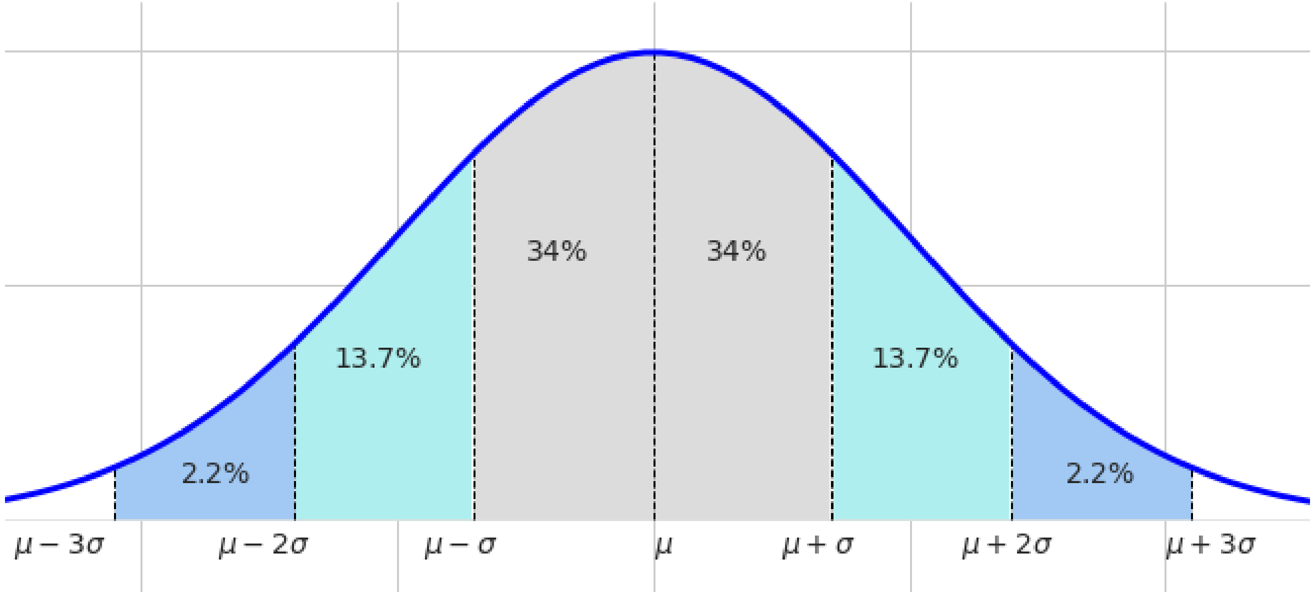


› STATISTIK

Statistik | Prof. Dr. Florian Kauffeldt

QUELLEN

Vorlesungsbegleitende Literatur:

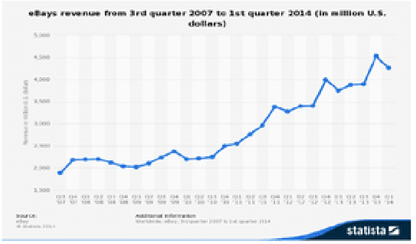
AUER, B. & ROTTMANN, H. (2015). Statistik und Ökonometrie für Wirtschaftswissenschaftler. Leipzig: Springer

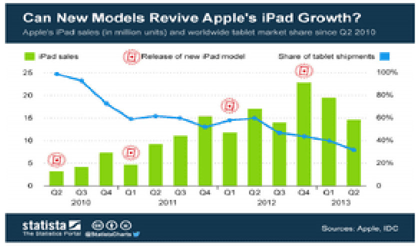
HÄRDLE, W.K., KLINKE, S., & RÖNZ, B. (2015). Introduction to Statistics. Heidelberg: Springer

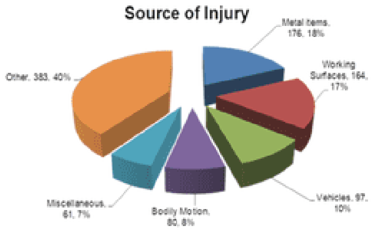
KOSFELD, R., ECKEY, H. & TÜRCK, M. (2019). Wahrscheinlichkeitsrechnung und induktive Statistik. Wiesbaden: Springer

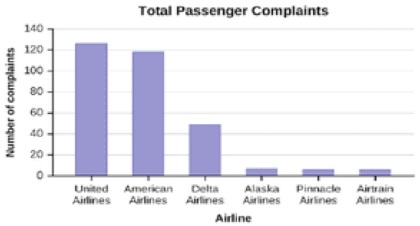
MATHAI, A.M. & HAUBOLD, H.J. (2018). Probability and Statistics. De Gruyter: Berlin

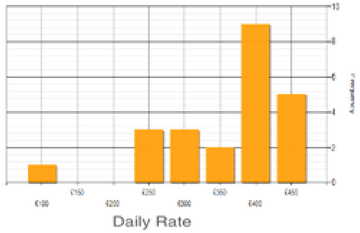
(siehe Syllabus und ILIAS-Link "Literatur")

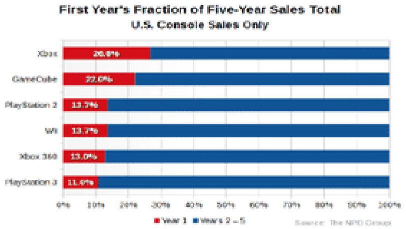












HOCHSCHULE HEILBRONN

› DESKRIPTIVE STATISTIK

Statistik | Prof. Dr. Florian Kauffeldt

TEILGEBIETE DER STATISTIK

HOCHSCHULE HEILBRONN

Statistik

Deskriptive Statistik	Inferenzstatistik
> Beschreibung von Daten	> Induktive Rückschlüsse basierend auf Daten

WAS IST DESKRIPTIVE STATISTIK?

HOCHSCHULE HEILBRONN

- Ziele Deskriptive Statistik: Beschreibung von Daten (keine induktiven Schlüsse)

- Deskriptive Statistik: Methoden, um Daten zusammenzufassen und zu visualisieren

Beispiel. Wir haben einen Datensatz mit dem Jahreseinkommen von 12'403 Menschen.

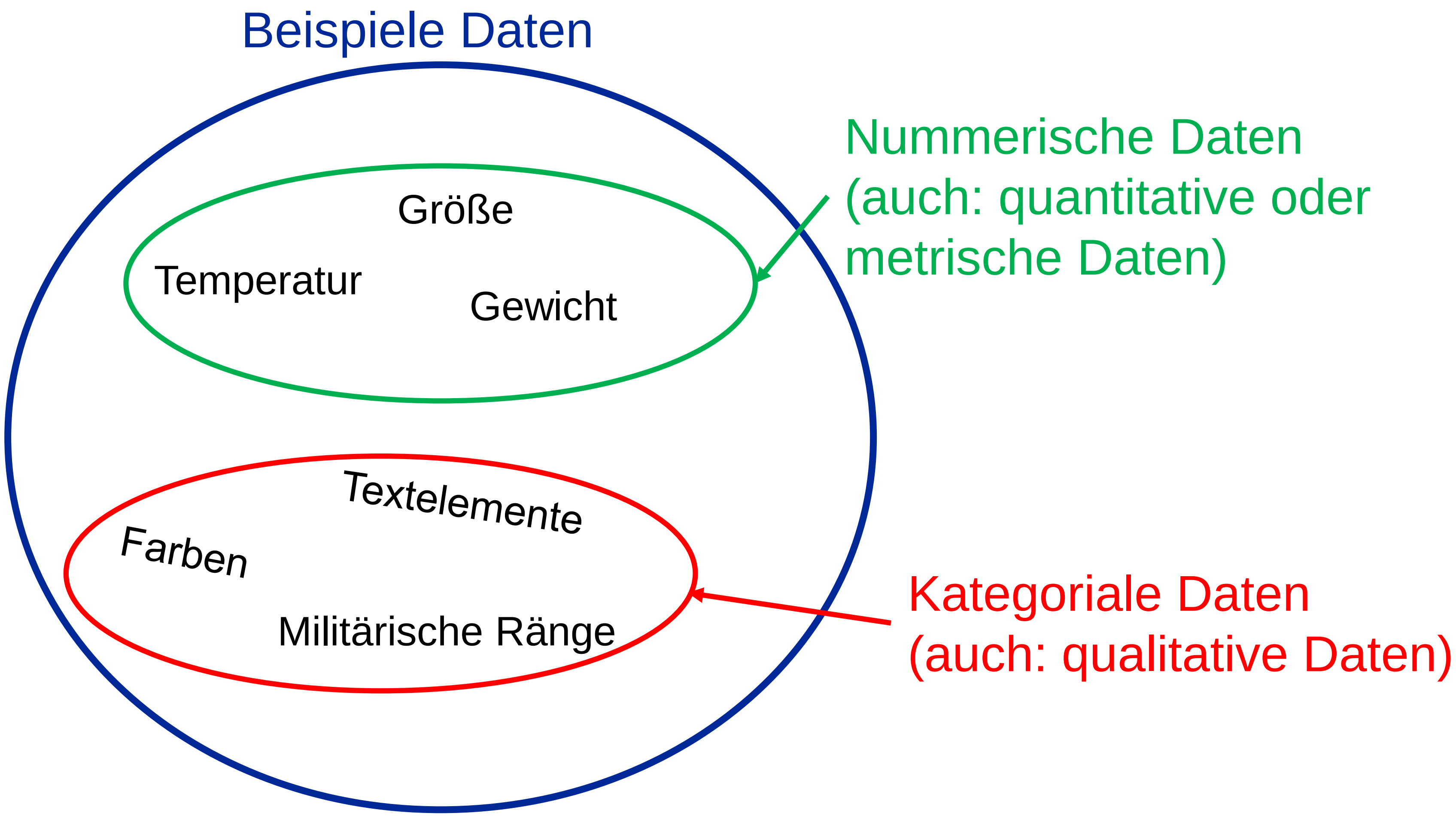
- Niemand möchte 12'403 Einzeleinträge lesen

- Stattdessen z.B. mittlere Jahreseinkommen anschauen

› DATEN

WAS SIND DATEN?

Daten sind Merkmale von Objekten, die durch Beobachtung erhoben werden



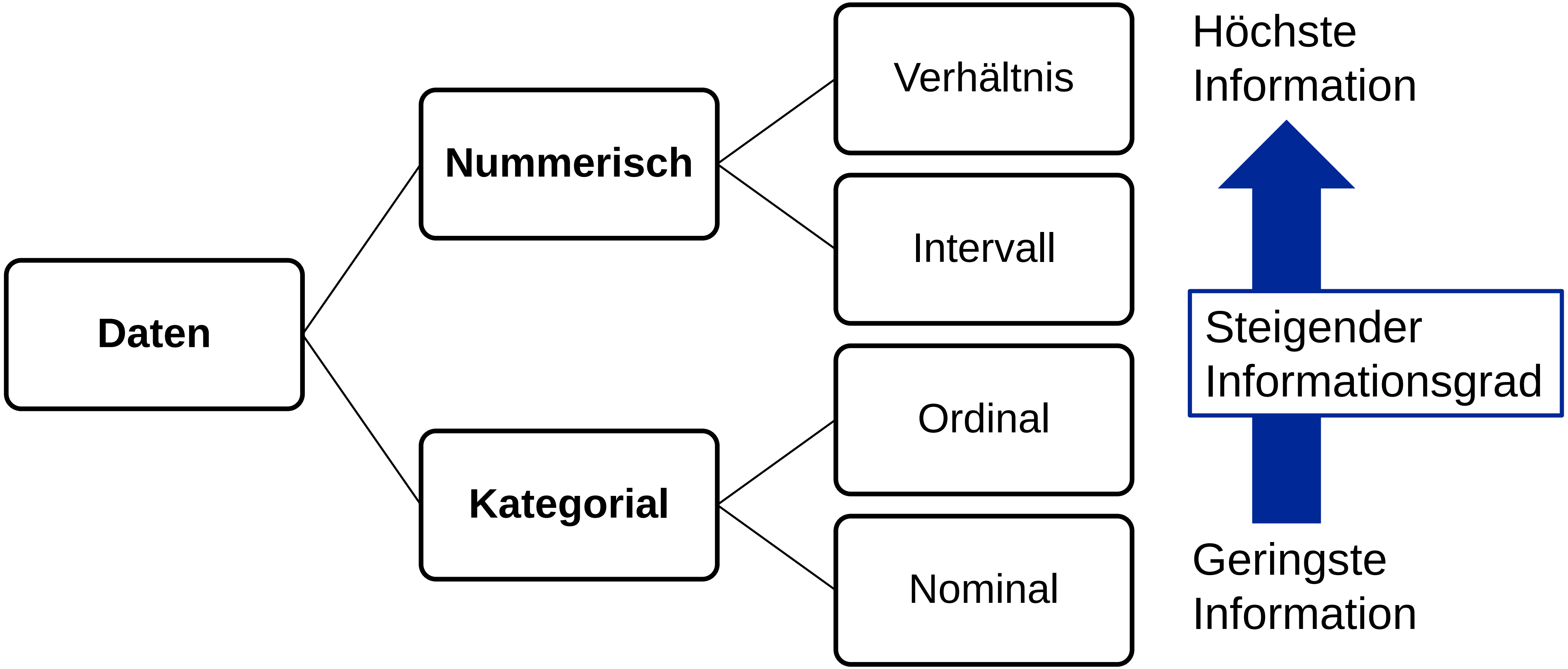
DATEN: MESSNIVEAUS

	Messniveau	Zulässige Operationen	Definition	Beispiele
Qualitative Daten	Nominal	$=, \neq$	Kategorien ohne Rangordnung	Geschlecht, Farben
	Ordinal	$=, \neq$ $\geq, \leq$	Kategorien mit Rangordnung	Sterne eines Hotels, Militärische Ränge
Quantitative Daten	Intervall	$=, \neq$ $\geq, \leq$ $+, -$	Nummerische Messungen. Differenzen sind interpretierbar. Kein absoluter Nullpunkt.	Temperatur in ℃, IQ-Test
	Verhältnis	$=, \neq$ $\geq, \leq$ $+, -$ $\times, \div$	Nummerische Messungen. Verhältnisse sind interpretierbar. Absoluter Nullpunkt.	Temperatur in K, Größe

DATEN: MESSNIVEAUS

HOCHSCHULE HEILBRONN

> Messniveaus haben unterschiedliche Informationsgrade:



DATEN: MESSNIVEAUS

HOCHSCHULE HEILBRONN

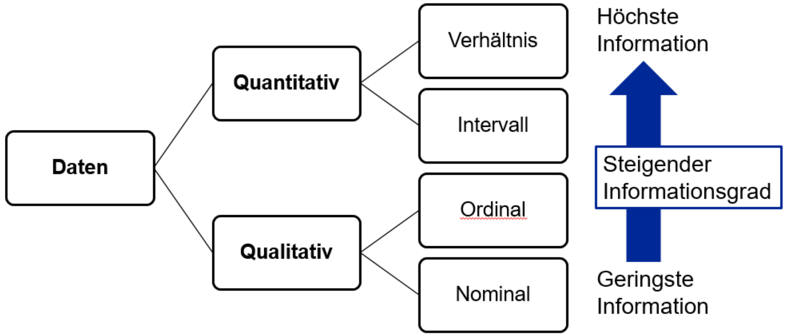
Beispiel. Wettrennen.

Erfassung auf Niveau

Teilnehmer	Nominal	Ordinal	Verhältnis
A	> Teilgenommen	> 2. Platz	> 16.2 Sek.
B	> Teilgenommen	> 1. Platz	> 11.3 Sek.
C	> Disqualifiziert	> -	> -
D	> Teilgenommen	> 3. Platz	> 16.9 Sek.

DATEN: MESSNIVEAUS

Informativere Messniveaus können in niedrigere konvertiert werden, aber nicht umgekehrt



Beispiel.

Teilnehmer	Nominal	Ordinal	Verhältnis
A	> Teilgenommen	⇏ > 2. Platz ⇐	> 16.2 Sek.
B	> Teilgenommen	⇏ > 1. Platz ⇐	> 11.3 Sek.
C	> Disqualifiziert	⇏ > - ⇐	> -
D	> Teilgenommen	⇏ > 3. Platz ⇐	> 16.9 Sek.

DATEN: MESSNIVEAUS

HOCHSCHULE HEILBRONN

Daten	Skalenniveau
Körpergröße (182cm, 167cm, …)	
Geschlecht	
IQ (100, 99, 116, 89,…)	
Prozentsatz korrekter Testantworten	
Platzierung in einem Schönheitswettbewerb	
Reiseziele (New York, Berlin)	
Zustimmungsgrade (stimmte nicht zu, …, stimme voll zu)	
Jahreszahlen (1640, 1920, 2020,…)	

Skalenniveau
Nominal
Ordinal
Intervall
Verhältnis

DATENARTEN: UNIVARIAT VS. MULTIVARIAT

HOCHSCHULE HEILBRONN

> **Univariate Daten:** messen ein **einziges** Merkmal

Beispiel. Merkmal = Größe:

Person i	1	2	3	4	5
Height in cm (x_i)	172	175	177	186	191

> **Multivariate Daten:** messen **mehrere** Merkmale

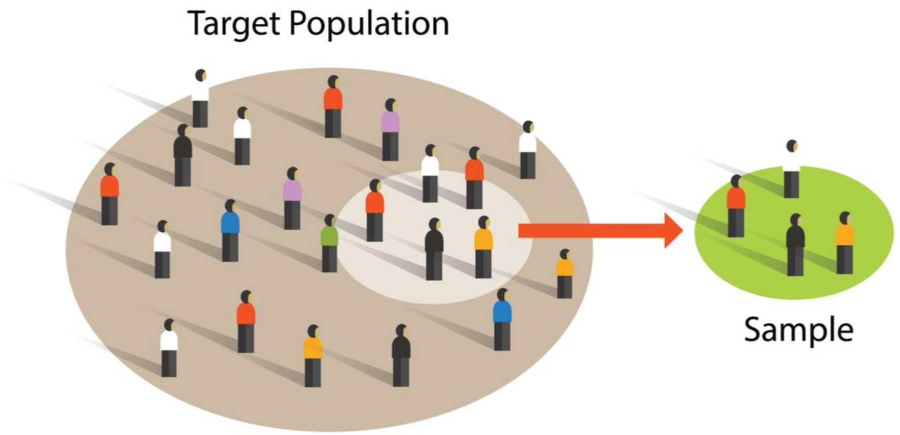
Beispiel. Merkmal = Größe + Gewicht:

Person i	1	2	3	4	5
Height in cm (x_i)	172	175	177	186	191
Weight in kg (y_i)	67	74	72	81	86

DATENERHEBUNG: ZUFALLSSTICHPROBEN

HOCHSCHULE HEILBRONN

Statistik: Erhebung von Daten über **Zufallsstichproben:**



Zufallsstichprobe (intuitiv):

> Stichprobe = Teilmenge einer Population (= Menge der Merkmalsträger)

> Zufall = Jedes Element der Population wird mit gleicher Wahrscheinlichkeit gezogen

› STATISTIKEN UND PARAMETER

STATISTIKEN UND PARAMETER

Kennzahlen

Statistiken	Parameter
> Kennzahlen einer **Stichprobe** > Gegenstand der deskriptiven Statistik	> Kennzahlen für eine **Populationsverteilung** > In der Regel nicht bekannt > Müssen geschätzt werden (Inferenzstatistik)

STATISTIKEN UND PARAMETER: BEISPIEL

HOCHSCHULE HEILBRONN

> Verteilung der Körpergröße von Frauen

> Population = Welt

> Stichprobe = Deutschland

Durchschnittsgröße Welt = 160 cm ≠ Durchschnittsgröße Deutschland = 165 cm

> Im Gegensatz zu Statistiken sind Parameter in der Regel **unbekannt**

> Die Inferenzstatistik zielt darauf ab Paramater zu **schätzen** (basierend auf Stichproben)

STATISTIKEN UND PARAMETER: ARTEN

<table>
<tr><th colspan="2">Univariat</th><th colspan="2">Multivariat</th></tr>
<tr><th>Art</th><th>Kennzahlen</th><th>Art</th><th>Kennzahlen</th></tr>
<tr><td>Lagemaße:
> Typische Werte einer Verteilung → charakterisieren deren Lage</td><td>> Modus / Modalwert
> Median
> Mittelwert / Erwartungswert</td><td rowspan="2">Zusammenhangsmaße:
> Messen die Stärke des linearen Zusammenhangs zwischen 2 Merkmalen</td><td rowspan="2">> Kovarianz
> Korrelations-koeffizient</td></tr>
<tr><td>Streuungsmaße:
> Messen Streuung um ein Lagemaß → charakterisieren die Form der Verteilung</td><td>> Varianz
> Standard-abweichung</td></tr>
</table>

AUSBLICK: PARAMETER UND STATISTIKEN DER VORLESUNG

Typ	Name	Notation in Lehrbüchern	
		Parameter	Statistik
Lage	Modus	-	$\bar{x}_M$
	Median	-	$\bar{x}_{0.5}$
	Mittelwert	μ	$\bar{x}$
	Anteil	π	$\bar{p}$
Streuung	Standard-abweichung	σ / SD	s
	Varianz	σ^2 / var	s^2
	Spannweite	-	$range$
Zusammenh.	Kovarianz	σ_{XY} / cov	s_{xy}
	Korrelationskoeff.	ρ_{XY}	r_{xy}

› UNIVARIATE STATISTIKEN / PARAMETER

LAGEMAẞE: TYPISCHE WERTE EINER VERTEILUNG

HOCHSCHULE HEILBRONN

Modus / Modalwert

Sei $(x_1, \dots, x_n)$ ein Datensatz der Länge n. Der *Modus* $\bar{x}_M$ ist derjenige Wert, der am häufigsten vorkommt (höchste absolute Häufigkeit).

Median

Sei $(x_1, \dots, x_n)$ ein Datensatz der Länge n. Der *Median* $\tilde{x}_{0.5}$ ist der "mittlere" Wert, wenn die Werte der Stichprobe von klein nach groß sortiert wurden.

Mittelwert

Sei $(x_1, \dots, x_n)$ ein Datensatz der Länge n. Der *Mittelwert* ist das arithmetische Mittel:

$$Mittelwert = \frac{x_1 + \cdots + x_n}{n}$$

LAGEMAßE: ERWARTUNGSWERT

Beispiel 1. Wette: Gewinn 2€ mit 70%, Verlust 3 € mit 30%.

> Was würde man „erwarten“, wenn man die sehr (unendlich) oft annimmt?

Beispiel 2. Würfelwurf.

> Welche Durchschnitts-Augenzahl würde man erwarten, bei unendlicher Wiederholung?

LAGEMAẞE: ERWARTUNGSWERT

Erwartungswert

Der Erwartungswert $E[X]$ eines Merkmals X ist die wahrscheinlichkeitsgewichtete Summe seiner Werte.

> Der Erwartungswert kann als "wahrer" langfristiger Durchschnittswert gesehen werden

LAGEMAßE: ERWARTUNGSWERT

HOCHSCHULE HEILBRONN

Beispiel 3. Durchschnittsnoten in Statistik der letzten Jahre:

Jahr	14	15	16	17	18	19	20	21	22	23
Durchschnittsnote	2,3	2,3	2,7	3,3	2,3	2,7	2,7	1,7	3,3	2,3

> Welche Durchschnittsnote würde man dieses Jahr „erwarten“?

LAGEMAẞE: ERWARTUNGSWERT

HOCHSCHULE HEILBRONN

Rechenregeln.

1. Der Erwartungswert einer Konstanten ist die Konstante:

$$E[Konstante] = Konstante$$

   Anmerkung. Deshalb: $E\big[E[X]\big] = E[X]$

2. Konstante Multiplikatoren können vor den Erwartungswert geschrieben werden:

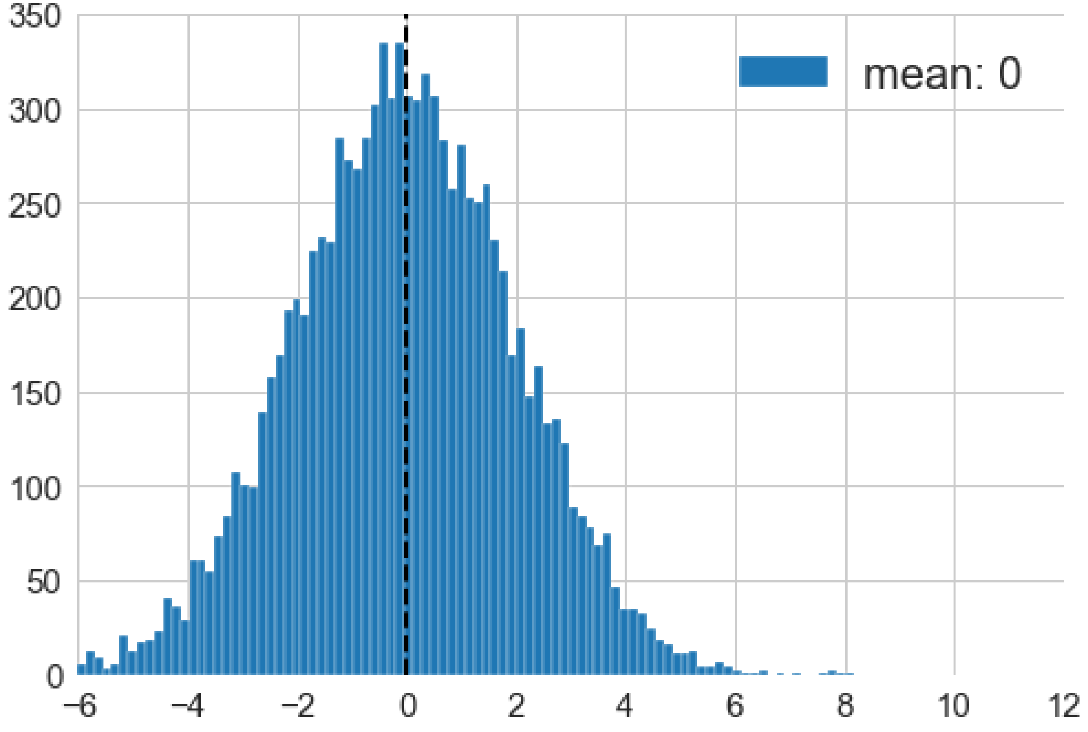
$$E[Konstante \cdot X] = Konstante \cdot E[X]$$

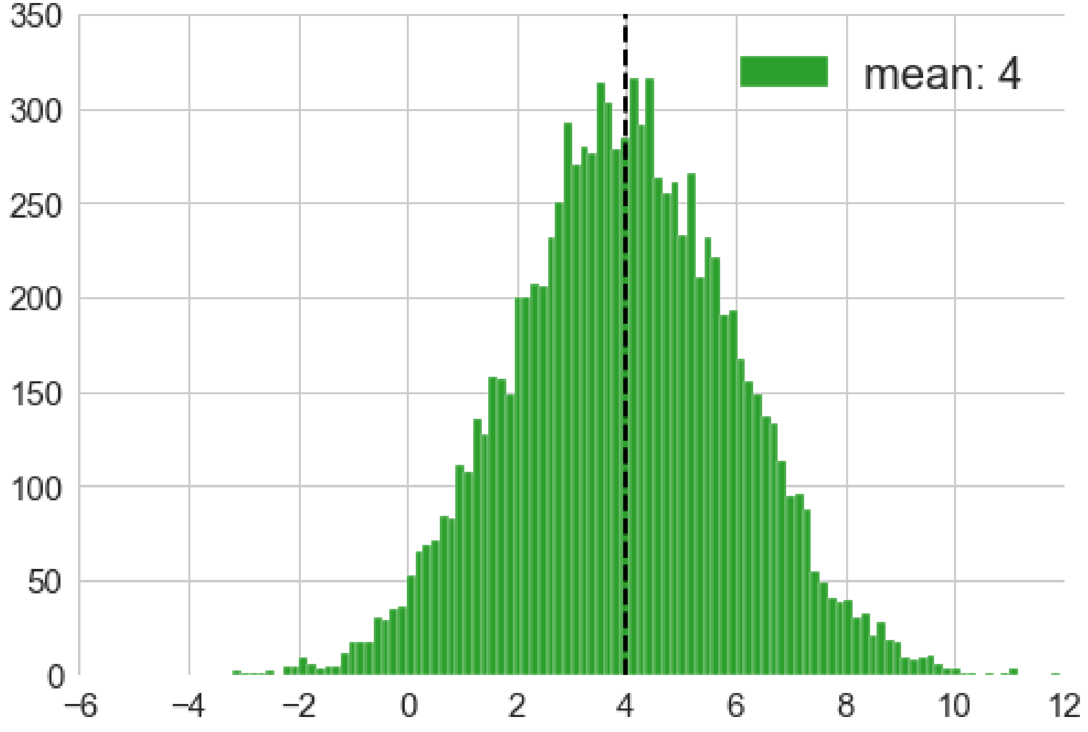
3. Der Erwartungswert ist additiv separabel:

$$E[X + Y] = E[X] + E[Y]$$

LAGEMAßE CHARAKTERISIEREN
DIE LAGE DER VERTEILUNG

HOCHSCHULE HEILBRONN

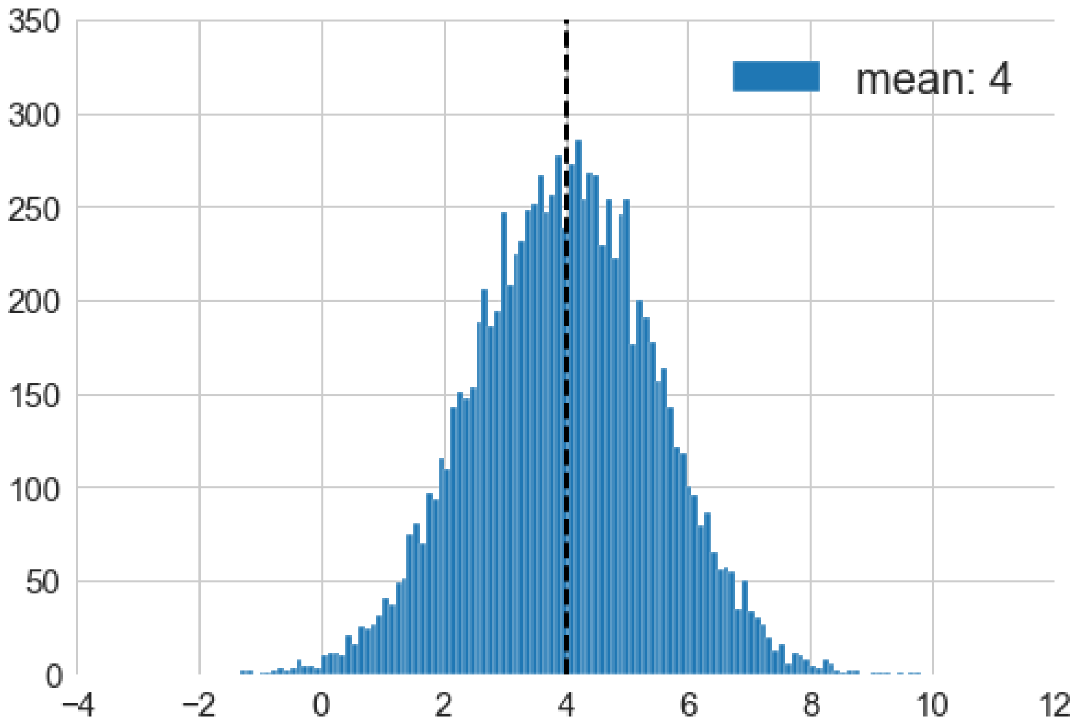




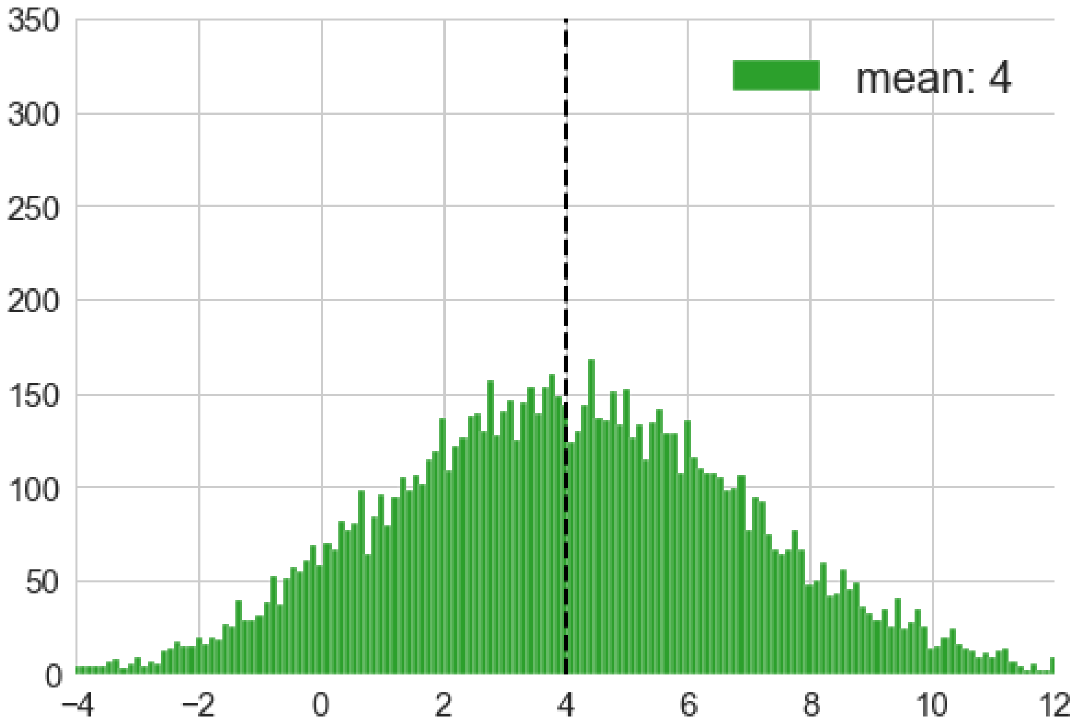
WARUM STREUUNGSMAßE?

HOCHSCHULE HEILBRONN

Niedrige Streuung um den Mittelwert



Hohe Streuung um den Mittelwert



Beide Verteilungen haben den gleichen Mittelwert (4), aber streuen unterschiedlich

WARUM STREUUNGSMAßE?

> Es gibt 2 Statistik-Kurse (A und B) → Durchschnittnote in beiden Kursen: 3.0

Class A	
Grade	**Abs_Frequency**
1	-
1.3	-
1.7	-
2	-
2.3	-
2.7	-
3	40
3.3	-
3.7	-
4	-
5	-
sum	40

Class B	
Grade	**Abs_Frequency**
1	20
1.3	-
1.7	-
2	-
2.3	-
2.7	-
3	-
3.3	-
3.7	-
4	-
5	20
sum	40

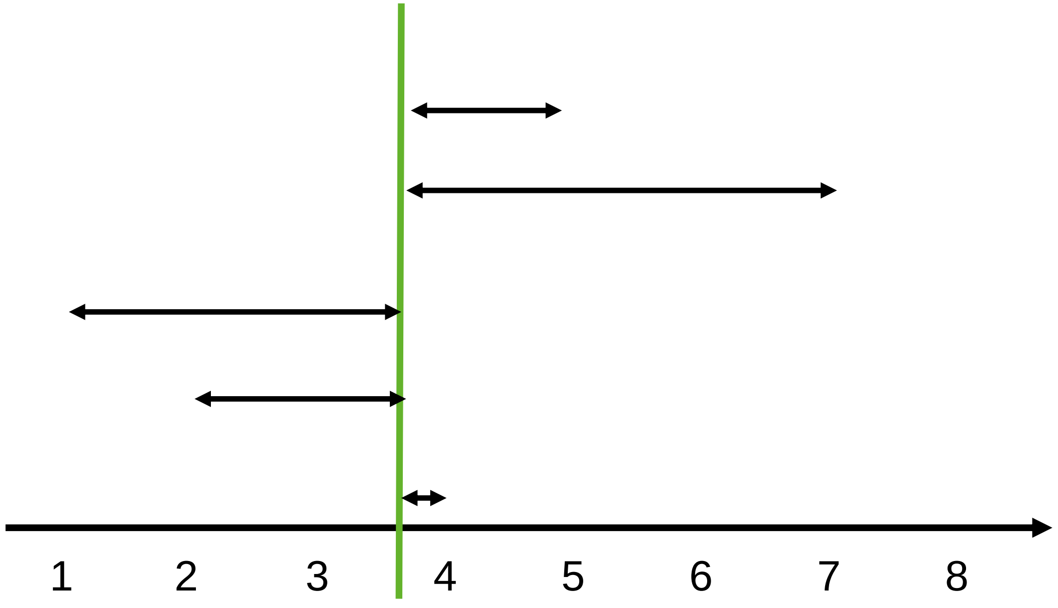
STREUUNGSMAßE: MESSEN STREUUNG UM DEN MITTELWERT

Maß für die Streuung der Daten um ihren Mittelwert

> Durchschnitt der Summe der Abweichungen vom Mittelwert?

