



# **Statistics for Diploma and PhD Students "Basics"**

**-- Descriptive Statistics--**

**Josef Fritz**

**Department for Medical Statistics, Informatics and Health Economics  
Medical University Innsbruck**





# The principles of statistical testing:

Formulating Hypothesis

&

Teststatistics

&

p-values

# Formulating Hypothesis & Statistical Tests

## Steps in conducting a statistical test:

- Quantify the scientific problem from a clinical / biological perspective
- Formulate the model assumptions (distribution of the variable of interest)
- Formulate the problem as a statistical testing problem:  
Nullhypothesis versus alternative hypothesis
- Define the „error“ you are willing to tolerate
- Calculate the appropriate test statistic
- Decide for the Nullhypothesis or against it

# Formulating Hypothesis & Statistical Tests

## Hypothesis Formulation:

- Null hypothesis **H0**: The conservative hypothesis you want to reject
- Alternative Hypothesis **H1**: The hypothesis you want to proof
- Examples:

### **Scientific hypothesis:**

A new therapy is assumed to better prevent myocardial infarctions in risk patients than the old therapy.

### **Scientific hypothesis:**

Women and men achieve equally good scores in the EMS-AT test

### **Statistical hypothesis:**

$$H_0: \pi_{\text{new}} \geq \pi_{\text{old}}$$

$$H_1: \pi_{\text{new}} < \pi_{\text{old}}$$

with

$\pi_{\text{new}}$  : the proportion of patients experiencing a MI during the study receiving the new therapy

$\pi_{\text{old}}$  : the proportion of patients experiencing a MI during the study receiving the old therapy

### **Statistical hypothesis:**

$$H_0: \mu_{\text{men}} = \mu_{\text{women}}$$

$$H_1: \mu_{\text{men}} \neq \mu_{\text{women}}$$

with

$\mu_{\text{men}}$  : mean scores for men

$\mu_{\text{women}}$  : mean scores for women

# Formulating Hypothesis & Statistical Tests

## Possible decisions in statistical tests:

		Decide for	
		$H_0$	$H_1$
Reality	$H_0$	Correct decision	Wrong decision: Type I error ( $\alpha$ )
	$H_1$	Wrong decision: Type II error ( $\beta$ )	Correct decision: Power ( $1-\beta$ )

- Statistical tests are constructed in that way, that the probability of a Type I error is not bigger than the significance level  $\alpha$  (typically set to 0.01 or 0.05)

### Example:

- Test the new MI-therapy on patients to a significance level of 5%.
- In reality,  $H_0$  is true and there is no difference between therapies.
- If the study is repeated 100 times on 100 different samples, the statistical test rejects the Nullhypothesis in maximum 5 of the 100 tests.



## **The most common statistical tests:**

Testing measures of location

&

Testing frequencies

# The most common statistical tests

	Quantitative Outcome variable		Qualitative Outcome variable	
	Normal distribution	Any other distribution	Expected frequency in each cell of the crosstable „high“	Expected frequency in each cell of the crosstable „low“
Compare 2 groups	t-test	Wilcoxon-test / Mann-Whitney U-Test	Chi-Square	Fishers exact test
Compare >2 groups	Analysis of Variance (ANOVA)	Kruskal-Wallis-Test	Chi-Square	Fishers exact test

## Testing measures of location:

Does the mean/median differ between groups

## Testing frequencies in a crosstable:

Are the rows and columns independent from each other?

# Testing measures of location

## The One-sample t-test (the “standard test” for mean comparisons):

- Situation: Compare the sample mean ( $\mu_{\text{sample}}$ ) with a specified mean ( $\mu_0$ )
- Assumption: normal distribution of the sample
- Hypothesis:  $H_0: \mu_{\text{sample}} = \mu_0$  versus  $H_1: \mu_{\text{sample}} \neq \mu_0$
- Teststatistic:

$$T = \frac{\bar{X} - \mu_0}{S} \sqrt{n} \sim t(n - 1)$$

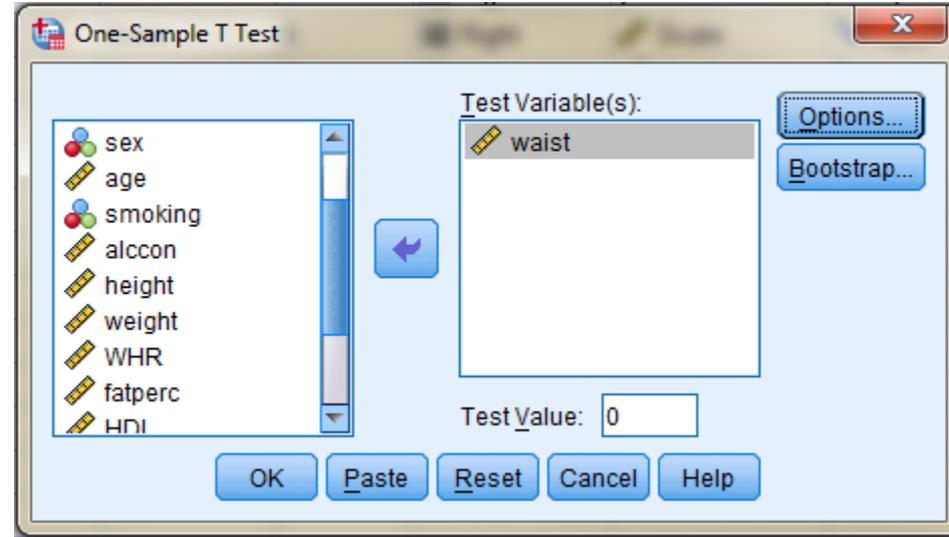
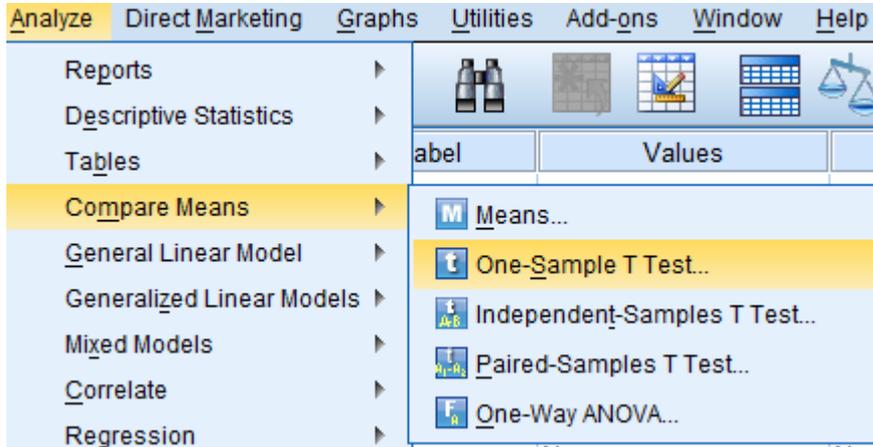
- Test decision for a two sided test:  $|T| > t_{n-1, 1-\alpha/2}$  : Reject  $H_0$
- Test decision for a one sided test:  $|T| > t_{n-1, 1-\alpha}$  : Reject  $H_0$

→ The higher T, the more likely it is that the Nullhypothesis can be rejected

- But in practice: Statistical programs give out **p-values**

# Testing measures of location

The One-sample t-test in SPSS (We use dataset “Alldata.sav”):



One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
waist	1453	90,945	12,0929	,3172

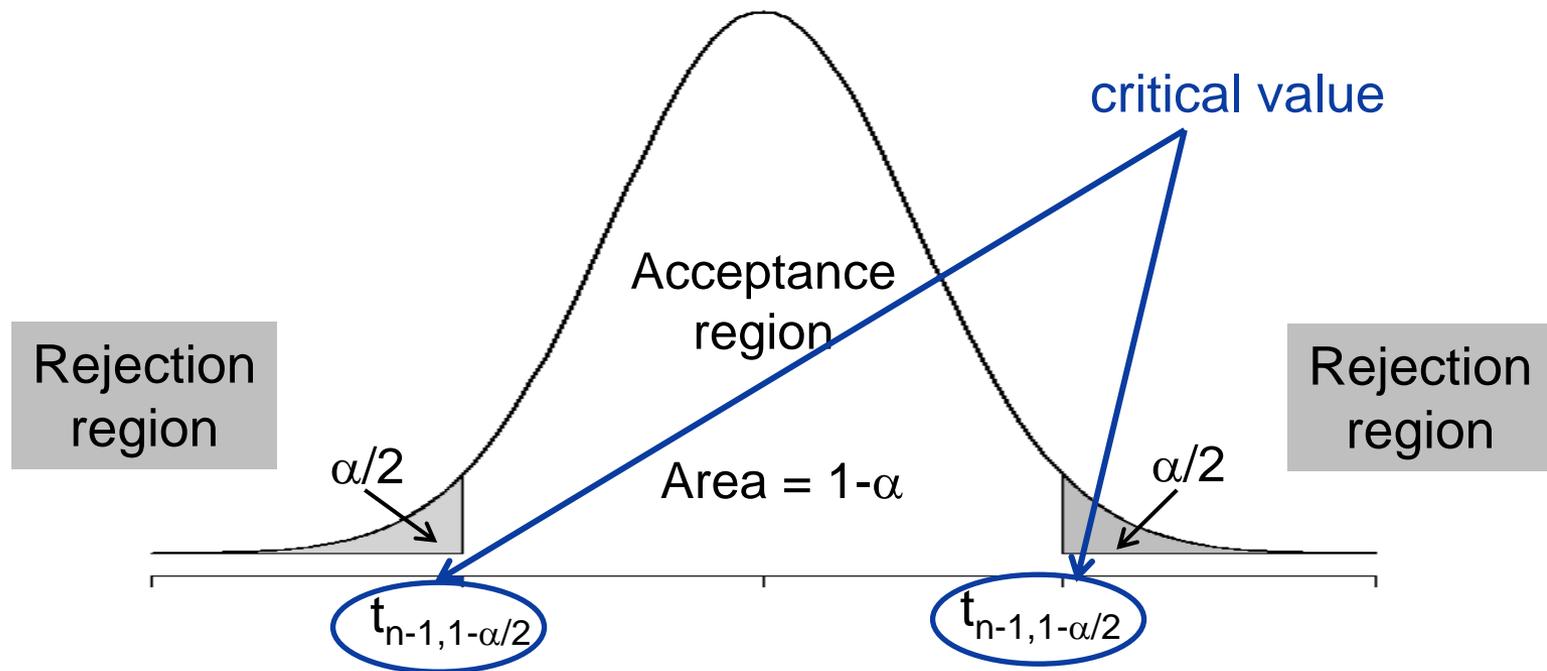
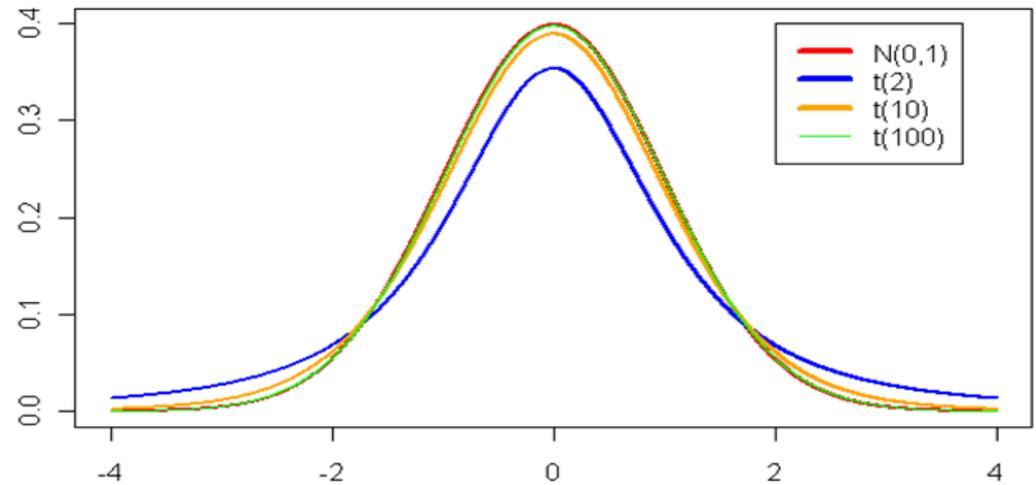
p-value

One-Sample Test

	Test Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
waist	286,669	1452	,000	90,9449	90,323	91,567

# Testing measures of location

- The T-statistic is t-distributed
- The t-distribution approximates the normal distribution:
- If a T-Statistic is very extreme (lower or higher than the critical value) → Null hypothesis will be rejected !

