

SW 1 - Einführung in R

Aufgabe 1.2

a) Bilden Sie einen Vektor x mit den Zahlen 4, 2, 1, 3, 3, 5, 7.
↳ $x <- c(4, 2, 1, 3, 3, 5, 7)$

b) Wählen Sie mit "R" den dritten Wert aus.
↳ $x[3]$

c) Wählen Sie mit "R" den ersten & vierten Wert aus.
↳ $x[c(1, 4)]$

d) Bestimmen Sie die Länge des Vektors x .
↳ $length(x)$

e) Was macht der Befehl $x+2$?

↳ addiert zu allen Werten im Vektor +2 $\Rightarrow x <- c(6, 4, 3, 5, 5, 7, 9)$

Sum($x+2$)

↳ gibt Gesamtsumme des Vektors, nachdem zu jedem Wert +2 gemacht wurde
 $x <= 3$

↳ erzeugt Vektor mit Länge 3. Für alle Werte die kleiner sind: false, sonst true. \Rightarrow FALSE TRUE
TRUE TRUE...

$x[x <= 3]$

↳ $x[...]$ wählt Elemente aus Vektor x aus. Es werden Werte mit TRUE ausgewählt ausschließlich.

SORT(x)

↳ Werte von x werden der Größe nach aufsteigend geordnet.

order(x)

↳ gibt Stellen an, wo sich die Werte von x der Größe nach befinden.

Aufgabe 1.3

Gegeben sind folgende Temperaturen in Grad Fahrenheit ($^{\circ}\text{F}$): 51.9, 51.8, 51.9, 53.

a) Bilden Sie den Vektor fahrenheit mit diesen Werten.

↳ $fahrenheit <- c(51.9, 51.8, 51.9, 53)$

b) Berechnen Sie die Temperaturen in Grad Celsius ($^{\circ}\text{C}$) um. Formel: $C = \frac{5}{9}(F - 32)$, Bilden Sie den Vektor celsius .
↳ $celsius <- 5/9 * (fahrenheit - 32)$

c) Gegeben sind weitere Temperaturen: 48, 48.2, 48, 48.7. Bestimmen Sie die Differenz zu den ursprünglichen Temperaturen.

↳ $fahrenheit_2 <- c(48, 48.2, 48, 48.7)$

$fahrenheit_3 <- fahrenheit - fahrenheit_2$

$fahrenheit_3$

Aufgabe 1.4

Wir haben 6 Personen mit kg: 60, 72, 57, 90, 95, 72 & m: 1.75, 1.80, 1.65, 1.90, 1.74, 1.91.
Wie wollen den BMI berechnen. Formel: $\frac{\text{gewicht}}{\text{grösse}^2}$

a) Berechnen Sie den BMI der 6 Personen gleichzeitig.

↳ $weight <- c(60, 72, 57, 90, 95, 72)$

$height <- c(1.75, 1.80, 1.65, 1.90, 1.74, 1.91)$

$bmi <- weight / height^{1/2}$ \Rightarrow dann: "bmi" eingeben.

- 1 Einführung in R
- 2 Deskriptive Statistik in \mathbb{R}^1
- 3 Histogram und desk. Statistik in \mathbb{R}^2
- 4 Korrelation & Wahrscheinlichkeitslehre
- 5 Zufallsvariablen, Wahrscheinlichkeitsverteilung
- 6 Bedingte Wahrscheinlichkeit
- 7 Normalverteilung
- 8 Zentraler Grenzwertsatz
- 9 Hypothesentest
- 10 Vertrauensintervall
- 11 Lineare Regression
- 12 Multiple Lineare Regression
- 13 Variablenelektion

Aufgabe 1.5

a) Was macht der Befehl $\text{seq}(\dots)$? bildet eine Folge von Zahlen

$\hookrightarrow \text{seq}(\text{from} = 3, \text{to} = 10, \text{by} = 2)$ \text{beginn damit} \text{schritt Länge}

$\text{\#\# [1] } 3 \ 5 \ 7 \ 9$ \text{hört damit auf, sofern das möglich ist}

auch möglich: $\text{seq}(3, 10, 2)$
gleicher Output.

$\hookrightarrow \text{seq}(\text{from} = 3, \text{to} = 10, \text{length.out} = 10)$ \text{macht den Abstand gleichmäßig}

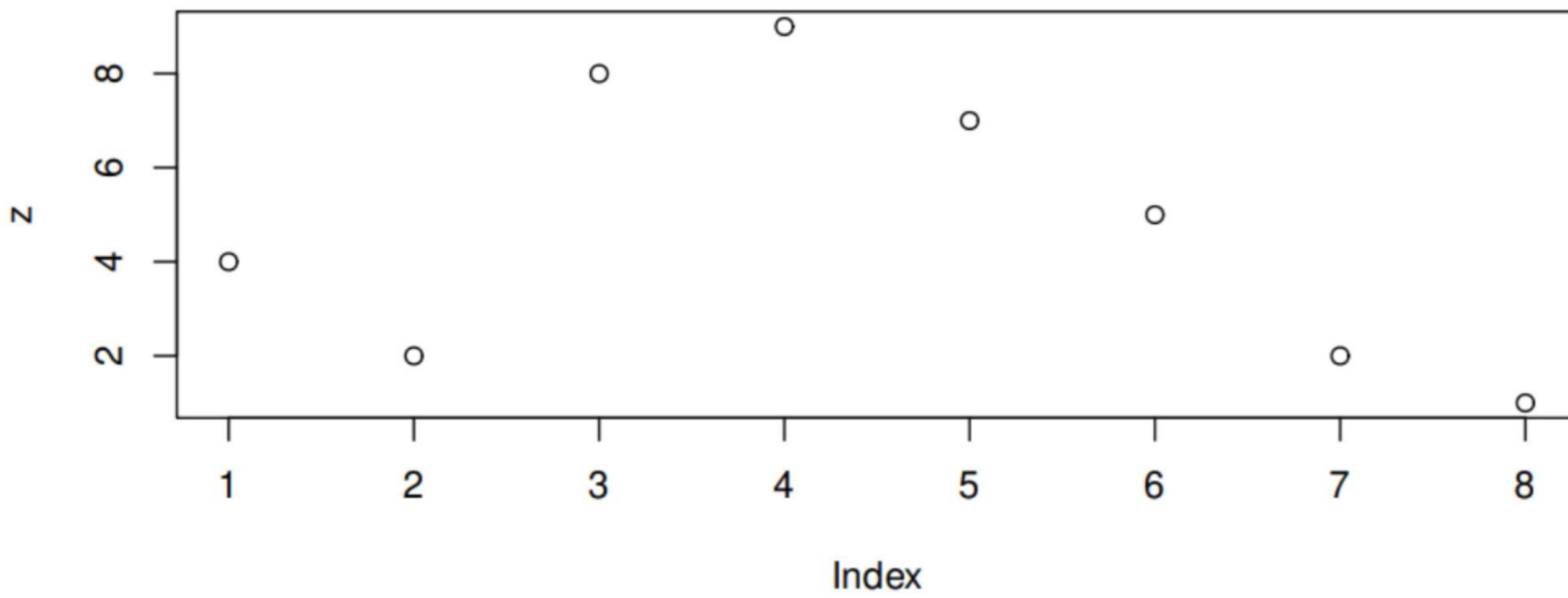
$\text{\#\# [1] } 3.000000 \ 3.777778 \ 4.555556 \ 5.333333 \ 6.111111 \ 6.888889$

$\text{\#\# [7] } 7.666667 \ 8.444444 \ 9.222222 \ 10.000000$

c) Plots spielen in der Statistik eine wichtige Rolle. Der folgende Plot ist zwar sehr einfach zu erstellen, sieht aber auch etwas gar schmucklos aus.

```
z <- c(4, 2, 8, 9, 7, 5, 2, 1)
```

```
plot(z)
```



d) Ändern Sie im Befehl diese Parameter ab! Erläutern Sie, was die machen.

```
plot(z,  
      type = "l",  
      col = "blue",  
      lty = 2, \rightarrow Linientyp, 1: —, 2: ....,  
      main = "Haupttitel",  
      xlab = "Ein paar Zahlen",  
      ylab = "Andere Zahlen")
```

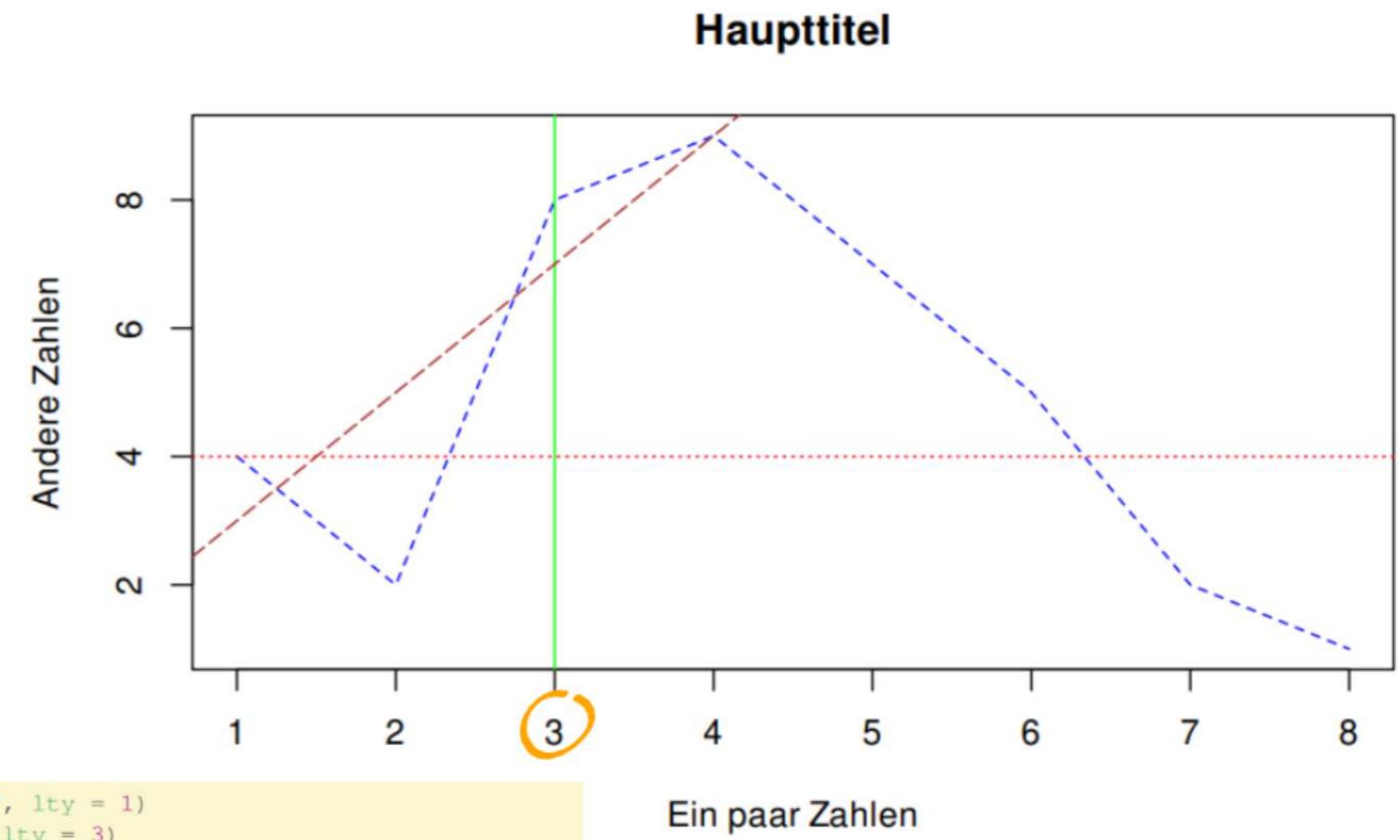
PYTHON EINFÜHRUNG

Nominal: Nur Gleichheit/Ungleichheit \rightarrow Geschlecht, Nationalität

Ordinal: Zusätzlich natürliche Reihenfolge \rightarrow Schulnoten 1-6

Metrisch: Zusätzlich mit Zahlen rechenbar \rightarrow Alter, Einkommen, etc.

```
abline(v = 3, col = "green", lty = 1)  
abline(h = 4, col = "red", lty = 3)  
abline(a = 1, b = 2, col = "brown", lty = 5)
```



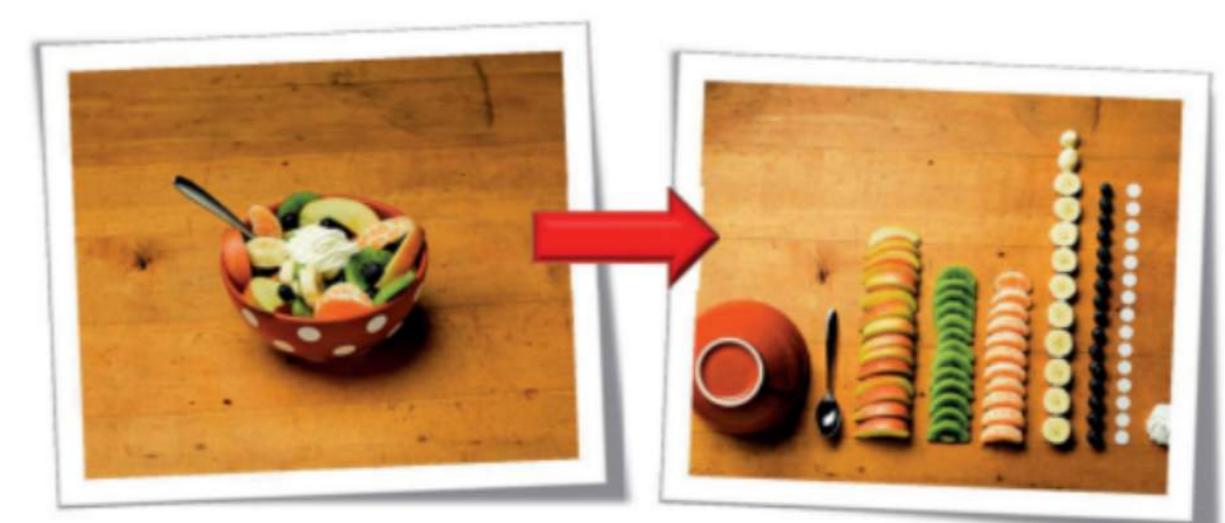
SW2 - EINDIMENSIONALE DESKRIPTIVE STATISTIK

- Was heißt eindimensional? → z.B. nur ein Array mit einer Form von DATENSÄTZE (z.B. nur Körpergrößen oder nur Gewicht.)
↳ zweidimensional: z.B. Gewicht & Größe gemeinsam. Meistens in Form einer Tabelle.

- Ziel der deskriptiven Statistik: Daten zusammenfassen & grafische Darstellung dieser Daten. z.B. wie beim Fruchtsalat:

- \bar{x} bezeichnet in der Statistik den Durchschnitt aus x_n Messungen

$$\hookrightarrow \bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} = \text{arithmetisches Mittel}$$



↑ "Wo ist die Mitte der Daten?"

- ↳ in R macht man das über `mean(...)`

```
waageA <- c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04,
          79.97, 80.05, 80.03, 80.02, 80.00, 80.02)
```

```
mean(waageA)
## [1] 80.02077
```

- Streung: z.B. bei Noten in einer Klasse: Klasse A hatte 2; 6; 3; 5 } Durchschnitt bei beiden Klassen wäre
B hatte 4; 4; 4; 4; 4.

↳ für die mittlere absolute Streung kann man die Streuungsweite in 11 setzen, damit sich negative Werte nicht aufheben.

Anstatt:
$$\frac{(2-4)+(6-4)+(3-4)+(5-4)}{4} = \frac{-2+2-1+1}{4} = \frac{0}{4} = 0$$

$$\frac{|(2-4)|+|(6-4)|+|(3-4)|+|(5-4)|}{4} = \frac{2+2+1+1}{4} = \underline{\underline{1.5}} \quad A$$

Ist die empirische V. (8 damit die Var(x)) gross, so ist die Streung der Messwerte um das arithmetische Mittel gross.

- weil mathematisch die Streung nicht immer korrekt ist, verwendet man empirische Varianz.

$$\text{Var}(x) = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1} \quad \text{Var}(\dots)$$

muss immer erfolgen nachher

↳ für die STANDARTabweichung: $\sqrt{\text{Var}(x)}$

→ Wieso macht man das? Damit man wieder bei der gleichen Einheit wie die Daten sind.

```
var(waageA)
## [1] 0.000574359
sd(waageA)
## [1] 0.02396579
```

- Median = Hälfte der Messwerte unter oder gleich diesem Wert. Andere Hälfte ist gleich oder über diesem Wert.

↳ Bei zwei gleichen Werten oder geraden Anzahl Messwerten: Durchschnitt nehmen.

↳ in R über `median(...)`

```
median(waageA)
## [1] 80.03
waageB <- c(80.02, 79.94, 79.98, 79.97, 79.97, 80.03, 79.95, 79.97)
median(waageB)
## [1] 79.97
```

$$\frac{79.97 + 79.97}{2} = \underline{\underline{79.97}}$$

↑ Median

- Was ist besser, Median oder arithmetisches Mittel? Der Median ist robust, d.h. wird viel weniger stark durch extreme Beobachtungen beeinflusst.

Das arithmetische Mittel wird durch Veränderung sehr stark beeinflusst.

↳ deshalb: es kommt darauf an!

- ⚠ QUARTILE ≠ QUATILE!!** → oberes Quartil für die obere $\frac{1}{4}$ des Medians, unteres Quartil für untere $\frac{1}{4}$ des Medians. → in R ist es der Befehl `quantile(waageA, p=0.25, Type 2)` → unteres Quartil (`waageA, p=0.75, Type 2`) → oberes Quartil

- QUARTILSDIFFERENZ = Ist ein Streuungsmass für die Daten (oberes Quartil - unteres Quartil).

Dieses Mass, umso näher liegt die Hälfte aller Werte um den Median & umso kleiner ist Streuung.

↳ in R über `IQR(...)`

Je kleiner

→ Quantitative Daten = actually gemessene Zahlen (z.B. Größe & Gewicht)
→ Qualitative Daten = andere Werte (z.B. Geschlecht & Nationalität)

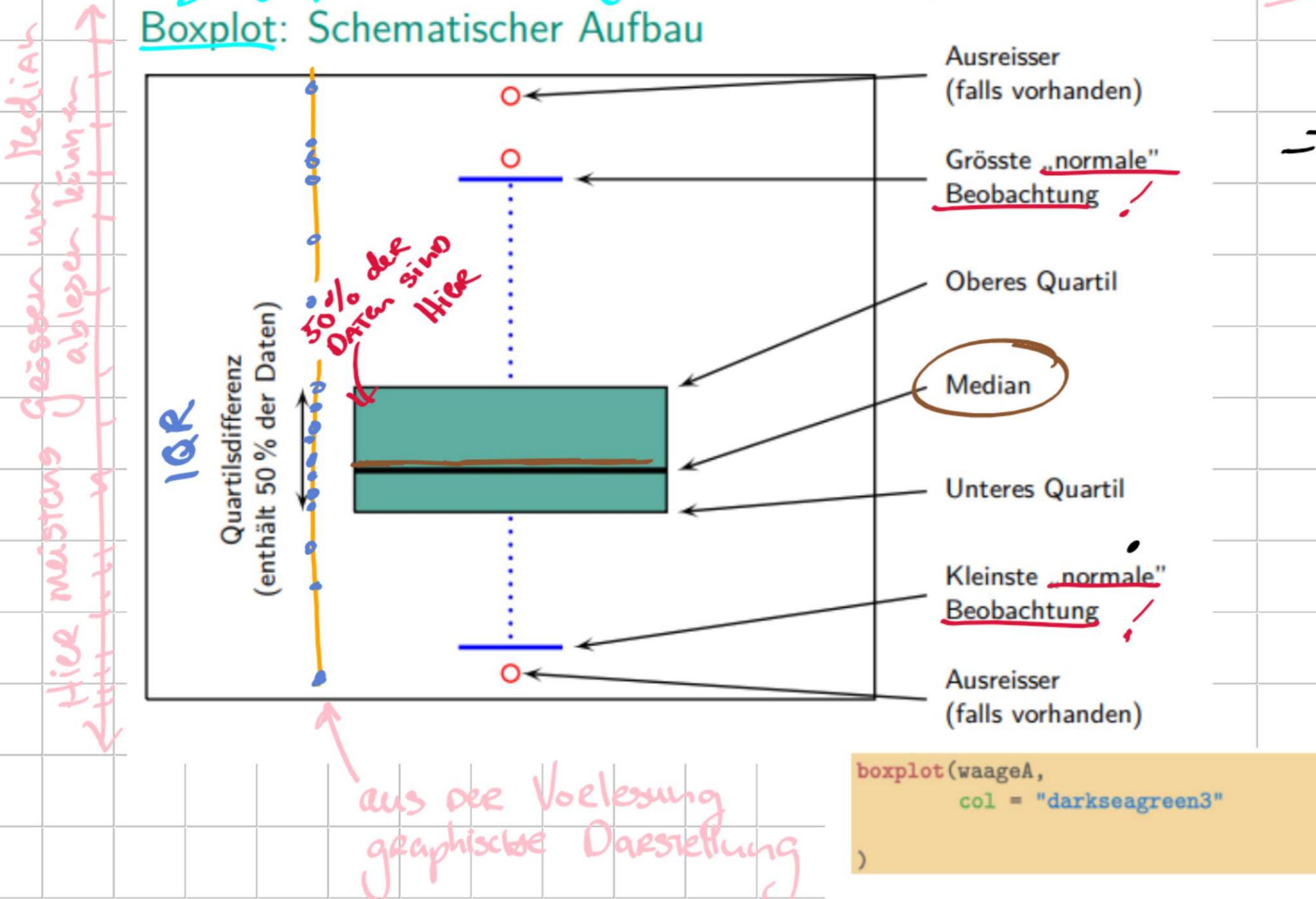
Syntax für das untere Quartil: p=0.25

```
quantile(waageA, p = 0.25, type = 2)
## 25%
## 80.02
quantile(waageB, p = 0.25, type = 2)
## 25%
## 79.96
# Syntax für das obere Quartil: p=0.75
```

- Boxplot - Fragen kommen immer an der Prüfung!

↳ = grafische Darstellung von Median & Quartilen

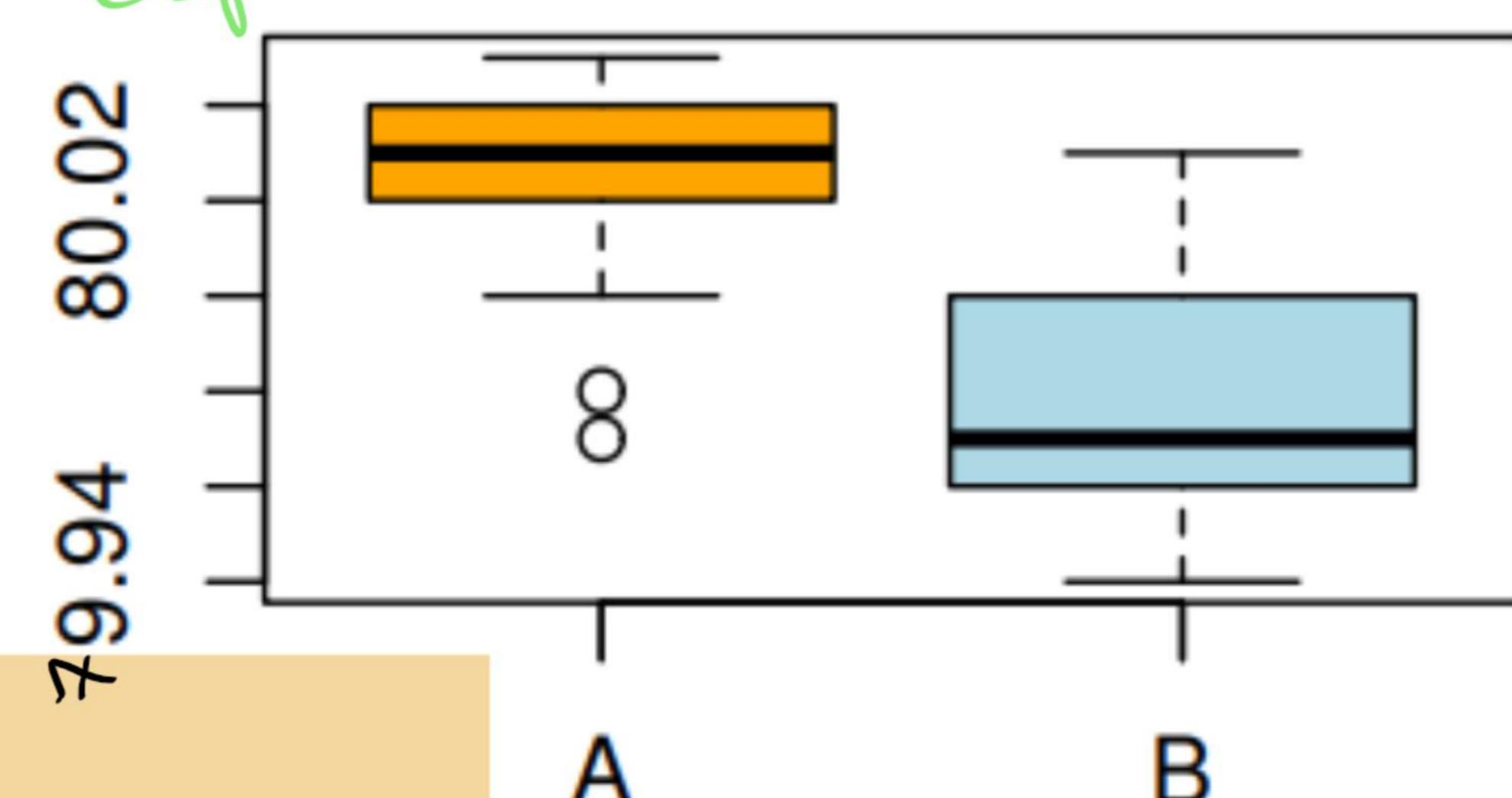
Boxplot: Schematischer Aufbau



→ kann auch waagrecht sein

→ in R heißt es `boxplot(waageA, col = "darkseagreen3")`

Bsp:

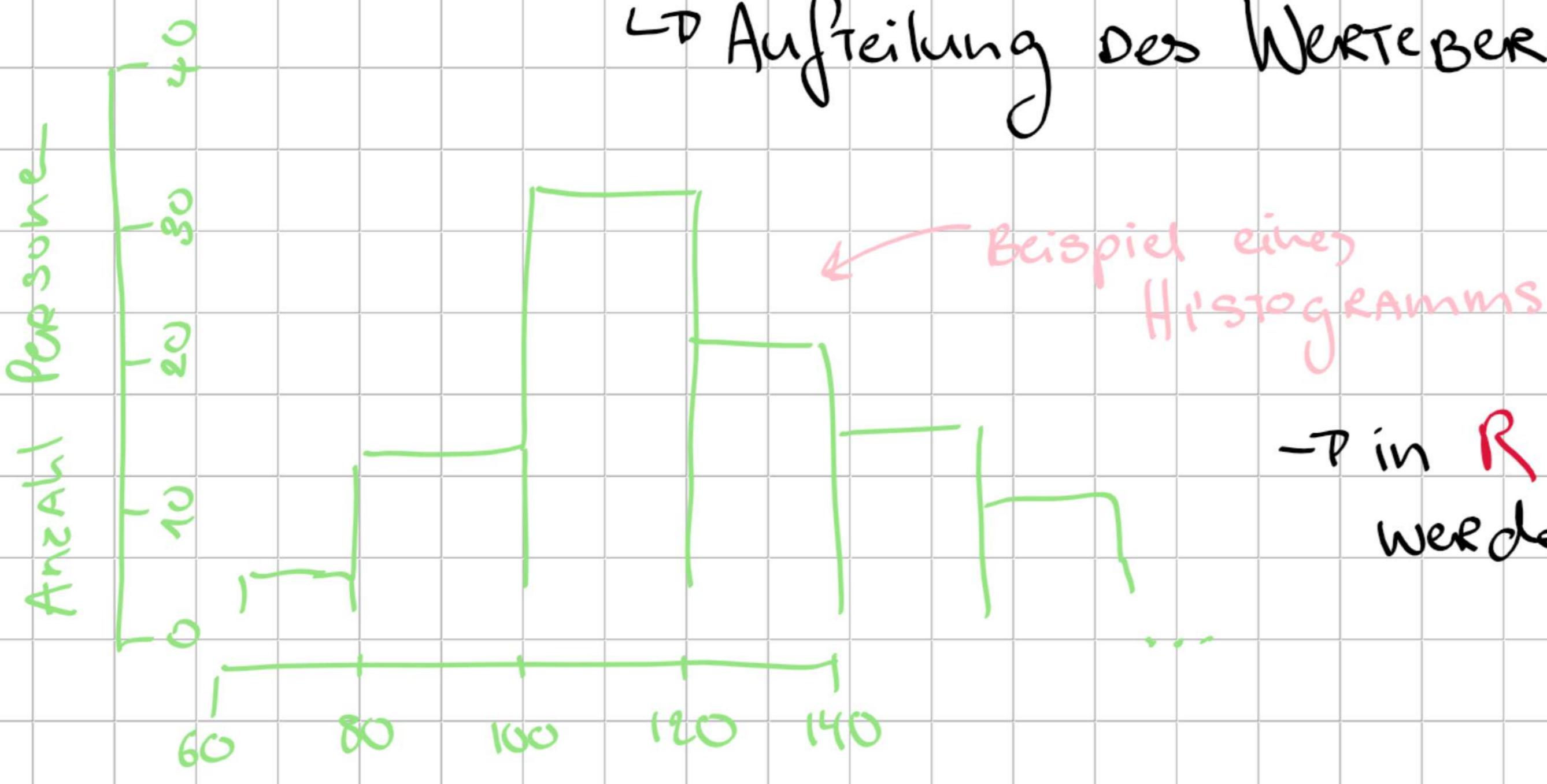


Waage B hat eine größere Streuung, ist aber ungenauer als A.

