



# **Statistik**

Foliensammlung

Regina Tüchler & Achim Zeileis

# Grundbegriffe

## Kapitel 1

## Begriff Statistik:

- Erhebung von Daten  
z. B. statistische Ämter
- Auswertung von Daten  
z. B. Maßzahlen aus vorhandenen Daten  
⇒ Methodenlehre der empirischen Wissenschaften

**Deskriptive Datenanalyse:** Beschreibung von vorhandenen Daten  
z. B. Maßzahlen, Tabellen, Graphiken

**Induktive Datenanalyse:** Gesetzmäßigkeiten und Ursachen, die  
hinter den Daten stehen werden untersucht:

- **Explorativ:** Ziel ist, Hypothesen für die Theoriebildung zu gewinnen
- **Konfirmativ:** Ziel ist es vorhandene Theorien zu präzisieren oder zu falsifizieren  
z. B. Parameterschätzung, Testen von Hypothesen

# Terminologie, Variablentypen, Datenmatrix

---

Ausschnitt aus dem Datensatz **Weltraum** dient zur Illustration der folgenden Begriffe:

Terminologie: Untersuchungseinheit, Variable, Ausprägung

Variablentypen:

quantitativ: diskret - kontinuierlich;

qualitativ: dichotom - polytom - ordinal

Aufbau einer Datenmatrix

# Terminologie, Variablentypen, Datenmatrix

---

Beobachtung	Geschlecht	Groesse	Sternzeichen	Gedaechtnis
1	Maennlich	165.11	4	16
2	Maennlich	154.53	10	17
3	Maennlich	167.99	4	20
4	Maennlich	163.89	9	14
5	Maennlich	164.5	1	11
6	Maennlich	174.16	9	17
7	Maennlich	166.07	3	15
8	Maennlich	161.48	3	16
9	Maennlich	163.63	5	14
10	Maennlich	165.02	5	18
51	Weiblich	162.09	3	17
52	Weiblich	152.25	2	17
53	Weiblich	162.61	8	17
54	Weiblich	151.95	7	16
55	Weiblich	167.55	7	14
56	Weiblich	156.53	5	15
57	Weiblich	169.65	11	18
58	Weiblich	162.3	8	13
59	Weiblich	154.68	8	16
60	Weiblich	160.12	6	14

# Terminologie, Variablentypen, Datenmatrix

---

Beobachtung	Geschlecht	Groesse	Sternzeichen	Gedaechtnis
1	Maennlich	165.11	4	16
2	Maennlich	154.53	10	17
3	Maennlich	167.99	4	20
4	Maennlich	163.89	9	14
5	Maennlich	164.5	1	11
51	Weiblich	162.09	3	17
52	Weiblich	152.25	2	17
53	Weiblich	162.61	8	17
54	Weiblich	151.95	7	16
55	Weiblich	167.55	7	14

# Häufigkeiten von Aussagen

---

$A$  ist eine Aussage über ein Merkmal der Untersuchungsobjekte.

$h(A)$  bezeichnet die **absolute Häufigkeit** = Anzahl der Untersuchungsobjekte, für die  $A$  zutrifft.

$f(A) = h(A)/n$  bezeichnet die **relative Häufigkeit** = Anteil (Prozentsatz) der Untersuchungsobjekte, für die  $A$  zutrifft.



# Rechnen mit Aussagen

---

## Verknüpfung von Aussagen

$A \cup B$        $A$  oder  $B$   
**mindestens**

$A \cap B$        $A$  und  $B$   
**gleichzeitig**

$A'$       nicht  $A$

**Gegenteil**

$A \subseteq B$       wenn  $A$ , dann  $B$

# Rechnen mit Aussagen

---

**Beispiel:** Skript

**(1.11)** Es seien  $A, B, C$  beliebige Aussagen. Beschreibe:

(a) Alle Aussagen treffen zu.

$$A \cap B \cap C$$

(b)  $A$  und  $B$  treffen ein,  $C$  nicht.

$$A \cap B \cap C'$$

(c) Genau eine der drei Aussagen trifft zu.

$$(A \cap B' \cap C') \cup (A' \cap B \cap C') \cup (A' \cap B' \cap C)$$

# Spezielle Aussagen

---

$$A = \emptyset$$

$A$  ist **unmöglich**. Die Aussage  $A$  trifft niemals zu.  $f(A) = 0$ .

$$A = \Omega$$

$A$  ist **sicher**. Die Aussage  $A$  trifft immer zu.  $f(A) = 1$ .

# Rechengesetze für Häufigkeiten

---

## Additionsgesetz:

$$A \cap B = \emptyset \Rightarrow \begin{cases} h(A \cup B) &= h(A) + h(B), \\ f(A \cup B) &= f(A) + f(B). \end{cases}$$

## Siebformel:

$$\begin{aligned} h(A \cup B) &= h(A) + h(B) - h(A \cap B), \\ f(A \cup B) &= f(A) + f(B) - f(A \cap B). \end{aligned}$$

# Rechengesetze für Häufigkeiten

---

**Beispiel:** Skript

**(1.23), 3.** Reaktion auf Videofilm

	Sehr positiv	Positiv	Negativ	Sehr negativ
Männer	1	3	5	10
Frauen	6	8	3	1
Jungen	5	5	3	2
Mädchen	8	5	1	1

# Rechengesetze für Häufigkeiten

---

**Beispiel:** Skript

**(1.23), 3.** Reaktion auf Videofilm

	Sehr positiv	Positiv	Negativ	Sehr negativ	Summe
Männer	1	3	5	10	19
Frauen	6	8	3	1	18
Jungen	5	5	3	2	15
Mädchen	8	5	1	1	15
Summe	20	21	12	14	67

# Rechengesetze für Häufigkeiten

---

## Beispiel:

Bankkunden:

28% der Kunden einer Bank haben einen Wohnungskredit;

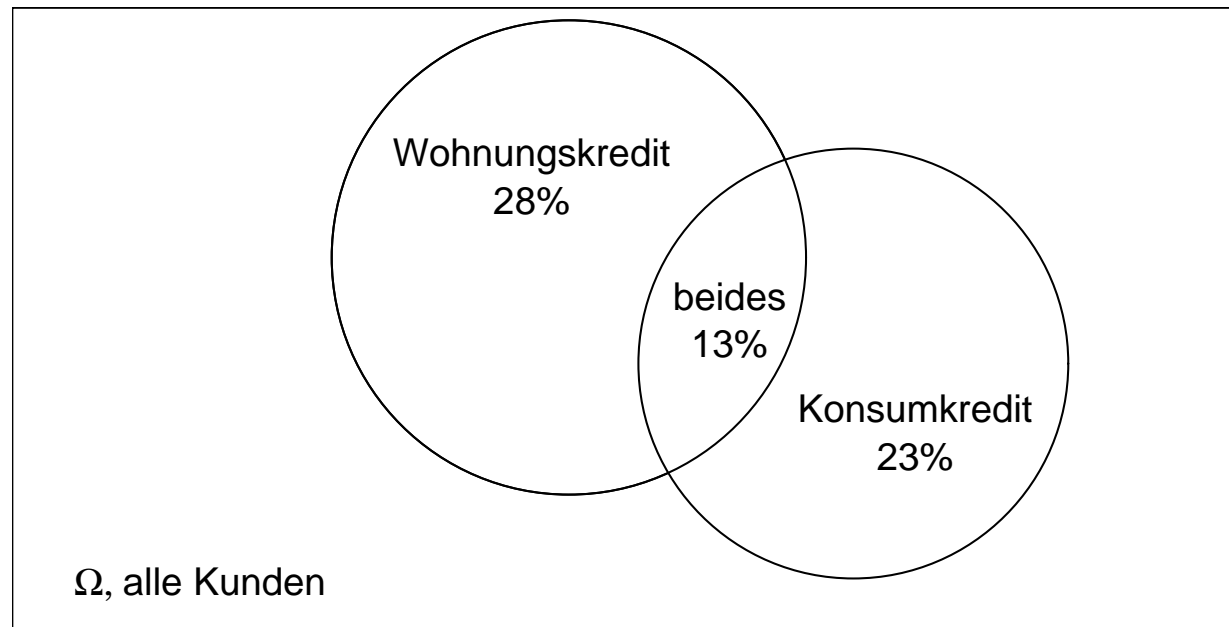
23% einen Kredit zur Finanzierung von Konsumgütern;

13% beides.

Wie groß ist der Anteil der Kunden, die weder Wohnungs- noch Konsumgüterkredit haben?

# Rechengesetze für Häufigkeiten

---





# Rechengesetze für Häufigkeiten

---



# Zusammenfassung Kapitel 1

---

- Terminologie, Variablentypen, Aufbau einer Datenmatrix
- Absolute und relative Häufigkeiten
- Aussagen und Rechenoperationen
- Gesetze für Häufigkeiten

# Beschreibung von Datenlisten

## Kapitel 2

# Skalentypen

---

- **Nominalskala:** Zahlenwerte sind nur Bezeichnung für sich ausschließende Kategorien. **Bsp. Familienstand:** ledig, verheiratet, geschieden, verwitwet.
- **Ordinalskala:** Ordnung der Zahlen ist sinnvoll interpretierbar. **Bsp. Leistungsbeurteilung:** Noten 1, 2, 3, 4, 5.
- **Intervallskala:** Ordnung und Abstände zwischen den Werten sinnvoll interpretierbar. **Bsp. Jahr:** 2003, 1997, ...
- **Ratioskala (Verhältnisskala):** Ordnung, Abstände und Größenverhältnisse. **Bsp. Alter:** Altersangabe in Jahren

# Qualitative Variablen

---

- Zerlegung: vollständig, alternativ
- Häufigkeitsverteilung:  $h(A_1), \dots, h(A_m)$  und  $f(A_1), \dots, f(A_m)$
- Häufigkeitstabelle
- Balkendiagramm, Sektorendiagramm

# Qualitative Variablen

---

**Beispiel:** “Bekenntnis” aus dem Datensatz Weltraum

- 1 Christliches Bekenntnis
- 2 Islamisches Bekenntnis
- 3 Buddhismus, Hindu oder Shintoismus
- 4 Sonstige
- 5 Ohne Bekenntnis

# Qualitative Variablen

---

Alle Personen

$A$	$h(A)$	$f(A)$
Christlich	33	0.33
Islam	27	0.27
Buddhismus	13	0.13
Sonstige	15	0.15
Ohne Bekenntnis	12	0.12
Summe	100	1

# Qualitative Variablen

---

Männer

$A$	$h(A)$	$f(A)$
Christlich	11	0.22
Islam	20	0.40
Buddhismus	8	0.16
Sonstige	6	0.13
Ohne Bekenntnis	5	0.10
Summe	50	1



# Qualitative Variablen

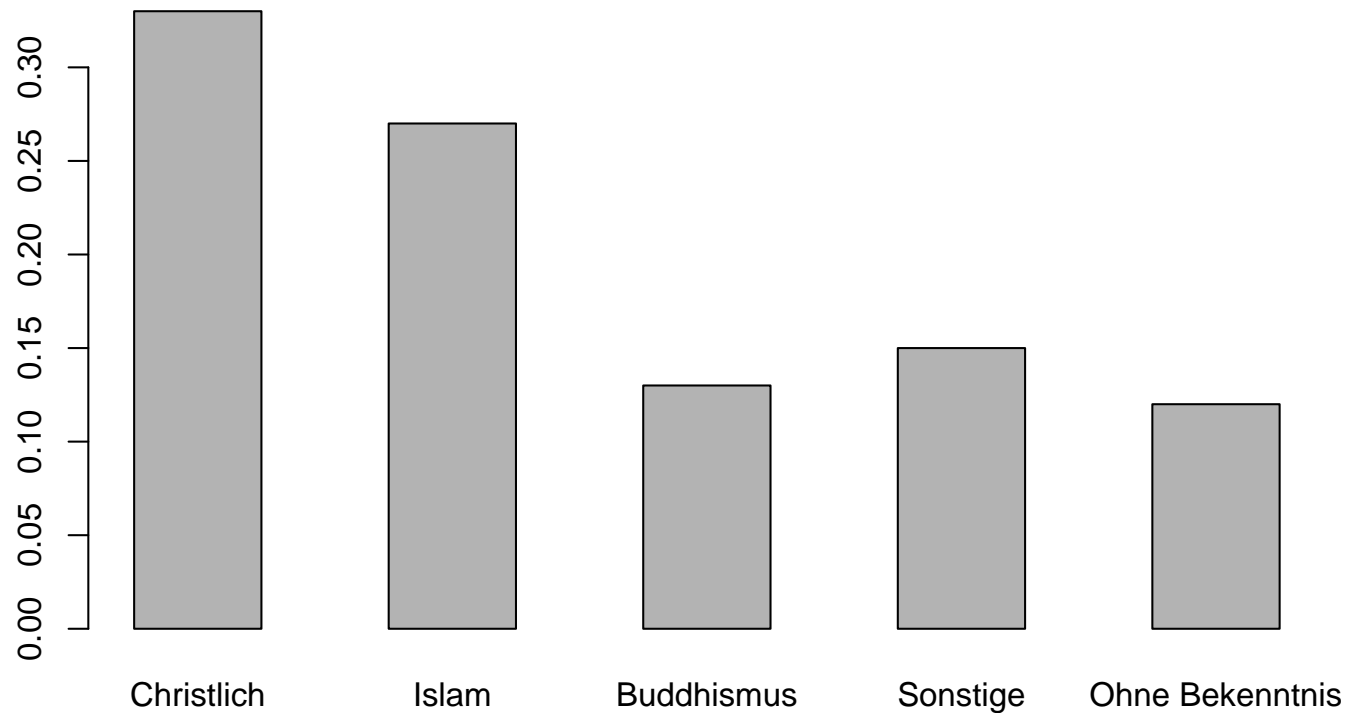
---

Frauen

$A$	$h(A)$	$f(A)$
Christlich	22	0.44
Islam	7	0.14
Buddhismus	5	0.10
Sonstige	9	0.18
Ohne Bekenntnis	7	0.14
Summe	50	1

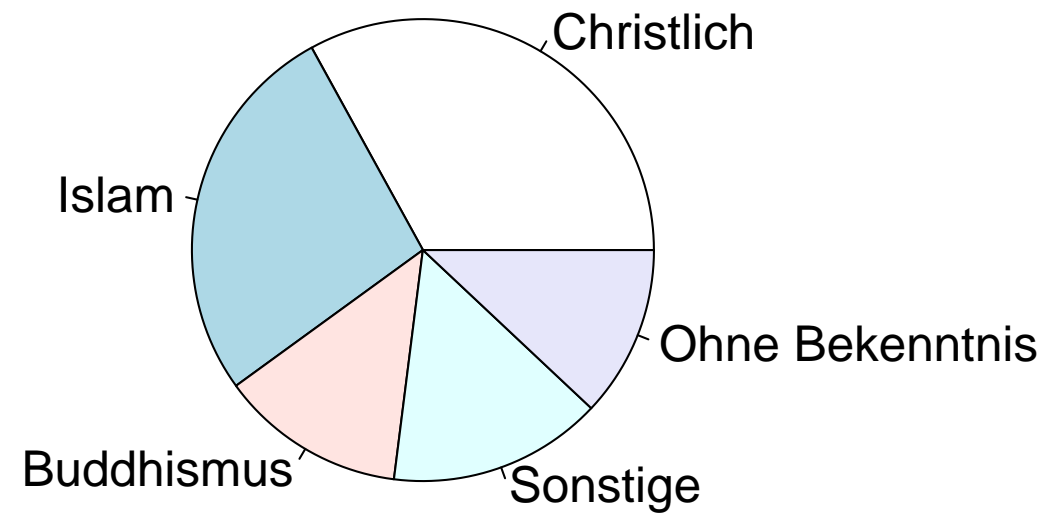
# Qualitative Variablen

---



# Qualitative Variablen

---



# Quantitative Variablen

---

Beobachtung	Geschlecht	Groesse	Sternzeichen	Gedaechnis
1	Maennlich	165.11	4	16
2	Maennlich	154.53	10	17
3	Maennlich	167.99	4	20
4	Maennlich	163.89	9	14
5	Maennlich	164.5	1	11
6	Maennlich	174.16	9	17
7	Maennlich	166.07	3	15
⋮	⋮	⋮	⋮	⋮

Quantitative Variable **Gedaechnis** mit möglichen Ausprägungen:  
0 bis 20 richtig beantwortete Fragen.

# Häufigkeitstabelle einer quantitativen Variablen

---

1. Ausprägungen der Größe nach ordnen  
⇒ Summenhäufigkeit sinnvoll (ab Ordinalskala)
2. absolute und relative Häufigkeit ( $h$ ,  $f$ ) und Summenhäufigkeit ( $H$ ,  $F$ ) berechnen und eintragen

## Häufigkeitstabelle einer quantitativen Variablen

---

$a$	$h$	$H$	$f$	$F$
10	1	1	0.01	0.01
11	1	2	0.01	0.02
12	3	5	0.03	0.05
13	6	11	0.06	0.11
14	12	23	0.12	0.23
15	15	38	0.15	0.38
16	19	57	0.19	0.57
17	17	74	0.17	0.74
18	11	85	0.11	0.85
19	8	93	0.08	0.93
20	7	100	0.07	1

# Empirische Verteilungsfunktion

---

Daten  $x_1, x_2, \dots, x_n$ :

$$F_n(x) = f_n(X \leq x) \quad x \in \mathbb{R}$$

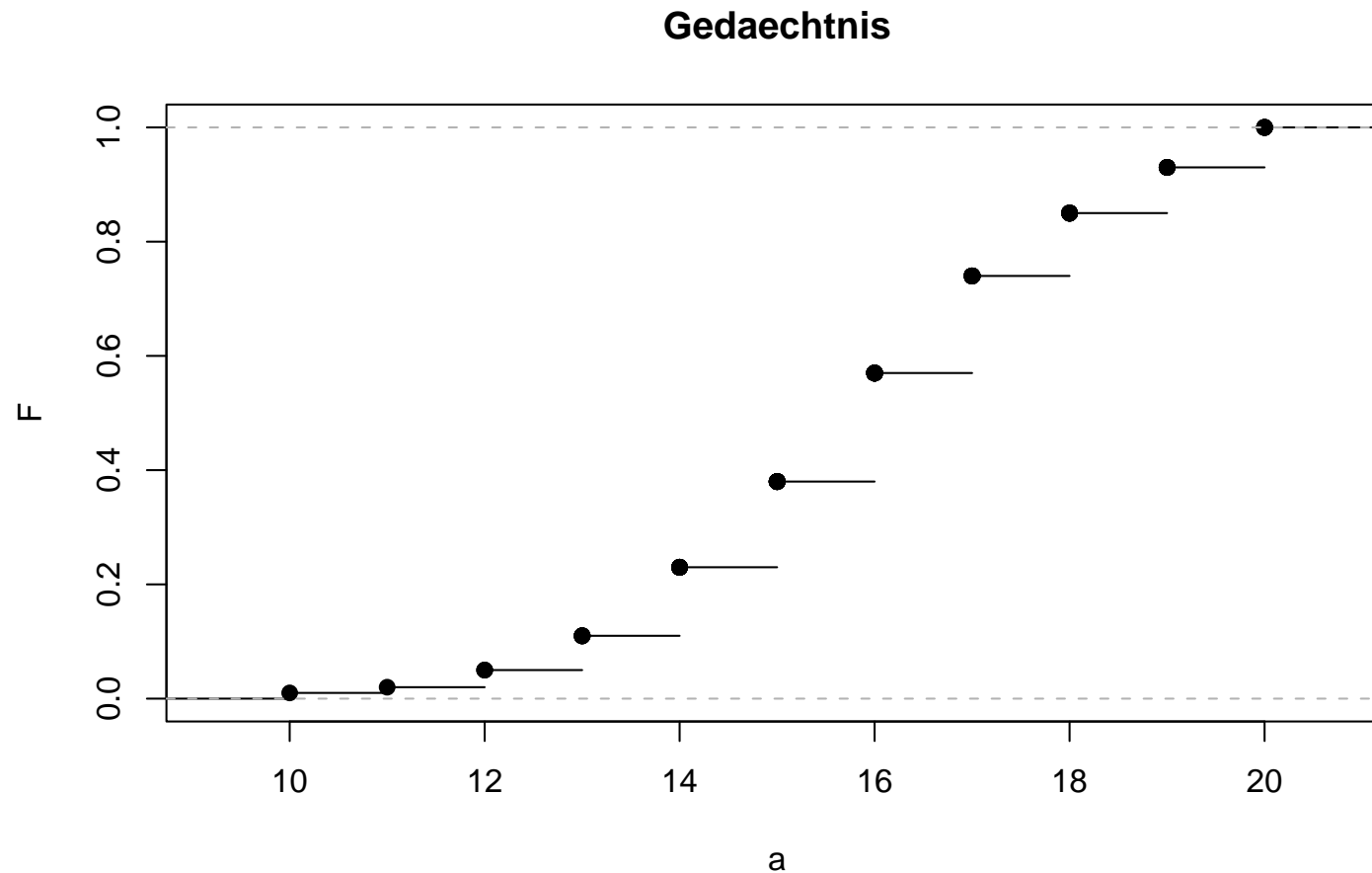
$$F_n(x) = F_i = f_1 + \dots + f_i \quad a_i \leq x < a_{i+1}$$

$F_n$  ist eine Sprungfunktion: Sprungstellen sind die Daten, Sprunghöhe sind die relativen Häufigkeiten.

Daher gibt der Niveauunterschied an den Endpunkten eines Intervalls die relative Häufigkeit der Daten in diesem Intervall an.

# Empirische Verteilungsfunktion

---





# Empirische Verteilungsfunktion

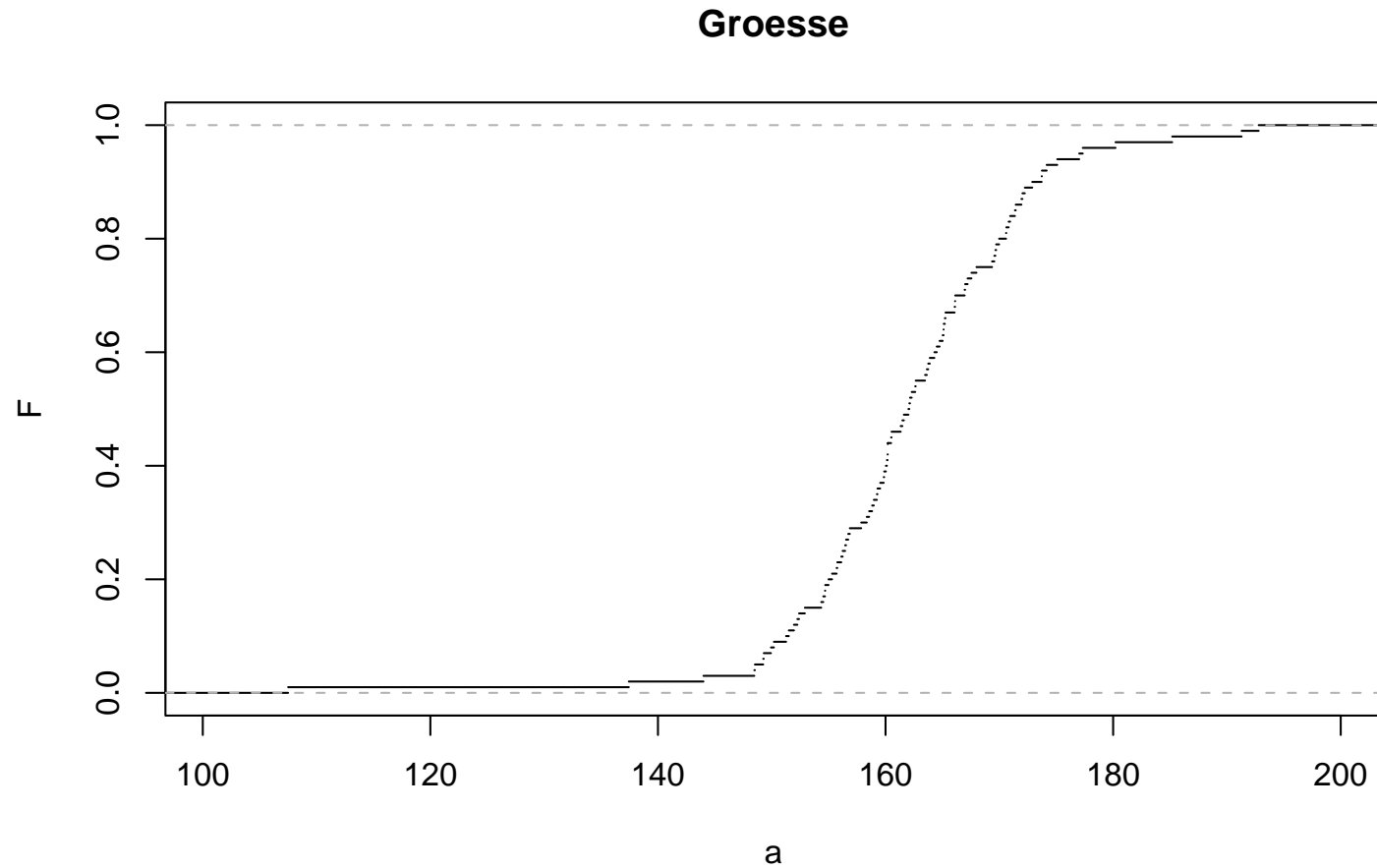
---

**Beispiel:** Größe aus Datensatz Weltraum

1. Welcher Anteil der Personen ist höchstens 1.50m gross?  
(Schluss von  $x$  auf Anteil)
2. Welche Größe wird von 20% der Personen überschritten?  
(Schluss von Anteil auf  $x$ )

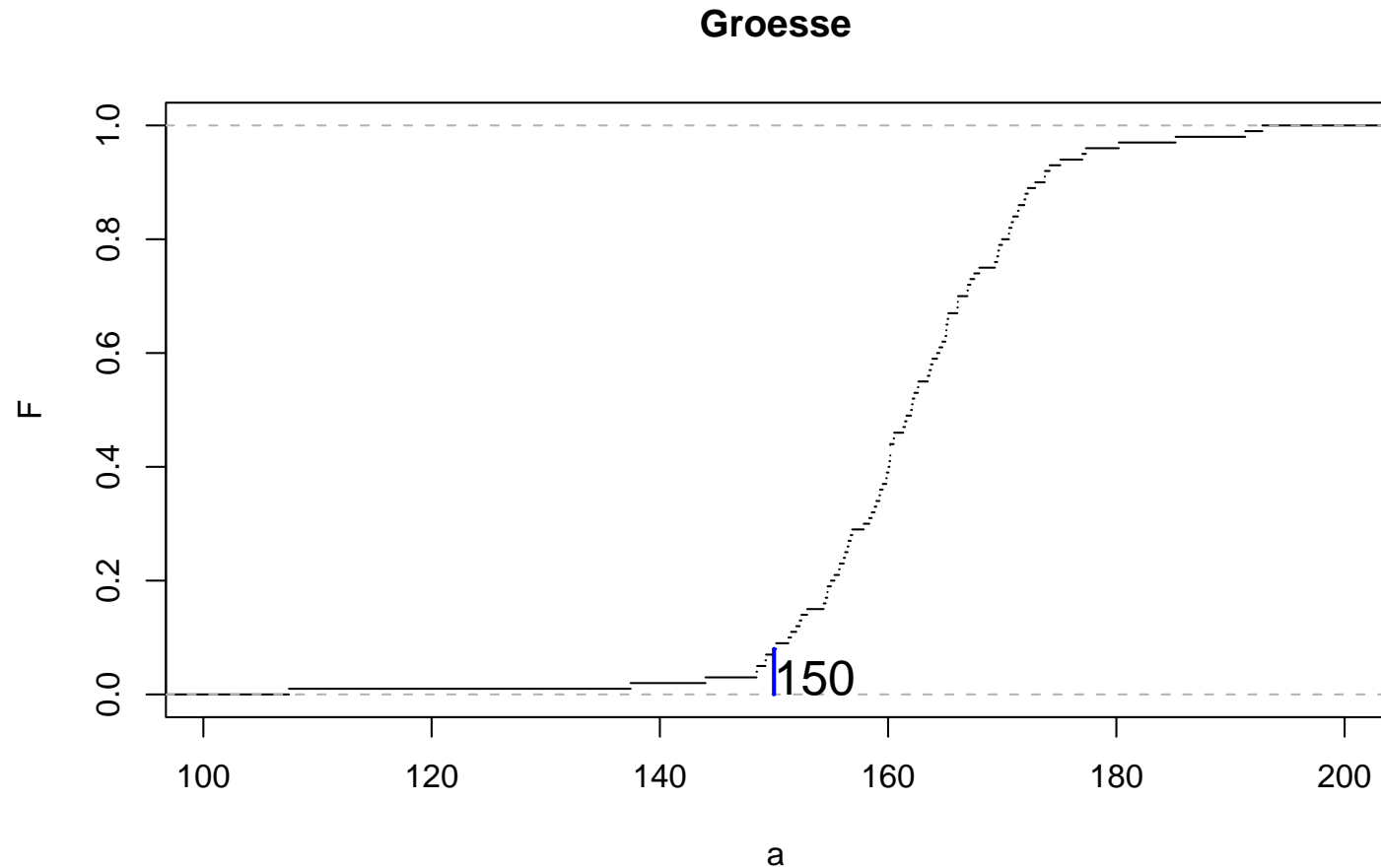
# Empirische Verteilungsfunktion

---



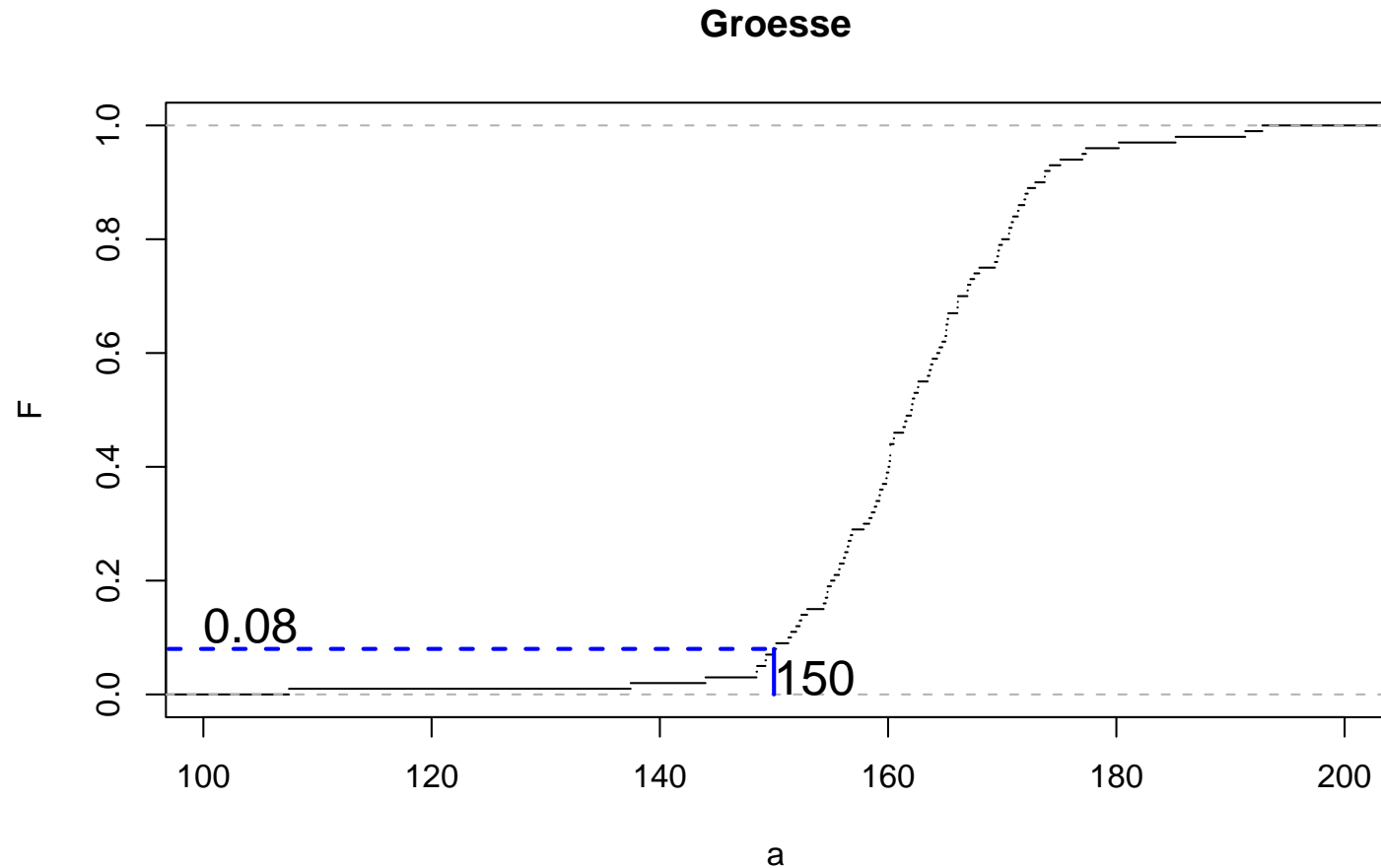
# Empirische Verteilungsfunktion

---



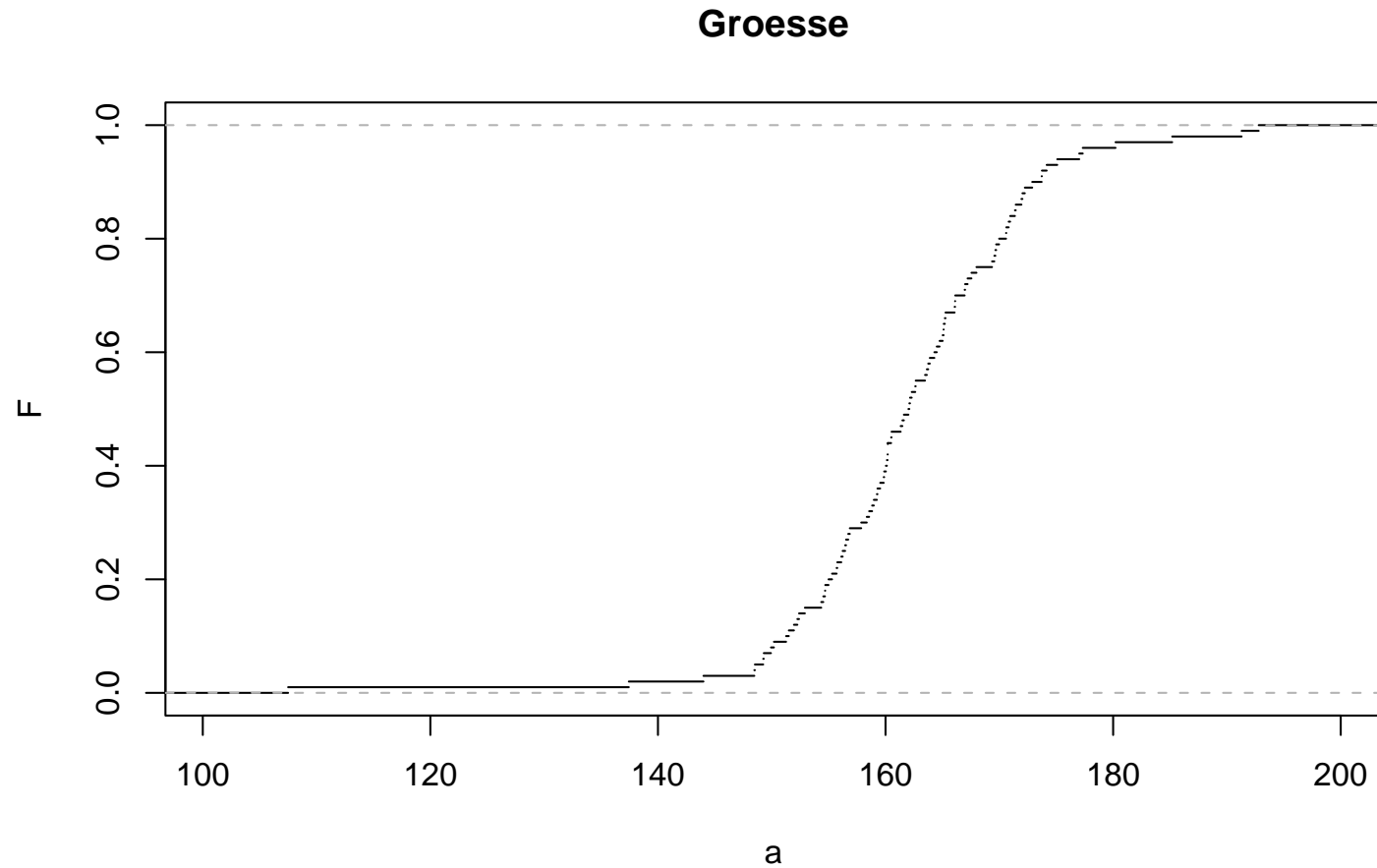
# Empirische Verteilungsfunktion

---



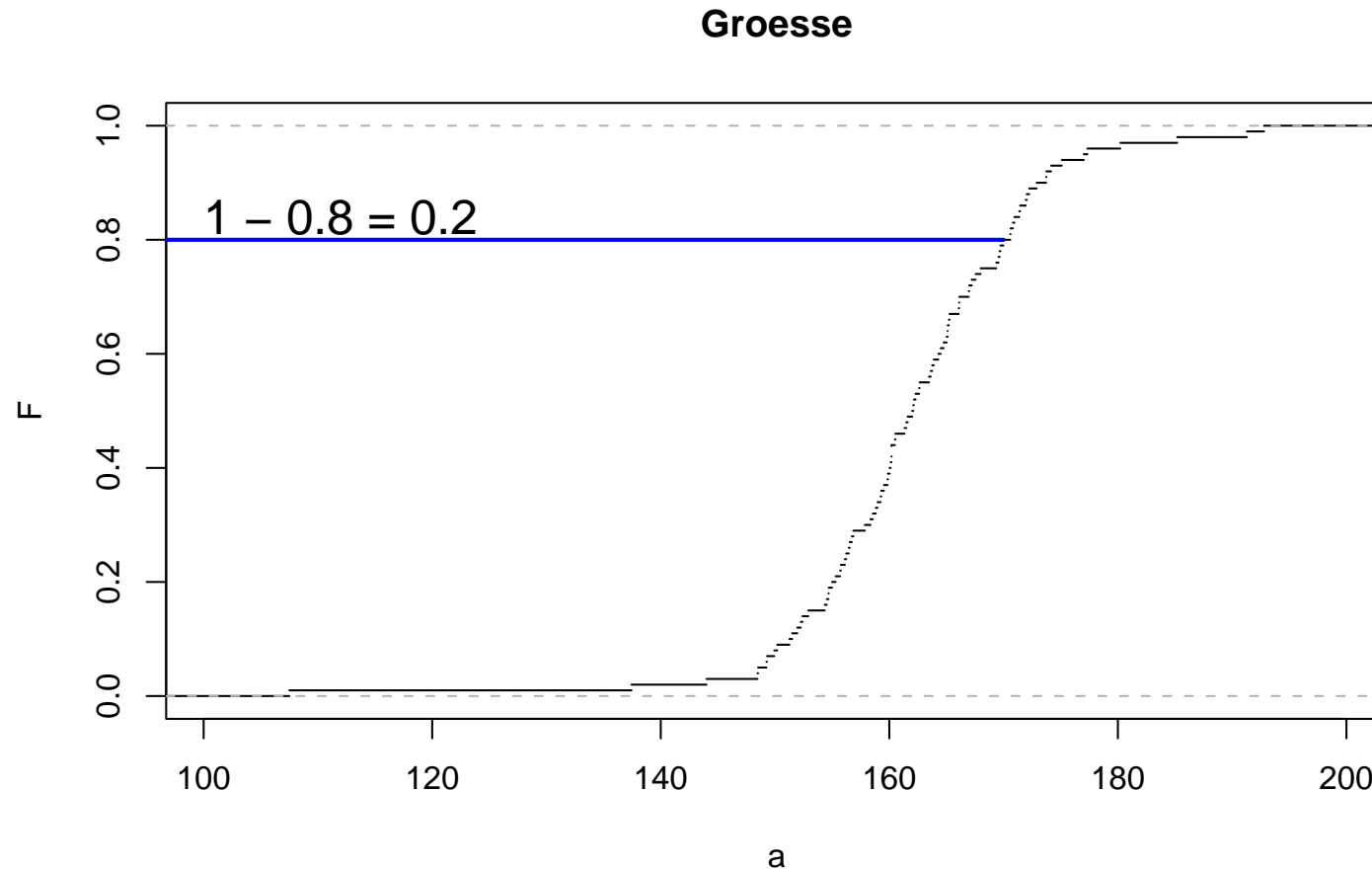
# Empirische Verteilungsfunktion

---



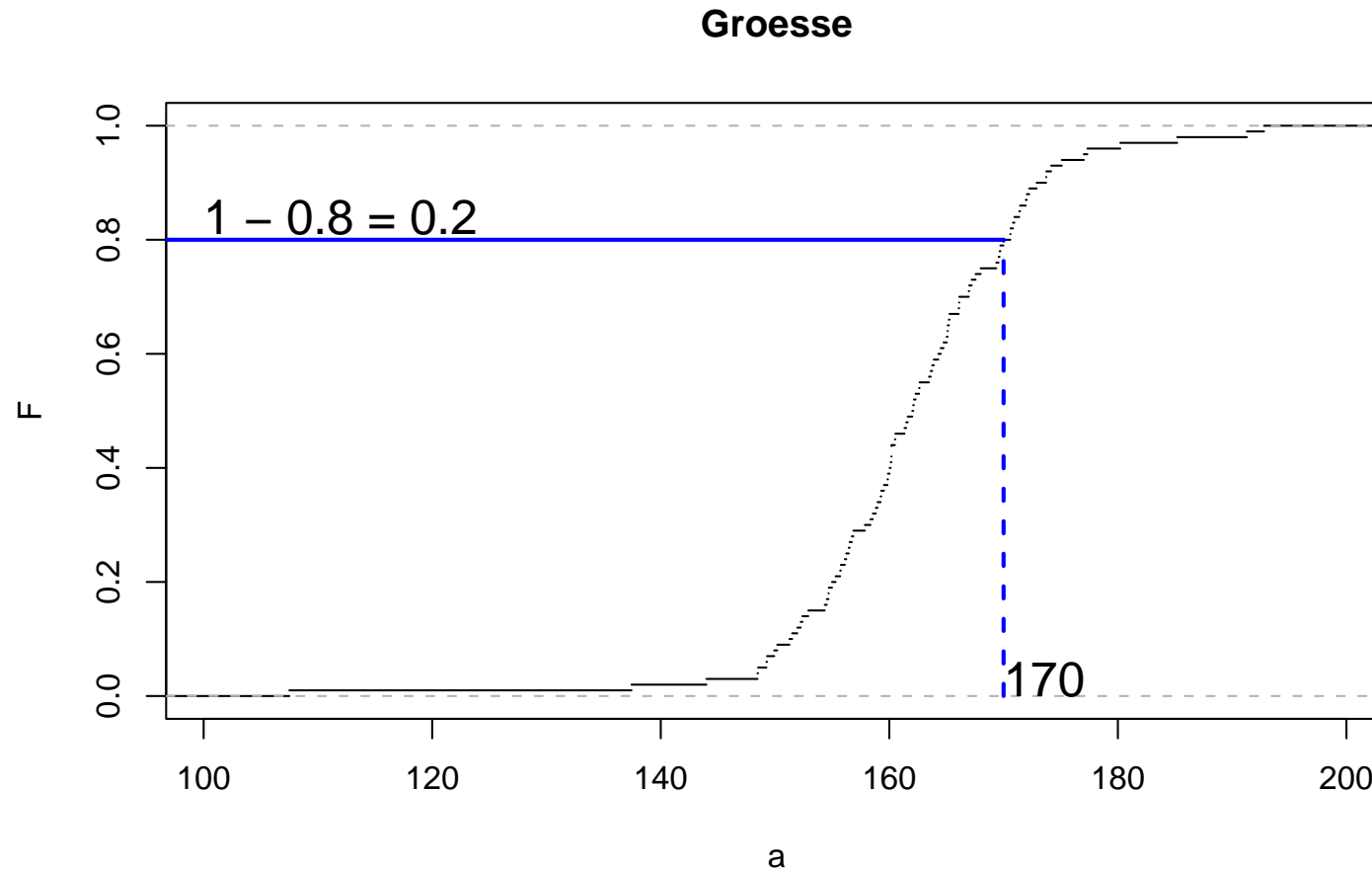
# Empirische Verteilungsfunktion

---



# Empirische Verteilungsfunktion

---



# Empirische Verteilungsfunktion

---

## Beispiel: Aufgabensammlung

**43.** Abbildung 41 zeigt die Verteilung der Einkommen (in 1000 GE) der in einem Betrieb Beschäftigten anhand der Perzentilsfunktion.

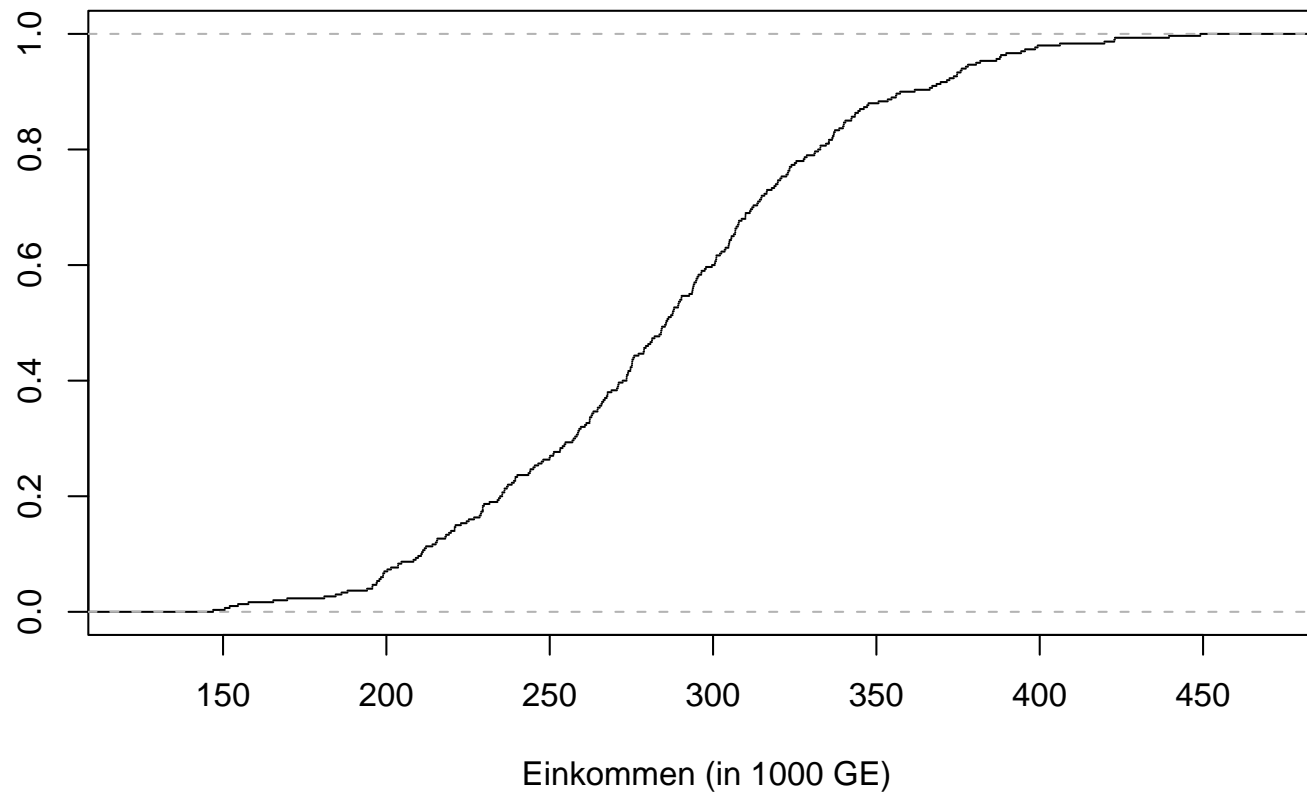
Wie groß ist ungefähr der Anteil der Beschäftigten, die mehr als 300000 GE verdienen?

(Schluss von  $x$  auf Anteil)



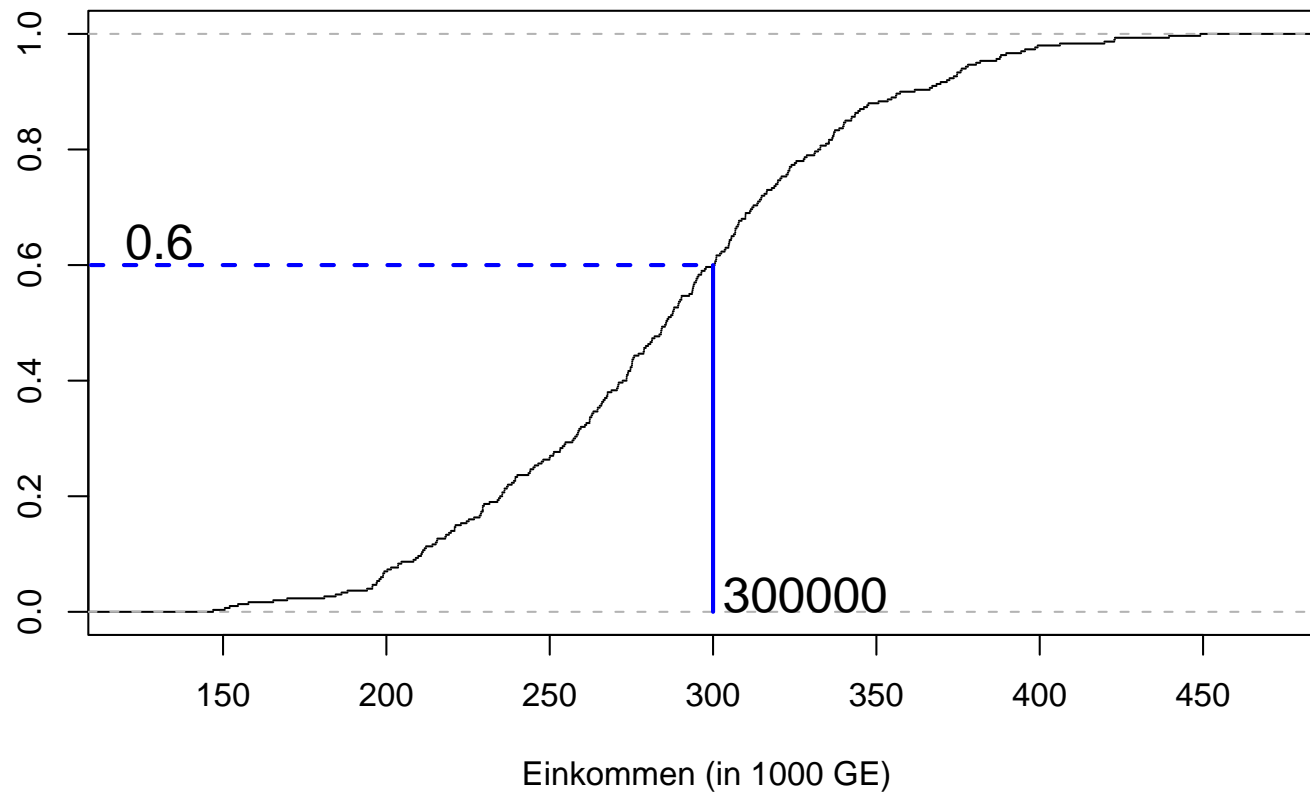
# Empirische Verteilungsfunktion

---



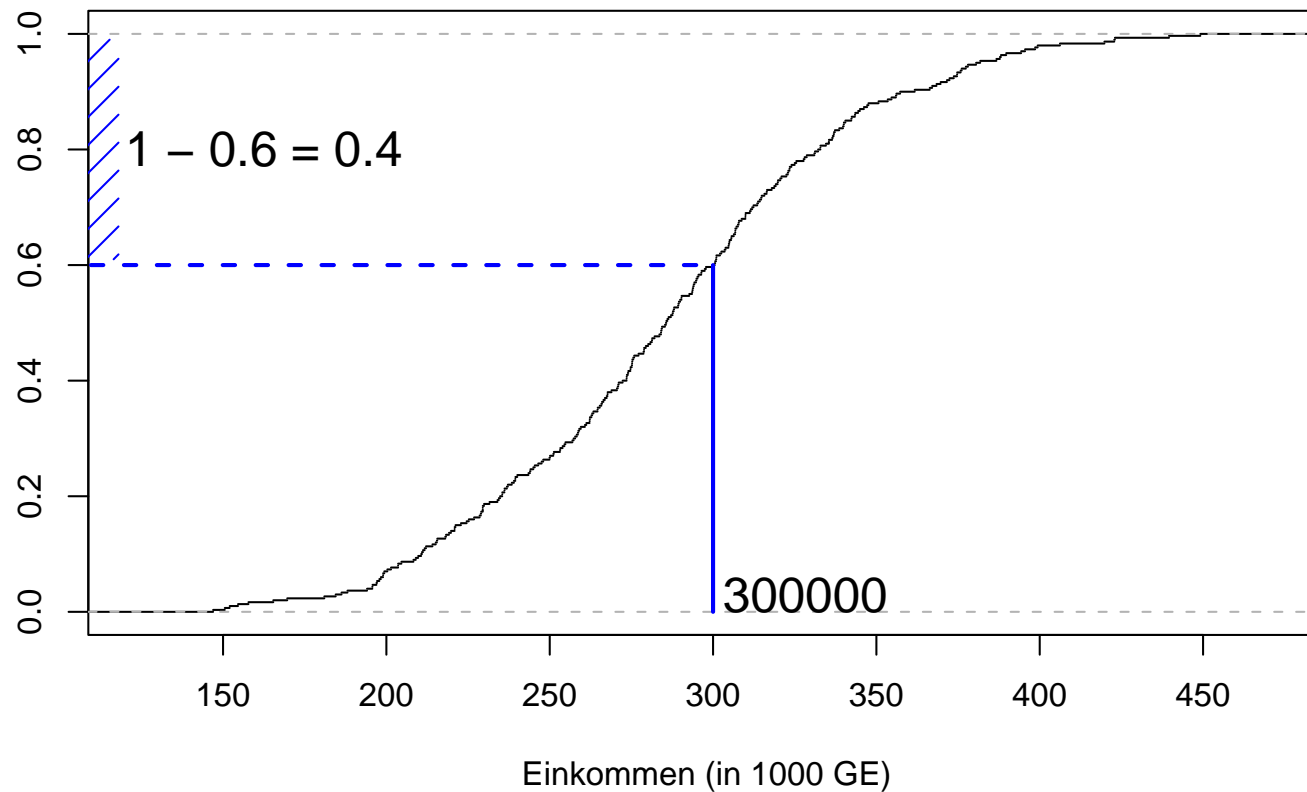
# Empirische Verteilungsfunktion

---



# Empirische Verteilungsfunktion

---



# Empirische Verteilungsfunktion

---

**Beispiel:** Aufgabensammlung

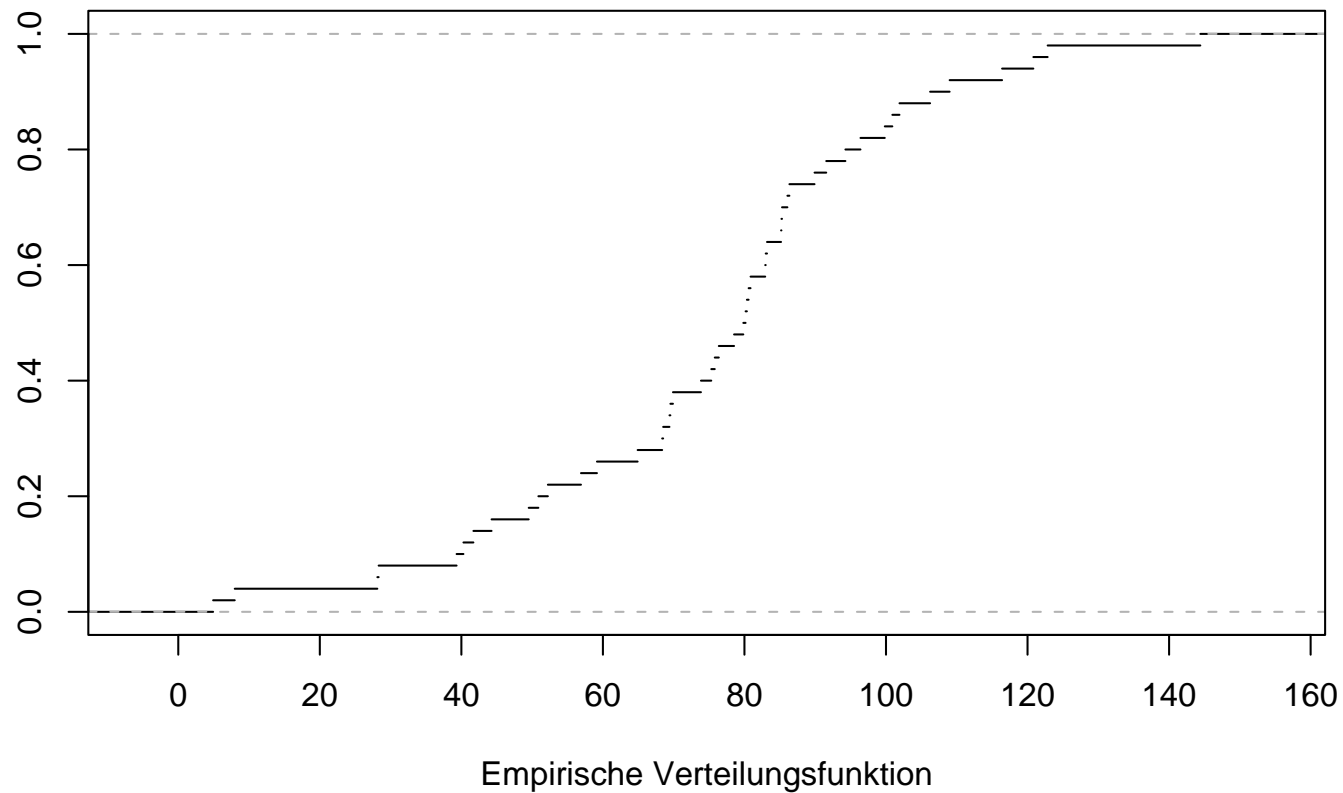
42. Abbildung 40 zeigt eine empirische Verteilungsfunktion.

Wie groß ist der Anteil der Beobachtungen, die zwischen 80 und 100 liegen?

(Schluss von  $x$  auf Anteil)

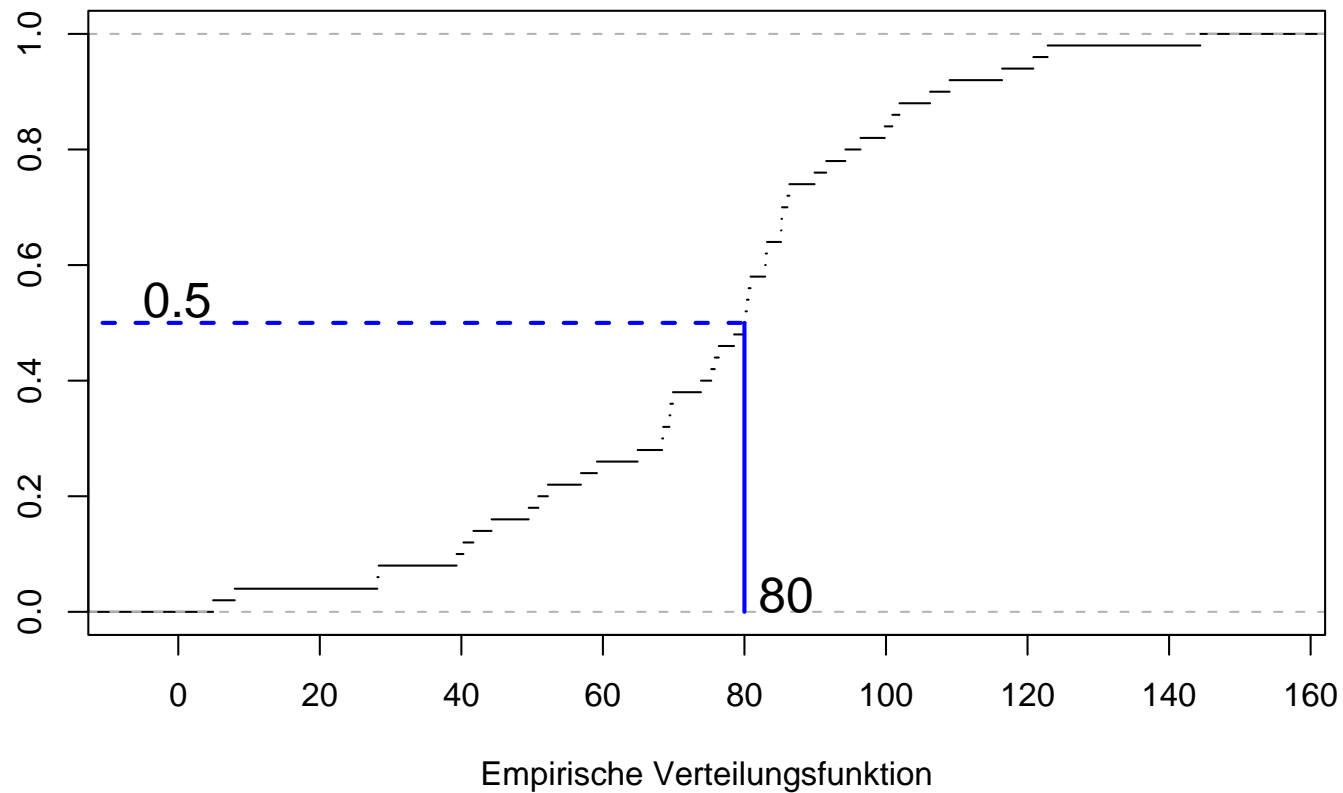
# Empirische Verteilungsfunktion

---



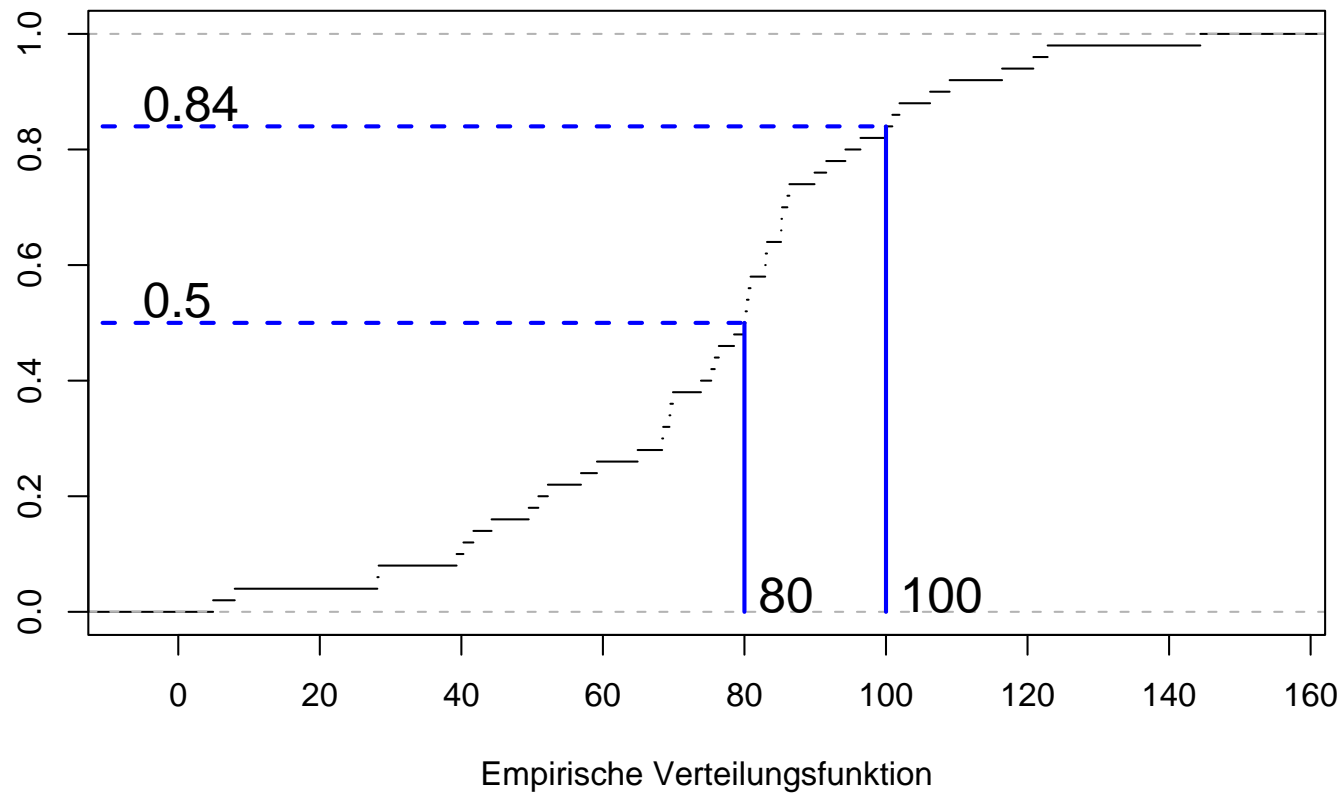
# Empirische Verteilungsfunktion

---

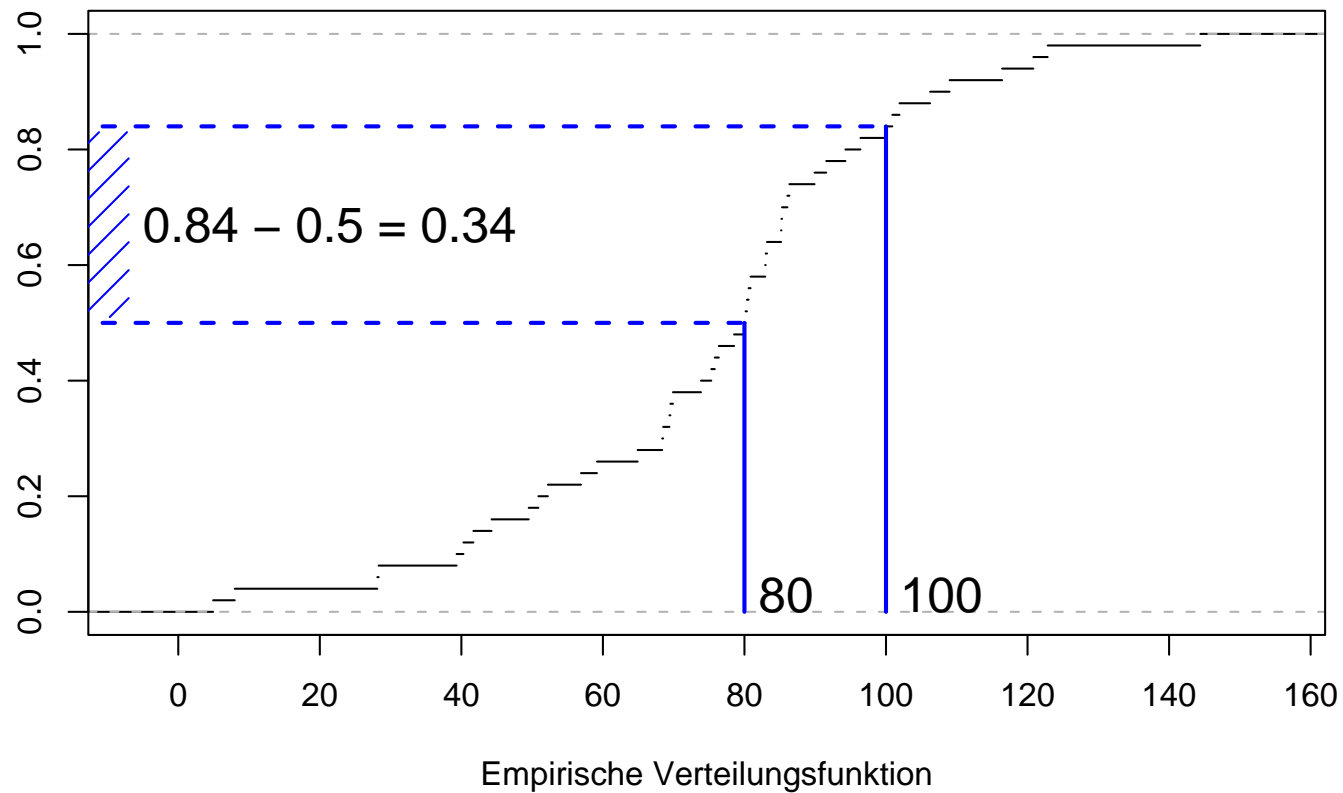


# Empirische Verteilungsfunktion

---



# Empirische Verteilungsfunktion





# Empirische Verteilungsfunktion

---

## Beispiel: Aufgabensammlung

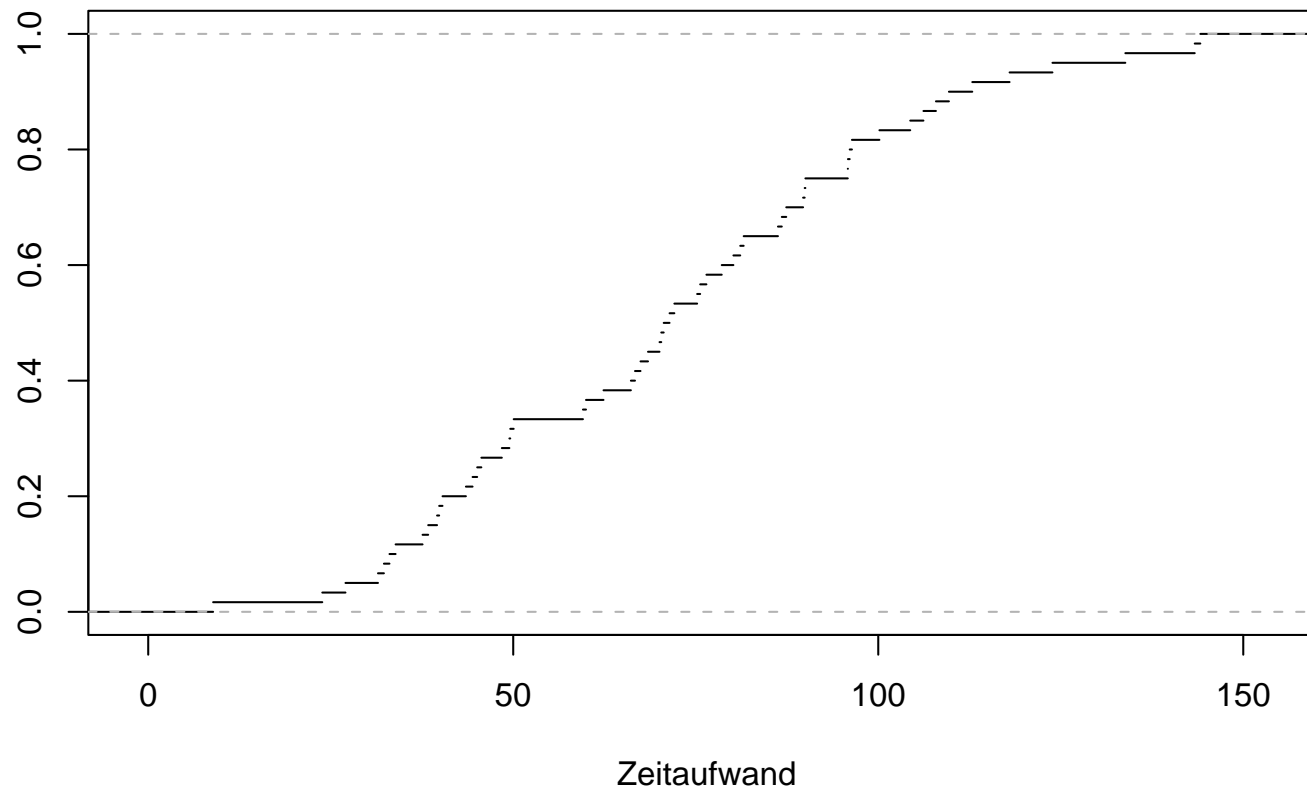
44. Im Zuge einer Umstellung des Arbeitsablaufes in einer Manufaktur für Hemden wurde 60 Mal erhoben, wieviel Zeit für das Annähen eines Kragens benötigt wird. Abbildung 42 zeigt die Ergebnisse anhand einer empirischen Verteilungsfunktion.

Ermitteln Sie den Zeitaufwand (in Sekunden), der von 90 Prozent der Versuche nicht überschritten wurde.

(Schluss von Anteil auf  $x$ )

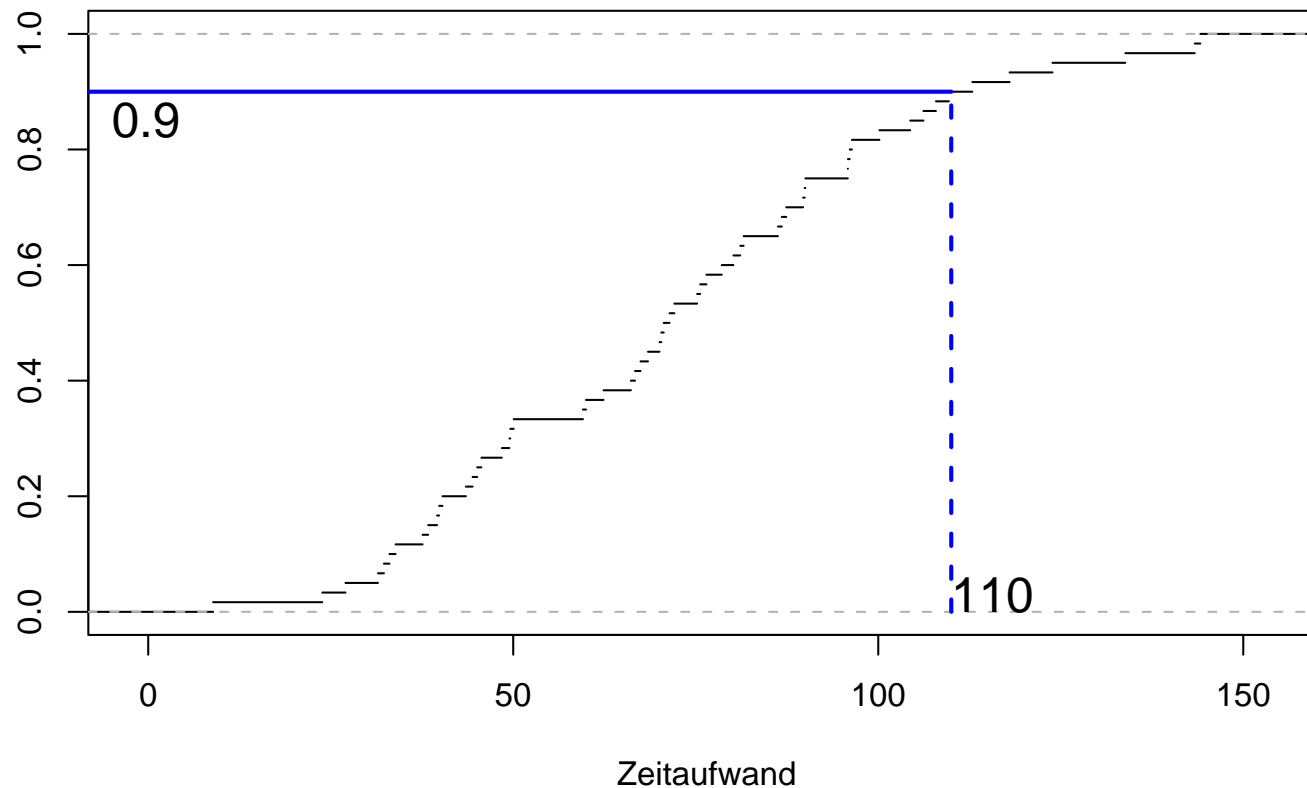
# Empirische Verteilungsfunktion

---



# Empirische Verteilungsfunktion

---



# Empirische Verteilungsfunktion

---

- Die Funktion  $F(x)$  gibt zu jedem Datenpunkt  $x$  die Summenhäufigkeit an,
- Anteile können direkt abgelesen werden,
- $x$ -Wertebereich-Achse gegen  $y$ -Anteilsachse.

# Histogramm

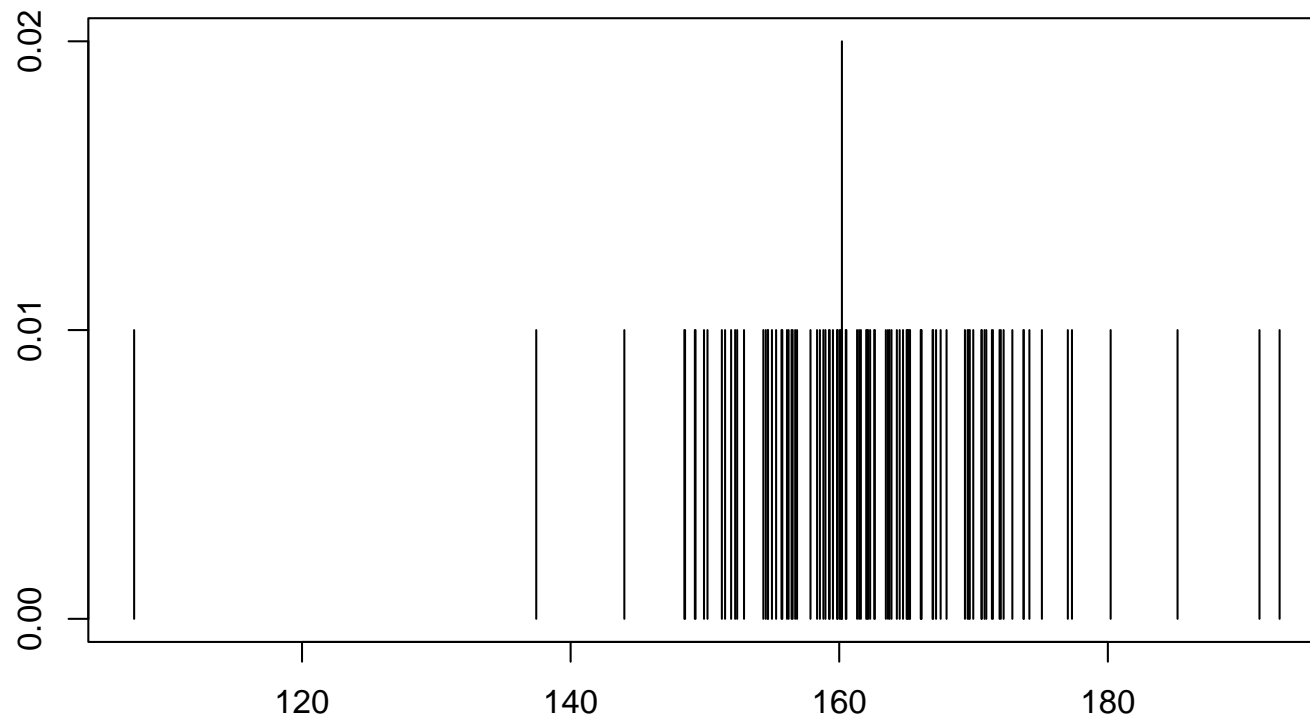
---

- Die Funktion  $f(x)$  gibt zu jedem möglichen Datenpunkt  $x$  die Dichte um  $x$  an,
- Anteile werden als Flächeninhalte dargestellt,
- Aufzeichnen von  $x$ -Wertebereich-Achse gegen  $f(x)$ -Dichteachse,
- Rechtecke so dass Fläche gleich Anteil,
- andere Kurven so dass Fläche gleich Anteil.

# Histogramm

---

Groesse



# Histogramm durch Gruppierung

---

## 1. Gruppierung in Intervalle

Intervalle müssen den Wertebereich abdecken: vollständig und alternativ.

## 2. Häufigkeitstabelle für gruppierte Daten

Skript, Musteraufgabe (2.12): **Größe** aus dem Datensatz **Weltraum**

# Histogramm durch Gruppierung

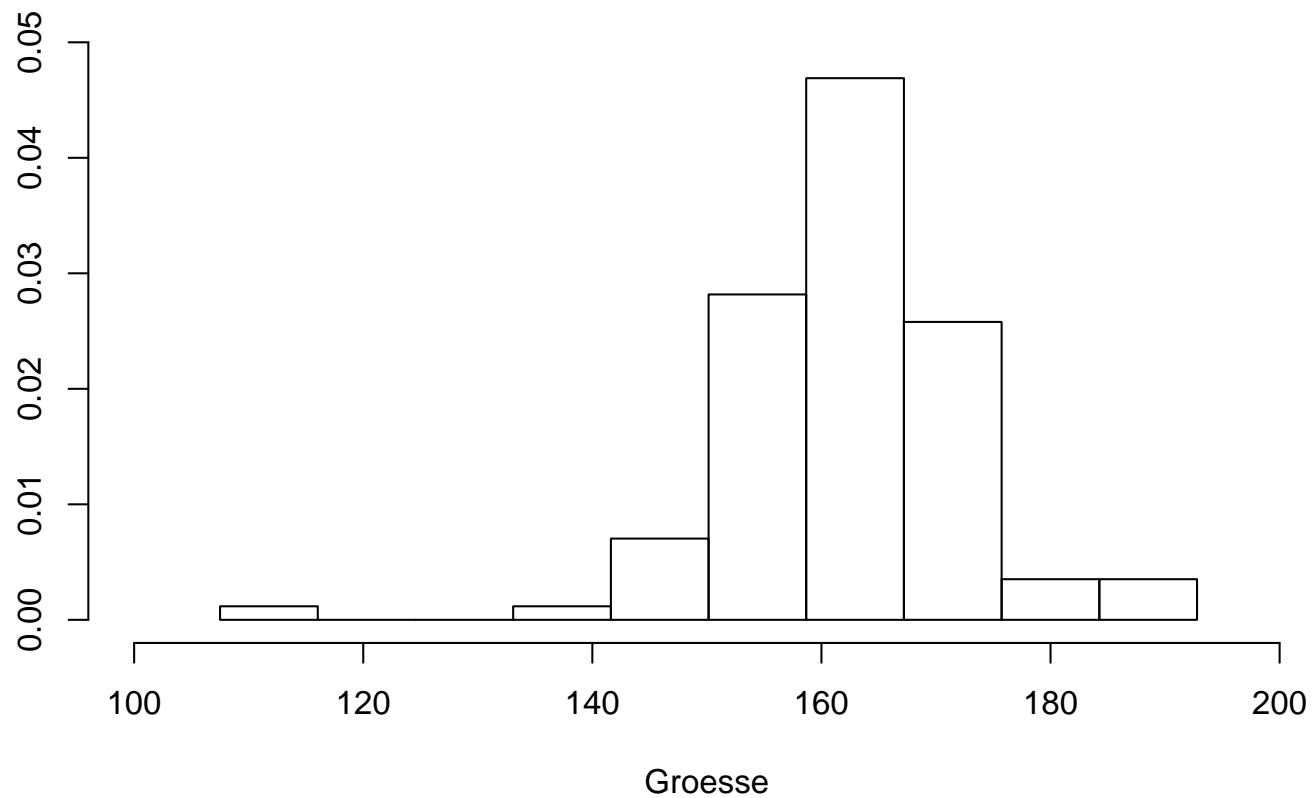
---

Intervalle	$h$	$H$	$f$	$F$
$107.50 \leq X \leq 116.03$	1	1	0.01	0.01
$116.03 < X \leq 124.56$	0	1	0.00	0.01
$124.56 < X \leq 133.09$	0	1	0.00	0.01
$133.09 < X \leq 141.62$	1	2	0.01	0.02
$141.62 < X \leq 150.15$	6	8	0.06	0.08
$150.15 < X \leq 158.67$	24	32	0.24	0.32
$158.67 < X \leq 167.20$	40	72	0.40	0.72
$167.20 < X \leq 175.73$	22	94	0.22	0.94
$175.73 < X \leq 184.26$	3	97	0.03	0.97
$184.26 < X \leq 192.79$	3	100	0.03	1.00



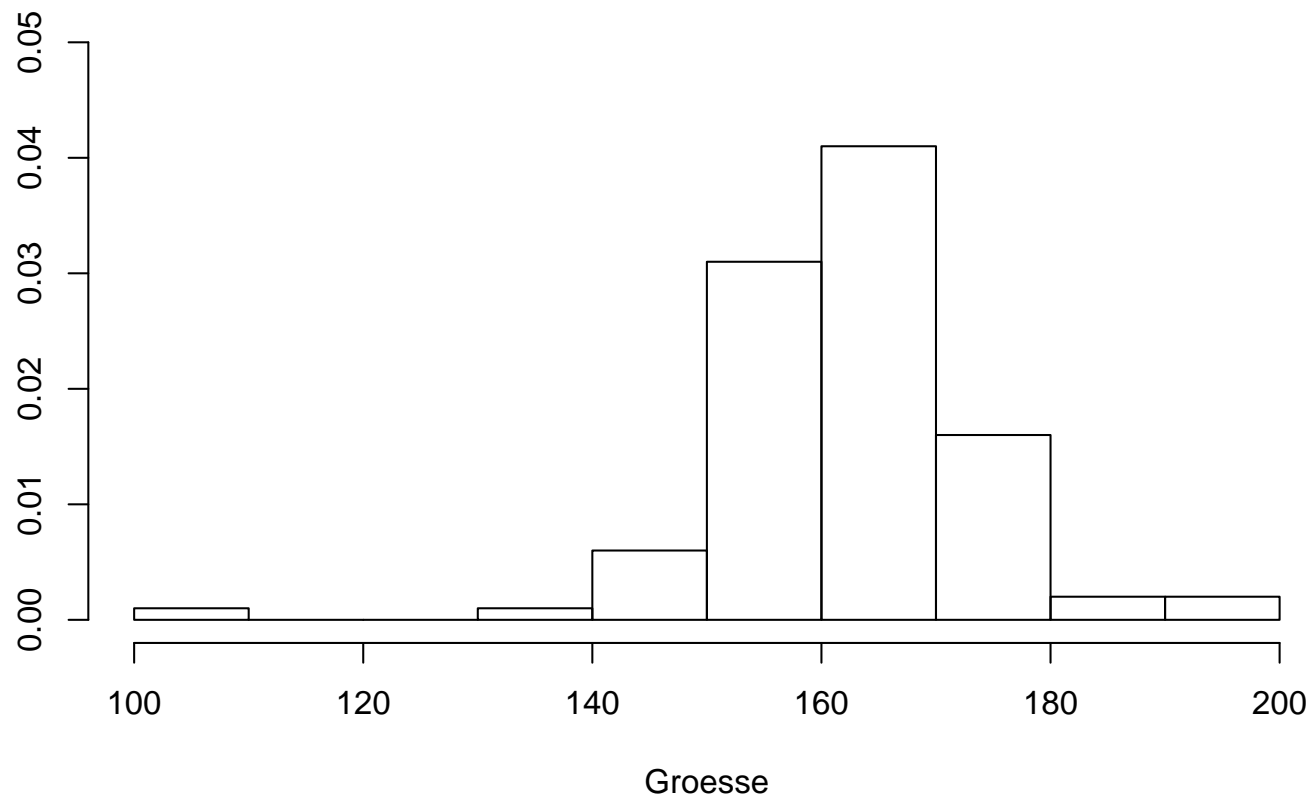
# Histogramm

---



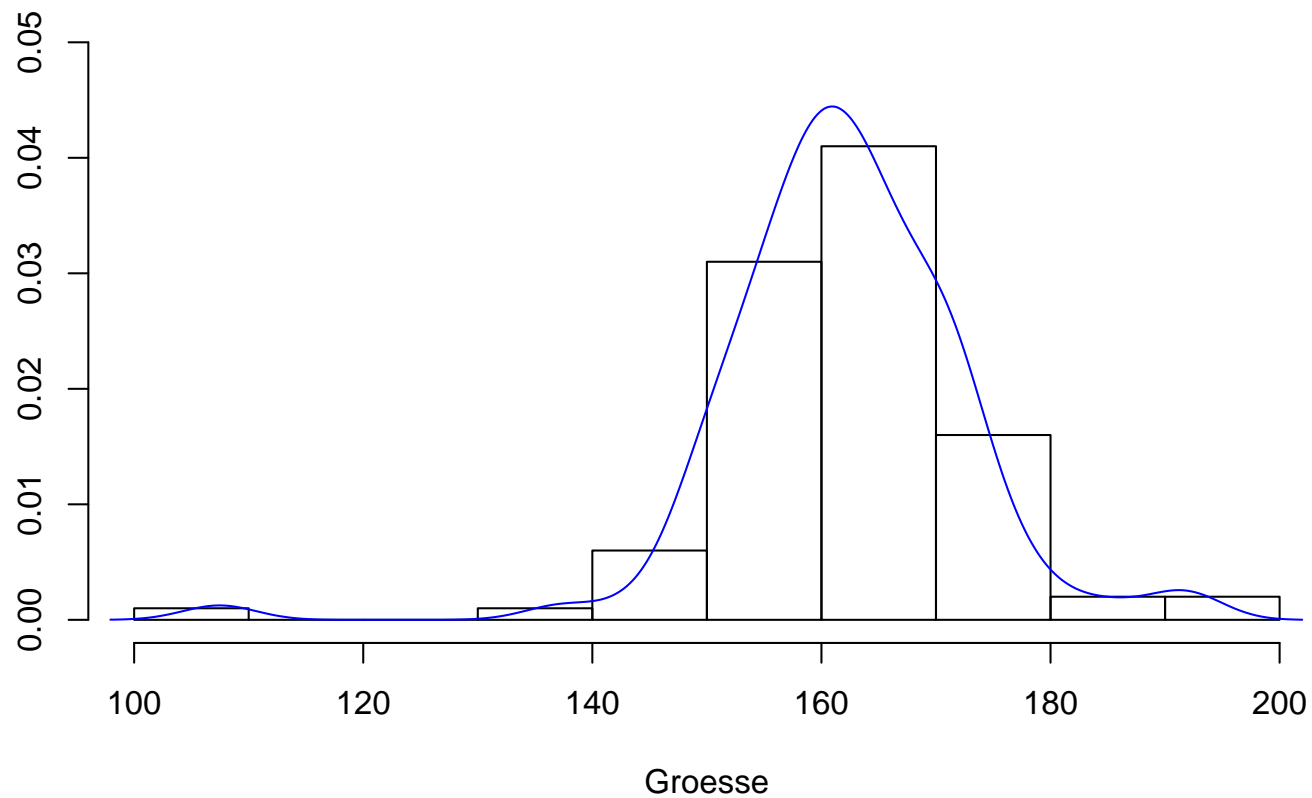
# Histogramm

---



# Histogramm

---



# Histogramm

---

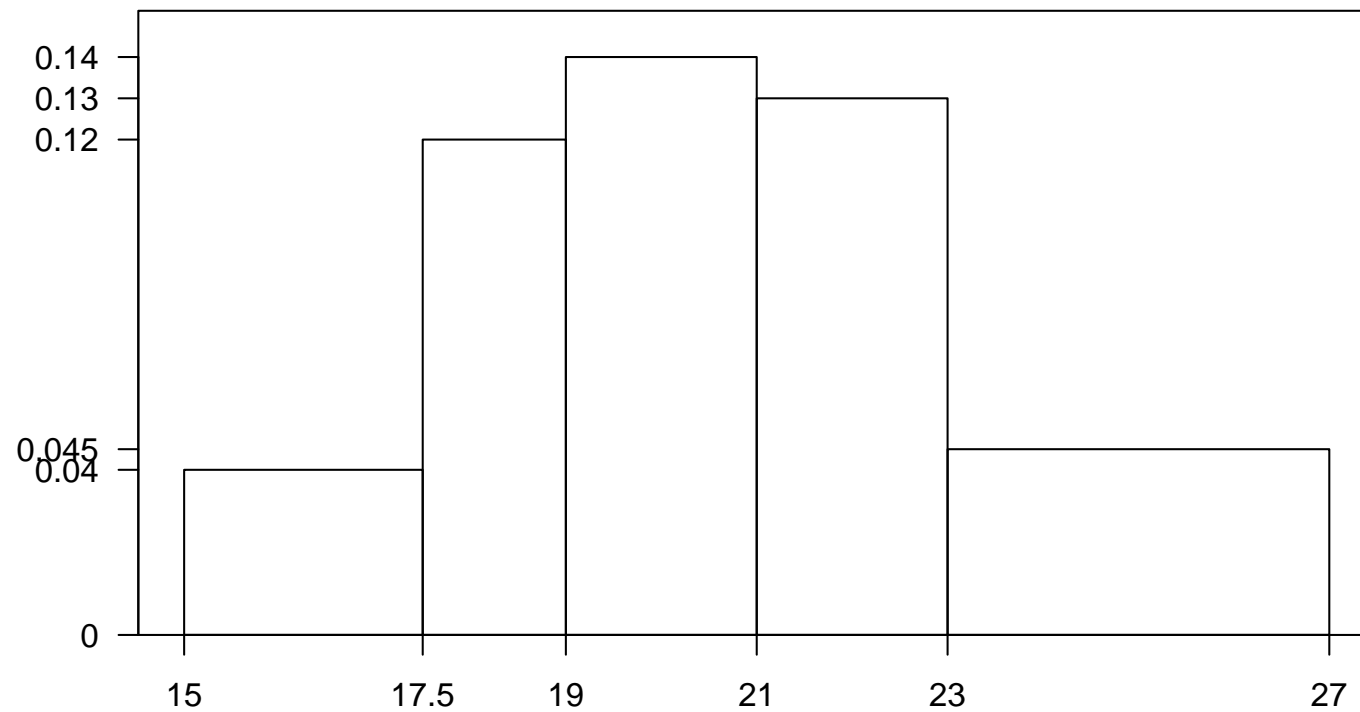
## Beispiel: Aufgabensammlung

**34.** Von einem Marktforschungsinstitut wurde der Preis für ein Kilogramm Tomaten in 120 Geschäften erhoben. Der Bericht enthielt ein Histogramm. Wie groß ist der Anteil der Geschäfte, in denen der Preis unter 19 GE liegt?

(Schluss von  $x$  auf Anteil)

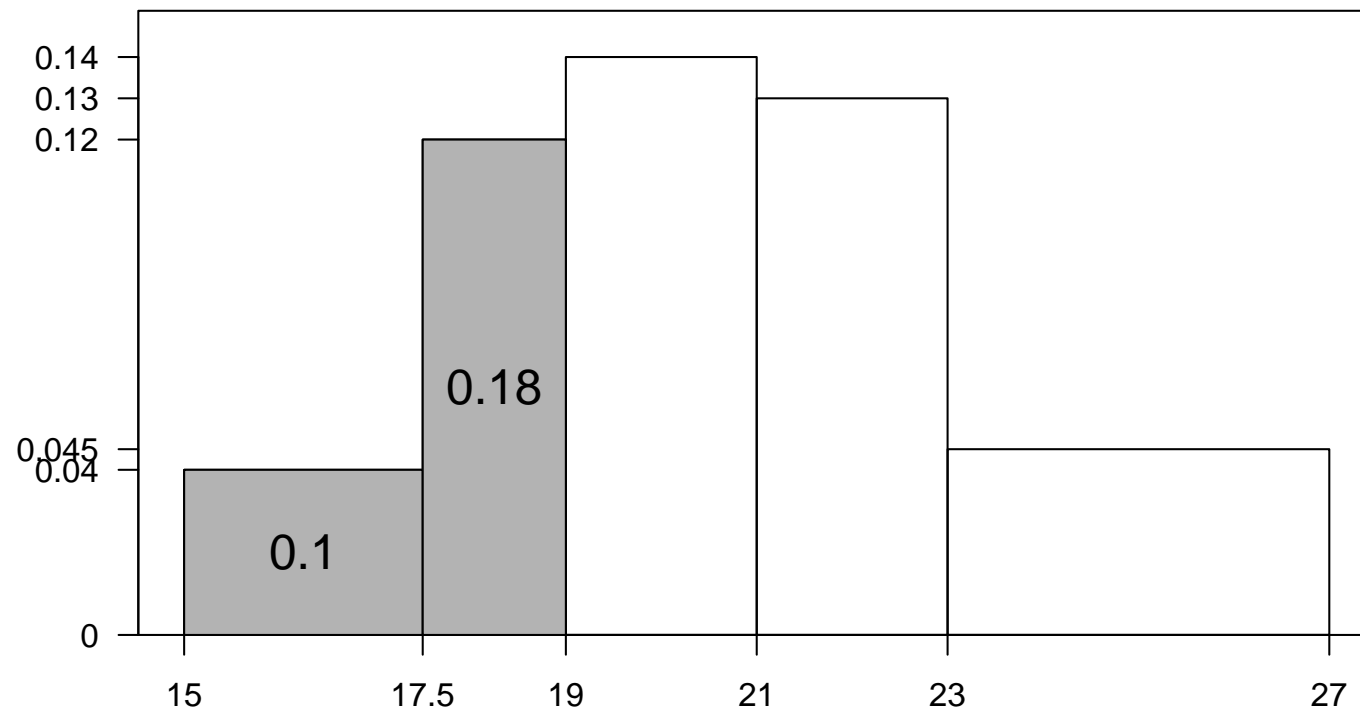
# Histogramm

---



# Histogramm

---



# Histogramm

---

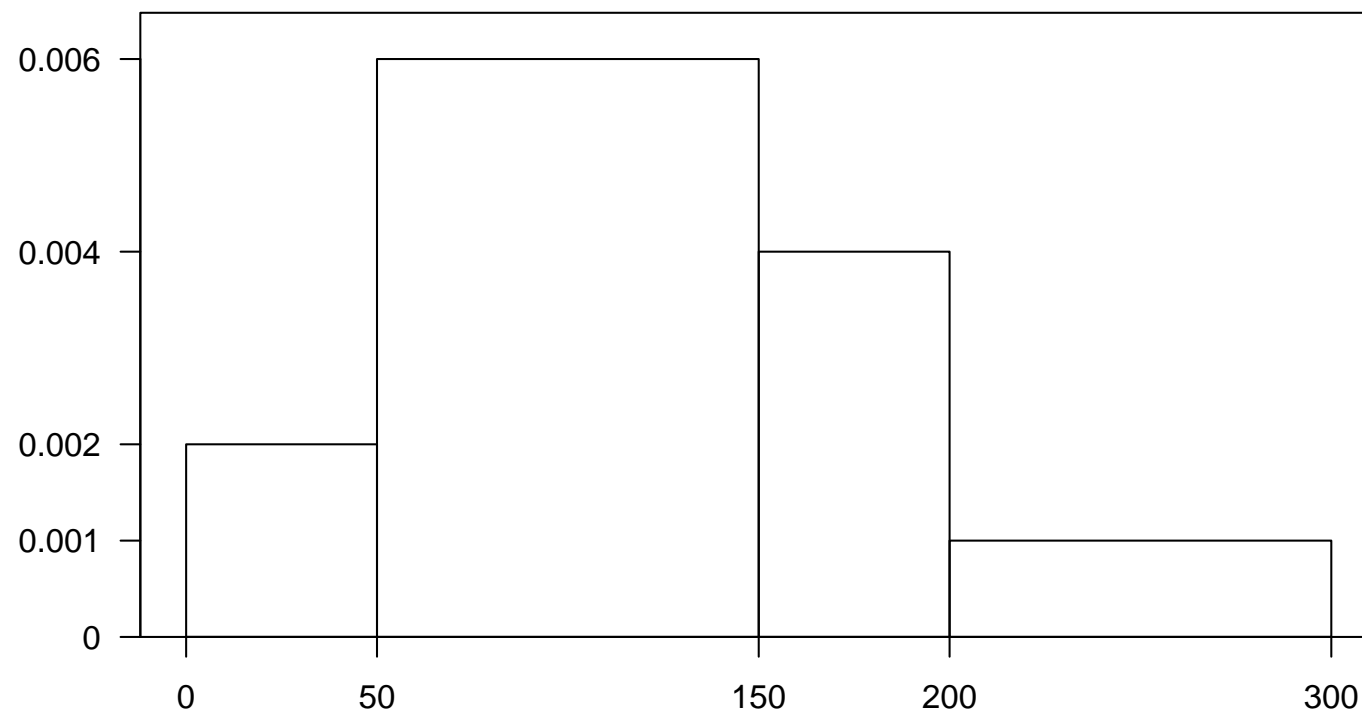
## Beispiel: Aufgabensammlung

**33.** Lesen Sie aus dem Histogramm, dessen Gesamtfläche 1 beträgt, die Häufigkeit ab, mit der der Wert 100 unterschritten wird.

(Schluss von  $x$  auf Anteil)

# Histogramm

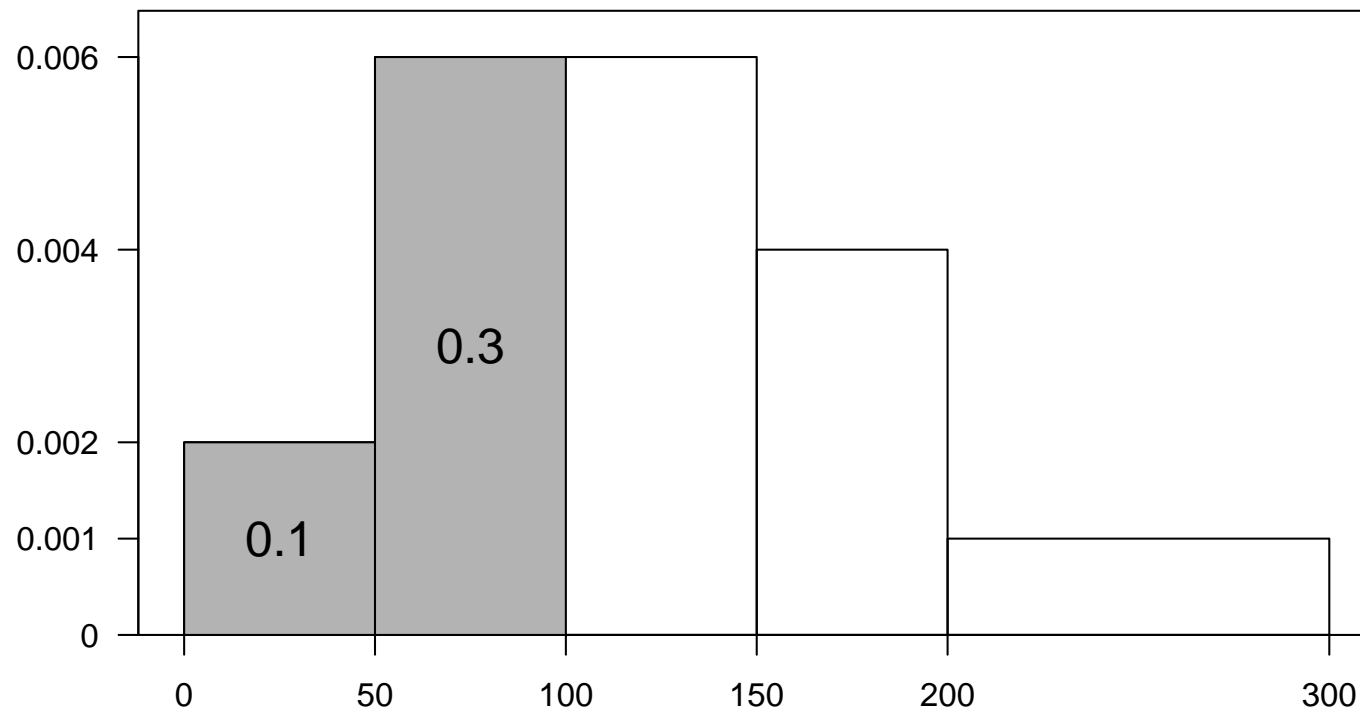
---





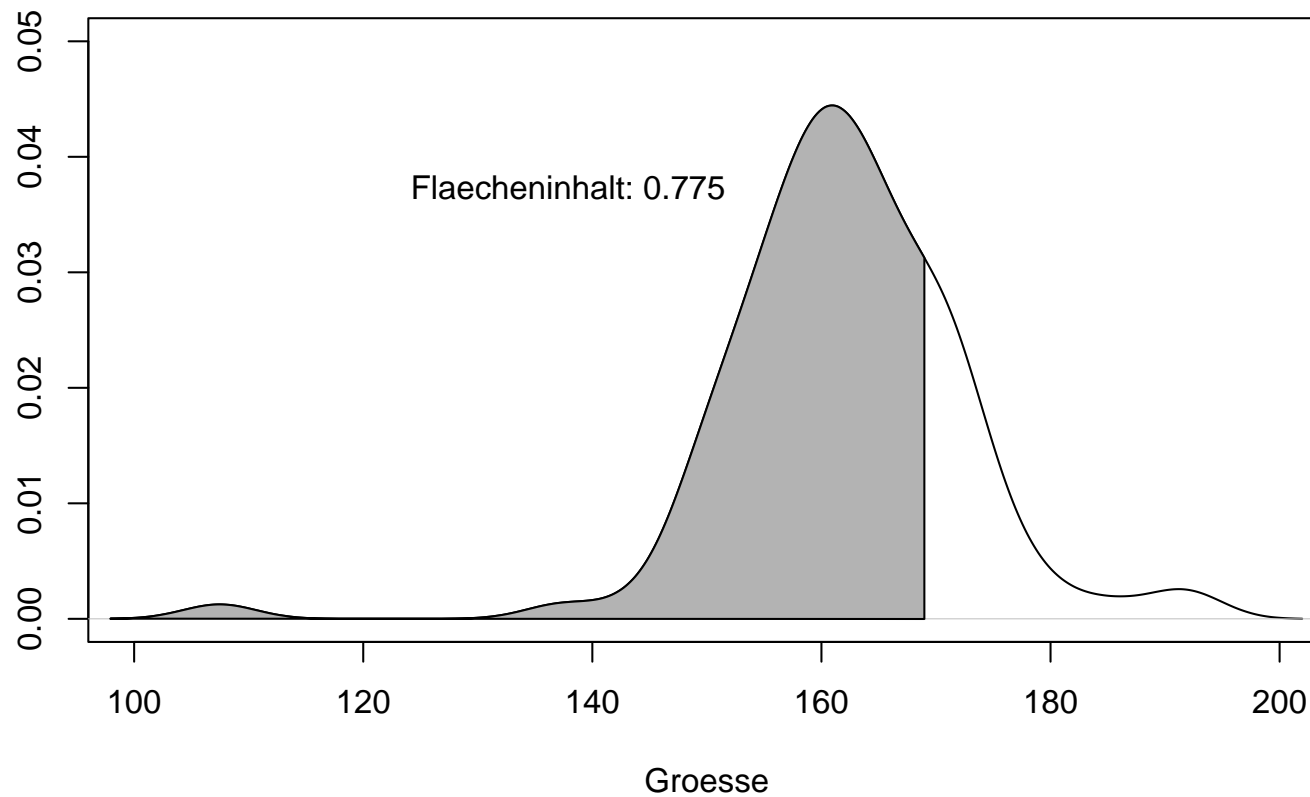
# Histogramm

---



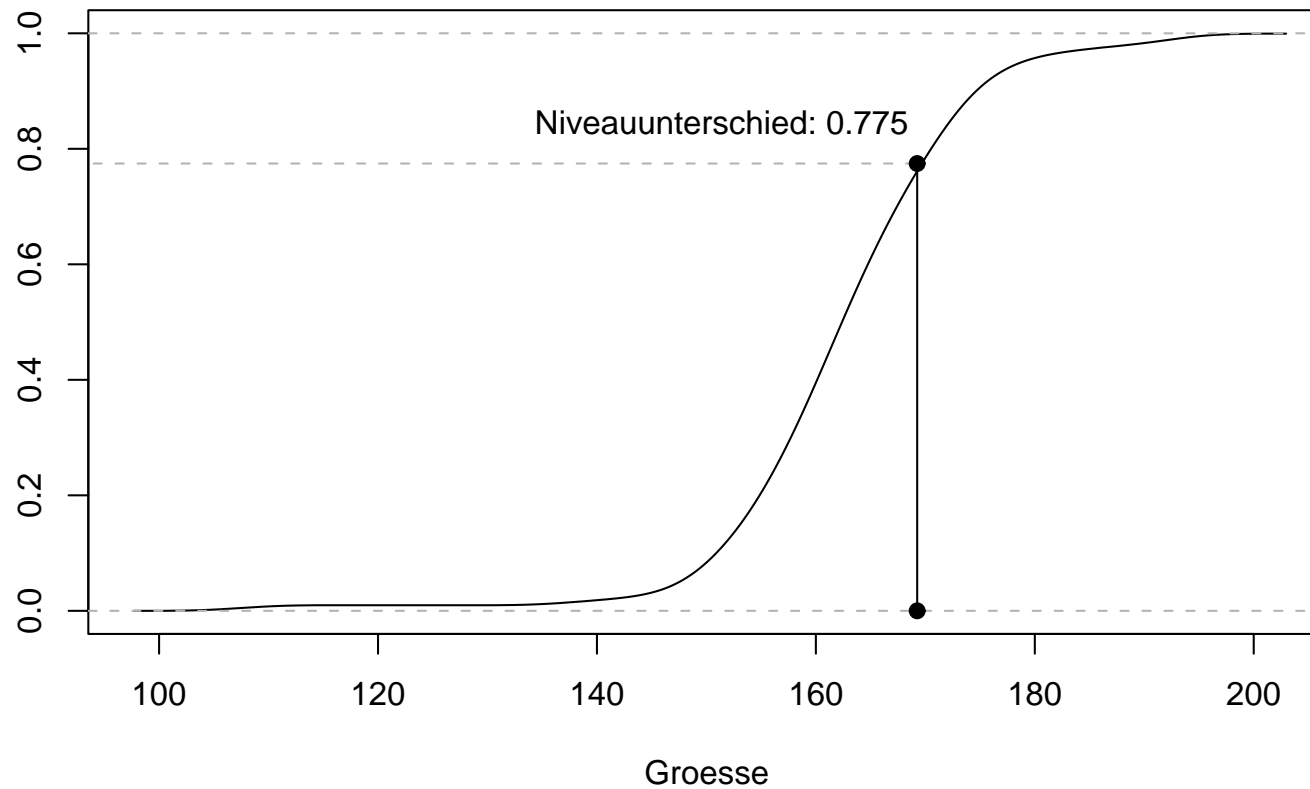
# Histogramm durch Glättung

---



# Summenkurve, Verteilungsfunktion

---



# Summenkurve, Verteilungsfunktion

---

Dichte des Histogramms sei  $f(x)$ ,  $x \in \mathbb{R}$ .

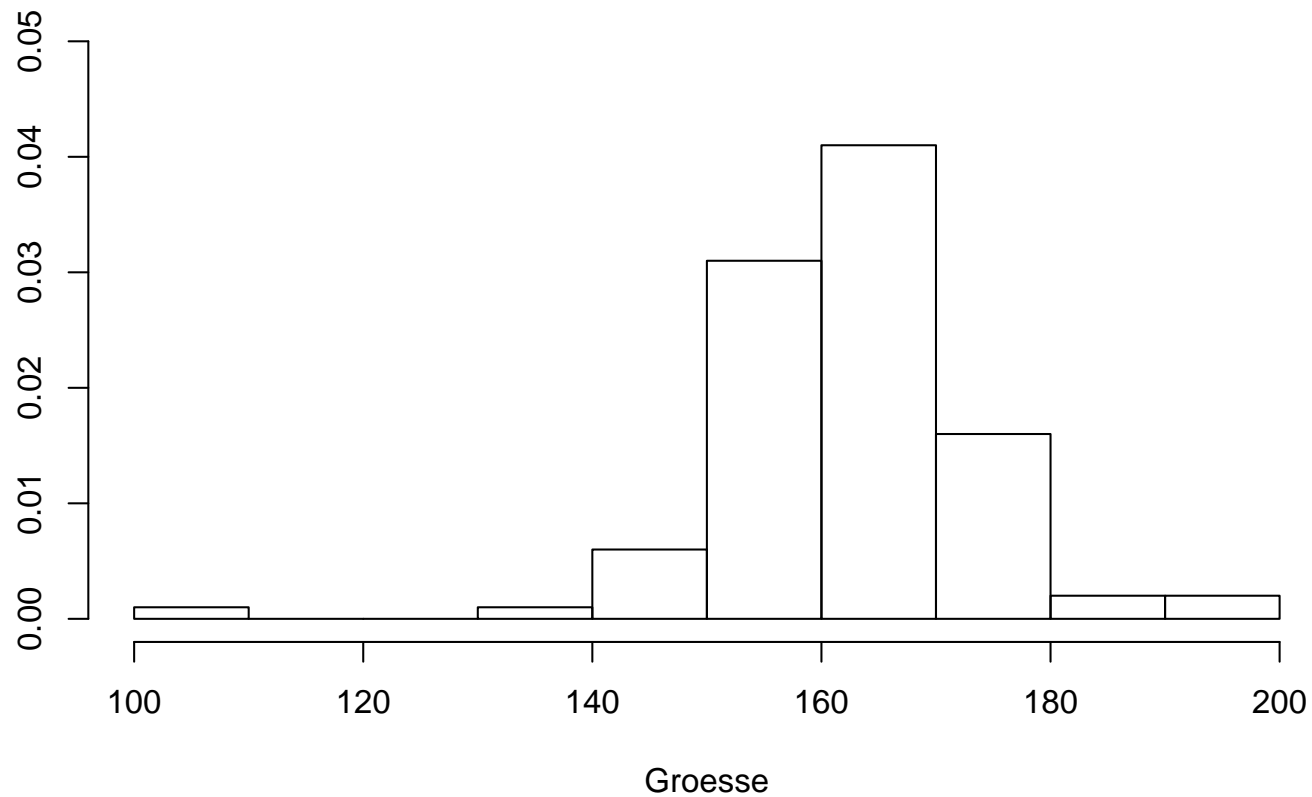
$F(x)$  ist der Flächeninhalt des Histogramms links von  $x$ .

$x \rightarrow F(x)$  heißt **Summenkurve** bzw. **Verteilungsfunktion** des Histogramms.

Die Summenkurve ist eine Approximation der empirischen Verteilungsfunktion.

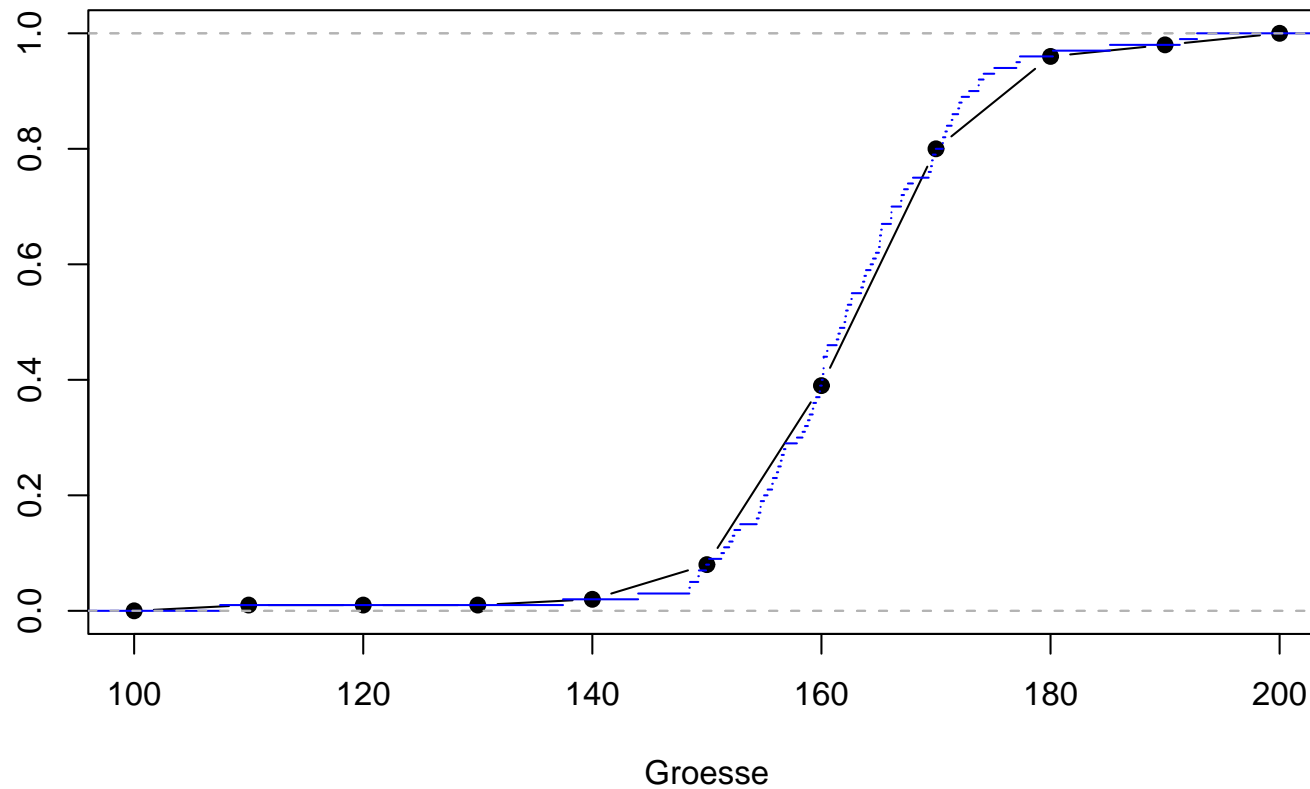
# Summenkurve, Verteilungsfunktion

---



# Summenkurve, Verteilungsfunktion

---



# Summenkurve, Verteilungsfunktion

---

zusammenfassend:

- Anteile auf Ordinate ablesbar:  
(empirische) Verteilungsfunktion, Summenkurve
- Anteile als Flächen:  
Histogramm

# Histogramm, Summenkurve

---

## Beispiel: Aufgabensammlung

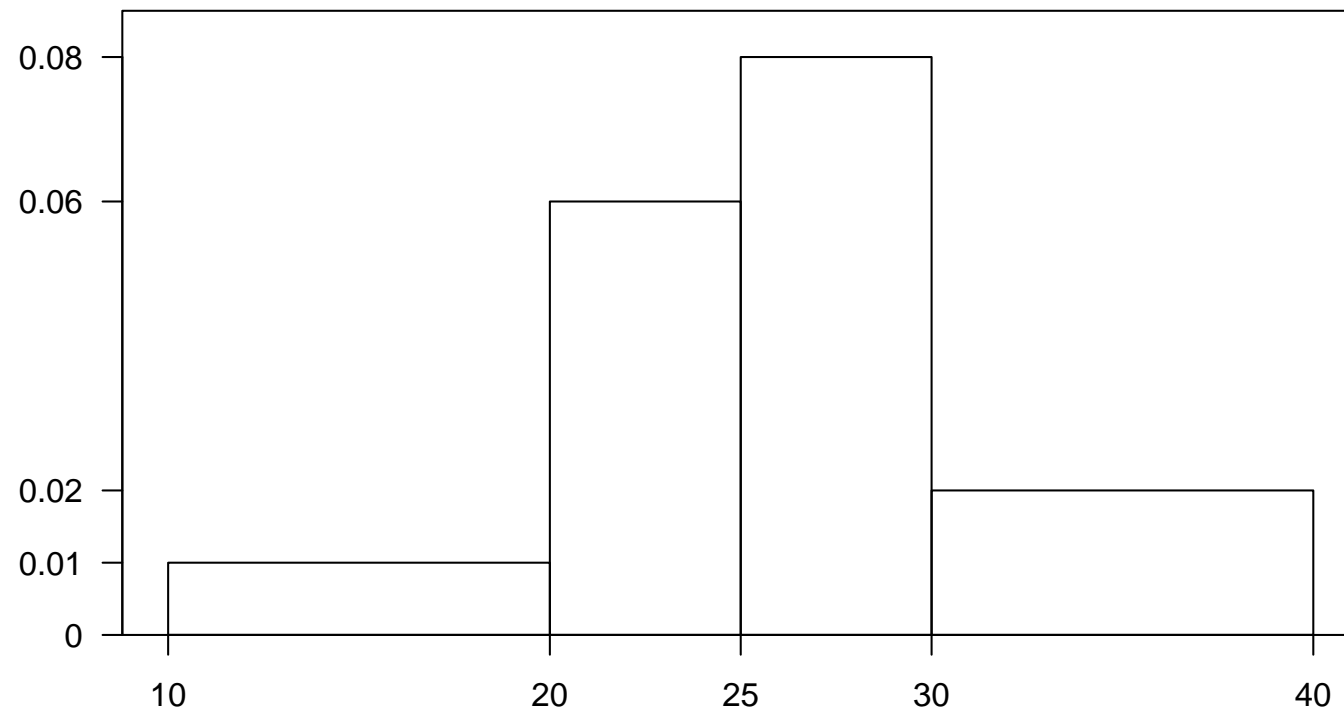
**26.** Abbildung 26 zeigt die Verteilung der Preise von Eigentumswohnungen (pro  $m^2$ ). Wie lautet die Summenkurve  $S(x)$  für  $x = 27$ ?

(Schluss von  $x$  auf Anteil  $S(x)$ )



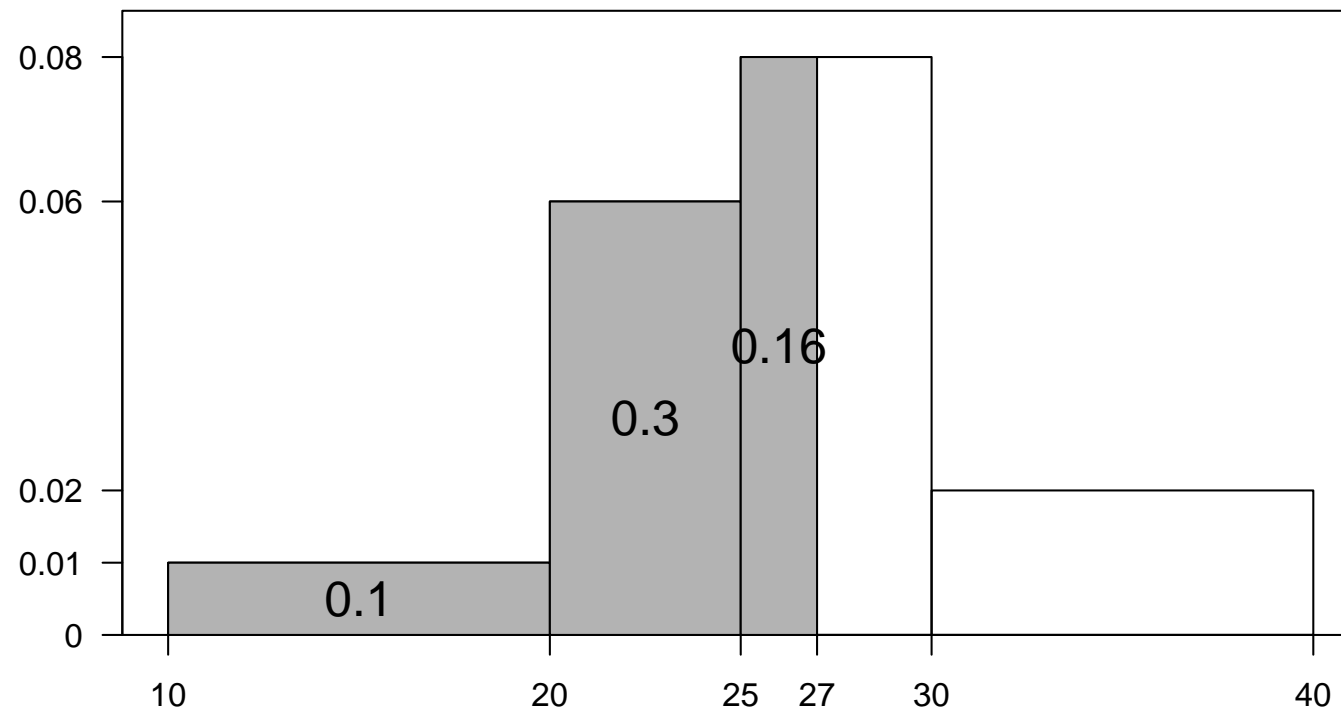
# Histogramm, Summenkurve

---



# Histogramm, Summenkurve

---



# Rangzahlen, Ordnungsgrößen

---

**Rangzahl:** Sind  $k$  Beobachtungen kleiner als  $x$ , so ist die Rangzahl  $r(x) = k + 1$ .

Die **Ordnungsgröße** der Daten bekommt man durch das Sortieren des Datensatzes:  $x_{(1)} < x_{(2)} < \dots < x_{(n)}$ .

# Rangzahlen, Ordnungsgrößen

---

## Beispiel:

Gegeben sei die Datenliste:

4.8, 6.6, 5.8, 4.8, 8.6, 4.3, 15.9, 4.3, 5.2, 10.3, 5.2, 4.5, 0.8

Wie lautet der Rang von 4.8?

Wie lautet die Ordnungsgröße  $x_{(3)}$ ?

# Quantile, Median, Five Point Summary

---

Das  $\alpha$ -**Quantil**  $Q_\alpha$  teilt die sortierte Datenliste in Teile vom Umfang

$$\alpha : 1 - \alpha$$

Der **Median** ist das Quantil mit  $\alpha = 1/2$ .

Die **Five Point Summary**:  $Q_0, Q_{0.25}, Q_{0.5}, Q_{0.75}, Q_1$ .

Die **Interquartilsdistanz** ist  $Q_{0.75} - Q_{0.25}$ .

# Quantile, Median, Five Point Summary

---

## Beispiel:

Gegeben sei die Datenliste:

4.8, 6.6, 5.8, 4.8, 8.6, 4.3, 15.9, 4.3, 5.2, 10.3, 5.2, 4.5, 0.8

Wie lautet der Median der Daten?

## Beispiel:

Preis 25 GE/Stück. Verkaufszahlen haben die Five Point Summary:

4, 25, 46, 62, 88

Welcher Erlös wurde an 25 Prozent der Tage überschritten?

Wie lautet die Interquartilsdistanz der Erlöse?

# Quantile, Median, Five Point Summary

---

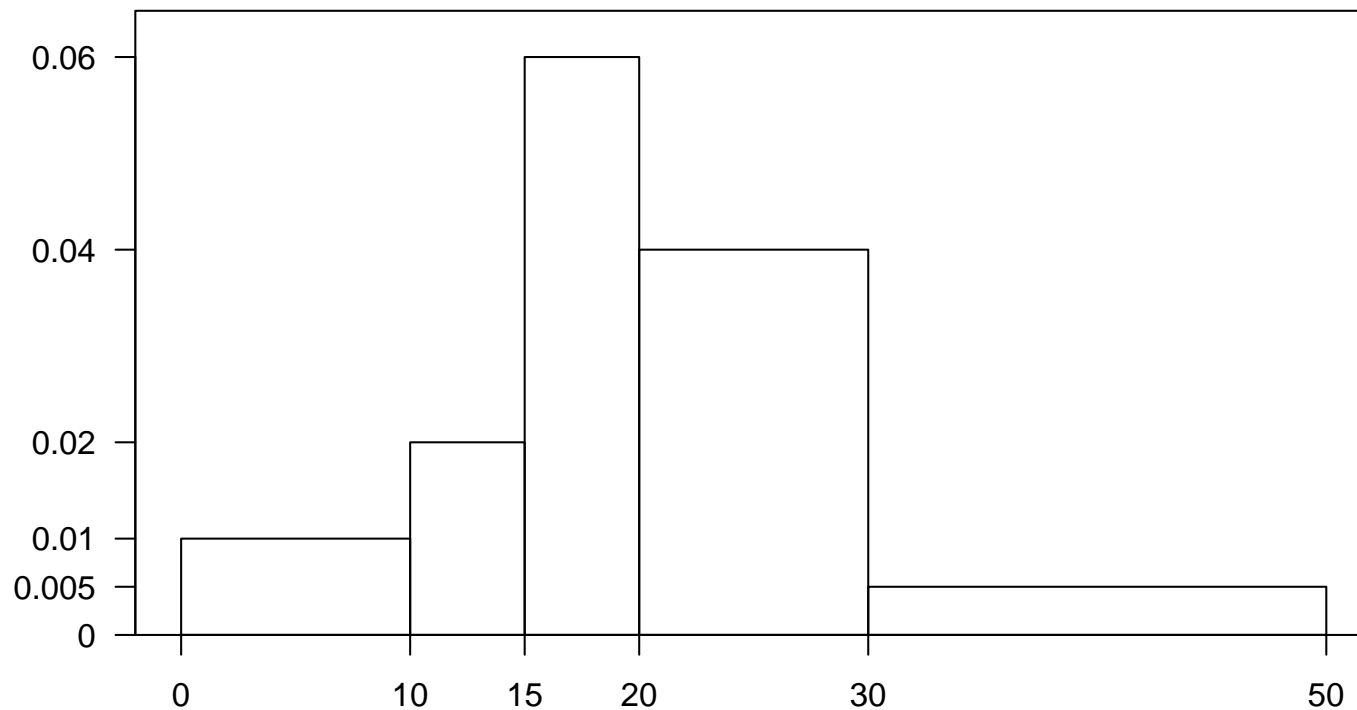
## Beispiel: Aufgabensammlung

**36.** Für einen Handwerksbetrieb wurden die Wegzeiten (vom Betrieb zu den Kunden, in Minuten) ermittelt und zu einem Histogramm zusammengefaßt. Wie lautet der Median des Histogramms?

(Schluss von Anteil auf  $x$ )

# Quantile, Median, Five Point Summary

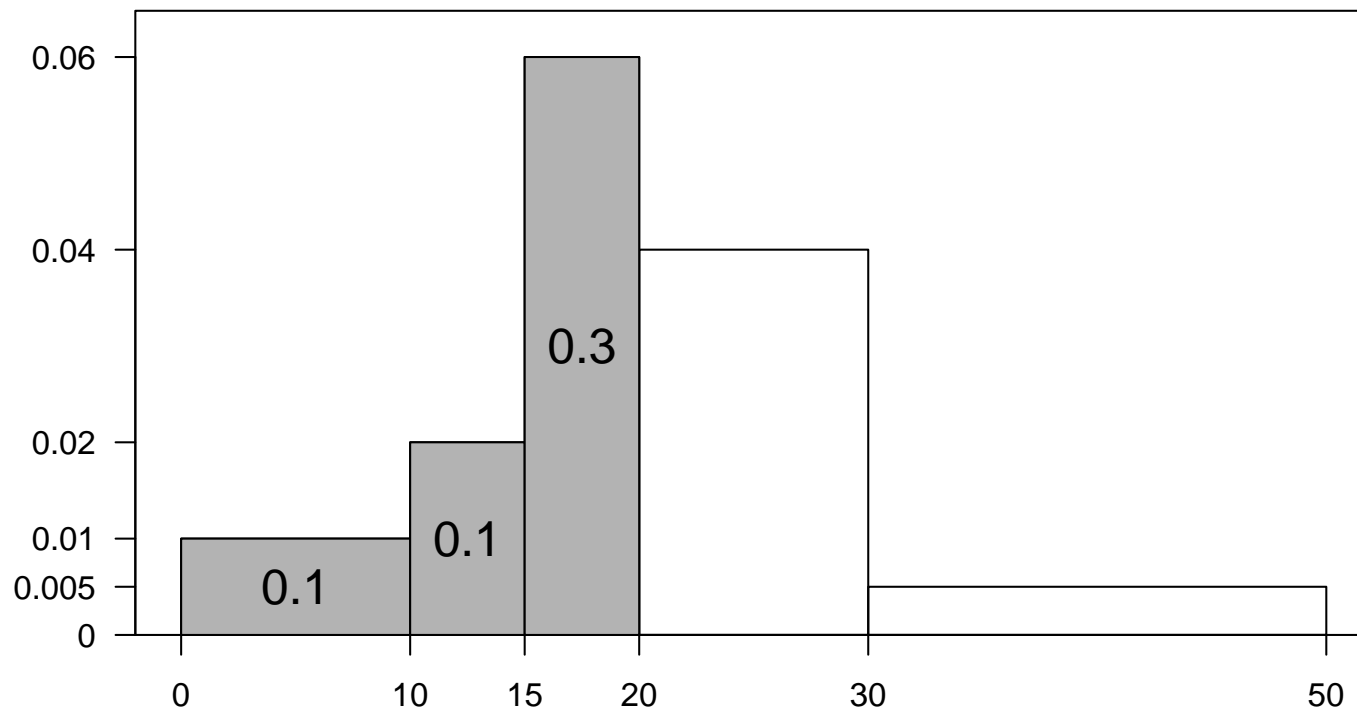
---





# Quantile, Median, Five Point Summary

---



# Quantile, Median, Five Point Summary

---

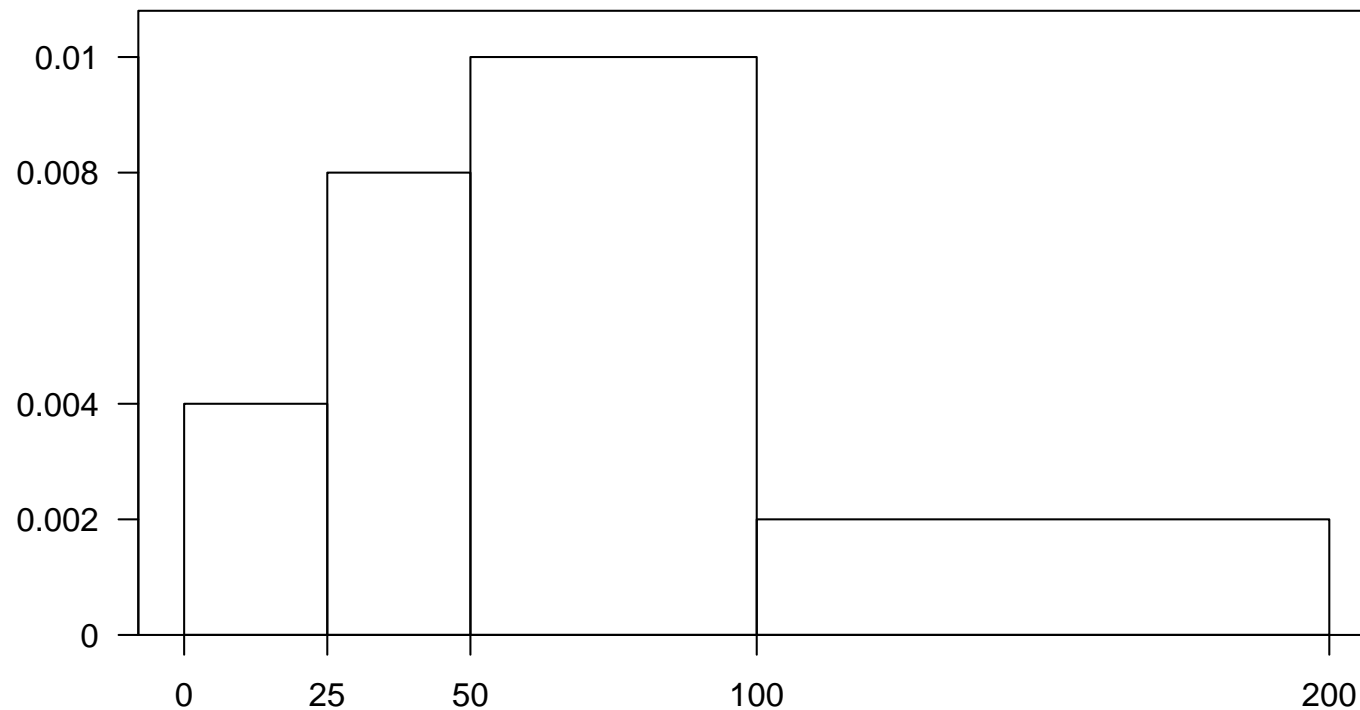
**Beispiel:** Aufgabensammlung

**12.** Abbildung 12 zeigt die Verteilung der Restlaufzeiten von Krediten einer Bank (in Monaten). Wie lautet der Median des Histogramms?

(Schluss von Anteil auf  $x$ )

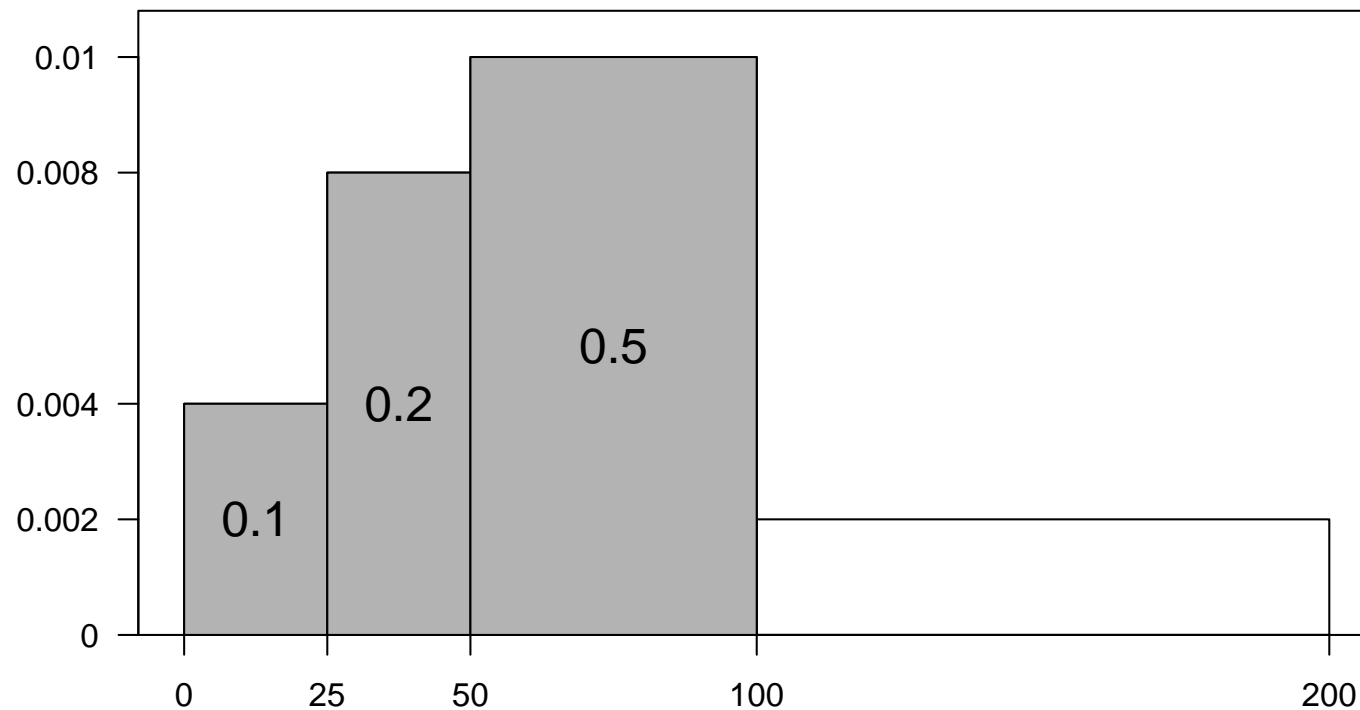
# Quantile, Median, Five Point Summary

---



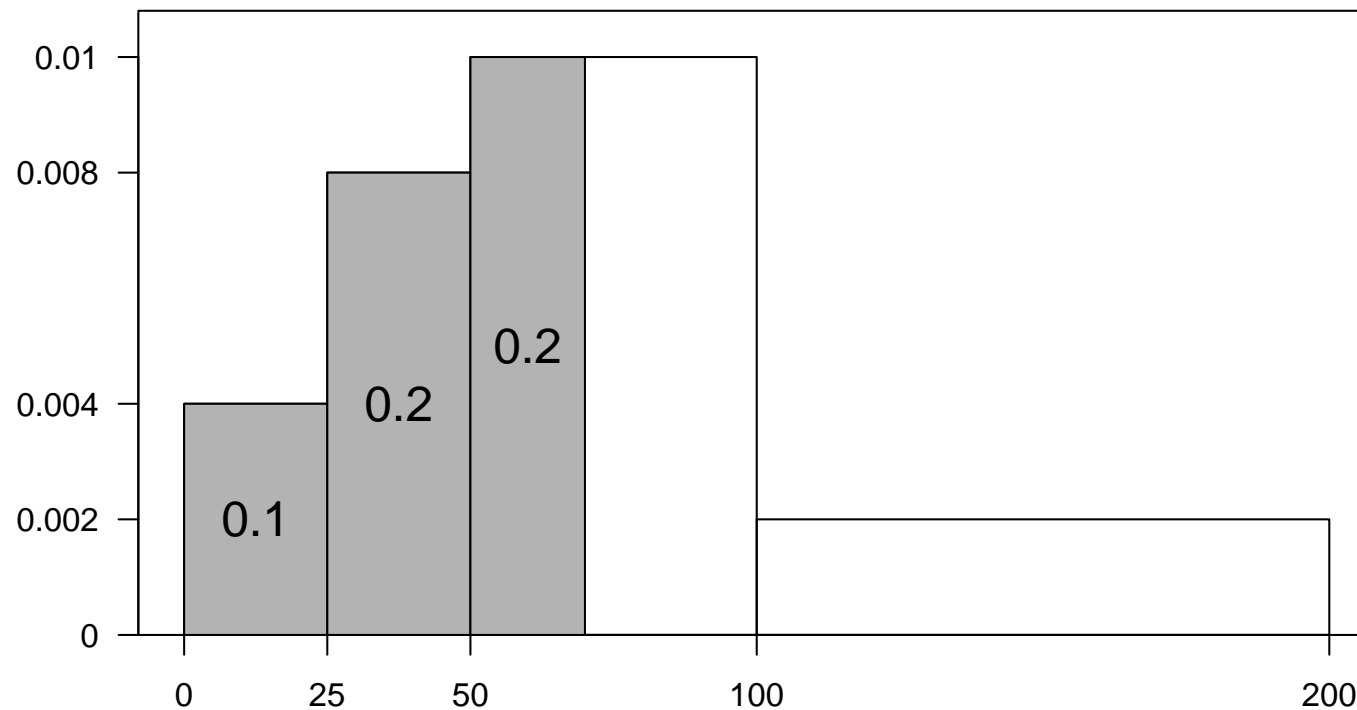
# Quantile, Median, Five Point Summary

---



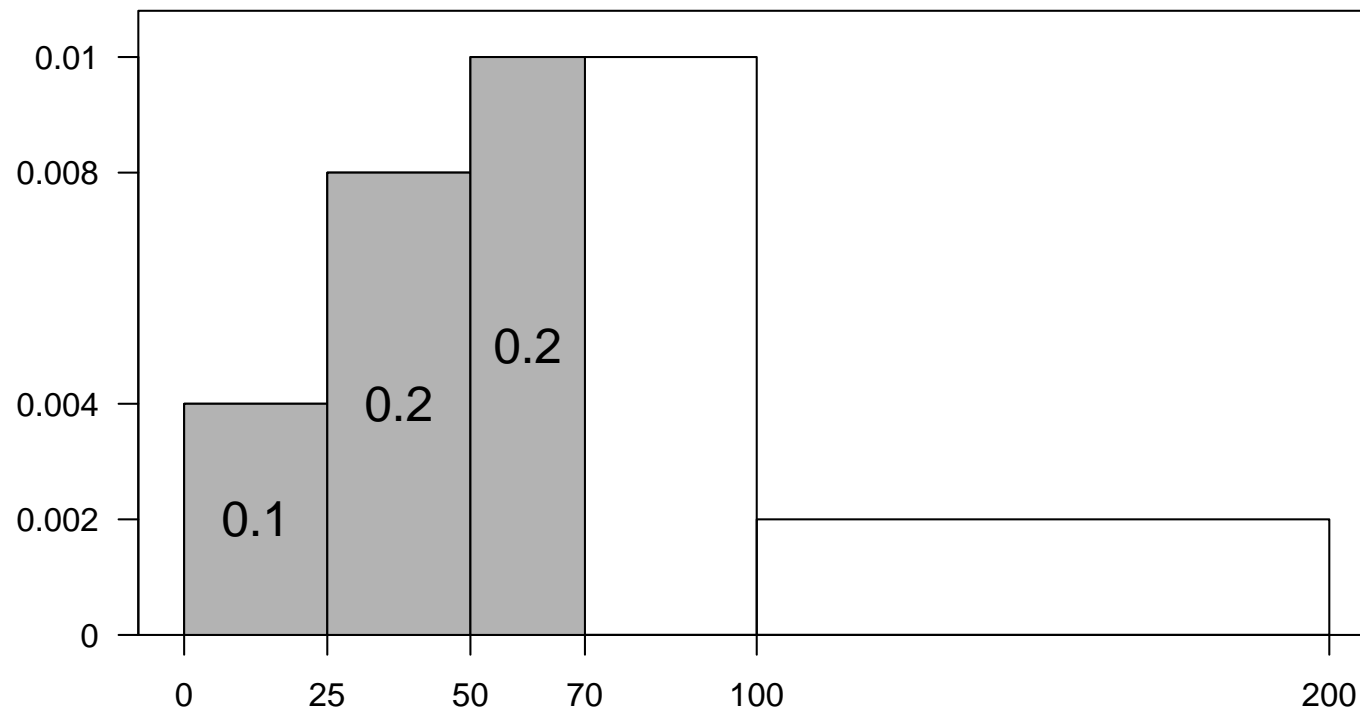
# Quantile, Median, Five Point Summary

---



# Quantile, Median, Five Point Summary

---



# Zusammenfassung Kapitel 2

---

- Skalentypen
- Häufigkeitstabellen: qualitativ - quantitativ
- Empirische Verteilungsfunktion
- Histogramm
- Summenkurve, Verteilungsfunktion
- Rangzahlen, Ordnungsgrößen
- Quantile, Median, Five point summary, Interquartilsdistanz

# Die Rolle des Zufalls

## Kapitel 3



# Erhebungsformen

---

- **Totalerhebung**

Bsp. Volkszählung

- **Repräsentative Erhebung**

- Stichprobenerhebung

Bsp. Mikrozensus, Marktforschung

- Zufallsexperimente

Bsp. Münzwurf

→ Wahrscheinlichkeitsrechnung, Stochastik

# Gesetz der Großen Zahlen, Wahrscheinlichkeit

---

**Beispiel:** Zufallsexperiment Münzwurf

**Ereignis:**  $A = \text{Zahl}$

**Wahrscheinlichkeit** für das Eintreten von  $A$ :

$$P(A) = 1/2$$

$n$  **unabhängige** Versuchswiederholungen  $\rightarrow$  statistische Auswertung

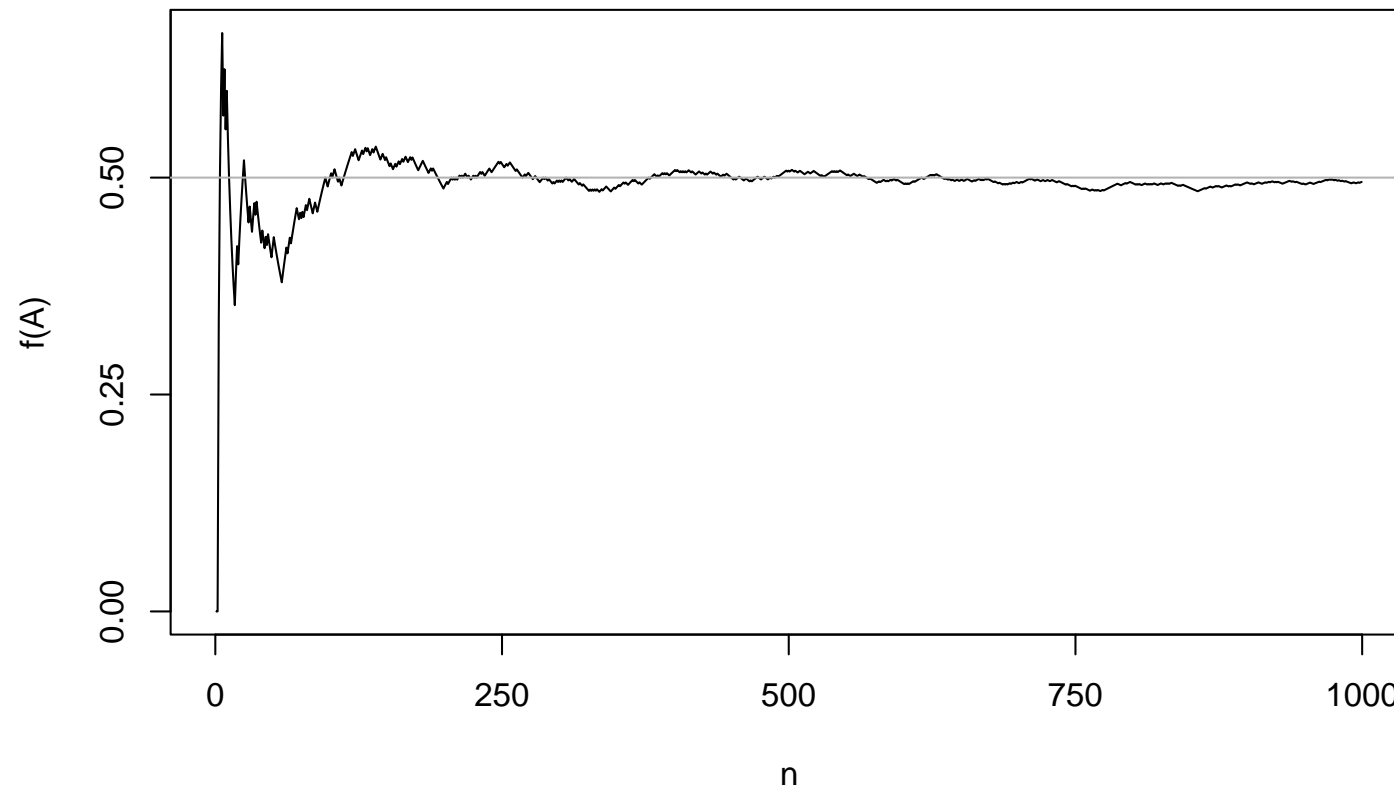
# Gesetz der Großen Zahlen, Wahrscheinlichkeit

---

$n$	$h(A)$	$f(A)$	$ f(A) - \frac{1}{2} $
10	3	0.3	0.2
100	47	0.47	0.03
500	254	0.508	0.008
1000	488	0.488	0.012
5000	2453	0.4906	0.0094

# Gesetz der Großen Zahlen, Wahrscheinlichkeit

---



# Gesetz der Großen Zahlen, Wahrscheinlichkeit

---

**Beispiel:** Zufallsexperiment Würfeln

**Ereignisse:**  $A$  = Augenzahl 1 und  $B$  = Augenzahl 5

**zusammengesetztes Ereignis:**  $A \cup B$ ,  $A$  oder  $B$

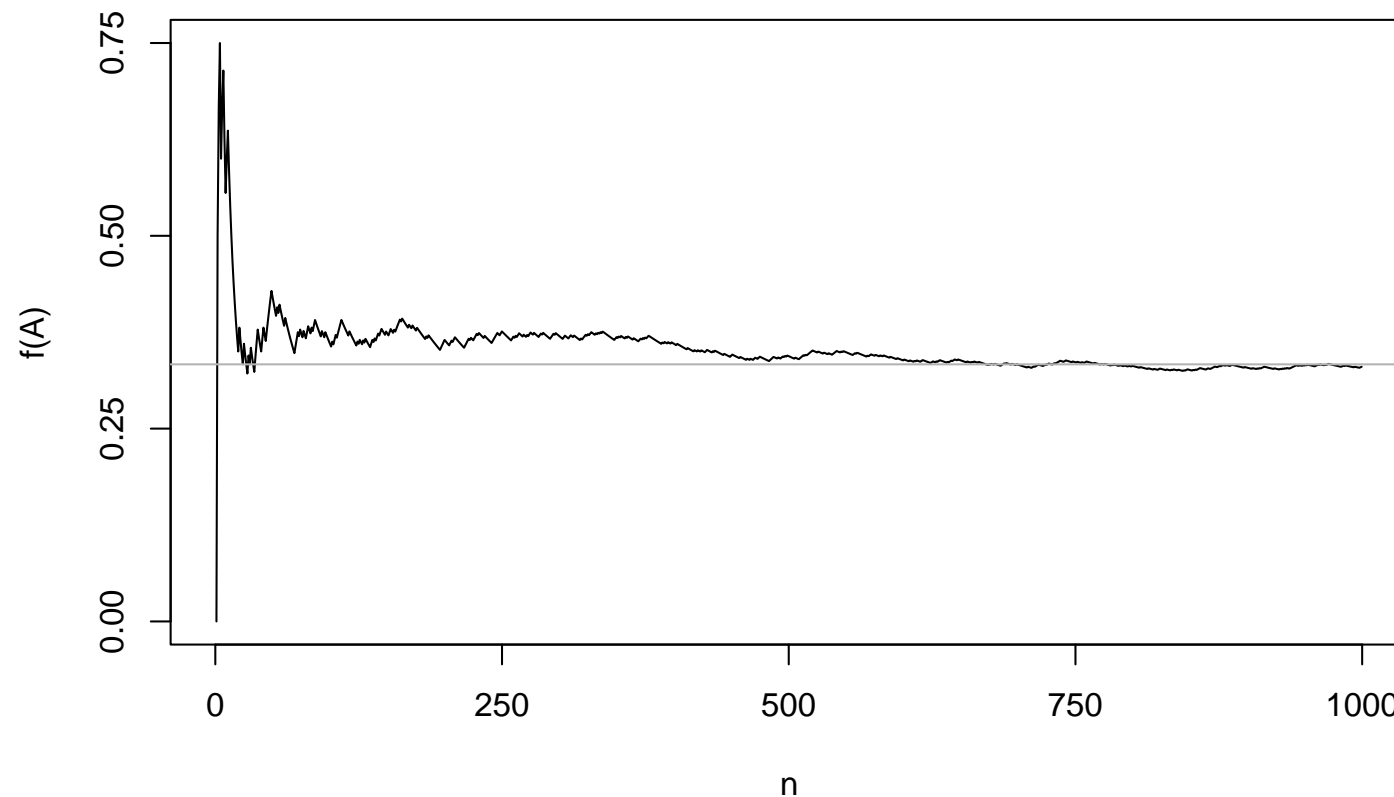
**Wahrscheinlichkeit** für das Eintreten von  $A \cup B$ :

$$P(A \cup B) = 1/3$$

$n$  **unabhängige** Versuchswiederholungen  $\rightarrow$  statistische Auswertung

# Gesetz der Großen Zahlen, Wahrscheinlichkeit

---



# Gesetz der Großen Zahlen, Wahrscheinlichkeit

---

## Empirisches Gesetz der großen Zahlen

Führe  $n$  Zufallsexperimente unter identischen Bedingungen durch.  
Die einzelnen Experimente sind voneinander unabhängig.

$$\lim_{n \rightarrow \infty} f_n(A) = P(A).$$

## Wahrscheinlichkeit

Die Wahrscheinlichkeit für das Ereignis  $A$  ist der Grenzwert der relativen Häufigkeiten  $f_n(A)$ .

# Wahrscheinlichkeit

---

## Rechengesetze:

- $0 \leq P(A) \leq 1$
- $P(\emptyset) = 0, P(\Omega) = 1$
- Monotoniegesetz:  $A \subseteq B \Rightarrow P(A) \leq P(B)$
- Additionsgesetz:  $A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$
- Siebformel:  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$



# Methode von Laplace

---

Ereignisse  $A_1, A_2, \dots, A_m$  bilden eine **Zerlegung** der Ergebnismenge  $\Omega$ :

**paarweise unvereinbar**, d. h.  $A_i \cap A_j = \emptyset$ ,  $i \neq j$

**vollständig**, d. h.  $A_1 \cup A_2 \cup \dots \cup A_m = \Omega$ .

Dann gilt:

$$P(A_1) + P(A_2) + \dots + P(A_m) = P(\Omega) = 1$$

# Methode von Laplace

---

Sind alle Ereignisse  $A_i$  gleichwahrscheinlich, dann gilt:

$$P(A_1) = P(A_2) = \dots = P(A_m) = \frac{1}{m}$$

Für ein **zusammengesetztes Ereignis**  $B = A_{i_1} \cup \dots \cup A_{i_g}$

$$P(B) = \frac{g}{m}$$

$g$  ... Anzahl der günstigen Fälle

$m$  ... Anzahl der möglichen Fälle

# Binomialverteilung

---

Zufallsexperiment mit:

- 2 alternativen Ereignissen  $A$  und  $A'$
- $n$  Wiederholungen
- unabhängige Versuche unter gleichen Bedingungen

Dann:

- $P(A) = p$
- $h_n(A)$  nimmt Werte  $0, 1, \dots, n$  an

# Binomialverteilung

---

Wahrscheinlichkeit, daß bei  $n$  Versuchen  $A$  genau  $k$  mal eintritt,  $k = 0, 1, \dots, n$ :

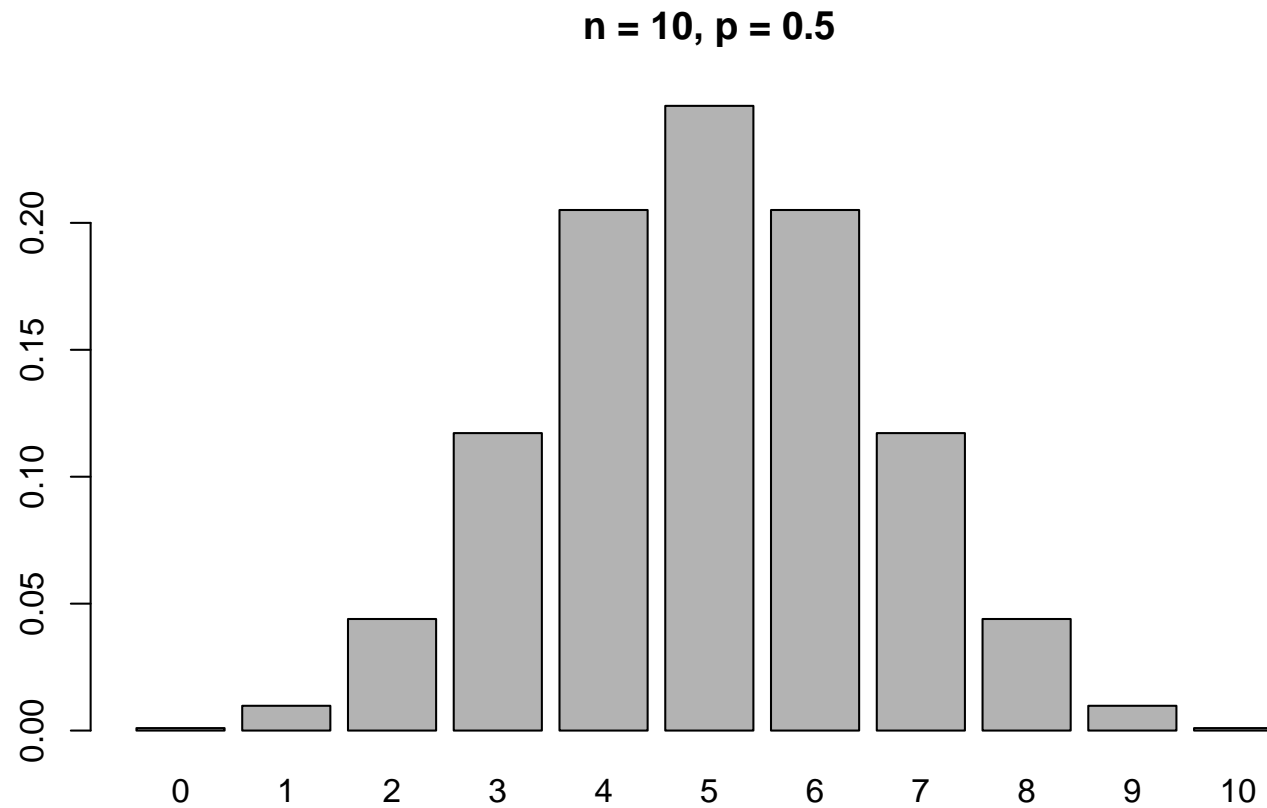
$$P(h_n(A) = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

Binomialkoeffizient:

$$\binom{n}{k} = \frac{n!}{(n - k)! k!}$$

# Binomialverteilung

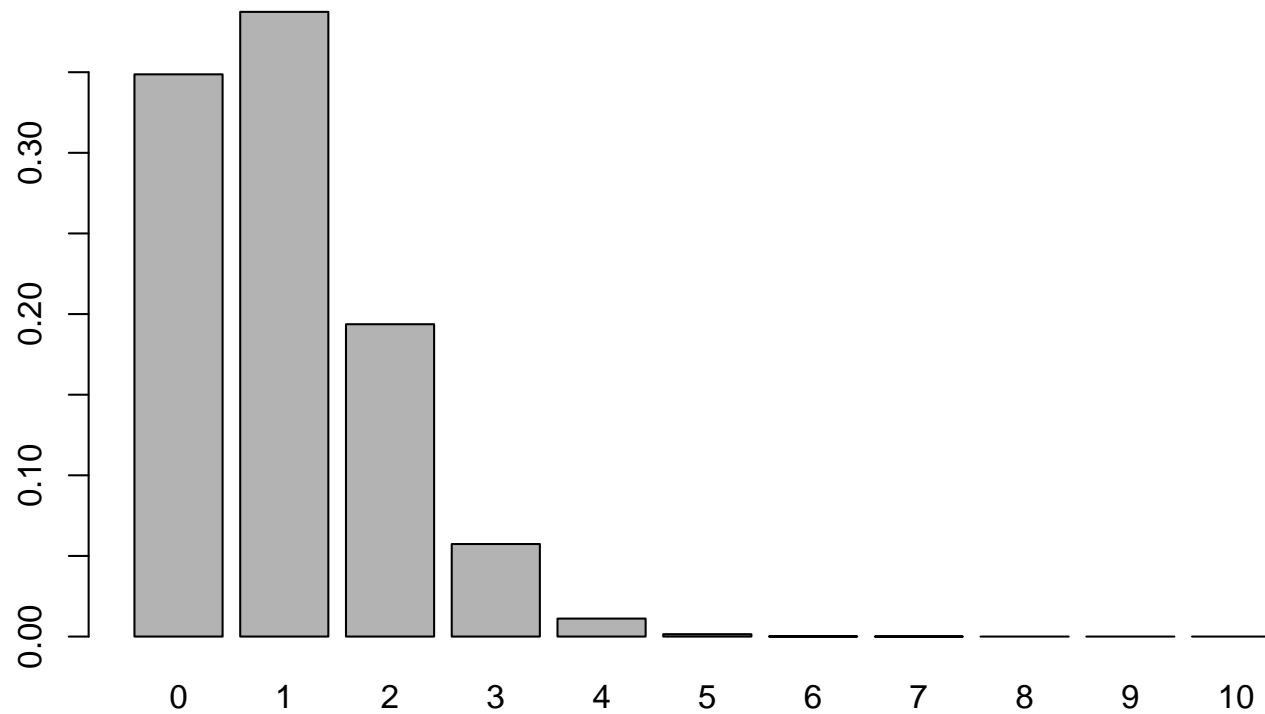
---



# Binomialverteilung

---

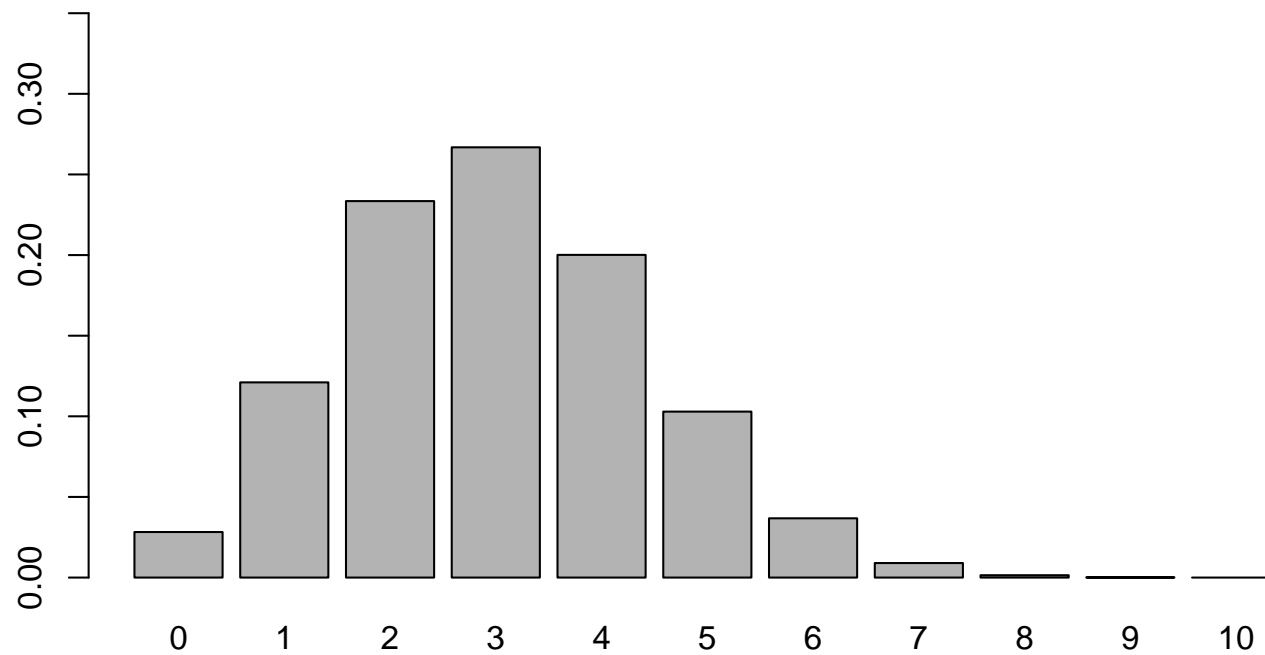
$n = 10, p = 0.1$



# Binomialverteilung

---

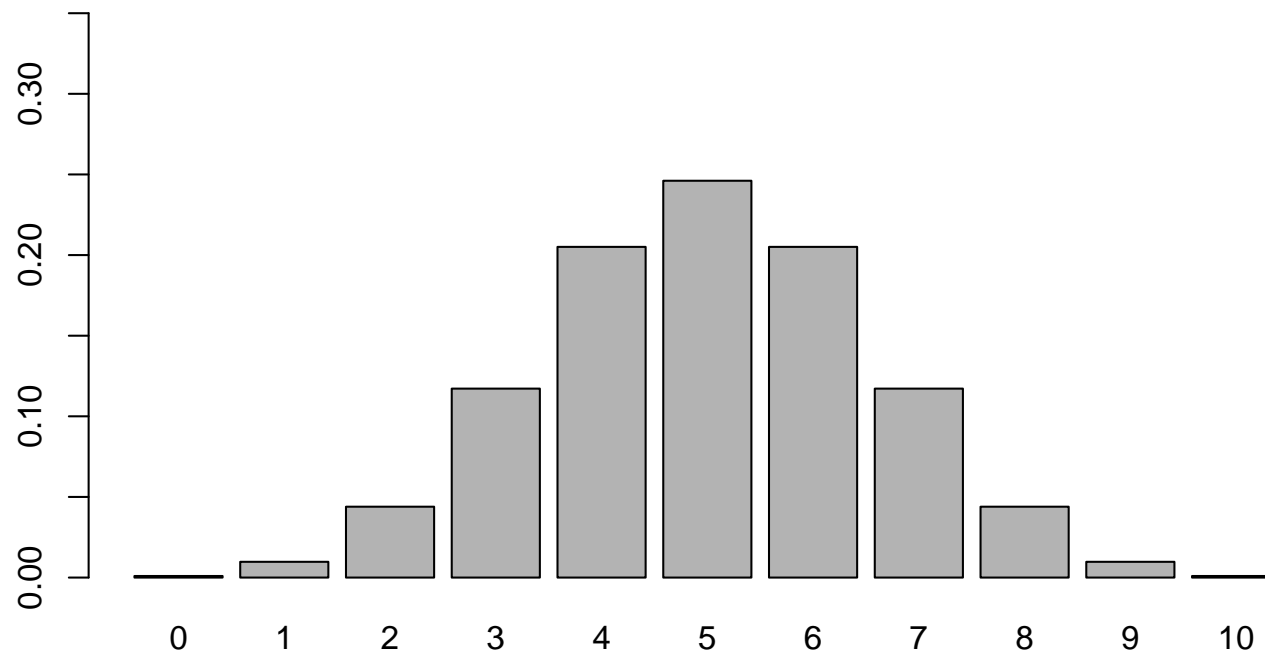
$n = 10, p = 0.3$



# Binomialverteilung

---

$n = 10, p = 0.5$

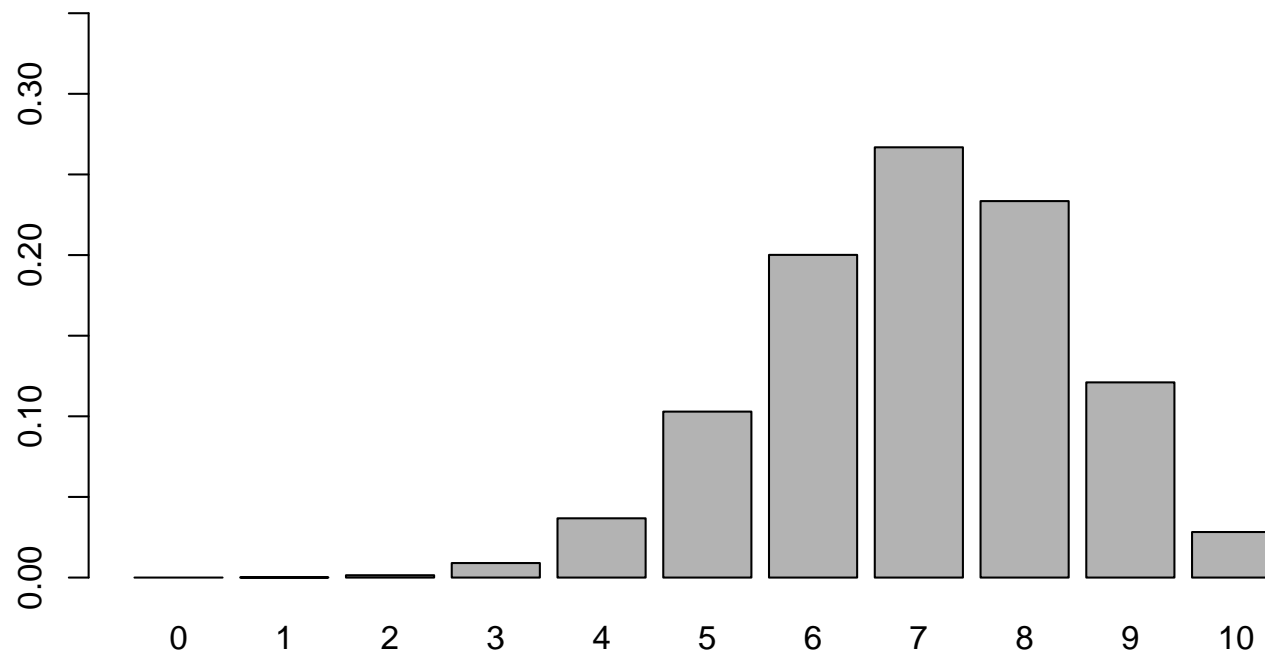




# Binomialverteilung

---

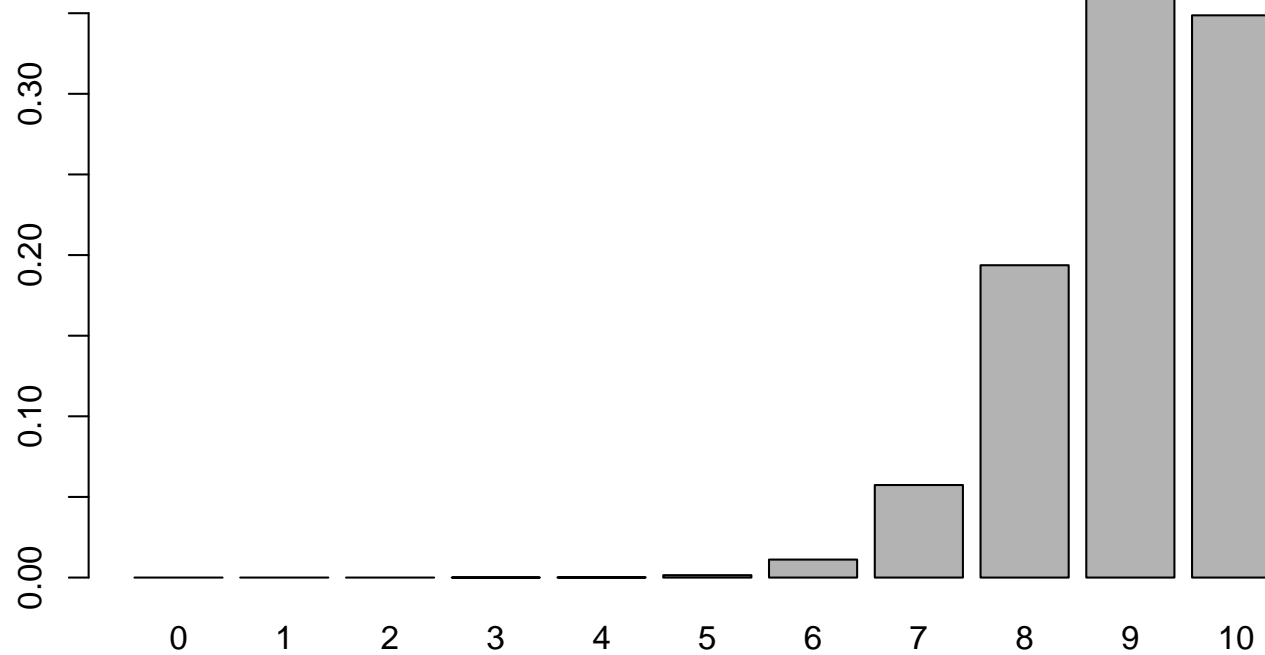
$n = 10, p = 0.7$



# Binomialverteilung

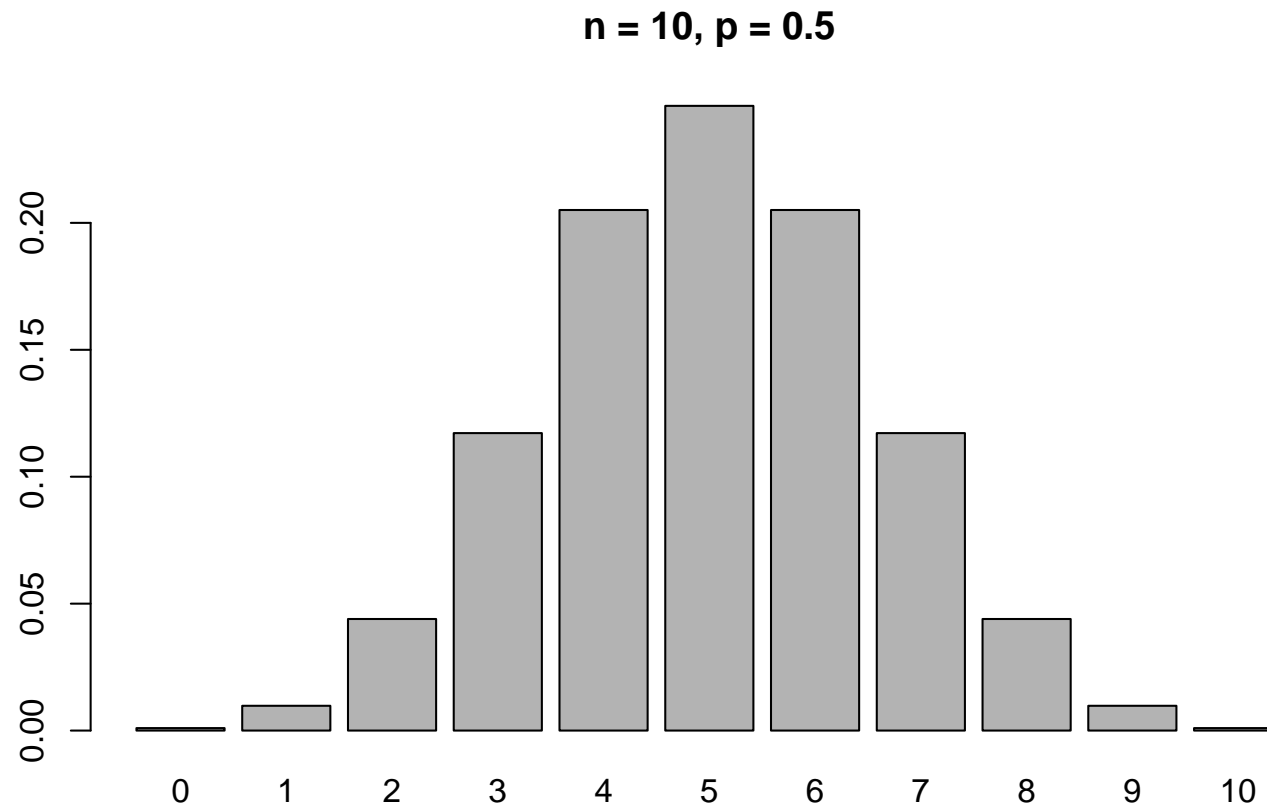
---

$n = 10, p = 0.9$



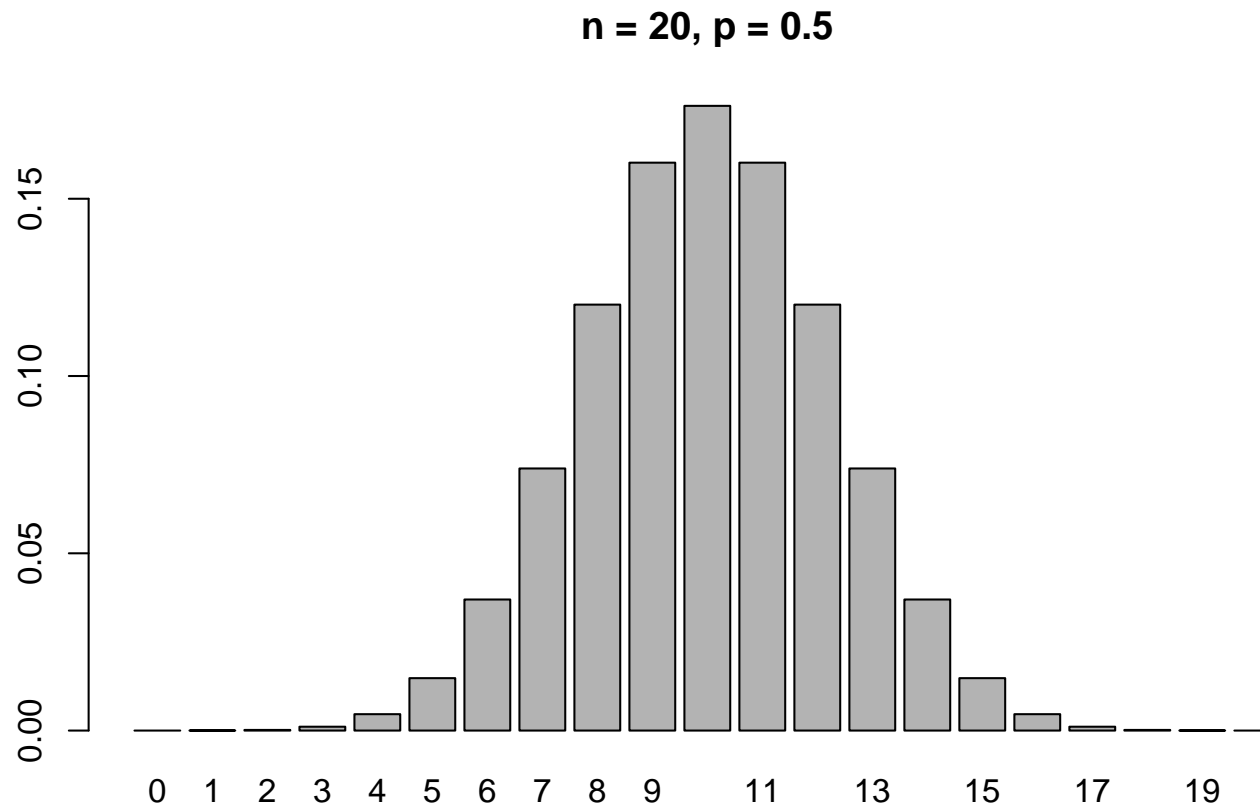
# Binomialverteilung

---



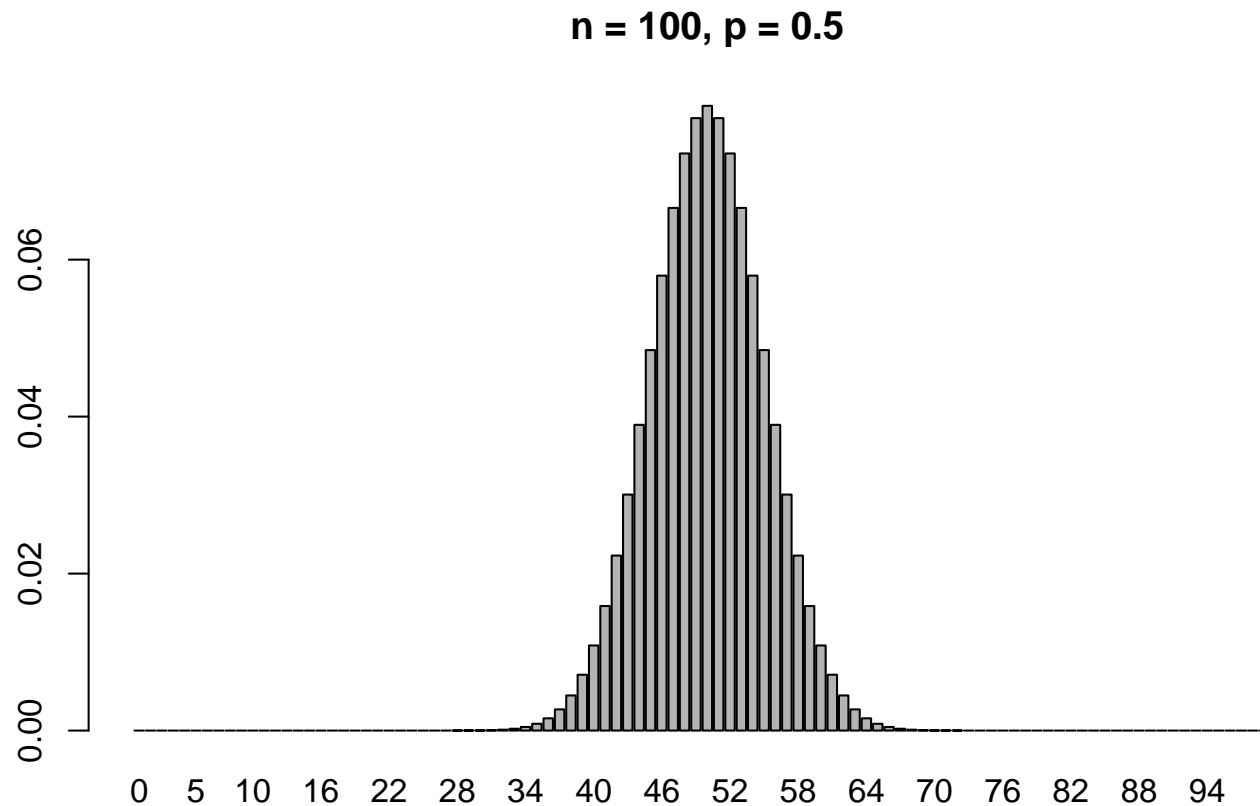
# Binomialverteilung

---



# Binomialverteilung

---



# Binomialverteilung

---

## Beispiel: Aufgabensammlung

**46.** 41% aller Kinder werden ohne Sicherheitssitz im Auto transportiert. Es werden zufällig 10 Autos, die Kinder befördern, kontrolliert.

Wie groß ist die Wahrscheinlichkeit, daß in genau 3 Autos das Kind ohne Sicherheitssitz befördert wird?

# Binomialverteilung

---

## Beispiel: Aufgabensammlung

**52.** 15% aller Hühner einer bestimmten Farm sind mit Salmonellen verseucht. Für ein Grillfest werden 10 Hühner dieser Farm eingekauft.

Wie groß ist die Wahrscheinlichkeit, daß mindestens 2 Hühner verseucht sind?

# Hypergeometrische Verteilung

---

Zufallsexperiment mit:

- Grundgesamtheit mit  $N$  Objekten
- $M$  Objekte mit einer Eigenschaft  $A$
- alle Objekte der Grundgesamtheit haben dieselbe Chance

**Formel von Laplace:**

$$P(A) = \frac{M}{N}$$



# Hypergeometrische Verteilung

---

Ziehung von  $n$  Objekten ohne Zurücklegen aus der Grundgesamtheit der Größe  $N$ .

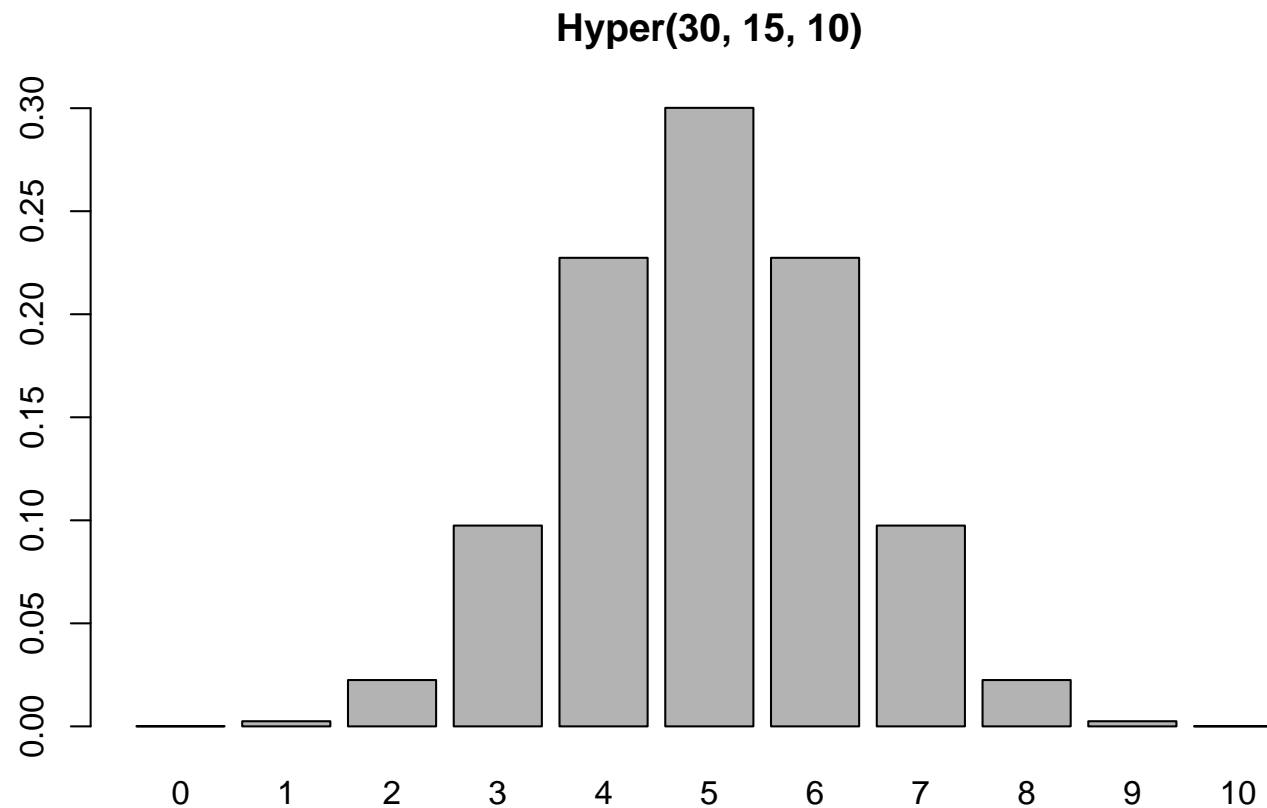
Die Wahrscheinlichkeit, daß genau  $k$  Objekte die Eigenschaft  $A$  haben ist:

$$P(h_n(A) = k) = \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}},$$

$$k = 0, 1, \dots, n.$$

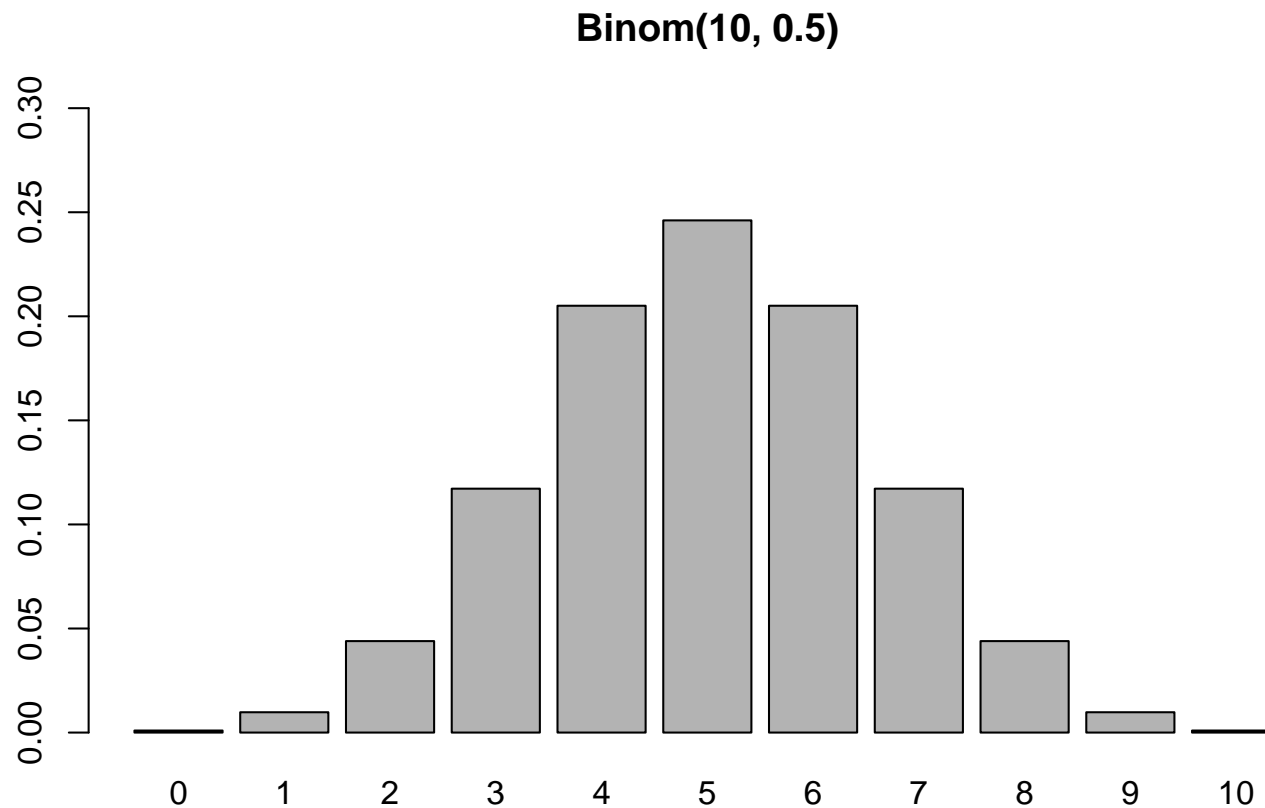
# Hypergeometrische Verteilung

---



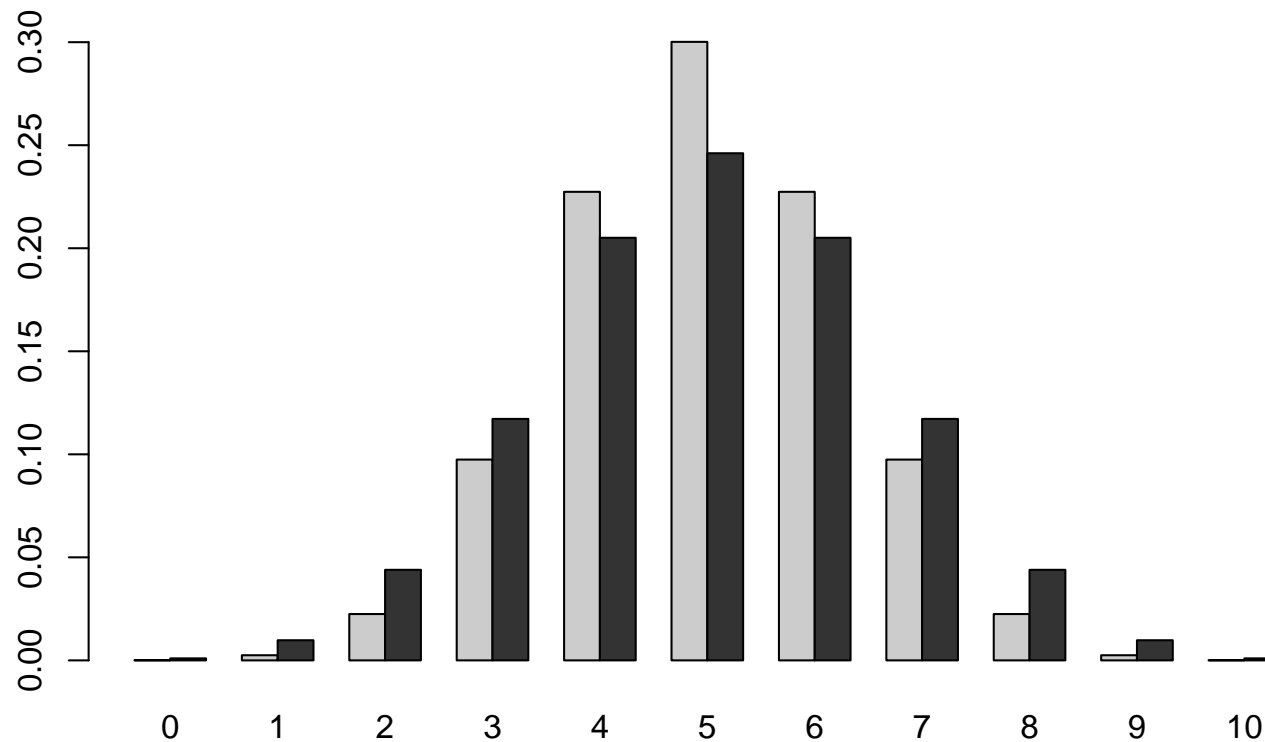
# Hypergeometrische Verteilung

---



# Hypergeometrische Verteilung

---



# Hypergeometrische Verteilung

---

## Beispiel: Aufgabensammlung

**51.** In einer Proseminargruppe mit 25 Studierenden haben 10 ein Beispiel vorbereitet. 3 Studierende werden zufällig ausgewählt, das Beispiel vorzutragen.

Wie groß ist die Wahrscheinlichkeit, daß von diesen drei keiner das Beispiel vorbereitet hat?

# Hypergeometrische Verteilung

---

## Beispiel: Aufgabensammlung

**49.** 12 Architekten und 3 Baumeister reichen Pläne für ein Bauprojekt ein. 5 der Bewerber werden zufällig ausgewählt, ihr Projekt zu präsentieren.

Wie groß ist die Wahrscheinlichkeit, daß mindestens ein Baumeister unter den Ausgewählten ist?

# Hypergeometrische Verteilung

---

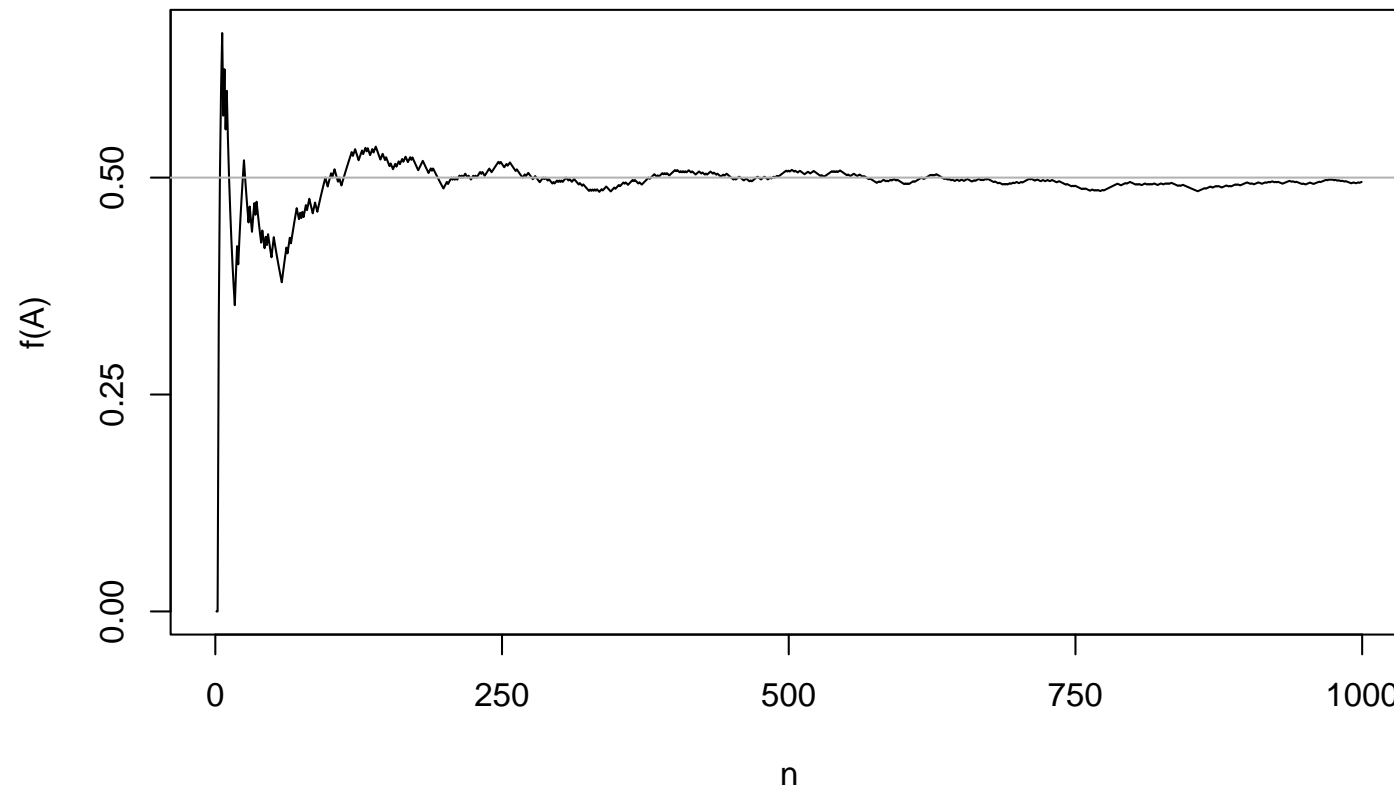
## Beispiel: Aufgabensammlung

**53.** Ein Vertreter für Zeitungsabonnements wählt in einem Wohnhaus zufällig 5 Wohnungen aus und fragt, ob dort ein Abo gewünscht wird. In dem Haus gibt es 10 Wohnungen und es ist bekannt, daß 5 Wohnparteien ein Abo wollen.

Wie groß ist die Wahrscheinlichkeit, daß der Vertreter mehr als 3 Abos verkauft?

# Prognoseintervalle, Zufallsschwankung Münze

---





# Prognoseintervalle

---

Das Ereignis  $A$  tritt mit theoretischer Wahrscheinlichkeit  $p$  ein.

Relative Häufigkeit von  $A$  in der Stichprobe ist  $f_n(A)$

**Prognoseintervall:**

$$P(c \leq f_n(A) \leq d) = \alpha$$

Das Intervall wird mit zunehmender Stichprobengröße  $n$  genauer.

## Prognoseintervalle

---

**Schwierig:** Berechnung von  $c$  und  $d$  aus der exakten Binomialverteilung von  $h_n(A)$  ist.

**Einfacher:** Normalapproximation. Standardisiere  $f_n(A)$  um den Mittelwert  $p$  und mit der Standardabweichung

$$SD = \sqrt{\frac{p(1-p)}{n}}.$$

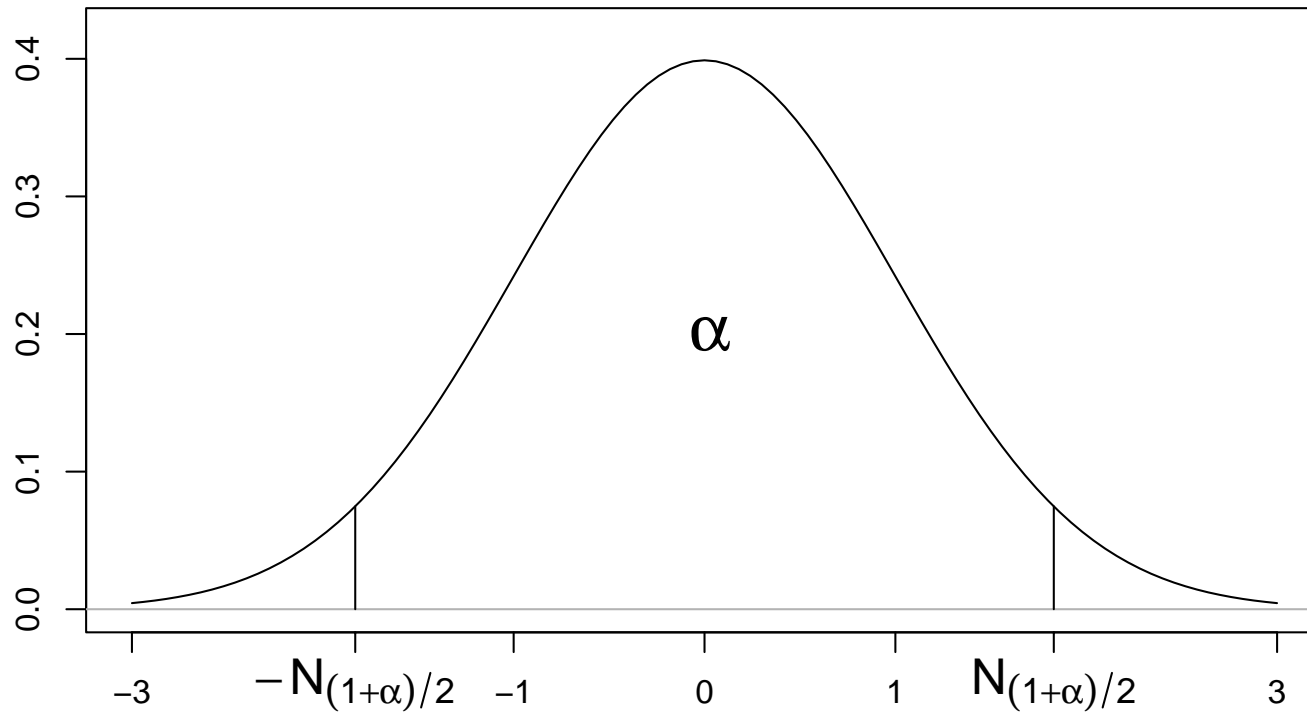
Die standardisierten relativen Häufigkeiten

$$\frac{f_n(A) - p}{SD}$$

sind annähernd standardnormalverteilt.

# Prognoseintervalle

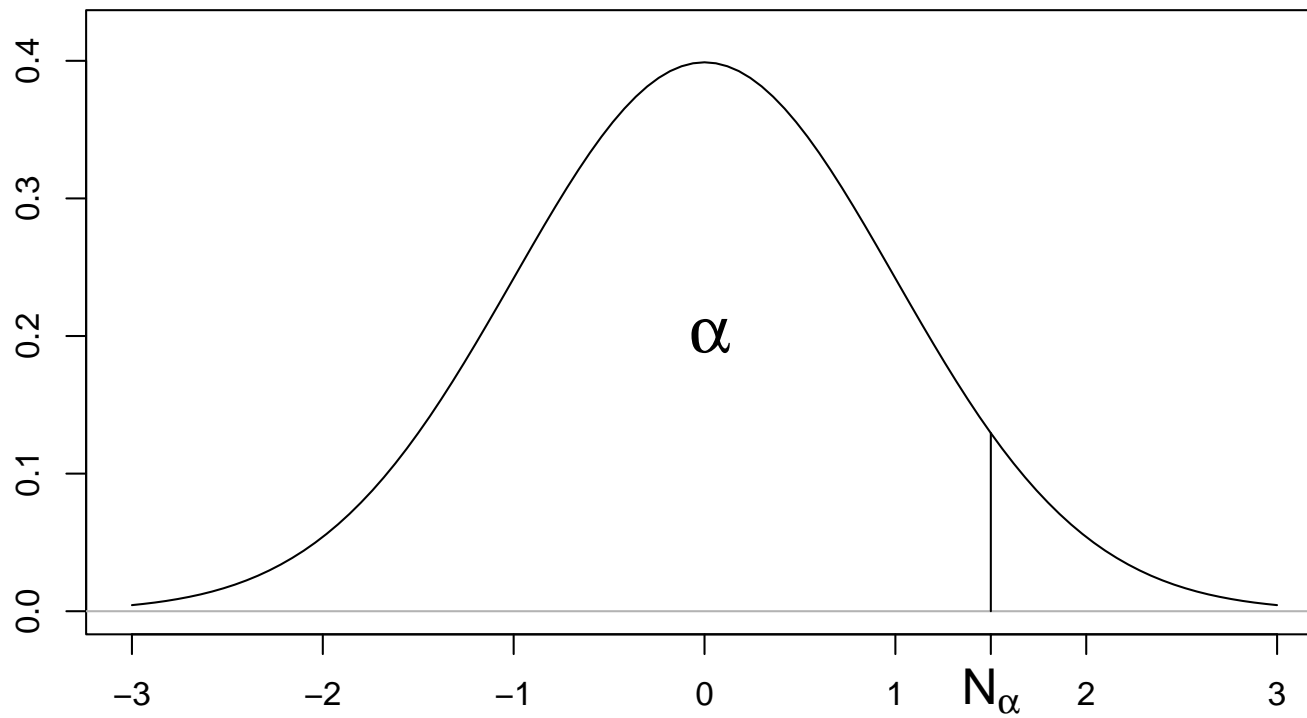
---



$$P \left( -N_{\frac{1+\alpha}{2}} \leq \frac{f_n(A) - p}{SD} \leq N_{\frac{1+\alpha}{2}} \right) \approx \alpha$$

# Prognoseintervalle

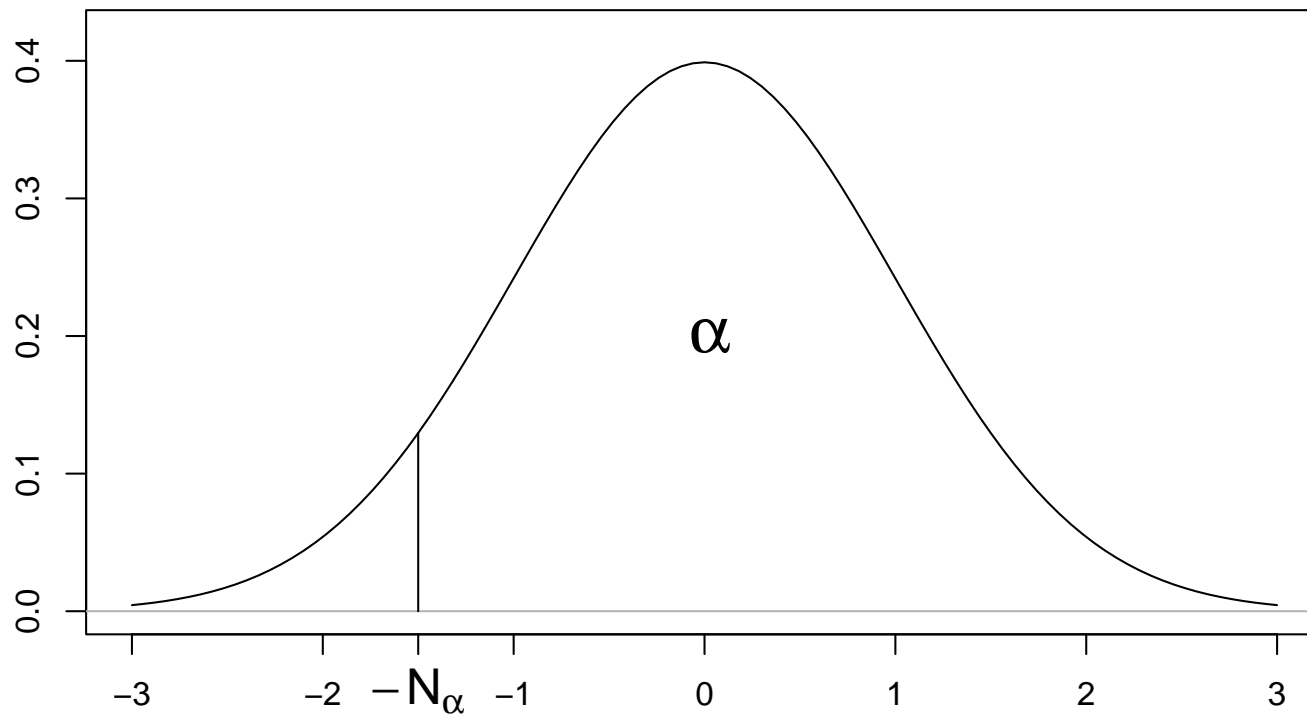
---



$$P\left(\frac{f_n(A) - p}{SD} \leq N_\alpha\right) \approx \alpha$$

# Prognoseintervalle

---



$$P\left(\frac{f_n(A) - p}{SD} \geq -N_\alpha\right) \approx \alpha$$

# Prognoseintervalle

---

einseitig:

$$P(f_n(A) \leq p + N_\alpha SD) \approx \alpha$$

$$P(f_n(A) \geq p - N_\alpha SD) \approx \alpha$$

beidseitig:

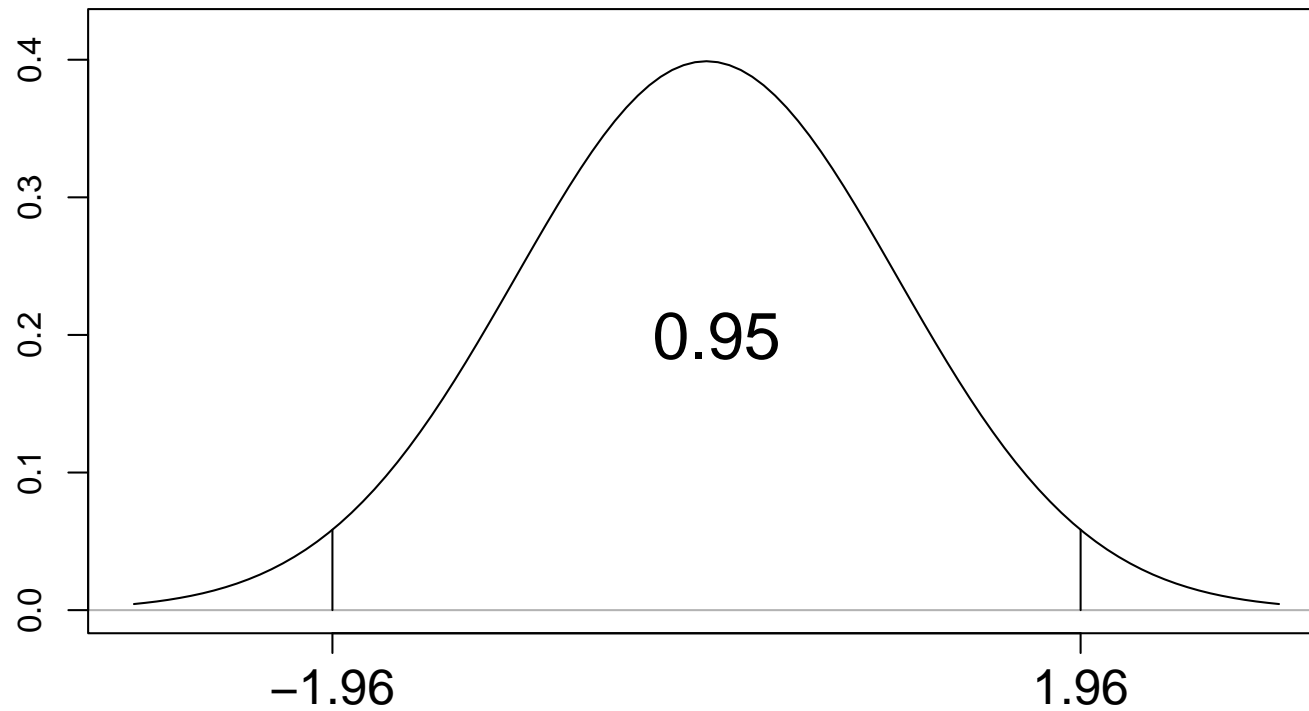
$$P\left(p - N_{\frac{1+\alpha}{2}} SD \leq f_n(A) \leq p + N_{\frac{1+\alpha}{2}} SD\right) \approx \alpha$$

mit

$$SD = \sqrt{\frac{p(1-p)}{n}}$$

# Prognoseintervalle

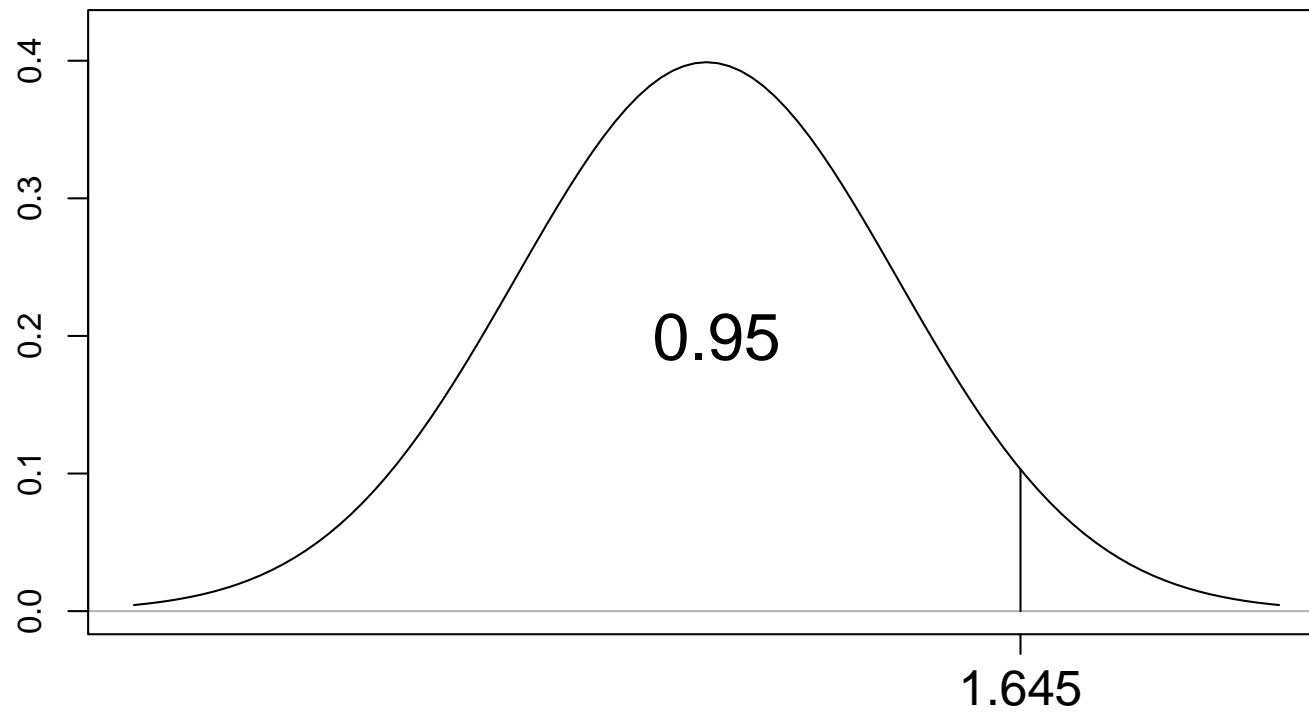
---



$$P(-1.96 SD \leq f_n(A) - p \leq 1.96 SD) \approx 0.95$$

# Prognoseintervalle

---

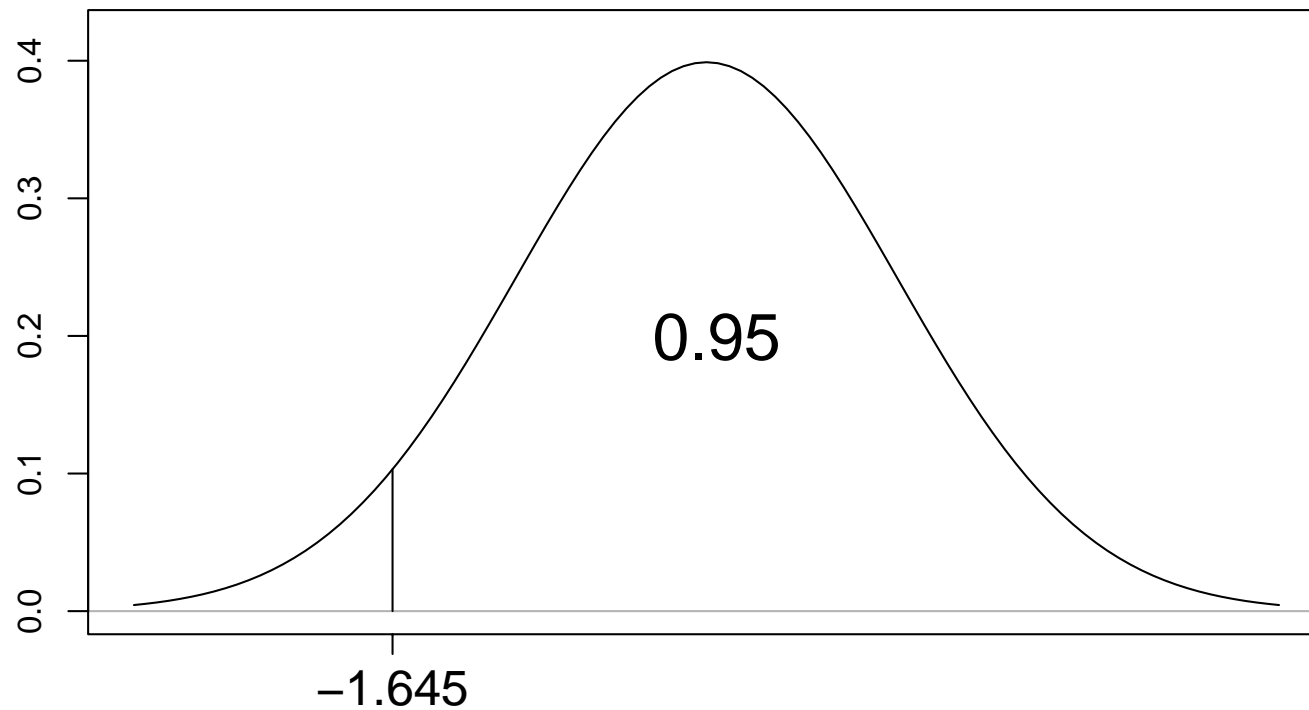


$$P(f_n(A) - p \leq 1.64 SD) \approx 0.95$$



# Prognoseintervalle

---



$$P(f_n(A) - p \geq -1.64 SD) \approx 0.95$$

# Prognoseintervalle

---

**einseitig**       $P(f_n(A) \leq p + 1.64 SD) \approx 0.95$

$$P(f_n(A) \geq p - 1.64 SD) \approx 0.95$$

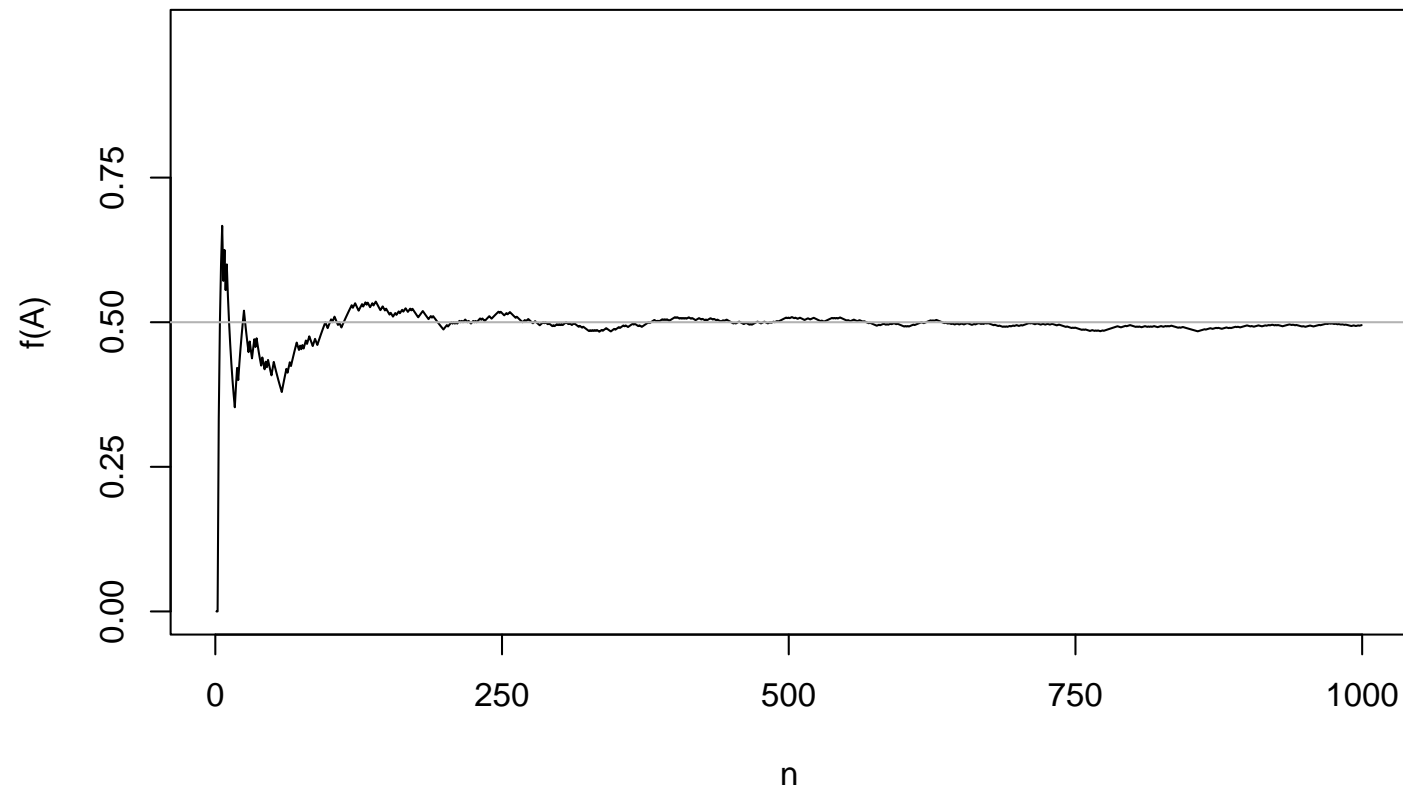
**beidseitig**       $P(|f_n(A) - p| \leq 1.96 SD) \approx 0.95$

**Faustregel**       $P(|f_n(A) - p| \leq 2 SD) \approx 0.95$

$$SD = \sqrt{\frac{p(1-p)}{n}}$$

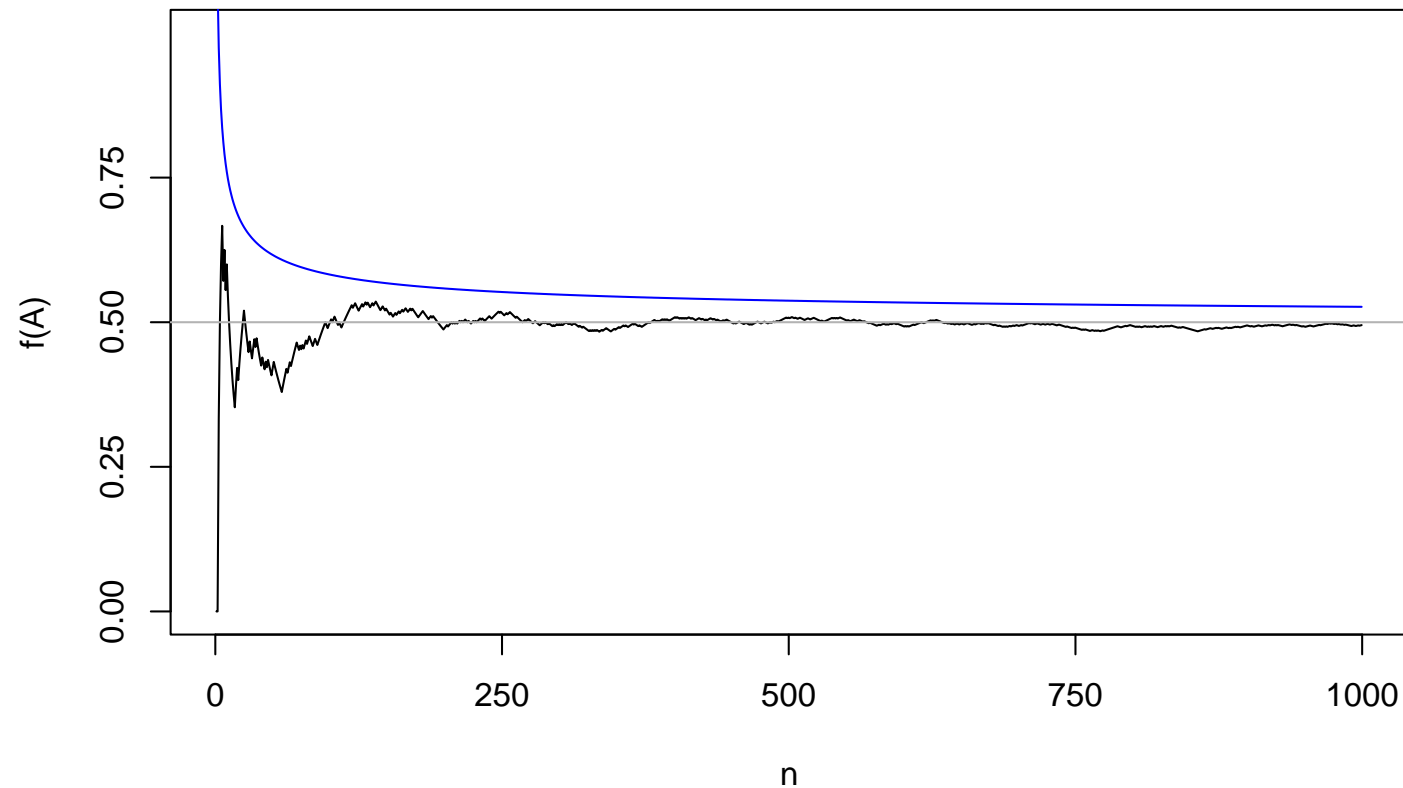
# Prognoseintervalle

---



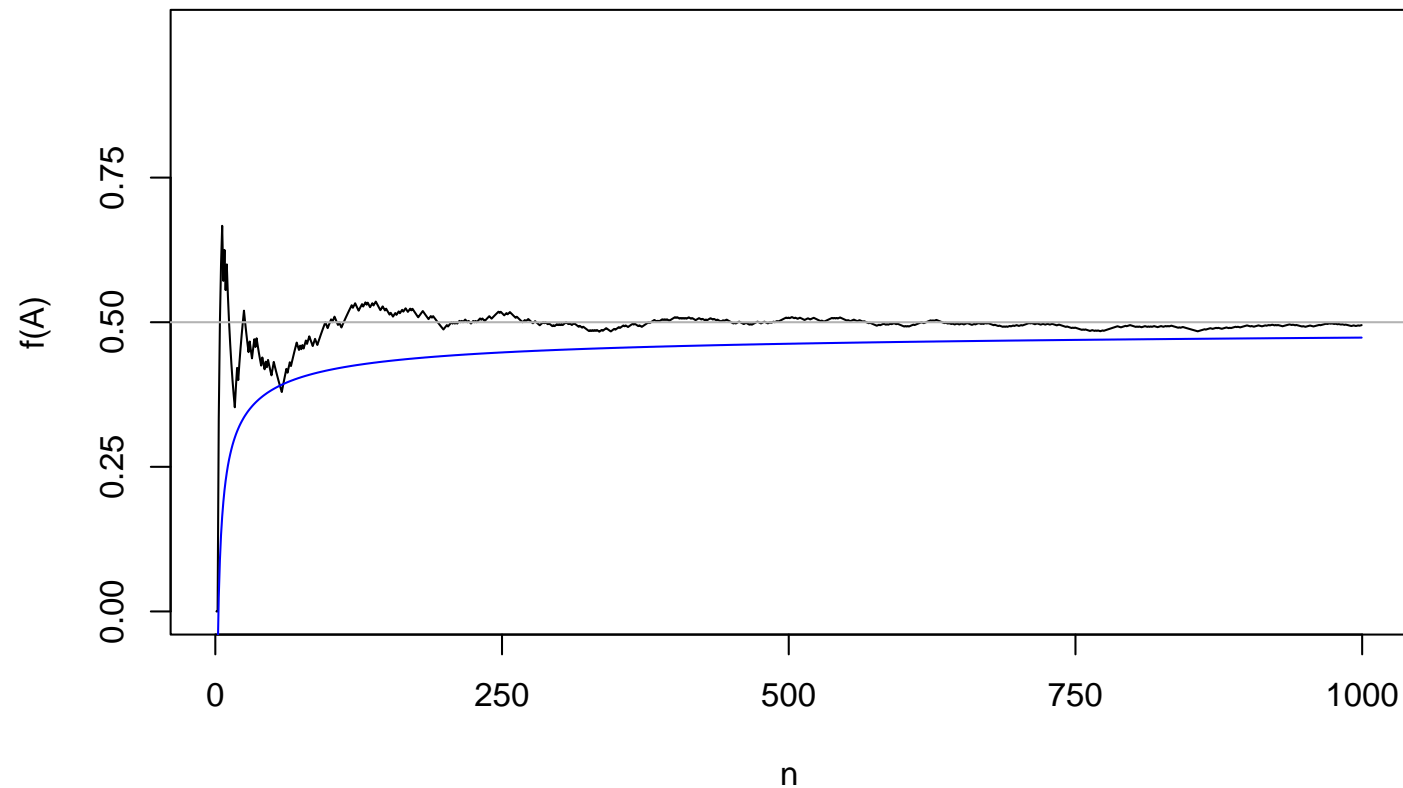
# Prognoseintervalle

---



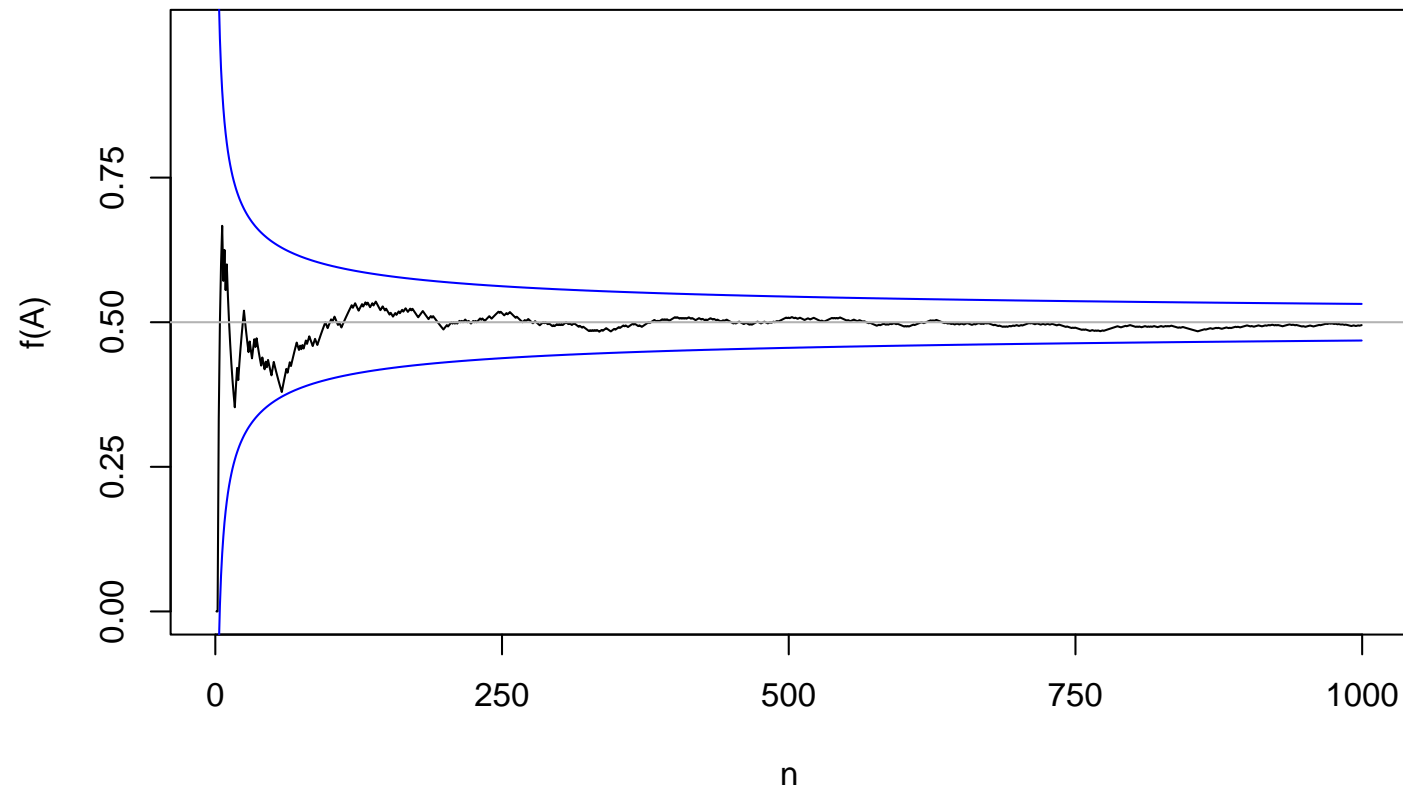
# Prognoseintervalle

---



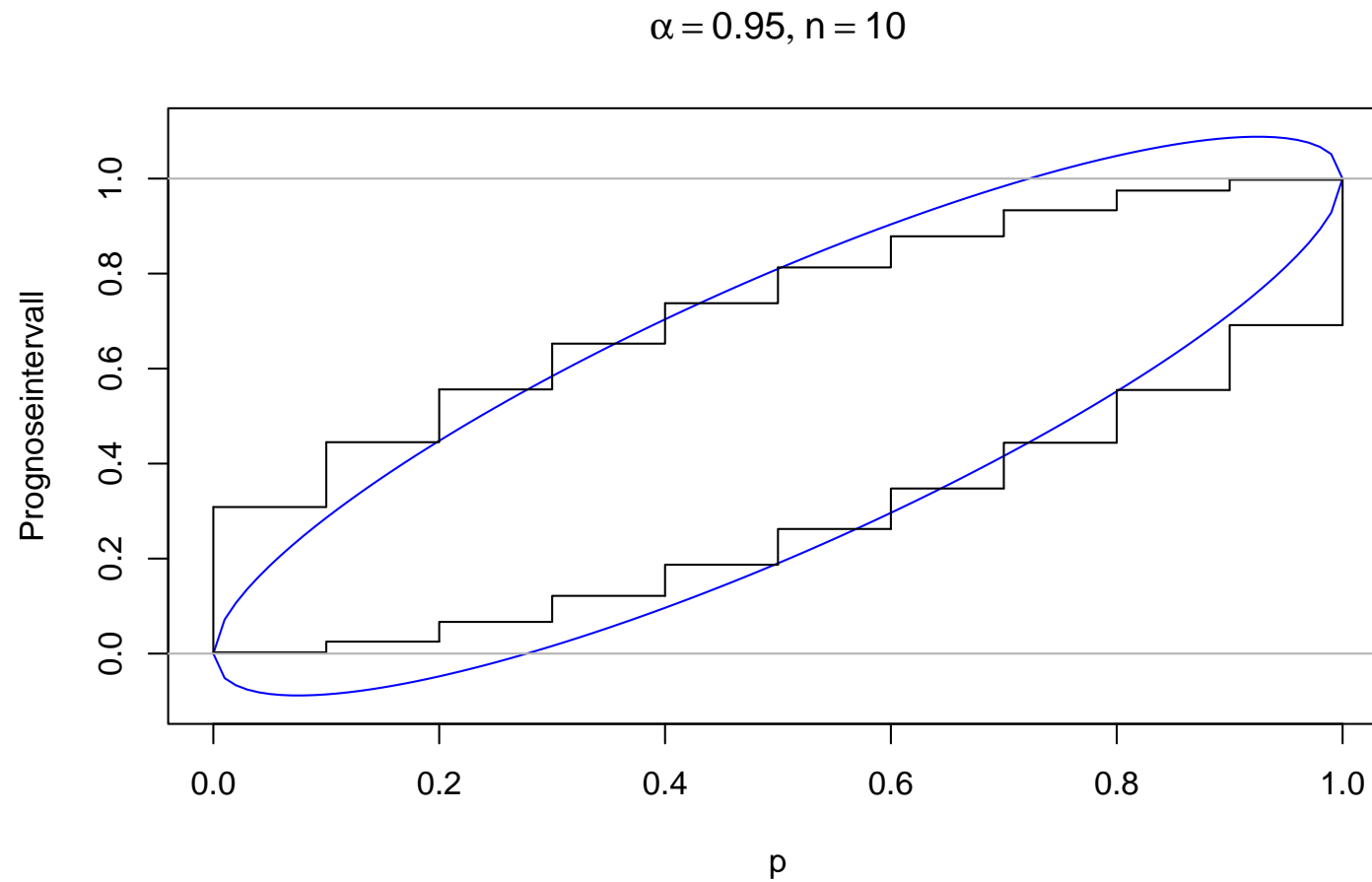
# Prognoseintervalle

---



# Prognoseintervalle

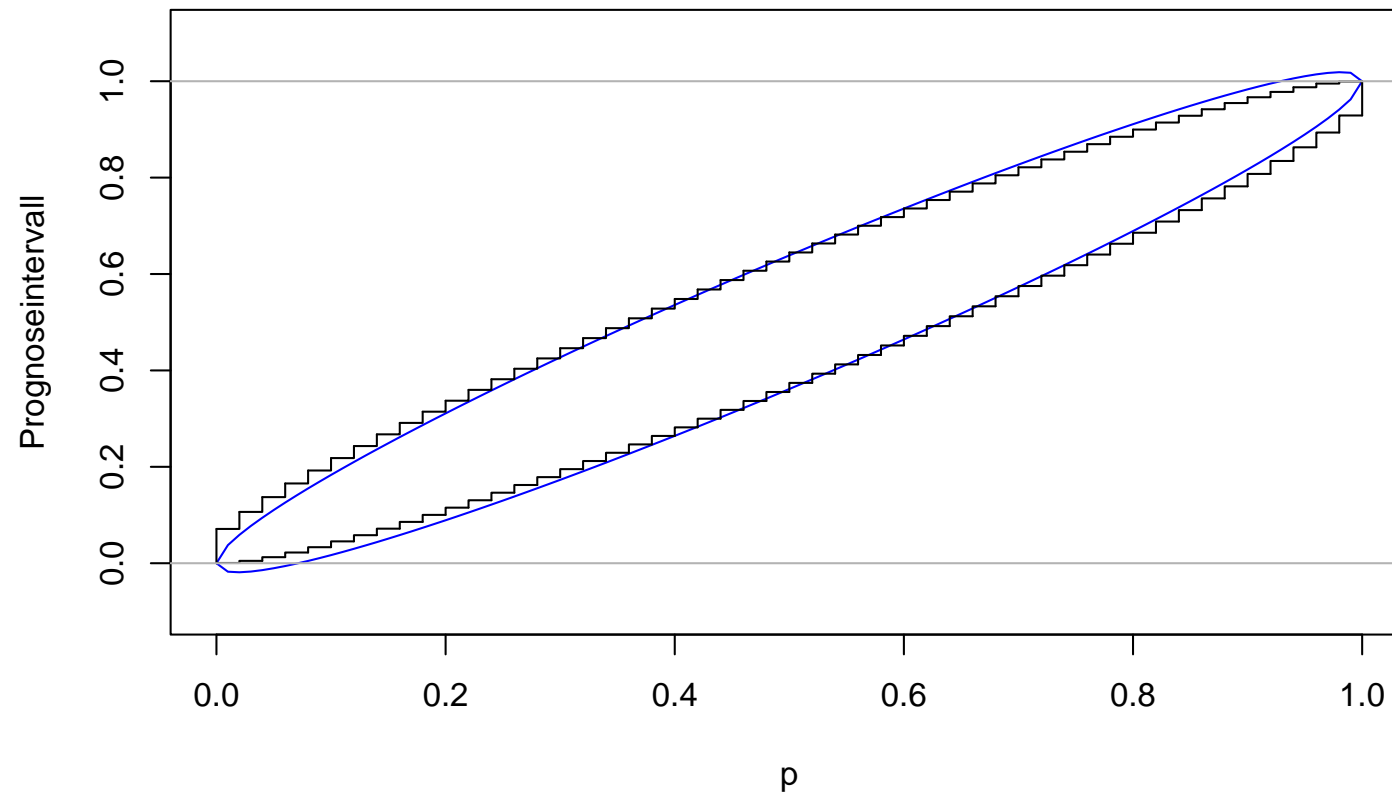
---



# Prognoseintervalle

---

$\alpha = 0.95, n = 50$

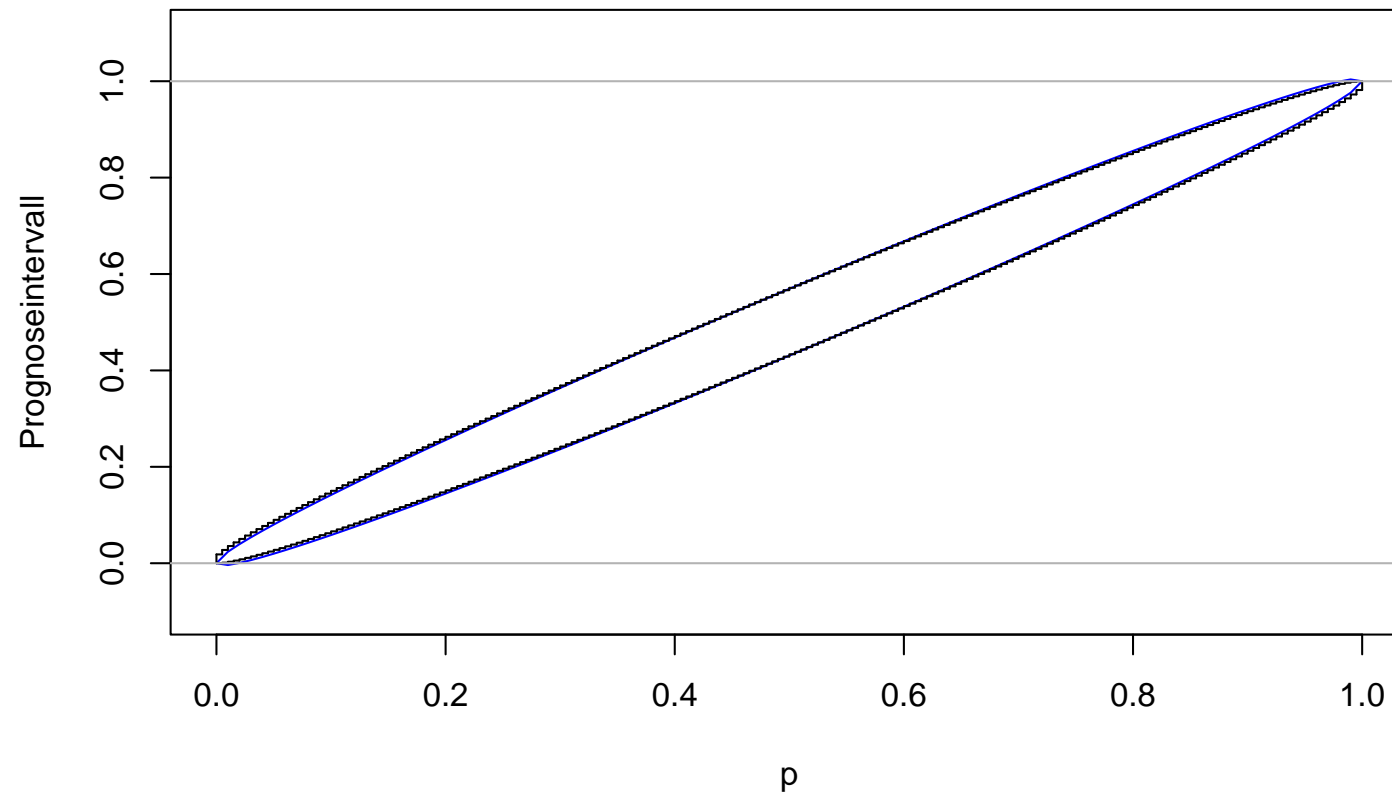




# Prognoseintervalle

---

$\alpha = 0.95, n = 200$



# Prognoseintervalle

---

## Beispiel: Skript

### (3.42) Fertigungskontrolle

Die von einer Maschine produzierten Werkstücke sind mit Wahrscheinlichkeit 0.8 tauglich. Die Werkstücke werden in Paketen zu je 100 Stück ausgeliefert. Für Garantiezusagen möchte man wissen, wie stark der Anteil an tauglichen Werkstücken in einem Paket schwanken kann (mit Sicherheit 95%).

# Prognoseintervalle

---

## Beispiel: Skript

### (3.43) Meinungsumfrage

Eine politische Partei geht davon aus, daß sie 40% Wähleranteil hat. Bei einer Meinungsumfrage mit 1000 Befragten geben 367 Personen an, Wähler dieser Partei zu sein.

Kann dieses Umfrageergebnis als zufällige Schwankung um die relative Häufigkeit 0.4 interpretiert werden, oder muss die Einschätzung der Partei revidiert werden?

# Prognoseintervalle und Stichprobengröße

---

## Beispiel: Aufgabensammlung

**60.** 30% aller Angestellten planen, in nächster Zeit einen PC zu erwerben. In einer Gemeinde mit 850 Angestellten soll ein EDV-Fachgeschäft eröffnet werden.

- Geben sie ein 95%-Prognoseintervall für den Anteil an potenziellen Käufern an.
- Ab welcher Gemeindegröße kann der Anteil mit einer Genauigkeit von  $\pm 2\%$  geschätzt werden?

# Prognoseintervalle und Stichprobengröße

---

## Beispiel: Aufgabensammlung

**59.** 19% aller Erwachsenen möchte abnehmen. In einer Firma mit 85 erwachsenen Mitarbeitern soll in der Kantine ein spezielles Menü zum Abnehmen angeboten werden.

- Der Anteil der potenziellen Interessenten an diesem Menü soll mit 95%-iger Sicherheit angegeben werden.
- Ab welcher Firmengröße könnte der Anteil der Personen, die abnehmen wollen mit einer Genauigkeit von  $\pm 5\%$  angegeben werden?

# Prognoseintervalle und Stichprobengröße

---

## Beispiel: Aufgabensammlung

**58.** Ein Fernsehteam will mittels Umfrage feststellen, ob die Mehrheit der Bevölkerung für eine bestimmte umweltpolitische Maßnahme ist. Es werden 50 Personen zufällig befragt. In Wirklichkeit sind nur 44% der Bevölkerung für die Maßnahme.

- Kann die Umfrage trotzdem ergeben, daß eine Mehrheit der Bevölkerung für die Maßnahme ist?
- Wie groß muss die Stichprobe sein, damit diese Maßnahme mit hoher Sicherheit (99%) keine Mehrheit findet?

# Prognoseintervalle und Stichprobengröße

---

## Beispiel: Aufgabensammlung

**57.** Laut Angabe des Versicherungsverbandes besitzen von 2.9 Millionen Haushalten in Österreich 2.7 Millionen eine Haushaltsversicherung. Ein Versicherungsvertreter ruft zufällig 70 Haushalte an.

- Geben sie an, wieviele Haushalte **ohne** Haushaltsversicherung mindestens unter den angerufenen Haushalten zu erwarten sind.
- Wieviele Haushalte müsste er anrufen, damit dieser Anteil an angerufenen Haushalten **ohne** Haushaltsversicherung mindestens 3% beträgt?

# Versuchspläne

---

**Zufallsexperiment** beliebig oft wiederholbar  $\rightarrow P(A)$  verändert sich nicht

**Ziehung mit Zurücklegen**  $\rightarrow P(A)$  verändert sich nicht

**Ziehung ohne Zurücklegen**  $\rightarrow P(A)$  verändert sich  
daher wird die Endlichkeitskorrektur verwendet, die das Prognoseintervall verkleinert

$$SD = \sqrt{\frac{p(1-p)}{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$



# Zusammenfassung Kapitel 3

---

- Erhebungsformen: Totalerhebung, Repräsentative Erhebung
- Wahrscheinlichkeit
- Gesetz der großen Zahlen
- Binomialverteilung
- Hypergeometrische Verteilung
- Prognose von Häufigkeiten: Intervalle, Stichprobengröße

# Testen und Schätzen

## Kapitel 4

# Induktive Statistik

---

## Grundsätzliches Problem:

Wie kann man aus der Beobachtung von Häufigkeiten in (endlichen) empirischen Stichproben Rückschlüsse über die zugrundeliegenden Wahrscheinlichkeiten ziehen?

## Fragestellungen:

- Testen von Hypothesen
- Schätzen von Parametern

# Testen von Hypothesen

---

## Beispiel: Skript

### (4.4) Mensabesucher

- Ziel einer Mensa ist es, daß Studenten und Studentinnen die Mensa gleich häufig besuchen.
- An einem zufällig ausgewählten Tag waren unter 425 Mensabesuchern 170 Studentinnen.
- Frage: Kann aus diesen Daten geschlossen werden, daß die Mensa ihr Ziel nicht erreicht hat?

# Testen von Hypothesen

---

## Formal:

Sei  $p$  der Anteil der Studentinnen.

Kann nachgewiesen werden, daß  $p \neq 0.5$ ?

Damit stellt sich die Frage:

Ist die relative Häufigkeit von  $\hat{p} = f_{425}(\text{Studentin}) = 0.4$  aus einer Stichprobe vom Umfang 425 unter einer (hypothetisch unterstellten) Wahrscheinlichkeit  $p_0 = 0.5$  erklärbar oder sehr unwahrscheinlich?

# Testen von Hypothesen

---

Ziel des Testens ist nicht die Schätzung einer Wahrscheinlichkeit  $p$  sondern eine Entscheidung zwischen zwei Alternativen über  $p$ .

## Zweiseitiges Testproblem:

Nullhypothese:  $p = p_0$

Alternative:  $p \neq p_0$

# Testen von Hypothesen

---

**Beispiel:** ähnlich (4.2)

- In einer Stadt ziehen 60% der Konsumenten Produkt A dem Produkt B vor.
- Nach Webekampagne für B: 25 von 50 befragten Konsumenten erklären, daß sie B vorziehen.
- Frage: War die Werbekampagne erfolgreich?

# Testen von Hypothesen

---

## Formal:

Sei  $p$  der Anteil der B-Konsumenten.

Kann nachgewiesen werden, daß  $p > 0.4$ ?

Damit stellt sich die Frage:

Ist die relative Häufigkeit von  $\hat{p} = f_{50}(\text{B vorgezogen}) = 0.5$  aus einer Stichprobe vom Umfang 50 unter einer (hypothetisch unterstellten) Wahrscheinlichkeit  $p_0 = 0.4$  erklärbar oder sehr unwahrscheinlich?



# Testen von Hypothesen

---

## Linksseitiges Testproblem:

Nullhypothese:  $p \leq p_0$

Alternative:  $p > p_0$

# Testen von Hypothesen

---

## Rechtsseitiges Testproblem:

Nullhypothese:  $p \geq p_0$

Alternative:  $p < p_0$

# Durchführung von Tests

---

1. Gültigkeit der Nullhypothese wird unterstellt.
2. Berechnung von Prognoseintervall für Daten.
3. Liegen die tatsächlich beobachteten Daten im Prognoseintervall?
  - (a) Ja. Die empirischen Daten widersprechen nicht den Annahmen  
⇒ **nicht-signifikantes** Resultat und **Beibehaltung der Nullhypothese**.
  - (b) Nein. Die empirischen Daten widersprechen den Annahmen  
⇒ **signifikantes** Ergebnis und **Verwerfen der Nullhypothese**.

# Durchführung von Tests

---

## Beispiel: (4.4)

Für die Wahrscheinlichkeit  $p$ , mit der Studentinnen die Mensa besuchen, wird  $p = p_0 = 0.5$  angenommen.

Die Standardabweichung beträgt dann

$$SD = \sqrt{\frac{p_0(1 - p_0)}{n}} = \sqrt{\frac{0.5 \cdot 0.5}{425}} \approx 0.0243.$$

Wir benutzen die Sicherheitswahrscheinlichkeit  $\alpha = 0.95$  und die Normalapproximation, damit ist  $N_{(1+\alpha)/2} = N_{0.975} = 1.96$ .

## Durchführung von Tests

---

Das Prognoseintervall für die relative Häufigkeit  $f_{425}(\text{Studentin})$  ist damit

$$\begin{aligned}[p_0 \pm N_{(1+\alpha)/2} \cdot SD] &= [0.5 \pm 1.96 \cdot 0.0243] \\ &= [0.452, 0.548]\end{aligned}$$

Die empirische Häufigkeit  $f_{425}(\text{Studentin}) = \hat{p} = 170/425 = 0.4$  liegt außerhalb dieses Intervalls und steht damit im Widerspruch zur Annahme die Wahrscheinlichkeit sei  $p = 0.5$ .

Also muss die Nullhypothese **verworfen** werden: die Mensa hat ihr Ziel nicht erreicht.

## Durchführung von Tests

---

Völlig analog kann man das Prognoseintervall für die absolute Häufigkeit  $h_{425}(\text{Studentin})$  berechnet werden:

$$[0.452, 0.548] \cdot 425 = [192.1, 232.9].$$

Aber  $h_{425}(\text{Studentin}) = 170$  liegt außerhalb dieses Intervalls.

# Durchführung von Tests

---

**Beispiel:** ähnlich (4.2)

Die Hypothese lautet  $H : p \leq 0.4$ ,  $p_0 = 0.4$ .

Die Standardabweichung beträgt dann

$$SD = \sqrt{\frac{p_0(1 - p_0)}{n}} = \sqrt{\frac{0.4 \cdot 0.6}{50}} \approx 0.0693.$$

Wir benutzen die Sicherheitswahrscheinlichkeit  $\alpha = 0.95$  und die Normalapproximation, damit ist  $N_\alpha = N_{0.95} = 1.645$ .

## Durchführung von Tests

---

Das einseitige Prognoseintervall für die relative Häufigkeit  $f_{50}(\text{B vorgezogen})$  ist damit

$$\begin{aligned}[0, p_0 + N_\alpha \cdot SD] &= [0, 0.4 + 1.645 \cdot 0.0693] \\ &= [0, 0.514]\end{aligned}$$

Die empirische Häufigkeit  $f_{50}(\text{B vorgezogen}) = \hat{p} = 25/50 = 0.5$  liegt innerhalb dieses Intervalls und steht **nicht** im Widerspruch zur Annahme, die Wahrscheinlichkeit sei  $p \leq 0.4$ .

Also kann die Nullhypothese **nicht verworfen** werden: es kann nicht ausgeschlossen werden, daß die Werbekampagne nutzlos war.



# Durchführung von Tests

---

Die Testprozedur verwirft die Nullhypothese, wenn bei  $p = p_0$

$$\begin{aligned} \hat{p} & \begin{matrix} > \\ < \end{matrix} p_0 \pm c \cdot SD \\ \frac{\hat{p} - p_0}{SD} & \begin{matrix} > \\ < \end{matrix} \pm c \end{aligned}$$

Den Standardscore der relativen Häufigkeit

$$T = \frac{\hat{p} - p_0}{SD}$$

nennt man **Testgröße** oder **Teststatistik**.

Den Wert  $c = c(\alpha)$  nennt man **kritischen Wert**.

# Durchführung von Tests

---

1. Berechne **Teststatistik**  $T$  aus Daten
2. Lege **Annahmebereich** fest, in den die Teststatistik mit hoher Wahrscheinlichkeit (= Signifikanzniveau  $\alpha$ ) fällt, falls die Nullhypothese wahr ist.
3. Die Grenzen des Annahmebereiches heißen **kritische Werte**. Überschreitet  $T$  einen kritischen Wert, so liegt ein **signifikantes** Ergebnis vor: die Nullhypothese wird **verworfen**.

# Durchführung von Tests

---

## Kritische Werte für die Testgrößen für 95% Niveau:

**zweiseitig**      $-2 \leq T \leq 2 \Rightarrow H: p = p_0$  wird beibehalten  
                     $T < -2 \Rightarrow p$  ist signifikant kleiner als  $p_0$   
                     $T > +2 \Rightarrow p$  ist signifikant größer als  $p_0$

**linksseitig**      $T \leq 1.64 \Rightarrow H: p \leq p_0$  wird beibehalten  
                     $T > 1.64 \Rightarrow p$  ist signifikant größer als  $p_0$

**rechtsseitig**    $T \geq -1.64 \Rightarrow H: p \geq p_0$  wird beibehalten  
                     $T < -1.64 \Rightarrow p$  ist signifikant kleiner als  $p_0$

# Fehlerarten

---

	H wird nicht verworfen	H wird verworfen
H trifft zu	Entscheidung richtig	Fehler 1. Art
H trifft nicht zu	Fehler 2. Art	Entscheidung richtig

# Fehlerarten

---

## **Signifikanzniveau:**

Die Sicherheit, mit der sich der Fehler 1. Art vermeiden läßt.

Das Signifikanzniveau wird beim Testen kontrolliert  $\Rightarrow$  Fehler 1. Art sind selten!

Das Verwerfen der Nullhypothese ist ein statistischer Beweis dafür, daß sie falsch ist.

# Fehlerarten

---

## Trennschärfe:

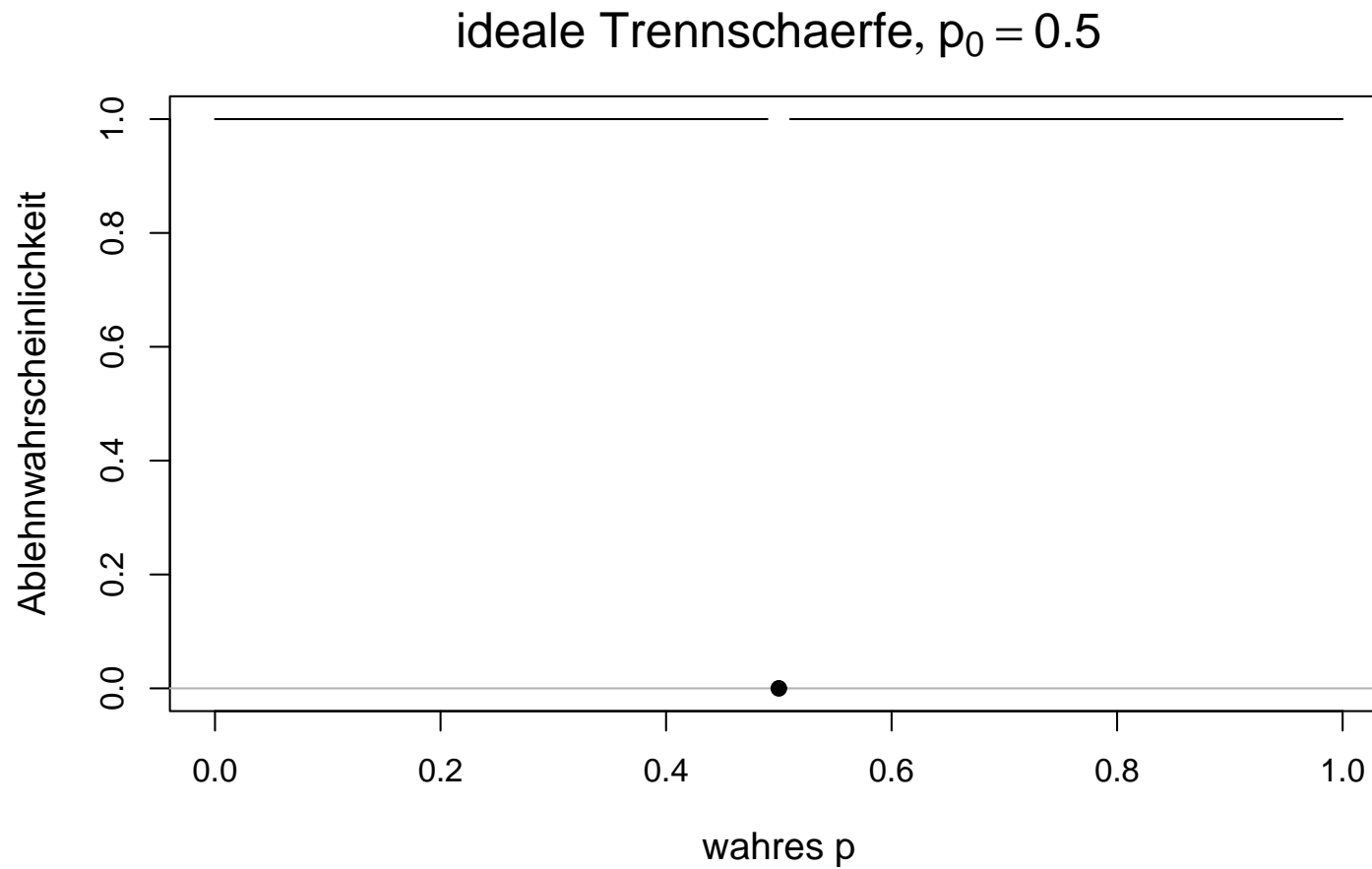
Die Sicherheit, mit der sich der Fehler 2. Art vermeiden läßt.

Die Trennschärfe hängt ab vom Signifikanzniveau  $\alpha$ , Stichprobenumfang  $n$  und der **wahren** Wahrscheinlichkeit  $p \Rightarrow$  kann in der Praxis nicht kontrolliert werden!

Das Beibehalten der Nullhypothese ist **kein** statistischer Beweis dafür, daß sie richtig ist.

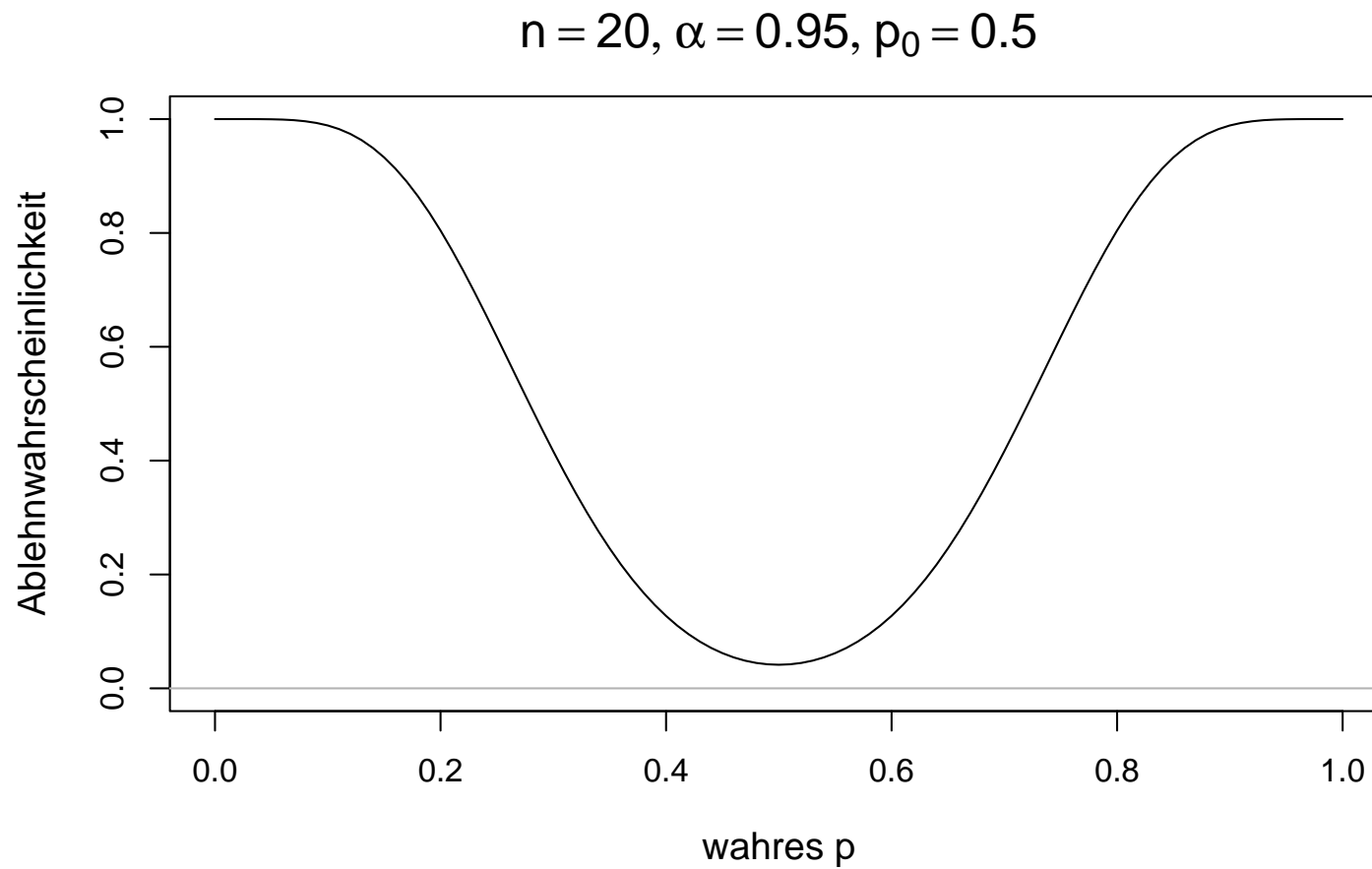
# Gütefunktion

---



# Gütefunktion

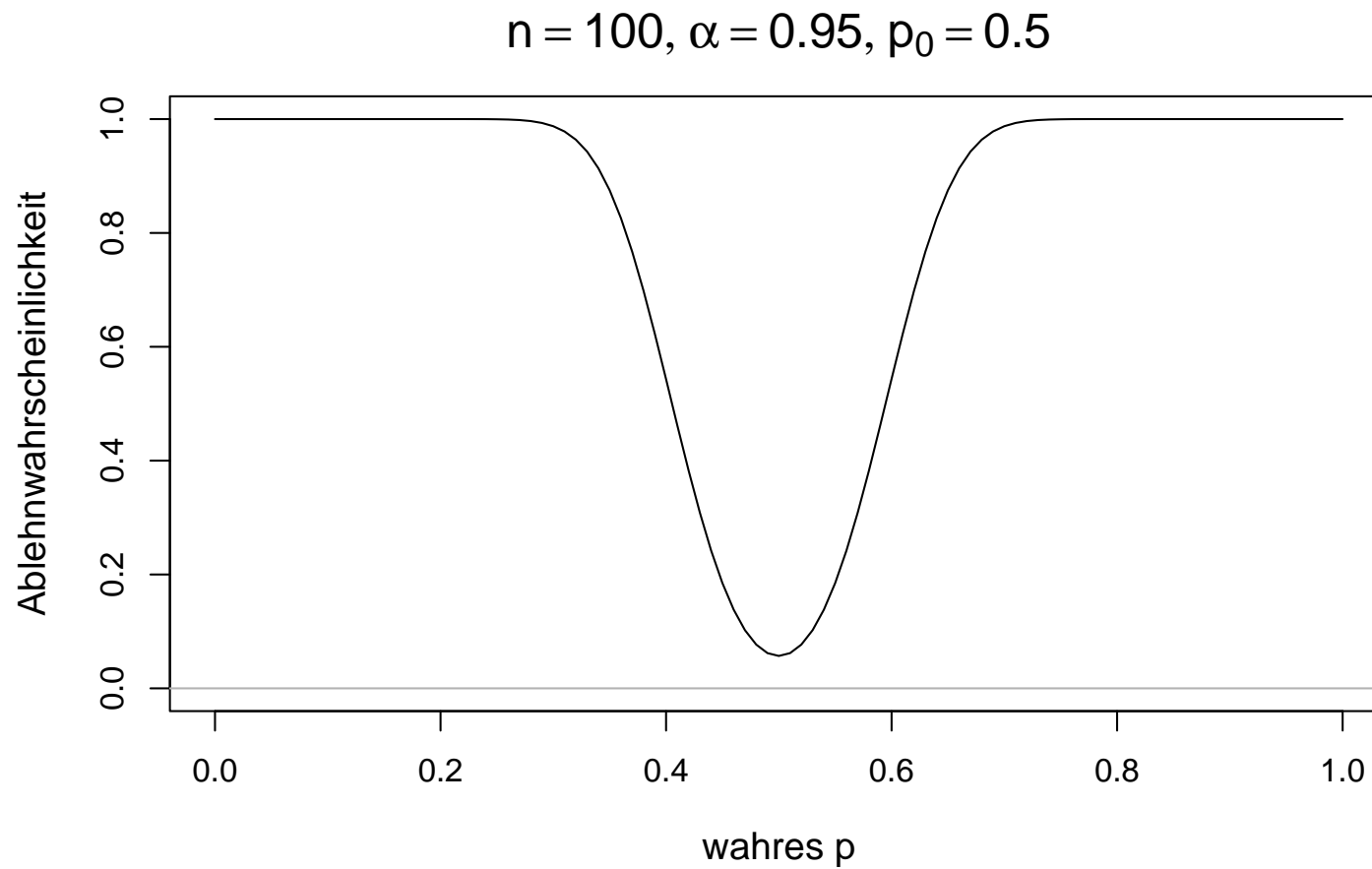
---





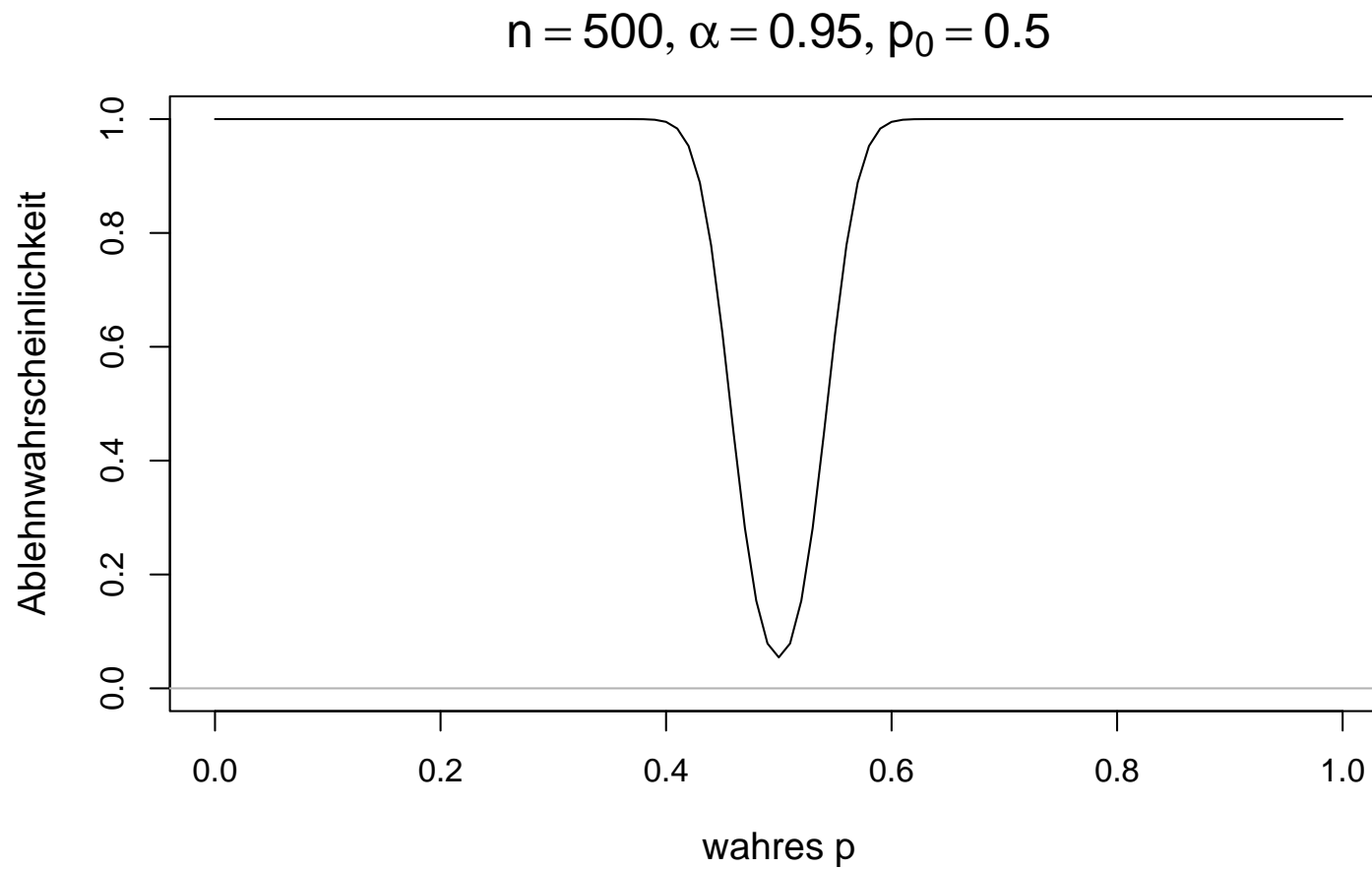
# Gütefunktion

---



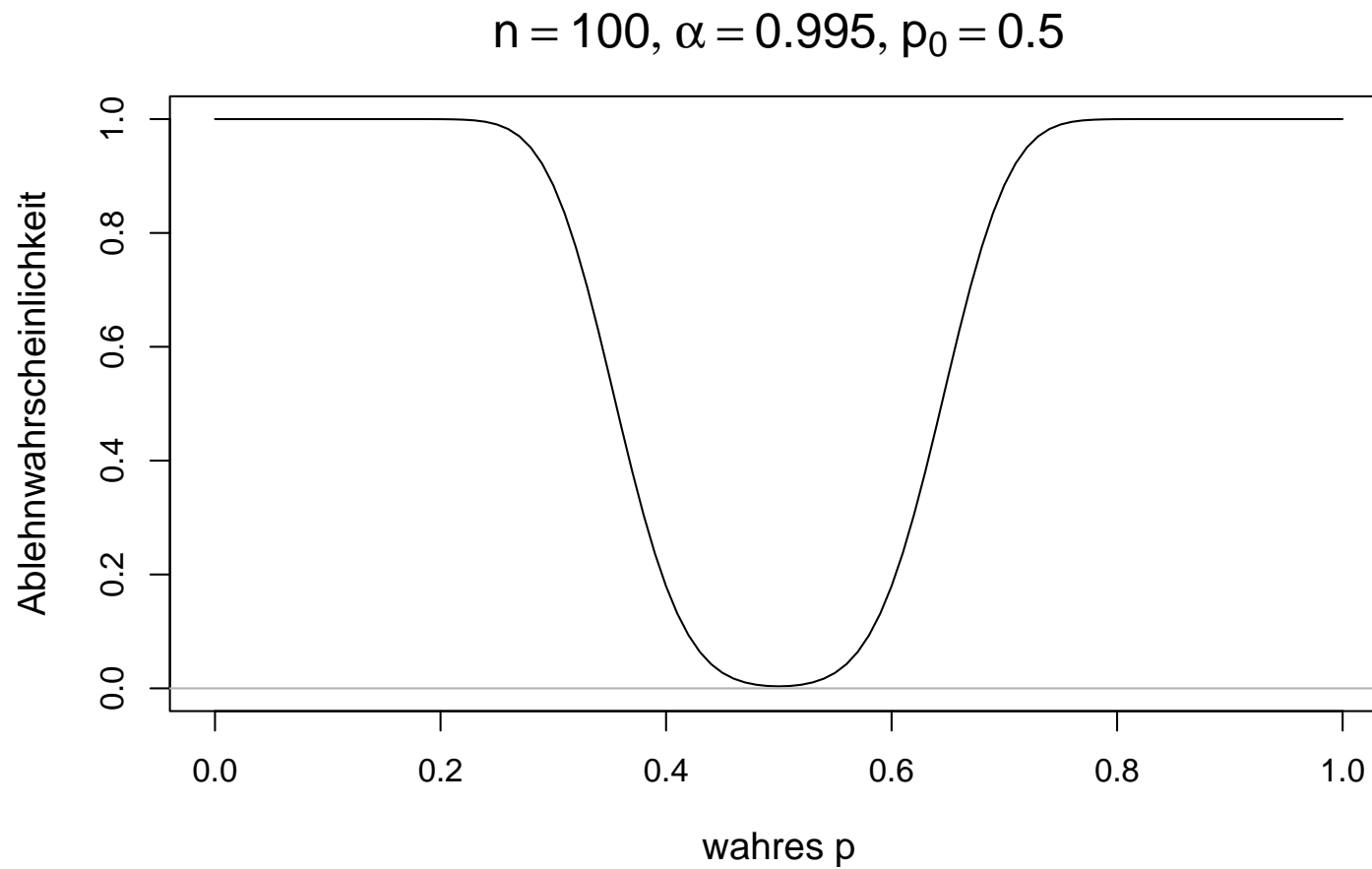
# Gütefunktion

---



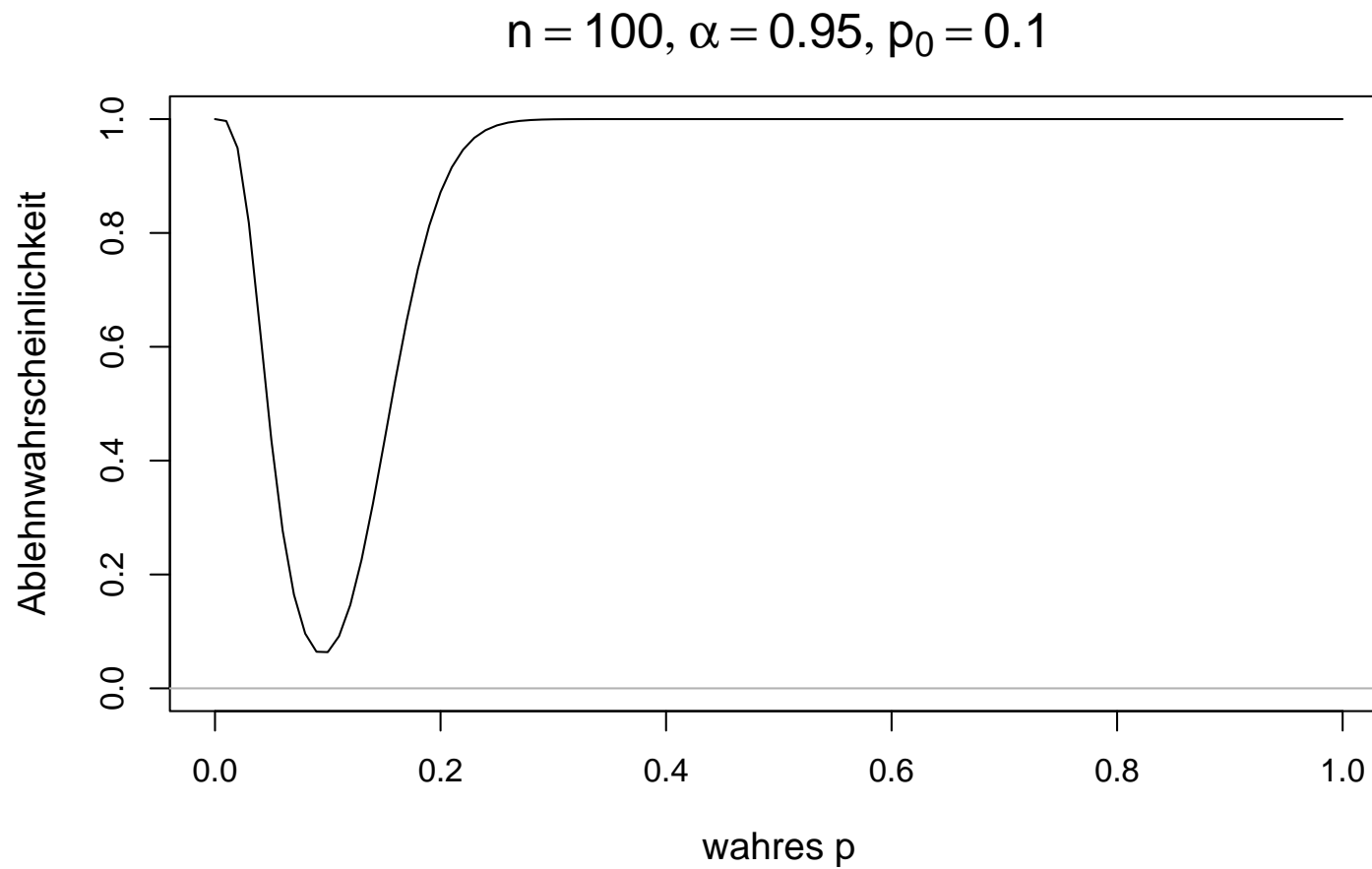
# Gütefunktion

---



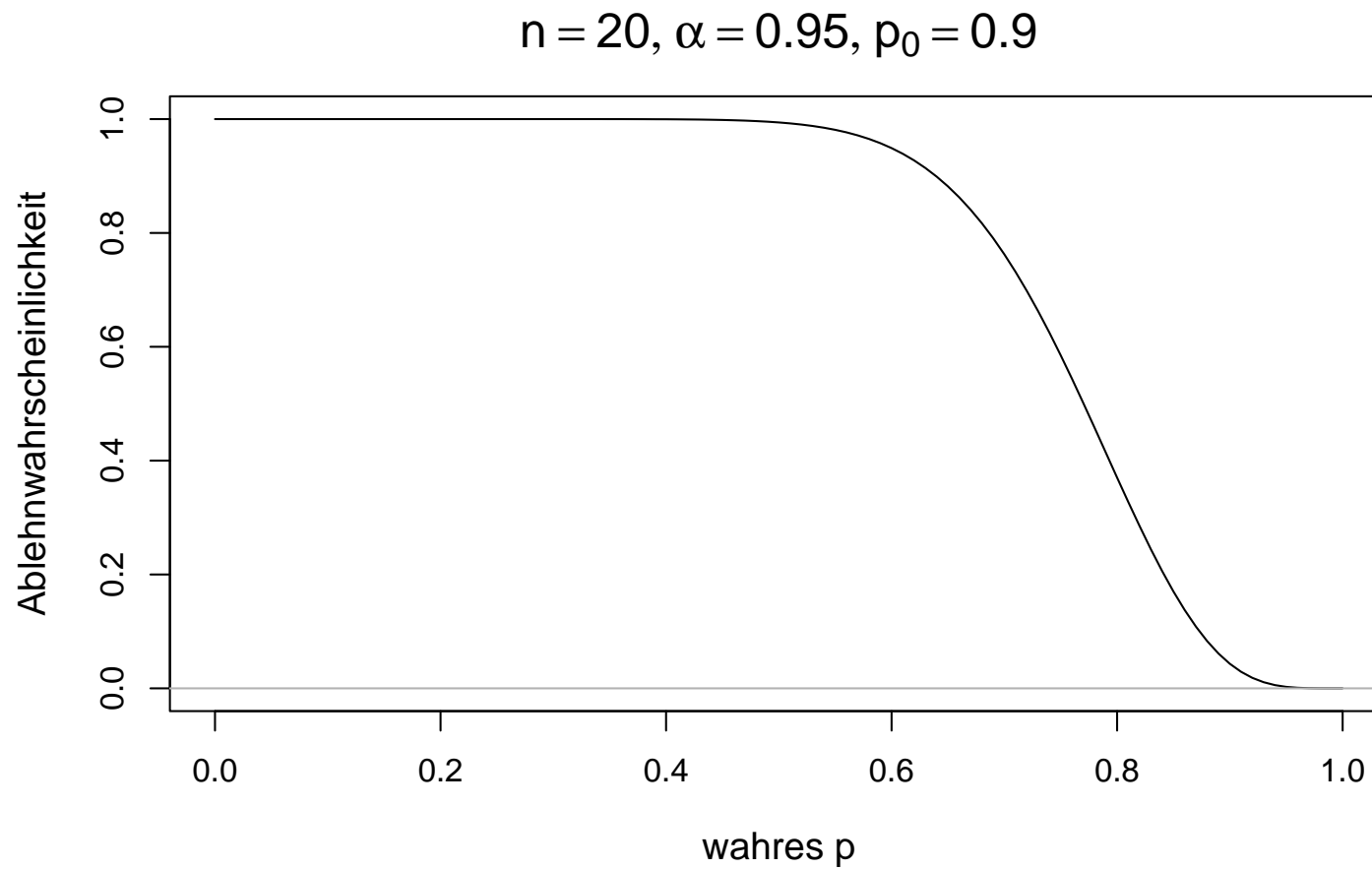
# Gütefunktion

---



# Gütefunktion

---



# Testen

---

## Beispiel: Aufgabensammlung

### 61. Meinungsumfrage: Beurteilung einer Pensionsreform:

	sehr positiv	positiv	negativ	sehr negativ
Ärzte	1	3	51	10
Rechtsanwälte	5	25	13	12
Krankenpfleger	11	21	20	16
Verkäufer	28	15	11	1

Für welche Berufsgruppe kann mittels eines statistischen Tests entschieden werden, ob der Anteil jener, die das Pensionssystem positiv oder sehr positiv beurteilen, von 50% verschieden ist?

# Testen

---

## Beispiel: Aufgabensammlung

### 61. Meinungsumfrage: Beurteilung einer Pensionsreform:

	sehr positiv	positiv	negativ	sehr negativ	Summe
Ärzte	1	3	51	10	65
Rechtsanwälte	5	25	13	12	55
Krankenpfleger	11	21	20	16	68
Verkäufer	28	15	11	1	55

Für welche Berufsgruppe kann mittels eines statistischen Tests entschieden werden, ob der Anteil jener, die das Pensionssystem positiv oder sehr positiv beurteilen, von 50% verschieden ist?

# Testen

---

## Beispiel: Aufgabensammlung

**76.** Gebrauchtwagenhändler ermittelt durch Kundenbefragung, ob im 1. Jahr ab Kauf eine Reparatur notwendig war.

Fahrzeugtyp	Anzahl der Kunden	Autos mit Reparatur
Buick Park Avenue	125	42
Geo Metro Coupé	55	13
Alfa Romeo	153	28
Golf Rabbit	89	13

Prüfen sie für die verschiedenen Fahrzeugtypen die Hypothese, daß die Wahrscheinlichkeit einer Reparatur größer als 25 % ist.



# Testen

---

## Beispiel: Aufgabensammlung

**76.** Gebrauchtwagenhändler ermittelt durch Kundenbefragung, ob im 1. Jahr ab Kauf eine Reparatur notwendig war.

Fahrzeugtyp	Anzahl der Kunden	Autos mit Reparatur	$\hat{p}$	$SD$	$T$
Buick Park Avenue	125	42	0.336	0.039	2.221
Geo Metro Coupé	55	13	0.236	0.058	-0.234
Alfa Romeo	153	28	0.183	0.035	-1.914
Golf Rabbit	89	13	0.146	0.046	-2.264

Prüfen sie für die verschiedenen Fahrzeugtypen die Hypothese, daß die Wahrscheinlichkeit einer Reparatur größer als 25 % ist.

# Testen

---

## Beispiel: Aufgabensammlung

### 67. Überprüfung von gemeldeten Schadensfälle:

Versicherungssparte	überprüft	betrügerisch	$T$
Privathaftpflicht	62	17	4.572
Haushaltsversicherung	84	10	
Hausversicherung	76	15	
Autokasko	110	10	-0.318
Lebensversicherung	60	1	-2.152

Für welche Sparte kann nachgewiesen werden, daß der Anteil an betrügerischen Fällen größer als 10% ist?

# Testen

---

## Beispiel: Aufgabensammlung

### 67. Überprüfung von gemeldeten Schadensfälle:

Versicherungssparte	überprüft	betrügerisch	$T$	$\hat{p}$
Privathaftpflicht	62	17	4.572	0.274
Haushaltsversicherung	84	10	0.582	0.119
Hausversicherung	76	15	2.829	0.197
Autokasko	110	10	-0.318	0.091
Lebensversicherung	60	1	-2.152	0.017

Für welche Sparte kann nachgewiesen werden, daß der Anteil an betrügerischen Fällen größer als 10% ist?

# Testen

---

- (a) Privathaftpflicht, Hausversicherung
- (b) Privathaftpflicht, Haushaltsversicherung, Hausversicherung
- (c) Lebensversicherung
- (d) Privathaftpflicht, Haushaltsversicherung, Hausversicherung, Autokasko
- (e) Lebensversicherung, Autokasko

# Testen

---

## Beispiel: Aufgabensammlung

**56.** In einer Fernsehdiskussion behauptet ein Politiker, seine Partei würde bei Neuwahlen 25% der Stimmen erhalten. Ein Meinungsforscher widerspricht, denn eine Befragung seines Instituts hätte einen Anteil von 21% ergeben, womit die Behauptung des Politikers zum Signifikanzniveau 95% statistisch widerlegt sei.

Wie groß muß die Stichprobe mindestens gewesen sein?

## P-Wert

---

Gegeben sei ein Datensatz:

Ist  $\alpha$  jenes Signifikanzniveau, zu dem die Nullhypothese gerade noch verworfen werden kann, dann nennt man  $1 - \alpha$  den **P-Wert** der Daten.

Das Signifikanzniveau wird vor dem Testen festgelegt. Daraus ergeben sich die kritischen Werte für die Testgröße.

Verschiedene Personen können ein unterschiedliches Signifikanzniveau festlegen und daher zu unterschiedlichen Entscheidungen kommen.

## P-Wert

---

### Beispiel (4.4)

Teststatistik:

$$T = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = -4.123$$

$P$ -Wert  $\approx 0.00004$ .

**Interpretation:** Entweder ist ein extrem unwahrscheinliches Ereignis eingetreten (Wahrscheinlichkeit 0.00004) oder die Nullhypothese  $p = 0.5$  ist falsch.

# $P$ -Wert

---

## Merkregel

- Kleine  $P$ -Werte sprechen gegen die Nullhypothese.
- Große  $P$ -Werte sprechen **nicht gegen** die Nullhypothese.



# Punktschätzung

---

## Bisher:

- Die Wahrscheinlichkeit  $p$  war bekannt (unter der Nullhypothese).
- Betrachte Eigenschaften von  $f_n(A)$  bzw.  $h_n(A)$  in Abhängigkeit von  $p$ .

**Jetzt:** Die Wahrscheinlichkeit  $p$  ist unbekannt und soll aus Daten geschätzt werden.

# Punktschätzung

---

Der beste Schätzer für eine unbekannte Wahrscheinlichkeit  $p$  ist die relative Häufigkeit  $\hat{p} = f_n(A)$ .

**Frage:** Wie genau ist die Schätzung  $\hat{p}$ ?

# Konfidenzintervalle

---

Einfache Antwort nach Faustregel: mit 95% Sicherheit liegt  $p$  im Intervall

$$| \hat{p} - p | \leq 2SD$$
$$\hat{p} - 2SD \leq p \leq \hat{p} + 2SD$$

Unterschied zum Prognoseintervall: Intervallgrenzen sind nun zufällig, eingeschlossener Wert  $p$  nun fest.

**Problem:**  $SD$  hängt vom unbekannten Wert  $p$  ab.

# Konfidenzintervalle

---

**Ausweg:** Verwende geschätztes  $\hat{p}$  (**Bootstrapmethode**).

$$\widehat{SD} = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

Das Intervall

$$p_{1,2} = \hat{p} \pm c \widehat{SD}$$

heißt **Konfidenzintervall**.

Dieses überdeckt mit hoher Sicherheit (bestimmt durch den kritischen Wert  $c = c(\alpha)$ ) die wahre Wahrscheinlichkeit  $p$ .

# Konfidenzintervalle

---

**Korrekte Methode:** Löse die quadratische Gleichung:

$$\begin{aligned} |\hat{p} - p| &= c SD \\ \Leftrightarrow (\hat{p} - p)^2 &= c^2 \frac{p(1-p)}{n} \end{aligned}$$

Das korrekte Konfidenzintervall ist

$$p_{1,2} = \frac{1}{1 + c^2/n} \left( \hat{p} + \frac{c^2}{2n} \pm \frac{c}{\sqrt{n}} \sqrt{\hat{p}(1 - \hat{p}) + \frac{c^2}{4n}} \right).$$

# Konfidenzintervalle

---

**Robuste Methode:** Die wahre  $SD$  kann abgeschätzt werden durch

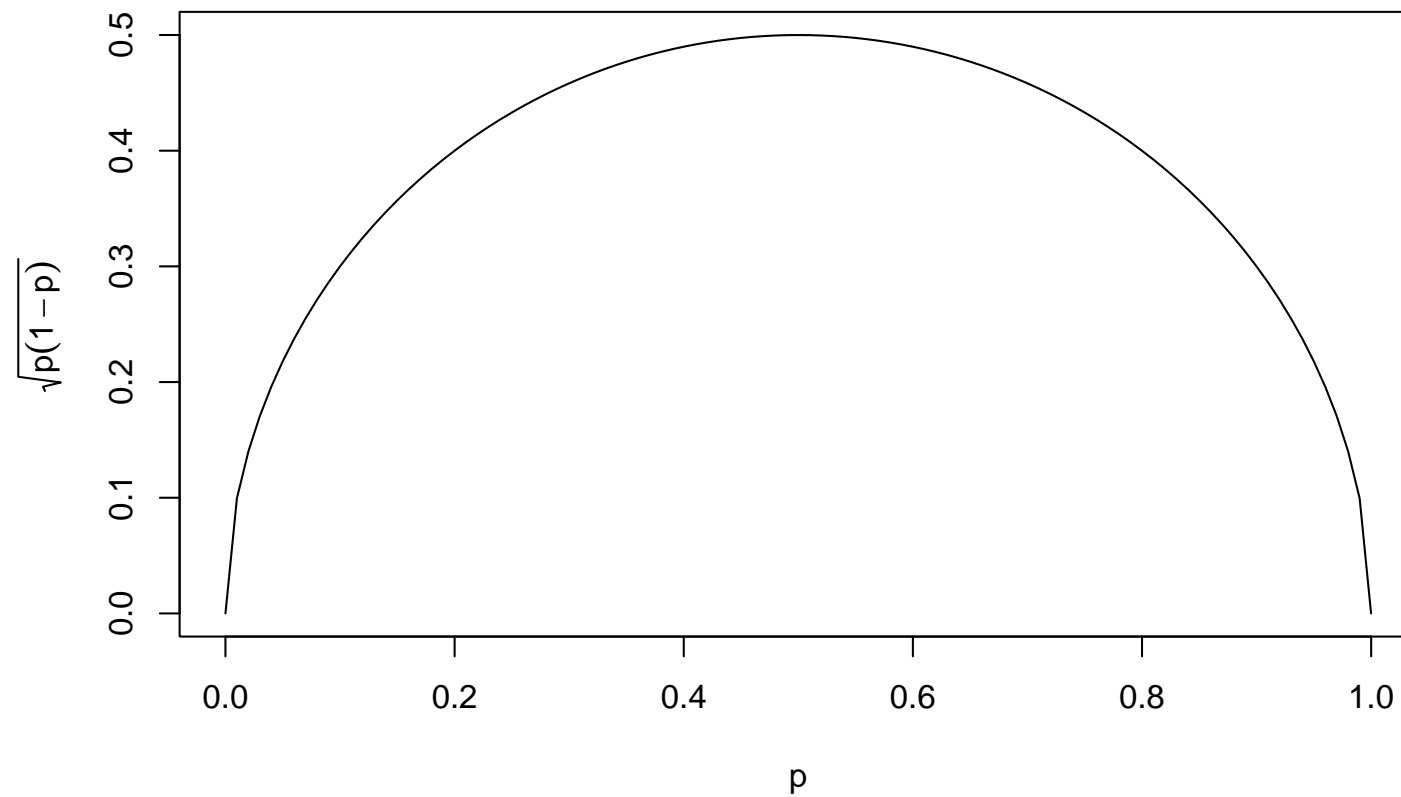
$$SD = \sqrt{\frac{p(1-p)}{n}} \leq SD_{\max} = \frac{1/2}{\sqrt{n}}.$$

Das robuste Konfidenzintervall ist

$$p_{1,2} = \hat{p} \pm c \frac{1}{2\sqrt{n}}.$$

# Konfidenzintervalle

---



# Konfidenzintervalle

---

## Beispiel: Aufgabensammlung

### 90. Autoreparaturen

Fahrzeugtyp	Anzahl der befragten Kunden	Autos mit Reparatur
Buick Park Avenue	125	42
Geo Metro Coupé	55	13
Alfa Romeo	153	28
Golf Rabbit	89	13

Geben sie für den Buick Park Avenue ein 95% Konfidenzintervall für die Wahrscheinlichkeit einer Reparatur an (Robuste Methode, Faustregel und Bootstrap Methode, Faustregel).



# Konfidenzintervalle

---

## Beispiel: Aufgabensammlung

### 61. Pensionsreform

	sehr positiv	positiv	negativ	sehr negativ
Ärzte	1	3	51	10
Rechtsanwälte	5	25	13	12
Krankenpfleger	11	21	20	16
Verkäufer	28	15	11	1

Wieviele Personen hätte man mindestens befragen müssen, um den Anteil jener, die der Reform pos./sehr pos. gegenüberstehen, mit  $\pm 5\%$  angeben zu können? (Robuste Methode, Faustregel)

## $\sqrt{n}$ -Gesetz

---

Das Prognoseintervall

$$p \pm c(\alpha) SD$$

hat die Länge  $\ell$  mit

$$\ell = 2 c(\alpha) SD = 2 c(\alpha) \sqrt{p(1-p)} \frac{1}{\sqrt{n}}.$$

Für jede Wahl von  $\alpha$  und  $p$  gilt: Die Intervallbreite eines Prognose- oder Konfidenzintervalls ist indirekt proportional zu  $\sqrt{n}$ .

## $\sqrt{n}$ -Gesetz

---

**Beispiel:** Genauigkeitsverdoppelung

$$\frac{\ell}{2} = 2 c(\alpha) \sqrt{p(1-p)} \frac{1}{\sqrt{4n}}.$$

Damit sich die Präzision der Schätzung (= 1/Intervallbreite) verdoppelt, muß der Stichprobenumfang vervierfacht werden.

## $\sqrt{p}$ -Gesetz

---

Der absolute Fehler eines Prognoseintervalls ist proportional zu  $\sqrt{p(1-p)}$ . Entscheidend ist jedoch der relative Fehler (in Relation zu  $p$ ).

Für jede Wahl von  $\alpha$  und  $n$  gilt: Der relative Fehler eines Prognoseintervalls für eine relative Häufigkeit ist bei kleinen Werten von  $p$  proportional zu  $1/\sqrt{p}$ .

# Zusammenfassung Kapitel 4

---

- Testen von Hypothesen
- Nullhypothese, Alternative
- (nicht-)signifikante Ergebnisse
- Beibehalten/Verwerfen der Nullhypothese
- Teststatistik, Signifikanzniveau, kritischer Wert
- Fehler 1./2. Art
- $P$ -Wert
- Konfidenzintervalle: korrekt, Bootstrap, robust

# Verteilungsmaßzahlen

## Kapitel 5

# Verteilungsmaßzahlen

---

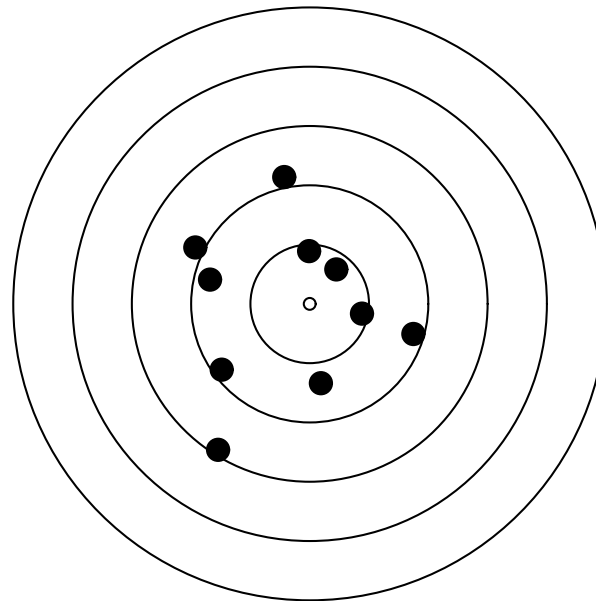
Gestalt von Verteilungen wird beschrieben:

- Lage (Mittelwert, Median)
- Streuung
- Schiefe
- Wölbung

# Verteilungsmaßzahlen

---

Lage: gut – Streuung: schlecht

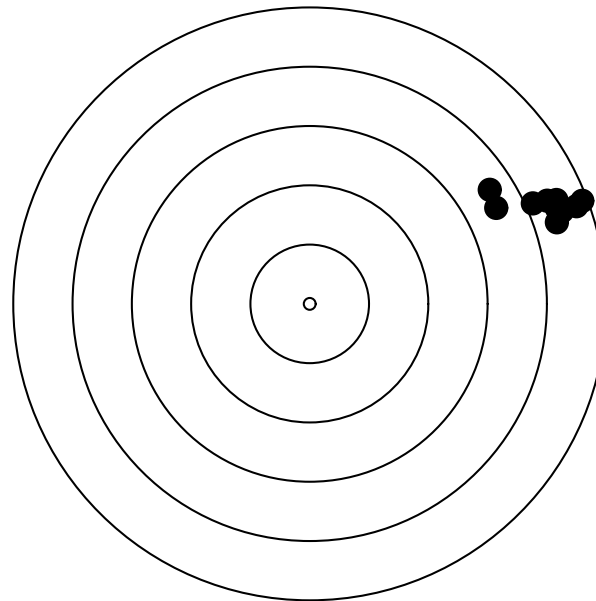




# Verteilungsmaßzahlen

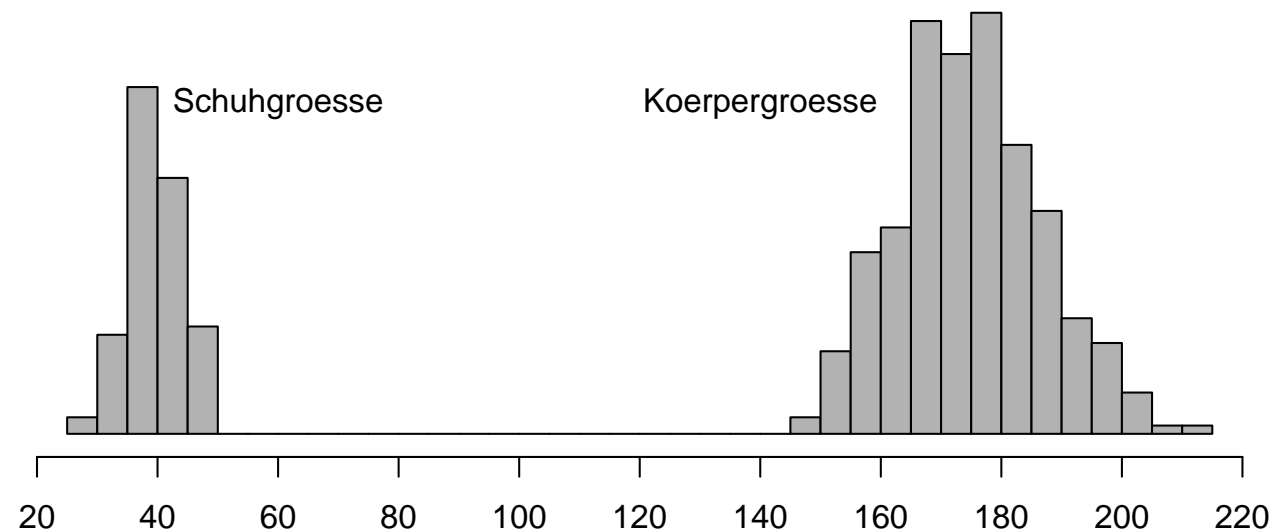
---

Lage: schlecht – Streuung: gut



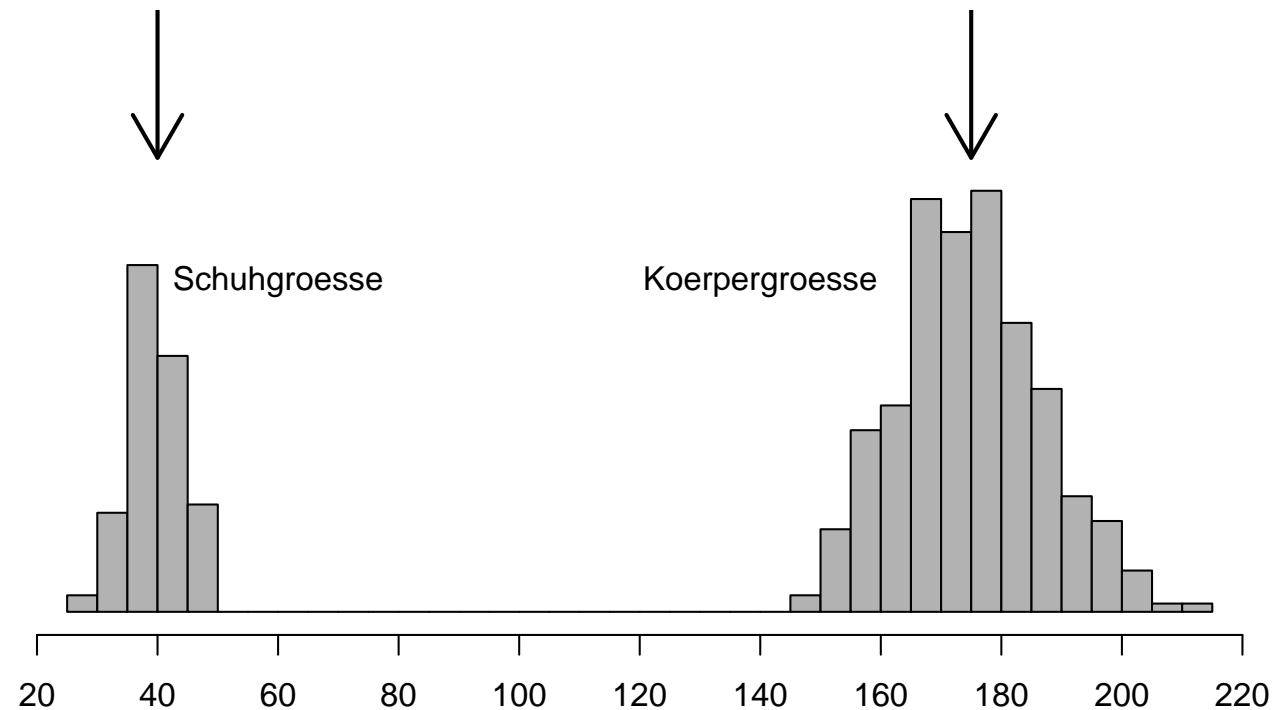
# Verteilungsmaßzahlen

---



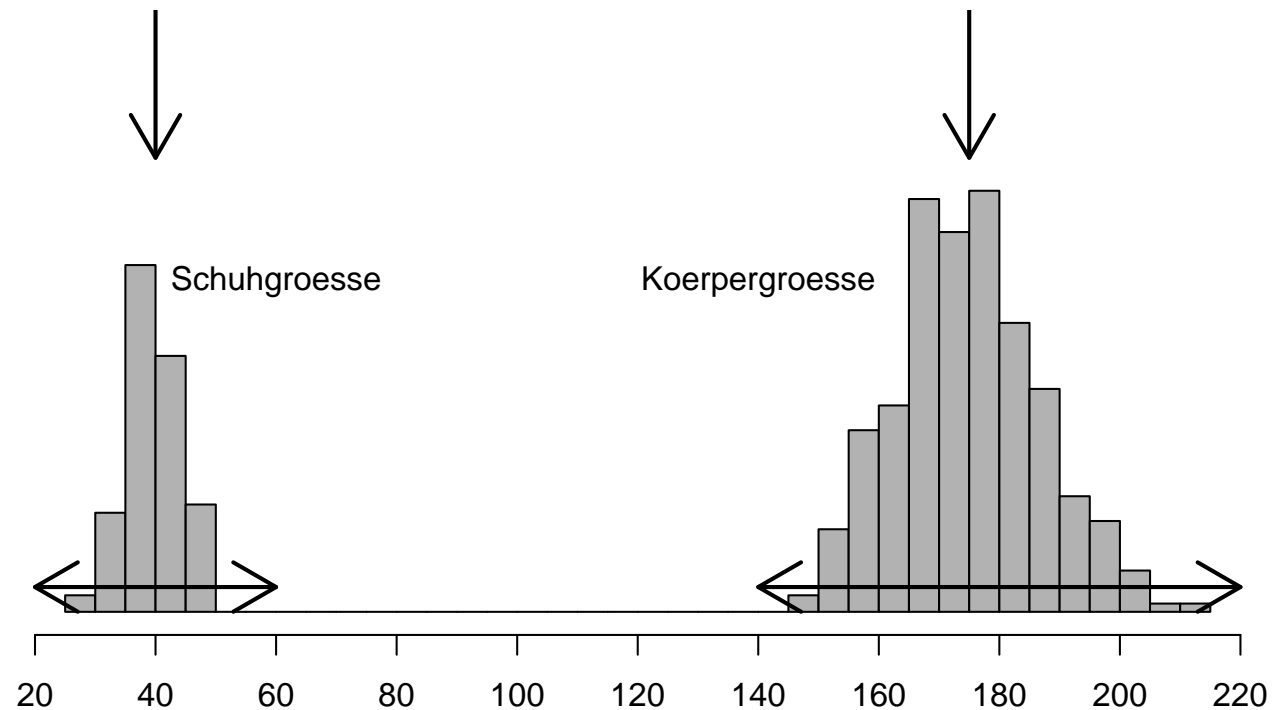
# Verteilungsmaßzahlen

---



# Verteilungsmaßzahlen

---



# Verteilungsmaßzahlen

---

**Illustration:**

**Lage** Körpergröße: Männer – Frauen

**Streuung** Vergleich zweier Meßverfahren

**Schiefe** Monatseinkommen: viele kleine Einkommen, wenige sehr große

**Wölbung**

# Verteilungsmaßzahlen

---

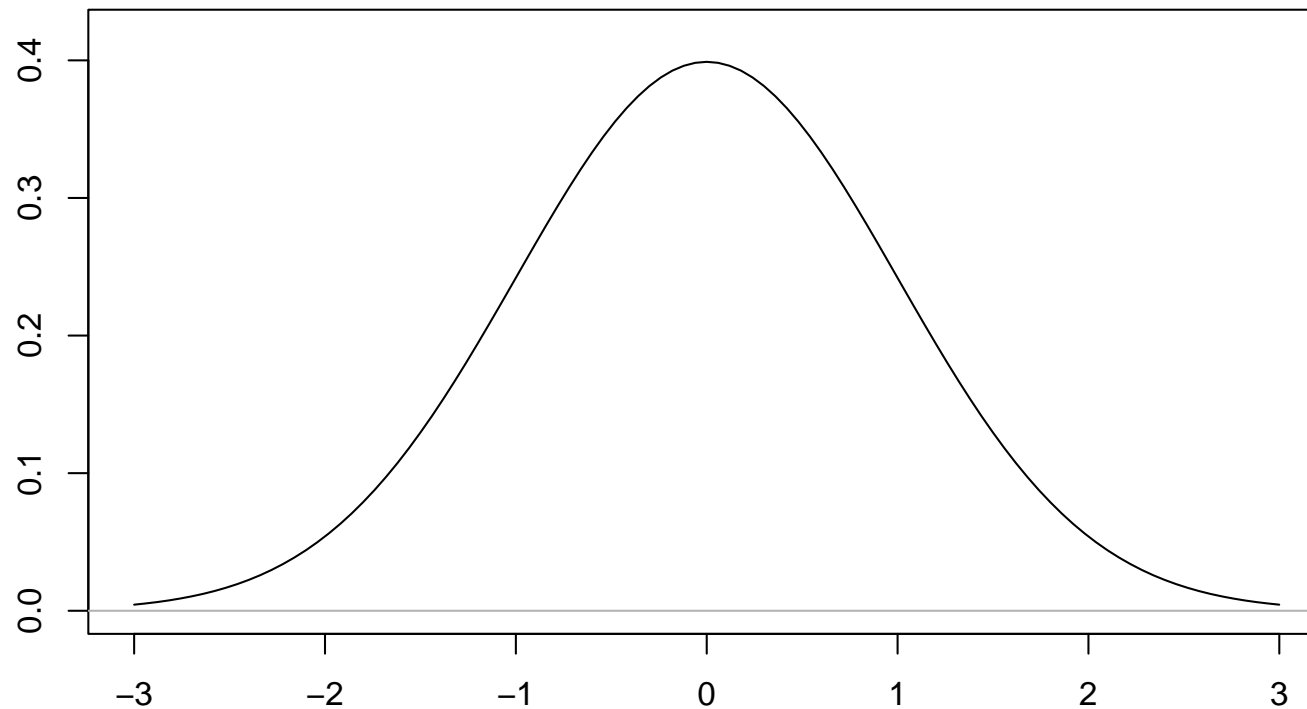


Abbildung 1: Standardnormalverteilung

# Verteilungsmaßzahlen

---

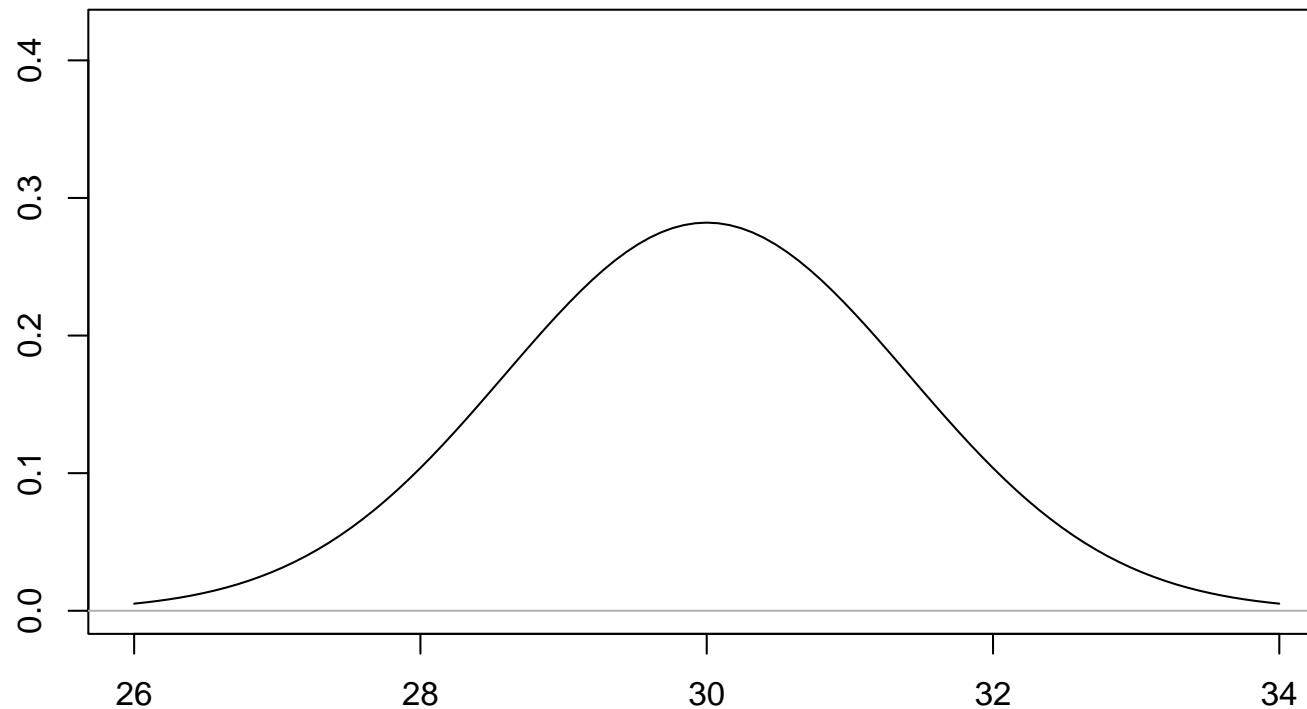


Abbildung 2: Normalverteilung:  $\mu = 30$ ,  $\sigma^2 = 2$

# Verteilungsmaßzahlen

---

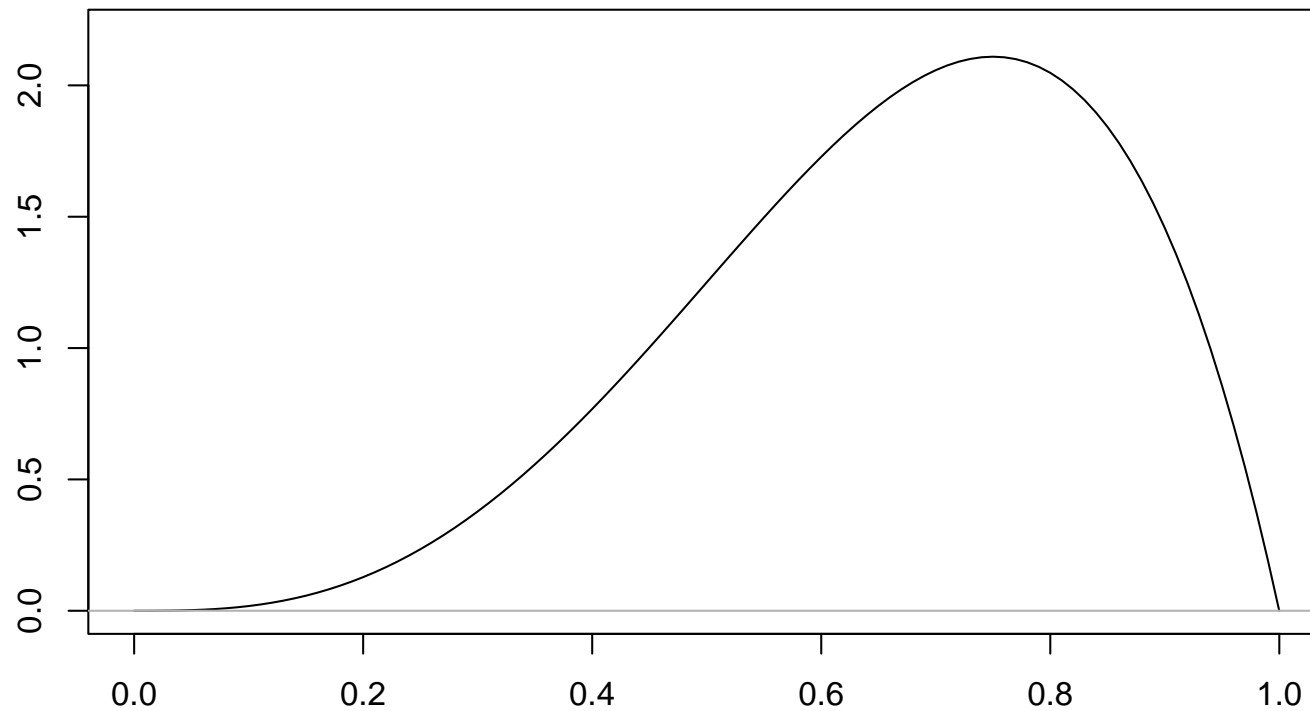


Abbildung 3: Betaverteilung



# Verteilungsmaßzahlen

---

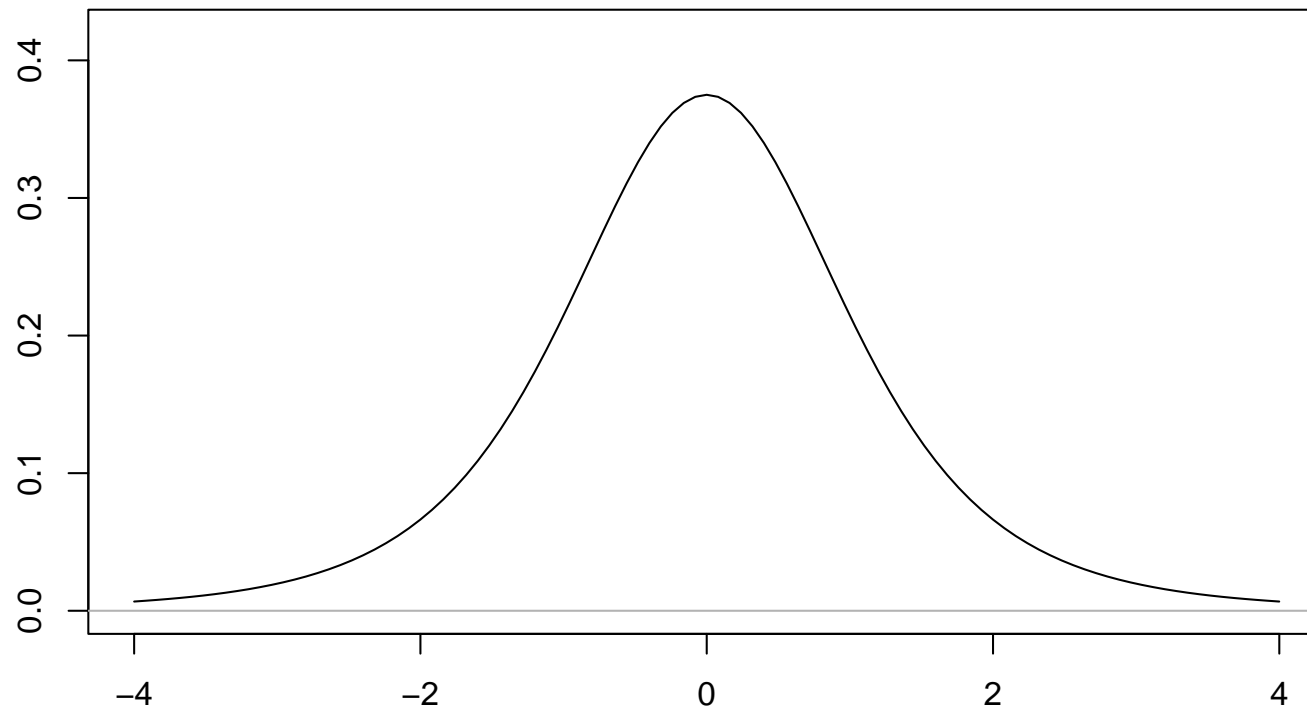


Abbildung 4: t-Verteilung

# Lagemaße

---

**Mittelwert**  $\bar{x} = \frac{1}{n}(x_1 + x_2 + \dots + x_n)$

Das Zentrum des Datensatzes wird durch den **Abstand** der Datenpunkte bestimmt.

**Median** Quantil  $Q_{0.5}$

Das Zentrum des Datensatzes wird durch die **Anzahl** der Datenpunkte bestimmt.

## Mittelwert einer Häufigkeitsverteilung

	$h$	$f$
$a_1$	$h_1$	$f_1$
$\vdots$	$\vdots$	$\vdots$
$a_m$	$h_m$	$f_m$

$$\begin{aligned}\bar{x} &= \frac{1}{n} (a_1 h_1 + a_2 h_2 + \dots + a_m h_m) \\ &= a_1 f_1 + a_2 f_2 + \dots + a_m f_m\end{aligned}$$

# Robuste Lagemaße

---

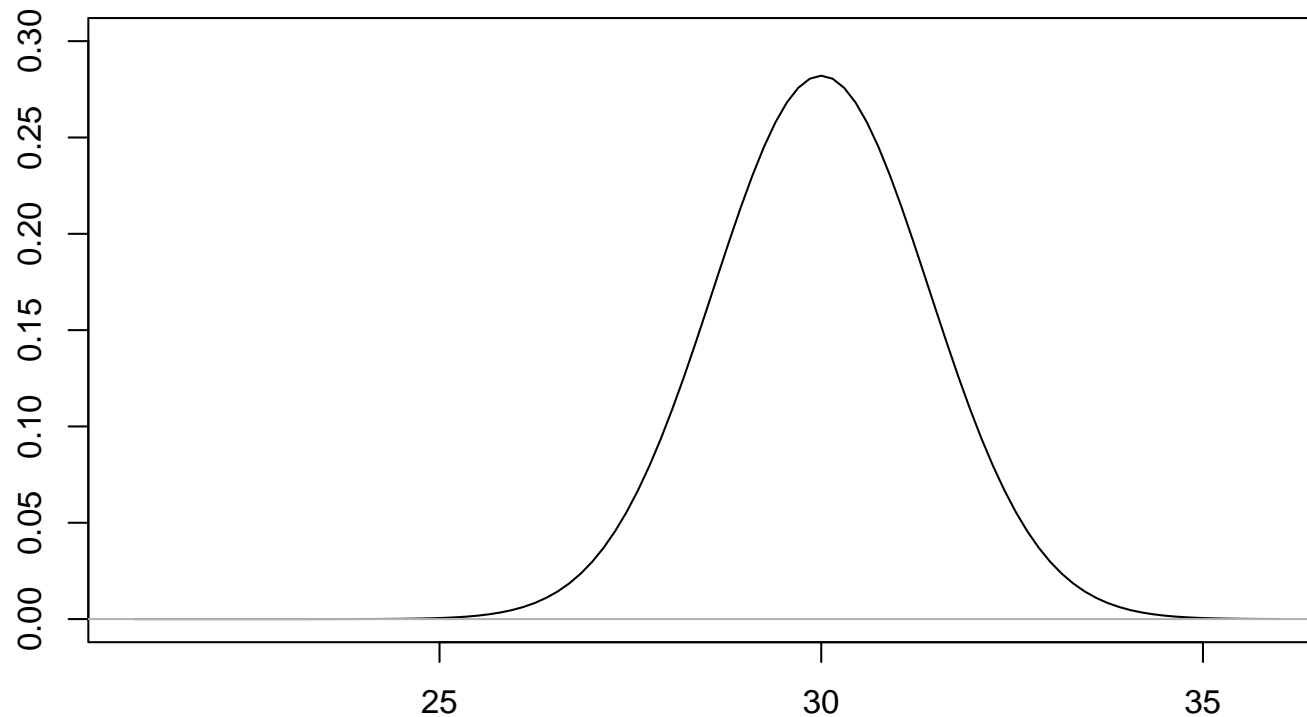


Abbildung 5: Normalverteilung:  $\mu = 30$ ,  $\sigma^2 = 2$

# Robuste Lagemaße

---

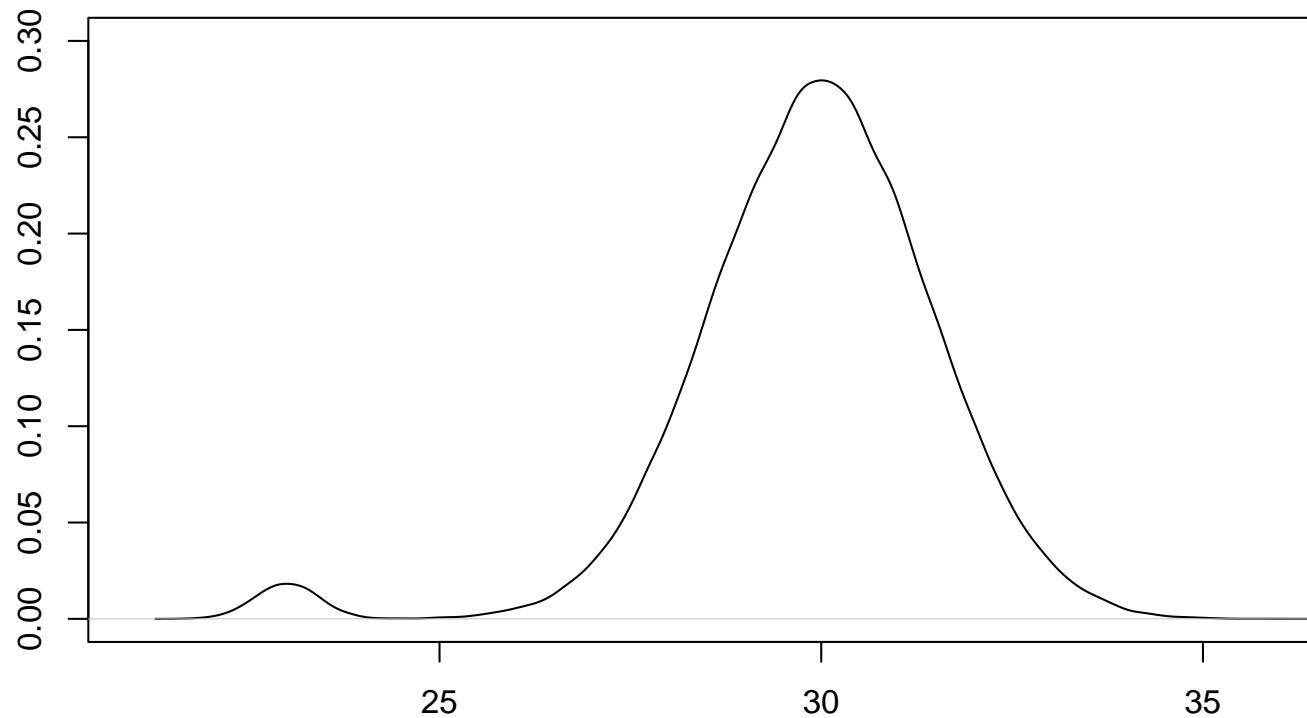


Abbildung 6: Normalverteilung:  $\mu = 30$ ,  $\sigma^2 = 2$  (mit Ausreißern)

# Robuste Lagemaße

---

Eine Maßzahl heißt **robust**, wenn sie sich bei Ausreißern nur beschränkt ändert.

## Beispiel:

Datenliste: 25, 29, 31, 33, 37

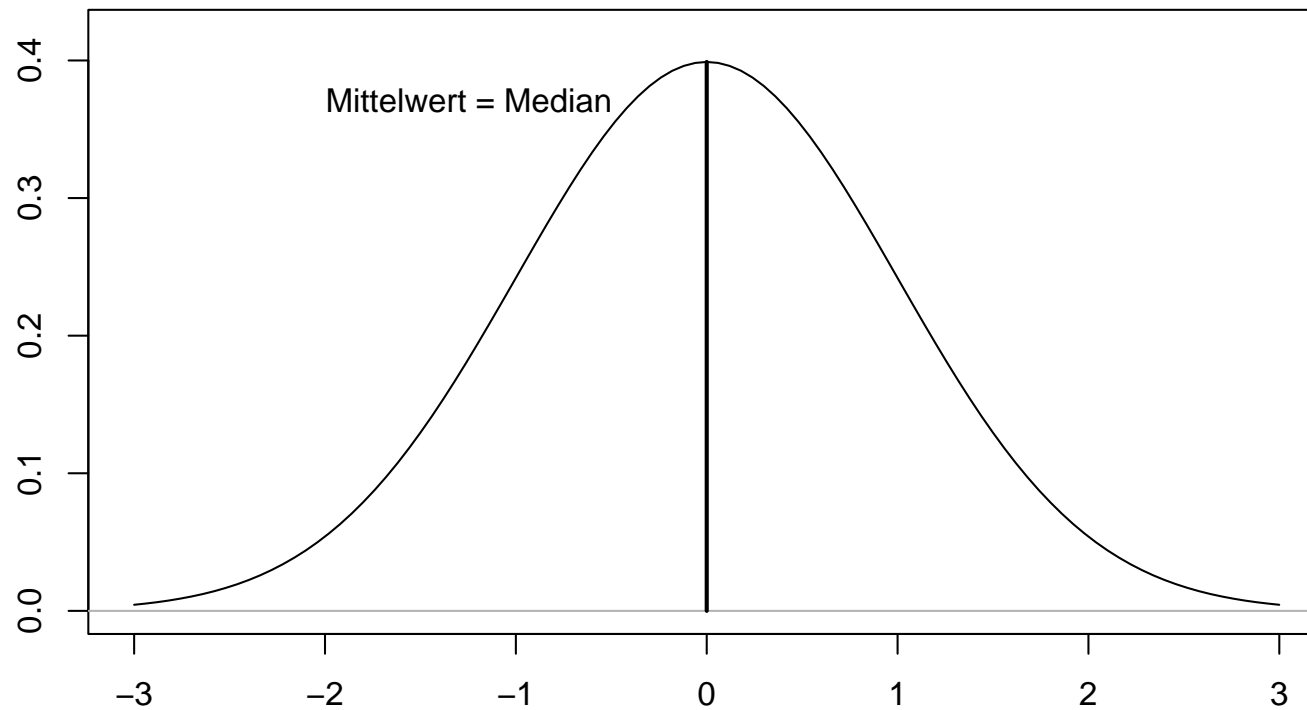
Mittelwert = Median = 31

Datenliste: 25, 29, 31, 33, 73

Median = 31  $\ll$  Mittelwert = 38.2

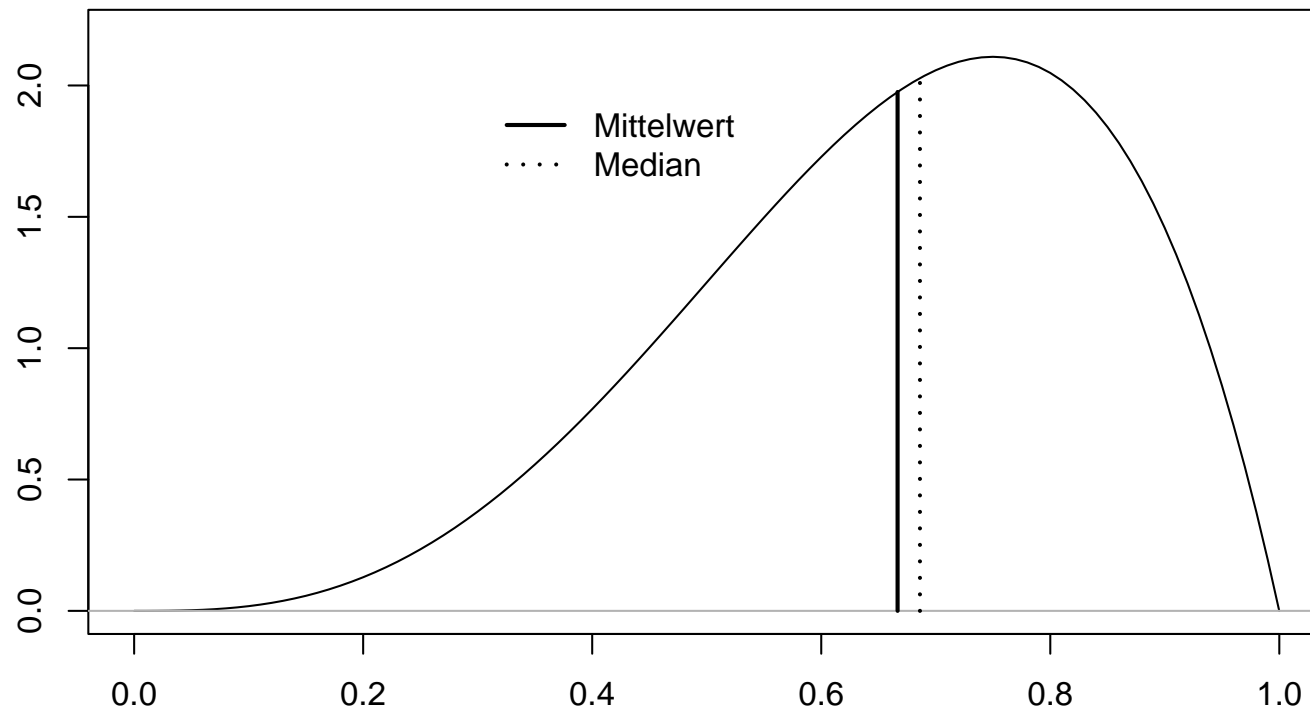
# Schiefte

---



# Schiefte

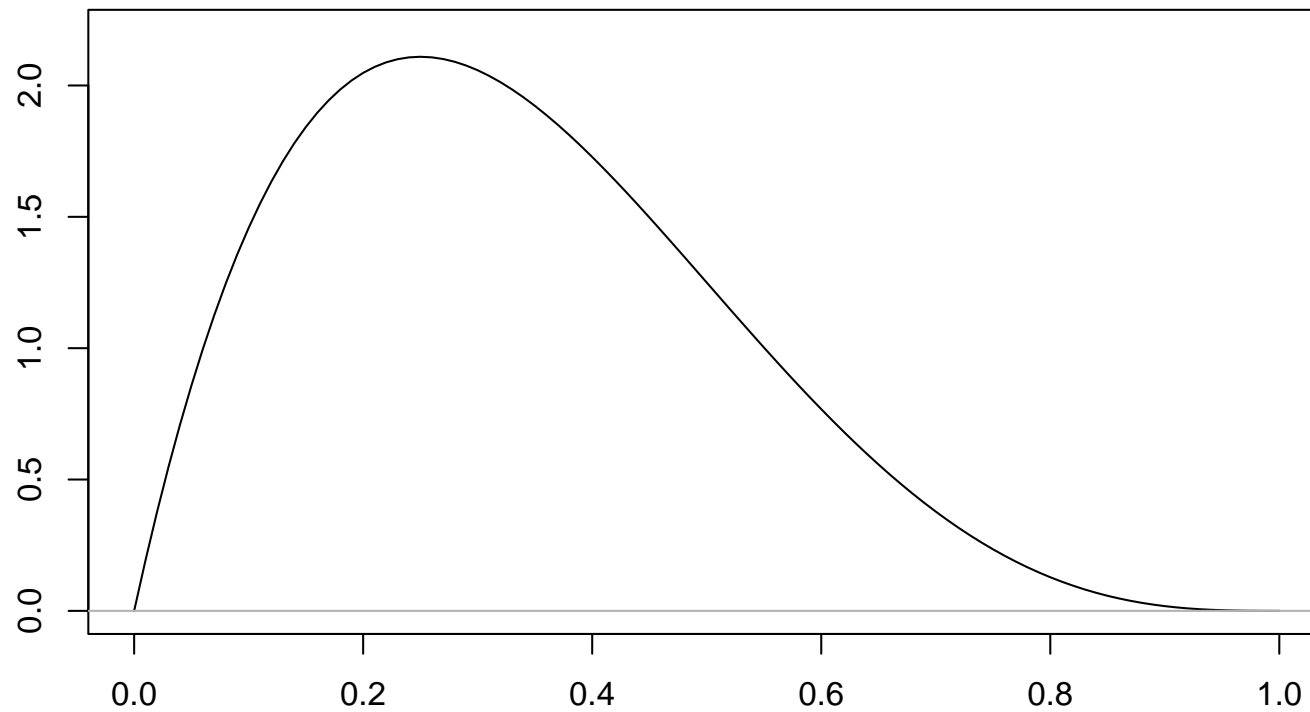
---





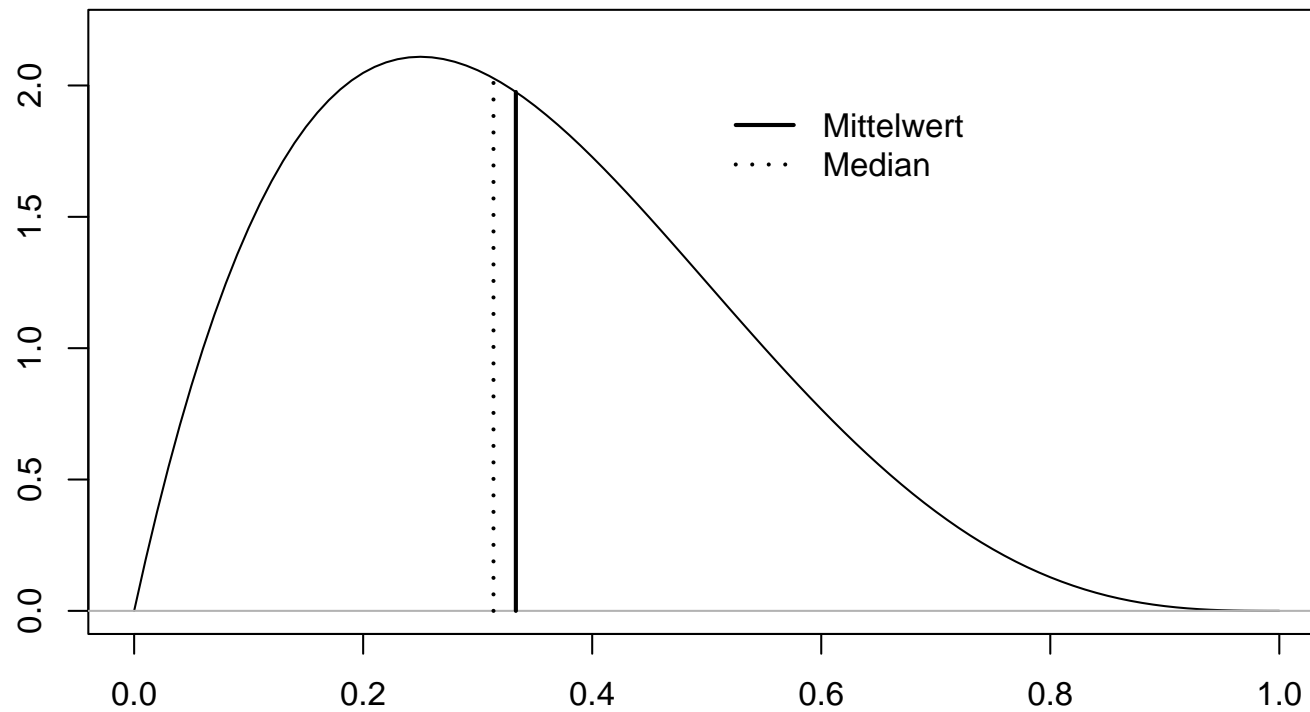
# Schiefe

---



# Schiefte

---



# Schiefe

---

Bei **symmetrischen** Verteilungen stimmen Mittelwert und Median überein. Der Median ist das Symmetriezentrum.

Sonst heißt die Verteilung **schief**. Es gibt **rechtsschiefe** und **linksschiefe** Verteilungen.

Der **Schiefekoeffizient**:

$$SK = \frac{R - L}{R + L}$$

$$R = Q_{0.75} - Q_{0.5}, L = Q_{0.5} - Q_{0.25}.$$

# Schiefe

---

Der  $SK$  liegt zwischen  $-1$  ( $R = 0$ ) und  $+1$  ( $L = 0$ ).  
Symmetrie bei  $SK = 0$ .

Es gilt:

- linksschief = rechtssteil ( $SK < 0$ , weil  $L > R$ )
- rechtsschief = linkssteil ( $SK > 0$ , weil  $L < R$ )

# Boxplots

---

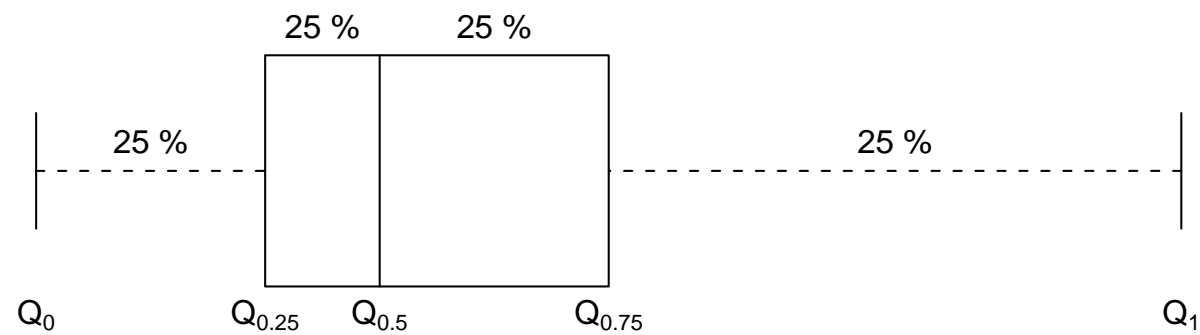
Der **Boxplot** ist eine Darstellung, aus der die Five Point Summary

$$Q_0 \quad Q_{0.25} \quad Q_{0.5} \quad Q_{0.75} \quad Q_1$$

direkt abgelesen werden kann.

# Boxplots

---



# Boxplots, Symmetrie

---

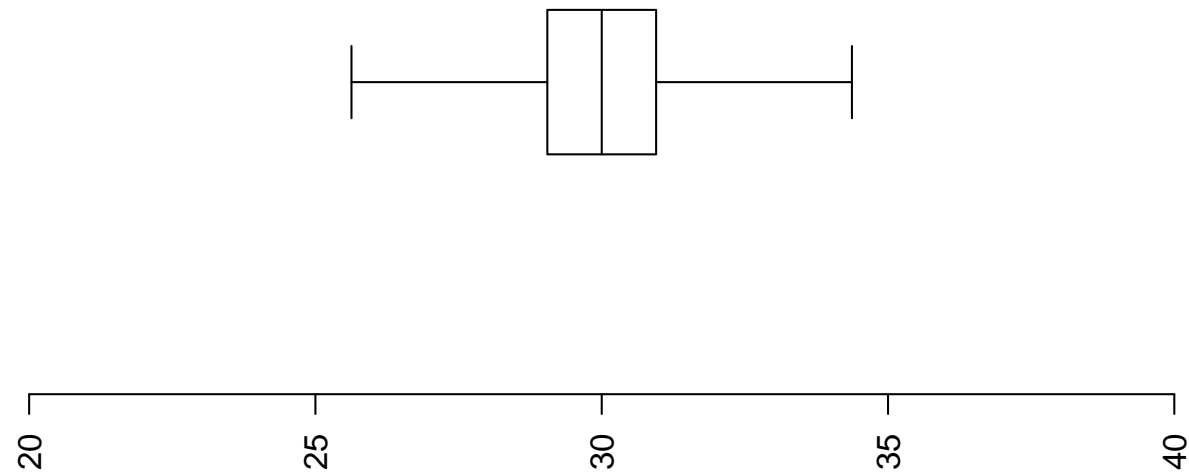


Abbildung 7: Normalverteilung:  $\mu = 30$ ,  $\sigma^2 = 2$

# Boxplots, Ausreißer

---

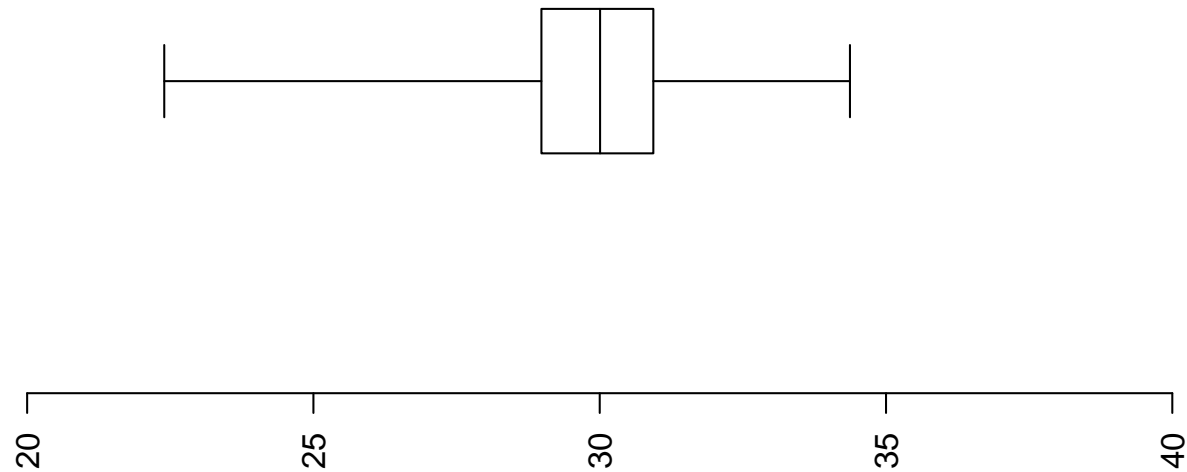


Abbildung 8: Normalverteilung:  $\mu = 30$ ,  $\sigma^2 = 2$  (mit Ausreißern)



# Boxplots, Schiefe

---

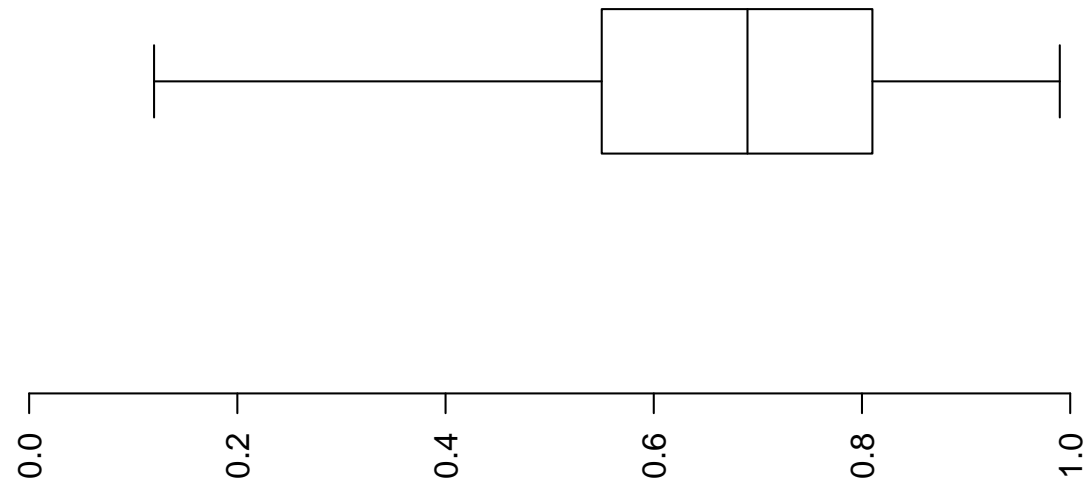
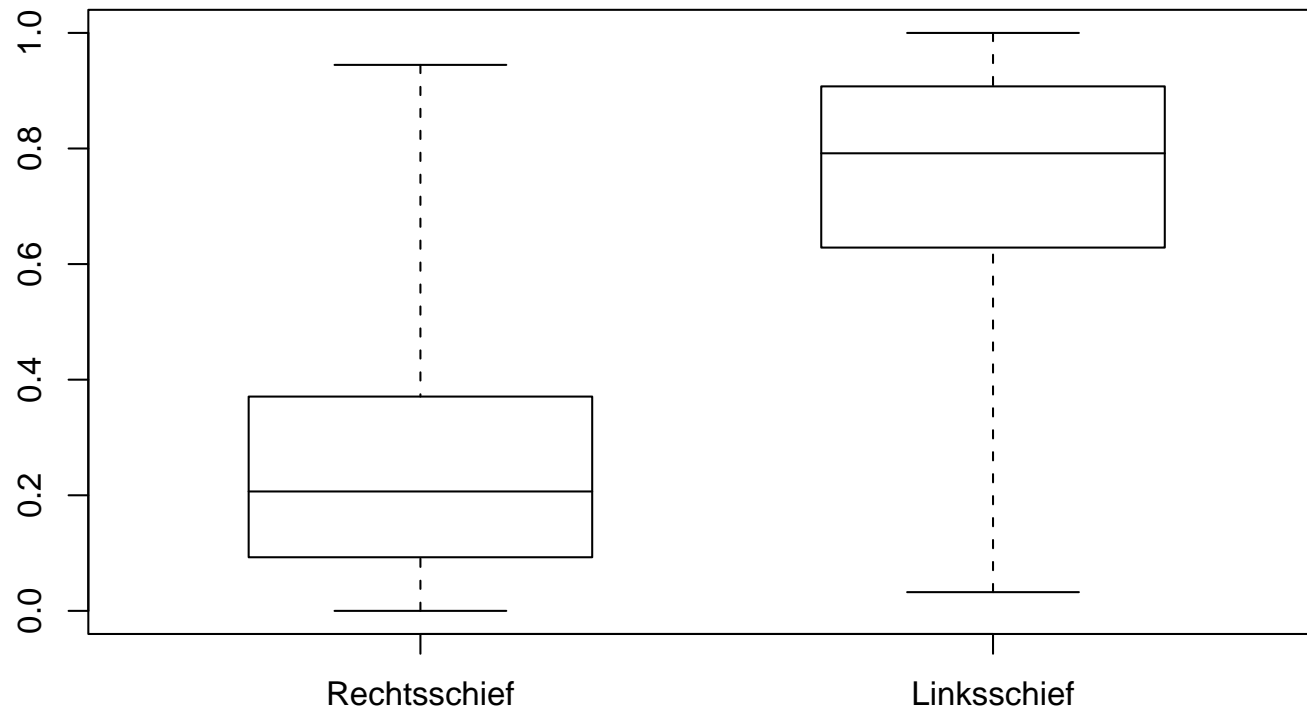


Abbildung 9: Betaverteilung

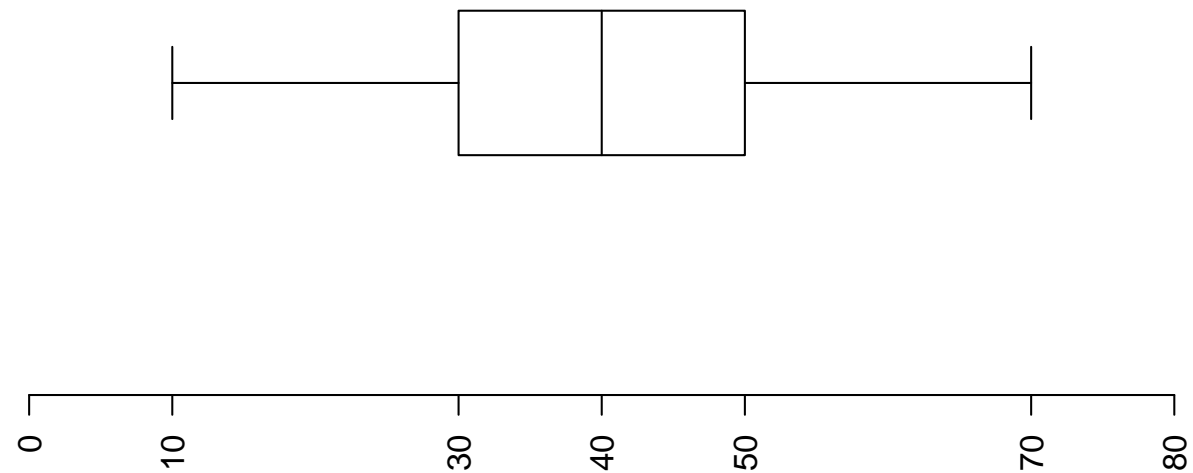
# Boxplots, Schiefe

---



# Boxplot, Beispiel 1

---



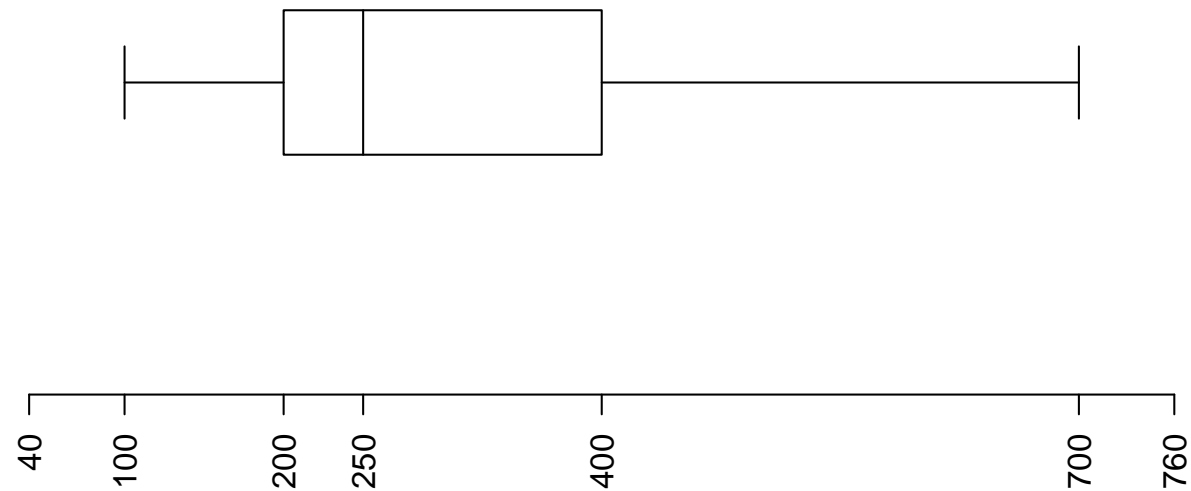
## Boxplot, Beispiel 1

---

- Die Verteilung ist linksschief. falsch,  $SK = 0$ .
- Der Schiefekoeffizient ist größer als 0.3. falsch,  $SK = 0$ .
- Der Median ist größer als der Mittelwert. falsch, da die Verteilung nicht linksschief.
- Mindestens 25% der Beobachtungen sind kleiner oder gleich 30. richtig,  $Q_{0.25} = 30$ .
- Die Interquartilsdistanz beträgt höchstens 30. richtig,  $QD = Q_{0.75} - Q_{0.25} = 50 - 30 = 20$ .

## Boxplot, Beispiel 2

---



Boxplot aus Abbildung 63 (Aufgabensammlung)

## Boxplot, Beispiel 2

---

- Wert 200 wird von 25% der Werte nicht überschritten. **richtig**,  $Q_{0.25} = 200$ .
- Der Schiefekoeffizient ist größer als 0.4. **richtig**,  $SK = 0.5$ .
- Maximaler Wert beträgt 400. **falsch**,  $400 = Q_{0.75}$ .
- In  $[250, 400]$  liegen 50% der Daten. **falsch**
- Die Verteilung ist rechtsschief. **richtig**
- Mittelwert ist kleiner als 250. **falsch**,  
denn Mittelwert  $>$  Median = 250.

# Streuungsmaßzahlen

---

- **Varianz:**

$$s_x^2 = \frac{1}{n} \left( (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2 \right)$$

vgl. Abb. 1 und 2.

- **Standardabweichung:**  $s_x$

- **Quartilsdistanz:**  $QD = Q_{0.75} - Q_{0.25}$  ist robust gegen Ausreißer, da ja die **zentralen** 50 % beschrieben werden.

vgl. Abb. 7 und 8.

# Standardisierung

---

Der Einfluß des Maßstabs wird herausgenommen.

- Rangzahlen
- Standardscores:

$$z = \frac{x - \bar{x}}{s_x}$$

vgl. Abb. 1 und 2.



# Normalverteilung

---

**Standardnormalverteilung,  $N(0, 1)$ :**

Dichtefunktion

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

vgl. Abb. [1](#)

**Normalverteilung,  $N(\mu, \sigma^2)$ :**

Dichtefunktion

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

vgl. Abb. [2](#)

# Normalverteilung

---

Ist die Größe  $X$  normalverteilt mit Mittelwert  $\mu$  und Varianz  $\sigma^2$  ( $N(\mu, \sigma^2)$ ), dann sind die Standardscores

$$Z = \frac{X - \mu}{\sigma}$$

standardnormalverteilt ( $N(0, 1)$ ).

# Normalverteilung

---

Die Verteilungsfunktion der Standardnormalverteilung ist in der Normalverteilungstabelle vertafelt. Die Tabelle gibt zu einem Quantil  $N_\alpha$  die zugehörige Wahrscheinlichkeit  $\alpha$  an.

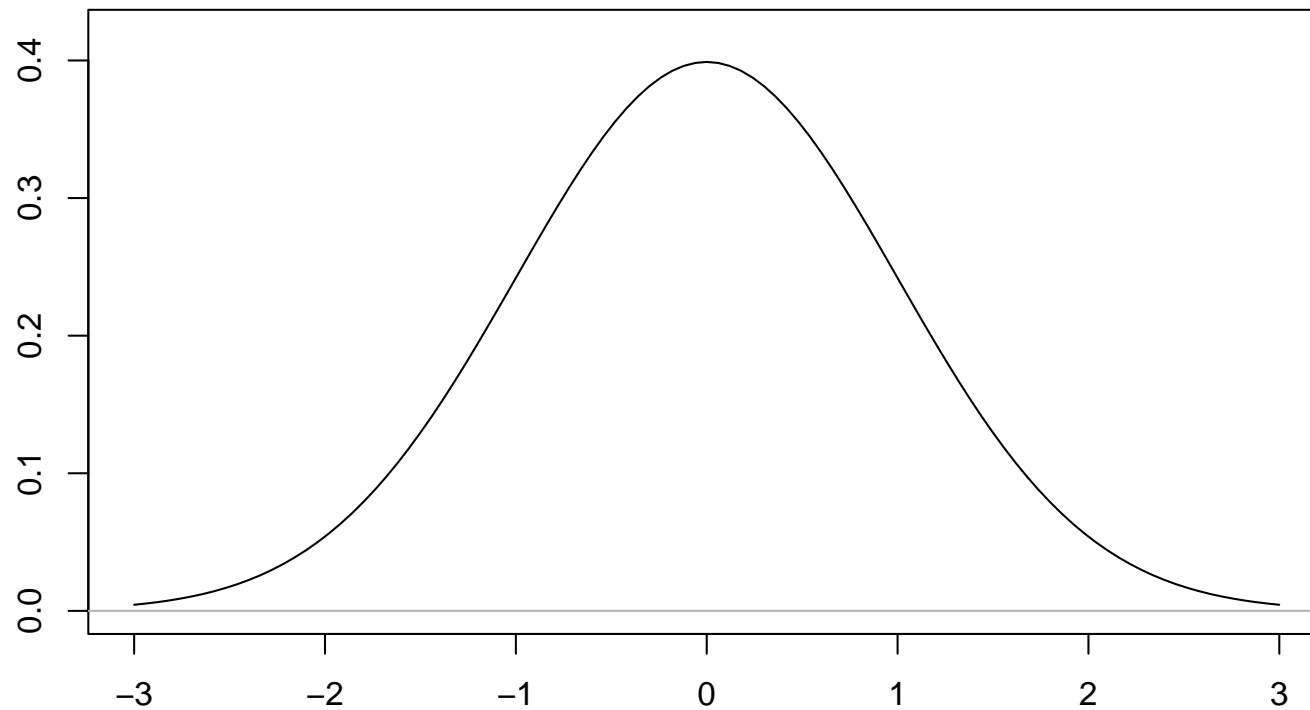
Man schreibt auch:

$$P(Z \leq z) = \Phi(z)$$

$$P(Z \leq N_\alpha) = \Phi(N_\alpha) = \alpha$$

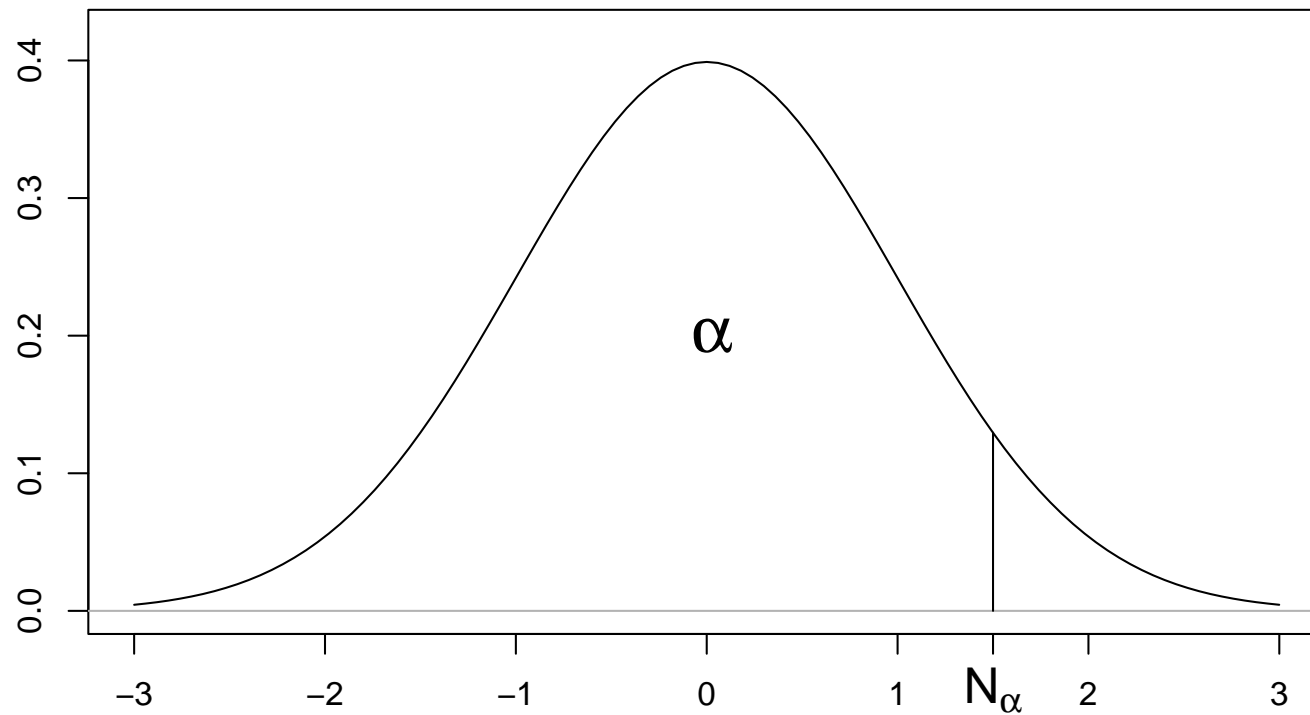
# Normalverteilung

---



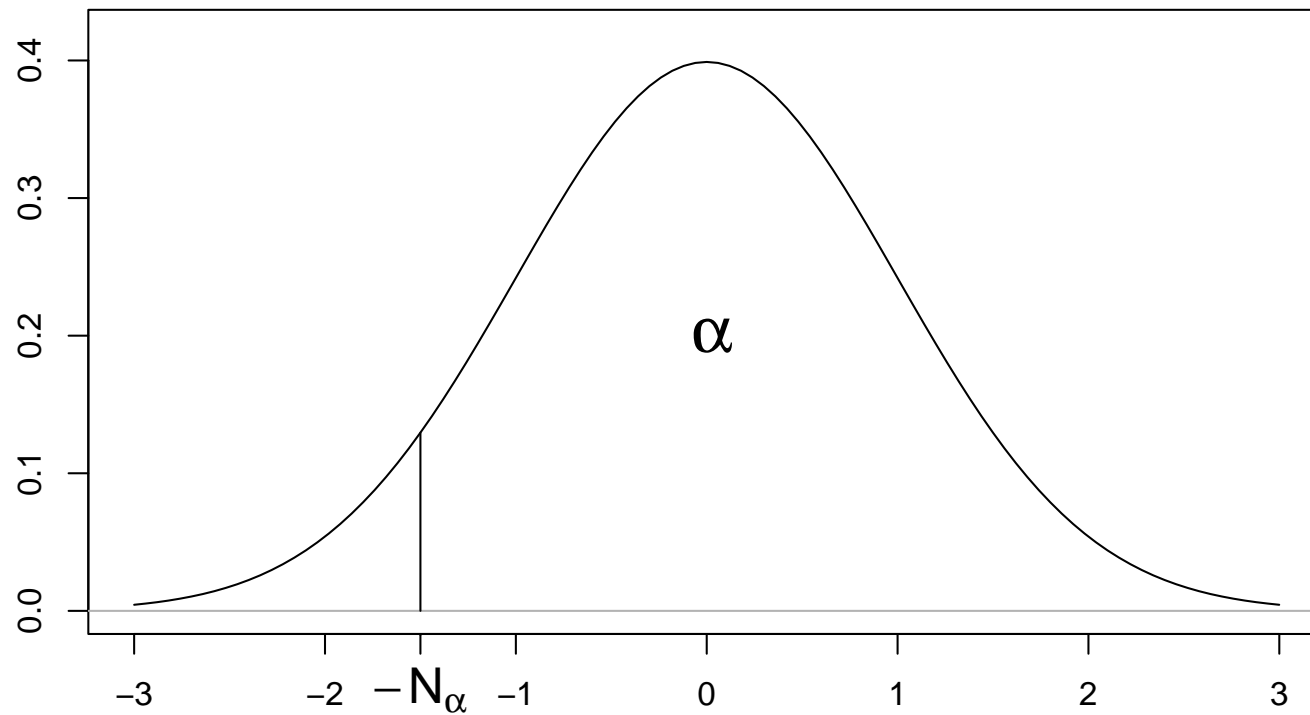
# Normalverteilung

---



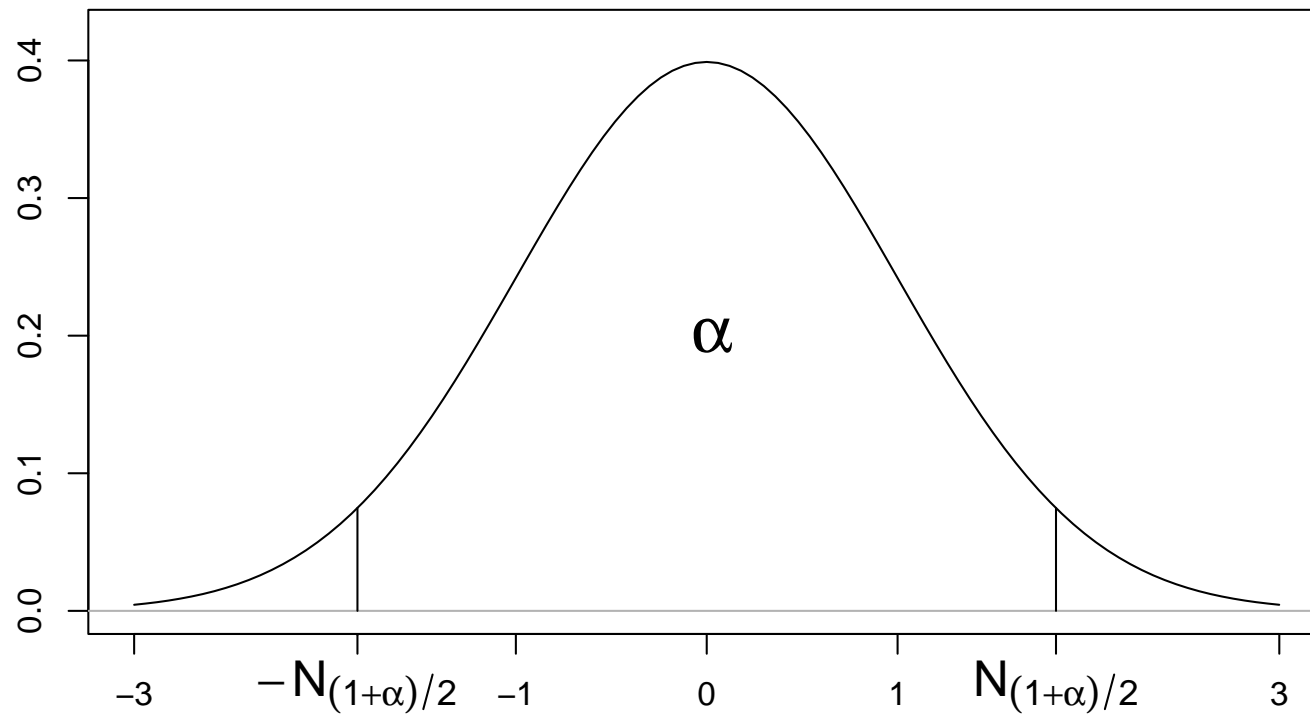
# Normalverteilung

---



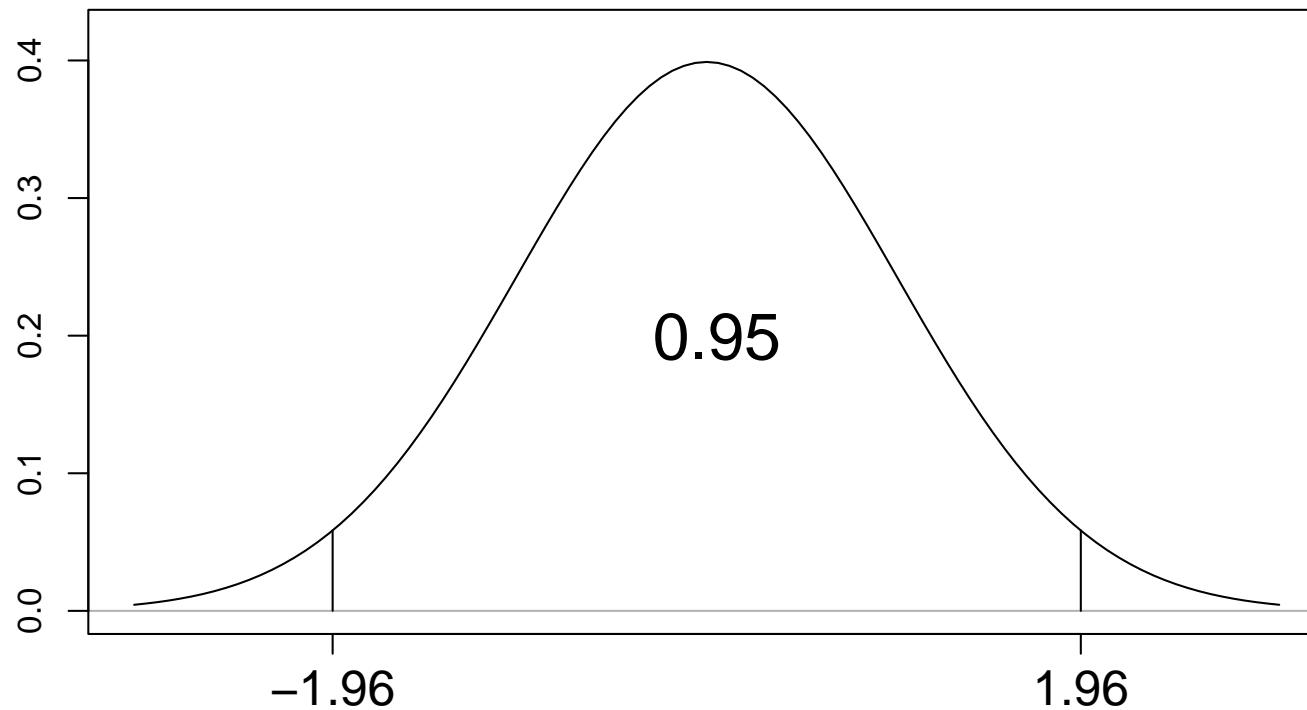
# Normalverteilung

---



# Normalverteilung

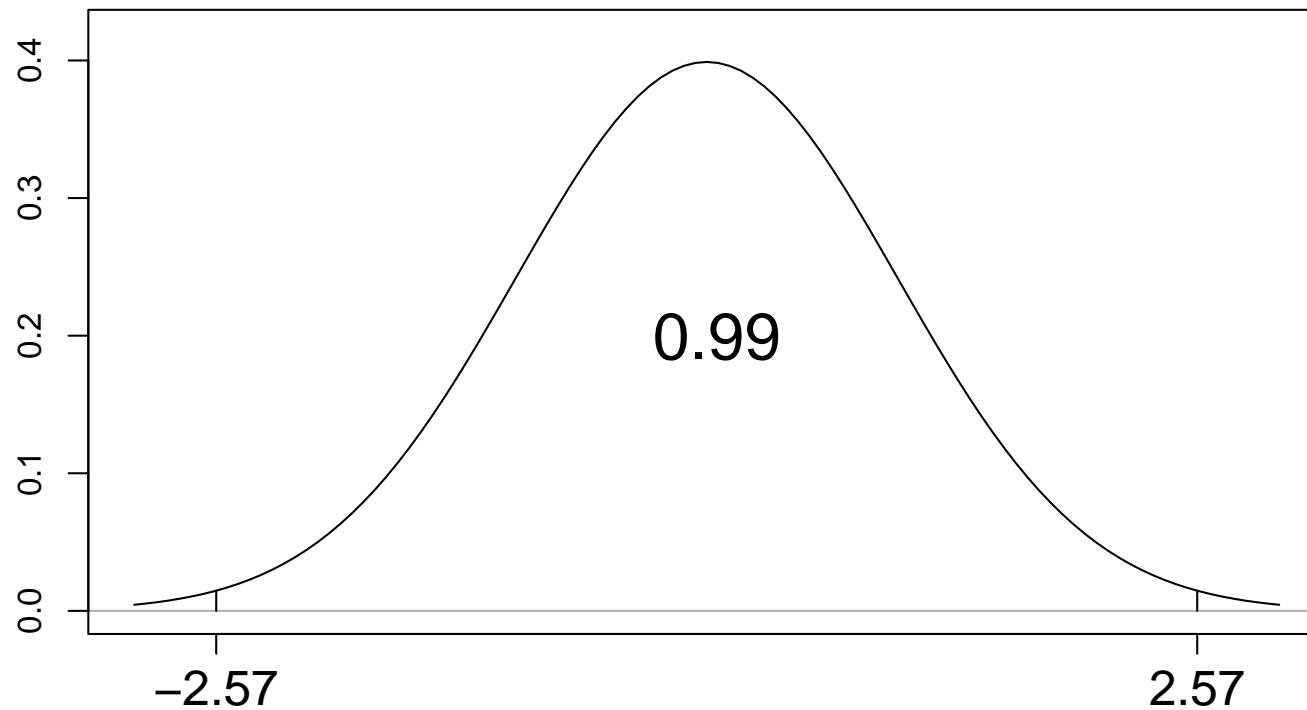
---





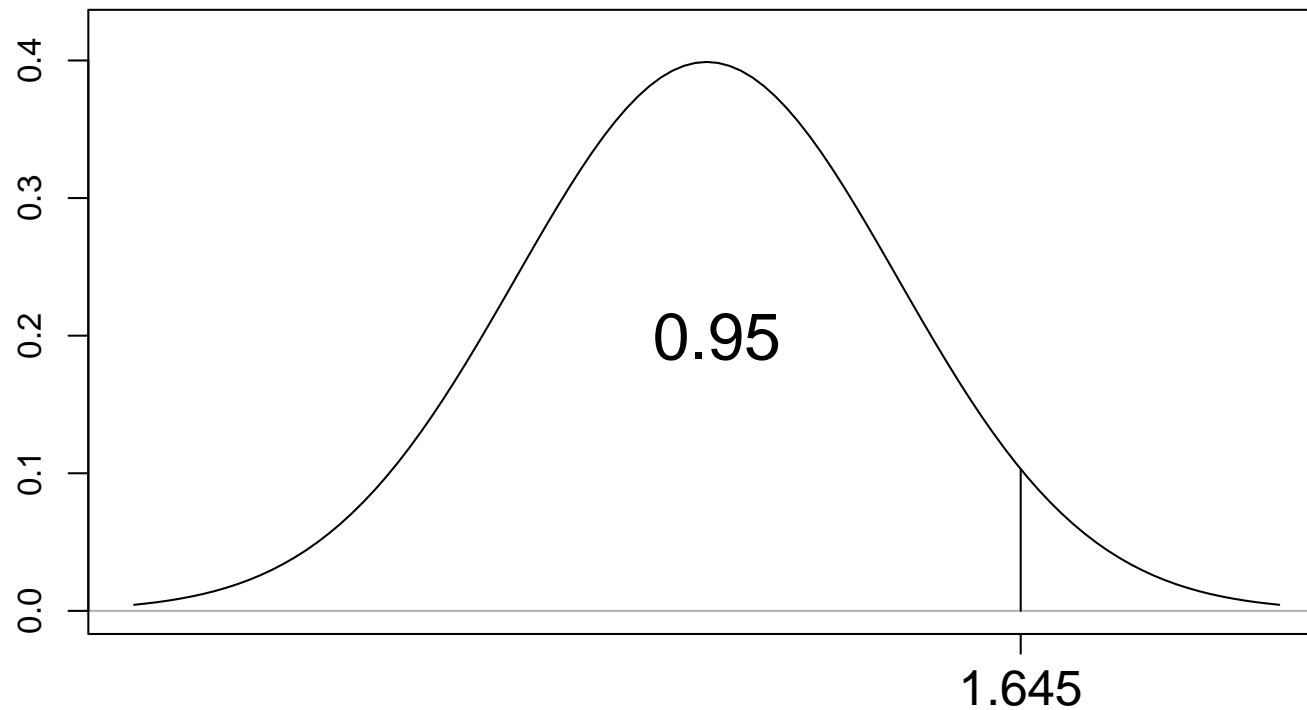
# Normalverteilung

---



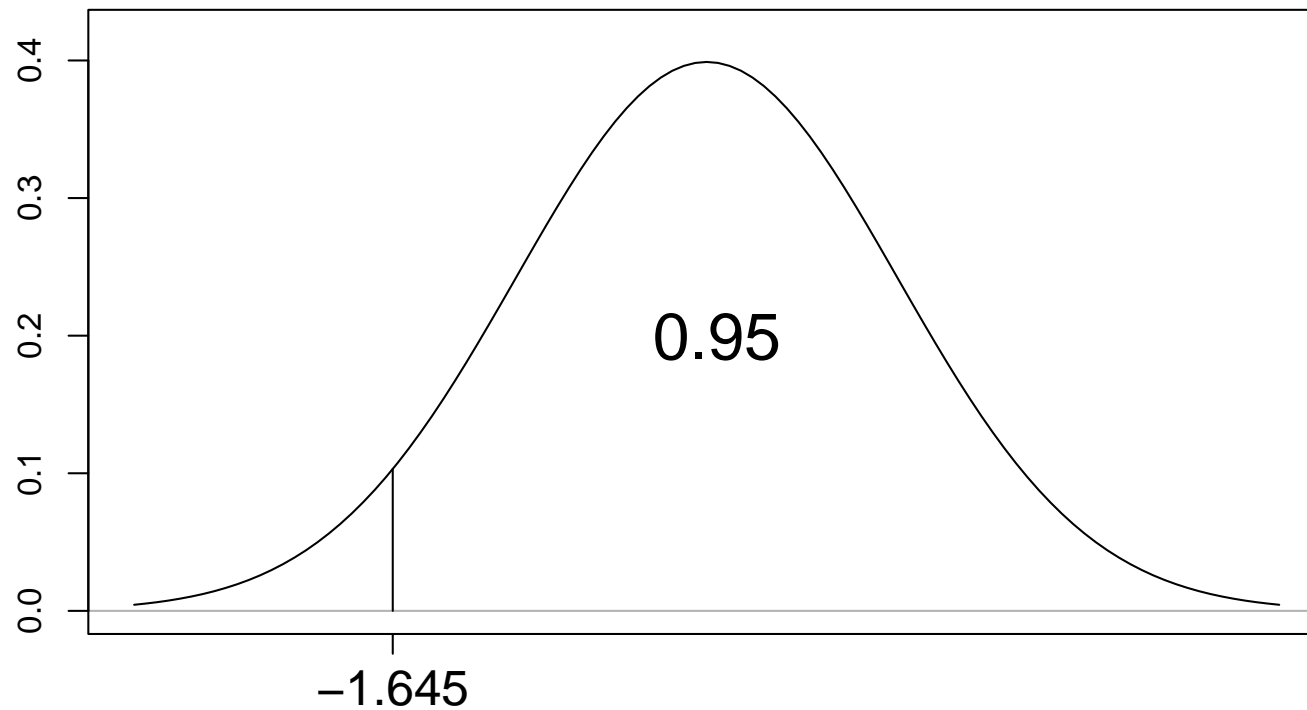
# Normalverteilung

---



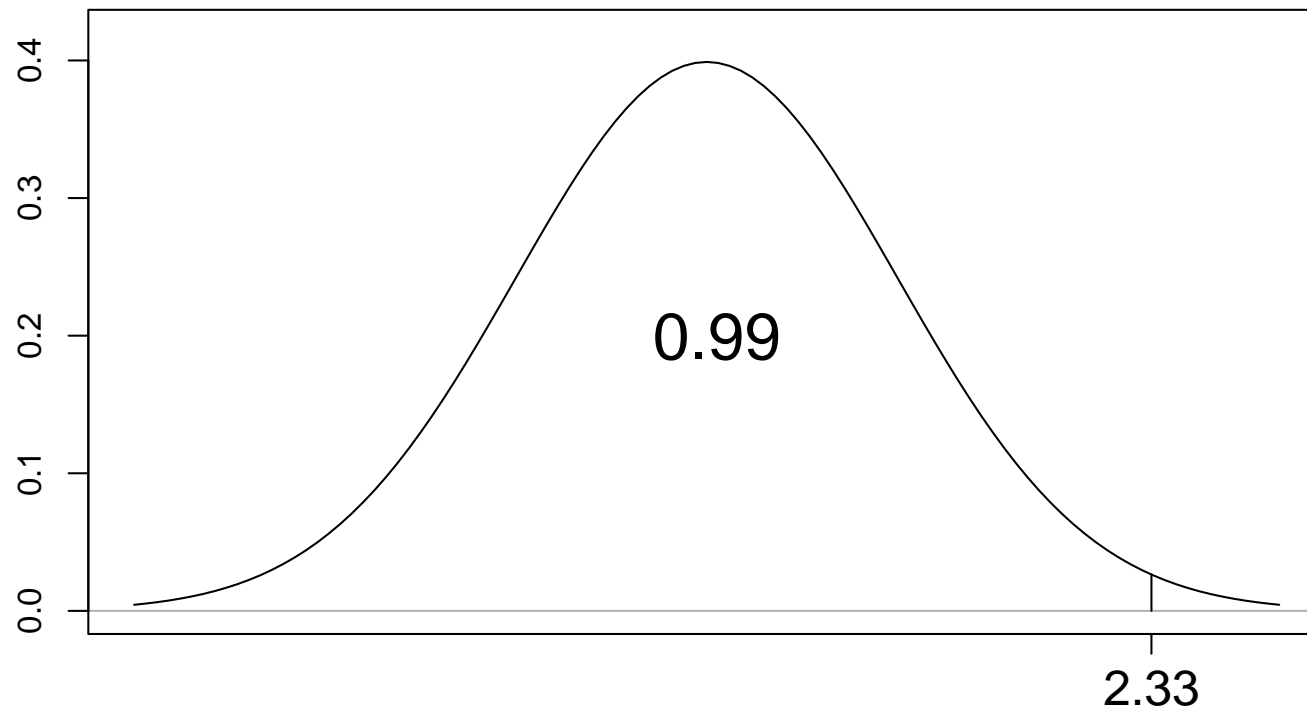
# Normalverteilung

---



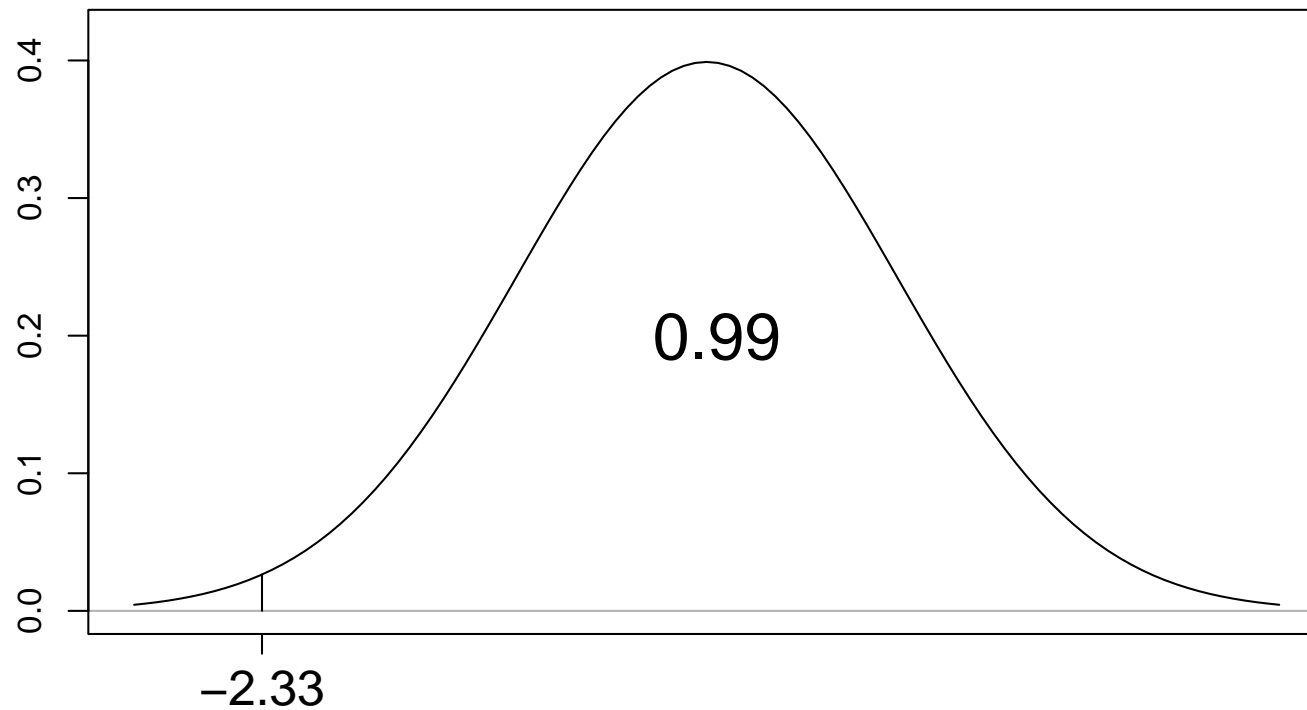
# Normalverteilung

---



# Normalverteilung

---



# Normalverteilung

---

## Beispiel: Aufgabensammlung

**112.** Wie groß ist die Wahrscheinlichkeit, dass eine standardnormalverteilte Zufallsgröße im Intervall  $[-1.25, 1.25]$  liegt?

Aus Tabelle ablesen: 1.25 ist 0.8944-Quantil.

# Normalverteilung

---

## Beispiel: Aufgabensammlung

**132.** Der tägliche Stromverbrauch eines Betriebes (in MWh) ist normalverteilt mit Mittelwert 6.5 und Varianz 4.6. Eine eigene Stromversorgung für diesen Betrieb liefert täglich 8 MWh. Wie groß ist die Wahrscheinlichkeit, dass sich der Betrieb an einem bestimmten Tag selbst versorgen kann?

# Normalverteilung

---

## Beispiel: Aufgabensammlung

**109.** Die Wahrscheinlichkeit, dass eine standardnormalverteilte Zufallsgröße im Intervall  $[-a, a]$  liegt, sei 0.90. Wie lautet  $a$ ?



# Normalverteilung

---

## Beispiel:

Die Wahrscheinlichkeit, dass eine normalverteilte Zufallsgröße mit Mittelwert  $\bar{x} = 5$  und Standardabweichung  $s_x = 1.4$  im Intervall  $[\bar{x} - cs_x, \bar{x} + cs_x]$  liegt, sei 0.90. Wie lautet  $c$ ?

# Normalverteilung

---

## Beispiel: Aufgabensammlung

**133.** Die Müllmenge, die in einem Ortsteil pro Tag anfällt, sei normalverteilt mit Mittel 2.5 und Standardabweichung 0.4 Tonnen. Die Beseitigung des Mülls verursacht tägliche Fixkosten von 2500 GE und variable Kosten von 1600 GE pro Tonne. Welche Kosten werden in 75% der Tage nicht überschritten?

# Normalverteilung

---

## Beispiel: Aufgabensammlung

**134.** Ein Bekannter beklagt sich über sein Einkommen. Er erzählt, daß mindestens 70% seiner Kollegen mehr verdienen als er. Wie viel kann der Bekannte höchstens verdienen, wenn vorausgesetzt wird, daß die Zufallsgröße Einkommen normalverteilt mit Erwartungswert 28000 und Standardabweichung 12000 ist?

# Normalverteilung

---

## Beispiel: Aufgabensammlung

**135.** Die jährlichen Spenden (in Mill.), die auf ein Konto einer karitativen Gesellschaft eingezahlt werden, sind etwa normalverteilt mit Erwartungswert 2.5 und Varianz 2.658. Welches jährliches Spendenaufkommen wird mit 50%-iger Wahrscheinlichkeit überschritten?

# Wölbung

---

Wölbungskoeffizient:

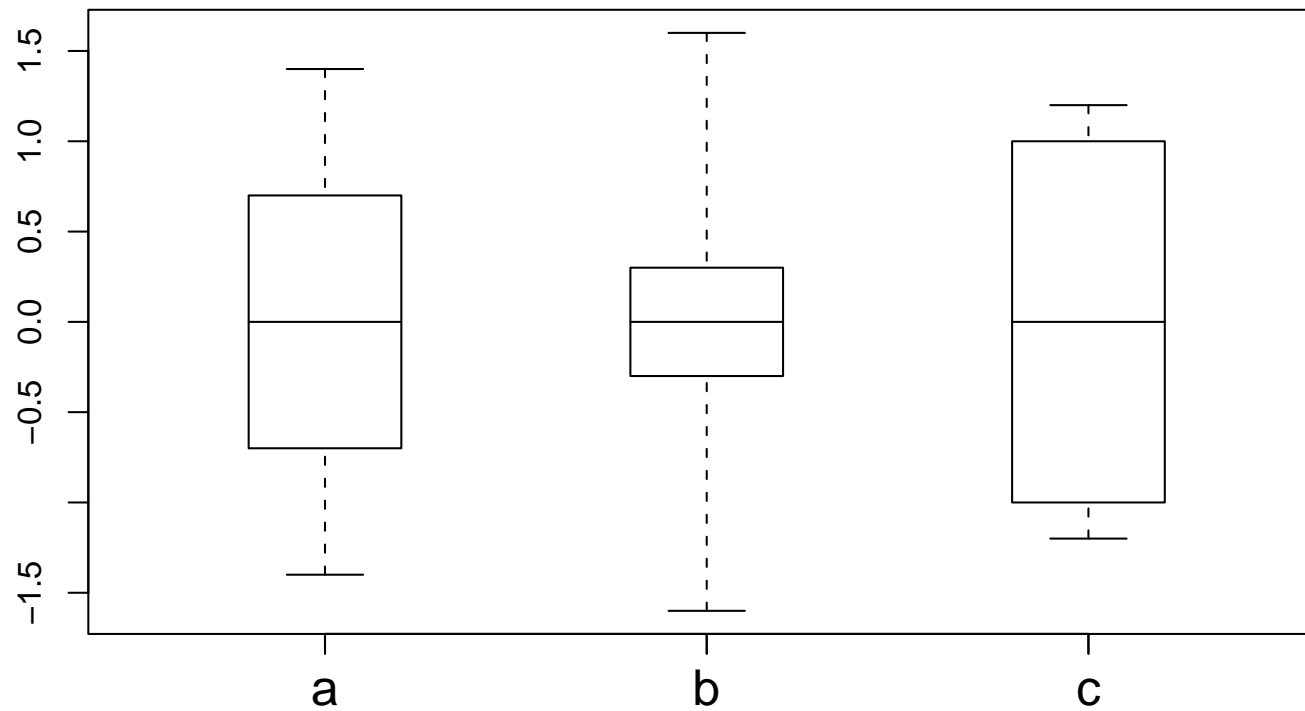
$$\frac{QD}{s_x}$$

Für Normalverteilung ist der Wert gleich 1.34.

- **Lange Enden:**  $< 1.34$ .
- **Kurze Enden:**  $> 1.34$ .

# Wölbung

---



## Zusammenfassung Kapitel 5:

---

- Lagemaße: Mittelwert, Median
- Schiefe
- Streuungsmaße: Varianz, Standardabweichung, QD
- Standardisierung
- Normalverteilung
- Wölbung

# Der Erwartungswert

## Kapitel 6



# Stochastische Grundbegriffe

---

**Zufallsgröße, -variable (ZG)** ist jene numerische Variable  $X$ , die jedem möglichen Ergebnis eines bestimmten Zufallsexperiments einen numerischen Wert  $x$  (= Realisation) zuordnet.

Sie kann einfach, diskret oder stetig sein.

Beispiel: gewürfelte Augenzahl, landwirtschaftlicher Ertrag einer bestimmten Fläche

**Wahrscheinlichkeitsverteilung von  $X$**  ordnet allen durch  $X$  bestimmten Ereignissen Wahrscheinlichkeiten zu. Grafische Darstellungen sind Stabdiagramme für diskrete, Histogramme und Dichtefunktionen für stetige ZG.

# Erwartungswert einer ZG

---

## Empirisches Gesetz der großen Zahlen:

Unabhängige Realisationen  $x_1, x_2, \dots$  einer ZG  $X$ . Wenn der Grenzwert

$$\mu = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n x_i = \lim_{n \rightarrow \infty} \bar{x}$$

existiert, dann nennt man ihn den **Erwartungswert** der ZG  $X$ :  
 $E(X) = \mu$ .

# Erwartungswert einer ZG

---

Der **Erwartungswert** einer ZG  $X$  ist der langfristige Durchschnitt von unabhängigen Realisationen der ZG.

Der **Erwartungswert** einer ZG  $X$  ist eine **Lagemaßzahl** für das Zentrum der Wahrscheinlichkeitsverteilung der ZG.

**Statistische Bestimmung:** Ein **Schätzer** für den Erwartungswert ist:  $\hat{\mu} = \bar{x}$ .

**Mathematische Bestimmung:** Bei bekannter Wahrscheinlichkeitsverteilung kann der Erwartungswert exakt bestimmt werden.

# Erwartungswert einer ZG

---

Der Erwartungswert hat folgende Eigenschaften:

$$E(X + Y) = E(X) + E(Y)$$

$$E(aX + b) = aE(X) + b$$

$$E(X^2) \neq E(X)^2$$

# Varianz einer ZG

---

## Empirisches Gesetz der großen Zahlen:

Unabhängige Realisationen  $x_1, x_2, \dots$  einer ZG  $X$ . Wenn der Grenzwert

$$\sigma^2 = \lim_{n \rightarrow \infty} s_x^2 = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = E((X - \mu)^2)$$

existiert, dann nennt man ihn die **Varianz** der ZG  $X$ :

$$V(X) = \sigma^2.$$

Die **Standardabweichung** von  $X$  ist  $\sqrt{V(X)} = \sigma$ .

# Varianz einer ZG

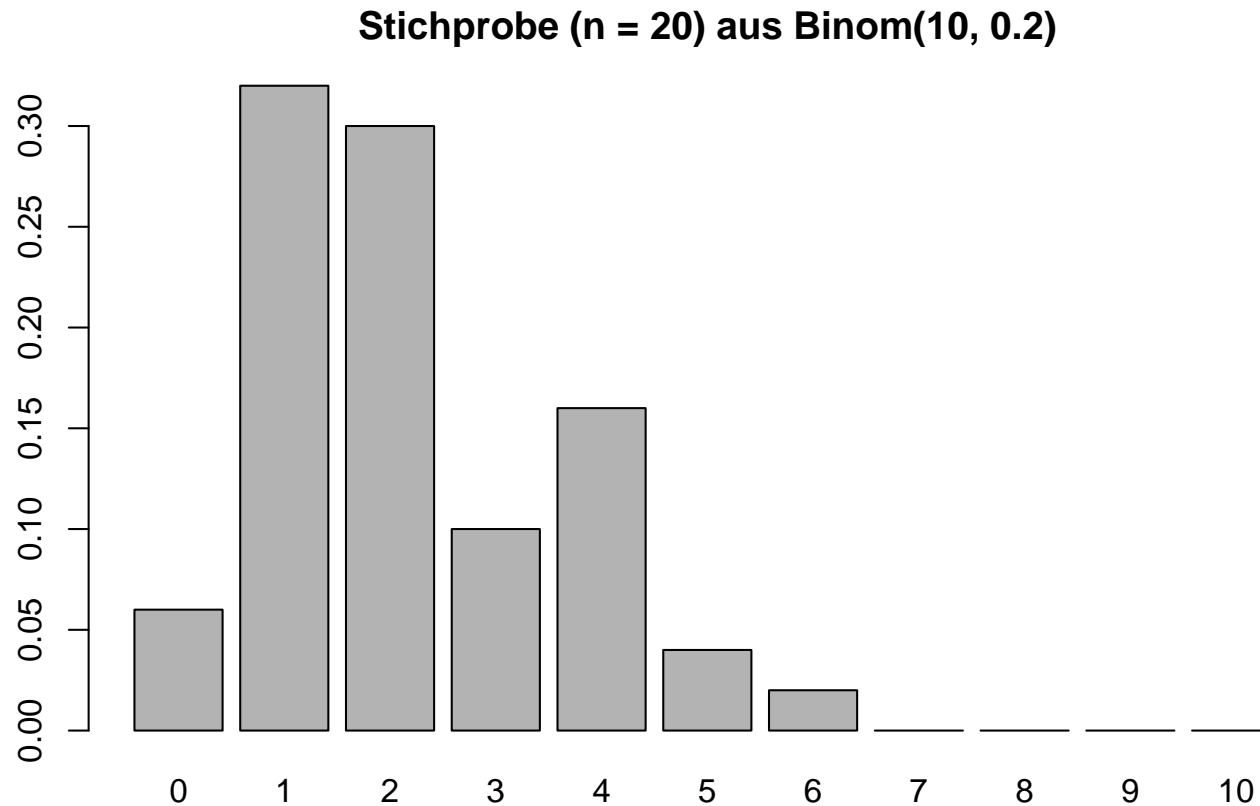
---

Die Varianz hat folgende Eigenschaft:

$$V(aX + b) = a^2 V(X)$$

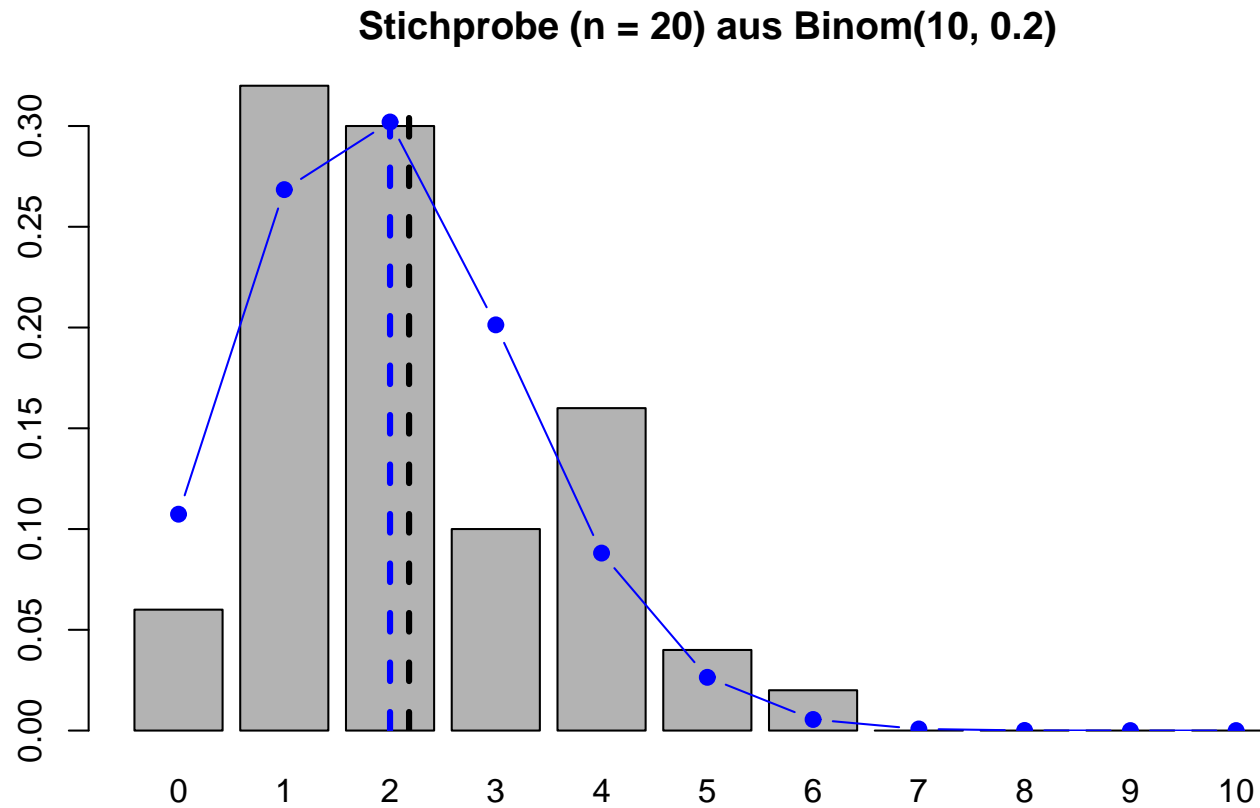
# Varianz einer ZG

---



# Varianz einer ZG

---

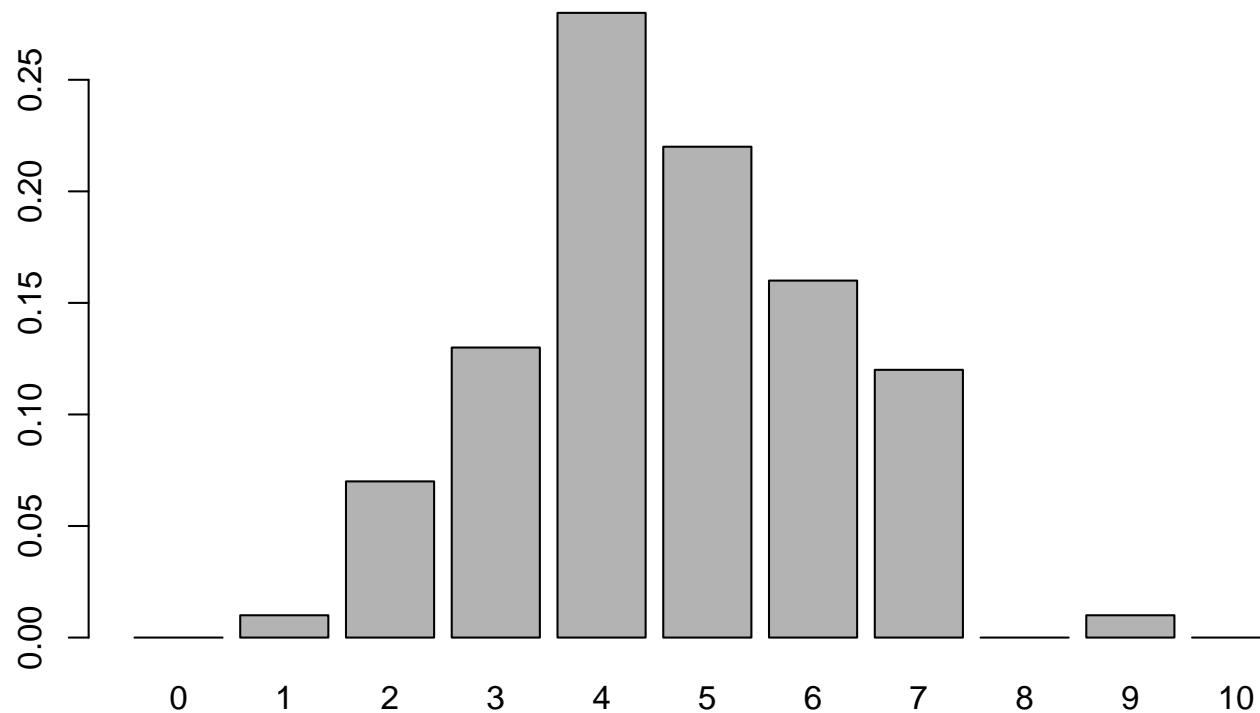




# Varianz einer ZG

---

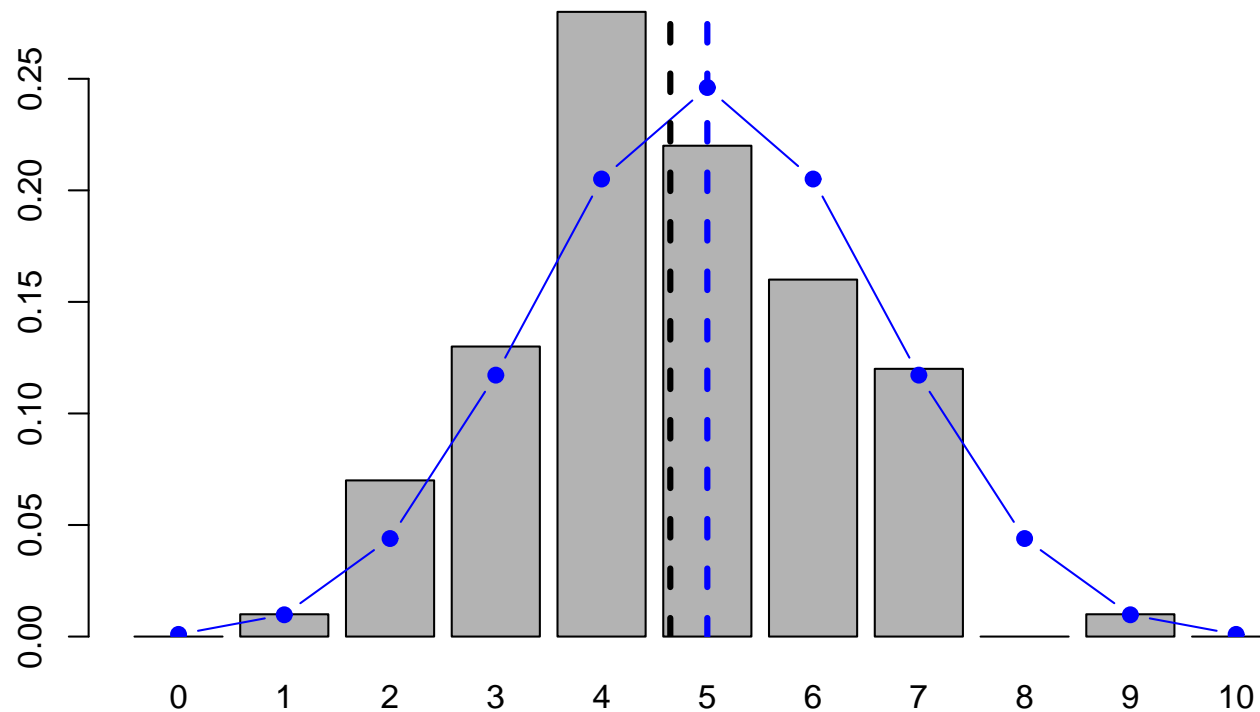
Stichprobe (n = 100) aus Binom(10, 0.5)



# Varianz einer ZG

---

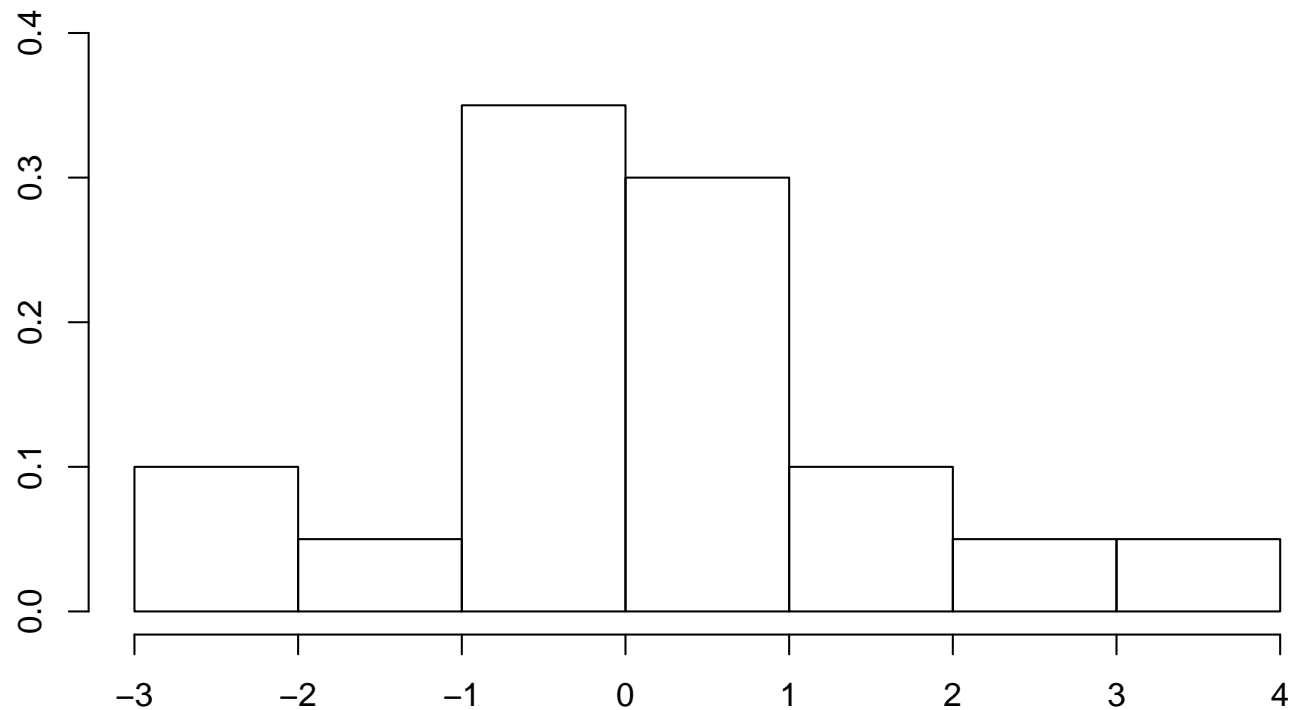
Stichprobe (n = 100) aus Binom(10, 0.5)



# Varianz einer ZG

---

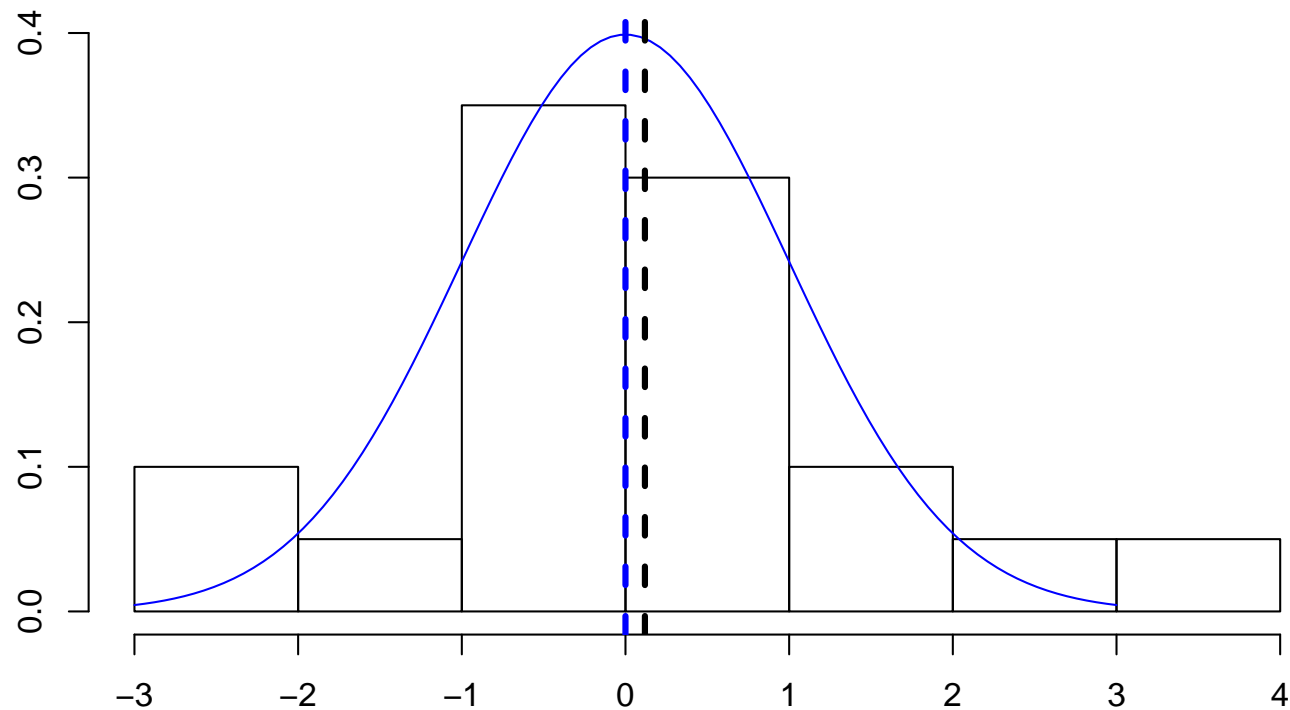
Stichprobe (n = 20) aus  $N(0,1)$



# Varianz einer ZG

---

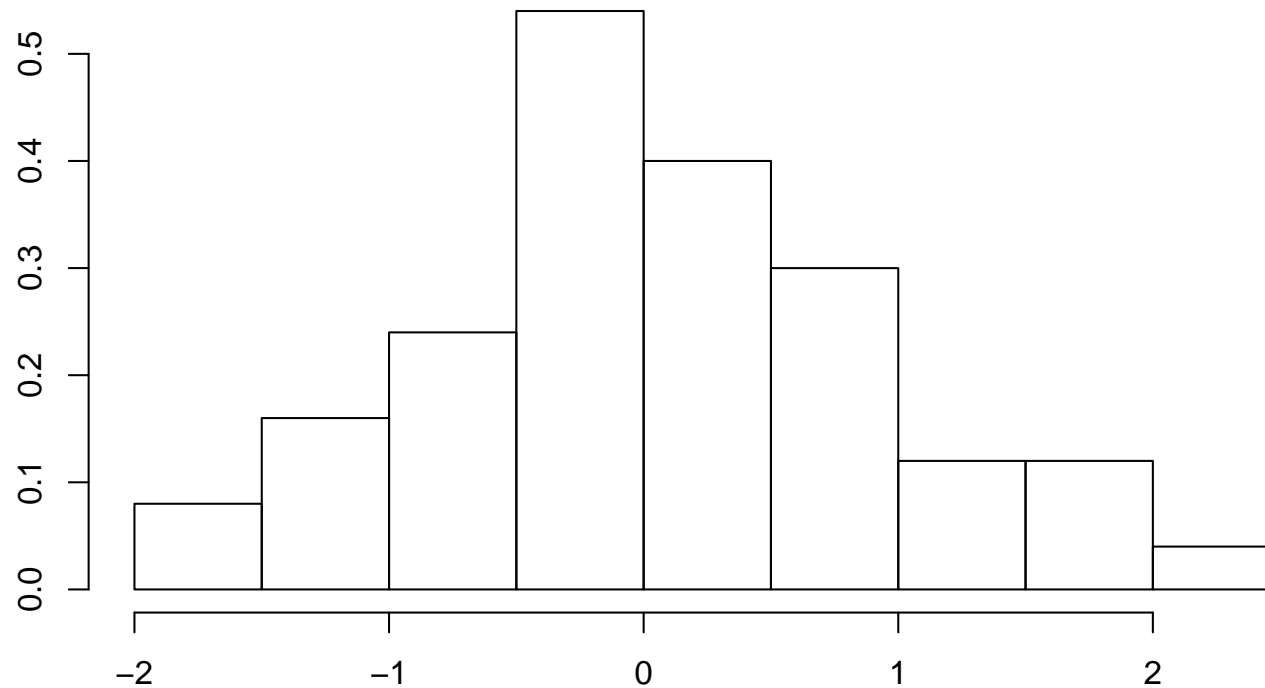
Stichprobe (n = 20) aus  $N(0,1)$



# Varianz einer ZG

---

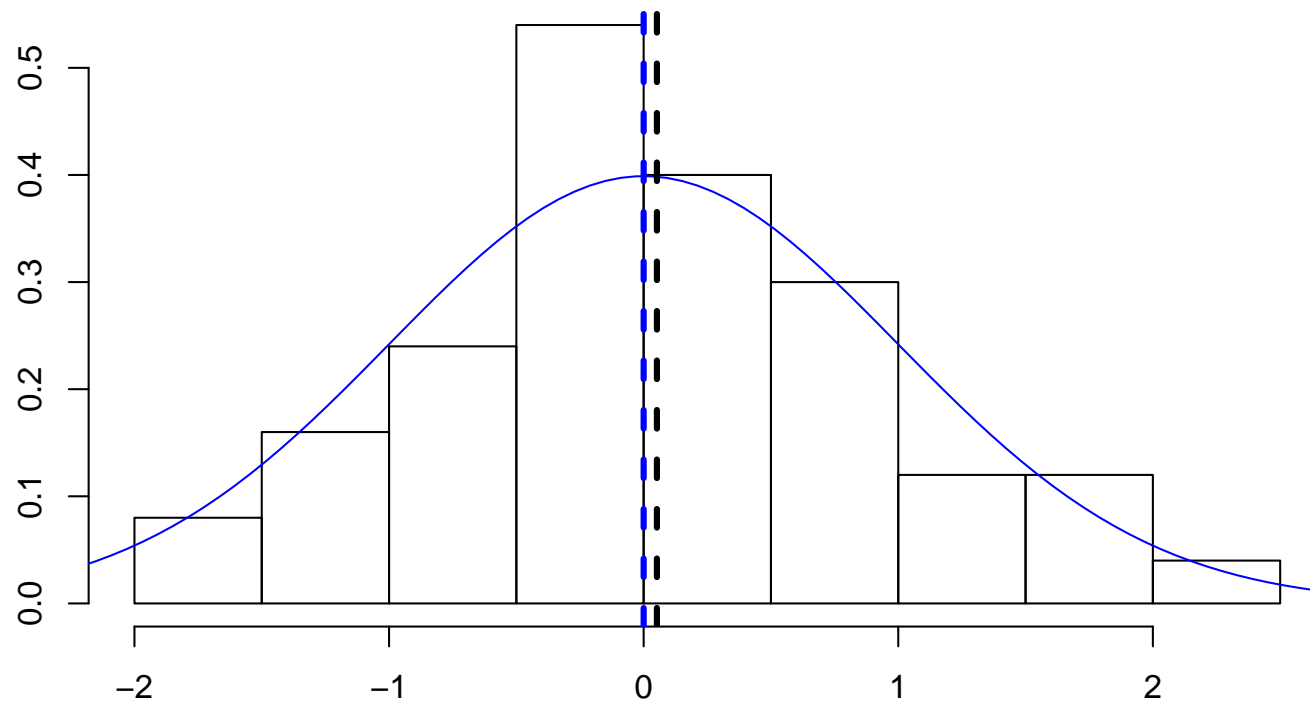
Stichprobe (n = 100) aus  $N(0,1)$



# Varianz einer ZG

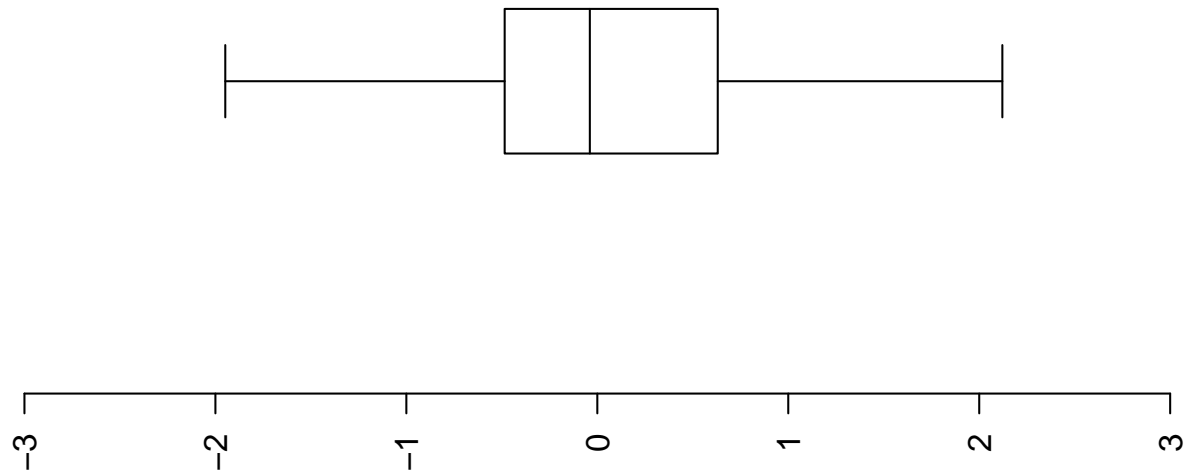
---

Stichprobe (n = 100) aus  $N(0,1)$



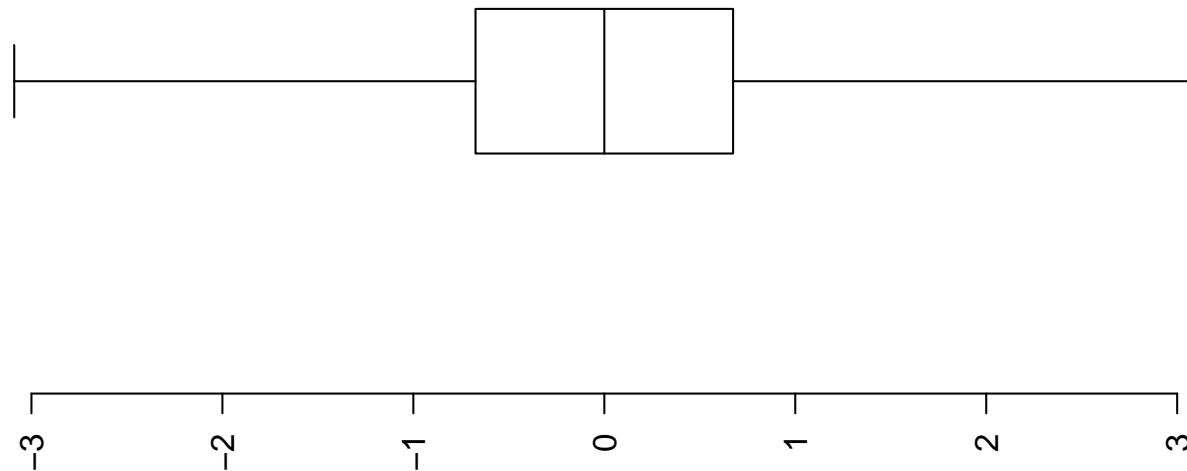
# Varianz einer ZG

---



# Varianz einer ZG

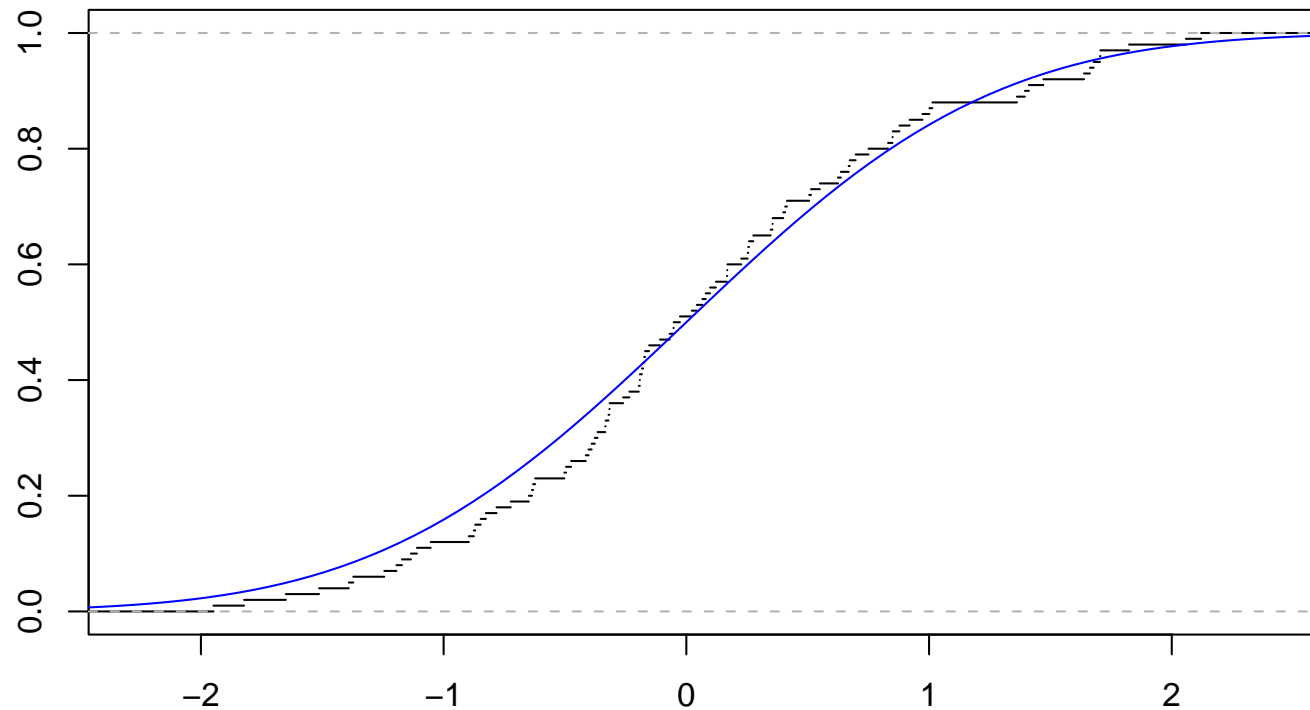
---





# Varianz einer ZG

---



# Zufallsschwankungen des Mittelwerts

---

ZG  $X$  mit Realisationen  $x_1, x_2, \dots, x_n$

Der Mittelwert  $\bar{x}$  der Realisationen schwankt von Stichprobe zu Stichprobe zufällig um den Erwartungswert  $\mu$ .

Der **Mittelwert** ist daher selbst eine **Zufallsgröße**,  
Bezeichnung:  $\bar{X}$ .

- In welcher Beziehung stehen  $\mu$  und  $\bar{X}$ ?
- Wie stark kann  $\bar{X}$  bei gegebenem  $\mu$  schwanken?
- Welche Schlüsse kann man von  $\bar{X}$  über  $\mu$  ziehen?

# Zufallsschwankungen des Mittelwerts

---

Besitzt ZG  $X$  den  $E(X) = \mu$  und die  $V(X) = \sigma^2$ , dann gilt:

$$\begin{aligned} E(\bar{X}) &= \mu \\ V(\bar{X}) &= \frac{\sigma^2}{n} \\ SD &= \frac{\sigma}{\sqrt{n}} \end{aligned}$$

Bei Ziehen ohne Zurücklegen aus einer endlichen Grundgesamt wird wieder die Endlichkeitskorrektur verwendet:

$$SD = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

# Zentraler Grenzwertsatz

---

Besitzt ZG  $X$  den  $E(X) = \mu$  und die  $V(X) = \sigma^2$ , dann ist der **standardisierte Mittelwert**

$$Z = \frac{\bar{X} - \mu}{SD}$$

annähernd **standardnormalverteilt**.

Die Genauigkeit dieser Aussage steigt mit dem Stichprobenumfang  $n$ .

# Anwendung des zentralen Grenzwertsatzes

---

Gegeben sei die ZG  $X$  mit  $E(X) = \mu$  und  $V(X) = \sigma^2$ .

Man kann die Wahrscheinlichkeit bestimmen, mit der ein Mittelwert  $\bar{X}$  in einem vorgegebenen Intervall liegt:

$$\begin{aligned} P(A \leq \bar{X} \leq B) &= P\left(\frac{A - \mu}{SD} \leq \frac{\bar{X} - \mu}{SD} \leq \frac{B - \mu}{SD}\right) \\ &\approx \Phi\left(\frac{B - \mu}{SD}\right) - \Phi\left(\frac{A - \mu}{SD}\right) \end{aligned}$$

# Anwendung des zentralen Grenzwertsatzes

---

## Beispiel: Skript

**(6.19)** Eine Firma produziert Kaffeepackungen mit je 250 g Inhalt und einer Standardabweichung  $\sigma = 10$  g. Die Qualitätskontrolle überprüft Kartons, die jeweils 20 Packungen enthalten. Kartons mit einem Gewicht von 4.9 – 5.1 kg werden ausgeliefert.

Mit welchem Anteil an nicht ausgelieferten Kartons muss gerechnet werden?

# Prognoseintervall

---

einseitig

$$P(\bar{X} \leq \mu + N_{\alpha} SD) \approx \alpha$$

$$P(\bar{X} \geq \mu - N_{\alpha} SD) \approx \alpha$$

beidseitig

$$P(|\bar{X} - \mu| \leq N_{(1+\alpha)/2} SD) \approx \alpha$$

$$P(\mu - N_{(1+\alpha)/2} SD \leq \bar{X} \leq \mu + N_{(1+\alpha)/2} SD) \approx \alpha$$

$$SD = \frac{\sigma}{\sqrt{n}}$$

# Prognoseintervall

---

**einseitig**      $P(\bar{X} \leq \mu + 1.64 SD) \approx 0.95$

$$P(\bar{X} \geq \mu - 1.64 SD) \approx 0.95$$

**beidseitig**      $P(|\bar{X} - \mu| \leq 1.96 SD) \approx 0.95$

**Faustregel**      $P(|\bar{X} - \mu| \leq 2 SD) \approx 0.95$

$$SD = \frac{\sigma}{\sqrt{n}}$$



# Prognoseintervall

---

## Beispiel: Aufgabensammlung

**167.–170.** Die Behörde führt Versuchsreihen mit je 50 Motoren durch und stuft sie ab einem Verbrauch von 10 l als umweltbelastend ein. Nur den Produzenten ist bekannt:

Motortyp	$E(X)$	$V(X)$
Tigerbaby	9.5	9.2
Brumm	12	10.2
Ökomot	8	16.4
Velocità	11	16.5

In welchem Bereich kann der Produzent das Ergebnis einer Versuchsreihe mit 95%-iger Sicherheit erwarten?

# Schätzung des Erwartungswerts

---

Der Erwartungswert  $\mu$  ist jetzt unbekannt und soll aus Daten geschätzt werden.

Der beste **Schätzer** für einen unbekannten Erwartungswert  $\mu$  ist bei Fehlen weiterer zusätzlicher Information der Mittelwert der Daten:

$$\hat{\mu} = \bar{x}.$$

**Frage:** Wie genau ist die Schätzung  $\hat{\mu}$ ?

→ **Konfidenzintervalle**

# Konfidenzintervalle

---

Das **Konfidenzintervall (KI)** für den unbekannten Erwartungswert  $E(X) = \mu$  lautet:

$$P(\bar{X} - N_{(1+\alpha)/2} SD \leq \mu \leq \bar{X} + N_{(1+\alpha)/2} SD) \approx \alpha$$

Die **statistische Sicherheit** wird durch die Überdeckungswahrscheinlichkeit  $\alpha$  bestimmt. Unter der **Genauigkeit** des KI versteht man seine Länge.

**Problem:**  $SD$  hängt von der unbekannten Varianz  $\sigma^2$  ab.

# Schätzung der Varianz

---

In den meisten Anwendungen ist die Varianz  $\sigma^2$  der ZG  $X$  unbekannt. Sie wird daher durch die **Varianz der Daten** geschätzt:

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

Da die Varianz der Daten  $s_x^2$  die wahre Varianz  $\sigma^2$  systematisch zu klein schätzt, verwendet man als Schätzer in Anwendungen die **Stichprobenvarianz**:

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

# Schätzung des Konfidenzintervalls

---

Wir verwenden für die Schätzung der Standardabweichung in der Formel des KI die Stichprobenstandardabweichung:

$$SD = \frac{\sigma}{\sqrt{n}} \approx \widehat{SD} = \frac{s_{n-1}}{\sqrt{n}}.$$

Das KI lautet dann:

$$\bar{x} - N_{(1+\alpha)/2} \frac{s_{n-1}}{\sqrt{n}} \leq \mu \leq \bar{x} + N_{(1+\alpha)/2} \frac{s_{n-1}}{\sqrt{n}}.$$

# Schätzung des Konfidenzintervalls

---

## Beispiel: Aufgabensammlung

**165.** Unbezahlt gebliebene Rechnungen in 1000 GE wurden in folgenden Branchen erfasst:

Branche	Anzahl überprüfter Betriebe	Mittelwert $\bar{x}$	Varianz $s_x^2$
Gastgewerbe	36	35.95	54.2
Bauwirtschaft	30	52.39	108.7
Metall	28	38.91	89.3
Uhren u. Schmuck	32	4.67	47.3
Leder u. Pelze	24	12.63	51.2

Geben sie 95% KI an (Faustregel).

# Testen von Hypothesen über $\mu$

---

**Zweiseitiges Testproblem:** Nullhypothese:  $\mu = \mu_0$   
Alternative:  $\mu \neq \mu_0$

**Linksseitiges Testproblem:** Nullhypothese:  $\mu \leq \mu_0$   
Alternative:  $\mu > \mu_0$

**Rechtsseitiges Testproblem:** Nullhypothese:  $\mu \geq \mu_0$   
Alternative:  $\mu < \mu_0$

# Durchführung von Tests

---

Die Nullhypothese wird unterstellt. Man kann wieder folgende äquivalente Methoden zum Testen wählen:

**A. Prognoseintervalle**

**B. Standardisierter Mittelwert** als Testgröße:

$$z = \frac{\bar{x} - \mu}{SD} \quad \text{bzw.} \quad z = \frac{\bar{x} - \mu}{\widehat{SD}}$$

**C.  $P$ -Werte**



# Durchführung von Tests

---

## Kritische Werte für die Testgrößen für 95% Niveau:

**zweiseitig**      $-2 \leq z \leq 2 \Rightarrow H : \mu = \mu_0$  wird beibehalten  
                     $z < -2 \Rightarrow \mu$  ist signifikant kleiner als  $\mu_0$   
                     $z > +2 \Rightarrow \mu$  ist signifikant größer als  $\mu_0$

**linksseitig**      $z \leq 1.64 \Rightarrow H : \mu \leq \mu_0$  wird beibehalten  
                     $z > 1.64 \Rightarrow \mu$  ist signifikant größer als  $\mu_0$

**rechtsseitig**    $z \geq -1.64 \Rightarrow H : \mu \geq \mu_0$  wird beibehalten  
                     $z < -1.64 \Rightarrow \mu$  ist signifikant kleiner als  $\mu_0$

# Durchführung von Tests

---

Kritische Werte als Quantile von  $N(0, 1)$

Signifikanzniveau	einseitige Fragestellung	zweiseitige Fragestellung
90%	$\pm 1.28$	$\pm 1.64$
95%	$\pm 1.64$	$\pm 1.96$
99%	$\pm 2.33$	$\pm 2.58$

# Durchführung von Tests

---

## Beispiel: Skript

**(6.28)** Zwanzig Bewerber für eine Spezialausbildung unterziehen sich einem Intelligenztest, der auf den Mittelwert 100 und die Standardabweichung 10 in der Gesamtbevölkerung geeicht ist.

Ihre IQ-Scores lauten:

110	95	103	127	90
93	130	97	111	109
89	115	114	122	91
88	103	105	101	111

Unterscheidet sich ihr IQ signifikant vom Mittelwert 100?

# Durchführung von Tests

---

## Beispiel: Aufgabensammlung

**142.** Unbezahlt gebliebene Rechnungen in 1000 GE wurden in folgenden Branchen erfasst:

Branche	Anzahl überprüfter Betriebe	Mittelwert $\bar{x}$	Varianz $s_x^2$
Gastgewerbe	36	35.95	54.2
Bauwirtschaft	30	52.39	108.7
Metall	28	38.91	89.3
Uhren u. Schmuck	32	4.67	47.3
Leder u. Pelze	24	12.63	51.2

# Durchführung von Tests

---

Verschiedene Hypothesen über die Höhe an unbezahlt gebliebenen Rechnungen sollen getestet werden (Fehler 1. Art: 5%). In der folgenden Tabelle sind Vergleichswerte  $\mu_0$  für die verschiedenen Branchen vorgegeben.

Branche	$\mu_0$	Testgröße
Gastgewerbe	30 GE	4.781
Bauwirtschaft	50 GE	
Metall	40 GE	-0.559
Uhren u. Schmuck	10 GE	-4.315
Leder u. Pelze	10 GE	

# Durchführung von Tests

---

Verschiedene Hypothesen über die Höhe an unbezahlt gebliebenen Rechnungen sollen getestet werden (Fehler 1. Art: 5%). In der folgenden Tabelle sind Vergleichswerte  $\mu_0$  für die verschiedenen Branchen vorgegeben.

Branche	$\mu_0$	Testgröße
Gastgewerbe	30 GE	4.781
Bauwirtschaft	50 GE	1.234
Metall	40 GE	-0.559
Uhren u. Schmuck	10 GE	-4.315
Leder u. Pelze	10 GE	1.763

## Durchführung von Tests

---

- (a) Im Gastgewerbe kann nachgewiesen werden, dass der Erwartungswert von  $X$  kleiner als 30 GE ist.
- (b) In der Bauwirtschaft kann keine Entscheidung darüber getroffen werden, ob der Erwartungswert ungleich 50 GE ist.
- (c) Beim Metall kann nachgewiesen werden, dass der Erwartungswert von  $X$  größer als 40 GE ist.
- (d) Bei Uhren und Schmuck kann nachgewiesen werden, dass der Erwartungswert von  $X$  größer als 10 GE ist.
- (e) Bei Leder und Pelze kann keine Entscheidung darüber getroffen werden, ob der Erwartungswert ungleich 10 GE ist.

# ANOVA

---

**Streuungszerlegung = Varianzanalyse = ANalysis Of VAriance:**

Die **totale Abweichung** wird in die Komponente der **systematischen Abweichung** und in die Komponente der **zufälligen Streuung** zerlegt.

$$SS_T = SS^* + SS_R$$

Systematische Abweichung	$SS^* = \sum_{i=1}^n (\bar{x} - \mu_0)^2 = n(\bar{x} - \mu_0)^2$
Zufällige Streuung	$SS_R = \sum_{i=1}^n (x_i - \bar{x})^2 = (n - 1)s_{n-1}^2$
Totale Abweichung	$SS_T = \sum_{i=1}^n (x_i - \mu_0)^2$



# ANOVA-Tabelle

---

	$SS$	$df$	$MSS$
*	$SS^*$	1	$MSS^* = \frac{SS^*}{1}$
$R$	$SS_R$	$n - 1$	$MSS_R = \frac{SS_R}{n-1}$
	$SS_T$	$n$	

$df$  = **d**egrees of **f**reedom = Anzahl der Freiheitsgrade

$MSS$  = mittlere Quadratsummen

# ANOVA-Tabelle

---

	$SS$	$df$	$MSS$
$*$	$\sum_{i=1}^n (\bar{x} - \mu_0)^2$	1	$n(\bar{x} - \mu_0)^2$
$R$	$\sum_{i=1}^n (x_i - \bar{x})^2$	$n - 1$	$s_{n-1}^2$
	$SS_T$	$n$	

$df$  = **d**egrees of **f**reedom = Anzahl der Freiheitsgrade

$MSS$  = mittlere Quadratsummen

# ANOVA

---

Zweiseitige Tests über den Erwartungswert können schematisch als Varianzanalyse durchgeführt werden.

Die Testgröße lautet

$$T = z = \frac{\bar{x} - \mu_0}{SD} = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s_{n-1}}$$

und ihr Quadrat lautet daher

$$z^2 = \frac{n(\bar{x} - \mu_0)^2}{s_{n-1}^2} = \frac{MSS^*}{MSS_R}$$

# ANOVA

---

Aus den Komponenten der Varianzanalyse kann daher leicht das **Signifikanzproblem**, d.h. unser bisheriges Testproblem mit der Hypothese  $H : \mu = \mu_0$ , beantwortet werden. Es wird die  $F$ -**Größe** berechnet:

$$F = \frac{SS^*}{\frac{SS_R}{n-1}} = \frac{MSS^*}{MSS_R} = z^2$$

Auf dem 95%-Niveau wird die Hypothese für  $F \leq 4$  angenommen. Für  $F > 4$  liegt ein signifikanter Wert vor und die Hypothese wird verworfen. (Faustregel)

# ANOVA

---

Beim **Relevanzproblem** geht es darum, den Anteil der systematischen Abweichung an der Gesamtstreuung anzugeben. Die Größe

$$\frac{SS^*}{SS_T}$$

heisst **Bestimmtheitsmaß**.

Durch das Bestimmtheitsmaß kann man ausrechnen, wieviel Prozent der Gesamtstreuung durch die systematische Abweichung erklärt werden kann.

# ANOVA

---

## Beispiel: Skript

**(6.38)** Fortsetzung (6.28): Zwanzig Bewerber für eine Spezialausbildung unterziehen sich einem Intelligenztest. Ihre IQ-Scores lauten:

110	95	103	127	90
93	130	97	111	109
89	115	114	122	91
88	103	105	101	111

Unterscheidet sich der IQ der Bewerber signifikant vom Mittelwert 100 in der Gesamtbevölkerung? Geben sie den Anteil der systematischen Abweichung an der Gesamtstreuung an.

# ANOVA

---

## Beispiel: Aufgabensammlung

**152.** Eine Pizzeria-Kette möchte in ihren Filialen durch freundliche Bedienung einen mittleren Rechnungswert von 200 GE erzielen.

Filiale	Capri	Gino	Baracca	Pescatore	Fiducia
Anzahl der Gäste	50	41	54	63	24
Mittelwert $\bar{x}$	180	205	210	210	225
Varianz $s_x^2$	446	556	640	640	480

Wieviel Prozent der Abweichung vom vorgegebenen Wert 200 GE lässt sich bei der Filiale Capri systematisch erklären?

## $t$ -Test und $F$ -Test

---

Wenn man keine zusätzlichen Informationen über die den Daten zugrunde liegende Verteilung hat, wendet man beim Testen den zentralen Grenzwertsatz an und berechnet kritische Werte aus der Standardnormalverteilung.

Wenn man aber unterstellt, dass die empirischen Daten exakt normalverteilt sind, kann man beim Testen exakte kritische Werte bestimmen. Die Testgröße  $z$  ist dann  $t$ -verteilt ( $t$ -**Test**) und die  $F$ -Größe ist  $F$ -verteilt ( $F$ -**Test**). Die kritischen Werte erhält man daher aus der  $t$ - bzw. aus der  $F$ -Verteilung.



# Zusammenfassung Kapitel 6

---

- Stochastische Grundbegriffe: Zufallsgröße, Verteilung
- Erwartungswert und Varianz einer ZG
- Zufallsschwankungen des Mittelwerts, Prognoseintervalle
- Konfidenzintervalle
- Testen von Hypothesen
- Varianzanalyse, ANOVA

# Regression und Korrelation

## Kapitel 7

# Überblick

---

## bisher:

- univariate Fragestellungen  
→ Eigenschaften der Verteilung einer einzigen Variablen

## jetzt:

- bivariate Fragestellungen
- Abhängigkeit, Koppelung  
→ Korrelation
- Prognose  
→ Regression

# Streudiagramm

---

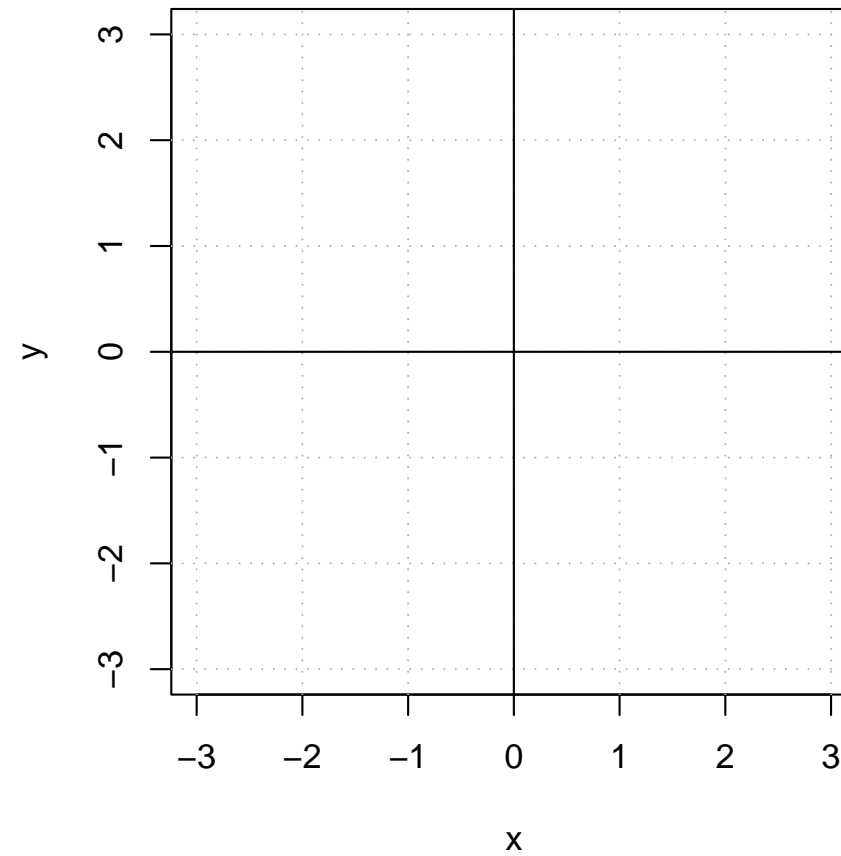
Bivariate Datenliste:

**Liste von Datenpaaren**  $(x_i, y_i), i = 1, \dots, n$ .

$i$	$x$	$y$
1	-0.61	-0.59
2	-1.25	0.34
3	1.45	-1.09
4	-0.10	1.16
5	0.71	1.56

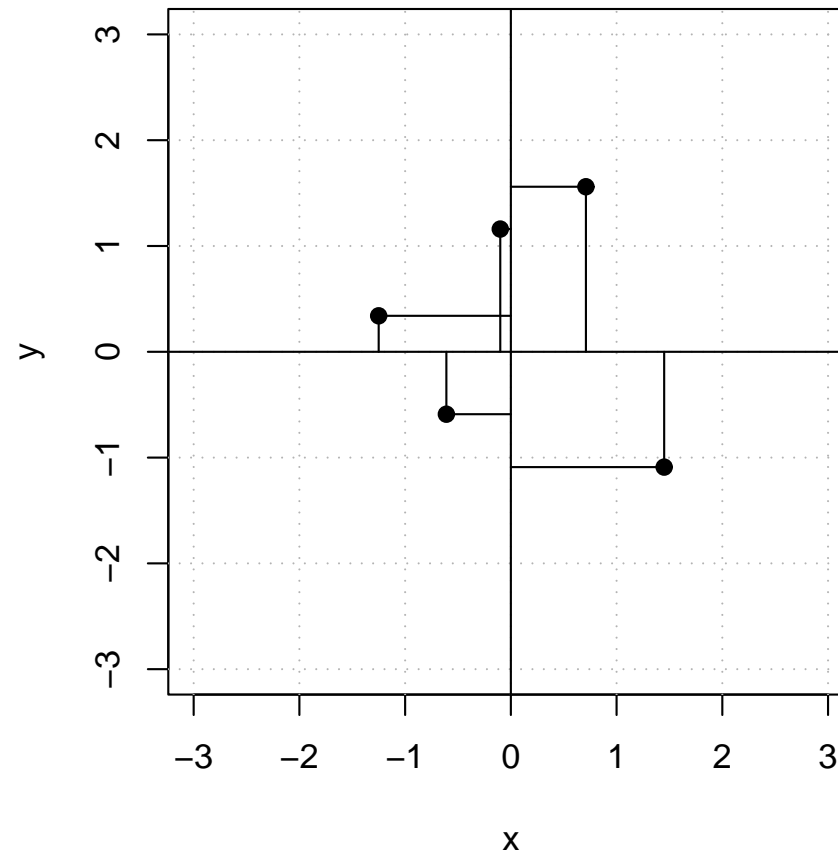
# Streudiagramm

---



# Streudiagramm

---



# Streudiagramm

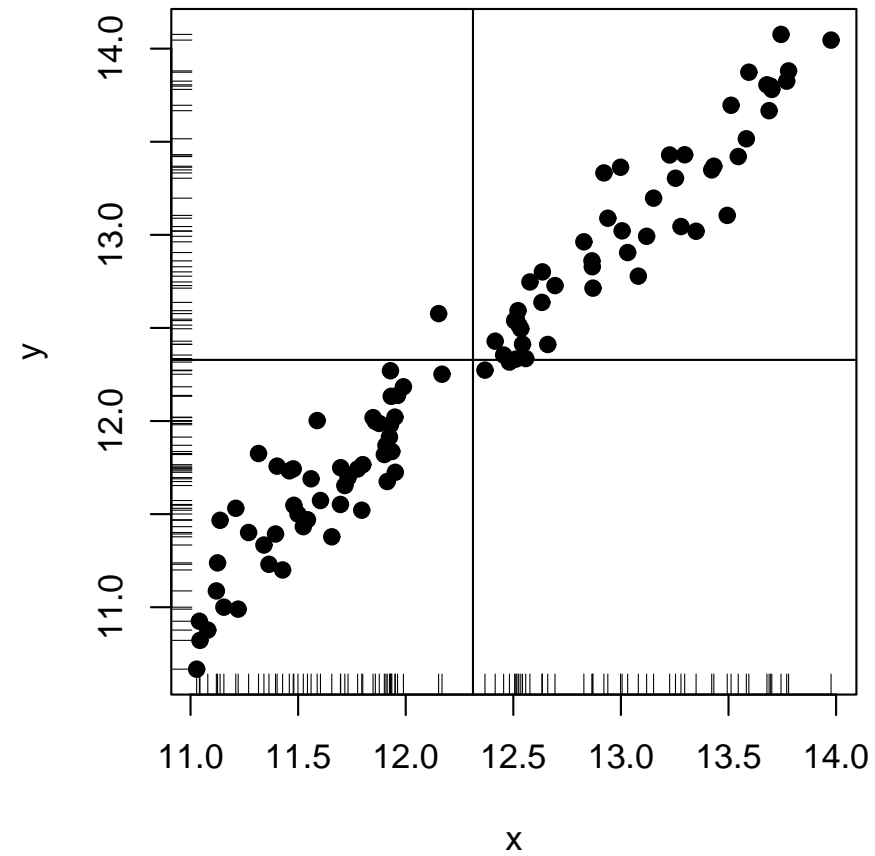
---

## Eigenschaften des Streudiagramms

- $(\bar{x}, \bar{y})$  ist der Mittelpunkt der Punktwolke
- Projektion der Punktwolke auf die  $x$ -Achse ergibt das Punktediagramm der Datenliste  $x_1, \dots, x_n$ .
- Projektion der Punktwolke auf die  $y$ -Achse ergibt das Punktediagramm der Datenliste  $y_1, \dots, y_n$ .

# Streudiagramm

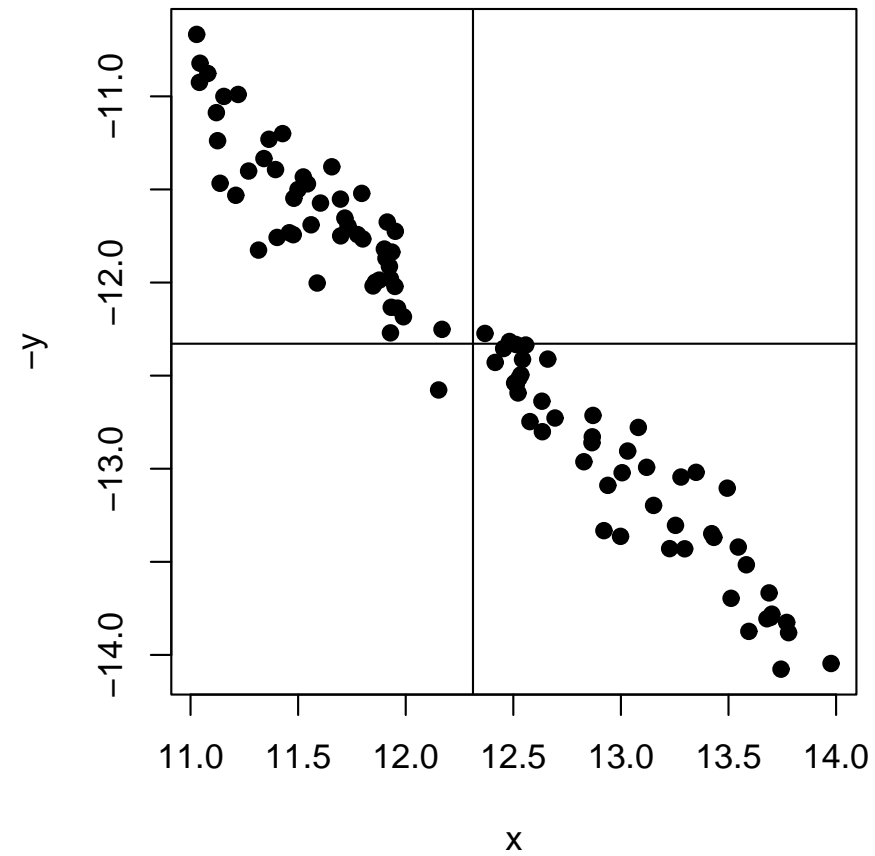
---





# Streudiagramm

---



# Streudiagramm

---

**Explorative Statistik:** Beurteilung, ob ein im Streudiagramm auftretendes Muster interessant ist, durch das menschliche Auge.

**Frage:** Ist das beobachtete Muster wirklich ungewöhnlich?

**Antwort:** Randomisierung der Zuordnung von  $x$  und  $y$  Werten.

**Problem:** Rechenintensiv.

**Ausweg:** Statistische Masszahlen für bestimmte Typen von Kopplung.

# Monotone Koppelung

---

$X$  und  $Y$  sind **positiv gekoppelt**, wenn *tendenziell* gilt

- je größer  $x$ , desto größer  $y$
- je kleiner  $x$ , desto kleiner  $y$

$X$  und  $Y$  sind **negativ gekoppelt**, wenn *tendenziell* gilt

- je größer  $x$ , desto kleiner  $y$
- je kleiner  $x$ , desto größer  $y$

# Korrelationskoeffizient

---

Betrachte Standard-Scores:

$$z_x = \frac{x - \bar{x}}{s_x}$$
$$z_y = \frac{y - \bar{y}}{s_y}$$

Es gilt:

- $z_x \cdot z_y > 0$ : Der Punkt  $(x, y)$  liegt im 1. oder 3. Quadranten
- $z_x \cdot z_y < 0$ : Der Punkt  $(x, y)$  liegt im 2. oder 4. Quadranten

# Korrelationskoeffizient

---

Berücksichtige zusätzlich zum Vorzeichen die Größe von  $z_x \cdot z_y$ .

Der **Korrelationskoeffizient** ist definiert als

$$r_{xy} = \frac{1}{n} (z_{x_1} z_{y_1} + \cdots + z_{x_n} z_{y_n})$$

Es gilt:

- symmetrisch:  $r_{xy} = r_{yx}$
- identische Datenlisten haben die Korrelation 1:  $r_{xx} = 1$
- $-1 \leq r_{xy} \leq 1$

# Korrelationskoeffizient

---

Der **Korrelationskoeffizient**  $r_{xy}$  ist eine Maßzahl für die Korrelation einer Punktwolke:

- Das Vorzeichen gibt die **Richtung** der Korrelation an.
- Der Absolutbetrag  $|r|$  gibt die **Stärke** der Bindung an die Hauptachse an.

# Korrelationskoeffizient

---

Der Korrelationskoeffizient mißt die **Korrelation** der Daten.

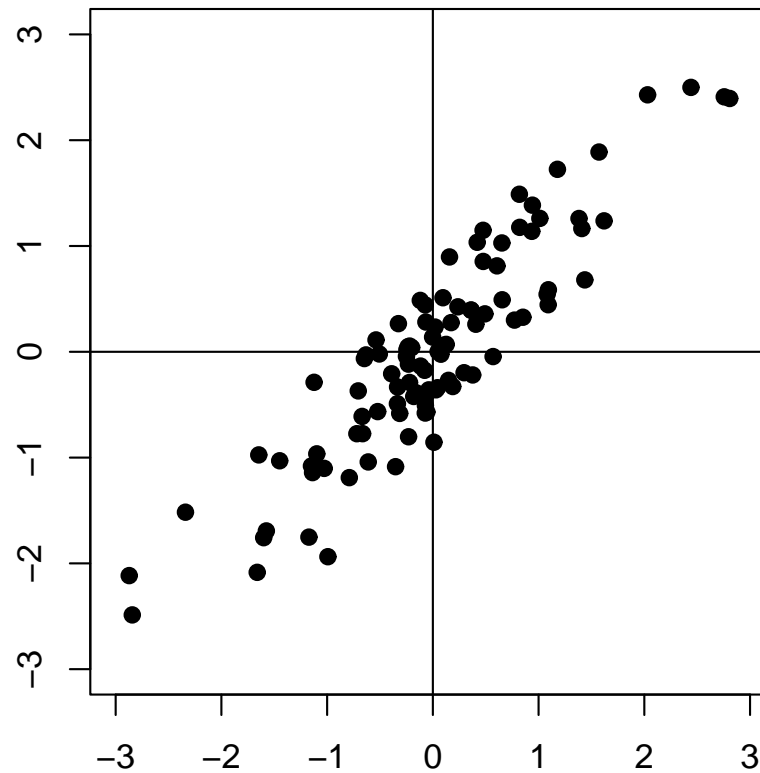
Die Korrelation gibt die Bindung der Punktwolke an eine steigende oder fallende **Hauptachse** an.

Die Korrelation gibt also das Ausmaß der **linearen Koppelung** an.

# Korrelationskoeffizient

---

Korrelation:  $r = 0.909$

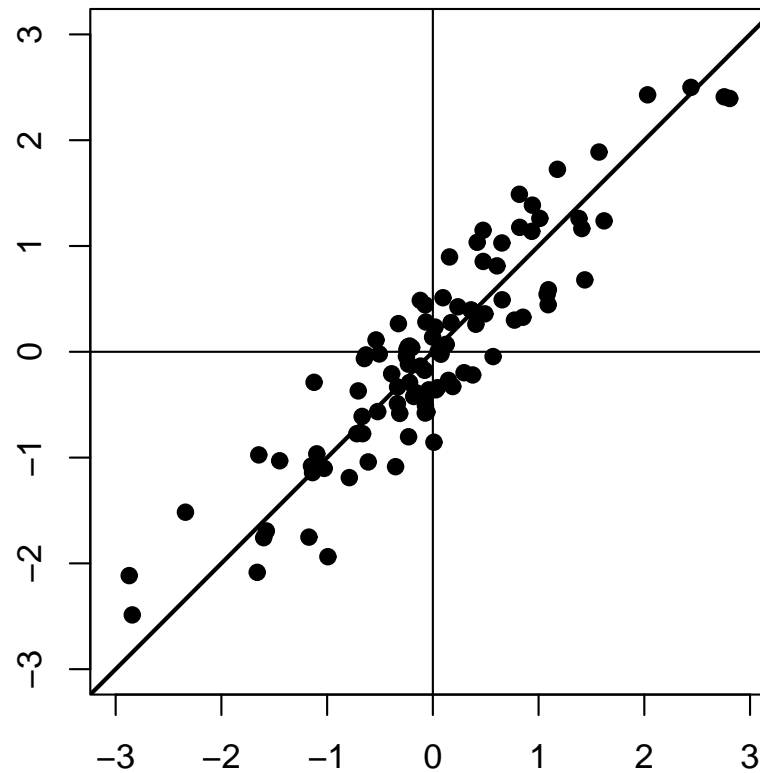




# Korrelationskoeffizient

---

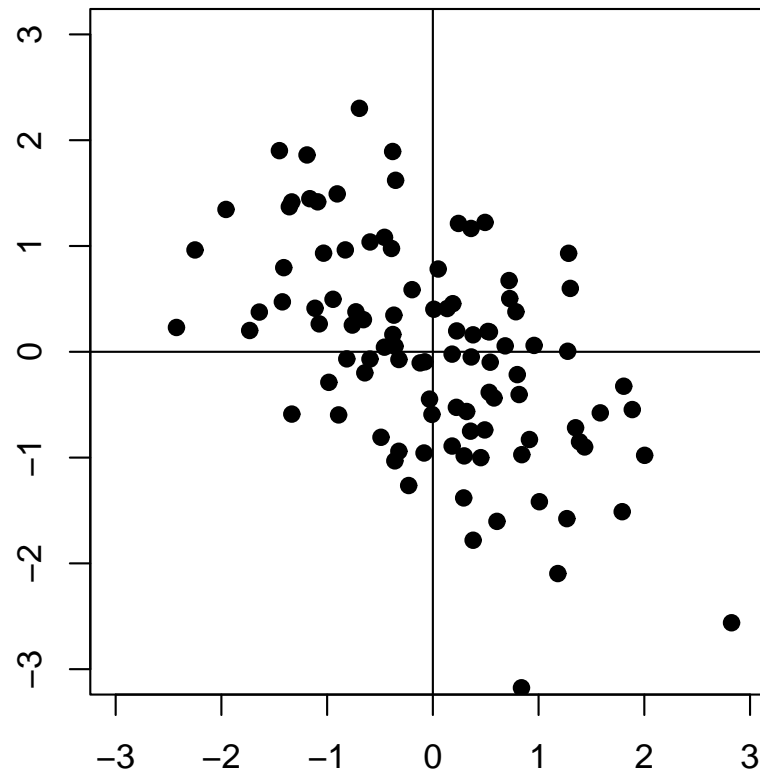
Korrelation:  $r = 0.909$



# Korrelationskoeffizient

---

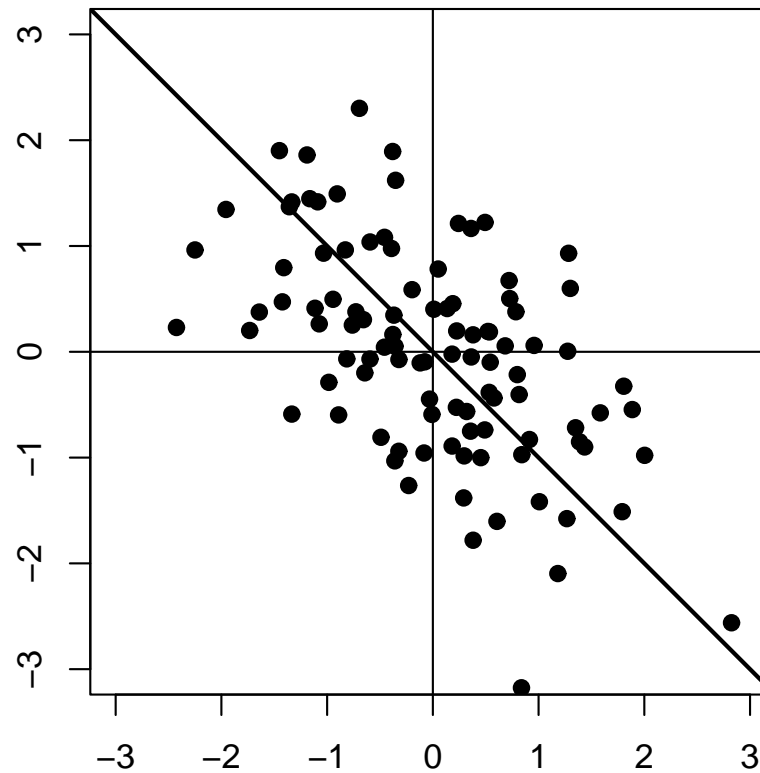
Korrelation:  $r = -0.5491$



# Korrelationskoeffizient

---

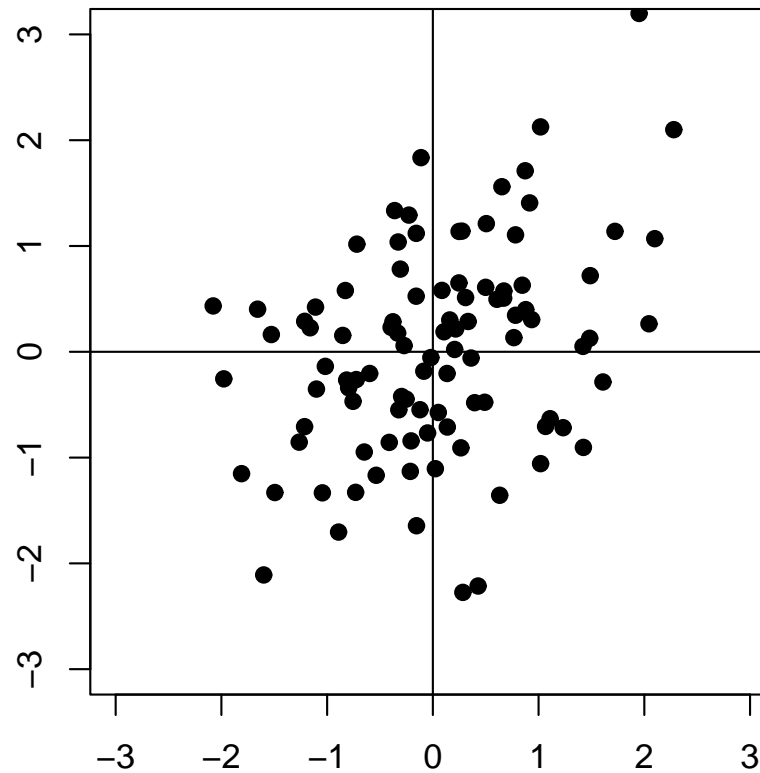
Korrelation:  $r = -0.5491$



# Korrelationskoeffizient

---

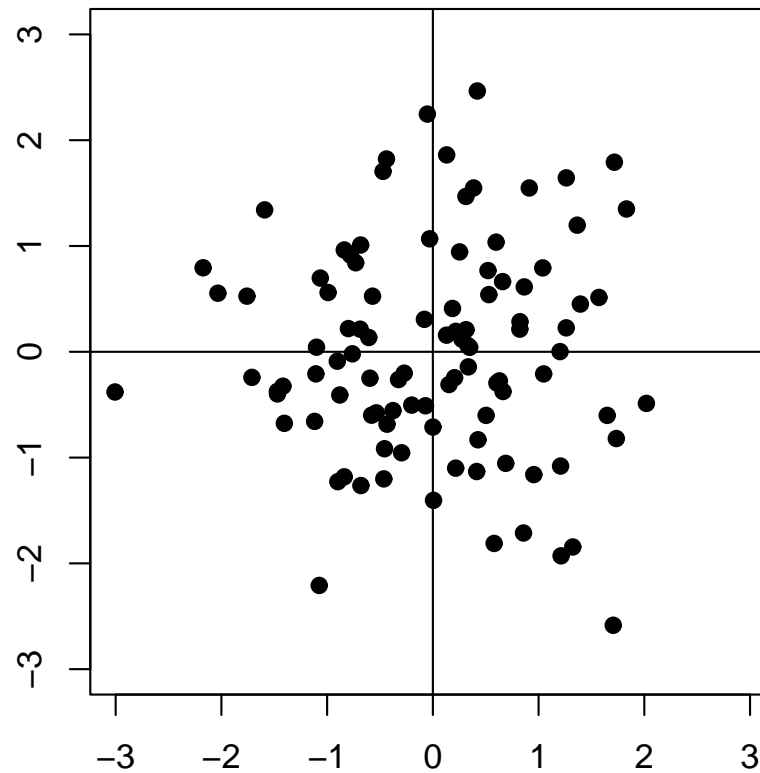
Korrelation:  $r = 0.3978$



# Korrelationskoeffizient

---

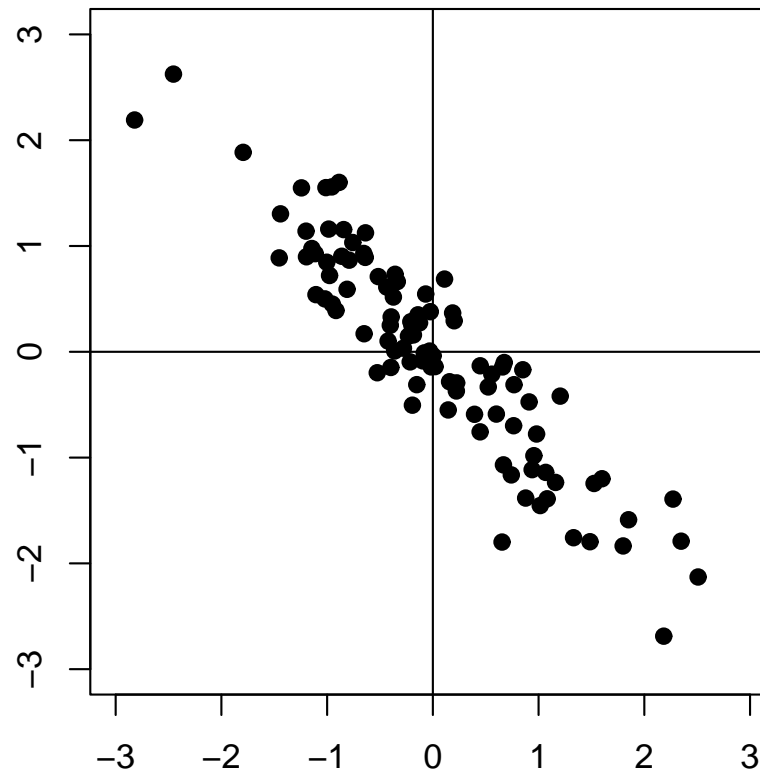
Korrelation:  $r = -0.0071$



# Korrelationskoeffizient

---

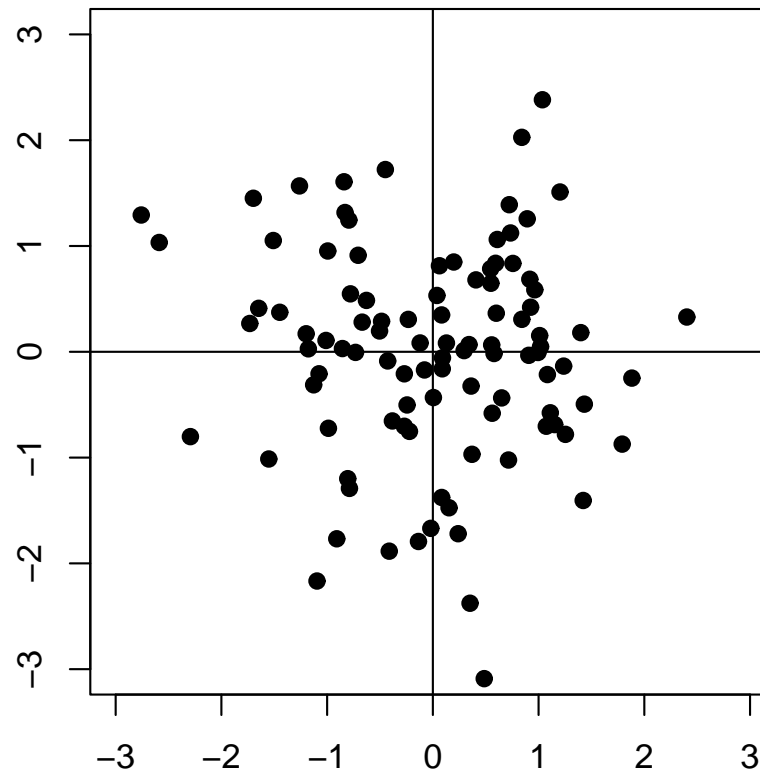
Korrelation:  $r = -0.9229$



# Korrelationskoeffizient

---

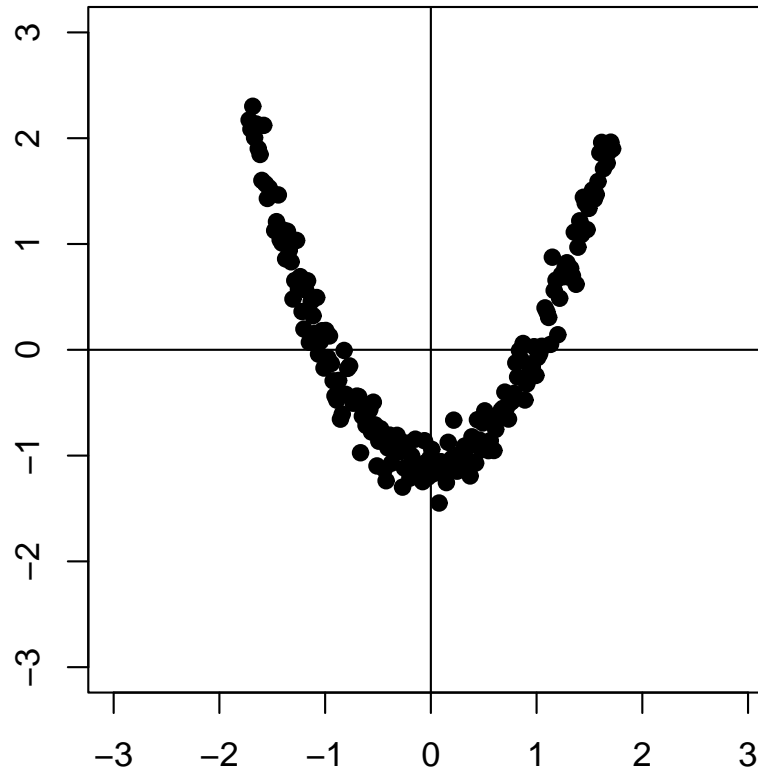
Korrelation:  $r = -0.0672$



# Korrelationskoeffizient

---

Korrelation:  $r = -0.0134$

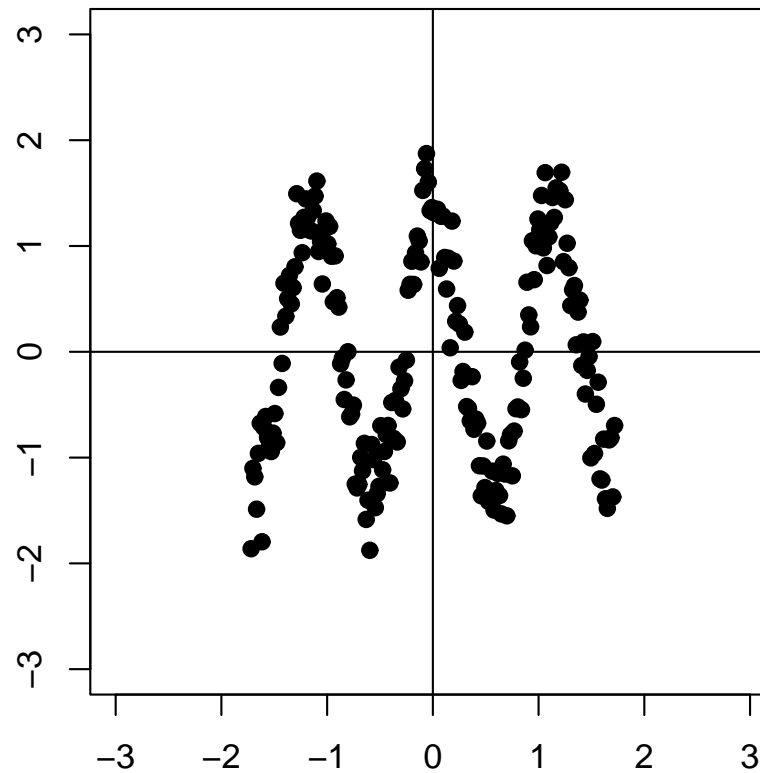




# Korrelationskoeffizient

---

Korrelation:  $r = 0.0086$



# Korrelationskoeffizient

---

Berechnung des Korrelationskoeffizienten  $r$  aus den Originaldaten

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

wobei

$$s_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

die **Kovarianz** der Daten bezeichnet.

# Korrelationskoeffizient

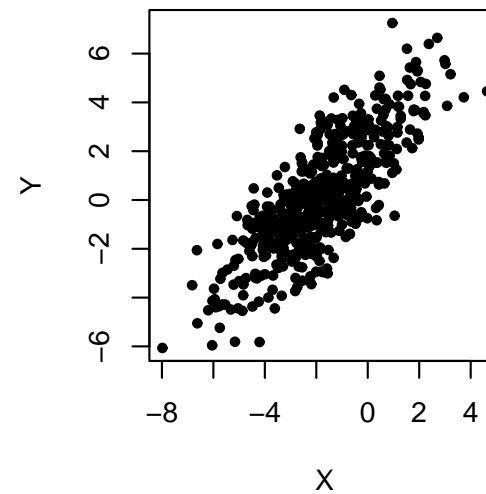
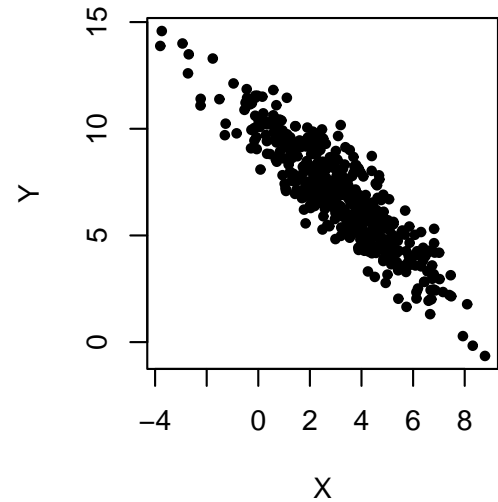
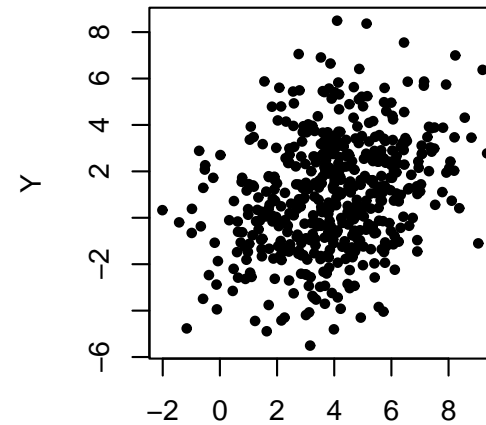
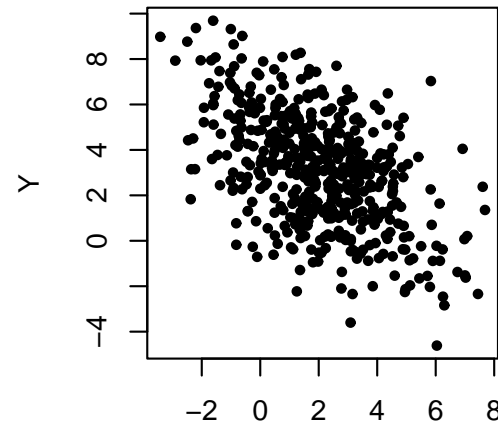
Zuordnen:

$$r = -0.9$$

$$r = -0.5$$

$$r = +0.3$$

$$r = +0.8$$



# Korrelationskoeffizient

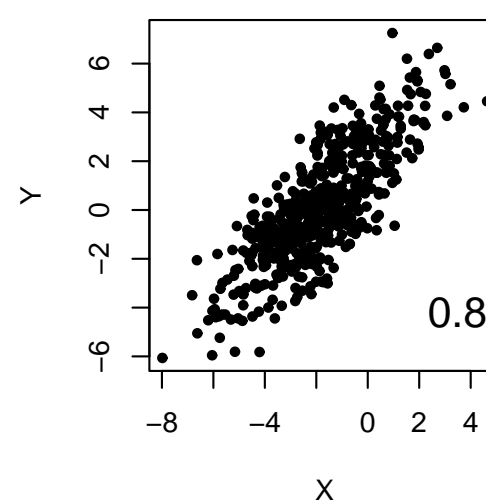
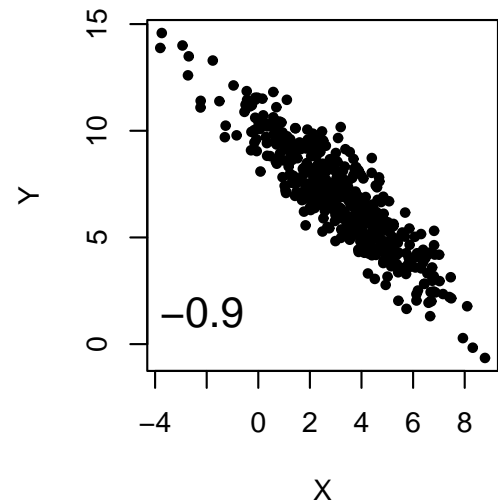
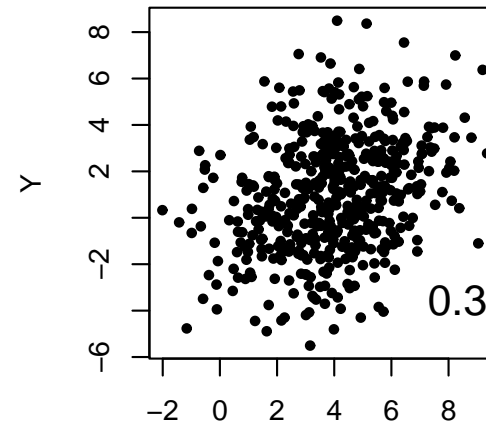
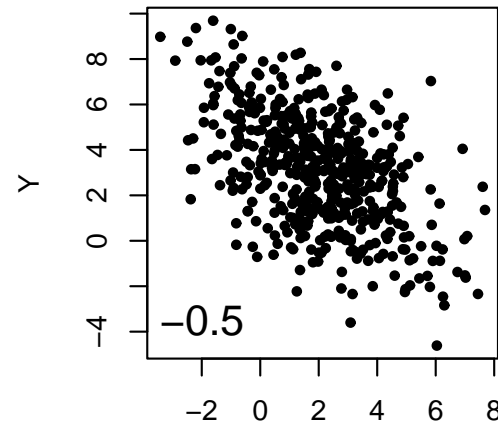
Zuordnen:

$$r = -0.9$$

$$r = -0.5$$

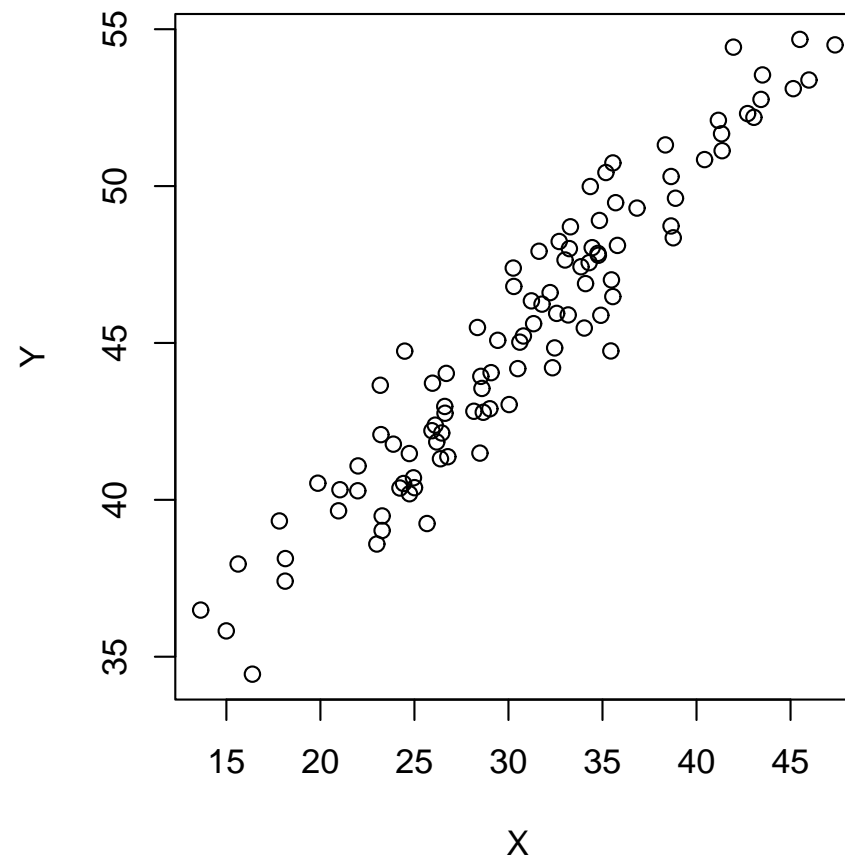
$$r = +0.3$$

$$r = +0.8$$



# Korrelationskoeffizient

---



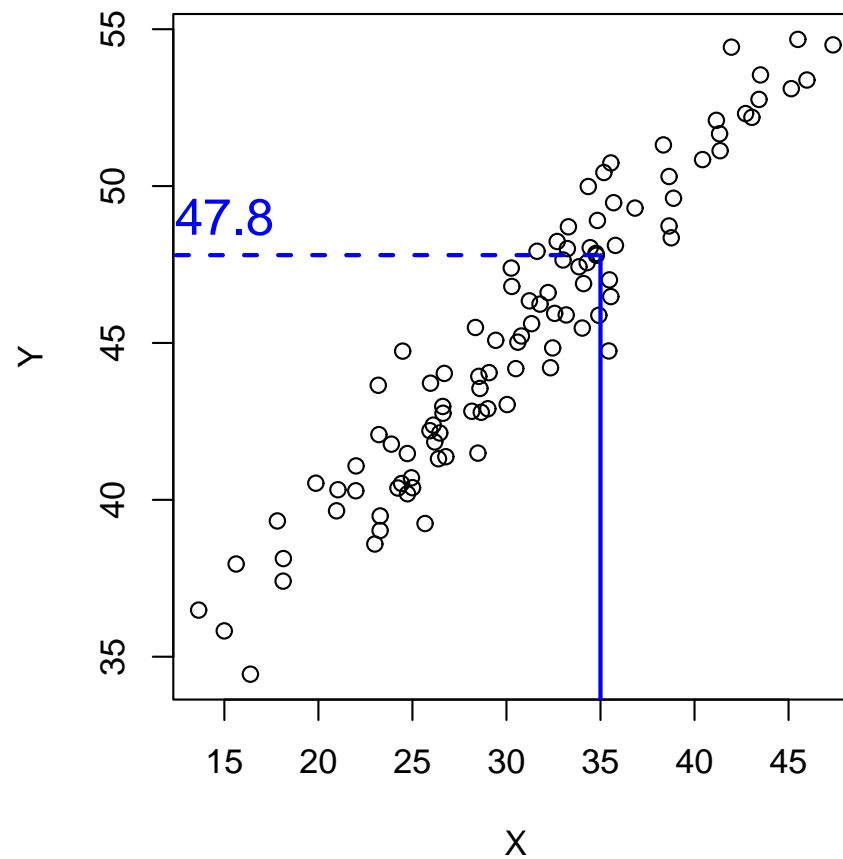
# Korrelationskoeffizient

---

- (a) Die Korrelation von  $X$  und  $Y$  ist positiv. **richtig**
- (b) Die Standardabweichung  $s_y$  ist größer als 15. **falsch**  $s_y = 4.64$
- (c) Der Absolutbetrag des Korrelationskoeffizienten ist höchstens 0.8. **falsch**  $r = 0.956$
- (d) Die Mittelwert  $\bar{x}$  ist kleiner als 47. **richtig**  $\bar{x} = 30.7$
- (e) Der Korrelationskoeffizient fällt um 3, wenn man von  $Y$  die Zahl 3 subtrahiert. **falsch**
- (f) Der Korrelationskoeffizient ändert das Vorzeichen, wenn man das Vorzeichen von  $X$  ändert. **richtig**
- (g) Für  $X = 35$  ist ungefähr  $Y = 47.8$  zu erwarten. **richtig**

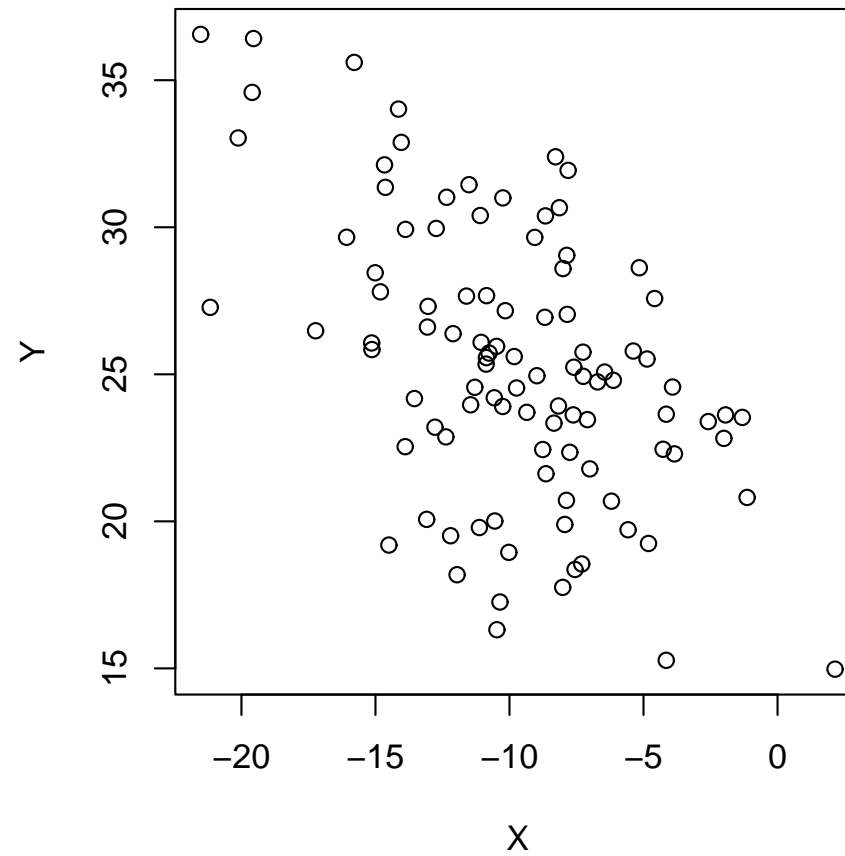
# Korrelationskoeffizient

---



# Korrelationskoeffizient

---





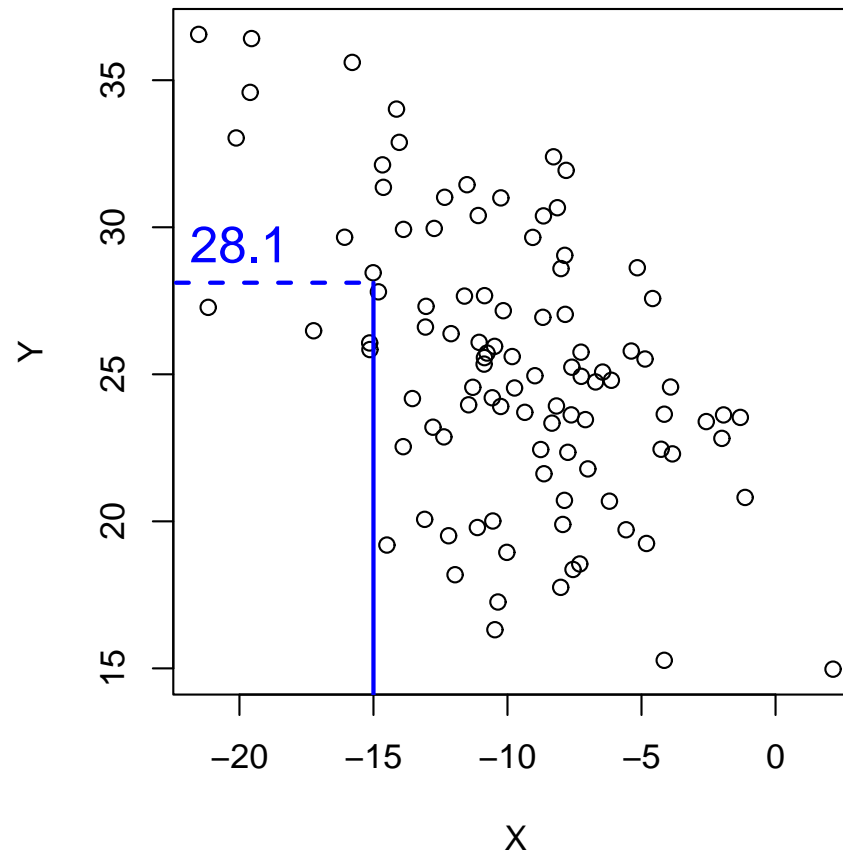
# Korrelationskoeffizient

---

- (a) Ist  $Y$  überdurchschnittlich groß, so ist tendenziell  $X$  unterdurchschnittlich groß. **richtig**
- (b) Das Mittel von  $X$  ist kleiner als 9.5. **richtig**  $\bar{x} = -9.851$
- (c) Der Absolutbetrag des Korrelationskoeffizienten ist höchstens 0.8. **richtig**  $r = -0.516$
- (d) Die Varianz von  $X$  ist kleiner als 50. **richtig**  $s_x^2 = 19.845$
- (e) Der Korrelationskoeffizient bleibt gleich, wenn man  $Y$  mit der Zahl 3 multipliziert. **richtig**
- (f) Der Korrelationskoeffizient ändert das Vorzeichen, wenn man das Vorzeichen von  $X$  und  $Y$  ändert. **falsch**
- (g) Für  $X = -15$  ist ungefähr  $Y = 17.3$  zu erwarten. **falsch**

# Korrelationskoeffizient

---



# Regression

---

**Bisher:** Betrachtung der Koppelung von  $X$  und  $Y$ , dabei  $X$  und  $Y$  gleichberechtigt.

**Jetzt:** Benutze  $X$  um  $Y$  vorherzusagen. Dabei ist:

- $X$  das **Prädiktormerkmal** oder die **unabhängige** oder **erklärende** Variable.
- $Y$  die **Responsevariable** oder die **abhängige** Variable.

# Regression

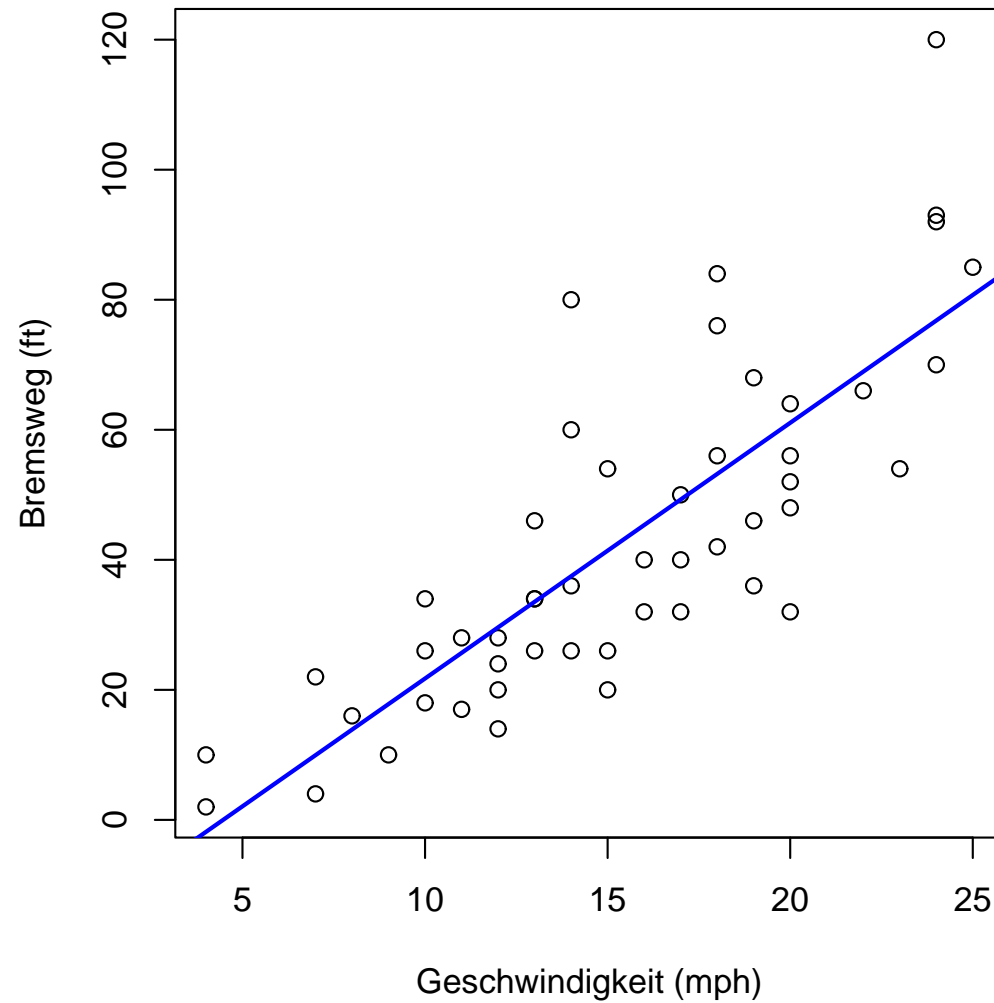
---

Typische Anwendung: **kausaler** Zusammenhang, d.h.  $X$  beeinflußt  $Y$ .

**Beispiel:** Geschwindigkeit und Bremsweg bei Autos.

# Regression

---

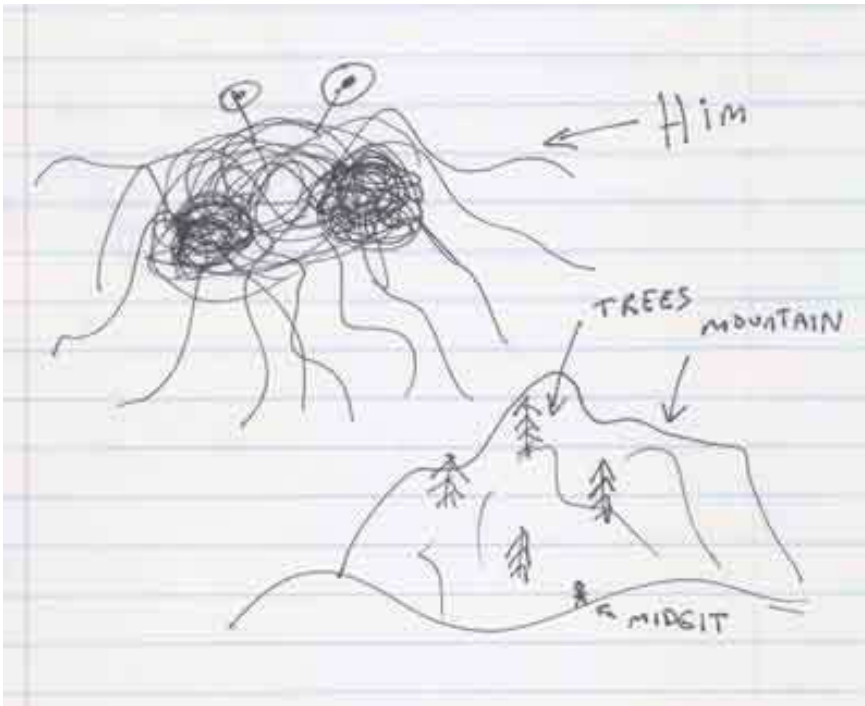


# Regression

---

**Aber:** Weder Korrelation noch Prognose/Regression können Kausalität nachweisen.

**Beispiel:** Globale Durchschnittstemperatur und Anzahl Piraten.

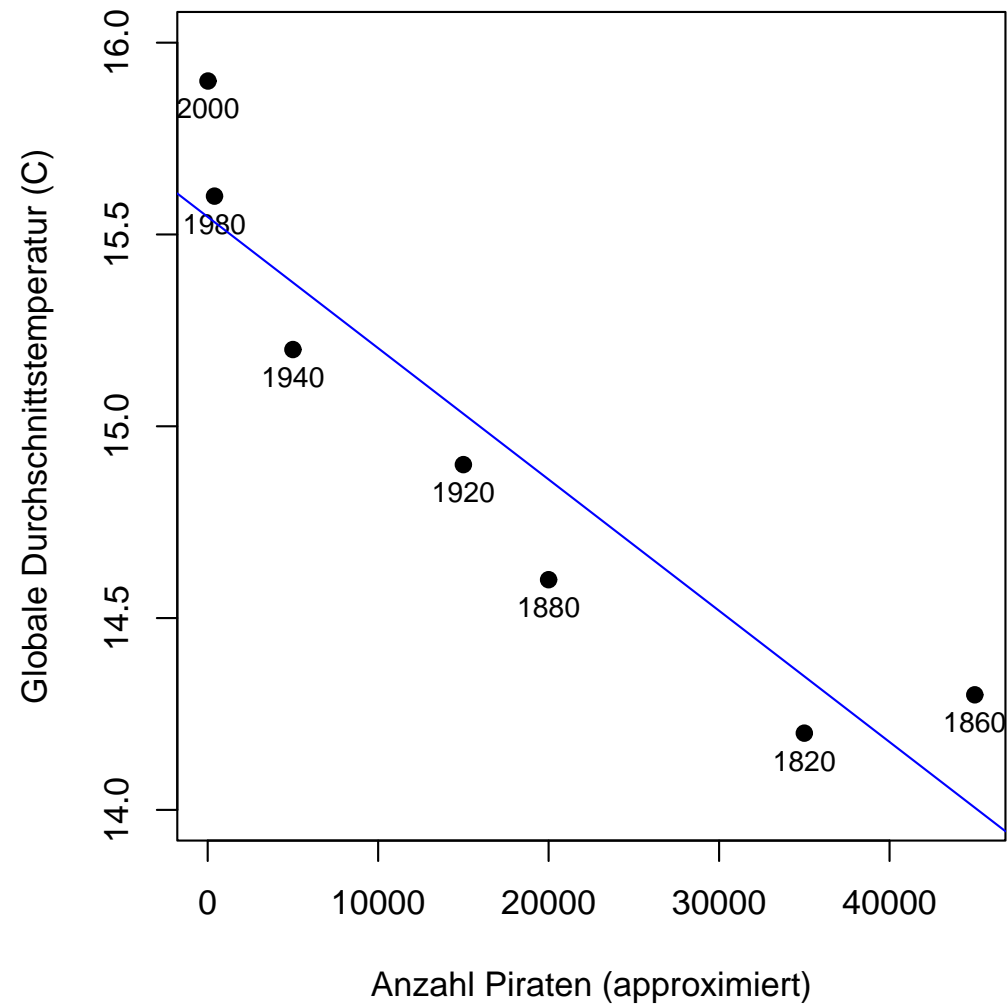


Quelle:

<http://www.venganza.org/>  
(*Church of the Flying Spaghetti Monster*)

# Regression

---



# Regression

---

**Gegeben:** Datenpaar  $(x_i, y_i)$

**Gesucht:** Prognosefunktion  $y = g(x)$ , die aus einem beobachteten Wert  $x$  eine möglichst gute Prognose  $y$  berechnet.

- $\hat{y}_i = g(x_i)$  heißt **Schätzwert**,
- $y_i - \hat{y}_i$  heißt **Prognosefehler**.

Was heißt *möglichst gut*? Minimiere die Quadratsumme der Prognosefehler

$$SS_R = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$



# Kleinste Quadrate

---

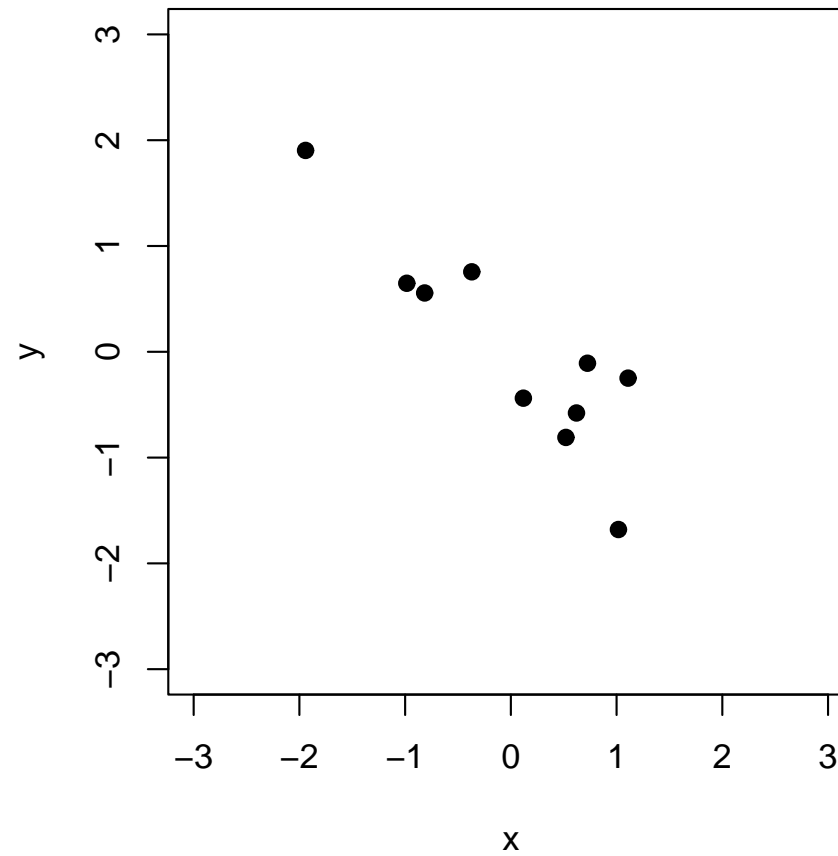
Wähle für  $g(x) = \hat{a} + \hat{b}x$  eine lineare Funktion. Dann wird  $SS_R$  minimal, wenn

$$\begin{aligned}\hat{b} &= r \frac{s_y}{s_x} = \frac{s_{xy}}{s_x^2} \\ \hat{a} &= \bar{y} - \hat{b}\bar{x}\end{aligned}$$

Dies ist das **Prinzip der kleinsten Quadrate** (LSQ, KQ oder OLS).  $g(x) = \hat{a} + \hat{b}x$  heißt empirische Regressionsgerade und  $\hat{b}$  empirischer Regressionskoeffizient.

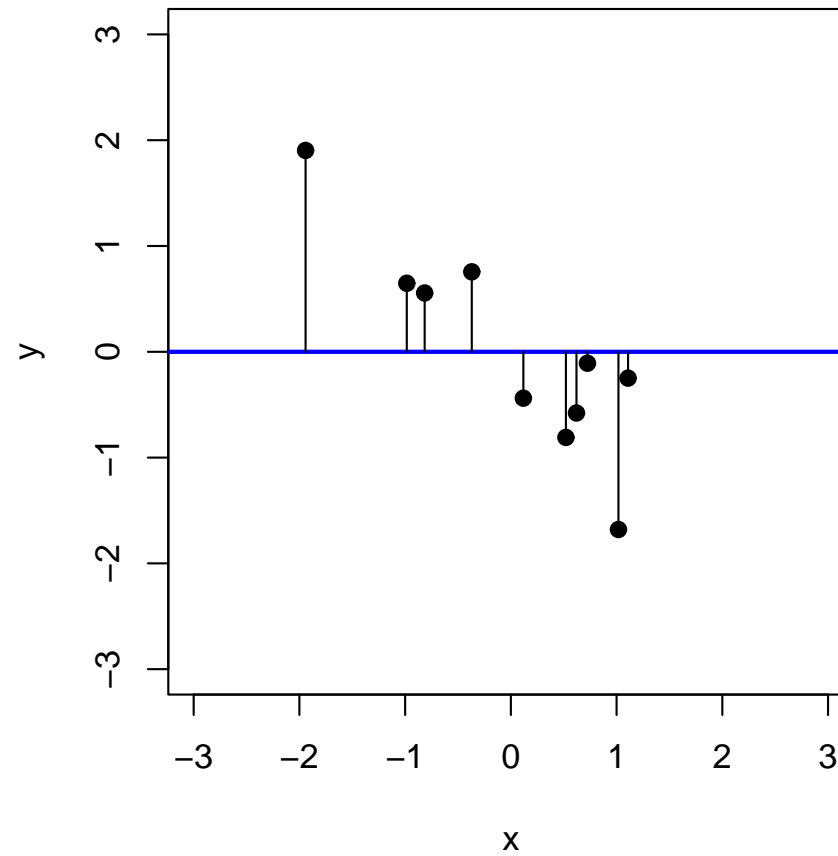
# Kleinste Quadrate

---



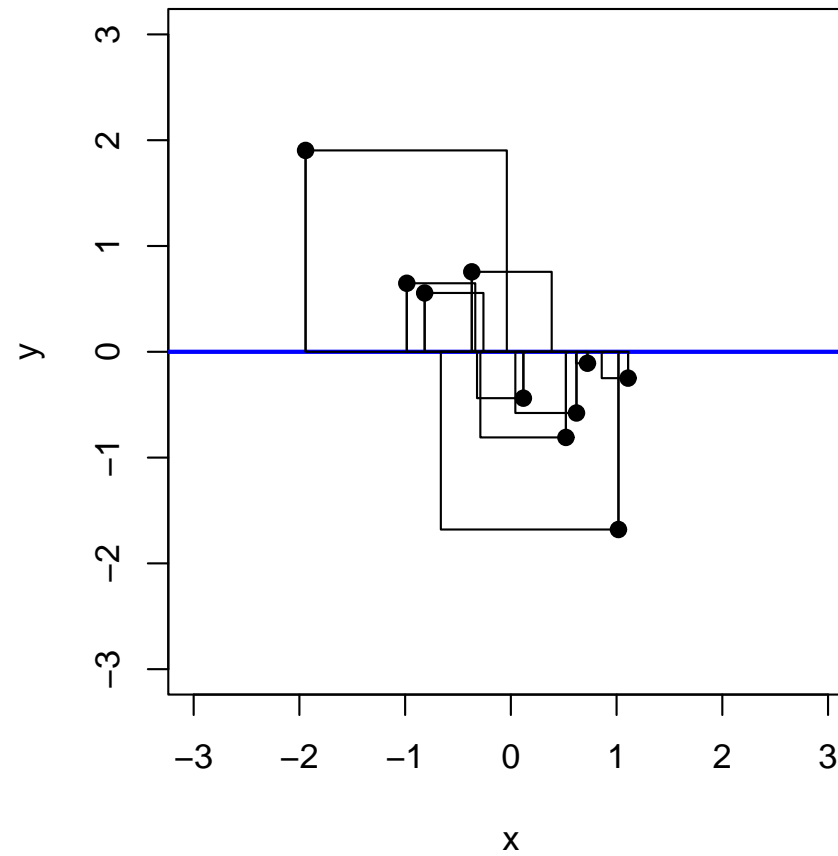
# Kleinste Quadrate

---



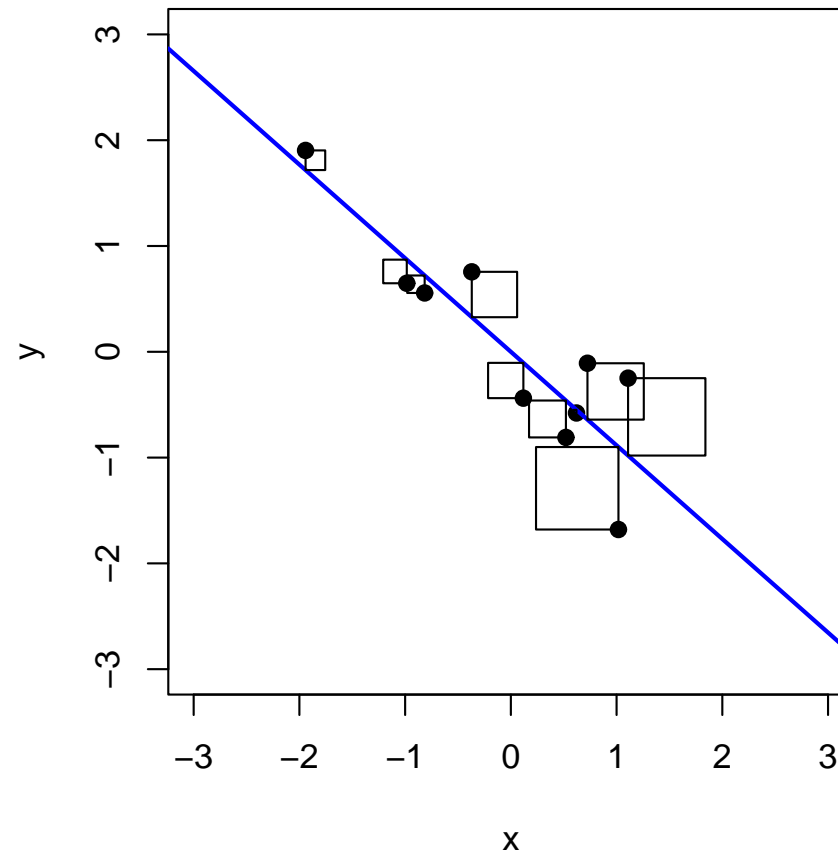
# Kleinste Quadrate

---



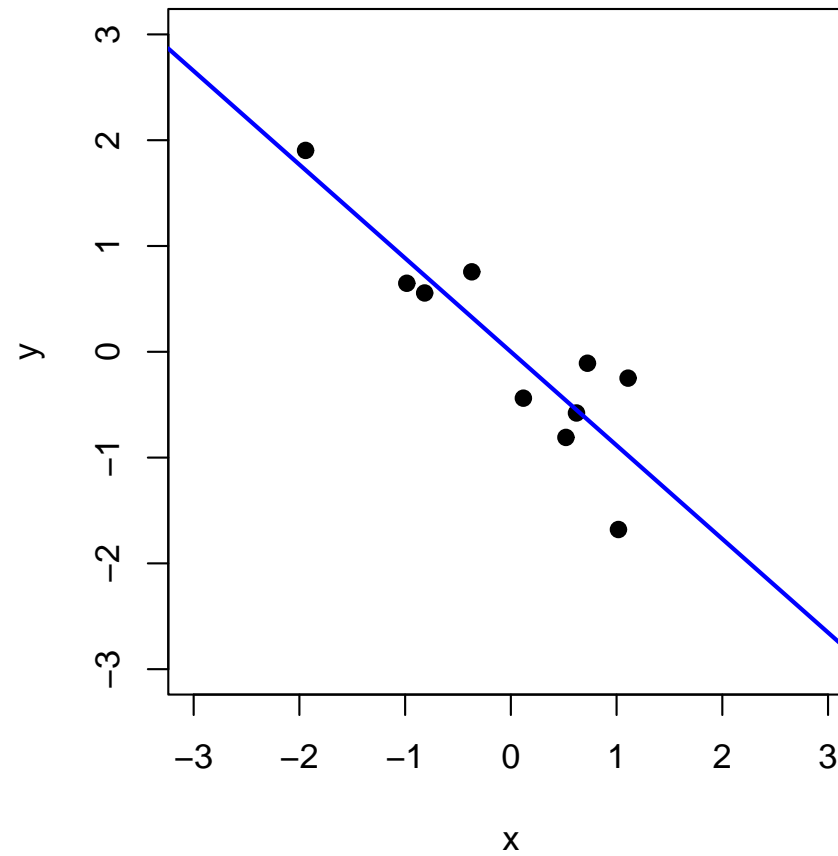
# Kleinste Quadrate

---



# Kleinste Quadrate

---



# Regression

---

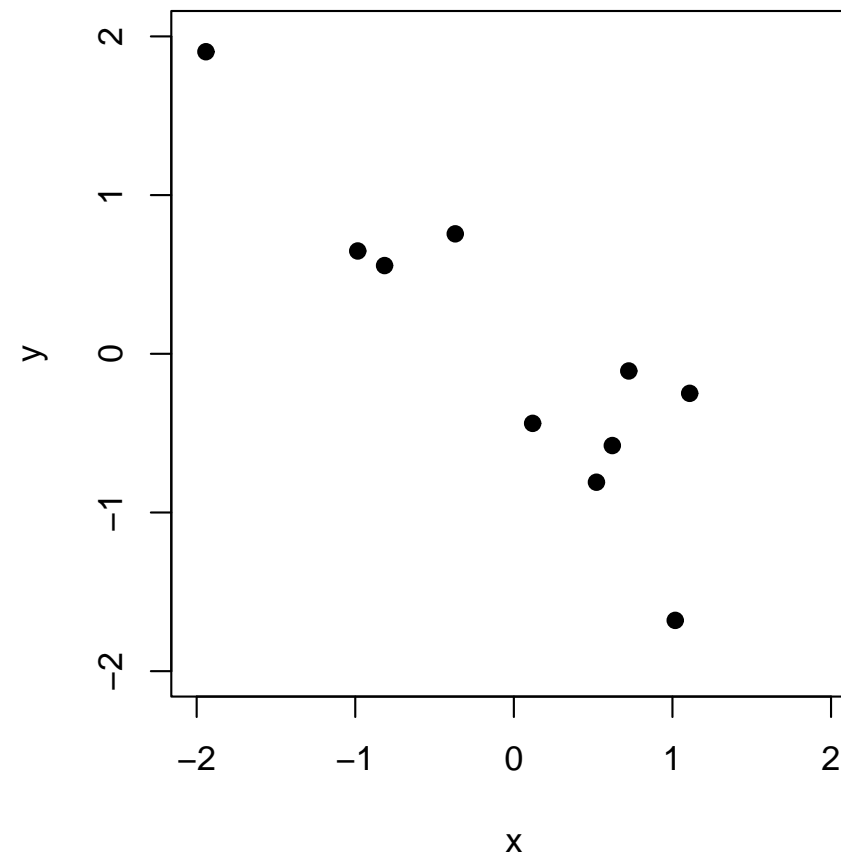
Bei standardisierten Daten gilt  $\bar{x} = \bar{y} = 0$  und  $s_x = s_y = 1$ , deshalb:

- $\hat{b} = r$  und  $\hat{a} = 0$ .
- Da  $|r| \leq 1$  ist die Regressionsgerade flacher als die Hauptachse.

→ Regressionsphänomen

# Regression

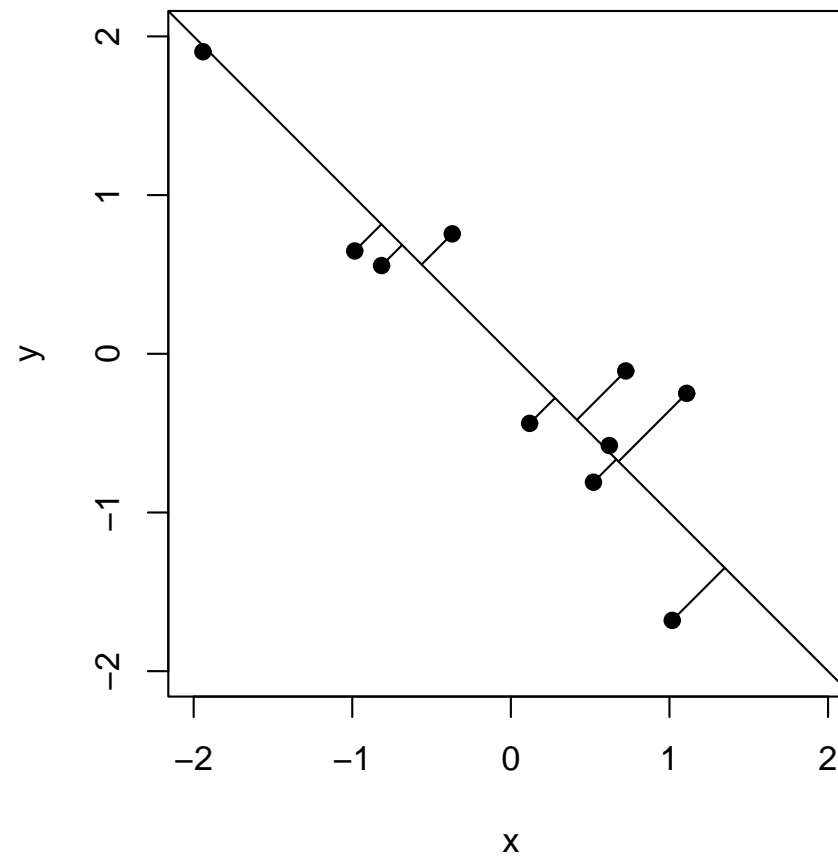
---





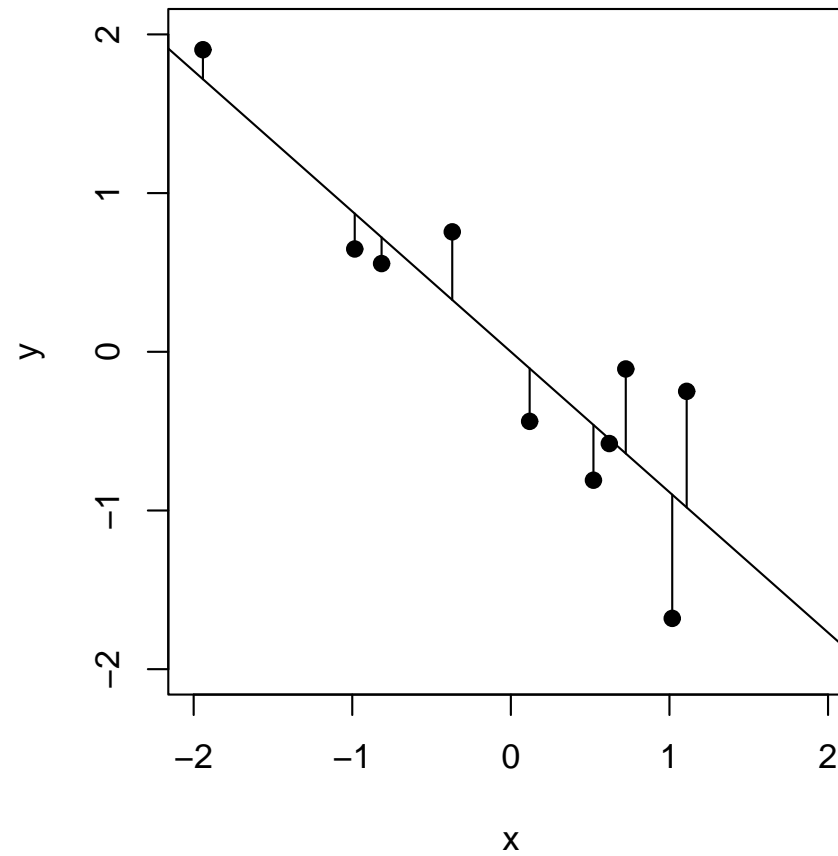
# Regression

---



# Regression

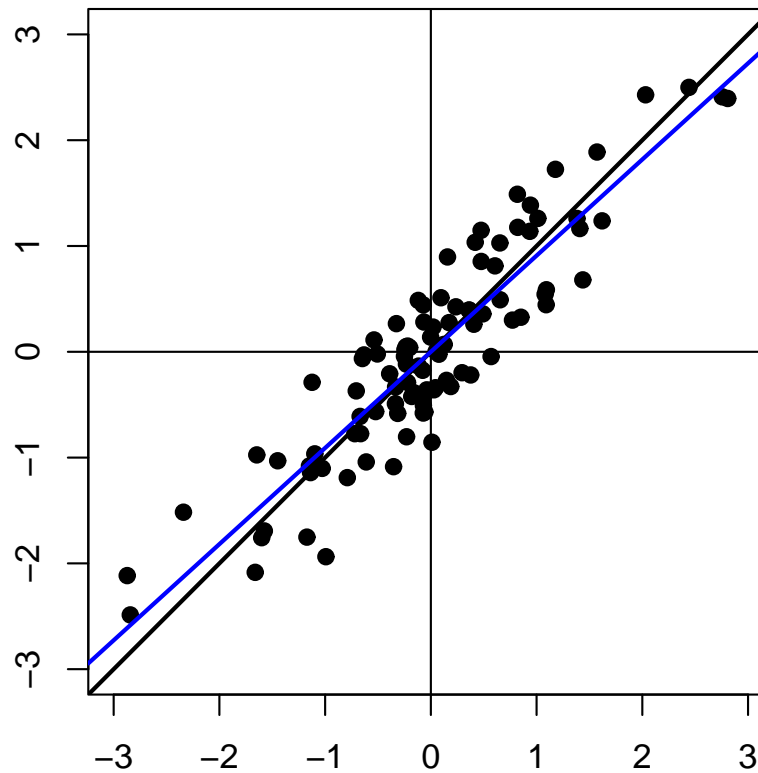
---



# Regression

---

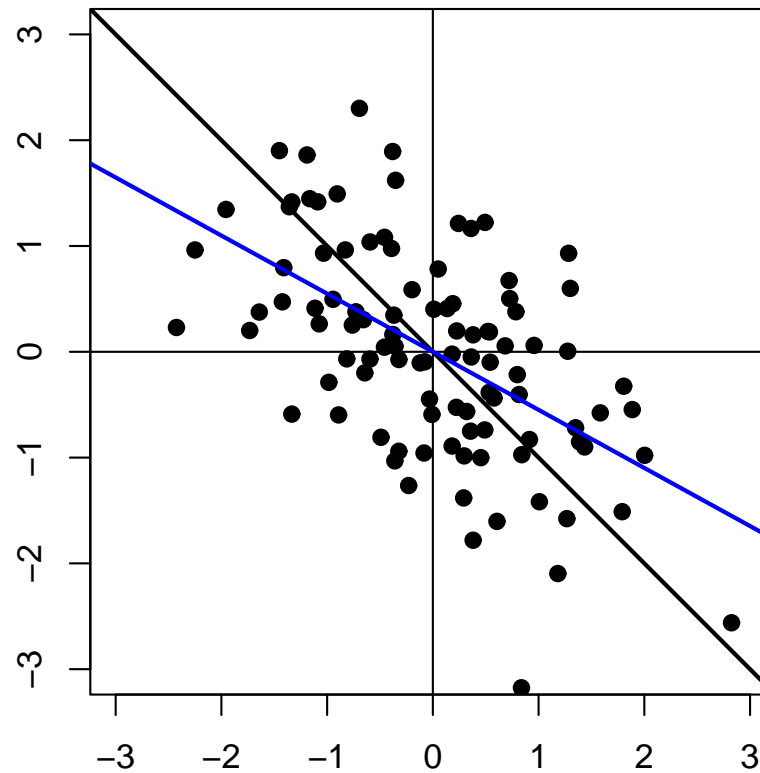
Korrelation:  $r = 0.909$



# Regression

---

Korrelation:  $r = -0.5491$



# Regression

---

## Beispiel: Aufgabensammlung

**189.** Für 20 Betriebe wurde die Anzahl  $X$  der Mitarbeiter und die Höhe der Aufwendungen für Fortbildungskosten  $Y$  (in GE) ermittelt.

	Merkmal $X$	Merkmal $Y$
Mittelwert	150	2000
Varianz	625	250000

Die Korrelation zwischen  $X$  und  $Y$  beträgt 0.65. Wie lautet die Gleichung der Regressionsgeraden?

# Regression

---

## Beispiel: Aufgabensammlung

**186.** Der Zusammenhang zwischen  $X$  = “jährliche Fahrleistung” (in 1000 km) und  $Y$  = “jährliche Schadenssumme” (in GE) soll mit einer Regressionsgeraden beschrieben werden.

	Merkmal $X$	Merkmal $Y$
Mittelwert	72.3	12.5
Varianz	236.5	6.4

Die Korrelation zwischen  $X$  und  $Y$  beträgt 0.8.

Welchen Schätzwert für die Schadenssumme erhält man für ein Fahrzeug, dessen Fahrleistung 90000 km beträgt?

# Varianzanalyse

---

Die Streuung der Werte  $y_i$  hat im Regressionsmodell unterschiedliche Ursachen:

- systematische Unterschiede durch unterschiedliche Prädiktoren  $x_i$ ,
- zufällige Streuung von Versuch zu Versuch.

$$\text{Erklärbare Streuung} \quad SS^* = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = r^2 n s_y^2$$

$$\text{Reststreuung} \quad SS_R = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = (1 - r^2) n s_y^2$$

$$\text{Gesamtstreuung} \quad SS_T = \sum_{i=1}^n (y_i - \bar{y})^2 = n s_y^2$$

Es gilt:

$$SS_T = SS^* + SS_R$$

# Varianzanalyse

---

Die Qualität der Regressionsgeraden kann wieder durch das **Bestimmtheitsmaß** angegeben werden:

$$\frac{SS^*}{SS_T} = r^2 = r_{xy}^2.$$

Es gibt also an, wie groß der Anteil an der Gesamtstreuung ist, der durch die Regressionsgerade erklärt werden kann.



# Regressionsmodelle

---

Ein **Regressionsmodell** ist ein Modell, in dem ein Prädiktor nur den Erwartungswert der Responsevariablen beeinflusst:

$$\begin{aligned}E_x(Y) &= f(x) \\ Y &= f(x) + U\end{aligned}$$

wobei die zufällige Störung den Erwartungswert  $E_x(U) = 0$  hat.  
 $f(x)$  heißt Regressionsfunktion.

# Regressionsmodelle

---

Im **linearen Regressionsmodell** ist die Regressionsfunktion  $f(x) = a + bx$  eine **Regressionsgerade**.

$$E_x(Y) = a + bx$$

$$Y = a + bx + U$$

Die statistische Aufgabe ist das Schätzen der **Parameter**  $a$  und  $b$  aus Daten. Die beste Schätzung ist gegeben durch die empirische Regressionsgerade.

# Signifikanzproblem

---

**Frage:** Hat  $X$  überhaupt einen Einfluß auf  $Y$ ?

Anders formuliert:

- Ist  $b \neq 0$ ?
- Ist  $\hat{b}$  **signifikant** von  $b = 0$  verschieden?

**Antwort:** Teste

Nullhypothese:  $b = 0$

Alternative:  $b \neq 0$

# ANOVA-Tabelle

---

Die Hypothese wird verworfen, falls die  $F$ -Größe aus

	$SS$	$df$	$MSS$
$*$	$r^2 n s_y^2$	1	$r^2 n s_y^2$
$R$	$(1 - r^2) n s_y^2$	$n - 2$	$\frac{(1 - r^2) n s_y^2}{n - 2}$
	$n s_y^2$	$n - 1$	

$$F = \frac{MSS^*}{MSS_R} = (n - 2) \frac{r^2}{1 - r^2}$$

den Wert 4 überschreitet (Faustregel).

# Regression

---

## Beispiel: Aufgabensammlung

**188.** Für 20 Betriebe wurde die Anzahl  $X$  der Mitarbeiter und die Höhe der Aufwendungen für Fortbildungskosten  $Y$  (in GE) ermittelt.

	Merkmal $X$	Merkmal $Y$
Mittelwert	150	2000
Varianz	625	250000

Die Korrelation zwischen  $X$  und  $Y$  beträgt 0.65. Die Hypothese, “ $X$  und  $Y$  sind nicht gekoppelt” soll getestet werden. Wie lautet die Testgröße?

# Regression

---

## Beispiel: Aufgabensammlung

**187.** Umfrage unter 13 internationalen Airlines über Anzahl der Passagiermeilen  $X$  (in Mrd. Meilen) und die Gewinne  $Y$  (in Mrd. US Dollar).

	Merkmal $X$	Merkmal $Y$
Mittelwert	49.4	0.475
Varianz	670	0.15

Die Korrelation zwischen  $X$  und  $Y$  beträgt 0.48. Die Hypothese, “ $X$  und  $Y$  sind nicht gekoppelt” soll getestet werden. Wie lautet die Testgröße?

## Verständnisfragen

---

Unter der Regressionsgeraden von  $Y$  nach  $X$  einer zweidimensionalen  $(x, y)$ -Punktwolke versteht man jene Gerade durch den Mittelpunkt, die die Quadratsumme der:

- (a) senkrechten Abstände (parallel zur  $y$ -Achse) zu den Daten minimiert. **richtig**
- (b) waagrechten Abstände (parallel zur  $x$ -Achse) zu den Daten minimiert. **falsch**
- (c) orthogonalen (dh. rechtwinkelligen) Abstände zu den Daten minimiert. **falsch**
- (d) maximalen Abstände zu den Daten minimiert. **falsch**
- (e) die signifikanten Abstände zu den Daten minimiert. **falsch**

## Verständnisfragen

---

Was ist die Streuungszerlegung bei einfacher Regression?

- (a) Die Aufteilung der Streuung der Responsevariablen um ihren Mittelwert in zwei Teile. **richtig**
- (b) Die Aufteilung der Streuung der Responsevariablen um einen Vergleichswert in zwei Teile. **falsch**
- (c) Die Aufteilung der Gesamtstreuung in einen Teil, der durch den Mittelwert verursacht wird, und einen Teil der durch die Erklärungsvariable verursacht wird. **falsch**
- (d) Die Aufteilung der Gesamtstreuung in die Streuungen um die beiden Mittelwerte der einzelnen Stichproben. **falsch**
- (e) Die Aufteilung der Reststreuung in die Varianzen der einzelnen Stichproben. **falsch**



# Verständnisfragen

---

Der Korrelationskoeffizient

- (a) ändert das Vorzeichen, wenn man das Vorzeichen von  $X$  ändert.  
richtig
- (b) ändert das Vorzeichen, wenn man das Vorzeichen von  $Y$  ändert.  
richtig
- (c) verdoppelt sich, wenn man  $X$  mit 2 multipliziert. falsch
- (d) fällt um 3, wenn man von  $Y$  die Zahl 3 subtrahiert. falsch
- (e) ändert das Vorzeichen, wenn man das Vorzeichen von  $X$  und  $Y$  ändert. falsch

## Verständnisfragen

---

Die Steigung der Regressionsgeraden von  $Y$  nach  $X$  (also der Rohdaten)

- (a) halbiert sich, wenn man  $Y$  durch 2 dividiert. richtig
- (b) halbiert sich, wenn man  $X$  mit 2 multipliziert. richtig
- (c) ändert das Vorzeichen, wenn man das Vorzeichen von  $X$  ändert. richtig
- (d) verdoppelt sich, wenn man  $X$  mit 2 multipliziert. falsch
- (e) wächst um 3, wenn man zu  $X$  die Zahl 3 addiert. falsch

## Verständnisfragen

---

Die Steigung der Regressionsgeraden im standardisierten Streudiagramm (also von  $Z_y$  nach  $Z_x$ )

- (a) ändert das Vorzeichen, wenn man das Vorzeichen von  $X$  ändert. **richtig**
- (b) halbiert sich, wenn man  $Y$  durch 2 dividiert. **falsch**
- (c) verdoppelt sich, wenn man  $X$  mit 2 multipliziert. **falsch**
- (d) halbiert sich, wenn man  $X$  mit 2 multipliziert. **falsch**
- (e) wächst um 3, wenn man zu  $X$  die Zahl 3 addiert. **falsch**

## Verständnisfragen

---

Der Korrelationskoeffizient einer bivariaten Datenliste  $(X, Y)$  beträgt  $-0.8$ . Die Varianz von  $X$  ist so groß wie die Varianz von  $Y$ .

- (a) Unterdurchschnittliche Werte von  $X$  treten häufig gemeinsam mit unterdurchschnittlichen Werten von  $Y$  auf. falsch
- (b) Es besteht eine undeutliche negative Kopplung, weil der Korrelationskoeffizient negativ ist, aber nahe bei 0 liegt. falsch
- (c) Die Regressionsgerade von  $Y$  nach  $X$  fällt. richtig
- (d) 64 Prozent der Varianz von  $Y$  werden durch die Varianz von  $X$  erklärt. richtig
- (e) Die  $X$ -Daten streuen um 30 Prozent stärker als die  $Y$ -Daten. falsch

## Verständnisfragen

---

Der Korrelationskoeffizient einer bivariaten Datenliste  $(X, Y)$  beträgt 0. Die Varianz von  $X$  ist kleiner als die Varianz von  $Y$ .

- (a) Unterdurchschnittliche Werte von  $X$  treten häufig gemeinsam mit unterdurchschnittlichen Werten von  $Y$  auf. falsch
- (b) Überdurchschnittliche Werte von  $X$  treten etwa gleich häufig gemeinsam mit überdurchschnittlichen und mit unterdurchschnittlichen Werten von  $Y$  auf. richtig
- (c) Es besteht eine deutliche monotone Kopplung, weil der Betrag des Korrelationskoeffizienten nahe bei 0 liegt. falsch
- (d) Es besteht keine Kopplung, weil der Korrelationskoeffizient kleiner als 1 ist. falsch
- (e) Die Regressionsgerade von  $Y$  nach  $X$  steigt. falsch

## Verständnisfragen

---

$r_{xy} = 0.1$ . Die Varianz von  $X$  ist größer als die Varianz von  $Y$ .

- (a) Die Standardscores der  $X$ -Daten sind durchschnittlich um 20 Prozent größer als die Standardscores der  $Y$ -Daten. falsch
- (b) Die Standardscores der  $X$ -Daten streuen um 20 Prozent stärker als die Standardscores der  $Y$ -Daten. falsch
- (c) Die  $X$ -Daten sind durchschnittlich um 40 Prozent kleiner als die  $Y$ -Daten. falsch
- (d) Überdurchschnittliche Werte von  $X$  treten häufig gemeinsam mit überdurchschnittlichen Werten von  $Y$  auf. falsch
- (e) Überdurchschnittliche Werte von  $X$  treten etwa gleich häufig gemeinsam mit überdurchschnittlichen und mit unterdurchschnittlichen Werten von  $Y$  auf. richtig

# Zusammenfassung Kapitel 7

---

- Streudiagramm von bivariaten Datenlisten
- Standardisiertes Streudiagramm
- Koppelung im Streudiagramm
- Korrelation und Korrelationskoeffizient
- Prognoseproblem
- Regressionsgerade
- Varianzanalyse

# Das Zweistichprobenproblem

## Kapitel 8



# Überblick

---

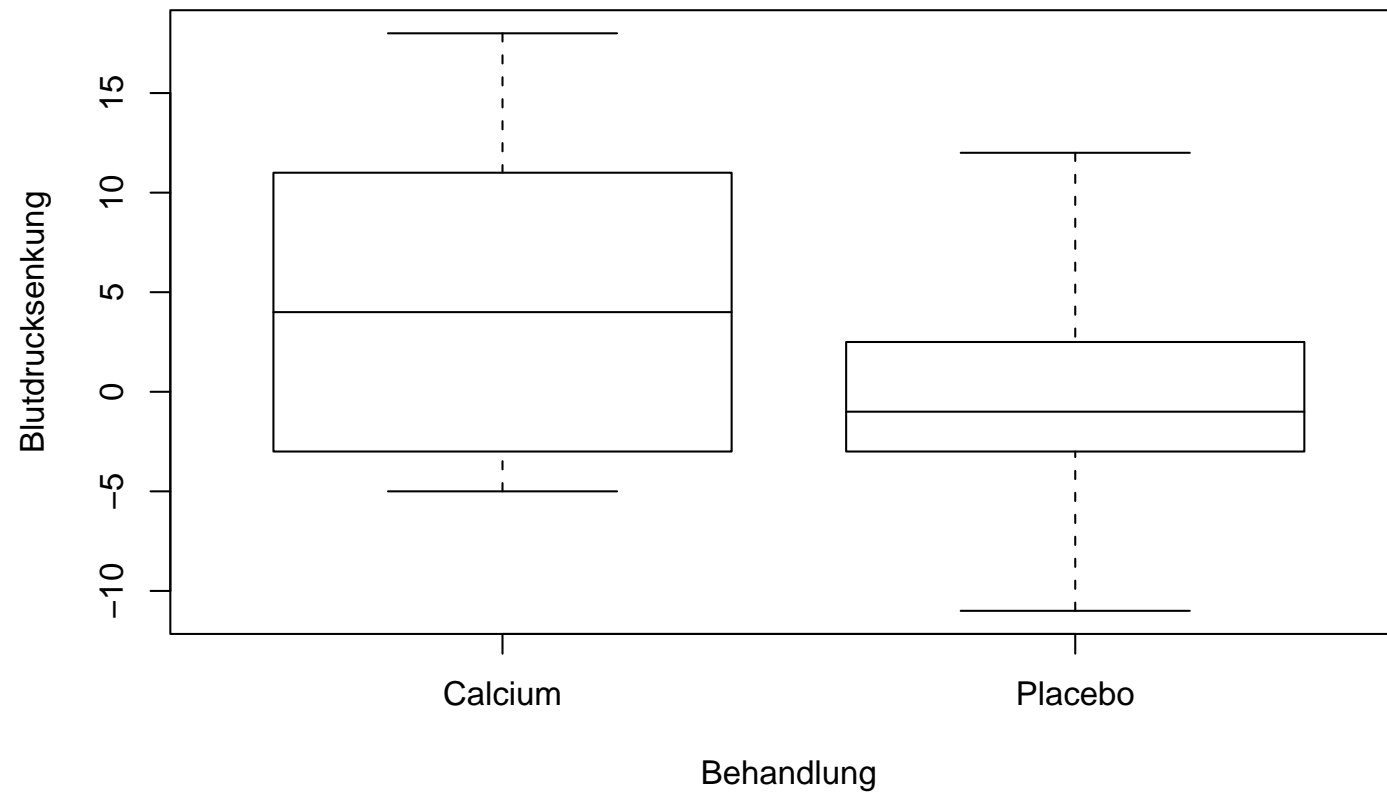
**Problem:** Eine quantitative Variable wird in zwei unabhängigen Stichproben gemessen. (Formal: Eine quantitative und eine qualitative Variable wird erhoben, wobei die qualitative Variable die Stichprobe in zwei Teile unterteilt.)

**Frage:** Unterscheidet sich der Mittelwert in den beiden Stichproben?

**Beispiel:** Blutdrucksenkung durch ein zwei Behandlungen, ein aktives Medikament (Calcium) und ein Placebo.

# Überblick

---



# Versuchsplan

---

Betrachte zwei unabhängige Zufallsvariablen  $X$  und  $Y$ . Diese haben die Erwartungswerte  $E(X) = \mu_1$  und  $E(Y) = \mu_2$  und die Varianzen  $V(X) = \sigma_1^2$  und  $V(Y) = \sigma_2^2$ .

Ziehe eine Stichprobe vom Umfang  $n_1$  aus  $X$  mit Realisationen  $x_1, \dots, x_{n_1}$ . Diese haben den Mittelwert  $\bar{x}$  und die Varianz  $s_x^2$ .

Ziehe eine Stichprobe vom Umfang  $n_2$  aus  $Y$  mit Realisationen  $y_1, \dots, y_{n_2}$ . Diese haben den Mittelwert  $\bar{y}$  und die Varianz  $s_y^2$ .

# Vergleich der Erwartungswerte

---

**Frage:** Stimmen die Erwartungswerte  $\mu_1$  und  $\mu_2$  überein?

**Antwort:** Teste

Nullhypothese:  $\mu_1 = \mu_2$

Alternative:  $\mu_1 \neq \mu_2$

auf Basis empirischer Daten.

# Vergleich der Erwartungswerte

---

**Intuitiv:** Betrachte die Differenz der Mittelwerte  $\bar{x} - \bar{y}$ . Es gilt:

$$E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2$$

$$V(\bar{X} - \bar{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

$$SD = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$\widehat{SD} = \sqrt{\frac{s_{x,n-1}^2}{n_1} + \frac{s_{y,n-1}^2}{n_2}}$$

# Vergleich der Erwartungswerte

---

Die Teststatistik

$$T = \frac{\bar{x} - \bar{y}}{\widehat{SD}}$$

hat approximativ eine Standardnormalverteilung.

Zur Entscheidung über Nullhypothese gegen Alternative vergleiche  $T$  mit den (einseitigen bzw. zweiseitigen) Quantilen der Standardnormalverteilung.

# Vergleich der Erwartungswerte

---

## Beispiel: Skript

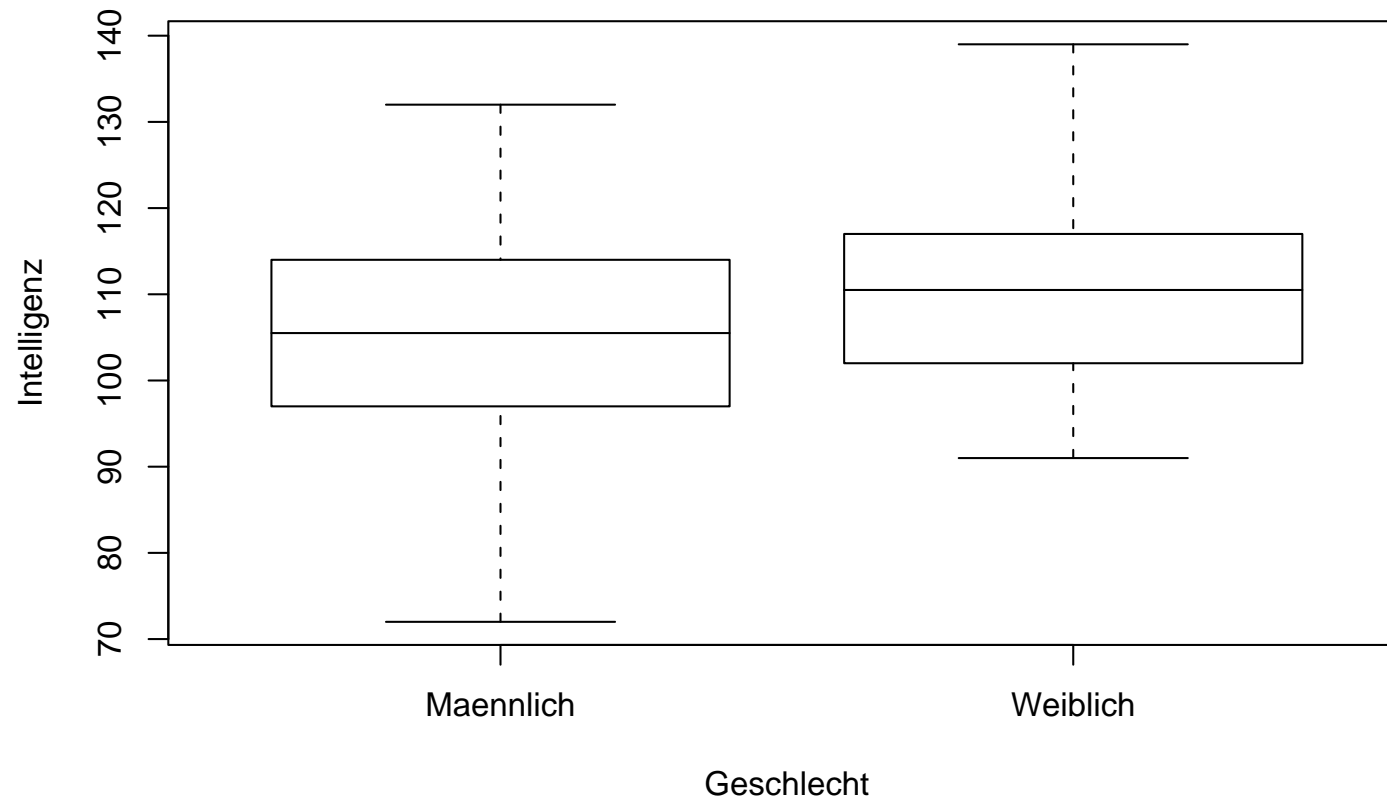
**(8.3)** Bestehen im Weltraum-Datensatz beim Merkmal Intelligenz signifikante Unterschiede zwischen Männern und Frauen?

Wir haben zwei Stichproben mit  $n_1 = 50$  und  $n_2 = 50$ . Dabei ist bei den Männern  $\bar{x} = 104.72$  und  $s_{x,n-1} = 12.387$  und bei den Frauen  $\bar{y} = 110.36$  und  $s_{y,n-1} = 11.133$ . Damit ist

$$\widehat{SD} = \sqrt{\frac{12.387^2}{50} + \frac{11.133^2}{50}} = 2.355$$

# Vergleich der Erwartungswerte

---





## Vergleich der Erwartungswerte

---

Die Teststatistik ist somit

$$T = \frac{\bar{x} - \bar{y}}{\widehat{SD}} = \frac{104.72 - 110.36}{2.355} = -2.39$$

Diese unterschreitet den kritischen Wert  $-2$  (Faustregel) und es kann  $\mu_1 < \mu_2$  nachgewiesen werden.

## Vergleich der Erwartungswerte

---

### Beispiel: Aufgabensammlung

**214.** Wartezeit an der Kassa bei zwei Supermärkten mit verschiedenem Kassensystem (in Minuten):

Supermarkt	Sparag	Consumo
Anzahl der befragten Kunden	50	80
Mittelwert	11.2	7.3
Stichprobenvarianz $s_{n-1}^2$	16.4	10.2

Überprüfen sie mittels eines statistischen Tests die Hypothese "Bei Consumo wartet man im Mittel solange an der Kassa wie bei Sparag." Führen sie einen Test auf Gleichheit der Erwartungswerte durch. Geben sie den Absolutbetrag der Testgröße an.

# Varianzanalyse

---

Das bisherige Prüfverfahren läßt zu, daß die Varianzen  $V(X) = \sigma_1^2$  und  $V(Y) = \sigma_2^2$  unterschiedlich sind.

Falls angenommen werden kann, daß  $\sigma_1^2 = \sigma_2^2$ , dann ist ein genaueres Prüfverfahren für den Vergleich von  $\mu_1$  und  $\mu_2$  möglich. Verwende:

$$\widehat{SD} = \sqrt{\frac{n_1 + n_2}{n_1 \cdot n_2} \cdot \frac{n_1 s_x^2 + n_2 s_y^2}{n_1 + n_2 - 2}}$$

Äquivalent dazu ist wieder eine Varianzanalyse.

# Varianzanalyse

---

Wenn die Hypothese zutrifft, dann ist  $\mu = \mu_1 = \mu_2$ .

Dieser **gemeinsame Erwartungswert** kann geschätzt werden durch den **Gesamtmittelwert**:

$$\bar{m} = \frac{n_1 \bar{x} + n_2 \bar{y}}{n_1 + n_2}$$

Die Abweichung der Daten um diesen Gesamtmittelwert kann wieder zerlegt werden.

# Varianzanalyse

---

$$\begin{aligned}SS_{ZW} &= \sum_{i=1}^{n_1} (\bar{x} - m)^2 + \sum_{i=1}^{n_2} (\bar{y} - m)^2 \\&= \frac{n_1 n_2}{n_1 + n_2} (\bar{x} - \bar{y})^2 \\&= \textbf{systematischer Unterschied} \text{ zwischen Stichproben}\end{aligned}$$

$$\begin{aligned}SS_{IN} &= \sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{i=1}^{n_2} (y_i - \bar{y})^2 \\&= n_1 s_x^2 + n_2 s_y^2 \\&= \textbf{innere Streuung} \text{ der Stichproben}\end{aligned}$$

# Varianzanalyse

---

$$\begin{aligned} SS_T &= \sum_{i=1}^{n_1} (x_i - m)^2 + \sum_{i=1}^{n_2} (y_i - m)^2 \\ &= \textbf{Gesamtstreuung} \text{ der Daten} \end{aligned}$$

Es gilt folgende Streuungszerlegung:

$$SS_T = SS_{ZW} + SS_{IN}$$

# ANOVA-Tabelle

---

	$SS$	$df$	$MSS$
$ZW$	$SS_{ZW}$	1	$MSS_{ZW} = \frac{SS_{ZW}}{1}$
$IN$	$SS_{IN}$	$n_1 + n_2 - 2$	$MSS_{IN} = \frac{SS_{IN}}{n_1 + n_2 - 2}$
	$SS_T$	$n_1 + n_2 - 1$	

Relevanzproblem

$$\frac{SS_{ZW}}{SS_T}$$

Signifikanzproblem

$$F = \frac{MSS_{ZW}}{MSS_{IN}}$$

# ANOVA-Tabelle

---

	$SS$	$df$	$MSS$
$ZW$	$\frac{n_1 n_2}{n_1 + n_2} (\bar{x} - \bar{y})^2$	1	$\frac{n_1 n_2}{n_1 + n_2} (\bar{x} - \bar{y})^2$
$IN$	$n_1 s_x^2 + n_2 s_y^2$	$n_1 + n_2 - 2$	$\frac{n_1 s_x^2 + n_2 s_y^2}{n_1 + n_2 - 2}$
	$SS_T$	$n_1 + n_2 - 1$	

Relevanzproblem

$$\frac{SS_{ZW}}{SS_T}$$

Signifikanzproblem

$$F = \frac{MSS_{ZW}}{MSS_{IN}}$$



# ANOVA

---

## Beispiel: Skript

(8.9) Bestehen im Weltraum-Datensatz beim Merkmal Intelligenz signifikante Unterschiede zwischen Männern und Frauen?

	$SS$	$df$	$MSS$
$ZW$	795.24	1	795.24
$IN$	13591.60	98	138.69
	14386.84	99	

# ANOVA

---

Damit ist die  $F$ -Größe

$$F = \frac{MSS_{ZW}}{MSS_{IN}} = \frac{795.24}{138.67} = 5.734$$

und signifikant, da sie den kritischen Wert 4 (Faustregel) überschreitet.

Das Bestimmtheitsmaß ist

$$\frac{SS_{ZW}}{SS_T} = \frac{795.24}{14386.84} = 0.055$$

Also sind nur etwa 5.5% der Streuung auf den Unterschied zwischen den Geschlechtern zurückzuführen.

# ANOVA

---

## Beispiel:

- 10 afro-amerikanische Männer erhalten Calcium zur Blutdrucksenkung, 11 erhalten ein Placebo.

Der Mittelwert des Blutdruckrückgangs bei den Calcium-Patienten ist  $\bar{x} = 5$  bei einer Varianz von  $s_x^2 = 68.8$ . Für die Placebo-Patienten war  $\bar{y} = -0.273$  und  $s_y^2 = 31.653$ .

Kann eine Wirksamkeit von Calcium als blutdrucksenkendes Mittel nachgewiesen werden?

# ANOVA

---

Die ANOVA-Tabelle ist

	$SS$	$df$	$MSS$
$ZW$	145.628	1	145.628
$IN$	1036.182	19	54.536
	1181.81	20	

Die  $F$ -Größe ist nicht signifikant.

$$F = \frac{MSS_{ZW}}{MSS_{IN}} = \frac{145.628}{54.536} = 2.67$$

Und das Bestimmtheitsmaß ist  $\frac{SS_{ZW}}{SS_T} = 0.123$ .

# ANOVA

---

## Beispiel: Aufgabensammlung

**221.** Wartezeit an der Kassa bei zwei Supermärkten mit verschiedenem Kassensystem (in Minuten):

Supermarkt	Sparag	Consumo
Anzahl der befragten Kunden	50	80
Mittelwert	11.2	7.3
Stichprobenvarianz $s_{n-1}^2$	16.4	10.2

Überprüfen sie mittels eines statistischen Tests die Hypothese "Bei Consumo wartet man im Mittel solange an der Kassa wie bei Sparag." Beantworten sie diese Frage mittels der ANOVA-Tabelle und geben sie die  $F$ -Größe an.

# ANOVA

---

## Beispiel:

- Bei einem Zweistichprobenproblem ergab sich die ANOVA-Tabelle

	$SS$	$df$	$MSS$
$ZW$	63.42	*	*
$IN$	71.77	5	*
	*	*	

# ANOVA

---

## Beispiel:

- Bei einem Zweistichprobenproblem ergab sich die ANOVA-Tabelle

	$SS$	$df$	$MSS$
$ZW$	63.42	1	63.42
$IN$	71.77	5	14.354
	135.19	6	

$$F = \frac{MSS_{ZW}}{MSS_{IN}} = \frac{63.42}{14.354} = 4.418$$

$$\frac{SS_{ZW}}{SS_T} = \frac{63.42}{135.19} = 0.469$$

# ANOVA

---

Welche der folgenden Behauptungen ist richtig?

- (a) Das Bestimmtheitsmaß liegt zwischen 0.1 und 0.5.
- (b) Die  $F$ -Größe ist größer als 4.
- (c) Das Bestimmtheitsmaß liegt zwischen 0.05 und 0.4.
- (d) Die Hypothese, daß die beiden Erwartungswerte gleich sind, kann bei einem Signifikanzniveau von 95% nicht verworfen werden.
- (e) Die  $F$ -Größe ist kleiner als 5.



# ANOVA

---

Welche der folgenden Behauptungen ist richtig?

- (a) Das Bestimmtheitsmaß liegt zwischen 0.1 und 0.5. **richtig**
- (b) Die  $F$ -Größe ist größer als 4. **richtig**
- (c) Das Bestimmtheitsmaß liegt zwischen 0.05 und 0.4. **falsch**
- (d) Die Hypothese, daß die beiden Erwartungswerte gleich sind, kann bei einem Signifikanzniveau von 95% nicht verworfen werden. **falsch**
- (e) Die  $F$ -Größe ist kleiner als 5. **richtig**

# Zusammenfassung Kapitel 8

---

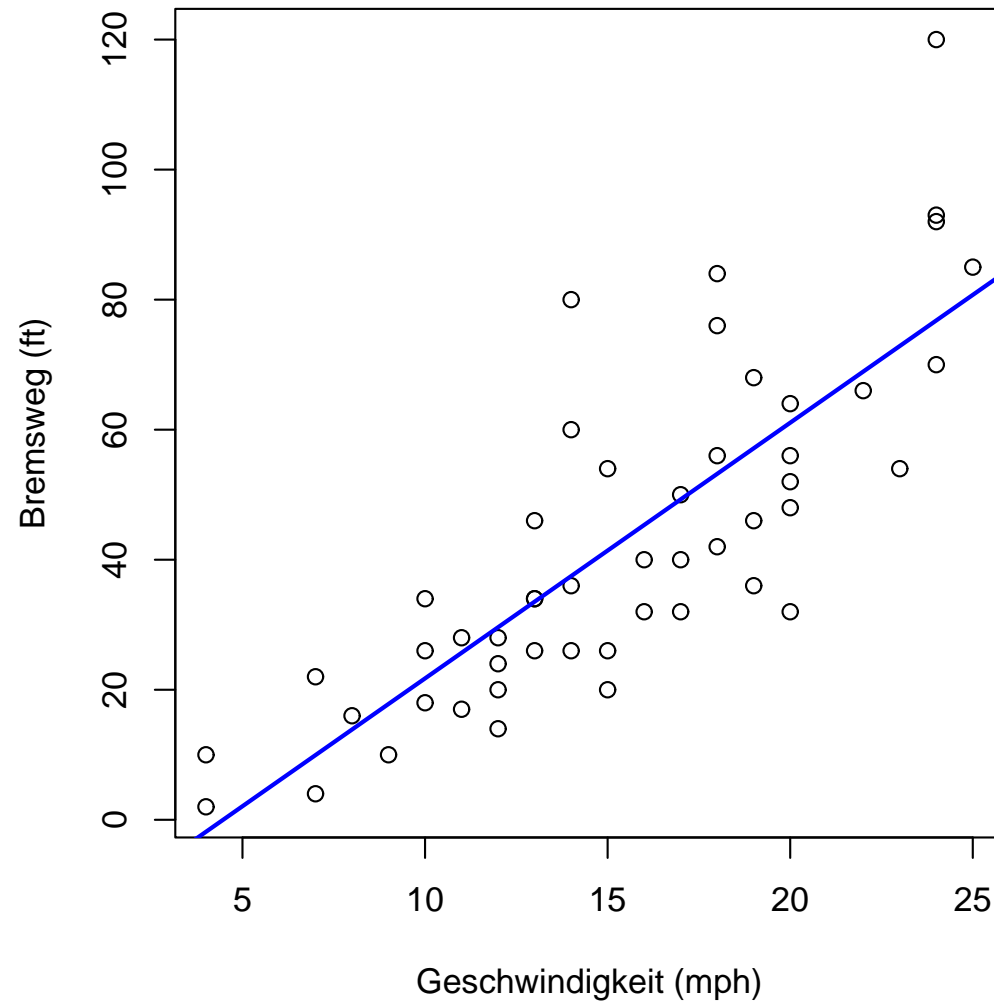
- Vergleich der Mittelwerte zweier Stichproben
- Annahme der Gleichheit der Varianzen
- Streuungszerlegung für das Zweistichprobenproblem
- Varianzanalyse,  $F$ -Größe, Bestimmtheitsmaß

# Kontingenztafeln

## Kapitel 9

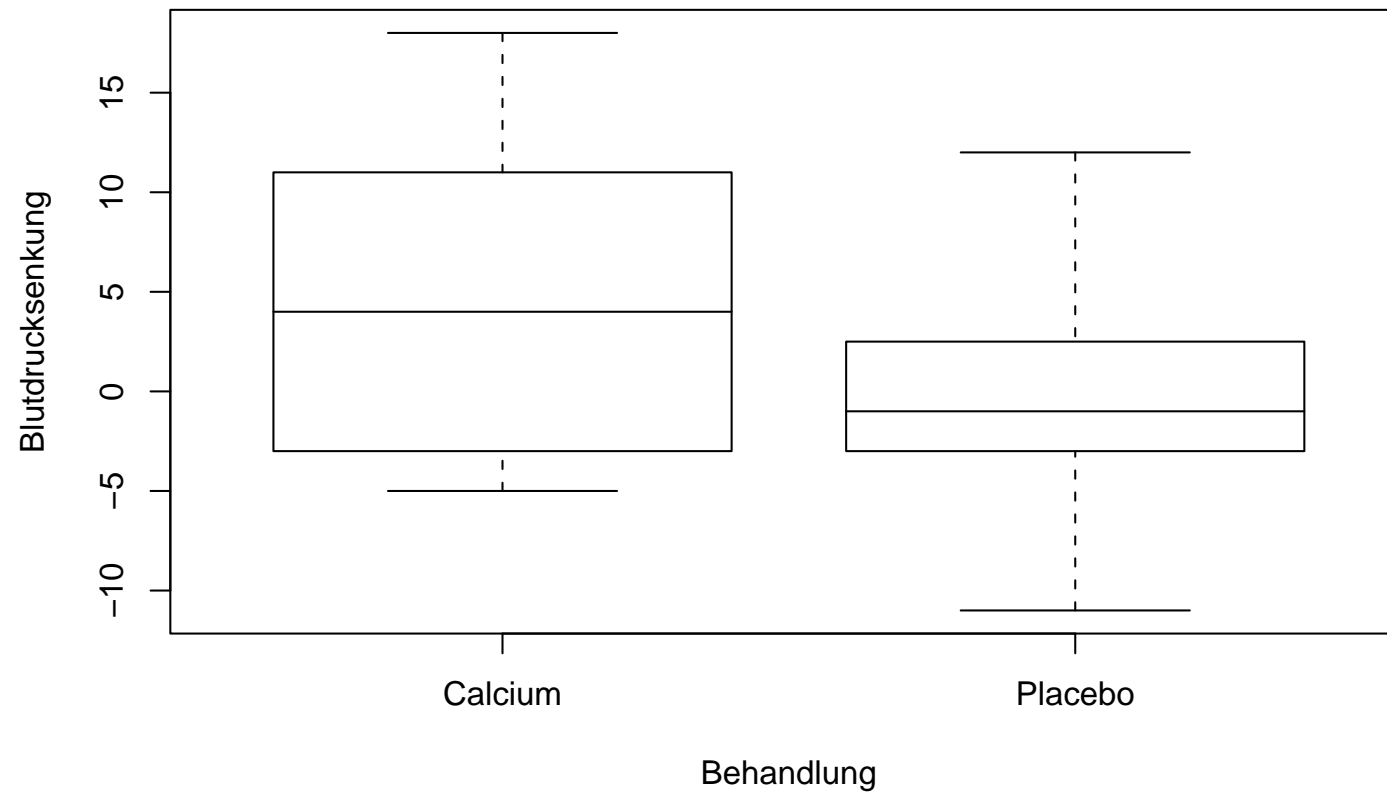
# Kapitel 7

---



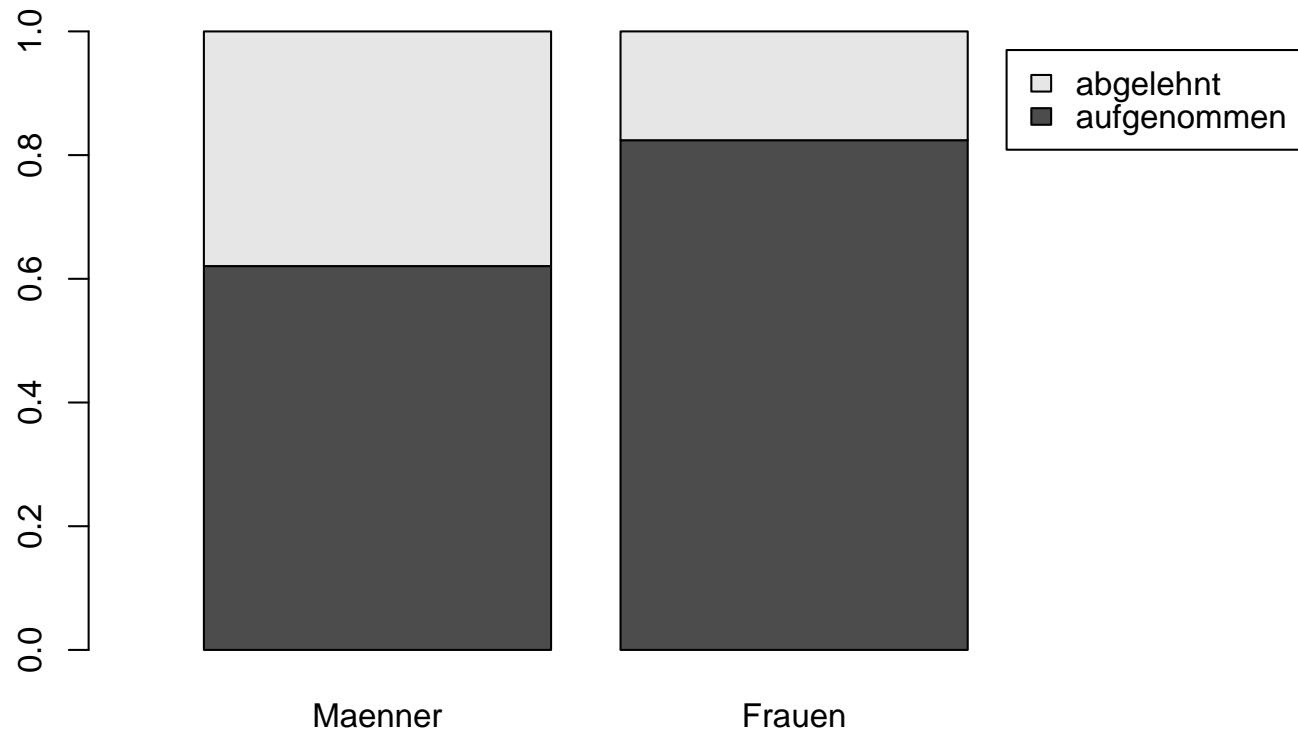
# Kapitel 8

---



# Kapitel 9

---



# Überblick

---

**Problem:** Betrachte Kopplung von 2 qualitativen Merkmalen in einer Häufigkeitstabelle.

**Frage:** Gibt es einen Zusammenhang zwischen den beiden Merkmalen?

**Beispiel:** Hohe Geschwindigkeit bei der sich ein Unfall ereignet und tödlicher Ausgang.

# Überblick

---

Es sei  $A$  = „Unfall endet tödlich“ und  
 $B$  = „Unfall ereignet sich bei mehr als 150 km/h“.

	$B$	$B'$	
$A$	80	200	
$A'$	20	700	



# Überblick

---

Es sei  $A$  = „Unfall endet tödlich“ und  
 $B$  = „Unfall ereignet sich bei mehr als 150 km/h“.

	$B$	$B'$	
$A$	80	200	280
$A'$	20	700	720
	100	900	1000

# Bedingte Wahrscheinlichkeiten

---

**Formal:** Die beiden Merkmale haben jeweils 2 mögliche Ausprägungen  $A, A'$  bzw.  $B, B'$ .

**Kontingenztafel (Vierfeldertafel):**

	$B$	$B'$	
$A$	$h(A \cap B)$	$h(A \cap B')$	$h(A)$
$A'$	$h(A' \cap B)$	$h(A' \cap B')$	$h(A')$
	$h(B)$	$h(B')$	$n$

Zeilen- und Spaltensummen sind die Häufigkeitsverteilungen der einzelnen Merkmale = **Randverteilungen**.

# Bedingte Wahrscheinlichkeiten

---

**Gewöhnliche relative Häufigkeiten** werden auf den Umfang  $n$  des gesamten Datensatzes bezogen:

$$f(A \cap B) = \frac{h(A \cap B)}{n}.$$

**Bedingte relative Häufigkeiten von  $A$  unter der Bedingung  $B$**

$$f(A|B) = \frac{h(A \cap B)}{h(B)} = \frac{f(A \cap B)}{f(B)}$$

.

Wie oft wird zusätzlich  $A$  beobachtet, unter all jenen Fällen, in denen  $B$  beobachtet wurde.

# Bedingte Wahrscheinlichkeiten

---

Stammen die Daten aus einem Zufallsexperiment, dann besitzen die Ereigniskombinationen auch Wahrscheinlichkeiten.

Wahrscheinlichkeitstabelle:

	$B$	$B'$	
$A$	$P(A \cap B)$	$P(A \cap B')$	$P(A)$
$A'$	$P(A' \cap B)$	$P(A' \cap B')$	$P(A')$
	$P(B)$	$P(B')$	1

Nach dem empirischen Gesetz der großen Zahl sind diese Wahrscheinlichkeiten die Grenzwerte der entsprechenden relativen Häufigkeiten.

# Bedingte Wahrscheinlichkeiten

---

Die bedingten relativen Häufigkeiten konvergieren für  $n \rightarrow \infty$  gegen einen Grenzwert:

$$f_n(A|B) = \frac{f_n(A \cap B)}{f_n(B)} \rightarrow P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

$P(A|B)$  heißt **bedingte Wahrscheinlichkeit von  $A$  unter der Bedingung  $B$** , wobei  $P(B) \neq 0$ .

**Produktformel:**  $P(A \cap B) = P(A|B) P(B)$ .

# Bedingte Wahrscheinlichkeiten

---

## Beispiel: Skript

(9.1) Ausgang und Geschwindigkeit von Unfällen mit absoluten Häufigkeiten

	$B$	$B'$	
$A$	80	200	280
$A'$	20	700	720
	100	900	1000

$$f(A|B) = \frac{h(A \cap B)}{h(B)} = 0.8 \quad \text{und} \quad f(A|B') = \frac{h(A \cap B')}{h(B')} = 0.222.$$

# Bedingte Wahrscheinlichkeiten

---

## Beispiel: Skript

(9.1) Ausgang und Geschwindigkeit von Unfällen mit relativen Häufigkeiten

	$B$	$B'$	
$A$	0.08	0.2	0.28
$A'$	0.02	0.7	0.72
	0.10	0.9	1

$$f(A|B) = \frac{f(A \cap B)}{f(B)} = 0.8 \quad \text{und} \quad f(A|B') = \frac{f(A \cap B')}{f(B')} = 0.222.$$

# Bedingte Wahrscheinlichkeiten

---

**Formel für die „inverse“ Wahrscheinlichkeit:**

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

**Beispiel:** Wie groß ist die Wahrscheinlichkeit, daß der Unfall bei mehr als 150 km/h stattfand, gegeben, daß er tödlich ausging?

$$f(B|A) = f(A|B) \frac{f(B)}{f(A)} = 0.8 \frac{0.1}{0.28} = 0.286$$



# Bedingte Wahrscheinlichkeiten

---

**Formel für die „totale“ Wahrscheinlichkeit:**

Es sei  $(B_1, \dots, B_m)$  eine Zerlegung. Dann gilt:

$$P(A) = P(A|B_1) P(B_1) + \dots + P(A|B_m) P(B_m).$$

**Beispiel:** Wie groß ist die Wahrscheinlichkeit für tödliche Unfälle?

$$\begin{aligned} f(A) &= f(A|B) f(B) + f(A|B') f(B') \\ &= 0.8 \cdot 0.1 + 0.222 \cdot 0.9 \\ &= 0.28 \end{aligned}$$

# Bedingte Wahrscheinlichkeiten

---

## Formel von Bayes:

Es sei  $(B_1, \dots, B_m)$  eine Zerlegung.

$$P(B_i|A) = \frac{P(A|B_i) P(B_i)}{P(A|B_1) P(B_1) + \dots + P(A|B_m) P(B_m)}$$

für  $i = 1, \dots, m$ .

# Bedingte Wahrscheinlichkeiten

---

## Beispiel: Skript

**(9.5), erweitert:** Ein Unternehmen produziert zwei Sorten von Produkten. 26% der Produkte gehören zur Sorte 2. 4% aller Produkte sind Ausschuß. Von den einwandfreien Produkten gehören 75% zur Sorte 1.

Wie groß ist die Wahrscheinlichkeit, daß ein zufällig ausgewähltes Produkt zur Sorte 1 gehört und einwandfrei ist?

Wie groß ist die Wahrscheinlichkeit, daß ein zufällig ausgewähltes Produkt der Sorte 1 einwandfrei ist?

# Bedingte Wahrscheinlichkeiten

---

## Beispiel:

Ein Unternehmen produziert zwei Sorten von Produkten. 18% der Produkte sind einwandfrei und gehören zur Sorte 1. 42% aller Produkte sind einwandfrei. 68% aller fehlerhaften Produkte gehören zur Sorte 2.

Wie groß ist die Wahrscheinlichkeit, daß ein zufällig ausgewähltes einwandfreies Produkt zur Sorte 2 gehört?

# Bedingte Wahrscheinlichkeiten

---

**Entscheidungsproblem:** Untersuchungsgegenstand besitzt  $n$  verschiedene Zustände  $Z_1, Z_2, \dots, Z_n$ , sind aber nicht direkt beobachtbar sind.

**Gesucht:** Entscheidungsmechanismus, der eine Entscheidung über den wahren Zustand trifft.

**Beispiele:** Eignungstests, medizinische Labortests, industrielle Qualitätskontrolle und alle statistischen Tests von Hypothesen.

## Bedingte Wahrscheinlichkeiten

---

**Beispiel, Labortest:** Ein medizinischer Labortest werde eingesetzt, um zu entscheiden, ob eine Untersuchungsperson an einer bestimmten Krankheit leidet ( $K_+$ ) oder nicht ( $K_-$ ). Der Labortest führt zu den Entscheidungen  $E_+$ : „Der Patient leidet an der Krankheit“, und  $E_-$ : „Der Patient leidet nicht an der Krankheit“. Die **Verlässlichkeit** des Labortests sei gegeben durch:

$$P(E_+|K_+) = 0.95 :$$

95% der tatsächlich Erkrankten werden als solche erkannt.

$$P(E_-|K_-) = 0.80 :$$

80% der nicht Erkrankten werden als solche erkannt.

# Bedingte Wahrscheinlichkeiten

---

Das ergibt die **Fehlerwahrscheinlichkeiten**

$$\begin{aligned}P(E_-|K_+) &= 1 - P(E_+|K_+) = 0.05 \\P(E_+|K_-) &= 1 - P(E_-|K_-) = 0.2\end{aligned}$$

Für die Patienten sind die folgenden **a posteriori Wahrscheinlichkeiten** interessant:

$P(K_+|E_+) = ?$  Wieviele der als krank eingestuften Untersuchungspersonen sind tatsächlich krank ?

$P(K_-|E_-) = ?$  Wieviele der als gesund eingestuften Untersuchungspersonen sind tatsächlich gesund ?

## Bedingte Wahrscheinlichkeiten

---

$$P(K_+|E_+) = \frac{P(E_+|K_+)P(K_+)}{P(E_+|K_+)P(K_+) + P(E_+|K_-)P(K_-)}$$
$$P(K_-|E_-) = \frac{P(E_-|K_-)P(K_-)}{P(E_-|K_+)P(K_+) + P(E_-|K_-)P(K_-)}$$

Offenbar genügt es nicht, die Verlässlichkeit bzw. die Fehlerwahrscheinlichkeiten des Labortests zu kennen. Es müssen darüber hinaus die **a priori Wahrscheinlichkeiten**  $P(K_+)$  und  $P(K_-)$  der mögliche Zustände des Patienten bekannt sein.



## Bedingte Wahrscheinlichkeiten

---

Annahme 1: Die Krankheit sei relativ häufig, z.B.  $P(K_+) = 0.7$ .

$$P(K_+|E_+) = \frac{0.95 \cdot 0.7}{0.95 \cdot 0.7 + 0.2 \cdot 0.3} = 0.91$$

$$P(K_-|E_-) = \frac{0.8 \cdot 0.3}{0.05 \cdot 0.7 + 0.8 \cdot 0.3} = 0.87$$

Es sind also beide Entscheidungen als einigermaßen verlässlich anzusehen.

## Bedingte Wahrscheinlichkeiten

---

Annahme 2: Die Krankheit sei eher selten, z.B.  $P(K_+) = 0.05$ .

$$P(K_+|E_+) = \frac{0.95 \cdot 0.05}{0.95 \cdot 0.05 + 0.2 \cdot 0.95} = 0.2$$

$$P(K_-|E_-) = \frac{0.8 \cdot 0.95}{0.05 \cdot 0.05 + 0.8 \cdot 0.95} = 0.997.$$

Ein pathologisches Testergebnis ist unter diesen Umständen sehr kritisch zu bewerten.

## Bedingte Wahrscheinlichkeiten

---

Die Indifferenzannahme setzt die a priori Wahrscheinlichkeiten der beiden Zustände gleich groß, also gleich 0.5 sind an. In unserem Fall gilt dann

$$P(K_+|E_+) = \frac{0.95 \cdot 0.5}{0.95 \cdot 0.5 + 0.2 \cdot 0.5} = \frac{1}{1 + \frac{0.2}{0.95}} = 0.83$$

$$P(K_-|E_-) = \frac{0.8 \cdot 0.5}{0.05 \cdot 0.5 + 0.8 \cdot 0.5} = \frac{1}{1 + \frac{0.05}{0.8}} = 0.94$$

# Bedingte Wahrscheinlichkeiten

---

## Beispiel: Aufgabensammlung

**238.** 20% eines Produkts weisen (nach der Produktion) Mängel auf. Das Produkt wird vor dem Verkauf getestet. Der Test stuft 80% der mangelhaften und 10% der fehlerfreien Produkte als mangelhaft ein. Wie groß ist unter den Produkten, die in den Handel kommen, der Anteil der fehlerhaften (in Prozent)?

# Bedingte Wahrscheinlichkeiten

---

## Beispiel:

Die Krankheit ASDS (Acute Statistics Deficiency Syndrom) verhindert, daß man ein Studium an der WU erfolgreich abschließen kann. 5% aller Studenten leiden an ASDS.

Es gibt einen Test, der diese Schwäche nachweisen kann: die Prüfung der Statistik-Vorlesung. Wer ASDS hat, wird mit 90%iger Sicherheit durchfallen – wer es nicht hat, wird mit 90%iger Sicherheit bestehen.

Das Unglück ist passiert: ein Student ist in der Prüfung durchgefallen. Sollte er nun das Studium beenden?

# Gekoppelte Ereignisse

---

Im Rahmen eines Zufallsexperiments untersuchen wir die Ereignisse  $A$  und  $B$ . Wir interessieren uns für die Frage, ob die Ereignisse einander behindern oder begünstigen.

Zwei Ereignisse  $A$  und  $B$  **begünstigen einander** oder sind **positiv gekoppelt**, wenn

$$P(A \cap B) > P(A) P(B).$$

Denn „ $B$  begünstigt  $A$ “, wenn  $P(A|B) > P(A)$  und „ $A$  begünstigt  $B$ “, wenn  $P(B|A) > P(B)$ .

# Gekoppelte Ereignisse

---

Zwei Ereignisse  $A$  und  $B$  **behindern einander** oder sind **negativ gekoppelt**, wenn

$$P(A \cap B) < P(A) P(B).$$

Zwei Ereignisse  $A$  und  $B$  heißen **gekoppelt** oder **stochastisch abhängig**, wenn

$$P(A \cap B) \neq P(A) P(B).$$

Sie heißen **stochastisch unabhängig**, wenn

$$P(A \cap B) = P(A) P(B).$$

# Richtung der Koppelung

---

tatsächlich

	$B$	$B'$	
$A$	$P(A \cap B)$	$P(A \cap B')$	$P(A)$
$A'$	$P(A' \cap B)$	$P(A' \cap B')$	$P(A')$
	$P(B)$	$P(B')$	

bei Unabhängigkeit

	$B$	$B'$	
$A$	$P(A)P(B)$	$P(A)P(B')$	$P(A)$
$A'$	$P(A')P(B)$	$P(A')P(B')$	$P(A')$
	$P(B)$	$P(B')$	



# Richtung der Koppelung

---

Vorzeichenmuster der Differenzen der tatsächlichen Tabelleneinträge und der Einträge bei Unabhängigkeit:

Positive Koppelung

	$B$	$B'$
$A$	+	-
$A'$	-	+

Negative Koppelung

	$B$	$B'$
$A$	-	+
$A'$	+	-

# Vierfelderkorrelation

---

## Vierfelderkorrelation:

$$\rho = \rho(A, B) = \frac{P(A \cap B) - P(A)P(B)}{\sqrt{P(A) P(A') P(B) P(B')}}.$$

- $-1 \leq \rho(A, B) \leq 1$ .
- $A$  und  $B$  sind genau dann **stochastisch unabhängig**, wenn  $\rho(A, B) = 0$  ist.
- $A$  und  $B$  sind genau dann **positiv gekoppelt**, wenn  $\rho(A, B) > 0$  ist. Speziell:  $\rho(A, B) = 1$  für  $A = B$ .
- $A$  und  $B$  sind genau dann **negativ gekoppelt**, wenn  $\rho(A, B) < 0$  ist. Speziell:  $\rho(A, B) = -1$  für  $A = B'$ .

# Vierfelderkorrelation

---

Das Vorzeichen  $\text{sign}\rho$  gibt also die **Richtung der Koppelung** an.

Der Betrag  $|\rho|$  gibt die **Stärke der Koppelung** an.

Eine bestehende Koppelung ist kein Beweis für einen kausalen Zusammenhang.

# Vierfelderkorrelation

---

**Beispiel:** Skript

**(9.22)**

$R$  ... Der PKW weist Rostschäden auf.

$S$  ... Der PKW besitzt eine Hohlraumversiegelung.

$P(R) = 0.37$ ,  $P(S) = 0.71$  und  $P(R \cap S) = 0.11$ .

Untersuchen Sie die Koppelung der Merkmale.

Lösung: Da  $P(R \cap S) = 0.11 < P(R)P(S) = 0.2626$ , sind die beiden Ereignisse  $R$  und  $S$  negativ gekoppelt, d.h. sie behindern einander.

## Vierfelderkorrelation

---

Vergleich der Vierfeldertafel der **Wahrscheinlichkeiten** mit der bei Unabhängigkeit erwarteten Vierfeldertafel:

	tatsächlich		
	$S$	$S'$	
$R$	0.11	0.26	0.37
$R'$	0.60	0.03	0.63
	0.71	0.29	

	bei Unabhängigkeit		
	$S$	$S'$	
$R$	0.2627	0.1073	0.37
$R'$	0.4473	0.1827	0.63
	0.71	0.29	

## Vierfelderkorrelation

---

Die Tafel der **Differenzen** lautet daher

	$S$	$S'$	
$R$	$-0.1527$	$0.1527$	$0$
$R'$	$0.1527$	$-0.1527$	$0$
	$0$	$0$	

## Vierfelderkorrelation

---

Interpretation der Koppelung: Aus dem Bestehen der negativen Koppelung allein ist nicht ersichtlich, ob und in welchem Sinn eine kausale Beziehung zwischen den Ereignissen besteht.

- Möglich: Hohlraumversiegelung eines Autos schützt vor Rost.
- Unplausibel: Geringe Rostschäden sind Ursache für Hohlraumversiegelungen.
- Versteckter Faktor Pflegeaufwand wirkt auf Rostschäden und auf Neigung zur Hohlraumversiegelung. Deshalb ist anzunehmen, daß ein gewisser Teil der negativen Koppelung zwischen Rostschäden und Hohlraumversiegelung eine Scheinkoppelung ist.

# Kontingenzproblem

---

Gegeben empirische relative Häufigkeiten.

**Kontingenztafel (Vierfeldertafel):**

	$B$	$B'$	
$A$	$f(A \cap B)$	$f(A \cap B')$	$f(A)$
$A'$	$f(A' \cap B)$	$f(A' \cap B')$	$f(A')$
	$f(B)$	$f(B')$	

Selbst wenn die Ereignisse  $A$  und  $B$  stochastisch unabhängig sind, wird die Kontingenztafel auf Grund von Zufallsschwankungen eine gewisse Koppelung aufweisen.



# Kontingenzproblem

---

## Empirische Vierfelderkorrelation:

$$\hat{\rho} = r = \frac{f(A \cap B) - f(A)f(B)}{\sqrt{f(A)f(A')f(B)f(B')}}.$$

$r$  ist ein Schätzer der Vierfelderkorrelation  $\rho$ .

Sind die Ereignisse stochastisch unabhängig, dann ist zwar  $\rho = 0$ , aber die empirische Vierfelderkorrelation kann auf Grund zufälliger Schwankungen von Null verschieden sein. Ihre Standardabweichung beträgt etwa  $\frac{1}{\sqrt{n}}$ .

# Kontingenzproblem

---

## Test für das Kontingenzproblem:

Nullhypothese:  $A$  und  $B$  sind stochastisch unabhängig

Alternative:  $A$  und  $B$  sind gekoppelt

Die Testgröße ist die standardisierte empirische Vierfelderkorrelation:

$$T = \sqrt{n}r.$$

# Kontingenzproblem

---

$-2 \leq \sqrt{n}r \leq 2$ : Das Ergebnis ist nicht signifikant.  $r$  widerspricht nicht der Nullhypothese.

$\sqrt{n}r < -2$ : Das Ergebnis ist signifikant. Nachweis einer negativen Koppelung.

$\sqrt{n}r > 2$ : Das Ergebnis ist signifikant. Nachweis einer positiven Koppelung.

# Kontingenzproblem

---

## Beispiel: Aufgabensammlung

231. erweitert:

$A$  ... Besitz eines Mobiltelefons

$B$  ... Nutzung von alternativen Festnetzangeboten

Befragte Personen	$A$	$B$	$A \cap B$
580	330	120	30

Wie lautet die empirische Vierfelderkorrelation? Sind die beiden Merkmale unabhängig?

# Kontingenzproblem

---

## Beispiel:

- Um zu beurteilen, ob sich die Werbekampagne für ein bestimmtes Produkt lohnt, befragt der Hersteller die potentiellen Kunden in folgender Studie:

In Stadt A führt man die Werbekampagne durch und befragt im Anschluss 100 zufällig ausgewählte Personen, ob sie das Produkt kaufen würden oder nicht. Als Vergleich werden in Stadt B 50 zufällig ausgewählte Personen befragt; dort wurde allerdings keinerlei Werbung für das Produkt gemacht.

# Kontingenzproblem

---

Das Ergebnis wurde in folgender Tabelle zusammengefaßt:

Würden Sie das Produkt kaufen?		
	Ja	Nein
Stadt A	74	26
Stadt B	19	31

War die Werbekampagne erfolgreich?

# Symmetrieproblem

---

Ein **Test für den Vergleich zweier Wahrscheinlichkeiten** im Rahmen eines **Symmetrieproblems** ist ein Verfahren, welches eine Entscheidung zwischen den Aussagen

$$\text{Nullhypothese: } P(A) = P(B)$$

$$\text{Alternative: } P(A) \neq P(B)$$

herbeiführt. Die Entscheidung wird auf Grund von empirischen Daten getroffen.

# Symmetrieproblem

---

Für den Test werden nur jene Daten betrachtet, wo genau eines der beiden Ereignisse  $A$  oder  $B$  eintritt:

$$C = (A \cap B') \cup (A' \cap B).$$

	$B$	$B'$	
$A$	$f(A \cap B)$	$f(A \cap B')$	$f(A)$
$A'$	$f(A' \cap B)$	$f(A' \cap B')$	$f(A')$
	$f(B)$	$f(B')$	



# Symmetrieproblem

---

Die Ereignisse  $A$  und  $B$  sind genau dann gleichwahrscheinlich, wenn  $P(A|C) = 0.5$ .

Testgröße für das Symmetrieproblem ist daher der Standardscore der relativen Häufigkeit  $f(A|C)$ :

$$\frac{f(A|C) - 0.5}{\sqrt{\frac{0.5 \cdot 0.5}{h(C)}}}.$$

# Symmetrieproblem

---

**Beispiel:** Aufgabensammlung

**232.**  $A$  ... Besitz eines Mobiltelefons

$B$  ... Nutzung von alternativen Festnetzangeboten

Befragte Personen	$A$	$B$	$A \cap B$
580	330	120	30

Entscheiden sie, ob der Anteil der Besitzer von Mobiltelefonen und Nutzern von alternativen Festnetzangeboten gleich groß ist.

# Symmetrieproblem

---

## Beispiel:

- Um zu beurteilen, ob sich die Werbekampagne für ein bestimmtes Produkt lohnt, befragt der Hersteller die potentiellen Kunden in folgender Studie:

Es werden 256 Personen zufällig ausgewählt und gefragt, ob sie das Produkt kaufen würden oder nicht. Im Monat nach dieser Befragung führt man die Werbekampagne durch und im darauffolgenden Personen befragt man dieselben Personen noch einmal, ob sie nun das Produkt kaufen würden oder nicht.

# Symmetrieproblem

---

Das Ergebnis der Befragung vor und nach der Kampagne wurde in folgender Tabelle zusammengefaßt:

		Würden Sie das Produkt kaufen?	
		Vorher:	
		Ja	Nein
Nachher:	Ja	116	44
	Nein	31	65

War die Werbekampagne erfolgreich?

# Zusammenfassung Kapitel 9

---

- Kontingenztafeln
- Bedingte Wahrscheinlichkeiten, Produktformel
- Formel “inverse” Wahrscheinlichkeit, Formel “totale” Wahrscheinlichkeit, Formel von Bayes
- Gekoppelte Ereignisse
- Kontingenzproblem
- Symmetrieproblem