Outline I

Einführung

Hal Varian (Google Chief Economist), 2009:

„The sexy job in the next ten years will be statisticians “

Die Statistik ist wie die Mathematik eine Grundlagenwissenschaft mit herausragender Bedeutung für die wissenschaftliche Auswertung von Datensätzen in der Ökonomie, den Naturwissenschaften und der Medizin.

Statistik ist die Wissenschaft vom Sammeln, Aufbereiten, Darstellen, Analysieren und Interpretieren von Fakten und Zahlen.

Problem: Statistische Auswertungen werden vielfach bewusst, aber auch unbewusst verfälscht.

- Traue keiner Statistik, die Du nicht selber gefälscht hast (Winston Churchill)

Wofür ist Statistik wichtig?

- Statistik & Mathe sind die Sprachen der Wirtschaftswissenschaften
- Besseres Verständnis anderer Fächer
- Wissenschaftliches Arbeiten (Seminararbeiten, Bachelor-/Masterarbeit, Forschung)

Was sind das für Berufe?

- Statistiker/Ökonometriker/Data Scientists
- Versicherungen: Finanzmathematik/Statistik
- Banken und Investment Fonds: empirische Finanzwirtschaft
- Marktforschung und Marketing
- Qualitätsmanagement

Unterteilung der Statistik

Deskriptive (Beschreibende) Statistik:

- Aufbereitung einer Datenmenge mittels Tabellen und Grafiken
- Charakterisierung der Daten durch einige wenige, jedoch aussagekräftige Kenngrößen

Wahrscheinlichkeitstheorie:

- Bietet Methoden für die Arbeit mit Zufallszahlen.

Induktive (Schließende) Statistik:

- Rückschluss von einer zufällig ausgewählten Teilmenge auf die Gesamtheit (z. B. Wahlverhalten)
- Verwendung von Modellbildungen, die mittels Methoden der **Wahrscheinlichkeitstheorie** untersucht werden

Übersicht

Vorlesung

Mo	10:00 - 11:30	Sigma Park

Übungen

Di	08:15 - 09:45	FW 1106
Di	10:00 - 11:30	HW 1002
Di	10:00 - 11:30	FW 2106
Mi	12:15 - 13:45	FW 1004
Mi	14:00 - 15:30	HW 1002
Mi	14:00 - 15:30	HW 1106

Tutorium

- Ab Mitte des Semesters in den Räumen des Lern- und Servicezentrums (LSZ)
- Termin: Siehe Website des LSZ

Klausur:

- 90-minütig; 5 CP; 4 Aufgaben; Nachholklausur im nachfolgenden Semester
- Hilfsmittel:
  - vom Lehrstuhl bereitgestellte Formelsammlung: in dieser sind Unterstreichungen und Hervorhebungen zulässig, aber keine eigenen Eintragungen, Zusatzseiten, beschriebene Post-Its etc.
  - nicht-programmierbarer Taschenrechner

Für die **erfolgreiche Teilnahme** empfehlen wir:

- Besuch der **Vorlesungen** sowie **Nacharbeit** des Stoffes;
- aktive Teilnahme an den **Übungen**;
- eigenständiges Lösen der **Altklausuren**.

Hilfreiche Unterlagen:

- Foliensatz Statistik I
- Übungsaufgabensammlung Statistik I
- Klausuraufgabensammlung Statistik I
- Formelsammlung

Empfohlene Literatur

- Bamberg, G., Baur, F. und Krapp, M.: Statistik, *Oldenbourg-Verlag, München, 17. Auflage, 2012*
- Bamberg, G., Baur, F. und Krapp, M.: Arbeitsbuch Statistik, *Oldenbourg-Verlag, München, 9. Auflage, 2012*
- Schira, J.: Statistische Methoden der VWL und BWL. *Pearson Studium, München, 2007*
- Fahrmeir, L., Künstler, R. , Pigeot, I. und Tutz, G.: Statistik. *Springer, Berlin, 2003*
- Schlittgen, R.: Einführung in die Statistik. *Oldenbourg Verlag, München, 2003*
- Dalgaard, P.: Introductory Statistics with R. *Springer, New York, 2008*

1. Deskriptive Statistik

1.1. Statistische Grundlagen

- **Statistische Einheiten** oder **Merkmalsträger**: die Objekte, deren Charakteristiken/Merkmale untersucht werden

  Beispiel: Branchen, Kraftfahrzeuge, Produkte, Menschen, Arbeitnehmer, usw.

- Die Menge aller Merkmalsträger, auf die sich die Untersuchung bezieht, heißt **Grundgesamtheit** bzw. **Population** (Symbol: Ω). Die Elemente von Ω bezeichnen wir mit ω.

  Beispiel: Kraftfahrzeuge eines bestimmten Typs, Verkehrsunfälle im Jahr 2009 in Bayern, Mitarbeiter eines Unternehmen

- **Beachte:** Man interessiert sich nicht direkt für die einzelnen Merkmalsträger, sondern für eine oder mehrere Eigenschaften/Größen X (**Merkmale** oder **Variablen**).

  Beispiel: Familienstand, Körpergewicht, Geschlecht

- Konkret beobachteter Wert $x = X(\omega)$ des Merkmals heißt **(Merkmals-) Ausprägung, Realisation** oder **Beobachtung**.

  Beispiel: {ledig, verheiratet, geschieden, verwitwet}, $x \in [40, 200]$, {männlich, weiblich}

1.1.1. Merkmalstypen

Typen von Merkmalen:

- **qualitativ vs. quantitativ**
  - **quantitativ**: die Ausprägungen sind Zahlen (Schuhgröße, Einkommen, Verkaufszahlen)
  - **qualitativ**: alle anderen (Familienstand, Geschlecht, Augenfarbe)
  - qualitative Merkmale sind quantifizierbar (weiblich 1, männlich 0)

- **diskret vs. stetig**
  - **diskret**: abzählbar viele unterschiedliche Ausprägungen (Schulnoten, Familienstand)
  - **stetig**: alle Zwischenwerte sind realisierbar (Körpergewicht, Temperatur)
  - manche diskrete Merkmale weisen sehr viele Ausprägungen auf (Preis, Einkommen) $\rightsquigarrow$ behandle derartige Merkmale wie stetige.

1.1.2. Datenerhebung und Datengewinnung

- **Totalerhebung**: Alle Untersuchungseinheiten aus Ω werden in der Studie erfasst (z. B. Volkszählung).

  Nachteil: zu teuer, zu aufwendig, prinzipiell nicht immer möglich (z. B. Lebensdauer einer Glühbirne)

- **Teilerhebung**: Greife auf eine Teilmenge der Grundgesamtheit zurück.

- Die Menge der erhaltenen Realisationen heißt **Stichprobe**.

1.1.3. Skalierung

- **Nominalskala** (**klassifikatorische Skala**):
  Bezeichnen x und y Realisationen des Merkmals, so ist lediglich

  $$x = y \text{ (Gleichheit) bzw. } x \neq y \text{ (Ungleichheit)}$$

  interpretierbar.

  Beispiel: Familienstand, Geschlecht, Beruf

- **Ordinalskala**: zusätzlich ist eine Ordnungsrelation erklärt, d. h. es ist „kleiner“ und „größer“ sinnvoll interpretierbar.

  Somit gilt für alle Realisationen x und y

  $$x = y \quad \text{oder} \quad x > y \quad \text{oder} \quad x < y.$$

  Beispiel: Noten, Handelsklassen, Rating-Urteile

- **Kardinalskala** oder **metrische Skala**: zusätzlich hat die absolute Differenz zwischen zwei Beobachtungen eine inhaltliche Bedeutung (ist interpretierbar).

  Beispiel: Temperaturangaben in Celsius, Geburtsjahrgang, Einkommen, Preis, Umsatz, Alter

Beispiel (Größte Unternehmen): Liste der größten Unternehmen gemessen am Gewinn, Umsatz, Vermögenswert und Marktwert (in Mrd. USD).

```
## install.packages("HSAUR")  # Nur bei der ersten Verwendung
data("Forbes2000", package = "HSAUR")  # Datensatz laden
## ??Forbes2000 # Öffnet Hilfeseite zum Datensatz
head(Forbes2000)  # Zeigt ersten 6 Zeilen des Datensatzes
##   rank                 name        country
## 1    1            Citigroup  United States
## 2    2     General Electric  United States
## 3    3 American Intl Group  United States
## 4    4           ExxonMobil  United States
## 5    5                   BP United Kingdom
## 6    6      Bank of America  United States
##                category  sales profits  assets marketvalue
## 1               Banking  94.71   17.85 1264.03      255.30
## 2        Conglomerates 134.19   15.59  626.93      328.54
## 3             Insurance  76.66    6.46  647.66      194.87
## 4 Oil & gas operations 222.88   20.96  166.99      277.02
## 5 Oil & gas operations 232.57   10.27  177.57      173.54
## 6               Banking  49.01   10.81  736.45      117.55
## View(Forbes2000)  # Zeigt den ganzen Datensatz
```

Beispiel: Größte Unternehmen

```
# Betrachte die 500 größten Unternehmen der G7 Länder
G7 <- c("Germany", "France", "Italy", "Japan", "Canada", "United Kingdom",
    "United States")
ForbesG7 <- Forbes2000[Forbes2000$country %in% G7, ]
ForbesG7 <- ForbesG7[1:500, ]
ForbesG7 <- droplevels(ForbesG7)

# Übersicht über den Datensatz
str(ForbesG7)
## 'data.frame': 500 obs. of  8 variables:
##  $ rank       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ name       : chr  "Citigroup" "General Electric" "American Intl Grou
##  $ country    : Factor w/ 7 levels "Canada","France",..: 7 7 7 7 6 7 6
##  $ category   : Factor w/ 27 levels "Aerospace & defense",..: 2 6 16 19
##  $ sales      : num  94.7 134.2 76.7 222.9 232.6 ...
##  $ profits    : num  17.85 15.59 6.46 20.96 10.27 ...
##  $ assets     : num  1264 627 648 167 178 ...
##  $ marketvalue: num  255 329 195 277 174 ...
```

Beispiel: Größte Unternehmen

```
summary(ForbesG7)
##       rank           name                      country
##  Min.   :  1.0   Length:500         Canada         : 23
##  1st Qu.:162.8   Class :character   France         : 33
##  Median :315.5   Mode  :character   Germany        : 31
##  Mean   :325.1                      Italy          : 14
##  3rd Qu.:493.2                      Japan          : 83
##  Max.   :664.0                      United Kingdom: 51
##                                     United States :265
##                    category       sales
##  Banking                : 66   Min.   :  1.470
##  Utilities              : 42   1st Qu.:  8.375
##  Insurance              : 37   Median : 14.190
##  Consumer durables      : 32   Mean   : 23.605
##  Diversified financials: 28   3rd Qu.: 27.540
##  Food drink & tobacco   : 28   Max.   :256.330
##  (Other)                :267
##     profits            assets          marketvalue
##  Min.   :-25.830   Min.   :   3.36   Min.   :  0.940
##  1st Qu.:  0.360   1st Qu.:  13.91   1st Qu.:  8.828
##  Median :  0.650   Median :  26.02   Median : 14.560
##  Mean   :  1.086   Mean   :  85.85   Mean   : 28.805
##  3rd Qu.:  1.383   3rd Qu.:  64.99   3rd Qu.: 29.858
##  Max.   : 20.960   Max.   :1264.03   Max.   :328.540
##
```

1.2. Darstellung und Kenngrößen univariater Datenmengen

Häufigkeitsverteilung

Beispiel (Lieferzeiten): In einem Unternehmen wird festgestellt, dass der Hauptlieferant für die letzten 50 Bestellungen folgende Lieferzeiten (in Tagen) benötigt hat:

7	8	7	3	8	7	5	7	8	9	9	8	8	7	10	7	9	8	9	7	8	7	10	8	8
9	10	7	10	9	9	10	7	8	7	10	10	8	8	8	8	9	9	7	8	5	8	7	10	8

- Merkmalsträger: Lieferungen des Hauptlieferanten
- Merkmal: Lieferzeit (in Tagen)
- Stichprobe: die ausgewählten Lieferzeiten
- Die Menge der beobachteten Ausprägungen: $\{3, 5, 7, 8, 9, 10\}$

Die Liste ist unübersichtlich $\rightsquigarrow$ grafische Darstellung

Ausgangspunkt: ein Merkmal X

- Stichprobe $x_1, ..., x_n$ mit i.a. $x_i \in \mathbb{R}$ (univariat)
- $a_1, ..., a_k$ bezeichnen die verschiedenen Ausprägungen in der Stichprobe ($k \leq n$)

absolute Häufigkeit von a_i:

$n(a_i)$ = Häufigkeit, mit der die Ausprägung a_i in der Stichprobe $x_1, ..., x_n$ auftritt

relative Häufigkeit von a_i:

$$h(a_i) = n(a_i)/n$$

Beachte: $\sum_{i=1}^{k} n(a_i) = n$ und $\sum_{i=1}^{k} h(a_i) = 1$.

1.2.1. Stab– und Säulendiagramm, Kreissektorendiagramm

Stab- und Säulendiagramm

Stabdiagramm: Über jeder Ausprägung werden die zugehörigen relativen (absoluten) Häufigkeiten in Form von Stäben aufgetragen.

Säulendiagramm: Anstelle von Stäben werden Säulen über den Ausprägungen abgetragen.

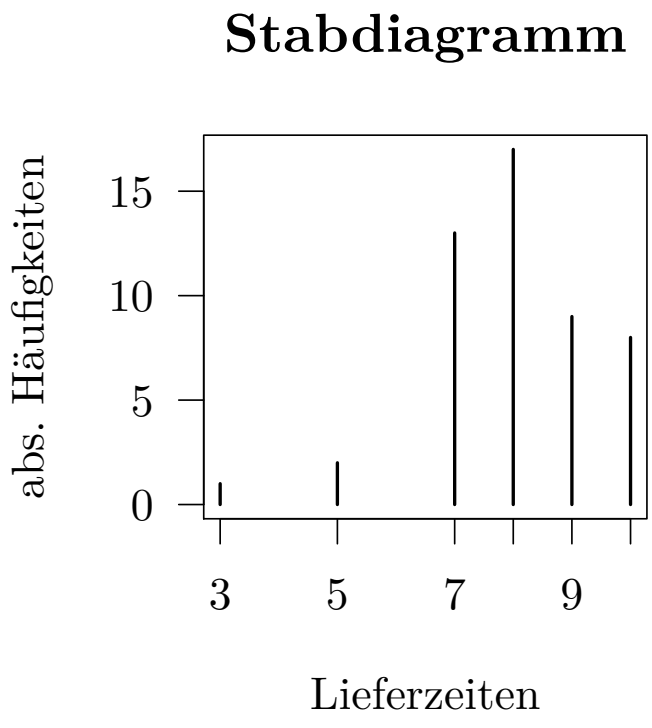
Beispiel: Lieferzeiten (vgl. F20)

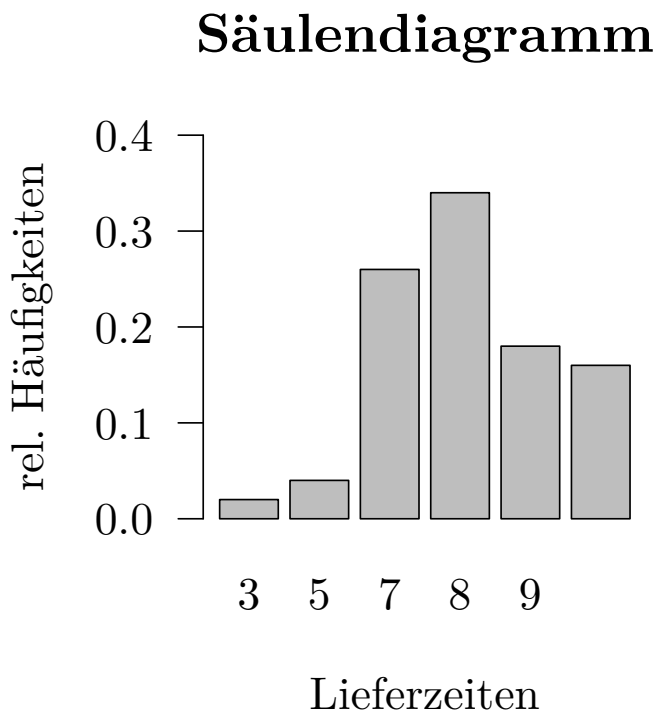
Ausprägung (sortiert) a_j		3	5	7	8	9	10	$\sum$
absolute Häufigkeit	$n(a_j) = n_j$	1	2	13	17	9	8	50
relative Häufigkeit	$h(a_j) = n(a_j)/n$	$\frac{1}{50}$	$\frac{2}{50}$	$\frac{13}{50}$	$\frac{17}{50}$	$\frac{9}{50}$	$\frac{8}{50}$	1

```
lieferzeiten <- c(7, 8, 7, 3, 8, 7, 5, 7, 8, 9, 9, 8, 8, 7,
    10, 7, 9, 8, 9, 7, 8, 7, 10, 8, 8, 9, 10, 7, 10, 9, 9, 10,
    7, 8, 7, 10, 10, 8, 8, 8, 8, 9, 9, 7, 8, 5, 8, 7, 10, 8)
table(lieferzeiten)  # Absolute Häufigkeiten
## lieferzeiten
##  3  5  7  8  9 10
##  1  2 13 17  9  8
table(lieferzeiten)/length(lieferzeiten)  # Relative Häufigkeiten
## lieferzeiten
##    3    5    7    8    9   10
## 0.02 0.04 0.26 0.34 0.18 0.16
prop.table(table(lieferzeiten))  # Relative Häufigkeiten II
## lieferzeiten
##    3    5    7    8    9   10
## 0.02 0.04 0.26 0.34 0.18 0.16
```

Beispiel: Lieferzeiten (vgl. F20)

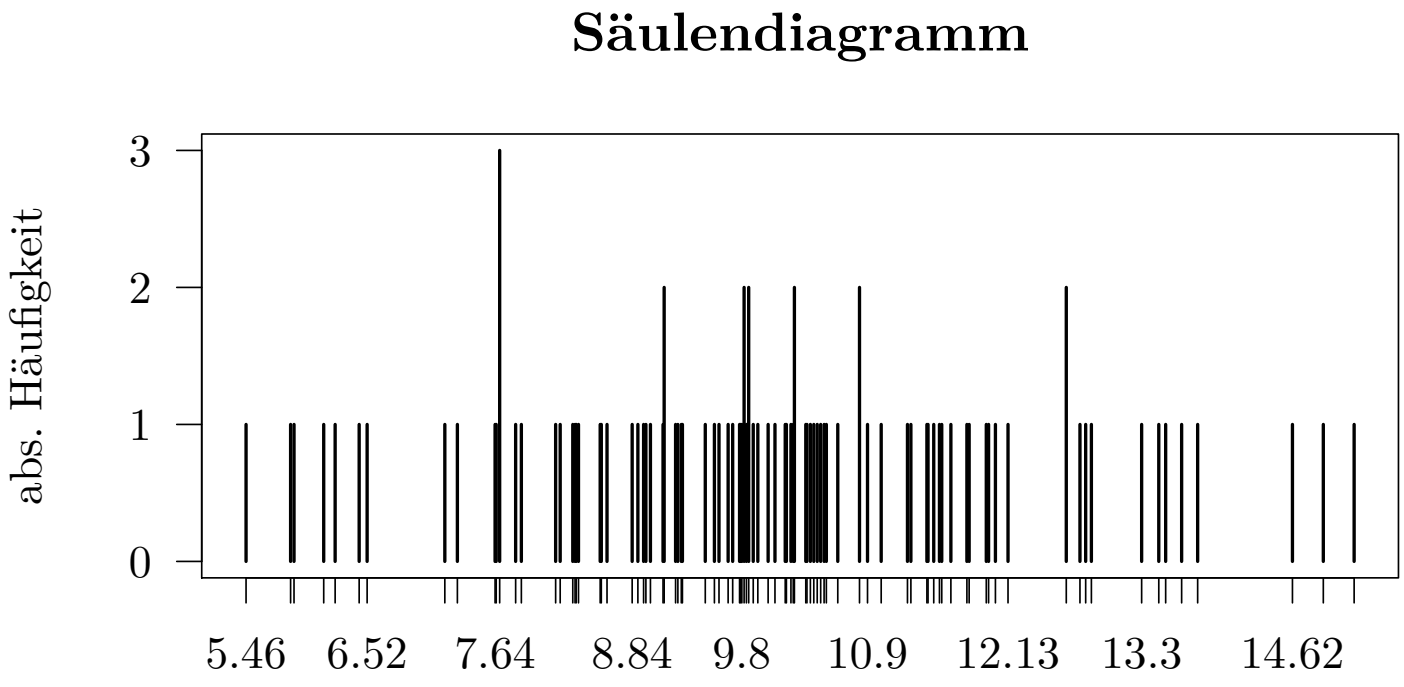
```
# Stabdiagramm für die absoluten Häufigkeiten
plot(table(lieferzeiten), type = "h", las = 1, xlab = "Lieferzeiten",
    ylab = "abs. Häufigkeiten", main = "Stabdiagramm")
# Säulendiagramm für die relativen Häufigkeiten
barplot(table(lieferzeiten)/length(lieferzeiten), xlab = "Lieferzeiten",
    ylab = "rel. Häufigkeiten", las = 1, main = "Säulendiagramm",
    ylim = c(0, 0.4))
```





Beachte: Darstellung mittels Säulen– bzw. Stabdiagramm geeignet für diskrete Merkmale, die nicht allzu viele verschiedene Ausprägungen besitzen. Bei stetigen Merkmalen ist nicht die Höhe der Stäbe von Interesse, da i. a. jede Ausprägung nur einmal auftritt, sondern die „Dichte“der Stäbe. $\rightsquigarrow$ Histogramm später (F28 ff.)

Beispiel für 100 Zufallszahlen



Kreissektorendiagramm

Winkel:

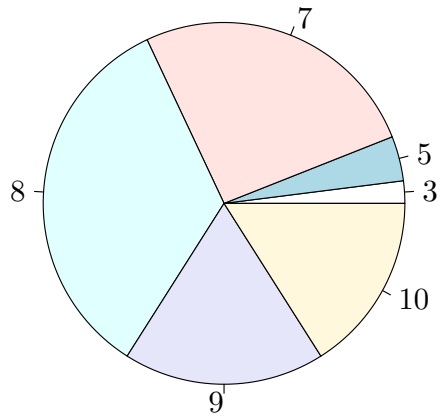
$$w_j = 360° \cdot h(a_j)$$

(Fläche proportional zu Häufigkeit)

Beispiel: Lieferzeiten (vgl. F20)

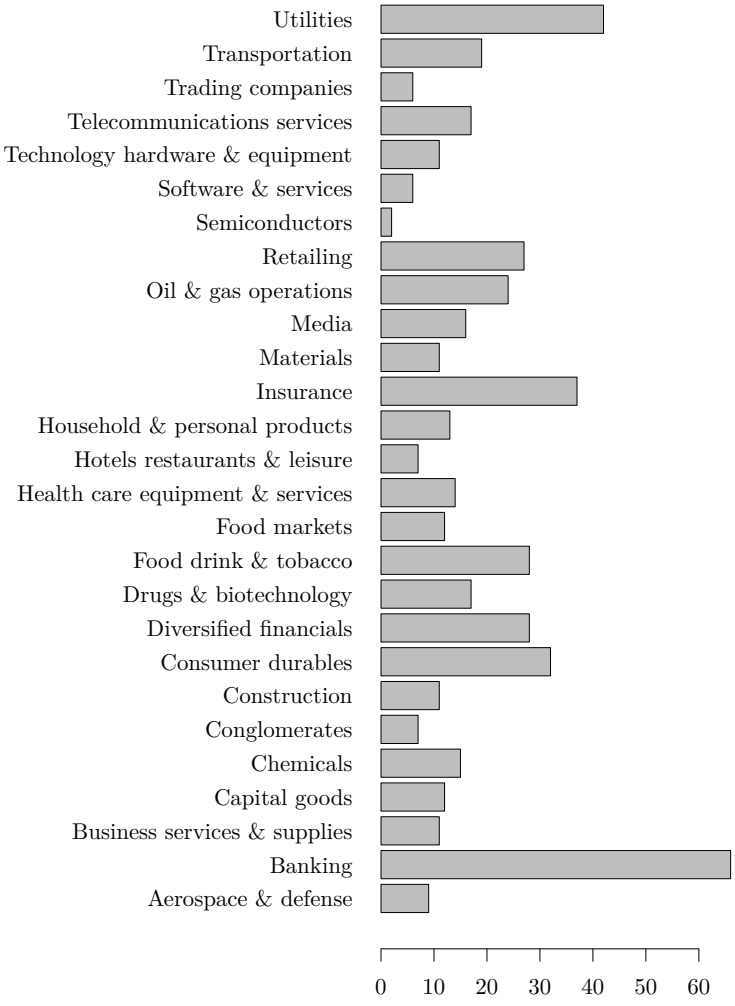
$$w_1 = 360° \cdot \tfrac{1}{50} = \quad 7.2° \quad \text{und} \quad w_6 = 360° \cdot \tfrac{10}{50} = 72°$$

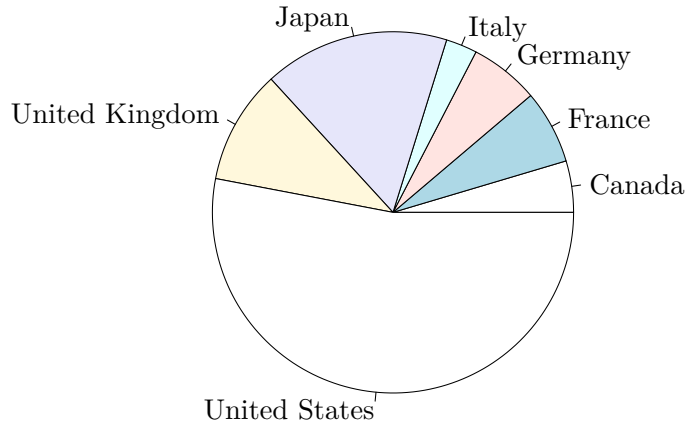
```
pie(table(lieferzeiten))
```



Beispiel: Größte Unternehmen (vgl. F19)

```
barplot(table(ForbesG7$category), horiz = TRUE, las = 1)
pie(table(ForbesG7$country))
```





1.2.2. Histogramm

Problem: Die Häufigkeitsdarstellung ist unübersichtlich, falls viele Ausprägungen vorliegen $\rightsquigarrow$ **Klassenbildung**

Beispiel:

a) Zeit (in Minuten) die benötigt wird um eine Maschine nach einem Ausfall wieder in Betrieb zu nehmen (stetiges Merkmal).
$K_1 = (0, 15]$, $K_2 = (15, 30]$,
$K_3 = (30, 45]$, $K_4 = (45, 60]$

b) Klicks auf ein Werbebanner pro Stunde (diskretes Merkmal) Messung zur Bestimmung der Wirksamkeit der Werbung.
$K_1 = [170, .., 200]$, $K_2 = [120, .., 170)$,
$K_3 = [75, .., 120)$, $K_4 = [50, .., 120)$,
$K_5 = [0, 50)$

Anforderungen an Klassenbildung:

- X sei reellwertig mit Werten in $S \subset \mathbb{R}$
- $K_i \cap K_j = \emptyset$ für alle i und j
- $S = \bigcup_{i=1}^{r} K_i$, Klassen seien Intervalle
- **wünschenswert:** Klassen gleich breit

absolute Klassenhäufigkeit von K_i:

$$n(K_i) = \text{Anzahl der Beobachtungen in } K_i$$

relative Klassenhäufigkeit von K_i:

$$h(K_i) = n(K_i)/n$$

Beispiel (Reparaturzeit):
5, 15, 28, 45, 20, 10, 15, 50, 55, 7, 12, 18, 35, 40, 17, 10, 38, 8, 25, 20

i	1 (kleine *)	2 (mittlere *)	3 (große *)	4 (erhebl. Störung)
K_i	$(0, 15]$	$(15, 30]$	$(30, 45]$	$(45, 60]$
$n(K_i)$	8	6	4	2
$h(K_i)$	0.4	0.3	0.2	0.1

```
dauer <- c(5, 15, 28, 45, 20, 10, 15, 50, 55, 7, 12, 18, 35,
    40, 17, 10, 38, 8, 25, 20)
table(cut(dauer, breaks = c(0, 15, 30, 45, 60)))
##
##  (0,15] (15,30] (30,45] (45,60]
##       8       6       4       2
table(cut(dauer, breaks = c(0, 15, 30, 45, 60)))/length(dauer)
##
##  (0,15] (15,30] (30,45] (45,60]
##     0.4     0.3     0.2     0.1
```

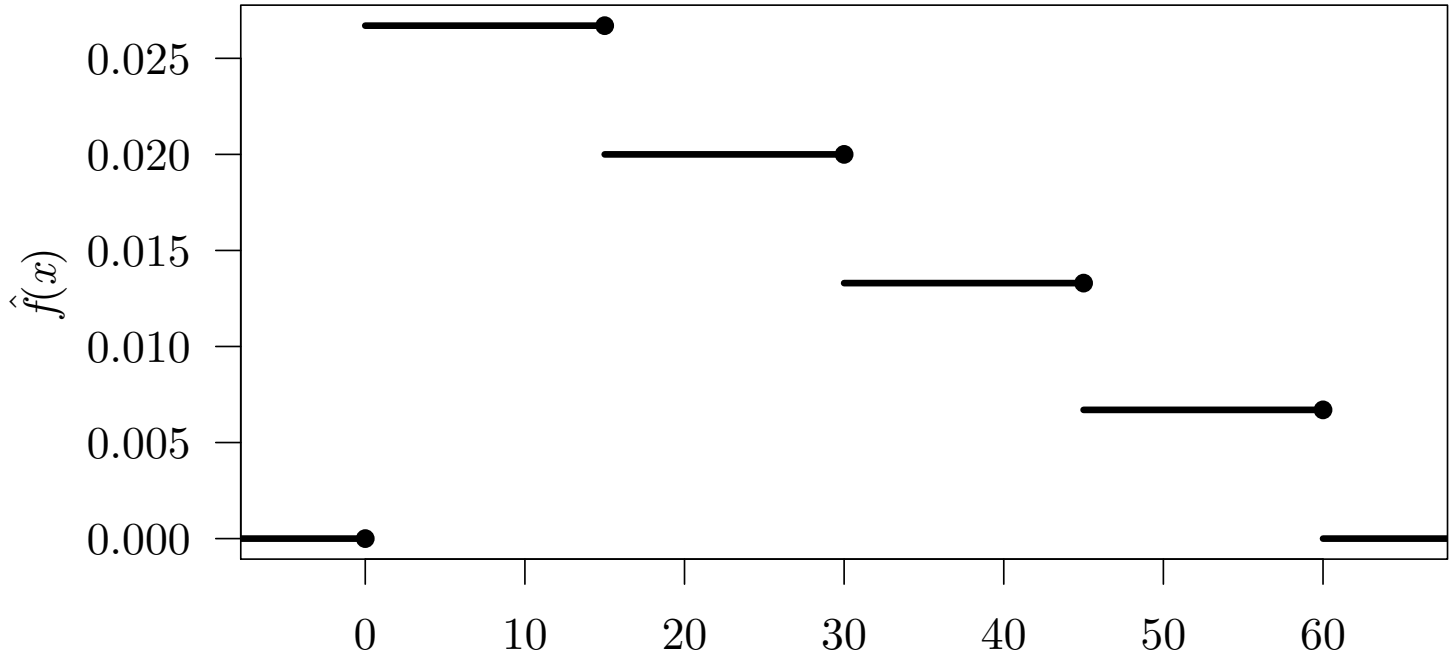
Histogramm:

$$\hat{f}(x) = \frac{\text{relative Klassenhäufigkeit}}{\text{Klassenbreite}}$$

$$= \frac{h(K_i)}{|K_i|} \quad \text{für} \quad x \in K_i$$

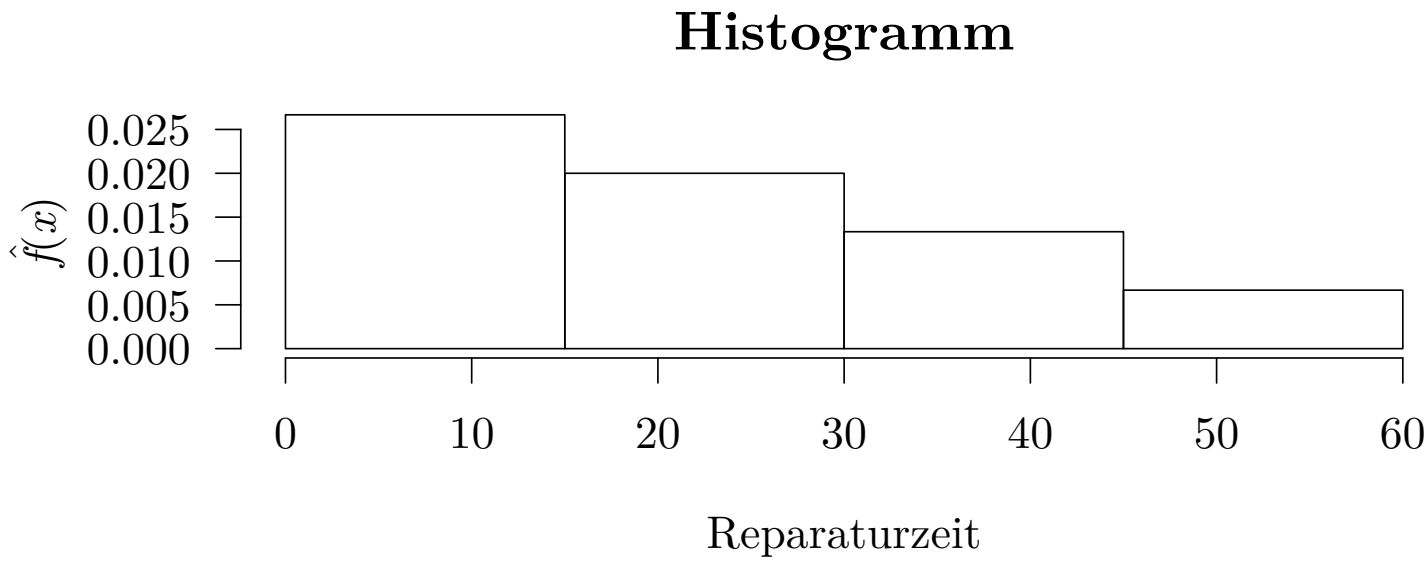
Beispiel: Reparaturzeit (vgl. F30)

Histogramm: $$\hat{f}(x) = \begin{cases} 0.0267 & \text{für} \quad x \in K_1 \\ 0.0200 & \text{für} \quad x \in K_2 \\ 0.0133 & \text{für} \quad x \in K_3 \\ 0.0067 & \text{für} \quad x \in K_4 \end{cases}$$



Beispiel: Reparaturzeit (vgl. F30)

```
hist(dauer, breaks = c(0, 15, 30, 45, 60), freq = FALSE, las = 1,
    xlab = "Reparaturzeit", ylab = expression(paste(hat(f),
        "(x)")), main = "Histogramm")
```




```
hist.info <- hist(dauer, breaks = c(0, 15, 30, 45, 60), plot = FALSE)
hist.info$counts
## [1] 8 6 4 2
hist.info$density
## [1] 0.026666667 0.020000000 0.013333333 0.006666667
```

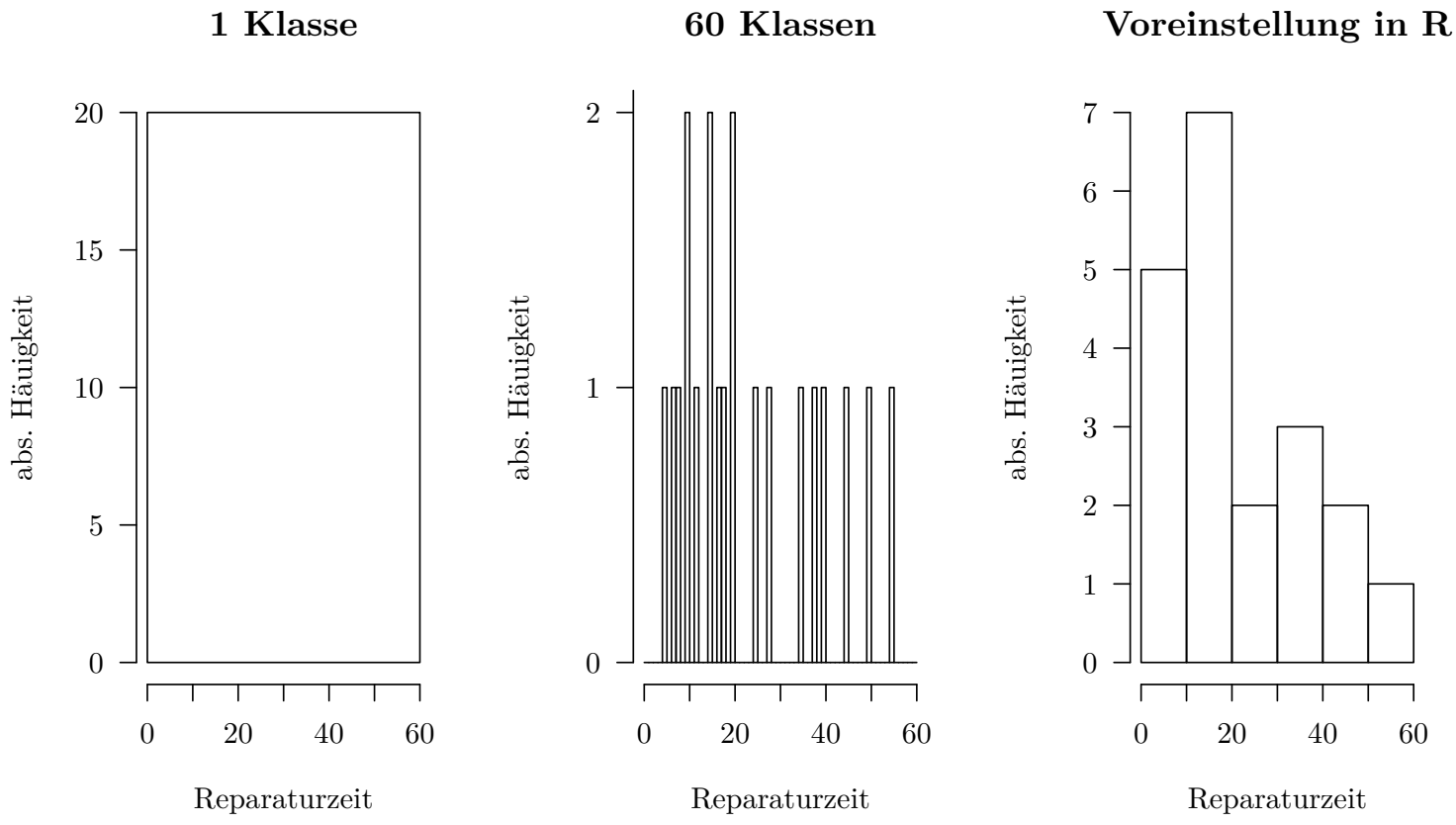
Beachte:

- Fläche zur Klasse K_i: $h(K_i)/|K_i| \cdot |K_i| = h(K_i)$,
  d. h. die entscheidende Information über das Histogramm ist die Fläche des Rechtecks!
- $$\int_{-\infty}^{\infty} \hat{f}(x)\, dx = \sum_{i=1}^{k} h(K_i) = 1$$

  $\rightsquigarrow$ Die Fläche unter dem Histogramm ist gleich 1.

Problem: Wahl der Klassenanzahl

Beispiel: Reparaturzeit (vgl. F30)



Problem: Wahl der Klassenanzahl

- Sind die Daten gleichmäßig über ein Intervall verteilt, so $k = \lfloor \sqrt{n} \rfloor$.
- Kompromissvorschlag: $k = \lfloor 10 \log_{10}(n) \rfloor$

  $\lfloor z \rfloor$ bezeichnet die größte ganze Zahl, die kleiner oder gleich z ist, z.B. $\lfloor 2.1 \rfloor = 2$, $\lfloor 4.9 \rfloor = 4$.

n	10	20	30	50	100	200	500
$\lfloor 10 \log_{10} n \rfloor$	10	13	14	16	20	23	26
$\lfloor \sqrt{n} \rfloor$	3	4	5	7	10	14	22
$2 \lfloor \sqrt{n} \rfloor$	6	8	10	14	20	28	44

1.2.3. Empirische Verteilungsfunktion

Voraussetzung: kardinales Messniveau ($a_i \in \mathbb{R}$, $a_1 < ... < a_k$)

empirische Verteilungsfunktion:

$\hat{F}(x) =$ relative Anzahl der Beobachtungen, die kleiner gleich x sind

$$\hat{F}(x) = \sum_{a_i \leq x} h(a_i)$$

Beispiel: Lieferzeiten (vgl. F20)

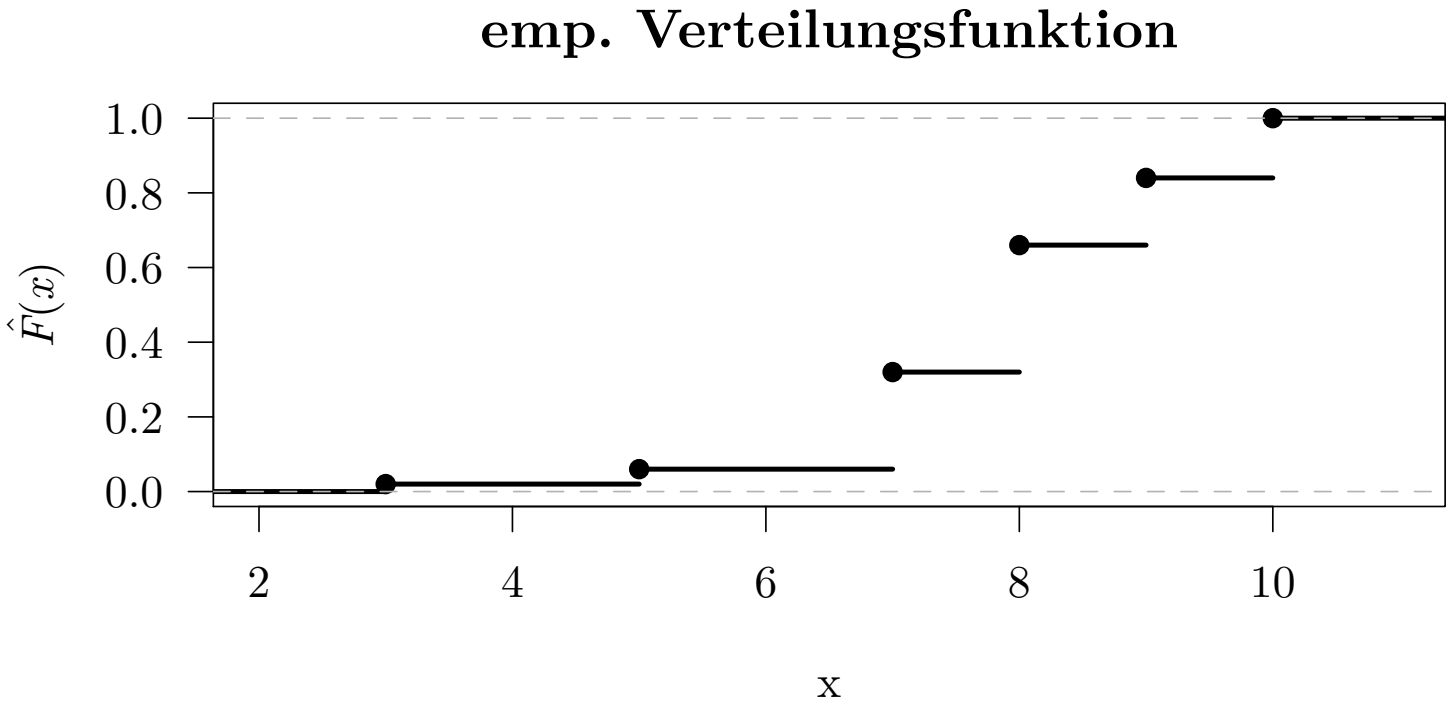
Ausprägung (sortiert)	a_j	3	5	7	8	9	10
absolute Häufigkeit	$n(a_j) = n_j$	1	2	13	17	9	8
relative Häufigkeit	$h(a_j) = n(a_j)/n$	$\frac{1}{50}$	$\frac{2}{50}$	$\frac{13}{50}$	$\frac{17}{50}$	$\frac{9}{50}$	$\frac{8}{50}$
kumulierte rel. Häufigkeit	$\sum_{i=1}^{j} h(a_i)$	$\frac{1}{50}$	$\frac{3}{50}$	$\frac{16}{50}$	$\frac{33}{50}$	$\frac{42}{50}$	1

$$\hat{F}(x) = \begin{cases} 0, & \text{für } x < 3 \\ \frac{1}{50} = 0.02, & \text{für } 3 \leq x < 5 \\ \frac{3}{50} = 0.06, & \text{für } 5 \leq x < 7 \\ \frac{16}{50} = 0.32, & \text{für } 7 \leq x < 8 \\ \frac{33}{50} = 0.66, & \text{für } 8 \leq x < 9 \\ \frac{42}{50} = 0.84, & \text{für } 9 \leq x < 10 \\ \frac{50}{50} = 1, & \text{für } 10 \leq x \end{cases}$$

```
cumsum(table(lieferzeiten)/length(lieferzeiten))
##    3    5    7    8    9   10
## 0.02 0.06 0.32 0.66 0.84 1.00
```

Beispiel: Lieferzeiten (vgl. F20)

```
plot(ecdf(lieferzeiten), lwd = 3, ylab = expression(paste(hat(F),
    "(x)")), las = 1, xlab = "x", main = "emp. Verteilungsfunktion")
```



$\hat{F}(7.5) = 0.32 \rightsquigarrow$ bei 32% der Lieferungen betrug der Lieferzeit weniger als 7.5 Tage oder genau 7.5 Tage.

Eigenschaften der empirischen Verteilungsfunktion:

a) $\hat{F}(x) = 0$ für $x < a_1$, $\hat{F}(x) = 1$ für $x \geq a_k$

b) $\hat{F}(x)$ ist monoton steigend

c) $\hat{F}(x)$ ist rechtsseitig stetig

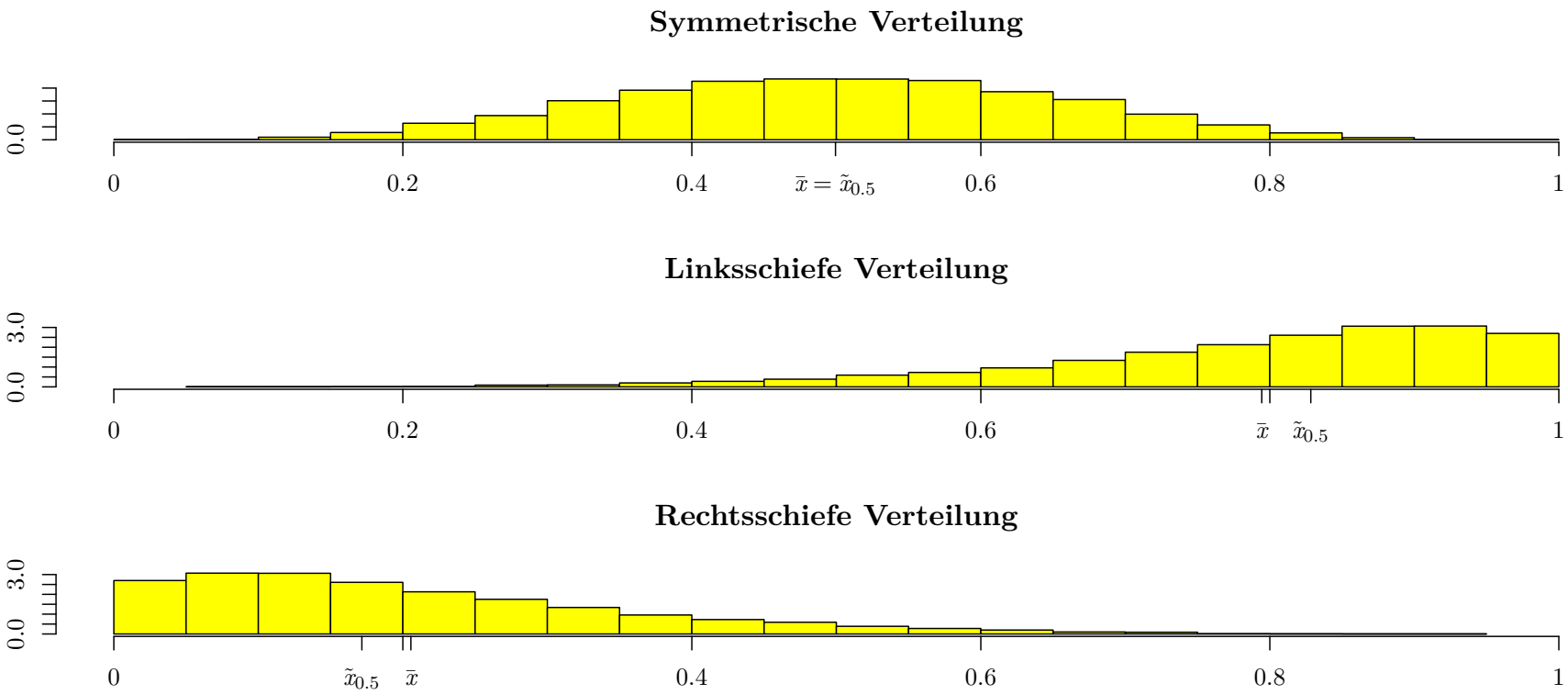
d) $\hat{F}(a_j) - \hat{F}(a_j-) = h(a_j) \rightsquigarrow$ die Sprünge nur an den Stellen $a_1, \ldots a_k$ und die Höhe des Sprunges ist gleich der relativen Häufigkeit der Ausprägung

Bezeichnung: $G(a-) = \lim_{\varepsilon>0,\, \varepsilon\to 0} G(a-\varepsilon)$

1.2.4. Lokalisationsmaße (Lagemaße)

Ziel: Aussage über das Zentrum der Stichprobe

Interpretation: Liegen die Daten in etwa symmetrisch um einen Punkt, so ist dieser das Zentrum.



(i) Modus x_{Mod}: häufigster Wert

Beispiel (Rohstoffbedarf): Bedarf an einem Rohstoff für die Produktion pro Periode
10, 23, 20, 33, 50, 20, 20, 13, 50, 33

a_j	10	13	20	23	33	50
$n(a_j)$	1	1	3	1	2	2

$\Rightarrow x_{\text{Mod}} = 20$

Sinnvoll bei allen Skalenniveaus.

```
rohbedarf <- c(10, 23, 20, 33, 50, 20, 20, 13, 50, 33)
tab.rohbedarf <- table(rohbedarf)
tab.rohbedarf[tab.rohbedarf == max(tab.rohbedarf)]
## 20
##  3
```

(ii) Mittelwert (arithmetisches Mittel, Durchschnitt)

Mittelwert

$$\begin{aligned} \bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ &= \frac{1}{n}\sum_{i=1}^{k} n(a_i)\, a_i \;=\; \sum_{i=1}^{k} h(a_i)\, a_i \end{aligned}$$

Sinnvoll nur bei kardinalem Skalenniveau.

Beispiel (Pizzapreise): Preise für eine Pizza Margherita in $n = 5$ verschiedenen Restaurants in Augsburg:

$$x_1 = 6,\ x_2 = 8,\ x_3 = 5,\ x_4 = 5,\ x_5 = 6.$$

Durchschnittspreis:

$$\bar{x} = \frac{1}{5}(2 \cdot 5 + 2 \cdot 6 + 8) = 6.$$

Beispiel: Pizzapreise (vgl. F43)

```
preise <- c(6, 8, 5, 5, 6)
mean(preise)
## [1] 6
```

```
# Mittelwertsberechnung bei einem fehlenden Wert
preiseNA <- c(6, 8, 5, 5, 6, NA)
mean(preiseNA)
## [1] NA
mean(preiseNA, na.rm = TRUE)
## [1] 6
```

Mittelwert für klassierte Daten (Näherung, Ersatzgröße)

Klassen K_1,..,K_r mit Klassenmittel $m_1, \ldots, m_r$

$$\bar{x}_K = \frac{1}{n} \sum_{i=1}^{r} n(K_i)\, m_i = \sum_{i=1}^{r} h(K_i)\, m_i$$

Beispiel: Reparaturzeit (vgl. F30)

K_i	(0,15]	(15,30]	(30,45]	(45,60]
$h(K_i)$	0.4	0.3	0.2	0.1

$$\begin{aligned} \bar{x}_K &= 0.4 \cdot 7.5 + 0.3 \cdot 22.5 + 0.2 \cdot 37.5 + 0.1 \cdot 52.5 \\ &= 22.5. \end{aligned}$$

Wäre $K_4 = (45, 90]$, so bleibt $h(K_4) = 0.1$, aber $\bar{x}_K = 24.0$ ($m_4 = 67.5$).

(iii) p–Quantil $\tilde{x}_p$ und Median

Sortiere die Stichprobe

$$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \cdots \leq x_{(n)},$$

d.h. $x_{(i)}$ bezeichnet die i-te Beobachtung in der geordneten Stichprobe.

p–Quantil

$$\tilde{x}_p = \begin{cases} x_{(\lfloor n\,p \rfloor + 1)} & \text{für } n\,p \notin \mathbb{Z} \\ \left(x_{(n\,p)} + x_{(n\,p+1)}\right)/2 & \text{für } n\,p \in \mathbb{Z} \end{cases}, \quad p \in (0, 1]$$

$\lfloor z \rfloor$ bezeichnet die größte ganze Zahl, die kleiner ist als z, z.B. $\lfloor 2.1 \rfloor = 2$, $\lfloor 4.9 \rfloor = 4$.

Interpretation: mindestens $100 \cdot p\%$ der Werte sind kleiner oder gleich $\tilde{x}_p$ und gleichzeitig sind mindestens $100 \cdot (1-p)\%$ größer oder gleich $\tilde{x}_p$.

Beachte: Sinnvoll nur ab ordinalem Skalenniveau.

