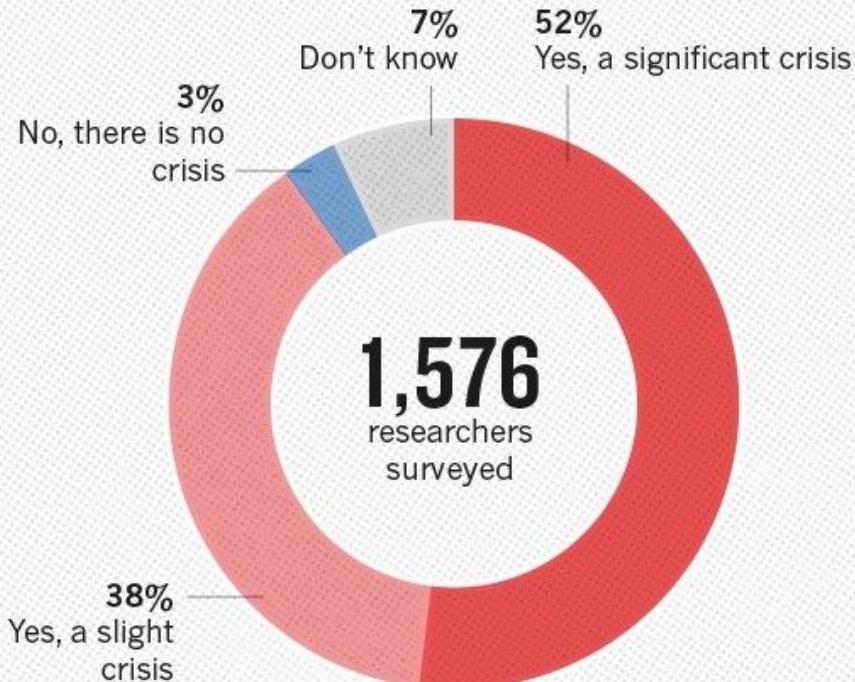


Data Descriptions & Summaries

Robin Roy, Florian Stein

IS THERE A REPRODUCIBILITY CRISIS?



©nature

Gliederung

1. Variablen
 - 1.1. Qualitative Variablen
 - 1.2. Quantitative Variablen
2. Lagemaß & Streuung
 - 2.1. Definition des Lagemaß
 - 2.2. Variabilitätsmaße
3. Korrelationen
 - 3.1. Pearson Korrelation
 - 3.2. Spearmen Korrelation
4. Fallstricke
 - 4.1. Ausreißer
 - 4.2. Schiefe
 - 4.3. Simpson-Paradoxon

Variablen

1. Variablen

- Grundlegende Daten einer Statistik
- Variablen: mess-/beobachtbare Merkmale von Merkmalsträgern
- Merkmalsträger: Statistische Einheiten mit relevanten Eigenschaften
- Merkmalsausprägungen: unterschiedliche Werte dieser Eigenschaften

1.1 Qualitative Variablen

Variable ist eine Kategorie

- Dichotom - “Ja/Nein-Fall”
 - Berufsausbildung - keine Berufsausbildung
 - Schwanger - nicht Schwanger
 - Krank - nicht krank
- Nicht-dichotome Variablen
- Alle qualitativen Variablen sind gleichzeitig diskret

1.1 Qualitative Variablen

- Nominale Variablen
 - Keine natürliche Reihenfolge
 - Religion, Staatsangehörigkeit
- Ordinale Variablen
 - Mit natürlicher Reihenfolge
 - Grad des Bildungsabschlusses, Schmerzniveau

1.2 Quantitative Variablen

Variable ist numerisch

- Diskrete Variablen
 - Zählbar, innerhalb eines begrenzten Intervalls
 - Schulnoten, Alter
- Stetige Variablen
 - Messbar, Intervallfrei
 - Temperatur, Körpergröße
 - Unterkategorie kontinuierlicher Variablen

Lagemaß & Streuung

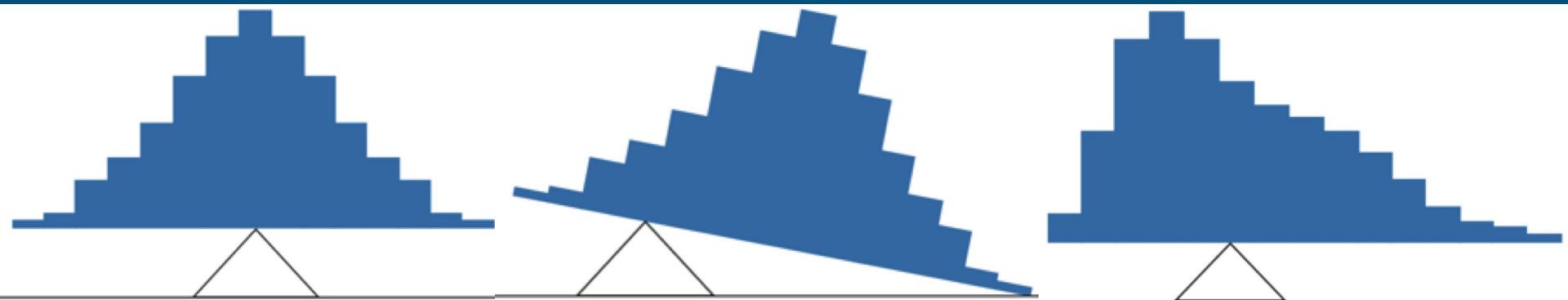
2.1 Definition des Lagemaß

- Vergleich einzelner Werte mit einer Verteilung an Werten
- Finden eines Mittelpunktes einer Verteilung
- Entfernung von Datenpunkten voneinander
- Feststellen der Variabilität eines Merkmals

