Thesis Seminar
Introduction to statistics

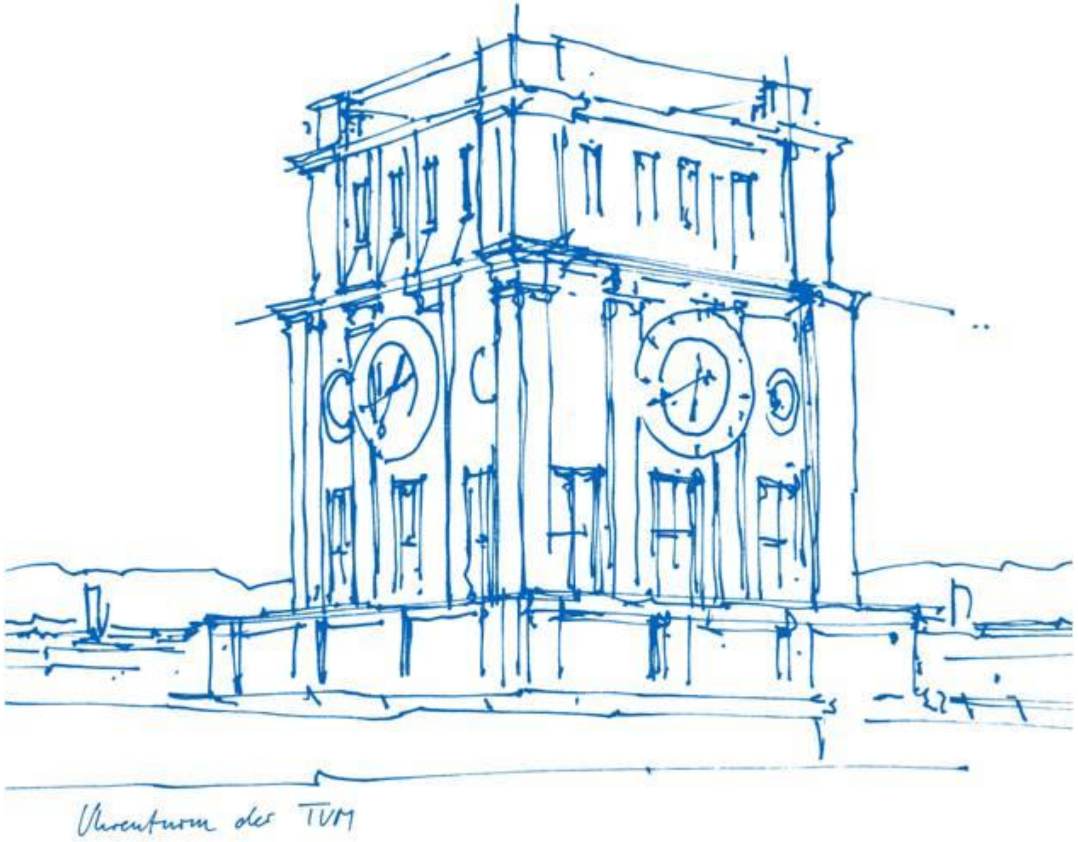
Dr. Theresa Treffers

Technical University of Munich

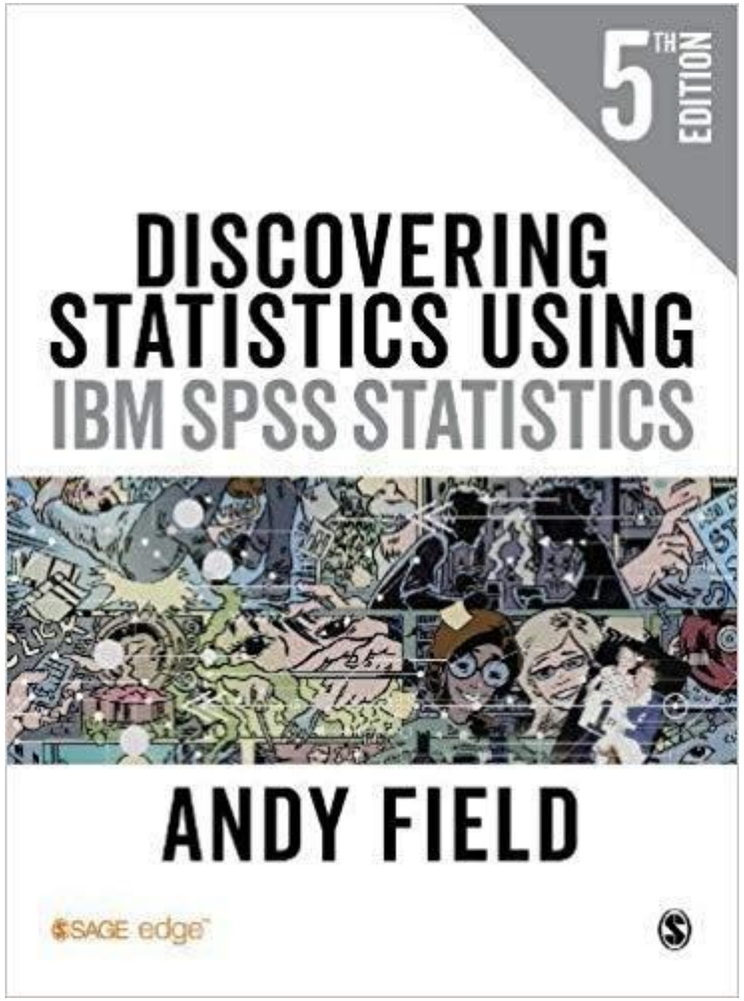
TUM School of Management

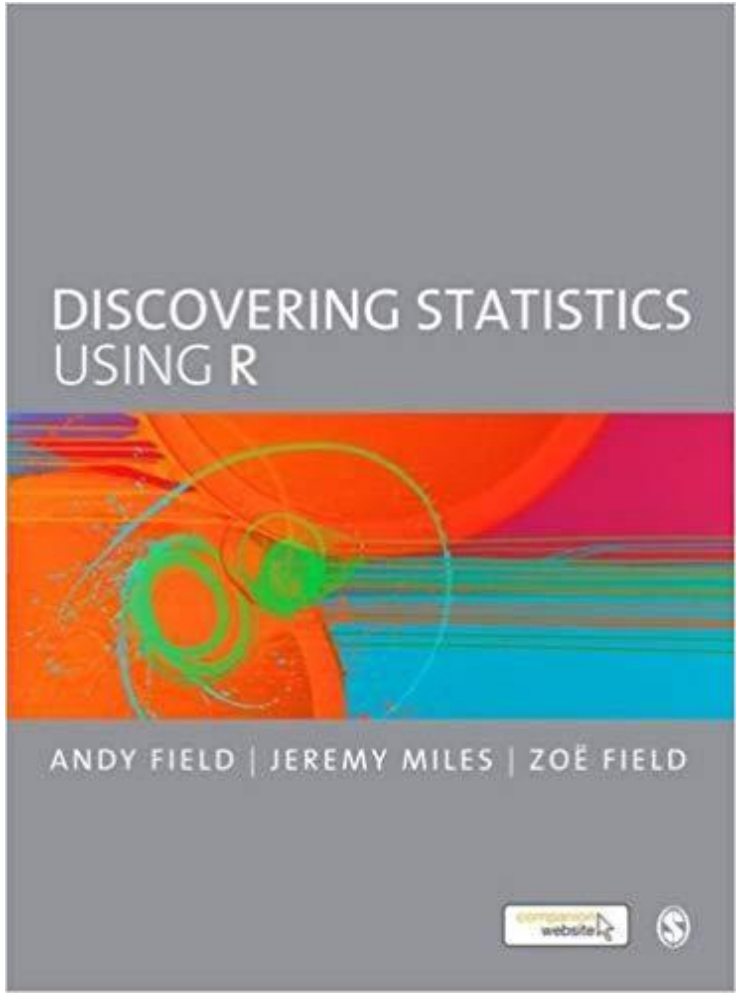
Chair for Strategy and Organization

Prof. Dr. Isabell M. Welpe



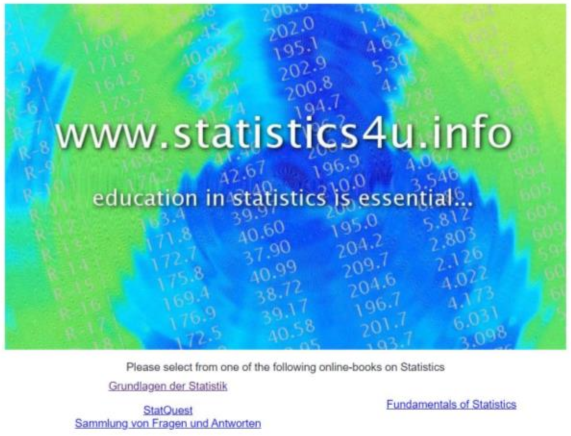
A helpful book:
Andy Field: Discovering statistics using SPSS or R




DISCOVERING STATISTICS USING R
ANDY FIELD | JEREMY MILES | ZOË FIELD
companion website

Some helpful links

http://www.statistics4u.info/



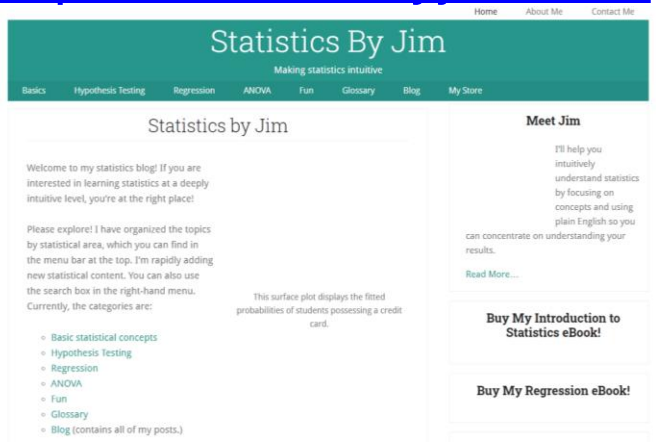
Please select from one of the following online-books on Statistics

Grundlagen der Statistik

StatQuest
Sammlung von Fragen und Antworten

Fundamentals of Statistics

https://statisticsbyjim.com/



Statistics By Jim

Making statistics intuitive

Statistics by Jim

Welcome to my statistics blog! If you are interested in learning statistics at a deeply intuitive level, you're at the right place!

Please explore! I have organized the topics by statistical area, which you can find in the menu bar at the top. I'm rapidly adding new statistical content. You can also use the search box in the right-hand menu. Currently, the categories are:

- Basic statistical concepts
- Hypothesis Testing
- Regression
- ANOVA
- Fun
- Glossary
- Blog (contains all of my posts.)

This surface plot displays the fitted probabilities of students possessing a credit card.

Meet Jim

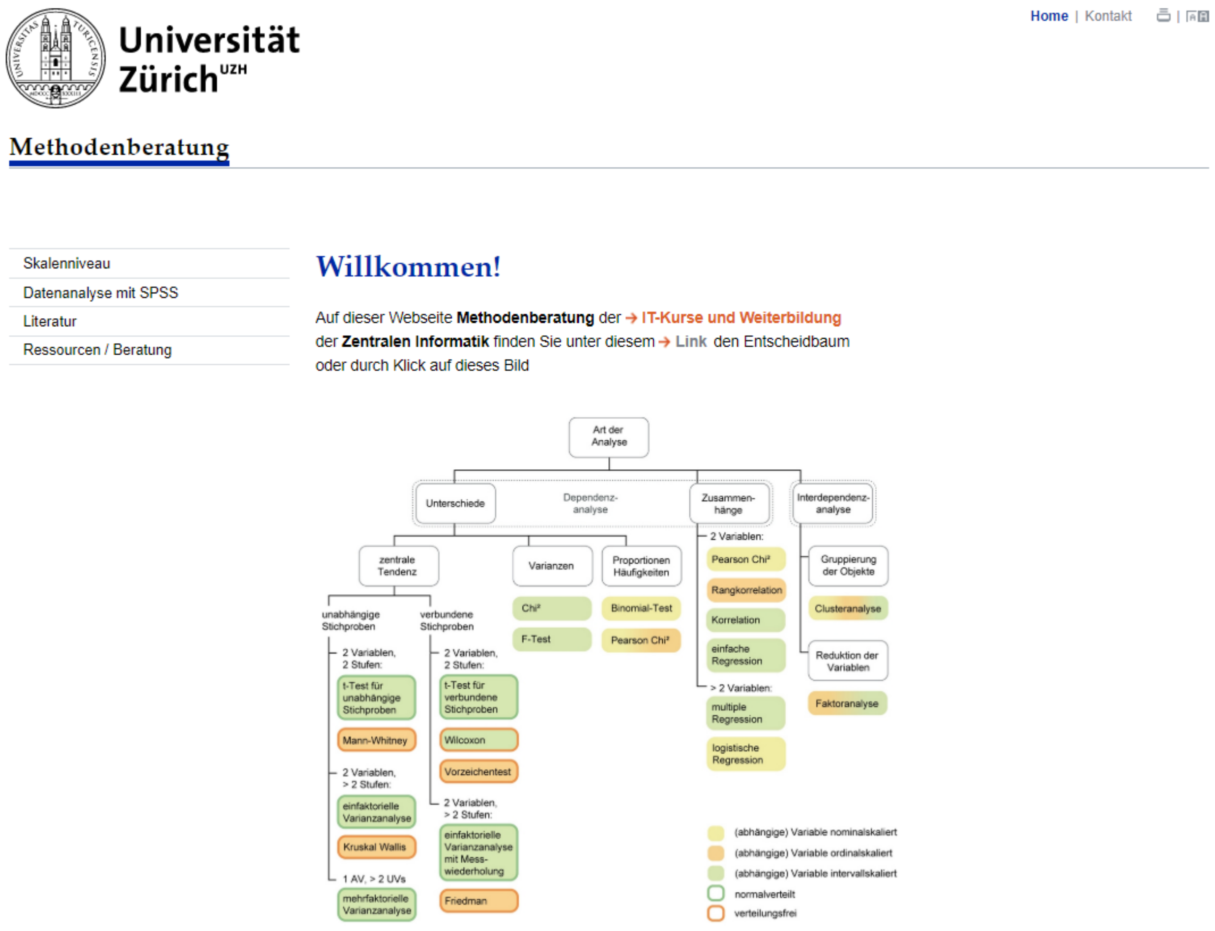
I'll help you intuitively understand statistics by focusing on concepts and using plain English so you can concentrate on understanding your results.

Read More...

Buy My Introduction to Statistics eBook!

Buy My Regression eBook!

http://www.methodenberatung.uzh.ch/index.html



Universität Zürich^{UZH}

Methodenberatung

Willkommen!

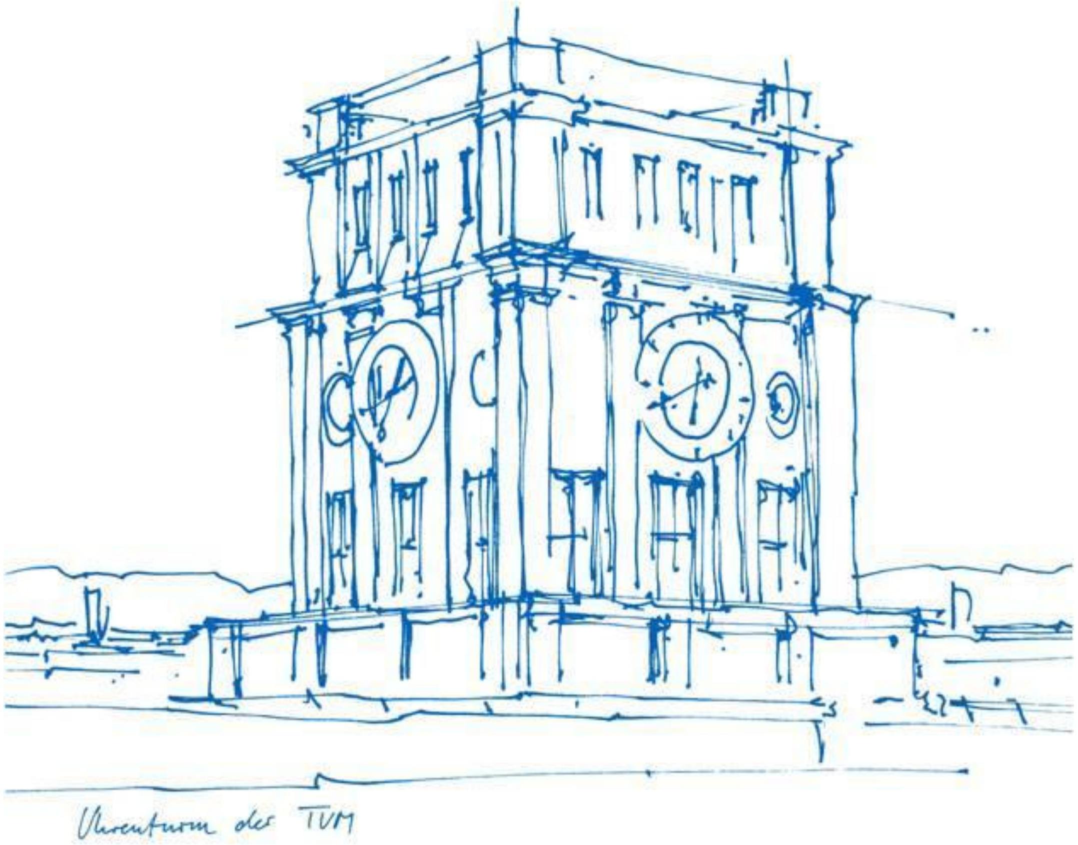
Auf dieser Webseite **Methodenberatung** der → IT-Kurse und Weiterbildung der **Zentralen Informatik** finden Sie unter diesem → Link den Entscheidbaum oder durch Klick auf dieses Bild

A short introduction to statistics

1. Data preparation
2. Data description / descriptive statistics
3. Data analysis / Inferential statistics

→ **DOCUMENT ALL STEPS TO REPRODUCE YOUR RESULTS!!!**

Data preparation

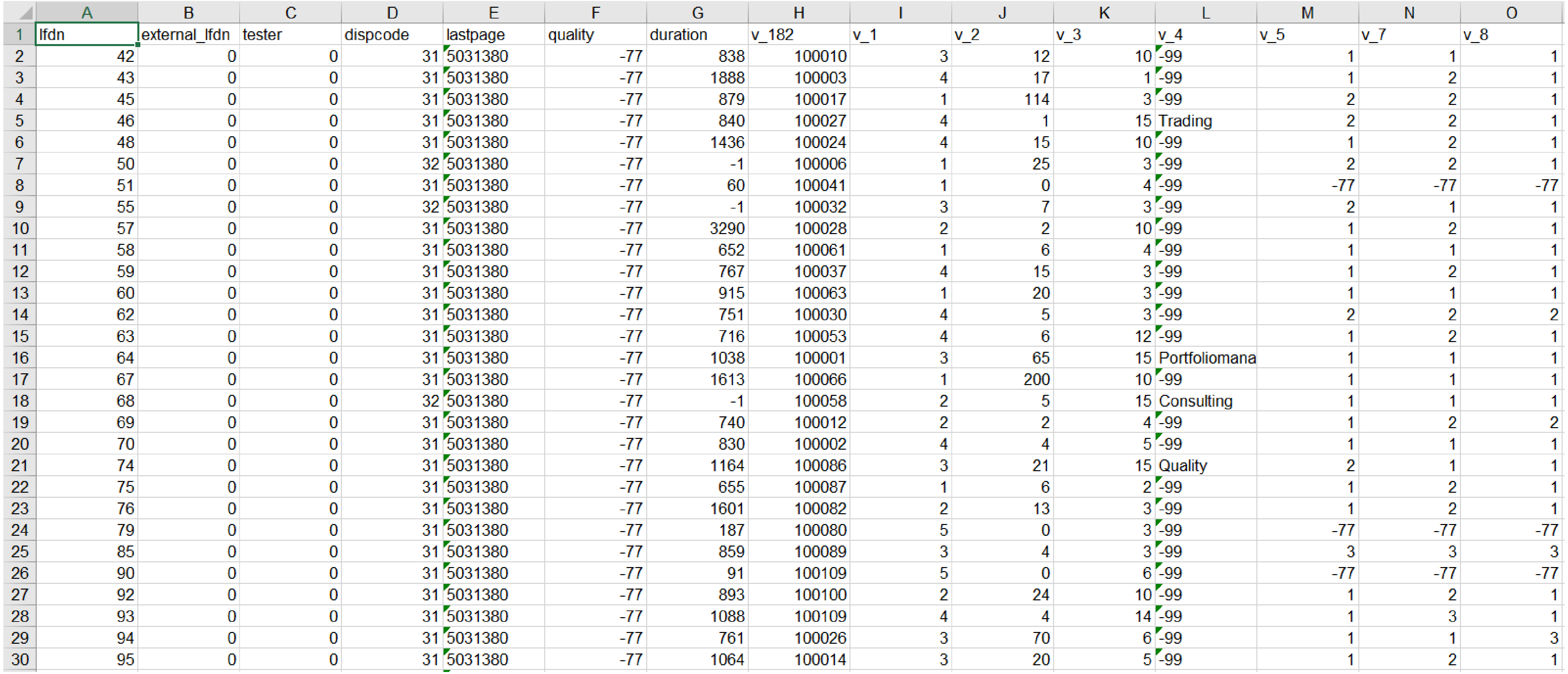


“Conducting data analysis is like drinking a fine wine. It is important to swirl and sniff the wine, to unpack the complex bouquet and to appreciate the experience. Gulping the wine doesn't work.”

