Statistics for Diploma Students "Basics"
-- Statistical inference and statistical tests--

Claudia Lamina

Medical University Innsbruck, Division of Genetic Epidemiology

GENEPI INNSBRUCK

The principles of statistical testing:

Formulating Hypothesis & Teststatistics & p-values

The most common statistical tests:

Testing measures of location

&

Testing frequencies

Formulating Hypothesis & Statistical Tests

Steps in conducting a statistical test:

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

Formulating Hypothesis & Statistical Tests

<u>Hypothesis Formulation:</u>

- Nullhypothesis **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: $\pi_{new} \geq \pi_{old}$ H_1: $\pi_{new} < \pi_{old}$ with π_{new} : the proportion of patients experiencing a MI during the study receiving the new therapy π_{old} : the proportion of patients experiencing a MI during the study receiving the old therapy	**Statistical hypothesis:** H_0: $\mu_{men}=\mu_{women}$ H_1: $\mu_{men}\neq\mu_{women}$ with μ_{men} : mean scores for men μ_{women} : mean scores for women
One-sided test	Two-sided test

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 (α)
	H_1	Wrong decision: Type II error (β)	Correct decision: Power ($1-\beta$)

- Type I and Type II error cannot be minimized simultaneously
- Statistical tests are constructed in that way, that the probability of a Type I error is not bigger than the significance level α (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 the100 tests.

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

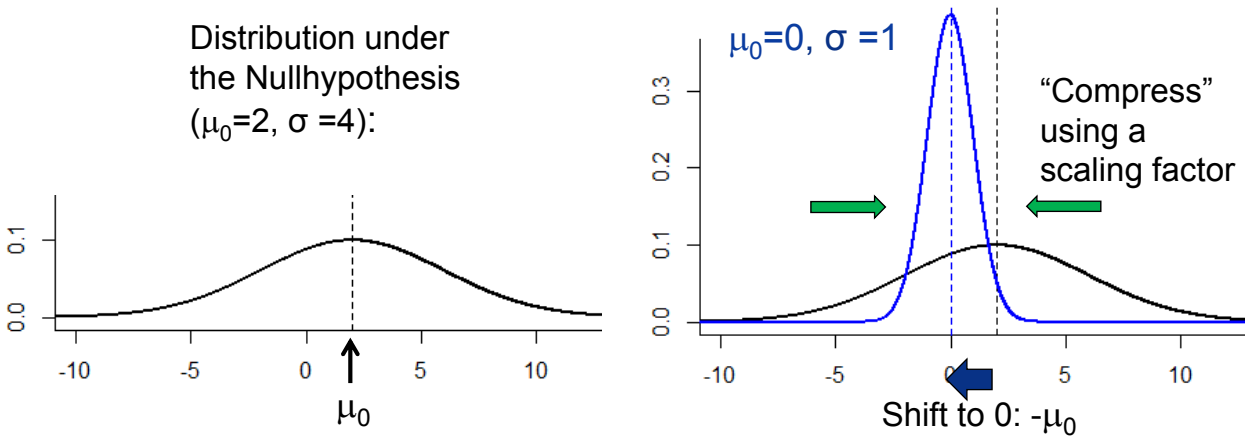
The One-sample t-test (the "standard test" for mean comparisons):

- Situation: Compare the sample mean (μ_{sample}) with a specified mean (μ_0)
- Hypothesis: H_0: $\mu_{sample} = \mu_0$ versus H_1: $\mu_{sample} \neq \mu_0$
- Assumption under the Nullhypothesis: normal distribution (mean μ_0 and known σ)

Testing measures of location

The One-sample t-test (the "standard test" for mean comparisons):

- Situation: Compare the sample mean (μ_{sample}) with a specified mean (μ_0)
- Hypothesis: H_0: $\mu_{sample} = \mu_0$ versus H_1: $\mu_{sample} \neq \mu_0$
- Assumption under the Nullhypothesis: normal distribution (mean μ_0 and known σ)

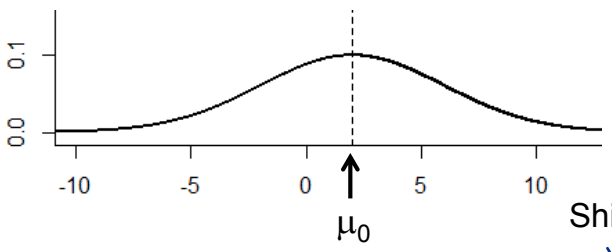


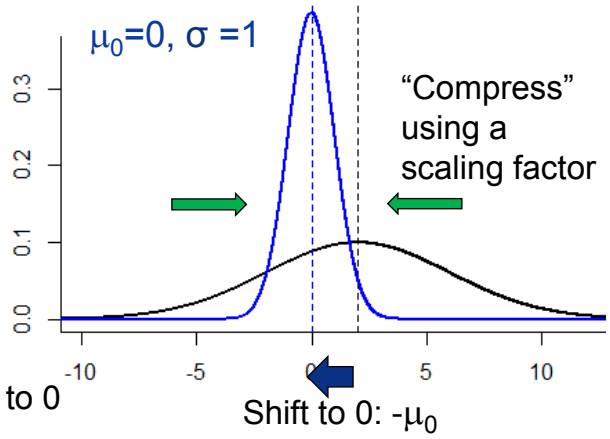
Testing measures of location

The One-sample t-test (the "standard test" for mean comparisons):

- Situation: Compare the sample mean (μ_{sample}) with a specified mean (μ_0)
- Hypothesis: H_0: $\mu_{sample} = \mu_0$ versus H_1: $\mu_{sample} \neq \mu_0$
- Assumption under the Nullhypothesis: normal distribution (mean μ_0 and variance σ)

Distribution under the Nullhypothesis (μ_0=2, σ =4):