SW3 - HISTOGRAMM, ZWEIDIMENSIONALE STATISTIK

- Histogramm = grafischer Überblick über auftretende Werte. Höhe der Balken entspricht Anzahl der Beobachtungen in einer Klasse.

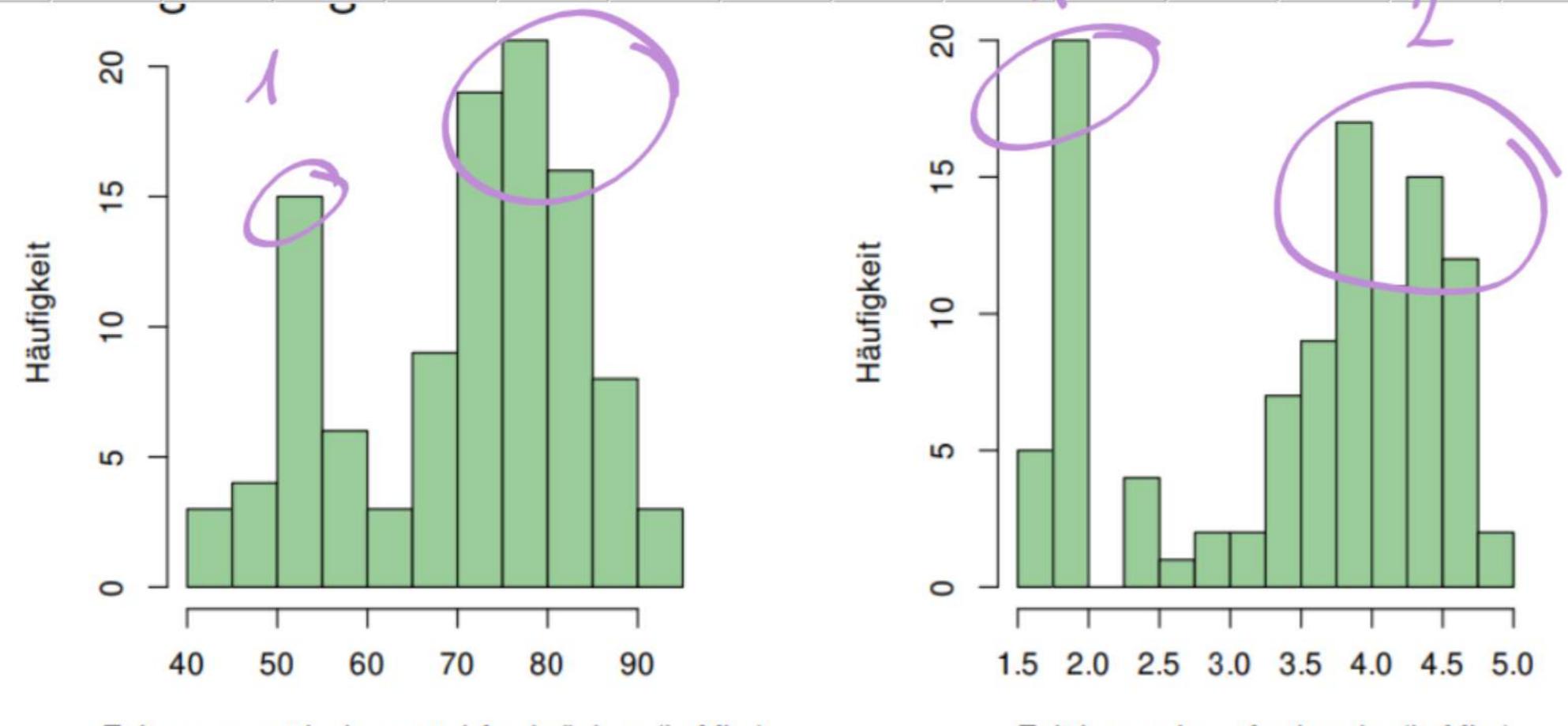
↳ Aufteilung des Wertebereichs in k Klassen (Intervalle)



→ Wahl der Anzahl Balken (Klassen) ist relevant für die Aussagekraft eines Histogramms

↳ Wenn zu viel, dann zu detailliert. Muster sind schwer erkennbar.

↳ Wenn zu wenig, dann zu ungenau.



Bei Beiden ist bimodales Verhalten (d.h. zwei Hügel)

→ in R kann Histogramm über `hist(...)` gemacht

• Code: Der R-Code für das Histogramm oben lautet wie folgt:

```
iq <- rnorm(n = 200, mean = 100, sd = 15)
```

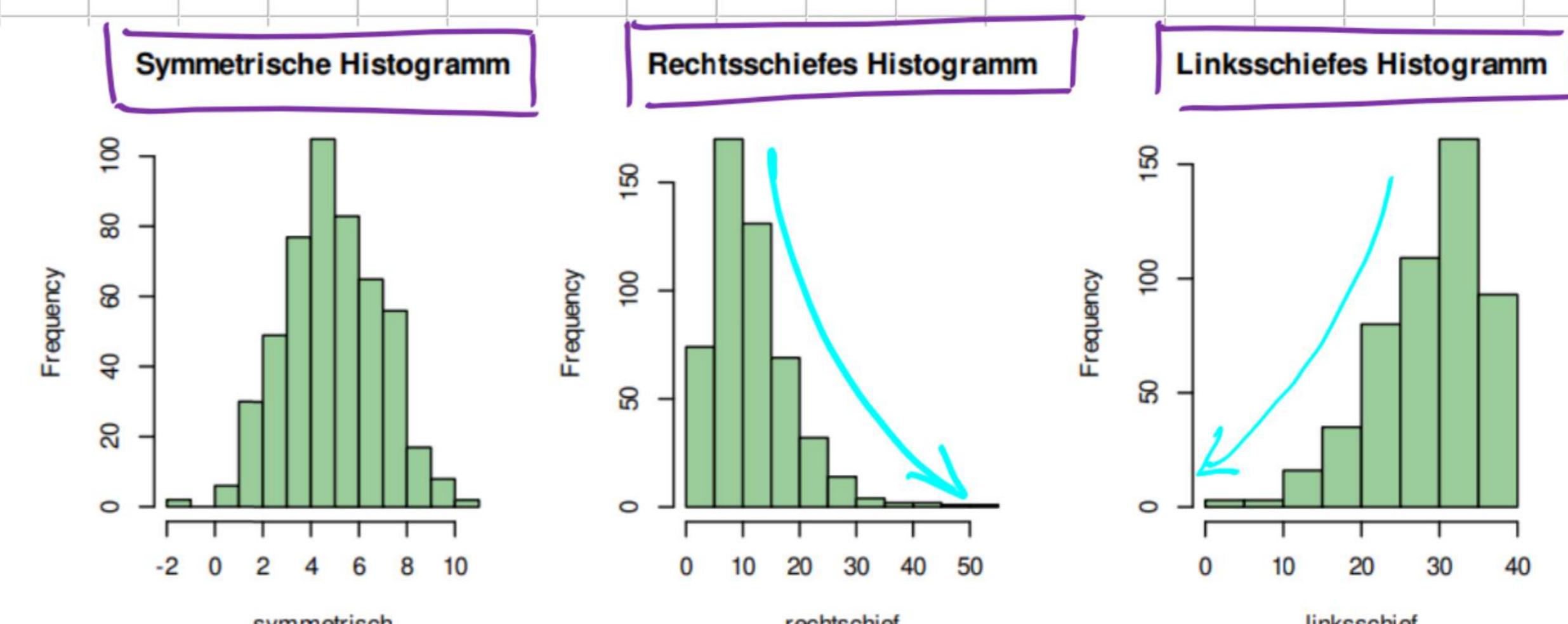
```
hist(iq,
      col = "darkseagreen3",
      xlab = "Punkte im IQ-Test",
      ylab = "Anzahl Personen",
      main = "Verteilung der Punkte in einem IQ-Test")
```

• Befehl `rnorm(n = 200, mean = 100, sd = 15)`: Wählt zufällig 200 normalverteilte Daten (siehe Kapitel Normalverteilung) mit Mittelwert 100 und Standardabweichung 15 aus

• Befehl `hist(iq, ...)`: Histogramm für die Daten `iq`

• Die weiteren Optionen sollten klar sein:

- `xlab` steht für x-Label, die Beschriftung der x-Achse
- `ylab` steht für y-Label, die Beschriftung der y-Achse
- `col` steht für color
- `main` steht für Haupttitel

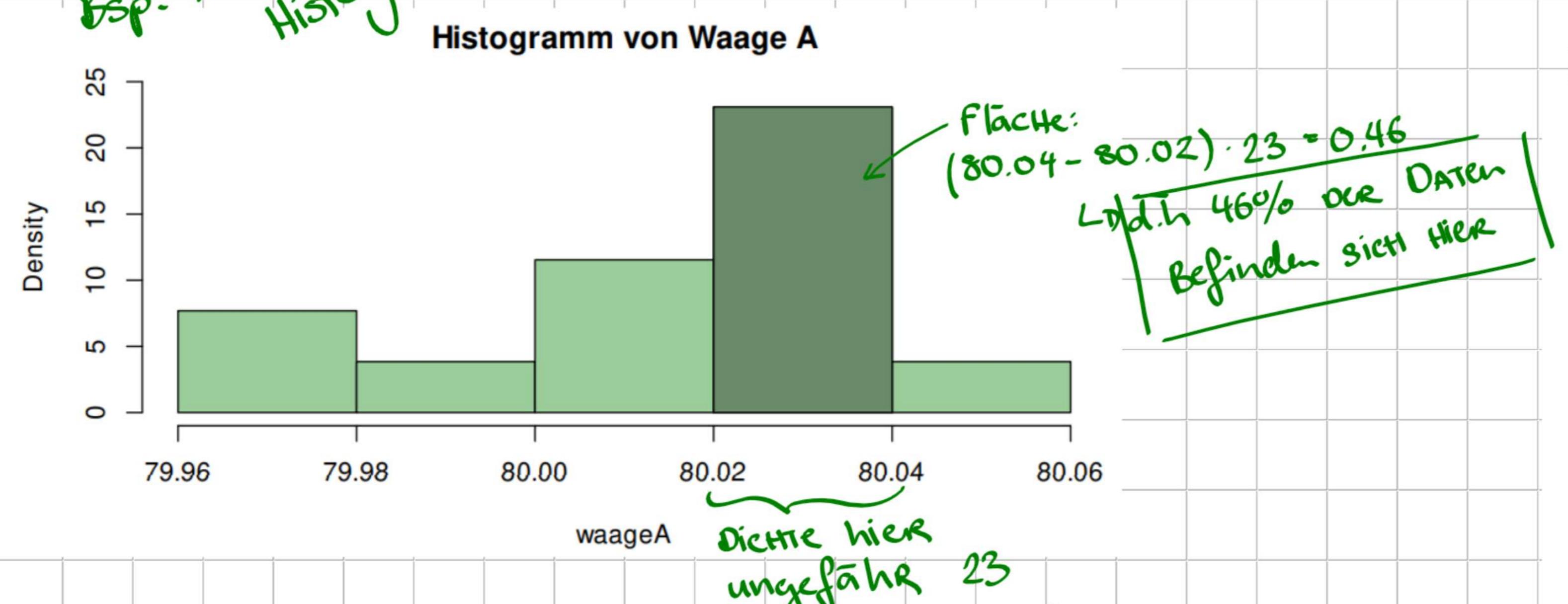


→ bezieht sich immer auf die Richtung wo es weniger Daten hat.

Normiertes Histogramm

- = anstatt Balken, die der Anzahl der Beobachtungen in einer Klasse entsprechen zu wählen: besser → Balkenhöhe so wählen, dass die Balkenfläche dem prozentualen Anteil der jeweiligen Beobachtungen entspricht.
 - ↳ auf der y-Achse wird Dichte angegeben
 - ↳ Gesamtfläche aller Balken muss 1 ergeben.

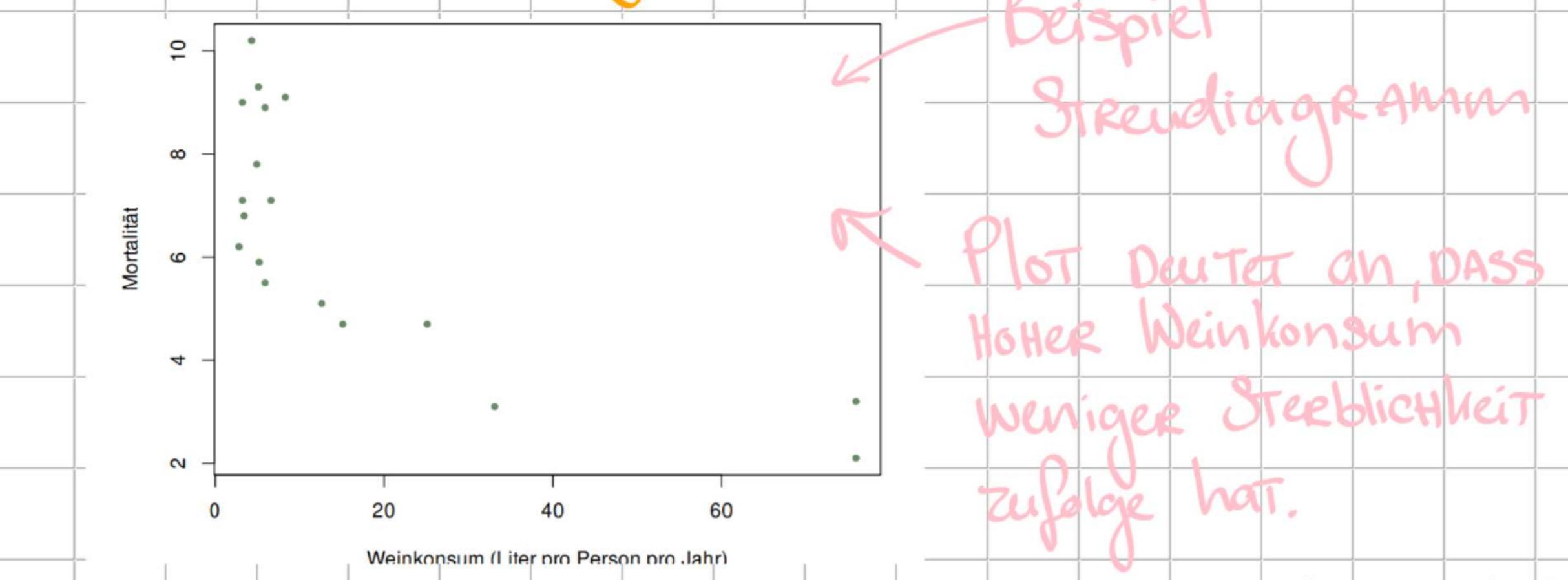
bsp.- normiertes Histogramm



(Zweidimensionale diskursive Statistik)

- Bei Untersuchung von zweidimensionalen Daten:

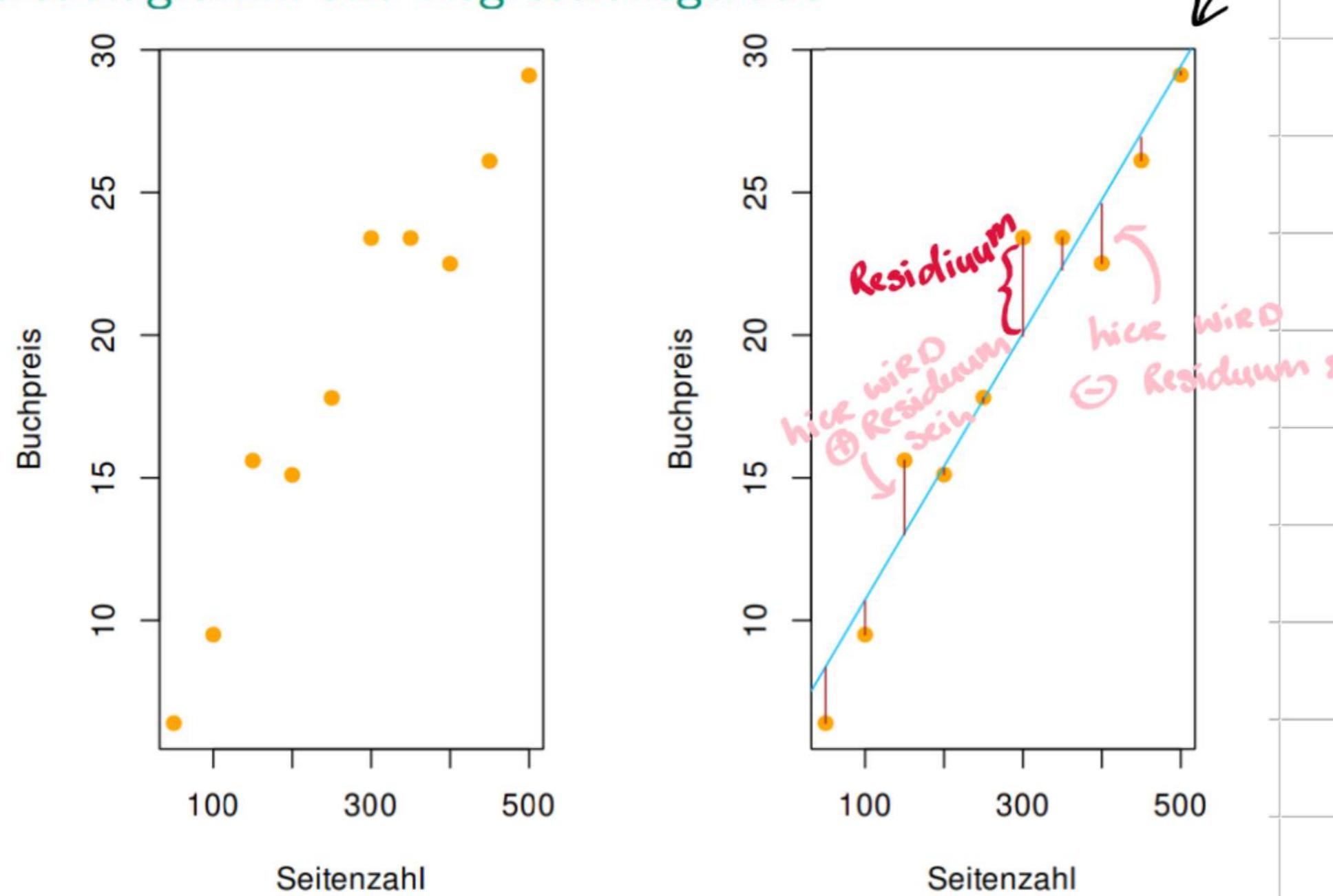
↳ Streudiagramm!



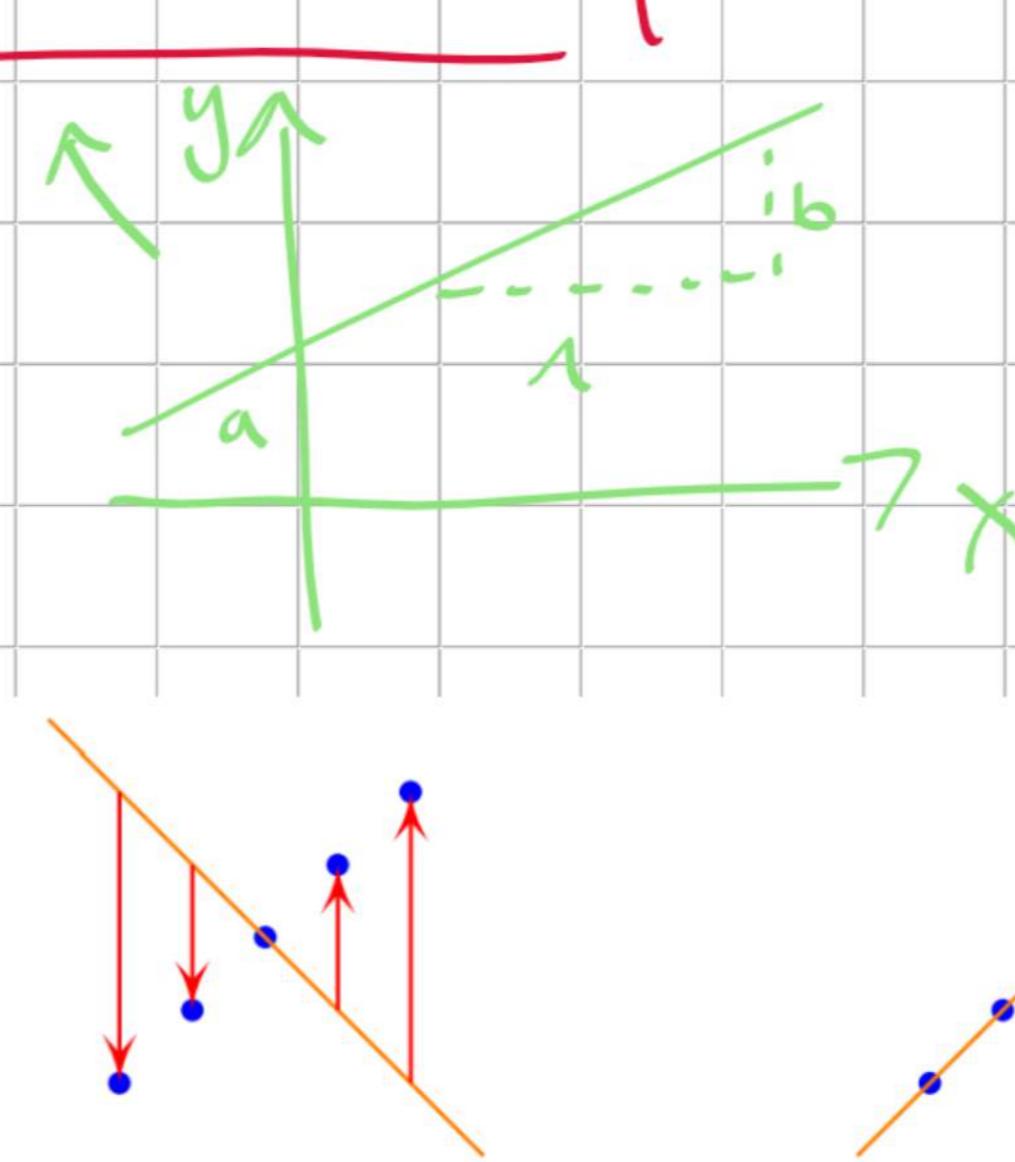
```
wein <- c(2.8, 3.2, 3.2, 3.4, 4.3, 4.9, 5.1, 5.2, 5.9, 5.9, 6.6, 8.3, 12.6, 15.1, 25.1, 33.1, 75.9, 75.9)
mort <- c(6.2, 9.0, 7.1, 6.8, 10.2, 7.8, 9.3, 5.9, 8.9, 5.5, 7.1, 9.1, 5.1, 4.7, 4.7, 3.1, 3.2, 2.1)

plot(wein, mort,
  xlab = "Weinkonsum (Liter pro Jahr)",
  ylab = "Mortalität",
  col = "blue",
  pch = 20)
```

Streudiagramm und Regressionsgerade



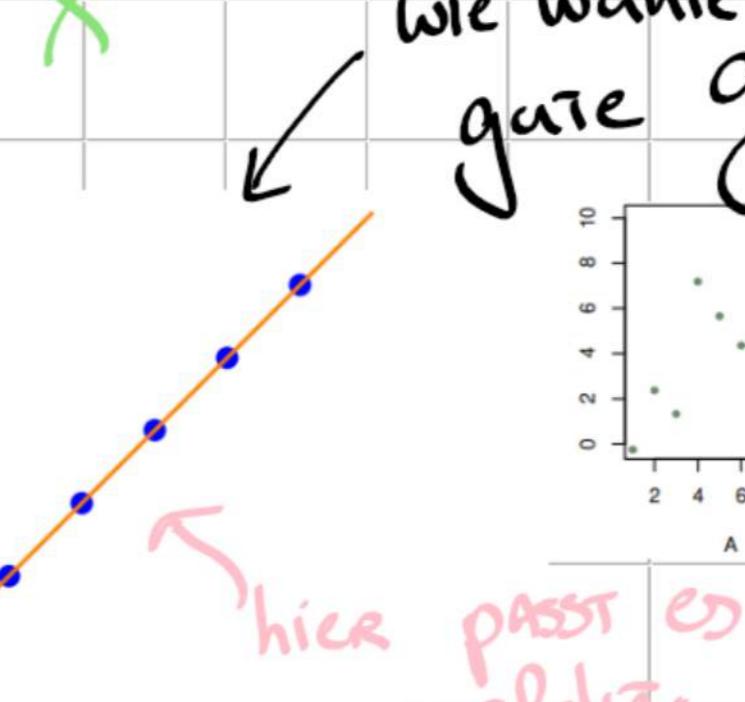
lässt sich mit Geradengleichung berechnen.

$$y = a + bx$$


Summe der Residuen hier 0, aber gerade passt absolut nicht.

wie wähle ich eine gute Gerade?

hier am besten



Spick

Steigung:

Die Steigung gibt an, wie stark die abhängige Variable (y) steigt oder fällt, wenn die unabhängige Variable (x) um eine Einheit zunimmt.

