



Intelligente Datenanalysen

Data Science

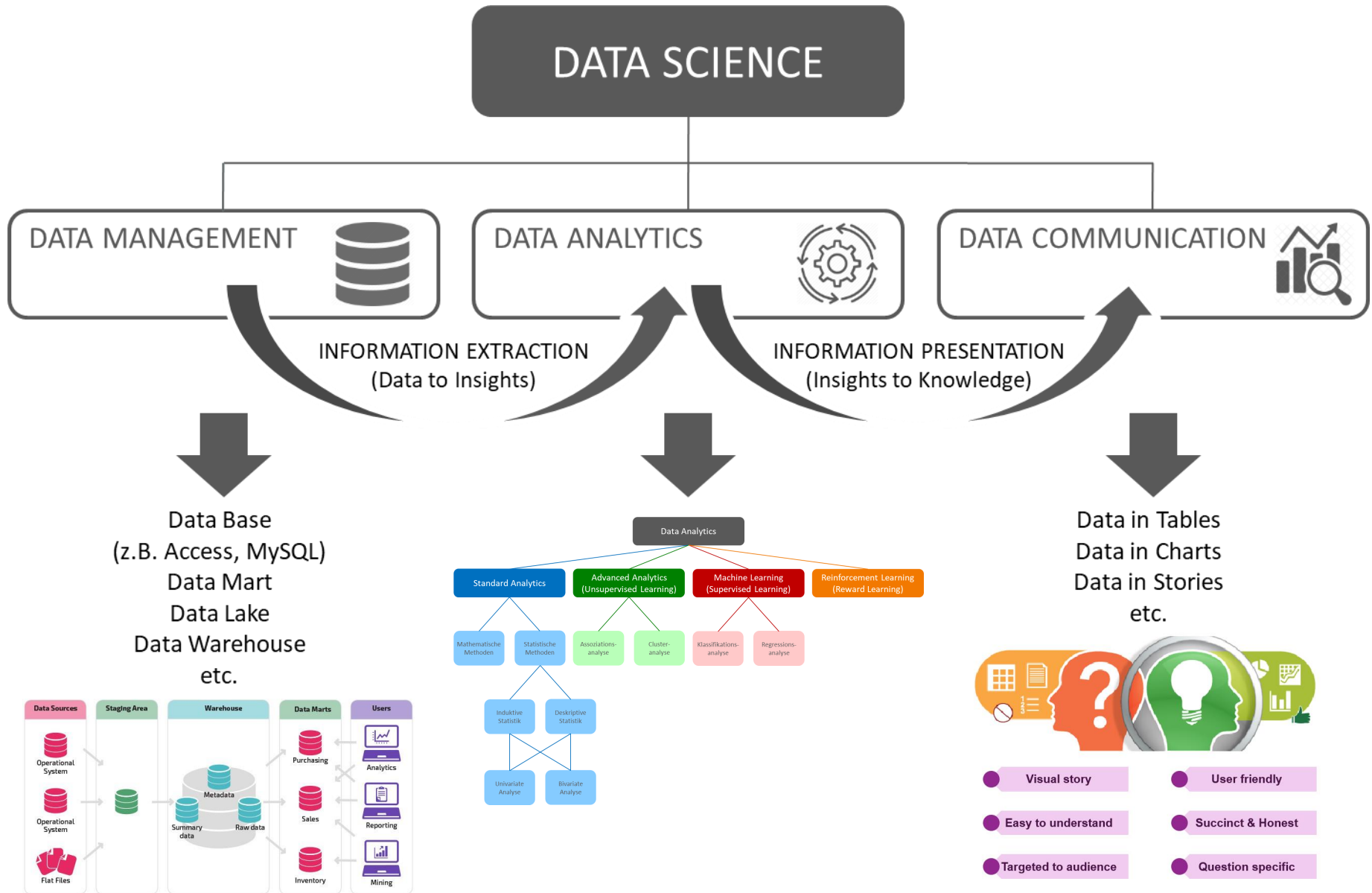
Überzeugende Datenvisualisierungen



Modul ST-01

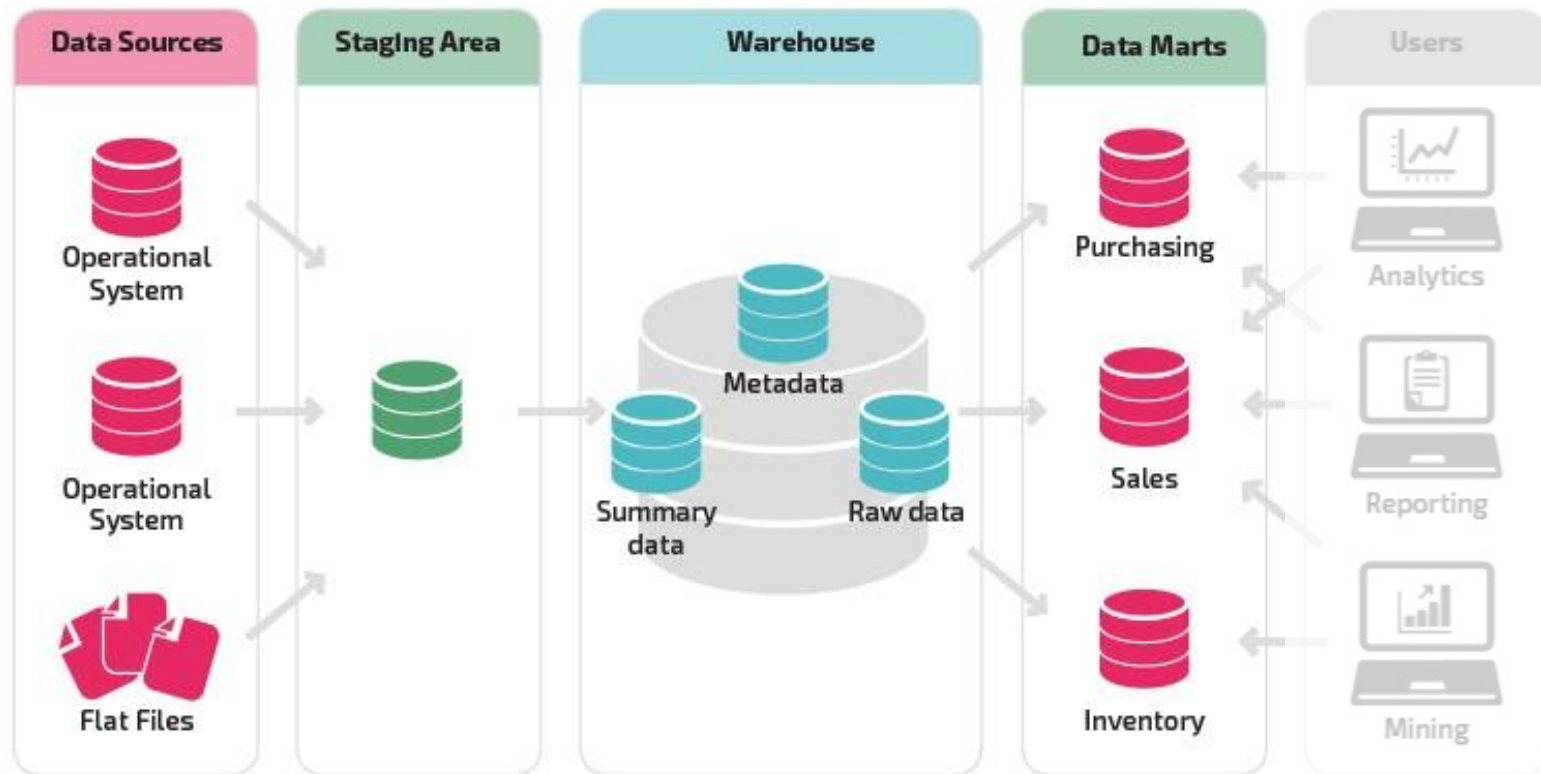
Praktische Statistik für Data Science

Was ist Data Science?

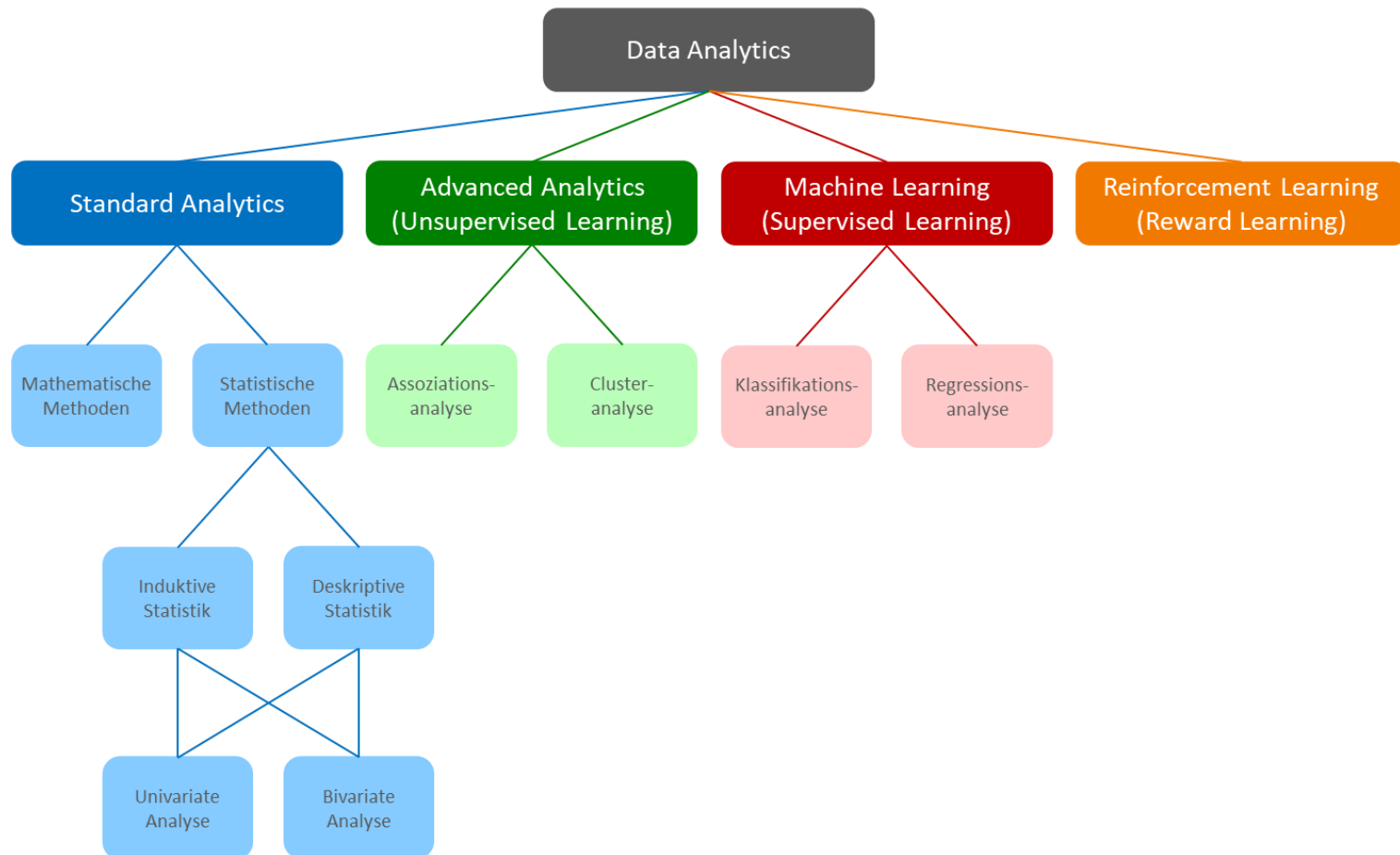


Was ist Data Science?

DATA MANAGEMENT



DATA ANALYTICS



DATA COMMUNICATION



Visual story

User friendly

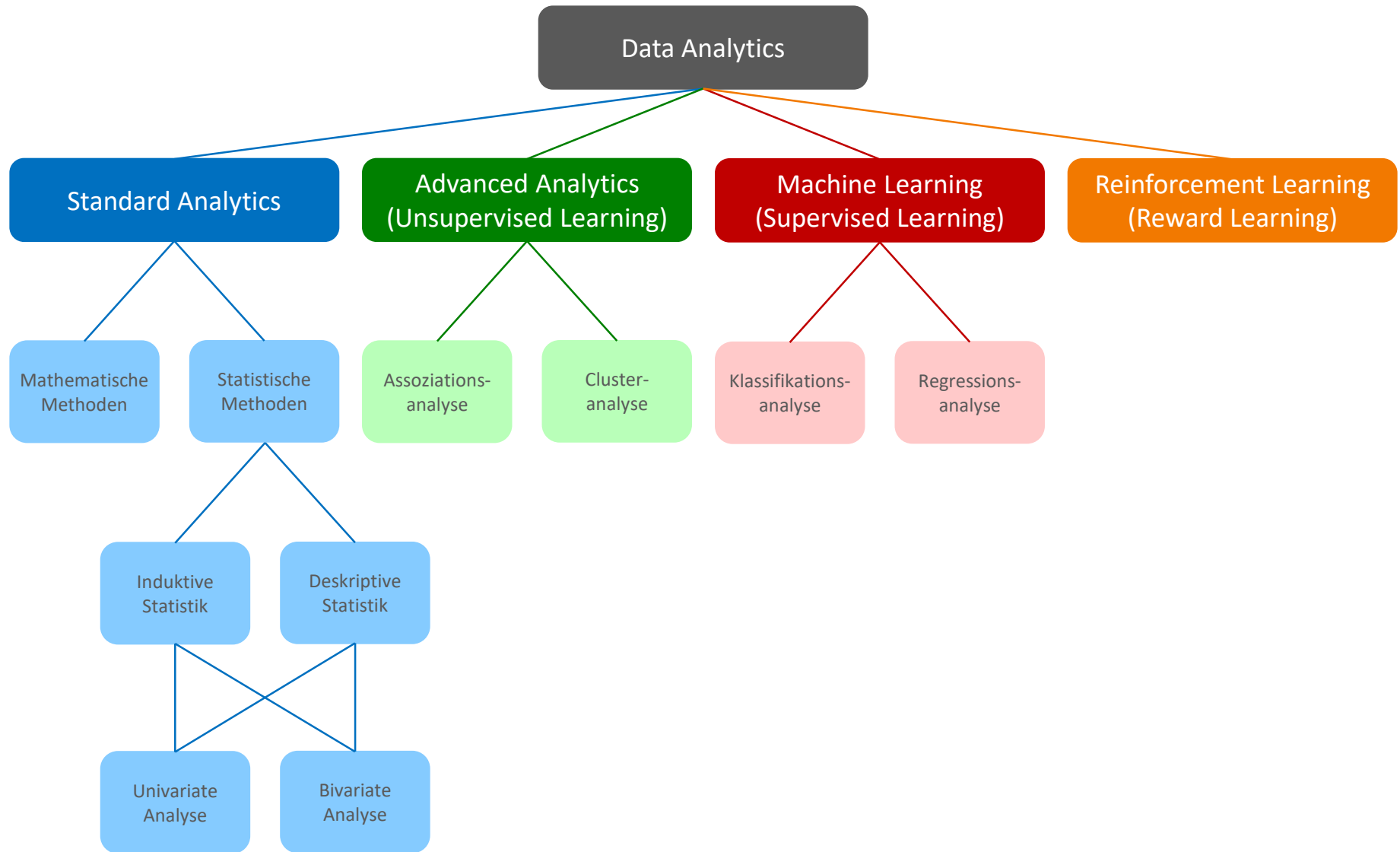
Easy to understand

Succinct & Honest

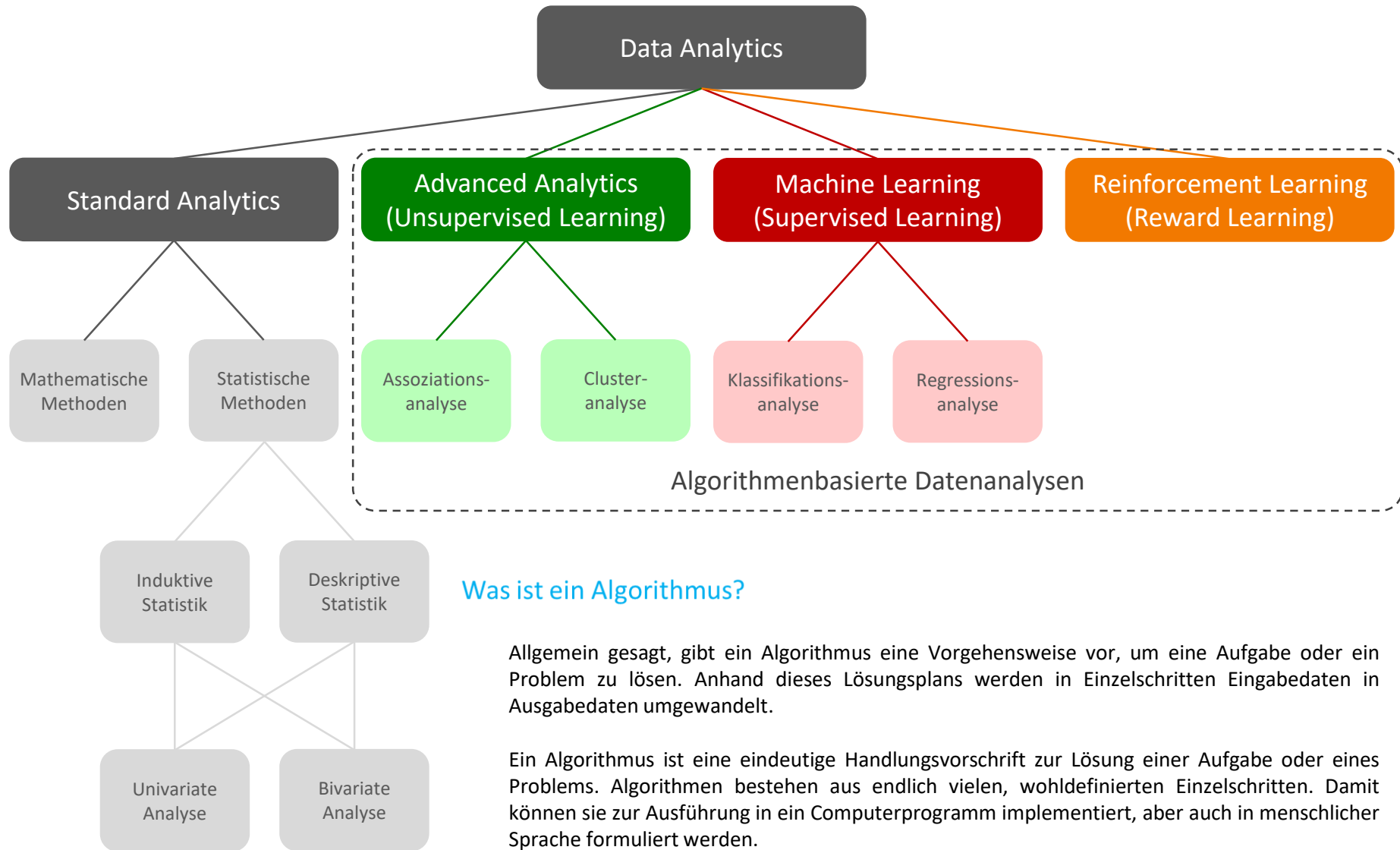
Targeted to audience

Question specific

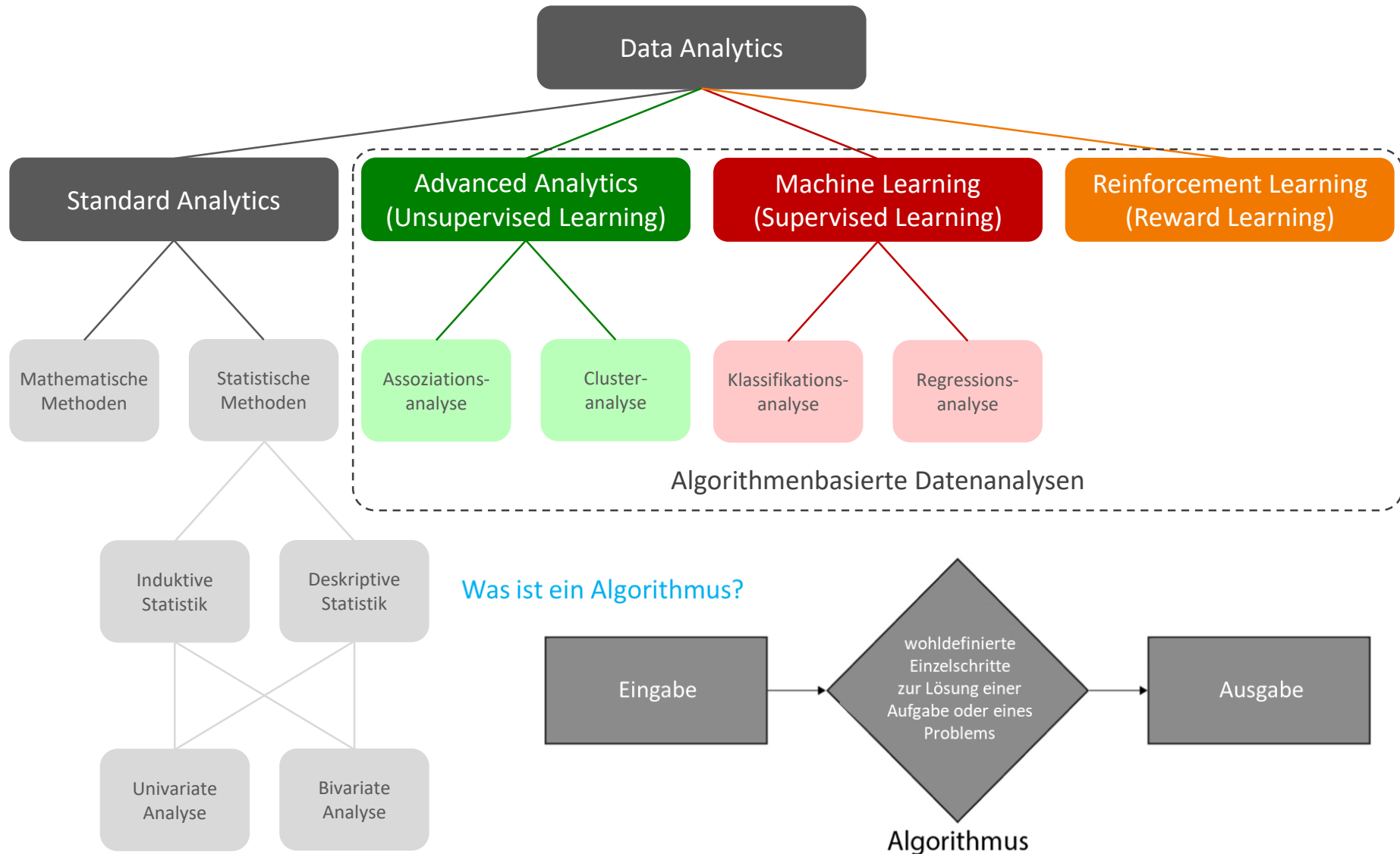
Was ist Data Analytics und welche Grenzen hat Data Analytics?