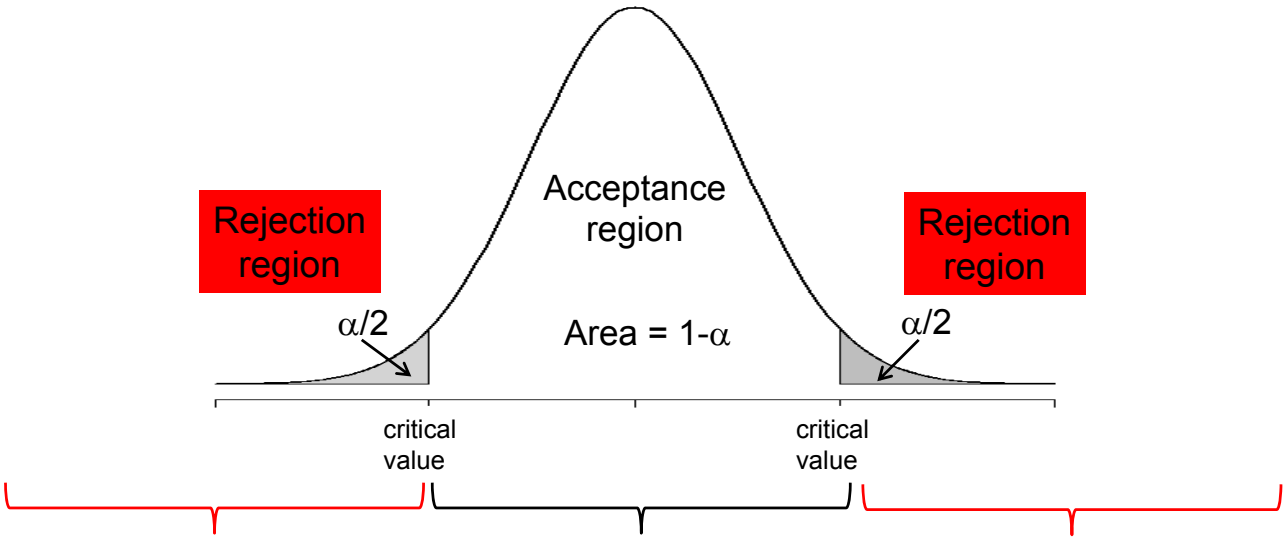
Shift to 0

Teststatistic: $T = \frac{\bar{X} - \mu_0}{S}\sqrt{n}$

Scaling factor

Testing measures of location

- If a T-Statistic is very extreme (lower or higher than the critical value) → it is very likely that it <u>does not</u> belong to the distribution under the nullhypothesis
  → reject H_0



T within rejection region → Reject H_0

T within acceptance region → Accept H_0

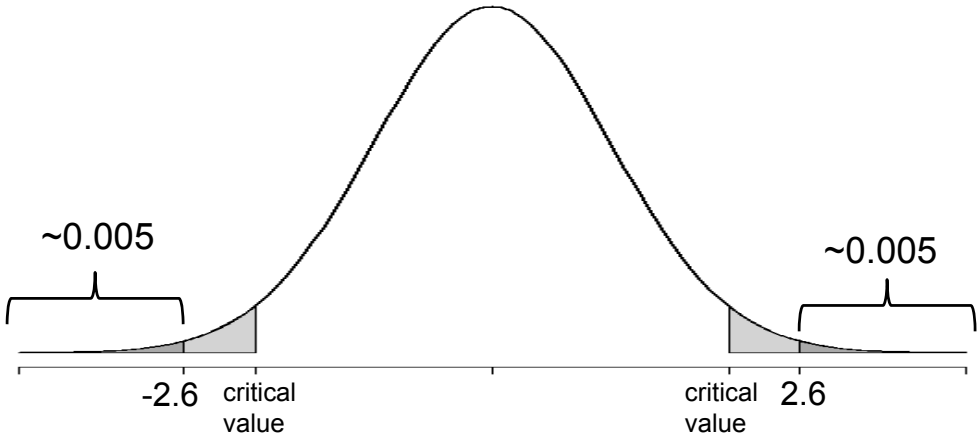
T within rejection region → Reject H_0

Testing measures of location

Example:

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

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



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

P-value (two-sided test)= 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 nullhypothesis.

Example:

A one sample t-Test comparing the sample mean to 0: H_0: $\mu_{sample} = 0$; H_1: $\mu_{sample} \neq 0$ 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 different from 0 is 1%“

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

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 (just by chance), which is ≥ |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

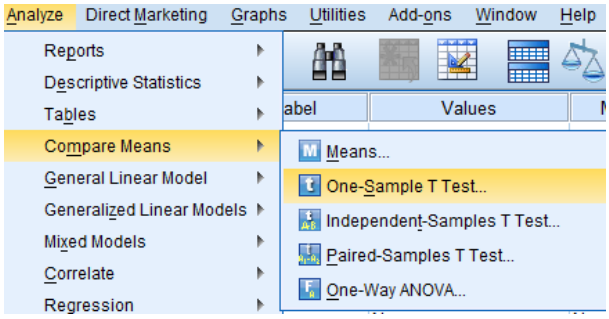
If $p < \alpha$, reject H_0

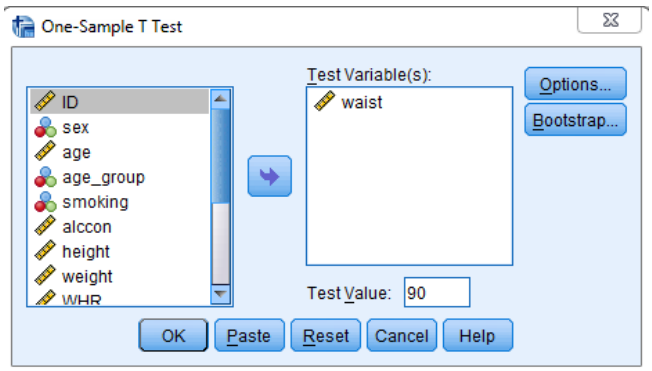
→ In most cases, a **p-value < 0.05** (or <5%) is said to be **statistically significant** !