Beispiel: Steigung = 2 → Für jede Erhöhung von x um 1 steigt y um 2.

SW04 - KORRELATION, WAHRSCHEINLICHKEITSMODELLE

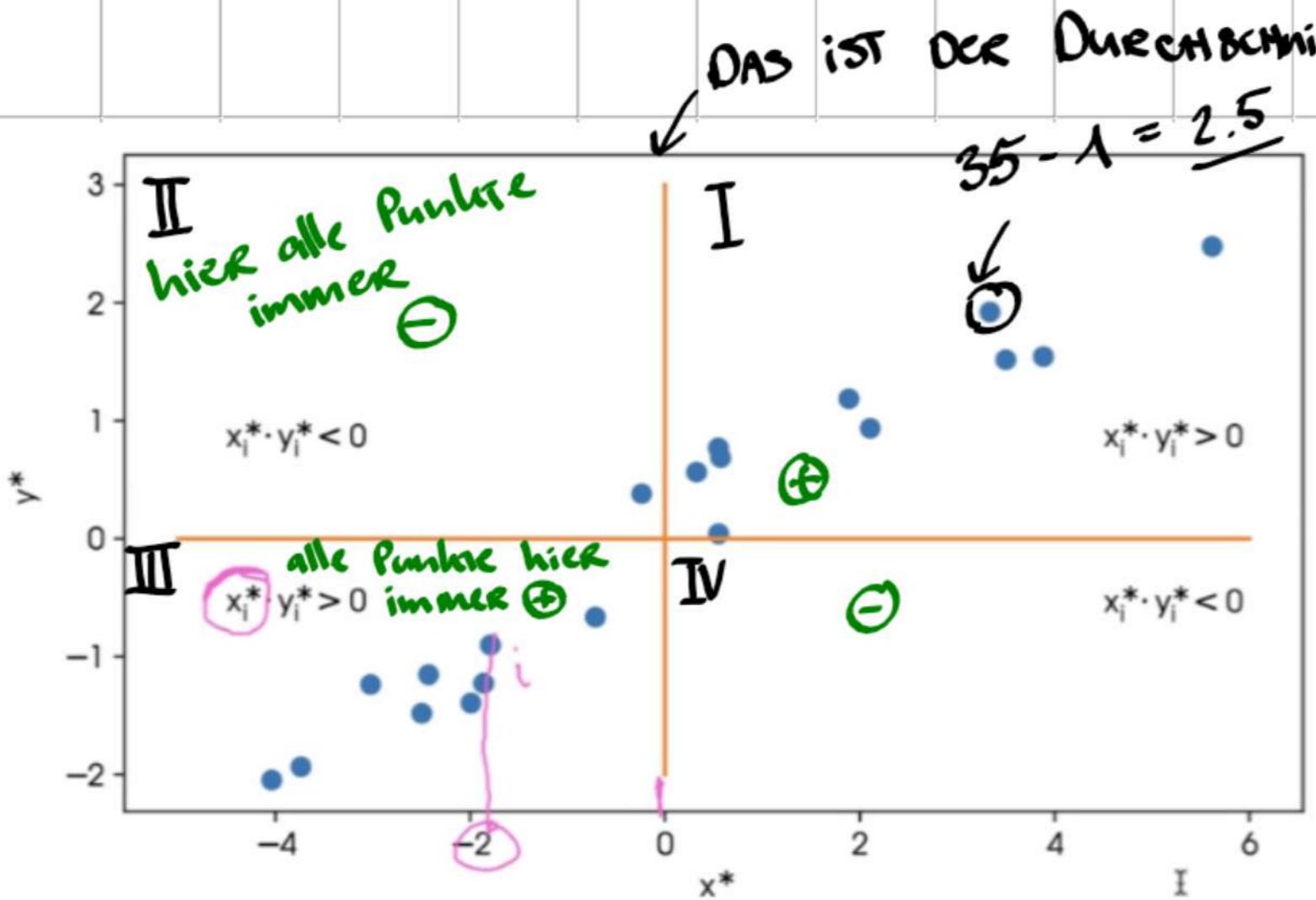
Empirische Korrelation

= Formel. Ist eine dimensionslose Zahl zwischen -1 & +1. Misst Stärke & Richtung der linearen Abhängigkeit zwischen Daten x & y

↳ Bei +1: steigende Gerade

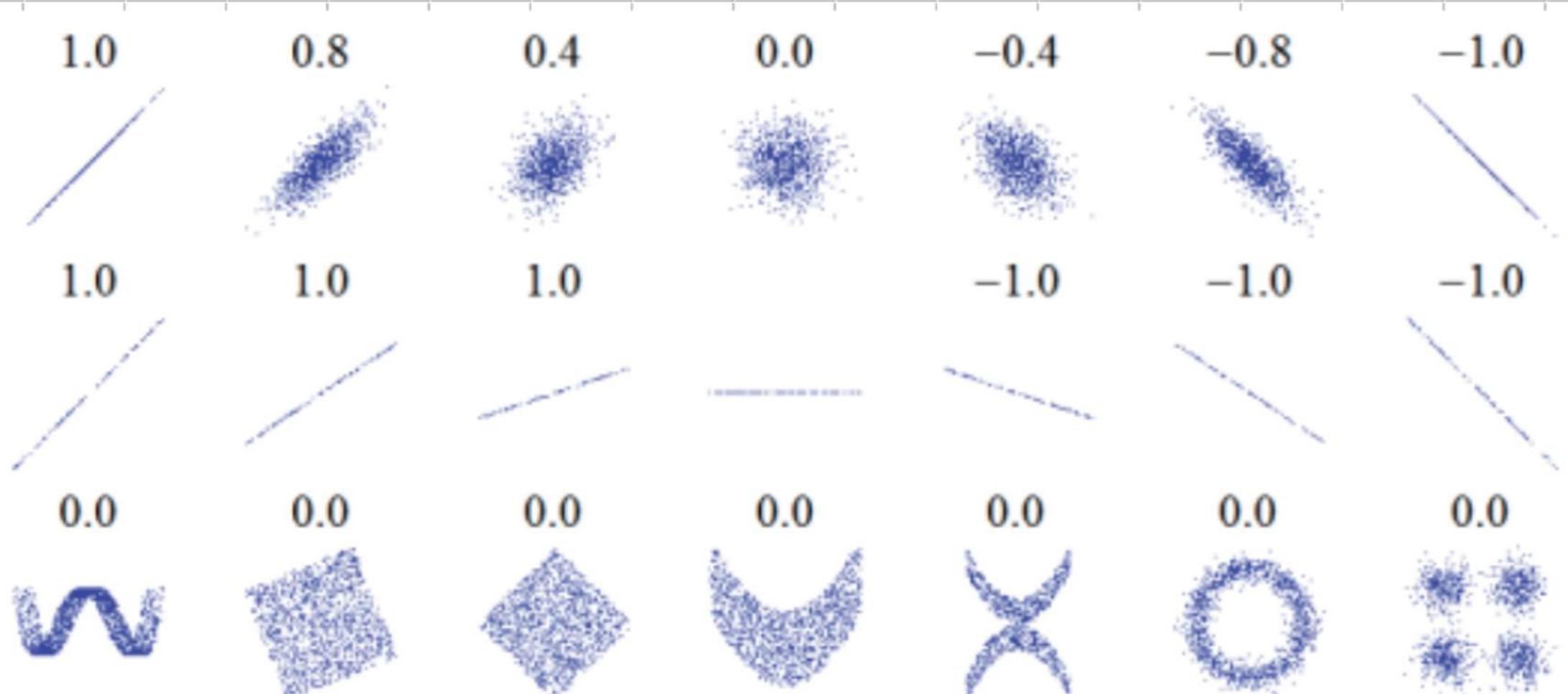
-1: fallende Gerade

0: x & y kein Zusammenhang (unabhängig)



$$x_i^* = (x_i - \bar{x}) = -1.8 - 0 = -1.8 \leq 0$$

$$y_i^* - (\bar{y}_i) = -1 - 0 = -1 < 0$$

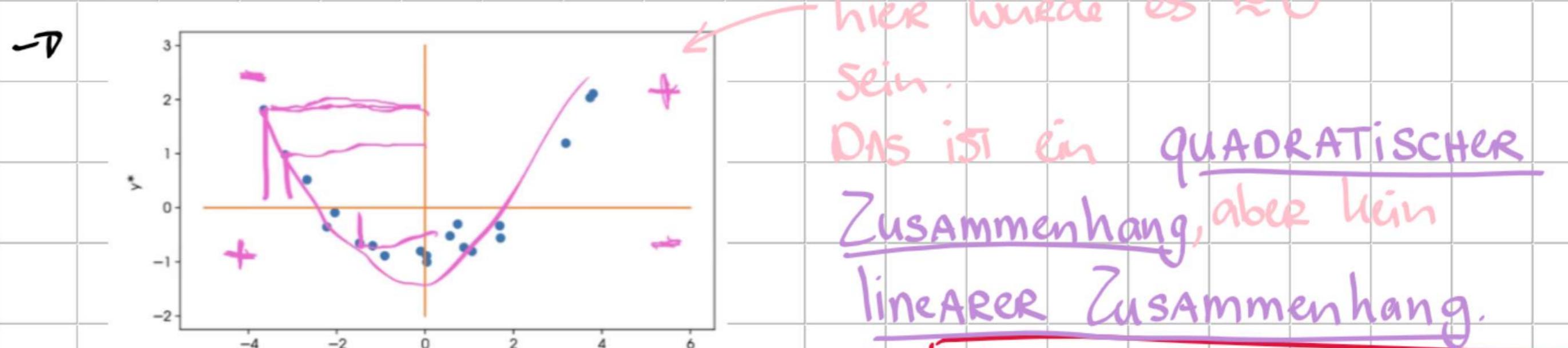


DATEN AUCH IMMER GRAPHISCH BETRACHTEN!

→ Wenn die meisten Punkte im QUADRANT I & III sind, ist die Steigung positiv.

→ Sind die meisten Punkte in II & IV, so ist die Steigung negativ.

→ Sind die Punkte einigermaßen gut verteilt (in einer Wolke), so gibt es keine Steigung.



→ mit R kann man COR(...) dafür beachten

→ es gibt eine Zahl als Output. Je näher diese Zahl bei 1 ist, desto stärker ist der lineare Zusammenhang: je mehr, desto mehr

cor(seiten, preis)

[1] 0.9681122

KORRELATIONSKoeffizient erkennet nur lineare Zusammenhänge

WAHRSCHEINLICHKEIT

- Grundraum Ω (Omega): alle möglichen Elementarereignisse w

- Ereignisse A, B, C: Teilmengen des Grundraums

- Wahrscheinlichkeiten p: gehören zu A, B, C.

- Elementarereignisse: kleinste mögliche Ergebnis eines Zufallsexperiments. z.B. beim Würfeln ist "1" ein Elementarereignis.

mögliche Ausgänge / Resultate

$\Omega = \{\text{mögliche Elementarereignisse}\}$

z.B. Würfel: $\Omega = \{1, 2, 3, 4, 5, 6\}$

ev = 2

τ = es wurde eine 2 geworfen

τ ist kein Elementarereignis DA NICHT im Grundraum Ω

Mengenlehre

Name	Symbol	Bedeutung
Vereinigung	$A \cup B$	A oder B, nicht-exklusives „oder“
Schnittmenge	$A \cap B$	A und B
Komplement	\bar{A}	nicht A
Differenz	$A \setminus B = A \cap \bar{B}$	A ohne B

Münze:

- Ereignis A, wo genau einmal K geworfen wird
- Ereignis A besteht aus Elementarereignissen KZ und ZK

• Ereignis A ist dann die Menge
 $\frac{1}{2} \quad \frac{1}{2}$
 $A = \{KZ, ZK\}$

- Werfen ZZ: Ereignis A tritt nicht ein

- W'keit, dass A eintritt (falls Münze fair):

$$P(A) = \frac{2}{4} = \frac{1}{2}$$

- Statistik: W'keiten oft mit P oder p bezeichnet

- mit der Gegenwahrsch. lässt es sich manchmal schneller berechnen. $P(B) = P(\bar{B}) = 1 - P(B) = 1 - \frac{1}{12} = \frac{11}{12}$
- Modell von Laplace = jedes Elementarereignis HAT die gleiche Wahrscheinlichkeit
- Stochastisch unabhängig: Ereignisse A & B sind stochast. unabhängig, wenn Ausgang des Ereignisses A keinen Einfluss auf den Ausgang des Ereignisses B hat & umgekehrt.

Sind die Ereignisse A und B stochastisch unabhängig, so gilt

$$P(A \cap B) = P(A) \cdot P(B)$$
 mit Schnittmenge

- Unabhängig = keine gegenseitige Beeinflussung, Modell oft „Ziehen mit Zurücklegen“.
- Abhängig = gegenseitige Beeinflussung, Modell oft „Ziehen ohne Zurücklegen“.

SW05 - ZUFALLS VARIABLE & WAHRSCHEINLICHKEITSVERTEILUNG

- Zufallsvariable = $X(w) = x$ werden mit grossen Buchstaben markiert! Kleinbuchstabe sieht konkreten Wert ORT zugeordnete Zahl jedes Elementarereignis funktion

z.B. $X = 11$ entspricht das Ziehen eines Asses.

Zufallsvariablen können wir selbst definieren

↪ z.B. bei Jasskarten.
6, 7, 8, 9 haben Wert 0.
10 hat Wert 10.
Dame hat Wert 3.
König hat Wert 4.

↪ jetzt sind Ziehungen miteinander vergleichbar

z.B. gesucht ist die Wahrscheinlichkeit dass ein König gezogen wird

↪ Resultat einer Zufallsvariable muss eine Zahl sein.

$$P(X=4) = P(\{\omega \mid \omega = \text{ein König}\}) = \frac{4}{36} = \frac{1}{9}$$

4 Farben
6 insgesamt Spielkarten

wie gross ist die Wahrscheinlichkeit, dass die gezogene Karte den Wert 4 hat?

Addieren aller Werte der Wahrscheinlichkeitsverteilung muss 1 ergeben!!

↪ Realisierung ist $X = 4$

↪ weil in 4 Farben gezogen werden kann.

↪ Sonst ist es keine Wahrscheinlichkeitsverteilung.

- Zufallsvariable X: Wert einer gezogenen Jasskarte

Weitere Bsp: • wie gross ist die Wkeit, genau die Augensumme 6 zu würfeln?

- Schon berechnet: Wkeit

↪ $P(X=6)$ gesucht

$$P(X=4) = \frac{1}{9}$$

$$P(X=6) = \frac{5}{36}$$

- Wkeit $P(X=0)$ mit Laplace-Wkeit:

$$P(X=0) = \frac{16}{36} = \frac{4}{9}$$

- Wkeit $P(X=2)$: Ziehen eines Buben:

$$P(X=2) = \frac{4}{36} = \frac{1}{9}$$

• Augensumme 6 oder 8 zu würfeln?

- Wkeitsverteilung von X als Tabelle:

x	0	2	3	4	10	11
$P(X=x)$	$\frac{4}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$

↪ $P(X=6) + P(X=8)$ gesucht

$$P(X=6) + P(X=8) = \frac{5}{36} + \frac{5}{36} = \frac{10}{36} = \frac{5}{18}$$

- mindestens die Augensumme 3 zu würfeln?

• höchstens die Augensumme 3 zu würfeln?

↪ $P(X \leq 3) = P(X=2) + P(X=3)$

↪ $P(X \geq 3) = P(X=3) + \dots + P(X=12)$ gesucht

↪ einfacher: $P(X \geq 3) = 1 - P(X=2)$

$$1 - P(X=2) = 1 - \frac{1}{36} = \frac{35}{36}$$

$$P(X=2) + P(X=3) = \frac{1}{36} + \frac{2}{36} = \frac{3}{36} - \frac{1}{12}$$

- Wie gross ist die W'keit eine Augensumme vom 3 bis 5 zu würfeln?

$$\Leftrightarrow P(3 \leq X \leq 5) = P(X=3) + P(X=4) + P(X=5) \text{ gesucht}$$

$$P(3 \leq X \leq 5) = \frac{2}{36} + \frac{3}{36} + \frac{4}{36} = \frac{9}{36} = \frac{1}{4}$$

- Erwartungswert $E(X)$: mittlere Lage der Verteilung \Rightarrow oft auch μ_X

Standardabweichung $\sigma(X)$: Streuung der Verteilung

Beispiel

- Wurf eines fairen Würfels: Alle 6 möglichen Zahlen gleiche W'keit geworfen zu werden

- Zufallsvariable X sei die geworfene Zahl

- Erwartungswert $E(X)$:

$$E(X) = x_1 \cdot P(X=x_1) + x_2 \cdot P(X=x_2) + \dots + x_6 \cdot P(X=x_6) \\ = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} \\ = \frac{1}{6}(1+2+3+4+5+6) \\ = 3.5$$

Standardabweichung mit R berechnen:

```
x <- 1 : 6
p <- 1 / 6
E_X <- sum(x * p)
var_X <- sum((x - E_X)^2 * p)
sd_X <- sqrt(var_X)
sd_X
## [1] 1.707825
```

\rightarrow bei 100 Mill. mal würfeln ist der Durchschnitt nicht EXAKT 3.5, aber sehr nahe.

- Dieser Erwartungswert 3.5 nicht anderes als der Durchschnitt der Augenzahlen

```
R:
x <- c(0, 2, 3, 4, 10, 11)
p <- 1 / 9 * c(4, 1, 1, 1, 1, 1)
E_X <- sum(x * p)
E_X
## [1] 3.333333
```

Beispiel: Spielkarten

- Verteilung:

x	0	2	3	4	10	11
P(X=x)	4/9	1/9	1/9	1/9	1/9	1/9

- Ziehen aus dem Stapel eine Karte

- Welches ist der durchschnittliche Wert der Karte, die gezogen wird?

- Berechnen Erwartungswert $E(X)$:

$$E(X) = 0 \cdot \frac{4}{9} + 2 \cdot \frac{1}{9} + 3 \cdot \frac{1}{9} + 4 \cdot \frac{1}{9} + 10 \cdot \frac{1}{9} + 11 \cdot \frac{1}{9} = 3.33$$

- Zahlen untereinander in Tabelle werden multipliziert und dann addiert

- Arithmetischer Mittelwert \bar{x} : Aus konkreten Daten berechnet. Aus Messwerten x_1, \dots, x_n wird nach Formel berechnet.

Erwartungswert $E(X)$: Theoretischer Wert, der sich aus dem Modell der W'keitsverteilung ergibt.

SW06 - BEDINGTE WAHRSCHEINLICHKEIT

IRI

z.B. Gruppe aus 20 Personen:

- einige sind Raucher, die anderen Nichtraucher
- einige sind Frauen, andere Männer

F: Frau, M: Mann, R: Raucher, \bar{R} : Nicht Raucher

		M	F	
R	3	1	4	
\bar{R}	9	7	16	
	12	8		20

- 0.15 heißt, dass die W'keit dass die Person Mann ist Raucht.

$$\Leftrightarrow \text{Berechnung: } P(R \cap M) = \frac{|R \cap M|}{|I\Omega|} = \frac{3}{20} = 0.15$$

für Wahrscheinlichkeit: dividieren aller Werte in Tabelle durch 20

- 0.2 heißt, dass eine zufällig ausgewählte Person ein Raucher ist.

$$\Leftrightarrow P(R) = \frac{|R|}{|I\Omega|} = 0.2$$

$$\Leftrightarrow \frac{|R \cap M|}{|R|} = \frac{3}{4} = 0.75 \Rightarrow \text{D.h. } 75\% \text{ der Raucher sind Männer}$$

		M	F	P(R)
R	0.15		0.05	0.2
\bar{R}	0.45	0.35		0.8
	0.6	0.4		1

- Bedingte Wahrscheinlichkeit heißt: NICHT die gesamte Grundmenge betrachten, sondern nur ein Teil davon.

Bezeichnung: $P(M|R)$

DAS IST DIE NEUE GRUNDMENGE: RAUCHER R

$$\text{Es gilt dann: } P(M|R) = \frac{P(R \cap M)}{P(R)}$$

*DAS WAS HINTER DEM STRICH IST
KOMMT UNTER DEN BRUCHSTRICH!*

z.B. $P(R|M) = \frac{P(M \cap R)}{P(M)} = \frac{P(R \cap M)}{P(M)} = \frac{0.15}{0.6} = \underline{\underline{0.25}}$

nur Männer als Grundraum hier

- Die bedingte W'keit ist die W'keit, dass das Ereignis A eintritt, wenn man schon weiß, dass B eingetreten ist
- Bezeichnung:
 $P(A|B)$
- Längstrich wird als „unter der Bedingung“ gelesen
- Bedingte W'keit $P(A|B)$ wird definiert durch
$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$
- Interpretation: $P(A|B)$ ist die W'keit für das Ereignis A, wenn man weiß, dass das Ereignis B schon eingetreten ist

- Weiteres Bsp für bedingte W'kei: a) W'keit, dass ein Kranke auch wirklich positiv getestet wird
b) W'keit, dass ein positiv getester auch wirklich krank ist.

- Bedingte W'keiten hängen auch zusammen:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A) \cdot P(A)}{P(B)}$$

$$\Rightarrow P(A \cap B) = P(A|B) \cdot P(B)$$

$$P(A \cap B) = P(B|A) \cdot P(A)$$

*hier umgedreht, sonst hebt sich
der Term auf & es ist 1.*

- Totale Wahrscheinlichkeit kommt fix an der Prüfung.

- Wenn Menge A in kleinere Mengen A_1, A_2, A_3, \dots unterteilt wird, heißt dies Partitionierung. Wichtig:
Sie haben keine Schnittmenge zusammen!

$$\Leftrightarrow \left| \begin{array}{l} A_1 \cap A_2 = \emptyset; \quad A_1 \cap A_3 = \emptyset; \quad A_2 \cap A_3 = \emptyset \\ \Omega \\ A_1 \cup A_2 \cup A_3 = A \end{array} \right.$$

- Fall $k = 2$
- Graphische Darstellung:
- Mengen A_1 und A_2 bilden eine Partition von Ω
- Es gilt also:
$$A_1 \cup A_2 = \Omega \quad \text{und} \quad A_1 \cap A_2 = \emptyset$$

Bsp. Spam-Filter

$$\begin{aligned} A = \text{Menge an Mails} : \quad A_1 &= \text{"spam"} \quad \rightarrow P(A_1) = 0.7 \\ A_2 &= \text{"niedrige Priorität"} \quad \rightarrow P(A_2) = 0.2 \\ A_3 &= \text{"hohe Priorität"} \quad \rightarrow P(A_3) = 0.1 \end{aligned} \quad \left. \right\} 1$$

Wie gross ist die W'keit, dass es sich um Spam handelt?