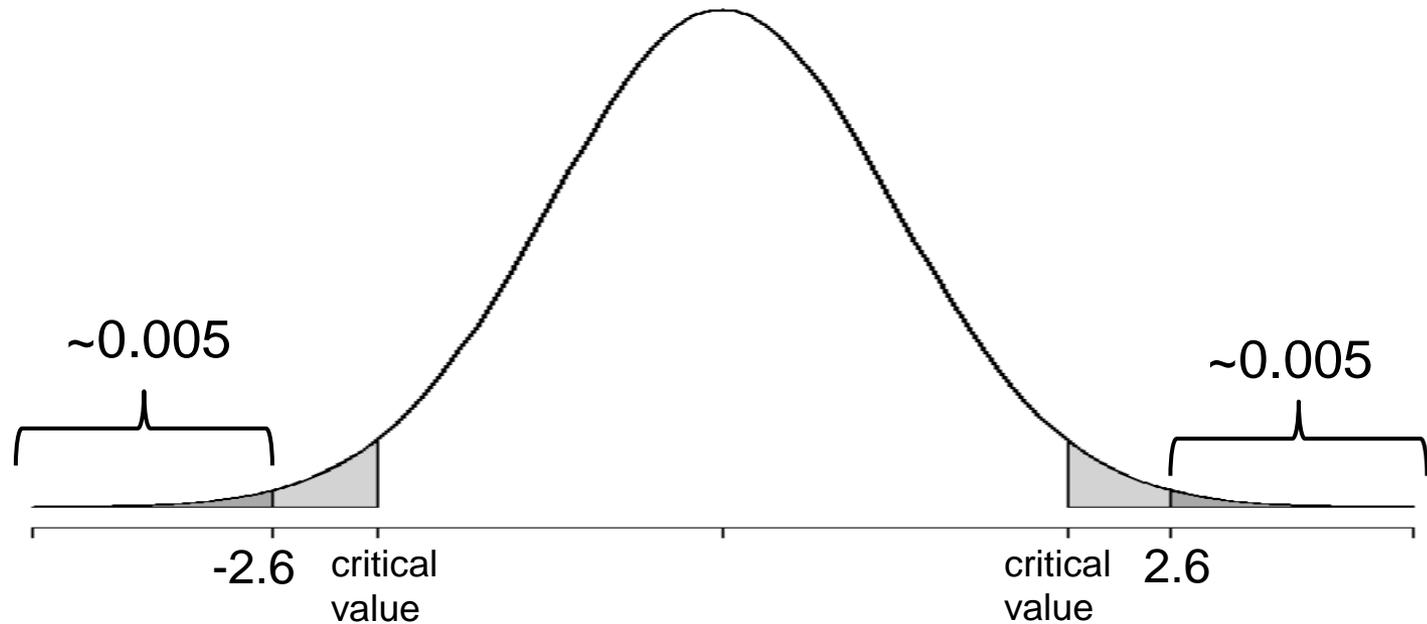


# Testing measures of location

## Example:

A one sample t-Test comparing the sample mean to 0:

$H_0: \mu_{\text{sample}} = 0$ ;  $H_1: \mu_{\text{sample}} \neq 0$  results in a test statistic  $|T| = 2.6$



P-value (one-sided test) = 0.005 (= Area under the curve)

P-value = 0.005 + 0.005 = 0.01 (= Area under the curve)

# Formulating Hypothesis & Statistical Tests

- The P-value  $p$  is a measure of certainty against the null hypothesis.

## Example:

A one sample t-Test comparing the sample mean to 0:  $H_0: \mu_{\text{sample}} = 0$ ;  $H_1: \mu_{\text{sample}} \neq 0$  results in a test statistic  $T=2.6$ , which corresponds to a p-value of 0.01.

## A popular interpretation, but wrong:

„The probability, that the sample mean is 0 is 1%“

The sample mean **does not have a probability**. It is 0 or not!

## What we want to know: is this result due to chance?

## Correct interpretation:

„A different random sample is drawn 100 times from the population of interest. The population mean is 0 (=Nullhypothesis). Maximum 1 of the 100 experiments results in a teststatistic, which is  $\geq |2.6|$ “

# Formulating Hypothesis & Statistical Tests

→ The smaller the p-value, the more certainty is given that the result is not only due to chance

→ P value 0.01: only in 1 of 100 experiments you get such a result just by chance

→ P value 0.001: only in 1 of 1000 experiments you get such a result just by chance  
→ really seldom

If  $p < \alpha$ , reject  $H_0$  → decision on the test can be based on T or p

→ In most cases, a **p-value < 0.05** (or <5%) is said to be **statistically significant** !

# Testing measures of location

## The two-sample t-test for unpaired samples:

- Situation: Compare the means ( $\mu_1, \mu_2$ ) of two unpaired samples
- Assumption: normal distribution of both samples,  $\sigma$  ( $= \sigma_1 = \sigma_2$ ) is not known

Here: Equal  $\sigma$  assumed, but there are methods (Welch t-test) for unequal  $\sigma$

### Hypothesis:

- Teststatistic: 
$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t(n_1 + n_2 - 2)$$

with the pooled variance 
$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

- Test decision for a two sided test:  $|T| > t_{n_1+n_2-2, 1-\alpha/2}$  : Reject  $H_0$
- Test decision for a one sided test:  $|T| > t_{n_1+n_2-2, 1-\alpha}$  : Reject  $H_0$

# Testing measures of location

**Example:** A biotech company claims that their new biomarker XY can distinguish diseased from non-diseased; A pilot study on 10 diseased and 10 healthy persons gives the following results:

	Labparameter XY in Diseased Patient	Labparameter XY in Healthy
	8.70	3.36
	11.28	18.35
	13.24	5.19
	8.37	8.35
	12.16	13.1
	11.04	15.65
	10.47	4.29
	11.16	11.36
	4.28	9.09
	19.54	(missing)
$\bar{X}$	11.024	9.86
$S^2$	15.227	27.038

# Testing measures of location

**Example:** A biotech company claims that their new biomarker XY can distinguish diseased from non-diseased; A pilot study on 10 diseased and 10 healthy persons gives the following results:

Labparameter XY in Diseased	Labparameter XY in Healthy Individ.	
8.70	3.36	
11.28	18.35	
13.24	5.19	
8.37	8.35	
12.16	13.1	
11.04	15.65	
10.47	4.29	
11.16	11.36	
4.28	9.09	
19.54	(missing)	
$\bar{X}$	11.024	9.86
$S^2$	15.227	27.038

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

$$s^2 = (9 \cdot 15.227 + 8 \cdot 27.038) / 17 = 20.78512$$

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t(n_1 + n_2 - 2)$$

$$T = \frac{(11.024 - 9.86)}{\sqrt{(20.78512 \cdot (1/10 + 1/9))}} = 0.556 \sim t(17)$$

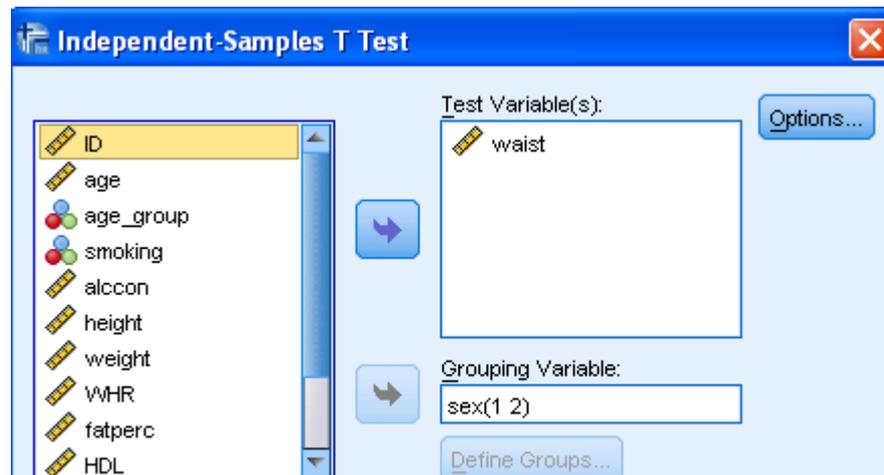
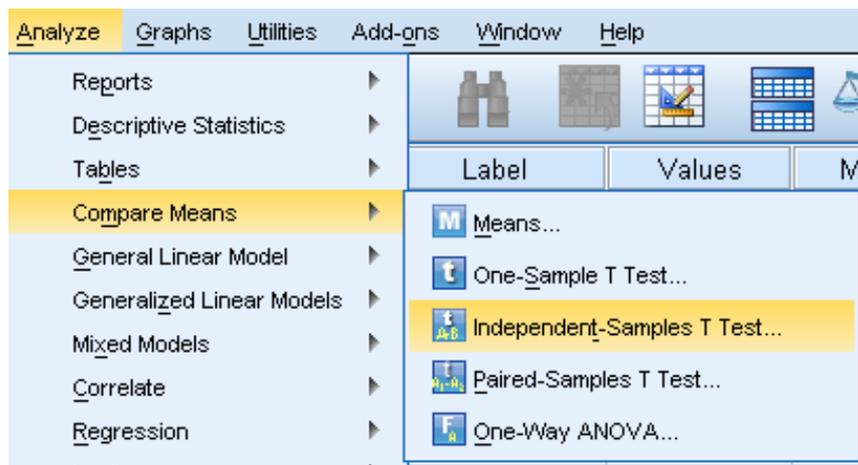
Critical value of a t(17)-distribution to  $\alpha = 5\%$  (two-sided test) = 2.11

$0.556 < 2.11 \rightarrow$  XY does not differ between diseased and non-diseased

P-value = 0.29

# Testing measures of location

- How to do unpaired T-Test in SPSS (We use dataset "Alldata.sav"):



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
waist	Equal variances assumed	23,266	,000	-23,787	1451	,000	-12,8064	,5384	-13,8625	-11,7503
	Equal variances not assumed			-23,811	1420,653	,000	-12,8064	,5378	-13,8615	-11,7514

P-value

# Testing measures of location

## The two-sample t-test for paired samples:

- Situation: Compare the means of two paired samples, e.g. compare the means of variables in the same patients before a treatment and after the treatment
- Assumption: normal distribution of both samples,  $\sigma$  ( $= \sigma_1 = \sigma_2$ ) is not known
- Hypothesis:  $H_0: \mu_{\text{before}} = \mu_{\text{after}}$  versus  $H_1: \mu_{\text{before}} \neq \mu_{\text{after}}$

Calculate  $d = x_{\text{before}} - x_{\text{after}}$  for each patient

→ new Hypothesis:  $H_0$ : The mean of the difference is 0:  $\mu_d = 0$

versus  $H_1$ : The mean of the difference is  $\neq 0$ :  $\mu_d \neq 0$

T-tests can also be used approximatively for any distribution, that is not too skewed.

