



Statistics for Diploma and PhD Students "Basics"

-- Descriptive Statistics--

Josef Fritz

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



The aim of this lecture:

- Get to know statistical terms and definitions as they are used specifically in medical science
- How to interpret statistical results in the literature
- Application of appropriate statistical methods for your own data and observations
- Hands-on experience in a well-known statistics software
- Where are the pitfalls and sources of error?



Introduction

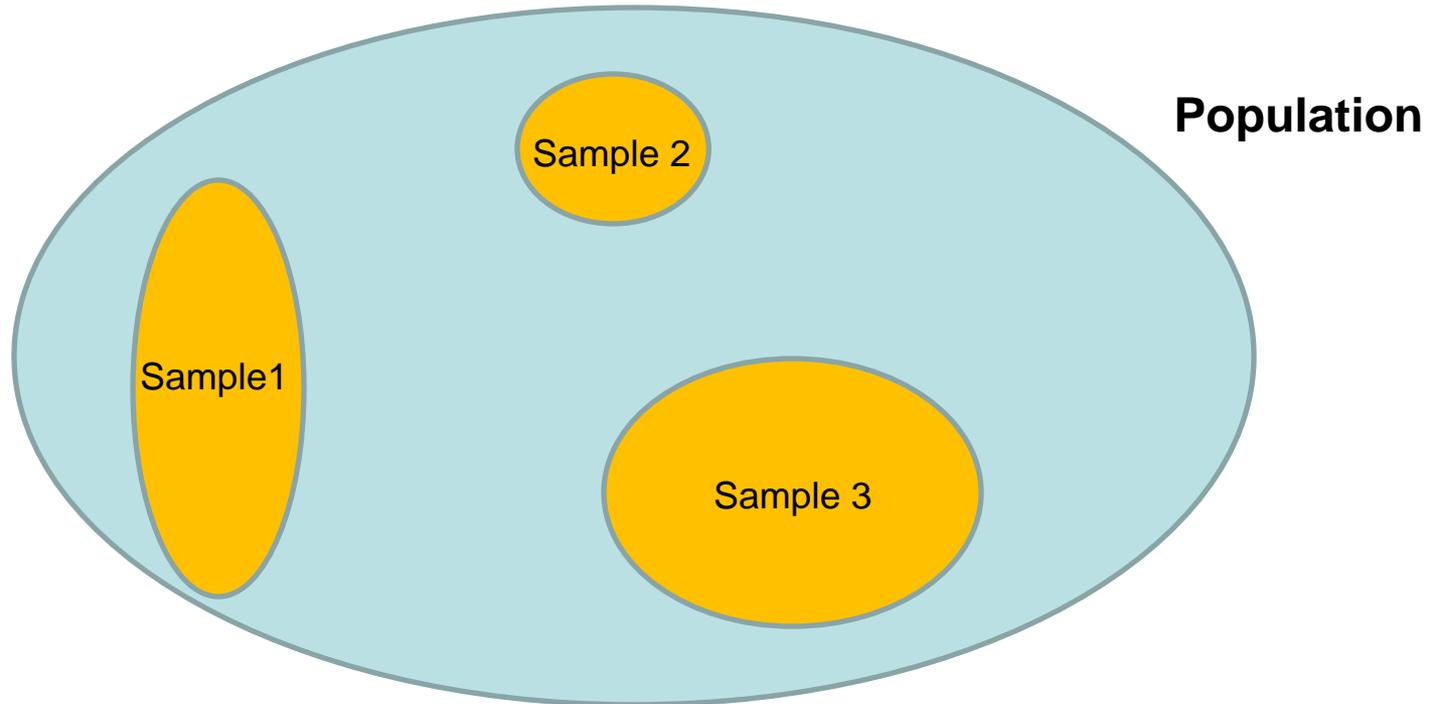
Statistics is one of the most important tools in science

The aim of statistical methods is:

- Describing data realistically (**Descriptive Statistics**)
- Formulate relations/correlations between observations and construct hypotheses (**Exploratory Statistics**)
- Draw conclusions from the observed sample to the underlying population (**Inductive Statistics**)

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

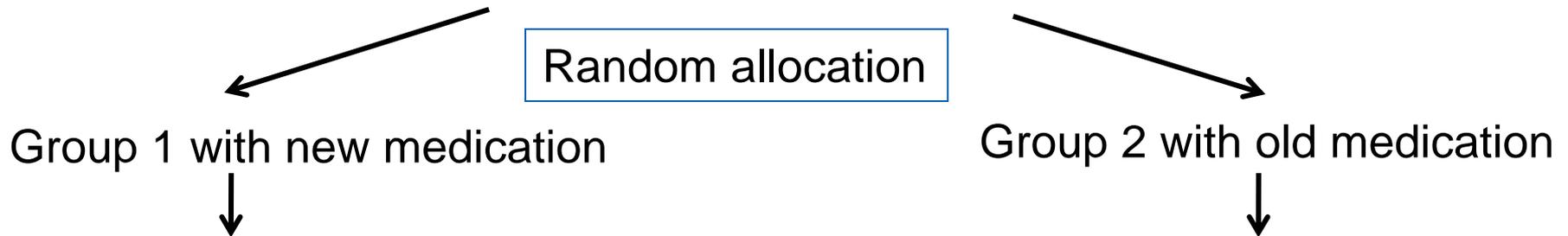
(Often, stages cannot be discriminated, however)

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



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 → no names !
- In such a dataset you have variables (PatID, Gender, Age, Disease)
- 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) has to be done

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, but have to be identifiable as such

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)

Data types and structuring of data

A special case: Repeated measurements per patient: A „typical“ example

Parameter1

	Measurement method		1	2
#1	Baseline		596	444.6
	1 month		517.8	387.4
	2 months		561.8	435.2
	Mean		558.533	422.4
#2	Baseline		633	653.2
	1 month		656.4	618.2
	2 months		625.4	542.4
	Mean		638.267	604.6
#3	Baseline		824.8	818.6
	1 month		815.2	851
	2 months		782.4	820.8
	Mean		807.467	830.1333

Data types and structuring of data

Parameter1

	Measurement method		1	2
#1	Baseline	596	444.6	
	1 month	517.8	387.4	
	2 months	561.8	435.2	
	Mean	558.533	422.4	
#2	Measurement method	Baseline	633	653.2
		1 month	656.4	618.2
		2 months	625.4	542.4
	Mean	638.267	604.6	
#3	Measurement method	Baseline	824.8	818.6
		1 month	815.2	851
		2 months	782.4	820.8
	Mean	807.467	830.1333	

There are 2 measurement methods which have to be compared

3 different time points of measurement

Data types and structuring of data

Parameter1

	Measurement method		1	2
#1	Baseline		596	444.6
	1 month		517.8	387.4
	2 months		561.8	435.2
	Mean		558.533	422.4
#2	Baseline		633	653.2
	1 month		656.4	618.2
	2 months		625.4	542.4
	Mean		638.267	604.6
#3	Baseline		824.8	818.6
	1 month		815.2	851
	2 months		782.4	820.8
	Mean		807.467	830.1333

2 „levels“ of data:
individual data &
aggregated data
(mean)

NEVER EVER in
one dataset!

Data types and structuring of data

Two possibilities for repeated measures:

- One line for each patient: „**Wide**“ data format

→ Unique Identifier: Patient ID

- One line for each observation: „**Long**“ data format

→ There is not one unique Identifier, but a combination of variables: In this case:

(Patient ID) x (Measurement method) x (Time point of measurement)

Data types and structuring of data

- One line for each patient: **Wide format**

This is an intuitive format, but: not very easy to read

Here: 6 columns per parameter → 18 columns only for the measurements

	A	B	C	D	E	F	G	H	I	J	K	L
1	Patient	Age	Sex	P1_M1_T1	P1_M1_T2	P1_M1_T3	P1_M2_T1	P1_M2_T2	P1_M2_T3	P2_M1_T1	P2_M1_T2	P2_M1_T3
2	1	46	m	596	517.8	561.8	444.6	387.4	435.2	662	693.8	621.8
3	2	77	f	633	656.4	625.4	653.2	618.2	542.4	508.6	593.6	542.4
4	3	62	m	824.8	815.2	782.4	818.6	851	820.8	454.8	556.2	460.8

Time 1 2 3 1 2 3 1 2

Method 1 2 1

Parameter 1 2

Metadata for each patient can be added

Data types and structuring of data

- One line for each observation: **Long format**

2 methods x 3 timepoints

- 6 lines per patient
- 1 column for each parameter

	A	B	C	D	E	F	G	H
1	Patient	Age	Sex	time	method	Parameter1	Parameter2	Parameter3
2	1	46	m	1	1	596	662	1008.6
3	1	46	m	2	1	517.8	693.8	1243.8
4	1	46	m	3	1	561.8	625.2	1053.4
5	1	46	m	1	2	444.6	616	1149.7
6	1	46	m	2	2	387.4	618.6	1180.6
7	1	46	m	3	2	435.2	511	1217
8	2	77	f	1	1	633	508.6	1163.8
9	2	77	f	2	1	656.4	593.6	1234.6
10	2	77	f	3	1	625.4	518	1240.2
11	2	77	f	1	2	653.2	522.2	1196.6
12	2	77	f	2	2	618.2	472.6	1042.6
13	2	77	f	3	2	542.4	548.6	1365.8
14	3	62	m	1	1	824.8	454.8	1406.4
15	3	62	m	2	1	815.2	556.2	1372.4
16	3	62	m	3	1	782.4	465.2	1406.4
17	3	62	m	1	2	818.6	483.8	1075.4
18	3	62	m	2	2	851	557.6	1218.8
19	3	62	m	3	2	820.8	455	1620.2

- Here: Metadata are redundant
- But: Best format for most analyses methods
- With SPSS it is possible to create a wide or long data format as long as there are key variables



**Import Data
and
Data Handling in SPSS**

Type in data from scratch

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

PatID	Gender	Age	Disease
1,00	1,00	52,00	,00
2,00	1,00	23,00	,00
3,00	2,00	79,00	1,00
4,00	2,00	64,00	.
5,00	1,00	55,00	,00
6,00	2,00	50,00	.
7,00	2,00	32,00	,00
8,00	1,00	44,00	1,00

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
PatID	Numeric	8	2		None	None	8	Right	Scale	Input
Gender	Numeric	8	0	Gender	{1, male}...	None	8	Right	Nominal	Input
Age	Numeric	8	2	Age	None	None	8	Right	Scale	Input
Disease	Numeric	8	0	Disease status	{0, healthy}...	None	8	Right	Nominal	Input

Attention: 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!

Data types and structuring of data

- 2 Examples in SPSS: Create long format (use test.sav): e.g used for repeated measure analysis

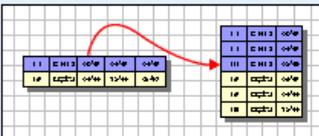
- Define Variable Properties...
- Set Measurement Level for Unknown...
- Copy Data Properties...
- New Custom Attribute...
- Define Dates...
- Define Multiple Response Sets...
- Identify Duplicate Cases...
- Sort Cases...
- Sort Variables...
- Transpose...
- Merge Files
- Restructure...**

Restructure Data Wizard

Welcome to the Restructure Data Wizard!

This wizard helps you to restructure your data from multiple variables (columns) in a single case to groups of related cases (rows) or vice versa, or you can choose to transpose your data.

i The wizard replaces the current data set with the restructured data. Note that data restructuring cannot be undone.



What do you want to do?

- Restructure selected variables into cases.

Use this when each case in your current data has some variables that you would like to rearrange into groups of related cases in the new data set.

repeated measure analysis

Restructure Data Wizard - Step 2 of 7

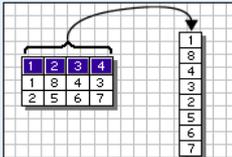
Variables to Cases: Number of Variable Groups

You have chosen to restructure selected variables into groups of related cases in the new file.

A group of related variables, called a variable group, represents measurements on one variable.

i For example, the variable may be width. If it is recorded in three separate measurements, each one representing a different point in time--w1, w2, and w3, then the data are arranged in a group of variables.

If there is more than one variable in the file often it is also recorded in a variable group, for example height, recorded in h1, h2, and h3.



How many variable groups do you want to restructure?

- One (for example, w1, w2, and w3)

Restructure Data Wizard - Step 3 of 7

Variables to Cases: Select Variables

For each variable group you have in the current data the restructured file will have one target variable.

In this step, choose how to identify case groups in the restructured data, and choose which variables belong with each target variable.

Optionally, you can also choose variables to copy to the new file as Fixed Variables.

Variables in the Current File:

- PatID
- Gender
- Age
- Disease
- P1
- P1_2
- P1_3

Case Group Identification:

Use selected variable

Variable: PatID

Variables to be Transposed:

Target Variable: P1_group

- P1
- P1_2
- P1_3

Fixed Variable(s):

- Gender
- Age
- Disease

Data types and structuring of data

Restructure Data Wizard - Step 4 of 7



Variables to Cases: Create Index Variables

In the current data, values for a variable group appear in a single case in multiple variables. For example, a single case contains the values for w1, w2, and w3.

In the new data, values for a variable group will appear in multiple cases in a single variable. For example, there will be three cases, one each for w1, w2, and w3.

An index is a new variable that identifies the group of new cases that was created from the original case. For example, an index named "w" would have the values 1, 2, and 3.

1	1	1	0.07
1	1	2	0.11
1	1	3	0.05
2	1	1	0.08
2	1	2	0.04
2	1	3	0.06

How many index variables do you want to create?

One

Use this when a variable group records the effects of a single factor, treatment or condition.

More than one How many?

Use this when a variable group records the effects of more than one factor, treatment or condition.

None

Use this if index information is stored in one of the sets of variables to be transposed.

1	1	1	1	0.07
1	1	1	2	0.11
1	1	1	3	0.05
1	1	2	1	0.08
1	1	2	2	0.04
1	1	2	3	0.06

1	1	0.08	2	0.07
2	1	0.11	2	0.11
3	1	0.07	2	0.05
4	1	0.06	2	0.08
5	1	0.09	2	0.04
6	1	0.02	2	0.06

< Back

Next >

Finish

Cancel

Help

Data types and structuring of data

Restructure Data Wizard - Step 6 of 7

Variables to Cases: Options

In this step you can set options that will be applied to the restructured data file.

Handling of Variables not Selected

- Drop variable(s) from the new data file
- Keep and treat as fixed variable(s)

System Missing or Blank Values in all Transposed Variables

- Create a case in the new file
- Discard the data

Case Count Variable

- Count the number of new cases created by the case in the current data

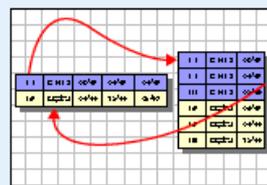
Name:

Label:

< Back Next > Finish Cancel Help

Restructure Data Wizard - Finish

Finish



What do you want to do?

- Restructure the data now
Use this when you want to replace the current file immediately.
- Paste the syntax generated by the wizard into a syntax window
Use this when you want to save or modify the syntax before you restructure the data.



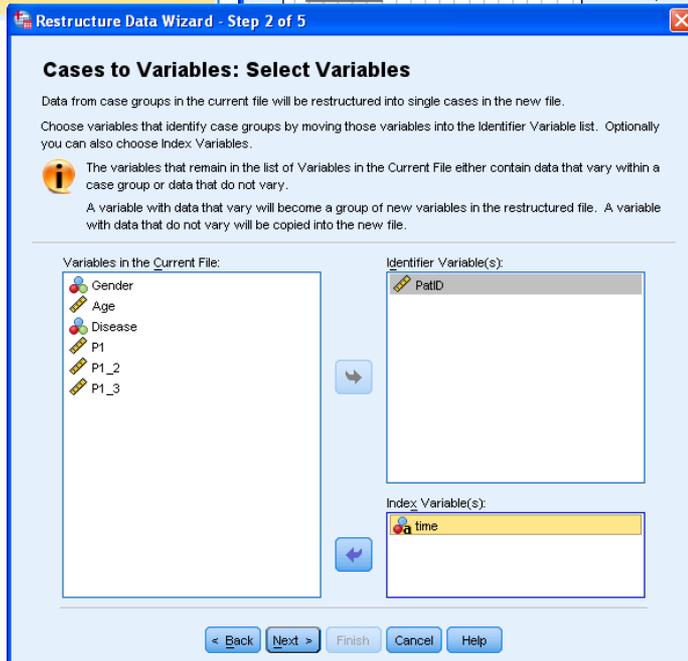
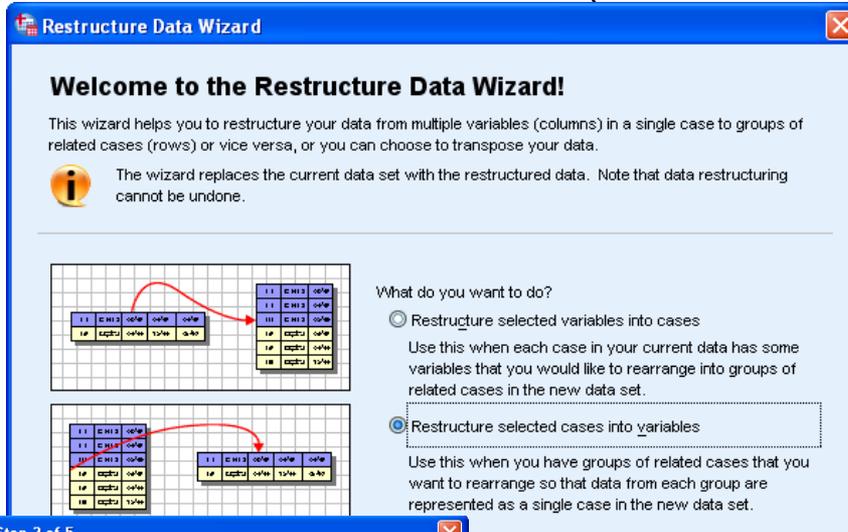
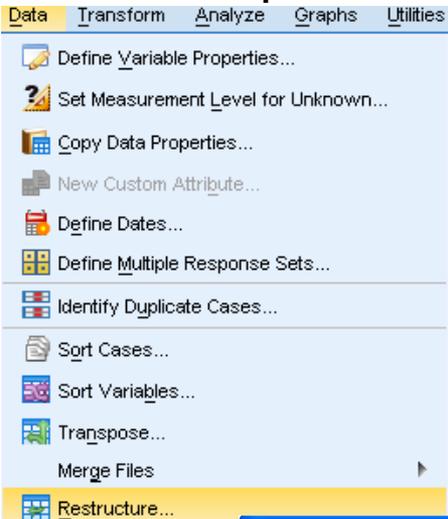
< Back Next > Finish Cancel Help

Data types and structuring of data

PatID	Gender	Age	Disease	P1_group
1,00	1	52,00	,00	1,00
1,00	1	52,00	,00	2,00
1,00	1	52,00	,00	3,00
2,00	1	23,00	,00	2,00
2,00	1	23,00	,00	3,00
2,00	1	23,00	,00	4,00
3,00	2	79,00	1,00	3,00
3,00	2	79,00	1,00	4,00
3,00	2	79,00	1,00	5,00
4,00	2	64,00	.	4,00
4,00	2	64,00	.	5,00
4,00	2	64,00	.	6,00
5,00	1	55,00	,00	5,00
5,00	1	55,00	,00	6,00
5,00	1	55,00	,00	7,00
6,00	2	50,00	.	6,00
6,00	2	50,00	.	7,00
6,00	2	50,00	.	8,00
7,00	2	32,00	,00	7,00
7,00	2	32,00	,00	8,00
7,00	2	32,00	,00	9,00
8,00	1	44,00	1,00	8,00
8,00	1	44,00	1,00	9,00
8,00	1	44,00	1,00	10,00

Data types and structuring of data

■ Example in SPSS: Create wide format (use test3.sav)



e.g. used for paired t-test
(comparing values before and
after treatment)

Data types and structuring of data

Restructure Data Wizard - Step 3 of 5



Cases to Variables: Sorting Data

The variables that you used to identify case groups in the current file need to be sorted before the file can be restructured. If you are not sure about your data, select "Yes".