- Lösung mit Bayes Theorem und Gesetz der totalen W'keit:

$$\begin{aligned}
 P(A_1|B) &= \frac{P(B|A_1)P(A_1)}{P(B)} \\
 &= \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3)} \\
 &= \frac{0.9 \cdot 0.7}{(0.9 \cdot 0.7) + (0.01 \cdot 0.2) + (0.01 \cdot 0.1)} \\
 &= 0.995
 \end{aligned}$$

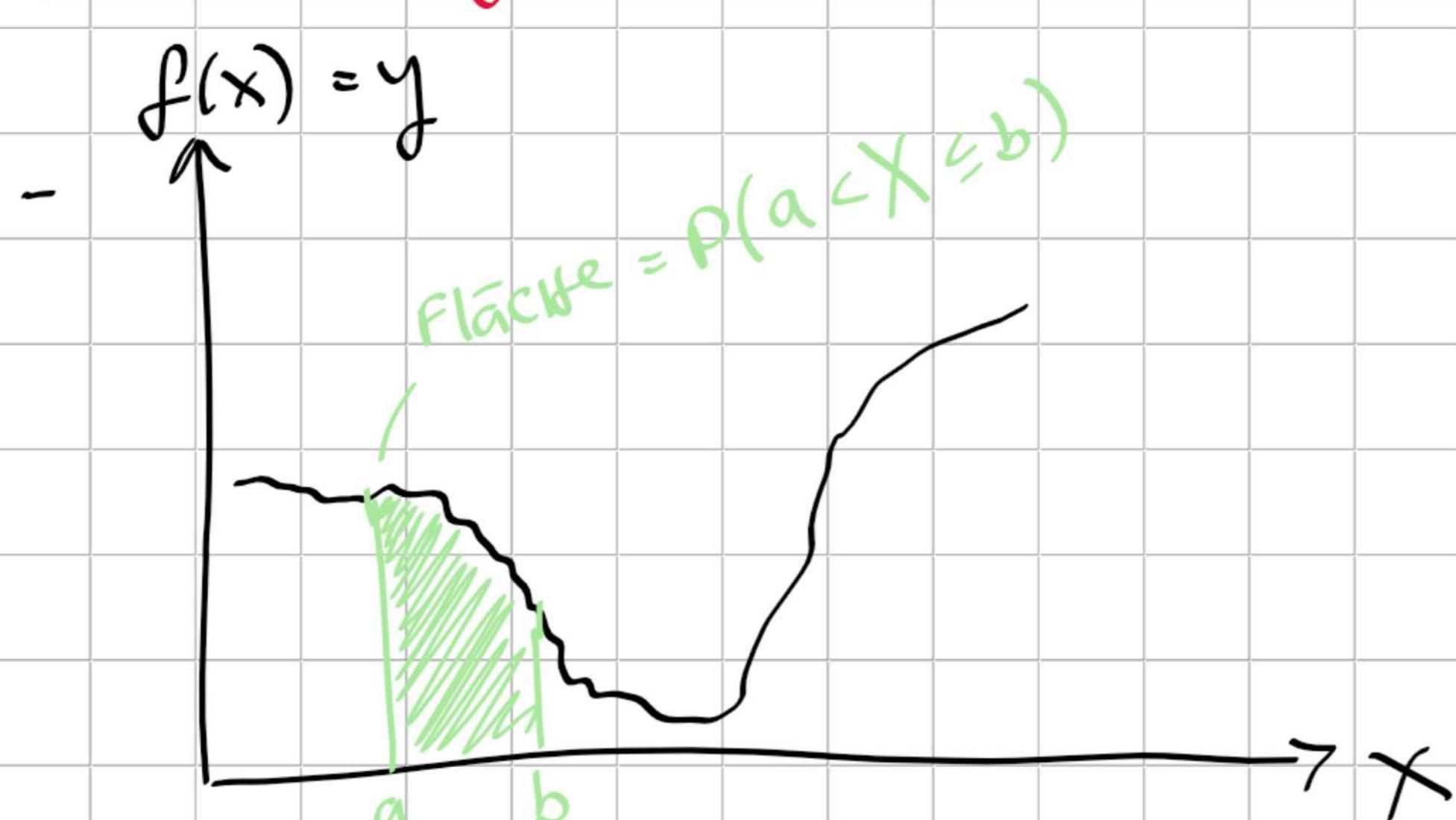
- Viele Spamfilter basieren tatsächlich auf diesem Prinzip

SWo7 - NORMALVERTEILUNG

= Gaussverteilung

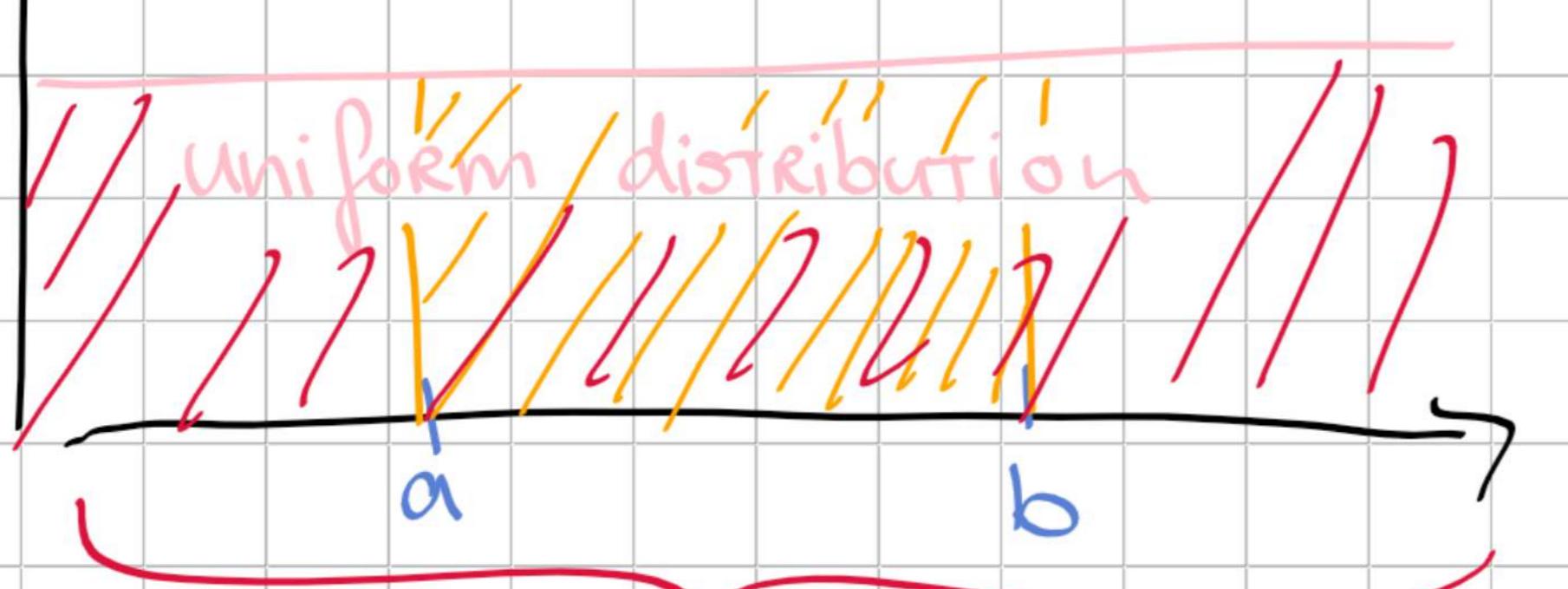
- Unterschied zu Daten aus vorherigen Vorlesungen: sind kontinuierlich! | vorher waren sie diskret, d.h. "lückig".
- $[0, 1]$ → alle Zahlen zwischen 0 & 1 mit 0 & 1.
- $(0, 1]$ → " ohne 0
- $]0, 1[$ → alles vor 0 & 1 aber nicht 0 & 1 & alles zwischen 0 & 1.

Bei diskreten Zufallsvariablen ist die W'keitsverteilung ein Punkt, d.h. $P(X=x)$.
 Bei stetigen " " " " " eine Fläche, d.h. $P(X=x) = 0$.



für eine W'keitsdichte gilt:

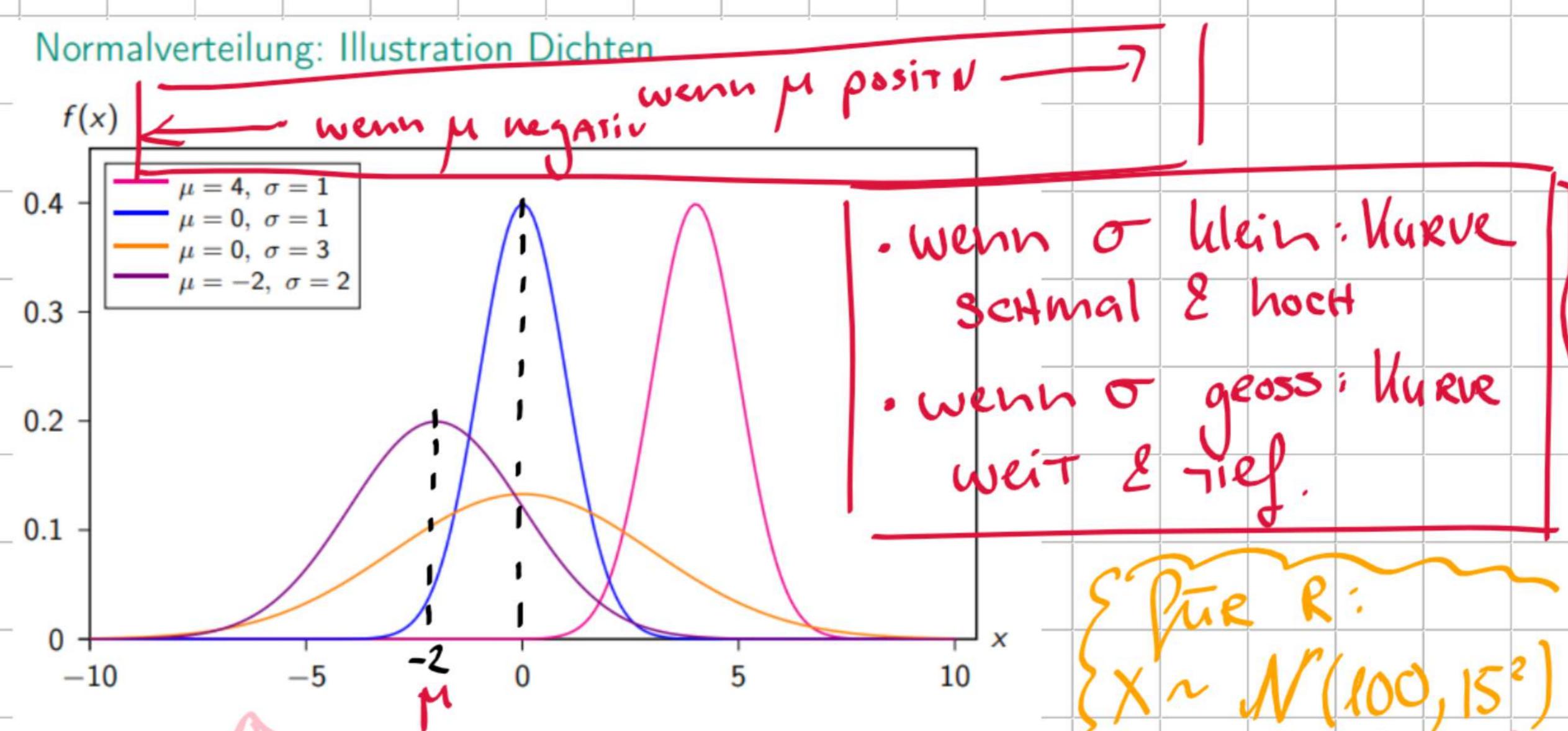
- 1) $f(x) \geq 0$ (d.h. sie fällt nie unter 0)
- 2) $P(a < X \leq b)$ (gesuchte Fläche unter $f(x)$)
- 3) Die gesamte Fläche unter der Kurve ist 1.
 (W'keit, dass irgendein Wert gemessen wird.)



die gesamte Fläche unter dem Graphen ist immer = 1

$$\begin{aligned}
 f(x) &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \\
 &\quad \left. \begin{array}{l} \text{Normalverteilungsparm.} \\ \text{dabei ist } \sigma = \text{STANDARDabweichung} \end{array} \right\}
 \end{aligned}$$

- $E[X] = \mu$ } ERWARTUNGSWERT

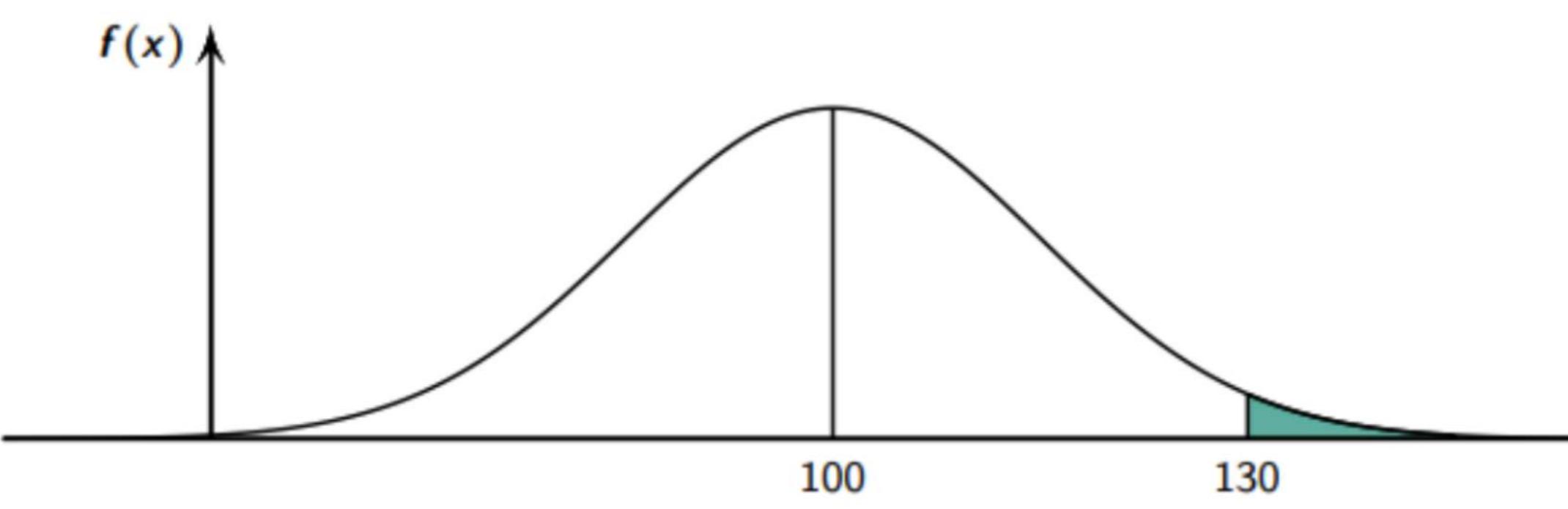


zugeordnungsfrage kommt fix an Prüfung

für R:
 $X \sim N(100, 15^2)$
 $\mu = 100$
 $\sigma^2 = 15^2$
 $N(100, 50)$
 $\sigma = \sqrt{15}$

Beispiel: IQ Tests folgen einer Normalverteilung mit Mittelwert 100 und Standardabweichung 15.
Notation: $X \sim N(100, 15^2)$

- Wie gross die W'keit ist, dass jemand einen IQ von mehr als 130 hat, also als hochbegabt gilt?
- $P(X > 130)$, wobei $X \sim N(100, 15^2)$
- Skizze:

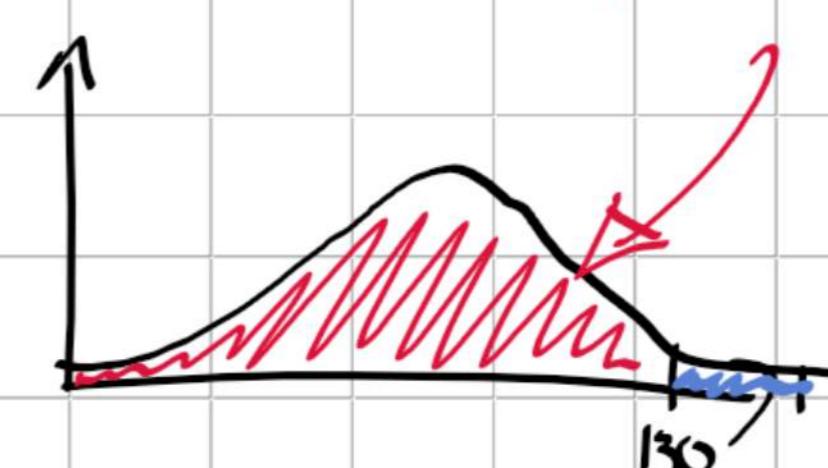


```
1 - pnorm(q = 130, mean = 100, sd = 15)
```

```
## [1] 0.02275013
```

⇒ Rund 2% der Bevölkerung sind hochbegabt.

A ABER: das würde nur die Fläche bis zu den 130 berechnen, d.h.



Wenn an der Prüfung die Normalverteilung so angegeben wird: $N(100, 30)$, dann in R so schreiben: `pnorm(q = 130, mean = 100, sd = 30)`

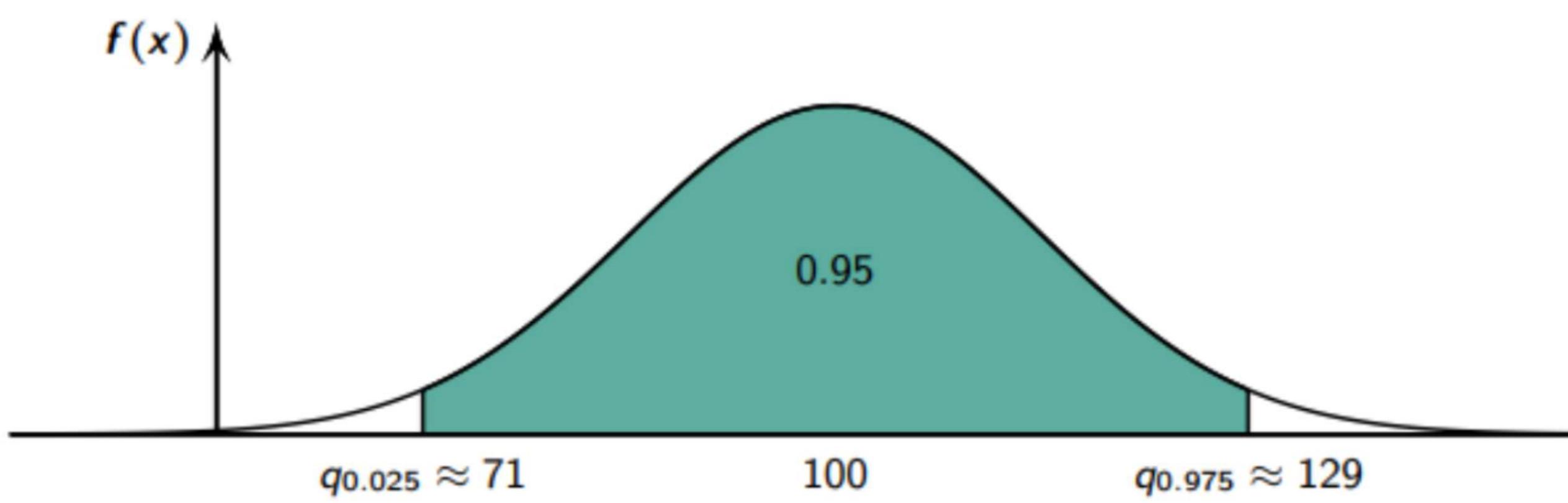
V damit DAS NICHT passiert:

$1 - \underbrace{\text{pnorm}(q = 130, mean = 100, sd = 15)}$

Befehl für probability Nicht zu verwechseln mit qnorm.
"quartil"

Beispiel 2: Welches Intervall enthält 95% der IQ's um den Mittelwert $\mu = 100$?

- W'keit als Fläche:



- Grüne Fläche: 95 % der Gesamtfläche

Hier sind W'keiten gegeben ⇒ Suche der Werte. Bestimmung der Quartile $q_{0.025}$ und $q_{0.975}$

in R: `qnorm(p = 0.025, mean = 100, sd = 15)`

`qnorm(p = 0.975, mean = 100, sd = 15)`

- Oder kürzer:

```
qnorm(p = c(0.025, 0.975), mean = 100, sd = 15)
## [1] 70.60054 129.39946
```

- 95 % der Menschen haben einen IQ zwischen ungefähr 70 und 130

in R: `pnorm(q = 115, mean = 100, sd = 15) - pnorm(85, 100, 15)`

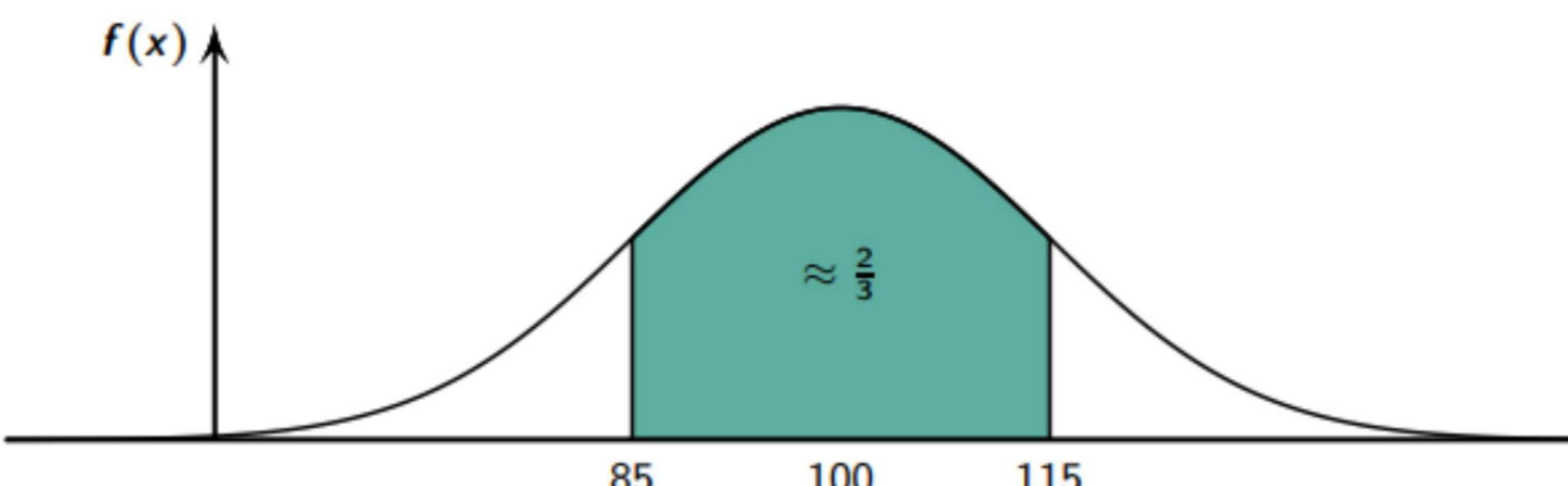
Beispiel 3:

- Wieviel Prozent der Bevölkerung liegen innerhalb einer Standardabweichung vom Mittelwert liegen?

- Gesucht W'keit:

$$P(85 \leq X \leq 115)$$

- W'keit als Fläche:



Funktion	Was wird berechnet?	Wann verwenden?
<code>pnorm(x)</code>	Wahrscheinlichkeit $P(X \leq x)$	Wenn du wissen möchtest, wie viele Werte kleiner/gleich x sind.
$1 - \text{pnorm}(x)$	Wahrscheinlichkeit $P(X > x)$	Wenn du wissen möchtest, wie viele Werte größer x sind.
<code>qnorm(p)</code>	Wert x , sodass $P(X \leq x) = p$	Wenn du den Wert für eine bestimmte Wahrscheinlichkeit suchst.

SW08 - GESETZ DER GROSSEN ZAHLEN

Funktion von mehreren Zufallsvariablen:

- hier 1 Messung mehrmals machen

$$X_1, X_2, X_3, \dots, X_n$$

- $X_i \rightarrow$ i-te Wiederholung des Zufallsexperiments

Summe: $S_n = X_1 + X_2 + \dots + X_n = \sum_{i=1}^n X_i$

Durchschnitt: $\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \dots + X_n) = \frac{1}{n} \cdot \sum_{i=1}^n X_i = \frac{1}{n} \cdot S_n$

STANDARDfehler: STANDARDabweichung des Durchschnitts nicht proportional zu $\frac{1}{n}$.

" \sqrt{n} -Gesetz"

ZENTRALER Grenzwertsatz:

↑ durchschnittliche Wahrscheinlichkeit, nähерungsweise Wahrscheinlichkeit

→ bekannt Kennzahlen $S_n \approx \bar{X}_n$

→ unbekannt Verteilung $S_n \approx \bar{X}_n$

→ Wie verteilt?

$$S_n \approx N(n \cdot \mu, n \cdot \sigma^2 \cdot x) \quad \left\{ \begin{array}{l} \text{Summe} \\ pnorm(q = \dots, mean = n \cdot \mu, sd = \sqrt{n} \cdot \sigma) \end{array} \right.$$

$$\bar{X}_n \approx N(\mu, \frac{\sigma^2 \cdot x}{n}) \quad \left\{ \begin{array}{l} \text{durchschnittlicher Mittelwert} \\ pnorm(q = \dots, mean = \mu, sd = \sqrt{\frac{\sigma^2 \cdot x}{n}}) \end{array} \right.$$



$$\text{Var}(\bar{X}_n) = \frac{\text{Var}(x)}{n}, \quad \sigma(\bar{X}_n) = \frac{\sigma \cdot x}{\sqrt{n}}$$

⚠

Summe → Befehl: `pnorm(a, mu * n, sigma * sqrt(n))`

Strassenverkehrsamt hat genug Streusalz gelagert, um mit einem Schneefall von insgesamt 80 cm pro Jahr fertigzuwerden. Täglich fallen im Mittel 1.5 cm mit einer Standardabweichung von 0.3 cm. Wie gross ist Wahrscheinlichkeit, dass das gelagerte Salz für die nächsten 50 Tage ausreicht?

```
pnorm(q = 80, mean = 50 * 1.5, sd = sqrt(50) * 0.3)
## [1] 0.9907889
```

SW09 - HYPOTHESENTEST

- ZGWS = zentraler Grenzwertsatz → Verteilung der Mittelwerte/Summen nähert sich mit wachsendem n einer Normalverteilung an.

- Nullhypothese = wir schreiben: $H_0: \mu = \mu_0 = 80$ bezeichnet.

- Alternativhypothese = $H_A: \mu \neq \mu_0 = 80$ oder " $<$ " / " $>$ "

Durchschnitt → Befehl: `pnorm(a, mu, sigma / sqrt(n))`

Die Lebensdauer eines bestimmten elektrischen Teils ist durchschnittlich 100 Stunden mit Standardabweichung von 20 Stunden. Wir testen 16 solcher Teile. Wie gross ist die Wahrscheinlichkeit, dass das Stichprobenmittel unter 104 Stunden liegt?

```
pnorm(q = 104, mean = 100, sd = 20/sqrt(16))
## [1] 0.7881446
```

Beispiel: wir kennen μ ($P(\bar{X}_6 \leq 79.98)$) und die σ ($= 0.02$) & haben 6x gemessen.
d.h. wir schreiben:

$$\bar{X}_6 \sim N(80, \frac{0.02^2}{6})$$

↳ Damit testen wir, ob die Annahme $\mu = 80$ gerechtfertigt ist.

in R: $pnorm(q = 79.98, mean = 80, sd = 0.02/sqrt(6))$

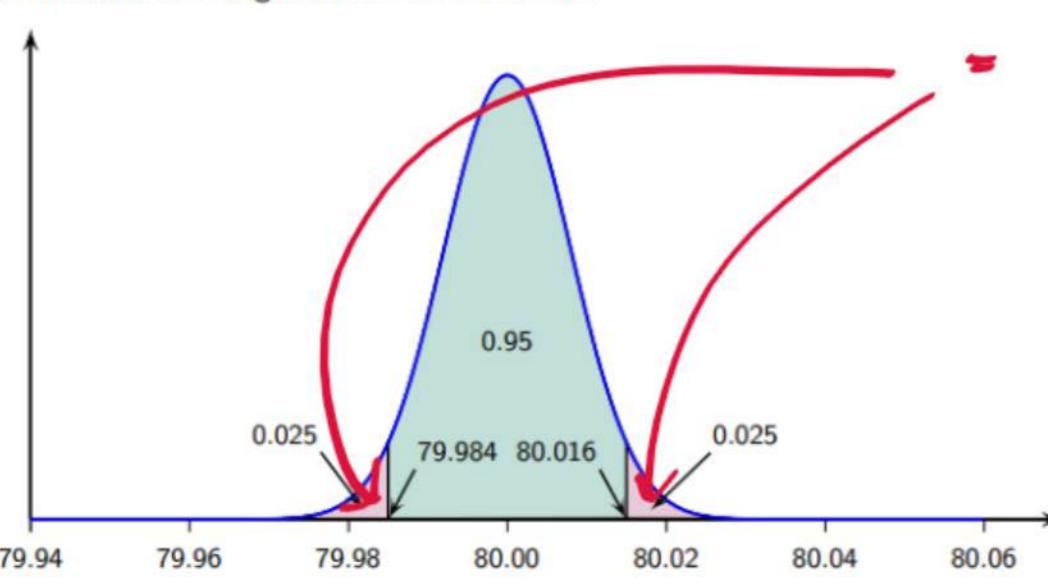
↳ hier kommt 0.007152939 raus. D.h. die W'keit ist 0.7%

In der Statistik hat man sich mal geeinigt, dass der Grenzwert bei 2.5% ist.

