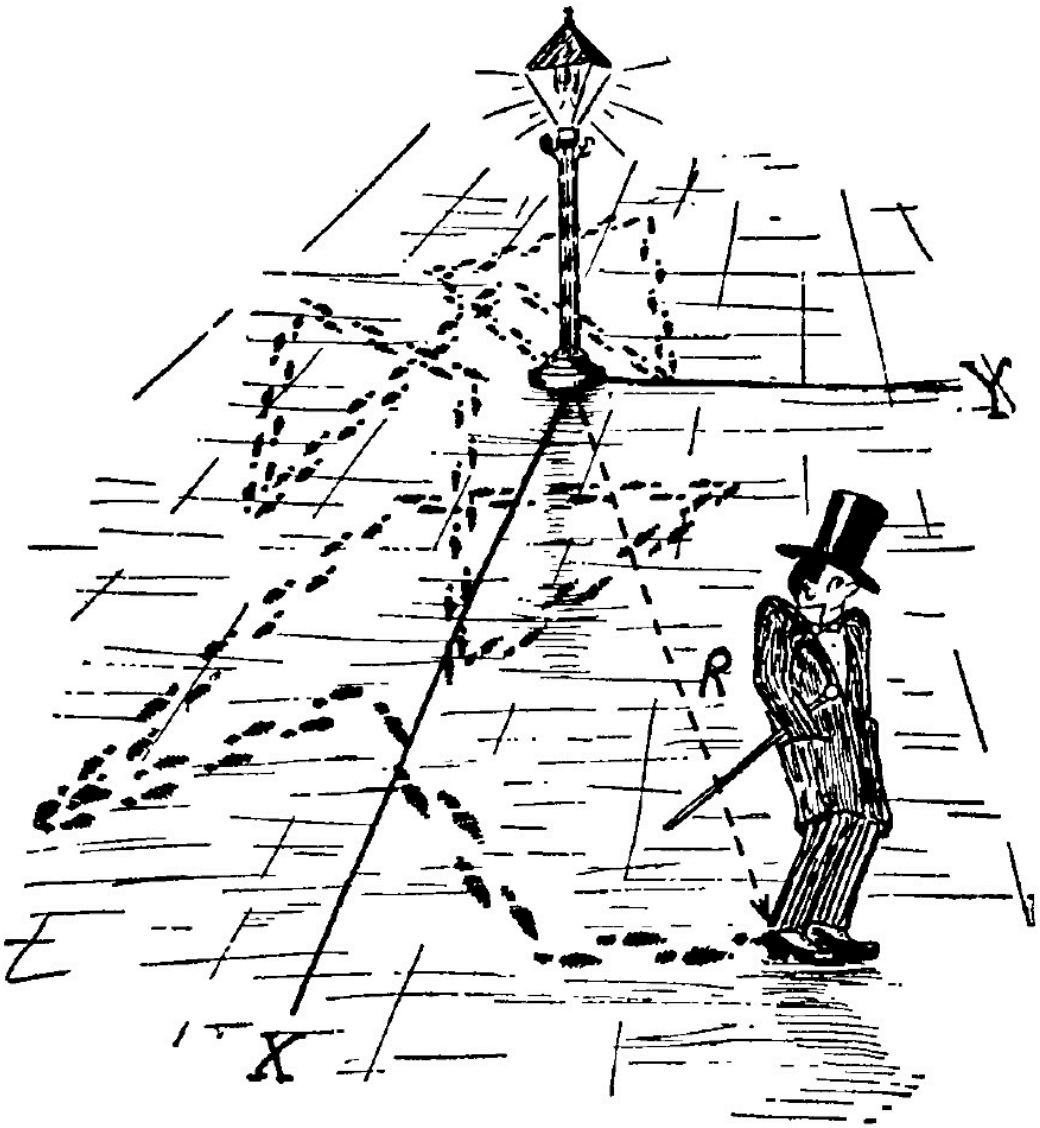
Mirko Birbaumer
Peter Büchel

Applied Statistics for Data Science

Vorlesungsskript HS 20



Hochschule Luzern Technik & Architektur

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Teil I.

Grundlagen

Kapitel 1.

Was ist applied statistics?

> *. . . as we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknown – the ones we don't know we don't know.*
>
> Donald Rumsfeld, former US Secretary of Defense

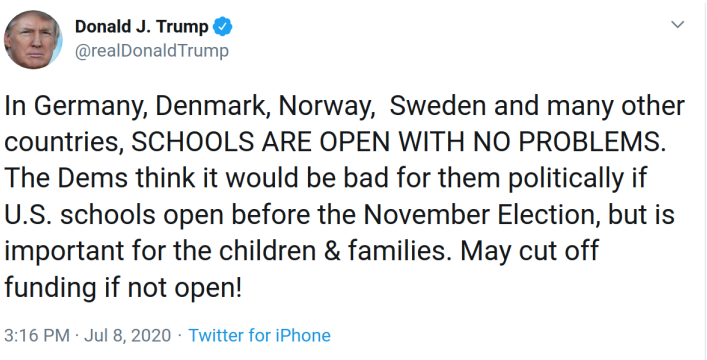
1.1. Einleitung

Applied statistics macht gerade das, was der Name besagt: Wir wenden Statistik an und zwar auf konkrete Alltagsprobleme. Hier ein Beispiel, das thematisch (*critical thinking*) zwar ein wenig ausserhalb dieses Modules liegt, aber den Vorteil hat, dass keine statistischen Vorkenntnisse verwendet werden.

Beispiel 1.1.1

Die folgenden Daten stammen alle vom 8. und 9. Juli 2020.

Es begann, wie sovieles, mit einem Tweet von Donald Trump:



Er wollte aus politischen Gründen unbedingt die Schulen, die aufgrund von Covid 19 geschlossen waren, wieder öffnen. Leider, für ihn, liegt das nicht in seiner Kompetenz, sondern ist Sache der Gouverneure. Es begründete die Schulöffnung damit, dass dies Deutschland, Dänemark, Norwegen und Schweden auch gemacht haben.

Anderson Cooper von CNN wollte zeigen, dass dieser Vergleich hinkt, was er auch tut. Seine Argumentation war aber ziemlich unsinnig. Er bezieht sich auf die Graphik in Abbildung 1.1.

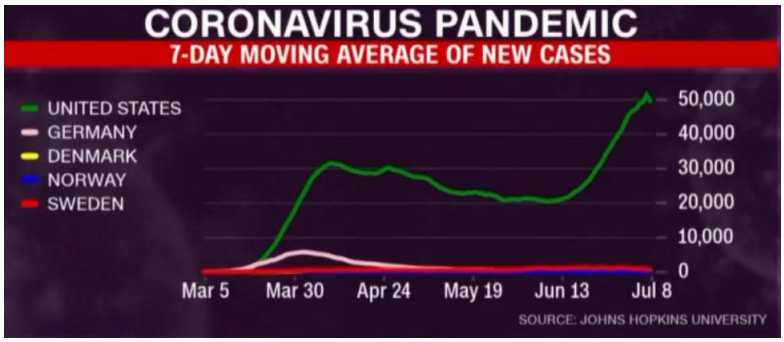


Abbildung 1.1. : Absolute Covid Fallzahlen

Cooper's Begründung war, dass die erwähnten europäischen Fallzahlen viel geringer sind als in den USA und somit es Sinn macht, dass die europäischen Länder die Schulen wieder öffnen, aber in den USA überhaupt nicht, wo die Fallzahlen *sehr* sind.

So plausibel das tönt, die Argumentation ist unsinnig, weil Cooper hier Äpfel mit Birnen vergleicht. Die Zahlen in Abbildung 1.1 sind *absolute* Zahlen. Nun hat aber Norwegen eine Bevölkerung von etwa 5.5 Millionen und die USA 330 Millionen. Notwendigerweise ist die Kurve von Norwegen viel tiefer und flacher als diejenige der USA. Auch wenn die Kurve von Norwegen ähnliche Form wie die Kurve der USA hätte, würde sie nicht viel anders aussehen, als in Abbildung 1.1.

Um fair zu sein, hat Cooper am Schluss erwähnt, dass es einen Unterschied in den Bevölkerungszahlen gibt, aber ging dann auch nicht genauer darauf ein.

Auf den gleichen (nehmen wir zumindest an) Daten beruhend hat die Zeitung *Washington Post* die Graphik in Abbildung 1.2 veröffentlicht.

Hier sind die Fallzahlen pro einer Million Einwohner aufgeführt. Das macht schon mehr Sinn. Diese Kurven können miteinander vergleichen, diejenigen in Abbildung 1.1 nicht.

Jetzt können wir argumentieren, dass Deutschland, Norwegen und Dänemark in Bezug auf die Fallzahlen in einer viel besseren Verfassung sind als die USA, aber Schweden ist es nicht. Die letztere Beobachtung ist in Abbildung 1.1 überhaupt nicht offensichtlich.

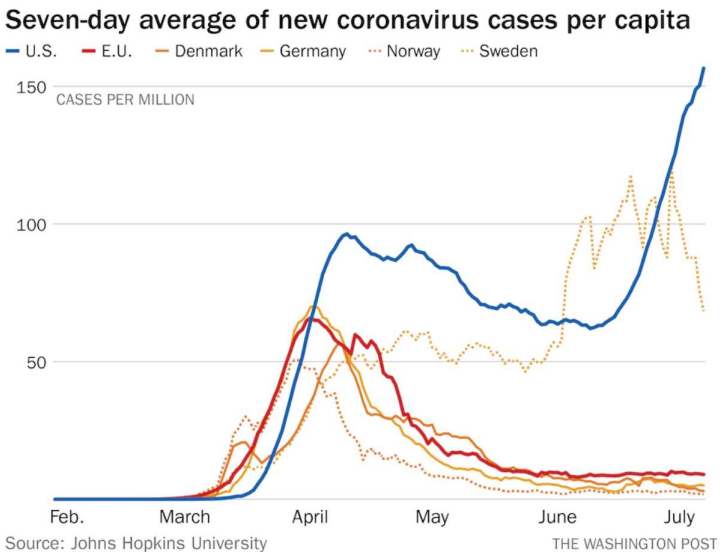


Abbildung 1.2. : Relative Covid Fallzahlen

Die Wiedereröffnung von Schulen ist in Deutschland, Norwegen und Dänemark sinnvoll, in den USA und Schweden jedoch vielleicht nicht.

Aber auch bei relativen Zahlen müssen wir aufpassen. In Abbildung 1.3 sind nochmals die absoluten *Todeszahlen* vom 9. September aufgeführt[1].

#	Country, Other	Total Cases	New Cases	Total Deaths	New Deaths	Total Recovered	Active Cases	Serious, Critical	Tot Cases/ 1M pop	Deaths/ 1M pop	Total Tests	Tests/ 1M pop	Population
	World	27,523,031	+42,072	897,625	+1,069	19,624,327	7,001,079	60,296	3,531	115.2			
1	USA	6,486,426	+851	193,586	+52	3,758,629	2,534,211	14,589	19,575	584	88,067,850	265,771	331,367,517
2	Brazil	4,147,794		127,001		3,355,564	665,229	8,318	19,488	597	14,408,116	67,694	212,842,596
3	India	4,284,103	+6,519	72,843	+27	3,324,060	887,200	8,944	3,099	53	50,650,128	36,636	1,382,530,286
4	Mexico	637,509	+3,486	67,781	+223	446,715	123,013	2,836	4,935	525	1,435,703	11,114	129,184,522
5	UK	350,100		41,554		N/A	N/A	69	5,152	612	17,619,897	259,295	67,953,144
6	Italy	278,784		35,553		210,238	32,993	142	4,612	588	9,271,810	153,393	60,444,825
7	France	328,980		30,726		87,836	210,418	537	5,038	471	8,500,000	130,167	65,300,897
8	Peru	691,575		29,976		522,251	139,348	1,488	20,921	907	3,386,625	102,450	33,056,487
9	Spain	525,549		29,516		N/A	N/A	1,034	11,240	631	9,987,326	213,595	46,758,226
10	Iran	391,112	+2,302	22,542	+132	337,414	31,156	3,713	4,645	268	3,431,646	40,760	84,191,615
11	Colombia	671,848		21,615		529,279	120,954	863	13,178	424	2,964,722	58,150	50,983,652
12	Russia	1,035,789	+5,099	17,993	+122	850,049	167,747	2,300	7,097	123	38,758,184	265,565	145,946,376
13	South Africa	639,362		15,004		566,555	57,803	539	10,755	252	3,808,949	64,073	59,446,940
14	Chile	424,274		11,652		395,717	16,905	930	22,159	609	2,641,589	137,965	19,146,844
15	Ecuador	110,092		10,576		91,242	8,274	424	6,223	598	330,998	18,709	17,692,261
16	Argentina	488,007		10,179	+50	366,590	111,238	2,698	10,779	225	1,412,149	31,192	45,273,109
17	Belgium	88,769	+402	9,909	+2	18,576	60,284	52	7,653	854	2,449,055	211,141	11,599,139
18	Germany	254,168	+543	9,407	+2	227,000	17,761	223	3,032	112	12,383,035	147,708	83,834,622
19	Canada	132,142		9,146		116,459	6,537	54	3,495	242	5,841,880	154,531	37,803,923
20	Indonesia	200,035	+3,046	8,230	+100	142,958	48,847		730	30	2,484,807	9,067	274,061,093

Abbildung 1.3. : Absolute Covid Fallzahlen

[1] Aus https://www.worldometers.info/coronavirus/

Die Länder in der ersten Spalte werden nach den absoluten Zahlen der 4. Spalte von rechts geordnet und wir sehen das, was wir aus den Nachrichten hören.

In Abbildung 1.4 sind die Daten nach den *relativen* Zahlen in der zweitletzten Spalte geordnet.

Die beiden Ranglisten sehen komplett anders aus. Es stand wohl noch nie in den Nachrichten, dass San Marino das Land ist, dass am schlimmsten von Covid 19 betroffen ist.

#	Country, Other	Total Cases	New Cases	Total Deaths	New Deaths	Total Recovered	Active Cases	Serious, Critical	Tot Cases/ 1M pop	Deaths/ 1M pop	Total Tests	Tests/ 1M pop	Population
1	San Marino	716		42		660	14	1	21,093	1,237	6,865	202,239	33,945
2	Peru	691,575		29,976		522,251	139,348	1,488	20,921	907	3,386,625	102,450	33,056,487
3	Belgium	88,769	+402	9,909	+2	18,576	60,284	52	7,653	854	2,449,055	211,141	11,599,139
4	Andorra	1,261		53		934	274	3	16,316	686	137,457	1,778,504	77,288
5	Spain	525,549		29,516		N/A	N/A	1,034	11,240	631	9,987,326	213,595	46,758,226
6	UK	350,100		41,554		N/A	N/A	69	5,152	612	17,619,897	259,295	67,953,144
7	Chile	424,274		11,652		395,717	16,905	930	22,159	609	2,641,589	137,965	19,146,844
8	Bolivia	121,604	+835	7,054	+46	73,150	41,400	71	10,391	603	253,647	21,675	11,702,383
9	Ecuador	110,092		10,576		91,242	8,274	424	6,223	598	330,998	18,709	17,692,261
10	Brazil	4,147,794		127,001		3,355,564	665,229	8,318	19,488	597	14,408,116	67,694	212,842,596
11	Italy	278,784		35,553		210,238	32,993	142	4,612	588	9,271,810	153,393	60,444,825
12	USA	6,486,426	+851	193,586	+52	3,758,629	2,534,211	14,589	19,575	584	88,067,850	265,771	331,367,517
13	Sweden	85,707		5,838	+4	N/A	N/A	13	8,477	577	1,124,269	111,192	10,111,092
14	Mexico	637,509	+3,486	67,781	+223	446,715	123,013	2,836	4,935	525	1,435,703	11,114	129,184,522
15	Panama	97,578		2,099		70,247	25,232	149	22,550	485	369,420	85,371	4,327,225
16	France	328,980		30,726		87,836	210,418	537	5,038	471	8,500,000	130,167	65,300,897
17	Sint Maarten	516		19		321	176	7	12,009	442	2,450	57,022	42,966
18	Colombia	671,848		21,615		529,279	120,954	863	13,178	424	2,964,722	58,150	50,983,652
19	Netherlands	76,548	+964	6,244	+1	N/A	N/A	45	4,466	364	1,648,103	96,144	17,142,061
20	Ireland	29,774		1,777		23,364	4,633	7	6,017	359	906,432	183,191	4,948,020

Abbildung 1.4. : Relative Covid Fallzahlen

Wie kann das sein? Sehen wir uns die letzte Spalte in Abbildung 1.4 an, so stellen wir fest, dass es sich bei den aufgeführten Länder um sehr bevölkerungsarme Länder handelt und da haben ein paar wenige absolute Fälle einen grossen Einfluss auf die relativen Fallzahlen.

In 1.3 sehen wir anhand der letzten Spalte, dass es sich hier grosse oder sehr grosse Länder bezüglich der Bevölkerung handelt. ◀

Die eben gemachte Beobachtung tritt bei angewandten Problemstellungen sehr oft auf: Klärten die relativen Fallzahlen (Abbildung 1.2) die Argumentation von CNN, so heisst das noch lange nicht, dass relative Fallzahlen *immer* besser oder aussagekräftiger sind (siehe Abbildung 1.4) als absolute.

Um es kurz zu machen: Bei angewandten Problemstellungen gibt es kein Kochrezept um das Problem zu lösen.

1.2. Problemlösung in applied statistics

Obwohl das Beispiel 1.1.1 oben sehr einfach ist, enthält es viele Aspekte, die für die Lösung von Problemen in applied statistics relevant sind.

1. Zuerst ist nicht klar, was eigentlich die Fragestellung war.

   Wir begannen mit dem Trump-Tweet. Und jetzt? CNN, gewöhnlich Anti-Trump, dachte wohl, man muss da was dagegen sagen.

2. Es ist nicht klar, wie der Lösungsweg aussieht.

   Wie soll man auf so einen Tweet reagieren? CNN entschied sich, mit Fallzahlen zu argumentieren.

3. Es ist nicht klar, welche Elemente der Lösungsweg enthalten.

   CNN entschied sich, die absoluten Fallzahlen zu verwenden, obwohl die relativen aussagekräftiger sind und auch eindrücklicher sind. Da kann man argumentieren, dass die europäischen Länder im April ähnliches Verhalten der relativen Fallzahlen hatte, aber dann sinkende Fallzahlen hatte (ausser Schweden).

4. Es ist nicht klar, wie man das Resultat interpretieren soll.

   Dies ist oft der schwierigste Punkte in der Lösung von Problemen in applied statistics. Und hier ging CNN in die falsche Richtung. Es wurden Dinge verglichen, die man so nicht miteinander vergleichen kann.

Bemerkung:

Die Graphik in Abbildung 1.1 ist korrekt produziert, aber die Interpretation ist falsch.

Oft sind Graphiken falsch konstruiert, die etwas suggerieren wollen, was nicht vorhanden ist[2]. ⧫

Aufgaben, der folgenden Art, die in der Schulmathematik oft vorkommen, gibt es in applied statistics nicht.

[2]Siehe zum Beispiel Daniel J. Levitin: *A Field Guide to Lies*. Das Buch ist *sehr* zu empfehlen.

Beispiel 1.2.1

Lösen die Sie die Gleichung

$$2x + 1 = 5$$

nach x auf.

Die Problemstellung ist klar, der Lösungsweg ist klar (oft geübt) und an der Lösung $x = 2$ gibt es nichts zu diskutieren. ◀

Sie[3] stossen täglich auf Probleme:

- Sie haben es eilig, irgendwo hinzukommen, können aber ihre Autoschlüssel nicht finden.
- Ihnen geht beim Backen eines Geburtstagskuchens das Mehl aus und der Supermarkt ist geschlossen.
- Ihr Flug wird auf dem Weg zu einem Vorstellungsgespräch gestrichen.
- Sie wollen diese tollen neuen Schuhe kaufen, aber es ist kein Geld auf der Bank.

Für alle diese Probleme gibt es keine Lösungsanweisung. Sie können auch nicht formelmässig gelöst werden.

In diesem Abschnitt entwickeln wir eine vierstufige Problemlösungsstrategie, die Ihnen bei der Lösung einer Vielzahl von Problemen helfen. Die meisten statistischen Probleme werden in Worten formuliert.

Vierstufige Problemlösungsstrategie

1. Erste Schritte

Es ist nicht klar, welcher Weg am effizientesten zur Antwort auf ein gegebenes Problem ist.

Der erste Schritt unserer Problemlösungsstrategie, der Beginn, ist der schwierigste.

Es ist daher sinnvoll, mit etwas zu beginnen, das Sie tun:

- Organisieren Sie die gegebenen Informationen und stellen Sie sicher, dass Ihnen klar ist, was genau in der Problemstellung gefragt ist.
- Machen Sie sich klar, welche Informationen in der Problemstellung enthalten sind.

[3] Aus: Eric Mazur; Principles & Practice of Physics

- Formulieren Sie das Problem mit ihren eigenen Worten.
- Schliesslich müssen Sie feststellen, ob Sie alle Informationen haben oder nicht. die zur Lösung des Problems notwendig sind.

2. **Plan ausarbeiten**

   Als Nächstes müssen Sie einen Plan zur Lösung Ihres Problems ausarbeiten, d.h. herausfinden, was Sie tun müssen, um das Problem zu lösen.

   Ein guter Plan ist es, die Schritte zu umreissen, die Sie unternehmen müssen, um um eine Lösung zu erhalten.

3. **Plan ausführen**

   Im dritten Schritt führen Sie Ihren Plan aus, indem Sie den von Ihnen umrissenen Schritten folgen.

4. **Resultat interpretieren**

   Sie mögen denken, dass Sie fertig sind, aber es gibt noch einen letzten – und sehr wichtigen – Schritt: Interpretieren Sie Ihre Antwort.

   - Überprüfen Sie, ob ihr Resultat überhaupt möglich ist.

     Wenn beispielsweise eine Wahrscheinlichkeit negativ wird, so ist sicher etwas falsch gelaufen.

   - Interpretieren Sie das Resultat in den Worten der Problemstellung.

In der Schulmathematik ist der 3. Punkt der wichtigste, in applied statistics eher weniger. Dies liegt daran, dass die meisten Berechnungen mit **R** durchgeführt werden.

Umso mehr ist der 4. Punkt der wichtigste. Um dies kurz zu illustieren, schauen wir uns folgenden **R**-Output an. Was der **R**-Befehl macht, ist jetzt nicht wichtig:

```
x <- c(1, 4, 6, 8)

t.test(x)


##
##  One Sample t-test
##
## data:  x
## t = 3.1814, df = 3, p-value = 0.05004
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.00151774  9.50151774
```

```
## sample estimates:
## mean of x
##      4.75
```

Dieser Output *muss* interpretiert werden.

Schlussendlich aber, müssen alle 4 Punkte korrekt durchgeführt werden. Auch wenn der 4. Punkt der wichtigste ist, nützt es nichts diesen korrekt zu machen, wenn bei der ersten 3 Schritten ein Fehler gemacht wurde.

Natürlich scheint die Anwendung dieser vier Punkte für sehr einfache Aufgaben überflüssig, aber im weiteren Verlauf dieses Modules, wenn die Probleme komplexer werden, bilden sie sehr gute Anhaltspunkte für die Lösung dieser Probleme. Wir wissen aus Erfahrung, dass es für die Studierenden schwierig ist aus der Problemstellung herauszulesen, was eigentlich gefragt ist.

1.3. Was ist Statistik, die nicht applied ist?

In applied statistics werden Verfahren und Methoden angewendet und beschrieben. Oft machen diese auch Sinn und kann diese auch einfach erklären. Was aber nicht gemacht wird, *warum* diese Verfahren und Methoden *genau* das machen, was sie machen sollen. Und das Problem ist, dass zwar das Prinzip einfach ist, aber die Detail schwierig. Diese Details werden dann *mathematischer Statistik* bewiesen und das sieht dann wie in Abbildung 1.5 aus.

$$S^2 := \frac{1}{m+n-2}\left\{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{j=1}^{m}(Y_j - \bar{Y})^2\right\}.$$

Note that $\bar{X}$ has expectation μ and variance σ^2/n, and $\bar{Y}$ has expectation $\mu+\gamma$ and variance σ^2/m. So $\bar{Y} - \bar{X}$ has expectation γ and variance

$$\frac{\sigma^2}{n} + \frac{\sigma^2}{m} = \sigma^2\left(\frac{n+m}{nm}\right).$$

The normality assumption implies that

$$\bar{Y} - \bar{X} \text{ is } \mathcal{N}\left(\gamma, \sigma^2\left(\frac{n+m}{nm}\right)\right)\text{–distributed.}$$

Hence

$$\sqrt{\frac{nm}{n+m}}\left(\frac{\bar{Y} - \bar{X} - \gamma}{\sigma}\right) \text{ is } \mathcal{N}(0,1)\text{–distributed.}$$

To arrive at a pivot, we now plug in the estimate S for the unknown σ:

$$Z(\mathbf{X}, \mathbf{Y}, \gamma) := \sqrt{\frac{nm}{n+m}}\left(\frac{\bar{Y} - \bar{X} - \gamma}{S}\right).$$

Abbildung 1.5. : Mathematische Statistik

Kapitel 2.

Deskriptive Statistik – Eindimensionale Daten

2.1. Einführung

2.1.1. Datensätze

In diesem Modul werden wir sehr viel mit *Datensätzen* zu tun haben. Aber was sind Datensätze? Datensätze sind Zusammenstellungen von Daten, die in vielen Formen vorkommen können. Die beiden folgenden Beispiele zeigen die wichtigsten Formen von Datensätzen dieses Moduls.

Beispiel 2.1.1

Eine *Liste* von Daten ist die einfachste Variante eines Datensatzes. So enthält beispielsweise die Liste

$$1.75, \quad 1.80, \quad 1.72, \quad 1.65, \quad 1.54$$

die Körpergrössen (in Meter) von 5 Personen.

Solche Listen heissen auch *eindimensionale Datensätze* oder *Messreihen*. ◀

Beispiel 2.1.2

Die häufigste Form von Datensätzen sind *Tabellen* oder *zweidimensionale Datensätze*. Ein einfaches Beispiel sehen wir in Tabelle 2.1.

Die Einträge in den Spalten Grösse und Gewicht sind sogenannte *quantitative* Daten, also Meeswerte (Zahlen). Diese können, zumindest theoretisch, jeden beliebigen Zahlwert annehmen.

Person	Grösse	Gewicht	Geschlecht	Nationalität
A	1.82	72	m	CH
B	1.75	82	w	D
C	1.61	70	w	CH
D	1.80	83	m	A
E	1.83	95	w	FL

Tabelle 2.1. : Beispiel für einen zweidimensionalen Datensatz

Die Einträge in den Spalten Geschlecht und Nationalität sind sogenannte *qualitative* Daten. Diese können nur eine bestimmte Anzahl Werte annehmen und müssen, wie in dem Beispiel hier, auch keine Zahlen sein.

Wir werden auf diese Unterschiede später noch genauer eingehen (siehe Kapitel 8).

◀

Wir werden in diesem Modul immer davon ausgehen, dass ein Datensatz in elektronischer Form als File vorhanden ist. Wie man Daten sammelt liegt ausserhalb dieses Moduls und ist Thema in *Design of Experiment* (DoE).

Die Website

www.kaggle.com

enthält unter anderem eine riesige Sammlung von Datensätzen.

2.1.2. Deskriptive Statistik

Die deskriptive (lat. describere „beschreiben") Statistik befasst sich mit der *Darstellung* von Datensätzen. Dabei werden diese Datensätze durch gewisse Zahlen charakterisiert (zum Beispiel dem Mittelwert) und/oder graphisch dargestellt.

Beispiel 2.1.3

Im Beispiel 2.1.1 können wir uns fragen, wie gross die durchschnittliche Grösse der fünf Personen ist. ◀

Wir befassen uns zunächst mit *eindimensionalen* Daten, wo *eine* Messgrösse an einem Untersuchungsobjekt ermittelt wird. Mit Hilfe des folgenden Beispiels werden wir die wichtigen Begriffe und Vorgehensweisen der deskriptiven Statistik genauer kennenlernen.

Beispiel 2.1.4 Messungen Körpergewicht

Eine Erfahrung, die wir wohl alle schon gemacht haben: Wir stehen am Morgen auf die Waage, merken uns das Gewicht[1], stehen nochmals auf die Waage und wir bekommen oft ein leicht anderes Gewicht.

Wir gehen in diesem Einführungsbeispiel von der folgenden Situation aus:

- Wir nehmen einen 80 Kilogramm schweren Metallblock, der geeicht ist, das heisst, er hat mit sehr grosser Genauigkeit ein Gewicht von 80 kg.
- Das Gewicht dieses Metallblocks wird mehrere Male mit zwei Waagen, Waage A und Waage B, gemessen.
- Dadurch erhalten wir zwei *Datensätze* mit Gewichten (in kg), die in Tabelle 2.3 aufgeführt sind.
- Die Daten sind auf 10 g genau angegeben.

Waage A	79.98	80.04	80.02	80.04	80.03	80.03	80.04	79.97	80.05
Waage A	80.03	80.02	80.00	80.02					
Waage B	80.02	79.94	79.98	79.97	79.97	80.03	79.95	79.97	

Tabelle 2.2. : Gewichtsmessungen eines Metallblocks mit zwei verschiedenen Waagen

Waage A	79.98	80.04	80.02	80.04	80.03	80.03	80.04	79.97	80.05
Waage A	80.03	80.02	80.00	80.02					
Waage B	80.02	79.94	79.98	79.97	79.97	80.03	79.95	79.97	

Tabelle 2.3. : Gewichtsmessungen eines Metallblocks mit zwei verschiedenen Waagen

◀

Zunächst stellt sich die Frage, warum Messungen, die am gleichen Objekt stattfinden, unterschiedliche Resultate ergeben. Messungen finden nie unter *exakt* denselben Bedingungen statt. Scheinbar genaue Angaben, wie die Kalorienzahl auf einer Packung Schokolade oder der Inhalt 500 ml einer Pet-Flasche sind nur *ungefähre* Angaben. Keine zwei Pet-Flaschen sind *absolut* gleich.

[1] Wir verwenden im folgenden Beispiel die umgangssprachliche Verwendung des Begriffes „Gewicht" und nicht die physikalische.

Beispiel 2.1.5

Zurück zu unserem Beispiel 2.1.4 der beiden Waagen: Obwohl die Messungen mit der grösstmöglichen Sorgfalt durchgeführt und alle Störeinflüsse so weit wie möglich ausgeschaltet wurden[2], variieren die Messwerte innerhalb beider Waagen. Es stellen sich hier nun die folgenden Fragen:

- Gibt es einen Unterschied zwischen den gemessenen Werten der Waage A und der Waage B?
- Falls ja, wie können wir diesen Unterschied bestimmen und darstellen?

Betrachten wir die Tabelle ein bisschen genauer, so stellen wir Folgendes fest:

- Bei beiden Waagen liegen die Messwerte in der Nähe von 80, was ja auch so sein sollte.
- Bei der Waage A liegen nur 2 Werte von 13 *unter* 80.
- Bei der Waage B liegen nur 2 von 8 Werten *über* 80 liegen.
- Die Werte der Waage A sind also eher grösser als die der Waage B.

Was heisst hier aber „eher"? Wie können wir die beiden Messreihen konkret miteinander vergleichen?

Eine Möglichkeit ist, die Messreihen irgendwie so *zusammenzufassen*, dass wir die beiden Waagen miteinander vergleichen können. Zusammenfassen heisst hier, dass wir nicht die ganze Messreihe betrachten, sondern dass wir *eine* Zahl berechnen, die die Messreihe beschreibt.

Die erste, allen bekannte Idee ist, dass wir für beide Messreihen den Durchschnitt berechnen und die Werte miteinander vergleichen. Damit sind wir bei der deskriptiven Statistik angelangt. ◀

Die *deskriptive Statistik* beschäftigt sich damit, wie (quantitative) Daten organisiert und zusammengefasst werden können. Die Interpretation und darauffolgende statistische Analyse dieser Daten soll damit vereinfacht werden. Wir machen dies mit Hilfe von

- graphischen Darstellungen, wie Histogramme und Boxplots (siehe Unterkapitel 2.3)
- *Kennzahlen*, die Daten numerisch zusammenfassen, wie Durchschnitt und Standardabweichung (siehe Unterkapitel 2.2)

[2] Wie dies gemacht wird, siehe DoE.

Bemerkungen:

i. Daten sollten immer mit Hilfe von geeigneten graphischen Mitteln *und* den Kennzahlen dargestellt werden, da man nur auf diese Weise (teils unerwartete) Strukturen und Besonderheiten entdecken kann. Wir werden noch einige Male sehen, dass die ausschliessliche Betrachtungen von Kennzahlen problematisch ist. Dasselbe ist der Fall, wenn wir nur graphische Darstellungen betrachten.

ii. Folgendes müssen wir uns immer bewusst sein: *Werden Daten zusammengefasst, gehen Informationen verloren.*

Wie dies gemeint ist, sehen wir am folgenden Beispiel.

Beispiel 2.1.6

In einer Schulklasse mit 24 Lernenden gab es an einer Prüfung folgende Noten[3]:

4.2, 2.3, 5.6, 4.5, 4.8, 3.9, 5.9, 2.4, 5.9, 6, 4, 3.7, 5, 5.2, 4.5, 3.6, 5, 6, 2.8, 3.3, 5.5, 4.2, 4.9, 5.1

Der Notendurchschnitt ist 4.51 (siehe Beispiel 2.2.3). Dieser Wert sagt zwar über die Klasse als *Ganzes* etwas aus, aber nichts mehr über die Noten der einzelnen Lernenden.

Kennen wir nur die Zahl 4.51, so haben wir keine Information mehr, wie die einzelnen Lernenden abgeschnitten haben. Wir wissen nicht einmal, wieviele Lernende an der Prüfung teilgenommen haben. ◀

Der Gewinn, die Aussage über das Ganze, hat einen Verlust, die Information über die Teile, zur Folge. ♦

Im Folgenden werden Daten mit

$$x_1,\ x_2,\ \ldots,\ x_n$$

bezeichnet, wobei n der *Umfang* der Messreihe genannt wird.

[3] Im Schweizerischen Bildungssystem werden Leistungen von 1 – 6 bewertet: 1 ist die schlechteste Note (Bewertung), 6 die beste.

Beispiel 2.1.7

Bei der Messreihe der Waage A aus Beispiel 2.1.4 schreiben wir

$$x_1 = 79.98, \quad x_2 = 80.04, \quad \ldots, \quad x_{13} = 80.02$$

Der Umfang der Messreihe ist dabei $n = 13$. ◀

Wir beginnen die deskriptive Statistik mit den Kennzahlen.

2.2. Kennzahlen

Oft ist es sinnvoll, Datensätze durch eine Zahl, also *numerisch* zusammenzufassen und zu beschreiben. Die Datensätze werden dabei auf eine oder mehrere Zahlen reduziert. Dazu verwenden wir zwei Arten[4] von *Kennzahlen*:

- Eine beschreibt beispielsweise die *mittlere Lage* der Messwerte. Damit ist *nicht* notwendigerweise der Durchschnitt gemeint.
- Die *Variabilität* oder *Streuung* dieser Messwerte gibt die „durchschnittliche" Abweichung der Messwerte von der mittleren Lage an.

Wir sprechen im ersten Fall allgemeiner von *Lageparametern* oder *Lagemassen*. Sie beschreiben, *wo* sich die Daten befinden.

Im zweiten Fall sprechen wir von *Streuungsparametern* oder *Streuungsmassen*. Sie beschreiben, *wie* sich die Daten um die mittlere Lage verteilen.

2.2.1. Arithmetisches Mittel

Die bekannteste Grösse für eine mittlere Lage ist der *Durchschnitt* oder das

Arithmetisches Mittel $\overline{x}$

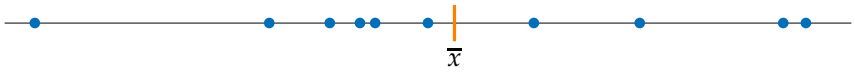
$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

[4] Es gibt noch weitere, aber wir beschränken uns hier auf die beiden wichtigsten.

Bemerkungen:

i. Manchmal verwenden wir die Notation $\overline{x}_n$, wobei n wieder den Umfang der Messreihe bezeichnet.

ii. Graphische Darstellung das arithmetischen Mittelwertes



iii. Es gibt noch weitere Mittel, wie das geometrische oder das harmonische Mittel, aber wir beschränken uns hier auf das arthmetische Mittel. ♦

Beispiel 2.2.1 Messungen mit Waage A

Das arithmetische Mittel der $n = 13$ Messungen ist

$$\overline{x}_{13} = \frac{79.98 + 80.04 + \cdots + 80.03 + 80.02 + 80.00 + 80.02}{13} = 80.02$$

Wir summieren alle Werte auf und dividieren die Summe durch die Anzahl der Werte.

