Statistics for Diploma Students "Basics"
-- Descriptive Statistics--

Claudia Lamina

Medical University Innsbruck, Division for Genetic Epidemiology

GENEPI INNSBRUCK

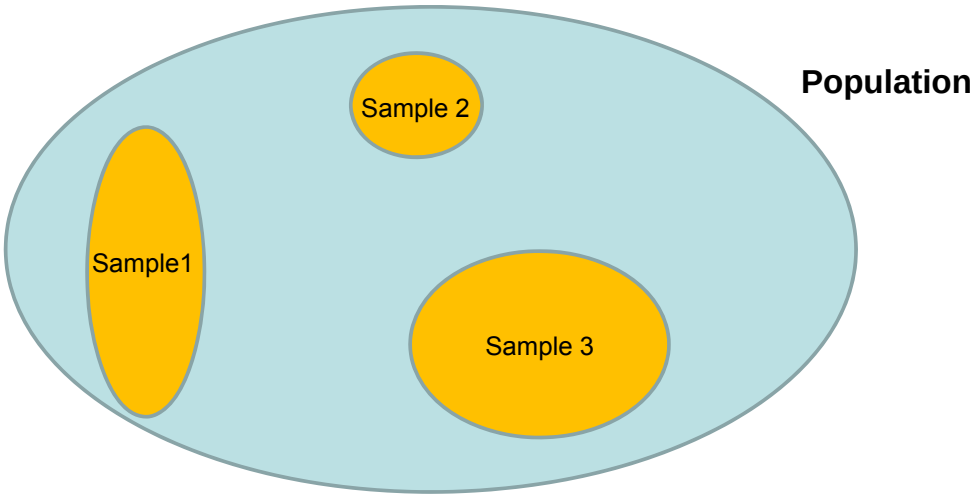
Contents

Aim:

- Get to know basic statistical terms and definitions as they are used specifically in medical science
- Application of appropriate statistical methods for your own data and observations: How can you present your data? Which statistical test should be used ?
- Hands-on experience in a well-known statistics software (SPSS)
- Where are the pitfalls and sources of error?

Introduction

- Intention: Conclude from **sample** on underlying **population**



- Complete population of interest (e.g. all Austrians, all patients with previous myocardial infarctions etc…) cannot be observed → samples drawn from the population
- Samples should be chosen to be representative for the population

Introduction

To draw statistical conclusions, all 3 steps are needed:

1. descriptive, 2. exploratory, 3. inductive

(Often, stages cannot be discriminated, however)

Example:

Population of interest = Patients with previous myocardial infarction (MI)

Aim: Blood pressure reduction via a new medication

Study design: Draw representative sample of all MI patients

Random allocation

Group 1 with new medication

Group 2 with old medication

Introduction

Group 1 with new medication

↓

Reduction of systolic blood pressure by 10 points in group 1 on average

Group 2 with old medication

↓

No reduction of systolic blood pressure in group 2 on average

Descriptive Statistics

↓

Exploratory Statistics:

Difference in mean between group 1 and group 2 in the study sample

↓

Generate Hypothesis in exploratory (study) sample or from the literature: The new medication is leading to a reduction of systolic blood pressure in patients with previous myocardial infarctions

↓

Inductive Statistics in validation sample:

Within this study it could be shown that the new medication is leading to a reduction of systolic blood pressure in patients with previous myocardial infarction

Data types and structuring of data

Data types and structuring of data

Different data types:

- **Qualitative:**
  - **Categorical / nominal :** e.g. binary traits (two possibilities, e.g. gender: 1=male, 2=female) or categorizations, which can not be ordered
  - **Ordinal**: can be ordered (e.g. educational status) or assigned ranks

- **Quantitative:**
  - **Count data** (e.g. number of deaths in a hospital)
  - **Continuous:** e.g. age, weight, glucose levels etc.
    - Continuous data can also be dichotomized: e.g. For defining hypertension, blood pressure is divided into high blood pressure (140/90 mmHg or above) or normal blood pressure (below 140/90).
    - Continuous data can also be divided into more than one category (e.g. Follow up in years: 1-4, 5-9, etc.)

Data types and structuring of data

Create a dataset for your patient data:

PatID	Gender (1=male, 2=female)	Age	Disease
1	1	52	0
2	1	23	0
3	2	79	1
4	2	64	1
5	1	55	0
6	2	50	0
7	2	32	0
8	1	44	1

- Unique identifier for your patient or observation (here: PatID)
- Datasets should be anonymized anyway → no names !
- The names of your variables ideally should have a meaning (PatID, Gender, Age, Disease and not Var1, Var2, Var3 etc….)
- Labeling of the codings (1=male, 2=female) should be saved in an extra file or should be declared as variable labeling in the statistics program
- All variables must have the same length

Data types and structuring of data

Create a dataset for your patient data:

PatID	Gender (1=male, 2=female)	Age	Disease
1	1	52	0
2	1	23	0
3	2	79	1
4	2	64	.
5	1	55	0
6	2	50	.
7	2	32	0
8	1	44	1

Since diagnosis was not sure, values have been set to missing;

. = missing value code in SPSS

Data view /Variable view in SPSS

- Missing values are allowed, <u>but have to be identifiable as such</u>

  Example: "." for a true missing value

  "-999" for a meaningful missing value (e.g. date of stroke for a person who never had a stroke)

Import Data

and

Data Handling in SPSS

Type in data from scratch

Open SPSS and type in the following data into the „**Datenansicht**“:

	PatID	Gender	Age	Disease
1	1.00	1	52	0
2	2.00	1	23	0
3	3.00	2	79	1
4	4.00	2	64	.
5	5.00	1	55	0
6	6.00	2	50	.
7	7.00	2	32	0
8	8.00	1	44	1

And define your variables correctly in the „**Variablenansicht**“

	Name	Typ	Spaltenf...	Dezimal...	Variablenlabel	Wertelabels	Fehlende W...	Spalten	Ausrichtung	Messniveau
1	PatID	Numerisch	8	2		Keine	Keine	8	Rechts	Skala
2	Gender	Numerisch	8	0		{1, male}...	Keine	8	Rechts	Nominal
3	Age	Numerisch	8	0		Keine	Keine	8	Rechts	Skala
4	Disease	Numerisch	8	0		{0, gesund}...	Keine	8	Rechts	Ordinal

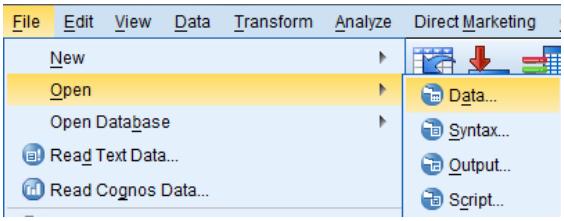
Data View | Variable View

<u>Attention:</u> Check, if numeric data are treated/read in correctly!

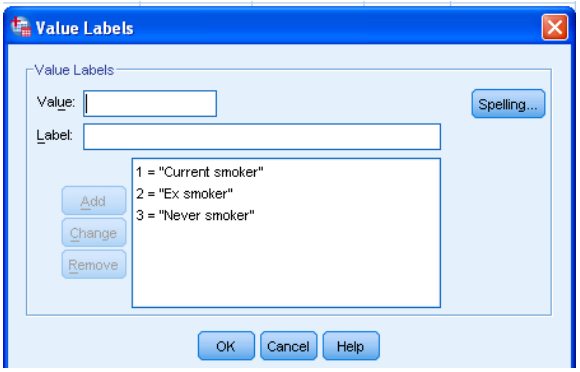
Whether you need , (german) or . (english) as a decimal point depends on the system control settings of the computer you are using !!!

Read in SPSS datafiles

- Read in the dataset Alldata_Tag1.sav:



- Check, if the variables have been defined correctly (Scale, nominal, ordinal)
- Check variable labels:



Quality control of data

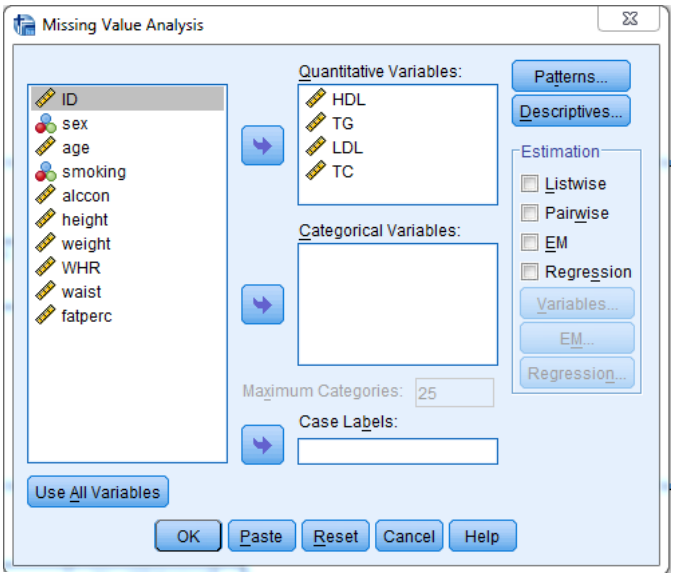
- Check for missing values:

→ Are all data read in correctly ?

