LV Datengestützte Analysemethoden

4. DATEN AUSWERTEN

Sommersemester 2018
FH Joanneum Graz
Studiengang Journalismus und Public Relations

Lehrender: Stefan Kasberger

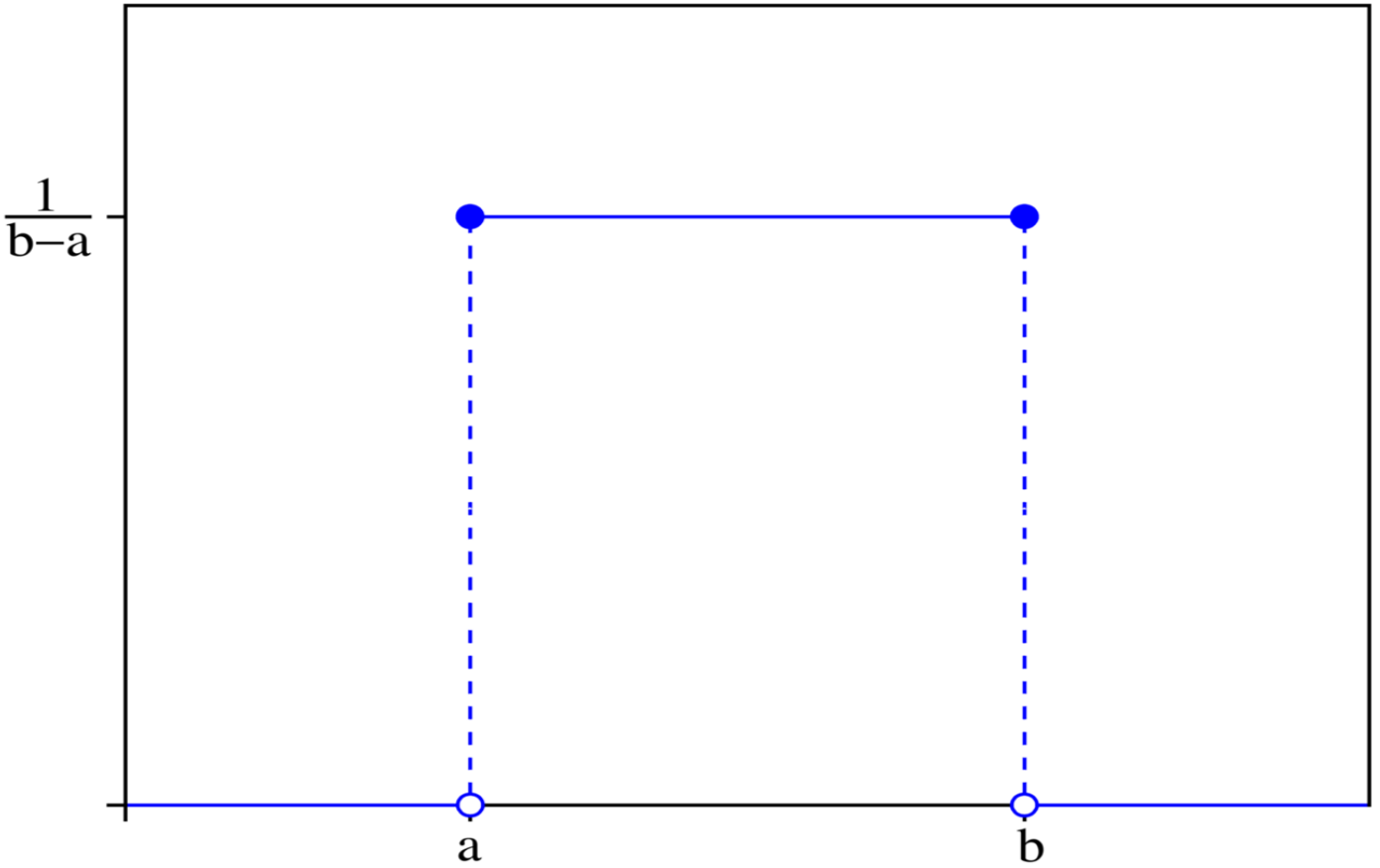
Stefan Kasberger
@stefankasberger

Warum analysieren?

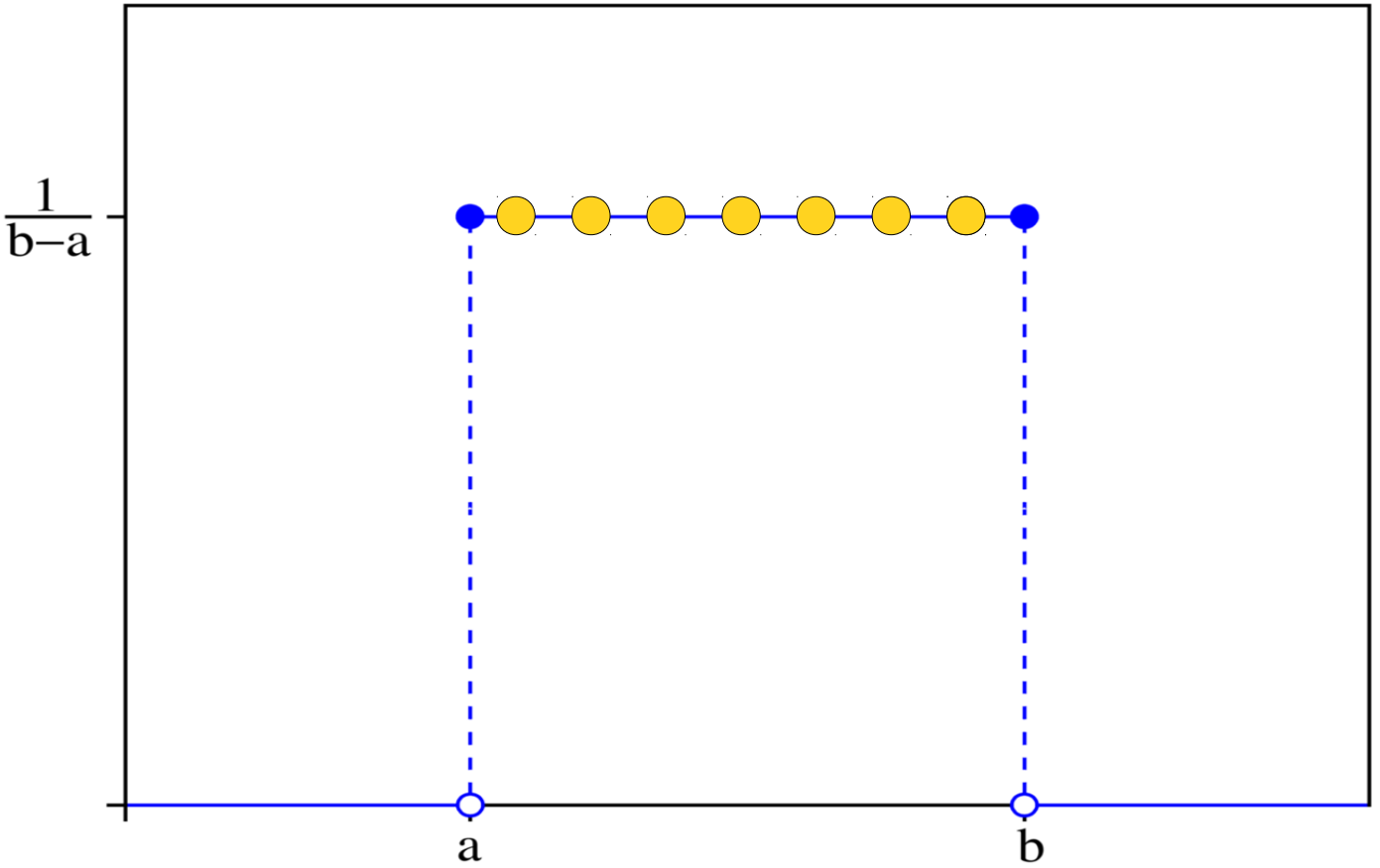
- Fragestellung beantworten
- Zusammenhänge finden
- Datensets verstehen
- Muster finden