Mit **R** berechnen wir den Mittelwert mit dem Befehl **mean(...)**:

```
waageA <- c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04, 79.97,
    80.05, 80.03, 80.02, 80, 80.02)

mean(waageA)


## [1] 80.02077
```

Der Durchschnitt der 13 Gewichte von Waage A ist 80.02 kg. ◀

Beispiel 2.2.2

Wir können also die Messwerte der beiden Waagen mit Hilfe des arithmetischen Mittels miteinander vergleichen.

```
waageB <- c(80.02, 79.94, 79.98, 79.97, 79.97, 80.03, 79.95, 79.97)

mean(waageB)


## [1] 79.97875
```

Waage B hat also durchschnittlich die tieferen Werte als die Waage A. ◀

Beispiel 2.2.3

Im Notenbeispiel 2.1.6 der Noten lautet der Durchschnitt:

```
noten <- c(4.2, 2.3, 5.6, 4.5, 4.8, 3.9, 5.9, 2.4, 5.9, 6, 4, 3.7,
    5, 5.2, 4.5, 3.6, 5, 6, 2.8, 3.3, 5.5, 4.2, 4.9, 5.1)

mean(noten)


## [1] 4.5125
```

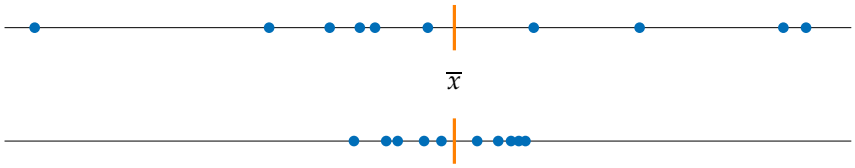
Der Notendurchschnitt ist also 4.51, auf zwei Nachkommastellen gerundet. ◀

2.2.2. Empirische Varianz und Standardabweichung

Obwohl das arithmetische Mittel schon einiges über einen Datensatz aussagt, beschreibt er diesen aber nur unvollständig.

Beispiel 2.2.4

Zunächst ein graphisches Beispiel:



Beide Datenreihen haben dasselbe arithmetische Mittel, aber die Punkte der zweiten Datenreihe liegen durchschnittlich viel näher beim Mittelwert $\overline{x}$ als die Punkte der ersten. Wir sprechen von unterschiedlicher *Streuung* der Daten um den Durchschnitt.

Bei der ersten Messreihe sprechen wir von einer „grossen" Streuung, bei der zweiten von einer „kleinen" Streuung. Was konkret „gross" und „klein" hier heisst, lässt sich nicht sagen. Was wir aber sagen können, dass die Streuung der ersten Messreihe *grösser* ist als diejenige der zweiten. ◀

Beispiel 2.2.5

Wir betrachten die folgenden beiden Datensätze von (fiktiven) Schulnoten von zwei Schulklassen:

$$2,\ 6,\ 3,\ 5 \quad \text{und} \quad 4,\ 4,\ 4,\ 4$$

Beide haben denselben Mittelwert 4, aber die Verteilung der Daten um den Mittelwert ist sehr unterschiedlich. In der ersten Klasse gibt es zwei gute und zwei schlechte Lernende und in der zweiten sind alle Lernenden gleich gut. Wir sagen, die Datensätze haben eine unterschiedliche *Streuung* um die Mittelwerte. Die erste Klasse hat die *grössere* Streuung als die zweite.

Wie können wir diesen Unterschied numerisch, also durch eine Zahl für jeden Datensatz ausdrücken?

Da wir uns für die *Unterschiede der Daten zum Durchschnitt* interessieren, wählen wir als ersten Ansatz für die Streuung, dass wir den *Durchschnitt der Unterschiede der Daten zum Mittelwert* betrachten.

Für den ersten Datensatz erhalten wir

$$\frac{(2-4)+(6-4)+(3-4)+(5-4)}{4} = \frac{-2+2-1+1}{4} = \frac{0}{4} = 0$$

Für den zweiten Datensatz ergibt dies ebenfalls 0:

$$\frac{(4-4)+(4-4)+(4-4)+(4-4)}{4} = \frac{0+0+0+0}{4} = \frac{0}{4} = 0$$

Da wir in beiden Klassen dasselbe Resultat erhalten, sagt diese Grösse *nichts* über die Streuung aus. Das Problem ist, dass die Unterschiede zum Mittelwert *negativ* werden können und sich diese wie im obigen Fall aufheben.

Der eben gesehene Ansatz für die Streuung bringt also nichts, aber wir können in leicht verbessern, damit sich die Unterschied nicht mehr aufheben. Dazu ersetzen die Unterschiede durch die Absolutwerte der Unterschied. Damit erhalten wir keine negativen solcher Werte.

Für den ersten Datensatz erhalten wir

$$\frac{\big|(2-4)\big|+\big|(6-4)\big|+\big|(3-4)\big|+\big|(5-4)\big|}{4} = \frac{2+2+1+1}{4} = \frac{6}{4} = 1.5$$

Die Noten weichen nun im Schnitt 1.5 Noten vom Mittelwert ab.

Für den zweiten Datensatz erhalten wir 0:

$$\frac{\big|(4-4)\big| + \big|(4-4)\big| + \big|(4-4)\big| + \big|(4-4)\big|}{4} = \frac{0+0+0+0}{4} = \frac{0}{4} = 0$$

Dies sollte auch so sein, da wir beim zweiten Datensatz keine Streuung um den Mittelwert haben.

Je grösser dieser Wert (der immer grösser gleich 0 ist), desto mehr unterscheiden sich die Daten bei gleichem Mittelwert voneinander. ◀

Diese eben berechnete Grösse für die Streuung heisst *mittlere absolute Abweichung*. Diese ist zwar von der Überlegung her recht einfach, hat aber schwerwiegende theoretische und rechnerische Nachteile, auf die wir hier nicht eingehen können.

Die gebräuchlichsten Masse für die Streuung oder Variabilität von Messwerten sind die (scheinbar kompliziertere) *empirische Varianz* und *empirische Standardabweichung*. Diese werden definiert durch

Empirische Varianz $\mathbf{Var}\,(x)$

$$\mathrm{Var}(x) = \frac{(x_1-\overline{x}_n)^2 + (x_2-\overline{x}_n)^2 + \cdots + (x_n-\overline{x}_n)^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x}_n)^2$$

Empirische Standardabweichung s_x

$$s_x = \sqrt{\mathrm{Var}(x)}$$

Bemerkungen:

i. Bei der Varianz quadrieren wir die Abweichungen vom Mittelwert $x_i - \overline{x}_n$, damit sich diese nicht gegenseitig aufheben können.

ii. Der Nenner $n-1$, anstelle von n, ist mathematisch begründet, worauf wir hier nicht eingehen können.

In einigen Büchern steht bei der Definition für die Varianz im Nenner n, anstatt $n-1$. Das mag bei kleinen Datensätzen eine Rolle spielen, aber für grosse Datensätze ist der Unterschied vernachlässigbar, wie wir im folgenden Beispiel sehen.

Beispiel 2.2.6

Wir wählen einen Datensatz mit drei Zahlen 1, 7 und 10 und bestimmen die Varianz einmal mit Division durch $n = 3$ und einmal mit Division $n - 1 = 2$:

```
# Definition des Vektors x
x <- c(1, 7, 10)

# Bestimmung des arithmetischen Mittels von x
mean_x <- mean(x)
mean_x

## [1] 6

# Bestimmung der Varianz von x mit n dividiert
var_x_n <- sum((x - mean_x)^2)/length(x)
var_x_n

## [1] 14

# Bestimmung der Varianz von x mit n-1 dividiert
var_x_n_1 <- sum((x - mean_x)^2)/(length(x) - 1)
var_x_n_1

## [1] 21
```

Hier ist der Unterschied doch beträchtlich zwischen 14 und 21, je nachdem, ob wir durch 3 oder durch 2 dividieren.

Nun wählen wir einen Vektor **y** der Länge 999, der aus 333 Kopien von **x** besteht. Das heisst, diesselben Daten werden wiederholt, der Mittelwert von **y** bleibt derselbe wie bei **x** und auch die Varianz bleibt dieselbe bei Division durch $n = 999$.

Der Vektor **y** mit dem Befehl **rep(...)** erzeugt. Der Vektor **x** wird **times = 333** wiederholt.

```
# Definition des Vektors y
y <- rep(x, times = 333)

# Bestimmung des arithmetischen Mittels von y
mean_y <- mean(y)
mean_y

## [1] 6

# Bestimmung der Varianz von y mit n dividiert
var_y_n <- sum((y - mean_y)^2)/length(y)
var_y_n

## [1] 14
```

```
# Bestimmung der Varianz von y mit n-1 dividiert
var_y_n_1 <- sum((y - mean_y)^2)/(length(y) - 1)
var_y_n_1

## [1] 14.01403
```

Das arithmetische Mittel von **y** ist dasselbe wie dasjenige von **x**. Der Unterschied bei der Varianz ist nun allerdings minim, 14 verglichen zu 14.014, je nachdem ob wir durch 999 oder 998 dividieren. ◀

In der Stochastik kommt es noch recht oft vor, dass die Definitionen von Konzepten nicht einheitlich sind. Dies unterschiedlichen Definitionen machen für kleine Datensätze zwar einen Unterschied aus, aber für grosse Datensätze sind Unterschiede vernachlässigbar.

R dividiert beim Befehl **var(...)** durch $n - 1$ (siehe Beispiel 2.2.7).

iii. Die Standardabweichung ist die Wurzel der Varianz. Da wir für die Berechnung der Varianz die Quadrate der Abstände zum Mittelwert verwendet haben, bekommen wir durch das Wurzelziehen wieder dieselbe Einheit wie bei den Daten selbst.

Sind die Daten beispielsweise in cm angegeben, so ist die Varianz wegen der Quadrierung der Daten ein Wert in cm^2. Damit wir wieder die Einheit der ursprünglichen Daten erhalten, müssen wir aus der Varianz die Wurzel ziehen und erhalten damit die Standardabweichung.

iv. Wichtig: *Nur die Standardabweichung s_x lässt sich konkret interpretieren.*

Der Wert der empirischen Varianz hat *keine* physikalische Bedeutung. Wir wissen nur, je grösser der Wert ist, umso grösser ist die Streuung.

v. Nach der Bemerkung iv kann man sich fragen, warum dann die Varianz überhaupt definiert wird, wenn sie schon keine praktische Bedeutung hat.

Der Grund dafür ist, dass die Varianz einfacher zu berechnen ist. Es kommen nur einfache Operationen, wie Addition, Quadrate und Division vor. So wird zuerst immer mit der Varianz gerechnet und erst ganz am Schluss wird die Wurzel gezogen.

Es mag nicht offensichtlich sein, aber Ausdrücke mit Wurzeln, die Summen enthalten sind schwierig umzuformen und man versucht diese so weit möglich zu vermeiden. Dieses Problem haben wir mit der Varianz nicht. Kurz: Quadrate hat man im Griff, Wurzeln aber nicht. □

vi. Für normalverteilte Daten hat die Standardabweichung noch eine schöne geometrische Interpretation (siehe Kapitel 5). ♦

Beispiel 2.2.7 Messungen des Metallblocks mit Waage A

Das arithmetische Mittel der $n = 13$ Messungen der Waage A ist $\overline{x}_{13} = 80.02$ (siehe Beispiel 2.2.1). Die empirische Varianz ist

$$\begin{aligned}\mathrm{Var}(x) &= \frac{(79.98 - 80.02)^2 + (80.04 - 80.02)^2 + \cdots + (80.00 - 80.02)^2 + (80.02 - 80.02)^2}{13 - 1} \\ &= 0.000\,574\,4\end{aligned}$$

und die empirische Standardabweichung:

$$s_x = \sqrt{0.000\,574} = 0.023\,97$$

Somit ist die „mittlere" Abweichung vom Mittelwert 0.023 97 kg.

Für Waage B finden wir $\overline{x}_8 = 79.98$ und $s_x = 0.03137$ mit der analogen Interpretation.

Die empirische Varianz bzw. Standardabweichung (englisch: standard deviation) ist von Hand mühsam auszurechnen, deswegen benützen wir **R**:

```
var(waageA)

## [1] 0.000574359

sd(waageA)

## [1] 0.02396579
```

◀

Beispiel 2.2.8

Für das Notenbeispiel 2.1.6 erhalten wir für die Standardabweichung

```
sd(noten)

## [1] 1.104265
```

Somit weichen die Noten „im Mittel" um 1.1 Noten vom Mittelwert 4.51 ab (siehe Beispiel 2.2.3). ◀

2.2.3. Median

Ein weiteres Lagemass für die mittlere Lage von quantitativen Daten ist der *Median*. Es handelt sich dabei grob gesagt um denjenigen Wert, bei dem rund die Hälfte der Messwerte kleiner oder gleich und die andere Hälfte grösser oder gleich diesem Wert sind.

Beispiel 2.2.9

Im Notenbeispiel 2.1.6 ist der Median 4.65. Die Hälfte der Klasse hat eine Note 4.65 oder schlechter. Umgekehrt hat die andere Hälfte der Klasse eine 4.65 oder besser. ◀

Um den *Median* zu bestimmen, müssen wir die Daten zuerst der Grösse nach ordnen:

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

Die runden Klammern im Index sollen andeuten, dass die Werte geordnet sind.

Beispiel 2.2.10

Wir ordnen die Werte der Waage A der Grösse nach:

79.97, 79.98, 80.00, 80.02, 80.02, 80.02, 80.03, 80.03, 80.03, 80.04, 80.04, 80.04, 80.05

Der Median von diesen 13 Messungen ist dann der Wert der *mittleren* Beobachtung. Dies ist in diesem Fall der Wert der 7. Beobachtung:

79.97, 79.98, 80.00, 80.02, 80.02, 80.02, 80.03, 80.03, 80.03, 80.04, 80.04, 80.04, 80.05

Der Median des Datensatzes der Waage A lautet somit 80.03. Es sind 6 Beobachtungen kleiner oder gleich 80.03 und 6 Messwerte grösser oder gleich 80.03. ◀

In diesem Beispiel ist die Anzahl der Daten ungerade, und somit gibt es eine mittlere Beobachtung. Ist die Anzahl der Daten gerade, so gibt es zwei gleichwertige mittlere Beobachtungen. Als Median benützen (definieren) wir in diesem Fall den Durchschnitt der beiden mittleren Beobachtungen.

Beispiel 2.2.11

Der Datensatz der Waage *B* hat 8 Beobachtungen. Wir ordnen den Datensatz der Grösse nach und *definieren* als Median den Durchschnitt von der 4. und 5. Beobachtung:

$$79.94,\ 79.95,\ 79.97,\ \underbrace{79.97,\ 79.97}_{\text{Median}},\ 79.94,\ 80.02,\ 80.03$$

Also

$$\frac{79.97 + 79.97}{2} = 79.97$$

Der Median der Daten der Waage *B* ist 79.97. Das heisst, die Hälfte der Messwerte kleiner oder gleich und die andere Hälfte grösser oder gleich diesem Wert sind. Die Werte der beiden mittleren Beobachtungen sind hier zufällig gleich, dies ist aber im Allgemeinen nicht so.

Mit **R** bestimmen wir den Median mit dem Befehl **median(...)**:

```
median(waageA)

## [1] 80.03

median(waageB)

## [1] 79.97
```

◀

Beispiel 2.2.12

Der Median im Notenbeispiel ist 4.65 (siehe Beispiel 2.2.9):

```
median(noten)

## [1] 4.65
```

◀

Bemerkungen:

i. Am Beispiel 2.2.12 sehen wir, dass der Median *kein* Wert aus dem Datensatz sein muss. Die Noten sind auf eine Stelle nach dem Komma angegeben.

ii. Der Median wird auch *Zentralwert* oder *mittlerer Wert* (nicht zu verwechseln mit dem Mittelwert) genannt.

iii. Die exakte Interpretation des Medians ist noch erstaunlich schwierig. Wir belassen es damit, dass die Hälfte der Werte kleiner oder gleich und die andere Hälfte grösser oder gleich dem Median sind. ♦

Wir haben nun zwei Lagemasse für die Mitte eines Datensatzes: das arithmetische Mittel und den Median. Welches sind die Vorzüge der jeweiligen Lagemasse? Oder ist das eine Lagemass „besser" als das andere?

Eine Eigenschaft des Medians ist die *Robustheit*. Der Median wird weniger stark durch extreme Beobachtungen beeinflusst als das arithmetische Mittel.

Beispiel 2.2.13 Messungen mit Waage A

Bei der grössten Beobachtung ($x_9 = 80.05$) aus 2.2.1 ist ein Tippfehler passiert und es wurde fälschlicherweise $x_9^* = 800.5$ für die Berechnung des Durchschnittes verwendet. Das arithmetische Mittel x_{13}^* wird dann zu

$$\overline{x}_{13}^* = 135.44$$

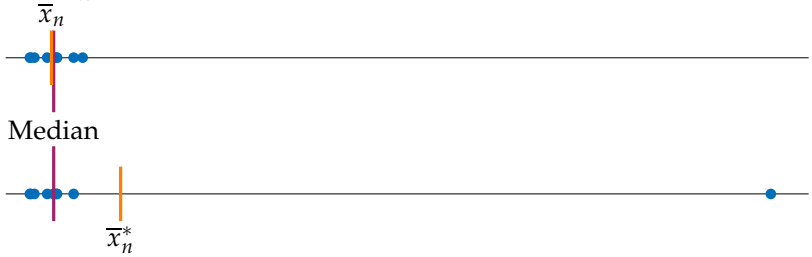
Der Median ist aber nach wie vor

$$x_{(7)} = 80.03$$

Das arithmetische Mittel wird also durch Veränderung *einer* Beobachtung sehr stark beeinflusst, während der Median hier gleich bleibt – er ist *robust*. ◀

Beispiel 2.2.14

Wir können die Robustheit noch graphisch darstellen. Der grösste der blauen Punkte wird weit nach rechts verschoben. Der Median ist verändert sich nicht, der Durchschnitt von x_n zu x_n^* aber schon.



◀

Beispiel 2.2.15

Nehmen wir das typische Haushaltseinkommen von Vororten von Seattle um Lake Washington. Die durchschnittlichen Einkommen von Medina und Windermere werden sehr unterschiedlich sein. Der Grund dafür ist, dass Bill Gates in Medina lebt.

Grundsätzlich wird für das mittlere Einkommen praktisch immer der Median gewählt und nicht der Durchschnitt, da der Median gerechter ist.

Zügelt Bill Gates von Medina nach Windermere, wird sich das Durchschnittseinkommen in beiden Orten mehr oder weniger stark verändern. Die Bewohner von Windermere haben aber nichts davon, dass das Durchschnittseinkommen grösser wurde.

Auf der anderen Seite merken die Bewohner von Medina nichts davon, dass das Durchschnittseinkommen gesunken ist. ◀

Bemerkungen:

i. Grundsätzlich sollte man beide Lagemasse für die mittlere Lage immer gleichmeinsam betrachten. Eine grosse Abweichung zwischen den Werten deutet auf eine besondere Verteilung der Daten hin.

ii. In der Stochastik ist es noch oft so, dass es kein „besser" oder „schlechter" für verschiedene Definitionen von Konzepten (hier die mittlere Lage) gibt. In gewissen Situationen eignen sich die eine Definition oder Methode besser als die andere. Dies hängt immer von der konkreten Problemstellung ab. ♦

2.2.4. Quartile

Die Quartile sind analog dem Median definiert, aber nicht für 50 % der Daten, die grösser oder kleiner gleich sind, sondern für 25 % bzw. 75 % der Daten.

Wir wollen zunächst die Interpretation der Quartile anhand des folgenden Beispiels betrachten. Die Definition und Berechnung der Quartile folgt nach diesem Beispiel.

Beispiel 2.2.16

Beim Notenbeispiel 2.1.6 ist das untere Quartil 3.8. Also sind 25 % der Noten von den Lernenden 3.8 und tiefer und für 75 % der Noten 3.8 und höher.

Das obere Quartil ist 5.35. Also sind 75 % der Noten der Lernenden 5.35 und tiefer und für 25 % der Lernenden 5.35 und höher.

Für die Berechnung dieser Werte siehe Beispiel 2.2.19. ◀

Das *untere Quartil* ist derjenige Wert, bei welchem 25 % aller Beobachtungen kleiner oder gleich und 75 % grösser oder gleich wie diesem Wert sind.

Entsprechend ist das *obere Quartil* derjenige Wert, bei dem 75 % aller Beobachtungen kleiner oder gleich und 25 % grösser oder gleich wie diesem Wert sind.

Allerdings gibt es für die meisten Datensätze nicht *exakt* 25 % der Anzahl Beobachtungen, wie folgendes Beispiel zeigt.

Beispiel 2.2.17

Die Waage A hat $n = 13$ Messpunkte und 25 % dieser Anzahl ist 3.25.

Wir *wählen* (definieren) in diesem Fall den nächstgrösseren Wert $x_{(4)}$ als unteres Quartil:

$$79.97,\ 79.98,\ 80.00,\ \underline{80.02},\ 80.02,\ 80.02,\ 80.03,\ 80.03,\ 80.03,\ 80.04,\ 80.04,\ 80.04,\ 80.05$$

Das untere Quartil ist somit 80.02. Rund ein Viertel der Messwerte sind kleiner oder gleich gross und drei Viertel der Messwerte sind grösser oder gleich gross wie diesem Wert 80.02.

Für das obere Quartil wählen wir $x_{(10)}$, da für $0.75 \cdot 13 = 9.75$ die Zahl 10 der nächsthöhere Wert ist.

$$79.97,\ 79.98,\ 80.00,\ 80.02,\ 80.02,\ 80.02;80.03,\ 80.03,\ 80.03,\ \underline{80.04},\ 80.04,\ 80.04;\ 80.05$$

Rund drei Viertel aller Messwerte sind also kleiner oder gleich 80.04 und ein Viertel grösser oder gleich diesem Wert.

Bei der Waage B sind 25 % von 8 Werten zwei Werte. Dies ist eine ganze Zahl, und wir wählen den Durchschnitt von $x_{(2)}$ und $x_{(3)}$ als unteres Quartil. Dann sind 2 Beobachtungen kleiner und 6 Beobachtungen grösser als dieser Wert.

$$79.94,\ \underline{79.95,\ 79.97},\ 79.97,\ 79.97,\ 79.94,\ 80.02,\ 80.03$$

Also

$$\frac{79.95 + 79.97}{2} = 79.96$$

Das untere Quartil der Waage B ist also 79.96. ◀

Bemerkungen:

i. Wir haben für den Fall, dass die Anzahl Beobachtungen keine ganze Zahl ist, jeweils aufgerundet. Wir hätten auch abrunden können oder den Mittelwert nehmen. Es ist nicht klar, was man in solchen Fällen machen soll.

   Entsprechend gibt es auch verschiedene Definitionen für die Quartile. *Es existiert in der Statistik keine einheitliche Definition für die Quartile*[5].

   Aber: *Für grosse Datensätze spielt es praktisch keine Rolle, ob wir auf-, abrunden oder irgendeine andere Definition wählen. Die Unterschiede für grosse Datensätze sind vernachlässigbar.*

ii. Die Standardeinstellung für die Quartile ist für unterschiedliche Software nicht einheitlich.

iii. Wichtig ist die Interpretation: Beim unteren Quartil sind rund 25 % aller Messwerte kleiner oder gleich diesem Wert und 75 % grösser oder gleich diesem Wert.

♦

Die Software **R** kennt keine eigenen Befehle für die Quartile. Wir können allerdings den allgemeineren Befehl **quantile** (siehe Abschnitt 2.2.6) verwenden. Damit **R** die Quartile nach unserer Definition berechnet, müssen wir die Option **type = 2** hinzufügen. Mit **p = ...** wird die Prozentzahl durch 100 dividiert angegeben. Für das untere Quartil ist dies **p = 0.25**.

```
# Syntax fuer das untere Quartil: p = 0.25

quantile(waageA, p = 0.25, type = 2)


##   25%
## 80.02


quantile(waageB, p = 0.25, type = 2)


##   25%
## 79.96
```

[5] Das heisst, es gibt eine exakte Definition für die Quartile. Diese ist aber recht kompliziert und für praktische Berechnung nicht von Belang.

```
# Syntax fuer das obere Quartil: p = 0.75

quantile(waageA, p = 0.75, type = 2)


##   75%
## 80.04
```

2.2.5. Quartilsdifferenz

Die Quartilsdifferenz ist definiert als die Differenz der beiden Quartile:

$$\text{oberes Quartil} - \text{unteres Quartil}$$

Sie ist ein Streuungsmass für die Daten. Es misst die Länge des Intervalls, das etwa die Hälfte der mittleren Beobachtungen enthält. Je kleiner dieses Mass, umso näher liegt die Hälfte aller Werte beim Median und umso kleiner ist die Streuung. Dieses Streuungsmass ist robust.

Beispiel 2.2.18

Die Quartilsdifferenz bei der Waage A

$$80.04 - 80.02 = 0.02$$

und wird mit dem R-Befehl `IQR(...)` (interquartile range) bestimmt:

```
IQR(waageA, type = 2)


## [1] 0.02
```

Rund die Hälfte aller Messwerte liegt also in einem Bereich der Länge 0.02. ◀

Beispiel 2.2.19

Wir berechnen für das Notenbeispiel 2.2.16 mit R die Quartile und die Quartilsdifferenz:

```
noten <- c(4.2, 2.3, 5.6, 4.5, 4.8, 3.9, 5.9, 2.4, 5.9, 6, 4, 3.7,
    5, 5.2, 4.5, 3.6, 5, 6, 2.8, 3.3, 5.5, 4.2, 4.9, 5.1)

quantile(noten, p = c(0.25, 0.75), type = 2)
```

```
##  25%  75%
## 3.80 5.35

IQR(noten, type = 2)

## [1] 1.55
```

Die mittlere Hälfte der Lernenden liegen innerhalb von 1.55 Noten, nämlich zwischen 3.8 und 5.35. Dementsprechend haben auch rund 25 % der Klasse 3.8 oder weniger und rund 25 % der Klasse 5.35 und mehr. ◀

Wir werden beim Boxplot (siehe Abschnitt 2.3.2) die Quartilsdifferenz noch geometrisch interpretieren.

2.2.6. Quantile

Mit den *Quantilen* können wir das Konzept der Quartile auf jede beliebige Prozentzahl verallgemeinern. So ist das 10 %-Quantile derjenige Wert, wo rund 10 % der Werte kleiner oder gleich und rund 90 % der Werte grösser oder gleich diesem Wert sind.

Das *empirische α-Quantil* ist anschaulich gesprochen derjenige Wert, bei dem $\alpha \times 100\,\%$ der Datenpunkte kleiner oder gleich und $(1-\alpha) \times 100\,\%$ der Punkte grösser oder gleich diesem Wert sind.

Das 10 %- und 70 %-Quantil der Waage A berechnen wir wie folgt:

```
quantile(waageA, p = 0.1, type = 2)

##   10%
## 79.98

quantile(waageA, p = 0.7, type = 2)

##   70%
## 80.04
```

Rund 10 % der Messwerte sind kleiner oder gleich 79.98 und rund 90 % grösser oder gleich diesem Wert . Entsprechend sind rund 70 % der Messwerte kleiner oder gleich 80.04 und 30 % grösser oder gleich diesem Wert.

Beispiel 2.2.20

Wir berechnen nun mit **R** verschiedene Quantile im Beispiel 2.2.19:

```
quantile(noten, p = seq(from = 0.2, to = 1, by = 0.2), type = 2)

##  20%  40%  60%  80% 100%
##  3.6  4.2  5.0  5.6  6.0
```

Rund 20 % der Lernenden haben also eine 3.6 oder waren schlechter und rund 80 % der Lernenden waren¨ gleich oder besser als dieser Wert. Genau 20 % der Lernenden ist nicht möglich, da dies 4.8 Lernenden entsprechen würde.

Das 60 %-Quantil besagt, dass rund 60 Prozent der Lernenden Noten von 5 oder weniger haben. Folglich haben rund 40 % eine 5 oder sind besser. ◀

2.3. Graphische Methoden

Daten graphisch darzustellen ist ein sehr wichtiger Aspekt der statistischen Datenanalyse, da dann oft die Struktur der Daten auf einen Blick ersichtlich ist. Zudem erkennt man häufig Muster, die aus den Kennzahlen nicht erkennbar sind.

Wir werden hier zwei Methoden kennenlernen, um eindimensionale Daten graphisch darzustellen: das Histogramm und der Boxplot.

Bevor wir mit dem Histogramm beginnen, das wohl allen bekannt ist, möchten wir zuerst ein Beispiel einer graphischen Darstellung zeigen, die *nicht* nützlich ist, da sie die Struktur der Daten nicht zeigt.

Beispiel 2.3.1

Wir simulieren den IQ-Test von 200 Personen (siehe Bemerkungen in Beispiel 2.3.2) und zeichnen die Resultate in ein Koordinatensystem und erhalten Abbildung 2.1.

Auf der horizontalen Achse (**Index**) sind die 200 Personen aufgeführt und auf der vertikalen Achse (**iq**) der zugehörige IQ. So entspricht beispielweise der rote Punkt dem IQ-Wert der 87. Person, also etwa 110.

Es ist offensichtlich kein klares Muster erkennbar. Der Grund dafür ist, das die Personen nicht der Grösse nach geordnet sind. ◀

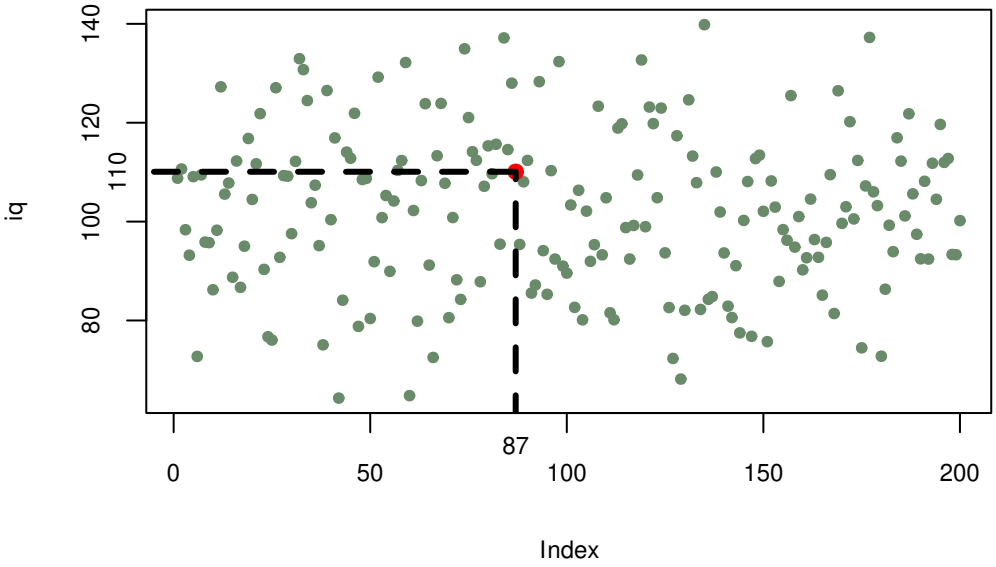


Abbildung 2.1. : IQ-Test Ergebnis von 200 Personen

Diese Art der Darstellung ist also nicht sinnvoll. Es ist nicht jede graphische Darstellung einfach hilfreich. Graphische Darstellungen können mehr verwirrend als hilfreich sein, falls sie ungeeignet erstellt werden.

Wir können dem Abhilfe schaffen, indem wir einen Ausschnitt des Beobachtungsbereiches (vertikale Achse `iq` in Beispiel 2.3.1) in sogenannte *Klassen* einteilen und dann die Anzahl Beobachtungen in dieser Klasse zählen.

Der Beobachtungsbereich liegt nun auf der *horizontalen* Achse.

2.3.1. Histogramm

Einen graphischen Überblick über die auftretenden Werte erhalten wir mit einem sogenannten *Histogramm*. Histogramme helfen uns bei der Frage, in welchem *Wertebereich* besonders viele Datenpunkte liegen. Ist die Datenmenge gross, so macht es keinen Sinn, alle Werte einzeln zu betrachten (siehe Beispiel 2.3.3 links oben.)

Beispiel 2.3.2

In Abbildung 2.2 sehen wir das Histogramm von dem Ergebnis des IQ-Testes in Beispiel 2.3.1.

- Die Breite der Klassen wurde mit 10 IQ-Punkten festgelegt und ist für jede Klasse gleich.

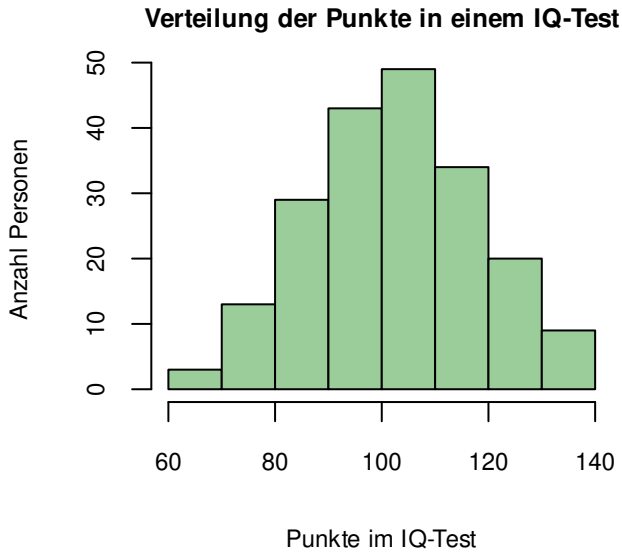


Abbildung 2.2. : Histogramm von dem IQ-Test Ergebnis von 200 Personen.

- Die Höhe der Balken gibt die Anzahl Personen an, die in diese Klasse fallen. Zum Beispiel fallen ca. 20 Personen in die Klasse zwischen 120 und 130 IQ-Punkten.

Die Form des Histogrammes in diesem Beispiel ist typisch für viele Histogramme. Es hat die Form einer sogenannten *Normalverteilung*. Wir werden auf dieses Beispiel nochmals im Kapitel 5 (Normalverteilung) zu sprechen kommen.

Bemerkungen: R-Code

i. Der **R**-Code für das Histogramm oben lautet wie folgt:

```
iq <- rnorm(n = 200, mean = 100, sd = 15)

hist(iq,
     col = "darkseagreen3",
     xlab = "Punkte im IQ-Test",
     ylab = "Anzahl Personen",
     main = "Verteilung der Punkte in einem IQ-Test"

)
```

ii. Der Befehl **`rnorm(n = 200, mean = 100, sd = 15)`** wählt zufällig 200 normalverteilte Daten (siehe Kapitel 5 mit Mittelwert 100 mit Standardabweichung 15 aus (siehe auch Beispiel 5.2.2).

iii. Der Befehl **`hist(iq, ...)`** zeichnet das Histogramm für die Daten **`iq`**.

iv. Die weiteren Optionen sollten klar sein:

- `xlab` steht für x-Label, die Beschriftung der x-Achse
- `ylab` steht für y-Label, die Beschriftung der y-Achse
- `col` steht für color
- `main` steht für Haupttitel

♦

◄

Allgemein wird ein Histogramm schrittweise wie folgt aufgebaut (**R** macht das natürlich automatisch):

- Wir teilen die Werte der Daten in Klassen ein.

  Für die Festlegung der *Anzahl* der Klassen bzw. Rechtecke existieren verschiedene Faustregeln: bei weniger als 50 Messungen ist die Klassenzahl 5 bis 7, bei mehr als 250 Messungen wählt man 10 bis 20 Klassen.

- Im einfachsten Fall wird die gleiche Breite für alle Klassen gewählt, was aber nicht unbedingt der Fall sein muss.

- Dann zeichnen wir für jede Klasse einen Balken, dessen Höhe proportional zur Anzahl Beobachtungen in dieser Klasse ist.

Die Wahl der Anzahl Klassen ist relevant für die Aussagekraft eines Histogrammes. Es gibt keine allgemeine Grundregel, wie wir die Anzahl Klassen wählen sollen.

Beispiel 2.3.3

In Abbildung 2.3 sehen wir Histogramme der IQ Daten von Beispiel 2.3.2 mit verschiedenen Anzahlen von Klassen aufgezeichnet.

Das Histogramm links oben ist viel zu detailiert, als dass man ein Muster erkennen könnte.

Auf der anderen Seite ist das Histogramm rechts unten zu ungenau.

Die beiden anderen Histogramm sind ähnlich und haben am meisten Aussagekraft. Es hängt auch vom konkreten Problem ab, wie detailiert das Histogramm sein soll.