2	1	3	.006
3	1	1	.010
1	1	1	.003
2	1	1	.008
2	1	2	.007
1	1	2	.004
1	1	3	.002

1	1	1	.003
1	1	2	.004
1	1	3	.002
2	1	1	.008
2	1	2	.007
2	1	3	.006
3	1	1	.010

Sort the current data?

Yes - data will be sorted by the Identifier and Index variabl...

2	1	3	.006
3	1	1	.010
1	1	1	.003
2	1	1	.008
2	1	2	.007
1	1	2	.004
1	1	3	.002

1	1	1	.003
1	1	2	.004
1	1	3	.002
2	1	1	.008
2	1	2	.007
2	1	3	.006
3	1	1	.010

No - use the data as currently sorted

< Back

Next >

Finish

Cancel

Help

Data types and structuring of data

Cases to Variables: Options

In this step you can set options that will be applied to the restructured data file.

Order of New Variable Groups

- Group by original variable (for example: w1 w2 w3, h1 h2 h3)
- Group by index (for example: w1 h1, w2 h2, w3 h3)

Case Count Variable

Count the number of cases in the current data used to create a new case

Name:

Label:

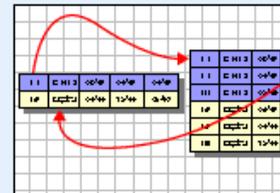
Indicator Variables

Create indicator variables

Root Name:

< Back Next > Finish Cancel Help

Finish



What do you want to do?

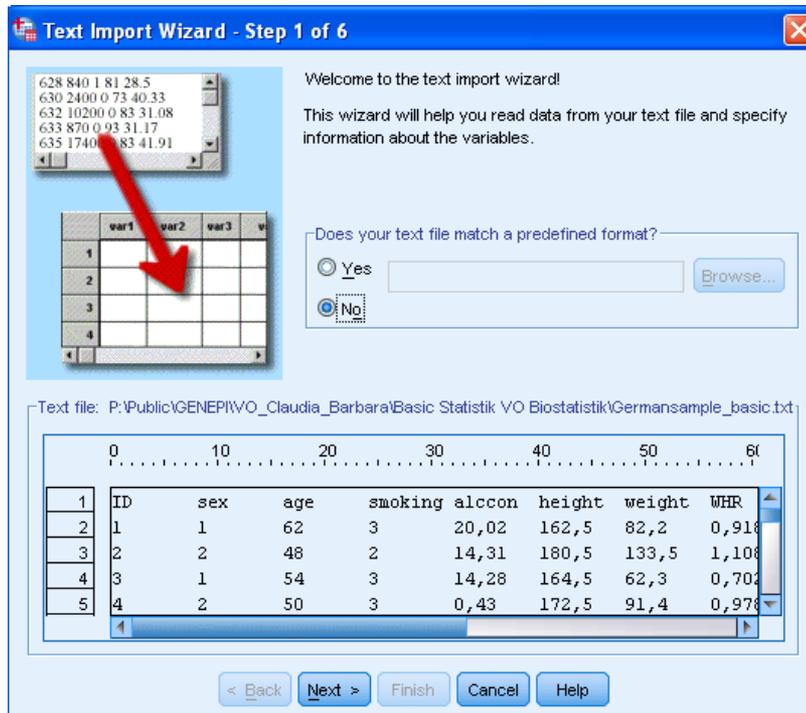
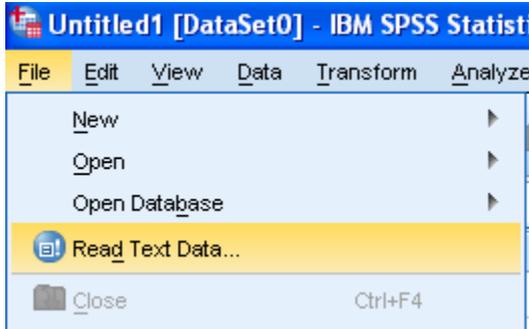
- Restructure the data now
Use this when you want to replace the current file immediately.
- Paste the syntax generated by the wizard into a syntax window
Use this when you want to save or modify the syntax before you restructure the data.

< Back Next > Finish Cancel Help

PatID	Gender	Age	Disease	P1.after	P1.before	P2.after	P2.before	P3.after	P3.before
1,00	1	52,00	,00	4,00	1,00	5,00	2,00	6,00	3,00
2,00	1	23,00	,00	7,00	2,00	8,00	3,00	9,00	4,00
3,00	2	79,00	1,00	6,00	3,00	7,00	4,00	8,00	5,00

Data types and structuring of data

■ Read in the dataset Basicdata.txt into SPSS



Data types and structuring of data

Text Import Wizard - Step 2 of 6

How are your variables arranged?

Delimited - Variables are delimited by a specific character (i.e., comma, tab).
 Fixed width - Variables are aligned in fixed width columns.

Are variable names included at the top of your file?

Yes
 No

Text file: P:\Public\GENEPIVO_Claudia_Barbara\Basic Statistik VO Biostatistik\Germansample_basic.txt

	0	10	20	30	40	50	60	
1	1	1	62	3	20,02	162,5	82,2	0,918
2	2	2	48	2	14,31	180,5	133,5	1,108
3	3	1	54	3	14,28	164,5	62,3	0,702
4	4	2	50	3	0,43	172,5	91,4	0,978
5	5	2	62	1	42	160,5	26,8	0,06

< Back Next > Finish Cancel Help

Text Import Wizard - Delimited Step 3 of 6

The first case of data begins on which line number? 2

How are your cases represented?

Each line represents a case
 A specific number of variables represents a case: 23

How many cases do you want to import?

All of the cases
 The first 1000 **cases.**
 A random percentage of the cases (approximate): 10 %

Data preview

	0	10	20	30	40	50	60	
1	1	1	62	3	20,02	162,5	82,2	0,918
2	2	2	48	2	14,31	180,5	133,5	1,108
3	3	1	54	3	14,28	164,5	62,3	0,702
4	4	2	50	3	0,43	172,5	91,4	0,978
5	5	2	62	1	42	160,5	26,8	0,06

< Back Next > Finish Cancel Help

Data types and structuring of data

Text Import Wizard - Delimited Step 4 of 6

Which delimiters appear between variables?

Tab Space
 Comma Semicolon
 Other:

What is the text qualifier?

None
 Single quote
 Double quote
 Other:

Data preview

ID	sex	age	smoking	alcocon	height	we
1	1	62	3	20,02	162,5	82,
2	2	48	2	14,31	180,5	133
3	1	54	3	14,28	164,5	62,
4	2	50	3	0,43	172,5	91,
5	2	67	1	43	169,5	76,
6	2	52	1	0	177	91,
7	2	34	2	0	178,5	75,
8	2	52	2	48,58	184,5	86,

< Back Next > Finish Cancel Help

Text Import Wizard - Step 5 of 6

Specifications for variable(s) selected in the data preview

Variable name: Original Name: alcocon

Data format:

Data preview

ID	sex	age	smoking	alcocon	height	we
1	1	62	3	20,02	162,5	82,
2	2	48	2	14,31	180,5	133
3	1	54	3	14,28	164,5	62,
4	2	50	3	0,43	172,5	91,
5	2	67	1	43	169,5	76,

< Back Next > Finish Cancel Help

Check data format for each variable, has to be numeric!

Data types and structuring of data

Text Import Wizard - Step 6 of 6

You have successfully defined the format of your text file.

Would you like to save this file format for future use?

Yes No

Would you like to paste the syntax?

Yes No Cache data locally

Press the Finish button to complete the text import wizard.

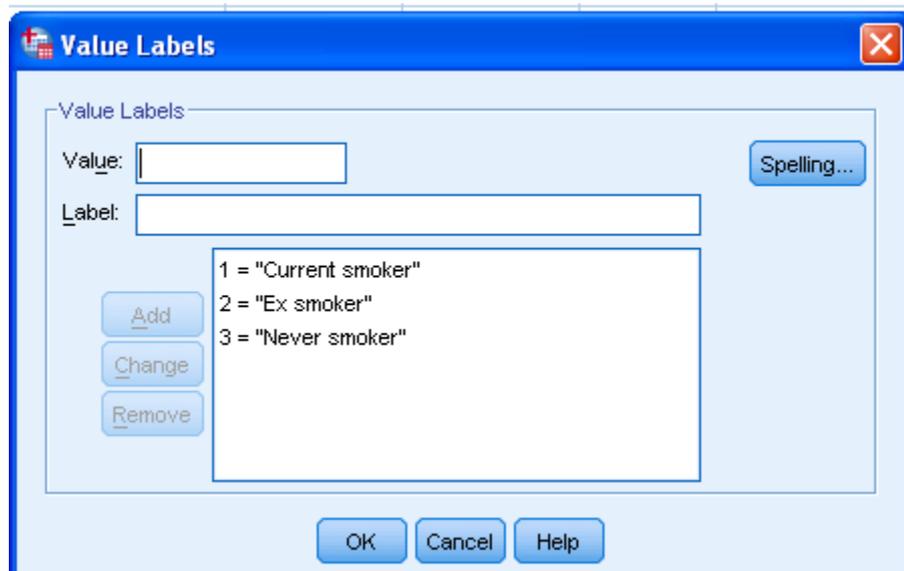
Data preview

ID	sex	age	smoking	alcco	height	we
1	1	62	3	20,02	162,5	82,
2	2	48	2	14,31	180,5	133
3	1	54	3	14,28	164,5	62,
4	2	50	3	0,43	172,5	91,
5	2	67	1	43	169,5	76,
6	2	52	1	0	177	91,
7	2	64	2	0	178,5	75,

< Back Next > Finish Cancel Help

Data types and structuring of data

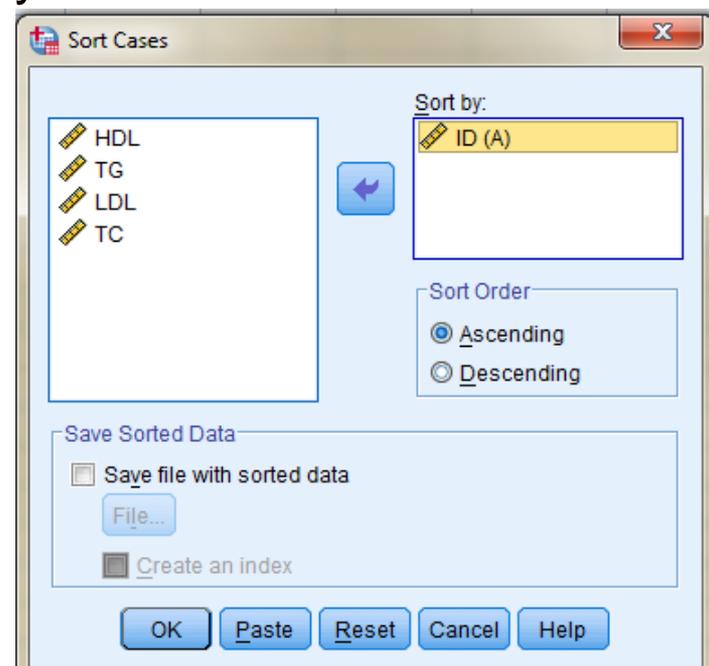
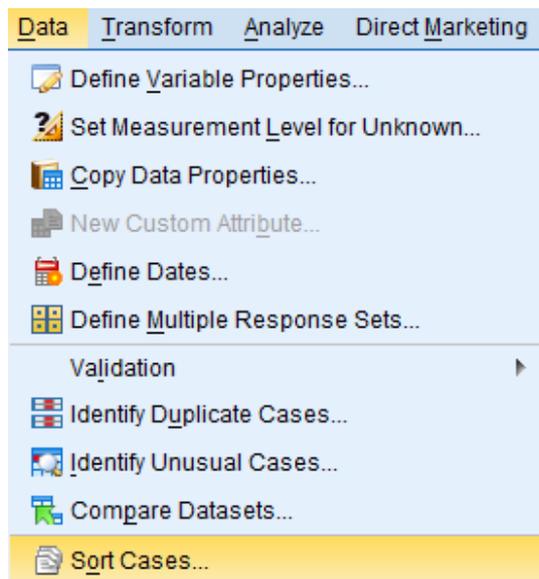
- Label the variable smoking: 1= Current smoker, 2= Ex-smoker, 3= Never smoker



- Label the variable sex: 1= female, 2= male
- Save your SPSS-Dataset under Basicdata.sav

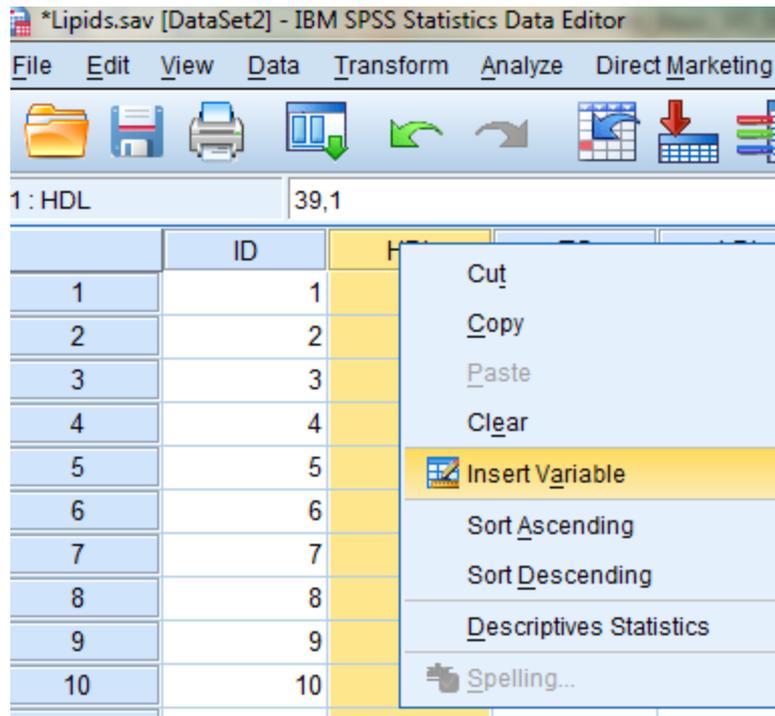
Read in data from other data sources

- The dataset Lipids.txt contains additional variables for the same patients.
- **Exercise:** Read the dataset Lipids.txt into SPSS and save it as Lipids.sav
- Now: We want to merge both datasets together into one dataset
 - what is the key variable, that identifies the observations?
 - both datasets have to be sorted by the key variable
 - here: Sort by “ID”

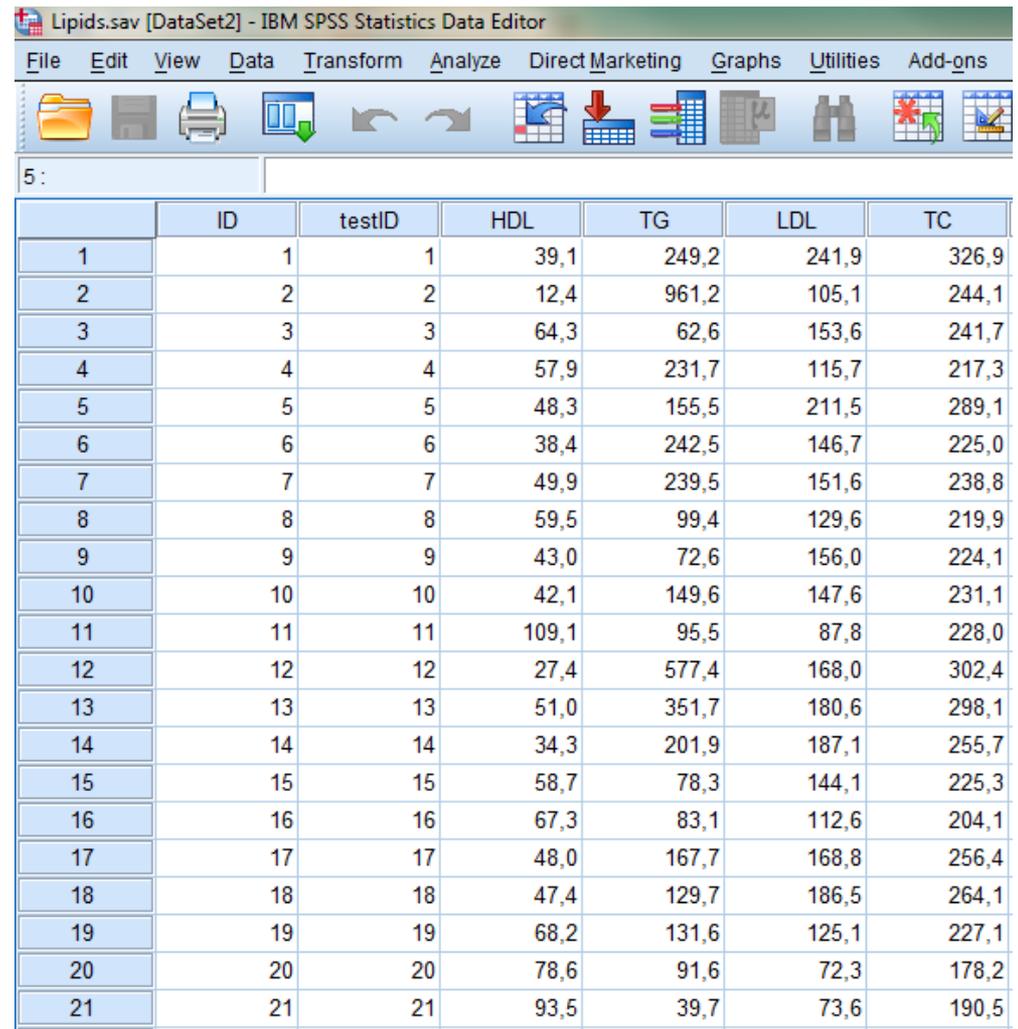


Merge two datasets

→ Create an additional variable in the Lipids.sav (testID)



The screenshot shows the IBM SPSS Statistics Data Editor window for the file *Lipids.sav [DataSet2]. The menu bar includes File, Edit, View, Data, Transform, Analyze, and Direct Marketing. The toolbar contains icons for file operations and data manipulation. The data grid shows a variable named 'HDL' with a value of 39,1 for the first row. A context menu is open over the first row, with the 'Insert Variable' option highlighted.



The screenshot shows the IBM SPSS Statistics Data Editor window for the file Lipids.sav [DataSet2]. The menu bar includes File, Edit, View, Data, Transform, Analyze, Direct Marketing, Graphs, Utilities, and Add-ons. The data grid displays 21 rows of data with the following columns: ID, testID, HDL, TG, LDL, and TC.