Beispiel.

	x_i	$x_i - Mittelwert$
	5	1.2
	7	3.2
	1	-2.8
	2	-1.8
	4	0.2
Summe / n	$Mittelwert = 19/5 = 3.8$	$0/5 = 0$



$Mittelwert = 3.8$

STREUUNGSMAẞE: MESSEN STREUUNG UM DEN MITTELWERT

HOCHSCHULE HEILBRONN

> Durchschnitt der Summe der Abweichungen vom Mittelwert ist immer 0

> Lösung: Durchschnitt der Summe der **quadrierten** Abweichungen vom Mittelwerten

Beispiel.

	Beobachtung: x_i	**Abweichungen:** $x_i - Mittelwert$	**Quadrierte Abweichungen:** $(x_i - Mittelwert)^2$
	5	1.2	1.44
	7	3.2	10.24
	1	-2.8	7.84
	2	-1.8	3.24
	4	0.2	0.04
Summe / n	$Mittelw. = 3.8$	0	$\frac{(x_1 - Mittelw.)^2 + \cdots + (x_5 - Mittelw.)^2}{n} = 4.56$

Varianz

STREUUNGSMAßE: MESSEN STREUUNG UM DEN MITTELWERT

HOCHSCHULE HEILBRONN

Beispiel. Kalorien von Lebensmitteln.

Lebensmittel	Kalorien (pro 10 Gramm) x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
Pizza Margherita	228	32	1024
Döner	215	19	361
Hamburger	241	45	2025
Spaghetti mit Pesto	181	-15	225
Salat mit Meeresfrüchten	115	-81	6561
Summe	980		10‘196
Summe durch Anzahl (n)	$\bar{x} = 196$		$s^2 = 2'039.2$
Einheit	**Kalorien**		$\boldsymbol{Kalorien^2}$

STREUUNGSMAßE:
MESSEN STREUUNG UM DEN MITTELWERT

HOCHSCHULE HEILBRONN

- Varianz = 2039.2 Kalorien^2
- Kann nicht mit ursprünglichen Messungen (in Kalorien) in Verbindung gesetzt werden
- Deshalb: Standardabweichung = positive Quadratwurzel der Varianz
- Im Beispiel:

$$Standardabweichung = \sqrt{2039.2} \approx 45$$

- 45 Kalorien = Durchschnittliche Entfernung der 5 Produkte vom Mittelwert

STREUUNGSMAẞE: MESSEN STREUUNG UM DEN MITTELWERT

HOCHSCHULE HEILBRONN

Varianz

Sei $(x_1, \ldots, x_n)$ ein Datensatz der Länge n. Die *Varianz* ist die durchschnittliche quadratische Abweichung vom Mittelwert:

$$Varianz = \frac{(x_1 - Mittelwert)^2 + \cdots + (x_n - Mittelwert)^2}{n}.$$

Standardabweichung

Sei $(x_1, \ldots, x_n)$ ein Datensatz der Länge n. Die *Standardabweichung* ist die positive Quadratwurzel der Varianz:

$$Standardabweichung = \sqrt{Varianz}$$

STREUUNGSMAẞE: MESSEN STREUUNG UM DEN MITTELWERT

HOCHSCHULE HEILBRONN

Eigenschaften. Die Varianz / Standardabweichung

> kann nur für nummerische / quantitative Daten bestimmt werden

> ist größer, je stärker die Daten um ihren Mittelwert streuen

VARIANZ: BERECHNUNG MITTELS ERWARTUNGSWERT

Die Varianz ist der Erwartungswert der quadratischen Abweichung vom Mittelwert:

$$Var(X) = E[(X - Mittelwert)^2]$$

Dies kann wie folgt geschrieben werden:

$$\begin{aligned}
&E[(X - Mittelwert)^2] \\
&= E[X^2 - 2 \cdot X \cdot Mittelwert + Mittelwert^2] \\
&= E[X^2] - 2 \cdot E[X] \cdot Mittelwert + E[Mittelwert^2] \leftarrow \text{Konstante Multiplikatoren vor den Erwartungswert} \\
&= \boldsymbol{E[X^2] - Mittelwert^2} \leftarrow E[X] = Mittelwert
\end{aligned}$$

VARIANZ: BERECHNUNG MITTELS ERWARTUNGSWERT

HOCHSCHULE HEILBRONN

Beispiel. Einkommensverteilung

Einkommen x (in 1000€)	$P(x)$	$P(x) \cdot x$	x^2	$P(x) \cdot x^2$
1	0.2	0.2	1	0.2
1.5	0.1	0.15	2.25	0.23
2	0.2	0.4	4	0.8
4	0.4	1.6	16	6.4
6	0.1	0.6	36	3.6
Summe		2.95		11.23

$Mittelwert$ $\qquad$ $E[X^2] = 11.23$

$$Varianz = E[X^2] - Mittelwert^2 = 11.23 - 2.95^2 = 2.52$$

VARIANZ: VEREINFACHTE BERECHNUNG BEI ANTEILSWERTEN

HOCHSCHULE HEILBRONN

Beispiel. Geschlechterverteilung:

Person *i*	1	2	3	4
Geschlecht	weiblich	männlich	weiblich	weiblich
Indikator Weiblich	1	0	1	1

> Hier Varianz $= \boldsymbol{Anteilswert} \cdot (\mathbf{1} - \boldsymbol{Anteilswert})$

> Beweis:

$$Varianz(X)$$
$$= E[X^2] - Anteilswert^2$$
$$= Anteilswert - Anteilswert^2$$
$$= Anteilswert \cdot (1 - Anteilswert)$$
$$= 0.75 \cdot 0.25 = 0.1875$$

$1^2 = 1\ und\ 0^2 = 0$
$\rightarrow$
$E[X^2] = E[X]$
$= Mittelwert = Anteilswert$

› MULTIVARIATE STATISTIKEN / PARAMETER

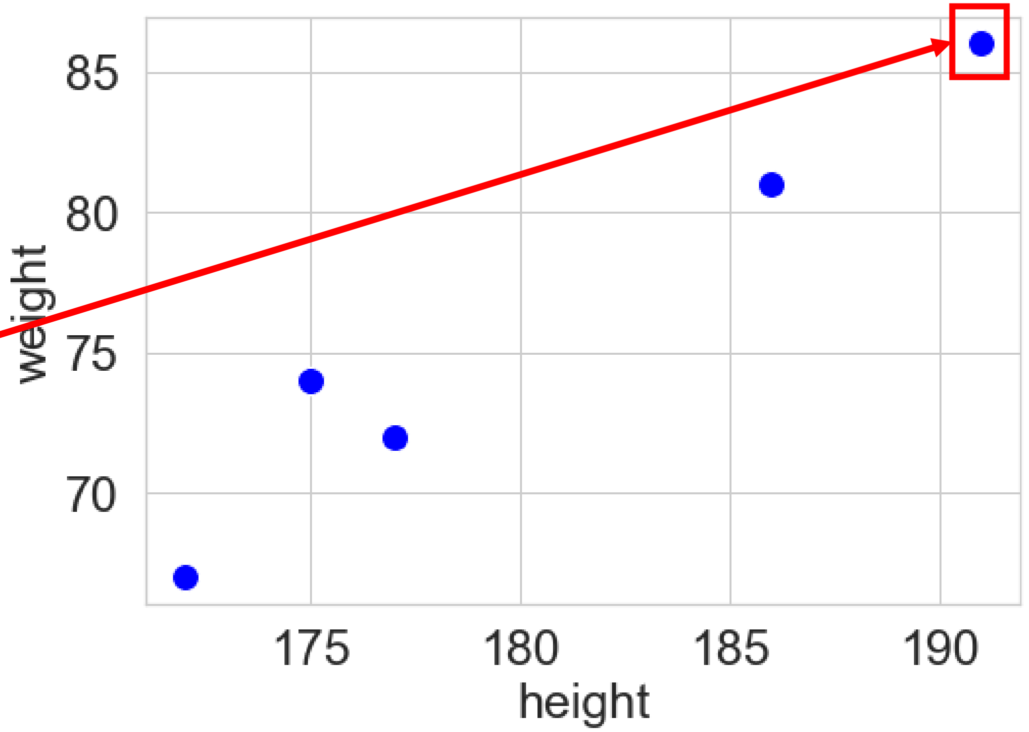
ZUSAMMENGHANGSMAßE FÜR BIVARIATE DATEN

HOCHSCHULE HEILBRONN

Häufig möchte man messen, ob zwei Merkmale eine Beziehung haben

Beispiel. Zusammenhang zwischen Größe und Gewicht.

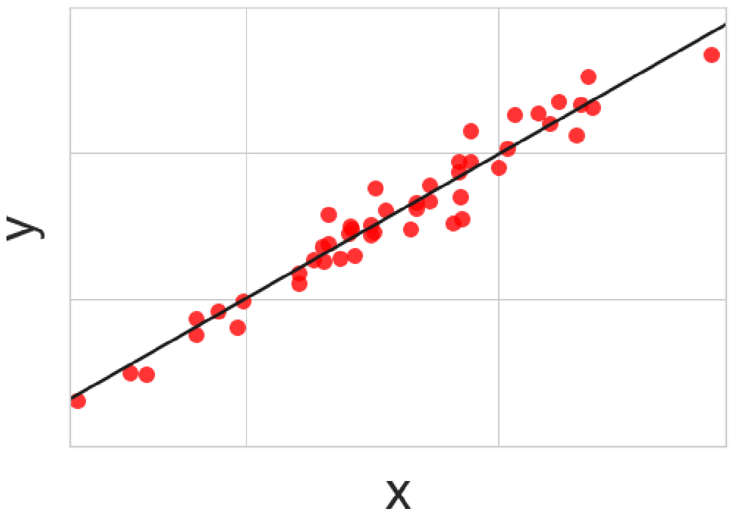
Person i	1	2	3	4	5
Height in cm (x_i)	172	175	177	186	191
Weight in kg (y_i)	67	74	72	81	86



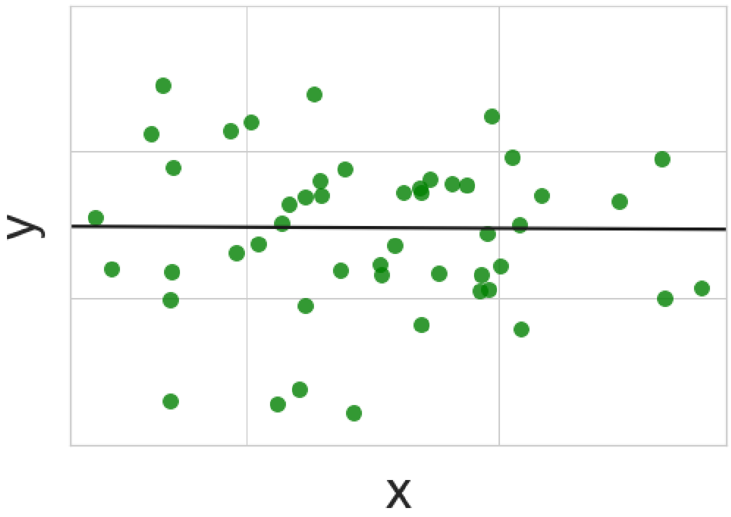
ZUSAMMENHANGSMAẞ: KOVARIANZ

Kovarianz: Misst Stärke der linearen Beziehung zwischen 2 Merkmalen X und Y

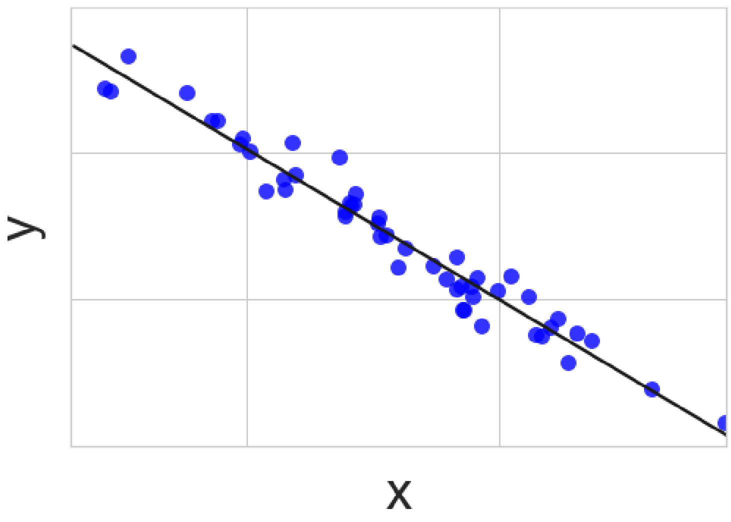
Positive Kovarianz:
$X \uparrow \rightarrow Y \uparrow$



Kovarianz nahe 0:
Keine lin. Beziehung



Negative Kovarianz:
$X \uparrow \rightarrow Y \downarrow$



ZUSAMMENHANGSMAß: KOVARIANZ

HOCHSCHULE HEILBRONN

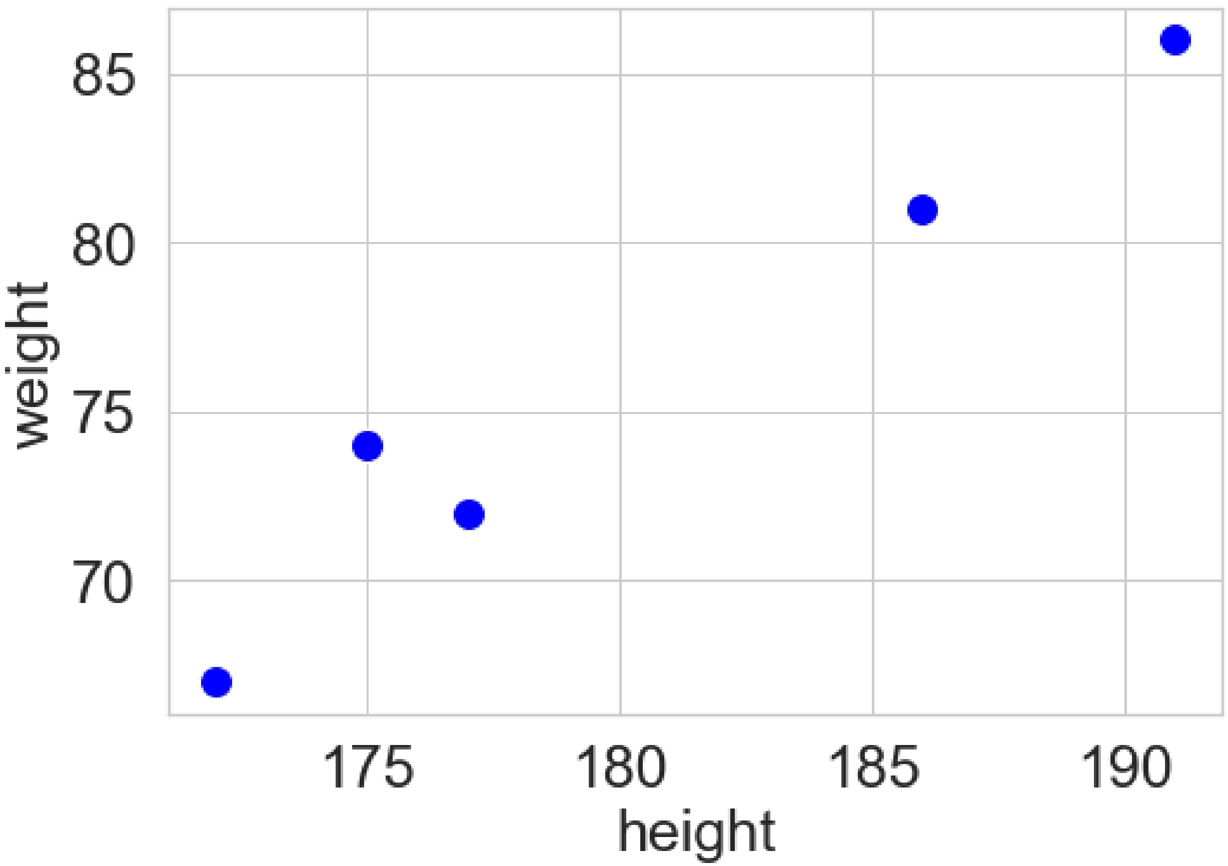
Kovarianz

Sei $((x_1, y_1) \dots, (x_n, y_n))$ Datensatz mit 2 Merkmalen der Größe n. Die *Kovarianz* der Merkmale ist der Durchschnitt des Produkts Ihrer Abweichungen vom Mittelwert:

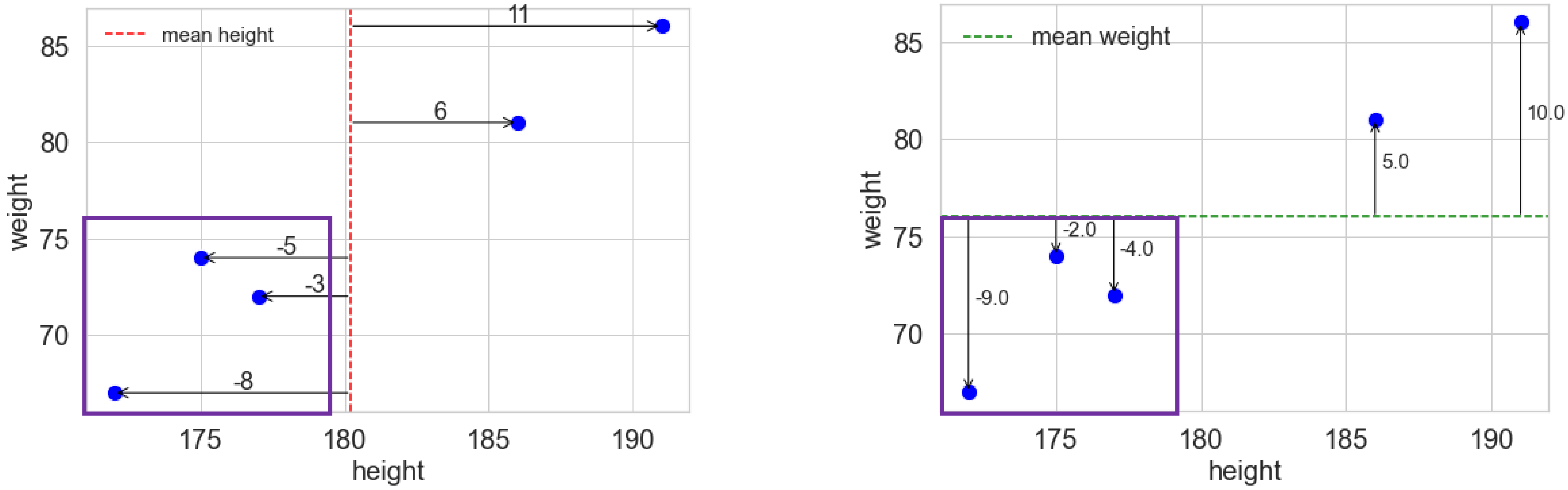
$$cov(X, Y) = \frac{(x_1 - Mwert_X) \cdot (y_1 - Mwert_Y) + \dots + (x_n - Mwert_X) \cdot (y_n - Mwert_Y)}{n}.$$

ZUSAMMENHANGSMAß: KOVARIANZ

Beispiel (Fortsetzung). Beziehung zw. Größe und Gewicht.

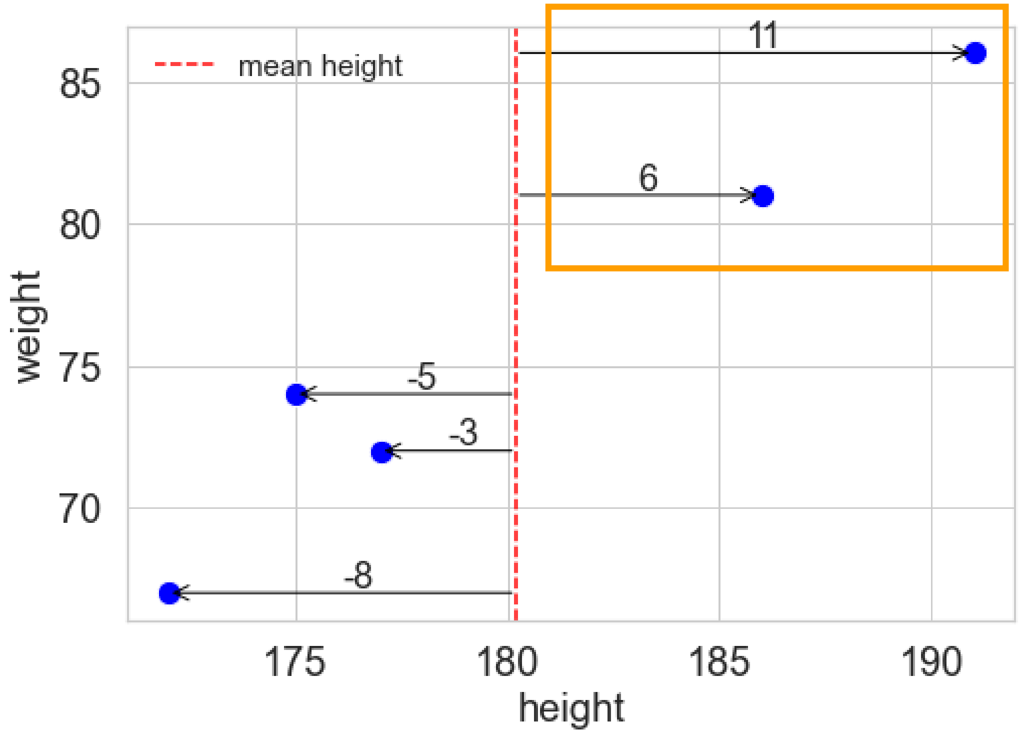


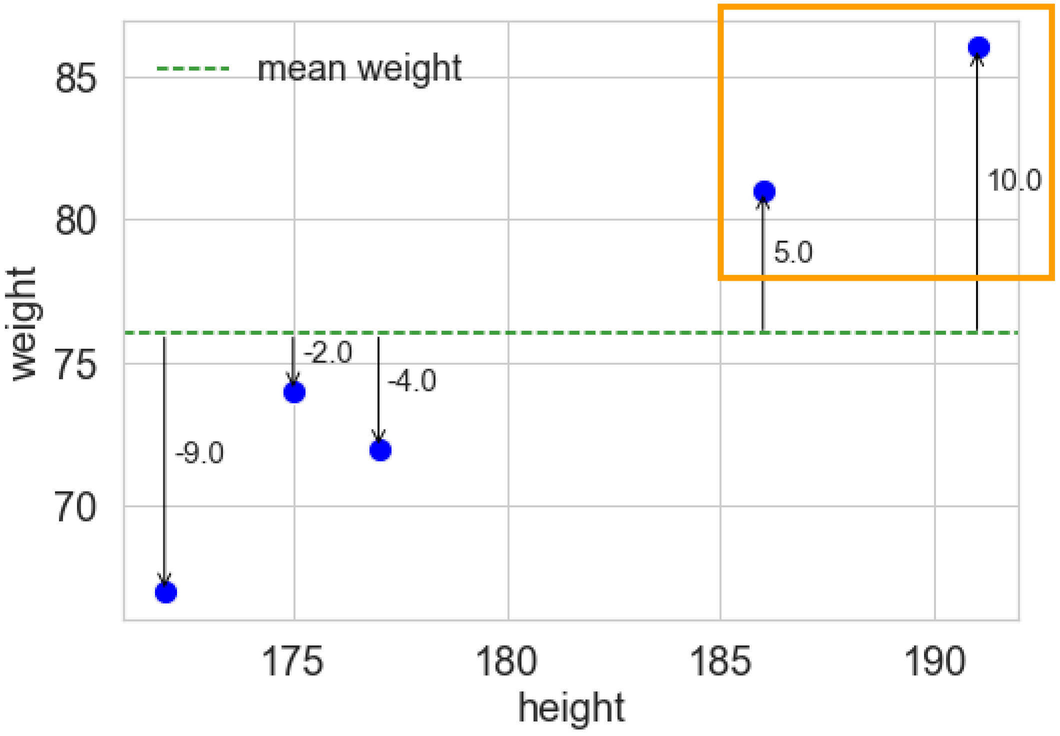
ZUSAMMENHANGSMAß: KOVARIANZ



Gewicht und Größer der **ersten 3 Personen** liegen **unter dem Durchschnitt**

ZUSAMMENHANGSMAß: KOVARIANZ

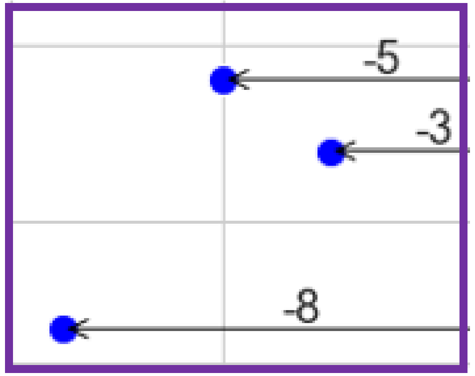




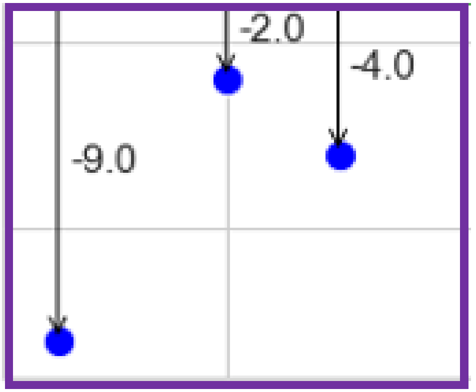
Gewicht und Größer der letzten **beiden Personen** liegen **über dem Durchschnitt**

ZUSAMMENHANGSMAß: KOVARIANZ

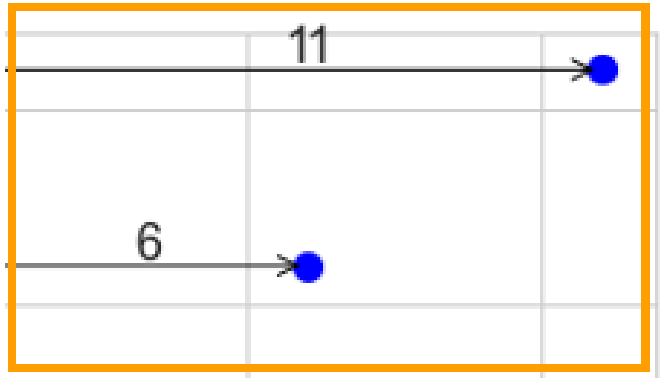
Berechnung.



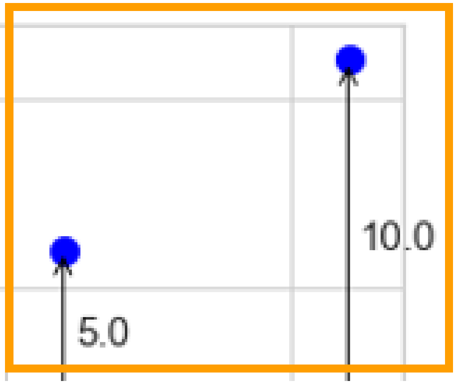
$\times$



$=$



$\times$



$=$

	Produkt Abweichungen: $(x_i - \bar{x}) \cdot (y_i - \bar{y})$
	72
	10
	12
	30
	110
Summe/ n	Kovarianz $s_{height,weight}$: $234/5 = 46.8$

ZUSAMMENHANGSMAß: KOVARIANZ

HOCHSCHULE HEILBRONN

Ausführliche Berechnung.

	Beobachtung: height x_i (in cm)	Abweichungen: $x_i - \bar{x}$	Beobachtung: weight y_i (in kg)	Abweichungen: $y_i - \bar{y}$	Produkt Abweichungen: $(x_i - \bar{x}) \cdot (y_i - \bar{y})$
	172	-8	67	-9	72
	175	-5	74	-2	10
	177	-3	72	-4	12
	186	6	81	5	30
	191	11	86	10	110
Summe / n	$Mittelwert_x = 180$		$Mittelwert_y = 76$		$Kovarianz = 46.8$

ZUSAMMENHANGSMAß: KOVARIANZ

HOCHSCHULE HEILBRONN

Die Kovarianz hängt von der Einheit der Merkmale ab.

Beispiel.

	Beobachtung: height x_i (m statt cm)	Abweichungen: $x_i - \bar{x}$	Beobachtung: weight y_i (in kg)	Abweichungen: $y_i - \bar{y}$	Produkt Abweichungen: $(x_i - \bar{x}) \cdot (y_i - \bar{y})$
	1.72	-0.08	67	-9	0.72
	1.75	-0.05	74	-2	0.1
	1.77	-0.03	72	-4	0.12
	1.86	0.06	81	5	0.3
	1.91	0.11	86	10	1.1
Summe / n	$Mittelwert_x = 1.8$		$Mittelwert_y = 76$		$Kovarianz = 0.468$ (vorher: 46.8)

ZUSAMMENHANGSMAß: KORRELATIONSKOEFFIZIENT

- Wir möchten ein normiertes Maß, das nicht von der Einheit abhängt
- Korrelationskoeffizient = Normierte Kovarianz
- Koeffizient = Kovarianz geteilt durch das Produkt der Standardabweichungen

Korrelationskoeffizient

Sei $((x_1, y_1) \ldots, (x_n, y_n))$ Datensatz mit 2 Merkmalen der Größe n. Der *Korrelationskoeffizient* ist die Kovarianz geteilt durch das Produkt der Standardabweichungen:

$$Korrelationskoeffizient = \frac{Kovarianz\ x\ und\ y}{Standardabweichung\ x \cdot Standardabweichung\ y}.$$

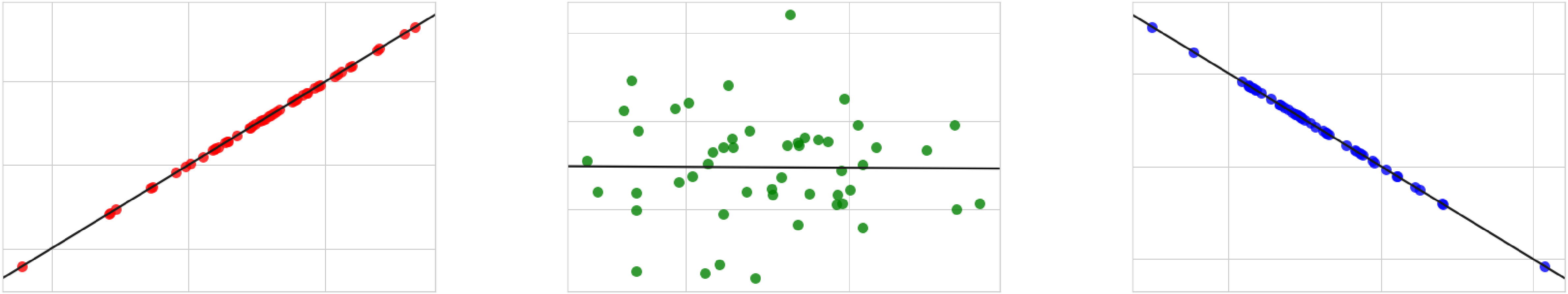
ZUSAMMENHANGSMAß: KORRELATIONSKOEFFIZIENT

HOCHSCHULE HEILBRONN

$r_{x,y} = 1$
perfekt positiv linear

$r_{x,y} = 0$
nicht linear

$r_{x,y} = -1$
perfekt negativ linear



ZUSAMMENHANGSMAß: KORRELATIONSKOEFFIZIENT

HOCHSCHULE HEILBRONN

Beispiel.

	height x_i (in cm)	height x_i (in m)	weight y_i (in kg)		
	172	1.72	67		
	175	1.75	74		
	177	1.77	72		
	186	1.86	81		
	191	1.91	86		
Standard-abweichung	7.139	0.071	6.723	**height in cm vs weight**	**height in m vs weight**
Kovarianz				46.8	0.468
Korrelations-koeffizienten				$\frac{46.8}{7.139 \times 6.723} = 0.975$	$\frac{0.468}{0.071 \times 6.723} = 0.975$

ZUSAMMENHANGSMAẞE: KOVARIANZ UND KORRELATIONSKOEFFIZIENT

HOCHSCHULE HEILBRONN

Eigenschaften. Der Korrelationskoeffizient / Die Kovarianz

- kann nur für nummerische / quantitative Daten bestimmt werden
- Kovarianz nimmt Werte zwischen $-\infty$ und ∞ an
- Korrelationskoeffizient nimmt Werte zwischen -1 und +1 an
- Kovarianz hängt von der Einheit der Messungen ab, Korrelationskoeffizient nicht
- Je negativer (positiver), desto stärker ist der negative (positive) lineare Zusammenhang

› STATISTIKEN UND MESSNIVEAUS

STATISTIKEN UND ERFORDERLICHE MESSNIVEAUS

		kann berechnet werden für		
Art	Statistik	Nummerische Daten	Kategoriale Daten	
			Ordinal	Nominal
Lagemaß	Modalwert	✓	✓	✓
	Median	✓	✓	✗
	Mittelwert	✓	✗	✗
Streuungsmaß	Alle	✓	✗	✗
Zusammenhang	Alle	✓	✗	✗

HOCHSCHULE HEILBRONN

› S.2 WAHRSCHEINLICH-KEITSTHEORIE

Statistik | Prof. Dr. Florian Kauffeldt

› ZUFALLSEXPERIMENT UND WAHRSCHEINLICHKEIT

UNSICHERHEIT UND WAHRSCHEINLICHKEIT

HOCHSCHULE HEILBRONN

In Alltag sagen wir Sätze wie:

- Ich bin sicher, dass Bayern München morgen gewinnt
- Wahrscheinlich werde ich Statistik bestehen
- Möglicherweise wird es morgen regnen

→ Ausdruck unterschiedlicher Grade von Unsicherheit über künftige Ereignisse

- Wahrscheinlichkeiten formalisieren dieses Konzept

UNSICHERHEIT UND WAHRSCHEINLICHKEIT

HOCHSCHULE HEILBRONN

- Wahrscheinlichkeit = Maß zwischen 0 und 1, das Unsicherheit repräsentiert
- Je höher die Wahrscheinlichkeit eines Ereignisses, desto eher wird es eintreten

Notation. Die Wahrscheinlichkeit eines Ereignisses E bezeichnen wir mit $P(E)$

- Ein Ereignis E mit $P(E) = 0$ heißt *unmöglich*
- Ein Ereignis E mit $P(E) = 1$ heißt *sicher*

STICHPROBEN-
RAUM S

Ein *Zufallsexperiment* ist ein Vorgang der

> unter identischen Bedingungen wiederholt werden kann und

> dessen mögliche Ergebnisse im Vorhinein bekannt und unsicher sind

Stichprobenraum

Die Menge aller möglicher Ergebnisse eines Zufallsexperiments heißt *Stichprobenraum* S.

Beispiel. Würfelwurf. Stichprobenraum:

$$S = \{⚀, ⚁, ⚂, ⚃, ⚄, ⚅\}$$

MEHRSTUFIGE ZUFALLSEXPERIMENTE

HOCHSCHULE HEILBRONN

Mehrstufiges Zufallsexperiment = fixierten Anzahl von Wiederholungen eines Experiments.

> Mehrstufige Experimente haben größere Stichprobenräume

Beispiel. Münzwurf.

> Stichprobenraum beim einfachen Münzwurf:

$$Stichprobenraum = \{H, T\},$$

wobei H = Kopf, T = Zahl

> Stichprobenraum beim zweifachen Münzwurf (mehrstufiges Experiment):

$$Stichprobenraum = \{HH, HT, TH, TT\},$$

wobei HT = Kopf im ersten Wurf, Zahl im zweiten Wurf usw.