VERTEILUNGEN

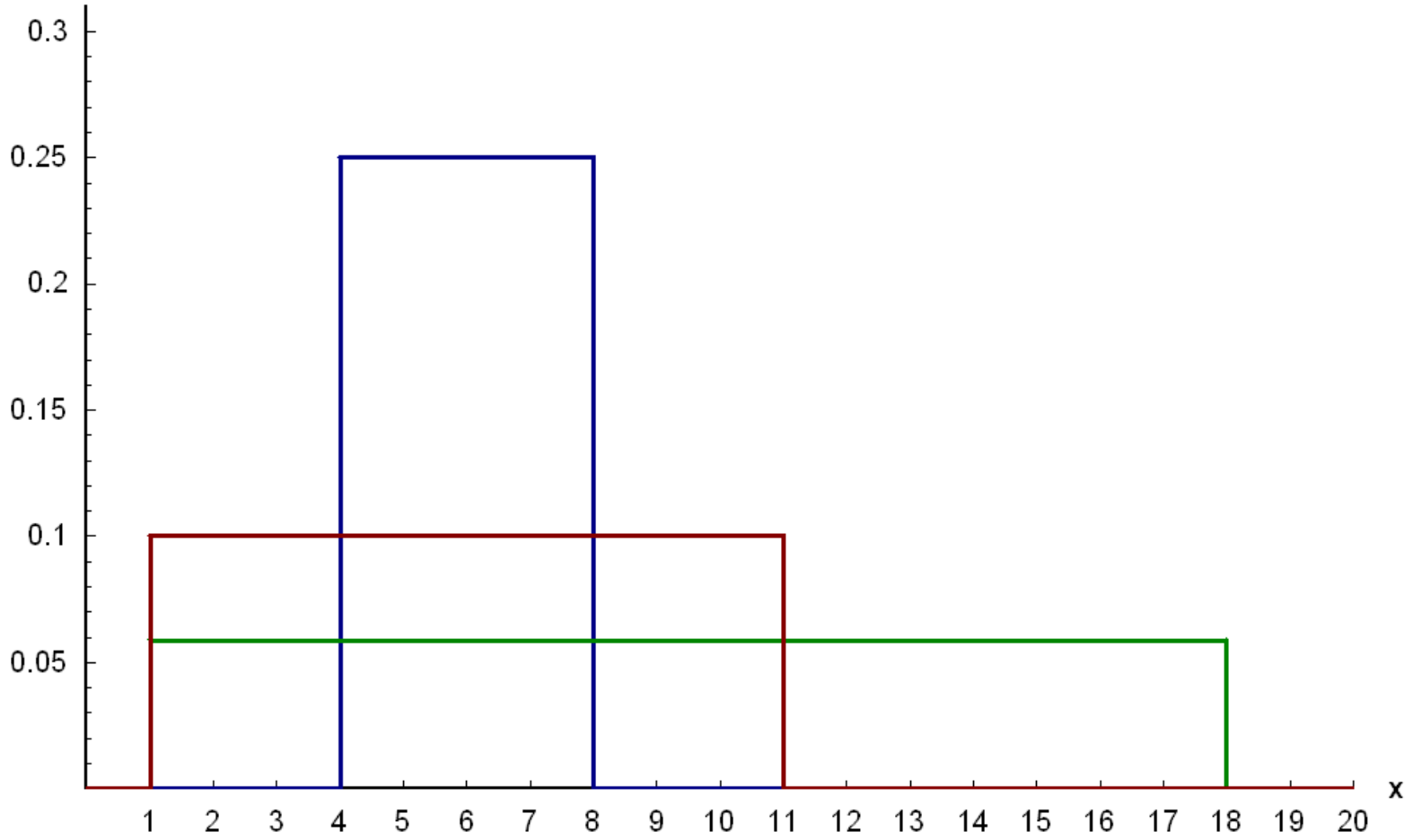
Uniformverteilung



Uniformverteilung



Uniformverteilung

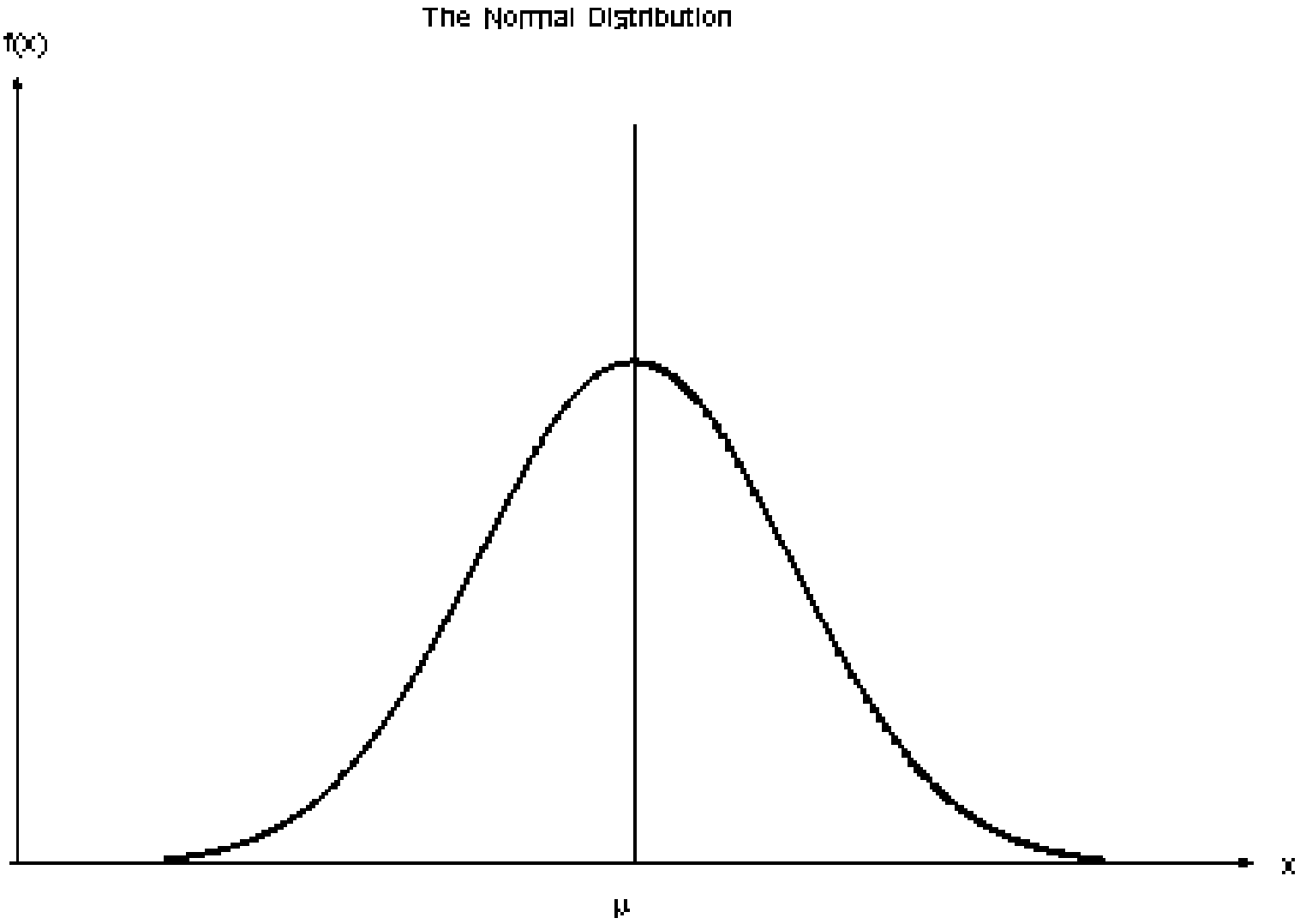


Uniformverteilung

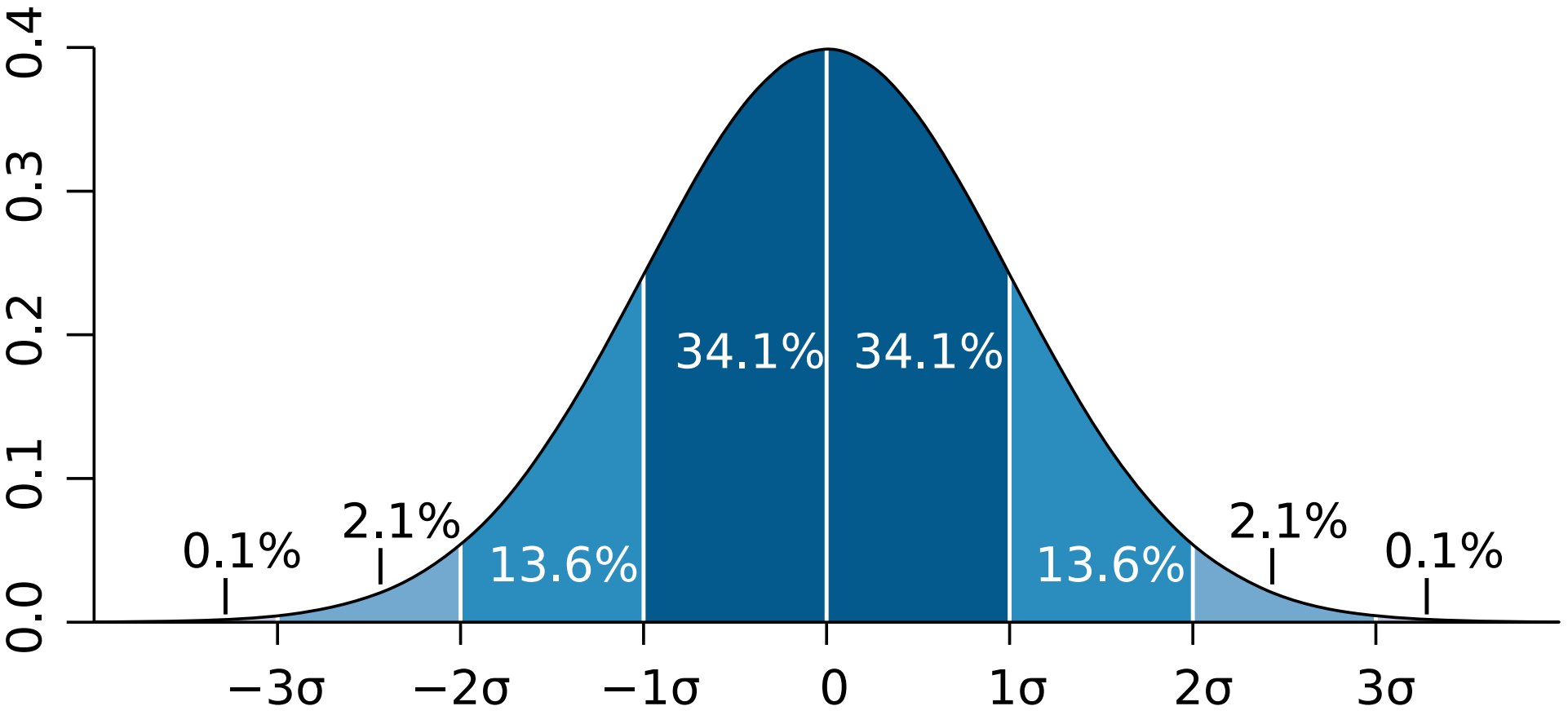
Beispiele

- Würfel
- Münzwurf

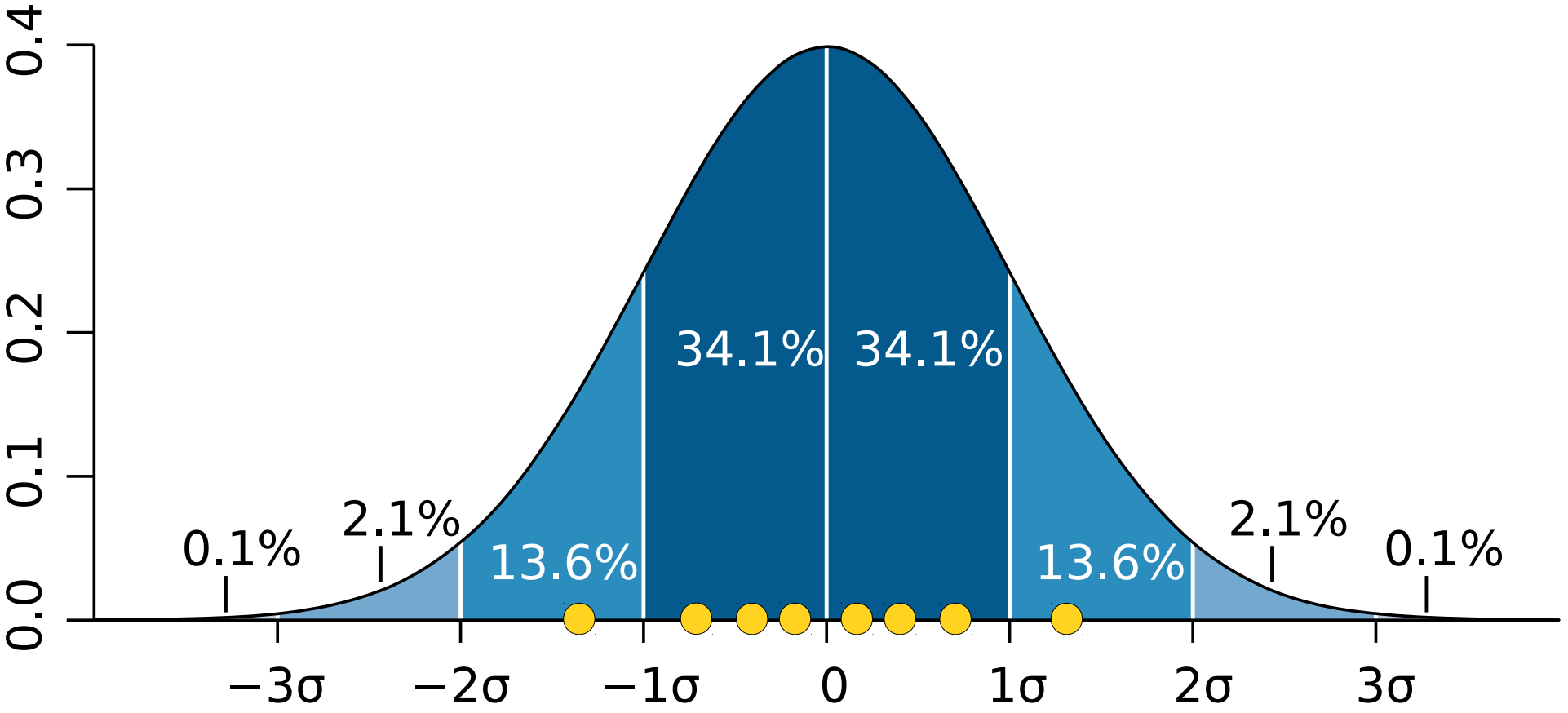
Normalverteilung (Gauß)



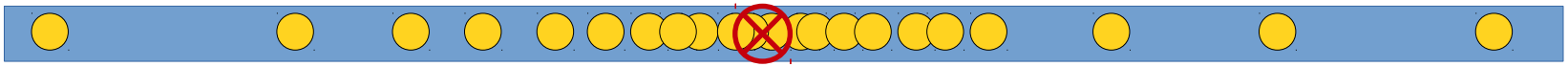
Normalverteilung (Gauß)



Normalverteilung (Gauß)



Normalverteilung

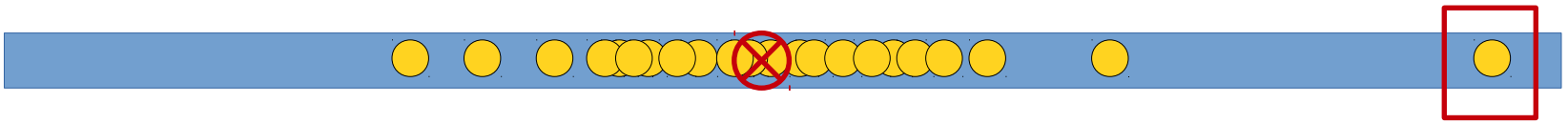


Beispiele

- Preise eines Produktes am Markt
- Dart Wurf
- Körpergröße
- Gewicht
- IQ
- Testpunkte

Ausreisser (Outlier)

Als Ausreißer wird ein Messwert bezeichnet, der nicht in eine erwartete Messreihe passt oder allgemein nicht den Erwartungen entspricht.