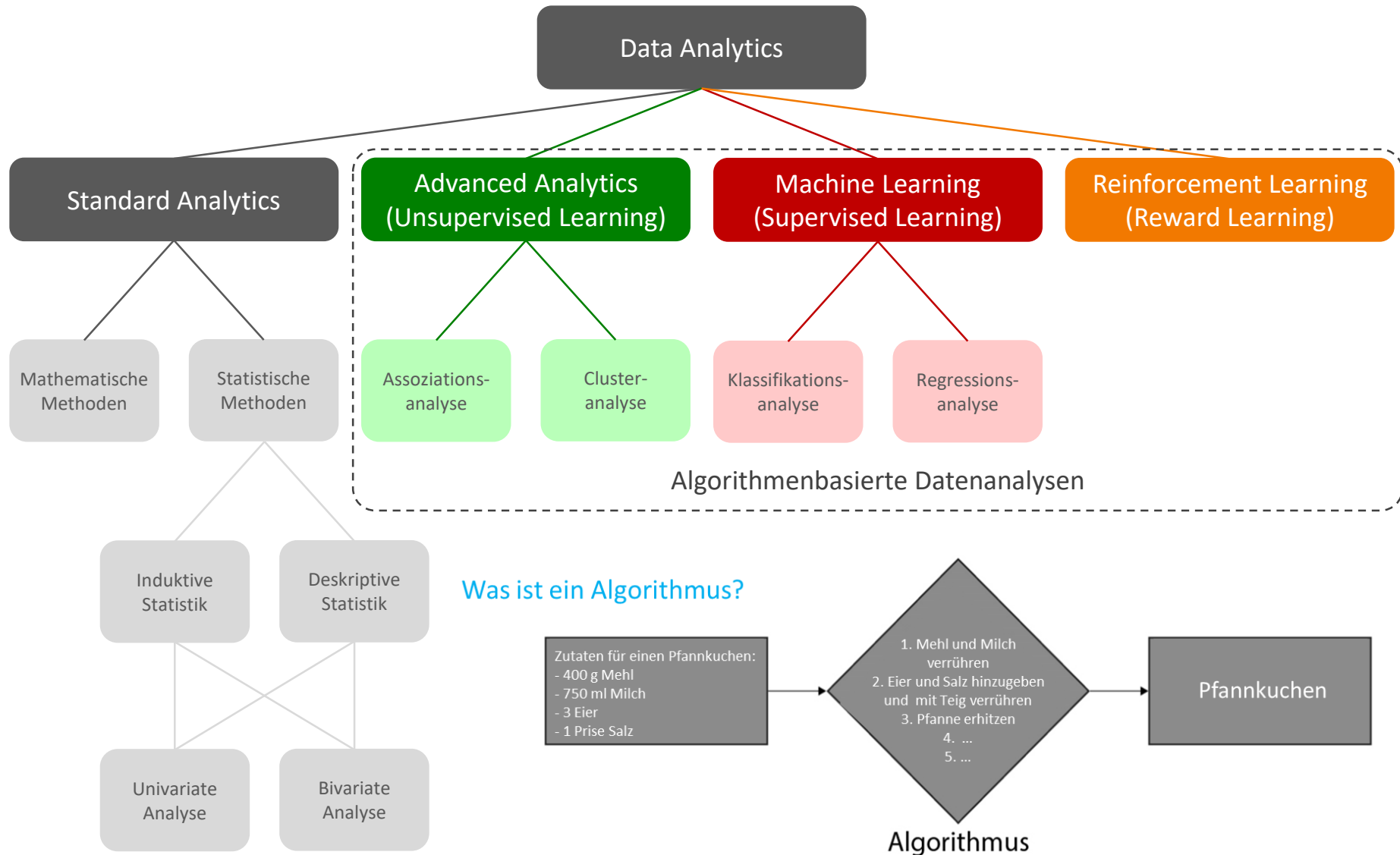
Was ist Data Analytics und welche Grenzen hat Data Analytics?



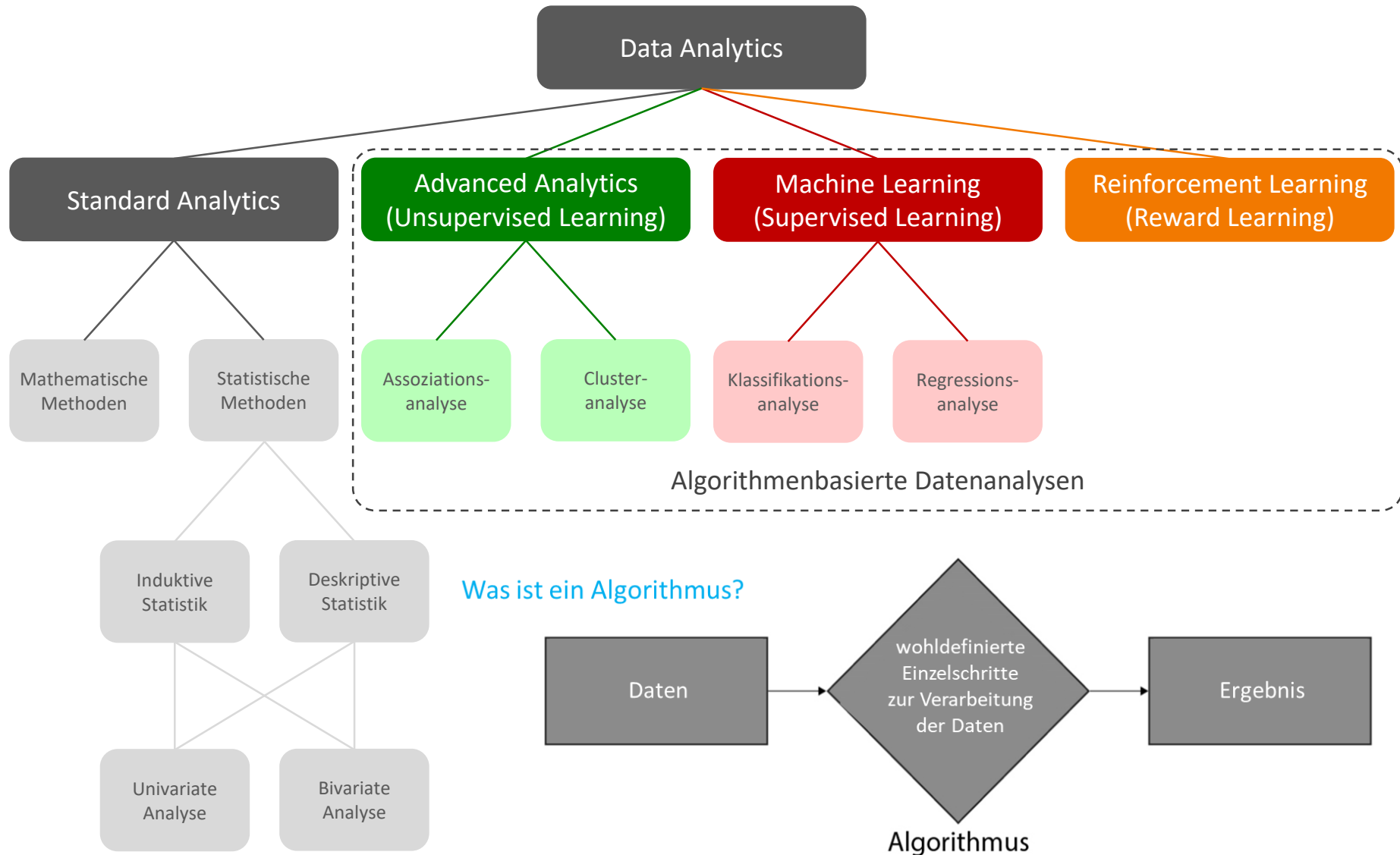
Was ist Data Analytics und welche Grenzen hat Data Analytics?



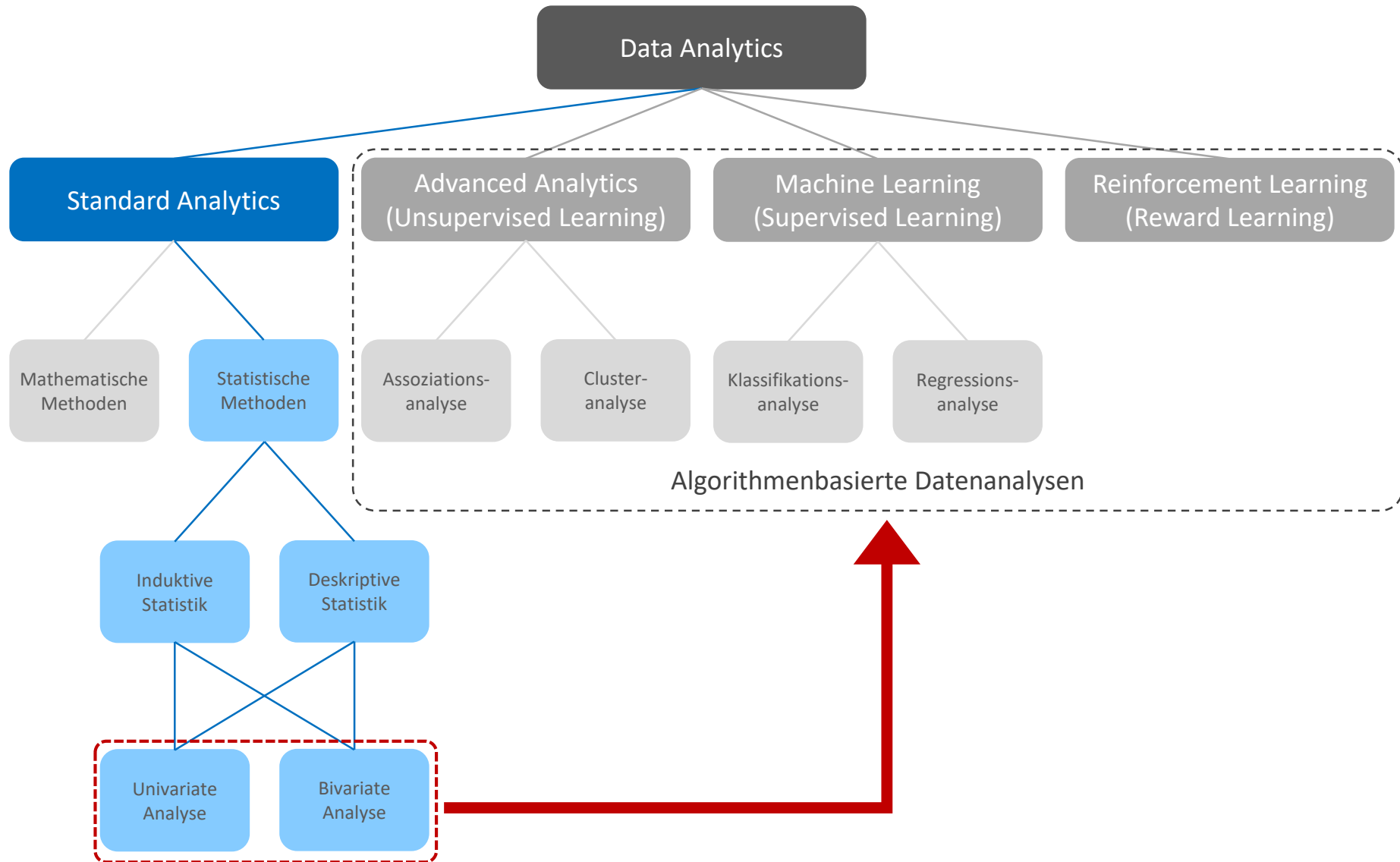
Was ist Data Analytics und welche Grenzen hat Data Analytics?



Was ist Data Analytics und welche Grenzen hat Data Analytics?



Was ist Data Analytics und welche Grenzen hat Data Analytics?



- Univariate Analyse
 - Skalenniveaus von Daten und statistische Lagemaße
 - Statistische Lagemaße
 - Modus, Median, Quartil, Interquartilsabstand, Spannweite und Mittelwert
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"
 - Skalenniveaus von Daten und statistische Streuungsmaße
 - Statistische Streuungsmaße
 - Varianz und Standardabweichung
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"

- Bivariate Analyse
 - Skalenniveaus von Daten und statistische Zusammenhangsmaße
 - Statistische Zusammenhangsmaße
 - χ^2 -Koeffizient und Cramer's V
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"
 - Korrelationskoeffizient (Pearson)
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"
 - t-Wert und Cohen's D
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"

- **Univariate Analyse**
 - Skalenniveaus von Daten und statistische Lagemaße
 - Statistische Lagemaße
 - Modus, Median, Quartil, Interquartilsabstand, Spannweite und Mittelwert
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"
 - Skalenniveaus von Daten und statistische Streuungsmaße
 - Statistische Streuungsmaße
 - Varianz und Standardabweichung
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"

Klassifikation von Daten

Nominalskala

$= / \neq$

Haarfarbe

Herkunft

- männlich / weiblich
- Berlin / Hamburg / Köln
- Lager 1 / Lager 2 / Lager 3

keine Beziehung

Ordinalskala

$> / <$

Tabellen-
-platz

Noten

- unfreundlich < freundlich
- AAA > AA+ > AA > A+ > A
- Schulnoten: 1 > 3 > 6

Ordnungsbeziehung

Kardinalskala

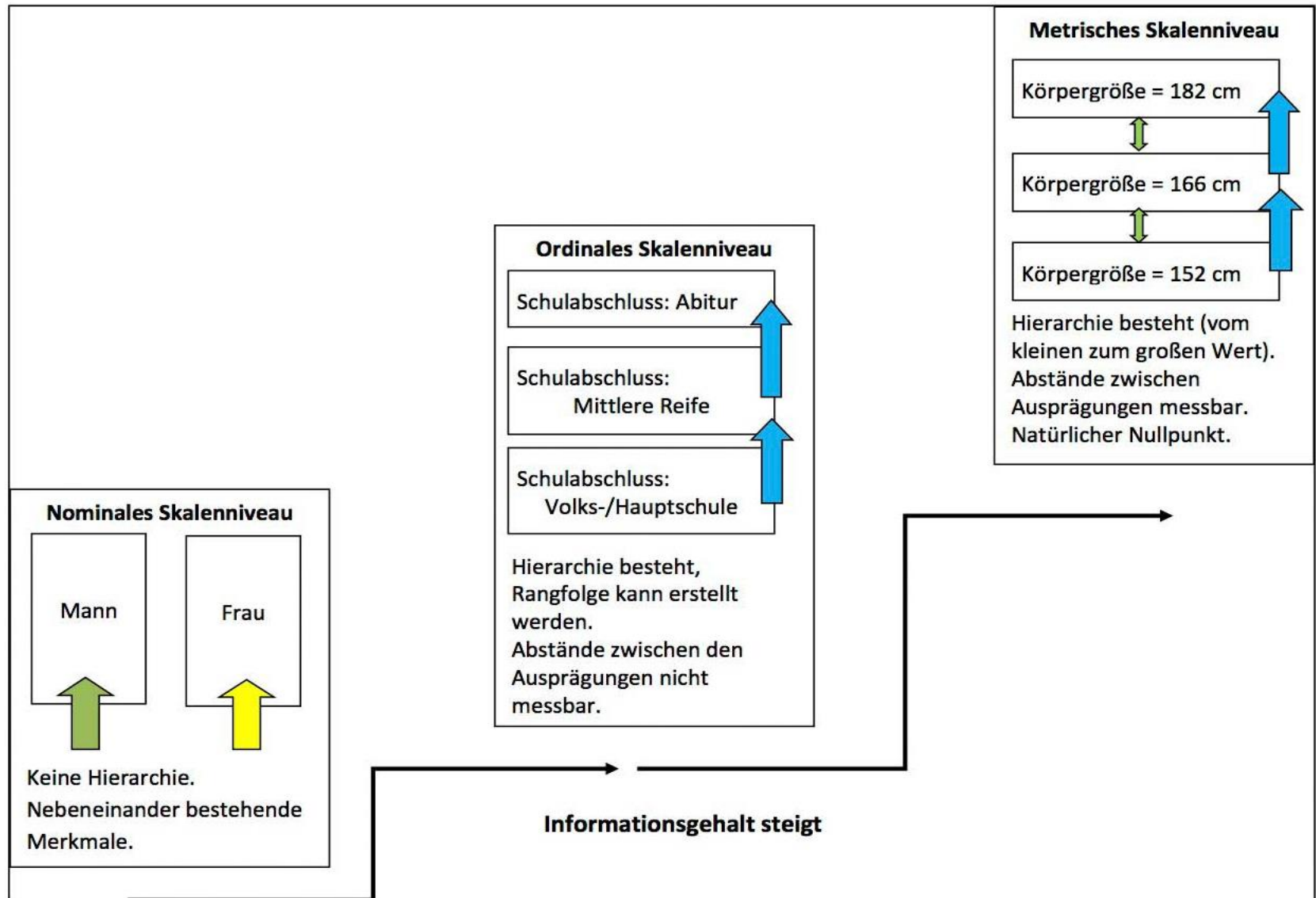
$+ / - (* / \div)$

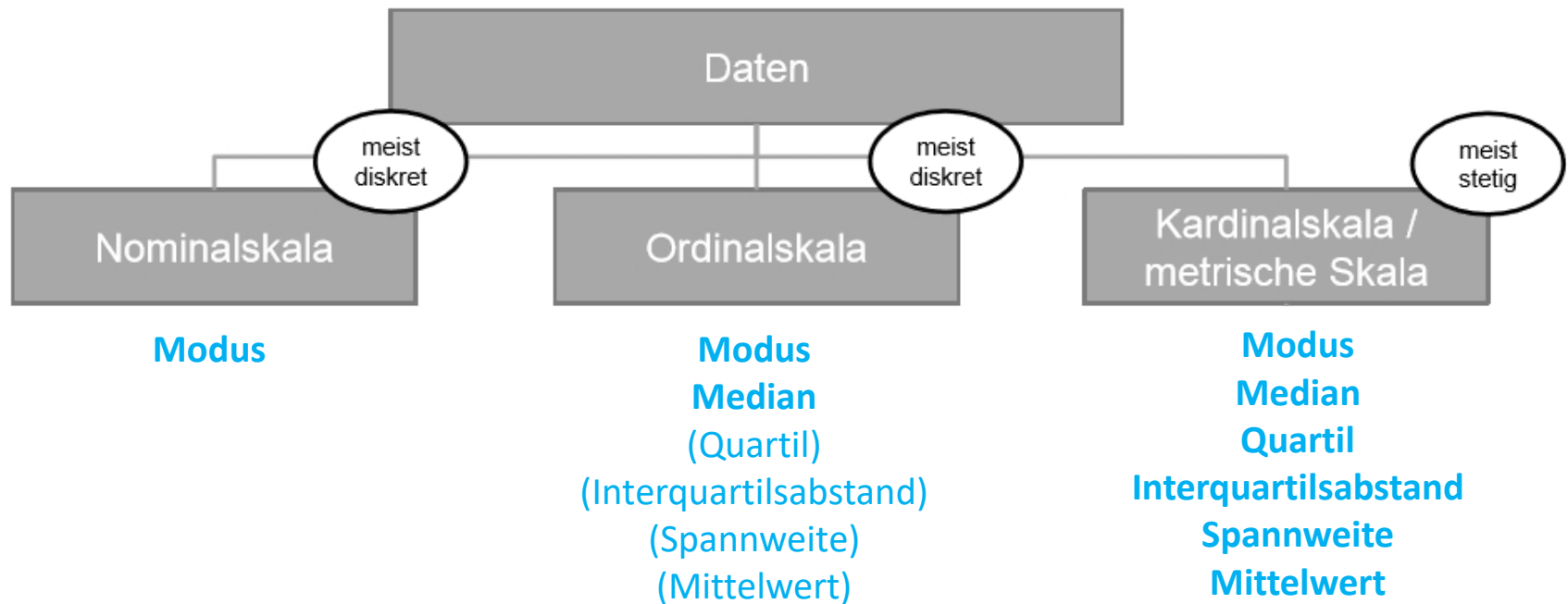
Zeit

Einkommen

- Einkommen
- Alter
- Umsatz

Arithmetische Beziehung

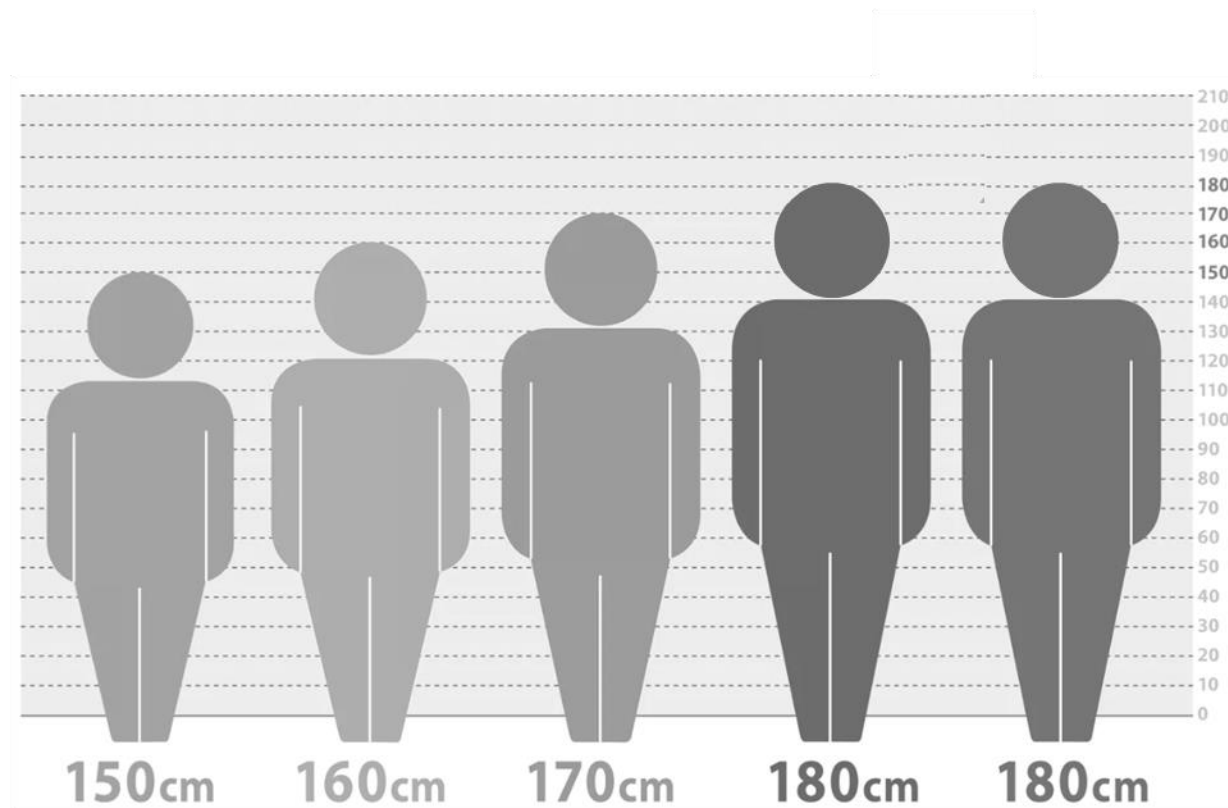




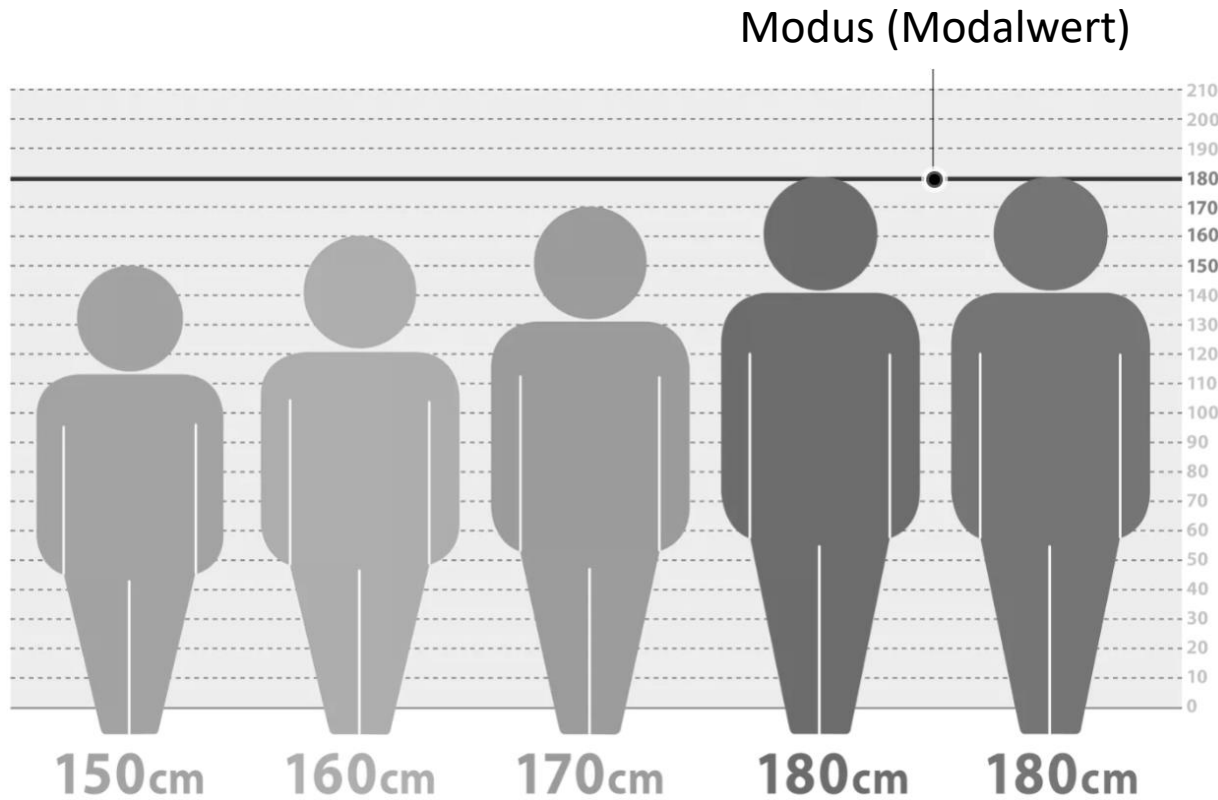
Lagemaße, die ein niedriges Skalenniveau voraussetzen, können problemlos für Datensätze mit einem höheren Skalenniveau berechnet werden

