



Teil 1

Grundlagen

Mathematische Grundlagen

5. November 2022

Summenzeichen • Produktzeichen • Potenzen • Wurzeln • Logarithmen •
Binomialkoeffizienten • Funktionen • Ableitungen • Integrale • Software

Summenzeichen

Definition:

Das Summenzeichen \sum verwendet man zur Abkürzung der Schreibweise von Summen:

$$a_1 + a_2 + \dots + a_n = \sum_{i=1}^n a_i$$

Es bezeichnet eine Summe von n Summanden a_i ($i = 1, 2, \dots, n$). Man nennt i *Laufindex* oder *Summenvariable*.

Rechenregeln:

1. Summe gleicher Summanden:

$$\sum_{i=1}^n a = na$$

2. Multiplikation mit einem konstanten Faktor:

$$\sum_{i=1}^n ka_i = k \sum_{i=1}^n a_i$$

3. Aufspalten einer Summe:

$$\sum_{i=1}^n a_i = \sum_{i=1}^m a_i + \sum_{i=m+1}^n a_i, \quad (1 < m < n)$$

4. Addition von Summanden gleicher Länge:

$$\sum_{i=1}^n (a_i + b_i + c_i \dots) = \sum_{i=1}^n a_i + \sum_{i=1}^n b_i + \sum_{i=1}^n c_i + \dots$$

5. Umnummerierung:

$$\sum_{i=1}^n a_i = \sum_{i=m}^{m+n-1} a_{i-m+1}, \quad \sum_{i=m}^n a_i = \sum_{i=l}^{n-m+l} a_{i+m-l},$$

6. Vertauschen der Summationsfolge bei Doppelsummen:

$$\sum_{i=1}^n \sum_{k=1}^m a_{ik} = \sum_{k=1}^m \sum_{i=1}^n a_{ik}$$



Listing 1.1: sum.R

```
1 i <- 1:5
2 sum(i)
3 sum(i^2)
```



Listing 1.2: sum.ma

```
1 Sum[i, {i ,1, 5}]
2 Sum[i^2, {i ,1, 5}]
```



Listing 1.3: sum.py

```
1 import numpy
2 i = numpy.arange(1, 6)
3 print(numpy.sum(i))
4 print(numpy.sum(i*i))
```

Produktzeichen

Definition:

Das Produktzeichen \prod verwendet man zur Abkürzung der Schreibweise von Produkten:

$$a_1 \cdot a_2 \cdot \dots \cdot a_n = \prod_{i=1}^n a_i$$

Es bezeichnet ein Produkt von n Faktoren a_i ($i = 1, 2, \dots, n$), wobei i Laufindex genannt wird.

Rechenregeln:

1. **Produkt gleicher Faktoren**, d.h. $a_i = a$ für $i = 1, 2, \dots, n$:

$$\prod_{i=1}^n a = a^n$$

2. **Vorzeichen konstanter Faktoren**:

$$\prod_{i=1}^n (ka_i) = k^n \prod_{i=1}^n a_i$$

3. **Aufspalten in Teilprodukte**:

$$\prod_{i=1}^n a_i = \prod_{i=1}^m a_i \prod_{i=m+1}^n a_i, \quad (1 < m < n)$$

4. Produkt von Produkten:

$$\prod_{i=1}^n a_i b_i c_i \dots = \prod_{i=1}^n a_i \prod_{i=1}^n b_i \prod_{i=1}^n c_i \dots$$

5. Umnummerierung:

$$\prod_{i=m}^n a_i = \prod_{i=l}^{n-m+l} a_{i+m-l}$$

6. Vertauschen der Produktzeichen bei Doppelprodukten:

$$\prod_{i=1}^n \prod_{k=1}^m a_{ik} = \prod_{k=1}^m \prod_{i=1}^n a_{ik}$$



Listing 1.4: product.R

```
1 i <- 1:5
2 prod(i)
3 prod(i^2)
```



Listing 1.5: product.ma

```
1 Product[i, {i ,1, 5}]
2 Product[i^2, {i ,1, 5}]
```



Listing 1.6: product.py

```
1 import numpy
2 i = numpy.arange(1, 6)
3 print(numpy.prod(i))
4 print(numpy.prod(i*i))
```

Potenzen

Definition:

Die Schreibweise a^x wird für die algebraische Operation des *Potenzierens* verwendet. Wobei a als *Basis* bezeichnet wird, x als *Exponent* und a^x als *Potenz*.

Rechenregeln:

Unter Beachtung der Definitionsbereiche für Basis und Exponent gilt

$$a^x \cdot a^y = a^{x+y}, \quad a^x : a^y = \frac{a^x}{a^y} = a^{x-y}$$

$$a^x \cdot b^x = (a \cdot b)^x, \quad a^x : b^x = \frac{a^x}{b^x} = \left(\frac{a}{b}\right)^x$$

$$(a^x)^y = (a^y)^x = a^{xy}$$

$$a^x = e^{x \log(a)}, (a > 0), \quad e = 2,71828182845\dots$$

Wurzeln

Definition:

Es sei $n > 1$ eine natürliche Zahl. Ist a eine nichtnegative reelle Zahl, so besitzt die Gleichung

$$x^n = a$$

genau eine nichtnegative reelle Lösung. Diese wird als n -te Wurzel aus a bezeichnet. Man schreibt dafür

$$x = \sqrt[n]{a}$$

Rechenregeln:

Für natürliche n , m und reelle a gilt:

$$a^{\frac{1}{n}} = \sqrt[n]{a}, \quad a^{\frac{m}{n}} = \sqrt[n]{a^m}, \quad a^{-\frac{m}{n}} = \frac{1}{\sqrt[n]{a^m}}$$

$$(\sqrt[n]{a})^n = a, \quad \sqrt[n]{a^n} = |a|$$

$$\sqrt[n]{a} \sqrt[n]{b} = \sqrt[n]{ab}, \quad \frac{\sqrt[n]{a}}{\sqrt[n]{b}} = \sqrt[n]{\frac{a}{b}}$$

$$\sqrt[m]{\sqrt[n]{a}} = \sqrt[mn]{a} = \sqrt[n]{\sqrt[m]{a}}$$

$$\sqrt[n]{a^m} = (\sqrt[n]{a})^m$$

$$\sqrt{a} = a^{\frac{1}{2}}, \quad \sqrt{a} \cdot \sqrt{a} = a$$

Logarithmen

Definition:

Unter dem *Logarithmus* einer Zahl $a > 0$ zur Basis $b > 0, \neq 1$, oder als Formel $x = \log_b a$, wird der Exponent der Potenz verstanden, in die b zu erheben ist, um die Zahl a zu erhalten. Somit ergibt sich aus der Gleichung

$$b^x = a \quad \text{die Gleichung} \quad \log_b a = x$$

und umgekehrt folgt aus der zweiten die erste Gleichung.

Es gilt:

- Es gibt keine Logarithmen von 0 oder von negativen Zahlen.
- Jede positive Zahl besitzt für jede beliebige positive Basis ihren Logarithmus, ausgenommen zur Basis $b = 1$.

Rechenregeln:

Das Zeichen \log wird ohne eine angegebene Basis verwendet, da die verwendete Basis keine Rolle spielt.

$$1. \log(a \cdot b) = \log a + \log b$$

$$\log\left(\frac{a}{b}\right) = \log a - \log b$$

$$2. \log a^x = x \cdot \log a$$

$$\log \sqrt[n]{a} = \frac{1}{n} \cdot \log a$$

$$3. \log \prod_{i=1}^n a_i = \sum_{i=1}^n \log a_i$$

4. **Spezielle Gleichungen:** \ln ist der *natürliche Logarithmus* (der Logarithmus zur Basis e). Er wird im Zusammenhang mit Exponentialfunktionen verwendet.

$$\ln e^x = x, \quad e^{\ln x} = x$$

$$\ln e = 1, \quad \ln 1 = 0$$

$$e^{i\pi} + 1 = 0 \quad (\text{Euler-Identität})$$

Binomialkoeffizienten

Definition:

Mit n und k positiv ist der Binomialkoeffizient n über k definiert als

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

wobei $n!$ das Produkt der ganzen positiven Zahlen von 1 bis n Fakultät genannt wird.

$$n! = 1 \cdot 2 \cdot 3 \cdot \dots \cdot n$$

Rechenregeln:

$$0! = 1$$

$$\binom{n}{0} = 1, \quad \binom{n}{1} = n, \quad \text{für alle } n \geq 0$$

$$\binom{n}{k} = \binom{n}{n-k} = \frac{n!}{k!(n-k)!}, \quad \text{für } 0 \leq k \leq n$$

$$\binom{n}{k} = 0, \quad \text{für } k > n$$



Listing 1.7: binomial.R

```
1 n <- 5
2 k <- 3
3 choose(n, k)
```



Listing 1.8: binomial.ma

```
1 Binomial[5, 3]
```



Listing 1.9: binomial.py

```
1 import math
2 print(math.comb(5, 3))
```

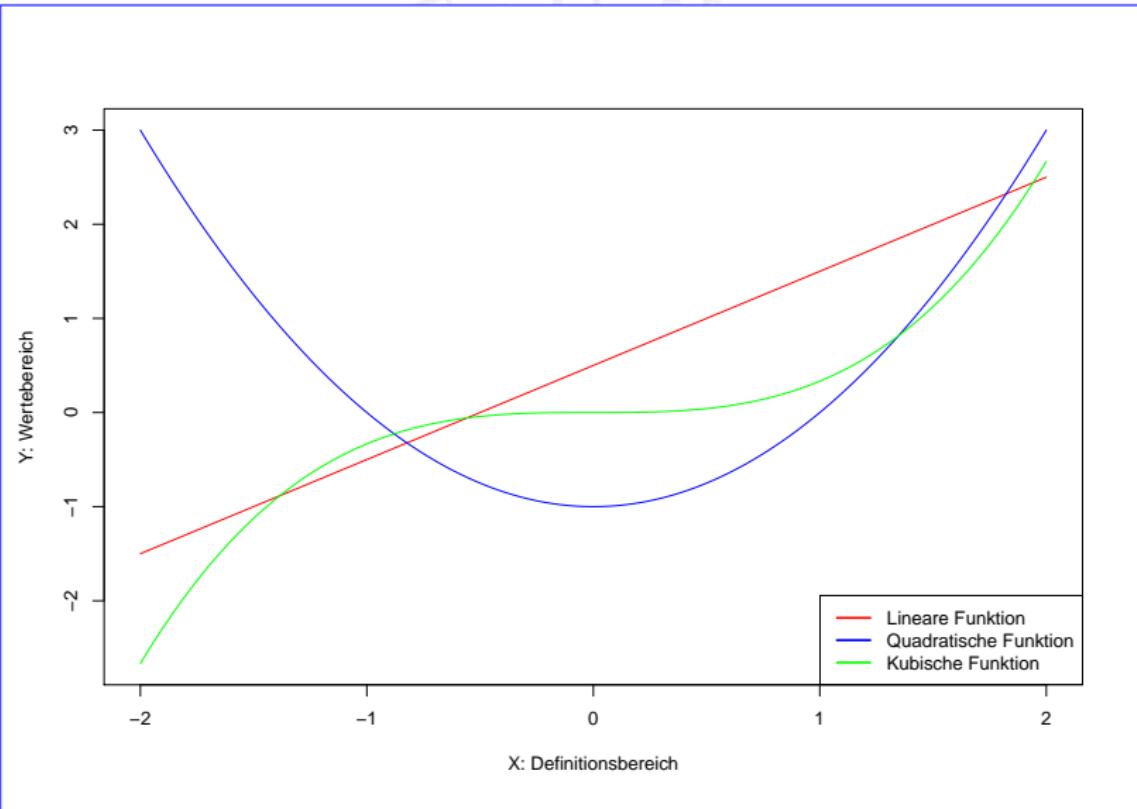
Funktionen

Definition:

Wenn x und y zwei variable Größen sind und wenn sich einem gegebenen x -Wert genau ein y -Wert zuordnen lässt, dann nennt man y eine Funktion von x und schreibt:

$$y = f(x)$$

Die veränderliche Größe x heißt *unabhängige Variable* oder *Argument* der Funktion y . Alle x -Werte, denen sich y -Werte zuordnen lassen, bilden den *Definitionsbereich* der Funktion y . Die veränderliche Größe y heißt *abhängige Variable*. Alle y -Werte bilden den *Wertebereich* der Funktion $f(x)$.

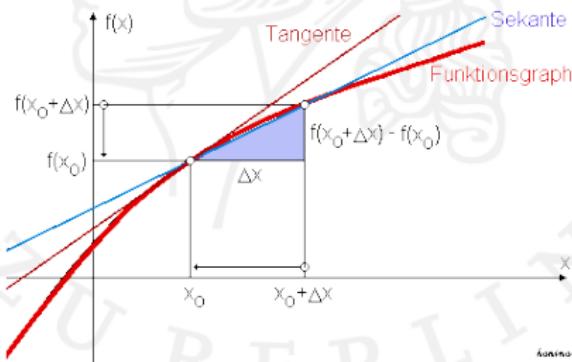


Ableitungen

Definition:

Die Ableitung einer Funktion $y = f(x)$ ist eine neue Funktion von x , die mit den Symbolen y' , $f'(x)$, Dy , $Df(x)$, $\frac{df(x)}{dx}$ oder $\frac{\partial f(x)}{\partial x}$ gekennzeichnet wird und die für jeden Wert x gleich dem Grenzwert des Quotienten aus dem Zuwachs der Funktion Δy und dem entsprechenden Zuwachs Δx für $\Delta x \rightarrow 0$ ist:

$$f'(x) = \lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}.$$



1. Potenzfunktionen

Funktion $f(x)$	Ableitung $f'(x)$
x^n	nx^{n-1}
x	1
1	0
x^2	$2x$
x^3	$3x^2$
$\frac{1}{x} = x^{-1}$	$-\frac{1}{x^2} = -x^{-2}$
$\sqrt{x} = x^{\frac{1}{2}}$	$\frac{1}{2\sqrt{x}} = \frac{1}{2}x^{-\frac{1}{2}}$

2. Logarithmus- und Exponentialfunktionen

Funktion $f(x)$	Ableitung $f'(x)$
$\ln x$	$\frac{1}{x}$
e^x	e^x
e^{kx}	ke^{kx}

3. Winkelfunktionen

Funktion $f(x)$	Ableitung $f'(x)$
$\sin x$	$\cos x$
$\cos x$	$-\sin x$
$\tan x$	$\frac{1}{\cos^2 x}$

Rechenregeln		
Funktion	Ableitung	Regel
$kf(x)$	$kf'(x)$	mit konstantem Faktor
$f(x) + g(x)$	$f'(x) + g'(x)$	Summenregel
$f(x) \cdot g(x)$	$f'(x) \cdot g(x) + f(x) \cdot g'(x)$	Produktregel
$\frac{f(x)}{g(x)}$	$\frac{f'(x) \cdot g(x) - f(x) \cdot g'(x)}{g(x)^2}$	Quotientenregel
$f(g(x))$	$f'(g(x))g'(x)$	Kettenregel

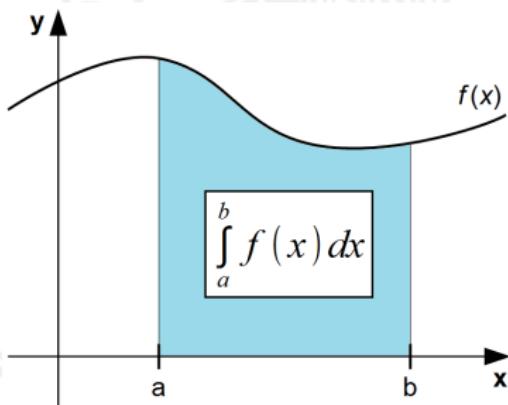
✳ Listing 1.10: `differentiate.ma`

```
1 f[x_] := 1/Sqrt[2*Pi]*Exp[-x^2/2]
2 D[f[x], x]
```

Integrale

Die Stammfunktion oder Integral einer gegebenen Funktion $y = f(x)$, die in einem zusammenhängenden Intervall (a, b) definiert ist, wird eine differenzierbare Funktion $F(x)$ genannt, die im selben Intervall definiert ist und deren Ableitung gleich $f(x)$ ist:

$$F'(x) = f(x)$$



$$\text{mit } \int_a^b f(x) dx = [F(x)]_a^b = F(b) - F(a)$$

Eigenschaften:

1. Vorzeichen: $\int_a^b f(x)dx = - \int_b^a f(x)dx$ und $\int_a^a f(x)dx = 0$

2. Additivität: $\int_a^b f(x)dx = \int_a^z f(x)dx + \int_z^b f(x)dx$

3. Linearität: $\int_a^b (f(x) + g(x))dx = \int_a^b f(x)dx + \int_a^b g(x)dx$

$$\int_a^b kf(x)dx = k \int_a^b f(x)dx$$

4. Monotonie: Gilt $f(x) \leq g(x)$ für alle $x \in [a, b]$, so auch

$$\int_a^b f(x)dx \leq \int_a^b g(x)dx$$

5. Partielle Integration:

$$\int_a^b f'(x)g(x)dx = [f(x) \cdot g(x)]_a^b - \int_a^b f(x)g'(x)dx$$

Stammfunktionen $F(x)$

$f(x)$	$F(x)$	$f(x)$	$F(x)$
k	$k \cdot x + C$	$\sin x$	$-\cos x + C$
x^n	$\frac{1}{n+1}x^{n+1} + C$	$\cos x$	$\sin x + C$
e^x	$e^x + C$	$\tan x$	$-\ln \cos(x) + C$
$\frac{1}{x}$	$\ln(x) + C$	$\sqrt[n]{x} = x^{\frac{1}{n}}$	$\frac{x^{\frac{1}{n}+1}}{\frac{1}{n}+1} + C$

Unbestimmtes Integral Stammfunktion plus eine Konstante C

$$\int f(x)dx = F(x) + C$$

(unpräzise Schreibweise)

Bestimmtes Integral mit Integralgrenzen a und b

$$\int_a^b f(x)dx = F(b) - F(a)$$

Beispiel 1.1

$$\int \frac{e^{-x}}{(1+e^{-x})^2} dx = -\frac{1}{1+e^x} + C$$

Logistische Verteilung: $C = 1$

$$\int_0^1 \frac{e^{-x}}{(1+e^{-x})^2} dx = -\frac{1}{1+e^x} \Big|_0^1 = F(1) - F(0) = -\frac{1}{1+e} + \frac{1}{2} \approx 0,2311$$

 Listing 1.11: `integrate.R`

```

1 f <- function(x) { exp(-x) / ((1+exp(-x))^2) }
2 integrate(f, 0, 1)

```

 Listing 1.12: `integrate.ma`

```

1 f[x_] := Exp[-x] / ((1+Exp[-x])^2)
2 Integrate[f[x], x]
3 Integrate[f[x], {x, 0, 1}]
4 N[1/2 - 1/(1+Exp[1])]

```

 Listing 1.13: `integrate.py`

```

1 import scipy.integrate
2 import math
3 def f(x):
4     return math.exp(-x)/(1+math.exp(-x))**2
5 print(scipy.integrate.quad(f, 0, 1))

```

Software

- Freie Software
 - ▶ R und RStudio Desktop
 - ▶ Python und verschiedene Python IDEs
 - Kostenfreie Software für Studenten der HU
 - ▶ SPSS
 - ▶ Mathematica
 - Im PC-Pool der Fakultät
 - ▶ Stata
 - Installation von R Paketen
-  Listing 1.14: `example_install.R`

```
1 # Allgemein
2 install.packages("plotrix")
3 # mmstat4 von Github
4 install.packages("devtools")
5 devtools::install_github("sigbertklinke/mmstat4")
```

- Nutzung von `mmstat4`

- ▶ die Link-Adresse kopieren
- ▶ in RStudio einfügen und z.B. `ghopen("...")` hinzufügen

 Listing 1.15: `use_mmstat4.R`

```
1 #install.packages("devtools")
2 #devtools::install_github("sigbertklinke/mmstat4")
3 run("stat/lottozahlen.R")
4 show("stat/lottozahlen.R")
```

Einführung

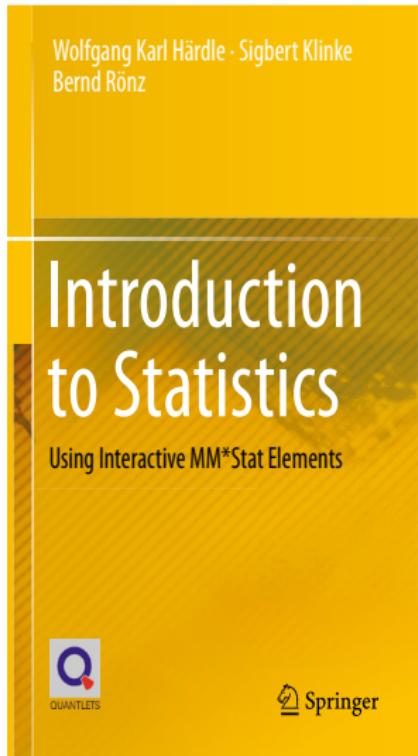
5. November 2022

Informationen • Warum Statistik? • Modellierung • Bevölkerungspyramide
• Studienfälle HU Berlin • Arbeitslose und offene Stellen •
Haushaltsgrößen • Volkszählung • Was ist Statistik? • Erkenntnisziele der
Statistik • Etappen statistischer Untersuchungen • Amtliche Statistik •
Nichtamtliche Statistik

Informationen

- Vorlesung: Fr 10:15-11:45
 - ▶ Online-Stream via Zoom: ([Meeting-ID: 699 8799 7856](#), Passwort: [431725](#))
- Übungen:
 - ▶ Aufgabenlisten werden nach der Vorlesung in Moodle veröffentlicht
 - ▶ Bearbeitung erfolgt selbstständig
 - ▶ Für Fragen stehen die Übungsleiter im Moodle-Forum [Diskussionsforum \(nur Statistikthemen\)](#) zu Verfügung
- Lehrmaterialien: Folien, Aufgabensammlung, Formelsammlung
 - ▶ In der HU-Box:
<https://box.hu-berlin.de/d/17f070caf3f04e26939f/>
 - ▶ Passwort: cF*AN#Cy
 - ▶ Ordner: Statistik_SS22+WS2223

- Aufgabensammlung
 - ▶ Aufgaben sind für Statistik I&II
 - ▶ zu allen Aufgaben gibt es Kurzlösungen
 - ▶ enthält Aufgaben für Übungen (und Tutorien)
 - ▶ zur Vorbereitung für die Klausur
 - ▶ **Videos mit Aufgabenlösungen**
- Online Tests
 - ▶ Moodle-Kurs "Tests für Statistik I+II"
 - ▶ für jedes Themengebiet
- MM*Stat: <http://www.mm-stat.org>
 - ▶ Skripte, Aufgaben und Klausuren anderer Universitäten
 - ▶ **Wiki (deutsch) mit den Vorlesungsinhalten**

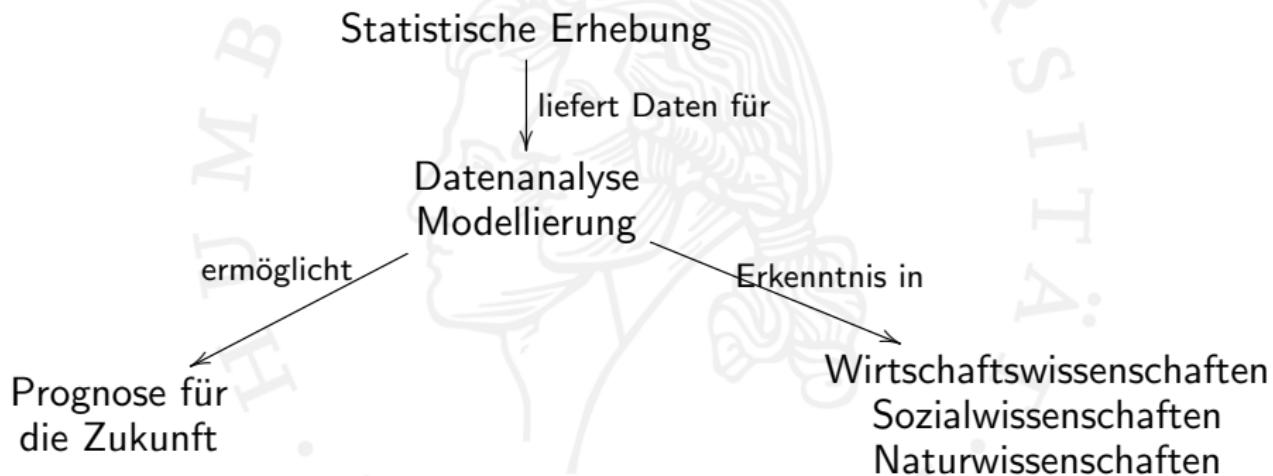


- In Englisch
- Deckt alle Themengebiete der Statistik I+II ab
- Themenreihenfolge entspricht nicht der Vorlesung
- Hardcover \approx 65 EUR
- PDF \approx 50 EUR
- erschienen im [Springer Verlag](#)

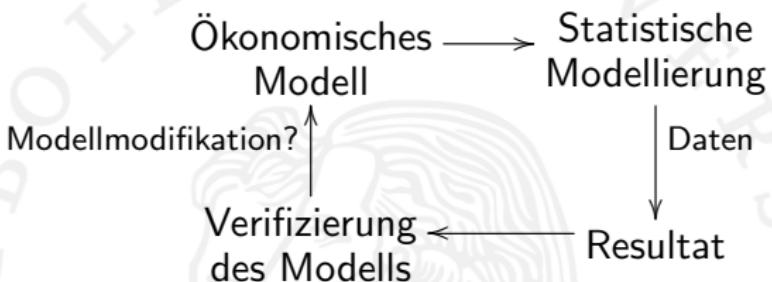
Statistik Software

- The R Project for Statistical Computing
<https://www.r-project.org/>
 - ▶ RStudio: Integrierte Entwicklungsumgebung für R
<https://www.rstudio.com/products/RStudio/>
 - ▶ Gelegentlich R Beispielprogramme in der Vorlesung
- Stata
 - ▶ Kommerzielle Software
 - ▶ Im PC-Pool der Fakultät installiert
 - ▶ Wird in anderen Lehrveranstaltungen benutzt
- SPSS
 - ▶ Kann kostenfrei installiert werden
 - ▶ Gelegentlich Beispiele in der Vorlesung

Warum Statistik?



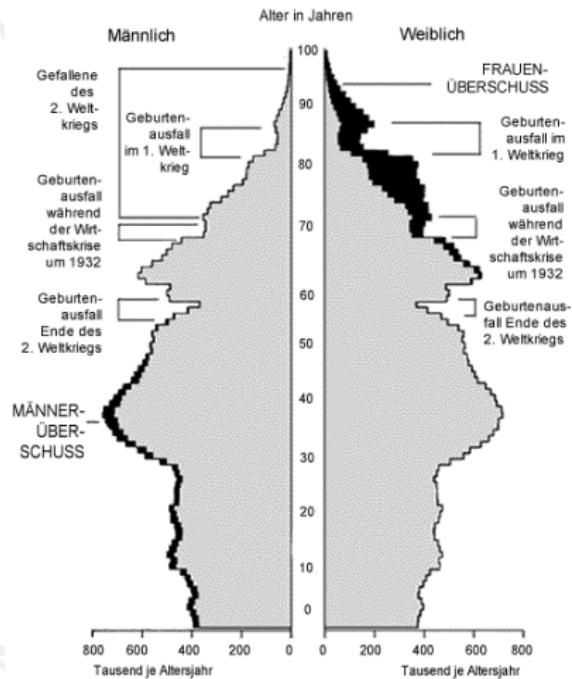
Modellierung



- Modelle sind Vereinfachungen des realen Sachverhaltes
- George Box (1976, 1978)
All models are wrong but some are useful
- Die Frage ist nicht: Ist das betrachtete Modell das wahre Modell?
- Die Frage ist: Ist das betrachtete Modell gut genug um den interessierenden Sachverhalt zu beschreiben?
- Ockhams Rasiermesser (William Ockham 1287-1347)

Bevölkerungspyramide

Altersaufbau der Bevölkerung Deutschlands am 31.12.2000



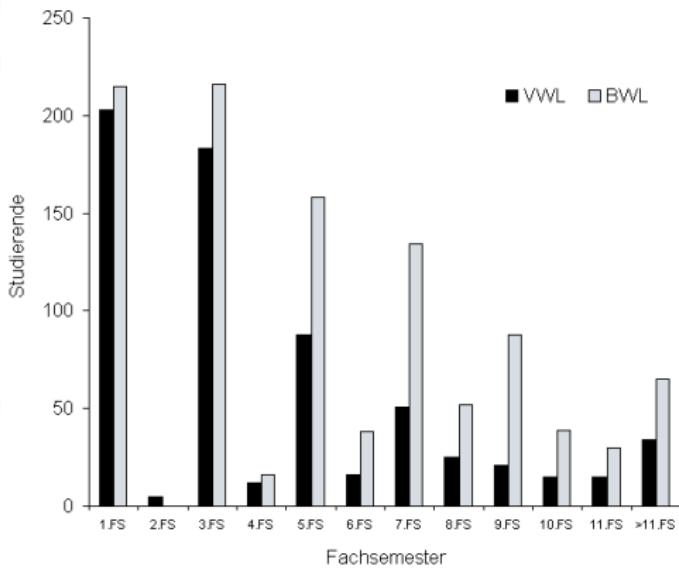
Quelle: Statistisches Bundesamt (Interaktives Beispiel)

R Listing 2.1: example_pyramid.R

```
1 # install.packages("plotrix")
2 library("plotrix")
3 #
4 x <- read.csv("12411-0006.csv")
5 pyramid.plot(x$M/1e5, x$W/1e5,
6               labels=c(1:85, ">85"), labelcex=0.65,
7               lxcol="blue", rxcol="red", unit="in 100000",
8               top.labels=c("Maennlich", "Alter", "Weiblich")
9 )
```

Studienfälle HU Berlin

Nach Fachsemestern, Wirtschaftswissenschaftliche Fakultät,
WS 2001/2002, Diplomstudiengänge



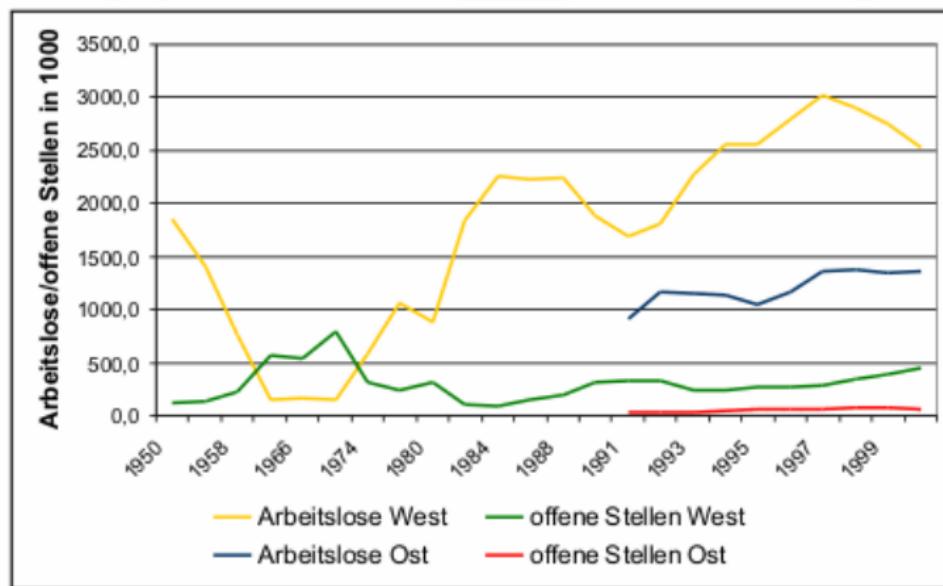
Quelle: Humboldt-Universität zu Berlin

R Listing 2.2: example_grouped_barchart.R

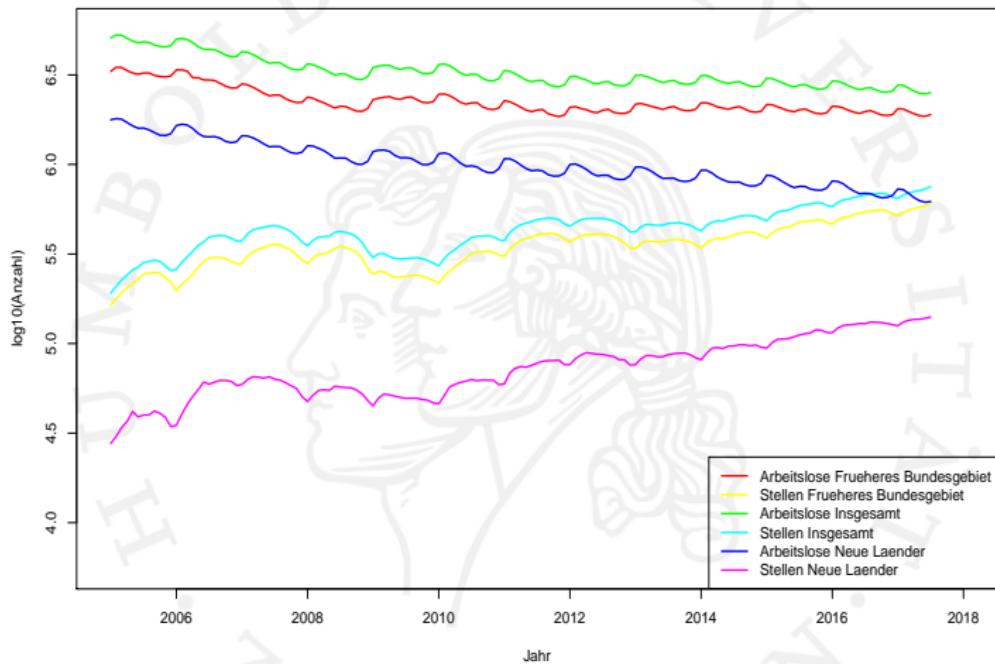
```
1 # Studierende nach FS (HU Berlin, WS 13/14)
2 # https://www.hu-berlin.de/de/studium/statistik
3 bsc_bwl <- c(171, 9, 174, 16, 115, 30, 90, 10, 46, 4, 8, 3, 21)
4 bsc_vwl <- c(135, 2, 111, 2, 75, 24, 60, 8, 22, 4, 7, 1, 18)
5 #
6 counts <- rbind(bsc_bwl, bsc_vwl)
7 colnames(counts) <- c(1:12, ">12")
8 barplot(counts, beside=T,
9         legend=c("B.Sc. BWL", "B.Sc. VWL"),
10        main="Studierende im WS 2013/14",
11        xlab="Fachsemester")
```

Arbeitslose und offene Stellen

Im früheren Bundesgebiet und in den neuen Ländern einschl. Berlin-Ost
1950–2000 (jeweils in 1000)



Quelle: Datenreport 2002, Bundeszentrale für politische Bildung,
Bonn 2002, S. 96ff.



Quelle: Statistisches Bundesamt, GENESIS online, Tabelle 13211-0002

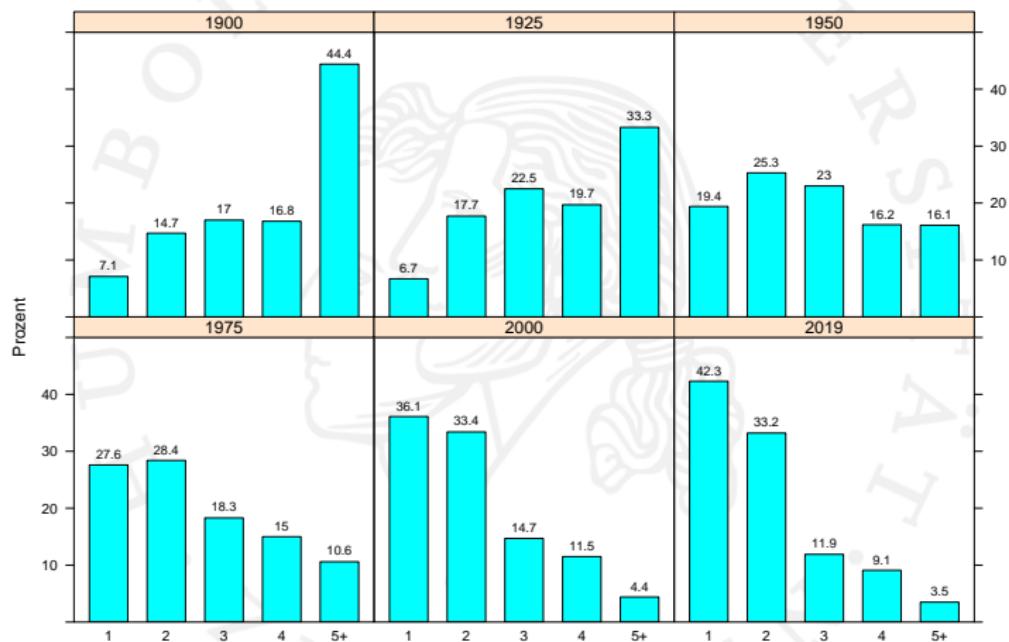


Listing 2.3: example_time_series.R

```
1 library("mmstat4")
2 #
3 data(data13211)
4 # Zeile 1-156 Deutschland
5 # Zeile 157-312 Alte Bundesländer
6 # Zeile 313-468 Neue Bundesländer
7 ger <- data13211[1:156,]
8 #
9 ts <- ts(ger$Arbeitslose/1e6, start=c(2005,1), frequency=12)
10 plot(ts, main="Arbeitslose in Deutschland (in Mio)", xlab="")
11 #
12 ts <- ts(ger$"Gemeldete Stellen"/1e6, start=c(2005,1),
13           frequency=12)
14 plot(ts, main="Gemeldete Arbeitsstellen (in Mio)", xlab="")
```

Haushaltsgrößen

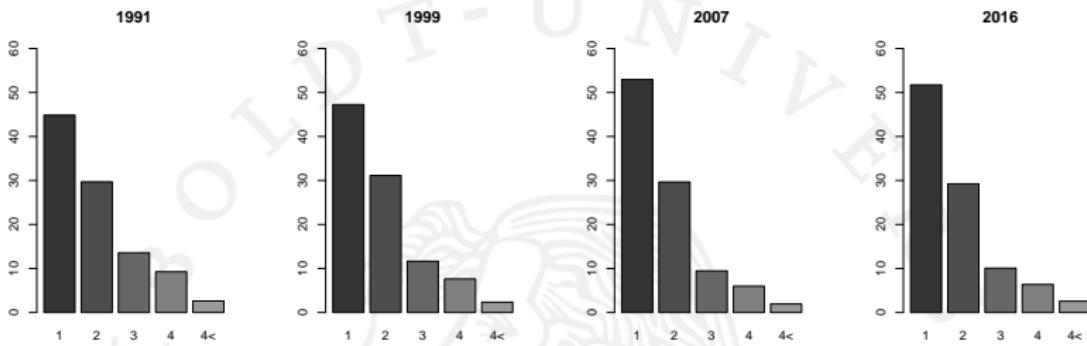
Anteil Haushaltsgrößen Deutschland



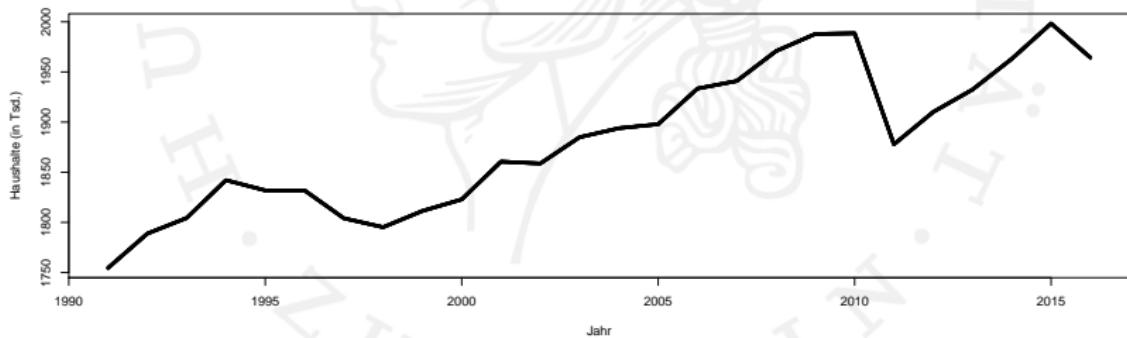
Quelle: Statistisches Bundesamt

R Listing 2.4: example_households.R

```
1 library("mmstat4")
2 data(hhD)
3 #
4 ylim <- c(0, max(hhD$Einpersonen, na.rm=T))
5 plot(hhD$Jahr, hhD$Einpersonen, xlab="", ylab="Anteil (in %)",
6       main="Anteil x Personen Haushalte (Deutschland)",
7       sub="Bis 1990 Frueheres Bundesgebiet",
8       pch=19, cex=0.75, ylim=ylim)
9 #
10 points(hhD$Jahr, hhD$Zweipersonen, pch=19, cex=0.75, col="red")
11 points(hhD$Jahr, hhD$Dreipersonen, pch=19, cex=0.75, col="blue")
12 points(hhD$Jahr, hhD$Vierpersonen, pch=19, cex=0.75, col="green")
13 points(hhD$Jahr, hhD$'Fuenf und mehr Personen',
14         pch=19, cex=0.75, col="orange")
15 #
16 legend("topleft", title="Haushalte mit", pch=19, cex=0.75,
17        col=c("black", "red", "blue", "green", "orange"),
18        legend=c("1 Person", "2 Personen", "3 Personen",
19                  "4 Personen", "5+ Personen"))
```



Privathaushalte in Berlin (oben: Anteile, unten: Gesamtsumme)



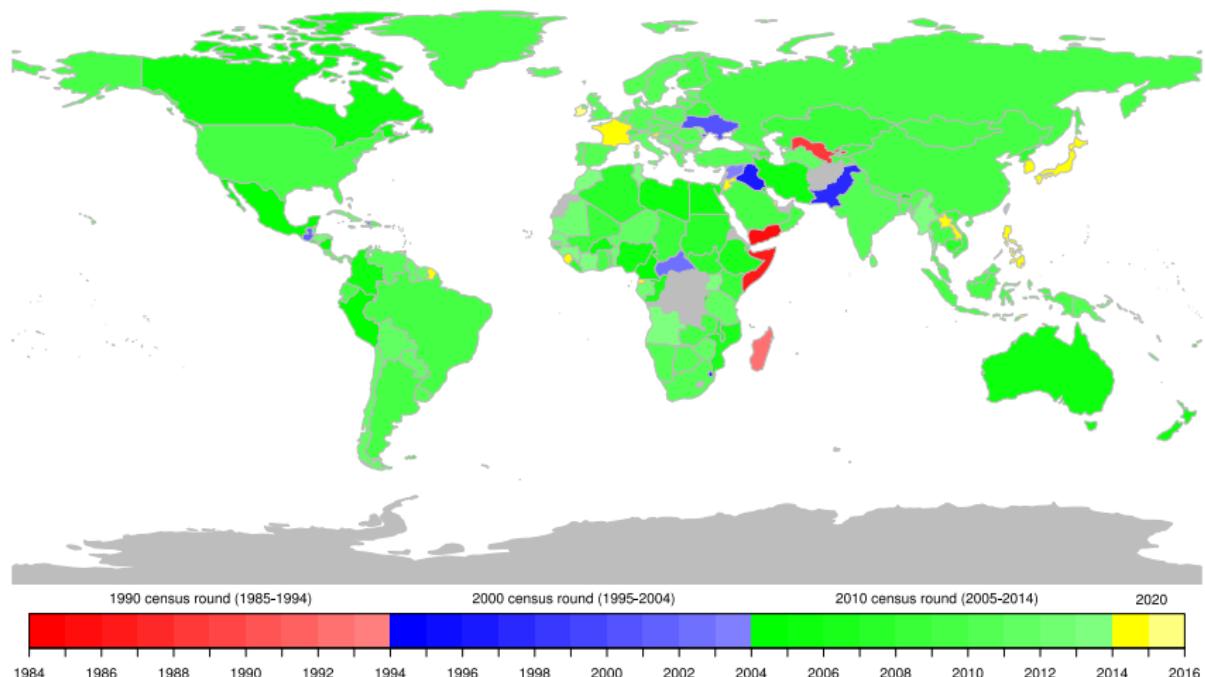
Quelle: Amt für Statistik Berlin-Brandenburg, Mikrozensus Berlin



Listing 2.5: example_households_berlin.R

```
1 library("mmstat4")
2 data(hhB)
3 #
4 jahr <- c(1991, 2014)
5 row <- hhB[,1]==jahr
6 hh <- as.matrix(100*hhB[row, 3:7]/hhB[row,2], nrow=5)
7 colnames(hh) <- c(1:4, "5+")
8 #
9 barplot(hh, main="Anteil x Personenhaushalte (Berlin)",
10         xlab="Haushaltsgroesse", ylab="Anteil (in %)",
11         beside=T, legend.text=jahr)
12 #
13 plot(hhB$Jahr, hhB$Privathaushalte, type="b",
14       main="Anzahl Privathaushalte (Berlin)",
15       xlab="", ylab="Privathaushalte (in Tsd.)")
```

Volkszählung



Quelle: Wikimedia Commons, [Lastcensus.svg](#)

Was ist Statistik?

Die Statistik ist die Wissenschaft

- der empirischen, objektivierten, am theoretischen Modell orientierten Information, die aus der Untersuchung von zum Teil zufallsbedingten Massenerscheinungen resultiert
- der Verfahren, nach denen Informationen gesammelt/erstellt, dargestellt, verarbeitet und analysiert werden

Erkenntnisziele der Statistik

1. Deskriptive Statistik (auch: beschreibende Statistik)

- umfasst für eine gegebene Zielstellung
 - ▶ Datenerhebung
 - ▶ Aufbereitung
 - ▶ Auswertung
 - ▶ statistische Verfahren
- quantitative Beschreibung empirischer Massenerscheinungen
- Ergebnisse und Aussagen beziehen sich grundsätzlich nur auf die untersuchte Datenmenge

2. Induktive Statistik (auch: statistische Inferenz)

- umfasst Verfahren und allgemeine Bedingungen (Sätze), die die Übertragung von Ergebnissen aus Untersuchungen einer Teilgesamtheit (Stichprobe) auf die Gesamtheit (Grundgesamtheit) erlauben
- stellt wahrscheinlichkeitstheoretisch fundierte Methoden bereit, mit denen der Rückschluss von Aussagen der Stichprobe auf die Gesamtheit unter Vorgabe einer gewissen Präzision vorgenommen werden kann.

Etappen statistischer Untersuchungen

1. Planung

2. Erhebung

- ▶ Schriftliche oder mündliche Befragung
- ▶ Beobachtung
- ▶ Experiment
- ▶ Automatische Erfassung

3. Aufbereitung

4. Analyse

5. Interpretation

Untersuchungen

- primärstatistische
- sekundärstatistische
- tertiärstatistische

Beispiel 2.1

Problem Beschaffung neuer Server (Computer) im Rahmen eines Forschungsprojektantrages:

SUN Solaris Server (Unix)		Linux Server	
Vorteile	Nachteil	Vorteil	Nachteile
wenig Ausfälle	teuer	billig	mehr Ausfälle
wenig Wartung			mehr Wartung

Planung entfällt aufgrund der Problemgröße

Lösung Prüfung der Auslastungsdaten der bestehenden Server → welche Art von Servern wird häufiger genutzt

Datenerhebung Log-files der Server von Januar bis einschließlich März 2002 liegen vor mit Auslastungsdaten, Plattenplatzverbrauch, etc.
(5000 - 7000 Beobachtungen pro Rechner)

Aufbereitung Übertragung eines Teils der Daten (Auslastung) in Excel

Analyse Zusammenfassung der Auslastungsdaten in Kennzahlen
(Minimum, Maximum, Mittelwert)

	Apus(S)	Columbus(L)	Hobbes(S)	Mars(S)	Pluto(L)
Min	0,00	0,00	0,01	0,01	0,00
Max	2,02	4,72	1,17	1,45	2,56
MW	0,27	0,28	0,13	0,34	0,67

Interpretation Rechner mit der höchsten durchschnittlichen Auslastung ist ein Linux Server

Entscheidung Linux Server

Amtliche Statistik

- Statistisches Bundesamt (www.destatis.de)
 - ▶ Genesis online (www-genesis.destatis.de)
- Statistische Landesämter (de.wikipedia.org/wiki/Statistisches_Landesamt)
- Ressortstatistik
 - ▶ Bundesagentur für Arbeit (www.arbeitsagentur.de)
 - ▶ Bundesinstitut für Bevölkerungsforschung (www.bib-demographie.de)
- Internationale Organisationen
 - ▶ United Nations (unstats.un.org) → World Statistics Day (20.10.2010, 20.10.2015, 20.10.2020)
 - ▶ EU (ec.europa.eu/eurostat)
 - ▶ OECD (www.oecd-ilibrary.org/statistics) → PISA (www.pisa.oecd.org)

Nichtamtliche Statistik

- Arbeitgeber- und Arbeitnehmerorganisationen
(www.berlin-brandenburg.dgb.de)
- Industrie- und Handelskammern (www.dihk.de)
- GESIS - Leibniz-Institut für Sozialwissenschaften (www.gesis.org)
 - ▶ Allgemeine Bevölkerungsumfrage der Sozialwissenschaften
(www.gesis.org/allbus)
 - ▶ International Social Survey Program (www.issp.org)
- Markt- und Meinungsforschungsinstitute
 - ▶ Forsa (www.forsa.de)
 - ▶ Emnid (www.tns-emnid.com)
 - ▶ Allensbach (www.ifd-allensbach.de)
- Wirtschaftswissenschaftliche Forschungsinstitute
 - ▶ Deutsches Institut für Wirtschaftsforschung (www.diw.de)
 - ▶ Institut für Wirtschaftsforschung (www.ifo.de)
 - ▶ Institut für Wirtschaftsforschung Halle (www.iwh-halle.de)
- größere Unternehmen

A faint watermark of the HU Berlin logo is visible in the background, featuring a circular emblem with a figure and the text "HUMBOLDT-UNIVERSITÄT ZU BERLIN".

Teil 2

Deskriptive Statistik

Grundbegriffe

5. November 2022

- Definitionen • Ausreißer • Skalierung von Variablen • Nominalskala •
- Dichotome oder binäre Variable • Häufbare Variable • Ordinalskala •
- Metrische Skala (Kardinalskala) • Intervallskala • Verhältnisskala •
- Absolutskala • Skalierung von Variablen • Diskrete und stetige Variablen •
- Zusammenfassung • Unklassierte & gepoolte Datensätze • Klassierung von Variablen

Definitionen

Statistische Einheit (Merkmalsträger)

- Gegenstand oder Vorgang
- eindeutig definiert auf Grund von Identifikationskriterien
- Träger der Informationen für die statistische Untersuchung: natürliche Einheiten (Personen, Tiere, Pflanzen), sozio-ökonomische Einheiten (Familien, Haushalte, Unternehmen).

Variable (Merkmal)

- Eigenschaft einer statistischen Einheit

Variablenausprägungen

- Werte, die die Variable bei einer statistischen Einheit annehmen kann

Variable Variablenausprägungen

X $x_1, x_2, x_3 \dots, x_n$

Y $y_1, y_2, y_3 \dots, y_n$

Grundgesamtheit

- Menge der statistischen Einheiten mit übereinstimmenden Identifikationskriterien

Stichprobe

- eine endliche Teilmenge der Elemente der Grundgesamtheit
- ausgewählt und erfasst für die statistische Untersuchung

Beispiel 3.1

- Grundgesamtheit: Einwohner der Stadt X
- Merkmalsträger: ein Einwohner
- Stichprobe: Gruppe der erfassten Einwohner
- Merkmal: Geschlecht
- Merkmalsausprägung: m, w



Listing 3.1: example_sample.R

```
1 library("MASS")      # Boston Housing Daten
2 # https://stat.ethz.ch/R-manual/R-devel/
3 #           library/MASS/html/Boston.html
4 N <- nrow(Boston)    # Umfang Grundgesamtheit
5 n <- 51                # Umfang Stichprobe
6 # Ziehen ohne Zurücklegen
7 sample(N, size=n)
8 # Ziehen mit Zurücklegen
9 sample(N, size=n, replace=TRUE)
```

Beispiel 3.2

- Frankfurter Allgemeine (26.03.14): Integration: Schlechtere Bewerbungschancen mit ausländischen Namen
- Süddeutsche Zeitung (26.03.14): Türkischer Name schmälert Chance auf Ausbildungsplatz
- Spiegel (26.03.14): Vornamen-Diskriminierung: "Keiner will einen Ali im Team haben"
- Der Tagesspiegel (27.03.14): Türkischer Name ist bei Bewerbungen ein Nachteil
- Sachverständigenrat deutscher Stiftungen für Integration und Migration (März 2014): Diskriminierung am Ausbildungsmarkt: Ausmaß, Ursachen und Handlungsperspektiven
- Statistische Einheit???

Ausreißer

- ein Messwert,
 - ▶ der weit weg von allen anderen Messwerten liegt oder
 - ▶ der unerwartet auftritt (oder fehlt)

Beispiel 3.3 (ALLBUS 2010)

Extremwerte

		Fallnummer	Wert
BEFR.: NETTOEINKOMMEN, OFFENE ABFRAGE	Größte Werte	1	136
		2	780
		3	569
		4	2249
		5	2692
	Kleinste Werte	1	1120
		2	1964
		3	1292
		4	1061
		5	1560



Listing 3.2: example_order.R

```
1 data(Boston, package="MASS")
2 # show first and last six observations
3 head(Boston)
4 tail(Boston)
5 # order by medv
6 index <- order(Boston$medv)
7 head(Boston[index,], n=10)
8 tail(Boston[index,], n=10)
```

Skalierung von Variablen

Skalierung

Die relationstreue Abbildung einer Zeichenmenge (Skala) auf eine Menge von statistisch erhobenen Einheiten bezüglich einer Variablen

Skalen

- Nominalskala
- Ordinalskala
- metrische Skala (Kardinalskala)
 - ▶ Intervallskala
 - ▶ Verhältnisskala
 - ▶ Absolutskala

Nominalskala

Eine Nominalskala liegt vor, wenn Variablenausprägungen durch zugeordnete Zahlen

- lediglich eine Verschiedenartigkeit zum Ausdruck bringen

⇒ die Variable heißt nominalskaliert

- nominalskalierte Variablen können

- ▶ binär (dichotom)
 - ▶ häufig

sein

- Zulässige Relationen: "gleich" oder "ungleich"



Die zugeordneten Zahlen (Nominalzahlen, Schlüsselzahlen) haben reine Bezeichnungsfunktion, es wird **keine** Reihenfolge gebildet!!!

Beispiel 3.4

Der Familienstand von verschiedenen Personen kann nur gleich oder verschieden, nicht aber objektiv besser oder schlechter bzw. größer oder kleiner sein.

Dichotome oder binäre Variable

Eine binäre Variable weist nur zwei sich gegenseitig ausschließende (disjunkte) Ausprägungen auf

Beispiel 3.5

- Geschlecht: männlich, weiblich
- Eine bestimmte Behandlung führt zum Erfolg oder nicht
- Ein Haushalt besitzt ein privaten PKW oder nicht
- Eine Bank stuft einen Kunden als kreditwürdig oder nicht kreditwürdig ein

Häufbare Variable

Häufbare Variable bedeutet, dass an derselben statistischen Einheit mehrere Ausprägungen beobachtet werden können

Beispiel 3.6

häufbare Variable:

- Vorname
- erlernter Beruf
- abonnierte Zeitungen

nicht häufbare Variable:

- Geschlecht
- Familienstand
- Postleitzahl
- Hauptwohnsitz

Ordinalskala

Eine Ordinalskala liegt vor, wenn Variablenausprägungen durch zugeordnete Zahlen

- eine Verschiedenartigkeit und
- eine natürliche Rangfolge

zum Ausdruck bringen.

- die Abstände sind nicht quantifizierbar
- Zulässige Relationen: "größer als" und "kleiner als"

Beispiel 3.7

- militärischer Dienstgrad
- Zensuren
- Erdbebenstärken
- Güteklassen für Produkte
- Aggressivität
- Intelligenz
- sozialer Status

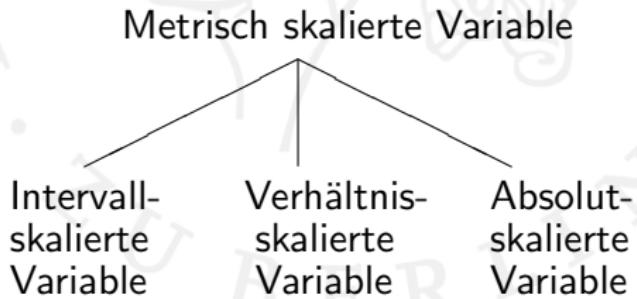
Metrische Skala (Kardinalskala)

Eine metrische Skala (Kardinalskala) liegt vor, wenn Variablenausprägungen durch zugeordnete Zahlen

- Verschiedenartigkeit und Rangfolge
- mess- und quantifizierbare Unterschiede

zum Ausdruck bringen.

- die Variable heißt **metrisch skaliert (kardinalskaliert, quantitativ)**
- die Variablenwerte sind im allgemeinen Ergebnis eines Zähl- oder Messvorgangs



Intervallskala

Eine Intervallskala liegt vor, wenn

- die Abstände (Differenzen) zwischen den Variablenwerten messbar und plausibel interpretierbar sind
- Quotienten dürfen nicht gebildet werden.

Sie besitzt

- keinen natürlichen Nullpunkt
 - keine natürliche Maßeinheit
- ⇒ Nullpunkt und Maßeinheit müssen Festlegungen sein.

Beispiel 3.8

- Temperatur in °C
- Kalenderzeitrechnung
- Breiten- und Längengrade der Erde

Verhältnisskala

Eine Verhältnisskala liegt vor, wenn

- Differenzen zwischen Variablenwerten
- Quotienten von Variablenwerten

berechenbar und plausibel interpretierbar sind
Sie besitzt

- einen natürlichen Nullpunkt
- keine natürliche Maßeinheit

Beispiel 3.9

- Wertvolumen eines Warenkorbes
- Längenmaße, Gewichtsmaße
- Alter, Einkommen

Absolutskala

Eine Absolutskala besitzt

- einen natürlichen Nullpunkt
- eine natürliche, maßstabsunabhängige Einheit

Beispiel 3.10

- Stückzahl
- Anzahl immatrikulierter Studenten an einer Universität

Skalierung von Variablen

R Listing 3.3: example_measurement.R

```
1  levels <- c(1, 1.3, 1.7, 2, 2.3, 2.7, 3, 3.3, 3.7, 4, 5)
2  noten  <- ordered(c(1.3, 3, 2.7, 2, 5), levels=levels)
3  noten
4  # Metrisch?
5  is.numeric(noten)
6  # Ordinal?
7  is.ordered(noten)
8  # Nominal?
9  is.factor(noten)
10 # Falsches Skalenniveau
11 mean(noten)
12 # Aendere Skalenniveau
13 noten_num <- as.numeric(levels(noten)[noten])
14 noten_num
15 # Kein Fehler, da das Skalenniveau nun Summation zulaesst.
16 mean(noten_num)
```

Diskrete und stetige Variablen

Metrische Variablen werden noch in stetig und diskret unterteilt.

Diskrete Variable

Eine Variable, die nur endlich oder abzählbar unendlich viele Werte annehmen kann

⇒ zwischen zwei benachbarten Merkmalsausprägungen gibt es keine weitere Merkmalsausprägung

Beispiel 3.11

- monatlicher Produktionsausstoß von PKW
- Anzahl der täglichen Anrufe bei einem Service-Point

Stetige Variable

Eine Variable, die in jedem beliebig kleinen Intervall überabzählbar unendlich viele Werte annehmen kann

⇒ zwischen zwei Merkmalsausprägungen gibt es immer eine weitere Merkmalsausprägung

Beispiel 3.12

- verkauft Menge von bleifreiem Benzin an einer Tankstelle pro Tag

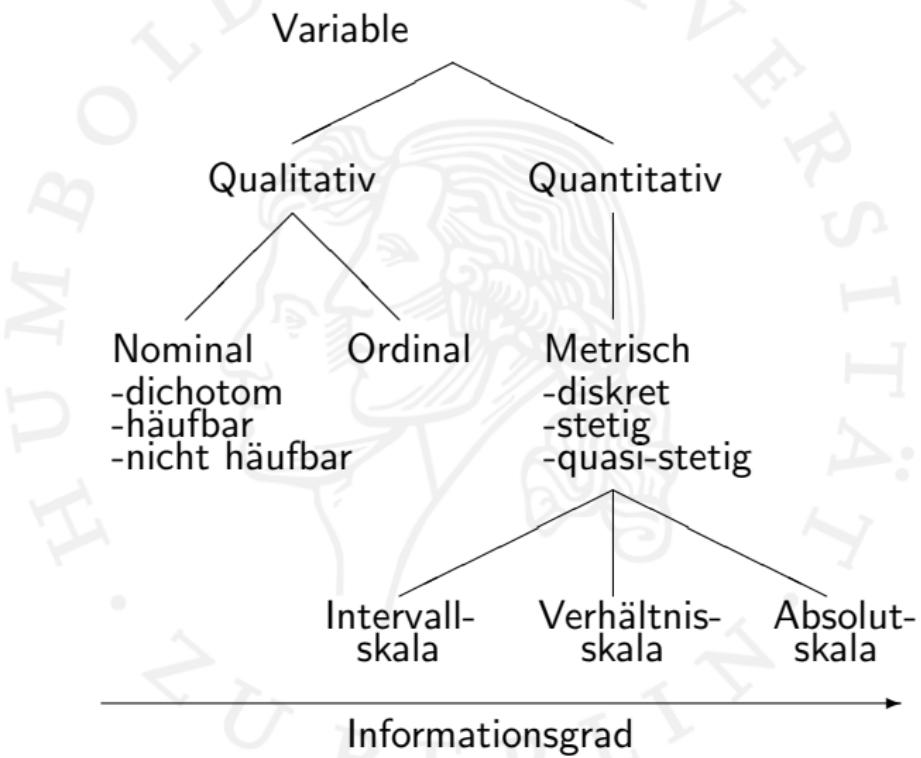
⇒ In der Praxis oft: **Quasi-stetige Merkmale**

Beispiel 3.13

- Preise
- Temperatur

Zusammenfassung

- Jeder Variablen wird genau ein Skalenniveau zugeordnet.
 - Das Skalenniveau hängt von den möglichen Ausprägungen ab.
 - Es gibt 3 wichtige Skalenniveaus mit steigendem Informationsgehalt der Variablen (nominal, ordinal, metrisch).
 - Metrische Variablen werden noch in stetig und diskret unterteilt.
 - Metrische diskrete Variablen mit vielen Ausprägungen werden auch als quasi-stetig bezeichnet.
-
- 
- Oft werden Voraussetzungen an das Skalenniveau einer Variablen für die Nutzung eines Koeffizienten oder einer Grafik gestellt
 - Die Voraussetzungen dienen dazu Fehlinterpretationen zu vermeiden



Unklassierte & gepoolte Datensätze

- Unklassierte Daten: Variablen liegen oft als Einzelwerte (unklassiert) vor
- Klassierte Daten: Variablen können auch in Klassenwerte vorliegen
- Gepoolte Daten: Ein Datensatz kann auch in mehreren Teildatensätzen vorliegen

Beispiel 3.14 (ALLBUS Daten 2014)

V417 BEFR.: NETTOEINKOMMEN, OFFENE ABFRAGE
 V418 BEFR.: NETTOEINKOMMEN, LISTENABFRAGE
 V419 BFR.:NETTOEINKOMMEN<OFFENE+LISTENANGABE>
 V420 NETTOEINKOMMEN<OFFENE+LISTENANGABE>,KAT.

V417	V418	V419	V420
1800	ANGABE SCHON DA	1800	1750 - 1999 EURO
1100	ANGABE SCHON DA	1100	1000 - 1124 EURO
2500	ANGABE SCHON DA	2500	2500 - 2749 EURO
VERWEIGERT	875 - 999 EURO	938	875 - 999 EURO
1000	ANGABE SCHON DA	1000	1000 - 1124 EURO
VERWEIGERT	5000 - 7499 EURO	6250	5000 - 7499 EURO
1400	ANGABE SCHON DA	1400	1375 - 1499 EURO

Klassierung von Variablen

Beispiel 3.15 (Einkommensverteilung)

statistische Einheit: Steuerpflichtiger

statistische Variable: steuerpflichtige Einkommen

Gesamtbetrag Einkünfte von ... bis unter ... Euro	Steuerpflichtige (1000)	\sum Einkommen (Mrd. Euro)
1 – 4 000	1445.2	2611.3
4 000 – 8 000	1455.5	8889.2
8 000 – 12 000	1240.5	12310.9
12 000 – 16 000	1110.7	15492.7
16 000 – 25 000	2762.9	57218.5
25 000 – 30 000	1915.1	52755.4
30 000 – 50 000	6923.7	270182.7
50 000 – 75 000	3876.9	234493.1
75 000 – 100 000	1239.7	105452.9
100 000 – 250 000	791.6	108065.7
250 000 – 500 000	93.7	31433.8
500 000 – 1 Mill.	26.6	17893.3
1 Mill. – 2 Mill.	8.6	11769.9
2 Mill. – 5 Mill.	3.7	10950.8
5 Mill. und mehr	1.4	16791.6

Klassierung (Gruppierung)

- die Zerlegung des Wertebereiches einer metrisch skalierten Variablen in mehrere Teilintervalle (Klassen oder Gruppen)
⇒ bessere Übersichtlichkeit bei großer Datenmenge

Klassenintervalle

- nicht überlappende (disjunkte) und aneinandergrenzende Intervalle von Variablenwerten
- n Anzahl der Beobachtungen
- k Anzahl der Klassen

Klassengrenze

Der Wert einer metrisch skalierten Variablen, der eine Klasse nach unten bzw. oben begrenzt

- untere Klassengrenze $x_j^u \quad j = 1, \dots, k$
- obere Klassengrenze $x_j^o \quad j = 1, \dots, k$

Eigenschaften

- $x_j^o = x_{j+1}^u, \quad j = 1, \dots, k - 1$
- $x_j^u < x \leq x_j^o \quad (\text{oder } x_j^u \leq x < x_j^o), \quad j = 1, \dots, k$

Klassenmitte

$$\bullet \quad x_j^m = \frac{1}{2}(x_j^u + x_j^o), \quad j = 1, \dots, k$$

Beispiel 3.16 (Weganteil nach Entfernungsklassen)

Aus dem [Mobilitätsreport 2013](#) der Berliner Senats:

statistische Einheit: Weg

statistische Variable: Weglänge

Von	Bis unter	Anteil	Klassenmitte
0	1	31%	0,5
1	3	22%	2,0
3	5	12%	4,0
5	10	17%	7,5
10 km und mehr		17%	??

Mit Hilfe weiterer Daten aus dem Mobilitätsreport kann für die letzte Klasse ein sinnvoller Wert bestimmt werden ($x_5^m \approx 21,5$ km).

R Listing 3.4: example_grouped.R

```
1 data(Boston, package="MASS")
2 # Häufigkeitstabelle einer stetigen Variablen
3 table(Boston$crim)
4 # Klassierung
5 cccrim <- cut(Boston$crim, c(0,0.1,0.2,0.3,0.4,0.5,1,2,10,100))
6 # Häufigkeitstabelle klassierter Variablen
7 table(ccrim)
```

Univariate Verteilungen

5. November 2022

- Notation
- Häufigkeit statistischer Variablen
- Grafische Darstellung der Häufigkeit
- Empirische Verteilungsfunktion
- Verteilung klassierter Variablen
- Grafische Darstellung klass. Daten
- Verteilungsfunktion klass. Variablen
- Interpolation von $F(x)$

Notation

- Variable: X
- Gesamtzahl der Beobachtungen: n
- Beobachtungswerte: x_i ($i = 1, \dots, n$)
- sich unterscheidende mögliche Variablenausprägungen (-werte): x_j ($j = 1, \dots, k$)

Beispiel 4.1 (10x Werfen einer „idealen“ Münze)

Variable: „sichtbare Seite der Münze“

Gesamtzahl der Beobachtungen: 10

sich unterscheidende mögliche Variablenausprägungen (-werte): „Kopf (K)“, „Zahl (Z)“

Beobachtungswerte: $K, Z, K, Z, Z, K, Z, K, K, Z$

Statistische Häufigkeit

Die Häufigkeit ist die Anzahl der Beobachtungen in einer Stichprobe mit der gleichen Ausprägung oder Klasse in einer Variablen.

Häufigkeit statistischer Variablen

Absolute Häufigkeit

- Anzahl der statistischen Einheiten mit einer bestimmten Variablenausprägung x_j ($j = 1, \dots, k$)

$$h(X = x_j) = h(x_j) = h_j = \sum_{i=1}^n I(x_i = x_j)$$

Wie viele Beobachtungswerte x_i sind gleich der j ten Merkmalsausprägung?

- Wobei I die Indikatorfunktion (charakteristische Funktion) ist.

$$I(x_i = x_j) = \begin{cases} 1 & \text{für } x_i = x_j \\ 0 & \text{für } x_i \neq x_j \end{cases}$$

- Eigenschaften: $0 \leq h(x_j) \leq n$, $\sum_{j=1}^k h(x_j) = n$

Relative Häufigkeit

- Anteil statistischen Einheiten mit einer bestimmten Variablenausprägung x_j ($j = 1, \dots, k$)

$$f(x_j) = \frac{h(x_j)}{n}$$

- Eigenschaften: $0 \leq f(x_j) \leq 1$, $\sum_{j=1}^k f(x_j) = 1$

Empirische Häufigkeitsverteilung

Die Häufigkeitsverteilung einer Variablen ergibt sich durch

- die geordneten Variablenausprägungen
- die Angabe der dazugehörigen absoluten bzw. relativen Häufigkeiten

Die Häufigkeitsverteilung gibt an, wie sich die statistischen Einheiten auf die beobachteten Variablenausprägungen verteilen

Allgemeine Häufigkeitstabelle

Variablenausprägung	abs. Häufigkeit	rel. Häufigkeit
x_1	$h(x_1)$	$f(x_1)$
\vdots	\vdots	\vdots
x_j	$h(x_j)$	$f(x_j)$
\vdots	\vdots	\vdots
x_k	$h(x_k)$	$f(x_k)$
Summe	n	1



Listing 4.1: example_freq.R

```
1 data(Boston, package="MASS")
2 # Häufigkeitstabellen (absolut & relativ)
3 table(Boston$rad)
4 prop.table(table(Boston$rad))
5 # Häufigkeitstabellen fuer stetige Daten
6 table(Boston$crim)
7 # Abhilfe kann Klassierung leisten:
8 table(cut(Boston$crim, c(0,0.1,0.2,0.3,0.4,0.5,1,2,10,100)))
```

Grafische Darstellung der Häufigkeit

Darstellung i.d.R. als

- Säulendiagramm
- Stabdiagramm

Abszisse: Variablenausprägungen x_j

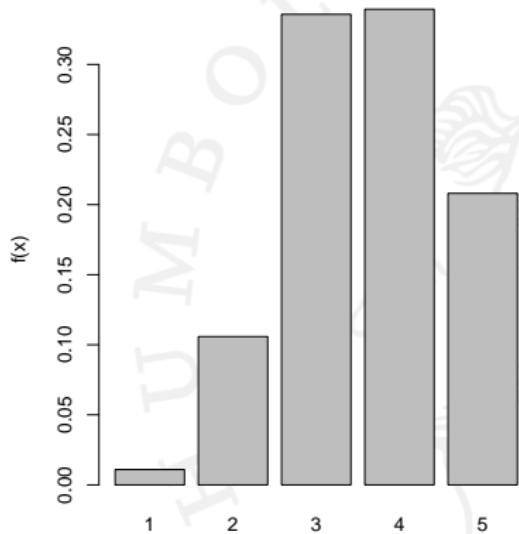
Ordinate: absolute oder relative Häufigkeit

Beispiel 4.2 (Note der Statistik I Klausur vom Juli 2002)

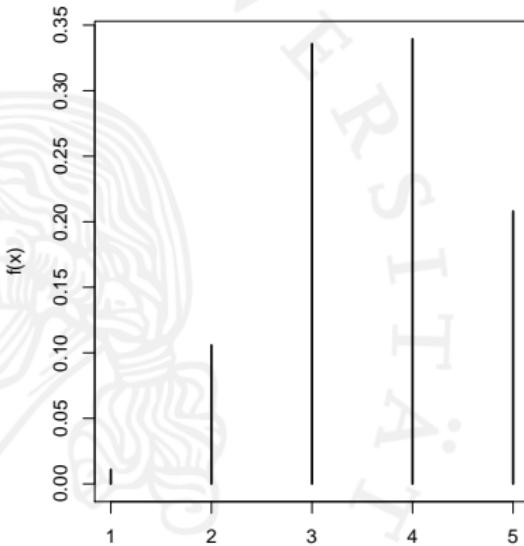
- 1 – sehr gut
- 2 – gut
- 3 – mangelhaft
- 4 – ausreichend
- 5 – nicht ausreichend

Note	abs. Häufigkeit	rel. Häufigkeit
1	3	0,01
2	29	0,11
3	92	0,33
4	93	0,34
5	57	0,21

Säulendiagramm

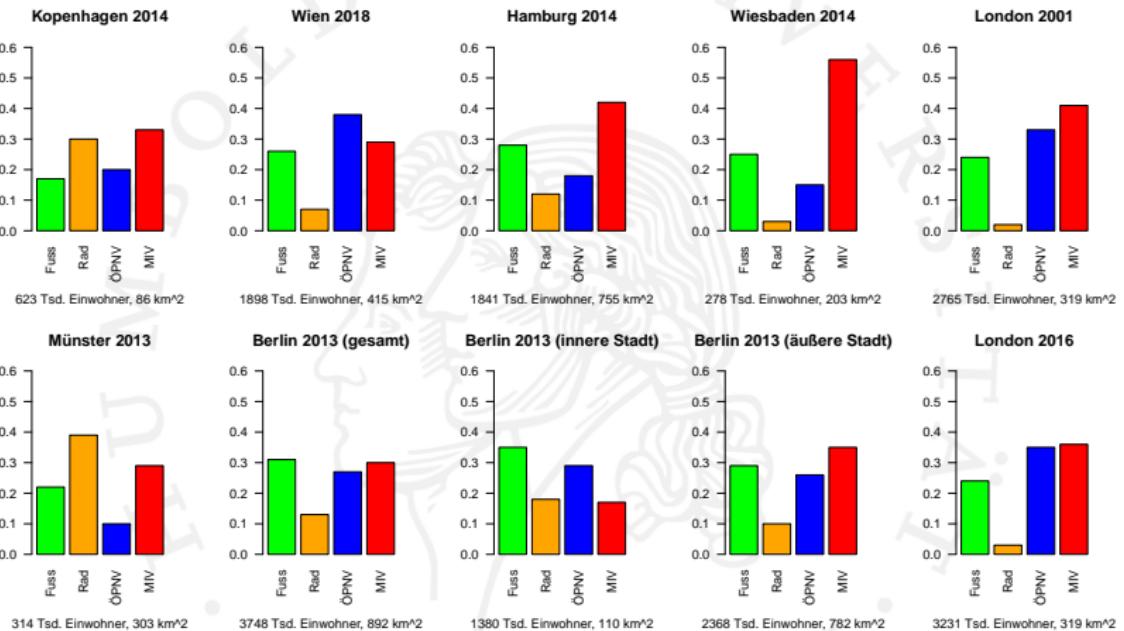


Stabdiagramm



höhenproportionale Darstellung

Modal splits für ausgewählte Städte



Modal split: Anteil an den Wegen, Flächen- und Einwohnerdaten aus 2017/2018

höhenproportionale Darstellung

 Listing 4.2: `example_barchart_needle.R`

```
1 data(Boston, package="MASS")
2 barplot(table(Boston$rad))
3 # Saeulendiagramm
4 plot(table(Boston$rad), type="h")
5 # Stabdiagramm
```

Empirische Verteilungsfunktion

Empirische Verteilungsfunktion $F(x)$ und Summenhäufigkeit $H(x)$

Setzt ordinal- oder metrischskalierte Variablen voraus

$$F(x) = \begin{cases} 0 & \text{für } x < x_1 \\ \sum_{i=1}^j f(x_i) & \text{für } x_j \leq x < x_{j+1} \\ 1 & \text{für } x_k \leq x \end{cases} \quad \text{bzw. } H(x) = \begin{cases} 0 \\ \sum_{i=1}^j h(x_i) \\ n \end{cases}$$

Berechnungen mittels der Verteilungsfunktion

$$f(x_j) = F(x_j) - F(x_{j-1}) \quad \text{für } j = 1, \dots, k \text{ mit } F(x_0) = 0$$

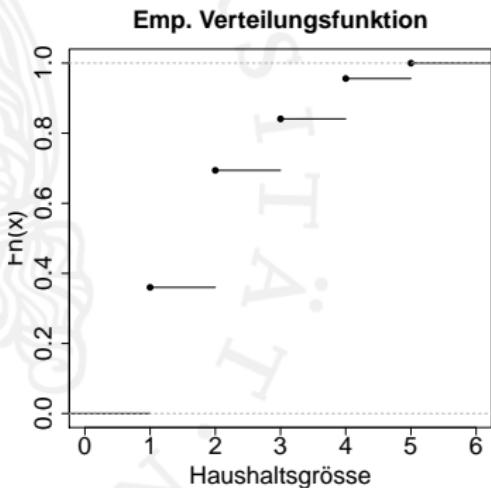
$$\begin{aligned} f(x_i < X < x_l) &= f(x_i < X \leq x_{l-1}) \\ &= F(x_{l-1}) - F(x_i) \end{aligned}$$

Grafische Darstellung der Verteilungsfunktion

→ monoton wachsende Treppenfunktion

Beispiel 4.3 (Verteilungsfunktion der Haushaltsgröße 2000)

HH-größe x_j	$f(x)$	$F(x)$
1	0,360	0,360
2	0,334	0,694
3	0,147	0,841
4	0,115	0,956
5 und mehr	0,044	1,000



$$f(2 < x \leq 4) = F(4) - F(2) = 0,956 - 0,694 = 0,262$$



Listing 4.3: example_ecdf.R

```
1 data(Boston, package="MASS")
2 # (diskrete) empirische kumulative Verteilungsfunktion
3 plot(ecdf(Boston$rad))
4 # empirische kumulative Verteilungsfunktion
5 plot(ecdf(Boston$lstat))
```

Verteilung klassierter Variablen

Empirische Häufigkeitsverteilung (klassierter Daten)

Beobachtungswerte einer **stetigen** Variablen

- x_1, x_2, \dots, x_n
- klassiert in k Klassen

Häufigkeitstabelle für klassierte Daten

j	Klassen $x_j^u < X \leq x_j^o$	absolute Klassenhäufigkeit $h(x_j) = h(x_j^u < X \leq x_j^o)$	relative Klassenhäufigkeit $f(x_j) = f(x_j^u < X \leq x_j^o)$
1	$x_1^u - x_1^o$	$h(x_1)$	$f(x_1)$
⋮	⋮	⋮	⋮
j	$x_j^u - x_j^o$	$h(x_j)$	$f(x_j)$
⋮	⋮	⋮	⋮
k	$x_k^u - x_k^o$	$h(x_k)$	f_k
Summe		n	1

Grafische Darstellung klass. Daten

Histogramm

- flächenproportionale Darstellung
 - ▶ Abszisse: Klassengrenzen x_j^u, x_j^o
 - ▶ Ordinate: Häufigkeitsdichte $h_K(x_j) = \frac{h(x_j)}{x_j^o - x_j^u}$ oder $f_K(x_j) = \frac{f(x_j)}{x_j^o - x_j^u}$
- Klassenhäufigkeit = Fläche des Rechtecks über der jeweiligen Klasse.
- Gesamtfläche unter dem Histogramm = 1

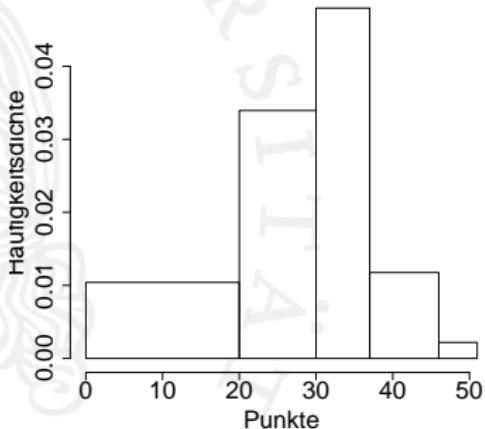
$$\sum_{j=1}^k f_K(x_j) \cdot (x_j^o - x_j^u) = \sum_{j=1}^k f(x_j) = 1$$

Beispiel 4.4 (Klausur)

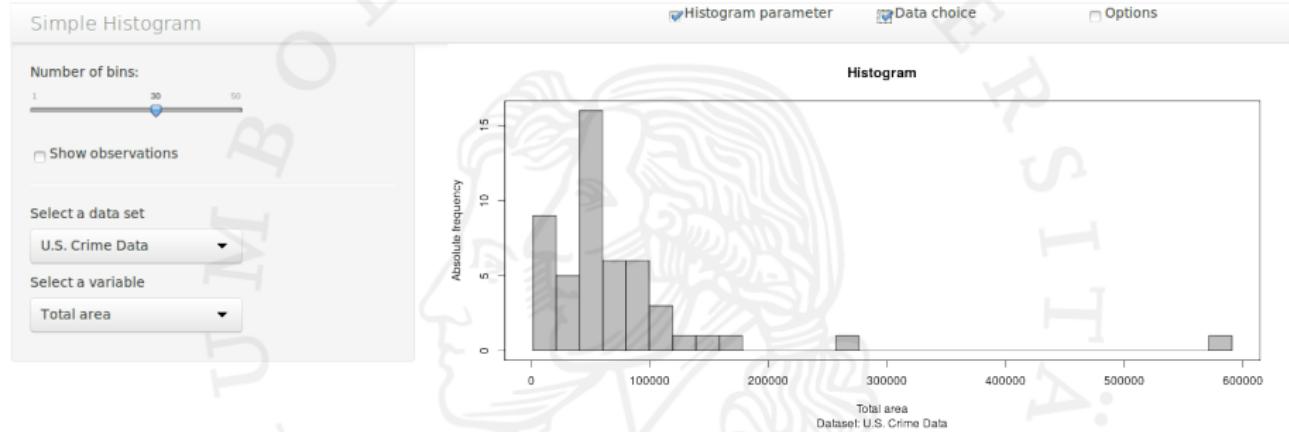
Klassierung anhand der Punkte aus der Klausur Statistik I vom Juli 2002

j	$x_j^u \leq X < x_j^o$	$h(x_j)$	$f(x_j)$	$f_K(x_j)$
1	0 – 20	57	0,208	0,010
2	20 – 30	93	0,339	0,034
3	30 – 37	92	0,336	0,048
4	37 – 46	29	0,106	0,012
5	46 – 51	3	0,011	0,002
Summe		274	1,000	

Histogramm nach Notenklassen



flächenproportionale Darstellung



http://u.hu-berlin.de/men_hist

Stamm-Blatt-Diagramm

- Halbgrafische Darstellung der Werte einer Beobachtungsreihe eines metrisch skalierten Merkmals
- Erste Ziffer eines Beobachtungswertes = Stamm (stem)
- Zweite Ziffer eines Beobachtungswertes = Blatt (leaf)
- stem width = Stammbreite
- Mit der Vergabe der Stamm- und Blattziffern ist eine Klassenbildung verbunden.

Konstruktion:

z.B.: Beobachtung: 47

stem width = 10, stem = 4, leaf = 7

$$4 \cdot 10 + 7 = 47$$

Beispiel 4.5

Beobachtungsreihe:

32, 32, 35, 36, 40, 44, 47, 48, 53, 57, 100, 105

Frequency Stem & Leaf

2,00 3 * 22

2,00 3 . 56

2,00 4 * 04

Stem width: 10,00

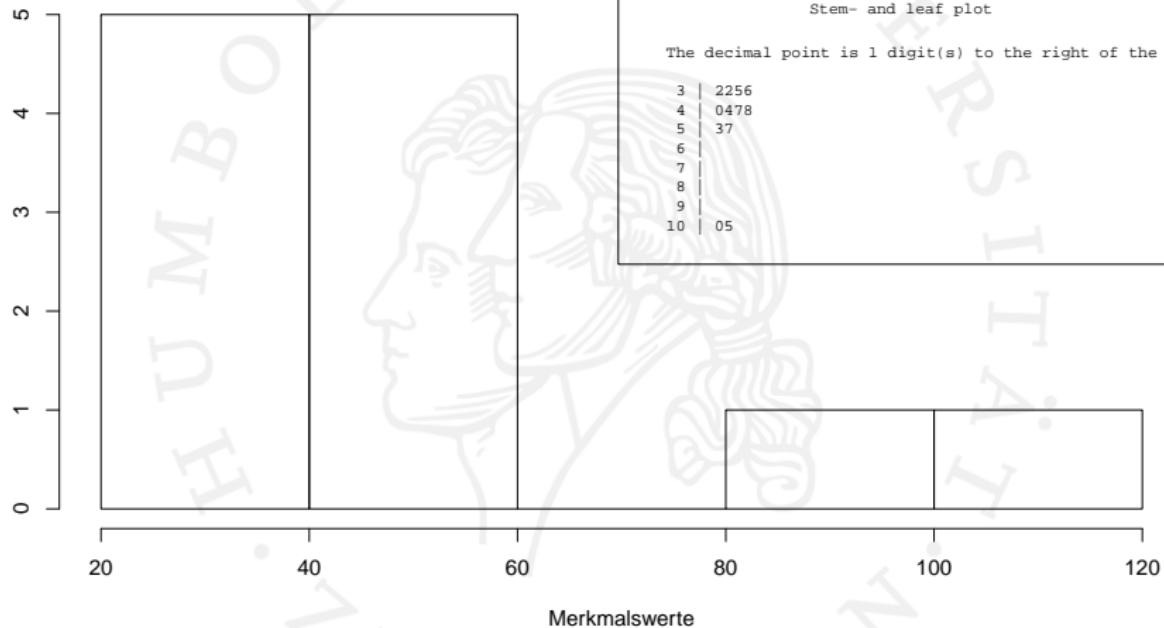
2,00 4 . 78

Each leaf: 1 case

1,00 5 * 3

1,00 5 . 7

2,00 Extremes (≥ 100)





Listing 4.4: example_stem.R

```
1 data(Boston, package="MASS")
2 stem(Boston$dis)
3 # fuer besseren Uebersichtlichkeit scale erhoeht
4 stem(Boston$dis, scale=2)
```

Verteilungsfunktion klass. Variablen

Empirische Verteilungsfunktion

$$F(x) = \begin{cases} 0 & \text{für } x \leq x_1^u \\ \sum_{i=1}^{j-1} f(x_i) + \frac{x - x_j^u}{x_j^o - x_j^u} f(x_j) & \text{für } x_j^u < x \leq x_j^o \\ 1 & \text{für } x_k^o < x \end{cases}$$

x_j^u untere Klassengrenze, x_j^o obere Klassengrenze

Grafische Darstellung: stückweise lineare Kurve (Polygonzug)

Beispiel 4.6 (Lampen)

Untersuchung der Lebensdauer (in Stunden) von 100 Glühlampen

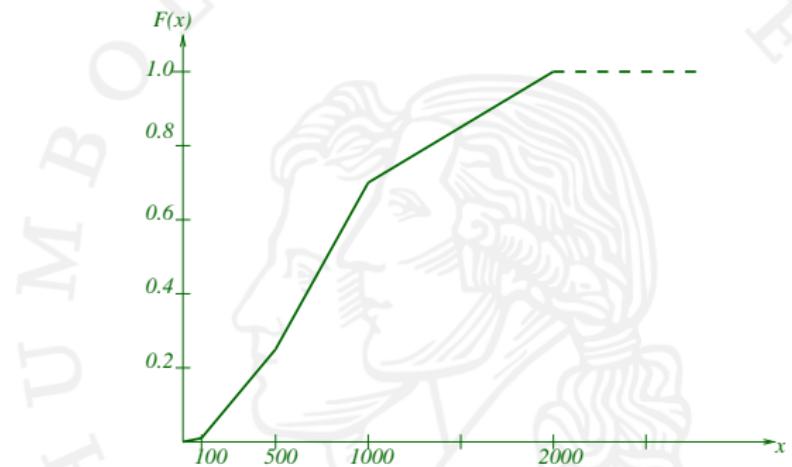
statistische Einheit: Glühlampe

Variable: Lebensdauer

metrisch, stetig

$x_j^u < X \leq x_j^o$	$h(x_j)$	$f(x_j)$	$H(x_j^o)$	$F(x_j^o)$
0 – 100	1	0.01	1	0.01
100 – 500	24	0.24	25	0.25
500 – 1000	45	0.45	70	0.70
1000 – 2000	30	0.30	100	1.00
Summe	100	1.0		

Verteilungsfunktion der Lebensdauer von Glühlampen

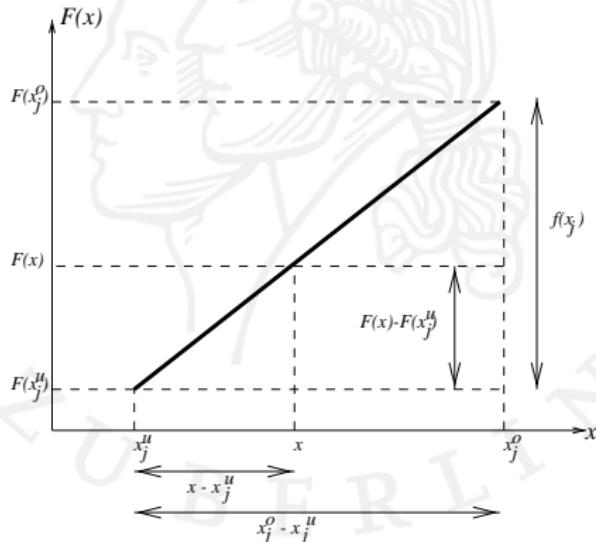


Annahme: gleichmäßige Verteilung der Beobachtungen innerhalb einer Klasse
⇒ geradlinige Verbindung der Punkte in der grafischen Darstellung

Interpolation von $F(x)$

Der Wert der Verteilungsfunktion $F(x)$ kann für jedes x im beobachteten Bereich des Merkmals X mithilfe einer Interpolation von $F(x)$ approximativ bestimmt werden:

$$F(x) = F(x_j^u) + \frac{x - x_j^u}{x_j^o - x_j^u} \cdot f(x_j)$$



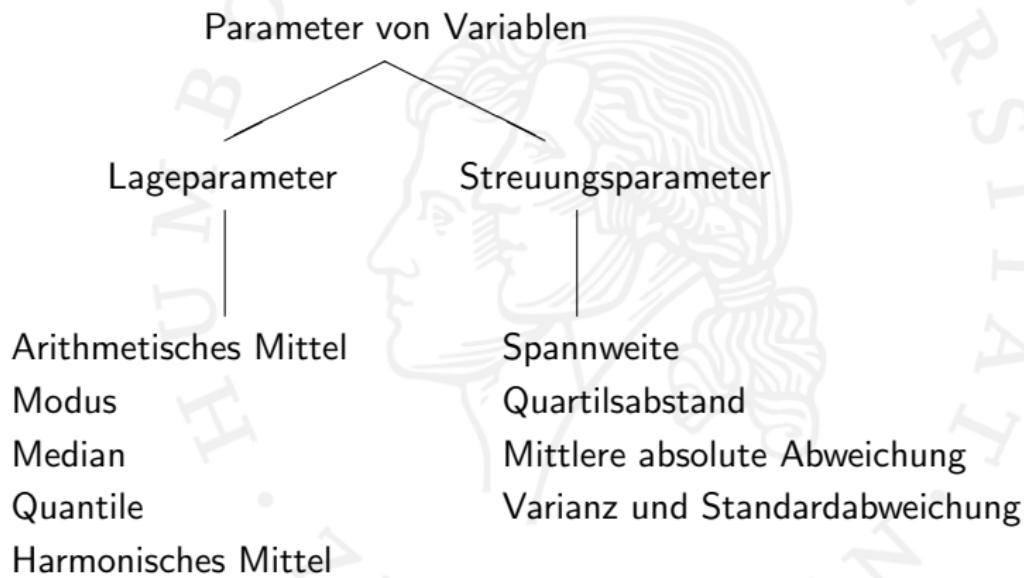
Parameter von univariaten Verteilungen

5. November 2022

- Parameter von Variablen • Lageparameter • Arithmetisches Mittel • Modus • Sortierte Beobachtungen • Median • Quantile • Lageparameter mit R • Harmonisches Mittel • Geometrisches Mittel • Streuungsparameter • Spannweite • Interquartilsabstand • Mittlere Abweichungen von c • Varianz und Standardabweichung • Relative Streuungsmaße • Streuungsmaße • Gepoolter Datensatz • Lineare Transformation • Lage- und Streuungsparametern • Fünf-Zahlen-Zusammenfassung • Boxplot • Darstellungen einer metr. Variable • Zusammenfassung

Parameter von Variablen

Parameter sind Maßzahlen, die wichtige Charakteristika einer Häufigkeitsverteilung beinhalten



Parameter

- **Lageparameter** geben an, wo die Verteilung liegt
- **Streuungsparameter** beinhalten eine Aussage über die Variabilität der Daten
- Weitere Parameter:
 - ▶ Schiefe
 - ▶ Wölbung

Robustheit

- Eine Kennzahl heißt robust, wenn sie relativ unempfindlich gegenüber Ausreißern ist

Lageparameter

Mittelwerte charakterisieren die Lage der Häufigkeitsverteilung auf der Variablenachse

Beispiel 5.1

Monatliches persönliches Einkommen in €

Mann:

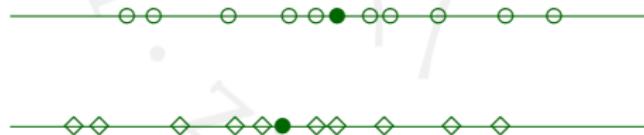
1000, 1200, 1750, 2200, 2400, 2800, 2950, 3300, 3800, 4150 (○)

$$\bar{x}_{mann} = 2555 \text{ € } (\bullet)$$

Frau:

600, 800, 1350, 1800, 2000, 2400, 2550, 2900, 3400, 3750 (◊)

$$\bar{x}_{frau} = 2155 \text{ € } (\bullet)$$



Arithmetisches Mittel

Arithmetisches Mittel \bar{x} einer empirischen Häufigkeitsverteilung

- Voraussetzung: metrisch skalierte Variable
- Ergibt sich, wenn die Summe aller beobachteten Variablenwerte gleichmäßig auf alle statistischen Einheiten aufgeteilt wird:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Klassierte Daten

- Für jede Klasse liegt vor
 - ▶ x_j^m Klassenmitte
 - ▶ n_j die Anzahl der Beobachtungen in Klasse j

$$\bar{x} = \frac{1}{n} \sum_{j=1}^k x_j^m n_j, \quad n = \sum_{j=1}^k n_j$$

Beispiel 5.2 (MHNE)

monatliches Haushaltsnettoeinkommen (MHNE) (bis unter 25 000 Euro)

MHNE von... bis unter ... Euro	Klassenmitte x_j	Anteil der HH $f(x_j)$	$F(x_j^o)$
1 – 800	400	0,044	0,044
800 – 1 400	1100	0,166	0,210
1 400 – 3 000	2200	0,471	0,681
3 000 – 5 000	4000	0,243	0,924
5 000 – 25 000	15000	0,076	1,000

$$\begin{aligned}\bar{x} &= 400 \cdot 0,044 + 1100 \cdot 0,166 + 2200 \cdot 0,471 + \\ &\quad 4000 \cdot 0,243 + 15000 \cdot 0,076 \\ &= 17,6 + 182,6 + 1036,2 + 972 + 1140 = 3348,4 \text{ Euro}\end{aligned}$$

Gewichtetes arithmetisches Mittel

- für jede Beobachtung liegt noch ein Gewicht w_i vor

$$\bar{x} = \frac{\sum_{i=1}^n x_i w_i}{\sum_{i=1}^n w_i}$$

Gepoolter Datensatz

- Datensatz zerfällt in r Teildatensätze $D = D_1 \cup \dots \cup D_r$
- Für jeden der Teildatensatz liegt vor
 - ▶ n_p die Anzahl der Beobachtungen in dem Teildatensatz
 - ▶ \bar{x}_p der Mittelwert in jedem Teildatensatz
- Arithmetisches Mittel des Gesamtdatensatzes kann ohne Kenntnis der Beobachtungswerte in den Teildatensätzen berechnet werden

$$\bar{x} = \frac{1}{n} \sum_{p=1}^r \bar{x}_p n_p, \quad n = \sum_{p=1}^r n_p$$

Beispiel 5.3 (Kinder, Volkszählung 2011)

Kinder pro Haushalt (x_i)	Anzahl Haushalte in Tsd. (w_i)	Anteil in %
0	27.800	70,4
1	6.144	15,6
2	4.205	10,6
3	1.070	2,7
4	215	0,5
5+	75	0,2
Σ	39.509	100,0

$$\bar{x} = \frac{0 \cdot 27.800 + 1 \cdot 6.144 + 2 \cdot 4.205 + 3 \cdot 1.070 + 4 \cdot 215 + 5 \cdot 75}{27.800 + 6.144 + 4.205 + 1.070 + 215 + 75} = 0,48$$

Beispiel 5.4 (Weganteil nach Entfernungsklassen)

Aus dem Mobilitätsreport 2013 der Berliner Senats:

Von	Bis unter	Anteil	Repräsentant
0	1	31%	0,5
1	3	22%	2,0
3	5	12%	4,0
5	10	17%	7,5
10 km und mehr		17%	??