2.1 Definition des Lagemaß

Definition 1: Waage

- Verteilung im Gleichgewicht



2.1 Definition des Lagemaß

Definition 2: Kleinste absolute Abweichungen

- Summe der absoluten Abweichungen
(Differenzen)

Werte	absolute Abweichung von 10	absolute Abweichung von 5
2	8	3
3	7	2
4	6	1
9	1	4
16	6	11
Summe	28	21

2.1 Definition des Lagemaß

Definition 3: Kleinste quadrierte Abweichungen

- Summe der absoluten quadrierten Abweichungen (Differenzen)

Werte	absolute Abweichung von 10	absolute Abweichung von 5
2	64	9
3	49	4
4	36	1
9	1	16
16	36	121
Summe	186	151

2.2 Mittelwert & Median

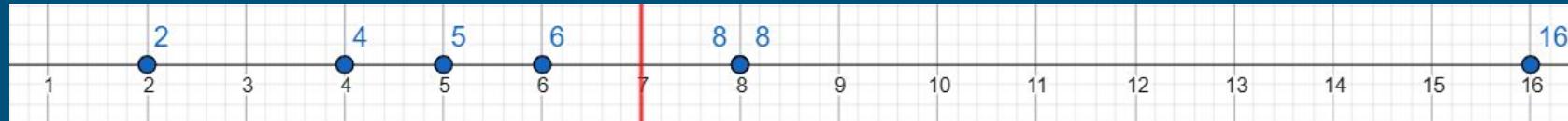
Arithmetischer Mittelwert:

- Durchschnitt einer Datenmenge

Verteilung: {2, 4, 5, 6, 8, 8, 16}

Mittelwert: $(2 + 4 + 5 + 6 + 8 + 8 + 16) / 7 = 7$

$$\mu = \frac{\sum X}{N}$$



2.2 Mittelwert & Median

Median:

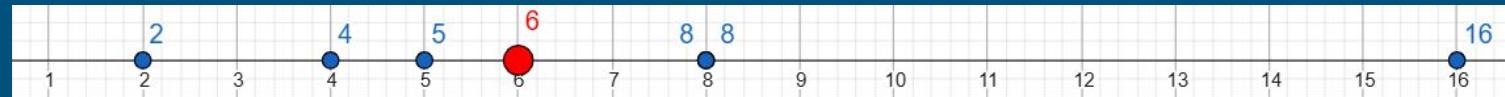
- Mitte einer geordneten Datenreihe
- Teilung in obere & untere Datenmenge

→ ungerader Anzahl an Zahlen: Median ist mittlere Zahl

→ gerader Anzahl an Zahlen: Median ist Mittelwert zwischen mittleren Zahlen

Verteilung: {2, 4, 5, **6**, 8, 8, 16}

Median: 6



2.3 Variabilitätsmaße: Interquartilsabstand

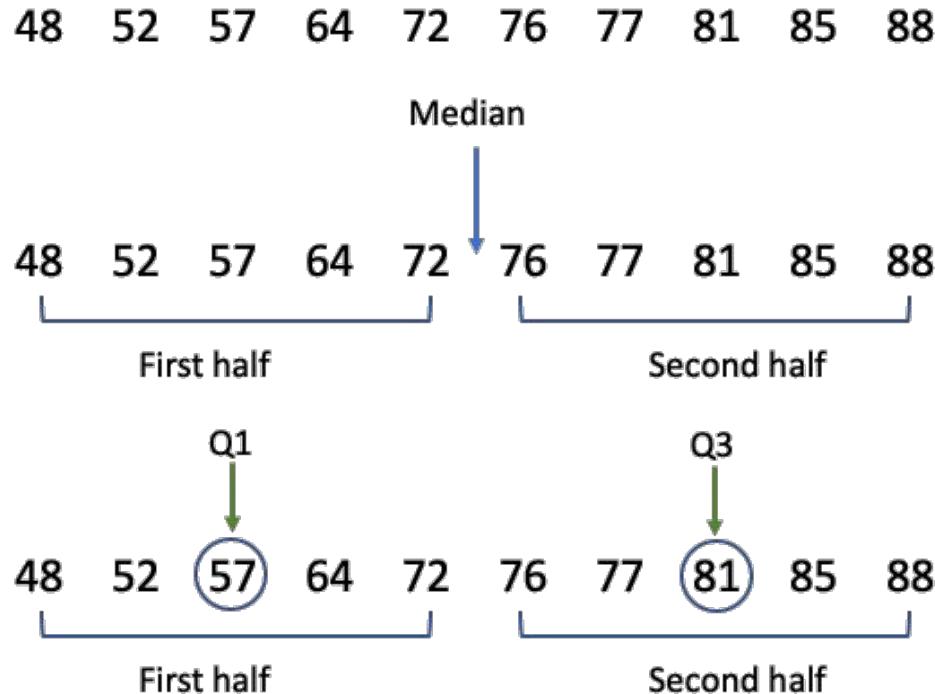
- Verteilung der mittleren Hälfte der Verteilung
- Quartile: Teilung der Verteilung in 4 gleiche Teile (klein nach groß)
- IQA: enthält 2. & 3. Quartil (Mitte der Verteilung)