	ID	testID	HDL	TG	LDL	TC
1	1	1	39,1	249,2	241,9	326,9
2	2	2	12,4	961,2	105,1	244,1
3	3	3	64,3	62,6	153,6	241,7
4	4	4	57,9	231,7	115,7	217,3
5	5	5	48,3	155,5	211,5	289,1
6	6	6	38,4	242,5	146,7	225,0
7	7	7	49,9	239,5	151,6	238,8
8	8	8	59,5	99,4	129,6	219,9
9	9	9	43,0	72,6	156,0	224,1
10	10	10	42,1	149,6	147,6	231,1
11	11	11	109,1	95,5	87,8	228,0
12	12	12	27,4	577,4	168,0	302,4
13	13	13	51,0	351,7	180,6	298,1
14	14	14	34,3	201,9	187,1	255,7
15	15	15	58,7	78,3	144,1	225,3
16	16	16	67,3	83,1	112,6	204,1
17	17	17	48,0	167,7	168,8	256,4
18	18	18	47,4	129,7	186,5	264,1
19	19	19	68,2	131,6	125,1	227,1
20	20	20	78,6	91,6	72,3	178,2
21	21	21	93,5	39,7	73,6	190,5

Merge two datasets

If you have sorted both datasets by the key variable:

→ Go into the Basicdata.sav and merge the additional dataset Lipids.sav

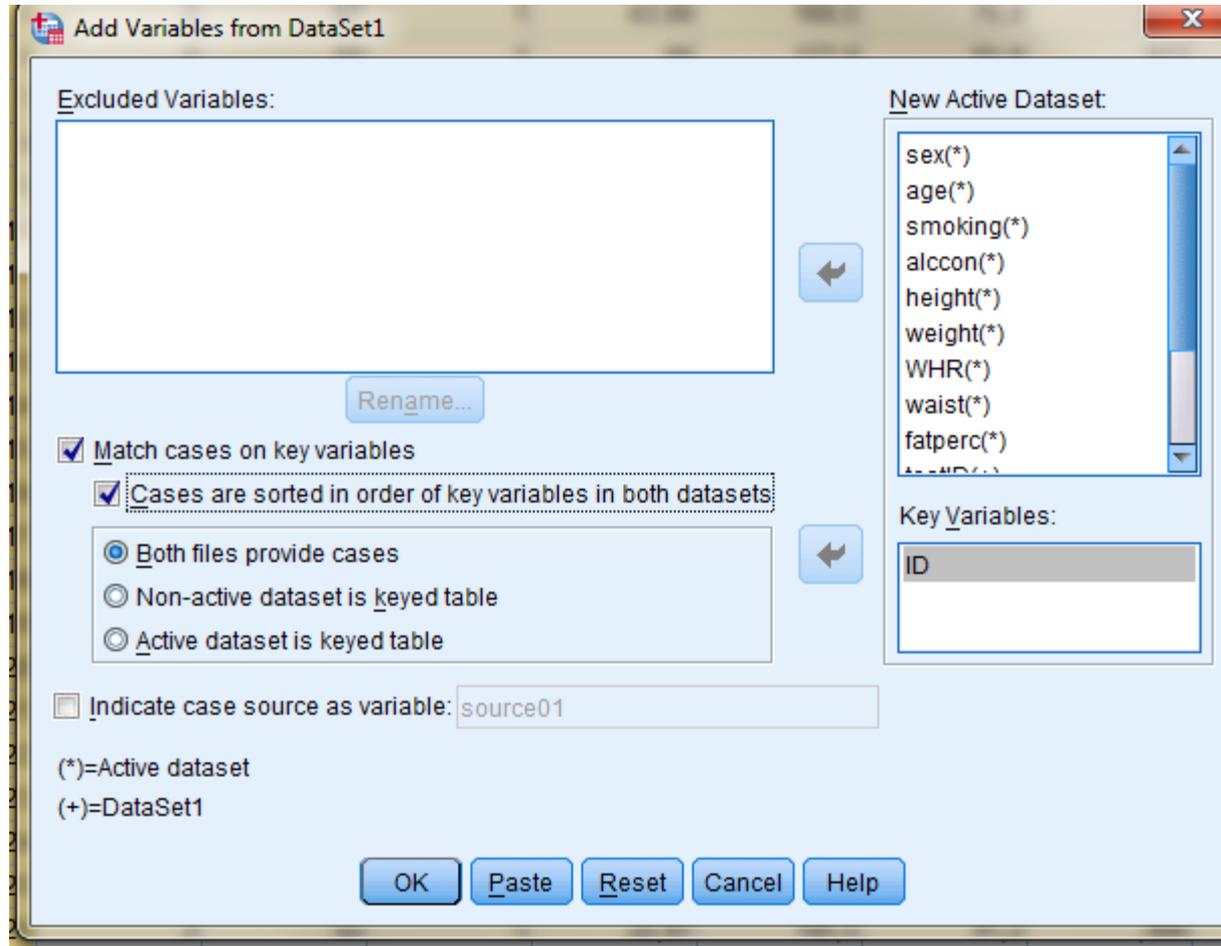
The screenshot shows the SPSS software interface. The 'Data' menu is open, and the 'Merge Files' option is selected. The 'Add Variables to Basicdata.sav' dialog box is open, showing the 'An open dataset' option selected. The list of open datasets includes 'Lipids.sav[DataSet2]'. The dialog box also includes a 'Browse...' button and a note: 'Non-SPSS Statistics data files must be opened in SPSS Statistics before they can be used as part of a merge.' The 'Continue', 'Cancel', and 'Help' buttons are visible at the bottom of the dialog box.

	alcon	height
3	20,02	16
2	14,31	15
3	14,28	
3	,43	
1	43,00	
1	,00	
2	,00	
2	48,58	
3	,00	
2	,00	
2	1,75	

Merge two datasets

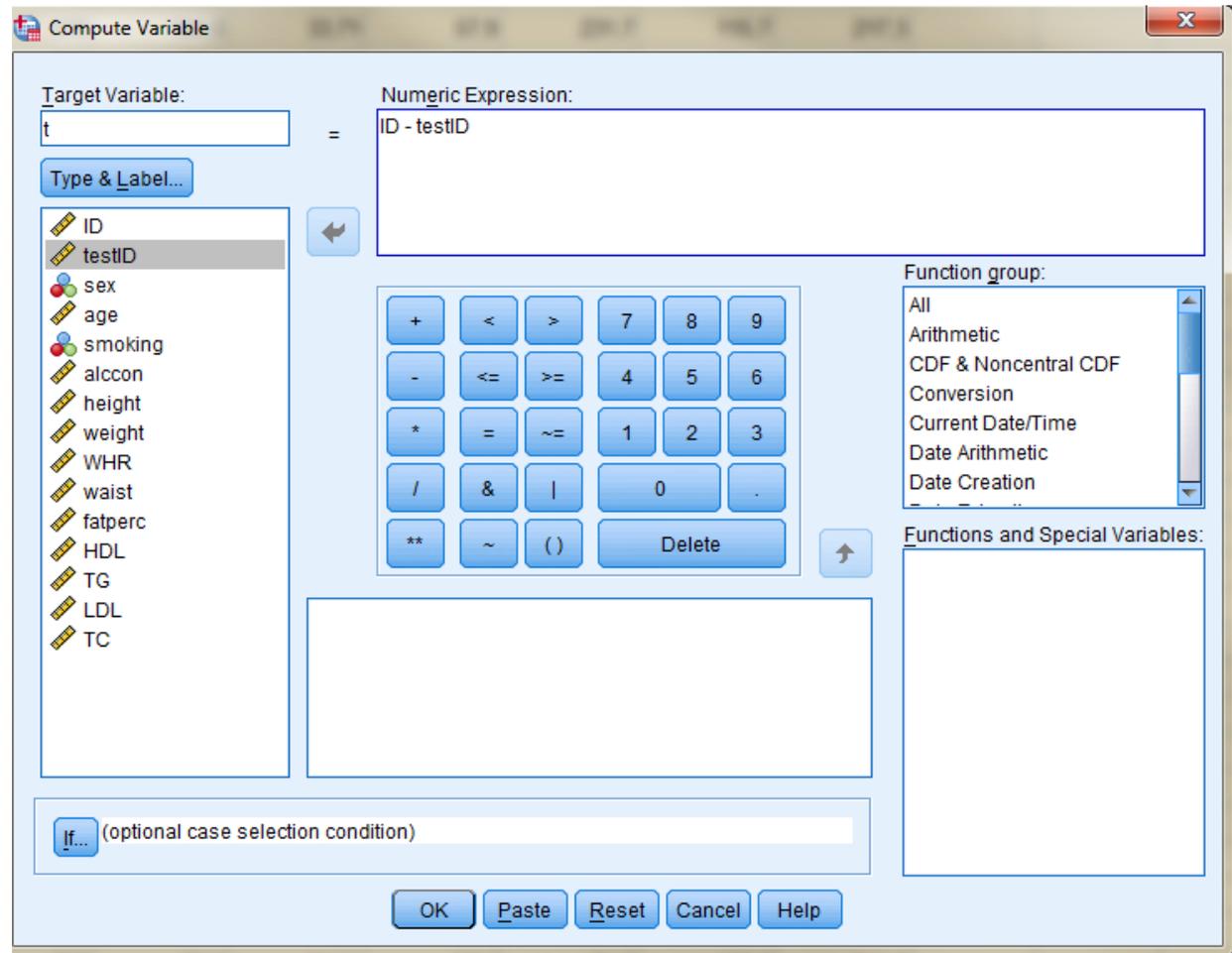
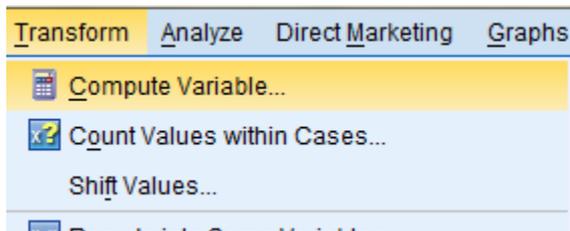
If you have sorted both datasets by the key variable:

→ Go into the Basicdata.sav and merge the additional dataset Lipids.sav



Merge two datasets

→ Create a variable „t“ and subtract testID from ID to check if the merge process was correct → t should be „0“ for all patients



Merge two datasets

Alldata.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

1 : testID 1

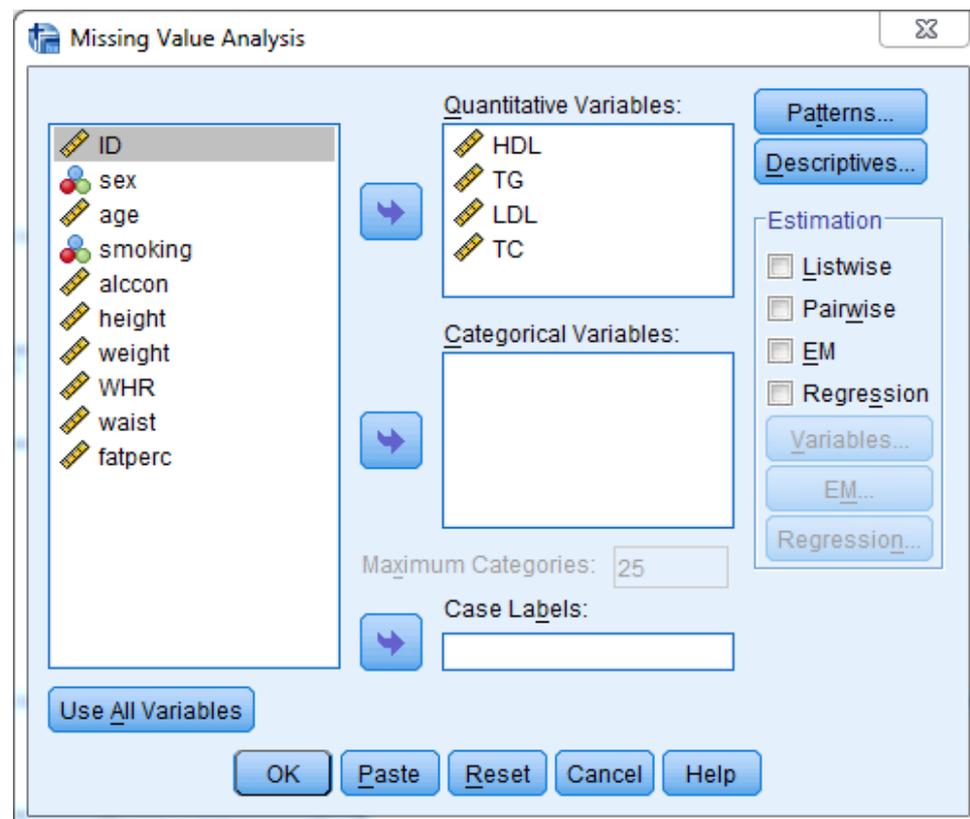
	ID	testID	t	sex	age	smoking	alcon	height	weight	WHR	waist	fatperc	HDL	TG	LDL	TC
1	1	1	,00	1	62	3	20,02	162,5	82,2	,918	104,0	.	39,1	249,2	241,9	326,9
2	2	2	,00	2	48	2	14,31	180,5	133,5	1,108	132,0	38,68	12,4	961,2	105,1	244,1
3	3	3	,00	1	54	3	14,28	164,5	62,3	,702	71,5	30,64	64,3	62,6	153,6	241,7
4	4	4	,00	2	50	3	,43	172,5	91,4	,978	104,5	33,71	57,9	231,7	115,7	217,3
5	5	5	,00	2	67	1	43,00	169,5	76,8	,960	95,5	33,03	48,3	155,5	211,5	289,1
6	6	6	,00	2	52	1	,00	177,0	91,8	,913	102,0	29,80	38,4	242,5	146,7	225,0
7	7	7	,00	2	34	2	,00	178,5	75,9	,869	82,5	20,03	49,9	239,5	151,6	238,8
8	8	8	,00	2	52	2	48,58	184,5	86,8	,908	98,0	26,65	59,5	99,4	129,6	219,9
9	9	9	,00	2	58	3	,00	166,5	61,3	,838	80,0	23,69	43,0	72,6	156,0	224,1
10	10	10	,00	1	46	2	,00	162,5	96,5	1,020	116,0	45,35	42,1	149,6	147,6	231,1
11	11	11	,00	1	37	2	1,75	162,0	58,5	,754	69,5	29,93	109,1	95,5	87,8	228,0
12	12	12	,00	2	64	2	51,45	168,5	93,5	1,026	115,0	36,06	27,4	577,4	168,0	302,4
13	13	13	,00	2	63	2	99,99	164,0	71,8	,917	90,0	29,28	51,0	351,7	180,6	298,1
14	14	14	,00	1	50	2	,00	160,5	73,6	,849	90,0	39,05	34,3	201,9	187,1	255,7
15	15	15	,00	2	65	3	15,86	158,5	62,9	,842	80,0	31,30	58,7	78,3	144,1	225,3
16	16	16	,00	2	63	3	20,02	173,5	80,7	,933	97,0	.	67,3	83,1	112,6	204,1
17	17	17	,00	2	55	1	83,15	173,5	80,9	,970	98,0	.	48,0	167,7	168,8	256,4
18	18	18	,00	1	56	3	,00	162,0	60,5	,767	69,0	32,19	47,4	129,7	186,5	264,1
19	19	19	,00	1	53	3	,00	163,5	55,0	,707	66,5	27,12	68,2	131,6	125,1	227,1
20	20	20	,00	1	47	1	9,94	171,0	87,0	,834	91,0	39,24	78,6	91,6	72,3	178,2
21	21	21	,00	1	31	3	2,86	169,0	54,0	,826	67,5	22,31	93,5	39,7	73,6	190,5
22	22	22	,00	1	38	3	,00	163,0	82,9	,758	84,0	.	64,1	107,5	134,0	241,5

Save the merged data set → „Alldata.sav“

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: Summarizing & 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

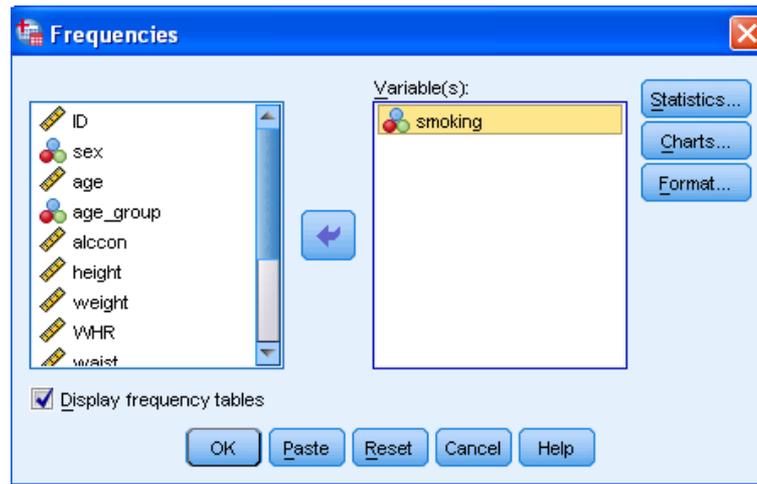
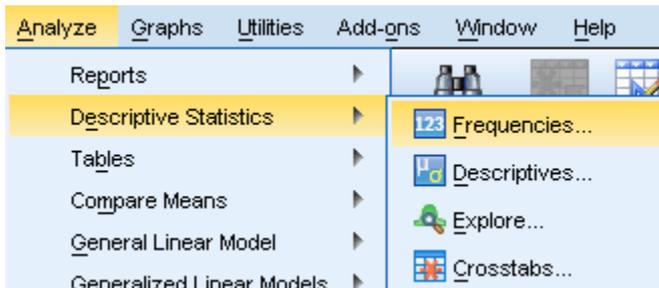
Descriptive Statistics

Example: A population-based study (n=1457) with the aim to identify atherosclerotic risk factors

Univariate methods:

For qualitative data or grouped quantitative data:

Simple tables for one variable



Descriptive Statistics

Example: A population-based study (n=1457) with the aim to identify atherosclerotic risk factors

Univariate methods:

For qualitative data or grouped quantitative data:

Simple tables for one variable

Sums up to 100%

		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	

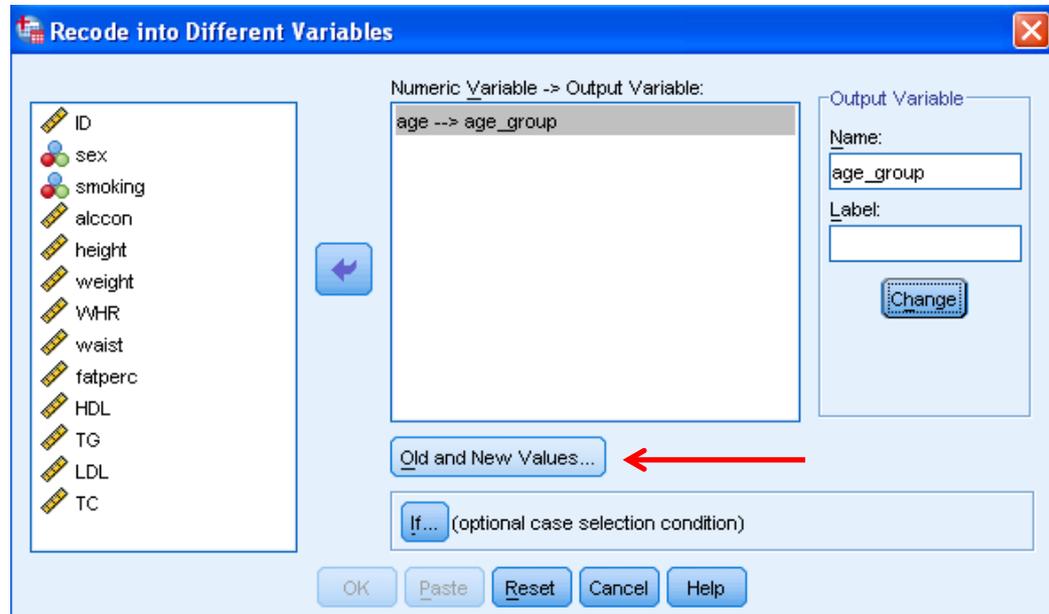
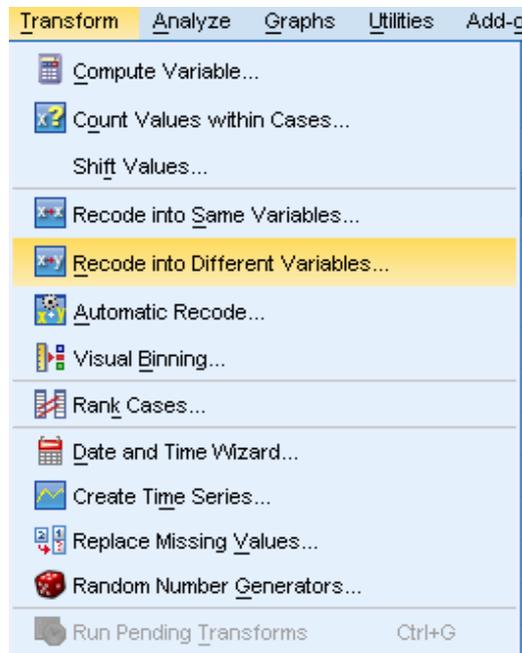


Descriptive Statistics

Example: A population-based study (n=1457) with the aim to identify atherosclerotic risk factors

How is the age distribution of study participants?