- **Univariate Analyse**
 - Skalenniveaus von Daten und statistische Lagemaße
 - **Statistische Lagemaße**
 - **Modus, Median, Quartil, Interquartilsabstand, Spannweite und Mittelwert**
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"
 - Skalenniveaus von Daten und statistische Streuungsmaße
 - **Statistische Streuungsmaße**
 - **Varianz und Standardabweichung**
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"

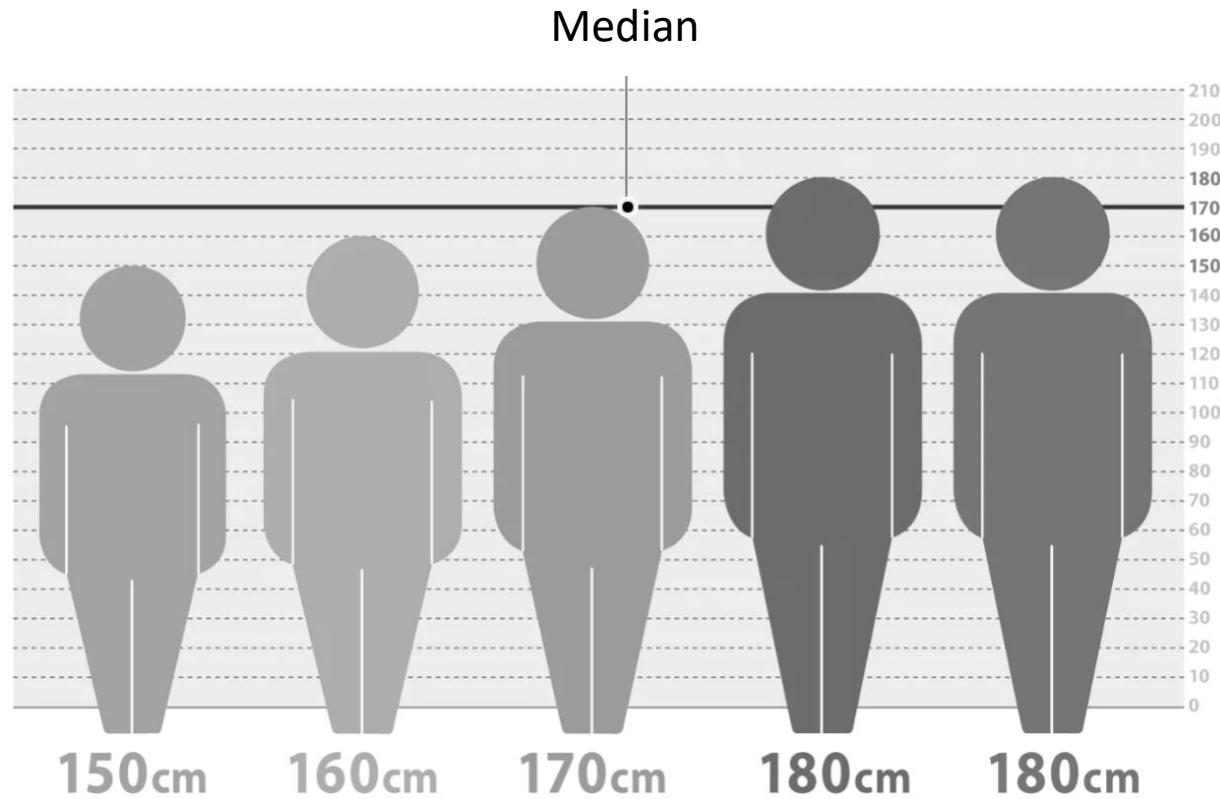
Modus, Median, Quartil, Interquartilsabstand, Spannweite und Mittelwert



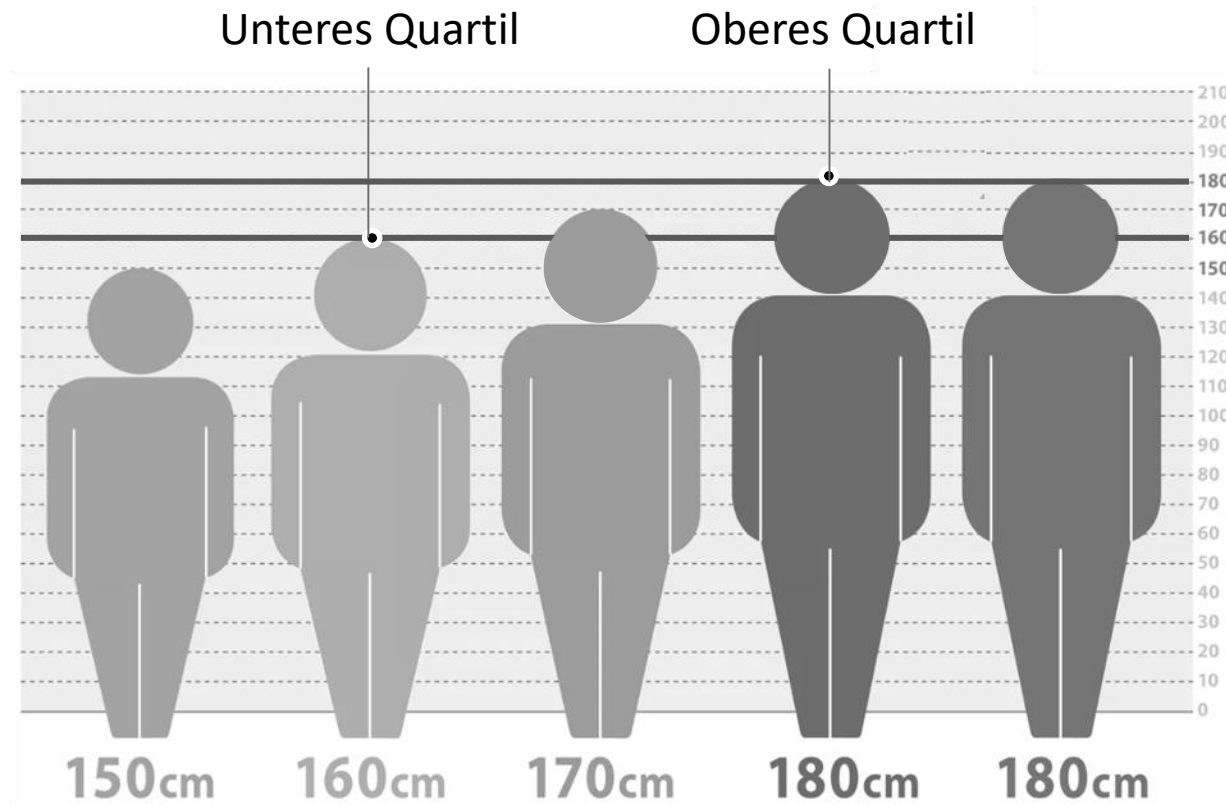
Wert, der am häufigsten vorkommt



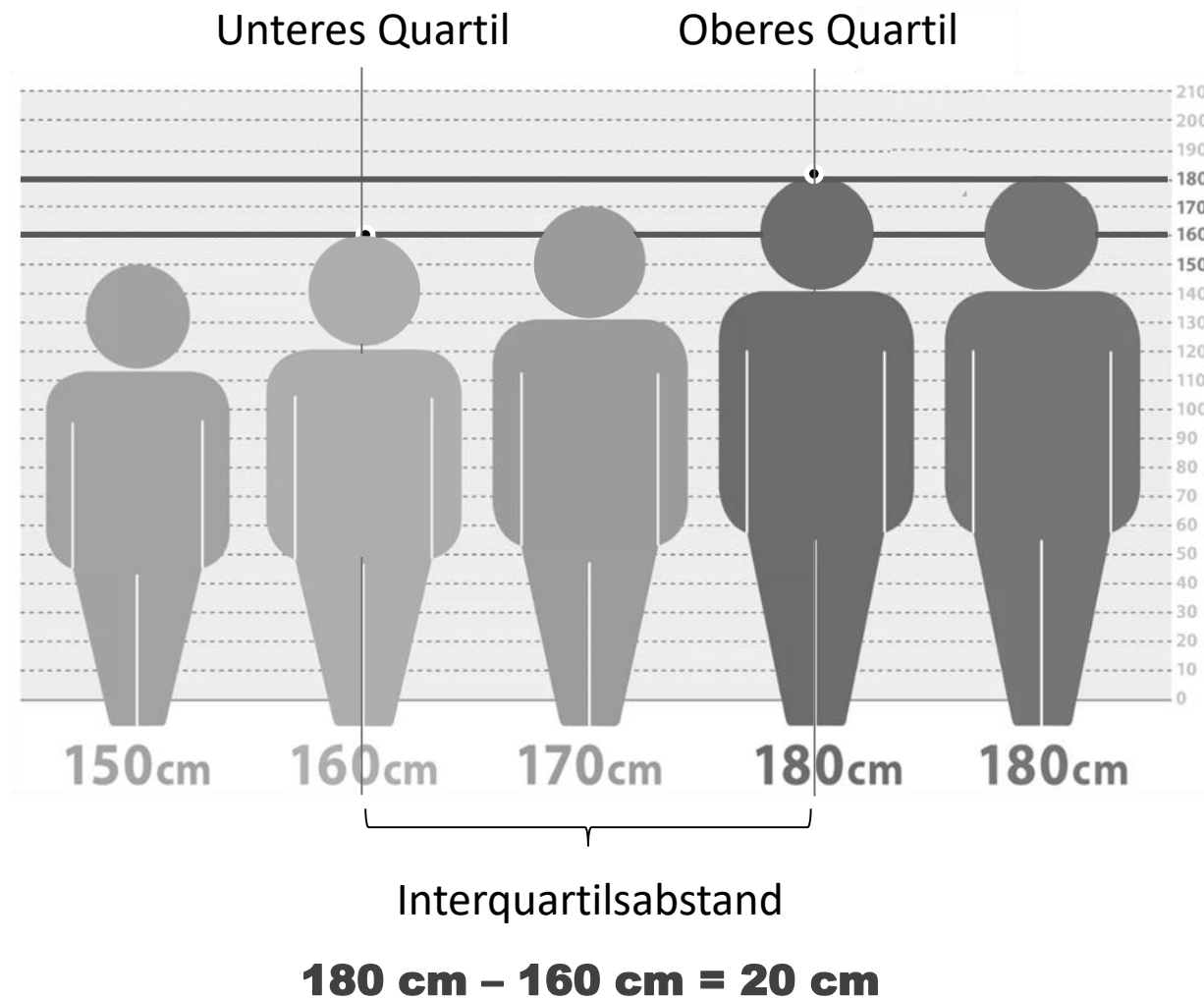
Wert, der in der Mitte liegt



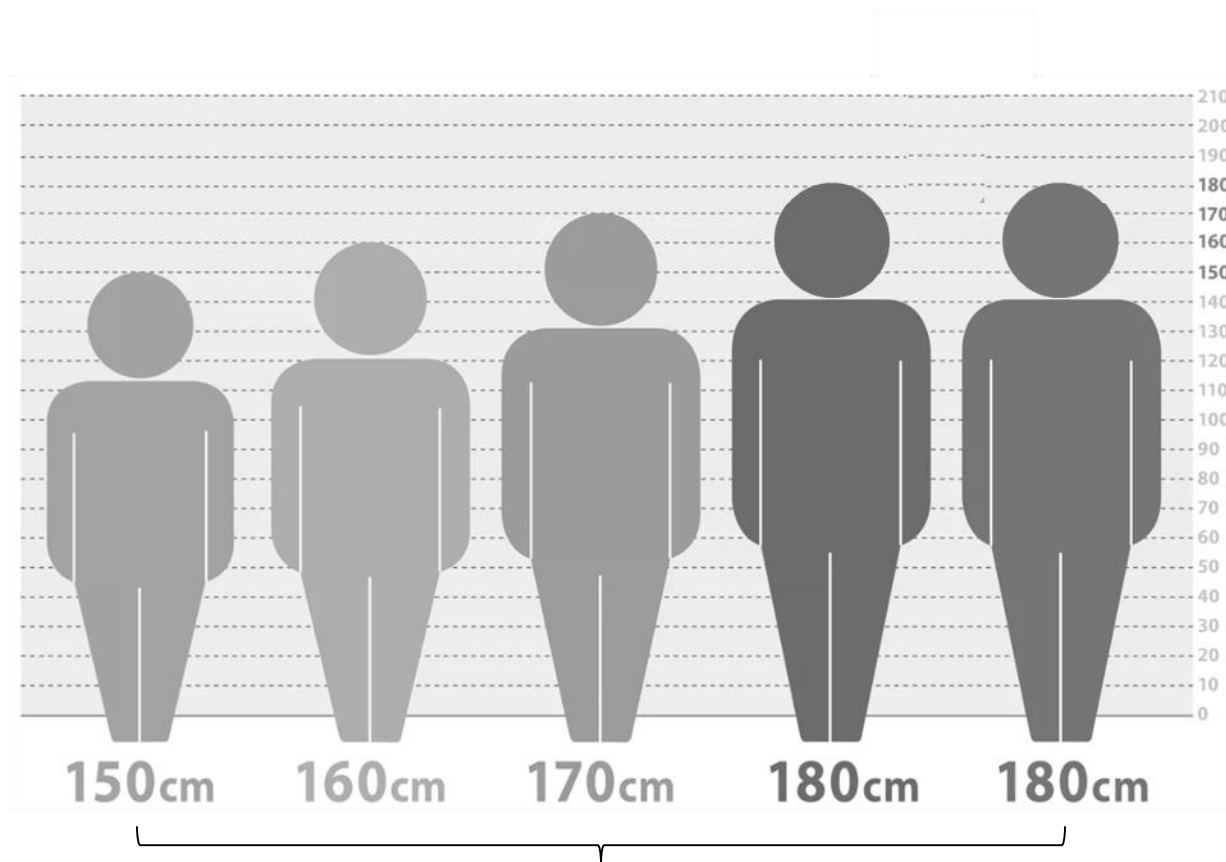
Werte, die in der Mitte der beiden Hälften liegen



Abstand (Differenz) zwischen oberem und unterem Quartil



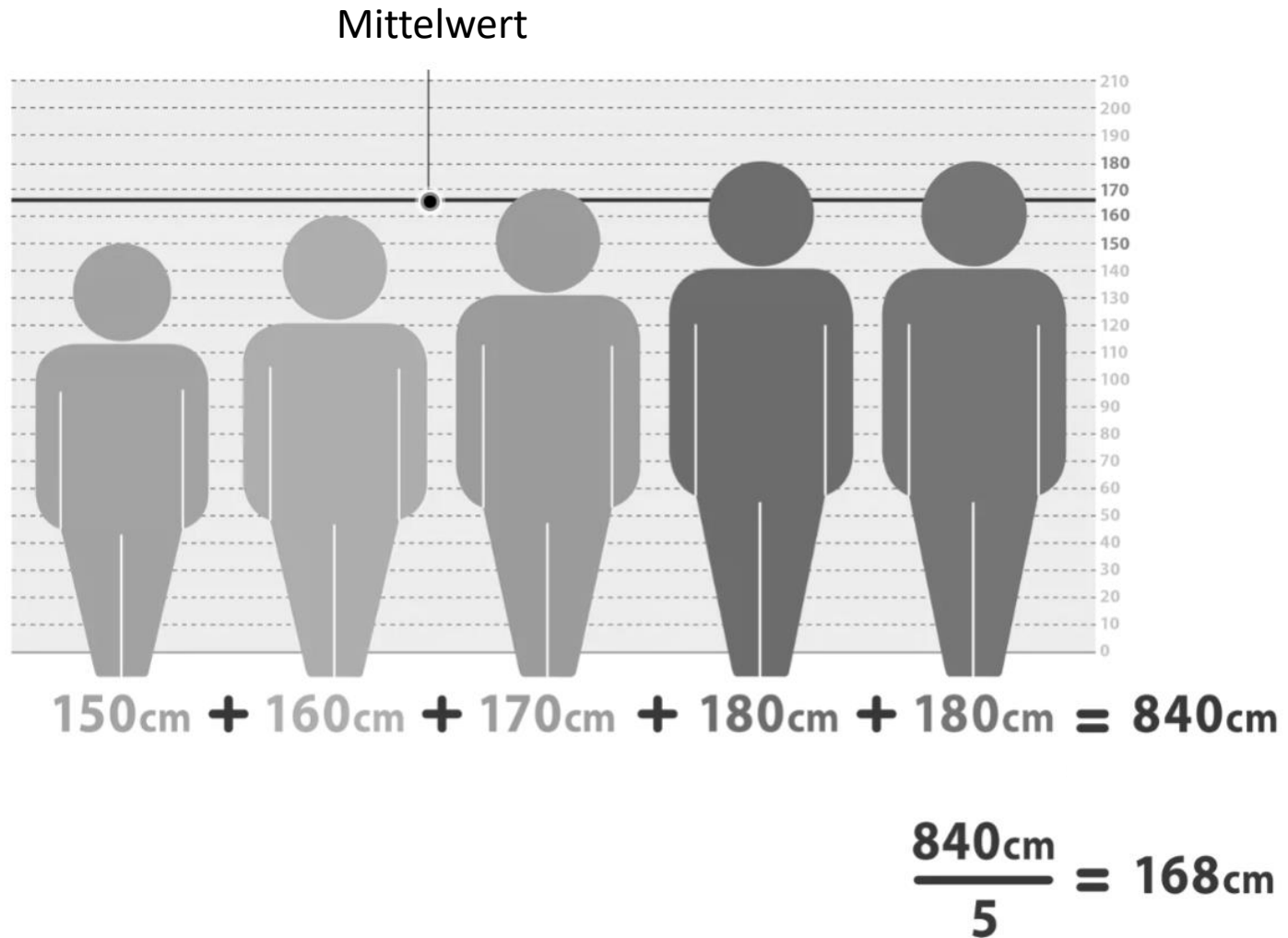
Abstand (Differenz) zwischen Maximal- und Minimalwert



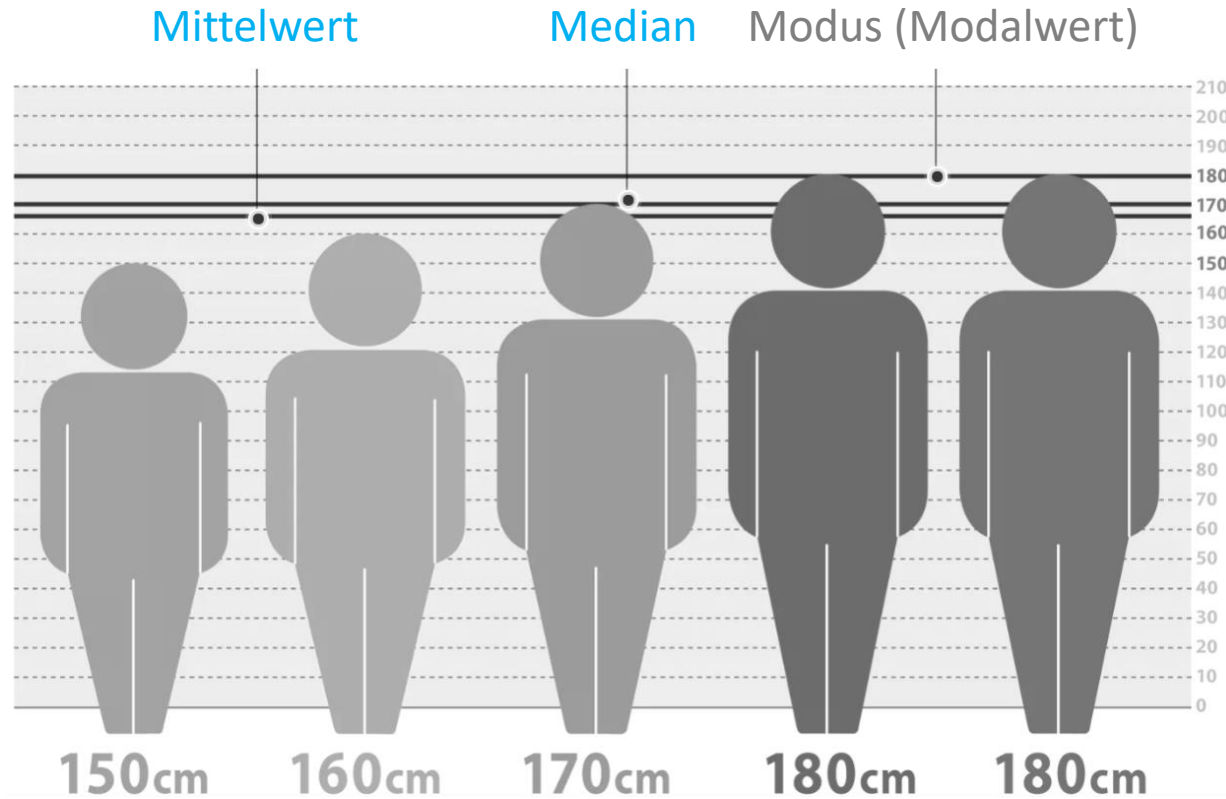
Spannweite

$$180 \text{ cm} - 150 \text{ cm} = 30 \text{ cm}$$

(ungewichteter) Durchschnittswert



Median vs. Mittelwert



Median vs. Mittelwert

Vorteil des Medians

- Unempfindlich gegenüber Extremwerten und **Ausreißern**

Nachteil des Medians

- Man kennt zwar die Mitte der Verteilung, aber nicht die Schwerpunkte der Daten

Vorteil des Mittelwerts

- Ein Vorteil des Mittelwerts ist, dass er sensitiv hinsichtlich jeden einzelnen Werts ist.
- Das bedeutet, dass wenn sich ein Wert ändert, sich auch der Mittelwert ändert.
- Aufgrund dieser Eigenschaft ist der Mittelwert ein guter Schätzer des Zentrums einer Verteilung.

Nachteil des Mittelwerts

- Dadurch, dass der Mittelwert sehr sensitiv hinsichtlich jeden Werts ist, kann er auch stark beeinflusst werden durch **Ausreißer.**

Median vs. Mittelwert

9 Personen haben das folgende Einkommen:

Person A = 10.000 €

Person B = 20.000 €

Person C = 20.000 €

Person D = 30.000 €

Person E = 30.000 €

Person F = 40.000 €

Person G = 40.000 €

Person H = 50.000 €

Person I = 660.000 €

Summe = 900.000 €

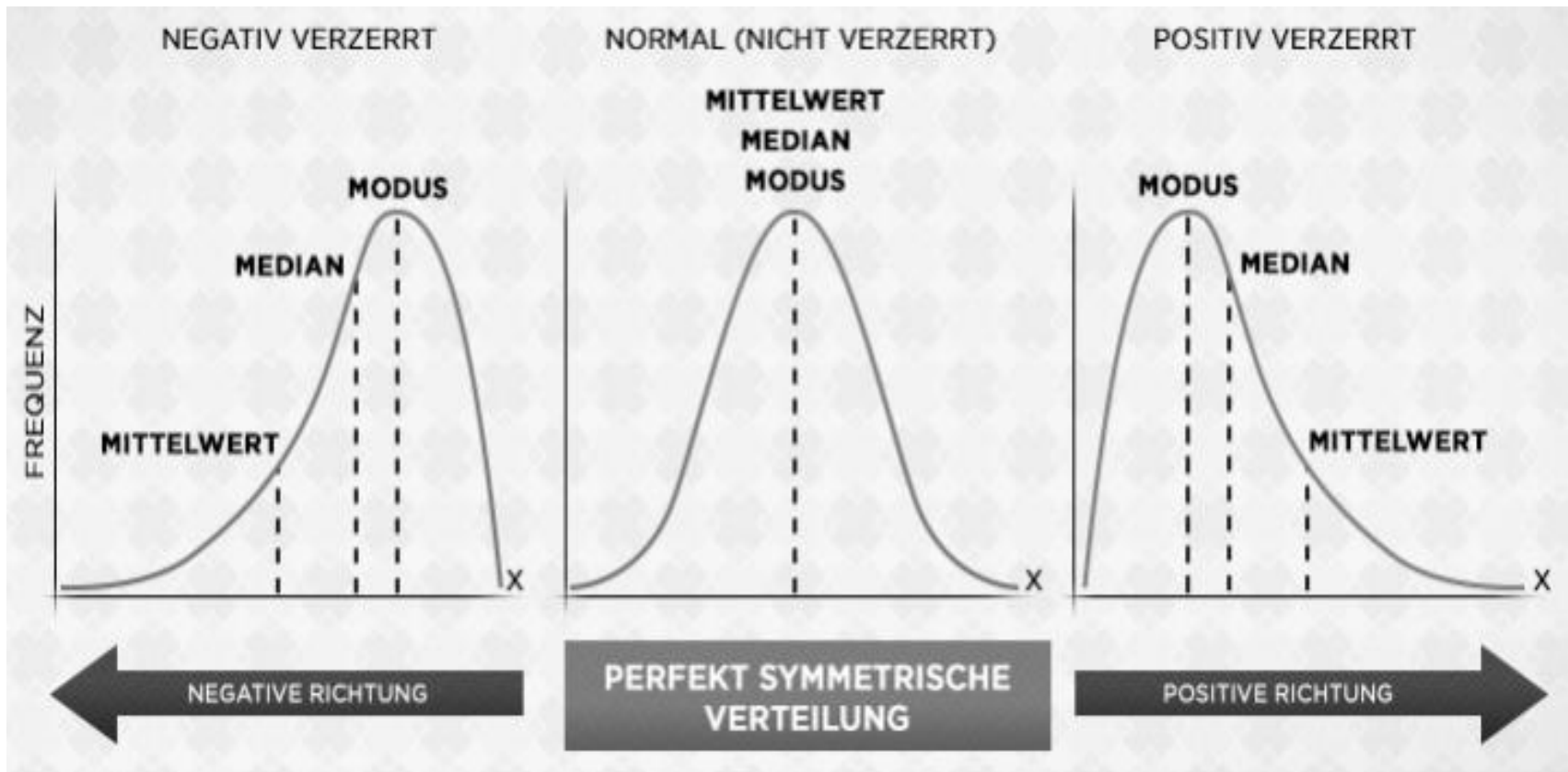
Wie hoch ist das „mittlere“ Einkommen?

