1867
1425	1625

t-Test: Two-Sample Assuming Unequal Variances

	Variable 1	*Variable 2*	
Mean	1392	1864.667	
Variance	21697	56886.33	
Observations	3	3	
Hypothesized Mea	0		
df	3		
t Stat	-2.920454		
P(T<=t) one-tail	0.030737		< 0.05
t Critical one-tail	2.353363		
P(T<=t) two-tail	0.061474		> 0.05
t Critical two-tail	3.182446		

NOTE: other (one tail) hypothesis can be applied as well, depending on the question.

HYPOTHESIS TESTING

Non-parametric method: U-test

Wilcoxon rank-sum test, also known as 'Mann-Whitney U' checks whether data for two sets come from the same distribution.

- Non-parametric methods do not put restrictions on the distribution of the data.
- Specifically the U-test can be used for ordinal data (e.g. "G", "S", "B" medals in sport)
- Robust to outliers
- Attention: U-test compares distributions, not specifically medians (as addressed usually)

Example in R

R programming language originally was developed to solve statistical tasks, it has much wider possibilities and consistency in comparison to Excel Data Analysis.

Let us apply U-test to the same data as t-test:

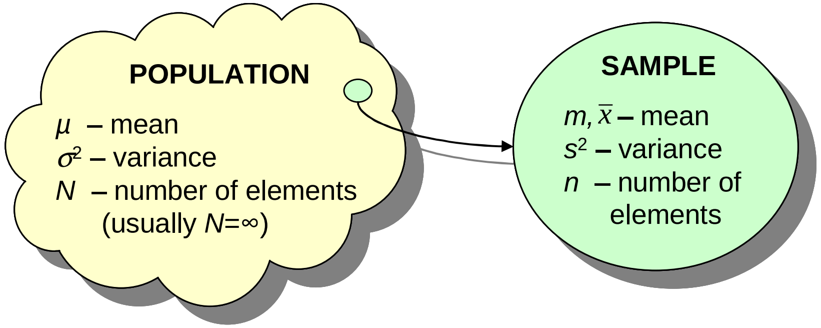
A	B
1520	2102
1231	1867
1425	1625

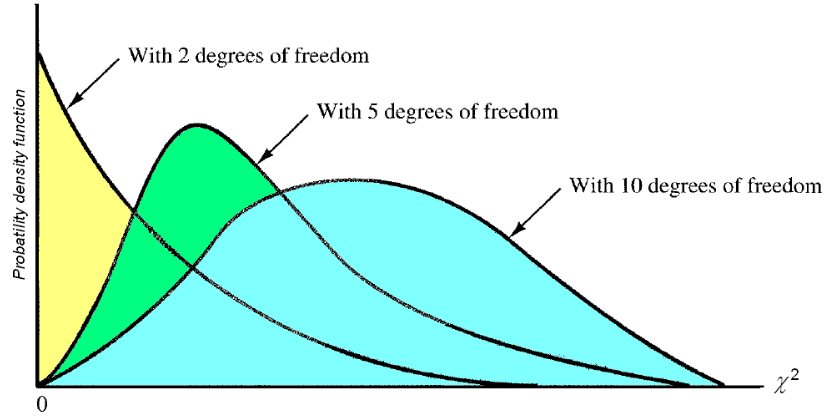
```
> x1=c(1520,1231,1425)
> x2=c(2102,1867,1625)
> wilcox.test(x1,x2)
        Wilcoxon rank sum test
data:  x1 and x2
W = 0, p-value = 0.1
alternative hypothesis: true location shift is not equal to 0

> wilcox.test(x1,x2, alternative="less")
        Wilcoxon rank sum test
data:  x1 and x2
W = 0, p-value = 0.05
alternative hypothesis: true location shift is less than 0
```