# Testing measures of location

**Example:** A doctor claims, that he has invented the perfect weight loss method; A pilot study on 10 obese individuals gives the following results:

ID	kg at baseline	kg after 6 months	Difference
1	108	90	18
2	97	97	0
3	88	91	-3
4	120	111	9
5	98	94	4
6	95	91	4
7	87	82	5
8	85	77	8
9	99	103	-4
10	134	127	7
$\bar{X}$	101.1	96.3	4.8
$S^2$	242.767	209.122	41.07 → s = 6.41

# Testing measures of location

**Example:** A doctor claims, that he has invented the perfect weight loss method; A pilot study on 10 obese individuals gives the following results:

ID	kg at baseline	kg after 6 months	Difference
1	108	90	18
2	97	97	0
3	88	91	-3
4	120	111	9
5	98	94	4
6	95	91	4
7	87	82	5
8	85	77	8
9	99	103	-4
10	134	127	7
$\bar{X}$	101.1	96.3	4.8
$S^2$	242.767	209.122	41.07 → s = 6.41

Paired t-test =

$$T = \frac{\bar{X} - \mu_0}{S} \sqrt{n} \sim t(n - 1)$$

$$T = \sqrt{10} * (4.8 - 0) / 6.41 = 2.368$$

$$t_{0.975}(9) = 2.262 \text{ (two-sided)}$$

→  $H_0$  can be rejected ( $p = 0.042$ )

Since you want to prove, that  
kg(before) > kg(after):

→ one-sided test more appropriate (more power)

$$\rightarrow t_{0.95}(9) = 1.833, p = 0.021$$

If you would have done a „normal“  
unpaired t-test:

$p = 0.484$  →  $H_0$  can not be rejected !

# Testing measures of location

- How to do a paired T-Test in SPSS: (We use test4.sav)

The screenshot shows the SPSS 'Paired-Samples T Test' dialog box. The 'Analyze' menu is open, and 'Paired-Samples T Test...' is selected. The dialog box shows a list of variables on the left, with 'P1\_2.after' selected. The 'Paired Variables' table on the right shows two pairs: Pair 1 with 'P1.after' and 'P1.bef...', and Pair 2 with empty fields. The 'OK' button is highlighted.

Pair	Variable1	Variable2
1	[P1.after]	[P1.bef...]
2		

In this example:

P1.before: First measurement of one parameter

P1.after: Second measurement of the same parameter in the same patient

# Testing measures of location

## ■ Result of a paired T-Test in SPSS

### T-Test

[DataSet1] P:\Public\GENEPI\VO\_Claudia\_Barbara\Basic Statistik VO Biostatistik\WS 2013\_2014\Tag 2\test4.sav

Paired Samples Statistics

	Mean	N	Std. Deviation	Std. Error Mean
Pair 1 P1_2.before	3,0000	3	1,00000	,57735
P1_2.after	6,6667	3	1,52753	,88192

Paired Samples Correlations

	N	Correlation	Sig.
Pair 1 P1_2.before & P1_2.after	3	,655	,546

Paired Samples Test

	Paired Differences						t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference					
				Lower	Upper				
Pair 1 P1_2.before - P1_2.after	-3,66667	1,15470	,66667	-6,53510	-,79823	-5,500	2	,032	

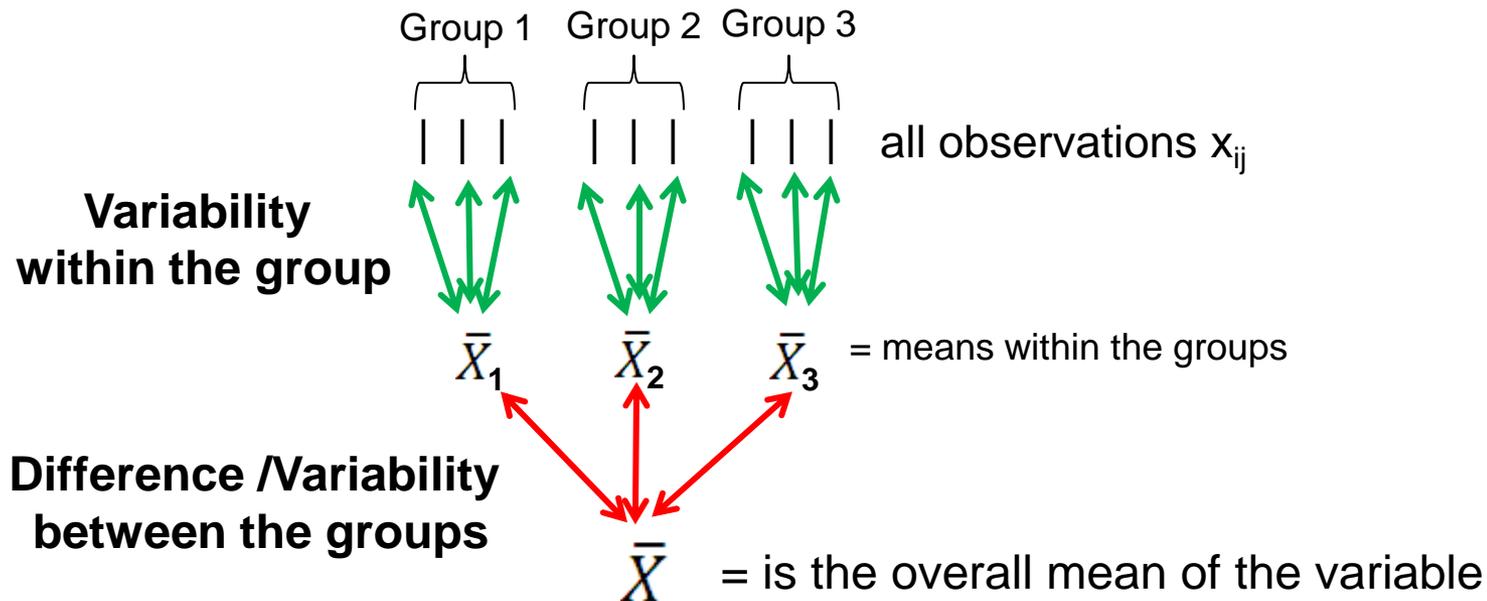
→ Difference between two means (before – after)

P-value

# Testing measures of location

## Analysis of Variance (ANOVA)

- Situation: Compare the means of  $k$  samples ( $k > 2$ )
- Assumption: normal distribution of the population,  $\sigma = \sigma_1 = \sigma_2 = \dots = \sigma_k$
- Hypothesis:  $H_0: \mu_1 = \mu_2 = \dots = \mu_k$  versus  $H_1: \mu_i \neq \mu_j$  ( $i \neq j$ ): At least two of the means differ



# Testing measures of location

- **Test statistic:**

$$F = \frac{S_{between}^2}{S_{within}^2} \sim F(k - 1, n - k)$$

- **Test decision** for a two sided test:  $F > F_{k-1, n-k, 1-\alpha}$  : Reject  $H_0$
- If  $H_0$  is rejected, you can tell, that there are at least two groups, which differ from each other significantly. You can't tell, which groups differ!  
→ perform pairwise t-tests after overall F-Test

## Example:

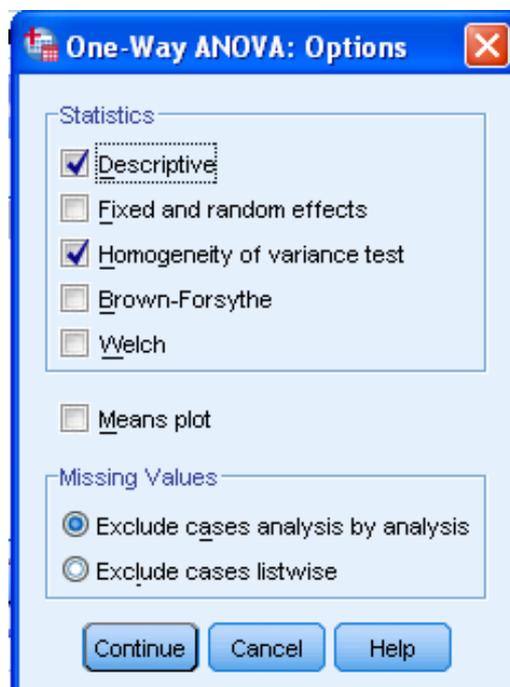
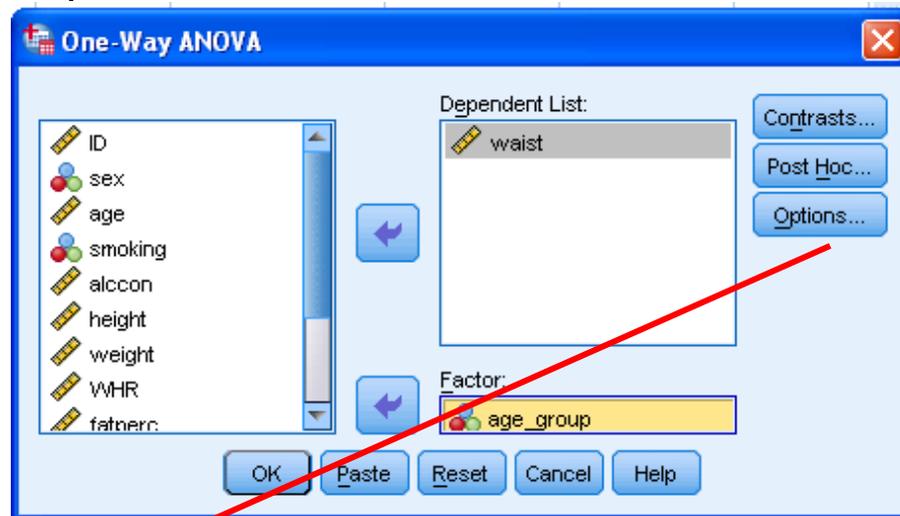
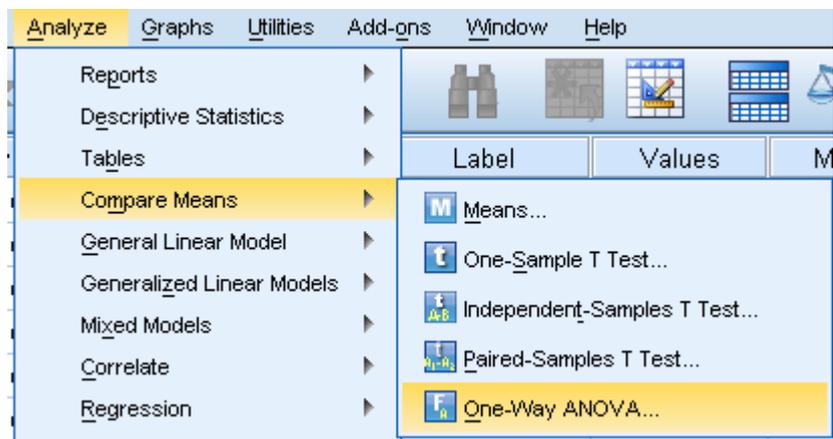
There are 3 different medications (Med1, Med2, Med3), which are intended to increase the HDL-cholesterol levels in patients