Aus dem Mobilitätsreport wissen wir, dass die mittlere Weglänge aller Wege 6 km beträgt, es gilt also

$$0,31 \cdot 0,5 + 0,22 \cdot 2,0 + 0,12 \cdot 4,0 + 0,17 \cdot 7,5 + 0,17 \cdot x_5^m = 6,0$$

$$\Rightarrow x_5^m = 21,12 \text{ km}$$

Null- oder Schwerpunkteigenschaft

$$\sum_{i=1}^n (x_i - \bar{x}) = 0$$

Additionseigenschaft

$$z_i = x_i + y_i \quad \bar{z} = \bar{x} + \bar{y}$$

Modus

Modus x_D (Modalwert, Dichtemittel, häufigster Wert)

- Voraussetzung: nominale, ordinale, metrisch diskrete oder klassierte Daten \Rightarrow nicht metrisch stetige Daten!

Modus nicht-klassierter Variablen

- Diejenige Variablenausprägung, die am häufigsten beobachtet wurde

$$x_D = \arg \max_{x_j} f(x_j)$$

Modus klassierter Variablen (Grobberechnung)

- Die Klassenmitte der Klasse mit der größten Häufigkeitsdichte

$$x_D = \{x_M^m \mid M = \arg \max_k f_K(x_k)\}$$

Feinberechnung

- Ziel: den Modus in der Modalklasse ein bisschen in Richtung der größten benachbarten Klasse zu schieben
- Hinweis: Die Modalklasse ist die Klasse mit der größten Häufigkeitsdichte, nicht der größten Häufigkeit!

$$x_D = x_M^u + \frac{f_K(x_M) - f_K(x_{M-1})}{2f_K(x_M) - f_K(x_{M-1}) - f_K(x_{M+1})} \cdot (x_M^o - x_M^u)$$

x_M^u, x_M^o untere/obere Grenze der Modalklasse

$f_K(x_M)$ Häufigkeitsdichte der Modalklasse

$f_K(x_{M-1})$ Häufigkeitsdichte der Klasse vor der Modalklasse

$f_K(x_{M+1})$ Häufigkeitsdichte der Klasse nach der Modalklasse

$$f_K(x_j) = \frac{f(x_j)}{x_j^o - x_j^u}$$

Beispiel 5.5 (Lampen)

$x_j^u < X \leq x_j^o$	$h(x_j)$	$f(x_j)$	$f_K(x_j)$
0 – 100	1	0,01	0,0001
100 – 500	24	0,24	0,0006
500 – 1000	45	0,45	0,0009
1000 – 2000	30	0,30	0,0003
Summe	100	1,00	

- Modalklasse: 500 – 1000 Stunden
- (grober) Modus: 750 Stunden
- Feinberechnung:

$$x_D = 500 + \frac{0,0009 - 0,0006}{2 \cdot 0,0009 - 0,0006 - 0,0003} \cdot 500 = 666,67$$

Beispiel 5.6 (Klausur)

Klassierung anhand der Punkte aus der Klausur Statistik I vom Juli 2002

$x_j^u < X \leq x_j^o$	$h(x_j)$	$f(x_j)$	$f_K(x_j)$
0 – 20	57	0,208	0,010
20 – 30	93	0,339	0,034
30 – 37	92	0,336	0,048
37 – 46	29	0,106	0,012
46 – 51	3	0,011	0,020
Summe	274	1,000	

Modalklasse: 30 – 37 Punkte, da dort die größte Häufigkeitsdichte ist.

$$x_D = 30 + \frac{0,048 - 0,034}{2 \cdot 0,048 - 0,034 - 0,012} \cdot (37 - 30) = 31,96$$

Beispiel 5.7 (Weganteil nach Entfernungsklassen)

Aus dem Mobilitätsreport 2013 der Berliner Senats:

Von	Bis unter	Anteil	Häufigkeitsdichte
0	1	31%	0,31
1	3	22%	0,11
3	5	12%	0,06
5	10	17%	0,03
10 km und mehr		17%	<0,03

Modalklasse: 0-1 km, da dort die größte Häufigkeitsdichte ist.

$$x_D = 0 + \frac{0,31 - 0,00}{2 \cdot 0,31 - 0,11 - 0,00} \cdot (1 - 0) = 0,61$$

Sortierte Beobachtungen

- x_i der i te Beobachtungswert
- $x_{(i)}$ der sortierte i te Beobachtungswert

Beispiel 5.8

i	x_i	$x_{(i)}$	
1	$x_1 = 0,69$	$x_{(1)} = 0,11$	$= \min_i x_i$
2	$x_2 = 0,11$	$x_{(2)} = 0,25$	
3	$x_3 = 0,34$	$x_{(3)} = 0,34$	
4	$x_4 = 0,25$	$x_{(4)} = 0,43$	
5	$x_5 = 0,43$	$x_{(5)} = 0,69$	$= \max_i x_i$

Median

- Median $x_{0,5}$ wird auch als 50% Quantil oder Zentralwert bezeichnet
- Links und rechts vom Median liegen jeweils 50% der Beobachtungswerte
- robuster Lageparameter
- Voraussetzung: mindestens ordinalskalierte Variablen

Median von nicht klassierten Variablen

- Falls n ungerade ist:

$$x_{0,5} = x_{\left(\frac{n+1}{2}\right)}$$

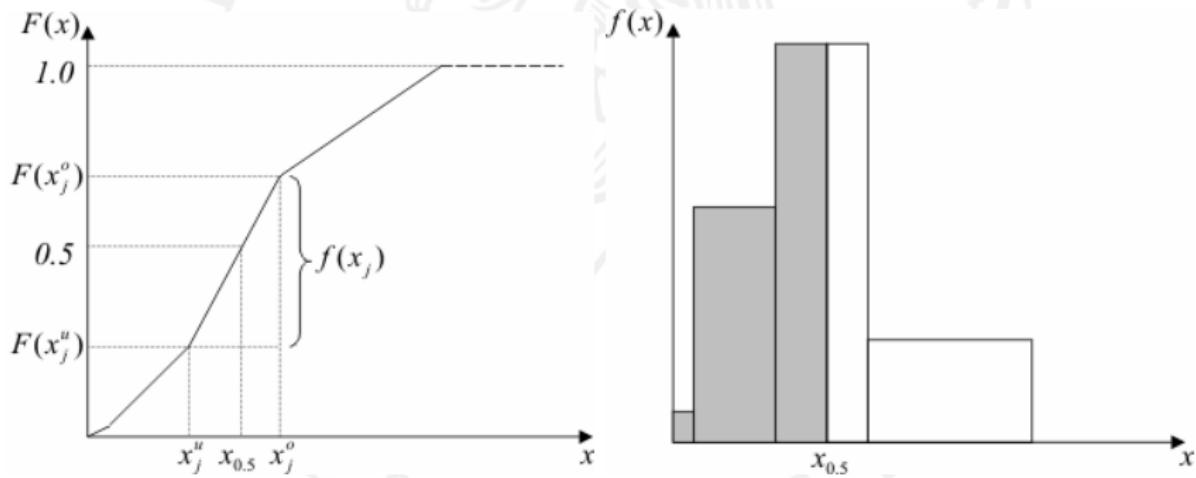
- Falls n gerade ist:

$$x_{0,5} = \frac{1}{2} \left\{ x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)} \right\}$$

Median von klassierten Variablen

$$F(x_{0,5}) = 0,5 \iff x_{0,5} = x_j^u + \frac{0,5 - F(x_j^u)}{f(x_j)} \cdot (x_j^o - x_j^u)$$

mit j die erste Klasse für die gilt $F(x_j^o) \geq 0,5$.



Der Median macht immer fifty-fifty

Die Unstatistik des Monats September ist die (inzwischen geänderte) Titelzeile einer Pressemitteilung des Deutschen Bundestages, die auf eine Antwort auf eine Kleine Anfrage der AfD-Fraktion hinwies. Wie ältere Bildschirmfotos zeigen, stand zum Erscheinungsdatum am 24. September dort noch die Überschrift: „Die Hälfte verdient weniger als das Medianentgelt“ und der letzte Satz lautete: „Der Anteil der sozialversicherungspflichtig Vollzeitbeschäftigte der Kerngruppe, die ein Bruttomonatsentgelt unterhalb des bundesweiten Medianentgelts erzielten, betrug laut Bundesregierung jeweils 50 Prozent.“ ...

Quelle: [Unstatistik 09/20](#)

Quantile

- Das Quantil x_p beschreibt den Punkt auf der Variablenachse, der eine der Größe nach in aufsteigender Folge geordnete Reihe von n Variablenwerten der Anzahl nach ungefähr oder genau im Verhältnis p zu $(1 - p)$ teilt ($0 \leq p \leq 1$)
- links vom Quantil x_p liegen also $p\%$ der Daten und rechts vom Quantil x_p liegen $(1 - p)\%$ der Daten

Quantile von nicht klassierten Variablen

- Ist $n \cdot p$ keine ganze Zahl und k die auf $n \cdot p$ folgende ganze Zahl, so ist das Quantil

$$x_p = x_{(k)}$$

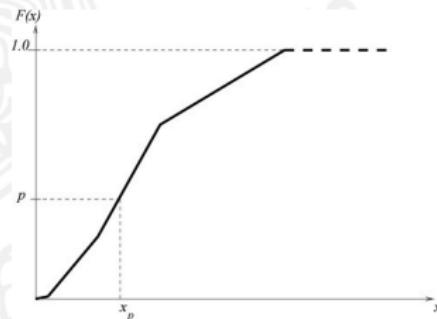
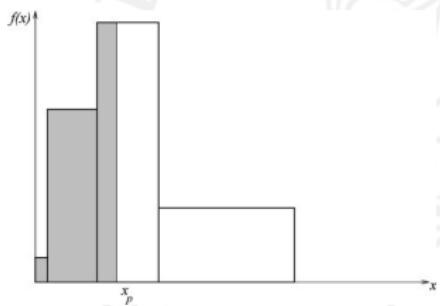
- Ist $n \cdot p$ eine ganze Zahl und $k = n \cdot p$, so könnte jeder Wert zwischen $x_{(k)}$ und $x_{(k+1)}$ als Quantil definiert werden

$$x_p = \frac{1}{2} \{ x_{(k)} + x_{(k+1)} \}$$

Quantile von klassierte Variablen

$$F(x_p) = p \iff x_p = x_j^u + \frac{p - F(x_j^u)}{f(x_j)} \cdot (x_j^o - x_j^u)$$

mit j die erste Klasse für die gilt $F(x_j^o) \geq p$.



Spezielle Quantile

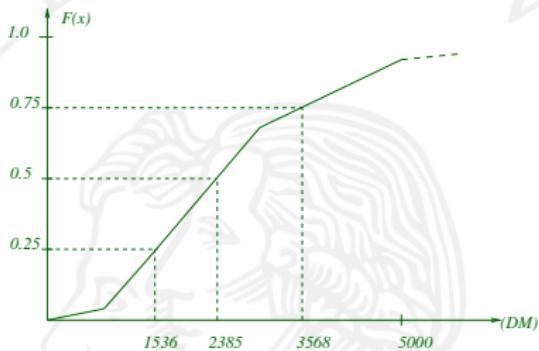
Dezile $p = s/10, \quad s = 1, \dots, 9$

Quartile $p = q/4, \quad q = 1, 2, 3$

Quintile $p = r/5, \quad r = 1, \dots, 4$

Median $p = 0,5$

Beispiel 5.9 (MHNE)



$$x_{0,25} = 1400 + 1600 \cdot \frac{(0,25 - 0,21)}{0,471} = 1535,88 \text{ Euro}$$

$$x_{0,50} = 1400 + 1600 \cdot \frac{(0,5 - 0,21)}{0,471} = 2385,14 \text{ Euro}$$

$$x_{0,75} = 3000 + 2000 \cdot \frac{(0,75 - 0,681)}{0,243} = 3567,90 \text{ Euro}$$

Beispiel 5.10 (Weganteil nach Entfernungsklassen)

Aus dem Mobilitätsreport 2013 der Berliner Senats:

Von	Bis unter	Anteil	Verteilungsfunktion
0	1	31%	0,31
1	3	22%	0,53
3	5	12%	0,65
5	10	17%	0,82
10 km und mehr		17%	1,00

$$x_{0,25} = 0 + 1 \cdot \frac{(0,25 - 0,00)}{0,31} = 0,81 \text{ km}$$

$$x_{0,50} = 1 + 2 \cdot \frac{(0,5 - 0,31)}{0,22} = 2,72 \text{ km}$$

$$x_{0,75} = 5 + 5 \cdot \frac{(0,75 - 0,65)}{0,12} = 9,17 \text{ km}$$

Lageparameter mit R

R Listing 5.1: example_location.R

```
1 library("MASS")
2 # Mittelwert
3 mean(Boston$tax)
4 # Median
5 median(Boston$tax)
6 # 75% Quantil
7 quantile(Boston$tax, 0.75)
8 # Kein Befehl fuer den Modus
9 tab <- table(Boston$rad)
10 tab
11 max(tab)
12 which.max(tab)
```

Harmonisches Mittel

Das harmonische Mittel wird verwendet, wenn der Mittelwert von Verhältniszahlen (Quotient zweier Größen) gesucht ist und gilt für Variablenwerten $\neq 0$

$$\bar{x}_H = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$$

Eigenschaft des harmonischen Mittels:

- $\min_i x_i \leq \bar{x}_H \leq n \min_i x_i$
- Wenn min. ein $x_i = 0$, dann ist $\bar{x}_H = 0$

Gewichtetes harmonisches Mittel

$$\bar{x}_H = \frac{\sum_{i=1}^n w_i}{\sum_{i=1}^n \frac{w_i}{x_i}} \quad w_i > 0$$

Beispiel 5.11 (Geschwindigkeit)

- statistische Variable: Geschwindigkeit
- Durchschnittsgeschwindigkeit: Gesamtstrecke dividiert durch die benötigte Gesamtzeit

Gegeben:

Teilstrecke i	1	2	3	4
Länge w_i in km	2	4	3	8
Geschwindigkeit x_i in km/h	40	50	80	100

Gesamtzeit: $\sum_{i=1}^k \frac{w_i}{x_i} = 0,2475$ (Stunden)

Gesamtstrecke: $\sum_{i=1}^k w_i = 17$ km

Durchschnittsgeschwindigkeit: $17/0,2475 = 68,687$ km/Std.

Direkte Anwendung des arithmetischen Mittels → falsch:

$$\bar{x} = \frac{40 \cdot 2 + 50 \cdot 4 + 80 \cdot 3 + 100 \cdot 8}{2 + 4 + 3 + 8} = 77,647$$

$$\bar{x} = \frac{\sum_{i=1}^k x_i w_i}{\sum_{i=1}^k w_i} = \frac{\sum \frac{\text{Länge}}{\text{Zeit}} \cdot \text{Länge}}{\sum \text{Länge}} \neq \frac{\text{Gesamtlänge}}{\text{Gesamtzeit}}$$

Anwendung des harmonischen Mittels → richtig:

$$\bar{x}_H = \frac{17}{0,2475} = \frac{2 + 4 + 3 + 8}{\frac{2}{40} + \frac{4}{50} + \frac{3}{80} + \frac{8}{100}} = 68,687 \frac{\text{km}}{\text{Std.}}$$

Es sind Informationen zum Zähler des Verhältnisses gegeben

- ein Durchschnitt aus Verhältniszahlen ist zu berechnen
- es sind Zusatzinformationen (Häufigkeiten, Gewichte) gegeben, die sich inhaltlich auf den Zähler der Verhältniszahlen beziehen
 - ⇒ das harmonische Mittel muss benutzt werden

Es sind Informationen zum Nenner des Verhältnisses gegeben

- ein Durchschnitt aus Verhältniszahlen ist zu berechnen
- es sind Zusatzinformationen (Häufigkeiten, Gewichte) gegeben, die sich inhaltlich auf den Nenner der Verhältniszahlen beziehen
 - ⇒ das arithmetische Mittel muss benutzt werden

Beispiel 5.12 (Kurs-Gewinn-Verhältnis)

Firma	A	B
Marktkapitalisierung (in Mio. EUR)	150	1
Gewinn (in Mio. EUR)	5	0,001
Kurs-Gewinn-Verhältnis (P/E)	30	1000
Aktienanteil im Fond (in %)	30	70

Mittleres Kurs-Gewinn-Verhältnis (für den Fond):

- Anwendung des arithmetischen Mittels → falsch

$$0,3 * 30 + 0,7 * 1000 = 710$$

- Anwendung des harmonischen Mittels → richtig:

$$\frac{0,3 + 0,7}{\frac{0,3}{30} + \frac{0,7}{1000}} \approx 93,46$$

Geometrisches Mittel

Voraussetzungen

- (mindestens) verhältnisskalierte Variablen
- nur positive Variablenwerte

Geometrisches Mittel

- n -te Wurzel aus dem Produkt der Variablenwerte x_1, \dots, x_n :

$$\bar{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$$

- arithmetisches Mittel der Logarithmen der Beobachtungswerte:

$$\log \bar{x}_G = \frac{1}{n} \sum_{i=1}^n \log x_i$$

Entwicklungsrationen

- y_0, y_1, \dots, y_T : zeitlich geordnete Beobachtungswerte von einem Basiszeitraum 0 bis zu einem Berichtszeitraum T
- Entwicklungsrationen von Zeitraum zu Zeitraum (verhältnisskaliert)

$$x_t = y_t / y_{t-1}$$

- relative Gesamtentwicklung: multiplikative Verknüpfung der Entwicklungsrationen

$$x_1 \cdot x_2 \cdot \dots \cdot x_T = \frac{y_1}{y_0} \cdot \frac{y_2}{y_1} \cdots \frac{y_T}{y_{T-1}} = \frac{y_T}{y_0}$$

- mittlere Entwicklungsrate: geometrisches Mittel aus den Entwicklungsrationen

$$\bar{x}_G = \sqrt[T]{x_1 \cdot x_2 \cdot \dots \cdot x_T} = \sqrt[T]{\frac{y_T}{y_0}}$$

Gewichtetes geometrisches Mittel

- Mit $w = w_1 + w_2 + \dots + w_T$

$$\bar{x}_G = \sqrt[w]{x_1^{w_1} \cdot x_2^{w_2} \cdot \dots \cdot x_T^{w_T}}$$

- Zusammenhang mit dem arithmetisches und harmonischen Mittel

$$\bar{x}_H \leq \bar{x}_G \leq \bar{x}$$

Streuungsparameter

Streuung (Dispersion) ist die Variabilität in den beobachteten Werten einer metrisch skalierten Variablen.

Beispiel 5.13

Monatliche Aufwendungen für Freizeitgüter und Urlaub in €

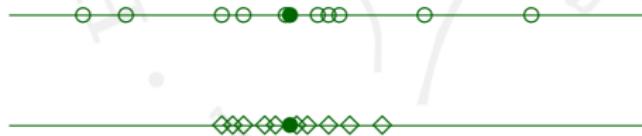
Zweipersonenhaushalte:

105, 125, 170, 180, 200, 215, 220, 225, 240, 315 (○)

Vierpersonenhaushalte:

170, 175, 180, 190, 195, 205, 210, 220, 230, 245 (◊)

$$\bar{x} = 202 \text{ €} (\bullet)$$



Spannweite

→ auch Range, Schwankungsbereich, Variationsbreite

1) Spannweite von nicht klassierten Variablen

$$R_x = x_{\max} - x_{\min} = x_{(n)} - x_{(1)}$$

wobei $x_{(1)}, \dots, x_{(n)}$ geordnete Beobachtungen sind

Beispiel 5.14 (Weiterführung Monatliche Aufwendungen für Freizeitgüter und Urlaub)

Spannweite für Zweipersonenhaushalte: $R_x = 315 - 105 = 210$

2) (approximative) Spannweite von klassierten Variablen

$$R_x = x_k^o - x_1^u$$

wobei x_k^o die obere Klassengrenze der letzten Klasse und x_1^u die untere Klassengrenze der ersten Klasse ist

Interquartilsabstand

- Der (**Inter-**)**Quartilsabstand** ist die Differenz zwischen dem dritten Quartil $x_{0,75}$ und dem ersten Quartil $x_{0,25}$:

$$QA_x = x_{0,75} - x_{0,25}$$

- Der Quartilsabstand ist ein robuster Parameter (im Gegensatz zur Spannweite)
- Der **Quartilsdispersionskoeffizient** ist der relativier Quartilsabstand bezogen auf den Median.

$$QA_{r;x} = \frac{QA_x}{x_{0,5}} \quad \left(\text{ähnlich: } \frac{x_{0,75} - x_{0,25}}{x_{0,75} + x_{0,25}} \right)$$

- Der Quartilsdispersionskoeffizient ist ein relativer robuster Streuungsparameter

Mittlere Abweichungen von c

Mittlere quadratische Abweichung

Die mittlere quadratische Abweichung ist das arithmetische Mittel aus den quadrierten Abweichungen der Variablenwerte von einem Bezugspunkt c auf der Variablenachse:

$$MQ_x(c) = \frac{1}{n} \sum_{i=1}^n (x_i - c)^2$$

Mittlere absolute Abweichung

Die mittlere absolute Abweichung ist das arithmetische Mittel aus den absoluten Abweichungen der Variablenwerte von einem Bezugspunkt c auf der Variablenachse:

$$MAD_x(c) = \frac{1}{n} \sum_{i=1}^n |x_i - c|$$

Im Allgemeinen wählt man: $c = \bar{x}$ oder $c = x_{0,5}$

Varianz und Standardabweichung

Die Varianz ist die mittlere quadratische Abweichung vom arithmetischen Mittel: $c = \bar{x}$

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

Die Standardabweichung ist die positive Quadratwurzel aus der Varianz:

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

Die Standardabweichung wird in der gleichen Einheit gemessen, wie die Daten (im Gegensatz zur Varianz)

Relative Streuungsmaße

- einheitslose Koeffizienten
- Variationskoeffizient (für $\bar{x} > 0$)

$$v = s/\bar{x}$$

- Normierter Variationskoeffizient (für $x_i \geq 0$)

$$v^* = \frac{v}{\sqrt{n-1}} \leq 1$$

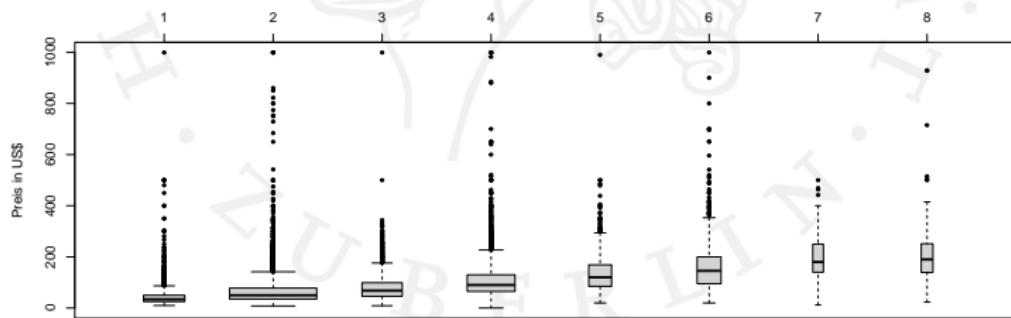
- Variationskoeffizienten klein $\Rightarrow \bar{x}$ ist guter Repräsentant der Daten
(Faustregel nach Eckstein: $v < 0,5$)
- Quartilsdispersionskoeffizient (robust, für $x_{0,5} > 0$)

$$v_r = QA/x_{0,5}$$

Beispiel 5.15

Übernachtungspreise (in US\$) in Berlin bei Airbnb nach max. Anzahl der Gäste

	1	2	3	4	5	6	7	8
n	1893	8467	1809	2545	579	640	125	152
\bar{x}	44.47	65.80	81.94	111.07	139.74	168.85	201.36	212.80
s	48.18	57.44	56.69	83.41	85.82	113.94	91.16	131.55
v	1.08	0.87	0.69	0.75	0.61	0.67	0.45	0.62
$x_{0.5}$	34.00	50.00	68.00	90.00	120.00	145.50	180.00	190.00
QA	25.00	43.00	53.00	65.00	84.00	105.00	110.00	111.75
v_r	0.74	0.86	0.78	0.72	0.70	0.72	0.61	0.59



Streuungsmaße

R Listing 5.2: example dispersion.R

```
1 library("MASS")
2 x <- Boston$crim
3 IQR(x)           # Interquartilsabstand
4 mad(x)           # mittlere absolute Abweichung
5 IQR(x)/median(x) # Quartilsdispersionskoeffizient
6
7 # Reskalierung, weil R die Stichprobenvarianz berechnet.
8 n      <- length(x)
9 sigma2 <- var(x)*((n-1)/n)
10 sigma2          # Varianz
11 sqrt(sigma2)    # Standardabweichung
12 sd(Boston$crim) #
13 sqrt(sigma2)/mean(x) # Variationskoeffizient
14 diff(range(x))   # Spannweite
```

Gepoolter Datensatz

$$D = D_1 \cup \dots \cup D_r \text{ mit } \bar{x}_1, \dots, \bar{x}_r \text{ und } s_1^2, \dots, s_r^2$$

und n_1, \dots, n_r mit $n = n_1 + \dots + n_r$

Dann gilt

$$\bar{x} = \frac{1}{n} \sum_{p=1}^r \bar{x}_p n_p, \quad n = \sum_{p=1}^r n_p$$

$$s^2 = \sum_{p=1}^r \frac{n_p}{n} s_p^2 + \sum_{p=1}^r \frac{n_p}{n} (\bar{x}_p - \bar{x})^2$$

Streuungszerlegung

gesamte Varianz = Varianz *innerhalb* der Teilmassen
 + Varianz *zwischen* den Teilmassen

Lineare Transformation

$$y_i = a + b \cdot x_i \quad (b \neq 0)$$

a : Verschiebung der Daten

$0 < b < 1$: Stauchung der Daten

$b > 1$: Streckung der Daten

$b < 0$: Spiegelung am Ursprung mit Stauchung
oder Streckung

Standardisierung

$$z_i = a + bx_i \quad \text{mit} \quad a = -\bar{x}/s_x, \quad b = 1/s_x$$

$$z_i = \frac{x_i - \bar{x}}{s_x} \Rightarrow \bar{z} = 0, \quad s_z^2 = 1$$



Listing 5.3: example_scale.R

```
1 library("MASS")
2 # Mittelwert
3 mean(Boston$crim)
4 # Standardabweichung
5 sd(Boston$crim)
6 plot(ecdf(Boston$crim))
7 # Standardisierung (Lineare Transformation)
8 z <- scale(Boston$crim)
9 # Mittelwert standardisiert
10 mean(z)
11 # Standardabweichung standardisiert
12 sd(z)
13 plot(ecdf(z))
```

Lineare Transformation

– des arithmetischen Mittels

$$\bar{y} = a + b\bar{x}$$

– des Medians

$$y_{0,5} = a + bx_{0,5}$$

– der Varianz

$$s_y^2 = b^2 s_x^2 \quad s_y = |b| s_x$$

– der mittleren absoluten Abweichung

$$MAD_y(c) = |b| \cdot MAD_x(c)$$

– der Spannweite

$$R_y = |b|R_x$$

– des Quartilsabstandes

$$QA_y = |b| QA_x$$

Lage- und Streuungsparametern

Mittelwert und Varianz: quadratische Minimumseigenschaft

$$\sum_{i=1}^n (x_i - \bar{x})^2 \leq \sum_{i=1}^n (x_i - c)^2$$

Median und MAD: lineare Minimumseigenschaft

$$\sum_{i=1}^n |x_i - x_{0,5}| \leq \sum_{i=1}^n |x_i - c|$$

Fünf-Zahlen-Zusammenfassung

- Fasst fünf Kennzahlen einer Verteilung zusammen
 - ▶ Minimum
 - ▶ Unteres Quartil (25% Quantil)
 - ▶ Median (50% Quantil)
 - ▶ Oberes Quartil (75% Quantil)
 - ▶ Maximum

Median	
Unteres Quartil	Oberes Quartil
Minimum	Maximum

- Vorläufer des Boxplots

Beispiel 5.16 (ALLBUS 2010, Nettoeinkommen, v612)

1200	
700	1750
22	10000

$$QA = 1750 - 700 = 1050$$

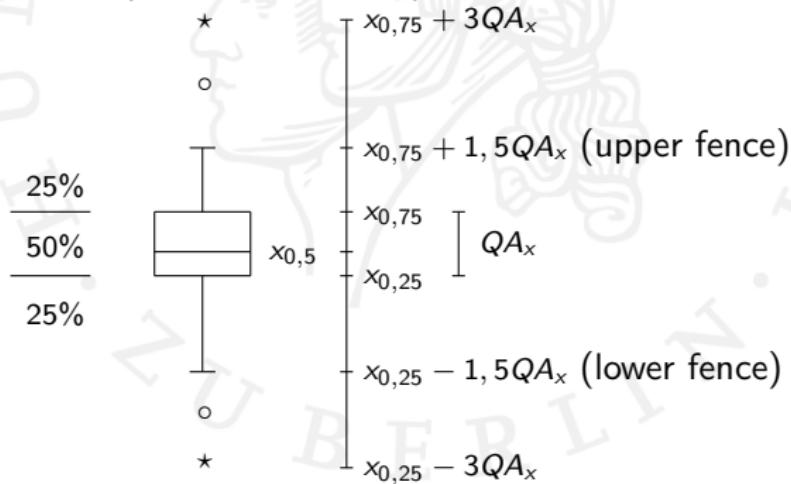
$$R = 10000 - 22 = 9978$$

R Listing 5.4: example_fivenum.R

```
1 library("MASS")
2 # Minimum, lower-hinge (1. Quartil), median,
3 # upper-hinge (3. Quartil), Maximum
4 fivenum(Boston$crim)
5 # Minimum, 1., 2., und 3. Quartil, Maximum
6 quantile(Boston$crim)
```

Boxplot

- auch Box-Whisker-Plot oder Schachtelzeichnung
- grafische Darstellung wesentlicher Kenngrößen einer Beobachtungsreihe bzw. einer Häufigkeitsverteilung einer metrisch skalierten Variablen X
- Werte, die ausserhalb des Intervalls $(x_{0,25} - 1,5QA_x, x_{0,75} + 1,5QA_x)$ liegen, werden (meist willkürlich) als Ausreißer bezeichnet



Beispiel 5.17 (ALLBUS 1980-1996)

- Statistische Einheit: befragte Person
- Statistische Variable: monatliches Nettoeinkommen der Befragten

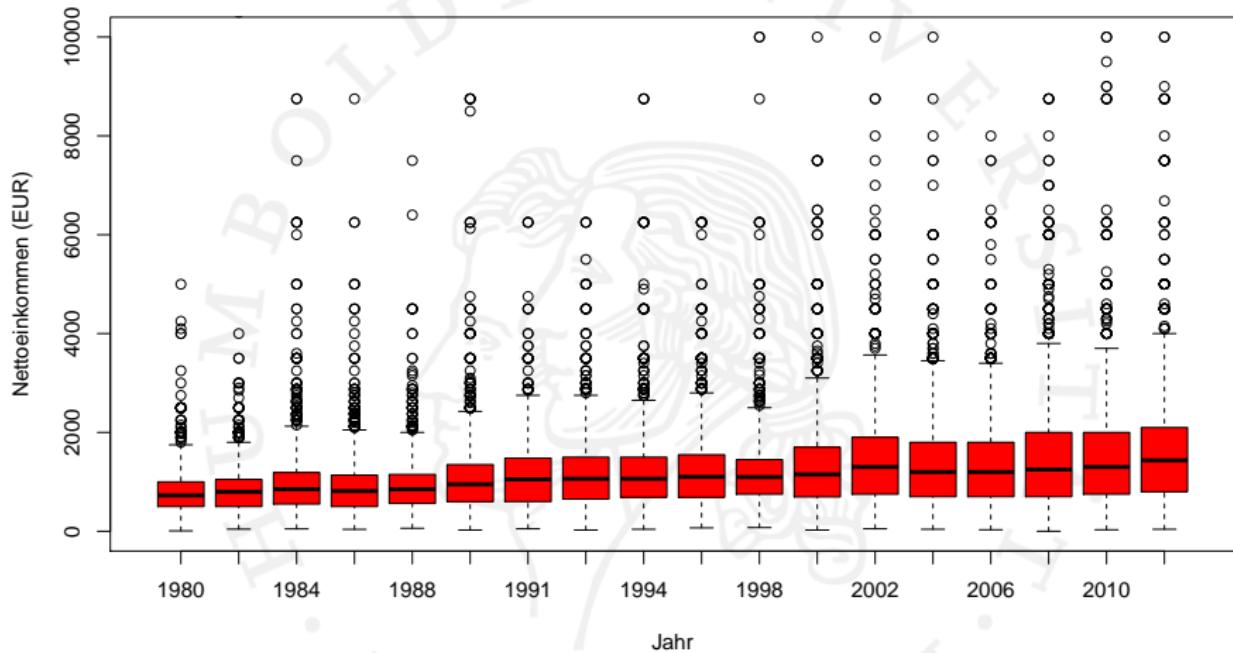
Neue Bundesländer

Jahr	\bar{x}	$x_{0,25}$	$x_{0,5}$	$x_{0,75}$	s_x	n
1991	1029	700	984	1232	503	1085
1992	1242	800	1150	1500	751	790
1994	1469	1000	1400	1800	729	726
1996	1725	1125	1625	2125	870	927

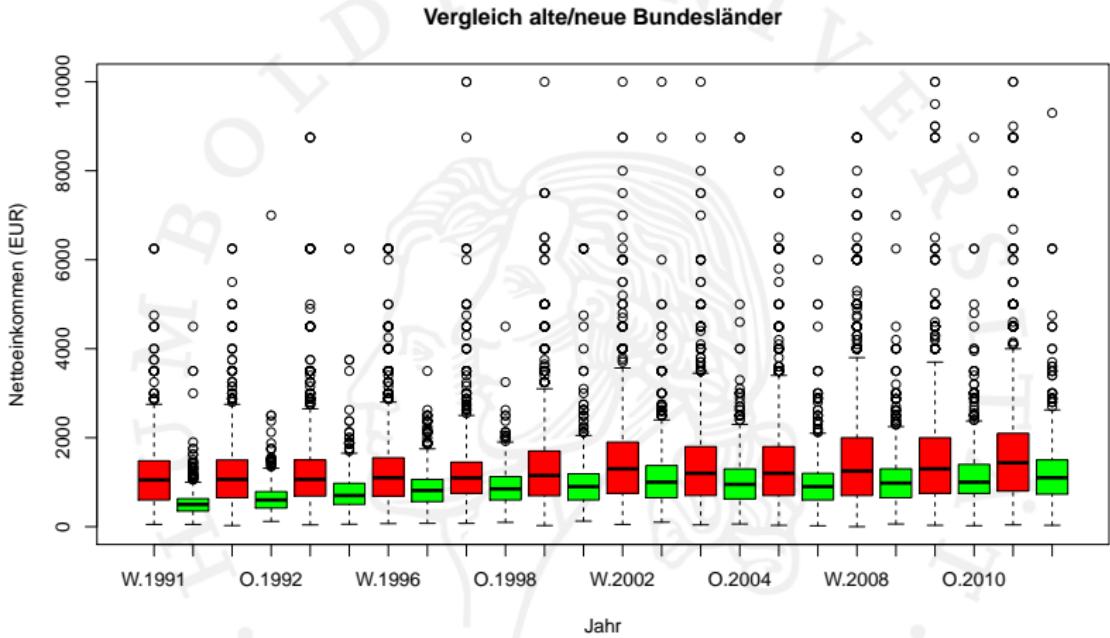
Alte Bundesländer

Jahr	\bar{x}	$x_{0,25}$	$x_{0,5}$	$x_{0,75}$	s_x	n
1980	1570	1000	1450	2000	912	1932
1982	1685	1000	1600	2100	1034	1721
1984	1761	1000	1600	2200	1199	1661
1986	1727	1000	1600	2125	1084	1697
1988	1807	1085	1650	2200	1126	1153
1990	2053	1200	1800	2600	1338	1583
1991	2110	1100	2000	2700	1307	709
1992	2232	1216	2000	2900	1399	1203
1994	2210	1200	2000	290	1343	1226
1996	2493	1375	2200	3100	1646	1689

Alte Bundesländer



Für die alten Bundesländer steigen der Median und die Streuung.

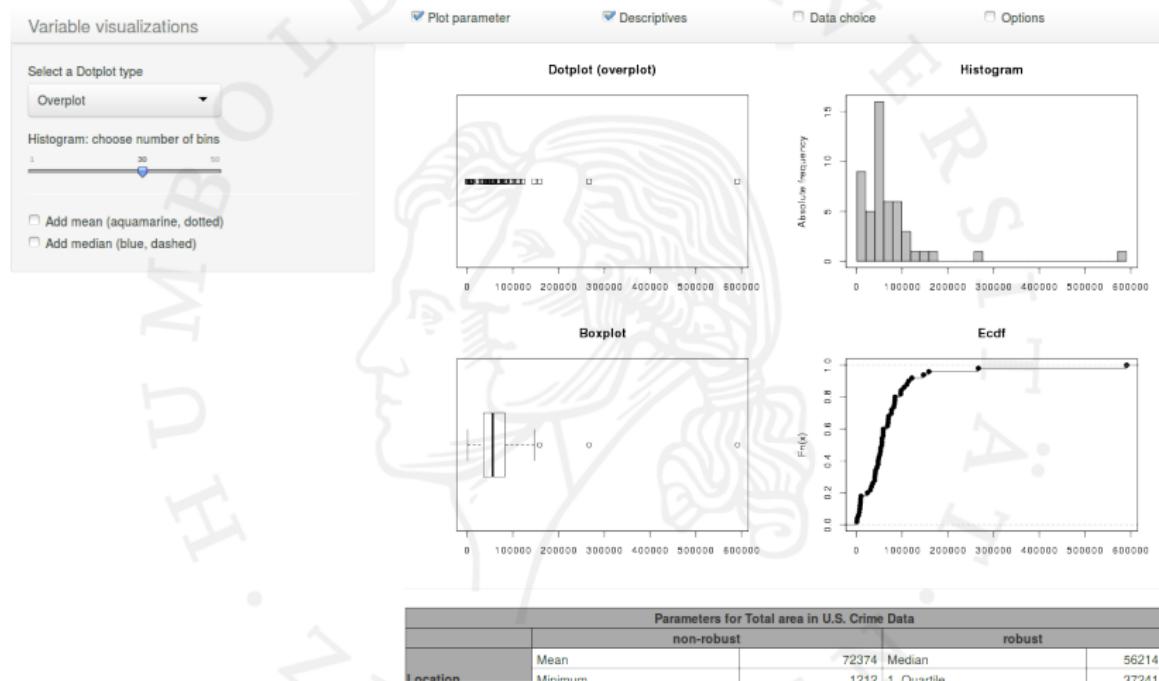


Die neuen Bundesländer nähern sich den alten Bundesländern an.
Sie sind aber noch immer unterschiedlich.

 Listing 5.5: [example_boxplot.R](#)

```
1 library("MASS")
2 boxplot(Boston$age)
```

Darstellungen einer metr. Variable



http://u.hu-berlin.de/men_vis

Zusammenfassung

		Skalenniveau					
		Nominal	Ordinal	Diskret	Stetig klassiert	Stetig unklassiert	Robust
Parameter	Parameterverwendung	meistens problemlos					
		problembehaftet					
		auf keinen Fall					
	Lage	Modus	■■■■■	■■■■■	■■■■■	■■■■■	■■■■■
	Streuung	Mittelwert	■■■■■	■■■■■	■■■■■	■■■■■	■■■■■
		Median	■■■■■	■■■■■	■■■■■	■■■■■	■■■■■
		Varianz ¹	■■■■■	■■■■■	■■■■■	■■■■■	■■■■■
		Spannweite	■■■■■	■■■■■	■■■■■	■■■■■	■■■■■
		QA	■■■■■	■■■■■	■■■■■	■■■■■	■■■■■
		MAD	■■■■■	■■■■■	■■■■■	■■■■■	■■■■■

¹ oder Standardabweichung

Bivariate Verteilungen

5. November 2022

Diskrete Variablen • Stetige Variablen • Gemeinsame Verteilung •
Randverteilung • Randverteilung Mittelwert • Randverteilung Varianz •
Bedingte empirische Verteilungen • Parameter bedingter Verteilungen

Diskrete Variablen

- X mit Merkmalsausprägungen $x_i \quad i = 1, \dots, m$
- Y mit Merkmalsausprägungen $y_j \quad j = 1, \dots, r$
- Anzahl der Paare von Variablenausprägungen ($m \cdot r$)

$$(x_i, y_j) = \{(X = x_i) \cap (Y = y_j)\}$$

Zweidimensionale Häufigkeitstabelle

auch Kontingenztabelle oder Kreuztabelle

Variable X	Variable Y					Randverteilung X
	y_1	...	y_j	...	y_r	
x_1	h_{11}	...	h_{1j}	...	h_{1r}	$h_{1\bullet}$
\vdots	\vdots	...	\vdots	...	\vdots	\vdots
x_i	h_{i1}	...	h_{ij}	...	h_{ir}	$h_{i\bullet}$
\vdots	\vdots	...	\vdots	...	\vdots	\vdots
x_m	h_{m1}	...	h_{mj}	...	h_{mr}	$h_{m\bullet}$
Randverteilung Y	$h_{\bullet 1}$...	$h_{\bullet j}$...	$h_{\bullet r}$	$h_{\bullet\bullet} = n$

Gesamtheit aller gemeinsamen Variablenausprägungen (x_i, y_j) und der dazugehörigen absoluten bzw. relativen Häufigkeiten.

- Absolute Häufigkeit: $h(x_i, y_j) = h_{ij}$
- Relative Häufigkeit: $f(x_i, y_j) = f_{ij} = \frac{h(x_i, y_j)}{n}$

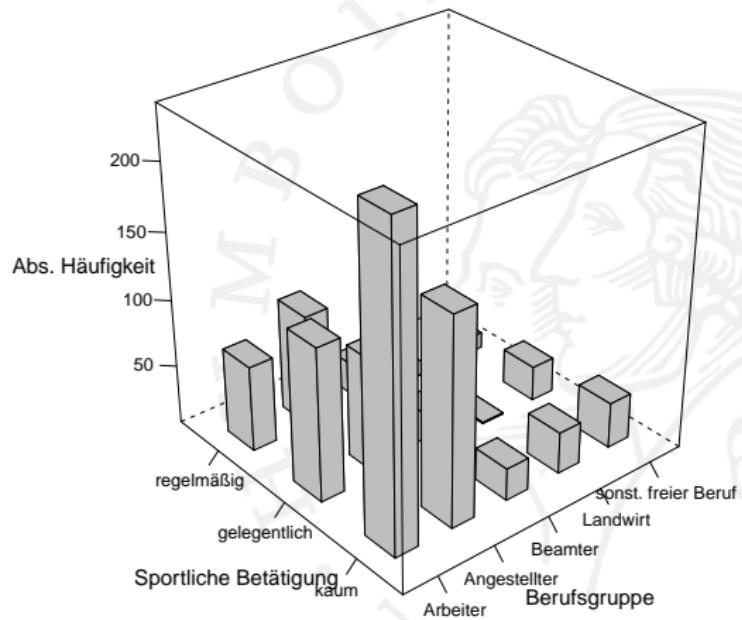
Eigenschaften:

- $\sum_{i=1}^m \sum_{j=1}^r h(x_i, y_j) = n$
- $\sum_{i=1}^m \sum_{j=1}^r f(x_i, y_j) = 1$
- $0 \leq h(x_i, y_j) \leq n$
- $0 \leq f(x_i, y_j) \leq 1$

Beispiel 6.1 (Sport)

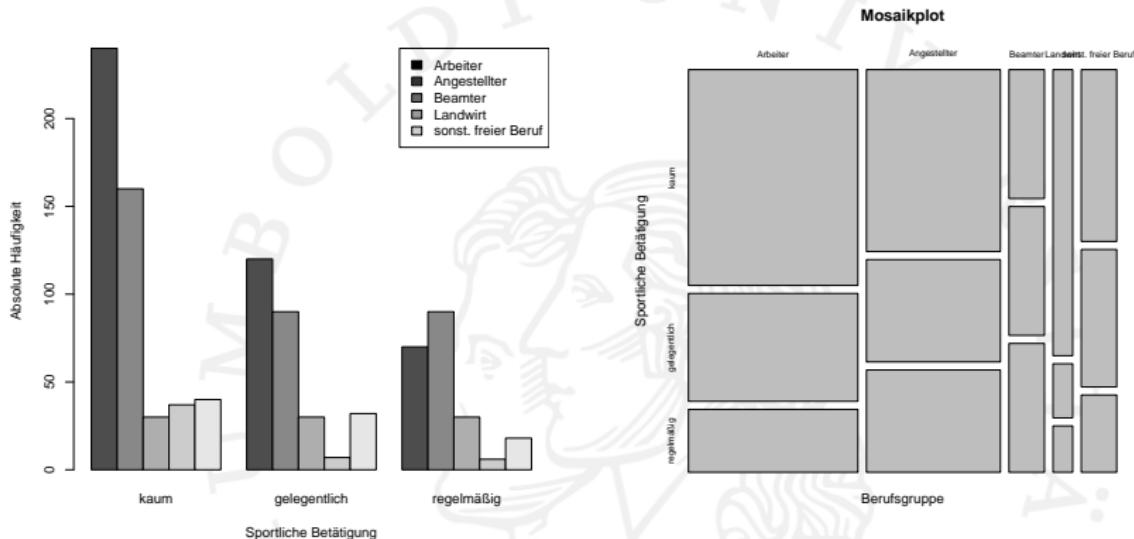
- X – Berufsgruppe (nominal) mit 5 Merkmalsausprägungen
- Y – sportliche Betätigung (nominal) mit 3 Merkmalsausprägungen
- Kontingenztabelle mit 5×3
- $n = 1000$ berufstätige Personen

Berufsgruppe (X)	sportliche Betätigung (Y)			Randverteilung X
	kaum	gelegentlich	regelmäßig	
Arbeiter	240	120	70	430
Angestellter	160	90	90	340
Beamter	30	30	30	90
Landwirt	37	7	6	50
sonst. freier Beruf	40	32	18	90
Randverteilung Y	507	279	214	1000

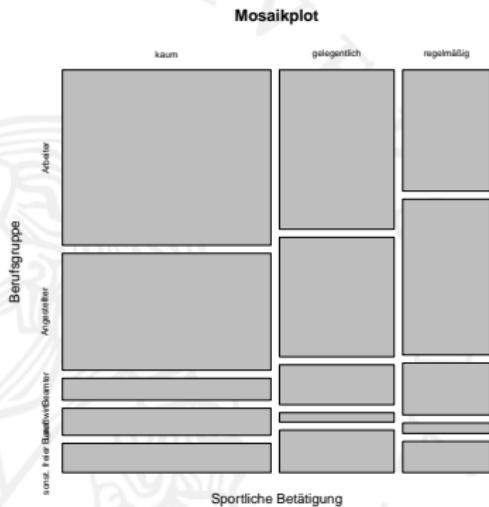
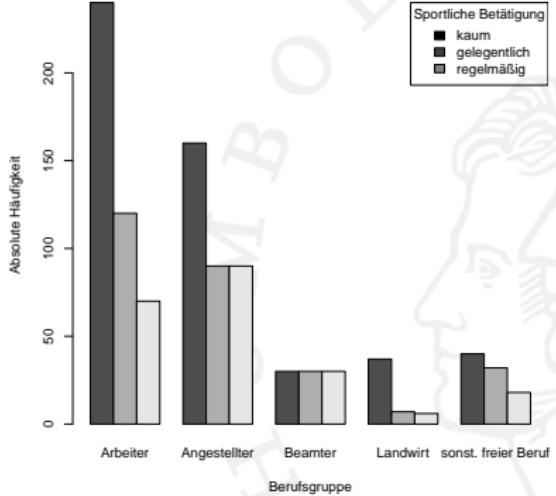


Die Interpretation
des Balkendiagramms
wird schwer, wenn

- viele Gruppen/Kategorien vorhanden sind oder
- es nicht richtig gedreht wird



- Gruppiertes Balkendiagramm: Die Höhe der Rechtecke entspricht den abs. Häufigkeiten
- Mosaikplot: Die Fläche der Rechtecke entspricht den abs. Häufigkeiten



- Beide Grafiktypen behandeln die Variablen ungleich

Beispiel 6.2 (HIV-Infektion)

- X Test auf HIV (positiv, negativ), Y – HIV Infektion (vorhanden, nicht vorhanden)
- X, Y nominalskaliert, $n = 100.000$ Personen
- 2×2 Kontingenztabelle

HIV-Test (X)	HIV Infektion (Y)		Randverteilung X
	vorhanden (y_1)	nicht vorhanden (y_2)	
positiv (x_1)	199	499	698 ($h_{1\bullet}$)
negativ (x_2)	1	99301	99302 ($h_{2\bullet}$)
Randverteilung Y	200 ($h_{\bullet 1}$)	99800 ($h_{\bullet 2}$)	100000 (n)

Stetige Variablen

Empirische Häufigkeitsverteilung
In Form einer Tabelle

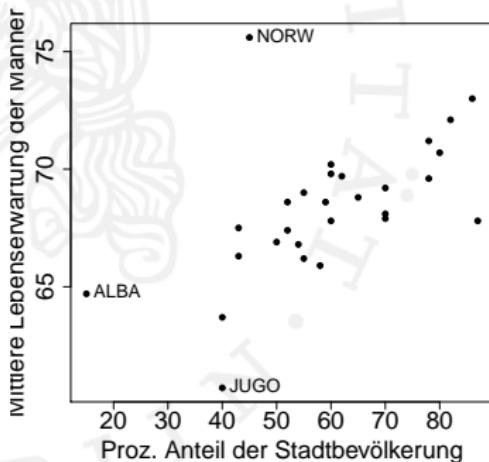
i	Variable X	Variable Y
1	x_1	y_1
\vdots	\vdots	\vdots
i	x_i	y_i
\vdots	\vdots	\vdots
n	x_n	y_n

⚠ Wenn sehr viele Daten vorhanden sind – Interpretation anhand der Tabelle schwierig.

Streudiagramm

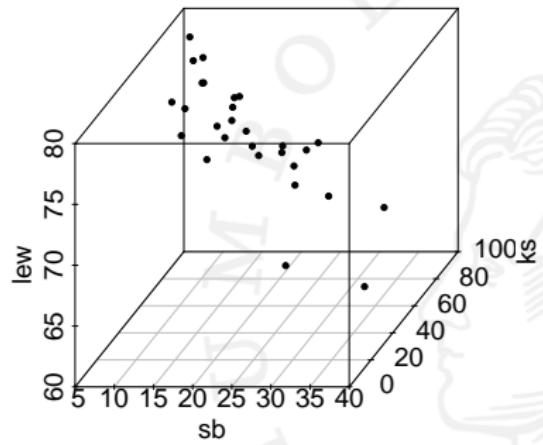
Beispiel 6.3

Anteil der Stadtbevölkerung vs.
Lebenserwartung der Männer (1992)

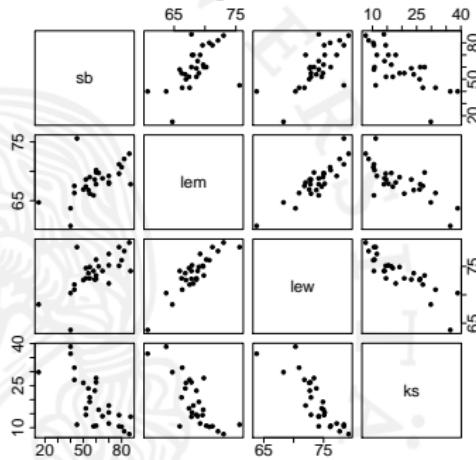


Streudiagramme für Multivariate Daten

3D-Streudiagramm

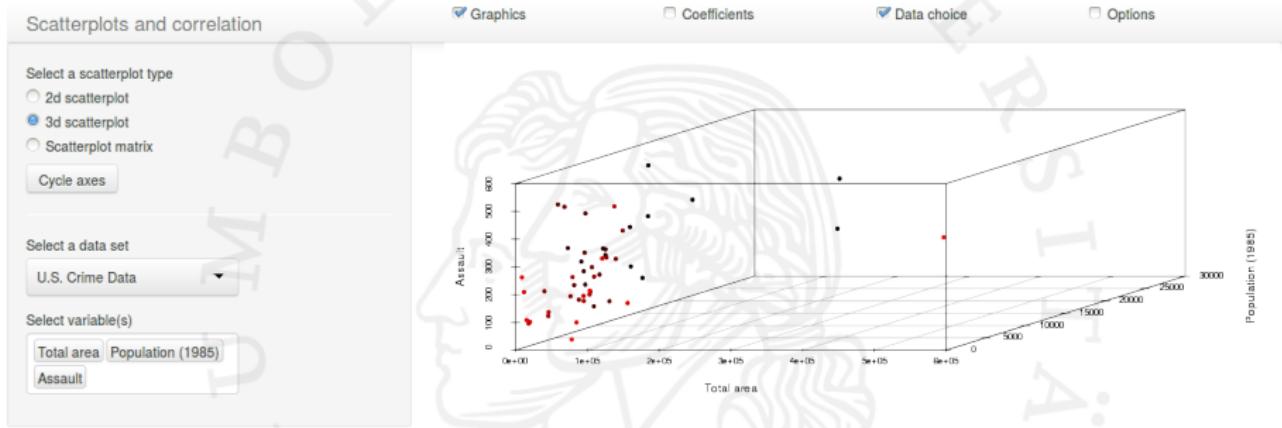


Streudiagramm-Matrix



Beispiel 6.4 (Europa Daten)

Variablen: Anteil der Stadtbewölkerung (sb), Mittlere Lebenserwartung der Männer (lem) und Frauen (lew), Kindersterblichkeit bei 1000 Geburten (ks) in 28 europäischen Ländern.



http://u.hu-berlin.de/men_corr

R Listing 6.1: `example_s3d.R`

```
1 #install.packages("scatterplot3d")
2 library("MASS")
3 library("scatterplot3d")
4 scatterplot3d(Boston$crim, Boston$black, Boston$indus)
```

Gemeinsame Verteilung

Variable X	Variable Y				Randverteilung X	
	y_1	...	y_j	...	y_r	
x_1	h_{11}	...	h_{1j}	...	h_{1r}	$h_{1\bullet}$
\vdots	\vdots	...	\vdots	...	\vdots	\vdots
x_i	h_{i1}	...	h_{ij}	...	h_{ir}	$h_{i\bullet}$
\vdots	\vdots	...	\vdots	...	\vdots	\vdots
x_m	h_{m1}	...	h_{mj}	...	h_{mr}	$h_{m\bullet}$
Randverteilung Y	$h_{\bullet 1}$...	$h_{\bullet j}$...	$h_{\bullet r}$	$h_{\bullet\bullet} = n$

Beispiel 6.5

Unterschiedliche gemeinsame Verteilungen, aber gleiche Randverteilungen

	y_1	y_2	y_3	y_4	
x_1	3	4	2	1	10
x_2	4	3	1	2	10
x_3	1	2	4	3	10
x_4	2	1	3	4	10
	10	10	10	10	40

	y_1	y_2	y_3	y_4	
x_1	4	3	1	2	10
x_2	3	4	2	1	10
x_3	1	2	3	4	10
x_4	2	1	4	3	10
	10	10	10	10	40

Randverteilung

- für eine empirische zweidimensionale Häufigkeitsverteilung
- auch “marginale Verteilung”

$$h_{i\bullet} = \sum_{j=1}^r h_{ij} \quad f_{i\bullet} = \sum_{j=1}^r f_{ij} \quad i = 1, \dots, m$$

$$h_{\bullet j} = \sum_{i=1}^m h_{ij} \quad f_{\bullet j} = \sum_{i=1}^m f_{ij} \quad j = 1, \dots, r$$

Eigenschaften:

$$\sum_{i=1}^m h_{i\bullet} = \sum_{j=1}^r h_{\bullet j} = n$$

$$\sum_{i=1}^m f_{i\bullet} = \sum_{j=1}^r f_{\bullet j} = 1$$

Randverteilung Mittelwert

$$\begin{aligned}\bar{x} &= \frac{1}{n} \sum_{i=1}^m \sum_{j=1}^r x_i \cdot h(x_i, y_j) & \bar{y} &= \frac{1}{n} \sum_{j=1}^r \sum_{i=1}^m y_j \cdot h(x_i, y_j) \\ &= \sum_{i=1}^m \sum_{j=1}^r x_i \cdot f(x_i, y_j) & &= \sum_{j=1}^r \sum_{i=1}^m y_j \cdot f(x_i, y_j) \\ &= \sum_{i=1}^m x_i \cdot f(x_i) & &= \sum_{j=1}^r y_j \cdot f(y_j)\end{aligned}$$

Beispiel 6.6 (Sport)

Angenommen der zeitliche Aufwand für die Variable Y (Sportliche Betätigung) setzt sich wie folgt zusammen:

Sportliche Betätigung (Y)	kaum	gelegentlich	regelmäßig
y_j	y_1	y_2	y_3
Aufwand in Stunden	$0 - 2$	$2 - 4$	$4 - 6$
$h_{\bullet j}$	507	279	214
$f(y_j)$	0,507	0,279	0,214

Mittelwertberechnung der empirischen Randverteilung von Y :

$$\begin{aligned}
 \bar{y} &= \sum_{j=1}^r y_j \cdot f(y_j) \\
 &= (1 \cdot 0,507 + 3 \cdot 0,279 + 5 \cdot 0,214) \\
 &= 2,414
 \end{aligned}$$

Randverteilung Varianz

$$\begin{aligned}s_x^2 &= \frac{1}{n} \sum_{i=1}^m \sum_{j=1}^r (x_i - \bar{x})^2 h(x_i, y_j) = \sum_{i=1}^m \sum_{j=1}^r (x_i - \bar{x})^2 f(x_i, y_j) \\&= \sum_{i=1}^m (x_i - \bar{x})^2 f(x_i)\end{aligned}$$

$$\begin{aligned}s_y^2 &= \frac{1}{n} \sum_{j=1}^r \sum_{i=1}^m (y_j - \bar{y})^2 h(x_i, y_j) = \sum_{j=1}^r \sum_{i=1}^m (y_j - \bar{y})^2 f(x_i, y_j) \\&= \sum_{j=1}^r (y_j - \bar{y})^2 f(y_j)\end{aligned}$$

Bedingte empirische Verteilungen

Bedingte Verteilung (der relativen Häufigkeiten)

- von X für gegebenes $Y = y_j$

$$f(x_i | Y = y_j) = f(x_i | y_j) = \frac{f_{ij}}{f_{\bullet j}} = \frac{h_{ij}}{h_{\bullet j}}$$

- von Y für gegebenes $X = x_i$

$$f(y_j | X = x_i) = f(y_j | x_i) = \frac{f_{ij}}{f_{i \bullet}} = \frac{h_{ij}}{h_{i \bullet}}$$

Beispiel 6.7 (Sport)

Bedingte Verteilung der Variablen Y (sportliche Betätigung) für gegebene x_i (Berufsgruppe) bei 1000 berufstätigen Personen

Berufsgruppe (X)	sportliche Betätigung (Y)			
	kaum	gelegentlich	regelmäßig	
Arbeiter	$0,56 = \frac{240}{430}$	$0,28 = \frac{120}{430}$	$0,16 = \frac{70}{430}$	1,00
Angestellter	$0,47 = \frac{160}{340}$	$0,26 = \frac{90}{340}$	$0,26 = \frac{90}{340}$	1,00
Beamter	$0,33 = \frac{30}{90}$	$0,33 = \frac{30}{90}$	$0,33 = \frac{30}{90}$	1,00
Landwirt	$0,74 = \frac{37}{50}$	$0,14 = \frac{7}{50}$	$0,12 = \frac{6}{50}$	1,00
sonst. freier Beruf	$0,44 = \frac{40}{90}$	$0,36 = \frac{32}{90}$	$0,20 = \frac{18}{90}$	1,00

Beispiel 6.8 (Sport)

Bedingte Verteilung der Variablen X (Berufsgruppe) für gegebene y_j (sportliche Betätigung) bei 1000 berufstätigen Personen

Berufsgruppe (X)	sportliche Betätigung (Y)		
	kaum	gelegentlich	regelmäßig
Arbeiter	$0,47 = \frac{240}{507}$	$0,43 = \frac{120}{279}$	$0,33 = \frac{70}{214}$
Angestellter	$0,32 = \frac{160}{507}$	$0,32 = \frac{90}{279}$	$0,42 = \frac{90}{214}$
Beamter	$0,06 = \frac{30}{507}$	$0,11 = \frac{30}{279}$	$0,14 = \frac{30}{214}$
Landwirt	$0,07 = \frac{37}{507}$	$0,03 = \frac{7}{279}$	$0,03 = \frac{6}{214}$
sonst. freier Beruf	$0,08 = \frac{40}{507}$	$0,11 = \frac{32}{279}$	$0,08 = \frac{18}{214}$
	1,00	1,00	1,00

Beispiel 6.9 (HIV-Infektion)

Bedingte Verteilung

- der Variablen X für gegebene y_j bei 100000 Personen.

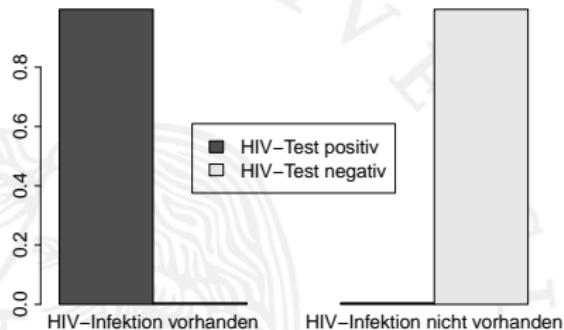
HIV-Test (X)	HIV-Infektion (Y)	
	vorhanden (y_1)	nicht vorhanden (y_2)
positiv (x_1)	0,995	0,005
negativ (x_2)	0,005	0,995
	1,000	1,000

- der Variablen Y für gegebene x_i bei 100000 Personen.

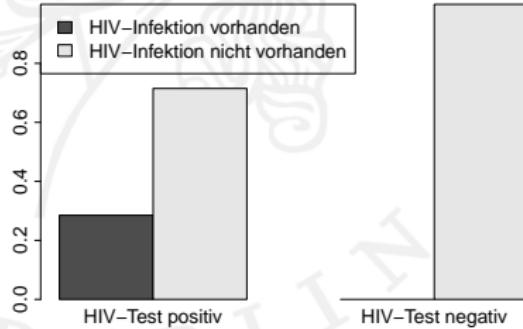
HIV-Test (X)	HIV-Infektion (Y)		
	vorhanden (y_1)	nicht vorhanden (y_2)	
positiv (x_1)	0,289	0,711	1,000
negativ (x_2)	0,001	0,999	1,000

Gruppiertes Balkendiagramm

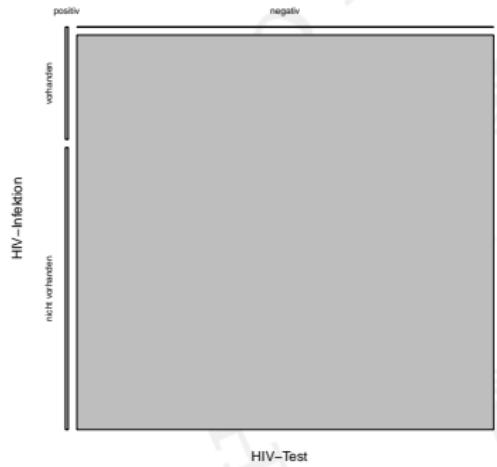
bedingte Verteilung von
HIV-Test gegeben
HIV-Infektion



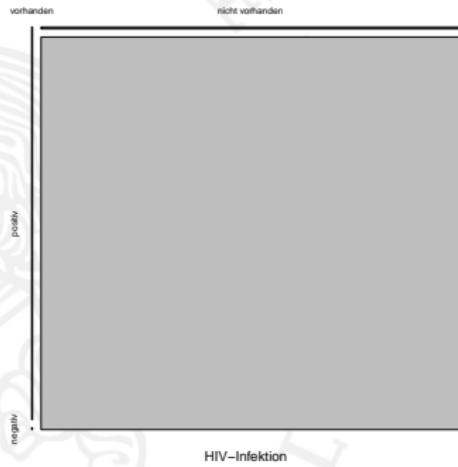
bedingte Verteilung von
HIV-Infektion gegeben
HIV-Test-Ergebnis



Mosaikplot



bedingte Verteilung von HIV-Infektion
gegeben HIV-Test-Ergebnis



bedingte Verteilung von HIV-Test
gegeben HIV-Infektion

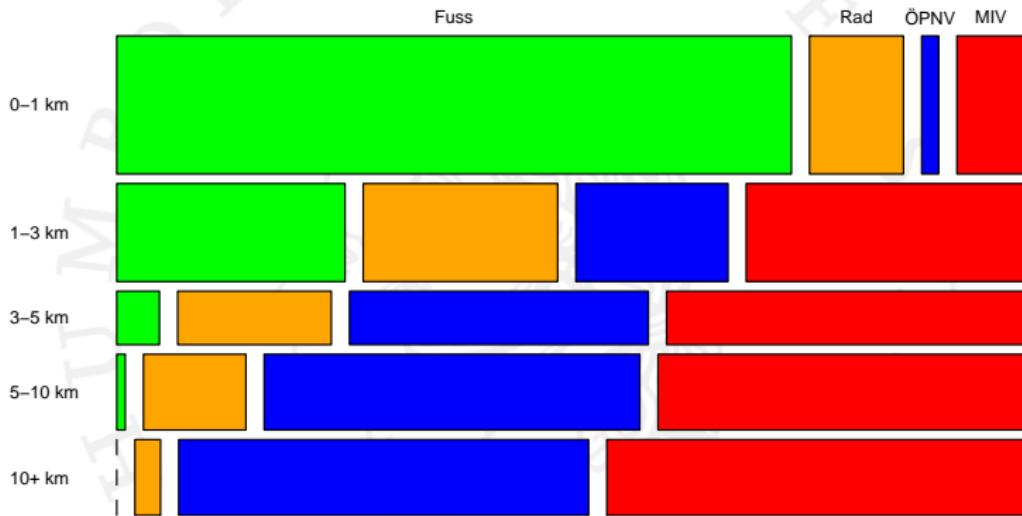
Beispiel 6.10 (Weganteil nach Entfernungsklassen und Verkehrsmittel)

Aus dem Mobilitätsreport 2013 der Berliner Senats:

Weglänge	Fuss	Rad	ÖPNV	MIV	Summe
0-1	0,79	0,11	0,02	0,08	1,00
1-3	0,27	0,23	0,18	0,33	1,00
3-5	0,05	0,18	0,35	0,42	1,00
5-10	0,01	0,12	0,44	0,43	1,00
10+	0,00	0,03	0,48	0,49	1,00

In den Spalten stimmen die Prozentzahlen nicht überein \Rightarrow die Wahl des Verkehrsmittel und die Weglänge hängen zusammen.

Mosaikplot von Weglängen und Verkehrsmittel



- Fläche: proportional zur gemeinsamen Häufigkeit
- X (Breite): proportional zur bed. Häufigkeit des Verkehrsmittels
- Y (Höhe): proportional zur Randhäufigkeit der Weglänge



Listing 6.2: example_barplot2.R

```
1 library("MASS")
2 tab <- table(Boston$chas, Boston$rad)
3 tab
4 # Gestapeltes Balkendiagramm
5 barplot(tab, legend=TRUE)
6 # Gruppiertes Balkendiagramm
7 barplot(tab, beside=TRUE, legend=TRUE)
8 # Bedingte Verteilung bzgl. CHAS
9 ctab <- prop.table(tab, 1)
10 ctab
11 barplot(ctab, beside=TRUE, legend=TRUE)
12 # Bedingte Verteilung bzgl. RAD
13 ctab <- prop.table(tab, 2)
14 ctab
15 barplot(ctab, legend=TRUE)
```

Parameter bedingter Verteilungen

$$\bar{x}|y_j = \sum_{i=1}^m x_i \cdot f(x_i|y_j)$$

$$\bar{y}|x_i = \sum_{j=1}^r y_j \cdot f(y_j|x_i)$$

$$s_{x|y_j}^2 = \sum_{i=1}^m (x_i - \bar{x})^2 f(x_i|y_j)$$

$$s_{y|x_i}^2 = \sum_{j=1}^r (y_j - \bar{y})^2 f(y_j|x_i)$$

Beispiel 6.11 (Sport)

Zeitlicher Aufwand für die Variable Y (Sportliche Betätigung):

Sportliche Betätigung (Y)	kaum	gelegentlich	regelmäßig
y_j	y_1	y_2	y_3
Aufwand in Stunden	0 – 2	2 – 4	4 – 6
Beamter $f(y_j x_3)$	0,33	0,33	0,33
Landwirt $f(y_j x_4)$	0,74	0,14	0,12

Wir möchten berechnen, wie viele Stunden die Beamten(x_3) bzw. Bauern(x_4) im Durchschnitt dem Sport widmen:

$$\bar{y}|x_3 = \sum_{j=1}^r y_j \cdot f(y_j|x_3) = 1 \cdot 0,33 + 3 \cdot 0,33 + 5 \cdot 0,33 = 2,97$$

$$\bar{y}|x_4 = \sum_{j=1}^r y_j \cdot f(y_j|x_4) = 1 \cdot 0,74 + 3 \cdot 0,14 + 5 \cdot 0,12 = 1,76$$

Beispiel 6.12 (Weganteil nach Entfernungsklassen und Verkehrsmittel)

Berechnung der gemeinsamen Häufigkeiten und der Randverteilungen mit $f(x_i, y_j) = f(x_i|y_j) \cdot f(y_j)$ (nicht im Mobilitätsreport enthalten)

Weglänge	Fuss	Rad	ÖPNV	MIV	Summe
0-1	0,25	0,03	0,01	0,03	0,31
1-3	0,06	0,05	0,04	0,07	0,22
3-5	0,01	0,02	0,04	0,05	0,12
5-10	0,00	0,02	0,07	0,07	0,17
10+	0,00	0,01	0,08	0,08	0,17
Summe	0,31	0,13	0,25	0,31	1,00
Modal split	0,31	0,13	0,27	0,30	1,00

Die Ungenauigkeiten kommen aus Rundungen in den ursprünglichen Daten, z.B.

$$0,31 + 0,22 + 0,12 + 0,17 + 0,17 = 0,99 \neq 1,00.$$

Berechnung der empirischen Verteilungen mit $f(y_j|x_i) = f(x_i, y_j)/f(x_i)$
 (nicht im Mobilitätsreport enthalten)

Weglänge	Fuss	Rad	ÖPNV	MIV	Repräsentant
0-1	0,79	0,26	0,03	0,08	0,5
1-3	0,19	0,38	0,16	0,24	2,0
3-5	0,02	0,16	0,17	0,17	4,0
5-10	0,01	0,16	0,31	0,24	7,5
10+	0,00	0,04	0,33	0,27	21,1
Summe	1,00	1,00	1,00	1,00	

$$\begin{aligned}\bar{y}|\text{Fuss} &= 0,5 \cdot 0,78 + 2,0 \cdot 0,19 + 4,0 \cdot 0,02 + 7,5 \cdot 0,01 + 21,1 \cdot 0,00 = & 0,89 \text{ km} \\ \bar{y}|\text{Rad} &= 0,5 \cdot 0,26 + 2,0 \cdot 0,38 + 4,0 \cdot 0,16 + 7,5 \cdot 0,15 + 21,1 \cdot 0,04 = & 3,54 \text{ km} \\ \bar{y}|\text{ÖPNV} &= 0,5 \cdot 0,03 + 2,0 \cdot 0,16 + 4,0 \cdot 0,17 + 7,5 \cdot 0,31 + 21,1 \cdot 0,33 = & 10,39 \text{ km} \\ \bar{y}|\text{MIV} &= 0,5 \cdot 0,08 + 2,0 \cdot 0,24 + 4,0 \cdot 0,17 + 7,5 \cdot 0,24 + 21,1 \cdot 0,27 = & 8,78 \text{ km} \\ && \bar{y} = & 6 \text{ km}\end{aligned}$$

Parameter bivariater Verteilungen

5. November 2022

- Empirische Unabhängigkeit • Für nominalskalierte Variablen • Empirische Kovarianz • Empirischer Korrelationskoeffizient • Für ordinalskalierte Variablen • Zusammenfassung • Kovarianz unter Unabhängigkeit • Bravais–Pearson–KK • Spearmanschen RangKK

Empirische Unabhängigkeit

Zwei Variablen X und Y heißen unabhängig, wenn man für eine Beobachtung

- aus der Kenntnis der Merkmalsausprägung x_k keinen Rückschluss auf die Merkmalsausprägung y_k ziehen kann
- und umgekehrt!

Beispiel 7.1 (HIV-Infektion)

- **Frage:** Können die beiden Variablen X : "Test auf HIV" und Y : "HIV Infektion" unabhängig sein?
- **Antwort:** Wenn es ein sinnvoller Test ist nicht. Wünschenswert wäre:
 - ▶ HIV Infektion vorhanden \Rightarrow Test positiv
 - ▶ HIV Infektion nicht vorhanden \Rightarrow Test negativ
 - ▶ Test positiv \Rightarrow HIV Infektion vorhanden
 - ▶ Test negativ \Rightarrow keine HIV Infektion vorhanden

- Bedingte Verteilung von X gegeben y_i

HIV-Test (X)	HIV-Infektion (Y)		Randvert. von X
	vorhanden (y_1)	nicht vorhanden (y_2)	
positiv (x_1)	0,995	0,005	0,007
negativ (x_2)	0,005	0,995	0,993
	1,000	1,000	1,000

- ▶ HIV-Infektion vorhanden \Rightarrow Test meist positiv
- ▶ HIV-Infektion nicht vorhanden \Rightarrow Test meist negativ

- Bedingte Verteilung von Y gegeben x_j

HIV-Test (X)	HIV-Infektion (Y)		
	vorhanden (y_1)	nicht vorhanden (y_2)	
positiv (x_1)	0,289	0,711	1,000
negativ (x_2)	0,001	0,999	1,000
Randvert. von Y	0,002	0,998	1,000

- ▶ Test positiv \Rightarrow in 28,9% der Fälle eine HIV Infektion
- ▶ Test negativ \Rightarrow in nur 0,1% der Fälle eine HIV Infektion

- Die Variablen X und Y sind unabhängig, wenn gilt:

$$f(x_i|y_j) = f(x_i|y_l) = f(x_i)$$

für alle $j, l = 1, \dots, r$ und für alle $i = 1, \dots, m$

- Die bedingten Verteilungen von Y , gegeben X , stimmen untereinander und mit der Randverteilung von Y überein.

$$f(y_j|x_i) = f(y_j|x_h) = f(y_j)$$

für alle $i, h = 1, \dots, m$ und für alle $j = 1, \dots, r$

$$f(x_i|y_j) \stackrel{\text{unabh.}}{=} f(x_i) = \frac{f(x_i, y_j)}{f(y_j)} \Rightarrow f(x_i, y_j) = f(x_i)f(y_j)$$

$$f(y_j|x_i) \stackrel{\text{unabh.}}{=} f(y_j) = \frac{f(x_i, y_j)}{f(x_i)} \Rightarrow f(x_i, y_j) = f(x_i)f(y_j)$$

- analog für absolute Häufigkeiten:

$$h(x_i, y_j) = \frac{h(x_i)h(y_j)}{n}$$

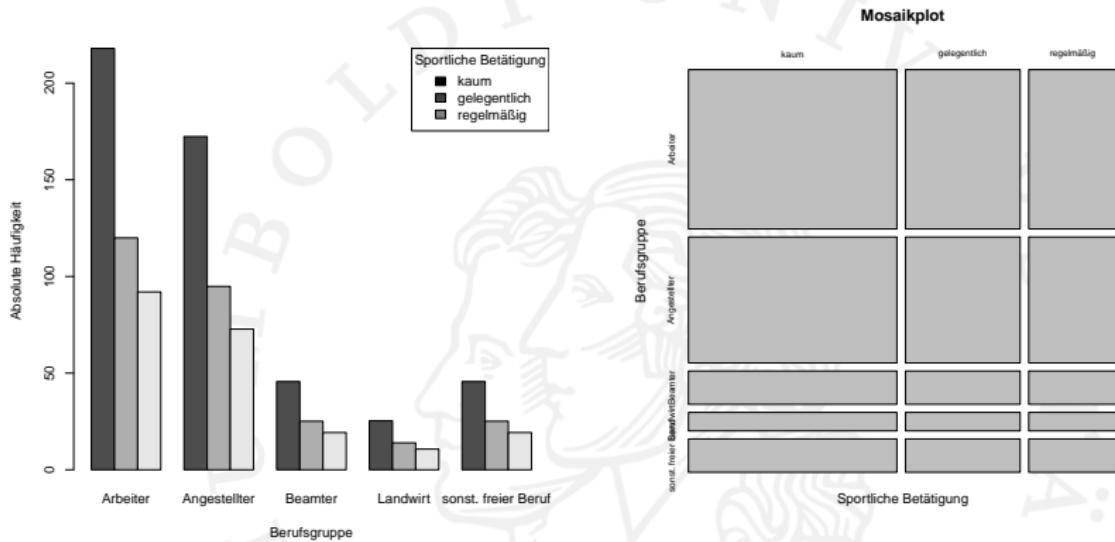
Prüfung der Unabhängigkeit – Vergleich der beobachteten relativen Häufigkeiten $f(x_i, y_j)$ mit theoretischen (erwarteten) Häufigkeiten $e_{i,j} = f(x_i) \cdot f(y_j)$

Variable X	Variable Y				Randverteilung X
	y_1	...	y_j	...	
x_1	$f(x_1, y_1) \stackrel{?}{=} e_{1,1}$...	$f(x_1, y_j) \stackrel{?}{=} e_{1,j}$...	$f(x_1)$
\vdots	\vdots	...	\vdots	...	\vdots
x_i	$f(x_i, y_1) \stackrel{?}{=} e_{i,1}$...	$f(x_i, y_j) \stackrel{?}{=} e_{i,j}$...	$f(x_i)$
\vdots	\vdots	...	\vdots	...	\vdots
Randverteilung Y	$f(y_1)$...	$f(y_j)$...	1

Beispiel 7.2 (HIV-Infektion)

HIV-Test (X)	HIV Infektion (Y)		Randverteilung X
	vorhanden (y_1)	nicht vorhanden(y_2)	
positiv (x_1)	0,001990	0,004990	0,006980
$e_{i,j}$	0,000014	0,006966	
negativ (x_2)	0,000010	0,993010	0,993020
$e_{i,j}$	0,001986	0,991034	
Randverteilung Y	0,002000	0,998000	1

- $e_{1,1} = f(x_1) \cdot f(y_1) = 0,006980 \cdot 0,00200 \approx 0,000014$
- Die beiden Variablen HIV-Test und HIV-Infektion sind nicht unabhängig.



- Gruppiertes Balkendiagramm: Bei Unabhängigkeit sehen die Balkendiagramme in jeder Teilgruppe ähnlich aus
- Mosaikplot: Bei Unabhängigkeit sieht man ein Schachbrettmuster

Für nominalskalierte Variablen

Kontingenz

Beobachtete Häufigkeit:

$$h_{ij} = h(x_i, y_j)$$

$$f_{ij} = f(x_i, y_j) = h(x_i, y_j)/n$$

Erwartete Häufigkeit
bei Unabhängigkeit:

$$e_{ij} = \frac{h_{i\bullet} h_{\bullet j}}{n} = n f_{i\bullet} f_{\bullet j}$$

$$e_{ij}/n = f_{i\bullet} f_{\bullet j}$$

Quadratische Kontingenz:

$$K^2 = \sum_{i=1}^m \sum_{j=1}^r \frac{\left(h_{ij} - \frac{h_{i\bullet} h_{\bullet j}}{n} \right)^2}{\frac{h_{i\bullet} h_{\bullet j}}{n}} = n \sum_{i=1}^m \sum_{j=1}^r \frac{(f_{ij} - f_{i\bullet} f_{\bullet j})^2}{f_{i\bullet} f_{\bullet j}}$$

Eigenschaften:

- $K^2 \geq 0$
- $K^2 = 0$, wenn $h_{ij} = e_{ij}$ für alle i und j .

Kontingenzkoeffizient

$$C = \sqrt{\frac{K^2}{n + K^2}}$$

$$0 \leq C \leq \sqrt{\frac{C^* - 1}{C^*}} < 1, \quad C^* = \min\{m, r\}$$

Korrigierter Kontingenzkoeffizient

$$C_{korr} = C \cdot \sqrt{\frac{C^*}{C^* - 1}}$$

$$0 \leq C_{korr} \leq 1$$

Beispiel 7.3 (HIV-Infektion)

HIV-Test (X)	HIV Infektion (Y)		Randverteilung X
	vorhanden (y_1)	nicht vorhanden (y_2)	
positiv (x_1)	0,00199	0,00499	0,00698
negativ (x_2)	0,00001	0,99301	0,99302
Randverteilung Y	0,00200	0,99800	1

$$\begin{aligned}
 K^2 &= 100000 \\
 &\cdot \left[\frac{(0,00199 - 0,00200 \cdot 0,00698)^2}{0,00200 \cdot 0,00698} + \frac{(0,00499 - 0,99800 \cdot 0,00698)^2}{0,99800 \cdot 0,00698} \right. \\
 &+ \left. \frac{(0,00001 - 0,00200 \cdot 0,99302)^2}{0,00200 \cdot 0,99302} + \frac{(0,99301 - 0,99800 \cdot 0,99302)^2}{0,99800 \cdot 0,99302} \right] \\
 &= 28223,93 \\
 C &= \sqrt{\frac{28223,93}{100000 + 28223,93}} = 0,47; \quad C_{\text{corr}} = 0,47 \cdot \sqrt{\frac{2}{2-1}} = 0,66
 \end{aligned}$$

Beispiel 7.4 (Wirtschaft)

X1 – Wie beurteilen Sie die heutige wirtschaftliche Lage in Deutschland?

Y – Erhebungsgebiet → nominalskaliert

Ausprägungen:

alte Bundesländer (West)

neue Bundesländer (Ost)

Besteht ein Zusammenhang zwischen der Einschätzung der Wirtschaftslage für die Bundesrepublik und dem Erhebungsgebiet? → Kontingenz

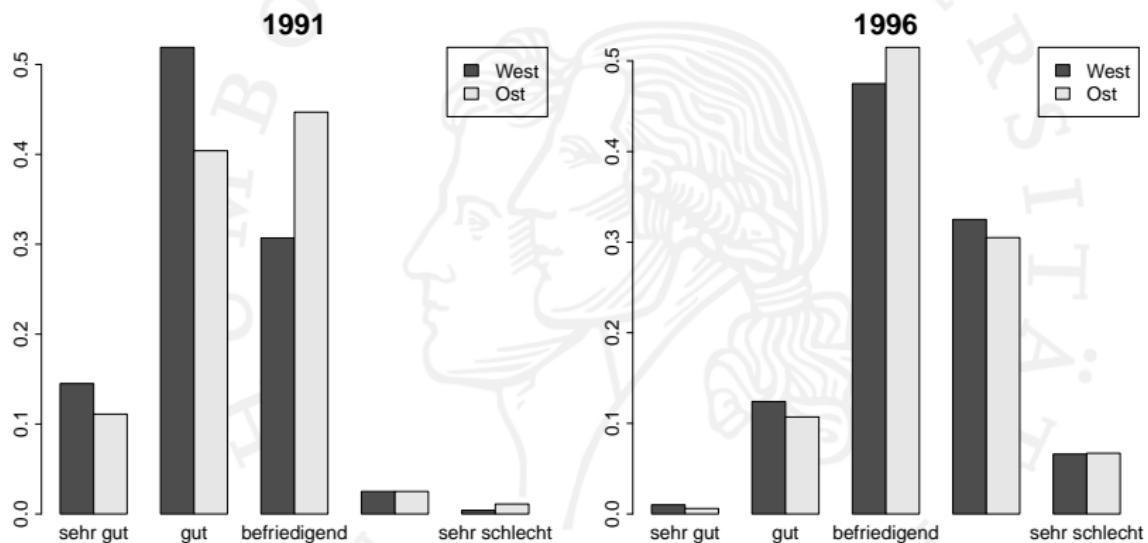
Kontingenztabelle, $n = 1000$

Einschätzung X1	1991		1996		RV X1
	Erhebungsgebiet Y West	Ost	Erhebungsgebiet Y West	Ost	
sehr gut	72	56	128	6	2
gut	256	205	461	82	36
teils teils	152	226	378	314	175
schlecht	12	14	26	215	104
sehr schlecht	2	5	7	44	22
RV Y	494	506	1000	661	339
					1000

Bedingte Verteilung:

Einschätzung X1	$f(x_{1i} y_j)$ 1991		$f(x_{1i} y_j)$ 1996		RV X1
	Erhebungsgebiet Y West	Ost	Erhebungsgebiet Y West	Ost	
sehr gut	0,145	0,111	0,128	0,010	0,006
gut	0,519	0,404	0,461	0,124	0,107
teils teils	0,307	0,447	0,378	0,475	0,515
schlecht	0,025	0,027	0,026	0,325	0,305
sehr schlecht	0,004	0,011	0,007	0,066	0,067
	1,000	1,000	1,000	1,000	1,000

Bedingte Verteilungen $f(x_1;|y_j)$



- X2 Wie beurteilen Sie Ihre eigene gegenwärtige wirtschaftliche Lage?
Y Erhebungsgebiet

1991: Kontingenztabelle, $n = 1000$

Einschätzung X2	Erhebungsgebiet Y		RV X2
	West	Ost	
sehr gut	30	6	36
gut	306	173	479
teils teils	119	227	346
schlecht	29	81	110
sehr schlecht	7	22	29
RV Y	491	509	1000

$$\begin{aligned} K^2 &= \frac{\left(30 - \frac{491 \cdot 36}{1000}\right)^2}{\frac{491 \cdot 36}{1000}} + \frac{\left(306 - \frac{491 \cdot 479}{1000}\right)^2}{\frac{491 \cdot 479}{1000}} + \frac{\left(119 - \frac{491 \cdot 346}{1000}\right)^2}{\frac{491 \cdot 346}{1000}} + \frac{\left(29 - \frac{491 \cdot 110}{1000}\right)^2}{\frac{491 \cdot 110}{1000}} \\ &\quad + \frac{\left(7 - \frac{491 \cdot 29}{1000}\right)^2}{\frac{491 \cdot 29}{1000}} + \frac{\left(6 - \frac{509 \cdot 36}{1000}\right)^2}{\frac{509 \cdot 36}{1000}} + \frac{\left(173 - \frac{509 \cdot 479}{1000}\right)^2}{\frac{509 \cdot 479}{1000}} + \frac{\left(227 - \frac{509 \cdot 346}{1000}\right)^2}{\frac{509 \cdot 346}{1000}} \\ &\quad + \frac{\left(81 - \frac{509 \cdot 110}{1000}\right)^2}{\frac{509 \cdot 110}{1000}} + \frac{\left(22 - \frac{509 \cdot 29}{1000}\right)^2}{\frac{509 \cdot 29}{1000}} = 118,83 \end{aligned}$$

$$C = \sqrt{\frac{118,83}{1000 + 118,83}} = 0,3259$$

$$C_{korr} = 0,3259 \sqrt{\frac{2}{2-1}} = 0,4609$$

Y – Erhebungsgebiet

X_1 – gegenwärtige Wirtschaftslage
in der Bundesrepublik

X_2 – eigene gegenwärtige
Wirtschaftslage

X_3 – zukünftige Wirtschaftslage in
der Bundesrepublik

X_4 – eigene zukünftige
Wirtschaftslage

	1991	1996
X_1	$C = 0,154$	$C = 0,044$
X_2	$C = 0,325$	$C = 0,116$
X_3	$C = 0,293$	$C = 0,071$
X_4	$C = 0,300$	$C = 0,061$

- deutlich geringere Assoziation 1996 gegenüber 1991
- Angleichung der Auffassungen zwischen West und Ost

Beispiel 7.5 (Weganteil nach Entfernungsklassen und Verkehrsmittel)

Beobachtete und erwartete Häufigkeiten unter Unabhängigkeit
(nicht im Mobilitätsreport enthalten)

Weglänge	Beobachtete Häufigkeiten				Summe	Erwartete Häufigkeiten			
	Fuss	Rad	ÖPNV	MIV		Fuss	Rad	ÖPNV	MIV
0-1	0,24	0,03	0,01	0,02	0,31	0,10	0,04	0,08	0,10
1-3	0,06	0,05	0,04	0,07	0,22	0,07	0,03	0,05	0,07
3-5	0,01	0,02	0,04	0,05	0,12	0,04	0,02	0,03	0,04
5-10	0,00	0,02	0,08	0,07	0,17	0,05	0,02	0,04	0,05
10+	0,00	0,01	0,08	0,08	0,17	0,05	0,02	0,04	0,05
Summe	0,31	0,13	0,25	0,31	1,00	0,31	0,13	0,25	0,31

$$n \approx 50.000 \text{ Wege}$$

$$K^2 = n \cdot \left(\frac{(0,24-0,10)^2}{0,10} + \dots + \frac{(0,08-0,05)^2}{0,05} \right) \approx 30.725 \quad C = 0,62 \quad C^* = 0,71$$

Crosstable

Select a data set: Hair and Eye Color of Statistics Students

Select column variable: Hair color

Select row variable: Eye color

Coefficients Data choice Options

Columns: Hair color

	Black	Brown	Red	Blond
Brown	68	119	26	7
Blue	20	84	17	94
Hazel	15	54	14	10
Green	5	29	14	16

Rows: Eye color

http://u.hu-berlin.de/men_asso

 Listing 7.1: example_cornom.R

```
1 HIV <- as.table(cbind(vorhanden=c(995,5),  
2  
3 # Kontingenztabelle mit Randverteilungen  
4 addmargins(HIV)  
5 # chi-quadrat Test, speichern der Ergebnisse  
6 kont <- chisq.test(HIV)  
7 # Erwartete Häufigkeiten  
8 kont$expected  
9 # Quadratische Kontingenz  
10 qk <- kont$statistic  
11 qk  
12 # Kontingenzkoeffizient  
13 kk <- sqrt(qk/(qk+sum(HIV)))  
14 kk  
15 # Korrigierter Kontingenzkoeffizient  
16 cs <- min(nrow(HIV), ncol(HIV))  
17 KK <- kk*sqrt(cs/(cs-1))  
18 KK
```

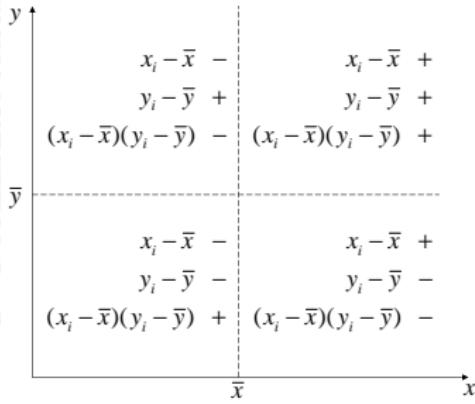
nicht

Empirische Kovarianz

Die Kovarianz ist ein Parameter für die gemeinsame Streuung zweier metrisch skalierter Variablen.

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

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



Vorzeichenwechsel der Kovarianz in Abhängigkeit der Vorzeichen von $(x_i - \bar{x})$ und $(y_i - \bar{y})$.

Empirischer Korrelationskoeffizient

Zentrierung

$$x_i^* = x_i - \bar{x}, \quad y_i^* = y_i - \bar{y}, \quad \sum x_i^* y_i^* = \sum (x_i - \bar{x})(y_i - \bar{y})$$

- ungeeignetes Maß für den Zusammenhang, da abhängig
 - ▶ von den Maßeinheiten der Variablen
 - ▶ und von der Anzahl n der statistischen Einheiten

Standardisierung

$$u_i = \frac{(x_i - \bar{x})}{s_x} \text{ mit } \bar{u} = 0, s_u^2 = 1; \quad v_i = \frac{(y_i - \bar{y})}{s_y} \text{ mit } \bar{v} = 0, s_v^2 = 1$$

$$\sum_{i=1}^n \frac{(x_i - \bar{x})}{s_x} \frac{(y_i - \bar{y})}{s_y} = \sum_{i=1}^n u_i \cdot v_i$$

- Durch die Standardisierung werden die Daten skalenunabhängig.
- Skalenunabhängige Größen sind besser zu interpretieren.

Bravais–Pearson–Korrelationskoeffizient

$$\begin{aligned}
 r_{xy} = r_{yx} &= \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{s_x \cdot s_y} = \frac{s_{xy}}{s_x \cdot s_y} \\
 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}
 \end{aligned}$$

Eigenschaften:

- $r_{yx} = r_{xy}$
- $-1 \leq r_{xy} \leq +1$

Beispiel 7.6 (Jahresmiete und Jahresgewinn)

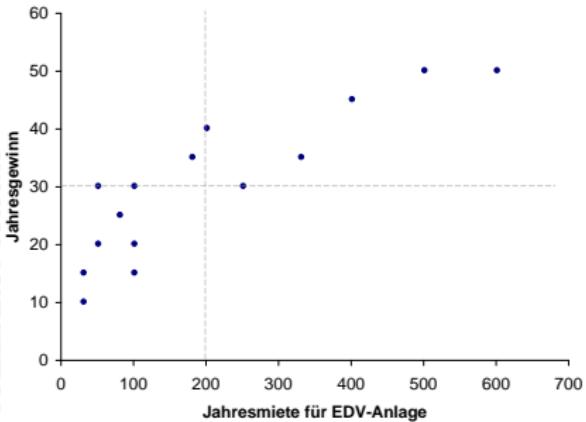
- Variable X – Jahresmiete für EDV-Anlage in 1000 Euro
- Variable Y – Jahresgewinn in Mio. Euro

$$\bar{x} = 200, \sum_{i=1}^{15} (x_i - \bar{x})^2 = 457\,000$$

$$\bar{y} = 30, \sum_{i=1}^{15} (y_i - \bar{y})^2 = 2\,250$$

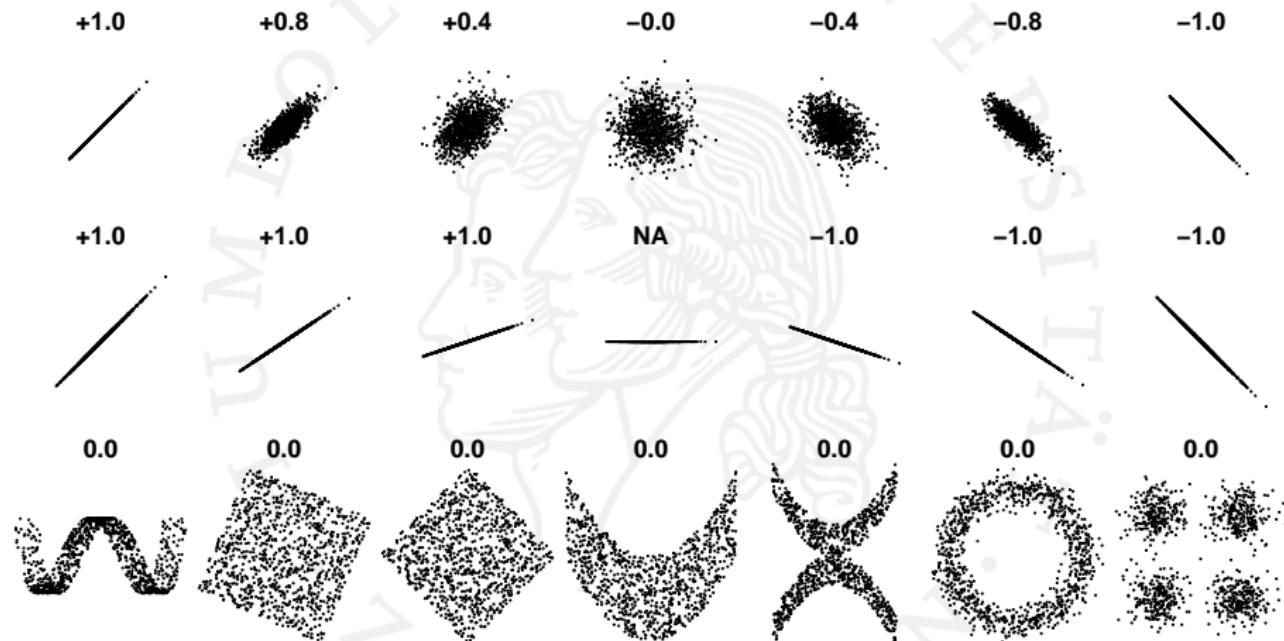
$$\sum_{i=1}^{15} (x_i - \bar{x})(y_i - \bar{y}) = 28\,100$$

$$r = \frac{28\,100}{\sqrt{457\,000 \cdot 2\,250}} = 0,8763$$



- $r = 0,8763$ - starke positive Korrelation
- Wenn X steigt, dann erhöht sich auch Y und umgekehrt.