INFERENCE ABOUT VARIANCES

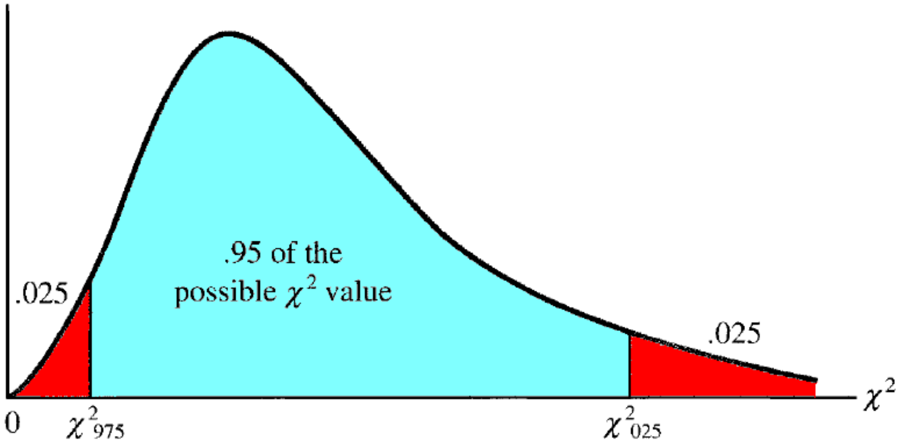
Interval estimation for the sample variance, χ^2 statistics





If x is a random variable, then s^2 is a random variable too. The interval estimation for it is build using chi-square statistics (χ^2).

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{(1-\alpha/2)}}$$



Example in Excel

x
4.38
2.18
2.21
3.29
2.50
2.85
2.67
2.30
4.06
3.26
1.83
2.73
2.59
1.56
2.76
3.99
3.14
2.79
3.43
2.56

m=	2.854241
s=	0.728399
s2=	0.530565
min_s2=	0.30685
max_s2=	1.131839

m = AVERAGE(A2:A21)

s = STDEV(A2:A21)

s^2 = VAR(A2:A21)

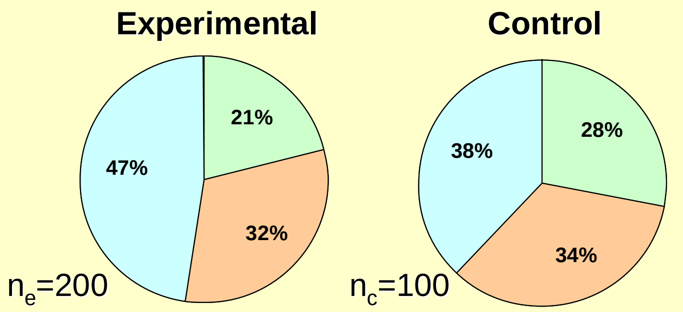
min_s^2 = 19*VAR(A2:A21)/**CHIINV**(0.025;19)

max_s^2 = 19*VAR(A2:A21)/**CHIINV**(0.975;19)

TEST OF GOODNESS OF FIT

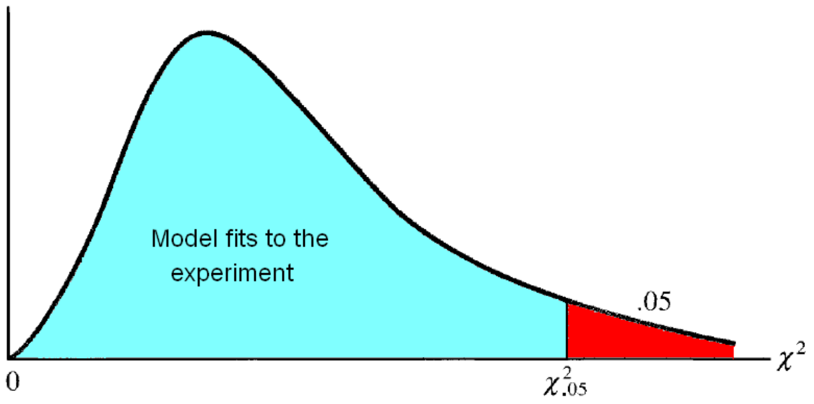
Application of χ^2 statistics for model testing

◆ The proportions for 3 "classes" of patients with and without treatment are:



Are the proportions ***significantly different*** in control and experimental groups?

◆ Goodness of fit hypothesis is always one tail!



◆ Build the model of the distribution and calculate *expected frequencies* using control group of patients. Each expected frequency must be ≥ 5.

Category	Control frequenc.	Distrib. model	Expected freq., e	Experim. freq.,f
A	28	0.28	56	42
B	34	0.34	68	64
C	38	0.38	76	94
Sum	**100**	**1**	**200**	**200**

◆ Calculate test χ^2 statistics using equation:

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

χ^2 degree of freedom = k-1

Category	(f-e)2/e
A	3.500
B	0.235
C	4.263
Chi2	**7.998**
p-value	**0.01833**

Chi2 = SUM(…)

p-value = **CHIDIST**(Chi2;2)

◆ Exactly the same approach can be applied for testing the independence. Difference: expected frequencies are calculated on all the data, instead of "control set".