→ You can also base your decision on the Confidence Interval!

Testing measures of location

The One-sample t-test in SPSS (We use dataset "Alldata_Tag2.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 = 90					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
waist	2,979	1452	,003	,945	,32	1,57

T-Statistics

Testing measures of location

The two-sample t-test for unpaired samples:

- Situation: Compare the means (μ_1, μ_2) of two unpaired samples
- Assumption: normal distribution of both samples, σ ($= \sigma_1 = \sigma_2$) is not known

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

Hypothesis:

- Teststatistic: $$T = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

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

- If T "too extreme → reject H_0
- If $p < \alpha$ → 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	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)
$\overline{X}$	11.024	9.86
S^2	15.227	27.038

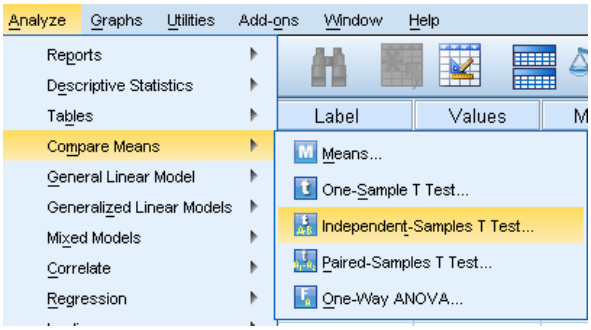
T = 0.556

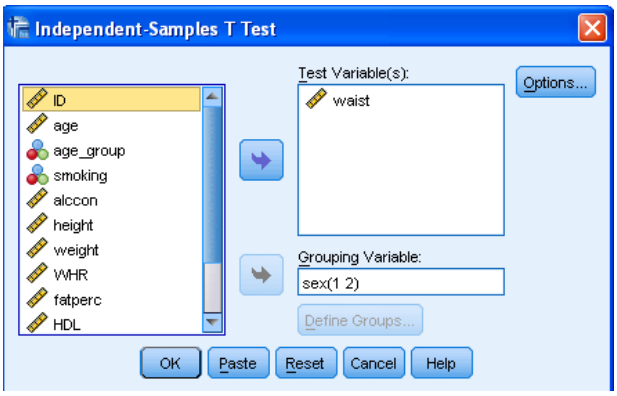
P-value = 0.29

→ XY does not differ between diseased and non-diseased

Testing measures of location

- How to do unpaired T-Test in SPSS:





Test bei unabhängigen Stichproben

		Levene-Test der Varianzgleichheit		T-Test für die Mittelwertgleichheit						
									95% Konfidenzintervall der Differenz	
		F	Signifikanz	T	df	Sig. (2-seitig)	Mittlere Differenz	Standardfehler der Differenz	Untere	Obere
waist	Varianzen sind gleich	23.266	.000	-23.787	1451	.000	-12.8064	.5384	-13.8625	-11.7503
	Varianzen sind nicht gleich			-23.811	1420.653	.000	-12.8064	.5378	-13.8615	-11.7514

P-value < 0.05

→ Waist does differ significantly between men and women

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, σ ($= \sigma_1 = \sigma_2$) is not known
- Hypothesis: H_0: $\mu_{before} = \mu_{after}$ versus H_1: $\mu_{before} \neq \mu_{after}$

  Calculate $d = x_{before} - x_{after}$ 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 ≠ 0: $\mu_d \neq 0$

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 = 2.368

p = 0.042

→ H_0 can be rejected

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

→one-sided test more appropriate (more power)

→p = 0.021

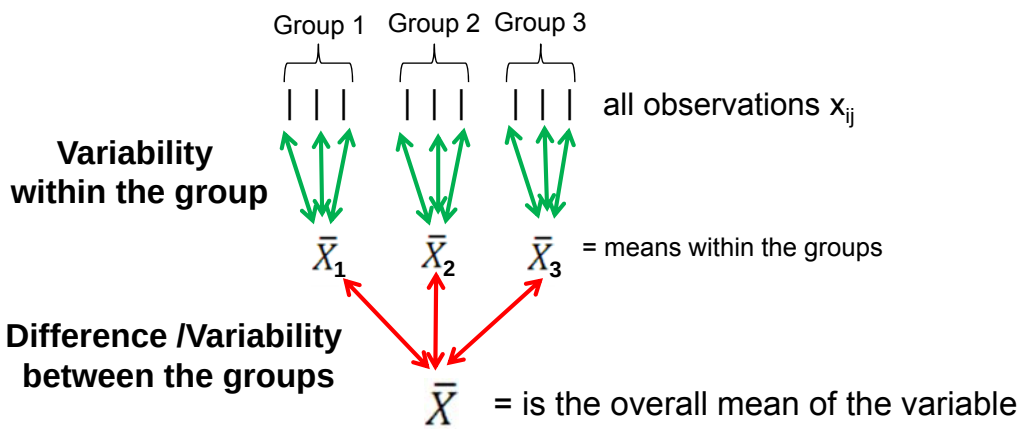
If you would have done a „normal“ unpaired t-test:

p = 0.484 →H_0 can not be rejected !