Mehrstufige Zufallsexperimente können in Baumdiagrammen dargestellt werden

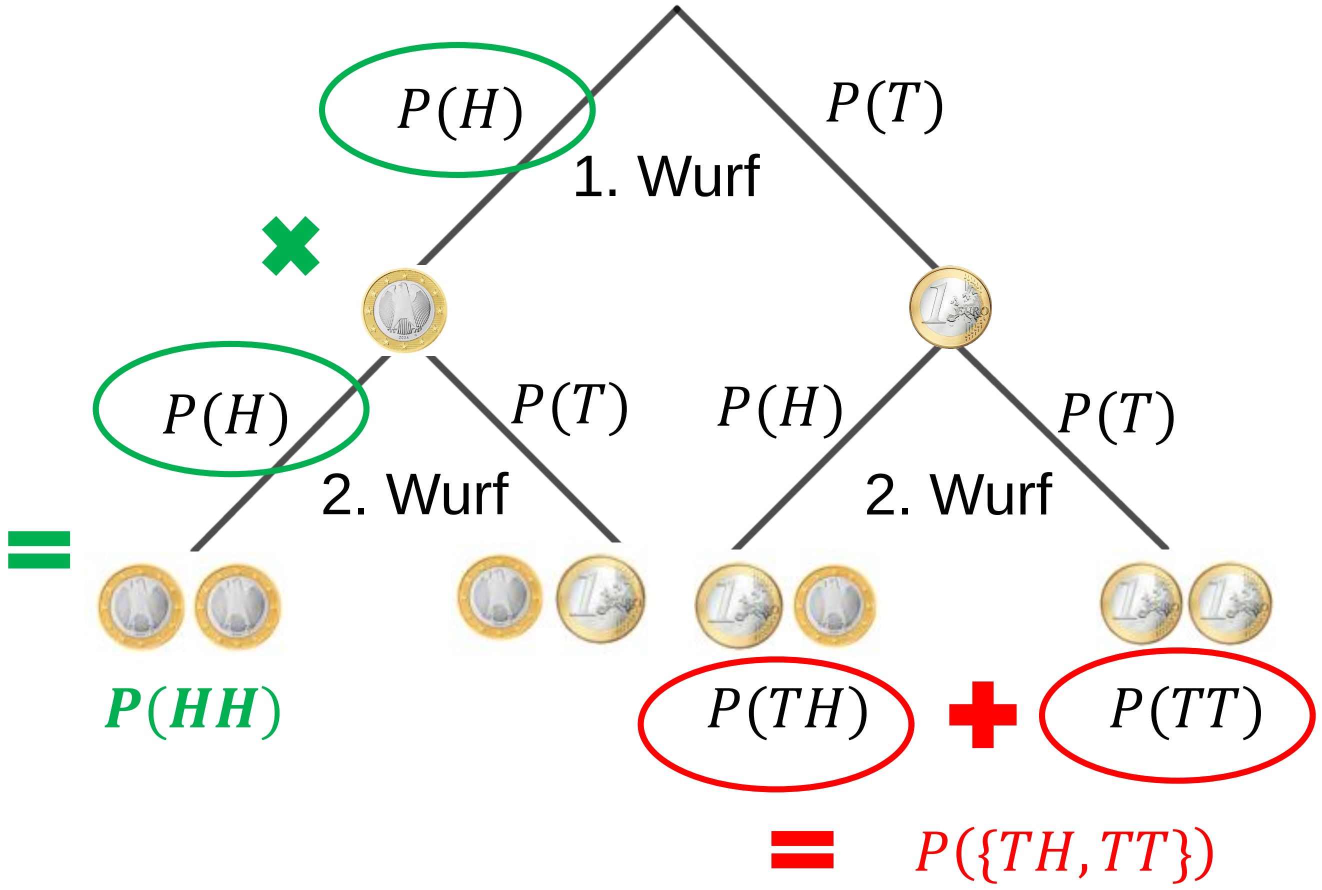
Beispiel. Zweimaliger Münzwurf.

> Produktregel:

$$P(HH) = P(H) \times P(H) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

> Summenregel:

$$P(\{TH, TT\}) = P(TH) + P(H) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$



WAS SIND EREIGNISSE?

Ereignis E
Ein *Ereignis* E ist eine Menge möglicher Ergebnisse eines Zufallsexperiments.

> Ein Ereignis ist also eine Teilmenge des Stichprobenraums: $E \subseteq S$

Beispiel. Würfelwurf.

> Ereignis $E =$ Würfelwurf: ⚁ oder ⚄ (Kurz: $E = \{$⚁, ⚄$\}$)

WAS SIND EREIGNISSE?

Besondere Ereignisse

> Einzelne Ergebnisse des Stichprobenraums $s \in S$ heißen *Elementarereignisse*

Beispiel. Würfelwurf. $s =$ Wurf ⚁ (Kurz: $s = \{$⚁$\}$)

> Der Stichprobenraum selbst ist ein Ereignis ("irgendetwas passiert"): $E = S$

> Die leere Menge $\emptyset$ ist ein Ereignis ("nichts passiert"): $E = \emptyset$

Anmerkung. $\emptyset$ ist eine Menge (die nichts enthält) und keine Zahl! $\emptyset \neq 0$

OPERATIONEN MIT EREIGNISSEN

Linda ist 31 Jahre alt, Single, offen und sehr intelligent. Sie hat Philosophie studiert. Als Studentin hat sie sich intensive mit Diskriminierung und sozialer Gerechtigkeit beschäftigt sowie an Anti-Atomkraft-Demonstrationen teilgenommen.

Was ist wahrscheinlicher?

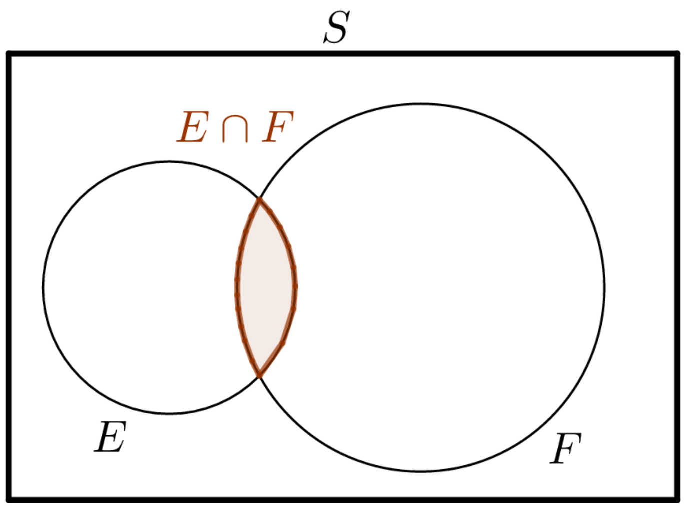
1. Linda ist eine Bankangestellte.
2. Linda ist eine Bankangestellte und aktive Feministin.

OPERATIONEN MIT EREIGNISSEN: SCHNITT ∩

HOCHSCHULE HEILBRONN

- Ereignis E = Bankangestellte
- Ereignis F = Aktive Feministin

Schnitt: $\boldsymbol{E \cap F} = \text{Bankangestellte } UND \text{ aktive Feministin}$



Wahrscheinlichkeit Ereignis
= Fläche Ereignis / Fläche von S

- Fläche von E ist größer als die Fläche von $E \cap F$ $\rightarrow P(E) > P(E \cap F)$

OPERATIONEN MIT EREIGNISSEN: SCHNITT ∩

HOCHSCHULE HEILBRONN

Weiteres Beispiel. Würfelwurf

- Ereignis E = ⚁ oder ⚂ oder ⚄ oder ⚅
- Ereignis F = ⚁ oder ⚃ oder ⚄

Schnitt?

- Ereignis $E \cap F$ = ⚁ oder ⚄

Wahrscheinlichkeiten?

- $P(E) = \frac{4}{6} = 2/3$
- $P(F) = \frac{3}{6} = 1/2$

Gemeinsame Wahrscheinlichkeit?

- $P(E \cap F) = \frac{2}{6} = 1/3 < P(E) = 2/3$

Anmerkung. Zwei Ereignisse heißen *disjunkt* wenn ihr Schnitt leer ist ($E \cap F = \emptyset$).

OPERATIONEN MIT EREIGNISSEN: VEREINIGUNG $\cup$

Linda ist 31 Jahre alt, Single, offen und sehr intelligent. Sie hat Philosophie studiert. Als Studentin hat sie sich intensive mit Diskriminierung und sozialer Gerechtigkeit beschäftigt sowie an Anti-Atomkraft-Demonstrationen teilgenommen.

Was ist wahrscheinlicher?

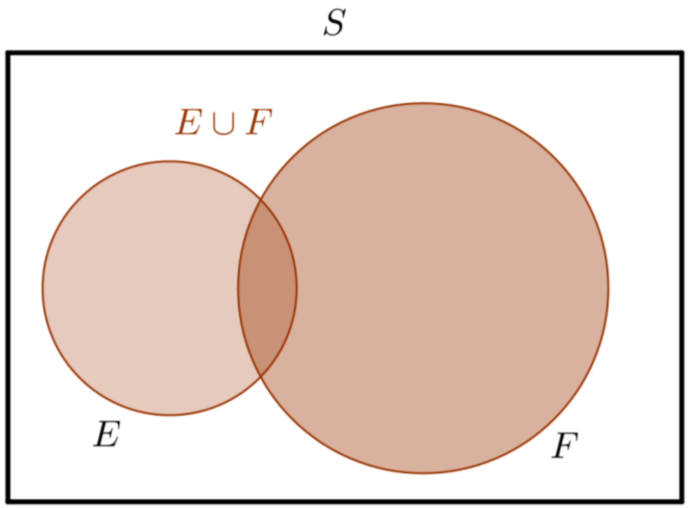
1. Linda ist eine Bankangestellte.
2. Linda ist eine Bankangestellte oder aktive Feministin.

OPERATIONEN MIT EREIGNISSEN: VEREINIGUNG ∪

HOCHSCHULE HEILBRONN

- Ereignis E = Bankangestellte
- Ereignis F = Aktive Feministin

Vereinigung: $\boldsymbol{E} \cup \boldsymbol{F} = \text{Bankangestellte } \mathit{ODER/UND} \text{ aktive Feministin}$



Wahrscheinlichkeit Ereignis
= Fläche Ereignis / Fläche von S

- Fläche von E ist kleiner als die Fläche von $E \cup F$ $\rightarrow P(E) < P(E \cup F)$

OPERATIONEN MIT EREIGNISSEN: VEREINIGUNG ∪

Weiteres Beispiel. Würfelwurf

- Ereignis E = ⚁ oder ⚂ oder ⚄ oder ⚅
- Ereignis F = ⚁ oder ⚃ oder ⚄

Vereinigung?

- Ereignis $E \cup F$ =⚁ oder ⚂ oder ⚃ oder ⚄ oder ⚅

Wahrscheinlichkeiten?

- $P(E) = \frac{4}{6} = 2/3$
- $P(F) = \frac{3}{6} = 1/2$

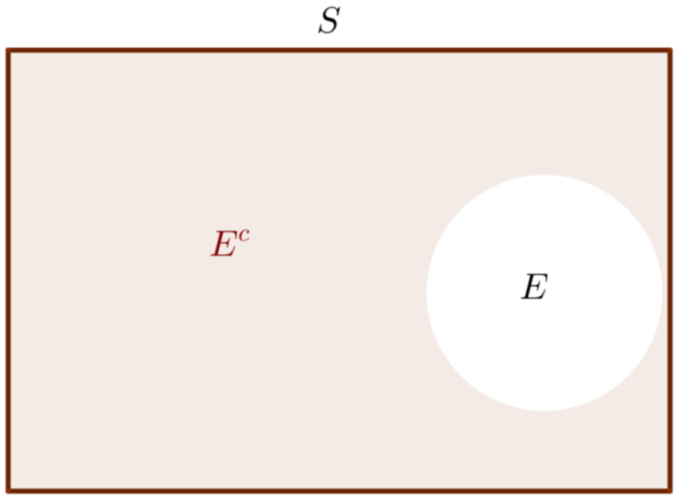
Vereinigungswahrscheinlichkeit?

- $P(E \cup F) = 5/6 > P(E) = 2/3$

OPERATIONEN MIT EREIGNISSEN: KOMPLEMENTÄREREIGNIS E^c

- Ereignis E = Bankangestellte

Komplement: E^c = $\textit{keine}$ Bankangestellte



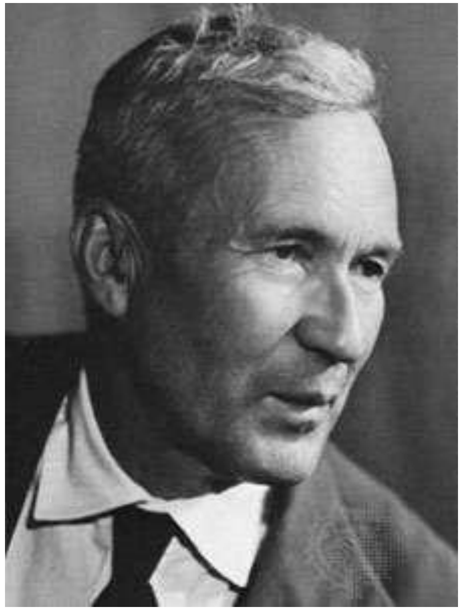
Weiteres Beispiel. Münzwurf.

- Ereignis E = ⚁ oder ⚂ oder ⚄ oder ⚅
- Komplementärereignis = E^c = ⚀ oder ⚃

WAS SIND WAHRSCHEINLICHKEITEN?

Andrey Kolmogorov (1903 – 1987)

- hat Wahrscheinlichkeiten anhand von mathematischen Eigenschaften (Axiome) definiert



WAS SIND WAHRSCHEINLICHKEITEN?

HOCHSCHULE HEILBRONN

Wahrscheinlichkeitsaxiome (Kolmogorov, 1928):

(A1) Nichtnegativität. Die Wahrscheinlichkeit jedes Ereignisses E ist größer oder gleich 0:

$$P(E) \geq 0.$$

(A2) Normalisierung. Irgendetwas muss passieren: Der Stichprobenraum S ist sicher:

$$P(S) = 1.$$

(A3) Additivität. Die Wahrscheinlichkeit von disjunkten Ereignissen E, F ist:

$$P(E \cup F) = P(E) + P(F).$$

Erinnerung: Zwei Ereignisse sind disjunkt, wenn $E \cap F = \emptyset$

WAS FOLGT AUS DEN WAHRSCHEINLICHKEITSAXIOMEN?

HOCHSCHULE HEILBRONN

> Irgendetwas muss passieren → Wahrscheinlichkeit, dass nichts passiert ist 0:

$$P(\emptyset) = 0$$

> Die Wahrscheinlichkeit jedes Ereignisses liegt zwischen 0 und 1:

$$0 \leq P(E) \leq 1$$

> "Kleinere" Ereignisse haben kleinere Wahrscheinlichkeiten:

$$Wenn\ E\ Teilmenge\ von\ F\ ist, dann\ P(E) \leq P(F)$$

WAS FOLGT AUS DEN WAHRSCHEINLICHKEITSAXIOMEN?

HOCHSCHULE HEILBRONN

> Die Vereinigungswahrscheinlichkeit von zwei Ereignissen E und F ist:

$$P(E \cup F) = P(E) + P(F) - P(E \cap F)$$

Beispiel. $E = \{$⚀, ⚁, ⚄$\}$ und $F = \{$⚁, ⚂, ⚄, ⚅$\}$:

- Wahrscheinlichkeiten? $P(E) = 1/2$ und $P(F) = 2/3$
- Schnitt? $E \cap F = \{$⚁, ⚄$\}$
- Gemeinsame Wahrscheinlichkeit? $P(E \cap F) = 1/3$
- Vereinigungswahrscheinlichkeit: $P(E \cup F) = P(E) + P(F) - P(E \cap F) = \frac{1}{2} + \frac{2}{3} - \frac{1}{3} = \frac{5}{6}$

EXKURS: PARTITION

Eine *Partition* des Stichprobenraums S ist eine Menge von Ereignissen,

- die **paarweise disjunkt** sind und
- deren **Vereinigung der Stichprobenraum** ist

Beispiele. Würfelwurf. $Stichprobenraum = \{$⚀,⚁,⚂,⚃,⚄,⚅$\}$. Partitionen:

- $E_1 = \{$⚀,⚁,⚂$\}$ und $E_2 = \{$⚃,⚄,⚅$\}$
- $E_1 = \{$⚀,⚂$\}$ und $E_2 = \{$⚁,⚃,⚄,⚅$\}$
- Weitere?

WAS FOLGT AUS DEN WAHRSCHEINLICHKEITSAXIOMEN?

HOCHSCHULE HEILBRONN

> Die Summe der Wahrscheinlichkeiten jeder Partition ist 1:

$$P(E_1) + \cdots + P(E_k) = 1$$

für jede Partition $\{E_1, \ldots, E_k\}$ von S.

Beispiele (Fortsetzung).

> $E_1 = \{$⚀,⚁,⚂$\}$ und $E_2 = \{$⚃,⚄,⚅$\} \rightarrow P(E_1) = P(E_2) = 1/2 \rightarrow Summe = 1$

> $E_1 = \{$⚀,⚂$\}$ und $E_2 = \{$⚁,⚃,⚄,⚅$\} \rightarrow P(E_1) = 1/3$ and $P(E_2) = 2/3 \rightarrow Summe = 1$

Anmerkung. Die einfachste Partition ist ein Ereignis E und dessen Komplement E^c. Also:

$$P(E) + P(E^c) = 1$$

WOHER KOMMEN WAHRSCHEINLICHKEITEN?

HOCHSCHULE HEILBRONN

Konzepte

- Klassisch (Laplace)
- Objektiv (Frequentistisch)
- Subjektiv (Bayesianisch)

WOHER KOMMEN WAHRSCHEINLICHKEITEN? KLASSISCHER ANSATZ

HOCHSCHULE HEILBRONN

Pierre-Simon Laplace (1749 – 1827)

Hypothese: Prinzip des unzureichenden Grundes

- Reduktion Ereignisse auf gleichwahrscheinliche symmetrische Fälle
- Prinzip des unzureichenden Grundes: Kein Grund für andere Wahrscheinlichkeit
- Wahrscheinlichkeit Ereignisses E = Verhältnis günstiger Fälle (E tritt ein) zur Gesamtzahl

$$\boxed{P(E) = \frac{Anzahl\ günstige\ Fälle\ für\ E}{Gesamtzahl\ Fälle}}$$

WOHER KOMMEN WAHRSCHEINLICHKEITEN? KLASSISCHER ANSATZ

Beispiel. Klassischer Ansatz: Würfelwurf.

> Wahrscheinlichkeit des Ereignisses $E = \{$⚁, ⚃$\}$ bestimmen (Augenzahl: 2 oder 4)?

> Gesamtanzahl an Fällen?

6 (⚀,⚁,⚂,⚃,⚄,⚅)

> Fälle, die für E günstig sind?

2 (⚁,⚃)

> Wahrscheinlichkeit:

$$P(E) = P(\{\text{⚁},\text{⚃}\}) = \frac{2}{6} \approx 33\%$$

WOHER KOMMEN WAHRSCHEINLICHKEITEN? FREQUENTISTISCHER ANSATZ

HOCHSCHULE HEILBRONN

Richard von Mises, Egon Pearson, John Venn,…

Hypothese: Relative Häufigkeiten & Gesetz der großen Zahl

> Häufige Wiederholung Zufallsexperiment:

Relative Häufigkeiten → "Wahre" Wahrscheinlichkeiten

> Die Wahrscheinlichkeit eines Ereignisses E kann wie folgt geschätzt werden:

$$\boxed{P(E) = \frac{Anzahl\ Versuche\ bei\ denen\ E\ eingetreten\ ist}{Gesamtanzahl\ Versuche}}$$

WOHER KOMMEN WAHRSCHEINLICHKEITEN? FREQUENTISTISCHER ANSATZ

Beispiel. Frequentistischer Ansatz: Würfelwurf.

- Wir möchten die Wahrscheinlichkeit des Ereignisses $E = \{⚁, ⚃\}$ bestimmen (Augenzahl: 2 oder 4)
- Der Würfel wird 600'000 Mal geworfen
- Eine 2 oder 4 wurde 197'123 geworfen
- Wahrscheinlichkeit:

$$P(E) = P(\{⚁,⚃\}) = \frac{197'123}{600'000} \approx 33\%$$

WOHER KOMMEN WAHRSCHEINLICHKEITEN? BAYESIANISCHER ANSATZ

HOCHSCHULE HEILBRONN

Thomas Bayes (1701 – 1761)

Hypothese: Subjektives Urteil

> Wahrscheinlichkeiten sind subjektive Einschätzungen über den Grad der Gewissheit mit welchem Ereignisse eintreten

> Subjektive Wahrscheinlichkeitsurteile basieren auf individuellem Vorwissen

WOHER KOMMEN WAHRSCHEINLICHKEITEN? BAYESIANISCHER ANSATZ

Bayesianischer Ansatz. Unterscheidung zwischen:

A−Priori Wahrscheinlichkeit:
Subjektive Wahrscheinlichkeit $P(E)$

Beobachtung
Ereignis F

A-Posteriori Wahrscheinlichkeit:
Bedingte Wahrscheinlichkeit $P(E \mid F)$

> Bedingte Wahrscheinlichkeit $P(E \mid F)$: W'keit von E nachdem F beobachtet wurde

WOHER KOMMEN WAHRSCHEINLICHKEITEN? BAYESIANISCHER ANSATZ

HOCHSCHULE HEILBRONN

Beispiel. Bayesianischer Ansatz: Würfelwurf: $E = \{$⚁, ⚃$\}$ und $F = \{$⚁$\}$.

> Wahrscheinlichkeiten von E und F?

> Wahrscheinlichkeit von E unter der Bedingung, dass F eingetreten ist?

> Wahrscheinlichkeit von F unter der Bedingung, dass E eingetreten ist?

WOHER KOMMEN WAHRSCHEINLICHKEITEN? BAYESIANISCHER ANSATZ

HOCHSCHULE HEILBRONN

Beispiel. Bayesianischer Ansatz: Würfelwurf: $E = \{⚁, ⚃\}$ und $F = \{⚁\}$.

> Wahrscheinlichkeiten von E und F?

$$P(E) = 1/3 \text{ und } P(F) = 1/6$$

> Wahrscheinlichkeit von E unter der Bedingung, dass F eingetreten ist?

$$P(E \mid F) = 1$$

> Wahrscheinlichkeit von F unter der Bedingung, dass E eingetreten ist?

$$P(F \mid E) = 1/2$$

WOHER KOMMEN WAHRSCHEINLICHKEITEN? BAYESIANISCHER ANSATZ

HOCHSCHULE HEILBRONN

Bayes-Theorem und totale Wahrscheinlichkeit.

- Das zentrale Konzept des Bayesianischen Ansatzes ist das Bayes-Theorem
- Bayes-Theorem: Anpassung A-priori Wahrscheinlichkeiten im Lichte neuer Information
- Theorem beschreibt also den Übergang von A-priori zu A-posteriori Wahrscheinlichkeiten
- Anwendung Bayes-Theorem erfordert häufig den Satz der totalen Wahrscheinlichkeit

WOHER KOMMEN WAHRSCHEINLICHKEITEN? BAYESIANISCHER ANSATZ

HOCHSCHULE HEILBRONN

Beispiel aus der Einführung. Erinnerung

- Texas: 38% fahren einen SUV, 57% wählen die Republikaner
- Umfrage unter SUV-Fahrern: 78% sind Republikaner

Wir müssen folgende Wahrscheinlichkeiten unterscheiden:

- Bedingte W'keit Republikaner gegeben man ist SUV-Fahrer:

$$P(Republikaner \mid SUV) = 78\%$$ **Das misst die Studie**

- Bedingte W'keit SUV-Fahrer gegeben man ist Republikanter:

$$P(SUV \mid Republikaner) = ?$$ **Das wollte die Studie messen**

WOHER KOMMEN WAHRSCHEINLICHKEITEN? BAYESIANISCHER ANSATZ

HOCHSCHULE HEILBRONN

Beispiel aus der Einführung (Fortsetzung). **Bayes-Theorem.**

Bayes-Theorem

Die bedingte Wahrscheinlichkeit von Ereignis E nachdem F beobachtet wurde ist:

$$P(F \mid E) = \frac{P(E \mid F) \cdot P(F)}{P(E)} = \frac{P(E \cap F)}{P(E)}.$$

> Wir können nun dieses Theorem anwenden, um den Machern der Studie zu helfen:

WOHER KOMMEN WAHRSCHEINLICHKEITEN? BAYESIANISCHER ANSATZ

Beispiel aus der Einführung (Fortsetzung). **Bayes-Theorem.**

Bayes-Theorem

Die bedingte Wahrscheinlichkeit von Ereignis E nachdem F beobachtet wurde ist:

$$P(F \mid E) = \frac{P(E \mid F) \cdot P(F)}{P(E)} = \frac{P(E \cap F)}{P(E)}.$$

> Wir können nun dieses Theorem anwenden, um den Machern der Studie zu helfen:

$$\boldsymbol{P(SUV \mid Republikaner)} = \frac{P(Republikaner \mid SUV) \cdot P(SUV)}{P(Republikaner)} = \frac{78\% \cdot 38\%}{57\%} = \mathbf{52\%}$$

WOHER KOMMEN WAHRSCHEINLICHKEITEN? BAYESIANISCHER ANSATZ

Betrachten Sie einen Corona-Test der 90% sensitiv ist und 80% spezifisch.

> 90% sensitiv: 90% der Leute, die Corona haben, erhalten ein positives Testresultat

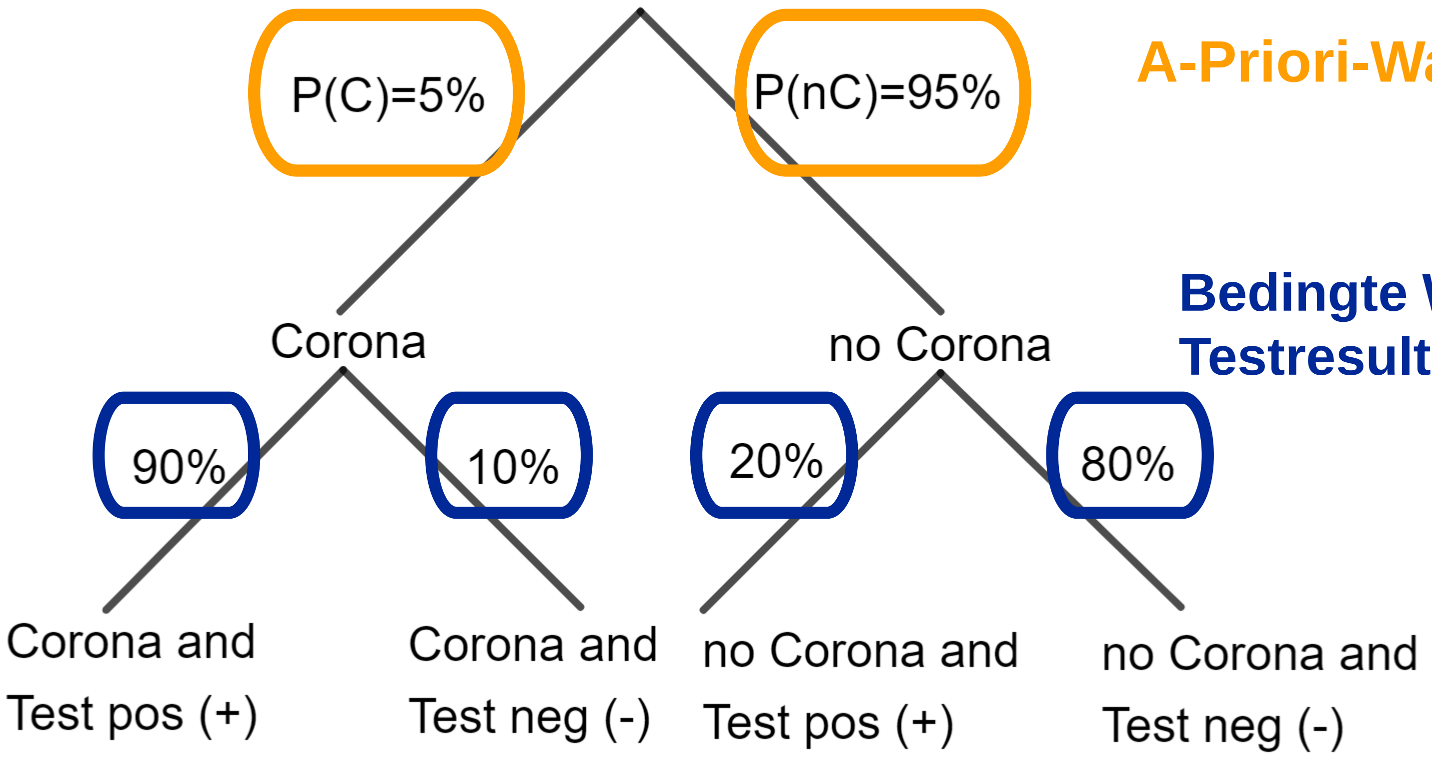
> 80% spezifisch: 80% der Leute, die gesund sind, erhalten ein negatives Testresultat

> Außerdem hat ca. 5% der Bevölkerung Corona

Frage: Wahrscheinlichkeit, dass man Corona hat bei positivem Testresultat?

WOHER KOMMEN WAHRSCHEINLICHKEITEN? BAYESIANISCHER ANSATZ

Beispiel (Fortsetzung). **Baumdiagramm**



A-Priori-Wahrscheinlichkeiten

Bedingte Wahrscheinlichkeiten der Testresultate gegeben Corona / kein Corona

WOHER KOMMEN WAHRSCHEINLICHKEITEN? BAYESIANISCHER ANSATZ

HOCHSCHULE HEILBRONN

Beispiel (Fortsetzung).

Wir müssen folgende Wahrscheinlichkeiten unterscheiden:

> Wahrscheinlichkeit einer Corona-Infektion (C):

$$P(C) = 5\%$$

> Bedingte Wahrscheinlichkeit eines positiven Tests (+) gegeben man hat Corona (C):

$$P(+ \mid C) = 90\%$$

> Bedingte Wahrscheinlichkeit einer Corona-Infektion (C) gegeben ein positiver Test (+):

$$P(C \mid +) = ?$$ Diese möchten wir wissen

WOHER KOMMEN WAHRSCHEINLICHKEITEN? BAYESIANISCHER ANSATZ

HOCHSCHULE HEILBRONN

Beispiel (Fortsetzung). **Anwendung Bayes-Theorem.**

> Wir möchten die bedingte Wahrscheinlichkeit $P(Corona \mid positiver\ Test)$ berechnen

> Bayes-Theorem:

$$P(Corona \mid +) = \frac{P(+ \mid Corona) \cdot P(Corona)}{P(+)}$$

HOCHSCHULE HEILBRONN

Exkurs. **Satz der totalen Wahrscheinlichkeit.**

Satz der totalen Wahrscheinlichkeit

Gegeben zwei Ereignisse E, F und deren Komplemente E^c, F^c. Die W'keit von E ist:

$$(Bed. W'keit\ E\ auf\ F) \times (W'keit\ F) + (Bed. W'keit\ E\ auf\ F^c) \times (W'keit\ F^c),$$

formal:

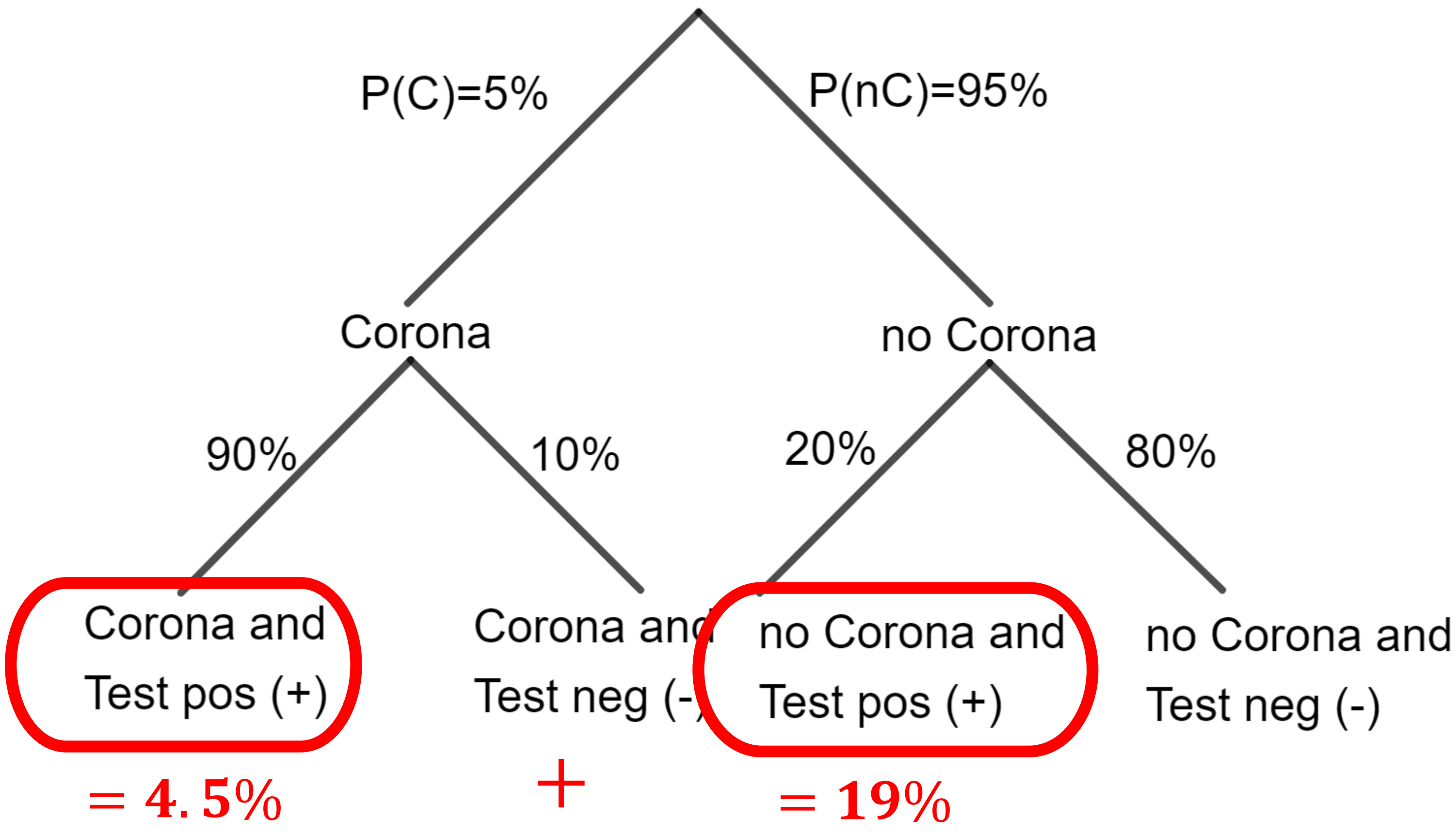
$$P(E) = P(E \mid F) \cdot P(F) + P(E \mid F^c) \cdot P(F^c) = P(E \cap F) + P(E \cap F^c).$$

WOHER KOMMEN WAHRSCHEINLICHKEITEN? BAYESIANISCHER ANSATZ

HOCHSCHULE HEILBRONN

Beispiel (Fortsetzung). **Anwendung Satz der totalen W‘keit zur Berechnung von P(+).**

> $P(+) = 4.5\% + 19\% = \mathbf{23.5\%}$



WOHER KOMMEN WAHRSCHEINLICHKEITEN? BAYESIANISCHER ANSATZ

HOCHSCHULE HEILBRONN

Beispiel (Fortsetzung). **Anwendung Bayes-Theorem.**

> Erinnerung: Wir möchten die bedingte W'ket $P(Corona \mid positive\ test)$ bestimmen

> Bayes-Theorem:

$$P(Corona \mid +) = \frac{\overbrace{\boxed{P(+ \mid Corona)}}^{90\%} \cdot \overbrace{\boxed{P(Corona)}}^{5\%}}{\underbrace{\boxed{P(+)}}_{23.5\%}} = 19.15\%$$

→ Die W'keit Corona zu haben bei einem positiven Test ist 19.15%

Beispiel. 200 Menschen. 47 (= 23.5%) werden positiv getestet → 10 (= 5%) haben Corona.

› ZUFALLS-VARIABLEN

WAS SIND ZUFALLSVARIABLEN?

HOCHSCHULE HEILBRONN

Meistens interessieren wir uns nicht für alle Details eines Zufallsexperiments.