HYPOTHESIS TESTING

Hypothesis testing for variances, F-statistics

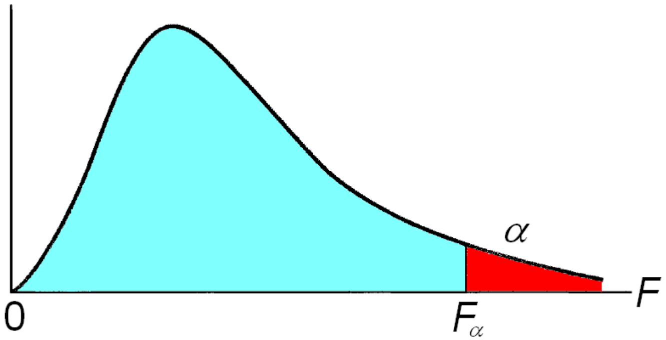
x1	x2
3.50	2.18
4.11	3.24
1.78	3.01
4.07	1.95
3.18	2.72
4.05	3.08
2.07	2.59
4.69	1.93
1.99	3.15
2.45	3.09

Two Tail

H_0: $\sigma_1 = \sigma_2$

H_a: $\sigma_1 \neq \sigma_2$

$$\frac{s_1^2}{s_2^2} = F$$



- The ratio of the sample variances is called *F-statistics*.
- As opposed to *t* and χ^2 it is has 2 degrees of freedom, called numerator and denominator degrees of freedom.

F numerator d.f.= n_1-1
F denominator d.f.= n_2-1

- Note: For the consistency the maximal *s* is put to numerator. Then $F > 1$.

Example in Excel

x1	x2
3.50	2.18
4.11	3.24
1.78	3.01
4.07	1.95
3.18	2.72
4.05	3.08
2.07	2.59
4.69	1.93
1.99	3.15
2.45	3.09

s2_1= 1.104897
s2_2= 0.257265

F= 4.294772
p-value= 0.020453

FTEST
p-value= 0.040907

s_1^2 = VAR(A2:A10)

s_2^2 = VAR(B2:B10)

F=MAX(s_1^2, s_2^2)/MIN(s_1^2, s_2^2)

p-value1 = **FDIST**(*F*;9;9)

p-value2 = **FTEST**(A2:A11;B2:B11)

ANOVA

ANOVA: first glance

- The behaviour of a cell line is studied, being affected by several factors (e.g. concentration, time of treatment, temperature).

Time	Concentration 0.1	0.2	0.5	1	2
1	21.11	23.74	22.19	24.45	24.32
2	24.02	25.19	25.44	26.59	27.43
5	25.43	25.58	25.30	24.74	28.59
10	22.48	22.84	24.01	26.04	26.60
30	25.77	26.52	25.43	25.39	30.75
60	28.76	31.08	28.97	28.74	34.96

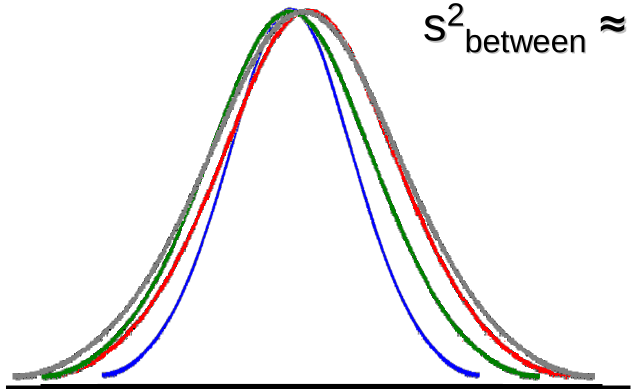
Which of the factors effect the behavior more and are more important?

The answer to this question can be given by the Analysis of Variance (ANOVA).

There are several explanation how does ANOVA works. The one related to within/between treatment distributions is given below.