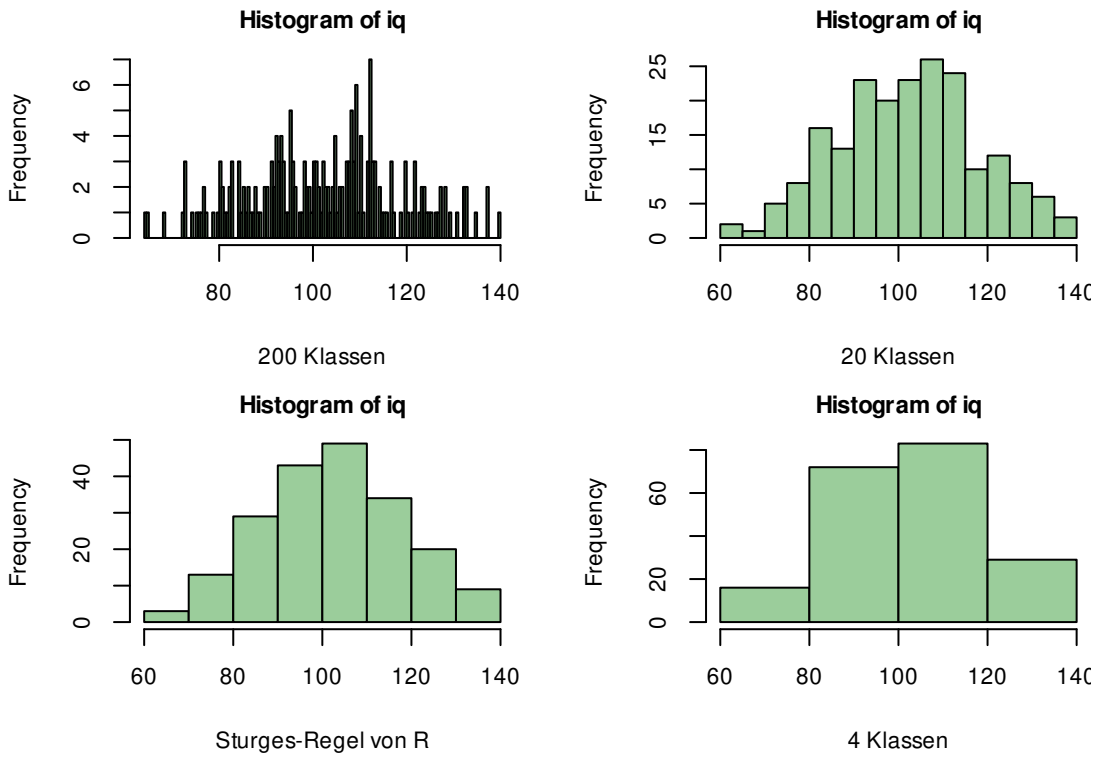


Abbildung 2.3. : Histogramme der IQ-Daten mit verschiedener Klassenwahl

Bemerkungen: R-Code

i. R wählt standardmässig die Anzahl Klassen nach der sogenannten *Sturges*-Regel[6]. Diese ist oft sehr gut, aber halt nicht immer.

ii. R-Code:

```
par(mfrow = c(2, 2))

hist(iq,
     breaks = 100,
     xlab = "200 Klassen",
     col = "darkseagreen3"
)
hist(iq,
     breaks = 20,
     xlab = "20 Klassen",
     col = "darkseagreen3"
)
hist(iq,
     breaks = "sturges", # default R
     xlab = "Sturges-Regel von R",
```

[6]Für diejenigen, die es genau wissen wollen: $k = 1 + \log_2 n$, wobei k die Anzahl Klassen und n die Anzahl Beobachtungen

```
     col = "darkseagreen3"
)
hist(iq,
     breaks = 3,
     xlab = "4 Klassen",
     col = "darkseagreen3"
)
```

iii. Der Befehl **`par(mfrow = c(2, 2))`** erreicht, dass die vier Histogramme in 2 Zeilen (das ist die erste 2) und 2 Spalten zeichnet (die zweite 2).

iv. Mit der Option **`breaks = ...`** wird die Anzahl Klassen festgelegt.

v. Beachten Sie, dass in letzter Graphik **`breaks = 3`** angegeben wurde, aber vier Klassen gezeichnet werden. **`R`** nimmt die Option **`breaks = ...`** nur als Vorschlag. ♦

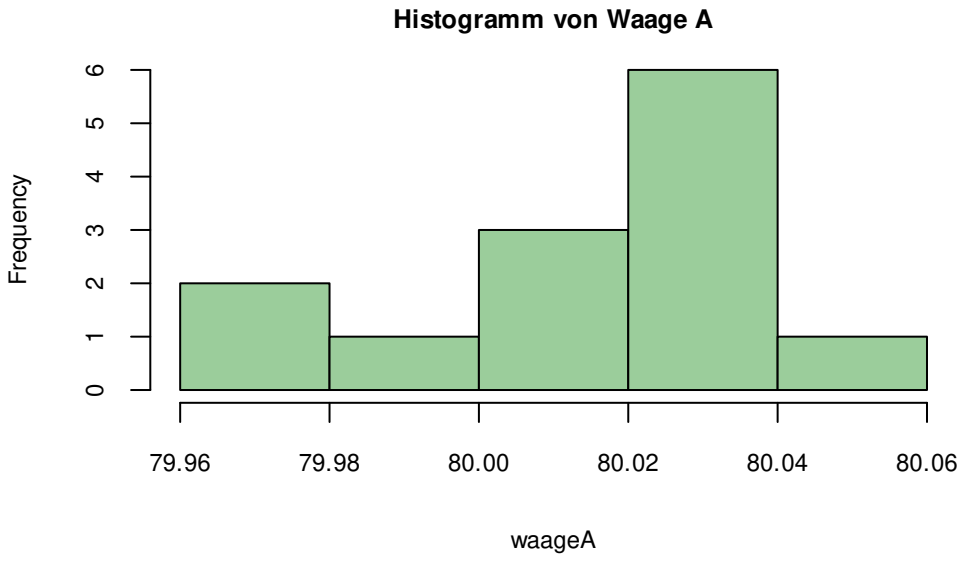
◀

Beispiel 2.3.4

Für das Beispiel der Waage *A* sieht das Histogramm wie folgt aus:

```
par(mar=c(4,4,2,0))
hist(waageA,
     main = "Histogramm von Waage A",
     col = "darkseagreen3"

)
```



Bemerkungen:

i. Bedeutung der Anzahlen (Frequency):

- In der 1. Klasse 79.96-79.98 sind die Beobachtungen mit den Werten 79.97 und 79.98 berücksichtigt.
- In der 2. Klasse 79.99 und 80.00; usw. Der linke Rand wird also nicht berücksichtigt, der rechte dagegen schon.

  Man hätte dies auch umgekehrt machen können, und das Histogramm würde etwas anders aussehen. Bei grossen Datensätzen spielen solche Überlegungen kaum eine Rolle. Die beiden Histogramme würden von Auge kaum erkennbar anders aussehen.

ii. Mit dem **R**-Befehl lassen sich auch die Anzahl und die Breiten der Klassen festlegen, Überschriften ändern, usw.

iii. Vorsicht bei der Interpretation eines Histogrammes (und auch beim Boxplot Abschnitt 2.3.2), wenn der Datenumfang klein ist, wie dies hier für $n = 13$ der Fall ist.

Da wird eine Genauigkeit suggeriert, die gar nicht vorhanden ist. Die Daten sind zufällig und machen wir erneut 13 Messungen, so sieht das Resultat womöglich ganz anders aus. ♦

◀

Beispiel 2.3.5

Der Geysir *Old Faithful* im Yellowstone National Park ist eine der bekanntesten heissen Quellen (*hot springs*). Für die Zuschauer und den Nationalparkdienst ist die Zeitspanne zwischen zwei Ausbrüchen und die Eruptionsdauer von grossem Interesse. Es wurden vom 1.8.1978 - 8.8.1978 insgesamt 107 Messungen von aufeinanderfolgenden Ausbrüchen gemacht. Die Daten sind in der Datei **geysir.txt** enthalten.

```
geysir <- read.table("../../../Themen/Deskriptive_Statistik/Skript_de/Daten/geysir.txt"
head(geysir)


##   X.Tag. Zeitspanne Eruptionsdauer
## 1      1         78            4.4
## 2      1         74            3.9
## 3      1         68            4.0
## 4      1         76            4.0
## 5      1         80            3.5
## 6      1         84            4.1
```

In Abbildung 2.4 sind die Histogramme von Ausbruchsdauer eines Ausbruchs (rechts) und Zeitspanne zwischen zwei Ausbrüchen (links) aufgezeichnet.

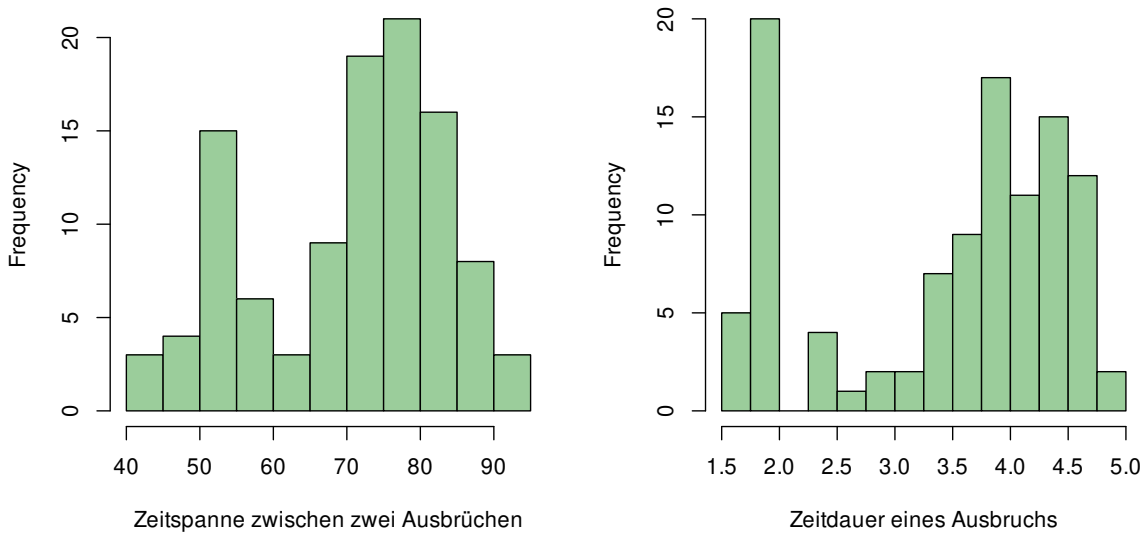


Abbildung 2.4. : Dauer und Zeitspanne von Ausbrüchen des Old aithful

Bei beiden Histogrammen ist ein sogenanntes *bimodales* Verhalten sichtbar: Es gibt zwei „Hügel" im Histogramm:

- Bei der Zeitspanne zwischen zwei Ausbrüchen ist die Dauer relativ kurz (um die 50 Minuten) oder eher lang (um die 80 Minuten).

  Die Zeitdauer zwischen zwei Ausbrüchen ist also nicht „gleichmässig" verteilt.

- Dasselbe Verhalten sehen wir bei der Zeitdauer eines Ausbruchs. Entweder ist der Ausbruch relativ kurz (um die 1.5 bis 2 Minuten) oder lang (um die 4 bis 4.5 Minuten).

Wir können uns jetzt natürlich fragen, ob es einen Zusammenhang zwischen Eruptionsdauer und Zeitspanne zwischen zwei Ausbrüchen gibt. Oder anders gefragt:

- Geht es nach einem langen Ausbruch länger, bis es wieder einen Ausbruch gibt?
- Oder kommt ein Ausbruch schon sehr schnell wieder?
- Oder gibt es gar keinen Zusammenhang?

Wir werden diese Frage im Kapitel 3 beantworten. ◀

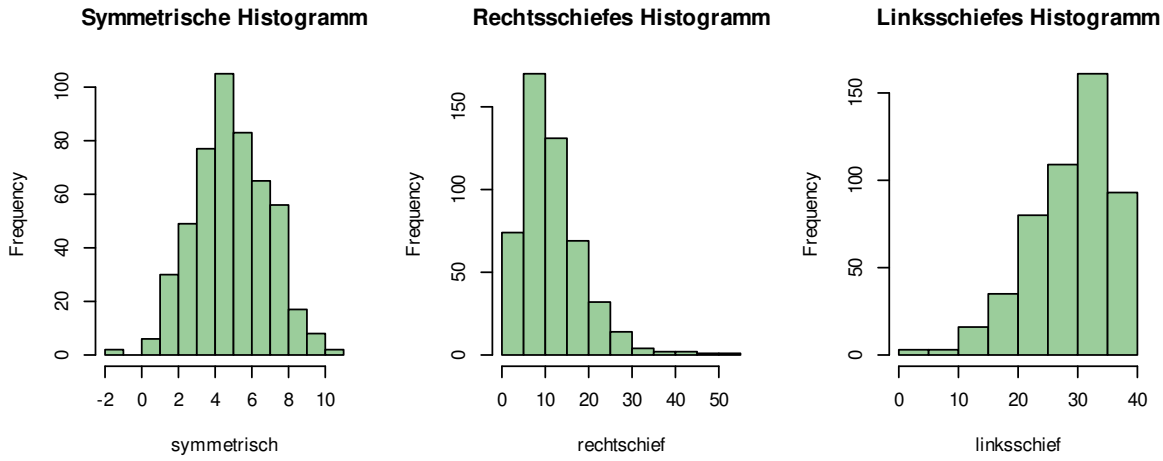


Abbildung 2.5. : Symmetrisches, rechts- und linksschiefes Histogram

Schiefe von Histogrammen

Wir betrachten die Histogramme in Abbildung 2.5.

- Das Histogramm links ist symmetrisch bezüglich (ungefähr) 5. Die Daten sind um 5 auf beiden Seiten ähnlich verteilt.
- Im mitteren Histogramm sind die meisten Daten links im Histogramm. Wir sprechen dann von einem *rechtsschiefen* Histogramm.
- Im rechten Histogramm sind die meisten Daten rechts im Histogramm. Wir sprechen dann von einem *linksschiefen* Histogramm.

Die Bezeichnung „rechts" und „links" bezieht sich immer auf die Richtung, wo *weniger* Daten vorhanden sind.

Normiertes Histogramm

In den Histogrammen bisher entspricht die Höhe der Balken gerade der Anzahl der Beobachtungen in einer Klasse. Oft ist es besser und übersichtlicher, wenn wir die Balkenhöhe so wählen, dass die *Balkenfläche* dem Anteil der jeweiligen Beobachtungen an der Gesamtanzahl Beobachtungen entspricht. Die Gesamtfläche aller Balken muss dann gleich eins sein.

Auf der vertikalen Achse ist dann die sogenannte *Dichte* aufgetragen. Wichtig: Dies sind *keine* Prozentzahlen, wie oft gemeint wird.

Beispiel 2.3.6

Im folgenden Histogramm der Messwerte der Waage A ist auf der vertikalen Achse nun die Dichte angegeben.

```
hist(waageA,
     freq = F,
     main = "Histogramm von Waage A",
     col = "darkseagreen3",
     ylim = c(0, 25)

)
rect(80.02, 0, 80.04, 23.1, col="darkseagreen4")
```

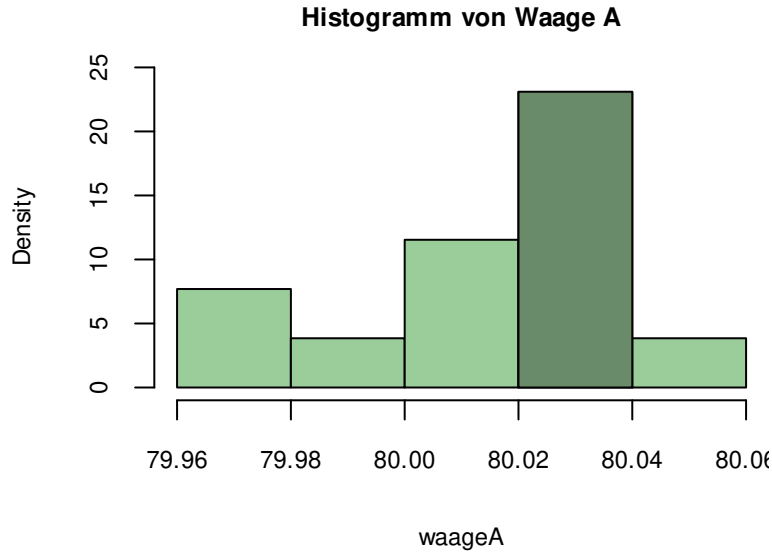


Abbildung 2.6. : Normiertes Histogramm der Waage A

Die Dichte der Klasse von $80.02 - 80.04$ ist etwa 23. Somit gilt für die Fläche dieses Balkens (die dunkelgrüne Fläche in Abbildung 2.6):

$$(80.04 - 80.02) \cdot 23 = 0.46$$

Diese Fläche mit 100 multipliziert entspricht dann gerade der Prozentzahl der Daten, die in diesem Abschnitt des Wertebereiches liegen. In diesem Beispiel befinden sich etwa 46 % der Daten zwischen 80.02 und 80.04.

Bemerkungen: R-Code

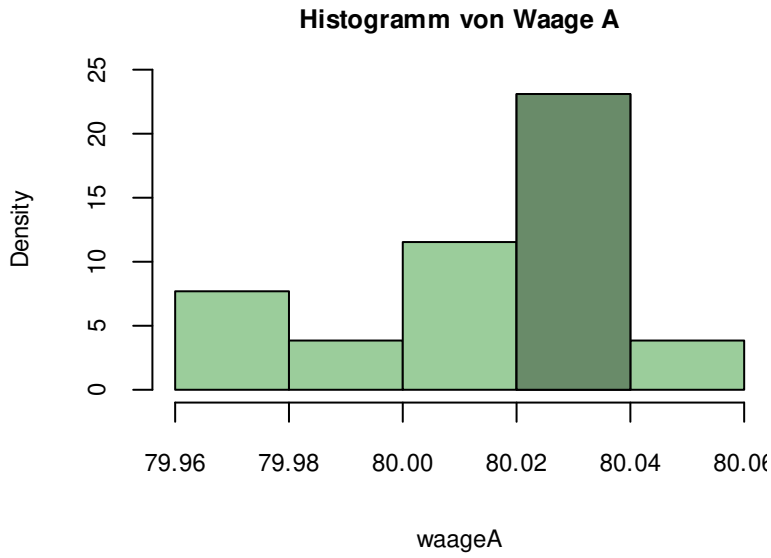
i. Mit der Option **`freq = F`** (*frequency false*) wird das Histogramm normiert gezeichnet.

ii. Die Option **`ylim = c(0, 25)`** gibt an, in welchem Bereich die vertikale Achse (y-Achse) gezeichnet werden soll.

R wählt diesen Bereich standardmässig selbst, ist aber nicht immer optimal. Ohne dieses Option **`ylim = c(0, 25)`** zeichnet das **R** das Histogramm wie folgt:

```
hist(waageA,
     freq = F,
     main = "Histogramm von Waage A",
     col = "darkseagreen3"

)
rect(80.02, 0, 80.04, 23.1, col="darkseagreen4")
```



Die y-Achse ragt nicht über den höchsten Balken hinaus und dessen Wert ist schwierig abzulesen.

iii. Mit dem Befehl **`rect(80.02, 0, 80.04, 23.1, col="darkseagreen4")`** wird ein Rechteck in eine vorgegebene Graphik gezeichnet.

Die ersten beiden Zahlen die Koordinaten des Punktes links unten des Rechteckes und die zweiten beiden Zahlen die Koordinaten des Punktes rechts oben.

♦

◀

Diese Darstellung hat den Vorteil, dass wir Messungen mit unterschiedlichen Umfängen besser miteinander vergleichen können. Würden wir also mit Waage A nun eine Messung mit 30 Beobachtungen durchführen, liessen sich mit Dichten die Verteilungen der Messwerten auf die jeweiligen Klassen besser miteinander vergleichen.

Beispiel 2.3.7

Wir nehmen die Schulnoten einer Klasse aus Beispiel 2.1.6. Diese Klasse vergleichen wir mit einer anderen (hypothetischen) Klasse mit 194 Lernenden, die dieselbe Prüfung gemacht haben.

In Abbildung 2.7 sind die Histogramme der beiden Klassen mit den jeweiligen Häufigkeiten aufgezeichnet. Es ist sehr schwierig die Histogramme miteinander zu vergleichen. Insbesondere kann man nicht sagen, welche Schulklasse die bessere ist.

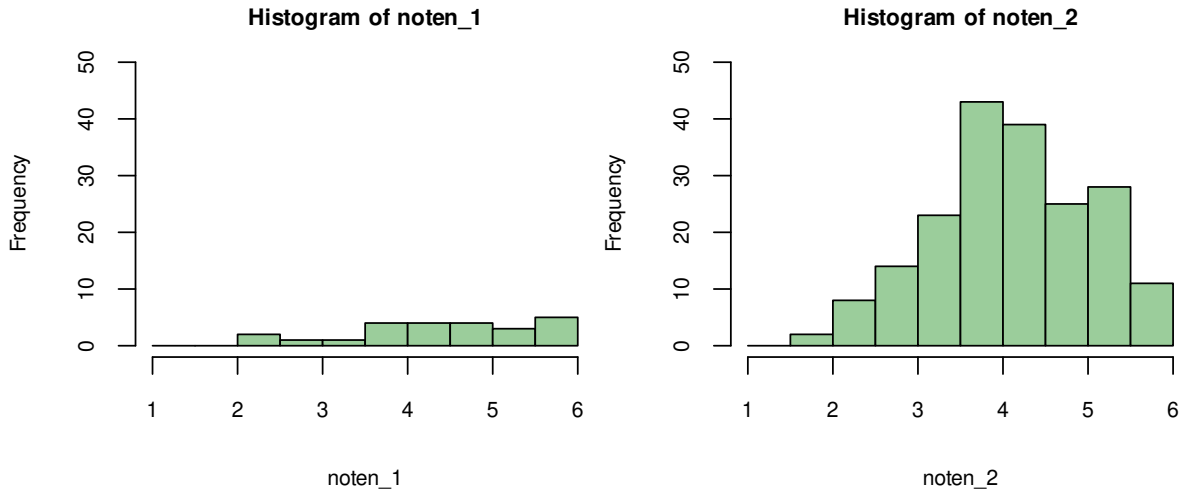


Abbildung 2.7. : Histogramm mit Häufigkeiten von Noten von zwei Schulklassen

Nehmen wir aber die normierten Histogramme, so können wir die Klassen miteinander vergleichen (siehe Abbildung 2.8):

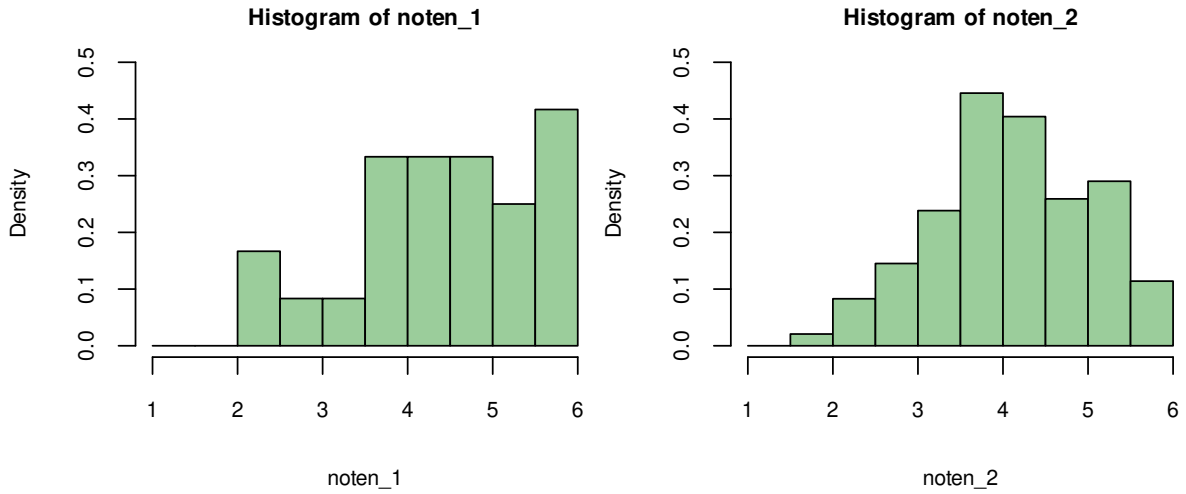


Abbildung 2.8. : Histogramm mit Dichten von Noten von zwei Schulklassen

- Klasse 1 hat im oberen Bereich mehr Anteile als Klasse 2. Vor allem der Balken von 5.5 bis 6 von Klasse 1 ist sehr viel höher als der von Klasse 2. Somit hat es in der Klasse 1 prozentual mehr starke Lernende als in Klasse 2.
- Bei den tiefen Noten sieht es ähnlich aus. Klasse 1 hat eher mehr schwächere Lernende als Klasse 2.
- Aus den ersten beiden Punkte (und den Histogrammen) können wir schliessen, dass im mittleren Bereich der Noten Klasse 2 prozentual mehr Lernende hat als Klasse 1.

Bemerkungen: R-Code

i. Der Code zu Abbildung 2.7 lautet wie folgt:

```
noten_2 <- round(rnorm(n = 200, mean= 4.2, sd = 1), digits = 1)

noten_2 <- noten_2[noten_2 <= 6 & noten_2 >= 1]

par(mfrow = c(1, 2))

hist(noten_1,
     breaks = seq(from = 1, to = 6, by = 0.5),
     ylim = c(0, 50),
     col = "darkseagreen3"

)

hist(noten_2,
     breaks = seq(from = 1, to = 6, by = 0.5),
     ylim = c(0, 50),
     col = "darkseagreen3"

)
```

ii. Mit **`rnorm(n = 200, mean= 4.2, sd = 1)`** wurden hier zufällig 200 normalverteilte Zahlen mit Mittelwert 4.2 und Standardabweichung 1 ausgewählt (siehe Kapitel 5).

iii. Mit **`round(..., digits = 1)`** werden diese 200 Zahlen auf eine Stelle nach dem Komma gerundet.

iv. Nun hat es in diesen 200 zufällig ausgewählten Zahlen noch solche, die grösser als 6 und kleiner als 1 sind.

```
##   [1] 3.4 5.6 2.9 4.3 5.9 3.6 3.7 3.6 3.9 4.3 5.4 3.4 3.1 4.0 3.1
##  [16] 4.1 3.6 2.0 4.4 3.9 5.1 5.1 5.7 4.9 5.0 3.9 5.6 5.7 3.5 3.3
##  [31] 4.5 5.3 6.4 5.4 5.7 5.2 3.2 2.2 2.4 4.1 5.8 3.4 4.1 6.1 3.7
##  [46] 4.8 3.3 3.7 3.5 4.1 5.7 4.4 5.2 3.6 4.1 3.3 5.0 4.1 4.1 4.4
##  [61] 3.1 5.1 3.6 4.7 3.4 3.9 2.1 3.9 2.9 3.9 4.0 4.0 4.5 4.2 4.6
##  [76] 4.0 5.2 4.3 4.4 3.6 4.7 2.5 5.2 4.2 4.9 3.5 6.6 3.7 4.1 3.7
```

```
##  [91] 5.1 3.1 4.8 5.1 5.2 4.6 3.9 3.7 4.0 4.1 2.2 5.3 4.9 4.4 4.1
## [106] 5.1 4.0 2.2 3.4 5.5 3.2 5.8 6.8 4.3 2.8 5.0 2.6 4.7 4.3 4.0
## [121] 5.4 5.1 2.9 2.6 5.3 4.5 3.8 5.4 2.8 2.8 5.5 2.2 3.0 4.1 4.9
## [136] 4.7 4.5 3.1 4.8 4.5 3.4 4.2 6.4 5.2 3.9 3.8 4.3 4.0 2.8 3.8
## [151] 5.1 5.0 3.7 5.0 3.6 5.4 3.9 5.4 3.8 4.4 1.6 6.4 4.3 5.8 3.7
## [166] 3.5 4.2 4.1 4.2 3.7 2.8 4.3 3.0 2.4 4.2 4.5 5.8 3.1 5.1 3.6
## [181] 6.4 3.5 5.0 5.2 2.9 3.9 4.0 3.8 4.9 3.0 4.9 4.6 3.7 4.7 4.0
## [196] 5.5 5.0 4.3 3.5 4.9
```

So ist die 33. Zahl 6.1. Alle diese müssen wir noch entfernen. Aus **`noten_2[...]`** werden nur diejenigen Werte ausgewählt, die kleiner gleich 6 (**`noten_2 <= 6`**) und (**`&`**) grösser gleich 1 (**`noten_2 >= 1`**) sind.

v. Dem Befehl **`par(mfrow = c(1, 2))`** sind wir schon begegnet. Hier werden die Graphiken in einer Zeile und zwei Spalten gezeichnet, also nebeneinander.

vi. Die Option **`breaks = ...`** hat hier eine andere Form als in Beispiel 2.3.3, wo die Anzahl Klassen angegeben wurde.

Hier wird **`R`** mit der Option **`breaks = seq(from = 1, to = 6, by = 0.5)`** angegeben, *wo* die Klassengrenzen (breaks) gemacht werden sollen. In diesem Fall von 1 bis 6 in Schritten von 0.5.

Allgemein kann man die Klassengrenzen mit **`breaks = c(...)`** fast beliebig wählen.

vii. Der Code für Abbildung 2.8 enthält noch zusätzlich **`freq = F`**:

```
hist(noten_1, breaks = seq(from = 1, to = 6, by = 0.5), freq = F,
    col = "darkseagreen3", ylim = c(0, 0.5))

hist(noten_2, breaks = seq(from = 1, to = 6, by = 0.5), freq = F,
    col = "darkseagreen3", ylim = c(0, 0.5))
```

2.3.2. Boxplot

Ein *Boxplot* ist in Abbildung 2.9 schematisch dargestellt.

Der Boxplot besteht aus:

- Einem Rechteck, dessen Höhe vom empirischen 25 %- und vom 75 %-Quantil begrenzt wird (grüne Fläche)

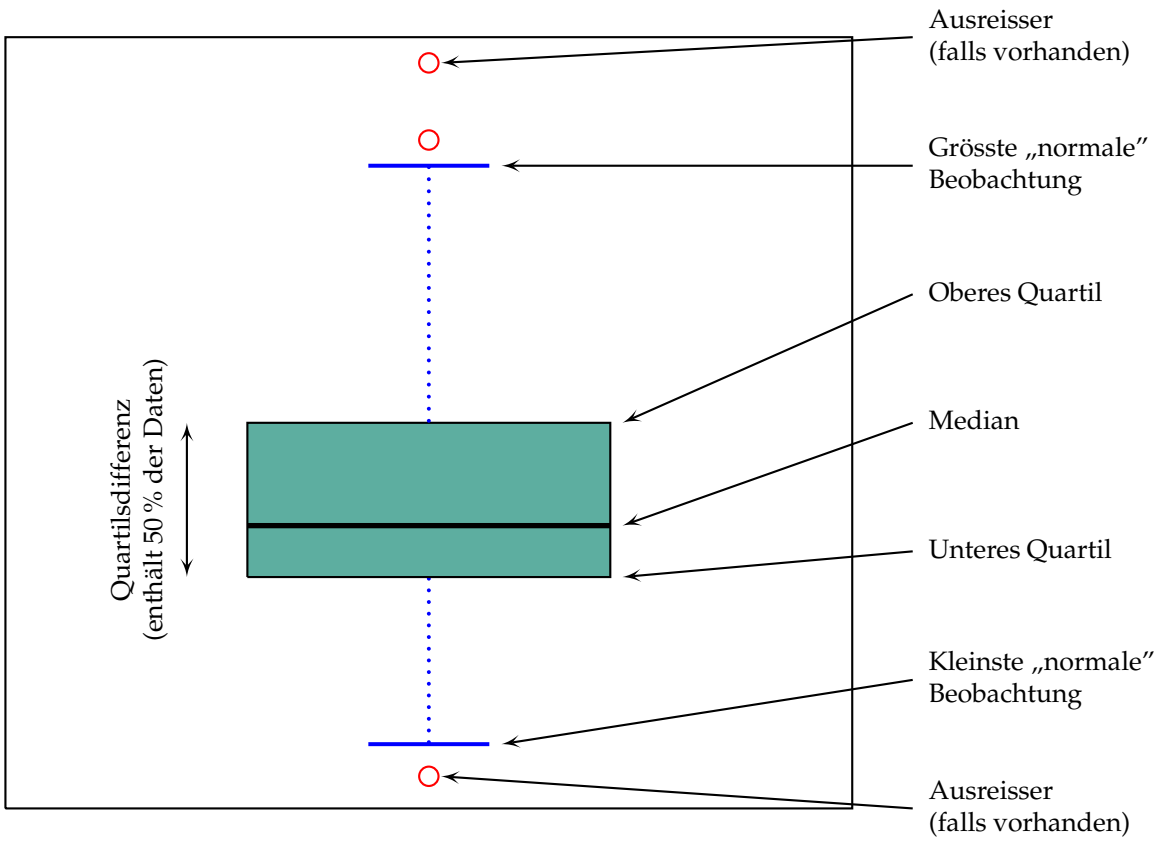


Abbildung 2.9. : Schematischer Aufbau eines Boxplots

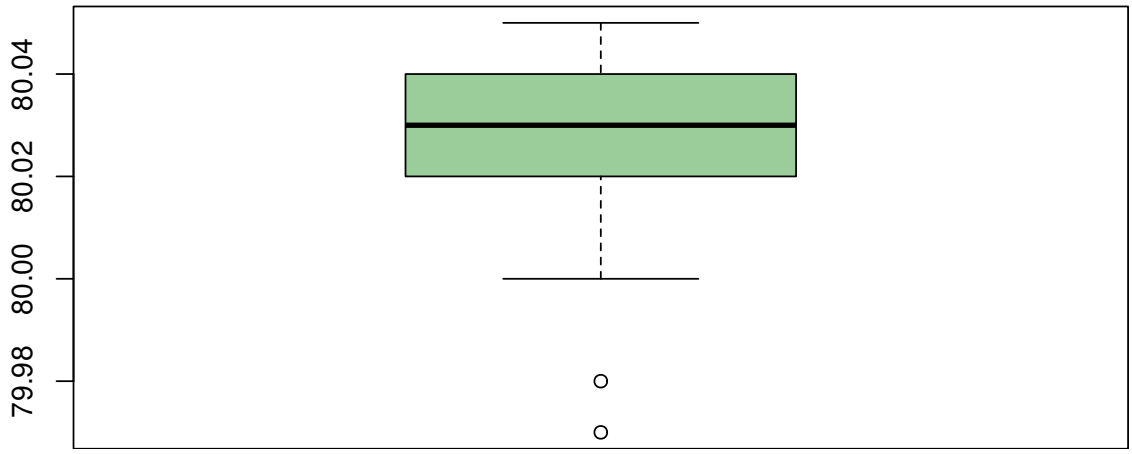
- Einem horizontalen Strich in der Box für den Median (schwarz)
- Linien, die von diesem Rechteck bis zur kleinsten- bzw. grössten „normalen" Beobachtung führen (blau eingezeichnet). Diese Linien werden *whiskers* genannt.

  Per Definition ist der grösste „normale" Wert *höchstens* 1.5 mal die Quartilsdifferenz vom oberen Quartil entfernt und entsprechend ist der kleinste normale Wert *höchstens* 1.5 mal die Quartilsdifferenz vom unteren Quartil entfernt.
- Kleinen Kreisen, die Ausreisser markieren (rot). Ausreisser liegen ausserhalb der normalen Beobachtungen.

Beispiel 2.3.8 Boxplot Waage *A*

Der **R**-Befehl für den Boxplot ist **`boxplot(...)`**:

```
boxplot(waageA, col = "darkseagreen3")
```



Bemerkungen:

i. Die Hälfte der Beobachtungen befindet sich zwischen dem oberen Quartil 80.04 und dem unteren Quartil 80.02, mit Quartilsdifferenz 0.02

ii. Der Median liegt bei 80.03.

iii. Der „normale" Bereich der Werte liegt zwischen 80.00 und 80.05.

iv. Wir haben zwei Ausreisser 79.97 und 79.98.

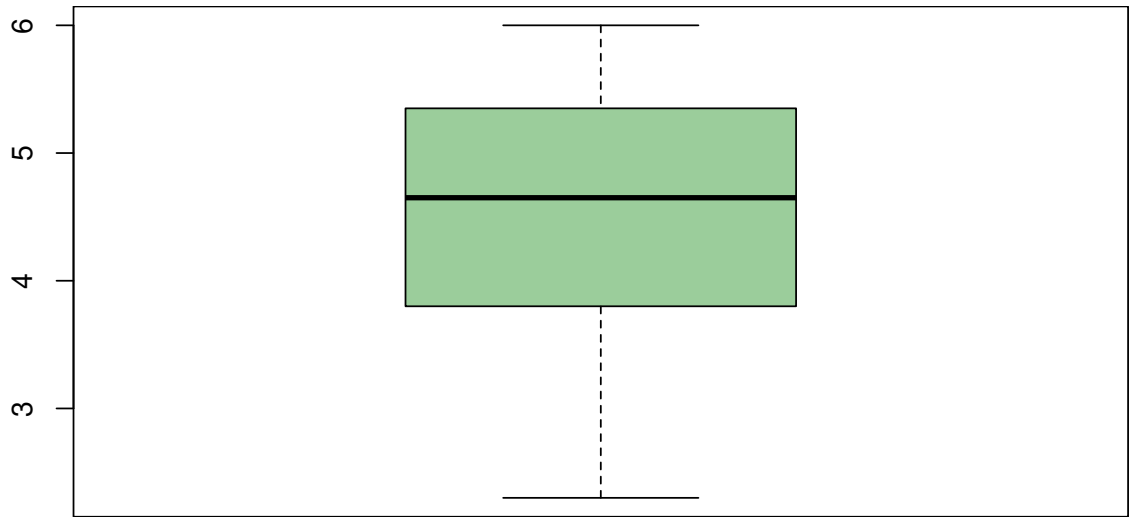
v. Die Punkte i) und ii) hatten wir schon bei den Quantilen berechnet (siehe Beispiel 2.2.17). Der Boxplot stellt somit unsere Berechnungen graphisch dar. ♦

◀

Beispiel 2.3.9

Hier noch der Boxplot der Noten aus Beispiel 2.1.6.

```
par(mar = c(0, 4, 0, 0))
boxplot(noten, col = "darkseagreen3")
```



Auch hier entsprechen die Werte für den Median und die Quartile den Werten, die wir schon in Beispiel 2.2.19 gefunden haben. ◀

Der Boxplot ist vor allem dann geeignet, wenn wir die Verteilungen der Daten in verschiedenen Gruppen, die im Allgemeinen verschiedenen Versuchsbedingungen entsprechen, vergleichen wollen.

Beispiel 2.3.10

Bei unserem Einführungsbeispiel 2.1.4 haben wir zwei Datensätze erhalten. Wir können die Boxplots der beiden Messreihen gegenüberstellen und die Messwerte der Waagen miteinander vergleichen (siehe Abbildung 2.10).

- Waage A liefert die grösseren Werte als Waage B, da der Median von A grösser ist.Der Median von Waage A ist grösser als der der Waage B. Auch überschneiden sich die Boxen nicht.
- Die Daten von Waage A haben weniger Streuung als die Daten von Waage B, da das Rechteck weniger hoch ist (Quartilsdifferenz!).