Zusammenhang von Korrelation und Punktewolke im Streudiagramm



Quelle: [Wikimedia Commons](#)



Listing 7.2: example_cormet.R

```
1 library("MASS")
2 # Kovarianz
3 cov(Boston$crim, Boston$tax)
4 # Korrelation
5 cor(Boston$crim, Boston$tax)
```

Für ordinalskalierte Variablen

Zwei ordinalskalierten Variablen X und Y

- Bei ordinalskalierten Variablen: Bildung von arithmetischen Mittel nicht legitim.
- Man kann jedoch Beobachtungswerten eine Rangordnung bzw. eine Rangzahl zuweisen.
- Korrelation bei ordinalskalierten Variablen bzw. Rangkorrelation wird anhand der Rangzahlen berechnet.
- Rangzahlen: $R(x_i)$, $R(y_i)$, $i = 1, \dots, n$
 - ▶ Dem kleinsten Beobachtungswert von X wird die Rangzahl 1 zugeordnet, der zweitkleinsten die Rangzahl 2 usw.
 - ▶ Entsprechend verfährt man auch bei der Zuordnung der Rangzahlen der Beobachtungswerte von Y

Spearmanscher Rangkorrelationskoeffizient

= Bravais-Pearson-Korrelationskoeffizient der Rangzahlen $R(x_i), R(y_i)$

$$\begin{aligned} r_s &= 1 - \frac{6 \sum_{i=1}^n \{R(x_i) - R(y_i)\}^2}{n(n^2 - 1)} \\ &= 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}, \quad d_i = R(x_i) - R(y_i) \end{aligned}$$

Eigenschaft:

$$-1 \leq r_s \leq 1$$

Beispiel 7.7

$n = 6$ Sportler

X – Plazierung des Sportlers in der Abfahrt

Y – Plazierung des Sportlers im Slalom

Sportler (i)	1	2	3	4	5	6
Abfahrt (X)	2	1	3	4	5	6
Slalom (Y)	2	3	1	5	4	6
d_i^2	0	4	4	1	1	0

Besteht ein Zusammenhang zwischen den Plazierungen in beiden Disziplinen?

$$r_s = 1 - \frac{6 \cdot 10}{6(36 - 1)} = 0,7143$$

Beispiel 7.8 (Wirtschaftslage)

X – Wie beurteilen Sie die gegenwärtige wirtschaftliche Lage in der Bundesrepublik?

Y – Wie beurteilen Sie Ihre gegenwärtige eigene wirtschaftliche Lage?

Einschätzungskala:

sehr gut	gut	teils gut/teils schlecht	schlecht	sehr schlecht
1	2	3	4	5

⇒ X, Y ordinalskaliert

Besteht ein Zusammenhang zwischen der Einschätzung der Wirtschaftslage für die Bundesrepublik und der eigenen Wirtschaftslage?

Jahr	n	r_s
1991	2958	0,195
1996	3268	0,306

Kendallscher Rangkorrelationskoeffizient τ

- Vergleich der Ordnungsrelation für alle möglichen Paare von beobachteten Werten zweier Merkmale
- Sortierung der Rangpaare $R(x_i), R(y_i)$ nach $R(x_i)$
- P - konkordante Merkmalspaare: weisen eine gleiche Ordnungsrelation auf, d.h. wenn $x_i < x_j$, dann gilt $y_i < y_j$
- Q - diskordante Merkmalspaare: weisen eine entgegengesetzte Ordnungsrelation auf, d.h. wenn $x_i < x_j$, dann gilt $y_i > y_j$
- p_i - Anzahl der Paare mit $R(x_i) < R(x_j)$ und $R(y_i) < R(y_j)$
- q_i - Anzahl der Paare mit $R(x_i) < R(x_j)$ und $R(y_i) > R(y_j)$

Definition I

$$\tau = \frac{P - Q}{\frac{n(n-1)}{2}} \quad \text{mit} \quad Q = \sum_{i=1}^n q_i, \quad P = \sum_{i=1}^n p_i$$

Definition II

$$\tau = 1 - \frac{4Q}{n(n-1)} = \frac{4P}{n(n-1)} - 1$$

Eigenschaft:

$$-1 \leq \tau \leq 1$$

- Anzahl möglicher Paarvergleiche: $1 + 2 + \dots + (n-1) = \frac{n(n-1)}{2}$

Beispiel 7.9 (Angestellte)

X – organisatorische Fähigkeiten, Y – Arbeitssorgfalt, $n = 10$ Angestellten

Sortierung nach Rang:

Angestellter i	5	9	2	7	6	8	1	10	3	4
$R(x_i)$	1	2	3	4	5	6	7	8	9	10
$R(y_i)$	7	2	9	5	1	4	3	6	10	8

Berechnung von p_i, q_i bzw. P, Q und Einsetzen in die Formel von τ :

Angestellter i	5	9	2	7	6	8	1	10	3	4	\sum
$R(x_i)$	1	2	3	4	5	6	7	8	9	10	
$R(y_i)$	7	2	9	5	1	4	3	6	10	8	
q_i (kleiner)	6	1	6	3	0	1	0	0	1	0	18
p_i (größer)	3	7	1	3	5	3	3	2	0	0	27

$$Q = 18, \quad P = 27, \quad \frac{n(n-1)}{2} = \frac{10 \cdot 9}{2} = 45, \quad \tau = \frac{27 - 18}{45} = \frac{9}{45} = 0,2$$

Beispiel 7.10 (Bundesliga)

- Gibt es eine (gute) Variable um am Anfang der Bundesliga-Saison vorherzusagen wer Meister wird?
- Ja, der Gesamtmarktwert in Mio. EUR (Stichtag: 1.10.2020)

Rang	Verein	Wert	Rang	Verein	Wert
1	FC Bayern München	868,95	10	FC Schalke 04	171,55
2	Borussia Dortmund	587,25	11	SC Freiburg	106,20
3	RB Leipzig	514,58	12	1.FSV Mainz 05	103,95
4	Bayer 04 Leverkusen	323,45	13	SV Werder Bremen	102,50
5	Borussia Mönchengladbach	313,38	14	1.FC Köln	101,83
6	Hertha BSC	234,63	15	FC Augsburg	96,43
7	TSG 1899 Hoffenheim	224,10	16	VfB Stuttgart	79,75
8	VfL Wolfsburg	209,25	17	1.FC Union Berlin	56,03
9	Eintracht Frankfurt	184,90	18	Arminia Bielefeld	47,15

- Der Spearman'sche Rangkorrelationswert zwischen der Rangfolge im Gesamtmarktwert und dem Rang am Saisonende beträgt ca. 0,8



Listing 7.3: example_corord.R

```
1 library("MASS")
2 # Spearmansche Rangkorrelation
3 cor(Boston$crim, Boston$tax, method="spearman")
4 # Kendalls Rangkorrelation
5 cor(Boston$crim, Boston$tax, method="kendall")
```

Zusammenfassung

Parameterverwendung	Skalenniveau					
	X nominal Y nominal	X nominal Y ordinal	X nominal Y metrisch	X ordinal Y ordinal	X ordinal Y metrisch	X metrisch Y metrisch
problemlos	green	yellow	red	yellow	red	red
problembehaftet	yellow	green	red	yellow	red	yellow
auf keinen Fall	red	red	red	green	yellow	green
K^2	yellow	yellow	red	yellow	red	red
C_{korr}	green	yellow	red	green	yellow	yellow
Kendall's τ	red	red	red	green	yellow	yellow
Spearman	red	red	red	yellow	yellow	green
Kovarianz	red	red	red	red	red	yellow
Bravais-Pearson	red	red	red	red	red	green

Generell: Falls die Skalenniveaus der Variablen unterschiedlich sind, dann versucht man einen Koeffizienten eines niedrigeren Skalenniveaus zu benutzen.

Kovarianz unter Unabhängigkeit

Beweis: Variablen X und Y unabhängig $\Rightarrow s_{xy} = 0$

$$\begin{aligned}
 s_{xy} &= \sum_{i=1}^m \sum_{j=1}^r (x_i - \bar{x})(y_j - \bar{y}) f_{ij} \\
 \text{wegen Unabh.} &= \sum_{i=1}^m \sum_{j=1}^r (x_i - \bar{x})(y_j - \bar{y}) f_{i\bullet} f_{\bullet j} \\
 &= \left\{ \sum_{i=1}^m (x_i - \bar{x}) f_{i\bullet} \right\} \left\{ \sum_{j=1}^r (y_j - \bar{y}) f_{\bullet j} \right\} \\
 &= \left\{ \sum_{i=1}^m x_i f_{i\bullet} - \bar{x} \sum_{i=1}^m f_{i\bullet} \right\} \left\{ \sum_{j=1}^r y_j f_{\bullet j} - \bar{y} \sum_{j=1}^r f_{\bullet j} \right\} \\
 &= (\bar{x} - \bar{x})(\bar{y} - \bar{y}) = 0
 \end{aligned}$$

Bravais–Pearson–KK

Möglichkeiten zur Berechnung:

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

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

$$r_{xy} = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{\left\{ n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 \right\} \left\{ n \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i \right)^2 \right\}}}$$

Spearmanschen RangKK

Spearmansche Rangkorrelationskoeffizient =

Bravais-Pearson-Korrelationskoeffizient der Rangzahlen $R(x_i), R(y_i)$

$$\begin{aligned}
 r_s &= \frac{\sum_{i=1}^n (R(x_i) - \overline{R(x)})(R(y_i) - \overline{R(y)})}{\sqrt{\sum_{i=1}^n (R(x_i) - \overline{R(x)})^2} \sqrt{\sum_{i=1}^n (R(y_i) - \overline{R(y)})^2}} \\
 &= 1 - \frac{6 \sum_{i=1}^n \{R(x_i) - R(y_i)\}^2}{n(n^2 - 1)}
 \end{aligned}$$

falls $i \neq j \Rightarrow R(x_i) \neq R(x_j)$

mit $n\overline{R(x)} = \sum_{i=1}^n R(x_i) = \frac{n \cdot (n-1)}{2}$, $\sum_{i=1}^n (R(x_i) - \overline{R(x)})^2 = \frac{(n-1) \cdot n \cdot (n+1)}{12}$
 und für y_i analog



Regressionsanalyse

5. November 2022

- Allgemeines Regressionsproblem
- Residuum
- Einfache lineare Regression
- Schätzung der Regressionsparameter
- Methode der kleinsten Quadrate
- Bestimmtheit (Güte) der Regression
- Bestimmtheitsmaß

Allgemeines Regressionsproblem

Ziel:

Schätzung der (mittleren) statistischen Abhängigkeit einer Variablen Y von Variablen X_1, X_2, \dots, X_p

- Y : abhängige Variable, erklärte Variable, endogene Variable, Regressand, metrisch skaliert
- X_1, X_2, \dots, X_p : unabhängige Variable, erklärende Variable, exogene Variable, Regressoren, metrisch skaliert

⇒ einseitig gerichtete Abhängigkeit

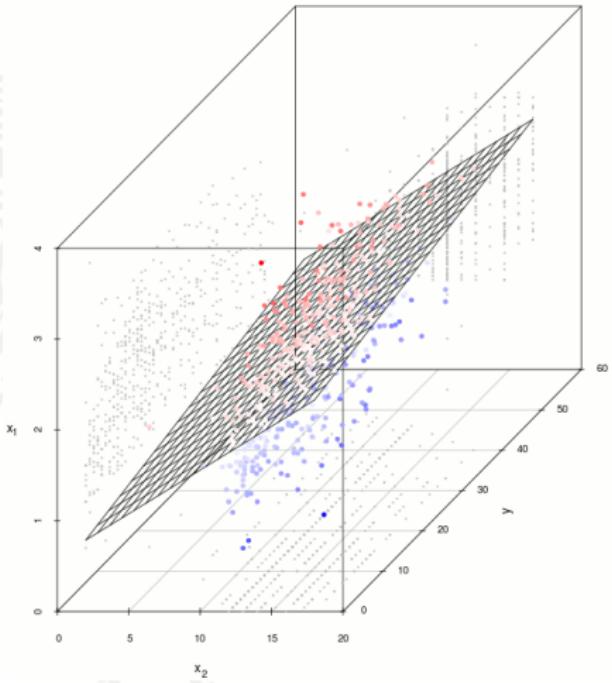
- Variablenwerte, Beobachtungen ($i = 1, \dots, n$):
 $y_i, x_{1i}, x_{2i}, \dots, x_{pi}$

Beispiel 8.1

$$Y = m(X_1, X_2) + \varepsilon$$

- Y : $\log(\text{Stundenlohn in Euro})$
- X_1 : Berufsausbildung in Jahren
- X_2 : Berufserfahrung in Jahren
- ε : Fehler

Regressionsfläche beschreibt den Zusammenhang zwischen den Variablen

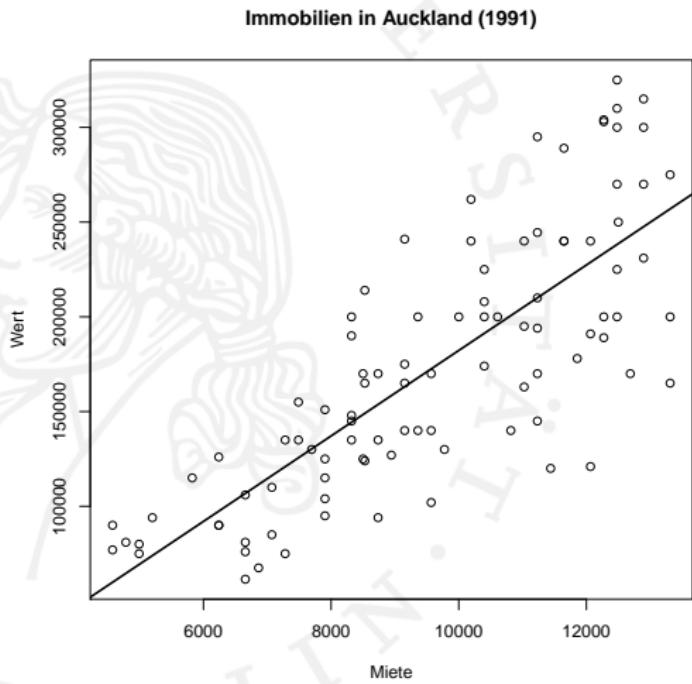


Beispiel 8.2

$$Y = m(X) + \varepsilon$$

- Y : Wert einer Immobilie
- X : Miete für die Immobilie
- ε : Fehler

Regressionsgerade beschreibt den Zusammenhang zwischen den Variablen



Residuum

Mögliche weitere Einflüsse auf Y :

- weitere erklärende X -Variablen (systematische Einflüsse)
- Zufallseinflüsse

Der Wert y_i setzt sich zusammen aus :

- einer Funktion der bekannten Einflüsse $x_{1i}, x_{2i}, \dots, x_{pi}$
- und dem Residuum \hat{u}_i

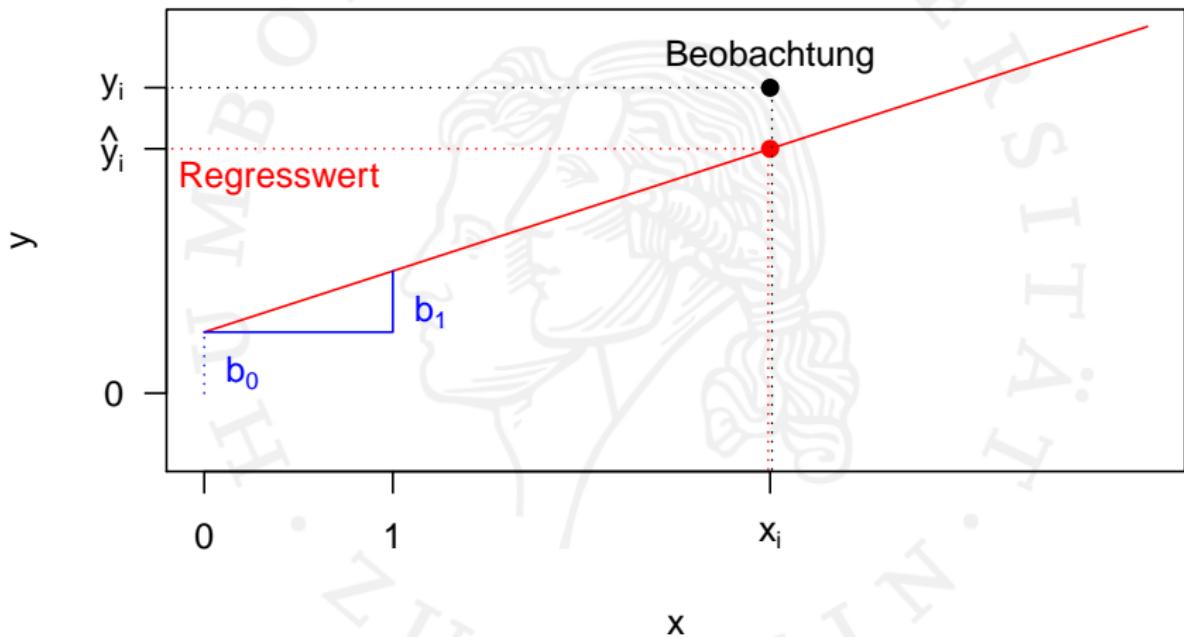
$$y_i = m(x_{1i}, x_{2i}, \dots, x_{pi}) + \hat{u}_i$$

- Die Residuen \hat{u}_i enthalten vor allem Zufallseinflüsse

Einfache lineare Regression

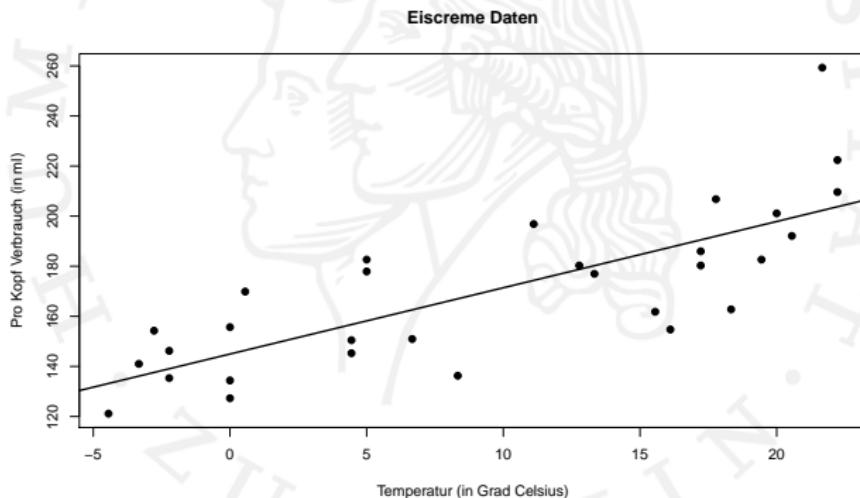
Regressionsgerade	Stichprobenregressionsmodell
$\hat{y}_i = b_0 + b_1 x_i$	$y_i = \hat{y}_i + \hat{u}_i = b_0 + b_1 x_i + \hat{u}_i$

- y_i - Beobachtungswerte der abhängigen Variablen Y
- x_i - Beobachtungswerte der unabhängigen Variablen X
- \hat{y}_i - Regresswerte (Funktionswerte)
- b_0 - geschätzter Regressionsparameter, Regressionskonstante, Schnittpunkt mit der Ordinatenachse ($x_i = 0$)
- b_1 - geschätzter Regressionsparameter, linearer Regressionskoeffizient, Anstieg der Regressionsgeraden
- $\hat{u}_i = y_i - \hat{y}_i$ - geschätzte Residuen, vertikaler Abstand zwischen Beobachtungswert y_i und Funktionswert \hat{y}_i



Beispiel 8.3 (Eiscremekonsum)

- Y - Pro-Kopf-Verbrauch von Eiscreme (ml)
- X - mittlere Monatstemperatur (Celsius)
- Datenbasis: Beobachtungen über ca. 30 Monate ($i = 1, \dots, 30$)



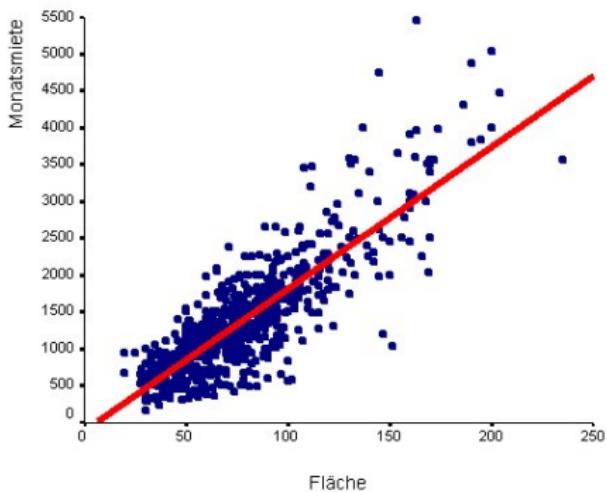
- Geschätzte Regressionsgrade:

$$\hat{y}_i = b_0 + b_1 x_i$$

- b_0 - geschätzte Regressionskonstante
- b_1 - geschätzter linearer Regressionskoeffizient
⇒ Wenn sich die mittlere Monatstemperatur um 1 Grad Celsius erhöht, erhöht sich **im Mittel** der Pro-Kopf-Eiscreme-Verbrauch um b_1 (ml)
- $\hat{u}_i = y_i - \hat{y}_i$ - geschätzte Residuen
⇒ Beinhalten weitere Einflüsse auf den Pro-Kopf-Eiscreme-Verbrauch (z.B. den Preis pro ml, das Familieneinkommen) sowie zufällige Einflüsse, die nur auf die einzelnen statistischen Einheiten wirken

Beispiel 8.4 (Monatsmiete)

- Y - Höhe der Monatsmiete (in €)
- X - Wohnfläche (in m^2)
- Datenbasis: Beobachtungen für 815 Berliner Mietwohnungen ($i = 1, \dots, 815$)



- Geschätzte Regressionsgrade:

$$\hat{y}_i = b_0 + b_1 x_i$$

- b_0 - geschätzte Regressionskonstante
- b_1 - geschätzter linearer Regressionskoeffizient
⇒ Wenn sich die Wohnfläche um 1 m^2 erhöht, erhöht sich (im Mittel) die Monatsmiete um $b_1 \text{ €}$
- $\hat{u}_i = y_i - \hat{y}_i$ - geschätzte Residuen
⇒ Beinhalten weitere Einflüsse auf die Monatsmiete (z.B. qualitative Ausstattungsmerkmale, Lage) sowie zufällige Einflüsse, die nur auf die einzelne statistische Einheit (Wohnung) wirken

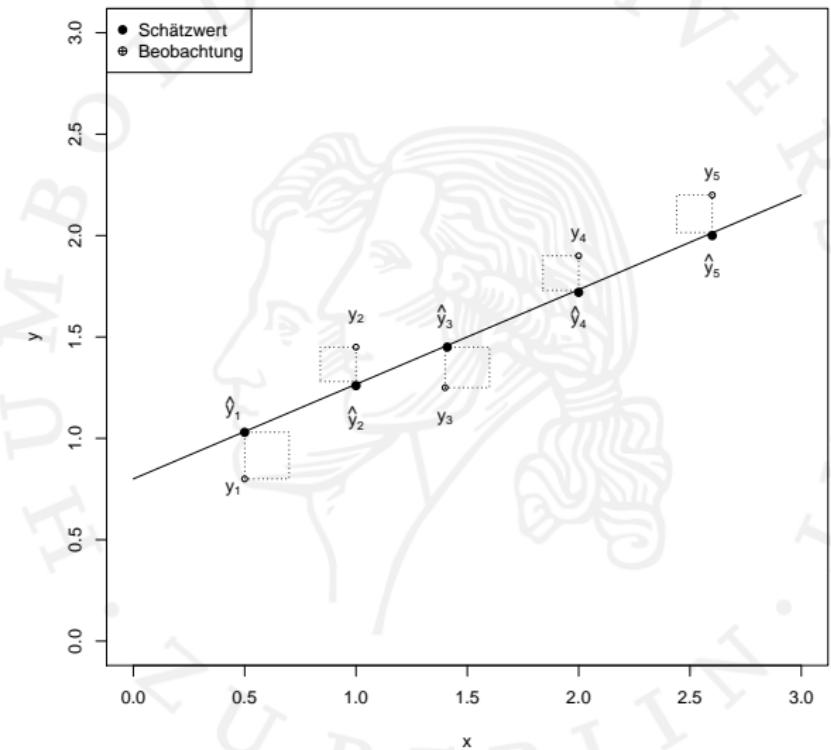
Schätzung der Regressionsparameter

- Regressionsfunktion $b_0 + b_1x$ soll so bestimmt werden, dass sie sich möglichst gut an die beobachteten Werte anpasst
- Abweichungen (Residuen) $\hat{u}_i = y_i - \hat{y}_i$ möglichst klein halten

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \hat{u}_i^2 \rightarrow \text{minimal}$$

- Prinzipiell möglich, aber nicht geschlossen lösbar

$$\sum_{i=1}^n |y_i - \hat{y}_i| \rightarrow \text{minimal}$$



Methode der kleinsten Quadrate

- Summe der quadrierten Abweichungen der Beobachtungswerte von den Regresswerten:

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 \rightarrow \min \quad | \quad \hat{y}_i = b_0 + b_1 x_i$$

- Minimum für $S(b_0, b_1)$ finden

$$S(b_0, b_1) = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2 \rightarrow \min_{b_0, b_1}$$

- ▶ Erste Ableitungen nach b_0 und b_1 gleich Null setzen
- ▶ Determinante der zweiten Ableitungen muss positiv sein

- Erste Ableitungen nach beiden Parametern b_0 und b_1 gleich Null setzen:

$$\frac{\partial S(b_0, b_1)}{\partial b_0} = -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i) \doteq 0$$

$$\frac{\partial S(b_0, b_1)}{\partial b_1} = -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i) x_i \doteq 0$$

- Zweite Ableitungen nach beiden Parametern b_0 und b_1 überprüfen:

$$\frac{\partial^2 S(b_0, b_1)}{\partial b_0^2} = 2n > 0, \quad \frac{\partial^2 S(b_0, b_1)}{\partial b_1^2} = 2 \sum_{i=1}^n x_i^2 > 0$$

$$\frac{\partial^2 S(b_0, b_1)}{\partial b_0 \partial b_1} = \frac{\partial^2 S(b_0, b_1)}{\partial b_1 \partial b_0} = 2 \sum_{i=1}^n x_i, \quad D = (2ns_x)^2 > 0$$

- Aufstellen der Normalgleichungen:

$$nb_0 + b_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \Leftrightarrow b_0 + b_1 \bar{x} = \bar{y}$$

$$b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i \Leftrightarrow b_0 \bar{x} + b_1 \bar{x^2} = \bar{xy}$$

- Nach der Lösung der Normalgleichungen (siehe Anhang) ergibt sich linearer Regressionskoeffizient b_1 und Regressionskonstante b_0 :

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

- Es folgt auch

$$\sum_{i=1}^n \hat{u}_i = 0$$

- Die Varianz von X muss größer als Null sein, um die Regressionsparameter schätzen zu können: $s_x^2 > 0$
- Eine nach der Methode der kleinsten Quadrate geschätzte Regressionsgerade verläuft stets durch den Punkt (\bar{x}, \bar{y}) , denn

$$\hat{y}_i = b_0 + b_1 x_i = \underbrace{\bar{y} - b_1 \bar{x}}_{b_0} + b_1 x_i = \bar{y} + b_1(x_i - \bar{x})$$

für $x_i = \bar{x}$ wird $\bar{y} + b_1(\bar{x} - \bar{x}) = \bar{y}$

- Regressions- und Korrelationsanalyse:

$$b_1 = \frac{s_{xy}}{s_x^2}, \quad r_{xy} = \frac{s_{xy}}{s_x s_y} \quad \Rightarrow b_1 = r_{xy} \frac{s_y}{s_x}$$

- Regression von x auf y \neq Regression von y auf x

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{s_{xy}}{s_x^2}$$

$$b_0^* = \bar{x} - b_1^* \bar{y}$$

$$b_1^* = \frac{s_{xy}}{s_y^2}$$

Fortsetzung des Beispiels Eiscreme

$$\bar{y} = 170,08 \text{ ml} \quad s_y^2 = 969,08 \text{ ml}^2 \Rightarrow s_y = 31,13 \text{ ml}$$

$$\bar{x} = 9,5 \text{ } ^\circ\text{C} \quad s_x^2 = 83,23 \text{ } ^\circ\text{C}^2 \Rightarrow s_x = 9,12 \text{ } ^\circ\text{C}$$

$$s_{yx} = 220,29$$

$$r_{yx} = 0,7756$$

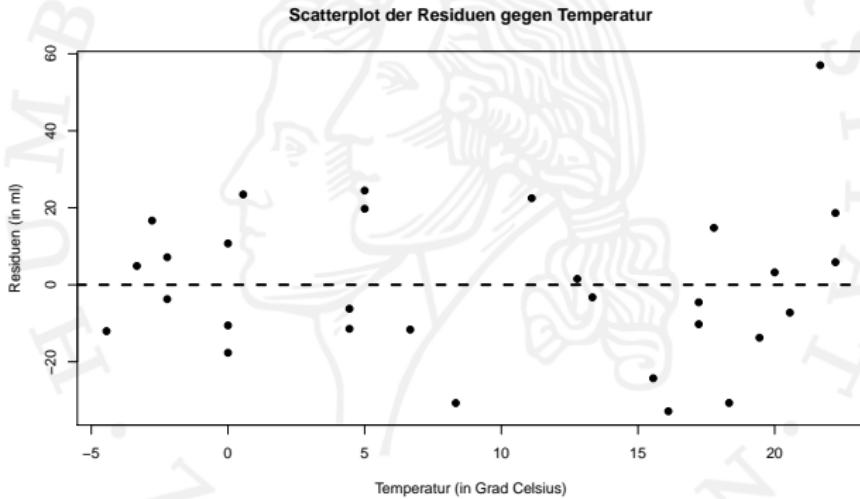
$$b_1 = \frac{s_{xy}}{s_x^2} = \frac{220,29}{83,23} = r_{yx} \cdot \frac{s_y}{s_x} = 0,7756 \cdot \frac{31,13}{9,12} = 2,65$$

$$b_0 = \bar{y} - b_1 \bar{x} = 170,08 - 2,65 \cdot 9,5 = 144,93$$

$$\hat{y}_i = b_0 + b_1 x_i = 144,93 + 2,65 x_i$$

Prüfung der Modellannahmen:

- Scatterplot der Residuen gegen die Temperatur
- Prüfung der Varianzhomogenität der Residuen



Fortsetzung des Beispiels Monatsmiete

$$\bar{y} = 1343,46 \text{ €} \quad s_y^2 = 544932,5 \text{ €}^2 \Rightarrow s_y = 738,20 \text{ €}$$

$$\bar{x} = 75,412 \text{ m}^2 \quad s_x^2 = 1043,61 \text{ m}^4 \Rightarrow s_x = 32,31 \text{ m}^2$$

$$s_{yx} = 20061,63$$

$$r_{yx} = 0,84$$

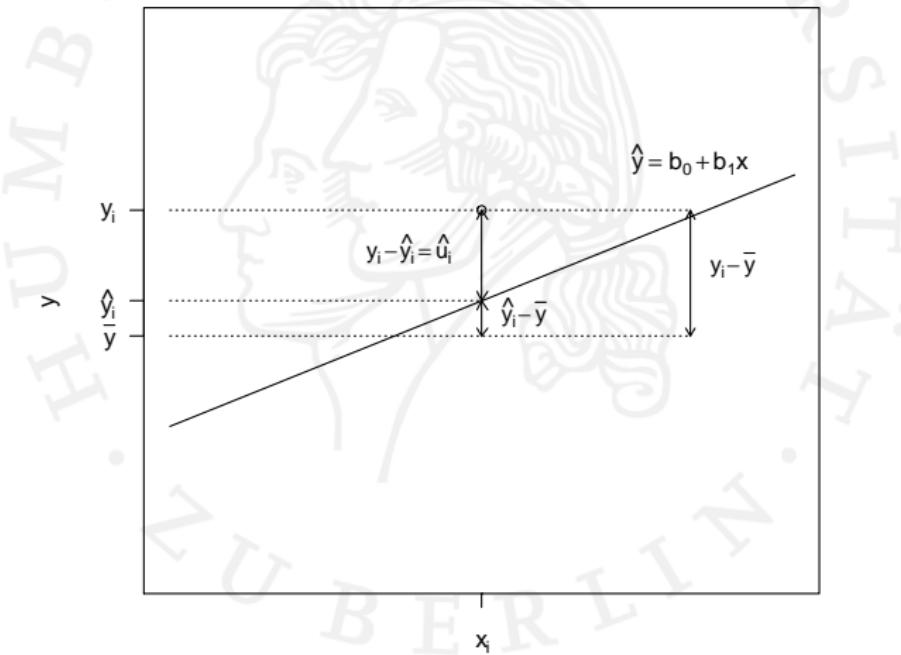
$$b_1 = \frac{s_{xy}}{s_x^2} = \frac{20061,63}{1043,61} = r_{yx} \cdot \frac{s_y}{s_x} = 0,84 \cdot \frac{738,20}{32,31} = 19,22$$

$$b_0 = \bar{y} - b_1 \bar{x} = 1343,46 - 19,22 \cdot 75,412 = -106,1$$

$$\hat{y}_i = b_0 + b_1 x_i = -106,18 + 19,22 x_i$$

Bestimmtheit (Güte) der Regression

- Wie gut beschreibt das Regressionsmodell die Beobachtungsdaten?



- Variation in Y

- | | |
|---|----------------------------|
| = Variation (Residuen) | + Variation (Regresswerte) |
| = durch Modell nicht erklärte Variation | + erklärte Variation |

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

- Varianz von Y:

$$s_y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}$$

Varianzzerlegung:

$$\begin{aligned}(y_i - \bar{y})^2 &= [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 \\ &= (y_i - \hat{y}_i)^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + (\hat{y}_i - \bar{y})^2\end{aligned}$$

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

Gemischter Term:

$$\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum_{i=1}^n \hat{u}_i(\hat{y}_i - \bar{y}) = \left(\underbrace{\sum_{i=1}^n \hat{y}_i \hat{u}_i}_{=0} \right) - \left(\bar{y} \underbrace{\sum_{i=1}^n \hat{u}_i}_{=0} \right) = 0$$

$$\sum_{i=1}^n \hat{u}_i \hat{y}_i = \sum_{i=1}^n \hat{u}_i (b_0 + b_1 x_i) = b_0 \underbrace{\sum_{i=1}^n \hat{u}_i}_{=0} + b_1 \sum_{i=1}^n \hat{u}_i x_i = b_1 \sum_{i=1}^n \hat{u}_i x_i$$

$$\frac{\partial S(b_0, b_1)}{\partial b_1} = -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i) x_i \doteq 0$$

$$\Rightarrow \sum_{i=1}^n (y_i - [b_0 + b_1 x_i]) x_i = \sum_{i=1}^n (y_i - \hat{y}_i) x_i = \sum_{i=1}^n \hat{u}_i x_i = 0$$

daher gilt:

$$\begin{aligned} \sum_i^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \\ &= \sum_{i=1}^n \hat{u}_i^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \end{aligned}$$

Bestimmtheitsmaß

$$\begin{aligned} R_{yx}^2 &= \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n \hat{u}_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= 1 - \frac{\sum_{i=1}^n \hat{u}_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \end{aligned}$$

$$R_{yx}^2 = \frac{\left[\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) \right]^2}{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_{yx}^2}{s_y^2 s_x^2} = r_{yx}^2$$



Zshg. mit Korrelation gilt nur für die einfache lineare Regression!

$$R_{yx}^2 = \frac{\left(n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i \right)^2}{\left[n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 \right] \left[n \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i \right)^2 \right]}$$

$$0 \leq R_{yx}^2 \leq 1, R_{yx}^2 = R_{xy}^2$$

Fortsetzung des Beispiels Eiscreme

$$s_y^2 = 969,08 \text{ ml}^2 \Rightarrow s_y = 31,13 \text{ ml}$$

$$s_x^2 = 83,23 \text{ }^\circ\text{C}^2 \Rightarrow s_x = 9,12 \text{ }^\circ\text{C}$$

$$s_{yx} = 220,29$$

$$r_{yx} = 0,7756$$

$$R_{yx}^2 = \frac{s_{yx}^2}{s_y^2 s_x^2} = \frac{220,29^2}{969,08 \cdot 83,23} = 0,6016$$

$$R_{yx}^2 = r_{yx}^2 = 0,7756^2 = 0,6016$$

Interpretation: 60,16% der Varianz des Pro-Kopf-Verbrauchs von Eiscreme wird durch die lineare Abhangigkeit von der mittleren Monatstemperatur erklart.

Fortsetzung des Beispiels Monatsmiete

$$s_y^2 = 544932,5 \Rightarrow s_y = 738,1954 \text{ €}$$

$$s_x^2 = 1043,608 \Rightarrow s_x = 32,305 \text{ m}^2$$

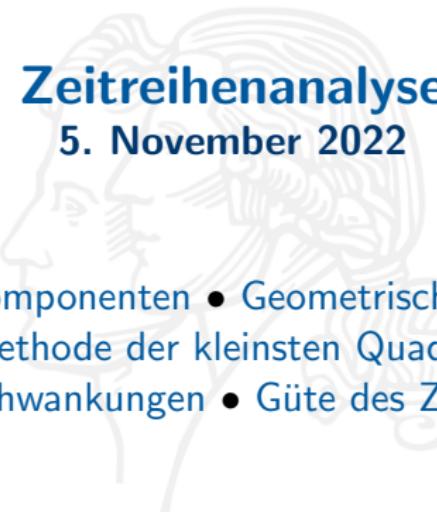
$$s_{yx} = 20061,626$$

$$r_{yx} = 0,8413$$

$$R_{yx}^2 = \frac{s_{yx}^2}{s_y^2 s_x^2} = \frac{20061,626^2}{544932,5 \cdot 1043,608} = 0,7077$$

$$R_{yx}^2 = r_{yx}^2 = 0,8413^2 = 0,7077$$

Interpretation: 70,77% der Varianz der Monatsmiete wird durch die lineare Abhangigkeit von der Wohnflache erklart.



Zeitreihenanalyse

5. November 2022

Zeitreihe und ihre Komponenten • Geometrisches Mittel • Gleitenden Durchschnitte • Methode der kleinsten Quadrate • Wahl von t • Periodische Schwankungen • Güte des Zeitreihenmodells

Zeitreihe und ihre Komponenten

Eine Zeitreihe ist eine Reihe von Werten einer Variablen X , die zu verschiedenen Zeitpunkten oder für verschiedene Zeitintervalle erhoben werden

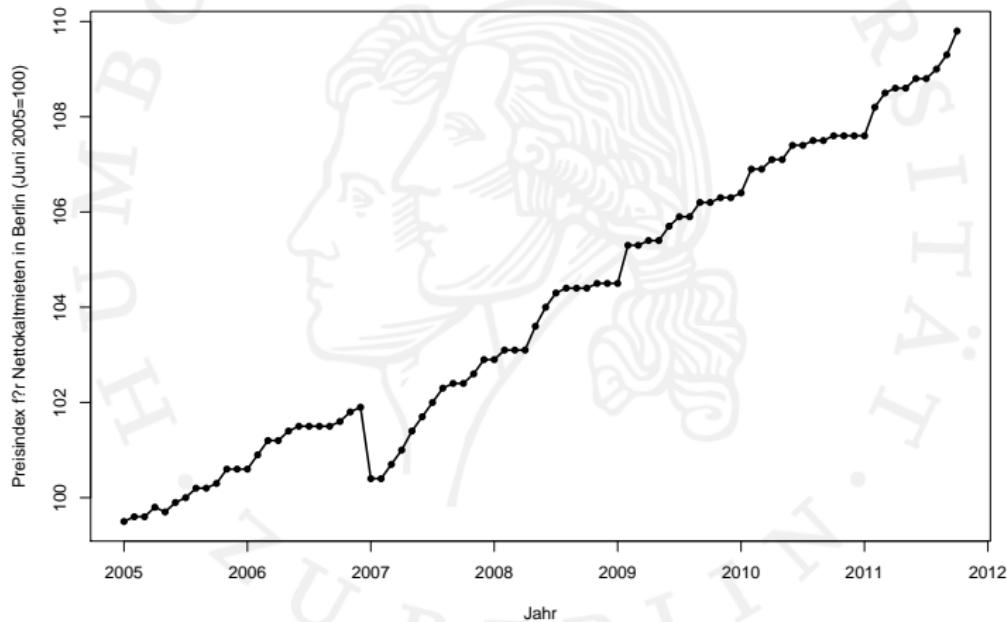
Ziel: Vorhersage der Zukunft (Prediktion)

Graphische Darstellung:

- Scatterplot:
 - ▶ Abszisse = Zeit
 - ▶ Ordinate = Variablenwerte x_t

Beispiel 9.1 (Preisindex)

Monatlicher Preisindex für Mietpreise in Berlin, Jan. 2005 - Okt. 2011



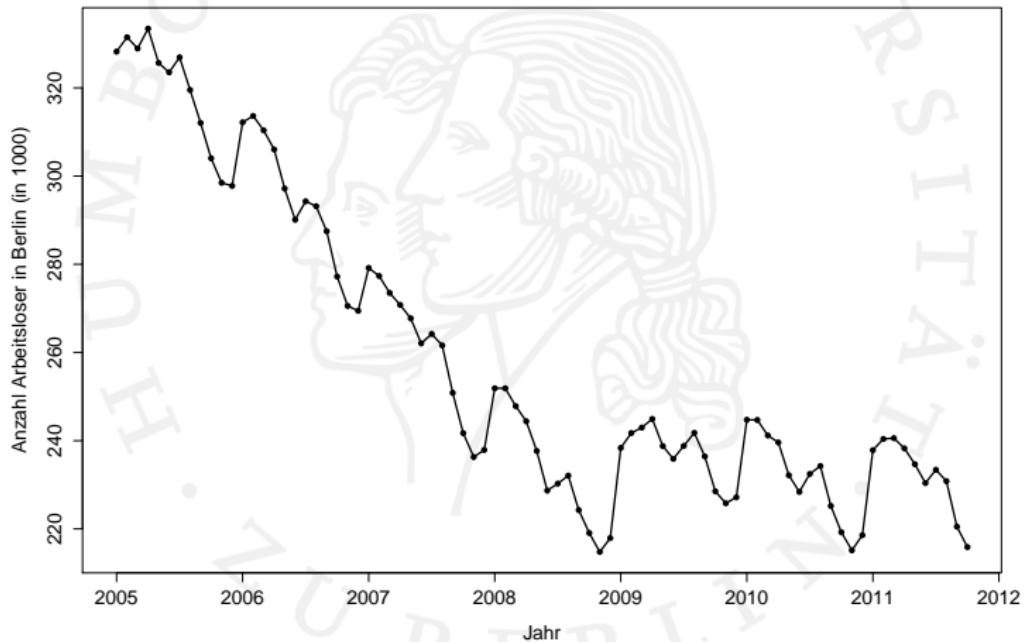
Beispiel 9.2 (Telefon)

Anzahl der Telefone in den USA (in Mio.) von 1900 - 1970



Beispiel 9.3 (Arbeitslosenzahlen)

Anzahl Arbeitsloser in Berlin, Jan. 2005 - Okt. 2011



Komponenten einer Zeitreihe

- Systematische Komponenten
 - ▶ Trend
 - ▶ Periodische Schwankungen
- ★ müssen aus der beobachteten Zeitreihe geschätzt werden
- Unregelmäßige Restschwankungen
 - ★ ergeben sich als Residuen

Geometrisches Mittel

Mittlere Entwicklung und Prognose

- mittlere Entwicklungsrate: geometrisches Mittel aus den Entwicklungsralten

$$i_G = \sqrt[T]{i_1 \cdot i_2 \cdot \dots \cdot i_T} = \sqrt[T]{\frac{x_T}{x_0}}$$

- Prognosewert $x_{T+\Delta T}^*$:

$$x_{T+\Delta T}^* = x_T \cdot i_G^{\Delta T}$$

- Bestimmung der Zeitdauer ΔT :

$$\Delta T = \frac{\log x_{T+\Delta T}^* - \log x_T}{\log i_G}$$

Beispiel 9.4

Bruttonationaleinkommen (BNE) der Bundesrepublik Deutschland von 2004 bis 2010 in Mrd. Euro

Jahr	t	BNE x_t	i_t	x_t^*
2004	0	2214,51	–	2214,51
2005	1	2249,59	1,0158	2263,13
2006	2	2361,03	1,0495	2312,83
2007	3	2470,33	1,0463	2363,61
2008	4	2505,50	1,0142	2415,51
2009	5	2424,85	0,9678	2468,55
2010	6	2522,75	1,0404	2522,75
2011	7	2668,92		2578,14
2012	8	2730,07		2634,75

$$\begin{aligned} i_G &= \sqrt[6]{\frac{2522,75}{2214,51}} = \sqrt[6]{1,1392} \\ &= 1,0219 \end{aligned}$$

$$\begin{aligned} x_{2014}^* &= 2522,75 \cdot 1,0219^4 \\ &= 2780,30 \text{ Mrd. €} \end{aligned}$$

$$\Delta T = \frac{\log(3000) - \log(2522,75)}{\log(1,0219)}$$

$$= 7,99 \text{ Jahre}$$

$$\Rightarrow 2010 + 8 = 2018$$

Gleitenden Durchschnitte

- Gegeben:

$$x_1, x_2, \dots, x_T \iff x_t, t = 1, \dots, T$$

- Ordnung des gleitenden Durchschnitts (Stützbereich):

- ▶ Anzahl k der Werte, die in die Mittelwertberechnung eingehen
- ▶ Ungerade Ordnung $2k + 1$:

$$x_t^* = \frac{1}{2k+1} \sum_{i=t-k}^{t+k} x_i \quad t = k+1, \dots, T-k$$

- ▶ Gerade Ordnung $2k$:

$$x_t^* = \frac{1}{2k} \left[\frac{1}{2} x_{t-k} + \frac{1}{2} x_{t+k} + \sum_{i=t-(k-1)}^{t+(k-1)} x_i \right] \quad t = k+1, \dots, T-k$$

Gleitende Durchschnitte mit ungerader Ordnung

k ungerade Ordnung:	1 $2k+1 = 3$	2 $2k+1 = 5$
x_1	$x_1^* = -$	$x_1^* = -$
x_2	$x_2^* = \frac{1}{3} \sum_{i=1}^3 x_i$	$x_2^* = -$
x_3	$x_3^* = \frac{1}{3} \sum_{i=2}^4 x_i$	$x_3^* = \frac{1}{5} \sum_{i=1}^5 x_i$
x_4	$x_4^* = \frac{1}{3} \sum_{i=3}^5 x_i$	$x_4^* = \frac{1}{5} \sum_{i=2}^6 x_i$
\vdots	\vdots	\vdots
x_{T-2}	$x_{T-2}^* = \frac{1}{3} \sum_{i=T-3}^{T-1} x_i$	$x_{T-2}^* = \frac{1}{5} \sum_{i=T-4}^T x_i$
x_{T-1}	$x_{T-1}^* = \frac{1}{3} \sum_{i=T-2}^T x_i$	$x_{T-1}^* = -$
x_T	$x_T^* = -$	$x_T^* = -$

Gleitende Durchschnitte mit gerader Ordnung

$$\begin{array}{ll} k & 1 \\ \text{gerade Ordnung:} & 2k = 2 \end{array}$$

$$x_1 \quad x_1^* -$$

$$x_2 \quad x_2^* = \frac{1}{2} \left[\frac{1}{2}x_1 + \frac{1}{2}x_3 + x_2 \right]$$

$$x_3 \quad x_3^* = \frac{1}{2} \left[\frac{1}{2}x_2 + \frac{1}{2}x_4 + x_3 \right]$$

$$x_4 \quad x_4^* = \frac{1}{2} \left[\frac{1}{2}x_3 + \frac{1}{2}x_5 + x_4 \right]$$

$$\vdots \quad \vdots$$

$$x_{T-2} \quad x_{T-2}^* = \frac{1}{2} \left[\frac{1}{2}x_{T-3} + \frac{1}{2}x_{T-1} + x_{T-2} \right]$$

$$x_{T-1} \quad x_{T-1}^* = \frac{1}{2} \left[\frac{1}{2}x_{T-2} + \frac{1}{2}x_T + x_{T-1} \right]$$

$$x_T \quad x_T^* -$$

$$\begin{array}{ll} 2 \\ 2k = 4 \end{array}$$

$$x_1^* -$$

$$x_2^* -$$

$$x_3^* = \frac{1}{4} \left[\frac{1}{2}x_1 + \frac{1}{2}x_5 + \sum_{i=2}^4 x_i \right]$$

$$x_4^* = \frac{1}{4} \left[\frac{1}{2}x_2 + \frac{1}{2}x_6 + \sum_{i=3}^5 x_i \right]$$

$$\vdots \quad \vdots$$

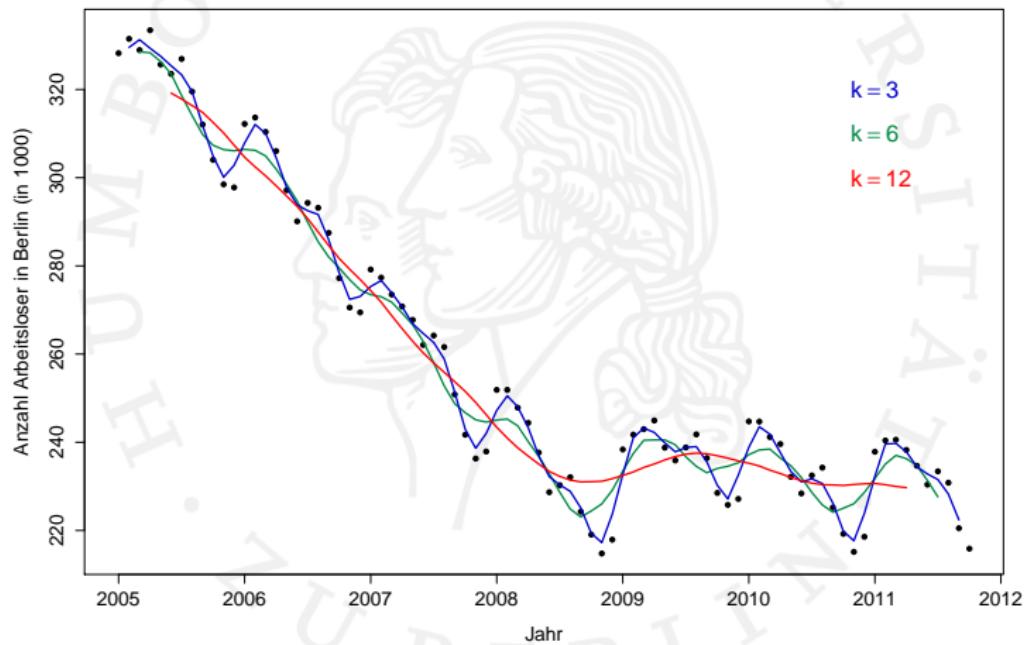
$$x_{T-2}^* = \frac{1}{4} \left[\frac{1}{2}x_{T-4} + \frac{1}{2}x_T + \sum_{i=T-3}^{T-1} x_i \right]$$

$$x_{T-1}^* -$$

$$x_T^* -$$

Fortsetzung des Beispiels Arbeitslosenzahlen

Anzahl Arbeitsloser in Berlin, Jan. 2005 - Okt. 2011



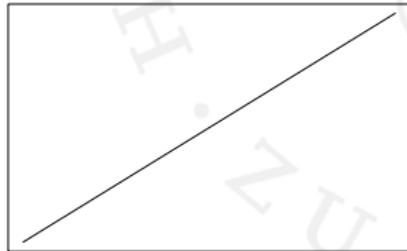
Methode der kleinsten Quadrate

- Wähle eine Trendfunktion \hat{x}_t , so dass der quadrierte Abstand der Beobachtungswerte x_t von der Trendfunktion \hat{x}_t minimal ist:

$$\sum_{t=1}^T (x_t - \hat{x}_t)^2 \rightarrow \min$$

- mögliche Trendfunktionen:

lineare Trendfunktion



exponentielle Trendfunktion



Lineare Trendfunktion

$$\hat{x}_t = a + b \cdot t$$

$$\begin{aligned} S(a, b) &= \sum_{t=1}^T (x_t - \hat{x}_t)^2 \\ &= \sum_{t=1}^T (x_t - a - b \cdot t)^2 \rightarrow \min_{a,b} \end{aligned}$$

$$a = \frac{\sum_{t=1}^T x_t \sum_{t=1}^T t^2 - \sum_{t=1}^T t \sum_{t=1}^T x_t t}{T \sum_{t=1}^T t^2 - \left(\sum_{t=1}^T t \right)^2}$$

$$b = \frac{T \sum_{t=1}^T x_t t - \sum_{t=1}^T x_t \sum_{t=1}^T t}{T \sum_{t=1}^T t^2 - \left(\sum_{t=1}^T t \right)^2}$$

Fortsetzung des Beispiels Preisindex

Monatlicher Preisindex für Mietpreise in Berlin, Jan. 2005 - Okt. 2011

$$\hat{x}_t = 98,7 + 0,1259t; \quad (t = 0 \hat{=} 2004.12, t = 1 \hat{=} 2005.1) \quad R^2 = 0,974$$



Exponentielle Trendfunktion

$$\hat{x}_t = ab^t$$

$$\Rightarrow \log \hat{x}_t = \log a + t \cdot \log b$$

$$\log a = \frac{\sum_{t=1}^T \log x_t \sum_{t=1}^T t^2 - \sum_{t=1}^T t \sum_{t=1}^T t \log x_t}{T \sum_{t=1}^T t^2 - \left(\sum_{t=1}^T t \right)^2}$$

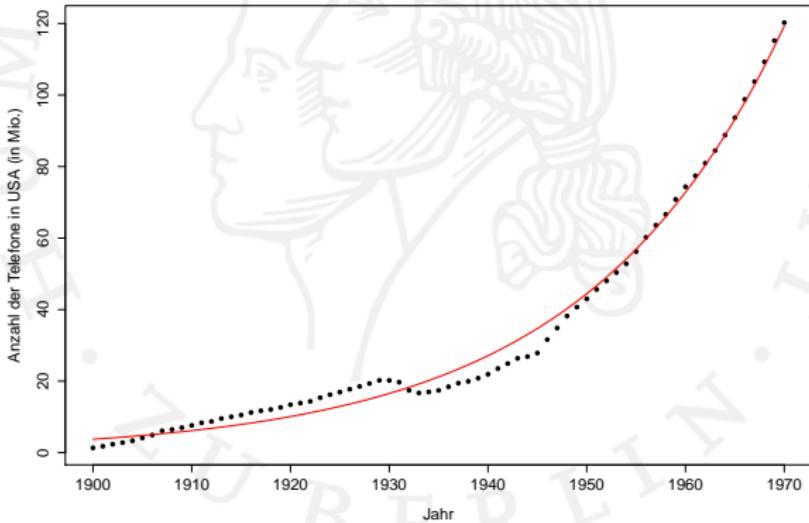
$$\log b = \frac{T \sum_{t=1}^T t \log x_t - \sum_{t=1}^T \log x_t \sum_{t=1}^T t}{T \sum_{t=1}^T t^2 - \left(\sum_{t=1}^T t \right)^2}$$

Fortsetzung des Beispiels Telefon

Anzahl der Telefone in den USA (in 1000), 1900 - 1970

$$\log \hat{x}_t = 8,18407 + 0,04937t \Leftrightarrow \hat{x}_t = 3583,41 \cdot 1,05^t$$

$$(t = 0 \hat{=} 1899, t = 1 \hat{=} 1900) \quad R^2 = 0,429$$



Wahl von t

Beispiel 9.5

Bruttonationaleinkommen (BNE) der Bundesrepublik Deutschland von 2004 bis 2010 in Mrd. Euro

Jahr	BNE x_t	t_1	t_2
2004	2214,51	0	-3
2005	2249,59	1	-2
2006	2361,03	2	-1
2007	2470,33	3	0
2008	2505,50	4	1
2009	2424,85	5	2
2010	2522,75	6	3
$\sum t_i$		21	0

- $t_1 = 0 \hat{=} 2004, t_1 = 1 \hat{=} 2005$

$$x_t = 2240,79 + t_1 \cdot 50,57$$

$$\begin{aligned}\hat{x}_{2011} &= 2240,79 + 7 \cdot 50,57 \\ &= 2595,47\end{aligned}$$

- $t_2 = 0 \hat{=} 2007, t_2 = 1 \hat{=} 2008$

$$x_t = 2392,79 + t_2 \cdot 50,57$$

$$\begin{aligned}\hat{x}_{2011} &= 2392,79 + 4 \cdot 50,57 \\ &= 2595,47\end{aligned}$$

Periodische Schwankungen

Notation

- Perioden: $P, i = 1, \dots, P$
- Unterzeiträume: $k, j = 1, \dots, k$
- Zeiträume: $T = k \cdot P$
- Trendwerte: $\hat{x}_{i,j}$
- Beobachtungswerte: $x_{i,j}$
- Schwankungskomponente: $s_{i,j}$

Additives Zeitreihenmodell

$$\begin{aligned}s_{i,j} &= x_{i,j} - \hat{x}_{i,j} \\ \bar{s}_j &= \frac{1}{P} \sum_{i=1}^P s_{i,j} \\ \hat{x}_{i,j}^{ZRM} &= \hat{x}_{i,j} + \bar{s}_j\end{aligned}$$

Multiplikatives Zeitreihenmodell

$$\begin{aligned}s_{i,j} &= \frac{x_{i,j}}{\hat{x}_{i,j}} \\ \bar{s}_j &= \frac{1}{P} \sum_{i=1}^P s_{i,j} \\ \hat{x}_{i,j}^{ZRM} &= \hat{x}_{i,j} \cdot \bar{s}_j\end{aligned}$$

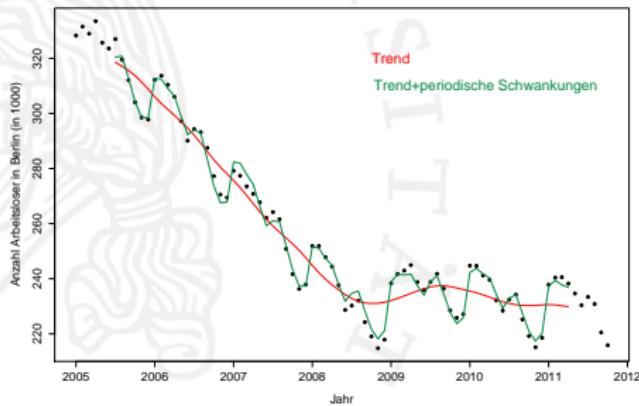
t	$\underbrace{1 \dots k}_1$										$\underbrace{k+1 \dots 2k}_2$										$\dots \underbrace{(P-1)k+1 \dots Pk}_P$									
Periode	1	2	\dots	P																										
Zeitreihe	$x_{1,1}$	\dots	$x_{1,k}$	$x_{2,1}$	\dots	$x_{2,k}$	\dots	$x_{P,1}$	\dots	$x_{P,k}$																				
Trend	$\hat{x}_{1,1}$	\dots	$\hat{x}_{1,k}$	$\hat{x}_{2,1}$	\dots	$\hat{x}_{2,k}$	\dots	$\hat{x}_{P,1}$	\dots	$\hat{x}_{P,k}$																				
1	$s_{1,1}$			$s_{2,1}$			\dots	$s_{P,1}$													\bar{s}_1									
\vdots	\vdots			\vdots			\vdots															\vdots								
k				$s_{1,k}$			$s_{2,k}$		\dots													\bar{s}_k								

Fortsetzung des Beispiels Arbeitslosenzahlen

Anzahl Arbeitsloser in Berlin, Jan. 2005 - Okt. 2011

Additives Zeitreihenmodell:

j	Summe	\bar{s}_j	P
1	47023,10	6717,586	7
2	62293,99	8899,142	7
3	55595,53	7942,218	7
\vdots	\vdots	\vdots	\vdots
12	-73383,85	-10483,407	7



- $P = 7$ Jahre
- $k = 12$ Unterzeiträume
- $T = 84$ Zeiträume

Güte des Zeitreihenmodells

- mittlere quadratische Streuung (Standardabweichung):

$$s_{ZRM} = \sqrt{\frac{1}{T} \sum_{i=1}^P \sum_{j=1}^k (x_{i,j} - \hat{x}_{i,j}^{ZRM})^2}$$

- Bestimmtheitsmaß:

$$R^2 = 1 - \frac{s_{ZRM}^2}{s_x^2}$$

$$0 \leq \frac{s_{ZRM}^2}{s_x^2} \leq 1 \quad \Rightarrow \quad 0 \leq R^2 \leq 1$$

$$s_x^2 = \frac{1}{T} \sum_{i=1}^P \sum_{j=1}^k (x_{i,j} - \bar{x})^2$$

Fortsetzung des Beispiels Arbeitslosenzahlen

- mittlere quadratische Streuung (Standardabweichung):

$$s_{ZRM} = \sqrt{\frac{1}{T} \sum_{i=1}^P \sum_{j=1}^k (x_{i,j} - \hat{x}_{i,j}^{ZRM})^2} = 2194,93$$

- Bestimmtheitsmaß:

$$R^2 = 1 - \frac{s_{ZRM}^2}{s_x^2} = 1 - \frac{2194,93}{34748,58} = 0,99$$

- ★ R^2 typischerweise sehr hoch bei Zeitreihen, da zusätzlich zum Trend periodische Schwankungen geschätzt werden

Indexzahlen

5. November 2022

- Indexzahlen • Notation • Ziel • Laspeyres-Preisindex • Paasche-Preisindex
- Beispiel • Laspeyres versus Paasche • Weitere Indizes • Fisher-Preisindex
- Mengenindizes • Wertindizes • Umbasieren von Indizes • Verketten von Indizes • Indexeigenschaften • Probleme mit Indizes • Der Fall Graciela Bevacqua • Globaler Wald-Notstand

Indexzahlen

- Bisher: Geometrisches Mittel und Zeitreihe
 - ▶ Zeitliche Entwicklung einer ökonomischen Größe über Messzahlen
- Jetzt: Zeitliche Entwicklung mehrerer Größen gleichzeitig

Beispiel 10.1

1. Preisentwicklung für Güter des privaten Konsums
 - ▶ Problem: Preise einiger Güter steigen, Preise anderer Güter fallen
→ Aggregation aller Messzahlen zu einer Indexzahl (Index)
2. Aktienindizes (DAX, Dow Jones, Euro Stoxx)
 - ▶ Aggregation von Kursen verschiedener Aktien zu einem Aktienkorb
 - ▶ Ziel: Darstellung der Entwicklung des Gesamtmarktes

Notation

- Betrachte einen Warenkorb (Kollektion von Gütern)
- Jedes Gut des Korbes hat einen Preis und eine Menge
 - ▶ n : Anzahl der Güter im Warenkorb
 - ▶ $p_t(i)$: Preis des Gutes i zur Zeit t
 - ▶ $q_t(i)$: Menge des Gutes i zur Zeit t
 - ▶ $v_t(i) = p_t(i)q_t(i)$: Wert des Gutes i zur Zeit t
- Benennungen
 - ▶ Preise: $\frac{\text{Geldeinheiten}}{\text{Mengeneinheit}}$ (z.B. 1 Euro/Liter)
 - ▶ Mengen: Mengeneinheiten (z.B. Liter, Kilogramm, Stück)
 - ▶ Wert: Geldeinheiten (z.B. Euro)
- Betrachtung zweier Zeitpunkte
 - ▶ Berichtszeit (notiert mit t)
 - ▶ Basiszeit (Setzung auf $t = 0$)

Ziel

- Beschreibung der Veränderungen von Preisen, Mengen und Werten des gesamten Warenkorbes zwischen der Berichtszeit t und dem Basiszeitpunkt 0
- Zunächst für einzelnes Gut i ($i = 1, \dots, n$)
 - ▶ Preismesszahl für das Gut i : $\frac{p_t(i)}{p_0(i)}$
 - ▶ Mengenmesszahl für das Gut i : $\frac{q_t(i)}{q_0(i)}$
 - ▶ Wertmesszahl für das Gut i : $\frac{v_t(i)}{v_0(i)} = \frac{p_t(i)}{p_0(i)} \frac{q_t(i)}{q_0(i)}$

Laspeyres-Preisindex

Die Mittelwertform des Preisindexes vom Typ Laspeyres ist definiert durch

$$I_{La;0,t}^p = \sum_{i=1}^n \frac{p_t(i)}{p_0(i)} \cdot \frac{p_0(i)q_0(i)}{\sum_{j=1}^n p_0(j)q_0(j)}$$

- gewogenes arithmetisches Mittel der Preismesszahlen $\frac{p_t(i)}{p_0(i)}$
- Gewichte $\frac{p_0(i)q_0(i)}{\sum_{j=1}^n p_0(j)q_0(j)}$ sind Ausgabenanteile für jedes Gut i zum Basiszeitpunkt 0

Aggregatform des Laspeyres-Indexes (Kürzen von $p_0(i)$)

$$I_{La;0,t}^p = \frac{\sum_{i=1}^n p_t(i)q_0(i)}{\sum_{i=1}^n p_0(i)q_0(i)}$$

Paasche-Preisindex

Die Mittelwertform des Preisindexes vom Typ Paasche ist definiert durch

$$I_{Pa;0,t}^P = \frac{1}{\sum_{i=1}^n \frac{1}{\frac{p_t(i)}{p_0(i)}} \cdot \frac{p_t(i)q_t(i)}{\sum_{j=1}^n p_t(j)q_t(j)}}$$

- gewogenes harmonisches Mittel der Preismesszahlen $\frac{p_t(i)}{p_0(i)}$
- Gewichte $\frac{p_t(i)q_t(i)}{\sum_{j=1}^n p_t(j)q_t(j)}$ sind Ausgabenanteile für jedes Gut i zum Berichtszeitpunkt t

Aggregatform des Paasche-Indexes (Umformung des Doppelbruchs)

$$I_{La;0,t}^P = \frac{\sum_{i=1}^n p_t(i)q_t(i)}{\sum_{i=1}^n p_0(i)q_t(i)}$$

Beispiel

- Warenkorb mit 3 Gütern

Gut i	Basiszeit $t = 0$ $p_0(i)$	Berichtszeit $t = 1$ $q_0(i)$	Berichtszeit $t = 1$ $p_1(i)$	Berichtszeit $t = 1$ $q_1(i)$	Preis- und Mengenmz. $p_1(i)/p_0(i)$	Preis- und Mengenmz. $q_1(i)/q_0(i)$
1	14,30	2,20	14,70	1,80	1,03	0,82
2	1,19	8,00	1,05	18,00	0,88	2,25
3	0,94	18,00	0,99	14,00	1,05	0,78

- Zwischenberechnung

i	$p_1(i)q_0(i)$	$p_0(i)q_0(i)$	$p_1(i)q_1(i)$	$p_0(i)q_1(i)$
1	32,34	31,46	26,46	25,74
2	8,40	9,52	18,90	21,42
3	17,82	16,92	13,86	13,16
\sum	58,56	57,90	59,22	60,32

- Indizes

$$I_{La,0,1} = \frac{58,56}{57,90} = 1,014 \quad > \quad I_{Pa,0,1} = \frac{59,22}{60,32} = 0,9818$$

Laspeyres versus Paasche

- Preisindex nach Laspeyres: Mengen aus dem Basiszeitraum
 - ▶ Die Werte Laspeyres-Preisindex bleiben vergleichbar über die Zeit, da sie alle den gleichen Nenner haben
- Preisindex nach Paasche: Mengen aus dem Berichtszeitraum
 - ▶ Ein praktisches Problem ist, dass die Gewichte (und auch der Warenkorb) zum Berichtszeitraum nicht bekannt sind und die Berechnung/Erstellung zeitaufwändig ist
 - ▶ Bruttoinlandsprodukt wird mit dem Paasche-Index berechnet
- Wenn die beiden Folgen der Preis- und Mengenmesszahlen

$$p_t(i)/p_0(i) \quad \text{und} \quad q_t(i)/q_0(i) \quad (i = 1, \dots, n)$$

negativ korreliert sind, dann gilt

$$I_{La;0,t}^P > I_{Pa,0,t}^P$$

- Wenn die beiden Folgen der Preis- und Mengenmesszahlen positiv korreliert sind, dann gilt $I_{La;0,t}^P < I_{Pa,0,t}^P$

Weitere Indizes

- Preisindex nach Carli (ungewichtetes Mittel)

$$I_{Ca;0,t}^p = \sum_{i=1}^n \frac{p_t(i)}{p_0(i)}$$

- Preisindex nach Drobisch (Mittelwert aus Laspeyre und Paasche)

$$I_{Dr;0,t}^p = \frac{I_{La;0,t}^p + I_{Pa;0,t}^p}{2}$$

Fisher-Preisindex

Der Preisindex vom Typ Fisher ist definiert durch

$$I_{Fi;0,t}^P = \sqrt{I_{La;0,t}^P I_{Pa;0,t}^P}$$

- Fisher-Index ist das geometrische Mittel aus Laspeyres- und Paasche-Index
- Es gilt:

$$\min(I_{La;0,t}^P, I_{Pa;0,t}^P) \leq I_{Fi;0,t}^P \leq \max(I_{La;0,t}^P, I_{Pa;0,t}^P)$$

- Für das Beispiel gilt

$$I_{Fi;0,1}^P = \sqrt{1,0114 \cdot 0,9818} = 0,9965$$

(Preisreduktion des Warenkorbes um 0,35%)

Mengenindizes

- Übertragung des Konzeptes der Preisindizes auf Mengenindizes durch einfache Vertauschung der Rollen von Preisen und Mengen
- Mengenindex nach Laspeyres

$$I_{La;0,t}^q = \sum_{i=1}^n \frac{q_t(i)}{q_0(i)} \cdot \frac{p_0(i)q_0(i)}{\sum_{j=1}^n p_0(j)q_0(j)} = \frac{\sum_{i=1}^n q_t(i)p_0(i)}{\sum_{i=1}^n q_0(i)p_0(i)}$$

- Mengenindex nach Paasche

$$I_{Pa;0,t}^q = \frac{1}{\sum_{i=1}^n \frac{1}{\frac{q_t(i)}{q_0(i)}} \cdot \frac{p_t(i)q_t(i)}{\sum_{j=1}^n p_t(j)q_t(j)}} = \frac{\sum_{i=1}^n q_t(i)p_t(i)}{\sum_{i=1}^n q_0(i)p_t(i)}$$

- Mengenindex nach Fisher

$$I_{Fi;0,t}^q = \sqrt{I_{La;0,t}^q I_{Pa;0,t}^q}$$

Wertindizes

- Kanonischer Wertindex

$$I_{0,t}^v = \frac{\sum_{i=1}^n v_t(i)}{\sum_{i=1}^n v_0(i)} = \frac{\sum_{i=1}^n p_t(i)q_t(i)}{\sum_{i=1}^n p_0(i)q_0(i)}$$

- Wertindizes nach Laspeyres, Paasche und Fisher

$$I_{La;0,t}^v = \sum_{i=1}^n \frac{v_t(i)}{v_0(i)} \cdot \frac{v_0(i)}{\sum_{j=1}^n v_0(j)}$$

$$I_{Pa;0,t}^v = \frac{1}{\sum_{i=1}^n \frac{1}{\frac{v_t(i)}{v_0(i)}} \cdot \frac{v_t(i)}{\sum_{j=1}^n v_0(j)}}$$

$$I_{Fi;0,t}^v = \sqrt{I_{La;0,t}^v I_{Pa;0,t}^v}$$

- Es gilt: $I_{0,t}^v = I_{La;0,t}^v = I_{Pa;0,t}^v = I_{Fi;0,t}^v$

Umbasieren von Indizes

- Gegeben sei eine Folge von Indizes zur Basiszeit s : $I_{s,t}^*$ mit $t = t_0, t_1, \dots, t_T$
- Eine Folge von Indizes $I_{r,t}$ zu einer alternativen Basiszeit $r \in \{t_0, t_1, \dots, t_T\}$ mit $r \neq s$, erhält man durch

$$I_{r,t} = \frac{I_{s,t}^*}{I_{s,r}^*}$$

- Beispiel: Der Verbraucherpreisindex wird auf Basis eines Warenkorbes berechnet und bezieht sich z.B. auf den Basiszeitraum 2010. So kann ein Verbraucherpreisindex für 2011, 2012, 2013, ... berechnet werden. Nun möchte man einen anderen Basiszeitraum wählen, z.B. 2015, so ergibt sich für 2018

$$I_{2015,2018} = \frac{I_{2010,2018}^*}{I_{2010,2015}^*}$$

Verketten von Indizes

- Gegeben seien zwei Folgen von Indizes zu äquidistanten Zeiten $I_{0,t}^*$ mit $t = 0, 1, \dots, s$ und $I_{s,t}^{**}$ für $t = s, s+1, \dots, T$
- Die verkette Folge zur Basiszeit 0 ist

$$I_{0,t} = \begin{cases} I_{0,t} & \text{für } t = 0, 1, \dots, s \\ I_{0,s}^* I_{s,t}^{**} & \text{für } t = s+1, \dots, T \end{cases}$$

- Beispiel: Der Verbraucherpreisindex wird auf Basis eines Warenkorbes berechnet und bezieht sich z.B. auf den Basiszeitraum 2010. So kann ein Verbraucherpreisindex für 2011, 2012, 2013, ... berechnet werden. Ab 2015 wird der Verbraucherpreisindex mit dem Basisjahr 2015 berechnet. Möchte man nun einen durchgehenden Index bis 2018 berechnen, so ergibt sich

$$I_{2010,t} = \begin{cases} I_{2010,t}^* & \text{für } t = 2010, \dots, 2015 \\ I_{2010,2015}^* I_{2015,t}^{**} & \text{für } t = 2015, 2016, 2017, 2018 \end{cases}$$

Indexeigenschaften

Ein Index $I_{s,t}$ (Basiszeitpunkt s , Berichtszeitpunkt t) sollte folgende Kriterien erfüllen (Fisher, 1922)

1. Identitätsprobe: $I_{t,t} = 1$
2. Zeitumkehrprobe: $I_{t,0} = \frac{1}{I_{0,t}}$
3. Rundprobe: $I_{t_1,t_T} = I_{t_1,t_2} I_{t_2,t_3} \dots I_{t_{T-1},t_T}$
4. Faktorumkehrprobe: $I_{0,t}^v = I_{0,t}^p I_{0,t}^q$
5. Proportionalitätsprobe: $I_{0,t}^P = 1 + \alpha$ (wenn alle Preise um α steigen)
6. Dimensionswechselprobe: Der Wert des Index hängt nicht von der Einheit ab
7. Bestimmtheitsprobe: Der Index soll auch dann bestimmt sein, wenn einzelne Preise oder Mengen gleich 0 sind

Fisherprobe	Laspeyres	Paasche	Fisher
Identitätsprobe	+	+	+
Zeitumkehrprobe	-	-	+
Rundprobe	-	-	-
Faktorumkehrprobe	-	-	+
Proportionalitätsprobe	+	+	+
Dimensionswechselsprobe	+	+	+
Bestimmtheitsprobe	+	+	+

- Frage: Welcher Index ist der 'beste'?
- Antwort: Der Fisher-Index erfüllt die meisten, aber auch nicht alle Kriterien (6 von 7)

Probleme mit Indizes

- Beispiel: Inflation

Am häufigsten wird zur Messung der Inflation der Verbraucherpreisindex herangezogen. Der Index wird mit Hilfe eines Warenkorbs berechnet, der in einem Basisjahr repräsentativ für einen durchschnittlichen Haushalt festgesetzt wird. Beim Verbraucherpreisindex wird der Warenkorb üblicherweise alle 5 Jahre angepasst.

- Problem 1: Mit zunehmendem Abstand zum Basisjahr ist der Warenkorb immer weniger repräsentativ, da das Konsumentenverhalten sich permanent ändert.

- ▶ Werden Produkte teurer, dann weicht der Konsument auf billigere Produkte aus (Substitutionseffekt)
- ▶ Neue Produkte finden keine Berücksichtigung.
- ▶ Produkte werden berücksichtigt, obwohl man sie nicht mehr kaufen kann.

- Problem 2: Die meisten Haushalte sind keine "durchschnittlichen" Haushalte (Größe, Einkommen).

Der Fall Graciela Bevacqua

- 1984 Beschäftigungsbeginn beim Instituto National de Estadistica y Censos (INDEC, äquivalent zum Statistischen Bundesamt), eingesetzt zur Berechnung des Verbraucherpreisindex
- 2005 Verbraucherpreisindex geschätzt auf +12,3% mit ansteigendem Trend
- Wirtschaftsminister stellt die Daten, Methoden und Ergebnisse in Frage

Die Forderungen von Moreno (Staatssekretär für Binnhandel) kamen täglich. Die Anrufe von Moreno waren 40-minütige geschriene Forderungen, ...

- ▶ Rundungen sollte immer Abrundungen sein, z.B. 2,599 zu 2,5
- ▶ Änderungen des Warenkorbes, z.B. nur "billige Kleidung", nur gekappte Preise verwenden
- ▶ Herausgabe von welchen Geschäften die Preise stammen

- 2007 suspendiert nach Urlaub, 2009 Kündigung
- vor 2006 waren Argentiniens Methoden vorbildlich in ganz Latinamerika
- seit 2012 verweigert "The Economist" die Aufnahme der offiziellen Inflationsrate des INDEC
- Bevacqua und andere versuchten unabhängige Berechnungen der Inflationsrate
- in 2011 wurden sieben Firmen angeklagt nicht "angemessene Methoden" zu benutzen (Strafe: ≈ 100.000 EUR)
- in 2012 wurde Bevacqua angeklagt "mit falschen Informationen den Markt zu verzerren" (2-6 Jahre Gefängnis)
- Anklagen wurden fallengelassen durch das Gericht

Globaler Wald-Notstand

WWF-Report: Tierbestände in Wäldern halbiert seit 1970

Die weltweiten Tierbestände in Wäldern sind seit 1970 durchschnittlich um mehr als die Hälfte zurückgegangen. Zu diesem Ergebnis kommt die am Dienstag vom WWF veröffentlichte Studie „Below The Canopy“. Es ist die erste Untersuchung ihrer Art, die sich speziell der Entwicklung der globalen Tierpopulationen in Wäldern widmet. Als Hauptursache für den Rückgang nennen die Umweltschützer den durch Menschen verursachten Lebensraumverlust. Entwaldung und Degradierung der Wälder seien zu 60 Prozent für den Einbruch der Tierbestände verantwortlich. Besonders dramatisch ist die Entwicklung laut WWF in den Tropen, wie etwa dem Amazonas-Regenwald.

Quellen:

- Pressemitteilung WWF vom 13.08.2019
- Below the Canopy - WWF publication
- Below the canopy: global trends in forest vertebrate populations and their drivers - PeerJ preprint

- Basiert auf dem Living Planet Index
- Gewichteter “Forest Specialist Index”, proportional zum Artenreichtum der einzelnen biogeografischen Gebiete und Taxa
- Forest Specialist Index basiert auf 268 “forest specialist” Arten (Art kommt hauptsächlich nur im Wald vor)
 - ▶ 135 Vogelarten,
 - ▶ 89 Säugetierarten,
 - ▶ 19 Reptilienarten und
 - ▶ 25 Amphibienarten.
- Probleme:
 1. komplexe Methodik
 - ★ im Kern gewichtete geometrische Mittel der Veränderung
 - ★ wie gut/stabil sind die Gewichte?
 2. Datengrundlage(!!)

- Sowohl Methodenentwicklung als auch Datenquelle (www.livingplanet.org) sind assoziiert mit dem WWF
- Verwendete Arten sind Wirbeltierarten
 - ▶ keine Würmer, Käfer, Schnecken, Spinnen, Insekten
 - ▶ nur etwa 4% aller Lebewesen sind Wirbeltiere (repräsentativ?)
- Basiszeitraum 1970
 - ▶ vor 1970 kaum Daten vorhanden
 - ▶ vor 1970 könnte sich die Anzahl der Lebewesen schon drastisch reduziert haben (Unterschätzung der Veränderung)
- Keine weltweite Tierarteninventur
 - ▶ Zählung bzw. Schätzungen nur an bestimmten Orten
 - ▶ Orte mit wenig Veränderung sind uninteressant für Wissenschaftler
 - ▶ Sekundärstatistische Untersuchung

A faint watermark of the HU Berlin logo is visible in the background, featuring a circular emblem with two figures and the text "HUMBOLDT-UNIVERSITÄT ZU BERLIN".

Teil 3

Induktive Statistik

Wahrscheinlichkeitsrechnung

5. November 2022

- Das Ziegenproblem • Zufallsexperimente und Ereignisse • Mengenlehre • Relationen und Operationen • Wahrscheinlichkeitsbegriffe • Binomialkoeffizient • Permutationen • Variationen • Kombinationen • Übersicht Kombinatorik • Wahrscheinlichkeitsbegriffe • Rechenregeln • Bedingte Wahrscheinlichkeit • Multiplikationssatz • Das Ziegenproblem • Unabhängige Ereignisse • Totale Wahrscheinlichkeit • Theorem von Bayes
 - Beweis: Binomialkoeffizienten • Beweis: Additionssatz • Beweis: Unabhängige Ereignisse

Das Ziegenproblem

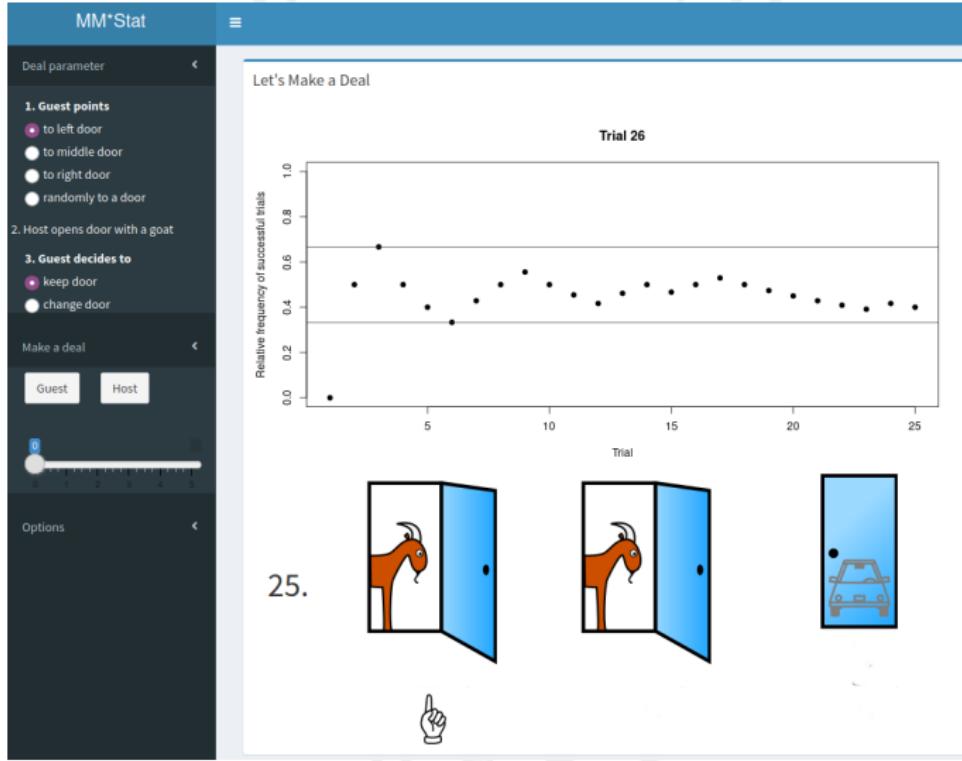
Beispiel 11.1

Nehmen Sie an, Sie wären in einer Spielshow und hätten die Wahl zwischen drei Toren. Hinter einem Tor ist ein Auto, hinter den anderen sind Ziegen. Sie wählen ein Tor, z.B. Tor Nummer 1, und der Showmaster, der weiß, was hinter den Toren ist, öffnet ein anderes Tor, z.B. Nummer 3 hinter dem eine Ziege steht. Er fragt Sie nun: „Möchten Sie das Tor Nummer Zwei?“ Ist es von Vorteil, die Wahl des Tors zu ändern?

(Auch Monty-Hall-Dilemma genannt, nach Monty Hall, dem Moderator der US-amerikanischen Spielshow 'Let's make a deal')



examples/stat/monty_hall



Zufallsexperimente und Ereignisse

- Zufallsexperiment: Ist ein Vorgang
 - ▶ der beliebig oft und gleichartig wiederholbar ist
 - ▶ der mindestens zwei mögliche verschiedene Ergebnisse hat
 - ▶ der nach einer ganz bestimmten Vorschrift ausgeführt wird
 - ▶ dessen Ergebnis vom Zufall abhängt
- Ergebnisse: möglichen Resultate des Zufallsexperimentes
- Stichprobenraum, Ereignisraum: Menge aller möglichen Ergebnisse des Zufallsexperimentes
- Ereignis: eine beliebige Zusammfassung von Ergebnissen des Zufallsexperimentes
- Elementarereignis: Ein Ereignis, dass nur ein Ergebnis enthält

Beispiel 11.2 (Einmaliges Werfen eines Würfels)

- Zufallsexperiment: Einmaliges Werfen eines Würfels
- Ergebnisse: 1, 2, 3, 4, 5 oder 6
- Stichprobenraum, Ereignisraum:
 $S = \{1, 2, 3, 4, 5, 6\}$
- Mögliche Ereignisse:
 - ▶ die Augenzahl ist eine gerade Zahl: $A = \{2, 4, 6\}$
 - ▶ die Augenzahl ist 5: $B = \{5\}$
- Sechs Elementarereignisse:
 $\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}$

Beispiel 11.3 (Einmaliges Werfen zweier Würfel)

- Zufallsexperiment: Einmaliges Werfen zweier Würfel
- Ergebnisse:
 $(1,1), (1,2), \dots, (6,6)$
- Stichprobenraum, Ereignisraum: $S = \{(1,1), (1,2), \dots, (6,6)\}$
- Mögliche Ereignisse:
 - ▶ Pasch $A = \{(1,1), (2,2), (3,3), (4,4), (5,5), (6,6)\}$
 - ▶ Augensumme kleiner vier $B = \{(1,1), (1,2), (2,1)\}$
 - ▶ Augensumme gleich vier $C = \{(1,3), (2,2), (3,1)\}$
- 36 Elementarereignisse:
 $\{(1,1)\}, \{(1,2)\}, \{(1,3)\}, \dots, \{(6,6)\}$

Vollständige Zerlegung

Die Ereignisse A_1, A_2, \dots, A_n bilden eine vollständige Zerlegung des Ereignisraumes S , wenn

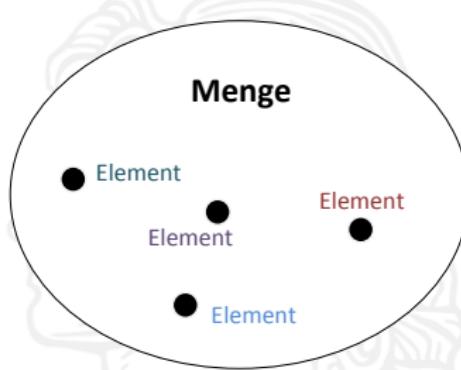
- $A_i \cap A_j = \emptyset$ für $i, j = 1, \dots, n, i \neq j$
- $A_1 \cup A_2 \cup \dots \cup A_n = S$
- $P(A_i) > 0$ für $i = 1, \dots, n$

Beispiel 11.4 (Einmaliges Werfen eines Würfels)

- Die Ereignisse $G = \{2, 4, 6\}$ (gerade Augenzahl) und $U = \{1, 3, 5\}$ (ungerade Augenzahl) sind eine Zerlegung von S
- $G \cap U = \emptyset$
- $G \cup U = S$

Mengenlehre

- Das Konzept formalisiert die Idee, Objekte zu gruppieren und sie als Einheit zu betrachten



- Elementarereignis \leftrightarrow ein Element
- Ereignis \leftrightarrow eine Menge von Elementarereignissen
- Ereignisraum \leftrightarrow die Menge aller möglichen Elementarereignisse

Leere Menge $A = \emptyset$

Die Menge, die kein Element enthält, heißt leere Menge

Teilmenge $A \subset B$

Eine Menge A heißt Teilmenge einer Menge B , wenn jedes Element von A auch Element von B ist

Schnittmenge $A \cap B$

Die Schnittmenge zweier Mengen A und B ist die Menge aller Elemente, die sowohl in A als auch in B enthalten sind

Vereinigungsmenge $A \cup B$

Die Vereinigungsmenge von A und B ist die Menge der Elemente, die in A oder in B enthalten sind

Differenzmenge $A \setminus B$

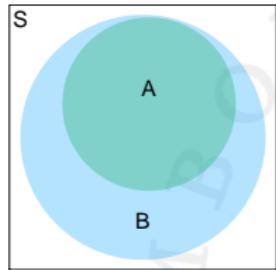
Die Differenzmenge von A und B ist die Menge der Elemente, die in A , aber nicht in B enthalten sind

Relationen und Operationen

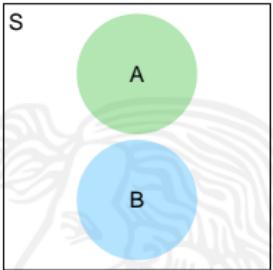
Das *Ereignis A tritt ein*, wenn das Zufallsexperiment das Ergebnis ω hat und ω in A enthalten ist.

Beschreibung des zugrundeliegenden Sachverhaltes	Bezeichnung (Sprechweise)	Darstellung
A tritt sicher ein	A ist sicheres Ereignis	$A = S$
A tritt sicher nicht ein	A ist unmögliches Ereignis	$A = \emptyset$
wenn A eintritt, tritt B ein	A ist Teilmenge von B	$A \subset B$
genau dann, wenn A eintritt, tritt B ein	A und B sind äquivalente Ereignisse	$A \equiv B$
wenn A eintritt, tritt B nicht ein genau dann, wenn A eintritt, tritt B nicht ein	A und B sind disjunkte Ereignisse A und B sind komplementäre Ereignisse	$A \cap B = \emptyset$ $B = \bar{A}$
genau dann, wenn mindestens ein A_i (A_1 oder A_2 oder ...) eintritt, tritt A ein	A ist Vereinigung der A_i	$A = \bigcup_i A_i$
genau dann, wenn alle A_i (A_1 und A_2 und ...) eintreten, tritt A ein	A ist Durchschnitt der A_i	$A = \bigcap_i A_i$

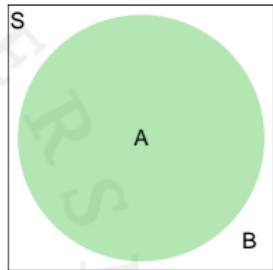
Venn-Diagramme



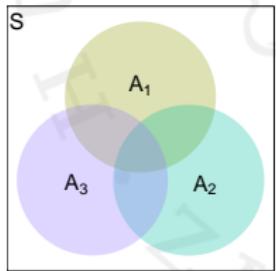
$$A \subset B$$



$$A \cap B = \emptyset$$



$$B = \overline{A}$$



$$A = \bigcup_i A_i$$



$$A = \bigcap_i A_i$$

Wahrscheinlichkeitsbegriffe

Was ist Wahrscheinlichkeit?



Laplace (1749-1827)



von Mises (1883-1953)



Kolmogorow (1903-1987)

Die Wahrscheinlichkeit ist ein Maß P zur Quantifizierung der Sicherheit bzw. Unsicherheit des Eintretens eines bestimmten Ereignisses A im Rahmen eines Zufallsexperimentes.

Beispiel 11.5

- Zufallsexperiment: Werfen eines Würfels
- Ereignis A: „gerade Augenzahl“
- $P(A)$ = Wahrscheinlichkeit, dass eine gerade Zahl gewürfelt wird

Welche Eigenschaften sollte P haben?

- $0 \leq P(A) \leq 1$
- $P(\emptyset) = 0$, $P(S) = 1$

Klassischer Wahrscheinlichkeitsbegriff nach Laplace

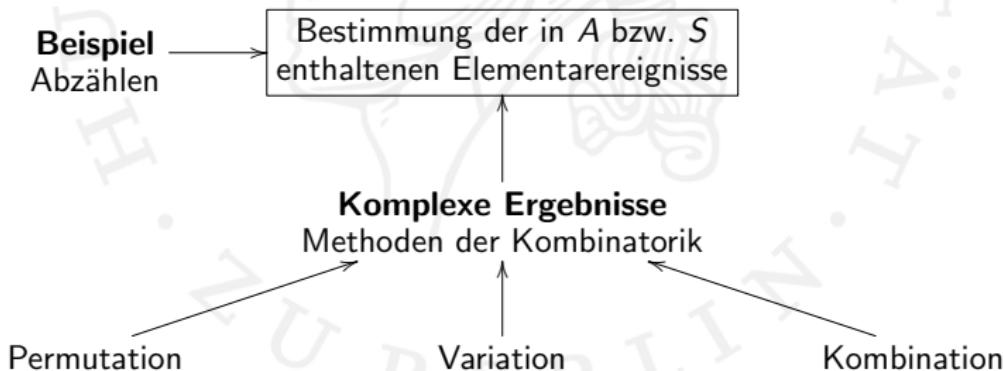
$$P(A) = \frac{\text{Anzahl der in } A \text{ enthaltenen Elementarereignisse}}{\text{Anzahl der Elementarereignisse in } S}$$

Voraussetzungen

- es gibt min. 2 möglichen Elementarereignisse
- genau eines der möglichen Elementarereignisse tritt ein
- Anzahl der möglichen Elementarereignisse ist endlich
- jedes Elementarereignis hat die gleiche Wahrscheinlichkeit des Eintretens

Beispiel 11.6

- Zufallsexperiment: Werfen eines idealen Würfels
- Ereignis A: „gerade Augenzahl“
- Elementarereignisse: $\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}$
- in A enthaltene Elementarereignisse: $\{2\}, \{4\}, \{6\}$
- $P(A) = 3/6 = 0,5$



Beispiel 11.7 (Medikamentennebenwirkung)

Auf den Beipackzetteln zu Medikamenten wird angegeben, wie häufig eine bestimmte Nebenwirkung auftritt:

Sehr häufig	mehr als 1 Behandelter von 10	>	10%
Häufig	1 bis 10 Behandelte von 100	1 -	10%
Gelegentlich	1 bis 10 Behandelte von 1.000	0,1 -	1%
Selten	1 bis 10 Behandelte von 10.000	0,01 -	0,1%
Sehr selten	weniger als 1 Behandelter von 10.000	<	0,01%
Nicht bekannt	Häufigkeit auf Grundlage der verfügbaren Daten nicht abschätzbar		

Zum Beispiel für Nebenwirkungen mit einer Häufigkeit von 1:1 Million müssen etwa sechs Millionen Anwendungen beobachtet werden. Dadurch besteht bei neuen oder wenig verbreiteten Arzneimitteln ein erhöhtes Risiko für bis dahin unbekannte Nebenwirkungen (Quelle: [Wikipedia](#)).

Beispiel 11.8 (Lottozahlen)

- Wie groß ist die Wahrscheinlichkeit von 6 Richtigen beim dem Lotto "6 aus 49"?
- Nach Laplace: $\frac{\text{Anzahl Gewinnziehung}}{\text{Anzahl möglicher Ziehungen}}$
- Zufallsexperiment: Ziehen von sechs Kugeln aus dem Ziehungsgerät
- Elementarereignis: Sechs gezogene Kugeln, z.B. $\{(24, 25, 26, 30, 31, 32)\}$ (23.01.1988)
- Wie wieviele Möglichkeiten gibt es 6 aus 49 Kugeln zu ziehen?

Einflußfaktoren:

- Wieviele Kugeln werden gezogen?
- Sind alle Kugeln unterscheidbar?
- Spielt die Reihenfolge der Ziehungen eine Rolle?
- Kann eine Kugel noch einmal gezogen werden?