Testing measures of location

Analysis of Variance (ANOVA)

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



Testing measures of location

- **Test statistic:** $F = \frac{S^2_{between}}{S^2_{within}}$
- **Test decision** for a two sided test: If F "too extreme": 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:

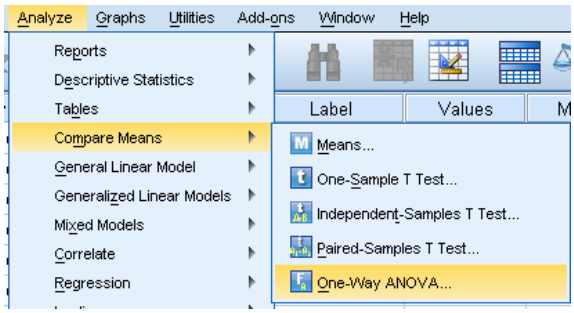
There are 3 different medications (Med1, Med2, Med3), which are intended to increase the HDL-cholesterol levels in patients

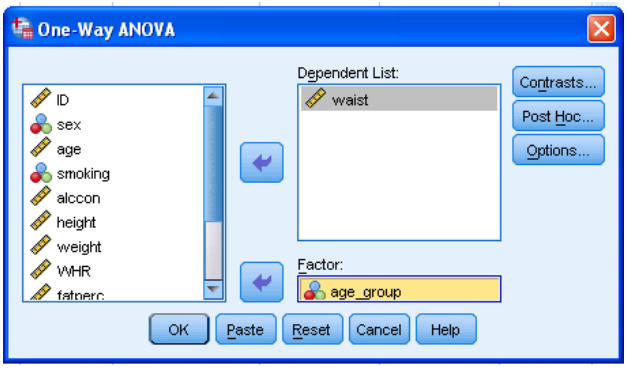
1. perform ANOVA as an overall test, if there is a difference between the groups
2. If the F-Test was significant, you know, that there is a difference
3. Test Med1 against Med2, Med1 against Med3, Med2 against Med3

→ 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:





ONEWAY ANOVA

waist

	Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
Zwischen den Gruppen	13726.175	3	4575.392	33.380	.000
Innerhalb der Gruppen	198611.920	1449	137.068		
Gesamt	212338.095	1452			

P-value

We only know, that there is a difference between the groups, but not between which groups

→ post-hoc tests

Testing measures of location

- ANOVA and Post-hoc tests:

Tukey: all pairwise comparisons

Dunnett: All groups are compared to one reference group ("Gold standard")

Dependent Variable: waist

Tukey HSD

(I) age_group	(J) age_group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
30-40	41-50	-5,511*	1,149	,000	-8,47	-2,56
	51-60	-8,057*	1,087	,000	-10,85	-5,26
	61-70	-10,931*	1,163	,000	-13,92	-7,94
41-50	30-40	5,511*	1,149	,000	2,56	8,47
	51-60	-2,546*	,774	,006	-4,54	-,56
	61-70	-5,420*	,876	,000	-7,67	-3,17
51-60	30-40	8,057*	1,087	,000	5,26	10,85
	41-50	2,546*	,774	,006	,56	4,54
	61-70	-2,873*	,794	,002	-4,92	-,83
61-70	30-40	10,931*	1,163	,000	7,94	13,92
	41-50	5,420*	,876	,000	3,17	7,67
	51-60	2,873*	,794	,002	,83	4,92

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

→ All groups differ from each other!

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 not sure about normal distribution:

Kolmogorov-Smirnov test to test normality assumption

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	• Wilcoxon-Test • Wilcoxon rank-sum test • Mann-Whitney U-Test
ANOVA	Kruskal-Wallis-Test

Testing measures of location

Two sample test on equality of distributions: Wilcoxon / Mann-Whitney U-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 → 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

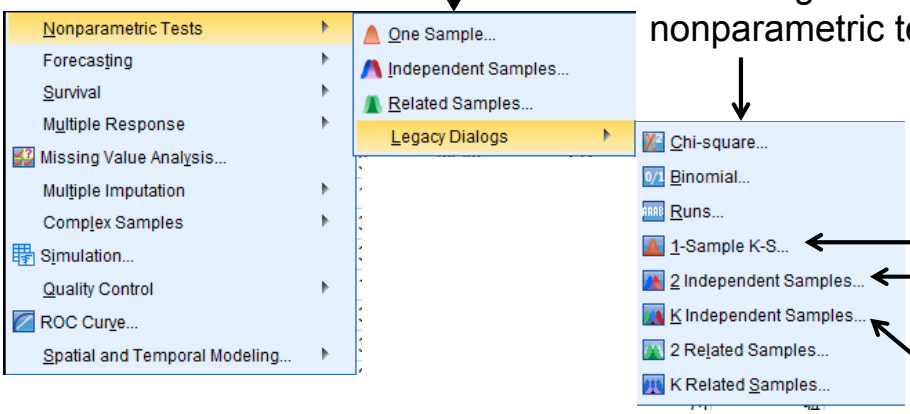
<u>Example for building ranks:</u>

Observations of Var x	Rank rank(x)
11	1
15	2
17	3.5
17	3.5
22	4

Testing measures of location

New dialogboxes for nonparametric tests

Old dialogboxes for nonparametric tests



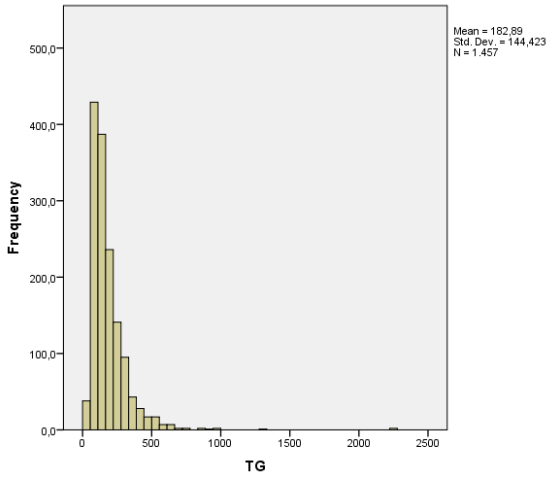
Kolmogorov-Smirnov test to test normality assumption

Mann-Whitney-U-Test

Kruskal-Wallis-Test

Testing measures of location

First step: Checking the normality assumption with Kolmogorov-Smirnov Test and histograms



One-Sample Kolmogorov-Smirnov Test

		TG
N		1457
Normal Parameters[a,b]	Mean	182,89
	Std. Deviation	144,423
Most Extreme Differences	Absolute	,165
	Positive	,152
	Negative	-,165
Test Statistic		,165
Asymp. Sig. (2-tailed)		,000[c]

a. Test distribution is Normal.
b. Calculated from data.
c. Lilliefors Significance Correction.