↳ d.h. $P(\bar{X}_6 \leq 79.98) < 0.025$. Damit passt die Annahme $\mu = 80$ nicht so.

Graphische Darstellung

- Normalverteilungskurve in drei Teile auf:



- Signifikanzniveau α (= p-Wert)

= gibt an, wie hoch das Risiko ist, das man bereit ist einzugehen, eine falsche Entscheidung zu treffen.

liegt der gemessene Mittelwert im roten Bereich, so zweifelt man an der Nullhypothese.

Wir verwirfen also die Nullhypothese $\mu = 80$.

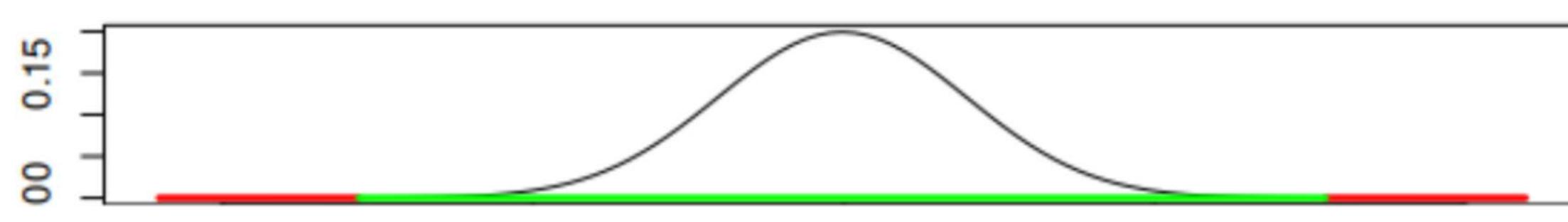
↳ Verwerfungsbereich: $K = (-\infty, 79.984] \cup [80.016, \infty)$

es gäbe noch zweiseitige Tests \rightarrow DAS haben wir vorher gemacht.

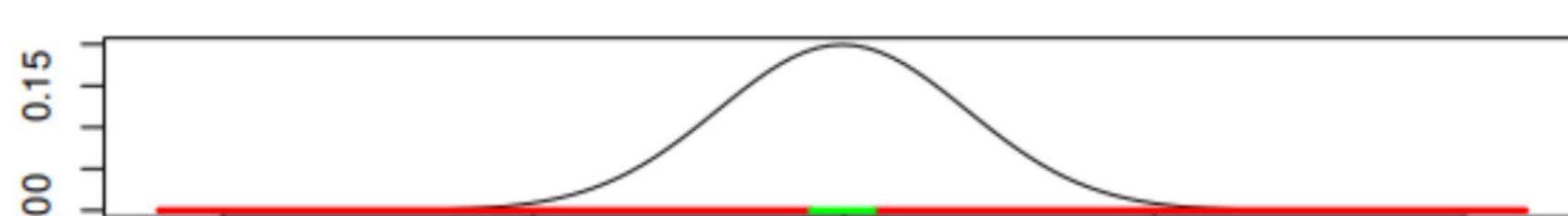
- einseitige Tests = Verwerfungsbereich nur einseitig getestet.

Wahl von Signifikanzniveau α

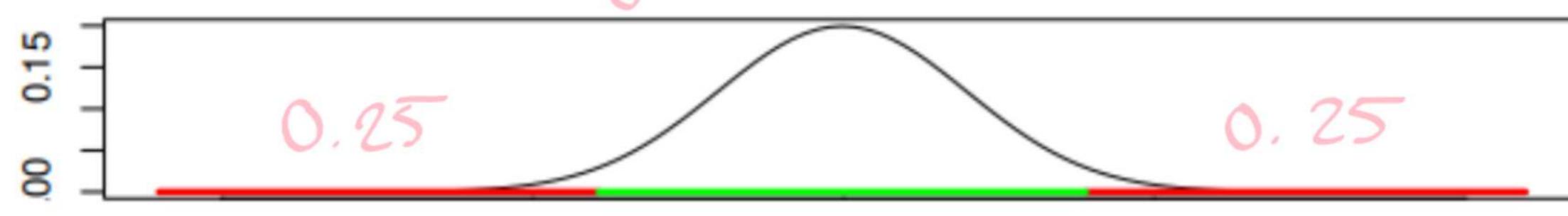
- Graphik: $\alpha = 0.0001$ (nahe bei 0)



- Graphik: $\alpha = 0.8$ (gross)



- Graphik: $\alpha = 0.05 \rightarrow$ gutes Kompromiss



Je kleiner das α , desto kleiner der rote Bereich (a.k.a. Verwerfungsbereich.)

\hookrightarrow bei Vermutungen machen wir nur einseitige Tests.

z.B. Vermutung: dieser Wert ist zu gross.

\hookrightarrow die Anzahl Messungen hat einen Einfluss auf den Verwerfungsbereich.

Je mehr Messungen, desto mehr wird verworfen, weil der Bereich engere wird.

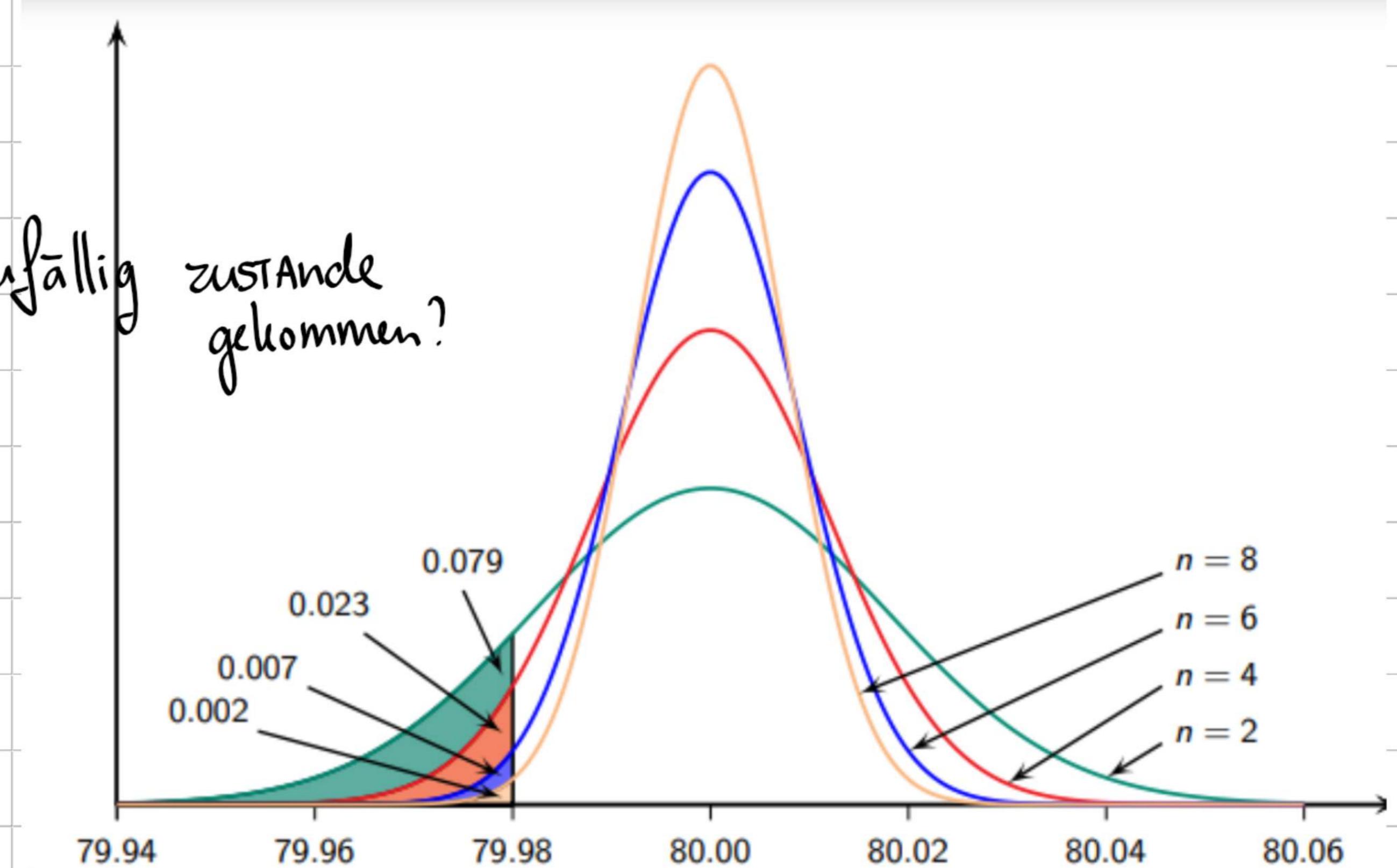
- p-Wert = ist ein Wert zwischen 0 und 1, der angibt, wie gut Nullhypothese & Daten zusammen passen. \rightarrow Ist das Ergebnis zufällig?

0: passt gar nicht.
1: passt sehr gut.

- z-Test = wichtig: Standardabw. bekannt.

- t-Test = setzt keine Standardabw. voraus. Setzt ähnlich zum z-Test, ist aber etwas unsicherer. Man sieht das an der Verteilung.

t-Verteilung
Die Verteilung der Teststatistik beim t-Test unter der Nullhypothese
 $H_0: \mu = \mu_0$
ist gegeben durch
 $T = \bar{X}_n \sim t_{n-1} \left(\mu, \frac{\hat{\sigma}_x}{\sqrt{n}} \right)$
wobei t_{n-1} eine t-Verteilung mit $n - 1$ Freiheitsgraden ist



Normalverteilung wird durch eine t-Verteilung ersetzt.

FÜR t-Test immer t_{n-1} verwenden!

in R mit `t.test(...)` arbeiten.

\hookrightarrow man setzt σ_x aus den Daten. Sodass im Output (z.B. `t.test(x, mu = 5)`) bei p-value mehr als 0.25, wird die Nullhypothese nicht verworfen. mean ist der Mittelwert von x im Output. Alle anderen Angaben im Output sind uninteressant.

z-Test

Beim z-Test ist die Standardabweichung im Voraus bekannt.

- Null- und Alternativhypothese aufstellen
- Verwerfungsbereich bestimmen
- p-Wert mit Messreihe berechnen
- Testentscheid fällen

Das Bundesamt für Statistik behauptet, dass die durchschnittliche Körpergrösse der erwachsenen Frauen in der Schweiz bei 180 cm mit einer Standardabweichung von 10 cm liegt. \rightarrow Einseitiger Test

Der p-Wert ist unter dem Signifikanzniveau von 0.05. Die Nullhypothese wird somit verworfen und die Alternativhypothese angenommen.

• Nullhypothese $H_0: \mu_0 = 180$
• Alternativhypothese $H_A: \mu < \mu_0 = 180$

```
qnorm(p = 0.05, mean = 180, sd = 10/sqrt(8))
## [1] 174.1846
```

Wählen zufällig 8 erwachsene Frauen aus, deren durchschnittliche Körpergrösse 171.54 cm beträgt.

```
pnorm(q = 171.54, mean = 180, sd = 10/sqrt(8))
## [1] 0.008359052
```

Wann? Große Stichprobe ($n > 30$), Varianzen bekannt, normalverteilt.

Beispiel: Vergleich des Durchschnittsgewichts einer Region mit einem bekannten Mittelwert von 70 kg.

t-Test

Beim t-Test ist die Standardabweichung im Voraus unbekannt. Es folgt eine zusätzliche Unsicherheit. Die t-Verteilung ist ähnlich der Normalverteilung, aber flacher, aufgrund der grösseren Unsicherheit. Das Bundesamt für Statistik behauptet, dass die durchschnittliche Körpergrösse der erwachsenen Frauen in der Schweiz bei 180 cm liegt. \rightarrow Einseitiger Test

Wählen zufällig 10 Frauen aus und messen deren Körpergrösse.

Der p-Wert ist unter dem Signifikanzniveau von 0.05. Die Nullhypothese wird somit verworfen und die Alternativhypothese angenommen.

```
groesse <- c(165.7, 156.7, 171.7, 180.3, 163.2, 166.7, 149.9,
170.4, 163.4, 152.5)
t.test(groesse, mu = 180, alternative = "less")
## One Sample t-test
## data: groesse
## t = -5.4836, df = 9, p-value = 0.0001942
## alternative hypothesis: true mean is less than 180
## 95 percent confidence interval:
##   -Inf 169.382
## sample estimates:
## mean of x
## 164.05
```

Wann? Kleine Stichprobe ($n < 30$), Varianzen unbekannt, normalverteilt.

Arten:

Gepaarter T-Test:

Beispiel: Gewichtsmessung vor und nach einer Diät.

Ungepaarter T-Test:

Beispiel: Vergleich der Durchschnittstemperatur in zwei Städten

SW10 - VERTRAUENSINTERVALL, Wilcoxon - Test

= t-Test, nur ohne Normalverteilte Daten.

μ steht immer für irgendeinen Durchschnitt.

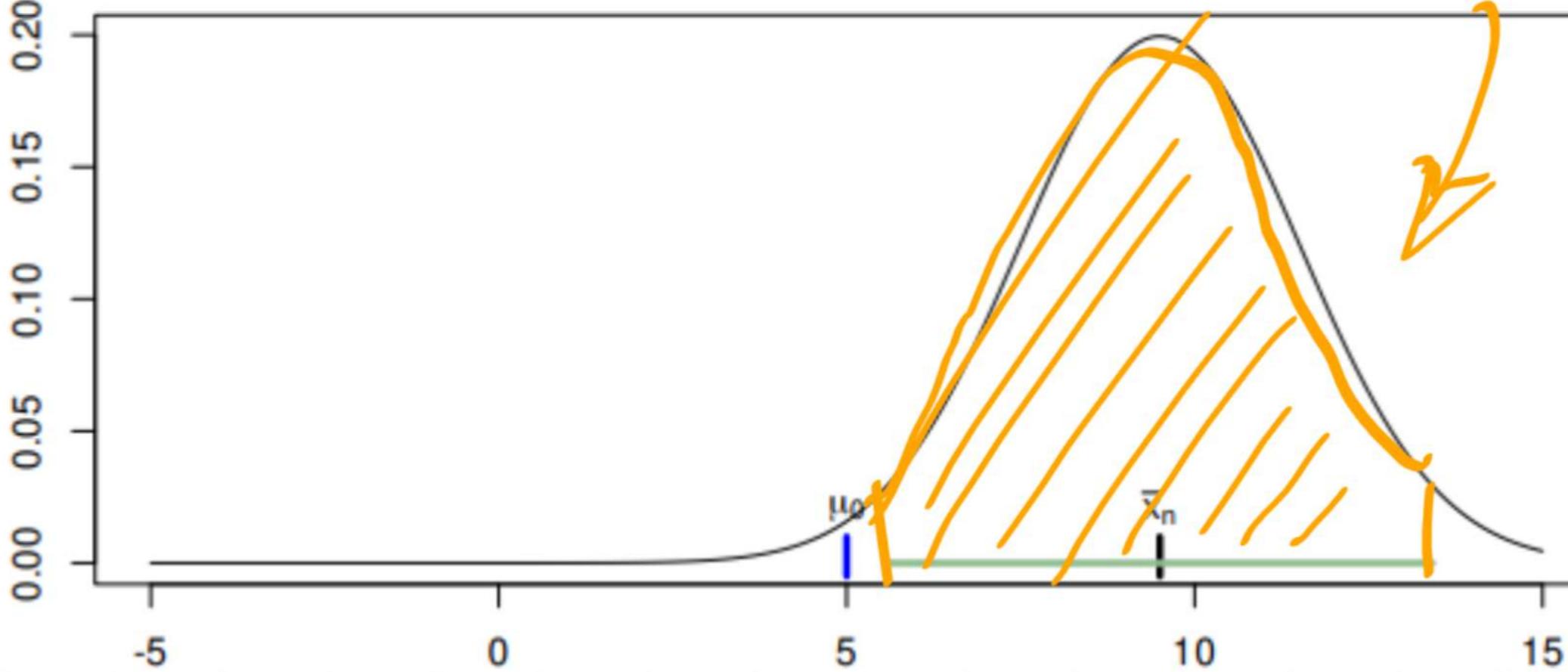
$\hat{\mu}$ = geschätzter Durchschnitt

\bar{x}_n = arithmetischer Durchschnitt, meist nur einer einzelnen Variablen
 n = Anzahl Messungen

Vertrauensintervall = Intervall, das angibt, wo grob gesagt, der wahre Mittelwert mit einer bestimmten Vorgegebene Wkheit liegt.

= Bereich, der nicht zum Verwerfungsbereich gehört.

- Dieses Intervall heißt Vertrauensintervall



in R: confidence interval. Damit lassen sich Testentscheidungen durchführen.

Zu 95% liegt der wahre Mittelwert im Vertrauensintervall.

auch: Enthält alle μ_0 's für die Nullhypothese nicht verworfen wird.
 Wahres μ_0 liegt zu 95% im Vertrauensintervall.

- Weitere Möglichkeit für Testentscheid:

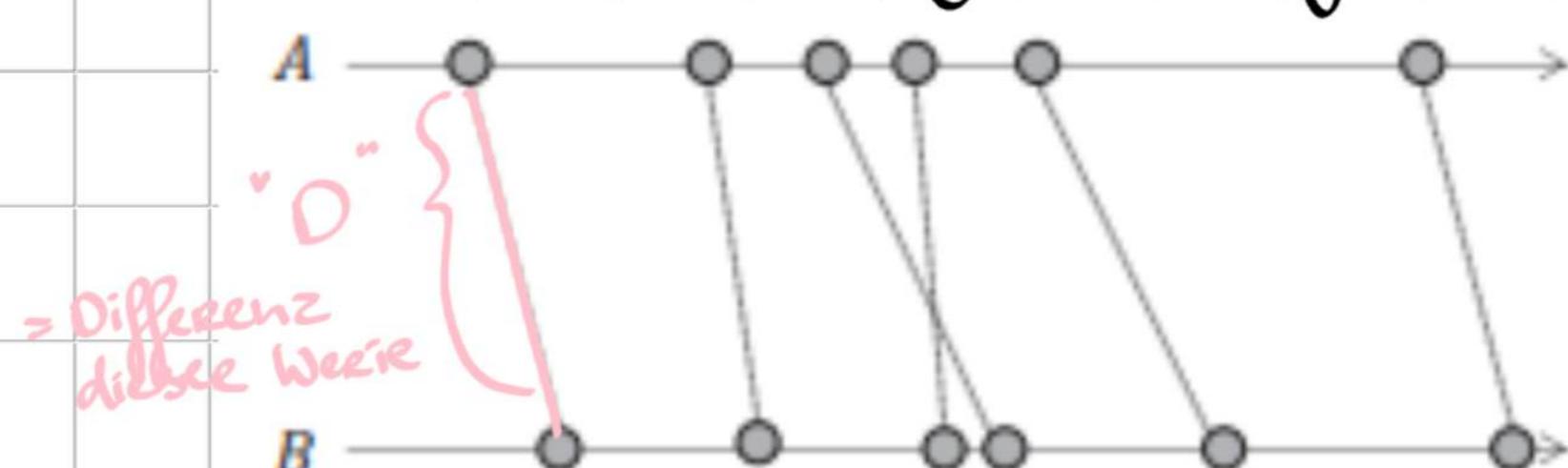
- Liegt μ_0 der Nullhypothese im Vertrauensintervall, so wird die Nullhypothese nicht verworfen
- Liegt μ_0 der Nullhypothese nicht im Vertrauensintervall, so wird die Nullhypothese verworfen

- Wie weiß man, dass Daten normalverteilt sind?

Am besten über Histogramm. Bei nicht normalverteilten Daten (z.B. Datensatz hat viele Nullen)

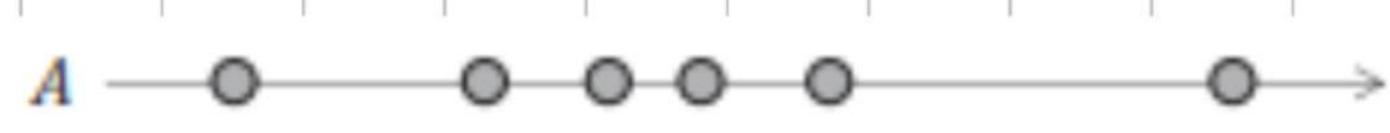
Gepaarte Stichproben = abhängige Daten voneinander. Stichprobengröße ist in beiden Gruppen gleich.

Jede Beobachtung einer Gruppe kann eindeutig der anderen Gruppe zugeordnet werden.



z.B. eine Auge wird betrachtet, das andere nicht. Resultate voneinander abhängig.

Ungepaarte Stichproben = unabhängige Daten voneinander. Keine Zuordnung von Beobachtungen möglich. Stichprobengrößen können auch versch. sein.



Differenz der Durchschnitte

Ungepaart:

$$\text{Intuition Teststatistik: } T = \frac{\bar{X} - \bar{Y}}{\sigma_{\bar{D}}}$$

Durchschnitt der Differenzen

Gepaart:

$$\text{Differenz } D_i = X_i - Y_i$$

$$\text{Teststatistik } T = \frac{\bar{D}}{\sigma_{\bar{D}}}$$

→ Durchschnitt der Differenzen

Wilcoxon-Test

Beim Wilcoxon-Test müssen die Daten nicht normalverteilt sein. Er ist eine Alternative zum t-Test mit weniger Voraussetzungen.

Lediglich die Verteilung unter der Nullhypothese muss symmetrisch sein.

```
x <- c(79.98, 80.04, 80.4, 80.05, 80.03, 80.02)
wilcox.test(x, mu = 80, alternative = "two.sided")
## Wilcoxon signed rank test with continuity correction
## data: x
## V = 69, p-value =
## alternative hypothesis
```

Wann?: Daten nicht-normalverteilt.

Arten:

Wilcoxon-Vorzeichen-Rang-Test (gepaart):

Beispiel: Vorher-Nachher-Vergleich der Schlafqualität bei nicht-normalverteilten Daten.

Wilcoxon-Mann-Whitney-Test (ungepaart):

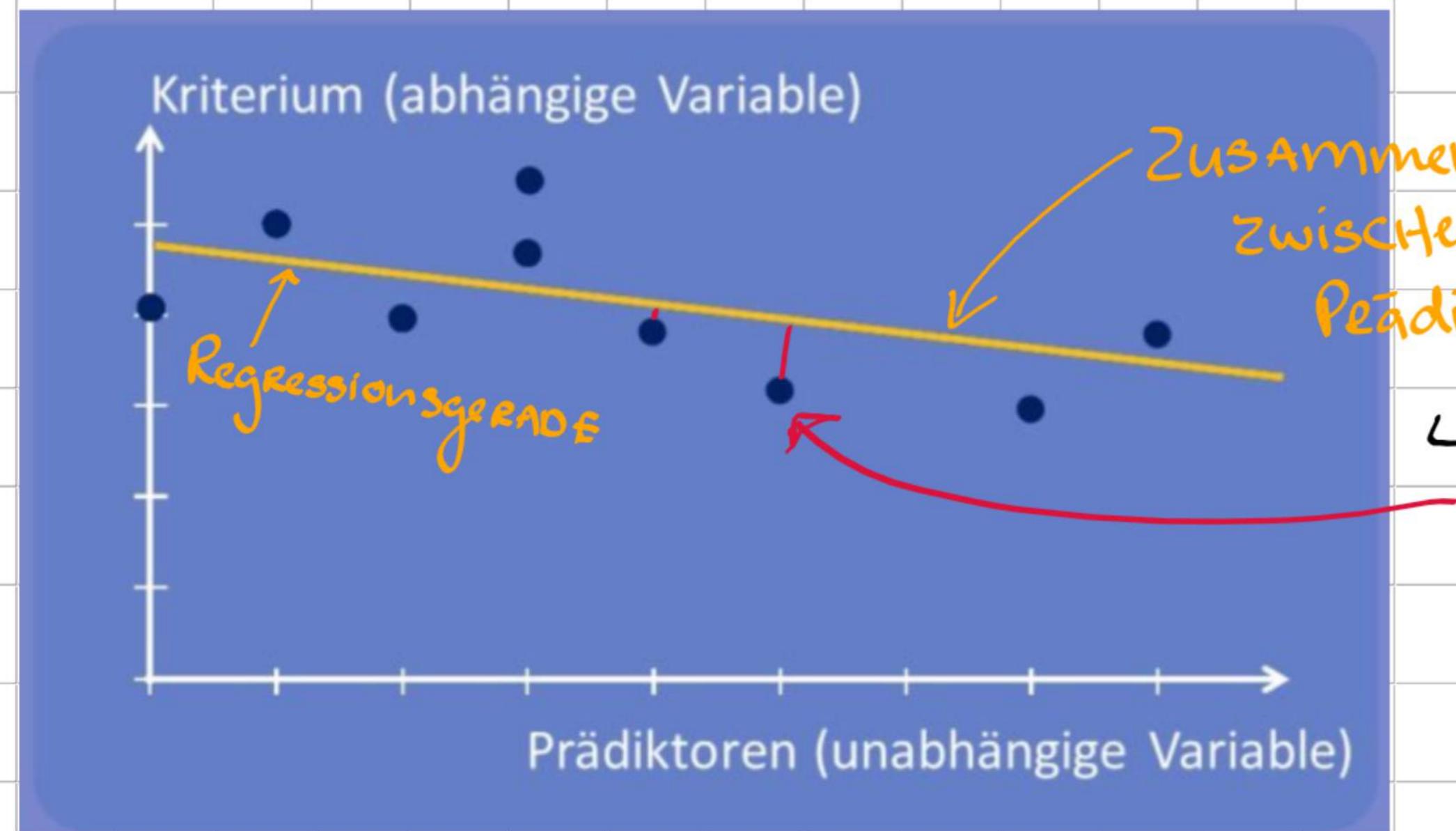
Beispiel: Vergleich von Gehältern in zwei Branchen bei nicht-normalverteilten Daten.

SW11 - LINEARE REGRESSION

Lineare Regression = ist einer der STARTpunkte im Machine Learning. Ziel: Zusammenhang herstellen, Modell entwickeln zur Vorhersage.

$$\rightarrow | Y \approx f(X_1, X_2, X_3) |$$

kein Gleichheitszeichen, da keine Graphen



Kriterium = Variable, die vorhergesagt werden soll.
(= abhängige Variable)

Prädiktoren = Variablen, die zur Vorhersage genutzt werden.
(= unabhängige Variable)

→ lineare Linie weil "lineare Reg." & nicht "kubische Reg."