Bemerkungen:

i. Wir werden den Vergleich von Boxplots im Kapitel über Varianzanalyse **??** wieder begegnen.

ii. `R`-Code: Mit dem Befehl `axis(...)` lässt sich die Beschriftung der Achsen ändern:

- Mit `side = ...` wird die Seite gewählt: `1` unten, `2` links, `3` oben, `4` rechts

```
boxplot(waageA, waageB,
        xlab = "Waage",
        col = c("orange", "lightblue")

)

axis(side = 1, at = c(1, 2), labels = c("A", "B"))
```

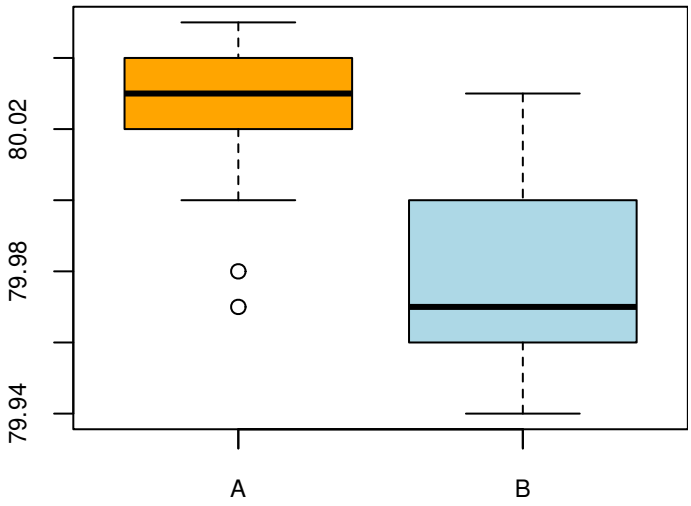


Abbildung 2.10. : Boxplots für die Gewichtsmessungen der zwei Waagen

- Mit `at = ...` werden die Stellen auf der jeweiligen Seite gewählt. In diesem Fall unten an der Stelle 1 und 2.
- Mit `labels = ...` wird angegeben, was an den entsprechenden Stellen geschrieben werden soll.

♦

◀

Beispiel 2.3.11

In Abbildung 2.5 haben Histogramme mit verschiedener Schiefe gesehen. In Abbildung 2.11 sind dazu noch die zugehörigen Boxplots gezeichnet.

- Beim symmetrischen Diagramm links ist der Median in der Mitte der Box.
- Beim rechtsschiefen Histogramm in der Mitte ist für den Boxplot der Median nicht mehr in der Mitte der Box, sondern nach links verschoben.

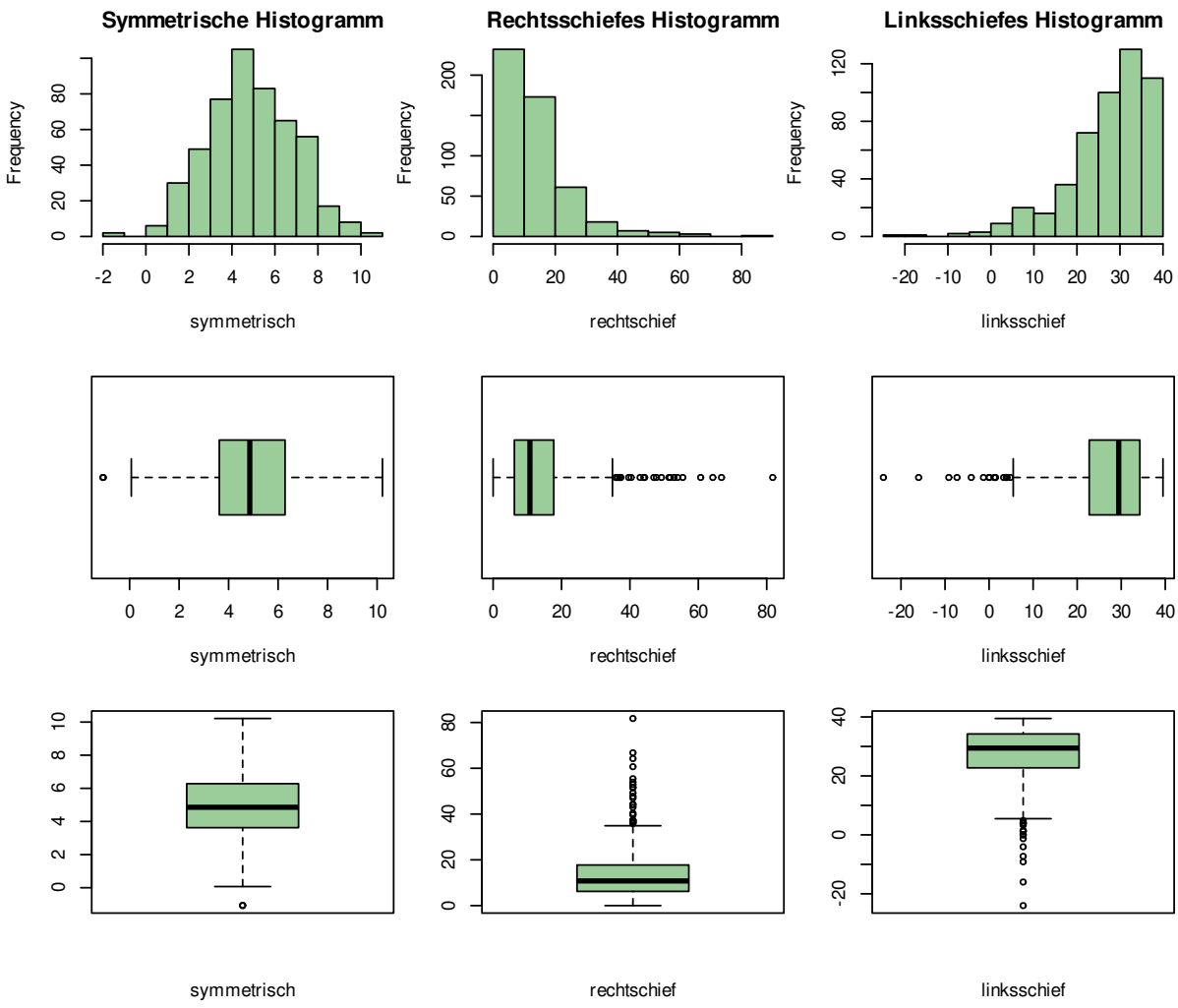


Abbildung 2.11. : Histogramme mit zu gehörigen Boxplots

Der Abstand vom unteren Quartil zum Median ist kleiner als der Abstand vom Median zum oberen Quartil.

Dies liegt daran, dass es im vom unteren Quartil zum Median viele Daten in kleinem Bereich liegen. Vom Median zum oberen Quartil braucht es ein viel grösserer Bereich bis 25 % der Daten in diesem Bereich liegen.

- Beim linksschiefen Histogramm ist die Interpretation gerade umgekehrt.

◀

Beispiel 2.3.12

In Abbildung 2.12 ist das Histogramm und der Boxplot von der Zeitspanne zwischen zwei Ausbrüchen von Old Faithful aus Beispiel 2.3.5 aufgeführt.

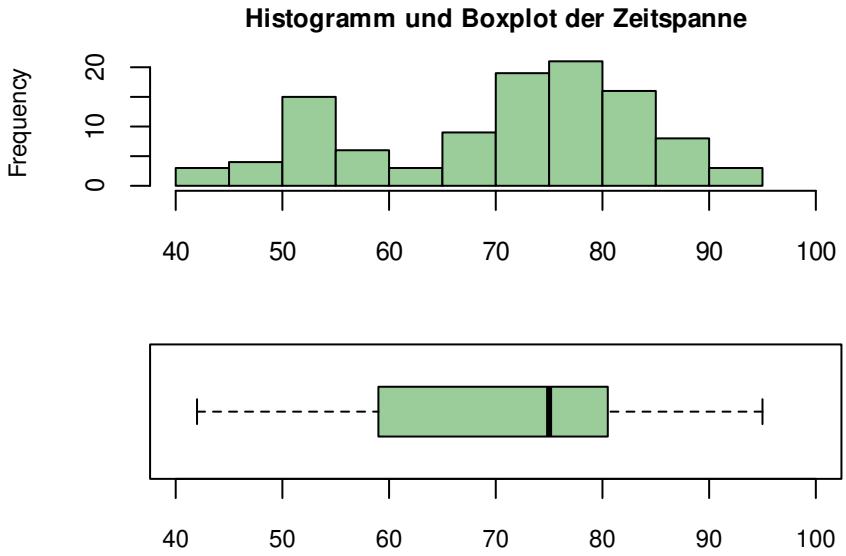


Abbildung 2.12. : Histogramm und Boxplot für die Zeitspanne zwischen zwei Ausbrüchen von Old Faithful

Wir können Folgendes herauslesen:

- Die Daten sind linksschief, was sowohl aus dem Histogramm als auch dem Boxplot deutlich zum Vorschein kommt.
- Aus dem Boxplot erkennen wir, dass die 50 % der mittleren Zeitspannen zwischen 60 und 80 Minuten liegen.
- Der Median liegt bei etwa 75 Minuten.
- Die Daten zwischen dem Median und dem oberen Quartil liegen in einem Bereich von 5 Minuten (von 75-80 Minuten). Das heisst in diesem Bereich befinden sich relativ viele Zeitspannen verglichen zu Abstand 15 Minuten vom unteren Quartil zum Median.
- Das bimodale Verhalten, dass im Histogramm erkennbar ist, erscheint im Boxplot *nicht* mehr.

◀

2.3.3. Intermezzo: Festlegungen in der Statistik

Sie wundern sich vielleicht, woher der Wert für das 1.5-fache der Quartilsdifferenz Abstand von den Quartilen als grösster bzw. kleinster „normale" Beobachtung für

den Boxplot kommt. Der Erfinder des Boxplots, Statistiker John Tukey, hat 1970 diesen Wert so *festgelegt* und seither machen es alle so.

Natürlich hätte er diesen Wert auch anders festlegen können, aber aus seiner Erfahrung als Statistiker hatte er diesen Wert als vernünftig erachtet. Hätte er diesen Wert kleiner gewählt, so hätte es zuviele Ausreisser gegeben; hätte er diesen grösser gewählt, so hätte es zuwenig Ausreisser gegeben.

Nun ist natürlich die Frage, was ist „zuviele" oder „zuwenig"? Diese Frage kann niemand beantworten.

Solche scheinbar willkürlichen Festlegungen kommen in der Statistik oft vor und wir diesen noch einige Male begegnen. Gewisse Grössen werden *festgelegt*, die man auch anders hätte wählen können.

Im Gegensatz zur restlichen Mathematik, wo etwas entweder richtig oder falsch ist, gibt es in der Statistik keine so klare Grenze. Es gibt bei vielen Resultaten erheblichen Interpretationsspielraum.

Kapitel 3.

Deskriptive Statistik – Zweidimensionale Daten

3.1. Einleitung

Bei zweidimensionalen Daten werden an *einem* Versuchsobjekt jeweils *zwei* verschiedene Grössen ermittelt.

Beispiel 3.1.1

An einer Gruppe von Menschen wird jeweils die Körpergrösse *und* das Körpergewicht gemessen.

Das Versuchsobjekt ist dann ein Mensch, zudem zwei Messungen gehören:

- die Körpergrösse
- das Körpergewicht

◀

Beispiel 3.1.2

In Beispiel 2.3.5 haben wir die Eruptionsdauer eines Ausbruchs und die Zeitspanne zwischen zwei Ausbrüchen des Old Faithful ein erstes Mal untersucht. Dort hatten wir die Zeitspanne und die Eruptionsdauer getrennt betrachtet.

Nun wollen diese Grössen *gemeinsam* untersuchen. Das Versuchsobjekt ist ein Ausbruch, zudem zwei Messungen gehören:

- die Eruptionsdauer
- die Zeit bis zum nächsten Ausbruch

◀

Beispiel 3.1.3 Weinkonsum und Mortalität

Wir betrachten als Beispiel einen Datensatz[1] (siehe Tabelle 3.1), der den durchschnittlichen Weinkonsum (in Liter pro Person und Jahr) und die Sterblichkeit (Mortalität) aufgrund von Herz- und Kreislauferkrankungen (Anzahl Todesfälle pro 1000 Personen zwischen 55 und 64 Jahren pro Jahr) in 18 industrialisierten Ländern umfasst.

Es stellt sich nun die Frage, ob diese Daten suggerieren, dass es einen Zusammenhang zwischen der Sterblichkeitsrate aufgrund von Herzkreislauferkrankung und dem Weinkonsum gibt.

Land	Weinkonsum	Mortalität Herzerkrankung
Norwegen	2.8	6.2
Schottland	3.2	9.0
Grossbritannien	3.2	7.1
Irland	3.4	6.8
Finnland	4.3	10.2
Kanada	4.9	7.8
Vereinigte Staaten	5.1	9.3
Niederlande	5.2	5.9
New Zealand	5.9	8.9
Dänemark	5.9	5.5
Schweden	6.6	7.1
Australien	8.3	9.1
Belgien	12.6	5.1
Deutschland	15.1	4.7
Österreich	25.1	4.7
Schweiz	33.1	3.1
Italien	75.9	3.2
Frankreich	75.9	2.1

Tabelle 3.1. : Weinkonsumation (Liter pro Person pro Jahr) und Mortalität aufgrund von Herzkreislauferkrankung (Todesfälle pro 1000) in 18 Ländern.

Ein kurzer Blick auf die Tabelle zeigt, dass ein höherer Weinkonsum eher weniger Todesfälle wegen Herz- und Kreislaufkrankheiten zur Folge hat.

Wir werden im Weiteren untersuchen, ob dieser Zusammenhang besteht oder nicht.

[1] A.S.St.Leger, A.L.Chochrane, and F.Moore, „Factors Associated with Cardiac Mortality in Developed Countries with Particular Reference to the Consumption of Wine."*Lancet, 1979*

◀

3.2. Graphische Darstellung: Streudiagramm

3.2.1. Streudiagramme

Ein erster wichtiger Schritt in der Untersuchung zweidimensionaler Daten ist die graphische Darstellung dieser Daten. Dies geschieht meist mit Hilfes eines sogenannten *Streudiagrammes* (engl.: *Scatterplot*). Dabei werden die beiden Messungen einer Versuchseinheit als *Koordinaten* von Punkten in einem Koordinatensystem interpretiert und dargestellt.

Beispiel 3.2.1

Das Streudiagramm für Beispiel 3.1.3 des Weinkonsums ist in Abbildung 3.1 dargestellt. Wie es zustande kommt, wird gleich erklärt.

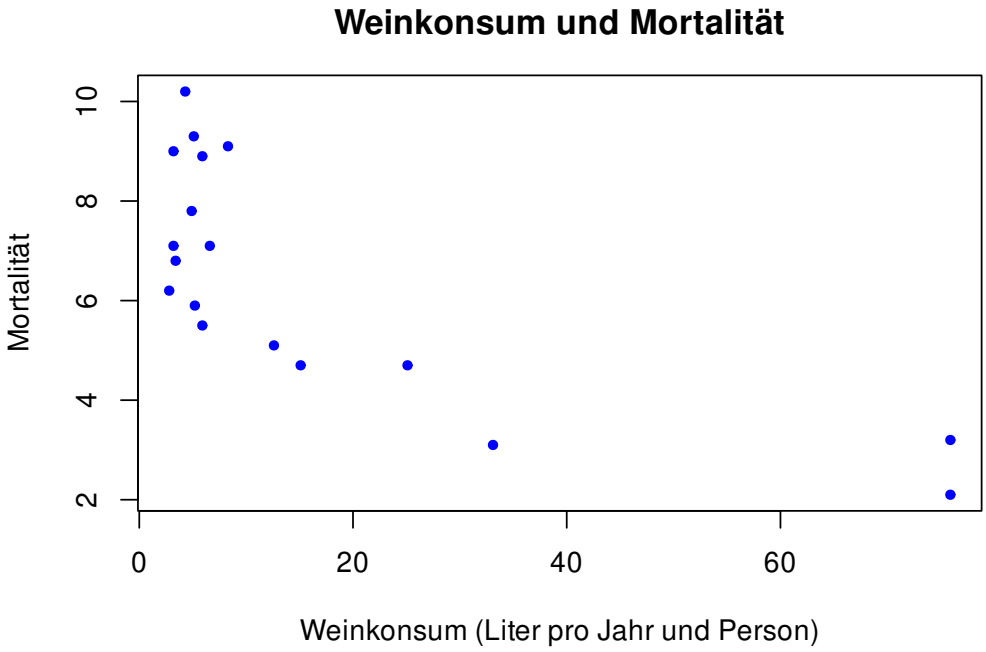


Abbildung 3.1. : Streudiagramm für die Mortalität und den Weinkonsum in 18 industrialisierten Ländern.

In Beispiel 3.1.3 stellt ein Land eine Versuchseinheit dar. Es wird die Grösse „Weinkonsum"

$$x_1,\ x_2,\ \ldots,\ x_{18}$$

und die zugehörige Grösse „Mortalität"

$$y_1,\ y_2,\ \ldots, y_{18}$$

gemessen. Wir können diese Daten als Koordinaten der Punkte

$$(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_{18}, y_{18})$$

interpretieren. So ist

$$(x_1, y_1) = (2.8,\ 6.2)$$

der Punkt mit den Koordinaten von Norwegen.

Sind die Daten in dieser Form gegeben, so interessieren uns in erster Linie für die *Zusammenhänge* und *Abhängigkeiten* zwischen den Variablen x und y. Hier konkret: Gibt es einen Zusammenhang zwischen dem Weinkonsum und der Mortalität wegen Herzkreislauferkrankungen?

Die Abhängigkeit zwischen den beiden Messgrössen können wir aus dem *Streudiagramm* ersehen, welches die Daten als Punkte in der Ebene darstellt (siehe Abbildung 3.1): Die i-te Beobachtung (i-tes Land) entspricht dem Punkt mit Koordinaten (x_i, y_i). Die Abbildung 3.1 zeigt das Streudiagramm für die Messgrössen „Weinkonsum" $(x_1, x_2, \ldots, x_{18})$ und „Mortalität" $(y_1, y_2, \ldots, y_{18})$.

Wir sehen einen klaren monoton fallenden Zusammenhang: Länder mit hohem Weinkonsum haben also eine Tendenz zu einer tieferen Mortalitätsrate wegen Herz- und Kreislaufkrankheiten.

Bemerkungen:

i. Die Schlussfolgerung, dass hoher Weinkonsum gesund *ist*, ist voreilig und vermutlich *falsch*.

   Es *scheint*, dass höherer Weinkonsum zu weniger Toten wegen Herz- und Kreislaufkrankheiten führt. Der Einfluss des höheren Weinkonsums auf andere Körperorgane (z.B. Leber) oder auf die Anzahl Verkehrsunfälle, wird hier *nicht* untersucht.

ii. Obwohl sich aufgrund des Streudiagramms ein Zusammenhang zwischen Weinkonsum und Mortalität *erahnen* lässt, muss nicht zwingend ein *kausaler* Zusammenhang zwischen den beiden Grössen vorhanden sein (siehe auch Abschnitt 3.2.2).

iii. Wir haben hier für die x-Koordinaten den Weinkonsum und für die y-Koordinaten die Mortalität gewählt. Dies hätten wir auch umgekehrt machen können.

Was wir als x- und y-Koordinate wählen, hängt von der Problemstellung ab. In diesem Fall wollen wir wissen, wie sich der Weinkonsum (unabhängige Variable x) auf die Mortalität (abhängige Variable y) auswirkt.

iv. Man bezeichnet die Punkte in einem Streudiagramm auch als *Punktwolke*.

v. R-Code:

```
wein <- c(2.8, 3.2, 3.2, 3.4, 4.3, 4.9, 5.1, 5.2, 5.9, 5.9, 6.6, 8.3,
    12.6, 15.1, 25.1, 33.1, 75.9, 75.9)
mort <- c(6.2, 9, 7.1, 6.8, 10.2, 7.8, 9.3, 5.9, 8.9, 5.5, 7.1, 9.1,
    5.1, 4.7, 4.7, 3.1, 3.2, 2.1)

plot(wein, mort, xlab = "Weinkonsum (Liter pro Jahr und Person)",
    ylab = "Mortalität", main = "Zusammenhang zwischen Weinkonsum und Mortalität",
    pch = 20, col = "blue")
```

Beispiel 3.2.2

Wir können die Daten des Old Faithful (siehe Beispiele 2.3.5 und 3.1.2) als Streudiagramm darstellen (siehe Abbildung 3.2).

Wir stellen drei Muster fest:

- Zunächst ist die Punktwolke steigend. Je länger die Zeitspanne zwischen den Ausbrüchen, umso länger dauert der Ausbruch.
- Im Gegensatz zu Abbildung 3.1, wo die Punkte einer krummen Kurve zu folgen scheinen, folgen die Punkte eher einer Geraden. Was diese Gerade bedeutet und wie man diese bestimmt, werden wir in Unterkapitel 3.3 kennenlernen.
- Im Streudiagramm hat es zwei Gruppen: Eine links unten und eine rechts oben.

  Dies bedeutet Folgendes: Entweder ist die Zeitspanne zwischen zwei Ausbrüchen kurz und die nächste Eruptionsdauer kurz oder die Zeitspanne ist lang und der Eruptionsdauer ist lang.

  Eine mittlere Zeitspanne (um die 70 Minuten) mit einer mittleren Ausbruchsdauer (um die 3 Minuten) gibt es nicht.

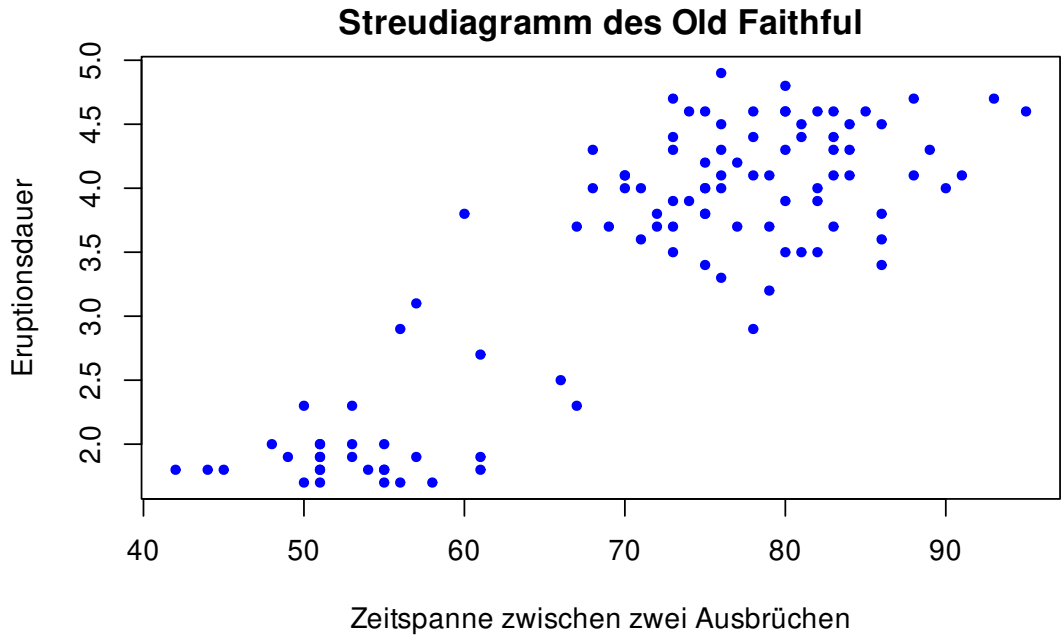


Abbildung 3.2. : Streudiagramm für die Eruptionsdauer und Zeitspanne zwischen zwei Ausbrüchen des Old Faithful

Es ist nicht ganz klar, welche Grösse wir als x- und welche als y-Koordinate wählen sollen. Hier ist x die Zeitspanne zwischen zwei Ausbrüchen, die einen Einfluss auf die Eruptionsdauer hat.

Wir können uns natürlich auch umgekehrt fragen, ob die Eruptionsdauer einen Einfluss auf die Zeitspanne bis zum nächsten Ausbruch hat.

Wie schon erwähnt, liegt es oft an der Fragestellung, welches die x- und welches die y-Koordinate ist. In diesem Fall stellt sich die Frage, was zuerst da war, das Huhn oder das Ei. ◀

3.2.2. Abhängigkeit und Kausalität

Bei Streudiagrammen müssen wir allerdings aufpassen, dass wir *Abhängigkeit* und *Kausalität* nicht miteinander verwechseln. Nur das eine Gesetzmässigkeit vorhanden ist, heisst noch lange nicht, dass diese Gesetzmässigkeit kausal auch erklärt werden kann.

Im Beispiel 3.2.2 des Old Faithful vorher hatten wir einen „je länger desto länger"-Zusammenhang festgestellt. Das Streudiagramm reicht nicht, dass dieser Zusammen-

hang *erklärt* werden kann. Dazu müssen andere Methoden verwendet werden, in diesem Fall geologische.

Folgendes Beispiel zeigt einen Zusammenhang, der nicht kausal erklärt werden kann.

Beispiel 3.2.3

In Abbildung 3.3 sehen wir zwei Kurven, die praktisch aufeinander liegen. Es handelt sich bei dieser Abbildung zwar nicht um ein Streudiagramm, aber dieses würde mit denselben Daten zeigen, dass die Punkte praktisch auf einer steigenden Geraden liegen. Also haben wir auch hier einen (linearen) „je mehr desto mehr"-Zusammenhang.

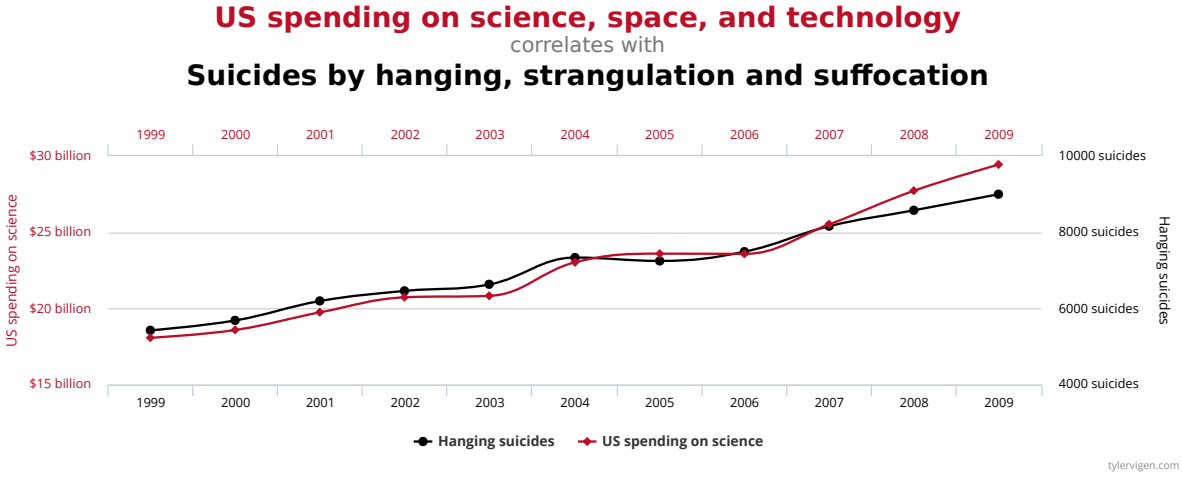


Abbildung 3.3. : Nichtvorhandene Kausalität

Betrachten wir nur ausschliesslich die Kurven, so könnten wir versucht sein zu sagen, dass eine Erhöhung der einen Grösse eine Erhöhung der anderen Grösse *zur Folge* hat, dass es also eine Kausalität zwischen den beiden Grössen gibt.

Betrachten wir aber die konkreten Grössen, die die entsprechenden Kurven beschreiben, so wird ein kausaler Zusammenhang eher schwierig zu erklären sein:

- Rote Kurve: US-Ausgaben für Wissenschaft, Raumfahrt und Technologie
- Schwarze Kurve: Selbstmorde durch Erhängen, Strangulation und Ersticken

Es wird wohl niemand behaupten, dass eine Erhöhung dieser US-Ausgaben für die verschiedenen Gebiete eine Erhöhung der Zahl dieser Selbstmorde zur Folge hat. Der Zusammenhang ist rein zufällig.

Dieses Beispiel stammt aus der Internetseite

https://www.tylervigen.com/spurious-correlations

wo es noch sehr viel mehr solcher Beispiele hat. ◀

Beispiel 3.2.4

Dieses Beispiel stammt aus dem „The NEW ENGLAND JOURNAL of MEDICINE" vom 18. Oktober 2012:

https://www.nejm.org/doi/full/10.1056/NEJMon1211064

Der Autor, Franz Messerli, versucht unter anderem aufgrund dem Streudiagramm in Abbildung 3.4 nachzuweisen, dass Schokolade Einfluss auf die kognitiven Eigenschaften hat.

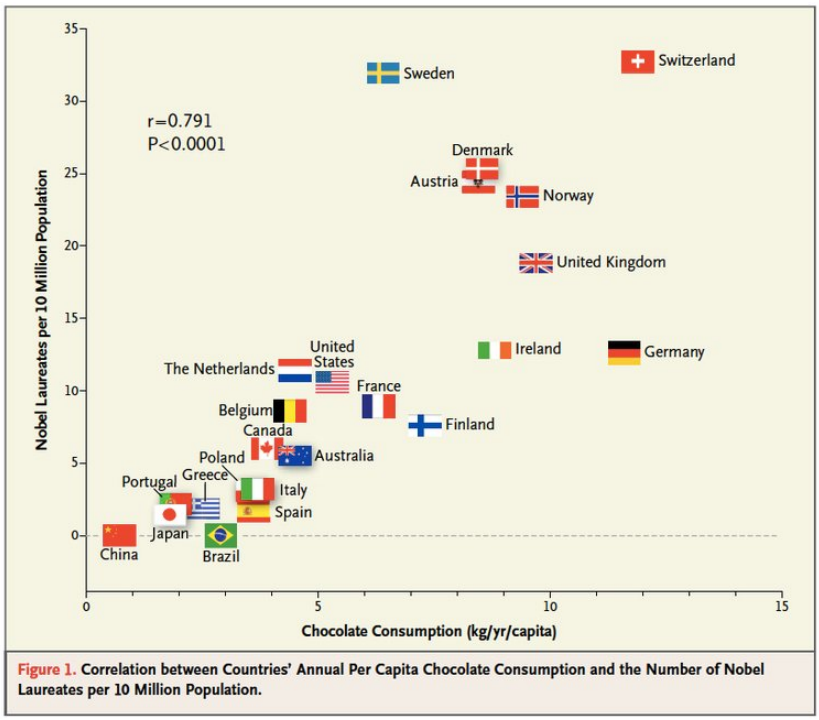


Abbildung 3.4. : Vergleich Schokoladenkonsum und der Zahl Nobelpreisträger.

Im Streudiagramm sind auf der horizontalen Achse der Schokoladenkonsum (Kilogramm pro Kopf pro Jahr) im entsprechenden Land und auf der vertikalen Achse die Anzahl Nobelpreisträger in den entsprechenden Ländern aufgeführt. Dabei wurde die Anzahl Nobelpreisträger auf eine Bevölkerung von 10 Millionen umgerechnet. So hat für die Schweiz mit einer Bevölkerung von etwa 8.5 Millionen ein Nobelpreisträger eine Gewichtung von $10/8.5 \approx 1.2$.

Nun könnten wir aus dem Streudiagramm schliessen, dass der Schokoladekonsum einen Einfluss auf die Anzahl Nobelpreisträger in einem Land hat.

Oder anders gefragt: Hat die Schweiz soviele Nobelpreisträger, *weil* die Bevölkerung der Schweiz soviel Schokolade isst? Und haben die Chinesen keinen Nobelpreis, *weil* sie kaum Schokolade essen?

Nun sind diese Feststellungen so eher zweifelhaft. Wir möchten hier nur einige Punkte aufführen:

- Wir wissen nicht, ob die entsprechenden Nobelpreisträger auch wirklich viel Schokolade gegessen haben. Vielleicht sind alle Nobelpreisträger der Schweiz allergisch auf Schokolade und somit hat das Essverhalten von Schokolade eines Landes keinen Einfluss auf die Leistung eines einzelnen Individuums.
- Schauen wir uns die vorkommenden Ländernamen an, so stellen wir fest, dass es sich fast ausschliesslich um reiche Länder handelt, die viele Nobelpreisträger haben.

  Reiche Länder haben ein besseres Schulsystem und Schokolade ist auch bezahlbar. Somit ist aber das bessere Schulsystem an der hohen Anzahl Nobelpreisträger verantwortlich.

  Wir sprechen hier von einem sogenannten *Confounder* (siehe Design of Experiment). Dies ist eine Grösse, die im Hintergrund für gewisse Effekte verantwortlich ist und dies ist in diesem Beispiel der Wohlstand.
- Warum hat Japan und China so wenige Nobelpreisträger?

  Eine *mögliche* Erklärung ist, dass sich ihre Forschung auf angewandte Probleme ausrichtet, die für die Wirtschaft wichtig sind und dafür bekommt man keine Nobelpreise[2].
- Wir haben noch einen statistischen Ausreisser Schweden, der bei eher geringem Schokoladekonsum sehr viele Nobelpreisträger hervorbringt.

  Nun der Schokoladekonsum ist hier wohl eher nebensächlich. Die Nobelpreiskommitees sind in Schweden beheimatet[3] und ein gewisser Heimvorteil nicht abzustreiten ist.

Diese Punkte sollen aufzeigen, dass nur das betrachten und interpretieren eines Streudiagrammes (oder irgendeiner anderen Graphik) aufgrund der vorkommenden Muster gefährlich ist. Wir *müssen* uns immer bewusst sein, *auf welchen Daten* das Streudiagramm basiert.

[2] Ausser dem Wirtschaftsnobelpreis natürlich, der für theoretische Forschung vergeben wird.
[3] Ausser dem Friedensnobelpreis, der in Norwegen vergeben wird.

Obwohl Graphiken oft sehr nützlich sind und Hinweise auf gewisse Gesetzmässigkeiten sind, sollten wir uns *nie* blindlings nur auf Graphiken verlassen.

Alle oben erwähnten Punkte des möglichen kausalen Zusammenhanges müssen auf einem anderen Wege untersucht werden. Der erwähnte Artikel geht kurz darauf ein.

Die Problematik der Datenerhebung ist Thema von *Design of Experiment* (DoE). ◀

3.3. Einfache lineare Regression

3.3.1. Einleitung

Wir haben im Beispiel 3.2.1 eine negative (je mehr desto weniger) Abhängigkeit zwischen Mortalität und Weinkonsum festgestellt, die einer gekrümmten Kurve folgt.

Oft wird angenommen, dass diese Abhängigkeit sehr einfach ist, nämlich *linear*. Die Punkte folgen dann einer Geraden.

Wir wollen die ganze Idee hinter der linearen Regression im folgenden Beispiel kurz beschreiben und nachher ausführlich darauf eingehen.

Beispiel 3.3.1

Im Beispiel 3.2.2 des Old Faithful hatten wir festgestellt, dass die Punkte mehr oder weniger einer Gerade folgen. Das heisst, wir können in das Streudiagramm eine Gerade einzeichnen (siehe Abbildung 3.5).

Wie man diese Gerade bestimmt, ist Thema dieses Unterkapitels. Bevor wir dies tun, beschäftigen wir uns mit der Frage, *warum* wir dies tun.

Es reicht nicht, dass wir ein Muster in den Daten erkennen, sondern wir wollen auch wissen, welche *Form* dieses Muster hat. Die Beschreibung dieser Form geschieht in der Sprache der Mathematik.

In diesem Fall hat die Gerade die Gleichung (wie man diese genau bestimmt, sehen wir in Beispiel 3.3.8)

$$y = -1.4 + 0.07x$$

oder

$$\texttt{Eruptionsdauer} = -1.4 + 0.07 \cdot \texttt{Zeitspanne}$$

Wir sprechen auch von einer *Modellierung* der Daten. Warum modellieren wir Daten?

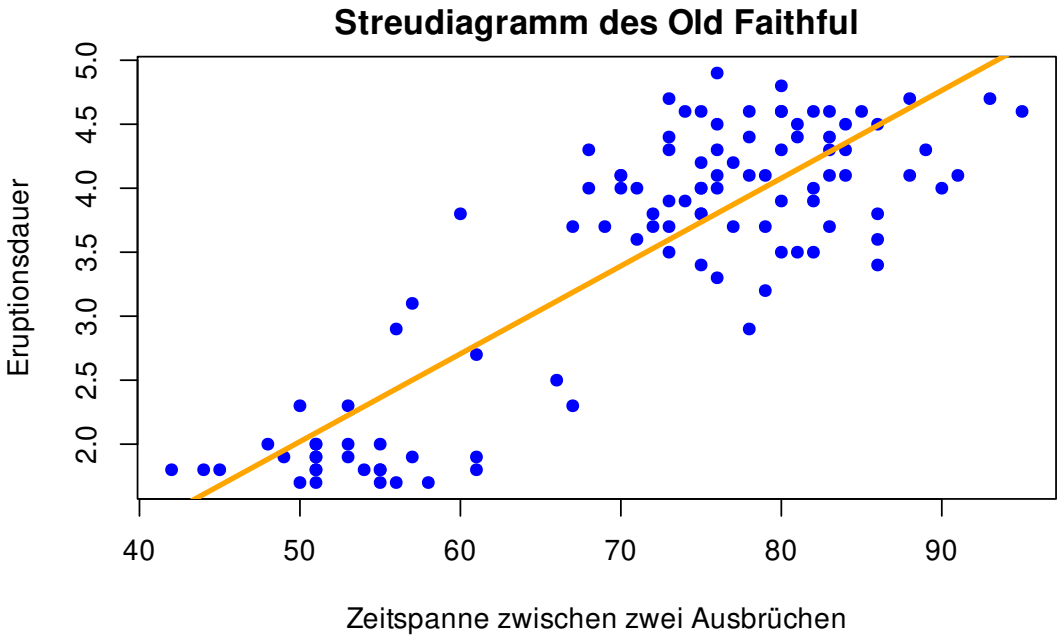


Abbildung 3.5. : Prinzip der linearen Regression für die Eruptionsdauer und Zeitspanne zwischen zwei Ausbrüchen des Old Faithful

- Wir wollen wissen, *wie* sich Daten verhalten. Mit einem Modell können wir weitere Analysen machen, die wir mit Daten wieder überprüfen können.