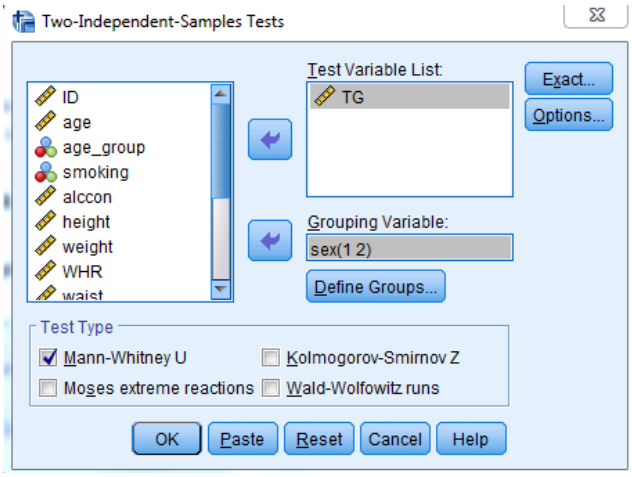
Here: Test significant

→ normality assumption <u>is not</u> fulfilled

→ Perform nonparametric tests

Testing measures of location

- Mann-Whitney-U-Test



Test Statistics[a]

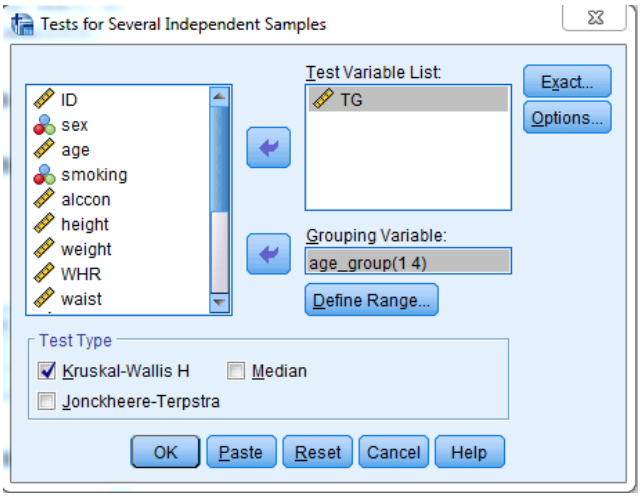
	TG
Mann-Whitney U	168556,000
Wilcoxon W	439036,000
Z	-12,053
Asymp. Sig. (2-tailed)	,000

a. Grouping Variable: sex

The distribution of TG differs significantly between men and women

Testing measures of location

- Kruskal-Wallis-Test



Test Statistics[a,b]

	TG
Chi-Square	65,077
df	3
Asymp. Sig.	,000

a. Kruskal Wallis Test

b. Grouping Variable: age_group

The distribution of TG differs significantly between agegroups

Testing frequencies

- Situation: Compare the frequencies between two groups

  Or: Test, if two categorical variables X (i=1,…k) and

  Y (j=1,…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: **χ^2-test of independence**:

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

  H_1: X and Y depend on each other

- Assumption:

→ none of the cells should have a very rare expectancy

(number of expected counts in each cell ≥ 1 and for at least 80% of the cells: ≥ 5)

→ 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 (under the assumption that the two factors are independent)

Testing frequencies

Table of observed numbers

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

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

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

X= Gender

Y= Smoking

Testing frequencies

X= Gender

Y= Smoking

Table of expected numbers:

		Y			
		1	…	m	Σ
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	$h_{k.}h_{.1}/n$	…	$h_{k.}h_{.m}/n$	$h_{k.}$
Σ		$h_{.1}$		$h_{.m}$	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:

(Row sum * Columns sum) / Total sum

Expected number in the upper left cell:

722*261/1457 = 129.336

Testing frequencies

Observed:

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

O_{ij}

Expected:

		Y			
		1	…	m	Σ
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	$h_{k.}h_{.1}/n$	…	$h_{k.}h_{.m}/n$	$h_{k.}$
Σ		$h_{.1}$		$h_{.m}$	n

E_{ij}

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

Test decision: If χ^2 „too extreme“ (or $p < \alpha$) → Reject H_0

Testing frequencies

Example:

Observed:

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

Expected:

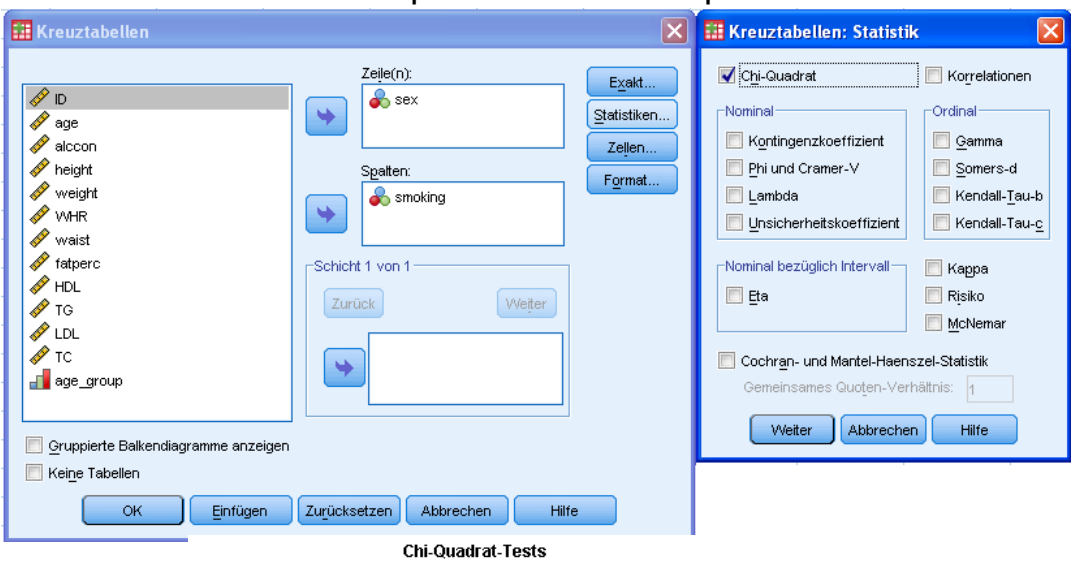
Smoking status / Gender	Current Smoker	Ex-Smoker	Never Smoker	Row Total
Men	129.336	224.479	368.185	722
Women	131.664	228.521	374.815	735
Column Total	261	453	743	1457