Mittelwert = $900.000 / 9 = 100.000$ €

Median = 30.000 € → näher an der „Wahrheit“

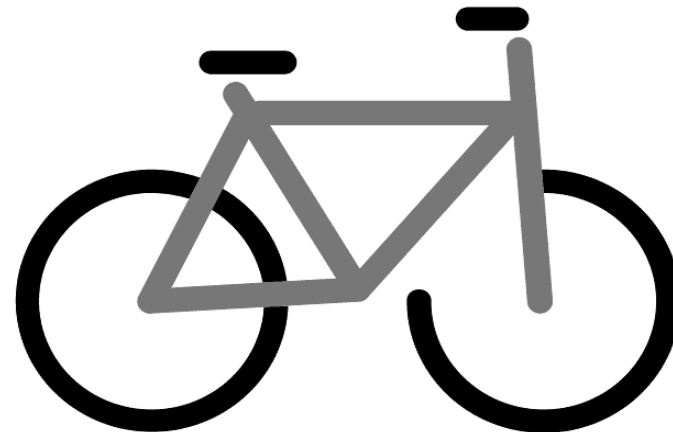
→ positiv verzerrte Verteilung

Median vs. Mittelwert




Fallstudie „Fahrradtour“

Tag	Kilometer	Höhenmeter
Tag 1	66	227
Tag 2	85	179
Tag 3	55	267
Tag 4	32	363
Tag 5	73	232
Tag 6	55	261
Tag 7	42	304
Tag 8	30	442
Tag 9	102	191
Tag 10	48	313
Tag 11	53	289
Tag 12	75	213
Tag 13	60	249
Tag 14	64	218



Fallstudie „Fahrradtour“



Tag	Kilometer	Höhenmeter
	30	179
	32	191
	42	213
	48	218
	53	227
	55	232
	55	249
	60	261
	64	267
	66	289
	73	304
	75	313
	85	363
	102	442

Modus (Modalwert)

Modus Kilometer = 55

Modus Höhenmeter = n. a. (nicht berechenbar)

Fallstudie „Fahrradtour“

Tag	Kilometer	Höhenmeter
	30	179
	32	191
	42	213
	48	218
	53	227
	55	232
	55	249
	60	261
	64	267
	66	289
	73	304
	75	313
	85	363
	102	442


Werte aufsteigend sortiert

Median

$$\text{Median}_{\text{Kilometer}} = \frac{55 + 60}{2} = 57,5$$

$$\text{Median}_{\text{Höhenmeter}} = \frac{249 + 261}{2} = 255,0$$

Fallstudie „Fahrradtour“



Tag	Kilometer	Höhenmeter
	30	179
	32	191
	42	213
	48	218
	53	227
	55	232
	55	249
	60	261
	64	267
	66	289
	73	304
	75	313
	85	363
	102	442

Unteres und oberes Quartil
(„ungewichtete“ Methode gem. „Orange“)


Unteres Quartil _{Kilometer} = 48

Unteres Quartil _{Höhenmeter} = 218

Oberes Quartil _{Kilometer} = 73

Oberes Quartil _{Höhenmeter} = 304

Fallstudie „Fahrradtour“



Tag	Kilometer	Höhenmeter
	30	179
	32	191
	42	213
	48	218
	53	227
	55	232
	55	249
	60	261
	64	267
	66	289
	73	304
	75	313
	85	363
	102	442

Interquartilsabstand

Interquartilsabstand _{Kilometer} =

Oberes Quartil _{Kilometer} – Unterres Quartil _{Kilometer} =

$$73 - 48 = 25$$

Interquartilsabstand _{Höhenmeter} =

Oberes Quartil _{Höhenmeter} – Unterres Quartil _{Höhenmeter} =

$$304 - 218 = 86$$

Fallstudie „Fahrradtour“

Tag	Kilometer	Höhenmeter
	30	179
	32	191
	42	213
	48	218
	53	227
	55	232
	55	249
	60	261
	64	267
	66	289
	73	304
	75	313
	85	363
	102	442

Werte aufsteigend sortiert

Spannweite

$$\text{Spannweite Kilometer} = 102 - 30 = 72$$

Maximalwert von Kilometer

Minimalwert von Kilometer

$$\text{Spannweite Höhenmeter} = 442 - 179 = 263$$

Maximalwert von Höhenmeter

Minimalwert von Höhenmeter

Fallstudie „Fahrradtour“

Tag	Kilometer	Höhenmeter
	30	179
	32	191
	42	213
	48	218
	53	227
	55	232
	55	249
	60	261
	64	267
	66	289
	73	304
	75	313
	85	363
	102	442

Werte aufsteigend sortiert

(arithmetischer) Mittelwert

Mittelwert Kilometer =

$$\frac{30 + 32 + 42 + 48 + 53 + 55 + 55 + 60 + 64 + 66 + 73 + 75 + 85 + 102}{14} = 60,0$$

Anzahl der Werte in der Tabelle (Urliste)

Mittelwert Höhenmeter =

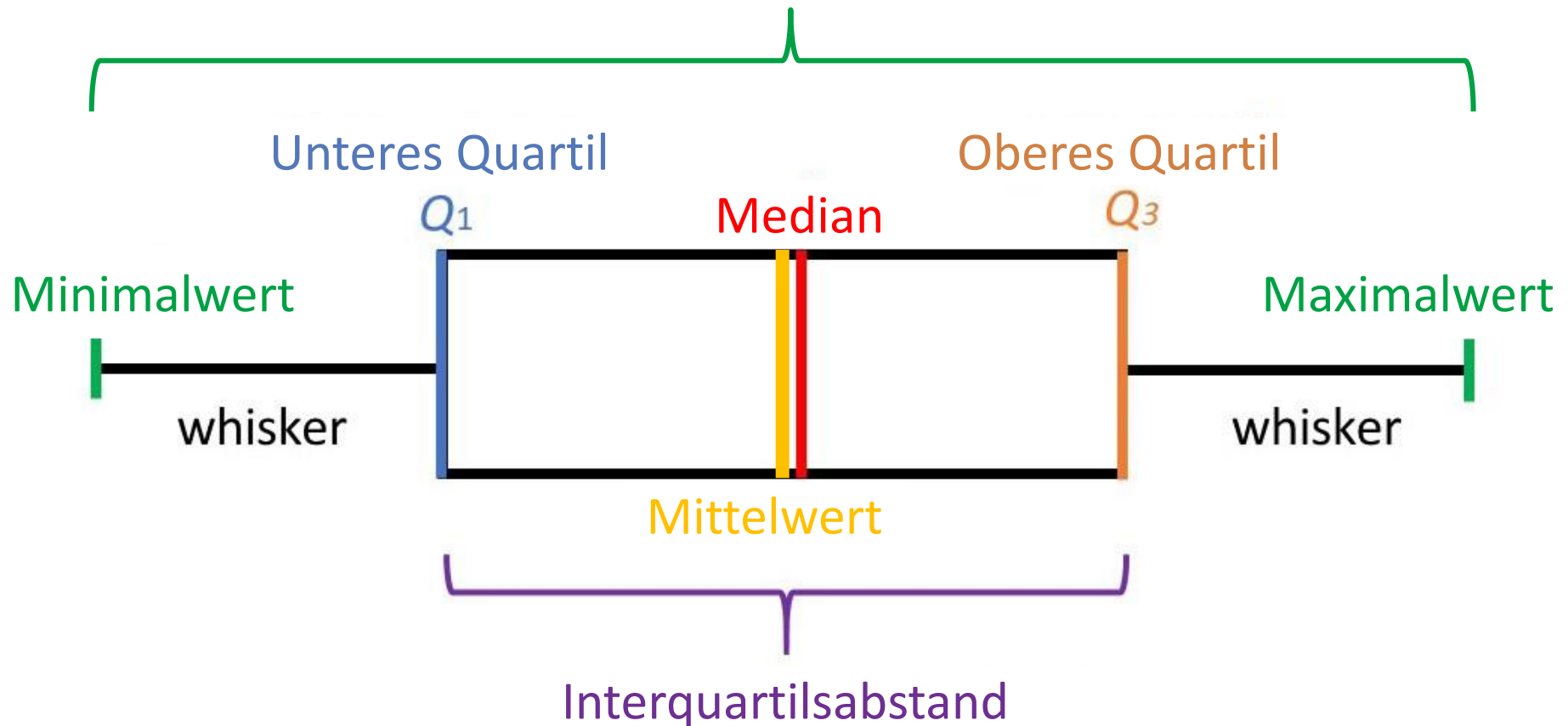
$$\frac{179 + 191 + 213 + 218 + 227 + 232 + 249 + 261 + 267 + 289 + 304 + 313 + 363 + 442}{14} = 267,7$$

Anzahl der Werte in der Tabelle (Urliste)

Box-Plot

(„5-Punkte-Zusammenfassung einer Verteilung“)

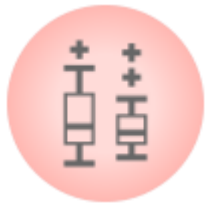
Spannweite



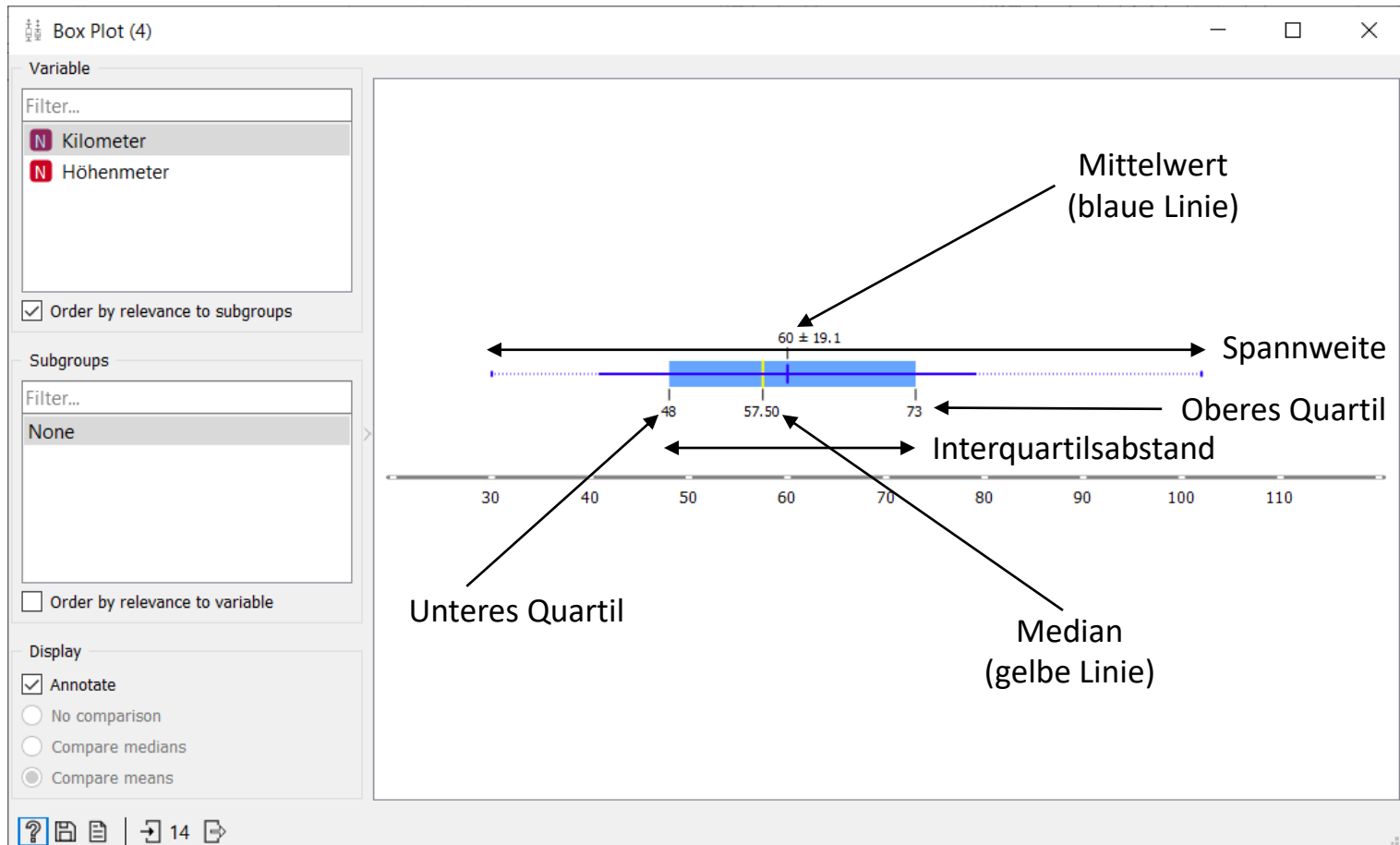
- **Univariate Analyse**
 - Skalenniveaus von Daten und statistische Lagemaße
 - **Statistische Lagemaße**
 - **Modus, Median, Quartil, Interquartilsabstand, Spannweite und Mittelwert**
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"
 - Skalenniveaus von Daten und statistische Streuungsmaße
 - **Statistische Streuungsmaße**
 - **Varianz und Standardabweichung**
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"

Statistische Excel-Standard-Funktionen (Auswahl):

Modus (Modalwert):	MODALWERT (<i>Matrix</i>)
Median:	MEDIAN (<i>Matrix</i>)
Quartile:	QUARTILE (<i>Matrix;Funktion</i>)
	Funktionen:
	1 = Unteres Quartil
	2 = Median
	3 = Oberes Quartil
Spannweite (Minimalwert):	MIN (<i>Matrix</i>)
Spannweite (Maximalwert):	MAX (<i>Matrix</i>)
(arithmetischer) Mittelwert:	MITTELWERT (<i>Matrix</i>)



Box Plot



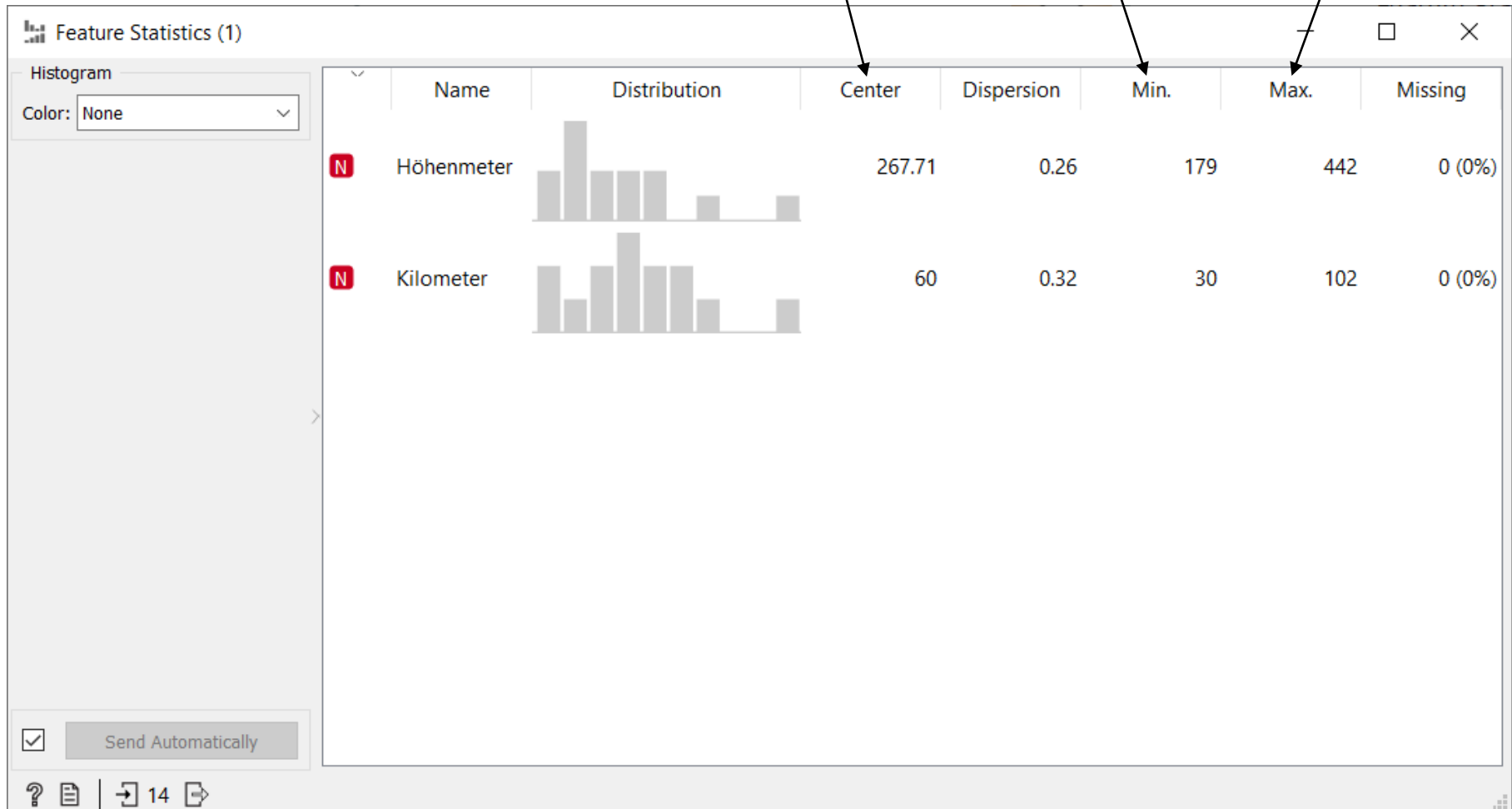


Feature Statistics

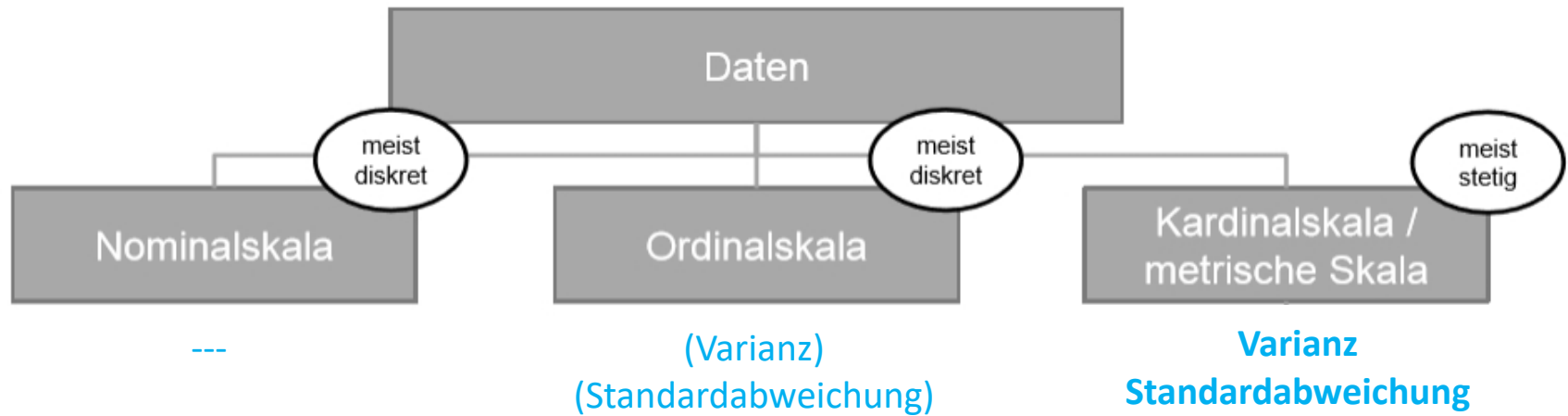
Mittelwert

Minimalwert
(Spannweite)

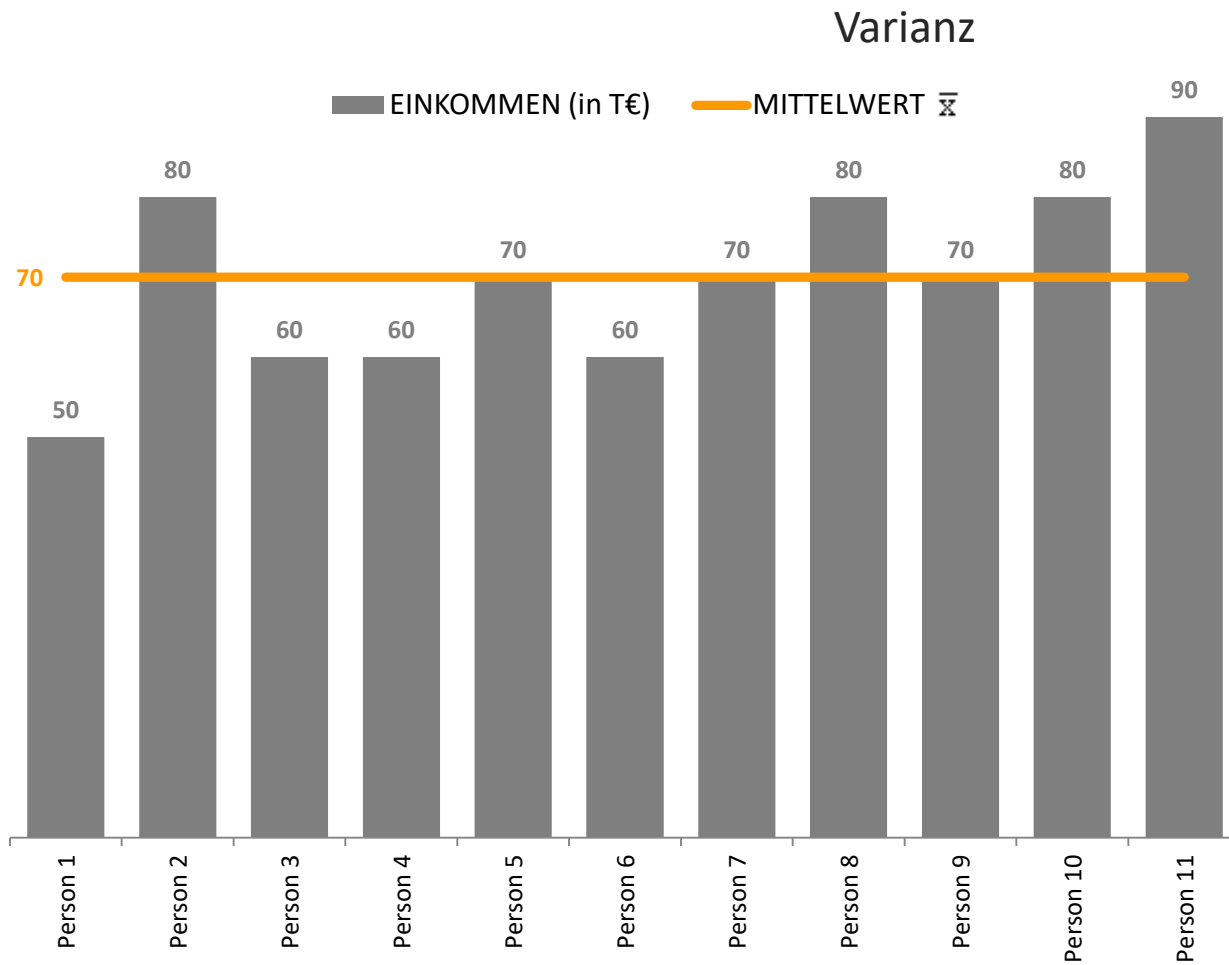
Maximalwert
(Spannweite)