  In diesem Beispiel lassen sich die Koeffizienten der Geraden *interpretieren*:

  – Der Koeffizient -1.4 ist der y-Achsenabschnitt der Geraden. Er beschreibt den Wert für

  $$x = \texttt{Zeitspanne} = 0$$

  Das heisst, wenn die Zeitspanne zwischen zwei Ausbrüchen 0 Minuten ist, dann dauert der Ausbruch -1.4 Minuten. Das macht natürlich keinen Sinn.

  – Wichtiger ist der Koeffizient 0.07. Dieser entspricht der Steigung der Geraden. Das heisst, nimmt die Zeitspanne zwischen zwei Ausbrüchen um eine Minute zu, so ist der Ausbruch entsprechend 0.07 Minuten länger.

- Haben wir ein Modell, so können wir *Vorhersagen* machen. Im Streudiagramm in Abbildung 3.5 können wir uns fragen, wie lange der Ausbruch dauert, wenn die Zeitspanne zwischen zwei Ausbrüchen 62 Minuten ist.

Ein weiterer Punkt, der wichtig ist: Warum wählen wir nicht ein Modell („Kurve"), das „besser" zu den Punkten passt?

- Das lineare Modell ist das *einfachste*. Vor allem lassen sich die Koeffizienten, wie oben gesehen, interpretieren: y-Achsenabschnitt, Steigung der Geraden.

  Wählen wir beispielsweise ein quadratisches Modell, wie

  $$y = ax^2 + bx + c$$

  so sind die Koeffizienten a, b und c nicht mehr einfach zu interpretieren.

- Kompliziertere Modelle haben den Nachteil des *overfitting*, auf das wir im Kapitel 8 noch zu sprechen kommen. Das Problem ist kurz gesagt, dass eine Genauigkeit vorgespielt wird, die nicht vorhanden ist. Wie gesagt, mehr dazu später.

- Die Erfahrung zeigt, dass lineare Modelle in sehr vielen Fällen zur Lösung von Problemen genügend oder gar am besten sind.

Keep it simple ist oft am besten, was wollen wir mehr.

Das Modell soll das Wirklichkeit, so gut wie möglich widerspiegeln. Es gibt immer ein Abwägen zwischen Genauigkeit und Einfachheit. Wenn das Modell sehr kompliziert ist, ist es auch schwierig zu handhaben und zu interpretieren.

Im Zweifelsfall entscheidet man sich darum eher für einfachere Modelle, die man im „Griff" hat. Und wie schon erwähnt, fahren wir damit, wie auch die Erfahrung mit Machine Learning zeigt, nicht schlecht. ◀

Hier noch ein weiteres Beispiel für den Nutzen einer Modellierung.

Beispiel 3.3.2

Ende 2019 brach in Wuhan, China, die Corona-Epidemie aus. Diese Epidemie entwickelte sich im Verlauf des Frühlings 2020 schnell zu einer Pandemie.

Die Einschnitte in das persönliche und allgemeine Leben waren massiv.

Die entsprechenden Massnahmen wurden in der Schweiz (und nicht nur dort) aufgrund von Modellen getroffen, die vor allem auf Daten aus China und Italien basierten. Mit denen konnte man vorhersagen, wie sich das Virus ausbreitet, wenn social distancing eingehalten wird oder wenn dies eben *nicht* eingehalten wird.

Aber auch in diesem Fall hat man gesehen, dass die Vorhersagen nicht immer korrekt waren. Modelle sollen die Wirklichkeit, so gut wie möglich widerspiegeln, aber sie entsprechen nicht eins-zu-eins der Realität. ◀

Beispiel 3.3.3

Wetterberichte basieren auf Modellen und wir alle wissen, dass diese die Wirklichkeit nicht immer sehr gut abbilden. ◀

Wir erklären das Modell der einfachen linearen Regression zunächst an einem fiktiven Beispiel. Dieses Beispiel entspricht zwar nicht ganz der Realität, veranschaulicht aber die Idee hinter der linearen Regression recht gut.

Beispiel 3.3.4 Zusammenhang Seitenzahl-Preis eines Buches

Wir gehen von der folgenden Annahme aus: Je mehr Seiten ein Roman hat, umso teurer ist er in der Regel. Es gibt also einen Zusammenhang zwischen Seitenzahl x und Buchpreis y.

Wir gehen in einen Buchladen und wählen zehn Romane mit den Seitenzahlen 50, 100, 150, ..., 450, 500 aus. Von jedem Buch notieren wir die Seitenzahl und den entsprechenden Buchpreis. Mit diesen Daten erstellen wir die Tabelle 3.2.

	Seitenzahl	Buchpreis (SFr)
Buch 1	50	6.4
Buch 2	100	9.5
Buch 3	150	15.6
Buch 4	200	15.1
Buch 5	250	17.8
Buch 6	300	23.4
Buch 7	350	23.4
Buch 8	400	22.5
Buch 9	450	26.1
Buch 10	500	29.1

Tabelle 3.2. : Zusammenhang zwischen Buchpreis und Seitenzahl (fiktiv).

Aus der Tabelle ist ersichtlich, dass Bücher mit mehr Seiten tendenziell mehr kosten.

Nun können wir uns folgende Fragen stellen:

- Wieviel kostet eine Seite?
- Wie teuer ist ein Buch mit „null" Seiten?

  Diese Frage ist nicht so unsinnig, wie sie im ersten Moment erscheint. Das wären nämlich die Grundkosten des Verlags, die unabhängig von der Seitenzahl anfallen: Einband, administrativer Aufwand, etc. für jedes Buch.

- Was würde dann voraussichtlich ein Buch mit 375 Seiten kosten?

  Diese Seitenzahl kommt in der Tabelle nicht vor.

Ziel ist es nun, einen formelmässigen Zusammenhang zwischen Buchpreis und Seitenzahl aufzustellen. Haben wir so einen Zusammenhang, so können wir obige Fragen auch beantworten.

Wie könnten wir diesen Zusammenhang mit einer Formel beschreiben? Das Streudiagramm in Abbildung 3.6 links zeigt diesen Zusammenhang graphisch deutlicher auf.

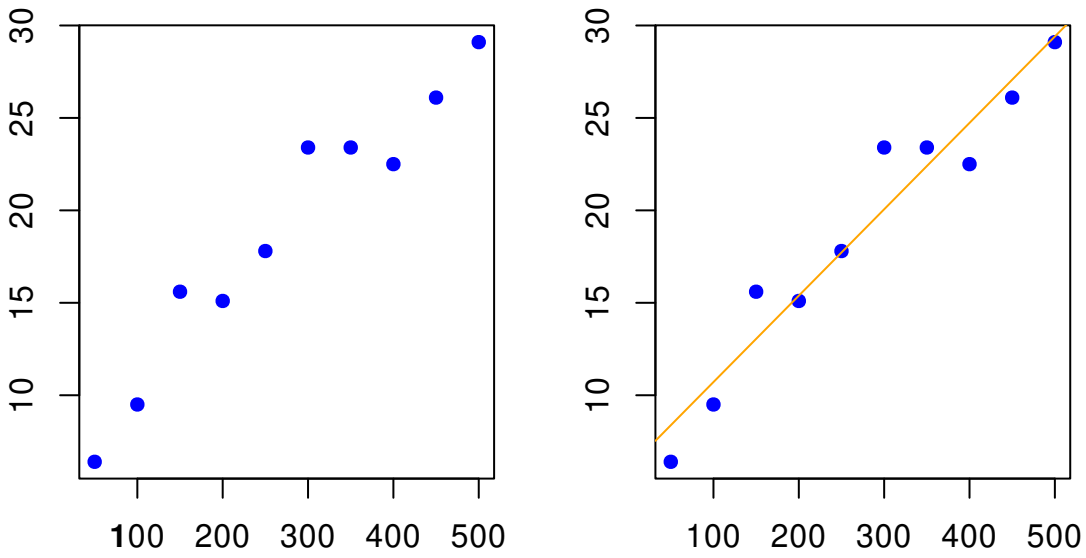


Abbildung 3.6. : Streudiagramm Seitenzahl - Buchpreis

Auf den ersten Blick scheint eine *Gerade* recht gut zu den Daten zu passen (siehe Abbildung 3.6) rechts.

Diese Gerade hat dann die Form

$$y = a + bx$$

wobei

- y: Buchpreis
- x: Seitenzahl
- Parameter a: Grundkosten des Verlags
- Parameter b: Kosten pro Seite

◀

Bemerkungen:

i. In der Mathematik wird die Geradengleichung meist in der Form

$$y = ax + b$$

geschrieben. Die Parameter a und b haben also gerade umgekehrte Bedeutung.

ii. Bezeichnungen in der Stochastik weichen noch oft von den „normalen" Bezeichnungen in der Mathematik ab.

Der Grund dafür liegt darin, dass die Stochastik bis in die 50er oder 60er Jahre des letzten Jahrhunderts *nicht* als Teil der Mathematik betrachtet wurde. Deswegen hielten sich die Stochastiker oft nicht an die Konventionen der Mathematiker. ♦

Die Frage ist nun, wie wir diese Gerade in Abbildung 3.6 rechts finden. Diese Gerade wird mit der sogenannten *Methode der kleinsten Quadrate* ermittelt.

3.3.2. Methode der kleinsten Quadrate

Versuchen wir mit einem Lineal eine Gerade durch *alle* Punkte in Abbildung 3.6 zu legen, so werden wir feststellen, dass das nicht möglich ist. Die Punkte folgen also nur *ungefähr* einer Geraden.

Da stellen sich uns zwei Fragen:

- Wie können wir eine Gerade finden, die *möglichst gut* zu allen Punkten passt?
- Was heisst „möglichst gut"?

Für alles Weitere brauchen wir einen neuen Begriff: das *Residuum*:

Residuum

Ein *Residuum* r_i ist die vertikale Differenz zwischen einem Datenpunkt (x_i, y_i) und dem Punkt $(x_i, a + bx_i)$ auf der gesuchten Geraden:

$$r_i = y_i - (a + bx_i) = y_i - a - bx_i$$

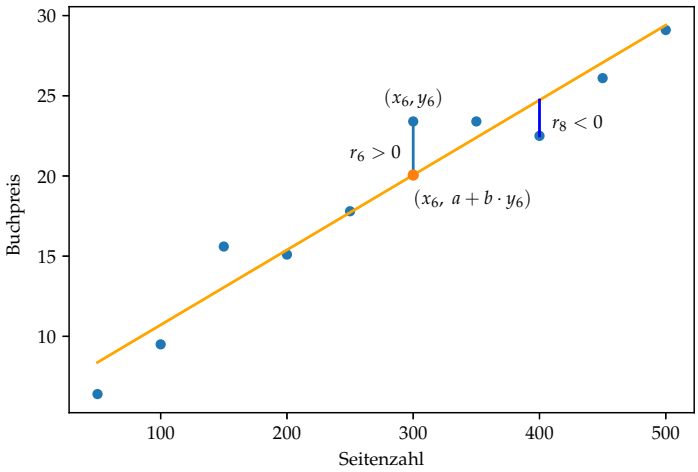


Abbildung 3.7. : Residuen für das Buchbeispiel

Beispiel 3.3.5

Für das Buchbeispiel sind die Residuen r_6 und r_8 für *diese* Gerade in Abbildung 3.7 dargestellt. Das Residuum r_6 ist positiv, da der Punkt oberhalb der Geraden liegt. Entsprechend ist $r_8 < 0$, da der Punkt unterhalb der Geraden liegt.

Wie wählen wir aber diese Gerade, damit sie „optimal" zur Punktwolke passt? Wir wollen hier nochmals auf die Ausgangssituation hinweisen: *Die Punktwolke ist gegeben und wir suchen die Gerade*.

Der erste Gedanke ist wohl, dass wir die Gerade wie folgt wählen: Wir zählen die vertikalen Differenzen zwischen Beobachtung und Gerade (siehe Abbildung 3.8) zusammen und gehen davon aus, dass eine kleine Summe der Differenzen (heisst, nahe bei 0) eine gute Anpassung bedeutet.

Das heisst, wir wählen die Gerade so, dass die Summe der roten Linien mit Vorzeichen so klein wie möglich wird. ◀

Allgemein

Wir möchten also die Gerade $y = a + bx$ so bestimmen, dass die Summe

$$r_1 + r_2 + \ldots + r_n = \sum_i r_i$$

minimal wird. Diese Methode hat aber eine gravierende Schwäche, wie folgendes Beispiel zeigt.

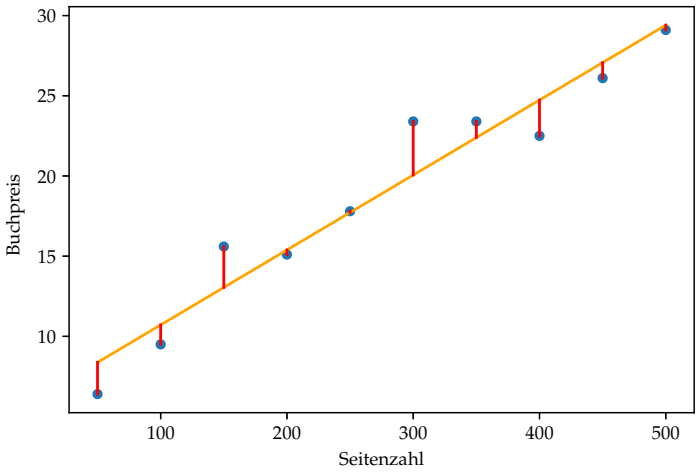


Abbildung 3.8. : Alle Residuen für das Buchbeispiel

Beispiel 3.3.6

Abbildung 3.9 links zeigt ein Beispiel einer „Punktwolke" mit einer Geraden, wo die Summe der Residuen 0 ist. Die positiven Abweichungen auf der rechten Seite (Pfeile nach oben) heben sich mit den negativen Abweichungen auf der linken Seite (Pfeile nach unten) auf.

Diese Gerade passt aber überhaupt nicht zu der Punktwolke, ist sogar jene, die am schlechtesten zu der Punktwolke passt. Trotzdem ist die Summe der Residuen 0.

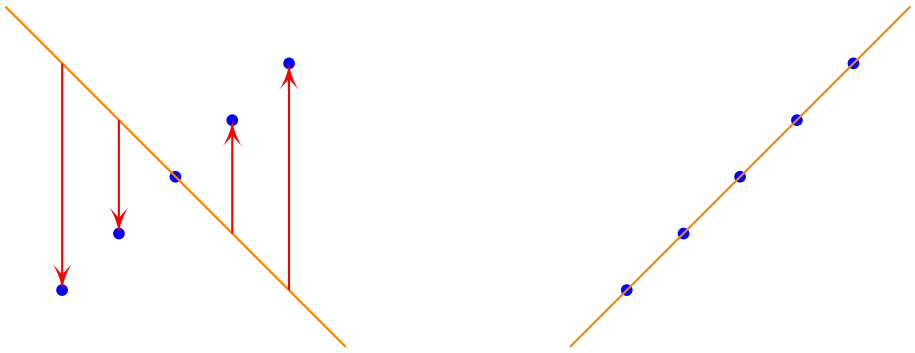


Abbildung 3.9. : Summe der Residuen ist 0

Natürlich können wir die Gerade auch wie in Abbildung 3.9 rechts einzeichnen. Auch hier ist die Summe der Residuen 0. Diese Gerade passt dann optimal zu der Punktwolke.

Aber welches ist die „richtige" Gerade, die am besten zu der Punktwolke passt? In beiden Fällen ist die Residuensumme 0. Hier scheint es klar zu sein, welches die „bes-

sere" Gerade ist, aber wir möchten ein Verfahren haben, das diese Gerade *eindeutig* festlegt. ◀

Wir müssen also das Vorzeichen der Abweichungen eliminieren, bevor wir addieren, damit sich die Beiträge in der Summe nicht aufheben können. Diesem Vorgehen sind wir schon bei der Definition der Streuung (siehe Beispiel 2.2.5) begegnet.

Eine Möglichkeit besteht darin, die Absolutbeträge der Abweichungen aufzusummieren, also

$$|r_1| + |r_2| + \ldots + |r_n| = \sum_i |r_i|$$

und diese Summe zu minimieren. Da es sich aber mit Absolutbeträgen nicht besonders bequem rechnen lässt, halten wir nach einer anderen Möglichkeit Ausschau.

Und wie schon bei der Definition der Varianz, summieren die *Quadrate der Abweichungen* auf:

$$r_1^2 + r_2^2 + \cdots + r_n^2 = \sum_i r_i^2$$

Die Parameter a und b werden so gewählt, dass diese Summe minimal wird. Eine Gerade passt (nach diesem Gütekriterium) also dann am besten zu den Punkten im Streudiagramm, wenn die Summe der Quadrate der vertikalen Abweichungen minimal ist.

Dieses Vorgehen ist unter dem Namen *Methode der kleinsten Quadrate*[4] bekannt.

Beispiel 3.3.7

In unserem Buchbeispiel erhalten wir mit **R** die Werte $a = 6.04$ und $b = 0.047$.

```
seitenzahl <- seq(50, 500, 50)

buchpreis <- c(6.4, 9.5, 15.6, 15.1, 17.8, 23.4, 23.4, 22.5, 26.1,
    29.1)

lm(buchpreis ~ seitenzahl)


##
## Call:
## lm(formula = buchpreis ~ seitenzahl)
##
## Coefficients:
## (Intercept)   seitenzahl
##     6.04000      0.04673
```

[4] Sie wurde unter anderen vom deutschen Mathematiker Carl Friedrich Gauss 1795 im Alter von 18 Jahren entwickelt.

Die Geradengleichung lautet

$$y = 6.04 + 0.047x$$

Bemerkungen:

i. Der Befehl `lm()` steht für „linear model".

ii. Mit dem Befehl `lm(y ~ x)` passt `R` ein Modell von der Form $y = a + bx$ an die Daten an.

iii. *Achtung:* Der Befehl für das Streudiagramm lautet `plot(x, y)`. Die Variablen verglichen zu dem `lm(...)` sind also vertauscht.

iv. Diese Gerade wird *Regressionsgerade* genannt. ♦

In Abbildung 3.10 ist nochmals das Streudiagramm mit die Regressionsgeraden dargestellt. Der Code befindet sich am Ende dieses Beispieles.

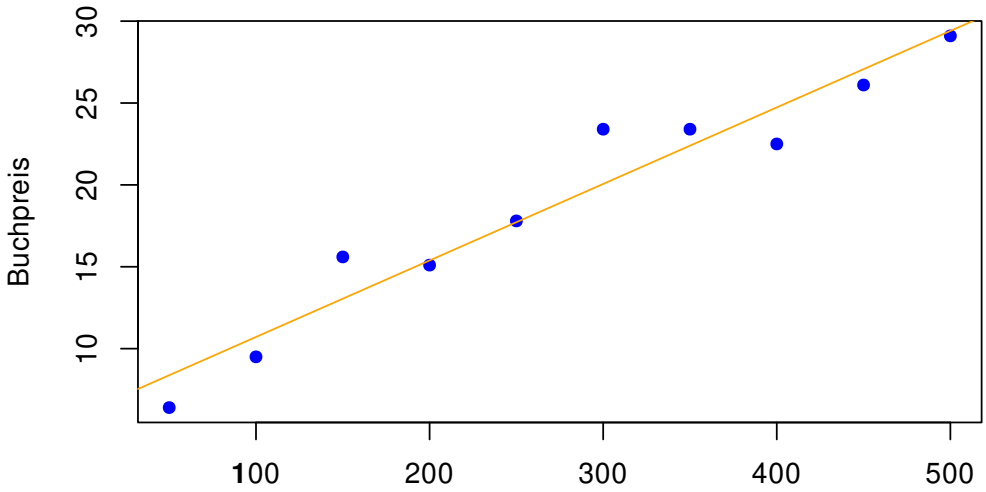


Abbildung 3.10. : Streudiagramm Seitenzahl - Buchpreis

Nun können wir die Fragen beantworten, die wir in Beispiel 3.3.4 gestellt haben:

- Der Parameter b entspricht dem Preis einer Seite. Pro Seite verlangt der Verlag rund 5 Rappen.

  Dieser Parameter entspricht geometrisch der *Steigung* der Regressionsgeraden in Abbildung 3.10.

- Der Parameter a entspricht dem Preis eines Buches mit 0 Seiten. Dieser Wert mit rund CHF 6 entspricht den Grundkosten des Verlags für jedes Buch.

  Der Parameter b entspricht geometrisch dem Schnittpunkt der Regressionsgeraden mit der vertikalen Achse (*y-Achsenabschnitt*) in Abbildung 3.10.

- Wie viel würde nach diesem Modell ein Buch mit 375 Seiten kosten?

  Dazu setzen wir $x = 375$ in die Geradengleichung oben ein und erhalten

  $$y = 6.04 + 0.04673 \cdot 375 \approx 23.60$$

  Das Buch dürfte also etwa CHF. 23.60 kosten.

Dieses Modell ist allerdings nur *begrenzt gültig*, vor allem müssen wir bei sogenannten *Extrapolationen* sehr vorsichtig sein.

Extrapolationen sind Vorhersagen der y-Werte (Preise) des Modelles , die für x-Werte (Seiten) gemacht werden, die *ausserhalb* des Bereichs (50 bis 500 Seiten) liegen, für die das Modell erstellt wurde .

Vor allem bei Extrapolationen, die sehr weit ausserhalb des gültigen Bereichs liegen, können die Vorhersagen sehr problematisch werden. Wir könnten ohne Probleme berechnen, wie viel ein Buch

- mit einer Million Seiten ($\approx$ CHF 50 000)
- mit -300 Seiten ($\approx$ CHF -9)

kostet. Aber diese Beträge entsprechen sicher nicht dem Preis von realen Büchern. Im zweiten Fall ist die Interpretation sogar sinnlos.

Unproblematischer sind hingegen *Interpolationen*, sofern das Modell vernünftig ist. Dies sind Vorhersagen für den Preis eines Buches, dessen Seitenzahl *innerhalb* des Bereiches liegt, für den das Modell erstellt wurde. Wir haben diesen Preis oben für die Seitenzahl 375 berechnet.

Diese Gerade in Abbildung 3.10 wird in **R** wie folgt erstellt:

```
seite <- c(seq(50, 500, 50))

preis <- c(6.4, 9.5, 15.6, 15.1, 17.8, 23.4, 23.4, 22.5, 26.1, 29.1)

plot(seite, preis, xlab = "Seitenzahl", ylab = "Buchpreis", pch = 16,
    col = "blue")

abline(lm(preis ~ seite), col = "orange")
```

Der Befehl zeichnet **abline(...)** zeichnet eine Gerade in eine schon gegebene Skizze. ◀

Der folgende Abschnitt ist für diejenigen gedacht, die genauer wissen wollen, wie die Parameter a und b bestimmt werden. Dies wird im Skript nicht mehr gebraucht und werden die Parameter a und b immer mit **R** berechnen.

Wie berechnet R die Parameter a und b? (Fakultativ!)

Die Parameter a und b werden wie folgt bestimmt:

Die Parameter a und b sollen den folgenden Ausdruck minimieren (Methode der Kleinsten-Quadrate)

$$\sum_{i=1}^{n}(y_i - (a + bx_i))^2$$

Die Lösung dieses Optimierungsproblems ergibt

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

$$\widehat{a} = \overline{y} - \widehat{b}\,\overline{x}$$

wobei $\overline{x}$ und $\overline{y}$ die entsprechenden Durchschnitte sind. $\widehat{a}$ und $\widehat{b}$ sind die Schätzer von den Parametern a und b, also die Werte, für welche $\sum_{i=1}^{n}(y_i - (a + bx_i))^2$ am kleinsten wird.

Wie man auf die Berechnung von a und b kommt, leiten wir hier nicht her. Nur soviel zur Idee: Da

$$\sum_i r_i^2 = \sum_{i=1}^{n}(y_i - (a + bx_i))^2$$

minimal werden muss, muss auch die Ableitung von $\sum_i r_i^2$ nach a *und* nach b gleich 0 sein. Wir erhalten also ein Gleichungssystem bestehend aus zwei Gleichungen und zwei Unbekannten:

$$\frac{\partial}{\partial a}\sum_{i=1}^{n}(y_i - (a + bx_i))^2 = \sum_{i=1}^{n} -2\,(y_i - a - bx_i) = 0$$

$$\frac{\partial}{\partial b}\sum_{i=1}^{n}(y_i - (a + bx_i))^2 = \sum_{i=1}^{n} -2\,(y_i - a - bx_i)\cdot x_i = 0$$

Die algebraischen Umformungen, die zu den Schätzern von a und b führen, sind dann etwas aufwendig und werden hier nicht aufgeführt. □

Beispiel 3.3.8

Wir kommen auf das Beispiel 3.2.2 des Old Faithful zurück und bestimmen die Regressionsgerade. In der Abbildung 3.11 ist noch die Regressionsgerade eingezeichnet.

Die Gleichung der Regressionsgeraden lautet

$$y = -1.4 + 0.07x$$

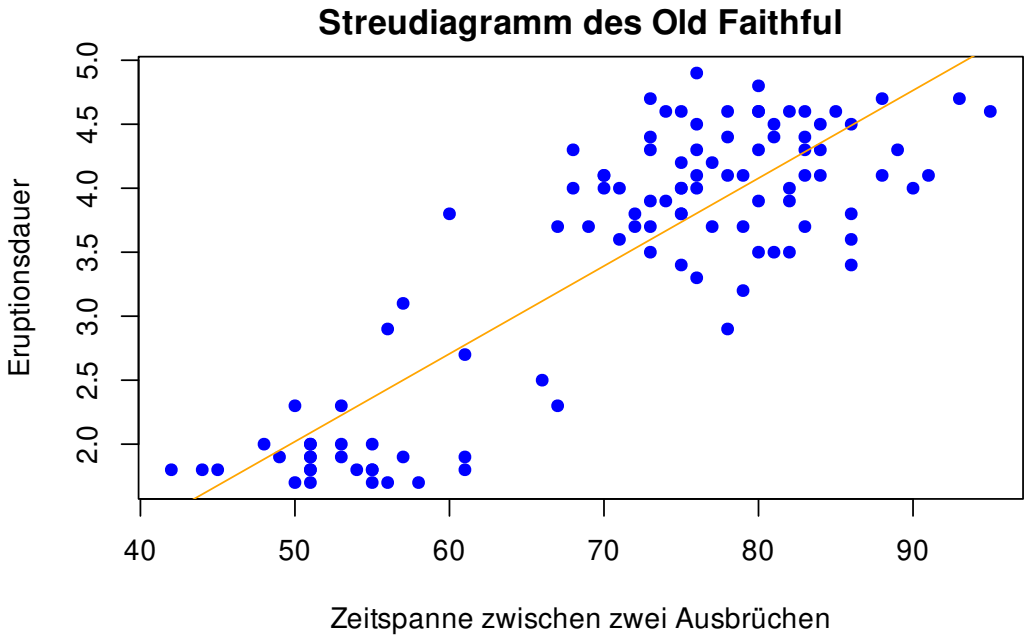


Abbildung 3.11. : Streudiagramm mit Regressionsgerade für die Eruptionsdauer und Zeitspanne zwischen zwei Ausbrüchen des Old Faithful

Was bedeuten nun diese Parameter?

- Der Wert -1.4 entspricht dem Schnittpunkt der Regressionsgeraden mit der vertikalen Achse (*intercept*). Das ist für $x = 0$ der Fall und entspricht demnach einer Eruptionsdauer, wenn die Ausbrüche 0 Minuten auseinanderliegen. Diese Interpretation macht natürlich keinen Sinn.
- Der Wert 0.07 entspricht der Steigung der Regressionsgeraden. Für jede zusätzliche Minute zwischen zwei Ausbrüchen nimmt die Eruptionsdauer um 0.07 Minuten zu.

Bemerkungen: R-Code

i. Plot:

```
geysir <- read.table("../../../Themen/Deskriptive_Statistik/Skript_de/Daten/geysir.

plot(geysir$Zeitspanne, geysir$Eruptionsdauer, xlab = "Zeitspanne zwischen zwei Aus
    ylab = "Eruptionsdauer", main = "Streudiagramm des Old Faithful",
    pch = 20, col = "blue")

abline(lm(geysir$Eruptionsdauer ~ geysir$Zeitspanne), col = "orange")
```

ii. Parameter der Regressionsgeraden:

```
lm(geysir$Eruptionsdauer ~ geysir$Zeitspanne)

##
## Call:
## lm(formula = geysir$Eruptionsdauer ~ geysir$Zeitspanne)
##
## Coefficients:
##       (Intercept)  geysir$Zeitspanne
##          -1.41029            0.06861
```

♦

◀

Beispiel 3.3.9

Es ist zu vermuten, dass es einen Zusammenhang zwischen der Körpergrösse der Väter und der Körpergrösse ihrer Söhne gibt. Der britische Statistiker Karl Pearson, einer der Väter der modernen Statistik, trug dazu um 1900 die Körpergrösse von 10 (in Wahrheit waren es 1078) zufällig ausgewählten Männern gegen die Grösse ihrer Väter auf. Dabei erhielt er die Daten von Tabelle 3.3.

Grösse des Vaters	152	157	163	165	168	170	173	178	183	188
Grösse des Sohnes	162	166	168	166	170	170	171	173	178	178

Tabelle 3.3. : Grössenvergleich von Vätern und Söhnen.

Es *scheint* hier tatsächlich einen Zusammenhang zu geben: je grösser der Vater, desto grösser der Sohn. Wenn wir noch das Streudiagramm aufzeichnen (siehe Abbildung 3.12), sehen wir, dass ein (möglicher) linearer Zusammenhang besteht.

Die Punktwolke „folgt" der Geraden

$$y = 0.445x + 94.7$$

wobei wir die Parameter mit der Methode der Kleinsten Quadrate aus den Daten mit **R** berechnet haben.

Wir interpretieren die Parameter:

- Der Wert 0.445 entspricht der Steigung der Regressionsgeraden. Bei jedem zusätzlichen Zentimeter Grösse des Vaters, nimmt die Grösse des Sohnes um etwa 0.45 cm zu.

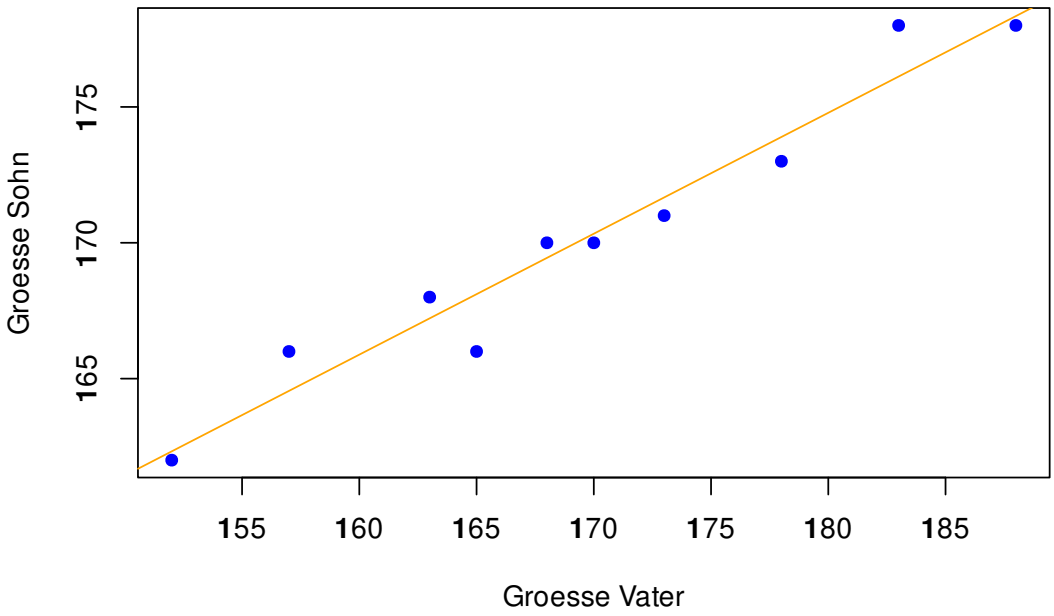


Abbildung 3.12. : Streudiagramm Körpergrössen Väter-Söhne.

- Der Wert 94.5 erhalten wir für $x = 0$. Was bedeutet dies aber? Wenn der Vater 0 cm gross ist, so wäre der Sohn gemäss dieses Modells ungefähr 95 cm gross, und das macht natürlich keinen Sinn.

Wir können also für die in der Tabelle 3.3 nicht vorkommende Grösse von 180 cm des Vaters den zu erwartenden Wert für die Grösse seines Sohnes berechnen.

$$y = 0.445 \cdot 180 + 94.7 \approx 175\,\text{cm}$$

Wir müssen bei dieser Formel allerdings wieder aufpassen, dass wir sie nicht dort anwenden, wo wir sie gar nicht dürfen. Wir können ohne weiteres die Körpergrösse des Sohnes für einen Vater mit 5 km Grösse berechnen. Ob das viel Sinn macht, sei dahingestellt. ◀

Beispiel 3.3.10

Die folgende Tabelle stellt einen Zusammenhang zwischen den Zahlen der Verkehrstoten her, die es 1988 und 1989 in zwölf Bezirken in den USA geben hat (die Bezirke weisen in etwa dieselbe Bevölkerungszahl auf).

Aus der Tabelle 3.4 ist kein offensichtlicher Zusammenhang ersichtlich.

Bezirk	1	2	3	4	5	6	7	8	9	10	11	12
Verkehrstote 1988	121	96	85	113	102	118	90	84	107	112	95	101
Verkehrstote 1989	104	91	101	110	117	108	96	102	114	96	88	106

Tabelle 3.4. : Verkehrstote in zwei aufeinanderfolgenden Jahren.

Betrachten wir das Streudiagramm in Abbildung 3.13, so sehen wir, dass kein Zusammenhang besteht. Dies war aber auch zu erwarten, wenn wir vernünftigerweise davon ausgehen können, dass es zwischen den Verkehrstoten der einzelnen Bezirke in einem Jahr keinen Zusammenhang gibt mit den Verkehrstoten im entsprechenden Bezirk im folgenden Jahr.

In Abbildung 3.13 ist die Regressionsgerade eingezeichnet. Diese können wir zwar berechnen und einzeichnen. Allerdings macht diese hier keinen Sinn, da es keinen linearen Zusammenhang zwischen den Messgrössen gibt.

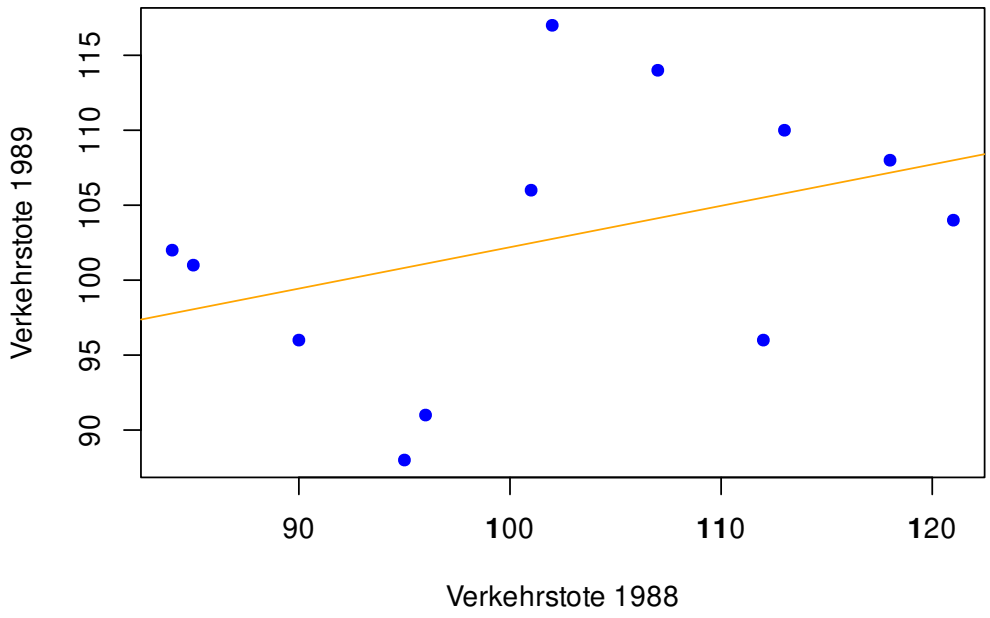


Abbildung 3.13. : Verkehrstote

Wir werden im nächsten Abschnitt 3.3.3 eine Grösse kennenlernen, mit der wir eine Aussage darüber machen können, wie stark der lineare Zusammenhang zwischen zwei Messgrössen ist. Oder anders gesagt: Diese Grösse, der Korrelationskoeffizient, beschreibt, wie gut die Regressionsgerade zur Punktwolke passt. ◀

Beispiel 3.3.11

Als weiteres Beispiel betrachten wir wieder die Erhebung, die den Zusammenhang zwischen Weinkonsum und der Sterblichkeit bezüglich Herz- und Kreislauferkrankungen untersucht (siehe Abbildung 3.14).