- First, create the variable `age_group` out of the variable `age`: 1=30-40, 2=41-50, 3=51-60, 4=61-70

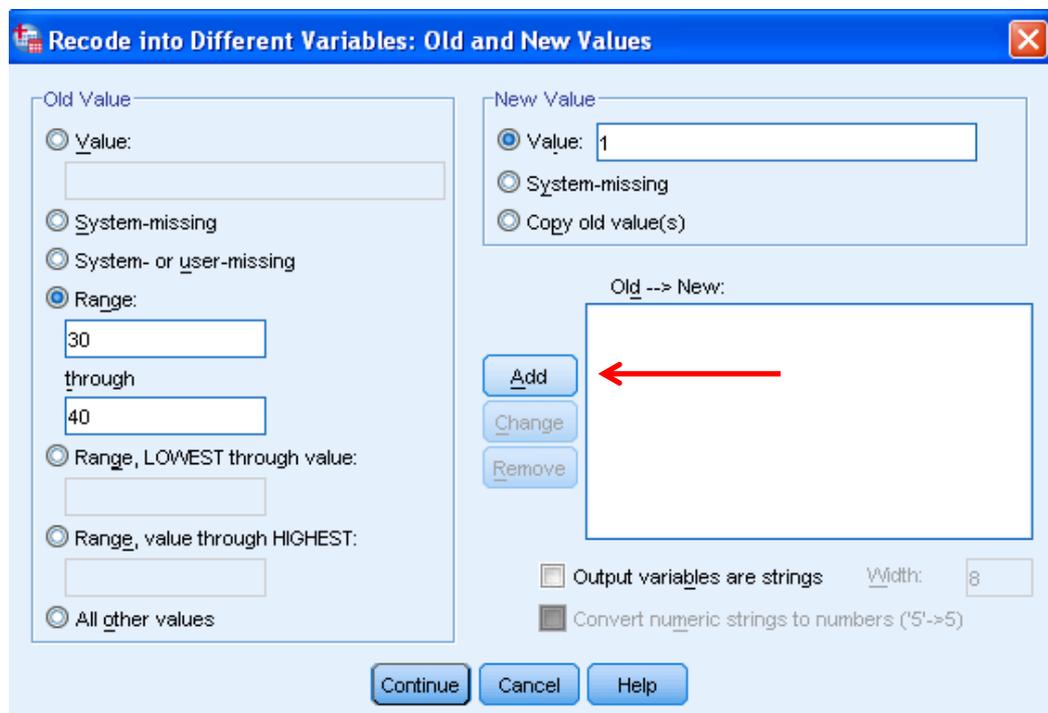


Descriptive Statistics

Example: A population-based study (n=1457) with the aim to identify atherosclerotic risk factors

How is the age distribution of study participants?

- First, create the variable age_group out of the variable age: 1=30-40, 2=41-50, 3=51-60, 4=61-70



The image shows the 'Recode into Different Variables: Old and New Values' dialog box in SPSS. The 'Old Value' section has 'Range' selected, with '30' in the 'through' field. The 'New Value' section has 'Value' selected, with '1' in the input field. A red arrow points to the 'Add' button. The 'Output variables are strings' checkbox is checked, and the 'Width' is set to 8. The 'Convert numeric strings to numbers' checkbox is unchecked. The 'Continue', 'Cancel', and 'Help' buttons are at the bottom.

Recode into Different Variables: Old and New Values

Old Value

Value:

System-missing

System- or user-missing

Range:

30

through

40

Range, LOWEST through value:

Range, value through HIGHEST:

All other values

New Value

Value: 1

System-missing

Copy old value(s)

Old --> New:

Add

Change

Remove

Output variables are strings Width: 8

Convert numeric strings to numbers ('5'->5)

Continue Cancel Help

Descriptive Statistics

Example: A population-based study (n=1457) with the aim to identify atherosclerotic risk factors

How is the age distribution of study participants?

- First, create the variable `age_group` out of the variable `age`: 1=30-40, 2=41-50, 3=51-60, 4=61-70
- Label the variable `age_group` (Variable view in SPSS → Values) : 1= 30-40, 2= 41-50, 3= 51-60, 4=61-70

		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:

- How to create a bar plot in SPSS

Absolute frequencies

Or relative frequencies (select percentage here)

The image shows the SPSS Chart Builder and Element Properties dialog boxes. The Chart Builder window displays a bar chart with three bars representing age groups: 30-40, 41-50, and [More...]. The Y-axis is labeled 'Count'. The Element Properties dialog box is open, showing the 'Statistics' section where 'Count' is selected as the statistic. The 'Bar Style' section shows 'Bar' selected. Red arrows point from the text 'Absolute frequencies' to the 'Count' dropdown and from 'Or relative frequencies (select percentage here)' to the 'Statistics' section.

Chart Builder

Variables:

- ID
- testID
- t
- sex
- age
- smoking
- alcon
- height
- weight
- WHR
- waist

Chart preview uses example data

Count

30-40 41-50 [More...]

age_group

No categories (scale variable)

Gallery Basic Elements Groups/Point ID Titles/Footnotes

Choose from:

- Favorites
- Bar
- Line
- Area
- Pie/Polar
- Scatter/Dot
- Histogram
- High-Low
- Boxplot
- Dual Axes

Element Properties

Edit Properties of:

- Bar1
- X-Axis1 (Bar1)
- Y-Axis1 (Bar1)

Statistics

Variable:

Statistic: Count

Set Parameters...

Display error bars

Error Bars Represent

- Confidence intervals
- Level (%): 95
- Standard error
- Multiplier: 2
- Standard deviation
- Multiplier: 2

Bar Style:

Bar

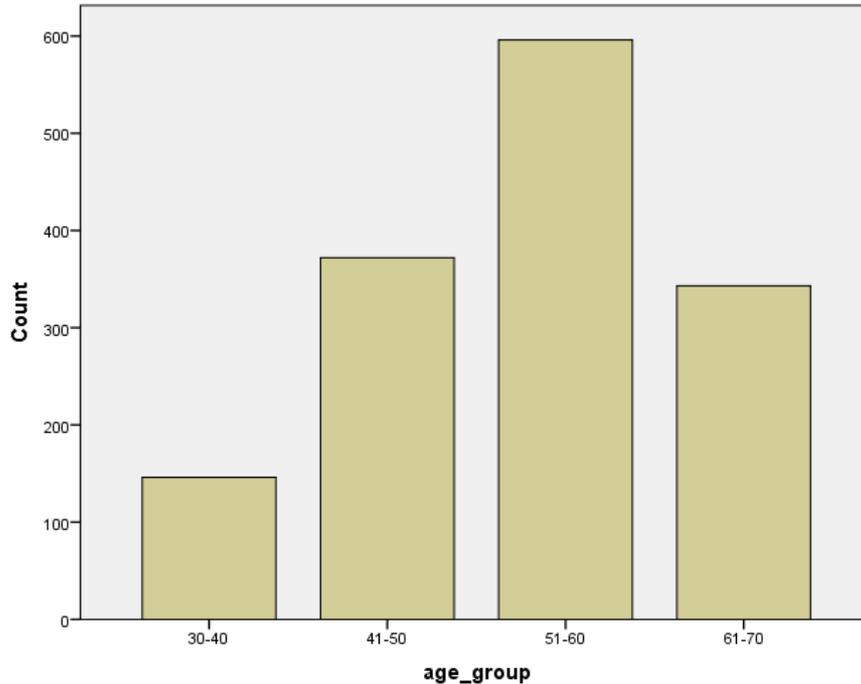
OK Paste Reset Cancel Help

Apply Close Help

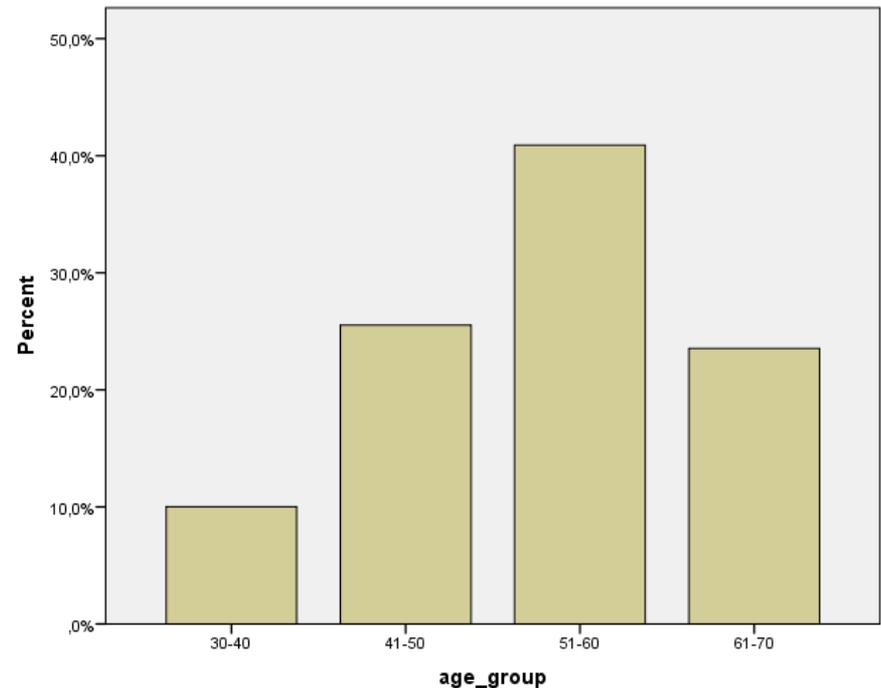
Descriptive Statistics

Barplots for illustrating tables:

Using absolute frequencies:



Using relative frequencies



Barplot: Graphical illustration of a categorical variable

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

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

$x_1, x_2, \dots, x_i, \dots, x_n, i=1, \dots, n$

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

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

The mean is invariant towards linear transformations:

Example:

In one European study, measurements of cholesterol levels are given in mg/dl, your US-American colleagues, however, want to see it in mmol/l :

Measurements in mg/dl * 0.0259 → Measurements in mmol/l

Mean (in mg/dl) = 225

→ Mean (in mmol/l) = $225 * 0.0259 = 5.8275$

It is not invariant against non-linear transformations,

e.g.: $\text{mean}(\log(\text{Cholesterol})) \neq \log(\text{mean}(\text{Cholesterol}))$

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 $1-\alpha$ 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:

$65, 66, 69, 70, 72, 75, 78 \rightarrow \text{Median} = 70$

50% of the data are lower,

50 % are higher than the median

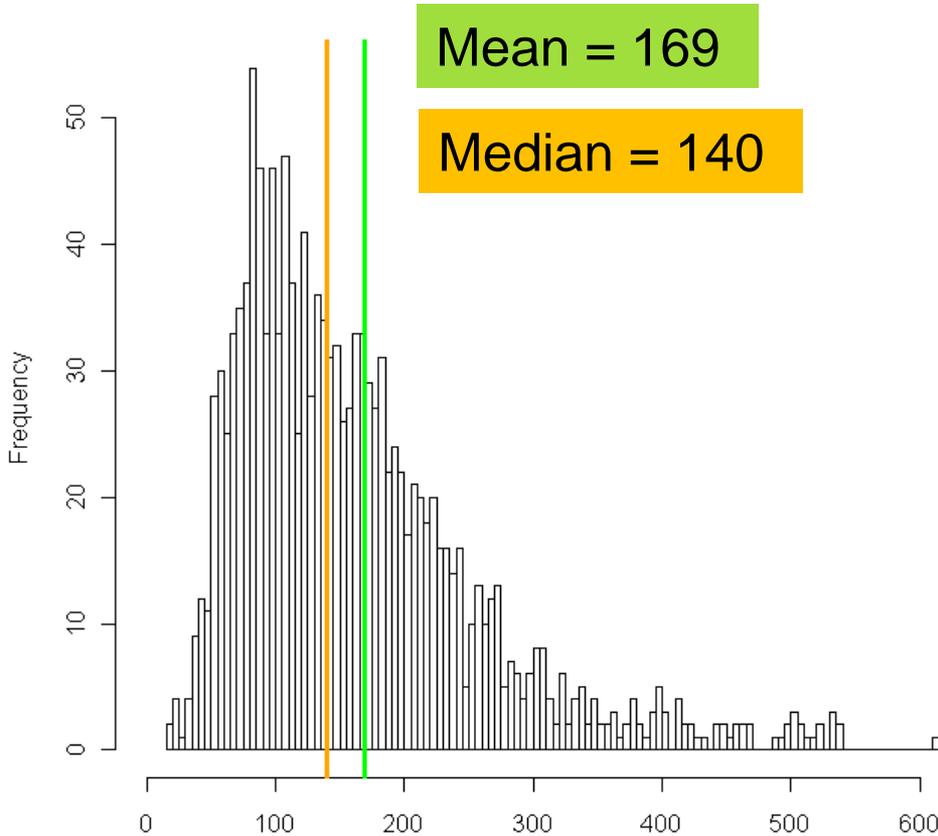
Including one young patient into this group:

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

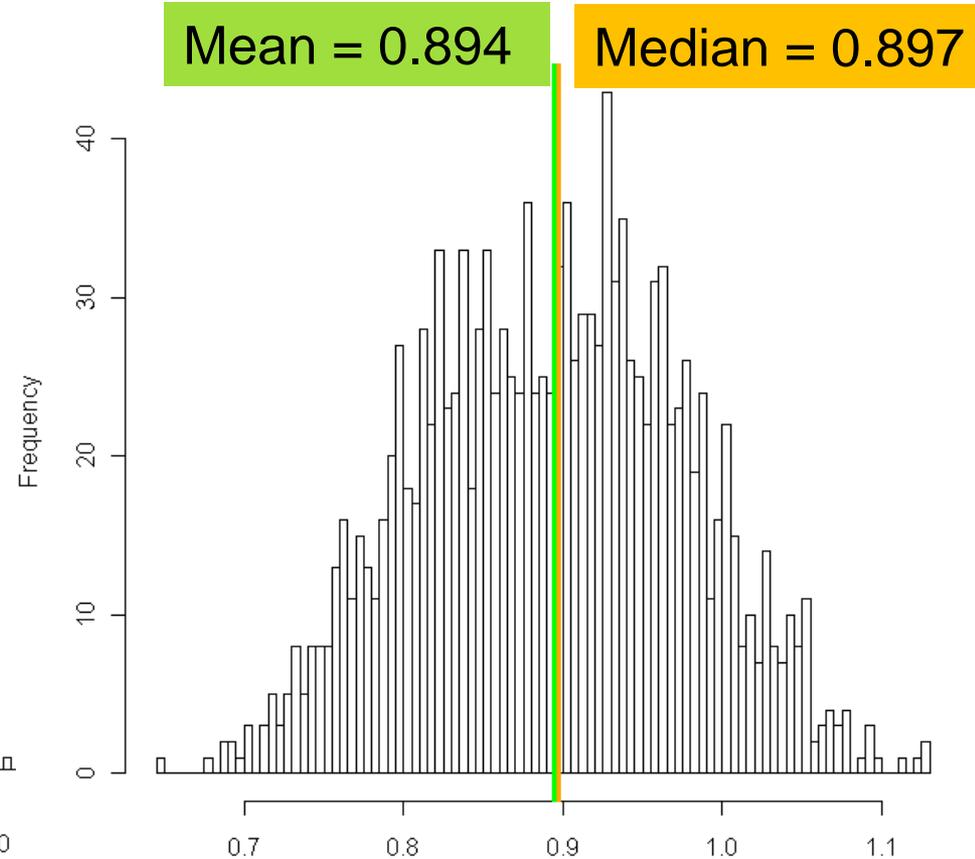
50% of the data are lower, 50 % are higher than the median

Descriptive Statistics

Right skewed distribution:



Symmetrically distributed:



For symmetrically distributed variables → Mean=Median

For skewed variables: Medians should be preferred

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:

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

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

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

Range = Maximum-Minimum

Interquartile range = 0.75-Quantile – 0.25-Quantile

Descriptive Statistics

Example:

Ages of Myocardial Infarction patient group: 65, 66, 69, 70, 72, 75, 78

$$\text{Sample variance} = S^2 = \frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$= [(65-70.71)^2 + (66-70.71)^2 + (69-70.71)^2 + (70-70.71)^2 + (72-70.71)^2 + (75-70.71)^2 + (78-70.71)^2] * 1/6$$
$$= 21.90$$

$$\text{Sample standard deviation} = \sqrt{21.90} = 4.68$$

- The standard-deviation is invariant against linear transformations

Descriptive Statistics

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

The image shows the SPSS 'Explore' dialog box and the 'Explore' window. In the dialog box, 'Descriptives' is checked, and the 'Confidence Interval for Mean' is set to 95%. In the 'Explore' window, 'age' is listed as the dependent variable, and the 'Statistics' radio button is selected under the 'Display' section.

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

Median

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		

Descriptive Statistics

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

How are such summary data typically presented in scientific publications?