Eine *Zufallsvariable*

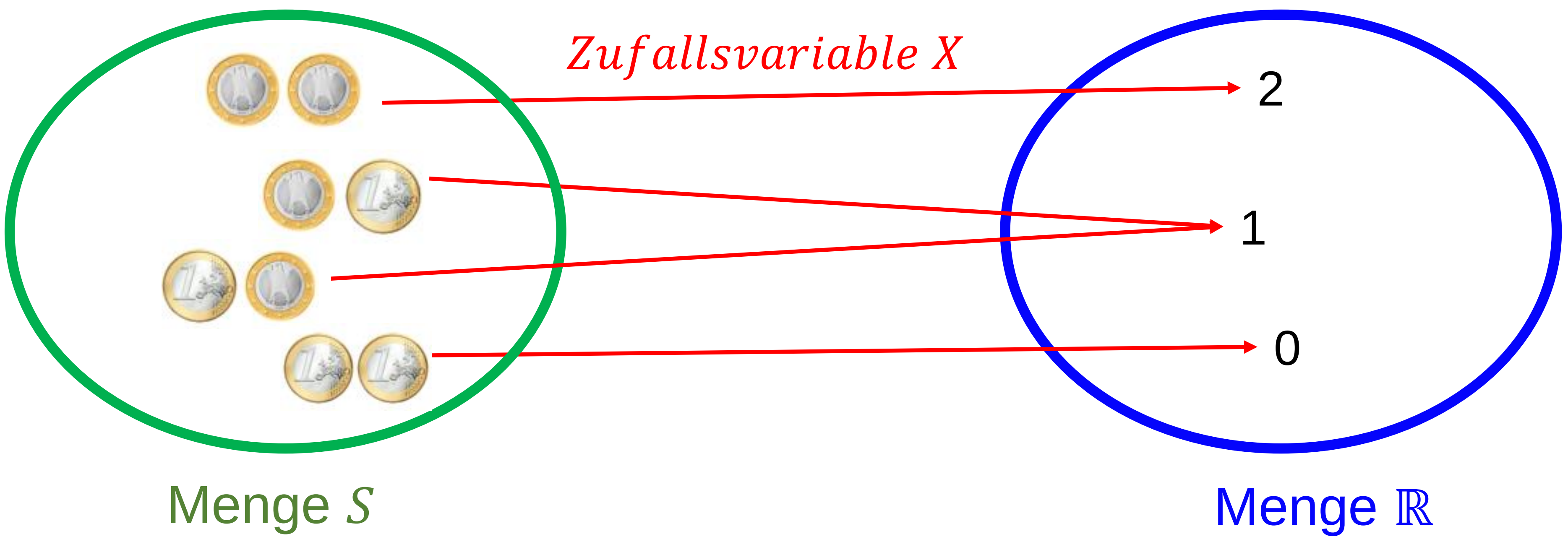
- hängt von den Ergebnissen eines Zufallsexperiments ab
- misst relevante Charakteristika des Experiments

Beispiel. Mehrfacher Münzwurf.

- Wie häufig wird "Kopf" geworfen?
- Zufallsvariable, die zählt wie häufig "Kopf" geworfen wurde

WAS SIND ZUFALLSVARIABLEN?

Beispiel. Zweifacher Münzwurf. Zufallsvariable: Funktion X = Anzahl "Kopf".



WAS SIND ZUFALLSVARIABLEN?

HOCHSCHULE HEILBRONN

Zufallsvariable

Eine *Zufallsvariable* X ist eine Funktion, die jedem Ergebnis $s \in S$ eines Zufalls-experiments ein Element s' einer Menge S' zuordnet (häufig eine Zahl: $S' = \mathbb{R}$):

$$X: S \to S'.$$

Beispiel. Zweimaliges Werfen einer Münze. X = Anzahl Kopf.

> X ist eine Stufenfunktion, die wie folgt geschrieben werden kann:

$$X(s) = \begin{cases} 2, & wenn\ s = HH \\ 1, & wenn\ s = HT\ oder\ TH \\ 0, & wenn\ s = TT \end{cases}$$

ZUFALLSVARIABLEN UND ZUFALLSEXPERIMENTE

Beispiel. Zweifacher Münzwurf.

Zufallsexperiment

$s \in S$	$P(s)$
	¼
	¼
	¼
	¼

X

Zufallsvariable X = Anzahl "Kopf"

$x \in S'$	$P'(X = x)$
2	¼
1	½
0	¼

ZUFALLSVARIABLEN UND ZUFALLSEXPERIMENTE

Das Konzept einer Zufallsvariablen = Natürliche Erweiterung Zufallsexperiment

$$Zufallsvariable\colon X\colon S \to S'$$

> Wahl $S' = S \to$ Ursprüngliches Zufallsexperiment

Ab jetzt beziehen wir uns auf Zufallsvariablen (statt auf Zufallsexperimente)

ZUFALLSVARIABLEN: ANMERKUNGEN

- Kombinationen von Zufalls:variablen

$$X + Y, X \cdot Y, \ldots$$

sind ebenfalls Zufallsvariablen

- Ebenso Kombinationen von Zufallsvariablen mit einer Konstanten

$$X + Konstante, X \cdot Konstante, \ldots$$

ZUFALLSVARIABLEN: NOTATION

Objekt	Notation
Zufallsvariable	$X, Y, Z, \dots\ (auch: X_1, X_2, \dots)$ (Großbuchstaben)
Platzhalter für einen Wert einer Zufallsvariablen	$x, y, z, \dots$ (Kleinbuchstaben)
Wahrscheinlichkeit, dass die Zufallsvariable X einen Wert x annimmt	$P(X = x)$ Beispiel: $P(X = 2)$ Kurz: $P(2)$
Wahrscheinlichkeit, dass die Zufallsvariable X in einer Menge M liegt	$P(X \in M)$ Beispiel: $P(X \in [145, 180])$ Kurz: $P([145, 180])$

ARTEN VON ZUFALLSVARIABLEN: DISKRET UND STETIG

HOCHSCHULE HEILBRONN

Diskrete Zufallsvariable

Nimmt eine **abzählbare** Anzahl (endlich oder unendlich) von Werten an.

Stetige Zufallsvariable

Nimmt eine **nicht abzählbare** unendliche Anzahl von Werten an.

Beispiel. Anzahl an Sechsen bei vierfachem Würfelwurf:

$$0, 1, 2, 3, 4$$

Beispiel. Größe einer zufällig ausgewählten Person:

$$145\ cm \leq X \leq 220\ cm$$

ZUFALLSVARIABLEN: DISKRET UND STETIG

Zufallsvariable	Art
Anzahl Verkäufe eines Verkäufers in einer Woche	
Die Größe eines zufällig ausgewählten Erwachsenen	
Die Zeit zwischen der Ankunft zweier Touristen an einem Urlaubsort	
Anzahl Kunden in einer Stichprobe, die ein bestimmtes Produkt gegenüber allen Wettbewerbern bevorzugen	
Neuer Wohnkomplex – die Zeit zwischen der Fertigstellung und dem Verkauf der letzten Wohnung	
Erwartete Anzahl Studierende in einer Vorlesung	
Die Tiefe, bis zu der gebohrt werden muss, bis man auf Öl stößt	

› WAHRSCHEINLICHKEITS-VERTEILUNGEN

DISKRETE UND STETIGE VERTEILUNGEN

- *Diskrete Verteilung* = Verteilung einer diskreten Zufallsvariablen

- *Stetige Verteilung* = Verteilung einer stetigen Zufallsvariablen

DISKRETE VERTEILUNGEN

HOCHSCHULE HEILBRONN

Eine diskrete Verteilung

- ist definiert durch eine *Wahrscheinlichkeitsfunktion* $P(X = x)$
- dies ist eine Liste aller Werte der Zufallsvariablen und Ihrer Wahrscheinlichkeiten

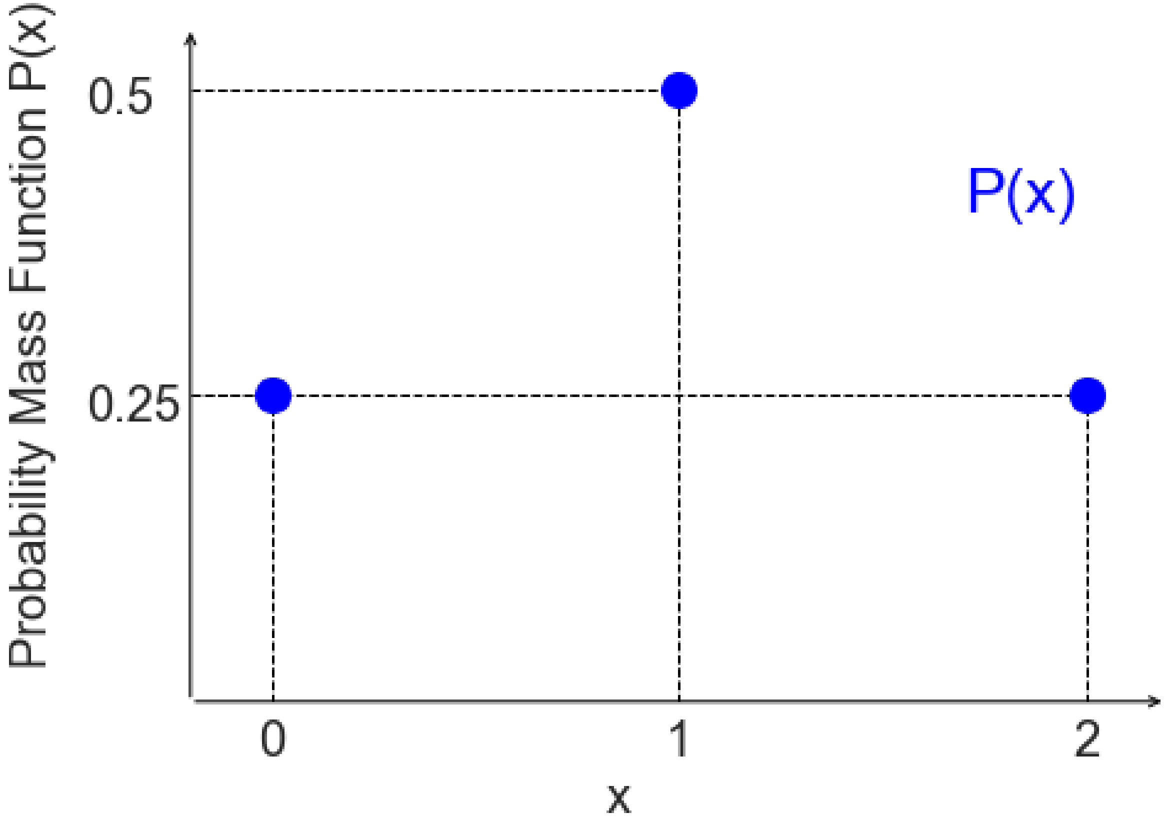
Beispiel. Zweifacher Münzwurf. X = Anzahl "Kopf". Wahrscheinlichkeitsverteilung:

x	$P(X = x)$
2	¼
1	½
0	¼

HOCHSCHULE HEILBRONN

Wahrscheinlichkeitsfunktion – graphisch:

x	$P(X = x)$
2	¼
1	½
0	¼



DISKRETE VERTEILUNGEN

HOCHSCHULE HEILBRONN

Berechnung von Wahrscheinlichkeiten mit der Wahrscheinlichkeitsfunktion

x	$P(X = x)$
2	¼
1	½
0	¼

> Was ist die Wahrscheinlichkeit, dass mindestens 1x "Kopf" geworden wird $P(X \geq 1)$:

AUSGEWÄHLTE DISKRETE VERTEILUNGEN: GLEICHVERTEILUNG

HOCHSCHULE HEILBRONN

Diskrete Gleichverteilung
Eine diskrete Zufallsvariable ist gleichverteilt, wenn alle ihre Werte gleichwahrscheinlich sind.

Beispiele.

AUSGEWÄHLTE DISKRETE VERTEILUNGEN: BERNOULLI-VERTEILUNG

HOCHSCHULE HEILBRONN

Bernoulli-Verteilung

Eine Zufallsvariable ist *Bernoulli-verteilt,* wenn sie nur zwei Werte hat (0 and 1). Wert 1 heißt *Erfolg* und Wert 0 *Misserfolg*.

Beispiel.

Die Wahrscheinlichkeitsfunktion der Bernoulli-Verteilung ist

$$P(x) = \begin{cases} \pi & für\ x = 1 \\ (1-\pi) & für\ x = 0 \end{cases},$$

> wobei $\pi =$ Erfolgswahrscheinlichkeit der einzige Parameter ist

AUSGEWÄHLTE DISKRETE VERTEILUNGEN: BINOMIALVERTEILUNG

HOCHSCHULE HEILBRONN

Bernoulli-Prozess = Folge von n unabhängig und identisch verteilten Bernoulli-Variablen

> Eigenschaften Bernoulli-Prozess:

1. Fixierte Länge n
2. Nur 2 mögliche Ergebnisse pro Versuch: Erfolg und Misserfolg (Bernoulli-Variable)
3. Erfolgsw'keit π ist in allen Versuchen identisch (unabhängig und identisch verteilt)

Binomialverteilung

Eine Zufallsvariable X ist *binomialverteilt,* wenn sie die Anzahl der Erfolge in einem Bernoulli-Prozess der Länge n zählt.

AUSGEWÄHLTE DISKRETE VERTEILUNGEN: BINOMIALVERTEILUNG

HOCHSCHULE HEILBRONN

Ist dies ein Bernoulli-Prozess?

> Ein kleines Reisebüro hat 20 Reisekataloge bestellt.

> Aufgrund schlechter Erfahrungen mit Fehldrucken möchte das Reisebüro 3 Kataloge auf Fehldruck prüfen, bevor es die Lieferung annimmt.

> 3 Kataloge werden sequentiell zufällig aus-gewählt und auf Fehldruck geprüft.

> Was das Reisebüro nicht weiß: 2 Kataloge sind Fehldrucke

AUSGEWÄHLTE DISKRETE VERTEILUNGEN: BINOMIALVERTEILUNG

HOCHSCHULE HEILBRONN

Beispiel. Fünfacher Münzwurf.

> Die Münze ist leicht unfair: W'keit Kopf = 2/5 (W'keit Zahl = 3/5)

> $X = Anzahl\ Kopf$

Was ist die Wahrscheinlichkeit 0-mal Kopf zu werfen $P(X = 0)$?

> 0-mal Kopf (H) = 5-mal Zahl (T): 1 Ereignis: $TTTTT$

$$P(X = 0) = \frac{3}{5} \cdot \frac{3}{5} \cdot \frac{3}{5} \cdot \frac{3}{5} \cdot \frac{3}{5} = \frac{3}{5}^5 = 1 \cdot \frac{2}{5}^0 \cdot \frac{3}{5}^5 \approx 7.78\%$$

AUSGEWÄHLTE DISKRETE VERTEILUNGEN: BINOMIALVERTEILUNG

HOCHSCHULE HEILBRONN

Beispiel (Fortsetzung). Fünfacher Münzwurf.

Was ist die Wahrscheinlichkeit exakt 1-mal Kopf zu werfen ($X = 1$)?

> 5 Ereignisse: $HTTTT, THTTT, TTHTT, TTTHT, TTTTH$

$$P(X = 1) = \frac{2}{5} \cdot \frac{3}{5} \cdot \frac{3}{5} \cdot \frac{3}{5} \cdot \frac{3}{5} + \cdots + \frac{3}{5} \cdot \frac{3}{5} \cdot \frac{3}{5} \cdot \frac{3}{5} \cdot \frac{2}{5} = 5 \cdot \frac{2}{5}^1 \cdot \frac{3}{5}^4 \approx 25.92\%$$

Beispiel (Fortsetzung). Fünfacher Münzwurf.

Was ist die Wahrscheinlichkeit exakt 2-mal Kopf zu werfen ($X = 2$)?

> 10 Ereignisse: $HHTTT, HTHTT, \ldots, TTTHH$

$$P(X = 2) = \frac{2}{5} \cdot \frac{2}{5} \cdot \frac{3}{5} \cdot \frac{3}{5} \cdot \frac{3}{5} + \cdots + \frac{3}{5} \cdot \frac{3}{5} \cdot \frac{3}{5} \cdot \frac{2}{5} \cdot \frac{2}{5} = 10 \cdot \frac{2}{5}^2 \cdot \frac{3}{5}^3 \approx 34.56\%$$

AUSGEWÄHLTE DISKRETE VERTEILUNGEN: BINOMIALVERTEILUNG

HOCHSCHULE HEILBRONN

Es existiert ein Muster!

Anzahl Erfolge (x)

Erfolgswahrscheinlichkeit (π)

$$P(X = 0) = 1 \cdot \frac{2}{5}^{0} \cdot \frac{3}{5}^{5}$$

Anzahl Misserfolge ($n - x$)

$$P(X = 1) = 5 \cdot \frac{2}{5}^{1} \cdot \frac{3}{5}^{4}$$

Misserfolgs-wahrscheinlichkeit ($1 - \pi$)

Anzahl Ereignisse

$$P(X = 2) = 10 \cdot \frac{2}{5}^{2} \cdot \frac{3}{5}^{3}$$

> Wie können wir die Anzahl der Ereignisse berechnen?

> Anzahl Möglichkeiten x Objekte (Erfolge) aus n Objekten (Versuche) zu ziehen:

$\boxed{= \textbf{Binomialkoeffizient}}$

AUSGEWÄHLTE DISKRETE VERTEILUNGEN: BINOMIALVERTEILUNG

HOCHSCHULE HEILBRONN

Binomialkoeffizient

Der *Binomialkoeffizient* $\binom{n}{x}$ (lies "*n* über x") bestimmt die Anzahl der Möglichkeiten *x* Objekte aus n Objekten ohne Reihenfolge zu ziehen:

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!}.$$

> wobei $n!$ (lies "*n* Fakultät") $= n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 2 \cdot 1$

Beispiel. 5 über 2:

$$\binom{5}{2} = \frac{5!}{2!\,3!} = \frac{5 \cdot 4 \cdot 3!}{2 \cdot 1 \cdot 3!} = \frac{5 \cdot 4}{2 \cdot 1} = 10$$

Wahrscheinlichkeitsfunktion Binomialverteilung

Die *Wahrscheinlichkeit, dass eine binomialverteilte Zufallsvariable* X den Wert x annimmt ist:

$$P_{\pi,n}(x) = Binomialkoeff. \times W'keit\ Erfolg^{Erfolge(x)} \times W'keit\ Misserfolg^{Misserfolge(n-x)},$$

formal:

$$P_{\pi,n}(x) = \binom{n}{x} \cdot \pi^x \cdot (1-\pi)^{n-x}.$$

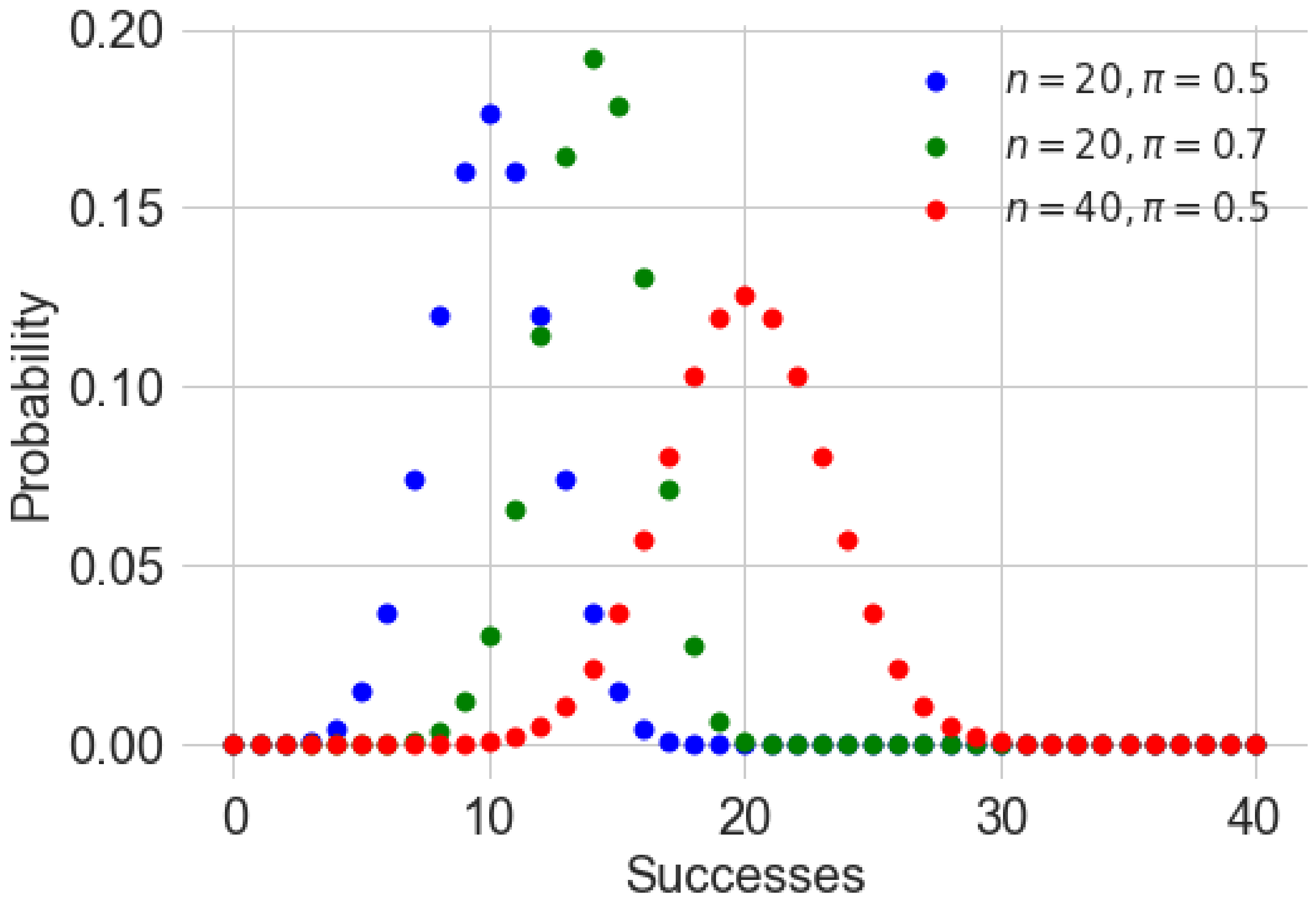
Anmerkung. Die Binomialverteilung hat zwei Parameter:

Erfolgswahrscheinlichkeit π und Länge des Bernoulli-Prozesses n

AUSGEWÄHLTE DISKRETE VERTEILUNGEN: BINOMIALVERTEILUNG

HOCHSCHULE HEILBRONN

Binomiale Wahrscheinlichkeitsfunktion für unterschiedliche Paramterkombinationen:



Berechnung binomialer Wahrscheinlichkeiten.

Beispiel. Einfluss von Frauen und Männern auf familiäre Kaufentscheidungen

> Studie: In 70% der Fällen haben Männer einen maßgeblichen Einfluss beim Autokauf

> 4 Familien sind im Begriff ein neues Auto zu kaufen

a) Was sind die Parameter dieser Binomialverteilung?

AUSGEWÄHLTE DISKRETE VERTEILUNGEN: BINOMIALVERTEILUNG

HOCHSCHULE HEILBRONN

Beispiel (Fortsetzung).

b) Wahrscheinlichkeit, dass der Mann in **genau** zwei Fällen die Entscheidung bestimmt?

c) W'keit, dass der Mann in höchstens 3 Familien die Kaufentscheidung bestimmt?

Hausaufgabe.

d) W'keit, dass der Mann in **mindestens** 2 Familien die Kaufentscheidung bestimmt?

Lösung: $P(X \geq 2) = 91.63\%$

STETIGE VERTEILUNGEN

Bei stetigen Verteilungen können wir keine Wahrscheinlichkeitsfunktion benutzen:

- Unendlich viele Werte → die Liste wäre unendlich lange
- Jeder einzelne Wert hätte eine Wahrscheinlichkeit von 0

Beispiel. X = Größe ($145\ cm \leq X \leq 220\ cm$) eine zufällig ausgewählten Person.

- Vereinfachende Annahme: Alle Größen sind gleichwahrscheinlich

Wahrscheinlichkeit, dass Person genau 180 cm groß ist?

- Anzahl günstige Fälle für 180 cm: 1, Gesamtanzahl Fälle: ∞ (Intervall: $[145{,}220]$)

Lösung: Wahrscheinlichkeit von Bereichen (Intervallen)

- Ist definiert durch eine Wahrscheinlichkeitsdichtefunktion $f(x)$
- W'keitsdichten identifizieren Regionen mit höherer bzw. niedrigerer W'keit:

  $f(2) > f(8)$: Wahrscheinlichkeit, dass Zufallsvariable Wert nahe 2 animmt ist höher

- Wahrscheinlichkeitsdichten $\neq$ W'keiten! Es ist z.B. möglich, dass $f(x) > 1$

Die Wahrscheinlichkeit eines Bereichs ist die Fläche unter der Dichtefunktion

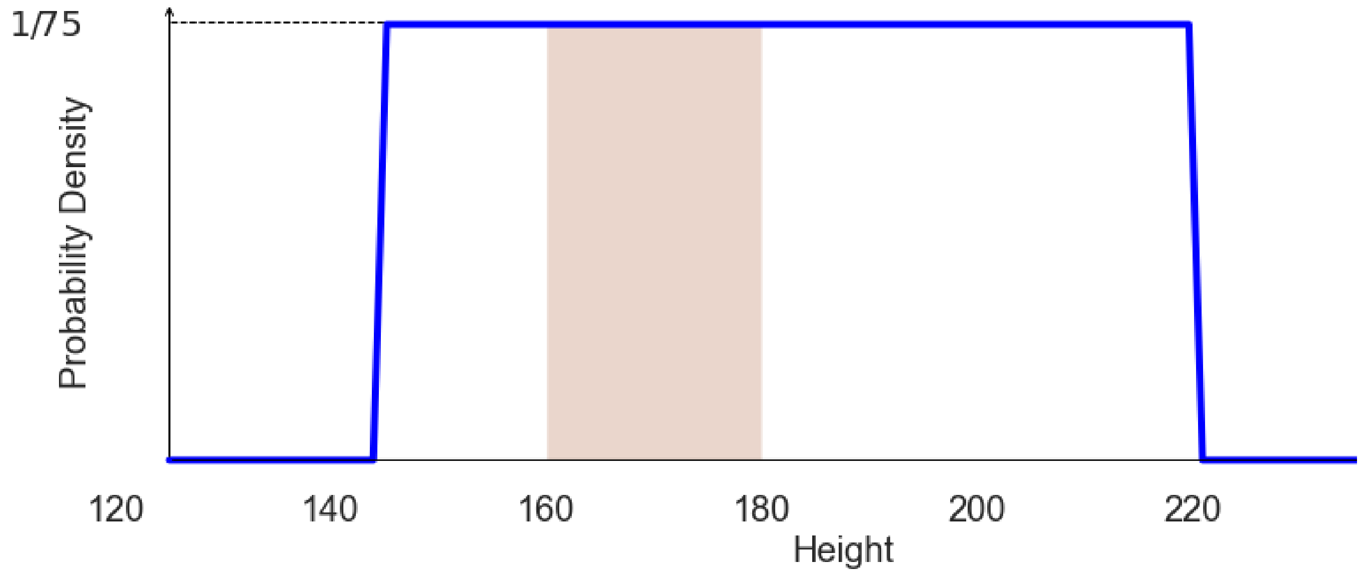
STETIGE VERTEILUNGEN

HOCHSCHULE HEILBRONN

Beispiel. X = Größe ($145\ cm \leq X \leq 220\ cm$) eine zufällig ausgewählten Person.

> Vereinfachende Annahme: Alle Größen sind gleichwahrscheinlich

> Dichtefunktion: $f(x) = \frac{1}{75}$ für alle Größen x



> P(Größe zwischen 160 und 180)?

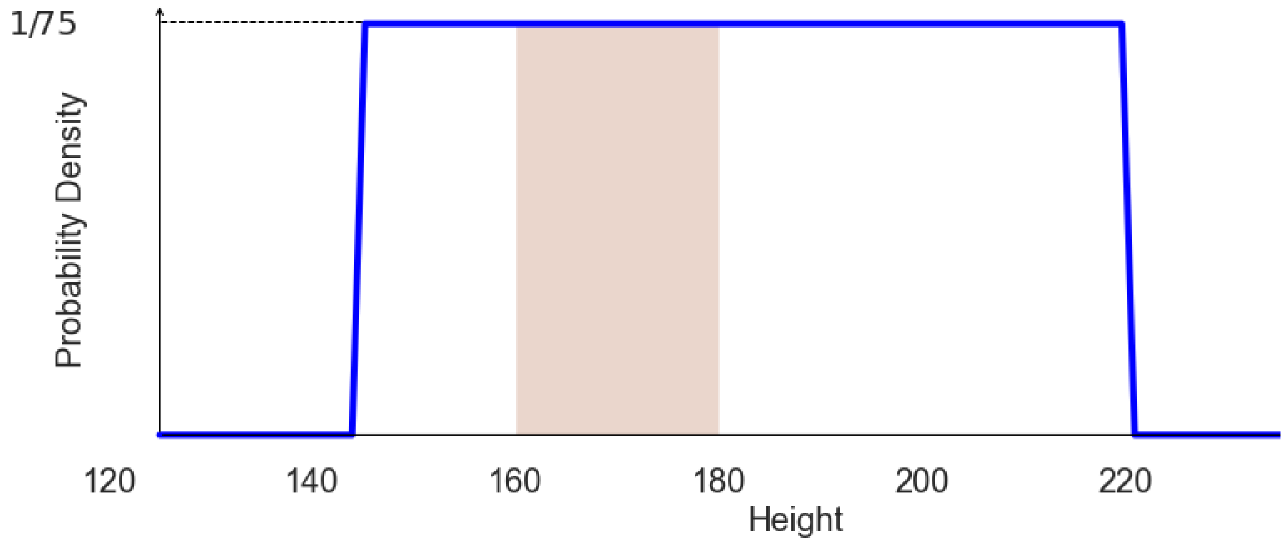
AUSGEWÄHLTE STETIGE VERTEILUNGEN: GLEICHVERTEILUNG

HOCHSCHULE HEILBRONN

Stetige Gleichverteilung

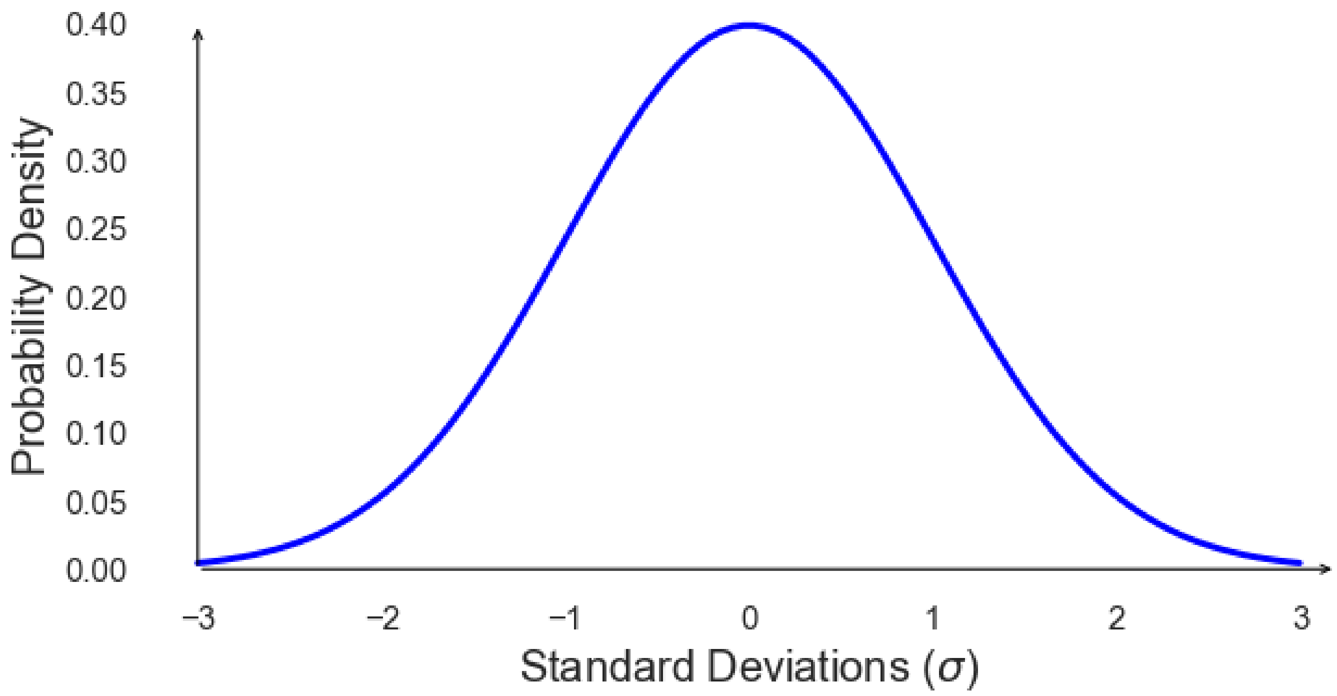
Eine stetige Zufallsvariable X ist gleichverteilt auf einem Intervall, wenn alle Werte x des Intervalls die gleiche Wahrscheinlichkeitsdichte $f(x)$ haben.

Beispiel. Siehe vorherige Folie



AUSGEWÄHLTE STETIGE VERTEILUNGEN: NORMALVERTEILUNG

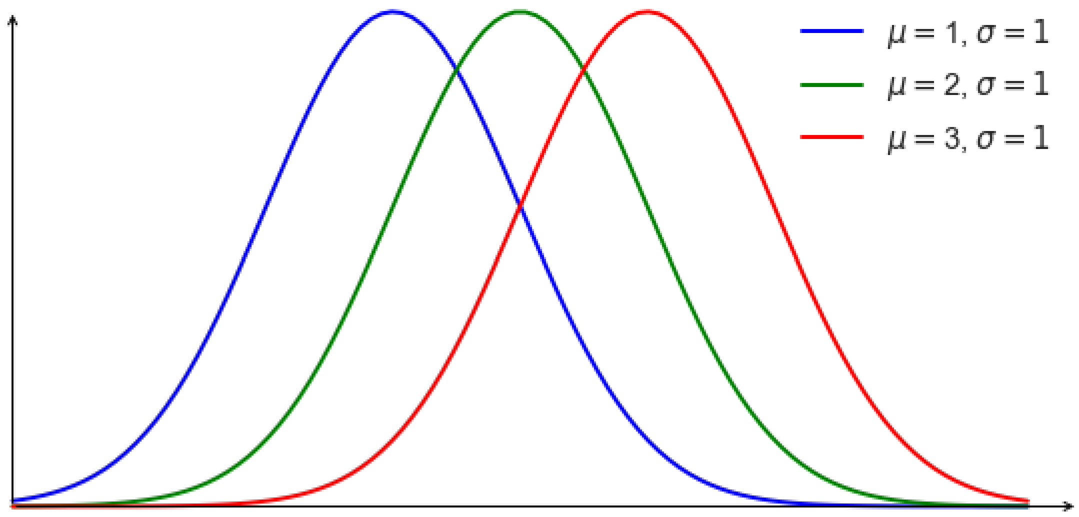
In der Realität haben die meisten Verteilungen eine Glockenform:

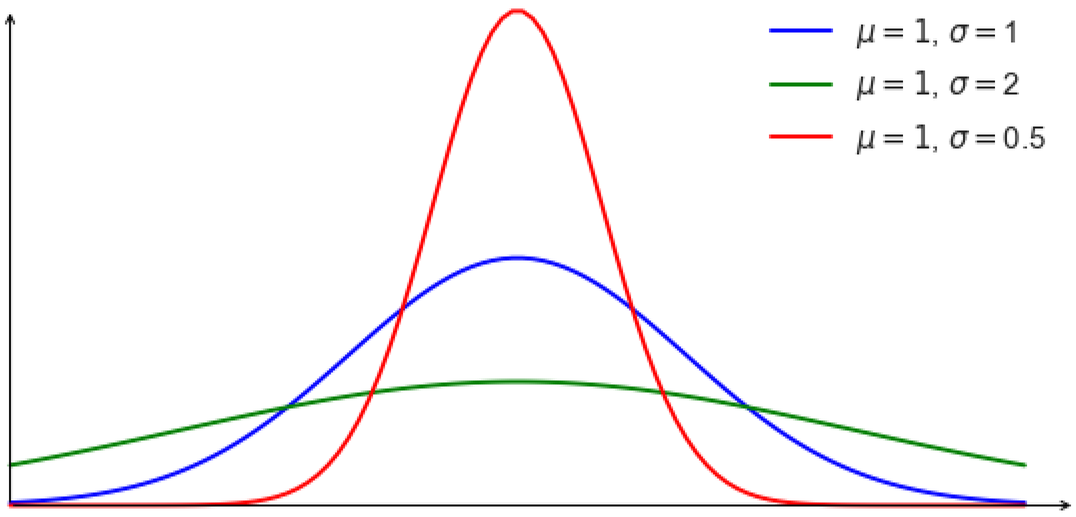


- > Derartige Verteilungen heißen *Normal*- oder *Gaußverteilung*
- > Glockenform: Extremwerte (Ränder) sind unwahrscheinlicher als mittlere Werte

AUSGEWÄHLTE STETIGE VERTEILUNGEN: NORMALVERTEILUNG

Parameter der Normalverteilung. Erwartungswert μ und Standardabweichung σ:



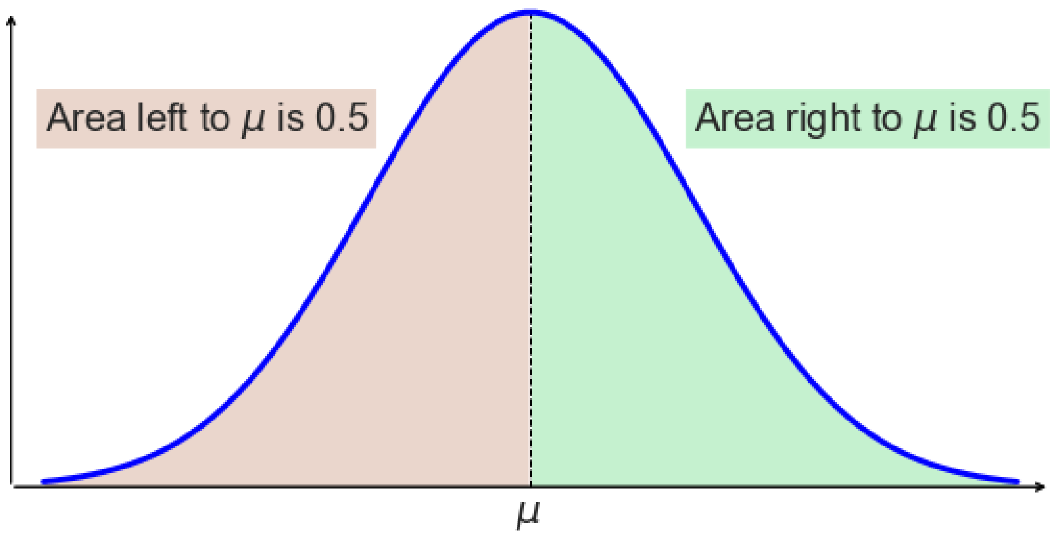


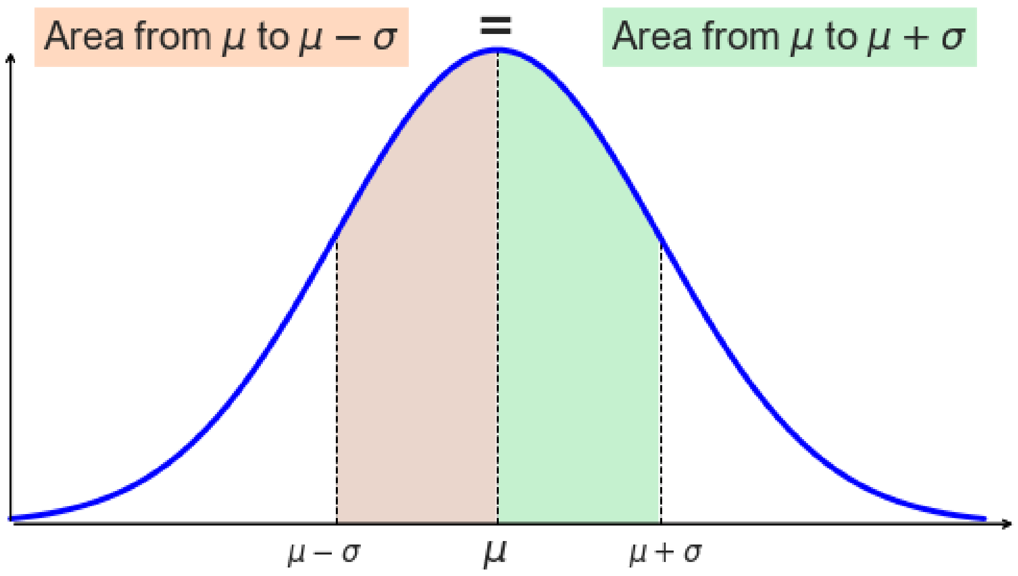
AUSGEWÄHLTE STETIGE VERTEILUNGEN: NORMALVERTEILUNG

HOCHSCHULE HEILBRONN

Symmetrie der Normalverteilung.