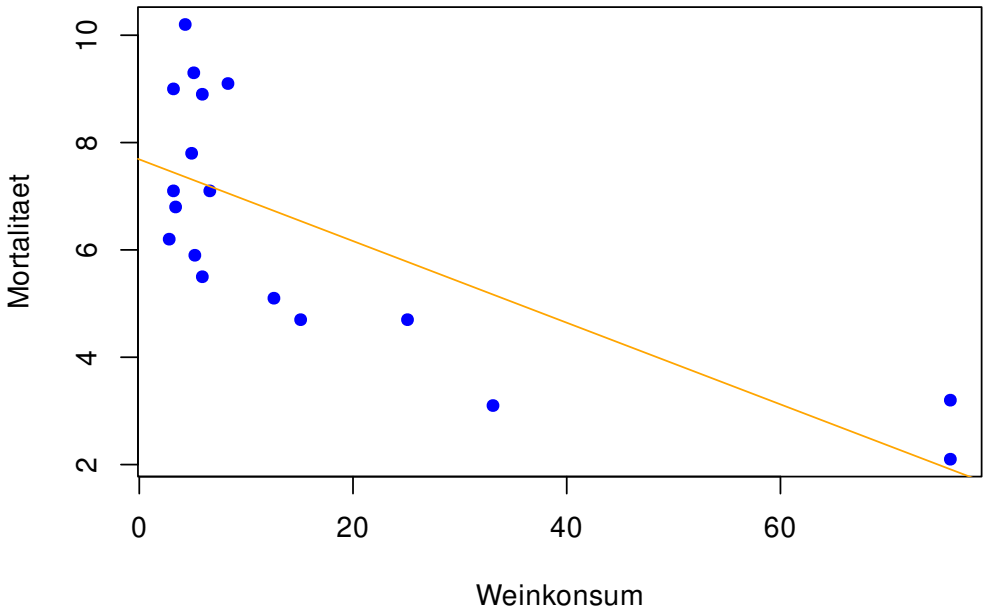


Abbildung 3.14. : Regressionsgerade Weinkonsum-Sterblichkeit

Wir legen den Daten ein lineares Modell zu Grunde

$$y = a + bx$$

wobei x den jährlichen Weinkonsum pro Person und y die Mortalität pro 1000 Personen bezeichnet. Ob dieses lineare Modell hier allerdings gerechfertigt ist, ist eine andere Frage.

Dann können wir aufgrund der Datenpunkte die Parameter a und b mit Hilfe der Methode der Kleinsten Quadrate mit **R** berechnen und erhalten die Regressionsgerade

$$y = 7.7 - 0.08x$$

Betrachten wir allerdings das Streudiagramm mit der Regressionsgerade (siehe Abbildung 3.14), so stellen wir fest, dass der Zusammenhang zwischen den Messgrössen *nicht* linear ist. Das Streudiagramm deutet eher auf eine, beispielsweise, Hyperbelfunktion hin.

Die Regressionsgerade sagt hier also wenig über den wahren Zusammenhang aus. ◀

Die Regressionsgerade können wir (fast[5]) immer bestimmen. In den letzten beiden Beispielen haben wir allerdings gesehen, dass die Regressionsgerade sehr wenig über die wirkliche Verteilung der Punkte im Streudiagramm aussagt. Dafür gibt es zwei Gründe:

- Die Punkte folgen scheinbar gar *keiner* Gesetzmässigkeit.
- Die Punkte folgen einer *nichtlinearen* Gesetzmässigkeit.

Wie können wir nun aber feststellen, ob ein linearer Zusammenhang der Daten besteht oder nicht? Eine Möglichkeit besteht sicher darin, die Daten graphisch zu betrachten, wie wir das eben gemacht haben.

Wir können aber auch einen Wert angeben, der den Zusammenhang numerisch beschreibt.

3.3.3. Empirische Korrelation

Für die quantitative Zusammenfassung der *linearen* Abhängigkeit von zwei Grössen ist die *empirische Korrelation* r als Kennzahl (oder auch mit $\widehat{\rho}$ bezeichnet) am gebräuchlichsten. Die Definition sieht wie folgt aus (sollte man einmal gesehen haben). Den Wert selbst berechnen wir wie üblich mit R.

Empirische Korrelation

$$r = \frac{(x_1-\overline{x})(y_1-\overline{y}) + \ldots + (x_n-\overline{x})(y_n-\overline{y})}{\sqrt{(x_1-\overline{x})^2+\ldots+(x_n-\overline{x})^2}\cdot\sqrt{(y_1-\overline{y})^2+\ldots+(y_n-\overline{y})^2}}$$

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

Die empirische Korrelation ist eine dimensionslose Zahl zwischen -1 und $+1$ und misst die Stärke und die Richtung der *linearen Abhängigkeit* zwischen den Daten x und y. Die empirische Korrelation hat folgende Eigenschaften:

Eigenschaften der Korrelation

1. Ist $r = +1$, dann liegen die Punkte auf einer steigenden Geraden ($y = a + bx$ mit $b > 0$) und umgekehrt.

[5] Wenn die Punkte *exakt* auf einer vertikalen Linie liegen, kann die Regressionsgerade mit unseren Mitteln nicht bestimmt werden. Aber dieser Fall praktisch keine Bedeutung.

2. Ist $r = -1$, dann liegen die Punkte auf einer fallenden Geraden ($y = a + bx$ mit $b < 0$) und umgekehrt.
3. Sind x und y unabhängig (d.h. es besteht kein Zusammenhang), so ist $r = 0$.

   Die Umkehrung gilt im Allgemeinen nicht: Ist $r = 0$, so heisst dies *nicht*, dass x und y unabhängig voneinander sind (siehe Abbildung 3.19).

Wichtig: Der Korrelationskoeffizient misst nur den linearen *Zusammenhang!*

Für diejenigen, die es ein bisschen genauer wissen wollen:

Wir werden diese Eigenschaften nicht herleiten, sondern mit Graphiken etwas veranschaulichen. Dazu betrachten wir nur den Zähler der Korrelation

$$(x_1 - \overline{x})(y_1 - \overline{y}) + \ldots + (x_n - \overline{x})(y_n - \overline{y}) = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

Wir werden zeigen, dass diese Grösse:

- grösser als 0 ist, wenn die Punktwolke einer steigenden Geraden folgt
- kleiner als 0 ist, wenn die Punktwolke einer fallenden Geraden folgt
- 0 ist, wenn kein Zusammenhang vorhanden ist
- 0 sein kann, wenn kein linearer Zusammenhang vorhanden ist

Der Nenner der Korrelation ist positiv und führt dazu, dass die Korrelation normalisiert ist, das heisst die Werte liegen zwischen -1 und 1

Beispiel 3.3.12

In Abbildung 3.15 sind einige Datenpunkte eingezeichnet, die mehr oder weniger einer Geraden folgen. Es wurden noch die zu den Koordinatenachsen parallelen Geraden $\overline{x}$ und $\overline{y}$ eingezeichnet.

Im Zähler der Korrelation

$$(x_1 - \overline{x})(y_1 - \overline{y}) + \ldots + (x_n - \overline{x})(y_n - \overline{y})$$

werden von den x-Koordinaten der Punkte der Durchschnitt $\overline{x}$ und von den y Koordinaten $\overline{y}$ subtrahiert. Das heisst, der Ursprung der Abbildung 3.15 wird in den Punkt $(\overline{x}, \overline{y})$ verschoben. Wir erhalten dann Abbildung 3.16.

Der Zähler der Korrelation sieht dann wie folgt aus:

$$x_1^* y_1^* + x_2^* y_2^* + \ldots + x_n^* y_n^*$$

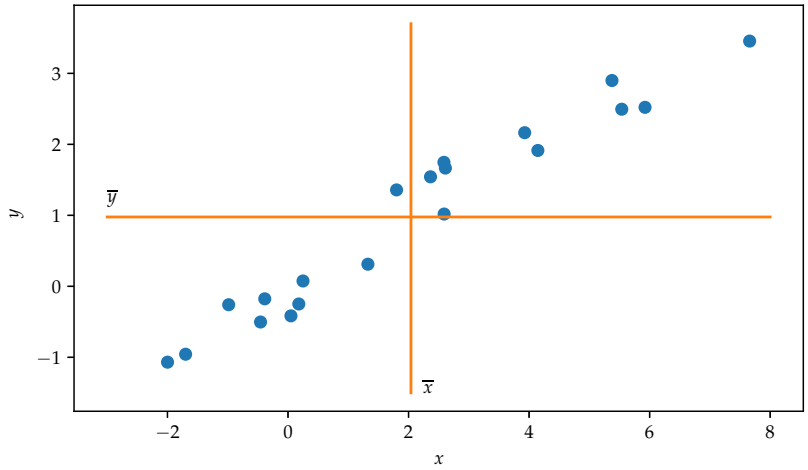


Abbildung 3.15. : Punkte, die fast auf einer Geraden liegen.

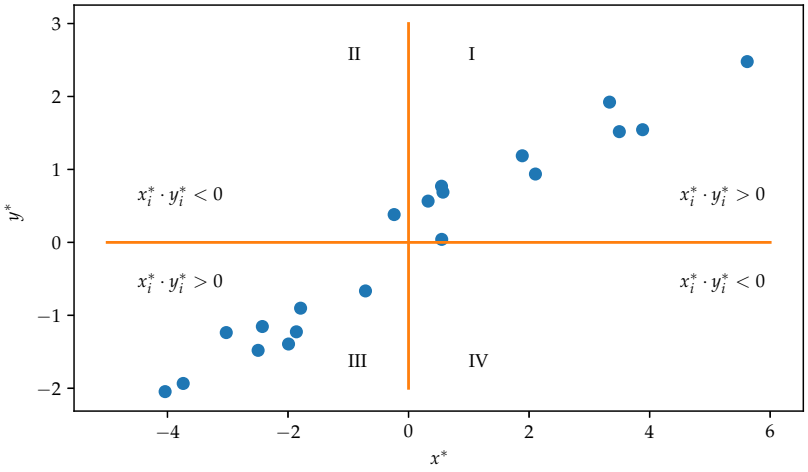


Abbildung 3.16. : Neuer Ursprung für Punkte, die fast auf einer Geraden liegen.

Nun, was bedeutet dies für die Punkte in Abbildung 3.16?

- Die Punkte im I. Quadranten (rechts oben) haben positive Koordinaten und somit ist auch $x_i^* y_i^*$ positiv.
- Die Punkte im II. Quadranten (links oben) haben eine negative x-Koordinate und eine positive y-Koordinate. Somit ist $x_i^* y_i^*$ *negativ*.
- Die Punkte im III. Quadranten (links unten) haben eine negative x-Koordinate und eine negative y-Koordinate. Somit ist $x_i^* y_i^*$ *positiv*.
- Die Punkte im IV. Quadranten (rechts unten) haben eine positive x-Koordinate und eine negative y-Koordinate. Somit ist $x_i^* y_i^*$ *negativ*.

In Abbildung 3.16 sind fast alle Punkte im I. und III. Quadranten, wo $x_i^* y_i^* > 0$ und somit ist auch

$$x_1^* y_1^* + x_2^* y_2^* + \ldots + x_n^* y_n^* > 0$$

Folgen die Punkte einer fallenden Geraden, so sind fast alle Punkte im II. und IV. Quadranten, wo $x_i^* y_i^* < 0$ und somit ist auch

$$x_1^* y_1^* + x_2^* y_2^* + \ldots + x_n^* y_n^* < 0$$

Was passiert nun, wenn die Punkte keinen Zusammenhang aufweisen (siehe Abbildung 3.17?

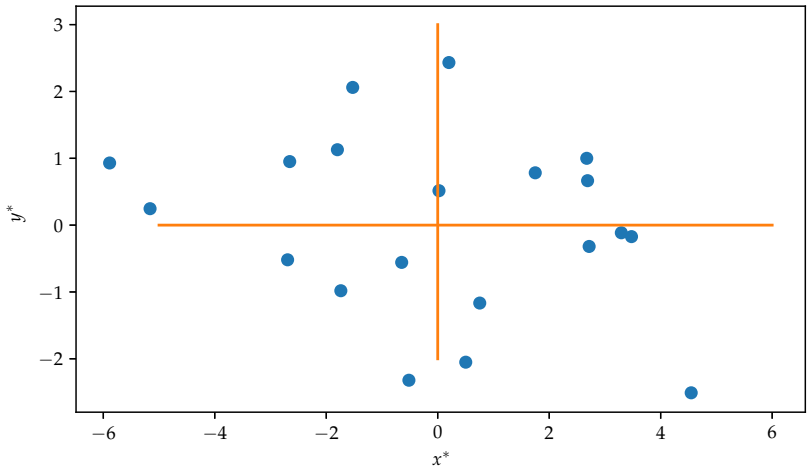


Abbildung 3.17. : Von den Koordinaten wurden die jeweiligen Mittelwerte subtrahiert.

In diesem Fall heben sich die Produkte $x_i^* y_i^*$ über alle Punkte aufaddiert in etwa auf, da die Hälfte aller Punkte im I. und III. Quadranten (Produkte positiv) und die andere Hälfte im II. und IV. Quadranten (Produkte negativ) liegen. Zudem sind die Produkte betragsmässig ähnlich. Damit gilt

$$x_1^* y_1^* + x_2^* y_2^* + \ldots + x_n^* y_n^* \approx 0$$

Wie sieht es aus, wenn wir einen *quadratischen* Zusammenhang haben (siehe Abbildung 3.18)?

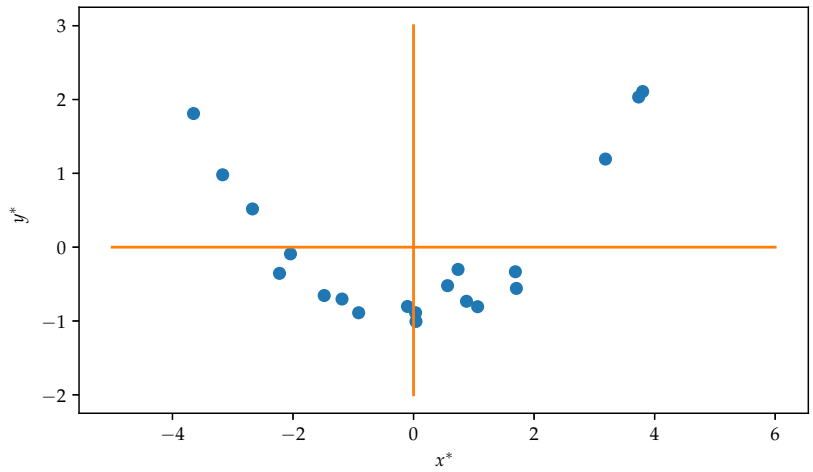


Abbildung 3.18. : Punkte, die fast auf einer Geraden liegen.

Auch hier heben sich die Beträge der Produkte links und rechts von der y-Achse auf. Es gilt

$$x_1^* y_1^* + x_2^* y_2^* + \ldots + x_n^* y_n^* \approx 0$$

Der Korrelationskoeffizient erkennt also nur *lineare* Zusammenhänge. ◀

□

Wir sollten jedoch nie nur einfach r berechnen, ohne einen Blick auf das Streudiagramm zu werfen, da ganz verschiedene Strukturen den gleichen Wert von r ergeben können. Siehe dazu Abbildung 3.19.

Beispiel 3.3.13

Für unser Seitenzahl-Preis-Beispiel 3.3.7 erhalten wir mit **R** mit dem Befehl **cor(...)**:

```
cor(seitenzahl, buchpreis)


## [1] 0.9681122
```

Der Wert liegt also sehr nahe bei 1 und somit besteht ein enger linearer Zusammenhang. Dazu ist der Wert positiv, was einem „je mehr, desto mehr", also einem positiven linearen Zusammenhang entspricht. ◀

Beispiel 3.3.14

Auch im Beispiel 3.3.9 der Körpergrösse von Vater und Sohn erwarten wir einen hohen Korrelationskoeffizienten. Wir erhalten 0.973. ◀

Beispiel 3.3.15

Bei den Verkehrsunfällen in Beispiel 3.3.10 haben wir keinen Zusammenhang und erwarten einen Korrelationskoeffizienten nahe null. Er beträgt 0.386. ◀

Beispiel 3.3.16

Im Beispiel 3.3.11 beim Weinkonsum erwarten wir einen negativen Korrelationskoeffizienten, da mit steigendem Weinkonsum die Mortalität sinkt und der nahe bei null liegt.

Er beträgt -0.746. Ohne die Daten in einem Streudiagramm darzustellen, würde man aufgrund dieses Wertes fälschlicherweise auf einen starken negativen linearen Zusammenhang schliessen. ◀

Beispiel 3.3.17

In Abbildung 3.19 sind 21 verschiedene Datensätze dargestellt, die je aus gleich vielen Beobachtungspaaren (x_i, y_i) mit den entsprechenden Punkten im Streudiagramm bestehen. Über jedem Datensatz steht jeweils die zugehörige empirische Korrelation.

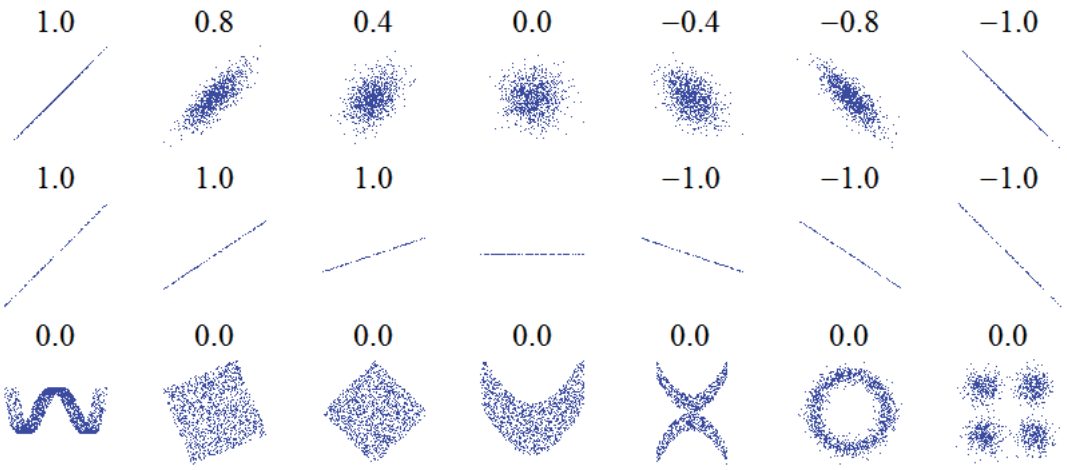


Abbildung 3.19. : 21 verschiedene Datensätze und deren empirische Korrelationskoeffizienten.

Bei perfektem linearen Zusammenhang ist die empirische Korrelation $+1$ oder -1 (je nachdem ob die Steigung positiv oder negativ; siehe zweite Zeile in Abbildung 3.19). Je mehr die Punkte um den linearen Zusammenhang streuen, desto kleiner wird der Betrag der empirischen Korrelation (siehe erste Zeile).

Da die empirische Korrelation nur den *linearen* Zusammenhang misst, kann es einen (nichtlinearen) Zusammenhang zwischen den beiden Variablen x und y geben, auch wenn die empirische Korrelation null ist (siehe unterste Zeile in Abbildung 3.19).

◀

Beispiel 3.3.18 Anscombe-Plot

In Abbildung 3.20 sehen wir den sogenannten *Anscombe*-Plot.

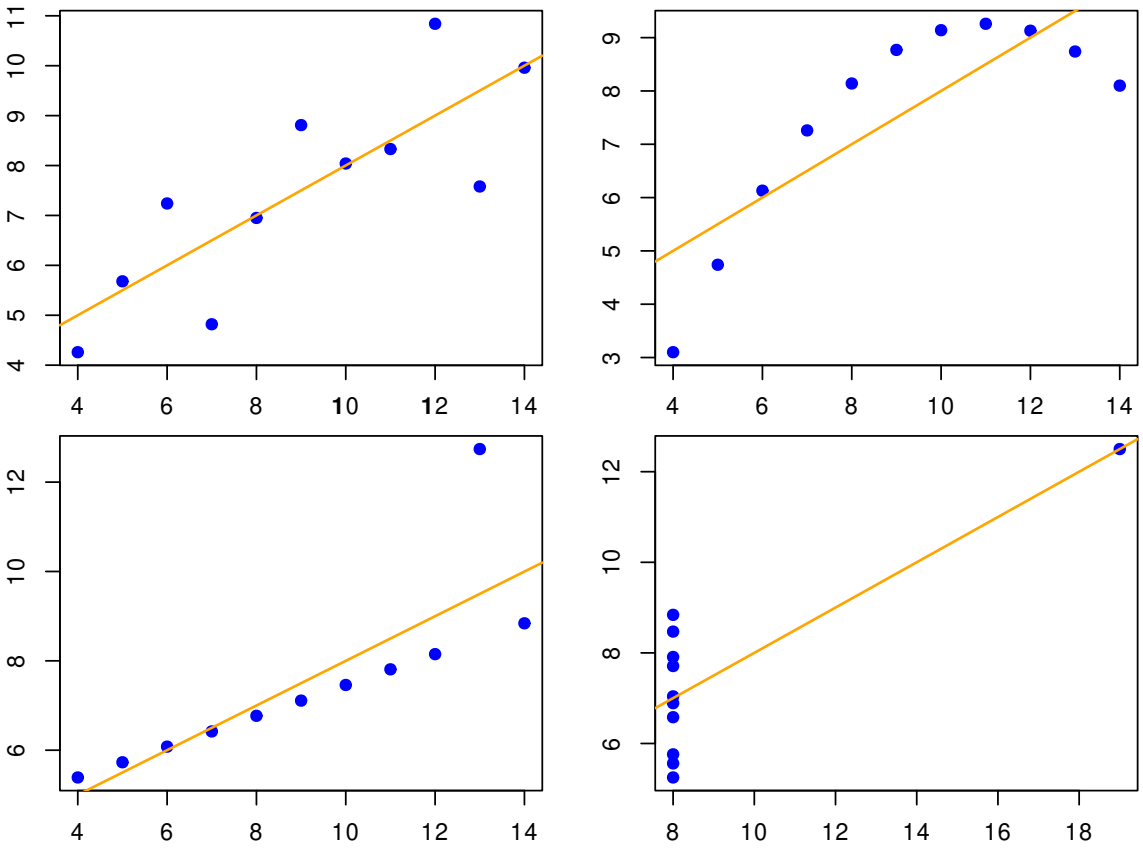


Abbildung 3.20. : Anscombe-Plot

Obwohl die Datensätze sehr unterschiedlich sind, haben sie sehr ähnlichen Korrelationskoeffizienten und Regressiongeraden:

- Das Streudiagramm links oben zeigt einen „normalen" Datensatz.

- Das Streudiagramm rechts oben zeigt einen nichtlinearen, quadratischen Zusammenhang.
- Das Streudiagramm links unten hat einen Ausreisser, alle andern Punkte liegen auf einer Geraden.
- Das Streudiagramm rechts unten ist ein sehr unwahrscheinlicher Datensatz.

Die Korrelationskoeffizienten sind bis auf die dritte Stelle nach dem Komma gleich, nämlich 0.816.

```
## [1] 0.8164205
## [1] 0.8162365
## [1] 0.8162867
## [1] 0.8165214
```

Auch die Regressionsgeraden sind praktisch identisch:

```
## (Intercept) anscombe$x1
##   3.0000909   0.5000909
## (Intercept) anscombe$x2
##    3.000909    0.500000
## (Intercept) anscombe$x3
##   3.0024545   0.4997273
## (Intercept) anscombe$x4
##   3.0017273   0.4999091
```

Diese Beispiel zeigt, dass wir uns nie nur auf Kennzahlen verlassen dürfen. Wir müssen *immer* zusätzlich die zugehörigen Plots auch noch betrachten.

Von John Tukey (Statistiker):

> *The greatest value of a picture is when it forces us to notice what we never expected to see.*

◀

Bemerkungen:

i. Wie „nahe" nahe bei 0 ist, kann man allgemein nicht sagen und ist von der Problemstellung abhängig.

ii. Zusätzlich haben verschiedene Wissenschaftsgebiete sehr unterschiedliche Vorstellungen, wann ein Regressionskoeffizient gross, also nahe bei 1, ist. In der Soziologie ist das wesentlich tiefer als beispielsweise in der Physik. ♦

Kapitel 4.

Wahrscheinlichkeit

> Everybody speaks of probability, but no one is able to say what it is, in a way which is satisfactory for others.
>
> *(Garrett Birkhoff)*

4.1. Einführung

Wir alle haben ein intuitives Gefühl, was Wahrscheinlichkeit ist. Die Wahrscheinlichkeit mit einem fairen Würfel eine 4 zu würfeln ist ein Sechstel.

Allerdings ist die exakte Definition und vor allem die Interpretation der Wahrscheinlichkeit überraschend schwierig.

Schon bei der Aussage „Es regnet morgen mit einer Wahrscheinlichkeit von 80 %" ist es alles andere als klar, was damit gemeint ist (siehe auch Schlussbemerkung 4.6).

Wir werden in diesem Kapitel die wichtigsten Begriffe der Wahrscheinlichkeitsrechnung kennenlernen. Dies ist vielleicht das theoretischste Kapitel des gesamten Moduls. Es enthält allerdings viele zentrale Begriffe der Wahrscheinlichkeitsrechnung, die immer wieder vorkommen.

4.2. Wahrscheinlichkeitsmodelle

4.2.1. Einleitung: Modelle vs. Realität

Bevor wir Wahrscheinlichkeits*modelle* definieren, wollen wir, wie in Beispiel 3.3.1, zuerst darauf eingehen, *warum* wir modellieren.

Wir verwenden oft Modelle, ohne dass uns dies bewusst ist. Ganz am Anfang der Einführung 4.1 haben wir erwähnt, dass die Wahrscheinlichkeit mit einem fairen Würfel eine 4 zu würfeln ein Sechstel ist.

Stimmt das aber überhaupt? Ja und nein. Schon diese sehr einfache Aussage beruht auf einem *Modell*. Wir gehen von der *Annahme* aus, dass der Würfel fair ist. Das heisst, der Würfel ist völlig symmetrisch und damit ist die Wahrscheinlichkeit eine bestimmte Zahl zu würfeln für alle Seiten gleich. Da wir 6 Zahlen haben, ist diese Wahrscheinlichkeit eben ein Sechstel.

Die eben gemachte Überlegung ist aber ein *Modell*. Real gibt es *keinen* Würfel, der absolut symmetrisch ist. Das heisst, die Wahrscheinlichkeit eine bestimmte Zahl zu würfeln ist *nie exakt* ein Sechstel (0.166666 . . .), sondern nur sehr nahe bei einem Sechstel sein, etwa 0.1666666521398 . . ..

Da wir nicht alle diese fast fairen Würfel getrennt untersuchen können, machen wir eben ein Modell, dass *alle* diese Würfel so gut wie möglich beschreibt.

Die Hoffnung oder Erwartung ist immer, dass das Modell die Realität so gut wie möglich beschreibt.

Mit dem Modell können wir dann natürlich auch untersuchen, ob ein Würfel fair ist. Würfeln wir sehr oft und die Zahl 2 kommt mit einer Wahrscheinlichkeit von 0.1 vor, so können wir annehmen, dass der Würfel nicht fair ist.

4.2.2. Definition Wahrscheinlichkeitsmodelle

Wir betrachten *Zufallsexperimente*, bei denen der Ausgang *nicht exakt* vorhersagbar ist. Beispiele dafür sind

- Anzahl Anrufe in einer Stunde in einem Callcenter
- geworfene Augenzahl bei einem Würfelwurf

Ein *Wahrscheinlichkeitsmodell* beschreibt grob, welche Ergebnisse in einem solchen Experiment möglich sind. Das Wahrscheinlichkeitsmodell beinhaltet zudem die Wahrscheinlichkeiten, mit denen die verschiedenen Ergebnisse eintreten können.

Beispiel 4.2.1

Beim Würfel sind die möglichen Ergebnisse $1, 2, 3, 4, 5, 6$ und die Wahrscheinlichkeit eine dieser Zahlen zu werfen, beträgt $1/6$, sofern der Würfel fair ist. ◀

Ein Wahrscheinlichkeitsmodell erlaubt uns dann, gewisse Vorhersagen zu machen, die wir experimentell überprüfen können. Damit können wir uns zum Beispiel bei Glücksspielen eine gute Spielstrategie erarbeiten.

Wir *definieren* nun, was ein Wahrscheinlichkeitsmodell ist.

Wahrscheinlichkeitsmodell

Ein Wahrscheinlichkeitsmodell hat die folgenden Komponenten:

- *Grundraum* Ω, der aus den *Elementarereignissen* ω besteht

  Elementarereignisse sind mögliche Ergebnisse oder Ausgänge des Experiments, die alle zusammen den Grundraum bilden:

$$\Omega = \{\underbrace{\text{mögliche Elementarereignisse } \omega}_{\text{mögliche Ausgänge/Resultate}}\}$$

- *Ereignisse* $A, B, C, \ldots$: Teilmengen von Ω
- *Wahrscheinlichkeiten* P, die zu den Ereignissen $A, B, C, \ldots$ gehören

Wir werden diese Begriffe mit Hilfe von Beispielen näher kennenlernen.

4.2.3. Grundraum, Elementarereignisse

Bei der Durchführung eines Experiments wird aus der Menge aller Elementarereignisse (Grundraum) *ein* Elementarereignis *zufällig* ausgewählt.

Beispiel 4.2.2

Beim Würfelwurf ist der Grundraum gegeben durch die möglichen Ergebnisse des Zufallsexperiments

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

Das Element $\omega = 2$ ist ein Elementarereignis. Der Ausdruck $\omega = 2$ hat die Bedeutung, dass beim Würfeln die Zahl 2 geworfen wurde.

Die Zahl 7 ist kein Elementarereignis, da sie nicht zum Grundraum Ω dieses Zufallsexperimentes gehört. ◀

Beispiel 4.2.3

Bei der Anzahl Anrufen während einer Stunde in einem Callcenter ist der Grundraum gegeben durch

$$\Omega = \{0, 1, 2, 3, 4, \dots\}$$

Zumindest theoretisch können in einer Stunde beliebig viele Anrufe eintreffen. Das Elementarereignis $\omega = 6$ bedeutet, dass in einer Stunde 6 Anrufe angekommen sind.

◀

Beispiel 4.2.4 2-maliges Werfen einer Münze

Wir wählen beim Werfen einer Münze die Bezeichnungen K für „Kopf" und Z für „Zahl" (dies gilt für das gesamte Kapitel).

Werfen wir eine Münze zweimal nacheinander, dann beschreibt das Elementarereignis ZK, dass *zuerst* Zahl und *dann* Kopf geworfen wurde.

Alle möglichen Ergebnisse beim 2-maligen Werfen einer Münze sind dann gegeben durch

$$\Omega = \{KK, KZ, ZK, ZZ\}$$

wobei Ω wieder den Grundraum bezeichnet. Ein Elementarereignis ist z.B. $\omega = KZ$.

◀

4.2.4. Ereignis

Ereignisse sind allgemeiner und wichtiger als Elementarereignisse, bestehen aber aus diesen.

Unter einem *Ereignis* A versteht man eine Teilmenge von Ω:

$$A \subset \Omega$$

Das Zeichen $\subset$ bedeutet „... ist Teilmenge von...".

„Ein Ereignis A tritt ein" bedeutet, dass das Ergebnis ω des Experiments zu A gehört.

Beispiel 4.2.5 2-maliges Werfen einer Münze

Wir werfen wie in Beispiel 4.2.4 eine Münze zweimal nacheinander und betrachten das Ereignis A, bei welchem genau einmal K geworfen wird.

Dieses Ereignis besteht aus den Elementarereignissen KZ und ZK. Das Ereignis A ist dann durch die Menge

$$A = \{KZ, ZK\}$$

gegeben.

Die Menge A ist deshalb ein Ereignis, da sie ein „Teil" der Grundmenge

$$\Omega = \{KK, KZ, ZK, ZZ\}$$

ist.

Werfen wir ZK, so tritt das Ereignis A ein. Werfen wir ZZ, so trifft das Ereignis A *nicht* ein, da ZZ nicht zur Menge A gehört.

Wir können nun diesem Ereignis A noch eine Wahrscheinlichkeit $P(A)$ zuordnen, die die Wahrscheinlichkeit angibt, mit der das Ereignis A eintritt.

Diese Wahrscheinlichkeit ist einfach zu berechnen, sofern die Münze *fair* ist (Z und K sind gleichwahrscheinlich): das Ereignis A hat zwei und der Grundraum Ω 4 Elemente. Also ist die Wahrscheinlichkeit, dass das Ereignis A eintritt (siehe auch Beispiel 4.2.9)

$$P(A) = \frac{2}{4} = \frac{1}{2}$$

In der Stochastik werden Wahrscheinlichkeiten oft mit P oder p bezeichnet.

Die hier berechnete Wahrscheinlichkeit ist eine sogenannte *Laplace-Wahrscheinlichkeit* (siehe Abschnitt 4.2.8). ◀

Beispiel 4.2.6 Würfeln

1. Das Ereignis A bezeichnet „eine ungerade Augenzahl würfeln". Dann ist

   $$A = \{1, 3, 5\}$$

   Das Ereignis A tritt ein, wenn z.B. die Zahl 5 gewürfelt wurde.

2. Wir bezeichnen mit B das Ereignis, eine Zahl kleiner als 7 zu würfeln. Das ist natürlich immer der Fall und somit ist in diesem Fall

$$B = \{1, 2, 3, 4, 5, 6\} = \Omega$$

   Wir sprechen von einem *sicheren* Ereignis.

3. Weiter sei C das Ereignis „die Zahl sieben zu würfeln". Dies ist unmöglich und wir schreiben

$$C = \{\}$$

   Dabei stellt $\{\}$ die leere Menge dar, die kein Element enthält. Wir sprechen in diesem Fall von einem *unmöglichen* Ereignis.

Bemerkungen:

i. Die Grundmenge Ω selbst ist also auch ein Ereignis

$$\Omega \subset \Omega$$

ii. Die leere Menge $\{\}$ ist also auch ein Ereignis

$$\{\} \subset \Omega$$

   Die leere Menge scheint ein sehr uninteressantes Ereignis zu sein. Sie spielt aber eine wichtige Rolle in der Stochastik (und nicht nur dort).

iii. Manchmal wird für die leere Menge auch die Bezeichnung $\emptyset$ verwendet. Wir werden diese Bezeichnung nicht verwenden. ♦

Ist der Würfel fair (alle Seiten haben die gleiche Wurfwahrscheinlichkeit), so können wir den obigen Ereignissen A, B und C wieder sehr einfach die entsprechende Wahrscheinlichkeit zuordnen.

1. Die Wahrscheinlichkeit $P(A)$ für das Eintreten des Ereignisses A eine ungerade Zahl zu würfeln, ist

$$P(A) = \frac{3}{6} = \frac{1}{2}$$

   Wir dividieren wieder die Anzahl Elemente von A, nämlich 3, durch die Anzahl Elemente von Ω, nämlich 6.

2. Die Wahrscheinlichkeit $P(B)$ für das Ereignis des Ereignisses B eine Zahl kleiner als 7 zu würfeln, ist

$$P(B) = \frac{6}{6} = 1$$

Wir dividieren wieder die Anzahl Elemente von B, nämlich 6, durch die Anzahl Elemente von Ω, nämlich 6.

Dies gilt allgemein (nicht nur für faire Würfel): Die Wahrscheinlichkeit, dass ein sicheres Ereignis eintritt, ist immer 1.

3. Die Wahrscheinlichkeit $P(C)$ für das Eintreten des Ereignisses C die Zahl 7 zu würfeln, ist
$$P(C) = \frac{0}{6} = 0$$
Allgemein: Die Wahrscheinlichkeit, dass ein unmögliches Ereignis eintritt, ist immer 0.

Die oben erfolgten Berechnungen der Wahrscheinlichkeiten sind nur erlaubt, falls der Würfel fair ist.

Falls der Würfel nicht fair ist, so müssen wir allgemeiner Vorgehen (siehe Beispiel 4.2.10). ◀

4.2.5. Neue Ereignisse aus schon bekannten

Für den Umgang mit Ereignissen ist es nützlich, sich die Operationen der Mengenlehre und deren Bedeutung in Erinnerung zu rufen.

Die Operationen der Mengenlehre (Vereinigung, Durchschnitt, Komplement) werden verwendet, um aus bereits definierten Ereignissen neue Ereignisse zu gewinnen (siehe Tabelle 4.1).

Name	Symbol	Bedeutung
Vereinigung	$A \cup B$	*A oder B*
Durchschnitt	$A \cap B$	*A und B*
Komplement	$\overline{A}$	*nicht A*
Differenz	$A \backslash B = A \cap \overline{B}$	*A ohne B*

Tabelle 4.1. : Operationen der Mengenlehre

Bemerkungen:

i. Das „oder" der Mengenoperation *Vereinigung* ist nicht exklusiv: Ein Element kann also in A *und* in B liegen.

Umgangssprachlich wird „oder" meist in der Bedeutung „entweder ... oder ..." gedeutet. Das ist beim „oder" der Mengenoperation Vereinigung nicht so.

ii. Anstelle des Begriffs *Durchschnitt* verwendet man oft auch den Begriff „Schnittmenge".

iii. Für das Komplement $\overline{A}$ wird auch die Bezeichnung A^c verwendet, die wir hier aber nicht gebrauchen werden. ♦

Venn-Diagramme

Alle diese Begriffe in Tabelle 4.1 lassen sich einfach mit sogenannten *Venn-Diagrammen* illustrieren (siehe Abbildung 4.1).

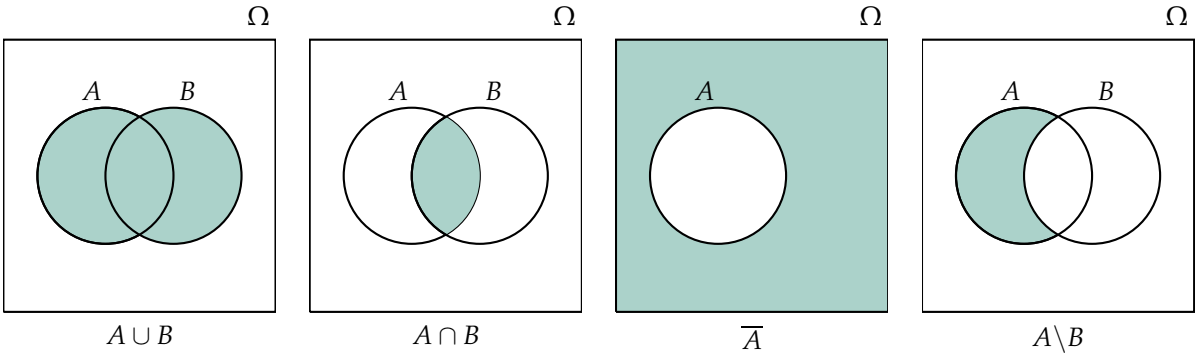


Abbildung 4.1. : *Links:* $A \cup B$, *Mitte links:* $A \cap B$, *Mitte rechts:* $\overline{A}$ und rechts $A \backslash B$

Beispiel 4.2.7

Bei einem Würfelwürf sind folgende Ereignisse definiert:

- Ereignis A: Die geworfene Zahl ist gerade.