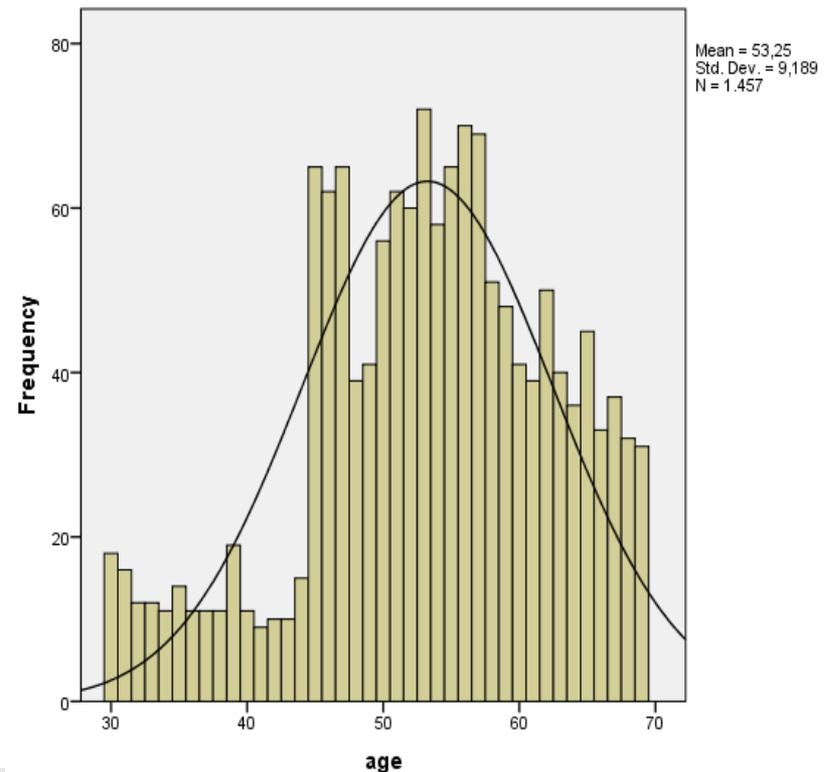
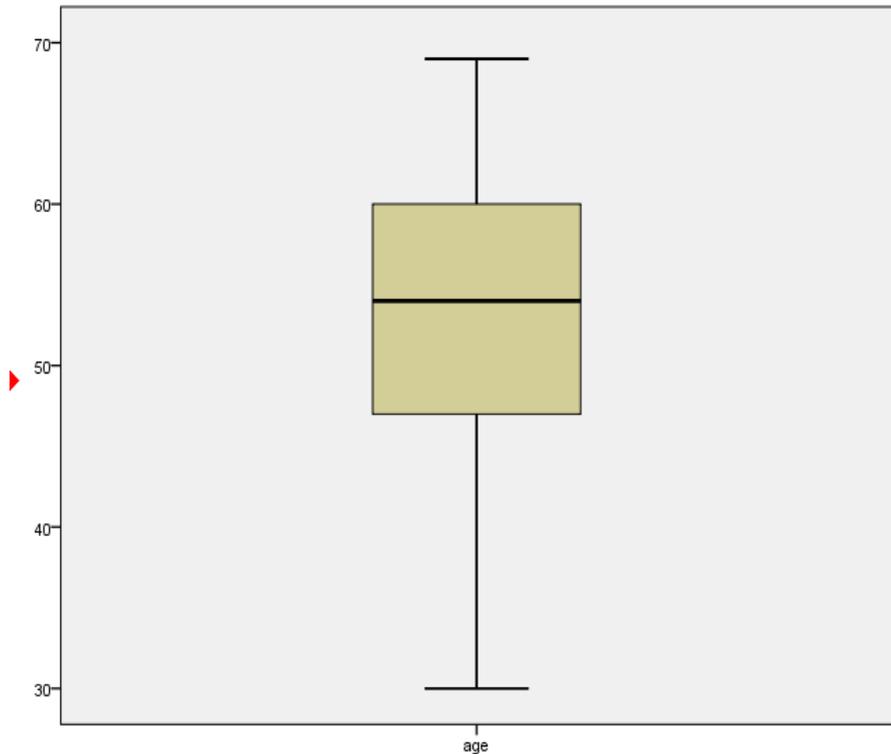
Characteristics of all participants (n=1457)	
Mean \pm SD [25., 50.; 75. percentile for non-normal distribution] or number (%)	
Age (years)	53.25 \pm 9.19
Sex (male/female), n (%)	722 (49.6) / 735 (50.4)
Smoking Status, n (%)	
Current Smoker	261 (17.9)
Ex-Smoker	453 (31.1)
Never Smoker	743 (51.0)
Measured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

Descriptive Statistics

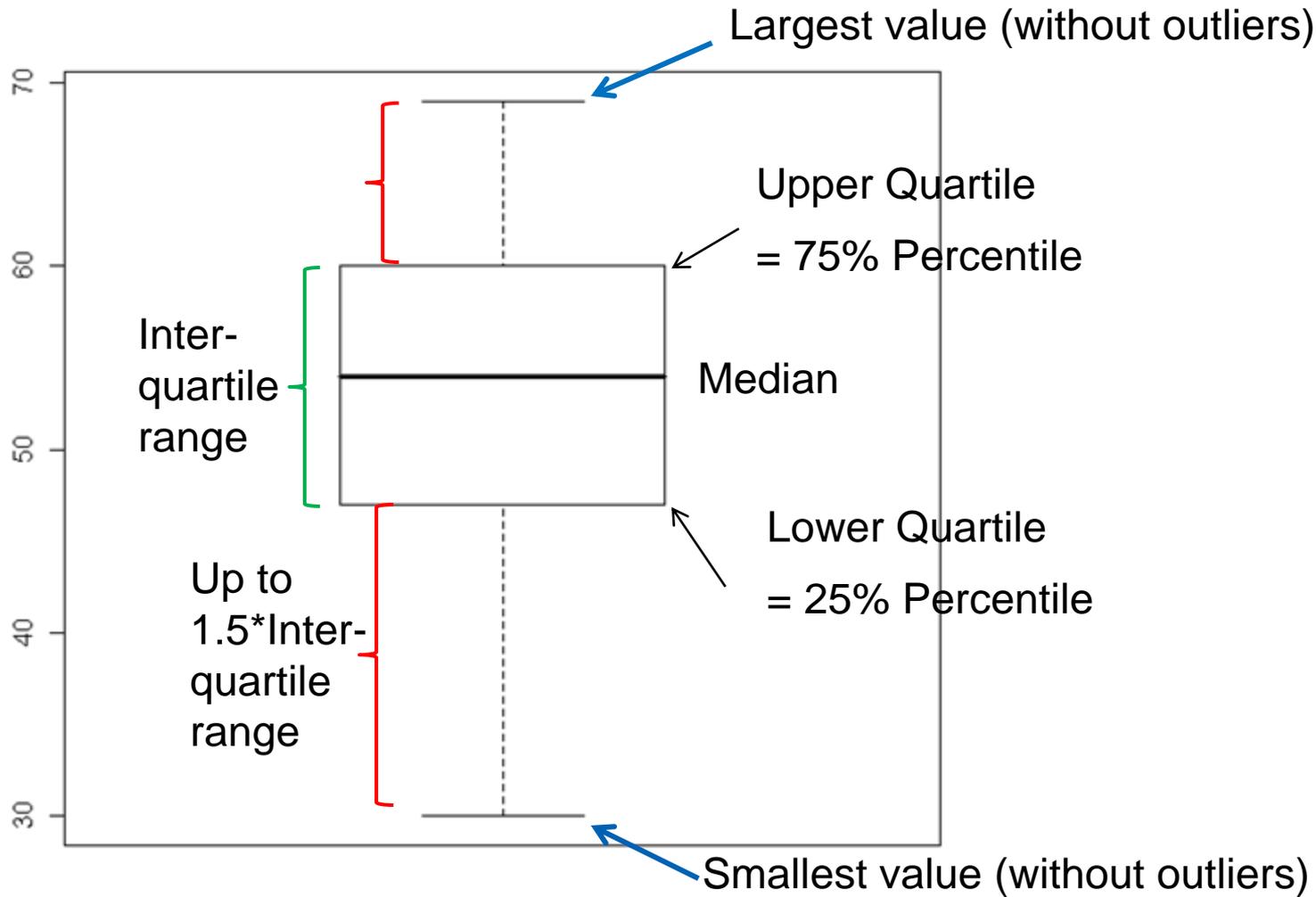
Point estimates are not sufficient to illustrate the complete distribution of a variable → Use Figures !

Boxplot

Histogram



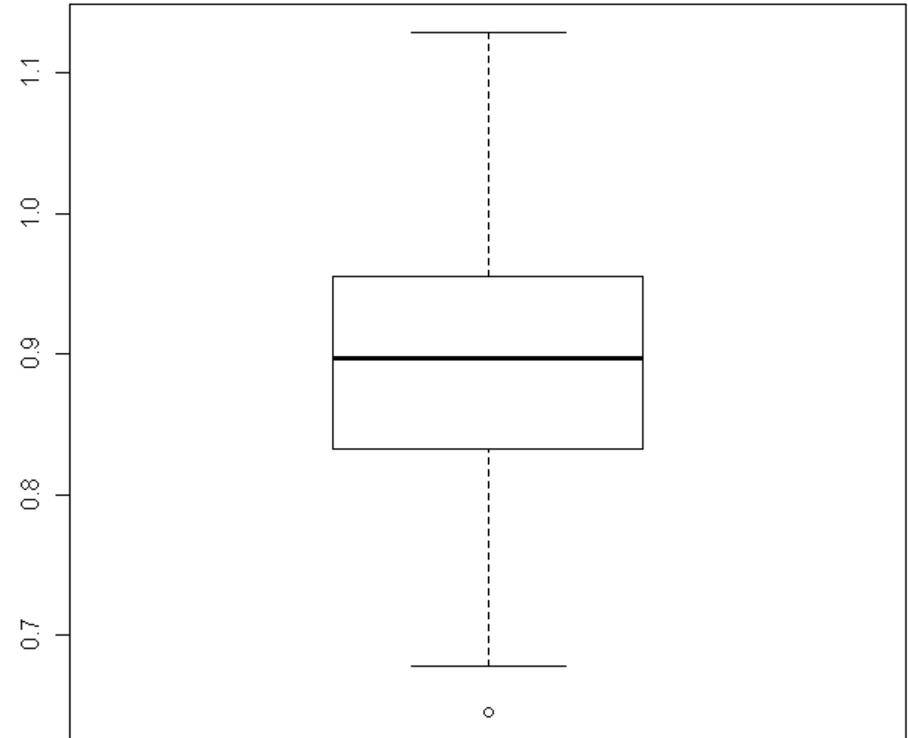
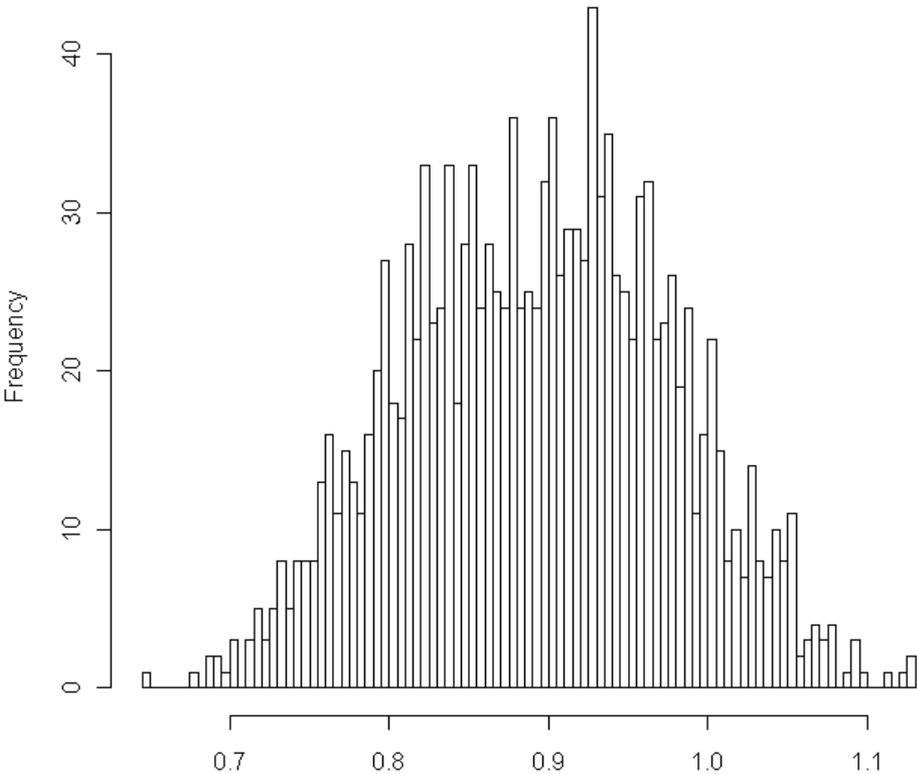
Descriptive Statistics



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

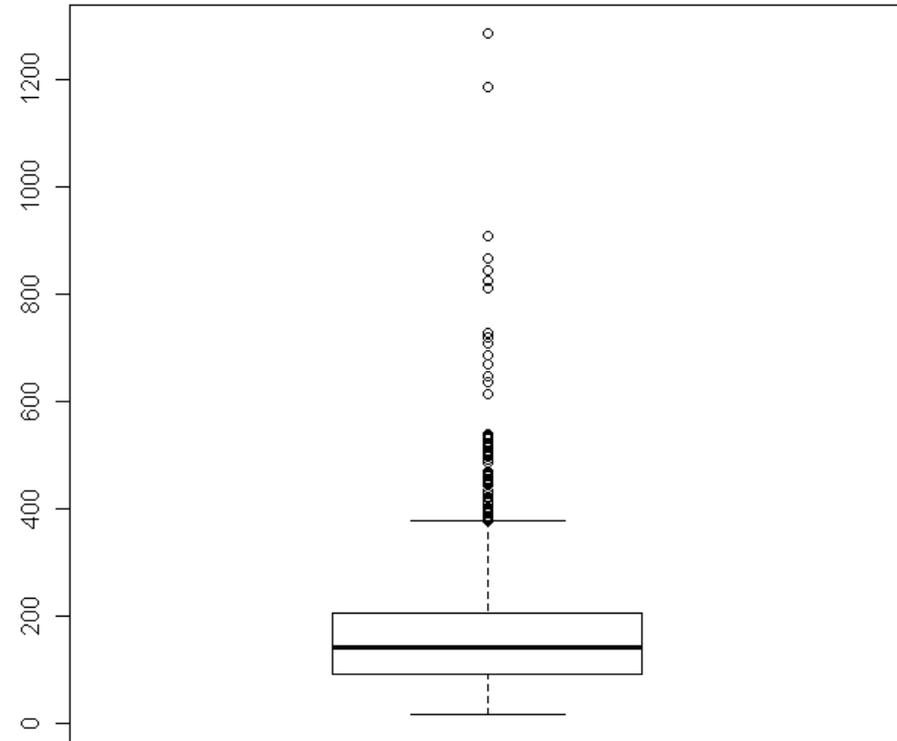
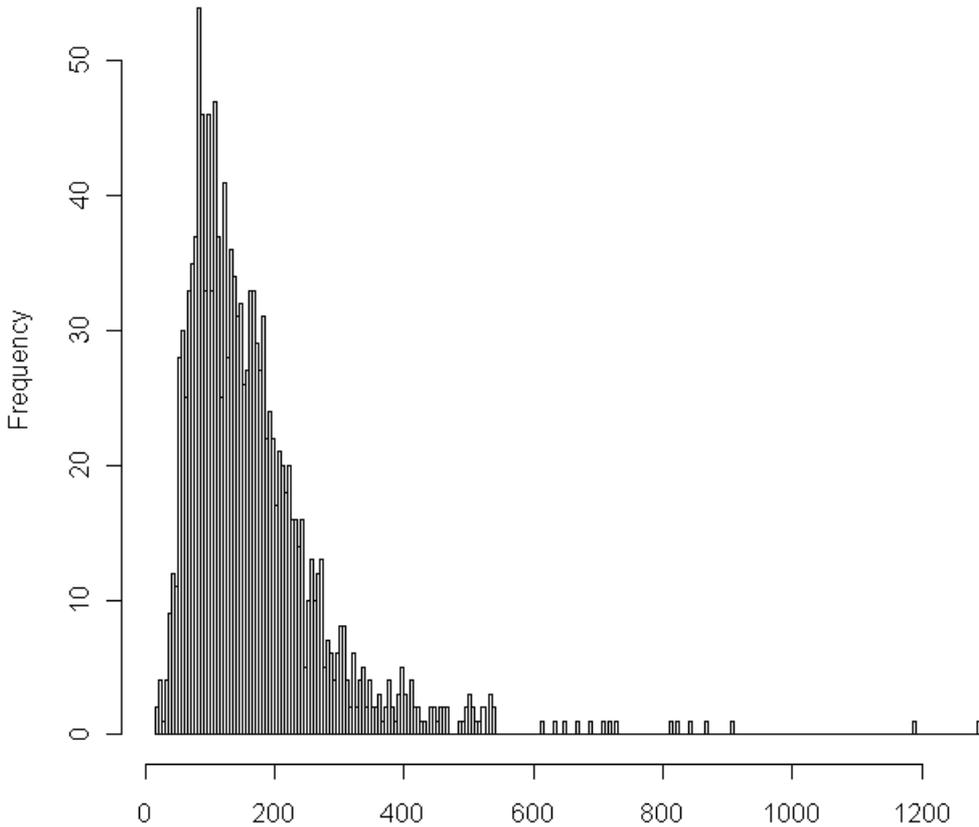
Descriptive Statistics

Histogram and Boxplot for a symmetrically distributed variable:



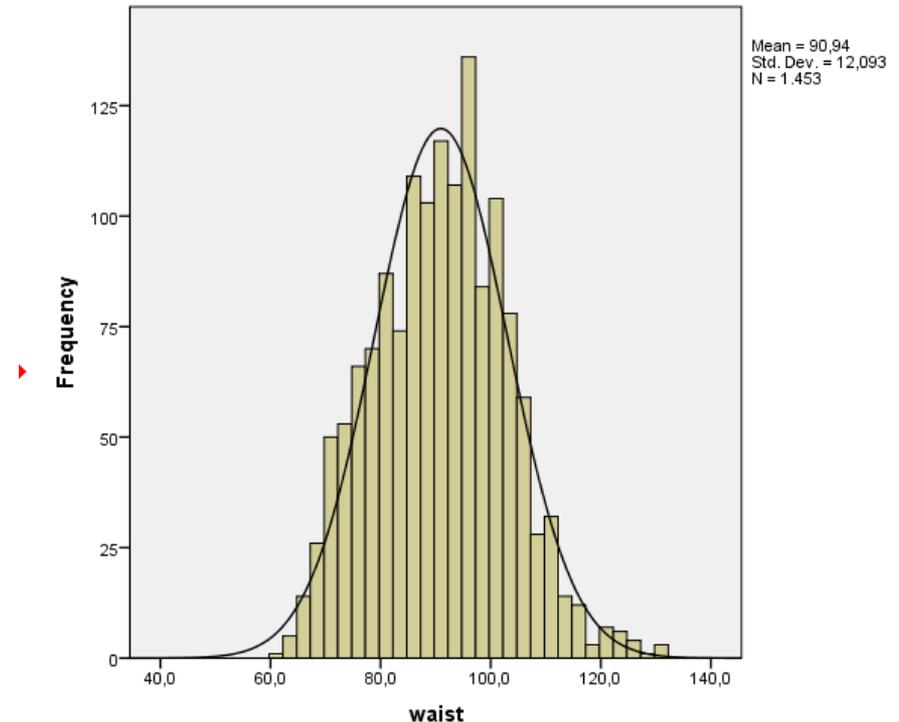
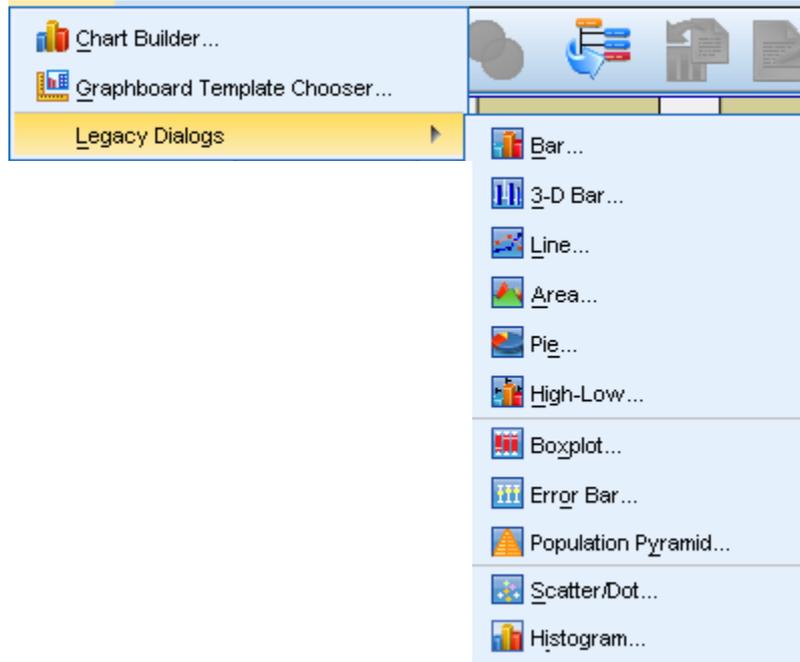
Descriptive Statistics

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



Descriptive Statistics

How to create such a plot (=Histogram) in SPSS: e.g. for the variable waist



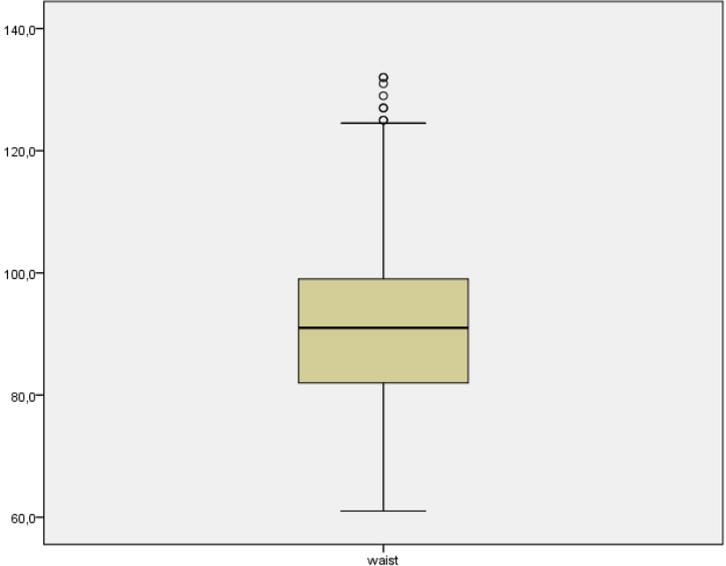
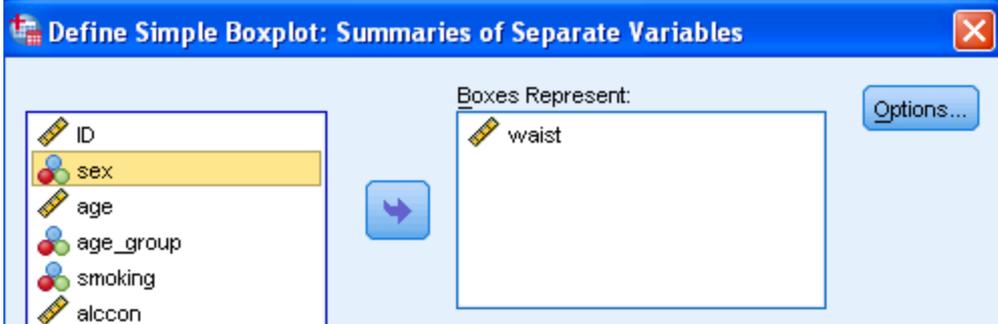
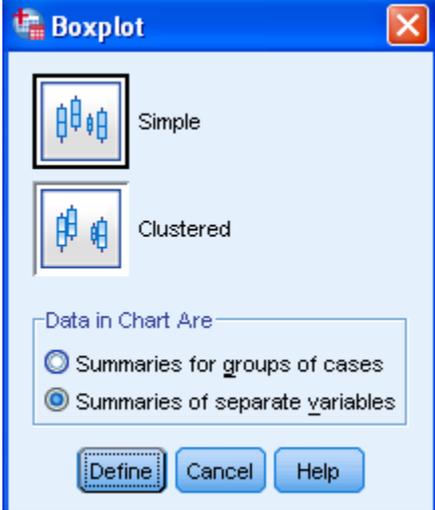
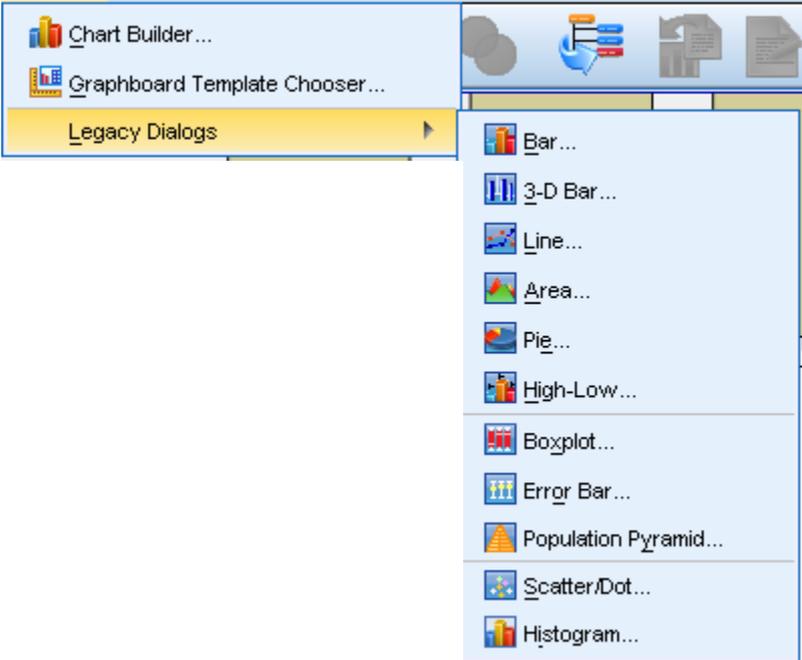
Histogram:

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

How to create a boxplot in SPSS: e.g. for the variable waist



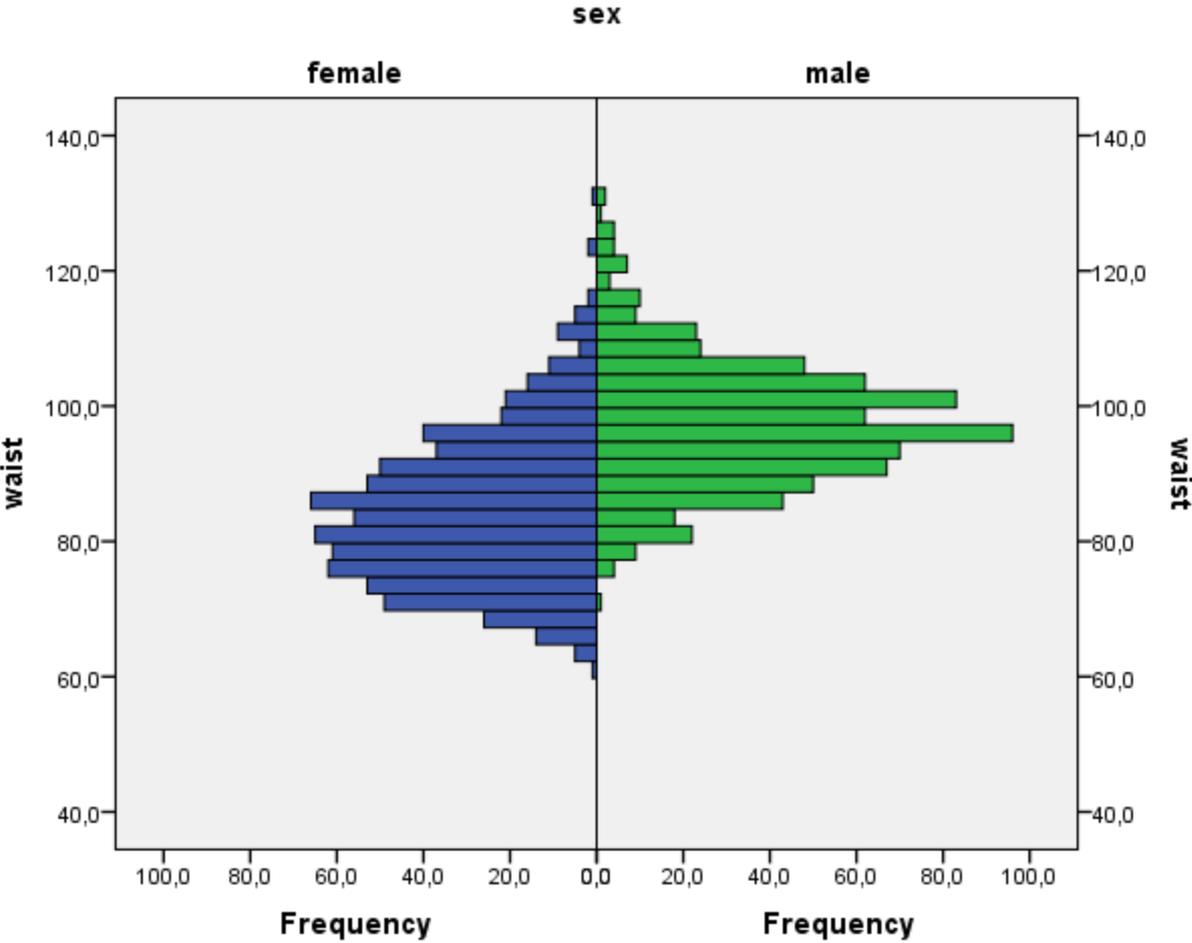
Descriptive Statistics

Histogram for waist separated by a grouping variable e.g.sex

The image shows two windows from the SPSS software interface. The 'Chart Builder' window on the left is used for creating charts. It features a 'Variables:' list on the left containing 'ID', 'testID', 't', 'sex', 'age', 'smoking', 'alcon', 'height', 'weight', 'WHR', and 'waist'. The 'sex' variable is placed in the 'Grouping Variable' box at the top, and 'waist' is placed in the 'Y-Axis' box. A preview of a horizontal bar chart is shown, with blue bars for one sex and green bars for the other. Below the variables list, there are tabs for 'Gallery', 'Basic Elements', 'Groups/Point ID', and 'Titles/Footnotes'. The 'Basic Elements' tab is active, showing a 'Choose from:' list with 'Histogram' selected. A 'Basic chart building blocks' section shows four icons: a bar chart, a stacked bar chart, a line graph, and a histogram. At the bottom of the Chart Builder are buttons for 'OK', 'Paste', 'Reset', 'Cancel', and 'Help'. The 'Element Properties' window on the right is titled 'Edit Properties of: Pyramid1'. It lists 'X-Axis1 (Pyramid1)', 'Y-Axis1 (Pyramid1)', and 'Split (Pyramid1)'. Under 'Scale Variable Distribution Options', there is a checkbox for 'Display normal curve' which is unchecked. The 'Anchor First Bin' section has 'Automatic' selected. The 'Bin Sizes' section has 'Automatic' selected. The 'Categorical Variable Distribution Options' section has 'Display error bars' checked and 'Confidence intervals level (%)' set to '95'. At the bottom of this window are buttons for 'Apply', 'Close', and 'Help'.

Descriptive Statistics

Histogram for waist seperated by a grouping variable e.g.sex



Descriptive Statistics

Boxplot for waist seperated by a grouping variable e.g.sex

The image shows two windows from the SPSS software interface. The left window is the 'Chart Builder' dialog, and the right window is the 'Element Properties' dialog.

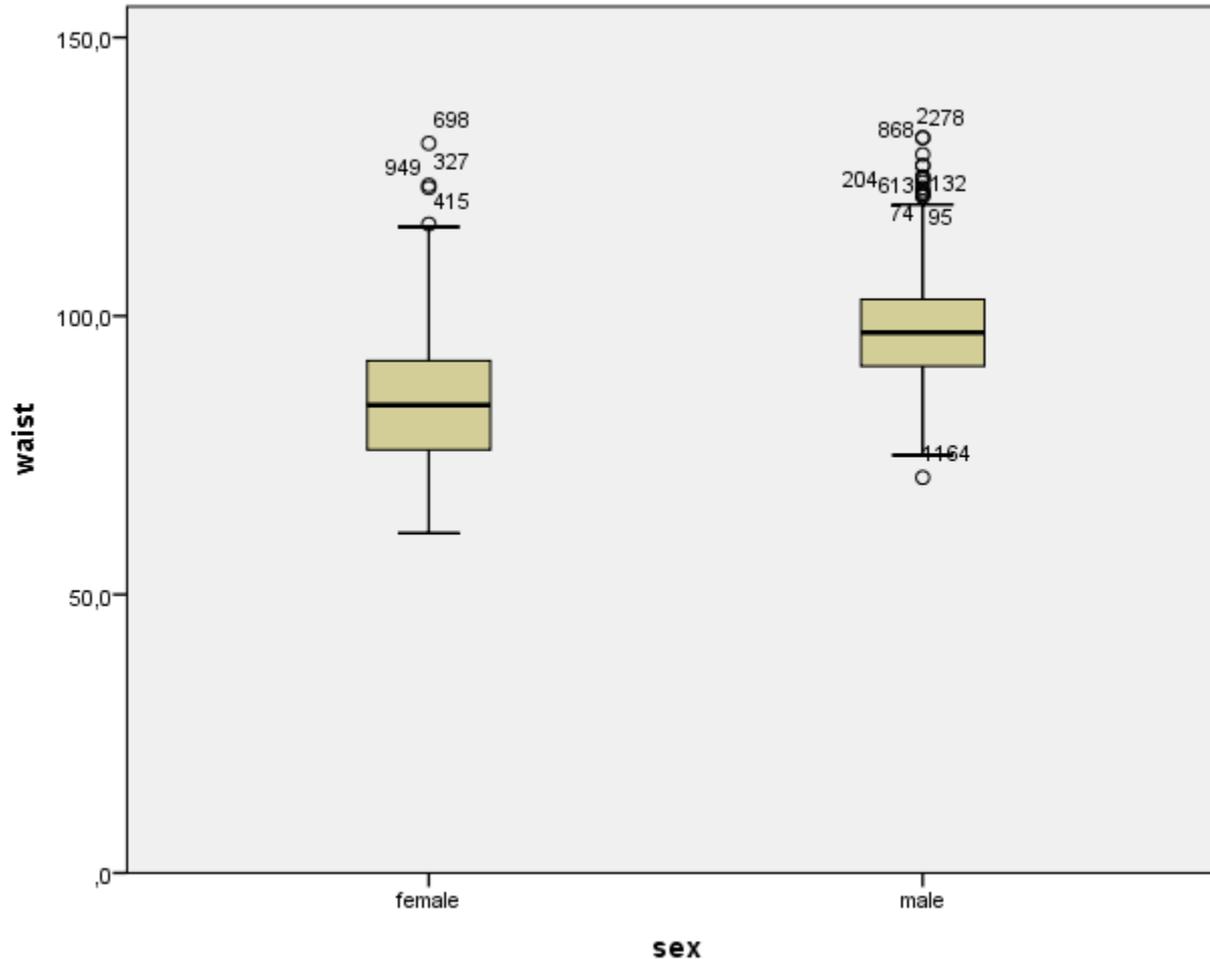
Chart Builder:

- Variables:** A list of variables is shown on the left, with 'waist' selected. Below the list, it says 'No categories (scale variable)'. The chart preview area shows a boxplot for 'waist' grouped by 'sex' (female and male).
- Gallery:** The 'Boxplot' option is selected under the 'Basic Elements' tab.
- Buttons:** 'OK', 'Paste', 'Reset', 'Cancel', and 'Help' are visible at the bottom.

Element Properties:

- Edit Properties of:** 'Box1' is selected.
- Statistics:** 'Variable: waist' and 'Statistic: Boxplot' are set.
- Display error bars:** This option is checked.
- Error Bars Represent:** 'Confidence intervals' is selected with a level of 95%. 'Standard error' and 'Standard deviation' are also listed with a multiplier of 2.
- Buttons:** 'Apply', 'Close', and 'Help' are visible at the bottom.