1. perform ANOVA as an overall test, if there is a difference between the groups
2. If the F-Test was significant, you know, that there is a difference
3. Test Med1 against Med2, Med1 against Med3, Med2 against Med3

→ If there are more than 3 groups this can not be done that way (e.g. ANOVA, Tukey test)

# Testing measures of location

How to do an ANOVA in SPSS (We use dataset "Alldata.sav"):



# Testing measures of location

## ■ Results of an ANOVA in SPSS: mean waist per age group:

### Descriptives

waist

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
					30-40	144		
41-50	372	89,171	12,4786	,6470	87,898	90,443	61,0	132,0
51-60	594	91,717	11,4026	,4679	90,798	92,636	65,0	132,0
61-70	343	94,590	11,3279	,6116	93,387	95,793	64,0	131,0
Total	1453	90,945	12,0929	,3172	90,323	91,567	61,0	132,0

### Test of Homogeneity of Variances

waist

Levene Statistic	df1	df2	Sig.
1,716	3	1449	,162

Homogeneity of variances is fulfilled  
 →  $p > 0.05$

### ANOVA

waist

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	13726,175	3	4575,392	33,380	,000
Within Groups	198611,920	1449	137,068		
Total	212338,095	1452			

P-value

# Testing measures of location

## ■ Results of an ANOVA in SPSS: mean waist per age group:

### Descriptives

waist

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
30-40	144	83,660	11,7830	,9819	81,719	85,601	64,5	120,0
41-50	372	89,171	12,4786	,6470	87,898	90,443	61,0	132,0
51-60	594	91,717	11,4026	,4679	90,798	92,636	65,0	132,0
61-70	343	94,590	11,3279	,6116	93,387	95,793	64,0	131,0
Total	1453	90,945	12,0929	,3172	90,323	91,567	61,0	132,0

### ANOVA

waist

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	13726,175	3	4575,392	33,380	,000
Within Groups	198611,920	1449	137,068		
Total	212338,095	1452			

P-value

We only know, that there is a difference between the groups, but not between which groups! → post-hoc tests (e.g.: Tukey: all pairwise comparisons; e.g.: Dunnett: assumes one reference group, e.g. the gold standard)

# Testing measures of location

All tests so far assumed a normally distributed variable → **parametric tests:**

- Should be preferred over nonparametric test, if appropriate, since they have the higher power

If the assumption does not hold → **nonparametric tests:**

- Application often for data that are rather ranks instead of numeric
- Robust against outliers and skewed distributions

Parametric Tests	Nonparametric Tests
T-Test	<ul style="list-style-type: none"><li>• Wilcoxon-Test</li><li>• Wilcoxon rank-sum test</li><li>• Mann-Whitney U-Test</li></ul>
ANOVA	Kruskal-Wallis-Test

# Testing measures of location

## Two sample test on equality of distributions: Wilcoxon Test

- Situation: Compare location measures of two unpaired samples X and Y, if the assumption of a t-test does not hold
- Assumption: the form of the continuous distributions of the variables X and Y is the same  $\rightarrow$  test on equality of distributions = test on equality of the medians
- Hypothesis:  $H_0: x_{med} = y_{med}$  versus  $H_1: x_{med} \neq y_{med}$
- Test is based on the ranks
- What are ranks?

# Testing measures of location

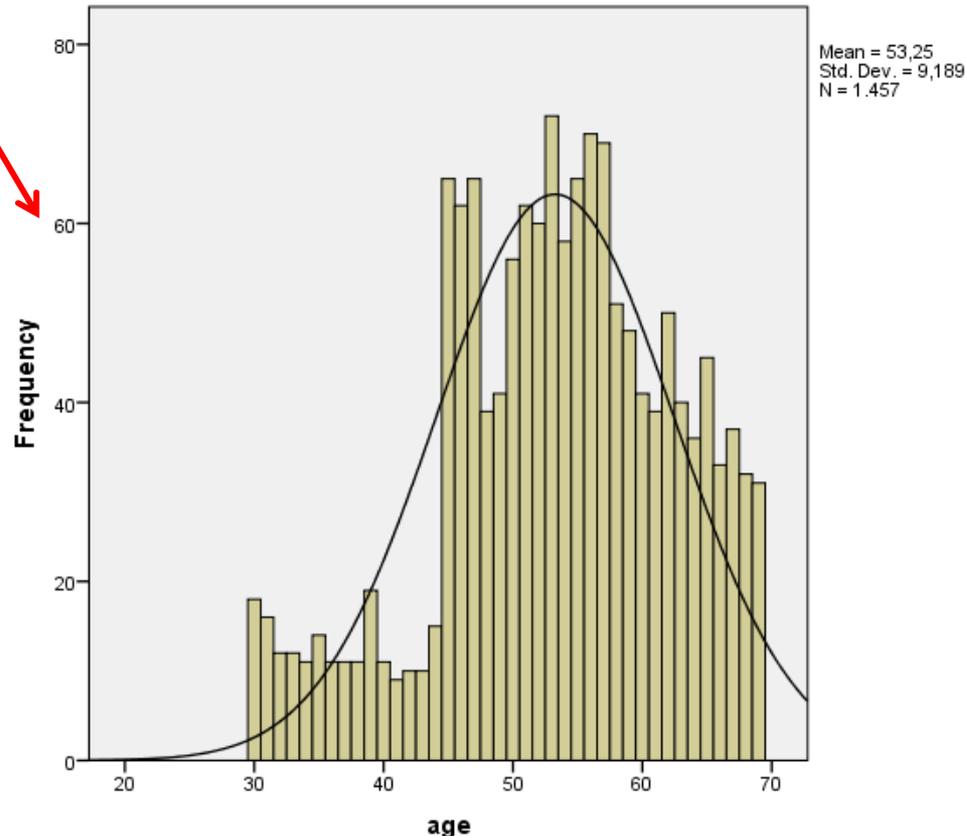
## Example: Wilcoxon Test

Original values:  $X=\{1,2,4,6,9,9,11\}$ ;  $x_{med}=6$ ;  $Y=\{1,3,4,5,6,7,8\}$ ;  $y_{med}=5$

- Sort both variables into one: 1/1,2,3,4/4,5,6/6,7,8,9/9,11
- Ranking: 1.5,1.5,3,4,5.5,5.5,7,8.5,8.5,10,11,12.5,12.5,14
- Sum the ranks:  $R_X=57.5$ ;  $R_Y=47.5$
- Teststatistic is calculated and p values are given out

# Testing measures of location

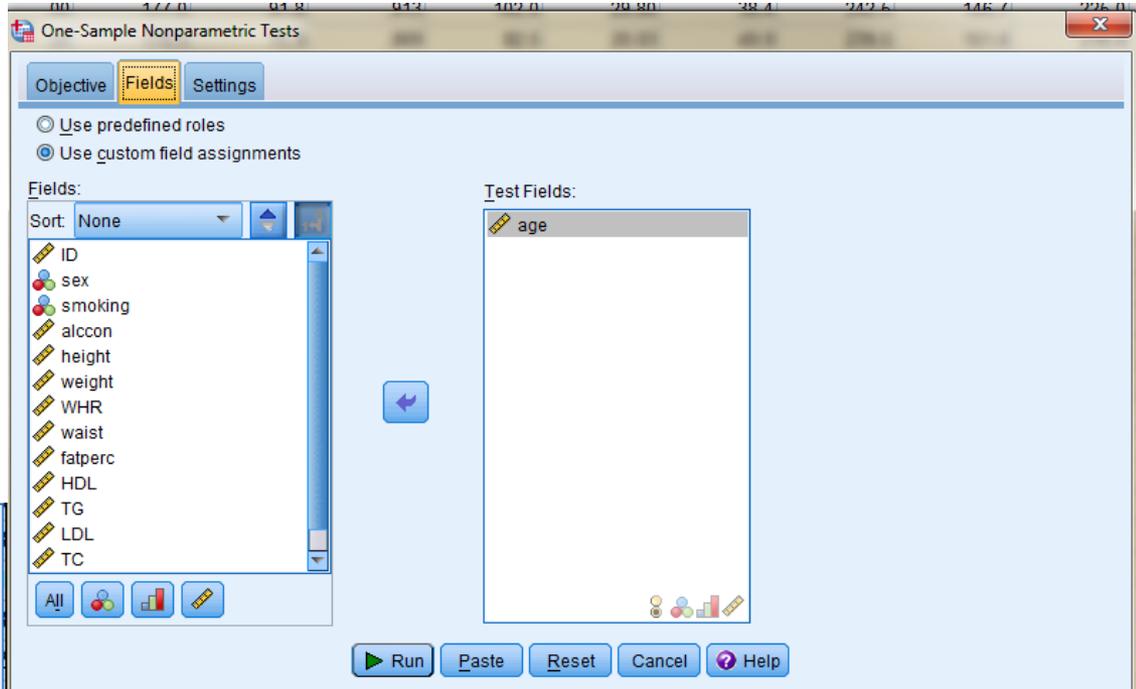
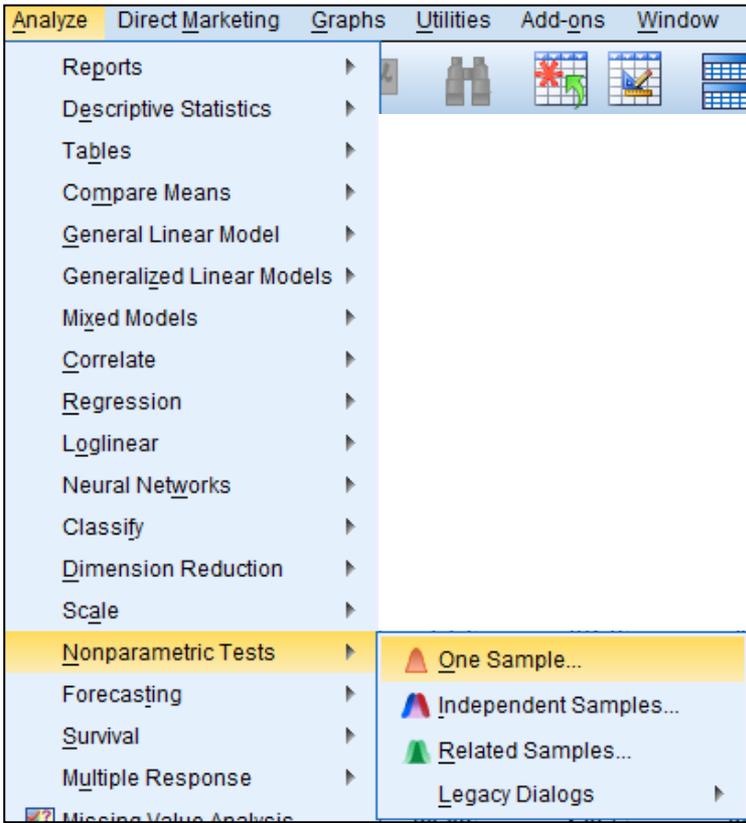
- Calculate the Wilcoxon Test in SPSS
- First: Why do this test here? How can you check this?  
→ Histogram and Kolmogorov-Smirnov Test to check if the variable is normally distributed