Q3: 3. Quartil (75%)

Q1: 1. Quartil (25%)

$$\text{IQR} = Q_3 - Q_1$$

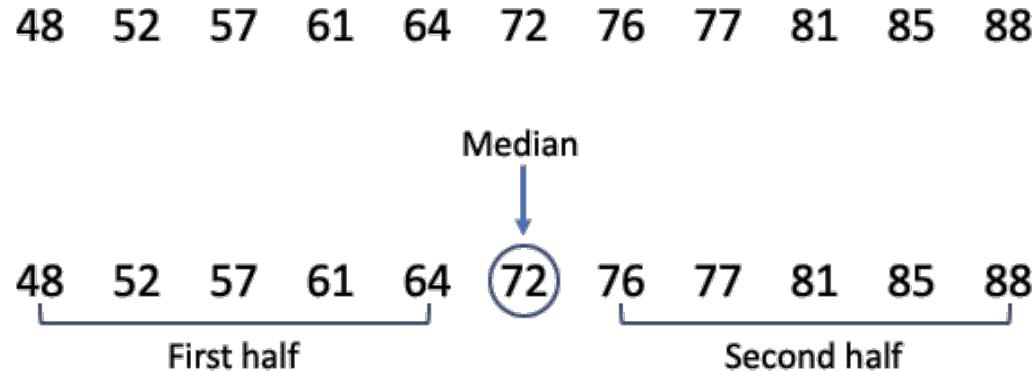
2.3 Variabilitätsmaße: Interquartilsabstand



$$\begin{aligned} \text{IQR} &= Q_3 - Q_1 \\ &= 81 - 57 \end{aligned}$$

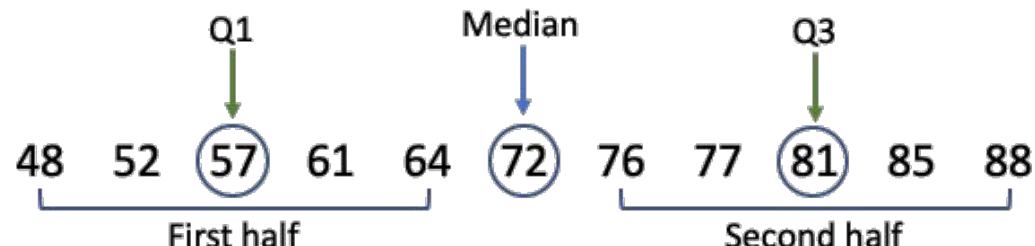
$$\text{IQR} = 24$$

2.3 Variabilitätsmaße: Interquartilsabstand



$$\begin{aligned} \text{IQR} &= Q_3 - Q_1 \\ &\equiv 81 - 57 \end{aligned}$$

$$\text{IQR} = 24$$



2.3 Variabilitätsmaße: Varianz

- mittlere quadratische Abweichung vom Erwartungswert
- Verteilung der Stichprobenwerte um arithmetischen Wert
- wichtigstes Streuungsmaß
- andere Dimension als Ausgangsdaten

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

2.3 Variabilitätsmaße: Varianz

Beispiel:

Verteilung: {2, 4, 5, 6, 8, 9, 14, 16}

Erwartungswert:

$$E(x) = (2 + 4 + 5 + 6 + 8 + 9 + 14 + 16) / 8 = 8$$

Varianz:

$$\begin{aligned} V(x) &= ((2 - 8)^2 + (4 - 8)^2 + (5 - 8)^2 + (6 - 8)^2 + (8 - 8)^2 + (9 - 8)^2 \\ &\quad + (14 - 8)^2 + (16 - 8)^2) / 8 \end{aligned}$$

$$V(x) = \underline{20,75}$$

2.3 Variabilitätsmaße: Standardabweichung

- mittlere einfache Abweichung vom Erwartungswert
- gleiche Dimension wie Ausgangsdaten

$$S(x) = \sqrt{V(x)}$$

2.3 Variabilitätsmaße: Standardabweichung

Beispiel:

Verteilung: {2, 4, 5, 6, 8, 9, 14, 16}

Erwartungswert:

$$E(x) = 8$$

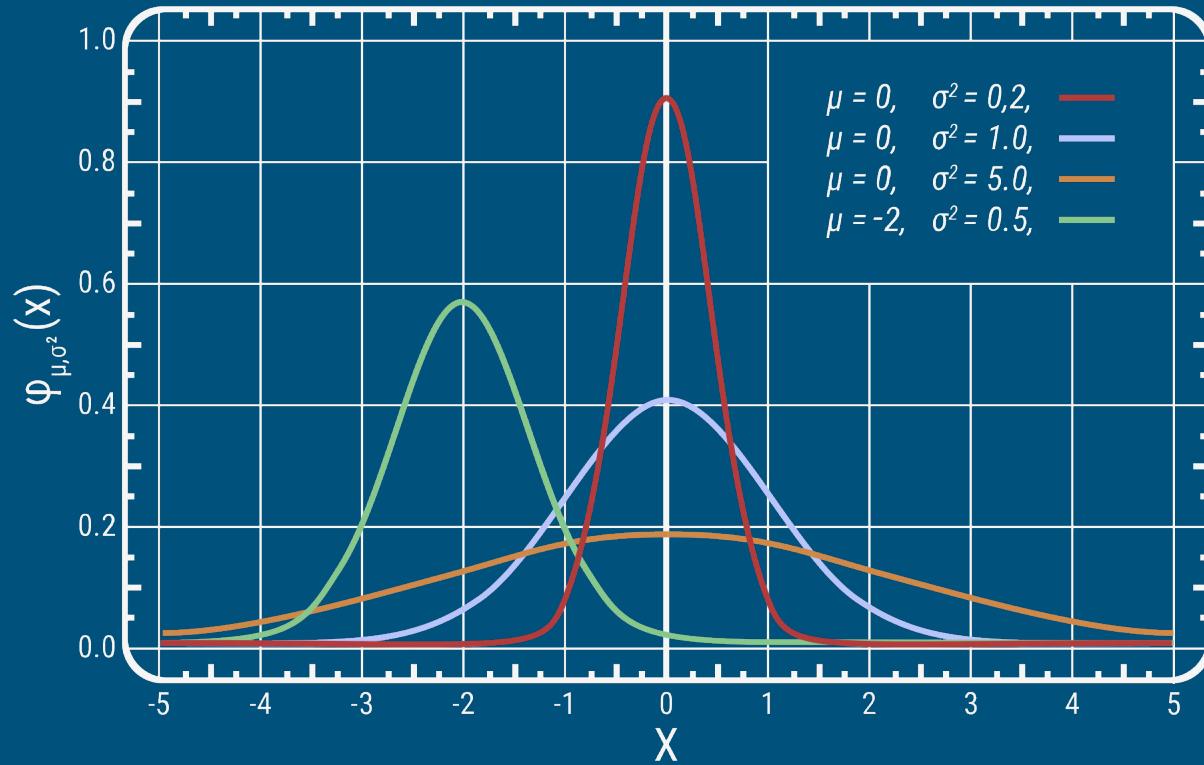
Varianz:

$$V(x) = 20,75$$

Standardabweichung:

$$S(x) = \sqrt{20,75} \approx \underline{4,555}$$

2.3 Variabilitätsmaße



2.3 Variabilitätsmaße: Beispiel 1



$$\text{Gelb} = 30^\circ \quad \text{Blau} = 150^\circ \quad \text{Rot} = 180^\circ$$

$$\text{Gelb: } 30^\circ / 360^\circ = 1/12$$

$$\text{Blau: } 150^\circ / 360^\circ = 5/12$$

$$\text{Rot: } 180^\circ / 360^\circ = 6/12 = 1/2$$

$x_i \in X: \text{Gewinn / Verlust}$	$p(x_i)$
-2€	1/2
1€	5/12
6€	1/12

2.3 Variabilitätsmaße: Beispiel 1

$$E(x) = \mu = -2\epsilon * 1/2 + 1\epsilon * 5/12 + 6\epsilon * 1/12$$

$$E(x) \approx \underline{-0,083\epsilon}$$

$$\begin{aligned}V(x) &= (-2 - (-0,083))^2 * 1/2 + (1 - (-0,083))^2 * 5/12 + (6 - (-0,083))^2 * 1/12 \\&= (-1,917)^2 * 1/2 + (1,083)^2 * 5/12 + (6,083)^2 * 1/12\end{aligned}$$

$$V(x) \approx \underline{5,41\epsilon^2}$$

$$S(x) = \sigma = \sqrt(V(x)) = \sqrt(5,41\epsilon^2) \approx \underline{2,33\epsilon}$$

2.3 Variabilitätsmaße: Beispiel 2

$$m_w = (135,41 \text{ cm} + 146,32 \text{ cm}) / 2 \approx 140,87 \text{ cm}$$

$$\mu_w \approx 140,97 \text{ cm}$$

$$m_m = (183,22 \text{ cm} + 184,25 \text{ cm}) / 2 \approx 183,74 \text{ cm}$$

$$\mu_m \approx 183,59 \text{ cm}$$

$$V_w(\text{Größe}) \approx 53,43 \text{ cm}^2$$

$$S_w(\text{Größe}) \approx 7,31 \text{ cm}$$

$$V_m(\text{Größe}) \approx 30,01 \text{ cm}^2$$

$$S_m(\text{Größe}) \approx 5,48 \text{ cm}$$

→ kein Vergleich zwischen Varianzen/SA möglich

Größe (cm)	Geschlecht
146,32	w
175,70	m
183,22	m
184,25	m
132,30	w
149,86	w
191,17	m
135,41	w

