Test in two goups

Question: May the samples derive from the same population or the parameters of the two population are the same?

parametric

$\mu_1 = \mu_2$?

Nullhypothesis: $\mu_1 = \mu_2$

2-sample t-test

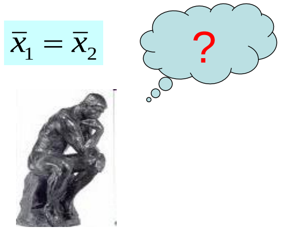
non-parametric

Nullhypothesis: same.

Mann-Whitney U-test

2-sample t-test

$\bar{x}_1 = \bar{x}_2$



Known distribution is necessary!

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s^* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$s^* = \sqrt{\frac{Q_1 + Q_2}{n_1 + n_2 - 2}}$$

Test

The t-vaue is same!

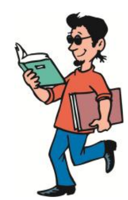
How much is the d.f.?

d.f. = n_1 + n_2 -2

$((n_1-1)+(n_2-1))$

conditions for the test

- task: comparison of two **independent** sample.
- The quantity has **normal distribution**.
- The sd are **same** in the two goups.

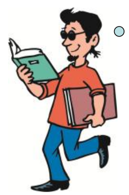


This is new!
How is it proved?

Test for standard deviations

How can I do?

Nullhypothesis: the two sd are the same the difference is random (sampling error).



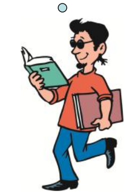
It is similar to the hypothesis testing!

F-test

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

A so-called F-distribution belongs to the nullhypothesis.

But which sd is in the nominator?



The larger variance is in the denominator! ($F \geq 1$)

decision

- 1. if the probability of the random deviation is small ($p(F>Fc_{rit}) \leq 5\%$) – ***reject*** the nullhypothesis.
- 2. if the probability of the random deviation is large ($p(F>F_{crit}) > 5\%$) – ***accept*** the nullhypothesis.

Mann-Whitney U-test

example: Is the painkiller effective?

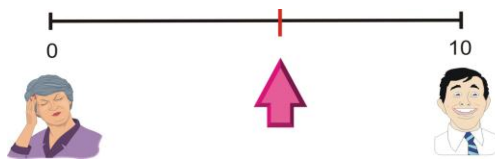
It has a serious pain.

How can we measure the effect?

Experiment

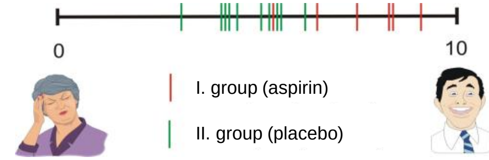
I. group:
(case)
aspirin

II. group:
(control)
placebo
(without agent)



This is an arbitrary scale.

Results



Value	3.1	4.1	4.2	4.3	4.5	5.1	5.3	5.4	5.5
Rank	1	2	3	4	5	6	7	8	9
value	5.6	6.2	6.2	6.5	7.5	8.3	8.3	8.4	9.1
rank	10	11.5	11.5	13	14	15.5	15.5	17	18

The nullhypothesis

The medicine is not effective.

The 2 groups belong to the same population.

The sum of the ranks (Gauss story)

Add the numbers from
1 to 100!

It is easy to calculate!

1 + 100 = 101
2 + 99 = 101 …

$$\sum_{i=1}^{n} i = \frac{n}{2}\cdot(n+1)$$

The transformation
(if n is enough large)

T – the sum of the ranks in the I. group. The expected value in the case of random distribution :

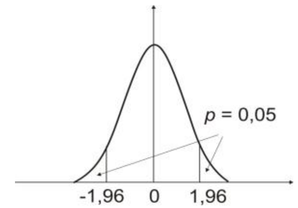
$$n_1 \cdot \frac{n_1 + n_2 + 1}{2}$$

z has standard normal distribution.

$$z = \frac{T - n_1(n_1 + n_2 + 1)/2}{s}$$

$$s = \sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}}$$

Decision



standard normal distribution

The calculated *z*-value: 3.24.

Larger then 1.96.

Conclusion: reject the nullhypothesis.

Calculated p-value < 0.1%.

Conclusion is same.

Analysis of variancia (ANOVA)

A

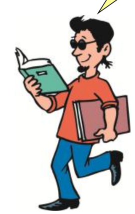
B

C

Is there difference among the groups?

No, There is only random deviations! This is the nullhypothesis.

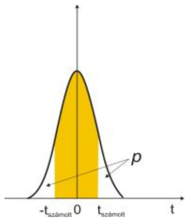
Why don't we do 2-sample t-tests?



The probabilty of the mistake increases rapidly with increasing number of groups!

Comparison of A – B and B – C and A – C. (not transitive)

How much is the chance of the mistake?



Suppose, that we reject the nullhypothesis in all cases.

p – probability being outside
(**1-p**) – probability being inside randomly.

Question: How much is the probability to have mistake at least one case?

How much is the chance that at least one is outside randomly?

1 test: ***p*** (let it be 5%). In the case of more than 1 the binomial distribution may be used to calculate.

$$1-(1-p)^3$$

In the case of 3 groups is about 15%!!!