- Daniel B. Wright -

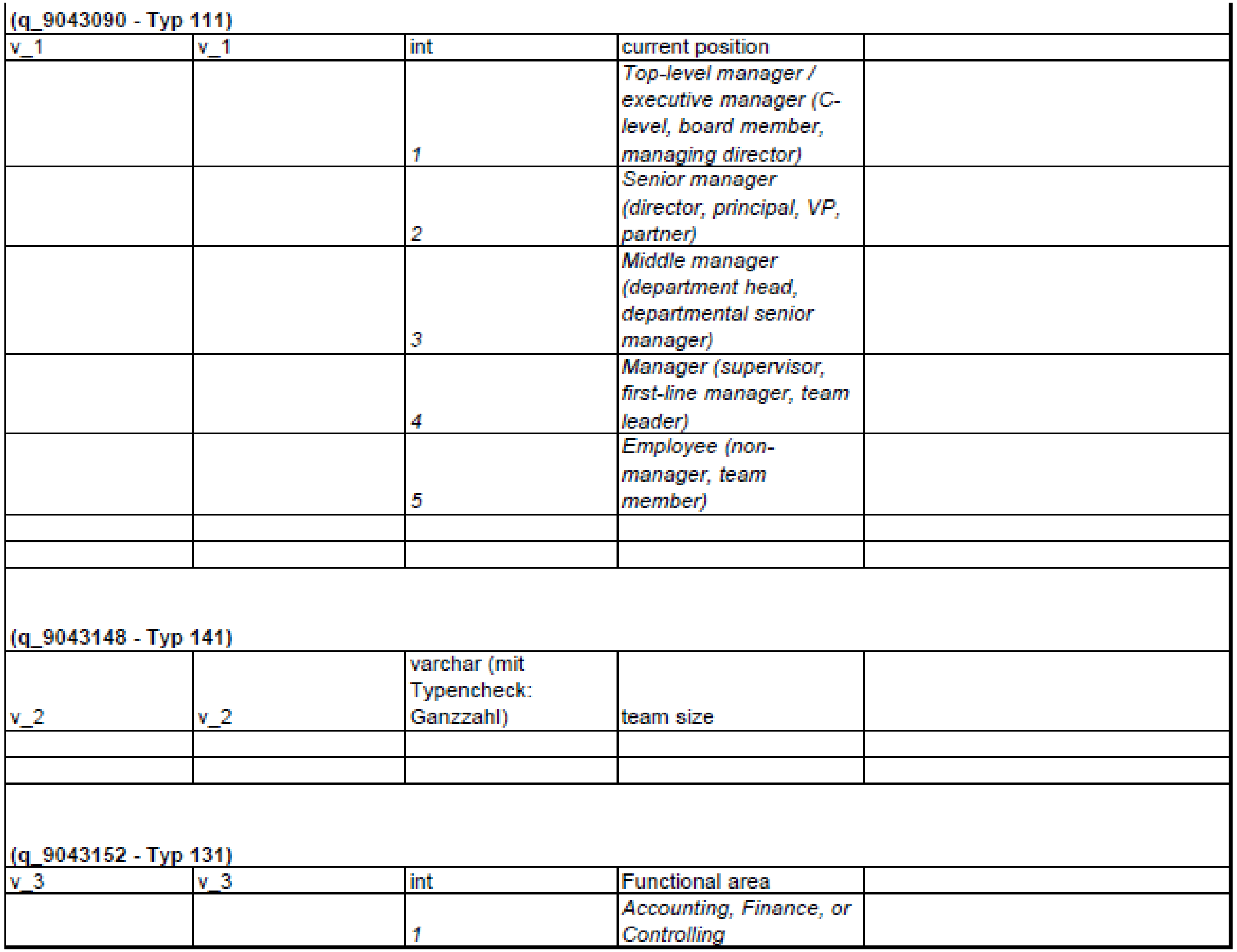
https://www.causeweb.org/cause/resources/fun/quotes/wright-data-analysis

Raw data set - example



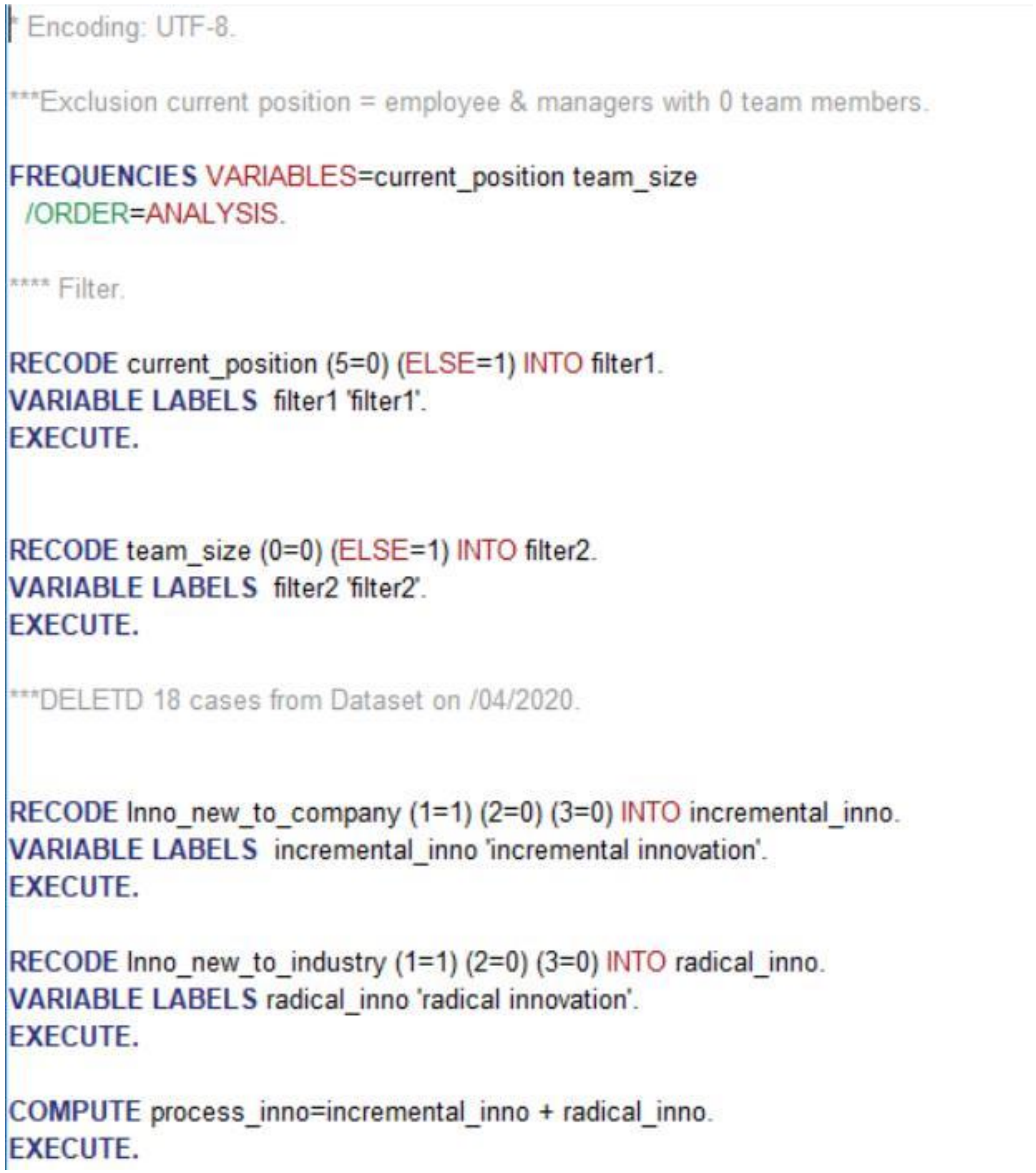
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	lfdn	external_lfdn	tester	dispcode	lastpage	quality	duration	v_182	v_1	v_2	v_3	v_4	v_5	v_7	v_8
2	42	0	0	31	5031380	-77	838	100010	3	12	10	-99	1	1	1
3	43	0	0	31	5031380	-77	1888	100003	4	17	1	-99	1	2	1
4	45	0	0	31	5031380	-77	879	100017	1	114	3	-99	2	2	1
5	46	0	0	31	5031380	-77	840	100027	4	1	15	Trading	2	2	1
6	48	0	0	31	5031380	-77	1436	100024	4	15	10	-99	1	2	1
7	50	0	0	32	5031380	-77	-1	100006	1	25	3	-99	2	2	1
8	51	0	0	31	5031380	-77	60	100041	1	0	4	-99	-77	-77	-77
9	55	0	0	32	5031380	-77	-1	100032	3	7	3	-99	2	1	1
10	57	0	0	31	5031380	-77	3290	100028	2	2	10	-99	1	2	1
11	58	0	0	31	5031380	-77	652	100061	1	6	4	-99	1	1	1
12	59	0	0	31	5031380	-77	767	100037	4	15	3	-99	1	2	1
13	60	0	0	31	5031380	-77	915	100063	1	20	3	-99	1	1	1
14	62	0	0	31	5031380	-77	751	100030	4	5	3	-99	2	2	2
15	63	0	0	31	5031380	-77	716	100053	4	6	12	-99	1	2	1
16	64	0	0	31	5031380	-77	1038	100001	3	65	15	Portfoliomana	1	1	1
17	67	0	0	31	5031380	-77	1613	100066	1	200	10	-99	1	1	1
18	68	0	0	32	5031380	-77	-1	100058	2	5	15	Consulting	1	1	1
19	69	0	0	31	5031380	-77	740	100012	2	2	4	-99	1	2	2
20	70	0	0	31	5031380	-77	830	100002	4	4	5	-99	1	1	1
21	74	0	0	31	5031380	-77	1164	100086	3	21	15	Quality	2	1	1
22	75	0	0	31	5031380	-77	655	100087	1	6	2	-99	1	2	1
23	76	0	0	31	5031380	-77	1601	100082	2	13	3	-99	1	2	1
24	79	0	0	31	5031380	-77	187	100080	5	0	3	-99	-77	-77	-77
25	85	0	0	31	5031380	-77	859	100089	3	4	3	-99	3	3	3
26	90	0	0	31	5031380	-77	91	100109	5	0	6	-99	-77	-77	-77
27	92	0	0	31	5031380	-77	893	100100	2	24	10	-99	1	2	1
28	93	0	0	31	5031380	-77	1088	100109	4	4	14	-99	1	3	1
29	94	0	0	31	5031380	-77	761	100026	3	70	6	-99	1	1	3
30	95	0	0	31	5031380	-77	1064	100014	3	20	5	-99	1	2	1

Codebook from Unipark - example



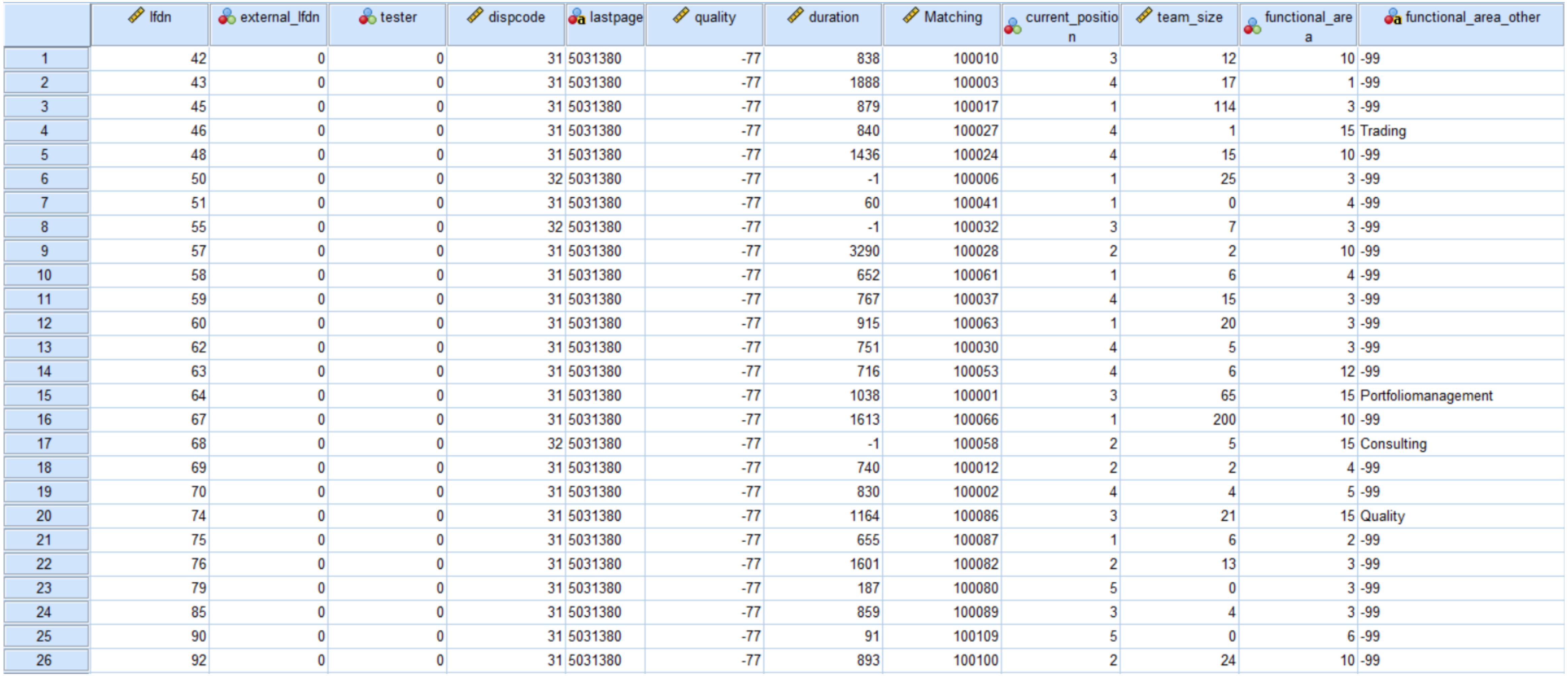
(q_9043090 - Typ 111)				
v_1	v_1	int	current position	
		1	*Top-level manager / executive manager (C-level, board member, managing director)*	
		2	*Senior manager (director, principal, VP, partner)*	
		3	*Middle manager (department head, departmental senior manager)*	
		4	*Manager (supervisor, first-line manager, team leader)*	
		5	*Employee (non-manager, team member)*	
(q_9043148 - Typ 141)				
v_2	v_2	varchar (mit Typencheck: Ganzzahl)	team size	
(q_9043152 - Typ 131)				
v_3	v_3	int	Functional area	
		1	*Accounting, Finance, or Controlling*	

SPSS Syntax - example



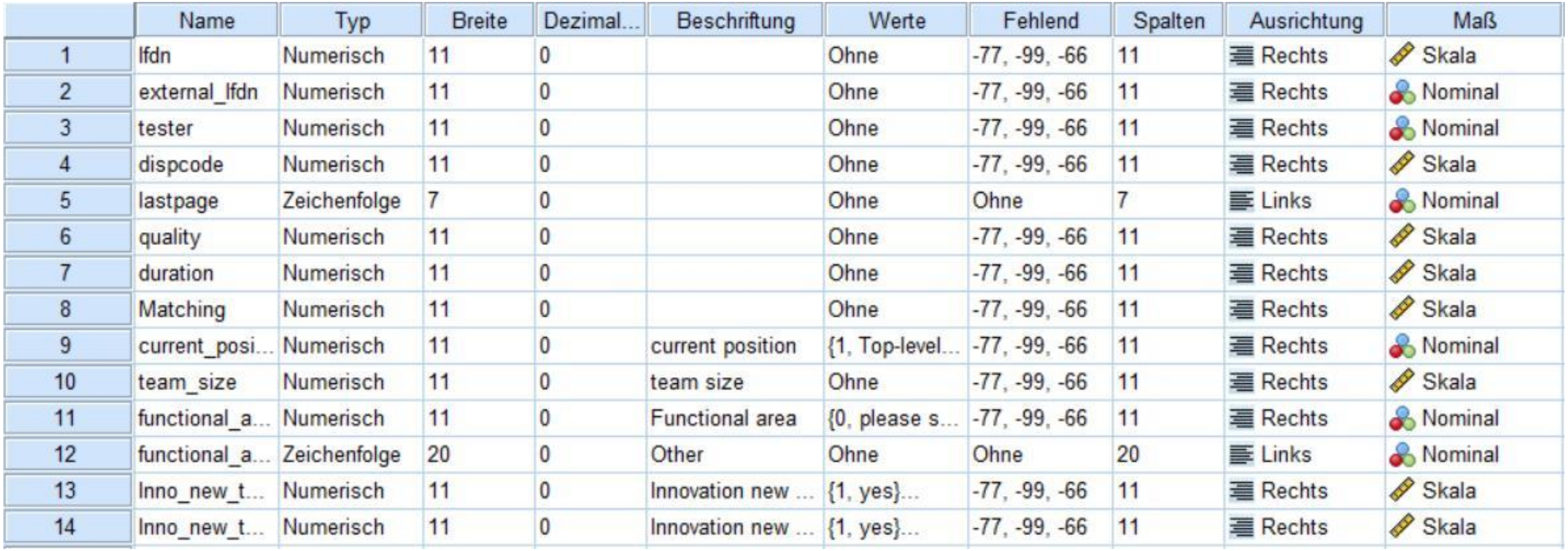
```
* Encoding: UTF-8.

***Exclusion current position = employee & managers with 0 team members.

FREQUENCIES VARIABLES=current_position team_size
  /ORDER=ANALYSIS.

**** Filter.

RECODE current_position (5=0) (ELSE=1) INTO filter1.
VARIABLE LABELS  filter1 'filter1'.
EXECUTE.


RECODE team_size (0=0) (ELSE=1) INTO filter2.
VARIABLE LABELS  filter2 'filter2'.
EXECUTE.

***DELETD 18 cases from Dataset on /04/2020.


RECODE Inno_new_to_company (1=1) (2=0) (3=0) INTO incremental_inno.
VARIABLE LABELS  incremental_inno 'incremental innovation'.
EXECUTE.

RECODE Inno_new_to_industry (1=1) (2=0) (3=0) INTO radical_inno.
VARIABLE LABELS radical_inno 'radical innovation'.
EXECUTE.

COMPUTE process_inno=incremental_inno + radical_inno.
EXECUTE.
```