2.3 Variabilitätsmaße: Stichproben

- repräsentative Stichprobe: Rückschluss von Stichprobe auf Gesamtheit möglich
- probabilistische Stichprobe: zufällige Ziehung
- Eigenschaftsverteilung in Stichprobe = Eigenschaftsverteilung in der Grundgesamtheit
- möglichst große Stichprobe → höhere Repräsentativität ($n \geq 30$)

Korrelationen

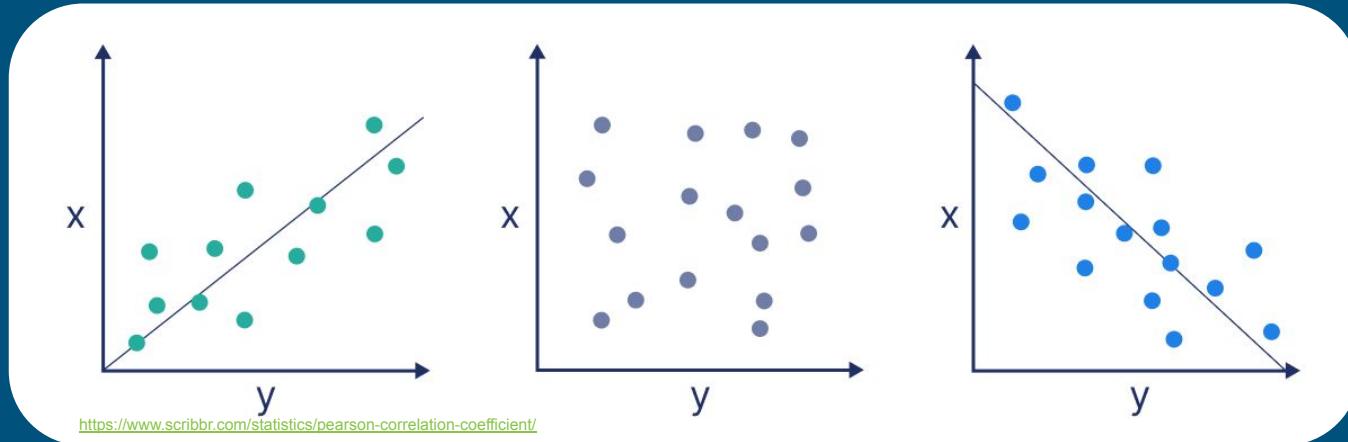
3. Korrelationen

- statistischer Zusammenhang zwischen mehreren Variablen
- Angabe der Stärke und Richtung
- lineare oder monotone Korrelation

3.1 Pearson Korrelation

Korrelationskoeffizient r : $[-1, 1]$

- $r \in]0,1]$ positive Korrelation (Änderung gleichermaßen)
- $r = 0$ keine Korrelation
- $r \in [-1,0[$ negative Korrelation (Änderung konträr)



3.1 Pearson Korrelation

- qualitative Variablen
- Normalverteilung der Daten
- keine Ausreißer
- lineare Zusammenhang

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

3.1 Pearson Korrelation

Gewicht = x

Größe = y

$$\begin{aligned}\Sigma x &= 3,63 + 3,02 + 3,82 + 3,42 + 3,59 + 2,87 \\ &\quad + 3,03 + 3,46 + 3,36 + 3,30\end{aligned}$$

$$\Sigma x = 33,5$$

$$\begin{aligned}\Sigma y &= 53,1 + 49,7 + 48,4 + 54,2 + 54,9 + 43,7 \\ &\quad + 47,2 + 45,2 + 54,4 + 50,4\end{aligned}$$

$$\Sigma y = 501,2$$

Gewicht (kg)	Größe (cm)
3,63	53,1
3,02	49,7
3,82	48,4
3,42	54,2
3,59	54,9
2,87	43,7
3,03	47,2
3,46	45,2
3,36	54,4
3,3	50,4

3.1 Pearson Korrelation

$$\Sigma x^2 = 13,18 + 9,12 + 14,59 + 11,70 + 12,89 \\ + 8,24 + 9,18 + 11,97 + 11,29 + 10,89$$

$$\Sigma x^2 = 113,05$$

$$\Sigma y^2 = 2\,819,6 + 2\,470,1 + 2\,342,6 \\ + 2\,937,6 + 3\,014,0 + 1\,909,7 + 2\,227,8 \\ + 2\,043,0 + 2\,959,4 + 2\,540,2$$

$$\Sigma y^2 = 25\,264$$

x	y	x^2	y^2
3,63	53,1	$(3,63)^2 = 13,18$	$(53,1)^2 = 2\,819,6$
3,02	49,7	9,12	2\,470,1
3,82	48,4	14,59	2\,342,6
3,42	54,2	11,7	2\,937,6
3,59	54,9	12,89	3\,014
2,87	43,7	8,24	1\,909,7
3,03	47,2	9,18	2\,227,8
3,46	45,2	11,97	2\,043
3,36	54,4	11,29	2\,959,4
3,3	50,4	10,89	2\,540,2