> Die Normalverteilung ist perfekt symmetrisch um Ihren Erwartungswert μ:





AUSGEWÄHLTE STETIGE VERTEILUNGEN: NORMALVERTEILUNG

HOCHSCHULE HEILBRONN

Normalverteilung

Eine Zufallsvariable X heißt *normalverteilt*, wenn sie folgende Dichtefunktion hat:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

Wenn $\mu = 0$ und $\sigma = 1$, dann heißt die Zufallsvariable standardnormalverteilt.

wobei

- e ($\approx$ 2.7183) und π ($\approx$ 3.1416) mathematische Konstanten sind
- μ der Erwartungswert und σ die Standardabweichung der Normalverteilung ist

AUSGEWÄHLTE STETIGE VERTEILUNGEN: NORMALVERTEILUNG

HOCHSCHULE HEILBRONN

Berechnung von Wahrscheinlichkeiten.

> Im Allgemeinen: Integrale bezüglich Dichtefunktion $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

=

Gute Nachrichten! Wir können Integralrechnung vermeiden:

> Tabelle mit Wahrscheinlichkeiten (Flächen) für Standardnormalverteilung

Z-Transformation: Konvertiert Normalverteilung in Standardnormalvereilung

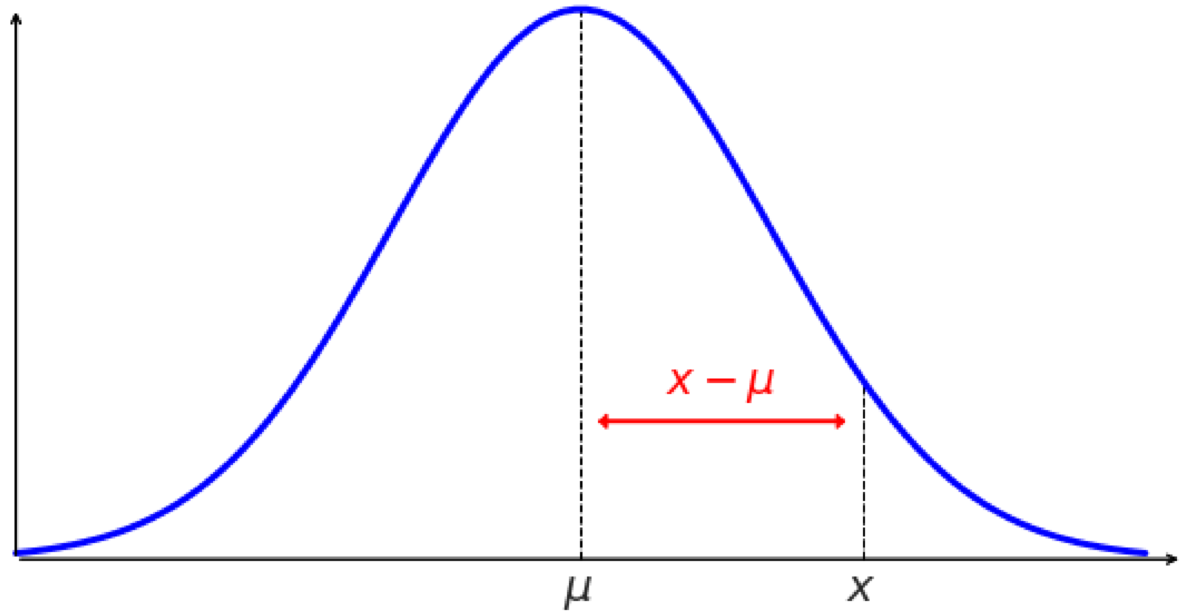
AUSGEWÄHLTE STETIGE VERTEILUNGEN: NORMALVERTEILUNG

Z-Transformation

Gegeben eine Zufallsvariable X mit Erwartungswert μ und Standardabweichung σ. Der z-Wert eines Wertes x von X ist:

$$z(x) = \frac{x-\mu}{\sigma}.$$

- $z(x)$ = Distanz zwischen μ und x in Standardabweichungen σ
- Zufallsvariable $Z = \frac{X-\mu}{\sigma}$ ist standard-normalverteilt, wenn X normalverteilt ist



AUSGEWÄHLTE STETIGE VERTEILUNGEN: NORMALVERTEILUNG

HOCHSCHULE HEILBRONN

Beispiel. Z-Transformation.

Wahrscheinlichkeitsverteilung:

x	p	$p \cdot x$
2	0.1	0.2
4	0.2	0.8
6	0.4	2.4
8	0.2	1.6
10	0.1	1
Summe		$\boldsymbol{\mu_x = 6}$
		$\boldsymbol{\sigma_X = 2.19}$

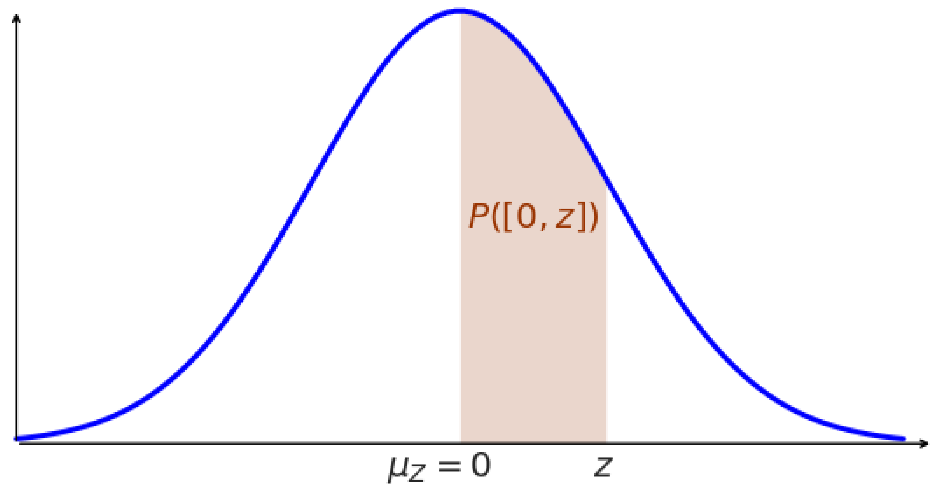
Z-transformierte Wahrscheinlichkeitsverteilung:

$z(x)$	p	$p \cdot z(x)$
-1.83	0.1	-0.18
-0.91	0.2	-0.18
0	0.4	0
0.91	0.2	0.18
1.83	0.1	0.18
Summe		$\boldsymbol{\mu_Z = 0}$
		$\boldsymbol{\sigma_Z = 1}$

$$z(x) = \frac{x - \mu_x}{\sigma_x}$$

AUSGEWÄHLTE STETIGE VERTEILUNGEN: NORMALVERTEILUNG

Z-Tabelle: Wahrscheinlichkeit $P([0,z])$



Z	0.00	0.01	0.02	0.03	0.04
0.00	0.0000	0.0040	0.0080	0.0120	0.0160
0.10	0.0398	0.0438	0.0478	0.0517	0.0557
0.20	0.0793	0.0832	0.0871	0.0910	0.0948
0.30	0.1179	0.1217	0.1255	0.1293	0.1331
0.40	0.1554	0.1591	0.1628	0.1664	0.1700
0.50	0.1915	0.1950	0.1985	0.2019	0.2054
0.60	0.2257	0.2291	0.2324	0.2357	0.2389
0.70	0.2580	0.2611	0.2642	0.2673	0.2704
0.80	0.2881	0.2910	0.2939	0.2967	0.2995
0.90	0.3159	0.3186	0.3212	0.3238	0.3264
1.00	0.3413	0.3438	0.3461	0.3485	0.3508
1.10	0.3643	0.3665	0.3686	0.3708	0.3729
1.20	0.3849	0.3869	0.3888	0.3907	0.3925
1.30	0.4032	0.4049	0.4066	0.4082	0.4099
1.40	0.4192	0.4207	0.4222	0.4236	0.4251
1.50	0.4332	0.4345	0.4357	0.4370	0.4382

> Z-Wert bis 1. Nachkommastelle

> 2. Nachkommastelle

Beispiel. $P([0,1.22])?$ $= 0.3888$

AUSGEWÄHLTE STETIGE VERTEILUNGEN: NORMALVERTEILUNG

HOCHSCHULE HEILBRONN

Berechnung Wahrscheinlichkeiten. Einzelwahrscheinlichkeiten.

Beispiel. Eine Person wird zufällig ausgewählt. $X = Größe$.

> Größe ist normalverteilt mit $\mu_X = 167cm$ und $\sigma_X = 10cm$.

a) Berechne die W'keit, dass die Person zwischen 167 und 178cm ist. $P_X([167,178])$

b) Wahrscheinlichkeit, dass X zwischen 150 und 181 cm ist?: $P_X([150,181])$

c) Bestimme die Wahrscheinlichkeit, dass die Person größer als 175 ist: $P_X(X > 175)$

d) Wahrscheinlichkeit, dass X zwischen 170 und 175 cm ist?: $P_X([170,175])$

AUSGEWÄHLTE STETIGE VERTEILUNGEN: NORMALVERTEILUNG

HOCHSCHULE HEILBRONN

Berechnung Wahrscheinlichkeiten. Vergleich von Gruppen.

Beispiel. Aggresivitäts-Score. $X = Score\ Frauen, Y = Score\ Männer$

- X ist normalverteilt mit $\mu_X = 38.82$ und $\sigma_X = 7.91$ (*)
- Y ist normalverteilt mit $\mu_Y = 40.86$ und $\sigma_Y = 8.69$ (*)

(*) Siehe "Gender Differences in Aggression-related Responses on EEG and ECG" in *Exp Neurobiol.* 27

Eine Frau und ein Mann werden zufällig ausgewählt.

- Wie wahrscheinlich ist es, dass der Mann aggressiver ist als die Frau?

ABSCHÄTZUNG STETIGER VERTEILUNGEN: EMPIRISCHE REGEL (NORMALVERTEILUNG)

Empirische Regel: Abschätzung der Normalverteilung

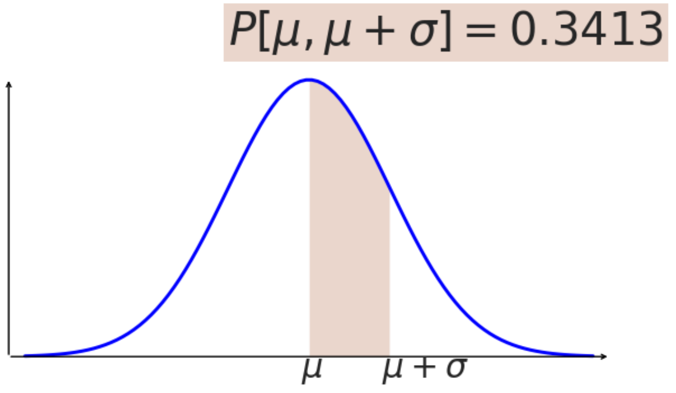
Wahrscheinlichkeit, dass eine normalverteilte Variable im Intervall

Mittelwert (μ) bis μ +1 Standardabweichung? +2 St.abw.? +3 St.abw.?

> Z-Werte für $[\mu, \mu + \sigma]$?

$$z(\mu) = 0$$

$$z(\mu + \sigma) = \frac{\mu + \sigma - \mu}{\sigma} = 1$$



ABSCHÄTZUNG STETIGER VERTEILUNGEN: EMPIRISCHE REGEL (NORMALVERTEILUNG)

HOCHSCHULE HEILBRONN

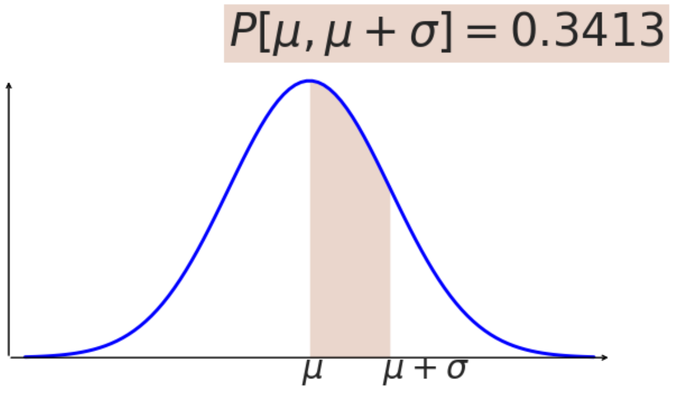
Empirische Regel: Abschätzung der Normalverteilung

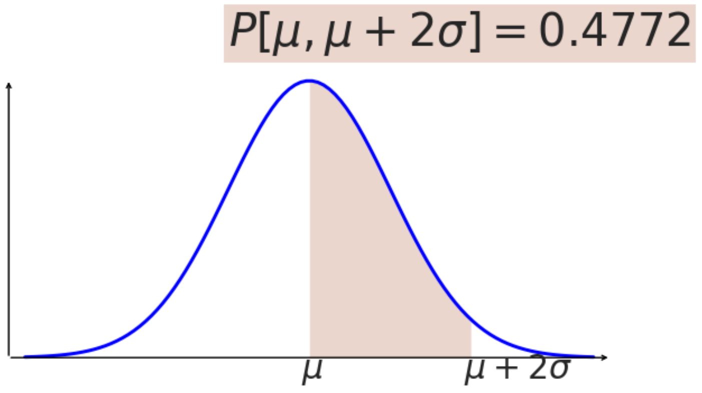
Analog:

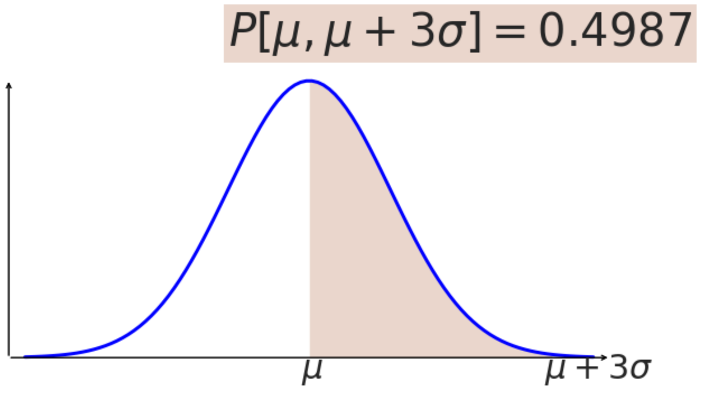
1.00	0.3413

2.00	0.4772

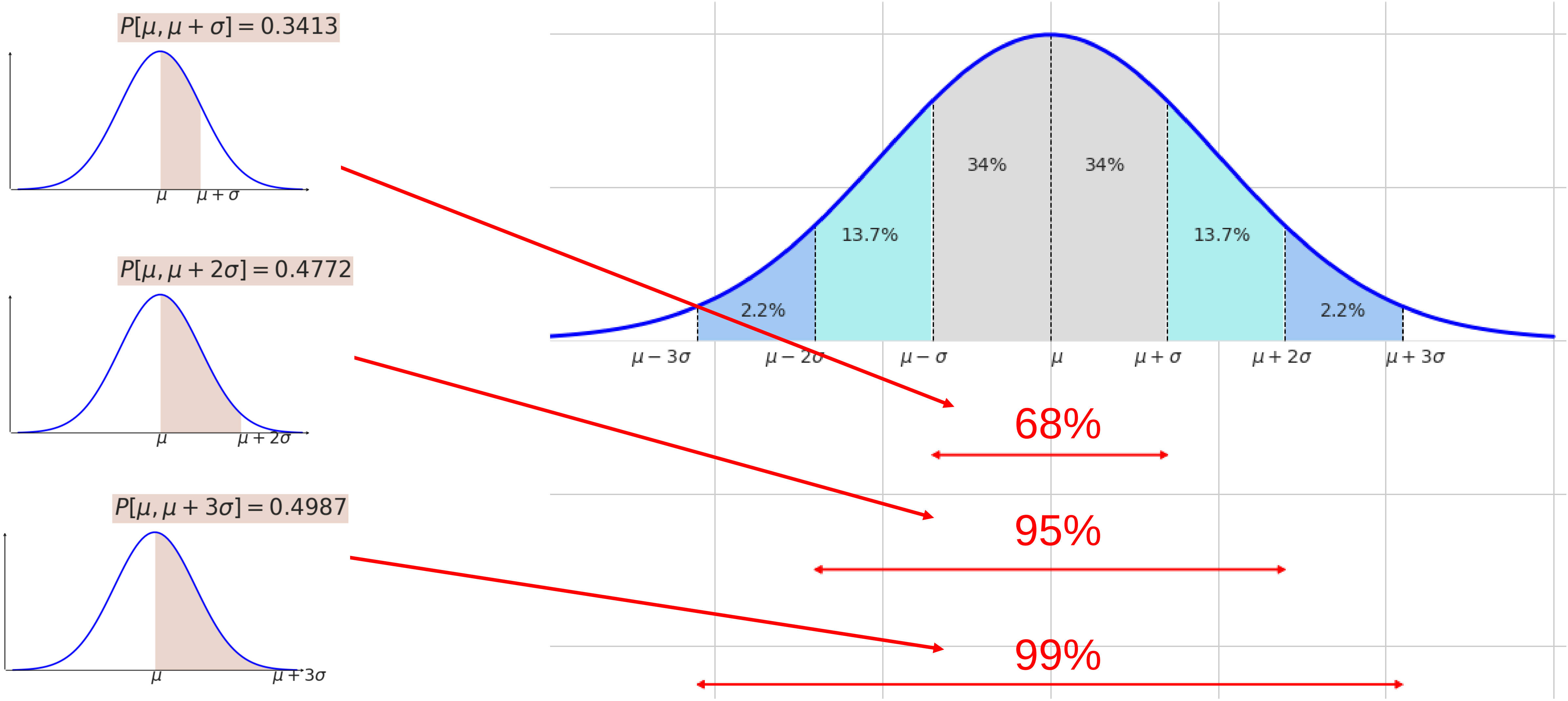
3.00	0.4987







ABSCHÄTZUNG STETIGER VERTEILUNGEN: EMPIRISCHE REGEL (NORMALVERTEILUNG)



ABSCHÄTZUNG STETIGER VERTEILUNGEN: EMPIRISCHE REGEL (NORMALVERTEILUNG)

HOCHSCHULE HEILBRONN

Zusammenfassung Empirische Regel. Gegeben eine Normalverteilung:

- **Ca. 68%** der Daten sind höchsten **1 Standardabweichung** vom Mittelwert entfernt:

$$P([\mu - \sigma, \mu + \sigma]) \approx 68\%$$

- **Ca. 95%** der Daten sind höchsten **2 Standardabweichungen** vom Mittelwert entfernt:

$$P([\mu - 2\sigma, \mu + 2\sigma]) \approx 95\%$$

- **Ca. 99%** der Daten sind höchsten **3 Standardabweichungen** vom Mittelwert entfernt:

$$P([\mu - 3\sigma, \mu + 3\sigma]) \approx 99\%$$

ABSCHÄTZUNG STETIGER VERTEILUNGEN: CHEBYSHEVS UNGLEICHUNG

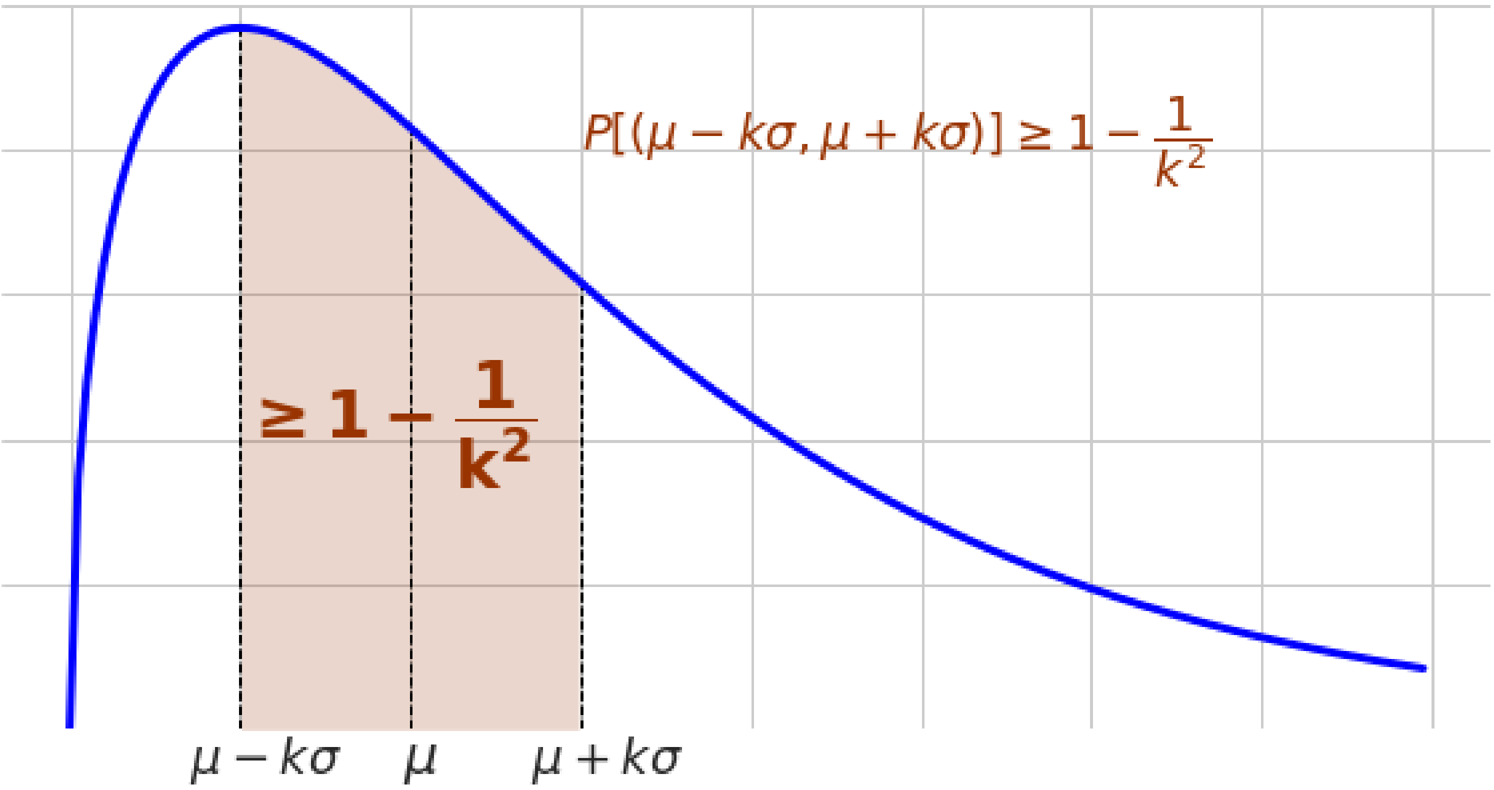
HOCHSCHULE HEILBRONN

> Die empirische Regel gilt nur für die Normalverteilung

> Gibt es eine Regel für andere Verteilungen

Ja! Chebyshevs Ungleichung:

> k = positive reelle Zahl



ABSCHÄTZUNG STETIGER VERTEILUNGEN: CHEBYSHEVS UNGLEICHUNG

HOCHSCHULE HEILBRONN

Anwendung von Chebyshevs Ungleichung

> k = Anzahl Standardabweichungen, die die Daten höchsten vom Mittelwert entfernt sind

k	Wahrscheinlichkeit
1	$\geq 0\%$
2	$\geq 75\%$
3	$\geq 88.89\%$

ABSCHÄTZUNG STETIGER VERTEILUNGEN: CHEBYSHEVS UNGLEICHUNG

HOCHSCHULE HEILBRONN

Beispiel. Nahrungsmittelunternehmen geben im Schnitt 135 € (Standardabweichung: 16 €) für Ads pro Internet-Nutzer aus. Die Art der Verteilung ist unbekannt.

a) Bestimmen Sie die Mindestwahrscheinlichkeit, dass eine zufällig ausgewählte Firma zwischen 115 € und 155 € ausgibt.

b) Bestimmen Sie die Ausgabenspanne (Intervall) in welche mindestens 50% der Unternehmen fallen

› MULTIVARIATE VERTEILUNG UND STOCHASTISCHE UNABHÄNGIGKEIT

MULTIVARIATE VERTEILUNG: GEMEINSAME VERTEILUNG ZWEIER VARIABLEN

HOCHSCHULE HEILBRONN

- Gemeinsame Wahrscheinlichkeitsfunktion $P_{X,Y}(x, y)$ von 2 Zufallsvariablen X und Y
- Gibt die gemeinsamen Wahrscheinlichkeiten an: $P(X = x, Y = y)$

Beispiel. Zweimaliger Münzwurf. $X = Anzahl\ Kopf\ in\ Wurf$ 1. $Y = Anzahl\ Kopf\ in\ Wurf$ 2.

		X = 1	X = 0	Randverteilung Y P_Y
Y	1	0.25	0.25	0.5
Y	0	0.25	0.25	0.5
Randverteilung X P_X		0.5	0.5	1

MULTIVARIATE VERTEILUNG: GEMEINSAME VERTEILUNG ZWEIER VARIABLEN

Beispiel. Statistik-Klausur. Zufallsvariablen:

> $X = \{Bestanden, Nicht\ Bestanden\}$ und $Y = \{Viel\ gelernt, Wenig\}$

		X		Randverteilung Y P_Y
		Bestanden	Nicht	
Y	Viel	0.4	0.1	0.5
	Wenig	0.2	0.3	0.5
Randverteilung X P_X		0.6	0.4	1

STOCHASTISCHE UNABHÄNGIGKEIT

Zwei Zufallsvariablen sind stochastisch unabhängig, wenn

> die Realisierung einer Zufallsvariable keinen Einfluss auf die W‘keiten der anderen hat

Beispiel. Zweifacher Münzwurf. Zufallsvariablen

$$X = Anzahl\ Kopf\ in\ Wurf\ 1$$

$$Y = Anzahl\ Kopf\ in\ Wurf\ 2$$

$$Z = Anzahl\ Zahl\ in\ beiden\ Würfen$$

> Wir beobachten $X = 1$

> Was sind nun die Wahrscheinlichkeiten $P(Y = 1)$ und $P(Z = 2)$?

STOCHASTISCHE UNABHÄNGIGKEIT

HOCHSCHULE HEILBRONN

Eine Zufallsvariable X ist unabhängig von Y, wenn bedingte W‘keit = unbedingte W‘keit:

$$P(X = x \mid Y = y) = P(X = x) \text{ für } alle\ x, y.$$

> Außerdem gemäß Bayes-Theorem:

$$P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

> Zusammen:

Stochastische Unabhängigkeit

Zwei Zufallsvariablen X und Y sind *stochastisch unabhängig*, wenn

$$P(X = x, Y = y) = P(X = x) \times P(Y = y) \text{ für } alle\ x, y.$$

STOCHASTISCHE UNABHÄNGIGKEIT: BEISPIEL

HOCHSCHULE HEILBRONN

Beispiel. Zweimaliger Münzwurf.

		X: 1	X: 0	P_Y
Y	1	0.25	0.25	0.5
	0	0.25	0.25	0.5
P_X		0.5	0.5	1

Stochastisch **Unabhängig**:

$$P_{X,Y}(x,y) = P_X(x) \cdot P_Y(y) = 0.5 \cdot 0.5 = 0.25$$

für alle $x, y \in \{0,1\}$

Beispiel. Klausur.

		X: Bestanden (B)	X: Nicht	P_Y
Y	Viel (V)	0.4	0.1	0.5
	Wenig	0.2	0.3	0.5
P_X		0.6	0.4	1

Stochastisch **Abhängig**:

$$P_{X,Y}(B,V) = 0.4 \neq P_X(B) \cdot P_Y(V) = 0.6 \cdot 0.5 = 0.3$$

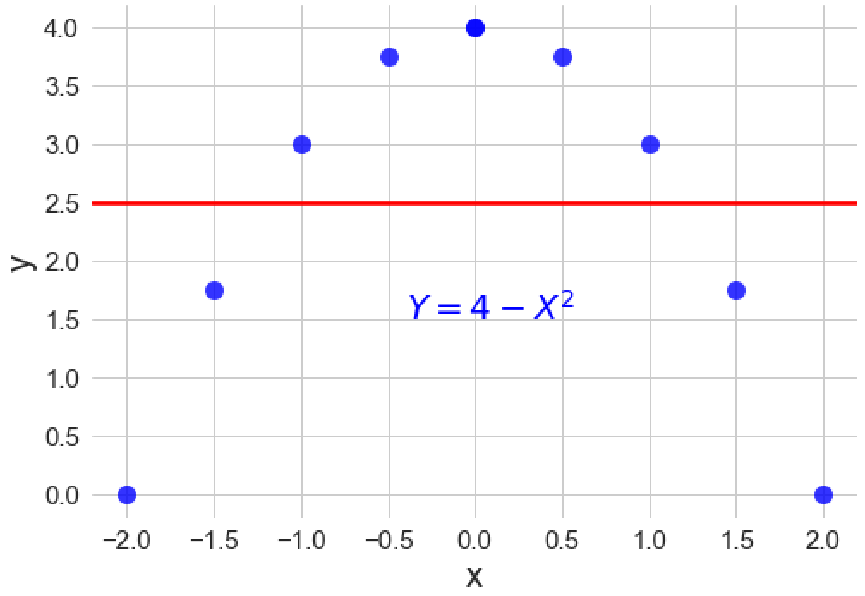
STOCHASTISCHE UNABHÄNGIGKEIT UND KORRELATION

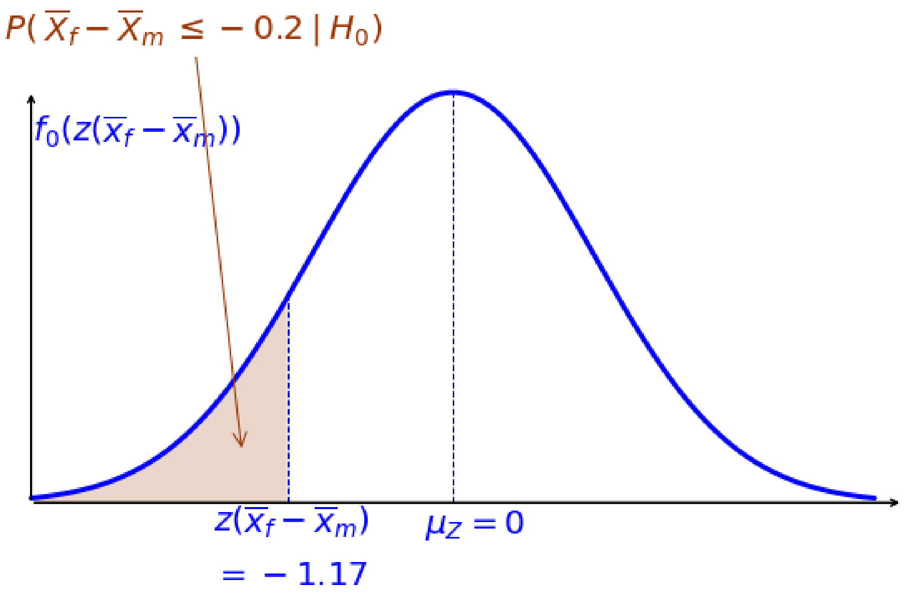
HOCHSCHULE HEILBRONN

- Zwei Zufallsvariablen X und Y sind korreliert, wenn deren Kovarianz ungleich 0 ist
- **Korrelierte** Zufallsvariablen sind immer **stochastisch abhängig**
- **Stochastisch abhängige** Zufallsvariablen können aber **unkorreliert sind**

Beispiel.

- X und Y sind perfekt abhängig aber unkorreliert
- Erinnerung: Korrelation = Lineare Beziehung
- Nicht alle Beziehungen sind linear!





› INFERENZSTATISTIK

TEILGEBIETE DER STATISTIK

HOCHSCHULE HEILBRONN

Statistik

Deskriptive Statistik

> Beschreibung von Daten

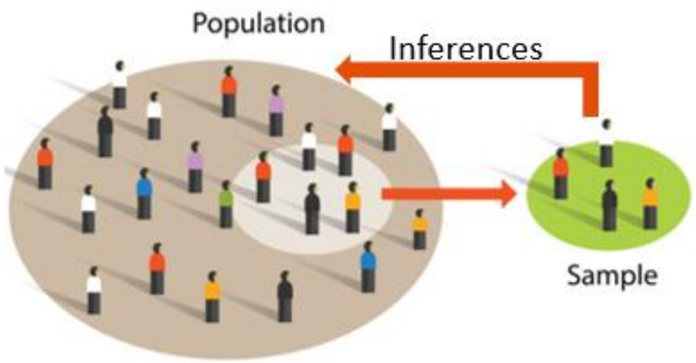
Inferenzstatistik

> Induktive Rückschlüsse basierend auf Daten

WAS IST INFERENZSTATISTIK?