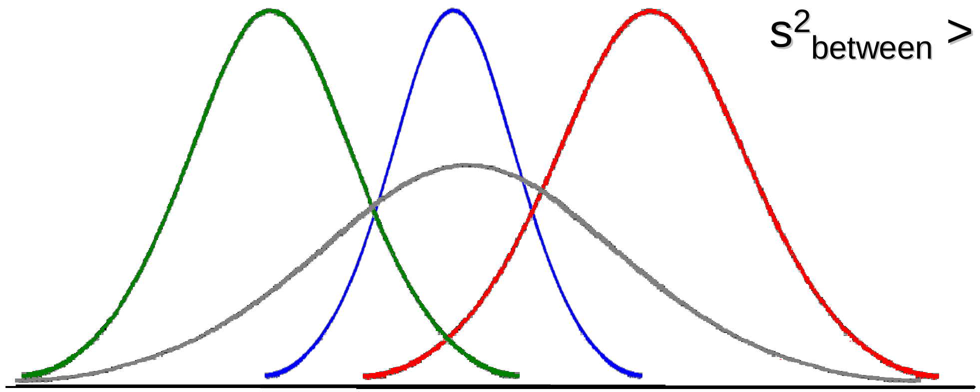
Assume that we have data recorded under 3 effects or ***treatments*** (red or green or blue)

- No significant effect.



$s^2_{between} \approx s^2_{within}$

- Presence of a significant effect.



$s^2_{between} > s^2_{within}$

ANOVA uses F statistics: $F = \frac{s^2_{between}}{s^2_{within}}$

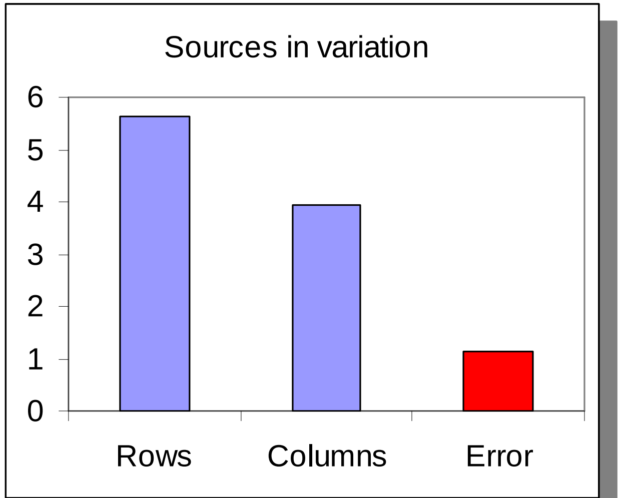
ANOVA

Application

◆ The behaviour of a cell line is studied, being affected by several factors (e.g. concentration, time of treatment, temperature).

Time	Concentration 0.1	0.2	0.5	1	2
1	21.11	23.74	22.19	24.45	24.32
2	24.02	25.19	25.44	26.59	27.43
5	25.43	25.58	25.30	24.74	28.59
10	22.48	22.84	24.01	26.04	26.60
30	25.77	26.52	25.43	25.39	30.75
60	28.76	31.08	28.97	28.74	34.96

Which of the factors effect the behavior more and are more important?



◆ If the number of factors is 1 or 2, Excel is an excellent tool for ANOVA.

◆ For more complex analysis (3 and more factors) other software tools should be used, including R and Partek®.

Anova: Two-Factor Without Replication

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
Row 1	5	115.81	23.162	2.12237
Row 2	5	128.67	25.734	1.73233
Row 3	5	129.64	25.928	2.31527
Row 4	5	121.97	24.394	3.45038
Row 5	5	133.86	26.772	5.15072
Row 6	5	152.51	30.502	7.17352
Column 1	6	147.57	24.595	7.27483
Column 2	6	154.95	25.825	8.36375
Column 3	6	151.34	25.22333	4.961267
Column 4	6	155.95	25.99167	2.443817
Column 5	6	172.65	28.775	13.71515