Final data set - example



	lfdn	external_lfdn	tester	dispcode	lastpage	quality	duration	Matching	current_position	team_size	functional_area	functional_area_other
1	42	0	0	31	5031380	-77	838	100010	3	12	10	-99
2	43	0	0	31	5031380	-77	1888	100003	4	17	1	-99
3	45	0	0	31	5031380	-77	879	100017	1	114	3	-99
4	46	0	0	31	5031380	-77	840	100027	4	1	15	Trading
5	48	0	0	31	5031380	-77	1436	100024	4	15	10	-99
6	50	0	0	32	5031380	-77	-1	100006	1	25	3	-99
7	51	0	0	31	5031380	-77	60	100041	1	0	4	-99
8	55	0	0	32	5031380	-77	-1	100032	3	7	3	-99
9	57	0	0	31	5031380	-77	3290	100028	2	2	10	-99
10	58	0	0	31	5031380	-77	652	100061	1	6	4	-99
11	59	0	0	31	5031380	-77	767	100037	4	15	3	-99
12	60	0	0	31	5031380	-77	915	100063	1	20	3	-99
13	62	0	0	31	5031380	-77	751	100030	4	5	3	-99
14	63	0	0	31	5031380	-77	716	100053	4	6	12	-99
15	64	0	0	31	5031380	-77	1038	100001	3	65	15	Portfoliomanagement
16	67	0	0	31	5031380	-77	1613	100066	1	200	10	-99
17	68	0	0	32	5031380	-77	-1	100058	2	5	15	Consulting
18	69	0	0	31	5031380	-77	740	100012	2	2	4	-99
19	70	0	0	31	5031380	-77	830	100002	4	4	5	-99
20	74	0	0	31	5031380	-77	1164	100086	3	21	15	Quality
21	75	0	0	31	5031380	-77	655	100087	1	6	2	-99
22	76	0	0	31	5031380	-77	1601	100082	2	13	3	-99
23	79	0	0	31	5031380	-77	187	100080	5	0	3	-99
24	85	0	0	31	5031380	-77	859	100089	3	4	3	-99
25	90	0	0	31	5031380	-77	91	100109	5	0	6	-99
26	92	0	0	31	5031380	-77	893	100100	2	24	10	-99

Final data set - example



	Name	Typ	Breite	Dezimal...	Beschriftung	Werte	Fehlend	Spalten	Ausrichtung	Maß
1	lfdn	Numerisch	11	0		Ohne	-77, -99, -66	11	Rechts	Skala
2	external_lfdn	Numerisch	11	0		Ohne	-77, -99, -66	11	Rechts	Nominal
3	tester	Numerisch	11	0		Ohne	-77, -99, -66	11	Rechts	Nominal
4	dispcode	Numerisch	11	0		Ohne	-77, -99, -66	11	Rechts	Skala
5	lastpage	Zeichenfolge	7	0		Ohne	Ohne	7	Links	Nominal
6	quality	Numerisch	11	0		Ohne	-77, -99, -66	11	Rechts	Skala
7	duration	Numerisch	11	0		Ohne	-77, -99, -66	11	Rechts	Skala
8	Matching	Numerisch	11	0		Ohne	-77, -99, -66	11	Rechts	Skala
9	current_posi...	Numerisch	11	0	current position	{1, Top-level...	-77, -99, -66	11	Rechts	Nominal
10	team_size	Numerisch	11	0	team size	Ohne	-77, -99, -66	11	Rechts	Skala
11	functional_a...	Numerisch	11	0	Functional area	{0, please s...	-77, -99, -66	11	Rechts	Nominal
12	functional_a...	Zeichenfolge	20	0	Other	Ohne	Ohne	20	Links	Nominal
13	Inno_new_t...	Numerisch	11	0	Innovation new ...	{1, yes}...	-77, -99, -66	11	Rechts	Skala
14	Inno_new_t...	Numerisch	11	0	Innovation new ...	{1, yes}...	-77, -99, -66	11	Rechts	Skala

Data preparation - getting acquainted with the data

- Label items/variables
- Label items/variable values
- Label measurement levels
- Recode reverse items
- Recode variables (e.g., age)
- Define missing values (→ maybe impute missing values)
- Define (don‘t delete!) participants who failed attention test, time test etc.
- Calculate construct reliabilities (→ Cronbachs Alpha for reflective measures)
- Calculate constructs from items

Hair/Anderson/Tatham/Black (1998)

Data preparation - getting more acquainted with the data

- Uncovering input errors (proofreading, graphical and statistical analysis)

- Checking the distribution form

- Are the conditions for certain statistical procedures fulfilled ?, e.g. normal distribution, homoscedasticity, linearity, etc.

- Description of the data by characteristic values

- Data encoding, recoding, indexing, dummy variable formation

Hair/Anderson/Tatham/Black (1998)

Discover outliers

- There are **four outlier classes** (1) input or coding errors (logically impossible), (2) exceptional events (explanation possible), (3) extraordinary events (explanation not possible), (4) unusual combination of occurrences across variables
- Discover outliers
  - Univariate discovery: Z-transformation of data, convention: values above / below 2.5 standard deviations (below n = 80) and 3-4 standard deviations (above n = 80)
  - Bivariate discovery: Variable pairs can be viewed in scatterplot; ellipse viewing
  - Mahalanobis D^2 measures the distance of each observation from the mean
- **There are many philosophies about how to handle outliers. Important: if the data is changed, this must be logged!**

Hair/Anderson/Tatham/Black (1998)

Missing values

- Missing values are a **frequently underestimated phenomenon** (few standards for dealing with missing values) in the analysis of empirical datasets. They influence the statistical results and the generalizability of these results and reduce the sample size

- The analysis of the causes of missing information should be **part of every empirical investigation**! Missing values have a significant influence on the result of the statistical analysis

- Common practice is **ignoring missing values - this is problematic** because the non-existent data cannot be used in the decision-making process for / against hypotheses!

⇒ Example: It should be recorded whether a company works profitably. Part of the companies refuses to answer, but now, if companies tend to refuse to provide information when they are not profitable, the share of profitable companies will be underestimated

Wirtz (2004)

Missing values in multivariate analyzes

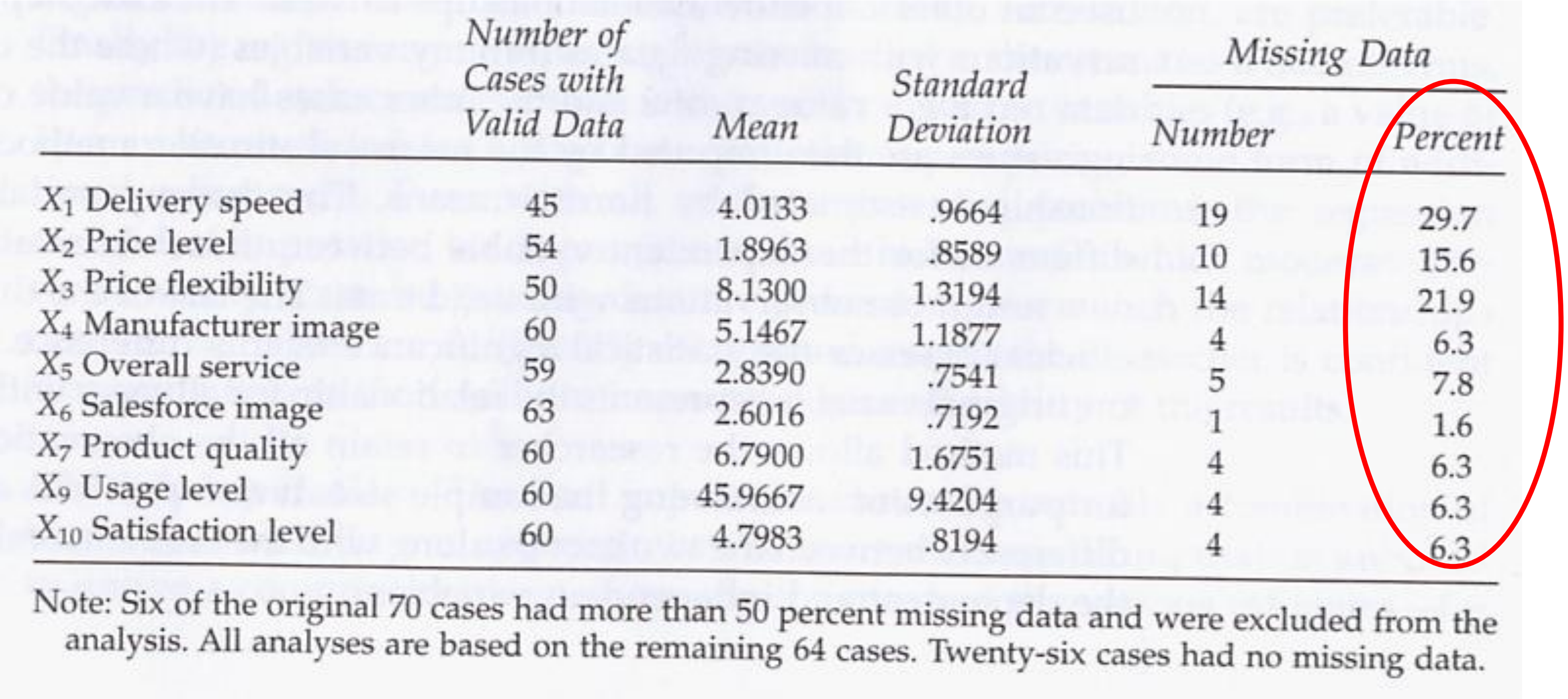
To be able to analyse a multivariate data set with missing values, two central questions must be clarified:

(1) What causes missing information? (If the missing information occurs accidentally or systematically?) Missing values that result from random sampling can be.

(2) On which assumptions are the methods available for handling missing data coupled?

Wirtz (2004)

Output missing values

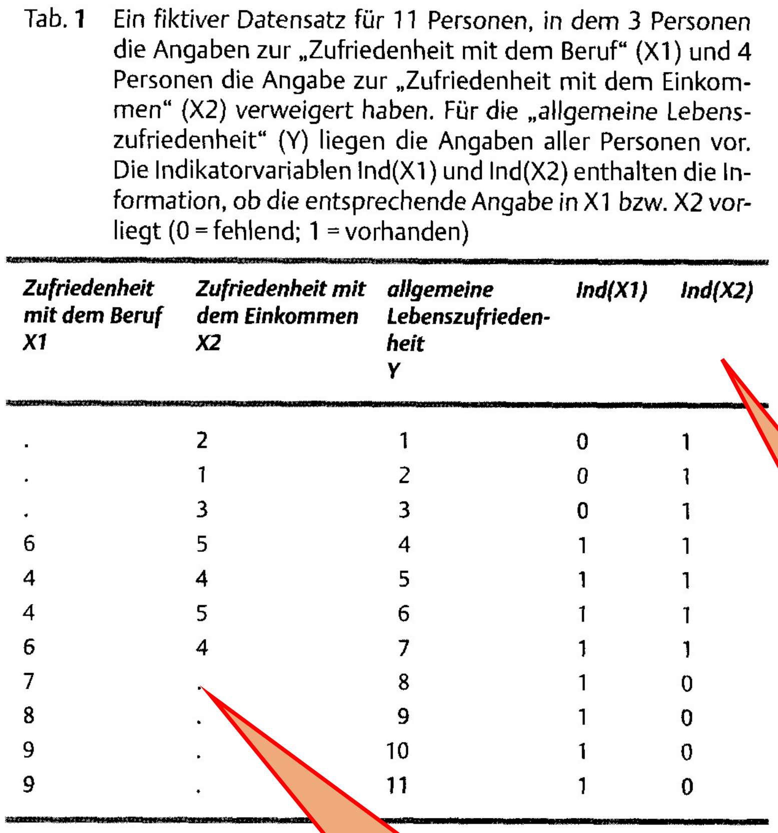


	Number of Cases with Valid Data	Mean	Standard Deviation	Missing Data	
				Number	Percent
X_1 Delivery speed	45	4.0133	.9664	19	29.7
X_2 Price level	54	1.8963	.8589	10	15.6
X_3 Price flexibility	50	8.1300	1.3194	14	21.9
X_4 Manufacturer image	60	5.1467	1.1877	4	6.3
X_5 Overall service	59	2.8390	.7541	5	7.8
X_6 Salesforce image	63	2.6016	.7192	1	1.6
X_7 Product quality	60	6.7900	1.6751	4	6.3
X_9 Usage level	60	45.9667	9.4204	4	6.3
X_{10} Satisfaction level	60	4.7983	.8194	4	6.3

Note: Six of the original 70 cases had more than 50 percent missing data and were excluded from the analysis. All analyses are based on the remaining 64 cases. Twenty-six cases had no missing data.

Hair/Anderson/Tatham/Black (1998)

Missing Data Diagnostics



Tab. 1 Ein fiktiver Datensatz für 11 Personen, in dem 3 Personen die Angaben zur „Zufriedenheit mit dem Beruf“ (X1) und 4 Personen die Angabe zur „Zufriedenheit mit dem Einkommen“ (X2) verweigert haben. Für die „allgemeine Lebenszufriedenheit“ (Y) liegen die Angaben aller Personen vor. Die Indikatorvariablen Ind(X1) und Ind(X2) enthalten die Information, ob die entsprechende Angabe in X1 bzw. X2 vorliegt (0 = fehlend; 1 = vorhanden)

Zufriedenheit mit dem Beruf X1	*Zufriedenheit mit dem Einkommen X2*	*allgemeine Lebenszufriedenheit Y*	*Ind(X1)*	*Ind(X2)*
.	2	1	0	1
.	1	2	0	1
.	3	3	0	1
6	5	4	1	1
4	4	5	1	1
4	5	6	1	1
6	4	7	1	1
7	.	8	1	0
8	.	9	1	0
9	.	10	1	0
9	.	11	1	0

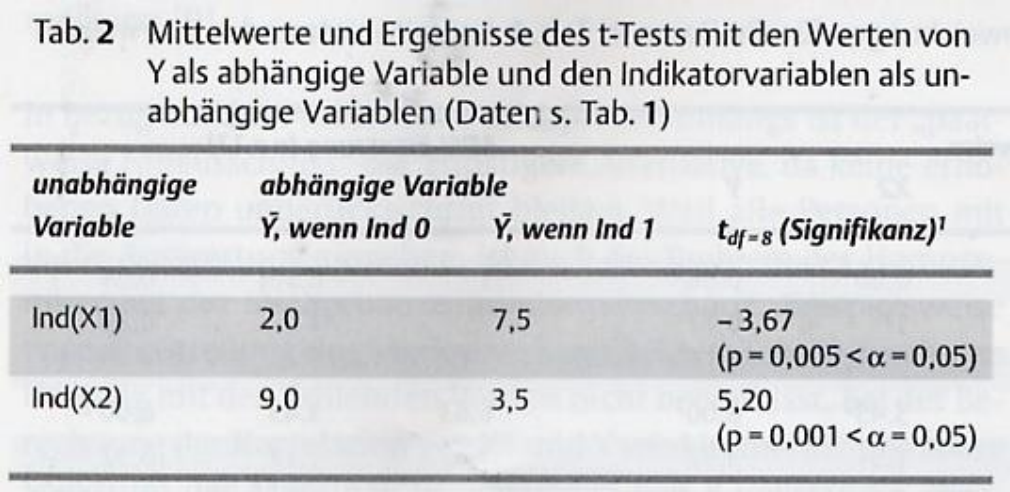
Wirtz (2004)

- Do individual persons or variables have high proportions of missing values? → elimination from 30% missing values more uncertainty and error as information gain, DOCUMENT, for example (X2)
- The missing data diagnosis analyzes whether missing values occur systematically in conjunction with other variables

Indicator variables that encode whether X1 or X2 contain missing values

36% missing values, X2 should be excluded

Missing Data Diagnostics



Tab. 2 Mittelwerte und Ergebnisse des t-Tests mit den Werten von Y als abhängige Variable und den Indikatorvariablen als unabhängige Variablen (Daten s. Tab. 1)

unabhängige	*abhängige Variable*		
Variable	*$\bar{Y}$, wenn Ind 0*	*$\bar{Y}$, wenn Ind 1*	*$t_{df=8}$ (Signifikanz)*[1]
Ind(X1)	2,0	7,5	– 3,67 ($p = 0{,}005 < \alpha = 0{,}05$)
Ind(X2)	9,0	3,5	5,20 ($p = 0{,}001 < \alpha = 0{,}05$)

- Missing data diagnostics should be performed if more than 5% of the values are missing, as a systematic failure can cause significant data distortion
- With a systematic system of missing values, the standard methods for dealing with missing values can no longer be applied
- Missing Data Diagnostics provides an overview!

- For the levels of the indicator variables, a check is made as to whether the existing information differs systematically in all other variables
- Persons with missing values in "job satisfaction" have a significantly lower value in "overall life satisfaction" than the other persons and persons with missing values in "satisfaction with income" have a significantly higher "overall life satisfaction" than those other people
- Missing information did not occur by chance
- Patterns mean that missing values occur together via variables (missing values in variable A, they are also missing in variable B)

Wirtz (2004)

Causes of Missing Data Processes

1. **Missing completely at random (MCAR)** if the missing information (a) does not depend on the expressions of other variables or (b) the values of the (unspecified) values of the variables (for example, errors in data entry or unsystematic return).

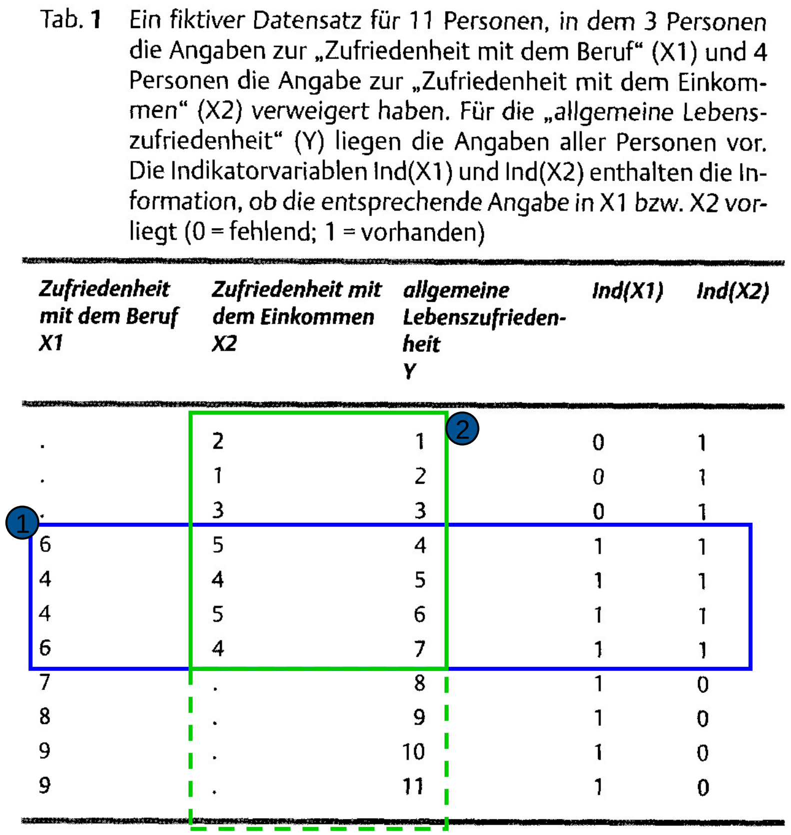
2. **(Conditional) missing at random (MAR),** if the missing information can be completely predicted by the rest of the information in the dataset (e.g. if the motivation to participate in the study is the key predictor of complete lack of information) Data at random (MAR) when the motivation to participate has been measured).

3. **Non Random Missing (NRM)** → It is not possible to decide empirically between MAR and NRM.