$\tilde{x}_{0.25}$ heißt **unteres Quartil**, $\tilde{x}_{0.5}$ **Median** und $\tilde{x}_{0.75}$ **oberes Quartil**

Alternative Schreibweise für den Median: x_{Med}

Hiermit gilt für den Median:

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

Beispiel: Rohstoffbedarf (vgl. F42), $n = 10$

Stichprobe x_i									
10	23	20	33	50	20	20	13	50	33
sortierte Stichprobe $x_{(i)}$									
10	13	20	20	20	23	33	33	50	50

Hiermit:

$$\tilde{x}_{0.25} \underset{10\cdot 0.25 = 2.5 \notin \mathbb{Z}}{=} x_{(\lfloor 2.5 \rfloor + 1)} = x_{(3)} = 20,$$

$$\tilde{x}_{0.5} \underset{10\cdot 0.5 = 5 \in \mathbb{Z}}{=} \frac{1}{2}(x_{(5)} + x_{(5+1)}) = \frac{1}{2}(20 + 23) = 21.5,$$

$$\tilde{x}_{0.75} \underset{10\cdot 0.75 = 7.5 \notin \mathbb{Z}}{=} x_{(\lfloor 7.5 \rfloor + 1)} = x_{(8)} = 33.$$

Beispiel: Rohstoffbedarf (vgl. F42)

```
rohbedarf <- c(10, 23, 20, 33, 50, 20, 20, 13, 50, 33)
median(rohbedarf)
## [1] 21.5
quantile(rohbedarf, type = 2)
##   0%  25%  50%  75% 100%
## 10.0 20.0 21.5 33.0 50.0
quantile(rohbedarf, probs = 0.6, type = 2)
## 60%
##  28
```

Eigenschaften des Mittelwertes:

- Der Mittelwert reagiert äußerst empfindlich auf **Ausreißer**.
  Z. B. Einkommen: (1000, 1000, 1000, 1000, 10000)

```
einkommen <- c(1000, 1000, 1000, 1000, 10000)
mean(einkommen)
## [1] 2800
median(einkommen)
## [1] 1000
# Lösung: Getrimmter Mittelwert
mean(c(1000, 1000, 1000, 1000, 10000), trim = 0.2)
## [1] 1000
```

- Mittelwert nur repräsentativ bei in etwa symmetrischen Daten, ansonsten schwer interpretierbar.
- Ist einfach bei klassierten Daten einzusetzen.

Eigenschaften des Medians:

- Der Median ist eine äußerst robuste Größe.
- Der Median kann auch als Kenngröße bei nichtsymmetrischen Daten verwendet werden.

Lineare Transformation der Lagemaße

Transformiert man die Beobachtungen gemäß

$$y_i = a + b \cdot x_i$$

so gilt für die Lageparameter

$$y_{\text{Mod}} = a + b \cdot x_{\text{Mod}}$$
$$y_{\text{Med}} = a + b \cdot x_{\text{Med}}$$
$$\tilde{y}_p = a + b \cdot \tilde{x}_p$$
$$\bar{y} = a + b \cdot \bar{x}$$

Beispiel: Rohstoffbedarf (vgl. F42), wobei die Produktion verdreifacht und 20 zusätzliche Reserveeinheiten eingeplant werden: $y_i = 20 + 3 \cdot x_i$

$x_{(i)}$	10	13	20	20	20	23	33	33	50	50
$y_{(i)}$	50	59	80	80	80	89	119	119	170	170

$$y_{\text{Mod}} = 80 \quad = 20 + 3 \cdot 20 \quad = 20 + 3 \cdot x_{\text{Mod}}$$
$$y_{\text{Med}} = 84.5 \quad = 20 + 3 \cdot 21.5 \quad = 20 + 3 \cdot x_{\text{Med}}$$
$$\bar{y} = 101.6 \quad = 20 + 3 \cdot 27.2 \quad = 20 + 3 \cdot \bar{x}$$

Optimalitätseigenschaften der Lagemaße

- Mittelwert:

$$\sum_{i=1}^{n}(x_i - \lambda)^2$$

  wird minimiert durch $\lambda = \bar{x}$.

- Median:

$$\sum_{i=1}^{n}|x_i - \lambda|$$

  wird minimiert durch $\lambda = x_{\text{Med}}$.

- Modus:

$$\sum_{i=1}^{n} s(x_i, \lambda) \qquad \text{mit} \qquad s(x_i, \lambda) = \begin{cases} 0, & \text{falls } x_i = \lambda \\ 1, & \text{falls } x_i \neq \lambda \end{cases}$$

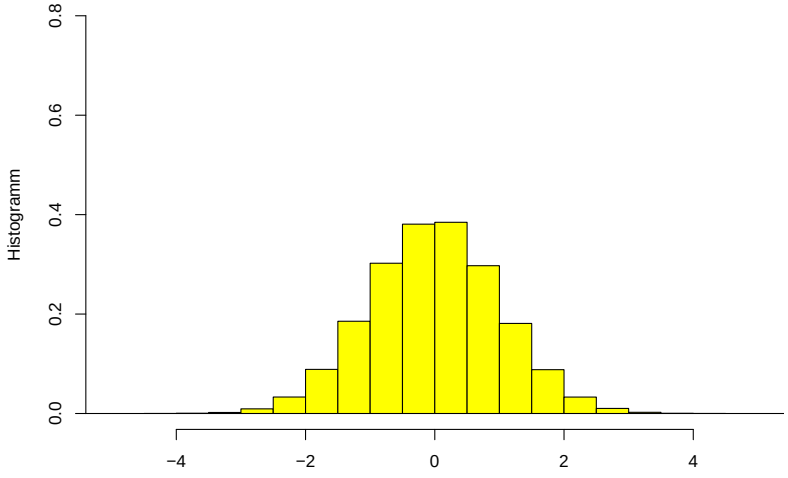
  wird minimiert durch $\lambda = x_{\text{Mod}}$.

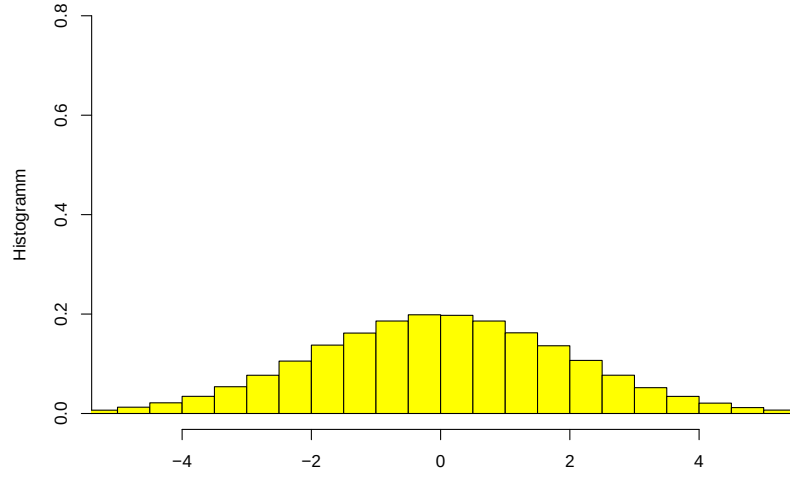
Beispiel: Größte Unternehmen (vgl. F19)

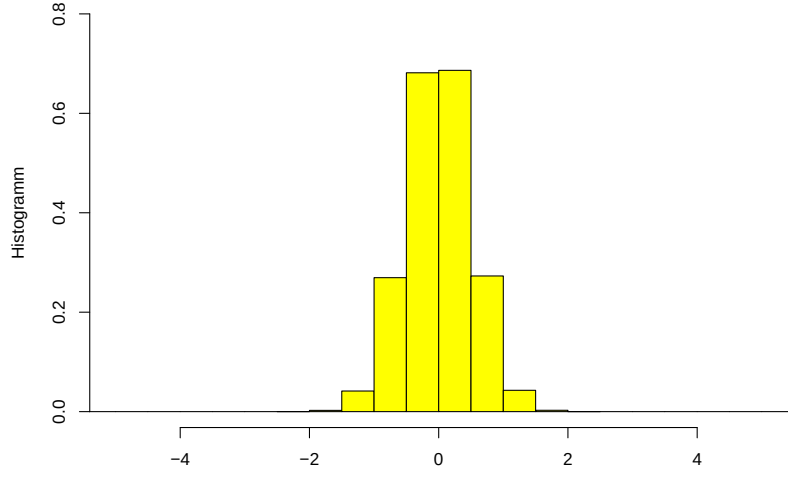
```
# Modus der nominalen Merkmale
tab.Country <- table(ForbesG7$country)
tab.Country[tab.Country == max(tab.Country)]
## United States
##           265
tab.Category <- table(ForbesG7$category)
tab.Category[tab.Category == max(tab.Category)]
## Banking
##      66
# Mittelwert der stetigen Merkmale
stetigeVar <- c("sales", "profits", "assets", "marketvalue")
apply(ForbesG7[, stetigeVar], 2, mean)
##       sales     profits      assets marketvalue
##    23.60454     1.08628    85.84856    28.80510
# Median der stetigen Merkmale
apply(ForbesG7[, stetigeVar], 2, median)
##       sales     profits      assets marketvalue
##      14.190       0.650      26.025      14.560
```

1.2.5. Streuungsmaße

Motivation für Streuungsmaße







Problem: Lokalisationsmaße beschreiben die Daten nur unzureichend

Ziel: Aussage über die Streuung der Daten um das Zentrum (Lagemaß)

Voraussetzung: kardinale Beobachtungen $x_1, \ldots, x_n$

(i) empirische Varianz (mittlere quadratische Abweichung)

$$
\begin{aligned}
s^2 &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
&= \frac{1}{n} \sum_{i=1}^{k} n(a_i) \, (a_i - \bar{x})^2 = \sum_{i=1}^{k} h(a_i) \, (a_i - \bar{x})^2
\end{aligned}
$$

s^2 gibt den mittleren quadratischen Abstand der Beobachtungen vom Mittelwert an.
$s = \sqrt{s^2}$ heißt **empirische Standardabweichung**.

Bemerkung: Häufig verwendet man anstelle des Vorfaktors $\frac{1}{n}$ auch den Vorfaktor $\frac{1}{n-1}$ (Statistik II). Dieser wird auch in der Funktion `var()` in R verwendet.

Beispiel: Pizzapreise (vgl. F43)

```
n <- length(preise)
var(preise)
## [1] 1.5
var(preise) * (n - 1)/n
## [1] 1.2
```

Eigenschaften von s^2

- **Verschiebungssatz:**

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

Beispiel: Pizzapreise (vgl. F43) $x_1 = (6, 8, 5, 5, 6)$ mit $\bar{x}_1 = 6$

$$s^2 = \frac{2 \cdot 5^2 + 2 \cdot 6^2 + 8^2}{5} - 6^2 = \frac{186}{5} - 36 = 37.2 - 36 = 1.2\,,$$

$$s \approx 1.095\,.$$

```
preise1 <- c(6, 8, 5, 5, 6)
n <- length(preise1)
var(preise1)
## [1] 1.5
(var.preise1 <- var(preise1) * (n - 1)/n)
## [1] 1.2
round(sqrt(var.preise1), 3)
## [1] 1.095
```

- Die empirische Varianz reagiert äußerst empfindlich auf Ausreißer.

Beispiel: Pizzapreise (vgl. F43) mit einem Ausreißer
$x_2 = (6, \mathbf{18}, 5, 5, 6) \quad \text{mit} \quad \bar{x}_2 = 8$

$$s^2 = \frac{2 \cdot 5^2 + 2 \cdot 6^2 + 18^2}{5} - 8^2 = \frac{446}{5} - 64 = 89.2 - 64 = 25.2\,,$$

$$s \approx 5.02\,.$$

```
preise2 <- c(6, 18, 5, 5, 6)
n <- length(preise2)
var(preise2)
## [1] 31.5
var(preise2) * (n - 1)/n
## [1] 25.2
sqrt((n - 1)/n) * sd(preise2)
## [1] 5.01996
```

(ii) **Spannweite**

$$SP = x_{(n)} - x_{(1)} = \max_i x_i - \min_i x_i$$

Bespiel: Pizzapreise (vgl. F43)

$$(a) \qquad SP = 8 - 5 = 3$$
$$(b) \qquad SP = 13 - 5 = 8$$

```
max(preise1) - min(preise1)
## [1] 3
max(preise2) - min(preise2)
## [1] 13
```

Beachte: Die Spannweite ist nicht robust.

(iii) Quartilabstand

$$QA = \tilde{x}_{0.75} - \tilde{x}_{0.25}$$

Bespiel: Pizzapreise (vgl. F43)

$$(a) \qquad QA = 6 - 5 = 1$$

$$(b) \qquad QA = 6 - 5 = 1$$

```
IQR(preise1, type = 2)
## [1] 1
```

Bemerkung:

- Der Quartilabstand ist resistent gegenüber Ausreißern.
- Es liegen mindestens $[n/2]$ aller Beobachtungen im Intervall $[\tilde{x}_{0.25}, \tilde{x}_{0.75}]$,

$$x_{([n/4])} \leq \tilde{x}_{0.25} \leq x_{([n/4]+1)} \leq \cdots \leq x_{([3n/4])} \leq \tilde{x}_{0.75} \leq x_{([3n/4]+1)}.$$

(iv) Median der absoluten Abweichungen vom Median (MAD)

MAD ist gleich dem

$$\text{Median von } |x_i - \tilde{x}_{0.5}|, \; i = 1, \ldots, n$$

Beispiel: Pizzapreise (vgl. F43) $x_1 = (6, 8, 5, 5, 6)$ mit $x_{med} = 6$

$\lvert x_i - x_{Med} \rvert$	0	2	1	1	0
sortierte Werte	0	0	1	1	2

Somit ist $MAD = 1$.

```
mad(preise1, constant = 1)
## [1] 1
med <- median(preise1)
median(abs(preise1 - med))
## [1] 1
```

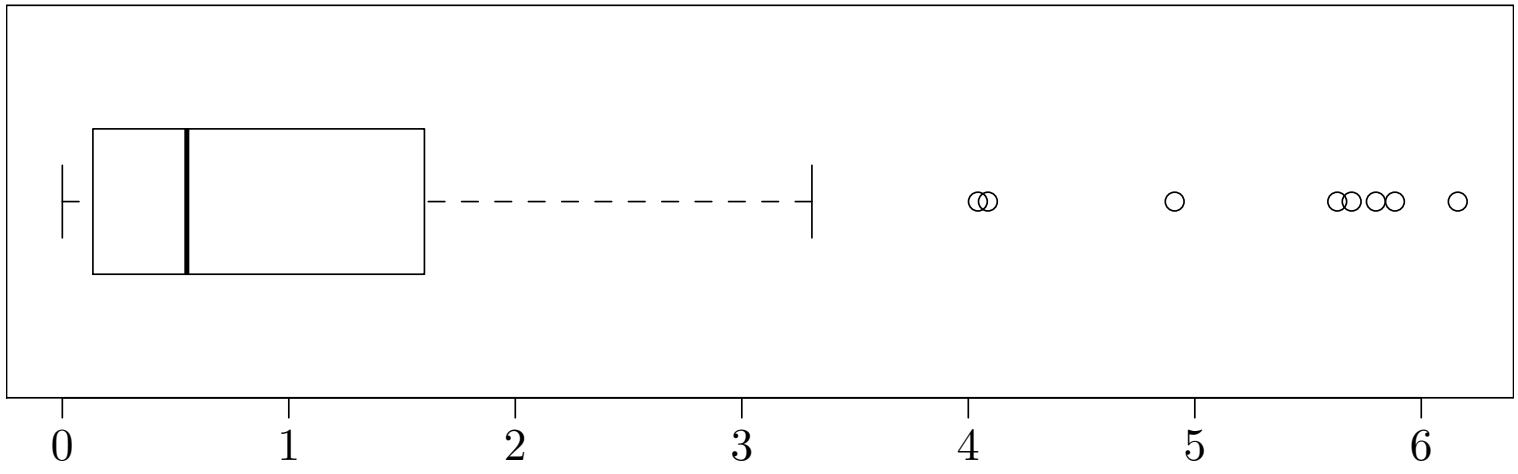
Beachte: MAD ist robust gegenüber Ausreißern.

Lineare Transformation der Streuungsmaße: $y_i = a + b \cdot x_i$

$$\begin{aligned} \mathrm{SP}_y &= |b| \cdot \mathrm{SP}_x \\ s_y^2 &= b^2 \cdot s_x^2 \\ s_y &= |b| \cdot s_x \\ MAD_y &= |b| \cdot MAD_x \end{aligned}$$

Das Box–Plot

Das **Box–Plot** ist eine Graphik, in der die wichtigsten Lagemaße und Streuungsmaße eingetragen werden.



- Linie in der Mitte: Median
- Unterer und oberer Rand der Box: Unteres und oberes Quartil
- Untere und obere Linie (=Whiskers): Kleinster und größter Wert, aber beschränkt durch das 1.5–fache des Quartilsabstands
- Punkte: Ausreißer, die außerhalb des 1.5–fachen des Quartilsabstandes liegen

Beispiel (Unternehmensgewinn): $n = 6$, Gewinn eines UN über mehrere Perioden (in Mio. €)

$$2.1,\ 2.5,\ 2.7,\ 3.0,\ 3.4,\ 6.7$$

$$\rightsquigarrow \quad x_{(1)} = 2.1\,,$$

$$\tilde{x}_{0.25} = x_{(\lfloor 0.25 \cdot 6 \rfloor + 1)} = x_{(2)} = 2.5\,,$$

$$\tilde{x}_{0.5} = \left(x_{(3)} + x_{(4)}\right)/2 = 2.85\,,$$

$$\tilde{x}_{0.75} = x_{(\lfloor 0.75 \cdot 6 \rfloor + 1)} = x_{(5)} = 3.4\,,$$

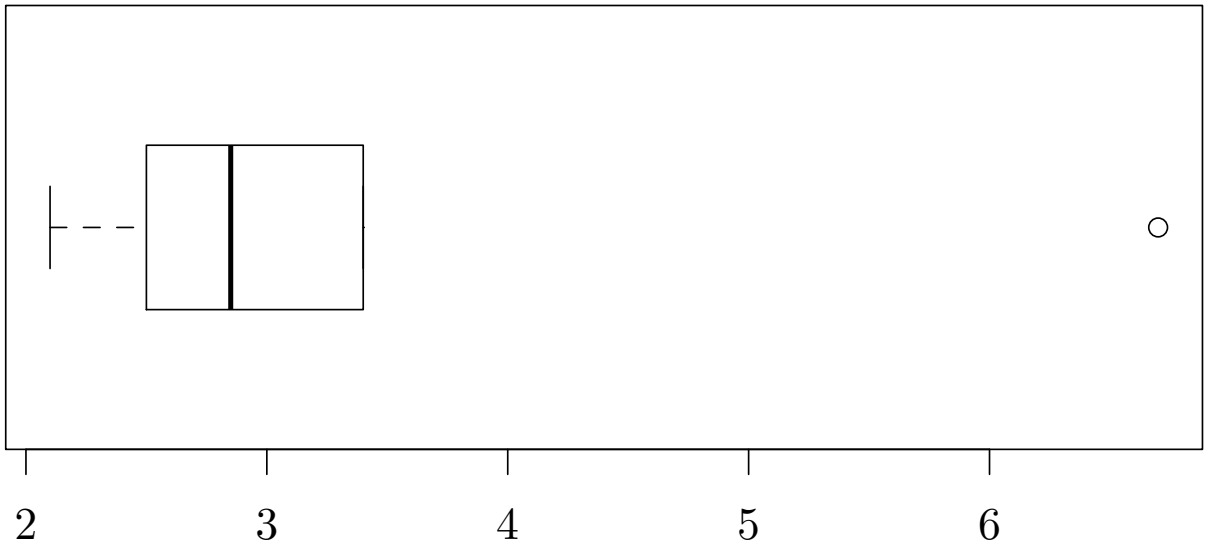
$$x_{(6)} = 6.7\,,$$

ABER: Ausreißer, da $x_{(6)}$ außerhalb des 1.5–fachen Quartilsabstandes liegt:

$$6.7 > \tilde{x}_{0.75} + 1.5(\tilde{x}_{0.75} - \tilde{x}_{0.25}) = 4.75$$

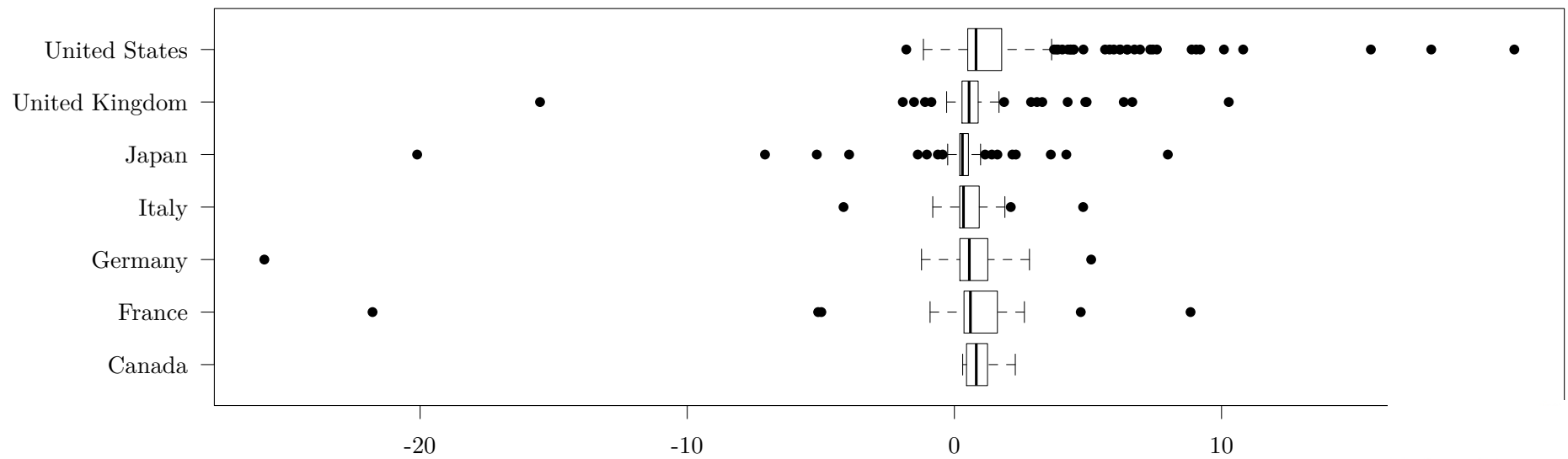
Beispiel: Unternehmensgewinn (vgl. F64)

```
gewinn <- c(2.1, 2.5, 2.7, 3, 3.4, 6.7)
boxplot(gewinn, horizontal = TRUE)
```



Beispiel: Größte Unternehmen (vgl. F19)

```
apply(ForbesG7[, stetigeVar], 2, sd)
##       sales     profits      assets marketvalue
##   29.463847    3.043527  169.299051   41.041540
apply(ForbesG7[, stetigeVar], 2, function(x) max(x) - min(x))
##       sales     profits      assets marketvalue
##      254.86       46.79     1260.67      327.60
apply(ForbesG7[, stetigeVar], 2, IQR)
##       sales     profits      assets marketvalue
##     19.1650      1.0225     51.0750     21.0300
boxplot(profits ~ country, data = ForbesG7, horizontal = TRUE,
    las = 1, pch = 16)
```



1.2.6. Konzentrationsmaße

Voraussetzungen: kardinale Werte $x_i \geq 0$, $i = 1, \cdots, n$

Beispiel: 5 Unternehmen, 25 Mio Kunden. Hat jedes Unternehmen 5 Mio Kunden, so keine Konzentration. Betreut dagegen ein Unternehmen 20 Mio Kunden, so starke Konzentration.

Problem: Wie stark wirkt sich jede Beobachtung auf die Summe aller Ausprägungen aus?

Ziel: Wie viel Prozent der Merkmalssummen entfällt auf die u Prozent kleinsten Merkmalsträger?

Achtung: Die Werte **müssen** aufsteigend sortiert werden: $x_i \mapsto x_{(i)}$!