HOCHSCHULE HEILBRONN

> Inferenzstatistik: Rückschlüsse von Stichproben auf die Population



Beispiel. In einer Stichprobe verdienen Frauen im Schnitt 211 € mehr als Männer.

> Können wir daraus schließen, dass Frauen im Allgemeinen mehr verdienen?

Solche und ähnliche Fragen werden mit inferenzstatistischen Methoden beantwortet.

› STATISTISCHE SCHÄTZER

WAS IST EIN SCHÄTZER?

Täglich werden Entscheidungen basierend auf Daten getroffen:

- Die Regierung möchte Güterströme von Auslandsmärkten vorhersagen
- Aktienhändler möchten Aktienentwicklungen vorhersagen
- Konsumenten möchten eine Einschätzung über Güter erlangen

Im Allgemeinen sind die Parameter dieser Verteilungen ($\mu, \sigma, \pi, \ldots$) unbekannt

Schätzer = Regel zur Schätzung eines Parameters basierend auf einer Stichprobe

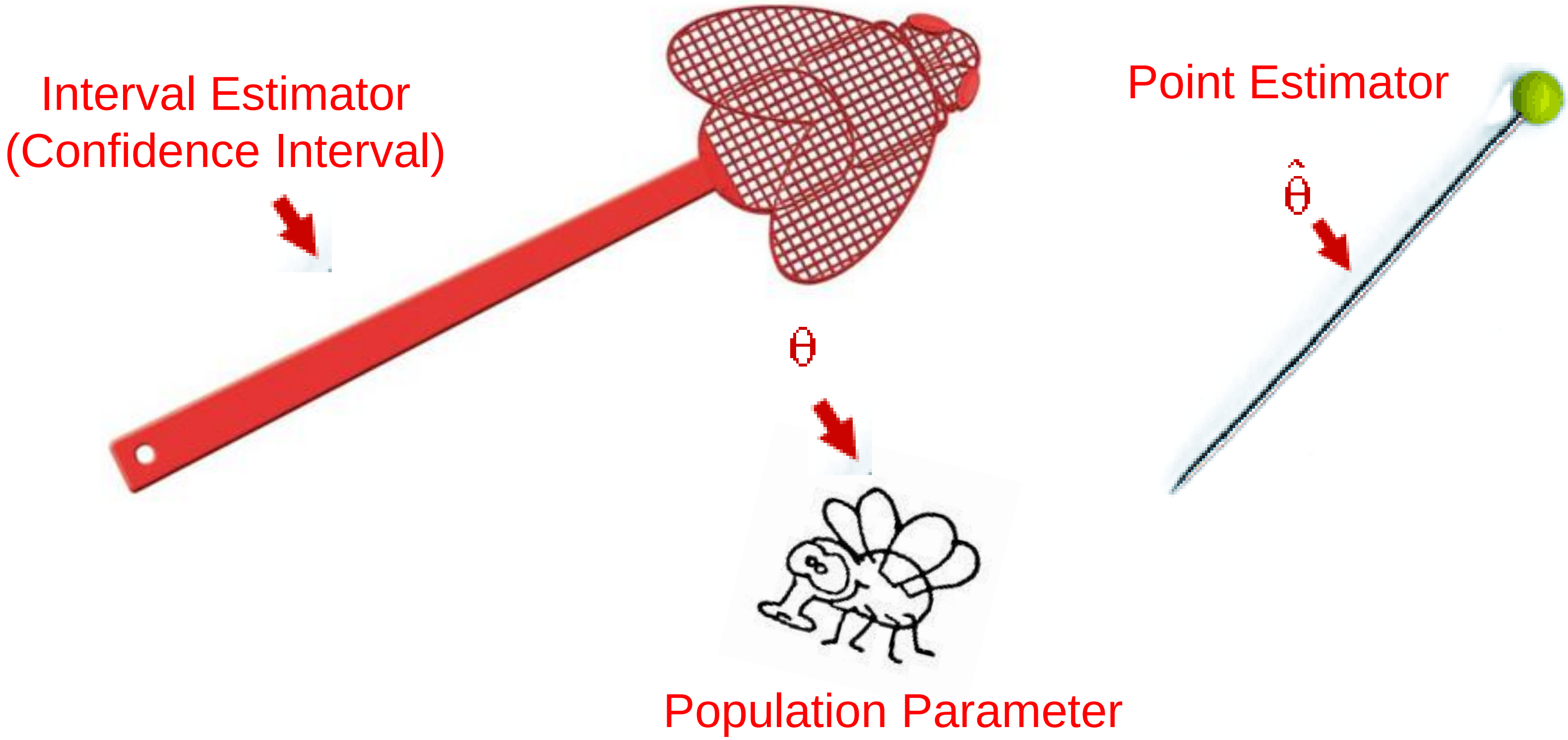
ARTEN VON SCHÄTZERN

HOCHSCHULE HEILBRONN

Schätzer

Punktschätzer	Intervallschätzer (Konfidenzintervall)
> Einzige Zahl als Schätzung	> Intervall: Obere und untere Schätzgrenze
Beispiel. Durchschnittsgröße = 176 cm	Beispiel. Durchschnittsgröße liegt zwischen 175 cm und 180 cm

ARTEN VON SCHÄTZERN



PUNKT-SCHÄTZER

HOCHSCHULE HEILBRONN

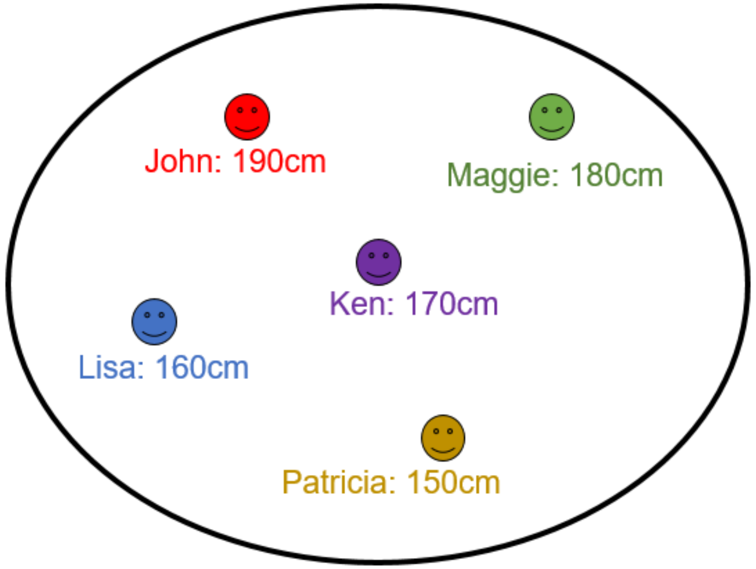
Notation. Schätzwerte werden mit überdachten griechischen Buchstaben $\widehat{\ }$ bezeichnet.

Beispiel. $\hat{\mu}$ ist der Schätzwert für den Erwartungswert μ.

PUNKTSCHÄTZER: MITTELWERT

HOCHSCHULE HEILBRONN

Beispiel. Betrachten Sie folgende Population:



> Der Populationsmittelwert ist:

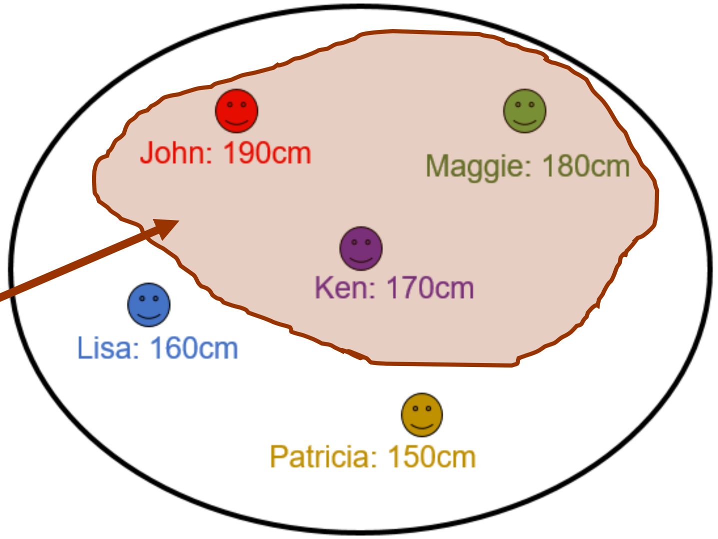
$$\mu = \frac{150 + 160 + 170 + 180 + 190}{5} = \mathbf{170}$$

PUNKTSCHÄTZER: MITTELWERT

HOCHSCHULE HEILBRONN

Beispiel (Fortsetzung).

- Der Mittelwert $\mu = 170$ sei **unbekannt.**
- Wir möchten den Mittelwert schätzen
- Ziehe eine Stichprobe, z.B. Größe $n = 3$



Der **Schätzer** für den **Populationsmittelwert** ist der **Stichprobenmittelwert:**

$$\hat{\mu} = \bar{x} = \frac{170 + 180 + 190}{3} = \mathbf{180}$$

PUNKTSCHÄTZER:
VERTEILUNG DER SCHÄTZWERTE

HOCHSCHULE HEILBRONN

- Unsere Stichprobe überschätzt den wahren Mittelwert
- Andere unterschätzen den wahren Mittelwert

Ist es möglich etwas über einen möglichen Schätzfehler zu sagen?

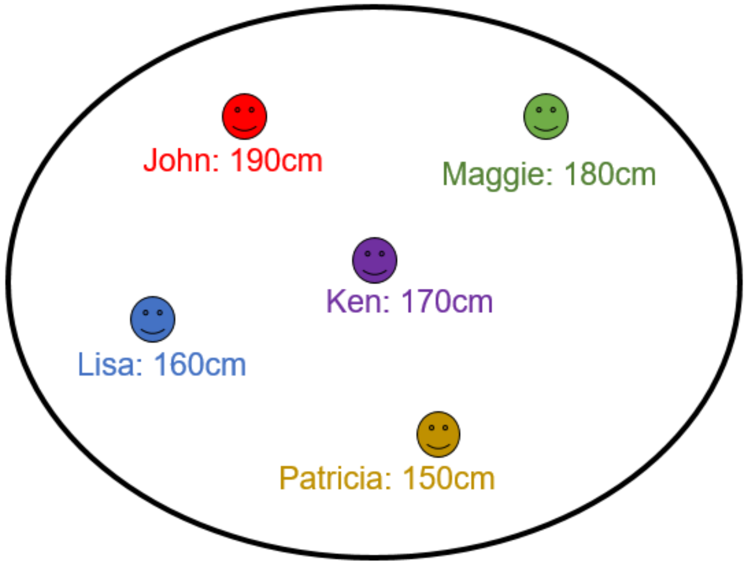
- Ja → Wir müssen die Verteilung der Schätzwerte untersuchen:

Je stärker die Schätzwerte um den Populationsparameter streuen,
desto wahrscheinlicher ist ein hoher Schätzfehler

PUNKTSCHÄTZER: VERTEILUNG DER SCHÄTZWERTE

HOCHSCHULE HEILBRONN

Beispiel.

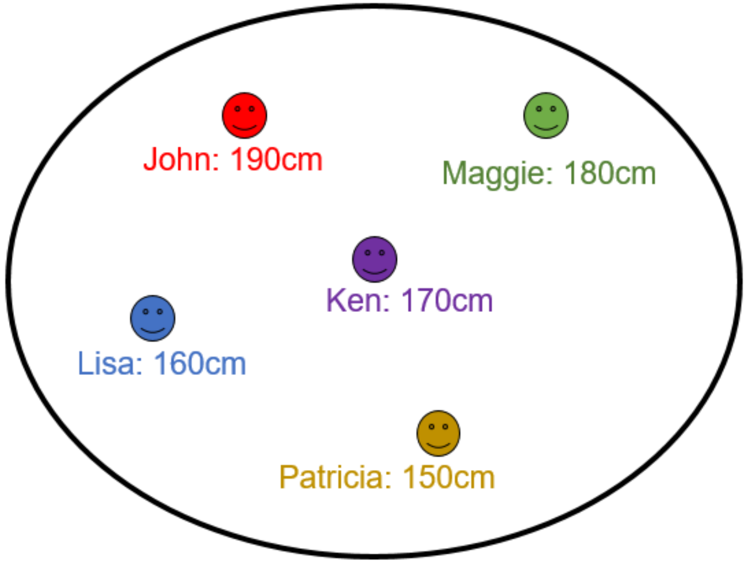


> Wie viele unterschiedliche Stichproben mit $n = 3$ können wir ziehen?

$$\binom{Populationsgröße}{Stichprobengröße} = \binom{5}{3} = \frac{5!}{3!\,2!} = \frac{5 \cdot 4}{2} = 10$$

PUNKTSCHÄTZER: VERTEILUNG DER SCHÄTZWERTE

Verteilung der Stichproben.



	Verteiling Stichproben	
i	Stichprobe i	$P(i)$
1	150, 160, 170	10%
2	150, 160, 180	10%
3	150, 160, 190	10%
4	150, 170, 180	10%
5	150, 170, 190	10%
6	150, 180, 190	10%
7	160, 170, 180	10%
8	160, 170, 190	10%
9	160, 180, 190	10%
10	170, 180, 190	10%

Zufallsstichprobe = Jedes Element der Population wird mit gleicher W'keit ausgewählt

PUNKTSCHÄTZER: VERTEILUNG DER SCHÄTZWERTE

HOCHSCHULE HEILBRONN

Verteilung Stichprobenmittelwerte (Stichprobenmittelw. = Schätzer Populationsmittelw.).

	Verteilung Stichproben		
i	Stichprobe i	$P(i)$	Mittelw. $\overline{x}_i$
1	150, 160, 170	10%	160
2	150, 160, 180	10%	163.34
3	150, 160, 190	10%	166.67
4	150, 170, 180	10%	166.67
5	150, 170, 190	10%	170
6	150, 180, 190	10%	173.34
7	160, 170, 180	10%	170
8	160, 170, 190	10%	173.34
9	160, 180, 190	10%	176.67
10	170, 180, 190	10%	180

→

Verteilung Stichprobenmittelwerte	
$\overline{x}$	$P(\overline{x})$
160	10%
163.34	10%
166.67	20%
170	20%
173.34	20%
176.67	10%
180	10%

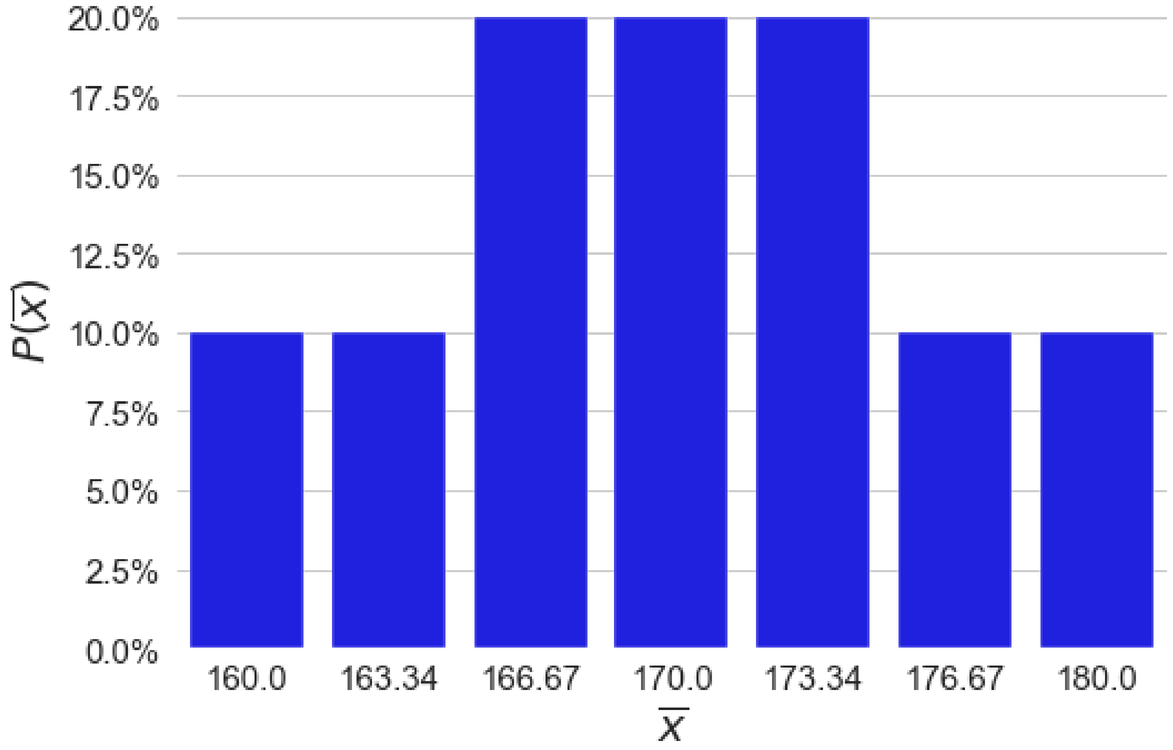
7 unterschiedliche Stichprobenmittelwerte

PUNKTSCHÄTZER: VERTEILUNG DER SCHÄTZWERTE

HOCHSCHULE HEILBRONN

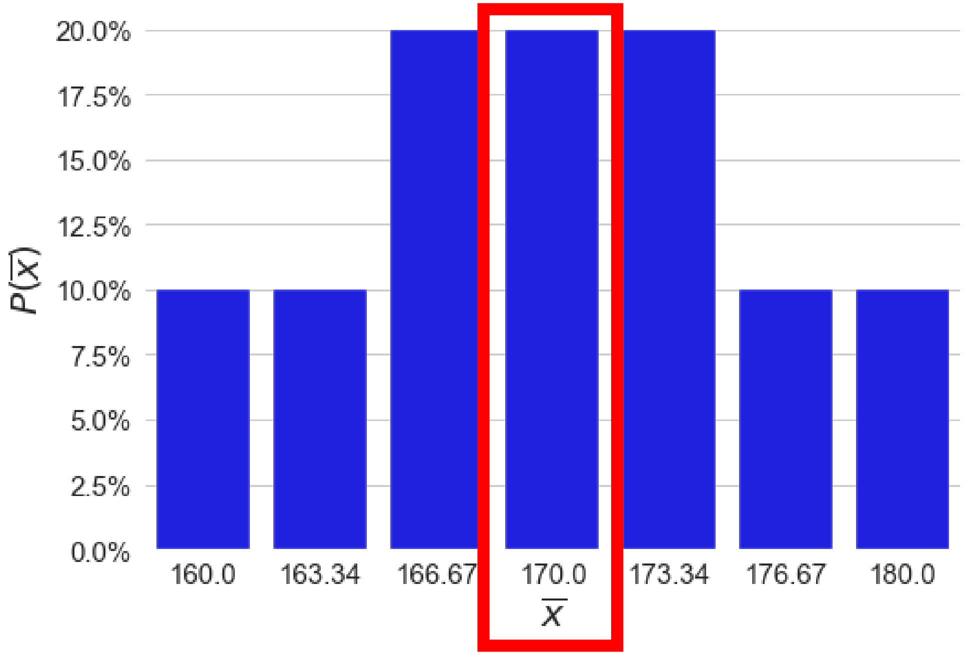
Verteilung Stichprobenmittelwerte (Stichprobenmittelw. = Schätzer Populationsmittelw.).

Verteilung Stichprobenmittelwerte	
$\overline{x}$	$P(\overline{x})$
160	10%
163.34	10%
166.67	20%
170	20%
173.34	20%
176.67	10%
180	10%



PUNKTSCHÄTZER:
VERTEILUNG DER SCHÄTZWERTE

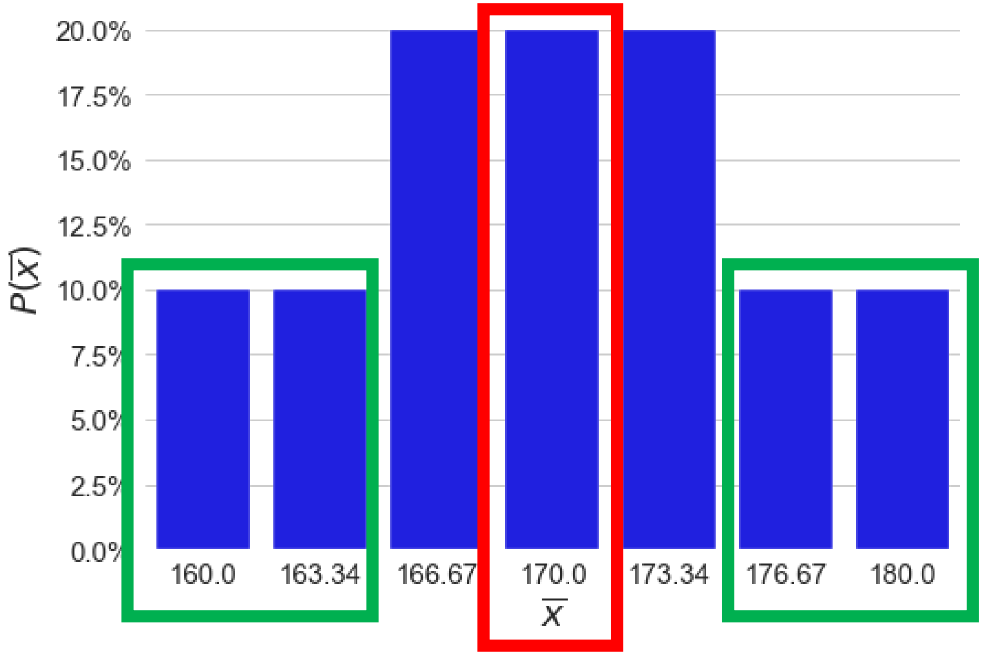
Verteilung Stichprobenmittelwerte (Stichprobenmittelw. = Schätzer Populationsmittelw.).



20% W'keit dass $Schätzwert\ \bar{x} = Populationsmittelwert\ \mu = 170$

PUNKTSCHÄTZER: VERTEILUNG DER SCHÄTZWERTE

Verteilung Stichprobenmittelwerte (Stichprobenmittelw. = Schätzer Populationsmittelw.).



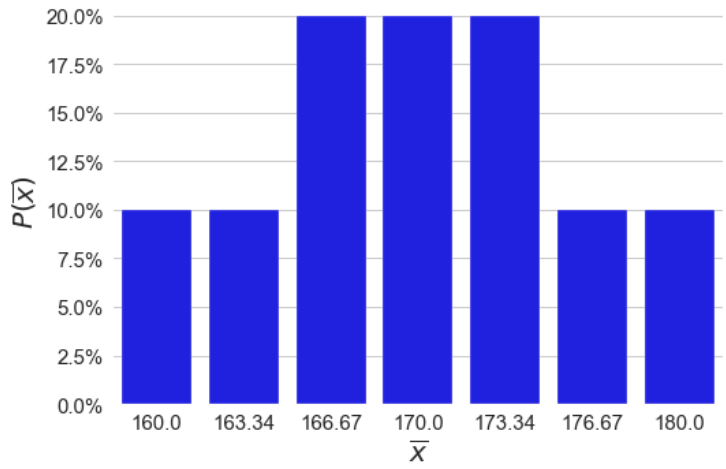
> Wahrscheinlichkeit, dass der Schätzfehler mindestens 5cm ist?

40%

PUNKTSCHÄTZER: VERTEILUNG DER SCHÄTZWERTE

HOCHSCHULE HEILBRONN

Verteilung Stichprobenmittelwerte (Stichprobenmittelw. = Schätzer Populationsmittelw.).



- Mittelwert der Verteilung der Stichprobenmittelwerte $\mu_{\bar{X}}$ = Populationsmittelwert μ
- **Standardfehler** = Standardabweichung der Verteilung → Streuung um μ
- Je höher der Standardfehler, desto wahrscheinlicher ist ein großer Schätzfehler

STANDARD-
FEHLER

HOCHSCHULE HEILBRONN

Standardfehler = Standardabweichung der Verteilung der Schätzwerte

- Hier: Schätzwerte = Stichprobenmittelwerte

Berechnung Standardfehler:

- Wenn Verteilung Schätzwerte bekannt → Berechnung Standardabweichung wie bekannt
- Üblicherweise nicht bekannt, dann Berechnung Standardfehler (standard error) SE:

$$SE_{\overline{X}} = \frac{Populationsstandardabweichung}{\sqrt{Stichprobengröße}} = \frac{\sigma}{\sqrt{n}}$$

STANDARD-FEHLER

Anmerkung. Berechnung des Standardfehlers.

Wenn die Stichprobe groß ist (n > 5% der Population), dann

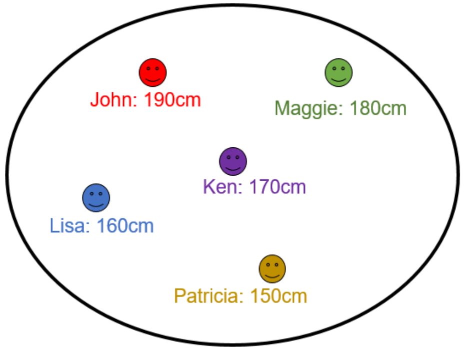
> müssen wir die Standardfehler mit folgendem **Korrekturfaktor** multiplizieren:

$$\sqrt{\frac{N-n}{N-1}},$$

N = Populationsgröße, n = Stichprobengröße

STANDARD-FEHLER

Beispiel. Berechnung Standardfehler (bekannte Pop.varianz).



- Pop.Varianz: $\sigma^2 = 200 \rightarrow$ Standardabweichung: $\sigma = \sqrt{\sigma^2} \approx 14.14$
- Stichprobengröße $n = 3$ groß oder klein?
- Groß ($n = 60\%$ der Population). Somit:

$$Standardfehler\ korrigiert = SE_{\bar{X}} \cdot Korrekturfaktor = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{14.14}{\sqrt{3}} \sqrt{\frac{5-3}{5-1}} \approx 5.77$$

PUNKTSCHÄTZER: DIFFERENZ VON MITTELWERTEN

HOCHSCHULE HEILBRONN

Gruppen werden häufig anhand der Differenz ihrer Mittelwerte verglichen

- Schätzer für Mittelwertdifferenz ($\mu = \mu_X - \mu_Y$) = Differenz Stichprobenmittelwerte:

$$\boldsymbol{\hat{\mu} = Stichprobenmittelwert\ x - Stichprobenmittelwert\ y}$$

- Standardfehler der Verteilung der Differenzen = gepoolter Standardfehler SE_{X-Y}:

$$\boldsymbol{SE_{X-Y} = \sqrt{(Standardfehler\ X)^2 + (Standardfehler\ Y)^2}}$$

PUNKTSCHÄTZER: VARIANZ

HOCHSCHULE HEILBRONN

> Der Schätzer für die Varianz $\hat{\sigma}^2$ ist die Bessel-korrigierte Stichprobenvarianz:

$$\boldsymbol{\hat{\sigma}^2 = Stichprobenvarianz \cdot \frac{n}{n-1}}$$

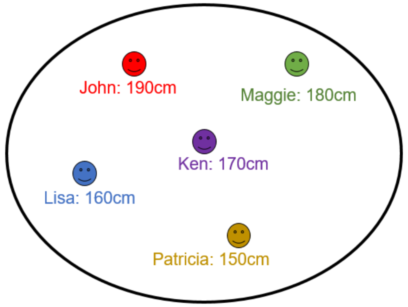
> Der Schätzer für die Standardabweichung $\hat{\sigma}$ ist somit:

$$\boldsymbol{\hat{\sigma} = Stichprobenstandardabweichung \cdot \sqrt{\frac{n}{n-1}}}$$

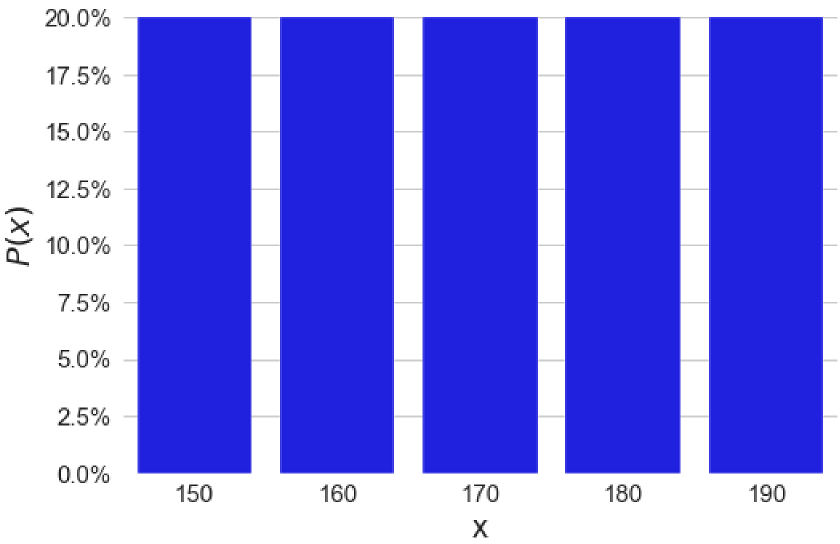
> Den komplizierten Standardfehler der Stichprobenvarianzverteilung lassen wir aus

EXKURS: ZENTRALER GRENZWERTSATZ

Beispiel 1:

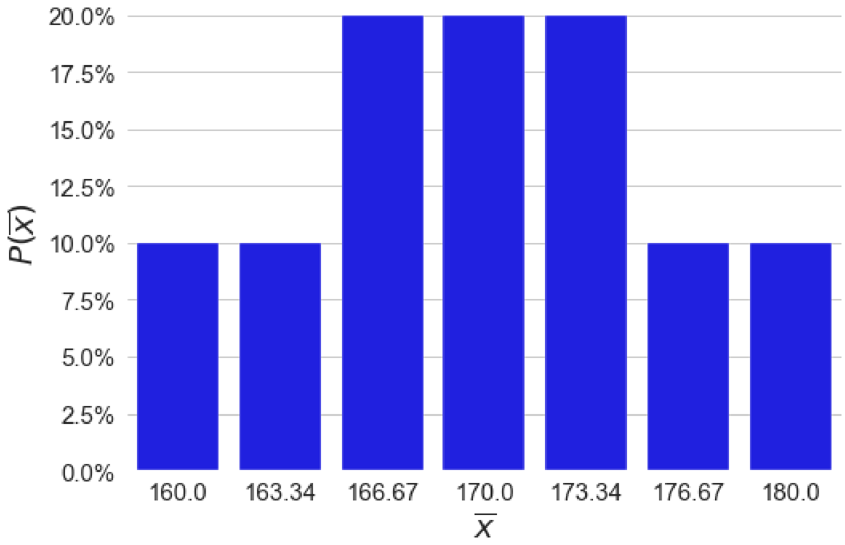


Population:
Gleichverteilung



Stichproben
$(n = 3)$

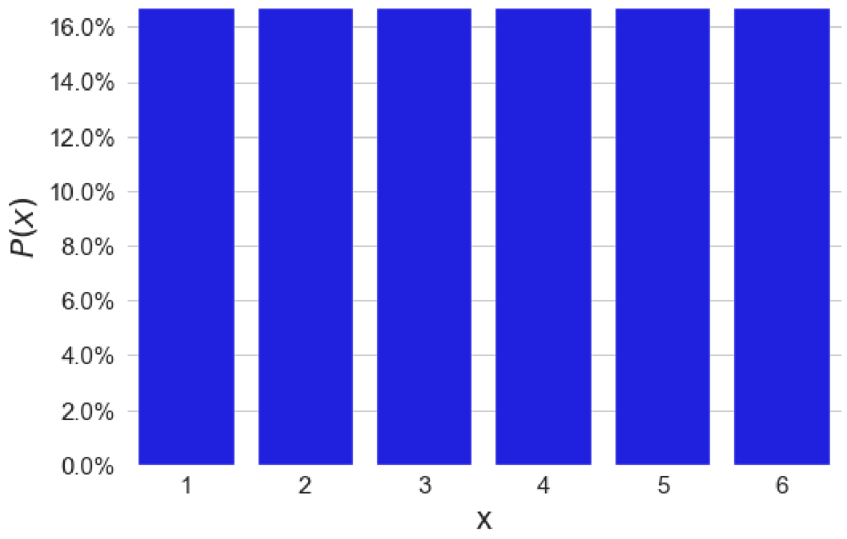
Verteilung Stichprobenmittelwerte:
Keine Gleichverteilung



EXKURS:
ZENTRALER GRENZWERTSATZ

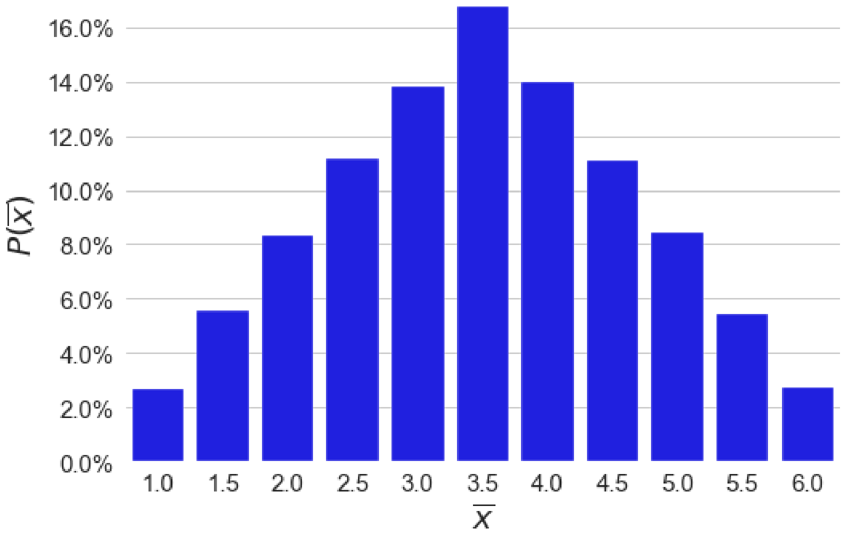
Beispiel 2. Zweimaliger Würfelwurf: Stichprobenmittelwert = Durchschnitt der Augenzahlen.

Population:
Gleichverteilung



100'000
Stichproben

Verteilung Stichprobenmittelwerte:
Keine Gleichverteilung



EXKURS: ZENTRALER GRENZWERTSATZ

Zentraler Grenzwertsatz

Summe und Mittelwert von n unabhängig und identisch verteilten Zufallsvariablen sind approximativ normalverteilt, wenn n hinreichend groß ist.