Wirtz (2004)

Dealing with missing values

1. **Case-by-rule exclusion:** A person drops out of the calculation of all statistics if one of the variables to be analyzed lacks a value for that person
2. **Pairwise Exclusion:** All available data information is used to calculate each statistic. For example, a person is excluded from calculating correlations only if one of the values of the directly affected variable is not present

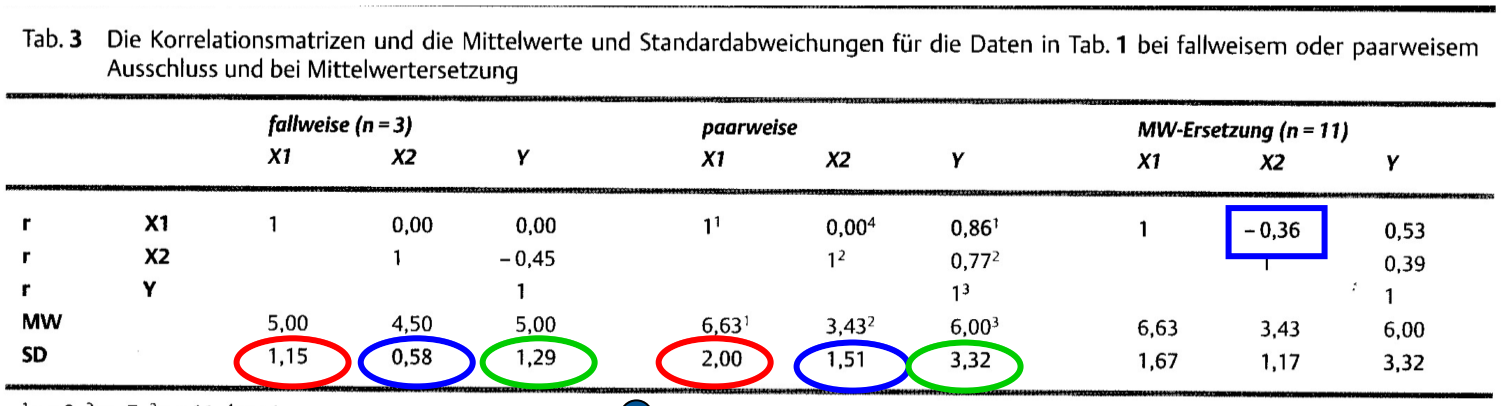


Tab. 1 Ein fiktiver Datensatz für 11 Personen, in dem 3 Personen die Angaben zur „Zufriedenheit mit dem Beruf" (X1) und 4 Personen die Angabe zur „Zufriedenheit mit dem Einkommen" (X2) verweigert haben. Für die „allgemeine Lebenszufriedenheit" (Y) liegen die Angaben aller Personen vor. Die Indikatorvariablen Ind(X1) und Ind(X2) enthalten die Information, ob die entsprechende Angabe in X1 bzw. X2 vorliegt (0 = fehlend; 1 = vorhanden)

Zufriedenheit mit dem Beruf X1	*Zufriedenheit mit dem Einkommen X2*	*allgemeine Lebenszufriedenheit Y*	*Ind(X1)*	*Ind(X2)*
.	2	1	0	1
.	1	2	0	1
.	3	3	0	1
6	5	4	1	1
4	4	5	1	1
4	5	6	1	1
6	4	7	1	1
7	.	8	1	0
8	.	9	1	0
9	.	10	1	0
9	.	11	1	0

Wirtz (2004)

Dealing with missing values



Tab. 3 Die Korrelationsmatrizen und die Mittelwerte und Standardabweichungen für die Daten in Tab. 1 bei fallweisem oder paarweisem Ausschluss und bei Mittelwertersetzung

		fallweise (n = 3)			*paarweise*			*MW-Ersetzung (n = 11)*		
		X1	*X2*	*Y*	*X1*	*X2*	*Y*	*X1*	*X2*	*Y*
r	**X1**	1	0,00	0,00	1[1]	0,00[4]	0,86[1]	1	− 0,36	0,53
r	**X2**		1	− 0,45		1[2]	0,77[2]		1	0,39
r	**Y**			1			1[3]			1
MW		5,00	4,50	5,00	6,63[1]	3,43[2]	6,00[3]	6,63	3,43	6,00
SD		1,15	0,58	1,29	2,00	1,51	3,32	1,67	1,17	3,32

[1] n = 8, [2] n = 7, [3] n = 11, [4] n = 4

Wirtz (2004)

1

- The case-by-case and pairwise exclusion leads to different results → lower correlations for persons with medium Y expression (case-by-case exclusion) → As individuals with extreme characteristics do not respond to certain variables, the sample becomes more homogeneous → *The more homogeneous a sample, the lower the correlations (lower standard deviations*)
- The case-by-case exclusion can change the properties of the sample and distort the statistics (if values are not completely random)
- The pairwise exclusion does not reduce the sample so much and does not homogenize the sample, but problem of sub-sample!
- Replacement by the mean is not recommended: (a) artificial homologation (b) completely random absence would have to exist; the overestimated values in X1 and underestimated in X2

Wirtz (2004)

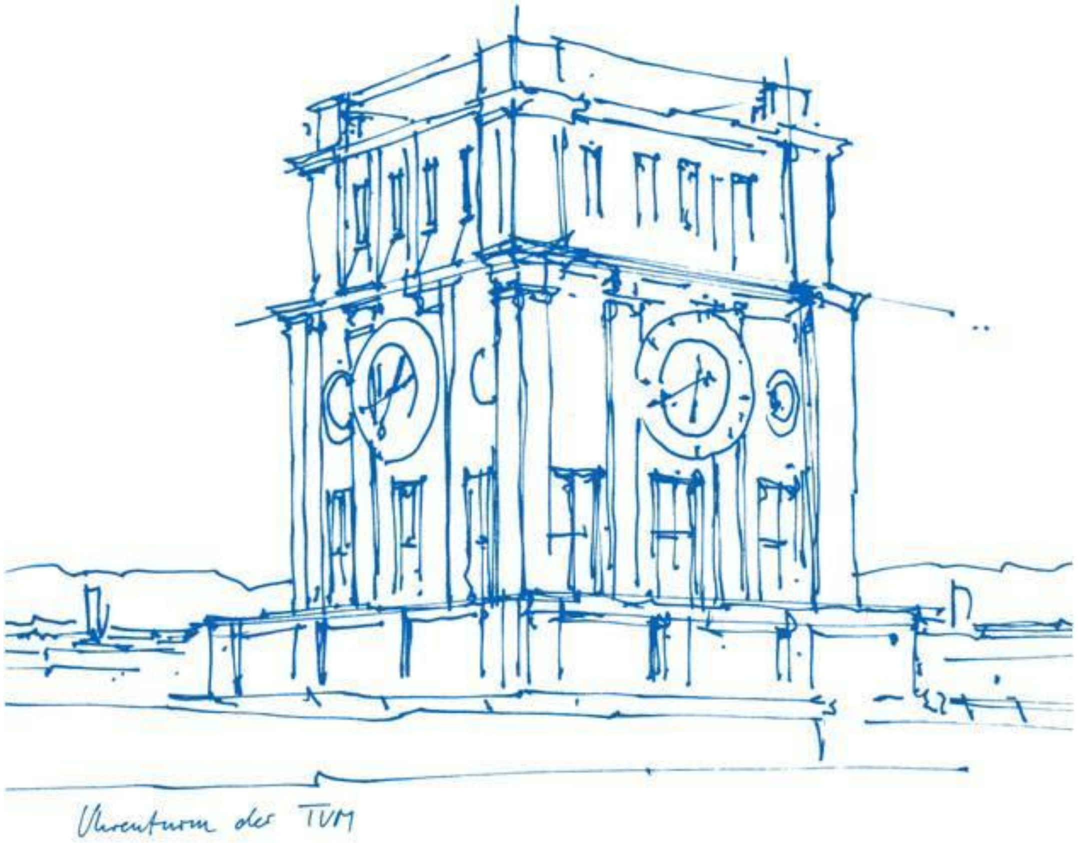
Dealing with missing values

1. Replace by mean should not be used.
2. Procedures "pairwise" and "case by case" exclusion should only be used if completely random.
3. (E)xpectation- (M)aximation algorithm and multiple substitutions can also be used with MAR (random absence).

- Estimation algorithm that replaces data so that the entire information in the record in itself is consistent and maximum plausible.
- Simulations have shown that the MCAR and MAR EM algorithm provides reliable results when up to 30% of values in individual variables are missing and sample sizes are large enough
- An extension of the EM algorithm is multiple substitution.
- The Missing Value Analysis (MVA) add-on module in SPSS allows missing data diagnosis and replacement with EM algorithm.
- EM-Algorithm and Imputation can also be used if MAR assumption is violated and the data are NMR. The more features are collected that correlate with variables with missing information, and the more correlated they are, the more certain it can be assumed that the condition "randomly missing" (MAR) is approximate.

Wirtz (2004)

Descriptive Statistics



What are Descriptive statistics?

- Descriptive statistics describe datasets through **tables and graphs** (→ Data visualization)
- Descriptive statistics **cannot test hypotheses**

Typical charts:

- Pie chart
- Bar chart

→Only use charts, if it‘s easier for the reader to understand a chart than the text

Which descriptive statistics do you know?

Arithmetic mean: average

- Sum all values and divide through number of observations, e.g. mean age in the seminar
- Problem: Outliers have strong effects

Median: medium value

- Sort all values from low to high and find the middle value, i.e., 50% of the values are smaller and 50% of the values are higher
- Advantage: relatively robust against outliers

x%-Quantile (frequently 25%- and 75%-Quantile)

- As median, but at 25% or 75% of all values

Mode

- The most frequent value

Scale	**Descriptive statistic**
Nominal	Mode
Ordinal	Mode, Median
Metric	Mode, Median, arithmetic Mean

Which descriptive statistics do you know?

Range: Difference between smallest and largest value

- Problem: Outliers have strong effects

Distribution: Average deviation from the mean

- Variance: average squared deviation of values from mean / Var = SD^2
- Standard deviation: deviation from mean within the sample / SD = √Var
- Standard error: deviation from mean within the population

Skewness (unimodal distribution):

- Right skewed: Mean > Mode
- Left skewed: Mean < Mode
- symmetric: Mean = Mode

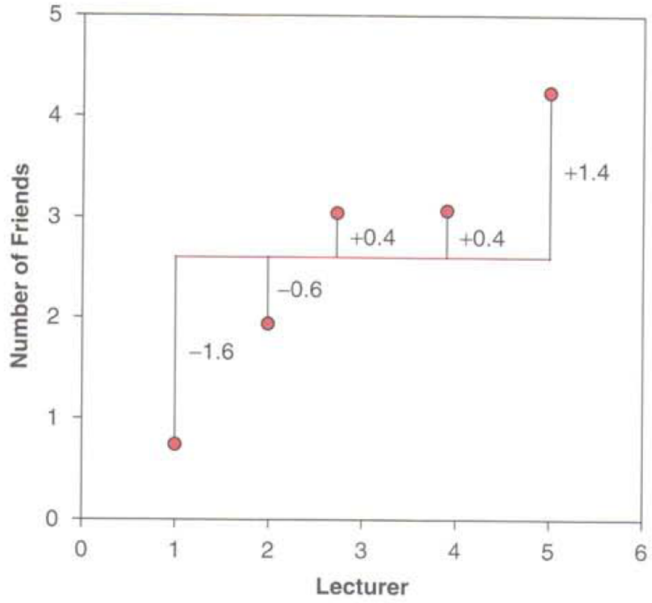
Scale	**Descriptive Statistics**
Nominal	Skewness
Ordinal	Range, Skewness
Metric	Range, Skewness, Distribution

Typical abbreviations

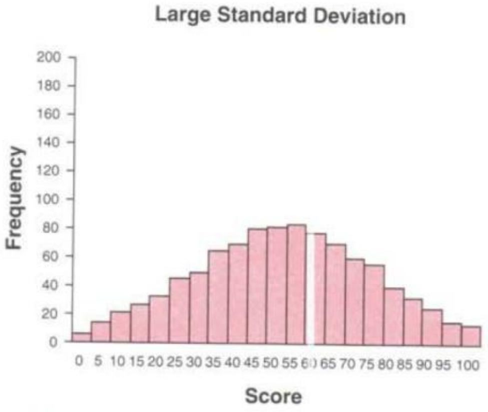
M	**mean**
Med	median
Mod	mode
SD	**standard deviation**
SE	**standard error**
MR	range
N	sample size
Df	degrees of freedom (Field, 2013, p. 37: „The number of observations that are free to vary.“)

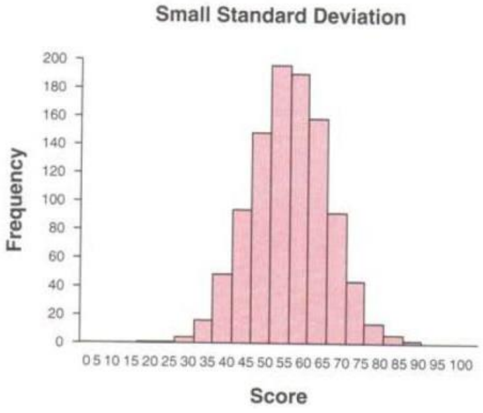
Examples of descriptive statistics

Mean



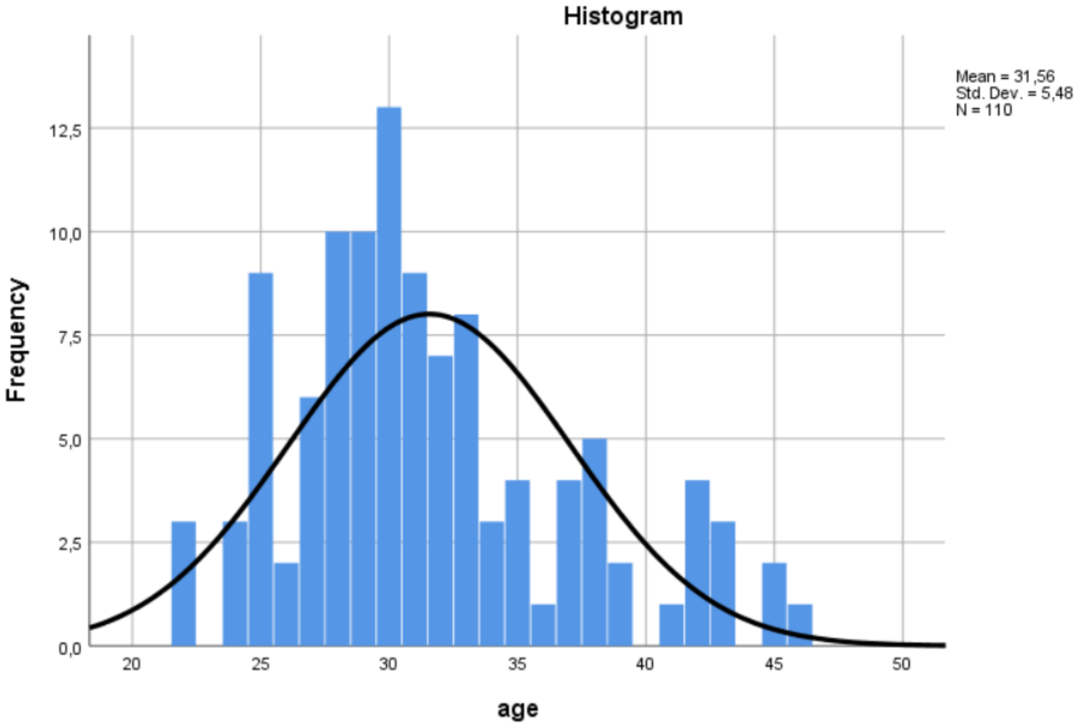
Standard Deviation





Field (2013)

Analyze distribution of variables



Check for normal distribution of variables: Kolmogoroff-Smirnov-Test

Test for normal distribution

Kolmogorov-Smirnov-Test & Shapiro-Wilk-Test:

- Testing the H0 (Null hypothesis): Variable is normally distributed. (!!!)
- If KS-Test significant (p < 0,001 / 0,01 / 0,05: reject H0 → Data is not normally distributed

→ use non-parametric tests (non-parametric tests are more conservative).

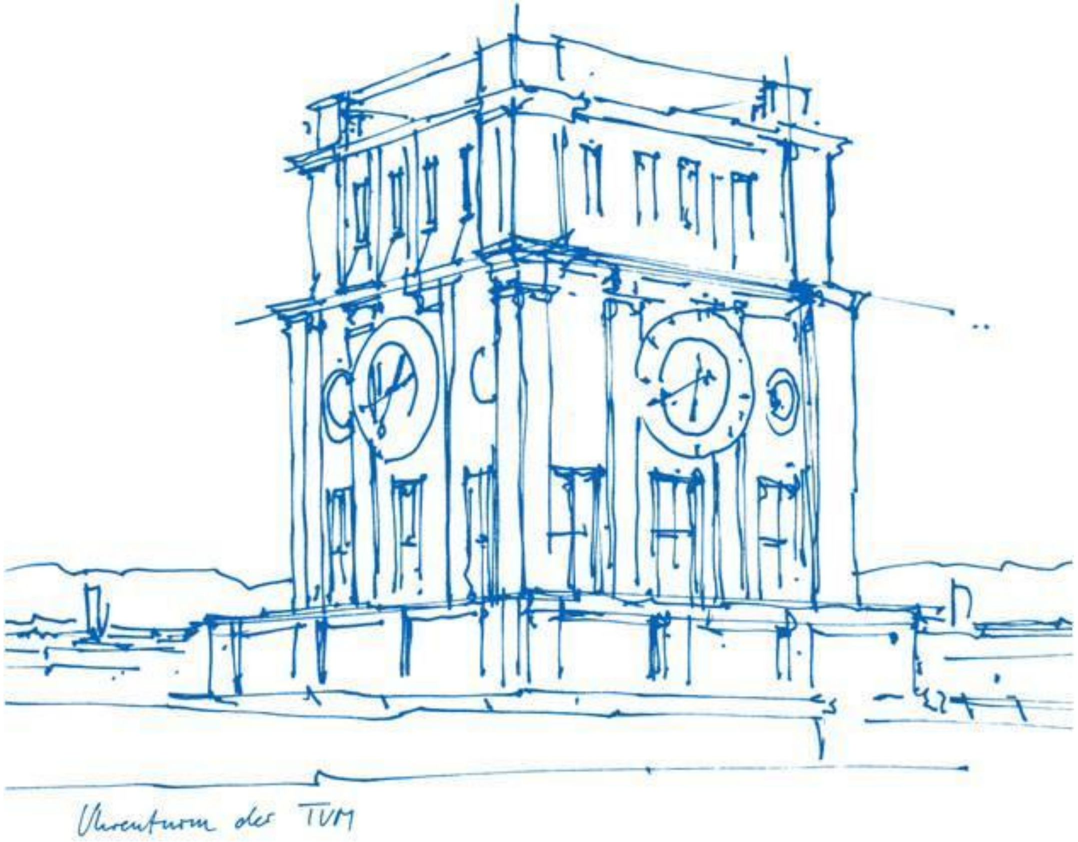
→ If this data is from a pretest, you can change your data collection strategy!

Descriptive Statistics: Reporting

Examples:

- How large is the sample size?
- What is the gender distribution?
- What is the average age? (*M* = XX, *SD* = XX)
- What is the age range? (*Min* = XX, *Max* = XX)
- What is the educationl background of the managers in your sample? (XX% have secondary schooling, XX% a Bachelor degree, XX% a Master degree, and XX% a PhD)

Inferential statistics



What is inferential statistics?