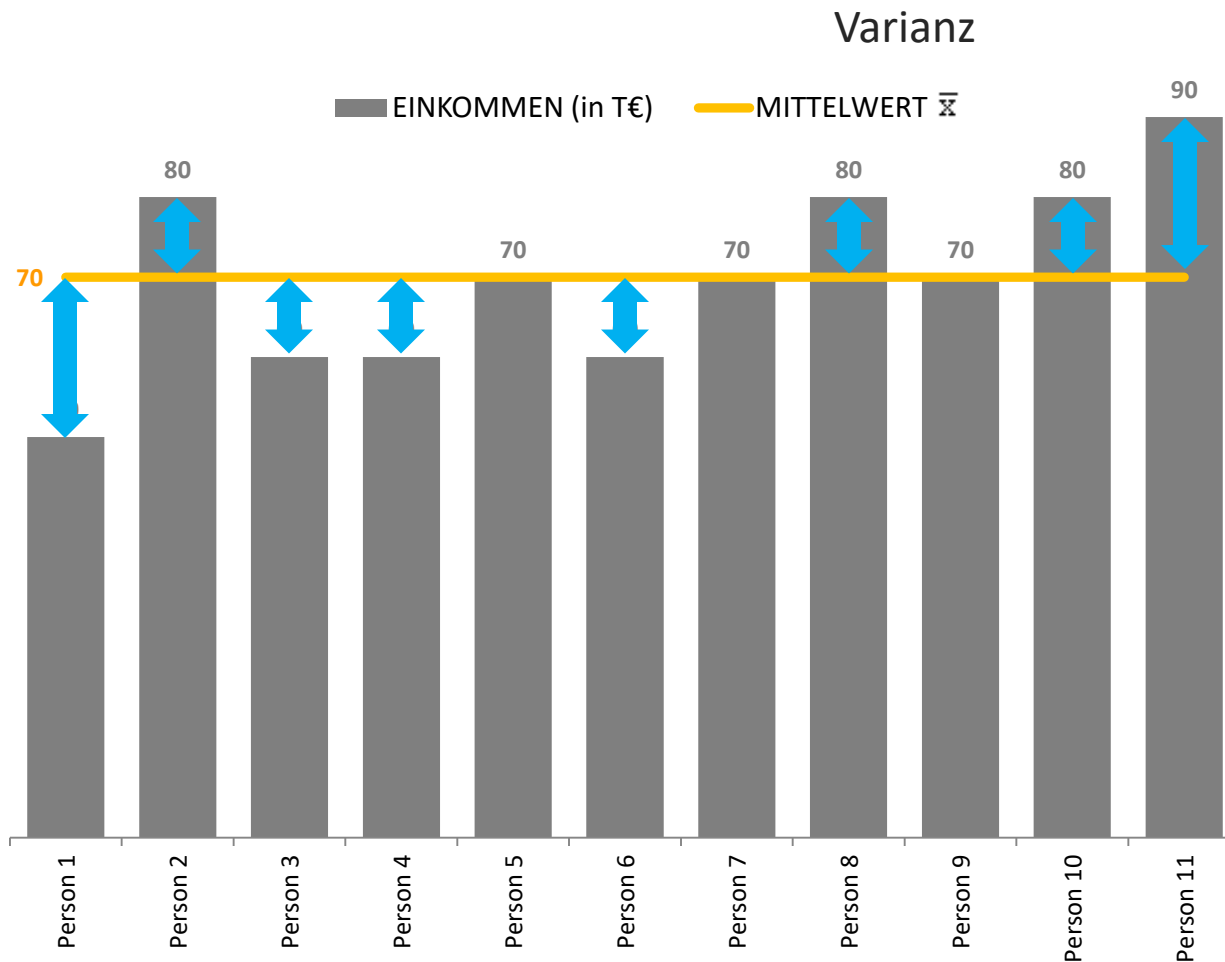


- **Univariate Analyse**
 - Skalenniveaus von Daten und statistische Lagemaße
 - Statistische Lagemaße
 - Modus, Median, Quartil, Interquartilsabstand, Spannweite und Mittelwert
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"
- **Skalenniveaus von Daten und statistische Streuungsmaße**
- Statistische Streuungsmaße
 - Varianz und Standardabweichung
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"



- **Univariate Analyse**
 - Skalenniveaus von Daten und statistische Lagemaße
 - Statistische Lagemaße
 - Modus, Median, Quartil, Interquartilsabstand, Spannweite und Mittelwert
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"
 - Skalenniveaus von Daten und statistische Streuungsmaße
 - **Statistische Streuungsmaße**
 - **Varianz und Standardabweichung**
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"





$$x_i - \bar{x}$$

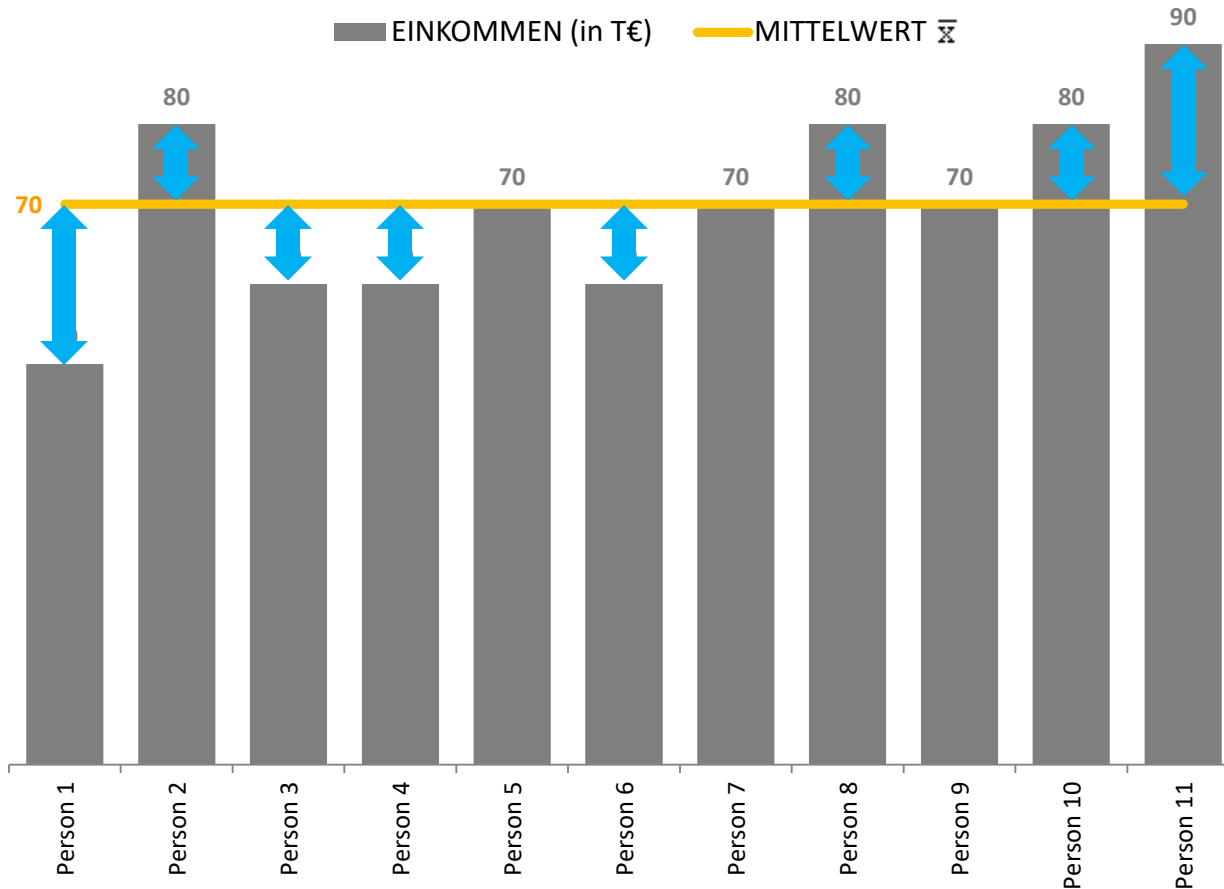
$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

Diagram explaining the variance formula:

- s^2 : Varianz
- $\sum_{i=1}^n$: Laufvariable
- x_i : i'ter Rohwert
- \bar{x} : arithmetischer Mittelwert aller Rohwerte
- n : Anzahl Rohwerte

$$s^2 = \frac{(50-70)^2 + (80-70)^2 + (60-70)^2 + \dots + (90-70)^2}{11} = 127,27$$

Standardabweichung (absolut)



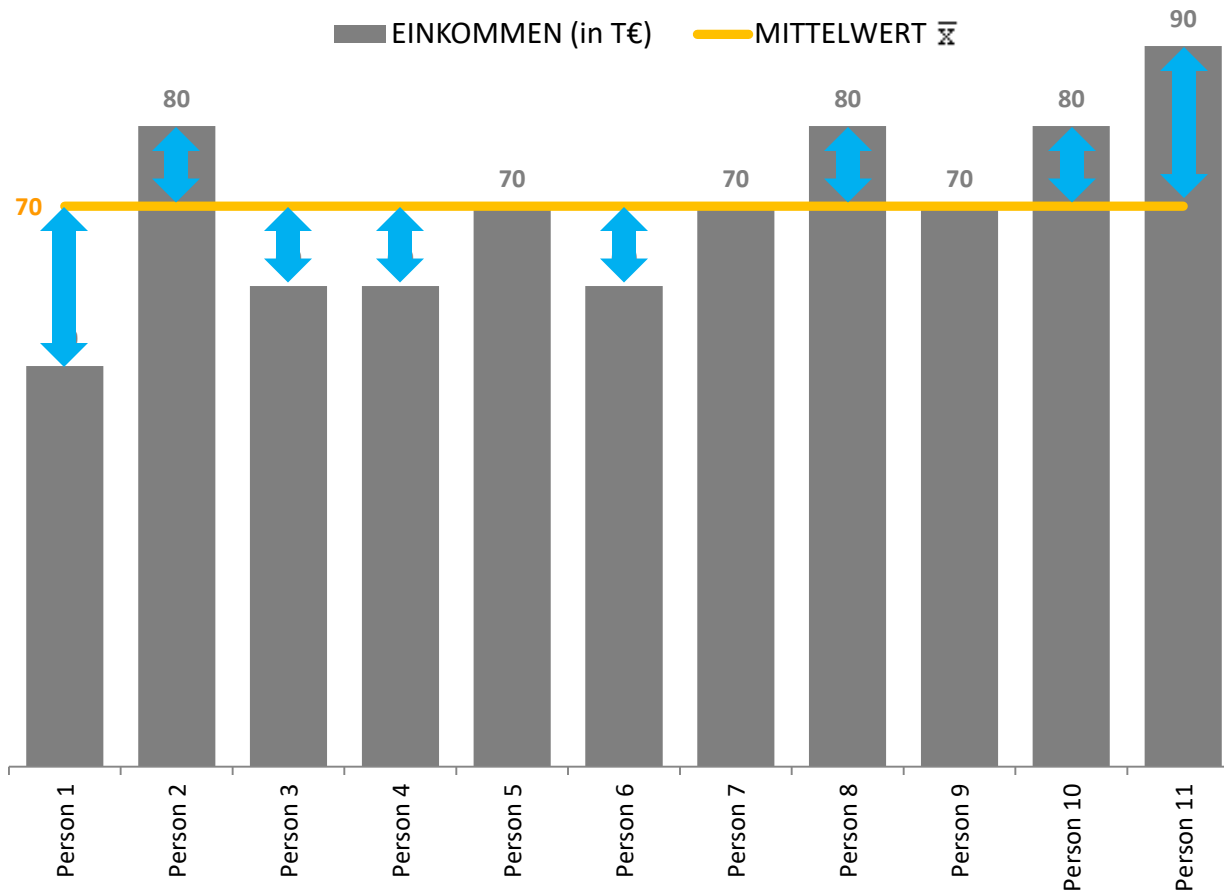
$$x_i - \bar{x}$$

$$s_{absolut} = \sqrt{\text{Varianz}}$$

$$s_{absolut} = \sqrt{s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}$$

$$s_{absolut} = \sqrt{127,27} = 11,28$$

Standardabweichung (relativ)



$$x_i - \bar{x}$$

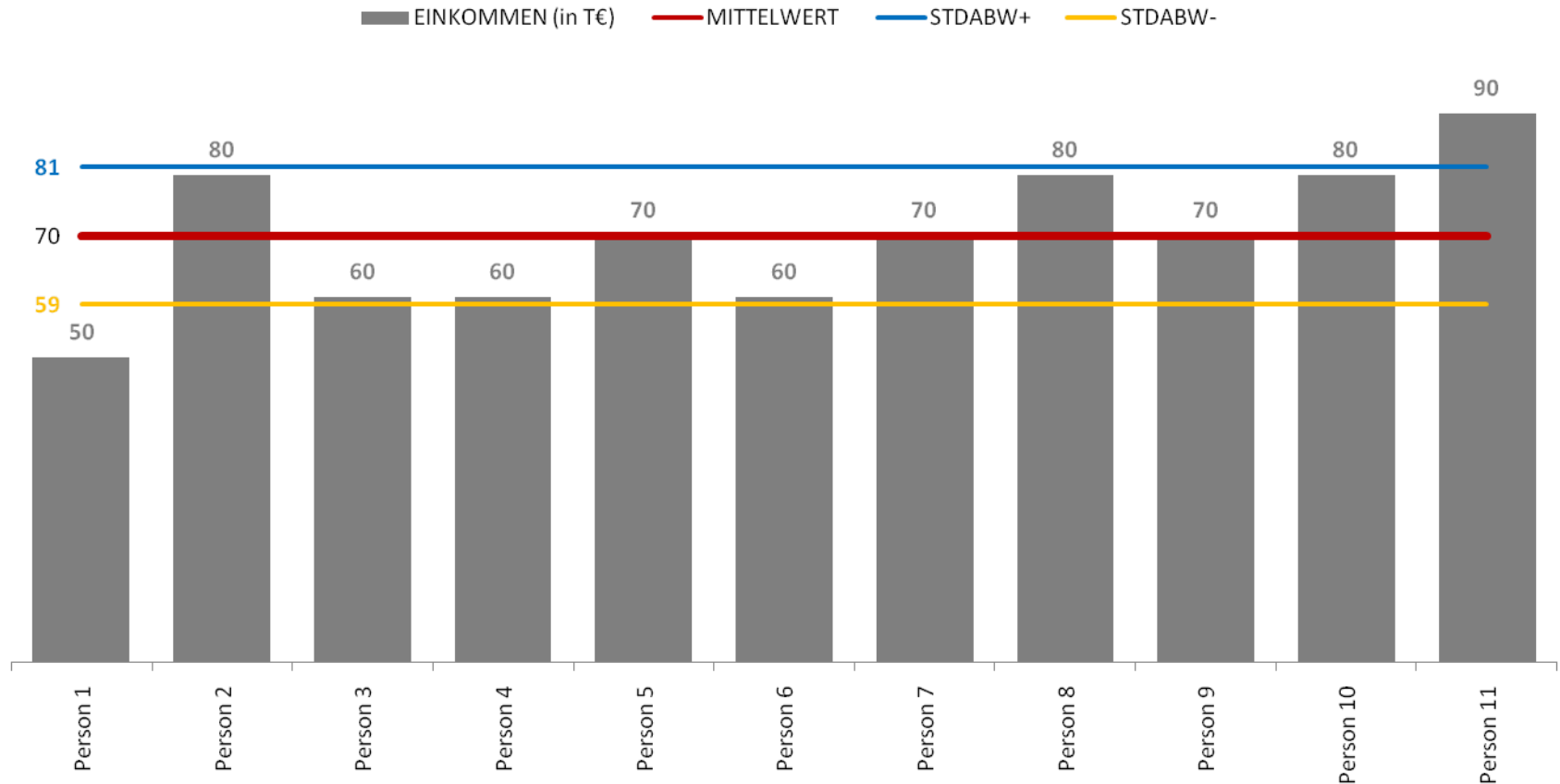
$$s_{absolut} = \sqrt{\text{Varianz}}$$

$$s_{absolut} = \sqrt{s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}$$

$$s_{relativ} = \frac{\sqrt{s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}}{\bar{x}}$$

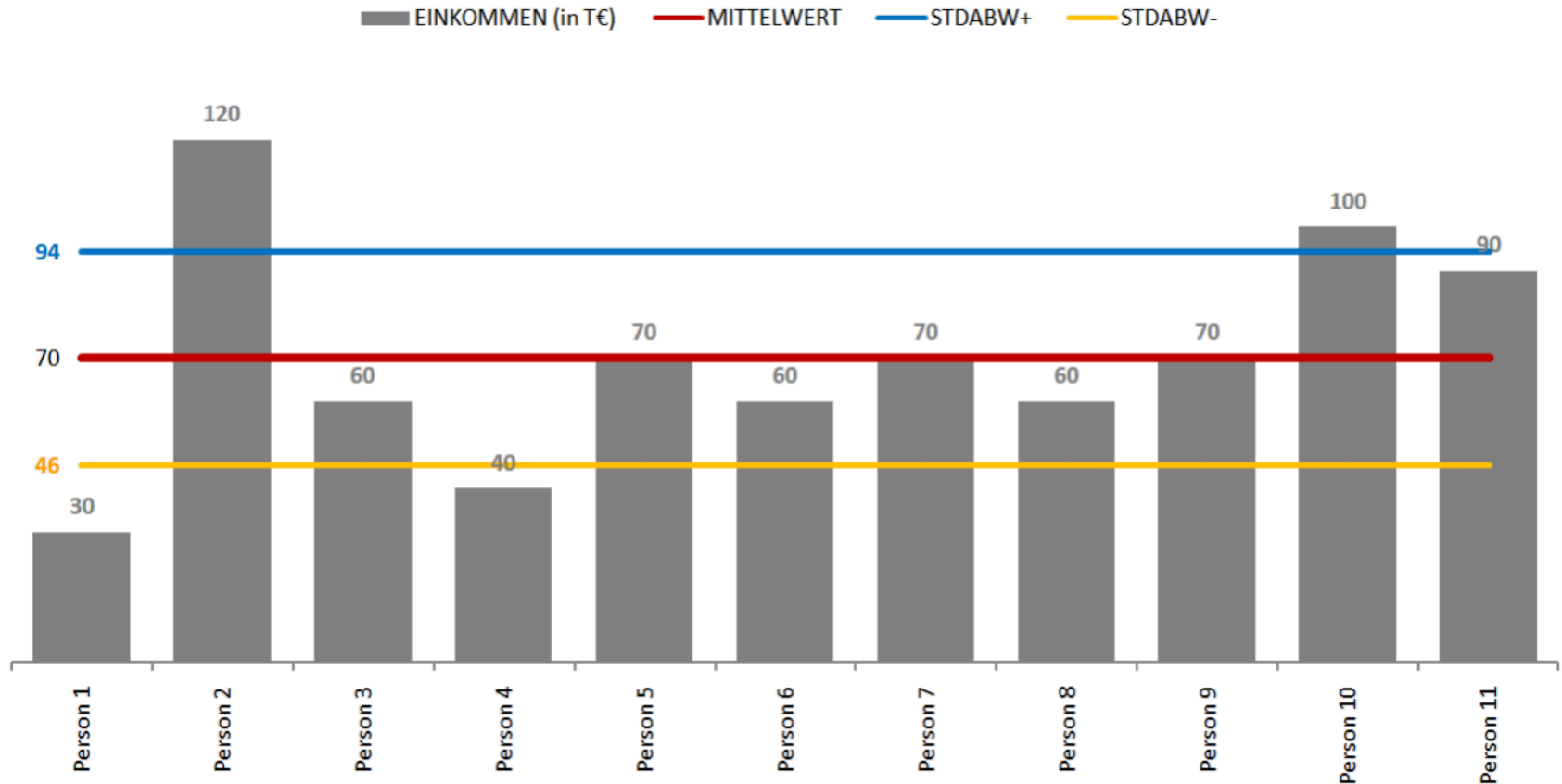
$$s_{relativ} = \frac{11,28}{70} = 0,16 = 16\%$$

Mittelwert und Standardabweichung (absolut)



(arithmetischer) Mittelwert: 70
Standardabweichung (absolut): 11

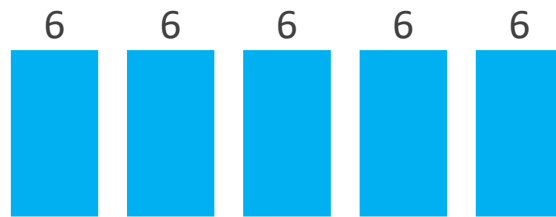
Mittelwert und Standardabweichung (absolut)



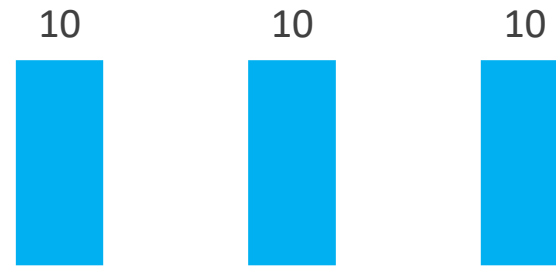
(arithmetischer) Mittelwert: 70
Standardabweichung (absolut): 24

Mittelwert und Standardabweichung (absolut)

Beispiel Schulnoten



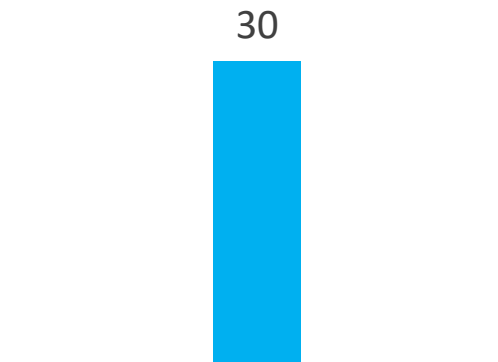
(arithmetischer) Mittelwert: 3,0
Standardabweichung (absolut): 1,4



(arithmetischer) Mittelwert: 3,0
Standardabweichung (absolut): 1,6



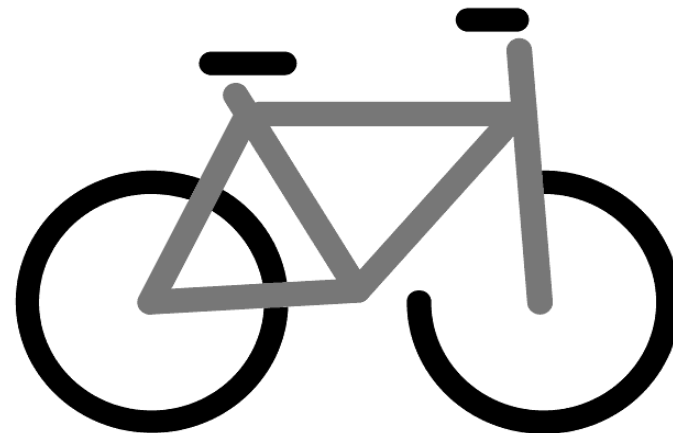
(arithmetischer) Mittelwert: 3,0
Standardabweichung (absolut): 2,0



(arithmetischer) Mittelwert: 3,0
Standardabweichung (absolut): 0

Fallstudie „Fahrradtour“

Tag	Kilometer	Höhenmeter
Tag 1	66	227
Tag 2	85	179
Tag 3	55	267
Tag 4	32	363
Tag 5	73	232
Tag 6	55	261
Tag 7	42	304
Tag 8	30	442
Tag 9	102	191
Tag 10	48	313
Tag 11	53	289
Tag 12	75	213
Tag 13	60	249
Tag 14	64	218



Fallstudie „Fahrradtour“

Tag	Kilometer	Höhenmeter
Tag 1	66	227
Tag 2	85	179
Tag 3	55	267
Tag 4	32	363
Tag 5	73	232
Tag 6	55	261
Tag 7	42	304
Tag 8	30	442
Tag 9	102	191
Tag 10	48	313
Tag 11	53	289
Tag 12	75	213
Tag 13	60	249
Tag 14	64	218

Varianz

Varianz Kilometer =

$$\frac{(66-60)^2 + (85-60)^2 + (55-60)^2 + \dots + (64-60)^2}{14} = 363,3$$

Anzahl der Werte in der Tabelle (Urliste)

Varianz Höhenmeter =

$$\frac{(227-267,7)^2 + (179-267,7)^2 + (267-267,7)^2 + \dots + (218-267,7)^2}{14} = 4.701,8$$

Anzahl der Werte in der Tabelle (Urliste)

Fallstudie „Fahrradtour“

Tag	Kilometer	Höhenmeter
Tag 1	66	227
Tag 2	85	179
Tag 3	55	267
Tag 4	32	363
Tag 5	73	232
Tag 6	55	261
Tag 7	42	304
Tag 8	30	442
Tag 9	102	191
Tag 10	48	313
Tag 11	53	289
Tag 12	75	213
Tag 13	60	249
Tag 14	64	218

Standardabweichung (absolut)

$$\text{Standardabweichung (absolut)}_{\text{Kilometer}} = \sqrt{\text{Varianz}_{\text{Kilometer}}} = \sqrt{363,3} = 19,1$$

$$\text{Standardabweichung (absolut)}_{\text{Höhenmeter}} = \sqrt{\text{Varianz}_{\text{Höhenmeter}}} = \sqrt{4.701,8} = 68,6$$