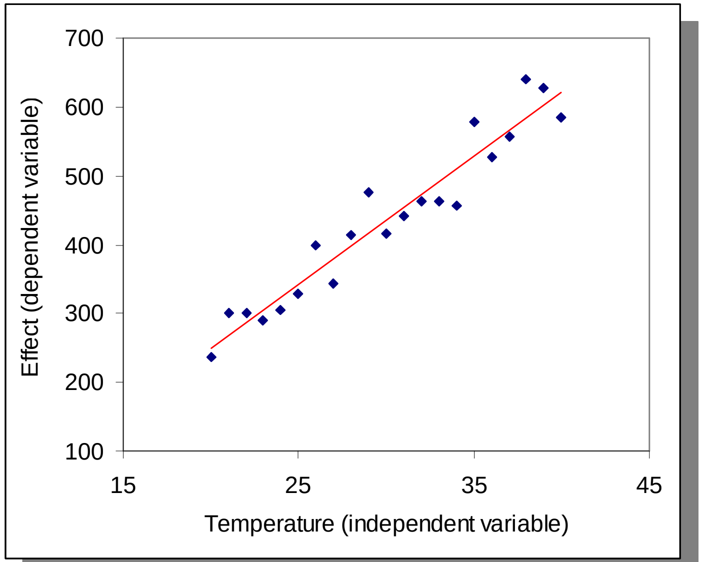
ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	157.6653	5	31.53306	24.13668	7.77E-08	2.71089
Columns	61.64961	4	15.4124	11.79728	4.38E-05	2.866081
Error	26.12875	20	1.306437			
Total	245.4437	29				

REGRESSION

Simple linear regression

Temper.	Effect
20	236
21	300
22	301
23	290
24	305
25	329
26	398
27	344
28	414
29	476
30	417
31	441
32	463
33	462
34	456
35	577
36	526
37	557
38	639
39	628
40	585



- Building a *regression* means finding and tuning the model to explain the behaviour of the data
- Model for a simple linear regression:

$$y(x) = \beta_1 x + \beta_0 + \varepsilon$$

$$y(x) = b_1 x + b_0$$

- b_1 and b_0 are random variables estimating β_1 and β_0. Interval estimations can be written for them.

Multiple linear regression

$$y(x_1,\dots,x_k) = \beta_1 x_1 + \dots + \beta_k x_k + \beta_0 + \varepsilon$$

- Linear regression (simple and multiple) is equivalent of ANOVA!

See the example: ↓

REGRESSION

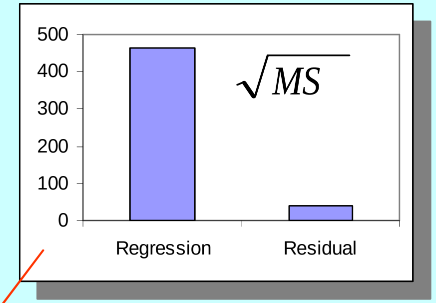
Simple linear regression in Excel

- Use Excel → Tools → Data Analysis → Regression.

Temper.	Effect
20	236
21	300
22	301
23	290
24	305
25	329
26	398
27	344
28	414
29	476
30	417
31	441
32	463
33	462
34	456
35	577
36	526
37	557
38	639
39	628
40	585

SUMMARY OUTPUT

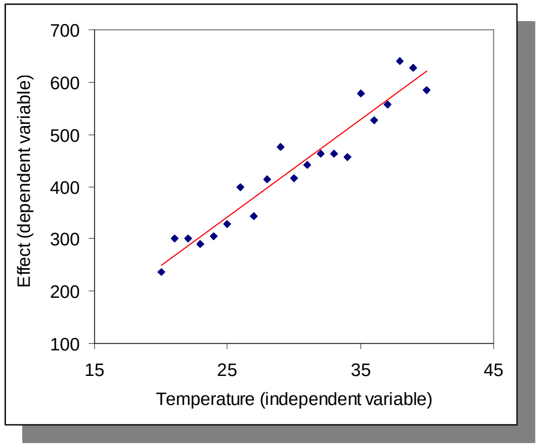
Regression Statistics	
Multiple R	0.933651166
R Square	0.871704499
Adjusted R Square	0.864952105
Standard Error	40.80172755
Observations	21