3.1 Pearson Korrelation

$$\begin{aligned}\Sigma xy &= 192,8 + 150,1 + 184,9 + 185,4 \\&\quad + 197,1 + 125,4 + 143,0 \\&\quad + 156,4 + 182,8 + 166,3\end{aligned}$$

$$\Sigma xy = 1\,684,2$$

x	y	x^2	y^2	$xy (x * y)$
3,63	53,1	13,18	2 819,6	3,63 * 53,1 = 192,8
3,02	49,7	9,12	2 470,1	150,1
3,82	48,4	14,59	2 342,6	184,9
3,42	54,2	11,7	2 937,6	185,4
3,59	54,9	12,89	3 014	197,1
2,87	43,7	8,24	1 909,7	125,4
3,03	47,2	9,18	2 227,8	143
3,46	45,2	11,97	2 043	156,4
3,36	54,4	11,29	2 959,4	182,8
3,3	50,4	10,89	2 540,2	166,3

3.1 Pearson Korrelation

$$n = 10$$

$$\Sigma x = 33,5$$

$$\Sigma y = 501,2$$

$$\Sigma x^2 = 113,05$$

$$\Sigma y^2 = 25\ 264$$

$$\Sigma xy = 1\ 684,2$$

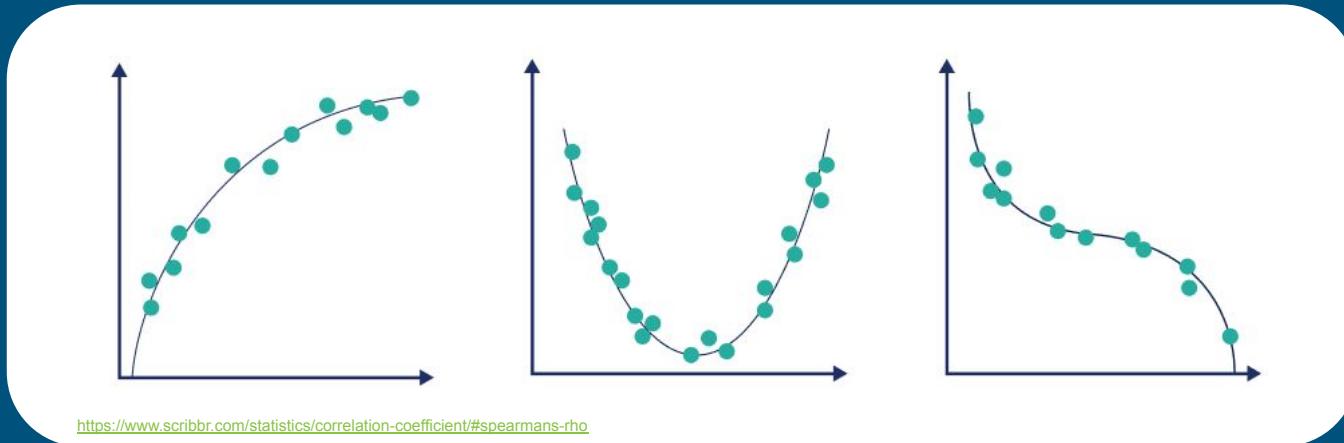
$$r = 0,47$$

$$r = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}}$$

3.2 Spearmen Korrelation

Spearmen-Koeffizient r_s : $[-1, 1]$

- $r_s \in]0,1]$ positive monotone Beziehung
- $r_s = 0$ keine monotone Beziehung
- $r_s \in [-1,0[$ negative monotone Beziehung



3.2 Spearmen Korrelation

- ordinale Daten
- keine Normalverteilung
- Daten enthalten Ausreißer
- Zusammenhang ist nicht linear und monoton

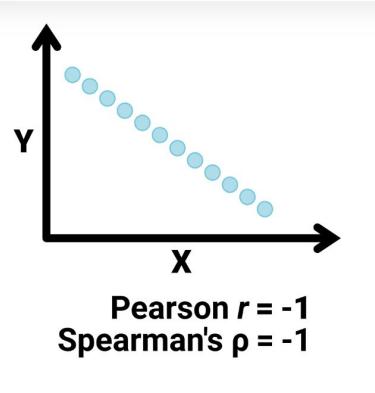
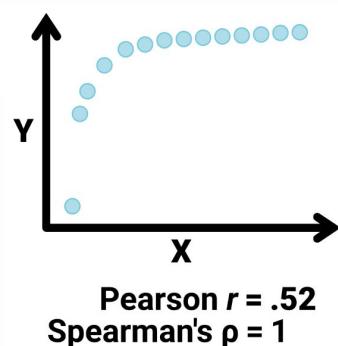
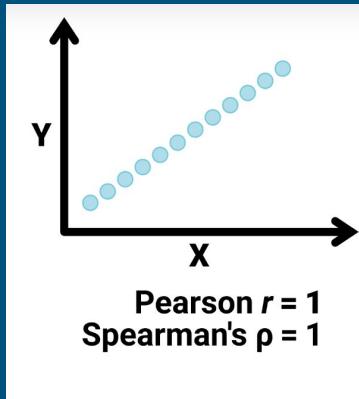
r_s = Stärke der Korrelation zwischen den Variablen

d = paarweise Differenz zwischen x und y

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

3. Korrelationen

- Pearson-Korrelation:
 - Lineare Beziehung kontinuierlicher Variablen
 - Änderung einer Variablen mit proportionaler Änderung einer anderen korreliert
- Spearman-Korrelation:
 - Nicht-konstante Beziehung zweier kontinuierlicher oder ordinaler Variablen
 - Änderung einer Variablen mit nicht-proportionaler Änderung einer anderen korreliert