$\chi^2 = 121.9218$ >> critical value → test is significant (p = 3.3e-27)

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

Testing frequencies

- How to calculate the Chi-squared-test of independence in SPSS:



Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (2-seitig)
Chi-Quadrat nach Pearson	121.922[a]	2	.000
Likelihood-Quotient	124.164	2	.000
Zusammenhang linear-mit-linear	62.436	1	.000
Anzahl der gültigen Fälle	1457		

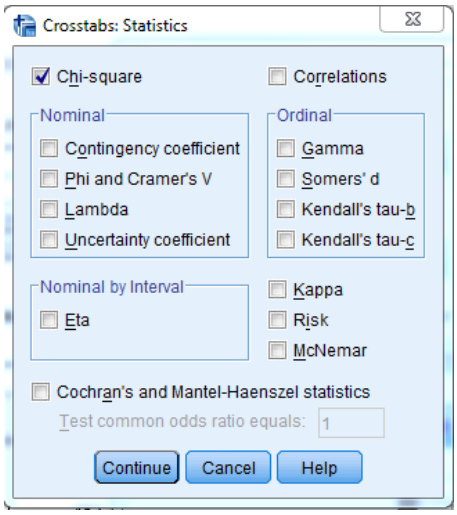
a. 0 Zellen (.0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 129.34.

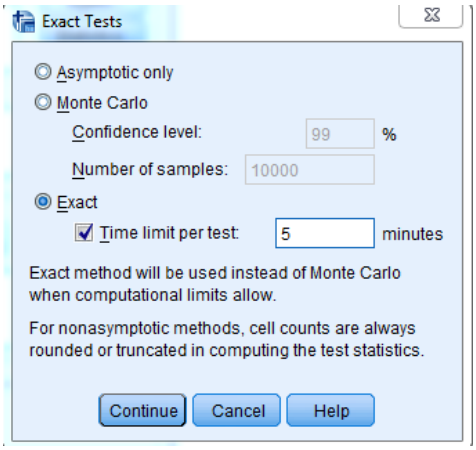
P-value

Assumption for Chi-Square test is fulfilled. If it is not fulfilled: 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[a]	2	,000	,000		
Likelihood Ratio	124,164	2	,000	,000		
Fisher's Exact Test	123,930			,000		
Linear-by-Linear Association	62,436[b]	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.

Risk estimates

Odds Ratio

- Often, you want to assess **risk**:

That is, risk to get a disease, if a risk factor is present

relative to persons not having this risk factor

(e.g. smoking, obesity etc…)

Risk factor	Disease Status	
	yes	no
present	a	b
not present	c	d

- **Risk estimates:**
  - **Relative Risk:** $RR = \frac{a/(a+b)}{c/(c+d)}$

    But: can only be estimated in prospective studies!

  - **Odds Ratio:** $OR = \frac{a/b}{c/d} = \frac{a \cdot d}{c \cdot b}$

    Approximates the RR and can be estimated in any kind of studies

  - **Hazard Ratio HR**: Can be estimated in "survival studies", where the time to event (e.g. death or a non-fatal event as MI/stroke) is known

Odds Ratio

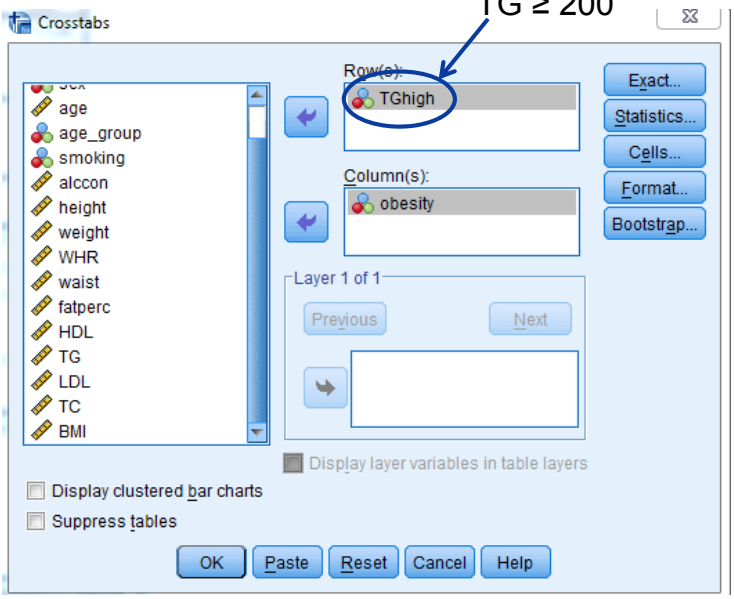
Interpretation of Odds Ratio:

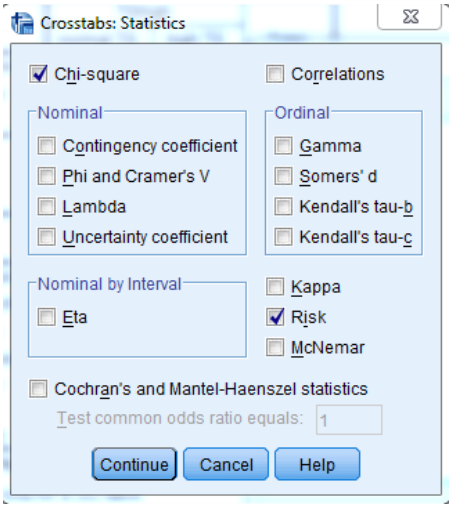
OR = 1	The risk factor is not associated with the disease
OR > 1	Positive association of risk factor with the disease (Persons with the risk factor have a higher probability for the disease as persons without)
OR < 1	Negative association of "risk" factor with the disease (Persons with the "risk" factor have a lower probability for the disease as persons without) → factor is protective

Odds Ratio

Calculation of Odds Ratio in SPSS using Crosstabs:

TG < 200 vs. TG ≥ 200





Risk Estimate

CI for OR

OR

	Value	95% Confidence Interval Lower	Upper
Odds Ratio for TGhigh (normal TG / high TG)	2,109	1,634	2,723
For cohort obesity = not obese	1,203	1,121	1,290
For cohort obesity = obese	,570	,473	,688
N of Valid Cases	1455		

→ Obese persons have a higher risk having Triglyceride values above 200

The multiple testing problem

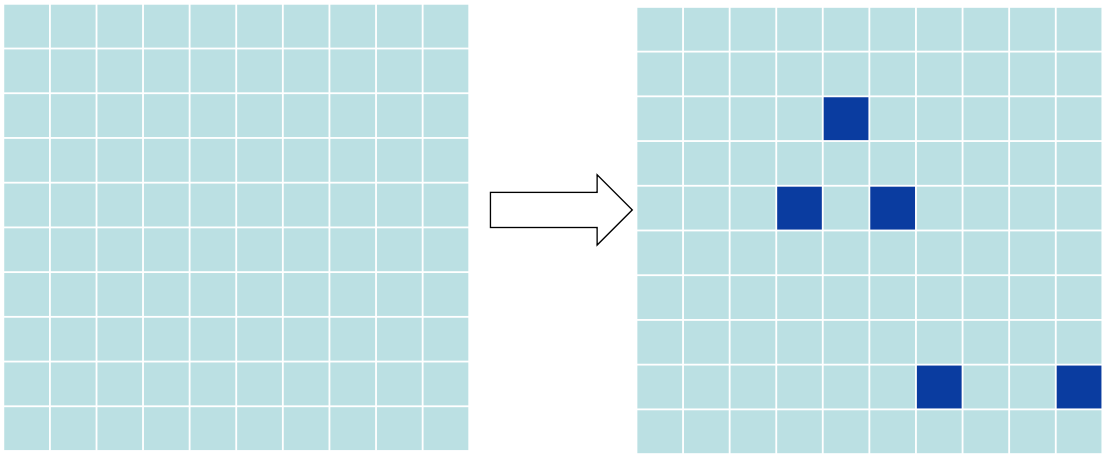
The multiple testing problem

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 α=0.05

→ The tests are constructed in that way, that maximum 5 of 100 tests reject the Nullhypothesis, although it is true



→ 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 α being the **comparison-wise error rate**

k	α^* (α=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	α/k (α=0.05)
1	0.05
5	0.05/5 = 0.01
10	0.05/10 = 0.005
100	0.05/100 = 0.0005

- **Other correction methods:** the post-hoc tests that can be performed after an ANOVA (e.g. Tukey, Dunnett) are already corrected 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 α-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

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 (α)
	H_1	Wrong: Type II error (β)	Correct: Power

→ while keeping the Type I error α 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 α (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

Sample size estimation

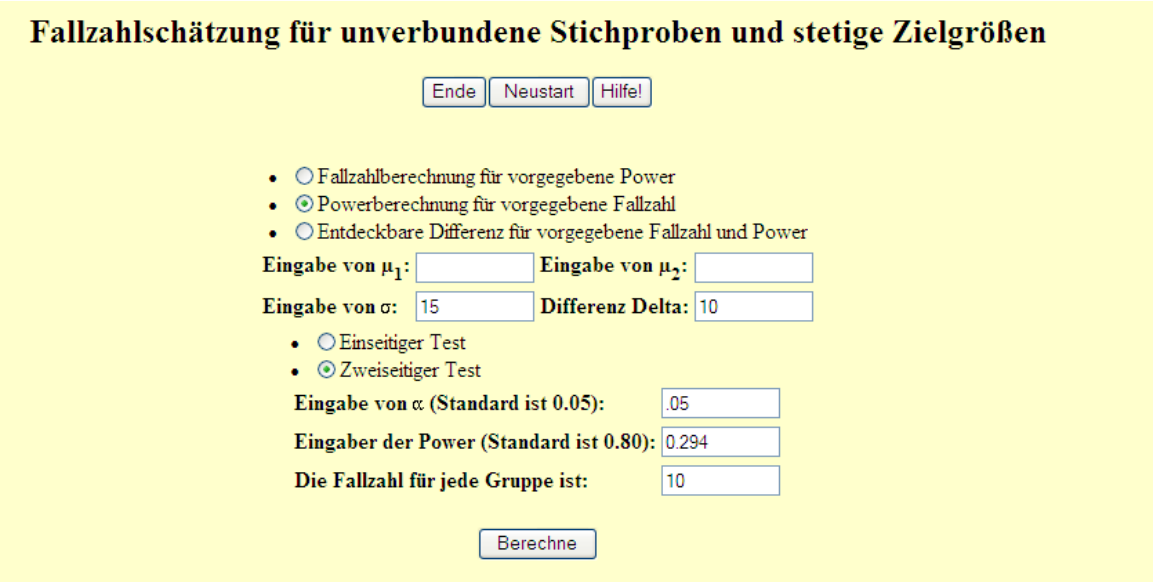
Example

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

→ What is your power?