- With inferential statistics, you can make **statements about a population.**
- Inferential statistics is **based on probability theory.**
- What is probabilistic theory? E.g., lotteries, dice, draw from an urn

Uni-, bi- and multivariate procedures

Univariate Analysis

Frequency distributions
-absolute
-relative
-cumulative
-class building

Location parameter
-Modus
-Median
-arithmetic mean

Scattering parameters
-span
-average deviation
-standard deviation
-Variance

Concentration parameters
-Lorenz curve
-Gini-Coefficient
-Concentration coefficient
-Herfindahl-Index H
-Exponential Index E

Bivariate Analysis

Contingency analysis
-Cross tables
-Edgy frequency
-Conditional frequency
-Scatter plot
-Chi-square
-Contingency coefficient

Correlation analysis
-Pearson-
correlation coefficient
-Spearman correlation coefficient
-contingency coefficient

Regression analysis
-Linear regression
-Non-linear
regression

Multivariate Analysis

Method of structural testing

-Multivariate regression analysis
-Variance analysis
-Discriminant analysis
-Contingency analysis
-analysis
-Conjoint Measurement

Method of structure discovery

-Factor analysis
-Cluster analysis
-Multidimensional
Scaling

Backhaus/Erichson/Plinke/Weiber (2003)

Overview of statistical tests - example 1

Measurement level of dependent variable	Sample (2 groups)		Sample (k groups)		Relationship test
	independent	dependent	independent	dependent	
Nominal	Chi^2 test for two groups	McNemar test	Chi^2 test for k groups	Cochran-Q-test	Contingency test (Phi-coefficient, Cramer's V) **Regression analysis**
Ordinal	Mann-Whitney-U-test	Wilcoxon-test	Kruskal-Wallis-H-test	Friedman test	Rank correlation coefficient (Spearman) **Regression analysis**
Continuous	t-test	t-test	Univariate analysis of variance (ANOVA)	Univariate analysis of variance (ANOVA)	Product-moment correlation coefficient (Pearson) **Regression analysis**
Tests for ...	**Differences**				**Relationships**

Dependent vs independent sample

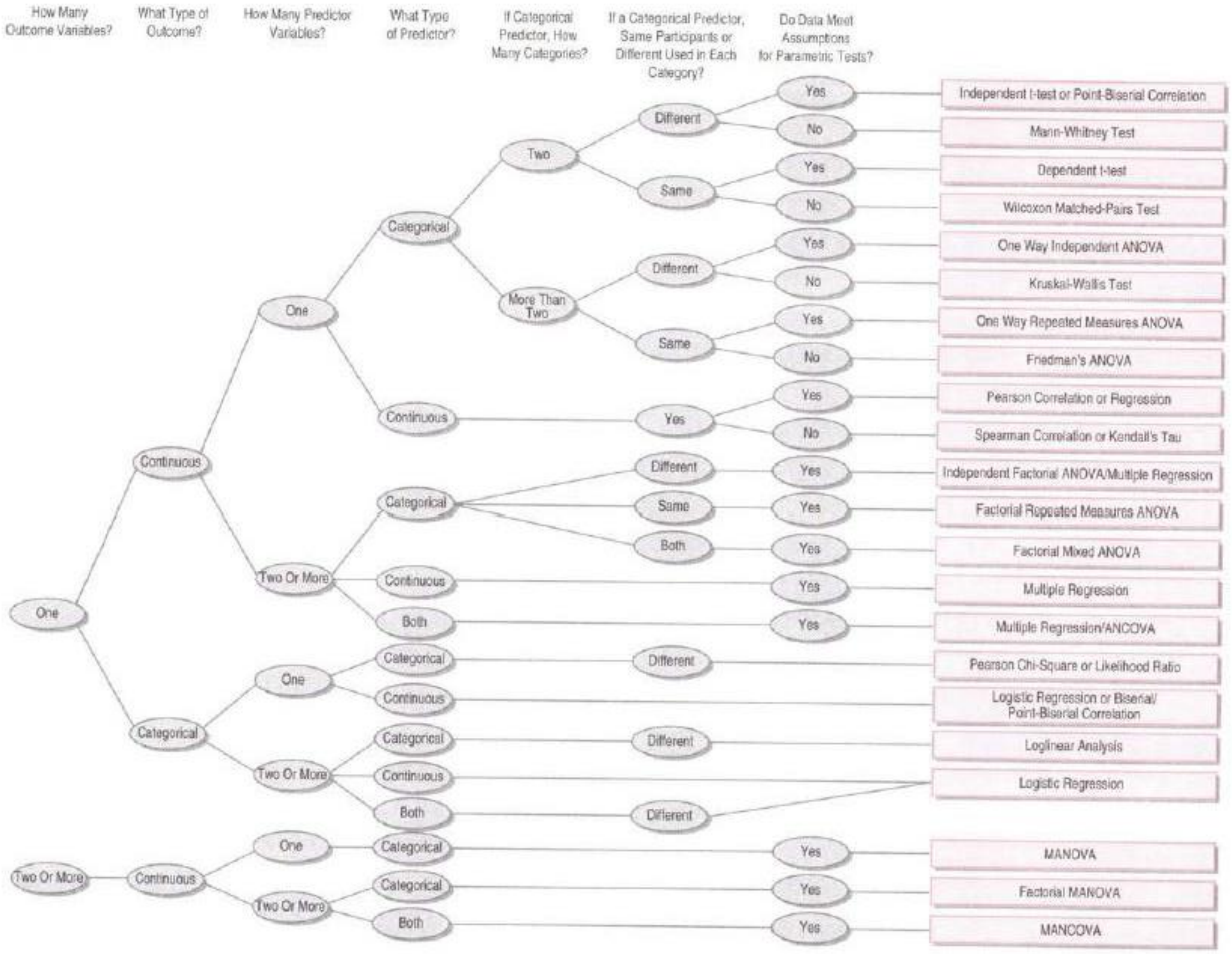
Dependent samples compare samples in which variables are paired

- E.g., variables that are measured at two points in time (longitudinal study: body weight before and after a diet).

Independent samples compare two independent groups based on a variable

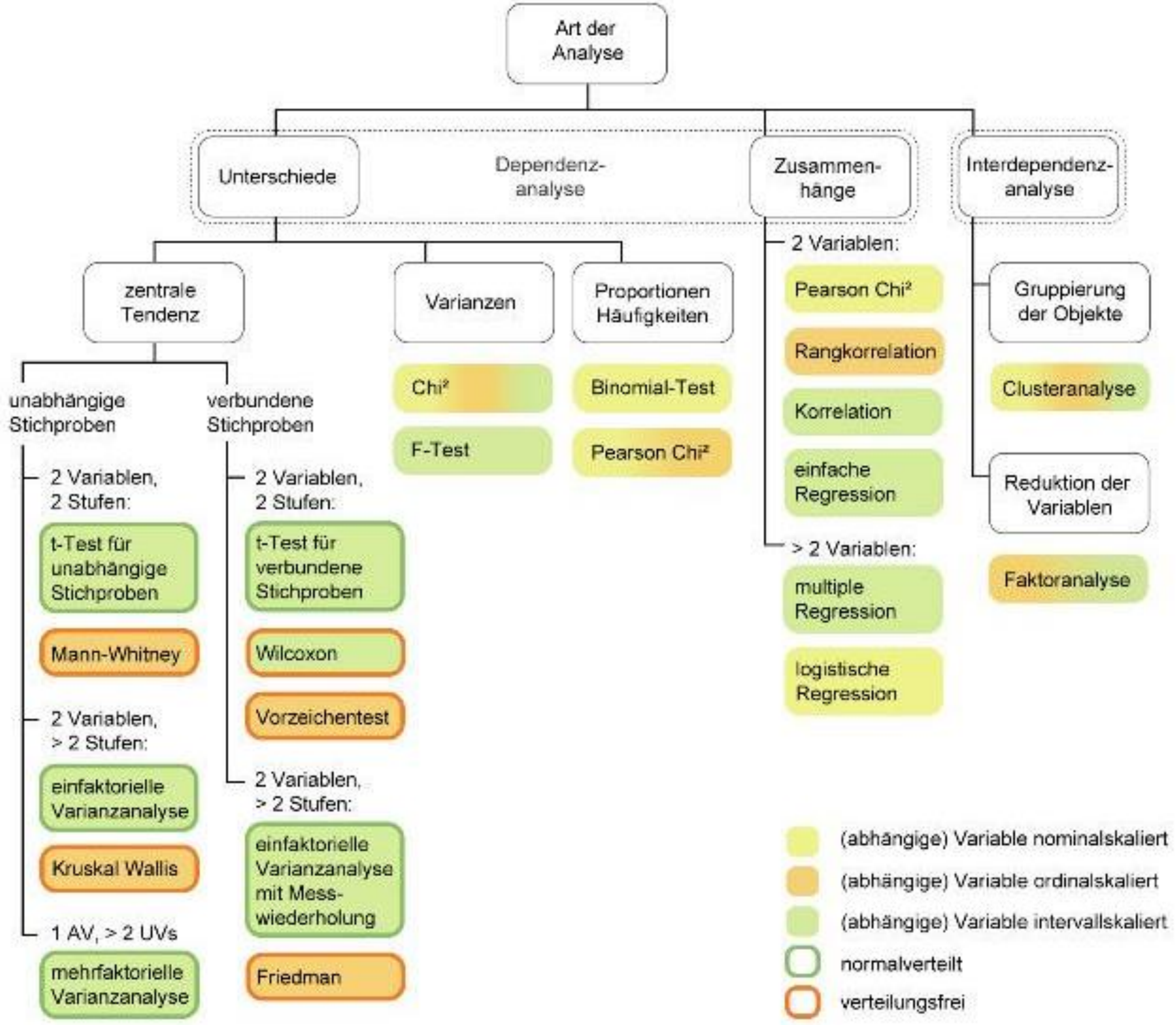
- E.g., A variable (neuroticism) is compared in two different groups (cross-sectional study: neuroticism in public servants and nuns).

Overview of statistical tests - example 2


How Many Outcome Variables?
What Type of Outcome?
How Many Predictor Variables?
What Type of Predictor?
If Categorical Predictor, How Many Categories?
If a Categorical Predictor, Same Participants or Different Used in Each Category?
Do Data Meet Assumptions for Parametric Tests?
One
Continuous
One
Categorical
Two
Different
Yes
Independent t-test or Point-Biserial Correlation
No
Mann-Whitney Test
Same
Yes
Dependent t-test
No
Wilcoxon Matched-Pairs Test
More Than Two
Different
Yes
One Way Independent ANOVA
No
Kruskal-Wallis Test
Same
Yes
One Way Repeated Measures ANOVA
No
Friedman's ANOVA
Continuous
Yes
Yes
Pearson Correlation or Regression
No
Spearman Correlation or Kendall's Tau
Two Or More
Categorical
Different
Yes
Independent Factorial ANOVA/Multiple Regression
Same
Yes
Factorial Repeated Measures ANOVA
Both
Yes
Factorial Mixed ANOVA
Continuous
Yes
Multiple Regression
Both
Yes
Multiple Regression/ANCOVA
Categorical
One
Categorical
Different
Pearson Chi-Square or Likelihood Ratio
Continuous
Logistic Regression or Biserial/ Point-Biserial Correlation
Two Or More
Categorical
Different
Loglinear Analysis
Continuous
Logistic Regression
Both
Different
Two Or More
Continuous
One
Categorical
Yes
MANOVA
Two Or More
Categorical
Yes
Factorial MANOVA
Both
Yes
MANCOVA

Field (2013)

Overview of statistical tests - example 3



Art der Analyse

Unterschiede

Dependenz-analyse

Zusammen-hänge

Interdependenz-analyse

zentrale Tendenz

Varianzen

Proportionen Häufigkeiten

unabhängige Stichproben

verbundene Stichproben

2 Variablen, 2 Stufen:

t-Test für unabhängige Stichproben

Mann-Whitney

2 Variablen, > 2 Stufen:

einfaktorielle Varianzanalyse

Kruskal Wallis

1 AV, > 2 UVs

mehrfaktorielle Varianzanalyse

2 Variablen, 2 Stufen:

t-Test für verbundene Stichproben

Wilcoxon

Vorzeichentest

2 Variablen, > 2 Stufen:

einfaktorielle Varianzanalyse mit Mess-wiederholung

Friedman

Chi²

F-Test

Binomial-Test

Pearson Chi²

2 Variablen:

Pearson Chi²

Rangkorrelation

Korrelation

einfache Regression

> 2 Variablen:

multiple Regression

logistische Regression

Gruppierung der Objekte

Clusteranalyse

Reduktion der Variablen

Faktoranalyse

🟨 (abhängige) Variable nominalskaliert

🟧 (abhängige) Variable ordinalskaliert

🟩 (abhängige) Variable intervallskaliert

○ normalverteilt

○ verteilungsfrei

http://www.methodenberatung.uzh.ch/index.html

Two types of hypotheses

Differences

There is a difference of at least two groups in one or more independent variables.

⇒ Group 1 is different than group 2 in variable X.

Hypothesis testing through analyzing means and variances.

Relation

There is a relation between at least two variables.

⇒ The higher variable x, the higher (lower) variable y.

Hypothesis testing through (correlation and) regression analysis.

Statistical hypothesis

Null hypothesis (equality): There is **no** difference between unit leaders and department leaders in their job satisfaction. (null difference / relationship → non significant result)

Alternative hypothesis (assumption to be tested): There is a difference between unit leaders and department leaders in their job satisfaction. (difference / relationship → significant result).

The goal of an empirical study is testing the probability of these two hypotheses and which hypothesis can be provisionally accepted or rejected.

⇒ Formulate the alternative hypothesis in your study, because it is easier to falsify an assumption than to confirm it.

Significance level

- The **significance level or α-error** level determines what is **considered likely and what is unlikely**; one speaks of a high level of significance when the α-error expectation is small.
- If the result is **significant, we decide for the rejection (!) of H_0**
  The probability that this decision is wrong is α.
- If the result is not significant, we decide to accept (!) The H0. This means that both groups with the expected error β do not differ by the amount specified by the effect size
- Classical significance testing only checks how well the data is compatible with H0 and not which of the two hypotheses H0 or H1 is more compatible with the data

→ **Statistical significance says nothing about meaning / real-life significance**

Wirtz/Nachtigall (2004), Hussey/Jain (2002)

Interpretation of significances

*** = Probability of error that there is a sample relationship that does not occur in the population

level of significance = **5%** = out of 100 results = 5 only by chance significant.

level of significance = **1%** = 1 out of 100 results happened to be significant.

level of significance= **0,1%** = 1 out of 1000 results coincidentally significant.

Typical expressions of significance:

ns	p > 0.10	(> 10%)	not significant
+	p < 0.10	(< 10%)	marginal significant
*	p < .05	(< 5%)	significant
**	p < .01	(< 1%)	very significant
***	p < .001	(<0,1%)	highly significant

Wirtz/Nachtigall (2004), Hussey/Jain (2002)

What does a correlation/regression coefficient of .553*** mean?

Two different messages: (1) *** (2) .553

(1) Findings are highly significant - probability of finding, although in reality no correlation is less than 0.001% (1 in 1000) says something about whether findings were randomly found in sample, so whether we may assume findings in the population.

(2) .553 says something about the strength of the context (effect size).

Does the result apply to the population?

→ The answer is in the stars ***

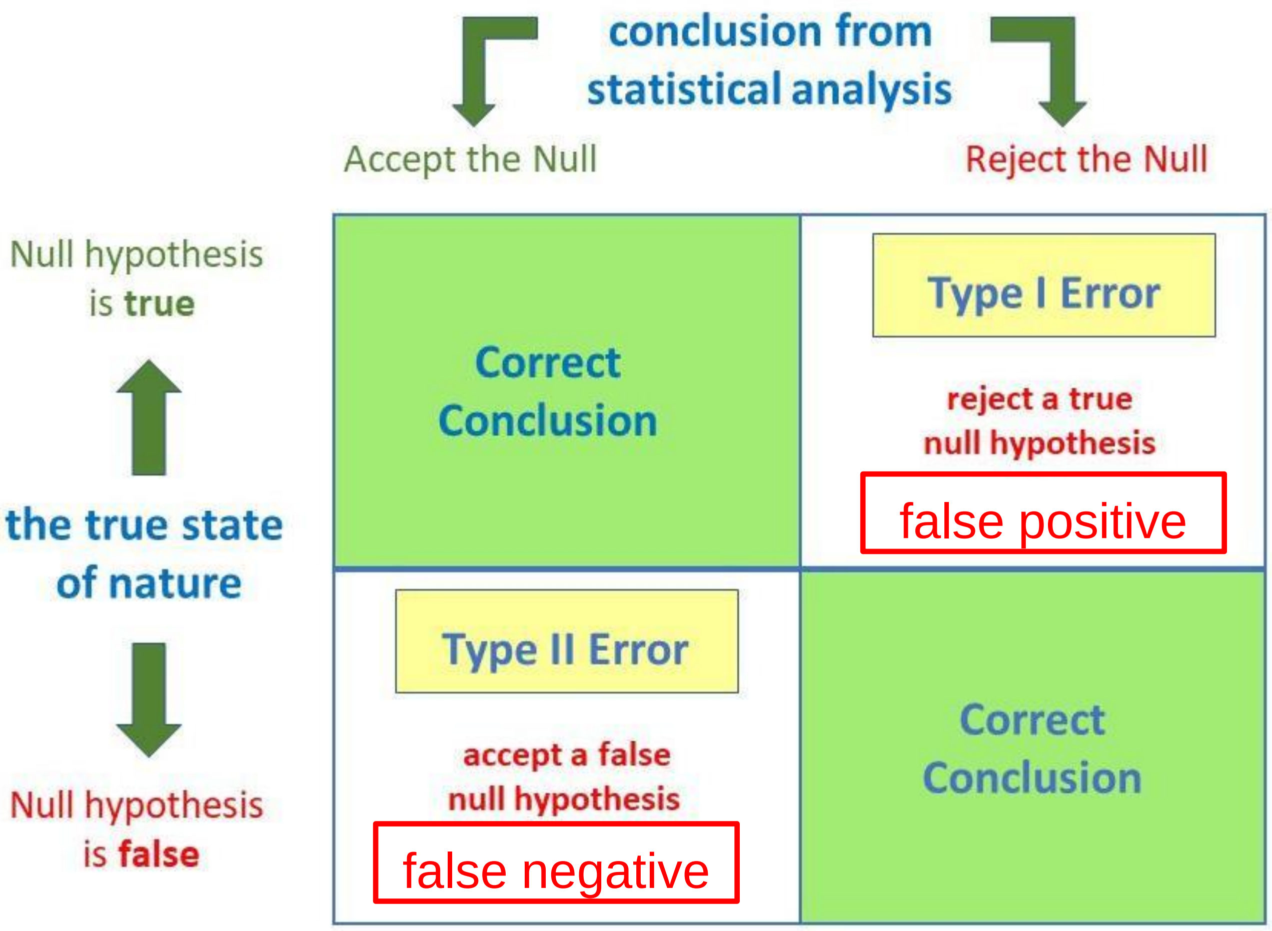
→ Relationship exists and is highly likely to be found in the population.

Significance versus effect size

- In order to have a uniform scale with which mean differences from different studies can be compared with different samples and measuring instruments, one can calculate **effect size measures**
- Effect size is a **measure of comparison that shows the size of the difference or relationship, regardless of scale and variance**
- The significance of a result makes a statement as to whether a found difference or coincidence occurs by chance. **The statement "The more significant a result, the more meaningful it is" is false.**
- The **effect size indicates how big and therefore how meaningful a difference or relationship is** → Under the same conditions an effect becomes all the more likely the bigger it is.
- The **statistical significance of a result is strongly determined by the size of the examined sample**! But the **effect size is largely independent of the size of the examined sample** and thus less **dependent on the realized empirical procedure.**
- Effect sizes can also be compared between different examinations (which does not apply to p).

Wirtz/Nachtigall (2004), Hussey/Jain (2002)

Type 1 and 2 error



conclusion from statistical analysis

the true state of nature

	Accept the Null	Reject the Null
Null hypothesis is true	Correct Conclusion	Type I Error reject a true null hypothesis false positive
Null hypothesis is false	Type II Error accept a false null hypothesis false negative	Correct Conclusion

$\alpha - error$

Covid-19 test is **positive**, although you are **not sick.**

β - error

Covid-19 test is **negative**, although you are **sick.**

What is worse? Alpha or Beta error?

→ Does the content of the question arise? Which error would be "worse"?

Example: misjudgment of the company situation

Alpha Errors: Risk of getting business consultants for a problem that does not exist

Beta Error: Risk of overlooking a problem

→ Where does the lack of test strength come from?

Too small samples, unfavorable measurement conditions

→ What can be the consequences of a lack of test strength?

- existing population effects are not detected
- Null effects can not be interpreted!