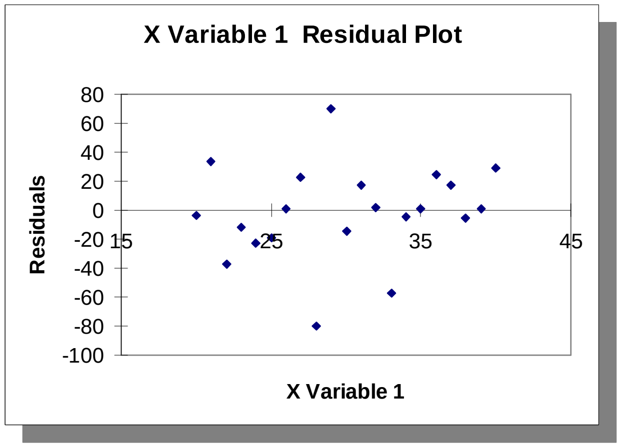


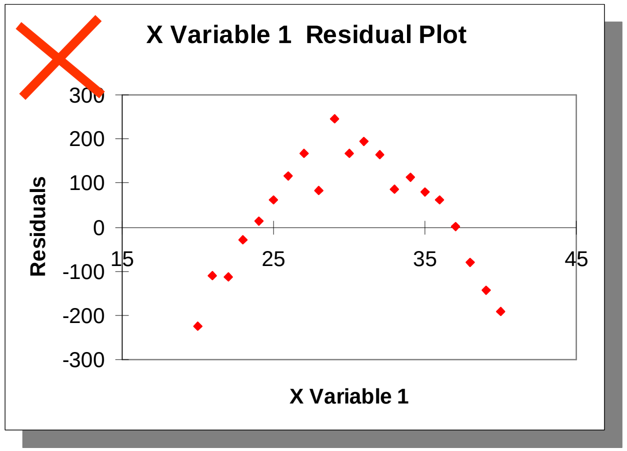
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	214915.9095	214916	129.09561	6.48154E-10
Residual	19	31630.83845	1664.78		
Total	20	246546.7479			

		Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
b_0	Intercept	-64.38300759	45.00136856	-1.4307	0.1687608	-158.5719543	29.80593909
b_1	X Variable 1	16.70663254	1.470392196	11.362	6.482E-10	13.62906631	19.78419877





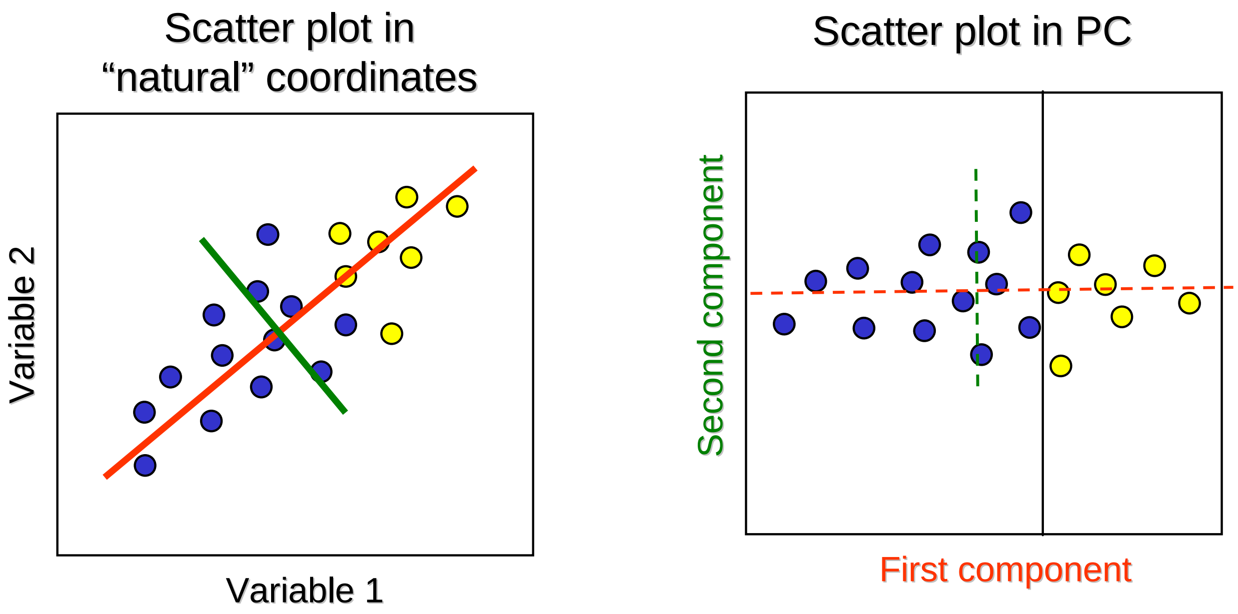


PRINCIPLE COMPONENT ANALYSIS

PCA basics

◆ 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.