Lorenzkurve:

Streckenzug: $(0,0),\ (u_1,v_1),\ \ldots,\ (u_n,v_n) = (1,1)$ mit

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

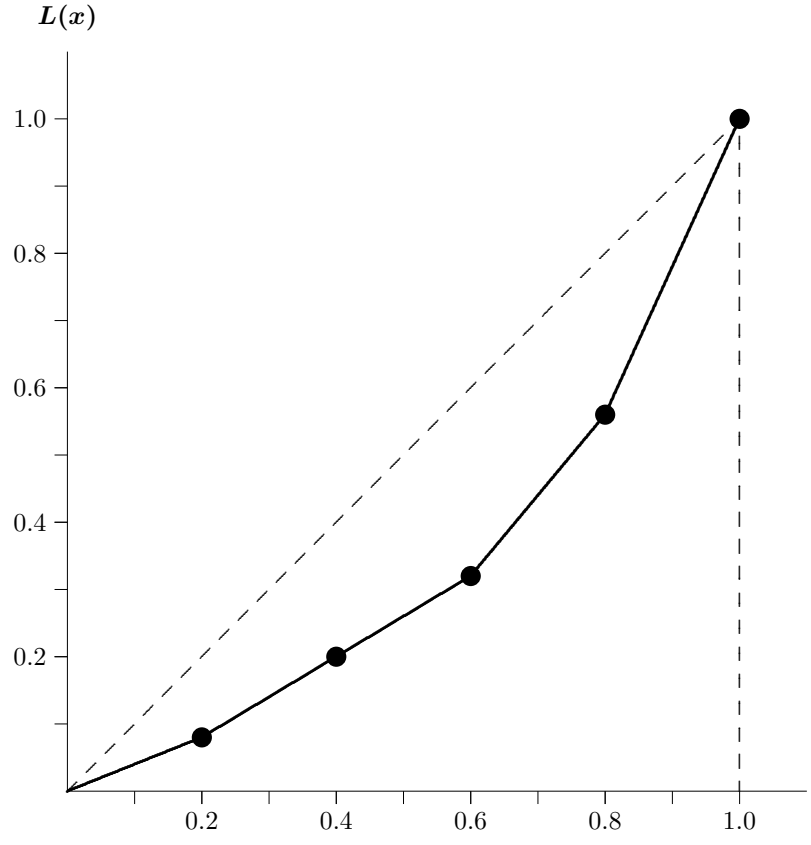
$$v_i = \text{Anteil der } \ i \ \text{ kleinsten Beob. an der Gesamtsumme} = \frac{\sum_{j=1}^{i} x_{(j)}}{\sum_{j=1}^{n} x_{(j)}}$$

Beispiel I:

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

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

i	1	2	3	4	5
$x_{(i)}$	2	3	3	6	11
u_i	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{3}{5}$	$\frac{4}{5}$	1
$v_i = \frac{\sum_{j=1}^{i} x_{(j)}}{\sum_{j=1}^{5} x_{(j)}}$	$\frac{2}{25}$	$\frac{5}{25}$	$\frac{8}{25}$	$\frac{14}{25}$	1

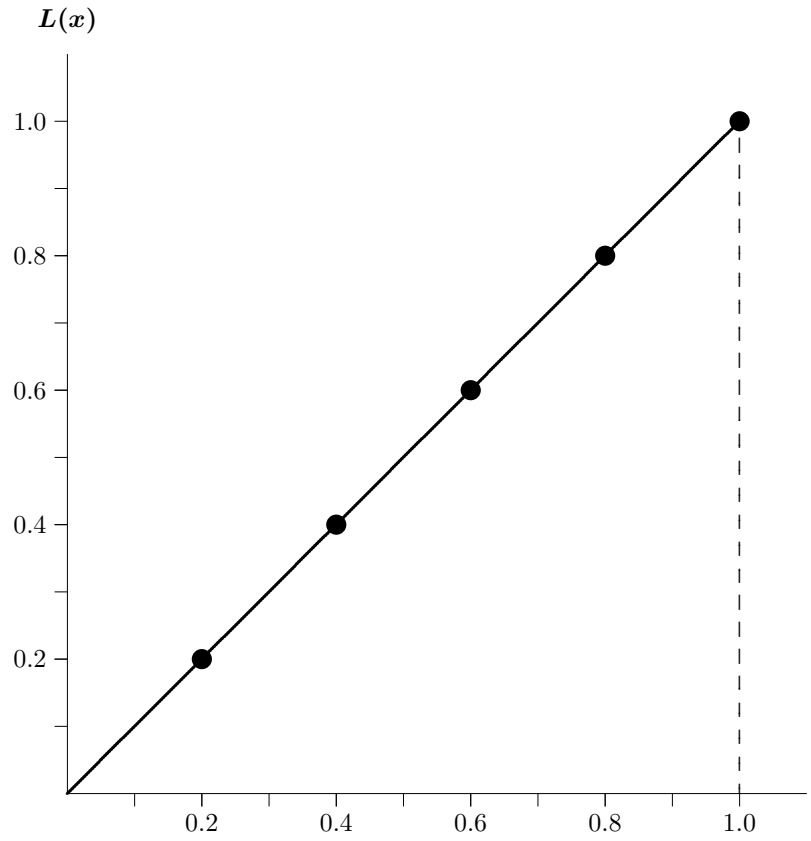


Beispiel II:

Markt mit fünf Unternehmen; Umsätze: 5, 5, 5, 5, 5 (Mio. €)

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

i	1	2	3	4	5
$x_{(i)}$	5	5	5	5	5
u_i	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{3}{5}$	$\frac{4}{5}$	1
$v_i = \frac{\sum_{j=1}^{i} x_{(j)}}{\sum_{j=1}^{5} x_{(j)}}$	$\frac{5}{25}$	$\frac{10}{25}$	$\frac{15}{25}$	$\frac{20}{25}$	$\frac{25}{25} = 1$



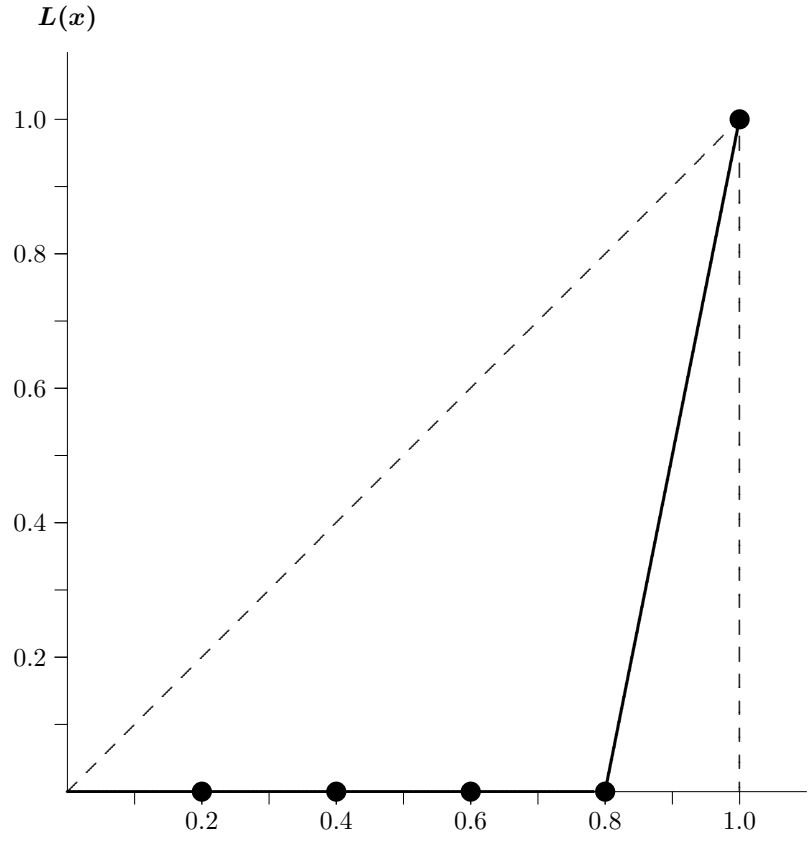
$\Rightarrow$ Gleichverteilung

Beispiel III:

Markt mit fünf Unternehmen; Umsätze: 0, 0, 0, 0, 25 (Mio. €)

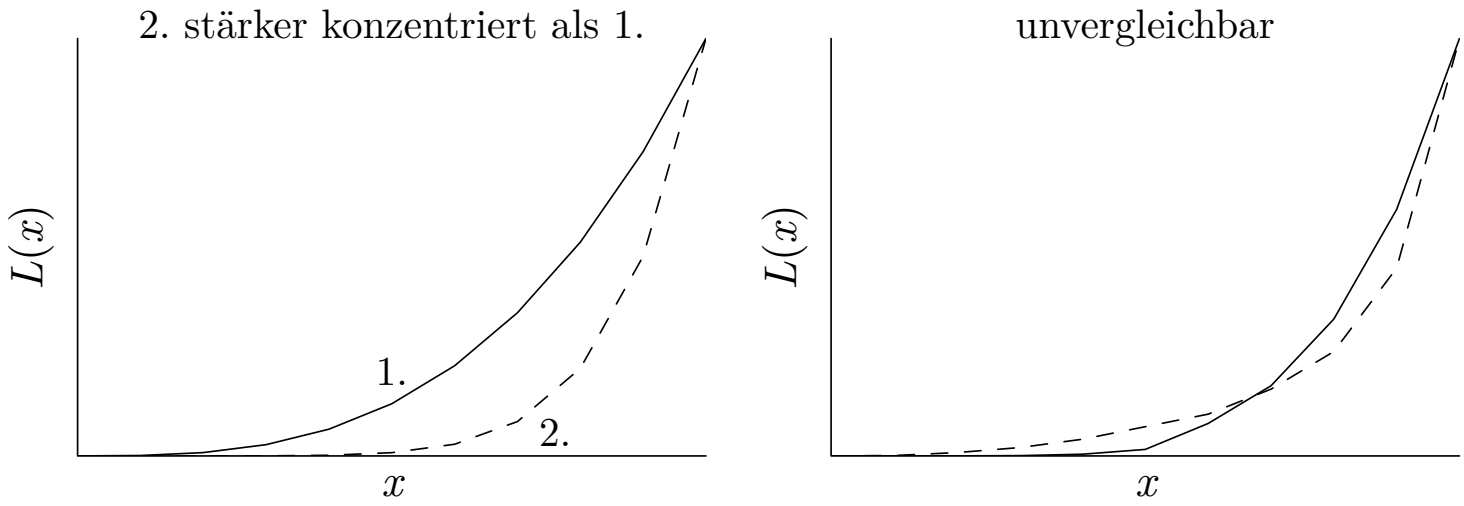
$\Rightarrow n = 5, \ \sum_{i=1}^{5} x_i = 25$

i	1	2	3	4	5
$x_{(i)}$	0	0	0	0	25
u_i	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{3}{5}$	$\frac{4}{5}$	1
$v_i = \frac{\sum_{j=1}^{i} x_{(j)}}{\sum_{j=1}^{n} x_{(j)}}$	0	0	0	0	$\frac{25}{25} = 1$



$\Rightarrow$ extreme Konzentration

Vergleich von Lorenzkurven:

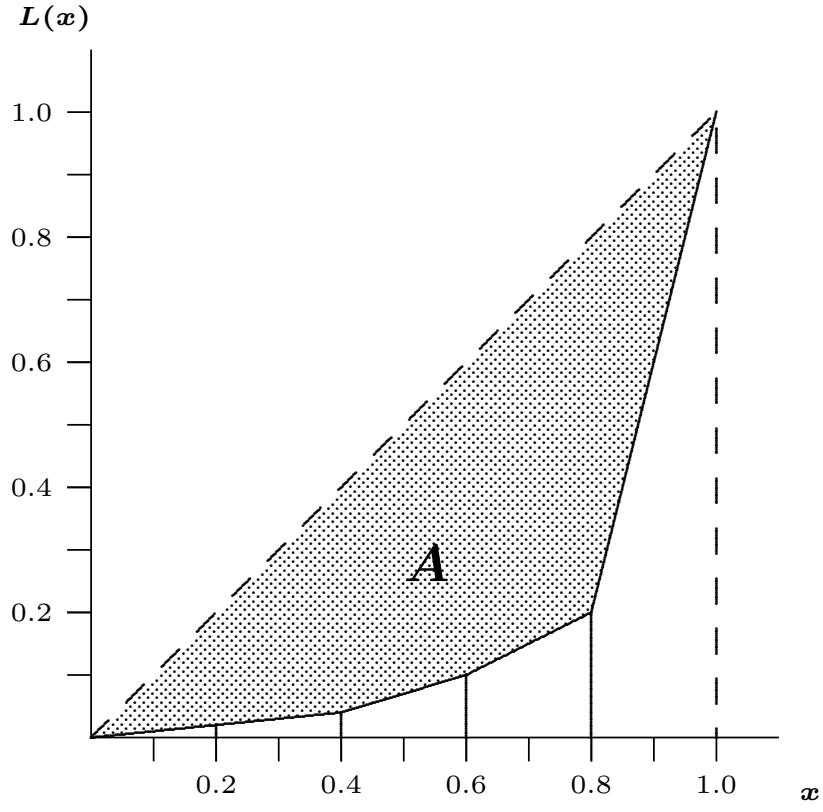


Eigenschaften der Lorenzkurve $L(x)$

- $0 \leq x \leq 1$
- $0 \leq L(x) \leq 1$ mit $L(0) = 0$ und $L(1) = 1$
- $L(x) \leq x$
- $L(x)$ ist konvex
- $L(x)$ ist eine monoton nichtfallende Funktion
- Knickstelle in $(u_i, v_i) \Longleftrightarrow x_{(i+1)} > x_{(i)}$

Gini-Koeffizient

Ziel: Maßzahl für die Konzentration



Verwende den Flächeninhalt A zwischen der ersten Winkelhalbierenden und der Lorenzkurve!

- Numerisches Maß der Konzentration:

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

$$G = \frac{2\sum_{i=1}^{n} i x_{(i)} - (n+1)\sum_{i=1}^{n} x_{(i)}}{n\sum_{i=1}^{n} x_{(i)}}$$

- **Problem:** $G_{\max} = \frac{n-1}{n}$
- **Normierter Gini-Koeffizient:**

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

- Je größer G_* ist, desto stärker ist die Konzentration.

Beispiel I:

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

i	1	2	3	4	5	$\sum$
$x_{(i)}$	2	3	3	6	11	25

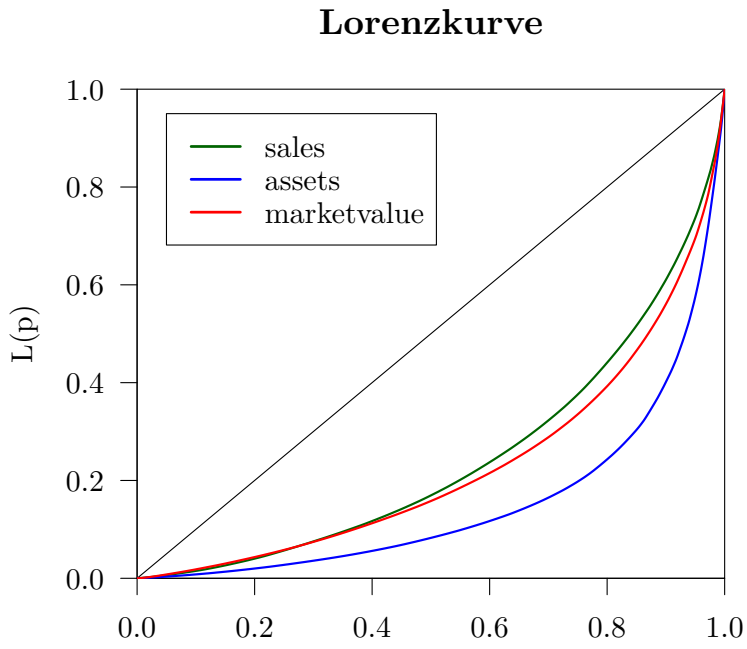
$$G = \frac{2 \cdot \big(1 \cdot 2 + 2 \cdot 3 + 3 \cdot 3 + 4 \cdot 6 + 5 \cdot 11\big) - 6 \cdot 25}{5 \cdot 25} = 0.336$$

Mit $G_{\max} = \frac{5-1}{5} = 0.8$ folgt $G_* = \frac{5}{5-1} \cdot 0.336 = 0.42$

```
## install.packages("ineq")  # Nur bei der ersten Verwendung
library(ineq)
umsatz <- c(6, 3, 11, 2, 3)
Gini(umsatz)
## [1] 0.336
Gini(umsatz, corr = TRUE)
## [1] 0.42
```

Beispiel: Größte Unternehmen (vgl. F19)

```
stetigeVar <- c("sales", "profits", "assets", "marketvalue")
apply(ForbesG7[, stetigeVar], 2, function(x) Gini(x, corr = TRUE))
##       sales     profits      assets marketvalue
##   0.5086780   0.9984003   0.6964531   0.5438243
plot(Lc(ForbesG7[, "sales"]), col = "darkgreen", main = "Lorenzkurve")
lines(Lc(ForbesG7[, "assets"]), col = "blue")
lines(Lc(ForbesG7[, "marketvalue"]), col = "red")
legend(0.05, 0.95, c("sales", "assets", "marketvalue"), lty = rep(1,
    3), lwd = rep(2.5, 3), col = c("darkgreen", "blue", "red"))
```



1.3. Darstellung und Kenngrößen bivariater Datenmengen

jetzt: 2 Merkmale X, Y. Stichprobe: $(x_1, y_1), \ldots, (x_n, y_n)$

- Für die beiden Variablen X und Y können einzeln die Kenngrößen für einen univariaten Datensatz bestimmt werden.
- Bei einem bivariaten Datensatz ist man allerdings insbesondere auch an der Beziehung zwischen den Variablen X und Y interessiert. Darauf gehen wir im folgenden näher ein.

Streuungsdiagramm

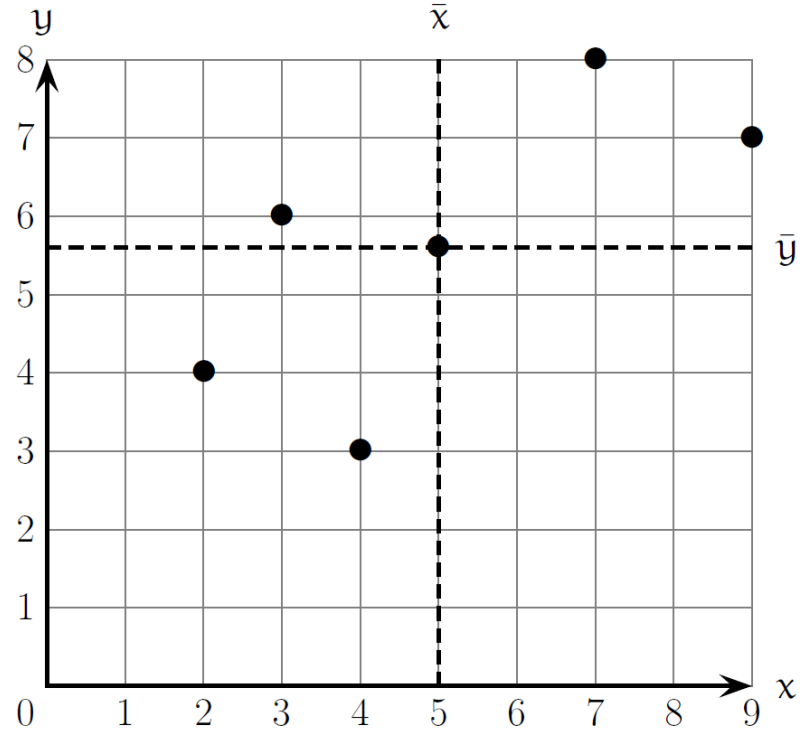
Sinnvoll bei vielen verschiedenen Ausprägungen (z.B. stetige Merkmale)

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

Beispiel (Nachfragemengen): Nachfrage nach den Gütern X und Y (in Tsd. Stück)

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

$$\Rightarrow \bar{x} = \frac{25}{5} = 5$$
$$\bar{y} = \frac{28}{5} = 5.6$$



Korrelationsrechnung

- Die Stärke des Zusammenhangs zwischen zwei Merkmalen X und Y wird mittels **Korrelationskoeffizienten** gemessen.
- Wahl abhängig vom Skalenniveau von X und Y:

Skalierung von X und Y	Korrelationskoeffizient
mind. nominal	Kontingenzkoeffizient
mind. ordinal	Rangkorrelationskoeffizient von Spearman
kardinal	Bravais–Pearson–Korrelationskoeffizient

1.3.1 Bravais–Pearson–Korrelation

Voraussetzung: X und Y kardinalskaliert

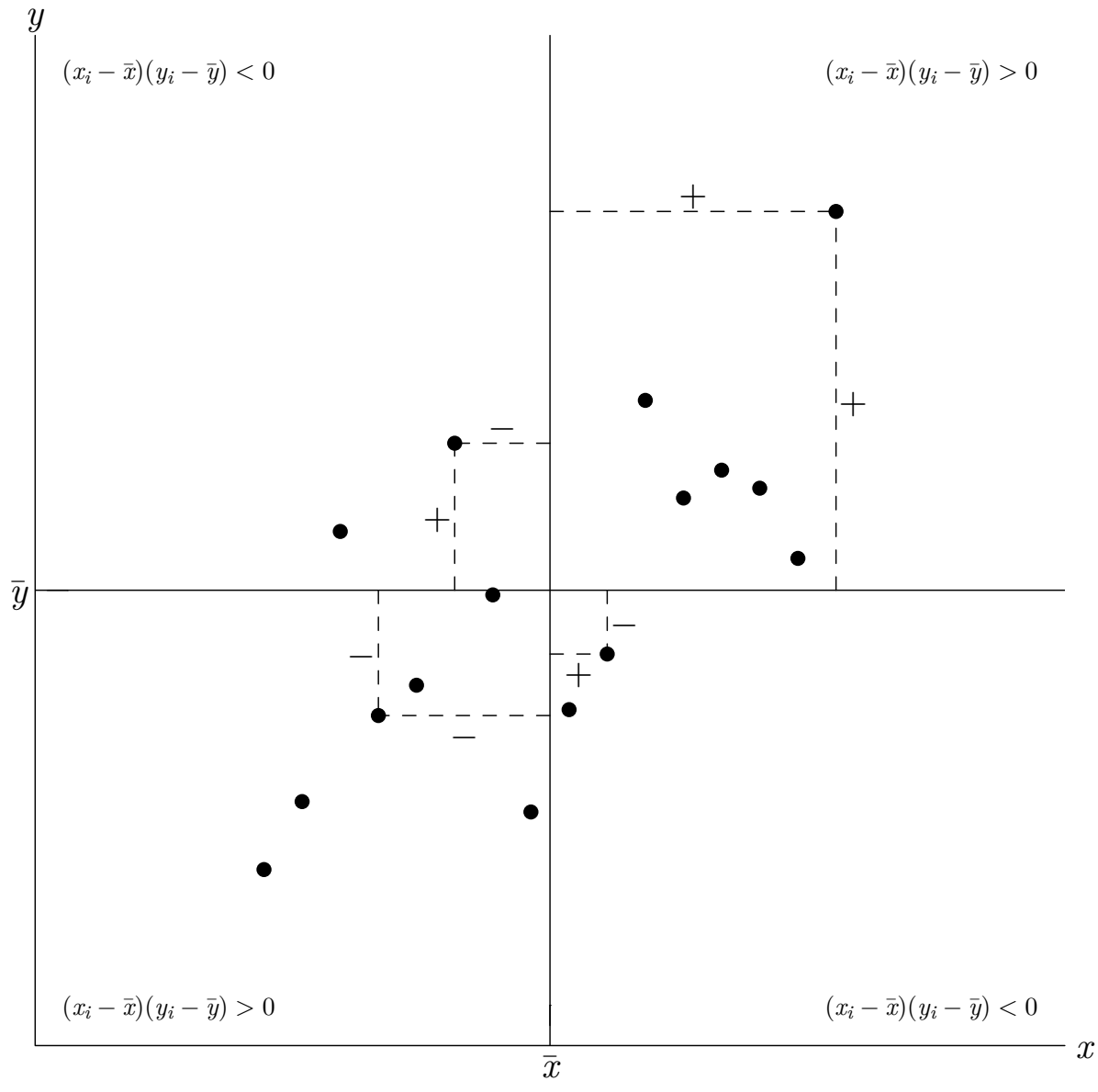
- **positiver** Zusammenhang: große (kleine) Werte von X gehen mit großen (kleinen) Werten von Y einher; die Datenpaare (x_i, y_i) lassen sich durch eine Gerade pos. Steigung beschreiben
- **negativer** Zusammenhang: bei umgekehrter Tendenz

Empirische Kovarianz

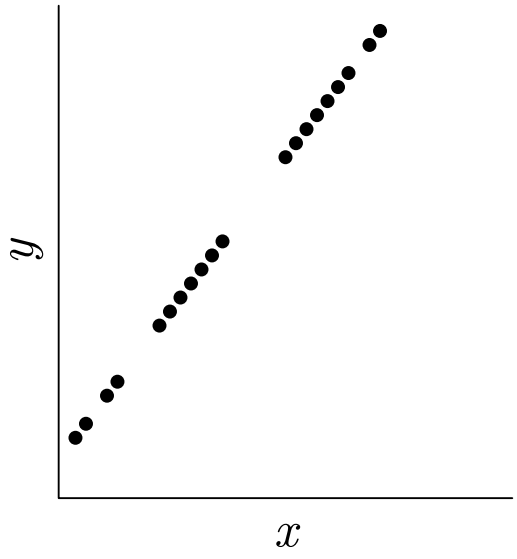
$$s_{XY} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Bravais-Pearson-Korrelationskoeffizient

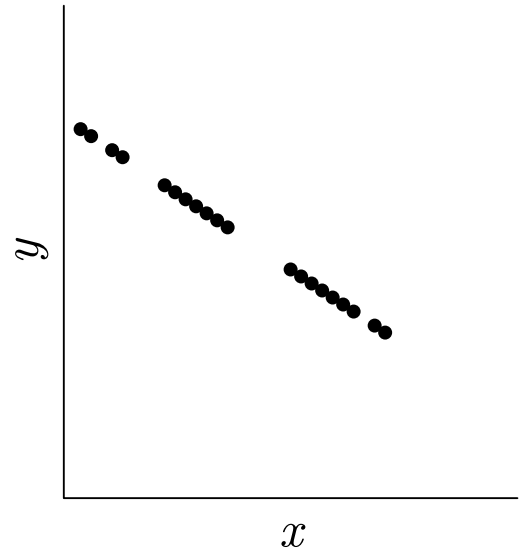
$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \sqrt{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}}$$


y
$(x_i - \bar{x})(y_i - \bar{y}) < 0$
$(x_i - \bar{x})(y_i - \bar{y}) > 0$
$\bar{y}$
$(x_i - \bar{x})(y_i - \bar{y}) > 0$
$(x_i - \bar{x})(y_i - \bar{y}) < 0$
$\bar{x}$
x

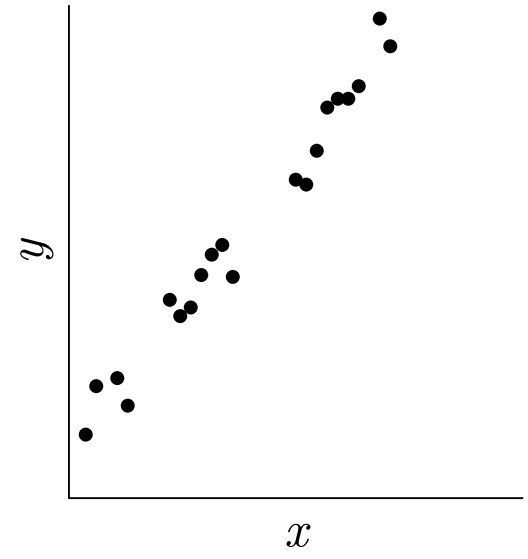
$r_{XY} = +1$



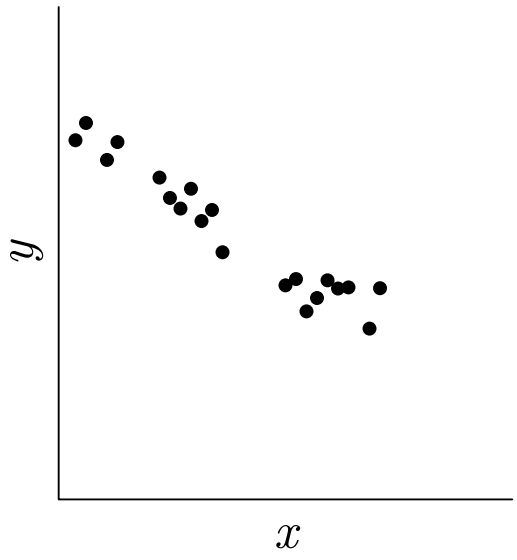
$r_{XY} = -1$



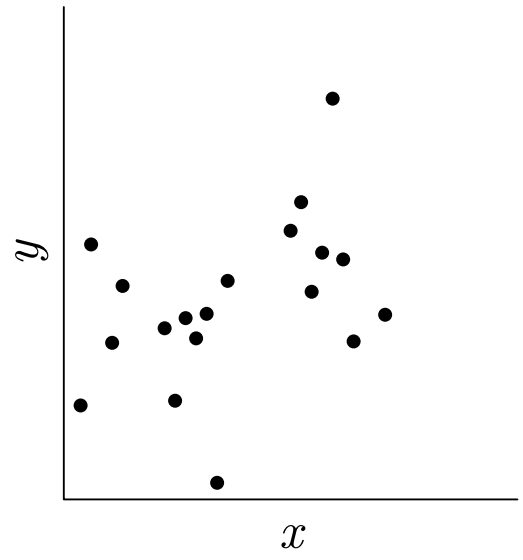
r_{XY} stark positiv



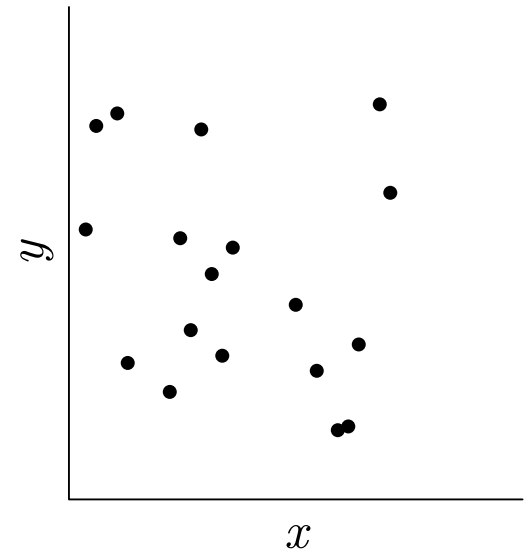
r_{XY} stark negativ



r_{XY} schwach positiv



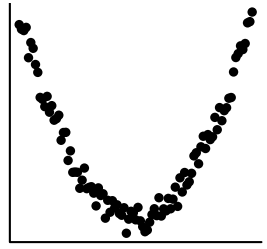
$r_{XY} \approx 0$



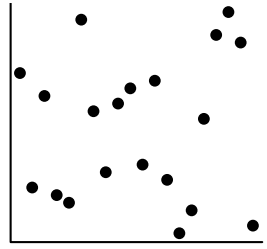
Eigenschaften:

- $r_{XY} = r_{YX}$
- Invarianz in Bezug auf lineare Transformationen:
  $r_{XY} = r_{X^*Y^*}$, für $X^* = a + bX$ und $Y^* = c + dY$ mit $b \cdot d > 0$
- $-1 \leq r_{XY} \leq 1$
- $r_{XY} = 1$ (bzw. -1), falls alle Beobachtungen auf einer Geraden mit positiver (bzw. negativer) Steigung liegen.
- **Beachte:** Es kann nicht auf einen kausalen Zusammenhang geschlossen werden!