- Zusammenhang approximativ darstellen zwischen Y und X_n

- Form lineare Regression = $| Y = \beta_0 + \beta_1 \cdot X + \varepsilon |$

→ Zielgröße, Output Variable
→ Prädiktoren, erklärende Variablen

→ Y = abhängende Variable

→ X = unabhängige Variable

→ β_0 = y -Achse Abschnitt (Konstante)

→ β_1 = Steigungsparameter

→ ε = Fehlerterm

BEI MEHREREN UNABHÄNGIGEN

$$| Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \dots + \beta_n \cdot X_n + \varepsilon |$$

→ X_n = versch. unabhängige Variablen

METHODE kleinste Quadrate = Technik zur Schätzung der Parameter in lin. Reg.

Ziel: Werte der unbekannten Parameter dass Summe quadrierter Residuen minimal wird.

RSS = Summe der Quadrate der Residuen

$$RSS = r_1^2 + r_2^2 + r_3^2 + \dots + r_n^2$$

$$\text{bzw. } RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_2)^2 + \dots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_n)^2$$

SW12 - Multiple lineare REGRESSION

→ einfache lineare Regression ist gut um Output aufgrund einer einzelnen Variablen vorherzusagen. Meistens aber von mehreren Variablen abhängig...

→ Prüfung auf Papier, closed-book, multiple choice

→ multiple lineare Regression nimmt mehrere lineare Graphen & versucht herauszufinden, wie sie zusammenhängen. So entsteht ein 3D-Modell.

Gleichung: $| Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon |$

Input Zielgröße Y

Einfache lineare Regression

Einfaches Verfahren, um einen quantitativen Output Y auf der Basis einer einzigen Inputvariable X vorherzusagen.

$$Y \approx \beta_0 + \beta_1 X$$

- ▶ β_0 ist der y-Achsenabschnitt
- ▶ β_1 die Steigung der Geraden

Für zusätzliche CHF 1'000 Werbeausgaben werden 47.5 zusätzliche Einheiten des Produktes verkauft.

Die Nullhypothese wird mit p-Wert $2 \cdot 10^{-16}$ verworfen.

Somit gibt es einen klaren Zusammenhang.

Die R2-Statistik erklärt 61.19% der Varianz durch das Modell.

Das Modell ist somit zu 2/3 akurat.

$$\text{Verkauf} \approx \beta_0 + \beta_1 \cdot \text{TV}$$

$$Y \approx 7.03 + 0.0475X$$

```
confint(lm(Verkauf ~ TV), level = 0.95)
##      2.5 %    97.5 %
## (Intercept) 6.12971927 7.93546783
## TV          0.04223072 0.05284256
```

Verkauf liegt ohne Werbung zwischen 6'130 und 7'935 Einheiten.

Für zusätzliche CHF 1'000 für TV-Werbung, werden durchschnittlich zwischen 42 und 53 Einheiten mehr verkauft.

```
summary(lm(Verkauf ~ TV))
##
## Call:
## lm(formula = Verkauf ~ TV)
##
## Residuals:
##   Min     1Q Median     3Q    Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.032594   0.457843 15.36 <2e-16 ***
## TV          0.047537   0.002691 17.67 <2e-16 ***
## ---
```

β_1 β_0

```
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF, p-value: < 2.2e-16
```

Multiple Lineare Regression

Das multiple lineare Modell verallgemeinert das einfache lineare Modell. Jeder erklärenden Variable wird ein eigener Steigungskoeffizient in einer Gleichung zugeordnet. Graphische Darstellung nur bis zwei erklärende Variablen möglich (Ebene).

β_i : Durchschnittliche Änderung der Zielgröße bei Änderung von X_i um eine Einheit, wenn alle anderen erklärenden Variablen gleichbleiben.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$$

Steigung für Zeitung beschreibt die Änderung der Zielgröße

Verkauf, wenn man CHF 1'000 mehr für Zeitungswerbung ausgibt, wobei die anderen erklärenden Variablen TV und Radio gleichbleiben.

R2 erhöht sich, je mehr erklärende Variablen berücksichtigt werden.

$$\text{Verkauf} \approx \beta_0 + \beta_1 \cdot \text{TV} + \beta_2 \cdot \text{Radio} + \beta_3 \cdot \text{Zeitung}$$

$$\text{Verkauf} \approx 2.94 + 0.046 \cdot \text{TV} + 0.189 \cdot \text{Radio} - 0.001 \cdot \text{Zeitung}$$

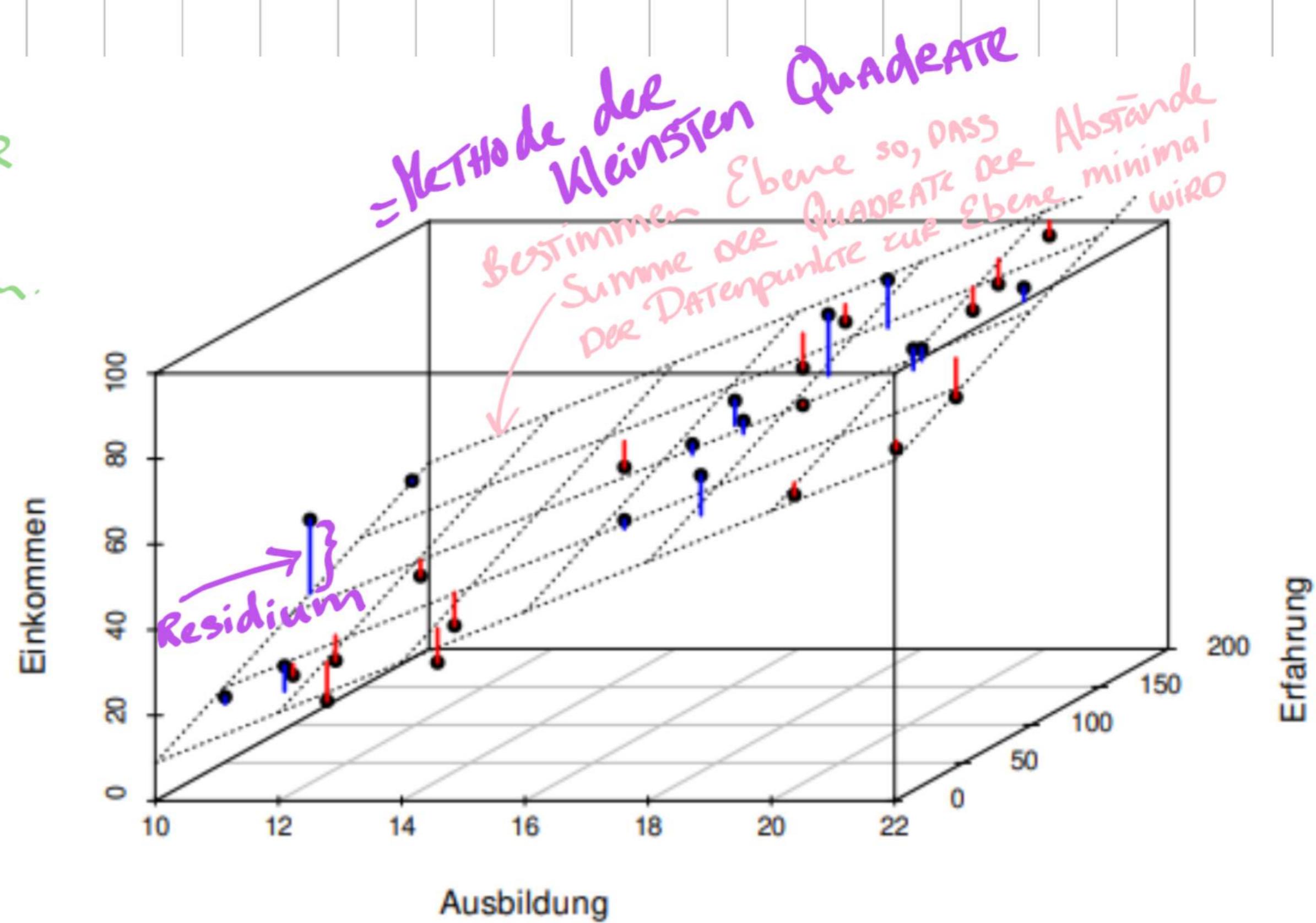
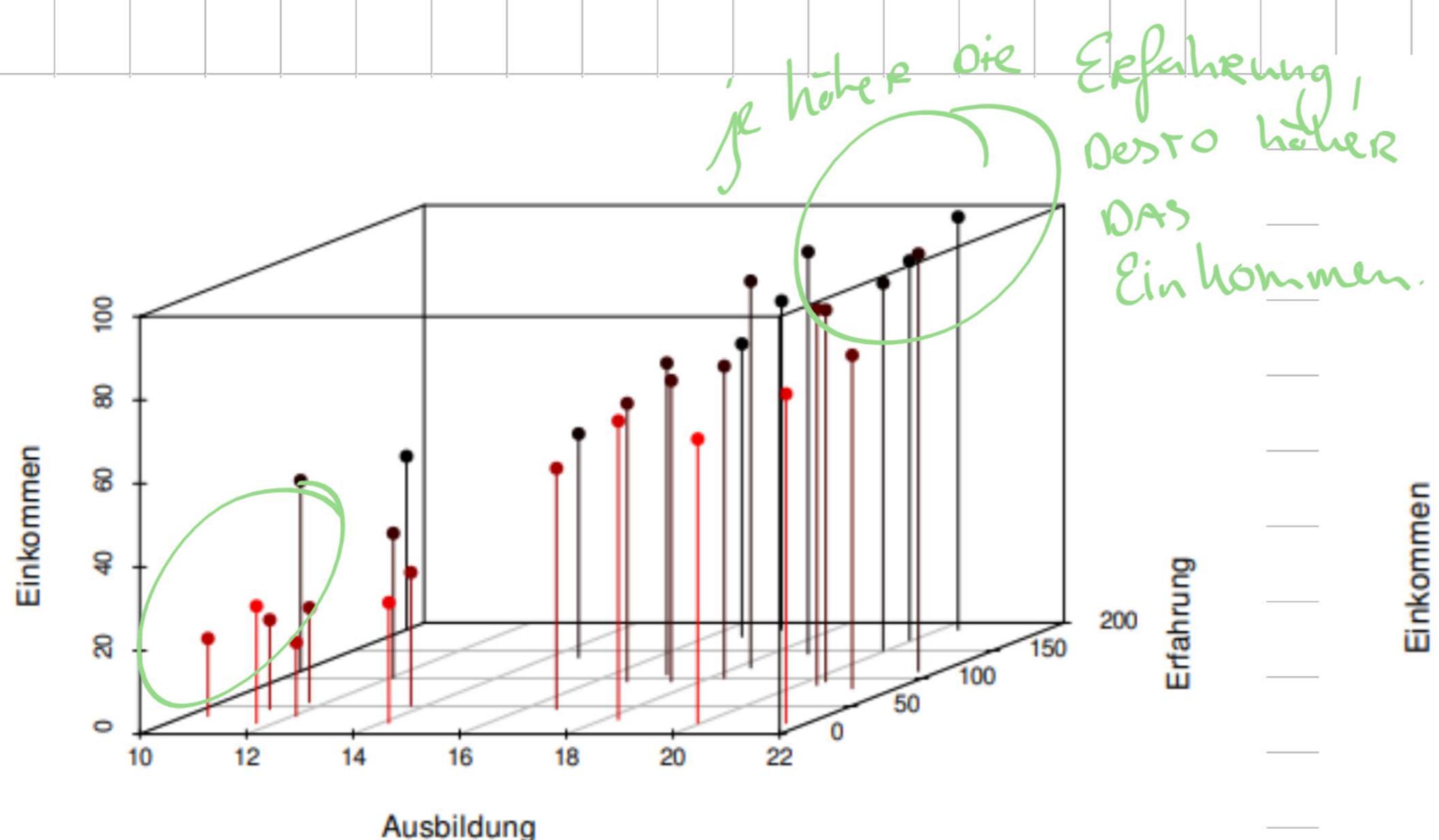
```
cor(data.frame(TV, Radio, Zeitung, Verkauf))
##           TV      Radio   Zeitung  Verkauf
## TV       1.0000000 0.05480866 0.05664787 0.7822244
## Radio    0.05480866 1.00000000 0.35410375 0.5762226
## Zeitung  0.05664787 0.35410375 1.00000000 0.2282990
## Verkauf  0.78222442 0.57622257 0.22829903 1.0000000
```

```
summary(lm(Verkauf ~ TV + Radio + Zeitung))
##
## Call:
## lm(formula = Verkauf ~ TV + Radio + Zeitung)
##
## Residuals:
##   Min     1Q Median     3Q    Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.938889  0.311908  9.422 <2e-16 ***
## TV          0.045765  0.001395 32.809 <2e-16 ***
## Radio       0.188530  0.008611 21.893 <2e-16 ***
## Zeitung     -0.001037  0.005871 -0.177  0.86
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF, p-value: < 2.2e-16
```

p zum F Wert
Signifikanz des Modells

In Märkten, wo mehr in die Werbung fürs Radio investiert wird, ist auch die Werbung für die Zeitung grösser, aufgrund des Korrelationskoeffizienten von 0.35. Aber Zeitungswerbung beeinflusst Verkäufe nicht. Zeitung schmückt sich hier mit fremden Lorbeeren, nämlich dem Erfolg von Radio auf Verkauf.

Zuerst entscheiden ob die erklärenden Variablen Einfluss auf die Zielgröße haben und dann ein Modell aufstellen, welches nur diese Variablen enthält. Interaktionseffekt: lm(medv~lstat*age)



z.B. anhand 2 Variablen "Erfahrung & Ausbildung" bereits 3D-Ebene.
d.h. wir können nicht über 2 Variablen multiple lineare Regression darstellen.

- Unterschied einfache Regression vs. multiple Regression:

z.B. anhand 3 Variablen: Verkauf, TV, Radio

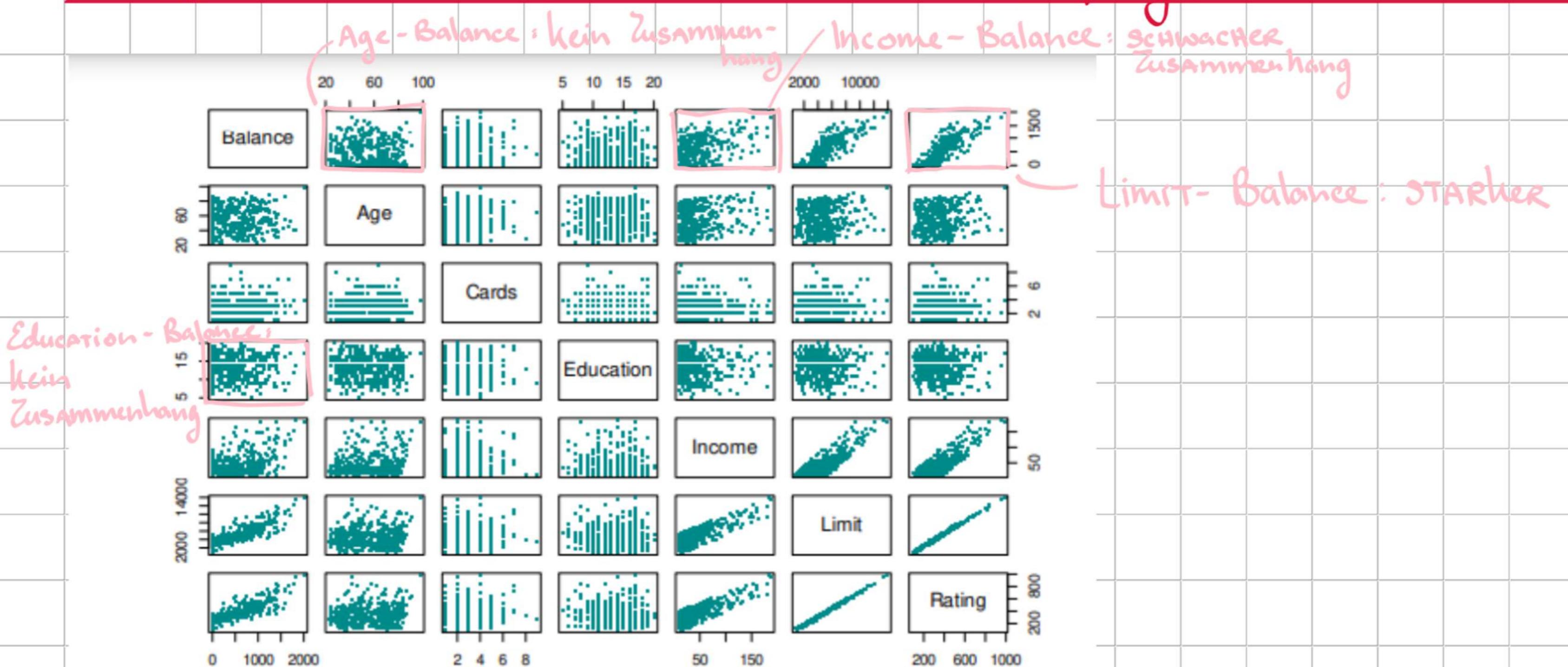
einfache R: TV & Radio werden ignoriert.

multiple lin. R: TV & Radio werden festgehalten (d.h. wir lassen die Variablen gleich)

absurdes Beispiel: Zusammenhang zwischen Haiattacken & Glaceverkäufen an einem bestimmten Strand. Je höher die Verkäufe desto mehr Attacken gibt es. Absurdo...
Aber wenn man jetzt Variable Temperatur noch rein nimmt, sieht die Geschichte anderes aus.

SW13 - qualitative Variablen & Variablenelektion

→ Qualitative Variablen kommt an der Prüfung, Variablenelektion nicht!



Beispiel

- Für Gender:

$$x_i = \begin{cases} 1 & \text{falls } i\text{-te Person weiblich} \\ 0 & \text{falls } i\text{-te Person männlich} \end{cases}$$

- Verwenden diese Variable als erklärende Variable im Regressionsmodell
- Modell:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{falls } i\text{-te Person weiblich} \\ \beta_0 + \varepsilon_i & \text{falls } i\text{-te Person männlich} \end{cases}$$

- β_0 : durchschn. Kreditkartenrechnungen der Männer
- $\beta_0 + \beta_1$: durchschn. Kreditkartenrechnungen der Frauen
- β_1 : durchschn. Unterschied der Rechnungen Männern/Frauen

One-Hot Encoding: Wir versuchen eine Variable auf maximal 2 Zustände zu beschränken (wie bei boolean false - true) (oder 1-0)

$\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$
nur einer dieser X darf 1 sein, Rest muss 0 sein

Beispiel
• Variable Ethnicity : Drei mögliche Levels
• Wählen zwei verschiedene Indikatorvariablen
• Wahl der 1. Indikatorvariable:
$x_{i1} = \begin{cases} 1 & \text{falls } i\text{-te Person asiatisch} \\ 0 & \text{falls } i\text{-te Person nicht asiatisch} \end{cases}$
• 2. Indikatorvariable:
$x_{i2} = \begin{cases} 1 & \text{falls } i\text{-te Person kaukasisch} \\ 0 & \text{falls } i\text{-te Person nicht kaukasisch} \end{cases}$
• Beide Variablen in Regressionsgleichung aufnehmen:
$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{falls } i\text{-te Person asiatisch} \\ \beta_0 + \beta_2 + \varepsilon_i & \text{falls } i\text{-te Person kaukasisch} \\ \beta_0 + \varepsilon_i & \text{falls } i\text{-te Person afroamerikanisch} \end{cases}$
• β_0 : Durchschn. Kreditkartenrechnungen von Afroamerikanern
• β_1 : Differenz der durchschn. Rechnungen von Afroamerikanern und Asiaten
• β_2 : Differenz der durchschn. Rechnungen von Afroamerikanern und Kaukasien

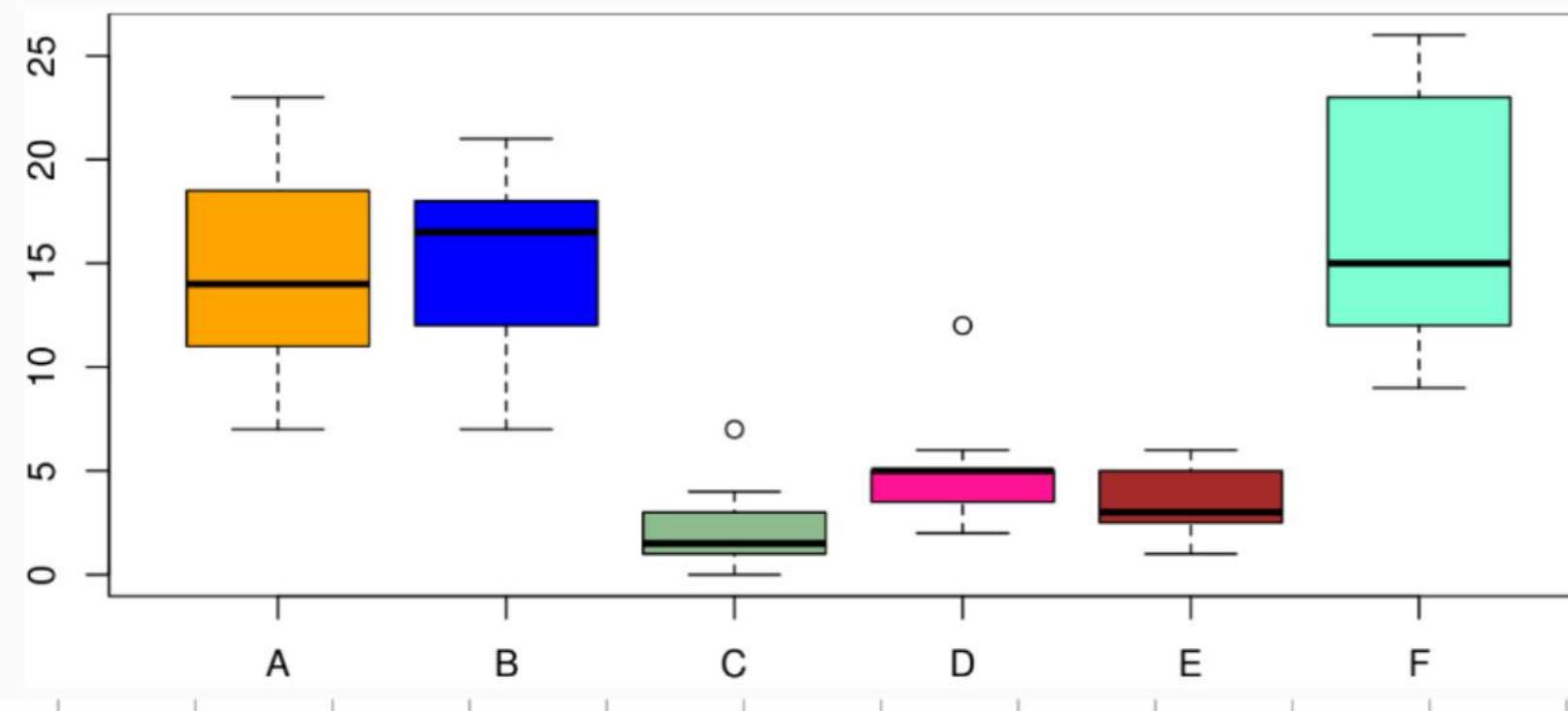
SWI4 - PROBEMEP

Boxplot

Frage 1 von 10 (4 Punkte)

Nicht beantwortet

Wir testen Insektsprays. Dabei wurden 6 verschiedene Insektsprays verwendet, die auf verschiedenen Feldern versprüht wurden. Danach wurde die Anzahl Insekten gezählt, die sich auf dementsprechenden Feld nach dem Besprühen befanden. Je kleiner die Anzahl, umso wirksamer der Spray.



Beachten Sie: Falsche Antworten geben Punkteabzug.

Für jede Aussage muss entschieden werden: [richtig] oder [falsch]

richtig	falsch
<input checked="" type="radio"/>	<input type="radio"/>
<input checked="" type="radio"/>	<input checked="" type="radio"/>
<input type="radio"/>	<input checked="" type="radio"/>

Etwa 75% der Messwerte von Spray A sind ungefähr 7 oder höher.
Spray A ist wirksamer als Spray C
50% der Messwerte von Spray A liegen zwischen etwa 11 und 18
Für Spray F sind die Hälfte der Messwerte etwa 15 oder kleiner.

Wahrscheinlichkeit

Frage 3 von 10 (4 Punkte)

Nicht beantwortet

Bei einem Zufallsexperiment werden ein roter (r) und ein blauer (b) Würfel gleichzeitig geworfen. Wir nehmen an, dass sie „fair“ sind, d. h. die Augenzahlen 1 bis 6 eines Würfels treten mit gleicher Wahrscheinlichkeit auf.

Beachten Sie: Falsche Antworten ergeben Punkteabzug.

Für jede Aussage muss entschieden werden: [richtig] oder [falsch]

richtig	falsch
<input type="radio"/>	<input checked="" type="radio"/>
<input checked="" type="radio"/>	<input type="radio"/>
<input type="radio"/>	<input checked="" type="radio"/>
<input checked="" type="radio"/>	<input type="radio"/>

Die Wahrscheinlichkeit, dass der rote Würfel 5 ist, ist 1/36
 $r_5 b_1$ ist ein mögliches Elementarereignis
Die Wahrscheinlichkeit, dass das Produkt der Augenzahlen 7 ist, ist 1/6
Die Wahrscheinlichkeit, dass die Augensumme grösser 2 ist, ist 35/36

Bedingte Wahrscheinlichkeit

Frage 5 von 10 (4 Punkte)

Nicht beantwortet

Der Serumtest untersucht schwangere Frauen auf Babys mit Down-Syndrom. Der Serumtest ist ein sehr guter, aber nicht perfekter Test. Etwa 1 % der Babys haben das Down-Syndrom. Wenn das Baby das Down-Syndrom hat, besteht eine 90-prozentige Wahrscheinlichkeit, dass das Ergebnis positiv ausfällt. Wenn das Baby nicht betroffen ist, besteht immer noch eine 1-prozentige Wahrscheinlichkeit, dass das Ergebnis positiv sein wird. Eine schwangere Frau wurde getestet und das Ergebnis ist negativ. Wie gross ist die Wahrscheinlichkeit, dass Ihr Baby das Down-Syndrom hat?

(Gerundet auf 5 Dezimalstellen)

- 0.99898
- 0.00102
- 0.00141
- 0.99859

Normalverteilung

Frage 6 von 10 (4 Punkte)

Nicht beantwortet

Für die Körpergrösse von 18-20jährigen Männern ergibt sich ein Mittelwert von 1.80 m bei einer Standardabweichung von 7.4 cm. Die Körpergrösse kann als normalverteilt angesehen werden.

Mit welcher Wahrscheinlichkeit ist ein zufällig ausgewählter Mann dieser Altersgruppe kleiner als 1.90cm?

X sei die Zufallsvariable für die Körpergrösse eines zufällig ausgewählten Mannes.

Welche der folgenden Aussagen beschreibt die gesuchte Wahrscheinlichkeit?

Für jede Aussage muss entschieden werden: [richtig] oder [falsch]

richtig	falsch
<input checked="" type="radio"/>	<input type="radio"/>
<input checked="" type="radio"/>	<input type="radio"/>
<input type="radio"/>	<input checked="" type="radio"/>
<input checked="" type="radio"/>	<input type="radio"/>

$1 - P(X \geq 190)$
 $1 - \text{pnorm}(q=190, \text{mean}=180, \text{sd}=7.4)$
 $P(X < 190)$
 $\text{qnorm}(p=190, \text{mean}=180, \text{sd}=7.4)$

Regressionsgerade

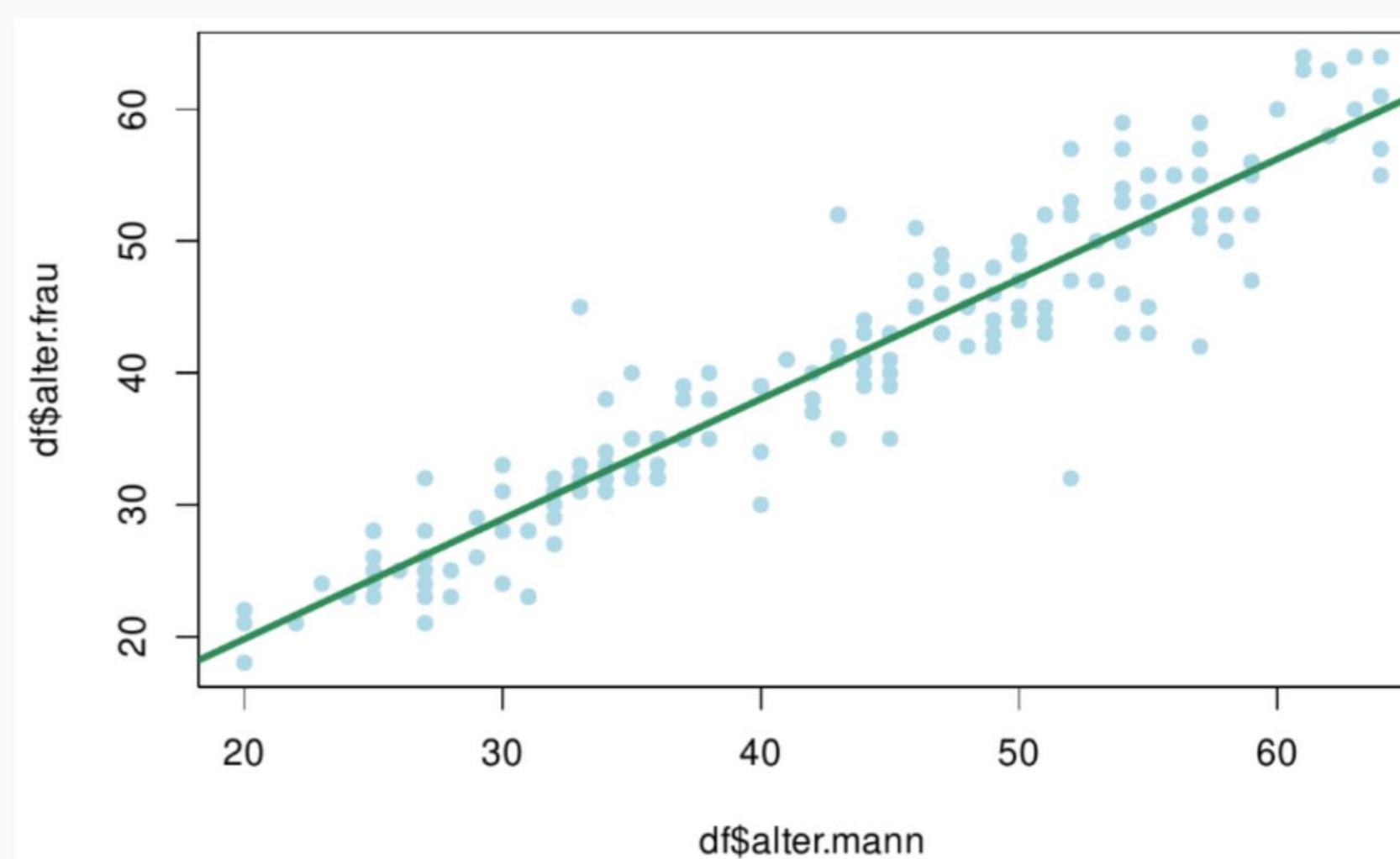
Frage 2 von 10 (4 Punkte)

Nicht beantwortet

Bitte lesen Sie die Aufgabe sorgfältig durch und überlegen Sie genau.

Wir haben aus eigener Erfahrung das Gefühl, dass bei Ehepaaren der Mann eher älter als die Frau ist. Nun wollen wir statistisch untersuchen, ob dem so ist. In einer Untersuchung in England wurden das Alter (in Jahren) und die Körpergrösse (in cm) von 170 Ehepaaren untersucht.

Das Streudiagramm sieht wie folgt aus:



Die Regressionsgerade wird wie folgt bestimmt:

```
lm(df$alter.frau ~ df$alter.mann)
##
## Call:
## lm(formula = df$alter.frau ~ df$alter.mann)
##
## Coefficients:
## (Intercept)  df$alter.mann
##           1.5740          0.9112
```

Beachten Sie: Falsche Antworten ergeben Punkteabzug.

Für jede Aussage muss entschieden werden: [richtig] oder [falsch]

richtig	falsch
<input checked="" type="radio"/>	<input type="radio"/>
<input checked="" type="radio"/>	<input type="radio"/>
<input checked="" type="radio"/>	<input type="radio"/>

Für jedes Jahr das der Ehemann älter ist, ist die Frau 0.911 Jahre älter
Die Regressionsgerade lautet $y = 1.574x + 0.9112$
Aus dem Streudiagramm ist ein klarer linearer Zusammenhang erkennbar.
Der Korrelationskoeffizient ist annähernd 0

Verteilung

Frage 4 von 10 (4 Punkte)

Nicht beantwortet

Ein Multiple-Choice-Test besteht aus 15 Fragen, mit jeweils 5 Antwortmöglichkeiten, von denen genau eine richtig ist. Die Wahrscheinlichkeit dafür, eine Aufgabe richtig zu beantworten, ist also 0.2. Die Wahrscheinlichkeits- und Verteilungsfunktion sind gegeben durch:

k	8	9	10	11	12	13	14	15
$P(X \leq k)$	0.711	0.939	0.969	0.982	0.989	0.992	0.999	1

Beachten Sie: Es handelt sich hier um die kumulierten Wahrscheinlichkeiten $P(X \leq k)$ und nicht $P(X = k)$.

Beachten Sie: Falsche Antworten geben Punkteabzug.

Für jede Aussage muss entschieden werden: [richtig] oder [falsch]

richtig	falsch
<input type="radio"/>	<input checked="" type="radio"/>
<input checked="" type="radio"/>	<input type="radio"/>
<input type="radio"/>	<input checked="" type="radio"/>

Die Wahrscheinlichkeit, dass genau 14 Fragen richtig beantwortet werden, ist 0.999
Die Wahrscheinlichkeit, dass mindestens 13 Fragen richtig beantwortet werden, ist 0.011
Die Wahrscheinlichkeit $P(X > 11)$ ist 0.989
Die Wahrscheinlichkeit $P(X \leq 10)$ ist 0.030

Hypothesentest 1

Frage 7 von 10 (4 Punkte)

Nicht beantwortet

Die Körpertemperatur von 10 Patienten wird zum Zeitpunkt der Verabreichung eines Medikaments (T₁) und 2 Stunden später (T₂) gemessen. Es soll geprüft werden, ob dieses Medikament eine fiebersenkende Wirkung hat.

Patient-Nr.	1	2	3	4	5	6	7	8	9	10
Temp. 1 in °C	39.1	39.3	38.9	40.6	39.5	38.4	38.6	39.0	38.6	39.2
Temp. 2 in °C	38.1	38.3	38.8	37.8	38.2	37.3	37.6	37.8	37.4	38.1

Wir führen einen Hypothesentest auf 5% durch um zu überprüfen, ob das Medikament fiebersenkend ist. Der R-Output zeigt:

```
## 
## Paired t-test
##
## data: t.1 and t.2
## t = 5.6569, df = 9, p-value = 0.0001554
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
## 0.7976252 Inf
## sample estimates:
## mean of the differences
## 1.18
```

Welche der folgenden Aussagen ist richtig?

Für jede Aussage muss entschieden werden: [richtig] oder [falsch]

- | richtig | falsch |
|----------------------------------|----------------------------------|
| <input type="radio"/> | <input checked="" type="radio"/> |
| <input type="radio"/> | <input checked="" type="radio"/> |
| <input checked="" type="radio"/> | <input type="radio"/> |
| <input type="radio"/> | <input checked="" type="radio"/> |
- Der Wert 1.18 in der letzten Linie bedeutet, dass die durchschnittliche Temperatur um 1.18 °C nach zwei Stunden grösser war.
Da 0 nicht im Vertrauensintervall ist, wird die Nullhypothese nicht verworfen.
Die Nullhypothese ist, dass das Medikament keine Wirkung hat.
Wir führen einen ungepaarten Test durch.

Einfache lineare Regression

Frage 9 von 10 (4 Punkte)

Nicht beantwortet

Die MASS-Bibliothek enthält den Boston-Datensatz, der medv (median house value in \$1000) für 506 Stadtviertel um Boston herum erfasst. Wir werden versuchen, medv mit 13 Prädiktoren wie rm (durchschnittliche Anzahl von Zimmern pro Haus), age (Durchschnittsalter der Häuser) und crim (Kriminalitätsrate) vorherzusagen.
Wir werden damit beginnen, die lm()-Funktion zu verwenden, um ein einfaches lineares Regressionsmodell mit medv als Zielvariable und crim als Prädiktor anzupassen. Wir erhalten folgenden Output:

```
lm.fit <- lm(medv ~ crim)
summary(lm.fit)

##
## Call:
## lm(formula = medv ~ crim)
##
## Residuals:
##   Min     1Q   Median     3Q    Max 
## -16.957 -5.449 -2.007  2.512 29.800 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 24.03311  0.40914  58.74   <2e-16 ***
## crim        -0.41519  0.04389 -9.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.484 on 504 degrees of freedom
## Multiple R-squared:  0.1508, Adjusted R-squared:  0.1491 
## F-statistic: 89.49 on 1 and 504 DF,  p-value: < 2.2e-16
```

Welche der folgenden Aussagen sind richtig?

Für jede Aussage muss entschieden werden: [richtig] oder [falsch]

- | richtig | falsch |
|----------------------------------|----------------------------------|
| <input type="radio"/> | <input checked="" type="radio"/> |
| <input type="radio"/> | <input checked="" type="radio"/> |
| <input type="radio"/> | <input checked="" type="radio"/> |
| <input checked="" type="radio"/> | <input type="radio"/> |
| <input type="radio"/> | <input checked="" type="radio"/> |
- Der y-Achsenabschnitt ist -0.4152
Das Modell lautet $\text{crim} = \beta_0 + \beta_1 \cdot \text{medv}$.
Der Wert von $R^2 = 0.1508$ bedeutet, dass die Daten gut zum Modell passen.
Die Steigung ist statistisch signifikant ungleich 0.

Hypothesentest 2

Aktionen ▾

Frage 8 von 10 (4 Punkte)

Nicht beantwortet

Ein U.S. Magazin, Consumer Reports, führte eine Untersuchung des Kalorien- und Salzgehaltes von verschiedenen Hotdog-Marken durch. Es gab drei verschiedene Typen von Hotdogs: Rind, „Fleisch“ (Rind, Schwein, Geflügel gemischt) und Geflügel.

Die Resultate unten führen den Kaloriengehalt verschiedener Marken von Rind- und Geflügel-Hotdogs auf.

Rinds-Hotdog: 186, 181, 176, 149, 184, 190, 158, 139, 175, 148, 152, 111, 141, 153, 190, 157, 131, 149, 135, 132

Geflügel-Hotdog: 129, 132, 102, 106, 94, 102, 87, 99, 170, 113, 135, 142, 86, 143, 152, 146, 144

Haben die beiden Hotdog-Arten verschiedenen Kaloriengehalt? Wir führen einen Hypothesentest auf 5% Signifikanzniveau durch und erhalten folgenden Output:

```
## 
## Wilcoxon rank sum test with continuity correction
##
## data: x and y
## W = 285.5, p-value = 0.0004549
## alternative hypothesis: true location shift is not equal to 0
```

x enthält die Rindsdaten und y die Geflügeldaten.

Welche der folgenden Aussagen sind richtig?

Für jede Aussage muss entschieden werden: [richtig] oder [falsch]

- | richtig | falsch |
|----------------------------------|----------------------------------|
| <input type="radio"/> | <input checked="" type="radio"/> |
| <input checked="" type="radio"/> | <input type="radio"/> |
| <input type="radio"/> | <input type="radio"/> |
| <input type="radio"/> | <input checked="" type="radio"/> |
- Da die Stichprobengrößen unterschiedlich sind, führen wir einen gepaarten Test durch.
Die Alternativhypothese ist $\mu_x \neq \mu_y$.
Wir nehmen an, dass die Daten nicht normalverteilt sind.
Der Unterschied zwischen x und y ist statistisch signifikant.

Multiple lineare Regression

Aktionen ▾

Frage 10 von 10 (4 Punkte)

Nicht beantwortet

Die MASS-Bibliothek enthält den Boston-Datensatz, der medv (median house value in \$1000)

für 506 Stadtviertel um Boston herum erfasst. Wir werden versuchen, medv mit 13 Prädiktoren wie rm (durchschnittliche Anzahl von Zimmern pro Haus), dis (Distanz vom Zentrum von Boston) und crim (Anteil der Kriminalität) vorherzusagen.

Wir passen ein multiples lineares Regressionsmodell mit der Zielvariable medv und den Prädiktoren crim, dis und rm. Das Signifikanzniveau ist 5%. Wir erhalten folgenden Output:

```
fit <- lm(medv ~ crim + rm + dis, data = Boston)
summary(fit)

##
## Call:
## lm(formula = medv ~ crim + rm + dis, data = Boston)
##
## Residuals:
##   Min     1Q   Median     3Q    Max 
## -21.247 -2.930 -0.572  2.390 39.072 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -29.45838  2.60010 -11.330 < 2e-16 ***
## crim         -0.25405  0.03532  -7.193 2.32e-12 ***
## rm            8.34257  0.40870  20.413 < 2e-16 ***
## dis           0.12627  0.14382   0.878   0.38    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.238 on 502 degrees of freedom
## Multiple R-squared:  0.5427, Adjusted R-squared:  0.5399 
## F-statistic: 198.6 on 3 and 502 DF,  p-value: < 2.2e-16
```

Welche der folgenden Aussagen sind richtig?

Für jede Aussage muss entschieden werden: [richtig] oder [falsch]

- | richtig | falsch |
|----------------------------------|----------------------------------|
| <input type="radio"/> | <input checked="" type="radio"/> |
| <input checked="" type="radio"/> | <input type="radio"/> |
| <input type="radio"/> | <input checked="" type="radio"/> |
| <input checked="" type="radio"/> | <input type="radio"/> |
- Der Wert 0.54 bedeutet, dass etwas 50% der Variablen einen Einfluss auf medv haben.
Der p-Wert um F-Wert ist signifikant und somit hängt mindestens ein Prädiktor mit der Zielvariable zusammen.
Der p-Wert für dis ist grösser als das Signifikanzniveau und somit wird die Alternativhypothese $\beta_3 = 0$ angenommen.
Der Koeffizient für rm bedeutet, dass pro Zimmer mehr, der Medianpreis um etwa \$8300 steigt.

Quiz SERIE 01

In R wird ein Vektor temp mit 10 Werten definiert.

Welche Behauptungen zu den R Befehlen plot(...) und abline(...) sind korrekt?

(Richtige Angaben geben 1 Punkt, falsche Angaben geben 0.5 Punkte Abzug.)

- Mit `plot(temp, type="p")` werden die Werte von `temp` als runde Kreise im Plot eingezeichnet. → Type "l" = Linie, Type "b" = Beides aka Linie & Punkte, Type "p" = Punkte
- Mit `abline(v=4)` wird in den durch `plot(temp)` erstellten Plot zusätzlich eine horizontale Linie mit Abstand 4 von der x Achse eingezeichnet. → es wäre vertikal.
- Mit `abline(a=0,b=1)` wird in den durch `plot(temp)` erstellten Plot zusätzlich eine Gerade mit der Gleichung $y = x$ eingezeichnet. → Gerade mit y-Achsenabschnitt $a=0$ & Steigung $b=1$
- Mit dem einzigen Befehl `abline(a=1,b=3)` wird ein Plot mit einer Linie erstellt, auch wenn vorher der Befehl `plot(temp)` nicht aufgerufen wird.
- Es ist nicht möglich einzig mit dem Befehl `plot(...)` die Werte von `temp` als runde Kreise zu zeichnen, so dass diese auch mit einer durchgezogenen Linie verbunden sind.
- Mit `plot(temp,type="l",lty=2)` werden die Werte von `temp` durch eine gestrichelte Linie verbunden eingezeichnet. → hy macht Linie gestrichelt.

Welche Behauptungen zum R Befehl `seq(...)` sind korrekt?

(Richtige Angaben geben 1 Punkt, falsche Angaben geben 0.5 Punkte Abzug.)

- Der Befehl `seq(from=10,to=20,length.out=11)` gibt die Werte 10 11 12 13 14 15 16 17 18 19 20 zurück. → macht den Abstand gleichmäßig
- Der Befehl `seq(from=10,to=20,length.out=100)` gibt 100 Werte zurück. → gibt immer diese Anzahl raus, egal was bei To = ... steht.
- Der Befehl `seq(from=10,to=20,by=3)` gibt fünf Werte zurück. → gibt 10, 13, 16, 19.
- Der Befehl `seq(from=10,to=20,by=2)` gibt die Werte 10 12 14 16 18 20 zurück.

Seq (from = 3, to = 10, by = 2)

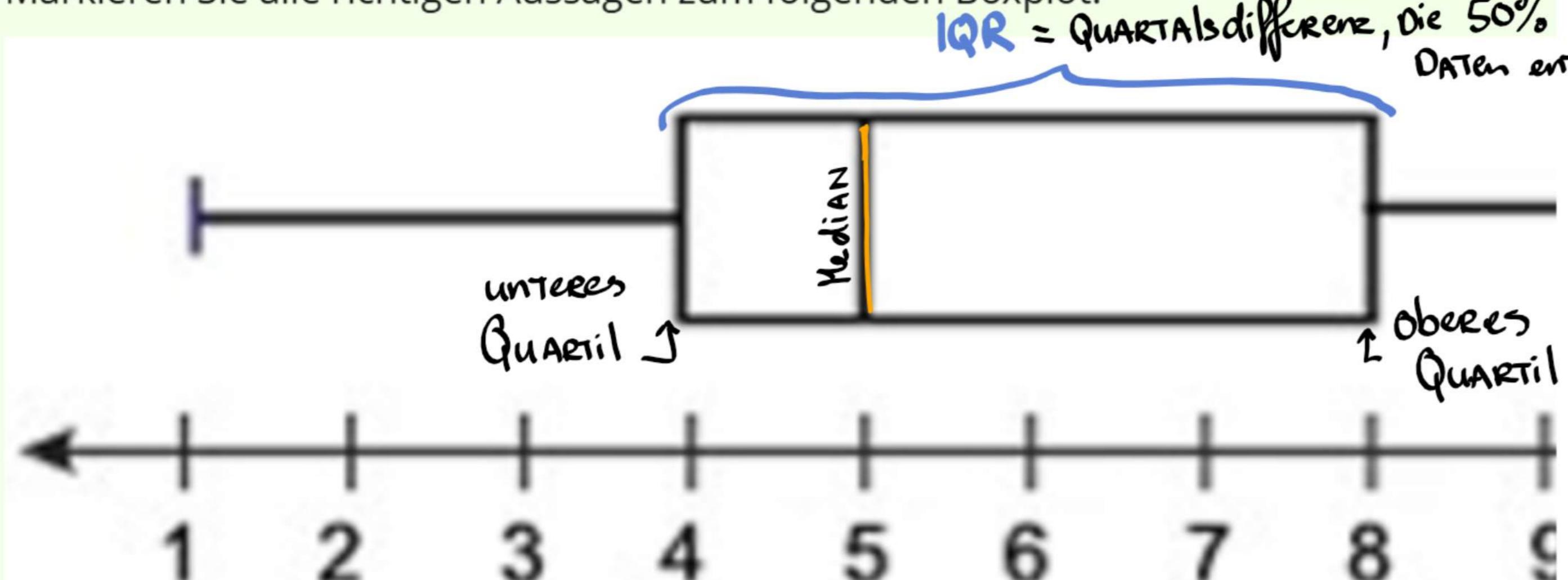
bildet eine Folge von Zahlen
beginnt hier ↑ SCHRITTlänge
hier damit auf, sofern möglich

Quiz SERIE 02

Boxplot

Markieren Sie alle richtigen Aussagen zum folgenden Boxplot:

IQR = QUARTALSdifferenz, die 50% aller Daten enthält.



- Der kleinste normale Wert ist 1. → alles, was hier ist, ist Teil des Normalbereichs.
- Der kleinste Ausreißer ist 1. → 1 ist noch im Normalbereich. Alles kleiner wahrscheinlich ja.
- $IQR = 8 \rightarrow Q3 - Q1 = 8 - 4 = 4$. Bestimme den Interquartilsabstand (IQR) des Datensatzes.
- $Q_3 = 5 \rightarrow Q_3$ fängt bei 8 an.
- Median = 5

$$\begin{array}{c} 5 \text{ Punkte} \\ 4, 5, 6, 8, 9, 10, 11, 13 \\ \frac{5+6}{2} = 5,5 \quad \frac{8+9}{2} = 8,5 \end{array}$$

$$IQR = 10,5 - 5,5 = 5$$

Median & IQR

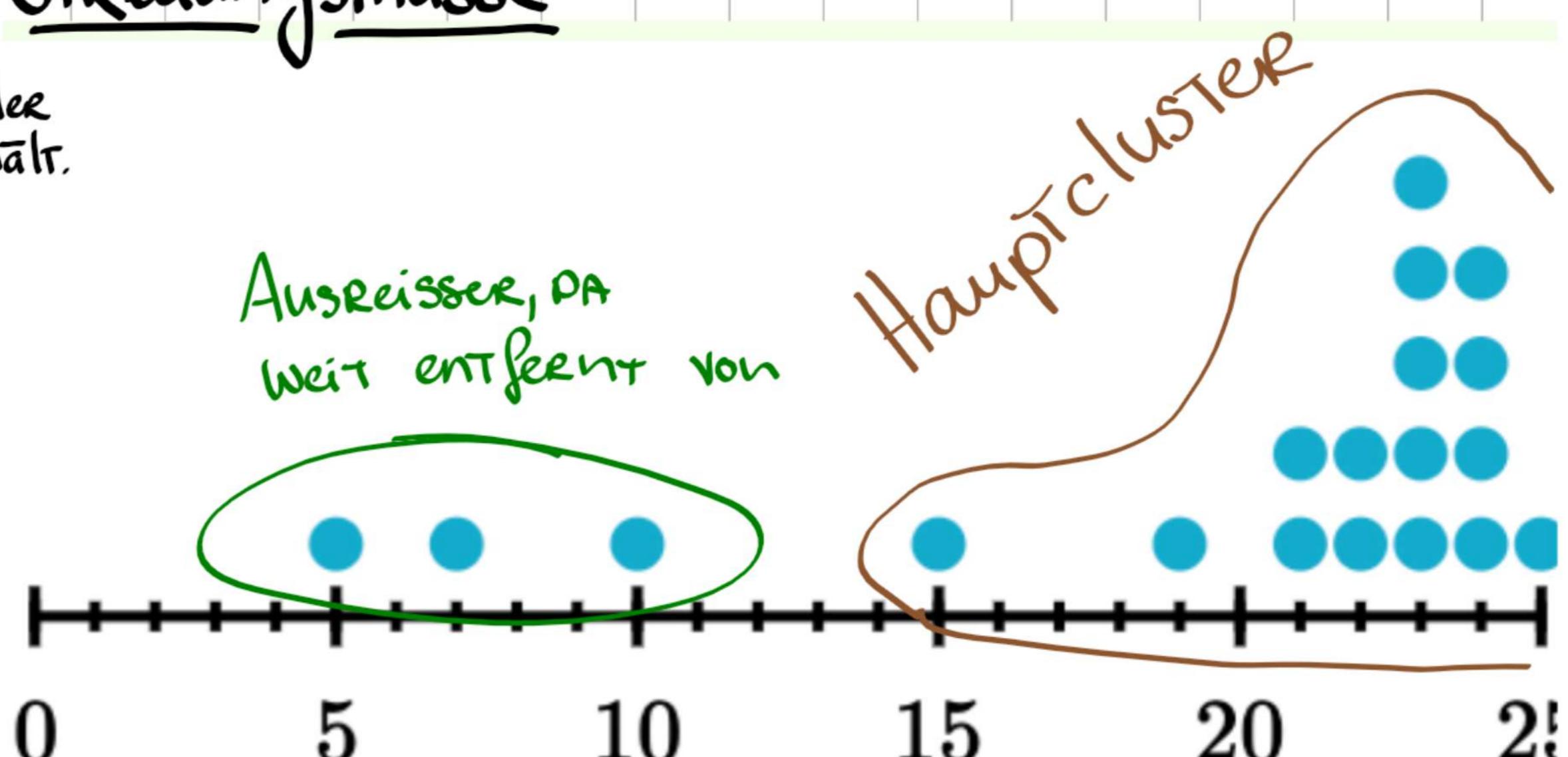
Bestmögliche Lösung:

$$\overbrace{20, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}^{Q_1} = \overbrace{13, 14, 15, 16, 17, 18, 19, 20}^{Q_3}, 50$$

Wir haben einen Datensatz mit 12 Zahlen, wobei die kleinste Zahl 20 ist, und die grösste Zahl 50 beträgt. Sei Q_1 das untere Quartile, und Q_3 das obere Quartile, sowie $Q_1 = Q_3$. Markieren Sie alle richtigen Aussagen:

- Median = $Q_3 \rightarrow$ wenn 75% aller Werte denselben Wert haben, dann ist $Q_1 = Q_2 = Q_3$. Aber eher ungewöhnlich.
- Es ist möglich, dass der Median 50 beträgt. → es kann sein, dass die 12 Ziffern 50 sind.
- Median $\neq Q_2 \rightarrow Q_2$ und Median SIND das Gleiche.
- Median = $\frac{Q_1+Q_3}{2}$

Streuungsmasse



Wieviele Werte können wir als untere Ausreißer bezeichnen:

- 2
- 5
- Keine
- 3
- 4

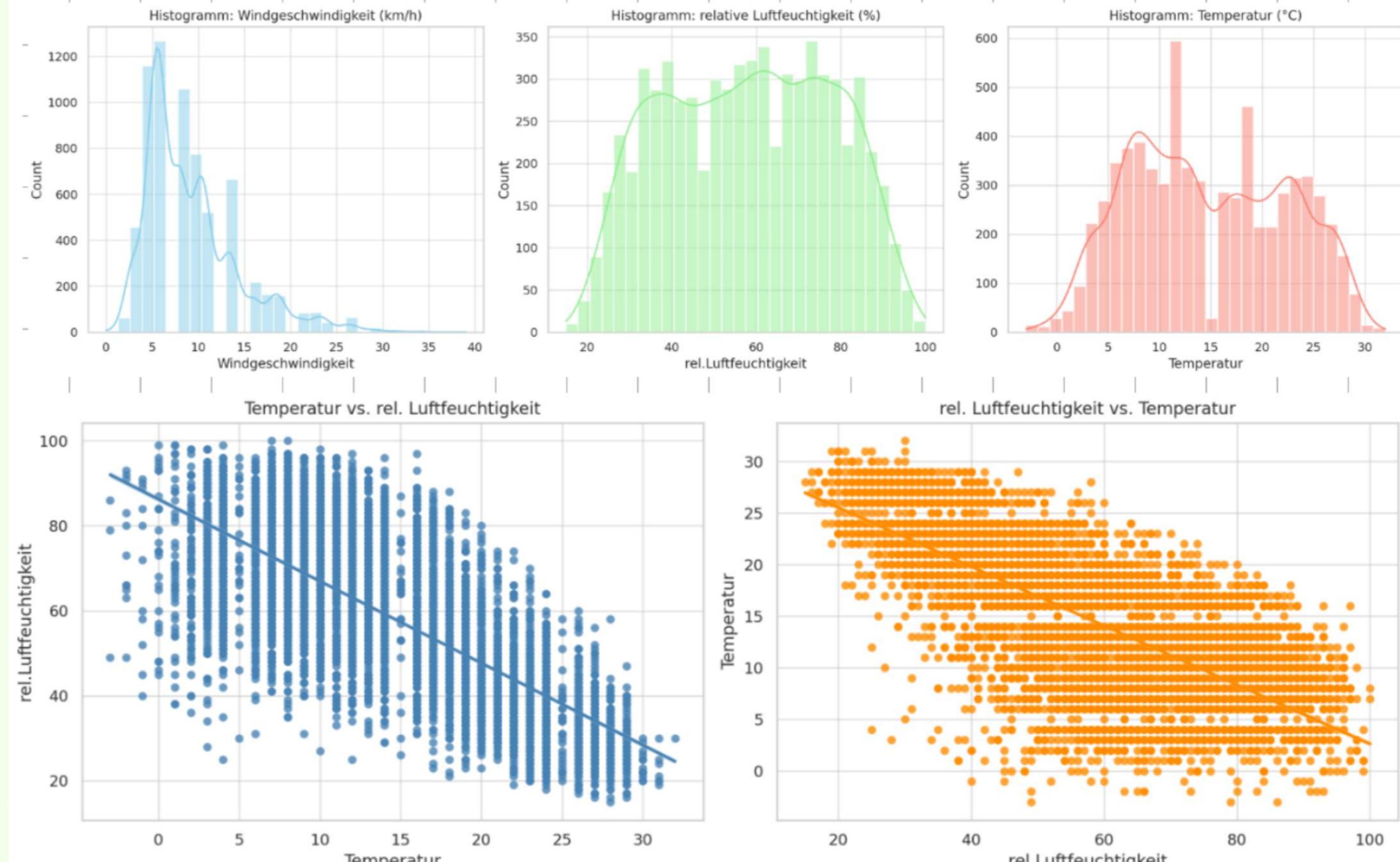
Quiz SERIE 03

Histogramme & Streudiagramme mit Regressionsgerade

In der csv Datei "MadridWetterdaten1997-2015" (download vor diesem Quiz in ILIAS) sind die täglich in Madrid gemessenen Werte für die Temperatur in Grad Celsius, für die relative Luftfeuchtigkeit in Prozent, für den Luftdruck in hPa und für die Windgeschwindigkeit in km/h von 1997 bis 2015 gespeichert.

Laden Sie diese Datei in R und entscheiden Sie durch eine entsprechende Datenanalyse in R, welche der folgenden Behauptungen wahr sind:

- Das Histogramm für die Windgeschwindigkeiten in Madrid ist rechtsschief.
- Wenn ein Histogramm für einen Datensatz eine symmetrische Form hat, dann muss auch das entsprechende normierte Histogramm eine symmetrische Form haben.
- Im Streudiagramm mit der Temperatur auf der horizontalen Achse und der relativen Luftfeuchtigkeit auf der vertikalen Achse hat die Regressionsgerade die Gleichung $y = -1.928 + 86.236x$.
- Im Streudiagramm mit der Temperatur auf der horizontalen Achse und der relativen Luftfeuchtigkeit auf der vertikalen Achse hat die Regressionsgerade die Gleichung $y = 86.236 + 1.928x$.
- Im Streudiagramm mit der Windgeschwindigkeit auf der horizontalen Achse und der relativen Luftfeuchtigkeit auf der vertikalen Achse sind die Punkte recht zerstreut, folgen aber einer leicht fallenden Regressionsgeraden.
- Die Daten zeigen, dass bei grösseren Temperaturen die relative Luftfeuchtigkeit kleiner ist.
- Mit dem Befehl `abline(lm(windspeed ~ temperature), col="red")` wird in das Streudiagramm mit windspeed auf der horizontalen Achse und temperature auf der vertikalen Achse die passende Regressionsgerade in roter Farbe gezeichnet.
- Das Histogramm für die relative Luftfeuchtigkeit in Madrid ist linksschief.
- Wenn die Windgeschwindigkeit in Madrid 0 km/h ist, beträgt der Luftdruck gemäss der Regressionsgerade schätzungsweise 1022.9076 hPa.
- Am häufigsten wurden Luftdruckwerte zwischen 1015 und 1020 hPa gemessen.
- Für die Temperaturwerte erstellt R mit `breaks=25` ein Histogramm mit weniger Klassen als mit `breaks=30`.
- Wenn in Madrid die Windgeschwindigkeit um 1 km/h steigt, dann nimmt der Luftdruck gemäss der Regressionsgerade schätzungsweise um 0.5381 hPa ab.



Quiz SERIE 04

Wahrscheinlichkeit 2

Bestmögliche Lösung:

$$\begin{aligned}
 &\cdot 300 \text{ Lose } \rightarrow \text{Einnahmen} = 300 \cdot 3 = 900 \\
 &\cdot \text{W'keit zu gewinnen} = 0.2 \rightarrow \text{Erwartete Gewinner} = 300 \cdot 0.2 = 60 \\
 &\cdot \text{Preis pro Gewinn} = 8 \rightarrow \text{Auszahlung} = 60 \cdot 8 = 480 \\
 &\cdot \text{Spende} = 900 - 480 = 420
 \end{aligned}$$

Wahrscheinlichkeit

Bestmögliche Lösung:

Ein Sack enthält 10 Holzscheiben. Die Scheiben sind beschriftet mit den Zahlen 1 bis 10. Eine Scheibe wird zufällig aus dem Sack gezogen. Die Wahrscheinlichkeit, als Dezimalzahl, dass die gezogene Scheibe Jede Zahl hat die W'keit von $\frac{1}{10} = 0.1$

- die Zahl 3 ist, beträgt $0.1 \rightarrow \frac{1}{10}$
- kleiner als 4 ist, beträgt $0.3 \rightarrow \text{Zahlen } 1, 2, 3 \rightarrow \frac{1}{10} + \frac{1}{10} + \frac{1}{10} = \frac{3}{10}$
- eine quadratische Zahl ist, beträgt $0.3 \rightarrow \text{Quadratische Zahlen: } 1, 4, 9 = \frac{3}{10}$
- eine Primzahl ist, beträgt $0.4 \rightarrow \text{Primzahlen } \leq 10: 2, 3, 5, 7 \rightarrow \frac{4}{10}$

Bestmögliche Lösung: Regression & Korrelation

Eine Regressionsgerade $y = a + bx$ beschreibt eine Korrelation. Kreuzen Sie alle richtigen Aussagen an:
Punkt wo Gerade y-Achse schneidet (also $x=0$)

- Wenn $y = 1$ für alle x eines Datensatzes, dann besteht keine Korrelation.
- Bei einem negativen Achsenabschnitt ist $y < 0$ wenn $x = 0 \rightarrow$ per Definition
- Gegeben ist ein Datenpunkt (x_4, y_4) und das Residuum $r_4 < 0$. Es ist daher zwingend dass $y_4 > a + bx_4 \rightarrow$ Ein negatives Residuum heißt (negativ) y liegt unter Geraden
- Bei einer negativen Korrelation ist $a < 0 \rightarrow$ hängt davon ab, wie die Gerade y trifft.

b = Steigung, lässt sich mit R vergleichen.

$r < 0 \rightarrow b < 0 \rightarrow$ Gerade fällt
 $r > 0 \rightarrow b > 0 \rightarrow$ Gerade steigt

Residuum = "Fehler" bei einer Vorhersage, also der Unterschied zwischen dem tatsächlichen Wert & dem Wert auf der Regressionsgeraden

Z.B. wenn $r = -0.5$, dann ist der tatsächliche Wert kleiner als der vorhergesagte Wert

$$\begin{aligned}
 r > 0 &\rightarrow \text{positiv steigend} \\
 r < 0 &\rightarrow \text{negativ fallend} \\
 r = 0 &\rightarrow \text{kein linearer Zusammenhang}
 \end{aligned}$$

Quiz SERIE 05

Adventslos von Swisslos

Swisslos verkauft jedes Jahr in der Weihnachtszeit ein Adventslos, bei dem unterschiedliche Geldbeträge k ausbezahlt werden. Aus den Angaben zur Losauflage kann man die Wahrscheinlichkeiten $P(X = k)$ für die Auszahlung k eines zufällig gekauften Loses ausrechnen. Die Daten zu den Auszahlungen k und ihre zugehörigen Wahrscheinlichkeiten P werden in R als Vektoren wie folgt eingelesen:

mögliche Auszahlungsbeträge \Rightarrow

```
k <- c(0,100,150,200,250,300,350,400,450,500,550,600,650,1000,1100,5000,100000,1000000)
P <- c(0.662256,0.28934,0.015883,0.029239,0.00085,0.000132,
      0.000132,0.000199,0.00011,0.000806,0.00011,0.000066,
      0.000022,0.000773,0.000055,0.000006,0.000015,0.000006)
```

jeweilige Wkheit
für diese
Auszahlung

Kopieren Sie diese Zuweisungen in R und berechnen Sie mit einer Genauigkeit auf zwei Stellen nach dem Komma für die Zufallsvariable X der Auszahlungen

- den Erwartungswert: 46.47255
- und die Standardabweichung: 2480.656

- Ein Los kostet 100 Franken. Der durchschnittlich zu erwartende Verlust pro Los ist

also: 53.52745

(1 Los = 100,-) \rightarrow Verlust = 100 - E(x) = 53.53

Kumulierte Wahrscheinlichkeiten

Bestmögliche Lösung:

Die **kumulierten** Wahrscheinlichkeiten einer Zufallsvariablen X , die vier verschiedene Werte $k = 3, 5, 9, 10$ realisieren kann, sind:

k	3	5	9	10
$P(X \leq k)$	0.3	0.5	0.9	1

Welche der folgenden Behauptungen sind wahr?

- $P(X = 8) = 0 \rightarrow 8$ kommt nicht im Definitionsbereich vor.
- $P(X \geq 10) = 0 \rightarrow$ Es gibt keine grösseren Werte als 10.
- $P(X > 5) = 0.5 \rightarrow$ JA, weil $1 - 0.5 = 0.5$.
- $P(X = 5) = 0.5 \rightarrow$ Nein, weil kumulierte Wkten.

Ein gezinkter Würfel

Bereits in altägyptischen Gräbern von 3500 v. Chr. wurden manipulierte Würfel gefunden. Aufgrund der Lage des Schwerpunkts hat jeder solche Würfel eine andere Wahrscheinlichkeitsverteilung für seine Auflageflächen. Die Wahrscheinlichkeiten für die Augenzahlen k bei einem Würfel aus dem antiken Rom sind:

k	1	2	3	4	5	6
$P(X = k)$	0.01	0.11	0.14	0.24	0.17	0.33

Berechnen Sie folgende Wahrscheinlichkeiten exakt als Dezimalzahl zwischen 0 und 1:

- Beim Werfen des Würfels kommt eine ungerade Zahl mit einer Wahrscheinlichkeit von 0.32 vor. Ungerade: 1,3,5 $\rightarrow P(X \text{ ungerade}) = P(1) + P(3) + P(5) = \frac{0.01 + 0.14 + 0.17}{0.32}$
- Die Wahrscheinlichkeit $P(X \geq 5 \text{ und } X \leq 2)$ beträgt 0. \rightarrow Unmöglich
- Die Wahrscheinlichkeit $P(3 \leq X < 5)$ beträgt 0.38. $\rightarrow P(3) + P(4) = 0.14 + 0.24 = 0.38$
- Höchstens eine Zahl 4 zu werfen kommt bei diesem Würfel mit Wahrscheinlichkeit 0.5 vor. $\rightarrow P(1) + P(2) + P(3) + P(4) = 0.01 + 0.11 + 0.14 + 0.24 = 0.5$

$$\text{In R: } \begin{aligned} \text{VarX} &\leftarrow \text{sum}(p \cdot (k - \bar{x})^2) \\ \text{SDX} &\leftarrow \text{sqrt}(\text{VarX}) \end{aligned}$$

Quiz SERIE 06

Stochastisch unabhängig

Angenommen die beiden Ereignisse A und B seien stochastisch unabhängig. Kreuzen Sie an, welche Aussagen richtig sind.

- $P(A|B) = P(A)$
- $P(A \cap B) = P(A) \cdot P(B)$
- Generell gilt, wenn A eintritt, dann ist $P(B) = 0$.
- Da stochastisch unabhängig, gilt $A \cap B = \{\}$ und daher $P(A \cap B) = 0$.

Bedingte Wahrscheinlichkeit mit Batterien

Ein Unternehmen, das Handy-Batterien herstellt, hat zwei Fabriken: Fabrik A und Fabrik B. Fabrik A produziert 60% der Batterien, während Fabrik B die restlichen 40% produziert. Die Batterien von Fabrik A haben eine Ausfallrate von 2%, während die Batterien von Fabrik B eine Ausfallrate von 3% haben.

Wenn zufällig eine Batterie aus dem Bestand des Unternehmens ausgewählt wird und sich als defekt erweist, wie hoch ist die Wahrscheinlichkeit, dass sie in Fabrik B hergestellt wurde? $P(B|D)$

Der Wert muss zwischen 0.4999 und 0.5001 liegen

Fabrik A : 60% Batterien $\rightarrow P(A) = 0.6$

Fabrik B : 40% Batterien $\rightarrow P(B) = 0.4$

Ausfallrate A: 2% $\rightarrow P(D|A) = 0.02$

Ausfallrate B: 3% $\rightarrow P(D|B) = 0.03$

Bedingte Wahrscheinlichkeit

Angenommen in einem totalitären Regime werden alle Personen eingesperrt, die demonstrieren. Weiter nehme man an, die Wahrscheinlichkeiten, dass eine beliebige Person eingesperrt ist oder demonstriert seien beide 0.01. Wir bezeichnen das Ereignis dass jemand eingesperrt ist als E und das Ereignis dass jemand demonstriert als D. Kreuzen Sie alle richtigen Aussagen an. Tip: Um die richtige Antwort einer der Aussagen herzuleiten, verwenden Sie am besten das Gesetz der Totalen Wahrscheinlichkeit, und lösen daraus für die bedingte Wahrscheinlichkeit.

- E und D sind stochastisch unabhängig. \rightarrow zu tun, d.h. Regierung speziell willkürlich ein.
- $P(E|\bar{D}) = 0 \rightarrow$ Wer nicht demonstriert, wird nicht eingesperrt
- $P(E) = 0.01 \rightarrow 1\%$ der Bevölkerung eingesperrt
- $P(E) = 0.99 \rightarrow$ Ein Normalo wird mit 99% eingesperrt.

$$P(E) = 0.01 \rightarrow \text{Normalo}$$

$$P(D) = 0.01 \rightarrow \text{Demonstrant}$$

$$P(B|D) = \frac{P(D|B) \cdot P(B)}{P(D)}$$

Satz von Bayes

$$\hookrightarrow P(B|D) = \frac{0.03 \cdot 0.4}{0.024} = \frac{0.12}{0.024} = 0.5$$

Quiz SERIE 07

Blutdruckwerte

In einer gross angelegten Studie konnte 2015 gezeigt werden, dass oszillometrische Messgeräte den Blutdruck angenähert **normalverteilt** messen. Beim **systolischen Wert** haben diese Geräte eine Standardabweichung von 4.4 mmHg und beim **diastolischen Wert** eine Standardabweichung von 3.4 mmHg.

1. Mit welcher Wahrscheinlichkeit (Zahl zwischen 0 und 1; auf vier Stellen nach dem Komma gerundet) misst ein solches Messgerät bei einem Mann mit einem systolischen Blutdruck von 120 mmHg Werte grösser als 130 mmHg? **0.0115**

2. Mit welcher Wahrscheinlichkeit (Zahl zwischen 0 und 1; auf vier Stellen nach dem Komma gerundet) misst ein solches Messgerät bei einer Frau mit einem diastolischen Blutdruck von 80 mmHg Werte zwischen 78 und 82 mmHg? **0.4436**

3. In welchem symmetrischen Bereich (Werte gerundet auf zwei Stellen nach dem Komma) um 120 mmHg beim systolischen bzw. um 80 mmHg beim diastolischen Blutdruck liegen die Werte eines oszillometrischen Messgeräts zu 90% Wahrscheinlichkeit?

- Beim systolischen Blutdruck im Bereich von **112.76** bis **127.24**.

- Beim diastolischen Blutdruck im Bereich von **74.41** bis **85.59**.

$$1) P(X > 130), \mu = 120, \sigma = 4.4 \rightarrow z = \frac{130 - 120}{4.4} \approx 2.27 \rightarrow P(z > 2.27) \approx 0.0115$$

$$2) P(78 \leq X \leq 82) \rightarrow z_1 = \frac{78 - 80}{3.4} = -0.59, z_2 = \frac{82 - 80}{3.4} = 0.59 \rightarrow P(-0.59 \leq z \leq 0.59) \approx 0.4436$$

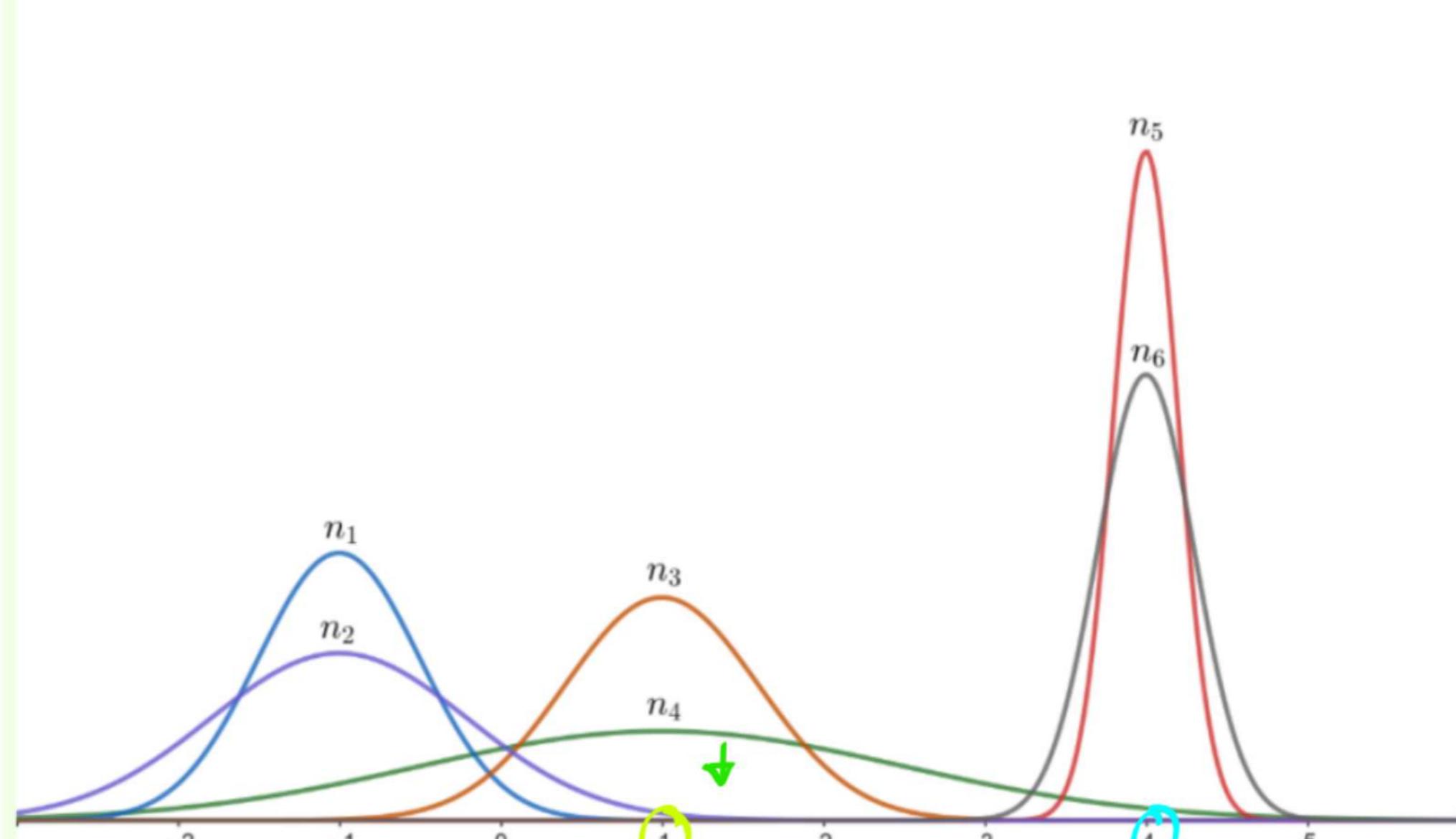
→ Für 90 % nimmst du z-Werte:
 $\mu \pm 1.645 \cdot \sigma$

Systolisch:
 $120 \pm 1.645 \cdot 4.4 \approx [112.76, 127.24] \quad \checkmark$

Diastolisch:
 $80 \pm 1.645 \cdot 3.4 \approx [74.41, 85.59] \quad \checkmark$

Dichte der Normalverteilung

Gegeben sind die Grafen von sechs verschiedenen Normalverteilungen nummeriert von n_1 bis n_6 :



oben gibt
x-Achse an
(Vorzeichen!)

Ordnen Sie die folgenden Dichtefunktionen diesen Normalverteilungen zu:

Normalverteilung Nummer 1

$$\frac{1}{0.5 \sqrt{2\pi}} e^{-0.5(\frac{x+1}{0.5})^2}$$

Normalverteilung Nummer 2

$$\frac{1}{0.8 \sqrt{2\pi}} e^{-0.5(\frac{x+1}{0.8})^2}$$

Normalverteilung Nummer 3

$$\frac{1}{0.6 \sqrt{2\pi}} e^{-0.5(\frac{x+1}{0.6})^2}$$

Normalverteilung Nummer 4

$$\frac{1}{1.5 \sqrt{2\pi}} e^{-0.5(\frac{x+1}{1.5})^2}$$

Normalverteilung Nummer 5

$$\frac{1}{0.2 \sqrt{2\pi}} e^{-0.5(\frac{x+1}{0.2})^2}$$

Normalverteilung Nummer 6

$$\frac{1}{0.3 \sqrt{2\pi}} e^{-0.5(\frac{x+1}{0.3})^2}$$

je grösser
hier
→ flatter
Kurve

Quiz SERIE 08

Streuung im Zentralen Grenzwertsatz

Bestmögliche Lösung:

Wenn die Varianz von \bar{X} gleich 25 ist und die Stichprobengröße $n = 6$, beträgt die Standardabweichung der ursprünglichen Verteilung (Populationsstandardabweichung):

- 4.17
- 5
- 25
- 2.04
- Kann nicht bestimmt werden.
- 12.25
- 150

Rechnungen

Bestmögliche Lösung:

Bei der Überprüfung der von einem Unternehmen ausgestellten Rechnungen stellt ein Prüfer fest, dass die Rechnungsbeträge einen Mittelwert von CHF 1'732 und eine Standardabweichung von CHF 298 aufweisen. Wie hoch ist die Wahrscheinlichkeit, dass der durchschnittliche Rechnungsbetrag für eine Stichprobe von 45 Rechnungen grösser als CHF 1'800 ist?

- 0.437
- Kann ohne den Populationsmittelwert nicht bestimmt werden.
- 0.937
- 0.063
- 0.563

STANDARDfehler

Bestmögliche Lösung:

Welche der folgenden Aussagen trifft auf den Standardfehler des Mittelwerts zu?

- Er ist kleiner als die Standardabweichung der Population.
- Er verringert sich, wenn die Stichprobengröße zunimmt.
- Er misst die Variabilität des Mittelwerts von Stichprobe zu Stichprobe.
- Alle oben genannten
- Keine der oben genannten

Zentraler Grenzwertsatz

Bestmögliche Lösung:

Ungefähr 10 % der Menschen sind Linkshänder. Wenn wir der Linkshändigkeit den Wert 1 und der Rechtshändigkeit den Wert 0 zuweisen, dann folgt die Wahrscheinlichkeitsverteilung der Linkshändigkeit für die gesamte Bevölkerung einer Bernoulli-Verteilung. Die Standardabweichung der Population beträgt 0.3. Sie machen eine Studie über Linkshänder in einem bestimmten Land und wählen zufällig 100 Personen aus. Kreuzen Sie die richtigen Antworten an.

- Der Zentrale Grenzwertsatz kann nicht angewendet werden, da die Anzahl Linkshänder nicht normalverteilt sind.
- Die Wahrscheinlichkeit dass nur 9.5% der Personen linkshändig sind beträgt 0.4338
- Die Wahrscheinlichkeit dass nur 9.5% der Personen linkshändig sind beträgt 0.0478