More than 2 groups

1. group

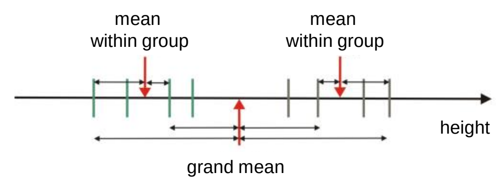
2. group

We can handle elements separated in groups or together.

mean within group: calculated from the elemetns of the given group.

grand mean: calculated from all elements.

Components of the variance



Remember:

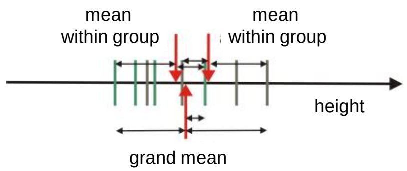
The variance is proportianal to the squared sum of the differences form the avgerage!

If the groups differ from each other significantly, the average squared differences from the grand mean are significanly higher than the differences in the groups!

The base of ANOVA

I this case the differences not so significant!

The total variance is the sum of the variance within groups and the variance between the groups!



Decomposition of the variance

(not neceesary to learn)

Let *N* data in *k* groups.

x_{ij} –j-th data in the i-th group,

n_i – no. of elements in the i-th group.

grand mean $\bar{x} = \frac{\sum_{i,j} x_{ij}}{N}$

mean within group $\bar{x}_i = \frac{\sum_j x_{ij}}{n_i}$

$$\left(\bar{x} - x_{ij}\right) = \left(\bar{x} - \bar{x}_i\right) + \left(\bar{x}_i - x_{ij}\right)$$

difference from the grand mean

difference of the mean within group from the grand mean

difference from the mean within group

Take the square!

$$\left(\bar{x} - x_{ij}\right)^2 = \left(\bar{x} - \bar{x}_i\right)^2 + \left(\bar{x}_i - x_{ij}\right)^2 + 2\left(\bar{x} - \bar{x}_i\right)\left(\bar{x}_i - x_{ij}\right)$$

$$\sum\left(\bar{x} - x_{ij}\right)^2 = \sum\left(\bar{x} - \bar{x}_i\right)^2 + \sum\left(\bar{x}_i - x_{ij}\right)^2 + \sum 2\left(\bar{x} - \bar{x}_i\right)\left(\bar{x}_i - x_{ij}\right)$$

covariance = 0 (indepent)

$$\sum 2\left(\bar{x} - \bar{x}_i\right)\left(\bar{x}_i - x_{ij}\right) = 0$$

$$\sum\left(\bar{x} - x_{ij}\right)^2 = \sum\left(\bar{x} - \bar{x}_i\right)^2 + \sum\left(\bar{x}_i - x_{ij}\right)^2$$

between groups

within groups

We can decompose the variance!

Nullhypothesis

There is no difference among the groups.

The difference among the averages of the groups are random deviation.

Decision: On the base of the comparison of the variance between groups and within groups.

Calculation of variances

Variance	Sum of squares	d.f.		F valsue
Among groups	$SS_A = \sum_j n_j (\bar{x}_j - \bar{x})^2$	k-1	$MS_A = \frac{SS_A}{k-1}$	$F = \frac{MS_A}{MS_E}$
Within groups	$SS_E = SS_T - SS_A$	N-k	$MS_E = \frac{SS_E}{N-k}$	
total	$SS_T = \sum_{i,j} (x_{i,j} - \bar{x})^2$	N-1		

k – no. of groups

Decision

- 1. if the probability of the random deviation is small ($p(F>F_{crit}) \leq 5\%$) – ***reject*** the nullhypothesis.
- 2. if the probability of the random deviation is large ($p(F>F_{crit}) > 5\%$) – ***accept*** the nullhypothesis.

(After decision, if it is necessary, we may do t-tests.)

Kruskal-Wallis test

Rank data without separating into groups, than sum ranks in each group!

If the variable has no normal distribution!

Nullhypothesis

There is no difference among the groups.

The difference among the averages of the groups is random deviation.

What kind of the distribution is suitable?

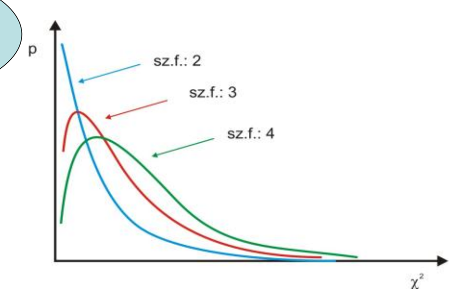
$$H = \frac{12}{N(N+1)} \sum_i \frac{R_i^2}{n_i} - 3(N+1)$$

N – no of elements
R_i – the sum of the ranks in the i-th group
n_i – no. of elementsin the i-th group

χ^2-distribution

The value of $H \geq 0$!

H variable has χ^2-distribuiton!



Decision

- 1. if the probability of the random deviation is small ($p(\chi^2 > \chi^2_{crit}) \leq 5\%$) – ***reject*** the nullhypothesis.
- 2. if the probability of the random deviation is large ($p(\chi^2 > \chi^2_{crit}) > 5\%$) – ***accept*** the nullhypothesis.

Hypothesis test?

- Set up the **nullhypothesis!**
- Look for a **variable having known distribution**.
- Calculate the **probability of the random deviation** on the base of the distribution.
- If it is smaller than the significance level **reject**, in opposite case **accept the nullhypothesis**.
- That's all!