Implikationen:

- Normalverteilung Population → Stichprobenverteilung normal (n egal)
- Keine Normalverteilung Population → Stichprobenverteilung normal (wenn n groß)

Daumenregel: Die Stichprobenverteilung ist annähernd normalverteilt, wenn

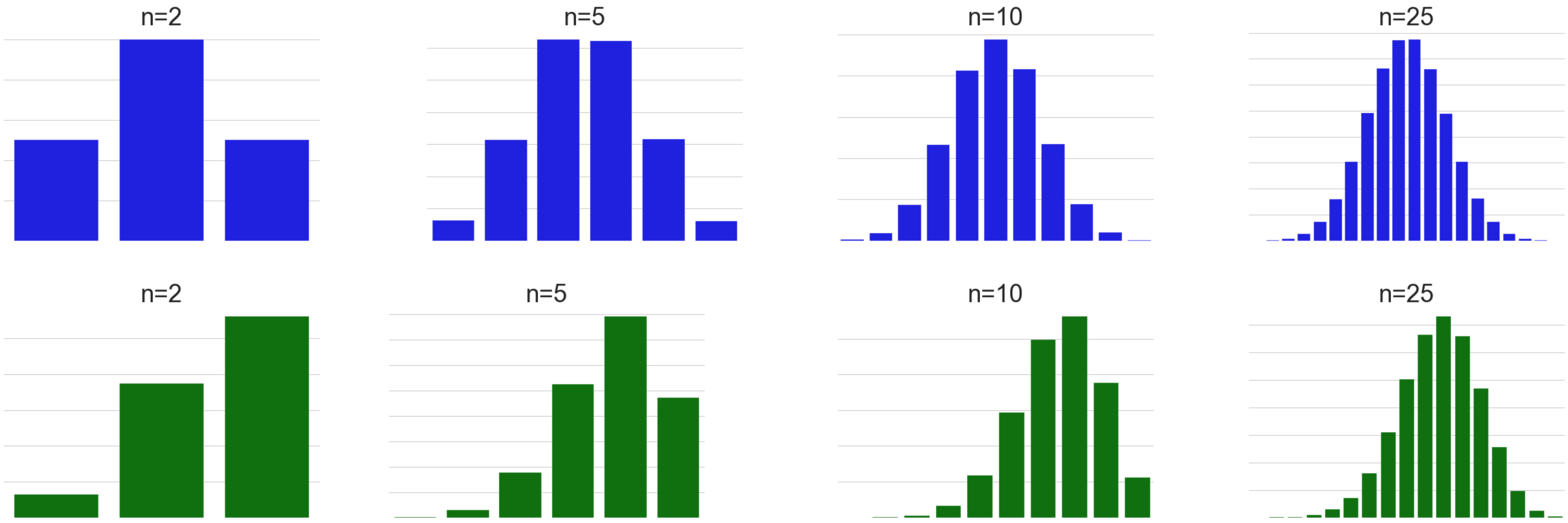
$$Stichprobengröße \geq 30$$

EXKURS: ZENTRALER GRENZWERTSATZ

HOCHSCHULE HEILBRONN

Illustration. Populationen: Bernoulli-verteilt mit $\pi = 50\%$ und $\pi = 75\%$

> Relative Häufigkeitsverteilung des Stichprobenmittels bei 100'000 Stichproben:



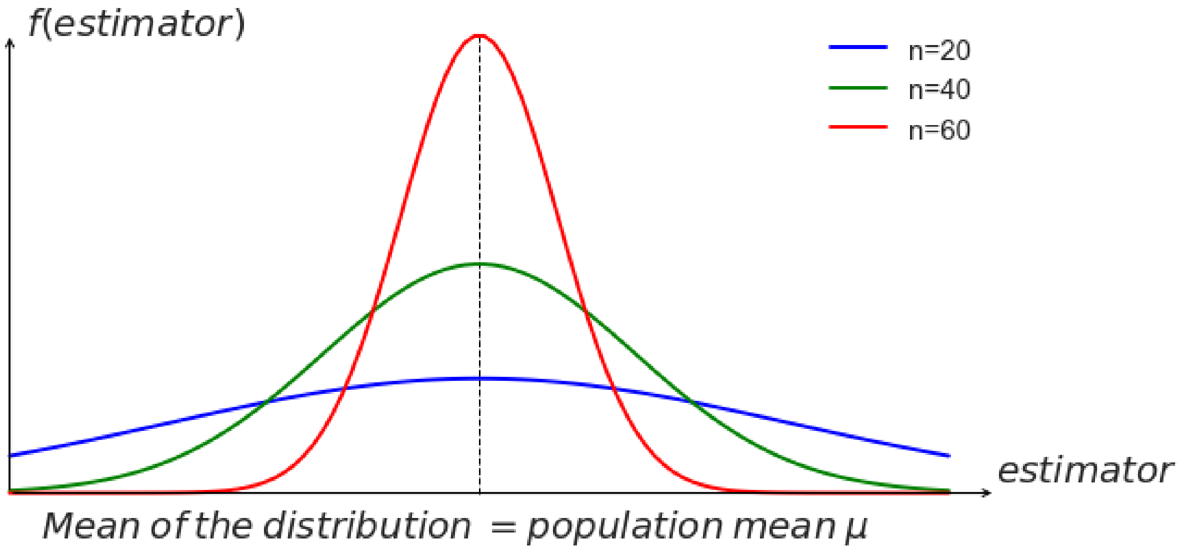
PUNKTSCHÄTZER: SCHÄTZFEHLER UND ZENTRALER GRENZWERTSATZ

HOCHSCHULE HEILBRONN

Schätzfehler = Schätzwert – Wahrer Parameter

> Mittelwert: $Schätzfehler = Schätzer - \mu =$ Stichprobenmittelwert $(\bar{x})$ $-\mu$

> Welche der folgenden Verteilungen ergibt im Schnitt den kleinsten Schätzfehler?



PUNKTSCHÄTZER: SCHÄTZFEHLER UND ZENTRALER GRENZWERTSATZ

HOCHSCHULE HEILBRONN

Beispiel. Körpergröße ist verteilt mit Mittelwert 167cm (Standardabweichung = 10cm).

> Stichprobe mit n = 36 Personen

> Wahrscheinlichkeit, dass der Mittelwert um mindestens 3cm überschätzt wird?

PUNKTSCHÄTZER: SCHÄTZFEHLER UND ZENTRALER GRENZWERTSATZ

HOCHSCHULE HEILBRONN

Beispiel. Monatliches Einkommen von Männern (M) und Frauen (W).

> Mittelwerte: $\mu_M = 3'201€$ und $\mu_W = 3'001€$

> Standardabweichungen: $\sigma_M = 420€$ und $\sigma_W = 305€$

> Stichprobe mit 36 Frauen und 49 Männer

> Wahrscheinlichkeit, dass Durchschnittseinkommen Frauen ≥ Männer?

MAXIMALER SCHÄTZFEHLER (= ERROR BOUND EB)

- In den meisten Fällen: Populationsmittelwert nicht bekannt
- Schätzfehler = Stichprobenmittelwert - Populationsmittelwert
- Ist es möglich den Schätzfehler bei unbekanntem Populationsmittelwert abzuschätzen?

Überraschenderweise ja

Betrachten Sie den z-Wert der Zufallsvariable des Stichprobenmittels ($\bar{X}$):

$$Z(\bar{X}) = \frac{\bar{X} - \mu}{Standardfehler\ (SE)}$$

Zufallsvariable des Schätzfehlers

MAXIMALER SCHÄTZFEHLER (= ERROR BOUND EB)

Betrachten Sie den z-Wert der Zufallsvariable des Stichprobenmittels ($\bar{X}$):

$$Z(\bar{X}) = \frac{\bar{X} - \mu}{SE}$$

Zufallsvariable des Schätzfehlers

Also:

$$Z = \frac{Schätzfehler}{SE} \Leftrightarrow Schätzfehler = Z \cdot SE$$

> Der Standardfehler SE ist bekannt

> ABER Z ist eine Zufallsvariable → bisher kann der Schätzfehler nicht berechnet werden

MAXIMALER SCHÄTZFEHLER (= ERROR BOUND EB)

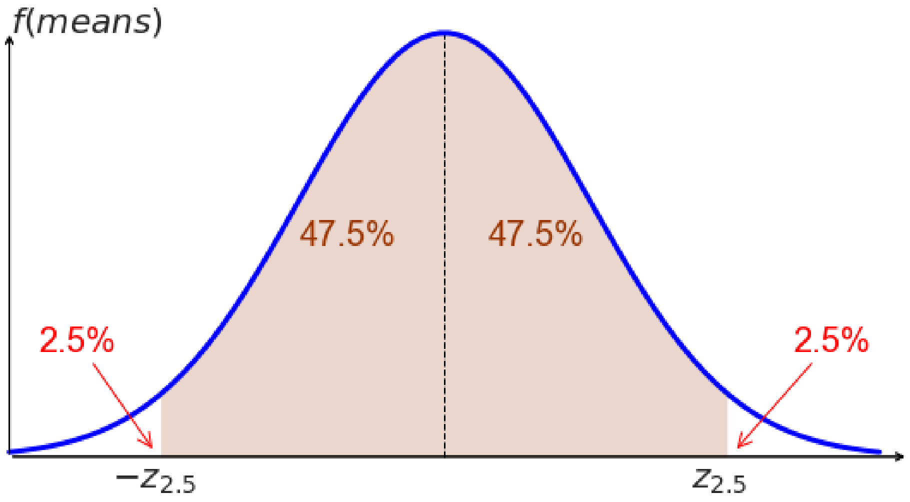
HOCHSCHULE HEILBRONN

Konfidenzniveau festlegen.

> Wir können ein Konfidenzniveau $(1-\alpha)$ festlegen

> $(1-\alpha)$ = 95% bedeutet z.B., dass der Schätzfehler zu 95% in dem Bereich liegt

> Wir lassen eine Fehlerwahrscheinlichkeit von $\alpha/2 = 2.5\%$ auf jeder Seite zu:



MAXIMALER SCHÄTZFEHLER (= ERROR BOUND EB)

HOCHSCHULE HEILBRONN

Bereich bestimmen.

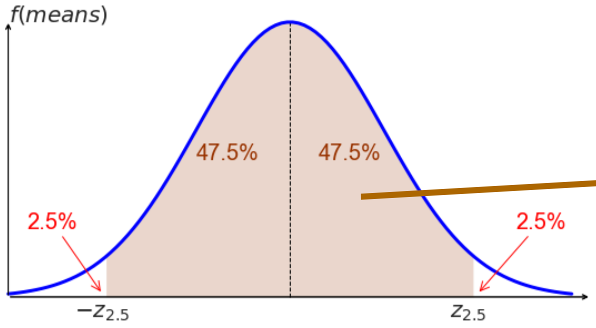
> Wie finden wir die Grenzen $-z_{\alpha/2}$ und $z_{\alpha/2}$ für ein festgelegtes Konfidenzniveau?

ERROR BOUND (EB)

Bereich bestimmen.

Beispiel. $\alpha = 5\% \rightarrow z_{2.5}$?

> Nun: Z-Wert für Fläche finden!



> Also:

$$\boldsymbol{z_{2.5} = 1.96}$$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.00	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.10	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.20	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.30	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.40	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.50	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.60	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.70	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.80	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.90	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.00	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.10	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.20	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.30	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.40	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.50	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.60	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.70	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.80	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.90	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.00	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.10	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.20	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.30	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.40	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.50	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.60	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.70	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.80	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.90	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.00	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

MAXIMALER SCHÄTZFEHLER (= ERROR BOUND EB)

HOCHSCHULE HEILBRONN

Fixiere ein Konfidenzniveau $(1-\alpha)$ und bestimme hierfür $z_{\alpha/2}$

> dann können wir den maximalen Schätzfehler (error bound eb) berechnen:

$$error\ bound\ für\ (1-\alpha) = z_{\alpha/2} \cdot Standardfehler$$

> Interpretation:

Mit W'keit $(1-\alpha)$ ist der maximale Schätzfehler nicht höher als $z_{\alpha/2} \cdot Standardfehler$

MAXIMALER SCHÄTZFEHLER (ERROR BOUND EB)

HOCHSCHULE HEILBRONN

Berechnung des error bound.

Beispiel. $z_{2.5} = 1.96$, Größe Stichprob. $n = 64$, Standardabweichung $\sigma = 16$.

$$eb = z_{2.5} \cdot SE = z_{2.5} \cdot \frac{\sigma}{\sqrt{n}} = 1.96 \cdot \frac{16}{\sqrt{64}} = 3.92$$

> Interpretation:

Maximaler Schätzfehler ist mit 95% W'keit nicht höher als 3.92

ERROR BOUND UND OPTIMALE STICHPROBENGRÖßE

Frage wenn Daten erhoben werden → Welche Stichprobengröße?

> Frage:

Wie akkurat sollen die Schätzungen sein?

> Wir können den error bound benutzen, um diese Frage zu beantworten

ERROR BOUND UND OPTIMALE STICHPROBENGRÖßE

HOCHSCHULE HEILBRONN

Beispiel. Sie besitzen eine große Brauerei und möchten Ihre Tagesproduktion schätzen.

- Ihre durchschnittliche Tagesproduktion der letzten 30 Tage ist 871 Kubikliter Bier
- Standardabweichung: 21 Kubikliter

Wie viele Tage müssen Sie die Tagesproduktion beobachten, um die langfristige Tagesproduktion mit 95% zu schätzen bei einem error bound von 4 Kubiklitern?

ERROR BOUND UND INTERVALLSCHÄTZER

Die folgenden Aussagen sind äquivalent:

1. Zu 95% ist der Schätzfehler nicht höher als 3.92 (Interpretation eb)

2. Zu 95% liegt der Populationsmittelwert in einem Intervall Stichprobenmittel $\pm$ 3.92

Aussage 2. legt einen Intervallschätzer nahe mit den folgenden Schätzgrenzen:

Punktschätzer ± Error Bound

INTERVALL-SCHÄTZER

HOCHSCHULE HEILBRONN

Beispiel. Zahlungsbereitschaft von Kunden für eine Übernachtung im Hotel bestimmen.

> In einer Stichprobe von n = 39 Befragten liegt die Zahlungsbereitschaft bei 86 €
> Standardabweichung = 26 €

a) Bestimmen Sie das 99%-Konfidenzintervall

Hausaufgabe.

b) Bestimmen Sie das 90%-Konfidenzintervall

Lösung: 90%-Konfidenzintervall = $[79, 93]$

INTERVALL-SCHÄTZER

HOCHSCHULE HEILBRONN

Beispiel. Wir möchten den durchschnittlichen IQ von Studierenden der Hochschule Heilbronn schätzen.

- In einer Stichprobe mit 47 Studierenden ist der durchschnittliche IQ 102
- Standardabweichung = 13

a) Bestimmen Sie das 90%-Konfidenzintervall

Hausaufgabe.

b) Bestimmen Sie das 95%-Konfidenzintervall

Lösung: 95%-Konfidenzintervall = $[98.28, 105.72]$

KONFIDENZ-INTERVALLE

Optimales Konfidenzintervall: So **schmal wie möglich** und **Konfidenzniveau nahe 1**

> Je größer das Konfidenzniveau, desto sicherer ist wahrer Parameter enthalten

> Je enger das Konfidenzintervall, desto exakter ist die Schätzung

> Je größer die Stichprobe, desto enger ist das Intervall

Aber:

> Je höher das Konfidenzniveau, desto breiter das Intervall

› HYPOTHESEN-TESTS

WARUM HYPOTHESENTESTS?

In einer Stichprobe mit 35 Frauen und 44 Männer ist das Durchschnittseinkommen der Frauen um 200 € niedriger als das der Männer

> Ist dies ein Beleg dafür, dass Frauen allgemein im Schnitt weniger verdienen?

Der Umsatz einer Firma war im letzten Jahr 300'000 € höher als der Durchschnitt ihres langfristigen jährlichen Umsatzes

> Ist dies ein Beleg dafür, dass sich die Nachfrage für die Produkte erhöht hat?

Derartige Fragen werden anhand von Hypothesentests beantwortet

HYPOTHESENTEST: NULL- UND ALTERNATIVHYPOTHESE

HOCHSCHULE HEILBRONN

Die Denklogik eines Hypothesentests ist anders als wir im Alltag gewöhnt sind:

- Keine direkten Belege für eine Hypothese (**Alternativ- bzw. Forschungshypothese**)
- Stattdessen: Ist das Gegenteil (**Nullhypothese)** unwahrscheinlich gegeben die Daten?

HYPOTHESENTEST: NULL- UND ALTERNATIVHYPOTHESE

HOCHSCHULE HEILBRONN

Beispiel. Wir möchten untersuchen, ob Frauen weniger verdienen als Männer.

- Alternativhypothese: $Durchschnittseinkommen_{Frauen} < Durchschnittseinkommen_{Männer}$

Wir untersuchen nun wie wahrscheinlich das Gegenteil ist:

- Nullhypothese: $Durchschnittseinkommen_{Frauen} \geq Durchschnittseinkommen_{Männer}$

HYPOTHESENTEST: NULL- UND ALTERNATIVHYPOTHESE

Hypothesen

Alternativhypothese H_a (auch H_1)

- Hypothese, die getestet werden soll
- Postuliert üblicherweise ein Resultat

Nullhypothese H_0

- Exaktes Gegenteil der Alternativhypothese
- Verneint das Resultat

HYPOTHESENTEST: NULL- UND ALTERNATIVHYPOTHESE

HOCHSCHULE HEILBRONN

Beispiele.

Wir möchten wissen…	Alternativhypothese (H_a)	Nullhypothese (H_0)
ob Absolventen der HS Heilbronn im Schnitt (μ_{Alter}) älter sind als 26 (μ_0)	$\mu_{Alter} > 26$	$\mu_{Alter} \leq 26$
ob sich der Durchschnitts-IQ (μ_{IQ}) von Studierenden verändert hat (langfristiger Durchschnitt: $\mu_0 = 101$)	$\mu_{IQ} \neq 101$	$\mu_{IQ} = 101$
ob die Nachfrage (μ_{Nach}) für das Produkt einer Firma abgenommen hat (langfristiger Durchschnitt: $\mu_0 = 41'000$)	$\mu_{Nach} < 41'000$	$\mu_{Nach} \geq 41'000$

HYPOTHESENTEST: NULL- UND ALTERNATIVHYPOTHESE

HOCHSCHULE HEILBRONN

Beispiel. In einer Stichprobe mit 35 Frauen und 44 Männer ist das monatliche Durchschnittseinkommen der Frauen $2.9\,T€$ $(Varianz = 0.7)$ und das der Männer $3.1\,T€$ $(Varianz = 0.4)$.

Ist dies ein Beleg dafür ist, dass Frauen im Schnitt weniger verdienen?

a) Schreiben Sie die Null- und die Alternativhypothese zu dieser Fragestellung auf.

> Alternativhypothese:

$$Durchschnittseinkommen\ Pop.Frauen\ \mu_f < Durchschnittseinkommen\ Pop.Männer\ \mu_m$$

> Nullhypothese:

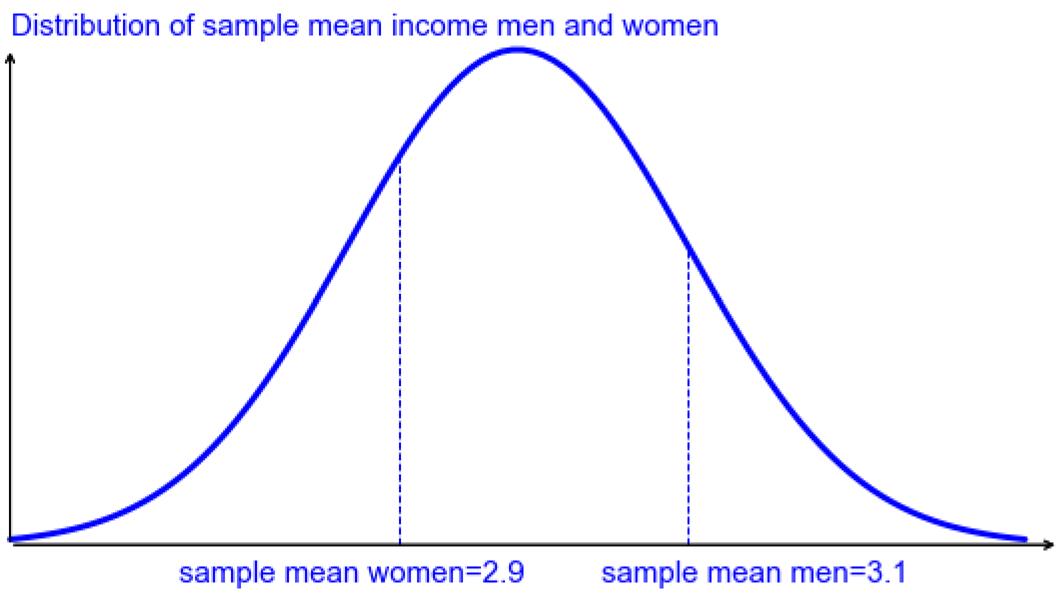
$$Durchschnittseinkommen\ Pop.Frauen\ \mu_f \geq Durchschnittseinkommen\ Pop.Männer\ \mu_m$$

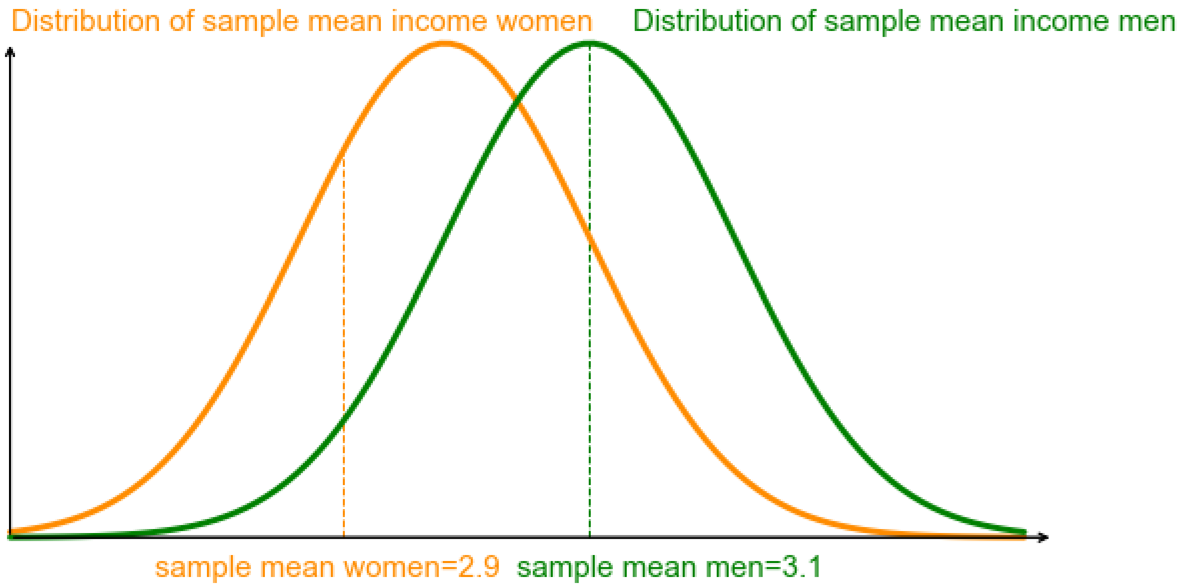
HYPOTHESENTEST: NULL- UND ALTERNATIVHYPOTHESE

HOCHSCHULE HEILBRONN

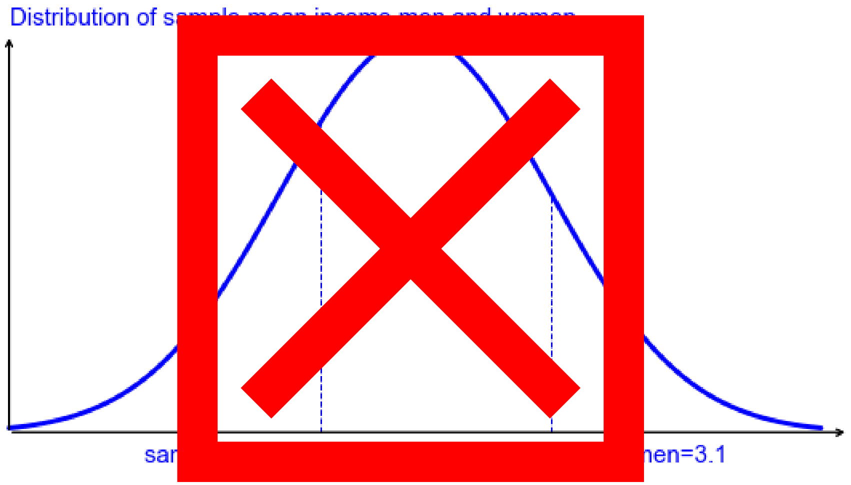
Null- und Alternativhypothese repräsentieren **zwei unterschiedliche Welten:**

Welt 0 (Nullhypothese H_0):	**Welt 1 (Alternativhypothese H_a):**
Frauen verdienen nicht weniger.	Frauen verdienen weniger.
> Gleiche Einkommensverteilung	> Unterschiedliche Einkommensverteilungen.

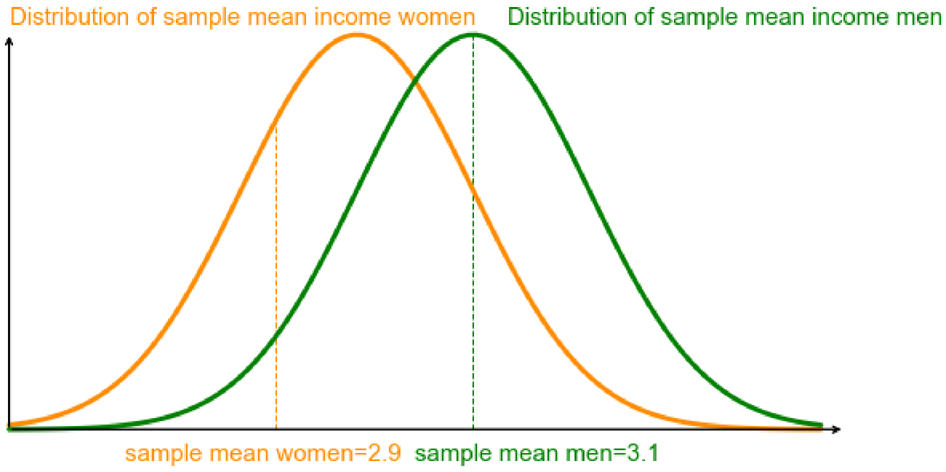




HYPOTHESENTEST:
NULL- UND ALTERNATIVHYPOTHESE



In welcher Welt leben wir?



Welt 0 (Nullhypothese H_0)

Welt 1 (Alternativhypothese H_a)

> Ein Hypothesentest sagt uns, ob wir die Nullhypothese ablehnen können

> Wenn wir die Nullhypothese ablehnen können, dann ist der Test erfolgreich

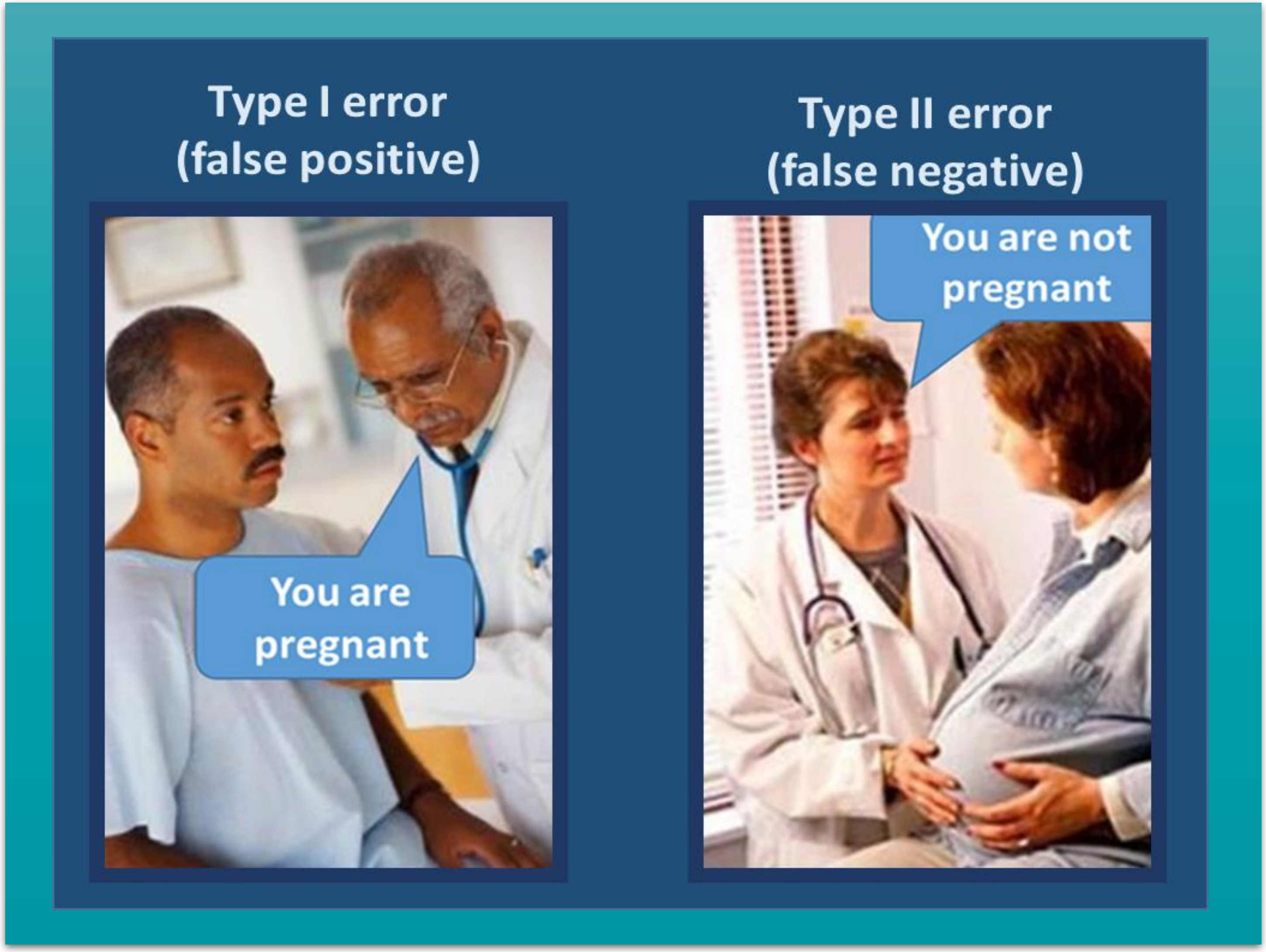
HYPOTHESENTEST: FEHLER 1. UND 2. ART

Welche Fehler sind möglich bei der Entscheidung, ob man die Nullhypothese ablehnt?

	Nullhypothese (H_0)	
Entscheidung	**Wahr**	**Falsch**
Nullhypothese ablehnen	**Fehler 1. Art (α-Fehler):** Falsch-positives Resultat (Wahrscheinlichkeit: α)	Richtige Entscheidung (Wahrscheinlichkeit: $1 - \beta$)
Nullhypothese nicht ablehnen	Richtige Entscheidung (Wahrscheinlichkeit: $1 - \alpha$)	**Fehler 2. Art (β-Fehler):** Falsch-negatives Resultat (Wahrscheinlichkeit: β)

HYPOTHESENTEST: FEHLER 1. UND 2. ART

HOCHSCHULE HEILBRONN



HYPOTHESENTEST: FEHLER 1. UND 2. ART

Welches Kriterium bestimmt, ob wir die Nullhypothese (Welt 1) ablehnen können?

Wir berechnen eine Statistik, genannt ***Teststatistik***, mit welcher wir die

> W'keit der beobachteten Daten oder extremeren Werten in Welt 0 berechnen (= ***p-Wert***)

Die Nullhypothese kann abgelehnt werden, wenn der p-Wert hinreichend gering ist

> Das heißt wir minimieren die W'keit eines Fehlers 1. Art

HYPOTHESENTEST: FEHLER 1. UND 2. ART

HOCHSCHULE HEILBRONN

Warum minimieren wir nicht auch die W'keit eines Fehlers 2. Art?

> Es ist nicht möglich beide zu minimieren: W'keit Fehler 1. Art $\downarrow \rightarrow$ W'keit Fehler 2. Art $\uparrow$

Vorteile Minimierung Fehler 1. Art (statt 2. Art):

- **konservativer** $\rightarrow$ Nullhypothese wird weniger oft abgelehnt
- **exakter** $\rightarrow$ α kann berechnet werden, β muss geschätzt werden

HYPOTHESENTEST: TESTSTATISTIK

HOCHSCHULE HEILBRONN

Erinnerung:

Teststatistik → p-Wert → Entscheidung, ob Nullhypothese abgelehnt werden kann

- Die Teststatistik ist eine Statistik (Zahl die Daten einer Stichprobe zusammenfasst)

HYPOTHESENTEST: TESTSTATISTIK

HOCHSCHULE HEILBRONN

Beispiel (Fortsetzung). In einer Stichprobe mit 35 Frauen und 44 Männer ist das monatliche Durchschnittseinkommen der Frauen $2.9\,T€$ $(Varianz = 0.7)$ und das der Männer $3.1\,T€$ $(Varianz = 0.4)$.

Ist dies ein Beleg dafür ist, dass Frauen im Schnitt weniger verdienen?

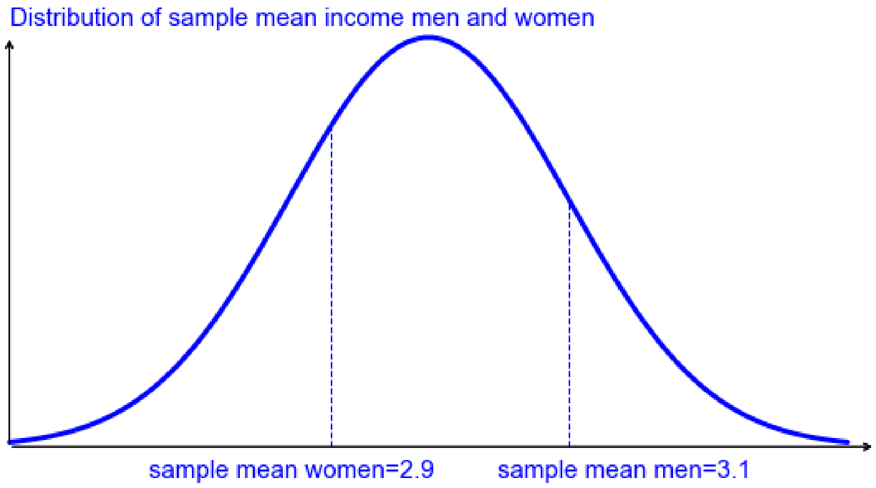
a) Schreiben Sie die Null- und die Alternativhypothese zu dieser Fragestellung auf.

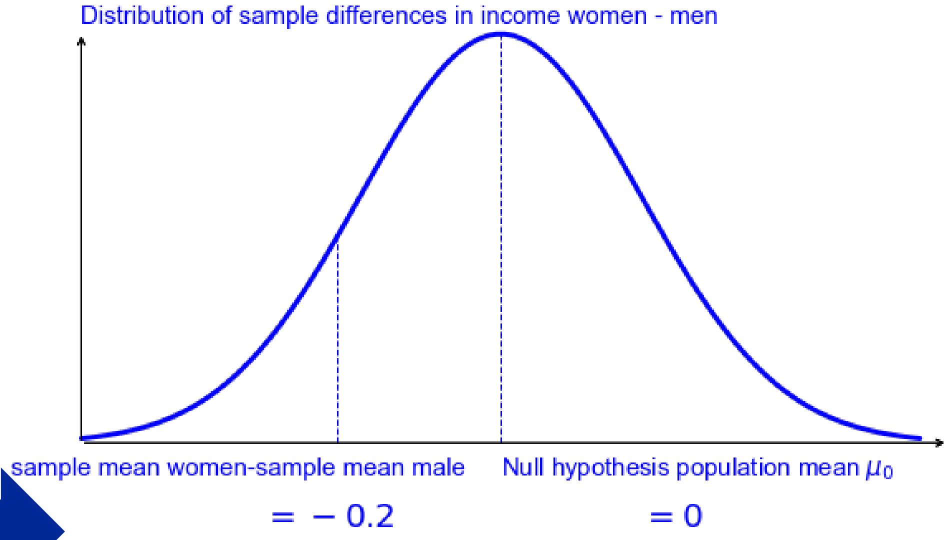
b) Führen Sie einen Test durch, um zu überprüfen, ob der Unterschied signifikant ist.