Fallstudie „Fahrradtour“

Tag	Kilometer	Höhenmeter
Tag 1	66	227
Tag 2	85	179
Tag 3	55	267
Tag 4	32	363
Tag 5	73	232
Tag 6	55	261
Tag 7	42	304
Tag 8	30	442
Tag 9	102	191
Tag 10	48	313
Tag 11	53	289
Tag 12	75	213
Tag 13	60	249
Tag 14	64	218

Standardabweichung (relativ)

Standardabweichung (relativ) Kilometer =

$$\frac{\text{Standardabweichung (absolut) Kilometer}}{\text{Mittelwert Kilometer}} = \frac{19,1}{60,0} = 0,32 = 32\%$$

Standardabweichung (relativ) Höhenmeter =

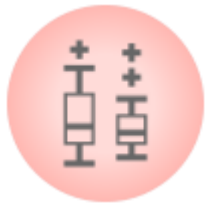
$$\frac{\text{Standardabweichung (absolut) Höhenmeter}}{\text{Mittelwert Höhenmeter}} = \frac{68,6}{267,7} = 0,26 = 26\%$$

- **Univariate Analyse**
 - Skalenniveaus von Daten und statistische Lagemaße
 - Statistische Lagemaße
 - Modus, Median, Quartil, Interquartilsabstand, Spannweite und Mittelwert
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"
 - Skalenniveaus von Daten und statistische Streuungsmaße
 - **Statistische Streuungsmaße**
 - **Varianz und Standardabweichung**
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"

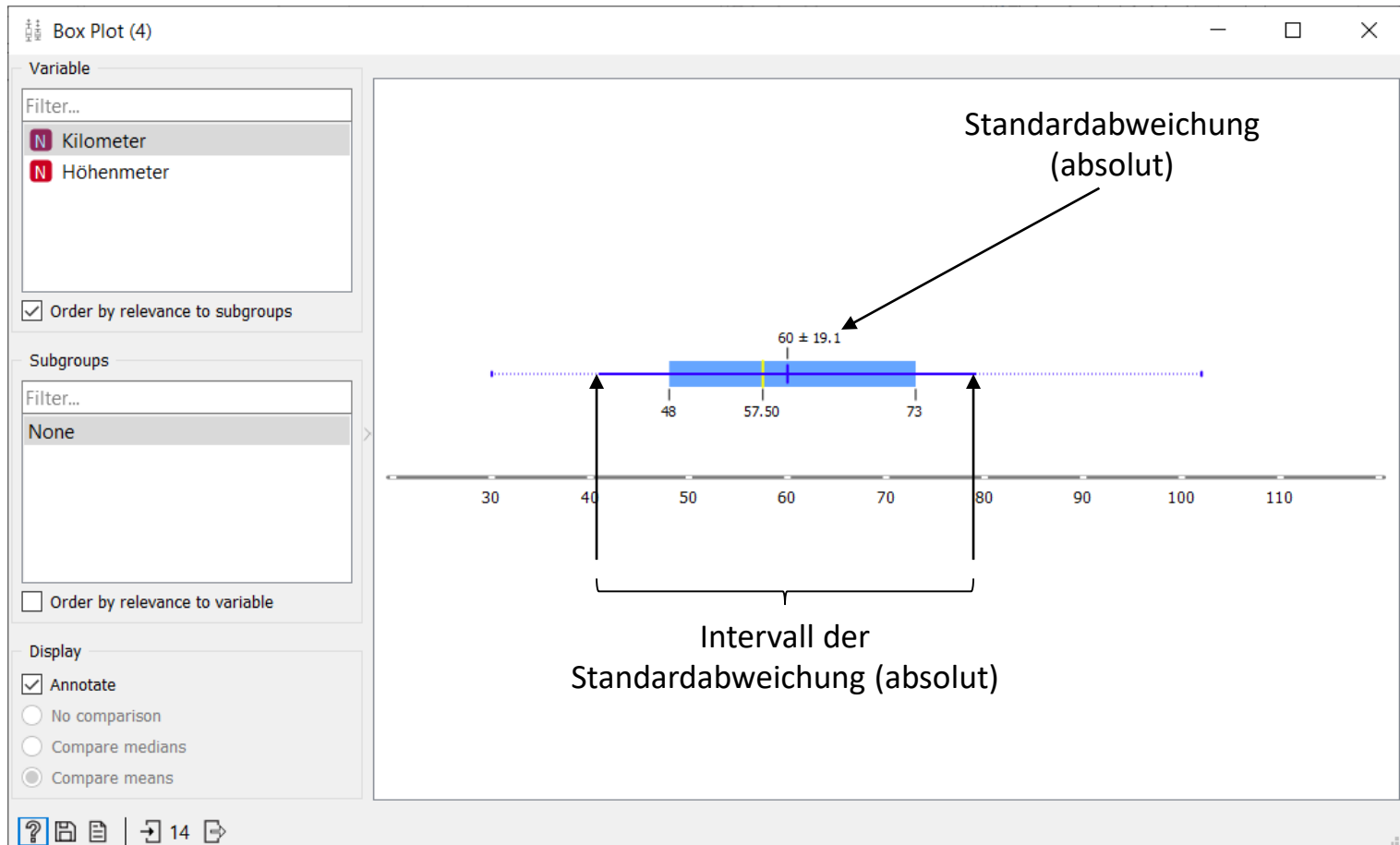
Statistische Excel-Standard-Funktionen (Auswahl):

Varianz: `VARIANZENA (Matrix)`

Standardabweichung (absolut): `STABWNA (Matrix)`



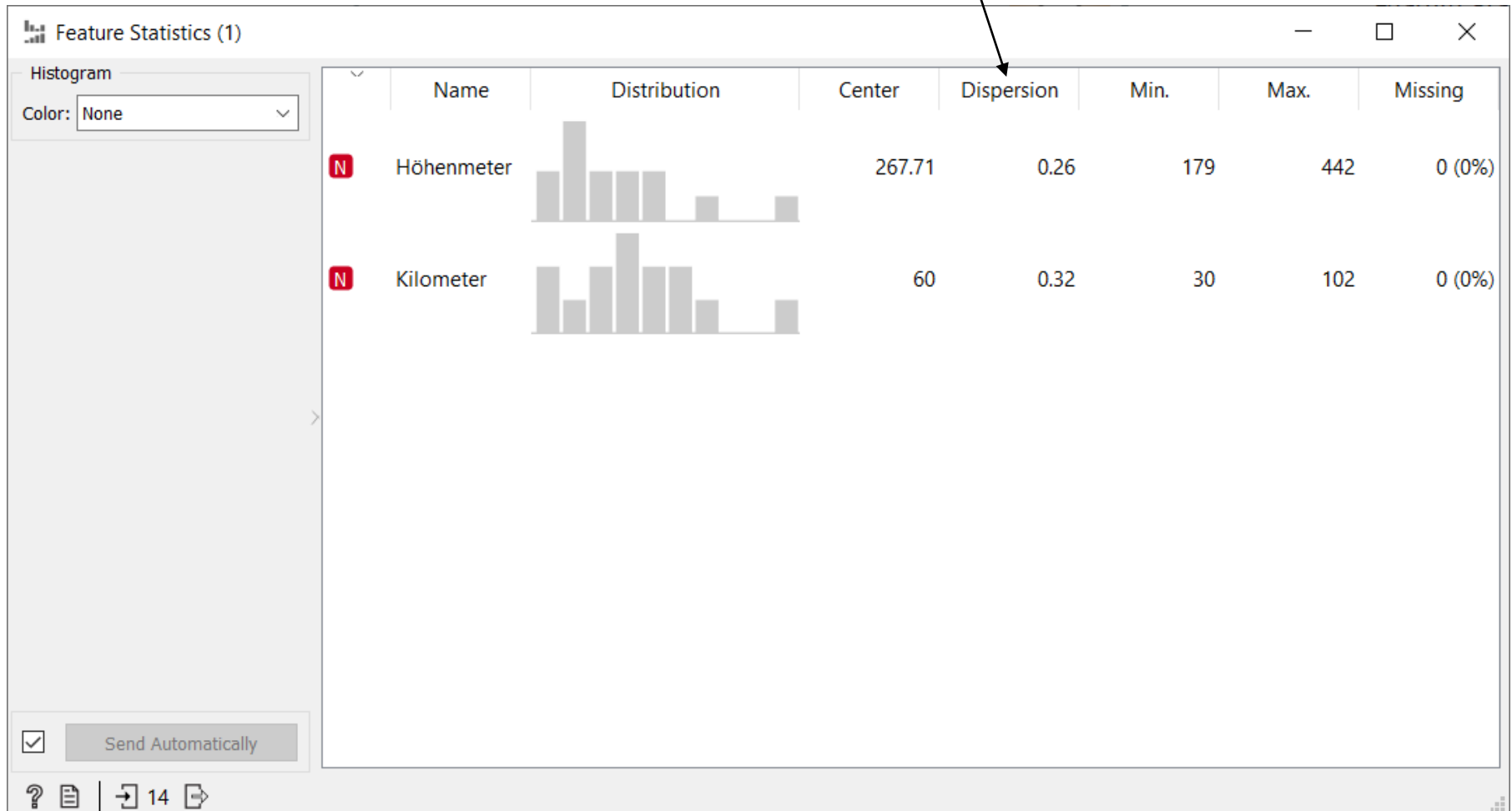
Box Plot



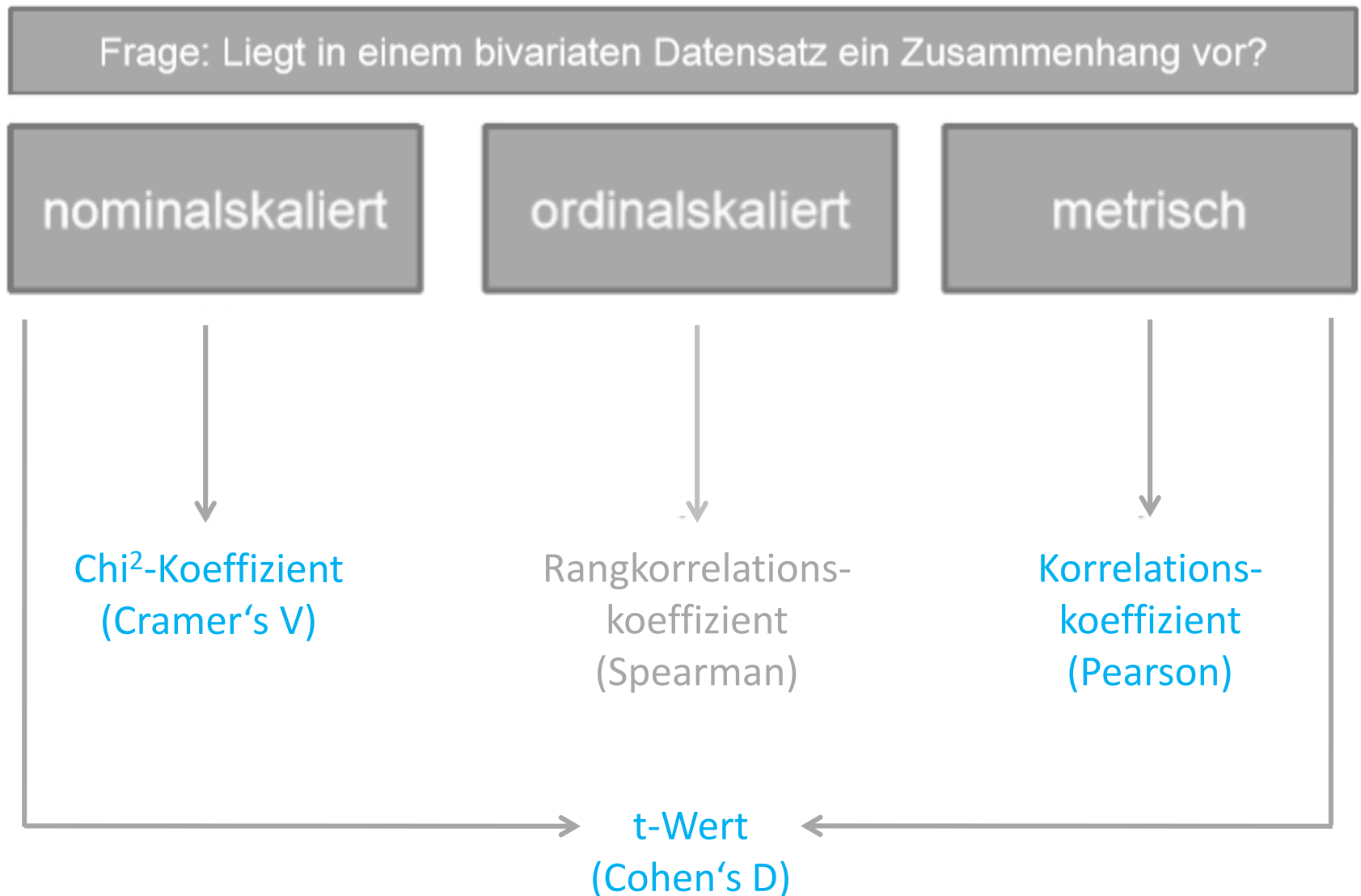


Feature Statistics

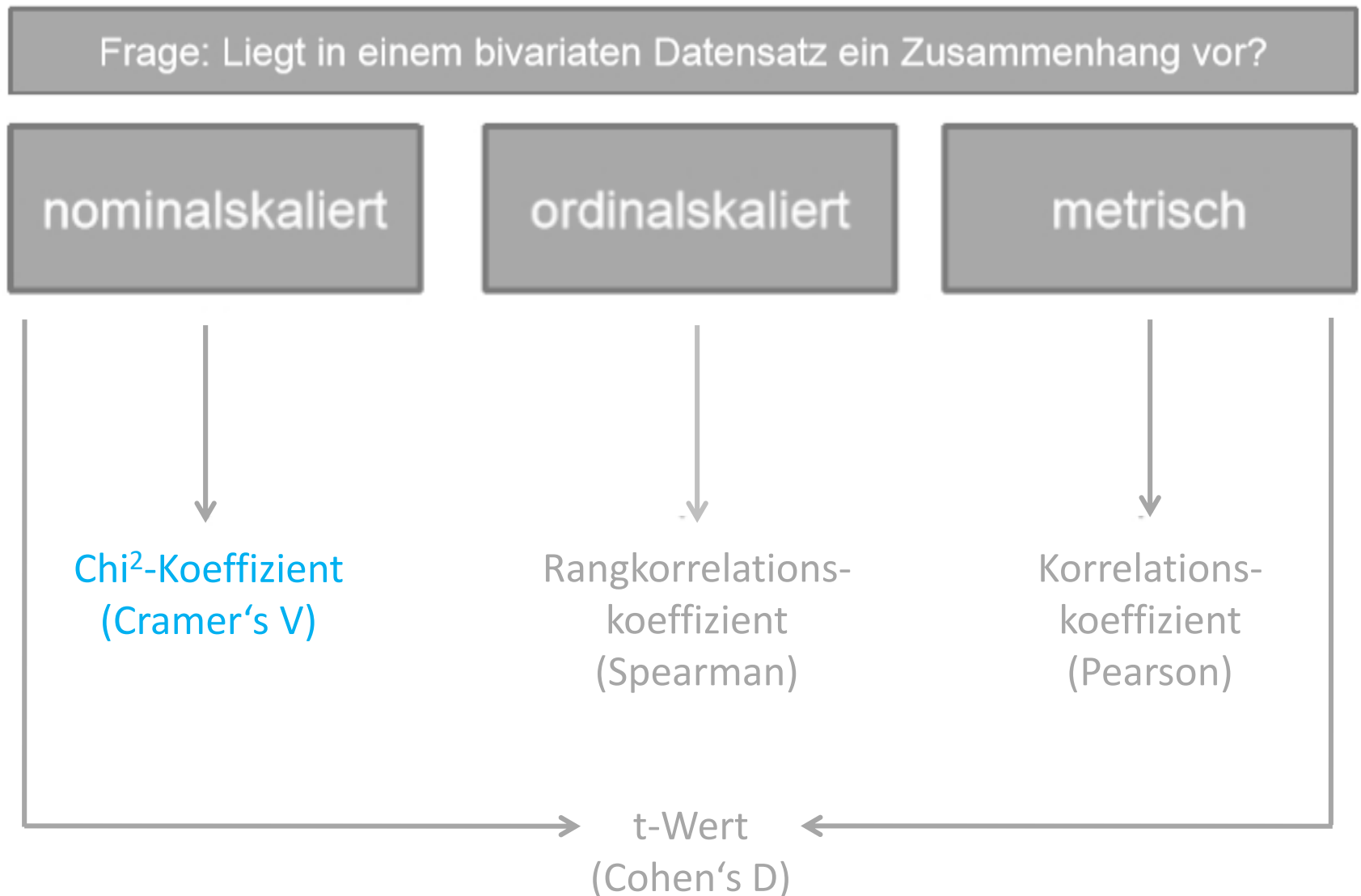
Standardabweichung
(relativ)



- Bivariate Analyse
 - Skalenniveaus von Daten und statistische Zusammenhangsmaße
 - Statistische Zusammenhangsmaße
 - Chi²-Koeffizient und Cramer's V
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"
 - Korrelationskoeffizient (Pearson)
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"
 - t-Wert und Cohen's D
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"



- Bivariate Analyse
 - Skalenniveaus von Daten und statistische Zusammenhangsmaße
 - Statistische Zusammenhangsmaße
 - χ^2 -Koeffizient und Cramer's V
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"
 - Korrelationskoeffizient (Pearson)
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"
 - t-Wert und Cohen's D
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"



Fallstudie „Studienfach“:

Gibt es bei 500 befragten Studenten/innen einen Zusammenhang zwischen dem Geschlecht und dem Studienfach?

beobachtet	BWL	Soz	VWL	SoWi	Stat	SUMME
männlich	110	120	20	30	20	300
weiblich	90	60	30	10	10	200
SUMME	200	180	50	40	30	500

1. Schritt: Berechnung der erwarteten Häufigkeiten

beobachtet	BWL	Soz	VWL	SoWi	Stat	SUMME
männlich	110	120	20	30	20	300
weiblich	90	60	30	10	10	200
SUMME	200	180	50	40	30	500



erwartet	BWL	Soz	VWL	SoWi	Stat	SUMME
männlich	120	108	30	24	18	300
weiblich	80	72	20	16	12	200
SUMME	200	180	50	40	30	500

1. Schritt: Berechnung der erwarteten Häufigkeiten

beobachtet	BWL	Soz	VWL	SoWi	Stat	SUMME
männlich	110	120	20	30	20	300
weiblich	90	60	30	10	10	200
SUMME	200	180	50	40	30	500

$$120 = \frac{200 \times 300}{500}$$

erwartet	BWL	Soz	VWL	SoWi	Stat	SUMME
männlich	120	108	30	24	18	300
weiblich	80	72	20	16	12	200
SUMME	200	180	50	40	30	500

1. Schritt: Berechnung der erwarteten Häufigkeiten

beobachtet	BWL	Soz	VWL	SoWi	Stat	SUMME
männlich	110	120	20	30	20	300
weiblich	90	60	30	10	10	200
SUMME	200	180	50	40	30	500

$$108 = \frac{180 \times 300}{500}$$

erwartet	BWL	Soz	VWL	SoWi	Stat	SUMME
männlich	120	108	30	24	18	300
weiblich	80	72	20	16	12	200
SUMME	200	180	50	40	30	500

1. Schritt: Berechnung der erwarteten Häufigkeiten

beobachtet	BWL	Soz	VWL	SoWi	Stat	SUMME
männlich	110	120	20	30	20	300
weiblich	90	60	30	10	10	200
SUMME	200	180	50	40	30	500

$$12 = \frac{30 \times 200}{500}$$

erwartet	BWL	Soz	VWL	SoWi	Stat	SUMME
männlich	120	108	30	24	18	300
weiblich	80	72	20	16	12	200
SUMME	200	180	50	40	30	500

2. Schritt: Berechnung des Chi²-Koeffizienten

beobachtet	BWL	Soz	VWL	SoWi	Stat	SUMME
männlich	110	120	20	30	20	300
weiblich	90	60	30	10	10	200
SUMME	200	180	50	40	30	500

erwartet	BWL	Soz	VWL	SoWi	Stat	SUMME
männlich	120	108	30	24	18	300
weiblich	80	72	20	16	12	200
SUMME	200	180	50	40	30	500

$$\text{Chi}^2\text{-Koeffizient} = \frac{(110 - 120)^2}{120} + \frac{(120 - 108)^2}{108} + \dots + \frac{(10 - 12)^2}{12} = 18,06$$

2. Schritt: Berechnung des Chi²-Koeffizienten

beobachtet	BWL	Soz	VWL	SoWi	Stat	SUMME
männlich	110	120	20	30	20	300
weiblich	90	60	30	10	10	200
SUMME	200	180	50	40	30	500

erwartet	BWL	Soz	VWL	SoWi	Stat	SUMME
männlich	120	108	30	24	18	300
weiblich	80	72	20	16	12	200
SUMME	200	180	50	40	30	500

$$\text{Chi}^2\text{-Koeffizient} = \frac{(110-120)^2}{120} + \frac{(120-108)^2}{108} + \dots + \frac{(10-12)^2}{12} = 18,06$$

2. Schritt: Berechnung des Chi²-Koeffizienten

beobachtet	BWL	Soz	VWL	SoWi	Stat	SUMME
männlich	110	120	20	30	20	300
weiblich	90	60	30	10	10	200
SUMME	200	180	50	40	30	500

erwartet	BWL	Soz	VWL	SoWi	Stat	SUMME
männlich	120	108	30	24	18	300
weiblich	80	72	20	16	12	200
SUMME	200	180	50	40	30	500

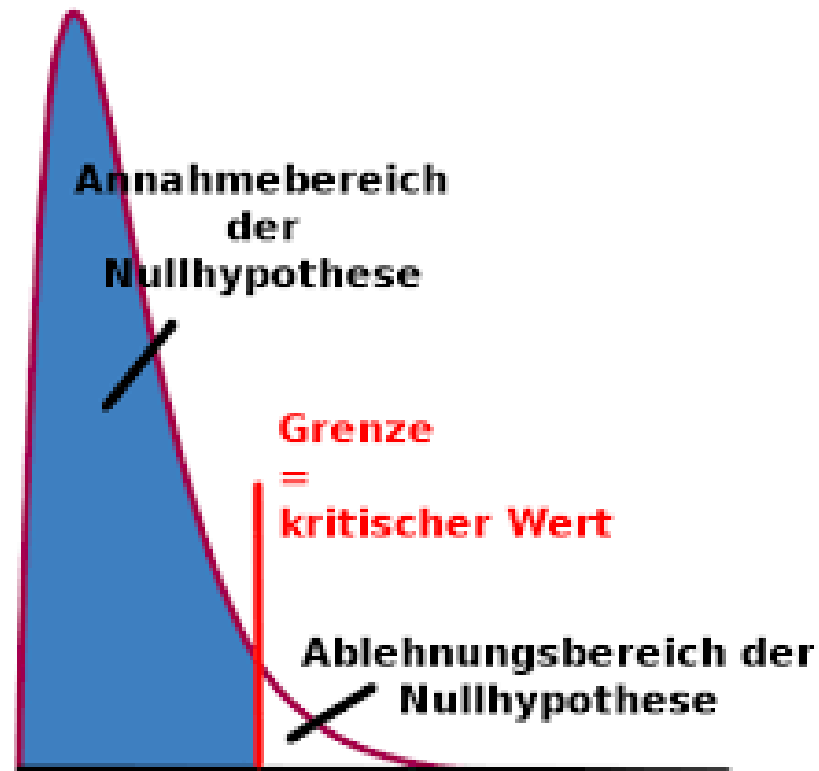
$$\text{Chi}^2\text{-Koeffizient} = \frac{(110-120)^2}{120} + \frac{(120-108)^2}{108} + \dots + \frac{(10-12)^2}{12} = 18,06$$

Frage: Sprechen 18,06 für einen Zusammenhang oder nicht? Problem: Der Chi²-Koeffizient ist für sich schwer interpretierbar!