Wirtz/Nachtigall (2004), Hussey/Jain (2002)

Statistical Power

Stronger alpha level → less errors 1st kind, more mistakes 2nd kind, less power

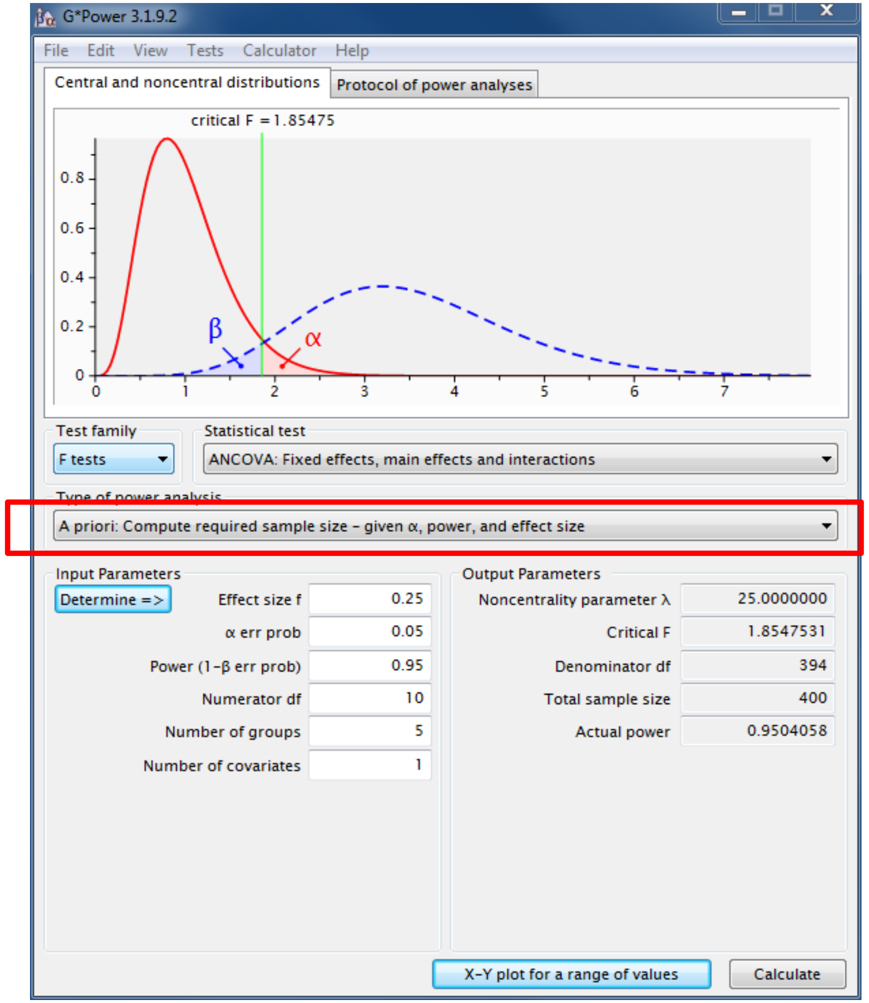
Smaller effect → more errors 2nd kind, more mistakes 1st kind, less power

Bigger effect → less errors 1st kind, less mistakes 2nd kind, more power

Bigger sample → less errors 2nd kind, more power

Wirtz/Nachtigall (2004), Hussey/Jain (2002)

Power analysis


G*Power 3.1.9.2
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
0.8
0.6
0.4
0.2
0
0 1 2 3 4 5 6 7
β
α
Determine =>
X-Y plot for a range of values
Calculate

critical F = 1.85475

Test family: F tests

Statistical test: ANCOVA: Fixed effects, main effects and interactions

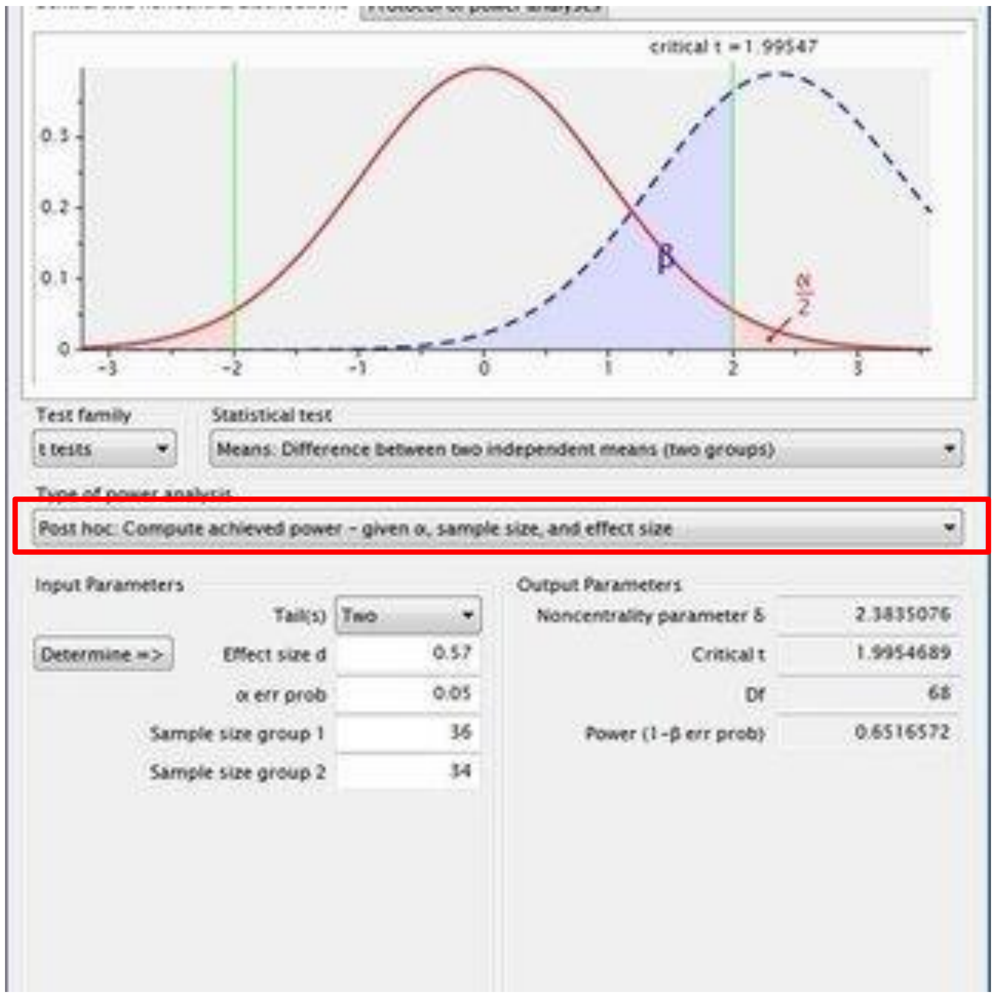
Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Effect size f: 0.25
α err prob: 0.05
Power (1–β err prob): 0.95
Numerator df: 10
Number of groups: 5
Number of covariates: 1

Output Parameters

Noncentrality parameter λ: 25.0000000
Critical F: 1.8547531
Denominator df: 394
Total sample size: 400
Actual power: 0.9504058



critical t = 1.99547

Test family: t tests

Statistical test: Means: Difference between two independent means (two groups)

Type of power analysis: Post hoc: Compute achieved power – given α, sample size, and effect size

Input Parameters

Tail(s): Two
Effect size d: 0.57
α err prob: 0.05
Sample size group 1: 36
Sample size group 2: 34

Output Parameters

Noncentrality parameter δ: 2.3835076
Critical t: 1.9954689
Df: 68
Power (1-β err prob): 0.6516572

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods, 39, 175-191.

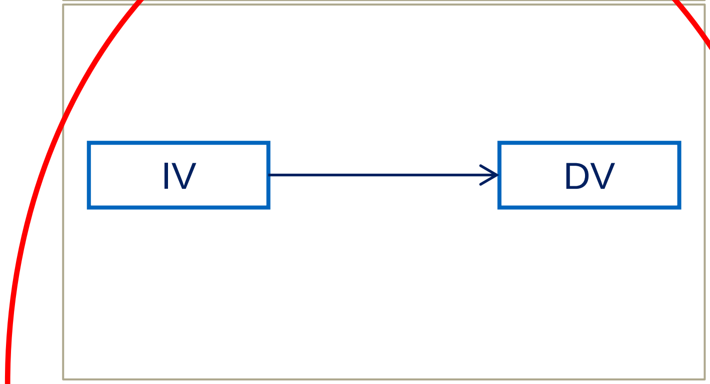
Criticism of NHST (null hypothesis significant testing)

- The beta error is ignored
- Significance tests promote the use of inaccurate hypotheses (Gigerenzer, 1998)
- Significance tests do not say anything about the probability of the alternative hypothesis
- With large samples you get every & each effect significantly
- Significance does not say anything about whether an effect is unimportant or important
- If Ho is rejected, does not say anything about establishing a theory.

Wirtz/Nachtigall (2004), Hussey/Jain (2002)

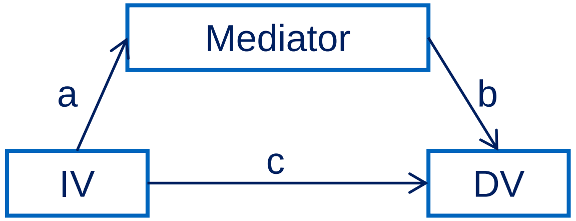
Which types of research models are there?

Simple model



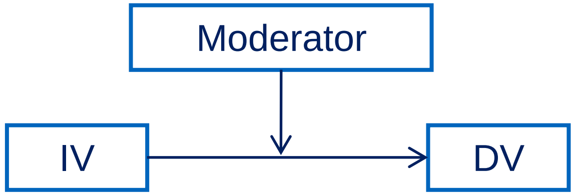
IV influences DV

Mediator



Are there third variables through which the IV influences the DV?

Moderator



Are there third variables that qualify the relationship between IV and DV?

Correlation

Correlation

- Quantitative measurement to describe linear relationships
- Correlation is characterized by correlation coefficients which can be between -1 and +1

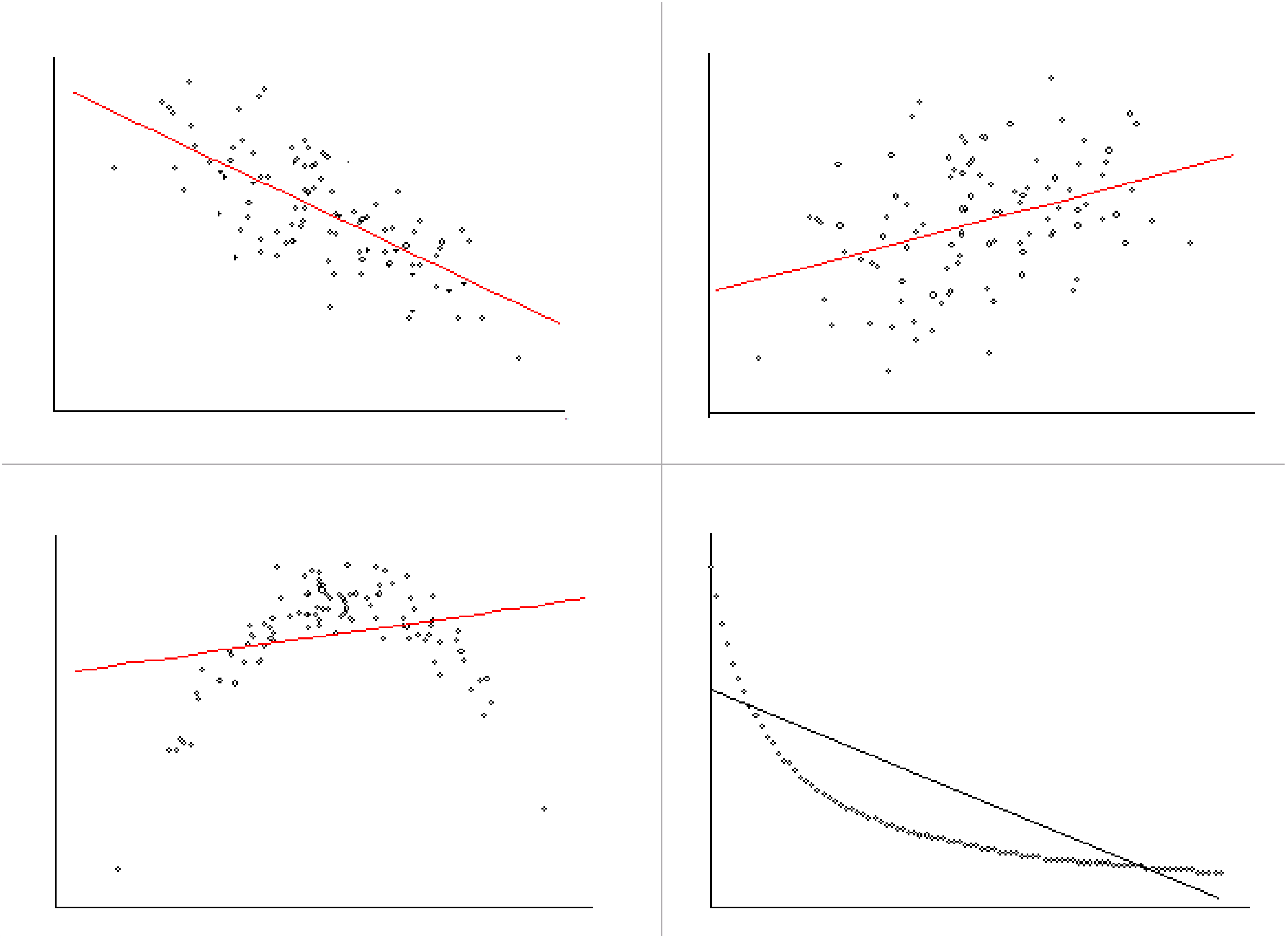
Positive correlation: if a person is above average in variable A, then she is also above average in B.

Example: height and weight

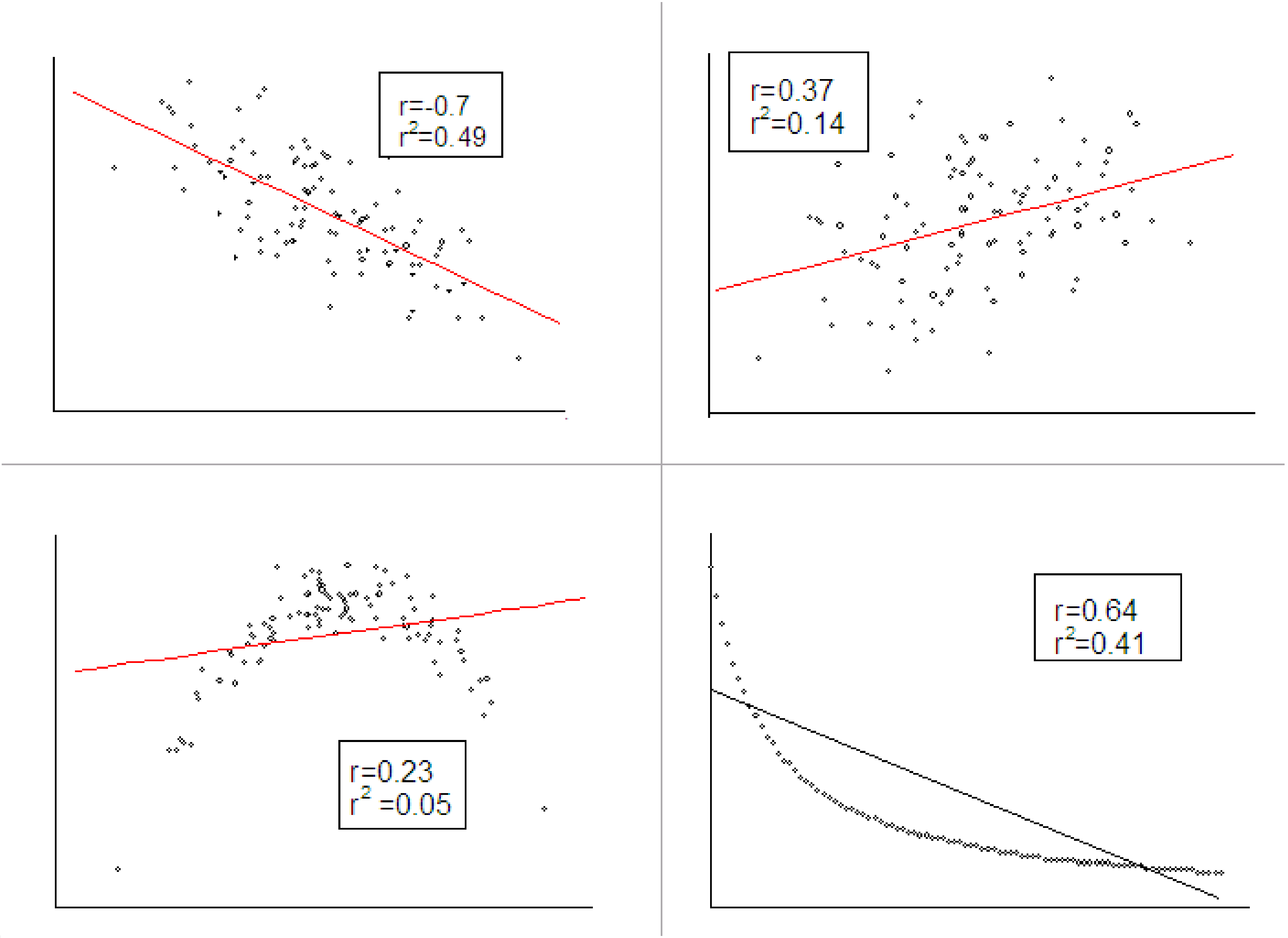
Negative correlation: if a person is above average in variable A, then she is below average in B.

Example: Age and memory

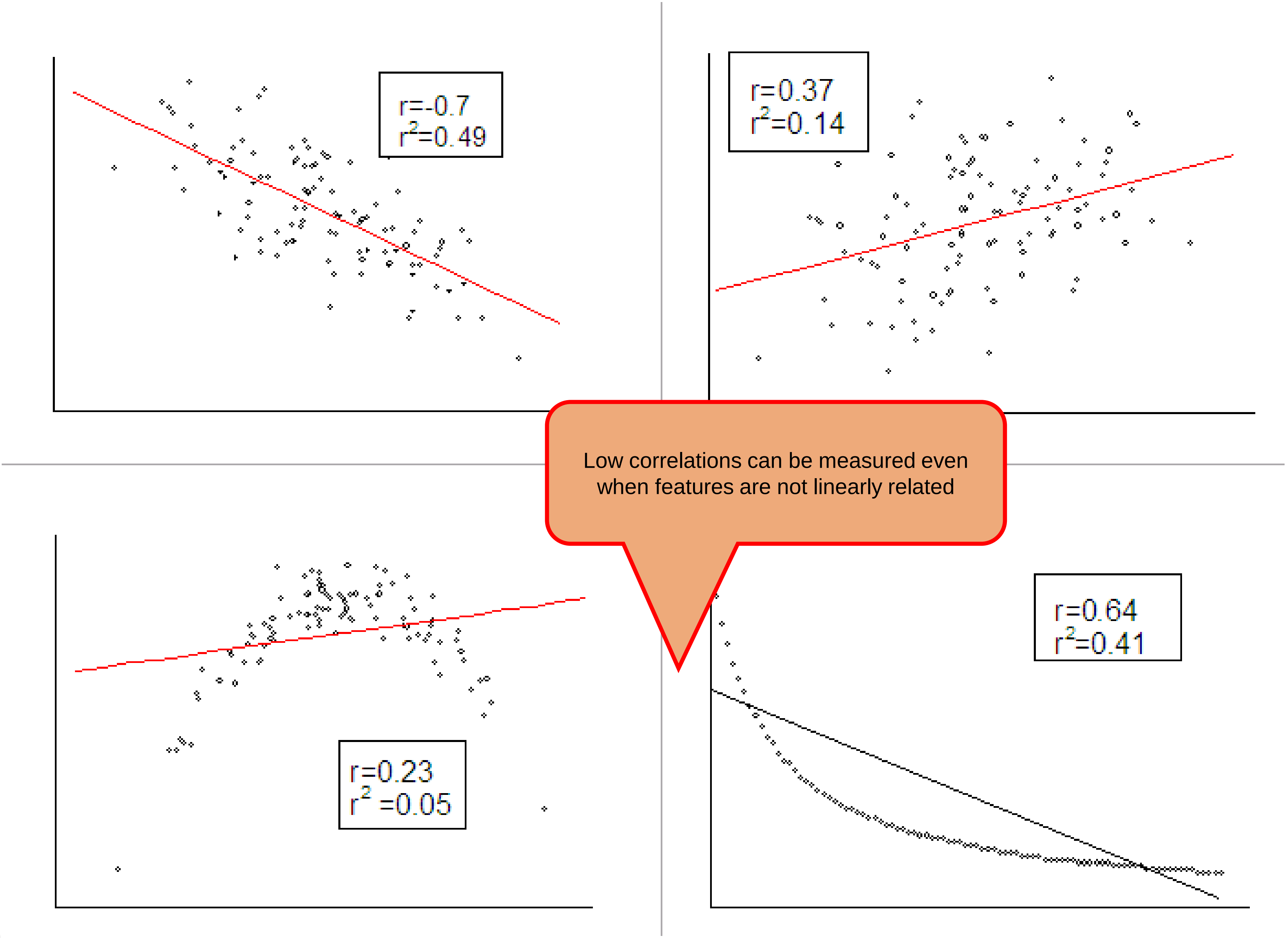
Backhaus/Erichson/Plinke/Weiber (2003)



Wirtz/Nachtigall (2004)



