Biometrie

Sebastian Mentemeier*

17. Juli 2019

Warum Biometrie - Einführung in die Statistik?

In dieser Vorlesung wollen wir die Natur des Zufalls verstehen und statistische Verfahren zum Umgang mit zufallsbehafteten Phänomenen behandeln. Dazu drei Beispiele:

1. Interpretation von Wachstumskurven: Durchschnittsgröße, Durchschnittsgewicht von Kleinkindern in einem gewissen Alter
2. Blutgruppen: Mutter Genotyp A0 (Phänotyp A), Vater Genotyp B0 (Phänotyp B). Genotyp und Phänotyp des Kindes sind zufällig! Mögliche Genotypen (Phänotypen) des Kindes: A0 (A), B0 (B), AB (AB), 00 (0).
3. Überprüfung von Hypothesen, z.B. „Der Nitratgehalt der Fulda überschreitet den gesetzlichen Grenzwert“

Analog zu diesen drei Beispielen ist auch die Vorlesung aufgeteilt in drei große Kapitel:

1. Explorative Datenanalyse (Grafische Darstellung, statistische Kennzahlen, empirische Korrelationen)
2. Grundlagen der Wahrscheinlichkeitstheorie (diskrete und stetige Wahrscheinlichkeitsverteilungen, bedingte Wahrscheinlichkeit)
3. Schließende Statistik (Schätzer, Hypothesentests, Regressionsanalyse)

Abschließend noch zum Begriff Biometrie: Heutzutage wird dieser zuerst mit Gesichtserkennung, Personalausweisen und Flughafenkontrollen in Verbindung gebracht; als Titel dieser Vorlesung meint er jedoch die Anwendung statistischer und wahrscheinlichkeitstheoretischer Verfahren in den Biowissenschaften. Man kann auch von *Biostatistik* sprechen.

Ich **danke** meinem Kollegen Felix Lindner für die freundliche Überlassung seines Vorlesungsmanuskripts, welches an vielen Stellen als Vorlage diente.

*Prof. Dr. Sebastian Mentemeier, Universität Kassel, FB 10, Institut für Mathematik; mentemeier@mathematik.uni-kassel.de

Inhaltsverzeichnis

Teil I.
Explorative Datenanalyse

Wir beginnen mit beschreibender Statistik und lernen verschiedene Methoden kennen, gegebene Daten darzustellen (in Tabellenform oder grafisch) und mittels aussagekräftiger Kennzahlen zu beschreiben. Darauf aufbauend sucht die explorative Datenanalyse nach Strukturen in den Daten, mit dem Ziel, Hypothesen über Eigenschaften der zugrundeliegenden Untersuchungsobjekte zu formulieren.

1. Grundbegriffe

Definition 1.1.

Statistische Einheiten:	Objekte (Personen, Lebewesen), an denen interessierende Größen erfasst werden
Grundgesamtheit:	Menge aller für die Fragestellung relevanten statistischen Einheiten
Stichprobe:	tatsächlich untersuchte Teilmenge der Grundgesamtheit
Merkmal:	interessierende Größe
Merkmalsausprägung:	konkreter Wert des Merkmals für eine bestimmte statistische Einheit

Der *Stichprobenumfang* (oft mit N bezeichnet) ist die Anzahl der in der Stichprobe enthaltenen Untersuchungseinheiten. Man spricht von *univariaten / bivariaten / multivariaten* Daten, je nachdem, ob ein / zwei / drei oder mehr Merkmale betrachet werden.

Beispiel 1.2.

- *Grundgesamtheit:* BSc-Biologie-Studierende im Sommersemester 2019 an der Uni Kassel.
- *Stichprobe:* Die ersten zehn zur Vorlesung „Biometrie“eintreffenden Studierenden werden befragt.
- $N = 10$.
- *Erhobene Merkmale und Merkmalsausprägungen:*

Merkmal	Ausprägungen
Haarfarbe	blond, braun, schwarz, rot, grau, ...
Note „Mathematik für Biologen“	0.7, 1.0, 1.3, ..., 3.7, 4.0, 5
Semesteranzahl	1, 2, 3, 4, ...
Körpergröße	alle Werte im Intervall [50, 250] (in cm)

Definition 1.3 (Merkmalstypen)**.**

diskret:	endlich oder abzählbar unendlich viele, isolierte Ausprägungen
stetig:	alle Werte eines Intervalls sind (prinzipiell) mögliche Ausprägungen
nominalskaliert:	Ausprägungen sind Namen, keine Ordnung möglich
ordinalskaliert:	Ausprägungen können geordnet, aber Abstände nicht interpretiert werden
intervallskaliert:	Ausprägungen sind Zahlen, Interpretationen der Abstände möglich
verhältnisskaliert:	Ausprägungen besitzen zusätlich sinnvollen absoluten Nullpunkt
metrisch:	intervall- oder verhältnisskaliert
qualitativ:	endlich viele Ausprägungen, höchstens Ordinalskala
quantitativ:	Ausprägungen sind Zahlen

Beispiel 1.4. Die Haarfarbe, Note und Semesteranzahl sind diskrete Merkmale; die Körpergröße ist ein stetiges Merkmal. Haarfarbe ist nominalskaliert, Note ist ordinalskaliert, Semesterzahl und Körpergröße sind verhältnisskaliert.

Die Einteilung ist nicht immer völlig eindeutig. So bezeichnet man Merkmale als *quasi-stetig*, wenn durch Begrenzung der Meßgenauigkeit nicht jeder beliebige Wert eines Intervalls, sondern nur endlich viele verschiedene Ausprägungen angenommen werden können. Dies trifft bspw. auf die Körpergröße zu.

Bemerkung 1.5. Je nach Skalenart sind verschiedene Berechnungen zulässig:

Skalenart	auszählen	ordnen	Differenzen	Quotienten
nominal	ja	nein	nein	nein
ordinal	ja	ja	nein	nein
intervall	ja	ja	ja	nein
verhältnis	ja	ja	ja	ja

Beispiel 1.6 (vgl. [6, S.12f])**.**
Ein Nährboden wurde 30 Minuten bei Zimmertemperatur offen stehen gelassen. Nach 3 Tagen Inkubationszeit waren 40 Pilz- bzw. Bakterienkolonien gewachsen. Es wurden folgende Merkmale bestimmt:

Merkmal	Typ	Erläuterung
Durchmesser	metrisch	in mm
Antibiotikaresistenz	ordinal	3 Ausprägungen: sensitiv, intermediär, resistent
Farbe	nominal	7 Ausprägungen: gelb, weißlich, braun, orange, farblos, rosa, grün

Da drei Merkmale erhoben wurden, handelt es sich um multivariate Daten.
Der folgende Datensatz wurde simuliert.

1.1. Erste Schritte in R

1.1.1. R herunterladen und installieren

Downloaden Sie R unter https://cran.r-project.org.

Starten Sie RGui. Nun können Sie in der sog. R-Konsole Befehle eingeben. Im Datei-Menü haben Sie mit `Neues Skript` bzw. `Öffne Skript ...` die Möglichkeit, in einem zweiten Fenster sog. R-Skripte zu erstellen. D.h. sie geben zeilenweise R-Befehle ein und können diese dann einzeln oder auch blockweise mit `Strg+R` ausführen lassen. Dies ist eine sehr bequeme Art, mit R zu arbeiten und diese Arbeit mit anderen zu teilen.

1.1.2. R als Taschenrechner

`+,-`	Addition, Subtraktion
`*, /`	Multiplikation, Division
`^`	Potenz
`exp(·)`	Exponentialfunktion
`sin(·), cos(·), tan(·)`	trigonometrische Funktionen

Beispiele:

`1+2*3`	liefert 7
`2*5^2`	liefert 50
`4 * sin(pi/2)`	liefert 4

1.1.3. Zuweisungen

`x <- 2.25`	dem Objekt x wird die Zahl 2.25 zugewiesen
x	Wert von x wird ausgegeben
`3 ->y`	funktioniert auch (beachte Pfeilrichtung)
x+y	Wert 5.25 wird ausgegeben

1.1.4. Generierung von Vektoren

`seq(0,1,0.1)`	generiert den Vektor `0 0.1 ... 0.9 1`
`1:10`	wie `seq(1,10,1)`
`c(1,2,3)`	generiert den Vektor `1 2 3`
`rep(c(2,7),2)`	erzeugt den Vektor `2 7 2 7`

Nr	Durchmesser	Resistenz	Farbe
1	10.8	intermediär	grün
2	3.3	sensitiv	weißlich
3	4.6	sensitiv	braun
4	7.0	sensitiv	farblos
5	10.9	intermediär	grün
6	2.6	sensitiv	weißlich
7	10.8	intermediär	grün
8	11.3	intermediär	grün
9	8.0	resistent	farblos
10	7.6	resistent	farblos
11	0.9	sensitiv	gelb
12	2.6	sensitiv	weißlich
13	2.3	sensitiv	weißlich
14	8.3	resistent	farblos
15	4.7	sensitiv	braun
16	9.3	resistent	rosa
17	6.1	sensitiv	orange
18	8.7	resistent	rosa
19	11.9	intermediär	grün
20	4.7	sensitiv	braun
21	9.4	resistent	rosa
22	11.2	intermediär	grün
23	2.7	sensitiv	weißlich
24	7.9	resistent	farblos
25	1.7	sensitiv	gelb
26	3.4	sensitiv	weißlich
27	4.8	sensitiv	braun
28	0.4	sensitiv	gelb
29	4.7	sensitiv	braun
30	10.5	intermediär	grün
31	4.2	sensitiv	braun
32	5.9	sensitiv	orange
33	7.3	resistent	farblos
34	6.0	sensitiv	orange
35	2.4	sensitiv	weißlich
36	10.0	intermediär	rosa
37	8.1	resistent	farblos
38	9.6	resistent	rosa
39	1.5	sensitiv	gelb
40	8.7	resistent	rosa

Tabelle 1: Messwerte zu Beispiel 1.6

Indizierung und Komponentenansteuerung: Vektoren

`x[i]`	gibt die `i`-te Komponente des Vektors `x` aus
`x[1:5]`	gibt die ersten 5 Komponenten von `x` aus
`x[c(2,3,5)]`	gibt die 2., 3. und 5. Komponente des Vektors `x` aus
`x[y<=30]`	gibt den Vektor derjenigen Komponenten $\mathtt{x_i}$ aus, für die $\mathtt{y_i} \leq 30$ ist
`which[y<=30]`	gibt die Positionen derjenigen Komponenten $\mathtt{y_i}$ aus, für die $\mathtt{y_i} \leq 30$ ist

1.1.5. Der wichtigste Befehl

`help(Befehlsname)` bzw. `?Befehlsname` ruft die Hilfeseite auf, z.B. `?seq`. Besonders nützlich ist dies, um die Syntax der Befehle nachzuschlagen. Falls `help` oder `?` kein Ergebnis liefern, `??Befehl` probieren, dies liefert eine erweiterte Suche.

1.2. Daten in R

1.2.1. Grundlegendes

- Dezimalzahlen werden mit Punkt notiert! Beispiel: 0.5 ist $\frac{1}{2}$.
  $0,5$ wird als die zwei Zahlen 0 und 5 interpretiert
- Sollen qualitative Merkmale (Namen, Farben, etc.) erfasst werden, so müssen die Ausprägungen jeweils in Anführungsstriche gesetzt werden. Beispiel: `x<-"Z"` weist der Variable x den Buchstaben „Z“ zu.
  `x<-Z` hingegen würde der Variable x den Wert der *Variablen* Z zuweisen (falls diese existiert).

1.2.2. Univariate Daten als Vektoren

Wir wollen die ersten 4 Datensätze aus dem Bakterien-Beispiel 1.6 als Vektoren ablegen. Grundsätzlich werden Vektoren mit

```
c(Erster Entrag, Zweiter Eintrag, ..., Letzter Eintrag)
```

erzeugt; abhängig vom Typ der Einträge haben die Vektoren dann den Typ `"numeric"` (Zahlen, also qualitative Merkmale) oder `"char"` (Zeichenketten). Mit Befehlen wie `ordered` (s.u.) kann der Typ des im Vektor abgelegten Merkmals genauer spezifiziert werden.

a) quantitative Merkmale:

`x<-c(10.8, 3.3, 4.6, 7.0)` `class(x)`	Vektor mit Durchmesser der ersten 4 Kolonien liefert `"numeric"`, dies entspricht dem quantitativen Merkmalstyp

b) ordinale Merkmale:

```
y<-ordered( c("intermediär", "sensitiv", "sensitiv", "sensitiv"),
   levels=c("sensitiv", "intermediär", "resistent") )
```

`ordered` sorgt für die Interpretation der Zeichenketten als ordinale Merkmale, `levels` legt die Rangfolge fest. Der Aufruf `y` liefert

```
[1] intermediär sensitiv    sensitiv    sensitiv
Levels: sensitiv < intermediär < resistent
```

`class(y)` liefert

```
[1] "ordered" "factor"
```

c) nominale Merkmale:

```
z<-factor( c("grün", "weißlich", "braun", "farblos"))
```

`factor` sorgt für die Interpretation der Zeichenketten als nominale Merkmale. Der Aufruf `z` liefert

```
[1] grün     weißlich braun    farblos
Levels: braun farblos grün weißlich
```

`class(z)` liefert

```
[1] "factor"
```

1.2.3. Bi- und multivariate Daten als Datentabellen (data frames)

Bi- und multivariate Daten werden in Tabellenform abgelegt. Dabei entspricht jede Zeile einer Untersuchungseinheit, jede Spalte entspricht einem Merkmal. Wir wollen die oben definierten Vektoren zu einer Datentabelle zusammenfassen. Der Aufruf

```
Tab <- data.frame(x,y,z)
```

erzeugt eine Datentabelle mit den Spalten x, y, z.

Der Aufruf `Tab` liefert dann

```
     x           y        z
1 10.8 intermediär     grün
2  3.3    sensitiv weißlich
3  4.6    sensitiv    braun
4  7.0    sensitiv  farblos
```

Indizierung und Komponentenansteuerung: Matrizen

`Tab$x`	gibt den Vektor `x` der Datentabelle `Tab` aus
`Tab[4,]`	gibt die 4. Zeile der Datentabelle `Tab` aus
`Tab[,3]`	gibt die 3. Spalte von `Tab` aus (liefert das gleiche Ergebnis wie `Tab$z`)
`Tab[4,3]`	gibt Eintrag 4. Zeile, 3. Spalte wieder
`Tab[Tab$x<=5]`	gibt alle Zeilen der Datentabelle `Tab` aus, die in der Spalte `x` einen Wert ≤ 5 haben
`subset(Tab,x<=70)`	wie oben, in vielen Situationen einfacher

1.3. Kurz-Befehlsreferenz

`x<-c(1,2,3)`	erzeugt einen Vektor (hier mit den Einträgen 1, 2 und 3) und weist diesen der Variable x zu.
`ordered(c("a","b"), levels=c("a", "b"))`	erzeugt einen Vektor mit ordinalskalierten Einträgen `"a"` und `"b"`, und legt die Rangfolge `"a" < "b"` fest.
`factor(c("a","b"))`	erzeugt einen Vektor mit nominalskalierten Einträgen `"a"` und `"b"`.
`data.frame(x,y)`	erzeugt eine Datentabelle mit den Spalten x und y.

2. Häufigkeitsverteilungen und die grafische Darstellung univariater Daten

Der elementarste Schritt zur Aufbereitung erhobener Daten ist das Auszählen. Im Folgenden gehen wir immer von einem univariaten Datensatz aus, und bezeichnen die (Merkmalsausprägungen in der) Stichprobe mit

$$(x_1, \ldots, x_N).$$

In Beispiel 1.6 könnten wir uns auf die Betrachtung der Antibiotikaresistenzen beschränken (um einen univariaten Datensatz zu erhalten), dann wäre der Stichprobenumfang $N = 40$ und

$$(x_1, x_2, \ldots, x_{39}, x_{40}) \;=\; (\text{intermediär, sensitiv, } \ldots, \text{ sensitiv, resistent}).$$

2.1. Diskrete Merkmale

Wir betrachten zuerst die Situation eines diskreten Merkmals, und nehmen zusätzlich an, dass nur endlich viele verschiedene Merkmalsausprägungen möglich sind. Wir bezeichnen

die verschiedenen möglichen Ausprägungen mit

$$a_1, \ldots, a_J;$$

J ist also die Anzahl der verschiedenen möglichen Ausprägungen. Betrachten wir wieder in Beispiel 1.6 das Merkmal der Antibiotikaresistenz, so wäre $J = 3$ und

$$a_1 = \text{sensitiv}, \quad a_2 = \text{intermediär}, \quad a_3 = \text{resistent}.$$

Definition 2.1 (Absolute und relative Häufigkeiten)**.**

$h(a_j) = h_j$	*absolute Häufigkeit* der Ausprägung a_j in der Stichprobe, d.h. Anzahl der x_i aus $x_1, \ldots, x_N$ mit $x_i = a_j$.
$f(a_j) = f_j := h_j/N$	*relative Häufigkeit* der Ausprägung a_j
$h_1, \ldots, h_J$	*absolute Häufigkeitsverteilung* des beobachteten Merkmals
$f_1, \ldots, f_J$	*relative Häufigkeitsverteilung* des beobachteten Merkmals

Beispiel 2.2. Wir betrachten weiterhin das Merkmal der Antibiotikaresistenz aus Beispiel 1.6, wobei wir uns auf die ersten 4 Kolonien beschränken, um alles von Hand zählen zu können - also $N = 4$ und

$$(x_1, x_2, x_3, x_4) \;=\; (\text{intermediär, sensitiv, sensitiv, sensitiv})$$

sowie

$$a_1 = \text{sensitiv}, \quad a_2 = \text{intermediär}, \quad a_3 = \text{resistent}.$$

Als absolute bzw. relative Häufigkeiten erhalten wir

$$h_1 = 3 \qquad h_2 = 1 \qquad h_3 = 0$$
$$f_1 = \frac{3}{4} \qquad f_2 = \frac{1}{4} \qquad f_3 = 0$$

In R liefert der Befehl `table(x)` die absolute Häufigkeitsverteilung des Vektors x; Division durch den Stichprobenumfang liefert dann die relativen Häufigkeiten. Die Länge (=Anzahl der Einträge) des Vektors x lässt sich mit dem Befehl `length(x)` abfragen, wir erhalten also die relativen Häufigkeiten mit dem Befehl

`table`

`length`

```
table(x)/length(x)
```

Definition 2.3. In einem *Säulendiagramm (Stabdiagramm)* wird über jeder möglichen Merkmalsausprägung eine Säule (ein Stab) in Höhe der entsprechenden absoluten Häufigkeit gezeichnet.

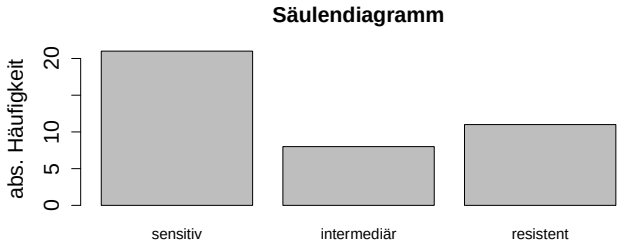


Abbildung 1: Säulendiagramm

Ein Stab- bzw. Säulendiagramm dient der grafischen Darstellung der Häufigkeitsverteilung qualitativer Merkmale (nominal- oder ordinalskaliert).

Nehmen wir alle 40 Beobachtungen aus Beispiel 1.6 für das Merkmal Farbe, so erhalten wir folgende absolute Häufigkeitsverteilung:

$$h_1 = 21, \quad h_2 = 8, \quad h_3 = 11.$$

Ein Säulendiagramm erhalten wir in R mit dem Befehl `barplot`

```
barplot(table(Resistenz), ylab="abs. Häufigkeit", main="Säulendiagramm")
```

Hierbei sorgen die Argumente `ylab` und `main` für die Beschriftung der y-Achse bzw. die Überschrift.

Für ein Stabdiagramm werden Striche (Stäbe) anstelle der Säulen gezeichnet. Dies geschieht mit dem Aufruf `plot`

```
plot(table(Resistenz), type="h", ylab="abs. Häufigkeit", main="Stabdiagramm")
```

Der Befehl `plot` ist sehr vielseitig und „intelligent“- abhängig von dem übergebenen Datensatz kann er verschiedene Resultate liefern! Hier wird mit `type="h"` spezifiziert, dass ein Stabdiagramm gezeichnet werden soll.

Bemerkung 2.4. Es gibt einen Unterschied zwischen Säulen- und Stabdiagramm: Sind die Merkmalsausprägungen Zahlen, so werden diese beim Stabdiagramm auf ihren Positionen auf der Zahlengeraden abgetragen, während sie beim Säulendiagramm als nominale Merkmale interpretiert werden. Dies verdeutlicht das folgende Beispiel.

Beispiel 2.5. Semesterzahl von $N = 5$ Studierenden des Biologie-Bachelors (vgl. Beispiel 1.2).
Mögliche Ausprägungen (Begrenzung z.B. Regelstudienzeit 6 Semester).

$$a_1 = 1,\ a_2 = 2,\ a_3 = 3,\ a_4 = 4,\ a_5 = 5,\ a_6 = 6$$

Erhobene Daten ($N = 10$)

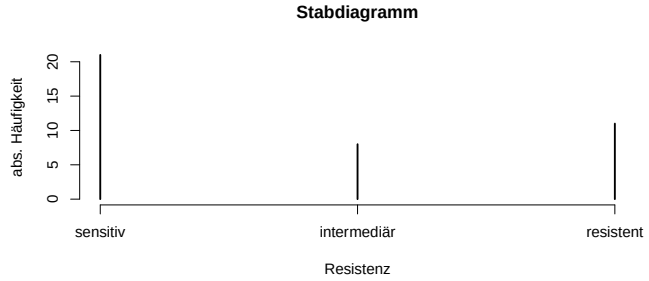


Abbildung 2: Stabdiagramm

i	1	2	3	4	5	6	7	8	9	10
x_i	2	3	1	4	6	2	2	2	4	4

Absolute / relative Häufigkeiten:

$$h_1 = 1,\ h_2 = 4,\ h_3 = 1,\ h_4 = 3,\ h_5 = 0, \quad h_6 = 1$$

$$f_1 = \frac{1}{10},\ h_2 = \frac{2}{5},\ h_3 = \frac{1}{10},\ h_4 = \frac{3}{10}, h_5 = 0,\ h_6 = \frac{1}{10}$$

Zum Zeichnen wurde folgender R-Code verwendet:

```
x<-c(2,3,1,4,6,2,2,2,4,4)
par(mfrow=c(1,2))
barplot(table(x),main="Säulendiagramm", xlab="Semesteranzahl", ylab="abs. Häufigkeiten")
plot(table(x),main="Stabdiagramm", xlab="Semesteranzahl", ylab="abs. Häufigkeiten")
```

Als letzte Darstellungsmöglichkeit für qualitative Daten betrachten wir Kreisdiagramme.

Definition 2.6. In einem *Kreisdiagramm* wird jeder Merkmalsausprägung ein Kreissektor zugewiesen, dessen Fläche proportional zur relativen Häufigkeit ist. Winkel des Kreissektors zu Ausprägung a_k: $f_k \cdot 360°$

Beispiel 2.7. Wir betrachten die Ergebnisse der Bundestagswahl 2017 (Zweitstimmen):

Mit folgenden R-Befehlen zeichnen wir ein Kreisdiagramm: `pie`

```
Parteien<-c("SPD","CDU","CSU","DIE LINKE","GRÜNE","FDP","AFD")
Ergebnisse<-c(11429231, 14030751, 3255487, 3966637, 3717922, 3249238, 5317499)
Farben<-c("red","black","grey","pink","green","yellow","blue")
pie(Ergebnisse,labels=Parteien,col=Farben,main="Kreisdiagramm")
```

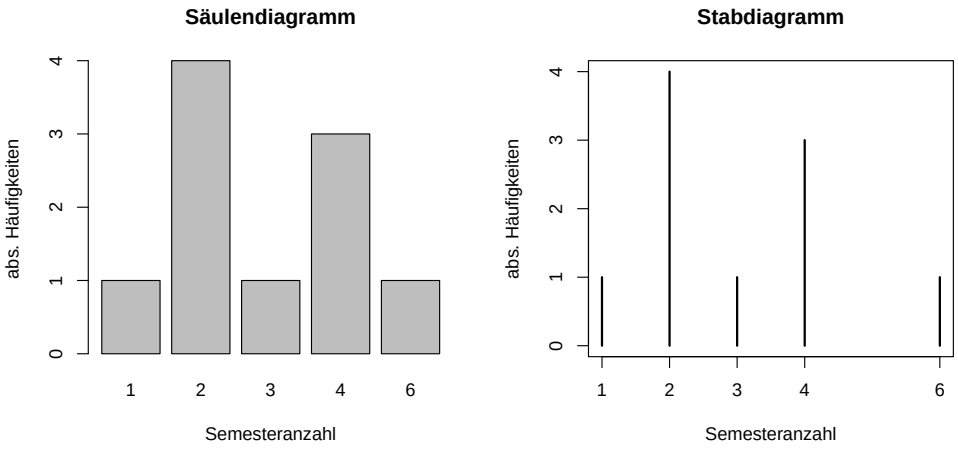


Abbildung 3: Säulen- und Stabdiagramm zu Beispiel 2.5

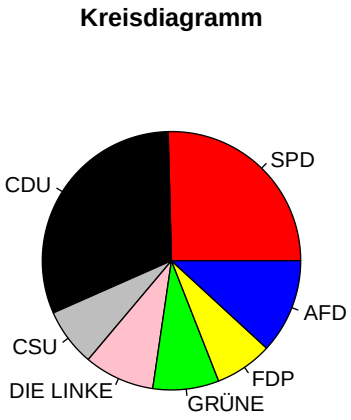


Abbildung 4: Kreisdiagramm: Ergebnisse der Bundestagswahl 2017

Partei	Zweitstimmen
SPD	11429231
CDU	14030751
CSU	3255487
DIE LINKE	3966637
GRÜNE	3717922
FDP	3249238
AFD	5317499

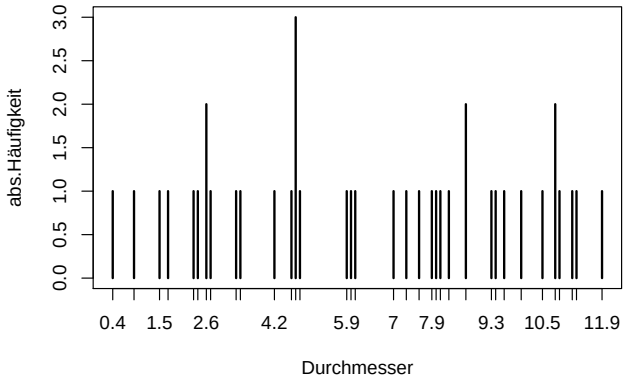


Abbildung 5: Säulendiagramm zum Merkmal Durchmesser im Beispiel 1.6

2.2. Stetige Merkmale

Wird ein stetiges Merkmal beobachtet (z.B. Längen-, Gewichtsmessungen), so wird nur in wenigen Fällen exakt dieselbe Ausprägung mehrfach angenommen. Dies verdeutlicht das Säulendiagramm (Abbildung 5) zum Merkmal Durchmesser aus Beispiel 1.6. Um dennoch sinnvoll von Häufigkeiten sprechen zu können, werden Ausprägungen in *Klassen* eingeteilt; und nur die Häufigkeiten dieser Klassen angegeben.

Definition 2.8. Wir betrachten ein stetiges Merkmal, dessen mögliche Ausprägungen Werte aus einem Intervall $I = (c_*, c^*]$, $c_* < c^* \in \mathbb{R}$, sind. Gegeben seien weiterhin *Klassengrenzen*

$$c_* = c_0 < c_1 < \cdots < c_{J-1} < c_J = c^*,$$

so dass das Intervall I als disjunkte Vereinigung der *Klassen* $K_j = (c_{j-1}, c_j]$, $1 \leq j \leq J$ darstellbar ist:

$$I \;=\; (c_0, c_1] \cup (c_1, c_2] \cup \cdots \cup (c_{J-1}, c_J].$$

Abbildung 6: Aufteilung eines Intervalls in $J = 5$ Klassen

Die Differenzen $b_j := c_j - c_{j-1}$ werden als *Klassenbreiten* bezeichnet.
Liegt eine Stichprobe $x = (x_1, \ldots, x_N)$ vom Umfang N vor, so bezeichnet

$h(K_j)$	*absolute Häufigkeit* der Klasse K_j in der Stichprobe, d.h. Anzahl der x_k aus $x_1, \ldots, x_N$ mit $c_{j-1} < x_k \leq c_j$.
$f(K_j) := h(K_j)/N$	*relative Häufigkeit* der Klasse K_j

Beispiel 2.9. Wir betrachten weiterhin den Durchmesser der Kolonien aus Beispiel 1.6. Gemäß der Devise: „Erst denken, dann messen"legen wir ZUERST die Klassengrenzen fest - denn offensichtlich kann manchmal eine kleine Verschiebung der Klassengrenzen deutliche Verschiebungen der Häufigkeiten bewirken! Wir wählen

$$c_0 = 0,\ c_1 = 3,\ c_2 = 6,\ c_3 = 9,\ c_4 = 12,$$

also

$$K_1 = (0, 3] \quad K_1 = (3, 6] \quad K_3 = (6, 9], \quad K_4 = (9, 12]$$

Zunächst beschränken wir uns auf die ersten 6 Kolonien, also $N = 6$, es liegen dann folgende Beobachtungen vor.

$$(x_1, x_2, x_3, x_4, x_5, x_6) \ = \ (10.8,\ 3.3,\ 4.6,\ 7.0,\ 10.9,\ 2.6)$$

Von Hand zählen wir nach, dass

$$h(K_1) = 1, \quad h(K_2) = 2, \quad h(K_3) = 1, \quad h(K_4) = 2.$$

Definition 2.10 (Histogramm)**.** In einem Histogramm wird über jeder Klasse $K_j = (c_{j-1}, c_j]$, $1 \leq j \leq J$ ein Rechteck gezeichnet, dessen Fläche proportional ist zur relativen (oder absoluten) Häufigkeit $f(K_j)$.
Konkret gilt für das Rechteck über Klasse K_j:

Breite	$b_j = c_j - c_{j-1}$, sowie
Höhe	$C \cdot f(K_j)/b_j$ für relative Häufigkeiten, bzw. $C \cdot h(K_j)/b_j$ für absolute Häufigkeiten; hierbei ist C eine Proportionalitätskonstante, die für alle Klassen gleich ist

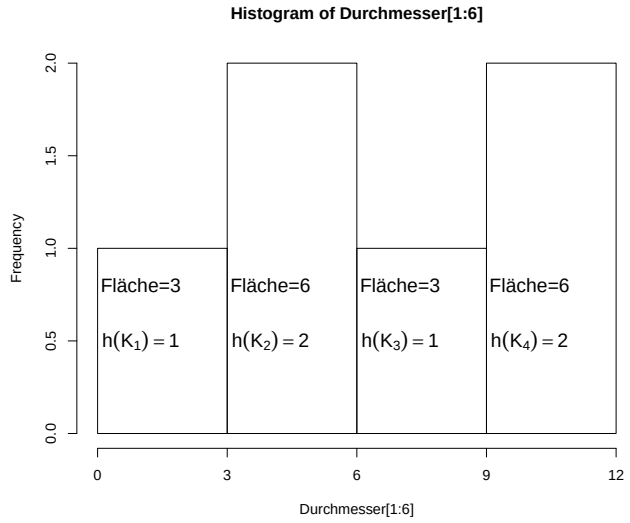


Abbildung 7: Histogramm mit den Daten aus Beispiel 2.9; hier ist $C = 3$.

Beispiel 2.11 (Fortsetzung von Beispiel 2.9)**.** Wir zeichnen in der Situation von Beispiel 2.9 ein Histogramm, dessen Rechtecke proportional zu den absoluten Häufigkeiten sind:

Bemerkung 2.12. Da die relativen und absoluten Häufigkeiten ebenfalls propotional zueinander sind (unterscheiden sich nur um den Faktor N), sind die Flächen eines Histogramms immer proportional sowohl zu den relativen als auch zu den absoluten Häufigkeiten. Der Unterschied liegt nur in der Beschriftung der y-Achse.

Offensichtlich hängt die Form eines Histogramms stark von der Wahl der Klassengrenzen ab. Zu kleine Klassenbreiten erzeugen unübersichtliche Darstellungen, zu große Klassenbreiten führen zu Informationsverlust. In jedem Fall sollte die Klassenbreite konstant gewählt werden; dies ist in R die Standardeinstellung[1].

Beispiel 2.13. Schließlich wollen wir uns das Histogramm zum Merkmal Durchmesser mit allen Beobachtungen aus Beispiel 1.6 zeichnen lassen. Wählen wir die Klassengrenzen und Klassen wie zuvor, also

$$c_0 = 0,\ c_1 = 3,\ c_2 = 6,\ c_3 = 9,\ c_4 = 12,$$

so lautet der R-Befehl: `hist`

```
hist(Durchmesser, breaks=c(0,3,6,9,12), freq=TRUE, xlab="Durchmesser",
ylab="abs. Häufigkeiten", main="Histogramm")
```

[1]Die Klassenbreite wird nach der Formel von *Sturges* gewählt: $b \approx \frac{x_{\max} - x_{\min}}{1+3.322 \log_{10} N}$, hierbei bezeichnen $x_{\max}$ und $x_{\min}$ die größte bzw. kleinste in der Stichprobe auftretende Ausprägung.

Hierbei sorgt das Argument `freq=TRUE` dafür, dass absolute Häufigkeiten auf der y-Achse abgetragen werden; mit `freq=FALSE` würden relative Häufigkeiten notiert.

Werden keine Klassengrenzen angegeben, so wählt R die Klassenbreite und -anzahl automatisch:

```
hist(Durchmesser, freq=TRUE, xlab="Durchmesser",
ylab="abs. Häufigkeiten", main="Histogramm")
```

Die Funktion `hist` kann auch genutzt werden, um die Häufigkeitsverteilungen der Klassen auszugeben. Dazu muss das automatische Zeichnen mit dem Argument `plot=FALSE` abgestellt werden:

```
hist(Durchmesser, freq=TRUE, plot=FALSE)
```

Man erhält (u.a.) folgende Ausgabe

```
$‘breaks‘
[1]  0  2  4  6  8 10 12

$counts
[1] 4 7 8 6 8 7

$density
[1] 0.0500 0.0875 0.1000 0.0750 0.1000 0.0875
```

Unter `$‘breaks‘` sind die von R gewählten Klassengrenzen aufgeführt (oder die selbst vorgegebenen); unter `$counts` dann die absoluten Häufigkeiten der einzelnen Klassen, gefolgt von den relativen Häufigkeiten unter `density`. Mit dem Aufruf

```
hist(Durchmesser, plot=FALSE)$counts
```

erhält man direkt einen Vektor mit den absoluten Häufigkeiten der Klassen (vergleichbar mit `table` im Falle diskreter Merkmale).

2.3. Kurz-Befehlsreferenz

`table`	erzeugt die absolute Häufigkeitsverteilung eines Vektors mit qualitativen Merkmalen
`length`	gibt die Anzahl der Einträge eines Vektors aus
`barplot`	erzeugt ein Säulendiagramm
`plot`	„Standard“-Zeichen-Funktion in R, erzeugt *kontextabhängig* verschiedene grafische Darstellungen
`pie`	erzeugt ein Kreisdiagramm
`hist`	erzeugt (zeichnet) ein Histogramm, automatische Wahl der Klassenbreiten. Kann auch zum Zählen von Klassenhäufigkeiten genutzt werden.

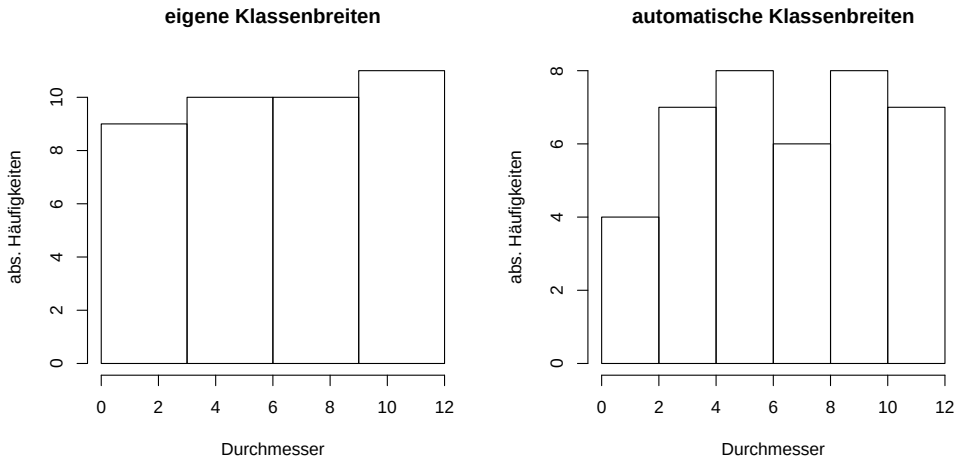


Abbildung 8: Histogramme zu Beispiel 2.13

3. Statistische Kennzahlen für Lage und Streuung

Ziel ist es Datensätze mittels weniger Kenngrößen zu beschreiben. Im Folgenden betrachten wir nur quantitative Merkmale, die Ausprägungen sind also Zahlen. Es sei stets eine Stichprobe

$$(x_1, \ldots, x_N)$$

gegeben. An verschiedenen Stellen werden wir mit der *geordneten Stichprobe* arbeiten, d.h., wir sortieren die beobachteten Ausprägungen nach ihrer Größe. Für die geordnete Stichprobe wird die Notation

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(N-1)} \leq x_{(N)}$$

verwendet.

Beispiel 3.1. Wir betrachten im Folgenden die originale (ungeordnete) Stichprobe, sowie darunter die geordnete Stichprobe.

i	1	2	3	4	5	6	7	8	9	10
x_i	9	8	15	7	2	1	9	9	9	6
$x_{(i)}$	1	2	6	7	8	9	9	9	9	15

Der R-Befehl zum Sortieren von Vektoren lautet `sort` . `sort`

Definition 3.2. In der geordneten Stichprobe bezeichnet $x_{(n)}$, $1 \leq n \leq N$ den *n-ten Rangwert*. $x_{(1)}$, $x_{(N)}$ heißen auch *Minimum* bzw. *Maximum* der Stichprobe, wir schreiben auch $x_{\min}$ bzw. $x_{\max}$. Die Differenz $R := x_{\max} - x_{\min}$ wird als *Spannweite* bezeichnet.

Minimum und Maximum geben nicht in allen Fällen einen sinnvollen Eindruck des Bereiches, in dem die Merkmalsausprägungen *üblicherweise* liegen. Z.B. kann es in einer Hockeymannschaft eine große, schussstarke Spielerin geben (also maximale Körpergröße 1,90 m); alle anderen Mitspielerinnen sind aber zwischen 1,50 und 1,75 m groß. Die Angabe eines *typischen* Bereiches, in dem ein Großteil der Ausprägungen liegt, ermöglichen Quantile: Ein 95%-Quantil einer Stichprobe ist eine Zahl, so dass 95% der beobachteten Werte unterhalb dieser Grenze liegen, und (nur) 5% oberhalb.

Definition 3.3. Für $p \in (0,1)$ ist das *p-Quantil* $\tilde{x}_p$ der Stichprobe $(x_1, \ldots, x_N)$ definiert durch

$$\tilde{x}_p \;=\; \begin{cases} x_{(k)}, & N \cdot p < k < N \cdot p + 1, \quad N \cdot p \notin \mathbb{N} \\ \frac{1}{2}\Big(x_{(k)} + x_{(k+1)}\Big), & k = N \cdot p \in \mathbb{N}. \end{cases}$$

Die p-Quantile für $p = \frac{1}{4}, \frac{1}{2}, \frac{3}{4}$ heißen *unteres Quartil, Median, oberes Quartil.*

Die Differenz

$$d_Q \;:=\; \tilde{x}_{.75} - \tilde{x}_{.25}$$

wird als Interquartilsabstand bezeichnet.

Zu Beginn der Vorlesung hatten wir Wachstumskurven für Säuglinge und Kleinkinder kennengelernt, die sog. Perzentilkurven. Hier ist Perzentil ein Synonym für Quantil. Für jede Altersstufe gibt also der Wert auf der P97-Kurve das 97%-Quantil der Körperlänge von Jungen an - 97% sind also kleiner oder höchstens so groß.

Beispiel 3.4. Wir berechnen unteres / oberes Quartil und Median für die Daten aus Beispiel 3.1. Wir hatten die folgende geordnete Stichprobe

i	1	2	3	4	5	6	7	8	9	10
$x_{(i)}$	1	2	6	7	8	9	9	9	9	15

Es ist $N = 10$. Unteres Quartil: $p = 0.25$, $N \cdot p = 2.5 \notin \mathbb{N}$, somit $\tilde{x}_{0.25} = x_{(3)} = 6$.
Median: $p = 0.5$, $N \cdot p = 5 \in \mathbb{N}$, somit

$$\tilde{x}_{0.5} \;=\; \frac{1}{2}\Big(x_{(5)} + x_{(6)}\Big) = \frac{1}{2}(8+9) \;=\; 8.5$$

Oberes Quartil: $p = 0.75$, $N \cdot p \notin \mathbb{N}$, somit $\tilde{x}_{0.75} = x_{(8)} = 9$.
Interquartilsabstand: $\tilde{x}_{.75} - \tilde{x}_{.25} = 9 - 6 = 3$.

Wollen wir obige Berechnungen in R durchführen, speichern wir zunächst die Stichprobe im Vektor `x` und lassen uns anschließend den Median und $\tilde{x}_{.4}$ ausgeben: `median`

```
> x<-c(1,2,6,7,8,9,9,9,9,15)
> median(x)
```

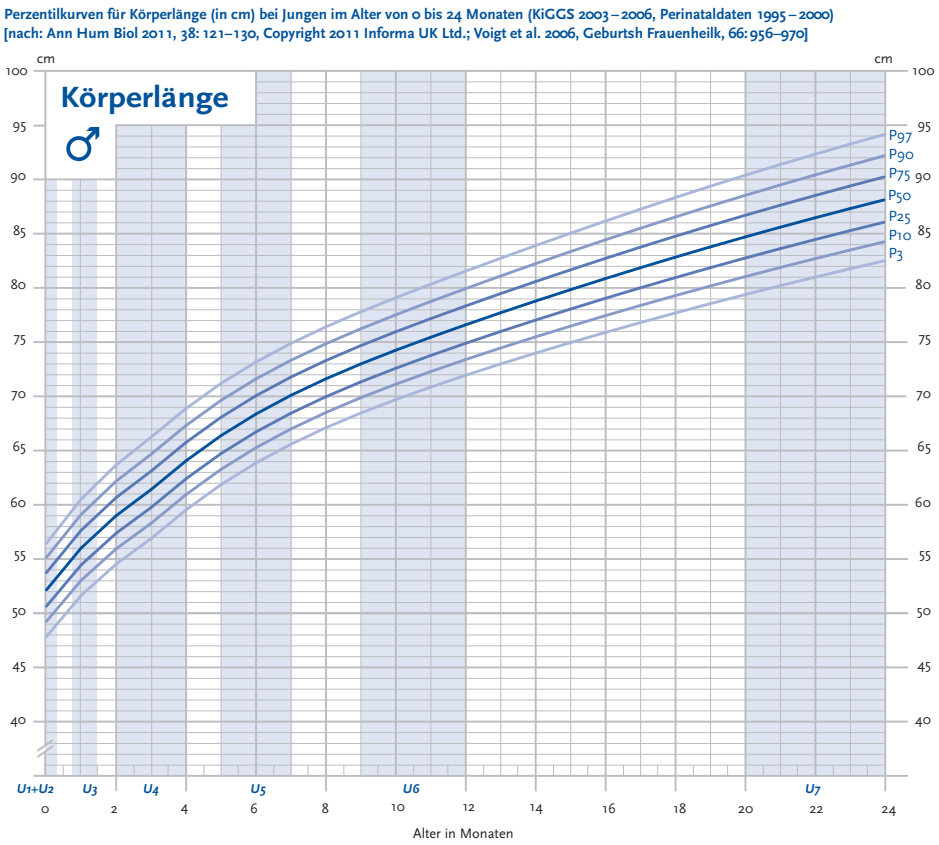


Abbildung 9: Perzentilkurven für die Körperlänge von Jungen 0-24 Monate, entnommen aus [4, S. 14]

```
[1] 8.5
> quantile(x, probs=0.4, type=2)
40%
7.5
```

Beim Aufruf `quantile` sorgt das Argument `type=2` dafür, dass die Quantile nach der obigen Regel bestimmt werden. Unter `probs=` muss der gewünschte Wert für p eingetragen werden. Der Befehl `quantile(x, type=2)` - also ohne Angabe eines Wertes für p, liefert die sogenannte *Fünf-Punkte-Zusammenfassung*, bestehend aus Minimum, unteren Quartil, Median, oberen Quartil und Maximum. Der Interquartilsabstand wird mit dem Befehl `IQR(x,type=2)` bestimmt.

`quantile`

`IQR`

Grafische Darstellung der Stichprobe - Box-Plot

Definition 3.5. Gegeben eine Stichprobe $(x_1, \ldots, x_N)$, bestimme $\tilde{x}_{.25}, \tilde{x}_{.5}, \tilde{x}_{.75}, d_Q$ sowie zusätzlich

- w_u kleinste Beobachtung, die größer ist als (unteres Quartil minus 1.5-facher Interquartilsabstand), d.h. $x_{(k)}$ mit $x_{(k-1)} \leq \tilde{x}_{.25} - 1.5 \cdot d_Q < x_{(k)}$
- w_o größte Beobachtung, die kleiner ist als (oberes Quartil plus 1.5-facher Interquartilsabstand), d.h. $x_{(k)}$ mit $x_{(k)} < \tilde{x}_{.75} + 1.5 \cdot d_Q \leq x_{(k+1)}$.

Trage diese Werte auf der y-Achse ab.

Zeichne eine Box von $\tilde{x}_{.25}$ bis $\tilde{x}_{.75}$ und einen waagerechten Strich auf der Höhe des Medians $\tilde{x}_{.5}$, anschließend waagerechte Striche bei w_u und w_o (den sog. *Whiskers*), diese werden mit der Box verbunden. Schließlich trage als Punkte alle Beobachtungen ein, die außerhalb von w_u oder w_o liegen.

Die Bedingungen für w_u und w_o sind durch Eigenschaften der Normalverteilung motiviert: Sind die Beobachtungen Realisierungen von standardnormalverteilten Zufallsvariablen, so würden je nur etwa 0.25% der Werte unterhalb bzw. oberhalb von w_u und w_o liegen.

Beispiel 3.6. Wir betrachten weiter den Datensatz aus Beispiel 3.1:

i	1	2	3	4	5	6	7	8	9	10
$x_{(i)}$	1	2	6	7	8	9	9	9	9	15

Wir hatten

$$\tilde{x}_{.25} = 6, \quad \tilde{x}_{.5} = 8.5, \quad \tilde{x}_{.75} = 9, \quad d_Q = 3.$$

Die relevante Grenzen für w_u und w_o sind somit:

$$\tilde{x}_{.25} - 1.5 \cdot d_Q \;=\; 6 - 1.5 \cdot 3 \;=\; 1.5; \qquad \tilde{x}_{.75} + 1.5 \cdot d_Q \;=\; 9 + 1.5 \cdot 3 \;=\; 13.5$$

Folglich

$$w_u = 2 = x_{(2)}, \qquad w_o = 9 = x_{(9)}$$

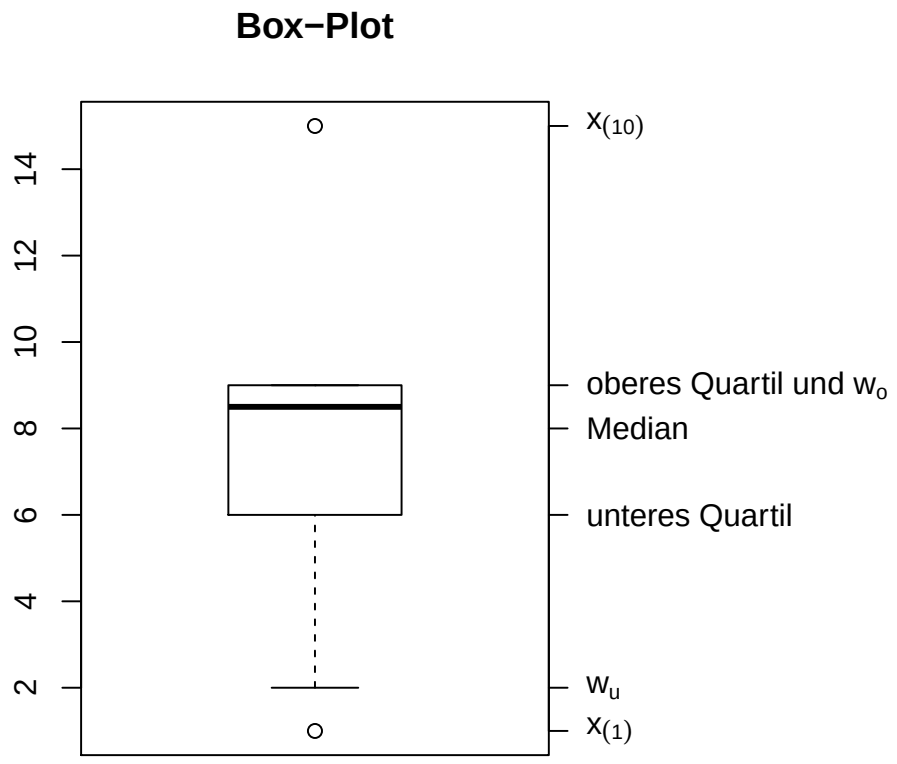


Abbildung 10: Box-Plot mit den Daten aus Beispiel 3.6.

Als Ausreißer (Werte kleiner als w_u oder größer als w_o) verbleiben

$$x_{(1)} = 1, \qquad x_{(10)} = 15.$$

Abbildung 10 zeigt den entsprechenden Box-Plot.

Beim Zeichnen von Hand dürfen Sie den Box-Plot auch waagerecht zeichnen (also die Werte auf der x-Achse abtragen). Der entsprechende R-Befehl lautet `boxplot(x)`, wenn der Vektor `x` die Beobachtungswerte enthält. Hierbei ist zu beachten, dass die R-Implementation mit etwas abgewandelten Definitionen arbeitet: statt der Quartile wird der Median der unteren bzw. oberen Hälfte der Beobachtungswerte verwendet (*left- bzw. right hinge*); die Werte unterscheiden sich aber nur marginal; da es beim Boxplot um einen qualitativen Eindruck geht, sind diese Unterschiede vernachlässigbar.

`boxplot`

Fortsetzung: Lage- und Streuungsparameter

Definition 3.7. Gegeben eine Stichprobe $(x_1, \ldots, x_N)$, bezeichnen wir

$$\bar{x} \;:=\; \frac{1}{N}\sum_{k=1}^{N} x_k \;=\; \frac{1}{N}\Big(x_1 + \cdots + x_N\Big)$$

als *Stichprobenmittel* oder *arithmetisches Mittel der Stichprobe.*

Bemerkung 3.8. Der Median ist robust gegenüber Ausreißern, das Stichprobenmittel nicht. Verändern wir im obigen Beispiel einen der Werte (durch falsches Übertragen?!) von 15 auf 150, so geschieht folgendes:

i	1	2	3	4	5	6	7	8	9	10	$\tilde{x}_{.5}$	$\bar{x}$
$x_{(i)}$	1	2	6	7	8	9	9	9	9	15	8.5	7.5
$x_{(i)}$	1	2	6	7	8	9	9	9	9	150	8.5	**21**

Das Stichprobenmittel (des Vektors x) berechnen wir in R mit dem Befehl `mean(x)`. `mean`

Definition 3.9. Gegeben eine Stichprobe $(x_1, \ldots, x_N)$, definieren wir:

$s := \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}(x_k - \bar{x})^2}$	empirische Standardabweichung
$s^2 = \frac{1}{N-1}\sum_{k=1}^{N}(x_k - \bar{x})^2$	Stichprobenvarianz

Die entsprechenden R-Befehle lauten `sd(x)` für die Standardabweichung, und `var(x)` `sd, var` für die Stichprobenvarianz.

Beispiel 3.10. Nutzen wir wieder die Beobachtungen zum Merkmal „Durchmesser“ aus Beispiel 1.6, so erhalten wir folgende Kennzahlen:

```
> quantile(Durchmesser,type=2)
   0%   25%   50%   75%  100%
 0.40  3.35  6.55  9.35 11.90
> IQR(Durchmesser,type=2)
[1] 6
> mean(Durchmesser)
[1] 6.42
> sd(Durchmesser)
[1] 3.37534
> mean(Durchmesser)
[1] 6.42
```

3.1. Kurz-Befehlsreferenz

`sort`	sortiert einen Vektor aufsteigend
`median`	bestimmt den Median
`quantile`	immer mit `type=2` verwenden; liefert bei Angabe von `probs=`p den Wert des p-Quantils; ohne weitere Angaben wird die Fünf-Punkte-Zusammenfassung ausgegeben
`IQR`	Interquartilsabstand. Beachte `type=2`.
`boxplot`	Zeichnet einen (oder mehrere, bei Angabe mehrerer Vektoren) Box-Plot
`mean`	berechnet das Stichprobenmittel / arithmetische Mittel eines Vektors
`sd`	Standardabweichung
`var`	Stichprobenvarianz (Quadrat der Standardabweichung)

4. Beschreibung und explorative Analyse bivariater Daten

Im Folgenden interessieren wir uns für die explorative Datenanalyse bivariater Daten. Wir unterscheiden, ob nur qualitative Merkmale, nur quantitative Merkmale oder beide Merkmalsarten zugleich beobachtet werden.

Den Fall bivariater qualitiativer Daten sparen wir hierbei aus, die entsprechende Darstellung mittels Kontingenztabellen werden wir im Kapitel über bedingte Wahrscheinlichkeiten nachholen.

4.1. Gemeinsame Beobachtung von qualitativen und quantitativen Merkmalen

Diese Situation liegt bspw. vor, wenn wir die Merkmale „Durchmesser“ und „Antibiotikaresistenz“ aus Beispiel 1.6 betrachten. Eine typische Fragestellung ist folgende: Gruppiert man die Beobachtungen anhand der Ausprägungen des qualitativen Merkmals, und bestimmt pro Gruppe die zuvor eingeführten Kennzahlen (Stichprobenmittel, Standardabweichung, ...); unterscheiden sich diese Kennzahlen? [Sind bspw. die Kolonien mit hoher Antibiotikaresistenz „im Mittel“ größer als solche mit geringer Antibiotikaresistenz?]

Beispiel 4.1. Die erste Möglichkeit, entsprechende Vergleiche durchzuführen, besteht darin, zunächst die Stichprobe in die oben genannten Gruppen aufzuspalten. Dazu werden neue Variablen eingeführt, die die jeweiligen Teildatensätze enthalten, z.B. erzeugt

```
Durchmesser.sensitiv<-Durchmesser[Resistenz=="sensitiv"]
```

einen Vektor, der Beobachtungen des Merkmals Durchmesser an denjenigen statistischen Einheiten enthält, bei denen das Merkmal Resistenz die Ausprägung „sensitiv“ aufweist. Gleiches führt man für die weiteren Stufen des Merkmals Antibiotikaresistenz („intermediär“, „resistent“); und bestimmt dann für die drei neuen Vektoren die jeweiligen Kennzahlen, und kann Box-Plots zeichnen.

Die zweite Möglichkeit ist etwas weniger „robust“, dafür effektiver:
`boxplot(Durchmesser~Resistenz)` erlaubt den Vergleich der Boxplots zu den Teilbeobachtungen des Merkmals „Durchmesser“, aufgeteilt nach den Ausprägungen des Merkmals Resistenz. Ein etwas länglicher Befehl erlaubt den automatischen Vergleich der Mittelwerte:

```
> model.tables(aov(Durchmesser~Resistenz), "means")
Tables of means
Grand mean

6.42

 Resistenz
    sensitiv intermediär resistent
       3.643       10.93     8.445
rep   21.000        8.00    11.000
```

Grand mean bezeichnet hier das arithmetische Mittel aller Beobachtungen, in der Tabelle sind anschließend wieder die Mittelwerte der Teilpopulationen aufgelistet, und unter *rep* wird angegeben, wieviele statistische Einheiten zu der entsprechenden Gruppe gehören - es gab also bspw. 21 Kolonien, die sensitiv auf Antibiotika reagieren, der mittlere Durchmesser dieser 21 Kolonien ist 3.643.

4.2. Bivariate quantitative Merkmale

Im Folgenden sei stets eine Stichprobe vom Umfang N gegeben, die nun aus Beobachtungspaaren

$$(x_1, y_1), \ldots, (x_N, y_N)$$

besteht.

Definition 4.2. Die Darstellung der Messwerte $(x_1, y_1), \ldots, (x_N, y_N)$ im $x-y$-Koordinatensystem heißt *Streudiagramm*.

Beispiel 4.3. An 20 Flüssen wurden die Sauerstoffkonzentration (in mg/l), die Fließgeschwindigkeit (in m/s) und die Wassertemperatur (in °C) gemessen. Die beobachteten Werte sind in gleichnamigen Vektoren abgelegt. Der R-Befehl

```
plot(Sauerstoff,Fliessgeschwindigkeit)
```

zeichnet ein Streudiagramm dieser beiden Merkmale. Sind alle Beobachtungen in einer Datentablle (bspw. `data.frame` „Wasser“) hinterlegt, so zeichnet der Aufruf `plot(Wasser)` Streudiagramme für jede mögliche Paarung.

Zwischen Sauerstoffkonzentration und Fließgeschwindigkeit scheint ein (positiver) linearer Zusammenhang zu bestehen, wohingegen zwischen Sauerstoffkonzentration und Wassertemperatur kein Zusammenhang erkennbar ist.

	Sauerstoff	Fließgeschwindigkeit	Wassertemperatur
1	12.1	0.90	10.4
2	2.9	0.27	11.3
3	5.8	0.37	13.2
4	8.1	0.57	16.2
5	11.3	0.91	9.8
6	0.9	0.20	16.1
7	9.9	0.90	16.5
8	11.0	0.94	13.9
9	7.9	0.66	13.7
10	10.0	0.63	8.6
11	1.5	0.06	9.9
12	1.7	0.21	9.6
13	1.0	0.18	14.2
14	8.0	0.69	11.5
15	4.3	0.38	14.9
16	8.8	0.77	12.5
17	6.3	0.50	14.5
18	7.7	0.72	16.9
19	12.3	0.99	11.4
20	3.3	0.38	15.0

Tabelle 2: Messwerte zu Beispiel 4.3

Definition 4.4. Gegeben Datenpaare $(x_1, y_1), \ldots, (x_N, y_N)$, ist der *(Bravais-Pearson)-Korrelationskoeffizient* definiert durch

$$r := \frac{\sum_{k=1}^{N}(x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{N}(x_k - \bar{x})^2 \sum_{k=1}^{N}(y_k - \bar{y})^2}}$$

Der Wertebereich ist $-1 \leq r \leq 1$,

- $r > 0$ gleichsinniger linearer Zusammenhang. Tendenz: Werte (x_i, y_i) liegen um eine Gerade positiver Steigung
- $r < 0$ gegensinninger linearer Zusammenhang. Tendenz: Werte (x_i, y_i) liegen um eine Gerade negativer Steigung
- $r = 0$ kein linearer Zusammenhang

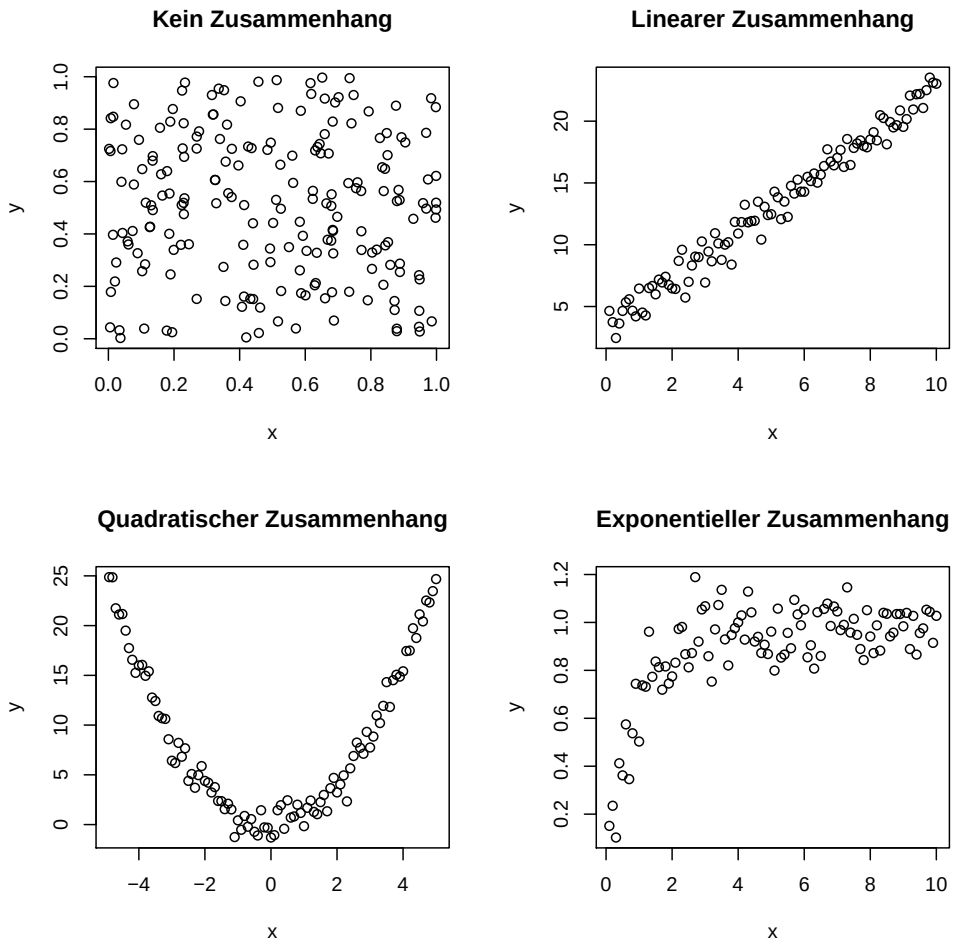


Abbildung 11: Generische Darstellung verschiedener Zusammenhangstypen im Streudiagramm

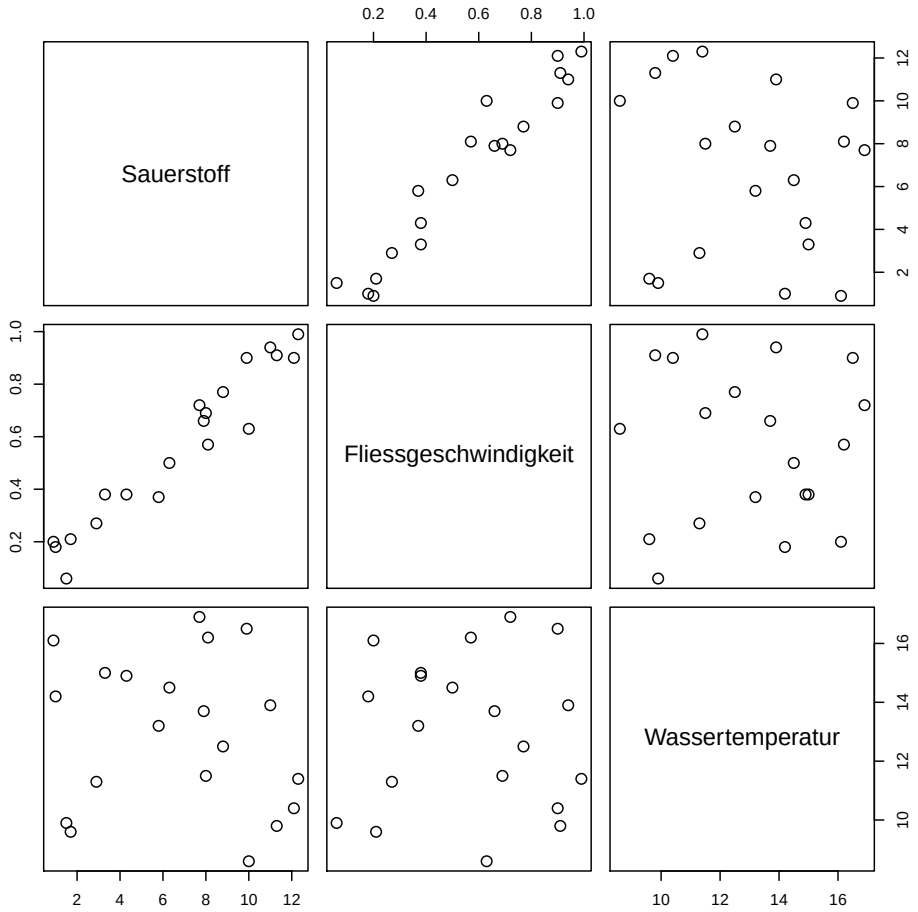


Abbildung 12: Streudiagramme zu Beispiel 4.3

Bemerkung 4.5. Der Korrelationskoeffizient beschreibt nur *lineare* Zusammenhänge! Würden (x_i, y_i) bspw. auf einer perfekten Parabel liegen, wäre dennoch $r = 0$, obwohl offensichtlich ein Zusammenhang besteht. Der Zahlenwert von r hat keine absolute Interpretation, eine grobe Orientierung ist aber wie folgt:

$\lvert r\rvert < 0.5$	schwache Korrelation
$0.5 \leq \lvert r\rvert < 0.8$	mittlere Korrelation
$\lvert r\rvert > 0.8$	starke Korrelation

Beispiel 4.6. Mit den Daten aus Beispiel 4.3 ergibt sich als Korrelationskoeffizient zwischen Sauerstoffkonzentration und Fließgeschwindigkeit `cor`

```
cor(Sauerstoff,Fliessgeschwindigkeit)
[1] 0.9653202
```

sowie zwischen Sauerstoffkonzentration und Wassertemperatur

```
cor(Sauerstoff,Wassertemperatur)
[1] -0.1504769
```

Bemerkung 4.7. Es gilt

$$r = \frac{\sum_{k=1}^{N} x_k y_k - N\bar{x}\bar{y}}{\sqrt{\left(\sum_{k=1}^{N} x_k^2 - N\bar{x}^2\right)\left(\sum_{k=1}^{N} y_k^2 - N\bar{y}^2\right)}}$$

Beispiel 4.8. Eine Kinderpsychologin vermutet, dass sich häufiges Fernsehen negativ auf das Tiefschlafverhalten von Kindern auswirkt. Um dieser Frage nachzugehen, wurden folgende Daten erhoben:

Kind i	1	2	3	4	5	6	7	8	9
Fernsehzeit x_i	0.3	2.2	0.5	0.7	1.0	1.8	3.0	0.2	2.3
Dauer Tiefschlaf y_i	5.8	4.4	6.5	5.8	5.6	5.0	4.8	6.0	6.1

Als Hilfsgrößen bestimmen wir

$$\sum_{k=1}^{N} x_k y_k = 62.96, \quad \bar{x} = 1.3\bar{3}, \quad \bar{y} = 5.5\bar{5}, \quad \sum_{k=1}^{N} x_k^2 = 24.24, \quad \sum_{k=1}^{N} y_k^2 = 281.5$$

und erhalten dann

$$r = \frac{62.96 - 9 \cdot 1.3\bar{3} \cdot 5.5\bar{5}}{\sqrt{(24.24 - 9 \cdot 1.7\bar{7})(281.5 - 9 \cdot 30.86)}} = -0.67$$

Bemerkung 4.9. Korrelation ist nicht mit Kausalität zu verwechseln. Korrelation ist nur ein Indikator auf einen möglichen Kausalzusammenhang. Der Korrelationskoeffizient gibt keine Auskunft über die Richtung einer Beeinflussung.

Beispiele für Korrelationen

- Die Zahl der Klapperstörche ist hoch mit den bundesdeutschen Geburten korreliert.
- Der Konsum von Südfrüchten ist positiv mit der deutschen Staatsverschuldung korreliert.
- Das Auftreten von Heuschnupfen ist negativ mit dem Weizenpreis korreliert.

Bei den oben angegebenen Beispielen handelt es sich um Beispiele der Auswirkungen einer vernachlässigten Hintergrundvariablen. Die Korrelation zwischen den Merkmalen X und Y lässt sich gegebenenfalls auf ein Merkmal Z zurückführen, das u. U. nicht erhoben wurde und X und Y beeinflusst.

Beispiel 4.10. Obwohl eine Korrelation vorliegt, kann die Korrelation verschwinden, wenn eine wichtige Variable übersehen wird: Ist der Zigarettenkonsum über lange konstant, so ist die Korrelation zwischen Zigarettenkonsum und Zeit 0. Dabei könnte sich nur der Effekt zwischen den Geschlechtern ausgleichen: Während der Zigarettenkonsum in der Gruppe der Frauen steigt, sinkt er in der Gruppe der Männer.

Beispiel 4.11. Selbst wenn ein Kausalzusammenhang vorliegt, ist es nicht klar, in welche Richtung er wirkt: Auf den Neuen Hebriden hielt sich einige Zeit der Aberglaube, Läuse vertrieben Krankheiten. Läuse und Gesundheit traten gehäuft zusammen auf: Gesunde Insulaner hatten Läuse, Kranke keine. Dabei vertrieben nicht die Läuse die Krankheit, sondern die Krankheit die Läuse.

Literaturtipp: Kraemer, So lügt man mit Statistik. [2]

4.3. Lineare Regression

Vermuten wir einen linearen Zusammenhang zwischen den beobachteten Daten - die Beobachtungspaare liegen also tendenziell auf einer Geraden, so sind wir daran interessiert, diese *Ausgleichsgerade* zu bestimmen.

Definition 4.12. Gegeben eine Stichprobe $(x_1, y_1), \ldots, (x_N, y_N)$, so heißt

$$y_k \;=\; \alpha + \beta x_k + \epsilon_k, \qquad 1 \leq k \leq N$$

lineare Einfachregression, wobei α den Achsenabschnitt, β den Steigungsparameter und ϵ_k den Fehler in der k-ten Beobachtung bezeichnen.

Beispiel 4.13. Wir zeichnen das Streudiagramm zu den Daten aus Beispiel 4.8. Nun wollen wir eine Gerade einzeichnen, die die Tendenz der Daten möglichst gut beschreibt...

Was ist nun eine gute Gerade?

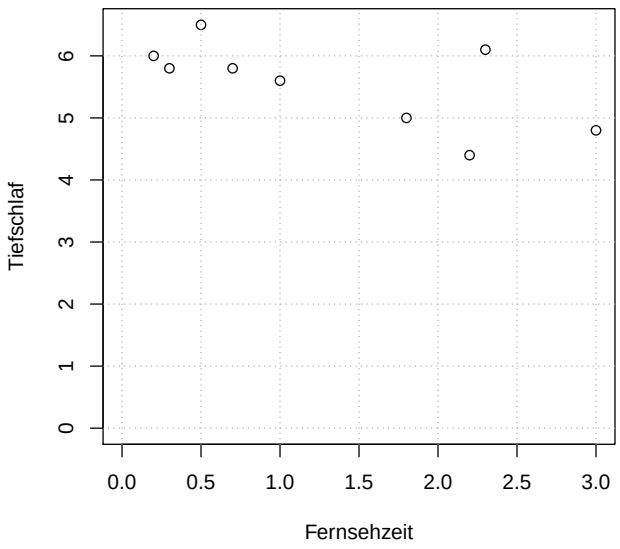


Abbildung 13: Streudiagramm zu den Daten aus Beispiel 4.8

Kriterium (Kleinste-Quadrate-Methode): Bestimme α und β so, dass der mittlere, quadrierte Fehler minimal wird, d.h., finde α und β so, dass

$$Q(\alpha,\beta) \;:=\; \frac{1}{N}\sum_{k=1}^{N}\Big(y_k-(\alpha+\beta x_k)\Big)^2$$

minimal wird.

Beispiel 4.14. Raten wir in obiger Situation $a = 6.5$ und $b = -1/3$, so erhalten wir

```
a=6.5
b=-1/3
mean((y-(a+b*x))^2)
[1] 0.4907407
```

Für die optimalen Werte $a = 6.15$ und $b = -0.45$ gilt

```
> a=6.15
> b=-0.45
>
> mean((y-(a+b*x))^2)
[1] 0.2283444
```

Die optimalen Werte werden mit folgender Formel bestimmt:

Satz 4.15. *In der Situation der linearen Einfachregression (Def. 4.12) sind die* Kleinste-Quadrate-Schätzer *für α und β gegeben durch*

$$\hat{\alpha} \;=\; \bar{y} - \hat{\beta}\bar{x},$$

$$\hat{\beta} \;=\; \frac{\sum_{k=1}^{N} x_k y_k - N\bar{x}\bar{y}}{\sum_{k=1}^{N} x_k^2 - N\bar{x}^2} \;=\; \frac{\sum_{k=1}^{N}(x_k - \bar{x})(y_k - \bar{y})}{\sum_{k=1}^{N}(x_k - \bar{x})^2}$$

In R werden die Parameter einer Regressionsgerade mit Hilfe des Aufrufs `lm` (für linear model) bestimmt: `lm`

```
> lm(Tiefschlaf~Fernsehzeit)$coefficients
(Intercept) Fernsehzeit
  6.1553398  -0.4498382
```

Unter (Intercept) steht der Schätzwert für den Achsenabschnitt, der zweite Wert ist die Steigung (=der Koeffizient des Wertes der Fernsehzeit).

```
abline(coef=lm(Tiefschlaf~Fernsehzeit)$coefficients)
```

zeichnet die Regressionsgerade in das Streudiagramm. **Nachtrag:** Es funktioniert auch der kürzere Befehl `abline`

```
abline(lm(Tiefschlaf~Fernsehzeit))
```

4.4. Nichtlineare Zusammenhänge

Neben linearen Zusammenhängen zwischen beobachteten Merkmalen können auch anders geartete funktionale Zusammenhänge auftreten, bspw. quadratisch oder exponentiell. In manchen Fällen können die Daten transformiert werden, so dass ein linearer Zusammenhang entsteht, dann können die Parameter des Modells wieder mit Hilfe der linearen Regression geschätzt werden[2]

Im Folgenden wird dargestellt, bei welchen (vermuteten) Zusammenhängen eine Transformation auf lineare Zusammenhänge möglich ist:

Vermutl. Zushg.	Transformation	Lin. Zushg.	Interpretation
$y_k = a \cdot \exp(b \cdot x_k)$	$z_k = \ln(y_k)$	$z_k = \ln(a) + b \cdot x_k$	$\alpha = \ln(a)$, $\beta = b$
$y_k = c \cdot x^d$	$z_k = \ln(y_k)$, $v_k = \ln(x_k)$	$z_k = \ln(c) + d \cdot v_k$	$\alpha = \ln(c)$, $\beta = d$
$y_k = e + f \cdot x^2$	$v_k = x_k^2$	$y_k = e + f \cdot v_k$	$\alpha = e$, $\beta = f$

[2]Ist solch eine Transformation nicht möglich, befinden wir uns im Bereich der nichtlinearen Regression.

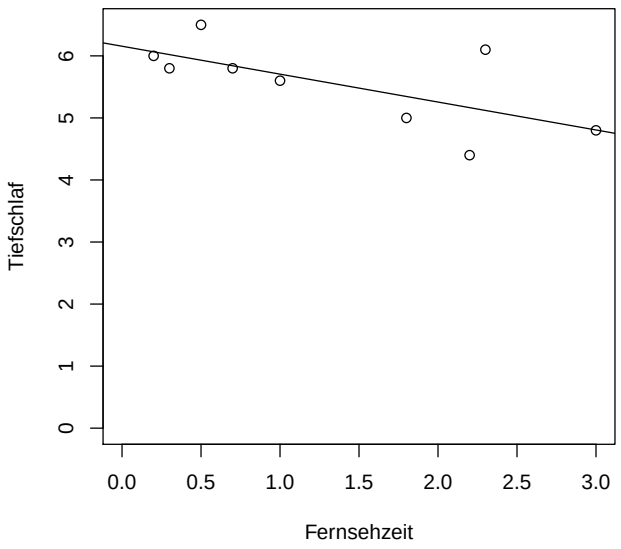


Abbildung 14: Streudiagramm zu den Daten aus Beispiel 4.8 mit Regressionsgerade

Beispiel 4.16. Sie vermuten bei einem Datensatz, bestehend aus den Beobachtungsvektoren `x` und `y` einen exponentiellen Zusammenhang der Form

$$y_k = a \cdot \exp(b \cdot x_k).$$

Gemäß obiger Tabelle definieren Sie einen neuen Vektor `z<-log(y)` (das ist in R der natürliche Logarithmus zur Basis e) und rufen folgenden Befehl auf:

```
lm(z~x)
```

Der Wert unter (Intercept) ist dann der Schätzwert für $\ln(a)$, der Wert des zweiten Koeffizienten ist dann der Schätzwert für b.

4.5. Kurz-Befehlsreferenz

Gegeben Beobachtungen `x` eines qualitativen Merkmals und `y` eines quantitativen Merkmals, zeichnet

```
boxplot(y~x)
```

die Boxplots der anhand des qualitativen Merkmals gruppierten Daten (in eine Grafik). In derselben Situation gibt

```
model.tables(aov(y~x), "means")
```

das Gesamt-Mittel (*grand mean*) und die arithmetischen Mittelwerte (des Merkmals `y`) der (anhand des qualitativen Merkmals `x` gebildeten) Gruppen aus.

`plot(x,y)`	zeichnet ein Streudiagramm
`plot(DF)`	ist `DF` ein `data.frame`, so werden Streudiagramme aller möglichen Paarungen von beobachteten Merkmalen, die in `DF` hinterlegt sind, gezeichnet
`cor`	berechnet den Bravais-Pearson-Korrelationskoeffizienten
`lm`	berechnet Koeffizienten der Regressionsgeraden
`abline`	zeichnet eine Gerade in eine bestehende Grafik

Teil II.
Grundlagen der Wahrscheinlichkeitstheorie

Idee: Beobachtete Daten sind Resultat von Zufallsmechanismen. Der Zufall kommt ins Spiel bspw. durch natürliche Schwankungen von Merkmalsausprägungen (Körpergröße), Messfehler (physikalische Experimente), zufällige Stichproben, ...

Um Aussagen treffen zu können, die über die bloße Beschreibung der Daten hinausgehen, benötigen wir mathematische Modelle zur Beschreibung zufälliger Phänomene.

5. Grundbegriffe und Kombinatorik

5.1. Grundbegriffe

Definition 5.1.

Ergebnisraum (Grundraum)	Menge Ω aller möglichen Ergebnisse eines Zufallsvorgangs
Ergebnisse	Elemente $\omega \in \Omega$
Ereignis	Teilmenge $A \subset \Omega$
Elementarereignis	Ereignis der Form $A = \{w\}$

Beispiel 5.2.

(i) Würfelwurf:

$$\begin{aligned} \Omega &= \{1,2,3,4,5,6\} \\ \omega \in \Omega &: \text{gewürfelte Augenzahl} \\ A = \{2,4,6\} &: \text{Ereignis „gerade Augenzahl“} \\ B = \{6\} &: \text{Elementarereignis „Augenzahl ist 6“} \end{aligned}$$

(ii) Dreifacher Würfelwurf:

$$\begin{aligned} \Omega &= \Big\{(\omega_1,\omega_2,\omega_3) : \omega_i \in \{1,2,\dots,6\}\Big\} \\ \omega = (\omega_1,\omega_2,\omega_3) \in \Omega &: \text{gewürfelte Augenzahlen} \\ A = \{(\omega_1,\omega_2,\omega_3) \in \Omega : \omega_3 = 6\} &: \text{Ereignis „Augenzahl 6 im dritten Wurf“} \\ B = \{(\omega_1,\omega_2,\omega_3) \in \Omega : \omega_1+\omega_2+\omega_3 \geq 12\} &: \text{Ereignis „Summe der Augenzahlen mindestens 12“} \end{aligned}$$

(iii) Geschlecht eines Neugeborenen:

$$\begin{aligned} \Omega &= \{m, w, d\} \\ A = \{w\} &: \text{Elementarereignis „Es ist ein Mädchen“} \end{aligned}$$

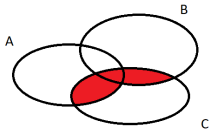
(iv) Körpergröße einer zufällig ausgewählten Person (in cm):

$$\begin{aligned} \Omega &= [0, \infty) \\ \omega \in \Omega &: \text{Körpergröße der Person} \\ A = [0, 180] &: \text{Ereignis „Person ist höchstens 180 cm groß“} \\ B = (160, \infty) &: \text{Ereignis „Person ist größer als 160 cm“} \end{aligned}$$

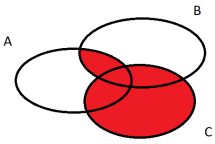
Verknüpfungen von Ereignissen (5.3)

Siehe Tabelle 3.
Außerdem gelten die folgenden Rechenregeln:

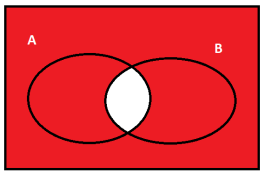
1. Distributivgesetz: $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$



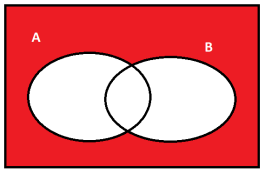
2. Distributivgesetz: $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$



1. Regel von de Morgan: $(A \cap B)^c = A^c \cup B^c$



2. Regel von de Morgan: $(A \cup B)^c = A^c \cap B^c$



Realität / Interpretation	Math. Modell	Venn-Diagramm
Ereignis E_1 *oder* Ereignis E_2 tritt ein	$E_1 \cup E_2$	
Ereignis E_1 *und* Ereignis E_2 tritt ein	$E_1 \cap E_2$	
Ereignis E tritt *nicht* ein	$E^c = \Omega \setminus E$	
Ereignis E_1 tritt ein, Ereignis E_2 aber nicht	$E_1 \cap E_2^c$	
das Eintreten von Ereignis E_1 impliziert das Eintreten von Ereignis E_2	$E_1 \subset E_2$	
die Ereignisse E_1 und E_2 sind unverträglich (**disjunkt**)	$E_1 \cap E_2 = \emptyset$	
eines der beiden unverträglichen Ereignisse E_1 und E_2 tritt ein	$E_1 + E_2$	
mindestens eines der Ereignisse E_i, $i \geq 1$, tritt ein	$\bigcup_{i\geq 1} E_i$	
alle Ereignisse E_i, $i \geq 1$, treten ein	$\bigcap_{i\geq 1} E_i$	

Tabelle 3: Sprechweisen für die Verknüpfung von Ereignissen

Beispiel 5.4. Zweifacher Würfelwurf:

$$\Omega \;=\; \Big\{(\omega_1,\omega_2) \,:\, \omega_i \in \{1,2,\ldots,6\}\Big\}$$

Ereignis A: „Summe der Augenzahlen ist kleiner oder gleich 3“ ist gegeben durch

$$A \;=\; \{(1,1),(1,2),(2,1)\},$$

das Ereignis B: „Erster Würfel zeigt Augenzahl 2“ ist gegeben durch

$$B \;=\; \Big\{(\omega_1,\omega_2) \in \Omega \,:\, \omega_1 = 2\Big\} \;=\; \Big\{(2,1),(2,2),(2,3),(2,4),(2,5),(2,6)\Big\}$$

Dann

$$\begin{aligned}
A \cap B &= \{(2,1)\} \\
A \cup B &= \Big\{(1,1),(1,2),(2,1),(2,2),(2,3),(2,4),(2,5),(2,6)\Big\} \\
A \setminus B &= \Big\{(1,1),(1,2)\Big\} \\
B^c &= \Big\{(\omega_1,\omega_2) \in \Omega \,:\, \omega_1 \neq 2\Big\}
\end{aligned}$$

Nächster Schritt: Ordne jedem Ereignis A eine Wahrscheinlichkeit $P(A)$ zu. Wie kann dies sinnvoll geschehen? Zwei Wege:

1. Wahrscheinlichkeiten ergeben sich aus empirischen Beobachtungen.
   Idee: Betrachte Zufallsexperiment, dass sich im Prinzip beliebig oft unabhängig voneinander unter identischen Bedingungen wiederholen lässt. Dann wird $P(A)$ festgelegt durch die relative Häufigkeit des Auftretens von A bei sehr vielen Wiederholungen des Experiments. Formal:
   $$P(A) \;=\; \lim_{N\to\infty} \frac{h_N(A)}{N},$$
   wobei $h_N(A)$ die absolute Häufigkeit des Auftretens von A bei N Durchführungen des Experimentes bezeichnet.

   Hierunter fällt auch die Situation, dass in einer großen Population die Anteile bestimmter Merkmalsausprägungen bekannt sind. Wird dann eine Person zufällig aus dieser Population ausgewählt, so entspricht die Wahrscheinlichkeit, bei dieser Person eine gewisse Merkmalsausprägung vorzufinden, gerade dem Anteil der Ausprägung in der Gesamtpopulation. **Beispiel**: In einer Schafherde mit 1000 Schafen gibt es 20 schwarze Schafe. Wie hoch ist die Wahrscheinlichkeit, dass ein zufällig (blind ;-) ausgewähltes Schaf ein schwarzes Schaf ist? Sie würde $20/1000 = 2\%$ betragen.

2. Wahrscheinlichkeiten ergeben sich aus theoretischen Überlegungen.
   In vielen Fällen ist die Annahme gerechtfertigt, dass alle möglichen Ergebnisse die gleiche Wahrscheinlichkeit haben (Laplace-Experiment, s.u.); dies trifft bspw. auf den fairen Würfel- oder Münzwurf zu. Ebenso kann ein Glücksrad mit verschieden großen Feldern gegeben sein; die Wahrscheinlichkeit eines Feldes würde dann als proportional zum Öffnungswinkel angenommen.

Für die mathematische Betrachtung ist die „Herkunft“ der Wahrscheinlichkeitswerte nicht wichtig, es wird nur festgehalten, welche Eigenschaften für das Rechnen mit Wahrscheinlichkeiten gelten müssen.

Definition 5.5. Sei Ω ein Ergebnisraum und $\mathcal{A}$ die Menge aller beobachtbaren Ereignisse[3] über Ω. Ein *Wahrscheinlichkeitsmaß* (bzw. eine *Wahrscheinlichkeitsverteilung*) ist eine Abbildung

$$P : \mathcal{A} \to [0,1]; \quad A \mapsto P(A)$$

mit folgenden Eigenschaften:

(i) $P(\emptyset) = 0$, $P(\Omega) = 1$,

(ii) $P(A \cup B) = P(A) + P(B)$ falls A und B disjunkt

(iii) $P\Big(\bigcup_{i\geq 1} A_i\Big) = \sum_{i=1}^{\infty} P(A_i)$ falls $A_1, A_2, A_3, \ldots$ paarweise disjunkt

$P(A)$ heißt *Wahrscheinlichkeit des Ereignisses A*.
$(\Omega, \mathcal{A}, P)$ heißt *Wahrscheinlichkeitsraum*.

Satz 5.6 (Eigenschaften von Wahrscheinlichkeitsmaßen)**.** *Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, und $A, B, C \in \mathcal{A}$ Ereignisse. Dann gilt:*

(i) $P(A^c) = 1 - P(A)$

(ii) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

(iii) $A \subset B \Rightarrow P(A) \leq P(B)$

(iv) $P(B \setminus A) = P(B) - P(A \cap B)$

(v) „Siebformel“:

$$\begin{aligned} P(A \cup B \cup C) \;=\; & P(A) + P(B) + P(C) \\ & - P(A \cap B) - P(B \cap C) - P(A \cap C) \\ & + P(A \cap B \cap C) \end{aligned}$$

Aufgabe: Zeichnen Sie Venn-Diagramme zu den Situationen (ii) - (v).

Beispiel 5.7. 41% der Bevölkerung haben die Blutgruppe 0, 85% der Bevölkerung haben den Rhesusfaktor positiv, 35% der Bevölkerung haben das Merkmal 0 positiv.

Wie hoch ist die Wahrscheinlichkeit, dass eine zufällig ausgewählte Person Blutgruppe 0 oder Rhesusfaktor positiv hat (also eines oder sogar beide Merkmale)?

Lösung:

Schritt 1: Führe geeignete Ereignisse ein. Ω=Gesamtbevölkerung.

A: Person hat Blutgruppe 0

B: Person hat Rhesusfaktor positiv

Schritt 2: Interpretiere die im Text genannten Anteile als Wahrscheinlichkeiten:

$$\begin{aligned} P(A) &= 41\% &&= 0.41 \\ P(B) &= 85\% &&= 0.85 \\ P(A \cap B) &= 35\% &&= 0.35 \end{aligned}$$

Schritt 3: Gesucht ist $P(A \cup B)$. Wende Rechenregel 5.6 (ii) an:

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= 0.41 + 0.85 - 0.35 \;=\; 0.91 \;=\; 91\% \end{aligned}$$

Definition 5.8. Ist $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum und hat Ω endlich viele oder abzählbar unendlich viele Elemente, dann spricht man von einem *diskreten Wahrscheinlichkeitsraum*. Für diskrete Wahrscheinlichkeitsräume gilt:
P ist bereits eindeutig bestimmt durch die Wahrscheinlichkeiten

$$p(\omega) \;:=\; P\big(\{\omega\}\big), \quad \omega \in \Omega,$$

denn für alle Ereignisse $A \subset \Omega$ gilt:

$$P(A) \;=\; \sum_{\omega \in A} p(\omega).$$

$p : \Omega \to [0, 1]$ heißt Wahrscheinlichkeitsfunktion. Beachte:

$$\sum_{\omega \in \Omega} p(\omega) \;=\; P(\Omega) \;=\; 1.$$

5.2. Laplace-Experimente

Definition 5.9. Sei Ω endlich. Dann heißt das durch

$$P(A) \;=\; \frac{\#A}{\#\Omega}, \qquad A \subset \Omega$$

definierte Wahrscheinlichkeitsmaß *Laplace-Verteilung* bzw. *diskrete Gleichverteilung* auf Ω, und $(\Omega, \mathcal{A}, P)$ *Laplaceraum*.

Hierbei bezeichnet $\#A$ die Anzahl der Elemente in A. Eine Laplace-Verteilung ordnet also jedem Ereignis eine Wahrscheinlichkeit entsprechend seiner relativen Größe (bezogen auf Ω) zu. Insbesondere gilt:

$$p(\omega) \;=\; P\big(\{\omega\}\big) \;=\; \frac{1}{\#\Omega},$$

d.h. **in einem Laplace-Experiment sind alle Elementarereignisse gleich wahrscheinlich.**

Beispiel 5.10. Der Ergebnisraum beim Würfelwurf ist

$$\Omega \;=\; \big\{1,2,3,4,5,6\big\}.$$

Beim *fairen* Würfel ist jede Augenzahl gleichwahrscheinlich, d.h. P ist die Gleichverteilung / Laplace-Verteilung auf Ω.

p(1)	p(2)	p(3)	p(4)	p(5)	p(6)
$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Mit den folgenden R-Befehlen zeichnen wir ein Stabdiagramm der Wahrscheinlichkeitsfunktion:

```
omega<-1:6
p<-rep(1/6,6)
plot(omega,p,type="h", ylim=c(0,0.5), xlab=expression(omega),
 ylab=expression(p(omega)), main="Wahrscheinlichkeitsfunktion",yaxt="n")
axis(2,at=c(0,1/6),labels=c(0,expression(1/6)))
```

Die Funktion `expression(...)` interpretiert ihr Argument, soweit möglich, als mathematischen Ausdruck und ersetzt bspw. `omega` durch den griechischen Buchstaben ω. Eine Übersicht der von R erkannten Notationen liefert der Hilfe-Aufruf `?plotmath` . Das Argument `yaxt="n"` verhindert das Zeichnen der y-Achsenbeschriftungen, die werden mit dem anschließenden Aufruf `axis(...)` von Hand an den gewünschten Stellen gesetzt.

`expression` `?plotmath` `axis`

Simulation endlicher Wahrscheinlichkeitsverteilungen (5.11)

Mit Hilfe des Befehls `sample` können Würfelwürfe (oder beliebige andere Zufallsexperimente mit nur endlich vielen möglichen Ergebnissen) in R simuliert werden. Als

`sample`

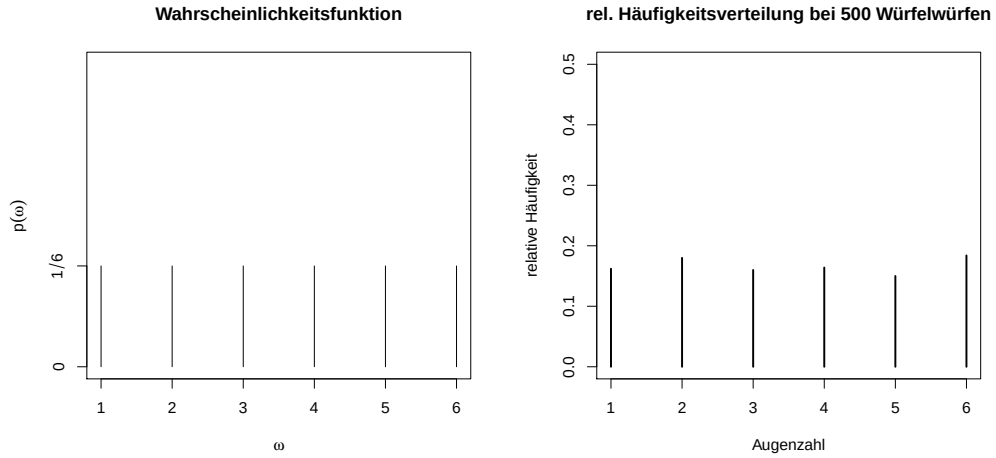


Abbildung 15: Wahrscheinlichkeitsfunktion des fairen Würfelwurfs und rel. Häufigkeitsverteilung bei 500 simulierten Würfen

Information benötigt die Funktion den Ergebnisraum (als Vektor der möglichen Ergebnisse); und die Anzahl der gewünschten Wiederholungen. Der folgende Aufruf simuliert 500 Würfelwürfe und speichert die Ausgänge im Vektor `x`. Anschließend zeichnen wir ein Stabdiagramm (vgl. 2.3).

```
x<-sample(omega, 500, replace=TRUE)
plot(table(x)/length(x), type="h", ylim=c(0,0.5), xlab="Augenzahl",
 ylab="relative Häufigkeit",main="rel. Häufigkeitsverteilung bei 500 Würfelwürfen")
```

Das Argument `replace=TRUE` bedeutet Ziehen mit Zurücklegen, dazu später mehr. Ohne weitere Angaben nimmt die Funktion `sample` immer eine Gleichverteilung an.

Beispiel 5.12. Als weiteres Beispiel betrachten wir das Werfen eines gezinkten Würfels mit Wahrscheinlichkeitsfunktion

p(1)	p(2)	p(3)	p(4)	p(5)	p(6)
$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{2}{7}$

Möchten wir diesen Würfel simulieren, müssen wir der Funktion `sample` die Wahrscheinlichkeitsfunktion explizit mit angeben. Das geschieht mit dem optionalen Argument `prob=...`:

```
p.gezinkt<-c(1/7,1/7,1/7,1/7,1/7,2/7)
x<-sample(omega, 500, replace=TRUE, prob=p.gezinkt)
```

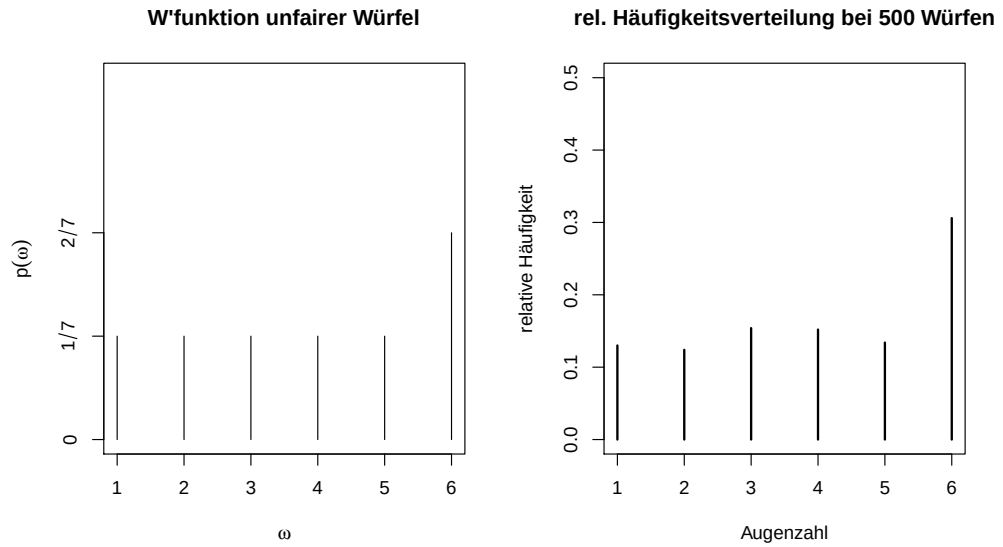


Abbildung 16: Wahrscheinlichkeitsfunktion des unfairen Würfelwurfs und rel. Häufigkeitsverteilung bei 500 simulierten Würfen

5.3. Kombinatorik

Nun beschäftigen wir uns mit der Berechnung von Wahrscheinlichkeiten in Laplace-Modellen,

$$P(A) \;=\; \frac{\#A}{\#\Omega};$$

d.h. wir lernen Verfahren kennen, um die Anzahl der Elemente in „typischen“ Mengen zu bestimmen. Wir behandeln die wichtigsten Abzählformen anhand des Urnenmodells, anschließend noch anhand des Teilchen-Fächer-Modells.

Kombinatorische Abzählformeln (5.13)

Ziehe k Kugeln aus einer Urne mit insgesamt n nummerierten Kugeln. Wie viele verschiedene Ergebnisse (=Kombinationen gezogener Kugeln) sind möglich?

Die Antwort hängt davon ab, ob

- mit oder ohne Zurücklegen gezogen wird? Legen wir eine Kugel zurück, bevor wir die nächste ziehen?
- wird die Reihenfolge, in der die Kugel gezogen werden, berücksichtigt?

I) **mit Zurücklegen, mit Reihenfolge**
Menge der möglichen Ergebnisse wird beschrieben durch

$$\Omega_I \;=\; \Big\{(\omega_1,\omega_2,\ldots,\omega_k)\,:\,1\le\omega_i\le n\Big\},$$

hierbei beschreibt ω_i die Nummer der i-ten gezogenen Kugel. Es gilt

$$\#\Omega_I \;=\; \underbrace{n\cdot n\cdots n}_{k\text{-mal}} \;=\; n^k$$

Beispiel: 4-maliger Würfelwurf:

$$\Omega \;=\; \Big\{(\omega_1,\omega_2,\omega_3,\omega_4)\,:\,1\le\omega_i\le 6\Big\};\qquad \#\Omega \;=\; 6^4 \;=\; 1296$$

$$P(\text{„4 mal die 6“}) \;=\; \frac{\#\{6,6,6,6\}}{\#} \;=\; \frac{1}{1296} \;\approx\; 0.00077$$

Simulation in R: Mit dem Befehl

```
sample(1:n, size=k, replace=TRUE)
```

II) **ohne Zurücklegen, mit Reihenfolge**

$$\Omega_{II} \;=\; \Big\{(\omega_1,\omega_2,\ldots,\omega_k)\,:\,1\le\omega_i\le n,\ \omega_i\ne\omega_j \text{ für } i\ne j\Big\}$$

Es gilt

$$\#\Omega_{II} \;=\; \underbrace{n\cdot(n-1)\cdots(n-k+1)}_{k\text{ Faktoren}} \;=:\; (n)_k \;=\; \frac{n!}{(n-k)!}$$

Beispiel: Turnierpaarungen werden ausgelost, wieviele Möglichkeiten gibt es, 16 Mannschaften auf 8 Spiele zu verteilen (wenn zwischen Heim- und Auswärtsrecht unterschieden wird, es also 16 unterschiedliche Startplätze gibt?)

$$(16)_{16} \;=\; 16! \approx 2.1\cdot 10^{13}$$

Simulation in R: Mit dem Befehl

```
sample(1:n, size=k, replace=FALSE)
```

Fakultäten können in R mit dem Aufruf `factorial` berechnet werden , hier also: `factorial(16)`

III) **ohne Zurücklegen, ohne Reihenfolge**

$$\Omega_{III} \;=\; \Big\{A \subset \{1,\ldots,n\} \,:\, \#A = k\Big\}.$$

Es gilt

$$\#\Omega_{III} \;=\; \frac{n!}{k!(n-k)!} \;=:\; \binom{n}{k}$$

$\binom{n}{k}$ heißt *Binomialkoeffizient* und gibt Anzahl der Möglichkeiten an, k Objekte aus einer Menge von n Objekten auszuwählen.

Beispiel: Lotto „6 aus 49“:

$$\Omega \;=\; \Big\{A \subset \{1,\ldots,49\} \,:\, \#A = 6\Big\}$$

$$\#\Omega \;=\; \frac{49\cdot 48\cdot 47\cdot 46\cdot 45\cdot 44}{6\cdot 5\cdot 4\cdot 3\cdot 2\cdot 1} \;=\; 13\,983\,816$$

$\binom{n}{k}$ lässt sich in R mit dem Befehl `choose(n,k)` berechnen. Für kleine Werte von n und k ($n \leq 10$) generiert der R-Befehl `combn(1:n,k)` alle möglichen Auswahlen von k Zahlen aus den Zahlen $1,\ldots,n$.

`choose`
`combn`

Teilchen-Fächer-Modelle (5.14)

Für Anwendungen ist es manchmal hilfreich, eine zweite Vorstellung der obigen Modelle zu haben: **Es werden k Teilchen auf n Fächer verteilt**. Dabei wird unterschieden, ob

- Mehrfachbelegungen erlaubt sind, oder nicht?
- die Teilchen unterscheidbar sind, oder nicht?

Die obigen Abzählformeln beschreiben dann folgende Situationen:

I) **Verteilen von k unterscheidbaren Teilchen auf n Fächer, Mehrfachbelegungen erlaubt.**
Ω_I wie oben, ω_i gibt dann die Nummer des Faches an, in welches das i-te Teilchen gelegt wird.

II) **Verteilen von k unterscheidbaren Teilchen auf n Fächer, Mehrfachbelegungen nicht erlaubt.**
Ω_{II} wie oben, ω_i gibt dann die Nummer des Faches an, in welches das i-te Teilchen gelegt wird.

III) **Verteilen von k nicht unterscheidbaren Teilchen auf n Fächer, Mehrfachbelegungen nicht erlaubt.**
Ω_{III} wie oben, die Teilmenge A enthält die Nummern der Fächern, in welche Teilchen gelegt werden.

Beispiel 5.15. Wie viele Möglichkeiten gibt es, 10 Personen auf 12 Stühle zu verteilen?

Wir fassen die Personen als Teilchen auf, die Stühle als Fächer. Es sind also $k = 10$ unterscheidbare Teilchen auf $n = 12$ Fächer zu verteilen, Mehrfachbelegungen nicht erlaubt. Das ist Situation II, die Anzahl der Möglichkeiten ist gegeben durch

$$(12)_{10} \;=\; 239\,500\,800$$

Es können auch Abzählformeln kombiniert werden:

Beispiel 5.16. Wahrscheinlichkeit für genau 2-mal die „6“ beim 10-maligen Würfeln?

$$P\Big(\text{„genau zweimal die Sechs“}\Big) \;=\; \frac{\binom{10}{2}\cdot 5^8}{6^{10}}$$

Hierbei ist 6^{10} die Größe des Ergebnisraums (Situation I), $\binom{10}{2}$ sind die Möglichkeiten, *wann* die Sechsen gewürfelt werden, 5^8 ist die Anzahl der möglichen Augenzahlkombinationen der verbleibenden 8 Würfe, in denen keine 6 fallen darf.

Beispiel 5.17 (Hypergeometrische Verteilung)**.** Gegeben ist eine Urne mit N Kugeln, davon R Rote und $N-R$ weiße Kugeln. Dies können wir so modellieren, dass die Kugeln mit den Nummern $1,\dots,R$ Rot sind, die verbleibenden Kugeln weiß.

Wir ziehen n Kugeln ohne Zurücklegen. Wie hoch ist die Wahrscheinlichkeit, dass genau r rote Kugeln gezogen werden?

Der zugrundeliegende Ergebnisraum ist (Situation III)

$$\Omega \;=\; \Big\{A\subset\{1,\dots,N\} \,:\, \#A = n\Big\}, \qquad \#\Omega \;=\; \binom{N}{n}.$$

Wir interessieren uns für das Ereignis

$$E_r \;:=\; \Big\{A\subset\{1,\dots,N\} \,:\, \#A\cap\{1,\dots,R\} = r\Big\}$$

$$\#E_r \;=\; \binom{R}{r}\cdot\binom{N-R}{n-r}$$

Hierbei ist $\binom{R}{r}$ die Anzahl möglicher Auswahlen von r roten Kugeln aus insgesamt R roten Kugeln; $\binom{N-R}{n-r}$ die Anzahl möglicher Auswahlen von $n-r$ weißen Kugeln aus insgesamt $N-R$ weißen Kugeln. Also

$$P(E_r) \;=\; \frac{\binom{R}{r}\cdot\binom{N-R}{n-r}}{\binom{N}{n}}$$

5.4. Kurz-Befehlsreferenz

`choose(n,k)`	berechnet $\binom{n}{k}$
`factorial(n)`	berechnet $n!$
`sample`	simuliert Ziehung mit / ohne Zurücklegen (mit Beachtung der Reihenfolge)
`combn(1:n,k)`	generiert alle möglichen Auswahlen von k Zahlen aus der Menge $\{1,\ldots,n\}$.

6. Bedingte Wahrscheinlichkeiten und stochastische Unabhängigkeit

Beispiel 6.1. Einfacher Würfelwurf:
$\Omega = \{1,\cdots,6\}$
$A = \{2\}$ = „Augenzahl 2“

Wahrscheinlichkeit von $A = P(A) = \frac{\#A}{\#\Omega} = \frac{1}{6}$

$B = \{2,4,6\}$ = „gerade Augenzahl“

Falls bekannt, dass B eingetreten, dann:

Wahrscheinlichkeit von $A = \frac{\#A}{\#B} = \frac{1}{3}$

Fazit: Vorabinformationen beeinflussen Einschätzung von Wahrscheinlichkeiten.

Definition 6.2. Seien A, B Ereignisse auf einem W'raum mit $P(B) > 0$. Dann heißt

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

die bedingte Wahrscheinlichkeit von A gegeben B.

Beispiel 6.3 ((Fortsetzung von Beispiel 6.1)). Berechnung bedingte Wahrscheinlichkeit beim Würfelwurf

$$P(A|B) \overset{A\subseteq B}{=} \frac{P(A)}{P(B)} = \frac{1/6}{3/6} = \frac{1}{3}$$

Satz 6.4 (Satz von der totalen Wahrscheinlichkeit). *Sei $B_1, ..., B_n$ Zerlegung von Ω, d.h. $B_1, ..., B_n$ paarweise disjunkt und $\bigcup_{i=1}^{n} B_i = \Omega$. Weiterhin sei $P(Bi) > 0 \; \forall_i$. Dann gilt für beliebiges Ereignis A:*

$$P(A) = \Sigma_{i=1}^{n} P(A|B_i) \, P(B_i)$$

Bemerkung: Satz 6.4 für $n = 2$ leicht einzusehen:

$$\begin{aligned} P(A) &= P((A \cap B) \cup (A \cap B^c)) \\ &= P(A \cap B) + P(A \cap B^c) \\ \text{(Def. 6.2)} \; &= P(A|B)P(B) + P(A|B^c)P(B^c) \end{aligned}$$

Beispiel 6.5. Von einer Form der Farbblindheit (anomale Trichomasie) sind betroffen:

6.3 Prozent der männlichen Bevölkerung

0.37 Prozent der weiblichen Bevölkerung
in der Altersklasse $\geq$ 65 Jahren.

Geschlechterverhältnis in dieser Altersklasse
0.67 : 1 (Männer : Frauen)

Gesucht: Wahrscheinlichkeit, dass eine zufällig ausgewählte Person in dieser Altersklasse farbenblind ist.

Schritt 1: Betrachte geeignete Ereignisse

F = „zufällig aus Altersgruppe ausgewählte Person ist farbenblind.“

M = „———————— “„———————— männlich “

W = „———————— “„———————— weiblich “

Schritt 2: Interpretiere gegebene Prozentsätze / Anteile als Wahrscheinlichkeiten

$$P(M) = \frac{0.67}{0.67 + 1} = 40.12\%$$

$$P(W) = \frac{1}{0.67+1} = 59.88\%$$
$$P(F|M) = 6.3\%$$
$$P(F|W) = 0.37\%$$

Schritt 3: Wende Satz von der totalen Wahrscheinlichkeit an.

$$\begin{aligned} P(F) &= P(F|M)P(M) + P(F|W) \cdot P(W) \\ &= 0.4012 \cdot 0.63 + 0.5988 \cdot 0.0037 \\ &= 2.75\% \end{aligned}$$

Satz 6.6 (Satz von Bayes)**.** *Sei* $B_1, \cdots, B_n$ *Zerlegung von* Ω *und gelte*

$$P(B_i) > 0 \; \forall_i, \; P(A) > 0.$$

Dann

$$\begin{aligned} P(B_i|A) &= \frac{P(A|B_i) \cdot P(B_i)}{P(A)} \\ &= \frac{P(A|B_i) \cdot P(B_i)}{\sum_{j=1}^{n} P(A|B_j)P(B_j)} \end{aligned}$$

Bemerkung: Die Aussage von Satz 6.6 sieht man so:

$$\begin{aligned} P(B_i|A) &\overset{\text{Def.}}{=} \frac{P(A \cap B_i)}{P(A)} \overset{\text{Def.}}{=} \frac{P(A|B_i)}{P(A)} \\ &\overset{\text{Satz 6.4}}{=} \frac{P(A|B_i)}{\sum_{j=1}^{n} P(A|B_j)P(B_j)} \end{aligned}$$

Beispiel 6.7 (Fortsetzung von Beispiel 6.5)**.** Gesucht: Wahrscheinlichkeit, dass eine zufällig ausgewählte Person aus der Altersklasse ≥ 65 Jahre eine Frau ist.

Schritt 1: Gesucht ist $P(W|F)$

Schritt 2: Wende Satz von Bayes an

$$P(W|F) = \frac{P(F|W) \cdot P(W)}{P(F)}$$

$$= \frac{0.0037 \cdot 0.5988}{0.0275}.$$

Nun: Unabhängigkeit

Intuitiv:

A, B unabhängig, wenn Eintreten von B die Wahrscheinlichkeit von A nicht beeinflusst.

Formal:
A, B unabhängig, wenn $P(A|B) = P(A)$,
d.h.

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

Definition 6.8 (Stochastische Unabhängigkeit für Ereignisse A, B)**.** Sei (Ω, A, P) W'raum, A, B Ereignisse. A, B heißen *(stochastisch) unabhängig*, wenn gilt:

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

Beispiel 6.9. Betrachte Urne mit 2 roten und 3 schwarzen Kugeln, ziehe 2 Kugeln mit Zurücklegen.

$A =$ „1. Kugel rot“, $B =$ „2. Kugel schwarz “

Modell: P Laplace-Verteilung auf

$$\Omega = \{(\omega_1, \omega_2) : \omega_i \, \epsilon \, \{1, \cdots, 6\}\} = \{1, \cdots, 6\}^2$$

$$P(A) = \frac{\#A}{\#\Omega} = \frac{2 \cdot 5}{5 \cdot 5} = \frac{2}{5}$$

$$P(B) = \frac{\#B}{\#\Omega} = \frac{5 \cdot 3}{5 \cdot 5} = \frac{3}{5}$$

$$P(A \cap B) = \frac{\#(A \cap B)}{\#\Omega} = \frac{2 \cdot 3}{5 \cdot 5} = \frac{6}{25}$$

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

$$\Rightarrow A, B \text{ sind unabhängig}$$

Vorsicht: **Keine Unabhängigkeit** bei Ziehen **ohne Zurücklegen**!

Bemerkung:

A, B *u.a.* $\Rightarrow$

$$A, B^c \text{ u.a.},$$
$$A^c, B \text{ u.a.},$$
$$A^c, B^c \text{ u.a.}$$

Definition 6.10 (Stochastische Unabhängigkeit für drei oder mehr Ereignisse, paarweise Unabhängigkeit)**.** Es sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, $A_1, \ldots, A_n \subset \Omega$ Ereignisse. Dann heißen

(i) $A_1, \ldots, A_n$ paarweise (stochastisch) unabhängig genau dann, wenn

$$P(A_j \cap A_k) = P(A_j) \cdot P(A_k) \qquad \forall j \neq k$$

(ii) $A_1, \ldots, A_n$ (gemeinsam stochastisch) unabhängig genau dann, wenn

$$P(A_{j_1} \cap A_{j_2} \cap \cdots \cap A_{j_m}) = P(A_{j_1}) \cdot P(A_{j_2}) \cdots P(A_{j_m})$$

für alle $2 \leq m \leq n$ und jede Auswahl $\{j_1, \ldots, j_m\} \subset \{1, \ldots, n\}$

Beispiel 6.11. Wie viele Identitäten muss man „pro m“ überprüfen? $\binom{n}{m}$! Und konkret für drei Ereignisse A, B, C muss folgendes geprüft werden:

$$\begin{aligned} P(A \cap B) &= P(A) \cdot P(B) \\ P(B \cap C) &= P(B) \cdot P(C) \\ P(C \cap A) &= P(C) \cdot P(A) \\ P(A \cap B \cap C) &= P(A) \cap P(B) \cap P(C) \end{aligned}$$

Bei paarweiser stochastischer Unabhängigkeit würden nur die ersten drei Gleichungen gelten.

Bemerkung 6.12. Es gilt: gemeinsam unabhängig $\Rightarrow$ paarweise unabhängig.
Es gilt **nicht**: paarweise unabhängig $\Rightarrow$ gemeinsam unabhängig.

Betrachte z.B. beim zweifachen Würfelwurf die Ereignisse A: 1. Augenzahl gerade, B: 2. Augenzahl gerade, C: Augensumme gerade. Diese Ereignisse sind paarweise unabhängig (nachrechnen!); aber

$$P(A \cap B \cap C) = P(A \cap B) = \frac{1}{4} \neq \frac{1}{8} = P(A) \cdot P(B) \cdot P(C)$$

Beispiel 6.13. Verfahren zur sterilen Abfüllung von Flaschen. Die langfriste Erfahrung besagt, dass ein Anteil von 0.1% der Flaschen Ausschuss ist, d.h. unsteril. Wir nehmen an, dass Verunreinigungen unabhängig voneinander auftreten. Untersuche Stichprobe von N Flaschen.

Gesucht: Wahrscheinlichkeit, dass alle Flaschen in der Stichprobe steril sind.

Betrachte geeignete Ereignisse

E_1: „erste Flasche der Stichprobe ist steril “

E_2: „zweite Flasche der Stichprobe ist steril “

$\vdots$

E_N: „N-te Flasche der Stichprobe ist steril “

Nach Annahme sind $E_1, \ldots, E_n$ unabhängig; außerdem

$$P(E_1) = P(E_2) = \cdots = P(E_N) = 1 - 0.001 = 0.999$$

Folglich

$$\begin{aligned} & P(\text{„alle Flaschen in der Stichprobe sind steril “}) \\ &= P(E_1 \cap E_2 \cap \cdots \cap E_N) \\ &\overset{(*)}{=} P(E_1) \cdot P(E_2) \cdots P(E_N) \\ &= (0.999)^N \end{aligned}$$

Bei $(*)$ haben wir die Unabhängigkeit benutzt.

Definition 6.14. Ein Zufallsexperiment mit nur zwei möglichen Ergebnissen (Erfolg, Misserfolg) heißt *Bernoulli-Experiment*. Sei p die Wahrscheinlichkeit für einen Erfolg, dann ist $1 - p$ die Wahrscheinlichkeit für einen Misserfolg.

Sehr viele Fragestellungen lassen sich auf ein Bernoulli-Experiment reduzieren: Fällt eine „Sechs“beim Würfelwurf? Bleibt es heute trocken? Gewinne ich im Lotto?

Definition 6.15. Die n-fache unabhängige Durchführung eines Bernoulli-Experimentes (mit Erfolgswahrscheinlichkeit p) heißt *Bernoulli-Kette* der Länge n (mit Erfolgswahrscheinlichkeit p). Der Ergebnisraum ist

$$\Omega \;=\; \{(\omega_1, \ldots, \omega_n) \,:\, \omega_i \in \{0, 1\},\; 1 \leq i \leq n\},$$

$\omega_i = 1$ entspricht hierbei einem Erfolg im i-ten Experiment.

Aufgrund der Unabhängigkeit ist die Wahrscheinlichkeitsfunktion gegeben durch

$$p\big((\omega_1,\ldots,\omega_n)\big) \;=\; p^k(1-p)^{n-k}; \qquad k=\sum_{i=1}^{n}\omega_i,$$

d.h., k gibt hier die Anzahl der Erfolge im Ergebnis $(\omega_1,\ldots,\omega_n)$ an.

Beispiel 6.16. Gesucht: Wahrscheinlichkeit, dass in einer Bernoulli-Kette der Länge n genau k Erfolge auftreten (egal wann)?

$$A_k = \text{„genau } k \text{ Erfolge“} = \big\{(\omega_1,\ldots,\omega_n)\in\Omega \,:\, \sum_{i=1}^{n}\omega_i = k\big\}$$

Jedes Ergebnis in A_k hat Wahrscheinlichkeit $p^k(1-p)^{n-k}$ und es gibt $\binom{n}{k}$ Ergebnisse in A_k - dies entspricht dem Verteilen von k nicht unterscheidbaren Teilchen (Erfolge) auf n Fächer (Durchführungen des Experimentes), ohne Mehrfachbelegung. Gemäß Def. 5.8 ist dann

$$P(\text{„genau } k \text{ Erfolge“}) \;=\; P(A_k) \;=\; \sum_{\omega\in A_k} p(\omega) \;=\; \binom{n}{k}p^k(1-p)^{n-k}.$$

Definition 6.17. Die Wahrscheinlichkeitsverteilung P auf $\Omega=\{0,\ldots,n\}$ mit Wahrscheinlichkeitsfunktion

$$p(k) \;=\; \binom{n}{k}p^k(1-p)^{n-k}$$

heißt *Binomialverteilung* mit Parametern n und p, kurz: $B(n,p)$-Verteilung.

7. Zufallsvariablen und ihre Kenngrößen

Bei der Einführung der Binomialverteilung haben wir jedem Ergebnis $(\omega_1,\ldots,\omega_n)$ eine Zahl zugeordnet, nämlich die Anzahl der Erfolge in diesem Ergebnis. Führen wir eine Abbildung

$$X\big((\omega_1,\ldots,\omega_n)\big) := \sum_{i=1}^{n}\omega_i$$

ein, so tritt das Ereignis A_k genau dann ein, wenn die *Zufallsvariable* X den Wert k annimmt; genauer:

$$A_k \;=\; \{\omega\in\Omega \,:\, X(\omega)=k\} \;=:\; \{X=k\}$$

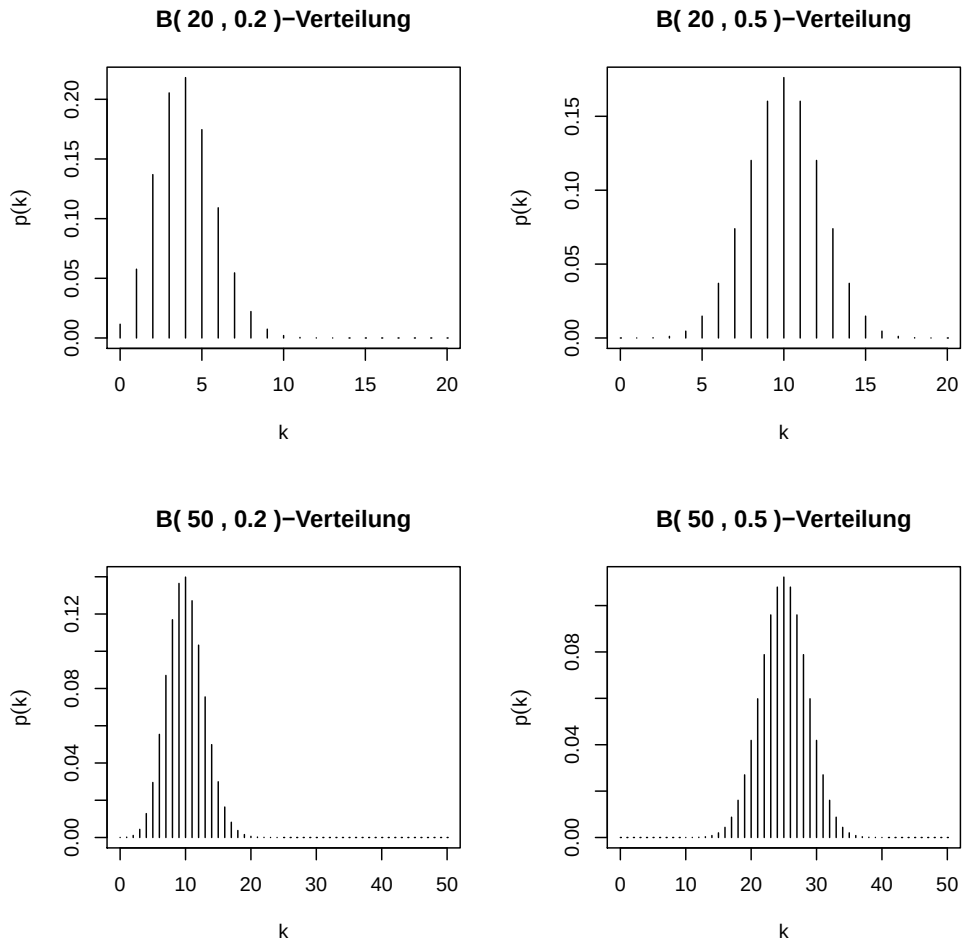


Abbildung 17: Darstellung der Wahrscheinlichkeitsfunktionen verschiedener Binomialverteilungen

Definition 7.1. Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum. Eine *Zufallsvariable* ist eine Abbildung

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

Notwendige technische Eigenschaft: Für alle $a \in \mathbb{R}$ ist

$$\{X \leq a\} = \{\omega \in \Omega : X(\omega) \leq a\} \in \mathcal{A}. \tag{1}$$

Interpretation: Eine Zufallsvariable ist eine Vorschrift, die jedem Ergebnis eines Zufallsexperimentes eine reelle Zahl zuordnet.

Beispiel 7.2 (Zweifacher Würfelwurf)**.**

$$\Omega = \{(\omega_1, \omega_2) : \omega_1, \omega_2 \in \{1, \ldots, 6\}\}$$

$$P = \text{Laplaceverteilung auf } \Omega, \quad \#\Omega = 36, \qquad p(\omega) = \frac{1}{36} \text{ für alle } \omega \in \Omega$$

Definiere Zufallsvariable $Y : \Omega \to \mathbb{R}$ durch

$$Y(\omega_1, \omega_2) := \omega_1 + \omega_2,$$

dann gibt Y die Summe der Augenzahlen an.

Gesucht: W'keit, dass die Augenzahl kleiner oder gleich 3 ist?

D.h., wir wollen $P(Y \leq 3)$ bestimmen.

$$\{\omega \in \Omega : Y(\omega) \leq 3\} = \big\{(1,1), (1,2), (2,1)\big\},$$

also

$$P(Y \leq 3) = P\big(\{\omega \in \Omega : Y(\omega) \leq 3\}\big) = 3 \cdot \frac{1}{36} = \frac{1}{12}.$$

Bemerkung 7.3. Bedingungen wie $Y = a$, $Y \leq b$, $Y > c$, $Y \in A$ etc. definieren stets Ereignisse! Schließen sich verschiedene Bedingungen gegenseitig aus, so sind die dadurch beschriebenen Ereignisse stets disjunkt. Beispiel:

$$\{Y = k\} \cap \{Y = j\} = \emptyset \qquad \text{für alle } k \neq j;$$

folglich $P(Y \in \{k, j\}) = P(Y = k) + P(Y = j)$

Definition 7.4. Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, $X : \Omega \to \mathbb{R}$ eine Zufallsvariable. Die Abbildung

$$A \mapsto P(X \in A) := P\big(\{\omega \in \Omega : X(\omega) \in A\}\big),$$

wobei $A \subset \mathbb{R}$ ein Intervall ist, heißt die *(Wahrscheinlichkeits-)Verteilung von X*.
Notation: $P_X(A) = P(X \in A)$.

Bemerkung 7.5. P_X ist wieder ein Wahrscheinlichkeitsmaß (bzw. lässt sich zu einem Wahrscheinlichkeitsmaß fortsetzen). Insbesondere gelten die allgemeinen Rechenregeln und Eigenschaften (siehe 5.5, 5.6).

Schreibweisen für Zufallsvariablen

A	Altern. Schreibw. für $X \in A$
B^c	$\{X \notin B\}$
$\{x\}$, $\{x\}^c$	$\{X = x\}$, $\{X \neq x\}$
$[a, b]$, (a, b)	$\{a \leq X \leq b\}$, $\{a < X < b\}$
$(a, b]$, $[a, b)$	$\{a < X \leq b\}$, $\{a \leq X < b\}$
$(-\infty, x]$, $(-\infty, x)$	$\{X \leq x\}$, $\{X < x\}$
$[x, \infty)$, (x, ∞)	$\{X \geq x\}$, $\{X > x\}$

Bei Wahrscheinlichkeiten verzichten wir auf die Verwendung von Mengenklammern, schreiben also

$$P(X \in A), \quad P(X = x), \quad \text{etc.}$$

für

$$P(\{X \in A\}), \quad P(\{X = x\}), \quad \text{etc..}$$

Definition 7.6. Zufallsvariablen $X_1, \ldots, X_n$ heißen (gemeinsam) stochastisch unabhängig genau dann, wenn für **jede** Auswahl von Intervallen $I_1, \ldots, I_n \subset \mathbb{R}$ gilt: Die Ereignisse $\{X_1 \in I_1\}, \ldots, \{X_n \in I_n\}$ sind stochastisch unabhängig (vgl. Def. 6.10).

Beispiel 7.7. Gegeben eine Bernoulli-Kette der Länge n (vgl. Def. 6.15), definiere Zufallsvariablen

$$X_i : \Omega \to \mathbb{R}, \quad (\omega_1, \ldots, \omega_n) \mapsto \omega_i.$$

Dann nimmt die Zufallsvariable X_i den Wert 1 an genau dann, wenn ein Erfolg im i-ten Experiment auftritt. Die Zufallsvariablen $X_1, \ldots, X_n$ sind stochastisch unabhängig.

Bemerkung 7.8. Sie werden in der Praxis nie stochastische Unabhängigkeit von Zufallsvariablen anhand dieser Definition nachprüfen, vielmehr werden Sie in der stochastischen Modellierung eines realen Experimentes sehr häufig die *Annahme* treffen, dass auftretende Zufallsvariablen stochastisch unabhängig sind.

7.1. Diskrete Zufallsvariablen

Definition 7.9. Nimmt eine Zufallsvariable X höchstens abzählbar viele Werte $\{a_1, a_2, \ldots\}$ an, so heißt X *diskrete Zufallsvariable.* In diesem Fall heißt

$$p_X(a_i) \ := \ P(X = a_i)$$

die *Wahrscheinlichkeitsfunktion von X*.

Bemerkung 7.10. Sei X eine diskrete Zufallsvariable (mit Wertebereich $\{a_1, a_2, \dots\}$).

(i) Für jedes Intervall $A \subset \mathbb{R}$ gilt dann

$$P(X \in A) = \sum_{a_i \in A} p_X(a_i).$$

(ii) Die Wahrscheinlichkeitsfunktion p_X von X entspricht der Wahrscheinlichkeitsfunktion des W'Maßes P_X, vgl. 5.8.

Beispiel 7.11. (a) Zweifacher Würfelwurf.

$$\Omega = \big\{(\omega_1, \omega_2) : \omega_1, \omega_2 \in \{1, \dots, 6\}\big\}, \qquad P = \text{Laplaceverteilung auf } \Omega, \qquad \#\Omega = 36$$

Wir interessieren uns für die Verteilung der Zufallsvariablen X: Summe der Augenzahlen. Da ein Laplace-Experiment vorliegt, gilt

$$p_X(i) = P(X = i) = \frac{\#\{\omega : X(\omega) = i\}}{\#\Omega}, \qquad i \in \{2, 3, \dots, 12\}.$$

Bestimme also für $i = 2, 3, \dots, 12$ die Mächtigkeit des Ereignisses $\{X = i\}$:

$$\begin{aligned}
\{X = 2\} &= \big\{(1,1)\big\} \Rightarrow \#\{X = 2\} = 1 \Rightarrow p_X(2) = \frac{1}{36}. \\
\{X = 3\} &= \big\{(1,2),(2,1)\big\} \Rightarrow \#\{X = 3\} = 2 \Rightarrow p_X(3) = \frac{2}{36} = \frac{1}{18}. \\
\{X = 4\} &= \big\{(1,3),(2,2),(3,1)\big\} \Rightarrow \#\{X = 4\} = 3 \Rightarrow p_X(4) = \frac{3}{36} = \frac{1}{12}. \\
\{X = 5\} &= \big\{(1,4),(2,3),(3,2),(4,1)\big\} \Rightarrow p_X(5) = \frac{4}{36} = \frac{1}{9}. \\
\{X = 6\} &= \big\{(1,5),(2,4),(3,3),(4,2),(5,1)\big\} \Rightarrow p_X(6) = \frac{5}{36}. \\
\{X = 7\} &= \big\{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\big\} \Rightarrow p_X(7) = \frac{6}{36} = \frac{1}{6}.
\end{aligned}$$

Analog erhält man

$$p_X(8) = \frac{5}{36},\ p_X(9) = \frac{1}{9},\ p_X(10) = \frac{1}{12},\ p_X(11) = \frac{1}{18},\ p_X(12) = \frac{1}{36}.$$

(b) Warten auf die Sechs: Werfe einen Würfel so lange, bis zum ersten Mal die Sechs erscheint. Einzelne Würfe werden unabhängig voneinander ausgeführt. Wir interessieren uns für die Verteilung der Zufallsvariable X: Anzahl der vergeblichen Versuche vor der ersten 6.

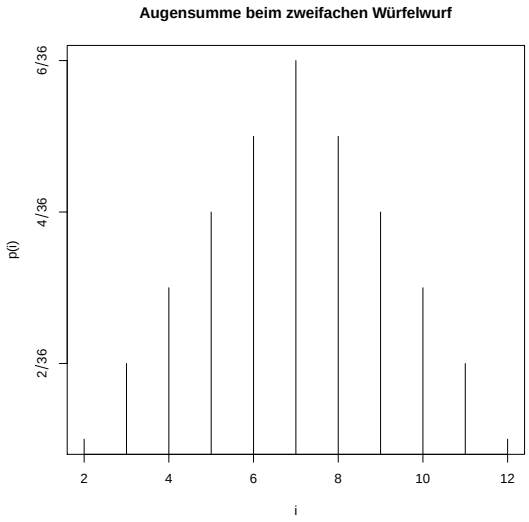


Abbildung 18: Darstellung der Wahrscheinlichkeitsfunktionen der Augensumme beim zweifachen Würfelwurf

$X = k$ gilt gdw. in den ersten k Würfen keine 6, im k-ten Wurf eine 6 fällt

Da die einzelnen Würfe unabhängig voneinander sind, gilt:

$$p_X(k) = P(X = k) \;=\; \left(\frac{5}{6}\right)^k \cdot \frac{1}{6}, \qquad k \in \{0, 1, 2, \dots\}$$

(c) Es bezeichne X: Anzahl Erfolge in einer Bernoulli-Kette der Länge n mit Erfolgswahrscheinlichkeit p. Nach Beispiel 6.16 gilt dann

$$p_X(k) = P(X = k) \;=\; \binom{n}{k} p^k (1-p)^{n-k} \qquad k \in \{0, \dots, n\}.$$

Gesucht: Wahrscheinlichkeit, dass *mindestens* ein Erfolg auftritt? D.h., gesucht ist $P(X \geq 1)$.

$$P(X \geq 1) = P\big(\{X = 0\}^c\big) = 1 - P(X = 0) = 1 - \binom{n}{0} p^0 (1-p)^n = 1 - (1-p)^n.$$

Wichtige diskrete Verteilungen (7.12)

Wir listen die wichtigsten diskreten Verteilungen auf, und ihre Implementation in R. Dabei gibt es ein generelles Schema: `d"Verteilungsname"` liefert die Wahrscheinlichkeitsfunktion (z.B. `dbinom`), `r"Verteilungsname"` erzeugt entsprechend verteilte Zufallsvariablen (z.B. `rbinom`), `p"Verteilungsname"` liefert die *Verteilungsfunktion*, d.h. $P(X \leq t)$, wobei t als Argument übergeben wird (z.B. `pbinom`).

1. **Gleichverteilung (Laplaceverteilung) auf** $\{1, \ldots, N\}$

$$p(k) = \frac{1}{N}, \qquad k \in \{1, 2, \ldots, N\}.$$

Kurzbezeichnung: Laplace($\{1, \ldots, N\}$)

Jede Zahl der Menge $\{1, \ldots, N\}$ tritt mit gleicher Wahrscheinlichkeit auf, wie etwa beim Münz- ($N = 2$) oder Würfelwurf ($N = 6$). Vgl. Def. 5.9.

R-Befehle: Simulation von mit Hilfe des Befehls `sample(1:N,size=100,replace=TRUE)`. Dies liefert 100 *Realisierungen* einer gleichverteilten Zufallsvariable auf $\{1, \ldots, N\}$. Die gewünschte Anzahl wird mit dem Argument `size` übergeben. `sample`

2. **Binomialverteilung mit Parametern** $n \in \mathbb{N}$, $p \in (0, 1)$

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k \in \{0, 1, \ldots, n\}.$$

Kurzbezeichnung: $B(n, p)$

Verteilung der Anzahl der Erfolge in einer Bernoulli-Kette der Länge n mit Erfolgswahrscheinlichkeit p (vgl. Bsp. 6.16). Die Normiertheit der Wahrscheinlichkeitsfunktion folgt aus dem Binomischen Lehrsatz.

R-Befehle: `dbinom(k, size=n, prob=p)` liefert $p(k)$ für gegebene Werte von n und p; `pbinom(k, size=n, prob=p)` liefert die Wahrscheinlichkeit, höchstens k Erfolge zu erzielen; `rbinom(L,size=n, prob=p)` erzeugt L Realisierungen von Binomial(n,p)-verteilten Zufallsvariablen. `dbinom`

3. **Geometrische Verteilung mit Parameter** $p \in (0, 1)$

$$p(k) = (1-p)^k \cdot p, \qquad k \in \{0, 1, 2, \ldots\} = \mathbb{N}_0$$

Kurzbezeichnung: Geom(p)

Verteilung der Anzahl der Fehlversuche vor dem ersten Erfolg bei unabhängigen Bernoulli-Experimenten mit Erfolgswahrscheinlichkeit p (vgl. Beispiel 7.11 (b)). Die Normiertheit der Wahrscheinlichkeitsfunktion folgt aus dem Grenzwert für die geometrische Reihe.

R-Befehle: `dgeom(k,prob=p)` liefert $p(k)$ für gegebenes p; `pgeom(k,prob=p)` lie- `dgeom`

fert die Wahrscheinlichkeit, höchstens k Durchführungen warten zu müssen, bis ein Erfolg auftritt; `rgeom(L,prob=p)` erzeugt L Realisierungen einer geometrisch verteilten Zufallsvariablen.

4. **Hypergeometrische Verteilung, Parameter** $N \in \mathbb{N}$, $R, n \in \{1, \ldots, N\}$

$$p(r) = \frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}}, \qquad r \in \{0, 1, \ldots, n\}$$

Kurzbezeichnung: $H(n, N, R)$

Verteilung der Anzahl roter Kugeln beim Ziehen ohne Zurücklegen von n Kugeln aus einer Urne, gefüllt mit R roten und $N - R$ schwarzen Kugeln (also insgesamt N Kugeln).

R-Befehle: `dhyper(r,m=R,n=N-R,k=n)` liefert $p(r)$ für gegebene Werte R, N und n; `phyper(r,m=R,n=N-R,k=n)` liefert die Wahrscheinlichkeit, höchstens r rote Kugeln zu ziehen; `rhyper(L,m=R,n=N-R,k=n)` erzeugt L Realisierungen einer hypergeometrisch verteilten Zufallsvariablen. `dhyper`

5. **Poisson-Verteilung mit Parameter** $\lambda \in (0, \infty)$

$$p(k) = e^{-\lambda}\frac{\lambda^k}{k!}, \qquad k \in \{0, 1, 2, \ldots\} = \mathbb{N}_0$$

Kurzbezeichnung: $\mathrm{Pois}(\lambda)$

Die Poisson-Verteilung ist die „Verteilung seltener Ereignisse“: Sie findet Verwendung als Annäherung der Binomialverteilung für großes n und kleines p (für $\lambda = n \cdot p$); außerdem beschreibt sie bspw. die Anzahl radioaktiver Zerfälle in einem Zeitintervall bei Zerfallsrate λ.

R-Befehle: `dpois(k,lambda=λ)` liefert $p(k)$ für gegebenes λ; `ppois(k,lambda=λ)` liefert die Wahrscheinlichkeit, höchstens k Ereignisse (Zerfälle) zu sehen; `rpois(k,lambda=λ)` erzeugt L Realisierungen einer Poisson-verteilten Zufallsvariablen. `dpois`

7.2. Kenngrößen für diskrete Verteilungen

Definition 7.13 (Erwartungswert einer diskreten Zufallsvariablen)**.** Sei X eine diskrete Zufallsvariable mit Wertebereich $\{a_1, a_2, \ldots\}$ und Wahrscheinlichkeitsfunktion $p_X(a_i)$. Dann definieren wir den *Erwartungswert* von X als

$$E(X) \;:=\; \sum_i a_i \cdot p_X(a_i).$$

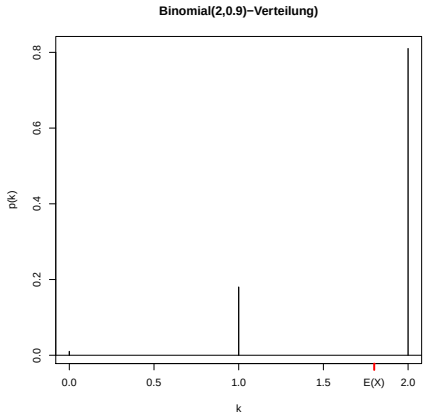


Abbildung 19: Zu Beispiel 7.14: Interpretation von $E(X)$ als Schwerpunkt eines Systems von Massepunkten

Notation $E(X) = EX = \mu_X = \mu$.

Der Erwartungswert ist eine **Kenngröße für die Lage**: wo liegen die Werte der Zufallsvariable im Mittel. Die empirische Entsprechung ist das *Stichprobenmittel*, siehe Def. 3.7.

Beispiel 7.14. Sei X binomialverteilt mit Parametern $n = 2$, $p = 0.9$, d.h.

$$P(X = k) = p_X(k) = \frac{2!}{k!(2-k)!} \cdot (0.9)^k \cdot (0.1)^{2-k}, \qquad k \in \{0, 1, 2\}.$$

Dann gilt:

$$E(X) = 0 \cdot p_X(0) + 1 \cdot p_X(1) + 2 \cdot p(2) = 0 + 1 \cdot 0.18 + 2 \cdot 0.81 = 1.8$$

Beachte: Die Zufallsvariable nimmt den Wert 1.8 überhaupt nicht an; der Erwartungswert ist das gewichtete Mittel der möglichen Werte.

Stellt man sich einen Stab vor, an dem an den Punkten a_i Gewichte der Masse $p(a_i)$ aufgehängt sind, so entspricht der Erwartungswert dieser Verteilung dem physikalischen Schwerpunkt dieses Systems von Massepunkten.

Satz 7.15 (Erwartungswert transformierter diskreter Zufallsvariablen)**.** *Sei X eine diskrete Zufallsvariable wie in Def. 7.13 und $f : \mathbb{R} \to \mathbb{R}$ eine Funktion, so gilt*

$$E\big(f(X)\big) \;=\; \sum_i f(a_i) \cdot p(a_i).$$

Satz 7.16 (Linearität des Erwartungswertes)**.** *Seien X, Y Zufallsvariablen und $a, b \in \mathbb{R}$. Dann gilt:*

$$\begin{aligned} E(X+Y) &= E(X) + E(Y) \\ E(aX+b) &= a \cdot E(X) + b \end{aligned}$$

Analog zur deskriptiven Statistik lernen wir neben dem Erwartungswert nun auch ein Streuungsmaß kennen, die Varianz.

Definition 7.17. Die *Varianz* einer Zufallsvariable ist definiert durch

$$\mathrm{Var}(X) \;=\; E\big((X - EX)^2\big).$$

Die nichtnegative Quadratwurzel der Varianz, $\sqrt{\mathrm{Var}(X)}$ heißt *Standardabweichung* von X.
Notation: $\mathrm{Var}(X) = \mathrm{Var}X = \sigma_X^2 = \sigma^2$.

Die empirische Entsprechung ist die Stichprobenvarianz bzw. die empirische Standardabweichung, siehe Def 3.9.

Beispiel 7.18. Sei X gleichverteilt auf $\{0, 1, \ldots, 4\}$, also $P(X = k) = p_X(k) = \frac{1}{5}$ für $k \in \{0, 1, \ldots, 4\}$. Dann gilt:

$$E(X) = \sum_{k=0}^{4} k \cdot p(k) \;=\; \frac{1}{5}\,(0+1+2+3+4) = \frac{10}{5} = 2.$$

Für die Varianz gilt:

$$\mathrm{Var}(X) = E\big((X - EX)^2\big) \;=\; E\big((X-2)^2\big)$$

Zur weiteren Berechnung wenden wir Satz 7.15 mit der Funktion $f(x) = (x-2)^2$ an, und erhalten

$$\begin{aligned} E\big((X-2)^2\big) &= \sum_{k=0}^{4} (k-2)^2 p(k) \\ &= \frac{1}{5}\Big((0-2)^2 + (1-2)^2 + (2-2)^2 + (3-2)^2 + (4-2)^2\Big) \\ &= \frac{1}{5}\big(4+1+0+1+4\big) \;=\; 2 \end{aligned}$$

Satz 7.19 (Eigenschaften der Varianz)**.** *Sei X eine Zufallsvariable und $a, b \in \mathbb{R}$ Dann gilt:*

$$\begin{aligned} \mathrm{Var}(X) &= E(X^2) - \big(EX\big)^2 \\ \mathrm{Var}(aX+b) &= a^2 \mathrm{Var}(X) \\ E\big((X-a)^2\big) &= \mathrm{Var}(X) + \big(E(X) - a\big)^2 \end{aligned}$$

Sind X, Y ***stochastisch unabhängige*** *Zufallsvariablen, so gilt auch:*

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y). \qquad (*)$$

Achtung: $(*)$ gilt nur unter der Voraussetzung, dass X und Y stochastisch unabhängig sind - im allgemeinen Fall gibt es noch einen Korrekturterm, die sog. Kovarianz.

Erwartungswert und Varianz wichtiger Verteilungen (7.20)

Die Zufallsvariable X habe eine ...

1. Laplaceverteilung auf $\{1, \ldots, N\}$:

$$EX = \frac{N+1}{2}, \qquad \mathrm{Var}(X) = \frac{N^2-1}{12}.$$

2. Binomialverteilung mit $n \in \mathbb{N}$, $p \in (0,1)$:

$$EX = np, \qquad \mathrm{Var}(X) = np(1-p).$$

3. Geometrische Verteilung mit $p \in (0,1)$:

$$EX = \frac{1-p}{p}, \qquad \mathrm{Var}(X) = \frac{1-p}{p^2}$$

4. Hypergeometrische Verteilung mit Parametern N, R, n:

$$EX = n\frac{R}{N}, \qquad \mathrm{Var}(X) = n\frac{R}{N}\Big(1-\frac{R}{N}\Big)\frac{N-n}{N-1}$$

5. Poisson-Verteilung mit $\lambda \in (0, \infty)$:

$$EX = \lambda, \qquad \mathrm{Var}(X) = \lambda.$$

Satz 7.21 (Ungleichung von Tschebyscheff)**.** *Sei X eine Zufallsvariable mit Erwartungswert $\mu = EX$ und Varianz $\sigma^2 = \mathrm{Var}(X)$. Dann gilt für jedes $c > 0$:*

$$P\big(|X-\mu| \geq c\big) \leq \frac{\sigma^2}{c^2}.$$

Beispielsweise gilt für $c = 3\sigma$:

$$P\big(|X-\mu| \geq 3\sigma\big) \leq \frac{\sigma^2}{(3\sigma)^2} = \frac{1}{9}$$

Mit anderen Worten: Mit einer Wahrscheinlichkeit von mindestens 88% nimmt die Zufallsvariable X Werte im Intervall $[\mu - 3\sigma, \mu + 3\sigma]$ an - dies zeigt die Funktion des Erwartungswertes als Lageparameter, sowie der Varianz als Streuungsparameter.

Die Tschebyscheff-Ungleichung ist sehr grob (für $c \leq \sigma$ liefert sie eine triviale obere Schranke), dafür gilt sie für **alle** Zufallsvariablen. In den allermeisten Fällen ist die Wahrscheinlichkeit, Werte aus dem sog. „3σ-Intervall“ zu beobachten, sogar deutlich höher, sie liegt bei über 99%.

Satz 7.22 (Gesetz der großen Zahl)**.** *Seien $X_1, \ldots, X_n$ stochastisch unabhängige, identisch verteilte Zufallsvariablen mit $EX_i = \mu$ und $\mathrm{Var}(X_i) = \sigma^2$. Dann gilt für jede beliebig kleine Schranke $\epsilon > 0$:*

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \epsilon\right) \;\leq\; \frac{\sigma^2}{n\epsilon} \quad \overset{n\to\infty}{\to} \quad 0.$$

Interpretation: Je mehr Realisierungen einer zufälligen Größe vorliegen, desto geringer ist die Wahrscheinlichkeit, dass der Mittelwert der Realisierungen vom Erwartungswert abweicht.

7.3. Stetige Zufallsvariablen

Definition 7.23. Eine Zufallsvariable X heißt *stetig*, wenn es eine integrierbare Funktion $f : \mathbb{R} \to [0, \infty)$ gibt, so dass für alle Intervalle $[a, b] \subset \mathbb{R}$ gilt

$$P(a \leq X \leq b) \;=\; \int_a^b f(x)\, dx.$$

Die Funktion f heißt *Dichtefunktion* oder *Dichte* von X. Notation: $f(x) = f_X(x)$.

Wichtige stetige Verteilungen (7.24)

Wir listen die wichtigsten stetigen Verteilungen auf, und ihre Implementation in R. Wie bei den diskreten Verteilungen gibt es ein generelles Schema: `d"Verteilungsname"` liefert die **Dichte**funktion (z.B. `dnorm`), `r"Verteilungsname"` erzeugt entsprechend verteilte Zufallsvariablen (z.B. `rnorm`), `p"Verteilungsname"` liefert die *Verteilungsfunktion*, d.h. $P(X \leq t)$, wobei t als Argument übergeben wird (z.B. `pnorm`).

1. **Gleichverteilung auf dem Intervall** $[a, b]$

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{für } x \in [a, b] \\ 0 & \text{sonst} \end{cases}$$

   Kurzbezeichnung: $R(a, b)$

R-Befehle: Verteilungsname ist `unif`, Argumente sind `min=a` und `max=b`, z.B. liefert `dunif(x,min=0,max=1)` den Wert von $f(x)$ für die Gleichverteilung auf $[0, 1]$. `dunif`

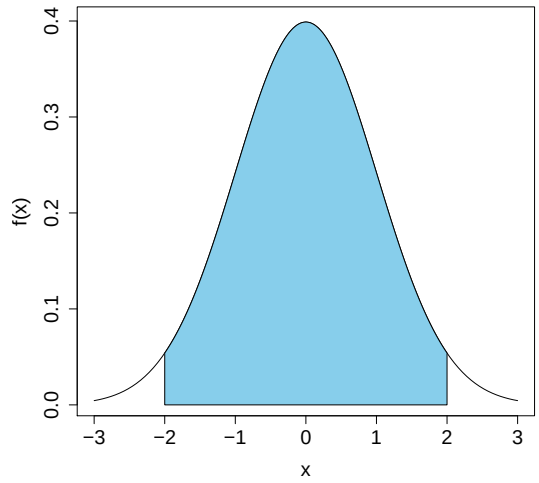


Abbildung 20: Zu Definition 7.23: Die Wahrscheinlichkeit eines Intervalls entspricht der Fläche unter dem Graphen von $f(x)$.

2. **Exponentialverteilung mit Parameter** $\lambda > 0$

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

Kurzbezeichnung: $\text{Exp}(\lambda)$

Anwendung: Lebensdauer, Wartezeit, stetiges Analogon der geometrischen Verteilung.

R-Befehle: Verteilungsname ist `exp`, Argument ist `rate=`λ, z.B. liefert `dexp(x,rate=1)` den Wert von $f(x)$ für die Exponentialverteilung mit Parameter $\lambda = 1$. `dexp`

3. **Normalverteilung mit Parametern** $\mu \in \mathbb{R}$**,** $\sigma^2 > 0$

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

Kurzbezeichnung: $\mathcal{N}(\mu, \sigma^2)$
Spezialfall $\mathcal{N}(0,1)$: *Standardnormalverteilung.*

Die Normalverteilung ist die wichtigste Verteilung überhaupt: Zufallsvariablen sind

normalverteilt, wenn sie eine Überlagerung vieler kleiner unabhängiger Zufallsvariablen sind (Zentraler Grenzwertsatz, s.u.)

R-Befehle: Verteilungsname ist `norm`, Argumente sind `mean=`μ und `sd=`σ, z.B. liefert `dnorm(x,mu=0,sd=1)` den Wert von $f(x)$ für die Standardnormalverteilung.

Darstellung stetiger Verteilungen in R (7.25)

Funktionsgraphen lassen sich in R mit dem Befehl `curve` zeichnen; das Intervall $[a, b]$, über dem die Funktion gezichnet werden soll, wird mit den Argumenten `from=`a und `to=`b übergeben. Dazu kann eine Funktionsvorschrift angegeben werden, z.B. zeichnet `curve`

```
curve(x^2,from=-1,to=1)
```

den Graphen einer Parabel über dem Intervall $[-1, 1]$. Alternativ kann der Name einer in R implementierten Funktion übergeben werden, so lassen sich insbesondere Dichten stetiger Verteilungen zeichnen. Beispielsweise zeichnet

```
curve(dnorm(x,mean=0,sd=1),from=-3,to=3)
```

den Graphen der Dichte der Standardnormalverteilung auf dem Intervall $[-3, 3]$. Hierbei können wie üblich die Achsenbeschriftungen und Überschriften angepasst werden.

Etwas „aufgeräumtere“ Aufrufe bekommt man, indem man zuerst die zu zeichnende Funktion als Funktion definiert, z.B. ist `function`

```
f<-function(x){dnorm(x,mean=0,sd=1)}
curve(f(x),from=-3,to=-3)
```

eine Langversion des vorherigen R-Befehls.

Definition 7.26 (Verteilungsfunktion)**.** Für eine Zufallsvariable X (diskret oder stetig) heißt

$$F_X(t) \;=\; P(X \leq t) \qquad t \in \mathbb{R}.$$

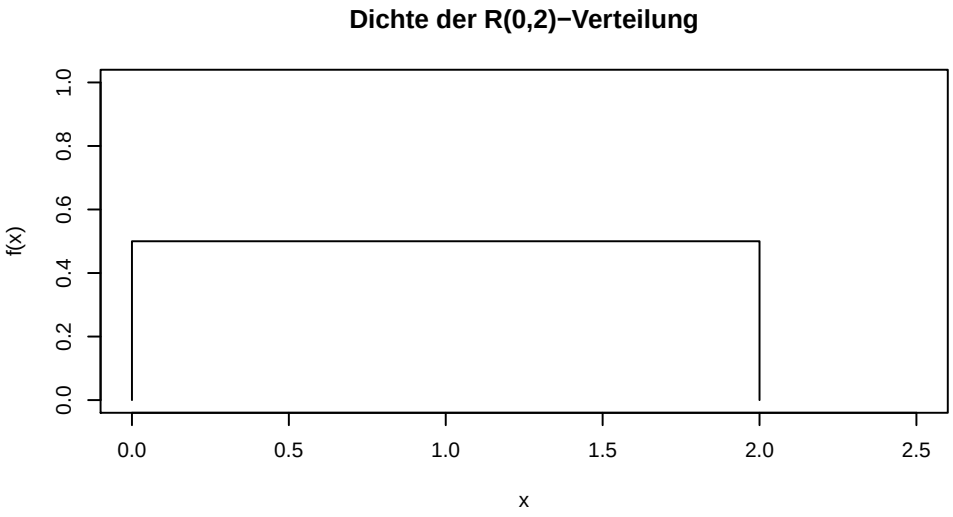
die *Verteilungsfunktion* von X.
Ist X standardnormalverteilt, so verwenden wir auch das Symbol $\Phi(t) = F_X(t)$ für die Verteilungsfunktion der Standardnormalverteilung.

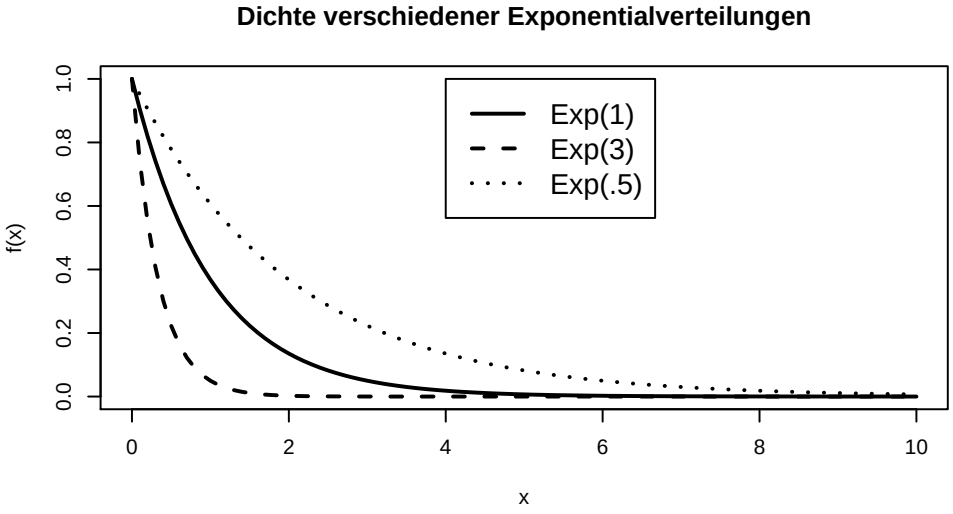
Bemerkung 7.27. Aus der Verteilungsfunktion lassen sich die Wahrscheinlichkeiten beliebiger Intervalle gewinnen:

$$P(X \in (a, b]) = P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a)$$

Ist X eine stetige Zufallsvariable, so gilt

$$F_X(t) = \int_{-\infty}^{t} f_X(y)\, dy.$$





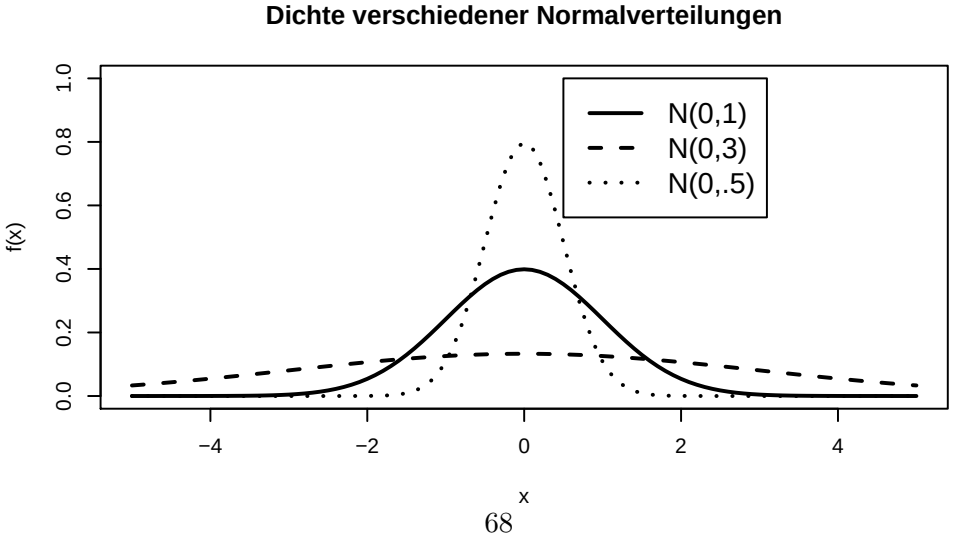


Abbildung 21: Darstellung verschiedener Dichten

In R erhält man die Werte der Verteilungsfunktion durch den Aufruf `p"Verteilungsname"`, z.B. liefert `pnorm(t,mean=0,sd=1)` den Wert $F_X(t)$ für eine standardnormalverteilte Zufallsvariable X.

Beispiel 7.28. (a) Sei X Exp(1)-verteilt, dann

$$\begin{aligned} P(X \leq 2) \; &= F_X(2) \; = \; \int_{-\infty}^{2} f_X(y)\,dy \; = \; \int_{0}^{2} e^{-y}\,dy \\ &= \big[-e^{-x} \big]_0^2 \; = \; -e^{-2} + 1 \; \approx \; 0.865 \end{aligned}$$

In R: `pexp(2,rate=1)`

(b) Sei X standardnormalverteilt, dann

$$P(-2 \leq X \leq 2) \; = \; P(-2 < X \leq 2) \; = \; \Phi(2) - \Phi(-2).$$

Es gibt keine elementar darstellbare Stammfunktion zur Dichte der Standardnormalverteilung; die Werte der Verteilungsfunktion Φ liegen tabelliert vor, bzw. sind in R implementiert: `pnorm(2,mean=0,sd=1)-pnorm(-2,mean=0,sd=1)` liefert ca. 0.954 als Ergebnis.

Definition 7.29 (Erwartungswert und Varianz für stetige Zufallsvariablen)**.** Für eine Zufallsvariable X mit Dichte $f(x)$ ist der *Erwartungswert* definiert als

$$E(X) \; = \; \int_{-\infty}^{\infty} x \cdot f(x)\,dx,$$

sofern das Integral auf der rechten Seite wohldefiniert ist.

Die *Varianz* ist definiert als

$$\mathrm{Var}(X) \; = \; \int_{-\infty}^{\infty} \big(x - E(X)\big)^2 \cdot f(x)\,dx,$$

sofern das Integral wohldefiniert ist.

Satz 7.30. *Ist $u : \mathbb{R} \to \mathbb{R}$ eine Funktion, so gilt*

$$E\big(u(X)\big) \; = \; \int_{-\infty}^{\infty} u(x) \cdot f(x)\,dx,$$

sofern das Integral wohldefiniert ist.

Bemerkung 7.31. Satz 7.16 und Satz 7.19 über die Eigenschaften des Erwartungswertes bzw. der Varianz gelten auch für stetige Zufallsvariablen, ebenso die Ungleichung von Tschebyscheff und das Gesetz der großen Zahl.

Erwartungswert und Varianz wichtiger stetiger Verteilungen (7.32)

Die Zufallsvariable X habe eine...

1. Gleichverteilung auf $[a, b]$:

$$E(X) = \frac{a+b}{2}, \qquad \mathrm{Var}(X) = \frac{(b-a)^2}{12}$$

2. Exponentialverteilung mit Parameter $\lambda > 0$:

$$E(X) = \frac{1}{\lambda}, \qquad \mathrm{Var}(X) = \frac{1}{\lambda^2}$$

3. Normalverteilung mit Parametern $\mu \in \mathbb{R}$, $\sigma^2 > 0$

$$E(X) = \mu, \qquad \mathrm{Var}(X) = \sigma^2.$$

Definition 7.33. Sei X eine Zufallsvariable mit stetiger Verteilung. Das α-*Quantil* von X ist der Wert q_α mit

$$F_X(q_\alpha) = P(X \leq q_\alpha) = \alpha$$

Die Quantile speziell der Standardnormalverteilung werden mit z_α bezeichnet.

In R werden Quantile mit `q"Verteilungsname"` bestimmt, z.B. liefert `qnorm(0.25,mean=0,sd=1)` das 25%-Quantil der Standardnormalverteilung. `qnorm`

Beispiel 7.34. Sei X standardnormalverteilt. Dann ist $q_{0.25} = -0.675$, $q_{.75} = 0.675$, somit ist der Interquartilsabstand $d_Q = q_{0.75} - q_{0.25} = 1.35$. Betrachte (wie beim Boxplot) die Schranke $q_{0.25} - 1.5 \cdot d_Q = -2.7$. Es ist $F_X(-2.7) = 0.0034$, d.h. eine standardnormalverteilte Zufallsvariable nimmt nur mit einer Wahrscheinlichkeit kleiner als 0.34% einen Wert unterhalb der Schranke des unteren Whiskers an.

7.4. Kurz-Befehlsreferenz

Im Folgenden kann anstelle von `binom` der Name beliebiger in R implementierter Verteilungsklassen benutzt werden (siehe Abschnitte „Wichtigste Verteilungen “für weitergehende Informationen).

`dbinom`	Wahrscheinlichkeitsfunktion der Binomialverteilung
`rbinom`	generiert binomialverteilte Zufallsvariablen
`pbinom`	Verteilungsfunktion der Binomialverteilung
`qbinom`	Quantile der Binomialverteilung
`curve`	zeichnet Graphen einer stetigen Funktion
`function`	ermöglicht das Definieren von Funktionen in R

Teil III.
Schließende Statistik

8. Testtheorie

Grundannahme: Die beobachteten Daten $x_1, \ldots, x_n$ sind Realisierungen unabhängiger, identisch verteilter Zufallsvariablen $X_1, \ldots, X_n$, d.h. $x_1 = X_1(\omega)$, $\ldots$, $x_n = X_n(\omega)$.
Notation: $x = (x_1, \ldots, x_n)$, $X = (X_1, \ldots, X_N)$.

Die Verteilung der X_i wird im Folgenden nur bis auf einen (oder mehrere) unbekannte Parameter gegeben sein, z.B.

$$X_i \sim \mathcal{N}(\mu, \sigma^2),$$

mit μ und / oder σ^2 unbekannt.
Ziel ist es, auf Grundlage der beobachteten Daten eine Entscheidung zwischen der *Nullhypothese* H_0 (über unbekannten Parameter) und *Alternativhypothese* H_1 zu treffen.

Beispiel 8.1. Ist eine Münze fair? Dazu werfen wir die Münze n-mal, $X_i = 1$ entspreche Kopf im i-ten Wurf. Dann sind $X_1, \ldots, X_n$ stochastisch unabhängig und identisch $B(1, p)$-verteilt; mit unbekanntem Parameter $p \in (0, 1)$. Aufgrund der beobachteten Ergebnisse $x_1, \ldots, x_n$ wollen wir zwischen

$$H_0 : \ p = \frac{1}{2} \qquad \text{„Münze ist fair“}$$

und

$$H_1 : \ p \neq \frac{1}{2} \qquad \text{„Münze ist unfair “}$$

entscheiden.

Definition 8.2 (Fehler 1./2. Art)**.** Bei der Entscheidung zwischen H_0 und H_1 können folgende Fehler auftreten:

Wirklichkeit \ Entscheidung für	Nullhypothese	Alternative
Nullhypothese	$\surd$	Fehler 1. Art
Alternative	Fehler 2. Art	$\surd$

Beispiel 8.3. In obigem Beispiel 8.1 kann also eine in Wirklichkeit faire Münze aufgrund der Beobachtungen irrtümlicherweise für unfair gehalten werden (Fehler 1. Art); oder eine in Wirklichkeit unfaire Münze aufgrund der Beobachtungen für fair gehalten werden (Fehler 2. Art).

Ablauf eines statistischen Tests (8.4)

(a) Treffe eine Annahme über die Verteilung der beobachteten Zufallsvariablen

(b) Formuliere H_0 und H_1 als Bedingungen an den unbekannten Parameter

(c) Lege ein *Irrtumsniveau* $\alpha \in (0,1)$ fest

(d) Wähle geeignete *Teststatistik* $T(X)$ und bestimme anhand der Teststatistik *Annahme-* und *Verwerfungsbereich* für H_0 derart, dass die Wahrscheinlichkeit des Fehlers 1. Art durch α beschränkt ist

(e) Berechne $T(x)$ anhand der Daten. H_0 wird beibehalten, wenn $T(x)$ im Annahmebereich liegt; H_0 wird abgelehnt und H_1 angenommen, wenn $T(x)$ im Verwerfungsbereich liegt.

Beispiel 8.5 (Zweiseitiger Gauß-Test)**.**

(a) Verteilungsannahme: $X_1, \ldots, X_n$ sind stochastisch unabhängig und identisch $\mathcal{N}(\mu, \sigma^2)$-verteilt, wobei $\sigma^2 > 0$ bekannt sei; der Parameter $\mu \in \mathbb{R}$ hingegen unbekannt.

(b) Getestet werden soll, ob der unbekannte Parameter μ einem Referenzwert μ_0 entspricht, oder von diesem abweicht: Teste

$$H_0 : \mu = \mu_0 \quad \text{gegen} \quad H_1 : \mu \neq \mu_0.$$

(c) Wir wählen als Irrtumsniveau $\alpha = 5\%$.

(d) Nach dem Gesetz der großen Zahl (Satz 7.22) ist das Stichprobenmittel $\bar{x}$ ein sinnvoller *Schätzer* für den Erwartungswert μ, den hier zu betrachtenden unbekannten Parameter. Es liegt also nahe, die Größe

$$|\bar{x} - \mu_0|$$

zu betrachten, und bei nur geringer Abweichung für H_0 zu entscheiden, bei größerer Abweichung für H_1. Aus „technischen“ Gründen ist es sinnvoll, die reskalierte Teststatistik

$$T(X) \;:=\; \sqrt{n} \cdot \frac{\overline{X} - \mu_0}{\sigma}$$

zu betrachten, da für diese gezeigt werden kann, dass $T(X)$ bei Vorliegen von H_0 eine $\mathcal{N}(0,1)$-Verteilung besitzt. Unverändert bleibt, dass kleine Werte von $T(X)$ für das Vorliegen von H_0 sprechen, große Werte dagegen.

Annahme- bzw. Verwerfungsbereich sollen also von folgender Form sein:

$$\begin{aligned} a \leq T(X) \leq b \quad &\Rightarrow \quad \text{Annahme von } H_0 \\ T(X) < a \text{ oder } T(X) > b \quad &\Rightarrow \quad \text{Verwerfung von } H_0\text{, Annahme von } H_1 \end{aligned}$$

Zur Bestimmung von a, b wird die Bedingung verwendet, dass die Wahrscheinlichkeit für den Fehler 1. Art durch α beschränkt sein soll. D.h., die Wahrscheinlichkeit (unter Vorliegen von H_0), dass $T(X)$ im Verwerfungsbereich liegt, soll durch α beschränkt sein. Dies ist gewährleistet, wenn a und b gerade das $\alpha/2$-Quantil $z_{\alpha/2}$ bzw. $1-\alpha/2$-Quantil $z_{1-\alpha/2}$der Standardnormalverteilung ist, siehe Abbildung 22. Die Quantile $z_{0.025}$ und $z_{0.975}$ erhalten wir in R mit dem Befehl `qnorm(0.025,mean=0,sd=1)` bzw. `qnorm(0.975,mean=0,sd=1)`. Somit sind Annahme- und Verwerfungsbereich wie folgt gegeben:

$$\begin{aligned} -1.96 = z_{\alpha/2} \leq T(X) \leq z_{1-\alpha/2} = 1.96 \quad &\Rightarrow \quad \text{Annahme von } H_0 \\ T(X) < -1.96 \text{ oder } T(X) > 1.96 \quad &\Rightarrow \quad \text{Verwerfung von } H_0\text{, Annahme von } H_1 \end{aligned}$$

(e) Gegeben seien nun folgende $n = 10$ Beobachtungen:

```
0.94 -2.73  4.42 -1.42 -0.38  2.66  3.34 -1.71  0.34  2.58
```

Wir wollen die Verträglichkeit mit der Nullhypothese $H_0 : \mu = 1$ testen, es sei $\sigma^2 = 4$ bekannt. Berechne die Teststatistik (die Daten mögen im Vektor `x` vorliegen)

```
> T<-sqrt(10)*(mean(x)-1)/2
> T
[1] -0.3099032
```

Der Wert von $T(x)$ liegt also innerhalb des Annahmebereichs, wir nehmen H_0 an.

8.1. Wichtige Tests

Im Folgenden stellen wir in kompakter Form die für Anwendungen wichtigsten Tests vor. Wir folgen dabei der exzellenten Darstellung in [1, Kapitel 10].

Gauß-Test (8.6)

Annahme / Voraussetzung: Es seien $X_1, \ldots, X_n$ stochastisch unabhängig und identisch $\mathcal{N}(\mu, \sigma^2)$-verteilt, σ^2 sei bekannt, $\mu \in \mathbb{R}$ sei unbekannt.

Betrachtet werde eines der folgenden Testprobleme:

(i) $H_0 : \mu = \mu_0$ gegen H_1: $\mu \neq \mu_0$

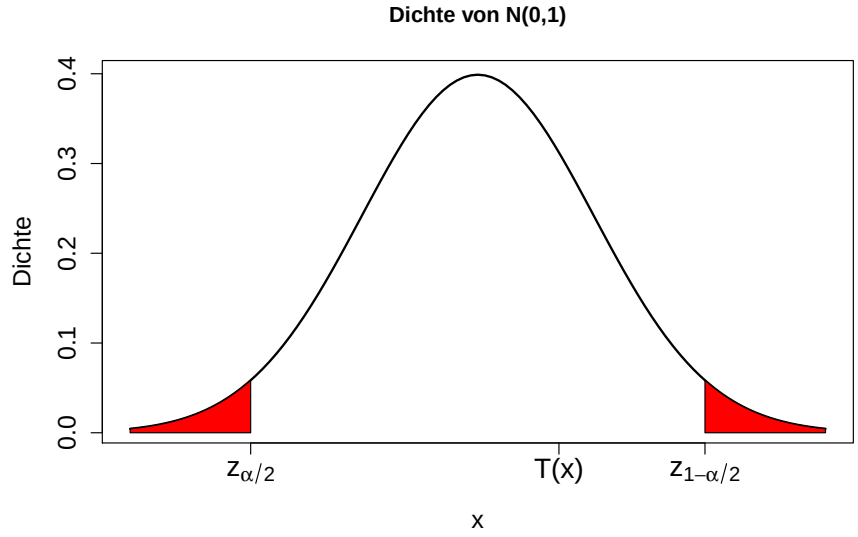


Abbildung 22: Zur Konstruktion des Annahme- und Verwerfungsbereiches. Die rot schraffierten Flächen addieren sich zu α.

(ii) $H_0 : \mu = \mu_0$ gegen H_1: $\mu < \mu_0$ ODER $H_0 : \mu \geq \mu_0$ gegen H_1: $\mu < \mu_0$

(iii) $H_0 : \mu = \mu_0$ gegen H_1: $\mu > \mu_0$ ODER $H_0 : \mu \leq \mu_0$ gegen H_1: $\mu > \mu_0$

Betrachte die Teststatistik

$$T(x) \;=\; \sqrt{n} \cdot \frac{\overline{x} - \mu_0}{\sigma}.$$

Die Nullhypothese H_0 wird **verworfen**, falls

(i) $|T(x)| > z_{1-\alpha/2}$

(ii) $T(x) < z_{\alpha}$

(iii) $T(x) > z_{1-\alpha}$

Bemerkung 8.7. Die obigen Testprobleme werden wie folgt bezeichnet:

(i) Zweiseitige Alternative,

(ii) Linksseitige Alternative,

(iii) Rechtsseitige Alternative.

In Beispiel 8.1 sind die beobachteten Zufallsvariablen $B(1,p)$-verteilt. In dieser Situation kann bei hinreichend großem Stichprobenumfang analog zum Gauß-Test verfahren werden; dies ist eine Konsequenz des Zentralen Grenzwertsatzes, welcher besagt, dass Summen unabhängig, identisch verteilter Zufallsvariablen bei hinreichend großer Anzahl von Summanden approximativ normalverteilt sind, siehe Satz 9.1 weiter unten.

Approximativer Binomial-Test (8.8)

Annahme / Voraussetzung: Es seien $X_1, \ldots, X_n$ stochastisch unabhängig und identisch $B(1,p)$-verteilt, $p \in (0,1)$ sei unbekannt. Der Stichprobenumfang sei hinreichend groß (Faustregel: $n \geq 30$)

Betrachtet werde eines der folgenden Testprobleme:

(i) $H_0 : p = p_0$ gegen H_1: $p \neq p_0$

(ii) $H_0 : p = p_0$ gegen H_1: $p < p_0$ ODER $H_0 : p \geq p_0$ gegen H_1: $p < p_0$

(iii) $H_0 : p = p_0$ gegen H_1: $p > p_0$ ODER $H_0 : p \leq p_0$ gegen H_1: $p > p_0$

Betrachte die Teststatistik

$$T(x) \;=\; \sqrt{n} \cdot \frac{\overline{x} - p_0}{\sqrt{p_0(1-p_0)}}.$$

Die Nullhypothese H_0 wird **verworfen**, falls

(i) $|T(x)| > z_{1-\alpha/2}$

(ii) $T(x) < z_\alpha$

(iii) $T(x) > z_{1-\alpha}$

In den meisten Anwendungssituationen wird die Standardabweichung der beobachteten Zufallsvariablen nicht bekannt sein, sie muss vielmehr durch die empirische Standardabweichung *geschätzt* werden (siehe Def. 3.9). Dies führt auf den t-Test. Hierbei sei an die Definition der empirischen Standardabweichung erinnert: Gegeben Daten $x = (x_1, \ldots, x_n)$ ist

$$s(x) = \sqrt{\frac{1}{n-1} \sum_{i=1}^{N} (x - \overline{x})^2}$$

Zur Formulierung des t-Tests benötigen wir noch folgende Definition:

Definition 8.9. Es seien $X_1, \ldots, X_n$ stochastisch unabhängige, standardnormalverteilte

Zufallsvariablen. Die Verteilung der Zufallsvariablen

$$\sqrt{n}\frac{\bar{X}}{S(X)} \;=\; \sqrt{n}\frac{\frac{1}{n}\sum_{i=1}^{n} X_i}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{N}(X-\frac{1}{n}\sum_{i=1}^{n} X_i)^2}}$$

heißt *Student'sche t-Verteilung* mit $n-1$ Freiheitsgraden, kurz t_{n-1}-Verteilung.
Die Quantile der t_{n-1}-Verteilung werden mit $t_{n-1,\alpha}$ bezeichnet.

Der Verteilungsname der t-Verteilung in R ist `t`. Beispielsweise lassen sich die Quantile der t_{n-1}-Verteilung mit Hilfe des Befehls `qt(`α`,df=`$n-1$`)` bestimmen. Für große Werte von n ($n \geq 30$) weichen sie nur noch sehr gering von den Quantilen der Standardnormalverteilung ab.

`qt`

t-Test (8.10)

Annahme / Voraussetzung: Es seien $X_1, \ldots, X_n$ stochastisch unabhängig und identisch $\mathcal{N}(\mu, \sigma^2)$-verteilt, sowohl μ als auch σ^2 seien unbekannt.

Betrachtet werde eines der folgenden Testprobleme:

(i) $H_0 : \mu = \mu_0$ gegen H_1: $\mu \neq \mu_0$

(ii) $H_0 : \mu \geq \mu_0$ gegen H_1: $\mu < \mu_0$

(iii) $H_0 : \mu \leq \mu_0$ gegen H_1: $\mu > \mu_0$

Betrachte die Teststatistik

$$T(x) \;=\; \sqrt{n} \cdot \frac{\overline{x} - \mu_0}{s(x)}.$$

Die Nullhypothese H_0 wird **verworfen**, falls

(i) $|T(x)| > t_{n-1,1-\alpha/2}$

(ii) $T(x) < t_{n-1,\alpha}$

(iii) $T(x) > t_{n-1,1-\alpha}$

Bemerkung 8.11. Die Verteilungen $\mathcal{N}(0,1)$ und t_{n-1} haben symmetrische Dichtefunktionen. Daher gilt für alle $\alpha \in (0,1)$

$$\begin{aligned} z_\alpha &= -\, z_{1-\alpha} \\ t_{N-1,\alpha} &= -\, t_{N-1,1-\alpha} \end{aligned}$$

Zur Wahl von H_0 und H_1 (8.12)

- Bei Ablehnung von H_0 (Annahme von H_1) können wir bis auf eine Irrtumswahrscheinlichkeit $\leq \alpha$ (Wahrscheinlichkeit des Fehlers 1. Art) sicher sein, dass wir die richtige Entscheidung getroffen haben, und H_1 tatsächlich gilt. Die Daten sprechen also *signifikant* gegen H_0 und für H_1.
- Bei Beibehaltung von H_0 wissen wir lediglich, dass die Daten nicht signifikant gegen H_0 sprechen. Wir können im Allgemeinen aber **nicht sicher sein**, dass H_0 tatsächlich gilt, da die Wahrscheinlichkeit für den Fehler 2. Art (irrtümliche Annahme von H_0) groß sein könnte; dies ist insbesondere für kleine Stichprobengrößen n der Fall.

$\Rightarrow$ Wenn wir also mittels der Daten nachweisen wollen, dass eine Aussage bis auf eine kleine Irrtumswahrscheinlichkeit α tatsächlich gilt, dann wählen wir diese Aussage als Alternativhypothese H_1, mit dem Ziel, dass die Daten zur Ablehnung von H_0 führen.

Beispiel 8.13. Die Wirkung eines Präparats auf den systolischen Blutdruck wurde durch Blutdruckmessungen an 20 Probanden vor und nach Gabe des Präparats ermittelt. Es ergaben sich die folgenden Werte für die Blutdruckänderung (Differenz aus dem End- und Anfangswert, in mmHg):

-23,	-5,	-18,	15,	-9,	-4,	-6,	6,	-12,	-11,
-6,	-28,	22,	3,	27,	-31,	2,	-33,	18,	-16.

Wie nehmen an, dass die Blutdruckänderung normalverteilt ist. Lässt sich aus den Daten mit einer Irrtumswahrscheinlichkeit von höchstens 5% schließen, dass die mittlere Blutdruckänderung eine signifikante Abnahme anzeigt?

Lösung: Die Blutdruckänderung ist $\mathcal{N}(\mu, \sigma^2)$-verteilt mit unbekannten μ und σ. Da nach einem Nachweis (mit geringer Irrtumswahrscheinlichkeit) einer bestimmten Aussage gefragt ist, wählen wir diese Aussage als Alternativhypothese H_1. Wir führen einen einseitigen t-Test durch und testen

$$H_0 \colon \mu \geq 0 \qquad \text{gegen} \qquad H_1 \colon \mu < 0$$

zum Niveau $\alpha = 5\%$. Der Stichprobe entnimmt man $N = 20$, $\bar{x} = -5.45$ und die empirische Standardabweichung

$$\begin{aligned} s(x) = \sqrt{s(x)^2} &= \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} \\ &= \sqrt{\frac{1}{19}\sum_{i=1}^{20}(x_i - (-5.45))^2} = 17.20. \end{aligned}$$

Damit findet man den Wert

$$\begin{aligned} T(x) &= \sqrt{N}\frac{(\bar{x} - \mu_0)}{s(x)} = \sqrt{20}\frac{-5.45}{17.20} \\ &= -1.417 \end{aligned}$$

Wir bestimmen $t_{19,0.05}$ mit dem R-Befehl `qt(0.05, df=19)`. Es ist $t_{19,0.05} = -1.729133$. Da $T(x) = -1.42 > -1.73 = t_{19,0.05}$, ist die beobachtete Unterschreitung des Sollwerts $\mu_0 = 0$ auf dem 5%-Niveau nicht signifikant. Wir behalten also H_0 bei.

8.2. Testen mit R

Die oben genannten Tests sind auch direkt in R implementiert. Die im folgenden beschriebenen Aufrufe liefern als Hauptinformation den sogenannten p-Wert:

Definition 8.14 (p-Wert)**.** Der *p-Wert* ist definiert als die Wahrscheinlichkeit, unter H_0 den beobachteten Teststatistik-Wert, oder einen in Richtung der Alternative extremeren Wert zu erhalten.

Ist der p-Wert kleiner oder gleich dem vorgegebenen Irrtumsniveau α, so wird H_0 verworfen. Ansonsten behält man H_0 bei.

Implementation der Tests in R (8.15)

Die Beobachtungen mögen in einem Vektor `x` vorliegen. Die Art des Testproblems wird jeweils über das Argument `alternative=...` spezifiziert; hierbei gibt es die Optionen `"two.sided"` für eine zweiseitige Alternative, `"less"` für eine linksseitige Alternative, sowie `"greater"` für eine rechtsseitige Alternative.

1. Der **Gauß-Test** ist in R nicht implementiert, da in praktischen Anwendungen stets σ^2 aus den Daten geschätzt, und somit der t-Test verwendet wird.

2. Der **exakte Binomialtest** lässt sich in R mit Hilfe des Befehls `binom.test` durchführen.`binom.test` Als Argumente werden `p=`p_0 benötigt, sowie entweder die *Anzahl* `x` der beobachteten Erfolge zusammen mit der Anzahl `n` der Durchführungen; oder ein Vektor `x` mit Einträgen 0 bzw. 1, der die Abfolge von Misserfolgen und Erfolgen wiedergibt.

   Beispiel: Beim $n = 30$-maligen Werfen einer Münze (vgl. Beispiel 8.1) haben wir $x = 18$ mal Kopf gesehen. Wir testen nun zum Irrtumsniveau $\alpha = 10\%$, ob die Münze fair ist $H_0 : p = p_0 := 1/2$ oder nicht.

```
> binom.test(x=18, n=30, alternative="two.sided")

        Exact binomial test

data:  18 and 30
number of successes = 18, number of trials = 30, p-value = 0.3616
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.4060349 0.7734424
```

```
sample estimates:
probability of success
                  0.6
```

Der p-Wert von 0.3616 liegt also über unserem Irrtumsniveau $\alpha = 0.1$, somit wird H_0 beibehalten.

R liefert hier noch weitere Informationen, nämlich ein *Konfidenzintervall* („95 percent confidence interval“), in dem anhand der Daten der wahre Parameter p mit einer Wahrscheinlichkeit von 95% liegt, sowie einen Schätzwert („sample estimates“) für den Parameter p.

3. Der **t-Test** lässt sich in R mit Hilfe des Befehls `t.test` durchführen. Als Argumente werden `mu=`μ_0 benötigt, sowie der Vektor `x` der beobachteten Daten. `t.test`

   Beispiel: Wir haben 25 Brötchen gekauft und wollen nun zum Irrtumsniveau $\alpha = 5\%$, ob das mittlere Brötchengewicht bei mindestens 100 g liegt, d.h. $H_0 : \mu \geq \mu_0 = 100$, oder darunter. Wir messen folgende Brötchengewichte:

```
104  78  88 101 111  87  81  73  96  90  48 103  88  62  85
86  72  92  98 103  79  67  63  94  82 108  81  97  93  92
```

   Diese seien im Vektor `x` abgelegt.

```
> t.test(x,mu=100)

        One Sample t-test

data:  round(x)
t = -4.9498, df = 29, p-value = 2.917e-05
alternative hypothesis: true mean is not equal to 100
95 percent confidence interval:
 81.25160 92.21507
sample estimates:
mean of x
 86.73333
```

   Der p-Wert $2.917 \cdot 10^{-5} = 0.00003$ liegt weit unterhalb des Irrtumsniveaus α, wir verwerfen also H_0.

Bemerkung 8.16. Eine weitere Interpretation des p-Wertes: Der p-Wert gibt die *a-posteriori* Wahrscheinlichkeit des Fehlers 1. Art an, d.h. bei Vorliegen der Beobachtungen wird die Wahrscheinlichkeit berechnet, dass diese bei Vorliegen der Hypothese entstanden sein könnten, man sich also irrtümlich für H_1 entscheiden würde.

8.3. Kurz-Befehlsreferenz

`binom.test`	Exakter Binomialtest
`t.test`	t-Test

9. Verknüpfung zur explorativen Datenanalyse

Wir erinnern an das Gesetz der großen Zahlen (Satz 7.22): Je mehr Realisierungen einer zufälligen Größe vorliegen, desto geringer ist die Wahrscheinlichkeit, dass der Mittelwert der Realisierungen vom Erwartungswert (um mehr als ϵ) abweicht.

Im Folgenden seien stets unabhängig, identisch verteilte Zufallsvariablen $X_1, \ldots, X_n$ gegeben; für einen Vektor von Realisierungen schreiben wir $x = (x_1, \ldots, x_n)$.

Es gelten folgende Verschärfungen von Satz 7.22:

Satz 9.1 (Version des zentralen Grenzwertsatzes)**.** *Seien $X_1, \ldots, X_n$ stochastisch unabhängige, identisch verteilte Zufallsvariablen mit $EX_i = \mu$ und $\mathrm{Var}(X_i) = \sigma^2$. Dann gilt für hinreichend großes n und jedes $c > 0$:*

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \leq \frac{c}{\sqrt{n}}\right) \;\approx\; \Phi\Big(\frac{c}{\sigma}\Big) - \Phi\Big(-\frac{c}{\sigma}\Big) \;=\; 2\Phi\Big(\frac{c}{\sigma}\Big) - 1.$$

D.h., die Wahrscheinlichkeit, eine Abweichung von höchstens $c/\sqrt{n}$ zu sehen (bei n Beobachtungen) ist näherungsweise durch die Wahrscheinlichkeit gegeben, dass eine standardnormalverteilte Zufallsvariable Werte im Bereich $[-c/\sigma, c/\sigma]$ annimmt. Interessant ist insbesondere $c = 3\sigma$, dann beträgt diese Wahrscheinlichkeit über 99%. Die auftretenden Abweichungen sind also mit sehr hoher Wahrscheinlichkeit kleiner als $3\sigma/\sqrt{n}$, bei n Beobachtungen. Obige Formel kann benutzt werden, um einen Mindeststichprobenumfang zu bestimmen, wenn eine geforderte Genauigkeit (der Approximation von μ durch das Stichprobenmittel) eingehalten werden soll.

Satz 9.2 (Starkes Gesetz der großen Zahl)**.** *Seien $X_1, X_2, \ldots,$ stochastisch unabhängige, identisch verteilte Zufallsvariablen mit $EX_i = \mu$ und $\mathrm{Var}(X_i) = \sigma^2$, $i \geq 1$. Dann gilt mit Wahrscheinlichkeit 1: Mit wachsendem n*

- *konvergiert das Stichprobenmittel $\overline{x}$ gegen μ,*
- *konvergiert die Stichprobenvarianz $s^2(x)$ gegen σ^2.*

D.h., durch Hinzunahme weiterer unabhängig wiederholter Beobachtungen kann eine Verbesserung der Schätzung erreicht werden. Wir sagen: $\overline{x}$ und $s^2(x)$ sind *stark konsistente Schätzer* für μ bzw. σ^2.

Nicht nur Erwartungswert und Varianz, auch die Verteilungsfunktion lässt sich aus den Daten konsistent schätzen, mit Hilfe der empirischen Verteilungsfunktion.

Definition 9.3. Gegeben Beobachtungen $(x_1, \ldots, x_n)$ eines quantitativen Merkmals, definiere die *empirische Verteilungsfunktion* $F_n : \mathbb{R} \to [0, 1]$ durch

$$F_n(x) = \frac{1}{n} \cdot \#\{i \,:\, x_i \leq x\}.$$

$F_n(x)$ gibt also die relative Häufigkeit (=den Anteil) von Beobachtungen kleiner gleich x an.

Satz 9.4 (Satz von Glivenko-Cantelli)**.** *Seien $X_1, \ldots, X_n$ stochastisch unabhängige, identisch verteilte Zufallsvariablen. Es bezeichne $F(x) = P(X_1 \leq x)$ die (für jede der beteiligten Zufallsvariablen identische) Verteilungsfunktion . Dann gilt mit Wahrscheinlichkeit 1: Für wachsendes n konvergiert $F_n(x)$ gegen $F(x)$; und dies sogar gleichmäßig in $x \in \mathbb{R}$.*

In R wird die empirische Verteilungsfunktion (für einen Datensatz `x`) mit Hilfe des Befehls `ecdf(x)` erzeugt. `ecdf`

Die rechte Graphik in Abbildung 23 wurde mit folgendem Befehl erzeugt:

```
x<-rnorm(20)
plot(ecdf(x),main="Empirische vs. Theoretische Verteilungsfunktion",ylab="")
y<-rnorm(50)
plot(ecdf(y),add=T, col="blue")
curve(pnorm(x),-3,4,add=T,col="red",lwd=2)
legend("topleft",,c(expression(F[20](x)),expression(F[50](x)),"F(x)"),
       col=c("black","blue","red"),lwd=c(1,1))
```

Durch graphischen Vergleich der empirischen Verteilungsfunktion mit verschiedenen (theoretischen) Verteilungsfunktionen kann eine Vermutung über die dem beobachteten Zufallsmechanismus zugrunde liegende Verteilung aufgestellt werden. Z.B. kann vermutet werden, dass die Körpergröße 14jähriger Jungen näherungsweise $\mathcal{N}(155, 100)$-verteilt ist.

Zur Überprüfung einer solchen Hypothese dient der Kolmogorov-Smirnov-Test.

Kolmogorov-Smirnov-Test (9.5)

Annahme / Voraussetzung: Es seien $X_1, \ldots, X_n$ stochastisch unabhängige, identisch verteilte Zufallsvariablen, die gemäß einer *unbekannten, stetigen* Verteilung verteilt seien. Für eine stetige Referenzverteilung Q wird folgendes Testproblem betrachtet:

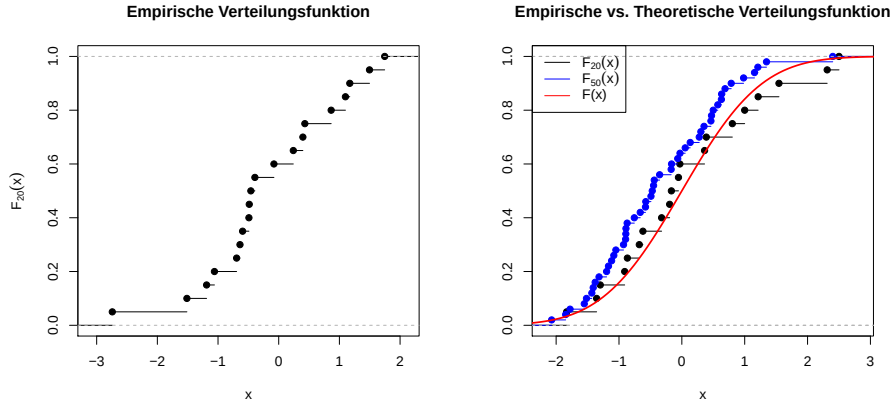


Abbildung 23: Empirische Verteilungsfunktion für Standardnormalverteilte Beobachtungen

H_0: Die Zufallsvariablen sind gemäß Q verteilt

gegen

H_1: Die Zufallsvariablen sind nicht gemäß Q verteilt.

Bezeichnet F_Q die Verteilungsfunktion einer gemäß Q verteilten Zufallsvariable, so wird folgende Teststatistik betrachtet:

$$T = \max_{x \in \mathbb{R}} |F_n(x) - F_Q(x)|$$

Die Nullhypothese wird **verworfen**, falls T große Werte annimmt.

In obigem Beispiel wäre $Q = \mathcal{N}(155, 100)$, es wird also eine konkrete Verteilung mit fixierten Parametern gewählt. Einen exakten Ablehnungsbereich können wir hier nicht angeben, stattdessen verweisen wir auf die Implementierung in R.

Der Befehl lautet `ks.test` ; als Argumente müssen der Beobachtungsvektor `x` sowie die Referenzverteilung Q in Form `p"Verteilungsname"`, sowie die zu wählenden Parameter von Q übergeben werden. Z.B. testet `ks.test`

```
ks.test(x,"pnorm",mean=155,sd=10)
```

ob die beobachteten Daten von einer Normalverteilung mit Parametern $\mu = 155$ und Standardabweichung $\sigma = 10$ stammen könnten.

Aus dem Satz von Glivenko-Cantelli folgt insbesondere, dass (bei wachsendem Stichprobenumfang n) mit Wahrscheinlichkeit 1 die empirischen Quantile (siehe Def. 3.3) gegen die theoretischen Quantile (siehe Def. 7.33) konvergieren; also stark konsistente

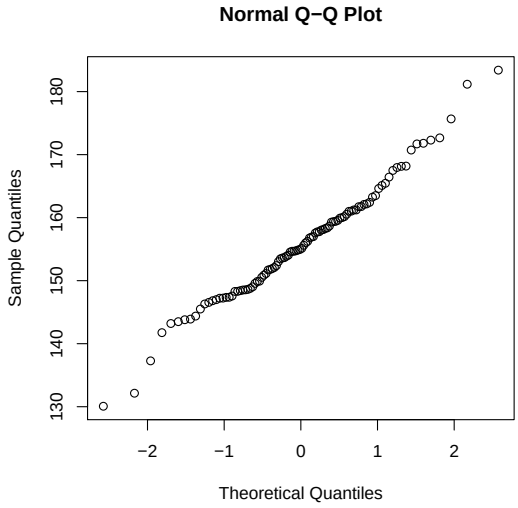


Abbildung 24: Normal-Quantil-Plot für 100 Realisierungen einer $\mathcal{N}(0,1)$-verteilten Zufallsvariable.

Schätzer darstellen. Dies liefert eine weitere Möglichkeit, die Daten auf Vorliegen einer bestimmten Verteilung zu untersuchen; im Folgenden insbesondere auf Vorliegen einer Normalverteilung.

Normal-Quantil-Plot (9.6) Gegeben eine *geordnete* Stichprobe $(x_{(1)}, \ldots, x_{(n)})$. Für $i = 1, \ldots, n$ berechne die $(i-0.5)/n$-Quantile $z_(i)$ der Standardnormalverteilung. Der *Normal-Quantil-Plot* besteht aus den Punkten

$$\big(z_{(1)}, x_{(1)}\big), \ldots, \big(z_{(n)}, x_{(n)}\big)$$

im $z-x$-Koordinatensystem.

In R wird ein Normal-Quantil-Plot zum Datensatz `x` mit dem Befehl `qqnorm(x)` gezeichnet. `qqnorm`

Bemerkung 9.7. Ist die beobachtete Zufallsgröße approximativ normalverteilt mit Parametern μ und σ^2, so liegen die Punkte $\big(z_{(i)}, x_{(i)}\big)$ des Normal-Quantil-Plots in etwa auf der Geraden

$$x = \mu + \sigma \cdot z.$$

Zum Schluss betrachten wir noch einmal bivariate Merkmale. In Def. 4.4 haben wir den Korrelationskoeffizienten eingeführt, der eine Stärke des linearen Zusammenhangs beschreibt. Mit Hilfe des *Korrelationstests* kann die Nullhypothese „Es liegt kein linearer

Zusammenhang vor“gegen die Alternative „Die Korrelation ist positiv “getestet werden. Hier spielt wieder die t-Verteilung (siehe Def. 8.9) eine wichtige Rolle.

Im Folgenden bezeichnet $r = r(x, y)$ den empirischen Korrelationskoeffizienten (siehe Def. 4.4), berechnet aus bivariaten Daten $(x, y) = \big((x_1, y_1), \ldots, (x_n, y_n)\big)$.

Um die Hypothesen präzise zu formulieren, benötigen wir noch die theoretische Entsprechung von r, den *Korrelationskoeffizienten*.

Definition 9.8. Seien X und Y Zufallsvariablen. Dann ist die *Kovarianz* von X und Y definiert durch

$$\mathrm{Cov}(X, Y) \;:=\; \frac{1}{4}\left(\mathrm{Var}(X + Y) - \mathrm{Var}(X - Y)\right),$$

und daraus abgeleitet der Korrelationskoeffizient

$$\rho = \rho_{X,Y} \;:=\; \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

Bemerkung 9.9. Aus Satz 7.19 folgt, dass die Kovarianz, und damit auch der Korrelationskoeffizient für unabhängige Zufallsvariablen X und Y gleich Null sind. Die Umkehrung gilt jedoch nicht! Ist $\rho_{X,Y} = 0$, so heißt dies nur, dass kein *linearer* Zusammenhang zwischen X und Y besteht.

Die Interpretation von ρ ist analog zur Interpretation des empirischen Korrelationskoeffizienten, $\rho > 0$ entspricht einem positiven linearen Zusammenhang usw. Insbesondere ist r ein stark konsistenter Schätzer für ρ.

Korrelationstest (9.10)

Annahme / Voraussetzung: Es seien $(X_1, Y_1), \ldots, (X_n, Y_n)$ stochastisch unabhängig und identisch verteilte (Paare von) Zufallsvariablen, die jeweils normalverteilt seien.

Betrachtet werde eines der folgenden Testprobleme (hierbei $\rho = \rho_{X_1,Y_1}$):

(i) $H_0 : \rho = 0$ gegen H_1: $\rho \neq 0$

(ii) $H_0 : \rho \geq 0$ gegen H_1: $\rho < 0$

(iii) $H_0 : \rho \leq 0$ gegen H_1: $\rho > 0$

Betrachte die Teststatistik

$$T(x, y) \;=\; \sqrt{n-2}\,\frac{r(x, y)}{\sqrt{1 - r(x, y)^2}}$$

Die Nullhypothese H_0 wird **verworfen**, falls

(i) $|T(x, y)| > t_{n-2,1-\alpha/2}$

(ii) $T(x, y) < t_{n-2,\alpha}$

(iii) $T(x, y) > t_{n-2,1-\alpha}$

Gegeben zwei gleichgroße Beobachtungsvektoren `x` und `y`, wird der Korrelationstest in R mit folgendem Befehl durchgeführt: `cor.test`

```
cor.test(x,y,alternative="two.sided",method="pearson")
```

9.1. Kurz-Befehlsreferenz

`ecdf`	Bestimmung der empirischen Verteilungsfunktion
`ks.test`	Kolmogorov-Smirnov-Test auf Vorliegen einer bestimmten Verteilung
`qqnorm`	Normal-Quantil-Plot
`cor.test`	Korrelationstest

Anhang

Quellcode zu Beispiel 1.6

```
set.seed(0)
Durchmesser<-round(runif(40, 0.2, 12),1)
set.seed(0)
Resistenz<-sample(c("sensitiv", "intermediär", "resistent"),size=40,replace=T,
prob=c(23/40,8/40,9/40))
Resistenz<-ordered(Resistenz, levels=c("sensitiv", "intermediär", "resistent"))
set.seed(0)
Farbe<-sample(c("gelb", "weißlich", "braun", "orange", "farblos", "rosa", "grün"),
size=40,replace=T)
Farbe<-factor(Farbe)

Bakterien<-data.frame(Durchmesser, Resistenz, Farbe)
```

Quellcode zu Beispiel 4.3

```
set.seed(0)
Fliessgeschwindigkeit<-round(runif(20,0,1),2)
set.seed(0)
Sauerstoff<-round(12*Fliessgeschwindigkeit+rnorm(20),1)
set.seed(1)
Wassertemperatur<-round(runif(20,8,17),1)

Wasser<-data.frame(Sauerstoff,Fliessgeschwindigkeit,Wassertemperatur)
```

Literatur

[1] Fahrmeir, L. et. al., *Statistik*, Springer 2007.

[2] Krämer, W. *So lügt man mit Statistik*, Campus Verlag Frankfurt 2015.

[3] Müller, C. und Denecke, L., *Stochastik in den Ingenieurwissenschaften. Eine Einführung mit R*, Springer 2013.

[4] Neuhauser, H. et. al., *Referenzperzentile für anthropometrische Maßzahlen und Blutdruck aus der Studie zur Gesundheit von Kindern und Jugendlichen in Deutschland (KiGGS)*, Robert Koch-Institut, Berlin 2013.

[5] Riede, A., *Mathematik für Biowissenschaftler*, Springer 2015.

[6] Rudolf, M. und Kuhlisch, W., *Biostatistik*, Pearson Studium 2008.

[7] Shababa, B. *Biostatistics with R*, Springer 2012.

[8] Timischl, W., *Angewandte Statistik*, Springer 2013.