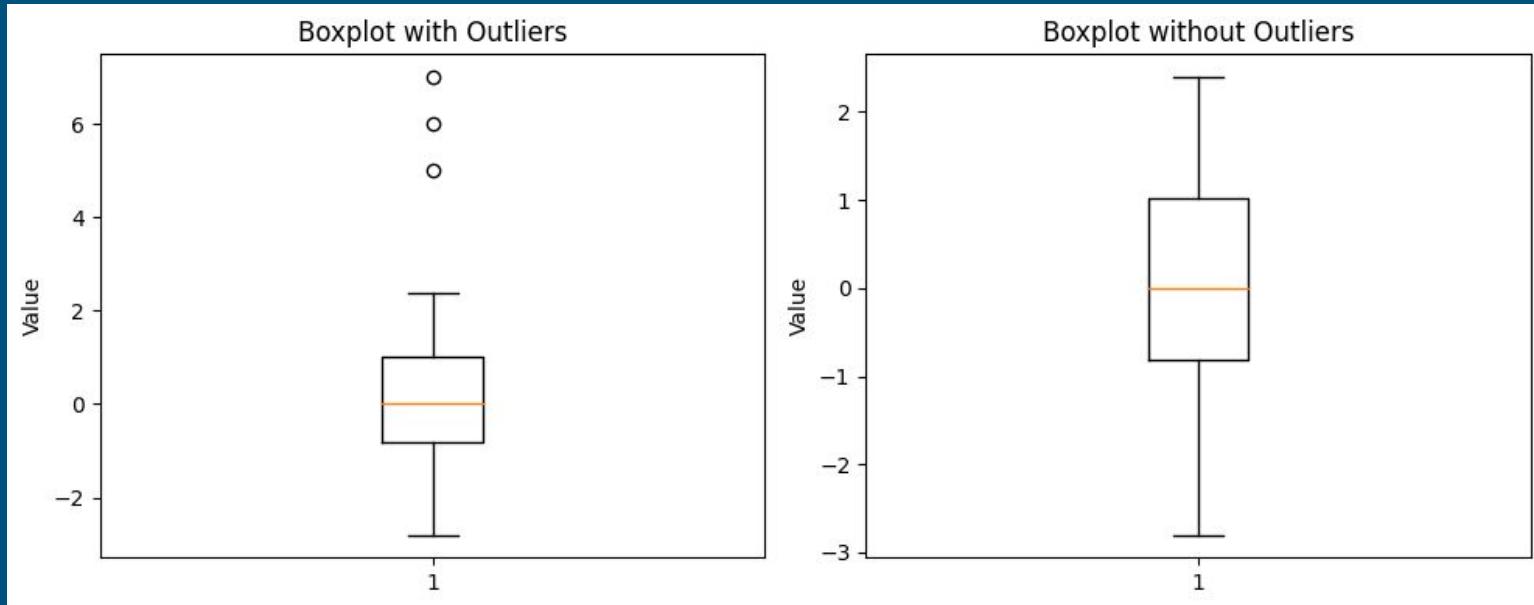


Fallstricke

4.1 Ausreißer

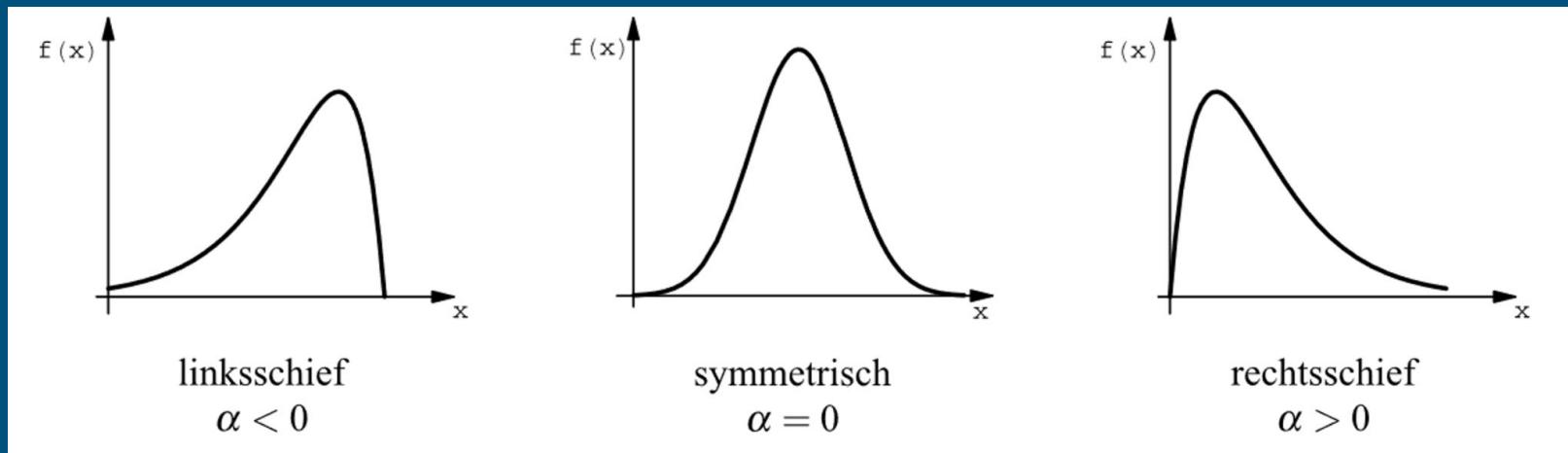
- Erheblich abweichende Datenpunkte
- Keine universelle Definition von Ausreißern möglich
- Entfernung nur mit Vorsicht und in begründeten Fällen
- Auswirkungen auf die statistische Verwertbarkeit sollten analysiert werden