METRIKEN

Min, Max und Range

- Min: kleinster Zahlenwert
- Max: größter Zahlenwert
- Range: Zahlen-Spannweite (Max - Min)

x	y
1	6
2	4
3	5
4	3
5	2
6	4

Min, Max und Range

- Min: kleinster Zahlenwert
- Max: größter Zahlenwert
- Range: Zahlen-Spannweite (Max - Min)

<u>Ergebnis:</u>
Min: 2
Max: 6
Range: 4

x	y
1	6
2	4
3	5
4	3
5	2
6	4

Mean

Summiere alle Werte und dividiere durch die Anzahl der Werte.

- Mean = Durchschnitt = Arithmetisches Mittel
- Aggregierte Größe: Ausreisser (Outlier) können verzerren.
- Bei Normalverteilung wahrscheinlich nahe des Erwartungswertes (Umso mehr Beobachtungen, umso wahrscheinlicher).

Mean

Summiere alle Werte und dividiere durch die Anzahl der Werte.

x	y
1	6
2	4
3	5
4	3
5	2
6	4

Mean

Summiere alle Werte und dividiere durch die Anzahl der Werte.

Mean:
= (6 + 4 + 5 + 3 + 2 + 4) / 6
= 24 / 6
= 4

x	y
1	6
2	4
3	5
4	3
5	2
6	4

Median

1. Sortiere die Werte in Reihenfolge.
2. Wähle den Wert des Datenpunktes in der Mitte. Bei einer geraden Anzahl an Datenpunkten, nimm den Mittelwert der beiden mittleren Datenpunkte.

→ Ist robuster gegenüber Ausreissern.

Median

Finde den Wert des in der Mitte liegenden Punktes.

x	y	y (sortiert)
1	6	2
2	4	3
3	5	4
4	3	5
5	2	6
6	4	6

Median

Finde den Wert des in der Mitte liegenden Punktes.

Median:
= (4 + 5) / 2
= 4.5

x	y	y (sortiert)
1	6	2
2	4	3
3	5	4
4	3	5
5	2	6
6	4	6

Modus

Modus ist der Wert, der am häufigsten im Datenset auftritt (nicht wie oft er auftritt).

- bei Gleichstand, sind alle gleich oft auftretenden Werte der Modus.
- geht auch für Nominal.
- bei Normalverteilung wahrscheinlich in der Nähe des Erwartungswertes.

Modus

Modus ist der Wert, der am häufigsten im Datenset auftritt.

x	y
1	6
2	4
3	5
4	3
5	2
6	4

Modus

Modus ist der Wert, der am häufigsten im Datenset auftritt.

Modus: 4

x	y
1	6
2	4
3	5
4	3
5	2
6	4

Zusammenhang Mean & Median

Wenn Mean höher ist als Median, deutet das auf einen oder mehrere relativ große Werte im Datensatz hin (Outlier?). Die Verteilung ist somit nicht normalverteilt, sondern nach rechts verzerrt (skewed right).

Übung

Diskutiert in den Gruppen in euren eigenen Worten was folgende Begriffe bedeuten:

- Mean
- Median
- Mode

(5min)

Übung

- Min?
- Max?
- Range?
- Mean?
- Median?
- Modus?

(10min)

Monat	Gitarren
Januar	5
Februar	8
März	10
April	8
Mai	7
Juni	9
Juli	3
August	11
September	10
Oktober	10
November	12
Dezember	15

In Libre Office

=MIN()
=MAX()
=AVERAGE()
=MEDIAN()
=MODE()

→ Beispiele Einkommens-Verteilung in Österreich.

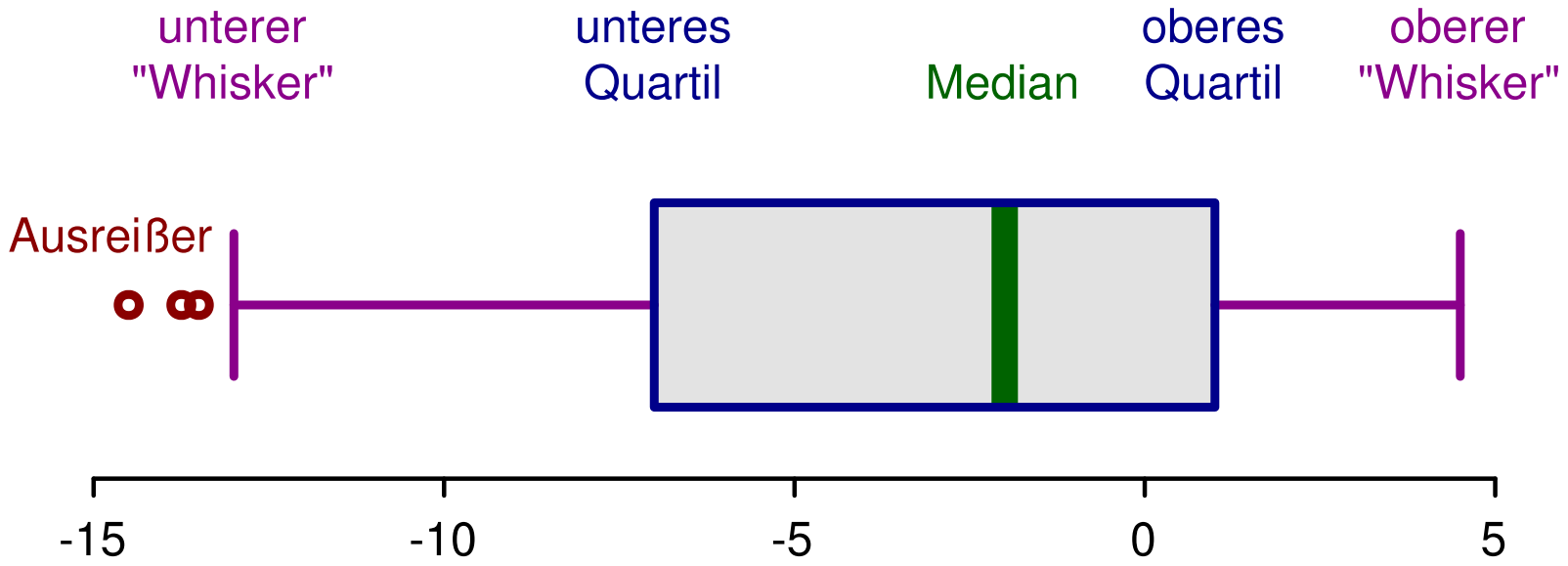
BOXPLOT

Boxplot

Ein Box-Plot soll schnell einen Eindruck darüber vermitteln, in welchem Bereich die Daten liegen und wie sie sich über diesen Bereich verteilen.

– Wikipedia

Boxplot



- Q1: 25% der Werte sind <= diesem Wert
- Median=Q2: 50% der Werte sind <= diesem Wert
- Q3: 75% der Werte sind <= diesem Wert
- Box: mittleren 50% der Daten (Interquartilsabstand → IQR)
- Antennen: links → Q1 – IQR * 1.5; rechts → Q3 + IQR * 1.5 (→ letzter Datenpunkt innerhalb des Intervalls)