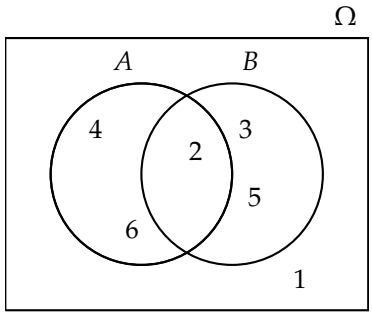
$$A = \{2, 4, 6\}$$

- Ereignis B: Die geworfene Zahl ist Primzahl.

$$B = \{2, 3, 5\}$$

Zudem ist Ω wie gewohnt

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

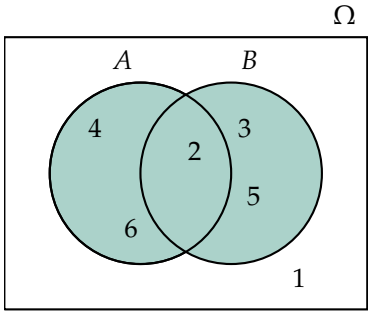


Nun können gemäss Tabelle 4.1 und Abbildung 4.1 vier neue Mengen bilden.

1. Vereinigung: Dies sind alle Elemente, die entweder in A *oder* in B *oder* in beiden Mengen vorkommen.

   $$A \cup B = \{2, 3, 4, 6\}$$

   Das Element 2 kommt sowohl in der Menge A als auch in der Menge B vor.

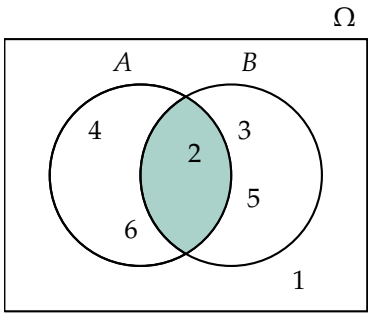


2. Durchschnitt, Schnittmenge: Dies sind alle Elemente, die in A *und* in B vorkommen.

   $$A \cap B = \{2\}$$

   Das Element 2 ist das einzige Element, das sowohl in der Menge A wie auch B

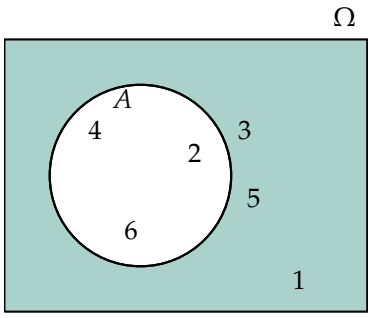
vorkommt.



3. Komplement: Dies sind alle Elemente von Ω, die nicht in der entsprechenden Menge vorkommen.

$$\overline{A} = \{1, 3, 5\}$$

Die sind die ungeraden Zahlen.



Und

$$\overline{B} = \{1, 4, 6\}$$

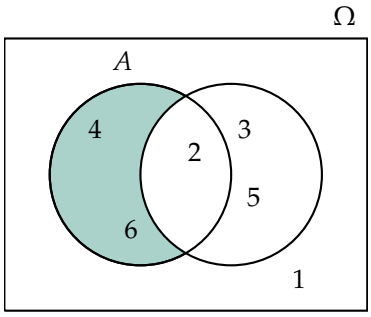
Dies sind die Zahlen von 1 bis 6, die keine Primzahlen sind[1].

4. Differenz: Dies sind alle Elemente der Menge A, die aber nicht in der Menge B vorkommen.

$$A \backslash B = \{4, 6\}$$

Das Element 2 kommt sowohl in A, wie auch in B vor und gehört deswegen *nicht* zur Differenz.

[1] Nein, 1 ist keine Primzahl, obwohl sie zum Teil bis ins 19. Jahrhundert als eine galt.



◀

Disjunkte Ereignisse

Für den nächsten Abschnitt brauchen wir noch einen neuen Begriff: *disjunkte Ereignisse*.

Zwei Ereignisse A und B heissen *disjunkt*, wenn sich A und B gegenseitig ausschliessen und daher nicht gemeinsam eintreten können. In diesem Fall gilt

$$A \cap B = \{\}$$

Dieses Ereignis ist also unmöglich.

Beispiel 4.2.8

Wir werfen zwei Würfel nacheinander und wählen folgende beiden Ereignisse:

- A: Die Wurfzahl des ersten Würfels ist 5
- B: Die Augensumme ist 3

Diese beiden Ereignisse können nicht gemeinsam eintreten, denn wenn A (Zahl 5) eintritt, dann kann B (Augensumme 3) nicht mehr eintreten. ◀

Wir haben bis jetzt noch kaum mit Wahrscheinlichkeiten gerechnet. Dazu brauchen wir Axiome (Grundregeln) und Rechenregeln.

4.2.6. Axiome und Rechenregeln der Wahrscheinlichkeitsrechnung

Axiome

Die Wahrscheinlichkeitsrechnung wird auf folgenden drei Grundregeln (Axiomen) aufgebaut:

Kolmogorov Axiome der Wahrscheinlichkeitsrechnung

Jedem Ereignis A wird eine *Wahrscheinlichkeit* $P(A)$ zugeordnet. Dabei müssen die folgenden drei grundlegenden Regeln - *Axiome der Wahrscheinlichkeitsrechnung* - erfüllt sein:

A1: $P(A) \geq 0$

A2: $P(\Omega) = 1$

A3: $P(A \cup B) = P(A) + P(B) \quad$ falls $A \cap B = \{\}$

Bemerkungen:

i. Die Bezeichnung $P(A)$ steht für die Wahrscheinlichkeit, dass das Ereignis A eintritt.

   Ist A das Ereignis, eine ungerade Zahl zu würfeln, so gilt bei einem fairen Würfel

   $$P(A) = \frac{1}{2}$$

ii. Der Buchstabe P steht für das englische Wort *probability*.

iii. A1 besagt, dass eine Wahrscheinlichkeit nicht negativ sein kann.

iv. Bei A2 wird die Wahrscheinlichkeit des sicheren Ereignisses mit $P(\Omega) = 1$ festgelegt. Dies bedeutet, dass zusammen mit A1 Wahrscheinlichkeiten zwischen 0 und 1 liegen müssen.

   Wahrscheinlichkeiten werden in der Mathematik und damit in der Statistik praktisch nie in Prozenten angeben.

v. A3 ist in Abbildung 4.2 dargestellt. Sind zwei Ereignisse *disjunkt*, so können wir die Wahrscheinlichkeit, dass eines der beiden eintritt, bestimmen, indem wir die Wahrscheinlichkeiten der beiden Ereignisse einfach addieren.

   Diese Regel gilt *nicht*, falls die Ereignisse nicht disjunkt sind. Die Rechenregel für diesen Fall folgt gleich. ⧫

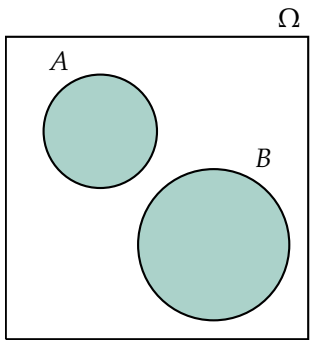


Abbildung 4.2. : Schematische Darstellung von Axiom A3

Beispiel 4.2.9

Beim Wurf zweier fairer Münzen ist es plausibel, dass alle 4 Elemente von

$$\Omega = \{KK, KZ, ZK, ZZ\}$$

gleich wahrscheinlich sind. Wegen $P(\Omega) = 1$ muss gelten

$$P(KK \cup KZ \cup ZK \cup ZZ) = 1$$

Nun sind alle diese Ereignisse disjunkt. So kann nicht ZZ und KZ gleichzeitig auftreten und damit können wir die Wahrscheinlichkeiten der entsprechenden Ereignisse einfach addieren. Das Ergebnis muss 1 ergeben:

$$P(KK) + P(KZ) + P(ZK) + P(ZZ) = 1$$

Somit gilt

$$P(KK) = P(KZ) = P(ZK) = P(ZZ) = \frac{1}{4}$$

◀

Rechenregeln

Weitere Regeln können aus den drei Axiomen von Kolmogorov hergeleitet werden. Wir werden allerdings nur die ersten beiden brauchen.

Rechenregeln

Sind A, B und $A_1, \ldots A_n$ Ereignisse, dann gilt

1.	$P(\overline{A})$	$= 1 - P(A)$	für jedes A
2.	$P(A \cup B)$	$= P(A) + P(B) - P(A \cap B)$	für beliebige A und B
3.	$P(A_1 \cup \ldots \cup A_n)$	$\leq P(A_1) + \ldots + P(A_n)$	für beliebige $A_1, \ldots, A_n$
4.	$P(B)$	$\leq P(A)$	für bel. A und B mit $B \subseteq A$
5.	$P(A \backslash B)$	$= P(A) - P(B)$	für bel. A und B mit $B \subseteq A$

Wir können uns die Wahrscheinlichkeiten als Flächen im Venn-Diagramm vorstellen. Dabei ist die Totalfläche von Ω gleich 1 oder $P(\Omega) = 1$. Dann erscheinen diese Rechenregeln ganz natürlich.

1. Regel folgt sofort aus Abbildung 4.3:

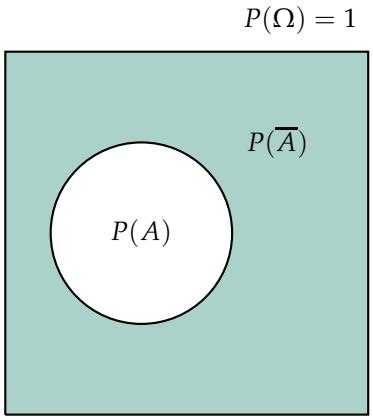


Abbildung 4.3. : Wahrscheinlichkeit für das Komplement

   Dabei ist $P(A)$ der Flächeninhalt der Fläche A und $P(\overline{A})$ ist der Flächeninhalt der restlichen Fläche in Ω. Es gilt also offensichtlich

$$P(A) + P(\overline{A}) = P(\Omega) = 1$$

   und damit

$$P(\overline{A}) = 1 - P(A)$$

2. Regel folgt sofort aus Abbildung 4.4:

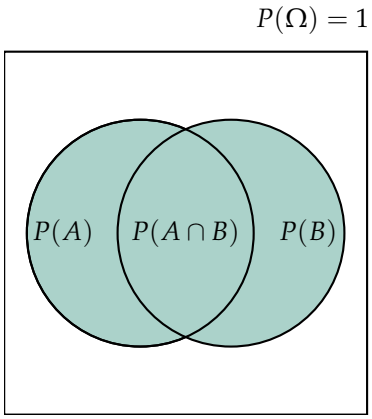


Abbildung 4.4. : Wahrscheinlichkeit für nicht disjunkte Ereignisse

Hier wird die Schnittmenge $A \cap B$ doppelt gezählt, also müssen wir diese einmal abziehen. Also gilt

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

Diese Regel ist die Verallgemeinerung von Axiom A3. Sind die Mengen A und B disjunkt, so gilt

$$A \cap B = \{\}$$

Mit der 2. Regel erhalten wir dann

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

Und dies ist genau Axiom A3, da die Wahrscheinlichkeit, dass das unmögliche Ereignis $\{\}$ eintritt, gleich 0 ist.

4.2.7. Diskrete Wahrscheinlichkeitsmodelle

Wir behandeln in diesem Kapitel ausschliesslich *diskrete* Wahrscheinlichkeitsmodelle, im Kapitel 5 werden wir uns dann mit *stetigen* Wahrscheinlichkeitsmodellen beschäftigen.

Bei einem diskreten Wahrscheinlichkeitsmodell ist der Grundraum Ω endlich oder unendlich und diskret ist. Mit dem Begriff „diskret" meinen wir folgende Mengen:

- Alle endlichen Mengen sind diskret, wie

$$\Omega = \{0, 1, \ldots, 10\}$$

- Unendliche, aber diskrete Mengen, wie

$$\Omega = \mathbb{N}_0 = \{0, 1, 2, \ldots\}$$

Ein bisschen salopp können wir diskrete Mengen als „löchrig" betrachten. Das heisst, dass beispielsweise der Wert 1.2 in der Menge $\{1, 2, 3\}$ nicht vorkommt, es hat zwischen 1 und 2 ein „Loch".

Die Menge $\Omega = \mathbb{R}$ (alle Dezimalbrüche oder alle Punkte der Zahlengerade) ist *nicht* diskret. Sie wird später im Kapitel 5 für Messdaten eine sehr wichtige Rolle spielen.

Im diskreten Fall ist die Wahrscheinlichkeit eines Ereignisses

$$A = \{\omega_1, \omega_2, \ldots, \omega_n\}$$

durch die Wahrscheinlichkeiten der zugehörigen Elementarereignisse $P(\omega)$ festgelegt:

$$P(A) = P(\omega_1) + P(\omega_2) + \ldots + P(\omega_n) = \sum_{\omega_i \in A} P(\omega_i)$$

Dies folgt aus Axiom A3. Alle Elementarereignisse sind disjunkt, da nur jeweils genau ein Elementarereignis eintreten kann, aber nicht mehrere zusammen. Damit können wir die Wahrscheinlichkeiten der entsprechenden Elementarereignisse einfach addieren.

Beispiel 4.2.10

Wir haben einen Würfel, der *nicht* fair ist. Die Wahrscheinlichkeiten, unterschiedliche Zahlen zu werfen, sind also *nicht* gleich. In Tabelle 4.2 sind die entsprechenden Wahrscheinlichkeiten für alle Wurfzahlen aufgeführt.

ω	1	2	3	4	5	6
$P(\omega)$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{4}$	$\frac{1}{12}$	$\frac{1}{12}$

Tabelle 4.2. : Wahrscheinlichkeiten für einen nicht-fairen Würfel

Es muss gelten

$$\begin{aligned} P(\Omega) &= P(1) + P(2) + P(3) + P(4) + P(5) + P(6) \\ &= \frac{1}{3} + \frac{1}{6} + \frac{1}{12} + \frac{1}{4} + \frac{1}{12} + \frac{1}{12} \\ &= 1 \end{aligned}$$

Für das Ereignis $A = \{1, 2, 4\}$ ist dann dessen Eintretenswahrscheinlichkeit

$$\begin{aligned} P(A) &= P(1) + P(2) + P(4) \\ &= \frac{1}{3} + \frac{1}{6} + \frac{1}{4} \\ &= \frac{3}{4} \end{aligned}$$

Man beachte, dass das Resultat nicht gleich ist, wenn der Würfel fair wäre. Das Ergebnis wäre dann $\frac{1}{2}$.

Wir können noch die Wahrscheinlichkeit berechnen, eine Zahl kleiner als 6 zu würfeln. Das entsprechende Ereignis B ist

$$B = \{1, 2, 3, 4, 5\}$$

mit der zugehörigen Eintretenswahrscheinlichkeit

$$\begin{aligned} P(B) &= P(1) + P(2) + P(3) + P(4) + P(5) \\ &= \frac{1}{3} + \frac{1}{6} + \frac{1}{12} + \frac{1}{4} + \frac{1}{12} \\ &= \frac{11}{12} \end{aligned}$$

In diesem Fall können wir diese Wahrscheinlichkeit allerdings wesentlich einfacher mit der sogenannten *Gegenwahrscheinlichkeit* berechnen. Dies ist nichts anderes als die Anwendung der 1. Rechenregel. Das Komplement $\overline{B}$ von B ist

$$\overline{B} = \{6\}$$

Dann gilt:

$$P(B) = 1 - P(\overline{B}) = 1 - P(6) = 1 - \frac{1}{12} = \frac{11}{12}$$

◀

4.2.8. Laplace-Wahrscheinlichkeit

In vielen Fällen ist es plausibel anzunehmen, dass jedes Elementarereignis die gleiche Wahrscheinlichkeit eintritt. Ein solches Wahrscheinlichkeitsmodell wird *Modell von Laplace* genannt. Das ist das, was sich die meisten unter Wahrscheinlichkeit vorstellen.

Beispiel 4.2.11

Das Standardbeispiel hierfür ist der Wurf eines fairen Würfels. Da wird *jede* Zahl mit der Wahrscheinlichkeit $1/6$ geworfen. ◀

In diesen Fällen, wo die Wahrscheinlichkeit für das Eintreten eines Elementarereignisse für alle Elementarereignisse gleich ist, gibt es eine besonders einfache Möglichkeit, die Wahrscheinlichkeit eines Ereignisses zu berechnen.

Ein Ereignis E besteht aus g verschiedenen Elementarereignissen

$$E = \{\omega_1, \omega_2, \ldots, \omega_g\}$$

und der Grundraum Ω aus m Elementarereignissen.

Da sich die Wahrscheinlichkeiten aller Elementarereignisse zu 1 addieren, muss für die Eintretenswahrscheinlichkeit $P(\omega_k)$ eines Elementarereignisses ω_k

$$P(\omega_k) = \frac{1}{|\Omega|} = \frac{1}{m}$$

sein.

Die Bezeichnung mit den Längsstrichen $|A|$ steht für die *Anzahl* Elemente einer Menge. Wir sprechen auch von der *Grösse* einer Menge.

Die Wahrscheinlichkeit, dass ein Elementarereignis eintritt, ist 1 durch die Anzahl der Elementarereignisse.

Somit gilt für $P(E)$:

$$\begin{aligned} P(E) &= P(\omega_1) + P(\omega_2) + \ldots + P(\omega_g) \\ &= \frac{1}{m} + \frac{1}{m} + \ldots + \frac{1}{m} \\ &= g \cdot \frac{1}{m} \\ &= \frac{g}{m} \end{aligned}$$

Für ein Ereignis E im Laplace Modell gilt also

$$P(E) = P(\omega_1) + P(\omega_2) + \ldots + P(\omega_g) = \sum_{\omega_k \in E} P(\omega_k) = \frac{g}{m}$$

Wir teilen die Anzahl der *„günstigen"* Elementarereignisse durch die Anzahl der *„möglichen"* Elementarereignisse.

Beispiel 4.2.12

Wir werfen einen blauen und einen roten Würfel, beide sind fair. Wie gross ist die Wahrscheinlichkeit, dass die Augensumme 7 der beiden Würfel ergibt?

Ein Elementarereignis beschreibt die Augenzahlen auf beiden Würfeln. Dieses Elementarereignis können wir in der Form 14 schreiben, wenn der blaue Würfel eine 1 und der rote eine 4 zeigt. Man beachte, dass 14 und 41 nicht dieselben Elementarereignisse sind. Im ersten Fall zeigt der blaue Würfel eine 1, im zweiten Fall eine 4.

Es sind insgesamt 36 Elementarereignisse möglich, wobei die erste Zahl immer die Wurfzahl des blauen Würfels angibt und die zweite Zahl die Wurfzahl des roten Würfels:

$$\Omega = \{11, 12, \ldots, 16, 21, 22, \ldots, 65, 66\}$$

Wir bezeichnen mit E das Ereignis, dass die Augensumme 7 erreicht wird. Es gibt davon 6 Elementarereignisse:

$$E = \{16, 25, 34, 43, 52, 61\}$$

Da alle Elementarereignisse gleich wahrscheinlich sind, ist die Wahrscheinlichkeit für das Eintreten des Ereignisses E nach dem Laplace-Modell:

$$P(E) = \frac{|E|}{|\Omega|} = \frac{6}{36} = \frac{1}{6}$$

◀

4.2.9. Der Begriff der Unabhängigkeit

Wir haben bei dem Axiom A3. in Unterabschnitt 4.2.6 gesehen, dass wir die Wahrscheinlichkeit von $P(A \cup B)$ berechnen können, sofern wir $P(A)$ und $P(B)$ kennen

und die Ereignisse A und B disjunkt sind:

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

Sind die Ereignisse nicht disjunkt, so haben wir gesehen, dass folgende Regel gilt:

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

Kennen wir $P(A \cap B)$, so können wir auch diese Wahrscheinlichkeit berechnen.

Es gibt leider *keine* allgemeine Regel, wie wir die Wahrscheinlichkeit $P(A \cap B)$ aus den Wahrscheinlichkeiten $P(A)$ und $P(B)$ berechnen können.

Ein wichtiger Spezialfall liegt vor, wenn die Berechnung von $P(A \cap B)$ aus $P(A)$ und $P(B)$ mit Hilfe folgender Produktformel möglich ist.

Sind die Ereignisse A und B *stochastisch unabhängig*, so gilt

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

Was heisst hier aber „stochastisch unabhängig"? Damit ist gemeint, dass der Ausgang des Ereignisses A keinen Einfluss auf den Ausgang des Ereignisses B hat und umgekehrt. Aber was heisst nun das?

Wir werden den Begriff der stochastischen Unabhängigkeit an Beispielen erläutern.

Beispiel 4.2.13

Ereignis A sei mit einem fairen Würfel eine eins oder zwei zu werfen und Ereignis B sei Kopf beim Werfen einer fairen Münze.

Nun hat das Werfen einer Münze wohl keinen Einfluss auf das Resultat beim Würfelwurf. Somit können wir die Formel oben verwenden:

$$P(12 \cap K) = P(12) \cdot P(K) = \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}$$

◀

Beispiel 4.2.14

Sei das Ereignis E, dass Tokyo an einem bestimmten Tag durch ein Erdbeben erschüttert wird und das Ereignis F, dass an demselben Tag ein Taifun über die Stadt fegt.

Nun hat ein Erdbeben wohl kaum Einfluss auf das Entstehen eines Taifuns. Die beiden Ereignisse sind also stochastisch unabhängig. ◀

Beispiel 4.2.15

Werfen wir eine Münze zweimal nacheinander, so hat das Resultat des ersten Wurfes kaum Einfluss auf das Resultat des zweiten Wurfes.

Dies ist allerdings nur richtig, wenn es sich um eine ideale Münze handelt. Werfen wir eine reale Münze, so ergeben sich durch den Aufprall minimalste Veränderungen an der Münze. Diese haben einen Einfluss auf die Wurfwahrscheinlichkeit für Kopf (oder Zahl) beim nächsten Wurf. Allerdings sind diese Veränderungen so klein, dass wir diese vernachlässigen können. ◀

Hier noch zwei Beispiele für Ereignisse, die *nicht* stochastisch unabhängig sind.

Beispiel 4.2.16

Wir haben in einem Topf 20 Lose mit 4 Gewinnen und ziehen zweimal hintereinander (ohne Zurücklegen).

Sei nun das Ereignis A, dass wir beim ersten Ziehen gewinnen und das Ereignis B, dass wir bei der zweiten Ziehung gewinnen.

Diese beiden Ereignisse sind *nicht* stochastisch unabhängig. Warum nicht?

Ziehen wir beim ersten Ziehen ein Gewinnlos, so ist die Wahrscheinlichkeit, dass A eintrifft

$$P(A) = \frac{4}{20}$$

Bei der 2. Ziehung „fehlt" dann ein Gewinn und die Wahrscheinlichkeit dann zu gewinnen ist

$$P(B) = \frac{3}{19}$$

Ziehen wir bei der ersten Ziehung eine Niete, so ist die Wahrscheinlichkeit bei der 2. Ziehung zu gewinnen

$$P(B) = \frac{4}{19}$$

Das heisst, je nachdem ob Ereignis A eintrifft oder nicht, ändert sich die Wahrscheinlichkeit, dass B eintritt. Die Ereignisse sind also *nicht* stochastisch unabhängig. ◀

Beispiel 4.2.17

Sei A das Ereignis, dass morgen schönes Wetter ist und B das Ereignis, dass eine Person morgen gute Laune hat.

Nun sind die meisten Menschen bei schönem Wetter besser aufgelegt, als bei schlechtem Wetter. Also hat das Eintreffen von A einen Einfluss auf B. Die Ereignisse sind nicht stochastisch unabhängig. ◀

Achtung

Die Formel

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

gilt *nur*, falls die Ereignisse A und B stochastisch unabhängig sind.

Sind die Ereignisse nicht stochastisch unabhängig, so gibt es keine allgemeine Formel für die Berechnung der Wahrscheinlichkeit der Schnittmenge von zwei Ereignissen.

4.3. Zufallsvariable

4.3.1. Einleitung, Definition

Der Begriff der *Zufallsvariable* spielt eine ganz zentrale Rolle in der Statistik.

Wir beginnen mit einem einfachen Beispiel.

Beispiel 4.3.1

Ein Pack Spielkarten besteht in der Schweiz aus 36 verschiedenen Karten, mit 4 „Farben" (Herz, Schaufel, Ecke und Kreuz) und je den Werten 6, 7, 8, 9, 10, Bube, Dame, König und Ass.

Wir ziehen zufällig aus ein Pack Spielkarten nacheinander drei Karten, wobei wir die Karten jeweils wieder zurücklegen. Dies machen wir zweimal und erhalten die beiden folgenden Resultate:

1. 6, Dame, König
2. 8, Bube, Ass

Die Frage ist nun, welcher der Versuche „besser" ist. Dies lässt sich aus den Resultaten oben nicht sagen.

Um diese Frage zu beantworten, welcher der Versuche „besser" ist, ordnen wir den einzelnen Spielkarten *Zahlen* zu. So haben in der Schweiz die Karten 6, 7, 8, 9 den Wert 0, die 10 hat den Wert 10, der Bube hat den Wert 2, die Dame den Wert 3, der König den Wert 4 und das Ass den Wert 11.

Wir haben also jede Spielkarte mit einer Zahl identifiziert.

Nun können wir die Ziehungen miteinander vergleichen:

1. 6, Dame, König $\rightarrow$ $0+3+4=7$
2. 8, Bube, Ass $\rightarrow$ $0+2+11=13$

Die zweite Ziehung ist also mit dieser Bewertung besser (sofern natürlich eine höhere Zahl „besser" als eine tiefe ist). ◀

Was wir im vorhergehenden Beispiel gesehen haben, kommt in der Stochastik häufig vor. Wir haben ein Zufallsexperiment mit dem Grundraum Ω. Dann ordnen wir jedem Elementarereignis von Ω einen *Zahlwert* zu. Zu jedem Elementarereignis ω gehört demnach ein Zahlwert

$$X(\omega) = x$$

Dabei ist X eine *Funktion*, die jedem Elementarereignis ω genau den Zahlwert x zuordnet.

Beispiel 4.3.2

Wir konkretisieren Beispiel 4.3.1 und ordnen jeder Karte eine Zahl zu:

$$\begin{array}{lcl} \omega = \text{As} & \mapsto & X(\omega) = 11 \\ \omega = \text{König} & \mapsto & X(\omega) = 4 \\ \vdots & & \vdots \\ \omega = \text{Sechs} & \mapsto & X(\omega) = 0 \end{array}$$

◀

Wie wir in diesem Beispiel gerade gesehen, ist X eine Funktion auf dem Grundraum Ω, wobei der Output der Funktion eine Zahl ist. Eine solche Funktion wird als *Zufallsvariable* bezeichnet. Sie ordnet jedem Element des Grundraumes *eine Zahl* zu.

Der Vorteil bei diesem Vorgehen liegt darin, dass wir mit den Werten der Zufallsvariable Berechnungen durchführen können.

Beispiel 4.3.3

Im Beispiel 4.3.1 oben ist es nun möglich, mit den Zahlenwerten $X(\omega)$ den „Durchschnitt" (den sogenannten *Erwartungswert*, den wir später im Unterkapitel 4.4 kennenlernen werden) der gezogenen Karten zu berechnen.

So ist der Durchschnitt von „6, Dame, König" gleich $\frac{7}{3}$.

Für die Elementarereignisse „6", „Dame" und „König" macht das Wort „Durchschnitt" gar keinen Sinn. ◀

Beispiel 4.3.4

Wir werfen gemeinsam einen blauen und roten Würfel. Der Grundraum Ω besteht aus den Augenzahlen der beiden Würfel. Diesen können wir wie folgt modellieren (siehe Beispiel 4.2.12):

$$\Omega = \{11, 12, \ldots 16, 21, 22, \ldots, 26, \ldots 66\}$$

Dabei sei die erste Zahl immer die Augenzahl des blauen und die zweite Zahl die Augenzahl des roten Würfels. Somit sind 23 und 32 *verschiedene* Elementarereignisse.

Nun können wir auf Ω verschiedene Zufallsvariablen definieren:

1. Sei X die Zufallsvariable für die Summe der Augenzahlen. Dann gilt:

   $$X(16) = 7 \qquad \text{oder} \qquad X(31) = 4$$

   Die Werte, die die Zufallsvariable annehmen kann, wird als *Wertemenge* bezeichnet. Für die Zufallsvariable X ist die Wertemenge:

   $$\mathbb{W}_X = \{2, 3, 4 \ldots, 11, 12\}$$

2. Sei Y die Augenzahl des blauen Würfels. Dann gilt:

   $$Y(16) = 6 \qquad \text{oder} \qquad Y(31) = 1 \qquad \text{oder} \qquad Y(13) = 1$$

Die Wertemenge von Y lautet

$$\mathbb{W}_Y = \{1, 2, \ldots, 6\}$$

3. Sei Z gleich 0 für alle Elementarereignisse.

$$Z(16) = 0 \qquad \text{oder} \qquad Z(31) = 0 \qquad \text{oder} \qquad Z(13) = 0$$

Die Wertemenge von Z lautet

$$\mathbb{W}_Z = \{0\}$$

Das 3. Beispiel ist eine völlig legitime Zufallsvariable. Wie sinnvoll diese ist, ist eine andere Frage. ◀

Beispiel 4.3.5

Wir wählen zufällig eine Person aus. Der Grundraum Ω besteht dann aus allen Personen dieses Planeten. Auch hier sind viele Zufallsvariablen denkbar.

1. Sei X die Zufallsvariable, die jeder Person das Einkommen zuordnet.
2. Sei Y die Zufallsvariable, die jeder Person die Körpergrösse zuordnet.
3. Sei Z die Zufallsvariable, die jeder Person das Alter zuordnet.

Folgende Variablen sind *keine* Zufallsvariablen:

4. Die Variable V ordnet jeder Person das Geschlecht zu.
5. Die Variable W ordnet jeder Person die zugehörige Nationalität zu.

Das „Resultat" einer Zufallsvariable *muss* eine *Zahl* sein. ◀

Bemerkungen:

i. Die Notation X (oder auch $Y, Z, \ldots$) ist eher ungewohnt für die Bezeichnung einer Funktion, ist aber üblich in der Wahrscheinlichkeitsrechnung und der Statistik.

ii. Die Zufallsvariable X ist eine *Funktion*. Diese werden in Mathematik oft in der Form

$$y = f(x)$$

geschrieben.

Auch hier hat sich in der Stochastik eine eigene Notation durchgesetzt.

iii. Folgende Unterscheidung ist *sehr wichtig*:

- Eine *Zufallsvariable* wird mit einem *Grossbuchstaben* X (oder Y, Z) bezeichnet.
- Der entsprechende *Kleinbuchstabe* x (oder y, z) stellt einen *konkreten Wert* dar, den die Zufallsvariable annehmen kann, also eine Zahl.
- Für das Ereignis, bei dem die Zufallsvariable X den Wert x annimmt, schreiben wir (siehe auch 4.3.2).

$$X = x$$

iv. Bei einer Zufallsvariable ist nicht die Funktion X zufällig, sondern nur das Argument ω: Je nach Ausgang des Zufallsexperiments ω erhalten wir einen anderen Wert $x = X(\omega)$.

v. Wir nennen x auch eine *Realisierung* der Zufallsvariablen X.

Falls das Experiment zweimal durchgeführt wird und wir zweimal das gleiche Ergebnis ω erhalten, dann sind auch die realisierten Werte von X gleich. ♦

Beispiel 4.3.6

Im Spielkartenbeispiel 4.3.2 entspricht die Realisierung $X = 11$ dem Ziehen eines Asses. ◀

Beispiel 4.3.7

Beim Würfelbeispiel 4.3.4 entspricht die Realisierung $X = 8$, dass die Augensumme 8 gewürfelt wurde. ◀

4.3.2. Wahrscheinlichkeitsverteilung einer Zufallsvariablen

Wir haben im Abschnitt 4.2.7 schon gesehen, wie wir die Wahrscheinlichkeit $P(E)$ eines Ereignisses E berechnen.

Entsprechend können wir auch die Wahrscheinlichkeit einer *allgemeinen* Realisierung x einer Zufallsvariable X definieren. Dazu zunächst folgendes Beispiel.

Beispiel 4.3.8

Wir definieren die Zufallsvariable X als den Wert einer gezogenen Spielkarte (siehe Beispiel 4.3.1 und fragen nun, wie gross die Wahrscheinlichkeit ist, dass die gezogene Karte den Wert 4 hat. Diese Realisierung bezeichnen wir mit

$$X = 4$$

Die zugehörige Eintretenswahrscheinlichkeit dieses Ereignisses bezeichnen wir mit

$$P(X = 4)$$

Die Realisation $X = 4$ entspricht dem Ziehen eines Königs. D.h., wir suchen die Wahrscheinlichkeit, dass ein König gezogen wird:

$$P(X = 4) = \frac{4}{36} = \frac{1}{9}$$

Wir können dies auch allgemeiner formulieren: Die Wahrscheinlichkeit, dass ein König gezogen wird, ist gleich der Summe der Wahrscheinlichkeiten, die verschiedenen Könige zu ziehen:

$$\begin{aligned} P(X = 4) &= P(\text{Herz-König}) + P(\text{Schaufel-König}) + P(\text{Kreuz-König}) + P(\text{Ecken-König}) \\ &= \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} \\ &= \frac{4}{36} = \frac{1}{9} \end{aligned}$$

Dieses Vorgehen ist hier zwar komplizierter als notwendig beschrieben, funktioniert aber auch, wenn wir Laplace-Wahrscheinlichkeit nicht verwenden können. ◀

Wahrscheinlichkeit einer Realisierung einer Zufallsvariablen

Die Werte einer Zufallsvariablen X (die möglichen Realisierungen von X) treten mit gewissen Wahrscheinlichkeiten auf. Die Wahrscheinlichkeit, dass X den Wert x annimmt, berechnet sich wie folgt:

$$P(X = x) = P(\{\omega \mid X(\omega) = x\}) = \sum_{\omega; X(\omega) = x} P(\omega)$$

Wir haben im Beispiel vorher die Wahrscheinlichkeit *einer* Realisierung berechnet. Natürlich können wir die Wahrscheinlichkeiten *aller* Realisierungen berechnen. Dies führt uns auf den *sehr wichtigen* Begriff einer *Wahrscheinlichkeitsverteilung*.

Wahrscheinlichkeitsverteilung

Berechnen wir für *jede* Realisierung einer Zufallsvariable die zugehörige Eintretenswahrscheinlichkeit, so bilden alle diese Wahrscheinlichkeiten zusammen die *Wahrscheinlichkeitsverteilung* dieser Zufallsvariablen.

Beispiel 4.3.9

Die Zufallsvariable X ist wieder der Wert einer gezogenen Jasskarte (siehe Beispiel 4.3.1. Die Wahrscheinlichkeit

$$P(X = 4) = \frac{1}{9}$$

haben wir schon berechnet.

Die Wahrscheinlichkeit $P(X = 0)$ können wir auch hier wieder mit der Laplace-Wahrscheinlichkeit berechnen. Es hat unter den 36 Karten genau 16 „leere" Karten. Somit gilt für die Wahrscheinlichkeit $P(X = 0)$:

$$P(X = 0) = \frac{16}{36} = \frac{4}{9}$$

Die Realisierung $X = 2$ entspricht dem Ziehen eines Unders. Da es 4 von denen gibt, gilt für die Wahrscheinlichkeit $P(X = 2)$:

$$P(X = 2) = \frac{4}{36} = \frac{1}{9}$$

Die Wahrscheinlichkeiten für die anderen Realisierungen berechnen wir analog. Wir haben dann *jeder* Realisierung einen Wahrscheinlichkeitswert zugeordnet. Wir sprechen dann eben von einer *Wahrscheinlichkeitsverteilung*. Die Wahrscheinlichkeitsverteilung von X ist in der Tabelle 4.3 aufgeführt.

x	0	2	3	4	10	11
$P(X = x)$	4/9	1/9	1/9	1/9	1/9	1/9

Tabelle 4.3. : Wahrscheinlichkeitsverteilung von gezogenen Jasskarten.

Die Wahrscheinlichkeiten für $X = 1$ oder $X = 178$ sind in Tabelle 4.3 *nicht* aufgeführt. Der Grund dafür ist natürlich, dass diese Werte nicht gezogen werden können. Wir können ihnen aber trotzdem eine Wahrscheinlichkeit zuordnen, nämlich die Zahl 0, da diese Realisierung nicht eintreten können:

$$P(X = 1) = 0 \quad \text{oder} \quad P(X = 178) = 0$$

Addieren wir alle Werte in unserer Wahrscheinlichkeitsverteilung auf, so müssen wir 1 erhalten. Der Grund dafür ist, dass wir eine Realisierung ziehen *müssen*.

$$P(X=0)+P(X=2)+P(X=3)+P(X=4)+P(X=10)+P(X=11)=1$$

◀

Beispiel 4.3.10

Die Wahrscheinlichkeitsverteilung für die Zufallsvariable X für die Augensumme zweier Würfel (siehe Beispiel 4.3.4) ist in Tabelle 4.4 aufgeführt (kontrollieren Sie dies nach):

x	2	3	4	5	6	7	8	9	10	11	12
$P(X=x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Tabelle 4.4. : Wahrscheinlichkeitsverteilung von gezogenen Jasskarten.