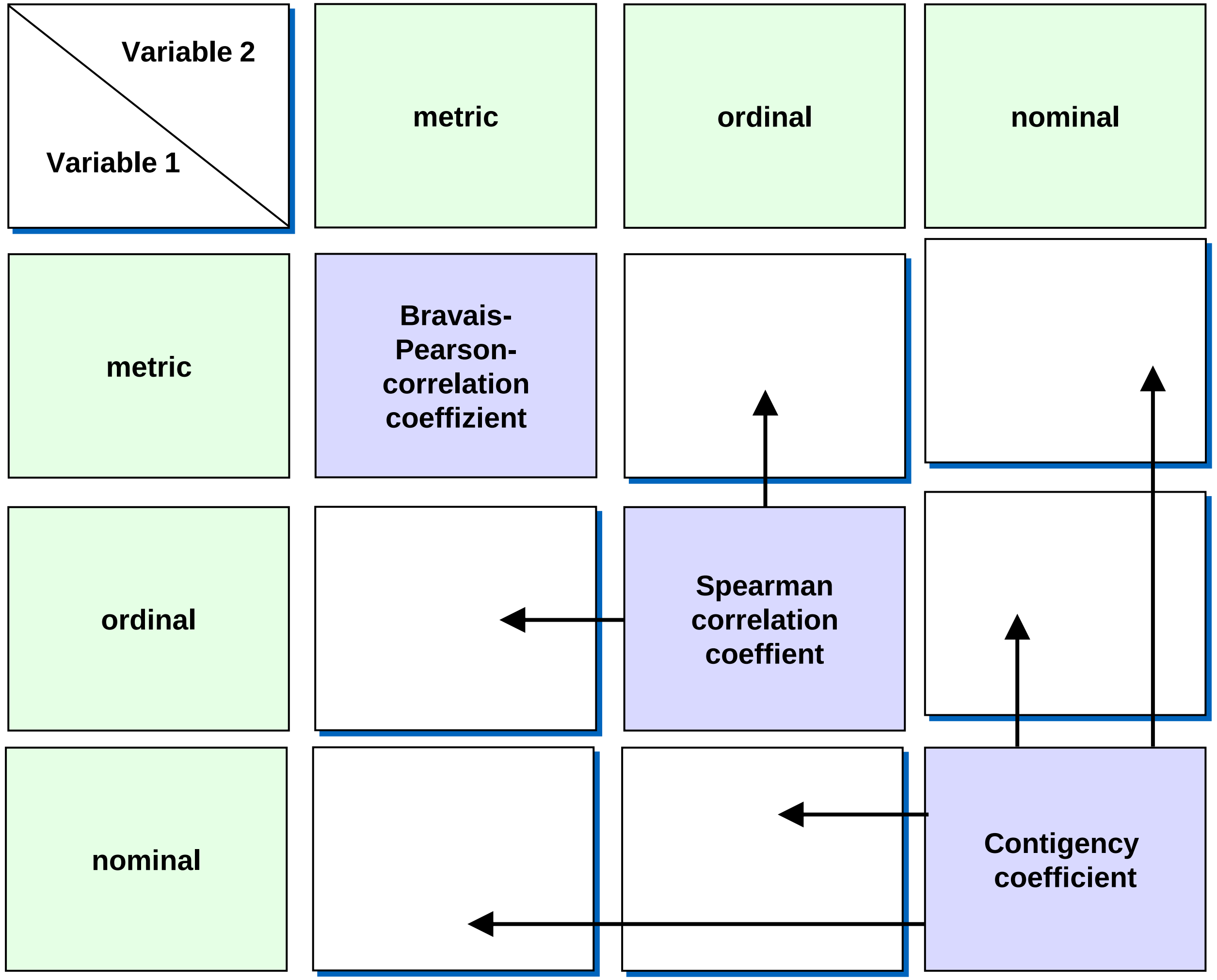
Wirtz/Nachtigall (2004)



Low correlations can be measured even when features are not linearly related

Wirtz/Nachtigall (2004)

Bivariate data analysis



Variable 2 Variable 1	metric	ordinal	nominal
metric	Bravais-Pearson-correlation coeffizient		
ordinal		Spearman correlation coeffient	
nominal			Contigency coefficient

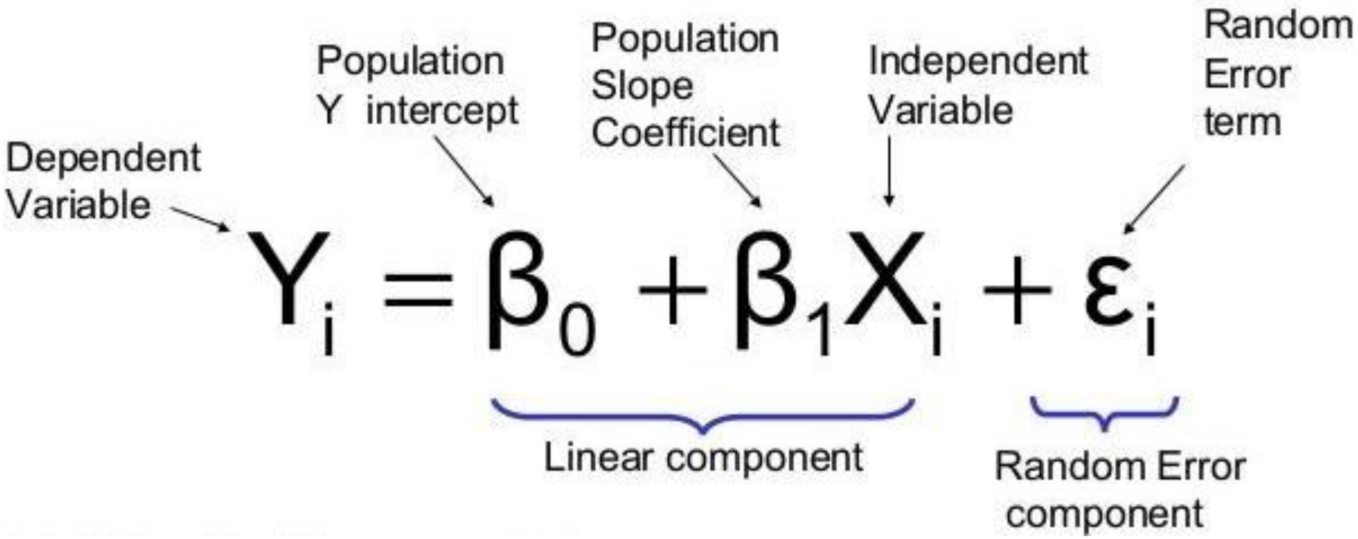
Backhaus/Erichson/Plinke/Weiber (2003)

What is a regression analysis?

Simple regression: Seeks to predict an outcome variable from a single predictor variable.

Multiple regression: Seeks to predict an outcome variable from several predictor variables

The simple regression function:



$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

Field (2013)

Why regression analysis?

Discover or explain relationships between two or more variables.

Values of a DV are estimated or predicted.

- **Root cause**: How strong is the influence of an IV on a DV? (e.g., different marketing strategies → sales)
- **Effect prediction:** How strong does the DV change in dependence of an IV? (e.g., drugs dosis → time to recovery, time until side effects)
- **Time series**: How does a DV change over time ceteris paribus (i.e., keeping everything else similar) ? (e.g., company shares)

Regression analysis complements correlations

- **Correlations** determine only the strength of a relationship (Variable 1 ↔ Variable 2)
- **Regressions** also determines the direction of the relationship (IV → DV).

Direction ≠ causality

Regression: theoretical example

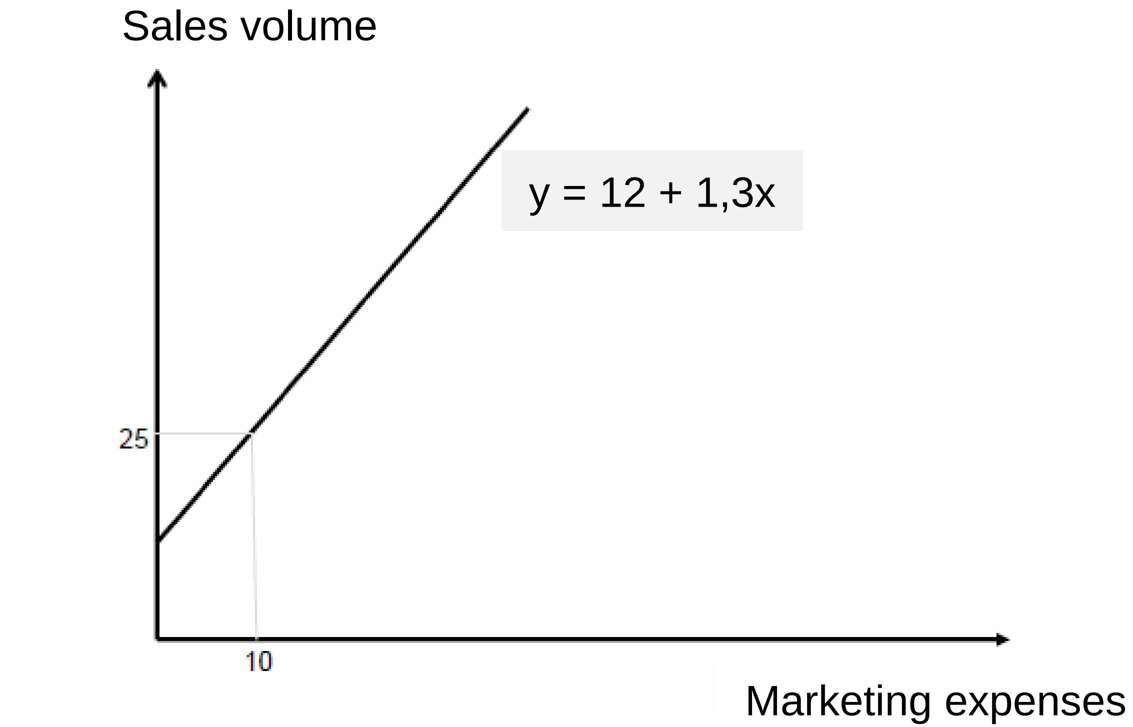
Assume the following function and graph is given:

y = 12 + 1,3x

If Marketing expenses are 10:

y = 12 + 1,3 * 10

then sales volume is y = 25.



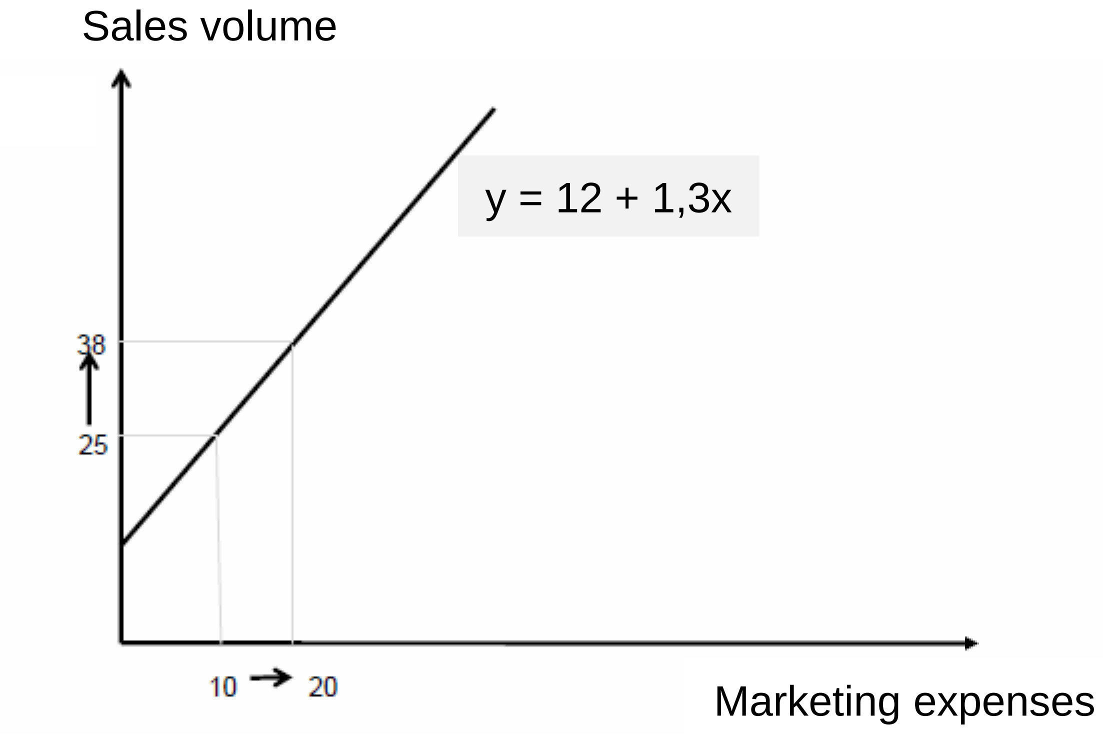
Regression: theoretical example

If we change the marketing expenses from 10 to 20, then

y = 12 + 1,3 * 20

sales volume is y = 38.

By changing marketing expenses from 10 to 20, we predict a change in sales volume of 13 units (from 25 to 38).



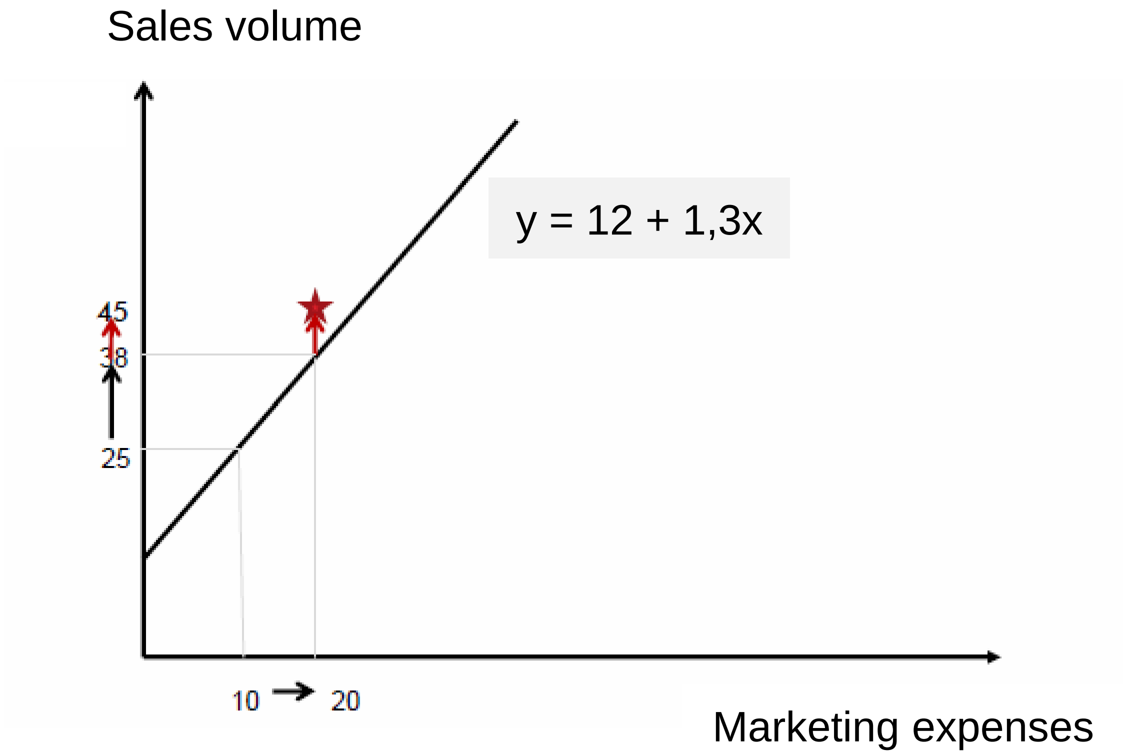
Regression: theoretical example

In reality, if marketing expenses are 20, sales volume is not 38, but 45.

Hence, there is a variance of 20 units (25-45).

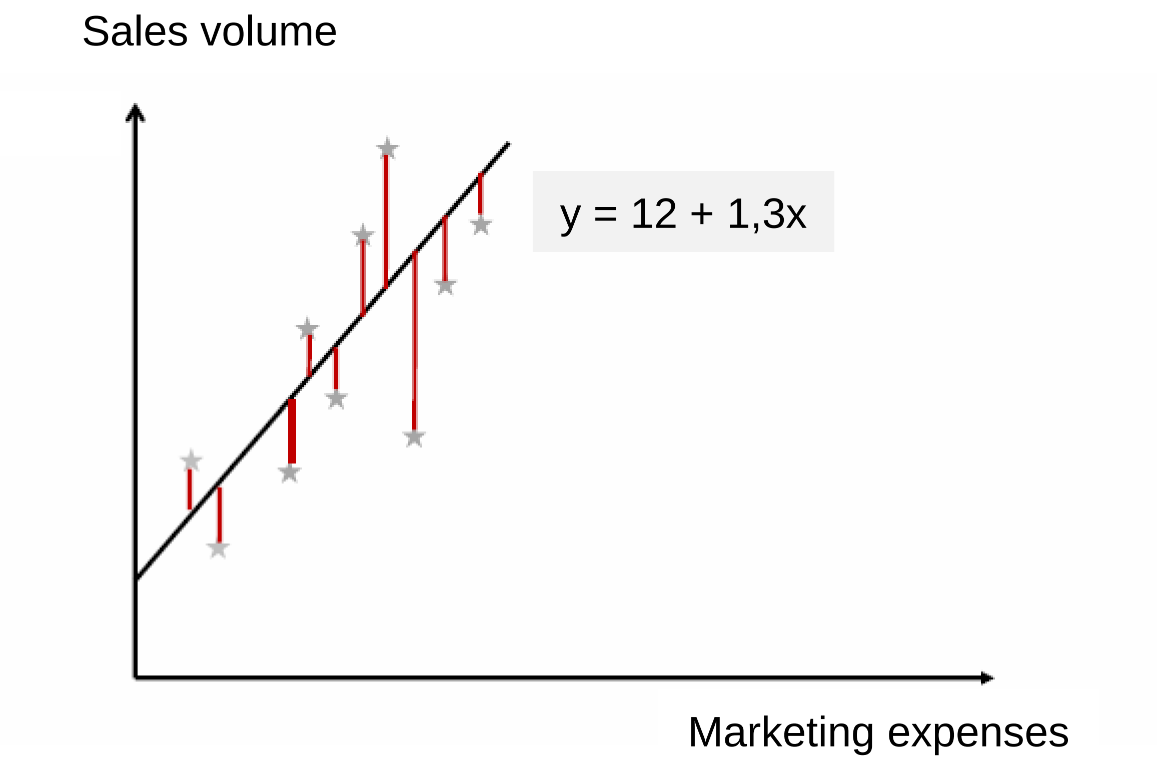
Our model only explains 13 units (= explained variance).

The rest (= residual) of 7 units is not explained variance.



Regression: Ordinary Least Squares (OLS)

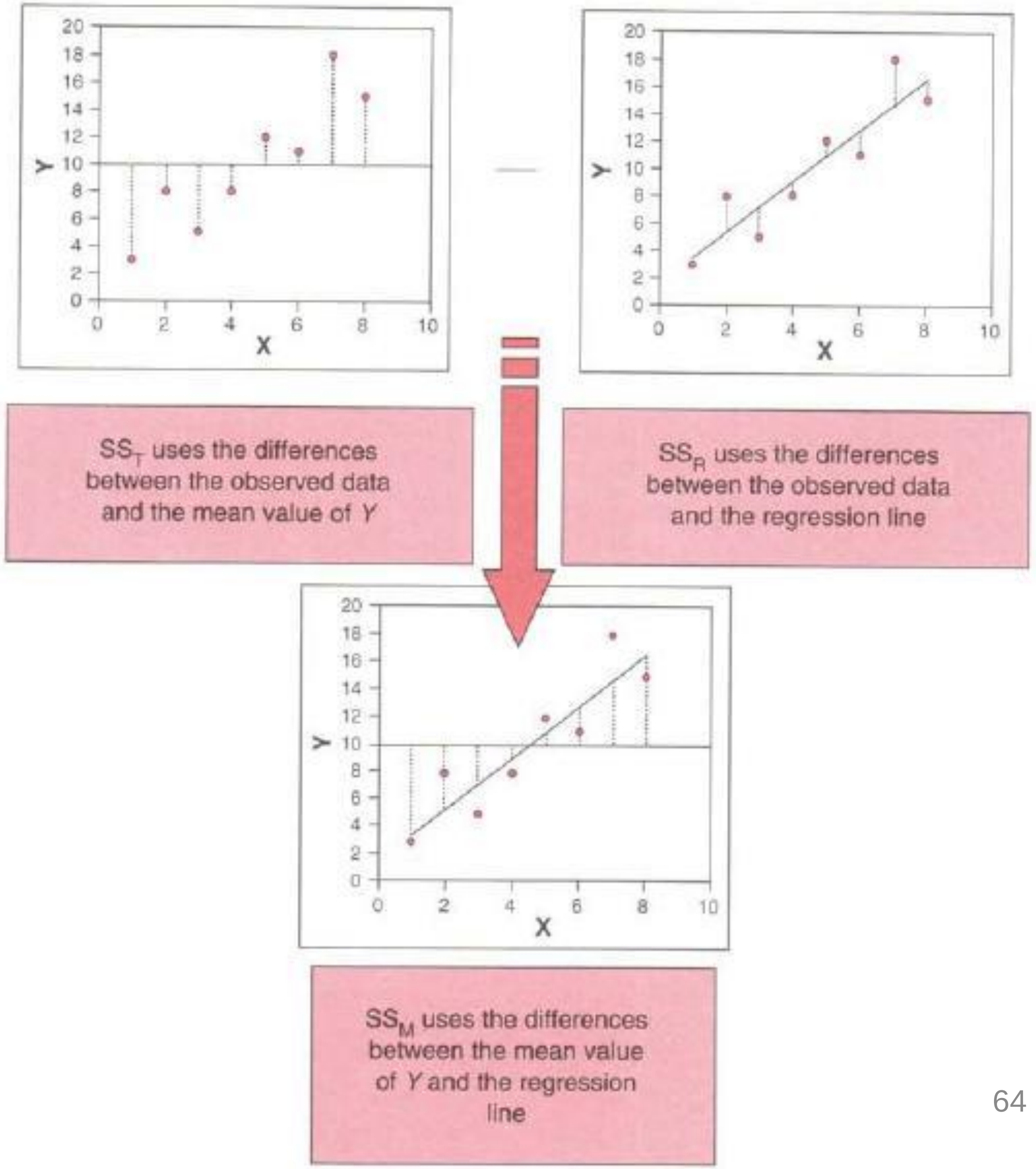
- The smaller the sum of all residuals, the better the model explains the reality (i.e., the better the model can explain sales volume).
- The red lines show deviations between the regression line and the observed values.
- The regression line is the line with the minimum sum of squared deviations (ordinary least squares). It is the line that is „closest“ to all points.