# Testing measures of location

→ Kolmogorov-Smirnov Test to check if the variable is normally distributed

(We use dataset “Alldata.sav”):



# Testing measures of location

→ Kolmogorov-Smirnov Test to check if the variable is normally distributed

What is the

Null Hypothesis of this test?



P-value

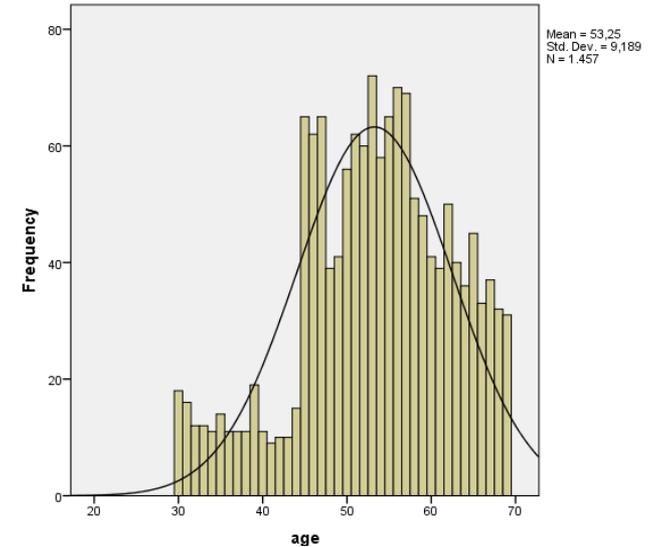


## Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of age is normal with mean 53,248 and standard deviation 9,19.	One-Sample Kolmogorov-Smirnov Test	,000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

How is the test called ?

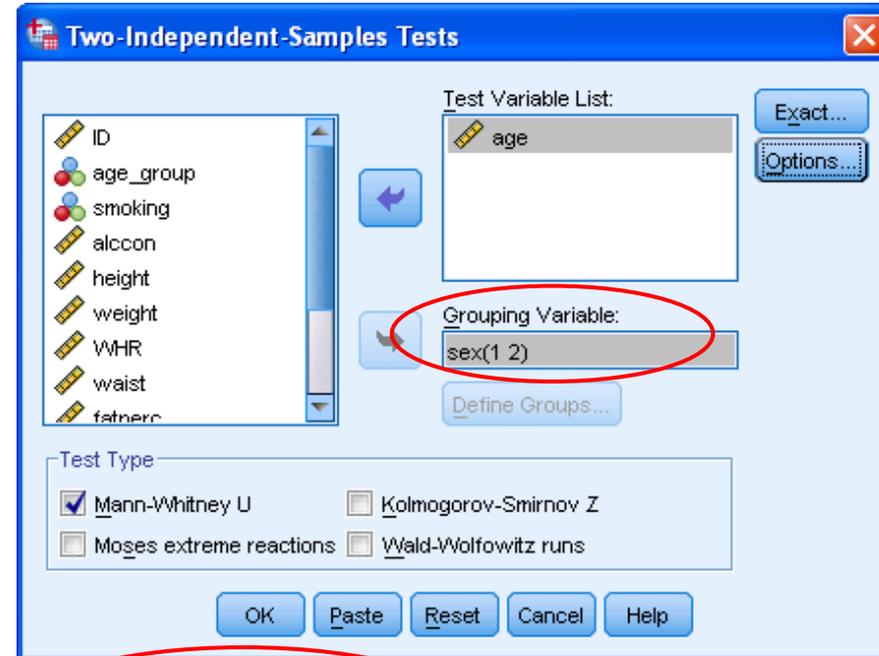
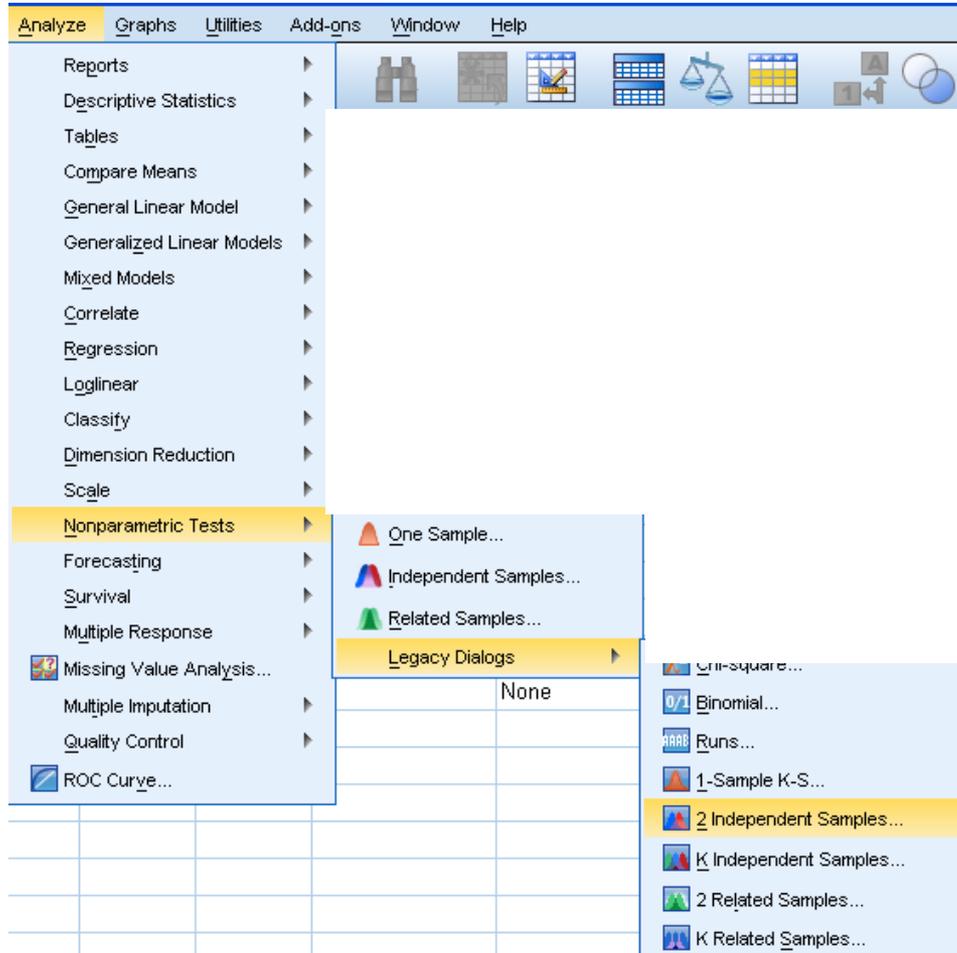


Conclusion

- normality assumption is not fulfilled
- Perform nonparametric tests

# Testing measures of location

How to calculate the Wilcoxon Test in SPSS (We use dataset "Alldata.sav"):

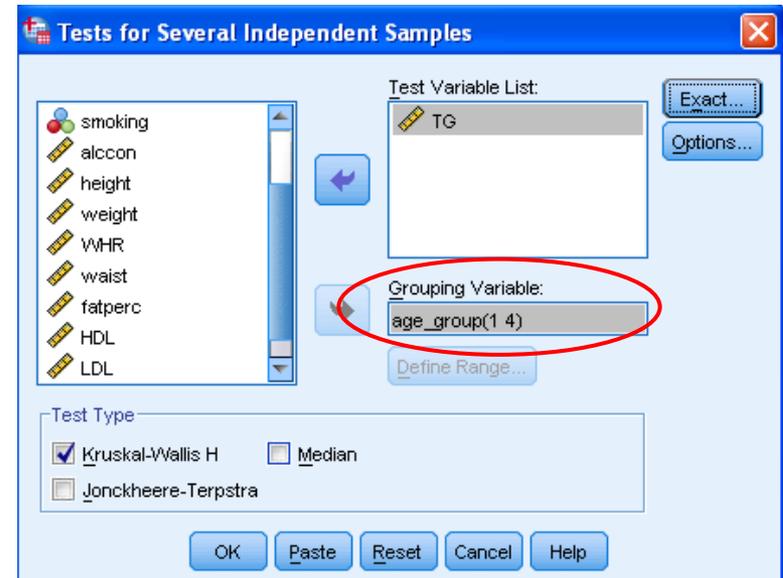
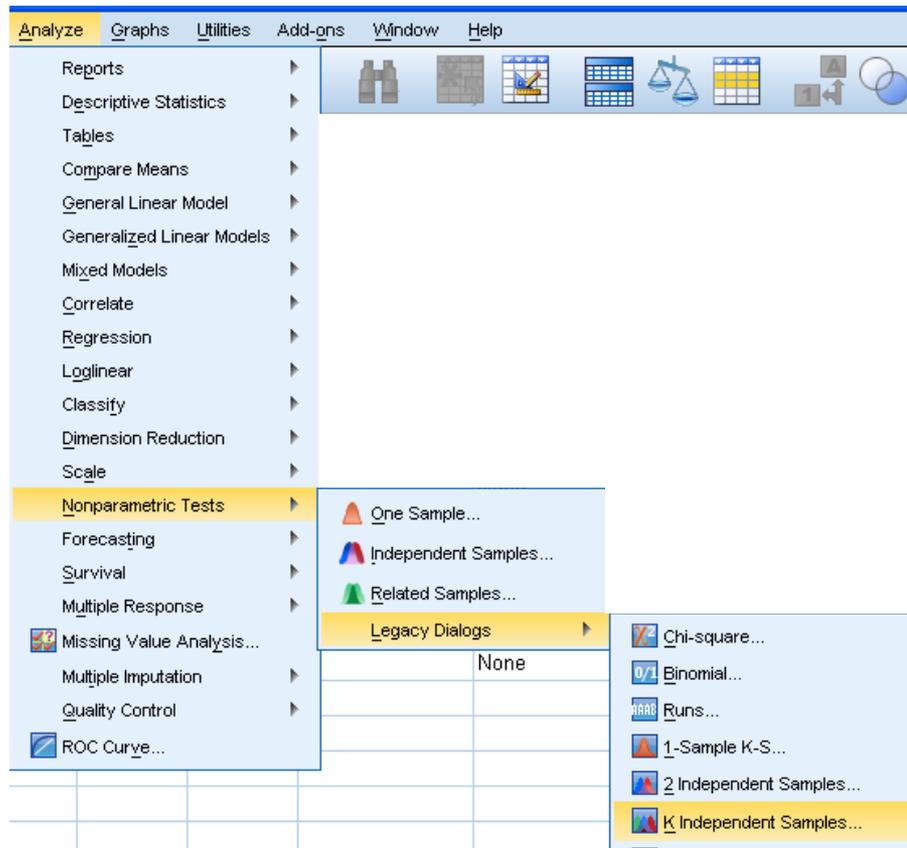


Grouping variable with:

2 categories: **Wilcoxon / Mann-Whitney U-Test**: Does age differ between men and women ?