Binomialkoeffizient

Fakultät

$$n! = 1 \cdot 2 \cdot 3 \cdot \dots \cdot (n-1) \cdot n$$

Binomialkoeffizient (Eulersches Symbol)

$$\binom{n}{k} = \frac{n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot (n-k+1)}{1 \cdot 2 \cdot 3 \cdot \dots \cdot (k-1) \cdot k} = \frac{n!}{k! \cdot (n-k)!}$$

Symmetrie-Eigenschaft

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

Spezialfälle

$$\binom{n}{0} = \frac{n!}{0!(n-0)!} = 1$$

$$\binom{n}{1} = \frac{n!}{1!(n-1)!} = n$$

$$\binom{n}{k} = 0 \quad \text{wenn } k > n \geq 0$$

Summen-Eigenschaft

$$\binom{n}{k} + \binom{n}{k+1} = \binom{n+1}{k+1}$$

Permutationen

Jede Zusammenstellung, in der alle n gegebenen Elemente in irgendeiner Anordnung stehen, heißt eine Permutation.

Permutationen ohne gleiche Elemente

Anzahl der Permutationen: $P(n) = 1 \cdot 2 \cdot 3 \cdots \cdots n = n!$

Ohne Wiederholung meint hier, dass alle n Elemente verschieden sind.

Permutationen mit Wiederholung

Anzahl der Permutationen: $P(n; g) = \frac{n!}{g!} \quad g \leq n$

Mit Wiederholung meint hier, dass g der n Elemente gleich sind.

Beispiel 11.9 (Permutationen ohne gleiche Elemente)

- 2 Elemente a, b : $P(2) = 2 \cdot 1$

Permutationen: ab, ba

- 3 Elemente a, b, c : $P(3) = 3 \cdot 2 \cdot 1$

Permutationen: $abc, bac, cab, acb, bca, cba$

Beispiel 11.10 (Permutationen mit gleichen Elementen)

- 2 Elemente a, b

wenn $a \neq b$	wenn $b = a$
$ab \ ba$	$aa \ aa$
$P(n) = 2! = 2$	$P(2; 2) = 2! / 2! = 1$

- 3 Elemente a, b, c

wenn $a \neq b \neq c \neq a$	wenn $b = a \neq c$
$abc \ bca \ cab$ $bac \ acb \ cba$	$aac \ aca \ caa$ $aac \ aca \ caa$
$P(3) = 3! = 6$	$P(3; 2) = 3! / 2! = 3$
	wenn $b = c = a$
	$aaa \ aaa \ aaa$ $aaa \ aaa \ aaa$
	$P(3; 3) = 3! / 3! = 1$

Mit mehreren Gruppen gleicher Elemente

$$g_1, g_2, \dots, g_r \text{ mit } g_1 + g_2 + \dots + g_r = n$$

Anzahl der Permutationen: $P(n; g_1, \dots, g_r) = \frac{n!}{g_1! \cdot g_2! \cdot \dots \cdot g_r!}$

Beispiel 11.11

Wie groß ist die Wahrscheinlichkeit, dass nach 5 (zufälligen) Zügen im Tic Tac Toe Spiel der 1. Spieler gewonnen hat?

0		
	X	0
0	X	X

- Nach Laplace: $\frac{\text{Anzahl Gewinnstellungen}}{\text{Anzahl möglicher Stellungen}}$
- Stellungen nach 5 Zügen: $P(9; 3; 2; 4) = \frac{9!}{3!2!4!} = 1260$
- 8 Gewinnstellungen $\Rightarrow P(\text{Gewinn}) = \frac{8}{1260} \approx 9,523\%$

- 8 Gewinnstellungen im 5. Zug:

x	x	x

x	x	x

x	x	x

x		
	x	
		x

x		
x		
x		

	x	
x		
	x	

		x
x		
	x	

		x
	x	
x		
		x

- Von den restlichen 6 Felder müssen 2 belegt sein durch den anderen Spieler und 4 noch frei sein, also

$$P(6; 4; 2) = \frac{6!}{4!2!} = \frac{5 \cdot 6}{1 \cdot 2} = 15$$

Variationen

Von n Elementen zur k -ten Klasse (Ordnung):

- jede Zusammenstellung von k Elementen aus n Elementen ($k \leq n$)
- unter Berücksichtigung ihrer Anordnung
- alle n Elemente der Ausgangsmenge sind verschieden

Mit Wiederholung

Anzahl der Variationen: $V^W(n, k) = n^k$

Mit Wiederholung meint hier, dass ein Element mehrmals ausgewählt werden kann.

Ohne Wiederholung

Anzahl der Variationen:

$$V(n, k) = n \cdot (n - 1) \cdot \dots \cdot (n - k + 2) \cdot (n - k + 1) = \frac{n!}{(n-k)!}$$

Ohne Wiederholung meint hier, dass jedes Element nur genau einmal ausgewählt werden kann.

Beispiel 11.12 (Variationen mit Wiederholung)

$n = 3$ Elemente: 1, 2, 3

- $k = 1$, Anzahl der Variationen: $V^W(3, 1) = 3$

1 2 3

- $k = 2$, Anzahl der Variationen: $V^W(3, 2) = 3 \cdot 3 = 3^2 = 9$

11 12 13

21 22 23

31 32 33

- $k = 3$, Anzahl: $V^W(3, 3) = 3 \cdot 3 \cdot 3 = 3^3 = 27$

111 112 113 211 212 213 311 312 313

121 122 123 221 222 223 321 322 323

131 132 133 231 232 233 331 332 333

Beispiel 11.13 (Variationen ohne Wiederholung)

$n = 3$ Elemente: 1, 2, 3

- $k = 1$, Anzahl: $V(3, 1) = 3$

1 2 3

- $k = 2$, Anzahl: $V(3, 2) = 3^2 - 3 = 3 \cdot 2 = 6$

11 12 13
21 22 23
31 32 33

- $k = 3$, Anzahl: $V(3, 3) = 3 \cdot 2 \cdot 1 = 6$

123 132
213 231
312 321

Beispiel 11.14 (Rennsportwette)

Wie groß ist die Gewinnwahrscheinlichkeit einer Viererwette (1., 2., 3. und 4. Platz müssen richtig vorhergesagt werden) bei 10 Kamelen?



- Nach Laplace: $\frac{\text{Anzahl Gewinnplatzierungen}}{\text{Anzahl möglicher Platzierungen}}$
- Die Reihenfolge der Ziehung spielt eine Rolle \Rightarrow Variation
- Wiederholung ist nicht möglich
- $n = 10$ Elemente, $k = 4$ Ziehungen $\Rightarrow V(10, 4) = \frac{10!}{6!} = 5040$
- 1 Gewinnplatzierung $\Rightarrow P(\text{Gewinn}) = \frac{1}{5040} \approx 0,02\%$

Kombinationen

Von n Elementen zur k -ten Klasse (Ordnung):

- jede Zusammenstellung von k Elementen aus n Elementen
- ohne Berücksichtigung ihrer Anordnung
- alle n Elemente der Ausgangsmenge sind verschieden

Mit Wiederholung

Anzahl der Kombinationen: $K^W(n, k) = \binom{n+k-1}{k}$

Mit Wiederholung meint hier, dass ein Element mehrmals ausgewählt werden kann.

Ohne Wiederholung

Anzahl der Kombinationen: $K(n, k) = \frac{V(n, k)}{P(k)} = \frac{n!}{k!(n-k)!} = \binom{n}{k}$

Ohne Wiederholung meint hier, dass jedes Element nur genau einmal ausgewählt werden kann.

Beispiel 11.15 (Kombination mit Wiederholung)

$n = 3$ Elemente: 1, 2, 3

- $k = 1$, Anzahl: $K^W(3, 1) = 3$

1 2 3

- $k = 2$, Anzahl: $K^W(3, 2) = 3 + 2 + 1 = 6$

11 12 13

21 22 23

31 32 33

Beispiel 11.16 (Kombination ohne Wiederholung)

$n = 3$ Elemente: 1, 2, 3

- $k = 1$, Anzahl: $K(3, 1) = 3$

1 2 3

- $k = 2$, Anzahl: $K(3, 2) = V(3, 2)/P(2) = 6/2 = 3$

11 12 13

21 22 23

31 32 33

- $k = 3$, Anzahl: $K(3, 3) = V(3, 3)/P(3) = 3/3 = 1$

123

Beispiel 11.17 (Lottozahlen)

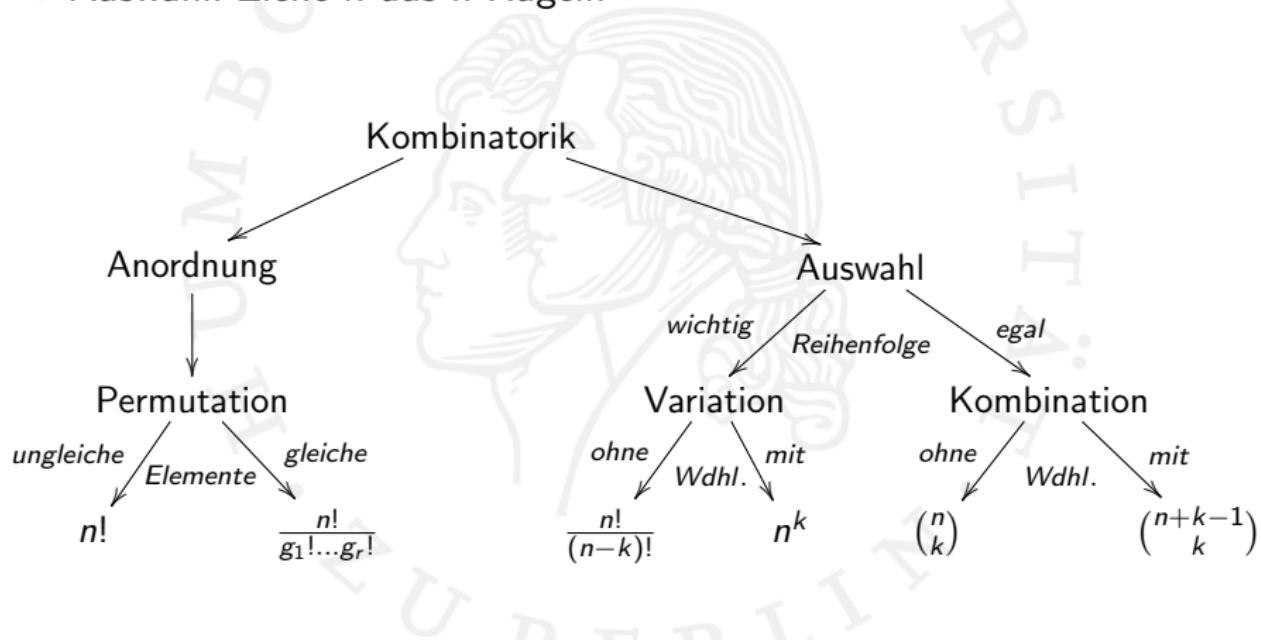
Wie groß ist die Wahrscheinlichkeit von 6 Richtigen beim Lotto "6 aus 49"?



- Nach Laplace: $\frac{\text{Anzahl Gewinnziehung}}{\text{Anzahl möglicher Ziehungen}}$
- Die Reihenfolge der Ziehung spielt keine Rolle \Rightarrow Kombination
- Wiederholung ist nicht möglich
- $n = 49$ Elemente, $k = 6$ Ziehungen $\Rightarrow K(49, 6) = \binom{49}{6} = 13.983.816$
- 1 Gewinnziehung $\Rightarrow P(\text{Gewinn}) = \frac{1}{13.983.816} \approx 0,0000007\%$

Übersicht Kombinatorik

- Urnenmodell mit n Kugeln
- Anordnung: Ordne alle n Kugeln an
- Auswahl: Ziehe k aus n Kugeln



R Listing 11.1: example_combinatorics.R

```
1 # 7 Fakultaet
2 # Anzahl Permutationen mit ungleichen Elementen
3 factorial(7)
4 # Logarithmus 7 Fakultaet
5 lfactorial(7)
6 # Binomialkoeffizient 10 ueber 3
7 # Anzahl Kombinationen ohne Wiederholung
8 choose(10, 3)
9 # Logarithmus des Binomialkoeffizient 10 ueber 3
10 lchoose(10, 3)
11 # Kombinationen ohne Wiederholung (3 aus 10)
12 combn(10, 3)
13 # Permutation mit 3 ungleichen Elementen
14 #install.packages("combinat")
15 library("combinat")
16 permn(3)
```

Wahrscheinlichkeitsbegriffe

Statistischer Wahrscheinlichkeitsbegriff nach von Mises

- Folge voneinander unabhängiger Versuche
 - ▶ unter identischen Bedingungen
 - ▶ beliebig oft wiederholbar
- n -malige Wiederholung des Zufallsexperimentes

Die Wahrscheinlichkeit $P(A)$ des Ereignisses A ist als der Grenzwert der relativen Häufigkeit des Auftretens $f_n(A)$ von A definiert:

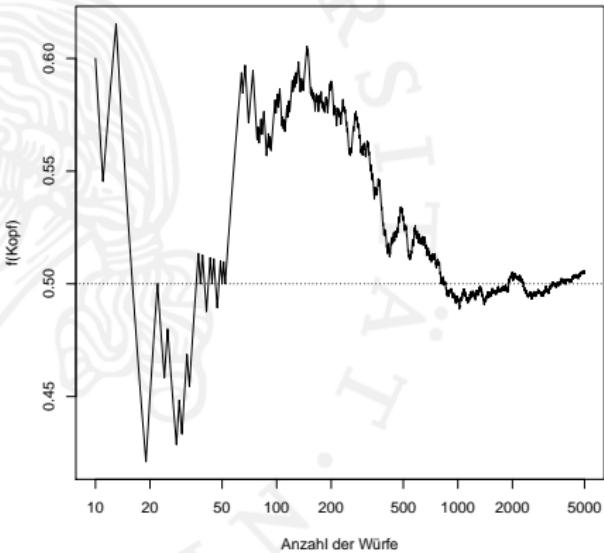
$$P(A) = \lim_{n \rightarrow \infty} f_n(A).$$

Wegen $0 \leq f_n(A) \leq 1$ gilt: $0 \leq P(A) \leq 1$

Beispiel 11.18

Zufallsexperiment: Werfen einer Münze, $A = \text{„Kopf“}$

n	$h(\text{Kopf})$	$f(\text{Kopf})$
10	6	0,600
20	9	0,450
40	20	0,500
80	45	0,562
100	58	0,580
200	118	0,590
1000	493	0,493
5000	2528	0,506



Axiomatischer Wahrscheinlichkeit nach Kolmogorow

Ein Wahrscheinlichkeitsmaß P ist eine Abbildung, die

- allen Ereignissen $A \subset S$ eines Ereignisraumes S (d.h. eines gegebenen Zufallsexperimentes) eine Zahl $P(A)$ zuordnet und
- folgende Bedingungen (Eigenschaften, Axiome) erfüllt:

Axiom 1: (Nichtnegativität)

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

Axiom 2: (Normierung)

$$P(S) = 1$$

Axiom 3: (Additivität) Wenn $A \cap B = \emptyset$, dann gilt

$$P(A \cup B) = P(A) + P(B)$$

Beispiel 11.19

Zufallsexperiment: Werfen einer nicht idealen Münze

Ereignisraum: $S = \{\text{Kopf}, \text{Zahl}\}$

$$P(\{\text{Kopf}\}) = 0,3$$

$$P(\{\text{Zahl}\}) = 0,7$$

Dann gilt:

- Axiom 1: $0 \leq P(\{\text{Kopf}\}) \leq 1$
 $0 \leq P(\{\text{Zahl}\}) \leq 1$
- Axiom 2: $P(S) = 1$
- Axiom 3: $P(\{\text{Kopf}\} \cup \{\text{Zahl}\}) = P(S) = 1 = P(\{\text{Kopf}\}) + P(\{\text{Zahl}\})$

Rechenregeln

$A, B, A_1, A_2, \dots \subset S$ seien Ereignisse und P ein Wahrscheinlichkeitsmaß zu einem betrachteten Zufallsexperiment

1. $P(\bar{A}) = 1 - P(A)$, wobei \bar{A} das Komplement von A ist
2. $P(\emptyset) = 1 - P(S) = 0$
3. Wenn $A \cap B = \emptyset$ gilt, folgt $P(A \cap B) = P(\emptyset) = 0$
4. Wenn $A \subset B$ gilt, folgt $P(A) \leq P(B)$

$$B = A \cup (B \setminus A)$$

A und $B \setminus A$ sind disjunkt

$$\Rightarrow P(B) = P(A) + P(B \setminus A)$$

$$P(B \setminus A) \geq 0 \quad \Rightarrow P(B) \geq P(A)$$

5. Für A_1, A_2, \dots mit $A_i \cap A_j = \emptyset$ ($i \neq j$) folgt

$$P(A_1 \cup A_2 \cup \dots) = P(A_1) + P(A_2) + \dots$$

6. $P(A \setminus B) = P(A) - P(A \cap B)$

Additionssatz

Sind A und B zwei beliebige Ereignisse eines Zufallsexperimentes, dann ist die Wahrscheinlichkeit des Ereignisses $A \cup B$:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Sind A und B disjunkt ($A \cap B = \emptyset$) dann gilt

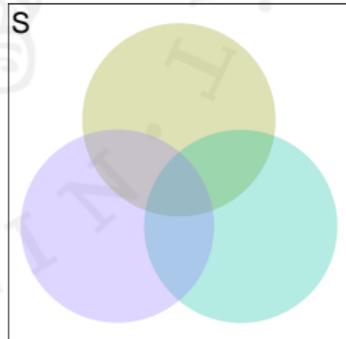
$$\Rightarrow P(A \cap B) = 0$$

$$\Rightarrow P(A \cup B) = P(A) + P(B)$$

Verallgemeinerung auf drei Ereignisse A , B und C :

$$P(A \cup B \cup C) =$$

$$\begin{aligned} & P(A) + P(B) + P(C) \\ & - P(A \cap B) - P(A \cap C) - P(B \cap C) \\ & + P(A \cap B \cap C) \end{aligned}$$



Beispiel 11.20 (Kartenspiel mit 32 Karten)

$A = \text{„Dame“}$

$B = \text{„Herz“}$

$A \cap B = \text{„Herzdame“}$

Nach der klassischen Definition der Wahrscheinlichkeit

$$P(A) = P(\text{„Dame“}) = 4/32$$

$$P(B) = P(\text{„Herz“}) = 8/32$$

$$P(A \cap B) = P(\text{„Herzdame“}) = 1/32$$

Anwendung des Additionssatzes:

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= 4/32 + 8/32 - 1/32 \\ &= 11/32 \end{aligned}$$

Bedingte Wahrscheinlichkeit

Wahrscheinlichkeit des Eintretens des Ereignisses A unter der Bedingung,
dass das Ereignis B bereits eingetreten ist:

bedingte Wahrscheinlichkeit $P(A|B)$

Beispiel 11.21 (Einmaliges Werfen eines Würfels)

$$P(\text{"Augenzahl}=6") = 1/6$$

$$P(\text{"Augenzahl}=6" | \text{"Augenzahl ist gerade"}) = 1/3$$

 Der Begriff der Bedingung ist nicht an einen zeitlichen Zusammenhang gebunden!

Gegeben seien die Ereignisse A und B eines Ereignisraumes S :

- Bedingte Wahrscheinlichkeit des Ereignisses A unter der Bedingung B :

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad P(B) > 0$$

Gegeben seien die Ereignisse A_1 , A_2 und A_3 eines Ereignisraumes S :

- Bedingte Wahrscheinlichkeit des Ereignisses A_1 unter der Bedingung A_2 und A_3 :

$$P(A_1|A_2 \cap A_3) = \frac{P(A_1 \cap A_2 \cap A_3)}{P(A_2 \cap A_3)} \quad P(A_2 \cap A_3) > 0$$

Beachte: $P(A|B) \neq P(B|A)$

Beispiel 11.22 (Einmaliges Werfen eines Würfels)

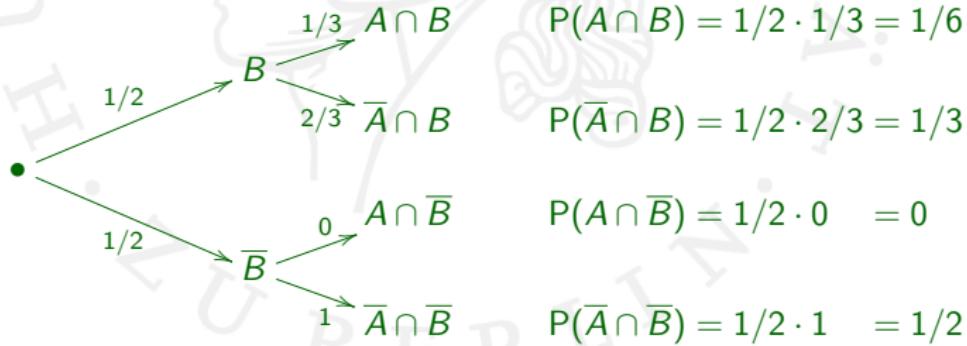
$A = \text{"Augenzahl}=5"$ und $B = \text{"Augenzahl ist größer als 3"}$

$$P(A) = 1/6$$

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

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{1/2} = 1/3$$

Wahrscheinlichkeitsbaum:



Multiplikationssatz

Für zwei Ereignisse A, B gilt:

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

Für drei Ereignisse A_1, A_2, A_3

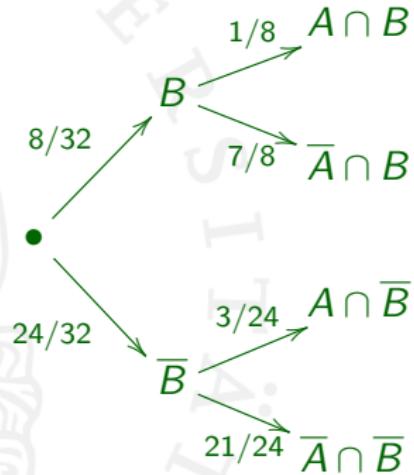
$$P(A_1 \cap A_2 \cap A_3) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2)$$

Für n Ereignisse A_1, \dots, A_n

$$\begin{aligned} P(A_1 \cap \dots \cap A_n) &= P(A_1) \cdot P(A_2|A_1) \\ &\quad \cdot P(A_3|A_1 \cap A_2) \cdot \dots \\ &\quad \cdot P(A_n|A_1 \cap \dots \cap A_{n-1}) \end{aligned}$$

Beispiel 11.23 (Kartenspiel mit 32 Karten)

- $A = \text{„Dame“}$
- $B = \text{„Herz“}$
- $P(A) = P(\text{„Dame“}) = 4/32$
- $P(B) = P(\text{„Herz“}) = 8/32$
- $P(A|B) = P(\text{„Dame“}|\text{„Herz“}) = 1/8$



Anwendung des Multiplikationssatzes:

$$P(A \cap B) = P(\text{„Herzdame“}) = P(B) P(A|B) = 8/32 \cdot 1/8 = 1/32$$

Beispiel 11.24 (Einmaliges Werfen eines Würfels)

A_1 = „gerade Augenzahl“, A_2 = „Augenzahl größer 2“, A_3 = „Augenzahl kleiner 5“

$$P(A_3) = P(\text{„Augenzahl kleiner 5“}) = 2/3$$

$$P(A_2|A_3) = P(\text{„Augenzahl größer 2“} | \text{„Augenzahl kleiner 5“}) = 1/2$$

$$P(A_1|A_2 \cap A_3) = P(\text{„gerade Augenzahl“} | \text{„Augenzahl zwischen 2 und 5“}) = 1/2$$

Anwendung des Multiplikationssatzes:

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= P(\text{„gerade Augenzahl zwischen 2 und 5“}) \\ &= P(A_3) P(A_2|A_3) P(A_1|A_2 \cap A_3) = 2/3 \cdot 1/2 \cdot 1/2 = 1/6 \end{aligned}$$

Das Ziegenproblem

Beispiel 11.25 (Ziegenproblem)

Der Kandidat hat Tor 1 gewählt, und der Moderator hat daraufhin Tor 3 geöffnet. Sollte der Kandidat das Tor wechseln? Wie groß ist die Wahrscheinlichkeit, dass das Auto hinter Tor 2 ist?

Ereignisse:

G_i : Der Gewinn ist hinter Tor i ($i = 1, 2, 3$)

M_j : Der Moderator hat Tor j geöffnet ($j = 1, 2, 3$)

⇒ Gesucht ist die bedingte Wahrscheinlichkeit $P(G_2|M_3)$, dass das Auto hinter Tor 2 ist, wenn bekannt ist, dass es nicht hinter Tor 3 ist.

Bekannt, wenn der Kandidat auf Tor 1 gezeigt hat:

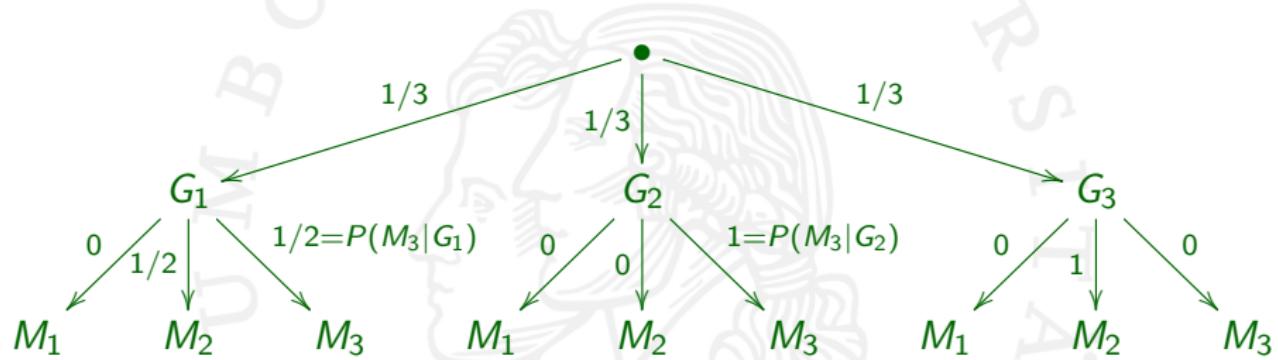
- $P(G_1) = P(G_2) = P(G_3) = 1/3$
- $P(M_3|G_2) = 1$
- $P(M_3|G_1) = 1/2$
- $P(M_3) = \frac{1}{2}$

$$P(G_2|M_3) = \frac{P(G_2 \cap M_3)}{P(M_3)} = \frac{P(M_3|G_2)P(G_2)}{P(M_3)} = \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2}} = 2/3$$

$$P(G_1|M_3) = \frac{P(G_1 \cap M_3)}{P(M_3)} = \frac{P(M_3|G_1)P(G_1)}{P(M_3)} = \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2}} = 1/3$$

Der Kandidat sollte also wechseln, um seine Gewinnchancen von anfangs $1/3$ auf nun $2/3$ zu verdoppeln.

Kandidat hat auf Tor 1 gezeigt



$$P(M_3) = \frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot 1 + \frac{1}{3} \cdot 0 = \frac{1}{6} + \frac{1}{3} + 0 = \frac{1}{2}$$

Let's Make a Deal

Make a deal

Specify speed

Options

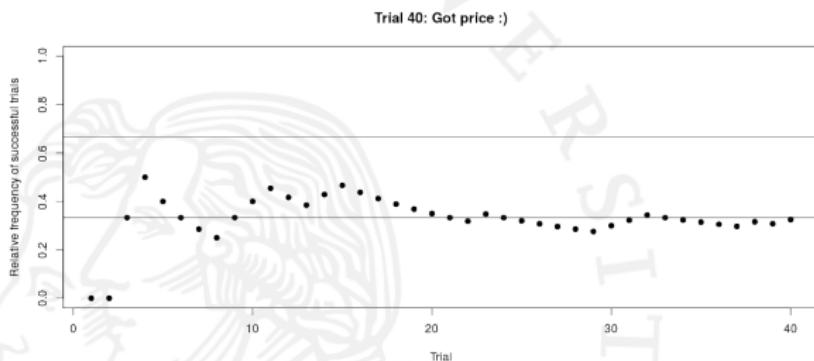
1. Guest points

- to left door
- to middle door
- to right door
- randomly to a door

2. Host opens door with a goat

3. Guest decides to

- keep door
- change door



3.



http://u.hu-berlin.de/men_hall

Unabhängige Ereignisse

Ein Ereignis A ist dann von einem Ereignis B stochastisch unabhängig, wenn das Eintreten des Ereignisses A

- nicht von dem Eintreten
- nicht von dem Nichteintreten

des Ereignisses B abhängt:

$$P(A|B) = P(A|\bar{B}) \text{ und } P(B|A) = P(B|\bar{A})$$

$$\Leftrightarrow P(A|B) = P(A) \text{ und } P(B|A) = P(B) \Leftrightarrow P(A \cap B) = P(A) \cdot P(B)$$



unabhängig \neq disjunkt

Unabhängigkeit von mehr als zwei Ereignissen

Die Ereignisse A_1, A_2, \dots, A_n heißen stochastisch unabhängig, wenn für jede Auswahl A_{i1}, \dots, A_{im} mit $m \leq n$ gilt

$$P(A_{i1} \cap \dots \cap A_{im}) = P(A_{i1}) \cdot P(A_{i2}) \cdot \dots \cdot P(A_{im})$$

Beispiel 11.26 (Vorlesung)

Gegeben seien die Wahrscheinlichkeiten, dass einer, zwei oder alle drei der Studenten X, Y und Z nicht zur Vorlesung erscheinen:

E_i	{-}	{X}	{Y}	{Z}	{XY}	{XZ}	{YZ}	{XYZ}
$P(E_i)$	0,751	0,100	0,063	0,061	0,011	0,008	0,005	0,001

Wir können verschiedene Wahrscheinlichkeiten berechnen:

$$P(X \text{ nicht da}) = 0,100 + 0,011 + 0,008 + 0,001 = 0,12$$

$$P(Y \text{ nicht da}) = 0,063 + 0,011 + 0,005 + 0,001 = 0,08$$

$$P(X \text{ und } Y \text{ nicht da}) = P(XY) + P(XYZ) = 0,012$$

$$\begin{aligned} P(X \text{ oder } Y \text{ nicht da}) &= P(X) + P(Y) + P(XY) + P(XZ) + P(YZ) \\ &\quad + P(XYZ) = 0,188 \end{aligned}$$

$$\begin{aligned} P(X \text{ oder } Y \text{ nicht da}) &= P(X \text{ nicht da}) + P(Y \text{ nicht da}) \\ &\quad - P(X \text{ und } Y \text{ nicht da}) \\ &= 0,12 + 0,08 - 0,012 = 0,188 \text{ (Additionssatz)} \end{aligned}$$

Ist das Erscheinen der Studenten X und Y unabhängig voneinander ?

Nein, da gilt

$$P(X \text{ und } Y \text{ nicht da}) = 0,012$$

$$\begin{aligned} P(X \text{ nicht da}) P(Y \text{ nicht da}) &= 0,12 \times 0,08 \\ &= 0,0096 \end{aligned}$$

$$P(X \text{ und } Y \text{ nicht da}) \neq P(X \text{ nicht da}) P(Y \text{ nicht da})$$

Ist das Erscheinen aller drei Studenten unabhängig voneinander ?

Nein, da das Erscheinen der Studenten X und Y nicht unabhängig voneinander ist.

Totale Wahrscheinlichkeit

Beispiel 11.27 (Weinkeller)

Im Weinkeller lagern:

- 10 Flaschen Deutscher Wein, Weißweinanteil: 1/5
- 6 Flaschen Französischer Wein, Weißweinanteil: 1/3
- 4 Flaschen Spanischer Wein, Weißweinanteil: 1/4

Wahrscheinlichkeiten für die Herkunft des Weins:

$$A_1 = \{\text{Deutscher Wein}\} \quad P(A_1) = 0,5$$

$$A_2 = \{\text{Französischer Wein}\} \quad P(A_2) = 0,3$$

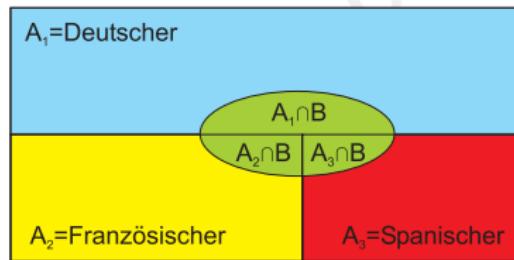
$$A_3 = \{\text{Spanischer Wein}\} \quad P(A_3) = 0,2$$

⇒ vollständige Zerlegung von $S \rightarrow A_1 \cup A_2 \cup A_3 = S$

Wie groß ist die Wahrscheinlichkeit, dass eine zufällig ausgewählte Weinflasche Weißwein ist?

Gesucht: $P(B) = P(\text{Weißwein})$

$$A_1 \cap A_2 = \emptyset, \quad A_1 \cap A_3 = \emptyset, \quad A_2 \cap A_3 = \emptyset$$



$$P(B|A_1) = 1/5$$

$$P(B|A_2) = 1/3$$

$$P(B|A_3) = 1/4$$

$$B = (B \cap A_1) \cup (B \cap A_2) \cup (B \cap A_3)$$

$$\begin{aligned} P(B) &= P[(A_1 \cap B) \cup (A_2 \cap B) \cup (A_3 \cap B)] \\ &= P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B) \end{aligned}$$

unbekannt: $P(A_i \cap B)$, $i = 1, 2, 3$;

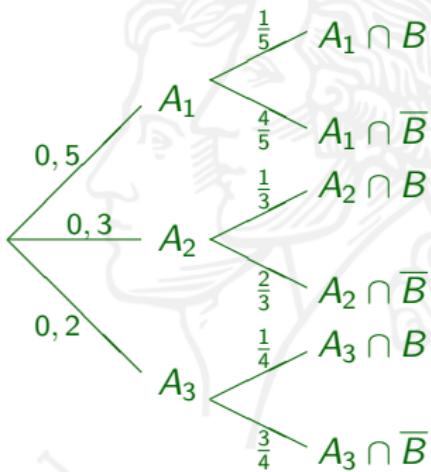
aber: $P(A_i \cap B) = P(B|A_i) P(A_i)$

⇒ Multiplikationssatz für beliebige Ereignisse

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3)$$

$$P(B) = 1/5 \cdot 0,5 + 1/3 \cdot 0,3 + 1/4 \cdot 0,2$$

$$= 0,1 + 0,1 + 0,05 = 0,25$$



Satz:

Die Ereignisse A_1, A_2, \dots, A_n bilden eine vollständige Zerlegung des Ereignisraumes S :

- $A_i \cap A_j = \emptyset$ für $i, j = 1, \dots, n, i \neq j$
- $A_1 \cup A_2 \cup \dots \cup A_n = S$
- $P(A_i) > 0$ für $i = 1, \dots, n$

Dann gilt für ein beliebiges Ereignis $B \subset S$ mit $P(B) > 0$

$$\begin{aligned} P(B) &= P(A_1 \cap B) + P(A_2 \cap B) + \dots + P(A_n \cap B) \\ &= \sum_{i=1}^n P(A_i \cap B) \\ &= P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \dots + P(B|A_n)P(A_n) \\ &= \sum_{i=1}^n P(B|A_i)P(A_i) \end{aligned}$$

Theorem von Bayes

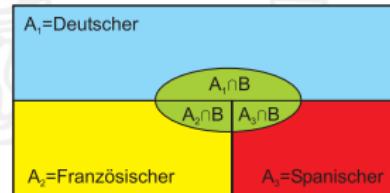
Beispiel 11.28 (Weinkeller)

Wie groß ist die Wahrscheinlichkeit, dass eine zufällig ausgewählte Weißweinflasche Deutscher Wein ist?

Gesucht: $P(A_1|B) = P(\text{Deutscher Wein}|Weißwein)$

$$P(A_1|B) = \frac{P(A_1 \cap B)}{P(B)} = \frac{P(B|A_1) P(A_1)}{P(B)}$$

$$P(B) = \sum_{i=1}^N P(B|A_i) P(A_i)$$



$$\Rightarrow P(A_1|B) = \frac{P(B|A_1) P(A_1)}{\sum_{i=1}^N P(B|A_i) P(A_i)} = \frac{0,2 \cdot 0,5}{0,25} = \frac{0,1}{0,25} = 0,4$$

Theorem:

Die Ereignisse A_1, A_2, \dots, A_n bilden eine vollständige Zerlegung des Ereignisraumes S :

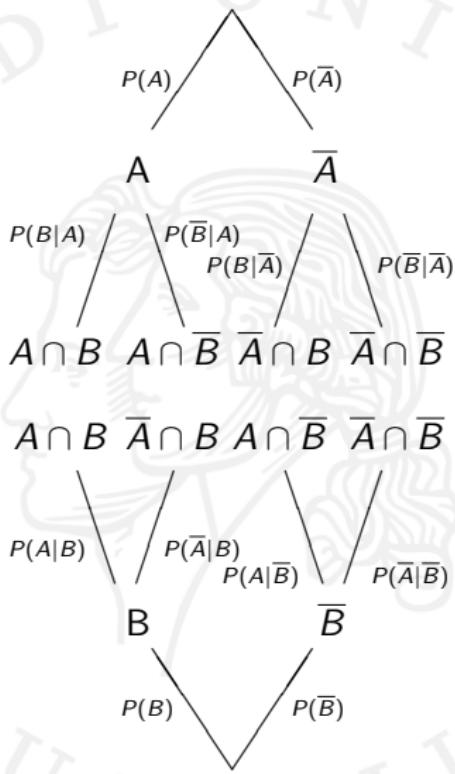
- $A_i \cap A_j = \emptyset$ für $i, j = 1, \dots, n, i \neq j$
- $A_1 \cup A_2 \cup \dots \cup A_n = S$
- $P(A_i) > 0$ für $i = 1, \dots, n$

Ferner sei ein zufälliges Ereignis B mit $P(B) > 0$ und $P(B|A_1), \dots, P(B|A_n)$ gegeben.

Dann gilt

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_i P(B|A_i)P(A_i)} \quad \forall j = 1, \dots, n$$

$P(A_j|B)$ a-posteriori-Wahrscheinlichkeit
 $P(A_j)$ a-priori-Wahrscheinlichkeit



Spam E-Mail

Beispiel 11.29

- Frage: Wie kann man Spam E-Mails in E-Mail-Programmen heraus filtern?
- Jede E-Mail ist entweder Spam oder nicht, also Ereignis $E = \{\text{E-Mail ist Spam}\}$ bzw. $\bar{E} = \{\text{E-Mail ist kein Spam}\}$
- Jede E-Mail besteht aus n Worten: W_1, \dots, W_n
- Gesucht $P(E|W_1 \cap \dots \cap W_n) = ?$
- Bekannt aus alten E-Mails ist
 - ▶ $P(E)$ - Gesamter Anteil der Spam-E-Mails in der Vergangenheit
 - ▶ $P(W_i|E)$ - Geschätzte Wahrscheinlichkeit mit der Wort W_i in der Vergangenheit in Spam-E-Mails enthalten war

- Klassifiziere eine E-Mail als Spam-E-Mail, wenn

$$\frac{P(E|W_1 \cap \dots \cap W_n)}{P(\bar{E}|W_1 \cap \dots \cap W_n)} > c \text{ z.B. mit } c = 10$$

⇒

$$\frac{P(E|W_1 \cap \dots \cap W_n)}{P(\bar{E}|W_1 \cap \dots \cap W_n)} = \frac{\frac{P(E \cap W_1 \cap \dots \cap W_n)}{P(W_1 \cap \dots \cap W_n)}}{\frac{P(\bar{E} \cap W_1 \cap \dots \cap W_n)}{P(W_1 \cap \dots \cap W_n)}}$$

- Die Wahrscheinlichkeit $P(W_1 \cap \dots \cap W_n)$ muss nicht angegeben werden, da sie sich herauskürzt

- Es gilt

$$\begin{aligned} P(E \cap W_1 \cap \dots \cap W_n) &= P(W_1 \cap \dots \cap W_n | E)P(E) \\ P(\overline{E} \cap W_1 \cap \dots \cap W_n) &= P(W_1 \cap \dots \cap W_n | \overline{E})P(\overline{E}) \end{aligned}$$

- Annahme: Die Worte treten in Spam- und Nicht-Spam-E-Mails unabhängig voneinander auf

$$\begin{aligned} P(W_1 \cap \dots \cap W_n | E) &= P(W_1 | E) \dots P(W_n | E) \\ P(W_1 \cap \dots \cap W_n | \overline{E}) &= P(W_1 | \overline{E}) \dots P(W_n | \overline{E}) \end{aligned}$$

- Klassifiziere eine E-Mail als Spam-E-Mail, wenn

$$\frac{P(E | W_1 \cap \dots \cap W_n)}{P(\overline{E} | W_1 \cap \dots \cap W_n)} = \frac{P(W_1 | E) \dots P(W_n | E)}{P(W_1 | \overline{E}) \dots P(W_n | \overline{E})} \frac{P(E)}{P(\overline{E})} > c$$

Beweis: Binomialkoeffizienten

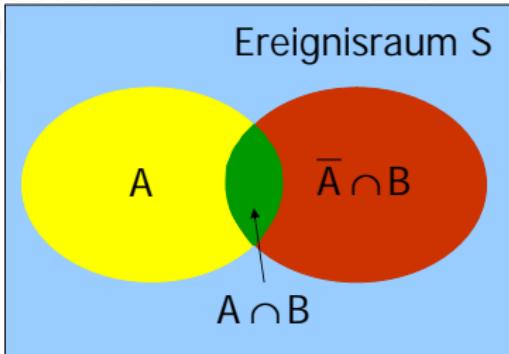
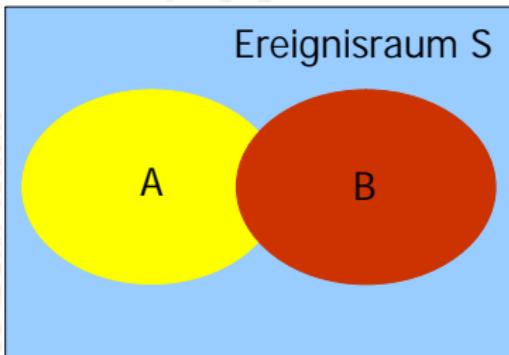
$$\begin{aligned}
 \binom{n}{k} + \binom{n}{k+1} &= \frac{n!}{k!(n-k)!} + \frac{n!}{(k+1)!\{n-(k+1)\}!} \\
 &= \frac{(k+1)}{(k+1)k!} \frac{n!}{(n-k)!} + \frac{n!}{(k+1)!} \frac{(n-k)}{\{n-(k+1)!\}(n-k)} \\
 &= \frac{(k+1)n!}{(k+1)!(n-k)!} + \frac{n!(n-k)}{(k+1)!(n-k)!} \\
 &= \frac{n!\{(k+1)+(n-k)\}}{(k+1)!(n-k)!} \\
 &= \frac{n!(n+1)}{(k+1)!\{(n+1)-(k+1)\}!} \\
 &= \frac{(n+1)!}{(k+1)!\{(n+1)-(k+1)\}!} = \binom{n+1}{k+1}
 \end{aligned}$$

Beweis: Additionssatz

- $B = (A \cap B) \cup (\bar{A} \cap B)$
 $(A \cap B)$ und $(\bar{A} \cap B)$ sind disjunkt

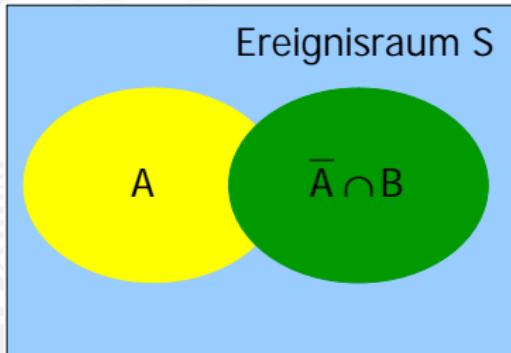
$$\begin{aligned}P(B) &= P[(A \cap B) \cup (\bar{A} \cap B)] \\&= P(A \cap B) + P(\bar{A} \cap B)\end{aligned}$$

$$P(\bar{A} \cap B) = P(B) - P(A \cap B)$$

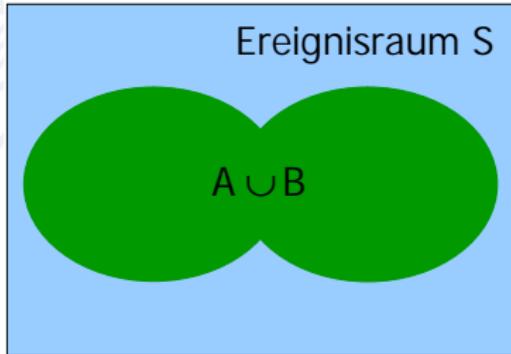


2. $A \cup B = A \cup (\bar{A} \cap B)$
 A und $(\bar{A} \cap B)$ sind disjunkt

$$\begin{aligned}P(A \cup B) &= P[A \cup (\bar{A} \cap B)] \\&= P(A) + P(\bar{A} \cap B)\end{aligned}$$



3. $P(A \cup B) =$
 $P(A) + P(B) - P(A \cap B)$



$$P(A \cup B) = P[(A \cap B) \cup (A \cap \bar{B}) \cup (\bar{A} \cap B)]$$

$$= P(A \cap B) + P(A \cap \bar{B}) + P(\bar{A} \cap B)$$

$$P(A) = P[(A \cap B) \cup (A \cap \bar{B})]$$

$$= P(A \cap B) + P(A \cap \bar{B})$$

$$P(B) = P[(A \cap B) \cup (\bar{A} \cap B)]$$

$$= P(A \cap B) + P(\bar{A} \cap B)$$

$$P(A) + P(B) = P(A \cap B) + P(A \cap \bar{B})$$

$$+ P(A \cap B) + P(\bar{A} \cap B)$$

$$= P(A \cap B) + P(A \cup B)$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Beweis: Unabhängige Ereignisse

Zu $P(A|B) = P(A|\bar{B})$

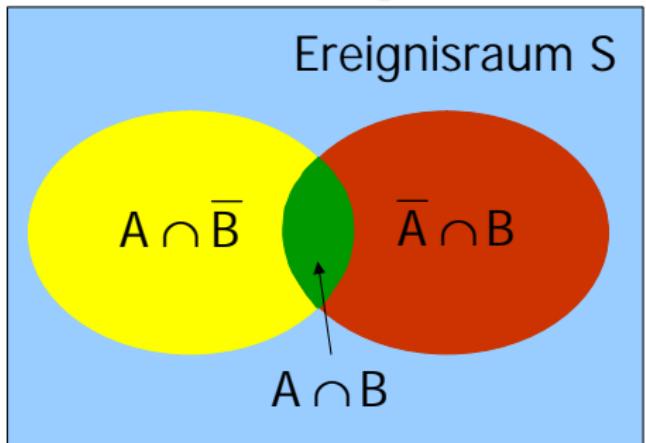
Ist A von B unabhängig, so gilt

$$P(A|B) = P(A|\bar{B}) = \frac{P(A \cap \bar{B})}{P(\bar{B})} = \frac{P(A \cap \bar{B})}{1 - P(B)}$$

$$P(A|B) \cdot [1 - P(B)] = P(A \cap \bar{B})$$

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

$$\begin{aligned} P(A|B) &= P(A|B) \cdot P(B) + P(A \cap \bar{B}) \\ &= \frac{P(A \cap B)}{P(B)} \cdot P(B) + P(A \cap \bar{B}) \\ &= P(A \cap B) + P(A \cap \bar{B}) = P(A) \end{aligned}$$



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

Ist A von B unabhängig, so gilt

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

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

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

Zufallsvariablen

5. November 2022

- Zufallsexperiment und -variable
- Zufallsvariablen und ihre Verteilungen
- Verteilung diskreter Zufallsvariablen
- Verteilung stetiger Zufallsvariablen
- Parameter von Zufallsvariablen
- Erwartungswert
- Varianz
- Standardisierung
- Tschebyscheff Ungleichung
- Diskrete Zufallsvariablen
- Stetige Zufallsvariablen
- Verteilungsfunktion zweier Zufallsvariablen
- Randverteilungen
- Bedingte Verteilungen
- Unabhängigkeit
- Kovarianz
- Theoretischer Korrelationskoeffizient
- Variablen vs. Zufallsvariablen
- Zusammenfassung

Definition

- Induktive Statistik
 - ▶ Rückschluss von der Stichprobe auf die Grundgesamtheit
 - ▶ Problem: Verteilung einer Variablen in der Grundgesamtheit ist unbekannt
- Zufallsvariable X
 - ▶ theoretisches Äquivalent zu einer Variablen X
 - ▶ hat die gleichen Merkmalsausprägungen
 - ▶ hat das gleiche Skalenniveau wie X
- Problem: Bestimme die Verteilung der Zufallsvariablen X
 - ▶ Lösung: Mache Annahme über die Verteilung einer Variablen der Grundgesamtheit \Rightarrow z.B. bestimme Wahrscheinlichkeiten ("relative Häufigkeiten") für jede Merkmalsausprägung
 - ▶ mittels historischer Daten
 - ▶ mittels Zufallsexperimente

Zufallsexperiment und -variable

Zufallsexperiment

- ein unter gleichen Bedingungen wiederholbarer Vorgang
- z.B. naturwissenschaftlicher Versuch, sozialwissenschaftliche Beobachtung oder Befragung
- Ergebnis oder Ausgang nicht mit Sicherheit vorhersagbar

Zufallsvariable

- Zufallsvariable X ist eine Abbildung, die den möglichen Ereignissen eines Zufallsexperiments (Ω) reelle Zahlen (\mathbb{R}) zuordnet ($X : \Omega \rightarrow \mathbb{R}$)

Ω : Definitionsbereich der Abbildung (Ereignisraum)

\mathbb{R} : Wertebereich der Abbildung

Beispiel 12.1 (3× Werfen einer “idealen“ Münze)

- mit Kopf (K) und Zahl (Z)
- Ereignisraum Ω :

$$\Omega = \{KKK, KKZ, KZK, ZKK, KZZ, ZKZ, ZZK, ZZZ\}$$

- genau einmal Z: $\{(KKZ) \cup (KZK) \cup (ZKK)\}$

Ereignis $\in \Omega$	Anzahl Z $\in \mathbb{R}$
KKK	0
KKZ	1
KZK	
ZKK	
KZZ	2
ZKZ	
ZZK	
ZZZ	3

jedem Elementarereignis $E \in \Omega$ wird
eine reelle Zahl zugeordnet

Vor der Durchführung des Zufallsexperiments:

- Zufallsvariable X mit zugehörigem Wertebereich
- die verschiedenen möglichen Werte der Zufallsvariablen treten mit bestimmten Wahrscheinlichkeiten ein
- Gesamtzahl der Merkmalsausprägungen: k
- sich unterscheidende Merkmalsausprägungen: x_j ($j = 1, \dots, k$)

Nach der Durchführung des Zufallsexperiments:

- Realisation x
- Wert, den eine Zufallsvariable X bei der Durchführung des Zufallsexperiments konkret angenommen hat
- Stichprobe
 - ▶ Gesamtzahl der Beobachtungen: n
 - ▶ Beobachtungswerte: x_i ($i = 1, \dots, n$)

Zufallsvariablen und ihre Verteilungen

Wahrscheinlichkeitsfunktion

- ordnet jedem möglichen Wert der Zufallsgröße eine Wahrscheinlichkeit zu

Verteilungsfunktion $F(x)$ einer Zufallsvariablen X

- Funktion, die die Wahrscheinlichkeit dafür angibt, dass die Zufallsvariable X höchstens den Wert x annimmt

$$F(x) = P(X \leq x)$$

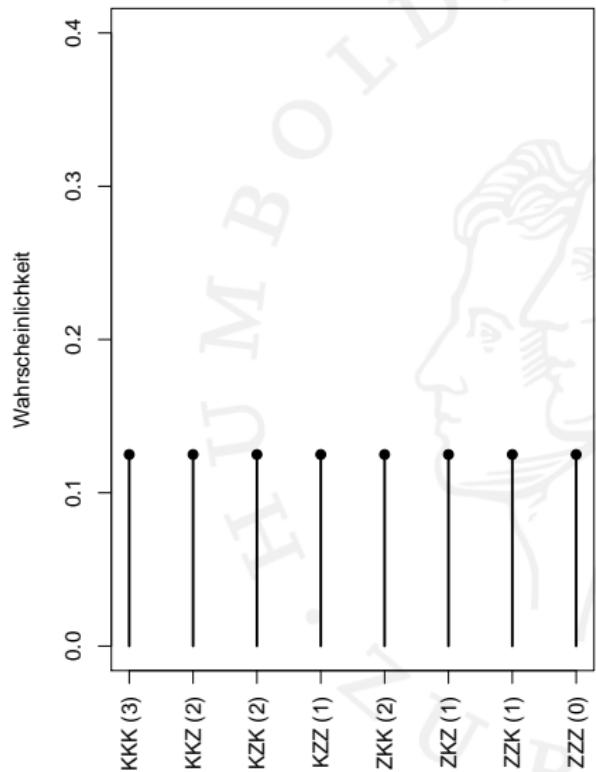
Beispiel 12.2 (Münze)

- Zufallsexperiment: 3× Werfen einer idealen Münze mit Kopf (K) und Zahl (Z)
- Zufallsvariable X : „Anzahl von Z beim dreimaligen Werfen der Münze“

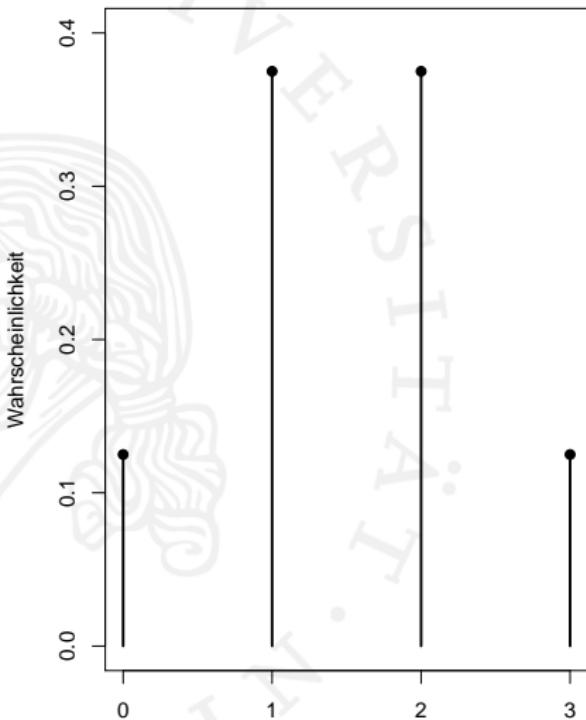
Elementar-ereignis E_j	Wahrscheinlichkeit $P(E_j)$	Anzahl der Z x_i	Wahrscheinlichkeitsfunktion $P(X = x_i) = f(x_i)$
$E_1 = \{KKK\}$	$P(E_1) = 0,125$	$x_1 = 0$	$f(x_1) = 0,125$
$E_2 = \{KKZ\}$	$P(E_2) = 0,125$	$x_2 = 1$	$f(x_2) = 0,375$
$E_3 = \{KZK\}$	$P(E_3) = 0,125$		
$E_4 = \{ZKK\}$	$P(E_4) = 0,125$		
$E_5 = \{KZZ\}$	$P(E_5) = 0,125$	$x_3 = 2$	$f(x_3) = 0,375$
$E_6 = \{ZKZ\}$	$P(E_6) = 0,125$		
$E_7 = \{ZZK\}$	$P(E_7) = 0,125$		
$E_8 = \{ZZZ\}$	$P(E_8) = 0,125$	$x_4 = 3$	$f(x_4) = 0,125$

$$P(E_j) = 1/2 \cdot 1/2 \cdot 1/2 = 1/8$$

Elementarereignisse



Anzahl Koepfe



Verteilung diskreter Zufallsvariablen

Verteilungsfunktion $F(x)$

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i) \quad \text{mit } f(x_i) = P(X = x_i)$$

Wahrscheinlichkeitsfunktion $f(x)$

$$P(X = x_i) = f(x_i)$$

Eigenschaften

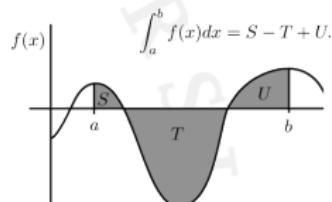
$$f(x_i) \geq 0, \quad \sum_{i=1}^{\infty} f(x_i) = 1$$

Die Wahrscheinlichkeitsfunktion gibt an, mit welcher Wahrscheinlichkeit die Zufallsvariable X **genau** den Wert x_i annimmt

Verteilung stetiger Zufallsvariablen

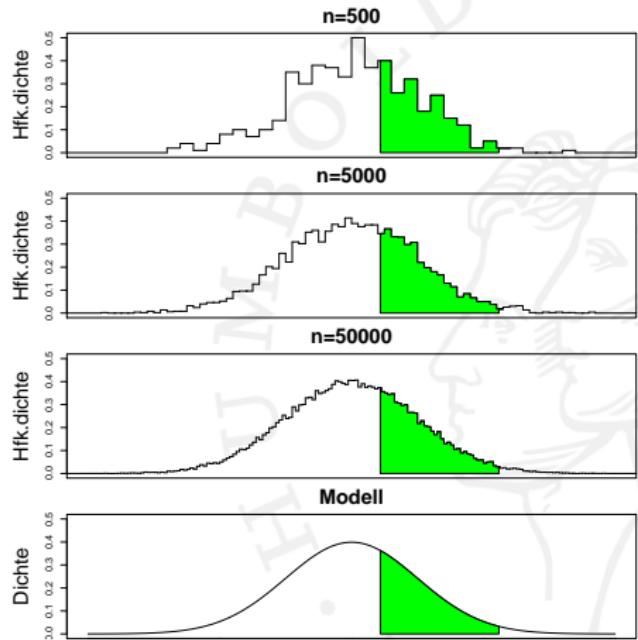
Wahrscheinlichkeitsdichte $f(x)$

$$P(a < X \leq b) = \int_a^b f(x)dx \quad \text{für } a \leq b$$



Schulmathematik

- $f(x)$ für sich alleine bedeutet keine Wahrscheinlichkeit
- $f(x)dx$ entspricht der Wahrscheinlichkeit, dass die stetige Zufallsvariable X einen Wert in einem beliebig kleinen Intervall $[x, x + dx]$ annimmt
- Die Wahrscheinlichkeit, dass eine stetige Zufallsvariable X einen bestimmten Wert x annimmt, ist stets Null. Das ergibt sich aus der Tatsache, dass die Fläche über einem Punkt x gleich Null ist.



- theoretisches Äquivalent der “Häufigkeitsdichte” (mit $dx \rightarrow 0$)
- Die grüne Fläche unter der Kurve entspricht der Wahrscheinlichkeit, einen Wert aus diesem Intervall zu erhalten.
- Die Wahrscheinlichkeit des Intervalls $[a, b]$ wird durch eine Fläche, d.h. formal durch ein Integral, beschrieben.

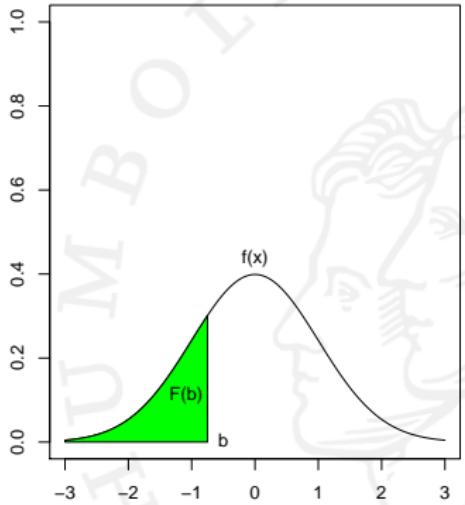
Verteilungsfunktion $F(x)$

$$F(x) = P(-\infty < X \leq x) = \int_{-\infty}^x f(t) dt$$

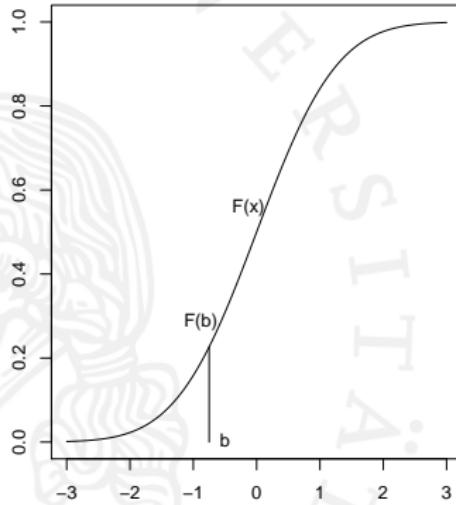
Eigenschaften

- $F(x)$ monoton wachsend
- $F(-\infty) = 0$
- $F(+\infty) = 1$
- $f(x) \geq 0$
- $\int_{-\infty}^{+\infty} f(x) dx = 1$

Dichtefunktion

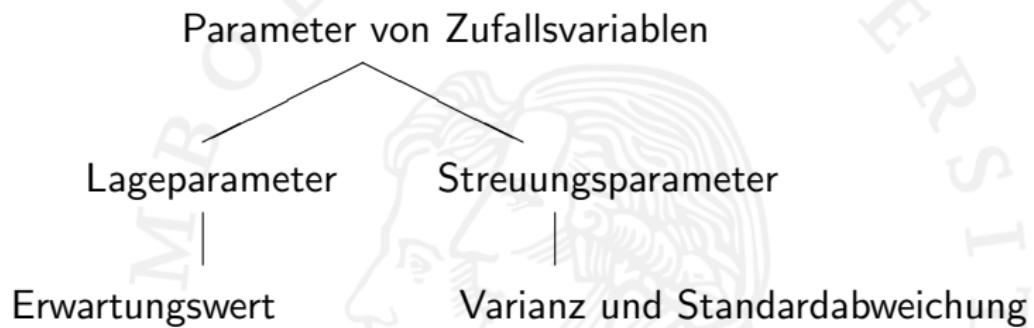


Verteilungsfunktion



$$\frac{\partial F(x)}{\partial x} = f(x) \Leftrightarrow F(x) = \int_{-\infty}^x f(t) dt$$

Parameter von Zufallsvariablen



- **Lageparameter** geben an, wo die Verteilung liegt
- **Streuungsparameter** beinhalten eine Aussage über die Variabilität innerhalb der Zufallsvariablen

Erwartungswert

Erwartungswert einer Zufallsvariablen X

- Wert der Zufallsvariablen, der im Mittel **vor** der Durchführung des Zufallsexperimentes zu erwarten ist
- diskrete Zufallsvariable:

$$E(X) = \mu_x = \sum_{i=1}^{\infty} x_i \cdot f(x_i)$$

- stetige Zufallsvariable:

$$E(X) = \mu_x = \int_{-\infty}^{+\infty} x \cdot f(x) dx$$

Beispiel 12.3

Wieviel Zahlwürfe kann man (im Mittel) erwarten, wenn man drei Münzen wirft. Also das Zufallsexperiment ganz oft durchführt?

Zufallsvariable X : "Anzahl der Zahlwürfe, wenn man drei Münzen wirft"

i	1	2	3	4
x_i	0	1	2	3
$f(x_i)$	0,125	0,375	0,375	0,125

$$E(X) = 0 \cdot 0,125 + 1 \cdot 0,375 + 2 \cdot 0,375 + 3 \cdot 0,125 = 1,5$$

Im Mittel wird man also 1,5 mal Zahl erwarten, wenn man drei Münzen wirft.

Beispiel 12.4

Welche weitere Lebenserwartung hat ein Mann, der seinen zwanzigsten Geburtstag feiert, in Deutschland?

Zufallsvariable X : „Lebensdauer in Dekaden“

$X = 1$: Mann überlebt eine Dekade, erreicht also das 30ste Lebensjahr, stirbt aber vor Erreichen des 40sten Lebensjahrs

i	1	2	3	4	5	6	7	8	9
x_i	0	1	2	3	4	5	6	7	8
$f(x_i)$	0,017	0,021	0,029	0,090	0,217	0,328	0,239	0,038	0,001

$$E(X) = 0 \cdot 0,017 + 1 \cdot 0,021 + 2 \cdot 0,029 + 3 \cdot 0,090 + 4 \cdot 0,217 + 5 \cdot 0,328 + 6 \cdot 0,239 + 7 \cdot 0,038 + 8 \cdot 0,001 = 4,705$$

Ein gerade 20 Jahre alt gewordener Mann kann danach erwarten, weitere 4,705 Dekaden zu leben, also 67 Jahre alt zu werden.

Linearität des Erwartungswertes

- Ergibt sich Y als lineare Transformation $Y = a + bX$, so ist

$$E(Y) = E(a + bX) = a + bE(X) \text{ mit } a, b \text{ beliebig}$$

- Für die Summe von Zufallsvariablen $Z = X + Y$ gilt:

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

Erwartungswerte von Funktionen von X

- diskrete Zufallsvariable:

$$E(g(X)) = \sum_{i=1}^{\infty} g(x_i) \cdot f(x_i)$$

- stetige Zufallsvariable:

$$E(g(X)) = \int_{-\infty}^{+\infty} g(x) \cdot f(x) dx$$

Varianz

Erwartungswert der quadrierten Abweichungen der Zufallsvariablen von ihrem Erwartungswert:

$$\begin{aligned} \text{Var}(X) = \sigma_x^2 &= E[\{X - E(X)\}^2] \\ &= E(X^2) - [E(X)]^2 \end{aligned}$$

- diskrete Zufallsvariable:

$$\text{Var}(X) = \sum_{i=1}^{\infty} [x_i - E(X)]^2 \cdot f(x_i) = \sum_{i=1}^{\infty} x_i^2 f(x_i) - [E(X)]^2$$

- stetige Zufallsvariable:

$$\text{Var}(X) = \int_{-\infty}^{+\infty} [x - E(X)]^2 \cdot f(x) dx = \int_{-\infty}^{+\infty} x^2 f(x) dx - [E(X)]^2$$

Eigenschaften der Varianz

X und Y sind zwei Zufallsvariablen mit Varianzen $\text{Var}(X)$ und $\text{Var}(Y)$. Dann gilt:

- Ergibt sich Y als lineare Transformation $Y = a + bX$, so ist

$$\text{Var}(Y) = \text{Var}(a + bX) = b^2 \text{Var}(X) \text{ mit } a, b \text{ beliebig}$$

Standardabweichung

σ_X : Wurzel aus der Varianz

Beispiel 12.5

Zufallsvariable X : "Anzahl der Zahlwürfe, wenn man drei Münzen wirft"

i	1	2	3	4
x_i	0	1	2	3
$f(x_i)$	0,125	0,375	0,375	0,125

$$E(X) = 1,5$$

$$\begin{aligned}Var(X) &= (0 - 1,5)^2 \cdot 0,125 + (1 - 1,5)^2 \cdot 0,375 + \\&= (2 - 1,5)^2 \cdot 0,375 + (3 - 1,5)^2 \cdot 0,125 = 0,75\end{aligned}$$

$$E(X^2) = 0^2 \cdot 0,125 + 1^2 \cdot 0,375 + 2^2 \cdot 0,375 + 3^2 \cdot 0,125 = 3$$

$$Var(X) = 3 - 1,5^2 = 0,75$$

Standardisierung

Betrachtet man die lineare Transformation

$$Z = \frac{X - E(X)}{\sqrt{\text{Var}(X)}}$$

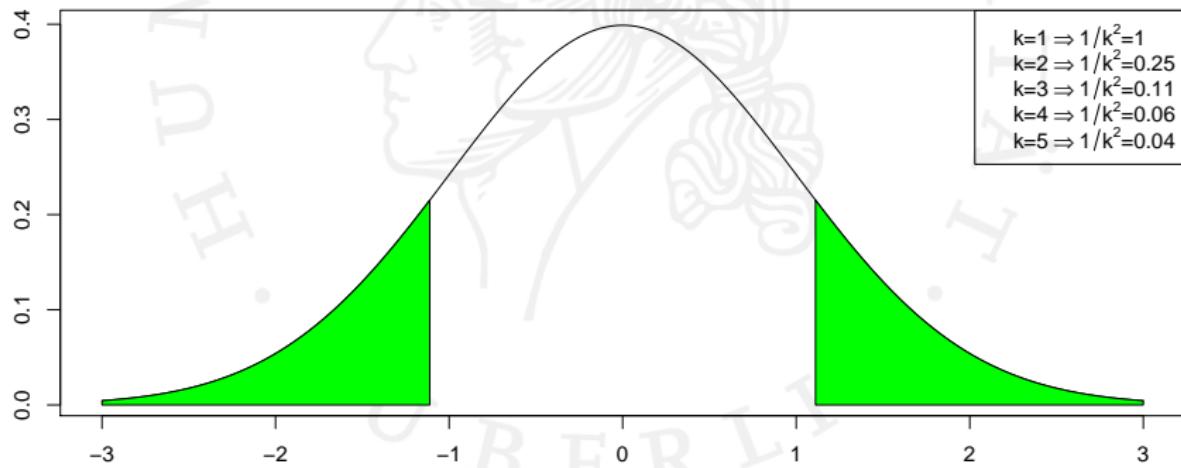
so gilt für die transformierte Zufallsvariable

$$\begin{aligned}E(Z) &= 0 \\ \text{Var}(Z) &= 1\end{aligned}$$

Tschebyscheff Ungleichung

Sei X eine beliebige diskrete oder stetige Zufallsvariable mit $E(X) = \mu_x$ und $\text{Var}(X) = \sigma_x^2$, dann gilt

$$P(|X - \mu_x| \geq k\sigma_x) \leq \frac{1}{k^2} \quad \text{bzw.} \quad P(|X - \mu_x| \leq k\sigma_x) \geq 1 - \frac{1}{k^2}$$



Diskrete Zufallsvariablen

Wahrscheinlichkeitsfunktion zweier diskreter Zufallsvariablen

$$P(X = x_i, Y = y_j) = f(x_i, y_j)$$

Eigenschaften:

- $\sum_i \sum_j f(x_i, y_j) = 1$
- $f(x_i, y_j) \geq 0$

Gemeinsame Wahrscheinlichkeitsverteilung

Variable X	Variable Y				Randverteilung X
	y_1	...	y_j	...	
x_1	$f(x_1, y_1)$...	$f(x_1, y_j)$...	$f(x_1)$
\vdots	\vdots	...	\vdots	...	\vdots
x_i	$f(x_i, y_1)$...	$f(x_i, y_j)$...	$f(x_i)$
\vdots	\vdots	...	\vdots	...	\vdots
Randverteilung Y	$f(y_1)$...	$f(y_j)$...	1

Stetige Zufallsvariablen

Wahrscheinlichkeitsdichte

$$P(x < X \leq x + \Delta x; y < Y \leq y + \Delta y) = f(x, y)$$

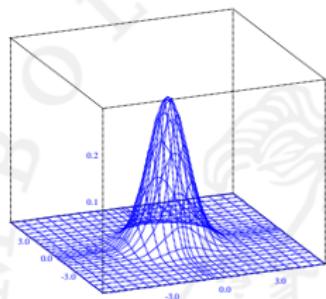
$$P(a < X \leq b; c < Y \leq d) = \int\limits_c^d \int\limits_a^b f(x, y) dx dy$$

Eigenschaften:

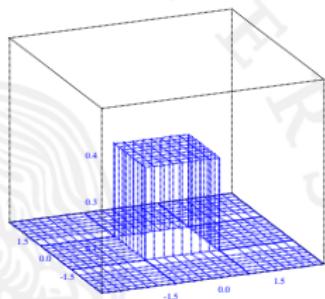
- $f(x, y) \geq 0$
- $\int\limits_{-\infty}^{+\infty} \int\limits_{-\infty}^{+\infty} f(x, y) dx dy = 1$

Wahrscheinlichkeitsdichte zweier Zufallsvariablen

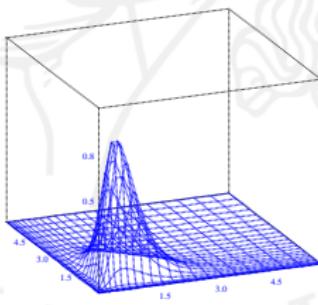
Wahrscheinlichkeitsdichte $f(x,y)$



Wahrscheinlichkeitsdichte $f(x,y)$



Wahrscheinlichkeitsdichte $f(x,y)$



Verteilungsfunktion zweier Zufallsvariablen

$$F(x, y) = P(X \leq x, Y \leq y)$$

- diskret:

$$F(x, y) = \sum_{x_i \leq x} \sum_{y_j \leq y} f(x_i, y_j)$$

- stetig:

$$F(x, y) = \int_{-\infty}^x \int_{-\infty}^y f(u, v) dudv$$

Eigenschaft:

$$0 \leq F(x, y) \leq 1$$

Randverteilungen

- Diskret

$$f(x_i) = P(X = x_i) = \sum_j f(x_i, y_j)$$

$$f(y_j) = P(Y = y_j) = \sum_i f(x_i, y_j)$$

- Stetig:

$$f(x) = \int_{-\infty}^{+\infty} f(x, y) dy$$

$$f(y) = \int_{-\infty}^{+\infty} f(x, y) dx$$

Verteilungsfunktion der Randverteilungen:

$$P(X \leq x) = F(x) = \begin{cases} \sum_{j=1}^r \sum_{x_i \leq x} f(x_i, y_j) & X, Y \text{ diskret} \\ \int_{-\infty}^{+\infty} \int_{-\infty}^x f(u, v) dudv & X, Y \text{ stetig} \end{cases}$$

$$P(Y \leq y) = F(y) = \begin{cases} \sum_{y_j \leq y} \sum_{i=1}^m f(x_i, y_j) & X, Y \text{ diskret} \\ \int_{-\infty}^y \int_{-\infty}^{+\infty} f(u, v) dudv & X, Y \text{ stetig} \end{cases}$$

Erwartungswerte der Randverteilungen:

Diskret:

$$\begin{aligned} E(X) &= \sum_{i=1}^m \sum_{j=1}^r x_i \cdot f(x_i, y_j) \\ &= \sum_{i=1}^m x_i \cdot f(x_i) \end{aligned}$$

Stetig:

$$\begin{aligned} E(X) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x \cdot f(x, y) dx dy \\ &= \int_{-\infty}^{+\infty} x \cdot f(x) dx \end{aligned}$$

$$\begin{aligned} E(Y) &= \sum_{j=1}^r \sum_{i=1}^m y_j \cdot f(x_i, y_j) \\ &= \sum_{j=1}^r y_j \cdot f(y_j) \end{aligned}$$

$$\begin{aligned} E(Y) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} y \cdot f(x, y) dx dy \\ &= \int_{-\infty}^{+\infty} y \cdot f(y) dy \end{aligned}$$

Varianz diskreter Randverteilungen:

$$\begin{aligned} \text{Var}(X) &= \sum_{i=1}^m \sum_{j=1}^r [x_i - E(X)]^2 f(x_i, y_j) \\ &= \sum_{i=1}^m [x_i - E(X)]^2 \sum_{j=1}^r f(x_i, y_j) \\ &= \sum_{i=1}^m [x_i - E(X)]^2 f(x_i) \end{aligned}$$

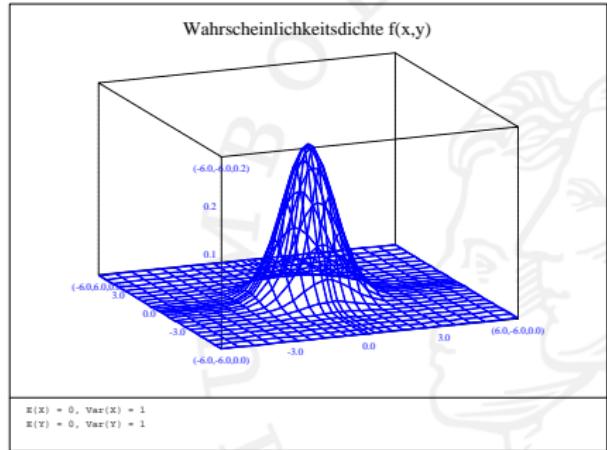
$$\begin{aligned} \text{Var}(Y) &= \sum_{j=1}^r \sum_{i=1}^m [y_j - E(Y)]^2 f(x_i, y_j) \\ &= \sum_{j=1}^r [y_j - E(Y)]^2 \sum_{i=1}^m f(x_i, y_j) \\ &= \sum_{j=1}^r [y_j - E(Y)]^2 f(y_j) \end{aligned}$$

Varianz stetiger Randverteilungen:

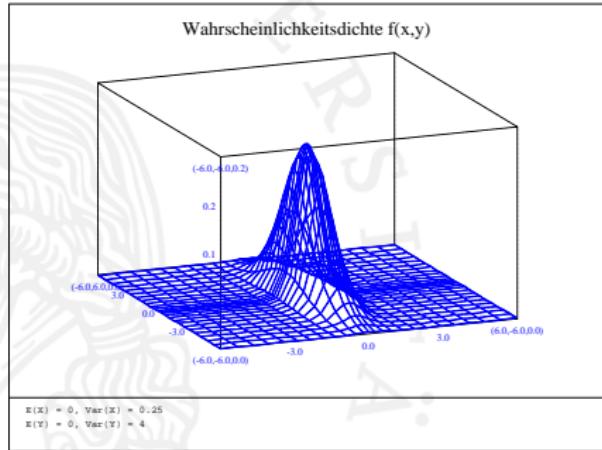
$$\begin{aligned} \text{Var}(X) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} [x - E(X)]^2 f(x, y) dx dy \\ &= \int_{-\infty}^{+\infty} [x - E(X)]^2 \int_{-\infty}^{+\infty} f(x, y) dx dy \\ &= \int_{-\infty}^{+\infty} [x - E(X)]^2 f(x) dx \end{aligned}$$

$$\begin{aligned} \text{Var}(Y) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} [y - E(Y)]^2 f(x, y) dx dy \\ &= \int_{-\infty}^{+\infty} [y - E(Y)]^2 \int_{-\infty}^{+\infty} f(x, y) dx dy \\ &= \int_{-\infty}^{+\infty} [y - E(Y)]^2 f(y) dy \end{aligned}$$

Grafische Darstellung



$$E(X) = 0, \quad \text{Var}(X) = 1$$
$$E(Y) = 0, \quad \text{Var}(Y) = 1$$



$$E(X) = 0, \quad \text{Var}(X) = 0.25$$
$$E(Y) = 0, \quad \text{Var}(Y) = 4$$

Bedingte Verteilungen

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{bzw.} \quad P(B|A) = \frac{P(A \cap B)}{P(A)}$$

- diskreter Fall:

$$P(X = x_i | Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} = \frac{f(x_i, y_j)}{f(y_j)} = f(x_i | y_j)$$

$$P(Y = y_j | X = x_i) = \frac{P(X = x_i, Y = y_j)}{P(X = x_i)} = \frac{f(x_i, y_j)}{f(x_i)} = f(y_j | x_i)$$

- stetiger Fall:

$$f(x|y) = \frac{f(x, y)}{f(y)} \qquad f(y|x) = \frac{f(x, y)}{f(x)}$$

Erwartungswert

- diskret:

$$E(X|y_j) = \sum_{i=1}^m x_i \cdot f(x_i|y_j)$$

$$E(Y|x_i) = \sum_{j=1}^r y_j \cdot f(y_j|x_i)$$

- stetig:

$$E(X|y) = \int_{-\infty}^{+\infty} x \cdot f(x|y) dx$$

$$E(Y|x) = \int_{-\infty}^{+\infty} y \cdot f(y|x) dy$$

Varianz

- diskret:

$$\begin{aligned} \text{Var}(X|y_j) &= \sum_{i=1}^m [x_i - E(X|y_j)]^2 f(x_i|y_j) \\ \text{Var}(Y|x_i) &= \sum_{j=1}^r [y_j - E(Y|x_i)]^2 f(y_j|x_i) \end{aligned}$$

- stetig:

$$\begin{aligned} \text{Var}(X|y) &= \int_{-\infty}^{+\infty} [x - E(X|y)]^2 f(x|y) dx \\ \text{Var}(Y|x) &= \int_{-\infty}^{+\infty} [y - E(Y|x)]^2 f(y|x) dy \end{aligned}$$

Unabhängigkeit

- Stochastische Unabhängigkeit

- diskrete Zufallsvariablen $A = \{X = x_i\}$ und $B = \{Y = y_j\}$
 - aus Wahrscheinlichkeitsrechnung:

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

- $\Rightarrow P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(Y = y_j)$

- Zwei Zufallsvariablen X und Y sind unabhängig, wenn gilt

- diskret:

$$f(x_i, y_j) = f(x_i) \cdot f(y_j) \quad \text{für alle } x_i, y_j$$

- stetig:

$$f(x, y) = f(x) \cdot f(y) \quad \text{für alle } x, y$$

Sind X und Y unabhängige Zufallsvariablen, so gilt

$$E(X \cdot Y) = E(X) \cdot E(Y)$$

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y)$$

Kovarianz

Die Kovarianz ist der Parameter, der die im Mittel zu erwartende gemeinsame Variation beinhaltet.

$$\begin{aligned}\text{Cov}(X, Y) &= E[\{X - E(X)\} \cdot \{Y - E(Y)\}] \\ &= E(XY) - E(X)E(Y)\end{aligned}$$

Linearkombinationen von Zufallsvariablen

- $Z = aX + bY$

$$\begin{aligned}E(Z) &= aE(X) + bE(Y) \\ \text{Var}(Z) &= a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab\text{Cov}(X, Y)\end{aligned}$$

- $Z = aX - bY$

$$\begin{aligned}E(Z) &= aE(X) - bE(Y) \\ \text{Var}(Z) &= a^2 \text{Var}(X) + b^2 \text{Var}(Y) - 2ab\text{Cov}(X, Y)\end{aligned}$$

Eigenschaften der Kovarianz:

- $\text{Var}(X) = \text{Cov}(X, X)$
- lineare Transformation:

$$X \rightarrow a + bX, Y \rightarrow c + dY$$

$$\text{Cov}(a + bX, c + dY) = b \cdot d \cdot \text{Cov}(X, Y)$$

- unter Unabhängigkeit $\text{Cov}(X, Y) = s_{xy} = 0$

 Aus $\text{Cov}(X, Y) = s_{xy} = 0$ folgt NICHT Unabhängigkeit!

Beispiel 12.6

$$Z = X + Y \quad (a = 1, b = 1)$$

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

$$\text{Var}(Z) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$$

$$Z = X - Y \quad (a = 1, b = -1)$$

$$E(Z) = E(X) - E(Y)$$

$$\text{Var}(Z) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y)$$

$$Z = \frac{1}{2}(X + Y) \quad (a = \frac{1}{2}, b = \frac{1}{2})$$

$$E(Z) = \frac{1}{2}[E(X) + E(Y)]$$

$$\text{Var}(Z) = \frac{1}{4}\text{Var}(X) + \frac{1}{4}\text{Var}(Y) + \frac{1}{2}\text{Cov}(X, Y)$$

Theoretischer Korrelationskoeffizient

$$\begin{aligned}\rho(X, Y) &= \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y} \\ &= E\left[\frac{\{X - E(X)\}}{\sigma_X} \cdot \frac{\{Y - E(Y)\}}{\sigma_Y}\right]\end{aligned}$$

Eigenschaft:

$$-1 \leq \rho(X, Y) \leq +1$$

Variablen vs. Zufallsvariablen

	Variable(n)	Zufallsvariable(n)
Lageparameter	\bar{x}	$E(X)$
Streuungsparameter	s^2 s	$Var(X)$ σ_X
Zusammenhangsparameter	s_{xy} r_{xy}	$Cov(X, Y)$ $\rho(X, Y)$

- Jeder Parameter einer Variablen kann auch für eine Zufallsvariable definiert werden.
 - ▶ Median
 - ▶ Interquartilsabstand
 - ▶ ...

Zusammenfassung

- Zufallsexperimente bilden die Grundlage von Zufallsvariablen
- Zufallsvariablen fassen die Ergebnisse von Zufallsexperimenten in "interessierender" Weise zusammen
- Zufallsvariablen werden eindeutig durch ihre Wahrscheinlichkeits- bzw. Dichtefunktion charakterisiert
- Mit Zufallsvariablen kann man nicht wie mit reellen Zahlen rechnen
- Zufallsvariablen werden genutzt werden um Modelle über die Verteilung einer Variablen in der Grundgesamtheit aufzustellen
- Sämtliche Definitionen und Begriffe für Variablen aus der deskriptiven Statistik können äquivalent für Zufallsvariablen übertragen werden

So lügt man mit Statistik

5. November 2022

- Bücher
- Unstatistik des Monats
- Illusion der Präzision
- Yin & Yang
- Betrügerische Basis
- Prozente
- Prozente
- Absolute Zahlen
- Mittelwerte
- Will-Rogers-Phänomen
- Simpson-Paradoxon
- Trendextrapolation
- Lange Zeiträume
- Synthetischer Superlativ
- Vorsortierte Stichprobe
- Wie es in den Wald hineinschallt...
- Definition & Operationalisierung
- Korrelation kontra Kausalität
- Bedingte Wahrscheinlichkeit
- Vertrauenswürdige Zahlen?
- Vertrauenswürdige Zahlen?
- Betrügerische Grafiken I
- Betrügerische Grafiken II
- Betrügerische Grafiken III
- Betrügerische Grafiken IV
- Betrügerische Grafiken V

Bücher

- Thomas Bauer, Gerd Gigerenzer, Walter Krämer, Katharina Schüller (2022) Grüne fahren SUV und Joggen macht unsterblich: Über Risiken und Nebenwirkungen der Unstatistik, Campus Verlag
- Gerd Bosbach, Jens Jürgen Korff (2017), Die Zahlentrickser: Das Märchen von den aussterbenden Deutschen und andere Statistiklügen, Heyne Verlag
- Thomas Bauer, Gerd Gigerenzer, Walter Krämer (2016), Warum dick nicht doof macht und Genmais nicht tötet: Über Risiken und Nebenwirkungen der Unstatistik, Goldmann Verlag
- Gerd Bosbach, Jens Jürgen Korff (2012), Lügen mit Zahlen: Wie wir mit Statistiken manipuliert werden, Heyne Verlag
- Walter Krämer (2011), Die Angst der Woche: Warum wir uns vor den falschen Dingen fürchten, Piper
- Walter Krämer (1997, 2009, 2015), So lügt man mit Statistik, Campus Verlag
- Walter Krämer, Gerald Mackenthun (2003), Die Panikmacher, Piper

Unstatistik des Monats

- Unstatistik des Monats
 - ▶ 30.06.2022: Falsch positive Chatkontrolle
 - ▶ 31.05.2022: WHO-Studie zur Corona-Übersterblichkeit nutzt störanfällige Methode
 - ▶ 27.04.2022: Künstliche Intelligenz schafft falsche Bilder
 - ▶ 29.03.2022: Jedes fünfte Kind in Deutschland gilt als arm – und das wird so bleiben
 - ▶ 25.02.2022: Höhere Straßenbaumdichte, weniger Depressionen?
 - ▶ 31.01.2022: Impfquote und Übersterblichkeit, eine “Spurious Correlation”
- Hinterfragung jüngst publizierte Zahlen als auch deren Interpretationen
- seit 2012 von Thomas Bauer, Gerd Gigerenzer, Walter Krämer, ab 2018 auch von Katharina Schüller

Quelle: <https://www.unstatistik.de>

Illusion der Präzision

- Je genauer Werte angegeben werden desto glaubwürdiger sind sie
- Methusalem wurde 969 Jahre alt, nicht ca. 1000
- Goliath ist sechs Ellen und eine Handbreit groß, nicht sehr groß oder riesengroß
- Peary am Nordpol
 - ▶ In Peary's Tagebuch (6.4.1909): Position 89 Grad 57 Minuten 11 Sekunden (5 km vom Pol)
 - ▶ 1 Bogensekunde = ca. 30 m
- Fahrradbestand in Deutschland 2020: 79,1 Mio. Stück (Statista/???)
- Wahlumfrage
 - ▶ bei 1.000 Befragten: $\pm 3\%$
 - ▶ bei 10.000 Befragten: $\pm 1\%$
 - ▶ Nachwahlumfrage: 100.000 Befragte = $\pm 0,3\%$

Yin & Yang

- Aktientrick
 - ▶ sechsmal richtigen Tip im Brief
 - ▶ 1. Brief an 16.000 potentielle Anleger, 2. Brief an 8.000, ...
- Wir haben letztes Jahr 2.200 Lehrer neu eingestellt
 - ▶ leider wurden auch 2.500 pensioniert...
- Sehr viele Raser auf der Autobahn/alle überholen mich
 - ▶ schnellere überholen mich
 - ▶ gleich schnell Fahrende sehe ich gar nicht
- 50% mehr betrunkene Autofahrer bei Polizeikontrollen
 - ▶ 1. Woche: 10 von 500 kontrollierten betrunken
 - ▶ 2. Woche: 15 von 500 kontrollierten betrunken
 - ▶ 1. Woche: 98% nüchtern, 2. Woche 97% nüchtern

Betrügerische Basis

- Trinker
 - ▶ New York Times: Zwei Drittel aller Trinker sind verheiratet, daraus wird gefolgert Eheleben fördert Alkoholismus
 - ▶ Problem: Wie gross ist der Anteil der Ehemänner an der Gruppe der Männer (im heiratsfähigen Alter) (> 2/3)?
- Höchste Kriminalitätsrate weltweit: Vatikanstadt
 - ▶ Viel Kriminalität bei Touristen
 - ▶ Nur 825 Einwohner
- Immer mehr Nitrat im Grundwasser
 - ▶ Mittlere Nitratgehalt an den 15 am stärksten belasteten deutschen Messpunkten von 2013 bis 2017 nimmt um rund 40 Milligramm pro Liter zu
 - ▶ 2013: Mittelwert über alle Meßstellen
 - ▶ 2017: Mittelwert über Höchstwerte an Meßstellen mit hoher Belastung

Prozente

- Schwangerschaften UK (pill scare)

- ▶ Warnung des britischen Ausschusses für Arzneimittelsicherheit im Oktober 1995, das bei bestimmten Antibabypillen das Thromboserisiko mehr als doppelt so hoch ist wie bei anderen ($> +100\%$)
- ▶ aufgegriffen von der Presse, Frauen setzten die Pille ab
- ▶ England & Wales 1996: 26.000 zusätzliche Geburten, 13.600 zusätzliche Abtreibungen (ähnliche Effekte in Deutschland und Norwegen)
- ▶ tatsächliche Risiken:
 - ★ an Thrombose zu sterben: 1 von 7000 Frauen auf 2 von 7000 Frauen
 - ★ bei der Geburt zu sterben: 80 von 1 Mio. schwangeren Frauen (Deutschland)
 - ★ bei der Abtreibung zu sterben: 6 von 1 Mio. schwangeren Frauen
- ▶ zusätzliche Kosten für NHS: mehr als 40 Mio. EUR

Prozente

- 25% weniger Brustkrebstote durch Mammographievorsorge bei Frauen über 50
 - ▶ mit jährlicher Mammographie: 3 von 1.000 Frauen sterben in zehn Jahren
 - ★ 50 Falsch-positive Befunde
 - ▶ ohne Mammographie: 4 von 1.000 Frauen sterben in zehn Jahren
- Manager der Firma XYZ verdienen 30% mehr
 - ▶ als letztes Jahr?
 - ▶ verglichen mit Managern von Konkurrenzunternehmen?
- Frage nach der Bezugsgröße

Absolute Zahlen

- Armin Laschet: 1.000 zusätzliche Lehrer in NRW eingestellt
 - ▶ aber 7.000 öffentliche Schulen in NRW
- Deutschland Exportweltmeister!
 - ▶ umgerechnet auf "pro Einwohner" nur Rang 16 (von 168)
- Rückgang der Jugendlichen unter 20 Jahre
 - ▶ 2005: 16,5 Mio., 2050: 10,9 Mio.
 - ▶ aber Bevölkerung schrumpft auch in Prognose
2005: 20% Anteil, 2050: 15,3%
 - ▶ Yin & Yang - 1900: 44,1%

Mittelwerte

- Mittlere Jahrestemperatur
 - ▶ Plymouth (UK): 13 Grad (+8 - +21 Grad)
 - ▶ Minneapolis (USA): 13 Grad (-15 - +40 Grad)
- Was ist sicherer: Bahn oder Flugzeug?
 - ▶ Bahn: 9 Tote pro 10 Mrd. Passagierkilometer, Flugzeug: 3 Tote pro 10 Mrd. Passagierkilometer
 - ▶ Bahn: 7 Tote pro 100 Mio. Passagierstunden, Flugzeug: 24 Tote pro 100 Mio. Passagierstunden
- Abiturnoten in Bayern besser als in Nordrhein-Westfalen
 - ▶ Abiturquote in Bayern: 10%
 - ▶ Abiturquote in Nordrhein-Westfalen: 40%

Will-Rogers-Phänomen

- Durch einen Wechsel eines Elements von einer zur anderen Gruppe kann der Mittelwert in beiden Gruppen steigen (oder fallen)

2020		2021	
Nord	Süd	Nord	Süd
5000	5000	5000	5000
6000	10000	6000	15000
7000	15000	7000	20000
	20000	10000	
6000	12500	7000	13333
	+14,1%		+6,6%

*When the Okies left Oklahoma and moved to California,
they raised the average intelligence level in both states*

Simpson-Paradoxon

- Bewertung verschiedener Gruppen fällt unterschiedlich aus, je nachdem ob man die Ergebnisse der Gruppen kombiniert oder nicht
- Diskriminierungsklage gegen die Universität Berkeley

Fach	Frauen	ang.	Männer	ang.	Quote	
					F	M
1	10	8	80	50	80%	63%
2	5	4	60	40	80%	67%
3	80	20	40	10	25%	25%
4	30	15	40	10	50%	25%
Σ	125	47	220	110	38%	50%

Trendextrapolation

- Malthus (1872)
 - ▶ Bevölkerung steigt exponentiell
 - ▶ Sozialprodukt steigt (nur) linear
 - ▶ Konsequenz: Hunger und Elend für die Masse der Menschen
- Club of Rome (1972)
 - ▶ Verfügbares Ackerland ist begrenzt, nutzbares Land nimmt sogar ab
 - ▶ Auch bei vervielfachter Produktivität steht nicht genug Ackerland zu Verfügung (bei gleicher Prod. 2005, bei doppelter Prod. 2040, ...)
- CO₂ Gehalt
 - ▶ in 400 Jahren besteht die Atmosphäre nur noch aus CO₂
- Club of Rome & IPCC: mehrere Alternativen berechnen

Lange Zeiträume

- Bildungsgipfel 2009: 18 Mrd. EUR für die Bildung
 - ▶ Zeitraum 2011-2018
 - ▶ Geld nur garantiert bis 2013
 - ▶ nur 2 Mrd. EUR/Jahre
 - ▶ ca. 11 Mio. Schüler & 3 Mio. Studenten
 - ▶ ca. 150 EUR pro Jahr und Schüler/Student
 - ▶ aber Bildung & Forschung
- ein großes Problem?
 - ▶ 218 Getötete bei rassistischer/rechter Motivation seit 1990 (Amadeu Antonio Stiftung)
 - ▶ pro Jahr(!): ca. 180 Getötete von (Ex-) Partnern (75% Frauen)
- Benzinpreise
 - ▶ 1970: 0,25 EUR/Liter, 2010: 1,40 EUR/Liter, 2019: 1,40 EUR/Liter

Synthetischer Superlativ

- Schränke die Vergleichsmenge soweit ein, dass jedes Datum ein Knaller wird
- Ephraim Kishon: Die “beste” Ehefrau von allen
- “I’m bigger and better than you” zeigt der Economist (vor 2000), daß jede der 24 OECD-Nationen die anderen in einer Konkurrenz mit Abstand schlägt:
 - ▶ Australien wohnen mehr Menschen als anderswo im eigenen Haus
 - ▶ Schweden zahlen die meisten Steuern
 - ▶ Deutsche trinken das meiste Bier
 - ▶ Schweizer haben das meiste Geld
 - ▶ Amerikaner haben das größte Sozialprodukt
 - ▶ ...
 - ▶ Island hat die größte Verbreitung des Spieles “Trivial Pursuit”
- sehr beliebt beim Sport

Vorsortierte Stichprobe

- Gute Studenten
 - ▶ Asiatische Studenten sind generell besser als deutsche Studenten
 - ▶ Nur erfolgreiche Studenten werden zum Auslandsstudium geschickt
- PISA: in VR China nur ausgewählte Schulen
- Times 1990: 60% Piloten sterben vor dem 65. Lebensjahr
 - ▶ eigentlich: 60% der im Vorjahr gestorbenen Piloten waren jünger als 65
 - ▶ Problem: es gab damals kaum Piloten älter als 65
- 2000: Fachhochschule Ansbach beste Hochschule Bayerns
 - ▶ Durchschnittsstudiendauer bis zum Abschluß ca. 8 Semester
 - ▶ gegründet 1996 ;)
- Gemälde sind eine gute Geldanlage
 - ▶ der Preisanstieg bei Auktionen schlägt jeden Aktienanstieg
 - ▶ ein Bild, das keiner haben will, wird nicht versteigert

Wie es in den Wald hineinschallt...