4.1 Ausreißer



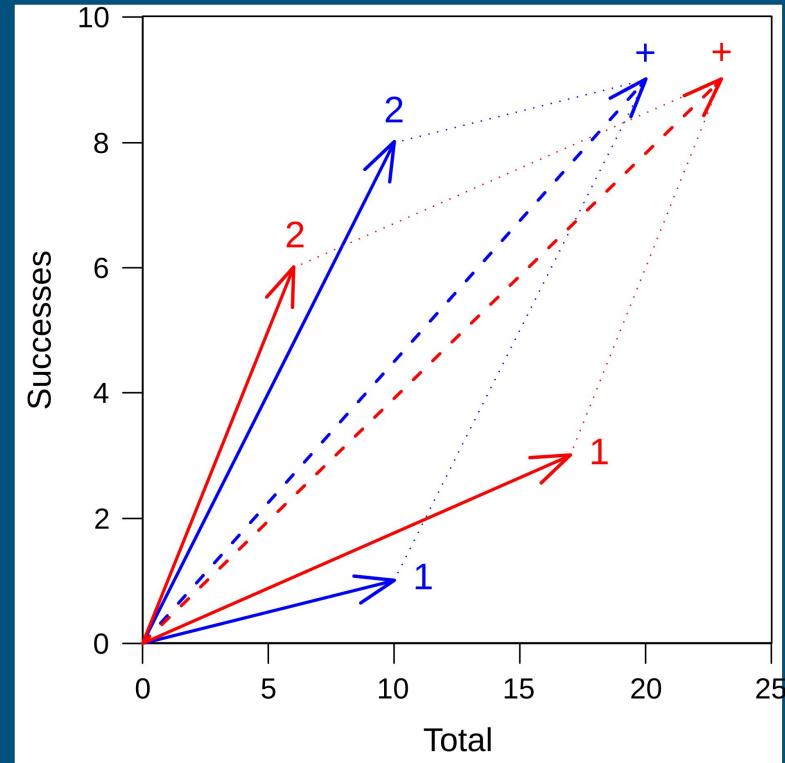
4.2 Schiefe

- Abweichung der Ergebnisverteilung vom Erwartungswert/Mittelwert
- links/negative - rechts/positive Schiefe



4.3 Simpson-Paradoxon

- Umkehrung eines Bewertungstrends bei Kombination verschiedener Gruppen
- Störvariablen als Ursache
 - Nicht-entdeckter Einflussfaktor auf Erfolgsquote
- Unterteilung einer Stichprobe führt zu umgekehrt anmutenden Trends, wenn das Kriterium der Unterteilung mit einer Untersuchungsvariable korreliert



4.3 Simpson-Paradoxon

UC Berkeley, Gender-Bias-Studie

	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
Total	12,763	41%	8,442	44%	4,321	35%

4.3 Simpson-Paradoxon

Department	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
A	933	64%	825	62%	108	82%
B	585	63%	560	63%	25	68%
C	918	35%	325	37%	593	34%
D	792	34%	417	33%	375	35%
E	584	25%	191	28%	393	24%
F	714	6%	373	6%	341	7%
Total	4526	39%	2691	45%	1835	30%

Quellen

1. Lane (Rice). Online Statistics Education – Descriptives: <https://onlinestatbook.com/lms/>
2. <https://statistikgrundlagen.de/ebook/chapter/chapter-1-2/>
3. <https://wissenschafts-thurm.de/grundlagen-der-statistik-wie-unterscheidet-man-zwischen-nominal-oder-kardinalskala/>
4. <https://fity.club/lists/suggestions/outlier-box-plot/>
5. <https://soilr.github.io/doku/datentransformation.html>
6. <https://de.wikipedia.org/wiki/Simpson-Paradoxon#/media/Datei:Simpsons-vector.svg>
7. https://service.destatis.de/eLearning/modul16/lm_pg_1856.html?up=1
8. <https://statistikguru.de/spss/spearman-korrelation/spearman-vs-pearson.html>
9. Bickel et. al.: Sex Bias in Graduate Admissions: Data from Berkeley. In: Science 187 (1975), Nr. 4175, S. 398–404
10. https://en.wikipedia.org/wiki/Simpson%27s_paradox
11. <https://simpleclub.com/lessons/mathematik-varianz-standardabweichung>
12. <https://www.scribbr.com/statistics/pearson-correlation-coefficient/>

Workshop

GitHub Repository

https://github.com/flo1304/data_descriptions_summaries