# Testing measures of location

## ■ How to calculate the Kruskal Wallis Test in SPSS:



Grouping variable with:

> 2 categories: **Kruskal-Wallis test**: Does TG differ between age groups?

# Testing frequencies

## Two sample test on frequencies: $\chi^2$ -test of independence:

- Situation: Compare the frequencies between two groups

Or: Test, if two categorical variables  $X$  ( $i=1, \dots, k$ ) and  $Y$  ( $j=1, \dots, m$ ) depend on each other

} All situations you can group into contingency tables

		Y			
		1	...	m	Row sum
X	1	$h_{11}$	...	$h_{1m}$	$h_{1.}$
	2	$h_{21}$	...	$h_{2m}$	$h_{2.}$
	:	:		:	:
	k	$h_{k1}$	...	$h_{km}$	$h_{k.}$
Column sum		$h_{.1}$		$h_{.m}$	n

- A possible scenario: Compare the number of smokers, ex-smokers and never-smokers (e.g. Y) between men and women (e.g. X)

# Testing frequencies

## Two sample test on frequencies: $\chi^2$ -test of independence:

- Hypothesis:  $H_0$ : X and Y are independent from each other

$H_1$ : X and Y are dependent from each other (are associated)

- Assumption: expected frequencies  $\geq 1$  for all & expected frequencies  $\geq 5$  for at least 80% of the cells

→ none of the cells should have a very rare expectancy

→ if assumption is not fulfilled → use Fishers exact test (also given out by SPSS)

- Idea to construct the teststatistic: Compare the observed numbers in each cell with the expected numbers, if  $H_0$  and therefore **independence of the two factor variables** is assumed

# Testing frequencies

Table of observed numbers

		Y			$\Sigma$
		1	...	m	
X	1	$h_{11}$	...	$h_{1m}$	$h_{1.}$
	2	$h_{21}$	...	$h_{2m}$	$h_{2.}$
				:	:
	k	$h_{k1}$	...	$h_{km}$	$h_{k.}$
$\Sigma$		$h_{.1}$		$h_{.m}$	n

$h_{1.} \dots h_{k.}$ ,  $h_{.1} \dots h_{.m}$  are the margin probabilities

	Smoking status (Y)	Current Smoker	Ex-Smoker	Never Smoker	Row Total
Gender (X)					
Men		144	310	268	722
Women		117	143	475	735
Column Total		261	453	743	1457

# Testing frequencies

**X= Gender**

**Y= Smoking**

Table of expected numbers:

		Y			
		1	...	m	$\Sigma$
X	1	$h_{1.}h_{.1}/n$	...	$h_{1.}h_{.m}/n$	$h_{1.}$
	2	$h_{2.}h_{.1}/n$	...	$h_{2.}h_{.m}/n$	$h_{2.}$
				:	:
	k	$h_{k.}h_{.1}/n$	...	$h_{k.}h_{.m}/n$	$h_{k.}$
$\Sigma$	$h_{.1}$			$h_{.m}$	$n$

	Smoking status	Current Smoker	Ex-Smoker	Never Smoker	Row Total
Gender					
Men		144	310	268	722
Women		117	143	475	735
Column Total	261	453	743	1457	

Expected number in each cell:

$$(\text{Row sum}) * (\text{Columns sum}) / \text{Total sum}$$

Expected number in the upper left cell:

$$722 * 261 / 1457 = 129.336$$

# Testing frequencies

Observed:

		Y			
		1	...	m	$\Sigma$
X	1	$h_{11}$	...	$h_{1m}$	$h_{1.}$
	2	$h_{21}$	...	$h_{2m}$	$h_{2.}$
				:	:
	k	$h_{k1}$	...	$h_{km}$	$h_{k.}$
$\Sigma$		$h_{.1}$		$h_{.m}$	n

$O_{ij}$

Expected:

		Y			
		1	...	m	$\Sigma$
X	1	$h_{1.}h_{.1}/n$	...	$h_{1.}h_{.m}/n$	$h_{1.}$
	2	$h_{2.}h_{.1}/n$	...	$h_{2.}h_{.m}/n$	$h_{2.}$
				:	:
	k	$h_{k.}h_{.1}/n$	...	$h_{k.}h_{.m}/n$	$h_{k.}$
$\Sigma$		$h_{.1}$		$h_{.m}$	n

$E_{ij}$

$$\text{Teststatistic: } \chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2((k-1)(m-1))$$

$$\text{Test decision: } \chi^2 > \chi_{1-\alpha}^2((k-1)(m-1)) : \text{Reject } H_0$$

# Testing frequencies

## Example:

Observed:

Smoking status	Current Smoker	Ex-Smoker	Never Smoker	Row Total
Gender				
Men	144	310	268	722
Women	117	143	475	735
Column Total	261	453	743	1457

Expected:

Smoking status	Current Smoker	Ex-Smoker	Never Smoker	Row Total
Gender				
Men	129.336	224.479	368.185	722
Women	131.664	228.521	374.815	735
Column Total	261	453	743	1457

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2((k-1)(m-1))$$

$$= (144-129.336)^2/129.336 + (310-224.479)^2/224.479 + (268-368.185)^2/368.185 + \\ + (117-131.664)^2/131.664 + (143-228.521)^2/228.521 + (475-374.815)^2/374.815 = 121.9218$$

$$\chi^2((2-1)*(3-1)) = \chi^2(2) = 5.99 \rightarrow 121.9218 \gg 5.99 \rightarrow \text{test is significant (p = 3.3e-27)}$$

→ the Null-Hypothesis, that gender and smoking status are independent can be rejected → gender and smoking are associated

# Testing frequencies

- How to calculate the Chi-squared-test of independence in SPSS (“We use dataset “Alldata.sav”):

The image shows the SPSS Crosstabs dialog box and its sub-dialogs. The main dialog box is titled "Crosstabs" and shows the following settings:

- Row(s): sex
- Column(s): age\_group
- Layer 1 of 1

The "Crosstabs: Statistics" sub-dialog box is open, showing the following settings:

- Chi-square
- Correlations
- Nominal**
  - Contingency coefficient
  - Phi and Cramer's V
  - Lambda
  - Uncertainty coefficient
- Ordinal**
  - Gamma
  - Somers' d
  - Kendall's tau-b
  - Kendall's tau-c
- Nominal by Interval**
  - Eta
- Kappa
- Risk
- McNemar
- Cochran's and Mantel-Haenszel statistics
- Test common odds ratio equals: 1

The "Crosstabs: Cell Display" sub-dialog box is open, showing the following settings:

- Counts**
  - Observed
  - Expected
  - Hide small counts (Less than 5)
- z-test**
  - Compare column proportions
  - Adjust p-values (Bonferroni method)
- Percentages**
  - Row
  - Column
  - Total
- Residuals**
  - Unstandardized
  - Standardized
  - Adjusted standardized
- Noninteger Weights**
  - Round cell counts
  - Round case weights
  - Truncate cell counts
  - Truncate case weights
  - No adjustments

# Testing frequencies

## ■ Result of Chi-squared-test of independence in SPSS:

sex \* smoking Crosstabulation

			smoking			Total
			Current smoker	Ex smoker	Never smoker	
sex	female	Count	117	143	475	735
		% within sex	15,9%	19,5%	64,6%	100,0%
		% within smoking	44,8%	31,6%	63,9%	50,4%
	male	Count	144	310	268	722
		% within sex	19,9%	42,9%	37,1%	100,0%
		% within smoking	55,2%	68,4%	36,1%	49,6%
Total		Count	261	453	743	1457
		% within sex	17,9%	31,1%	51,0%	100,0%
		% within smoking	100,0%	100,0%	100,0%	100,0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	121,922 <sup>a</sup>	2	,000
Likelihood Ratio	124,164	2	,000
Linear-by-Linear Association	62,436	1	,000
N of Valid Cases	1457		

P-value

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 129,34.

Assumption for Chi-Square test is fulfilled. If not: Fishers exact test !