→ Can missing values be filled ?

- Check for outliers:

→ Are the outliers real observations or are they possible typos?



Univariate Statistics

	N	Mean	Std. Deviation	Missing		No. of Extremes[a]	
				Count	Percent	Low	High
HDL	1457	53,99	16,505	0	,0	0	30
TG	1457	182,89	144,423	0	,0	0	78
LDL	1454	146,71	40,743	3	,2	3	18
TC	1457	237,10	43,277	0	,0	6	13

a. Number of cases outside the range (Q1 - 1.5*IQR, Q3 + 1.5*IQR).

Descriptive statistics: Summaryzing & Visualizing data

Descriptive Statistics

Descriptive statistics and data summaries are used to

Characterize the study:

Give the reader the possibility to compare studies and interpret the results

Example:

Results might be interpreted differently when generated either in a study of young patients without any serious previous diseases or in a study with older patients with serious previous diseases

Univariate methods (one variable)

Generate hypotheses:

In epidemiologic studies, research hypotheses and questions are often not prespecified or have to be refined

Example:

A new marker for disease progression has been detected in small experimental studies and needs to be proved on a population level

Multivariate methods (> one variable)

→ This is already explorative

Descriptive Statistics

For qualitative data or grouped quantitative data:

Simple tables with numbers and percentages:

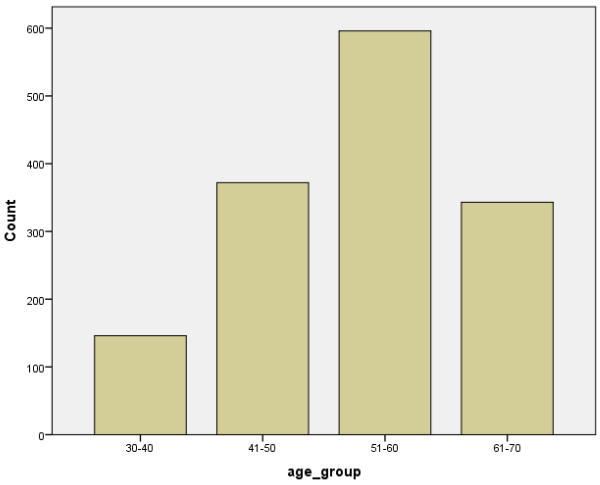
age_group

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	30-40	146	10,0	10,0	10,0
	41-50	372	25,5	25,5	35,6
	51-60	596	40,9	40,9	76,5
	61-70	343	23,5	23,5	100,0
	Total	1457	100,0	100,0	

Barplots for illustrating tables:

Graphical illustration of a categorial variable

But: Categorizations of continuous variables leads to loss of information !

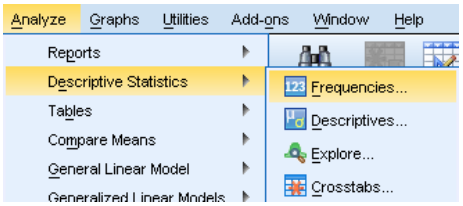


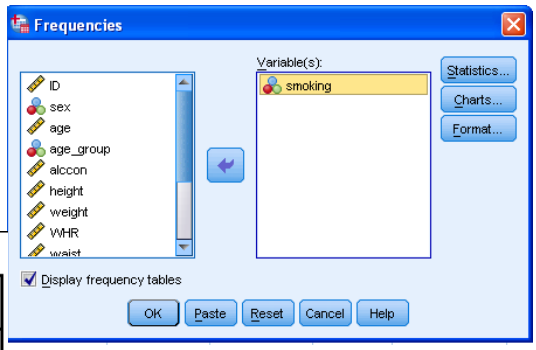
Descriptive Statistics

Example: A population-based study (n=1457) with the aim to identify athersoclerostic risk factors (use dataset Alldata.sav)

Univariate methods: For qualitative data or grouped quantitative data:

Simple tables:





Smoking:

smoking

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Current smoker	261	17,9	17,9	17,9
	Ex smoker	453	31,1	31,1	49,0
	Never smoker	743	51,0	51,0	100,0
	Total	1457	100,0	100,0	

Age-distribution:

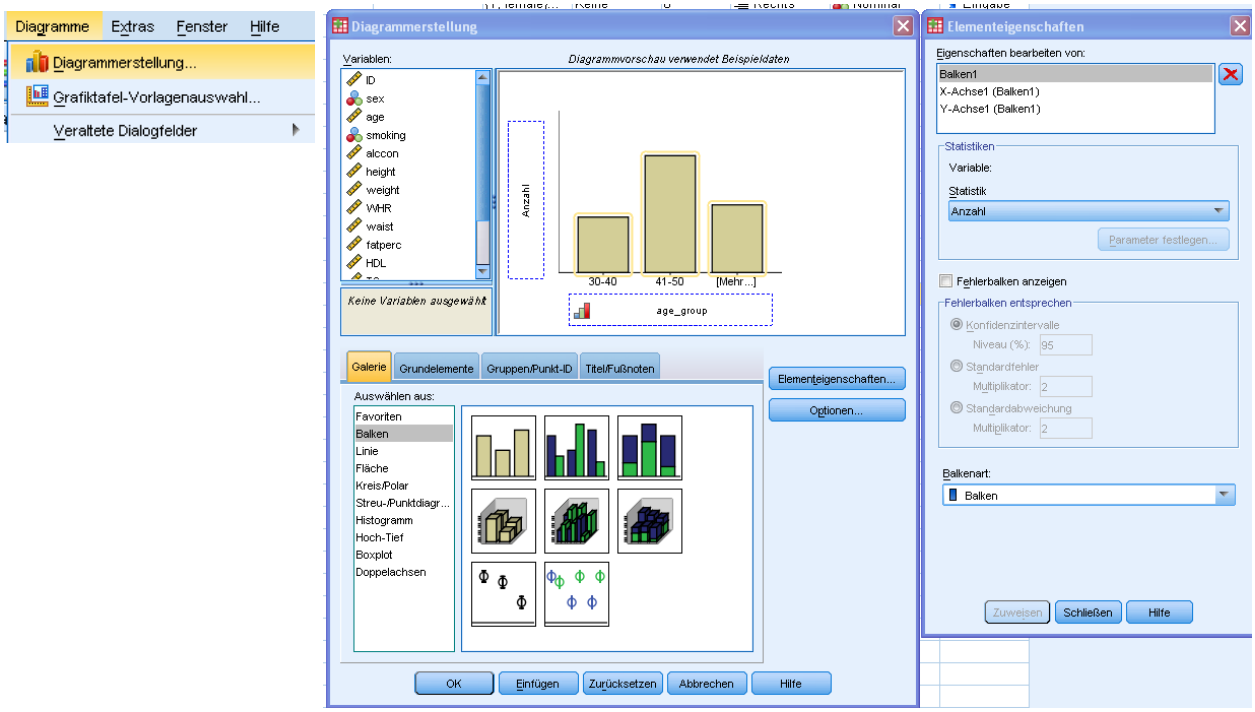
age_group

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	30-40	146	10,0	10,0	10,0
	41-50	372	25,5	25,5	35,6
	51-60	596	40,9	40,9	76,5
	61-70	343	23,5	23,5	100,0
	Total	1457	100,0	100,0	

Descriptive Statistics

Where possible, use figures to illustrate your data !

Barplots for illustrating tables:



Descriptive Statistics

For quantitative data: Measures of location (point estimates)

Mean $\bar{X}$: sum of observations divided by number of observations

Assume, that you have a variable X (e.g. age) (sample size n) with values

$x_1, x_2, \ldots, x_i, \ldots, x_n$, i=1,….n

$$\bar{X} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Example:

Given are the ages of a myocardial infarction patient group:

65, 66, 69, 70, 72, 75, 78

→ Mean= (65 + 66 + 69 + 70 + 72 + 75 + 78)/7 = 70.71

Descriptive Statistics

Attention: The mean is very sensitive to outliers.