Descriptive Statistics



Such plots can be used to create hypotheses →

Waist circumference seems to differ between men and women

Descriptive 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):

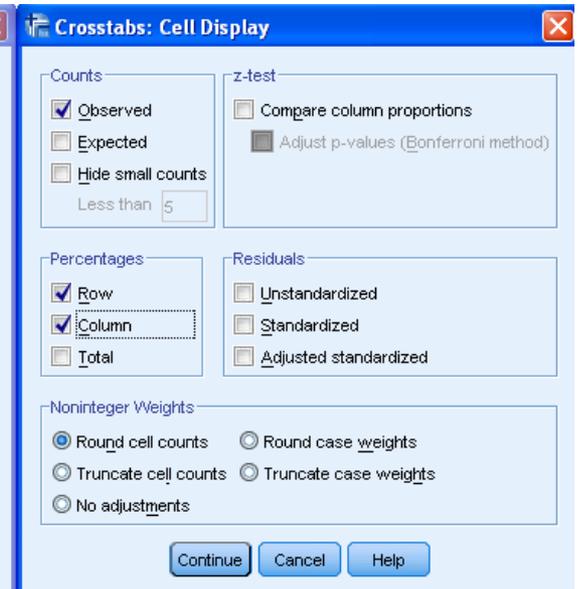
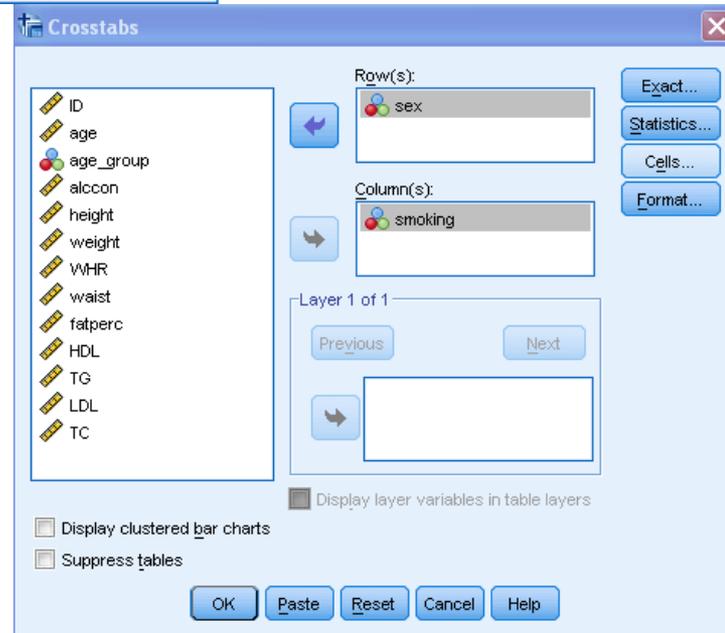
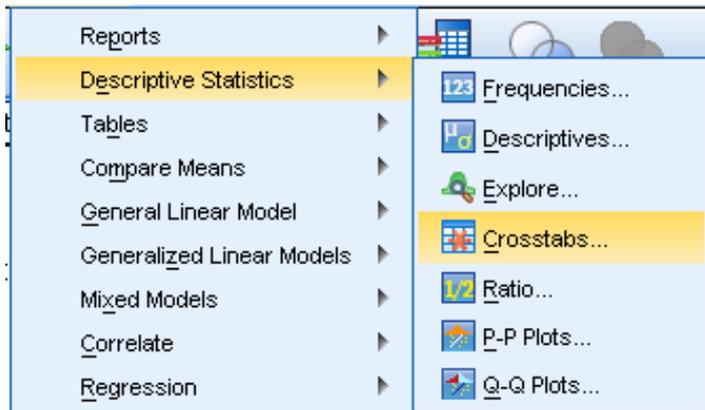
E.g. Comparison of measures of location of the quantitative variable between levels of the qualitative variable	Simple case: Does the mean of group 1 differ from the mean of group 2?
---	--

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

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

Descriptive Statistics

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



Descriptive Statistics

2 qualitative variables: Crosstable including absolute and relative frequencies

Example:

Gender
X
Smoking

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%	

Different numbers have a different emphasis and interpretation. Examples:

1. Altogether, there are 475 women in the study, which have never smoked (475/1457:32.6%).
1. 37.1% of all men have never smoked, but 64.6% women
2. 55.2% of all current smokers in the study are male

Descriptive Statistics

2-dimensional tables by barplots next to each other: e.g.: table on Gender x Smoking

The image shows two windows from the SPSS software interface. The left window is the 'Chart Builder' dialog, and the right window is the 'Element Properties' dialog.

Chart Builder:

- Variables:** A list of variables including ID, testID, t, sex, age, smoking, alcon, height, weight, WHR, waist, and fatness.
- Chart preview:** A bar chart showing the count of individuals for three smoking categories: 'Current smoker', 'Ex smoker', and 'Never smoker'. The bars are grouped by sex (male and female). The y-axis is labeled 'Count'. A legend indicates 'Cluster on X: set color' with 'sex' as the variable. Another legend indicates 'smoking' as the variable for the bars.
- Gallery:** A section with tabs for 'Gallery', 'Basic Elements', 'Groups/Point ID', and 'Titles/Footnotes'. Under 'Choose from:', a list of chart types is shown, with 'Bar' selected. A 'Clustered Bar' chart type is highlighted.
- Buttons:** 'Element Properties...', 'Options...', 'OK', 'Paste', 'Reset', 'Cancel', and 'Help'.

Element Properties:

- Edit Properties of:** A list of elements including 'Bar1', 'X-Axis1 (Bar1)', 'Y-Axis1 (Bar1)', and 'GroupColor (Bar1)'. 'Bar1' is selected.
- Statistics:** A section with 'Variable:' set to 'Count' and a 'Set Parameters...' button.
- Display error bars:** A checkbox that is currently unchecked.
- Error Bars Represent:** A section with three radio buttons: 'Confidence intervals' (selected), 'Standard error', and 'Standard deviation'. Each has a 'Multiplier:' field with the value '2'.
- Bar Style:** A dropdown menu set to 'Bar'.
- Buttons:** 'Apply', 'Close', and 'Help'.

Descriptive Statistics

2-dimensional tables by stacked barplot: e.g.: table on Gender x Smoking

The image shows the Minitab Chart Builder interface and its associated dialog boxes. The Chart Builder window displays a stacked bar chart with three categories on the X-axis: 'Current smoker', 'Ex smoker', and 'Never smoker'. The Y-axis is labeled 'Percentage'. The chart is stacked by 'sex' (male/female) and colored by 'sex'. The 'Variables' list includes ID, testID, t, sex, age, smoking, alcon, height, weight, WHR, and waist. The 'Gallery' shows various chart types, with 'Bar' selected. The 'Element Properties' dialog is open, showing the 'Statistics' section with 'Percentage ()' selected. The 'Set Parameters...' button is highlighted with a red arrow. The 'Element Properties: Set Parameters' dialog is also open, showing the 'Denominator for Computing Percentage' set to 'Total for Each X-Axis Category'.

Chart Builder

Variables:

- ID
- testID
- t
- sex
- age
- smoking
- alcon
- height
- weight
- WHR
- waist

Chart preview uses example data

Stack: set color

sex

Percentage

Current smoker Ex smoker Never smoker

smoking

No categories (scale variable)

Element Properties

Edit Properties of:

- Bar1
- X-Axis1 (Bar1)
- Y-Axis1 (Bar1)
- GroupColor (Bar1)

Statistics

Variable:

Statistic: Percentage ()

Set Parameters...

Display error bars

Error Bars Represent

Element Properties: Set Parameters

Denominator for Computing Percentage:

Total for Each X-Axis Category

Continue Cancel Help

Bar Style:

Bar

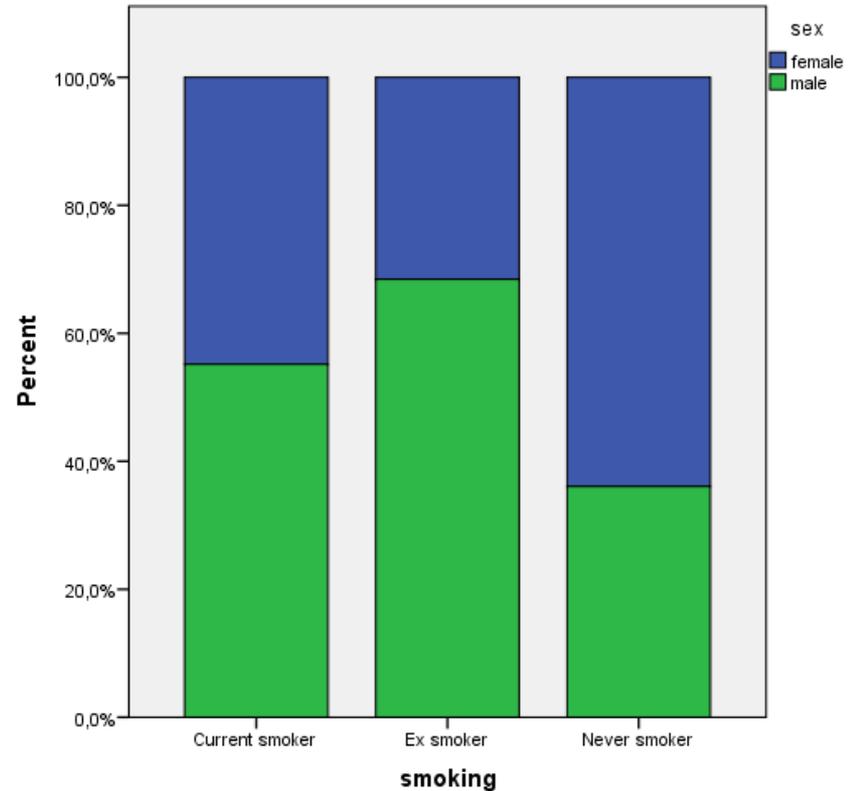
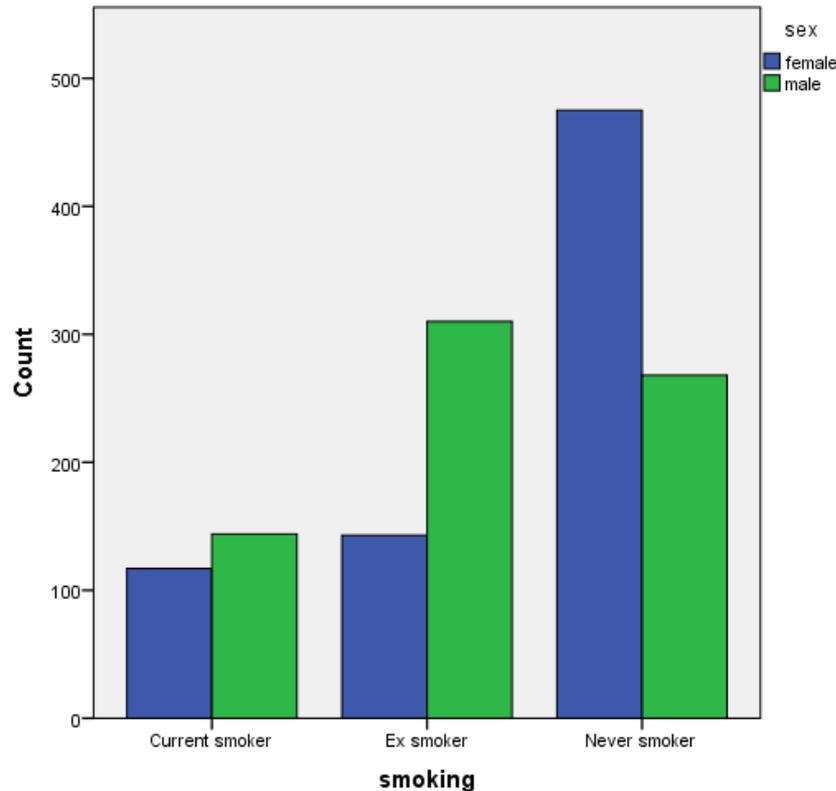
Apply Close Help

Descriptive Statistics

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

Barplots next to each other, absolute frequencies

Stacked barplot, relative frequencies (summing up to 100 for for each smoking category)

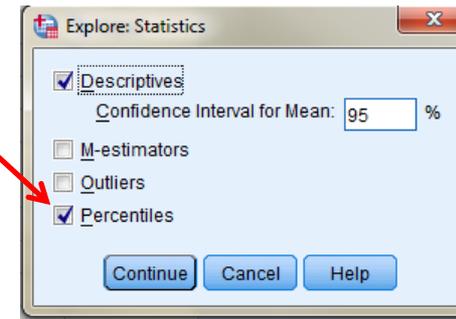
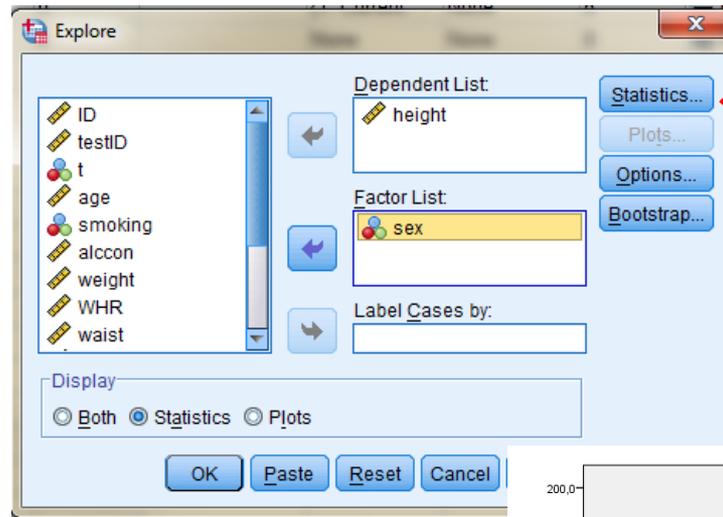


Descriptive Statistics

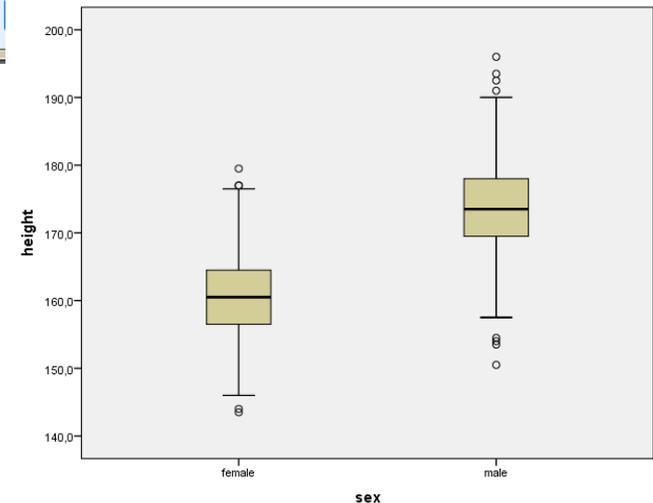
1 qualitative variable, 1 quantitative variable:

Example: Comparing the height between men and women: Calculate measures for location and variance, as well as their graphical illustrations separately for the levels of the qualitative/grouping variable: Analyze → Descriptive Statistics → Explore

sex		Statistic	Std. Error	
height	female	Mean	160,822	
		95% Confidence Interval for Mean	Lower Bound 160,389 Upper Bound 161,254	
	5% Trimmed Mean	160,776		
	Median	160,500		
	Variance	35,687		
	Std. Deviation	5,9739		
	Minimum	143,5		
	Maximum	179,5		
	Range	36,0		
	Interquartile Range	8,0		
	Skewness	,140	,090	
	Kurtosis	-,092	,180	
	male	Mean	173,772	,2437
			95% Confidence Interval for Mean	Lower Bound 173,294 Upper Bound 174,251
5% Trimmed Mean		173,781		
Median		173,500		
Variance		42,887		
Std. Deviation		6,5488		
Minimum		150,5		
Maximum		196,0		
Range		45,5		
Interquartile Range		8,5		
Skewness		,020	,091	
Kurtosis		,246	,182	

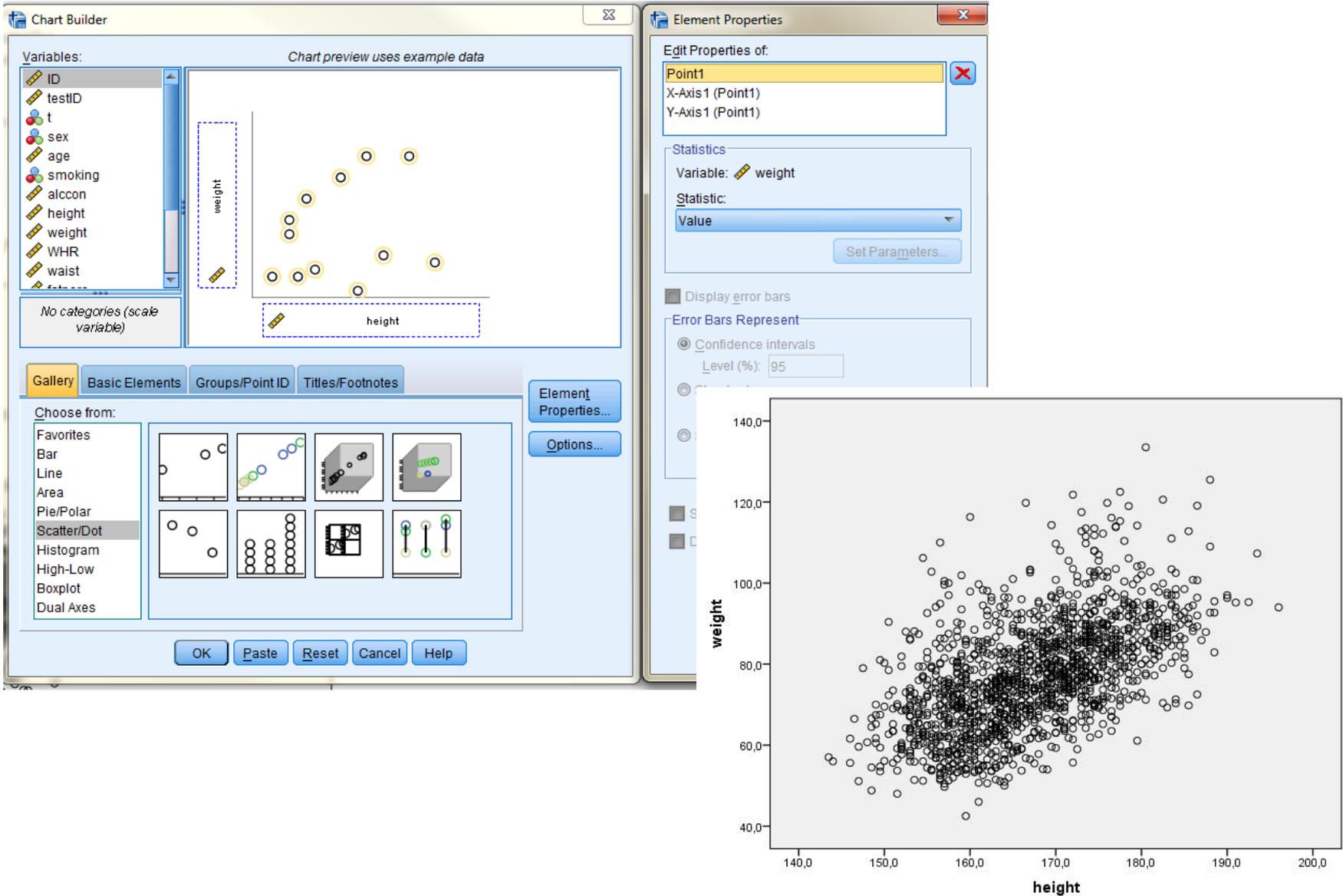


		Percentiles						
		5	10	25	50	75	90	95
Weighted Average (Definition 1)	height female	151,500	153,000	156,500	160,500	164,500	169,000	171,125
	height male	163,000	165,650	169,500	173,500	178,000	182,500	185,000
Tukey's Hinges	height female			156,500	160,500	164,500		
	height male			169,500	173,500	178,000		



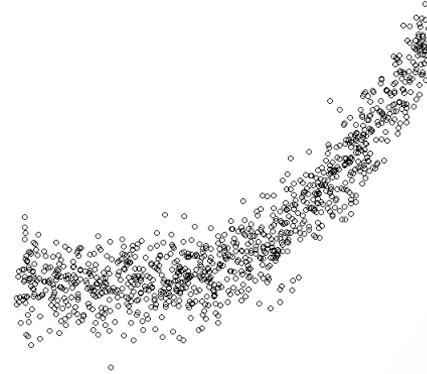
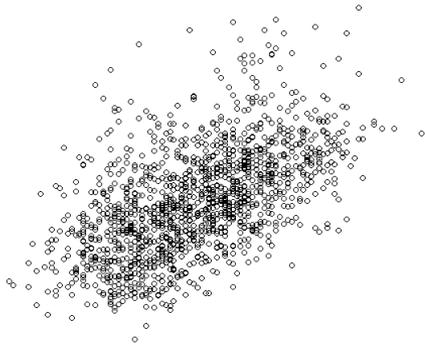
Descriptive Statistics

2 quantitative variables: Simple scatter plots between two variables:



Descriptive Statistics

2 quantitative variables: Correlation

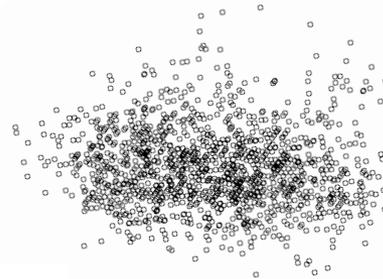


Positive correlation:

$$r > 0$$

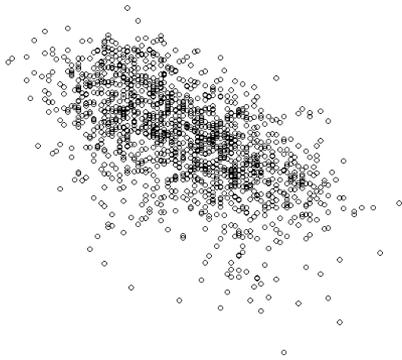
No correlation:

$$r \sim 0$$



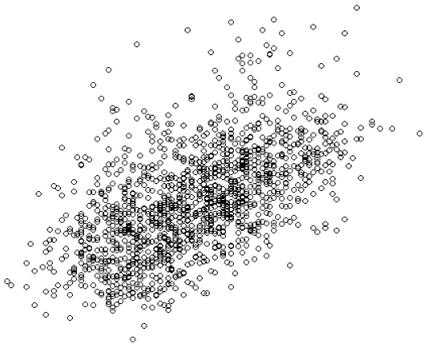
Negative correlation:

$$r < 0$$

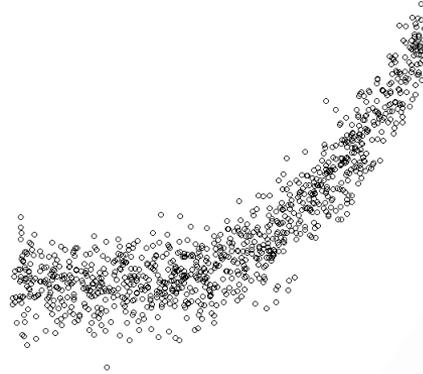


Descriptive Statistics

2 quantitative variables: Correlation

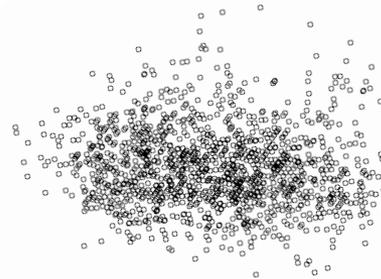


No correlation:
 $r \sim 0$

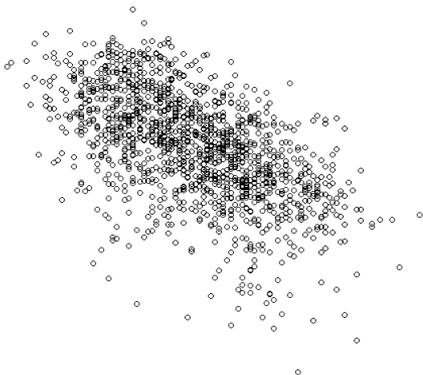


Positive correlation:
 $r > 0$

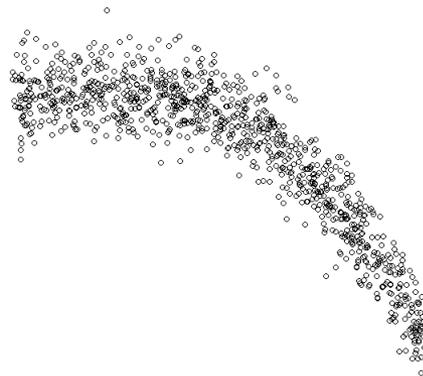
There is a structure
in the data, but



Correlation coefficient
is not able to find it



Negative correlation:
 $r < 0$



Descriptive Statistics

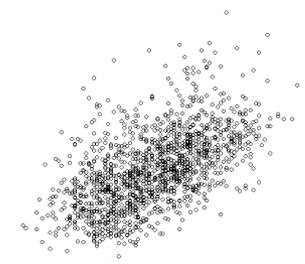
2 quantitative variables: Which correlation coefficient to use?

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

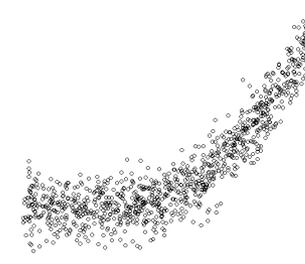


Spearman correlation coefficient:

Estimate a rank-based measure of association.

Observations of Var x	Rank rank(x)
22	1
18	2
15	3.5
15	3.5
11	4

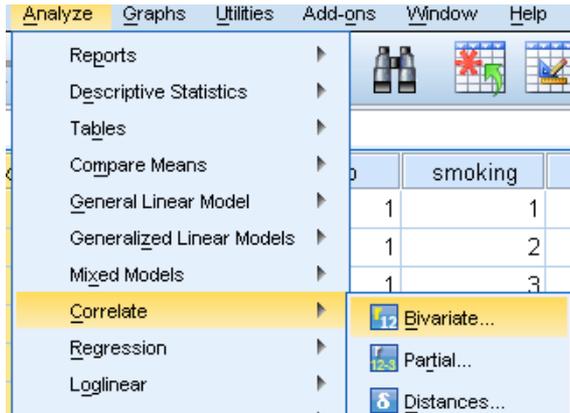
Useful also for nonlinear but monotonic relationships. Data can also be ranks or ordinal.



The correlation coefficient r ranges between -1 and 1

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

Descriptive Statistics



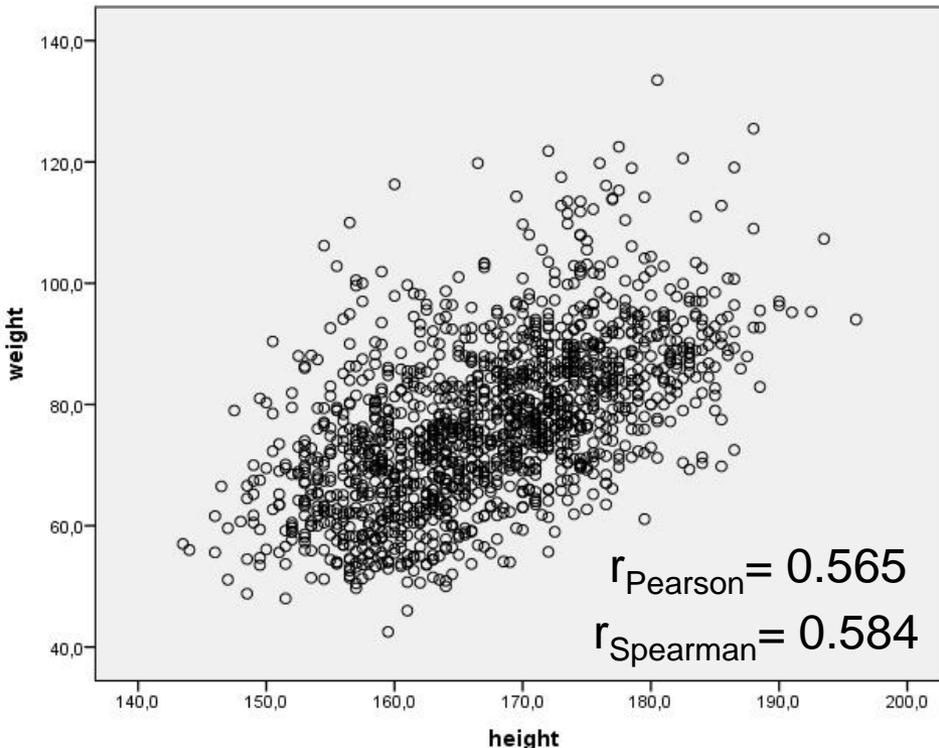
Roughly, correlations can be interpreted in the following way:

$|r| < 0.5$: weak correlation

$0.5 \leq |r| < 0.8$: moderate correlation

$0.8 \geq |r|$: 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).

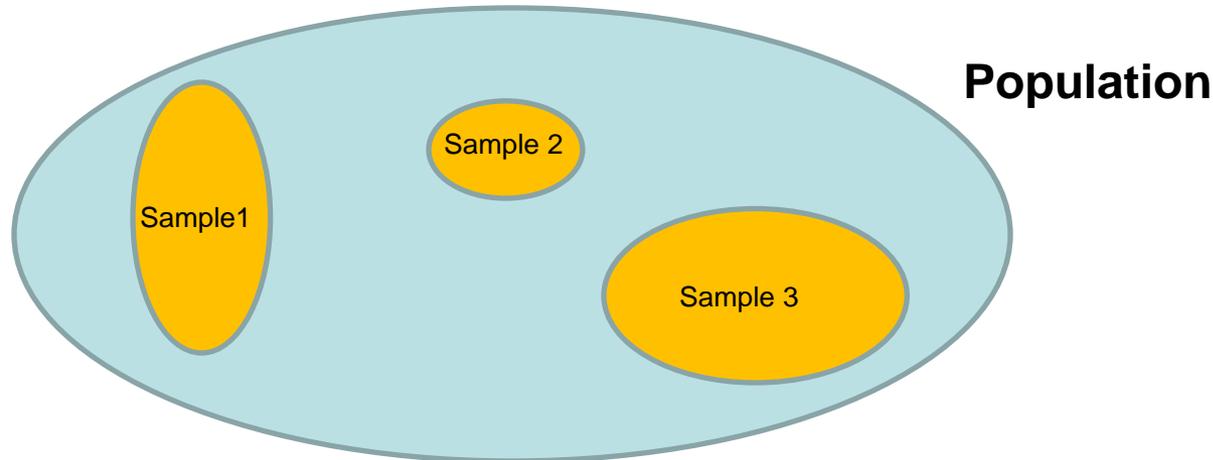




Point and Confidence Estimates

Point and confidence estimates

- We have calculated measures of location (e.g. mean) in our study sample
- But: 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 underlying population → standard error / confidence intervals

Point and confidence estimates

Descriptive measures in the study sample

↓
Conclude on
↓

unknown parameter / characteristic of the underlying population

For example:

Arithmetic Mean in the study sample:

$$\bar{X} = \sum_{i=1}^n X_i$$

↓
is the estimate of
↓

Expectation value μ of the underlying population

Point and confidence estimates

- Example:

Population of interest = All patients with previous MI

„Parameter“ of interest: blood pressure

Study sample: Representative sample of all patients with previous MI

↓
Arithmetic Mean of blood pressure in the study sample:

↓
is the estimate of

↓
Expected value of the blood pressure in the underlying population, that cannot be observed

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 population parameter with given certainty: **Confidence Interval CI**
- The measure of certainty is given by the error probability α

$\alpha = 5\%$: 95% CI; $\alpha = 1\%$: 99% CI etc.

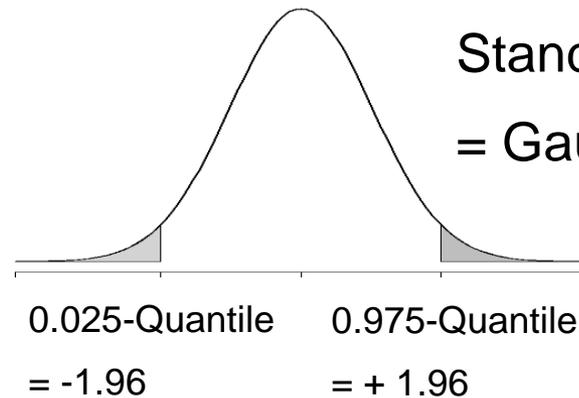
Interpretation of the confidence interval in general:

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

Point and confidence estimates

Example: **CI for the Mean:**

Mean	+/-	Quantiles of the Standardnormal-Distribution	* Standarderror of the Mean
------	-----	--	-----------------------------



Standard Normal distribution
= Gaussian curve

For a 95% confidence interval:

$$\alpha = 5\% \rightarrow Z_{1-\frac{\alpha}{2}} \rightarrow z_{0.975} \rightarrow \sim 1.96$$

Point and confidence estimates

$$\left[\bar{X} - z_{1-\frac{\alpha}{2}} \overbrace{\frac{S}{\sqrt{n}}}^{\text{SE}^*}, \bar{X} + z_{1-\frac{\alpha}{2}} \overbrace{\frac{S}{\sqrt{n}}}^{\text{SE}^*} \right]$$

Example for 95% confidence interval for the mean of total cholesterol within a sample of n=1475 individuals

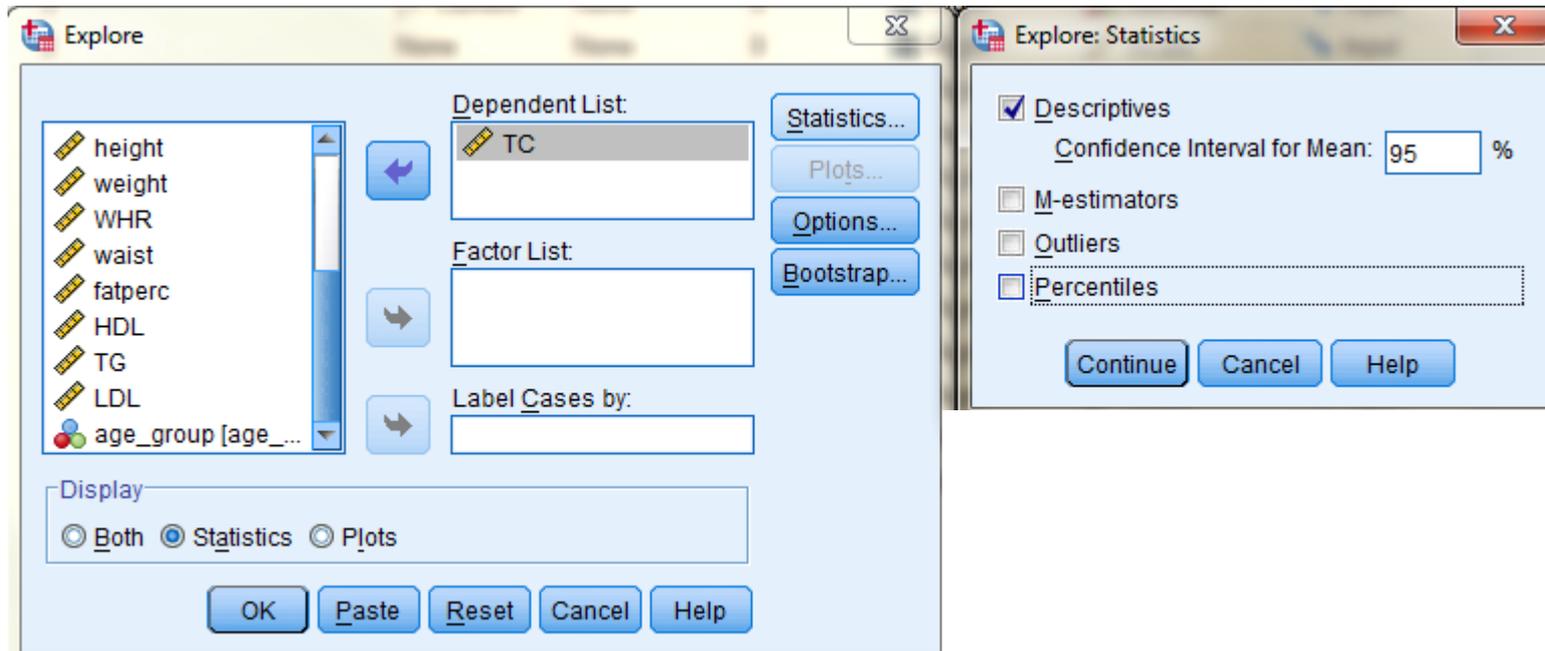
$$\left[237.1 - 1.96 \frac{43.28}{\sqrt{1457}}, 237.1 + 1.96 \frac{43.28}{\sqrt{1457}} \right]$$

$$[234.88, 239.32]$$

If this study is repeated 100 times, it is expected that the mean of total cholesterol falls into this range in 95 times

SE*=standard error of the mean

Point and confidence estimates



Descriptives

		Statistic	Std. Error
TC	Mean	237,100	1,1338
	95% Confidence Interval for Mean	Lower Bound	234,876
		Upper Bound	239,324
	5% Trimmed Mean	235,974	
	Median	233,900	
	Variance	1872,925	
	Std. Deviation	43,2773	
	Minimum	88,1	
	Maximum	755,6	
	Range	667,5	
	Interquartile Range	53,5	
	Skewness	1,446	,064
	Kurtosis	14,042	,128

Confidence intervals can be used to compare groups

→ Do CIs overlap or not ?

If yes: groups do not differ

If no: groups do differ



Formulating a statistical hypothesis

Formulating a statistical hypothesis

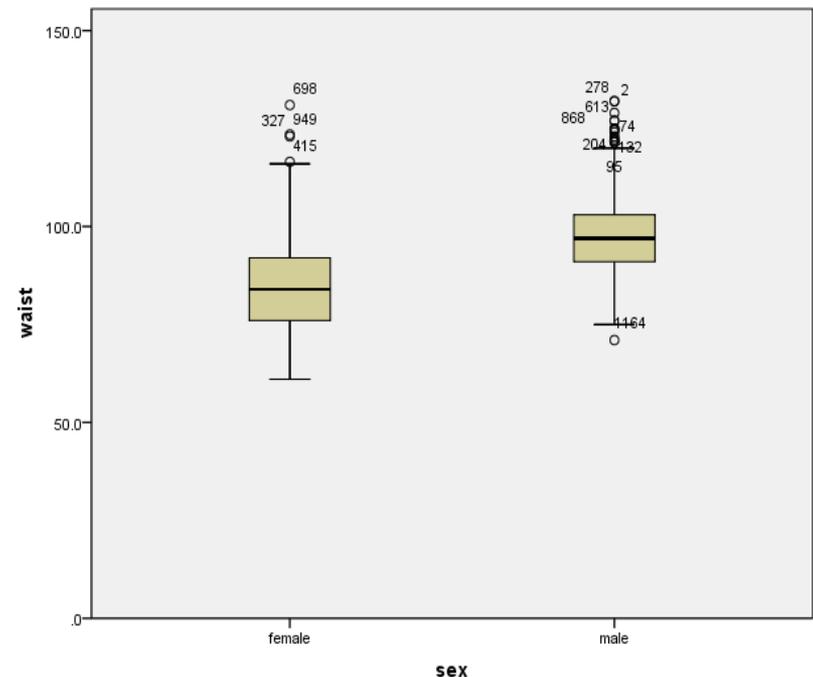
- If one has no a priori hypothesis, results from descriptive / explorative analyses can be used to formulate hypotheses (ideally on different datasets!)

Example:

Waist circumference seems to differ between men and women.

Is this difference statistically significant?

But: With statistics per se, it cannot be answered, if this difference is meaningful or clinically relevant !



Formulating a statistical hypothesis

First steps in conducting a statistical test:

- Quantify the scientific problem from a clinical / biological perspective
- Formulate the problem as a statistical testing problem:
Null-hypothesis versus alternative hypothesis

The current example:

- ▶ Scientific hypothesis: Waist circumference differs between men and women
- ▶ Statistical hypothesis:
 - **Null hypothesis:** Waist circumference does not differ between men and women
 - **Alternative hypothesis:** Waist circumference differs between men and women → **This is the hypothesis you want to proof**