- Darf ich beim Beten rauchen? / Darf ich beim Rauchen beten?
- IG Metall 95% gegen Samstagsarbeit, Marplan: 72% bereit am Wochenende zu arbeiten (S. 123)
- Nachrüstung (2 Emnid-Umfragen mit einer Woche Abstand)
 - ▶ Panorama: 14% dafür
 - ▶ Bundesverteidigungsministerium: 54% dafür
- Gallup: Impeachment/Untersuchungsausschuss
- Telefonumfrage: Werden zu viele englische Wörter verwendet? 97,4% stimmen zu (lieber Ja als Nein)
- Sexualpartner pro Leben: Frauen 2,9, Männer 11 (soziale Erwünschtheit)
- Wie lange brauchen Sie für die Statistik-Hausaufgaben?
0-1/2 Stunde, 1/2-1 Stunde, 1-2 Stunden, mehr als 2 Stunden

Definition & Operationalisierung

- Problem: Unkenntnis der Definition
- Mehrere Arbeitslosigkeitsdefinitionen in Deutschland
- Säuglingssterblichkeit (D: innerhalb des 1. Lebensjahres, am 1. Tag/bis zu Taufe Verstorbene)
- Analphabeten (Kaiserreich: jemand der seinen Namen nicht schreiben konnte)
- Kriminalität bei Amoklauf (London: 10 Morde, Schottland: 1 Mord)
- Armut: weniger als 60% des Medianeinkommens
- Anteil Homosexueller in D (Verband: 2 gleichgeschlechtliche Personen in einem Haushalt ⇒ 10%)
- “Industrieroboter”: Definition in DDR “alle Anlagen mit numerischer Steuerung”
- DDR 1989: “3 Mio. fertiggestellte Wohnungen”, aber das schloß auch modernisierte Wohnungen ein

Korrelation kontra Kausalität

- Die Suche nach dem Warum?
- Paracetamol verursacht Zwölffingerdarmgeschwüre?
 - ▶ Patienten mit Zwölffingerdarmgeschwüre vertragen andere Schmerzmittelwirkstoffe nicht
- PVC Fußbodenbeläge (Phthalate = Weichmacher) verursachen Allergien
 - ▶ Patienten mit Allergien verlegen oft Fußbodenbeläge
- Abgeordnetendiäten und Bierpreise steigen
 - ▶ dritte Variable: die Zeit
- Schuhgröße und Gehalt sind korreliert
 - ▶ dritte Variable: Frauen und junge Männer (kleinere Schuhgrößen) verdienen in der Regel weniger
- Ehemänner leben länger
 - ▶ Faktoren, die Ehen begünstigen, begünstigen auch längeres Leben (Reichtum, Gesundheit)

Bedingte Wahrscheinlichkeit

$$P(A|B) = \begin{cases} P(A) & \text{wenn } A \text{ und } B \text{ unabhängig} \\ \frac{P(A \cap B)}{P(B)} & \text{wenn } A \text{ und } B \text{ abhängig} \end{cases}$$

- Roulette
 - ▶ Spieler notieren die Zahlen um zu sehen, welche Zahlen noch nicht gekommen sind
 - ▶ Implikation: die noch nicht gekommen Zahlen müssten häufiger kommen, da alle Zahlen gleich wahrscheinlich sind
- Junge & Mädchen
 - ▶ Mutter glaubt, dass nach drei Söhnen, dass das nächste Kind ein Mädchen wird
- Leukämie wird durch Atomkraftwerke verursacht
 - ▶ Vergleiche die Zahl der Leukämiefälle pro 1000 Einwohner in der Umgebung aller Atomkraftwerke
 - ▶ Stattdessen: wähle ein Atomkraftwerk mit einer hohen Zahl von Leukämiefällen pro 1000 Einwohner aus

Vertrauenswürdige Zahlen?

- Schlacht an den Thermopylen (480 v. Chr)
 - ▶ Herodot: Armee des Xerxes mit 5.283.220 Mann
 - ▶ Problem 1: Schlachtfeld zu klein
 - ▶ Problem 2: Versorgung unmöglich
 - ▶ Problem 3: Die Kette von Menschen würde von den Thermopylen bis nach Susa (Persien) reichen
- Aussenhandelszahlen DDR
 - ▶ 1987: Genosse Mittag möchte einen Überschuss von 910 Mio. Mark statt 521 Mio. in der Statistik
 - ▶ Problem 1: der Überschuss von 521 Mio. war schon falsch, das tatsächliche Defizit betrug 579 Mio.
 - ▶ Problem 2: jede Fälschung erzwingt eine Fälschung im nächsten Jahr, sonst gibt es einen starken Einbruch

Vertrauenswürdige Zahlen?

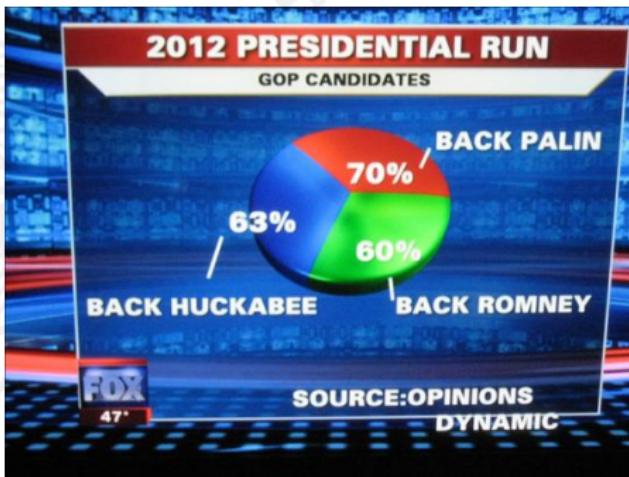
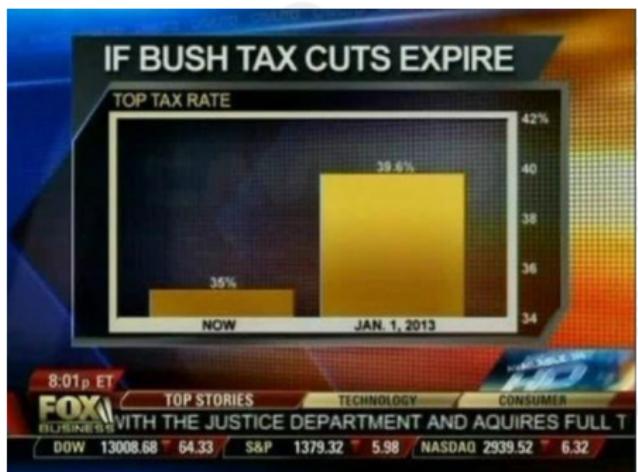
- Zahlen, denen wir glauben können
Zahl der formell unehelichen Geburten
- Zahlen, die einigermaßen genau sind
Alter, in dem eine (durchschnittliche) Frau/ein Mann zum ersten Mal Sex hatte
- Zahlen, die ziemlich weit daneben liegen können
Prozentsatz der Ehemänner, die außerehelichen Sex hatten (50%)
- Zahlen, die unzuverlässig sind
Prozentsatz der Frauen, die in den ersten fünf Jahren nach der Heirat Affären hatten (70%)
- Zahlen, die nur erfunden sind
Zahl der gehandelten "Sexsklaven" im Vereinigten Königreich (25.000)

Quellen, Konsistenz, Details, ... ?

Betrügerische Grafiken I

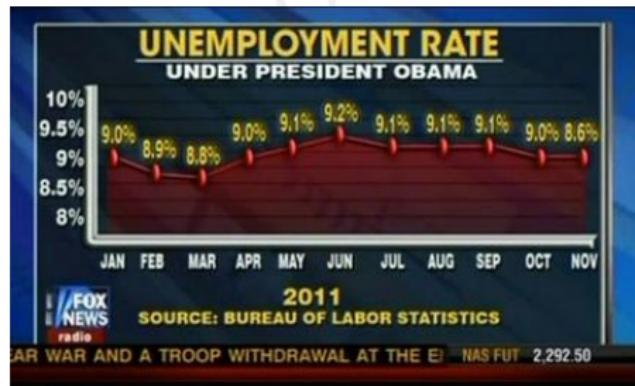
- **Ein Blick ist mehr wert als tausend Worte**
(Japanischer Philosoph)
 - **Ein Bild sagt mehr als zehntausend Worte**
Fred R. Barnyard (1927)
 - **Ein Bild lügt schneller als als tausend Worte**
-
- Abschneiden von Achsen
 - Skalen auseinander ziehen
 - Flächen/Symbole vervierfachen/verachtfachen
 - Perspektiven
 - Farbskalen (rot!)

Betrügerische Grafiken II



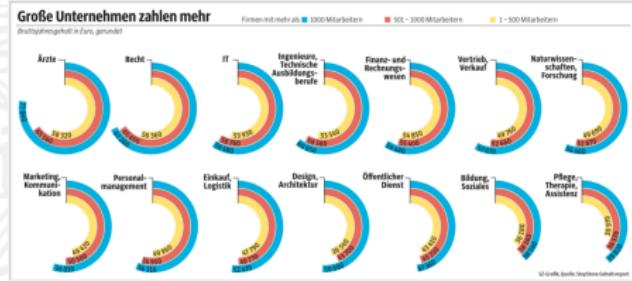
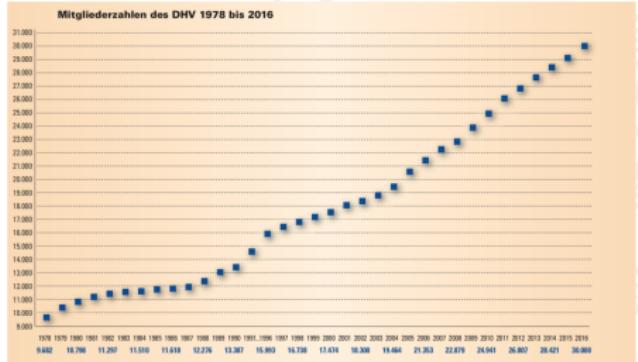
Source: simplystatistics.org (used in Fox news)

Betrügerische Grafiken III



Source: simplystatistics.org (used in Fox news)

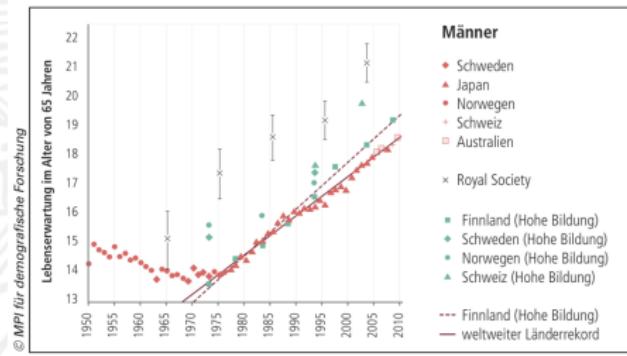
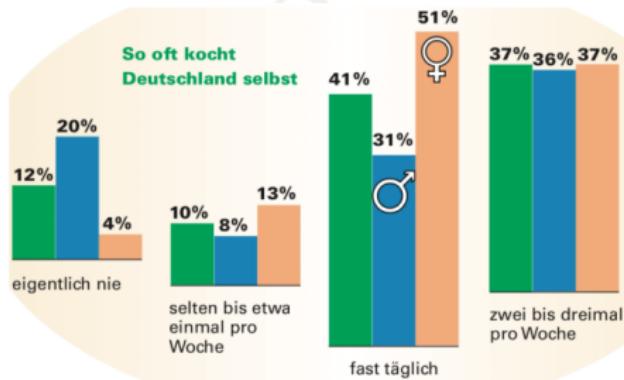
Betrügerische Grafiken IV



Big companies pay more

Source: Graphics in Research and Teaching illustrated in "Forschung & Lehre",
Vortrag DAGSTAT 2019, Prof. A. Unwin (Augsburg)

Betrügerische Grafiken V



Source: Graphics in Research and Teaching illustrated in "Forschung & Lehre",
Vortrag DAGSTAT 2019, Prof. A. Unwin (Augsburg)

Wichtige Verteilungsmodelle

5. November 2022

- Wiederholung: diskrete Zufallsvariablen • Diskrete Gleichverteilung • Bernoulli-Verteilung • Binomialverteilung • Hypergeometrische Verteilung
• Poisson-Verteilung • Bundesliga Ergebnisse • Wiederholung: stetige Zufallsvariablen • Stetige Gleichverteilung • Exponentialverteilung • Normalverteilung • Standardnormalverteilung • Zentraler Grenzwertsatz • Chi-Quadrat-Verteilung • t -Verteilung • F -Verteilung • Verteilungen • Approximation von Verteilungen • $B(n, p) \rightarrow Po(\lambda)$ • $B(n, p) \rightarrow N(\mu, \sigma)$
• $t_n \rightarrow N(0, 1)$ • $H(N, M, n) \rightarrow B(n, p)$ • $H(N, M, n) \rightarrow N(\mu, \sigma)$ •
Stetigkeitskorrektur • Zusammenfassung

Wiederholung: diskrete Zufallsvariablen

- Zufallsvariable: X
- Verteilungsfunktion:

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i)$$

- Wahrscheinlichkeitsfunktion:

$$f(x_i) = P(X = x_i)$$

- Erwartungswert:

$$E(X) = \mu_x = \sum_{i=1}^n x_i f(x_i)$$

- Varianz:

$$\text{Var}(X) = \sigma_x^2 = \sum_{i=1}^n [x_i - E(X)]^2 \cdot f(x_i) = \sum_{i=1}^n x_i^2 f(x_i) - [E(X)]^2$$

Diskrete Gleichverteilung

Situation

Eine Zufallsvariable X heißt diskret gleichverteilt, wenn

- sie nur n Werte x_1, \dots, x_n annehmen kann und
- jeder Wert wird mit der gleichen Wahrscheinlichkeit realisiert.

Notation: $X \sim U(n)$

Parameter: n

Wahrscheinlichkeitsfunktion

$$f(x_i) = \begin{cases} \frac{1}{n} & \text{für } i = 1, \dots, n \\ 0 & \text{sonst} \end{cases}$$

Verteilungsfunktion

$$F(x) = \begin{cases} 0 & \text{für } x < x_1 \\ \frac{i}{n} & \text{für } x_i \leq x < x_{i+1}; \quad i = 1, \dots, n-1 \\ 1 & \text{für } x_n \leq x \end{cases}$$

Erwartungswert und Varianz

$$E(X) = \mu_x = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{Var}(X) = \sigma_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_x)^2$$

im Spezialfall $x_1 = 1, \dots, x_n = n$

$$E(X) = \frac{n+1}{2} \quad \text{Var}(X) = \frac{(n+1)(n-1)}{12}$$

Beispiel 14.1 (Einmaliges Werfen eines idealen Würfels)

Zufallsvariable: X : "Augenzahl" auch $X \sim U(6)$

Wertebereich: 1, 2, 3, 4, 5, 6

klassische Definition der Wahrscheinlichkeit: $f(x_i) = 1/6$ für $i = 1, \dots, 6$

$E(X) = 3,5$ und $\text{Var}(X) = 2,916667$.

 Listing 14.1: example_dist_duniform.R

```
1 m <- 6 # Würfel ;)
2 par(mfrow=c(2,2))
3 plot(1:m, rep(1/m, m), type="h", xlab="x", ylab="Wahrscheinlichkeit",
4       main="Wahrscheinlichkeitsfunktion f(x)")
5 plot(ecdf(1:m), xlab="x", ylab="Kum. Wahrscheinlichkeit",
6       main="Verteilungsfunktion F(x)")
7 # Stichprobe
8 n <- 20
9 x <- sample(1:m, size=n, replace=TRUE)
10 tab <- table(factor(x, levels=1:m))
11 barplot(table(x), main="Häufigkeit h(x)", xlab="x", ylab="Anzahl",
12         sub=sprintf("n=%,.0f", n))
13 stripchart(x, method="jitter", ylim=c(0.75,1.25), pch=19,
14             main="Dotplot")
```

Bernoulli-Verteilung

Situation

- Zufallsexperiment mit zwei möglichen Ereignissen: A und \bar{A}
- konstante Wahrscheinlichkeiten des Eintretens der Ereignisse:
 $P(A) = p$ und $P(\bar{A}) = 1 - p$
- Versuche sind unabhängig (Modell mit Zurücklegen)

Die Zufallsvariable X heißt Bernoulli verteilt, wenn sie den Wert

- 1 annimmt, falls A eintritt
- 0 annimmt, falls \bar{A} eintritt

Notation: $X \sim B(p)$

Parameter: p

Wahrscheinlichkeitsfunktion

$$f(x; p) = \begin{cases} p^x \cdot (1 - p)^{1-x} & \text{für } x = 0, 1 \\ 0 & \text{sonst} \end{cases}$$

Verteilungsfunktion

$$F(x) = \begin{cases} 0 & \text{für } x < 0 \\ 1 - p & \text{für } 0 \leq x < 1 \\ 1 & \text{für } 1 \leq x \end{cases}$$

Erwartungswert und Varianz

$$E(X) = \mu_x = p \quad \text{Var}(X) = \sigma_x^2 = p(1 - p)$$

Binomialverteilung

Situation

- n -malige Durchführung eines Bernoulli-Versuchs mit $P(A) = p$ und $P(\bar{A}) = 1 - p$
- n dichotome unabhängige Zufallsvariablen $X_1, X_2, \dots, X_i, \dots, X_n$

$$X_i \sim B(p); \quad i = 1, \dots, n$$

⇒ Zufallsvariable X : "Anzahl des Auftretens von A bei n Versuchen"
heißt binomial verteilt:

$$X = \sum_{i=1}^n X_i$$

Notation: $X \sim B(n; p)$

Parameter: n, p

Herleitung der Wahrscheinlichkeitsfunktion

- Eintreten der Realisation $X = x$, z.B. wenn Ereignisfolge

$$\underbrace{A_1 \cap A_2 \cap \dots \cap A_x}_{x \text{ mal } A} \cap \underbrace{\bar{A}_{x+1} \cap \bar{A}_{x+2} \cap \dots \cap \bar{A}_n}_{(n-x) \text{ mal } \bar{A}}$$

eintritt.

- Reihenfolge der Anordnung spielt keine Rolle

$$\bar{A}_{x+1}, A_1, A_2, \dots, A_x, \bar{A}_{x+2}, \dots, \bar{A}_n$$

- Wahrscheinlichkeit, diese Ereignisfolge zu erhalten:

$$\begin{aligned}
 & P(A_1 \cap A_2 \cap \dots \cap A_x \cap \bar{A}_{x+1} \cap \bar{A}_{x+2} \cap \dots \cap \bar{A}_n) \\
 &= P(A_1) \cdot \dots \cdot P(A_x) \cdot P(\bar{A}_{x+1}) \cdot \dots \cdot P(\bar{A}_n) \\
 &= \underbrace{p \cdot \dots \cdot p}_x \cdot \underbrace{(1-p) \cdot \dots \cdot (1-p)}_{n-x} \\
 &= p^x \cdot (1-p)^{n-x}
 \end{aligned}$$

- Anzahl der verschiedenen Ereignisfolgen, x -mal das Ereignis A bei n Versuchen zu erhalten = Anzahl der Kombinationen ohne Wiederholung

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \quad (\text{Formel aus Kombinatorik})$$

$$\Rightarrow P(X = x) = f(x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}$$

Wahrscheinlichkeitsfunktion

$$f_B(x; n, p) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{für } x = 0, 1, \dots, n \\ 0 & \text{sonst} \end{cases}$$

Verteilungsfunktion

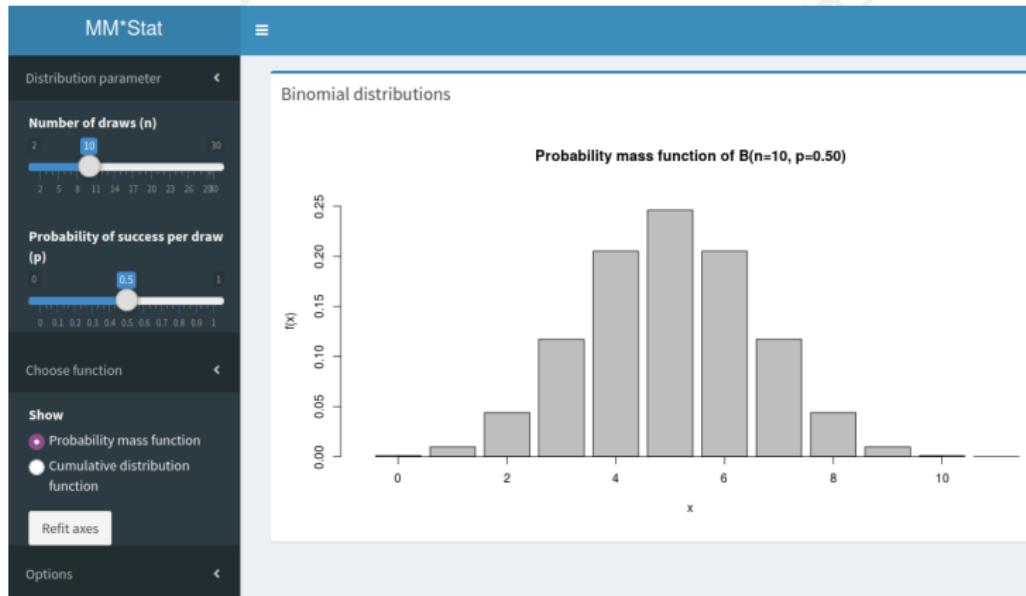
Werte für $F_B(x; n, p)$ siehe die Tabellen in der Formelsammlung

Erwartungswert und Varianz

$$E(X) = \mu_x = n \cdot p \quad \text{Var}(X) = \sigma_x^2 = n \cdot p \cdot (1 - p)$$



examples/stat/distribution_binom



R Listing 14.2: example_dist_binomial.R

```
1 library("mmstat4")
2 n <- 10
3 p <- 0.2
4 x <- 0:n
5 # Zufallsvariable
6 par(mfrow=c(2,2))
7 plot(x, dbinom(x, size=n, prob=p), type="h", xlab="x",
8       ylab="Wahrscheinlichkeit",
9       main="Wahrscheinlichkeitsfunktion f(x)")
10 df <- cdf(x, pbinom(x, size=n, prob=p))
11 plot(df, xlab="x", ylab="Kum. Wahrscheinlichkeit",
12       main="Verteilungsfunktion F(x)")
13 # Stichprobe
14 xs <- rbinom(20, size=n, prob=p)
15 tab <- table(factor(xs, levels=0:n))
16 barplot(tab, main="Häufigkeit h(x)", xlab="x", ylab="Anzahl",
17           sub=sprintf("n=%d", n))
18 stripchart(xs, method="jitter", ylim=c(0.75,1.25), pch=19,
19           main="Dotplot", xlim=range(x))
```

Beispiel 14.2 (Urnenmodell)

Situation:

- Urne mit $N = 10$ Kugeln, $M = 3$ weiße Kugeln, $N - M = 7$ rote Kugeln, $n = 5$ Ziehungen mit Zurücklegen
- W_i : "Anzahl des Auftretens einer weißen Kugel bei der i -ten Ziehung" ($i = 1, \dots, 5$)

W_i	$P(W_i = w_i) = f(w_i)$
0	0,7
1	0,3

- W : "Anzahl des Auftretens weißer Kugeln bei 5 Ziehungen mit Zurücklegen"
- $W = \sum_i W_i \implies W \sim B(n; p) = B(5; 0,3)$

Frage: Wie groß ist die Wahrscheinlichkeit, dass genau zwei weiße Kugel gezogen werden?

Lösung:

- Berechnung mit Wahrscheinlichkeitsfunktion:

$$P(W = 2) = f_B(2; 5; 0, 3) = \binom{5}{2} \cdot 0,3^2 \cdot 0,7^3 = 0,3087$$

- Berechnung mit Verteilungsfunktion:

$$f_B(2; 5; 0, 3) = F_B(2; 5; 0, 3) - F_B(1; 5; 0, 3) = 0,8369 - 0,5282 = 0,3087$$

x	$f_B(x; n, p)$	$F_B(x; n, p)$
0	0,1681	0,1681
1	0,3601	0,5282
2	0,3087	0,8369
3	0,1323	0,9692
4	0,0284	0,9976
5	0,0024	1,0000

Beispiel 14.3 (Urnenmodell)

Situation:

- Urne mit $N = 10$ Kugeln, $M = 3$ weiße Kugeln, $N - M = 7$ rote Kugeln, $n = 5$ Ziehungen mit Zurücklegen
- R_i : "Anzahl des Auftretens einer roten Kugel bei der i -ten Ziehung" ($i = 1, \dots, 5$)

R_i	$P(R_i = r_i) = f(r_i)$
0	0,3
1	0,7

- R : "Anzahl des Auftretens roter Kugeln bei 5 Ziehungen mit Zurücklegen"
- $R = \sum_i R_i \implies R \sim B(n; p) = B(5; 0,7)$

Frage: Wie groß ist die Wahrscheinlichkeit, dass genau zwei rote Kugel gezogen werden?

Lösung:

- Berechnung mit Wahrscheinlichkeitsfunktion:

$$P(R = 2) = f_B(2; 5; 0,7) = \binom{5}{2} \cdot 0,7^2 \cdot 0,3^3 = 0,1323$$

- Berechnung mit Verteilungsfunktion:

- ▶ Problem: Die Tabellen in der Formelsammlung enthalten nur Werte für $p \leq 0,5$
- ▶ Lösung: Symmetrie-Eigenschaft der Binomialverteilung:

$$\begin{aligned} f_B(x; n, p) &= f_B(n - x; n, 1 - p) \\ &= F_B(n - x; n, 1 - p) - F_B(n - x - 1; n, 1 - p) \end{aligned}$$

Wahrscheinlichkeits- und Verteilungsfunktion der Binomialverteilung $B(5; 0,3)$

x	$f_B(x; n, p)$	$F_B(x; n.p)$
0	0,1681	0,1681
1	0,3601	0,5282
2	0,3087	0,8369
3	0,1323	0,9692
4	0,0284	0,9976
5	0,0024	1,0000

$$\begin{aligned}f_B(2; 5; 0,7) &= F_B(3; 5; 0,3) - F_B(2; 5; 0,3) \\&= 0,9692 - 0,8369 \\&= 0,1323\end{aligned}$$

Hypergeometrische Verteilung

Situation

Zufallsexperiment:

- Gesamtheit: endliche Anzahl N von Objekten
 - ▶ M Objekte mit Eigenschaft A
 - ▶ $N - M$ Objekte ohne Eigenschaft A
- zwei Ereignisse A und \bar{A}
- zufällige Auswahl von n Objekten
- Zufallsauswahlmodell ohne Zurücklegen
 - keine Unabhängigkeit der Ziehungen
 - $P(A)$, $P(\bar{A})$ nicht konstant

⇒ Zufallsvariable X : "Anzahl des Auftretens von A bei n Versuchen ohne Zurücklegen" heißt hypergeometrisch verteilt

Notation: $X \sim H(N, M, n)$

Parameter: N, M, n

Herleitung der Wahrscheinlichkeitsfunktion

1. Anzahl der Möglichkeiten n Objekte aus N Objekten zu ziehen
(ohne Zurücklegen und ohne Berücksichtigung der Anordnung)
⇒ Kombinationen ohne Wiederholung

$$\binom{N}{n}$$

2. Anzahl der Möglichkeiten, aus den M Objekten mit A genau x auszuwählen

- $x \leq M$, da nicht mehr Objekte mit A gezogen werden können, als in der Gesamtheit vorhanden sind

$$\binom{M}{x}$$

3. Anzahl der Möglichkeiten, aus den $N - M$ Objekten ohne die Eigenschaft A genau $n - x$ auszuwählen

- $n - x \leq N - M$, da nicht mehr Objekte mit \bar{A} gezogen werden können, als in der Gesamtheit vorhanden sind

$$\binom{N - M}{n - x}$$

4. Anzahl der für $\{X = x\}$ günstigen Möglichkeiten

$$\binom{M}{x} \cdot \binom{N - M}{n - x}$$

Daraus folgt nach der klassischen Definition der Wahrscheinlichkeit

$$P(X = x) = f(x) = \frac{\binom{M}{x} \cdot \binom{N - M}{n - x}}{\binom{N}{n}}$$

5. Bestimmung des Wertebereichs von X

Der größtmögliche Wert von X ist

- ▶ n , wenn $n \leq M$ ist
- ▶ M , wenn $M < n$ ist

$$\Rightarrow x_{\max} = \min(n, M)$$

Für den kleinstmöglichen Wert von X ergibt sich:

- ▶ es ist $x \geq 0$
- ▶ falls n jedoch größer ist als die Anzahl der Objekte ohne A , gilt
 $x \geq n - (N - M)$

$$\Rightarrow x_{\min} = \max[0, n - (N - M)]$$

Wahrscheinlichkeitsfunktion

$$f_H(x; N, M, n) = \begin{cases} \frac{\binom{M}{x} \cdot \binom{N-M}{n-x}}{\binom{N}{n}} & (*) \\ 0 & \text{sonst} \end{cases}$$

(*) für $x = \max[0, n - (N - M)], \dots, \min[n, M]$

Verteilungsfunktion

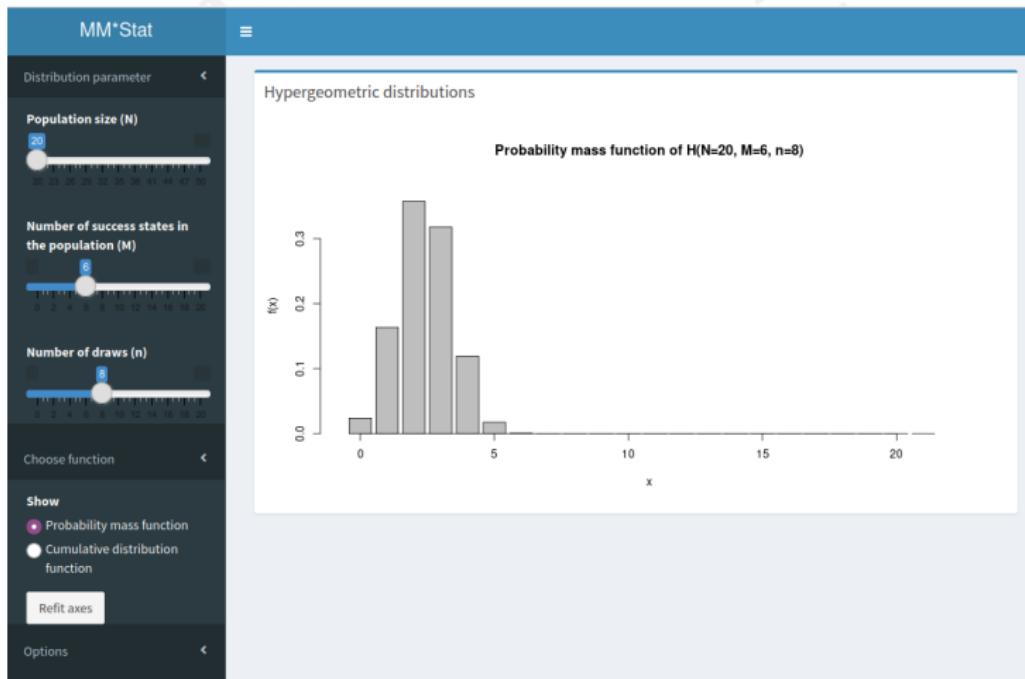
Die Werte von $F_H(x; N, M, n)$ können mithilfe der Normalverteilung approximiert werden (siehe Kapitel Approximationen)

Erwartungswert und Varianz

$$E(X) = \mu_x = n \cdot \frac{M}{N} \quad \text{Var}(X) = \sigma_x^2 = n \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right) \left(\frac{N-n}{N-1}\right)$$



examples/stat/distribution_hyper



Beispiel 14.4 (Urnenmodell)

Situation:

- Urne mit $N = 10$ Kugeln, $M = 3$ weiße Kugeln, $N - M = 7$ rote Kugeln, $n = 5$ Ziehungen ohne Zurücklegen
- $A = \{\text{Ziehen einer weißen Kugel}\}$, $\bar{A} = \{\text{Ziehen einer roten Kugel}\}$
- Ziehungen ohne Zurücklegen $\Rightarrow P(A)$ und $P(\bar{A})$ nicht konstant \Rightarrow Versuche nicht unabhängig
- W : "Anzahl des Auftretens weißer Kugeln bei $n = 5$ Ziehungen ohne Zurücklegen"

Frage: Wie groß ist die Wahrscheinlichkeit, dass genau zwei weiße Kugeln gezogen werden?

Lösung:

$$P(W = 2) = f_H(2; 10, 3, 5) = \frac{\binom{3}{2} \cdot \binom{10-3}{5-2}}{\binom{10}{5}} = \frac{3 \cdot 35}{252} = 0,4167$$

Poisson-Verteilung

Situation

- Ein Ereignis tritt auf:
 - ▶ wiederholt
 - ▶ zufällig
 - ▶ unabhängig voneinander
 - ▶ in einem Kontinuum vorgegebenen Umfangs
- ⇒ Zufallsvariable X : "Anzahl der eingetretenen Ereignisse in einem Kontinuum" heißt Poisson verteilt

Notation: $X \sim PO(\lambda)$

Parameter: λ

Wahrscheinlichkeitsfunktion

$$f_{PO}(x; \lambda) = \frac{\lambda^x}{x!} e^{-\lambda} \quad \text{für } x = 0, 1, 2, \dots; \lambda > 0$$

Verteilungsfunktion

Werte für $F_{PO}(x; \lambda)$ siehe die Tabellen in der Formelsammlung

Erwartungswert und Varianz

$$E(X) = \mu_x = \lambda \quad \text{Var}(X) = \sigma_x^2 = \lambda$$

Reproduktivitätseigenschaft

- $X \sim PO(\lambda_1)$ und $Y \sim PO(\lambda_2)$ verteilt
 - X, Y unabhängig voneinander
- $\Rightarrow Z = X + Y \sim PO(\lambda_1 + \lambda_2)$

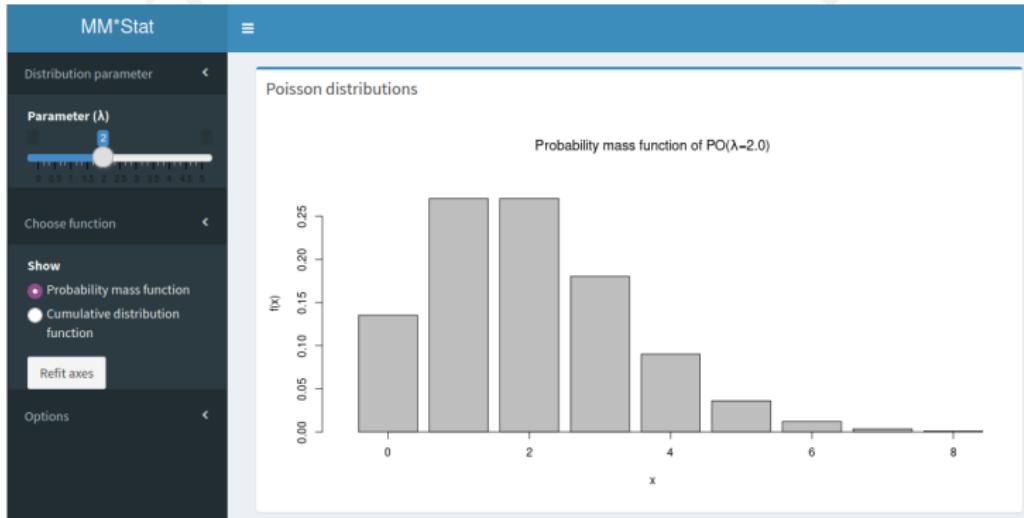
Poisson-Verteilung für Intervalle beliebigen Umfangs

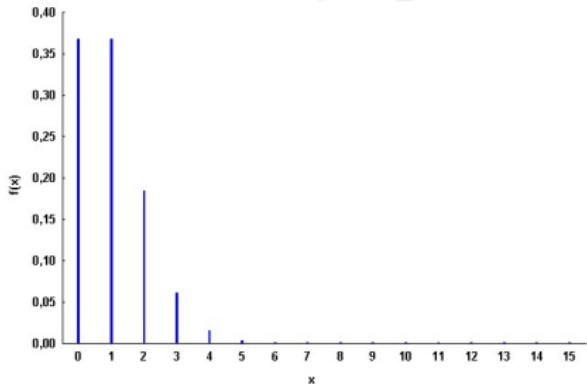
- X : "Anzahl von Ereignissen im einem Kontinuum" $PO(\lambda)$ verteilt,
- Y : "Anzahl von Ereignissen in einem Kontinuum des Umfangs t " Poisson verteilt mit dem Parameter $\lambda \cdot t$:

$$Y \sim PO(\lambda \cdot t)$$



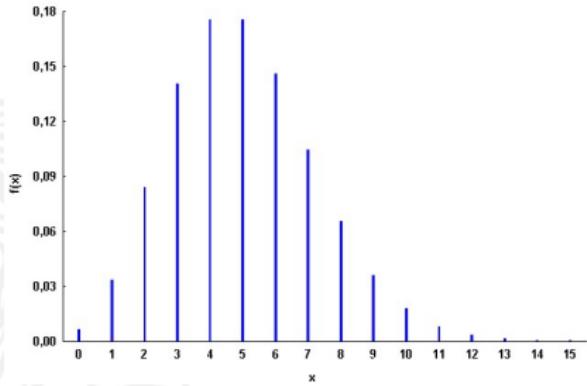
examples/stat/distribution_poisson





Poisson-Verteilung für $\lambda = 1$

Je kleiner λ , desto linkssteiler ist Verteilung



Poisson-Verteilung für $\lambda = 5$

Je größer λ , desto mehr nähert sich die Verteilung einer symmetrischen Verteilung

Beispiel 14.5 (Kundenservice eines großen Kaufhauses)

Situation

- 9 bis 14 Uhr: im Mittel 1 Kunde pro Stunde
- 14 bis 19 Uhr: im Mittel 2 Kunden pro Stunde
- Z_1 : "Anzahl der Kunden pro Stunde von 9 bis 14 Uhr" $\sim PO(\lambda_1 = 1)$
Kontinuum entspricht einer Stunde
- X_1 : "Anzahl der Kunden von 9 bis 14 Uhr" $\sim PO(\lambda_a = \lambda_1 \cdot t = 1 \cdot 5)$
Kontinuum entspricht fünf Stunden
- Z_2 : "Anzahl der Kunden pro Stunde von 14 bis 19 Uhr" $\sim PO(\lambda_2 = 2)$
Kontinuum entspricht einer Stunde
- X_2 : "Anzahl der Kunden von 14 bis 19 Uhr" $\sim PO(\lambda_b = \lambda_2 \cdot t = 2 \cdot 5)$
Kontinuum entspricht fünf Stunden

Frage 1a

Wie groß ist die Wahrscheinlichkeit, dass genau 6 Kunden in der Zeit von 9 bis 14 Uhr den Kundenservice in Anspruch nehmen?

Lösung

$$P(X_1 = 6) = f_{PO}(6; 1 \cdot 5) = \frac{(\lambda t)^x}{x!} e^{-\lambda t} = \frac{(1 \cdot 5)^6}{6!} e^{-1 \cdot 5} = 0,1462$$

Frage 1b

Wie groß ist die Wahrscheinlichkeit, dass mehr als 4 Kunden in der Zeit von

9 bis 14 Uhr den Kundenservice in Anspruch nehmen?

Lösung

$$\begin{aligned} P(X_1 > 4) &= 1 - P(X_1 \leq 4) \\ &= 1 - e^{-5} \left(\frac{5^0}{0!} + \frac{5^1}{1!} + \frac{5^2}{2!} + \frac{5^3}{3!} + \frac{5^4}{4!} \right) \\ &= 1 - 0,4405 = 0,5595 \end{aligned}$$

Frage 2a

Wie groß ist die Wahrscheinlichkeit, dass genau 6 Kunden in der Zeit von 14 bis 19 Uhr den Kundenservice in Anspruch nehmen?

Lösung

$$P(X_2 = 6) = f_{PO}(6; 2 \cdot 5) = \frac{(\lambda t)^x}{x!} e^{-\lambda t} = \frac{(2 \cdot 5)^6}{6!} e^{-2 \cdot 5} = 0,063$$

Frage 2b

Wie groß ist die Wahrscheinlichkeit, dass mehr als 4 Kunden in der Zeit von 14 bis 19 Uhr den Kundenservice in Anspruch nehmen?

Lösung

$$\begin{aligned} P(X_2 > 4) &= 1 - P(X_2 \leq 4) \\ &= 1 - e^{-10} \left(\frac{10^0}{0!} + \frac{10^1}{1!} + \frac{10^2}{2!} + \frac{10^3}{3!} + \frac{10^4}{4!} \right) \\ &= 1 - 0,0293 = 0,9707 \end{aligned}$$

Frage 3

Wie groß ist die Wahrscheinlichkeit, dass genau 12 Kunden in der Zeit von 9 bis 19 Uhr den Kundenservice in Anspruch nehmen?

Lösung

Y : Anzahl der Kunden in der gesamten Öffnungszeit von 9 bis 19 Uhr

= Anzahl der Kunden von 9 bis 14 Uhr

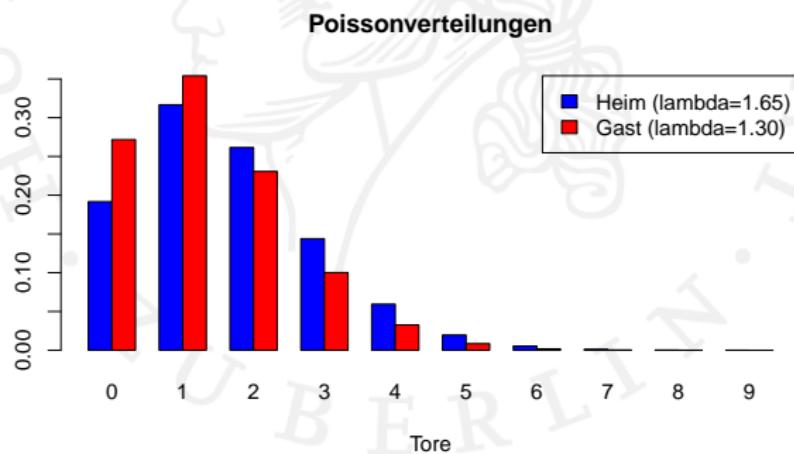
+ Anzahl der Kunden von 14 bis 19 Uhr

$$Y = X_1 + X_2, \quad Y \sim PO(\lambda_a + \lambda_b = 15) \quad (\text{Reproduktivitätseigenschaft})$$

$$P(Y = 12) = f_{PO}(12; 15) = \frac{\lambda^y}{y!} e^{-\lambda} = \frac{15^{12}}{12!} e^{-15} = 0,0829$$

Bundesliga Ergebnisse

- Spielergebnis Tore Heimmannschaft : Tore Gastmannschaft
- Tore Heimmannschaft und Tore Gastmannschaft werden modelliert mit Poissonverteilungen
- Aus historischen Daten (1. Bundesliga Saison 2010/2011 bis 2019/2020, 2754 Spiele)
 - ▶ $\lambda_{\text{Heim}} = 1,65$
 - ▶ $\lambda_{\text{Gast}} = 1,30$



- Heimmannschaft i Tore, Gastmannschaft j Tore:
 $P(H = i, G = j) = P(H = i) \cdot P(G = j)$
- Ergebnis basierend auf dem Produkt der Wk. der Poissonverteilungen
(in Prozent)

Heim \ Gast	0	1	2	3	4	5	6	7	8	9
0	5.2	6.8	4.4	1.9	0.6	0.2	0.0	0.0	0.0	0.0
1	8.6	11.2	7.3	3.2	1.0	0.3	0.1	0.0	0.0	0.0
2	7.1	9.3	6.0	2.6	0.9	0.2	0.0	0.0	0.0	0.0
3	3.9	5.1	3.3	1.4	0.5	0.1	0.0	0.0	0.0	0.0
4	1.6	2.1	1.4	0.6	0.2	0.1	0.0	0.0	0.0	0.0
5	0.5	0.7	0.5	0.2	0.1	0.0	0.0	0.0	0.0	0.0
6	0.1	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

- Kreuztabelle der tatsächlich beobachteten Spielergebnisse in der 1. Bundesliga (Saison 2010/2011 bis 2019/2020, 2754 Spiele)

Heim \ Gast	0	1	2	3	4	5	6	7	8	9
0	6.4	5.6	4.6	2.7	1.2	0.3	0.1	0.0	0.0	0.0
1	7.5	11.2	6.4	3.5	1.7	0.4	0.2	0.0	0.0	0.0
2	7.6	8.6	5.1	2.0	0.7	0.3	0.1	0.0	0.0	0.0
3	4.5	5.4	2.5	1.5	0.2	0.1	0.0	0.0	0.0	0.0
4	2.2	2.0	1.1	0.2	0.2	0.1	0.0	0.0	0.0	0.0
5	0.9	0.8	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
6	0.4	0.3	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
8	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

$$K^2 = 0.03, C = 0.18$$

- Beide Tabellen stimmen (fast) überein

Wiederholung: stetige Zufallsvariablen

- Zufallsvariable: X
- Verteilungsfunktion

$$F(x) = P(-\infty < X \leq x) = \int_{-\infty}^x f(t) dt$$

- Dichtefunktion:

$$f(x), \text{ so dass } P(a < X \leq b) = \int_a^b f(x) dx \quad \text{für } a \leq b$$

- Erwartungswert:

$$E(X) = \mu_x = \int_{-\infty}^{+\infty} x \cdot f(x) dx$$

- Varianz:

$$\text{Var}(X) = \sigma_x^2 = \int_{-\infty}^{+\infty} [x - E(X)]^2 \cdot f(x) dx = \int_{-\infty}^{+\infty} x^2 f(x) dx - [E(X)]^2$$

Stetige Gleichverteilung

Situation

Eine Zufallsvariable X heißt stetig gleichverteilt, wenn

- sie nur Werte im Intervall $[a, b]$ annimmt und
- sie hat eine positive und konstante Wahrscheinlichkeitsdichte in diesem Intervall

Notation: $X \sim U(a, b)$

Parameter: a, b

Dichtefunktion

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{für } a \leq x \leq b \\ 0 & \text{sonst} \end{cases}$$

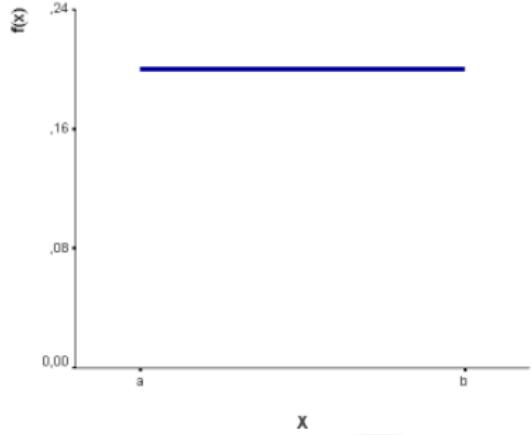
Verteilungsfunktion II

$$F(x) = \begin{cases} 0 & \text{für } x < a \\ \frac{x-a}{b-a} & \text{für } a \leq x < b \\ 1 & \text{für } b \leq x \end{cases}$$

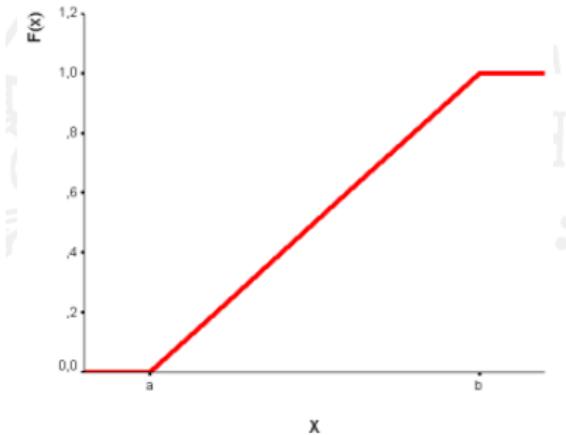
Erwartungswert und Varianz

$$E(X) = \mu_x = \frac{b+a}{2} \quad \text{Var}(X) = \sigma_x^2 = \frac{(b-a)^2}{12}$$

Dichtefunktion



Verteilungsfunktion





Beispiel 14.6 (Wartezeit auf den nächsten S-Bahn-Zug)

Situation

- Die S-Bahn fährt regelmäßig alle 20 Minuten
- X : "Wartezeit auf den nächsten S-Bahn-Zug in Minuten" $\sim U(0, 20)$

Frage

Wie sehen die Dichte- und Verteilungsfunktion aus? Was ist der Erwartungswert und die Varianz?

Lösung

$$f(x) = \begin{cases} \frac{1}{20} & 0 < x \leq 20 \\ 0 & \text{sonst} \end{cases}$$

$$E(X) = \frac{20+0}{2} = 10$$

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{20} \cdot x & 0 \leq x < 20 \\ 1 & \text{sonst} \end{cases}$$

$$\text{Var}(X) = \frac{(20-0)^2}{12} = 33,33$$

Exponentialverteilung

Situation

- Ein Ereignis tritt auf:
 - ▶ wiederholt
 - ▶ zufällig
 - ▶ unabhängig voneinander
 - ▶ in einem Kontinuum vorgegebenen Umfangs

⇒ Zufallsvariable X : "Wartezeit auf das nächste Ereignis in einem Kontinuum" heißt exponential verteilt

Notation: $X \sim EX(\lambda)$

Parameter: λ

 Obige Voraussetzungen sind identisch zu den Voraussetzungen der Poissonverteilung, aber der Inhalt der Zufallvariablen unterscheidet sich

Herleitung der Verteilungsfunktion

- Y : "Anzahl des Ereignisses in einem Kontinuum" $\sim PO(\lambda)$
- Z : "Anzahl des Ereignisses in einem Kontinuum der Länge t "
 $\sim PO(\lambda \cdot t)$
- X : "Intervall zwischen zwei aufeinander folgenden Ereignissen"

$P(\text{Intervall zwischen zwei aufeinanderfolgenden Ereignissen} \leq t)$

$$\begin{aligned}&= P(X \leq t) = 1 - P(\text{kein Ereignis im Intervall der Länge } t) \\&= 1 - P(Z = 0) = 1 - f_{PO}(0; \lambda t) \\&= 1 - \frac{(\lambda t)^0}{0!} e^{-\lambda t} = 1 - e^{-\lambda t} \\&\Rightarrow P(X \leq t) = F(t) = 1 - e^{-\lambda xt}\end{aligned}$$

Dichtefunktion

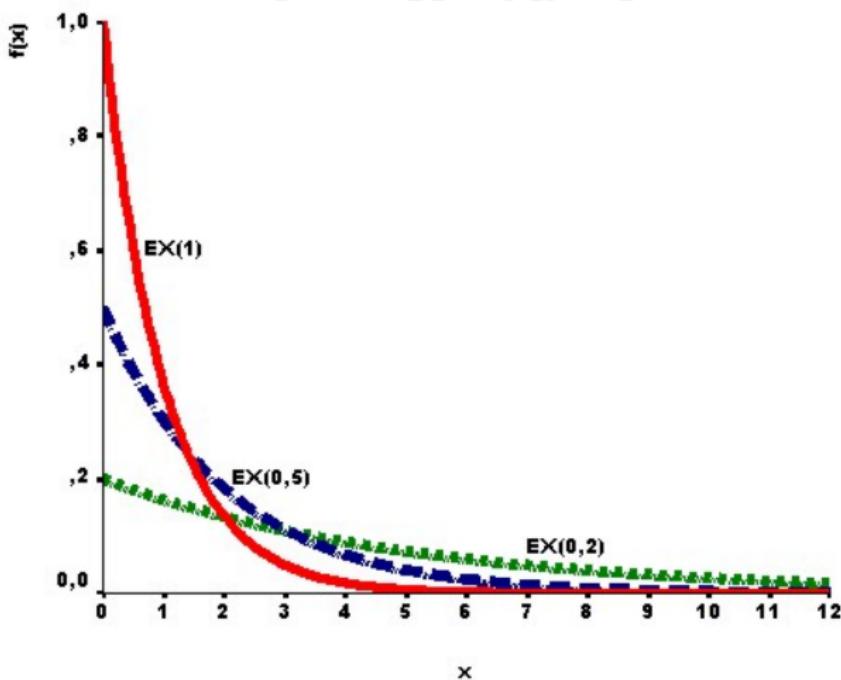
$$f_{EX}(x; \lambda) = \begin{cases} \lambda \cdot e^{-\lambda x} & \text{für } x \geq 0, \lambda > 0 \\ 0 & \text{für } x < 0 \end{cases}$$

Verteilungsfunktion

$$F_{EX}(x; \lambda) = \begin{cases} 1 - e^{-\lambda x} & \text{für } x \geq 0 \\ 0 & \text{für } x < 0 \end{cases}$$

Erwartungswert und Varianz

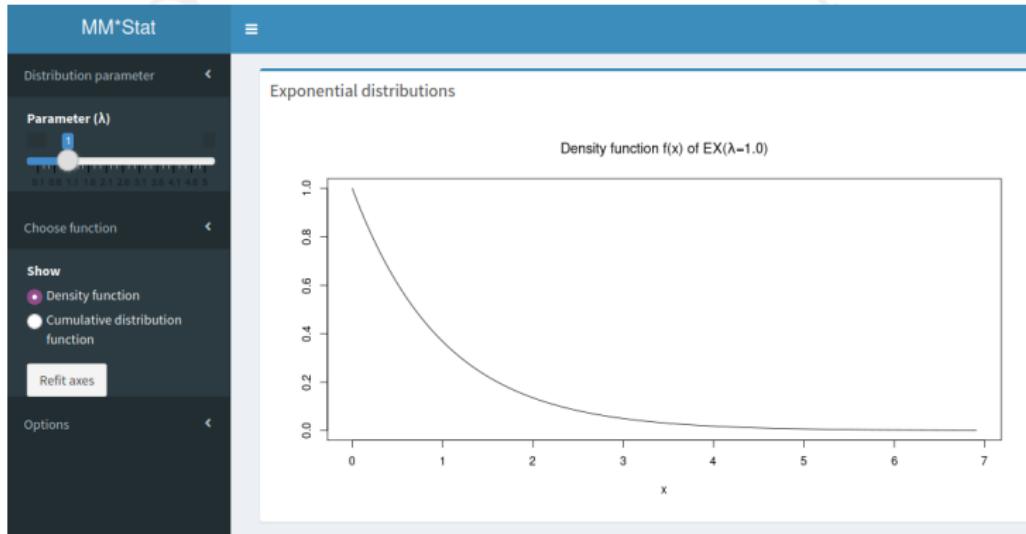
$$E(X) = \mu_x = \frac{1}{\lambda} \quad \text{Var}(X) = \sigma_x^2 = \frac{1}{\lambda^2}$$



für $\lambda = 1$, $\lambda = 0,5$ und $\lambda = 0,2$



examples/stat/distribution_exponential



Beispiel 14.7 (Defekte an einer Maschine)

Situation

Pro Woche gibt es durchschnittlich 2 Defekte an einer Maschine.

Frage 1

Wie groß ist die Wahrscheinlichkeit, dass es in einer Woche, in zwei Wochen bzw. in t Wochen keinen Defekt gibt?

Lösung

- $P(\text{in einer Woche kein Defekt}) = ?$

Y_1 : "Anzahl der Defekte pro Woche"

$$t = 1 \text{ Woche}; \quad E(Y_1) = \lambda t = 2; \quad Y_1 \sim PO(2)$$

$$P(Y_1 = 0) = \frac{(\lambda t)^{y_1}}{y_1!} e^{-\lambda t} = \frac{(2 \cdot 1)^0}{0!} e^{-2 \cdot 1} = e^{-2} = 0,1353$$

- $P(\text{in zwei Wochen kein Defekt}) = ?$

Y_2 : "Anzahl der Defekte in zwei Wochen"

$$t = 2 \text{ Wochen}; \quad E(Y_2) = \lambda t = 2 \cdot 2 = 4; \quad Y_2 \sim PO(4)$$

$$P(Y_2 = 0) = \frac{(2 \cdot 2)^0}{0!} e^{-2 \cdot 2} = \frac{4^0}{0!} e^{-4} = e^{-4} = 0,0183$$

- $P(\text{in } t \text{ Wochen kein Defekt}) = ?$

Y_t : "Anzahl der Defekte in t Wochen"

$$E(Y_t) = \lambda t \quad Y_t \sim PO(\lambda t)$$

$$P(Y_t = 0) = \frac{(\lambda t)^0}{0!} e^{-\lambda t} = e^{-\lambda t}$$

Frage 2

Wie groß ist die Wahrscheinlichkeit, dass die Wartezeit bis zum nächsten Defekt mehr als 2 Wochen beträgt?

Lösung

$P(\text{Wartezeit bis zum nächsten Defekt beträgt mehr als 2 Wochen}) = ?$

X : "Wartezeit bis zum nächsten Defekt"

$X \sim EX(\lambda)$ mit $\lambda = 2$

$$\begin{aligned} P(X > 2) &= 1 - P(X \leq 2) = 1 - F_{EX}(x; \lambda) \\ &= 1 - (1 - e^{-\lambda x}) = e^{-\lambda x} = e^{-2 \cdot 2} = 0,0183 \end{aligned}$$

$$\Rightarrow P(X > 2) = P(Y_2 = 0)$$

Normalverteilung

Notation: $X \sim N(\mu_x, \sigma_x)$

Parameter: μ_x, σ_x

Dichtefunktion

$$f_{NV}(x; \mu_x, \sigma_x) = \frac{1}{\sigma_x \sqrt{2\pi}} e^{-(x-\mu_x)^2/2\sigma_x^2} \quad \text{für } -\infty < x < +\infty, \quad \sigma_x > 0$$

Verteilungsfunktion

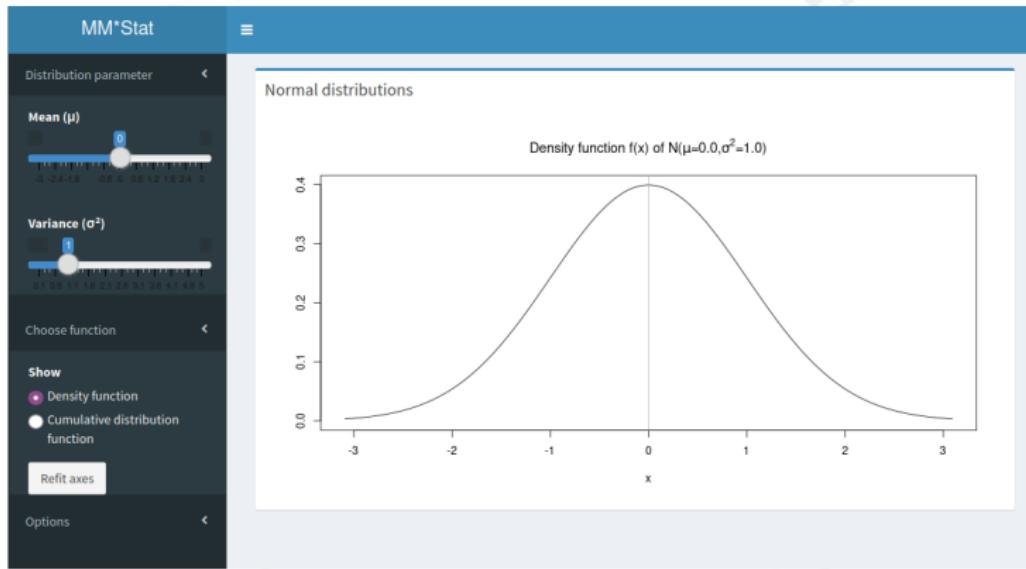
$$F_{NV}(x; \mu_x, \sigma_x) = \frac{1}{\sigma_x \sqrt{2\pi}} \int_{-\infty}^x e^{-(t-\mu_x)^2/2\sigma_x^2} dt$$

Erwartungswert und Varianz

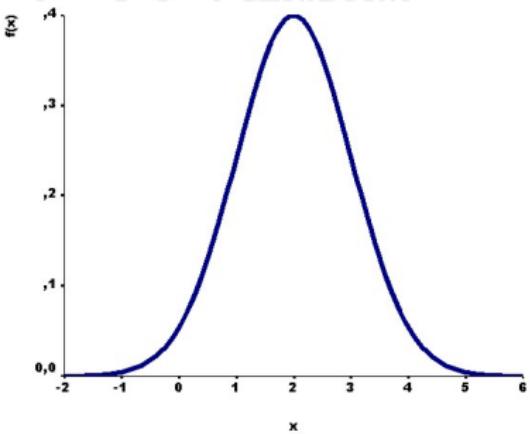
$$E(X) = \mu_x \quad \text{Var}(X) = \sigma_x^2$$



examples/stat/distribution_normal



1. Die Gestalt der Normalverteilung hängt von den Parametern μ_x und σ_x ab
2. Für $x \rightarrow -\infty$ und $x \rightarrow \infty$ nähert sich die Dichtefunktion asymptotisch dem Wert 0
3. Globales Maximum im Punkt $x = \mu_x$ mit $f_{NV}(x; \mu_x, \sigma_x) = \frac{1}{\sigma_x \sqrt{2\pi}}$



4. Die Dichtefunktion ist symmetrisch im Punkt $x = \mu_x$
5. Linearkombination

Sei

- ▶ $X \sim N(\mu_x, \sigma_x)$
- ▶ $Y = a + bX$ mit $b \neq 0$

Dann ist

- ▶ $Y \sim N(a + b\mu_x; |b| \cdot \sigma_x)$
mit
 - ★ $E(Y) = E(a + bX) = a + b \cdot E(X) = a + b \cdot \mu_x$
 - ★ $\text{Var}(Y) = \text{Var}(a + bX) = b^2 \cdot \text{Var}(X) = b^2 \sigma_x^2$

Reproduktivitätseigenschaft

- zwei Zufallsvariablen X und Y
 - ▶ unabhängig voneinander
 - ▶ normalverteilt: $X \sim N(\mu_x, \sigma_x^2)$, $Y \sim N(\mu_y, \sigma_y^2)$

\Rightarrow

$$Z = X + Y \sim N\left(\mu_x + \mu_y, \sqrt{\sigma_x^2 + \sigma_y^2}\right)$$

Gewichtete Linearkombination

- n Zufallsvariablen X_1, \dots, X_n
 - ▶ unabhängig voneinander
 - ▶ normalverteilt: $X_i \sim N(\mu_i, \sigma_i^2)$, $E(X_i) = \mu_i$, $Var(X_i) = \sigma_i^2$
- $Y = a_1X_1 + \dots + a_nX_n$ mit $a_i \neq 0$ für mindestens ein i

\Rightarrow

$$Y = \sum_{i=1}^n a_i X_i \sim N\left(\sum_{i=1}^n a_i \mu_i, \sqrt{\sum_{i=1}^n a_i^2 \sigma_i^2}\right)$$

Standardnormalverteilung

Situation

Eine Zufallsvariable Z

- ist normalverteilt
 - mit Erwartungswert $\mu_x = 0$ und Varianz $\sigma_x^2 = 1$
- ⇒ Z heißt standardnormal verteilt

Notation: $Z \sim N(0, 1)$

Parameter: keine

Dichtefunktion

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

Erwartungswert und Varianz

$$E(Z) = \mu_z = 0 \quad \text{Var}(Z) = \sigma_z^2 = 1$$

Verteilungsfunktion $\Phi(z)$

Werte für $\Phi(z)$ siehe Tabelle in der Formelsammlung

Normalverteilung \leftrightarrow Standardnormalverteilung

$$X \sim N(\mu_x; \sigma_x) \Rightarrow Z = \frac{X - \mu_x}{\sigma_x} \sim N(0; 1) \quad (\text{Standardisierung})$$

$$Z \sim N(0; 1) \Rightarrow X = \mu_x + \sigma_x Z \sim N(\mu_x; \sigma_x)$$

$$\begin{aligned} F_{NV}(x; \mu_x, \sigma_x) &= P(X \leq x) \\ &= P\left(\frac{X - \mu_x}{\sigma_x} \leq \frac{x - \mu_x}{\sigma_x}\right) \\ &= P(Z \leq z) = \Phi(z) \end{aligned}$$



Auf diese Art können für jede beliebige Normalverteilung die Werte der Verteilungsfunktion berechnet werden

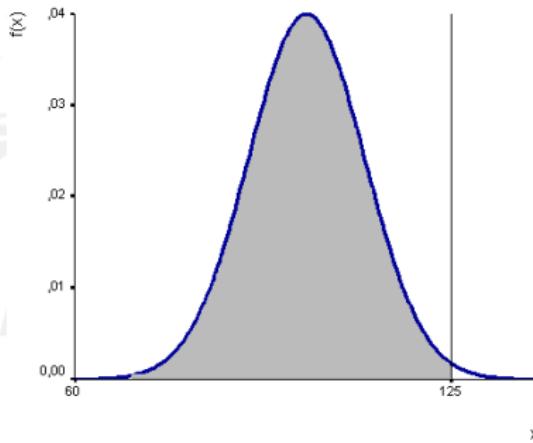
Beispiel 14.8

Gegeben: $X \sim N(100; 10)$

a) Gesucht ist die Wahrscheinlichkeit $P(X \leq x)$ mit $x = 125$

$$z = \frac{x - \mu_x}{\sigma_x} = \frac{125 - 100}{10} = 2,5$$

$$\begin{aligned} P(X \leq 125) &= F(125) \\ &= \Phi\left(\frac{125 - 100}{10}\right) \\ &= \Phi(2,5) \\ &= 0,99379 \end{aligned}$$



b) Gesucht ist die Wahrscheinlichkeit $P(X \geq x)$ mit $x = 115,6$:

$$z = \frac{x - \mu_x}{\sigma_x} = \frac{115,6 - 100}{10} = 1,56$$

$$P(X \geq 115,6)$$

$$= 1 - P(X \leq 115,6)$$

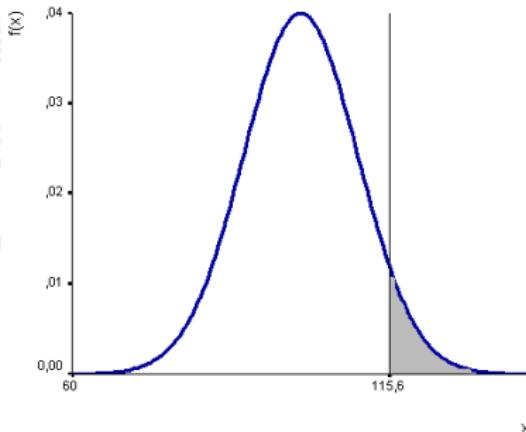
$$= 1 - F(115,6)$$

$$= 1 - \Phi\left(\frac{115,6 - 100}{10}\right)$$

$$= 1 - \Phi(1,56)$$

$$= 1 - 0,94062$$

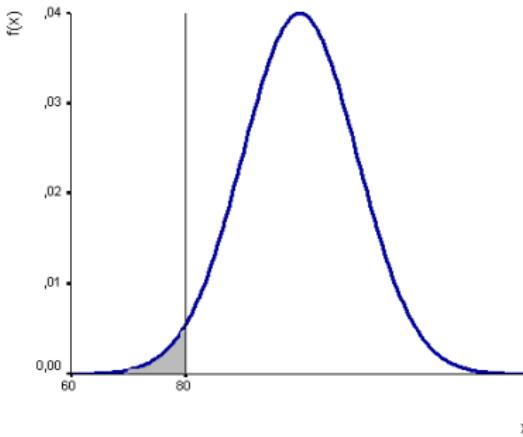
$$= 0,05938$$



c) Gesucht ist die Wahrscheinlichkeit $P(X \leq x)$ mit $x = 80$

$$z = \frac{x - \mu_x}{\sigma_x} = \frac{80 - 100}{10} = -2$$

$$\begin{aligned} P(X \leq 80) &= F(80) \\ &= \Phi\left(\frac{80 - 100}{10}\right) \\ &= \Phi(-2) \\ &= 1 - \Phi(2) \\ &= 1 - 0,97725 \\ &= 0,02275 \end{aligned}$$

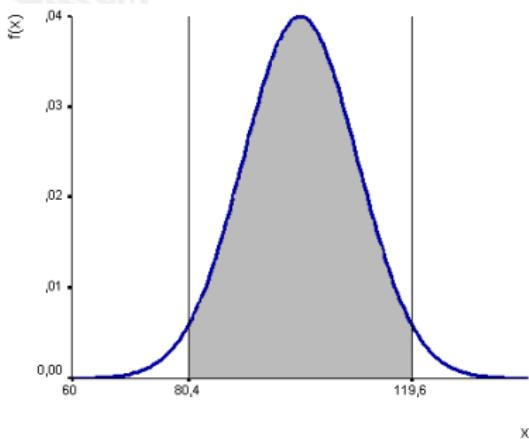


d) Gesucht ist die Wahrscheinlichkeit $P(x_u \leq X \leq x_o)$ mit $x_u = 80,4$ und $x_o = 119,6$

$$z_u = (x_u - \mu_x)/\sigma_x = (80,4 - 100)/10 = -1,96$$

$$z_o = (x_o - \mu_x)/\sigma_x = (119,6 - 100)/10 = 1,96$$

$$\begin{aligned} P(80,4 \leq X \leq 119,6) &= P(X \leq 119,6) - P(X \leq 80,4) \\ &= F(119,6) - F(80,4) \\ &= \Phi(1,96) - \Phi(-1,96) \\ &= \Phi(1,96) - (1 - \Phi(1,96)) \\ &= 2\Phi(1,96) - 1 \\ &= 2 \cdot 0,975 - 1 = 0,95 \end{aligned}$$



Zentraler Grenzwertsatz

Voraussetzungen

- X_1, X_2, \dots, X_n seien identisch verteilte, unabhängige Zufallsvariablen
- $E(X_i) = \mu_x$ und $\text{Var}(X_i) = \sigma_x^2 > 0$ für $i = 1, \dots, n$
- $|\mu_x| < \infty$ und $\sigma_x^2 < \infty$

Satz

Die Verteilung der standardisierten Zufallsvariablen konvergiert mit steigendem n gegen die Standardnormalverteilung:

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{X_i - \mu_x}{\sigma_x} \Rightarrow \lim_{n \rightarrow \infty} P(Z_n \leq z) = \Phi(z).$$

Konsequenz

$$S_n = \sum_{i=1}^n X_i \approx N(n\mu_x; \sigma_x \cdot \sqrt{n})$$

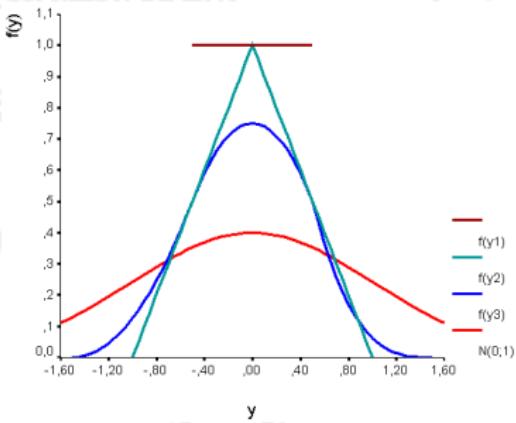
$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \approx N(\mu_x; \sigma_x / \sqrt{n})$$

Beispiel 14.9 (ZGS bei gleichverteilten Zufallsvariablen)

X_1, X_2, \dots seien stetige Zufallsvariablen, die unabhängig und identisch gleichverteilt im Intervall $[-0,5; 0,5]$ sind:

$$f(x) = \begin{cases} 1 & \text{für } -0,5 \leq x \leq 0,5 \\ 0 & \text{sonst.} \end{cases}$$

$$\begin{aligned} Y_1 &= X_1 \\ Y_2 &= X_1 + X_2 \\ Y_3 &= X_1 + X_2 + X_3 \end{aligned}$$



Chi-Quadrat-Verteilung

Situation

Die Zufallsvariablen Z_1, \dots, Z_n , sind

- voneinander unabhängig und
 - standardnormalverteilt: $Z_i \sim N(0; 1)$
- ⇒ Die Zufallsvariable

$$X = Z_1^2 + \dots + Z_n^2$$

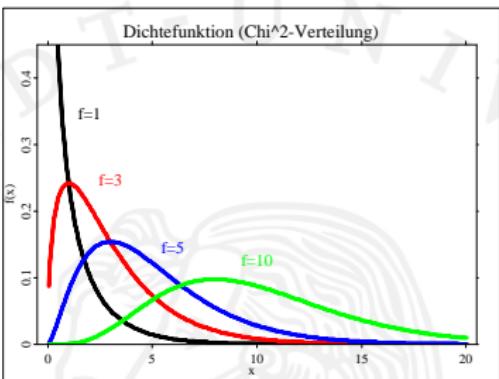
heißt Chi-Quadrat-verteilt mit $f = n$ Freiheitsgraden

Notation: $X \sim \chi_f^2$

Parameter: f

Erwartungswert und Varianz

$$E(X) = \mu_x = f \quad \text{Var}(X) = \sigma_x^2 = 2f$$



- die Zufallsvariable X kann nur positive Werte annehmen
- die Form der Wahrscheinlichkeitsdichte der Chi-Quadrat-Verteilung hängt von dem Parameter f ab
 - ▶ f ist die Anzahl n der unabhängigen Zufallsvariablen Z_i , die in die Summenbildung eingehen, d.h. die Anzahl der Freiheitsgrade
 - ▶ für $f = 1$ und $f = 2$ fällt die Dichtefunktion monoton
 - ▶ für kleine Werte von f sind die Dichtefunktionen deutlich rechtsschief
 - ▶ für wachsende Werte von f strebt die Dichte gegen die Dichtefunktion der Normalverteilung (Zentraler Grenzwertsatz !)

t-Verteilung

Situation

- Z ist eine standardnormalverteilte Zufallsvariable: $Z \sim N(0, 1)$
 - Y ist eine
 - ▶ von Z unabhängige
 - ▶ $Y \sim \chi_f^2$
- ⇒ Die Zufallsvariable

$$T = \frac{Z}{\sqrt{\frac{Y}{f}}}$$

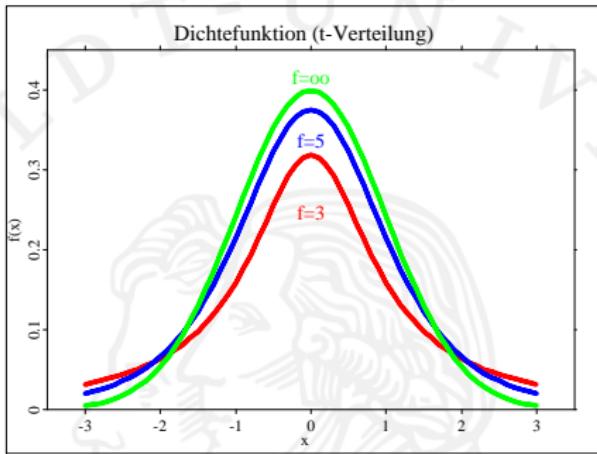
heißt t verteilt mit f Freiheitsgraden

Notation: $T \sim t_f$

Parameter: f

Erwartungswert und Varianz

$$E(T) = \mu_T = 0 \quad \text{Var}(T) = \sigma_T^2 = \frac{f}{f-2} \quad \text{für } f > 2$$



- die Dichtefunktion der t-Verteilung ist eine symmetrische Glockenkurve zum Erwartungswert $E(T) = 0$
- die Dichtefunktion der t-Verteilung ist flacher als die der Standardnormalverteilung
- für $f \rightarrow \infty$ konvergiert die Dichtefunktion der t-Verteilung gegen die Dichtefunktion der Standardnormalverteilung

F-Verteilung

Situation

Zwei Zufallsvariablen Y_1 und Y_2 sind

- unabhängig voneinander
- chi-quadrat-verteilt mit f_1 bzw. f_2 Freiheitsgraden
 - ▶ $Y_1 \sim \chi^2(f_1)$
 - ▶ $Y_2 \sim \chi^2(f_2)$

⇒ Die Zufallsvariable

$$F = \frac{Y_1/f_1}{Y_2/f_2}$$

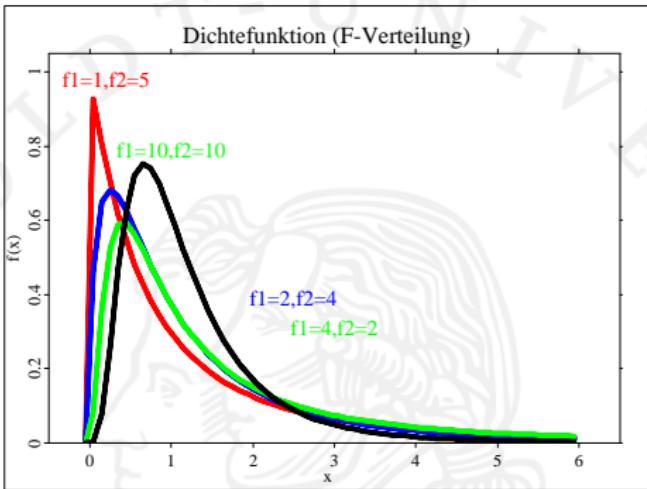
heißt F verteilt mit f_1 und f_2 Freiheitsgraden

Notation: $F \sim F_{f_1, f_2}$

Parameter: f_1, f_2

Erwartungswert und Varianz

$$E(F) = \mu_F = \frac{f_2}{f_2 - 2} \quad \text{Var}(F) = \sigma_F^2 = \frac{2f_2^2(f_1 + f_2 - 2)}{f_1(f_2 - 2)^2(f_2 - 4)}$$



- die Zufallsvariable F nimmt nur positive Werte an
- die Dichtefunktion der F-Verteilung ist rechtsschief
- für $f = f_1 = f_2$ und $f \rightarrow \infty$ strebt die Dichte der F -Verteilung gegen die Dichtefunktion einer Normalverteilung
- für $f_2 \leq 4$ sind die Bedingungen des ZGS nicht erfüllt

Verteilungen

Mathematik	Induktive Statistik
Zahlen	Verteilungen
0	Binomialverteilung $B(N; p)$
1	Hypergeometrische Verteilung $H(N; M; p)$
2	Poisson-Verteilung $Po(\lambda)$
π	Exponentialverteilung $Exp(\lambda)$
e	Normalverteilung $N(\mu, \sigma)$
...	...
Rechenregeln	(Voraussetzungen beachten!)
$1 + 1 = 2$	$N(0; 1) + N(0; 1) = N(0; \sqrt{2})$
$2 + 2 = 4$	$Po(2) + Po(2) = Po(4)$
$3 + 2 \cdot 1 = 5$	$3 + 2 \cdot N(0; 1) = N(3; 2)$
$1^2 + 1^2 = 2$	$N(0; 1)^2 + N(0; 1)^2 = \chi_2^2$
$\frac{1}{2} = 0,5$	$\frac{N(0; 1)}{\sqrt{\chi_2^2/2}} = t_2$
$e^{i\pi} + 1 = 0$	$X_1 + \dots + X_n \approx N(\bullet; \bullet)$ (wenn ZGS gilt)

Approximation von Verteilungen

- unter bestimmten Bedingungen kann statt der Ausgangsverteilung eine einfacher handhabbare Verteilung verwendet werden
- entsprechende Grenzwertsätze (z.B. der zentrale Grenzwertsatz) liefern die theoretischen Grundlagen
- für eine hinreichend gute Approximation müssen entsprechende Voraussetzungen eingehalten werden (siehe Formelsammlung)
 - ▶ Je nach Genauigkeit der Approximation sind andere Approximationsbedingungen möglich
- Allgemein gilt:
 - ▶ die Erwartungswerte der beiden Verteilungen müssen übereinstimmen
 - ▶ die Varianzen der beiden Verteilungen müssen übereinstimmen

$$B(n, p) \rightarrow Po(\lambda)$$

Voraussetzungen

Bernoulli-Experiment mit

- Anzahl n der unabhängigen Versuche sehr gross: $n \rightarrow \infty$
- Wahrscheinlichkeit $P(A) = p$ für das Eintreten des Ereignisses A sehr klein: $p \rightarrow 0$

Faustregel: $n > 10$ und $p < 0,05$

Approximation

$$X \sim B(n; p) \Rightarrow X \approx Po(\lambda = n \cdot p)$$

Begründung

Die mittlere Anzahl der Ereignisse pro Kontinuum sei λ , also $X \sim Po(\lambda)$

- \Leftrightarrow Unterteile das Kontinuum in viele gleichgroße Teile I_1, \dots, I_n ($n \gg \lambda$)
- \Leftrightarrow Die Wahrscheinlichkeit, dass ein Ereignis in einem Teil I_i auftritt ist λ/n
- \Leftrightarrow Die Anzahl der Ereignisse in n Teilen (Versuchen) ist $X \sim B(n; \lambda/n)$



Listing 14.3: example_approx_binomial_poisson.R

```
1 n <- 20
2 p <- 0.01
3 (n>10) && (p<0.05)
4 x <- 0:n
5 freq <- dbinom(0:n, n, p)
6 plot(x, freq, type="h",
7           main="Binomialverteilung -> Poissonverteilung")
8 points(x, freq, pch=19)
9 pois <- dpois(x, lambda=n*p)
10 lines(x+0.1, pois, type="h", col="red")
11 points(x+0.1, pois, pch=19, col="red")

---


```

Beispiel 14.10 (Impfschaden)

$P(\text{Person erleidet Impfschäden}) = 0,001$

Bernoulli-Experiment:

- $A = \{\text{Eintreten eines Impfschadens}\}$
 $\bar{A} = \{\text{kein Impfschaden}\}$
- $P(A) = 0,001$ ist konstant
- Unabhängigkeit der Versuche
- $n = 2000$ zufällig ausgewählte Personen

$X = \{\text{Anzahl der Patienten mit Impfschaden bei } n = 2000 \text{ Patienten}\}$

$X \sim B(2000; 0,001)$

Frage

Wie groß ist die Wahrscheinlichkeit, dass genau 3 Personen einen Impfschaden erleiden?

Gesucht: $P(X = 3)$

Lösung

p sehr klein, n sehr gross: Approximation durch die Poisson-Verteilung
 $\lambda = np = 2000 \cdot 0,001 = 2$ (im Mittel zu erwartende Anzahl von Impfschäden)

$$f_{PO}(x; \lambda) = \frac{\lambda^x}{x!} e^{-\lambda}$$

$$f_{PO}(3; 2) = \frac{2^3}{3!} e^{-2} = 0,18045$$

$$f_B(3; 2000; 0,001) = \binom{2000}{3} \cdot 0,001^3 \cdot 0,999^{1997} = 0,1805$$

$$B(n, p) \rightarrow N(\mu, \sigma)$$

Voraussetzungen

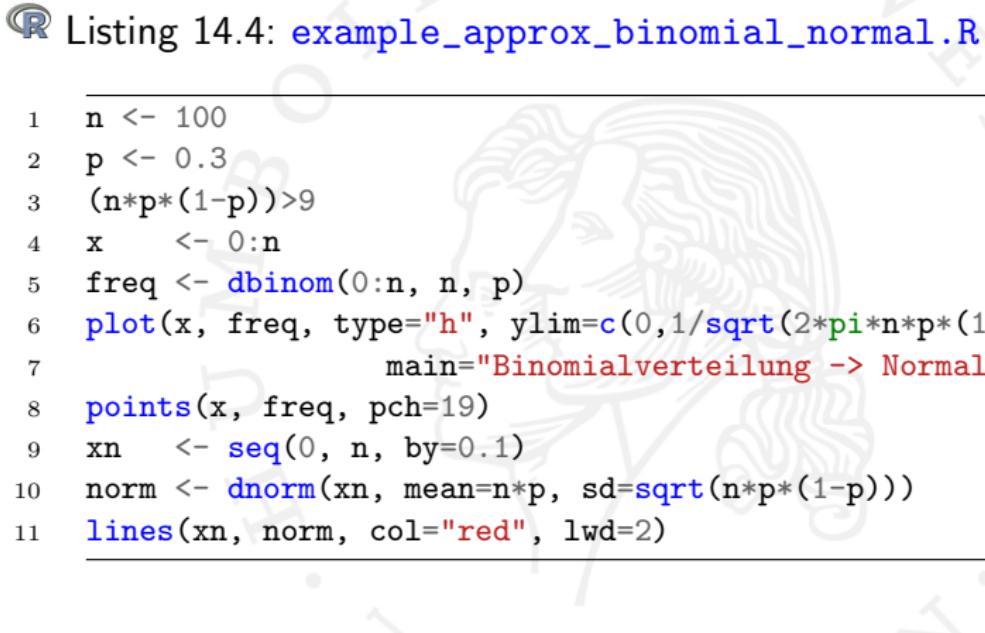
Bernoulli-Experiment mit

- Anzahl n der unabhängigen Versuche sehr groß: $n \rightarrow \infty$
- Wahrscheinlichkeit $P(A) = p$ für das Eintreten eines Ereignisses A nicht zu klein

Faustregel: $n \cdot p \cdot (1 - p) \geq 9$

Approximation

$$X \sim B(n; p) \Rightarrow X \approx N\left(\mu_x = n \cdot p, \sigma_x = \sqrt{n \cdot p \cdot (1 - p)}\right)$$

R Listing 14.4: example_approx_binomial_normal.R

```
1 n <- 100
2 p <- 0.3
3 (n*p*(1-p))>9
4 x <- 0:n
5 freq <- dbinom(0:n, n, p)
6 plot(x, freq, type="h", ylim=c(0,1/sqrt(2*pi*n*p*(1-p))),
7           main="Binomialverteilung -> Normalverteilung")
8 points(x, freq, pch=19)
9 xn <- seq(0, n, by=0.1)
10 norm <- dnorm(xn, mean=n*p, sd=sqrt(n*p*(1-p)))
11 lines(xn, norm, col="red", lwd=2)
```

$$t_n \rightarrow N(0, 1)$$

Voraussetzung

- Anzahl der Freiheitsgrade f sehr groß: $f \rightarrow \infty$

Faustregel: $f \geq 30$

Approximation

$$X \sim t_f \Rightarrow X \approx N(0; 1)$$

$$H(N, M, n) \rightarrow B(n, p)$$

Voraussetzungen

Zufallsexperiment mit

- Anzahl N der Objekte sehr groß: $N \rightarrow \infty$
- Anzahl M der Objekte mit Eigenschaft A sehr groß: $M \rightarrow \infty$

Faustregel: $\frac{n}{N} \leq 0,05$

Approximation

$$X \sim H(N, M, n) \Rightarrow X \approx B\left(n; p = \frac{M}{N}\right)$$



Listing 14.5: example_approx_hyper_binomial.R

```
1 N <- 1000
2 M <- 400
3 n <- 15
4 x <- 0:n
5 (n/N<0.05)
6 freq <- dhyper(x, m=M, n=N-M, k=n)
7 plot(x, freq, type="h",
8           main="Hypergeo. Verteilung -> Binomialverteilung")
9 points(x, freq, pch=19)
10 bin <- dbinom(x, size=n, prob=M/N)
11 lines(x+0.1, bin, type="h", col="red")
12 points(x+0.1, bin, pch=19, col="red")

---


```

$$H(N, M, n) \rightarrow N(\mu, \sigma)$$

Voraussetzungen

Zufallsexperiment mit

- Anzahl n der Versuche sehr groß: $n \rightarrow \infty$
- Anzahl N der Objekte sehr groß: $N \rightarrow \infty$
- Anzahl M der Objekte mit Eigenschaft A sehr groß: $M \rightarrow \infty$

Faustregel: $n \frac{M}{N} \left(1 - \frac{M}{N}\right) \geq 9, \quad N \geq 2$

Approximation

$$X \sim H(N, M, n) \Rightarrow X \approx N \left(\mu_x = n \cdot \frac{M}{N}; \sigma_x = \sqrt{n \frac{M}{N} \left(1 - \frac{M}{N}\right) \frac{N-n}{N-1}} \right)$$

Stetigkeitskorrektur

Grund für eine Stetigkeitskorrektur:

- genauere Approximation

Eine Stetigkeitskorrektur ist notwendig, wenn

- eine Binomialverteilung,
- eine Hypergeometrische Verteilung oder
- eine Poissonverteilung durch eine Normalverteilung approximiert wird und
- die Varianz der Normalverteilung $\sigma^2 \leq 9$ ist

Eine Stetigkeitskorrektur wird durchgeführt, indem

- von der unteren Grenze 0,5 abgezogen wird
- zu der oberen Grenze 0,5 hinzugeaddiert wird

Zusammenfassung

Verteilungen	$E(X)$	$Var(X)$
diskret		
$U(n)$	$\frac{1}{n} \sum_{i=1}^n x_i$	$\frac{1}{n} \sum_{i=1}^n (x_i - \mu_x)^2$
$B(p)$	p	$p(1-p)$
$B(n; p)$	np	$np(1-p)$
$H(N; M; n)$	$n \frac{M}{N}$	$n \frac{N}{M} \left(1 - \frac{N}{M}\right)$
$PO(\lambda)$	λ	λ
stetig		
$U(a, b)$	$\frac{b+a}{2}$	$\frac{(b-a)^2}{12}$
$EX(\lambda)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
$N(\mu; \sigma)$	μ	σ^2
χ_f^2	f	$2f$
t_f	0	$\frac{f}{f-2}$ (für $f > 2$)
$F_{f_1; f_2}$	$\frac{f_2}{f_2-2}$ (für $f_2 > 2$)	$\frac{2f_2^2(f_1+f_2-2)}{f_1(f_2-2)^2(f_2-4)}$ (für $f_2 > 4$)

Stichprobentheorie

5. November 2022

- Grundgesamtheit • Parameter der Grundgesamtheit •
- Erhebungsmöglichkeiten • Das Literary Digest Desaster • Auswahl •
- Induktive Statistik • Stichprobenziehung • "Repräsentative" Stichprobe •
- Andere Auswahlverfahren • Zusammenfassung der Verfahren •
- Stichprobenvariable • Einfache Zufallsstichprobe • Uneingeschränkte Zufallsstichprobe • Stichprobenfunktion • Verteilung von \bar{X} • Parameter von \bar{X} • Verteilung von \bar{X} • Stichprobenanteilswert • Parameter von S^2 •
- Parameter von S^2 • Parameter von S'^2 • Übersicht Stichprobentheorie •
- Herleitung der Verteilung von \bar{X} • Herleitung der Verteilung von S^2

Grundgesamtheit

- Menge aller für eine statistische Analyse relevanten statistischen Einheiten mit übereinstimmenden Identifikationskriterien
- Untersuchung hinsichtlich (mindestens) einer festgelegten Variable

Notation

X	Variable/ Merkmal
$x_j \ j = 1, \dots, K$	Merkmalsausprägungen
K	Anzahl der Merkmalsausprägungen
$x_i \ i = 1, \dots, N$	Statistisches Element
N	Anzahl der statistischen Elemente
$f(x_j)$ bzw. $f(x_i)$	Umfang der Grundgesamtheit (GG)
$h(x_j)$ bzw. $h(x_i)$	relative Häufigkeit in der Grundgesamtheit
	absolute Häufigkeit in der Grundgesamtheit

Parameter der Grundgesamtheit

- Mittelwert

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i = \frac{1}{N} \sum_{j=1}^K x_j h(x_j) = \sum_{j=1}^K x_j f(x_j)$$

- Varianz

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 = \frac{1}{N} \sum_{j=1}^K (x_j - \mu)^2 h(x_j) = \sum_{j=1}^K (x_j - \mu)^2 f(x_j)$$

- Anteilswert (in dichotomen Grundgesamtheit)

$$\pi = \frac{1}{N} \sum_{i=1}^N x_i \quad x_i = \begin{cases} 1 \\ 0 \end{cases}$$

- Die Verteilung $f(x_j)$ und/oder Parameter der GG sind in der Regel unbekannt

Erhebungsmöglichkeiten

- Voll- oder Totalerhebung
- Teilerhebung (Stichprobe)

Beispiel 15.1 (Ausgaben für Brot)

Angenommen, wir möchten die Ausgaben für Brot in der Bundesrepublik Deutschland untersuchen:

- **Vollerhebung** – Befragung aller Bürger der Bundesrepublik zu ihren tatsächlichen Ausgaben für Brot
- **Teilerhebung** – Befragung einer Teilmenge der Grundgesamtheit bezüglich ihrer Ausgaben für das Brot und auf Basis dieser Ergebnisse eine Aussage über das Merkmal in der Grundgesamtheit

Quelle: Schira J., Statistische Methoden der VWL und BWL (2003)

- Erhebung eines Merkmals in der GG (Vollerhebung) → Zeit- und Kostenaufwand
- Versuch mit einer Teilerhebung die Aussagen über die Grundgesamtheit zu machen

Stichprobe

- Endliche Teilmenge der Elemente der Grundgesamtheit
- Auswahl und Erfassung für die statistische Untersuchung

Stichprobenumfang: n

Auswahlsatz: n/N mit $n \leq N$

Das Literary Digest Desaster

Beispiel 15.2 (Literary Digest)

- Wochenzeitschrift "The Literary Digest"
- Befragung von 10 Mio Wählern zur US-Präsidentenwahl 1936
 - ▶ Auswahl aus Telefon-, Auto- und Abonnentenregister
- 2,3 Mio. Antworten erhalten
- Wahlergebnis: http://en.wikipedia.org/wiki/United_States_presidential_election,_1936
- Vorhersage 60% Republikaner, 40% Demokraten war falsch
- Das Gallup Institut → richtige Prognose mit nur 50 Tausend Befragten
- "The Literary Digest" wurde bald darauf eingestellt

Was kann man aus Literary Digest Beispiel lernen?

- Einige Verzerrungsfehler (die nicht korrigiert werden können!)
 - ▶ Auswahlfehler: vor allem Leser mit überdurchschnittlichem Einkommen während der Weltwirtschaftskrise
 - ▶ Selektionsfehler: Motivation zu antworten z.B. 10 Mio → 2,3 Mio



Große Anzahl von Befragten garantiert keine Unverzerrtheit

Auswahl

Beispiel 15.3

- Spiegel (16.09.2009): Ungerechte Grundschullehrer: "Kevin ist kein Name, sondern eine Diagnose"
- Spiegel (24.08.2010): Grundschullehrer-Vorurteile: Kevins bekommen schlechtere Noten
- Julia Kube (2009): Vornamensforschung: Fragebogenuntersuchung bei Lehrerinnen und Lehrern, ob Vorurteile bezüglich spezifischer Vornamen von Grundschülern und davon abgeleitete erwartete spezifische Persönlichkeitsmerkmale vorliegen.
 - ▶ Onlinebefragung = zu viele junge Lehrer?
 - ▶ Unterschichtvorname = bildungsferne Schicht?
- Kirsten Becker (2010): Vornamengebundene Vorurteile von Grundschullehrerinnen und -lehrern
 - ▶ Unterschiede sind kleiner

Induktive Statistik

- Unvollständige Informationen der Stichprobe über die Verteilung der Zufallsvariablen X in der Grundgesamtheit
 - ▶ Risiko eines Fehlers
- Rückschluss von den Aussagen der Stichprobe auf die Grundgesamtheit unter Vorgabe einer gewissen Präzision
 - ▶ induktiver Schluss
- Zufallsprinzip der Stichprobenziehung
 - ▶ Wahrscheinlichkeitsrechnung
 - ▶ Wahrscheinlichkeitstheoretisch fundierte Methoden

Stichprobenziehung

- Zufallsauswahl
 - ▶ Jedes Element der Grundgesamtheit hat eine von Null verschiedene, aber nicht notwendig gleiche Wahrscheinlichkeit, in die Stichprobe zu gelangen
- Uneingeschränkte Zufallsauswahl
 - ▶ Eine Zufallsauswahl
 - ▶ Jedes Element der Grundgesamtheit hat die gleiche Wahrscheinlichkeit, in die Stichprobe zu gelangen
- Einfache Zufallsauswahl
 - ▶ Eine Zufallsauswahl
 - ▶ Erfüllt die Bedingungen der uneingeschränkten Zufallsauswahl
 - ▶ Unabhängigkeit der Ziehungen der Elemente der Stichprobe

- mit Zurücklegen der gezogenen Elemente:
 - ▶ Garantierte Unabhängigkeit der Ziehungen und konstante Verteilung der Zufallsvariablen in der Grundgesamtheit
 - ▶ einfache Zufallsstichprobe
- ohne Zurücklegen der gezogenen Elemente:
 - ▶ Abhängigkeit der Ziehungen, Veränderung der Verteilung der Zufallsvariablen in der Grundgesamtheit
 - ▶ uneingeschränkte Zufallsstichprobe

“Repräsentative” Stichprobe

Beispiel 15.4

- Grundgesamtheit: 50% Frauen und 50% Männer
- Stichprobe: $n=100$

“Repräsentative” Stichprobe

- z.B. Sicherstellung von 50 Frauen und 50 Männern in der Stichprobe

Zufallsstichprobe

- ungleiche Anzahl der Frauen und Männer in der Stichprobe möglich
- Anzahl Männer/Frauen in der Stichprobe $\sim B(100; 0,5)$
 - ▶ $P(M = 50) \approx 7,96\%$
 - ▶ $P(45 \leq M \leq 55) = 72,87\%$

- "Repräsentative" Stichprobe \neq Zufallsstichprobe
- ⇒ Induktive Statistik nicht anwendbar
- Da Verteilung der untersuchten Variablen in GG nicht bekannt ist, weiß man nicht, ob die erhobenen Merkmale "repräsentativ" bezüglich ihrer Verteilung und Relevanz sind.
- In der Praxis wird meistens auch nur für ein Teil der erhobenen Merkmale Repräsentativität eingefordert
 - ▶ z. B. Alter, Geschlecht, Studiengang



Der Begriff der "repräsentativen" Stichprobe ist nicht eindeutig definiert.

Andere Auswahlverfahren

Quotenstichprobe

- repräsentative Zusammensetzung der Stichprobe
- vorgebene Quote an bestimmten Merkmalen (aus GG bekannt)
 - ▶ z.B. aus Zensusdaten: Alter, Geschlecht und Wohnortgröße
- Problem
 - ▶ Erfüllbarkeit der Quoten
- Keine Zufallsstichprobe!

Geschichtete Zufallsstichprobe

- Unterteilung der Grundgesamtheit in mehrere kleinere Gruppierungen (Schichten), z.b. Geschlecht(m/w)
- Separate Ziehung einer einfachen Zufallsstichprobe aus jeder Gruppierung
- Bei günstiger Auswahl genauere Ergebnisse

Klumpen-Stichprobe

- Zerlegung der Grundgesamtheit in viele kleine, oft geografisch abgegrenzte Teilgesamtheiten (Klumpen)
- Zufällige Auswahl von einem Teil der Klumpen für die Stichprobe
- Vollständige Erfassung der ausgewählten Klumpen
- Erhöhter Stichprobenfehler (Klumpeneffekt)

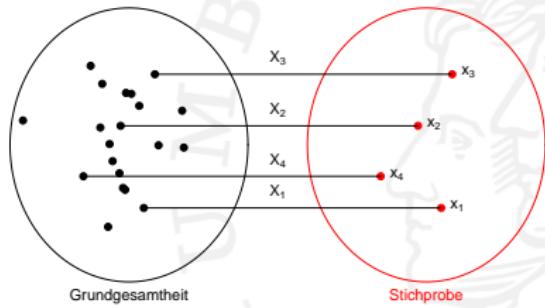
Judgement-Stichprobe

- Auswahl von Elementen abhängig von der Entscheidung der Experten
- Auswahl von typischen Fällen
- In den Marketingumfragen sehr verbreitet
- Keine Zufallsstichprobe

Zusammenfassung der Verfahren

Methode	benötigter Informationsgrad über Grundgesamtheit	Statistik	
		deskriptive	induktive
Einfache Zufallsstichprobe	Gering	Ja	Ja
Uneing. Zufallsstichprobe	Gering	Ja	Ja
Geschichtete Zufallsstichprobe	Mittelmäßig	Ja	Ja
Klumpen-Stichprobe	Mittelmäßig	Ja	Möglich
Judgement-Stichprobe	Hoch	Ja	Nein
“Repräsentative” Stichprobe	Sehr hoch	Ja	Nein

Stichprobenvariable



- Ziehen einer Stichprobe vom Umfang n
- n -malige Durchführung eines Zufallsexperiments
- Hier: $n = 4$
- X_i sind Stichprobenvariablen
- x_i Realisationen der Stichprobenvariablen

Vor der Ziehung

- n StichprobenvARIABLEN
 $X_1, \dots, X_i, \dots, X_n$
- Zufallsvariable X_i entspricht potentieller Realisation der Zufallsvariablen X der Grundgesamtheit bei i -ten Durchführung des Zufallsexperiments

Nach der Ziehung

- StichprobenWERTE
 $x_1, \dots, x_i, \dots, x_n$
- x_i nehmen konkrete Werte an
- Realisationen der Zufallsvariablen
 $X_1, \dots, X_i, \dots, X_n$

Einfache Zufallsstichprobe

Beispiel 15.5 (Urnenmodell 1)

- Grundgesamtheit: Urne mit $N = 7$ Kugeln
- Zufallsvariable X : "Zahl auf der Kugel"
- Wahrscheinlichkeitsverteilung

Kugel	x	$h(x)$	$f(x) = h(x)/N$	$F(x)$
A	10	1	1/7	1/7
B,C	11	2	2/7	3/7
D,E,F	12	3	3/7	6/7
G	16	1	1/7	7/7

- Mittelwert, Varianz, Standardabweichung:
 $\mu = 12$, $\sigma^2 = 22/7 = 3,143$, $\sigma = 1,773$

- Umfang $n = 2$
- Zufallsauswahl mit Zurücklegen
- Stichprobenvariablen:
 X_1 = "Zahl der ersten gezogenen Kugel"
 X_2 = "Zahl der zweiten gezogenen Kugel"
- Anzahl der möglichen Stichproben:
 $V^W(7; 2) = 7^2 = 49$
- Wahrscheinlichkeit, eine dieser Stichproben zu erhalten: $1/49$

Mögliche Realisationen: Urnenmodell 1

1.Kugel	2.Kugel						
	10	11	11	12	12	12	16
10	10;10	10;11	10;11	10;12	10;12	10;12	10;16
11	11;10	11;11	11;11	11;12	11;12	11;12	11;16
11	11;10	11;11	11;11	11;12	11;12	11;12	11;16
12	12;10	12;11	12;11	12;12	12;12	12;12	12;16
12	12;10	12;11	12;11	12;12	12;12	12;12	12;16
12	12;10	12;11	12;11	12;12	12;12	12;12	12;16
16	16;10	16;11	16;11	16;12	16;12	16;12	16;16

Wahrscheinlichkeitsfunktionen für X_1 und X_2

x_1	$h(x_1)$	$f(x_1)$
10	7	$7/49 = 1/7$
11	14	$14/49 = 2/7$
12	21	$21/49 = 3/7$
16	7	$7/49 = 1/7$

x_2	$h(x_2)$	$f(x_2)$
10	7	$7/49 = 1/7$
11	14	$14/49 = 2/7$
12	21	$21/49 = 3/7$
16	7	$7/49 = 1/7$

- X_1 und X_2 sind identisch verteilt
- gleiche Verteilung wie die Variable X in der GG

Zweidimensionale Verteilung $f(x_1, x_2)$

		X_2				
		10	11	12	16	$f(x_1)$
X_1	10	$1/49$	$2/49$	$3/49$	$1/49$	$1/7$
	11	$2/49$	$4/49$	$6/49$	$2/49$	$2/7$
X_1	12	$3/49$	$6/49$	$9/49$	$3/49$	$3/7$
	16	$1/49$	$2/49$	$3/49$	$1/49$	$1/7$
$f(x_2)$		$1/7$	$2/7$	$3/7$	$1/7$	1

Für alle Wertepaare (x_1, x_2) gilt: $f(x_1, x_2) = f(x_1) \cdot f(x_2)$.

→ Die Stichprobenvariablen X_1 und X_2 sind unabhängig.

- Stichprobenvariablen X_1 und X_2 :

- ▶ identisch verteilt
- ▶ gleiche Verteilung wie die Variable X in der Grundgesamtheit
- ▶ unabhängig voneinander

Uneingeschränkte Zufallsstichprobe

Beispiel 15.6 (Urnenmodell 2)

- Umfang $n = 2$
- Zufallsauswahl ohne Zurücklegen
- StichprobenvARIABLEN:
 X_1 = "Zahl der ersten gezogenen Kugel"
 X_2 = "Zahl der zweiten gezogenen Kugel"
- Anzahl der möglichen Stichproben:
 $V(7; 2) = 7!/(7 - 2)! = 42$
- Wahrscheinlichkeit, eine dieser Stichproben zu erhalten: $1/42$

Mögliche Realisationen: Urnenmodell 2

1.Kugel	2.Kugel						
	10	11	11	12	12	12	16
10		10;11	10;11	10;12	10;12	10;12	10;16
11	11;10		11;11	11;12	11;12	11;12	11;16
11	11;10	11;11		11;12	11;12	11;12	11;16
12	12;10	12;11	12;11		12;12	12;12	12;16
12	12;10	12;11	12;11	12;12		12;12	12;16
12	12;10	12;11	12;11	12;12	12;12		12;16
16	16;10	16;11	16;11	16;12	16;12	16;12	

Wahrscheinlichkeitsfunktionen für X_1 und X_2

x_1	$h(x_1)$	$f(x_1)$
10	6	$6/42 = 1/7$
11	12	$12/42 = 2/7$
12	18	$18/42 = 3/7$
16	6	$6/42 = 1/7$

x_2	$h(x_2)$	$f(x_2)$
10	6	$6/42 = 1/7$
11	12	$12/42 = 2/7$
12	18	$18/42 = 3/7$
16	6	$6/42 = 1/7$

- X_1 und X_2 sind identisch verteilt
- gleiche Verteilung wie die Variable X in der GG.

Zweidimensionale Verteilung $f(x_1, x_2)$

		X_2					
		X_1	10	11	12	16	$f(x_1)$
X_1	10	0	$2/42$	$3/42$	$1/42$	$1/7$	
	11	$2/42$	$2/42$	$6/42$	$2/42$	$2/7$	
	12	$3/42$	$6/42$	$6/42$	$3/42$	$3/7$	
	16	$1/42$	$2/42$	$3/42$	0	$1/7$	
		$f(x_2)$	$1/7$	$2/7$	$3/7$	$1/7$	1

Es ist: $f(x_1, x_2) \neq f(x_1) \cdot f(x_2)$.

→ Die Stichprobenvariablen X_1 und X_2 sind nicht unabhängig.

- Stichprobenvariablen X_1 und X_2 :

- ▶ identisch verteilt
- ▶ gleiche Verteilung wie die Variable X in der GG
- ▶ nicht unabhängig

Stichprobenfunktion

- Funktion $U = U(X_1, \dots, X_n)$ der Stichprobenvariablen X_1, \dots, X_n
- Funktion von Zufallsvariablen $\rightarrow U$ ist eine Zufallsvariable
- Verteilung in der GG \rightarrow Verteilung der Stichprobenvariablen

Erwartungswert

$$E(U) = \mu_U$$

Varianz

$$\text{Var}(U) = \sigma_U^2 = \sigma^2(U)$$

Standardabweichung

$$\sigma_U = \sigma(U)$$

Realisation der Stichprobenfunktion

$$u = U(x_1, \dots, x_n)$$

Berechnung der Verteilung von U

- Analyse aller möglichen Stichproben
- Verteilungsannahme an Stichprobenvariablen machen und Rechenregeln für Verteilungen nutzen
 - ▶ Reproduktivitätseigenschaften
 - ▶ Zentraler Grenzwertsatz

Stichprobenfunktionen dienen zur Gewinnung von Informationen über unbekannte Parameter der Grundgesamtheit:

- Stichprobenmittelwert

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

- Stichprobenanteilswert

$$\hat{\pi} = \frac{1}{n} \sum_{i=1}^n X_i$$

Da die Stichprobenvariablen X_i Zufallsvariablen sind \rightarrow
Stichprobenfunktionen $\bar{X}, \hat{\pi}$ sind auch Zufallsvariablen.

- Stichprobenvarianz falls der Erwartungswert $E(X) = \mu$ der Grundgesamtheit bekannt ist:

$$S^{*2} = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$$

- Stichprobenvarianz falls der Erwartungswert $E(X) = \mu$ der Grundgesamtheit unbekannt ist:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

$$S'^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

- Da die Stichprobenvariablen X_i und \bar{X} Zufallsvariablen sind \rightarrow Stichprobenfunktionen S^2, S'^2, S^{*2} sind auch Zufallsvariablen.

Verteilung von \bar{X}

Vor der Ziehung der Stichprobe

- \bar{X} als Stichprobenfunktion → eine Zufallsvariable
- Stichprobenvariablen: X_i ($i = 1, \dots, n$) mit $E(X_i) = \mu, \text{Var}(X_i) = \sigma^2$
- Das arithmetische Mittel der Stichprobe → Funktion der Stichprobenvariablen X_1, \dots, X_n :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

Nach der Ziehung der Stichprobe

- \bar{x} ist eine reelle Zahl
- Realisation der Stichprobenfunktion \bar{X}
- Konkrete Stichprobenwerte x_1, \dots, x_n
- Stichprobenmittelwert \bar{x}

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Beispiel 15.7 (Urnenmodell 1)

- Urne mit $N = 7$ Kugeln
- Zufallsvariable X : "Zahl auf der Kugel"
- Stichprobenvariablen:
 - ▶ X_1 = "Zahl der ersten gezogenen Kugel"
 - ▶ X_2 = "Zahl der zweiten gezogenen Kugel"

Mögliche Realisationen

1. Kugel	2. Kugel						
	10	11	11	12	12	12	16
10	10;10	10;11	10;11	10;12	10;12	10;12	10;16
11	11;10	11;11	11;11	11;12	11;12	11;12	11;16
11	11;10	11;11	11;11	11;12	11;12	11;12	11;16
12	12;10	12;11	12;11	12;12	12;12	12;12	12;16
12	12;10	12;11	12;11	12;12	12;12	12;12	12;16
12	12;10	12;11	12;11	12;12	12;12	12;12	12;16
16	16;10	16;11	16;11	16;12	16;12	16;12	16;16

Stichprobenmittelwerte

1. Kugel	2. Kugel						
	10	11	11	12	12	12	16
10	10,0	10,5	10,5	11	11	11	13
11	10,5	11	11	11,5	11,5	11,5	13,5
11	10,5	11	11	11,5	11,5	11,5	13,5
12	11	11,5	11,5	12	12	12	14
12	11	11,5	11,5	12	12	12	14
12	11	11,5	11,5	12	12	12	14
16	13	13,5	13,5	14	14	14	16

Verteilung von \bar{X} bei den Stichproben

	mit Zurücklegen	ohne Zurücklegen
\bar{x}	$P(\bar{X} = \bar{x})$	$P(\bar{X} = \bar{x})$
10	1/49	
10,5	4/49	10,5
11	10/49	11
11,5	12/49	11,5
12	9/49	12
13	2/49	13
13,5	4/49	13,5
14	6/49	14
16	1/49	

$$E(\bar{X}) = 12$$

$$E(\bar{X}) = 12$$

$$\text{Var}(\bar{X}) = 22/14 \approx 1,57 > \text{Var}(\bar{X}) = 55/42 \approx 1,31$$

Parameter von \bar{X}

Einfache Zufallsstichprobe (Ziehen mit Zurücklegen)

$$E(\bar{X}) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} \cdot n \cdot \mu = \mu$$

$$\begin{aligned} Var(\bar{X}) &= Var\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \stackrel{X_i \text{ unabh.}}{=} \frac{1}{n^2} \sum_{i=1}^n Var(X_i) \\ &= \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{1}{n^2} n \cdot \sigma^2 = \frac{\sigma^2}{n} \\ \sigma(\bar{X}) &= \frac{\sigma}{\sqrt{n}} \end{aligned}$$

- wenn σ^2 der Grundgesamtheit bekannt: $\hat{\sigma}^2(\bar{X}) = \frac{\sigma^2}{n}$
- wenn σ^2 der Grundgesamtheit unbekannt: $\hat{\sigma}^2(\bar{X}) = \frac{s^2}{n}$

Uneingeschränkte Zufallsstichprobe (Ziehen ohne Zurücklegen)

$$E(\bar{X}) = \mu$$

$$\sigma^2(\bar{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

$$\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

- wenn σ^2 der Grundgesamtheit bekannt: $\hat{\sigma}^2(\bar{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$
- wenn σ^2 der Grundgesamtheit unbekannt: $\hat{\sigma}^2(\bar{X}) = \frac{s^2}{n} \cdot \frac{N-n}{N-1}$

Beispiel 15.8 (Urnenmodell 1 und 2)

Grundgesamtheit

$$\mu = 12 \quad \sigma^2 = \frac{22}{7} = 3,143$$

Verteilung von \bar{X}

- mit Zurücklegen:

$$E(\bar{X}) = 12 = \mu$$

$$\text{Var}(\bar{X}) = \sigma^2(\bar{X}) = 22/14 = (22/7)/2 = \frac{\sigma^2}{n}$$

- ohne Zurücklegen:

$$E(\bar{X}) = 12 = \mu$$

$$\text{Var}(\bar{X}) = \sigma^2(\bar{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1} = \frac{22}{7} \cdot \frac{1}{2} \cdot \frac{7-2}{7-1} = \frac{55}{42}$$

Verteilung von \bar{X}

Fall 1

Falls,

- eine einfache Zufallsstichprobe vorliegt
- das Merkmal X der Grundgesamtheit $N(\mu, \sigma)$ -verteilt
- σ^2 der Grundgesamtheit bekannt

so gilt:

- ▷ die Stichprobenfunktion \bar{X} ist im wesentlichen eine Summe der normalverteilten Zufallsvariablen $\Rightarrow \bar{X} \sim N(\mu, \sigma(\bar{X}))$ und
- ▷ die standardisierte Zufallsvariable

$$Z = \frac{\bar{X} - \mu}{\sigma(\bar{X})} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0; 1)$$

Fall 2

Falls,

- eine einfache Zufallsstichprobe vorliegt
- das Merkmal X der Grundgesamtheit $N(\mu, \sigma)$ -verteilt
- σ^2 der Grundgesamtheit unbekannt

so sind (Beweise siehe Anhang):

- ▷ die Stichprobenfunktion \bar{X} ist im wesentlichen eine Summe der normalverteilten Zufallsvariablen $\Rightarrow \bar{X} \sim N(\mu, \sigma(\bar{X}))$
- ▷ Da σ^2 der Grundgesamtheit nicht bekannt ist, wird S^2 benutzt
- ▷ die standardisierte Zufallsvariable

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

ist t -verteilt mit $f = n - 1$ Freiheitsgraden.

- ▷ Wenn der Stichprobenumfang $n > 30$ ist, dann $T \approx N(0, 1)$.

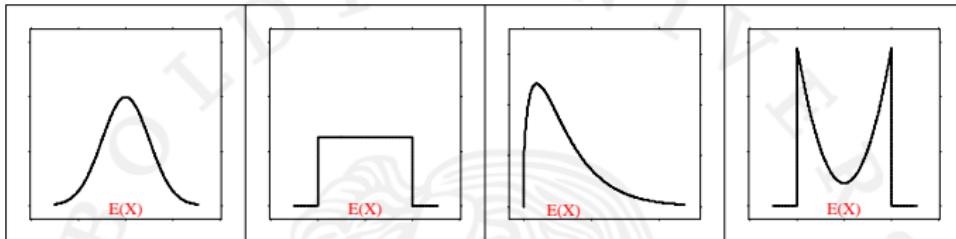
Fall 3

Falls,

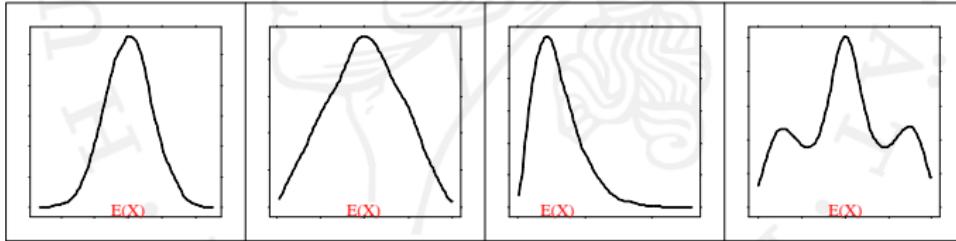
- Stichprobenvariablen X_i ($i = 1, \dots, n$) identisch und beliebig verteilt,
- mit $E(X_i) = \mu$ und $\text{Var}(X_i)) = \sigma^2$,
- $n > 30$

so sind:

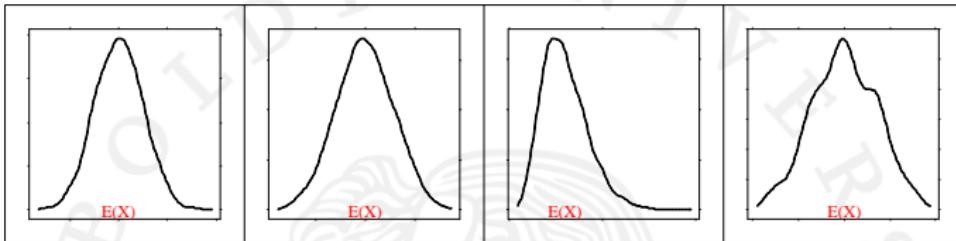
- ▷ Zentraler Grenzwertsatz $\rightarrow \sum X_i$ ist approximativ normalverteilt
- ▷ $\bar{X} \approx N(\mu; \sigma(\bar{X}))$
 - ▶ wenn σ^2 bekannt
 - ★ die standardisierte Zufallsvariable Z approximativ $N(0; 1)$ -verteilt
 - ▶ wenn σ^2 unbekannt
 - ★ die standardisierte Zufallsvariable T approximativ $N(0; 1)$ -verteilt



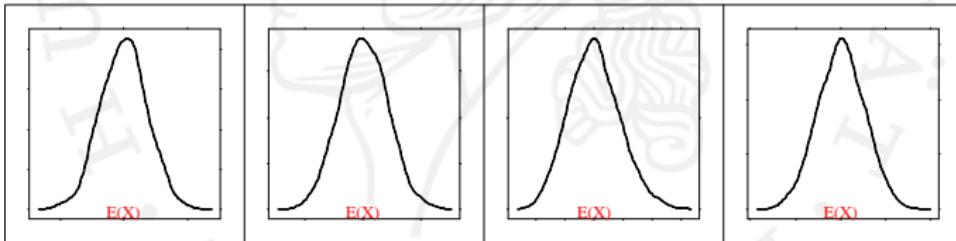
Verteilung der Zufallsvariablen X , $n=1$



Verteilung des Mittelwertes \bar{X} , $n=2$



Verteilung des Mittelwertes \bar{X} , $n=4$



Verteilung des Mittelwertes \bar{X} , $n=30$

Stichprobenanteilswert

Ziehen mit Zurücklegen

- dichotome Grundgesamtheit:

- ▶ A und \bar{A} mit $P(A) = \pi$ und $P(\bar{A}) = 1 - \pi$ konstant
- ▶ N - endlicher Umfang der Grundgesamtheit
- ▶ M - Anzahl der Elemente mit Eigenschaft A
- ▶ $\pi = M/N$ Anteilswert der Grundgesamtheit
- ▶ n - Stichprobenumfang

- Stichprobenvariablen X_i mit

$$X_i = \begin{cases} 1 \\ 0 \end{cases}$$

- Stichprobenfunktion X : "Anzahl der Elemente mit der Eigenschaft A in der Stichprobe"

$$X = \sum_{i=1}^n X_i$$

Ziehen mit Zurücklegen

- Realisierung mehrerer Bernoulli-Versuche
- Verteilung von X : $X \sim B(n; \pi)$

$$f_B(x|n; \pi) = \begin{cases} \binom{n}{x} \pi^x (1 - \pi)^{n-x}; & x = 0, 1, \dots, n \\ 0 & \text{sonst} \end{cases}$$

- Parameter von X :

$$E(X) = n\pi$$

$$\text{Var}(X) = n\pi(1 - \pi)$$

Parameter von $\hat{\pi}$ beim Ziehen mit Zurücklegen

- Stichprobenanteilswert: $\hat{\pi} = X/n$ ist eine Stichprobenfunktion

$$E(\hat{\pi}) = \pi$$

$$\begin{aligned}Var(\hat{\pi}) &= Var\left(\frac{X}{n}\right) = \frac{1}{n^2} Var(X) \\&= \frac{1}{n^2} n\pi(1 - \pi) = \frac{\pi(1 - \pi)}{n}\end{aligned}$$

$$\sigma(\hat{\pi}) = \sqrt{\frac{\pi(1 - \pi)}{n}}$$

Ziehen ohne Zurücklegen

- dichotome Grundgesamtheit:
 - ▶ A und \bar{A}
 - ▶ $P(A)$ und $P(\bar{A})$ nicht konstant
 - ▶ N - endlicher Umfang der Grundgesamtheit
 - ▶ M - Anzahl der Elemente mit Eigenschaft A
 - ▶ $\pi = M/N$ Anteilswert der Grundgesamtheit
 - ▶ n - Stichprobenumfang
- Stichprobenvariablen X_i mit

$$X_i = \begin{cases} 1 \\ 0 \end{cases}$$

- Stichprobenfunktion X : "Anzahl der Elemente mit der Eigenschaft A in der Stichprobe"

$$X = \sum_{i=1}^n X_i$$

Verteilung von X und $\hat{\pi}$

- Stichprobenanteilswert: $\hat{\pi} = X/n$ ist eine Stichprobenfunktion
- $X \sim H(N; M; n)$

$$P(X = x) = f_H(x|N; n; M) = \begin{cases} \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}}; & x = 0, 1, \dots, n \\ 0 & \text{sonst} \end{cases}$$

- $\hat{\pi}$ ist nicht hypergeometrisch verteilt: $P(\hat{\pi} = p) = P(X/n = p) = P(X = np)$

$$P(\hat{\pi} = p) = f_H(np|N; n; M) = \frac{\binom{M}{np} \binom{N-M}{n-np}}{\binom{N}{n}}$$

Parameter von X

$$E(X) = n\pi$$

$$Var(X) = n\pi(1 - \pi) \frac{N - n}{N - 1}$$

Parameter von $\hat{\pi}$

$$E(\hat{\pi}) = \frac{1}{n} E(X) = \frac{1}{n} n\pi = \pi$$

$$Var(\hat{\pi}) = \frac{1}{n^2} Var(X) = \frac{\pi(1 - \pi)}{n} \frac{N - n}{N - 1}$$

μ der Grundgesamtheit bekannt

$$S^{*2} = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 \quad \frac{nS^{*2}}{\sigma^2} \sim \chi_n^2$$

μ der Grundgesamtheit unbekannt

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

$$S'^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \quad \frac{nS'^2}{\sigma^2} \sim \chi_{n-1}^2$$

Herleitungen für die Chi-Quadrat-Verteilungen siehe Anhang.

Parameter von S^{*2}

Erwartungswert von S^{*2}

$$E\left(\frac{nS^{*2}}{\sigma^2}\right) = E(\chi_n^2) = n$$

$$E(S^{*2}) \cdot \frac{n}{\sigma^2} = n$$

$$E(S^{*2}) = \sigma^2$$

Varianz von S^{*2}

$$\text{Var}\left(\frac{nS^{*2}}{\sigma^2}\right) = \text{Var}(\chi_n^2) = 2n$$

$$\text{Var}(S^{*2}) \frac{n^2}{\sigma^4} = 2n$$

$$\text{Var}(S^{*2}) = \frac{2\sigma^4}{n}$$

Parameter von S^2

Erwartungswert von S^2

$$E\left(\frac{(n-1)S^2}{\sigma^2}\right) = E(\chi_{n-1}^2) = n-1$$

$$E(S^2)\frac{n-1}{\sigma^2} = n-1$$

$$E(S^2) = \sigma^2$$

Varianz von S^2

$$\text{Var}\left(\frac{(n-1)S^2}{\sigma^2}\right) = \text{Var}(\chi_{n-1}^2) = 2(n-1)$$

$$\text{Var}(S^2)\frac{(n-1)^2}{\sigma^4} = 2(n-1)$$

$$\text{Var}(S^2) = \frac{2\sigma^4}{n-1}$$

Parameter von S'^2

Erwartungswert von S'^2

$$E\left(\frac{nS'^2}{\sigma^2}\right) = E(\chi_{n-1}^2) = n - 1$$

$$E(S'^2) \cdot \frac{n}{\sigma^2} = n - 1$$

$$E(S'^2) = \frac{n-1}{n} \sigma^2$$

Varianz von S'^2

$$\text{Var}\left(\frac{nS'^2}{\sigma^2}\right) = \text{Var}(\chi_{n-1}^2) = 2(n - 1)$$

$$\text{Var}(S'^2) \cdot \frac{n^2}{\sigma^4} = 2(n - 1)$$

$$\text{Var}(S'^2) = \frac{2\sigma^4(n - 1)}{n^2}$$

Übersicht Stichprobentheorie

- Die wahren Parameter, z.B. μ , σ^2 , π , der Grundgesamtheit sind fixe Größen (reelle Zahlen).
- Jede Größe, die sich von Stichprobenziehung zu Stichprobenziehung ändert, wird mit einer Zufallsvariablen beschrieben.
- Um für die Zufallsvariablen den Verteilungstyp zu bestimmen bzw. die Verteilungsparameter berechnen zu können ist eine einfache (oder uneingeschränkte) Zufallsstichprobe notwendig.
- Die Wahrscheinlichkeitsrechnung ermöglicht uns die Abweichung zwischen Parameter berechnet aus der Stichprobe und wahren Parameter zu quantifizieren.
- In der Statistik II werden nur Unsicherheiten, die sich aus der Stichprobenziehung ergeben betrachtet.
 - ▶ Es gibt andere Arten der Unsicherheit, z.B. in der Praxis die Meßungenauigkeit der Stichprobenwerte.

Stichprobenfunktion U	Bedingung	$E(U)$	$Var(U)$	Verteilung
$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$	einf. ZA, σ bek.	μ	σ^2/n	$\bar{X} \sim N(\cdot, \cdot)$
	uneing. ZA, σ bek.	μ	$\sigma^2/n \cdot \frac{n-n}{N-1}$	
$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$	$X \sim N(\mu; \sigma)$, σ bek.	μ	σ^2/n	$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0; 1)$
	$n > 30$, σ bek.	μ	σ^2/n	$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \approx N(0; 1)$
	$X \sim N(\mu; \sigma)$, σ unbek.	μ	s^2/n	$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$
	$n > 30$, σ unbek.	μ	s^2/n	$\frac{\bar{X}-\mu}{s/\sqrt{n}} \approx N(0; 1)$

μ, σ sind die unbekannten Parameter der Grundgesamtheit

Stichprobenfunktion U	Bedingung	$E(U)$	$Var(U)$	Verteilung
$\hat{\pi} = \frac{1}{n} \sum_{i=1}^n X_i$	einf. ZA	π	$\frac{\pi(1-\pi)}{n}$	$n\hat{\pi} \sim B(n, \pi)$
$\pi^* = \frac{M}{N}$	uneing. ZA	π^*	$\frac{\pi^*(1-\pi^*)}{n} \frac{N-n}{N-1}$	$n\hat{\pi} \sim H(n, N, M)$
$S^{*2} = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$	(μ bek.)	σ^2	$\frac{2\sigma^4}{n}$	$\frac{nS^{*2}}{\sigma^2} \sim \chi_n^2$
$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$		σ^2	$\frac{2\sigma^4}{n-1}$	$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$
$S'^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$		$\frac{n-1}{n} \sigma^2$	$\frac{2\sigma^4(n-1)}{n^2}$	$\frac{nS'^2}{\sigma^2} \sim \chi_{n-1}^2$

μ, σ und π sind die unbekannten Parameter der Grundgesamtheit

Herleitung der Verteilung von \bar{X}

Standardisierte Zufallsvariable

$$T = \frac{\bar{X} - \mu}{S} \sqrt{n} \approx N(0, 1) \quad \text{für} \quad n > 30$$

Da σ^2 der Grundgesamtheit nicht bekannt, wird S^2 herangezogen

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

$$\frac{S^2}{\sigma^2} = \frac{1}{\sigma^2} \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

$$\frac{n - 1}{\sigma^2} S^2 = \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2$$

Summe von quadrierten unabhängigen standardnormalverteilten
Zufallsvariablen folgt einer χ^2 -Verteilung mit dem Parameter $f = n - 1$.

$$T = \frac{\frac{\bar{X} - \mu}{\sigma} \sqrt{n}}{\sqrt{\frac{1}{n-1} \left(\frac{n-1}{\sigma^2} S^2 \right)}} = \frac{\bar{X} - \mu}{S} \sqrt{n}$$

$$T = \frac{Z}{\sqrt{\frac{Y}{f}}} \approx t - \text{verteilt}$$

Herleitung der Verteilung von S^{*2}

- Einfache Zufallsstichprobe vom Umfang n
- Stichprobenvariablen: $X_1, \dots, X_n \quad X_i, i = 1, \dots, n$
- $E(X_i) = \mu \quad \text{Var}(X_i) = \sigma^2$
- $X_i \sim N(\mu; \sigma^2) \Rightarrow Z_i = (X_i - \mu)/\sigma \sim N(0; 1)$ Z_i unabhängig
- \bar{X} normalverteilt mit $E(\bar{X}) = \mu$ und $\text{Var}(\bar{X}) = \sigma^2(\bar{X}) = \sigma^2/n$
 $\bar{X} \sim N(\mu, \sigma^2(\bar{X}))$

$$S^{*2} = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$$

$$\frac{nS^{*2}}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2$$

$$\frac{nS^{*2}}{\sigma^2} = \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2$$

$$\frac{nS^{*2}}{\sigma^2} = \sum_{i=1}^n Z_i^2$$

nS^{*2}/σ^2 beinhaltet die Summe von quadrierten unabhängigen standardnormalverteilten Zufallsvariablen

$$\frac{nS^{*2}}{\sigma^2} \sim \chi^2(f = n)$$

Anzahl der Freiheitsgrade $f =$ Anzahl der unabhängigen Summanden = Anzahl der unabhängigen standardisierten Zufallsvariablen Z_i

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2$$

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2$$

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^n Z_i^2$$

$(n-1)S^2/\sigma^2$ beinhaltet die Summe von unabhängigen quadrierten standardnormalverteilten Zufallsvariablen

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(f = n-1)$$

Anzahl der Freiheitsgrade: $f = \text{Anzahl der unabhängigen Summanden}$

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

Bestimmung der Freiheitsgrade

Aus Nulleigenschaft des arithmetischen Mittels $\sum_{i=1}^n (X_i - \bar{X}) = 0$ folgt:

- Zufallsvariablen $X_i - \bar{X}$ insgesamt nicht unabhängig
- nur $n - 1$ dieser Zufallsvariablen sind unabhängig
- Zufallsvariablen $X_i - \bar{X}$ in S^2 enthalten

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

- $f = n - 1$

Herleitung der Verteilung von S'^2

$$S'^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

$$\frac{nS'^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2$$

$$\frac{nS'^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2$$

$$\frac{nS'^2}{\sigma^2} = \sum_{i=1}^n Z_i^2$$

nS'^2/σ^2 beinhaltet die Summe von unabhängigen quadrierten standardnormalverteilten Zufallsvariablen

$$\chi'^2 = \frac{nS'^2}{\sigma^2} \sim \chi^2(f = n - 1)$$

Anzahl der Freiheitsgrade: $f = n - 1$

Statistische Schätzverfahren

5. November 2022

- Grundbegriffe • Notation • Mittlere quadratische Abweichung •
- Erwartungstreue • Interpretation von Erwartungstreue • Effizienz •
- Konsistenz • Maximum-Likelihood-Methode • Likelihood-Funktion •
- Vorgehensweise ML-Schätzung • ML-Schätzer für λ in $PO(\lambda)$ •
- ML-Schätzer für π in $B(n, \pi)$ • ML-Schätzer für λ in $Exp(\lambda)$ •
- ML-Schätzer für μ & σ^2 in $N(\mu; \sigma^2)$ • ML-Schätzer für μ • ML-Schätzer für σ^2 • Grundgesamtheit und ML-Schätzung • Methode der kleinsten Quadrate • Vorgehensweise KQ-Schätzung • KQ-Schätzer für μ •
- ML-Schätzung vs. KQ-Schätzung • Konstruktion von Schätzfunktionen

Grundbegriffe

- Grundgesamtheit
 - ▶ Zufallsvariable X mit Verteilung $F(x)$ und zugehörigen Parametern
 - ▶ Parameter der Grundgesamtheit: $\vartheta \rightarrow$ unbekannt
- einfache Zufallsstichprobe vom Umfang n
 - ▶ Stichprobenvariablen: X_1, \dots, X_n
- Schätzung
 - ▶ Approximation der unbekannten Parameter der Grundgesamtheit auf der Basis von Stichproben
- Schätzverfahren
 - ▶ Vorschrift zur Schätzung

- Schätzfunktion (Schätzer)

- ▶ Eine Stichprobenfunktion $\hat{\theta}_n = g(X_1, \dots, X_n)$, die aufgrund ihrer Eigenschaften zur Schätzung eines Parameters der Grundgesamtheit geeignet ist
- ▶ Vor der Ziehung der Zufallsstichprobe ist die Schätzfunktion $\hat{\theta}_n$ eine Zufallsvariable
- ▶ Nach der Ziehung der Zufallsstichprobe nimmt die Schätzfunktion $\hat{\theta}_n$ einen bestimmten Schätzwert an

- Schätzwert

- ▶ Realisation der Schätzfunktion aufgrund einer konkreten Zufallsstichprobe mit den Stichprobenwerten x_1, \dots, x_n

$$\hat{\vartheta}_n = g(x_1, \dots, x_n)$$

Notation

Verwendung	Symbol	Beispiel
Zufallsvariable	Latein, groß	X, Y
Stichprobenvariable	Latein, groß, mit Index	X_i, Y_i
Beobachtung	Latein, klein, mit Index	x_i, y_i
Allg. Parameter der Grundgesamtheit	Griechisch, klein	ϑ
Spez. Parameter der Grundgesamtheit	Griechisch, klein	μ, σ^2
Allg. Schätzfunktion	Griechisch, klein, mit Dach	$\hat{\theta}, \hat{\theta}_n$
Spez. Schätzfunktion	aus historischen Gründen nicht einheitlich	$\hat{\Pi}, \bar{X}, S^2$

Mittlere quadratische Abweichung

- generelles Maß zur Beurteilung der Qualität einer Schätzfunktion (Mean Square Error)

$$MSE = E[(\hat{\theta}_n - \vartheta)^2] \quad (1)$$

- mittlerer quadratischer Abstand zwischen der Schätzfunktion $\hat{\theta}_n$ und dem wahren Parameter ϑ in der Grundgesamtheit

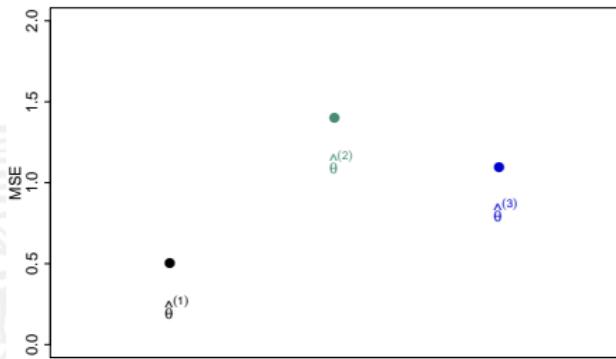
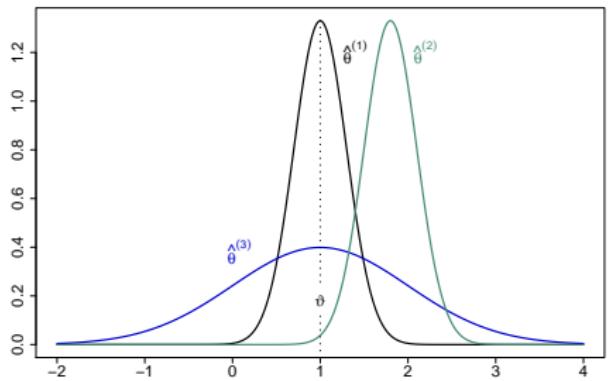
$$\text{Var}(\hat{\theta}_n) = E[\{\hat{\theta}_n - E(\hat{\theta}_n)\}^2] = E[(\hat{\theta}_n - \vartheta)^2] - [E(\hat{\theta}_n) - \vartheta]^2$$

Beweis dieser Behauptung siehe Anhang

(2)

$$E[(\hat{\theta}_n - \vartheta)^2] = E[\{\hat{\theta}_n - E(\hat{\theta}_n)\}^2] + [E(\hat{\theta}_n) - \vartheta]^2 \quad (3)$$

$$\text{MSE} = \text{Var}(\hat{\theta}_n) + \text{Verzerrung}^2 \quad (4)$$



1. $E(\hat{\theta}^{(1)}) = \vartheta$
2. $E(\hat{\theta}^{(2)}) \neq \vartheta, \text{Var}(\hat{\theta}^{(2)}) = \text{Var}(\hat{\theta}^{(1)}) \Rightarrow \text{MSE}(\hat{\theta}^{(2)}) > \text{MSE}(\hat{\theta}^{(1)})$
3. $E(\hat{\theta}^{(3)}) = \vartheta, \text{Var}(\hat{\theta}^{(3)}) > \text{Var}(\hat{\theta}^{(1)}) \Rightarrow \text{MSE}(\hat{\theta}^{(3)}) > \text{MSE}(\hat{\theta}^{(1)})$

$\hat{\theta}_n$	Bedingung	$\text{Var}(\hat{\theta}_n)$	Verzerr.	MSE
$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$	σ bek.	$\frac{\sigma^2}{n}$	0	$\frac{\sigma^2}{n}$
	σ unbek.	$\frac{s^2}{n}$	0	$\frac{s^2}{n}$
$\hat{\pi} = \frac{1}{n} \sum_{i=1}^n X_i$	einf. ZA	$\frac{\pi(1-\pi)}{n}$	0	$\frac{\pi(1-\pi)}{n}$
$S^{*2} = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$	μ bek.	$\frac{2\sigma^4}{n}$	0	$\frac{2\sigma^4}{n}$
$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$	μ unbek.	$\frac{2\sigma^4}{n-1}$	0	$\frac{2\sigma^4}{n-1}$
$S'^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$	μ unbek.	$\frac{2\sigma^4(n-1)}{n^2}$	$-\frac{\sigma^2}{n}$	$\frac{\sigma^4(2n-1)}{n^2}$

Für $n \rightarrow \infty$ geht der MSE bei allen obigen Schätzfunktionen gegen Null.

Erwartungstreue

Eine Schätzfunktion $\hat{\theta}_n$ des unbekannten Parameters ϑ heißt

- **erwartungstreu oder unverzerrt (unbiased)** wenn der Erwartungswert der Schätzfunktion mit dem wahren Parameter übereinstimmt:

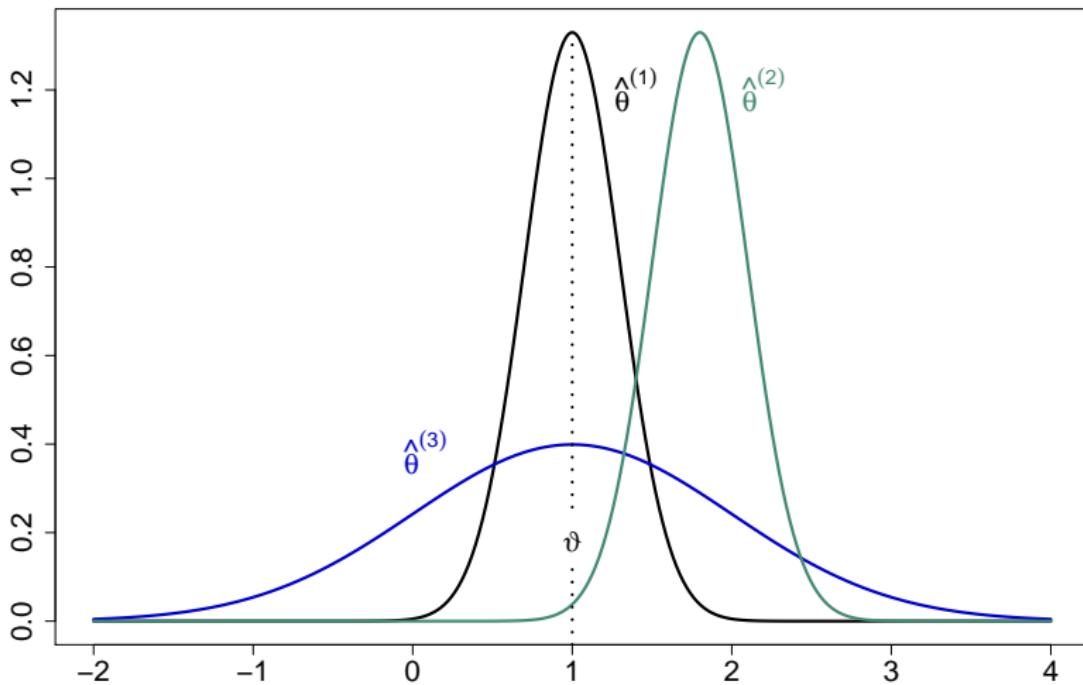
$$E(\hat{\theta}_n) = E[g(X_1, \dots, X_n)] = \vartheta \quad (1)$$

$$MSE = \text{Var}(\hat{\theta}_n) \quad (2)$$

- **asymptotisch erwartungstreu** wenn gilt:

$$\lim_{n \rightarrow \infty} E(\hat{\theta}_n) = \vartheta \quad (3)$$

Die Verzerrung geht mit wachsendem Stichprobenumfang n gegen Null



Beispiel 16.1

- Erwartungstreue Schätzfunktionen:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad E(\bar{X}) = \mu \quad (1)$$

$$\hat{\pi} = \frac{1}{n} \sum_{i=1}^n X_i, \quad X_i = \begin{cases} 1 \\ 0 \end{cases} \quad E(\hat{\pi}) = \pi \quad (2)$$

$$S^{*2} = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 \quad E(S^{*2}) = \sigma^2 \quad (3)$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad E(S^2) = \sigma^2 \quad (4)$$

- Nicht erwartungstreue Schätzfunktion:

$$S'^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \quad E(S'^2) = \frac{n-1}{n} \sigma^2 \quad (1)$$

- ▶ Verzerrung:

$$E(S'^2) - \sigma^2 = \frac{n-1}{n} \sigma^2 - \sigma^2 = -\frac{1}{n} \sigma^2 \quad (2)$$

Interpretation von Erwartungstreue

- Erwartungstreue (Unverzerrtheit) heißt, dass bei einer großen Anzahl der Stichproben der Durchschnittswert aller Schätzungen nahe dem wahren Parameter liegt
- Sind zwei Schätzfunktionen in allen Eigenschaften gleich bis auf die Erwartungstreue, dann ist der erwartungstreue Schätzer vorzuziehen; also z.B. S^2 statt S'^2

Effizienz

Voraussetzung Erwartungstreue Schätzfunktionen mit gleichem Stichprobenumfang für den unbekannten Parameter ϑ

$$E(\hat{\theta}_n^{(1)}) = \vartheta \quad E(\hat{\theta}_n^{(2)}) = \vartheta \quad (1)$$

Relative Effizienz Die Schätzfunktion $\hat{\theta}_n^{(1)}$ heißt relativ effizient zu $\hat{\theta}_n^{(2)}$, wenn die Varianz von $\hat{\theta}_n^{(1)}$ kleiner ist als die Varianz von $\hat{\theta}_n^{(2)}$:

$$\text{Var}(\hat{\theta}_n^{(1)}) \leq \text{Var}(\hat{\theta}_n^{(2)}) \quad (2)$$

Absolute Effizienz Die Schätzfunktion $\hat{\theta}_n^{(1)}$ heißt absolut effizient für ϑ , wenn sie im Vergleich zu jeder anderen erwartungstreuen Schätzfunktion für ϑ die kleinste Varianz aufweist

Konsistenz

Eine Schätzfunktion $\hat{\theta}_n$ des unbekannten Parameters ϑ heißt konsistent, wenn die beiden Bedingungen

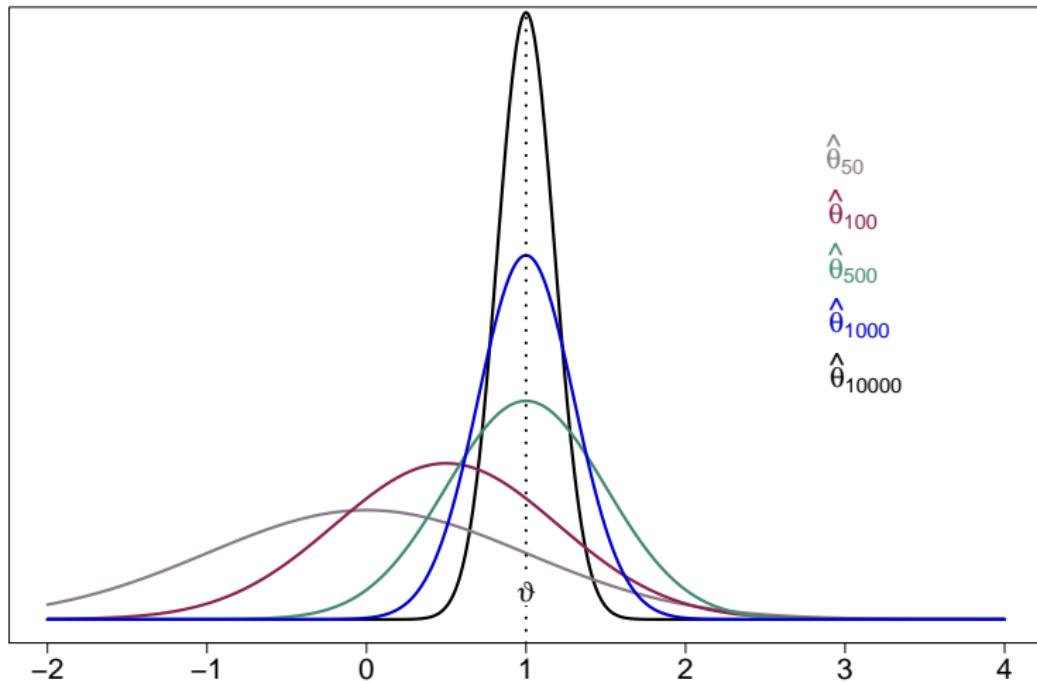
$$\lim_{n \rightarrow \infty} E[\hat{\theta}_n] = \vartheta \quad (1)$$

und

$$\lim_{n \rightarrow \infty} \text{Var}(\hat{\theta}_n) = 0 \quad (2)$$

gelten.

D.h. bei steigendem Stichprobenumfang n geht sowohl die Verzerrung, als auch die Varianz der Schätzfunktion gegen Null.



Maximum-Likelihood-Methode

Voraussetzungen:

- Verteilung der Zufallsvariablen X in der Grundgesamtheit: $f(x|\vartheta)$
 - ▶ Verteilung hängt von einem unbekannten Parameter ϑ ab
 - ▶ Verteilung muß vom Typ bekannt sein
- Einfache Zufallsstichprobe vom Umfang n
- Stichprobenvariablen X_1, \dots, X_n unabhängig und identisch verteilt wie X in der Grundgesamtheit:

$$P(X_i = x_i) = f(x_i|\vartheta) \quad \text{für alle } i = 1, \dots, n \quad (1)$$

- Gesamte Stichprobe:

$$\{X_1 = x_1\} \cap \{X_2 = x_2\} \cap \dots \cap \{X_n = x_n\} \quad (1)$$

Wie groß ist die Wahrscheinlichkeit, diese Stichprobe zu erhalten?

- gemeinsame Verteilung aller Stichprobenvariablen:

$$P(\{X_1 = x_1\} \cap \dots \cap \{X_n = x_n\}) = f(x_1, \dots, x_n | \vartheta) \quad (2)$$

$$\stackrel{*}{=} f(x_1 | \vartheta) \cdot \dots \cdot f(x_n | \vartheta) = \prod_{i=1}^n f(x_i | \vartheta) \quad (3)$$

* gilt wegen Unabhängigkeit von X_1, X_2, \dots, X_n

- Vor der Ziehung der Stichprobe:
 - ▶ $f(x_1, \dots, x_n | \vartheta)$ hängt
 - ★ von den konkreten Realisierungen x_1, \dots, x_n der Stichprobenvariablen
 - ★ vom unbekannten Parameter ϑ ab
- Nach der Ziehung der Stichprobe:
 - ▶ Stichprobenwerte x_1, \dots, x_n liegen vor
 - ▶ $f(x_1, \dots, x_n | \vartheta)$ hängt nur noch vom Parameter ϑ ab

Likelihood-Funktion

$$L(\vartheta) = L(\vartheta|x_1, \dots, x_n) = \prod_{i=1}^n f(x_i|\vartheta) \quad (1)$$

Für jeden möglichen Wert ϑ gibt $L(\vartheta)$ die Wahrscheinlichkeit für die konkret realisierte Stichprobe (x_1, \dots, x_n) an.

Stichprobe	ϑ_1	ϑ_2	ϑ_3	...
x_1	$f(x_1 \vartheta_1)$	$f(x_1 \vartheta_2)$	$f(x_1 \vartheta_3)$...
x_2	$f(x_2 \vartheta_1)$	$f(x_2 \vartheta_2)$	$f(x_2 \vartheta_3)$...
x_3	$f(x_3 \vartheta_1)$	$f(x_3 \vartheta_2)$	$f(x_3 \vartheta_3)$...
\vdots	\vdots	\vdots	\vdots	\vdots
x_n	$f(x_n \vartheta_1)$	$f(x_n \vartheta_2)$	$f(x_n \vartheta_3)$...
$L(\vartheta)$	$\prod_{i=1}^n f(x_i \vartheta_1)$	$\prod_{i=1}^n f(x_i \vartheta_2)$	$\prod_{i=1}^n f(x_i \vartheta_3)$...

- Wähle den Wert $\hat{\vartheta}$ als Schätzung, der die beobachtete Stichprobe am wahrscheinlichsten macht:

$$\hat{\vartheta} = \arg \max_{\vartheta} L(\vartheta) \quad (1)$$

⇒ d.h. finde das Maximum der Likelihood-Funktion

- Bedingungen für ein Maximum:

$$\frac{\partial L(\vartheta)}{\partial \vartheta} \Big|_{\vartheta=\hat{\vartheta}} = 0 \quad (2)$$

$$\frac{\partial^2 L(\vartheta)}{\partial^2 \vartheta} \Big|_{\vartheta=\hat{\vartheta}} < 0 \quad (3)$$

- Die Log-Likelihood-Funktion $\ln L(\vartheta)$ ist der natürliche Logarithmus der Likelihood-Funktion
- die Likelihood-Funktion und die Log-Likelihood-Funktion haben ihr Maximum an der gleichen Stelle

$$\frac{\partial L(\vartheta)}{\partial \vartheta} \Big|_{\vartheta=\hat{\vartheta}} = 0 \iff \frac{\partial \ln L(\vartheta)}{\partial \vartheta} \Big|_{\vartheta=\hat{\vartheta}} = 0 \quad (1)$$

⇒ maximiere die Log-Likelihood-Funktion anstatt der Likelihood-Funktion

- Warum? Das Maximum der Log-Likelihood-Funktion ist oft einfacher zu berechnen

Beispiel 16.2 (Call-Center)

Situation:

- In einem Call-Center sitzen mehrere Telefonistinnen
- die Telefonistinnen erhalten unabhängig voneinander Anrufe
- Telefonistin 1 erhält 3 Anrufe pro Stunde
 - ▶ X_1 : "Zahl der Anrufe pro Stunde bei Telefonistin 1" $\sim PO(\lambda)$
- Telefonistin 2 erhält 5 Anrufe pro Stunde
 - ▶ X_2 : "Zahl der Anrufe pro Stunde bei Telefonistin 2" $\sim PO(\lambda)$

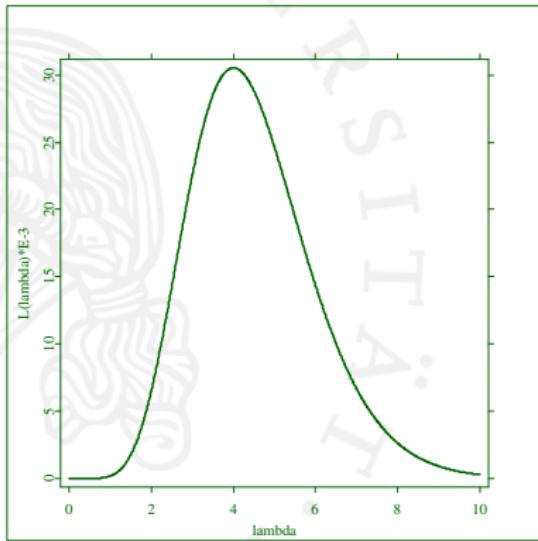
Likelihood-Funktion:

$$\begin{aligned} f_1(3|\lambda) &= \frac{\lambda^3}{3!} e^{-\lambda} \text{ und } f_2(5|\lambda) = \frac{\lambda^5}{5!} e^{-\lambda} \\ \Rightarrow L(\lambda) &= f_1(3|\lambda) \cdot f_2(5|\lambda) = \frac{\lambda^3}{3!} e^{-\lambda} \frac{\lambda^5}{5!} e^{-\lambda} \end{aligned}$$

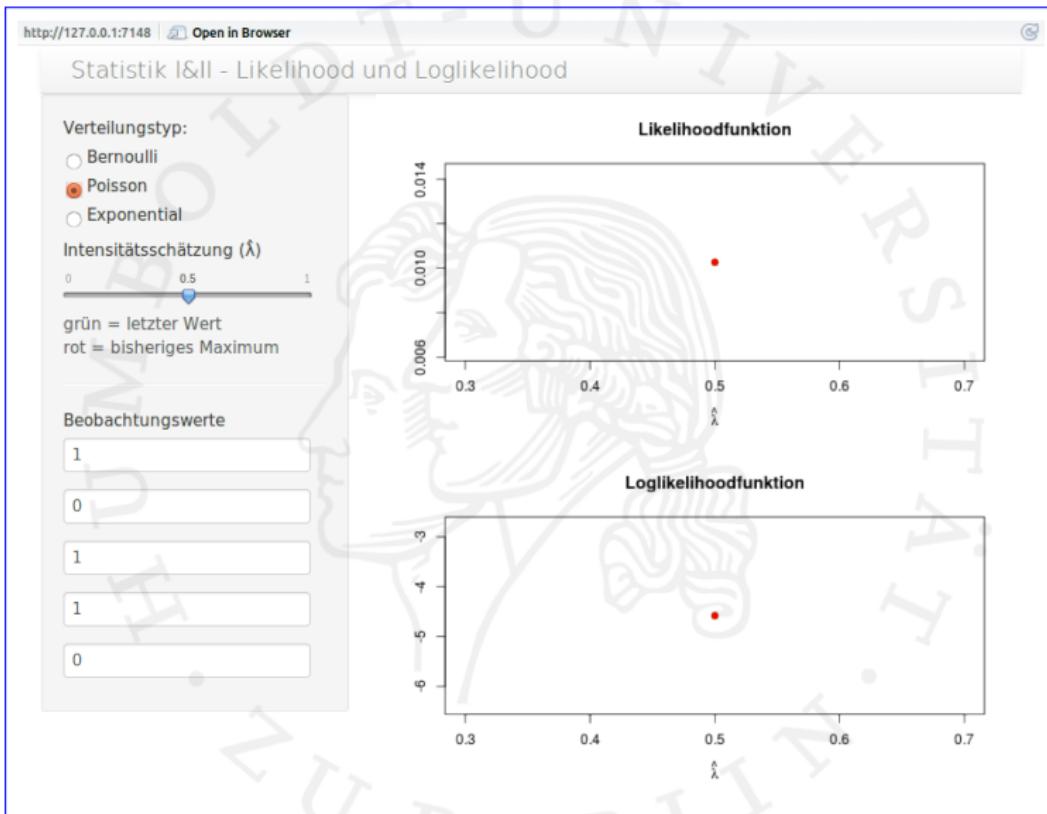
Wahl von λ :

- Wähle λ so, daß $L(\lambda)$ maximal!

λ	1	2	3
$L(\lambda)$	0,0002	0,0065	0,0226
λ	4	5	6
$L(\lambda)$	0,0305	0,0246	0,0143



- Aus der Tabelle und aus der Grafik: $\lambda = 4$



Beispiel 16.3 (Urnenmodell)

Situation:

- Ziehe 5 Kugeln (mit Zurücklegen) aus einer Urne mit roten und weißen Kugeln
- Drei der gezogenen Kugeln sind rot
- X : "Anzahl der roten Kugeln" $\sim B(n; \pi)$ mit $n = 5$

Likelihood-Funktion:

- die Reihenfolge der gezogenen Kugeln spielt keine Rolle
⇒ es gibt $\binom{5}{3}$ Möglichkeiten bei fünfmaligem Ziehen drei rote Kugeln zu ziehen (Kombinatorik)
- jede dieser Ziehungen tritt mit einer Wahrscheinlichkeit von $\pi^3 \cdot (1 - \pi)^{5-3}$ auf

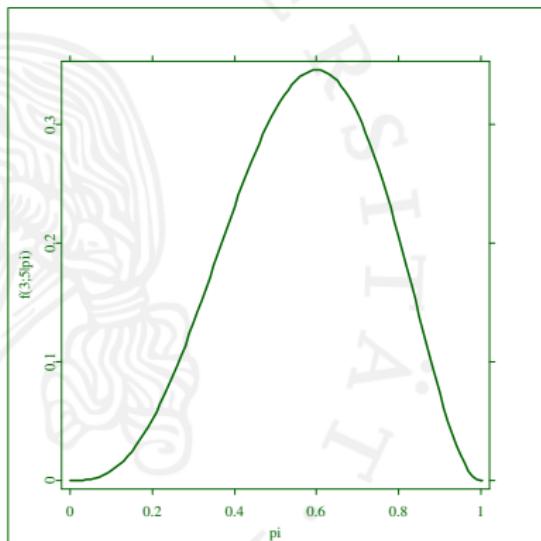
Daraus ergibt sich die Likelihood-Funktion

$$L(\pi) = \binom{5}{3} \pi^3 (1 - \pi)^{5-3} = \binom{5}{3} \pi^3 (1 - \pi)^2 = f(3; 5|\pi) \quad (1)$$

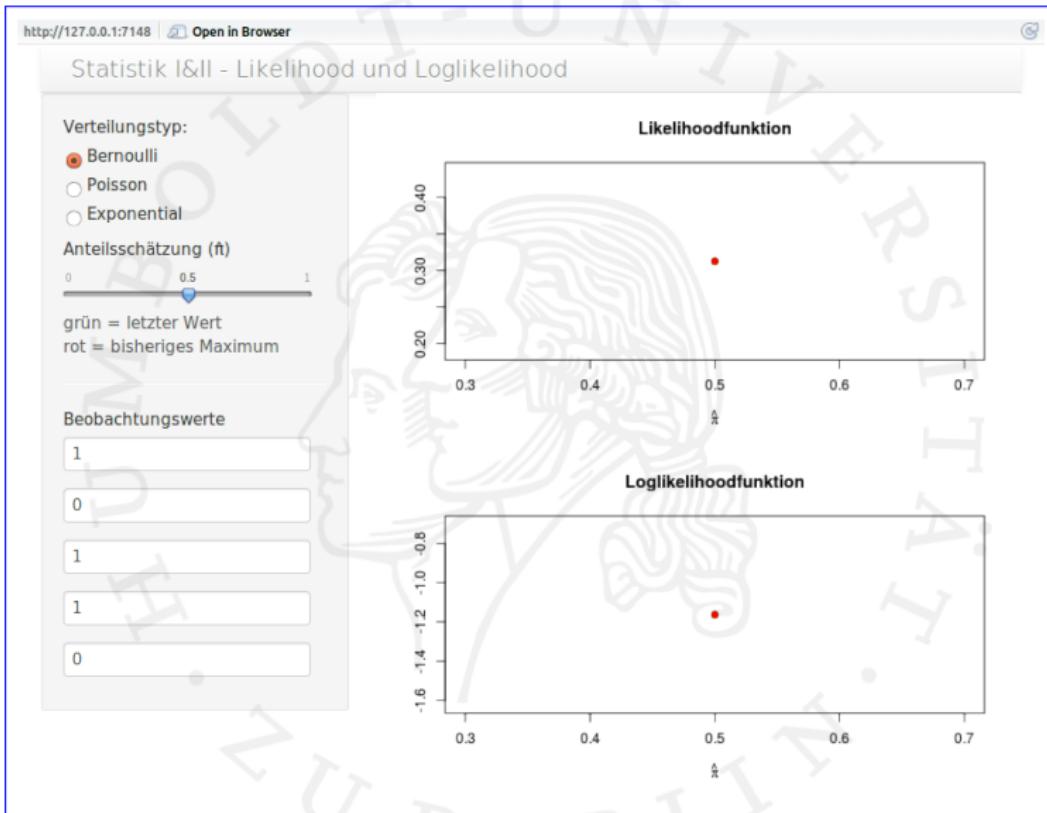
Wahl von π :

- Wähle π so, daß $f(3; 5|\pi)$ maximal!

π	$f(3; 5 \pi)$
0,1	0,0081
0,3	0,1323
0,6	0,3456
0,9	0,0729



- Aus der Tabelle und aus der Grafik: $\pi = 0,6$



- Für diskrete Verteilungen gilt

$$L(\vartheta|x_1, \dots, x_n) = \prod_{i=1}^n P(X_i = x_i|\vartheta) \quad (1)$$

- ▶ Wahrscheinlichkeit des Eintretens einer bestimmten Stichprobe unter allen möglichen Stichproben

- Für stetige Verteilungen gilt

$$P(x_i - \delta_i/2 \leq X_i|\vartheta \leq x_i + \delta_i/2) \approx f(x_i|\vartheta)\delta_i \quad (2)$$

$$L(\vartheta|x_1, \dots, x_n) \propto \prod_{i=1}^n P(x_i - \delta_i/2 \leq X_i|\vartheta \leq x_i + \delta_i/2) \quad (3)$$

- ▶ deswegen „Likelihood“ und keine Wahrscheinlichkeit!

Vorgehensweise ML-Schätzung

- 1a bei stetigen Variablen: Dichtefunktion der einzelnen StichprobenvARIABLEN aufstellen
- 1b bei diskreten Variablen: Wahrscheinlichkeitsfunktion der einzelnen StichprobenvARIABLEN aufstellen
2. Likelihood-Funktion bilden
3. Log-Likelihood-Funktion berechnen
4. 1. Ableitung gleich Null setzen
5. nach $\hat{\theta}$ auflösen
6. überprüfen ob 2. Ableitung < 0 damit Maximum vorliegt

ML-Schätzer für λ in $PO(\lambda)$

- Wahrscheinlichkeitsfunktion von X_i ($i = 1, \dots, n$)

$$f_{PO}(x_i; \lambda) = \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \quad (1)$$

- Likelihood-Funktion für die Stichprobe x_1, \dots, x_n

$$L(\lambda | x_1, \dots, x_n) = \prod_{i=1}^n \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = \frac{\lambda^{x_1 + \dots + x_n}}{x_1! \cdot \dots \cdot x_n!} e^{-n\lambda} \quad (2)$$

- Log-Likelihood-Funktion

$$\ln L = \sum_{i=1}^n \ln \left(\frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \right) = \sum_{i=1}^n (x_i \ln \lambda - \ln(x_i!) - \lambda) \quad (3)$$

4. 1. Ableitung gleich Null setzen

$$\frac{\partial \ln L}{\partial \lambda} = \sum_{i=1}^n \left(\frac{x_i}{\lambda} - 1 \right) \stackrel{!}{=} 0 \quad (1)$$

5. nach $\hat{\lambda}$ auflösen

$$\frac{1}{\hat{\lambda}} \sum_{i=1}^n x_i = n \quad (2)$$

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x} \quad (3)$$

6. 2. Ableitung überprüfen

$$\frac{\partial^2 \ln L}{\partial \lambda^2} = - \sum_{i=1}^n \frac{x_i}{\lambda^2} < 0 \quad (4)$$

ML-Schätzer für π in $B(n, \pi)$

1. Wahrscheinlichkeitsfunktion von X_i ($i = 1, \dots, n$)

$$f(x_i|\pi) = \pi^{x_i} \cdot (1 - \pi)^{1-x_i} \quad (1)$$

2. Likelihood-Funktion von der Stichprobe x_1, \dots, x_n

$$L(\pi|x_i) = \binom{n}{x_i} \cdot \pi^{x_i} \cdot (1 - \pi)^{n-x_i} \quad (2)$$

(Herleitung siehe Beispiel 7.3)

3. Log-Likelihood-Funktion

$$\ln L(\pi|x_i) = \ln \binom{n}{x_i} + x_i \ln \pi + (n - x_i) \ln(1 - \pi) \quad (3)$$

4. 1. Ableitung gleich Null setzen

$$\frac{\partial L}{\partial \pi} = \frac{x_i}{\pi} - \frac{n - x_i}{1 - \pi} \stackrel{!}{=} 0 \quad (1)$$

5. nach $\hat{\pi}$ auflösen

$$x(1 - \hat{\pi}) = (n - x_i)\hat{\pi} \quad (2)$$

$$\hat{\pi} = \frac{x_i}{n} = p \quad (3)$$

6. 2. Ableitung überprüfen

$$\frac{\partial^2 L}{\partial \pi^2} = -\frac{x_i}{\pi^2} - \frac{n - x_i}{(1 - \pi)^2} < 0 \quad (4)$$

ML-Schätzer für λ in $Exp(\lambda)$

1. Dichtefunktion von X_i ($i = 1, \dots, n$)

$$f_{EX}(x_i|\lambda) = \begin{cases} \lambda e^{-\lambda x_i} & \text{für } x_i \geq 0, \lambda > 0 \\ 0 & \text{für } x_i < 0 \end{cases} \quad (1)$$

2. Likelihood-Funktion für die Stichprobe x_1, \dots, x_n

$$L(\lambda|x_1, \dots, x_n) = \prod_{i=1}^n \lambda e^{-\lambda x_i} = \lambda^n \prod_{i=1}^n e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^n x_i} \quad (2)$$

3. Log-Likelihood-Funktion

$$\ln L(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^n x_i \quad (3)$$

4. 1. Ableitung gleich Null setzen

$$\frac{\partial \ln L(\lambda)}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^n x_i \stackrel{!}{=} 0 \quad (1)$$

5. nach $\hat{\lambda}$ auflösen

$$\frac{n}{\hat{\lambda}} = \sum_{i=1}^n x_i \Leftrightarrow \hat{\lambda} = \frac{n}{\sum_{i=1}^n x_i} = \frac{1}{\bar{x}} \quad (2)$$

6. 2. Ableitung überprüfen

$$\frac{\partial^2 \ln L(\lambda)}{\partial \lambda^2} = -\frac{n}{\lambda^2} < 0 \quad (3)$$

<http://127.0.0.1:7148>

[Open in Browser](#)

Statistik I&II - Likelihood und Loglikelihood

Verteilungstyp:

- Bernoulli
- Poisson
- Exponential

Intensitätsschätzung ($\hat{\lambda}$)

grün = letzter Wert

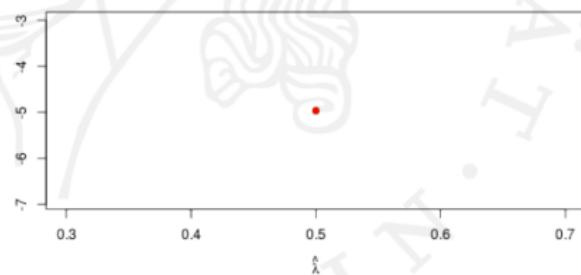
rot = bisheriges Maximum

Beobachtungswerte

Likelihoodfunktion



Loglikelihoodfunktion



ML-Schätzer für μ & σ^2 in $N(\mu; \sigma^2)$

1. Dichtefunktion von X_i ($i = 1, \dots, n$)

$$f(x_i|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \quad (1)$$

2. Likelihood-Funktion

$$\begin{aligned} L(\mu, \sigma^2 | x_1, \dots, x_n) &= \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n e^{-\frac{1}{2\sigma^2} \cdot \sum_{i=1}^n (x_i - \mu)^2} \\ &= (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2} \end{aligned}$$

3. Log-Likelihood-Funktion

$$\ln L(\mu, \sigma^2 | x_1, \dots, x_n) = -\frac{n}{2} \cdot \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \cdot \sum_{i=1}^n (x_i - \mu)^2$$

im Folgenden: Substitution $\sigma^2 = \psi$

ML-Schätzer für μ

4. 1. Ableitung gleich Null setzen

$$\frac{\partial \ln L}{\partial \mu} = -\frac{2 \cdot \sum_{i=1}^n (x_i - \mu) \cdot (-1)}{2\psi} \stackrel{!}{=} 0 \quad (1)$$

5. nach $\hat{\mu}$ auflösen

$$\sum_{i=1}^n (x_i - \hat{\mu}) = 0 \quad (2)$$

$$\hat{\mu} = \frac{\sum_{i=1}^n x_i}{n} = \bar{x} \quad (3)$$

6. 2. Ableitung überprüfen

$$\frac{\partial^2 \ln L}{\partial \mu^2} = -\frac{n}{\psi} < 0 \quad (4)$$

ML-Schätzer für σ^2

4. 1. Ableitung gleich Null setzen

$$\frac{\partial \ln L}{\partial \psi} = -\frac{n}{2} \cdot \frac{1}{\psi} + \frac{1}{2} \cdot \frac{1}{\psi^2} \cdot \sum_{i=1}^n (x_i - \mu)^2 \stackrel{!}{=} 0 \quad (1)$$

5. nach $\widehat{\psi}$ auflösen

$$\frac{n}{2\widehat{\psi}} = \frac{1}{2\widehat{\psi}^2} \sum_{i=1}^n (x_i - \mu)^2 \quad (2)$$

$$\widehat{\psi} = \widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 \quad (3)$$

6. 2. Ableitung überprüfen

$$\frac{\partial^2 \ln L}{\partial \psi^2} = \frac{n}{2\psi^2} - \frac{1}{\psi^3} \sum_{i=1}^n (x_i - \mu)^2 < 0 \quad (4)$$

Grundgesamtheit und ML-Schätzung

Grundgesamtheit



Verteilung von X

$$\mu, \sigma^2$$

Maximum-Likelihood



Einfache Zufallsstichprobe



Verteilung von X_i

$$E(X_i) = \mu, \text{Var}(X_i) = \sigma^2$$

z.B. Annahme $X_i \sim Po(\lambda)$

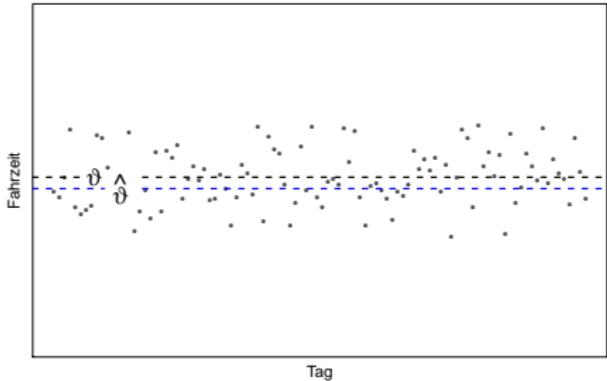
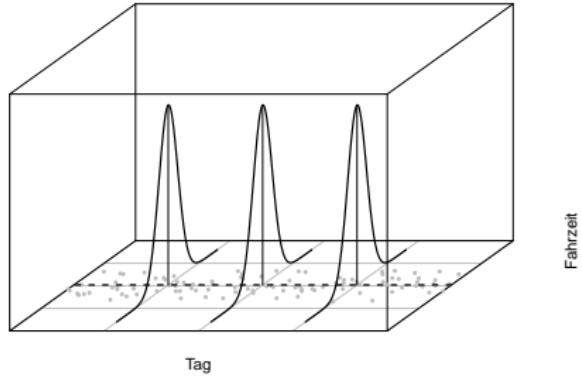
$$E(X_i) = \lambda, \text{Var}(X_i) = \lambda$$

$$\hat{\lambda} = \bar{x} \stackrel{!}{=} \hat{\mu}$$

Methode der kleinsten Quadrate

Beispiel 16.4

Herr Müller fährt jeden Tag mit dem Auto zur Arbeit. Daher möchte er gerne wissen, wie lange diese Fahrt dauert. Er misst seine Fahrzeit an 100 Tagen. Aufgrund von Faktoren wie Verkehrsstärke, Ampelschaltung etc. schwankt seine Fahrzeit allerdings. Wie kann Herr Müller herausfinden, was seine zu erwartende Fahrzeit ist?



- Die Erwartungswerte der Stichprobenvariablen X_1, \dots, X_n hängen über eine bekannte Funktion von dem unbekannten Parameter ϑ der Grundgesamtheit ab:

$$E(X_i) = g_i(\vartheta) \quad i = 1, \dots, n \quad (1)$$

- Summe der quadrierten Abweichungen der Stichprobenvariablen von $g_i(\hat{\vartheta})$:

$$Q(\hat{\vartheta}) = \sum_{i=1}^n (X_i - E(X_i))^2 = \sum_{i=1}^n (X_i - g_i(\hat{\vartheta}))^2 \quad (2)$$

- Schätzung $\hat{\vartheta}$ wird so gewählt, dass die Summe der quadrierten Abweichungen zwischen den Stichprobenwerten und $g_i(\hat{\vartheta})$ klein wird
- Keine Voraussetzung über die Verteilung der Grundgesamtheit notwendig!

Vorgehensweise KQ-Schätzung

1. Zusammenhang zwischen Erwartungswert und unbekanntem Parameter der Grundgesamtheit bestimmen
2. Zielfunktion aufstellen (Summe der quadrierten Abweichungen)
3. 1. Ableitung gleich Null setzen
4. nach $\hat{\theta}$ auflösen
5. überprüfen ob 2. Ableitung > 0 damit Minimum vorliegt

KQ-Schätzer für μ

1. Zusammenhang zwischen Erwartungswert und unbekanntem Parameter der Grundgesamtheit bestimmen
 - ▶ Der unbekannte Parameter ϑ ist der Mittelwert μ der Grundgesamtheit
- $$\vartheta = \mu \quad (1)$$
- $$\Rightarrow E(X_i) = g_i(\mu) = \mu \text{ für alle } i \quad (2)$$
2. Die Summe der Quadrate der Abweichungen der Stichprobenvariablen vom Erwartungswert μ

$$Q(\mu) = \sum_{i=1}^n (X_i - \mu)^2 \rightarrow \min \quad (3)$$

3. 1. Ableitung gleich Null setzen

$$\frac{\partial Q(\mu)}{\partial \mu} = -2 \sum_{i=1}^n (X_i - \mu) \stackrel{!}{=} 0 \quad (1)$$

4. nach $\hat{\mu}$ auflösen

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x} \quad (2)$$

5. 2. Ableitung überprüfen

$$\frac{\partial^2 Q(\mu)}{\partial \mu^2} = 2n > 0 \quad (3)$$

ML-Schätzung vs. KQ-Schätzung

Beispiel 16.5

Ein Spielautomat besitzt folgende Wahrscheinlichkeitsverteilung für den Gewinn X pro Spiel (in €):

x	-1	0	+1
$P(X = x)$	p	p	$1 - 2p$

Es wird eine Zufallsstichprobe vom Umfang $n = 6$ gezogen, die sich folgendermaßen realisiert:

x_1	x_2	x_3	x_4	x_5	x_6
-1	1	-1	0	1	1

Frage:

Schätzen Sie p aufgrund der vorliegenden Stichprobe nach der ML-Methode und nach der KQ-Methode.

Lösung:

1. ML-Methode

Sei

- ▶ v = Anzahl der verlorenen Spiele
- ▶ u = Anzahl der unentschiedenen Spiele
- ▶ g = Anzahl der gewonnenen Spiele

$$\Rightarrow n = v + u + g$$

$$L(p) = p^v \cdot p^u \cdot (1 - 2p)^{n-(v+u)} \quad (1)$$

$$\ln L(p) = (v + u) \ln(p) + (n - (v + u)) \ln(1 - 2p) \quad (2)$$

$$\frac{\partial \ln L(p)}{\partial p} = (v + u) \cdot \frac{1}{p} + (n - (v + u)) \cdot \frac{-2}{1 - 2p} \stackrel{!}{=} 0 \quad (3)$$

$$\hat{p} = \frac{v + u}{2n} = \frac{2 + 1}{2 \cdot 6} = \frac{1}{4} = \frac{4,5}{18} \quad (4)$$

2. KQ-Methode

$$E(X_i) = -1 \cdot p + 0 \cdot p + 1 \cdot (1 - 2p) = 1 - 3p \quad (1)$$

$$= g_i(p) = 1 - 3p \quad (2)$$

$$Q(p) = \sum_{i=1}^n (x_i - E(X_i))^2 = \sum_{i=1}^n (x_i - (1 - 3p))^2 \quad (3)$$

$$\frac{\partial Q(p)}{\partial p} = \sum_{i=1}^n (6x_i - 6 + 18p) = 6 \sum_{i=1}^n x_i - 6n + 18np \stackrel{!}{=} 0 \quad (4)$$

$$\hat{p} = \frac{6n - 6 \sum_{i=1}^n x_i}{18n} = \frac{1 - \frac{g-v}{n}}{3} = \frac{5}{18} \quad (5)$$

Aber:

	x	-1	0	$+1$	$E(X)$	
P($X = x$)	$p/2$	$2p$	$1 - 5p/2$	$1 - 3p$	$\Rightarrow \hat{p} = 5/18$	

- ML-Schätzung
 - ▶ Vorteil: liefert in der Regel konsistente Schätzer
 - ▶ Nachteil: der Verteilungstyp in der Grundgesamtheit muss festgelegt werden
- KQ-Schätzung
 - ▶ Vorteil: benötigt nur Informationen über der Erwartungswert in der Grundgesamtheit
 - ▶ Nachteil: liefert u.U. inkonsistenten Schätzer
- Begründung
 - ▶ der ML-Schätzung benötigt mehr Informationen (Verteilungsfunktion) über die Grundgesamtheit und liefert i.A. bessere Schätzer
 - ▶ der KQ-Schätzung benötigt weniger Informationen (Erwartungswert) über die Grundgesamtheit und liefert u.U. schlechtere Schätzer

Konstruktion von Schätzfunktionen

1. Konstruiere eine Lösung zur Schätzung von Parameter für n Beobachtungen, z.B.

$$X_i \sim N(\mu, \sigma^2) \text{ mit } \mu, \sigma \text{ unbekannt} \quad (1)$$

2. Maximum-Likelihood-Lösung:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (2)$$

3. Ersetze x_i durch die Stichprobenvariablen X_i

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad S'^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (3)$$

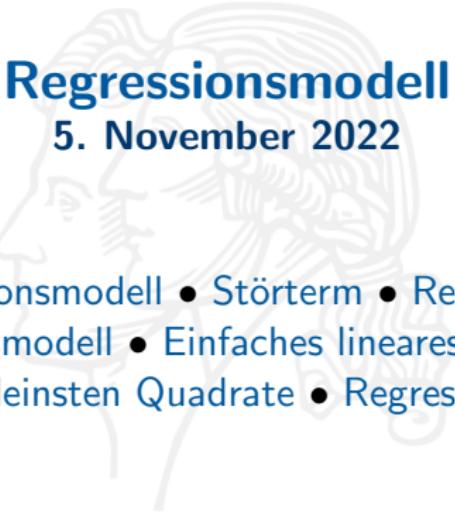
4. Analysiere die Eigenschaften der Schätzfunktionen

$$E(\bar{X}) = \mu \quad E(S'^2) = \frac{n-1}{n} \sigma^2 \quad (1)$$

5. Modifizierte die Schätzfunktion, z.B. um Unverzerrtheit zu erreichen

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (2)$$

6. Weitere Analysen, z.B. Konsistenzrate, also wie schnell konvergiert der geschätzte Parameter gegen den wahren Parameter in Abhängigkeit von n (\rightarrow benötigte Stichprobenumfänge?)



Regressionsmodell

5. November 2022

- Allgemeines Regressionsmodell
- Störterm
- Residuum
- Klassisches (lineares) Regressionsmodell
- Einfaches lineares Regressionsmodell
- Methode der kleinsten Quadrate
- Regressionsdiagnostik

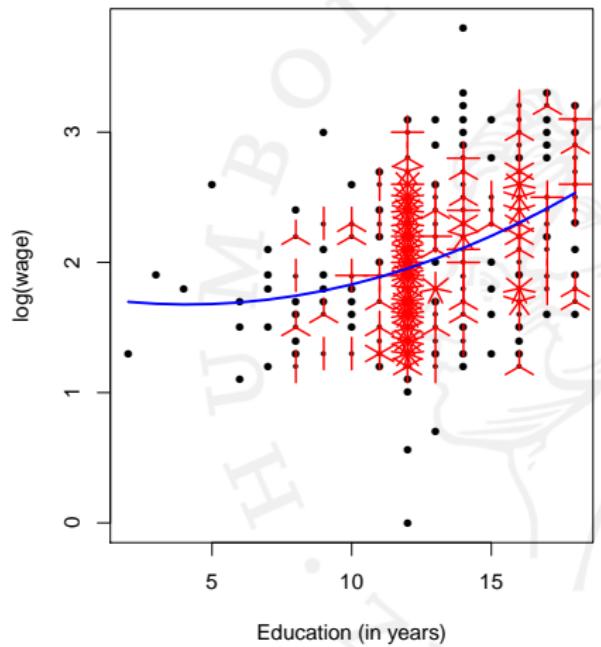
Allgemeines Regressionsmodell

- Für alle gegebene Variablenwerte $x_{1i}, x_{2i}, \dots, x_{pi}$ definiert man im Regressionsmodell Zufallsvariablen Y_i
- Darstellung der **mittleren** statistischen Abhängigkeit der endogenen Variablen Y von den exogenen Variablen X_1, X_2, \dots, X_p mittels einer Funktion

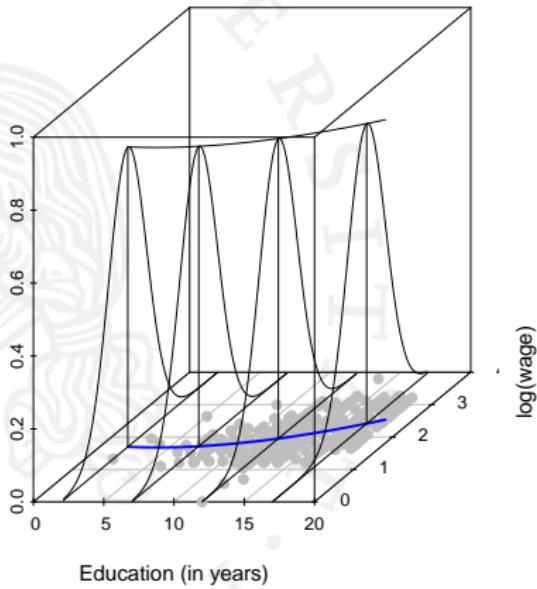
$$E(Y|x_{1i}, x_{2i}, \dots, x_{pi}) = E(Y_i) = m(x_{1i}, x_{2i}, \dots, x_{pi})$$

- $E(Y_i)$: Funktionswert (Regresswert), der im Mittel zu erwartende Wert von Y bei gegebenen Variablenwerten $x_{1i}, x_{2i}, \dots, x_{pi}$ und gegebener Funktion $m(\bullet)$

Sunflower plot of CPS 1985 data



Regression model



Störterm

Mögliche weitere Einflüsse auf Y :

- weitere erklärende X -Variablen (systematische Einflüsse)
- Zufallseinflüsse

Die Zufallsvariable Y_i setzt sich zusammen aus :

- einer Funktion der bekannten Einflüsse $x_{1i}, x_{2i}, \dots, x_{pi}$
- und dem nicht beobachtbaren Störterm $\varepsilon_i = \varepsilon | x_{i1}, \dots, x_{pi}$:

$$Y_i = m(x_{1i}, x_{2i}, \dots, x_{pi}) + \varepsilon_i$$

- Die Störterme ε_i enthalten vor allem Zufallseinflüsse
⇒ Die Störterme ε_i werden mit Zufallsvariablen U_i modelliert

Residuum

$$U_i, i = 1, \dots, n$$

- Abweichung zwischen der Zufallsvariablen Y_i und dem Regresswert $E(Y_i)$ in der zufällig gezogenen Stichprobe

$$U_i = Y_i - E(Y_i)$$

- Alle Einflüsse auf die Variable Y an der Stelle i , die nicht durch die Regressionsfunktion und damit durch die darin enthaltenen Variablen X_1, X_2, \dots, X_p erfasst werden

Klassisches (lineares) Regressionsmodell

Modellannahmen:

1. Die n Zufallsvariablen U_i haben den Erwartungswert Null:
 $E(U_i) = 0 \quad \text{für } i = 1, \dots, n.$
2. Die Varianz der n Zufallsvariablen U_i ist bei allen statistischen Einheiten gleich und konstant:
 $\text{Var}(U_i) = \sigma_u^2 \quad \text{für } i = 1, \dots, n$
3. Die n Zufallsvariablen U_i sind nicht korreliert (stochastisch unabhängig):
 $\text{Cov}(U_i, U_j) = \sigma_{ij} = 0 \quad \text{für } i, j = 1, \dots, n; i \neq j$
4. Die Zufallsvariablen U_i ($i = 1, \dots, n$) sind normalverteilt:
 $U_i \sim N(0; \sigma_u^2)$
5. Die Werte x_{ki} der Variablen X_k ($i = 1, \dots, n; k = 1, \dots, p$) sind feste Größen unabhängig von der Stichprobenziehung.

Konsequenz:

$$\underbrace{Y_i}_{\text{Zufallsvariable}} = \underbrace{m(x_{1i}, \dots, x_{pi})}_{\text{konstant}} + \underbrace{U_i}_{\text{Zufallsvariable}}$$

$$U_i \sim (0, \sigma_u^2) \implies Y_i \sim (m(x_{1i}, \dots, x_{pi}), \sigma_u^2)$$

$$U_i \sim N(0, \sigma_u^2) \implies Y_i \sim N(m(x_{1i}, \dots, x_{pi}), \sigma_u^2)$$

- Die Y_i sind normalverteilt (folgt aus Modellannahme 4)
- Da die U_i Zufallsvariablen sind, sind auch die Werte der Variablen Y an der Stelle i ($i = 1, \dots, n$) Zufallsvariablen:
 $Y_i \Rightarrow$ Zufallsvariablen

Einfaches lineares Regressionsmodell

Einschränkung der weiteren Betrachtungen:

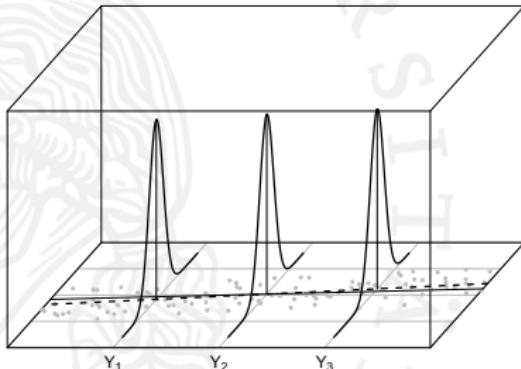
- eine unabhängige Variable X
- Wahl einer linearen Funktion
- Regressionsmodell:

$$Y_i = E(Y_i) + U_i \quad i = 1, \dots, n$$

- Wahre Regressionsgerade:

$$E(Y_i) = \beta_0 + \beta_1 x_i \quad i = 1, \dots, n$$

Die Parameter β_0 und β_1 sind unbekannt, müssen aus der Stichprobe als b_0 und b_1 geschätzt werden.



Methode der kleinsten Quadrate

- Zusammenhang zwischen dem Erwartungswert und dem Regresswert bestimmen:

$$\begin{aligned}\sum_{i=1}^n U_i^2 &= \sum_{i=1}^n ((\underbrace{\beta_0 + \beta_1 x_i + U_i}_{Y_i}) - (\underbrace{\beta_0 + \beta_1 x_i}_{E(Y_i)}))^2 \\ &= \sum_{i=1}^n (Y_i - E(Y_i))^2\end{aligned}$$

- Zielfunktion aufstellen:

$$Q(b_0, b_1) = \sum_{i=1}^n (y_i - \underbrace{f(b_0, b_1)}_{b_0 + b_1 x_i = \hat{y}_i})^2 \rightarrow \text{minimal}$$

Regressionsdiagnostik

Probleme bei Verletzung der Modellannahmen

1. Verzerrte Schätzung von $\beta_0 \rightarrow$ noch Struktur in den Residuen
2. Verzerrte Schätzung von $\text{Var}(B_i) \rightarrow$ Ineffiziente Schätzung von β_i
3. Verzerrte Schätzung von $\text{Var}(B_i) \rightarrow$ Ineffiziente Schätzung von β_i
4. Verteilung von B_i wird falsch bestimmt
5. kann in der Regel nicht geprüft werden

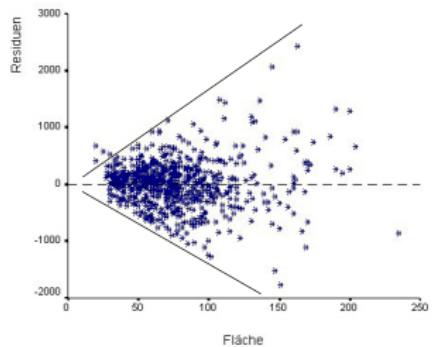
Mögliche Ursachen

1. Untergruppen in den Daten
2. Zeitliche oder räumliche Abhängigkeit
3. Linearkombination der erklärenden Variablen bildet abhängige Variable nicht gut genug ab
4. X_i sind auch zufällige Größen

Prüfung der Modellannahmen durch geschätzte Werte

1. Analyse des Medians der geschätzten Residuen \hat{u}_i ;
2. Plots der geschätzten Residuen \hat{u}_i gegen \hat{y}_i , y_i und x_i
3. Scatterplot der Residuen gegeneinander \hat{u}_i gegen \hat{u}_{i-k}
4. Plots der geschätzten Residuen \hat{u}_i gegen Normalverteilung (Q-Q-Plot)
5. kann in der Regel nicht geprüft werden

Beispiel 17.1



nicht homogen, da keine gleichmäßige Streuung
der Residuen

Konfidenzintervalle

5. November 2022

- Punkts- vs. Intervallschätzung • Zentrales Schwankungsintervall •
- Zentrales Schwankungsintervall für μ • Schwankungsintervall vs. Konfidenzintervall • Konfidenzintervall für μ • Bestimmung eines Konfidenzintervalls • Schwankungs- & Konfidenzintervall • Zentrales Konfidenzintervall • Konfidenzintervall für μ • Konfidenzintervall für π •
- Einfluß von α und n • Bestimmung von n • Konfidenzintervalle für β_i •
- Konfidenzintervall für $E(Y|x_h)$ • Häufige Missverständnisse • Übersicht Konfidenzintervalle • Anhang

Punkts- vs. Intervallschätzung

- Punktschätzung
 - ▶ man erhält für den unbekannten Parameter ϑ einen Schätzwert $\hat{\vartheta}$ als Realisation einer Zufallsvariablen (der Schätzfunktion)
- Intervallschätzung
 - ▶ Schwankungsintervall: man erhält ein Intervall um den unbekannten Parameter ϑ , dass den geschätzten Parameter $\hat{\vartheta}$ mit einer bestimmtem Wahrscheinlichkeit enthält
 - ▶ Konfidenzintervall: man erhält ein Intervall um den geschätzten Parameter $\hat{\vartheta}$, dass den unbekannten Parameter ϑ mit einer bestimmten Wahrscheinlichkeit überdeckt

Zentrales Schwankungsintervall

Zentrales Schwankungsintervall für eine normalverteilte oder t-verteilte Schätzfunktion $\hat{\theta}$:

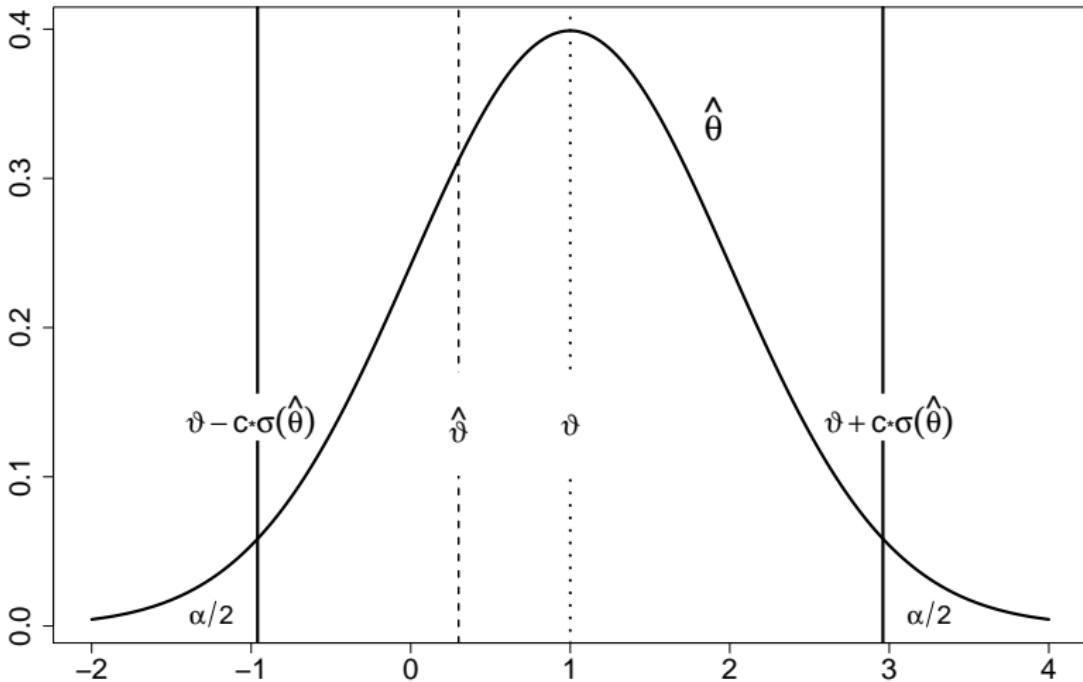
$$P(\vartheta - k \leq \hat{\theta} \leq \vartheta + k) = 1 - \alpha$$

$$P(\vartheta - c \cdot \sigma(\hat{\theta}) \leq \hat{\theta} \leq \vartheta + c \cdot \sigma(\hat{\theta})) = 1 - \alpha$$

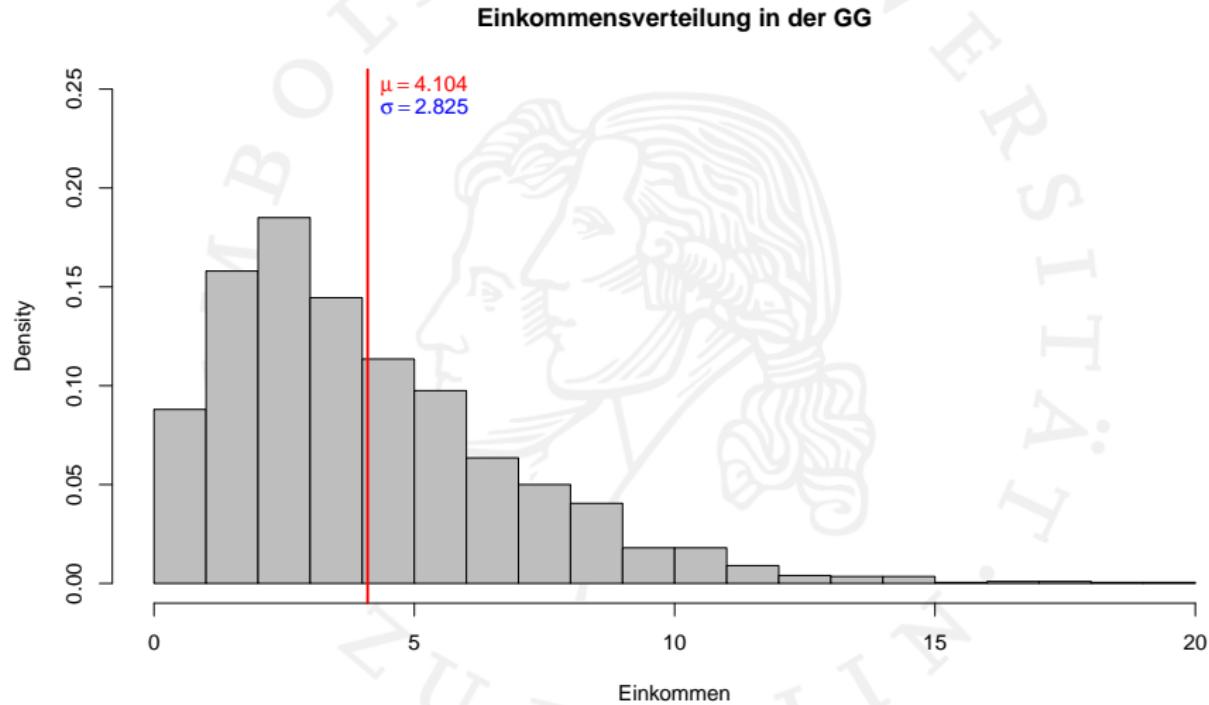
$$[\vartheta - c \cdot \sigma(\hat{\theta}) ; \vartheta + c \cdot \sigma(\hat{\theta})]$$

Bereich mit festen Grenzen um den Parameter ϑ

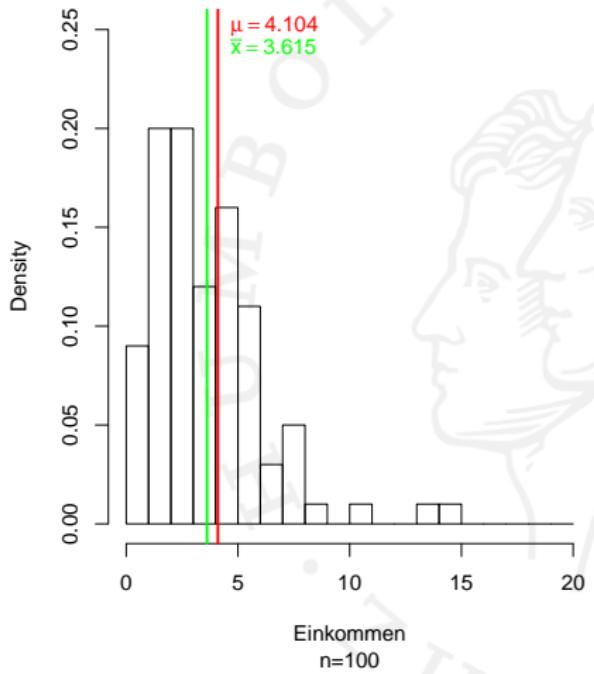
- in dem die Zufallsvariable $\hat{\theta}$ Realisationen mit einer vorgegebenen Sicherheitswahrscheinlichkeit $1 - \alpha$ annimmt
- wobei den beiden Bereichen außerhalb der Grenzen des Intervalls jeweils die gleiche Wahrscheinlichkeit $\alpha/2$ zugeordnet ist
- die Grenzen des Intervalls (i.d.R.) unter Verwendung des Standardfehlers der Schätzfunktion bestimmt werden



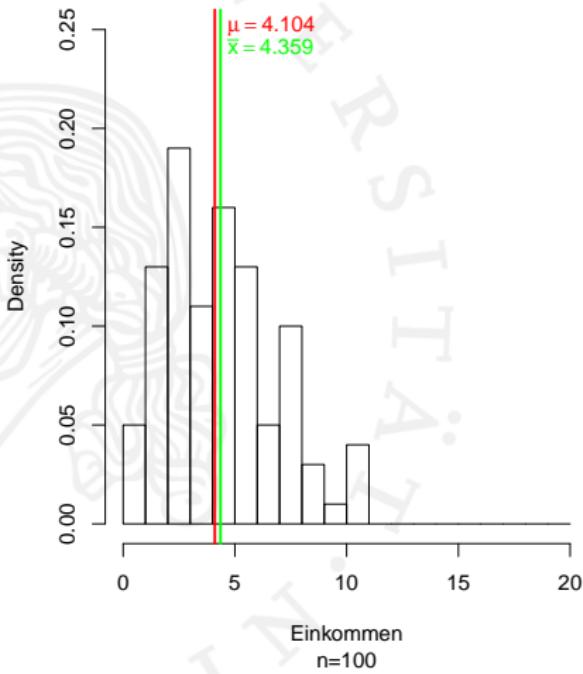
Zentrales Schwankungsintervall für μ



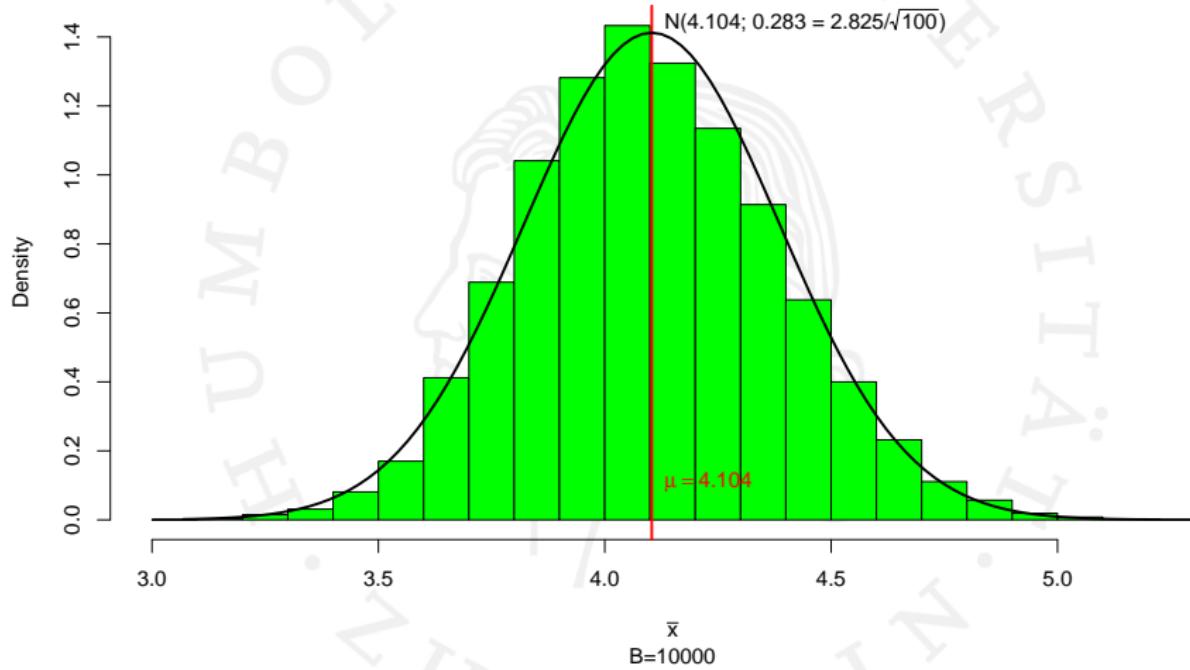
Einkommensverteilung in einer Stichprobe

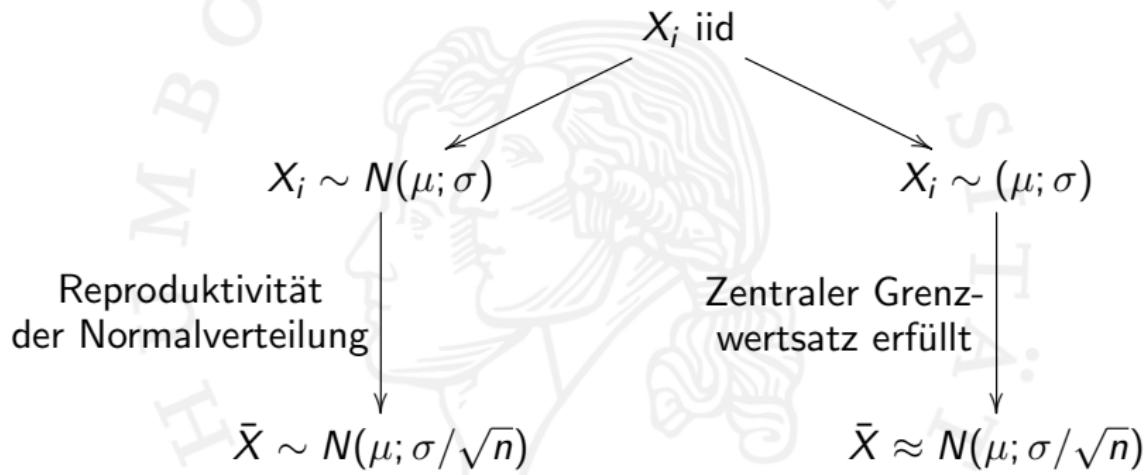


Einkommensverteilung in einer Stichprobe

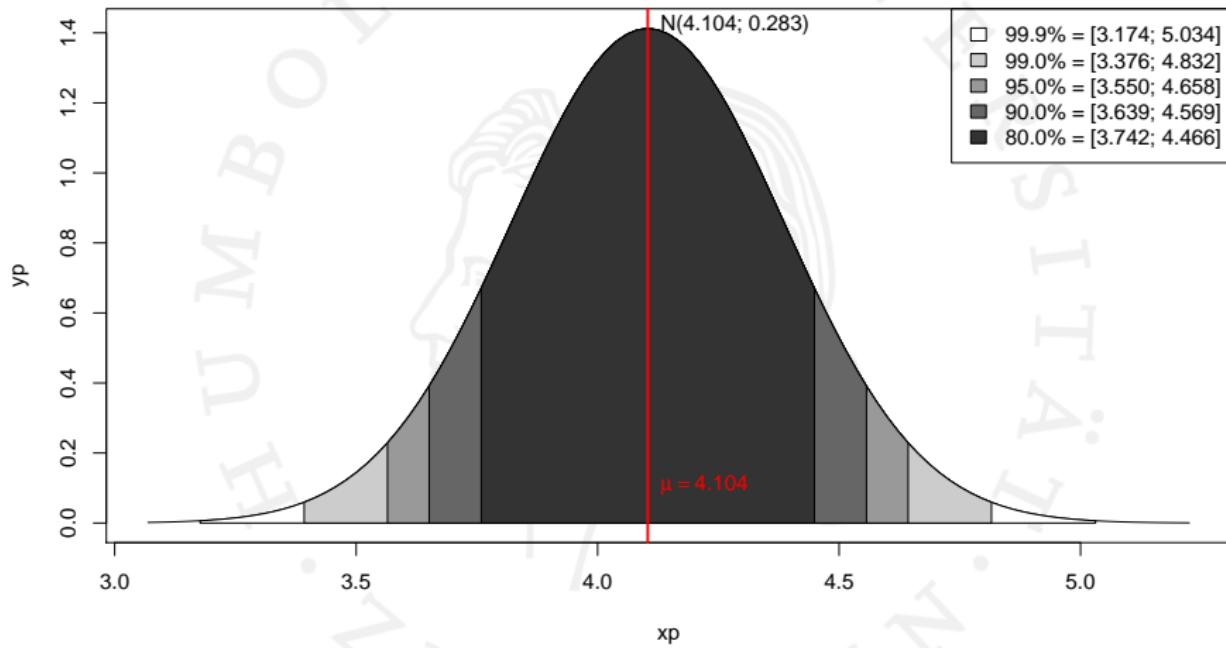


Verteilung der Stichprobenmittelwerte

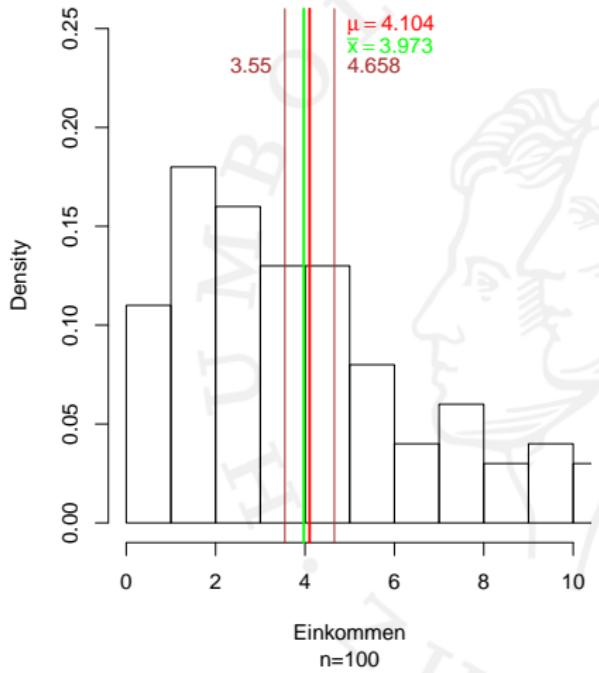




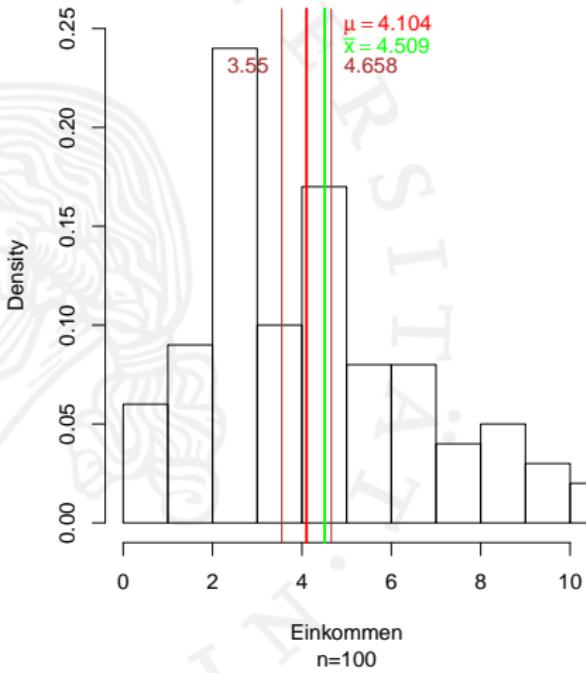
Zentrales Schwankungsintervall



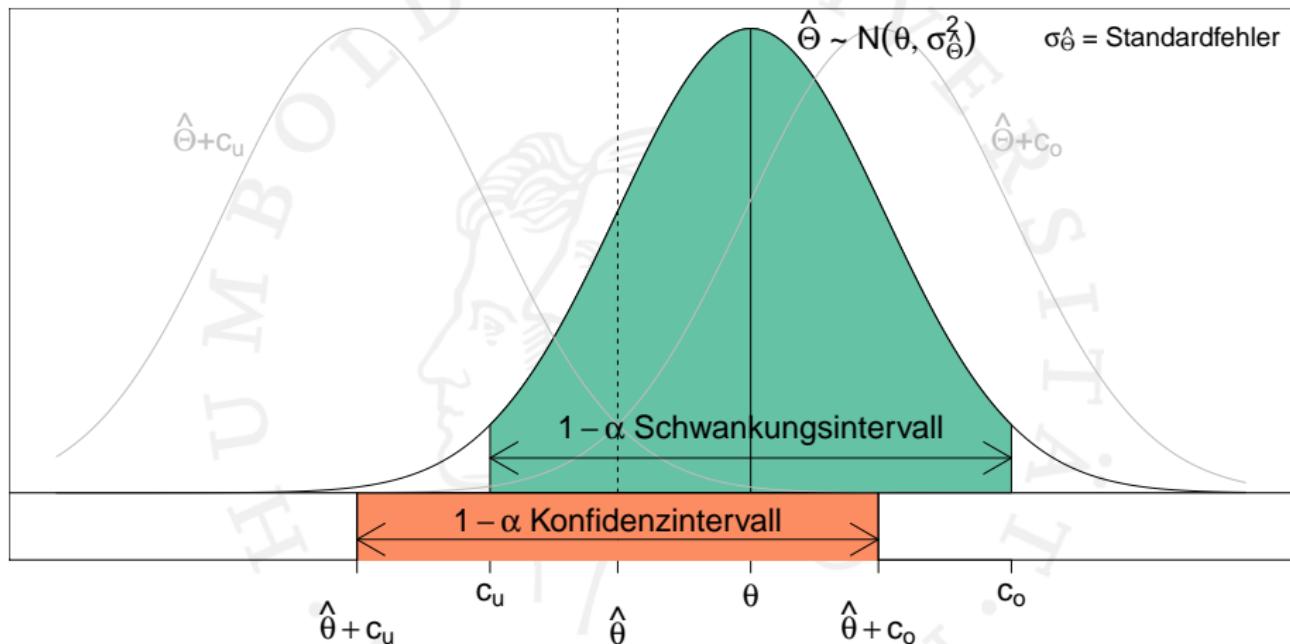
95% Schwankungsintervall



95% Schwankungsintervall



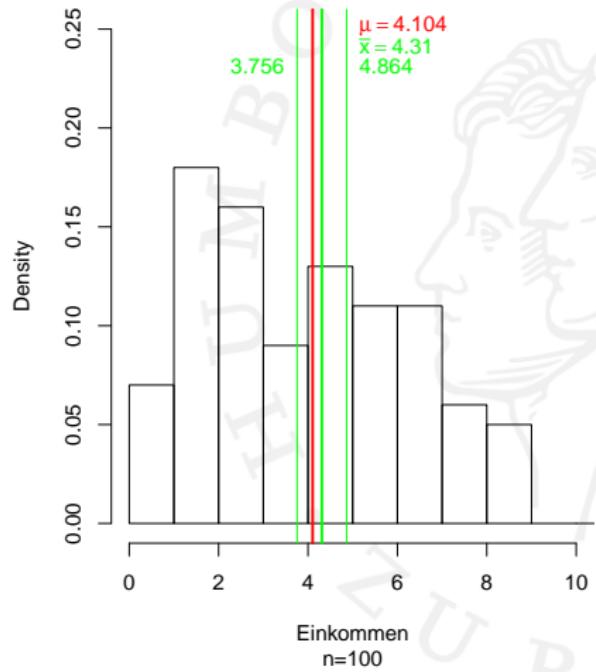
Schwankungsintervall vs. Konfidenzintervall



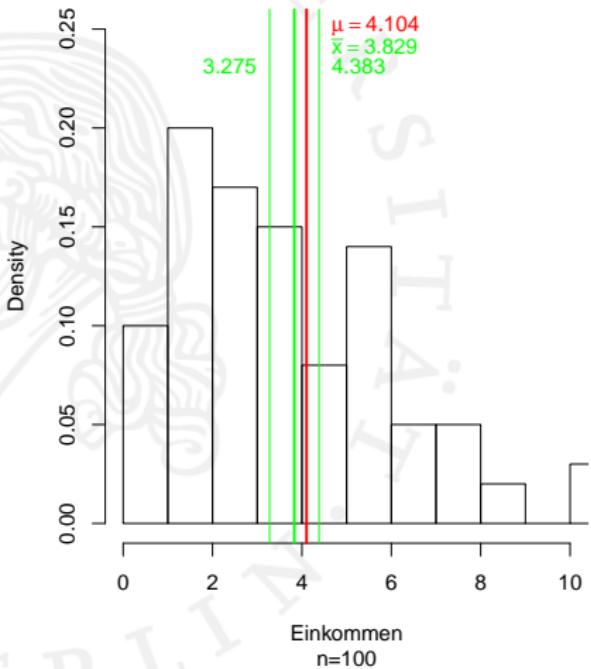
$\hat{\theta}$ liegt im Schwankungsintervall $\Leftrightarrow \theta$ wird vom Konfidenzintervall überdeckt

Konfidenzintervall für μ

95% Konfidenzintervall



95% Konfidenzintervall



Bestimmung eines Konfidenzintervalls

Basiert auf

- einer einfachen Zufallsstichprobe vom Umfang n mit den Stichprobenvariablen X_1, \dots, X_n ,
- der Festlegung zweier Stichprobenfunktionen $V_u = g_1(X_1, \dots, X_n)$ und $V_o = g_2(X_1, \dots, X_n)$ für die untere bzw. obere Intervallgrenze mit $V_u \leq V_o$
- der Wahrscheinlichkeit $P(V_u \leq \vartheta \leq V_o) \geq 1 - \alpha$, wobei die Wahrscheinlichkeit tatsächlich (bzw. approximativ) und ohne Kenntnis des wahren Wertes des Parameters ϑ bestimmbar sein muss

- $[V_u, V_o] \rightarrow$ Konfidenzintervall für den Parameter ϑ zum Konfidenzniveau $1 - \alpha$ (Vertrauenswahrscheinlichkeit)
- je größer α , desto kleiner das Konfidenzintervall
- α wird festgelegt (i.d.R. wählt man $1 - \alpha = 95\%$)

Schwankungs- & Konfidenzintervall

Schwankungsintervall

$$P\left(\vartheta - k \leq \hat{\theta} \leq \vartheta + k\right) = 1 - \alpha$$

Umformung

$$\begin{aligned} \vartheta - k &\leq \hat{\theta} \leq \vartheta + k & | -\hat{\theta}, -\vartheta \\ -\hat{\theta} - k &\leq -\vartheta \leq -\hat{\theta} + k & | \cdot -1 \\ \hat{\theta} + k &\geq \vartheta \geq \hat{\theta} - k \\ \hat{\theta} - k &\leq \vartheta \leq \hat{\theta} + k \end{aligned}$$

Konfidenzintervall

$$P\left(\hat{\theta} - k \leq \vartheta \leq \hat{\theta} + k\right) = 1 - \alpha$$

$$\text{mit } k = c \cdot \sigma(\hat{\theta})$$

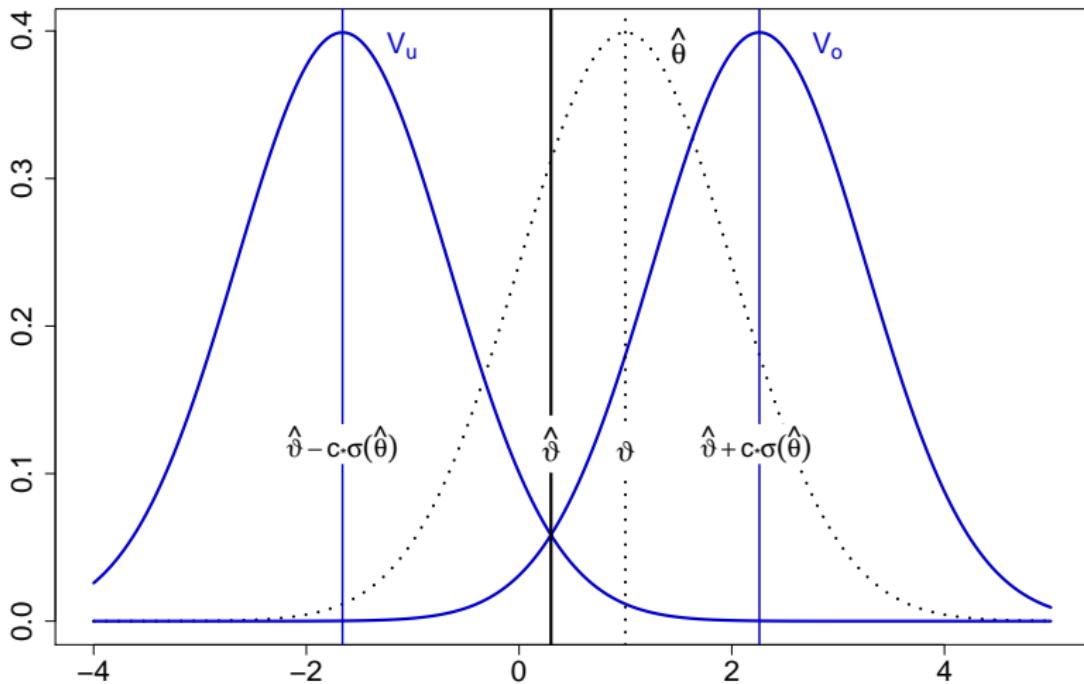
Zentrales Konfidenzintervall

Symmetrisches Konfidenzintervall

- $P(\hat{\theta} - |c_{\alpha/2}| \cdot \sigma(\hat{\theta}) \leq \vartheta \leq \hat{\theta} + |c_{1-\alpha/2}| \cdot \sigma(\hat{\theta})) = 1 - \alpha$
- $[V_u, V_o] = [\hat{\theta} - |c_{\alpha/2}| \cdot \sigma(\hat{\theta}), \hat{\theta} + |c_{1-\alpha/2}| \cdot \sigma(\hat{\theta})]$

Schätzintervall oder realisiertes Konfidenzintervall $[v_u, v_o]$

- ergibt sich durch konkrete Stichprobe mit den Stichprobenwerten x_1, \dots, x_n



Grenzen V_u und V_o hängen ab von:

1. den Stichprobenvariablen X_1, \dots, X_n , da die Schätzfunktion $\hat{\theta}$ eine Funktion dieser Stichprobenvariablen ist
2. der Standardabweichung der Schätzfunktion $\sigma(\hat{\theta})$
3. dem festgelegten Konfidenzniveau $1 - \alpha$ (über $c_{\alpha/2}$ und $c_{1-\alpha/2}$)
4. der Verteilung der Schätzfunktion
5. dem Stichprobenumfang

Vor der Ziehung der Stichprobe:

- Grenzen des Konfidenzintervalls sind Zufallsvariablen
- $1 - \alpha$ ist die Wahrscheinlichkeit, dass das Schätzverfahren zu Intervallen führt, die den wahren Wert des Parameters ϑ der Grundgesamtheit enthalten
- Das Konfidenzniveau $1 - \alpha$ gibt den Anteil aller möglichen Schätzintervalle $[v_u, v_o]$ an, die den unbekannten Wert des Parameters ϑ überdecken.

Nach der Ziehung der Stichprobe:

Grenzen des Schätzintervalls v_u und v_o sind nunmehr feste Größen

$$[v_u, v_o] = [\hat{\vartheta} - |c_{\alpha/2}| \cdot \hat{\sigma}(\hat{\theta}), \hat{\vartheta} + |c_{1-\alpha/2}| \cdot \hat{\sigma}(\hat{\theta})]$$

Konfidenzintervall für μ

- Grundgesamtheit:
 - ▶ Zufallsvariable X mit $E(X) = \mu$
 - ▶ μ unbekannt
- einfache Zufallsstichprobe vom Umfang n :
 - ▶ Stichprobenvariablen: X_1, \dots, X_n , unabhängig und identisch verteilt
- geeignete Schätzfunktion:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

mit

$$Var(\bar{X}) = \sigma^2(\bar{X}) = \frac{\sigma^2}{n} \quad \Leftrightarrow \quad \sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Schwankungsintervall

$$P\left(\underbrace{\mu - c \cdot \sigma(\bar{X}) < \bar{X} < \mu + c \cdot \sigma(\bar{X})}_{\text{Schwankungsintervall}}\right) = 1 - \alpha$$

Umformung

$$\begin{aligned} \mu - c \cdot \sigma(\bar{X}) &< \bar{X} &< \mu + c \cdot \sigma(\bar{X}) & | - \mu \\ -c \cdot \sigma(\bar{X}) &< \bar{X} - \mu &< c \cdot \sigma(\bar{X}) & | - \bar{X} \\ -\bar{X} - c \cdot \sigma(\bar{X}) &< -\mu &< -\bar{X} + c \cdot \sigma(\bar{X}) & | \cdot (-1) \\ \bar{X} + c \cdot \sigma(\bar{X}) &> \mu &> \bar{X} - c \cdot \sigma(\bar{X}) \\ \bar{X} - c \cdot \sigma(\bar{X}) &< \mu &< \bar{X} + c \cdot \sigma(\bar{X}) \end{aligned}$$

Konfidenzintervall

$$P\left(\overbrace{\bar{X} - c \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + c \cdot \sigma(\bar{X})}^{\text{Konfidenzintervall}}\right) = 1 - \alpha$$

Konfidenzintervall

$$[V_u, V_o] = \left[\bar{X} - c \cdot \frac{\sigma}{\sqrt{n}}; \bar{X} + c \cdot \frac{\sigma}{\sqrt{n}} \right]$$

Probleme:

- Verteilung des Stichprobenmittelwertes
- Information über Varianz der Grundgesamtheit σ^2

Voraussetzung:

- Varianz σ^2 bekannt
- Grundgesamtheit normalverteilt

$$X \sim N(\mu; \sigma) \Rightarrow X_i \sim N(\mu; \sigma) \quad \forall i$$

$$\Rightarrow \bar{X} \sim N\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$$

$$\Rightarrow Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0; 1)$$

$$\Rightarrow c = z_{1-\alpha/2} \text{ aus } N(0; 1)$$

Konfidenzintervall:

- $P\left(\bar{X} - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$
- $[V_u, V_o] = \left[\bar{X} - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right]$

Voraussetzung:

- Varianz σ^2 bekannt
- Verteilung der Grundgesamtheit unbekannt

→ keine direkte Aussage über die Verteilung des Stichprobenmittelwertes \bar{X} möglich, aber wenn die Bedingungen des zentralen Grenzwertssatzes erfüllt sind:

$$\bar{X} \approx N(\mu; \sigma/\sqrt{n}) \implies Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx N(0, 1)$$

approximatives Konfidenzintervall für $n > 30$:

- $P\left(\bar{X} - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$
- $[V_u, V_o] = \left[\bar{X} - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right]$

Voraussetzung:

- Varianz σ^2 unbekannt
- Grundgesamtheit normalverteilt

$$X \sim N(\mu; \sigma) \Rightarrow X_i \sim N(\mu; \sigma) \quad \forall i$$

$$\Rightarrow \bar{X} \sim N\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$$

$$\Rightarrow Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0; 1)$$

Schätzfunktion für unbekannte Varianz σ^2 ist die Stichprobenvarianz:

$$\begin{aligned} S^2 &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \\ \Rightarrow T &= \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \quad \text{Herleitung siehe Anhang} \\ \Rightarrow c &= t_{1-\alpha/2; n-1} \end{aligned}$$

Konfidenzintervall

- $P\left(\bar{X} - t_{1-\frac{\alpha}{2}; n-1} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{1-\frac{\alpha}{2}; n-1} \frac{S}{\sqrt{n}}\right) = 1 - \alpha$
- $[V_u, V_o] = \left[\bar{X} - t_{1-\frac{\alpha}{2}; n-1} \cdot \frac{S}{\sqrt{n}}, \bar{X} + t_{1-\frac{\alpha}{2}; n-1} \cdot \frac{S}{\sqrt{n}}\right]$

Beispiel 18.1

Grundgesamtheit

- X - Haushaltsnettoeinkommen
- $X \sim N(\mu; \sigma)$
- Varianz bekannt: $\sigma^2 = 1025770,43$ ($\sigma = 1012,8$)

einfache Zufallsstichprobe vom Umfang $n = 20$

Punktschätzung:

$$\bar{x} = 2405,25$$

Intervallschätzung:

- Konfidenzniveau: $1 - \alpha = 0,95$
- $z_{0,975} = 1,96$ aus $N(0; 1)$
- Konfidenzintervall:

$$\left[\bar{X} - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right]$$

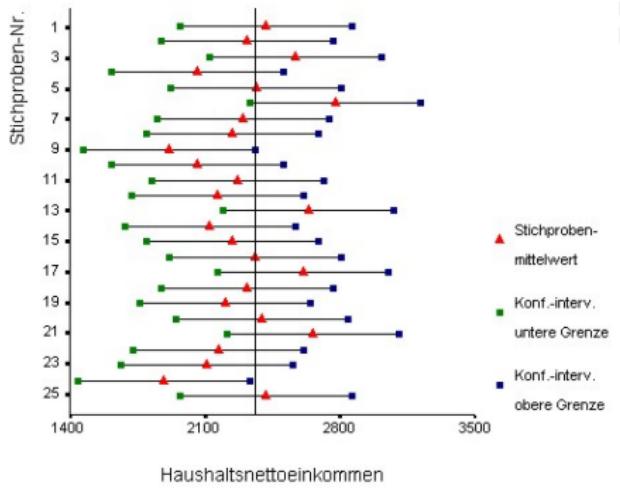
- Schätzintervall:

$$2405,25 - 1,96 \cdot \frac{1012,8}{\sqrt{20}} = 1961,37$$

$$2405,25 + 1,96 \cdot \frac{1012,8}{\sqrt{20}} = 2849,13$$

$$[1961,37; 2849,13]$$

Schätzintervalle von 25 einfache Zufallsstichproben vom Umfang $n = 20$



$$\left[\bar{X} - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right]$$

- alle Schätzintervalle sind gleich lang (887,76)
- nur 1 Schätzintervall überdeckt μ nicht

Beispiel 18.2

Grundgesamtheit:

- X - Haushaltsnettoeinkommen
- Verteilung von X unbekannt
- Varianz bekannt: $\sigma^2 = 1025770,43$ ($\sigma = 1012,8$)

einfache Zufallsstichprobe vom Umfang $n = 100$

Punktschätzung:

$$\bar{x} = 2454,16$$

Intervallschätzung:

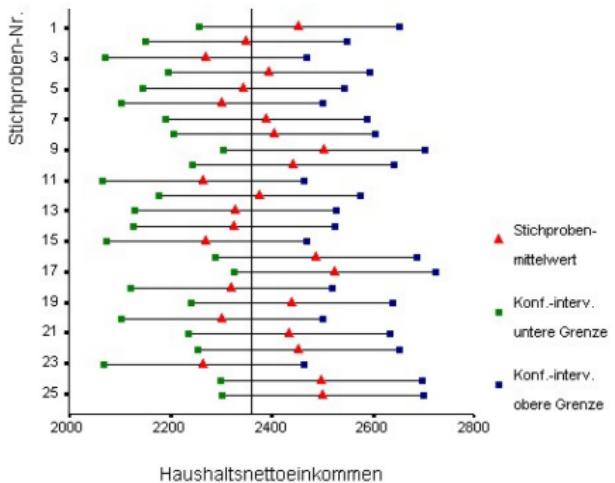
- Konfidenzniveau: $1 - \alpha = 0,95$
- $z_{0,975} = 1,96$ aus $N(0; 1)$
- approximatives Schätzintervall:

$$2454,16 - 1,96 \cdot \frac{1012,8}{\sqrt{100}} = 2255,65$$

$$2454,16 + 1,96 \cdot \frac{1012,8}{\sqrt{100}} = 2652,67$$

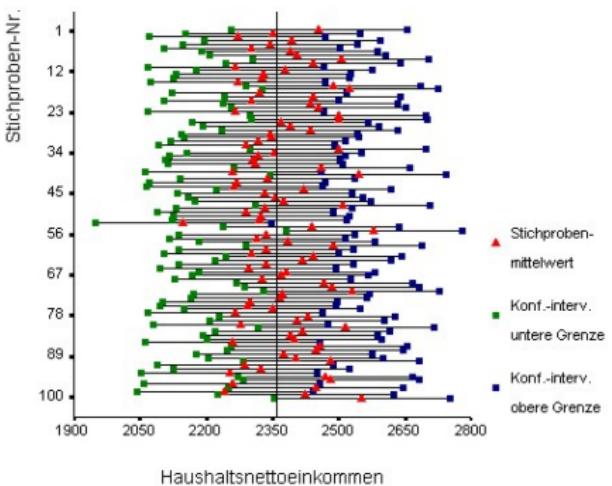
$$[2255,65; 2652,67]$$

Schätzintervalle von 25 einfache Zufallsstichproben vom Umfang $n = 100$



- alle Schätzintervalle sind gleich lang (397,02)
- Länge der Schätzintervalle ist kleiner als bei $n = 20$

Schätzintervalle von 100 einfache Zufallsstichproben vom Umfang $n = 100$



- alle Schätzintervalle sind gleich lang (397,02)
- nur 2 Schätzintervalle überdecken μ nicht

Beispiel 18.3

Grundgesamtheit:

- X - Haushaltsnettoeinkommen
- $X \sim N(\mu; \sigma)$
- Varianz σ^2 unbekannt

einfache Zufallsstichprobe vom Umfang $n = 20$

$t_{0,975} = 2,093$ aus t -Verteilung

1. aus Stichprobe: $\bar{x} = 2413,4$ $s = 1032,15$

► Schätzintervall:

$$2413,4 - 2,093 \cdot \frac{1032,15}{\sqrt{20}} = 2413,4 - 483,06 = 1930,34$$

$$2413,4 + 2,093 \cdot \frac{1032,15}{\sqrt{20}} = 2413,4 + 483,06 = 2896,46$$

$$[1930,34; 2896,46]$$

2. aus Stichprobe: $\bar{x} = 2317$ $s = 872,32$

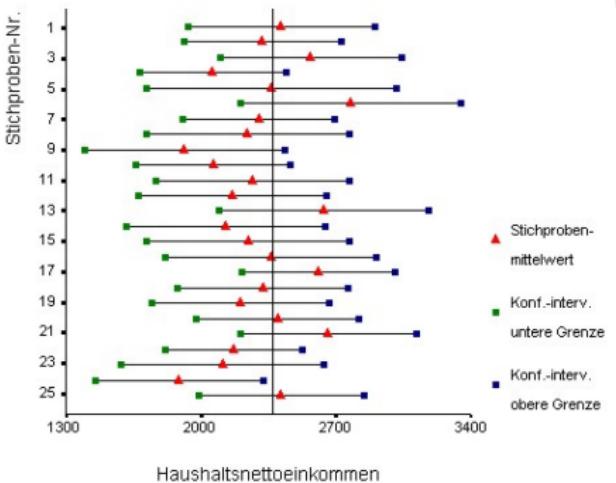
► Schätzintervall:

$$2317 - 2,093 \cdot \frac{872,32}{\sqrt{20}} = 2317 - 408,25 = 1908,75$$

$$2317 + 2,093 \cdot \frac{872,32}{\sqrt{20}} = 2317 + 408,25 = 2725,25$$

$$[1908,75; 2725,25]$$

Schätzintervalle für 25 einfache Zufallsstichproben vom Umfang $n = 20$



- Schätzintervalle sind nicht gleich lang
- nur 1 Schätzintervall überdeckt μ nicht

Konfidenzintervall für π

- dichotome Grundgesamtheit: $P(A) = \pi \quad P(\bar{A}) = 1 - \pi$
- einfache Zufallsstichprobe vom Umfang n
- X_i : {Anzahl der Erfolge beim i -ten Versuch}
 - ▶ $x_i = 0$ oder 1
 - ▶ $X_i \sim B(1; \pi)$ Bernoulli-verteilt
- X : {Anzahl der Erfolge bei einf. Zufallsstichprobe vom Umfang n }
 - ▶ $X = \sum_i X_i$
 - ▶ $X \sim B(n; \pi)$
 - ▶ $E(X) = n\pi, \quad Var(X) = n\pi(1 - \pi)$
- falls $n\pi \geq 5$ und $n(1 - \pi) \geq 5$; $n > 30$

$$X \approx N\left(n\pi; \sqrt{n\pi(1 - \pi)}\right)$$

- $\hat{\pi}$: {Anteil der Erfolge bei einfacher Zufallsstichprobe vom Umfang n }
- $\hat{\pi} = X/n$
 - ▶ $\hat{\pi} \approx N\left(\pi; \sqrt{\frac{\pi(1-\pi)}{n}}\right)$
 - ▶ $E(\hat{\pi}) = E(X/n) = \pi, \quad Var(\hat{\pi}) = Var(X/n) = \pi(1-\pi)/n$
- $Z = \frac{\hat{\pi} - \pi}{\sigma(\hat{\pi})} \approx N(0; 1) \Rightarrow c = z_{1-\alpha/2}$ aus $N(0; 1)$
- approximatives Konfidenzintervall

$$P\left(\hat{\pi} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\pi(1-\pi)}{n}} \leq \pi \leq \hat{\pi} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\pi(1-\pi)}{n}}\right) = 1 - \alpha$$

$$\left[\hat{\pi} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\pi(1-\pi)}{n}}, \hat{\pi} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\pi(1-\pi)}{n}}\right]$$

Problem: In $\text{Var}(\hat{\pi}) = \text{Var}(X/n) = \pi(1 - \pi)$ ist noch das unbekannte π enthalten

Lösung: Aufgrund von

$$\lim_{n \rightarrow \infty} E[\hat{\pi}(1 - \hat{\pi})] = \pi(1 - \pi)$$

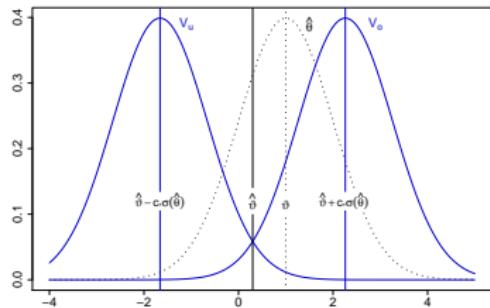
kann auch in $\text{Var}(\hat{\pi})$ π durch $\hat{\pi}$ ersetzt werden

- approximatives Konfidenzintervall

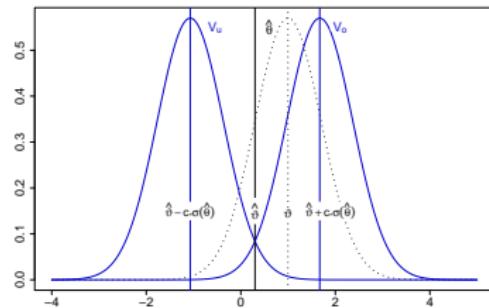
$$P\left(\hat{\pi} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \leq \pi \leq \hat{\pi} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right) \approx 1 - \alpha$$

$$\left[\hat{\pi} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}, \hat{\pi} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \right]$$

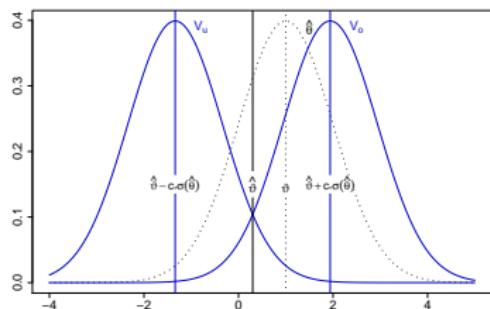
Einfluß von α und n



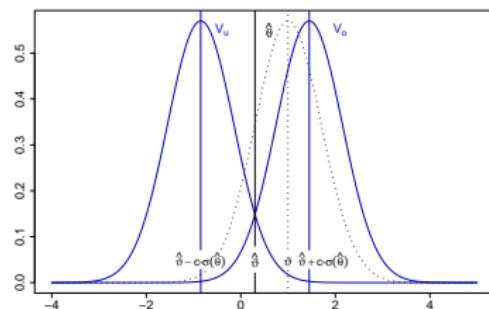
$\alpha = 0, 05; n = 100$



$\alpha = 0, 05; n = 1000$



$\alpha = 0, 1; n = 100$



$\alpha = 0, 1; n = 1000$

Bestimmung von n

- Länge des Konfidenzintervalls: $V_o - V_u$
- Schätzfehler e : halbe Länge des Intervalls
- Länge und Schätzfehler hängen i.A. ab von:
 - ▶ Konfidenzniveau $1 - \alpha$
 - ▶ Stichprobenumfang n
- Vorgabe von Konfidenzniveau und Länge des Konfidenzintervalls
⇒ Wie groß muss der Stichprobenumfang gewählt werden, um diese Vorgaben zu erfüllen?

Konfidenzintervall für μ

- normalverteilte Grundgesamtheit
- bekannte Varianz σ^2 der Grundgesamtheit

$$\begin{aligned}\ell = 2 \cdot e &= \left(\bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right) - \left(\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right) \\ &= 2 z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\end{aligned}$$

Länge hängt ab von

- Konfidenzniveau $1 - \alpha$ (wird vorgegeben)
- Stichprobenumfang n (ist zu bestimmen)

$$n \geq \frac{4 \cdot \sigma^2 \cdot z_{1-\frac{\alpha}{2}}^2}{\ell^2} = \frac{\sigma^2 \cdot z_{1-\frac{\alpha}{2}}^2}{e^2}$$

Beispiel 18.4

Grundgesamtheit:

- X - Haushaltsnettoeinkommen
- $X \sim N(\mu, \sigma)$
- Varianz bekannt: $\sigma^2 = 1025770,43$ ($\sigma = 1012,8$)

Frage:

Wie groß muss der Stichprobenumfang n sein, um bei einem Konfidenzniveau von 95%

1. das Einkommen auf ± 100 € genau zu schätzen?
2. das Einkommen auf ± 10 € genau zu schätzen?

Lösung zu 1.:

$$n \geq \frac{4 \cdot \sigma^2 \cdot z_{1-\frac{\alpha}{2}}^2}{\ell^2} = \frac{\sigma^2 \cdot z_{1-\frac{\alpha}{2}}^2}{e^2}$$

- $\sigma^2 = 1025770,43$
- $\ell = 200$
- $z_{1-\frac{\alpha}{2}} = 1,96$

$$n \geq \frac{4 \cdot 1025770,43 \cdot 1,96^2}{200^2} = \frac{1025770,43 \cdot 1,96^2}{100^2} = 394,06$$

Es müssen mindestens 395 Haushalte befragt werden.

Lösung zu 2.:

$$n \geq \frac{4 \cdot \sigma^2 \cdot z_{1-\frac{\alpha}{2}}^2}{\ell^2} = \frac{\sigma^2 \cdot z_{1-\frac{\alpha}{2}}^2}{e^2}$$

- $\sigma^2 = 1025770,43$
- $\ell = 20$
- $z_{1-\frac{\alpha}{2}} = 1,96$

$$n \geq \frac{4 \cdot 1025770,43 \cdot 1,96^2}{20^2} = \frac{1025770,43 \cdot 1,96^2}{10^2} = 39405,99$$

Es müssen mindestens 39406 Haushalte befragt werden.

Konfidenzintervall für π

- bei Approximation durch die Normalverteilung

$$\begin{aligned}\ell &= \left(\hat{\pi} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\pi(1-\pi)}{n}} \right) - \left(\hat{\pi} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\pi(1-\pi)}{n}} \right) \\ &= 2 \cdot z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\pi(1-\pi)}{n}} \\ n &\geq \frac{4 \cdot z^2 \cdot \pi \cdot (1-\pi)}{\ell^2}\end{aligned}$$

- π jedoch unbekannt \Rightarrow einfache Abschätzung

$$\sqrt{\pi(1-\pi)} \leq \sqrt{\frac{1}{2} \cdot \frac{1}{2}} = \frac{1}{2}$$

$$n \geq \frac{4 \cdot z^2 \cdot \frac{1}{2} \cdot \frac{1}{2}}{\ell^2} = \frac{z^2}{\ell^2} = \frac{z^2}{4 \cdot e^2}$$

Konfidenzintervalle für β_i

Für β_1 :

- $U_i \sim N(0, \sigma_U^2) \Rightarrow B_1 \sim N(\beta_1; \sigma_{B_1}^2)$
- $T = \frac{B_1 - \beta_1}{\hat{\sigma}_{B_1}}$ ist t-verteilt mit $f = n - 2$ Freiheitsgraden.

$$P\left(-t_{1-\alpha/2; n-2} \leq \frac{B_1 - \beta_1}{\hat{\sigma}_{B_1}} \leq t_{1-\alpha/2; n-2}\right) = 1 - \alpha$$

- Konfidenzintervall:

$$P(B_1 - t_{1-\alpha/2; n-2} \cdot \hat{\sigma}_{B_1} \leq \beta_1 \leq B_1 + t_{1-\alpha/2; n-2} \cdot \hat{\sigma}_{B_1}) = 1 - \alpha$$

- Schätzintervall:

$$[b_1 - t_{1-\alpha/2; n-2} \cdot \hat{\sigma}_{B_1}; b_1 + t_{1-\alpha/2; n-2} \cdot \hat{\sigma}_{B_1}]$$

Für β_0 :

- $U_i \sim N(0, \sigma_U^2) \Rightarrow B_0 \sim N(\beta_0; \sigma_{B_1}^2)$
- $T = \frac{B_0 - \beta_0}{\hat{\sigma}_{B_0}}$ ist t-verteilt mit $f = n - 2$ Freiheitsgraden.

$$P\left(-t_{1-\alpha/2; n-2} \leq \frac{B_0 - \beta_0}{\hat{\sigma}_{B_0}} \leq t_{1-\alpha/2; n-2}\right) = 1 - \alpha$$

- Konfidenzintervall:

$$P(B_0 - t_{1-\alpha/2; n-2} \cdot \hat{\sigma}_{B_0} \leq \beta_0 \leq B_0 + t_{1-\alpha/2; n-2} \cdot \hat{\sigma}_{B_0}) = 1 - \alpha$$

- Schätzintervall:

$$[b_0 - t_{1-\alpha/2; n-2} \cdot \hat{\sigma}_{B_0}; b_0 + t_{1-\alpha/2; n-2} \cdot \hat{\sigma}_{B_0}]$$

Fortsetzung des Beispiels Eiscreme

- $\hat{y}_i = b_0 + b_1 x_i = 144,93 + 2,65 x_i$

$$n = 30$$

$$\hat{\sigma}_{B_1} = 0,4332; \hat{\sigma}_{B_0} = 5,318$$

- Konfidenzniveau: $1 - \alpha = 0,95 \quad t_{0,975;28} = 2,048$
- Schätzintervalle:

$$\triangleright [b_1 - t_{1-\alpha/2;n-2} \cdot \hat{\sigma}_{B_1}; b_1 + t_{1-\alpha/2;n-2} \cdot \hat{\sigma}_{B_1}]$$

$$[2,65 - 2,048 \cdot 0,4332; 2,65 + 2,048 \cdot 0,4332]$$

$$= [1,8130; 3,480]$$

$$\triangleright [b_0 - t_{1-\alpha/2;n-2} \cdot \hat{\sigma}_{B_0}; b_0 + t_{1-\alpha/2;n-2} \cdot \hat{\sigma}_{B_0}]$$

$$[144,93 - 2,048 \cdot 5,318; 144,93 + 2,048 \cdot 5,318, 13]$$

$$= [134,04; 155,82]$$

Fortsetzung des Beispiels Monatsmiete

- $\hat{y}_i = b_0 + b_1 x_i = -106,188 + 19,223 x_i$

$$n = 815$$

$$\hat{\sigma}_{B_1} = 0,4332; \hat{\sigma}_{B_0} = 35,543$$

- Konfidenzniveau: $1 - \alpha = 0,95$ $t_{0,975;813} \approx z_{0,975} = 1,96$
- Schätzintervalle:

- ▶
$$\begin{aligned} & [b_1 - t_{1-\alpha/2;n-2} \cdot \hat{\sigma}_{B_1}; b_1 + t_{1-\alpha/2;n-2} \cdot \hat{\sigma}_{B_1}] \\ & [19,223 - 1,96 \cdot 0,4332; 19,223 + 1,96 \cdot 0,4332] \\ & = [18,374; 20,072] \end{aligned}$$
- ▶
$$\begin{aligned} & [b_0 - t_{1-\alpha/2;n-2} \cdot \hat{\sigma}_{B_0}; b_0 + t_{1-\alpha/2;n-2} \cdot \hat{\sigma}_{B_0}] \\ & [-106,188 - 1,96 \cdot 35,543; -106,188 + 1,96 \cdot 35,543] \\ & = [-175,85; -36,52] \end{aligned}$$

Konfidenzintervall für $E(Y|x_h)$

- Bedingter Erwartungswert in der Grundgesamtheit:

$$E(Y|x_i) = E(Y_i) = \beta_0 + \beta_1 x_i$$

- Stichprobenregressionsfunktion:

$$\hat{Y}_h = B_0 + B_1 x_h$$

$$B_0 \sim N(\beta_0; \sigma_{B_0}^2) \text{ und } B_1 \sim N(\beta_1; \sigma_{B_1}^2) \Rightarrow \hat{Y}_h \sim N(\bullet; \bullet)$$

$$E(\hat{Y}_h) = \beta_0 + \beta_1 x_h$$

$$Var(\hat{Y}_h) = \sigma^2(\hat{Y}_h) = \sigma_u^2 \left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)$$

- Geschätzte Varianz:

$$\hat{\sigma}^2(\hat{Y}_h) = s_u^2 \left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum(x_i - \bar{x})^2} \right)$$

- $T = \frac{\hat{Y}_h - E(\hat{Y}_h)}{\hat{\sigma}(\hat{Y}_h)}$ ist t-verteilt mit $f = n - 2$ Freiheitsgraden
- Konfidenzintervall für $E(\hat{Y}_h)$ zum vorgegebenen Konfidenzniveau $1 - \alpha$:

$$P \left(\hat{Y}_h - t_{1-\alpha/2; n-2} \cdot \hat{\sigma}(\hat{Y}_h); \hat{Y}_h + t_{1-\alpha/2; n-2} \cdot \hat{\sigma}(\hat{Y}_h) \right) = 1 - \alpha$$

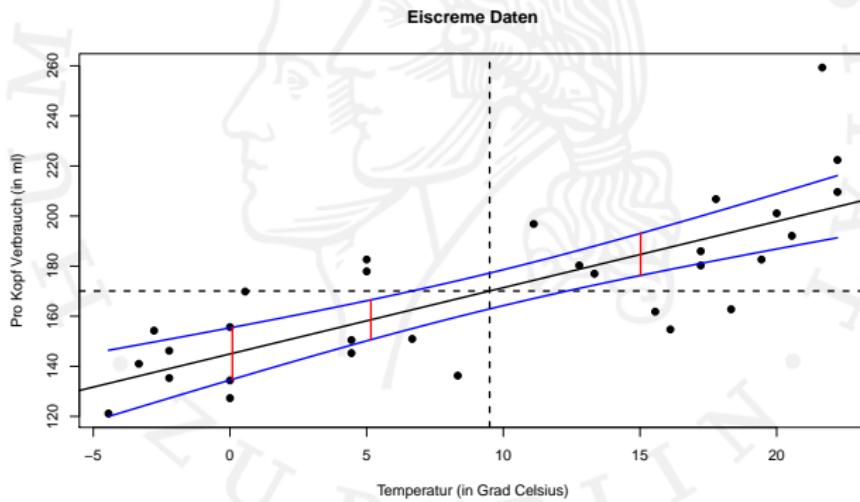
- Schätzintervall für $E(\hat{Y}_h)$

$$\left[b_0 + b_1 x_0 \pm t_{1-\alpha/2; n-2} \cdot s_u \sqrt{\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum(x_i - \bar{x})^2}} \right]$$

Fortsetzung des Beispiels Eiscreme

- $n = 30$
- $\hat{y}_i = b_0 + b_1 x_i = 144, 93 + 2, 65 x_i$
- Konfidenzniveau: $1 - \alpha = 0, 95$

Konfidenzintervalle (rot) für beliebiges x_h :



Häufige Missverständnisse

- Ein Konfidenzniveau von $1 - \alpha$ bedeutet nicht, dass $1 - \alpha$ der Stichprobendaten innerhalb des Konfidenzintervalls liegen.
- Ein Konfidenzniveau von $1 - \alpha$ bedeutet nicht, dass für ein bestimmtes Schätzintervall eine Wahrscheinlichkeit von $1 - \alpha$ besteht, dass der wahre Parameter innerhalb des Intervalls liegt.
 - ▶ Sobald ein Schätzintervall berechnet wurde, überdeckt dieses Intervall entweder den wahren Parameter oder nicht; es handelt sich nicht mehr um eine Frage der Wahrscheinlichkeit. Der wahre Parameter ist nicht zufällig!
 - ▶ Das Konfidenzniveau bezieht sich auf die Zuverlässigkeit des Schätzverfahrens, nicht auf ein bestimmtes Schätzintervall: den Anteil der Schätzintervalle, die den wahren Parameter überdecken.
- Ein Schätzintervall ist kein endgültiger Bereich plausibler Werte für den Stichprobenparameter, obwohl es heuristisch oft als Bereich plausibler Werte angesehen wird.
 - ▶ Die Länge des Schätzintervalles liefert eine Information darüber wie genau wir den wahren Parameter eingrenzen können.

Übersicht Konfidenzintervalle

Konfidenzintervall allgemein	Intervallgrenzen untere obere	
	$P(\hat{\Theta} - c \cdot \sigma(\hat{\Theta}) \leq \vartheta \leq \hat{\Theta} + c \cdot \sigma(\hat{\Theta}))$	$= 1 - \alpha$
für $\hat{\Theta} = \bar{X}$ allg.	$P(\bar{X} - c \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + c \frac{\sigma}{\sqrt{n}})$	$= 1 - \alpha$
$X \sim N(\mu, \sigma)$ σ bek.	$P(\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}})$ $z_{1-\frac{\alpha}{2}}$ aus Tab. 12.17	$= 1 - \alpha$
$X \sim N(\mu, \sigma)$ σ unbek.	$P(\bar{X} - t_{1-\frac{\alpha}{2};n-1} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{1-\frac{\alpha}{2};n-1} \frac{s}{\sqrt{n}})$ $t_{1-\frac{\alpha}{2};n-1}$ aus Tab. 11.17	$= 1 - \alpha$
Vert. von X unbek. σ bek.; $n > 30$	$P(\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}})$	$= 1 - \alpha$
$n\hat{\pi} > 5$ $n(1 - \hat{\pi}) > 5$ $n > 30$	$P\left(\hat{\pi} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \leq \pi \leq \hat{\pi} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right)$	$= 1 - \alpha$

Anhang

Beweis der Behauptung von Folie 600:

$$\text{Var}(\hat{\theta}) = E[(\hat{\theta} - \vartheta)^2] - [E(\hat{\theta}) - \vartheta]^2$$

Beweis:

$$\begin{aligned}
 E[\{\hat{\theta} - E(\hat{\theta})\}^2] &= E[\{\hat{\theta} - \vartheta + \vartheta - E(\hat{\theta})\}^2] \\
 &= E[\{(\hat{\theta} - \vartheta) - (E(\hat{\theta}) - \vartheta)\}^2] \\
 &= E[(\hat{\theta} - \vartheta)^2 - 2(\hat{\theta} - \vartheta)(E(\hat{\theta}) - \vartheta) + (E(\hat{\theta}) - \vartheta)^2] \\
 &= E[(\hat{\theta} - \vartheta)^2] - 2E(\hat{\theta} - \vartheta)(E(\hat{\theta}) - \vartheta) + (E(\hat{\theta}) - \vartheta)^2 \\
 &= E[(\hat{\theta} - \vartheta)^2] - 2(E(\hat{\theta}) - \vartheta)(E(\hat{\theta}) - \vartheta) + (E(\hat{\theta}) - \vartheta)^2 \\
 &= E[(\hat{\theta} - \vartheta)^2] - 2(E(\hat{\theta}) - \vartheta)^2 + (E(\hat{\theta}) - \vartheta)^2 \\
 &= E[(\hat{\theta} - \vartheta)^2] - (E(\hat{\theta}) - \vartheta)^2 \\
 &= \text{MSE} - (E(\hat{\theta}) - \vartheta)^2
 \end{aligned}$$

Herleitung t-Verteilung:

Welche Verteilung weist T auf?

$$T = \frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{\frac{S}{\sigma}} = \frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{S^2}{\sigma^2}}} = \frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{\left(\frac{(n-1)S^2}{\sigma^2}\right)}{(n-1)}}} = \frac{Z}{\sqrt{\frac{Y}{n-1}}}$$

$$Z \sim N(0, 1)$$

$$Y \sim \chi_{n-1}^2$$

$$\Rightarrow T \sim t \quad (f = n - 1)$$

Statistische Testverfahren

5. November 2022

Grundbegriffe • Testentscheidung • Kritische Werte • Fehler 2. Art •
Testentscheidung und Interpretation • Allgemeiner Ablauf eines Tests •
Allgemeine Handlungsweise • Testverfahren

Grundbegriffe

Statistische Tests

Verfahren zur Überprüfung von Annahmen (Hypothesen) über

- die unbekannte Verteilung oder
- unbekannte Parameter (Parametertest)

in der Grundgesamtheit unter Verwendung der Ergebnisse einer Zufallsstichprobe

Voraussetzungen

Gegeben

- eine Grundgesamtheit mit einer oder mehreren Variablen
- eine Zufallsstichprobe mit Stichprobenvariablen X_i (bzw. X_{ij}), die
 - ▶ unabhängig und
 - ▶ identisch verteilt sind

mit der Verteilung $F(x)$ (bzw. $F_j(x)$).

Parametertest

Verfahren zur Überprüfung von Annahmen (Hypothesen) über den unbekannten Parameter ϑ der Grundgesamtheit

ϑ	Parameter
μ	Mittelwert
σ^2	Varianz
π	Anteilswert

Nullhypothese H_0 und Alternativhypothese H_1

- Relation zwischen dem wahren Parameterwert ϑ und einem hypothetischen vorgegebenen Wert ϑ_0
- stets ein disjunktes Hypothesenpaar
- erfasst alle zulässigen Werte des Parameters ϑ
- nach der Testdurchführung ist nur eine Hypothese gültig

Nullhypothese H_0

- die statistische Formulierung der zu prüfenden Annahme
- stets die Nullhypothese geprüft
- das Gleichheitszeichen immer in der Nullhypothese

Alternativhypothese H_1

- die der Nullhypothese entgegengestellte Hypothese

Test	Nullhypothese	Alternativhypothese
zweiseitig	$H_0 : \vartheta = \vartheta_0$	$H_1 : \vartheta \neq \vartheta_0$
einseitig		
rechtsseitig	$H_0 : \vartheta \leq \vartheta_0$	$H_1 : \vartheta > \vartheta_0$
linksseitig	$H_0 : \vartheta \geq \vartheta_0$	$H_1 : \vartheta < \vartheta_0$

Teststatistik V

- für die Überprüfung der Nullhypothese
- StichprobenvARIABLEN X_1, \dots, X_n
 - ▶ Verteilung von X_i enthält den Parameter ϑ
- Stichprobenfunktion $V = V(X_1, \dots, X_n)$
- Zufallsvariable mit einer Verteilung $F(v)$
 - ▶ Verteilung von V unter der Annahme der Gültigkeit der Nullhypothese muss (zumindest approximativ) bekannt sein

Prüfwert v

- Stichprobenwerte x_1, \dots, x_n liegen konkret vor
- Einsetzen der Stichprobenwerte in die Teststatistik V führt zu einer Realisation $v = v(x_1, \dots, x_n)$ (Prüfwert)
- Reelle Zahl

Beispiel 19.1

Zu prüfen: Das mittlere Nettoeinkommen in Deutschland 2011 beträgt 3000 EUR

Variable:

Nettoeinkommen in Deutschland 2011

Unbekannter Parameter:

mittleres Nettoeinkommen μ

Hypothetischer Wert:

$\mu_0 = 3000$ EUR

Nullhypothese H_0 :

$\mu = \mu_0 = 3000$ EUR

Alternativhypothese H_1 :

$\mu \neq \mu_0 = 3000$ EUR

Teststatistik:

$$V = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

Prüfwert:

$$v = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Testentscheidung

Wie wahrscheinlich ist der Prüfwert v wenn H_0 wahr wäre?

- Geringe Wahrscheinlichkeit: Ablehnung von H_0
- Hohe Wahrscheinlichkeit: Nichtablehnung von H_0
- Möglichkeit der Fehlentscheidung!

Notation

- " H_1 " : Ablehnung von H_0 im Ergebnis der Testdurchführung
- " H_0 " : Nichtablehnung von H_0 im Ergebnis der Testdurchführung
- " $H_1|H_0$ " : Ablehnung von H_0 , obwohl H_0 wahr ist
- " $H_0|H_1$ " : Nichtablehnung von H_0 , obwohl H_1 wahr ist

Testentscheidung	Wahrer Zustand	
H_0 wird	H_0 trifft zu	H_0 trifft nicht zu
nicht abgelehnt “ H_0 ”	Richtige Entscheidung “ H_0 ” H_0 $P(\text{“}H_0\text{”} H_0) = 1 - \alpha$	Fehler 2. Art “ H_0 ” H_1 $P(\text{“}H_0\text{”} H_1) = \beta$
abgelehnt “ H_1 ”	Fehler 1. Art “ H_1 ” H_0 $P(\text{“}H_1\text{”} H_0) = \alpha$	Richtige Entscheidung “ H_1 ” H_1 $P(\text{“}H_1\text{”} H_1) = 1 - \beta$

- **Fehler 1. Art** beinhaltet die fälschliche Ablehnung der Nullhypothese, d.h. die Ablehnung der Nullhypothese, obwohl sie wahr ist
- **Fehler 2. Art** beinhaltet die fälschliche Beibehaltung der Nullhypothese, d.h. die Nichtablehnung der Nullhypothese, obwohl sie falsch ist

Beispiel 19.2

Entscheidung	Wirklichkeit	
	Regen	kein Regen
Schirm mitnehmen	richtige Entscheidung	Fehler
Schirm nicht mitnehmen	Fehler	richtige Entscheidung

kleine Abweichung zwischen Prüfwert v und hypothetischem Wert ϑ_0

- Abweichung zwischen v und ϑ_0 wird als zufällig angesehen
- H_0 wird nicht abgelehnt
- das bedeutet nicht, dass H_0 richtig ist

große Abweichung zwischen Prüfwert v und hypothetischem Wert ϑ_0

- v ist ein für die Gültigkeit der Nullhypothese unplausibel Wert
- H_0 wird abgelehnt
- Teststatistik V verteilt mit einem Parameterwert verschieden von dem unter H_0
- das bedeutet nicht, dass H_0 falsch ist

Die Menge der Realisationen der Teststatistik V wird in zwei komplementäre Bereiche unterteilt:

- **Nicht–Ablehnungsbereich von H_0 :** die Menge der Realisationen der Teststatistik V , bei denen man sich nicht für die Ablehnung von H_0 entscheidet
- **Ablehnungsbereich von H_0 :** die Menge der Realisationen der Teststatistik V , bei denen man sich für die Ablehnung von H_0 entscheidet

Kritischer Wert

- trennt den Nichtablehnungs- und Ablehnungsbereich der H_0 voneinander
- gehört zum Nichtablehnungsbereich von H_0

• Zweiseitiger Test

- ▶ zu große Abweichungen von ϑ_0 nach beiden Seiten sind nicht akzeptabel

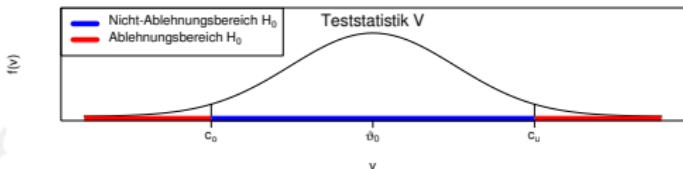
• Rechtsseitiger Test

- ▶ zu große Abweichungen von ϑ_0 nach rechts sind nicht akzeptabel

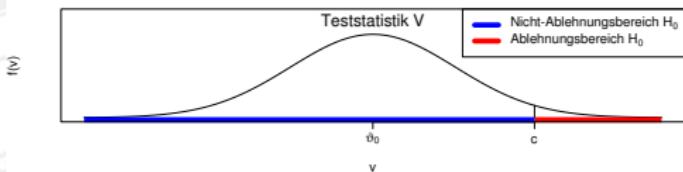
• Linksseitiger Test

- ▶ zu große Abweichungen von ϑ_0 nach links sind nicht akzeptabel

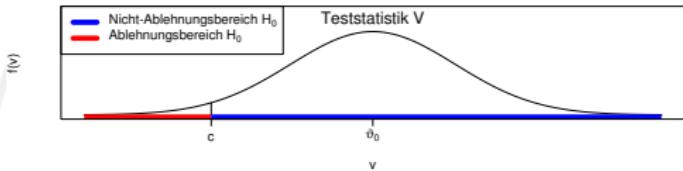
Zweiseitiger Test: $H_0: \vartheta = \vartheta_0$ vs. $H_1: \vartheta \neq \vartheta_0$



Rechtsseitiger Test: $H_0: \vartheta \leq \vartheta_0$ vs. $H_1: \vartheta > \vartheta_0$



Linksseitiger Test: $H_0: \vartheta \geq \vartheta_0$ vs. $H_1: \vartheta < \vartheta_0$



Kritische Werte

Signifikanzniveau α

- wird vor dem Test festgelegt (Konvention: $\alpha = 5\%$)
- gibt an, wie groß der Fehler 1. Art maximal sein darf:

$$P(V \in \text{Ablehnungsbereich von } H_0 | \vartheta_0) \leq \alpha$$

Die Wahrscheinlichkeit, dass bei Gültigkeit der Nullhypothese die Teststatistik V Realisationen im Ablehnungsbereich der H_0 annimmt, soll nicht größer als ein vorgegebenes α sein

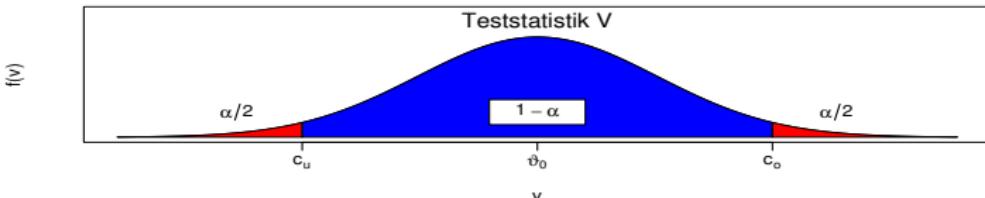
$$P(V \in \text{Nichtablehnungsbereich von } H_0 | \vartheta_0) \geq 1 - \alpha$$

Die Wahrscheinlichkeit, dass bei Gültigkeit der Nullhypothese die Teststatistik V Realisationen im Nichtablehnungsbereich der H_0 annimmt, soll mindestens $1 - \alpha$ sein

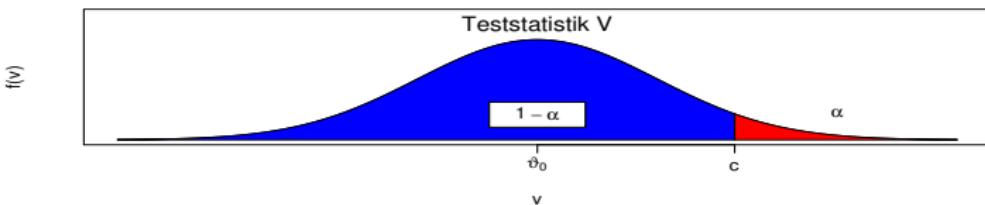
- Mit dem Signifikanzniveau (Wahrscheinlichkeit) α können aus der Verteilung der Teststatistik V die kritischen Werte bestimmt werden
- die Verteilung von V muss bei der Gültigkeit der Nullhypothese H_0 (zumindest approximativ) bekannt sein

Test	Hypothesen		Kritischer Wert
Zweiseitig	$H_0 : \vartheta = \vartheta_0$	$H_1 : \vartheta \neq \vartheta_0$	$P(V < c_u \vartheta_0) \leq \alpha/2$ $P(V > c_o \vartheta_0) \leq \alpha/2$
Rechtsseitig	$H_0 : \vartheta \leq \vartheta_0$	$H_1 : \vartheta > \vartheta_0$	$P(V > c \vartheta_0) \leq \alpha$
Linksseitig	$H_0 : \vartheta \geq \vartheta_0$	$H_1 : \vartheta < \vartheta_0$	$P(V < c \vartheta_0) \leq \alpha$
Nichtablehnungsbereich H_0		Ablehnungsbereich H_0	
Zweiseitig	$\{v \mid c_u \leq v \leq c_o\}$		$\{v \mid v < c_u \text{ oder } v > c_o\}$
Rechtsseitig	$\{v \mid v \leq c\}$		$\{v \mid v > c\}$
Linksseitig	$\{v \mid v \geq c\}$		$\{v \mid v < c\}$

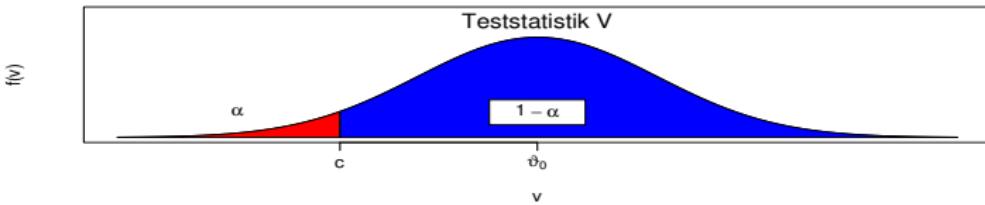
Zweiseitiger Test: $H_0: \vartheta = \vartheta_0$ vs. $H_1: \vartheta \neq \vartheta_0$



Rechtsseitiger Test: $H_0: \vartheta \leq \vartheta_0$ vs. $H_1: \vartheta > \vartheta_0$



Linksseitiger Test: $H_0: \vartheta \geq \vartheta_0$ vs. $H_1: \vartheta < \vartheta_0$



Fehler 2. Art

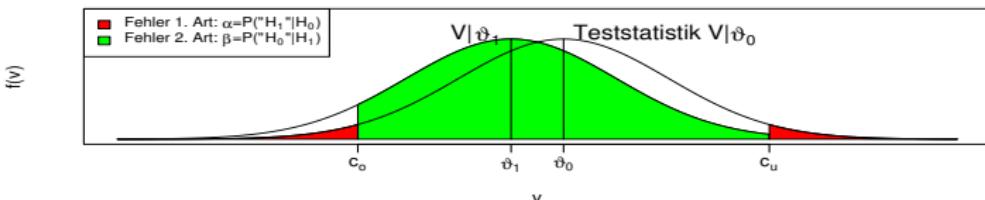
$$P(V \in \text{Nichtablehnungsbereich von } H_0 \mid \vartheta_1) = P(\text{"}H_0\text{"} \mid H_1) = \beta(\vartheta_1)$$

- hängt ab vom vorgegebenen Signifikanzniveau α
 - ▶ kritische Werte des Tests werden zur Berechnung benötigt

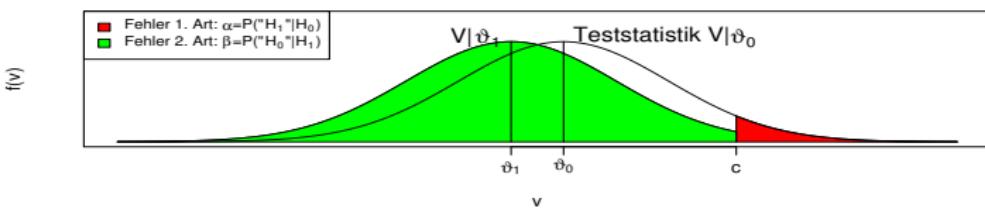
Test	$\beta(\vartheta_1) = P(\text{"}H_0\text{"} \mid H_1)$
Zweiseitig	$P(c_u \leq V \leq c_o \mid \vartheta_1)$
Linksseitig	$P(V \geq c \mid \vartheta_1)$
Rechtsseitig	$P(V \leq c \mid \vartheta_1)$

- hängt ab vom Stichprobenumfang n
- hängt ab von der Lage des wahren Parameterwertes ϑ_1 gegenüber dem hypothetischen Wert ϑ_0 unter H_0

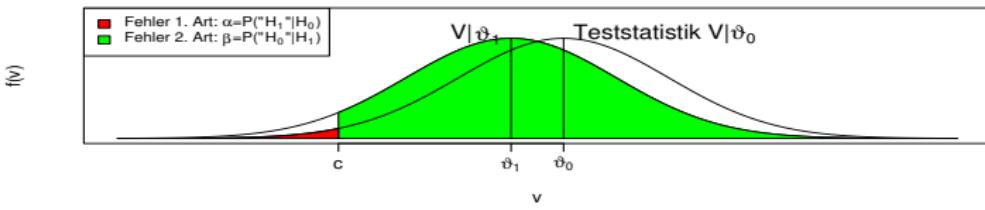
Zweiseitiger Test: $H_0: \vartheta = \vartheta_0$ vs. $H_1: \vartheta \neq \vartheta_0$



Rechtsseitiger Test: $H_0: \vartheta \leq \vartheta_0$ vs. $H_1: \vartheta > \vartheta_0$



Linksseitiger Test: $H_0: \vartheta \geq \vartheta_0$ vs. $H_1: \vartheta < \vartheta_0$



Parameter	Situation	Verhalten $\beta(\vartheta_1)$
Signifikanzniveau α (Fehler 1. Art)	Verminderung	Vergrößert sich
	Erhöhung	Verringert sich
Stichprobenumfang n	Verminderung	Vergrößert sich
	Erhöhung	Verringert sich
Abstand wahrer Parameterwert ϑ_1 und hypothetischer Wert ϑ_0	Groß	Kleiner
	Klein	Größer

- Signifikanzniveau α und Stichprobenumfang n werden festgelegt und können beeinflusst werden
- Der Abstand zwischen dem wahren Parameterwert ϑ_1 und dem hypothetischen Wert ϑ_0 ist unbekannt und kann nicht beeinflusst werden

Testentscheidung und Interpretation

- Prüfwert $v = V(x_1, \dots, x_n)$ berechnen
- $v \in$ Ablehnungsbereich der $H_0 \Rightarrow "H_1"$
- Das bedeutet nicht, dass H_1 richtig ist
- Standardisierter Antwortsatz:

Es konnte statistisch auf dem Signifikanzniveau α und basierend auf der einfachen Zufallsstichprobe vom Umfang n gezeigt werden, dass H_1 gilt.

- ▶ statistisch: Möglichkeit, eine Fehlentscheidung zu begehen
- ▶ gezeigt werden, dass H_1 gilt: Immer Bezug auf H_1
- Möglichkeit, einen Fehler 1. Art ($"H_1" | H_0$) zu begehen, wenn in Wirklichkeit die Nullhypothese richtig ist

- Prüfwert $v = V(x_1, \dots, x_n)$ berechnen
- $v \in$ Nichtablehnungsbereich der $H_0 \Rightarrow "H_0"$
- Das bedeutet nicht, dass H_0 richtig ist
- Standardisierter Antwortsatz:

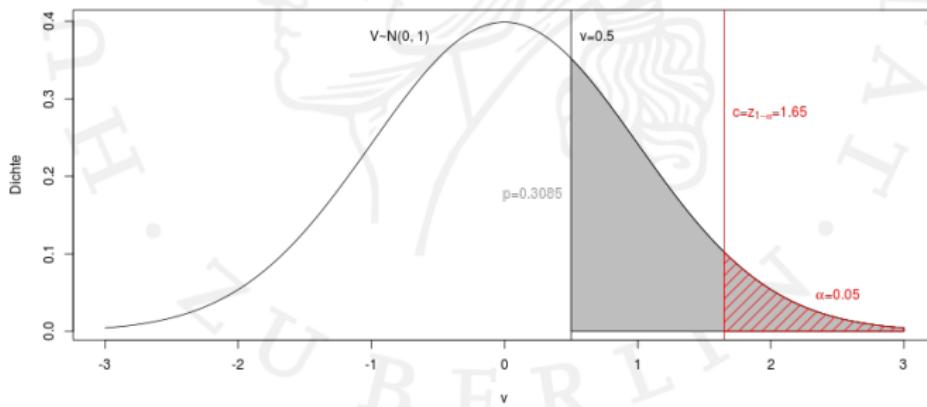
Es konnte statistisch auf dem Signifikanzniveau α und basierend auf der einfachen Zufallsstichprobe vom Umfang n nicht gezeigt werden, dass H_1 gilt.

- ▶ statistisch: Möglichkeit, eine Fehlentscheidung zu begehen
- ▶ nicht gezeigt werden, dass H_1 gilt: Immer Bezug auf H_1
- Möglichkeit, einen Fehler 2. Art ($"H_0" | H_1$) zu begehen, wenn in Wirklichkeit die Alternativhypothese richtig ist

- p -Wert als Alternative zur Berechnung des Prüf- und des kritischen Wertes

$$p\text{-Wert} = \begin{cases} 2 \min\{P(V \leq v), P(V \geq v)\} & \text{beidseitiger Test} \\ P(V \leq v) & \text{linksseitiger Test} \\ P(V \geq v) & \text{rechtsseitiger Test} \end{cases}$$

- p -Wert $\geq \alpha$, dann Nullhypothese beibehalten (" H_0 ")
- p -Wert $< \alpha$, dann Nullhypothese verwerfen (" H_1 ")



Allgemeiner Ablauf eines Tests

1. Prüfung von evtl. Voraussetzungen

- ▶ u.a. zur Berechnung der Verteilung der Teststatistik V unter H_0 benötigt

2. Festlegung der Hypothesen: Null- (H_0) und Alternativhypothese (H_1)

- ▶ einige Tests erlauben zwei- oder einseitige Hypothesen
- ▶ für einige Tests gibt es nur ein mögliches Hypothesenpaar

3. Berechnung der Verteilung der Teststatistik V unter H_0

4. Festlegung des Signifikanzniveaus α und Berechnung der kritischen Werte

- ▶ Konvention: $\alpha = 5\%$

5. Ziehung einer Zufallsstichprobe und Berechnung des Prüfwertes

6. Testentscheidung und Interpretation

Allgemeine Handlungsweise

- Hypothesenaufbau (grob):
 - ▶ H_0 : es gibt keinen Effekt vs. H_1 : es gibt einen Effekt
- In der Praxis passiert nach der Testdurchführung oft folgendes
 - ▶ " H_1 " : wir verhalten uns so als ob H_1 wahr ist
 - ▶ " H_0 " : wir verhalten uns so als ob H_0 wahr ist
- Es gilt jedoch
 - ▶ " H_1 " : unsere Daten zeigen, das ein Effekt existiert (Irrtumswk. α !)
 - ▶ " H_0 " : mögliche Schußfolgerungen
 - ★ Es gibt keinen Effekt
 - ★ Es gibt einen Effekt, aber wir können ihn mit den Daten nicht nachweisen, also H_1 wahr
- Wie sollten wir uns nach der Testdurchführung verhalten?
Sachgebietsinformationen heranziehen!
 - ▶ " H_1 " : wir verhalten uns so als ob H_1 wahr ist; je größer die Folgekosten für den Fehler 1. Art, desto vorsichtiger (α kleiner machen?)
 - ▶ " H_0 " : Folgekosten eines Irrtums abschätzen

Testverfahren

- Parametrische Tests

Annahme einer Verteilung in der Grundgesamtheit

- ▶ Test auf Mittelwert
- ▶ Test auf Differenz von zwei Mittelwerten
- ▶ Test auf Anteilswert

- Nichtparametrische Tests

Keine Annahme einer Verteilung in der Grundgesamtheit

- ▶ χ^2 Anpassungstest
- ▶ χ^2 Unabhängigkeitstest

Parametrische Tests

5. November 2022

Tests für den Mittelwert • Gauß-Test • Einstichproben t-Test •
Hypothesenwahl für einseitige Tests • Test für Anteilswert π • Test auf
Differenz zweier Mittelwerte • Zwei-Stichproben Gauß-Test •
Zwei-Stichproben t-Test • Test der Regressionsparameter • Test des
Regressionskoeffizienten β_1 • Drost vs. Kekulé • Gütfunktion (Macht)
eines Tests • Gütfunktion für Test auf Mittelwert • Gütfunktion für Test
auf Anteilswert

Tests für den Mittelwert

Für Hypothesen bezüglich

- des Mittelwerts einer Grundgesamtheit
 - ▶ Gauß Test
 - ▶ Einstichproben t-Test
 - ▶ Kruskal-Wallis-Test
 - ▶ Mann-Whitney-U-Test/Wilcoxon-Rangsummen Test
 - ▶ Median-Test
- der Mittelwerte zweier Grundgesamtheiten
 - ▶ Zweistichproben t-Test
 - ▶ Welch-Test
 - ▶ Zweistichproben t-Test für abhängige Stichproben
- der Mittelwerte mehrerer Grundgesamtheiten
 - ▶ ANalysis Of VAriance (ANOVA)
 - ▶ Friedman Test

Werden in der Vorlesung "Datenanalyse I" behandelt

- Verwendete Stichprobenfunktion

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

- Verteilung von \bar{X} , wenn X_i unabhängig und identisch verteilt
 - ▶ X_i normalverteilt

$$X_i \sim N(\mu; \sigma^2) \implies \bar{X} \sim N(\mu; \sigma^2/n)$$

- ▶ X_i beliebig verteilt und die Voraussetzungen des Zentralen Grenzwertsatzes erfüllt

$$X_i \sim (\mu; \sigma^2) \implies \bar{X} \approx N(\mu; \sigma^2/n)$$

Gauß-Test

1. Voraussetzungen:

- Stichprobe vom Umfang n ist eine Zufallsstichprobe
- Verteilung der Stichprobenvariablen X_i i.i.d. und
 - ▶ normalverteilt oder
 - ▶ beliebig verteilt, aber ZGS erfüllt
- $\text{Var}(X_i) = \sigma^2$ bekannt

2. Hypothesen:

Test	Zweiseitig	Einseitig	
		Linksseitig	Rechtsseitig
Nullhypothese H_0	$\mu = \mu_0$	$\mu \geq \mu_0$	$\mu \leq \mu_0$
Alternativhypothese H_1	$\mu \neq \mu_0$	$\mu < \mu_0$	$\mu > \mu_0$



3. Verteilung der Teststatistik V unter H_0 :

- μ in der Grundgesamtheit unbekannt
- In der Nullhypothese wird unterstellt, dass der hypothetische Wert μ_0 der wahre Mittelwert in der Grundgesamtheit ist

$$\mu = \mu_0 \implies \bar{X} \approx N(\mu_0; \sigma^2/n)$$

$$E(\bar{X}) = \mu_0, \quad \text{Var}(\bar{X}) = \sigma^2/n$$

$$V = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \approx N(0; 1)$$

$$E(V) = 0, \quad \text{Var}(V) = 1$$

4. Berechnung der kritischen Werte:

- Festlegung des Signifikanzniveaus α (Wahrscheinlichkeit eines Fehlers 1. Art)

$$\alpha = P("H_1" \mid H_0)$$

Test	Zweiseitig	Linksseitig	Rechtsseitig
$N(0; 1)$			
$P(\text{wei\beta})$	$P(c_u \leq V \leq c_o)$ $= 1 - \alpha$ $c_u = -c, c_o = +c$	$P(c_l \leq V)$ $= 1 - \alpha$ $c_l = -c^*$	$P(V \leq c_r)$ $= 1 - \alpha$ $c_r = c^*$
$P(\text{rot})$	$P(V < -c) + P(c < V)$ $= \alpha/2$	$P(V < c)$ $= \alpha$	$P(c < V)$ $= \alpha$

Test	Zweiseitig	Linksseitig	Rechtsseitig
$N(0; 1)$			
Kritische Werte	$c = z_{1-\alpha/2}$		$c^* = z_{1-\alpha}$
Ablehnungsbereich H_0	$\{v v < -c \text{ oder } v > +c\}$	$\{v v < -c^*\}$	$\{v v > +c^*\}$
Nichtablehnungsbereich H_0	$\{v -c \leq v \leq +c\}$	$\{v v \geq -c^*\}$	$\{v v \leq +c^*\}$

Quantile der Standardnormalverteilung:

q	0,900	0,950	0,975	0,990	0,995
z_q	1,28	1,64	1,96	2,33	2,58

$$V = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \iff \bar{X} = \mu_0 + \frac{\sigma}{\sqrt{n}} V$$

Zweiseitiger Test

$$c_u = -z_{1-\alpha/2} \iff \bar{x}_u = \mu_0 - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$c_o = +z_{1-\alpha/2} \iff \bar{x}_o = \mu_0 + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

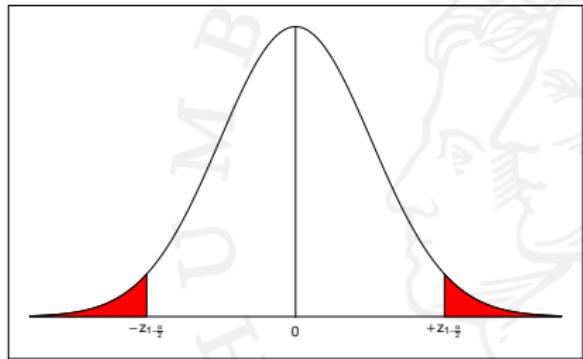
Linksseitiger Test

$$c_l = -z_{1-\alpha} \iff \bar{x}_l = \mu_0 - z_{1-\alpha} \frac{\sigma}{\sqrt{n}}$$

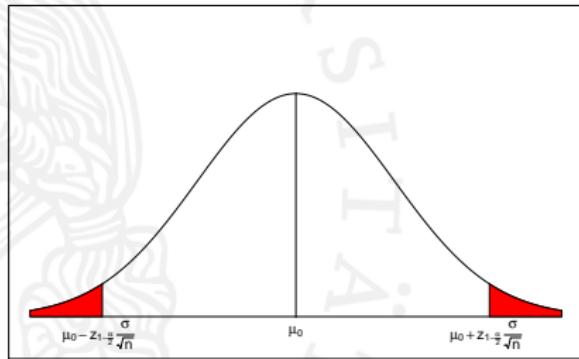
Rechtsseitiger Test

$$c_r = +z_{1-\alpha} \iff \bar{x}_r = \mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}}$$

Verteilung der Teststatistik V



Verteilung der Schätzfunktion \bar{X}



Beispiel 20.1

Problem: Mehlproduzent möchte sicher sein, dass in den Mehltüten genau ein Kilogramm Mehl ist

- zu wenig: Kunden beschweren sich
- zu viel: entgangener Gewinn

1. Voraussetzungen:

- ▶ Es ist bekannt, dass $X_i \sim N(\mu; \sigma^2 = 25g^2)$
- ▶ σ^2 bekannt → Gauß-Test

2. Hypothesen:

- ▶ $H_0 : \mu = \mu_0 = 1kg$
- ▶ $H_1 : \mu \neq \mu_0 = 1kg$

3. Verteilung der Teststatistik: $V = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \sim N(0; 1)$

4. Kritische Werte:

- ▶ Signifikanzniveau festgelegt mit $\alpha = 5\%$
- ▶ Kritische Werte aus Standardnormalverteilung

$$c_u = -z_{0,975} = -1,96 \quad c_o = +z_{0,975} = +1,96$$

- ▶ Ablehnungsbereich der H_0

$$\{v | v < -1,96 \text{ oder } v > +1,96\}$$

$$\{\bar{x} | \bar{x} < \mu_0 - 1,96 \frac{5}{\sqrt{n}} \text{ oder } \bar{x} > \mu_0 + 1,96 \frac{5}{\sqrt{n}}\}$$

- ▶ Nichtablehnungsbereich der H_0

$$\{v | -1,96 \leq v \leq +1,96\}$$

$$\{\bar{x} | \mu_0 - 1,96 \frac{5}{\sqrt{n}} \leq \bar{x} \leq \mu_0 + 1,96 \frac{5}{\sqrt{n}}\}$$

5. Ziehung einer einfachen Zufallsstichprobe mit Umfang $n = 25$

- ▶ Aus Stichprobe berechnet: $\bar{x} = 1002g$

6. Testentscheidung und Interpretation:

- ▶ Testentscheidung: $v \in$ Ablehnungsbereich der H_0

$$v = \frac{1002 - 1000}{5/\sqrt{25}} = 2 \Rightarrow "H_1"$$

- ▶ Interpretation:

Es konnte statistisch auf dem Signifikanzniveau $\alpha = 5\%$ und basierend auf der einfachen Zufallsstichprobe vom Umfang $n = 25$ gezeigt werden, dass der Mittelwert μ in der Grundgesamtheit verschieden vom hypothetischen Wert $\mu_0 = 1kg$ ist.

- ★ Das bedeutet nicht, dass H_1 richtig ist
- ★ Mit der maximalen Wahrscheinlichkeit $\alpha = 5\%$ haben wir den Fehler 1. Art begangen

5. Ziehung einer weiteren einfachen Zufallsstichprobe mit Umfang $n = 25$

- ▶ Aus Stichprobe berechnet: $\bar{x} = 999g$

6. Testentscheidung und Interpretation

- ▶ Testentscheidung: $v \in \text{Nicht-Ablehnungsbereich der } H_0$

$$v = \frac{999 - 1000}{5/\sqrt{25}} = -1 \Rightarrow "H_0"$$

- ▶ Interpretation:

Es konnte statistisch auf dem Signifikanzniveau $\alpha = 5\%$ und basierend auf der einfachen Zufallsstichprobe vom Umfang $n = 25$ nicht gezeigt werden, dass der Mittelwert μ in der Grundgesamtheit verschieden vom hypothetischen Wert $\mu_0 = 1kg$ ist.