Example:

Given are the ages of a myocardial infarction patient group:

65, 66, 69, 70, 72, 75, 78 → Mean= 70.71

Including one young patient into this group:

25, 65, 66, 69, 70, 72, 75, 78 → Mean= 65 → does not reflect the real structure in the data

Descriptive Statistics

Robust measures against outliers and skewed distributions:

Quantile: An α-Quantile is a value dividing data in a way that the proportion α of the data is smaller and the proportion 1-α of the data is larger. In the case of a 0.95-Quantile, 95% of the values are smaller than the quantile and 5% of the values are larger.

Percentile: α -Quantile = $\alpha * 100\%$-Percentile, e.g. 0.95-Quantile = 95% Percentile

Terciles: 33.3% and 66.6%-Percentile

Quartiles: 25%, 50% and 75%- Percentile

Median: 50%- Percentile

Descriptive Statistics

Example for Median: 50%-Percentile

Given are the ages of a Myocardial Infarction patient group:

$\underbrace{65, 66, 69}$, 70, $\underbrace{72, 75, 78}$ → Median= 70

50% of the data are lower,
50 % are higher than the median

Including one young patient into this group:

$\underbrace{25, 65, 66, 69}$, $\underbrace{70, 72, 75, 78}$ → Median= 69.5

50% of the data are lower, 50 % are higher than the median

Descriptive Statistics

For quantitative data: Measures of dispersion or variability

The **Variance** S^2 and **Standard deviation** S measure the scattering of data around their mean:

$$S^2 = \frac{1}{n}((x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

A modified version: $S^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2$ → **sample variance**

The variance and standard deviation are also sensitive against outliers and skewed distributions → robust measures:

Range = Maximum-Minimum

Interquartile range = 0.75-Quantile – 0.25-Quantile

Descriptive Statistics

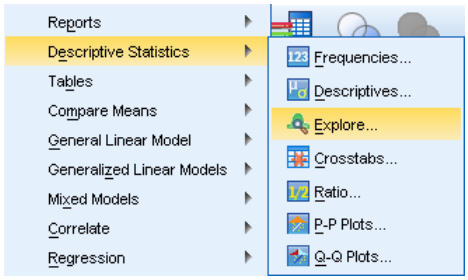
Example: Mean, quantiles and sd to summarize the age-distribution of a population-based study (n=1457)

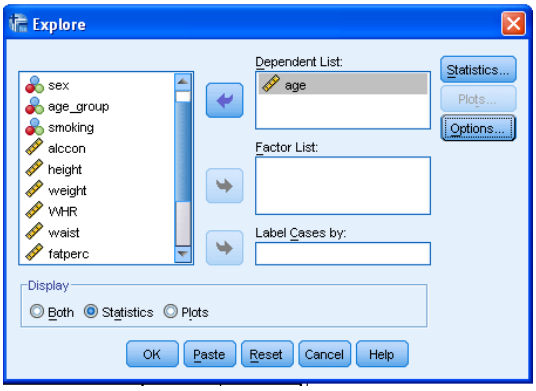
How are such summary data typically presented in scientific publications?

Characteristics of all participants (n=1457) Mean ± SD [25., 50.; 75. percentile for non-normal distribution] or number (%)		
Age (years)		53.25 ± 9.19
Sex (male/female), n (%)		722 (49.6) / 735 (50.4)
Smoking Status, n (%)	Current Smoker	261 (17.9)
	Ex-Smoker	453 (31.1)
	Never Smoker	743 (51.0)
Measured Parameter1 (g/dL)		3.61 ± 0.65 [3.30; 3.70; 4.20]
Measured Parameter2 (mmol/L)		35.5 ± 16.9 [15.5; 28.2; 46.7]

Descriptive Statistics

Example: Mean, quantiles and sd to summarize the age-distribution of a population-based study (n=1457), in SPSS:





Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
age	1457	100,0%	0	0,0%	1457	100,0%

Number of missing values

Descriptives

			Statistic	Std. Error
age	Mean		53,25	,241
	95% Confidence Interval for Mean	Lower Bound	52,78	
		Upper Bound	53,72	
	5% Trimmed Mean		53,60	
	Median		54,00	
	Variance		84,438	
	Std. Deviation		9,189	
	Minimum		30	
	Maximum		69	
	Range		39	
	Interquartile Range		13	
	Skewness		-,458	,064
	Kurtosis		-,172	,128

Measures of location and variability

Percentiles

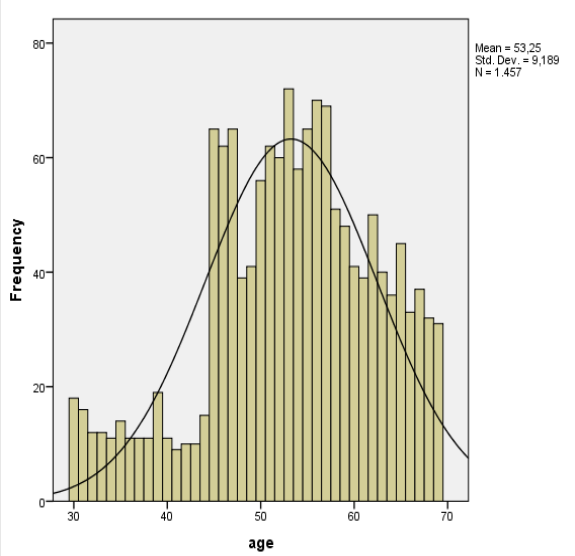
		Percentiles						
		5	10	25	50	75	90	95
Weighted Average (Definition 1)	age	35,00	40,00	47,00	54,00	60,00	65,00	67,00
Tukey's Hinges	age			47,00	54,00	60,00		

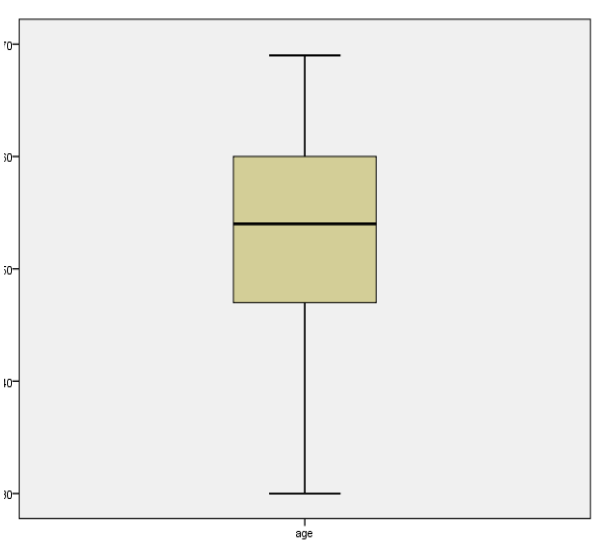
Percentiles

Descriptive Statistics

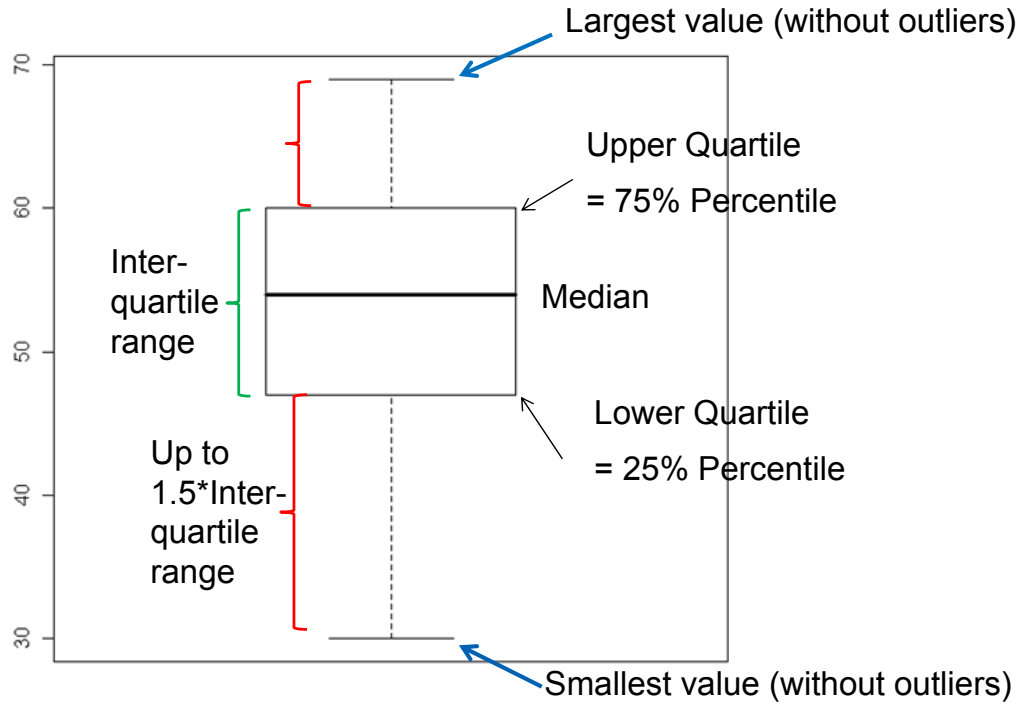
Point estimates are not sufficient to illustrate the complete distribution of a variable → Use Figures !