Tipp: Vorgangsweise

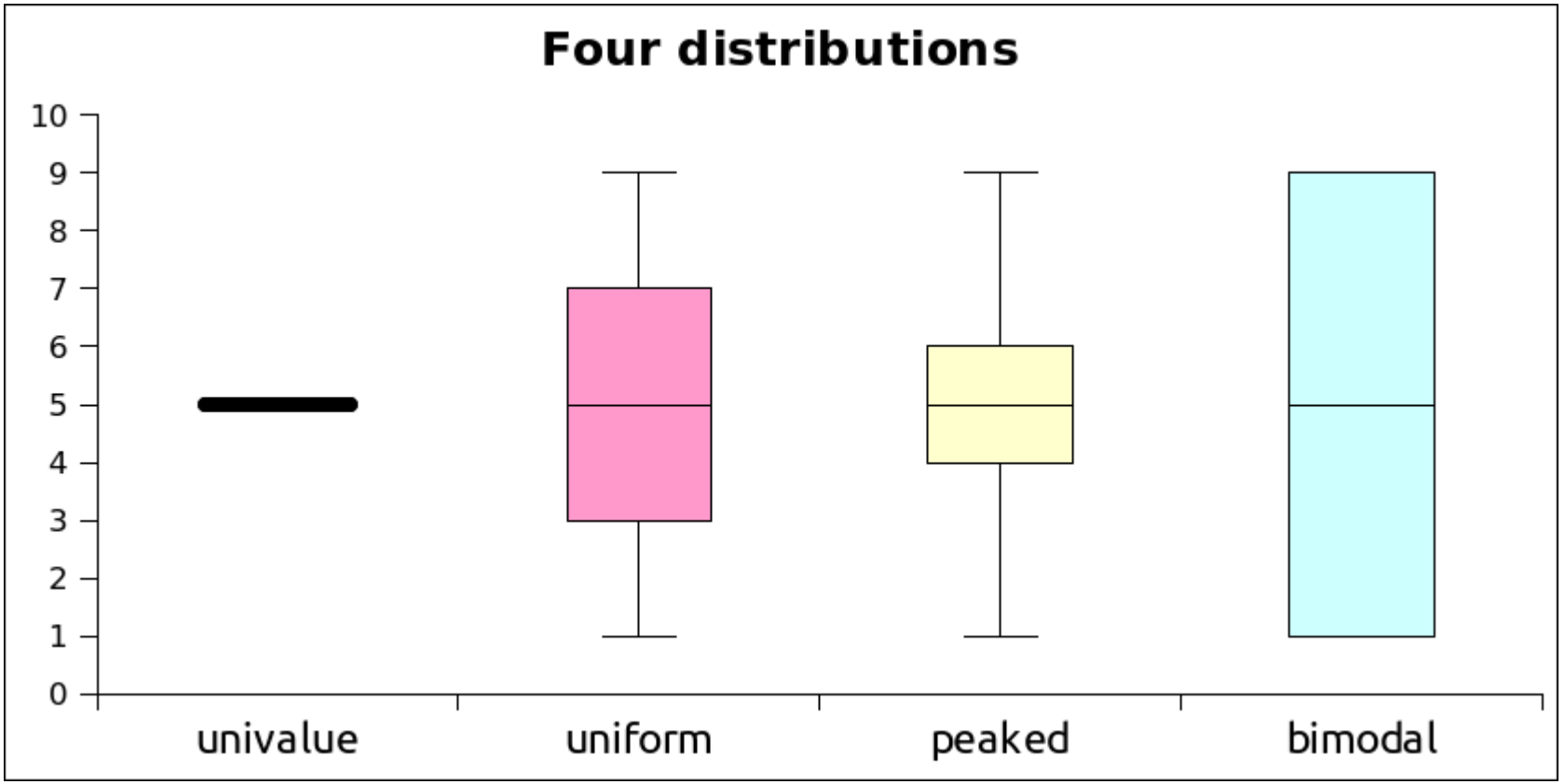
1) Range ansehen
2) Box ansehen
3) Lage Median zu Box ansehen
4) Lage Box zu Range ansehen
5) Lage Antennen zu Range und Box ansehen

Interpretation

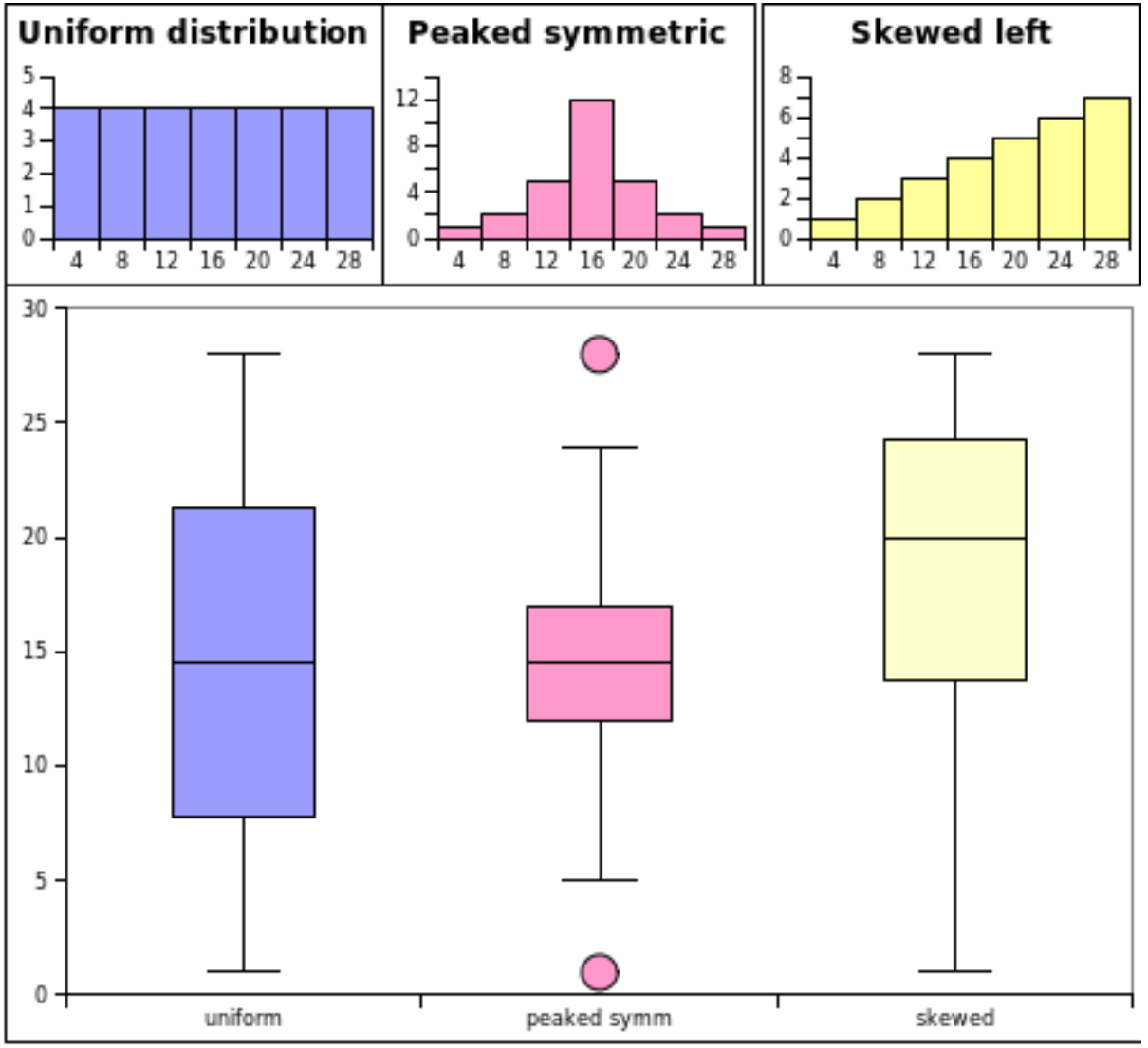
Median ist verschoben von Box-Mittelpunkt
Bsp.: Wenn Median links vom Box-Mittelpunkt, sind die Daten rechts verschoben (vermutlich Ausreisser rechts). Umso weiter links, umso mehr deutet es auf starke Ausreisser hin. Gilt auch für entgegengesetzte Richtung.

Lage Box zu Range: Wenn Abstand gleich, sind Daten gleichverteilt. Wenn Box = schmal, sind Daten stark zentriert. Wenn Box = Range, sind Daten bimodal.

Verteilungen

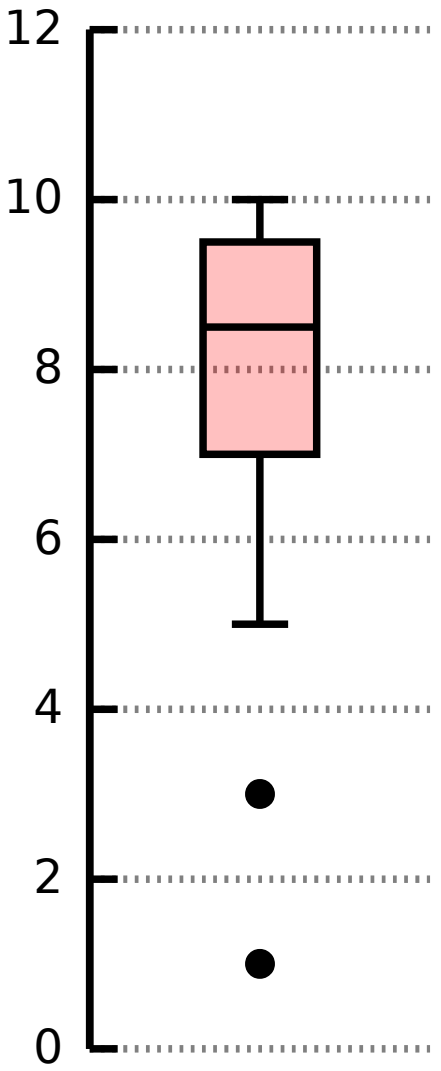


Verteilungen



Beispiel

Min = 1
Max = 10
Range = 9
Mean = 7.75
Median = 8.5
Modus = 9 und 10
Q1 = 7
Q3 = 9.5
IQR = 2.5