Exkurs: H_0 -Hypothesentest

H_0 = Es besteht kein Zusammenhang zwischen Geschlecht und Studiengang

H_1 = Es besteht ein Zusammenhang zwischen Geschlecht und Studiengang



Exkurs: H₀-Hypothesentest

H₀ = Es besteht kein Zusammenhang zwischen Geschlecht und Studiengang

H₁ = Es besteht ein Zusammenhang zwischen Geschlecht und Studiengang

	P										
DF	0.995	0.975	0.2	0.1	0.05	0.025	0.02	0.01	0.005	0.002	0.001
1	.0004	.00016	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.55	10.828
2	0.01	0.0506	3.219	4.605	5.991	7.378	7.824	9.21	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.86	16.924	18.467
5	0.412	0.831	7.289	9.236	11.07	12.833	13.388	15.086	16.75	18.907	20.515
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458
7	0.989	1.69	9.803	12.017	14.067	16.013	16.622	18.475	20.278	22.601	24.322
8	1.344	2.18	11.03	13.362	15.507	17.535	18.168	20.09	21.955	24.352	26.124
9	1.735	2.7	12.242	14.684	16.919	19.023	19.679	21.666	23.589	26.056	27.877
10	2.156	3.247	13.442	15.987	18.307	20.483	21.161	23.209	25.188	27.722	29.588
11	2.603	3.816	14.631	17.275	19.675	21.92	22.618	24.725	26.757	29.354	31.264
12	3.074	4.404	15.812	18.549	21.026	23.337	24.054	26.217	28.3	30.957	32.909
13	3.565	5.009	16.985	19.812	22.362	24.736	25.472	27.688	29.819	32.535	34.528

Chi²-Verteilungstabelle:

P = Irrtumswahrscheinlichkeit

DF = Freiheitsgrad (Spalten – 1 x Zeilen – 1)

→ z.B. 0,05 = 5%

→ im Beispiel (5 – 1) x (2 – 1) = 4

Ergebnis: Da 18,06 größer ist der kritische Wert von 9,488, wird die Nullhypothese H₀ verworfen und H₁ angenommen.

Frage: Wie stark ist dieser Zusammenhang, d.h. wie groß ist die sog. „Effektstärke“?

Lösung: Cramer's V als normalisierte Metrik zur Messung der Effektstärke des Zusammenhangs!

3. Schritt: Berechnung Cramer's V

$$\text{Cramer's } V = \sqrt{\frac{\text{Chi}^2}{n \times (m - 1)}}$$

n = Gesamtzahl der Fälle (Beobachtungen)

m = Minimum von Anzahl Zeilen bzw. Spalten

3. Schritt: Berechnung Cramer's V

$$\text{Cramer's } V = \sqrt{\frac{\text{Chi}^2}{n \times (m - 1)}}$$

n = Gesamtzahl der Fälle (Beobachtungen)

m = Minimum von Anzahl Zeilen bzw. Spalten

beobachtet	BWL	Soz	VWL	SoWi	Stat	SUMME
männlich	110	120	20	30	20	300
weiblich	90	60	30	10	10	200
SUMME	200	180	50	40	30	500

$$\text{Cramer's } V = \sqrt{\frac{18,06}{500 \times (2 - 1)}} = 0,19$$

3. Schritt: Berechnung Cramer's V

$$\text{Cramer's } V = \sqrt{\frac{\text{Chi}^2}{n \times (m - 1)}}$$

n = Gesamtzahl der Fälle (Beobachtungen)

m = Minimum von Anzahl Zeilen bzw. Spalten

beobachtet	BWL	Soz	VWL	SoWi	Stat	SUMME
männlich	110	120	20	30	20	300
weiblich	90	60	30	10	10	200
SUMME	200	180	50	40	30	500

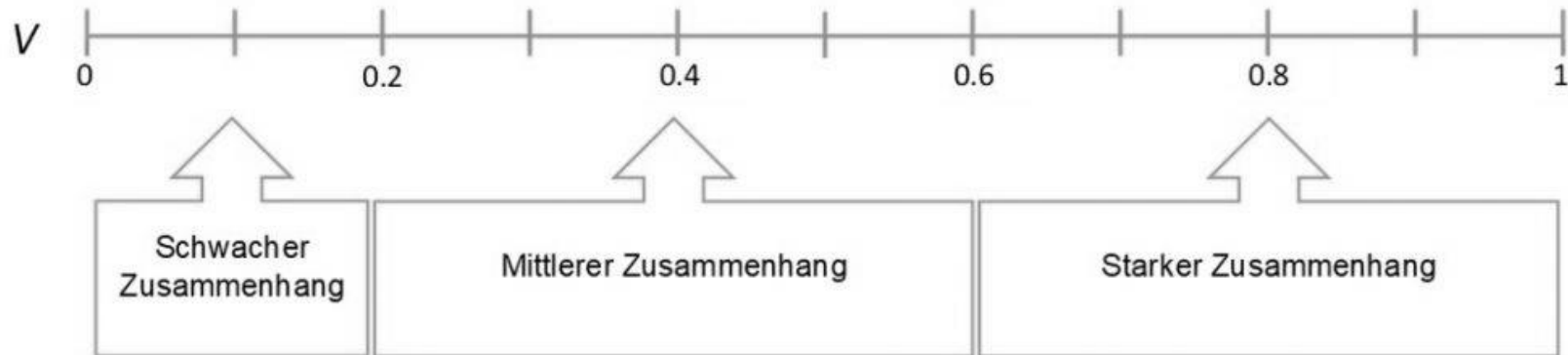
$$\text{Cramer's } V = \sqrt{\frac{18,06}{500 \times (2 - 1)}} = 0,19$$

4. Schritt: Interpretation von Cramer's V

$< 0,2$	schwacher Zusammenhang
$0,2 - 0,6$	mittlerer Zusammenhang
$> 0,6$	starker Zusammenhang

Merke: Cramer's V liegt immer zwischen 0 und 1 und ist daher eine normalisierte Metrik!

4. Schritt: Interpretation von Cramer's V



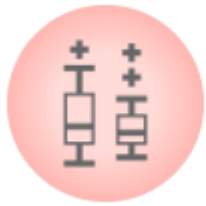
Merke: Cramer's V liegt immer zwischen 0 und 1 und ist daher eine normalisierte Metrik!

- Bivariate Analyse
 - Skalenniveaus von Daten und statistische Zusammenhangsmaße
 - Statistische Zusammenhangsmaße
 - χ^2 -Koeffizient und Cramer's V
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"
 - Korrelationskoeffizient (Pearson)
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"
 - t-Wert und Cohen's D
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"

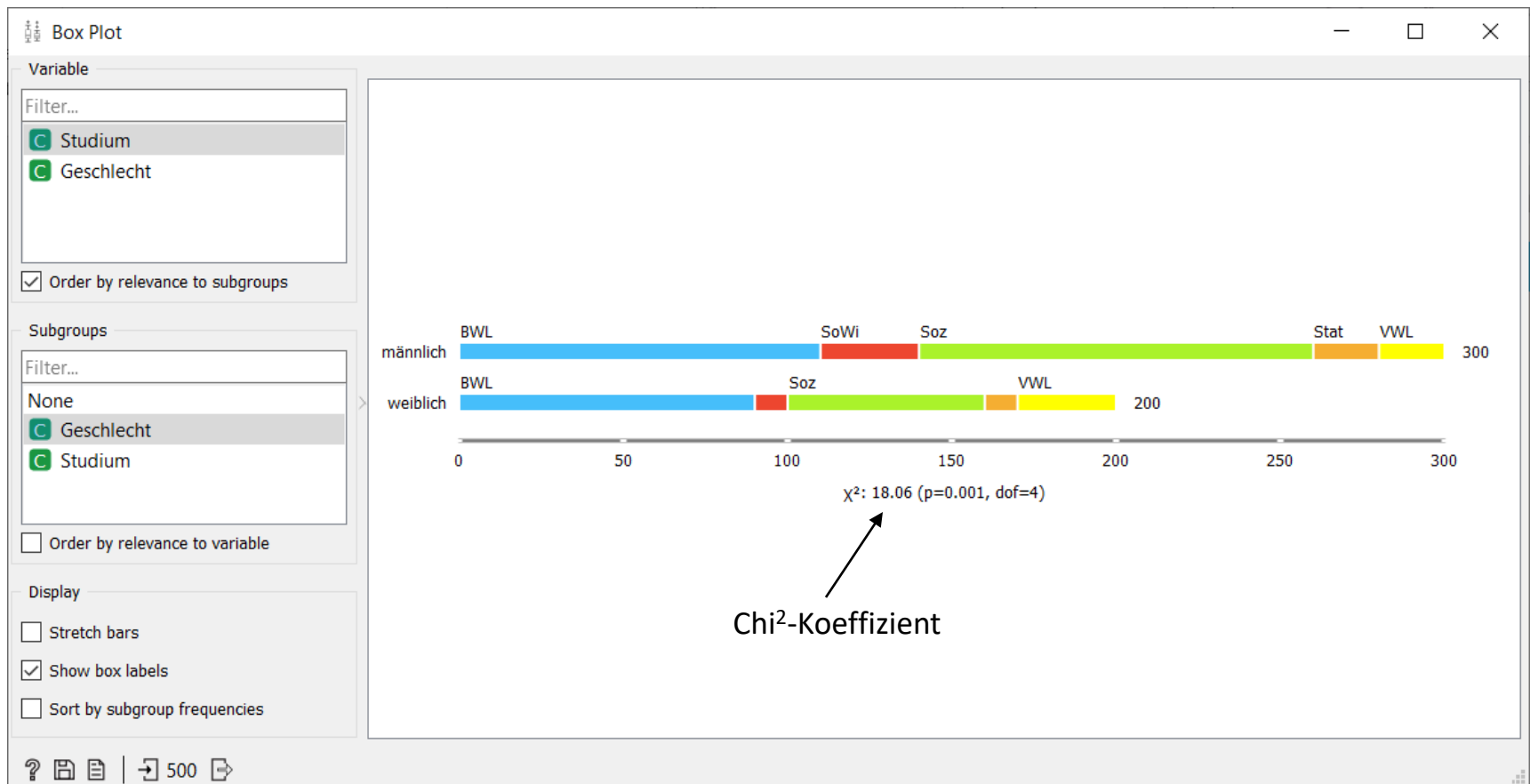
Statistische Excel-Standard-Funktionen (Auswahl):

Es existiert keine EXCEL-Standard-Funktion, die den Chi^2 -Koeffizienten oder Cramer's V berechnet!

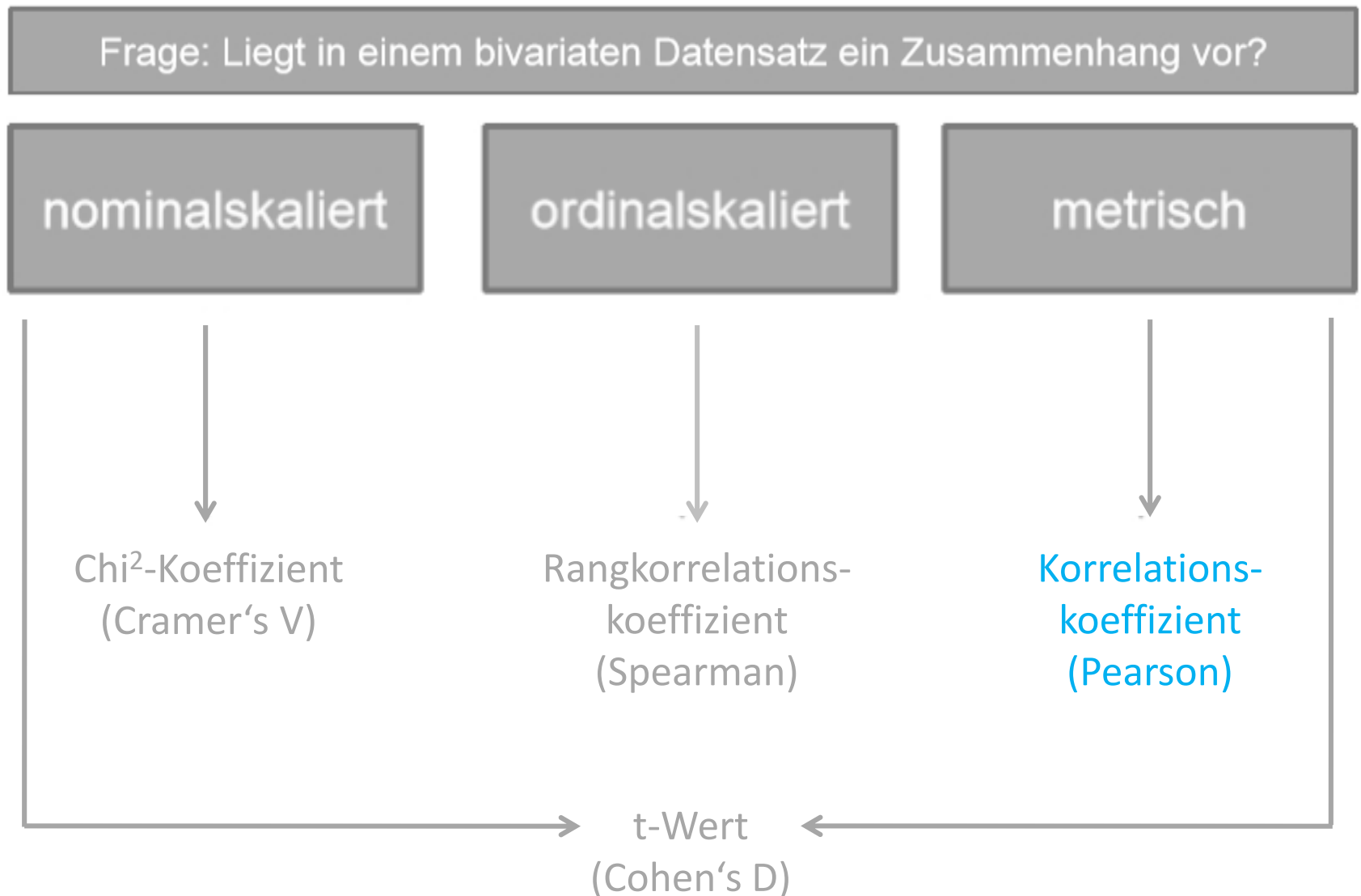




Box Plot



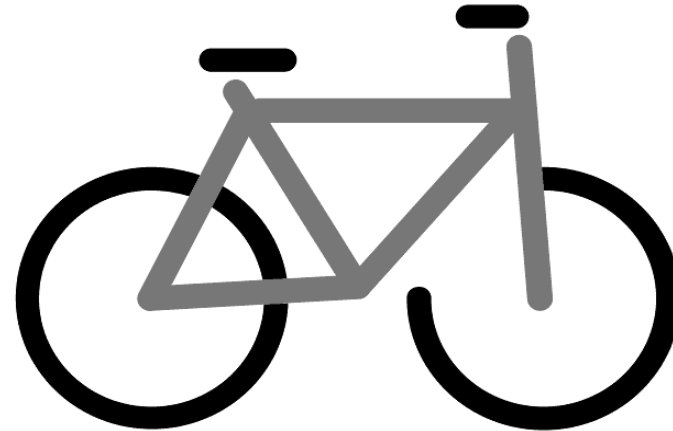
- Bivariate Analyse
 - Skalenniveaus von Daten und statistische Zusammenhangsmaße
 - Statistische Zusammenhangsmaße
 - Chi²-Koeffizient und Cramer's V
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"
 - Korrelationskoeffizient (Pearson)
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"
 - t-Wert und Cohen's D
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"



Fallstudie „Fahrradtour“:

Besteht ein Zusammenhang zwischen den täglich gefahrenen Kilometern und Höhenmetern?

Tag	Kilometer	Höhenmeter
Tag 1	66	227
Tag 2	85	179
Tag 3	55	267
Tag 4	32	363
Tag 5	73	232
Tag 6	55	261
Tag 7	42	304
Tag 8	30	442
Tag 9	102	191
Tag 10	48	313
Tag 11	53	289
Tag 12	75	213
Tag 13	60	249
Tag 14	64	218



Fallstudie „Fahrradtour“:

Besteht ein Zusammenhang zwischen den täglich gefahrenen Kilometern und Höhenmetern?

Tag	Kilometer	Höhenmeter
Tag 1	66	227
Tag 2	85	179
Tag 3	55	267
Tag 4	32	363
Tag 5	73	232
Tag 6	55	261
Tag 7	42	304
Tag 8	30	442
Tag 9	102	191
Tag 10	48	313
Tag 11	53	289
Tag 12	75	213
Tag 13	60	249
Tag 14	64	218



Lagemaß:

Mittelwert_{Kilometer} = 60,0

Mittelwert_{Höhenmeter} = 267,7



Streuungsmaß:

Standardabweichung (absolut)_{Kilometer} = 19,1

Standardabweichung (absolut)_{Höhenmeter} = 68,6

Fallstudie „Fahrradtour“:

Besteht ein Zusammenhang zwischen den täglich gefahrenen Kilometern und Höhenmetern?

Tag	Kilometer	Höhenmeter	Kilometer x Höhenmeter
Tag 1	66	227	14.982
Tag 2	85	179	15.215
Tag 3	55	267	14.685
Tag 4	32	363	11.616
Tag 5	73	232	16.936
Tag 6	55	261	14.355
Tag 7	42	304	12.768
Tag 8	30	442	13.260
Tag 9	102	191	19.482
Tag 10	48	313	15.024
Tag 11	53	289	15.317
Tag 12	75	213	15.975
Tag 13	60	249	14.940
Tag 14	64	218	13.952

Lagemaß:

Mittelwert_{Kilometer} = 60,0

Mittelwert_{Höhenmeter} = 267,7

Streuungsmaß:

Standardabweichung (absolut)_{Kilometer} = 19,1

Standardabweichung (absolut)_{Höhenmeter} = 68,6



Zusammenhangsmaß:

$$\text{Korrelationskoeffizient (Pearson)} = \frac{\text{Mittelwert (Kilometer} \times \text{Höhenmeter)} - \text{Mittelwert}_{\text{Kilometer}} \times \text{Mittelwert}_{\text{Höhenmeter}}}{\text{Standardabweichung(absolut)}_{\text{Kilometer}} \times \text{Standardabweichung(absolut)}_{\text{Höhenmeter}}}$$

Fallstudie „Fahrradtour“:

Besteht ein Zusammenhang zwischen den täglich gefahrenen Kilometern und Höhenmetern?

Tag	Kilometer	Höhenmeter	Kilometer x Höhenmeter
Tag 1	66	227	14.982
Tag 2	85	179	15.215
Tag 3	55	267	14.685
Tag 4	32	363	11.616
Tag 5	73	232	16.936
Tag 6	55	261	14.355
Tag 7	42	304	12.768
Tag 8	30	442	13.260
Tag 9	102	191	19.482
Tag 10	48	313	15.024
Tag 11	53	289	15.317
Tag 12	75	213	15.975
Tag 13	60	249	14.940
Tag 14	64	218	13.952

Lagemaß:

Mittelwert_{Kilometer} = 60,0

Mittelwert_{Höhenmeter} = 267,7

Streuungsmaß:

Standardabweichung (absolut)_{Kilometer} = 19,1

Standardabweichung (absolut)_{Höhenmeter} = 68,6



Kovarianz (skalierungsabhängig)

Zusammenhangsmaß:

Korrelationskoeffizient (Pearson) =
$$\frac{\text{Mittelwert (Kilometer} \times \text{Höhenmeter)} - \text{Mittelwert}_{\text{Kilometer}} \times \text{Mittelwert}_{\text{Höhenmeter}}}{\text{Standardabweichung(absolut)}_{\text{Kilometer}} \times \text{Standardabweichung(absolut)}_{\text{Höhenmeter}}}$$

Fallstudie „Fahrradtour“:

Besteht ein Zusammenhang zwischen den täglich gefahrenen Kilometern und Höhenmetern?

Tag	Kilometer	Höhenmeter	Kilometer x Höhenmeter
Tag 1	66	227	14.982
Tag 2	85	179	15.215
Tag 3	55	267	14.685
Tag 4	32	363	11.616
Tag 5	73	232	16.936
Tag 6	55	261	14.355
Tag 7	42	304	12.768
Tag 8	30	442	13.260
Tag 9	102	191	19.482
Tag 10	48	313	15.024
Tag 11	53	289	15.317
Tag 12	75	213	15.975
Tag 13	60	249	14.940
Tag 14	64	218	13.952

Lagemaß:

Mittelwert_{Kilometer} = 60,0

Mittelwert_{Höhenmeter} = 267,7

Streuungsmaß:

Standardabweichung (absolut)_{Kilometer} = 19,1

Standardabweichung (absolut)_{Höhenmeter} = 68,6



Zusammenhangsmaß:

$$\text{Korrelationskoeffizient (Pearson)} = \frac{\text{Mittelwert (Kilometer} \times \text{Höhenmeter)} - \text{Mittelwert}_{\text{Kilometer}} \times \text{Mittelwert}_{\text{Höhenmeter}}}{\text{Standardabweichung(absolut)}_{\text{Kilometer}} \times \text{Standardabweichung(absolut)}_{\text{Höhenmeter}}}$$

Normierung auf Wertebereich [-1,+1]

Fallstudie „Fahrradtour“:

Besteht ein Zusammenhang zwischen den täglich gefahrenen Kilometern und Höhenmetern?

Tag	Kilometer	Höhenmeter	Kilometer x Höhenmeter
Tag 1	66	227	14.982
Tag 2	85	179	15.215
Tag 3	55	267	14.685
Tag 4	32	363	11.616
Tag 5	73	232	16.936
Tag 6	55	261	14.355
Tag 7	42	304	12.768
Tag 8	30	442	13.260
Tag 9	102	191	19.482
Tag 10	48	313	15.024
Tag 11	53	289	15.317
Tag 12	75	213	15.975
Tag 13	60	249	14.940
Tag 14	64	218	13.952

Lagemaß:

$$\text{Mittelwert}_{\text{Kilometer}} = 60,0$$

$$\text{Mittelwert}_{\text{Höhenmeter}} = 267,7$$

Streuungsmaß:

$$\text{Standardabweichung (absolut)}_{\text{Kilometer}} = 19,1$$

$$\text{Standardabweichung (absolut)}_{\text{Höhenmeter}} = 68,6$$

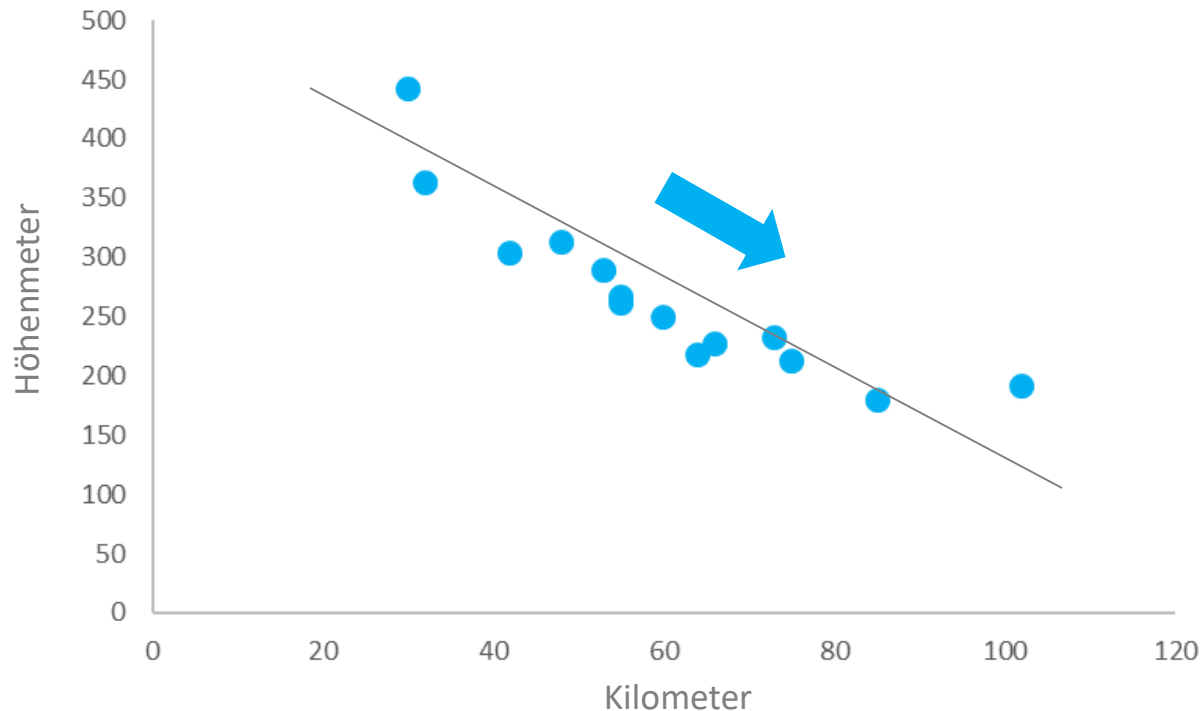


Zusammenhangsmaß:

$$\text{Korrelationskoeffizient (Pearson)} = \frac{14.893 - 60,0 \times 267,7}{19,1 \times 68,6} = \frac{-1.069,5}{1.306,9} = -0,895$$

Fallstudie „Fahrradtour“:

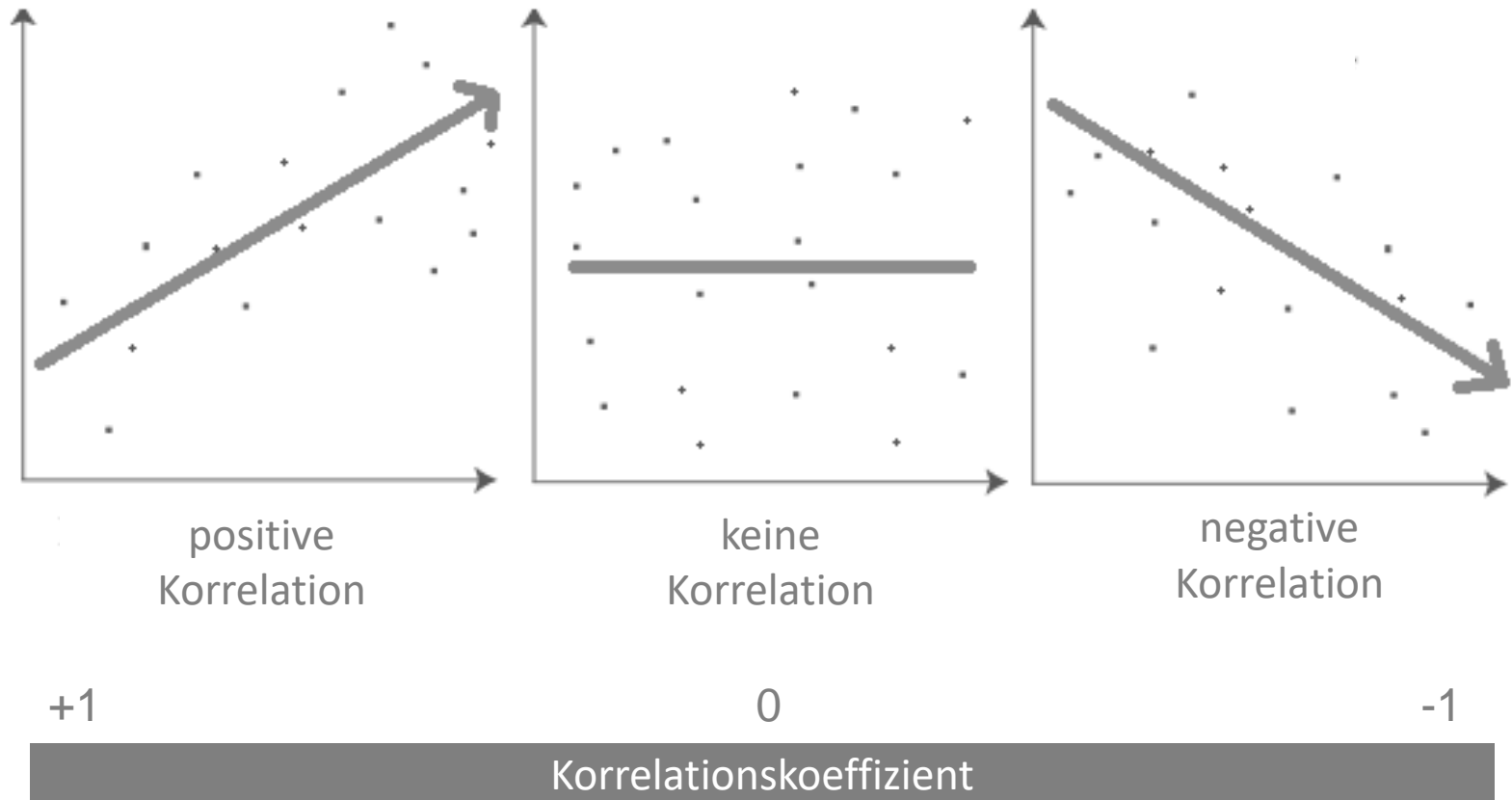
Besteht ein Zusammenhang zwischen den täglich gefahrenen Kilometern und Höhenmetern?