Histogram	Boxplot





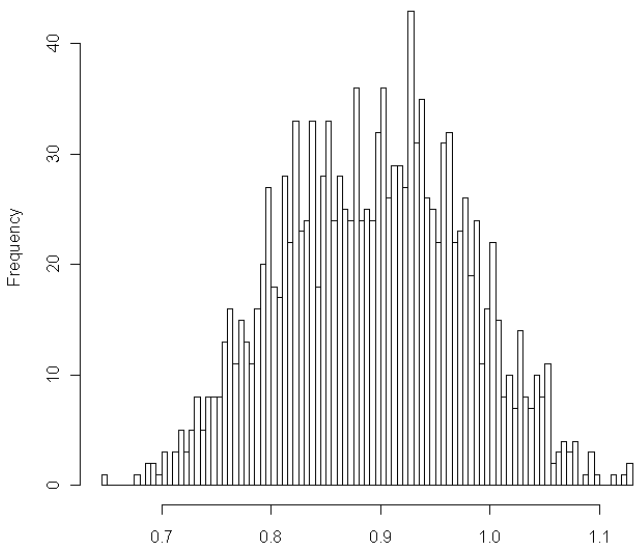
Descriptive Statistics

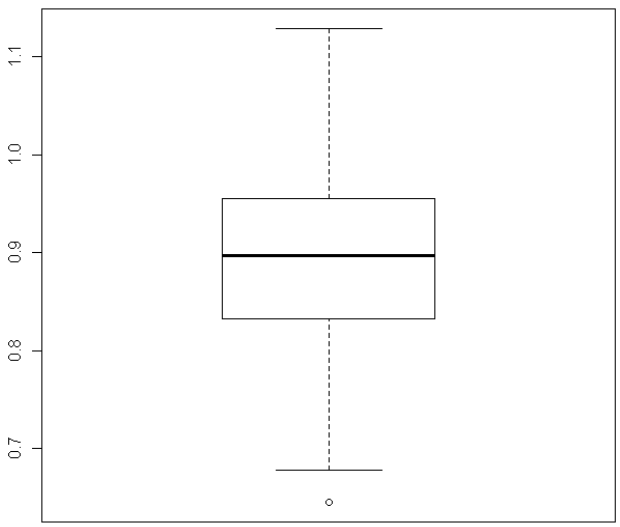


„Outliers“: Values above or below the whiskers are shown as single points and are denoted as outliers.

Descriptive Statistics

Histogram and Boxplot for a symmetrically distributed variable:





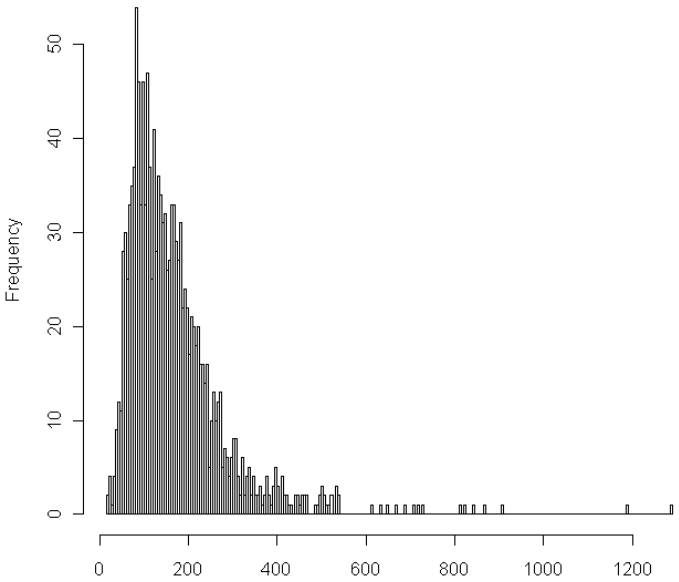
Histogram:

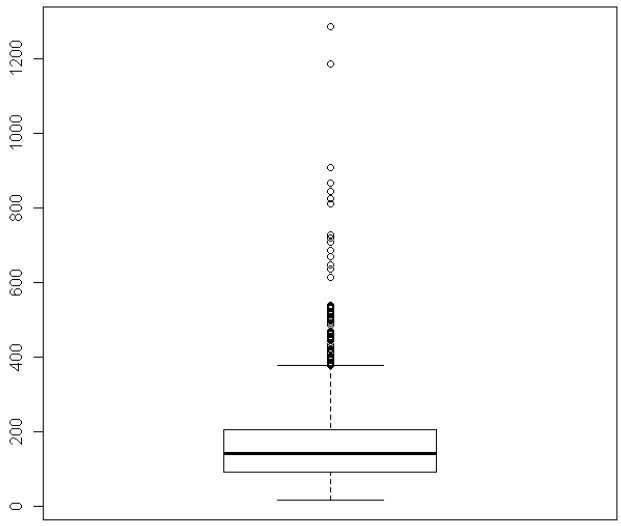
Graphical illustration of the distribution of a continuous variable

→ Useful to see, if a variable is symmetric, skewed or follows a specific distribution (e.g. normal distribution)

Descriptive Statistics

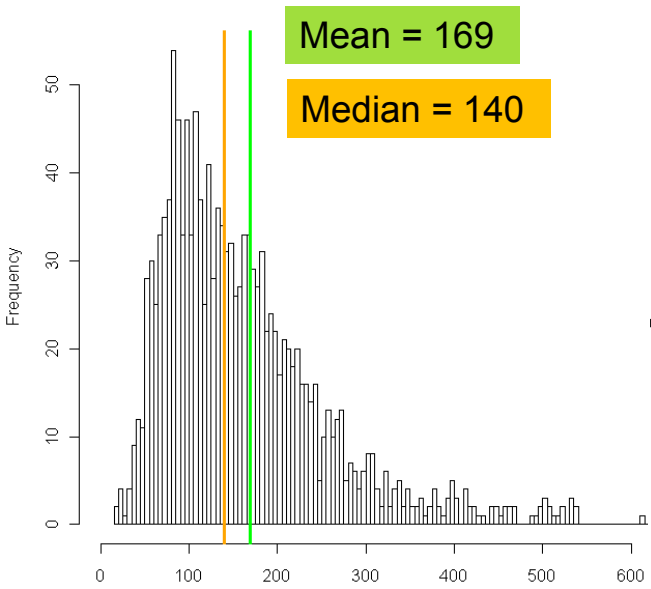
Histogram and Boxplot for a extremely right-skewed variable with outliers:



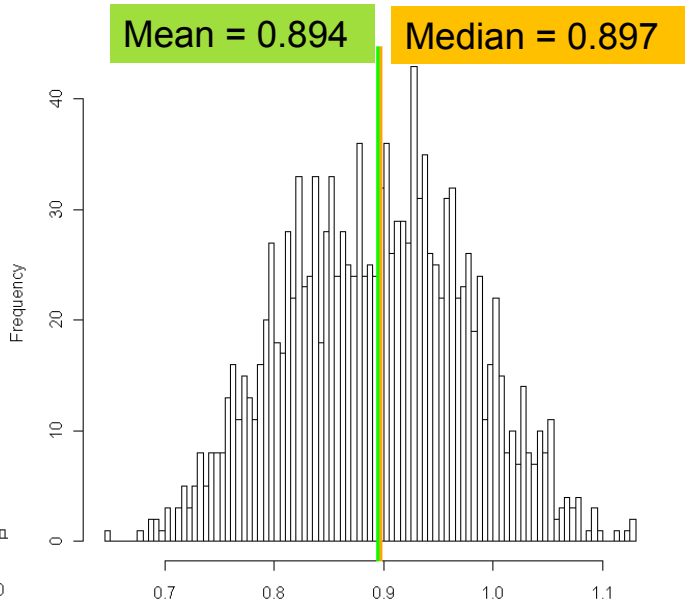


Descriptive Statistics

Right skewed distribution:



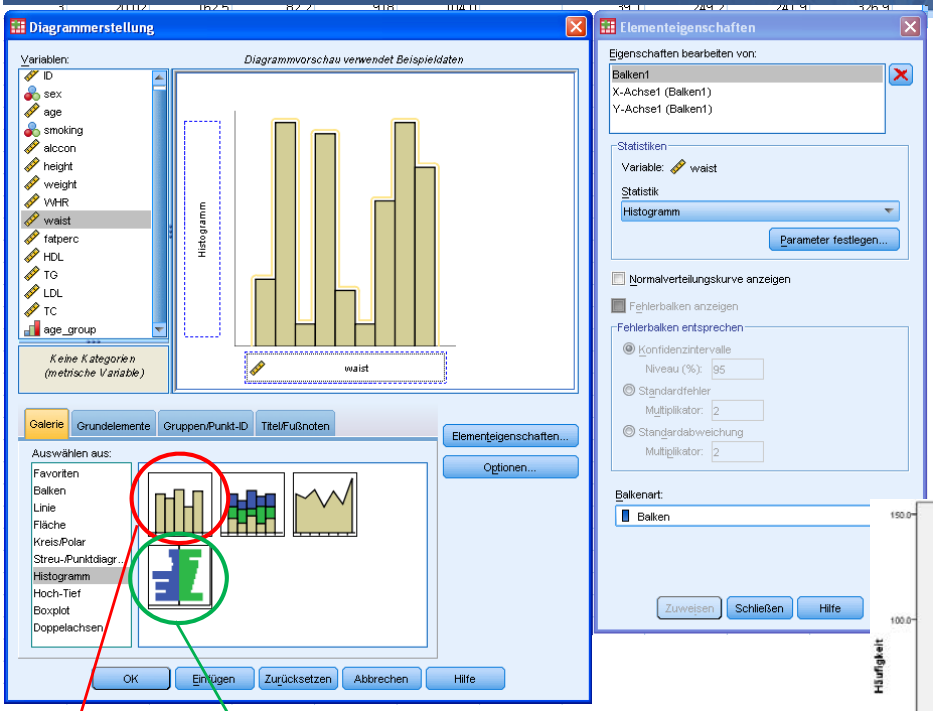
Symmetrically distributed:



For symmetrically distributed variables → Mean=Median

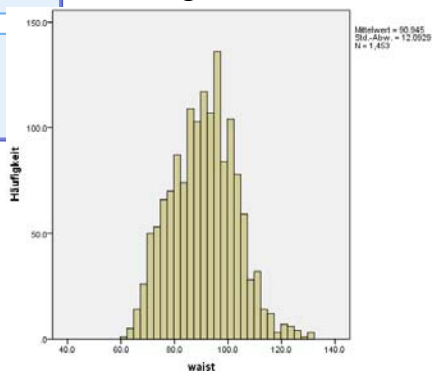
For skewed variables: Medians should be preferred

Descriptive Statistics



Histogram for waist in SPSS

Simple Histogramm:



Simple Histogram

Histogram seperated by a grouping variable, e.g. sex

Explorative Statistics

Describing the relationship between two variables:

Methods used	**Hypothesis that might be created**

- 2 qualitative variables (e.g. gender and smoking):

2-dimensional tables	Is variable 1 related to the other variable and vice versa ?

- 1 qualitative variable (e.g. gender), 1 quantitative variable (e.g. cholesterol):

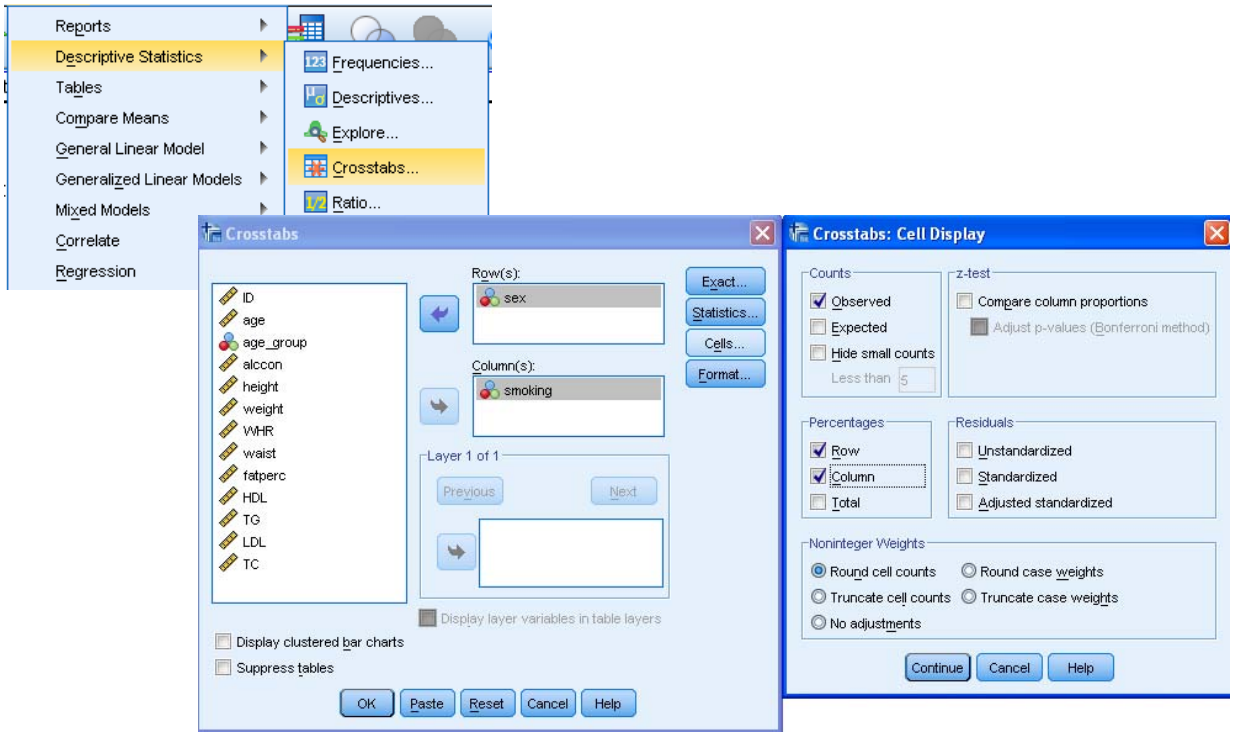
E.g. Comparison of measures of location of the quantitative variable between levels of the qualitative variable	Simple case: Does the mean of group 1 differ from the mean of group 2?

- 2 quantitative variables (e.g. cholesterol and age)

Correlation and scatterplots	Are the two variables associated with each other?

Explorative Statistics

2 qualitative variables: Crosstable including absolute and relative frequencies: do it in SPSS: gender and smoking



Explorative Statistics

2 qualitative variables: Crosstable including absolute and relative frequencies

Example:

Gender
x
Smoking

sex * smoking Crosstabulation

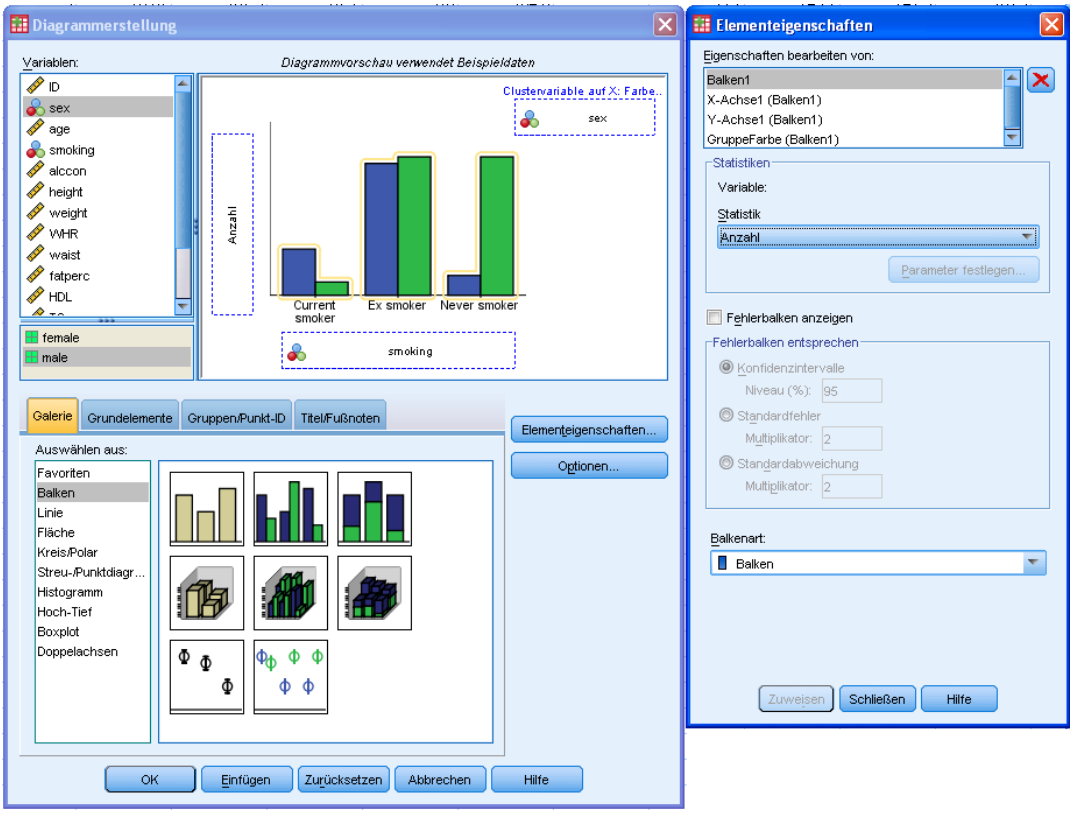
			smoking			
			Current smoker	Ex smoker	Never smoker	Total
sex	female	Count	117	143	475	735
		% within sex	15,9%	19,5%	64,6%	100,0%
		% within smoking	44,8%	31,6%	63,9%	50,4%
	male	Count	144	310	268	722
		% within sex	19,9%	42,9%	37,1%	100,0%
		% within smoking	55,2%	68,4%	36,1%	49,6%
Total		Count	261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