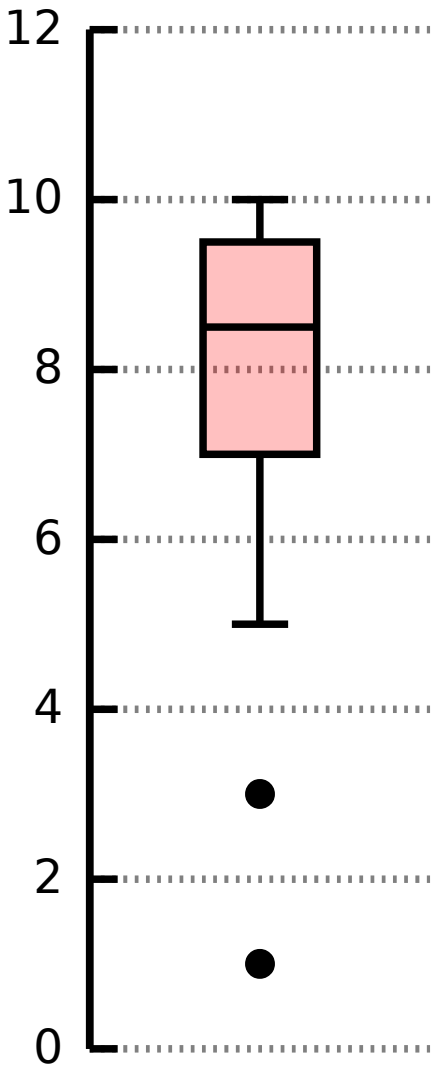


y	sort
9	1
6	3
7	5
7	6
3	7
9	7
10	7
1	8
8	8
7	8
9	9
9	9
8	9
10	9
5	9
10	10
10	10
9	10
10	10
8	10

Beispiel

Interpretation:

Verteilung wegen Ausreissern nach Links verschoben und somit nicht symmetrisch bzw. normalverteilt.



y	sort
9	1
6	3
7	5
7	6
3	7
9	7
10	7
1	8
8	8
7	8
9	9
9	9
8	9
10	9
5	9
10	10
10	10
9	10
10	10
8	10

Beispiele

- Uniform
- Normal
- Skewed links und rechts
- Ausreisser rechts
- Bimodal

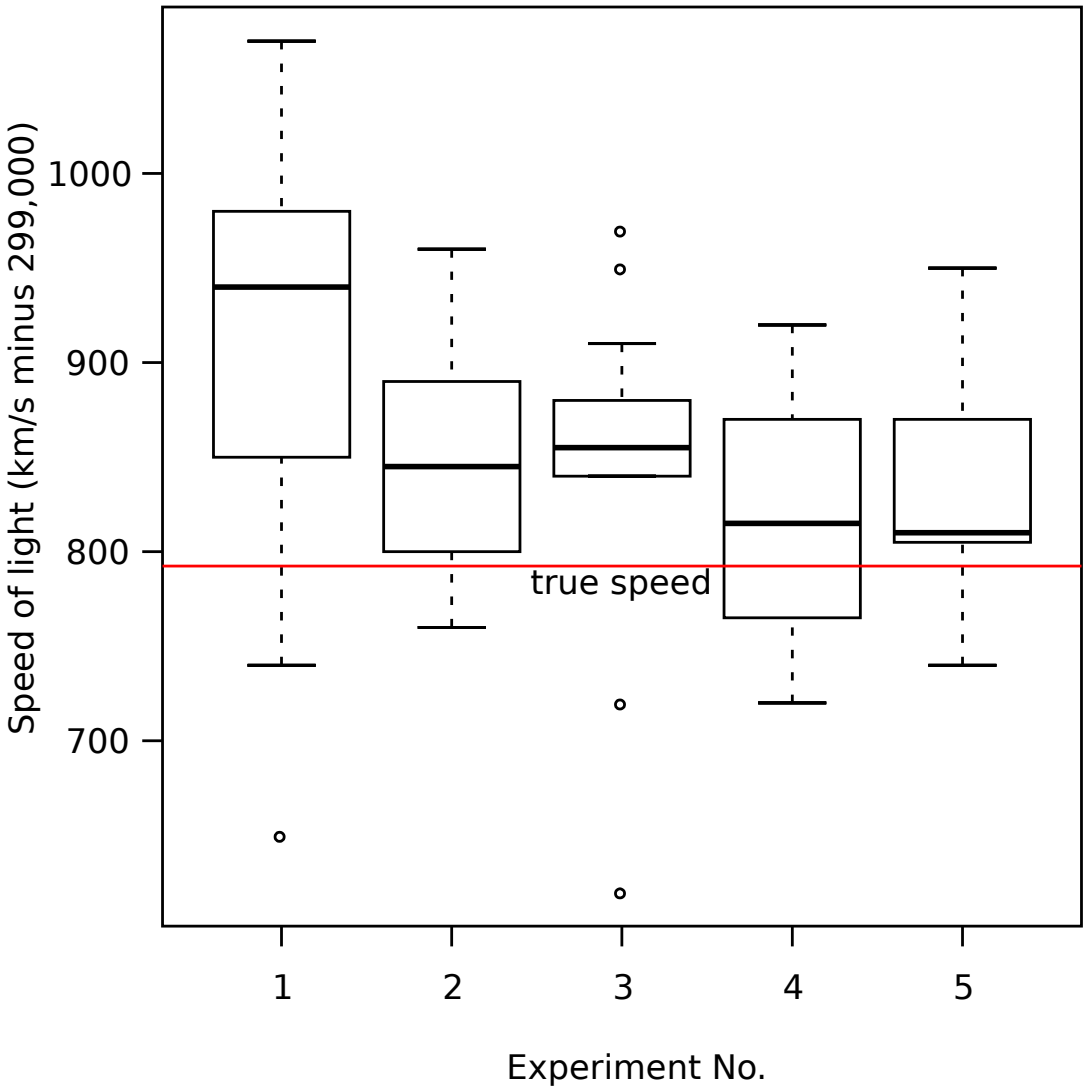
Übung

2er Gruppen, 15min Zeit

- Sortiern
- Berechne: Min, Max, Mean, Median, Modus, Q1, Q3, IQR
- Erstelle Boxplot: bit.ly/boxplotr
- Interpretiere Daten

7
12
5
8
10
8
9
10
11
9
8
1

Übung


Speed of light (km/s minus 299,000)
1000
900
800
700
true speed
1
2
3
4
5
Experiment No.