Der empirische Korrelationskoeffizient stellt ein Maß für den **linearen** Zusammenhang zweier Merkmale dar.



$r_{XY} \approx 0$



$r_{XY} \approx 0$

$\lvert r \rvert$	Interpretation
≈ 0	keine Korrelation
$(\ldots, 0.3)$	schwache Korrelation
$[0.3, 0.7)$	mittlere Korrelation
$[0.7, 1)$	starke Korrelation
1	perfekte Korrelation

Beispiel (Qualitätsmanagement-Reklamationen): Zusammenhang zwischen den Investitionen in das Qualitätsmanagement (X, in Tsd. €) und den Kosten für Reklamationen (Y, in Tsd. €)

i	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	2	10	4	100	20
2	4	7	16	49	28
3	3	7	9	49	21
4	9	3	81	9	27
5	7	5	49	25	35
$\sum$	25	32	159	232	131

$$\Rightarrow \quad \bar{x} = \frac{25}{5} = 5$$

$$\bar{y} = \frac{32}{5} = 6.4$$

$$r = \frac{131 - 5 \cdot 5 \cdot 6.4}{\sqrt{159 - 5 \cdot 5^2}\sqrt{232 - 5 \cdot 6.4^2}} = -0.95$$

(starker negativer linearer Zusammenhang)

```
x <- c(2, 4, 3, 9, 7)
y <- c(10, 7, 7, 3, 5)
cor(x, y, method = "pearson")
## [1] -0.9536172
```

1.3.2. Rangkorrelationskoeffizient von Spearman

Voraussetzung: X und Y (mindestens) ordinalskaliert

Idee der Rangbildung: Ordne jeder Beobachtung der Stichprobe $x_1, \ldots, x_n$ ihre Position in der geordneten Stichprobe $x_{(1)}, \ldots, x_{(n)}$ zu:

$$R(x_j) = v \quad \Leftrightarrow \quad x_j = x_{(v)} \quad (\text{d.h. } R(x_{(j)}) = j)$$

$R(x_j)$ heißt **Rang** der Beobachtung x_j.

Beispiel: Nutzen eines Gutes für verschiedene Personen

- $x_1 = 2, x_2 = 5, x_3 = 1, x_4 = 3$
- Geordnete Stichprobe: $x_3 < x_1 < x_4 < x_2$.
- Somit ist $R(x_1) = 2$, $R(x_2) = 4$, $R(x_3) = 1$, $R(x_4) = 3$.

```
rank(c(2, 5, 1, 3))
## [1] 2 4 1 3
```

Treten gleiche Beobachtungen (**Bindungen**) auf, so wird meistens die Methode der **Durchschnittsränge** („geteilte Plätze“) verwendet.

Beispiel: Nutzen eines Gutes für verschiedene Personen

- $x_1 = 3$, $x_2 = 2$, $x_3 = 3$, $x_4 = 5$, $x_5 = 2$
- Geordnete Stichprobe: $x_2 = x_5 < x_1 = x_3 < x_4$ (nicht eindeutig)
- $$R(x_1) = \frac{3+4}{2} = 3.5, \quad R(x_2) = \frac{1+2}{2} = 1.5,$$
  $$R(x_3) = \frac{3+4}{2} = 3.5, \quad R(x_4) = 5,$$
  $$R(x_5) = \frac{1+2}{2} = 1.5$$

```
rank(c(3, 2, 3, 5, 2))
## [1] 3.5 1.5 3.5 5.0 1.5
```

Beachte:

$$\sum_{i=1}^{n} R(x_i) = \frac{1}{2} n(n+1)$$

Vorgehensweise: Stichprobe $(x_1, y_1), \ldots, (x_n, y_n)$. Ordne $x_1, .., x_n$ die Ränge $R(x_1), .., R(x_n)$ zu und $y_1, .., y_n$ die Ränge $R(y_1), .., R(y_n)$.

Rangkorrelationskoeffizient von Spearman

$$r_{SP} = r_{R(X),R(Y)} = \frac{\sum_{i=1}^{n} \left(R(x_i) - \bar{R}\right)\left(R(y_i) - \bar{R}\right)}{\sqrt{\sum_{i=1}^{n} \left(R(x_i) - \bar{R}\right)^2 \sum_{i=1}^{n} \left(R(y_i) - \bar{R}\right)^2}} =$$

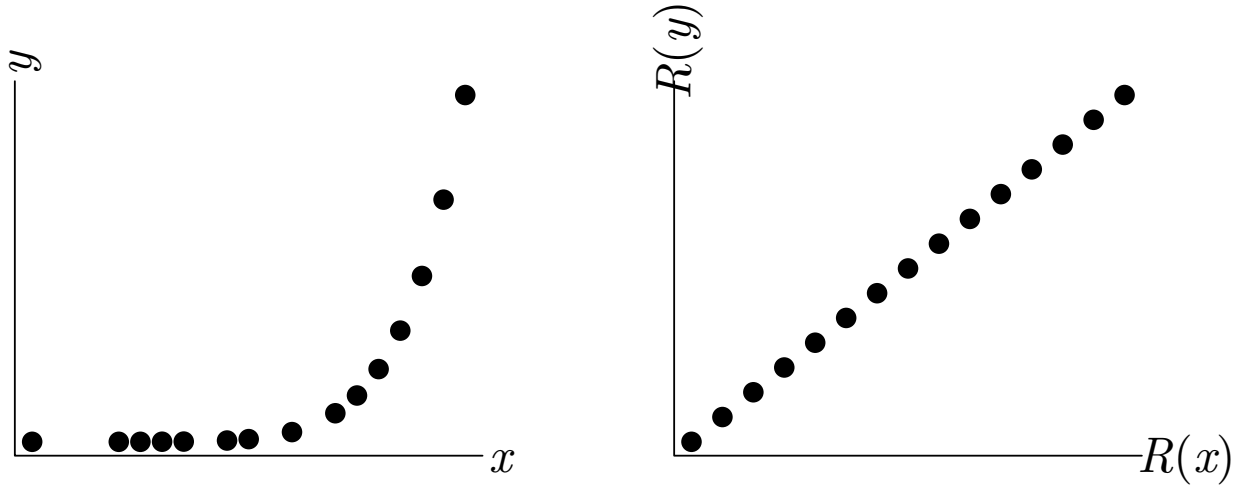
$$= \frac{\sum_{i=1}^{n} R(x_i)\, R(y_i) \; - \; n\, \bar{R}^2}{\sqrt{\left(\sum_{i=1}^{n} R(x_i)^2 \; - \; n\, \bar{R}^2\right)\left(\sum_{i=1}^{n} R(y_i)^2 \; - \; n\, \bar{R}^2\right)}}$$

Dabei ist $\bar{R} = (n+1)/2$.

Falls keine Bindungen vorliegen, verwendet man:

$$r_{SP} = 1 - \frac{6 \sum_{i=1}^{n} (R(x_i) - R(y_i))^2}{(n-1)n(n+1)}$$

Der Spearman–Korrelationskoeffizient stellt ein Maß für den **monotonen** Zusammenhang zweier Merkmale dar.



$$r_{XY} = 0.77, \; r_{SP} = r_{R(X),R(X)} = 1$$

- **positiver monotoner** Zusammenhang: niedrige (hohe) Ränge von X gehen mit niedrigen (hohen) Ränge von Y einher; die Datenpaare $(R(x_i), R(y_i))$ lassen sich durch eine Gerade positiver Steigung beschreiben; die Datenpaare (x_i, y_i) lassen sich durch eine streng monoton steigende Kurve beschreiben
- **negativer montoner** Zusammenhang: bei umgekehrter Tendenz

Eigenschaften:

- $-1 \leq r_{SP} \leq 1$
- Invarianz in Bezug auf monotone Transformationen (z.B. log)
- $r_{SP} = +1$ wird erreicht bei $R(x_i) = R(y_i) \qquad \forall\, i = 1, \ldots, n$
- $r_{SP} = -1$ wird erreicht bei $R(x_i) = n + 1 - R(y_i) \qquad \forall\, i = 1, \ldots, n$
- Rangfolge bei X **und** Y „umdrehen“ $\rightsquigarrow$ r_{SP} ändert sich nicht

Beispiel: Qualitätsmanagement-Reklamationen (vgl. F85)

i	x_i	y_i	$R(x_i)$	$R(y_i)$	$R(x_i)^2$	$R(y_i)^2$	$R(x_i)R(y_i)$
1	2	10	1	5	1	25	5
2	4	7	3	$\frac{1}{2}(3+4)=3.5$	9	12.25	10.5
3	3	7	2	$\frac{1}{2}(3+4)=3.5$	4	12.25	7
4	9	3	5	1	25	1	5
5	7	5	4	2	16	4	8
$\sum$			15	15	55	54.5	35.5

$$\bar{R} = \frac{5+1}{2} = 3$$

$$r_{SP} = \frac{35.5 - 5 \cdot 3^2}{\sqrt{55 - 5 \cdot 3^2}\sqrt{54.5 - 5 \cdot 3^2}} = -0.97$$

(starker negativer monotoner Zusammenhang)

```
cor(x, y, method = "spearman")
## [1] -0.9746794
```

1.3.3. Korrelationsmaße bei nominal skalierten Merkmalen

jetzt: 2 nominale oder klassierte Merkmale X und Y mit Ausprägungen $a_1, \ldots, a_k$ zu X und $b_1, \ldots, b_l$ zu Y und Beobachtungspaaren (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n)

Kontingenztabelle

		Ausprägungen von Y				
		b_1	b_2	$\ldots$	b_ℓ	Σ
Ausprägungen von X	a_1	n_{11}	n_{12}	$\ldots$	$n_{1\ell}$	$n_{1\bullet}$
	a_2	n_{21}	n_{22}	$\ldots$	$n_{2\ell}$	$n_{2\bullet}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	a_k	n_{k1}	n_{k2}	$\ldots$	$n_{k\ell}$	$n_{k\bullet}$
	Σ	$n_{\bullet 1}$	$n_{\bullet 2}$	$\ldots$	$n_{\bullet \ell}$	n

- **absolute Häufigkeit für (a_i, b_j):**

$n_{ij} = n(X = a_i, Y = b_j) \quad \rightsquigarrow \quad$ Anzahl der Fälle, in denen das Paar (a_i, b_j) in der Stichprobe auftritt

- **absolute Randhäufigkeit von a_i:**

$n_{i\bullet} = \sum_{j=1}^{\ell} n_{ij} \quad \rightsquigarrow \quad$ Anzahl der Fälle, in denen die Ausprägung a_i in $x_1, \ldots, x_n$ auftritt

- **absolute Randhäufigkeit von b_j:**

$n_{\bullet j} = \sum_{i=1}^{k} n_{ij} \quad \rightsquigarrow \quad$ Anzahl der Fälle, in denen die Ausprägung b_j in $y_1, \ldots, y_n$ auftritt

Beispiel (Neukundengewinnung): Zusammenhang zwischen dem Erfolg der Neukundengewinnung und der durchgeführten Maßnahme

X	Y			$n_{i\bullet}$
	telefonisch	E-Mail	Werbung per Post	
	$(= b_1)$	$(= b_2)$	$(= b_3)$	
Neukunde	264	90	6	360
$(= a_1)$	$(= n_{11})$	$(= n_{12})$	$(= n_{13})$	$(= n_{1\bullet})$
kein Neukunde	2	34	4	40
$(= a_2)$	$(= n_{21})$	$(= n_{22})$	$(= n_{23})$	$(= n_{2\bullet})$
$n_{\bullet j}$	266	124	10	400
	$(= n_{\bullet 1})$	$(= n_{\bullet 2})$	$(= n_{\bullet 3})$	$(= n)$

- **relative Häufigkeit** für (a_i, b_j):

  $h_{ij} = h(X = a_i, Y = b_j) = n_{ij}/n$

- **relative Randhäufigkeit** von a_i (b_j):

  $h_{i\bullet} = h(X = a_i) = n_{i\bullet}/n \qquad (h_{\bullet j} = n_{\bullet j}/n)$

	b_1	b_2	$\dots$	b_ℓ	Σ
a_1	h_{11}	h_{12}	$\dots$	$h_{1\ell}$	$h_{1\bullet}$
a_2	h_{21}	h_{22}	$\dots$	$h_{2\ell}$	$h_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
a_k	h_{k1}	h_{k2}	$\dots$	$h_{k\ell}$	$h_{k\bullet}$
Σ	$h_{\bullet 1}$	$h_{\bullet 2}$	$\dots$	$h_{\bullet\ell}$	1

X	Y Tel.	E-Mail	Post	$n_{i\bullet}$
	$(= b_1)$	$(= b_2)$	$(= b_3)$	
NK	0.66	0.225	0.015	0.90
$(= a_1)$	$(= h_{11})$	$(= h_{12})$	$(= h_{13})$	$(= h_{1\bullet})$
kein NK	0.005	0.085	0.01	0.10
$(= a_2)$	$(= h_{21})$	$(= h_{22})$	$(= h_{23})$	$(= h_{2\bullet})$
$h_{\bullet j}$	0.665	0.31	0.025	1
	$(= h_{\bullet 1})$	$(= h_{\bullet 2})$	$(= h_{\bullet 3})$	

bedingte relative Häufigkeit für a_i von X, falls Y den Wert b_j besitzt:

$$h(X = a_i \,|\, Y = b_j) = \frac{n(X = a_i, Y = b_j)}{n(Y = b_j)} = \frac{n_{ij}}{n_{\bullet j}}$$

bedingte relative Häufigkeit für b_j von Y, falls X den Wert a_i besitzt:

$$h(Y = b_j \,|\, X = a_i) = \frac{n(X = a_i, Y = b_j)}{n(X = a_i)} = \frac{n_{ij}}{n_{i\bullet}} = \frac{h(X = a_i, Y = b_j)}{h(X = a_i)}$$

Beispiel: Neukundengewinnung (vgl. F94)

- $h(b_1|a_1) = \frac{n_{11}}{n_{1\bullet}} = \frac{264}{360} = 0.73 \quad \rightsquigarrow \quad$ 73% derjenigen, die als NK gewonnen wurden, wurden telefonisch kontaktiert.
- $h(a_1|b_1) = \frac{n_{11}}{n_{\bullet 1}} = \frac{264}{266} = 0.99 \quad \rightsquigarrow \quad$ 99% derjenigen, die telefonsich kontaktiert wurde, wurden als NK gewonnen.

X heißt unabhängig von Y, falls für alle $i \in \{1, \ldots, k\}$ gilt

$$h(X = a_i \,|\, Y = b_1) = h(X = a_i \,|\, Y = b_2) = \ldots = h(X = a_i \,|\, Y = b_l)\,.$$

Eigenschaften:

Ist X unabhängig von Y, so gilt

- $h(X = a_i) = h(X = a_i | Y = b_j)$.
- Y ist unabhängig von X. Deshalb: **X und Y sind unabhängig.**
- X und Y sind genau dann unabhängig, wenn **für alle** i, j gilt

$$h_{ij} = h_{i\bullet} \cdot h_{\bullet j} =: \tilde{h}_{ij} \quad \text{bzw.} \quad n_{ij} = n_{i\bullet} \cdot n_{\bullet j}/n =: \tilde{n}_{ij}.$$

Beispiel: Neukundengewinnung (vgl. F94)

$$n_{11} = 264 \neq 239.4 = \frac{360 \cdot 266}{400} = \frac{n_{1\bullet} \cdot n_{\bullet 1}}{n}$$

$\rightsquigarrow$ Die Merkmale sind nicht unabhängig.

Ziel: Maß für die Abhängigkeit

Vorgehensweise:

- Kontingenztabelle mit Randhäufigkeiten
- Berechne „theoretische Häufigkeiten“(optional)

$$\tilde{n}_{ij} := \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}$$

- Berechne:

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{\ell} \frac{(n_{ij} - \tilde{n}_{ij})^2}{\tilde{n}_{ij}} = n \left(\sum_{i=1}^{k} \sum_{j=1}^{l} \frac{n_{ij}^2}{n_{i\bullet} \cdot n_{\bullet j}} - 1 \right)$$

- Im Fall $k = \ell = 2$ gilt:

$$\chi^2 = \frac{n \cdot (n_{11}n_{22} - n_{12}n_{21})^2}{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}}$$

Eigenschaften

- $\chi^2 \geq 0$
- X, Y unabhängig $\Rightarrow \chi^2 = 0$
- χ^2 wächst mit Interdependenz
- Aber χ^2 hängt von n ab! ($h_{ij} \mapsto 2 \cdot h_{ij} \Rightarrow \chi^2 \mapsto 2 \cdot \chi^2$)

Kontingenzkoeffizient:

$$K = \sqrt{\frac{\chi^2}{n + \chi^2}} \in [0; K_{\max}], \qquad \text{wobei} \quad K_{\max} = \sqrt{\frac{\min\{k, l\} - 1}{\min\{k, l\}}}$$

Normierter Kontingenzkoeffizient:

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

$K_* = 1 \iff$ Rückschluss $x_i \leftrightarrow y_i$ möglich (zumindest in einer Richtung)

Beispiel: Neukundengewinnung (vgl. F94)

$$\begin{aligned}\chi^2 &= n\left(\sum_{i=1}^{k}\sum_{j=1}^{l}\frac{n_{ij}^2}{n_{i\bullet}\cdot n_{\bullet j}} - 1\right)\\ &= 400\cdot\left(\frac{264^2}{360\cdot 266}+\frac{90^2}{360\cdot 124}+\frac{6^2}{360\cdot 10}\right.\\ &+ \left.\frac{2^2}{40\cdot 266}+\frac{34^2}{40\cdot 124}+\frac{4^2}{40\cdot 10}-1\right) = 77.085\end{aligned}$$

Hiermit

$$K = \sqrt{\frac{\chi^2}{\chi^2+n}} = 0.402,$$

$$K_{\max} = \sqrt{\frac{\min\{k,\ell\}-1}{\min\{k,\ell\}}} = \sqrt{\frac{\min\{2,3\}-1}{\min\{2,3\}}} = \sqrt{\frac{2-1}{2}} = 0.707$$

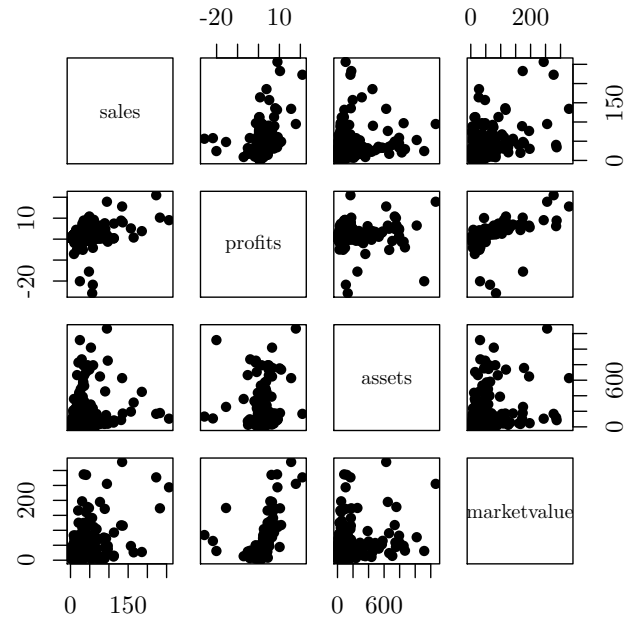
$\rightsquigarrow$ $K_* = K/K_{\max} = 0.402/0.707 = 0.569 \rightsquigarrow$ mittlerer Zusammenhang

Beispiel: Neukundengewinnung (vgl. F94)

```
## install.packages("vcd")  # Nur bei der ersten Verwendung
library(vcd)
## kundengewinnung <- read.table("kundengewinnung.txt", header = TRUE)
(tab.KundeMassnahme <- table(kundengewinnung$Neukunde, kundengewinnung$Massnahme))
##
##        Mail Post Telefon
##   ja     90    6     264
##   nein   34    4       2
apply(tab.KundeMassnahme, 1, sum)  # Randhäufigkeiten
##   ja nein
##  360   40
prop.table(tab.KundeMassnahme)  # Relative Häufigkeitstabelle
##
##         Mail  Post Telefon
##   ja   0.225 0.015   0.660
##   nein 0.085 0.010   0.005
assocstats(tab.KundeMassnahme)  # Kontingenzkoeffizient
##                     X^2 df P(> X^2)
## Likelihood Ratio 77.388  2        0
## Pearson          77.085  2        0
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.402
## Cramer's V        : 0.439
```

Beispiel: Größte Unternehmen (vgl. F19)

```
pairs(~sales + profits + assets + marketvalue, data = ForbesG7,
    pch = 16)
```




```
cor(ForbesG7[, stetigeVar])
##                  sales   profits    assets marketvalue
## sales       1.0000000 0.3692856 0.3169091   0.5522812
## profits     0.3692856 1.0000000 0.1555089   0.5308211
## assets      0.3169091 0.1555089 1.0000000   0.3815484
## marketvalue 0.5522812 0.5308211 0.3815484   1.0000000
```

Beispiel: Größte Unternehmen (vgl. F19)

```
cor(ForbesG7[, stetigeVar], method = "spearman")
##                  sales   profits    assets marketvalue
## sales       1.0000000 0.2602629 0.4738247   0.4856336
## profits     0.2602629 1.0000000 0.2636245   0.6450604
## assets      0.4738247 0.2636245 1.0000000   0.4716245
## marketvalue 0.4856336 0.6450604 0.4716245   1.0000000
tab.CountryCategory <- table(ForbesG7$country, ForbesG7$category)
assocstats(tab.CountryCategory)$cont
## [1] 0.5632128
```

1.4. Regressionsanalyse

- **Ziel:** Den Einfluss eines erklärenden Merkmals auf ein abhängiges Merkmal zu modellieren.
- Interpretiere Y als Funktion von X:

$$Y = f(X)$$

- X heißt **Regressor** bzw. unabhängige Variable
  Y heißt **Regressand** bzw. abhängige Variable
- Es gibt unendlich viele Funktionen $\rightsquigarrow$ der wichtigste Fall: f ist eine lineare Funktion

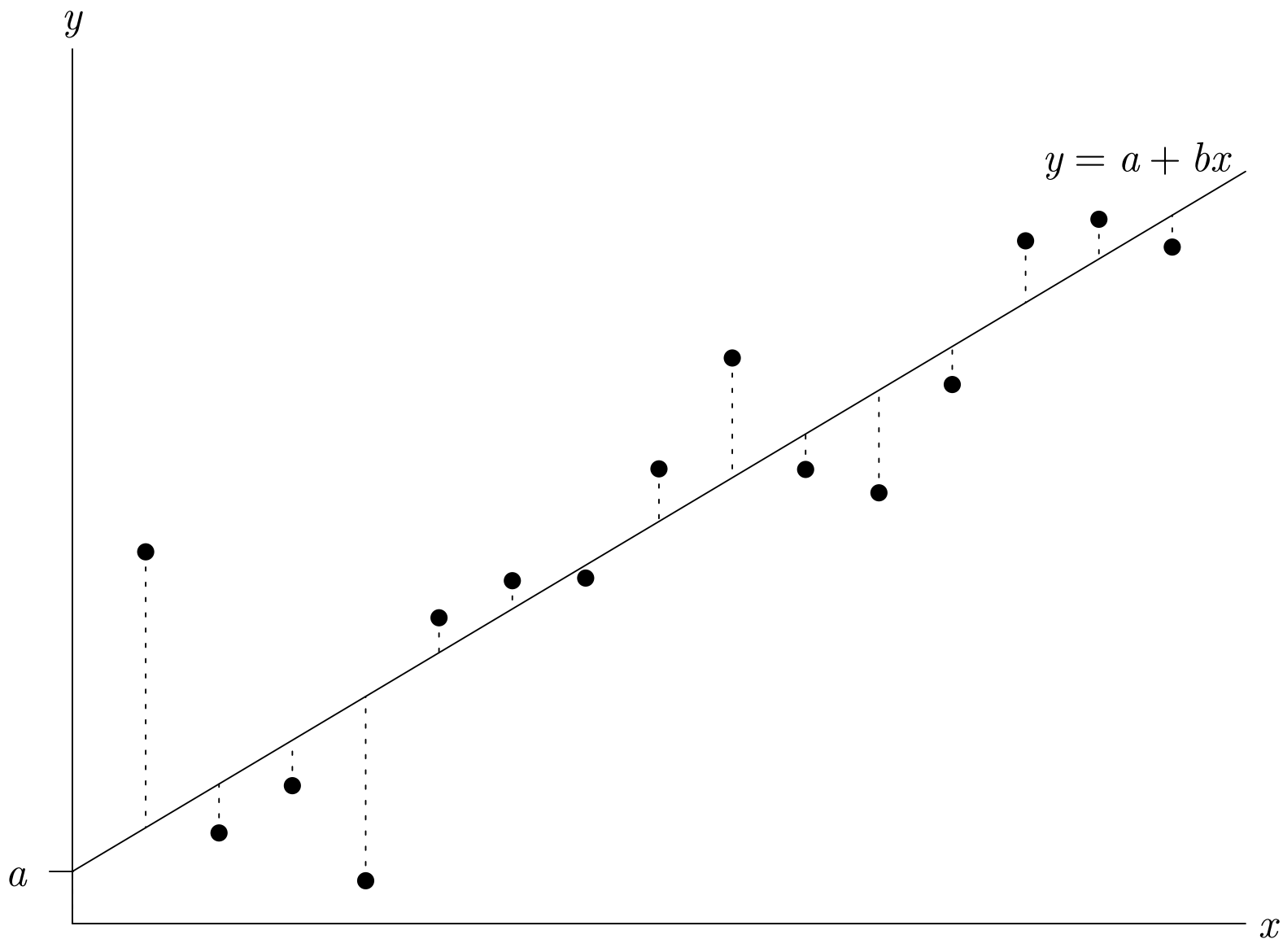
$$y = a + bx$$

- **Beachte:** die lineare Funktion f kann als Approximation einer unbekannten wahren Funktion angesehen werden.

- **Lineare Regression:** Da a und b unbekannt sind, müssen sie geschätzt werden.
- Kriterium für die Schätzung: **die Methode der kleinsten Quadrate (MKQ)**.

  Wähle a, b so, dass die Summe der quadrierten Abweichungen zwischen den beobachteten Werten und prognostizierten Werten aus dem Modell am kleinsten ist, d.h.