http://campus.uni-muenster.de/fileadmin/einrichtung/imib/lehre/skripte/biomathe/bio/fallz.html

Sample size estimation

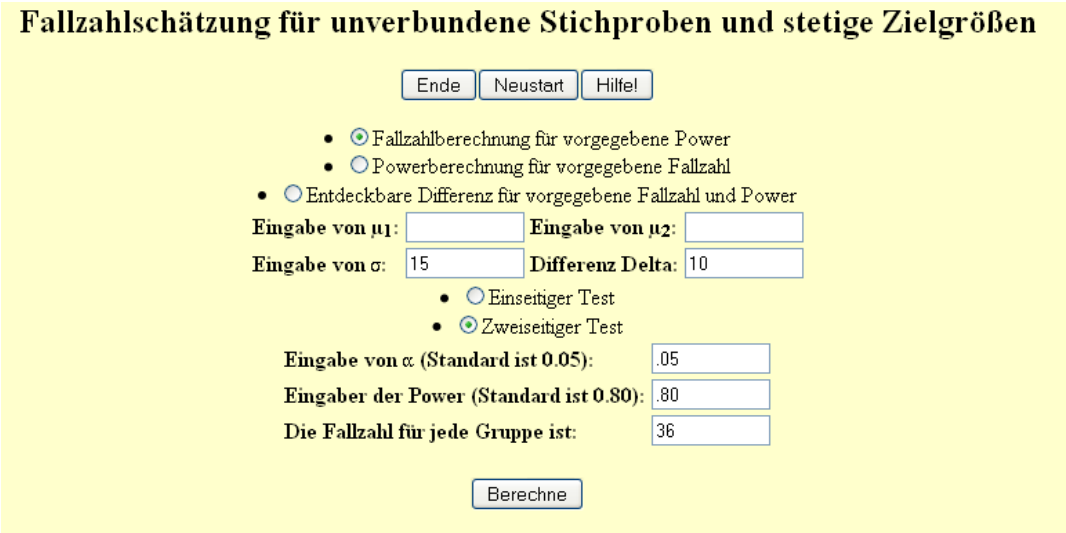


How to increase the power:

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

Sample size estimation

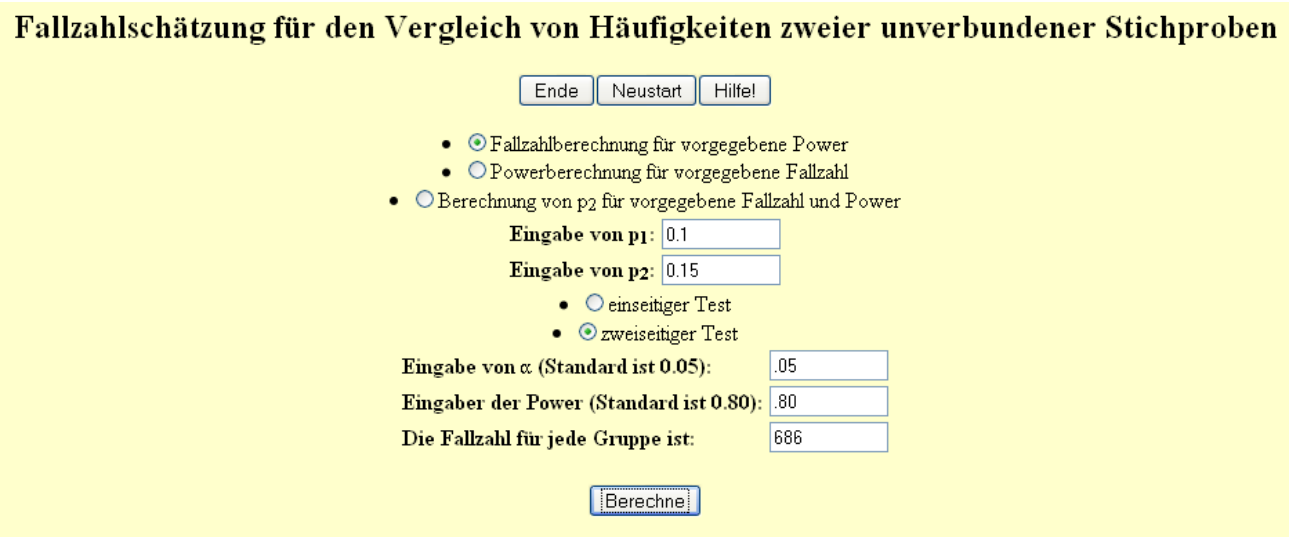
How many patients do you need to reach a power of 80%?



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: $\pi_A \neq \pi_B$ → χ^2-test



http://campus.uni-muenster.de/fileadmin/einrichtung/imib/lehre/skripte/biomathe/bio/fallb.html

Sample size estimation

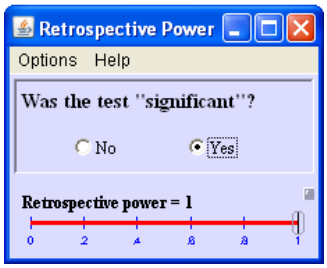
Some remarks for sample size estimation / power calculation:

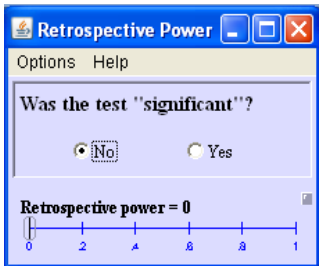
- Sample size estimation is <u>not exact</u>, it is not more than an <u>educated guess</u> ! Why? You have to provide the difference you want to test & the standard deviation → Based on experience, former studies, gut feeling etc…

- **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.





- **Freeware program for more sophisticated statistical models: G*Power 3**

http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/