KORRELATION

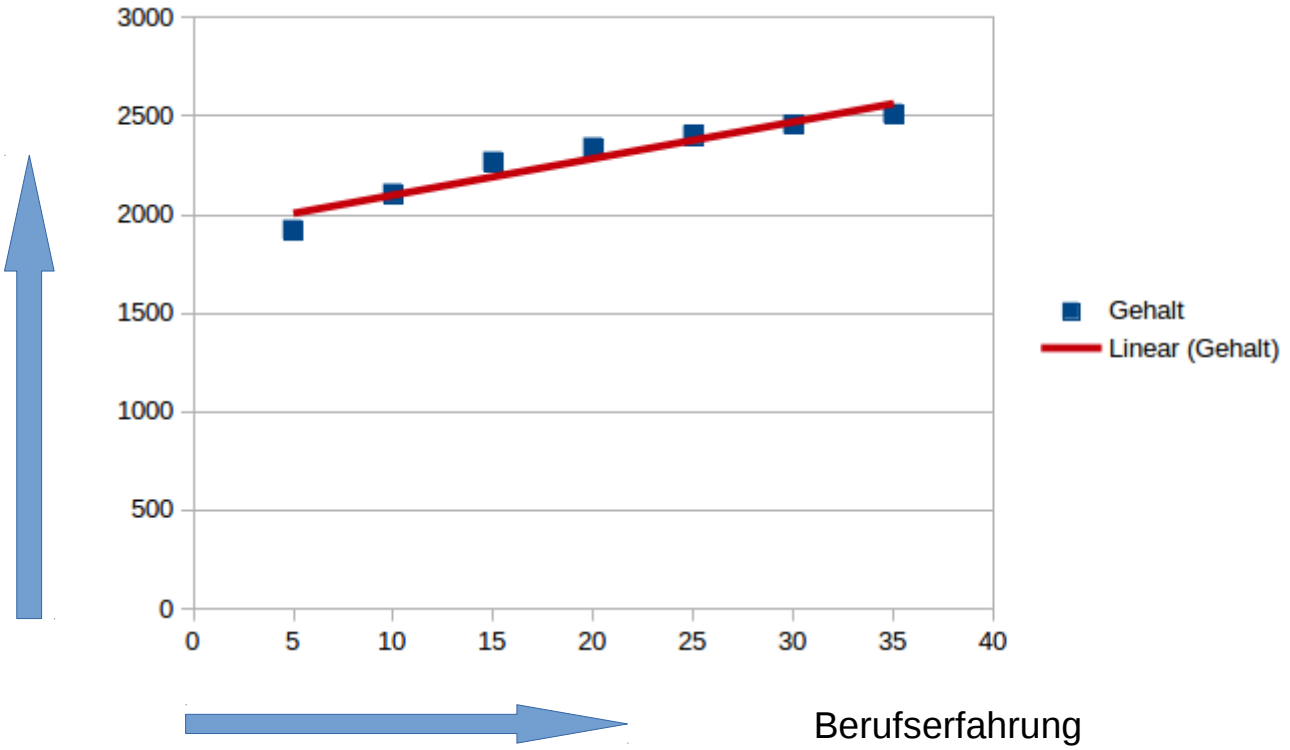
Definition

Eine Korrelation beschreibt eine Beziehung zwischen zwei oder mehreren Merkmalen, Ereignissen, Zuständen oder Funktionen. Die Beziehung muss keine kausale Beziehung sein.

– Wikipedia

Beispiel

Gehalt [€]



Berufserfahrung [Jahre]

Berufs-erfahrung	Gehalt
5	1921
10	2105
15	2267
20	2339
25	2401
30	2456
35	2510

r = 0.96

Korrelation und Kausalität

Korrelation ist nicht gleich kausaler Zusammenhang!

- Korrelation zeigt einen statistischen Zusammenhang, keinen kausalen (Ursache → Wirkung).
- Zusammenhang kann zufällig auftreten → statistische Tests nötig
- Zusammenhang kann indirekt durch einen anderen Zusammenhang vorhanden sein.
- Korrelation zwischen A und B = Korrelation zwischen B und A

→ **immer hinterfragen, ob Zusammenhang plausibel erscheint!**

Korrelations-Koeffizient

- Ko-Relation: Ko=Zusammen, Relation=Verbindung
- Pearson Korellations-Koeffizient
- Korrelationen ist nur bei linearen Beziehungen möglich
- gibt Richtung und Stärke des Zusammenhangs an
- Wert von -1 bis +1:
  - 0 ist kein linearer Zusammenhang
  - +1 ist perfekter positiver linearer Zusammenhang
  - -1 ist perfekter negativer linearer Zusammenhang
- Umso mehr Datenpunkte, umso statistisch gesicherter ist der Zusammenhang

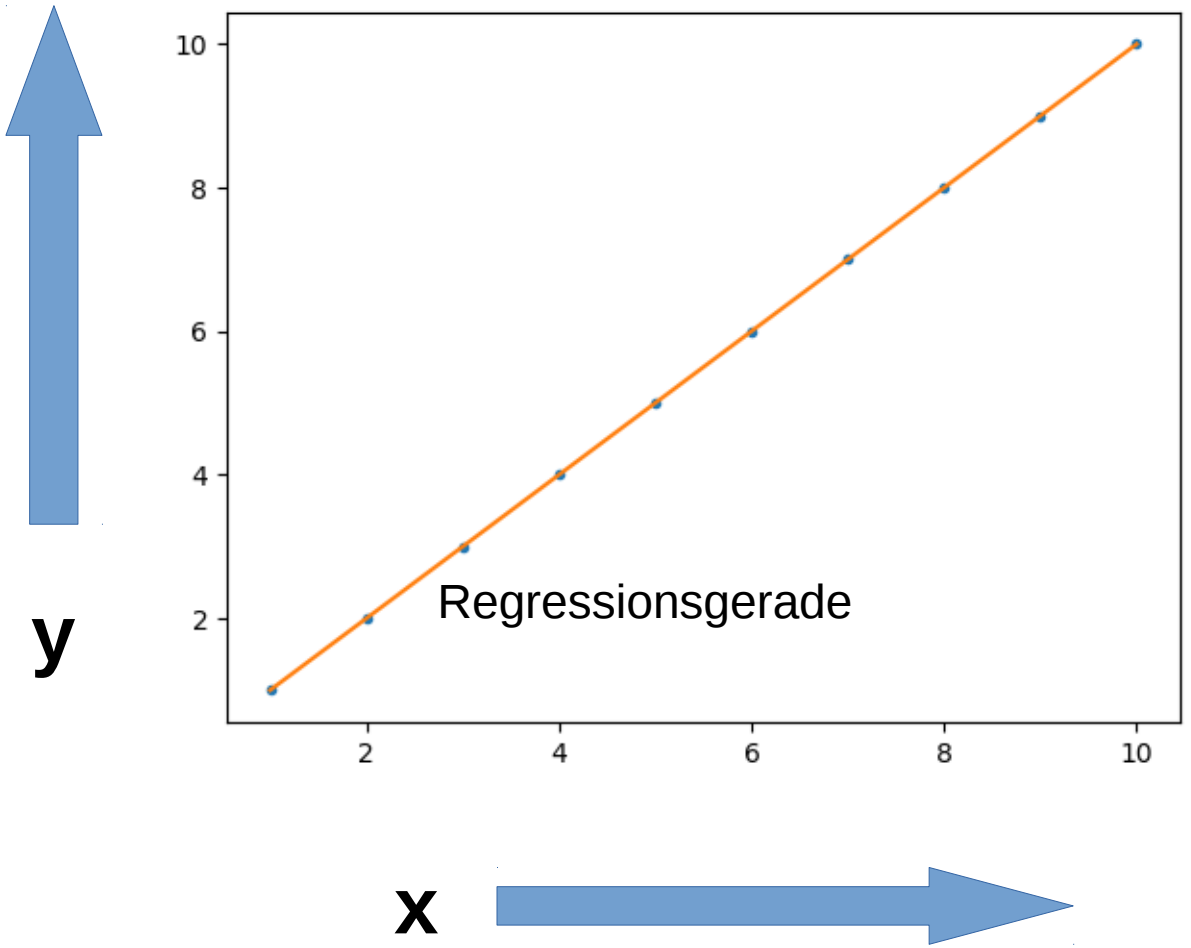
Korrelations-Koeffizient

Stärke

- umso größer, umso stärker ist der lineare Zusammenhang
- Stärke trifft keine Aussage über die statistische Signifikanz