Regression: How does the model fit the data?

$$\boxed{R^2 = SS_M / SS_T}$$

- r tells us how good the linear regression is, i.e., how good can the IV predict the DV.
- R^2 is the squared correlation coefficient.
- $0 < R^2 < 1$.
- $R^2 = 0{,}65$ means that we can explain 65% of the variance of sales volume with marketing expenses as predictor (13 out of 20 units).



SS_T uses the differences between the observed data and the mean value of Y

SS_R uses the differences between the observed data and the regression line

SS_M uses the differences between the mean value of Y and the regression line

Field (2013)

Interpretation of R^2

The total variance of a dependent variable (here: sales volume) can be split in two parts:

1. **Explained variance:** The regression function predicts how y changes when x is varied.

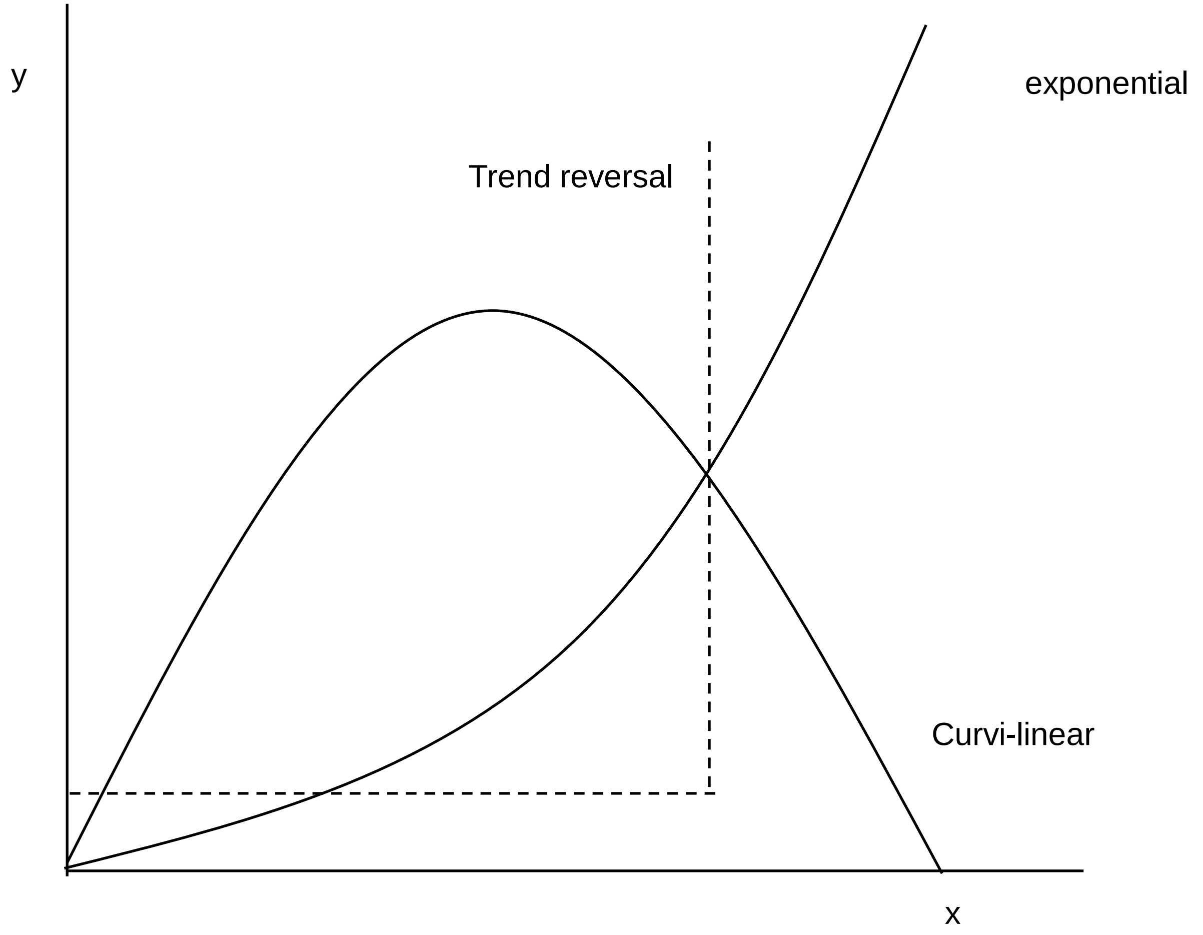
2. **Residual:** But not all points (observations) lie on the regression lie, hence they are the residual variance.

OLS Regression

Preconditions:

1. **Continuous DV:** DV must have continuous measurement.
2. **Normal distribution:** Values for x (IVs) and y (DV) must be approximately normally distributed → check histograms and Kolmogorof-Smirnov test.
3. **Linearity:** The relationship between x and y should be approximately linear → check scatterplots.
4. **Homoscedasticity:** The variance around the **regression** line must be the same for all values of x, i.e., the variance of y must not be wider with higher values of x → check scatterplots.
5. **Independence of data and error (autocorrelation):** all data should be independent from each other, i.e., cases should not correlate / x_4 should not follow from x_3 → Durbin/Watson-Test should not be significant.

Non-linear relationships



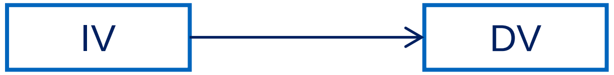
Linear statistical models discover the relationships described above in non-chaos theoretic considerations, non-linear structural equations

Other types of regression analyses

- Multivariate (linear) regression: more than one independent variable
- Logit / Probit Regression: nominal DV, e.g., probability to pass the course (1) or not (0).
- Poisson Regression: count data, e.g., number of creative ideas.
- Tobit Regression: censored variables, e.g., weekly work hours as employee.
- …

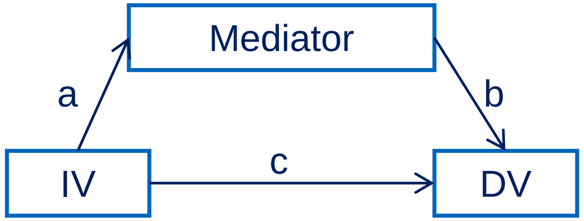
Which types of research models are there?

Simple model



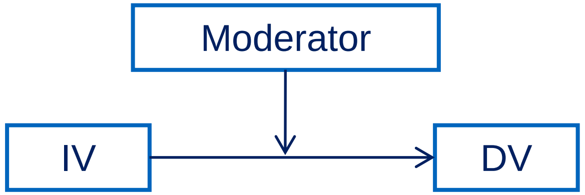
IV influences DV

Mediator



Are there third variables through which the IV influences the DV?

Moderator



Are there third variables that qualify the relationship between IV and DV?

Mediation analysis

1. **Classic step-wise approach of Baron and Kenny (1986)**
   - c-path is significant → a-path is significant → b-path is significant → c‘-path (when mediator is included in the regression) is no longer significant
2. **Sobel**
   - Calculate ratio compared to the critical value from the standard normal distribution
3. **Structural equation modeling**
4. **Bootstrapping approaches**

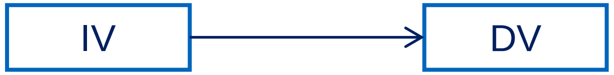
PROCESS (in SPSS)

R code

Baron und Kenny (1986), Preacher & Hayes (2004)

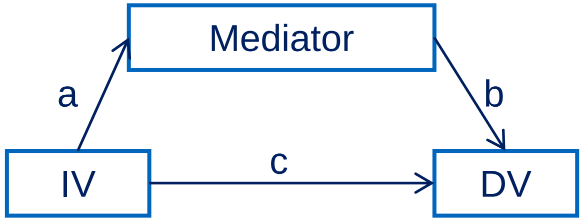
Which types of research models are there?

Simple model



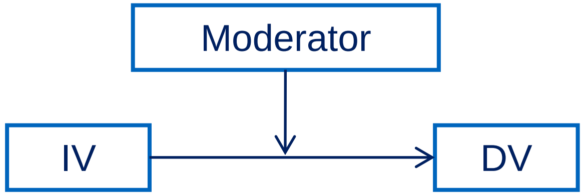
IV influences DV

Mediator



Are there third variables through which the IV influences the DV?

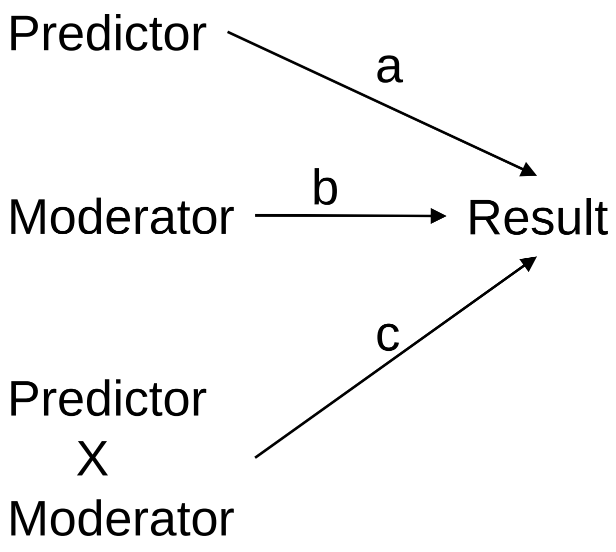
Moderator



Are there third variables that qualify the relationship between IV and DV?

Moderator Modell

- The predictor variable (a)
- The moderator variable (b)
- The interaction effect (c)
- **The moderation hypothesis is supported when the interaction (c) becomes significant**
- The **moderator should be uncorrelated with the UV and AV** to allow for interpretation
- **Moderators and UVs are on the same level** (unlike mediators who switch between UV and AV, depending on their perspective)
- **Moderation means that the relationship between two variables changes as a function of the moderator variable**; when there are unexpectedly weak or inconsistent relationships between variables, it is common to look for moderators
- For strong relationships between UV and AV, one is looking for mediators



Baron/Kenny (1986)

There are different ways to calculate moderations

1a. Interaction terms (multiplication) with mean-centered variables
(if scale of IV and DV is the same)

OR

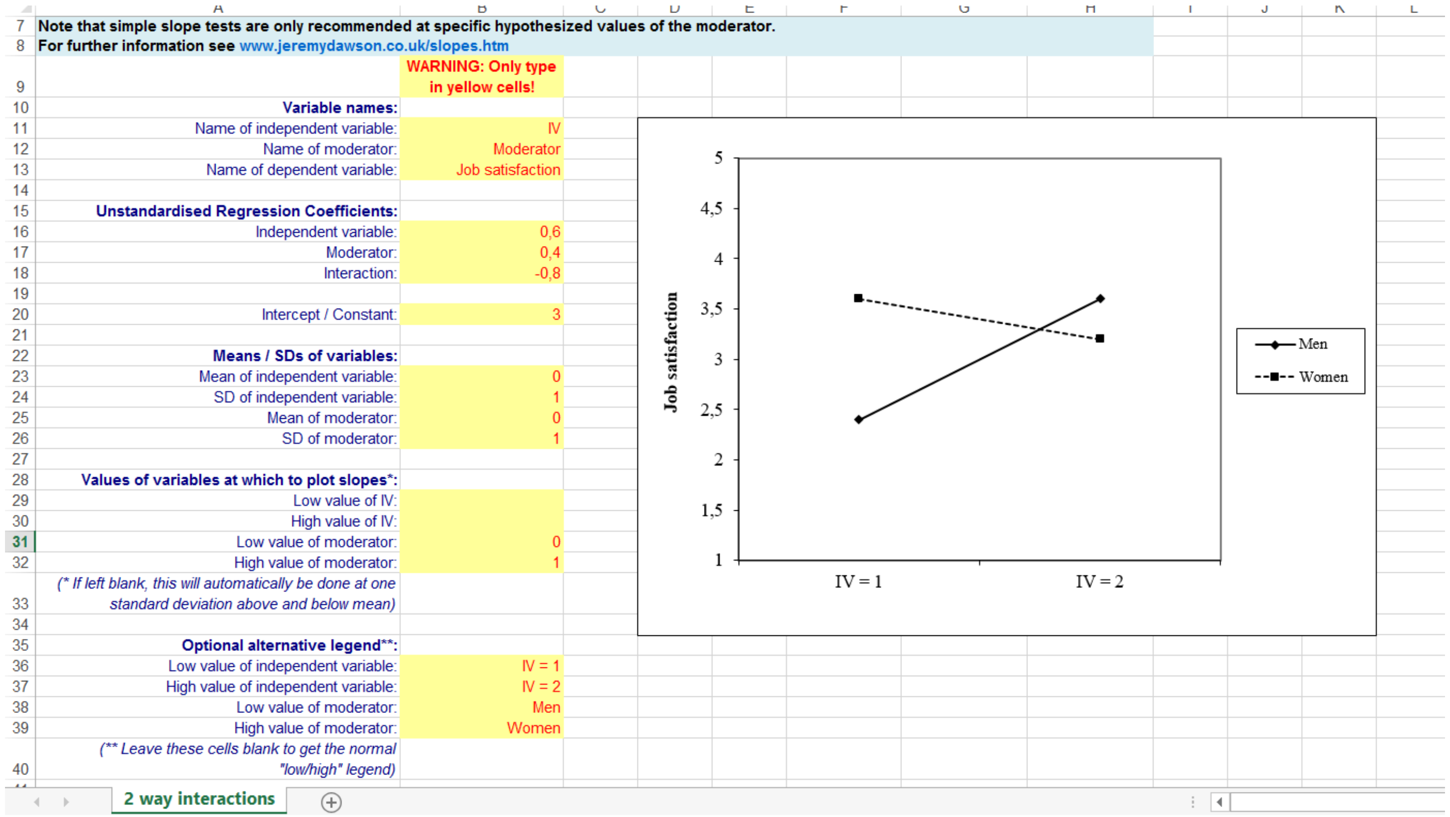
1b. Interaction terms (multiplication) with standardized (z-transformed) variables
(if scale of IV and DV is different)

2. Regression analysis or PROCESS (e.g. model 1)

All lower-order effects **always** have to be included in the regression analysis

Plotting significant interactions

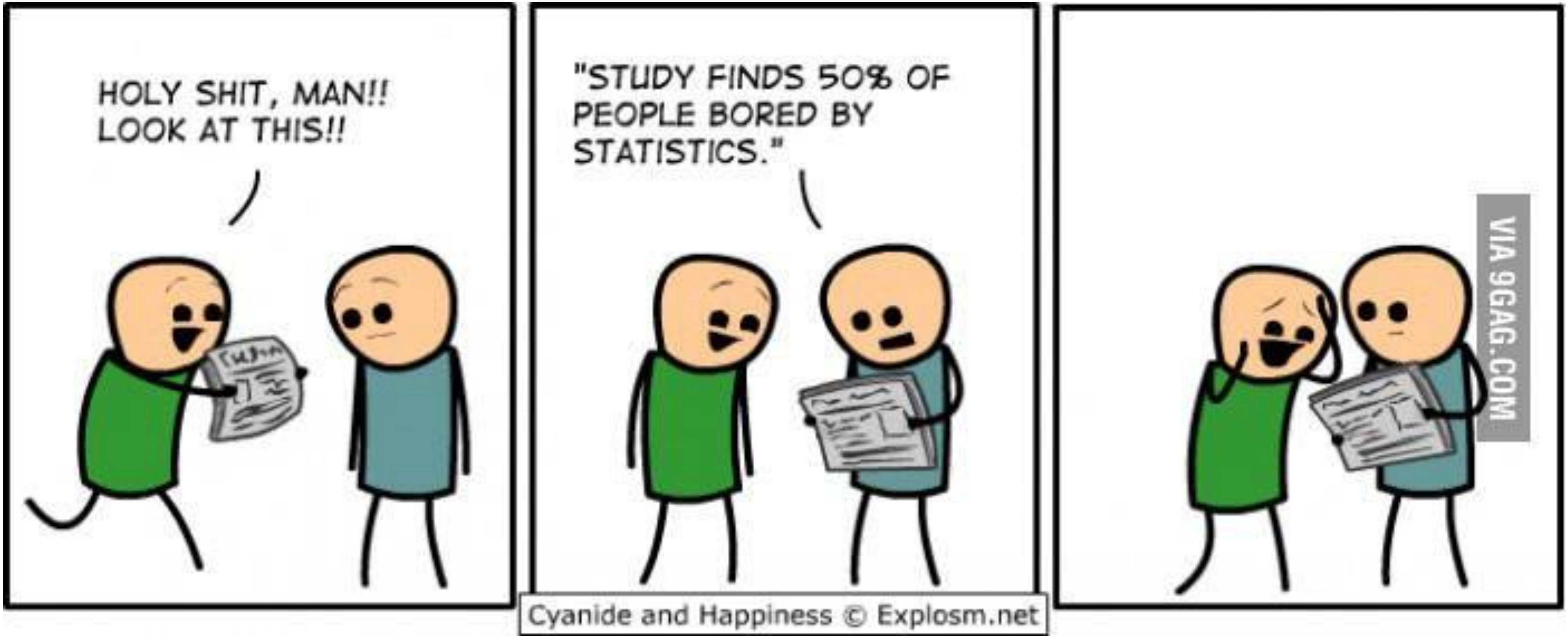
- Analyze the relationship between predictor and criterion for low (-1 SD) and high (+1 SD) values of the moderator (Aiken & West, 1991)
- Use the Excel tools provided by Jeremy Dawson on http://www.jeremydawson.com/slopes.htm
- Calculate simple slopes tests



	A	B
7	**Note that simple slope tests are only recommended at specific hypothesized values of the moderator.**	
8	**For further information see www.jeremydawson.co.uk/slopes.htm**	
9		**WARNING: Only type in yellow cells!**
10	**Variable names:**	
11	Name of independent variable:	IV
12	Name of moderator:	Moderator
13	Name of dependent variable:	Job satisfaction
14		
15	**Unstandardised Regression Coefficients:**	
16	Independent variable:	0,6
17	Moderator:	0,4
18	Interaction:	-0,8
19		
20	Intercept / Constant:	3
21		
22	**Means / SDs of variables:**	
23	Mean of independent variable:	0
24	SD of independent variable:	1
25	Mean of moderator:	0
26	SD of moderator:	1
27		
28	**Values of variables at which to plot slopes*:**	
29	Low value of IV:	
30	High value of IV:	
31	Low value of moderator:	0
32	High value of moderator:	1
33	*(* If left blank, this will automatically be done at one standard deviation above and below mean)*	
34		
35	**Optional alternative legend**:**	
36	Low value of independent variable:	IV = 1
37	High value of independent variable:	IV = 2
38	Low value of moderator:	Men
39	High value of moderator:	Women
40	*(** Leave these cells blank to get the normal "low/high" legend)*	

Dawson, J. F. (2014). Moderation in management research: What, why, when, and how. *Journal of Business and Psychology, 29*, 1–19.

Thank you!


HOLY SHIT, MAN!! LOOK AT THIS!!
"STUDY FINDS 50% OF PEOPLE BORED BY STATISTICS."
VIA 9GAG.COM
Cyanide and Happiness © Explosm.net

Snack Break

Source of image: https://kaboompics.com/download/d7fb73b7d9ea97ddf97046b0279af9e7/medium