HYPOTHESENTEST: TESTSTATISTIK

HOCHSCHULE HEILBRONN

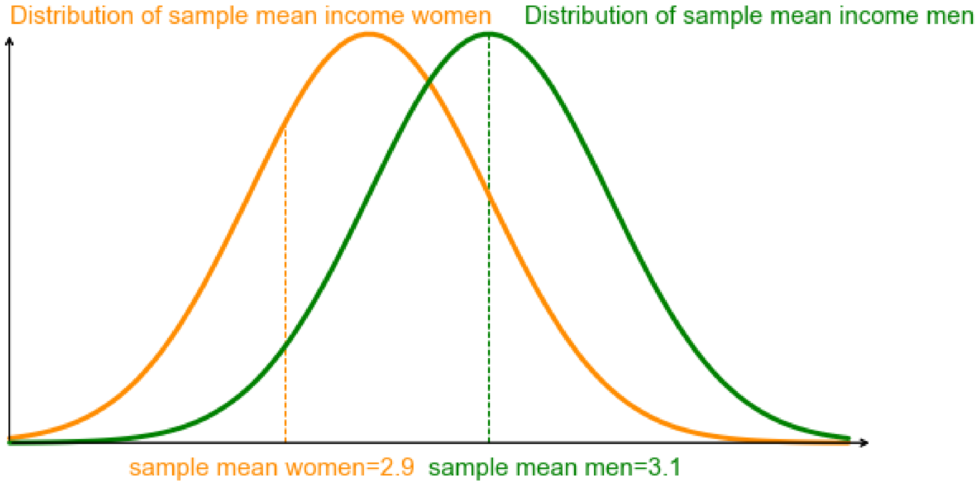
Welt 0 (H_0)

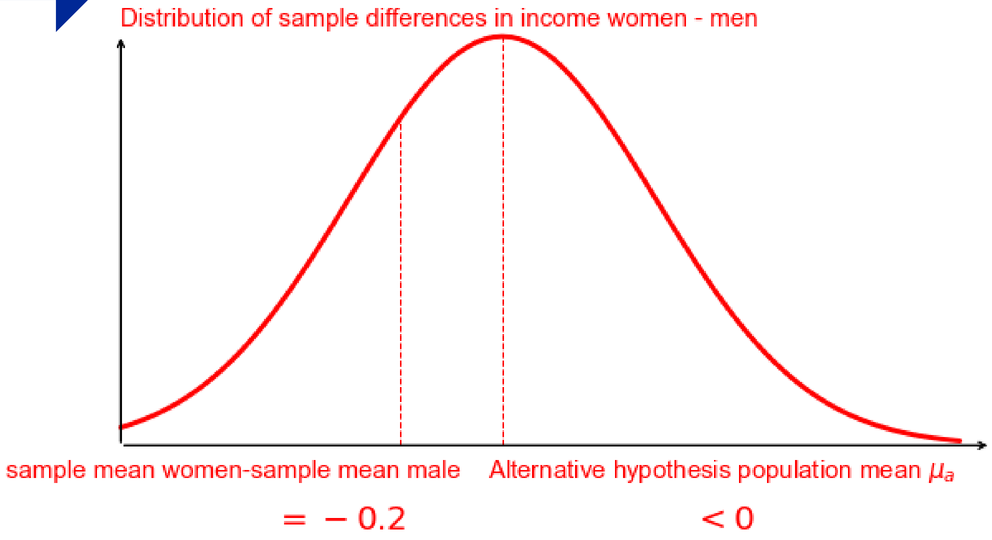




Differenzen

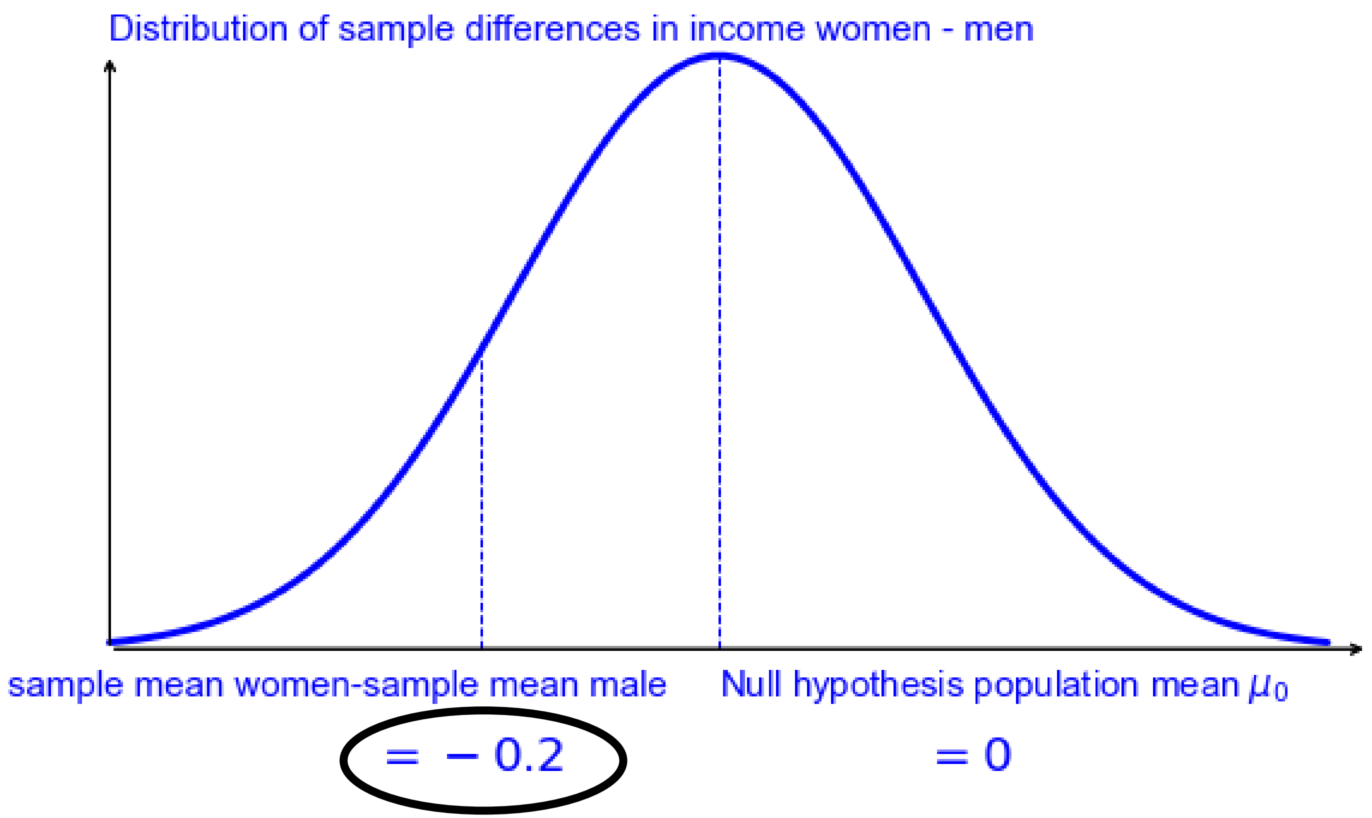
Welt 1 (H_a)





HYPOTHESENTEST: TESTSTATISTIK

Wenn Nullhypothese = wahr, dann ist die wahre Verteilung:



Teststatistik = z-Wert der Differenz der Stichprobenmittelwerte

HYPOTHESENTEST: TESTSTATISTIK

HOCHSCHULE HEILBRONN

Beispiel (Fortsetzung). In einer Stichprobe mit 35 Frauen und 44 Männer ist das monatliche Durchschnittseinkommen der Frauen $2.9\,T€$ $(Varianz = 0.7)$ und das der Männer $3.1\,T€$ $(Varianz = 0.4)$.

Ist dies ein Beleg dafür ist, dass Frauen im Schnitt weniger verdienen?

b) Führen Sie einen Test durch, um zu überprüfen, ob der Unterschied signifikant ist.

Was wissen wir?

- Differenz Stichprobenmittel: $\bar{x}_f - \bar{x}_m = 2.9 - 3.1 = -0.2$
- Stichprobengrößen: $n_f = 35$ und $n_m = 44$
- Varianzen: $\sigma_f^2 = 0.7$ und $\sigma_f^2 = 0.4$

Gepoolter Standardfehler:

$$SE_{f-m} = \sqrt{\frac{\sigma_f^2}{n_f} + \frac{\sigma_f^2}{n_m}} = 0.17$$

HYPOTHESENTEST: TESTSTATISTIK

HOCHSCHULE HEILBRONN

b) Führen Sie einen Test durch, um zu überprüfen, ob der Unterschied signifikant ist.

Was wissen wir?

- Differenz Stichprobenmittel: $\bar{x}_f - \bar{x}_m = 2.9 - 3.1 = -0.2$
- Gepoolter Standardfehler: $SE_{f-m} = 0.17$

Teststatistik = z-Wert der Differenz der Stichprobenmittel:

$$z(\mathit{Differenz\ der\ Stichprobenmittel}) = \frac{\mathit{Differenz\ der\ Stichprobenmittel} - \mathit{Pop.Diff.unter}\ H_0}{\mathit{Gepoolter\ Standardfehler}}$$

$$= \frac{-0.2 - 0}{0.17} = -1.17$$

HYPOTHESENTEST: P-WERT

HOCHSCHULE HEILBRONN

Basierend auf der Teststatistik können wir den p-Wert berechnen

P-Wert
Der *p-Wert* ist die Wahrscheinlichkeit (unter der Nullhypothese), dass die Teststatistik einen mindestens so extremen Wert annimmt wie denjenigen, der beobachtet wurde.

> Der p-Wert ist also eine bedingte Wahrscheinlichkeit:

$$P(Teststatistik = wie\ beobachtet\ oder\ extremer \mid Nullhypothese\ wahr)$$

HYPOTHESENTEST: P-WERT

Beispiel (Fortsetzung). In einer Stichprobe mit 35 Frauen und 44 Männer ist das monatliche Durchschnittseinkommen der Frauen $2.9\,T€$ $(Varianz = 0.7)$ und das der Männer $3.1\,T€$ $(Varianz = 0.4)$.

Ist dies ein Beleg dafür ist, dass Frauen im Schnitt weniger verdienen?

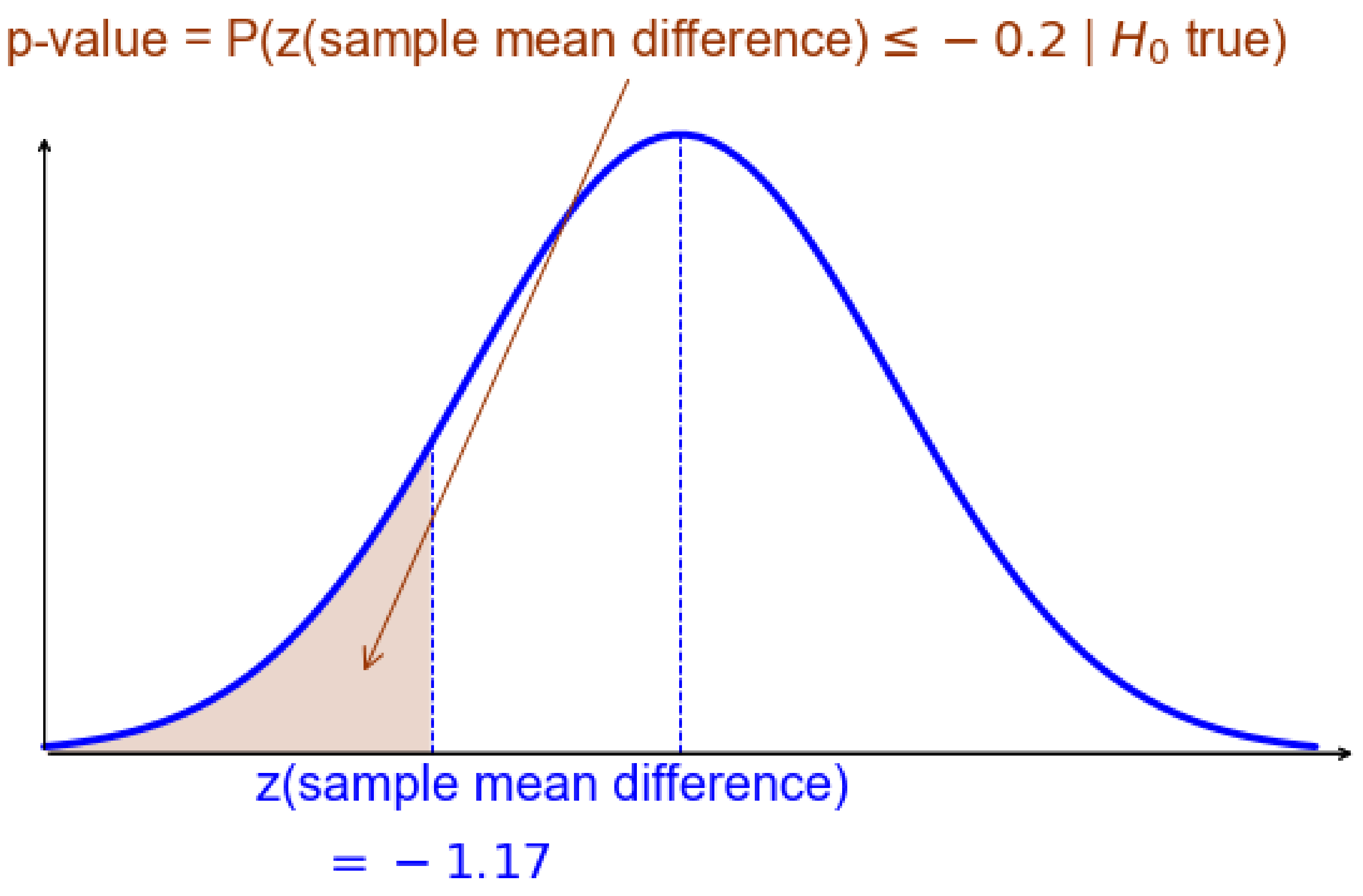
b) Führen Sie einen Test durch, um zu überprüfen, ob der Unterschied signifikant ist.

Schritt 1. Teststatistik = $z(Differenz\ Stichprobenmittel) = -1.17$ ✓

Schritt 2. Berechne p-Wert.

HYPOTHESENTEST: P-WERT

HOCHSCHULE HEILBRONN



HYPOTHESENTEST: P-WERT

HOCHSCHULE HEILBRONN

b) Führen Sie einen Test durch, um zu überprüfen, ob der Unterschied signifikant ist.

Schritt 1. Teststatistik = $z(Differenz\ Stichprobenmittel) = -1.17$ ✓

Schritt 2. Berechne p-Wert.

> Fläche Tabelle:

1.00	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577
1.10	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790
1.20	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980

> Damit:

$$\boldsymbol{p - Wert = 0.5 - 0.379 = 12.1\%}$$

HYPOTHESENTEST: P-WERT

HOCHSCHULE HEILBRONN

Was sagt uns p-Wert = 12.1%?

> Wenn p-Wert ≤ konventionelle Obergrenzen → Ablehnung Nullhypothese

p-Wert ≤ Obergrenze α (*)	Interpretation
5%	Statistisch signifikant
1%	Stark statistisch signifikant

(*) Die Obergrenze α heißt *Signifikanzniveau*

p-Wert = 12.1% > 5% → der Einkommensunterschied ist nicht statistisch signifikant

> Nullhypothese kann nicht abgelehnt werden → keine Evidenz für einen Unterschied

› ARTEN VON HYPOTHESENTESTS

ARTEN VON HYPOTHESENTESTS: EINSEITIG UND ZWEISEITIG

HOCHSCHULE HEILBRONN

> *Einseitige Tests* haben eine Richtung: mehr/weniger, höher/niedriger, …

> *Zweiseitige Tests* haben keine Richtung: Unterschied ja/nein

Beispiel. Überprüfen Sie, ob die folgenden Tests einseitig oder zweiseitig sind:

Wir möchten wissen…	Alternativhypothese (H_a)	Nullhypothese (H_0)
ob Absolventen der HS Heilbronn im Schnitt (μ_{Alter}) älter sind als 26 (μ_0)	$\mu_{Alter} > 26$	$\mu_{Alter} \leq 26$
ob sich der Durchschnitts-IQ (μ_{IQ}) von Studierenden verändert hat (langfristiger Durchschnitt: $\mu_0 = 101$)	$\mu_{IQ} \neq 101$	$\mu_{IQ} = 101$
ob die Nachfrage (μ_{Nach}) für das Produkt einer Firma abgenommen hat (langfristiger Durchschnitt: $\mu_0 = 41'000$)	$\mu_{Nach} < 41'000$	$\mu_{Nach} > 41'000$

ARTEN VON HYPOTHESENTESTS: EINSEITIG UND ZWEISEITIG

HOCHSCHULE HEILBRONN

> *Einseitige Tests* haben eine Richtung: mehr/weniger, höher/niedriger, …

> *Zweiseitige Tests* haben keine Richtung: Unterschied ja/nein

Beispiel. Überprüfen Sie, ob die folgenden Tests einseitig oder zweiseitig sind:

	Wir möchten wissen…	Alternativhypothese (H_a)	Nullhypothese (H_0)
einseitig	ob Absolventen der HS Heilbronn im Schnitt (μ_{Alter}) älter sind als 26 (μ_0)	$\mu_{Alter} > 26$	$\mu_{Alter} \leq 26$
zweiseitig	ob sich der Durchschnitts-IQ (μ_{IQ}) von Studierenden verändert hat (langfristiger Durchschnitt: $\mu_0 = 101$)	$\mu_{IQ} \neq 101$	$\mu_{IQ} = 101$
einseitig	ob die Nachfrage (μ_{Nach}) für das Produkt einer Firma abgenommen hat (langfristiger Durchschnitt: $\mu_0 = 41'000$)	$\mu_{Nach} < 41'000$	$\mu_{Nach} > 41'000$

ARTEN VON HYPOTHESENTESTS: LINKSSEITIG UND RECHTSSEITIG

HOCHSCHULE HEILBRONN

Einseitige Tests können links- oder rechtsseitig sein:

- *Linksseitiger Test* ("**<**" in Alternativhypothese) → Ablehnungsbereich/p-Wert links
- *Rechtsseitiger Test* ("**>**" in Alternativhypothese) → Ablehnungsbereich/p-Wert rechts

ARTEN VON HYPOTHESENTESTS: EINSTICHPROBEN- UND ZWEISTICHPROBEN

HOCHSCHULE HEILBRONN

Vergleich Mittelwert **einer Population** mit einer Zahl → **Einstichproben-Test**

Beispiele. Fragen:

> Hat sich der durchschnittliche IQ (bisher: 101 = μ_0) von Studierenden verändert?

> Verbringen Studierende mehr als 4h / Tag (= μ_0) in sozialen Netzwerken?

Vergleich von **zwei Populationen** → **Zweistichproben-Test**

Beispiele. Fragen:

> Gibt es einen Unterschied im Einkommen von Frauen und Männern?

> Sind Nutzer von sozialen Netzwerken glücklicher?

ARTEN VON HYPOTHESENTESTS: Z-TESTS

HOCHSCHULE HEILBRONN

Tests werden oft nach der Verteilung der Teststatistik benannt

- Z-Tests sind eine Familie von Tests mit z-verteilter Teststatistik (= standardnormalverteilt)

Wir fokussieren uns auf z-Tests, aber für den fortgeschrittenen Kurs:

- Wenn die **Varianz unbekannt** ist und geschätzt werden muss → t-verteilte Teststatistik
- Damit: t-Test → Einziger Unterschied zum Vorgehen beim z-Test:

  Benutze t-Tabelle (statt z-Tabelle), um p-Wert / kritischen Wert zu bestimmen

ARTEN VON HYPOTHESENTESTS: Z-TESTS

HOCHSCHULE HEILBRONN

Voraussetzungen Ein- und Zweistichproben Z-Test
(Z1) Die Daten sind quantitativ (Intervall- oder Verhältnisskala).
(Z2) Die Daten sind (in jeder Gruppe) normalverteilt.
(Z3) Die Standardabweichung(en) der Population(en) ist (sind) bekannt.
(Z4) Die Messungen innerhalb der Gruppen und zwischen den Gruppen sind unabhängig.
(Z5) Keine Extremwerte (Outlier).

Anmerkung. Anteilswert-Test: Die ursprünglichen Merkmale können binär qualitativ sein

> Anteilswerte = Verhältnisskala

› BEISPIELAUFGABEN

EINSTICHPROBEN Z-TEST: BEISPIEL 1

HOCHSCHULE HEILBRONN

Das Einstiegsgehalten von Absolventen der Hochschule Heilbronn betrug im langfristigen Durchschnitt ca. 31 T€/Jahr (Standardabweichung: 21 T€). Das durchschnittliche Einstiegsgehalt in einer Stichprobe mit 44 Absolventen ist 26 T€/Jahr.

Ist das ein Beleg, dass sich das Einstiegsgehalt signifikant verschlechtert hat?

a) Handelt es sich um einen einseitigen oder zweiseitigen Test?

b) Schreiben Sie die Null- und Alternativhypothese auf.

c) Führen Sie einen Test durch, um die Frage zu beantworten.

EINSTICHPROBEN Z-TEST: BEISPIEL 2

HOCHSCHULE HEILBRONN

Bei einem Online-Anbieter waren 54 der letzten 87 Kunden unter 30.

Ist dies Evidenz, dass die Kaufrate bei unter 30 Jährigen von 50% abweicht?

a) Handelt es sich um einen einseitigen oder zweiseitigen Test?

b) Schreiben Sie die Null- und Alternativhypothese auf.

c) Führen Sie einen Test durch, um die Frage zu beantworten.

ZWEISTICHPROBEN Z-TEST
BEISPIEL 1

In einer Stichprobe mit 32 Mitarbeitern der Firma A und 35 Mitarbeitern der Firma B werden folgende Durchschnittsjahresgehälter gemessen: Firma A: 43 T€ (Varianz: 450.11), Firma B: 29.8 T€ (Varianz: 294.05).

Unterscheiden sich die Durchschnittsgehälter von Firma A und B (1%-Niveau)?

a) Handelt es sich um einen einseitigen oder zweiseitigen Test?

b) Schreiben Sie die Null- und Alternativhypothese auf.

c) Führen Sie einen Test durch, um die Frage zu beantworten.

ZWEISTICHPROBEN Z-TEST
BEISPIEL 2

HOCHSCHULE HEILBRONN

In einer Stichprobe sind 63 von 90 Frauen und 27 von 52 Männer umweltbewusst. Die Stichprobenvarianz entspricht jeweils ungefähr der Populationsvarianz.

Sind Frauen umweltbewusster?

a) Handelt es sich um einen einseitigen oder zweiseitigen Test?

b) Schreiben Sie die Null- und Alternativhypothese auf.

c) Führen Sie einen Test durch, um die Frage zu beantworten.

Z-TESTS: ZUSAMMENFASSUNG

Test auf	Einstichproben-Tests Teststatistik	Zweistichproben-Tests Teststatistik
Mittelwert	$z = \dfrac{\bar{x} - \mu_0}{SE_{\bar{X}}}$	$z = \dfrac{\bar{x}_A - \bar{x}_B - \mu_0}{SE_{A-B}}$
Anteilswert	$z = \dfrac{\bar{p} - \pi_0}{\sqrt{\dfrac{\pi_0\,(1-\pi_0)}{n}}}$	$z = \dfrac{\bar{p}_A - \bar{p}_B - \pi_0}{SE_{A-B}}$

› TESTEN VON ZUSAMMENHÄNGEN

TESTEN VON ZUSAMMENHÄNGE

HOCHSCHULE HEILBRONN

Im Falle von bivariaten Daten, möchten wir häufig überprüfen, ob ein Zusammenhang da ist

Zwei kategoriale Variablen → Chi²-Test

> Beantwortet die Frage, ob zwei Variablen stochastisch abhängig sind

Beispiel. Frage: Hängen Krebserkranungen mit Rauchen zusammen?

Zwei nummerische Variablen → Korrelationstest

> Beantwortet die Frage, ob zwei Variablen korreliert sind

Beispiel. Frage: Sind Marketingausgaben mit Verkaufsumsätzen korreliert?

TESTEN VON ZUSAMMENHÄNGEN: CHI2-TEST (KATEGORIALE VARIABLEN)

Beispiel. Stichprobe mit 50 Männern und 52 Frauen. Jeder Teilnehmer gibt an, ob er/sie mehr als 2h/Tag auf Social-Media-Plattformen verbringt oder nicht:

	Männer	Frauen	Gesamt
$\leq$ 2h	33	27	60
> 2h	17	25	42
Gesamt	50	52	102

Sind Geschlecht und Social-Media-Nutzungsdauer unabhängig (5%)?

a) Schreiben Sie die Null- und die Alternativhypothese auf.

b) Führen Sie einen Test durch, um die Frage zu beantworten.

TESTEN VON ZUSAMMENHÄNGEN: CHI2-TEST (KATEGORIALE VARIABLEN)

HOCHSCHULE HEILBRONN

Sind Geschlecht und Social-Media-Nutzungsdauer unabhängig (5%)?

a) Schreiben Sie die Null- und die Alternativhypothese auf.

$$H_0: Social\!-\!Media\!-\!Nutzungsdauer\ und\ Geschlecht\ sind\ unabhängig$$

$$H_a: Social\!-\!Media\!-\!Nutzungsdauer\ und\ Geschlecht\ sind\ abhängig$$

TESTEN VON ZUSAMMENHÄNGEN: CHI2-TEST (KATEGORIALE VARIABLEN)

HOCHSCHULE HEILBRONN

Sind Geschlecht und Social-Media-Nutzungsdauer unabhängig (5%)?

b) Führen Sie einen Test durch, um die Frage zu beantworten.

Idee Chi2-Test:

> Berechne erwartete Häufigkeiten bei Unabhängigkeit

> Großer Unterschied erwartete und beobachtete Häufigkeit → Abhängigkeit

> Kleiner Unterschied erwartete und beobachtete Häufigkeit → Unabhängigkeit

TESTEN VON ZUSAMMENHÄNGEN: CHI2-TEST (KATEGORIALE VARIABLEN)

HOCHSCHULE HEILBRONN

b) Führen Sie einen Test durch, um die Frage zu beantworten.

Schritt 1. Berechne erwartete Häufigkeiten bei Unabhängigkeit

Beispiel. Spalte 1, Zeile 1. Beobachtet:

> P(Geschlecht = Mann) = 50/102

> P(Nutzung = " $\leq 2h$") = 60/102

	M	**F**	$\sum$
$\leq$ 2h	33	27	60
> 2h	17	25	42
$\sum$	50	52	102

Für Unabhängigkeit muss gelten (siehe stochastische Unabhängigkeit!):

$$P(G = Mann, N = "\leq 2h") = P(G = Mann) \times P(N = "\leq 2\text{h}") = \frac{50 \times 60}{102}$$

TESTEN VON ZUSAMMENHÄNGEN: CHI2-TEST

HOCHSCHULE HEILBRONN

b) Führen Sie einen Test durch, um die Frage zu beantworten.

Schritt 1. Berechne erwartete Häufigkeiten bei Unabhängigkeit

Beobachtete Häufigkeiten:

	M	F	$\sum$
$\leq$ 2h	33	27	60
> 2h	17	25	42
$\sum$	50	52	102

Erwartete Häufigkeiten bei Unabhängigkeit:

	M	F	$\sum$
$\leq$ 2h	$E_{11} = \frac{60 \cdot 50}{102} = \mathbf{29.41}$	$E_{12} = \mathbf{30.59}$	60
> 2h	$E_{21} = \mathbf{20.59}$	$E_{22} = \mathbf{21.41}$	42
$\sum$	50	52	102

TESTEN VON ZUSAMMENHÄNGEN: CHI2-TEST

HOCHSCHULE HEILBRONN

b) Führen Sie einen Test durch, um die Frage zu beantworten .

Schritt 1. Berechne erwartete Häufigkeiten bei Unabhängigkeit ✓

Schritt 2. Berechne Teststatistik:

$$\chi^2 = \sum \frac{\left(Beobachtet_{ij} - Erwartet_{ij}\right)^2}{Erwartet_{ij}}$$

$$\chi^2 = \frac{(33 - 29.41)^2}{29.41} + \frac{(17 - 20.59)^2}{20.59} + \frac{(27 - 30.59)^2}{30.59} + \frac{(25 - 21.41)^2}{21.41} = 2.09$$

TESTEN VON ZUSAMMENHÄNGEN: CHI2-TEST

HOCHSCHULE HEILBRONN

b) Führen Sie einen Test durch, um die Frage zu beantworten.

Schritt 1. Berechne erwartete Häufigkeiten bei Unabhängigkeit ✓

Schritt 2. Berechne Teststatistik. $\chi^2 = 2.09$ ✓

Schritt 3. Bestimme kritischen Wert:

> Signifikanzniveau: $\alpha = 5\%$ (siehe Aufgabe)

> Freiheitsgrade (Degrees of freedom df):

$$df = (Anzahl\ Zeilen - 1) \times (Anzahl\ Spalten - 1) = 1$$

Chi2-Tabelle:

DF / α	0.1	0.05	0.025
1	2.70554	3.84146	5.02389
2	4.60517	5.99146	7.37776
3	6.25139	7.81473	9.34840
4	7.77944	9.48773	11.14329
5	9.23636	11.07050	12.83250

$$\boldsymbol{Kritischer\ Wert:\ \chi^2_{0.05,1} \approx 3.84}$$

TESTEN VON ZUSAMMENHÄNGEN: CHI2-TEST

HOCHSCHULE HEILBRONN

b) Führen Sie einen Test durch, um die Frage zu beantworten.

Schritt 1. Berechne erwartete Häufigkeiten, wenn unabhängig ✓

Schritt 2. Berechne Teststatistik. $\chi^2 = 2.09$ ✓

Schritt 3. Bestimme kritischen Wert. Kritischer Wert = 3.84 ✓

Schritt 4. Vergleiche Teststatistik und kritischer Wert (immer rechtsseitig):

$$\chi^2 = 2.09 < 3.84 = kritischer\ Wert$$

Wir können die Nullhypothese nicht auf 5% Niveau ablehnen

→ keine Evidenz für Abhängigkeit

CHI2-TEST: ZUSAMMENFASSUNG

HOCHSCHULE HEILBRONN

Null- und Alternativhypothese	Teststatistik
$H_0: X, Y \ stochastisch \ unabhängig$ $H_a: X, Y stochastisch \ abhängig$	$$\chi^2 = \sum \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$ > O_{ij}: Beobachtete Häufigkeit Zeile i, Spalte j > E_{ij}: Erwartete Häufigkeit Zeile i, Spalte j

CHI2-TEST: VORAUSSETZUNGEN

HOCHSCHULE HEILBRONN

Voraussetzungen Chi²-Unabhängigkeitstest
(C1) Beide Variablen sind kategorial (Nominal- oder Ordinalskala).
(C2) Die Stichprobe ist hinreichend groß ($n > 50$).
(C3) Die erwarteten Zellhäufigkeiten sind > 5.
(C4) Die Messungen innerhalb der Gruppen und zwischen den Gruppen sind unabhängig.

› APPENDIX: KRITISCHE WERTE

HYPOTHESENTEST: KRITISCHE WERTE

Eine Alternative zum p-Wert sind sogen. *kritische Werte*

> Beide Wege sind äquivalent

Vorgehen mit kritischen Werten:

> Fixiere ein Signifikanzniveau (5%, 1%)

> Bestimme Bereich für Teststatistik, um das Signifikanzniveau zu erreichen

HYPOTHESENTEST: KRITISCHE WERTE

HOCHSCHULE HEILBRONN

Beispiel (Fortsetzung). In einer Stichprobe mit 35 Frauen und 44 Männer ist das monatliche Durchschnittseinkommen der Frauen $2.9\, T€$ $(Varianz = 0.7)$ und das der Männer $3.1\, T€$ $(Varianz = 0.4)$.

Ist dies ein Beleg dafür ist, dass Frauen im Schnitt weniger verdienen?

b) Führen Sie einen Test durch, um zu überprüfen, ob der Unterschied signifikant ist.

Schritt 1. Teststatistik = $z(Differenz\, Stichprobenmittel) = -1.17$ ✓

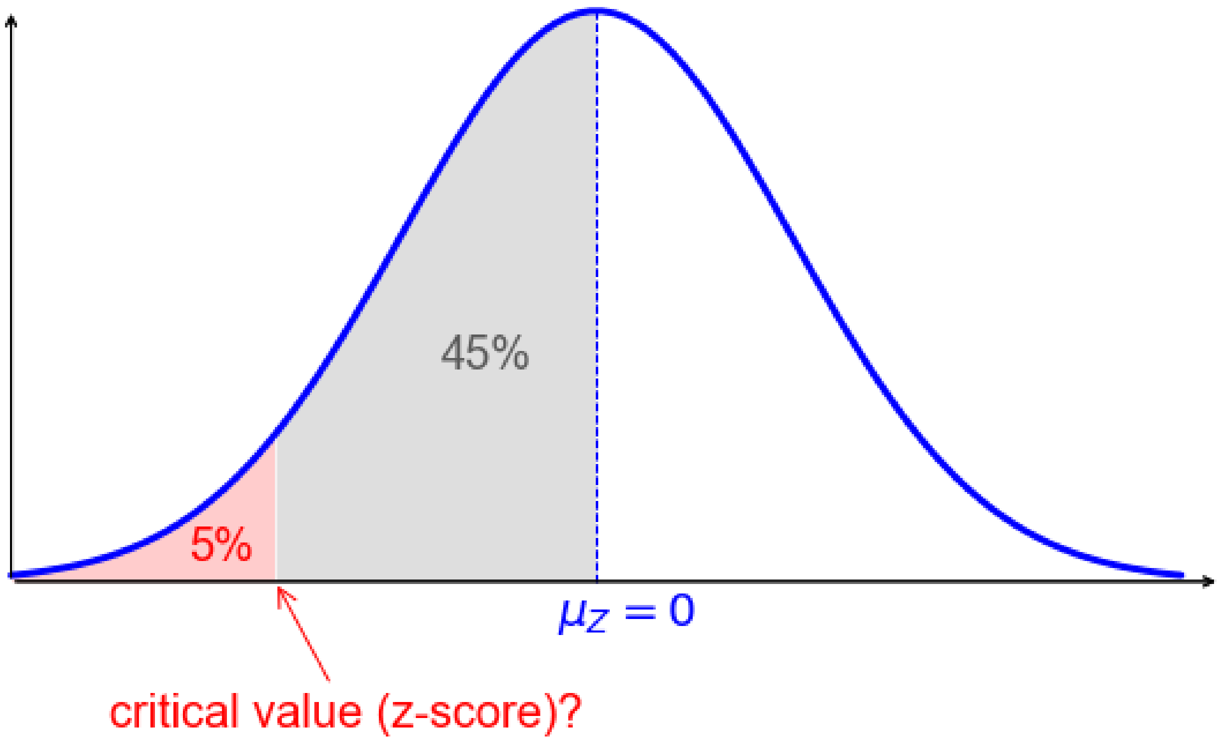
Schritt 2. Berechne kritischen Wert für Signifikanzniveau 5%.

HYPOTHESENTEST: KRITISCHE WERTE

HOCHSCHULE HEILBRONN

Schritt 2. Berechne kritischen Wert für Signifikanzniveau 5%.

> Kritischer Wert ist links (H_a: Stichprobenmittel Frauen < Stichprobenmittel Männer)



KRITISCHE WERTE

> Z-Wert für 0.45:

$$z_5 = 1.645$$

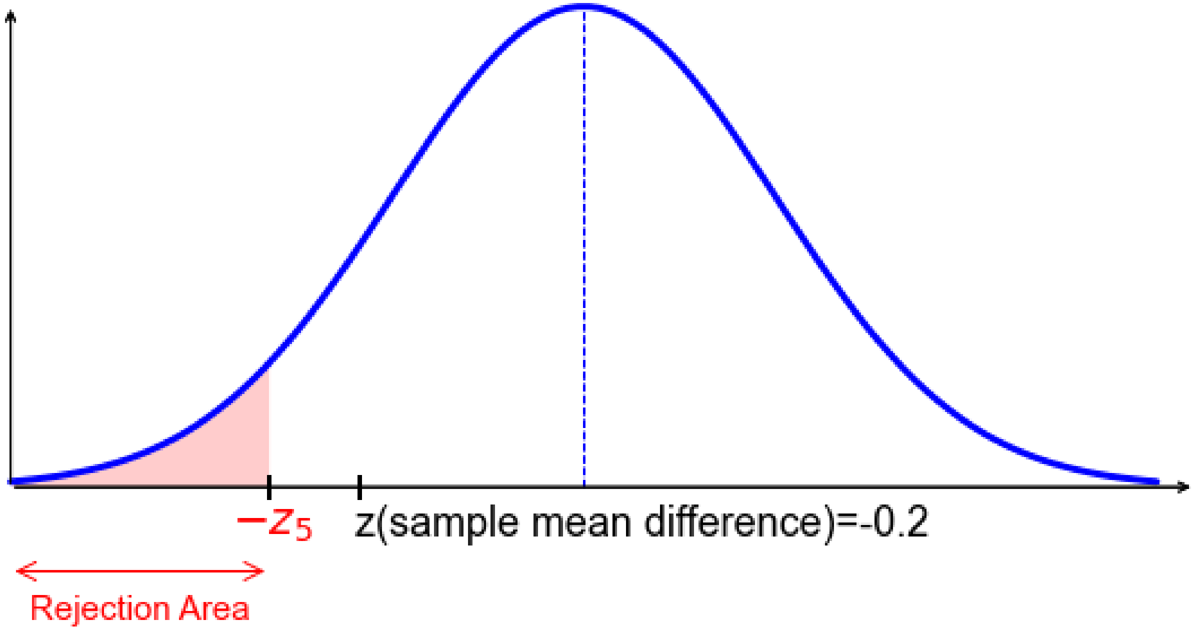
> Linksseitig:

$$-z_5 = -1.645$$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.00	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.10	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.20	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.30	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.40	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.50	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.60	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.70	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.80	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.90	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.00	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.10	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.20	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.30	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.40	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.50	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.60	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.70	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.80	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.90	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.00	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.10	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.20	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.30	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.40	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.50	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.60	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.70	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.80	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.90	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.00	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

HYPOTHESENTEST: KRITISCHE WERTE

Schritt 3. Vergleiche Teststatistik und kritische Werte:



- Teststatistik: $z(Differenz) = -1.17 > -z_5 = -1.645$
- Damit können wir die Nullhypothese nicht auf 5% Niveau ablehnen

P-WERT UND KRITISCHE WERTE: ZUSAMMENFASSUNG RECHENSCHRITTE

HOCHSCHULE HEILBRONN

P-Wert:

1. Berechne Teststatistik (z-Wert)
2. Berechne p-Wert
3. Vergleiche p-Wert mit konventionellen Obergrenzen für das Signifikanzniveau

Kritische Werte:

1. Berechne Teststatistik (z-Wert)
2. Berechne kritischen Wert für α
3. Vergleiche Teststatistik mit kritischem Wert

THE END!

Please refer any questions to:
Prof. Dr. Florian Kauffeldt
Faculty of International Business
Florian.kauffeldt@hs-heilbronn.de