Korrelationskoeffizient (Pearson): -0,895

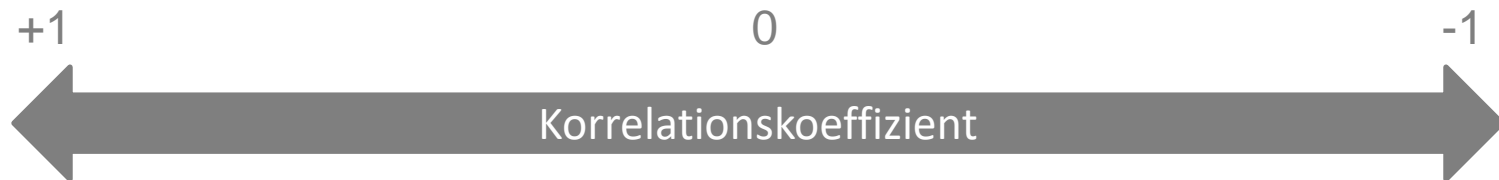
Achtung: (negativer) statistischer Zusammenhang, aber keine Kausalität!

Korrelationskoeffizient (Pearson)



Interpretation des Korrelationskoeffizienten (Pearson)

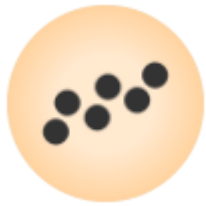
$M = 1$	perfekter positiver Zusammenhang	„Je mehr X, desto mehr Y“
$0,7 < M < 0,99$	sehr starker positiver Zusammenhang	
$0,5 < M < 0,69$	starker positiver Zusammenhang	
$0,3 < M < 0,49$	mittelstarker positiver Zusammenhang	
$0,2 < M < 0,29$	schwacher positiver Zusammenhang	
$M = 0$	statistische Unabhängigkeit, d.h. es besteht kein Zusammenhang	
$-0,2 < M < -0,29$	schwacher negativer Zusammenhang	
$-0,3 < M < -0,49$	mittelstarker negativer Zusammenhang	
$-0,5 < M < -0,69$	starker negativer Zusammenhang	
$-0,7 < M < -0,99$	sehr starker negativer Zusammenhang	
$M = -1$	perfekter negativer Zusammenhang	„Je mehr X, desto weniger Y“



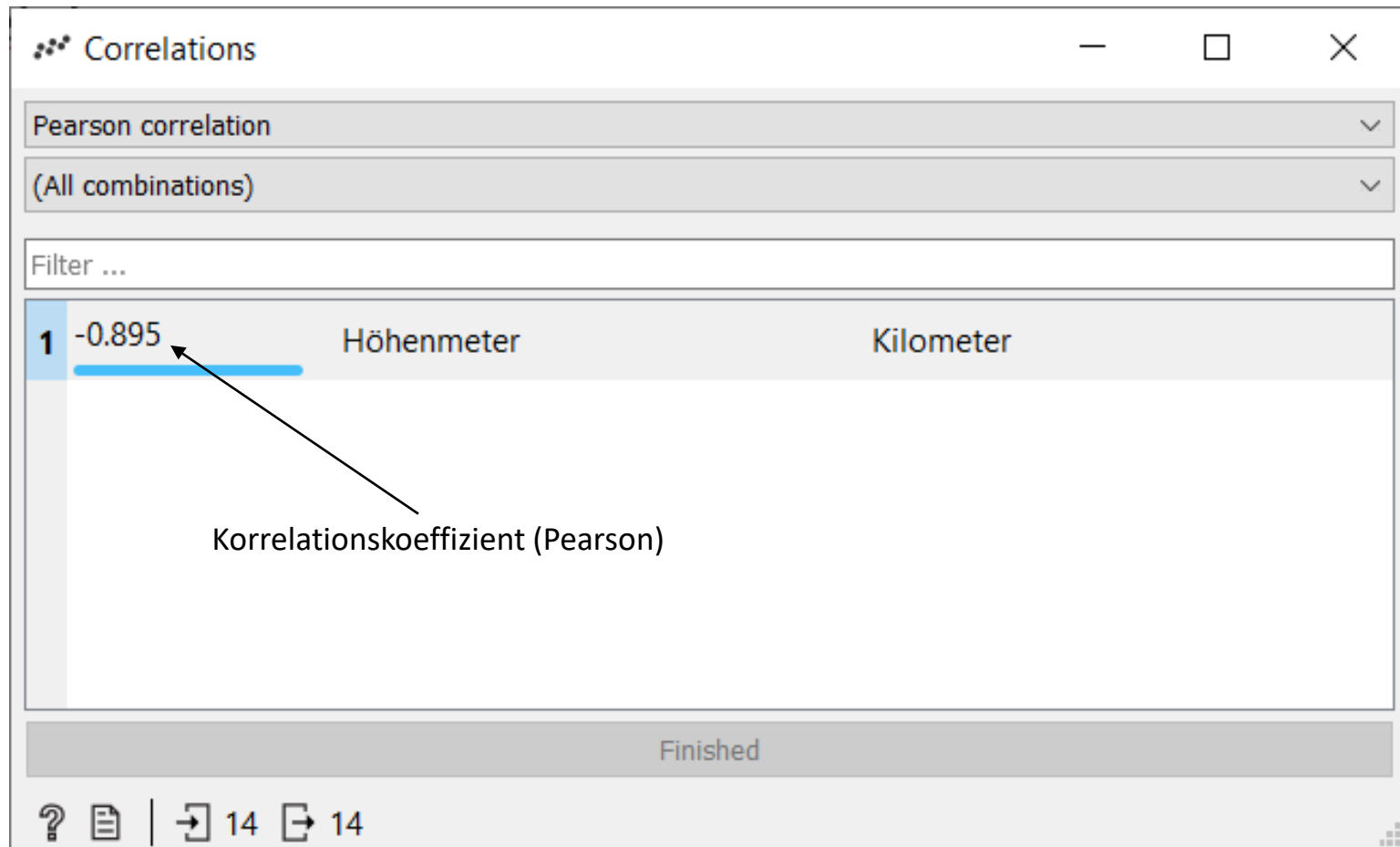
- Bivariate Analyse
 - Skalenniveaus von Daten und statistische Zusammenhangsmaße
 - Statistische Zusammenhangsmaße
 - Chi²-Koeffizient und Cramer's V
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"
 - Korrelationskoeffizient (Pearson)
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"
 - t-Wert und Cohen's D
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"

Statistische Excel-Standard-Funktionen (Auswahl):

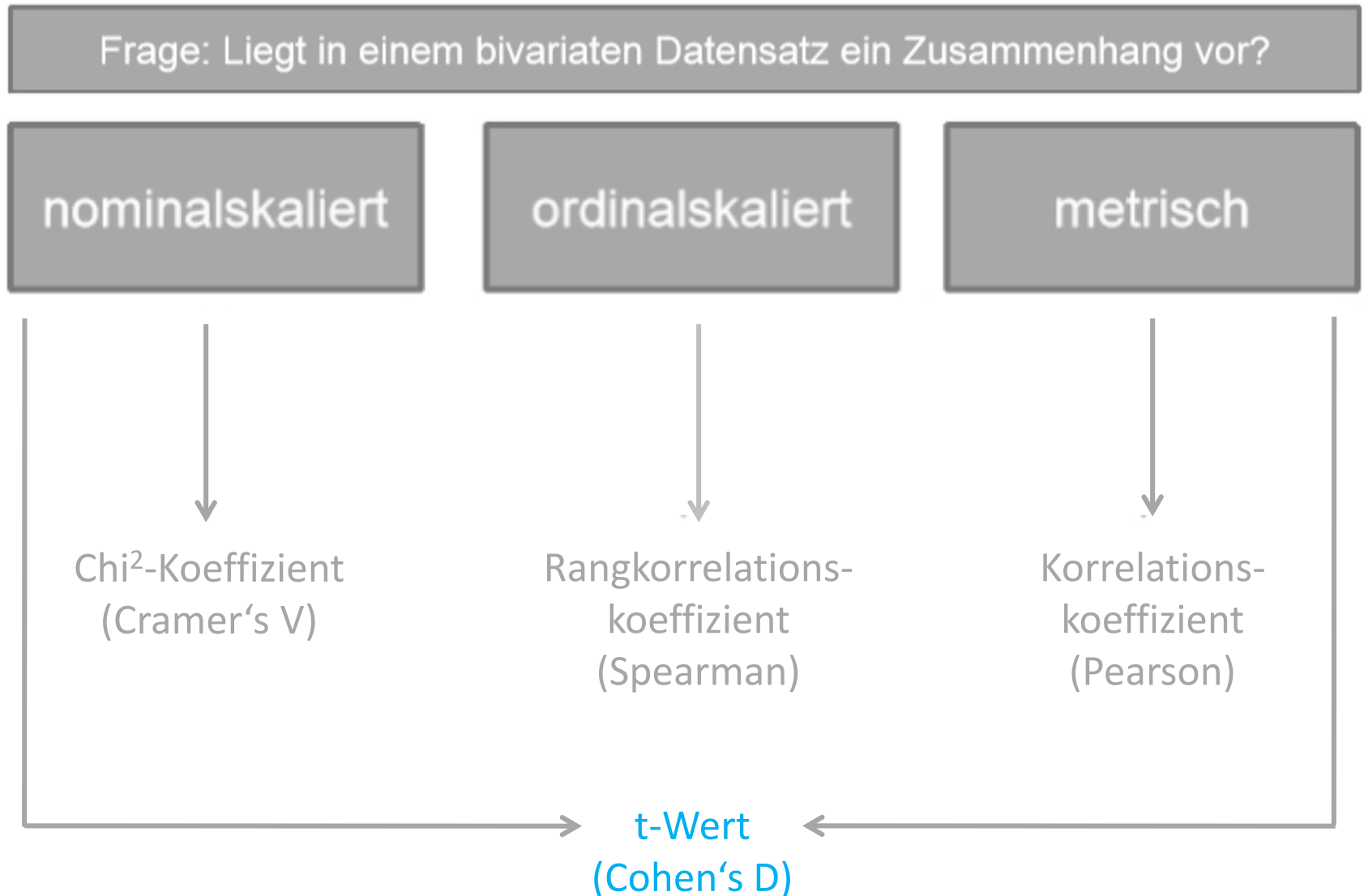
Korrelationskoeffizient (Pearson): **KORREL** (*Matrix1;Matrix2*)



Correlations



- **Bivariate Analyse**
 - Skalenniveaus von Daten und statistische Zusammenhangsmaße
 - **Statistische Zusammenhangsmaße**
 - Chi²-Koeffizient und Cramer's V
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"
 - Korrelationskoeffizient (Pearson)
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"
 - **t-Wert und Cohen's D**
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"



Der t-Wert gibt an, ob ein Zusammenhang zwischen einem nominal („kategorial“) skalierten und einem metrisch skalierten Merkmal besteht:

$$t - \text{Wert} = \frac{|\text{Mittelwert}_{\text{Kategorie 1}} - \text{Mittelwert}_{\text{Kategorie 2}}|}{\sqrt{\frac{\text{Varianz}_{\text{Kategorie 1}}}{n_1} + \frac{\text{Varianz}_{\text{Kategorie 2}}}{n_2}}}$$

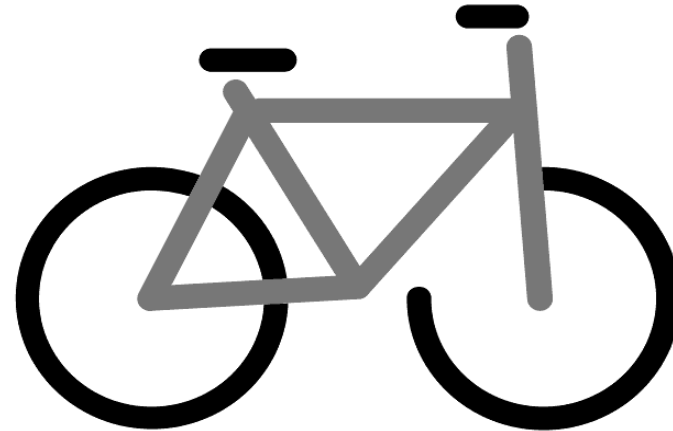
Anzahl der Beobachtungen in den jeweiligen (nominalen) Kategorien

Je höher der t-Wert, desto mehr unterscheiden sich die beiden Kategorien voneinander und desto stärker der Zusammenhang zwischen dem nominal („kategorial“) skalierten und dem metrisch skalierten Merkmal.

Fallstudie „Fahrradtour“:

Besteht ein Zusammenhang zwischen den täglich gefahrenen Kilometern (metrisch skaliert) und Regen-Tagen (nominal skaliert)?

Tag	Kilometer	Regen
Tag 1	66	nein
Tag 2	85	nein
Tag 3	55	ja
Tag 4	32	ja
Tag 5	73	ja
Tag 6	55	ja
Tag 7	42	nein
Tag 8	30	nein
Tag 9	102	ja
Tag 10	48	nein
Tag 11	53	ja
Tag 12	75	nein
Tag 13	60	nein
Tag 14	64	ja



Fallstudie „Fahrradtour“:

Besteht ein Zusammenhang zwischen den täglich gefahrenen Kilometern (metrisch skaliert) und Regen-Tagen (nominal skaliert)?

Tag	Kilometer	Regen
Tag 1	66	nein
Tag 2	85	nein
Tag 3	55	ja
Tag 4	32	ja
Tag 5	73	ja
Tag 6	55	ja
Tag 7	42	nein
Tag 8	30	nein
Tag 9	102	ja
Tag 10	48	nein
Tag 11	53	ja
Tag 12	75	nein
Tag 13	60	nein
Tag 14	64	ja

$$t - \text{Wert} = \frac{|\text{Mittelwert}_{\text{Kategorie 1}} - \text{Mittelwert}_{\text{Kategorie 2}}|}{\sqrt{\frac{\text{Varianz}_{\text{Kategorie 1}}}{n_1} + \frac{\text{Varianz}_{\text{Kategorie 2}}}{n_2}}}$$

Fallstudie „Fahrradtour“:

Besteht ein Zusammenhang zwischen den täglich gefahrenen Kilometern (metrisch skaliert) und Regen-Tagen (nominal skaliert)?

Tag	Kilometer	Regen
Tag 1	66	nein
Tag 2	85	nein
Tag 3	55	ja
Tag 4	32	ja
Tag 5	73	ja
Tag 6	55	ja
Tag 7	42	nein
Tag 8	30	nein
Tag 9	102	ja
Tag 10	48	nein
Tag 11	53	ja
Tag 12	75	nein
Tag 13	60	nein
Tag 14	64	ja

$$t - \text{Wert} = \frac{|\text{Mittelwert}_{\text{Regen=nein}} - \text{Mittelwert}_{\text{Regen=ja}}|}{\sqrt{\frac{\text{Varianz}_{\text{Regen=nein}}}{n_{\text{Regen=nein}}} + \frac{\text{Varianz}_{\text{Regen=ja}}}{n_{\text{Regen=ja}}}}}$$

$$t - \text{Wert} = \frac{|\frac{(66 + \dots + 60)}{7} - \frac{(55 + \dots + 64)}{7}|}{\sqrt{\frac{\text{Varianz}_{\text{Regen=nein}}}{7} + \frac{\text{Varianz}_{\text{Regen=ja}}}{7}}}$$

$$t - \text{Wert} = \frac{|58 - 62|}{\sqrt{\frac{(66 - 58)^2 + \dots + (60 - 58)^2}{7} + \frac{(55 - 62)^2 + \dots + (64 - 62)^2}{7}}}$$

$$t - \text{Wert} = \frac{4}{\sqrt{\frac{318,0}{7} + \frac{400,6}{7}}} = \frac{4}{10,1} = 0,39$$

Frage: Sprechen 0,39 für einen Zusammenhang oder nicht? **Problem:** Der t-Wert ist für sich schwer interpretierbar!

Lösung: Cohen's D als Metrik zur Messung der Effektstärke des Zusammenhangs!

Fallstudie „Fahrradtour“:

Besteht ein Zusammenhang zwischen den täglich gefahrenen Kilometern (metrisch skaliert) und Regen-Tagen (nominal skaliert)?

Tag	Kilometer	Regen
Tag 1	66	nein
Tag 2	85	nein
Tag 3	55	ja
Tag 4	32	ja
Tag 5	73	ja
Tag 6	55	ja
Tag 7	42	nein
Tag 8	30	nein
Tag 9	102	ja
Tag 10	48	nein
Tag 11	53	ja
Tag 12	75	nein
Tag 13	60	nein
Tag 14	64	ja

$$\text{Cohen's } D^* = \frac{t - \text{Wert}}{\sqrt{\frac{n}{4}}}$$

n = Anzahl der Werte in der Tabelle (Urliste)

$$\text{Cohen's } D^* = \frac{0,39}{\sqrt{\frac{14}{4}}} = \frac{0,39}{1,87} = 0,21$$

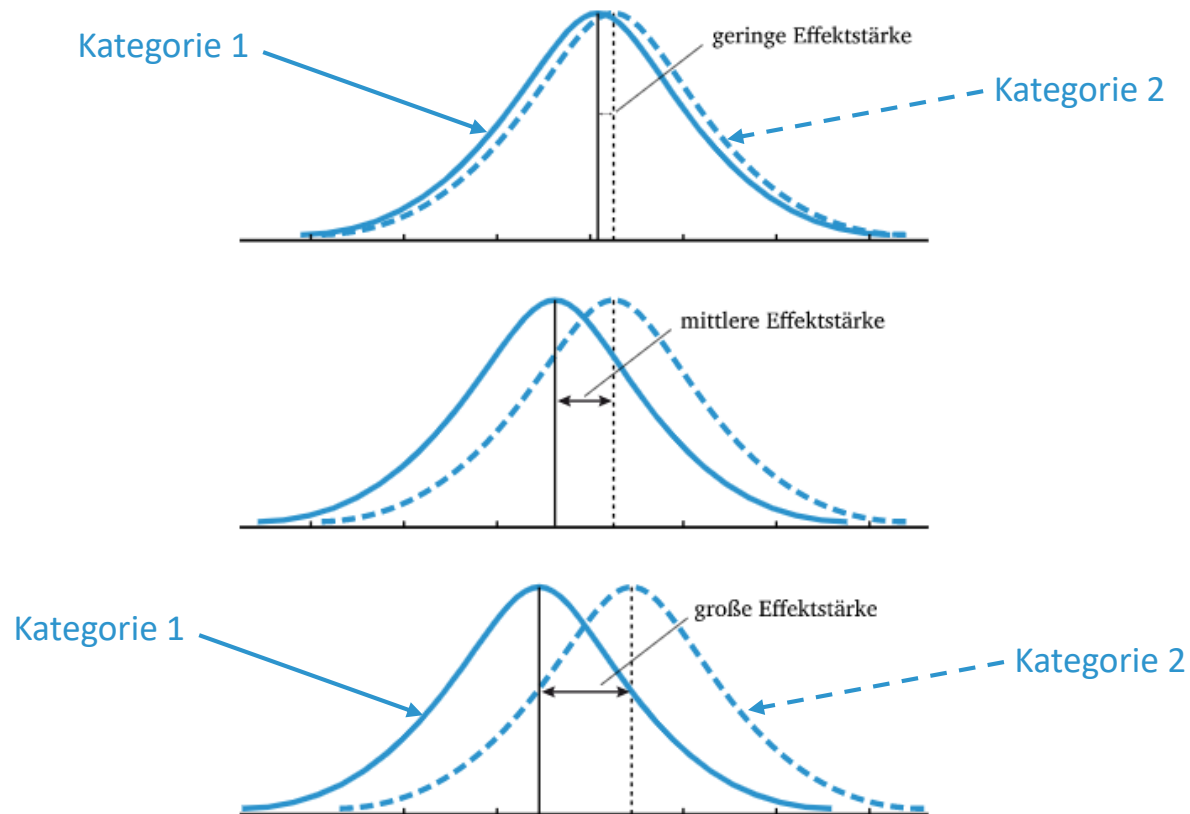
* = für den Fall $n_1 = n_2$

Interpretation von Cohen's D

$< 0,2$	kein Zusammenhang
$0,2 - 0,5$	schwacher Zusammenhang
$0,5 - 0,8$	mittlerer Zusammenhang
$> 0,8$	starker Zusammenhang

Merke: Cohen's D (gem. vorstehender Formeln) liegt zwischen 0 und ∞ und ist daher keine normalisierte Metrik!

Interpretation von Cohen's D



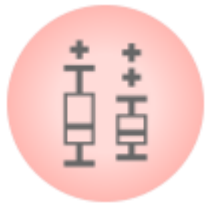
Merke: Cohen's D (gem. vorstehender Formeln) liegt zwischen 0 und ∞ und ist daher keine normalisierte Metrik!

- Bivariate Analyse
 - Skalenniveaus von Daten und statistische Zusammenhangsmaße
 - Statistische Zusammenhangsmaße
 - Chi²-Koeffizient und Cramer's V
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"
 - Korrelationskoeffizient (Pearson)
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"
 - t-Wert und Cohen's D
 - Fallstudie in "EXCEL"
 - Fallstudie in "ORANGE"

Statistische Excel-Standard-Funktionen (Auswahl):

Es existiert keine EXCEL-Standard-Funktion, die den t -Wert oder Cohen's D berechnet!





Box Plot