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


y
y = a + bx
a
x

- Das Optimierungsproblem nach zwei Parametern a und b ergibt eine eindeutige Lösung:

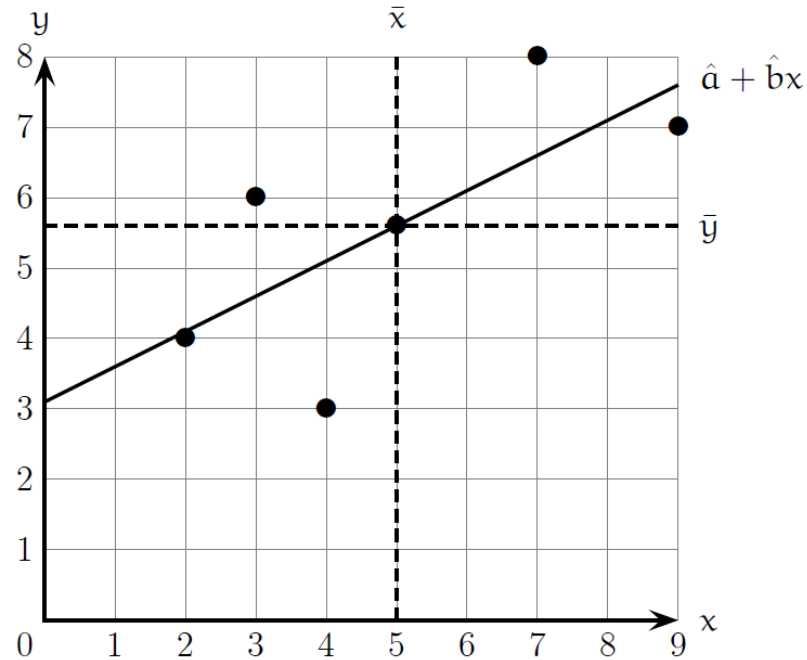
$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \qquad \text{und} \qquad \hat{a} = \bar{y} - \hat{b}\bar{x}$$

- Regressionsgerade: $y = \hat{a} + \hat{b}x$

Beispiel: Nachfragemengen (vgl. F78)

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

$$
\begin{aligned}
n &= 5 \\
\bar{x} &= 5 \\
\bar{y} &= 5.6 \\
\Rightarrow \hat{b} &= \frac{157 - 5 \cdot 5 \cdot 5.6}{159 - 5 \cdot 5^2} = 0.5 \\
\hat{a} &= 5.6 - 0.5 \cdot 5 = 3.1 \\
\Rightarrow y &= 3.1 + 0.5x
\end{aligned}
$$



Determinationskoeffizient

- **Frage:** Wie gut beschreibt $\hat{a} + \hat{b}x$ den Zusammenhang von X und Y?
- Durch die Regression prognostizierten Werte:

$$\hat{y}_i = \hat{a} + \hat{b}x_i$$

- **Prognosefehler**: die Abweichungen zwischen den wahren und den Prognosewerten

$$\hat{u}_i = y_i - \hat{y}_i$$

- $Q(\hat{a}, \hat{b}) = \sum_{i=1}^{n} \hat{u}_i^2$ als Gütemaß ungeeignet, da beliebig groß
- **Determinationskoeffizient:**

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

Eigenschaften von R^2

- $R^2 \in [0, 1]$
- $R^2 = 0$ wird erreicht wenn X, Y unkorreliert
- $R^2 = 1$ wird erreicht wenn $\hat{y}_i = y_i \; \forall \, i$ (alle Punkte auf Regressionsgerade)
- R^2 heißt auch ...
  - quadrierter multipler Korrelationskoeffizient ($R^2 = r^2_{XY}$)
  - durch die Regression erklärter Anteil der Varianz

Beispiel: Nachfragemengen (vgl. F78, F108)
$\hat{y}_i = 3.1 + 0.5x_i, \;\; n = 5, \;\; \bar{y} = 5.6, \;\; \sum y_i^2 = 174$

i	1	2	3	4	5
x_i	2	4	3	9	7
$\hat{y}_i$	4.1	5.1	4.6	7.6	6.6

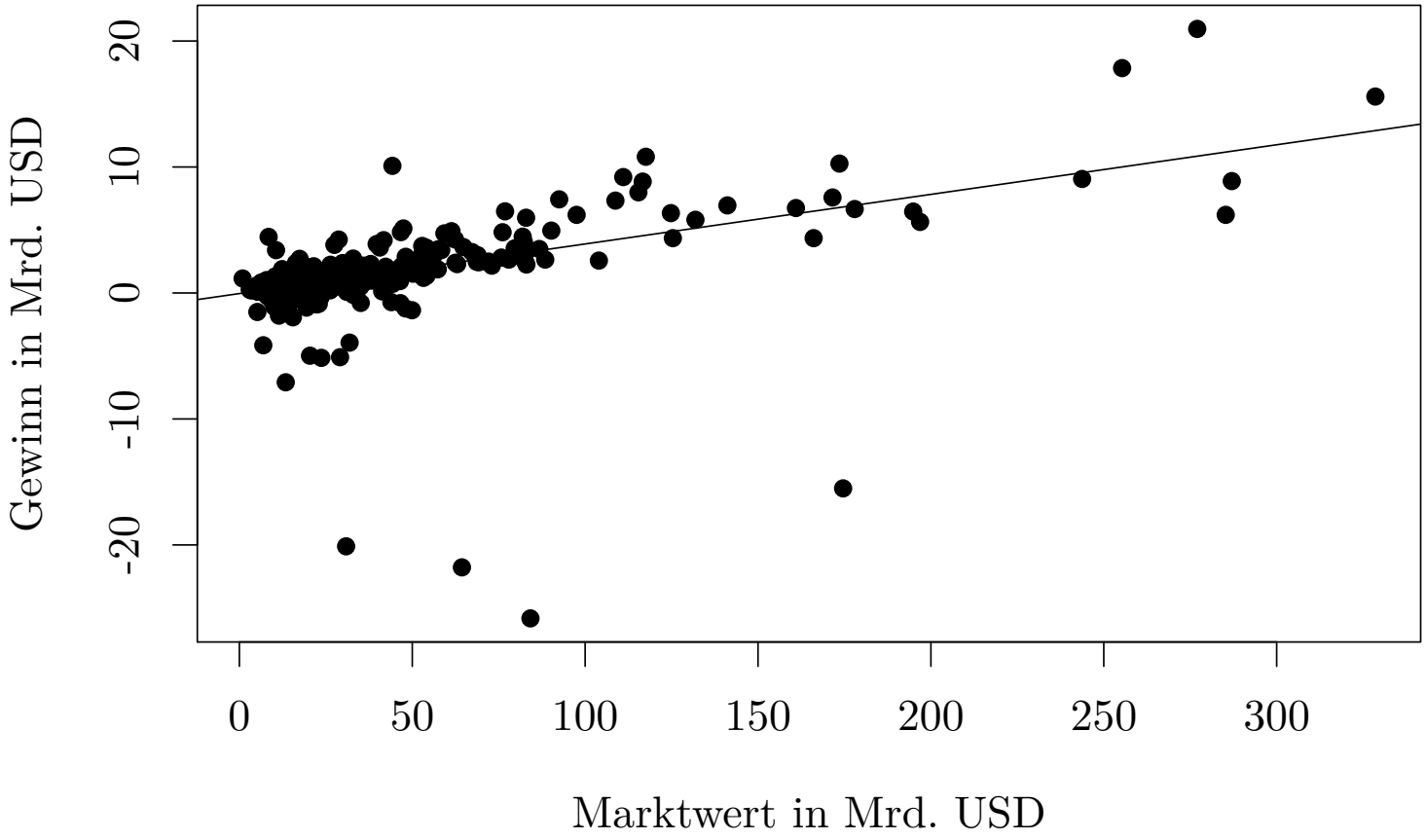
$$\Rightarrow \begin{aligned} R^2 &= \tfrac{4.1^2 + \cdots + 6.6^2 - 5 \cdot 5.6^2}{174 - 5 \cdot 5.6^2} = 0.4942 \\ R^2 &= r^2_{XY} = 0.703^2 = 0.4942 \end{aligned}$$

Beispiel: Größte Unternehmen (vgl. F19)

```
model <- lm(profits ~ marketvalue, data = ForbesG7)
summary(model)
##
## Call:
## lm(formula = profits ~ marketvalue, data = ForbesG7)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.0961  -0.0721   0.1634   0.3839  10.1029
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.047610   0.141115  -0.337    0.736
## marketvalue  0.039364   0.002816  13.978   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.582 on 498 degrees of freedom
## Multiple R-squared:  0.2818,Adjusted R-squared:  0.2803
## F-statistic: 195.4 on 1 and 498 DF,  p-value: < 2.2e-16
```

Beispiel: Größte Unternehmen (vgl. F19)

```
plot(ForbesG7$marketvalue, ForbesG7$profits, xlab = "Marktwert in Mrd. USD
    ylab = "Gewinn in Mrd. USD", pch = 16)
abline(model)
```



Beispiel: Größte Unternehmen (vgl. F19)

```
model$coefficients
## (Intercept) marketvalue
## -0.04761045  0.03936423
head(model$fitted.values)
##         1         2         3         4         5
## 10.002076 12.885112  7.623296 10.857067  6.783657
##         6
##  4.579654
head(ForbesG7$profits)
## [1] 17.85 15.59  6.46 20.96 10.27 10.81
head(model$residuals)
##         1         2         3         4         5
##  7.847924  2.704888 -1.163296 10.102933  3.486343
##         6
##  6.230346
```

2. Wahrscheinlichkeitstheorie

2.1. Kombinatorik

Die **Kombinatorik** ist ein Teilgebiet der Mathematik, das sich mit der Anzahl der möglichen Zusammensetzungen von Mengen beschäftigt. Derartige Aussagen sind für die Wahrscheinlichkeitstheorie von Interesse (vgl. Mathematik I/II).

Modell: Urne mit n gleichen von 1 bis n nummerierten Kugeln

- Man unterscheidet, ob eine Kugel nach dem Ziehen zurückgelegt wird (**Ziehungen mit Zurücklegen**) bzw. ob eine Kugel nicht zurückgelegt wird (**Ziehungen ohne Zurücklegen**).
- Das Ergebnis von k Ziehungen ist eine Zahlenfolge deren Elemente aus den Zahlen $1, ..., n$ stammen.
- Es wird nun weiterhin unterschieden, ob die Reihenfolge der gezogenen Zahlen von Bedeutung ist oder nicht (**Ziehungen mit (ohne) Berücksichtigung der Reihenfolge**).

(a) Ziehungen mit Berücksichtigung der Reihenfolge und mit Zurücklegen

Die Anzahl möglicher Ergebnisse von k Ziehungen ist gleich n^k.

Beispiel: Neukundengewinnung (vgl. F94)
Kunden können auf drei verschiedenen Kanälen (Telefon, Mail, Post) kontaktiert werden. Wie viele Möglichkeiten gibt es, zu 11 verschiedenen Kunden Kontakt aufzunehmen?

$$\text{Anzahl d. Möglichkeiten} = \underbrace{3 \cdot \ldots \cdot 3}_{11\times} = 3^{11} = 177\,147$$

(b) Ziehungen mit Berücksichtigung der Reihenfolge und ohne Zurücklegen

Eine so entstehende Zahlenfolge heißt **Permutation**.

Die Anzahl der möglichen Permutationen ist gleich
$$n(n-1)\cdot\ldots\cdot(n-k+1).$$

Notation: $n!$ heißt die **Fakultät** von $n \in \mathbb{N}$. Es ist $n! = n(n-1)\ldots 1$. Ferner setzt man $0! = 1$.

Folglich ist $n(n-1)\ldots(n-k+1) = \frac{n!}{(n-k)!}$.

Beispiel: Wie viele Möglichkeiten hat ein Supermarkt, aus 10 verschiedenen Produkten 6 auszuwählen und sie in einem Regal nebeneinander anzuordnen?
hier: Reihenfolge wichtig, aber keine Wiederholung
Gesuchte Anzahl $= 10\cdot 9\cdot\ldots\cdot 5 = 151\,200$.

(c) Ziehungen ohne Berücksichtigung der Reihenfolge und ohne Zurücklegen

Eine so entstehende Zahlenfolge heißt **Kombination**. Jeweils alle $k!$ Permutationen (vgl. (b)), die sich aus denselben Zahlen zusammensetzen, bilden eine Kombination.

Die Anzahl der möglichen Kombinationen ist gleich

$$\frac{n!}{k!(n-k)!}.$$

Notation:

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

Diese Größe heißt **Binomialkoeffizient.**

Insbesondere ist (mit $0 \leq k \leq n$)

$$\binom{n}{0} = 1\,, \quad \binom{n}{1} = n\,, \quad \binom{n}{k} = \binom{n}{n-k}$$

Beispiel: Wie viele Möglichkeiten gibt es, aus 10 Aktien ein Portfolio mit 5 Aktien zusammenzustellen?

hier: ohne Reihenfolge, ohne Zurücklegen

Anzahl der Ziehungen:

$$\binom{10}{5} = \frac{10 \cdot 9 \cdot \ldots \cdot 6}{5!} = 252$$

(d) Ziehungen ohne Berücksichtigung der Reihenfolge und mit Zurücklegen

Die Anzahl der möglichen Ergebnisse von k Ziehungen ist gleich

$$\binom{n+k-1}{k}.$$

Beispiel: Wie viele mögliche Abstimmungsergebnisse gibt es im Rahmen einer geheimen Abstimmung des Managements (6 Personen), bei der sich jeder für eine von 3 Investitionsalternativen entscheiden muss? Die Ergebnisse unterscheiden sich nur in der Anzahl an Stimmen, die jede Alternative bekommen hat.
hier: ohne Reihenfolge, mit Zurücklegen

Folglich ist die gesuchte Anzahl gleich $(n = 3, k = 6)$

$$\binom{3+6-1}{6} = \frac{8 \cdot 7}{2!} = 28$$

	ohne Zurücklegen	mit Zurücklegen
ohne Berücksichtigung der Reihenfolge	$\binom{n}{k} = \frac{n!}{k!(n-k)!}$	$\binom{n+k-1}{k}$
mit Berücksichtigung der Reihenfolge	$\frac{n!}{(n-k)!}$	n^k

2.2. Wahrscheinlichkeit von Ereignissen

Ursprünge der Wahrscheinlichkeitstheorie: Jakob Bernoulli (1655-1705), **Pierre-Simon de Laplace** (1749-1827)

Die Wahrscheinlichkeitstheorie entstand aus der Analyse von Glücksspielen.

Ziel: Aussage über Wahrscheinlichkeiten

Beispiel: Wahrscheinlichkeit für

- 6 Richtige im Lotto (reiner Zufallsprozess)
- Niederschlag von mehr als 100 mm im Laufe der nächsten Woche
- Rendite einer Aktie im kommenden Jahr größer als 8 %
- mindestens einen Gewinn in den kommenden beiden Quartalen.

- **Zufallsvorgang:** Geschehen mit ungewissem Ausgang, z.B. Quartalsergebnis
- **Elementarereignis** ω: Ein möglicher Ausgang, z.B. „Gewinn“

  Elementarereignisse schließen sich gegenseitig aus („Gewinn“ *oder* „Verlust“)!
- **Ergebnismenge** Ω: Menge aller ω
- **Ergebnis:** Tatsächlich eingetretenes Elementarereignis
- **Ereignis**: Eine Teilmenge der Grundgesamtheit Ω. Ereignisse müssen sich **nicht** gegenseitig ausschließen!

Beispiel: Ergebnis zweier aufeinanderfolgender Perioden (G, V)

- Elementarereignisse: $\{(G,G)\}, \{(G,V)\}, \{(V,G)\}, \{(V,V)\}$
- Grundgesamtheit: $\Omega = \{(G,G),(G,V),(V,G),(V,V)\}$
- Ereignis: $A = \{(G,G),(G,V),(V,G)\}$, d.h. „in zwei aufeinanderfolgenden Perioden wird mindestens einmal ein Gewinn erwirtschaftet“

Sprech- und Schreibweisen bei Ereignissen

Beschreibung	Bezeichnung	Darstellung in Ω
1. A tritt sicher ein	A ist **sicheres** Ereignis	$A = \Omega$
2. A tritt sicher nicht ein	A ist **unmögliches** Ereignis	$A = \emptyset$
3. wenn A eintritt, tritt B ein	A ist **Teilergebnis** von B	$A \subset B$
4. genau dann, wenn A eintritt, tritt B ein	A und B sind **äquivalente** Ereignisse	$A = B$
5. wenn A eintritt, tritt B nicht ein	A und B sind **disjunkte** Ereignisse	$A \cap B = \emptyset$
6. genau dann, wenn A eintritt, tritt B nicht ein	A und B sind **komplementäre** Ereignisse	$B = \bar{A}$
7. genau dann, wenn **mindestens ein** A_j eintritt (auch: genau dann, wenn A_1 **oder** A_2 oder ... eintritt), tritt A ein	A ist **Vereinigung** der A_j	$A = \bigcup_j A_j$
8. genau dann, wenn **alle** A_j eintreten (auch: genau dann, wenn A_1 **und** A_2 **und** ... eintreten), tritt A ein	A ist **Durchschnitt** der A_j	$A = \bigcap_j A_j$

Beispiel: Quartalsergebnis

- $A = \{(G, G), (G, V)\}$: Das Ereignis, dass in der ersten Periode ein Gewinn erwirtschaftet wird.
- $B = \{(G, G), (V, G)\}$: Das Ereignis, dass in der zweiten Perode ein Gewinn erwirtschaftet wird.

Mengendarstellung	Ereignis, dass
$A \cup B$	mindestens ein Gewinn auftritt
$A \cap B$	zweimal ein Gewinn auftritt
$\bar{A} \cup \bar{B}$	mindestens ein Verlust auftritt
$\bar{A} \cap \bar{B}$	zweimal ein Verlust auftritt
$(A \cap \bar{B}) \cup (\bar{A} \cap B)$	sowohl Gewinn als auch Verlust auftreten
$(A \cap B) \cup (\bar{A} \cap \bar{B})$	zweimal hintereinander dasselbe Elementarereignis auftritt

Gegeben: Menge (Ereignis) A

Ziel: Was ist $P(A)$ (P für probability)?

Die Laplace-Wahrscheinlichkeit

Ausgangspunkt: Alle Elementarereignisse besitzen die gleiche Wahrscheinlichkeit.

Ist Ω endlich, so gilt

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

wobei $|A|$ die Anzahl der Elemente von A bezeichne, analog $|\Omega|$.

Beispiel: Lebensdauer eines Bauteils in Monaten ($\Omega = \{1, .., 36\}$)

$A =$ Lebensdauer größer 9 Monate
$B =$ Lebensdauer zw. 5 und 15 Monaten
Es wird angenommen, dass $P\big(\{1\}\big) = P\big(\{2\}\big) = \cdots = P\big(\{36\}\big) = 1/36$, d.h. es liegt ein Laplace–Experiment vor. Somit ist

$$P(A) = \frac{|A|}{|\Omega|} = \frac{27}{36} \, .$$

Die Wahrscheinlichkeit, dass eine die Lebensdauer größer als 9 ist, aber nicht zwischen 5 und 15 liegt, ist

$$P(A \cap \bar{B}) = \frac{|\{16, 17, ..., 36\}|}{36} = \frac{21}{36} \, .$$

Die statistische Wahrscheinlichkeit

Es sei $A \subset \Omega$. Das Zufallsexperiment wird n–mal wiederholt. $h_n(A)$ bezeichne die relative Häufigkeit von A.

Beispiel: Lebensdauer ($\Omega = \{1, 2, \ldots, 36\}$)
A sei das Ereignis, dass die Lebensdauer kleiner gleich 12 ist, d.h. $A = \{1, 2, .., 12\}$.

Bei 16 Bauteilen wird beobachtet, nach wie vielen Monaten sie ausfallen:

23	34	13	11	28	9	8	21
16	33	31	15	3	13	23	32

Dann ist $h_{16}(A) = \frac{4}{16} = 0.25$.

Beispiel: Eine faire Münze, deren beiden Seiten einen Kopf (K) und eine Zahl (Z) zeigen, werde n–mal hintereinander geworfen. Man erhält:

n	$n(K)$	$h_n(K)$
10	7	0.700
20	11	0.550
100	47	0.470
400	204	0.510
1000	492	0.492
2000	1010	0.505

Die relativen Häufigkeiten streben gegen die exakte Wahrscheinlichkeit, hier 0.5.

Richard von Mises (1931)

Wahrscheinlichkeit für das Eintreten des Ereignisses A:

$$P(A) := \lim_{n\to\infty} h_n(A)$$

Nachteil: nicht praktikabel

Axiomatik der Wahrscheinlichkeitstheorie

- Ansätze über Laplace-Wahrscheinlichkeit und statistische Wahrscheinlichkeit haben Vor–und Nachteile
- Allgemeine axiomatische Definition der Wahrscheinlichkeit von Kolmogorov (1933)

A. N. Kolmogorov (1933)

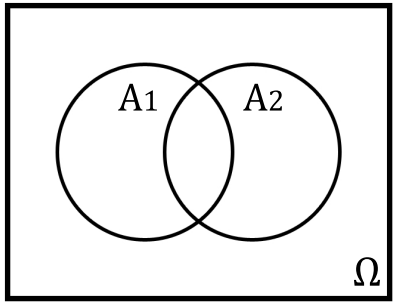
Unter einem **Wahrscheinlichkeitsmaß** P versteht man eine Abbildung, die (nahezu!) allen Ereignissen $A \subseteq \Omega$ eine Zahl zuordnet (nämlich $P(A)$) und die die folgenden Bedingungen erfüllt:

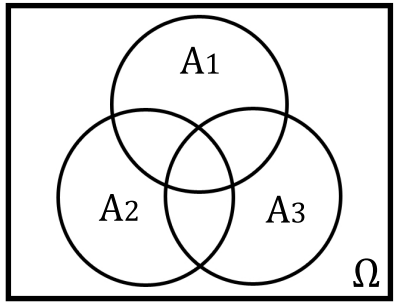
- $0 \leq P(A) \leq 1$
- $P(\Omega) = 1$
- $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ für alle $A_i \subset \Omega$ mit $A_i \cap A_j = \emptyset$ für $i \neq j$.

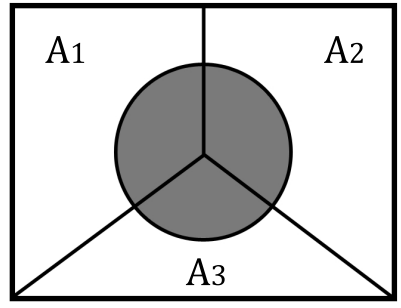
$P(A)$ heißt die **Wahrscheinlichkeit des Ereignisses** A.

Rechenregeln für Wahrscheinlichkeiten

- $P(A) \leq 1$
- $P(\emptyset) = 0$
- $A \subset B \Rightarrow P(A) \leq P(B)$
- $P(\bar{A}) = 1 - P(A)$
- $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$
  $P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3)$
  $- P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3)$
  $+ P(A_1 \cap A_2 \cap A_3)$
- $P(B) = \sum_i P(B \cap A_i)$, falls $A_i \cap A_j = \emptyset \; \forall \, i \neq j, \bigcup_i A_i = \Omega$







A1∪A2∪A3=Ω ●=B

Beispiel I : Anzahl defekter Teile einer Warensendung mit $P\big(\{0\}\big) = 0.1$, $P\big(\{1\}\big) = P\big(\{2\}\big) = P\big(\{3\}\big) = P\big(\{4\}\big) = 0.15$, $P\big(\{5\}\big) = 0.3$.

Wahrscheinlichkeit für drei oder weniger defekte Teile:

$$P\big(\{0,1,2,3\}\big) = P\big(\{0\}\big) + P\big(\{1\}\big) + P\big(\{2\}\big) + P\big(\{3\}\big) = 0.55.$$

Ferner ist:

$$\begin{aligned} P\big(\{4,5\}\big) &= 1 - 0.55 = 0.45, \\ P\big(\{0,1,2\}\big) &= 0.1 + 2 \cdot 0.15 = 0.4, \\ P\big(\{3,4,5\}\big) &= 0.6. \end{aligned}$$

Beispiel II: Zweimaliges Werfen eines fairen Würfels, mit $P\big(\{1\}\big) = P\big(\{2\}\big) = \ldots = P\big(\{6\}\big) = \frac{1}{6}$.

$$\begin{aligned} P(C) &= P(\text{„Augensumme} \leq 10\text{“}) \\ &= 1 - P(\text{„Augensumme} \geq 11\text{“}) = 1 - P(\bar{C}) \\ &= 1 - P(\{(5,6),\ (6,5),\ (6,6)\}) \\ &= 1 - \tfrac{3}{36} = \tfrac{11}{12} = 0.917 \end{aligned}$$

2.3. Bedingte W'keit und Unabhängigkeit

bisher: bedingte relative Häufigkeit für die Ausprägung b_j von Y unter der Bedingung, dass X die Ausprägung a_i besitzt

$$h(Y = b_j \mid X = a_i) = \frac{h(X = a_i, Y = b_j)}{h(X = a_i)}$$

jetzt: W'keit von A hängt von anderem Ereignis B ab.

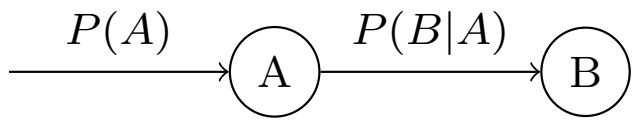
Beispiel: W'keit für eine Statistiknote hängt von der Mathenote ab.

Bedingte Wahrscheinlichkeit für ein Ereignis A unter der Bedingung B (für A, wenn B vorliegt)

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \quad \text{für} \quad P(B) > 0$$

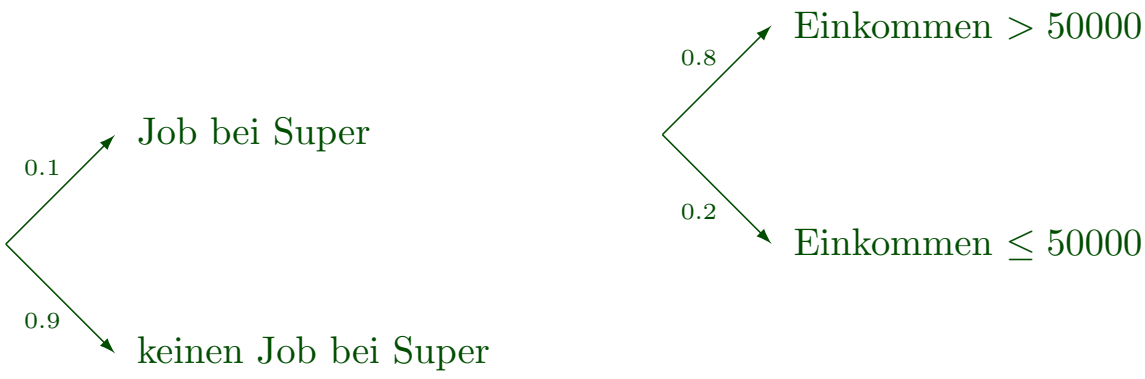
Beachte: $P(A \mid B) + P(\bar{A} \mid B) = 1$,
aber: $P(A \mid B) + P(A \mid \bar{B}) \neq 1$

Baumdiagramm:



Beispiel: Jobsuche und Einkommen
A: Job bei Firma „Super“
B: Einkommen größer als 50 000 Euro

$P(A) = 0.1,\ P(B \mid A) = 0.8$

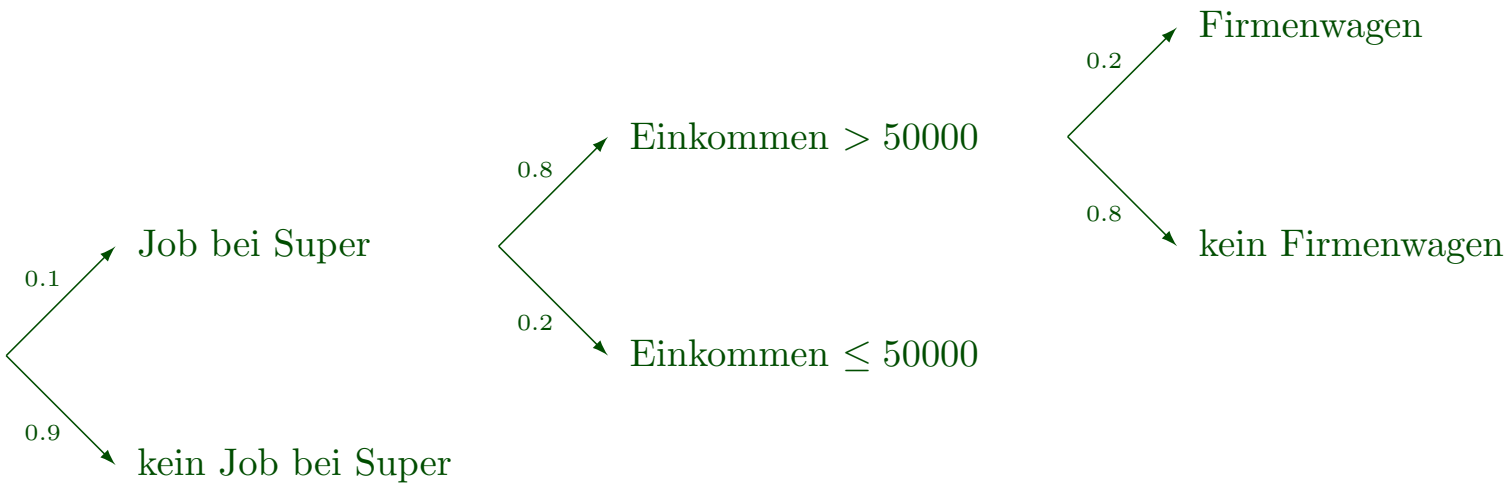


$$\begin{aligned} \text{Man erhält} \quad P(A \cap B) &= P(B|A) \cdot P(A) = 0.8 \cdot 0.1 = 0.08, \\ P(A \cap \bar{B}) &= P(\bar{B}|A) \cdot P(A) = 0.2 \cdot 0.1 = 0.02. \end{aligned}$$

Ziel: Verallgemeinerung auf mehrere Ereignisse, d.h.
$P(A_1 \cap A_2 \cdots \cap A_k)$

Beispiel: A: Job bei „Super“ B : Einkommen > 50000
C: Firmenwagen nach 3 Jahren

$P(A) = 0.1,\ P(B \mid A) = 0.8,\ P(C \mid A \cap B) = 0.2$



Man erhält

$$\begin{aligned} P(A \cap B \cap C) &= \underbrace{P(A \cap B)}_{=P(A)\cdot P(B \mid A)} \cdot P(C \mid A \cap B) \\ &= 0.1 \cdot 0.8 \cdot 0.2 = 0.016\,. \end{aligned}$$

Multiplikationssatz

Es seien $A_1, \ldots, A_k$ Ereignisse mit $P(A_1 \cap \cdots \cap A_{k-1}) > 0$. Dann gilt für $k \geq 2$

$$\begin{aligned} &P(A_1 \cap \cdots \cap A_k) \\ &= P(A_1) \cdot P(A_2 \mid A_1) \cdot P(A_3 \mid A_1 \cap A_2) \\ &\quad \cdot \ldots \cdot P(A_k \mid A_1 \cap \cdots \cap A_{k-1}). \end{aligned}$$

2.3.1. Der Satz von der totalen Wahrscheinlichkeit

Ziel: Nicht-bedingte W'keit aus bedingten Wahrscheinlichkeiten erschließen

Satz von der totalen Wahrscheinlichkeit

Es seien $A_1, \ldots, A_k$ paarweise disjunkte Ereignisse mit $A_1 \cup \cdots \cup A_k = \Omega$. Dann gilt für ein beliebiges Ereignis B

$$P(B) = \sum_{i=1}^{k} P(B \cap A_i) = \sum_{i=1}^{k} P(B \mid A_i) \cdot P(A_i)$$

Beispiel: Ausschuss 1

Drei Maschinen fertigen ein Produkt. Die erste Maschine fertigt 20% der gesamten Produktion und die zweite und dritte Maschine produzieren jeweils 40%. Aus Erfahrung ist bekannt, dass die erste Maschine 5%, die zweite 10% und die dritte 20% Ausschuß produziert.

Wie groß ist die W.keit, dass ein zufällig ausgewähltes Produkt defekt ist?

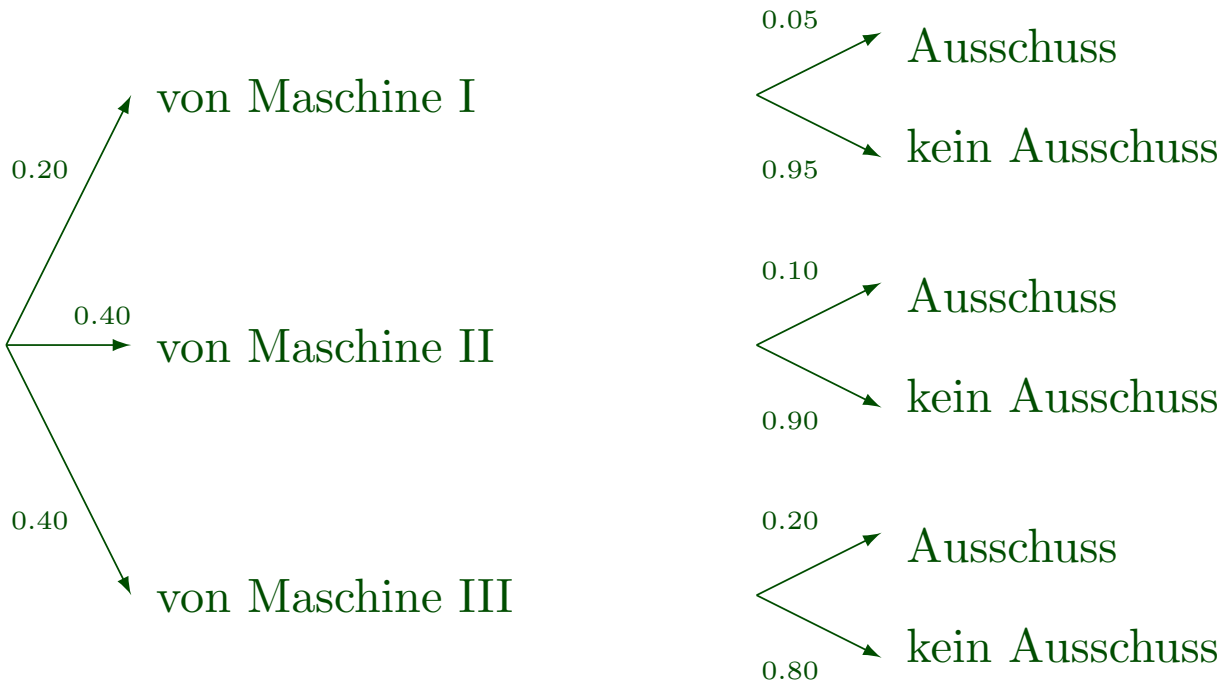
B : "Ausschussprodukt"

A_1 : "Produkt kommt von Maschine I "

A_2 : "Produkt kommt von Maschine II"

A_3 : "Produkt kommt von Maschine III"

$$P(A_1) = 0.2,\ P(A_2) = 0.4,\ P(A_3) = 0.4,$$
$$P(B|A_1) = 0.05,\ P(B|A_2) = 0.1,\ P(B|A_3) = 0.2.$$



Der Satz von der totalen Wahrscheinlichkeit liefert:

$$P(B) = \sum_{i=1}^{3} P(B|A_i)P(A_i) = 0.05 \cdot 0.2 + 0.1 \cdot 0.4 + 0.2 \cdot 0.4 = 0.13$$

2.3.2. Der Satz von Bayes

Ziel: Umkehrung von Argument und Bedingung

Satz von Bayes (1702 – 1761)

Es seien $A_1, \ldots, A_k$ paarweise disjunkte Ereignisse mit $A_1 \cup \cdots \cup A_k = \Omega$. Ferner sei B ein beliebiges Ereignis. Dann gilt für $i \in \{1, \ldots, k\}$

$$P(A_i \mid B) = \frac{P(B \mid A_i) \cdot P(A_i)}{P(B)} = \frac{P(B \mid A_i) \cdot P(A_i)}{\sum_{j=1}^{k} P(B \mid A_j) \cdot P(A_j)}.$$

Beispiel: Ausschuss 2

Bei der zufälliger Entnahme **eines** Produktes erhält man **ein** Ausschußstück. Wie groß ist die Wahrscheinlichkeit, dass das gezogene Produkt von Maschine I kommt?

Mit dem Satz von Bayes erhält man

$$\begin{aligned} P(A_1|B) &= \frac{P(B|A_1)P(B_1)}{P(B)} \\ &= \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3)} \\ &= \frac{0.05 \cdot 0.2}{0.05 \cdot 0.2 + 0.1 \cdot 0.4 + 0.2 \cdot 0.4} = \frac{0.01}{0.13} = 0.077 \end{aligned}$$

Beispiel: Ausschuss 3

Bei der zufälliger Entnahme **eines** Produktes erhält man **kein** Ausschußstück. Wie groß ist die Wahrscheinlichkeit, dass das gezogene Produkt von Maschine I kommt?

Mit dem Satz von Bayes erhält man

$$
\begin{aligned}
P(A_1|\bar{B}) &= \frac{P(\bar{B}|A_1)P(A_1)}{P(\bar{B})} \\
&= \frac{P(\bar{B}|A_1)P(A_1)}{P(\bar{B}|A_1)P(A_1) + P(\bar{B}|A_2)P(A_2) + P(\bar{B}|A_3)P(A_3)} \\
&= \frac{(1-0.05)\cdot 0.2}{(1-0.05)\cdot 0.2 + (1-0.1)\cdot 0.4 + (1-0.2)\cdot 0.4} \\
&= \frac{0.19}{0.87} = 0.218
\end{aligned}
$$

2.3.3. Unabhängigkeit von Ereignissen

- A, B unabhängig: A liefert keine Information über $P(B)$ und umgekehrt
- Formal:

$$P(B|A) = P(B) \qquad \text{bzw.} \qquad P(A|B) = P(A)$$

A und B sind unabhängig, wenn

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

- Dann gilt:

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

Beispiel: Für einen Produkttestreihe stehen 5 Personen zur Auswahl, 3 männliche und 2 weibliche.

A_i: für das $i$te Produkt wird eine weibliche Person ausgewählt

- **Eine Person darf Tester für mehr als ein Produkt sein (Ziehung mit Zurücklegen)**: A_1 und A_2 sind unabhängig, denn

$$P(A_1 \cap A_2) = P(A_1) \cdot P(A_2 \mid A_1) = \frac{2}{5} \cdot \frac{2}{5}$$
$$= \frac{4}{25} = P(A_1) \cdot P(A_2)\,,$$

- **Höchstens ein Produkt pro Person (Ziehung ohne Zurücklegen)**: A_1 und A_2 sind abhängig, denn

$$\begin{aligned} P(A_1 \cap A_2) &= P(A_1) \cdot P(A_2 \mid A_1) = \frac{2}{5} \cdot \frac{1}{4} = \frac{1}{10}\,, \\ P(A_1) \cdot P(A_2) &= \frac{2}{5} \cdot \left[P(A_1) \cdot P(A_2 \mid A_1) + P(\bar{A}_1) \cdot P(A_2 \mid \bar{A}_1)\right] \\ &= \frac{2}{5} \cdot \left[\frac{2}{5} \cdot \frac{1}{4} + \frac{3}{5} \cdot \frac{2}{4}\right] = \frac{2}{5} \cdot \frac{2}{5} = \frac{4}{25} \neq \frac{1}{10}\,. \end{aligned}$$

Unabhängigkeit von Ereignissen ($n \geq 2$)

Allgemein heißen $n \geq 2$ Ereignisse $A_1, \ldots, A_n$ **(stochastisch) unabhängig**, falls für alle Teilmengen $\{i_1, \ldots, i_r\}$ von $\{1, \ldots, n\}$ gilt :

$$P(A_{i_1} \cap \cdots \cap A_{i_r}) = P(A_{i_1}) \cdot \ldots \cdot P(A_{i_r}) \,.$$

Beispiel: Für $n = 3$ muss gelten:

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= P(A_1) \cdot P(A_2) \cdot P(A_3) \\ P(A_1 \cap A_2) &= P(A_1) \cdot P(A_2) \\ P(A_1 \cap A_3) &= P(A_1) \cdot P(A_3) \\ P(A_2 \cap A_3) &= P(A_2) \cdot P(A_3) \end{aligned}$$

2.4. Zufallsvariablen und Verteilungen

Zufallsvariablen

- Beschreibung von Ereignissen durch reelle Zahlen
- Formal: die **Zufallsvariable** X ist eine Abbildung von der Grundgesamtheit Ω in einen Bildraum S

$$X : \Omega \to S.$$

- Im allgemeinen ist $S \subset \mathbb{R}$ oder $S \subset \mathbb{R}^n$ (Zufallsvektor).
- Nach Durchführung des Zufallsvorgangs:

$$\textbf{Realisation:} \qquad x = X(\omega)$$

- Vor Durchführung des Zufallsvorgangs:

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

- **Beispiel**: Bauteil, X: Lebensdauer, $X(\Omega) = \{1, 2, \ldots, 36\}, \ \ x = 4$ (z.B.) $P(X = 4) = \frac{1}{36}, \quad P(X \leq 3) = \frac{3}{36} = \frac{1}{12}$

Verteilungsfunktion

Ziel: Zuweisung von Wahrscheinlichkeiten zu Realisationen

Die **Verteilungsfunktion F_X der ZV X** ist gegeben durch

$$F_X(x) = P(\{\omega : X(\omega) \leq x\})$$

- Man schreibt dafür häufig nur kurz $F(x) = P(X \leq x)$.
- Wir verwenden das Symbol $X \sim F_X$.

Eigenschaften:

- $F(x) \in [0;1]$
- Definitionsbereich $= \mathbb{R}$ mit:
  $F(-\infty) = \lim_{x\to-\infty} F(x) = 0, \quad F(\infty) = \lim_{x\to\infty} F(x) = 1$
- monoton wachsend: $x_1 < x_2 \Rightarrow F(x_1) \leq F(x_2)$
- F ist rechtsseitig stetig: $\lim_{\varepsilon\to 0} F(x+\varepsilon) = F(x),$ für $\varepsilon > 0$

Beachte I: Jede Funktion F, die die obigen Eigenschaften erfüllt, heisst eine Verteilungsfunktion.

Beachte II: Ist eine Funktion gegeben, die obige Eigenschaften erfüllt, so kann man eine Zufallsvariable konstruieren, sodass die Verteilungsfunktion dieser Zufallsvariablen gleich der vorgegebenen Funktion ist.

Berechnung von Wahrscheinlichkeiten

Die Verteilungsfunktion enthält sämtliche für den Statistiker relevante Information. Mit ihr können die Wahrscheinlichkeiten für alle Ereignisse berechnet werden, die die Zufallsvariable betreffen.

Es gilt für $a < b$:

- $P(a < X \leq b) = F(b) - F(a)$
- $P(a \leq X \leq b) = F(b) - F(a-)$
- $P(X > a) = 1 - P(X \leq a) = 1 - F(a)$
- $P(X \geq a) = 1 - P(X < a) = 1 - F(a-)$.

Dabei bezeichnet $F(a-)$ den linksseitigen Grenzwert von F gegen a, d.h. $F(a-) := \lim_{\varepsilon \to 0} F(a - \varepsilon)$, wobei $\varepsilon > 0$ ist.

2.4.1. Diskrete Zufallsvariablen

- Nimmt X höchstens abzählbar viele verschiedene Werte an, so heißt X eine **diskrete Zufallsvariable**, d.h. $X(\Omega) = \{x_1, x_2, \dots\}$. F_X heißt dann eine **diskrete Verteilungsfunktion**.
- $P(X = x_i) = p_i$ mit $\sum_i p_i = 1$

Die Funktion

$$f(x) = P(X = x) = \begin{cases} p_i, & \text{falls } x = x_i \\ 0, & \text{sonst} \end{cases}$$

heißt **Wahrscheinlichkeitsfunktion** von X.

- Es sei $x_1 < x_2 < \ldots$ und $x_i \leq x < x_{i+1}$, dann gilt für die Verteilungsfunktion:

$$\begin{aligned} F_X(x) &= P(X \leq x) = \sum_{x_v \leq x} f(x_v) \\ &= \sum_{v=1}^{i} f(x_v) = P(X = x_1) + \cdots + P(X = x_i)\,. \end{aligned}$$

Beispiel: Roulettespiel

Beim Roulette gibt es insgesamt 37 Zahlen $(0 - 36)$. Man setzt auf einzelne Zahlen oder bestimmte Eigenschaften von Zahlen. Die Gewinnquote (das Verhältnis von Nettogewinn und Einsatz) variiert je nach Wette. Wir setzen 100 € auf die Null (Gewinnquote $35:1$) und 100 € auf das Eintreten einer geraden Zahl (Gewinnquote $1:1$). Der Gewinn (= Auszahlung - Einsatz (hier 200)) ist gegeben durch

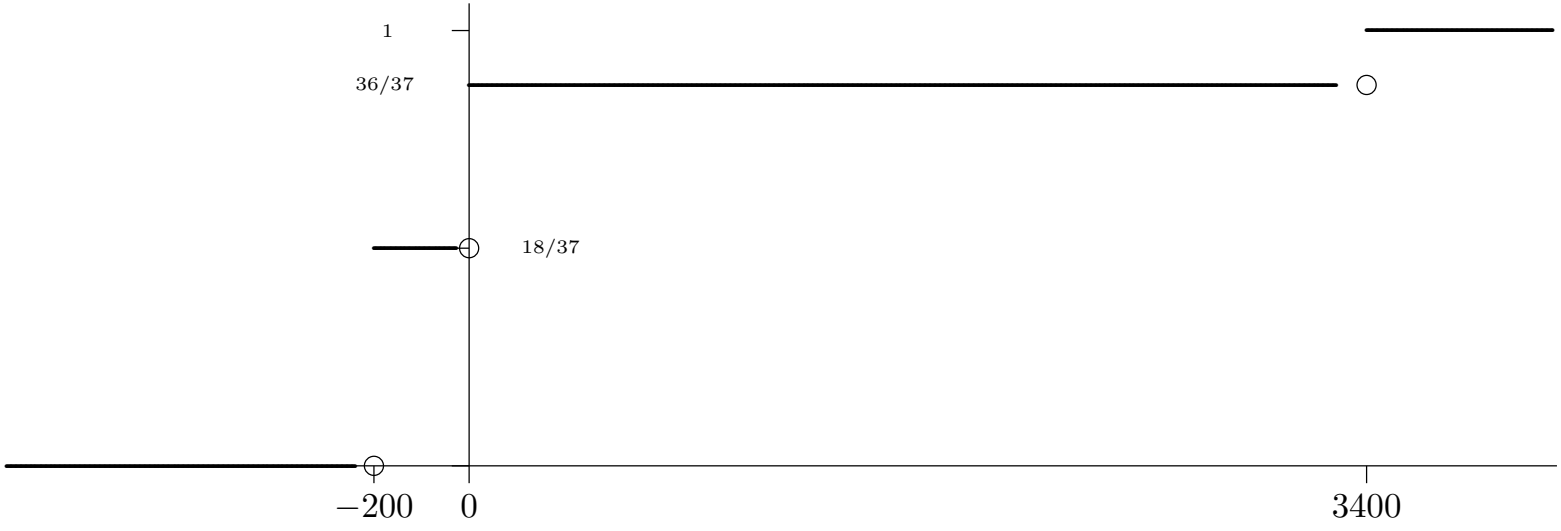
$$X = \begin{cases} 0 & \text{falls Gewinnzahl } i \text{ gerade} \\ -200 & \text{falls Gewinnzahl } i \text{ ungerade} \\ 3400 & \text{falls Gewinnzahl } i = 0 \end{cases}.$$

Es gilt: $P(X = 0) = \frac{18}{37}$, $P(X = -200) = \frac{18}{37}$ und $P(X = 3400) = \frac{1}{37}$. Folglich ist

$$F_X(x) = P(X \leq x) = \begin{cases} 0 & \text{falls } x < -200 \\ \frac{18}{37} & \text{falls } -200 \leq x < 0 \\ \frac{36}{37} & \text{falls } 0 \leq x < 3400 \\ 1 & \text{falls } x \geq 3400 \end{cases}.$$

Bild zu $P(X = x)$:

Schaubild der Verteilungsfunktion:



Weitere Beispiele für bekannte diskrete Verteilungen

- **Binomialverteilung**
- **Hypergeometrische Verteilung**
- **Poisson-Verteilung**

Binomialverteilung

- Ein Zufallsvorgang wird n-mal durchgeführt. Die W'keit des Eintreten des Ereignisses A pro Durchführung sei $P(A) = p$ (Ziehen mit Zurücklegen)
- Schreibe:

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

- Dann gibt

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

  an, wie oft A bei n Zufallsvorgängen eintritt.
- **Ziel:** Wahrscheinlichkeitsfunktion von X, d.h. wie groß ist die W'keit, dass A k-mal bei n Zufallsvorgängen eintritt.

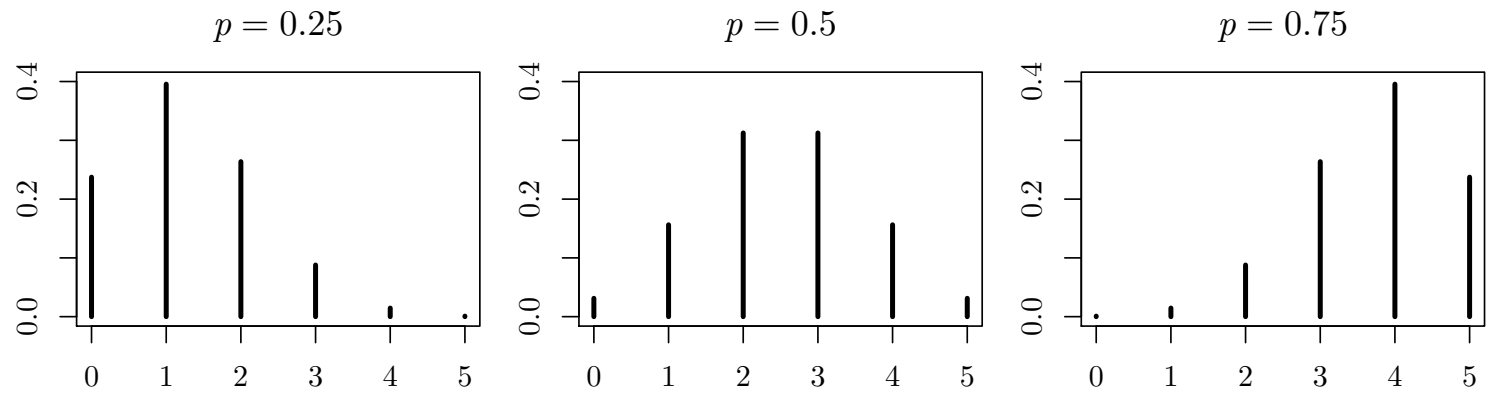
- Herleitung:
  - $P(Z_i = 1) = P(A) = p$, $P(Z_i = 0) = P(\bar{A}) = 1 - p$
  - $\sum_{i=1}^{n} z_i = x$ entspricht „x mal Ereignis A und $n - x$ mal $\bar{A}$“
    Wahrscheinlichkeit (bei Unabhängigkeit): $p^x \cdot (1-p)^{n-x}$
  - Aber: Reihenfolge irrelevant! Anzahl Anordnungen: $\binom{n}{x}$

Wahrscheinlichkeitsfunktion der Binomialverteilung

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

- Kurzschreibweise: $X \sim B(n; p)$
- $F(x)$ bestimmt man mithilfe der allgemeinen Vorgehensweise für diskrete Verteilungen (d.h. $F(x) = \sum_{x_i \leq x} f(x_i)$) oder aus Tabellen; für $f(x)$ gilt: $f(x) = F(x) - F(x-1)$

Wahrscheinlichkeitsfunktion der Binomialverteilung für $n = 5$




```
x <- 0:5
# Wahrscheinlichkeitsfunktion P(X=x): dbinom
(p.25 <- dbinom(x = x, size = 5, p = 0.25))
## [1] 0.2373047 0.3955078 0.2636719 0.0878906 0.0146484
## [6] 0.0009766
cumsum(p.25)
## [1] 0.2373 0.6328 0.8965 0.9844 0.9990 1.0000
# Verteilungsfunktion F(x) = P(X<=x): pbinom
(F.25 <- pbinom(q = x, size = 5, p = 0.25))
## [1] 0.2373 0.6328 0.8965 0.9844 0.9990 1.0000
```

Verteilungsfunktion der Binomialverteilung: $F(x) = P(X \leq x)$, mit $X \sim B(n; p)$ und $p = 0.25$

p = 0.25

x \ 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	0.4449	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	0.6785	0.6007	0.5256	0.4552	0.3907	0.3326	0.2811	0.2361
3	.	.	1.0000	0.9961	0.9844	0.9624	0.9294	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	0.9857	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	0.9978	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 \ 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	0.1113	0.0913	0.0745	0.0606	0.0492	0.0398	0.0321	0.0258	0.0207	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	0.0374
4	0.6302	0.5739	0.5187	0.4654	0.4148	0.3674	0.3235	0.2832	0.2466	0.2137	0.1844	0.1583	0.1354	0.1153	0.0979
5	0.8103	0.7653	0.7175	0.6678	0.6172	0.5666	0.5168	0.4685	0.4222	0.3783	0.3371	0.2989	0.2638	0.2317	0.2026
6	0.9204	0.8929	0.8610	0.8251	0.7858	0.7436	0.6994	0.6537	0.6074	0.5611	0.5154	0.4708	0.4279	0.3868	0.3481
7	0.9729	0.9598	0.9431	0.9225	0.8982	0.8701	0.8385	0.8037	0.7662	0.7265	0.6852	0.6427	0.5997	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	0.8195	0.7859	0.7501	0.7125	0.6736
9	0.9984	0.9969	0.9946	0.9911	0.9861	0.9794	0.9705	0.9592	0.9453	0.9287	0.9091	0.8867	0.8615	0.8337	0.8034
10	0.9997	0.9994	0.9988	0.9977	0.9961	0.9936	0.9900	0.9851	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	0.9493
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	0.9999	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	0.9999
18	.	.	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.	