# Testing frequencies

## Fishers exact test:

**Chi-Square Tests**

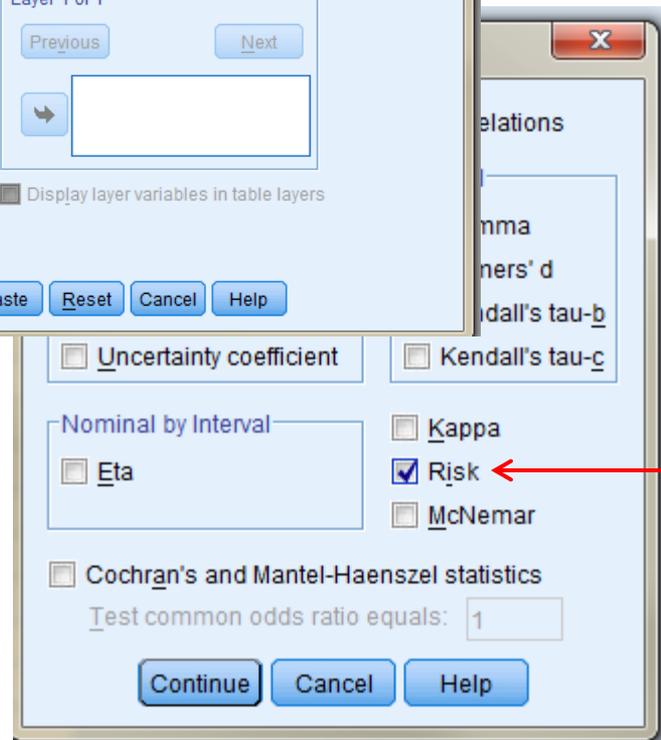
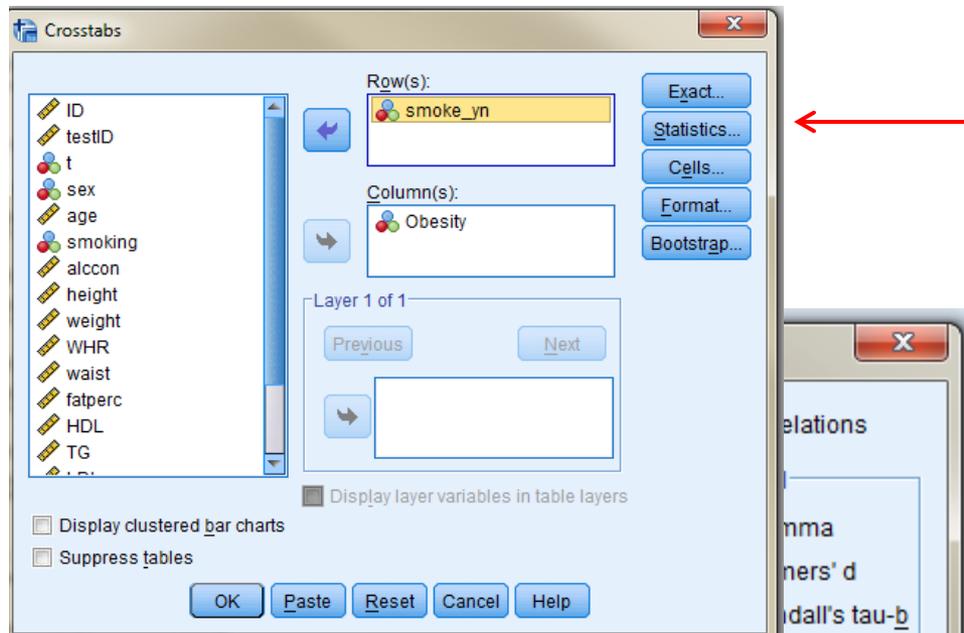
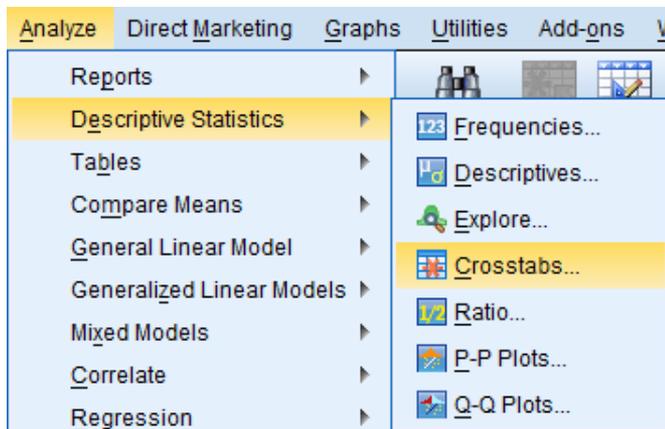
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	121,922 <sup>a</sup>	2	,000	,000		
Likelihood Ratio	124,164	2	,000	,000		
Fisher's Exact Test	123,930			,000		
Linear-by-Linear Association	62,436 <sup>b</sup>	1	,000	,000	,000	,000
N of Valid Cases	1457					

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 129,34.

b. The standardized statistic is -7,902.

# Calculate Odds Ratio out of a 2x2 contingency table

- Following variables are used: smoke\_yn (1=current smoker, 0=ex-and non-smoker; Obesity:  $\geq 30$  BMI = 1,  $<30$  BMI =0)



# Calculate Odds Ratio out of a 2x2 contingency table

smoke\_yn \* Obesity Crosstabulation

			Obesity		Total
			<30 BMI	>=30 BMI	
smoke_yn	ex- and non-smoker	Count	915	279	1194
		% within smoke_yn	76,6%	23,4%	100,0%
		% within Obesity	81,1%	85,3%	82,1%
	current smoker	Count	213	48	261
		% within smoke_yn	81,6%	18,4%	100,0%
		% within Obesity	18,9%	14,7%	17,9%
Total	Count	1128	327	1455	
	% within smoke_yn	77,5%	22,5%	100,0%	
	% within Obesity	100,0%	100,0%	100,0%	

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for smoke_yn (ex- and non-smoker / current smoker)	,739	,526	1,039
For cohort Obesity = <30 BMI	,939	,879	1,003
For cohort Obesity = >=30 BMI	1,271	,965	1,673
N of Valid Cases	1455		

**Odds ratio:** The odds ratio is a ratio of event odds. The odds of an event is the ratio of the probability that the event occurs, to the probability that the event does not occur.

# Calculate Odds Ratio out of a 2x2 contingency table

smoke\_yn \* Obesity Crosstabulation

			Obesity		Total
			<30 BMI	>=30 BMI	
smoke_yn	ex- and non-smoker	Count	915	279	1194
		% within smoke_yn	76,6%	23,4%	100,0%
		% within Obesity	81,1%	85,3%	82,1%
	current smoker	Count	213	48	261
		% within smoke_yn	81,6%	18,4%	100,0%
		% within Obesity	18,9%	14,7%	17,9%
Total	Count	1128	327	1455	
	% within smoke_yn	77,5%	22,5%	100,0%	
	% within Obesity	100,0%	100,0%	100,0%	

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for smoke_yn (ex- and non-smoker / current smoker)	,739	,526	1,039
For cohort Obesity = <30 BMI	,939	,879	1,003
For cohort Obesity = >=30 BMI	1,271	,965	1,673
N of Valid Cases	1455		

**Odds ratio** is calculated by:

Odds that ex-/non smokers have BMI < 30 is:  $76.6\%/23.4\%=3.27$

Odds that smokers have BMI < 30 is:  $81.6\%/18.4\%=4.43$

Odds ratio =  $3.27/4.43 \sim 0.739$

→ ex-/non smokers have a lower probability for BMI < 30 compared to smokers, but **not significant**

→ why? → **"1"** is contained in the **confidence interval**:

0.739 (0.526-1.039)

# Calculate Odds Ratio out of a 2x2 contingency table

- The values of the odds ratio can be interpreted as follows:

<b>OR = 1</b>	The exposition is not associated with the disease
<b>OR &gt; 1</b>	Positive association of exposition with the disease, higher probability for the disease
<b>OR &lt; 1</b>	Negative association of exposition with the disease, lower probability for the disease



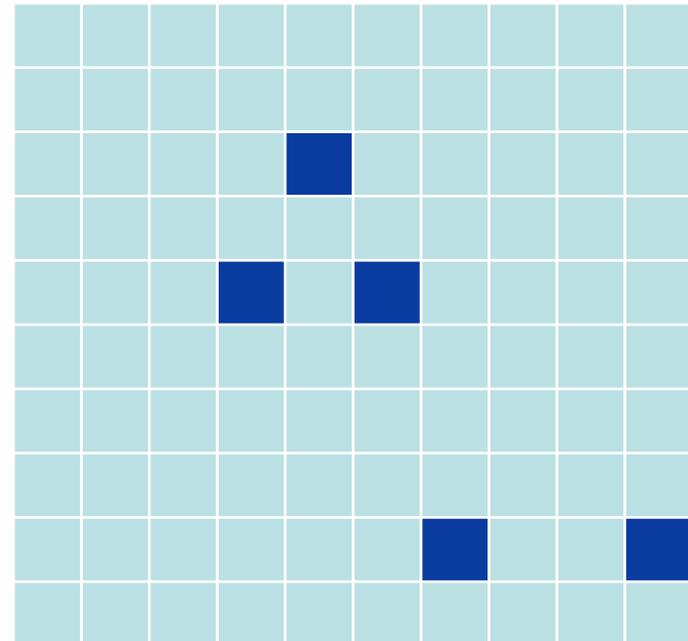
# **The multiple testing problem**

# The multiple testing problem

## The situation:

- Consider a dataset with 100 independent parameters, which do not play a role in the etiology of the disease of interest (what you don't know, of course)
- 100 statistical tests are performed with a significance level of  $\alpha=0.05$
- The tests are constructed in that way, that maximum 5 of 100 tests reject the Nullhypothesis, although it is true (which is the case in this example)

→ You expect 5 tests to be significant just by chance



# The multiple testing problem

- The probability to get at least one Type I error increases with increasing number of tests.
- **Family-wise error rate** (the error rate for the complete family of tests performed):  $\alpha^* = 1 - (1 - \alpha)^k$ , with  $\alpha$  being the **comparison-wise error rate**

k	$\alpha^*$ ( $\alpha=0.05$ )
1	0.05
5	0.226
10	0.401
100	0.994

The probability to get one or more false discoveries (Type I error)