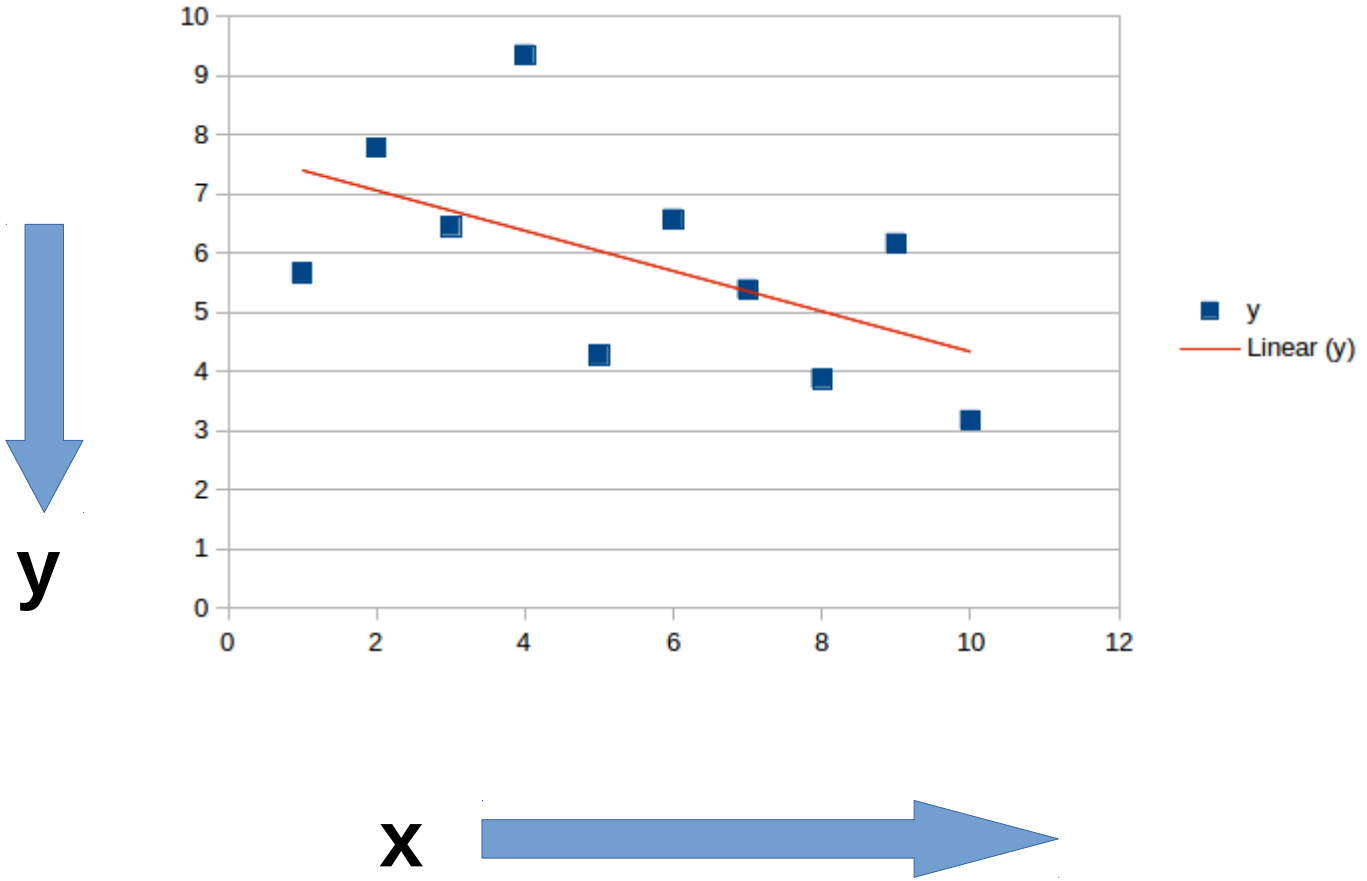
0:	kein Zusammenhang
0-0.4:	schwach
0.4-0.6:	mittel
0.6-0.8:	stark
0.8-1:	sehr stark

Scatter Plot (r = +1)



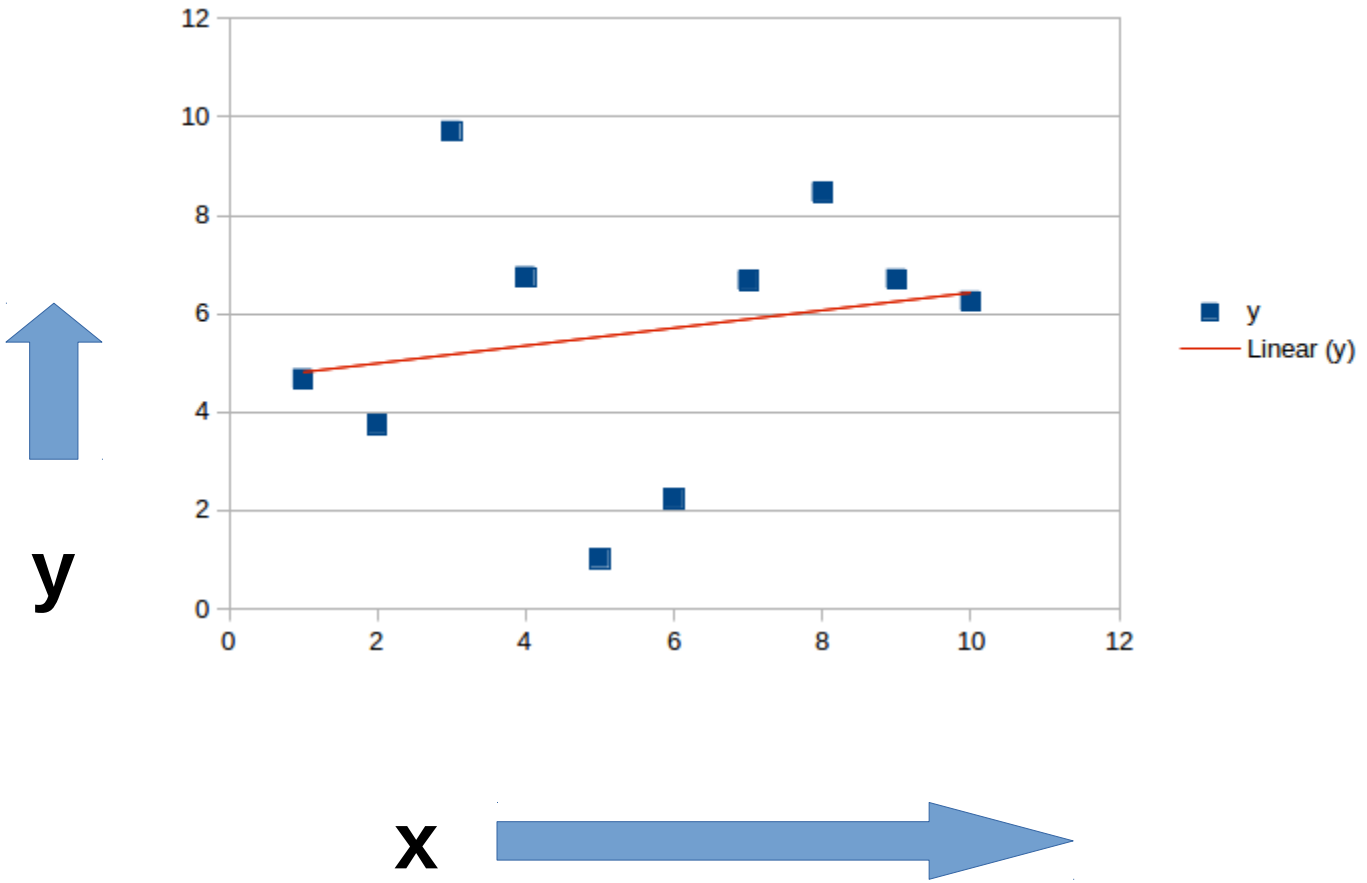
x	y
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10

Scatter Plot (r = -0.55)



x	y
1	5.66
2	7.78
3	6.45
4	9.35
5	4.27
6	6.57
7	5.38
8	3.87
9	6.16
10	3.17

Scatter Plot (r = +0.2)



x	y
1	5.66
2	7.78
3	6.45
4	9.35
5	4.27
6	6.57
7	5.38
8	3.87
9	6.16
10	3.17

Übung

Temperatur und Eisverkauf

1. Suche den Befehl für die Pearson Korrelation für deine Spreadsheet Software
2. Trage die Daten ein und berechne den Pearson Korrellations-Koeffizienten
3. Erstelle den Scatterplot + Regressionsgerade
4. Ist der Zusammenhang positiv oder negativ?
5. Wie stark ist der Zusammenhang (quantitativ und qualitativ)?
6. Interpretiere den Zusammenhang in eigenen Worten

(10min)

T	€
14.2	215
16.4	235
11.9	185
15.2	332
18.5	406
22.1	522
19.4	412
25.1	614
23.4	544
18.1	421
22.6	445
17.2	408

Übung

1. Überleg dir ein Beispiel von zwei Merkmalen, die einen stark negativen Zusammenhang haben.
2. Fabrizieren 12 Datenpunkte, die realistisch sind.
3. Trage die Daten in die Spreadsheet-Software. Passe die Daten so an, dass der gewünschte Zusammenhang heraus kommt.
4. Berechne den Pearson Korrelations-Koeffizienten und erstelle das Diagramm.

(10min)

Kontakt

www.offenewahlen.at
@stefankasberger
stefan.kasberger@okfn.at
www.ofkn.at

UrheberInnenrecht:

Markenrecht: