Statistik Workshop

Mini-Einführung und Auffrischung zu einigen Teilen der angewandten Statistik

12. und 14. Januar 2015

Prof. Dr. Stefan Etschberger

Outline

Material zur Vorlesung

Kursmaterial:

- Handout der Folien
- Alle Folien inklusive Anmerkungen (am Abend)
- Beispieldaten
- Alle Auswertungen als R-Datei

Literatur:

Bamberg, Günter, Franz Baur und Michael Krapp (2011). **Statistik**. 16. Aufl. München: Oldenbourg Verlag. ISBN: 3486702580.

Dalgaard, Peter (2002). **Introductory Statistics with R**. New York: Springer.

Fahrmeir, Ludwig, Rita Künstler, Iris Pigeot und Gerhard Tutz (2009). **Statistik: Der Weg zur Datenanalyse**. 7. Aufl. Berlin, Heidelberg: Springer. ISBN: 3642019382.

Datenbasis

Fragebogen

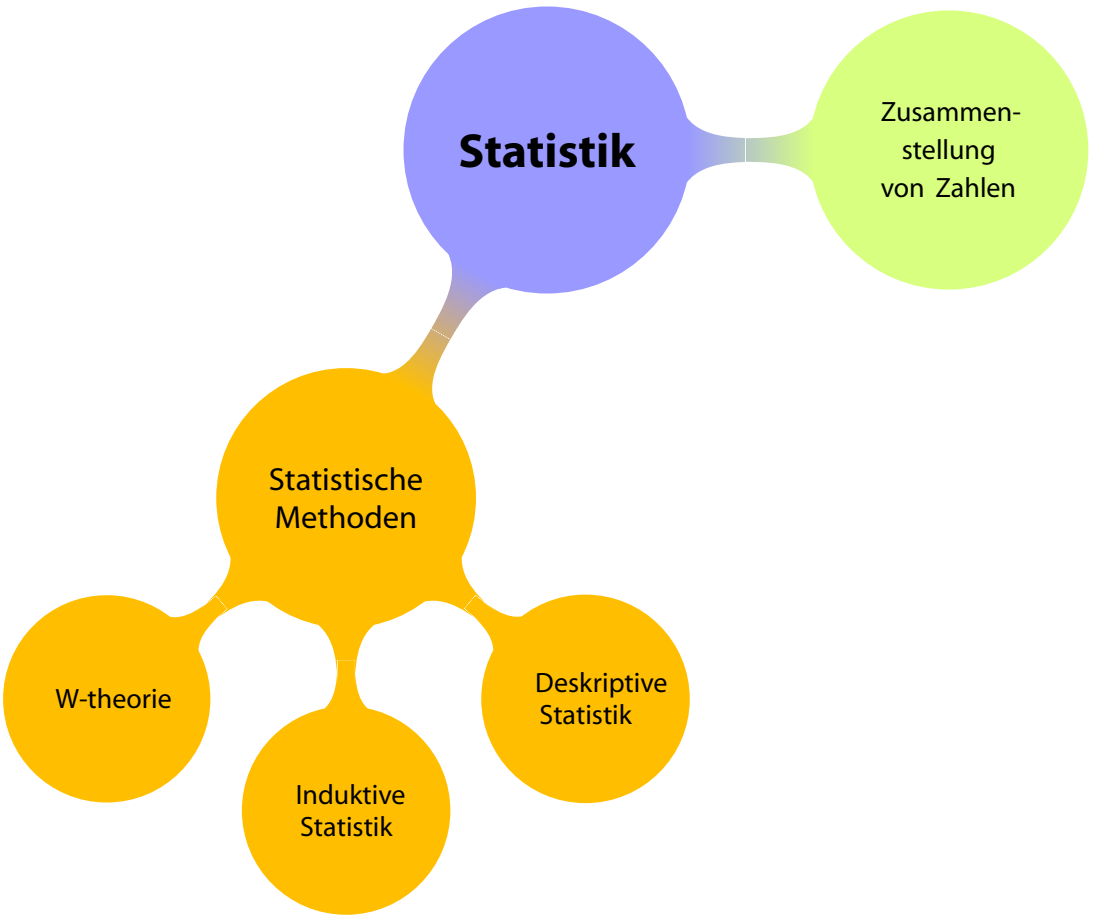
Umfrage Statistik Sommersemester 2014

Bitte beantworten Sie folgende Fragen vollständig und füllen Sie jeweils *nur eine* beliebige Spalte *leserlich* aus.

Ihr Alter (in Jahren)										
Ihre Größe (in cm)										
Ihr Geschlecht (m/w)										
Wie alt ist (bzw. wäre) Ihr Vater heute?										
Wie alt ist (bzw. wäre) Ihre Mutter heute?										
Größe Ihres Vaters (cm)										
Größe Ihrer Mutter (cm)										
Wie viele Geschwister haben Sie?										
Wunschfarbe für Ihr nächstes Smartphone; mögliche Auswahl: (si)lber, (sc)hwarz, (w)eiß, (g)elb, (b)lau, (r)ot										
Ausgaben für Ihre mobile Kommunikation (egal wer bezahlt hat) in den vergangenen 12 Monaten inkl. Hardware (Smartphone, Zubehör), Software (Apps), Vertrag										
Wie viele Paar Schuhe besitzen Sie?										
Ausgaben für Ihre Schuhe (egal wer bezahlt hat) in den letzten 12 Monaten										
Ihre Note in der Matheklausur										
Waren Sie mir Ihrer Leistung in Mathe zufrieden? Antworten: (s)ehr zufrieden, (z)ufrieden, (g)eht so, (n)icht zufrieden										

Statistik: Table of Contents

Bedeutungen des Begriffs Statistik



Einfaches Beispiel

Beispiel

12 Beschäftigte werden nach der Entfernung zum Arbeitsplatz (in km) befragt.
Antworten: 4, 11, 1, 3, 5, 4, 20, 4, 6, 16, 10, 6

- deskriptiv:
  - Durchschnittliche Entfernung: 7,5
  - Klassenbildung:

Klasse	$[0;5)$	$[5;15)$	$[15;30)$
Häufigkeit	5	5	2

- induktiv:
  - Schätze die mittlere Entfernung **aller** Beschäftigten.
  - Prüfe, ob die mittlere Entfernung geringer als 10 km ist.

Merkmale

Statistik
Etschberger -
Januar 2015

1. Einführung
Fehler durch Statistik
Gute und schlechte Grafiken
Begriff Statistik
Grundbegriffe der Datenerhebung
R und RStudio
2. Deskriptive Statistik
3. W-Theorie
4. Induktive Statistik
5. Datenanalyse

- Merkmalsträger: Untersuchte statistische Einheit
- Merkmal: Interessierende Eigenschaft des Merkmalträgers
- (Merkmals-)Ausprägung: Konkret beobachteter Wert des Merkmals
- Grundgesamtheit: Menge aller relevanten Merkmalsträger
- Typen von Merkmalen:

a) qualitativ – quantitativ
   - qualitativ: z.B. Geschlecht
   - quantitativ: z.B. Schuhgröße
   - Qualitative Merkmale sind quantifizierbar (z.B.: weiblich 1, männlich 0)

b) diskret – stetig
   - diskret: Abzählbar viele unterschiedliche Ausprägungen
   - stetig: Alle Zwischenwerte realisierbar

Skalenniveaus

Nominalskala:

- Zahlen haben nur Bezeichnungsfunktion
- z.B. Artikelnummern

Ordinalskala:

- zusätzlich Rangbildung möglich
- z.B. Schulnoten
- Differenzen sind aber **nicht** interpretierbar!
  ⇛ Addition usw. ist unzulässig.

Kardinalskala:

- zusätzlich Differenzbildung sinnvoll
- z.B. Gewinn
- Noch feinere Unterscheidung in: Absolutskala, Verhältnisskala, Intervallskala

Skalendegression und Skalenprogression

Ziel der Skalierung: Gegebene Information angemessen abbilden, möglichst ohne Über- bzw. Unterschätzungen

Es gilt:

- Grundsätzlich können alle Merkmale nominal skaliert werden.
- Grundsätzlich kann jedes metrische Merkmal ordinal skaliert werden.

Das nennt man Skalendegression. Dabei: Informationsverlust

Aber:

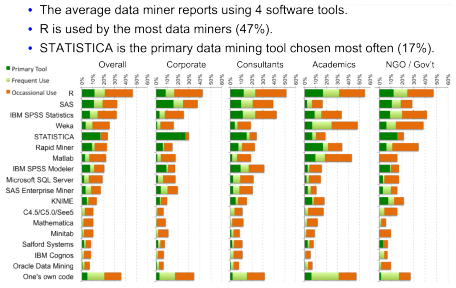
- Nominale Merkmale dürfen **nicht** ordinal- oder metrisch skaliert werden.
- Ordinale Merkmale dürfen **nicht** metrisch skaliert werden.

Das nennt nennt man Skalenprogression. Dabei: Interpretation von **mehr Informationen** in die Merkmale, als inhaltlich vertretbar.
(Gefahr der **Fehlinterpretation**)

Was ist R und warum soll man es benutzen?

- R ist ein freies Softwarepaket zu Statistik und Datenanalyse
- R ist sehr mächtig und weit verbreitet in Wissenschaft und Industrie (sogar von mehr Leuten benutzt als z.B. SPSS)
- Ursprung von R: 1993 an der Universität Auckland von Ross Ihaka and Robert Gentleman entwickelt
- Seitdem: Viele Leute haben R verbessert mit tausenden von Paketen für viele Anwendungen
- Nachteil (auf den ersten Blick): Kein point und click tool
- Großer Vorteil (auf den zweiten Blick): Kein point und click tool

graphics source: `http://goo.gl/W70kms`

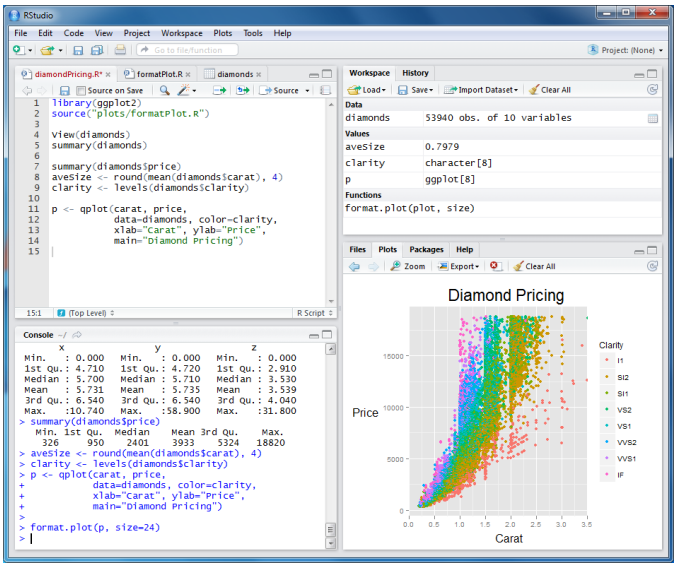


source: `http://goo.gl/axhGhh`

Whas ist RStudio?

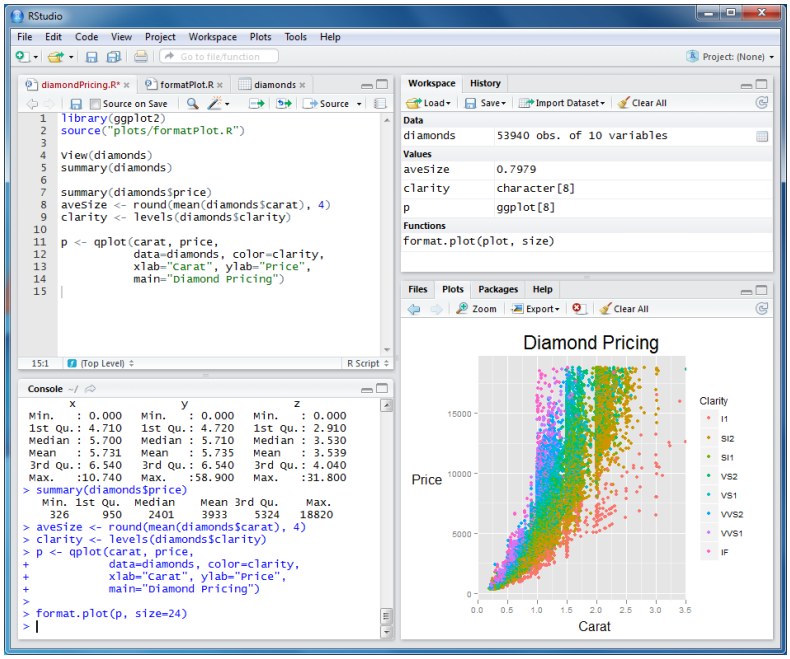
- RStudio ist ein Integrated Development Environment (IDE) um R leichter benutzen zu können.
- Gibt's für OSX, Linux und Windows
- Ist auch frei
- Trotzdem: Sie müssen Kommandos schreiben
- Aber: RStudio unterstützt Sie dabei
- **Download:** `RStudio.com`

RStudio™
Free & Open-Source IDE for R



RStudio Kennenlernen

- Code
- Console
- Workspace
- History
- Files
- Plots
- Packages
- Help
- Auto-Completion
- Data Import



Statistik
Etschberger - Januar 2015

1. Einführung
Fehler durch Statistik
Gute und schlechte Grafiken
Begriff Statistik
Grundbegriffe der Datenerhebung
R und RStudio
2. Deskriptive Statistik
3. W-Theorie
4. Induktive Statistik
5. Datenanalyse

```
# Arbeitsverzeichnis setzen (alternativ über Menü)
setwd("C:/ste/work/vorlesungen/2014WS_Doktorandenworkshop/2015_01_Statistik_Workshop")

# Daten einlesen aus einer csv-Datei (Excel)
MyData = read.csv2(file="../Daten/Umfrage_HSA_2014_03.csv", header=TRUE)
```

```
# inspect structure of data
str(MyData)

## 'data.frame': 205 obs. of  10 variables:
##  $ Alter      : int  21 20 19 20 20 24 20 27 23 21 ...
##  $ Geschlecht : Factor w/ 2 levels "Frau","Mann": 1 1 1 1 1 2 1 1 2 2 ...
##  $ AlterV     : int  54 57 49 45 43 54 49 53 52 55 ...
##  $ AlterM     : int  51 57 58 49 42 52 53 53 48 55 ...
##  $ Geschwister: int  1 0 3 3 5 2 2 1 2 1 ...
##  $ Farbe      : Factor w/ 6 levels "blau","gelb",..: 6 6 4 4 6 4 3 6 4 6 ...
##  $ AusgSchuhe : int  50 500 400 100 450 90 250 200 300 200 ...
##  $ AnzSchuhe  : int  17 22 15 15 22 8 20 10 3 7 ...
##  $ AusgKomm   : num  156 450 240 35.8 450 250 100 300 450 1300 ...
##  $ MatheZufr  : Ord.factor w/ 4 levels "nicht"<"geht so"<..: 1 4 4 4 4 2 1 1 3 3 ...
```

Erste Zeilen der Datentabelle


```
# Erste Zeilen in Datentabelle
head(MyData, 6)
```

```
##   Alter Geschlecht AlterV AlterM Geschwister   Farbe AusgSchuhe AnzSchuhe AusgKomm MatheZufr
## 1    21       Frau     54     51           1   weiss         50        17    156.0     nicht
## 2    20       Frau     57     57           0   weiss        500        22    450.0      sehr
## 3    19       Frau     49     58           3 schwarz        400        15    240.0      sehr
## 4    20       Frau     45     49           3 schwarz        100        15     35.8      sehr
## 5    20       Frau     43     42           5   weiss        450        22    450.0      sehr
## 6    24       Mann     54     52           2 schwarz         90         8    250.0   geht so
```

```
# lege MyData als den "Standard"-Datensatz fest
attach(MyData)
```

```
# Wie Viele Objekte gibt's im Datensatz?
nrow(MyData)
```

```
## [1] 205
```

```
# Wie Viele Merkmale?
ncol(MyData)
```

```
## [1] 10
```


Daten kennenlernen

Statistik
Etschberger -
Januar 2015

1. Einführung
Fehler durch Statistik
Gute und schlechte Grafiken
Begriff Statistik
Grundbegriffe der Datenerhebung
R und RStudio
2. Deskriptive Statistik
3. W-Theorie
4. Induktive Statistik
5. Datenanalyse

```
# Auswahl spezieller Objekte und Merkmale über [Zeile, Spalte]
MyData[1:3, 2:5]
```

```
##   Geschlecht AlterV AlterM Geschwister
## 1       Frau     54     51           1
## 2       Frau     57     57           0
## 3       Frau     49     58           3
```

```
# Auswahl von Objekten über logische Ausdrücke
head(Geschlecht=="Frau" & Alter<19, 30)
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [17] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
# Einsetzen in Klammern und Ausgabe von Alter des Studenten, seines Vaters und seiner Mutter
MyData[Geschlecht=="Frau" & Alter<19, # Objektauswahl
      c("Alter", "AlterM", "AlterV")  # Welche Merkmale anzeigen?
      ]
```

```
##     Alter AlterM AlterV
## 23     18     50     52
## 44     18     37     43
## 51     18     51     54
## 57     18     53     57
## 74     18     53     49
## 126    18     44     45
## 139    18     51     58
## 185    18     46     48
## 193    18     49     47
```

Daten kennenlernen

```
# Zeige die Männer, die mehr als 1000 Euro für Schuhe
# und Mobilfunk zusammen ausgegeben haben
MyData[Geschlecht=="Mann" & AusgSchuhe + AusgKomm > 1000,
       c("Alter", "Geschwister", "Farbe", "AusgSchuhe", "AusgKomm")]
```

```
##     Alter Geschwister   Farbe AusgSchuhe AusgKomm
## 10     21           1   weiss        200     1300
## 15     20           1     rot        400      815
## 26     20           1 schwarz        200     1250
## 40     21           0  silber        300      825
## 87     20           1    blau       1000      350
## 113    25           0 schwarz        280     1200
## 146    24           1 schwarz        300      900
## 177    19           2 schwarz        500      720
## 178    23           1 schwarz        450      630
## 192    20           0 schwarz        400      950
```


Statistik: Table of Contents

Auswertungsmethoden für eindimensionales Datenmaterial

- Merkmal X wird an n Merkmalsträgern beobachtet ➠ **Urliste** $(x_1, \ldots, x_n)$
  Im Beispiel: $x_1 = 4,\ x_2 = 11, \ldots, x_{12} = 6$
- Urlisten sind oft unübersichtlich, z.B.:

```
##  [1] 4 5 4 1 5 4 3 4 5 6 6 5 5 4 7 4 6 5 6 4 5 4 7 5 5 6 7 3 7 6 6 7 4 5 4 7 7 5 5 5 6 6 4 5 2 5 4
## [49] 7 5
```

- Dann zweckmäßig: Häufigkeitsverteilungen

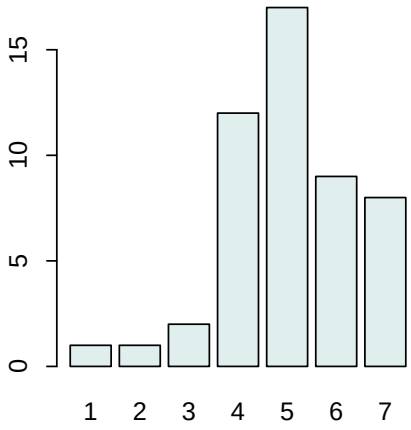
Ausprägung (sortiert)	a_j	1	2	3	4	5	6	7	$\sum$
absolute Häufigkeit	$h(a_j) = h_j$	1	1	2	12	17	9	8	50
kumulierte abs. H.	$H(a_j) = \sum_{i=1}^{j} h(a_i)$	1	2	4	16	33	42	50	—
relative Häufigkeit	$f(a_j) = h(a_j)/n$	$\frac{1}{50}$	$\frac{1}{50}$	$\frac{2}{50}$	$\frac{12}{50}$	$\frac{17}{50}$	$\frac{9}{50}$	$\frac{8}{50}$	1
kumulierte rel. H.	$F(a_j) = \sum_{i=1}^{j} f(a_i)$	$\frac{1}{50}$	$\frac{2}{50}$	$\frac{4}{50}$	$\frac{16}{50}$	$\frac{33}{50}$	$\frac{42}{50}$	1	—

❶ Balkendiagramm

```
table(x)

## x
##  1  2  3  4  5  6  7
##  1  1  2 12 17  9  8
```

```
barplot(table(x), col="azure2")
```



(Höhe proportional zu Häufigkeit)

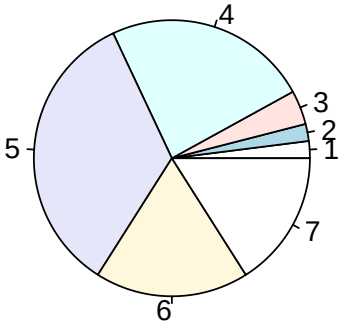
❷ Kreissektorendiagramm

Winkel: $w_j = 360° \cdot f(a_j)$

z.B. $w_1 = 360° \cdot \frac{1}{50} = 7{,}2°$

$w_7 = 360° \cdot \frac{8}{50} = 57{,}6°$

```
pie(table(x))
```



(Fläche proportional zu Häufigkeit)

❸ Histogramm

- für klassierte Daten
- Fläche proportional zu Häufigkeit:

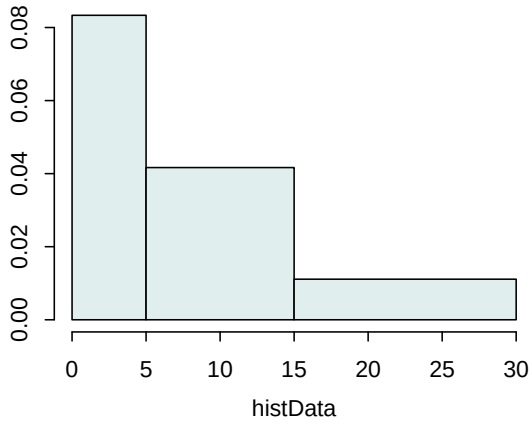
$$\text{Höhe}_j \cdot \text{Breite}_j = c \cdot h(a_j)$$

$$\Rightarrow \quad \text{Höhe}_j = c \cdot \frac{h(a_j)}{\text{Breite}_j}$$

- Im Beispiel mit $c = \frac{1}{12}$:

Klasse	$[0;5)$	$[5;15)$	$[15;30]$
$h(a_j)$	5	5	2
Breite_j	5	10	15
Höhe_j	$\frac{1}{12}$	$\frac{1}{24}$	$\frac{1}{90}$

```
histData <- c(0,1,2,3,4,
              5,6,7,10,14,
              15,30)
truehist(histData,
         breaks=c(0, 4.999, 14.999, 30),
         col="azure2", ylab='')
```



Modus x_{Mod}: häufigster Wert
Beispiel:

a_j	1	2	4
$h(a_j)$	4	3	1

$\Rightarrow x_{\text{Mod}} = 1$

Sinnvoll bei allen Skalenniveaus.

Median x_{Med}: ‚mittlerer Wert', d.h.

1. Urliste aufsteigend sortieren: $x_1 \leqq x_2 \leqq \cdots \leqq x_n$
2. Dann

$$x_{\text{Med}} \begin{cases} = x_{\frac{n+1}{2}}, & \text{falls } n \text{ ungerade} \\ \in [x_{\frac{n}{2}}; x_{\frac{n}{2}+1}], & \text{falls } n \text{ gerade (meist } x_{\text{Med}} = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right)) \end{cases}$$

Im Beispiel oben:
$1,\ 1,\ 1,\ 1,\ 2,\ 2,\ 2,\ 4 \Rightarrow x_{\text{Med}} \in [1;2]$, z.B. $x_{\text{Med}} = 1{,}5$

Sinnvoll ab ordinalem Skalenniveau.

Lageparameter (2)

- Arithmetisches Mittel $\bar{x}$: Durchschnitt, d.h.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{j=1}^{k} a_j \cdot h(a_j)$$

Im Beispiel:

$$\bar{x} = \tfrac{1}{8} \cdot (\underbrace{1+1+1+1}_{1\cdot 4} + \underbrace{2+2+2}_{2\cdot 3} + \underbrace{4}_{4\cdot 1}) = 1{,}75$$

Sinnvoll nur bei kardinalem Skalenniveau.
Bei klassierten Daten:

$$\bar{x}^* = \tfrac{1}{n} \sum \text{Klassenmitte} \cdot \text{Klassenhäufigkeit}$$

Im Beispiel:

$$\bar{x}^* = \tfrac{1}{12} \cdot (2{,}5 \cdot 5 + 10 \cdot 5 + 22{,}5 \cdot 2) = 8{,}96 \neq 7{,}5 = \bar{x}$$

Streuungsparameter

- Voraussetzung: kardinale Werte $x_1, \ldots, x_n$
- **Beispiel:**

a)	x_i	1950	2000	2050	} je $\bar{x} = 2000$
b)	x_i	0	0	6000	}

- Spannweite: $SP = \max_i x_i - \min_i x_i$

Im Beispiel:

a) $SP = 2050 - 1950 = 100$
b) $SP = 6000 - 0 \quad\;\; = 6000$

- Mittlere quadratische Abweichung:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \underbrace{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2}_{\text{Verschiebungssatz}}$$

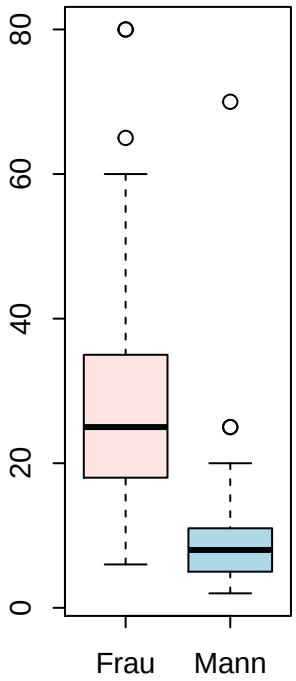
Streuungsparameter (2)

- **Mittlere quadratische Abweichung** im Beispiel:
  a) $s^2 = \frac{1}{3} \cdot (50^2 + 0^2 + 50^2)$
  $= \frac{1}{3} \cdot (1950^2 + 2000^2 + 2050^2) - 2000^2 = 1666{,}67$
  b) $s^2 = \frac{1}{3} \cdot (2000^2 + 2000^2 + 4000^2)$
  $= \frac{1}{3} \cdot (0^2 + 0^2 + 6000^2) - 2000^2 = 8000000$
- **Standardabweichung:** $s = \sqrt{s^2}$
  Im Beispiel:
  a) $s = \sqrt{1666{,}67} = 40{,}82$
  b) $s = \sqrt{8000000} = 2828{,}43$
- **Variationskoeffizient:** $V = \frac{s}{\bar{x}}$ (maßstabsunabhängig)
  Im Beispiel:
  a) $V = \frac{40{,}82}{2000} = 0{,}02\ (\widehat{=} 2\,\%)$
  b) $V = \frac{2828{,}43}{2000} = 1{,}41\ (\widehat{=} 141\,\%)$

Lage und Streuung als Grafik: Boxplot

- Graphische Darstellung von Lage und Streuung
- Box: Oberer/Unterer Rand: 3. bzw. 1. Quartil,
- Linie in Mitte: Median
- Whiskers: Länge: Max./Min Wert, aber beschränkt durch das 1,5-fache des Quartilsabstands (falls größter/kleinster Wert größeren/kleineren Abstand von Box: Länge Whiskers durch größten/kleinsten Wert innerhalb dieser Schranken)
- Ausreißer: Alle Objekte außerhalb der Whisker-Grenzen

```
boxplot(AnzSchuhe ~ Geschlecht,
        col=c("mistyrose", "lightblue"),
        data=MyData, main="")
```



„Wieviel Paar Schuhe besitzen Sie?“

Statistik
Etschberger - Januar 2015

1. Einführung
2. Deskriptive Statistik
Häufigkeiten
Lage und Streuung
Konzentration
Zwei Merkmale
Korrelation
Lineare Regression
3. W-Theorie
4. Induktive Statistik
5. Datenanalyse

Dateninspektion: Überblick über alle Variablen

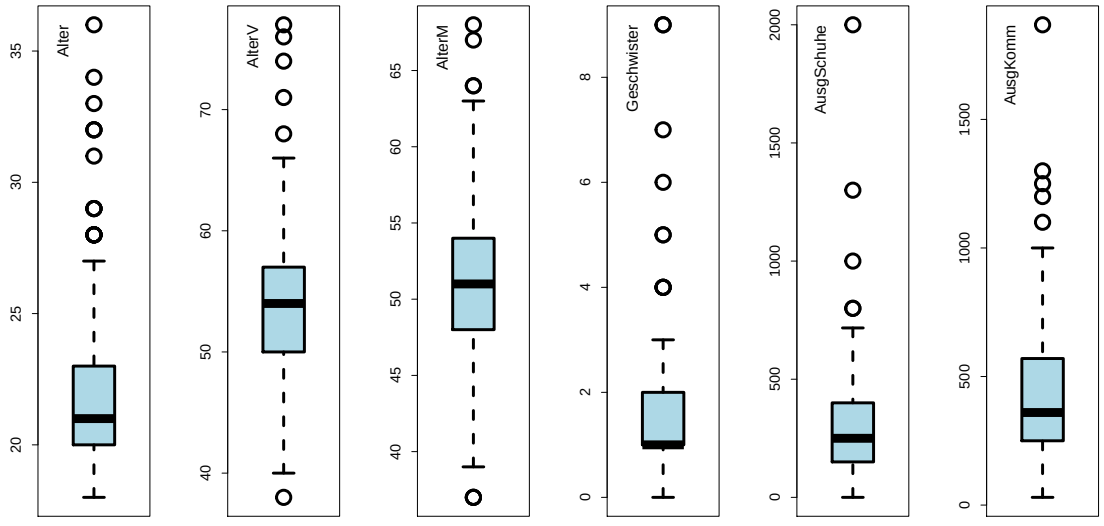
```
summary(MyData)
```

```
##     Alter       Geschlecht    AlterV          AlterM     Geschwister       Farbe
##  Min.   :18.00   Frau:134   Min.   :38.00   Min.   :37.0   Min.   :0.000   blau   :11
##  1st Qu.:20.00   Mann: 71   1st Qu.:50.00   1st Qu.:48.0   1st Qu.:1.000   gelb   : 4
##  Median :21.00              Median :54.00   Median :51.0   Median :1.000   rot    :13
##  Mean   :22.22              Mean   :53.95   Mean   :51.5   Mean   :1.473   schwarz:97
##  3rd Qu.:23.00              3rd Qu.:57.00   3rd Qu.:54.0   3rd Qu.:2.000   silber :17
##  Max.   :36.00              Max.   :77.00   Max.   :68.0   Max.   :9.000   weiss  :63
##    AusgSchuhe       AnzSchuhe        AusgKomm          MatheZufr
##  Min.   :   0.0   Min.   : 2.00   Min.   :  30.0   nicht    :68
##  1st Qu.: 150.0   1st Qu.:10.00   1st Qu.: 250.0   geht so  :47
##  Median : 250.0   Median :20.00   Median : 360.0   zufrieden:43
##  Mean   : 296.6   Mean   :21.58   Mean   : 429.4   sehr     :26
##  3rd Qu.: 400.0   3rd Qu.:30.00   3rd Qu.: 570.0   NA's     :21
##  Max.   :2000.0   Max.   :80.00   Max.   :1868.0
```


Dateninspektion

Boxplots

```
for(attribute in c("Alter", "AlterV", "AlterM", "Geschwister",
                   "AusgSchuhe", "AusgKomm")) {
  data=MyData[, attribute]
  boxplot(data, # all rows, column of attribute
          col="lightblue",      # fill color
          lwd=3,                # line width
          cex=2,                 # character size
          oma=c(1,1,2,1)
          )
  text(0.7,max(data), attribute, srt=90, adj=1)
}
```



Statistik
Etschberger - Januar 2015

1. Einführung
2. Deskriptive Statistik
Häufigkeiten
Lage und Streuung
Konzentration
Zwei Merkmale
Korrelation
Lineare Regression
3. W-Theorie
4. Induktive Statistik
5. Datenanalyse

Konzentrationsmaße

- Gegeben: kardinale Werte $0 \leqq x_1 \leqq x_2 \leqq \cdots \leqq x_n$
- **Achtung!** Die Werte müssen aufsteigend sortiert werden!
- Lorenzkurve:

> Wieviel Prozent der Merkmalssumme entfällt auf die x Prozent kleinsten Merkmalsträger?

- **Beispiel:** Die 90 % ärmsten besitzen 20 % des Gesamtvermögens.
- Streckenzug: $(0,0),\ (u_1,v_1),\ \ldots,\ (u_n,v_n) = (1,1)$ mit

$$v_k = \text{Anteil der } k \text{ kleinsten MM-Träger an der MM-Summe} = \frac{\sum_{i=1}^{k} x_i}{\sum_{i=1}^{n} x_i}$$

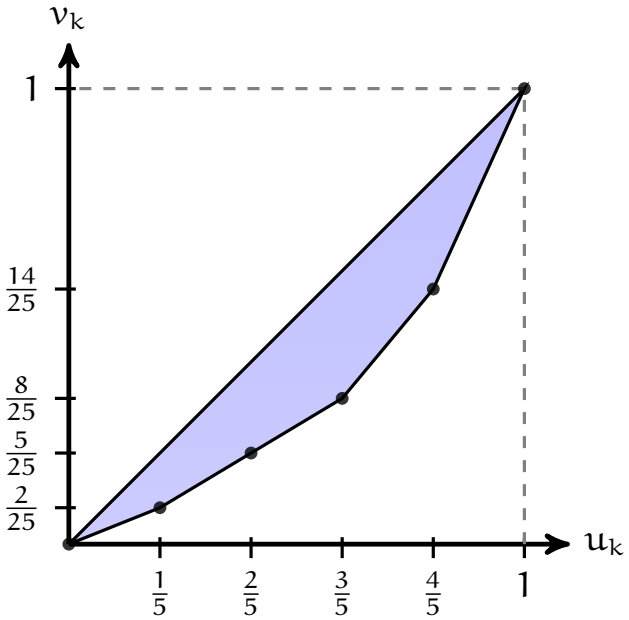
$$u_k = \text{Anteil der } k \text{ kleinsten an der Gesamtzahl der MM-Träger} = \frac{k}{n}$$

Lorenzkurve: Beispiel

Markt mit fünf Unternehmen; Umsätze: 6, 3, 11, 2, 3 (Mio. €)

$\Rightarrow n = 5,\ \sum_{k=1}^{5} x_k = 25$

k	1	2	3	4	5
x_k	2	3	3	6	11
p_k	$\frac{2}{25}$	$\frac{3}{25}$	$\frac{3}{25}$	$\frac{6}{25}$	$\frac{11}{25}$
v_k	$\frac{2}{25}$	$\frac{5}{25}$	$\frac{8}{25}$	$\frac{14}{25}$	1
u_k	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{3}{5}$	$\frac{4}{5}$	1



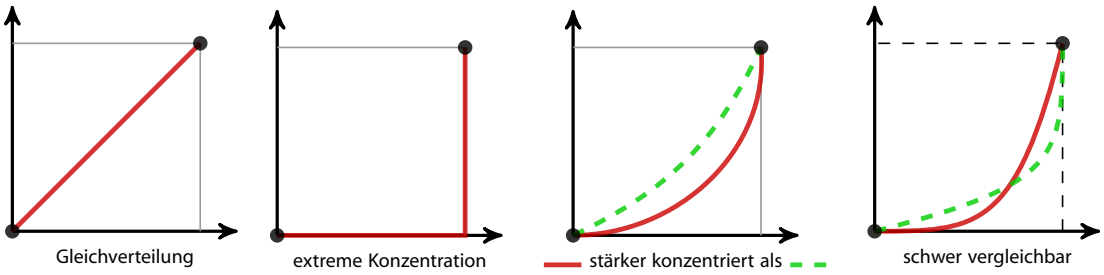
Lorenzkurve

Knickstellen:

- Bei i-tem Merkmalsträger $\iff x_{i+1} > x_i$
- Empirische Verteilungsfunktion liefert Knickstellen:

a_j	2	3	6	11
$h(a_j)$	1	2	1	1
$f(a_j)$	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{1}{5}$
$F(a_j)$	$\frac{1}{5}$	$\frac{3}{5}$	$\frac{4}{5}$	1

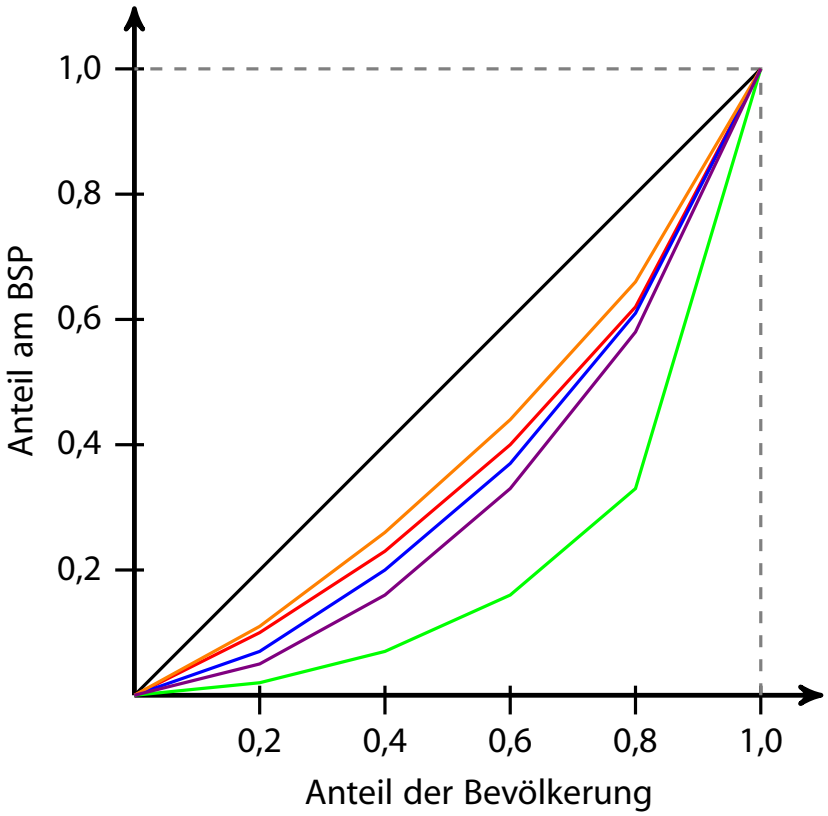
Vergleich von Lorenzkurven:



Lorenzkurve: Beispiel Bevölkerungsanteil gegen BSP

Bangladesch
Brasilien
Deutschland
Ungarn
USA

(Stand 2000)



Gini-Koeffizient

- Numerisches Maß der Konzentration: Gini-Koeffizient G

$$G = \frac{\text{Fläche zwischen } 45°\text{-Linie und L}}{\text{Fläche unter } 45°\text{-Linie}} = \frac{\text{◢}}{\text{◢}}$$

- Aus den Daten:

$$G = \frac{2\sum_{i=1}^{n} i\,x_i - (n+1)\sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i} = \frac{2\sum_{i=1}^{n} i\,p_i - (n+1)}{n} \quad \text{wobei} \quad p_i = \frac{x_i}{\sum_{i=1}^{n} x_i}$$

- Problem: $G_{max} = \frac{n-1}{n}$

➠ Normierter Gini-Koeffizient:

$$G_* = \frac{n}{n-1} \cdot G \in [0;1]$$

Gini-Koeffizient: Beispiel

Beispiel:

i	1	2	3	4	$\sum$
x_i	1	2	2	15	20
p_i	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{2}{20}$	$\frac{15}{20}$	1

$$G = \frac{2 \cdot \left(1 \cdot \frac{1}{20} + 2 \cdot \frac{2}{20} + 3 \cdot \frac{2}{20} + 4 \cdot \frac{15}{20}\right) - (4+1)}{4} = 0{,}525$$

Mit $G_{max} = \frac{4-1}{4} = 0{,}75$ folgt

$$G_* = \frac{4}{4-1} \cdot 0{,}525 = 0{,}7$$

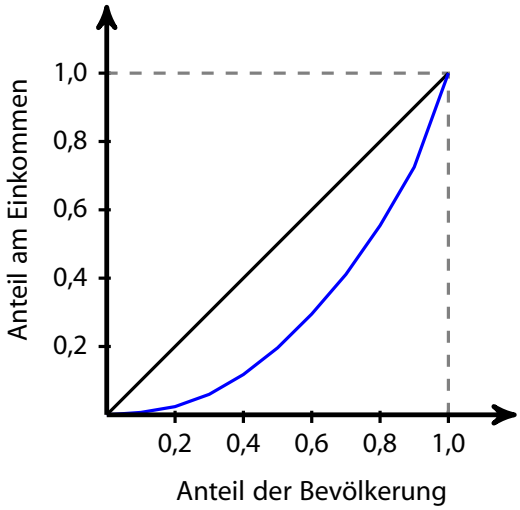
Statistik
Etschberger - Januar 2015

1. Einführung
2. Deskriptive Statistik
Häufigkeiten
Lage und Streuung
Konzentration
Zwei Merkmale
Korrelation
Lineare Regression
3. W-Theorie
4. Induktive Statistik
5. Datenanalyse

Konzentrationsmaße: Beispiel

Armutsbericht der Bundesregierung 2008

- Verteilung der Bruttoeinkommen in Preisen von 2000
- aus unselbständiger Arbeit der Arbeitnehmer/-innen insgesamt

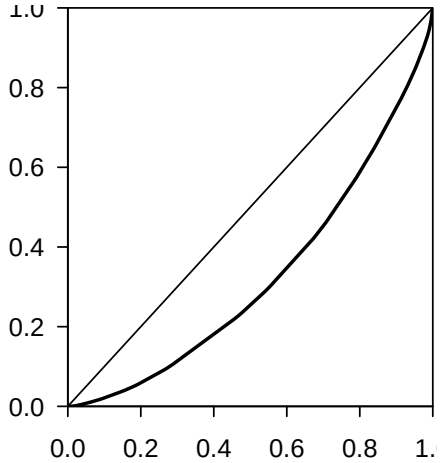


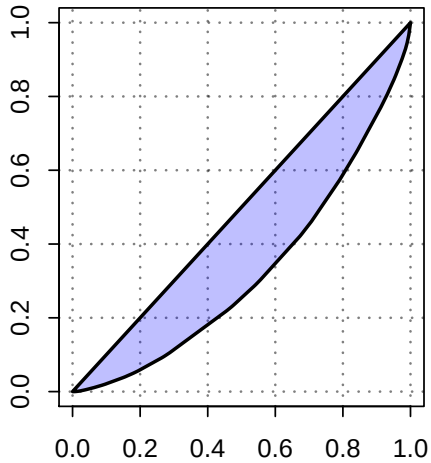
	2002	2003	2004	2005
Arithmetisches Mittel	24.873	24.563	23.987	23.648
Median	21.857	21.531	20.438	20.089
Gini-Koeffizient	0,433	0,441	0,448	0,453

Lorenzkurve mit R

```
require(ineq) # inequality Paket
Lorenz = Lc(AusgSchuhe)
plot(Lorenz, xlab="", ylab="", main="") # Standard plot

plot(c(0,1), c(0,1), type="n", # bisschen netter
    panel.first=grid(lwd=1.5, col=rgb(0,0,0,1/2)),
    xlab="", main="", ylab="")
polygon(Lorenz$p, Lorenz$L, density=-1, col=rgb(0,0,1,1/4), lwd=2)
```






```
Gini(AusgSchuhe) # Gini-Koeffizient

## [1] 0.3556353
```

Weitere Konzentrationsmaße

- Konzentrationskoeffizient:

$$CR_g = \text{Anteil, der auf die } g \text{ größten entfällt} = \sum_{i=n-g+1}^{n} p_i = 1 - \nu_{n-g}$$

- Herfindahl-Index:

$$H = \sum_{i=1}^{n} p_i^2 \qquad (\in [\tfrac{1}{n}; 1])$$

Es gilt: $H = \frac{1}{n}(V^2 + 1)$ bzw. $V = \sqrt{n \cdot H - 1}$

- Exponentialindex:

$$E = \prod_{i=1}^{n} p_i^{p_i} \qquad \left(\in [\tfrac{1}{n}; 1]\right) \qquad \text{wobei} \qquad 0^0 = 1$$

- Im Beispiel mit $x = (1, 2, 2, 15)$:

$$CR_2 = \tfrac{17}{20} = 0{,}85$$
$$H = \left(\frac{1}{20}\right)^2 + \cdots + \left(\frac{15}{20}\right)^2 = 0{,}59$$
$$E = \left(\frac{1}{20}\right)^{\frac{1}{20}} \cdots \left(\frac{15}{20}\right)^{\frac{15}{20}} = 0{,}44$$

Auswertungsmethoden für zweidimensionale Daten

Zweidimensionale Urliste

Urliste vom Umfang n zu **zwei** Merkmalen X und Y:

$$(x_1, y_1), \quad (x_2, y_2), \quad \ldots, \quad (x_n, y_n)$$

Kontingenztabelle:

Sinnvoll bei wenigen Ausprägungen bzw. bei klassierten Daten.

	Ausprägungen von Y			
Ausprägungen von X	b_1	b_2	$\ldots$	b_l
a_1	h_{11}	h_{12}	$\ldots$	h_{1l}
a_2	h_{21}	h_{22}	$\ldots$	h_{2l}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
a_k	h_{k1}	h_{k2}	$\ldots$	h_{kl}

Kontingenztabelle

Unterscheide:

- **Gemeinsame Häufigkeiten:**

$$h_{ij} = h(a_i, b_j)$$

- **Randhäufigkeiten:**

$$h_{i\cdot} = \sum_{j=1}^{l} h_{ij} \qquad \text{und} \qquad h_{\cdot j} = \sum_{i=1}^{k} h_{ij}$$

- **Bedingte (relative) Häufigkeiten:**

$$f_1(a_i \mid b_j) = \frac{h_{ij}}{h_{\cdot j}} \qquad \text{und} \qquad f_2(b_j \mid a_i) = \frac{h_{ij}}{h_{i\cdot}}$$

Häufigkeiten

Beispiel: 400 unfallbeteiligte Autoinsassen:

	leicht verletzt $(= b_1)$	schwer verletzt $(= b_2)$	tot $(= b_3)$	
angegurtet $(= a_1)$	264 $(= h_{11})$	90 $(= h_{12})$	6 $(= h_{13})$	360 $(= h_{1\cdot})$
nicht angegurtet $(= a_2)$	2 $(= h_{21})$	34 $(= h_{22})$	4 $(= h_{23})$	40 $(= h_{2\cdot})$
	266 $(= h_{\cdot 1})$	124 $(= h_{\cdot 2})$	10 $(= h_{\cdot 3})$	400 $(= n)$

$f_2(b_3 \mid a_2) = \frac{4}{40} = 0{,}1$ (10 % der nicht angegurteten starben.)

$f_1(a_2 \mid b_3) = \frac{4}{10} = 0{,}4$ (40 % der Todesopfer waren nicht angegurtet.)

Streuungsdiagramm

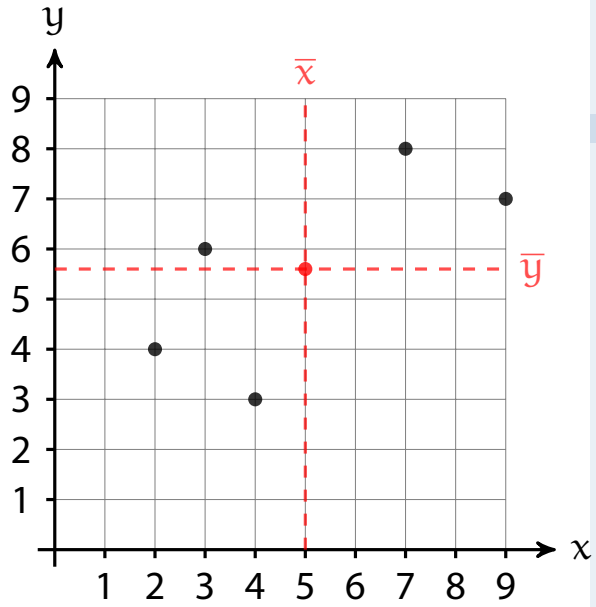
Streuungsdiagramm sinnvoll bei vielen verschiedenen Ausprägungen (z.B. stetige Merkmale)

➠ Alle (x_i, y_i) sowie $(\bar{x}, \bar{y})$ in Koordinatensystem eintragen.

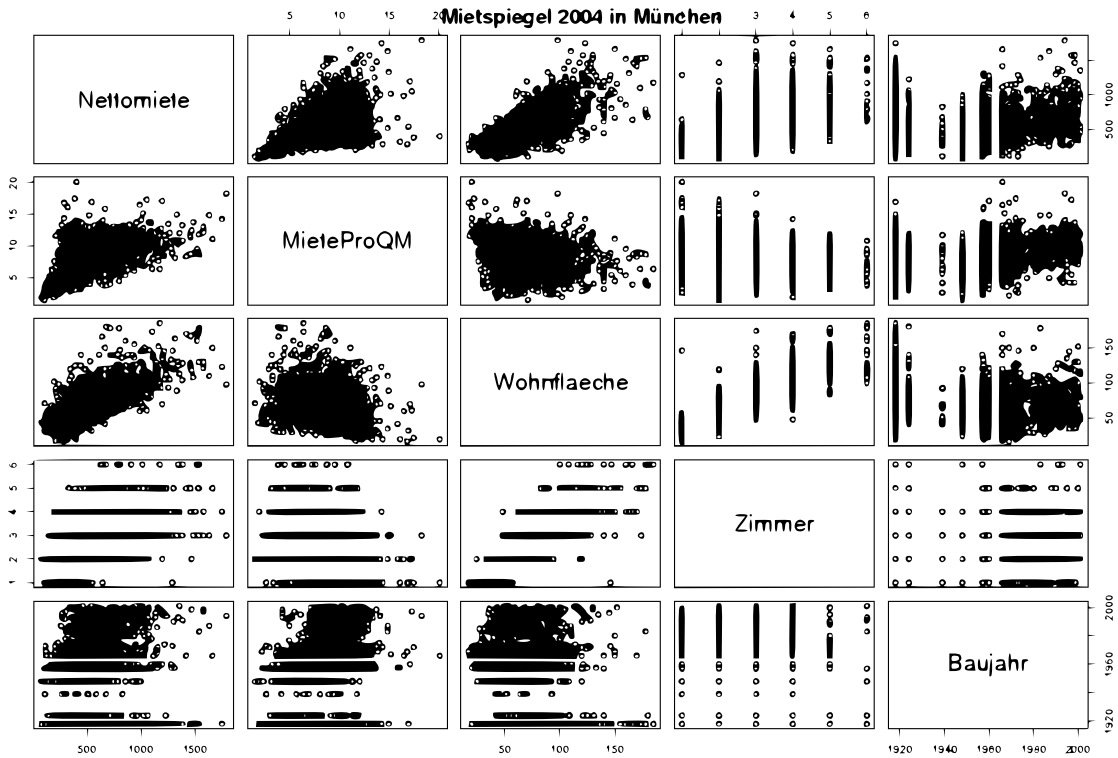
Beispiel:

i	1	2	3	4	5	$\sum$
x_i	2	4	3	9	7	25
y_i	4	3	6	7	8	28

$$\Rightarrow \bar{x} = \frac{25}{5} = 5$$
$$\bar{y} = \frac{28}{5} = 5{,}6$$



Beispiel Streuungsdiagramm

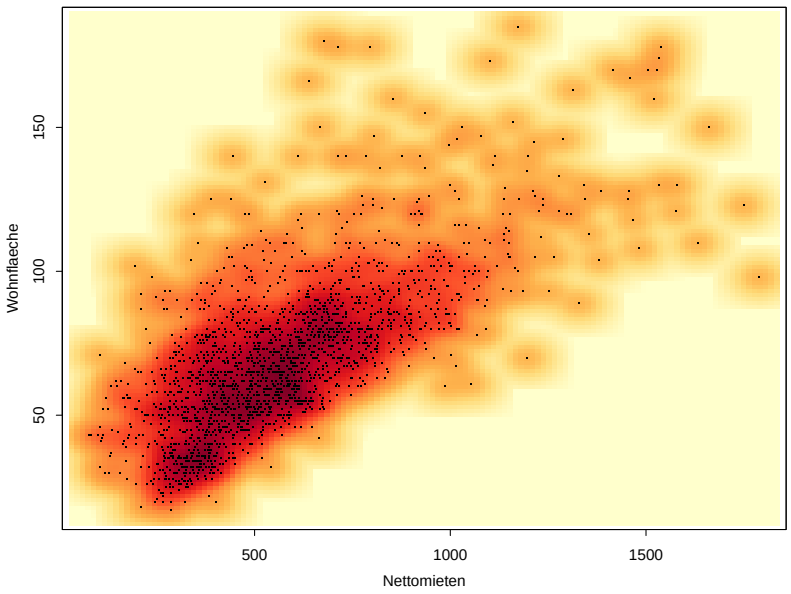


(Datenquelle: Fahrmeir u. a. (2009))

Beispiel Streuungsdiagramm

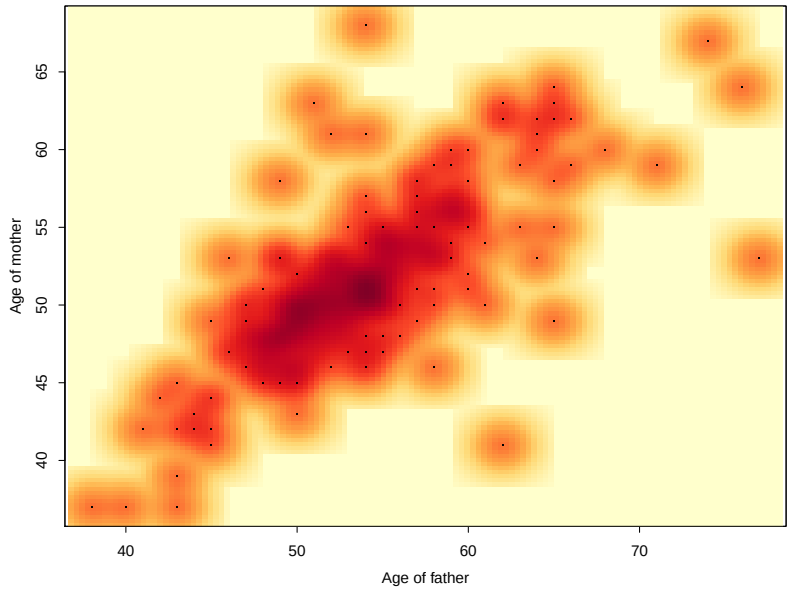
```
mieten <- read.table('../../_data/MietenMuenchen.csv',
                     header=TRUE, sep='\t',
                     check.names=TRUE, fill=TRUE,
                     na.strings=c('',''))
x   <- cbind(Nettomieten=mieten$nm, Wohnflaeche=mieten$wfl)

library("geneplotter")  ## from BioConductor
smoothScatter(x, nrpoints=Inf,
              colramp=colorRampPalette(brewer.pal(9,"YlOrRd")),
              bandwidth=c(30,3))
```



Beispiel Streuungsdiagramm

```
x = cbind("Age of father"=AlterV, "Age of mother"=AlterM)
require("geneplotter") ## from BioConductor
smoothScatter(x, colramp=colorRampPalette(brewer.pal(9,"YlOrRd")) )
```



Statistik
Etschberger -
Januar 2015

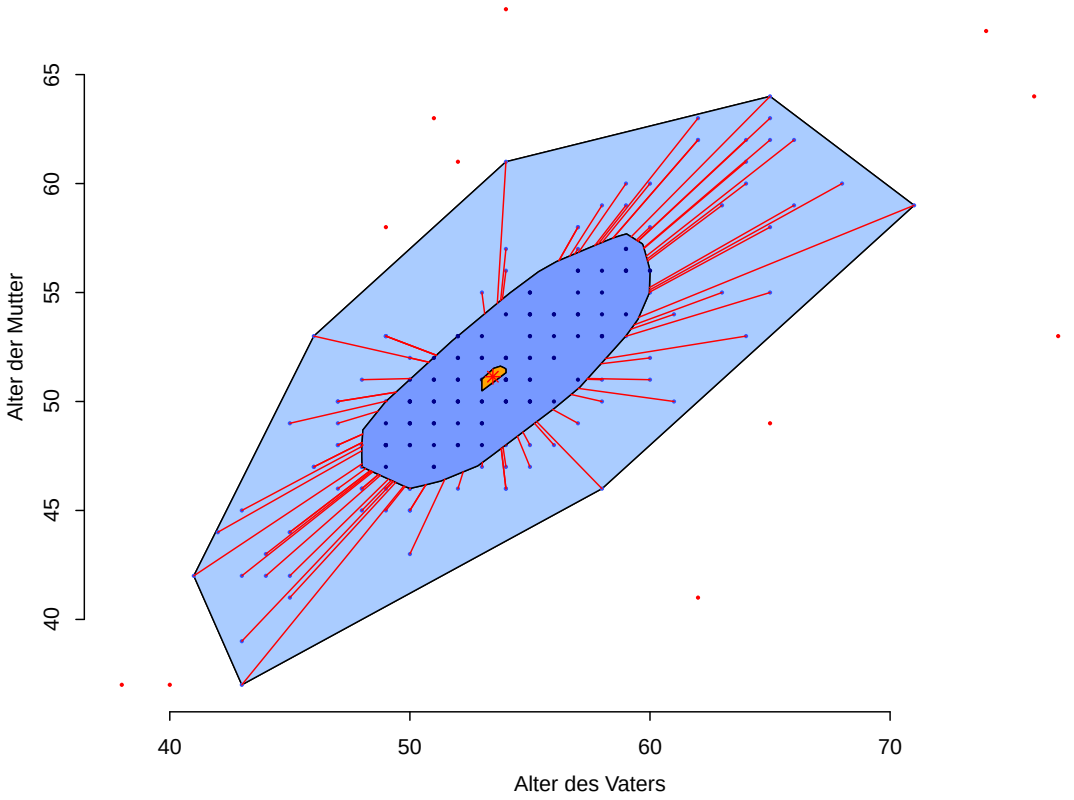
1. Einführung
2. Deskriptive Statistik
Häufigkeiten
Lage und Streuung
Konzentration
Zwei Merkmale
Korrelation
Lineare Regression
3. W-Theorie
4. Induktive Statistik
5. Datenanalyse

Visualisiere Paare

```
require(GGally)
ggpairs(MyData[, -c(5, 6, 10)], colour='Geschlecht', alpha=0.4)
```

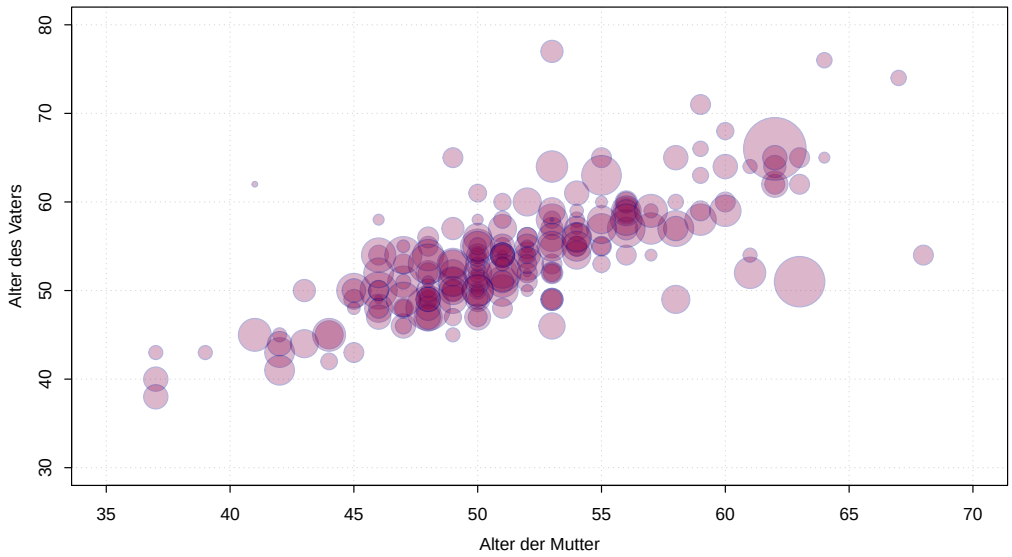
Bagplot: Boxplot in 2 Dimensionen

```
require(aplpack)
bagplot(AlterV, AlterM, xlab="Alter des Vaters", ylab="Alter der Mutter")
```



Bubbleplot: 3 metrische Variablen

```
require(DescTools)
PlotBubble(AlterM, AlterV, AusgSchuhe/400,
col=SetAlpha("deeppink4",0.3),
border=SetAlpha("darkblue",0.3),
xlab="Alter der Mutter", ylab="Alter des Vaters",
panel.first=grid(),
main="")
```



Größe der Blasen: Ausgaben für Schuhe

Korrelationsrechnung

- Frage: Wie stark ist der Zusammenhang zwischen X und Y?
- Dazu: Korrelationskoeffizienten
- Verschiedene Varianten: Wahl abhängig vom Skalenniveau von X und Y:

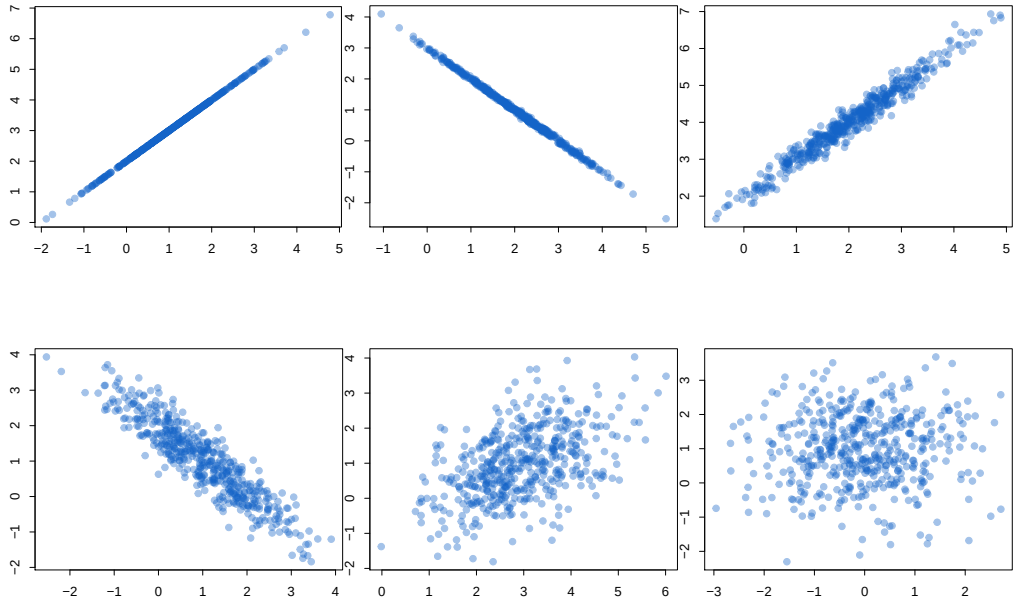
	Skalierung von Y		
Skalierung von X	kardinal	ordinal	nominal
kardinal	Bravais-Pearson-Korrelations-koeffizient	Rangkorrelations-koeffizient von Spearman	Kontingenz-koeffizient
ordinal			
nominal			

Korrelationskoeffizient von Bravais und Pearson

Bravais-Pearson-Korrelationskoeffizient

Voraussetzung: X, Y kardinalskaliert

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}} \in [-1;+1]$$



Bravais-Pearson-Korrelationskoeffizient

Im Beispiel:

i	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	2	4	4	16	8
2	4	3	16	9	12
3	3	6	9	36	18
4	9	7	81	49	63
5	7	8	49	64	56
$\sum$	25	28	159	174	157

$$\Rightarrow \quad \bar{x} = 25/5 = 5 \qquad \bar{y} = 28/5 = 5{,}6$$

$$r = \frac{157 - 5 \cdot 5 \cdot 5{,}6}{\sqrt{159 - 5 \cdot 5^2}\sqrt{174 - 5 \cdot 5{,}6^2}} = 0{,}703$$

(deutliche positive Korrelation)

Statistik
Etschberger - Januar 2015

1. Einführung
2. Deskriptive Statistik
Häufigkeiten
Lage und Streuung
Konzentration
Zwei Merkmale
Korrelation
Lineare Regression
3. W-Theorie
4. Induktive Statistik
5. Datenanalyse

Rangkorrelationskoeffizient von Spearman

- Voraussetzungen: X, Y (mindestens) ordinalskaliert, Ränge eindeutig (keine Doppelbelegung von Rängen)
- Vorgehensweise:
  1. Rangnummern R_i (X) bzw. R'_i (Y) mit $R_i^{(\prime)} = 1$ bei größtem Wert usw.
  2. Berechne

$$r_{SP} = 1 - \frac{6 \sum_{i=1}^{n} (R_i - R'_i)^2}{(n-1)\,n\,(n+1)} \in [-1; +1]$$

- Hinweise:
  - $r_{SP} = +1$ wird erreicht bei $R_i = R'_i \qquad \forall\, i = 1, \ldots, n$
  - $r_{SP} = -1$ wird erreicht bei $R_i = n + 1 - R'_i \quad \forall\, i = 1, \ldots, n$

Rangkorrelationskoeffizient von Spearman

Im Beispiel:

x_i	R_i	y_i	R'_i
2	5	4	4
4	3	3	5
3	4	6	3
9	1	7	2
7	2	8	1

$$r_{SP} = 1 - \frac{6 \cdot [(5-4)^2 + (3-5)^2 + (4-3)^2 + (1-2)^2 + (2-1)^2]}{(5-1) \cdot 5 \cdot (5+1)} = 0{,}6$$

Statistik
Etschberger - Januar 2015

1. Einführung
2. Deskriptive Statistik
Häufigkeiten
Lage und Streuung
Konzentration
Zwei Merkmale
Korrelation
Lineare Regression
3. W-Theorie
4. Induktive Statistik
5. Datenanalyse

Kontingenzkoeffizient

- Gegeben: Kontingenztabelle mit k Zeilen und l Spalten (vgl. hier)
- Vorgehensweise:
  1. Ergänze Randhäufigkeiten

$$h_{i\cdot} = \sum_{j=1}^{l} h_{ij} \qquad \text{und} \qquad h_{\cdot j} = \sum_{i=1}^{k} h_{ij}$$

  2. Berechne theoretische Häufigkeiten

$$\tilde{h}_{ij} = \frac{h_{i\cdot} \cdot h_{\cdot j}}{n}$$

  3. Berechne

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}}$$

χ^2 hängt von n ab! $(h_{ij} \mapsto 2 \cdot h_{ij} \Rightarrow \chi^2 \mapsto 2 \cdot \chi^2)$

Kontingenzkoeffizient

4. **Kontingenzkoeffizient:**

$$K = \sqrt{\frac{\chi^2}{n + \chi^2}} \qquad \in [0; K_{max}]$$

wobei

$$K_{max} = \sqrt{\frac{M-1}{M}} \qquad \text{mit} \qquad M = \min\{k, l\}$$

5. **Normierter Kontingenzkoeffizient:**

$$K_* = \frac{K}{K_{max}} \qquad \in [0; 1]$$

$K_* = +1 \iff$
bei Kenntnis von x_i kann y_i erschlossen werden u.u.

Kontingenzkoeffizient

Beispiel

X : Staatsangehörigkeit (d,a)
Y : Geschlecht (m,w)

h_{ij}	m	w	$h_{i.}$
d	30	30	60
a	10	30	40
$h_{.j}$	40	60	100

$\Rightarrow$

$\tilde{h}_{ij}$	m	w
d	24	36
a	16	24

wobei $\tilde{h}_{11} = \frac{60 \cdot 40}{100} = 24$ usw.

$$\chi^2 = \frac{(30-24)^2}{24} + \frac{(30-36)^2}{36} + \frac{(10-16)^2}{16} + \frac{(30-24)^2}{24} = 6{,}25$$

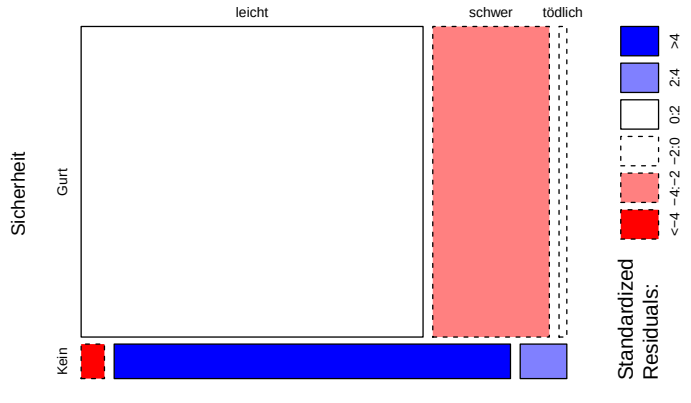
$$K = \sqrt{\frac{6{,}25}{100+6{,}25}} = 0{,}2425; \quad M = \min\{2,2\} = 2; \quad K_{max} = \sqrt{\frac{2-1}{2}} = 0{,}7071$$

$$K_* = \frac{0{,}2425}{0{,}7071} = 0{,}3430$$

Graphische Repräsentation von Kontingenztabellen

Beispiel Autounfälle

	Verletzung			
	leicht	schwer	tödlich	
angegurtet	264	90	6	360
nicht angegurtet	2	34	4	40
	266	124	10	400



Mosaikplot Autounfälle

Ausgangsdaten

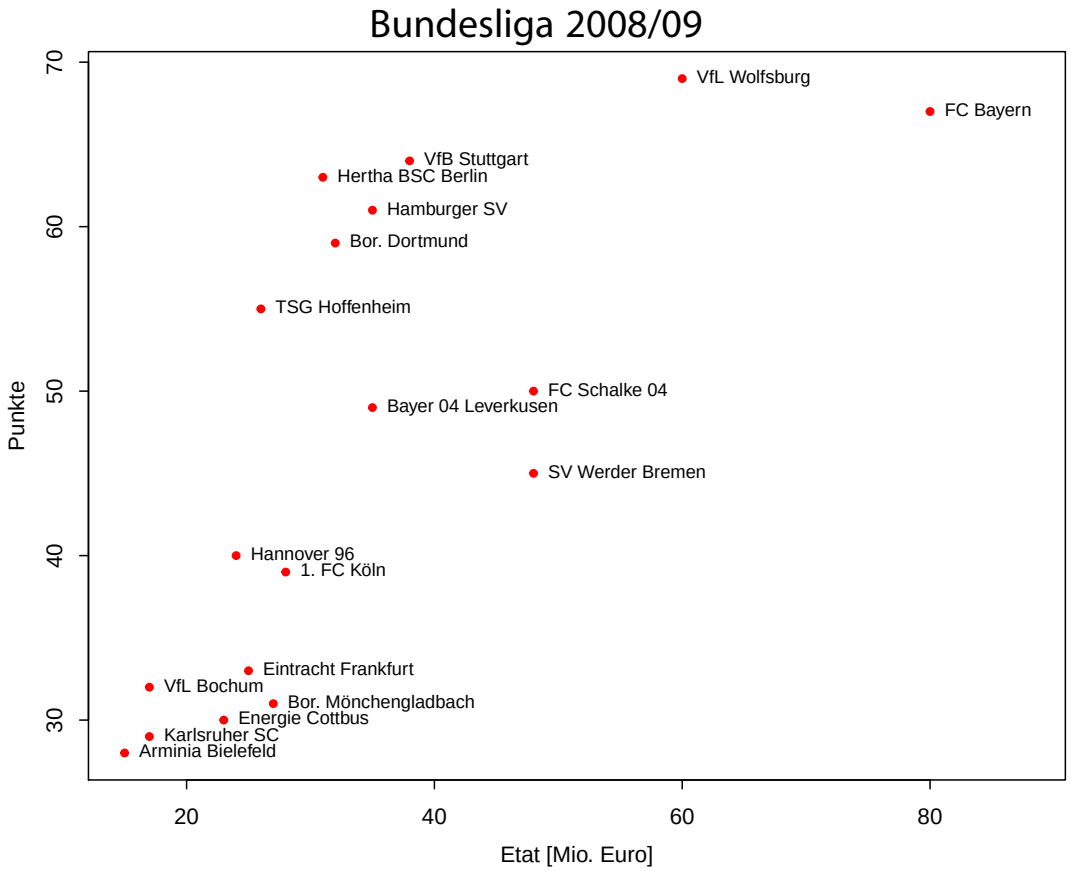
Bundesliga 2008/2009

- Gegeben: Daten zu den 18 Vereinen der ersten Bundesliga in der Saison 2008/09
- Merkmale: **Vereinssetat** für Saison (nur direkte Gehälter und Spielergehälter)
- und **Ergebnispunkte** in Tabelle am Ende der Saison

	Etat	Punkte
FC Bayern	80	67
VfL Wolfsburg	60	69
SV Werder Bremen	48	45
FC Schalke 04	48	50
VfB Stuttgart	38	64
Hamburger SV	35	61
Bayer 04 Leverkusen	35	49
Bor. Dortmund	32	59
Hertha BSC Berlin	31	63
1. FC Köln	28	39
Bor. Mönchengladbach	27	31
TSG Hoffenheim	26	55
Eintracht Frankfurt	25	33
Hannover 96	24	40
Energie Cottbus	23	30
VfL Bochum	17	32
Karlsruher SC	17	29
Arminia Bielefeld	15	28

(Quelle: Welt)

Darstellung der Daten in Streuplot



Trend als lineares Modell

- Kann man die **Tabellenpunkte** näherungsweise über einfache Funktion **in Abhängigkeit des Vereinsetats** darstellen?
- Allgemein: Darstellung einer Variablen Y als Funktion von X:

$$y = f(x)$$

- Dabei:
  - X heißt Regressor bzw. unabhängige Variable
  - Y heißt Regressand bzw. abhängige Variable
- Wichtiger (und einfachster) Spezialfall: f beschreibt einen linearen Trend:

$$y = a + b\,x$$

- Dabei anhand der Daten zu schätzen: a (Achsenabschnitt) und b (Steigung)
- Schätzung von a und b: Lineare Regression

Fehlerquadratsumme

- Pro Datenpunkt gilt mit Regressionsmodell:

$$y_i = a + bx_i + \epsilon_i$$

- Dabei: ϵ_i ist jeweils Fehler (der Grundgesamtheit),
- mit $e_i = y_i - (\hat{a} + \hat{b}x_i)$: Abweichung (Residuen) zwischen gegebenen Daten der Stichprobe und durch Modell geschätzten Werten
- Modell gut wenn alle Residuen e_i zusammen möglichst klein
- Einfache Summe aber nicht möglich, denn e_i positiv oder negativ
- Deswegen: Summe der Quadrate von e_i
- Prinzip der kleinsten Quadrate: Wähle a und b so, dass

$$Q(a,b) = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2 \to \min$$

Beste Lösung

- Beste und eindeutige Lösung:

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

$$= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$\hat{a} = \bar{y} - \hat{b}\,\bar{x}$$

- Regressionsgerade:

$$\hat{y} = \hat{a} + \hat{b}\,x$$

Bundesligabeispiel

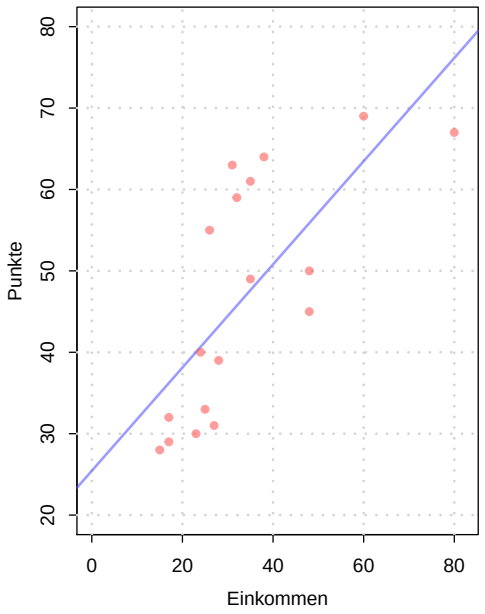
- Berechnung eines linearen Modells der Bundesligadaten
- dabei: Punkte $\hat{=} y$ und Etat $\hat{=} x$:

$\bar{x}$	33,83
$\bar{y}$	46,89
$\sum x_i^2$	25209
$\sum x_i y_i$	31474
n	18

$$\Rightarrow \hat{b} = \frac{31474 - 18 \cdot 33{,}83 \cdot 46{,}89}{25209 - 18 \cdot 33{,}83^2}$$
$$\approx 0{,}634$$

$$\Rightarrow \hat{a} = 46{,}89 - \hat{b} \cdot 33{,}83$$
$$\approx 25{,}443$$

- Modell: $\hat{y} = 25{,}443 + 0{,}634 \cdot x$

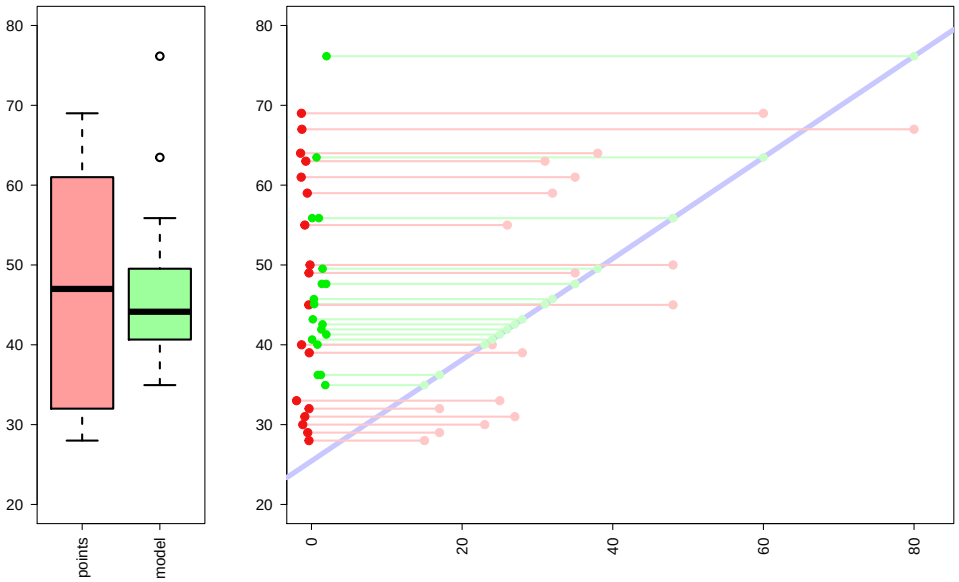


- Prognosewert für Etat $= 30$:

$$\hat{y}(30) = 25{,}443 + 0{,}634 \cdot 30$$
$$\approx 44{,}463$$

Varianz und Information

- **Varianz** der Daten in abhängiger Variablen y_i als Repräsentant des **Information**sgehalts
- Ein Bruchteil davon kann in Modellwerten $\hat{y}_i$ abgebildet werden



- Empirische Varianz (mittlere quadratische Abweichung) für „rot“ bzw. „grün“ ergibt jeweils

$$\frac{1}{18}\sum_{i=1}^{18}(y_i-\bar{y})^2 \approx 200{,}77 \qquad \text{bzw.} \qquad \frac{1}{18}\sum_{i=1}^{18}(\hat{y}_i-\bar{y})^2 \approx 102{,}78$$

Determinationskoeffizient

- Gütemaß für die Regression: Determinationskoeffizient (Bestimmtheitskoeffizient):

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = \frac{\sum_{i=1}^{n}\hat{y}_i^2 - n\bar{y}^2}{\sum_{i=1}^{n}y_i^2 - n\bar{y}^2} = r^2 \in [0;1]$$

- Mögliche Interpretation von R^2:
  Durch die Regression erklärter Anteil der Varianz
- $R^2 = 0$ wird erreicht wenn X, Y unkorreliert
  $R^2 = 1$ wird erreicht wenn $\hat{y}_i = y_i \; \forall i$ (alle Punkte auf Regressionsgerade)
- Im (Bundesliga-)Beispiel:

$$R^2 = \frac{\sum_{i=1}^{18}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{18}(y_i-\bar{y})^2} \approx \frac{102{,}78}{200{,}77} \approx 51{,}19\,\%$$

Regression: 4 eindimensionale Beispiele

- Berühmte Daten aus den 1970er Jahren:

i	x_{1i}	x_{2i}	x_{3i}	x_{4i}	y_{1i}	y_{2i}	y_{3i}	y_{4i}
1	10	10	10	8	8,04	9,14	7,46	6,58
2	8	8	8	8	6,95	8,14	6,77	5,76
3	13	13	13	8	7,58	8,74	12,74	7,71
4	9	9	9	8	8,81	8,77	7,11	8,84
5	11	11	11	8	8,33	9,26	7,81	8,47
6	14	14	14	8	9,96	8,10	8,84	7,04
7	6	6	6	8	7,24	6,13	6,08	5,25
8	4	4	4	19	4,26	3,10	5,39	12,50
9	12	12	12	8	10,84	9,13	8,15	5,56
10	7	7	7	8	4,82	7,26	6,42	7,91
11	5	5	5	8	5,68	4,74	5,73	6,89

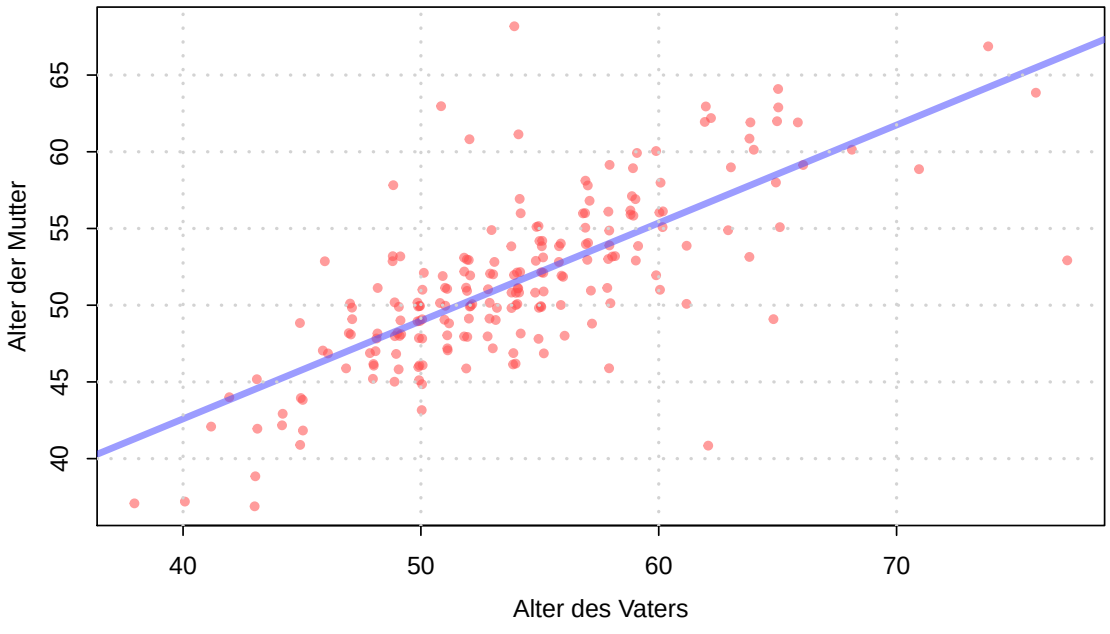
(Quelle: Anscombe (1973))

Beispieldaten


```
meineRegression = lm(AlterM ~ AlterV)
meineRegression

plot(AlterV, AlterM,
     xlab="Alter des Vaters",
     ylab="Alter der Mutter")
abline(meineRegression)
```

```
##
## Call:
## lm(formula = AlterM ~ AlterV)
##
## Coefficients:
## (Intercept)       AlterV
##     17.0537       0.6384
```



Cook's Distanz

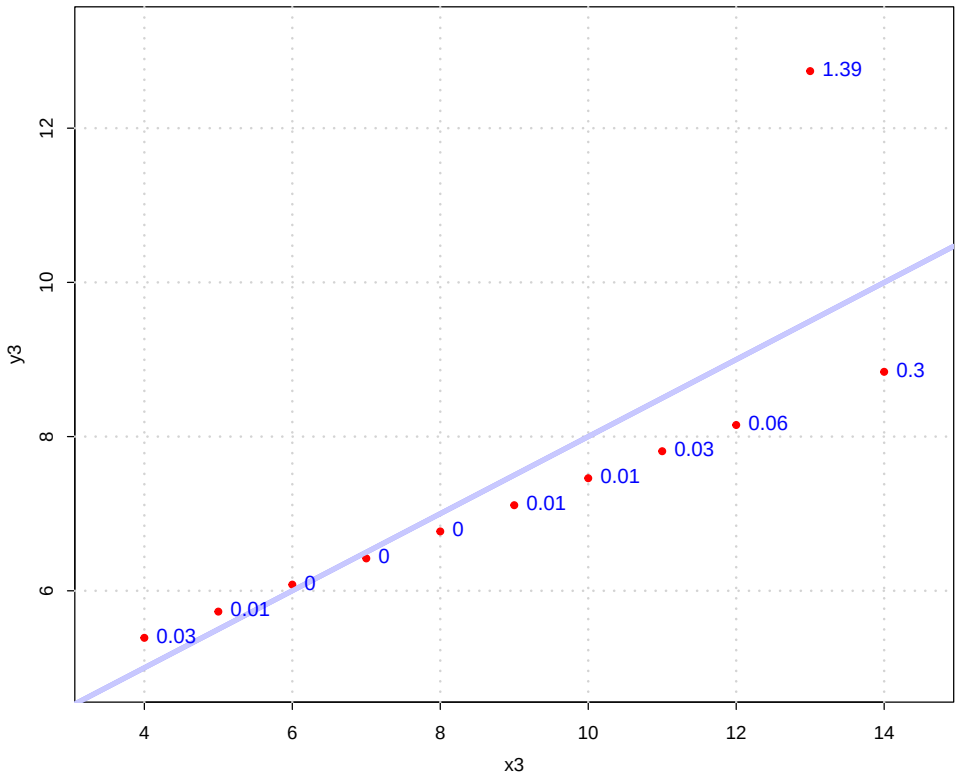
- Oft Kritisch: Einzelne Punkte, die Modell stark beeinflussen
- Idee: Was würde sich ändern, wenn solche Punkte weggelassen würden?
- Cook-Distanz: Misst den Effekt eines gelöschten Objekts
- Formel für ein lineares Modell mit einem unabh. Merkmal:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(\text{ohne } i)})^2}{\text{MSE}}$$

- Dabei bedeutet:
  - $\hat{y}_j$: Prognosewert des kompletten Modells für das j-te Objekt
  - $\hat{y}_{j(\text{ohne } i)}$: Prognosewert des Modells ohne Objekt i für das j-te Objekt
  - $\text{MSE} = \frac{1}{n} \cdot \sum (\hat{y}_i - y_i)^2$: Normierender Term (Schätzwert für Fehlerstreuung)

Ausreißer?

- Anscombe-Daten: Regressionsmodell Nr. 3
- Darstellung der Cook-Distanz neben Punkten
- Faustformel: Werte über 1 sollten genau untersucht werden

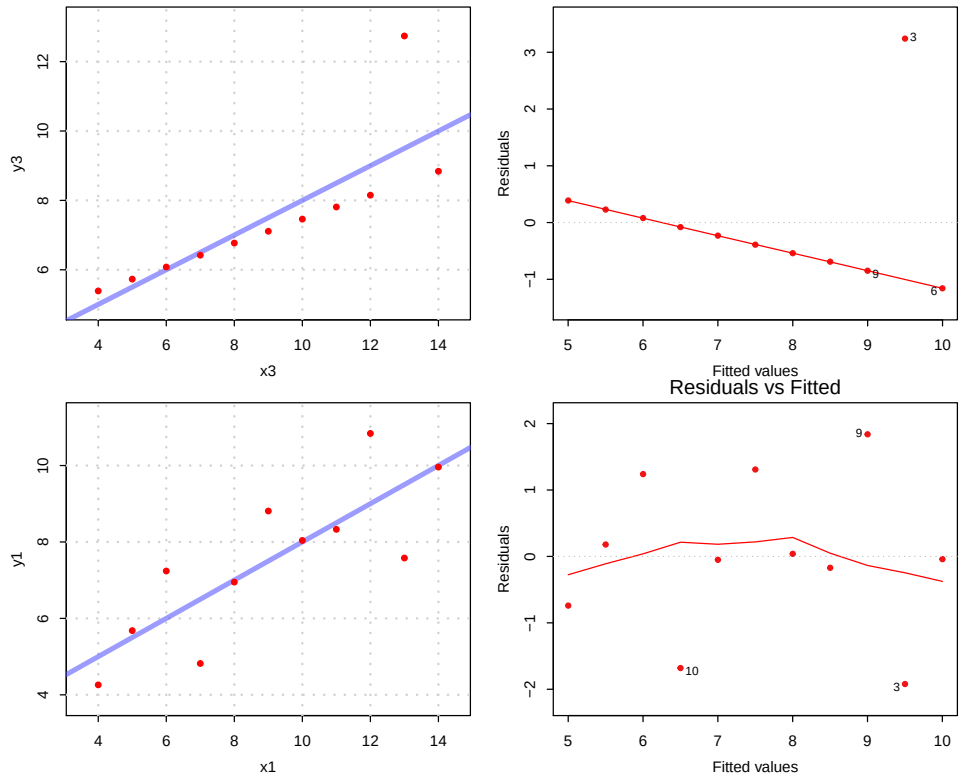


PLUS
Statistik
Etschberger - Januar 2015

1. Einführung
2. Deskriptive Statistik
Häufigkeiten
Lage und Streuung
Konzentration
Zwei Merkmale
Korrelation
Lineare Regression
3. W-Theorie
4. Induktive Statistik
5. Datenanalyse

Residualanalyse

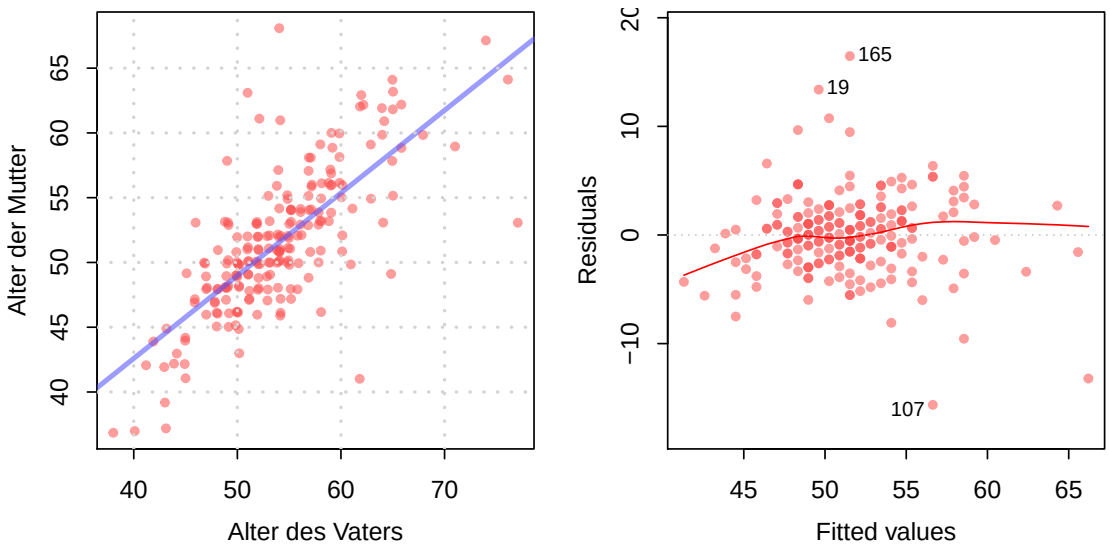
- Oft aufschlussreich: Verteilung der Residuen e_i
- Verbreitet: Graphische Darstellungen der Residuen
- Z.B.: e_i über $\hat{y}_i$



Residualanalyse

Wichtige Eigenschaften der Residuenverteilung

- Möglichst keine systematischen Muster
- Keine Änderung der Varianz in Abhängigkeit von $\hat{y}_i$ (Homoskedastizität)
- Nötig für inferentielle Analysen: Näherungsweise Normalverteilung der Residuen (q-q-plots)



Exkurs: Kausalität vs. Korrelation

- Meist wichtig für sinnvolle Regressionsanalysen:
- Kausale Verbindung zwischen unabhängigem und abhängigem Merkmal
- Sonst bei Änderung der unabhängigen Variablen keine sinnvollen Prognosen möglich
- Oft: Latente Variablen im Hintergrund

Statistik: Table of Contents

Kombinatorik: Anzahl von Kombinationen bei Auswahl

2-mal Würfeln, das heißt Auswahl von $k = 2$ aus $n = 6$ Zahlen.

$\rightarrow$

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

- mit WH, mit RF: alle Möglichkeiten, $6^2 = 36$
- ohne WH, mit RF: Diagonale entfällt, $36 - 6 = 30 = 6 \cdot 5 = \frac{6!}{(6-2)!}$
- ohne WH, ohne RF: Hälfte des letzten Ergebnisses: $\frac{30}{2} = 15 = \frac{6!}{4!\,2!} = \binom{6}{2}$
- mit WH, ohne RF: Letztes Ergebnis plus Diagonale, $15 + 6 = 21 = \binom{7}{2}$

Auswahl von k aus n Dingen		
	mit Wiederholung	ohne Wiederholung
mit Reihenfolge	n^k	$\frac{n!}{(n-k)!}$
ohne Reihenfolge	$\binom{n+k-1}{k}$	$\binom{n}{k}$

Zufallsvorgänge, Ereignisse und Wahrscheinlichkeiten

- Zufallsvorgang: Geschehen mit ungewissem Ausgang, z.B. Münzwurf
- Elementarereignis ω: Ein möglicher Ausgang, z.B. „ Kopf “ Elementarereignisse schließen sich gegenseitig aus („ Kopf “ oder „ Zahl “)!
- Ergebnismenge Ω: Menge aller ω
- **Beispiel:** Werfen zweier Würfel:

$$\Omega: \quad \begin{Bmatrix} (1,1) & (1,2) & \cdots & (1,6) \\ (2,1) & (2,2) & \cdots & (2,6) \\ \vdots & \vdots & \ddots & \vdots \\ (6,1) & (6,2) & \cdots & (6,6) \end{Bmatrix}$$

$$\Rightarrow \Omega = \{(x_1, x_2) : \; x_1, x_2 \in \{1, \ldots, 6\}\}$$

Ereignisse und Wahrscheinlichkeiten

- Ereignis A: Folgeerscheinung eines Elementarereignisses
- Formal:

$$A \subset \Omega$$

- Ereignisse schließen sich nicht gegenseitig aus!
- **Beispiel:** Werfen zweier Würfel:

Ereignis	verbal	formal
A	Augensumme $= 4$	$\{(1,3),\ (2,2),\ (3,1)\}$
B	Erste Zahl $= 2$	$\{(2,1),\ (2,2),\ \ldots,\ (2,6)\}$

- Wahrscheinlichkeit $P(A)$: Chance für das Eintreten von A
- Laplace-Wahrscheinlichkeit:

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{Anzahl der für } A \text{ günstigen Fälle}}{\text{Anzahl aller möglichen Fälle}}$$

Laplace Wahrscheinlichkeit und Urnenmodell

- **Beispiel:** Werfen zweier Würfel:

$$\text{Augensumme} = 4: \quad A = \{(1,3),\ (2,2),\ (3,1)\}$$

$$|\Omega| = 36, \quad |A| = 3 \Rightarrow P(A) = \frac{3}{36} = \frac{1}{12} = 0{,}083$$

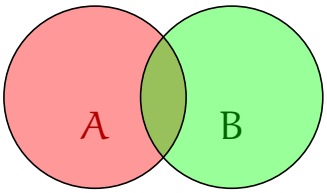
- Urnenmodell: Ziehe n Objekte aus einer Menge mit N Objekten
Anzahl Möglichkeiten:

mit Zurücklegen: N^n
ohne Zurücklegen: $N \cdot (N-1) \cdots (N-(n-1)) = \frac{N!}{(N-n)!}$

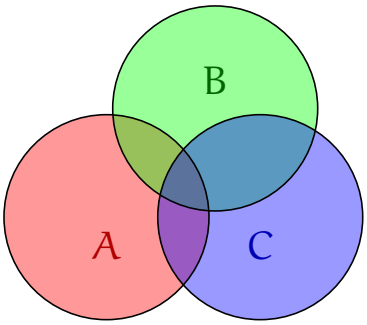
- **Beispiel:**
Wie groß ist die Wahrscheinlichkeit, aus einem gut gemischten 32-er Kartenblatt bei viermaligem Ziehen vier Asse zu bekommen?
a) Ziehen mit Zurücklegen,
b) Ziehen ohne Zurücklegen

Rechenregeln für Wahrscheinlichkeiten

- Wichtige Rechenregeln:



1. $P(A) \leqq 1$
2. $P(\emptyset) = 0$
3. $A \subset B \Rightarrow P(A) \leqq P(B)$
4. $P(\bar{A}) = 1 - P(A)$
5. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$



- **Beispiel:**

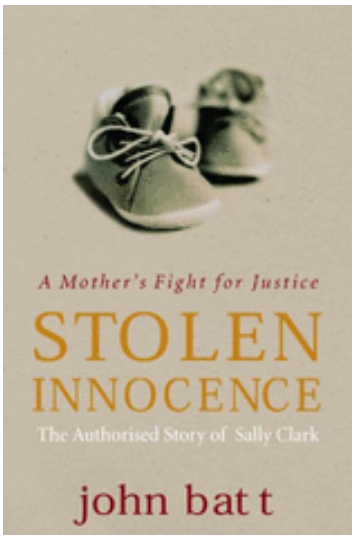
$$P(\text{„Augenzahl} \leqq 5\text{“}) = 1 - P(\text{„Augenzahl} = 6\text{“}) = 1 - \frac{1}{6} = \frac{5}{6}$$

Beispiel Gegenereignis

Der Fall Sally Clark

- Sally Clarks Söhne Christopher und Harry sterben 1996 und 1997 beide kurz nach der Geburt an plötzlichem Kindstod.
- Kinderarzt: „Wahrscheinlich Mord, da 2 maliger plötzlicher Kindstod sehr unwahrscheinlich!“ (ohne konkrete Hinweise)
- Gerichtliche Untersuchung
- Hauptargument der Anklage gestützt durch Gerichtsgutachter Sir Roy Meadow (renommierter Facharzt für Kinderheilkunde): Wahrscheinlichkeit für plötzlichen Kindstod ist 1:8500, d.h. Wahrscheinlichkeit für 2 maliges Auftreten in einer Familie

$$p = \left(\frac{1}{8500}\right)^2 \approx 1 : 72\,000\,000$$



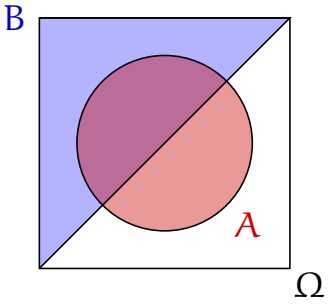
- Urteil: Doppelmord; Strafe: 2 mal lebenslang; Inhaftierung von Sally Clark 1999

Bedingte Wahrscheinlichkeiten

- Wahrscheinlichkeit von A hängt von anderem Ereignis B ab. (B kann zeitlich vor A liegen, muss aber nicht!)
- **Beispiel:** Wahrscheinlichkeit für Statistiknote hängt von Mathenote ab.
- Formal:

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

- Im Venndiagramm:



Unabhängigkeit von Ereignissen

- A, B unabhängig: Eintreten von A liefert keine Information über $P(B)$.
- Formal:

$$P(A \mid B) = P(A)$$

- Bei Unabhängigkeit ist äquivalent dazu:

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

- Dann gilt:

$$P(A \cup B) = P(A) + P(B) - P(A) \cdot P(B)$$

- **Beispiel:** Werfen zweier Würfel:

$$\left.\begin{array}{ll} A: & \text{"'erster Würfel gleich 6"'} \\ B: & \text{"'zweiter Würfel gleich 6"'} \end{array}\right\} \Rightarrow P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

$$= \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6}$$

$$= P(A)$$

Statistik
Etschberger - Januar 2015

1. Einführung
2. Deskriptive Statistik
3. W-Theorie
Kombinatorik
Zufall und Wahrscheinlichkeit
Zufallsvariablen und Verteilungen
Verteilungsparameter
4. Induktive Statistik
5. Datenanalyse

Zufallsvariablen und Verteilungen

- Beschreibung von Ereignissen durch reelle Zahlen
- Formal: Zufallsvariable ist Abbildung von Ereignisraum in reelle Zahlen:

$$X: \Omega \to \mathbb{R}$$

- **Nach** Durchführung des Zufallsvorgangs:

Realisation: $x = X(\omega)$

- **Vor** Durchführung des Zufallsvorgangs:

Wertebereich: $X(\Omega) = \{x: x = X(\omega), \omega \in \Omega\}$

- **Beispiel:** Würfeln, X: Augenzahl, $X(\Omega) = \{1,2,\ldots,6\}$, $x = 4$ (z.B.)

$$P(X = 4) = \frac{1}{6}, \quad P(X \leqq 3) = \frac{3}{6} = \frac{1}{2}$$

Statistik
Etschberger - Januar 2015

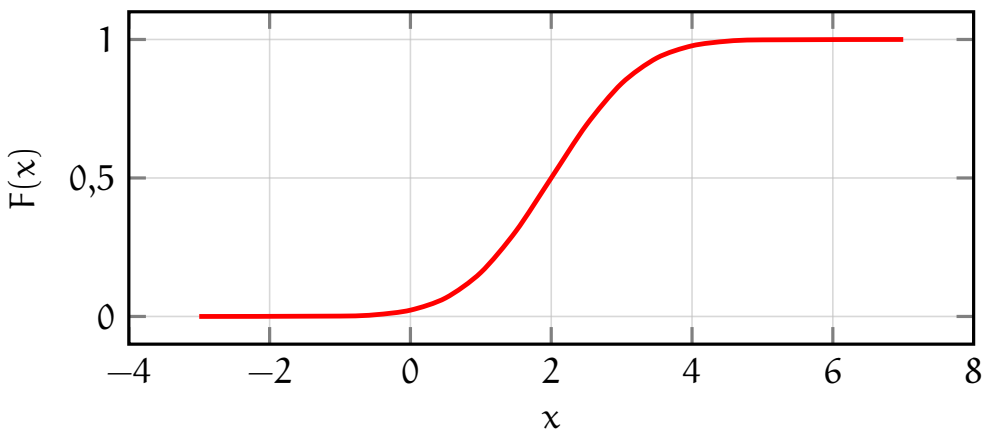
1. Einführung
2. Deskriptive Statistik
3. W-Theorie
Kombinatorik
Zufall und Wahrscheinlichkeit
Zufallsvariablen und Verteilungen
Verteilungsparameter
4. Induktive Statistik
5. Datenanalyse

- Zuweisung von Wahrscheinlichkeiten zu Realisationen
- Formal:

$$F(x) = P(X \leqq x)$$

- Eigenschaften der Verteilungsfunktion:
  - $F(x) \in [0;1]$
  - Definitionsbereich: $\mathbb{R}$ mit $F(-\infty) = 0$, $F(\infty) = 1$
  - monoton wachsend, d.h. $x_1 < x_2 \Rightarrow F(x_1) \leqq F(x_2)$
  - Es gilt:

$$P(a < X \leqq b) = F(b) - F(a)$$



Beispiel einer Verteilungsfunktion

Diskrete Zufallsvariablen

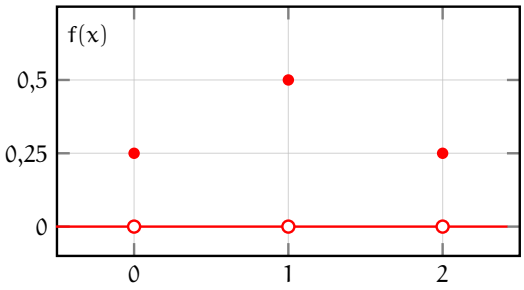
- X heißt diskret, wenn $X(\Omega) = \{x_1, x_2, \ldots\}$ endlich ist.
- Wahrscheinlichkeitsfunktion dann:

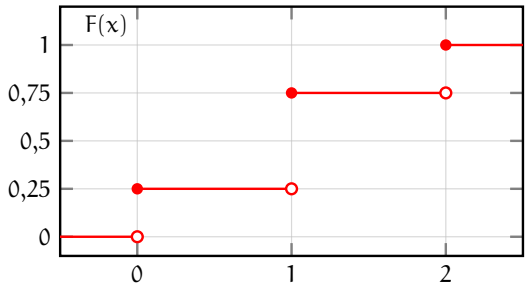
$$f(x) = P(X = x)$$

Beispiel: Münze 2 mal werfen; X: Anzahl "'Kopf"'

	(Z, Z)	(Z, K), (K, Z)	(K, K)
x_i	0	1	2
$f(x_i)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

$$F(x) = \begin{cases} 0, & \text{falls } x < 0 \\ \frac{1}{4}, & \text{falls } 0 \leqq x < 1 \\ \frac{3}{4}, & \text{falls } 1 \leqq x < 2 \\ 1, & \text{falls } x \geqq 2 \end{cases}$$





Binomialverteilung

- **Wiederholter** Zufallsvorgang
- n Durchführungen (jeweils unabhängig)
- Pro Durchführung: A oder $\bar{A}$ mit $P(A) = p$ ($\widehat{=}$ Ziehen mit Zurücklegen)
- Schreibe:

$$X_i = \begin{cases} 1, & \text{falls } A \text{ bei } i\text{-ter Durchführung eintritt} \\ 0, & \text{falls } \bar{A} \text{ bei } i\text{-ter Durchführung eintritt} \end{cases}$$

- Dann gibt

$$X = \sum_{i=1}^{n} X_i$$

an, wie oft A eintritt.

- Gesucht: Wahrscheinlichkeitsfunktion von X

Binomialverteilung

- Herleitung:
  1) $P(X_i = 1) = P(A) = p,\ P(X_i = 0) = P(\bar{A}) = 1 - p$
  2) $\sum_{i=1}^{n} x_i = x$ entspricht "'x mal Ereignis A und $n - x$ mal $\bar{A}$"'
     Wahrscheinlichkeit (bei Unabhängigkeit): $p^x \cdot (1-p)^{n-x}$
  3) Aber: Reihenfolge irrelevant! Anzahl Anordnungen: $\binom{n}{x}$

➠ Wahrscheinlichkeitsfunktion der Binomialverteilung:

$$f(x) = \begin{cases} \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}, & \text{falls } x \in \{0,1,\ldots,n\} \\ 0, & \text{sonst} \end{cases}$$

- Kurzschreibweise: $X \sim B(n;p)$
  X ist binomialverteilt mit Parametern n und p
- Tabellen zeigen meist $F(x)$
- für $f(x)$ gilt: $f(x) = F(x) - F(x-1)$

$X \sim B(n, 0.25)$, Tabelle der Binomialverteilung $F(x) = P(X \le x)$

$x \backslash n$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	0.7500	0.5625	0.4219	0.3164	0.2373	0.1780	0.1335	0.1001	0.0751	0.0563	0.0422	0.0317	0.0238	0.0178	0.0134
1	1.0000	0.9375	0.8438	0.7383	0.6328	0.5339	0.4450	0.3671	0.3003	0.2440	0.1971	0.1584	0.1267	0.1010	0.0802
2		1.0000	0.9844	0.9492	0.8965	0.8306	0.7564	0.6786	0.6007	0.5256	0.4552	0.3907	0.3326	0.2811	0.2361
3			1.0000	0.9961	0.9844	0.9624	0.9295	0.8862	0.8343	0.7759	0.7133	0.6488	0.5843	0.5213	0.4613
4				1.0000	0.9990	0.9954	0.9871	0.9727	0.9511	0.9219	0.8854	0.8424	0.7940	0.7415	0.6865
5					1.0000	0.9998	0.9987	0.9958	0.9900	0.9803	0.9657	0.9456	0.9198	0.8883	0.8516
6						1.0000	0.9999	0.9996	0.9987	0.9965	0.9924	0.9858	0.9757	0.9617	0.9434
7							1.0000	1.0000	0.9999	0.9996	0.9988	0.9972	0.9944	0.9897	0.9827
8								1.0000	1.0000	1.0000	0.9999	0.9996	0.9990	0.9979	0.9958
9									1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9992
10										1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
11											1.0000	1.0000	1.0000	1.0000	1.0000

$x \backslash n$	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
0	0.0100	0.0075	0.0056	0.0042	0.0032	0.0024	0.0018	0.0013	0.0010	0.0008	0.0006	0.0004	0.0003	0.0002	0.0002
1	0.0635	0.0501	0.0395	0.0310	0.0243	0.0190	0.0149	0.0116	0.0090	0.0070	0.0055	0.0042	0.0033	0.0025	0.0020
2	0.1971	0.1637	0.1353	0.1114	0.0913	0.0745	0.0607	0.0492	0.0398	0.0321	0.0258	0.0208	0.0166	0.0133	0.0106
3	0.4050	0.3530	0.3057	0.2631	0.2252	0.1917	0.1624	0.1370	0.1150	0.0962	0.0802	0.0666	0.0551	0.0455	0.0375
4	0.6302	0.5739	0.5187	0.4654	0.4149	0.3674	0.3235	0.2832	0.2467	0.2138	0.1844	0.1583	0.1354	0.1153	0.0979
5	0.8104	0.7653	0.7175	0.6678	0.6172	0.5666	0.5168	0.4685	0.4222	0.3783	0.3372	0.2990	0.2638	0.2317	0.2026
6	0.9205	0.8929	0.8610	0.8251	0.7858	0.7436	0.6994	0.6537	0.6074	0.5611	0.5154	0.4708	0.4279	0.3869	0.3481
7	0.9729	0.9598	0.9431	0.9226	0.8982	0.8701	0.8385	0.8037	0.7662	0.7265	0.6852	0.6427	0.5998	0.5568	0.5143
8	0.9925	0.9876	0.9807	0.9713	0.9591	0.9439	0.9254	0.9037	0.8787	0.8506	0.8196	0.7860	0.7502	0.7126	0.6736
9	0.9984	0.9969	0.9946	0.9911	0.9861	0.9794	0.9705	0.9592	0.9453	0.9287	0.9092	0.8868	0.8616	0.8337	0.8034
10	0.9997	0.9994	0.9988	0.9977	0.9961	0.9936	0.9900	0.9852	0.9787	0.9703	0.9599	0.9472	0.9321	0.9145	0.8943
11	1.0000	0.9999	0.9998	0.9995	0.9991	0.9983	0.9971	0.9954	0.9928	0.9893	0.9845	0.9784	0.9706	0.9610	0.9494
12	1.0000	1.0000	1.0000	0.9999	0.9998	0.9996	0.9993	0.9988	0.9979	0.9966	0.9948	0.9922	0.9888	0.9842	0.9784
13	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9997	0.9995	0.9991	0.9985	0.9976	0.9962	0.9944	0.9918
14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9996	0.9993	0.9989	0.9982	0.9973
15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9997	0.9995	0.9992
16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9998
17		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
18			1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Statistik
Etschberger - Januar 2015

1. Einführung
2. Deskriptive Statistik
3. W-Theorie
Kombinatorik
Zufall und Wahrscheinlichkeit
Zufallsvariablen und Verteilungen
Verteilungsparameter
4. Induktive Statistik
5. Datenanalyse

Binomialverteilung: Beispiel

Beispiel

Aus einem 32-er Kartenblatt wird 3-mal eine Karte mit Zurücklegen gezogen.
Wie wahrscheinlich ist es, 2-mal Herz zu ziehen?

$$X_i = \begin{cases} 1, & \text{falls i-te Karte Herz} \\ 0, & \text{sonst} \end{cases} \quad \Rightarrow \quad X_i \sim B(1; \tfrac{8}{32})$$

$$X = \sum_{i=1}^{n} X_i = X_1 + X_2 + X_3 \quad \Rightarrow \quad X \sim B(3; \tfrac{1}{4})$$

Mithilfe der Wahrscheinlichkeitsfunktion:

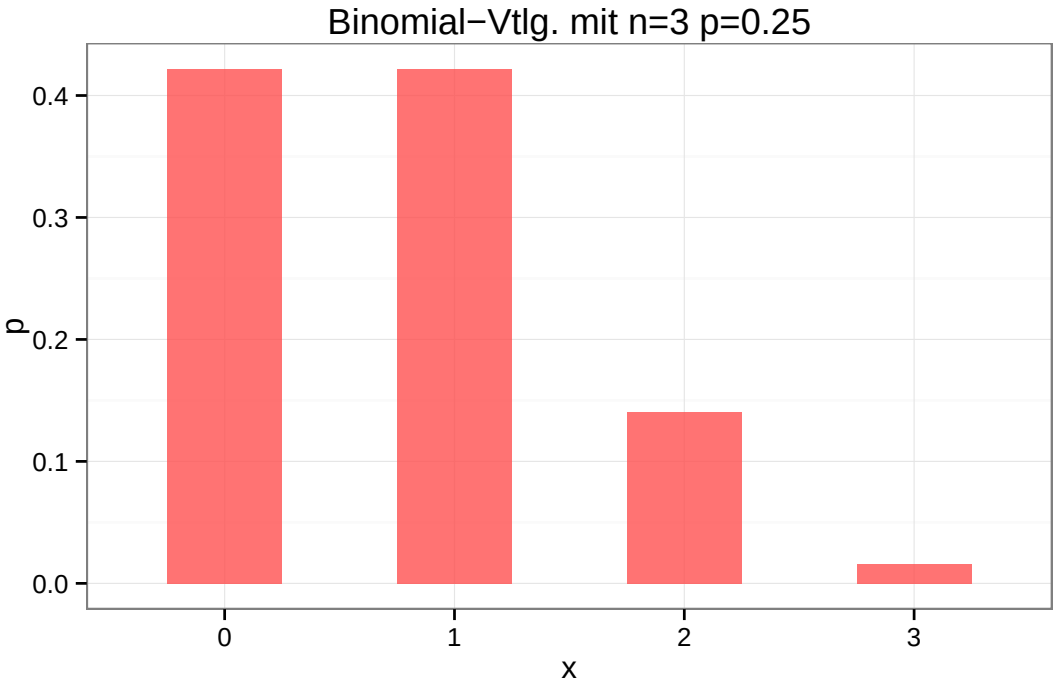
$$P(X = 2) = f(2) = \binom{3}{2} \cdot 0{,}25^2 \cdot 0{,}75^1 = 0{,}1406$$

Mithilfe der Tabelle ($n = 3$):

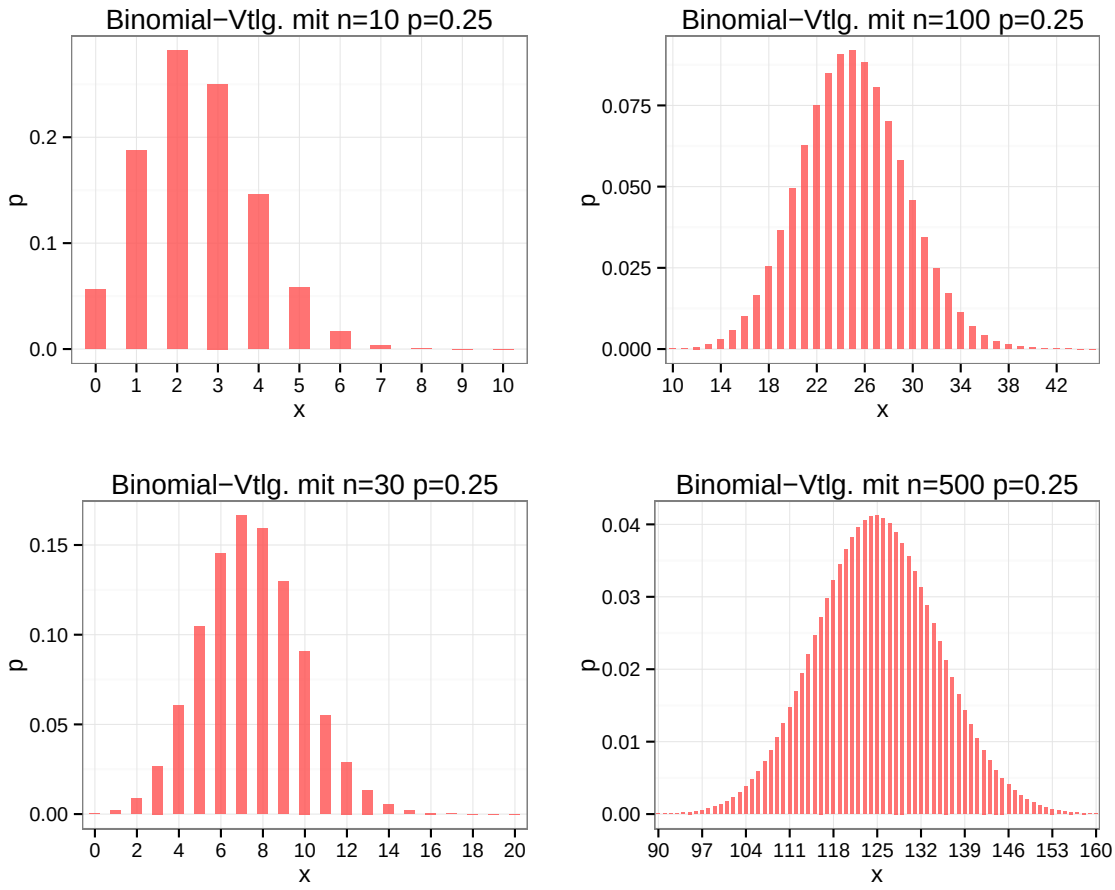
$$P(X = 2) = F(2) - F(1) = 0{,}9844 - 0{,}8438 = 0{,}1406$$

Binomialverteilung: Wahrscheinlichkeitsfunktion

- $X \sim B(3, \tfrac{1}{4})$



Binomialverteilung: Wahrscheinlichkeitsfunktion



Hypergeometrische Verteilung

- n-faches Ziehen **ohne** Zurücklegen aus N Objekten, davon M markiert.

  $$X = \text{Anzahl gezogener Objekte mit Markierung}$$

  heißt hypergeometrisch verteilt mit den Parametern N, M, n.
- Kurzschreibweise: $X \sim \text{Hyp}(N; M; n)$
- Wahrscheinlichkeitsfunktion:

$$f(x) = \begin{cases} \dfrac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, & \text{falls } x \text{ möglich} \\ 0, & \text{sonst} \end{cases}$$

- Ist $n \leq \frac{N}{20}$, so gilt: $\text{Hyp}(N; M; n) \approx B(n; \frac{M}{N})$

Beispiel: Hypergeometrische Verteilung

- Aus einem 32-Kartenblatt wird 3-mal eine Karte ohne Zurücklegen gezogen.
- Wie wahrscheinlich ist es, 2-mal "'Herz"' zu ziehen?
- D.h.: $N = 32,\ M = 8,\ n = 3,\ x = 2.$

$$P(X = 2) = f(2) = \frac{\binom{8}{2}\binom{32-8}{3-2}}{\binom{32}{3}} = \frac{\binom{8}{2}\binom{24}{1}}{\binom{32}{3}} = \frac{\frac{8!}{2!\cdot 6!}\cdot 24}{\frac{32!}{3!\cdot 29!}}$$

$$= \frac{29!\cdot 8!\cdot 3!\cdot 24}{32!\cdot 6!\cdot 2!} = \frac{8\cdot 7\cdot 3\cdot 24}{32\cdot 31\cdot 30} = \frac{4032}{29760} = \frac{21}{155}$$

$$= 0{,}1355$$

Dabei wurde verwendet:

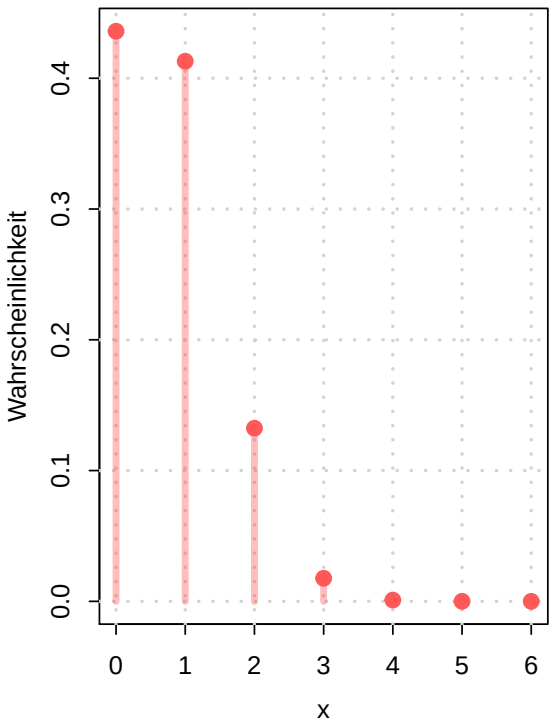
$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \quad \text{und} \quad \binom{n}{1} = n.$$

Hypergeometrische Verteilung

Beispiel: x Treffer im Lotto 6 aus 49

- $X \sim \text{Hyp}(49, 6, 6)$

x	$P(X = x)$ (in %)
0	43.596498
1	41.301945
2	13.237803
3	1.765040
4	0.096862
5	0.001845
6	0.000007



Poisson-Verteilung

- Approximation für $B(n;p)$ und $Hyp(N;M;n)$
- Geeignet, wenn
  p klein ($\leqq 0{,}1$), n groß ($\geqq 50$) und $np \leqq 10$.
- „Verteilung der seltenen Ereignisse“
  (z.B. Anzahl 6-er pro Lottoausspielung)
- X ist poissonverteilt mit Parameter λ: $X \sim P(\lambda)$
- Wahrscheinlichkeitsfunktion:

$$f(x) = \begin{cases} \dfrac{\lambda^x}{x!} \cdot e^{-\lambda}, & \text{falls } x = 0,1,2,\ldots \\ 0, & \text{sonst} \end{cases}$$

- $F(x)$ in Tabelle
- Überblick: Approximation

$$Hyp(N;M;n) \xrightarrow{p = \frac{M}{N}} B(n;p) \xrightarrow{\lambda = np = n\frac{M}{N}} P(\lambda)$$

Poissonverteilung: $X \sim P(\lambda)$, Tabelle der Verteilungsfunktionen

$x \setminus \lambda$	1.6	1.7	1.8	1.9	2	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3
0	0.2019	0.1827	0.1653	0.1496	0.1353	0.1225	0.1108	0.1003	0.0907	0.0821	0.0743	0.0672	0.0608	0.0550	0.0498
1	0.5249	0.4933	0.4628	0.4338	0.4060	0.3796	0.3546	0.3309	0.3085	0.2873	0.2674	0.2487	0.2311	0.2146	0.1992
2	0.7834	0.7572	0.7306	0.7037	0.6767	0.6496	0.6227	0.5960	0.5697	0.5438	0.5184	0.4936	0.4695	0.4460	0.4232
3	0.9212	0.9068	0.8913	0.8747	0.8571	0.8387	0.8194	0.7994	0.7787	0.7576	0.7360	0.7141	0.6919	0.6696	0.6472
4	0.9763	0.9704	0.9636	0.9559	0.9474	0.9379	0.9275	0.9163	0.9041	0.8912	0.8774	0.8629	0.8477	0.8318	0.8153
5	0.9940	0.9920	0.9896	0.9868	0.9834	0.9796	0.9751	0.9700	0.9643	0.9580	0.9510	0.9433	0.9349	0.9258	0.9161
6	0.9987	0.9981	0.9974	0.9966	0.9955	0.9941	0.9925	0.9906	0.9884	0.9858	0.9828	0.9794	0.9756	0.9713	0.9665
7	0.9997	0.9996	0.9994	0.9992	0.9989	0.9985	0.9980	0.9974	0.9967	0.9958	0.9947	0.9934	0.9919	0.9901	0.9881
8	1.0000	0.9999	0.9999	0.9998	0.9998	0.9997	0.9995	0.9994	0.9991	0.9989	0.9985	0.9981	0.9976	0.9970	0.9962
9	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9999	0.9998	0.9997	0.9996	0.9995	0.9993	0.9992	0.9989
10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9999	0.9998	0.9998	0.9997
11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

$x \setminus \lambda$	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4	4.1	4.2	4.3	4.4	4.5
0	0.0451	0.0408	0.0369	0.0334	0.0302	0.0273	0.0247	0.0224	0.0203	0.0183	0.0166	0.0150	0.0136	0.0123	0.0111
1	0.1847	0.1712	0.1586	0.1469	0.1359	0.1257	0.1162	0.1074	0.0992	0.0916	0.0845	0.0780	0.0719	0.0663	0.0611
2	0.4012	0.3799	0.3594	0.3397	0.3209	0.3028	0.2854	0.2689	0.2531	0.2381	0.2238	0.2102	0.1974	0.1852	0.1736
3	0.6248	0.6025	0.5803	0.5584	0.5366	0.5152	0.4942	0.4735	0.4533	0.4335	0.4142	0.3954	0.3772	0.3595	0.3423
4	0.7982	0.7806	0.7626	0.7442	0.7255	0.7064	0.6872	0.6679	0.6484	0.6288	0.6093	0.5898	0.5704	0.5512	0.5321
5	0.9057	0.8946	0.8829	0.8706	0.8576	0.8441	0.8301	0.8156	0.8006	0.7851	0.7693	0.7532	0.7367	0.7199	0.7029
6	0.9612	0.9554	0.9490	0.9422	0.9347	0.9267	0.9182	0.9091	0.8995	0.8893	0.8787	0.8675	0.8558	0.8437	0.8311
7	0.9858	0.9832	0.9802	0.9769	0.9733	0.9692	0.9648	0.9599	0.9546	0.9489	0.9427	0.9361	0.9290	0.9214	0.9134
8	0.9953	0.9943	0.9931	0.9917	0.9901	0.9883	0.9863	0.9840	0.9815	0.9786	0.9755	0.9721	0.9683	0.9642	0.9598
9	0.9986	0.9982	0.9978	0.9973	0.9967	0.9960	0.9952	0.9942	0.9931	0.9919	0.9905	0.9889	0.9871	0.9851	0.9829
10	0.9996	0.9995	0.9994	0.9992	0.9990	0.9987	0.9984	0.9981	0.9977	0.9972	0.9966	0.9959	0.9952	0.9943	0.9933
11	0.9999	0.9999	0.9998	0.9998	0.9997	0.9996	0.9995	0.9994	0.9993	0.9991	0.9989	0.9986	0.9983	0.9980	0.9976
12	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9999	0.9998	0.9998	0.9997	0.9997	0.9996	0.9995	0.9994	0.9992
13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9999	0.9999	0.9998	0.9998
14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Statistik
Etschberger - Januar 2015

1. Einführung
2. Deskriptive Statistik
3. W-Theorie
Kombinatorik
Zufall und Wahrscheinlichkeit
Zufallsvariablen und Verteilungen
Verteilungsparameter
4. Induktive Statistik
5. Datenanalyse

Poisson-Verteilung: Beispiel

Beispiel

- $X \sim B(10\,000; 0{,}0003)$; In Tabelle der Binomialverteilung nicht vertafelt! Approximation:

$$\left.\begin{array}{rl} p = 0{,}0003 & < 0{,}1 \\ n = 10\,000 & > 50 \\ np = 3 & < 10 \end{array}\right\} \Rightarrow B(10\,000; 0{,}0003) \approx P(3)$$

- Mithilfe der Wahrscheinlichkeitsfunktion:

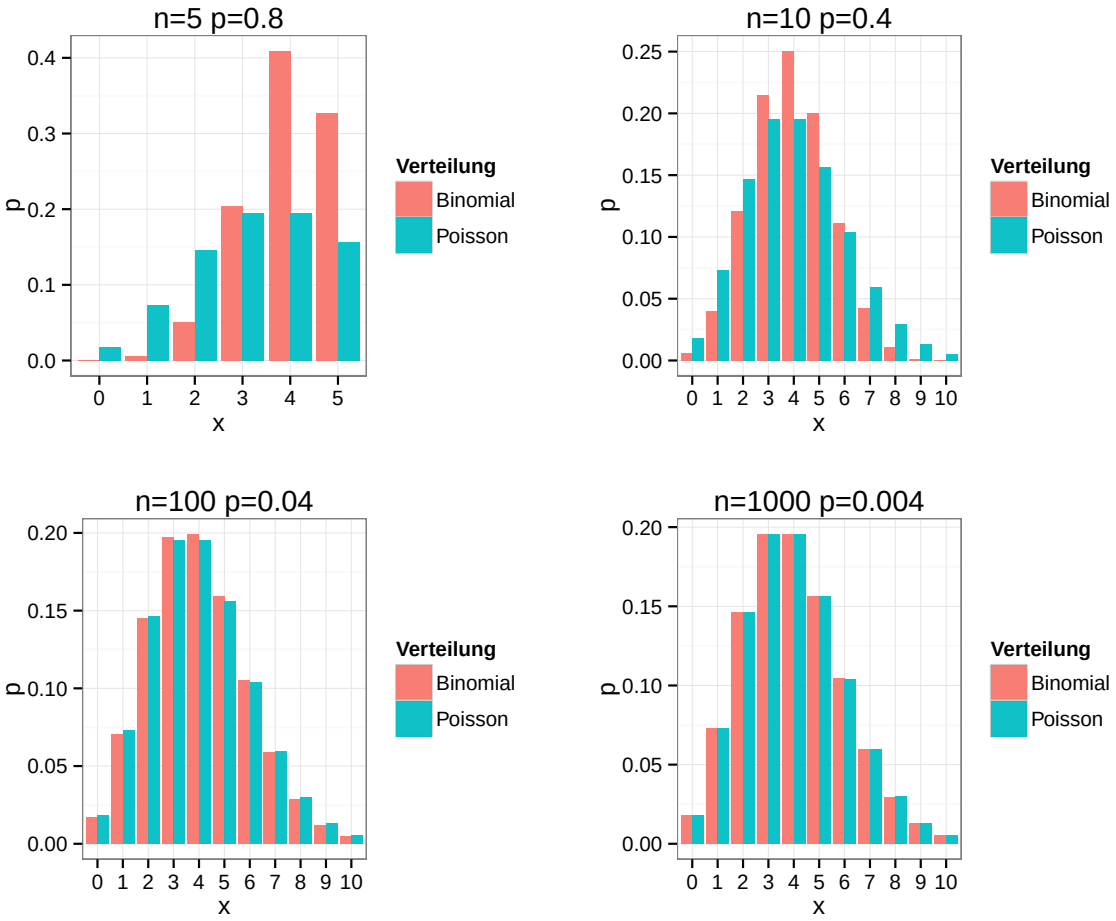
$$P(X = 5) = \frac{3^5}{5!} \cdot e^{-3} = 0{,}1008188$$

- Mithilfe der Tabelle der Poissonverteilung:

$$P(X = 5) = F(5) - F(4) = 0{,}9161 - 0{,}8153 = 0{,}1008$$

- Exakter Wert: $P(X = 5) = 0{,}1008239$

Poisson- versus Binomialverteilung: Vergleich

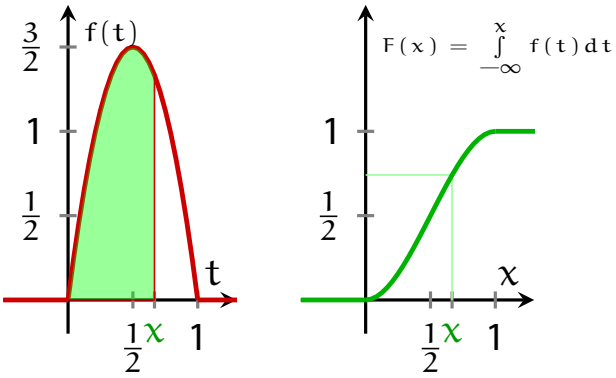


Stetige Zufallsvariablen

- X heißt stetig, wenn $F(x)$ stetig ist.
- Dann existiert ein $f(t)$ mit:

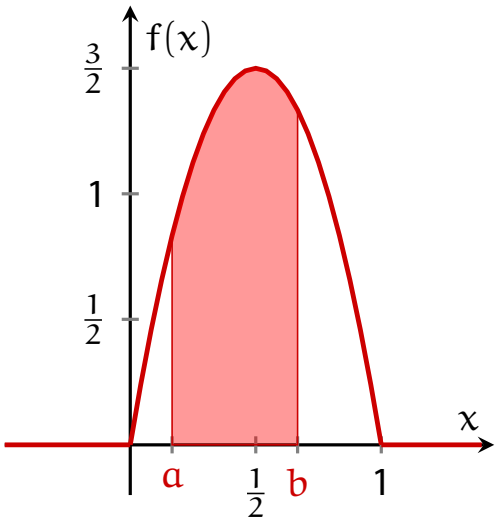
$$F(x) = \int_{-\infty}^{x} f(t)\,dt$$

$f(x)$ heißt Dichtefunktion von X.



- Dann:

$$\begin{aligned} P(a < X < b) &= P(a \leqq X < b) \\ &= P(a < X \leqq b) \\ &= P(a \leqq X \leqq b) \\ &= \int_a^b f(x)\,dx \\ &= F(b) - F(a) \end{aligned}$$



Dichtefunktion

Eigenschaften der Dichtefunktion

- $f(x) \geqq 0$ für alle $x \in \mathbb{R}$
- Wegen $F(\infty) = 1$ muss stets gelten:

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

- $P(X = x) = 0$ für alle $x \in \mathbb{R}$
- $f(x) > 1$ ist möglich
- für $x \in \mathbb{R}$ ist $F(x)$ differenzierbar $\Rightarrow F'(x) = f(x)$.
- Intervallgrenzen spielen keine Rolle:

$$\begin{aligned} P(X \in [a;b]) &= P(X \in (a;b]) \\ &= P(X \in [a;b)) \\ &= P(X \in (a;b)) \\ &= F(b) - F(a) \end{aligned}$$

Dichtefunktion: Beispiel

Beispiel

$$f(x) = \begin{cases} 0, & \text{falls} \quad x < 0 \\ \frac{1}{10}, & \text{falls} \quad 0 \leqq x \leqq 10 \\ 0, & \text{falls} \quad x > 10 \end{cases}$$

Verteilungsfunktion:

$$\int_0^x f(t)\,dt = \int_0^x \frac{1}{10}\,dt = \left[\frac{t}{10}\right]_0^x = \frac{x}{10} \Rightarrow$$

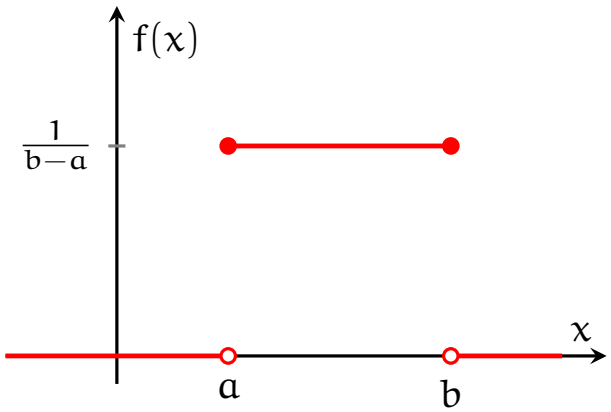
$$F(x) = \begin{cases} 0, & \text{falls} \quad x < 0 \\ \frac{x}{10}, & \text{falls} \quad 0 \leqq x \leqq 10 \\ 1, & \text{falls} \quad x > 10 \end{cases}$$

Gleichverteilung

Eine Zufallsvariable X mit

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{falls} \quad a \leqq x \leqq b \\ 0, & \text{sonst} \end{cases}$$

heißt gleichverteilt im Intervall $[a; b]$.



Gleichverteilung

- Verteilungsfunktion der Gleichverteilung:

$$F(x) = \begin{cases} 0 & , \text{ falls } x < a \\ \dfrac{x-a}{b-a} & , \text{ falls } a \leqq x \leqq b \\ 1 & , \text{ falls } x > b \end{cases}$$

- **Beispiel:** X gleichverteilt in $[1;20]$

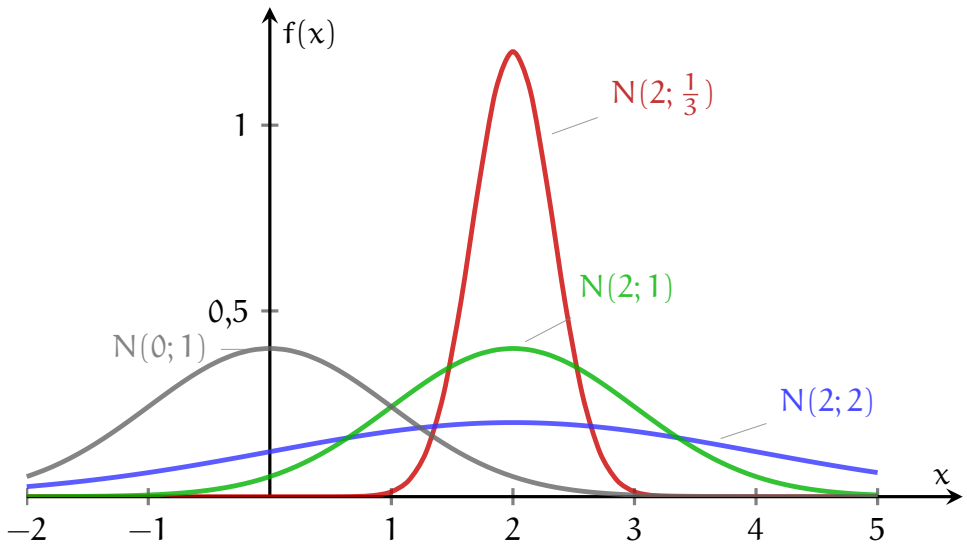
$$\begin{aligned} P(2 \leqq X \leqq 12) &= F(12) - F(2) = \frac{12-1}{20-1} - \frac{2-1}{20-1} \\ &= \frac{12-2}{20-1} = \frac{10}{19} \\ &= 0{,}5263 \end{aligned}$$

Normalverteilung

Eine Zufallsvariable X mit einer Dichtefunktion

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

und $\sigma > 0$ heißt normalverteilt.



Kurzschreibweise: $X \sim N(\mu;\sigma)$

Normalverteilung: Gaußkurve

Verteilungsfunktion Φ der Standardnormalverteilung

Statistik
Etschberger -
Januar 2015

1. Einführung
2. Deskriptive Statistik
3. W-Theorie
Kombinatorik
Zufall und Wahrscheinlichkeit
Zufallsvariablen und Verteilungen
Verteilungsparameter
4. Induktive Statistik
5. Datenanalyse

Dabei bedeutet $\Phi(x)$ zum Beispiel: $\Phi(2{,}13) = \Phi(2{,}1 + 0{,}03) = 0{,}9834$. Diesen Wert findet man in der Zeile mit $x_1 = 2{,}1$ und der Spalte mit $x_2 = 0{,}03$.

$x_1 \setminus x_2$	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5754
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6737	0.6773	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7020	0.7054	0.7089	0.7123	0.7157	0.7191	0.7224
0.6	0.7258	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7518	0.7549
0.7	0.7580	0.7612	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7882	0.7910	0.7939	0.7967	0.7996	0.8023	0.8051	0.8079	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8290	0.8315	0.8340	0.8365	0.8389
1	0.8414	0.8438	0.8461	0.8485	0.8508	0.8532	0.8554	0.8577	0.8599	0.8622
1.1	0.8643	0.8665	0.8687	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9083	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9193	0.9207	0.9222	0.9237	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9358	0.9370	0.9382	0.9394	0.9406	0.9418	0.9430	0.9441
1.6	0.9452	0.9463	0.9474	0.9485	0.9495	0.9505	0.9516	0.9526	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9600	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9679	0.9686	0.9693	0.9700	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9762	0.9767
2	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9914	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9933	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9975	0.9975	0.9976	0.9977	0.9978	0.9978	0.9979	0.9980	0.9980	0.9981
2.9	0.9981	0.9982	0.9983	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995	0.9995

Eigenschaften der Normalverteilung

- Dichte ist symmetrisch zu μ:

$$f(\mu - x) = f(\mu + x)$$

- μ ist Lage-, σ ist Streuungsparameter
- Standardnormalverteilung:
  $N(0;1)$ mit Verteilungsfunktion $\Phi(x)$ ($\rightarrow$ Tabelle 3)
- Kenntnis von $\Phi(x)$, μ und σ genügt, denn:
  $X \sim N(\mu;\sigma) \iff \frac{X-\mu}{\sigma} \sim N(0;1) \quad \Rightarrow$

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

- Tabelle enthält nur positive x: Deswegen

$$\Phi(-x) = 1 - \Phi(x)$$

Normalverteilung: Beispiel

Beispiel:

Projektdauer $X \sim N(39;2)$.

Wahrscheinlichkeit für Projektdauer zwischen 37 und 41 Wochen?

Lösung:

$$\begin{aligned}
P(37 \leqq X \leqq 41) &= F(41) - F(37) \\
&= \Phi\left(\tfrac{41-39}{2}\right) - \Phi\left(\tfrac{37-39}{2}\right) \\
&= \Phi(1) - \Phi(-1) \\
&= \Phi(1) - [1 - \Phi(1)] \\
&= 2 \cdot \Phi(1) - 1 \\
&= 2 \cdot 0{,}8413 - 1 \\
&= 0{,}6826
\end{aligned}$$

Lageparameter

a) **Modus** x_{Mod}: $f(x_{Mod}) \geqq f(x)$ für alle x
(i.A. nicht eindeutig, z.B. Gleichverteilung)

Beispiele:

- Normalverteilung: $x_{Mod} = \mu$
- Diskrete Verteilung mit:

x	0	1	2
$f(x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

$\Rightarrow x_{Mod} = 1$

b) **Median** x_{Med}: $F(x_{Med}) = \frac{1}{2}$ bzw. kleinstes x mit $F(x) > \frac{1}{2}$

Beispiele:

- Normalverteilung: $x_{Med} = \mu$
- Diskrete Verteilung
  oben: $F(0) = \frac{1}{4} < \frac{1}{2}$, $F(1) = \frac{3}{4} > \frac{1}{2} \Rightarrow x_{Med} = 1$

Lageparameter: Fraktile

c) **α-Fraktil** x_α: $F(x_\alpha) = \alpha$ (für stetige Verteilungen)

Beispiel: $X \sim N(0;1)$, $Y \sim N(3;2)$

$$\begin{aligned} x_{0,975} &= 1{,}96 \quad \text{(Tab. 3)} \\ x_{0,025} &= -x_{0,975} = -1{,}96 \\ y_{0,025} &= 2 \cdot x_{0,025} + 3 = -0{,}92 \end{aligned}$$

Hinweise:

- $x_{Med} = x_{0,5}$
- Wenn x_α nicht vertafelt $\rightarrow$ Interpolation:

$$x_\alpha \approx x_a + (x_b - x_a) \cdot \frac{\alpha - a}{b - a}$$

mit
a : größte vertafelte Zahl $< \alpha$
b : kleinste vertafelte Zahl $> \alpha$

Beispiel: $X \sim N(0;1)$; $x_{0,6} \approx 0{,}25 + (0{,}26 - 0{,}25) \cdot \frac{0{,}6 - 0{,}5987}{0{,}6026 - 0{,}5987} = 0{,}2533$

Statistik
Etschberger - Januar 2015

1. Einführung
2. Deskriptive Statistik
3. W-Theorie
Kombinatorik
Zufall und Wahrscheinlichkeit
Zufallsvariablen und Verteilungen
Verteilungsparameter
4. Induktive Statistik
5. Datenanalyse

Lageparameter: Erwartungswert

d) Erwartungswert $E(X)$ bzw. μ:

$$E(X) = \begin{cases} \sum_i x_i f(x_i), & \text{falls } X \text{ diskret} \\ \int_{-\infty}^{\infty} x f(x)\,dx, & \text{falls } X \text{ stetig} \end{cases}$$

Beispiel: Diskrete Verteilung mit

x	0	1	2
$f(x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

$$\Rightarrow \quad E(X) = 0 \cdot \tfrac{1}{4} + 1 \cdot \tfrac{1}{2} + 2 \cdot \tfrac{1}{4} = 1$$

Beispiel: Für eine exponentialverteilte Zufallsvariable X mit der Dichte

$$f(x) = \begin{cases} \lambda \cdot e^{-\lambda x} & \text{für } x \geq 0 \\ 0 & \text{sonst} \end{cases} \quad \text{folgt}$$

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx = \lambda \int_0^{\infty} x \cdot e^{-\lambda x} dx = \lambda \left[-\frac{1}{\lambda} x e^{-\lambda x} - \int_0^{\infty} 1 \cdot \left(-\frac{1}{\lambda} e^{-\lambda x} \right) dx \right]$$

$$= -x e^{-\lambda x} - \frac{1}{\lambda} e^{-\lambda x} \Big|_0^{\infty} = -0 - \left(-0 - \frac{1}{\lambda} \right) = \frac{1}{\lambda}$$

Rechenregeln für den Erwartungswert

- Ist f symmetrisch bzgl. a, so gilt $E(X) = a$
  Beispiel: f der Gleichverteilung symmetrisch bzgl. $\frac{a+b}{2} \Rightarrow E(X) = \frac{a+b}{2}$
- Lineare Transformation:

$$E(a + bX) = a + b \cdot E(X)$$

- Summenbildung:

$$E\left(\sum_{i=1}^{n} X_i \right) = \sum_{i=1}^{n} E(X_i)$$

Beispiel: X gleichverteilt in $[0;10]$, $Y \sim N(1;1)$; $Z = X + 5Y$

$$E(Z) = E(X+5Y) = E(X)+E(5Y) = E(X)+5 \cdot E(Y) = \tfrac{10+0}{2} + 5 \cdot 1 = 10$$

- Unabhängigkeit:

$$X,\ Y \text{ unabhängig} \Rightarrow E(X \cdot Y) = E(X) \cdot E(Y)$$

Statistik
Etschberger - Januar 2015

1. Einführung
2. Deskriptive Statistik
3. W-Theorie
Kombinatorik
Zufall und Wahrscheinlichkeit
Zufallsvariablen und Verteilungen
Verteilungsparameter
4. Induktive Statistik
5. Datenanalyse

Streuungsparameter

- Varianz $\mathrm{Var}(X)$ bzw. σ^2:

$$\mathrm{Var}(X) = \mathrm{E}([X - \mathrm{E}(X)]^2) = \begin{cases} \sum_i [x_i - \mathrm{E}(X)]^2 f(x_i), & \text{wenn } X \text{ diskret} \\ \int_{-\infty}^{\infty} [x - \mathrm{E}(X)]^2 f(x)\,dx, & \text{wenn } X \text{ stetig} \end{cases}$$

- Standardabweichung $\mathrm{Sta}(X)$ bzw. σ: $\mathrm{Sta}(X) = \sqrt{\mathrm{Var}(X)}$

Beispiel: Diskrete Verteilung

x	0	1	2
$f(x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

:

$$\mathrm{Var}(X) = (0-1)^2 \cdot \frac{1}{4} + (1-1)^2 \cdot \frac{1}{2} + (2-1)^2 \cdot \frac{1}{4} = \frac{1}{2}$$

Beispiel: Für eine exponentialverteilte Zufallsvariable X (Dichte siehe Erwartungswert) folgt

$$\begin{aligned} \mathrm{Var}(X) &= \int_{-\infty}^{\infty} (x - \mathrm{E}(X)) f(x)\,dx = \lambda \int_0^{\infty} \left(x - \tfrac{1}{\lambda}\right)^2 \cdot e^{-\lambda x}\,dx \\ &= e^{-\lambda x}\left(-x^2 + \tfrac{2x}{\lambda} - \left(\tfrac{1}{\lambda}\right)^2 - \tfrac{2}{\lambda^2} - \tfrac{2x}{\lambda} + \tfrac{2}{\lambda^2}\right)\Big|_0^{\infty} \\ &= 0 - \left(-0^2 - \left(\tfrac{1}{\lambda}\right)^2\right) = \frac{1}{\lambda^2} \end{aligned}$$

Rechenregeln für die Varianz

- Verschiebungssatz:

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$$

Beispiel: Diskrete Verteilung

x	0	1	2
$f(x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

:

$$\begin{aligned} \mathrm{E}(X^2) &= 0^2 \cdot \tfrac{1}{4} + 1^2 \cdot \tfrac{1}{2} + 2^2 \cdot \tfrac{1}{4} \\ &= \tfrac{3}{2} \\ \Rightarrow \quad \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 &= \tfrac{3}{2} - 1^2 = \tfrac{1}{2} = \mathrm{Var}(X) \end{aligned}$$

- Lineare Transformation:

$$\mathrm{Var}(a + bX) = b^2\,\mathrm{Var}(X)$$

- Summenbildung gilt nur, wenn die X_i unabhängig! Dann:

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i)$$

Erwartungswerte und Varianzen wichtiger Verteilungen

Verteilung von X	$E(X)$	$\mathrm{Var}(X)$
Binomialverteilung $B(n;p)$	np	$np(1-p)$
Hypergeometrische Verteilung mit den Parametern N, M, n	$n\frac{M}{N}$	$n\frac{M}{N}\frac{N-M}{N}\frac{N-n}{N-1}$
Poisson-Verteilung $P(\lambda)$	λ	λ
Gleichverteilung in $[a;b]$ mit $a<b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normalverteilung $N(\mu;\sigma)$	μ	σ^2

Anwendung: Ungleichung von Tschebyschow

- Für beliebige Zufallsvariablen X und $\varepsilon > 0$ gilt die Ungleichung von Tschebyschow:

$$P\big(|X - E[X]| \geq \varepsilon\big) \leq \frac{\mathrm{Var}[X]}{\varepsilon^2}$$

Beispiele:

- X ist gleichverteilt mit Parametern a, b und $\varepsilon = \frac{1}{3}(a-b)$, also $E[X] = \frac{1}{2}(a+b)$ und $\mathrm{Var}[X] = \frac{1}{12}(a-b)^2$

$$\Rightarrow P\Big(\big|X - \tfrac{1}{2}(a+b)\big| \geq \tfrac{1}{3}(a-b)\Big) \leq \frac{(a-b)^2}{12} \cdot \frac{3^2}{(a-b)^2} = 3/4$$

- $X \sim B(100;0{,}2)$ und $\varepsilon = 10$
damit: $E[X] = 100 \cdot 0{,}2 = 20$ und $\mathrm{Var}[X] = 100 \cdot 0{,}2 \cdot (1-0{,}2) = 16$

$$\Rightarrow P\big(|X - 20| \geq 10\big) \leq \frac{16}{10^2} = 0{,}16$$

Statistik
Etschberger - Januar 2015

1. Einführung
2. Deskriptive Statistik
3. W-Theorie
Kombinatorik
Zufall und Wahrscheinlichkeit
Zufallsvariablen und Verteilungen
Verteilungsparameter
4. Induktive Statistik
5. Datenanalyse

Kovarianz und Korrelation

- Kovarianz:

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \mathrm{E}[(X-\mathrm{E}(X))(Y-\mathrm{E}(Y))] \\ &= \mathrm{E}(X\cdot Y)-\mathrm{E}(X)\cdot\mathrm{E}(Y) \\ &\text{(Verschiebungssatz)} \end{aligned}$$

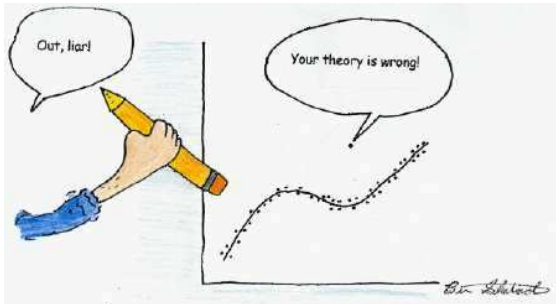
- Korrelationskoeffizient:

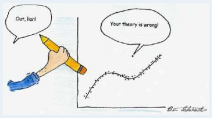
$$\rho(X,Y)=\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\cdot\mathrm{Var}(Y)}}$$

- **Bemerkungen:**
  - ρ ist r nachgebildet $\Rightarrow \rho\in[-1;1]$
  - $|\rho|=1 \iff Y=a+bX$ (mit $b\neq 0$)
  - $\rho=0 \iff X,\ Y$ **unkorreliert**
- Varianz einer Summe zweier ZV:

$$\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)+2\,\mathrm{Cov}(X,Y)$$

Statistik: Table of Contents



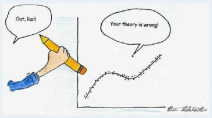


- Vollerhebung of unmöglich,
- Deshalb: Beobachte Teilgesamtheit und schließe auf Grundgesamtheit

Beispiel

Warensendung von 1000 Stück; darunter M Stück Ausschuss.
M ist unbekannt.
$\rightarrow$ Zufällige Entnahme von $n = 30$ Stück („Stichprobe“).
Darunter 2 Stück Ausschuss.
Denkbare Zielsetzungen:

- Schätze M durch eine Zahl (z.B. $\frac{2}{30} \cdot 1000 = 66{,}67$)
- Schätze ein Intervall für M (z.B. $M \in [58; 84]$)
- Teste die Hypothese, dass $M > 50$ ist.

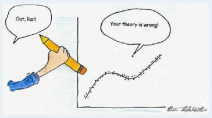


- Grundgesamtheit (G): Menge aller relevanten Merkmalsträger.
- Verteilung von G: $F(x) = P(X \leqq x)$ = Wahrscheinlichkeit, dass ein Merkmalsträger ausgewählt wird, der beim untersuchten Merkmal maximal die Ausprägung x aufweist.
- Uneingeschränkte (reine) Zufallsauswahl:
  Jedes Element von G hat die selbe Chance, ausgewählt zu werden.
- Stichprobenumfang (n): Anzahl der Merkmalsträger in der Stichprobe.
- Einfache Stichprobe:
  Uneingeschränkte Zufallsauswahl und unabhängige Ziehung.
  $\rightarrow$ Alle Stichprobenvariablen $X_1, \ldots, X_n$ sind iid.
- Stichprobenergebnis:
  n-Tupel der Realisationen der Stichprobenvariablen, $(x_1, \ldots, x_n)$.

Wichtige Stichprobenfunktionen

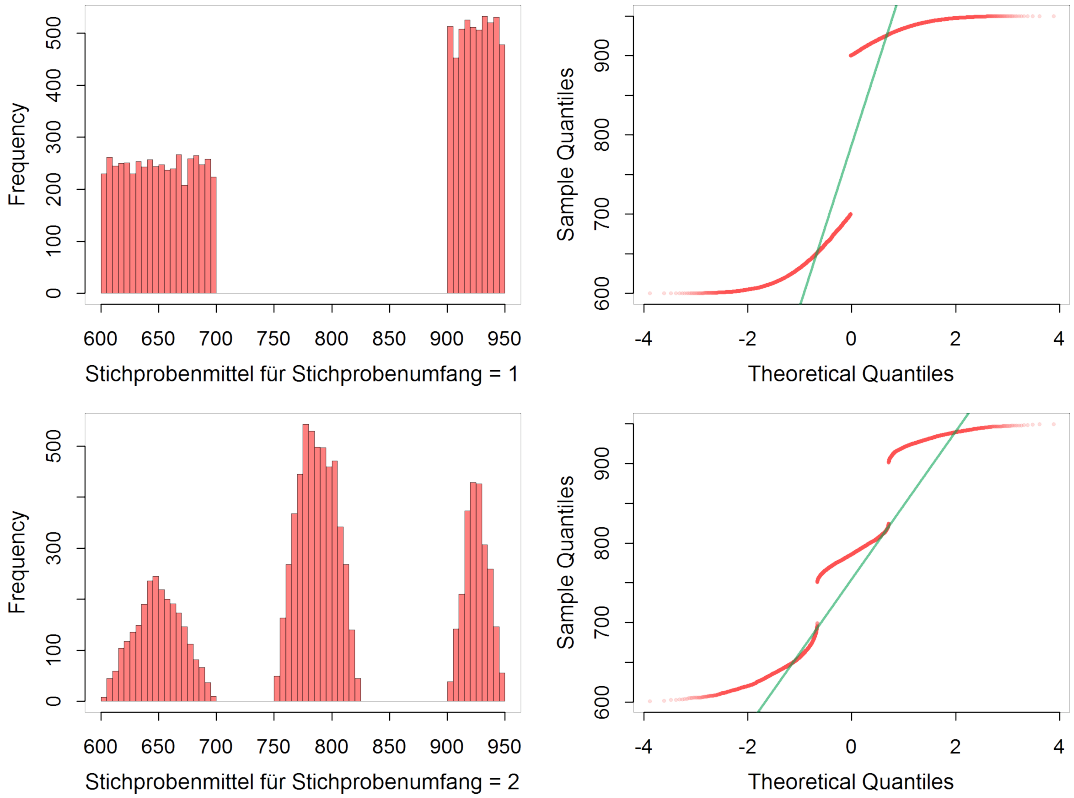
- Gegeben: Einfache Stichprobe $X_1, \ldots, X_n$, Beliebige Verteilung, mit $E(X_i) = \mu$, $Var(X_i) = \sigma^2$

Stichprobenfunktion V	Bezeichnung	E(V)	Var(V)
$\sum_{i=1}^{n} X_i$	Merkmalssumme	$n\mu$	$n\sigma^2$
$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	Stichprobenmittel	μ	$\frac{\sigma^2}{n}$
$\frac{\bar{X}-\mu}{\sigma}\sqrt{n}$	Gauß-Statistik	0	1
$\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$	mittlere quadratische Abweichung bezüglich μ	σ^2	
$\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$	mittlere quadratische Abweichung	$\frac{n-1}{n}\sigma^2$	
$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$	Stichprobenvarianz	σ^2	
$S = \sqrt{S^2}$	Stichproben-Standardabweichung		
$\frac{\bar{X}-\mu}{S}\sqrt{n}$	t-Statistik		

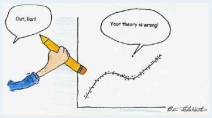


Auswirkungen der Stichprobengröße

Ziehen von 10.000 Stichproben (jeweils vom Umfang n) und Berechnung der Stichprobenmittel (Verteilung: zwei überlagerte Gleichverteilungen):

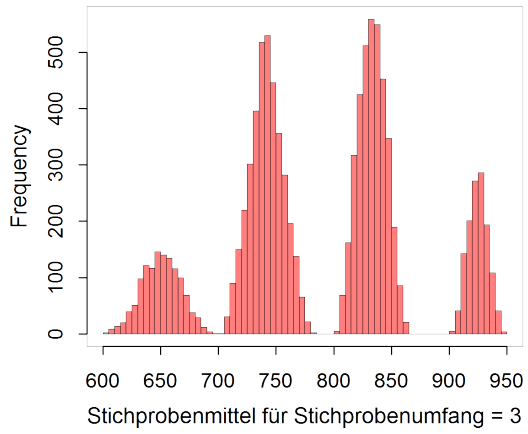


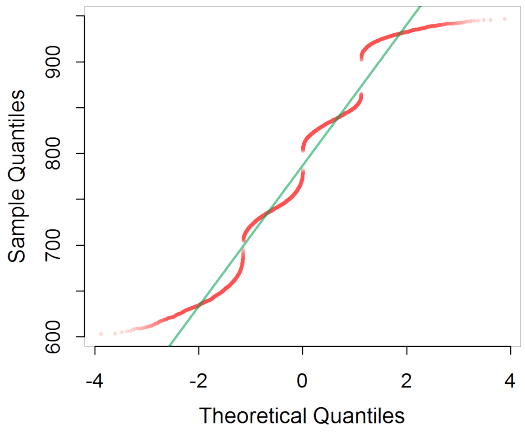
Statistik
Etschberger - Januar 2015

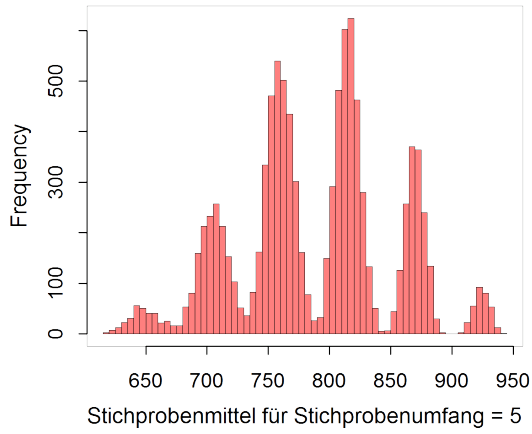


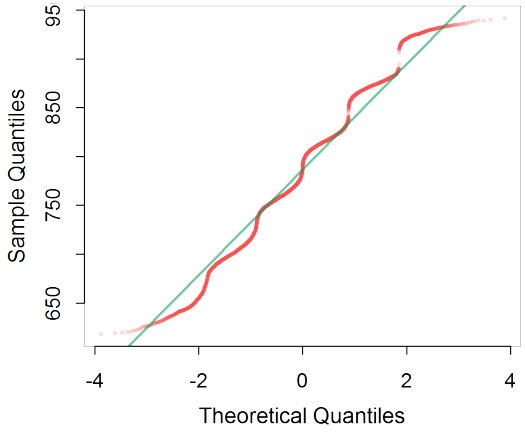
1. Einführung
2. Deskriptive Statistik
3. W-Theorie
4. Induktive Statistik
Grundlagen
Punkt-Schätzung
Intervall-Schätzung
Signifikanztests
5. Datenanalyse

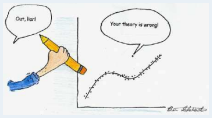
Auswirkungen der Stichprobengröße



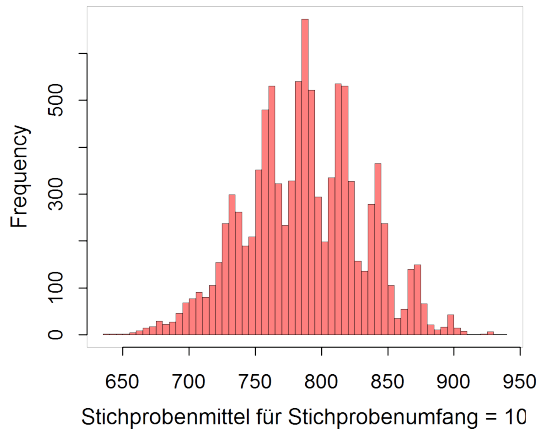


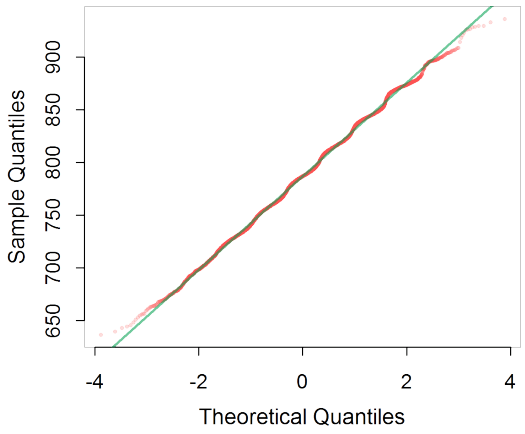


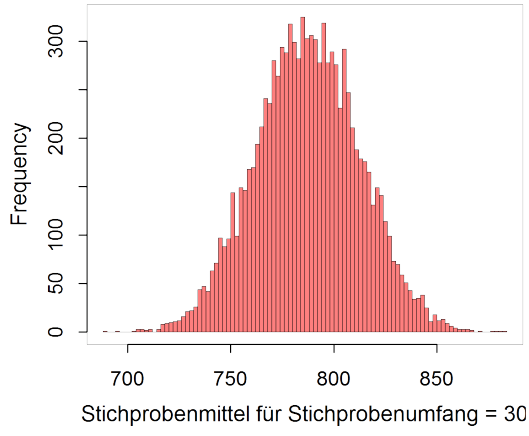


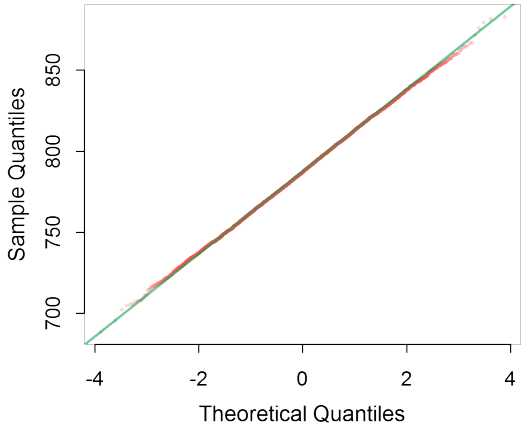


Auswirkungen der Stichprobengröße

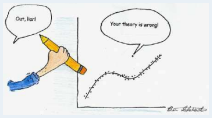




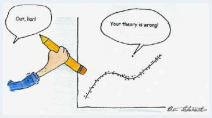




Statistik
Etschberger -
Januar 2015



1. Einführung
2. Deskriptive Statistik
3. W-Theorie
4. Induktive Statistik
Grundlagen
Punkt-Schätzung
Intervall-Schätzung
Signifikanztests
5. Datenanalyse

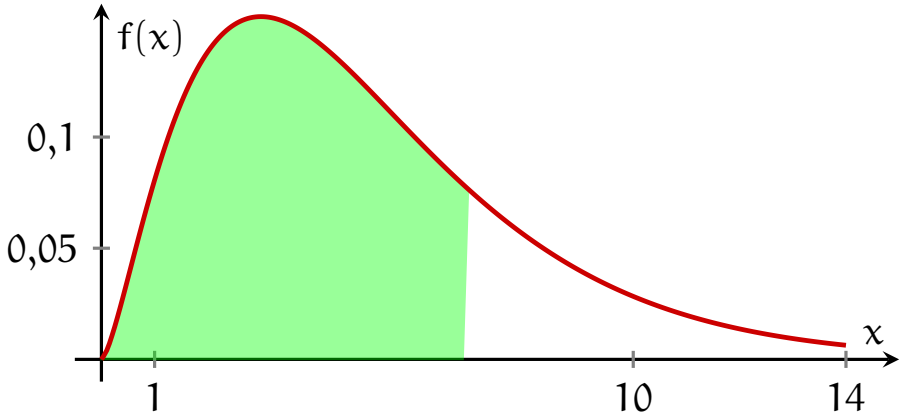


Chi-Quadrat-Verteilung

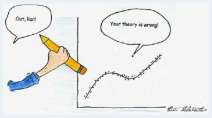
- Sind $X_1, \ldots, X_n$ iid $N(0;1)$-verteilte Zufallsvariablen, so wird die Verteilung von

$$Z = \sum_{i=1}^{n} X_i^2$$

als Chi-Quadrat-Verteilung mit n Freiheitsgraden bezeichnet.

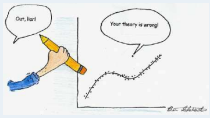


- Kurzschreibweise: $Z \sim \chi^2(n)$
- **Beispiel:** $\chi^2(30)$: $x_{0,975} = 46{,}98$



$\alpha \backslash n$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0.005	0.00	0.01	0.07	0.21	0.41	0.68	0.99	1.34	1.73	2.16	2.60	3.07	3.56	4.07	4.60
0.01	0.00	0.02	0.11	0.30	0.55	0.87	1.24	1.65	2.09	2.56	3.05	3.57	4.11	4.66	5.23
0.025	0.00	0.05	0.22	0.48	0.83	1.24	1.69	2.18	2.70	3.25	3.82	4.40	5.01	5.63	6.26
0.05	0.00	0.10	0.35	0.71	1.15	1.64	2.17	2.73	3.33	3.94	4.57	5.23	5.89	6.57	7.26
0.1	0.02	0.21	0.58	1.06	1.61	2.20	2.83	3.49	4.17	4.87	5.58	6.30	7.04	7.79	8.55
0.2	0.06	0.45	1.01	1.65	2.34	3.07	3.82	4.59	5.38	6.18	6.99	7.81	8.63	9.47	10.31
0.25	0.10	0.58	1.21	1.92	2.67	3.45	4.25	5.07	5.90	6.74	7.58	8.44	9.30	10.17	11.04
0.4	0.28	1.02	1.87	2.75	3.66	4.57	5.49	6.42	7.36	8.30	9.24	10.18	11.13	12.08	13.03
0.5	0.45	1.39	2.37	3.36	4.35	5.35	6.35	7.34	8.34	9.34	10.34	11.34	12.34	13.34	14.34
0.6	0.71	1.83	2.95	4.04	5.13	6.21	7.28	8.35	9.41	10.47	11.53	12.58	13.64	14.69	15.73
0.75	1.32	2.77	4.11	5.39	6.63	7.84	9.04	10.22	11.39	12.55	13.70	14.85	15.98	17.12	18.25
0.8	1.64	3.22	4.64	5.99	7.29	8.56	9.80	11.03	12.24	13.44	14.63	15.81	16.98	18.15	19.31
0.9	2.71	4.61	6.25	7.78	9.24	10.64	12.02	13.36	14.68	15.99	17.27	18.55	19.81	21.06	22.31
0.95	3.84	5.99	7.81	9.49	11.07	12.59	14.07	15.51	16.92	18.31	19.68	21.03	22.36	23.68	25.00
0.975	5.02	7.38	9.35	11.14	12.83	14.45	16.01	17.53	19.02	20.48	21.92	23.34	24.74	26.12	27.49
0.99	6.63	9.21	11.34	13.28	15.09	16.81	18.48	20.09	21.67	23.21	24.73	26.22	27.69	29.14	30.58
0.995	7.88	10.60	12.84	14.86	16.75	18.55	20.28	21.95	23.59	25.19	26.76	28.30	29.82	31.32	32.80

$\alpha \backslash n$	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
0.005	5.14	5.70	6.26	6.84	7.43	8.03	8.64	9.26	9.89	10.52	11.16	11.81	12.46	13.12	13.79
0.01	5.81	6.41	7.01	7.63	8.26	8.90	9.54	10.20	10.86	11.52	12.20	12.88	13.56	14.26	14.95
0.025	6.91	7.56	8.23	8.91	9.59	10.28	10.98	11.69	12.40	13.12	13.84	14.57	15.31	16.05	16.79
0.05	7.96	8.67	9.39	10.12	10.85	11.59	12.34	13.09	13.85	14.61	15.38	16.15	16.93	17.71	18.49
0.1	9.31	10.09	10.86	11.65	12.44	13.24	14.04	14.85	15.66	16.47	17.29	18.11	18.94	19.77	20.60
0.2	11.15	12.00	12.86	13.72	14.58	15.44	16.31	17.19	18.06	18.94	19.82	20.70	21.59	22.48	23.36
0.25	11.91	12.79	13.68	14.56	15.45	16.34	17.24	18.14	19.04	19.94	20.84	21.75	22.66	23.57	24.48
0.4	13.98	14.94	15.89	16.85	17.81	18.77	19.73	20.69	21.65	22.62	23.58	24.54	25.51	26.48	27.44
0.5	15.34	16.34	17.34	18.34	19.34	20.34	21.34	22.34	23.34	24.34	25.34	26.34	27.34	28.34	29.34
0.6	16.78	17.82	18.87	19.91	20.95	21.99	23.03	24.07	25.11	26.14	27.18	28.21	29.25	30.28	31.32
0.75	19.37	20.49	21.60	22.72	23.83	24.93	26.04	27.14	28.24	29.34	30.43	31.53	32.62	33.71	34.80
0.8	20.47	21.61	22.76	23.90	25.04	26.17	27.30	28.43	29.55	30.68	31.79	32.91	34.03	35.14	36.25
0.9	23.54	24.77	25.99	27.20	28.41	29.62	30.81	32.01	33.20	34.38	35.56	36.74	37.92	39.09	40.26
0.95	26.30	27.59	28.87	30.14	31.41	32.67	33.92	35.17	36.41	37.65	38.89	40.11	41.34	42.56	43.77
0.975	28.85	30.19	31.53	32.85	34.17	35.48	36.78	38.08	39.36	40.65	41.92	43.19	44.46	45.72	46.98
0.99	32.00	33.41	34.81	36.19	37.57	38.93	40.29	41.64	42.98	44.31	45.64	46.96	48.28	49.59	50.89
0.995	34.27	35.72	37.16	38.58	40.00	41.40	42.80	44.18	45.56	46.93	48.29	49.64	50.99	52.34	53.67



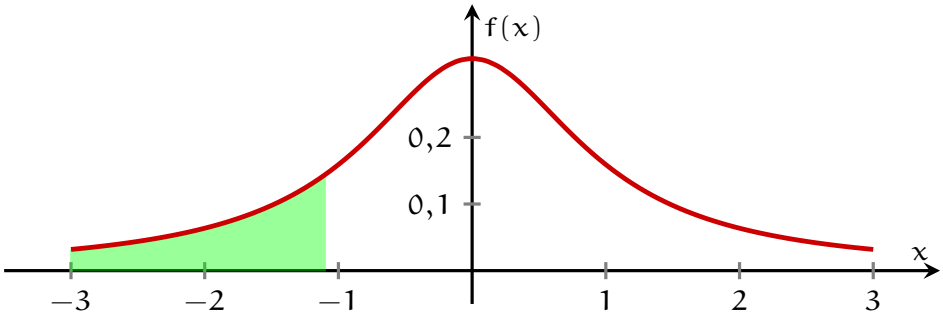
Testverteilungen: t-Verteilung

- Ist $X \sim N(0;1)$, $Z \sim \chi^2(n)$, X, Z unabhängig, so wird die Verteilung von

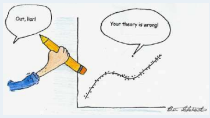
$$T = \frac{X}{\sqrt{\frac{1}{n} Z}}$$

als t-Verteilung mit n Freiheitsgraden bezeichnet.

William Sealy Gosset
1876 – 1937

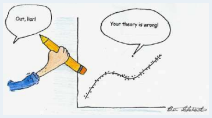


- Kurzschreibweise: $T \sim t(n)$
- **Beispiel:** $t(10)$ $x_{0,6} = 0{,}260$, $x_{0,5} = 0$, $x_{0,1} = -x_{0,9} = -1{,}372$



Quantilstabelle der t-Verteilung mit n Freiheitsgraden

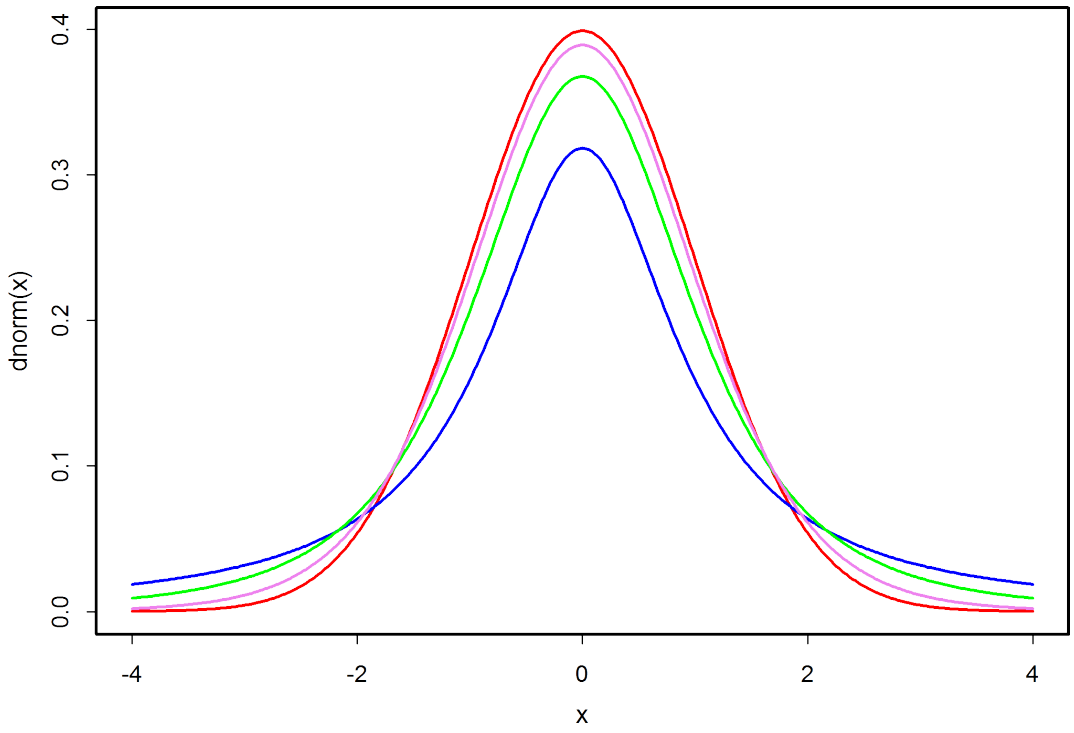
$\alpha \setminus n$	0.6	0.75	0.8	0.9	0.95	0.975	0.99	0.995
1	0.325	1.000	1.376	3.078	6.314	12.706	31.820	63.657
2	0.289	0.816	1.061	1.886	2.920	4.303	6.965	9.925
3	0.277	0.765	0.979	1.638	2.353	3.183	4.541	5.841
4	0.271	0.741	0.941	1.533	2.132	2.776	3.747	4.604
5	0.267	0.727	0.920	1.476	2.015	2.571	3.365	4.032
6	0.265	0.718	0.906	1.440	1.943	2.447	3.143	3.707
7	0.263	0.711	0.896	1.415	1.895	2.365	2.998	3.499
8	0.262	0.706	0.889	1.397	1.860	2.306	2.897	3.355
9	0.261	0.703	0.883	1.383	1.833	2.262	2.821	3.250
10	0.260	0.700	0.879	1.372	1.812	2.228	2.764	3.169
11	0.260	0.698	0.875	1.363	1.796	2.201	2.718	3.106
12	0.259	0.696	0.873	1.356	1.782	2.179	2.681	3.054
13	0.259	0.694	0.870	1.350	1.771	2.160	2.650	3.012
14	0.258	0.692	0.868	1.345	1.761	2.145	2.624	2.977
15	0.258	0.691	0.866	1.341	1.753	2.131	2.603	2.947
16	0.258	0.690	0.865	1.337	1.746	2.120	2.583	2.921
17	0.257	0.689	0.863	1.333	1.740	2.110	2.567	2.898
18	0.257	0.688	0.862	1.330	1.734	2.101	2.552	2.878
19	0.257	0.688	0.861	1.328	1.729	2.093	2.539	2.861
20	0.257	0.687	0.860	1.325	1.725	2.086	2.528	2.845
21	0.257	0.686	0.859	1.323	1.721	2.080	2.518	2.831
22	0.256	0.686	0.858	1.321	1.717	2.074	2.508	2.819
23	0.256	0.685	0.858	1.319	1.714	2.069	2.500	2.807
24	0.256	0.685	0.857	1.318	1.711	2.064	2.492	2.797
25	0.256	0.684	0.856	1.316	1.708	2.059	2.485	2.787
26	0.256	0.684	0.856	1.315	1.706	2.055	2.479	2.779
27	0.256	0.684	0.855	1.314	1.703	2.052	2.473	2.771
28	0.256	0.683	0.855	1.312	1.701	2.048	2.467	2.763
29	0.256	0.683	0.854	1.311	1.699	2.045	2.462	2.756
30	0.256	0.683	0.854	1.310	1.697	2.042	2.457	2.750

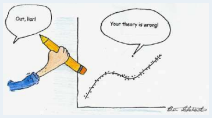


t-Verteilung vs. Normalverteilung

Dichtefunktion

- t-Verteilung mit 1 (blau), 3 (grün) und 10 (lila) Freiheitsgraden
- Standardnormalverteilung (rot)





Punkt-Schätzung

- Ein unbekannter Parameter ϑ der Verteilung von G soll auf Basis einer Stichprobe geschätzt werden.
- Zum Beispiel: σ von $N(10;\sigma)$
- Schätzwert: $\hat{\vartheta}$
- Vorgehen: Verwendung einer Schätzfunktion

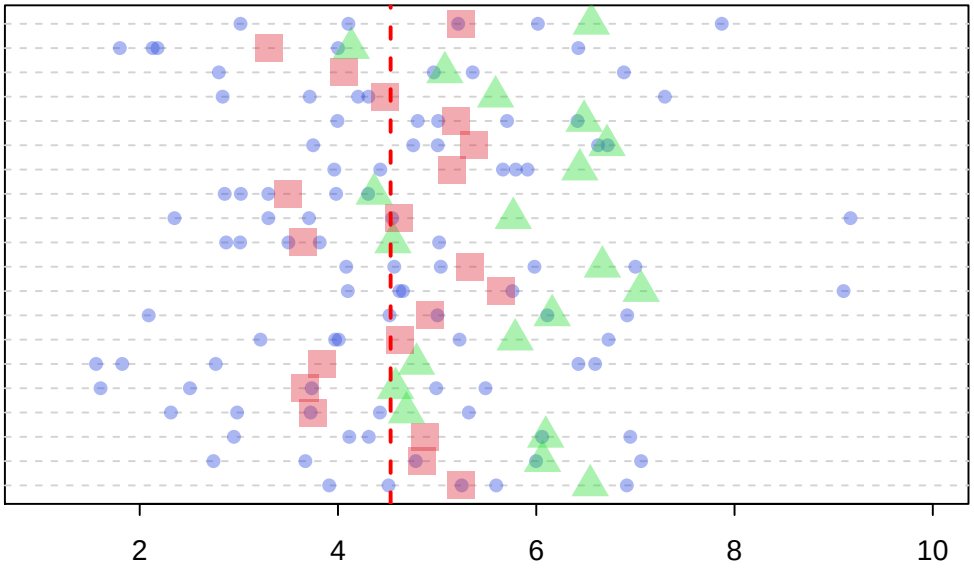
$$\hat{\Theta} = g(X_1, \ldots, X_n)$$

  Beachte: Der Schätzwert $\hat{\vartheta}$ ist die Realisierung der ZV (!) $\hat{\Theta}$.
- Frage: Welche Stichprobenfunktion ist zur Schätzung geeignet?
- ➠ Kriterien für die Beurteilung/Konstruktion von Schätzfunktionen!
- Im Folgenden: Vorliegen einer einfachen Stichprobe, d.h. $X_1, \ldots, X_n$ iid.

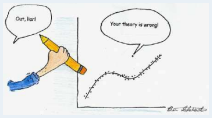
Beispiel

- Schätzen des Mittelwertes einer Grundgesamtheit
- dazu: Einfache Stichprobe vom Umfang 5
- und den beiden Stichprobenfunktionen

$$\hat{\Theta}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \hat{\Theta}_2 = \frac{1}{n-1}\sum_{i=1}^{n} X_i$$



Mittelwert Grundgesamtheit = 4.53



Erwartungstreue und Wirksamkeit

- Eine Schätzfunktion $\hat{\Theta} = g(X_1, \ldots, X_n)$ heißt erwartungstreu oder unverzerrt für ϑ, wenn unabhängig vom numerischen Wert von ϑ gilt:

$$E(\hat{\Theta}) = \vartheta$$

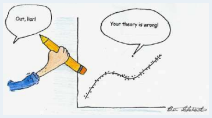
Beispiel

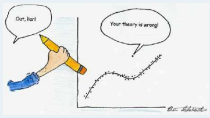
Sind $\hat{\Theta}_1 = \bar{X}$, $\hat{\Theta}_2 = \frac{X_1+X_n}{2}$, $\hat{\Theta}_3 = \frac{1}{n-1}\sum_{i=1}^{n} X_i$ erwartungstreu für μ?

a) $\hat{\Theta}_1$: $E(\bar{X}) = \mu$
$\Rightarrow \hat{\Theta}_1$ ist erwartungstreu.

b) $\hat{\Theta}_2$: $E\left(\frac{X_1+X_n}{2}\right) = \frac{1}{2}[E(X_1) + E(X_n)] = \frac{1}{2}(\mu+\mu) = \mu$
$\Rightarrow \hat{\Theta}_2$ ist erwartungstreu.

c) $\hat{\Theta}_3$: $E\left(\frac{1}{n-1}\sum_{i=1}^{n} X_i\right) = \frac{1}{n-1}\sum_{i=1}^{n} E(X_i) = \frac{1}{n-1}\sum_{i=1}^{n} \mu = \frac{n}{n-1}\mu \neq \mu$
$\Rightarrow \hat{\Theta}_3$ ist nicht erwartungstreu





Erwartungstreue und Wirksamkeit

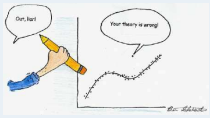
- Welche der erwartungstreuen Schätzfunktionen $\hat{\Theta}_1, \hat{\Theta}_2$ ist „besser“?
- Von zwei erwartungstreuen Schätzfunktionen $\hat{\Theta}_1$, $\hat{\Theta}_2$ für ϑ heißt $\hat{\Theta}_1$ wirksamer als $\hat{\Theta}_2$, wenn unabhängig vom numerischen Wert von ϑ gilt:

$$\mathrm{Var}(\hat{\Theta}_1) < \mathrm{Var}(\hat{\Theta}_2)$$

Beispiel: ($\hat{\Theta}_1 = \bar{X}$, $\hat{\Theta}_2 = \frac{X_1+X_n}{2}$)
Wegen

$$\left.\begin{aligned} \mathrm{Var}(\hat{\Theta}_1) &= \mathrm{Var}(\bar{X}) &&= \frac{\sigma^2}{n} \\ \mathrm{Var}(\hat{\Theta}_2) &= \mathrm{Var}\left(\tfrac{X_1+X_n}{2}\right) = \tfrac{1}{4}(\sigma^2+\sigma^2) &&= \frac{\sigma^2}{2} \end{aligned}\right\} \Rightarrow \mathrm{Var}(\hat{\Theta}_1) < \mathrm{Var}(\hat{\Theta}_2)$$

(falls $n > 2$) ist $\hat{\Theta}_1$ wirksamer als $\hat{\Theta}_2$.

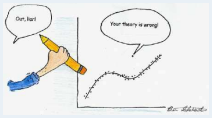


Intervall-Schätzung

- Für einen unbekannten Verteilungsparameter ϑ soll auf Basis einer Stichprobe ein Intervall geschätzt werden.
- Verwendung der Stichprobenfunktionen V_u, V_o, so dass $V_u \leqq V_o$ und

$$P(V_u \leqq \vartheta \leqq V_o) = 1 - \alpha$$

  stets gelten.
  $[V_u; V_o]$ heißt Konfidenzintervall (KI) für ϑ zum Konfidenzniveau $1-\alpha$.
- Beachte: Das Schätzintervall $[v_u; v_o]$ ist Realisierung der Zufallsvariablen (!) V_u, V_o.
  ➠ Irrtumswahrscheinlichkeit α (klein, i.d.R. $\alpha \leqq 0{,}1$)
- Frage: Welche Konfidenzintervalle sind zur Schätzung geeignet?
  ➠ Hängt von Verteilung von G sowie vom unbekannten Parameter (μ, σ^2) ab!
- Im Folgenden: Einfache Stichprobe $X_1, \ldots, X_n$ mit $E(X_i) = \mu$, $\mathrm{Var}(X_i) = \sigma^2$

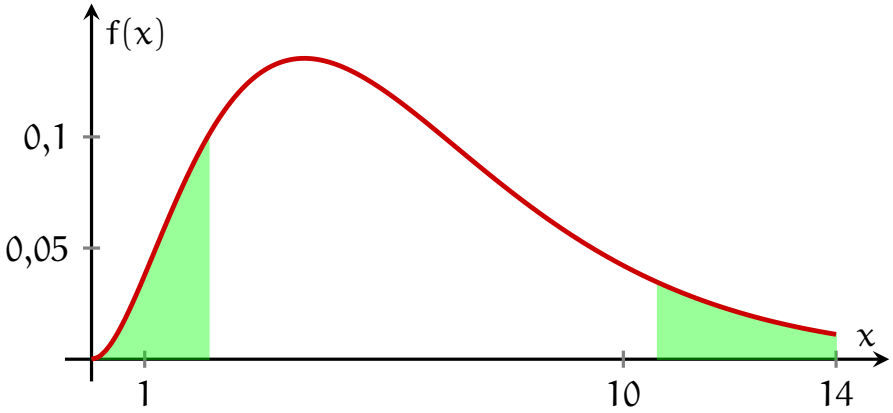


Intervall-Schätzung

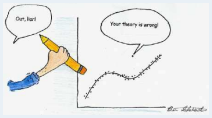
Wichtiger Spezialfall: Symmetrische Konfidenzintervalle

- Symmetrisch heißt **nicht**, dass die Dichte symmetrisch ist, sondern
- übereinstimmende Wahrscheinlichkeiten für Über-/Unterschreiten des Konfidenzintervalls, d.h.

$$P(V_u > \vartheta) = P(V_o < \vartheta) = \frac{\alpha}{2}$$



- **Wichtig:** Eine Verkleinerung von α bewirkt eine Vergrößerung des Konfidenzintervalls.

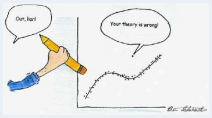


Konfidenzintervall für μ bei Normalverteilung mit bekanntem σ^2

Vorgehensweise:

1. Festlegen des Konfidenzniveaus $1 - \alpha$
2. Bestimmung des $\left(1 - \frac{\alpha}{2}\right)$-Fraktils c der $N(0,1)$-Verteilung
3. Berechnen des Stichprobenmittels $\bar{x}$
4. Berechnen des Wertes $\frac{\sigma c}{\sqrt{n}}$
5. Ergebnis der Intervall-Schätzung:

$$\left[\bar{x} - \frac{\sigma c}{\sqrt{n}} \; ; \quad \bar{x} + \frac{\sigma c}{\sqrt{n}}\right]$$



Beispiel

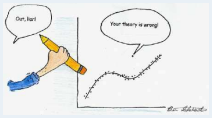
Normalverteilung mit $\sigma = 2{,}4$
$(x_1, \ldots, x_9)$ = (184.2, 182.6, 185.3, 184.5, 186.2, 183.9, 185.0, 187.1, 184.4)

Gesucht: Konfidenzintervall für μ zum Konfidenzniveau $1 - \alpha = 0{,}99$

1. $1 - \alpha = 0{,}99$
2. $N(0;1)$: $c = x_{1-\frac{\alpha}{2}} = x_{1-\frac{0{,}01}{2}} = x_{0{,}995} = 2{,}576$ (Tab. 3; Interpolation)
3. $\bar{x} = \frac{1}{9}(184{,}2 + \cdots + 184{,}4) = 184{,}8$
4. $\frac{\sigma c}{\sqrt{n}} = \frac{2{,}4 \cdot 2{,}576}{\sqrt{9}} = 2{,}06$
5. $KI = [184{,}8 - 2{,}06; 184{,}8 + 2{,}06] = [182{,}74; 186{,}86]$

Interpretation: Mit 99 % Wahrscheinlichkeit ist $\mu \in [182{,}74; 186{,}86]$.

Statistik
Etschberger -
Januar 2015



1. Einführung
2. Deskriptive Statistik
3. W-Theorie
4. Induktive Statistik
Grundlagen
Punkt-Schätzung
Intervall-Schätzung
Signifikanztests
5. Datenanalyse

Wichtige $N(0;1)$-Fraktilswerte:

α	x_α
0,9	1,281552
0,95	1,644854
0,975	1,959964
0,99	2,326348
0,995	2,575829

(I.d.R. genügen drei Nachkommastellen.)

Intervalllänge

- Bei bekannter Standardabweichung gilt offenkundig

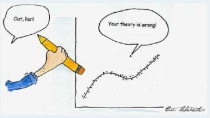
$$L = V_o - V_u = \frac{2\sigma c}{\sqrt{n}}$$

- Welcher Stichprobenumfang n sichert eine vorgegebene (Maximal-)Länge L? $\Rightarrow$ Nach n auflösen! $\Rightarrow$

$$n \geqq \left(\frac{2\sigma c}{L}\right)^2$$

- Eine Halbierung von L erfordert eine Vervierfachung von n!
- Angewendet auf letztes **Beispiel**:

$$L = 4 \Rightarrow n \geqq \left(\frac{2 \cdot 2{,}4 \cdot 2{,}576}{4}\right)^2 = 9{,}556 \Rightarrow n \geqq 10$$
$$L = 2 \Rightarrow n \geqq \left(\frac{2 \cdot 2{,}4 \cdot 2{,}576}{2}\right)^2 = 38{,}222 \Rightarrow n \geqq 39$$



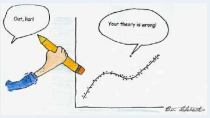
Konfidenzintervall

Konfidenzintervall für μ bei Normalverteilung mit unbekanntem σ^2

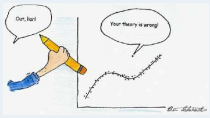
- Vorgehensweise:
  1. Festlegen des Konfidenzniveaus $1-\alpha$
  2. Bestimmung des $\left(1-\frac{\alpha}{2}\right)$-Fraktils c der $t(n-1)$-Verteilung
  3. Berechnen des Stichprobenmittels $\bar{x}$ und der Stichproben-Standardabweichung s
  4. Berechnen des Wertes $\frac{sc}{\sqrt{n}}$
  5. Ergebnis der Intervall-Schätzung:

$$\left[\bar{x} - \frac{sc}{\sqrt{n}}\ ;\quad \bar{x} + \frac{sc}{\sqrt{n}}\right]$$

- Zu Schritt 2: Falls $n-1 > 30$ wird die $N(0;1)$-Verteilung verwendet.



Konfidenzintervalllänge

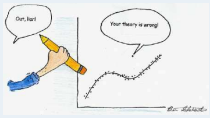


Beispiel:
Wie das letzte Beispiel, jedoch σ unbekannt.

1. $1-\alpha = 0{,}99$
2. $t(8)$: $c = x_{1-\frac{\alpha}{2}} = x_{1-\frac{0{,}01}{2}} = x_{0{,}995} = 3{,}355$ (Tab. 4)
3. $\bar{x} = \frac{1}{9}(184{,}2 + \cdots + 184{,}4) = 184{,}8$
   $s = \sqrt{\frac{1}{8}\left[(184{,}2^2 + \cdots + 184{,}4^2) - 9 \cdot 184{,}8^2\right]} = 1{,}31$
4. $\frac{sc}{\sqrt{n}} = \frac{1{,}31 \cdot 3{,}355}{\sqrt{9}} = 1{,}47$
5. $KI = [184{,}8 - 1{,}47; 184{,}8 + 1{,}47] = [183{,}33; 186{,}27]$

Interpretation: Mit 99 % Wahrscheinlichkeit ist $\mu \in [183{,}33; 186{,}27]$.

R Beispiel




```
x <- c(184.2, 182.6, 185.3, 184.5, 186.2,
       183.9, 185.0, 187.1, 184.4)
t.test(x,conf.level=.99)

##
##  One Sample t-test
##
## data:  x
## t = 422.1129, df = 8, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##  183.331 186.269
## sample estimates:
## mean of x
##     184.8
```

Konfidenzintervall für μ bei beliebiger Verteilung

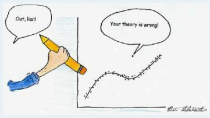
- Voraussetzung: $n > 30$, bzw. falls G dichotom: $5 \leqq \sum_{i=1}^{n} x_i \leqq n-5$
- Vorgehensweise:
  1. Festlegen des Konfidenzniveaus $1-\alpha$
  2. Bestimmung des $\left(1-\frac{\alpha}{2}\right)$-Fraktils c der Standardnormalverteilung $N(0;1)$
  3. Berechnung des Stichprobenmittels $\bar{x}$ sowe eines Schätzwertes $\hat{\sigma}$ für die Standardabweichung σ der GG mittels

$$\hat{\sigma} = \begin{cases} \sigma, & \text{falls } \sigma \text{ bekannt} \\ \sqrt{\bar{x}(1-\bar{x})}, & \text{falls GG dichotom} \\ s, & \text{sonst} \end{cases}$$

  4. Berechnung von $\frac{\hat{\sigma} c}{\sqrt{n}}$
  5. Ergebnis der Intervallschätzung:

$$\left[\bar{x} - \frac{\hat{\sigma} c}{\sqrt{n}}; \bar{x} + \frac{\hat{\sigma} c}{\sqrt{n}}\right]$$

- Zu Schritt 3: Manchmal kann anderer Schätzwert $\hat{\sigma}$ sinnvoller sein.

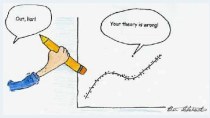


Konfidenzintervall für μ bei beliebiger Verteilung

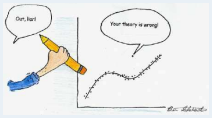
Beispiel:
Poisson-Verteilung mit $\lambda\ (= \mu = \sigma^2)$ unbekannt.
$(x_1, \ldots, x_{40}) = (3; 8; \ldots; 6)$
Gesucht: KI für λ zum Konfidenzniveau $1-\alpha = 0{,}9$

1. $1-\alpha = 0{,}9$
2. $N(0;1): \ c = x_{1-\frac{\alpha}{2}} = x_{1-\frac{0{,}1}{2}} = x_{0{,}95} = 1{,}645$
3. $\bar{x} = \frac{1}{40}(3+8+\cdots+6) = 6{,}5$
   $\hat{\sigma} = \sqrt{\bar{x}} = \sqrt{6{,}5} = 2{,}55 \ \ (\text{da} \ \ \sigma^2 = \lambda)$
4. $\frac{\hat{\sigma} c}{\sqrt{n}} = \frac{2{,}55 \cdot 1{,}645}{\sqrt{40}} = 0{,}66$
5. $KI = [6{,}5 - 0{,}66; 6{,}5 + 0{,}66] = [5{,}84; 7{,}16]$

Statistik
Etschberger - Januar 2015



1. Einführung
2. Deskriptive Statistik
3. W-Theorie
4. Induktive Statistik
Grundlagen
Punkt-Schätzung
Intervall-Schätzung
Signifikanztests
5. Datenanalyse



Konfidenzintervall für σ^2 bei Normalverteilung

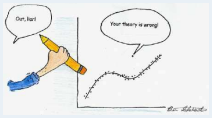
Vorgehensweise

1. Festlegen eines Konfidenzniveaus $1-\alpha$
2. Bestimmung der $\frac{\alpha}{2}$- bzw. $(1-\frac{\alpha}{2})$-Fraktile (c_1 bzw. c_2) der $\chi^2(n-1)$-Verteilung
3. Aus der Stichprobe: Berechnung der Größe

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

4. Berechnung des Konfidenzintervalls

$$\left[\frac{(n-1)s^2}{c_2};\frac{(n-1)s^2}{c_1}\right]$$



KI für σ^2 bei Normalverteilung

Beispiel:

$G \sim N(\mu;\sigma)$;

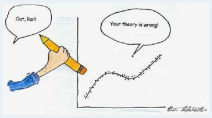
$(x_1,\ldots,x_5) = (1,\ 1.5,\ 2.5,\ 3,\ 2)$

Gesucht: KI für σ^2 zum Konfidenzniveau $1-\alpha = 0{,}99$

1. $1-\alpha = 0{,}99$
2. $\chi^2(5-1):\ c_1 = x_{\frac{\alpha}{2}} = x_{0{,}005} = 0{,}21$
   $c_2 = x_{1-\frac{\alpha}{2}} = x_{0{,}995} = 14{,}86$
3. $\bar{x} = \frac{1}{5}(1+1{,}5+2{,}5+3+2) = 2$
   $\sum_{i=1}^{5}x_i^2 - 5\cdot\bar{x}^2 = 1^2+1{,}5^2+2{,}5^2+3^2+2^2 - 5\cdot 2^2 = 2{,}5$
4. KI $= \left[\frac{2{,}5}{14{,}86};\frac{2{,}5}{0{,}21}\right] = [0{,}17;11{,}9]$

(Extrem groß, da n klein.)

Signifikanztests

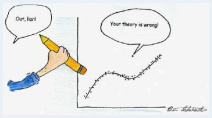


- Vorliegen einer Hypothese über die Verteilung(en) der Grundgesamtheit(en).
- Beispiele:
  - „Der Würfel ist fair."
  - „Die Brenndauern zweier unterschiedlicher Glühbirnensorten sind gleich."
- Hypothese soll anhand einer Stichprobe überprüft werden.
- Prinzip:
  - Hypothese verwerfen, wenn „signifikanter" Widerspruch zur Stichprobe.
  - Ansonsten: Hypothese nicht verwerfen.
- Eine verworfene Hypothese gilt als statistisch widerlegt.
- Nicht-Verwerfung ist dagegen ein „Freispruch aus Mangel an Beweisen".

Zu Beachten:

Nicht-Verwerfung ist **kein** „statistischer Beweis", dass Hypothese wahr ist!
(„Trick": Hypothese falsch $\Longleftrightarrow$ Gegenhypothese wahr!)

Test des Erwartungswertes bei bekannter Varianz in der Grundgesamtheit



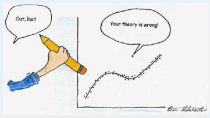
- Zunächst:
  - $G \sim N(\mu;\sigma)$ mit σ bekannt
  - Einfache Stichprobe $X_1,\ldots,X_n$
  - (Null-)Hypothese $H_0: \mu = \mu_0$
- **Beispiel:**
  $X_1,\ldots,X_{25}$ mit X_i = Füllmenge der i-ten Flasche $\sim N(\mu;1{,}5)$
  Nullhypothese $H_0: \mu = 500$, d.h. $\mu_0 = 500$
- Je nach Interessenlage sind unterschiedliche **Gegenhypothesen** möglich:

$$\begin{array}{ll} \text{a)} & H_1: \mu \neq \mu_0 \\ \text{b)} & H_1: \mu < \mu_0 \\ \text{c)} & H_1: \mu > \mu_0 \end{array}$$

- Entscheidung:

$$\begin{array}{lll} & H_0: \mu = \mu_0 & \text{wird abgelehnt gegenüber} \\ \text{a)} & H_1: \mu \neq \mu_0, & \text{wenn } |\bar{x} - \mu_0| \text{ „sehr groß" ist} \\ \text{b)} & H_1: \mu < \mu_0, & \text{wenn } \bar{x} \text{ „weit kleiner" als } \mu_0 \text{ ist} \\ \text{c)} & H_1: \mu > \mu_0, & \text{wenn } \bar{x} \text{ „weit größer" als } \mu_0 \text{ ist} \end{array}$$

Test des Erwartungswertes bei bekannter Varianz in der Grundgesamtheit



Entscheidungskriterium aus Stichprobe:

$$\nu = \frac{\bar{x} - \mu_0}{\sigma}\sqrt{n}$$

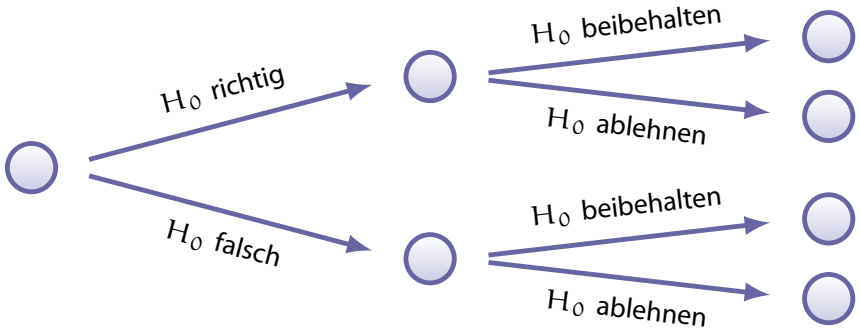
Mögliche Fehlentscheidungen

- **Ablehnung von** H_0, obwohl H_0 richtig ist: Fehler 1. Art
- **Nicht-Ablehnung von** H_0, obwohl H_0 falsch ist: Fehler 2. Art

- Vorteil: Verteilung bekannt: $N(0;1)$
- Dann:

  $H_0 : \mu = \mu_0$ wird abgelehnt gegenüber

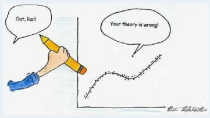
a) $H_1 : \mu \neq \mu_0$, wenn $|\nu|$ „sehr groß“ ist
b) $H_1 : \mu < \mu_0$, wenn ν „sehr negativ“ ist
c) $H_1 : \mu > \mu_0$, wenn ν „sehr positiv“ ist



- Signifikanzniveau α: Maximal erlaubte Wahrscheinlichkeit für einen Fehler 1. Art.

Test des Erwartungswertes bei bekannter Varianz in der Grundgesamtheit

Statistik
Etschberger - Januar 2015



1. Einführung
2. Deskriptive Statistik
3. W-Theorie
4. Induktive Statistik
Grundlagen
Punkt-Schätzung
Intervall-Schätzung
Signifikanztests
5. Datenanalyse

- Mithilfe von α und V kann geklärt werden, was „sehr groß“ usw. heißt:
  Wahrscheinlichkeit für Fehler 1. Art im Fall a): $|\nu| > x$, obwohl H_0 richtig:

$$\begin{aligned} P(|V| > x) &= P(V > x) + P(V < -x) \\ &= 2 \cdot P(V > x) \qquad \text{(Symmetrie der Normalverteilung)} \\ &= 2 \cdot [1 - P(V \leqq x)] = 2 \cdot [1 - \Phi(x)] \overset{!}{=} \alpha \\ &\iff \Phi(x) = 1 - \tfrac{\alpha}{2} \\ &\iff x = x_{1-\frac{\alpha}{2}} \end{aligned}$$

  H_0 wird demnach verworfen,
  wenn $|\nu| > x_{1-\frac{\alpha}{2}}$ bzw. $\nu \in B$ ist.
  $B = (-\infty; -x_{1-\frac{\alpha}{2}}) \cup (x_{1-\frac{\alpha}{2}}; \infty)$ heißt Verwerfungsbereich.
- Analoge Vorgehensweise für die Fälle b) und c)

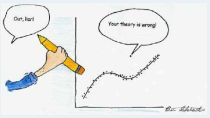
Test des Erwartungswertes bei bekannter Varianz in der GG

Rezept

1. Ein Signifikanzniveau α wird festgelegt.
2. Der Verwerfungsbereich

$$B = (-\infty;\ -x_{1-\alpha/2}) \cup (x_{1-\alpha/2};\ \infty) \quad \text{im Fall a)}$$
$$B = (-\infty;\ -x_{1-\alpha}) \quad \text{im Fall b)}$$
$$B = (x_{1-\alpha};\ \infty) \quad \text{im Fall c)}$$

   wird festgelegt, wobei $x_{1-\alpha/2}$ bzw. $x_{1-\alpha}$ das $(1-\alpha/2)$- bzw. das $(1-\alpha)$-Fraktil der $N(0,1)$-Verteilung ist. (Wichtig: Der Ablehnungsbereich ist also unabhängig von der Stichprobe)
3. Wichtig: Erst jetzt werden die Daten der Stichprobe erhoben/beachtet:
   Der Testfunktionswert $v = \dfrac{\bar{x} - \mu_0}{\sigma}\sqrt{n}$ wird berechnet.
4. H_0 wird genau dann verworfen, wenn $v \in B$ gilt.



Einstichproben-Gaußtest

Beispiel:

$X_1, \ldots, X_{25}$ mit $X_i \sim N(\mu; 1{,}5)$ und $\bar{x} = 499{,}28$

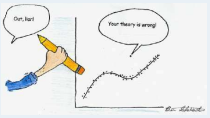
Prüfe $H_0: \mu = 500$, $H_1: \mu \neq 500$ zum Signifikanzniveau $\alpha = 0{,}01$

Lösung: Einstichproben-Gaußtest, Fall a)

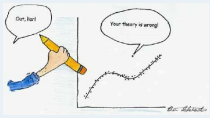
1. $\alpha = 0{,}01$
2. $N(0;1): x_{1-\frac{\alpha}{2}} = x_{1-0{,}005} = x_{0{,}995} = 2{,}576$
   $\Rightarrow B = (-\infty; -2{,}576) \cup (2{,}576; \infty)$
3. $v = \frac{499{,}28-500}{1{,}5} \cdot \sqrt{25} = -2{,}4$
4. $v \notin B \Rightarrow H_0$ nicht verwerfen

Interpretation: Zum Signifikanzniveau 1 % kann der Brauerei keine Abweichung vom Sollwert $\mu_0 = 500$ nachgewiesen werden.

Statistik
Etschberger - Januar 2015



1. Einführung
2. Deskriptive Statistik
3. W-Theorie
4. Induktive Statistik
Grundlagen
Punkt-Schätzung
Intervall-Schätzung
Signifikanztests
5. Datenanalyse

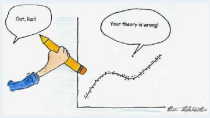


Aufbau und Klassifikation von Signifikanztests

Der jeweils geeignete Test hängt ab von ...

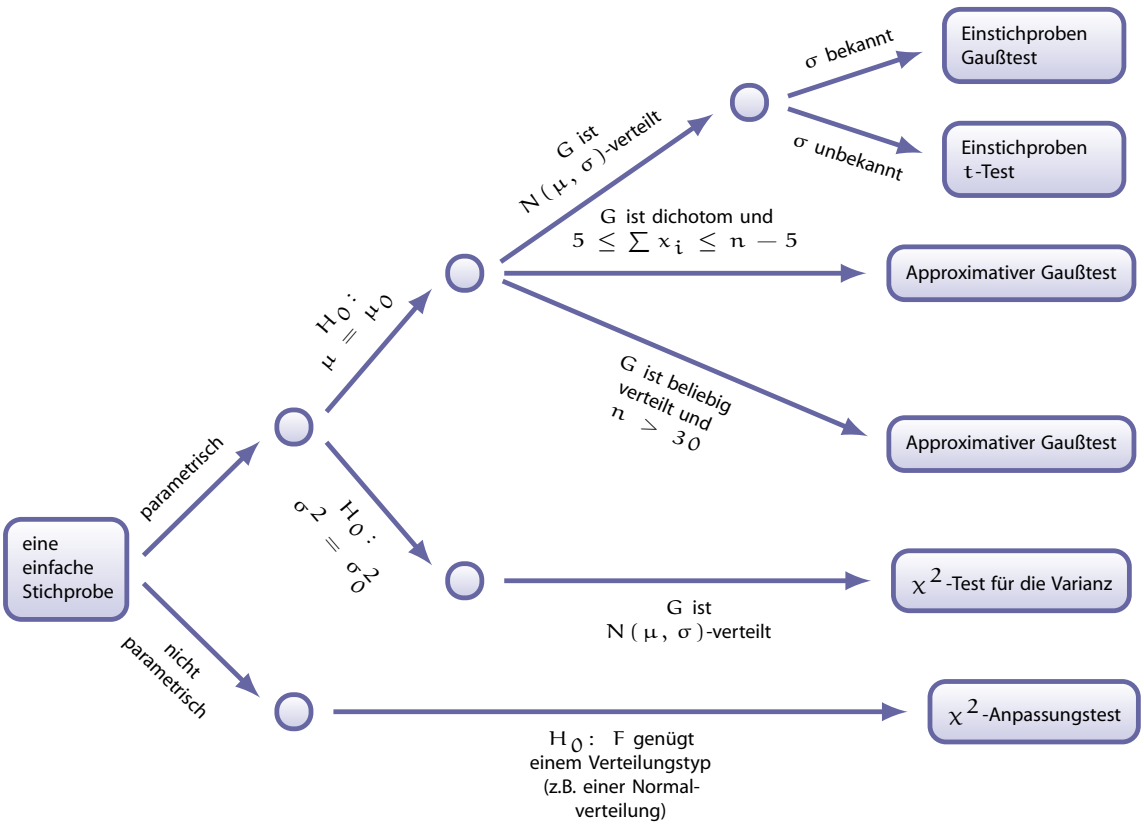
- dem zu testenden Hypothesenpaar H_0, H_1; unterscheide:
  - Parametrische Hypothesen:
    Beziehen sich auf unbekannte(n) Verteilungsparameter $(\mu, \sigma^2, \ldots)$
  - Nichtparametrische Hypothesen:
    Beinhalten sonstige Aussagen, z.B. „Alter und Einkommen sind unabh."
- den Voraussetzungen an die Verteilung/parameter (z.B. $G \sim N(\mu; \sigma)$)
- den Voraussetzungen an den Stichprobenumfang (z.B. $n > 30$)
- Art und Anzahl der Stichproben; unterscheide:
  - Signifikanztests bei einer einfachen Stichprobe
  - Signifikanztests bei mehreren unabhängigen Stichproben
  - Signifikanztests bei zwei verbundenen Stichproben

In dieser Vorlesung: Nur **einfache Stichproben**

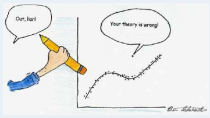


Klassifizierung von Signifikanztests

Signifikanztests bei einer einfachen Stichprobe



(Umfangreichere Übersicht über alle möglichen Fälle siehe Bamberg u. a. (2011), Seite 171ff.)



Einstichproben-t-Test und approximativer Gaußtest

Gegeben:

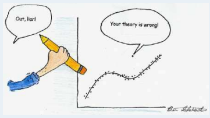
- Einfache Stichprobe $X_1, \ldots, X_n$ mit
- $E(X_i) = \mu$, $Var(X_i) = \sigma^2$

Hypothesenpaare:

a)	$H_0: \mu = \mu_0$		$H_1: \mu \neq \mu_0$
b)	$H_0: \mu = \mu_0$	(oder $\mu \geqq \mu_0$),	$H_1: \mu < \mu_0$
c)	$H_0: \mu = \mu_0$	(oder $\mu \leqq \mu_0$),	$H_1: \mu > \mu_0$

Voraussetzungen:

1. Normalverteilung mit σ unbekannt (Einstichproben-t-Test)
   oder
2. Beliebige Verteilung
   mit $n > 30$ bzw. $5 \leqq \sum x_i \leqq n - 5$ (bei $B(1;p)$)
   (approximativer Gaußtest)



Einstichproben-t-Test, approximativer Gaußtest

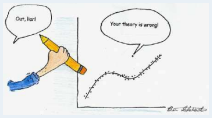
Ablauf:

1. Festlegen des Signifikanzniveaus α
2. Festlegen des Verwerfungsbereichs B:
   - Falls $H_1: \mu \neq \mu_0$: $B = (-\infty; -x_{1-\alpha/2}) \cup (x_{1-\alpha/2}; \infty)$
   - Falls $H_1: \mu < \mu_0$: $B = (-\infty; -x_{1-\alpha})$
   - Falls $H_1: \mu > \mu_0$: $B = (x_{1-\alpha}; \infty)$

   Dabei steht $x_{1-\alpha/2}$ bzw. $x_{1-\alpha}$ für das jeweilige Fraktil
   - der $t(n-1)$-Verteilung bei $n \leq 29$ bzw.
   - der $N(0;1)$-Verteilung bei $n \geq 30$.
3. Berechnen des Testfunktionswertes:

$$\nu = \begin{cases} \dfrac{\bar{x} - \mu_0}{s}\sqrt{n} & \text{falls Grundgesamtheit } N(\mu;\sigma)\text{-verteilt, } \sigma \text{ unbekannt oder falls Verteilung der GG beliebig, } n > 30, \sigma \text{ unbekannt} \\ \dfrac{\bar{x} - \mu_0}{\sigma}\sqrt{n} & \text{falls Verteilung der GG beliebig, } n > 30, \sigma \text{ bekannt} \\ \dfrac{\bar{x} - \mu_0}{\sqrt{\mu_0(1-\mu_0)}}\sqrt{n} & \text{falls GG gemäß } B(1;\mu)\text{-verteilt, } n > 30 \end{cases}$$

Einstichproben-t-Test: Beispiel



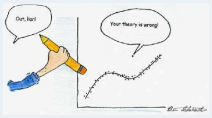
Beispiel t-Test: Energieaufnahme von Frauen

- Empfohlene täglich Energieaufnahme für Frauen: 7724 kJ (1845 kcal)
- Nehme einfache Stichprobe von 11 Frauen und teste zum Signifkanzniveau $\alpha = 0{,}05$ für
- H_0: „Der Erwartungswert der täglichen Energieaufnahme für Frauen ist 7724 kJ“ (μ_0)
- gegen $H_1: \mu \neq \mu_0$

```
daily.intake <- c(5260, 5470, 5640, 6180, 6390, 6515, 6805, 7515, 7515, 8230, 8770)
t.test(daily.intake, alternative="two.sided", mu=7724, conf.level=0.95)
```

```
##
##  One Sample t-test
##
## data:  daily.intake
## t = -2.8179, df = 10, p-value = 0.01823
## alternative hypothesis: true mean is not equal to 7724
## 95 percent confidence interval:
##  5986.348 7520.925
## sample estimates:
## mean of x
##  6753.636
```


Einstichproben-t-Test, approx. Gaußtest



Beispiel:

$X_1, \ldots, X_{2000} \sim B(1; p)$ mit

$$X_i = \begin{cases} 1, & \text{falls i-te Person Wähler einer bestimmten Partei} \\ 0, & \text{sonst} \end{cases}$$

Ergebnis der Stichprobe: $\sum_{i=1}^{2000} x_i = 108$

Prüfe $H_0: p \leqq 0{,}05$ gegen $H_1: p > 0{,}05$ zum Signifikanzniveau 2 %

Lösung:

approximativer Gaußtest bei dichotomer (zweiwertiger) Verteilung; Voraussetzung 2 erfüllt: $5 \leqq 108 \leqq 2000 - 5$

1. $\alpha = 0{,}02$
2. $N(0; 1): \; x_{1-\alpha} = x_{0{,}98} = 2{,}05$ (Tabelle) $\Rightarrow B = (2{,}05; \infty)$
3. $v = \frac{\frac{108}{2000} - 0{,}05}{\sqrt{0{,}05 \cdot (1 - 0{,}05)}} \sqrt{2000} = 0{,}82$
4. $v \notin B \Rightarrow H_0$ nicht verwerfen

Zusatzfrage: Entscheidung, falls $\alpha = 0{,}01$? $\rightarrow$ Keine Änderung!

Chi-Quadrat-Test für die Varianz

- Gegeben: Einfache Stichprobe $X_1, \ldots, X_n \sim N(\mu; \sigma)$
- Hypothesenpaare:

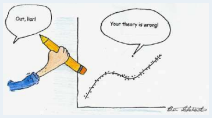
a)	$H_0: \sigma^2 = \sigma_0^2$		$H_1: \sigma^2 \neq \sigma_0^2$
b)	$H_0: \sigma^2 = \sigma_0^2$	(oder $\sigma^2 \geqq \sigma_0^2$),	$H_1: \sigma^2 < \sigma_0^2$
c)	$H_0: \sigma^2 = \sigma_0^2$	(oder $\sigma^2 \leqq \sigma_0^2$),	$H_1: \sigma^2 > \sigma_0^2$

- **Vorgehensweise:**
  1. Festlegen des Signifikanzniveaus α.
  2. Festlegen des Verwerfungsbereichs:

$$B = [0; x_{\alpha/2}) \cup (x_{1-\alpha/2}; \infty) \quad \text{im Fall a)}$$
$$B = [0; x_{\alpha}) \quad \text{im Fall b)}$$
$$B = (x_{1-\alpha}; \infty) \quad \text{im Fall c)}$$

  3. Berechnung des Testfunktionswertes:

$$\nu = \frac{(n-1)s^2}{\sigma_0^2} = \frac{1}{\sigma_0^2} \sum_{i=1}^{n} (x_i - \bar{x})^2$$



Chi-Quadrat-Test für die Varianz

Beispiel: $G \sim N(\mu; \sigma)$

$$(x_1, \ldots, x_{10}) = (2100; 2130; 2150; 2170; 2210; 2070; 2230; 2150; 2230; 2200)$$

Prüfe $H_0: \sigma = 40$, $H_1: \sigma \neq 40$ zum Signifikanzniveau $\alpha = 0{,}1$

Lösung: χ^2-Test für die Varianz, Hypothese Fall a);
Voraussetzungen sind erfüllt

1. $\alpha = 0{,}1$
2. $\chi^2(9): \; x_{\frac{\alpha}{2}} = x_{0{,}05} = 3{,}33; \; x_{1-\frac{\alpha}{2}} = x_{0{,}95} = 16{,}92$
   (Tabelle der χ^2-Verteilung)

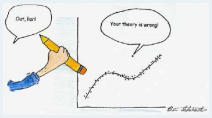
$$\Rightarrow B = [0; 3{,}33) \cup (16{,}92; \infty)$$

3. $\bar{x} = \frac{1}{10}(2100 + 2130 + \cdots + 2200) = 2164$

   $\nu = \frac{1}{40^2}[(2100 - 2164)^2 + \cdots + (2200 - 2164)^2] = 16{,}65$

   $\Rightarrow \nu \notin B \Rightarrow H_0$ nicht verwerfen

Statistik
Etschberger - Januar 2015



1. Einführung
2. Deskriptive Statistik
3. W-Theorie
4. Induktive Statistik
Grundlagen
Punkt-Schätzung
Intervall-Schätzung
Signifikanztests
5. Datenanalyse

Zwei verbundene einfache Stichproben: Kontingenztest

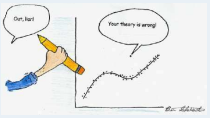
- Situation: In Grundgesamtheit G: Zwei verbundene einfache Stichproben, also Beobachtung zweier Merkmale X, Y
- Hypothese:

H_0 : Die beiden Merkmale X und Y sind in G unabhängig.
H_1 : X und Y sind in G abhängig.

Vorgehensweise Kontingenztest:

1. Festlegen des Signifikanzniveaus α.
2. Unterteilung der x-Achse in $k \geq 2$ und die y-Achse in $l \geq 2$ disjunkte, aneinander angrenzende Intervalle $A_1, \ldots, A_k$ bzw. $B_1, \ldots, B_l$
3. Erstellen einer Kontingenztabelle mit Randhäufigkeiten:

$x\downarrow$ $y\rightarrow$	B_1	B_2	$\cdots$	B_l	
A_1	h_{11}	h_{12}	$\cdots$	h_{1l}	$h_{1\bullet}$
A_2	h_{21}	h_{22}	$\cdots$	h_{2l}	$h_{2\bullet}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	
A_k	h_{k1}	h_{k2}	$\cdots$	h_{kl}	$h_{k\bullet}$
	$h_{\bullet 1}$	$h_{\bullet 2}$	$\cdots$	$h_{\bullet l}$	n



Zwei verbundene einfache Stichproben: Kontingenztest

Vorgehensweise Kontingenztest (Fortsetzung):

4. Mit dem Fraktilswert $x_{1-\alpha}$ der χ^2-Verteilung mit $(k-1)\cdot(l-1)$ Freiheitsgraden: Berechnung des Verwerfungsbereichs

$$B = (x_{1-\alpha}; \infty)$$

5. Zu jeder Kombination aus $i = 1, \ldots, k$ und $j = 1, \ldots, l$: Berechnung der Größe

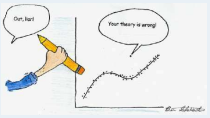
$$\tilde{h}_{ij} = \frac{h_{i\bullet} \cdot h_{\bullet j}}{n}$$

6. Berechnung des Testfunktionswerts v:

$$v = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(\tilde{h}_{ij} - h_{ij}\right)^2}{\tilde{h}_{ij}} = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{h_{ij}^2}{\tilde{h}_{ij}} - n$$

7. Ablehnung von H_0 genau dann, wenn $v \in B$.

Statistik
Etschberger - Januar 2015



1. Einführung
2. Deskriptive Statistik
3. W-Theorie
4. Induktive Statistik
Grundlagen
Punkt-Schätzung
Intervall-Schätzung
Signifikanztests
5. Datenanalyse

Zwei verbundene einfache Stichproben: Kontingenztest

Kontingenztest: Beispiel

- 400 Erstkandidaten einer praktischen Führerscheinprüfung schneiden abhängig von der besuchten Fahrschule folgendermaßen ab:

	Fahrschule		
	A	B	C
bestanden	130	88	62
durchgefallen	70	38	12

- Zum Signifikanzniveau von 5 % soll getestet werden, ob das Bestehen der Prüfung unabhängig von der besuchten Fahrschule ist.

Testdurchführung

1. Signifikanzniveau $\alpha = 5\,\%$
2. entfällt, da Skalenniveau nominal
3. Kontingenztabelle:

	A	B	C	$\sum$
best.	130	88	62	280
durchg.	70	38	12	120
$\sum$	200	126	74	400

4. Berechnung der $\tilde{h}_{ij}$:

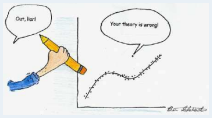
	A	B	C
best.	140	88,2	51,8
durchg.	60	37,8	22,2

5. χ^2-Verteilung mit $(3-1)\cdot(2-1) = 2$ Freiheitsgraden: $x_{1-0,05} = x_{0,95} = 5,99$:

$$B = (5,99; \infty)$$

6. $$\nu = \frac{(130-140)^2}{140} + \ldots + \frac{(12-22,2)^2}{22,2} \approx 9,077$$

7. $\nu \in B$: Also wird H_0 abgelehnt, die Prüfungsergebnisse sind abhängig von der Fahrschule.



Statistik: Table of Contents

Problemstellung

- Synonym: Multivariate Datenanalyse, Numerische Taxonomie, Multivariatenanalyse
- Aufgaben: Analyse von **Zusammenhängen und Ähnlichkeitsbeziehungen** zwischen Elementen einer bestimmten Menge
- Teilgebiet der **Statistik**
- Einsatz sinnvoll bei großen Datenmengen mit mehr als einem Merkmal
- Ausgangspunkt: Datenmatrix oder Distanzmatrix.

Die Datenmatrix

- enthält zeilenweise Objekte (Merkmalsträger, cases)
- enthält spaltenweise Merkmale (variables, items)

Beispiel

	type	income	education	prestige
engineer	prof	72	86	88
insurance.agent	wc	55	71	41
lawyer	prof	76	98	89
dentist	prof	80	100	90
mail.carrier	wc	48	55	34

(Auszug aus Daten von Duncan (1961))

Die Distanzmatrix

- enthält zeilen- und spaltenweise Objekte.
- Die Einträge der Matrix sind Werte für die Verschiedenheit (Distanzen) zweier Objekte.

Beispiel

	engineer	insurance.agent	lawyer	dentist	mail.carrier
engineer	0.000	0.676	0.102	0.149	0.851
insurance.agent	0.676	0.000	0.778	0.825	0.175
lawyer	0.102	0.778	0.000	0.047	0.953
dentist	0.149	0.825	0.047	0.000	1.000
mail.carrier	0.851	0.175	0.953	1.000	0.000

3 Teilbereiche der Datenanalyse nach dem Zweck der Anwendung

Datenverdichtende Verfahren (**deskriptiv**)

- Kennzahlen
- Indizes
- Faktorenanalyse

Strukturaufdeckende Verfahren (**explorativ**)

- Kreuztabellen
- Faktorenanalyse
- Clusteranalyse
- MDS
- Korrespondenzanalyse

Strukturprüfende Verfahren (**induktiv**)

- Varianzanalyse
- Regressionsanalyse
- logistische Regression
- Diskriminanzanalyse
- Conjoint-Analyse

Klassische Anwendungsbereiche der Datenanalyse

Marketing/ Marktforschung

- Marktsegmentierung
- Kundentypisierung
- Aufdecken von Marktnischen
- Ermittlung von Marktreaktionen

Sozialwissenschaften

- Einstellungsanalysen
- Qualifikationsprofile

Biologie

- Zuordnung von Pflanzen oder Tieren zu Gattungen

Medizin

- Hilfe bei Diagnosen
- Überprüfung von Therapieerfolgen

Volkswirtschaft

- Input-Output-Analysen zur Abgrenzung und Aggregation von Wirtschaftssektoren

Bibliothekswesen

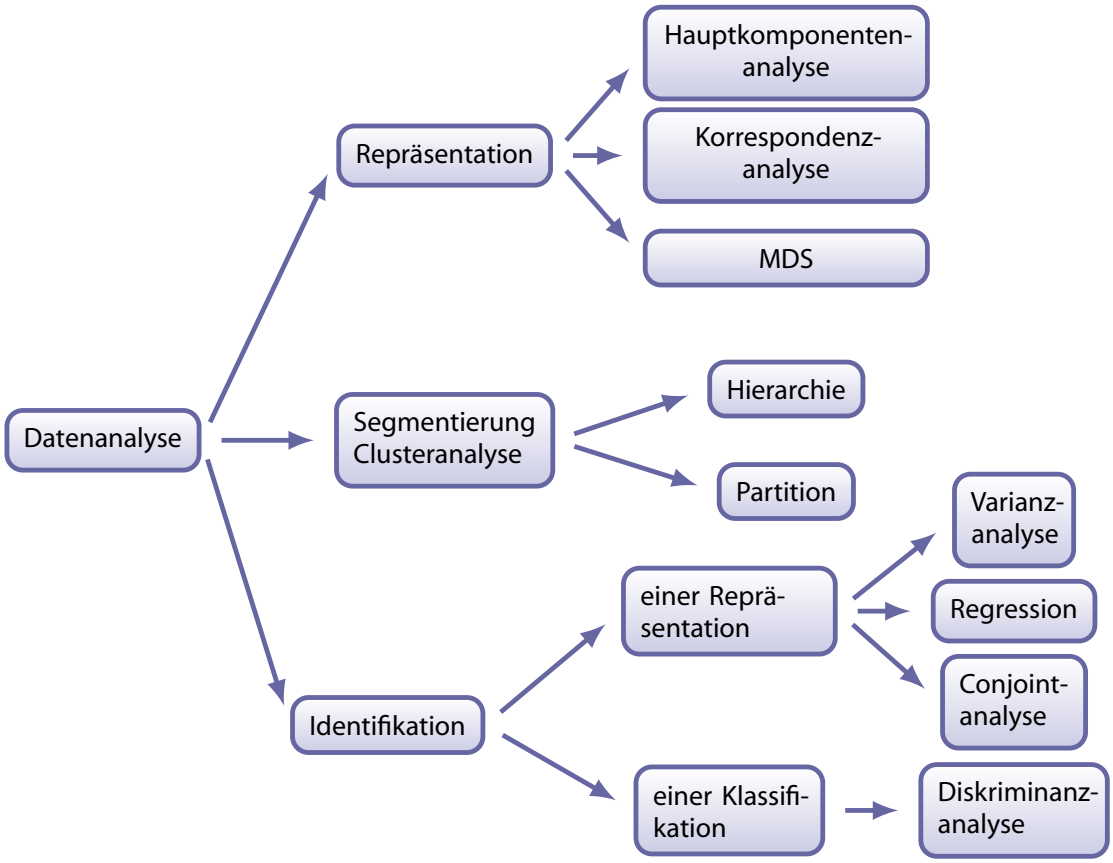
- Katalogisierung
- Auffinden von ähnlichen Werken

Dreiteilung

Die klassische Dreiteilung der Datenanalyse

- Segmentierung (Clusteranalyse): Zusammenfassung von Objekten zu homogenen Klassen aufgrund von Ähnlichkeiten in wichtigen Merkmalsbereichen
- Repräsentation: Darstellung von Objekten durch Punkte im 2- oder 3-dimensionalen Raum, wobei Ähnlichkeitsbeziehungen durch räumliche Nähe zum Ausdruck kommen sollen
- Identifikation: Reproduktion einer gegebenen Segmentierung oder Repräsentation mit Hilfe weniger aussagekräftiger Merkmale (Ziel: Prognose, Klassifikation)

Dreiteilung: Methoden



Ablauf einer datenanalytischen Untersuchung

1. Präzisierung des Untersuchungsziels

- Formulierung der Zielsetzung
- Abgrenzung der Untersuchungsobjekte
- Ableitung der taxonomischen Aufgabenstellung
  - Segmentierung
  - Repräsentation
  - Identifikation

2. Diskussion der Datenbasis

- Auswahl der Merkmale
- Festlegung des Skalenniveaus oder
- Charakterisierung der Objekte durch direkte Vergleiche

3. Datenerhebung und -erfassung

- Primär- oder Sekundärerhebung
- Vollerhebung oder Teilerhebung (Stichprobenauswahl!)
- Datencodierung und ggf. Dateneingabe in DV-Systeme

4. Datenanalyse

- Univariate Datenanalyse
  (Screening, erster Einblick in die Merkmalsstruktur, Plausibilitätsprüfung)
  $\longrightarrow$ Deskriptive Verfahren
- Multivariate Datenanalyse
  (nicht 'statistics all', sondern Verfahrenseinsatz nach Aufgabenstellung und Zielsetzung)
  $\longrightarrow$ Explorative und induktive Verfahren

5. Interpretation der Ergebnisse

- Klassenstatistiken und Bezeichnungen bei Clusteranalysen
- Benennung der Achsen bei Repräsentationsverfahren
- Überprüfen der Modellqualität z.B. mittels Test- bzw. Validierungsdaten bei Identifikationsverfahren