→ The significance level has to be modified for multiple testing situations

# The multiple testing problem

## The Bonferroni correction method:

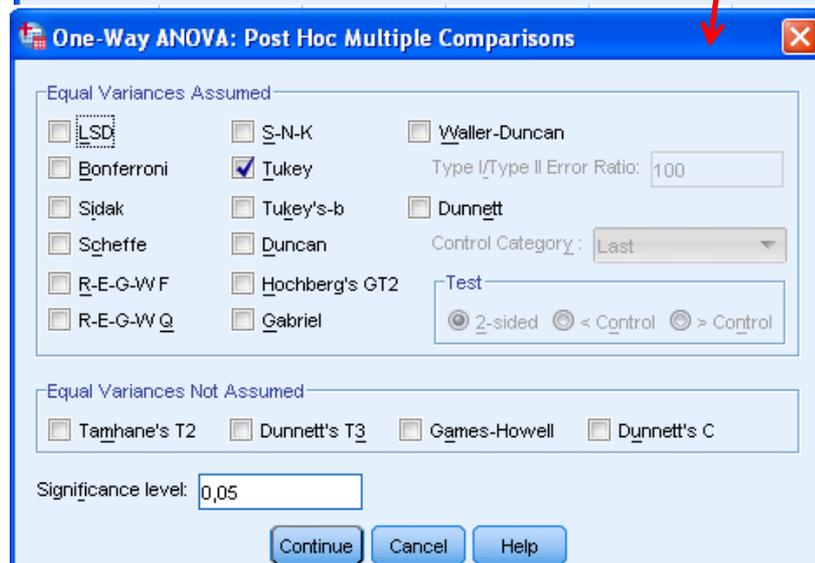
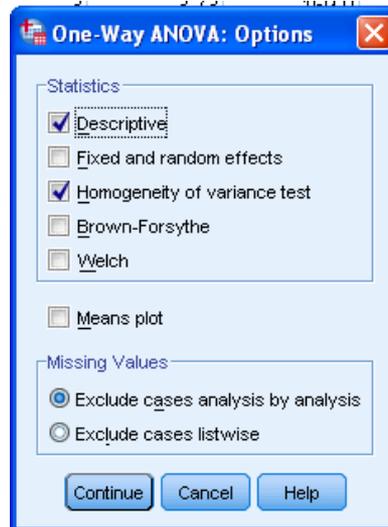
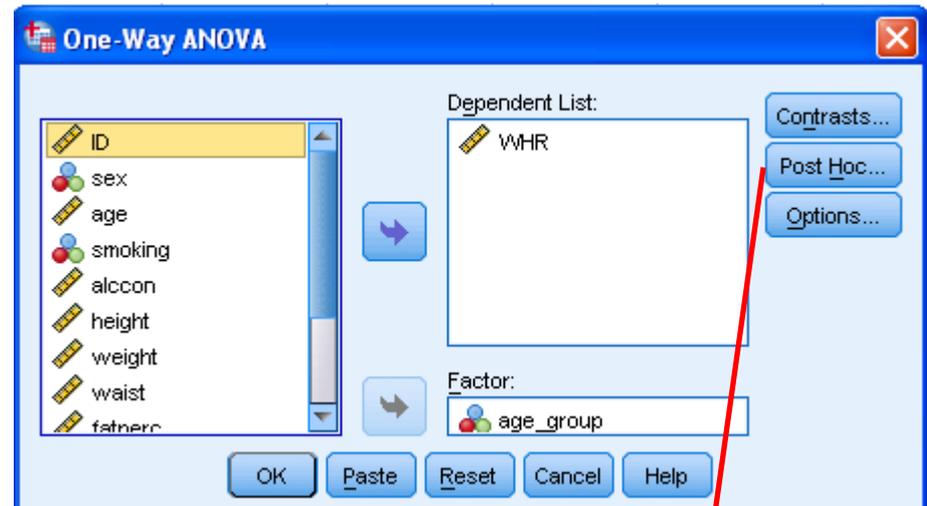
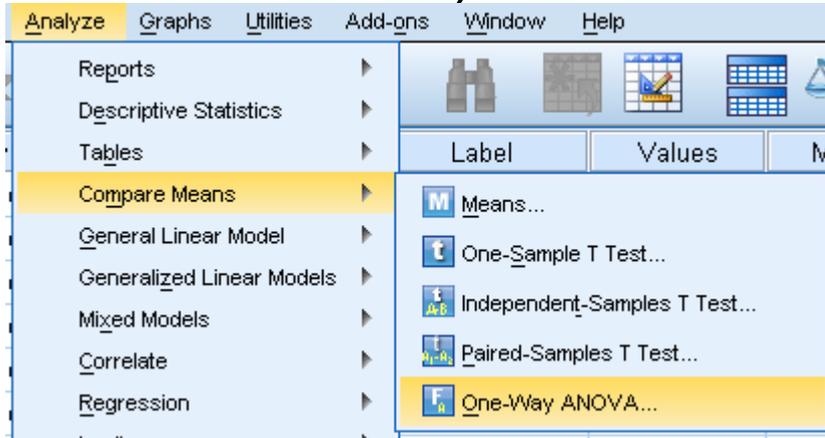
- Control the comparison-wise error rate: Reject  $H_0$ , if  $p < \alpha$
- Control the family-wise error rate (including  $k$  tests): Reject  $H_0$ , if  $p < \alpha/k$ 
  - **Advantage: simple**
- Problem: Bonferroni-correction increases the probability of a type II error
  - the power of detecting a true association is reduced → **Disadvantage: too conservative**

k	$\alpha/k$ ( $\alpha=0.05$ )
1	0.05
5	0.01
10	0.005
100	0.0005

→  **$0.05/5=0.01$**

# The multiple testing problem: ANOVA -Tukey

- How to do such an ANOVA in SPSS (We use dataset “Alldata.sav”):

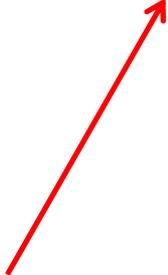


# The multiple testing problem: ANOVA -Tukey

## Post Hoc Tests

### Multiple Comparisons

These p-values are corrected for multiple testing



Dependent Variable: WHR

Tukey HSD

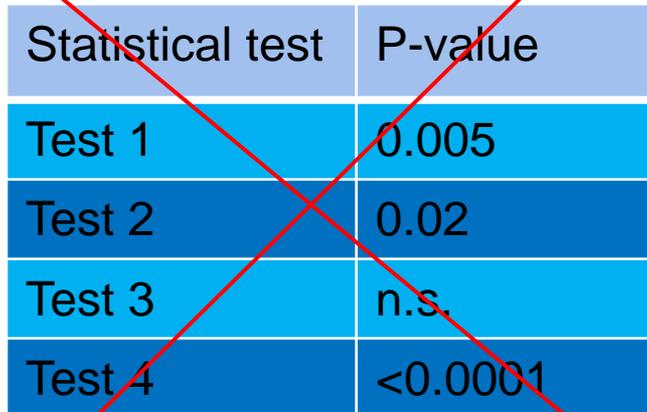
(I) smoking	(J) smoking	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Current smoker	Ex smoker	-,012888	,006540	,120	-,02823	,00245
	Never smoker	,032481*	,006057	,000	,01827	,04669
Ex smoker	Current smoker	,012888	,006540	,120	-,00245	,02823
	Never smoker	,045370*	,005016	,000	,03360	,05714
Never smoker	Current smoker	-,032481*	,006057	,000	-,04669	-,01827
	Ex smoker	-,045370*	,005016	,000	-,05714	-,03360

\*. The mean difference is significant at the 0.05 level.

# The multiple testing problem

How to report p-values / results of significance tests in papers:

- If you are only interested in test decisions (significant or not) to a pre-specified  $\alpha$ -level → report only the decision
- If you are interested in the „certainty“ of your test decision → report **all** p-values (can be interpreted as strength of evidence against the Nullhypothesis)
- In the case of multiple testing: report all raw p-values + a reasonable correction



Statistical test	P-value
Test 1	0.005
Test 2	0.02
Test 3	n.s.
Test 4	<0.0001

## How it should be (1):

Statistical test	P-value
Test 1	0.005*
Test 2	0.02
Test 3	0.2
Test 4	0.00008*

\*still significant even after Bonferroni correction for multiple testing

# The multiple testing problem

How to report p-values / results of significance tests in papers:

- If you are only interested in test decisions (significant or not) to a pre-specified  $\alpha$ -level → report only the decision
- If you are interested in the „certainty“ of your test decision → report **all** p-values (can be interpreted as strength of evidence against the Nullhypothesis)
- In the case of multiple testing: report all raw p-values + a reasonable correction

## How it should be (1):

Statistical test	P-value
Test 1	0.005*
Test 2	0.02
Test 3	0.2
Test 4	0.00008*

\*still significant even after Bonferroni correction for multiple testing

## How it should be (2):

Statistical test	P-value	P-value corrected
Test 1	0.005	0.02
Test 2	0.02	0.08
Test 3	0.2	0.8
Test 4	0.00008	0.00032



# **Sample size estimation**

# Sample size estimation

**Question:** How many individuals do you have to include in your study to get a reliable result ?

→ We want to **maximize the probability** for rejecting  $H_0$ , if  $H_1$  is true

		Decide for	
		$H_0$	$H_1$
Reality	$H_0$	Correct	Wrong: Type I error ( $\alpha$ )
	$H_1$	Wrong: Type II error ( $\beta$ )	Correct: <b>Power</b>

→ while keeping the **Type I error  $\alpha$**  fixed

**What do you have to know to calculate the sample size needed?**

1. Power (typically set to 80% or 90%)
2. Type I error  $\alpha$  (typically set to  $\alpha = 0.05$ )
3. The difference you want to find (for t-tests: the mean difference between groups)
4. standard deviation / measure of variance

## Example

- Hypothesis:  $H_0: \mu_A = \mu_B$  versus  $H_1: \mu_A \neq \mu_B \rightarrow$  two-sided t-test
  - You consider a difference of 10 as relevant
  - From former studies, you know, that the standard deviation is  $\sim 15$  mmHG
  - So far, you have recruited 20 patients (10 in each treatment arm)
- $\rightarrow$  What is your power?

## Fallzahlschätzung für unverbundene Stichproben und stetige Zielgrößen

- Fallzahlberechnung für vorgegebene Power
- Powerberechnung für vorgegebene Fallzahl
- Entdeckbare Differenz für vorgegebene Fallzahl und Power

Eingabe von  $\mu_1$ :  Eingabe von  $\mu_2$ :

Eingabe von  $\sigma$ :  Differenz Delta:

- Einseitiger Test
- Zweiseitiger Test

Eingabe von  $\alpha$  (Standard ist 0.05):

Eingabe der Power (Standard ist 0.80):

Die Fallzahl für jede Gruppe ist:

# Sample size estimation

How many patients do you need to reach a power of 80%?

## Fallzahlschätzung für unverbundene Stichproben und stetige Zielgrößen

Ende

Neustart

Hilfe!

- Fallzahlberechnung für vorgegebene Power
- Powerberechnung für vorgegebene Fallzahl
- Entdeckbare Differenz für vorgegebene Fallzahl und Power

Eingabe von  $\mu_1$ :  Eingabe von  $\mu_2$ :

Eingabe von  $\sigma$ :  Differenz Delta:

- Einseitiger Test
- Zweiseitiger Test

Eingabe von  $\alpha$  (Standard ist 0.05):

Eingabe der Power (Standard ist 0.80):

Die Fallzahl für jede Gruppe ist:

Berechne

# Sample size estimation

How to increase the power:

1. Increase the sample size  $n$
2. Increase the difference you want to show

## Fallzahlschätzung für unverbundene Stichproben und stetige Zielgrößen

Ende Neustart Hilfe!

- Fallzahlberechnung für vorgegebene Power
- Powerberechnung für vorgegebene Fallzahl
- Entdeckbare Differenz für vorgegebene Fallzahl und Power

Eingabe von  $\mu_1$ :  Eingabe von  $\mu_2$ :

Eingabe von  $\sigma$ :  Differenz Delta:

- Einseitiger Test
- Zweiseitiger Test

Eingabe von  $\alpha$  (Standard ist 0.05):

Eingabe der Power (Standard ist 0.80):

Die Fallzahl für jede Gruppe ist:

Berechne

# Sample size estimation

## Comparing proportions:

**Example:** Test differences in the proportions of Myocardial Infarctions between treatment A and B; Hypothesis:  $H_0: \pi_A = \pi_B$  versus  $H_1: \pi_A \neq \pi_B \rightarrow \chi^2$ -test

### Fallzahlschätzung für den Vergleich von Häufigkeiten zweier unverbundener Stichproben

Ende

Neustart

Hilfe!

- Fallzahlberechnung für vorgegebene Power
- Powerberechnung für vorgegebene Fallzahl
- Berechnung von  $p_2$  für vorgegebene Fallzahl und Power

Eingabe von  $p_1$ :

Eingabe von  $p_2$ :

- einseitiger Test
- zweiseitiger Test

Eingabe von  $\alpha$  (Standard ist 0.05):

Eingabe der Power (Standard ist 0.80):

Die Fallzahl für jede Gruppe ist:

Berechne

# Sample size estimation

## Some remarks for sample size estimation / power calculation:

- Sample size estimation is not exact, it is not more than an educated guess !  
Why? You have to provide the difference you want to test & the standard deviation → Based on experience, former studies, gut feeling etc...

## ■ Post-hoc power:

Power analysis is useless, if the analysis has already been performed!

Power is a probability. Retrospectively, the outcome of the test is known → the retrospective power is 1, if the test was significant, and 0 otherwise.