- ★ Das bedeutet nicht, dass H_0 richtig ist
- ★ Mit einer unbekannten Wahrscheinlichkeit β haben wir den Fehler 2. Art begangen

```
> library("compositions") # Gauss.test  
> library("foreign")      # read.spss  
> child <- read.spss("child_data.sav", to.data.frame=T)  
> Gauss.test(child$iq, mean=100, sd=15)
```

Ein Stichproben Gauss-Test

```
data: child$iq  
T = 96.5, mean = 100, sd = 15, p-value = 0.2967  
alternative hypothesis: two.sided
```

Einstichproben t-Test

1. Voraussetzungen:

- Stichprobe vom Umfang n ist eine Zufallsstichprobe
- Verteilung der StichprobenvARIABLEN X_i i.i.d. und
 - ▶ normalverteilt oder
 - ▶ beliebig verteilt, aber ZGS erfüllt
- $\text{Var}(X_i) = \sigma^2$ unbekannt (trifft meistens in der Praxis zu)

2. Hypothesen:

Test	Zweiseitig	Einseitige	
		Linksseitig	Rechtsseitig
Nullhypothese H_0	$\mu = \mu_0$	$\mu \geq \mu_0$	$\mu \leq \mu_0$
Alternativhypothese H_1	$\mu \neq \mu_0$	$\mu < \mu_0$	$\mu > \mu_0$

3. Verteilung der Teststatistik V unter H_0

- σ jetzt unbekannt in

$$V = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

- Ersetze σ^2 durch eine erwartungstreue Schätzfunktion

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

- Verteilung der Teststatistik ergibt sich zu

$$V = \frac{\bar{X} - \mu_0}{S / \sqrt{n}} \sim t_{n-1}$$

Herleitung der Verteilung

$$V = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}}{\frac{S}{\sigma}} = \frac{\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}}{\sqrt{\frac{1}{n-1} \frac{(n-1)S^2}{\sigma^2}}} = \frac{Z}{\sqrt{\frac{C}{n-1}}}$$

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0; 1) \quad C = \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

- Definition der t-Verteilung: X ist t_n verteilt, wenn

$$X \sim \frac{Z}{\sqrt{\frac{C}{n}}} \text{ mit } Z \sim N(0; 1) \text{ und } C \sim \chi_n^2$$

- Daraus folgt $V \sim t_{n-1}$

4. Berechnung der kritischen Werte:

- Festlegung des Signifikanzniveaus α (Wahrscheinlichkeit eines Fehlers 1. Art)

$$\alpha = P("H_1" \mid H_0)$$

Test	Zweiseitig	Linksseitig	Rechtsseitig
$P(\text{weiß})$	$P(c_u \leq V \leq c_o)$ $= 1 - \alpha$ $c_u = -c, c_o = +c$	$P(c_l \leq V)$ $= 1 - \alpha$ $c_l = -c^*$	$P(V \leq c_r)$ $= 1 - \alpha$ $c_r = c^*$
$P(\text{rot})$	$P(V < -c) + P(c < V)$ $= \alpha/2$	$P(V < c)$ $= \alpha$	$P(c < V)$ $= \alpha$

Test	Zweiseitig	Linksseitig	Rechtsseitig
Kritische Werte	$c = t_{n-1;1-\alpha/2}$		$c^* = t_{n-1;1-\alpha}$
Ablehnungsbereich H_0	$\{v v < -c \text{ oder } v > +c\}$	$\{v v < -c^*\}$	$\{v v > +c^*\}$
Nichtablehnungsbereich H_0	$\{v -c \leq v \leq +c\}$	$\{v v \geq -c^*\}$	$\{v v \leq +c^*\}$

Quantile der t_n -Verteilung ($t_n \approx N(0; 1)$ für $n > 30$)

q	0,900	0,950	0,975	0,990	0,995
$t_{10;q}$	1,37	1,81	2,23	2,76	3,17
$t_{30;q} \approx z_q$	1,31	1,70	2,04	2,46	2,75
$t_{\infty;q} = z_q$	1,28	1,64	1,96	2,33	2,58

Beispiel 20.2

Problem: Mehlproduzent möchte sicher sein, dass in den Mehltüten genau ein Kilogramm Mehl ist

- zu wenig: Kunden beschweren sich
- zu viel: entgangener Gewinn

1. Voraussetzungen:

- ▶ Es ist bekannt, dass $X_i \sim N(\mu; \sigma^2)$
- ▶ σ^2 unbekannt \rightarrow Einstichproben t -Test
- ▶ Zufallsstichprobe vom Umfang n

2. Hypothesen:

- ▶ $H_0 : \mu = \mu_0 = 1\text{kg}$
- ▶ $H_1 : \mu \neq \mu_0 = 1\text{kg}$

3. Verteilung der Teststatistik: $V = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$

4. Kritische Werte:

- ▶ Signifikanzniveau festgelegt mit $\alpha = 5\%$
- ▶ Kritische Werte aus t_{24} -Verteilung

$$c_u = -t_{24;0,975} = -2,064 \quad c_o = +t_{24;0,975} = +2,064$$

- ▶ Ablehnungsbereich der H_0

$$\{v | v < -2,064 \text{ oder } v > +2,064\}$$

$$\{\bar{x} | \bar{x} < \mu_0 - 2,064 \frac{s}{\sqrt{n}} \text{ oder } \bar{x} > \mu_0 + 2,064 \frac{s}{\sqrt{n}}\}$$

- ▶ Nichtablehnungsbereich der H_0

$$\{v | -2,064 \leq v \leq +2,064\}$$

$$\{\bar{x} | \mu_0 - 2,064 \frac{s}{\sqrt{n}} \leq \bar{x} \leq \mu_0 + 2,064 \frac{s}{\sqrt{n}}\}$$

5. Ziehung einer einfachen Zufallsstichprobe mit Umfang $n = 25$ Aus der Stichprobe berechnet

$$\bar{x} = 1002 \text{ g} \quad s = 4,9 \text{ g}$$

6. Testentscheidung und Interpretation:

- ▶ Testentscheidung: $v \in \text{Nichtablehnungsbereich der } H_0$

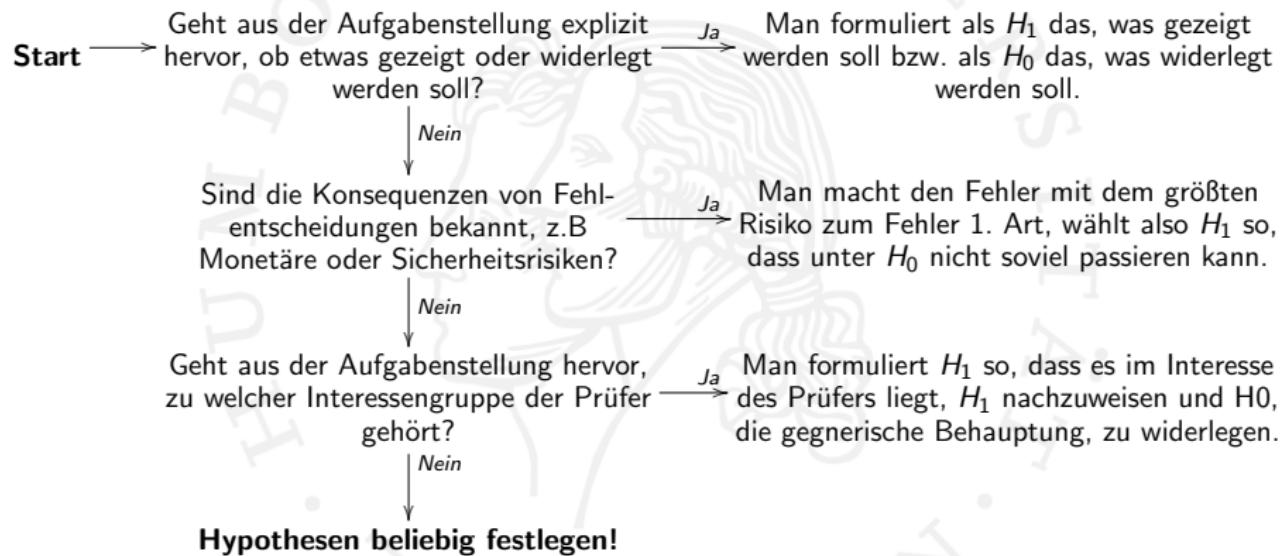
$$v = \frac{1002 - 1000}{4,9/\sqrt{25}} = 2,04 \Rightarrow "H_0"$$

- ▶ Interpretation:

Es konnte statistisch auf dem Signifikanzniveau $\alpha = 5\%$ und basierend auf der einfachen Zufallsstichprobe vom Umfang $n = 25$ nicht gezeigt werden, dass der Mittelwert μ in der Grundgesamtheit verschieden vom hypothetischen Wert $\mu_0 = 1 \text{ kg}$ ist.

- ★ Das bedeutet nicht, dass H_0 richtig ist
- ★ Mit einer unbekannten Wahrscheinlichkeit β haben wir den Fehler 2. Art begangen

Hypothesenwahl für einseitige Tests



Quelle: H. Rinne (1997), Taschenbuch der Statistik, 2. Auflage, Verlag Harri Deutsch, S. 528

Man formuliert als H_1 das, was gezeigt werden soll bzw. als H_0 das, was widerlegt werden soll.

- Man nimmt als Ergebnis der Testdurchführung " H_0 " an
 - ▶ Es könnte der Fehler 2. Art (" H_0 "| H_1) unterlaufen sein
 - ▶ Die Wahrscheinlichkeit dafür ist jedoch unbekannt
 - ⇒ es könnte H_0 oder H_1 richtig sein
- Man nimmt als Ergebnis der Testdurchführung " H_1 " an
 - ▶ Es könnte der Fehler 1. Art (" H_1 "| H_0) unterlaufen sein
 - ▶ Die Wahrscheinlichkeit dafür ist kleiner gleich dem Signifikanzniveau α
 - ⇒ " H_1 "| H_0 kann nur mit kleiner Wahrscheinlichkeit richtig sein

Man macht den Fehler mit dem größten Risiko zum Fehler 1. Art, wählt also H_1 so, dass unter H_0 nicht soviel passieren kann.

Beispiel 20.3

- Bankkunde will einen Kredit von 1000 EUR
 - ▶ Bank gibt Kredit und Kunde insolvent \Rightarrow Verlust von 1000 EUR
 - ▶ Bank gibt keinen Kredit und Kunde solvent \Rightarrow Verlust von 100 EUR
- Betrachte Fehler 1. Art und 2. Art: " $H_1|H_0$ " bzw. " $H_1|H_0$ "

H_0 :	Kunde insolvent	Kunde solvent
H_1 :	Kunde solvent	Kunde insolvent
" $H_1 H_0$ "	"Kunde solvent" Kunde insolvent	"Kunde insolvent" Kunde solvent
Verlust	1000 EUR	100 EUR
" $H_0 H_1$ "	"Kunde insolvent" Kunde solvent	"Kunde solvent" Kunde insolvent
Verlust	100 EUR	1000 EUR

- $H_0 : \mu \leq \mu_0$ vs. $H_1 : \mu > \mu_0$
- Betrachte Fehler 1. Art und 2. Art: " H_1 "| H_0 bzw. " H_0 "| H_1

Beispiel 20.4

- " H_1 "| H_0 = "Gewicht zu groß" | Gewicht zu gering oder gleich
 - ▶ Nichterkennen eines zu geringen Gewichts
⇒ Schlecht für Käufer, gut für Produzent
 - " H_0 "| H_1 = "Gewicht zu gering oder gleich" | Gewicht zu groß
 - ▶ Nichterkennen eines zu großen Gewichts
⇒ Gut für Käufer, schlecht für Produzent
- ⇒ Test gut für Käufer, schlecht für Produzent, da β -Fehler unbekannt

- $H_0 : \mu \geq \mu_0$ vs. $H_1 : \mu < \mu_0$
- Betrachte Fehler 1. Art und 2. Art: " H_1 "| H_0 bzw. " H_0 "| H_1

Beispiel 20.5

- " H_1 "| H_0 = "Gewicht zu gering"|"Gewicht zu groß oder gleich"
 - Fehler 1. Art (festgelegter α -Fehler)
 - ▶ Nichterkennen eines zu großen Gewichts
 - ⇒ Schlecht für Produzent, gut für Käufer
 - " H_0 "| H_1 = "Gewicht zu groß oder gleich"|"Gewicht zu gering"
 - Fehler 2. Art (unbekannter β -Fehler)
 - ▶ Nichterkennen eines zu geringen Gewichts
 - ⇒ Test gut für Produzent, schlecht für Käufer
- ⇒ Test gut für Produzent, schlecht für Käufer, da β -Fehler unbekannt

Test für Anteilswert π

Verwendete Stichprobenfunktion

$$\hat{\pi} = \frac{X}{n} = \frac{1}{n} \sum_{i=1}^n X_i$$

- X_i : „Anzahl der Elemente mit Eigenschaft beim i -ten Zug“
- $P(X_i = 1) = \pi, \quad P(X_i = 0) = 1 - \pi$
- $X_i \sim \text{Bernoulli}(1; \pi)$
- $X = \sum_{i=1}^n X_i$: „Anzahl der Elemente mit Eigenschaft in einer Zufallsstichprobe vom Umfang n “ $\rightarrow X$ diskrete Zufallsvariable

$$\Rightarrow X \sim B(n; \pi)$$

1. Voraussetzungen:

- dichotome Grundgesamtheit
- unbekannter Anteil π von Elementen mit einer interessierenden Eigenschaft

2. Hypothesen:

Test	Zweiseitig	Einseitig	
		Linksseitig	Rechtsseitig
Nullhypothese H_0	$\pi = \pi_0$	$\pi \geq \pi_0$	$\pi \leq \pi_0$
Alternativhypothese H_1	$\pi \neq \pi_0$	$\pi < \pi_0$	$\pi > \pi_0$



3. Verteilung der Teststatistik V unter H_0

- π in der Grundgesamtheit unbekannt
- In der Nullhypothese wird unterstellt, dass der hypothetische Wert π_0 der wahre Anteilswert der Grundgesamtheit ist

$$\pi = \pi_0 \quad \Rightarrow \quad X \sim B(n, \pi_0)$$

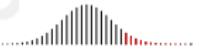
$$V = X \sim B(n, \pi_0)$$

4. Berechnung der kritischen Werte

- Festlegung des Signifikanzniveaus α (Wahrscheinlichkeit eines Fehlers 1. Art)

$$\alpha = P("H_1" \mid H_0)$$

Test	Zweiseitig	Linksseitig	Rechtsseitig
$B(n; \pi_0)$			
$P(\text{weiß})$	$\underbrace{P(x_u \leq V \leq x_o)}_{\geq 1-\alpha}$	$\underbrace{P(x_l \leq V)}_{\geq 1-\alpha}$	$\underbrace{P(V \leq x_r)}_{\geq 1-\alpha}$
$P(\text{rot})$	$\underbrace{P(V < x_u)}_{\leq \alpha/2} + \underbrace{P(x_o < V)}_{\leq \alpha/2}$	$\underbrace{P(V < x_l)}_{\leq \alpha}$	$\underbrace{P(x_r < V)}_{\leq \alpha}$

Test	Zweiseitig	Linksseitig	Rechtsseitig
$B(n; \pi_0)$			
Kritische Werte	x_u und x_o	x_l	x_r
Ablehnungsbereich H_0	$\{v v < x_u$ oder $v > x_o\}$	$\{v v < x_l\}$	$\{v v > x_r\}$
Nichtablehnungsbereich H_0	$\{v x_u \leq v \leq x_o\}$	$\{v v \geq x_l\}$	$\{v v \leq x_r\}$

- Zweiseitiger Test:

- ▶ unterer kritischer Wert x_u : diejenige Realisation vom X , für die $F_B(x)$ den Wert $\alpha/2$ gerade überschreitet, so dass gilt:

$$F_B(x_u - 1) \leq \alpha/2 \text{ und } F_B(x_u) > \alpha/2$$

- ▶ oberer kritischer Wert x_o : diejenige Realisation von X , für die $F_B(x)$ den Wert $1 - \alpha/2$ gerade erreicht oder überschreitet, so dass gilt:

$$F_B(x_o - 1) < 1 - \alpha/2 \text{ und } F_B(x_o) \geq 1 - \alpha/2$$

- Linksseitiger Test

- ▶ kritischer Wert x_l : diejenige Realisation vom X , für die $F_B(x)$ den Wert α gerade überschreitet, so dass gilt:

$$F_B(x_l - 1) \leq \alpha \text{ und } F_B(x_l) > \alpha$$

- Rechtsseitiger Test

- ▶ kritischer Wert x_r : diejenige Realisation vom X , für die $F_B(x)$ den Wert $1 - \alpha$ gerade erreicht oder überschreitet, so dass gilt:

$$F_B(x_r - 1) < 1 - \alpha \text{ und } F_B(x_r) \geq 1 - \alpha$$

Approximation durch die Normalverteilung

- Voraussetzung: genügend großer Stichprobenumfang n
⇒ Stichprobenfunktion $\hat{\pi}$ approximativ normalverteilt
- Teststatistik:

$$V = \frac{\hat{\pi} - \pi_0}{\sigma_0(\hat{\pi})} = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \approx N(0; 1) \text{ unter } H_0$$

- ▶ $\sigma_0(\hat{\pi})$ - Standardabweichung der Schätzfunktion $\hat{\pi}$ unter H_0
- kritische Werte $z_{1-\alpha/2}$ (zweiseitiger Test) bzw. $z_{1-\alpha}$ (einseitiger Test) aus der Tabelle der Verteilungsfunktion der $N(0; 1)$ entnehmen

5. Ziehung einer Zufallsstichprobe und Berechnung des Prüfwertes

- Stichprobenwerte: x_1, \dots, x_n
- Prüfwert: $v = x$

6. Testentscheidung und Interpretation

- Analog zum Test auf den Mittelwert μ
- Beachten:
 - ▶ Fehler 1. Art: " H_1 " | H_0
unberechtigte Ablehnung der H_0
 - ▶ Fehler 2. Art: " H_0 " | H_1
unberechtigte Beibehaltung der H_0

Beispiel 20.6

Problem: Es soll untersucht werden, ob eine Münze „fair“ ist.

1. Hypothesen:

- ▶ $H_0 : \pi = \pi_0 = 0,5$
- ▶ $H_1 : \pi \neq \pi_0 = 0,5$

2. Voraussetzungen:

- ▶ dichotome Grundgesamtheit: „Zahl“, „Kopf“
- ▶ interessierendes Ereignis „Zahl“ tritt mit Wahrscheinlichkeit π ein

3. Verteilung der Teststatistik unter H_0 :

$$V = X = \sum_{i=1}^n X_i \sim B(n; 0,5)$$

4. Kritische Werte:

- ▶ Signifikanzniveau festgelegt mit $\alpha = 0,05$
- ▶ Stichprobenumfang $n = 10$

Entscheidungsbereiche für $B(10; 0,5)$

x	$P(X \leq x)$	$P(X = x)$	
0	0,0010	0,0010	
1	0,0107	0,0097	Ablehnungsbereich von H_0
2	0,0547	0,0440	
3	0,1719	0,1172	
4	0,3770	0,2051	
5	0,6230	0,2460	Nichtabl.bereich von H_0
6	0,8281	0,2051	
7	0,9453	0,1172	
8	0,9893	0,0440	
9	0,9990	0,0097	Ablehnungsbereich von H_0
10	1,0000	0,0010	

- unterer kritischer Wert x_u : diejenige Realisation von X , für die $F_B(x)$ den Wert $\alpha/2 = 0,025$ gerade überschreitet, so dass gilt:

$$F_B(x_u - 1) \leq 0,025 \text{ und } F_B(x_u) > 0,025 \Rightarrow x_u = 2$$

- oberer kritischer Wert x_o : diejenige Realisation von X , für die $F_B(x)$ den Wert $1 - \alpha/2 = 0,975$ gerade erreicht oder überschreitet, so dass gilt:

$$F_B(x_o - 1) < 0,975 \text{ und } F_B(x_o) \geq 0,975 \Rightarrow x_o = 8$$

- Ablehnungsbereich der H_0 :

$$\{x | x < 2 \text{ oder } x > 8\} = \{0; 1\} \text{ oder } \{9; 10\}$$

- Nichtablehnungsbereich der H_0 :

$$\{x | 2 \leq x \leq 8\} = \{2; 3; 4; 5; 6; 7; 8\}$$

5. Ziehen einer einfachen Zufallsstichprobe vom Umfang $n = 10 \rightarrow$
10-maliges Werfen der Münze

- ▶ Prüfwert: $x = 9$

6. Testentscheidung und Interpretation

- ▶ Testentscheidung: $x = 9 \in$ Ablehnungsbereich der $H_0 \Rightarrow "H_1"$
- ▶ Interpretation:

*Es konnte statistisch auf dem Signifikanzniveau $\alpha_{\text{exakt}} = 0,0214$
und basierend auf einer Zufallsstichprobe vom Umfang $n = 10$
gezeigt werden, dass die verwendete Münze nicht "fair" ist.*

- ★ Das bedeutet nicht, dass H_1 richtig ist
- ★ Mit einer Wahrscheinlichkeit von $\alpha_{\text{exakt}} = 0,0214$ haben wir den Fehler
1. Art begangen

Beispiel 20.7

Problem:

- Ein Fabrikant von Massenartikeln möchte den Ausschussanteil seiner Produktion erfahren. Ist dieser größer als 20%, soll das Produktionsverfahren geändert werden.
- Das größte Risiko des Fabrikanten ist eine Entscheidung gegen Neuinvestitionen, obwohl mehr als 20% seiner Produkte fehlerhaft sind
- Geprüft werden soll auf einem Signifikanzniveau von $\alpha = 0,05$ und basierend auf einer einfachen Zufallsstichprobe vom Umfang $n = 30$

1. Hypothesen:

- ▶ Fehler mit größerem Risiko soll Fehler 1. Art sein (damit kontrollierbar)
- ▶ mögliche Hypothesen:

a) $H_0: \pi \geq \pi_0 = 0,2$ vs. $H_1: \pi < \pi_0 = 0,2$

Fehler 1. Art: „ $H_1|H_0$ “ = „Entscheidung gegen Neuinvestitionen“ | mehr als 20% der Produkte sind fehlerhaft

b) $H_0: \pi \leq \pi_0 = 0,2$ vs. $H_1: \pi > \pi_0 = 0,2$

Fehler 1. Art: „ $H_1|H_0$ “ = „Entscheidung für Neuinvestitionen“ | weniger als 20% der Produkte sind fehlerhaft

- ▶ Entscheidung für Hypothesenpaar a)
- ⇒ $H_0: \pi \geq \pi_0 = 0,2$
 $H_1: \pi < \pi_0 = 0,2$

2. Voraussetzungen:

- dichotome Grundgesamtheit: „fehlerhaftes Produkt“, „nicht fehlerhaftes Produkt“
- interessierendes Ereignis „fehlerhaftes Produkt“ tritt mit Wahrscheinlichkeit π ein

3. Verteilung der Teststatistik:

- X : „Anzahl defekter Teile bei einer Zufallsstichprobe vom Umfang n “

$$V = X = \sum_{i=1}^n X_i \sim B(n; 0, 2)$$

4. Kritischer Wert:

- ▶ Signifikanzniveau festgelegt mit $\alpha = 0,05$
- ▶ Stichprobenumfang $n = 30$

Entscheidungsbereiche für $B(30; 0,2)$

x	$P(X \leq x)$	$P(X = x)$	
0	0,0012	0,0012	
1	0,0105	0,0093	↑
2	0,0442	0,0337	Abl.bereich
3	0,1227	0,0785	Nichtabl.bereich
4	0,2552	0,1325	↓
5	0,4275	0,1723	
6	0,6070	0,1795	
.	.	.	

- kritischer Wert x_l : diejenige Realisation von X , für die $F_B(x)$ den Wert α gerade überschreitet, so dass gilt:

$$F_B(x_l - 1) \leq \alpha \text{ und } F_B(x_l) > \alpha \quad \Rightarrow x_l = 3$$

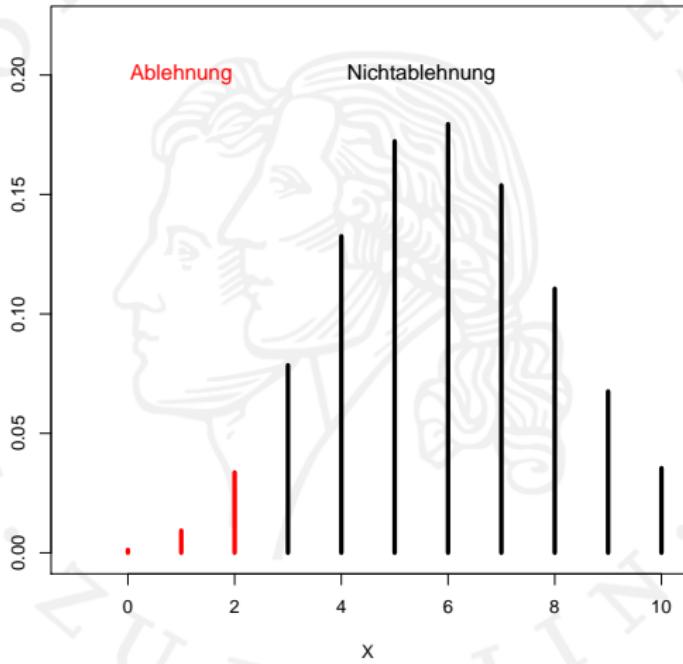
- Ablehnungsbereich der H_0 :

$$\{x \mid x \leq 2\} = \{0, 1, 2\}$$

- Nichtablehnungsbereich der H_0 :

$$\{x \mid x > 2\} = \{3, 4, \dots, 30\}$$

Wahrscheinlichkeitsfunktion der $B(30;0,2)$



5. Ziehung einer einfachen Stichprobe vom Umfang $n = 10$

- ▶ Prüfwert: $x = 5$

6. Testentscheidung und Interpretation:

- ▶ Testentscheidung: $x = 5 \in \text{Nichtablehnungsbereich der } H_0 \Rightarrow "H_0"$
- ▶ Interpretation:

Es konnte statistisch auf dem Signifikanzniveau $\alpha_{\text{exakt}} = 0,0442$ und basierend auf einer Zufallsstichprobe vom Umfang $n = 30$ nicht gezeigt werden, dass weniger als 20% der Teile defekt sind. Das Produktionsverfahren muss verändert werden

- ★ Das bedeutet nicht, dass H_0 richtig ist
- ★ Mit einer unbekannten Wahrscheinlichkeit β haben wir den Fehler 2. Art begangen

Test auf Differenz zweier Mittelwerte

- Hypothesenpaare:

zweiseitig

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

linksseitig

$$H_0 : \mu_1 \geq \mu_2$$

$$H_1 : \mu_1 < \mu_2$$

rechtsseitig

$$H_0 : \mu_1 \leq \mu_2$$

$$H_1 : \mu_1 > \mu_2$$

- Stichprobenfunktion für μ_1 und μ_2

$$\bar{X}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{1i} \quad E(\bar{X}_1) = \mu_1 \quad Var(\bar{X}_1) = \frac{\sigma_1^2}{n_1} \Rightarrow \bar{X}_1 \sim N\left(\mu_1; \frac{\sigma_1^2}{n_1}\right)$$

$$\bar{X}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} X_{2i} \quad E(\bar{X}_2) = \mu_2 \quad Var(\bar{X}_2) = \frac{\sigma_2^2}{n_2} \Rightarrow \bar{X}_2 \sim N\left(\mu_2; \frac{\sigma_2^2}{n_2}\right)$$

- Stichprobenfunktion für die Differenz $\mu_1 - \mu_2$

$$D = \bar{X}_1 - \bar{X}_2$$

$$E(D) = \omega = \mu_1 - \mu_2$$

$$\text{Var}(D) = \sigma_D^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

$$\Rightarrow D \sim N(\omega, \sigma_D^2)$$

- Tests

- ▶ σ_1^2 und σ_2^2 bekannt \Rightarrow Zwei-Stichproben Gauß-Test
- ▶ σ_1^2 und σ_2^2 unbekannt \Rightarrow Zwei-Stichproben t-Test
 - ★ $\sigma_1^2 \neq \sigma_2^2 \Rightarrow$ Welch-Test

Zwei-Stichproben Gauß-Test

1. Voraussetzungen

- zwei Grundgesamtheiten
- erste Grundgesamtheit: Zufallsvariable X_1
 - ▶ $E(X_1) = \mu_1$
 - ▶ $\text{Var}(X_1) = \sigma_1^2$ bekannt
 - ▶ $X_1 \sim N(\mu_1, \sigma_1)$
- zweite Grundgesamtheit: Zufallsvariable X_2
 - ▶ $E(X_2) = \mu_2$
 - ▶ $\text{Var}(X_2) = \sigma_2^2$ bekannt
 - ▶ $X_2 \sim N(\mu_2, \sigma_2)$
- μ_1 und μ_2 unbekannt
- einfache Zufallsstichprobe aus jeder Grundgesamtheit mit Stichprobenumfang n_1 bzw. $n_2 \Rightarrow$ Zweistichprobentest
- die beiden Zufallsstichproben sind unabhängig voneinander

2. Hypothesen:

Test	Zweiseitig	Einseitig	
	Linksseitig	Rechtsseitig	
Nullhypothese H_0	$\mu_1 - \mu_2 = \omega_0$	$\mu_1 - \mu_2 \geq \omega_0$	$\mu_1 - \mu_2 \leq \omega_0$
Alternativhypothese H_1	$\mu_1 - \mu_2 \neq \omega_0$	$\mu_1 - \mu_2 < \omega_0$	$\mu_1 - \mu_2 > \omega_0$

3. Verteilung der Teststatistik V unter H_0 :

- ω in der Grundgesamtheit unbekannt
- In der Nullhypothese wird unterstellt, dass der hypothetische Wert ω_0 der wahre Mittelwert in der Grundgesamtheit ist

$$\omega = \omega_0 \quad \Rightarrow E(D) = \omega_0, \quad \text{Var}(D) = \sigma_D^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

$$V = \frac{D - \omega_0}{\sigma_D} = \frac{(\bar{X}_1 - \bar{X}_2) - \omega_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

4. Berechnung der kritischen Werte:

- Festlegung des Signifikanzniveaus α (Wahrscheinlichkeit eines Fehlers 1. Art)

$$\alpha = P("H_1" \mid H_0)$$

Test	Zweiseitig	Linksseitig	Rechtsseitig
$N(0; 1)$			
$P(\text{weiß})$	$P(c_u \leq V \leq c_o) = 1 - \alpha$ $c_u = -c, c_o = +c$	$P(c_l \leq V) = 1 - \alpha$ $c_l = -c^*$	$P(V \leq c_r) = 1 - \alpha$ $c_r = c^*$
$P(\text{rot})$	$P(V < -c) + P(c < V) = \alpha/2$	$P(V < c) = \alpha$	$P(c < V) = \alpha$

Test	Zweiseitig	Linksseitig	Rechtsseitig
$N(0; 1)$			
Kritische Werte	$c = z_{1-\alpha/2}$		$c^* = z_{1-\alpha}$
Ablehnungsbereich H_0	$\{v v < -c \text{ oder } v > +c\}$	$\{v v < -c^*\}$	$\{v v > +c^*\}$
Nichtablehnungsbereich H_0	$\{v -c \leq v \leq +c\}$	$\{v v \geq -c^*\}$	$\{v v \leq +c^*\}$

Zwei-Stichproben t-Test

1. Voraussetzungen

- zwei Grundgesamtheiten
- erste Grundgesamtheit: Zufallsvariable X_1
 - ▶ $E(X_1) = \mu_1$
 - ▶ $\text{Var}(X_1) = \sigma_1^2$ unbekannt
 - ▶ $X_1 \sim N(\mu_1, \sigma_1)$
- zweite Grundgesamtheit: Zufallsvariable X_2
 - ▶ $E(X_2) = \mu_2$
 - ▶ $\text{Var}(X_2) = \sigma_2^2$ unbekannt
 - ▶ $X_2 \sim N(\mu_2, \sigma_2)$
- μ_1 und μ_2 unbekannt
- einfache Zufallsstichprobe aus jeder Grundgesamtheit mit Stichprobenumfang n_1 bzw. $n_2 \Rightarrow$ Zweistichprobentest
- die beiden Zufallsstichproben sind unabhängig voneinander

2. Hypothesen:

Test	Zweiseitig	Einseitig	
	Linksseitig	Rechtsseitig	
Nullhypothese H_0	$\mu_1 - \mu_2 = \omega_0$	$\mu_1 - \mu_2 \geq \omega_0$	$\mu_1 - \mu_2 \leq \omega_0$
Alternativhypothese H_1	$\mu_1 - \mu_2 \neq \omega_0$	$\mu_1 - \mu_2 < \omega_0$	$\mu_1 - \mu_2 > \omega_0$

3. Verteilung der Teststatistik V unter H_0 :

- ω in der Grundgesamtheit unbekannt
- In der Nullhypothese wird unterstellt, dass der hypothetische Wert ω_0 der wahre Mittelwert in der Grundgesamtheit ist

$$\omega = \omega_0 \quad \Rightarrow E(D) = \omega_0, \quad Var(D) = \sigma_D^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

- σ_1 und σ_2 sind unbekannt $\Rightarrow \sigma_D^2$ muss geschätzt werden

- Schätzfunktionen für σ_1^2 und σ_2^2

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_{1i} - \bar{X}_1)^2$$

$$S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (X_{2i} - \bar{X}_2)^2$$

Annahme der Varianzhomogenität: $\sigma_1^2 = \sigma_2^2$

- Schätzung S^2 für die gemeinsame Varianz σ^2

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

- Schätzfunktion S_D^2 für σ_D^2

$$S_D^2 = S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right) = \frac{n_1 + n_2}{n_1 n_2} \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

- Teststatistik V

$$V = \frac{D - \omega_0}{S_D} = \frac{(\bar{X}_1 - \bar{X}_2) - \omega_0}{\sqrt{\frac{n_1 + n_2}{n_1 n_2} \cdot \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}}$$

- V ist unter H_0 t_f -verteilt mit $f = n_1 + n_2 - 2$

Annahme der Varianzheterogenität: $\sigma_1^2 \neq \sigma_2^2$

- nur Näherungslösung möglich
- Schätzfunktion S_D^2 für σ_D^2

$$S_D^2 = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$$

- Teststatistik

$$V = \frac{D - \omega_0}{S_D} = \frac{(\bar{X}_1 - \bar{X}_2) - \omega_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

- V ist unter H_0 approximativ t_f -verteilt mit

$$f = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{1}{n_1-1} \left(\frac{s_1^2}{n_1} \right)^2 + \frac{1}{n_2-1} \left(\frac{s_2^2}{n_2} \right)^2}$$

- Im Fall der Varianzheterogenität wird der Test auch Welch-Test genannt

4. Berechnung der kritischen Werte:

- Festlegung des Signifikanzniveaus α (Wahrscheinlichkeit eines Fehlers 1. Art)

$$\alpha = P("H_1" \mid H_0)$$

Test	Zweiseitig	Linksseitig	Rechtsseitig
$P(\text{weiß})$	$P(c_u \leq V \leq c_o)$ $= 1 - \alpha$ $c_u = -c, c_o = +c$	$P(c_l \leq V)$ $= 1 - \alpha$ $c_l = -c^*$	$P(V \leq c_r)$ $= 1 - \alpha$ $c_r = c^*$
$P(\text{rot})$	$P(V < -c) + P(c < V)$ $= \alpha/2$	$P(V < c)$ $= \alpha$	$P(c < V)$ $= \alpha$

Test	Zweiseitig	Linksseitig	Rechtsseitig
Kritische Werte	$c = t_{1-\alpha/2;f}$	$c^* = t_{1-\alpha;f}$	
Ablehnungsbereich H_0	$\{v v < -c \text{ oder } v > +c\}$	$\{v v < -c^*\}$	$\{v v > +c^*\}$
Nichtablehnungsbereich H_0	$\{v -c \leq v \leq +c\}$	$\{v v \geq -c^*\}$	$\{v v \leq +c^*\}$

- für $n_1 > 30$ und $n_2 > 30 \rightarrow t_{1-\alpha;f} \approx N(0, 1)$

 - ▶ $t_{1-\alpha/2;f} \approx z_{1-\alpha/2}$
 - ▶ $t_{1-\alpha;f} \approx z_{1-\alpha}$

5. Ziehung einer Zufallsstichprobe und Berechnung des Prüfwertes:

- Ziehen der beiden einfachen Zufallsstichproben vom Umfang n_1 und n_2
- konkrete Stichprobenwerte $x_{11}, \dots, x_{1;n_1}$ und $x_{21}, \dots, x_{2;n_2}$ liegen dann vor
- Schätzwert \bar{x}_1 und \bar{x}_2 für die Stichprobenmittelwerte berechnen
- gegebenfalls Schätzwert s_1 und s_2 für die Standardabweichungen berechnen
- Einsetzen in die erforderliche Teststatistik V führt zu einem Prüfwert v

Testentscheidung und Interpretation

1. Fall

- Testentscheidung: $v \in$ Ablehnungsbereich der $H_0 \rightarrow "H_1"$
- Interpretation:

Es konnte statistisch auf dem Signifikanzniveau α und basierend auf den beiden einfachen Zufallsstichproben vom Umfang n_1 und n_2 gezeigt werden, dass der wahre Erwartungswert $E(D) = \omega$ in der Grundgesamtheit nicht gleich dem hypothetischen Wert ω_0 ist.

- ★ das bedeutet nicht, dass H_1 richtig ist
- ★ mit der Wahrscheinlichkeit α haben wir einen Fehler 1. Art ($"H_1" | H_0$) begangen

2. Fall

- Testentscheidung: $v \in \text{Nichtablehnungsbereich der } H_0 \rightarrow "H_0"$
- Interpretation:

Es konnte statistisch auf dem Signifikanzniveau α und basierend auf den beiden einfachen Zufallsstichproben vom Umfang n_1 und n_2 nicht gezeigt werden, dass der wahre Erwartungswert $E(D) = \omega$ in der Grundgesamtheit nicht gleich dem hypothetischen Wert ω_0 ist.

- ★ das bedeutet nicht, dass H_0 richtig ist
- ★ mit einer unbekannten Wahrscheinlichkeit β haben wir einen Fehler 2. Art (" H_0 "| H_1) begangen

Test der Regressionsparameter

Wahres einfaches lineares Regressionsmodell:

- $y_i = \beta_0 + \beta_1 x_i + u_i$

Wenn $\beta_1 = 0 \Rightarrow$ keine (lineare) Abhangigkeit Y von X

Annahmen:

- $E(U_i) = 0;$
- $Var(U_i) = \sigma_U^2; \quad i, j = 1, \dots, n; i \neq j$
- $Cov(U_i U_j) = \sigma_{ij} = 0$
- $U_i \sim N(0; \sigma_U^2) \quad !!!$

Stichprobenregressionsmodell:

$$y_i = b_0 + b_1 x_i + \hat{u}_i$$

1. Y_i ($i = 1, \dots, n$) \Rightarrow Zufallsvariable;
 Y_i normalverteilt
2. Regressionsparameter sind eine Linearkombination der Zufallsvariablen Y_i .
 - b_0 und b_1 können von Stichprobe zu Stichprobe verschiedene Werte annehmen.
 - Realisationen von B_0 und B_1 sind b_0 und b_1

$$E(B_0) = \beta_0$$

$$E(B_1) = \beta_1$$

$$Var(B_0) = \sigma_{B_0}^2 = \frac{\sigma_u^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2} \quad Var(B_1) = \sigma_{B_1}^2 = \frac{\sigma_u^2}{\sum (x_i - \bar{x})^2}$$

Stichprobenverteilung der Regressionsparameter:

$$B_0 \sim N(\beta_0; \sigma_{B_0}^2) \quad B_1 \sim N(\beta_1; \sigma_{B_1}^2)$$

Test des Regressionskoeffizienten β_1

- Bestimmung der Teststatistik:

$$Z = \frac{B_1 - E(B_1)}{\sigma_{B_1}} = \frac{B_1 - \beta_1}{\sigma_{B_1}} \sim N(0; 1)$$

- Problem:

$$\sigma_{B_1}^2 = \frac{\sigma_U^2}{\sum(x_i - \bar{x})^2} \Rightarrow \text{unbekannt, da } \sigma_U^2 \text{ unbekannt}$$

- Schätzung der Varianz der Residuen:

$$s_u^2 = \frac{\sum \hat{u}_i^2}{n - 2}$$

- Somit neue Teststatistik:

$$T = \frac{B_1 - \beta_1}{\hat{\sigma}_{B_1}} \sim t_{n-2}$$

$$\hat{\sigma}_{B_1}^2 = \frac{s_u^2}{\sum(x_i - \bar{x})^2}$$

1. Hypothesenformulierung: $H_0 : \beta_1 = 0$ $H_1 : \beta_1 \neq 0$

2. Bestimmung der Teststatistik:

$$V = \frac{B_1}{\hat{\sigma}_{B_1}} \quad V \text{ ist unter } H_0 \text{ t-verteilt mit } n - 2 \text{ Freiheitsgraden.}$$

3. Testentscheidung: $|v| > t_{1-\alpha/2; n-2}$ \rightarrow Ablehnung der H_0

Anmerkung:

Für großes n ($n > 30$) kann approximativ die Standardnormalverteilung verwendet werden; \rightarrow kritischer Wert: $z_{1-\alpha/2}$

Fortsetzung des Beispiels Eiscreme

- Geschätzte Regressionsgrade:

$$\hat{y}_i = b_0 + b_1 x_i = 144,93 + 2,65 x_i$$

- Stichprobenumfang: $n = 30$

- Hypothesenformulierung:

- ▶ $H_0 : \beta_1 = 0$
 - ▶ $H_1 : \beta_1 \neq 0$

- Bestimmung der Entscheidungsbereiche:

- ▶ $\alpha = 0,05$
 - ▶ $t_{0,975;28} = 2,048$

- ▶ Ablehnungsbereich der H_0 : $\{v | v < -2,048 \text{ oder } v > 2,048\}$
 - ▶ Nichtablehnungsbereich der H_0 : $\{v | -2,048 \leq v \leq 2,048\}$

- Bestimmung der Teststatistik:

$$\sum \hat{u}_i^2 = 11196,89$$

$$s_u^2 = 11196,89 / 28 = 399,89$$

$$\sum (x_i - \bar{x})^2 = 2413,76$$

$$\hat{\sigma}_{b_1}^2 = \frac{399,89}{2413,76} = 0,1657$$

$$\hat{\sigma}_{b_1} = 0,407$$

$$v = \frac{b_1}{\hat{\sigma}_{b_1}} = \frac{2,65}{0,407} = 6,502$$

- Testentscheidung und Interpretation:

$v \in$ Ablehnungsbereiches der $H_0 \rightarrow H_0$ ablehnen

Fortsetzung des Beispiels Monatsmiete

- Geschätzte Regressionsgrade:

$$\hat{y}_i = b_0 + b_1 x_i = -106,188 + 19,223 x_i$$

- Stichprobenumfang: $n = 815$
- Hypothesenformulierung:
 - ▶ $H_0 : \beta_1 = 0$
 - ▶ $H_1 : \beta_1 \neq 0$
- Bestimmung der Entscheidungsbereiche:
 - ▶ $\alpha = 0,05$
 - ▶ $t_{0,975;813} \approx z_{0,975} = 1,96$
 - ▶ Ablehnungsbereich der H_0 : $\{v | v < -1,96 \text{ oder } v > 1,96\}$
 - ▶ Nichtablehnungsbereich der H_0 : $\{v | -1,96 \leq v \leq 1,96\}$

- Bestimmung der Teststatistik:

$$\sum \hat{u}_i^2 = 129654890$$

$$s_u^2 = 129654890/813 = 159477,11$$

$$\sum (x_i - \bar{x})^2 = 849497$$

$$\hat{\sigma}_{B_1}^2 = 0,1877$$

$$\hat{\sigma}_{B_1} = 0,4332$$

$$v = \frac{b_1}{\hat{\sigma}_{b_1}} = \frac{19,223}{0,4332} = 44,37$$

- Testentscheidung und Interpretation:

$v \in$ Ablehnungsbereiches der $H_0 \rightarrow H_0$ ablehnen

Residuals:

Min	1Q	Median	3Q	Max
-15.168	-3.990	-1.318	2.034	24.500

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	34.55384	0.56263	61.41	<2e-16 ***
lstat	-0.95005	0.03873	-24.53	<2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.216 on 504 degrees of freedom

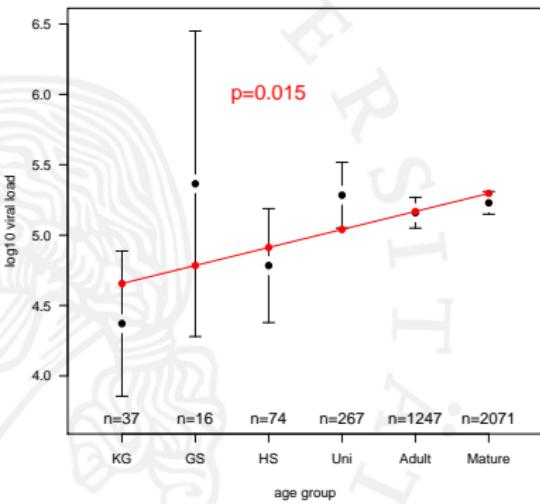
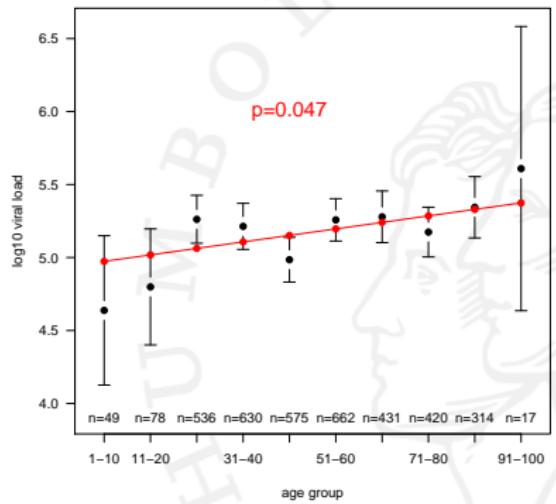
Multiple R-squared: 0.5441, Adjusted R-squared: 0.5432

F-statistic: 601.6 on 1 and 504 DF, p-value: < 2.2e-16

- b_0 und Test $H_0 : \beta_0 = 0$ vs. $H_1 : \beta_0 \neq 0$: (Intercept) Zeile
- b_1 und Test $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$: lstat zeile
- Bestimmtheitsmaß: Multiple R-squared

Drosten vs. Kekulé

- Prof. Drosten publiziert einen Preprint "An analysis of SARS-CoV-2 viral load by patient age" (29.04.2020)
- Aufgrund der Testergebnisse kommt er zu dem Ergebnis, dass es keinen Unterschied gibt zwischen der Viruslast eines Kindes und eines Erwachsenen
- Deswegen empfiehlt er Kindergärten und Schulen weiterhin geschlossen zu halten
- Prof. Held kann mit einem anderen Verfahren einen Unterschied in den Altersgruppen nachweisen
- Prof. Kekulé attackiert Prof. Drosten in einem Zeitungsinterview massiv



Quelle: "A discussion and reanalysis of the results reported in Jones et al. (2020)"
von Leonhard Held (Universität Zürich)

Gütfunktion (Macht) eines Tests

Die Gütfunktion

- ist eine Funktion vom Parameter ϑ
- gibt die Wahrscheinlichkeit der Ablehnung der Nullhypothese in Abhängigkeit von allen unter der Null- und Alternativhypothese zulässigen Parameterwerten an

$$G(\vartheta) = P(V \in \text{Ablehnungsbereich von } H_0 | \vartheta) = P("H_1" | \vartheta)$$

- Gehört der wahre Parameterwert ϑ zu den zulässigen Parameterwerten unter der Nullhypothese H_0 , so wurde im Ergebnis der Testdurchführung eine **falsche** Entscheidung getroffen ($"H_1" | H_0$).
- Die Gütfunktion gibt in diesem Fall die Wahrscheinlichkeit für den Fehler 1. Art an:

$$G(\vartheta) = P("H_1" | H_0) \leq \alpha \text{ für alle } \vartheta \in \Theta_0$$

- Gehört der wahre Parameterwert ϑ zu den zulässigen Parameterwerten unter der Alternativhypothese H_1 , so wurde im Ergebnis der Testdurchführung eine **richtige** Entscheidung getroffen (" H_1 "| H_1).
- Die Gütfunktion gibt in diesem Fall die Wahrscheinlichkeit für die berechtigte Ablehnung der Nullhypothese (berechtigte Annahme der Alternativhypothese) an:

$$G(\vartheta) = P(\text{"}H_1\text{"}|H_1) = 1 - \beta(\vartheta) \text{ für alle } \vartheta \in \Theta_1$$

- Wahrscheinlichkeit für einen Fehler 2. Art:

$$1 - G(\vartheta) = P(\text{"}H_0\text{"}|H_1) = \beta \text{ für alle } \vartheta \in \Theta_1$$

- erlaubt den Vergleich von Teststatistiken V_1 und V_2 für das gleiche Testproblem
- V_1 ist besser als V_2 , wenn für $\vartheta \in \Theta_1$ gilt
 $G_1(\vartheta) \geq G_2(\vartheta) \implies \beta_1 \leq \beta_2$
für mindestens ein $\vartheta \in \Theta_1$ muss gelten $G_1(\vartheta) > G_2(\vartheta)$
- Für $\vartheta \in \Theta_0$ gilt: $G_i(\vartheta) \leq \alpha$

Gütfunktion für Test auf Mittelwert

- hypothetischer Wert μ_0
- Signifikanzniveau α und Stichprobenumfang n vorgegeben
- σ bekannt

Gütfunktion $G(\mu)$ gibt die Wahrscheinlichkeit der Ablehnung von H_0 in Abhängigkeit vom Parameterwert μ an:

$$\begin{aligned} G(\mu) &= P(V \in \text{Ablehnungsbereich der } H_0 | \mu) \\ &= P("H_1" | \mu) \\ &= 1 - P(V \in \text{Nichtablehnungsbereich der } H_0 | \mu) \\ &= 1 - P("H_0" | \mu) \end{aligned}$$

Zweiseitiger Test

Nullhypothese nur wahr, wenn $\mu = \mu_0$

$$G(\mu) = \begin{cases} P("H_1" | H_0) = \alpha & \text{für } \mu = \mu_0 \\ P("H_1" | H_1) = 1 - \beta & \text{für alle } \mu \neq \mu_0 \end{cases}$$

$$G(\mu) = 1 - \left[P\left(V \leq z_{1-\alpha/2} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right) - P\left(V < -z_{1-\alpha/2} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right) \right]$$

Wahrscheinlichkeit für einen Fehler 2. Art:

$$P("H_0" | H_1) = 1 - G(\mu \neq \mu_0) = \beta$$

Herleitung:

$$G(\mu) = 1 - P(\bar{x}_u \leq \bar{X} \leq \bar{x}_o | \mu)$$

$$= 1 - P\left(\frac{\bar{x}_u - \mu}{\sigma/\sqrt{n}} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{\bar{x}_o - \mu}{\sigma/\sqrt{n}}\right)$$

$$\bar{x}_u = \mu_0 - c \frac{\sigma}{\sqrt{n}}$$

$$\bar{x}_o = \mu_0 + c \frac{\sigma}{\sqrt{n}}$$

$$G(\mu) = 1 - P\left(\frac{\mu_0 - c \frac{\sigma}{\sqrt{n}} - \mu}{\sigma/\sqrt{n}} \leq V \leq \frac{\mu_0 + c \frac{\sigma}{\sqrt{n}} - \mu}{\sigma/\sqrt{n}}\right)$$

$$G(\mu) = 1 - P\left(-c - \frac{-\mu - \mu_0}{\sigma/\sqrt{n}} \leq V \leq c - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right)$$

$$= 1 - P\left(-z_{1-\alpha/2} - \frac{-\mu - \mu_0}{\sigma/\sqrt{n}} \leq V \leq z_{1-\alpha/2} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right)$$

$$= \begin{cases} 1 - (1 - \alpha) = \alpha & \text{für } \mu = \mu_0 \\ 1 - \beta & \text{für } \mu \neq \mu_0 \end{cases}$$

$$G(\mu) = 1 - \left[P\left(V \leq z_{1-\alpha/2} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right) - P\left(V < -z_{1-\alpha/2} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right) \right]$$

Beispiel 20.8 (Fortsetzung Beispiel 8.3: Mehltüten)

- $\mu_0 = 1000$
- $\sigma = 10$
- $n = 25$
- $\alpha = 0,05$

Angenommen, der wahre Wert des Parameters ist $\mu = 1002$ g. Wie groß ist die Wahrscheinlichkeit eines Fehlers 2. Art?

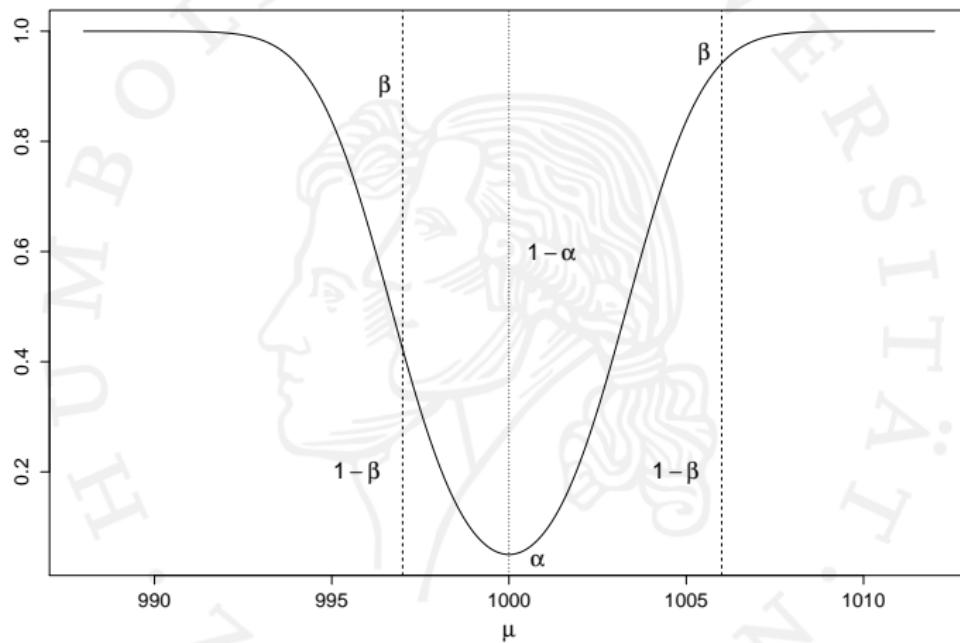
$$G(\mu) = 1 - \left[P\left(V \leq z_{1-\alpha/2} - \frac{\mu-\mu_0}{\sigma/\sqrt{n}}\right) - P\left(V < -z_{1-\alpha/2} - \frac{\mu-\mu_0}{\sigma/\sqrt{n}}\right) \right]$$

$$\begin{aligned}G(1002) &= 1 - [P(V \leq 1,96 - \frac{1002 - 1000}{2}) - P(V < -1,96 - \frac{1002 - 1000}{2})] \\&= 1 - [P(V \leq 0,96) - P(V < -2,96)] \\&= 1 - [P(V \leq 0,96) - (1 - P(V \leq 2,96))] \\&= 1 - [0,831472 - (1 - 0,998462)] \\&= 1 - (0,831472 - 0,001538) \\&= 1 - 0,829934 \\&= 0,170066 = 1 - \beta\end{aligned}$$

$$\Rightarrow P("H_0" | H_1) = \beta(\mu = 1002) = 1 - G(\mu = 1002) = 0,83$$

Wenn das tatsächliche mittlere Ist-Gewicht $\mu = 1002$ g beträgt, wird in rund 83% aller Stichproben vom Umfang $n = 25$ die Abweichung vom Sollgewicht 1000 g durch den Test nicht aufgedeckt.

Die Wahrscheinlichkeit für einen Fehler 2. Art ist sehr hoch, da die Differenz $\mu - \mu_0 = 1002 - 1000$ relativ klein ist.



Angenommen, der wahre Wert des Parameters ist $\mu = 989$ g. Wie groß ist die Wahrscheinlichkeit eines Fehlers 2. Art?

$$\begin{aligned}G(989) &= 1 - [P(V \leq 1,96 - \frac{989 - 1000}{2}) - P(V < -1,96 - \frac{989 - 1000}{2})] \\&= 1 - [P(V \leq 7,46) - P(V < 3,54)] \\&= 1 - (1 - 0,9998) \\&= 1 - 0,0002 = 0,9998 = 1 - \beta \\ \Rightarrow P("H_0" | H_1) &= \beta(\mu = 989) = 1 - G(\mu = 989) = 0,0002\end{aligned}$$

Wenn das tatsächliche durchschnittliche Ist-Gewicht $\mu = 989$ g beträgt, wird in nur rund 0,02% aller Stichproben vom Umfang $n = 25$ die Abweichung vom Sollgewicht 1000 g durch den Test nicht aufgedeckt.

Die Wahrscheinlichkeit für einen Fehler 2. Art ist sehr klein, da die Differenz $\mu - \mu_0 = 989 - 1000$ groß ist.

μ	Gültigkeit von	$G(\mu)$	$1 - G(\mu)$
988,00	H_1	$0,999973 = 1 - \beta$	$0,000027 = \beta$
990,40	H_1	$0,997744 = 1 - \beta$	$0,002256 = \beta$
992,80	H_1	$0,949497 = 1 - \beta$	$0,050503 = \beta$
995,20	H_1	$0,670038 = 1 - \beta$	$0,329962 = \beta$
997,60	H_1	$0,224416 = 1 - \beta$	$0,775584 = \beta$
1000,00	H_0	$0,049996 = \alpha$	$0,950004 = 1 - \alpha$
1002,40	H_1	$0,224416 = 1 - \beta$	$0,775584 = \beta$
1004,80	H_1	$0,670038 = 1 - \beta$	$0,329962 = \beta$
1007,20	H_1	$0,949497 = 1 - \beta$	$0,050503 = \beta$
1009,60	H_1	$0,997744 = 1 - \beta$	$0,002256 = \beta$
1012,00	H_1	$0,999973 = 1 - \beta$	$0,000027 = \beta$

Rechtsseitiger Test

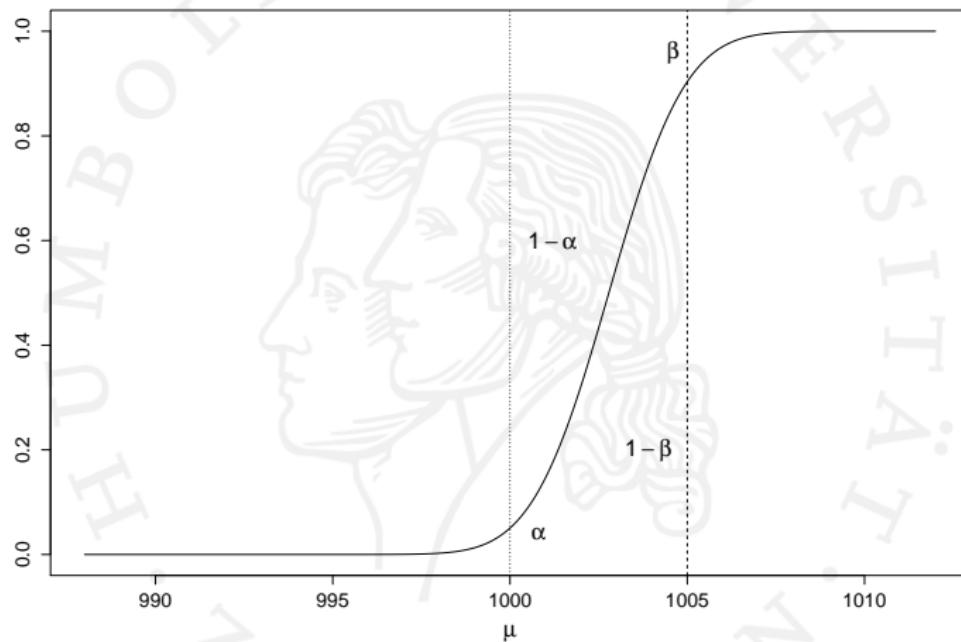
Nullhypothese in Wirklichkeit nur wahr, wenn $\mu \leq \mu_0$

$$G(\mu) = \begin{cases} P("H_1" | H_0) \leq \alpha & \text{für alle } \mu \leq \mu_0 \\ P("H_1" | H_1) = 1 - \beta & \text{für alle } \mu > \mu_0 \end{cases}$$

$$G(\mu) = 1 - P\left(V \leq z_{1-\alpha} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right)$$

Wahrscheinlichkeit für einen Fehler 2. Art:

$$P("H_0" | H_1) = 1 - G(\mu > \mu_0) = \beta$$



Linksseitiger Test

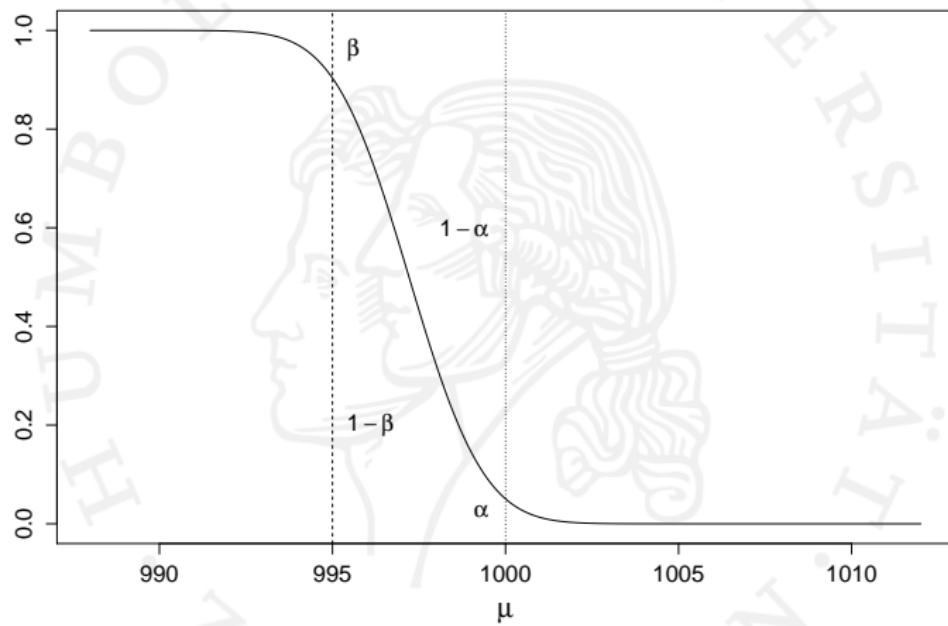
Damit ist also nicht gesagt, dass dieser Unterschied nicht existiert, sondern nur, dass die Daten nicht ausgereicht haben. Nullhypothese nur wahr, wenn $\mu \geq \mu_0$

$$G(\mu) = \begin{cases} P("H_1" | H_0) \leq \alpha & \text{für alle } \mu \geq \mu_0 \\ P("H_1" | H_1) = 1 - \beta & \text{für alle } \mu < \mu_0 \end{cases}$$

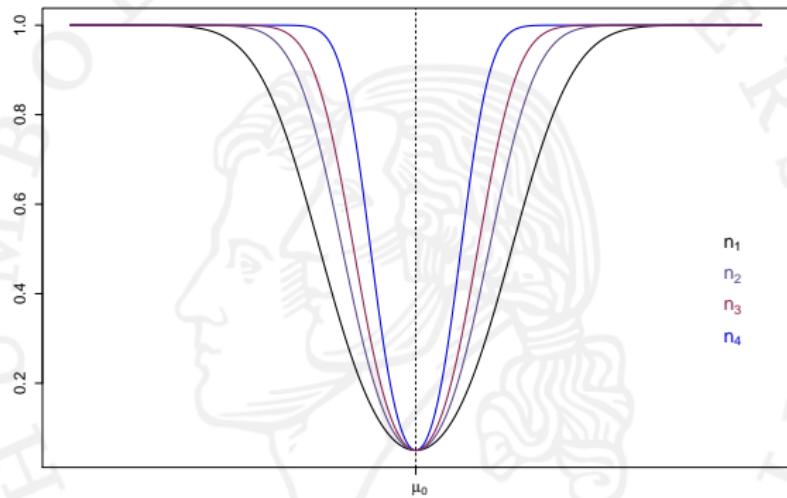
$$G(\mu) = P\left(V \leq -z_{1-\alpha} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right)$$

Wahrscheinlichkeit für einen Fehler 2. Art:

$$P("H_0" | H_1) = 1 - G(\mu < \mu_0) = \beta$$

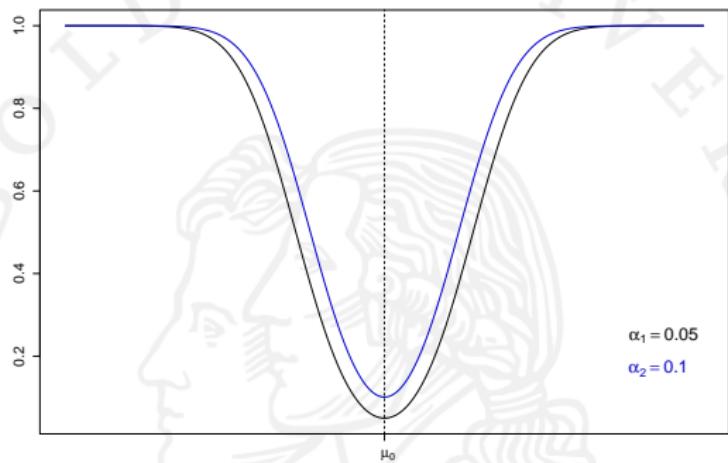


Die Gütfunktion hängt außer an der Stelle $\mu = \mu_0$ vom Stichprobenumfang n ab $\rightarrow n_1 < n_2 < n_3 < n_4$



- Mit Vergrößerung des Stichprobenumfanges n kann unter sonst gleichen Bedingungen die Wahrscheinlichkeit $\beta(\mu)$ für den Fehler 2. Art verringert werden

Gütfunktion hängt vom Signifikanzniveau α ab



- Mit Vergrößerung des Signifikanzniveaus α kann unter sonst gleichen Bedingungen die Wahrscheinlichkeit $\beta(\mu)$ für den Fehler 2. Art verringert werden
- Die beiden Fehlerwahrscheinlichkeiten können nicht gleichzeitig niedrig gehalten werden.

Gütfunktion für Test auf Anteilswert

- hypothetischer Wert π_0
- Signifikanzniveau α und Stichprobenumfang n vorgegeben

Gütfunktion $G(\pi)$ gibt die Wahrscheinlichkeit der Ablehnung von H_0 in Abhängigkeit vom Parameterwert π an:

$$\begin{aligned} G(\pi) &= P(V \in \text{Ablehnungsbereich der } H_0 | \pi) \\ &= P("H_1" | \pi) \\ &= 1 - P(V \in \text{Nichtablehnungsbereich der } H_0 | \pi) \\ &= 1 - P("H_0" | \pi) \end{aligned}$$

Zweiseitiger Test:

$$\begin{aligned}G(\pi) &= P(V < x_u | \pi) + P(V > x_o | \pi) \\&= P(V \leq x_u - 1 | \pi) + [1 - P(V \leq x_o | \pi)]\end{aligned}$$

Rechtsseitiger Test:

$$G(\pi) = P(V > x_c | \pi) = 1 - P(V \leq x_c | \pi)$$

Linksseitiger Test:

$$G(\pi) = P(V < x_c | \pi) = P(V \leq x_c - 1 | \pi)$$

- Wahrscheinlichkeiten aus der Tabelle der Verteilungsfunktion der $B(n; \pi)$
- Gütfunktion an der Stelle $\pi = \pi_0$ entspricht stets dem exakten Signifikanzniveau α_{exakt}

Beispiel 20.9 (Fortsetzung des Beispiel Werfen einer Münze)

- $\pi_0 = 0,5$
- $n = 10$
- $\alpha = 0,05$

$$G(\pi) = P(V = X \in \text{Ablehnungsbereich der } H_0 | \pi)$$

$$\begin{aligned} G(\pi) &= P(V < x_u | \pi) + P(V > x_o | \pi) \\ &= P(V \leq x_u - 1 | \pi) + [1 - P(V \leq x_o | \pi)] \end{aligned}$$

$$x_u = 2$$

$$x_o = 8$$

$$\begin{aligned} G(\pi) &= P(X \leq 1 | \pi) + P(X > 8 | \pi) \\ &= P(X \leq 1 | \pi) + [1 - P(X \leq 8 | \pi)] \\ &= B(1; 10; \pi) + [1 - B(8; 10; \pi)] \end{aligned}$$

- $\pi = 0$

$$G(0) = P("H_1" \mid \pi = 0) = 1$$

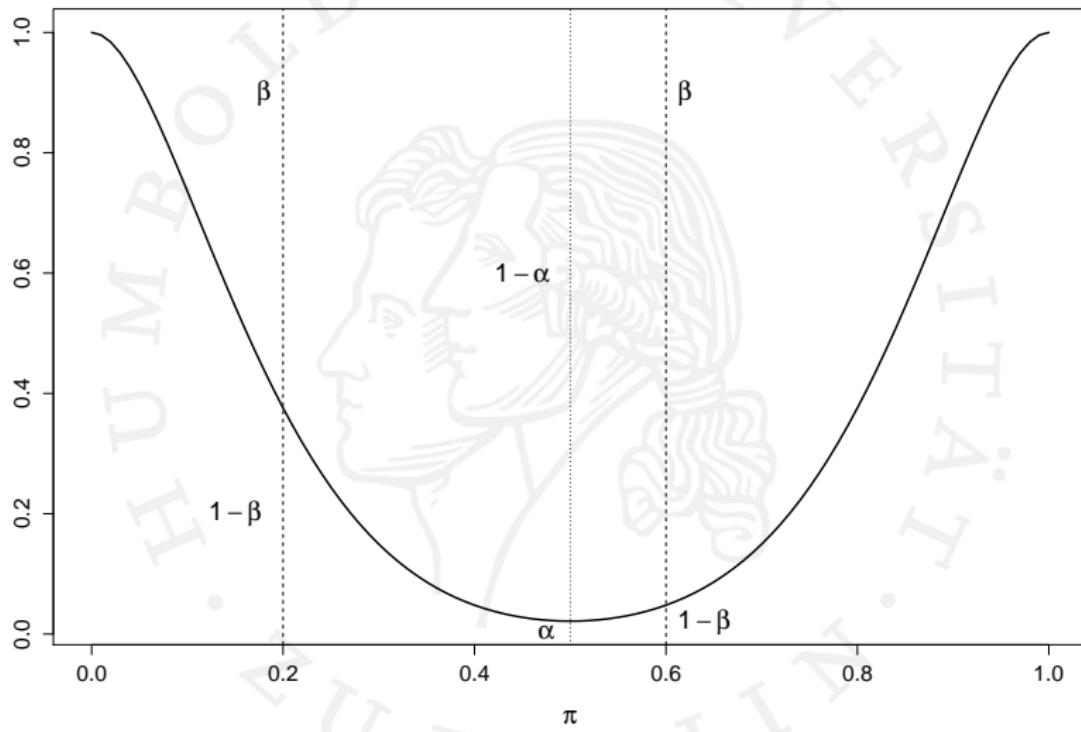
- $\pi = 0, 2$

$$\begin{aligned} G(0, 2) &= B(1; 10; \pi = 0, 2) + [1 - B(8; 10; \pi = 0, 2)] \\ &= 0,3758 + (1 - 1) = 0,3758 \end{aligned}$$

- $\pi = 0, 4$

$$\begin{aligned} G(0, 4) &= B(1; 10; \pi = 0, 4) + [1 - B(8; 10; \pi = 0, 4)] \\ &= 0,0464 + (1 - 0,9983) = 0,0481 \end{aligned}$$

π	$G(\pi)$	$\beta(\pi)$
0	1	0
0,05	0,9139	0,0861
0,1	0,7361	0,2639
0,15	0,5443	0,4547
0,2	0,3758	0,6242
0,25	0,2441	0,7559
0,3	0,1495	0,8505
0,35	0,0865	0,9135
0,4	0,0480	0,9520
0,45	0,0278	0,9722
0,5	0,0214	0,9786



Nichtparametrische Tests

5. November 2022

χ^2 -Anpassungstest • χ^2 -Anpassungstest für $U(a, b)$ • χ^2 -Anpassungstest
für $Po(\lambda)$ • χ^2 -Anpassungstest für $N(\mu; \sigma^2)$ • χ^2 Unabhängigkeitstest

χ^2 -Anpassungstest

- Test einer Hypothese über die unbekannte Verteilung der Zufallsvariablen X in der Grundgesamtheit
- nichtparametrischer Test

1. Voraussetzungen:

- Grundgesamtheit:
 - ▶ Zufallsvariable X mit unbekannter Verteilung $F(x)$
 - ▶ keine Voraussetzungen an das Skalenniveau der Variablen X
 - ▶ hypothetische Verteilung: $F_0(x)$
- Stichprobe vom Umfang n ist eine Zufallsstichprobe
- bei stetigen Variablen:
 - ▶ Klassierung der Beobachtungswerte in disjunkte, aneinander angrenzende Klassen $[x_0^*, x_1^*], [x_1^*, x_2^*], \dots, [x_{i-1}^*, x_i^*], \dots, [x_{I-1}^*, x_I^*]$
 - ▶ Zahl der Klassen: $I (\geq 2)$

Notation

- X - diskrete Zufallsvariable
 - ▶ beobachtete absolute Häufigkeit des Wertes x_j in der Stichprobe:
$$h(x_i) = h_i \quad i = 1, \dots, I$$
 - ▶ Wahrscheinlichkeit aufgrund der hypothetischen Verteilung $F_0(x)$, dass die Zufallsvariable X den Wert x_j annimmt:
$$p_i = P(X = x_i | F_0(x)) \quad i = 1, \dots, I$$
- X - stetige Zufallsvariable
 - ▶ beobachtete absolute Häufigkeit der j -ten Klasse in der Stichprobe:
$$h(x_{i-1}^* < X \leq x_i^*) = h_i \quad i = 1, \dots, I$$
 - ▶ Wahrscheinlichkeit aufgrund der hypothetischen Verteilung $F_0(x)$, dass die Zufallsvariable X einen Wert aus der Klasse $[x_{i-1}^*, x_i^*]$ annimmt:
$$P(x_{i-1}^* < X \leq x_i^*) = p_i \quad i = 1, \dots, I$$
- erwartete absolute Häufigkeit: np_i

2. Hypothesen:

- Nullhypothese H_0 : X folgt der hypothetischen Verteilung $F_0(x)$
- Alternativhypothese H_1 : X folgt nicht der hypothetischen Verteilung $F_0(x)$

Hypothesenpaar lautet konkret:

- $H_0 : h_i = np_i$ für alle $i = 1, \dots, I$
- $H_1 : h_i \neq np_i$ für mindestens ein i
 - ▶ wenn X diskret ist: $p_i = P(X = x_i)$
 - ▶ wenn X stetig ist: $p_i = P(x_{i-1}^* < X \leq x_i^*)$
- Indirekte Prüfung

3. Bestimmung der Teststatistik und ihrer Verteilung unter H_0 :

- Teststatistik:

- ▶ Vergleich der in der Stichprobe beobachteten Verteilung und der bei Gültigkeit von H_0 in der Stichprobe erwarteten Verteilung

$$V = \sum_{i=1}^I \frac{(h_i - n \cdot p_i)^2}{n \cdot p_i}$$

- ▶ V misst die Größe der quadrierten Abweichungen

- Verteilung der Teststatistik unter H_0 :

Bei

- ▶ hinreichend großem Stichprobenumfang n
 - ▶ Einhaltung der Approximationsbedingungen

ist V approximativ χ^2 -verteilt mit $f = I - k - 1$ Freiheitsgraden

- Approximationsbedingungen

- $np_i \geq 1$ für alle $i = 1, \dots, I$
- $np_i \geq 5$ für mindestens 80% der erwarteten absoluten Häufigkeiten

- Anzahl der Freiheitsgrade

- ▶ ein Freiheitsgrad geht grundsätzlich verloren, weil die beobachteten absoluten Häufigkeiten nicht unabhängig voneinander sind. Für vorgegebenen Stichprobenumfang n und aufgrund der Bedingung $\sum_i h_i = n$ folgt, dass jede Häufigkeit h_i durch die anderen $I - 1$ Häufigkeiten bestimmt ist
- ▶ I ist die Anzahl der verbliebenen Werte bzw. Klassen nach einer eventuell notwendigen Zusammenfassung
- ▶ k ist die Anzahl der unbekannten und aus der Stichprobe zu schätzenden Parameter der hypothetischen Verteilung

4. Berechnung der kritischen Werte

- Festlegung des Signifikanzniveaus α (Wahrscheinlichkeit eines Fehlers 1. Art)

$$\alpha = P(\text{"}H_1\text{"}|H_0)$$

- kritischer Wert: $c = \chi^2_{1-\alpha;f}$
→ Quantile der χ^2 -Verteilung siehe Tabelle in der Formelsammlung
- Ablehnungsbereich der H_0 : $\{v \mid v > \chi^2_{1-\alpha;f}\}$
- Nichtablehnungsbereich der H_0 : $\{v \mid v \leq \chi^2_{1-\alpha;f}\}$

5. Ziehung einer Zufallsstichprobe und Berechnung des Prüfwertes:

- absolute Häufigkeiten h_i ermitteln
- gegebenenfalls unbekannte Parameter der hypothetischen Verteilung aus der Stichprobe schätzen
- erwartete Häufigkeiten np_i berechnen
- Approximationsbedingungen überprüfen !!!
- Prüfwert v ermitteln

6. Testentscheidung und Interpretation:

1. Fall

- Testentscheidung: $v \in$ Ablehnungsbereich der $H_0 \rightarrow "H_1"$
- Interpretation:

Es konnte statistisch auf dem Signifikanzniveau α und basierend auf der einfachen Zufallsstichprobe vom Umfang n gezeigt werden, dass die Verteilung der Zufallsvariablen X in der Grundgesamtheit nicht der hypothetischen Verteilung $F_0(x)$ entspricht.

- ★ das bedeutet nicht, dass H_1 richtig ist
- ★ mit der Wahrscheinlichkeit α haben Sie einen Fehler 1. Art begangen

2. Fall

- Testentscheidung: $v \in$ Nichtablehnungsbereich der $H_0 \rightarrow "H_0"$
- Interpretation:

Es konnte statistisch auf dem Signifikanzniveau α und basierend auf der einfachen Zufallsstichprobe vom Umfang n nicht gezeigt werden, dass die Verteilung der Zufallsvariablen X in der Grundgesamtheit von der hypothetischen Verteilung $F_0(x)$ abweicht.

- ★ das bedeutet nicht, dass H_0 richtig ist
- ★ mit einer unbekannten Wahrscheinlichkeit β haben Sie einen Fehler 2. Art begangen

χ^2 -Anpassungstest für $U(a, b)$

Beispiel 21.1

1. Voraussetzungen:

- ▶ X : "Geworfene Augenzahl beim Werfen mit einem Würfel"
- ▶ Wertebereich: $\{1, 2, 3, 4, 5, 6\} \rightarrow X$ ist diskret
- ▶ Verteilung von X unbekannt
- ▶ Zufallsstichprobe vom Umfang $n = 90$

2. Hypothesen:

- ▶ H_0 : X ist diskret gleichverteilt
- ▶ H_1 : X ist nicht diskret gleichverteilt

3. Verteilung der Teststatistik:

- ▶ Anzahl der Merkmalsausprägungen: $I = 6$
- ▶ Anzahl der zu schätzenden Parameter: $k = 0$
- ▶ Approximationsbedingungen

$$p_i = P(X = x_i) = 1/6 \text{ für alle } i$$

$$np_i = 90 \cdot 1/6 = 15 \text{ für alle } i$$

$$np_i \geq 5 \text{ für alle } i \text{ erfüllt}$$

- ▶ Freiheitsgrade
 - ★ $k = 0, I = 6 \Rightarrow f = 6 - 1 = 5$

$$V = \sum_{i=1}^I \frac{(h_i - n \cdot p_i)^2}{n \cdot p_i} \approx \chi_5^2$$

4. Berechnung der kritischen Werte

- ▶ Signifikanzniveau festgelegt mit $\alpha = 0,05$
- ▶ kritischer Wert aus der Chi-Quadrat Verteilung

$$\chi^2_{0,95,5} = 11,07$$

- ▶ Ablehnungsbereich H_0

$$\{v \mid v > 11,07\}$$

- ▶ Nichtablehnungsbereich H_0

$$\{v \mid v \leq 11,07\}$$

5. Ziehung einer einfachen Zufallsstichprobe vom Umfang $n = 90$

Augenzahl	1	2	3	4	5	6
h_i	19	13	14	12	17	15
np_i	15	15	15	15	15	15
$h_i - np_i$	4	-2	-1	-3	2	0
$(h_i - np_i)^2$	16	4	1	9	4	0

- ▶ Prüfwert: $v = \frac{34}{15} = 2,267$

6. Testentscheidung und Interpretation:

- ▶ Testentscheidung: $v \in \text{Nichtablehnungsbereich der } H_0 \Rightarrow "H_0"$
- ▶ Interpretation:

Es konnte statistisch auf dem Signifikanzniveau $\alpha = 0,05$ und basierend auf der einfachen Zufallsstichprobe vom Umfang $n = 90$ nicht gezeigt werden, dass die Verteilung der Zufallsvariablen X in der Grundgesamtheit von einer diskreten Gleichverteilung abweicht.

- ★ das bedeutet nicht, dass H_0 richtig ist
- ★ mit einer unbekannten Wahrscheinlichkeit β haben Sie einen Fehler 2. Art begangen

Beispiel 21.2

- Betriebsprüfung eines Betriebes, zu dem eine Metzgerei (Zerlegebetrieb) und mehrere Schnellimbisse gehörten
- Bons oder Quittungen waren nicht vorhanden, jedoch Kassenbuch mit täglichen Aufzeichnungen (365 Einträge/Jahr)
- U.a. Auswertung der Häufigkeiten der beiden Ziffern vor dem Komma und der ersten Ziffer nach dem Komma
- χ^2 Anpassungstest für die zehn Ziffern unter Annahme der Gleichverteilung: $\chi_{9;0,95}^2 = 16,72$, $\chi_{9;0,99}^2 = 21,67$
- Prüfwerte pro Ziffer: 23,46-46,73 (1998), 29,66-35,65 (1999), 32,95-75,24 (2000)
- Angeklagter wurde zur Steuernachzahlung verurteilt

Beschluss des FG Münster vom 10. November 2003 (Az:6 V 4562/03 E,U)

χ^2 -Anpassungstest für $Po(\lambda)$

Beispiel 21.3

1. Voraussetzungen:

- ▶ X : "Anzahl der Kunden im 5-Minuten-Intervall"
- ▶ X ist diskret
- ▶ Verteilung von X unbekannt
- ▶ $n = 50$ Intervalle

2. Hypothesen:

- ▶ H_0 : X ist poissonverteilt
- ▶ H_1 : X ist nicht poissonverteilt

Für die weitere Vorgehensweise sind Informationen aus der Stichprobe notwendig

3. Ziehung einer einfachen Zufallsstichprobe vom Umfang $n = 50$

i	x_i	h_i	$x_i h_i$	p_i	np_i
1	0	3	0	0,0608	3,040
2	1	6	6	0,1703	8,515
3	2	10	20	0,2384	11,920
4	3	17	51	0,2224	11,120
5	4	7	28	0,1558	7,790
6	5	7	35	0,0872	4,360
7	≥ 6	0	0	0,0651	3,255
\sum		50	140	1,000	50

- Schätzung der Parameter aus Stichprobe:

$$\bar{x} = \hat{\lambda} = \sum_i x_i h_i / n = 140 / 50 = 2,8$$

- Die 1. und 2. Klasse und die 6. und 7. Klasse müssen zusammengefasst werden, da ansonsten die Approximationsbedingungen nicht erfüllt sind

i	x_i	h_i	p_i	np_i	$h_i - np_i$	$(h_i - np_i)^2 / np_i$
1	≤ 1	9	0,2311	11,555	-2,555	0,565
2	2	10	0,2384	11,920	-1,920	0,309
3	3	17	0,2224	11,120	5,880	3,109
4	4	7	0,1558	7,790	-0,790	0,080
5	≥ 5	7	0,1523	7,615	-0,615	0,050
Σ		50	1	50		4,113

- Überprüfung der Approximationsbedingungen:

$$np_i \geq 1 \text{ für alle } i$$

$$np_i \geq 5 \text{ für } 80\% \text{ der } i \text{ erfüllt}$$

- Prüfwert: $v = 4,113$

3. Verteilung der Teststatistik:

- ▶ Anzahl der Klassen: $I = 5$
- ▶ Anzahl der zu schätzenden Parameter: $k = 1$
- ▶ Freiheitsgrade

$$\star \quad k = 1, I = 5 \Rightarrow f = 5 - 1 - 1 = 3$$

$$V = \sum_{i=1}^I \frac{(h_i - n \cdot p_i)^2}{n \cdot p_i} \approx \chi_3^2$$

4. Berechnung der kritischen Werte

- ▶ Signifikanzniveau festgelegt mit $\alpha = 0,01$
- ▶ kritischer Wert aus der Chi-Quadrat Verteilung

$$\chi^2_{0,99;3} = 11,34$$

- ▶ Ablehnungsbereich H_0

$$\{v \mid v > 11,34\}$$

- ▶ Nichtablehnungsbereich H_0

$$\{v \mid v \leq 11,34\}$$

6. Testentscheidung und Interpretation:

- ▶ Testentscheidung: $v \in \text{Nichtablehnungsbereich der } H_0 \Rightarrow "H_0"$
- ▶ Interpretation:

Es konnte statistisch auf dem Signifikanzniveau $\alpha = 0,01$ und basierend auf der einfachen Zufallsstichprobe vom Umfang $n = 50$ nicht gezeigt werden, dass die Verteilung der Zufallsvariablen X in der Grundgesamtheit von einer Poissonverteilung abweicht.

- ★ das bedeutet nicht, dass H_0 richtig ist
- ★ mit einer unbekannten Wahrscheinlichkeit β haben Sie einen Fehler 2. Art begangen

χ^2 -Anpassungstest für $N(\mu; \sigma^2)$

Beispiel 21.4

1. Voraussetzungen:

- ▶ X : „Lebensdauer einer Glühbirne in Jahren“
- ▶ stetige Zufallsvariable → Klassierung erforderlich
- ▶ Verteilung von X unbekannt, aber Vermutung: Normalverteilung ist ein adäquates Verteilungsmodell
- ▶ Zufallsstichprobe vom Umfang $n = 80$

2. Hypothesen:

- ▶ H_0 : X ist normalverteilt
- ▶ H_1 : X ist nicht normalverteilt

Für die weitere Vorgehensweise sind Informationen aus der Stichprobe notwendig

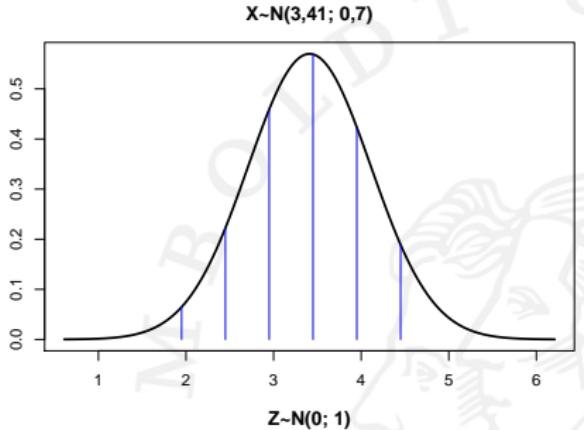
3. Ziehung einer Stichprobe vom Umfang $n = 80$

x_i	h_i	z_i	p_i	np_i	$(h_i - np_i)^2 / np_i$
-1,95	4	-2,09	0,0183	1,464	4,3929
1,95-2,45	2	-1,37	0,0670	5,360	2,1063
2,45-2,95	8	-0,66	0,1693	13,544	2,2693
2,95-3,45	30	0,06	0,2693	21,544	3,3190
3,45-3,95	20	0,77	0,2555	20,440	0,0095
3,95-4,45	10	1,49	0,1525	12,200	0,3967
4,45-	6	∞	0,0681	5,448	0,0559
	80			80	12,5496

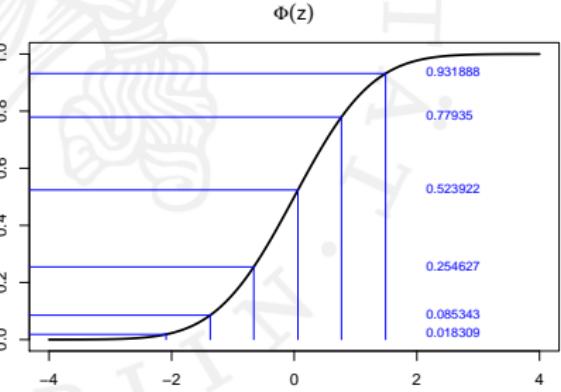
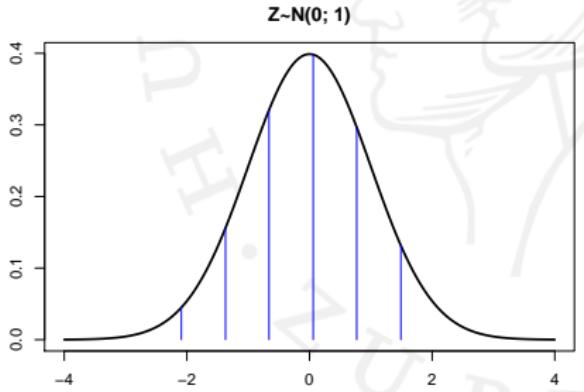
- Parameter aus Stichprobe geschätzt:

$$\bar{x} = 3,41 \text{ Jahre} \quad s = 0,7 \text{ Jahre}$$

- Prüfwert: $v = 12,55$



Klassengrenzen in	x	z	$\Phi(z)$
	1.95	-2.09	0.018309
	2.45	-1.37	0.085343
	2.95	-0.66	0.254627
	3.45	0.06	0.523922
	3.95	0.77	0.77935
	4.45	1.49	0.931888



3. Verteilung der Teststatistik

- ▶ Anzahl der Klassen: $I = 7$
- ▶ Anzahl der zu schätzenden Parameter: $k = 2$
- ▶ Freiheitsgrade

★ $k = 2, I = 7 \Rightarrow f = 7 - 1 - 2 = 4$

$$V = \sum_{i=1}^I \frac{(h_i - n \cdot p_i)^2}{n \cdot p_i} \approx \chi^2_4$$

- ▶ Überprüfung der Approximationsbedingungen:

$$np_i \geq 1 \text{ für alle } i \text{ erfüllt}$$

$$np_i \geq 5 \text{ für 80\% der } i \text{ erfüllt}$$

3. Berechnung der kritischen Werte

- ▶ Signifikanzniveau festgelegt mit $\alpha = 0,05$
- ▶ kritischer Wert aus der Chi-Quadrat Verteilung

$$\chi^2_{0,95;4} = 9,49$$

- ▶ Ablehnungsbereich H_0

$$\{v \mid v > 9,49\}$$

- ▶ Nichtablehnungsbereich H_0

$$\{v \mid v \leq 9,49\}$$

6. Testentscheidung und Interpretation:

- ▶ Testentscheidung: $v \in$ Ablehnungsbereich der $H_0 \Rightarrow "H_1"$
- ▶ Interpretation:

Es konnte statistisch auf dem Signifikanzniveau $\alpha = 0,05$ und basierend auf der einfachen Zufallsstichprobe vom Umfang $n = 80$ gezeigt werden, dass die Verteilung der Zufallsvariablen X in der Grundgesamtheit von einer Normalverteilung abweicht.

- ★ das bedeutet nicht, dass H_1 richtig ist
- ★ mit der Wahrscheinlichkeit α haben Sie einen Fehler 1. Art begangen

χ^2 Unabhängigkeitstest

- Zu prüfen: sind zwei Zufallsvariablen stochastisch unabhängig?
- nichtparametrischer Test

1. Voraussetzungen:

- Grundgesamtheit:
 - ▶ zwei Zufallsvariablen X und Y
 - ▶ keine Voraussetzungen an Skalenniveau der Zufallsvariablen
- einfache Zufallsstichprobe vom Umfang n
- bei stetigen Variablen:
 - ▶ Klassierung der Beobachtungswerte in disjunkte Klassen
 - ▶ Zahl der Klassen: I, J (≥ 2)
 - ▶ Beobachtungswerte bzw. Klassenmitten: x_i ($i = 1, \dots, I$), y_j ($j = 1, \dots, J$)

2. Hypothesen:

- H_0 : X und Y sind stochastisch unabhängig, d.h. $p_{ij} = p_{i\bullet} \cdot p_{\bullet j}$ für alle i,j
- H_1 : X und Y sind stochastisch nicht unabhängig, d.h. $p_{ij} \neq p_{i\bullet} \cdot p_{\bullet j}$ für mindestens ein Paar (i,j)

3. Herleitung der Teststatistik:

Zweidimensionale Häufigkeitstabelle

$X \setminus Y$	y_1	\dots	y_j	\dots	y_J	$h_{i\bullet}$
x_1	h_{11}	\dots	h_{1j}	\dots	h_{1J}	$h_{1\bullet}$
\vdots	\vdots		\vdots		\vdots	\vdots
x_i	h_{i1}	\dots	h_{ij}	\dots	h_{iJ}	$h_{i\bullet}$
\vdots	\vdots		\vdots		\vdots	\vdots
x_I	h_{I1}	\dots	h_{IJ}	\dots	h_{IJ}	$h_{I\bullet}$
$h_{\bullet j}$	$h_{\bullet 1}$	\dots	$h_{\bullet j}$	\dots	$h_{\bullet J}$	$h_{\bullet\bullet} = n$

Zur Erinnerung:

$$h_{i\bullet} = \sum_{j=1}^J h_{ij}; \quad i = 1, \dots, I$$

$$h_{\bullet j} = \sum_{i=1}^I h_{ij}; \quad j = 1, \dots, J$$

$$h_{\bullet\bullet} = \sum_{i=1}^I h_{i\bullet} = \sum_{j=1}^J h_{\bullet j} = \sum_{i=1}^I \sum_{j=1}^J h_{ij} = n$$

- Unter Unabhängigkeit gilt:

$$\begin{aligned} P(\{X = x_i\} \cap \{Y = y_j\}) &= P(X = x_i) \cdot P(Y = y_j) \\ &= p_{i\bullet} \cdot p_{\bullet j} = p_{ij} \end{aligned}$$

- Unter H_0 erwartete absolute Häufigkeiten:

$$e_{ij} = n \cdot p_{ij} = n \cdot p_{i\bullet} \cdot p_{\bullet j}$$

- Geschätzte absolute Häufigkeiten unter Unabhängigkeit:

$$\tilde{h}_{ij} = n \cdot f_{i\bullet} \cdot f_{\bullet j} = n \cdot \frac{h_{i\bullet}}{n} \cdot \frac{h_{\bullet j}}{n} = \frac{h_{i\bullet} \cdot h_{\bullet j}}{n}$$

- Beobachtete absolute Häufigkeiten: h_{ij}
- Teststatistik:

$$V = \sum_{i=1}^I \sum_{j=1}^J \frac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}} \approx \chi_f^2 \text{ unter } H_0$$

- Anzahl der Freiheitsgrade: $f = (I - 1) \cdot (J - 1)$
- Approximationsbedingungen:
 - $\tilde{h}_{ij} \geq 5$ für alle i, j
 - hinreichend großer Stichprobenumfang n

4. Berechnung der kritischen Werte:

- Festlegung des Signifikanzniveaus α
- kritischer Wert: $c = \chi^2_{1-\alpha;f}$
- Ablehnungsbereich der H_0 : $\{v \mid v > \chi^2_{1-\alpha;f}\}$
- Nichtablehnungsbereich der H_0 : $\{v \mid v \leq \chi^2_{1-\alpha;f}\}$

5. Ziehung einer Zufallsstichprobe und Berechnung des Prüfwertes

6. Testentscheidung und Interpretation:

1. Fall

- Testentscheidung: $v \in$ Ablehnungsbereich der $H_0 \rightarrow "H_1"$
- Interpretation:

Es konnte statistisch auf dem Signifikanzniveau α und basierend auf der einfachen Zufallsstichprobe vom Umfang n gezeigt werden, dass die Zufallsvariablen X und Y nicht stochastisch unabhängig voneinander sind.

- ★ das bedeutet nicht, dass H_1 richtig ist
- ★ mit der Wahrscheinlichkeit α haben Sie einen Fehler 1. Art begangen

2. Fall

- Testentscheidung: $v \in$ Nichtablehnungsbereich der $H_0 \rightarrow "H_0"$
- Interpretation:

Es konnte statistisch auf dem Signifikanzniveau α und basierend auf der einfachen Zufallsstichprobe vom Umfang n nicht gezeigt werden, dass die Zufallsvariablen X und Y nicht stochastisch unabhängig voneinander sind.

- ★ das bedeutet nicht, dass H_0 richtig ist
- ★ mit einer unbekannten Wahrscheinlichkeit β haben Sie einen Fehler 2. Art begangen

Beispiel 21.5

1. Voraussetzungen:

- ▶ X : „Geschlecht des Probanden“
- ▶ Y : „Fahrfähigkeit des Probanden“

2. Hypothesen:

- ▶ H_0 : Geschlecht und Fahrfähigkeit des Probanden sind stochastisch unabhängig voneinander
- ▶ H_1 : Geschlecht und Fahrfähigkeit des Probanden sind nicht stochastisch unabhängig voneinander

3. Verteilung der Teststatistik unter H_0 :

$$V = \sum_{i=1}^I \sum_{j=1}^J \frac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}} \approx \chi_f^2$$

- ▶ $f = (2 - 1)(3 - 1) = 2$

4. Berechnung des kritischen Wertes:

- ▶ Signifikanzniveau festgelegt mit $\alpha = 0,05$
- ▶ kritischer Wert aus der Chi-Quadrat Verteilung:

$$c = \chi^2_{0,95;2} = 5,99$$

- ▶ Ablehnungsbereich H_0

$$\{v \mid v > 5,99\}$$

- ▶ Nicht–Ablehnungsbereich H_0

$$\{v \mid v \leq 5,99\}$$

5. Ziehung einer Stichprobe vom Umfang $n = 100$:

Beobachtete Häufigkeiten (Häufigkeiten unter Unabhängigkeit):

$X \setminus Y$	gut	mittel	schlecht	$h_{i\bullet}$
männlich	14 (12)	10 (14)	16 (14)	40
weiblich	16 (18)	25 (21)	19 (21)	60
$h_{\bullet j}$	30	35	35	100

Prüfwert:

$$\nu = \frac{(14 - 12)^2}{12} + \frac{(10 - 14)^2}{14} + \frac{(16 - 14)^2}{14} \\ + \frac{(16 - 18)^2}{18} + \frac{(25 - 21)^2}{21} + \frac{(19 - 21)^2}{21} = 2,94$$

6. Testentscheidung und Interpretation:

- ▶ Testentscheidung: $v \in$ Nichtablehnungsbereich der $H_0 \rightarrow "H_0"$
- ▶ Interpretation:

Es konnte statistisch auf dem Signifikanzniveau $\alpha = 0,05$ und basierend auf der einfachen Zufallsstichprobe vom Umfang $n = 100$ nicht gezeigt werden, dass das Geschlecht eines Probanden und seine Fahrfähigkeit stochastisch abhängig voneinander sind.

- ★ das bedeutet nicht, dass H_0 richtig ist
- ★ mit einer unbekannten Wahrscheinlichkeit β haben Sie einen Fehler 2. Art begangen