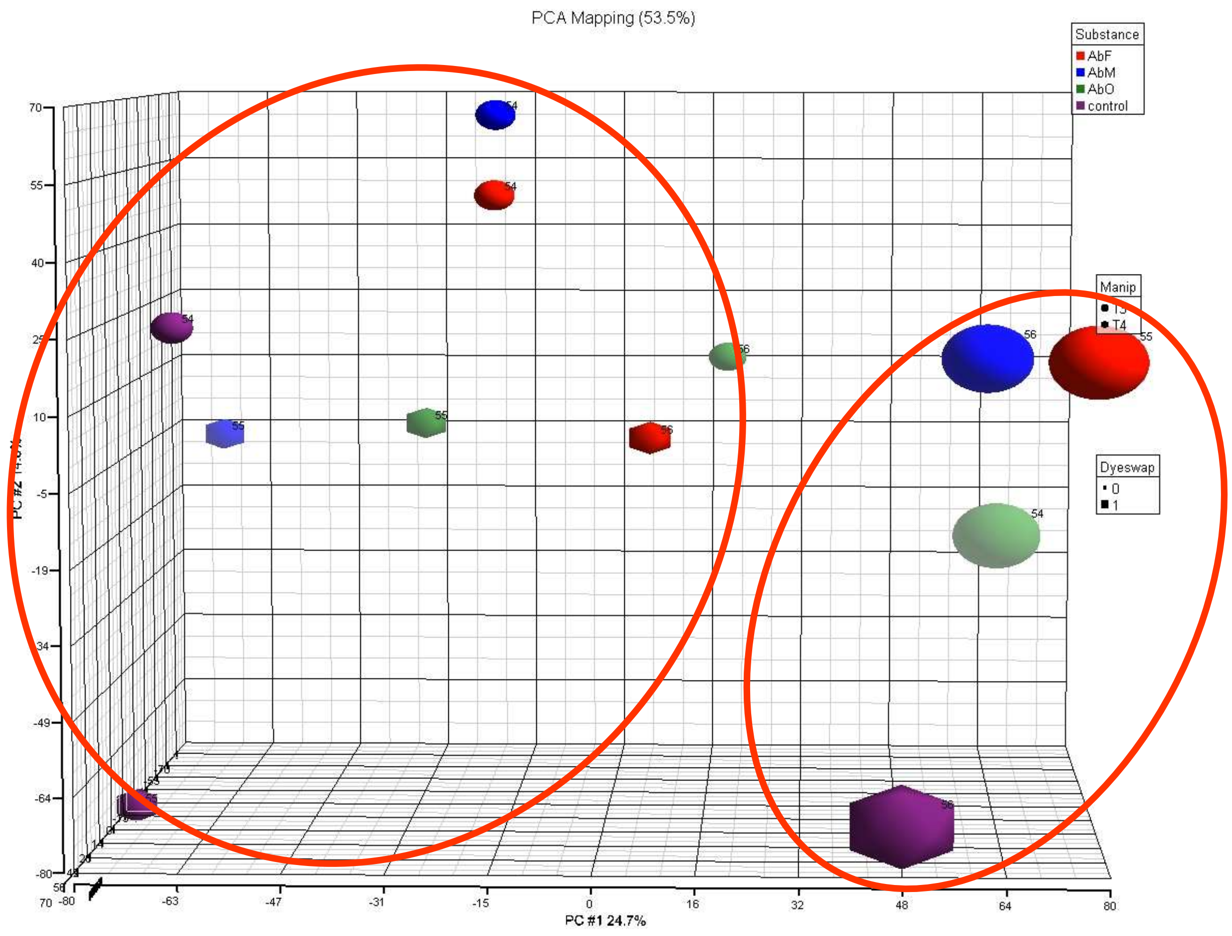
◆ Instead of using 2 "natural" parameters for the classification, we can use the first component!

PRINCIPLE COMPONENT ANALYSIS

PCA in Partek Genomic Suite

◆ 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

NORMALIZATION

An example of correction of the batch-effect

- Normalization can be considered as a correction for unwanted and artificial effects, e.g. batch effect, day effect, mood effect ☺ ☹.
- If effects are believed to be linear, the normalization can be performed using ANOVA or (equivalently) multiple regression.

$$y(x_1, x_2) = \beta_1 x_1 + \beta_2 x_2 + \beta_0 + \varepsilon \quad \Rightarrow \quad y^*(x_1) = y(x_1, x_2) - b_2 x_2 = \beta_1 x_1 + \beta_0 + \varepsilon^*$$