Beispiel: Karten

Aus einem 32er Kartenblatt werden drei Karten mit Zurücklegen gezogen. Wie wahrscheinlich ist es, zweimal „Herz“ zu ziehen?

$$
\begin{aligned}
X_i &= \begin{cases} 1, \text{ falls } i\text{-te Karte Herz} \\ 0, \text{ sonst} \end{cases} \\
X &= \sum_{i=1}^{n} X_i = X_1 + X_2 + X_3 \\
X &\sim B(3; \tfrac{1}{4})
\end{aligned}
$$

Mithilfe der Wahrscheinlichkeitsfunktion:

$$P(X = 2) = f(2) = \binom{3}{2} \cdot 0.25^2 \cdot 0.75^1 = 0.1406$$

Mithilfe von Tabelle:

$$P(X = 2) = F(2) - F(1) = 0.9844 - 0.8438 = 0.1406$$

```
dbinom(x = 2, size = 3, p = 0.25)
## [1] 0.1406
pbinom(2, 3, 0.25) - pbinom(1, 3, 0.25)
## [1] 0.1406
```

Beispiel: Kreditausfälle

Aus langjähriger Erfahrung ist bekannt, dass ein Kredit mit einer W'keit von 0.1 ausfällt. Wie groß ist die W'keit, dass genau 48 von 50 Krediten nicht ausfallen?

$$
\begin{aligned}
P(X = 48) &= \binom{50}{48} \cdot 0.9^{48} \cdot 0.1^2 \\
&= 49 \cdot 25 \cdot 0.9^{48} \cdot 0.1^2 \\
&\approx 0.078
\end{aligned}
$$

```
dbinom(x = 48, size = 50, p = 0.9)
## [1] 0.07794
```

Anmerkungen zur Binomialverteilung

- In Tabelle nur $p \leq 0.5$
- Falls $p > 0.5$:

$$F_{B(n;p)}(x) = 1 - F_{B(n;1-p)}(n - x - 1)$$

- **Beispiel**: $X \sim B(20; \frac{3}{4})$

$$\begin{aligned} P(X \leq 10) &= F_{B(20;\frac{3}{4})}(10) = 1 - F_{B(20;\frac{1}{4})}(20 - 10 - 1) \\ &= 1 - F_{B(20;\frac{1}{4})}(9) = 1 - 0.9861 = 0.0139 \end{aligned}$$

- Zusammenhang mit Urnenmodell:
  - N Objekte, davon M besonders markiert.
  - Ziehe n Objekte **mit** Zurücklegen.
  - X: Anzahl gezogener Objekte mit Markierung $\Rightarrow X \sim B(n; \frac{M}{N})$

Hypergeometrische Verteilung

- n-faches Ziehen **ohne** Zurücklegen aus N Objekten, davon M markiert (und $N - M$ nicht markiert).

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

heißt **hypergeometrisch verteilt** mit den Parametern N, M, n.

Wahrscheinlichkeitsfunktion der hypergeometrischen Verteilung

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

- Kurzschreibweise: $X \sim H(n, M, N)$
- **Beachte:** Approximation durch $B(n; \frac{M}{N})$ möglich falls $n \leq \frac{N}{20}$.

Beispiel: Versicherungsvertreter

Ein Versicherungsvertreter konnte in 32 Gesprächen mit potenziellen Neukunden 8 Versicherungen abschließen.Wie wahrscheinlich ist es, dass, wenn von den potenziellen Neukunden 3 zufällig ausgewählt werden, 2 eine Versicherung abgeschlossen haben?

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 wurden verwendet:

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

```
dhyper(x = 2, m = 8, n = 24, k = 3)
## [1] 0.1355
```

Poisson-Verteilung

- Häufig ist bei der Binomialverteilung n groß und p klein (z.B. Anzahl 6er pro Lottoausspielung) $\rightsquigarrow$ man untersucht die Verteilung seltener Ereignisse
-

Wahrscheinlichkeitsfunktion der Poisson-Verteilung

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

- Kurzschreibweise: $X \sim P(\lambda)$
- $F(x)$ ist tabelliert, bzw. Software
- Approximation:

$$B(n; p) \approx P(np)$$

Voraussetzungen:

$$p \text{ klein } (\leq 0,1), \quad n \text{ groß } (\geq 50) \text{ und } np \leq 10$$

Verteilungsfunktion der Poisson–Verteilung: $F(x) = P(X \leq x)$, mit $X \sim P(\lambda)$

x \ λ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5
0	0.9048	0.8187	0.7408	0.6703	0.6065	0.5488	0.4966	0.4493	0.4066	0.3679	0.3329	0.3012	0.2725	0.2466	0.2231
1	0.9953	0.9825	0.9631	0.9384	0.9098	0.8781	0.8442	0.8088	0.7725	0.7358	0.6990	0.6626	0.6268	0.5918	0.5578
2	0.9998	0.9989	0.9964	0.9921	0.9856	0.9769	0.9659	0.9526	0.9371	0.9197	0.9004	0.8795	0.8571	0.8335	0.8088
3	1.0000	0.9999	0.9997	0.9992	0.9982	0.9966	0.9942	0.9909	0.9865	0.9810	0.9743	0.9662	0.9569	0.9463	0.9344
4	1.0000	1.0000	1.0000	0.9999	0.9998	0.9996	0.9992	0.9986	0.9977	0.9963	0.9946	0.9923	0.9893	0.9857	0.9814
5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9997	0.9994	0.9990	0.9985	0.9978	0.9968	0.9955
6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9997	0.9996	0.9994	0.9991
7	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
8	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 \ λ	1.6	1.7	1.8	1.9	2.0	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0
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	0.4932	0.4628	0.4337	0.4060	0.3796	0.3546	0.3309	0.3084	0.2873	0.2674	0.2487	0.2311	0.2146	0.1991
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	0.8386	0.8194	0.7993	0.7787	0.7576	0.7360	0.7141	0.6919	0.6696	0.6472
4	0.9763	0.9704	0.9636	0.9559	0.9473	0.9379	0.9275	0.9162	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	0.9969	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	0.9991	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	0.9999	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

Beispiel: Kundendatenbank mit 10000 Einträgen, W'keit für einen Fehler = 0.03%
$X \sim B(10000; 0.0003)$ ist nicht vertafelt!

$$\left.\begin{array}{rl} p = 0.0003 & < 0.1 \\ n = 10000 & > 50 \\ np = 3 & < 10 \end{array}\right\} \Rightarrow B(10000; 0.0003) \approx P(3)$$

Mithilfe der Wahrscheinlichkeitsfunktion:

$$P(X = 5) = \frac{3^5}{5!} \cdot e^{-3} = 0.1008188$$

Mithilfe von Tabelle für Binomialverteilung:

$$P(X = 5) = F(5) - F(4) = 0.9161 - 0.8153 = 0.1008$$

Exakter Wert:

$$P(X = 5) = \binom{10000}{5} \cdot 0.0003^5 \cdot 0.9997^{9995} = 0.1008239$$

```
dpois(x = 5, lambda = 3)
## [1] 0.1008188
dbinom(5, 10000, 3e-04)
## [1] 0.1008239
```

2.4.2. Stetige Zufallsvariablen

Die ZV X heißt **stetig**, wenn die VF $F(x)$ eine stetige Funktion ist. Dann existiert eine nicht-negative Funktion $f(x)$ (d.h. $f(x) \geq 0$ für alle $x \in \mathbb{R}$) mit

$$F(x) = \int_{-\infty}^{x} f(t)dt \quad \text{für alle} \quad x \in \mathbb{R}.$$

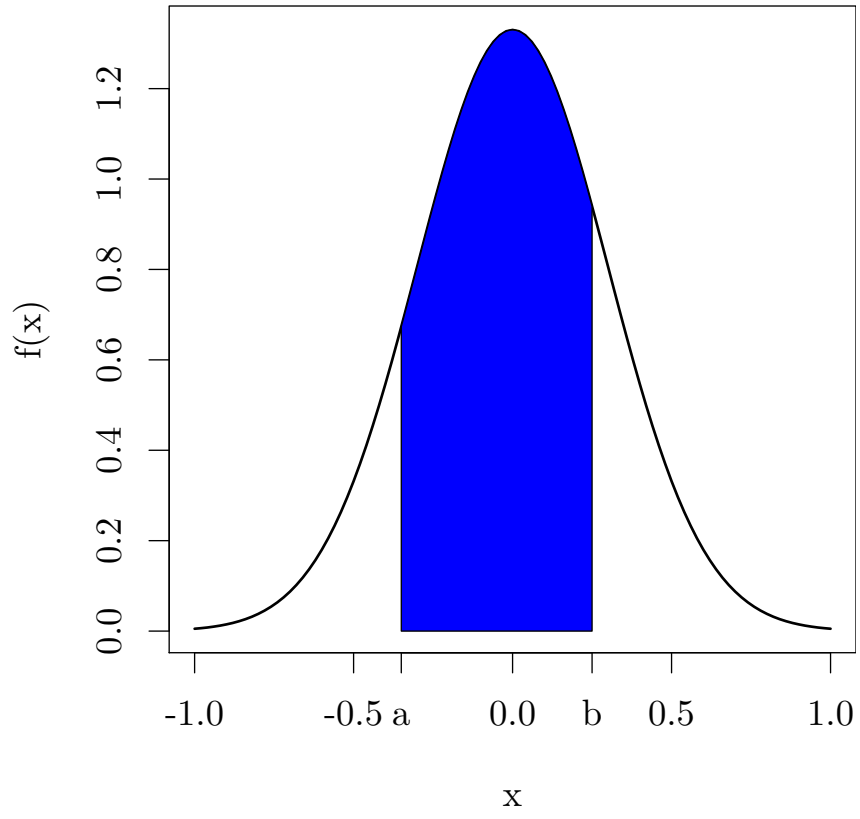
Die Funktion $f(x)$ heißt die **Dichtefunktion** (oder kurz **Dichte**) der ZV X.

Es gilt:

- $f(x) \geq 0$ für alle $x \in \mathbb{R}$ (**Beachte:** $f(x) \geq 1$ ist möglich!)
- $\int_{-\infty}^{+\infty} f(t)dt = 1$
- $F'(x) = f(x)$

Berechnung der W'keiten bei stetigen ZVn

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



Beachte: für eine stetige ZV

$$P(X = x) = P(x \leq X \leq x) = F(x) - F(x) = 0.$$

Die wichtigsten stetigen Verteilungen

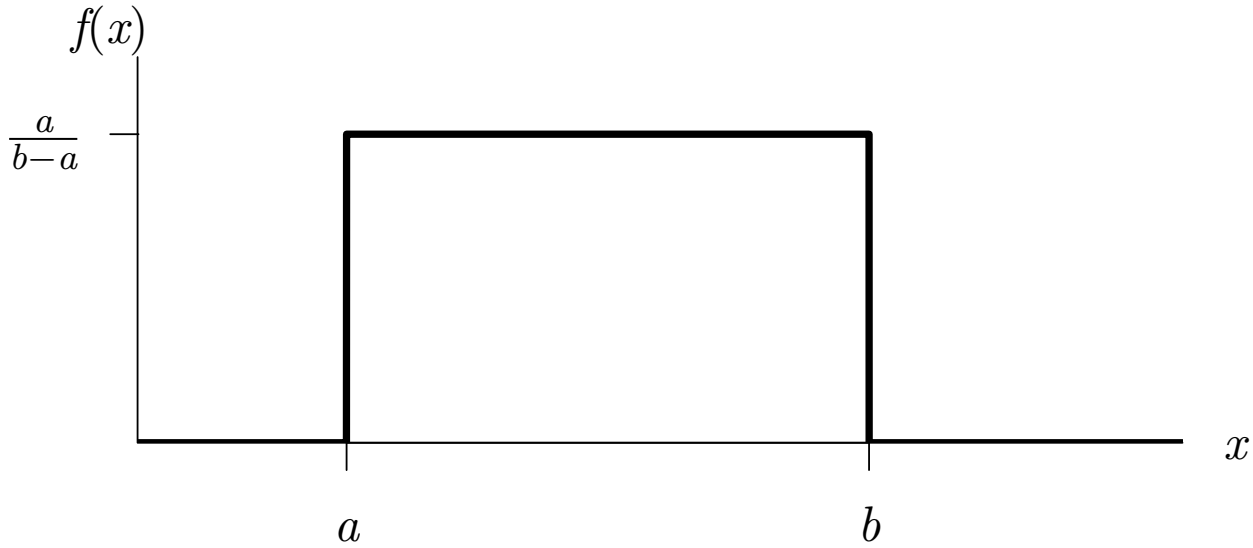
- Gleichverteilung
- Exponentialverteilung
- Normalverteilung (Gaussche Verteilung)
- t-Verteilung
- χ^2-Verteilung
- F-Verteilung

Gleichverteilung

Eine Zufallsvariable X mit

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

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



Verteilungsfunktion:

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

Da für $x \in [a; b]$ gilt:

$$\int_a^x f(t)\,\mathrm{d}t = \int_a^x \frac{1}{b-a}\,\mathrm{d}t = \left[\frac{t}{b-a}\right]_0^x = \frac{x-a}{b-a}.$$

Beispiel: X ist die Bearbeitungsdauer eines Kundenauftrags und wird als gleichverteilt in $[1;10]$ angenommen:

$$f(x) = \begin{cases} 0, & \text{falls} \quad x < 1 \\ \frac{1}{9}, & \text{falls} \quad 1 \leq x \leq 10 \\ 0, & \text{falls} \quad x > 10 \end{cases} \quad \Rightarrow \quad F(x) = \begin{cases} 0, & \text{falls} \quad x < 1 \\ \frac{x-1}{9}, & \text{falls} \quad 1 \leq x \leq 10 \\ 1, & \text{falls} \quad x > 10 \end{cases}$$

$$\begin{aligned} P(2 \leq X \leq 5) &= F(5) - F(2) = \frac{5-1}{9} - \frac{2-1}{9} = \frac{1}{3} \\ P(X = 5) &= \frac{1}{9} \\ P(X = -5) &= 0 \end{aligned}$$

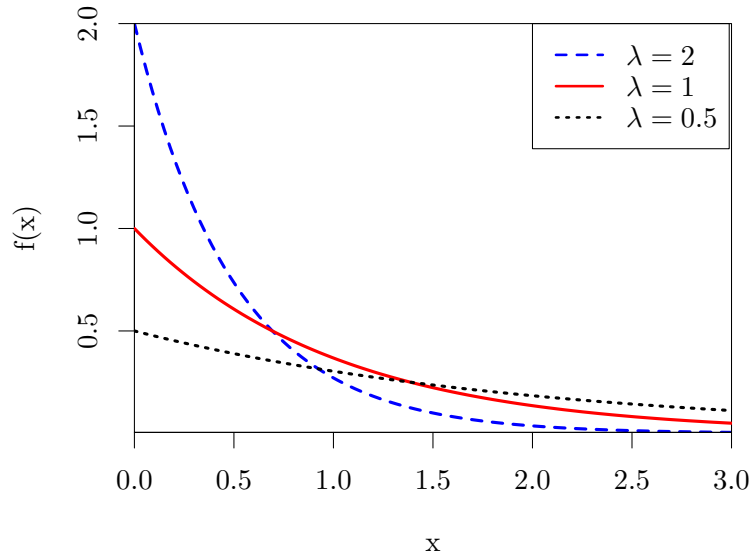
```
punif(q = 5, min = 1, max = 10) - punif(q = 2, min = 1, max = 10)
## [1] 0.3333333
dunif(x = 5, min = 1, max = 10)
## [1] 0.1111111
dunif(x = -5, min = 1, max = 10)
## [1] 0
```

Exponentialverteilung

Eine Zufallsvariable X mit

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

und $\lambda > 0$ heißt **exponentialverteilt**.



Kurzschreibweise: $X \sim \text{Exp}(\lambda)$.

- Verteilungsfunktion:

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, & \text{falls } x \geq 0 \\ 0, & \text{falls } x < 0 \end{cases}$$

- Die **Exponentialverteilung** wird oft zur Modellierung von Lebensdauern verwendet.

Beispiel (Computerbildschirm): Die Lebensdauer eines Computerbildschirms folge einer Exponentialverteilung mit $\lambda = 0.08$.

Wie groß ist die Wahrscheinlichkeit, dass er eine Lebensdauer X größer als 10 Jahre besitzt? Es ist

$$P(X > 10) = 1 - F(10) = e^{-0.08 \cdot 10} = e^{-0.8} \approx 0.45$$

```
1 - pexp(q = 10, rate = 0.08)
## [1] 0.449329
```

- Exponentialverteilung ist „gedächtnislos“:

$$P(X \leq t + s | X \geq t) = P(X \leq s)$$

Beispiel: Der Computerbildschirm ist bereits 7 Jahre alt. Wie groß ist die Wahrscheinlichkeit, dass er innerhalb zwei Jahren kaputt geht? ($X \sim \text{Exp}(0.08)$, vgl. F174)

$$P(X \leq 9 | X \geq 7) = \frac{P(7 \leq X \leq 9)}{P(X \geq 7)} = \frac{F(9) - F(7)}{1 - F(7)}$$

$$= \frac{(1 - e^{-0.08 \cdot 9}) - (1 - e^{-0.08 \cdot 7})}{1 - (1 - e^{-0.08 \cdot 7})} = \frac{e^{-0.56} - e^{-0.72}}{e^{-0.56}}$$

$$= 1 - \frac{e^{-0.72}}{e^{-0.56}} = 1 - e^{-0.16} = 1 - e^{-0.08 \cdot 2}$$

$$= F(2) = P(X \leq 2) = 0.1479$$

```
(pexp(9, rate = 0.08) - pexp(7, rate = 0.08))/(1 - pexp(7, rate = 0.08))
## [1] 0.1478562
pexp(q = 2, rate = 0.08)
## [1] 0.1478562
```

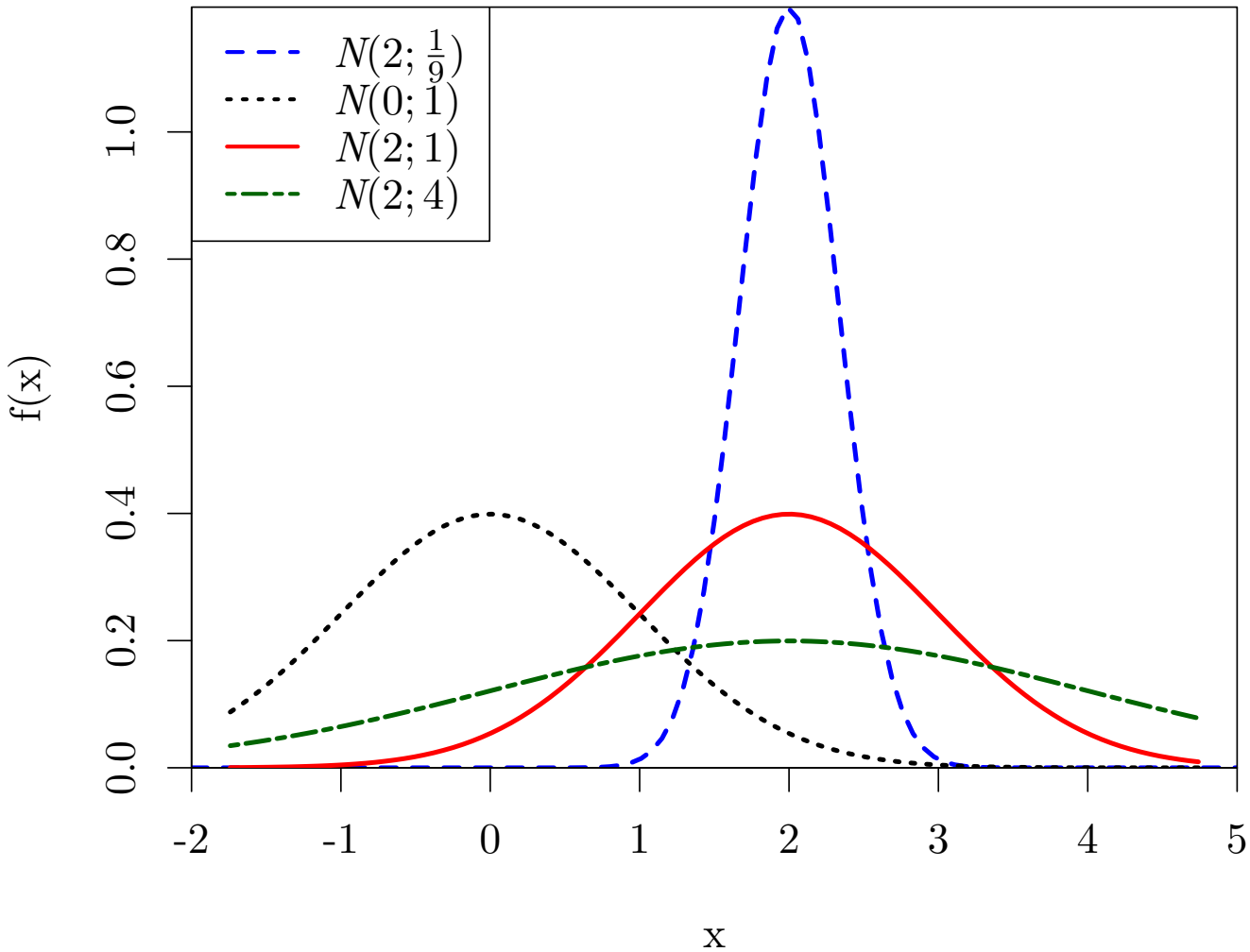
Normalverteilung

Eine Zufallsvariable X mit der Dichtefunktion

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

mit $\mu \in \mathbb{R}$ und $\sigma > 0$ heißt **normalverteilt**.

- Kurzschreibweise: $X \sim N(\mu; \sigma^2)$
- Die Normalverteilung ist die wichtigste stetige Verteilungsfunktion.


$N(2;\frac{1}{9})$
$N(0;1)$
$N(2;1)$
$N(2;4)$
f(x)
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
3
4
5
x

Eigenschaften der Normalverteilung

- Die Dichtefunktion ist achsensymmetrisch zu μ:

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

- Das Maximum von f wird an der Stelle $x = \mu$ angenommen.
- f hat zwei Wendepunkte, nämlich an den Stellen $\mu - \sigma$ und $\mu + \sigma$
- Man bezeichnet die Normalverteilung mit $\mu = 0$ und $\sigma = 1$ als **Standardnormalverteilung**. Wir schreiben Φ für die Verteilungsfunktion der Standardnormalverteilung und ϕ für deren Dichte.

Eigenschaften der Standardnormalverteilung:

Ist $X \sim N(\mu; \sigma^2)$, so ist

$$\frac{X-\mu}{\sigma} \sim \Phi \quad (\ N(0;1)\ ).$$

Folglich ist

$$F_X(x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

- Die Funktion Φ ist vertafelt, bzw. Software
- Kenntnis von Φ genügt um die VF einer beliebigen Normalverteilung $N(\mu, \sigma)$ zu bestimmen.
- Wegen $\phi(x) = \phi(-x)$ ist $\Phi(-x) = 1 - \Phi(x) \rightsquigarrow$ die Tabelle für Φ enthält nur positive Werte von x

Beispiel: Projektdauer

Sei X die Dauer des Projektes in Wochen. Es wird angenommen, dass $X \sim N(39; 2^2)$. Wie hoch ist die Wahrscheinlichkeit, dass die Projektdauer zwischen 37 und 41 Wochen liegt?

$$\begin{aligned} P(37 \leq X \leq 41) &= F(41) - F(37) \\ &= \Phi\left(\tfrac{41-39}{2}\right) - \Phi\left(\tfrac{37-39}{2}\right) \\ &= \Phi(1) - \Phi(-1) \\ &= \Phi(1) - [1 - \Phi(1)] \\ &= 2 \cdot \Phi(1) - 1 \\ &= 2 \cdot 0.8413 - 1 \\ &= 0.6826 \end{aligned}$$

(„1σ-Bereich“)

```
pnorm(q = 41, mean = 39, sd = 2) - pnorm(q = 37, mean = 39,
    sd = 2)
## [1] 0.6826895
```

Verteilungsfunktion der Standardnormalverteilung: $\Phi(x) = P(X \leq x)$, mit $X \sim N(0,1)$

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	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	0.5753
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	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	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	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	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	0.9864	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	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936

Eigenschaften der Normalverteilung II

- Wahrscheinlichkeit, um höchstens c von μ abzuweichen:

$$\begin{aligned} P(\mu - c \leq X \leq \mu + c) &= F(\mu + c) - F(\mu - c) \\ &= \Phi\left(\tfrac{\mu+c-\mu}{\sigma}\right) - \Phi\left(\tfrac{\mu-c-\mu}{\sigma}\right) \\ &= \Phi\left(\tfrac{c}{\sigma}\right) - \Phi\left(-\tfrac{c}{\sigma}\right) \\ &= \Phi\left(\tfrac{c}{\sigma}\right) - [1 - \Phi\left(\tfrac{c}{\sigma}\right)] \\ &= 2 \cdot \Phi\left(\tfrac{c}{\sigma}\right) - 1 \end{aligned}$$

$k\sigma$-Bereich $[\mu - k\sigma, \mu + k\sigma]$:

$$P(\mu - k\sigma \leq X \leq \mu + k\sigma) = 2\Phi(k) - 1 = \begin{cases} 0.683, & \text{falls } k = 1 \\ 0.954, & \text{falls } k = 2 \\ 0.997, & \text{falls } k = 3 \end{cases}$$

Reproduktionseigenschaft der Normalverteilung

- Ist $X \sim \Phi$, so ist $\mu + \sigma\, X \sim N(\mu; \sigma^2)$.
- Ist $X \sim N(\mu; \sigma^2)$, so ist $a\, X + b \sim N(a\, \mu + b; a^2\sigma^2)$.
- Sind $X_1, \ldots, X_n$ beliebig normalverteilt, so ist

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

ebenfalls normalverteilt.

2.5. Zweidimensionale Verteilungen

Es sei $X = (X_1, X_2)'$ (z.B.: Renditen von Daimler und BMW, Wechselkurse Euro/US$ und Euro/CHF). Dann heißt

$$F_X(x_1, x_2) = P\Big(\{\omega \in \Omega : X_1(\omega) \leq x_1, X_2(\omega) \leq x_2\}\Big), \quad x_1, x_2 \in I\!R$$

die **(2–dimensionale) Verteilungsfunktion** des Zufallsvektors X. Wir schreiben hierfür kurz $F(x_1, x_2) = P(X_1 \leq x_1, X_2 \leq x_2)$.

$F_X(x_1, \infty)$ heißt die **Randverteilungsfunktion von X_1** und $F_X(\infty, x_2)$ heißt die **Randverteilungsfunktion von X_2.**

Beachte: Es ist

$$\begin{aligned} F_X(x_1, \infty) &= P(X_1 \leq x_1) =: F_1(x_1) \quad \text{und} \\ F_X(\infty, x_2) &= P(X_2 \leq x_2) =: F_2(x_2). \end{aligned}$$

bisher: Unabhängigkeit von Ereignissen

jetzt: Unabhängigkeit von Zufallsvariablen

X_1 und X_2 sind stochastisch unabhängig, wenn

$$P(X_1 \leq x_1; X_2 \leq x_2) = P(X_1 \leq x_1) \cdot P(X_2 \leq x_2).$$

Dies entspricht auch der Bedingung

$$\begin{aligned} F(x_1, x_2) &= F_{X_1}(x_1) \cdot F_{X_2}(x_2), \\ f(x_1, x_2) &= f_{X_1}(x_1) \cdot f_{X_2}(x_2). \end{aligned}$$