Different numbers have a different emphasis and interpretation. Examples:

1. Altogether, there are 475 women in the study, which have never smoked
2. 37.1% of all men have never smoked, but 64.6% women
3. 55.2% of all current smokers in the study are male

Explorative Statistics

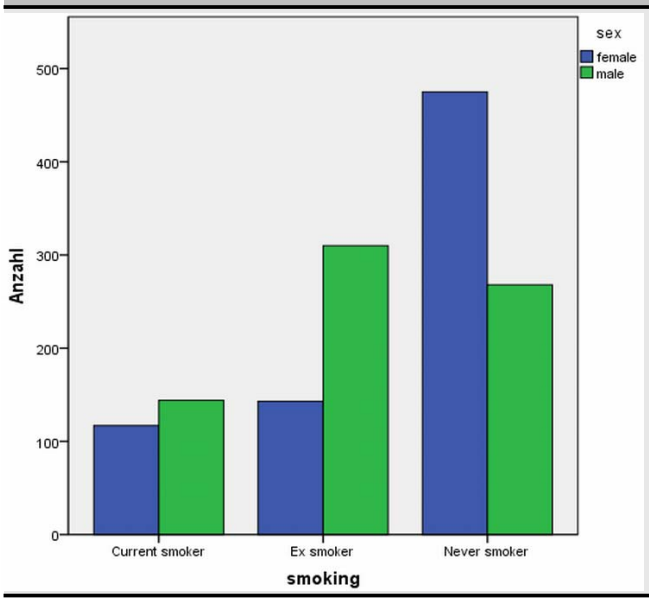
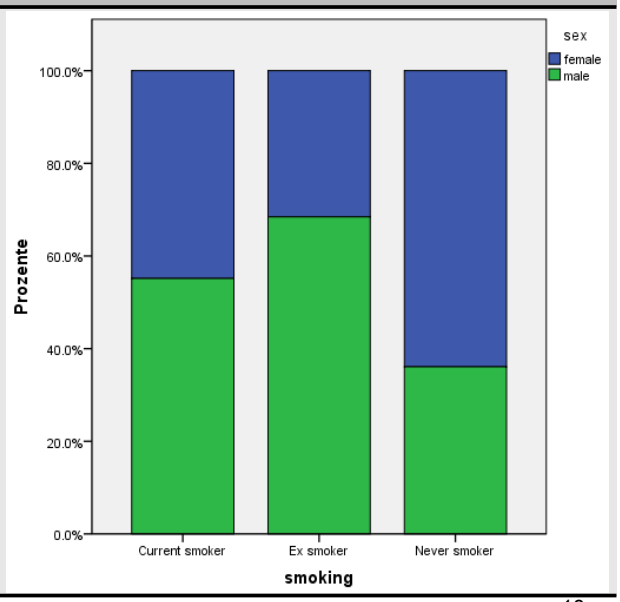
Illustrating 2-dimensional tables by barplots: Example: table on Gender x Smoking



Explorative Statistics

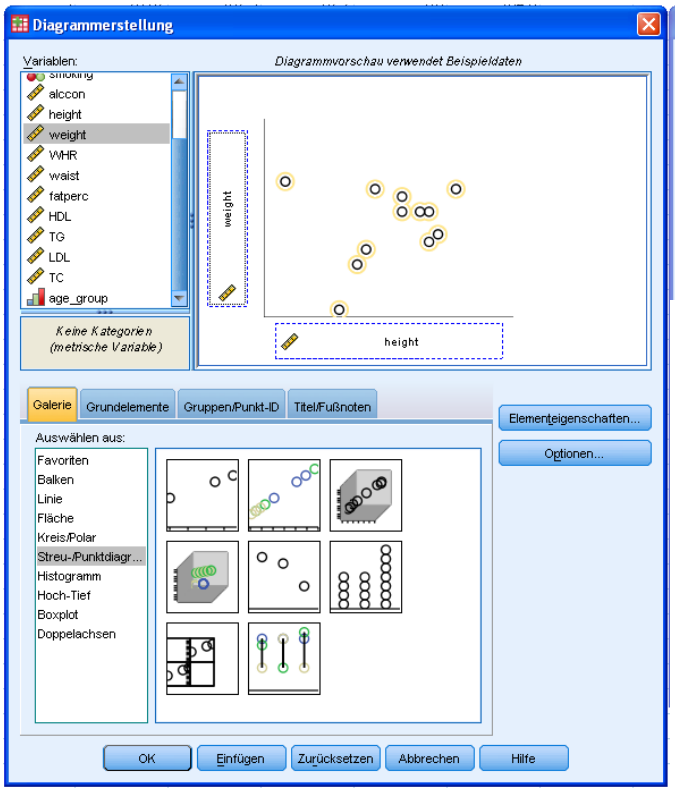
Illustrating 2-dimensional tables by barplots:

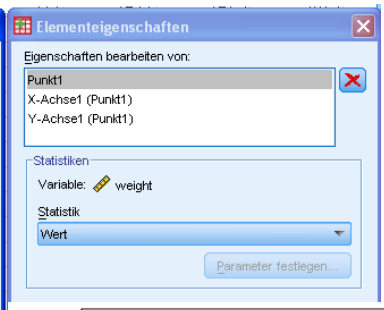
Example: table on Gender x Smoking: Different presentations for different statements

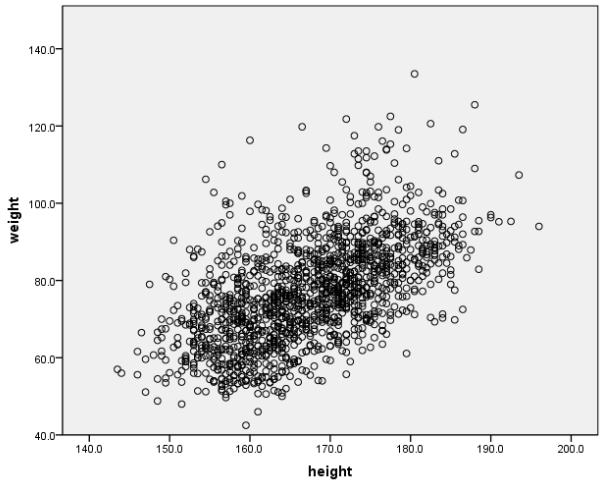
Barplots next to each other, absolute frequencies	Stacked barplot, relative frequencies (summing up to 100 for each smoking category)
	

Explorative Statistics

2 quantitative variables: Simple **scatter plots** between two variables:

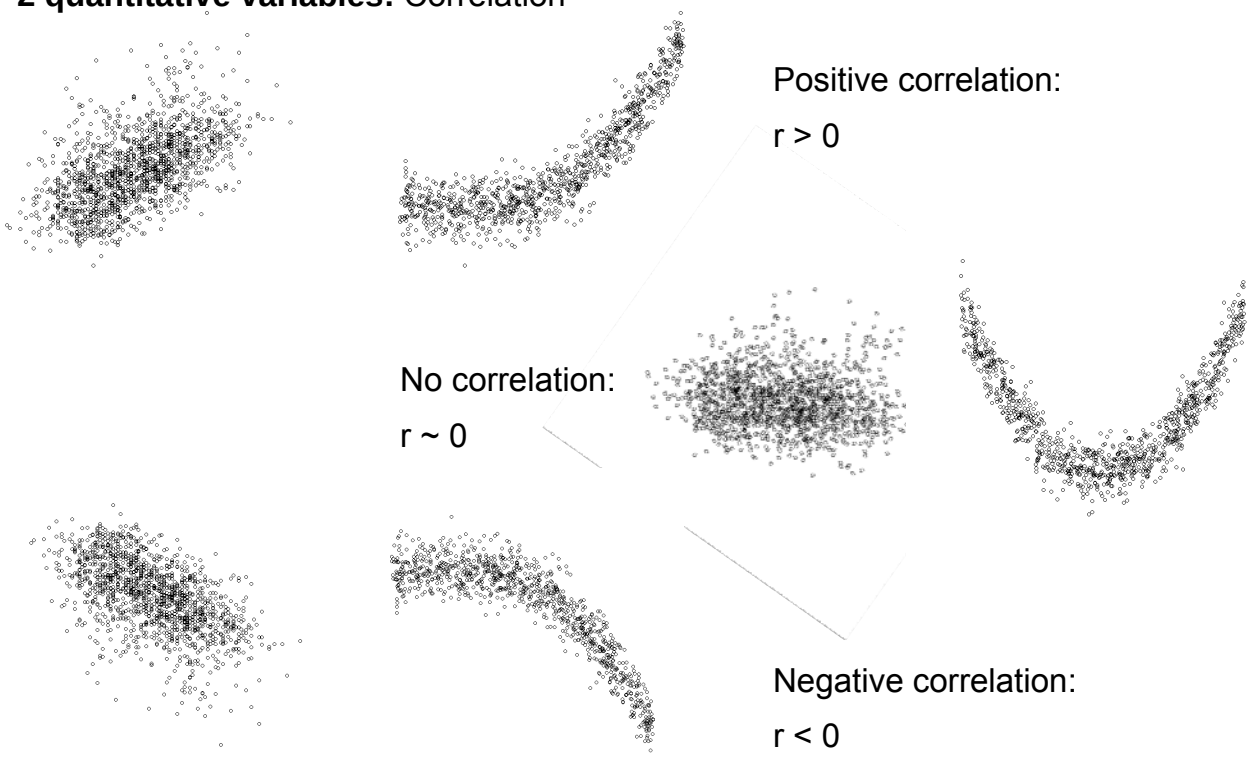






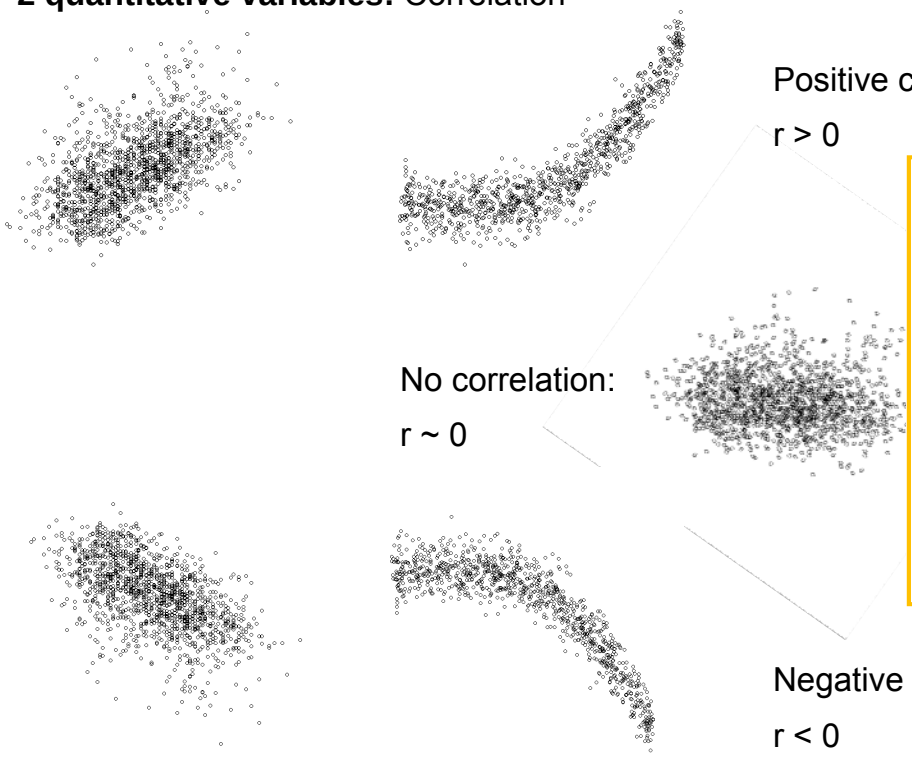
Explorative Statistics

2 quantitative variables: Correlation



Explorative Statistics

2 quantitative variables: Correlation



Positive correlation:

r > 0

No correlation:

r ~ 0

Negative correlation:

r < 0

There is a structure in the data, but

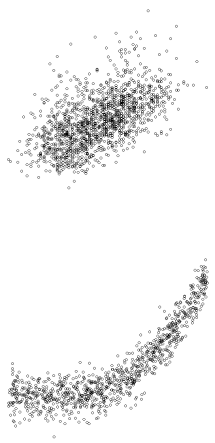
Correlation coefficient is not able to find it

Explorative Statistics

2 quantitative variables: Which correlation coefficient to use?

Pearson correlation coefficient: Test statistic is based on Pearson's product moment correlation, that follows a t-distribution (~approximation of normal distribution) **Measure of linear association**!	Can be used if data follow a bivariate normal distribution
Spearman correlation coefficient: Estimate a rank-based measure of association.	Useful also for nonlinear but monotonic relationships. Data can also be ranks or ordinal.

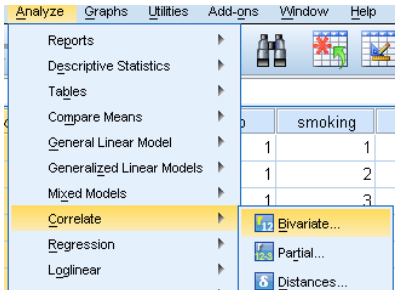
Observations of Var x	Rank rank(x)
11	1
15	2
17	3.5
17	3.5
22	4

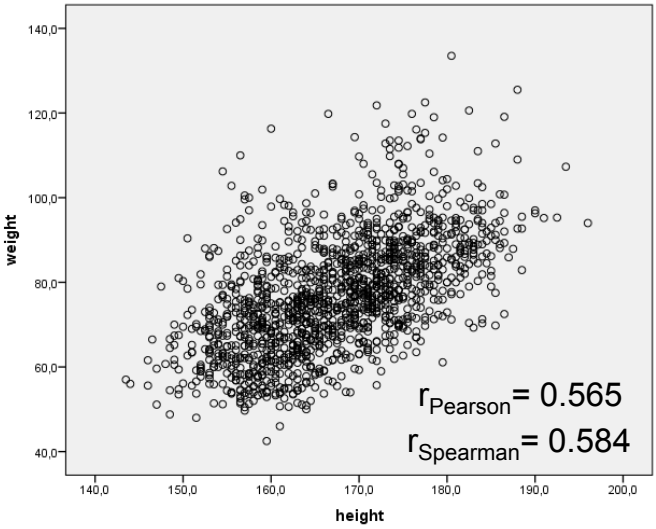


The correlation coefficient r ranges between -1 and 1

Often, the squared correlation coefficient r^2 is given, that ranges between 0 and 1

Explorative Statistics





Roughly, correlations can be interpreted in the following way:

|r| < 0.5 : weak correlation

0.5 ≤ |r| < 0.8 : moderate correlation

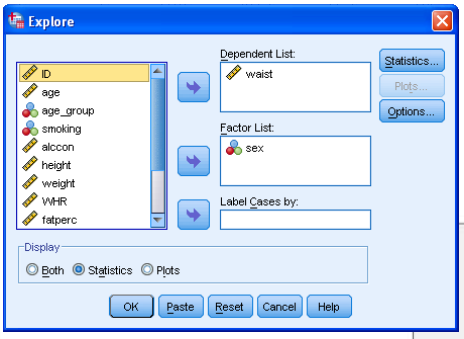
0.8 ≥ |r| : strong correlation

This interpretation always depends on the kind of data. For „weak“ variables (e.g. in social sciences), high correlations cannot be reached, in contrast to „strong“ variables (e.g. laboratory measurements).

Explorative Statistics

1 qualitative variable,
1 quantitative variable:

Example: Comparing waist between men and women

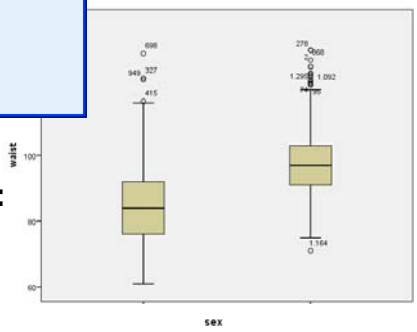


Descriptives

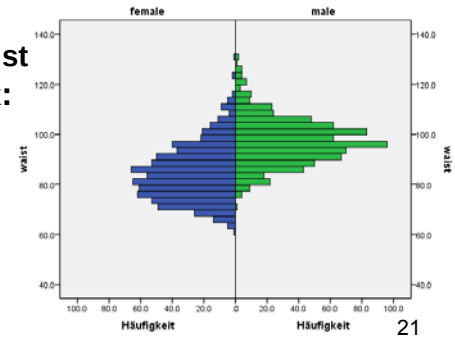
	sex			Statistic	Std. Error
waist	female	Mean		84,58	,408
		95% Confidence Interval for Mean	Lower Bound	83,78	
			Upper Bound	85,38	
		5% Trimmed Mean		84,14	
		Median		84,00	
		Variance		121,840	
		Std. Deviation		11,038	
		Minimum		61	
		Maximum		131	
		Range		70	
		Interquartile Range		16	
		Skewness		,585	,090
		Kurtosis		,277	,181
	male	Mean		97,39	,350
		95% Confidence Interval for Mean	Lower Bound	96,70	
			Upper Bound	98,08	
		5% Trimmed Mean		97,11	
		Median		97,00	
		Variance		88,519	
		Std. Deviation		9,408	
		Minimum		71	
		Maximum		132	
		Range		61	
		Interquartile Range		12	
		Skewness		,438	,091
		Kurtosis		,694	,182

Is it just by chance?

Boxplots for waist seperated by sex:



Histogramms for waist seperated by sex:



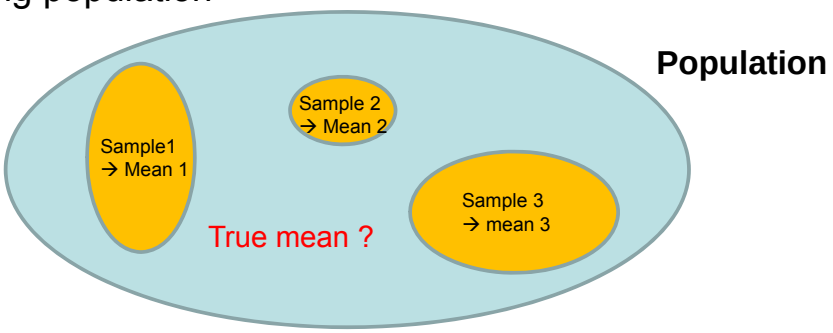
Percentiles

		sex	Percentiles						
			5	10	25	50	75	90	95
Weighted Average (Definition 1)	waist	female	68,80	71,00	76,00	84,00	92,00	99,40	104,20
		male	82,08	86,00	90,88	97,00	103,00	108,50	113,85
Tukey's Hinges	waist	female			76,00	84,00	92,00		
		male			91,00	97,00	103,00		

Point and Confidence Estimates

Point and confidence estimates

- So far: We have calculated measures of location (e.g. mean) in our study sample
- But remember: Our intention is to conclude from our sample on the underlying population



- There is uncertainty involved in the estimation of a population mean from the population mean → Standard error / Confidence Intervals

Point and confidence estimates

- There is uncertainty in parameter estimation because it is based on a random sample of finite size from the population of interest

→ Construct an interval, that includes the true population parameter with given certainty: **Confidence Interval CI**

→ The measure of certainty is given by the error probability α

α = 5% → 95% CI

α = 1% → 99% CI etc.

Interpretation of the confidence interval of the mean:

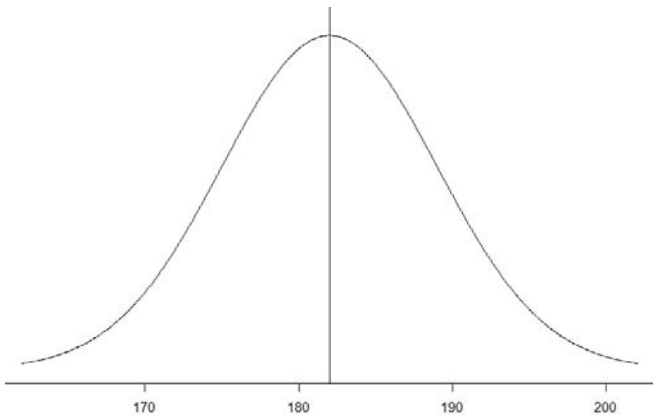
If the study is repeated 100 times on 100 different samples, the true mean will be within this range in $(1-\alpha = 95\%)$ percent of the studies

Point and confidence estimates

Example: 95% CI for height in Austrian men born 1992

True mean and true sd are known: mean=182 cm, sd = 7

and data are normally distributed!



Point and confidence estimates

Example: 95% CI for height in Austrian men born 1992

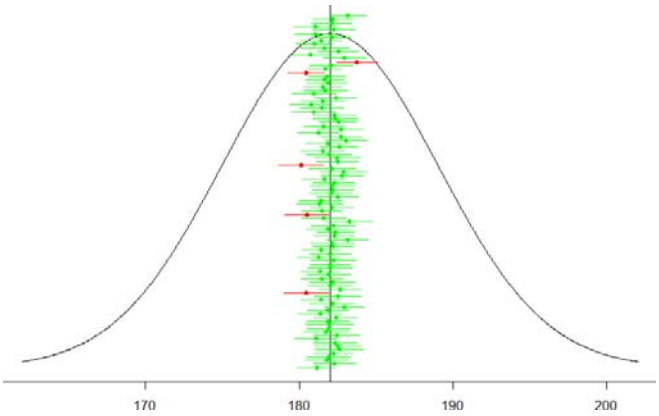
True mean and true sd are known: mean=182 cm, sd = 7

and data are normally distributed!

Experiment 1:

100 men are drawn randomly 100 times

→ 100 different mean values and 100 CIs



Point and confidence estimates

Example: 95% CI for height in Austrian men born 1992

True mean and true sd are known: mean=182 cm, sd = 7

and data are normally distributed!

Experiment 1:

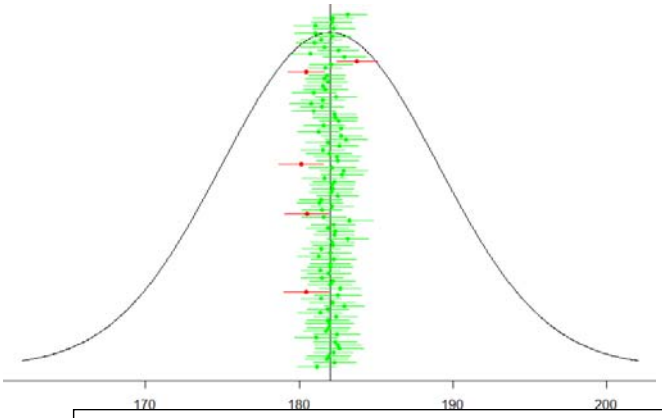
100 men are drawn randomly 100 times

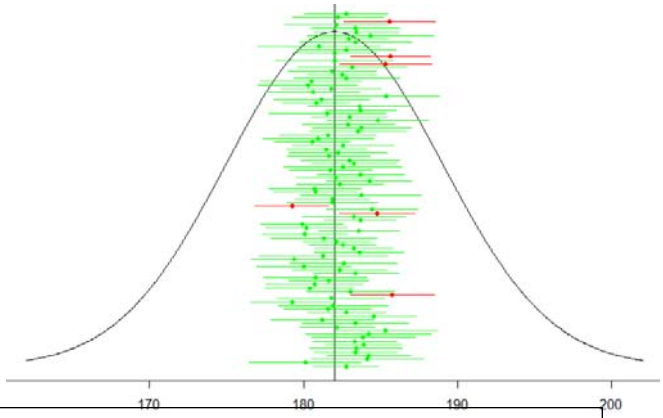
→ 100 different mean values and 100 CIs

Experiment 2:

20 men are drawn randomly 100 times

→ 100 different mean values and 100 CIs





5 out 100 95% CI do not cover the true mean of 182 cm just by chance!