Die Brüche sind hier ungekürzt, damit man besser sieht, dass die Summe der Wahrscheinlichkeiten wieder 1 ergeben muss. ◀

Wahrscheinlichkeitsverteilung

Die „Liste" von $P(X=x)$ für alle *möglichen* Werte $x_1, x_2, \ldots, x_n$ heisst diskrete *(Wahrscheinlichkeits-) Verteilung* der diskreten Zufallsvariablen X.

Dabei gilt immer

$$P(X=x_1)+P(X=x_1)+\ldots+P(X=x_n)=1$$

oder mit dem Summenzeichen

$$\sum_{\text{alle möglichen } x} P(X=x)=1$$

Alle Wahrscheinlichkeiten einer Wahrscheinlichkeitsverteilung aufaddiert müssen 1 ergeben.

Wahrscheinlichkeitsverteilungen für endlich diskrete Zufallsvariablen werden durch Tabellen wie Tabelle 4.3 in Beispiel 4.3.9 und Tabelle 4.3.10 in Beispiel 4.4 dargestellt.

4.4. Kennzahlen einer Verteilung

Eine beliebige (diskrete) Verteilung kann vereinfachend zusammengefasst werden durch 2 Kennzahlen, den *Erwartungswert* $\mathrm{E}(X)$ und die *Standardabweichung* $\sigma(X)$.

Erwartungswert einer Zufallsvariable

Der *Erwartungswert* einer diskreten Zufallsvariable X mit den möglichen Werten $x_1, x_2, \ldots, x_n$ beschreibt die mittlere Lage der Verteilung und ist wie folgt definiert:

$$\begin{aligned}\mathrm{E}(X) &= x_1 \cdot P(X = x_1) + x_2 \cdot P(X = x_2) + \ldots + x_n \cdot P(X = x_n) \\ &= \sum_{\text{alle möglichen } x} xP(X = x)\end{aligned}$$

Varianz und Standardabweichung einer Zufallsvariable

Die *Standardabweichung* beschreibt die Streuung der Verteilung. Rechnerisch ist das Quadrat der Standardabweichung, die sogenannte *Varianz* bequemer:

$$\begin{aligned}\mathrm{Var}(X) &= (x_1 - \mathrm{E}(X))^2 \cdot P(X = x_1) + \ldots + (x_1 - \mathrm{E}(X))^2 \cdot P(X = x_1) \\ &= \sum_{\text{alle möglichen } x} (x - \mathrm{E}(X))^2 P(X = x)\end{aligned}$$

$$\sigma(X) = \sqrt{\mathrm{Var}(X)}$$

Beispiel 4.4.1

Beim Wurf eines fairen Würfels haben alle 6 möglichen Zahlen die gleiche Wahrscheinlichkeit geworfen zu werden. Die Zufallsvariable X sei die geworfene Zahl. Dann ergibt sich für den Erwartungswert $\mathrm{E}(X)$:

$$\begin{aligned}\mathrm{E}(X) &= x_1 \cdot P(X = x_1) + x_2 \cdot P(X = x_2) + \ldots + x_6 \cdot P(X = x_6) \\ &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} \\ &= \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) \\ &= 3.5\end{aligned}$$

Dieser Erwartungswert 3.5 ist nicht anderes als der Durchschnitt der Augenzahlen.

Wie lässt sich dieser Wert nun interpretieren? Wir würfeln mit diesem fairen Würfel 100 Mal und nehmen den Durchschnitt, so wird dieser meistens *nicht* exakt, aber wohl in der *Nähe* von 3.5 sein.

Diese Annäherung sollte immer besser werden, je mehr wir würfeln. Auch wenn wir 100 Milliarden Mal würfeln, so wird der Durchschnitt praktisch nie *exakt* 3.5 sein, aber sehr nahe dran.

Die Interpretation für den Erwartungswert ist die folgende: Würfeln wir sehr viele Male, so wird der Durchschnitt sehr nahe beim Erwartungswert liegen (siehe die Diskussion im Beispiel 4.4.5).

Wir können noch die Standardabweichung berechnen und machen dies mit **R**, da die Berechnung von Hand zu mühsam und zu sehr fehlerbehaftet ist.

```
x <- 1:6
p <- 1/6

E_X <- sum(x * p)

var_X <- sum((x - E_X)^2 * p)
sd_X <- sqrt(var_X)

sd_X


## [1] 1.707825
```

Das heisst, die Abweichung vom Erwartungswert 3.5 ist „durchschnittlich" 1.7. ◀

Beispiel 4.4.2

Wir berechnen noch den Erwartungswert und die Standardabweichung des Beispiels 4.2.10 eines unfairen Würfels. Die Zufallsvariable X ist die geworfene Augenzahl.

ω	1	2	3	4	5	6
$P(\omega)$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{4}$	$\frac{1}{12}$	$\frac{1}{12}$

Tabelle 4.5. : Wahrscheinlichkeiten für einen nicht-fairen Würfel

Wir berechnen den Erwartungswert, indem wir die Werte in der Tabelle 4.5 untereinander miteinander multiplizieren und alle diese Multiplikationen aufaddieren:

$$\mathrm{E}(X) = 1 \cdot \frac{1}{3} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{12} + 4 \cdot \frac{1}{4} + 5 \cdot \frac{1}{12} + 6 \cdot \frac{1}{12} = 2.833$$

```
x <- 1:6
p <- c(4, 2, 1, 3, 1, 1)/12

E_X <- sum(x * p)
E_X


## [1] 2.833333


var_X <- sum((x - E_X)^2 * p)
sd_X <- sqrt(var_X)

sd_X


## [1] 1.674979
```

Der Erwartungswert ist 2.833 und die Standardabweichung ist 1.675. Würfeln wir mit diesem Würfel sehr viele Male, so wird der Mittelwert all dieser Würfe nahe bei 2.833 liegen und die Werte weichen im „Durchschnitt" 1.675 vom Erwartungswert 2.833 ab. ◀

Beispiel 4.4.3

Wir betrachten wiederum das Spielkartenbeispiel mit der Verteilung

x	0	2	3	4	10	11
$P(X = x)$	4/9	1/9	1/9	1/9	1/9	1/9

Wir ziehen aus dem Stapel eine Karte. Welches ist der durchschnittliche Wert der Karte, die wir ziehen?

Dazu berechnen wir den Erwartungswert $\mathrm{E}(X)$:

$$\mathrm{E}(X) = 0 \cdot \frac{4}{9} + 2 \cdot \frac{1}{9} + 3 \cdot \frac{1}{9} + 4 \cdot \frac{1}{9} + 10 \cdot \frac{1}{9} + 11 \cdot \frac{1}{9} = 3.33$$

```
x <- c(0, 2, 3, 4, 10, 11)
p <- 1/9 * c(4, 1, 1, 1, 1, 1)

E_X <- sum(x * p)
E_X


## [1] 3.333333
```

Dieser Wert ist eher tief, weil relativ viele Karten den Wert 0 haben.

Der Erwartungswert ist der durchschnittliche Wert, den wir erwarten können, wenn wir sehr oft eine Karte ziehen und diese wieder in den Stapel zurücklegen.

Wir berechnen noch die Varianz und die Standardabweichung:

$$\begin{aligned}\text{Var}(X) &= (0-3.33)^2 \cdot \frac{4}{9} + (2-3.33)^2 \cdot \frac{1}{9} + (3-3.33)^2 \cdot \frac{1}{9} \\ &\quad + (4-3.33)^2 \cdot \frac{1}{9} + (10-3.33)^2 \cdot \frac{1}{9} + (11-3.33)^2 \cdot \frac{1}{9} \\ &= 16.67\end{aligned}$$

und

$$\sigma(X) = \sqrt{16.67} = 4.08$$

```
var_X <- sum((x - E_X)^2 * p)
sd_X <- sqrt(var_X)

sd_X


## [1] 4.082483
```

Dieser Wert ist verglichen zum Erwartungswert eher gross, da mit den Werten 10 und 11 eine relative grosse Abweichung vom Erwartungswert vorhanden ist. Die Abweichung ist hier in erster Linie nach oben vom Erwartungswert und nicht nach unten .

Die Standardabweichung ist die „mittlere" Abweichung vom Mittelwert, wobei diese vor allem nach oben möglich ist. ◀

Beispiel 4.4.4

Der Erwartungswert von X für das Werfen von zwei Würfeln in Beispiel 4.3.10 ist 7 und die Standardabweichung ist 2.415. ◀

Bemerkungen:

i. Der Erwartungswert einer diskreten Zufallsvariable ist das gewichtete arithmetische Mittel von allen möglichen Werten, wobei die Werte mit ihrer Wahrscheinlichkeit gewichtet werden.

ii. Der Erwartungswert wird oft auch mit μ_X bezeichnet. Der Index X wird oft weggelassen, wenn klar ist, um welche Zufallsvariable es sich handelt.

iii. Sind die Wahrscheinlichkeiten für alle Werte $x_1, x_2, \ldots, x_n$ gleich (Laplace-Wahrscheinlichkeit), so entspricht der Erwartungswert gerade dem arithmetischen Mittel der Werte (siehe Beispiel 4.4.1).

iv. Ebenso wird bei der Varianz das Quadrat der Abweichung eines Wertes der Zufallsvariable vom Erwartungswert mit der Wahrscheinlichkeit des Wertes gewichtet.

   Die Standardabweichung hat dieselbe Einheit wie X, während die Einheit der Varianz deren Quadrat ist: Wird z. B. X in Metern (m) gemessen, so besitzt $\mathrm{Var}(X)$ die Dimension Quadratmeter (m^2) und $\sigma(X)$ wiederum die Dimension Meter (m.

   Wir haben dies schon bei der empirischen Standardabweichung in Abschnitt 2.2.2 gesehen. ♦

4.4.1. Unterschied empirischer und theoretischer Kennzahlen

Wir haben in Abschnitt 2.2.1 das arithmetische Mittel

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

und in Abschnitt 2.2.2 die *empirische* Varianz

$$\mathrm{Var}(x) = \frac{(x_1 - \overline{x}_n)^2 + (x_2 - \overline{x}_n)^2 + \cdots + (x_n - \overline{x}_n)^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2$$

und die empirische Standardabweichung s_x

$$s_x = \sqrt{\mathrm{Var}(x)} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}$$

kennengelernt.

Wie hängen diese Definitionen mit den Definitionen für Erwartungswert und Standardabweichung einer Zufallsvariable X zusammen? Wir müssen hier sehr genau zwischen den verschiedenen Definitionen unterscheiden.

Wir beginnen zuerst mit dem Unterschied zwischen Erwartungswert und arithmetischem Mittel:

Unterschied Mittelwert und Erwartungswert

- Der arithmetische Mittelwert $\overline{x}$ wird aus *konkreten* Daten gewonnen. Wir haben also Messwerte $x_1, \ldots, x_n$ und können nach der Formel oben $\overline{x}_n$ berechnen.
- Die Erwartungswert σ_X ist ein *theoretischer* Wert, der sich aus dem Modell der Wahrscheinlichkeitsverteilung ergibt.

Die Hoffnung ist nun, dass das arithmetische Mittel $\overline{x}$ für immer mehr Versuche den theoretischen Wert

$$\mu_X = \mathrm{E}(x)$$

immer besser annähert, sofern die konkreten Daten $x_1, \ldots, x_n$ der Wahrscheinlichkeitsverteilung von X folgen.

Wie dies gemeint ist, sehen wir am folgenden Beispiel.

Beispiel 4.4.5

Wir kommen auf das Beispiel 4.4.1 des fairen Würfels zurück. Dort berechneten wir den Erwartungswert

$$\mathrm{E}(X) = 3.5$$

wobei X die Zufallsvariable für die Augenzahl ist.

Nun nehmen wir so einen idealen, fairen Würfel und würfeln $n = 10$ Mal. Obwohl der Würfel fair ist, wird der Durchschnitt der Wurfzahlne sehr selten genau 3.5 sein. Führen wir das konkret durch und wählen als fairen Würfel **R**:

```
x <- sample(1:6, size = 10, replace = T)
x


##  [1] 2 3 1 3 1 1 5 6 3 3


mean(x)


## [1] 2.8
```

Der Durchschnitt 2.8 dieser 10 Würfe liegt hier noch einiges neben dem zu *erwartenden* Wert von 3.5.

Wenn wir den Würfel nochmals 10 Mal werfen, so erhalten wir normalerweise ein anderes Resultat für den Durchschnitt. Mit **R** simulieren wir 10 Mal 10 Würfe und berechnen jeweils die entsprechenden Durchschnitte.

```
for (i in 1:10) {
    x <- sample(1:6, size = 10, replace = T)
    cat(mean(x), " ")
}


## 3.3  3.5  4.2  3.4  2.9  3  3.9  2.9  4.1  2.7
```

Die Durchschnitte aus den jeweils 10 Würfen liegen zwischen 2.7 und 4.1. Zwar liegen diese Durchschnitte in der „Nähe" von 3.5, aber nicht sehr nahe.

Bemerkungen: R-Code

i. Der Befehl R-Befehl `sample(...)` wählt aus den Zahlen 1 bis 6 (`1:6`) die Anzahl `size = 10` aus.

ii. Die Option `replace = T` stellt sicher, dass die Zahl nach dem Ziehen wieder zurückgelegt (verfügbar) ist.

iii. Die `for`-Schlaufe

```
for (i in 1:10) {
    ...
}
```

erreicht, dass die `...` 10 Mal (`1:10`) durchgeführt wird.

iv. Der Befehl `cat(...)` ist ein Ausgabebefehl, wie `print(...)`. Der Leerschlag in `cat(..., " ")` erreicht, dass zwischen den Outputs ein Leerschlag eingefügt wird. ♦

Würfeln wir mit $n = 100$ Würfel, so erhalten wir folgendes Resultat:

```
x <- sample(1:6, size = 100, replace = T)
x


##   [1] 5 6 6 1 5 1 4 5 1 2 3 1 3 6 2 3 1 6 1 4 3 6 1 6 5 6 6 3 1 5
##  [31] 5 6 6 2 2 3 4 3 1 1 5 1 2 4 5 6 5 4 2 5 6 5 2 6 4 4 4 4 1 2
##  [61] 2 6 6 3 5 3 6 5 5 1 5 6 1 2 1 5 4 1 6 1 5 3 1 2 6 5 3 1 4 1
##  [91] 2 1 4 4 1 4 6 1 5 6


mean(x)


## [1] 3.57
```

Der Mittelwert 3.57 dieser 100 Würfe ist schon relativ nahe beim theoretischen Wert von 3.5. Auch dies können wir noch 10 Mal machen und erhalten:

```
for (i in 1:10) {
    x <- sample(1:6, size = 100, replace = T)
    cat(mean(x), " ")
}


## 3.39  3.43  3.55  3.5  3.48  3.61  3.46  3.64  3.28  3.41
```

Die Durchschnitte aus diesen jeweils 100 Würfen liegen zwischen 3.28 und 3.64.

Dasselbe für $n = 1000$ Würfe:

```
## 3.475  3.49  3.435  3.437  3.407  3.479  3.567  3.474  3.498  3.565
```

Und noch für $n = 1\,000\,000$. Hier sind die Durchschnitte auf 3 Nachkommastellen gerundet.

```
## 3.498  3.497  3.501  3.496  3.501  3.5  3.497  3.5  3.503  3.499
```

Wir sehen, dass der Durchschnitt konkreter Zahlen für immer grössere n immer näher bei 3.5 liegt. ◀

Derselbe Unterschied wie für Mittelwert und Erwartungswert gibt es auch für die empirische Standardabweichung und die Standardabweichung einer Zufallsvariablen.

- Die empirische Standardabweichung s_X wird aus *konkreten* Daten gewonnen. Wir haben also Messwerte $x_1, \ldots, x_n$ und können s_X nach der Formel aus Abschnitt 2.2.2 berechnen.
- Die Standardabweichung σ_X ist ein theoretischer Wert, der sich aus dem Modell der Wahrscheinlichkeitsverteilung ergibt.

Die Hoffnung ist nun, dass die empirische Standardabweichung s_X für immer mehr Versuche den theoretischen Wert σ_X immer besser annähert, sofern die Daten der Wahrscheinlichkeitsverteilung von X folgen.

4.5. Bedingte Wahrscheinlichkeit

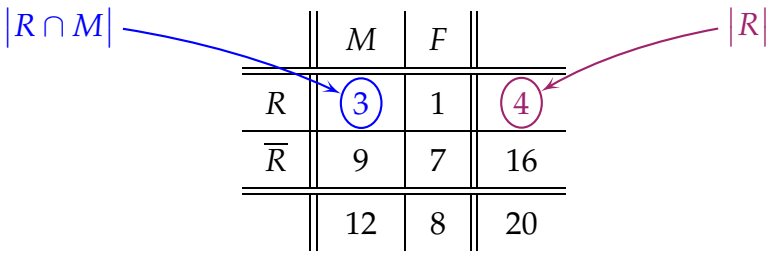
4.5.1. Einleitung, Definition

Wir wollen zuerst den Begriff der bedingten Wahrscheinlichkeit an einem Beispiel kennenlernen.

Beispiel 4.5.1

Wir betrachten eine Gruppe von 20 Personen, von denen einige Raucher bzw. Nichtraucher sind. Einige sind Frauen, der Resten Männer.

Eine Befragung der Personen ergibt die Daten in Tabelle 4.6.



	M	F	
R	3	1	4
$\overline{R}$	9	7	16
	12	8	20

Tabelle 4.6. : Raucher und Nichtraucher nach Geschlecht getrennt

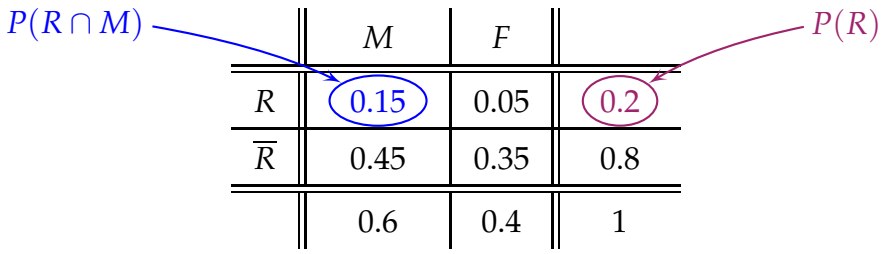
mit den Bezeichnungen

F: Frau, M: Mann, R: Raucher, $\overline{R}$: Nichtraucher

Somit hat es 4 Raucher und 16 Nichtraucher, 8 Frauen und 12 Männer. Die Werte in der Tabelle selbst haben folgende Bedeutung: Der Wert 3 links oben ist die Anzahl Personen, die Männer sind *und* rauchen. Wir können dies in der Mengenschreibweise auch wie folgt schreiben:

$$|R \cap M| = 3$$

Dividieren wir alle Werte in der Tabelle 4.6 oben durch 20, so erhalten wir relative Häufigkeiten (Tabelle 4.7).



	M	F	
R	0.15	0.05	0.2
$\overline{R}$	0.45	0.35	0.8
	0.6	0.4	1

Tabelle 4.7. : Wahrscheinlichkeiten von Raucher und Nichtrauchern nach Geschlecht getrennt

Die Werte in dieser Tabelle 4.7 haben folgende Bedeutung: Der Wert 0.15 links oben ist die Wahrscheinlichkeit, dass eine zufällig ausgewählte Person ein Mann ist und raucht. Diese Wahrscheinlichkeit berechnen wir wie folgt:

$$P(R \cap M) = \frac{|R \cap M|}{|\Omega|} = \frac{3}{20} = 0.15$$

Der Wert 0.2 links aussen entspricht der Wahrscheinlichkeit, dass eine zufällig ausgewählte Person ein Raucher ist. Also

$$P(R) = \frac{|R|}{|\Omega|} = \frac{4}{20} = 0.2$$

Nun betrachten wir nur einen Teil der Tabelle 4.7, nämlich die Raucher (siehe Tabelle 4.8):

	M	F	
R	0.15	0.05	0.2
$\overline{R}$	0.45	0.35	0.8
	0.6	0.4	1

Tabelle 4.8. : Wahrscheinlichkeiten der Männer unter den Rauchern

Wir können nun nach der Wahrscheinlichkeit fragen, dass eine zufällig ausgewählte Person *unter den Rauchern* ein Mann ist. Nach Tabelle 4.6 ist diese Wahrscheinlichkeit:

$$\frac{|R \cap M|}{|R|} = \frac{3}{4} = 0.75$$

Dasselbe Resultat lässt sich auch mit Tabelle 4.7 erreichen:

$$\frac{P(R \cap M)}{P(R)} = \frac{0.15}{0.20} = 0.75$$

Das heisst, dass 75 % der Raucher sind Männer. Oder: 75 % unter den Rauchern sind Männer.

Diese Wahrscheinlichkeit nennen wir *bedingte Wahrscheinlichkeit* und bezeichnen sie mit

$$P(M \mid R)$$

Der Begriff „bedingt" deutet an, dass wir nicht die gesamte Grundmenge betrachten, sondern nur einen *Teil* davon. Die neue Grundmenge hier sind die Raucher R. Dies ist in $P(M \mid R)$ die Variable *nach* dem Längsstrich.

Es gilt dann:

$$P(M \mid R) = \frac{P(R \cap M)}{P(R)} \qquad (*)$$

Diese Formel werden wir dann gleich für die Definition der bedingten Wahrscheinlichkeit verwenden.

Wir können nun natürlich auch die bedingte Wahrscheinlichkeit

$$P(R \mid M)$$

berechnen. Dies ist die Wahrscheinlichkeit, dass ein zufällig ausgewählter Mann ein Raucher ist. Wir können dies wieder in Tabelle 4.7 veranschaulichen, in der wir nur die Männer berücksichtigen (siehe Tabelle 4.9).

	M	F	
R	0.15	0.05	0.2
$\overline{R}$	0.45	0.35	0.8
	0.6	0.4	1

Tabelle 4.9. : Wahrscheinlichkeiten der Raucher unter den Männern

Um diese bedingte Wahrscheinlichkeit zu bestimmen, vertauschen wir in der Gleichung $(*)$ die Variable M durch R und erhalten

$$P(R \mid M) = \frac{P(M \cap R)}{P(M)} = \frac{P(R \cap M)}{P(M)} = \frac{0.15}{0.6} = 0.25$$

Die Wahrscheinlichkeit, dass ein zufällig ausgewählter Mann raucht, ist 0.25.

Wichtig ist die Beobachtung, dass die beiden bedingten Wahrscheinlichkeiten $P(M \mid R$ und $P(R \mid M)$ *verschieden* sind:

$$0.75 = P(M \mid R) \neq P(R \mid M) = 0.25$$

Der Anteil der Männer unter den Rauchern ist nicht gleich dem Anteil der Raucher unter den Männern. ◀

Die Überlegungen vom Beispiel 4.5.1 vorher verwenden wir für die Definition der bedingten Wahrscheinlichkeit.

Definition: Bedingte Wahrscheinlichkeit

Die *bedingte Wahrscheinlichkeit* ist die Wahrscheinlichkeit, dass das Ereignis A eintritt, wenn wir schon wissen, dass B eingetreten ist. Diese Wahrscheinlichkeit

wird mit

$$P(A \mid B)$$

bezeichnet. Der Längsstrich wird als „unter der Bedingung" gelesen.

Die bedingte Wahrscheinlichkeit $P(A \mid B)$ wird definiert durch

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

Die Interpretation ist folgendermassen: $P(A \mid B)$ ist die Wahrscheinlichkeit für das Ereignis A, wenn wir schon wissen, dass das Ereignis B eingetroffen ist.

Wie ist diese Formel zu verstehen? Da wir wissen, dass B schon eingetreten ist, haben wir einen neuen Grundraum

$$\Omega' = B$$

Damit müssen wir von A nur noch denjenigen Teil anschauen, der auch in B auftritt (daher $A \cap B$). Dies müssen wir jetzt noch in Relation zur Wahrscheinlichkeit von B, der neuen Grundmenge, setzen.

Das lässt sich anschaulich einfach im Venn-Diagramm in Abbildung 4.5 verdeutlichen.

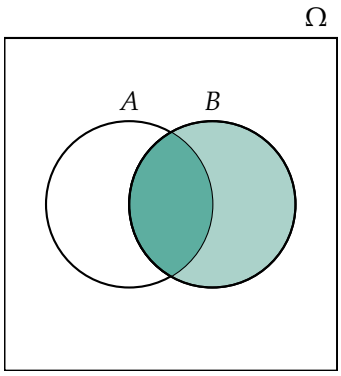


Abbildung 4.5. : Hilfsillustration für bedingte Wahrscheinlichkeiten.

Wenn wir die Wahrscheinlichkeiten wieder als Flächen in Abbildung 4.5 betrachten ($|\Omega| = 1$), dann ist die Wahrscheinlichkeit $P(A \cap B)$ der Flächeninhalt der dunkel gefärbten Fläche, während $P(B)$ der Flächeninhalt der gesamten gefärbten Fläche B ist. Die bedingte Wahrscheinlichkeit ist dann gerade der Anteil der dunkelgefärbten Fläche zur gesamten gefärbten Fläche, also

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

Bemerkungen:

i. In der Definition für die bedingte Wahrscheinlichkeit wird stillschweigend davon ausgegangen, dass $P(B) > 0$ gilt. Das muss auch so sein, da wir sonst eine Division durch 0 hätten, was nicht definiert ist.

ii. Für alle diejenigen, die bedingte Wahrscheinlichkeiten schon mit Hilfe von Baumdiagrammen kennengelernt haben: Beispiel 4.5.1 sieht mit einem Baumdiagramm wie folgt aus (siehe Abbildung 4.6):

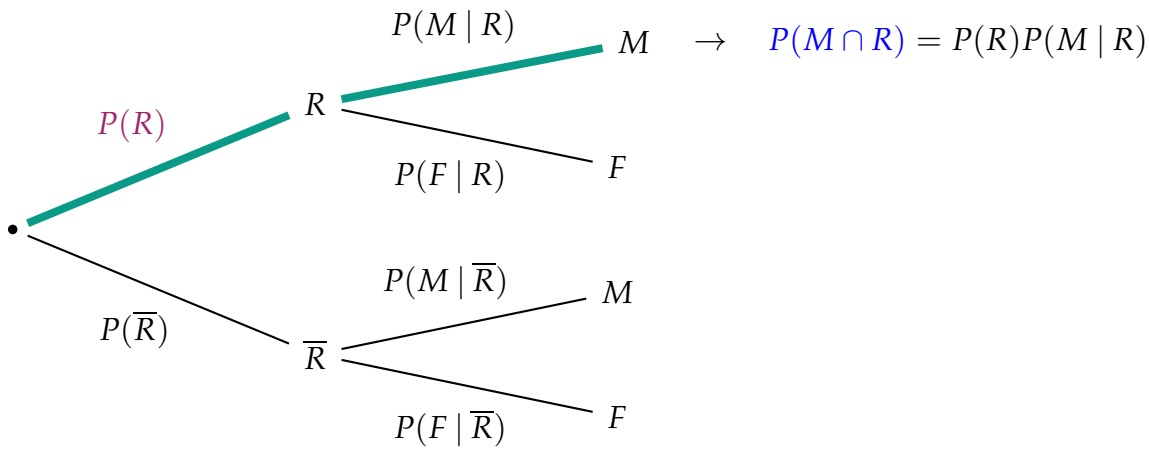


Abbildung 4.6. : Baumdiagramm zur bedingten Wahrscheinlichkeit

Wir werden Baumdiagramme im Weiteren aber nicht mehr verwenden. ♦

Bedingte Wahrscheinlichkeiten sind Wahrscheinlichkeiten für spezielle Situationen, wo der Grundraum eingeschränkt ist. Es gelten wieder entsprechende Rechenregeln.

Rechenregeln

$0 \leq P(A \mid B) \leq 1$	für jedes Ereignis A
$P(B \mid B) = 1$	für jedes Ereignis B
$P(\mid A_1 \cup A_2 \mid B) = P(A_1 \mid B) + P(A_2 \mid B)$	für A_1, A_2 disjunkt $(A_1 \cap A_2 = \{\})$
$P(\overline{A} \mid B) = 1 - P(A \mid B)$	für jedes Ereignis A

Wie in Beispiel 4.5.1 schon erwähnt, ist allgemein $P(A \mid B)$ *nicht* gleich $P(B \mid A)$:

Im Allgemeinen:

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

Diese Wahrscheinlichkeiten beschreiben völlig verschiedene Ereignissen, werden aber oft verwechselt. Diese Verwechslung wird vor Gericht (und nicht nur da) so oft gemacht, dass sie schon einen eigenen Namen hat: Prosecutor's fallacy.

Wir wollen die bedingte Wahrscheinlichkeit nun an einem realistischen Beispiel untersuchen, das ein überraschendes Ergebnis liefert.

Beispiel 4.5.2

Ein medizinischer Test auf eine seltene (tödliche) Krankheit soll feststellen, ob eine Person an dieser Krankheit erkrankt ist oder nicht.

Natürlich ist dieser Test nicht ganz genau. Manchmal zeigt er die Krankheit an, obwohl die Person gesund ist, oder er zeigt die Krankheit nicht an, obwohl die Person krank ist.

Uns interessiert folgende Frage: Sie gehen zum Arzt und machen diesen Test auf die Krankheit. Der Test ist positiv: Sie haben gemäss dem Test die Krankheit. Wie gross ist die Wahrscheinlichkeit, dass Sie auch wirklich diese Krankheit haben?

Um diese Frage zu beantworten, führen wir folgende Bezeichnungen ein:

- D: Krankheit ist vorhanden; $\overline{D}$: Krankheit ist nicht vorhanden
- $+$: Test zeigt Krankheit an; $-$: Test zeigt Krankheit nicht an

Die Wahrscheinlichkeiten in Tabelle 4.10 sind durch Versuche bekannt.

	D	$\overline{D}$	
$+$	0.009	0.099	0.108
$-$	0.001	0.891	0.892
	0.01	0.99	1

Tabelle 4.10. : Wahrscheinlichkeit für eine Krankheit.

Beispielsweise ist die Wahrscheinlichkeit, dass die Krankheit vorhanden ist *und* der Test positiv ausfällt

$$P(D \cap +) = 0.009$$

Diese Wahrscheinlichkeit ist recht klein. Der Grund dafür liegt darin, dass nur ein kleiner Prozentsatz der Bevölkerung diese Krankheit hat, wie wir gleich sehen werden.

Hier können wir verschiedene bedingte Wahrscheinlichkeiten berechnen:

- $P(+ \mid D)$: W'keit, dass ein Kranker auch wirklich positiv getestet wird
- $P(+ \mid \overline{D})$: W'keit, dass ein Gesunder fälschlicherweise positiv getestet wird
- $P(- \mid D)$: W'keit, dass ein Kranker fälschlicherweise negativ getestet wird
- $P(- \mid \overline{D})$: W'keit, dass ein Gesunder richtigerweise negativ getestet wird
- $P(D \mid +)$: W'keit, dass ein positiv Getesteter auch wirklich krank ist
- $P(D \mid -)$: W'keit, dass ein negativ Getesteter fälschlicherweise krank ist
- $P(\overline{D} \mid +)$: W'keit, dass ein positiv Getesteter fälschlicherweise gesund ist
- $P(\overline{D} \mid -)$: W'keit, dass ein negativ Getesteter wirklich auch gesund ist

Wir berechnen zuerst die Wahrscheinlichkeit $P(+ \mid D)$ wie folgt:

$$P(+ \mid D) = \frac{P(+ \cap D)}{P(D)} = \frac{0.009}{0.009 + 0.001} = 0.9$$

Dabei haben wir für $P(D)$ folgende Tatsache benützt:

$$P(D) = P(D \cap +) + P(D \cap -) = 0.009 + 0.001 = 0.01$$

Dies ist die Summe der Einträge in der Tabelle in der Spalte unter D in Tabelle 4.10. Die Kranken sind entweder positiv oder negativ getestet. Das heisst, 1 % der Bevölkerung hat diese Krankheit. Die Krankheit ist also nicht sehr weit verbreitet.

Die bedingte Wahrscheinlichkeit $P(- \mid \overline{D})$ ist die Wahrscheinlichkeit, dass eine gesunde Person auch wirklich negativ (gemäss Test ist die Krankheit nicht vorhanden) getestet wird:

$$P(- \mid \overline{D}) = \frac{P(- \cap \overline{D})}{P(\overline{D})} = \frac{0.891}{0.891 + 0.099} = 0.9$$

Dieser Test scheint recht genau zu sein. Kranke Personen werden zu 90 % als positiv eingestuft, und gesunde Personen werden zu 90 % als negativ eingestuft.

Dies ist aber *nicht* unsere Fragestellung. Wir wollen wissen, wie gross die Wahrscheinlichkeit ist, dass man krank ist, *wenn der Test positiv* ist. Die Wahrscheinlichkeiten oben geben an, wie gut der Test ist, *wenn man krank/gesund* ist.

Angenommen, Sie gehen zu einem Test, und dieser wird als positiv eingestuft. Wie gross ist die Wahrscheinlichkeit, dass Sie die Krankheit wirklich haben?

Die meisten Leute würden 0.9 antworten. Müssen Sie sich also grosse Sorgen machen und das Testament schreiben oder einer Sterbehilfeorganisation beitreten?

Schauen wir uns nun die *richtige* Antwort an, und zwar ist das bedingte Wahrscheinlichkeit $P(D \mid +)$:

$$P(D \mid +) = \frac{P(+ \cap D)}{P(+)} = \frac{0.009}{0.009 + 0.099} = 0.08$$

Was bedeutet nun dieses Resultat?

Die bedingte Wahrscheinlichkeit $P(D \mid +)$ ist die Wahrscheinlichkeit, dass man bei einem positiven Test auch wirklich krank ist. Diese beträgt aber nur 8 %. Man hat bei einem positiven Test also nur zu 8 % auch wirklich die Krankheit. Ein positiver Test sagt hier also sehr wenig darüber aus, ob man die Krankheit hat oder nicht.

Die Frage ist nun, warum ist dies so. Der Grund liegt darin, dass die Krankheit *selten* ist. Machen wir ein numerisches Beispiel:

Wir untersuchen 100 000 Personen. Dann erhalten wir folgende Zahlen:

- 1000 Personen haben die Krankheit (1 %)
- 90 % dieser Personen werden positiv getestet: 900 Personen
- 99 000 haben die Krankheit nicht
- 10 % dieser Personen werden positiv getestet: 9900 Personen
- Die Anzahl positiv getesteter ist

$$900 + 9900 = 10\,800$$

- Unter diesen positiv getesteten sind aber bei weitem mehr Gesunde, die fälschlicherweise positiv getestet wurden.
- Das heisst, dass die Wahrscheinlichkeit, dass eine positiv getestete Person auch wirklich krank ist, ist

$$\frac{900}{10\,800} = 0.0833$$

Wir haben also sehr viele Gesunde unter denen der Test fälschlicherweise positiv ausgefallen ist. Dies sind aber viel mehr als die Kranken, die positiv getestet wurden.

Das heisst, von allen positiven Tests sind die meisten fälschlicherweise positiv und somit hat ein positiver Test wenig Aussagekraft. ◀

Die Lektion ist: Bei bedingten Wahrscheinlichkeiten dürfen wir unserer Intuition nicht vertrauen, sondern wir müssen die entsprechenden Wahrscheinlichkeiten wirklich ausrechnen.

4.5.2. Theorem von Bayes und totale Wahrscheinlichkeit

Bayes' Theorem

Das *Theorem von Bayes* ist oft sehr nützlich. Es erlaubt uns die Wahrscheinlichkeit $P(A \mid B)$ zu berechnen, falls $P(B \mid A)$ bekannt ist.

Bayes Theorem

Zwischen $P(A \mid B)$ und $P(B \mid A)$ gibt es folgenden Zusammenhang:

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

Hier noch die Herleitung, für diejenigen, die es genauer wissen wollen.

Das Bayes Theorem erhalten wir durch zweimalige Anwendung der Definition der bedingten Wahrscheinlichkeit. Es gilt

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

und

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

Da

$$A \cap B = B \cap A$$

gilt auch

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

Und somit gilt auch

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

Dividieren wir beide Seiten durch $P(B)$ so erhalten wir die Formel von Bayes:

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

□

Beispiel 4.5.3

Das Bayes Theorem liefert die gleiche Lösung wie in Beispiel 4.5.2:

$$P(D \mid +) = \frac{P(+ \mid D)P(D)}{P(+)} = \frac{0.9 \cdot (0.009 + 0.001)}{0.009 + 0.099} = \frac{0.009}{0.009 + 0.099} = 0.08$$

◀

Totale Wahrscheinlichkeit

Ein weiterer nützlicher Begriff ist derjenige der *totalen Wahrscheinlichkeit*. Dabei wird eine Menge A zunächst in Mengen $A_1, \ldots, A_k$ unterteilt, die miteinander keine Schnittmenge haben und zusammen (Vereinigung) die ganze Menge A bilden. Eine solche Aufteilung nennen wir eine *Partitionierung*.

Beispiel 4.5.4

Eine mögliche Partitionierung für den Würfelwurf:

$$A_1 = \{1\}, \qquad A_2 = \{2, 4\}, \qquad A_3 = \{3, 5, 6\}$$

Es gilt also

$$A_1 \cap A_2 = \{\}; \quad A_1 \cap A_3 = \{\}; \quad A_2 \cap A_3 = \{\}$$

und

$$A_1 \cup A_2 \cup A_3 = A$$

◀

Es gilt dann das

Gesetz der totalen Wahrscheinlichkeit

Für die Partitionierung $A_1, \ldots, A_k$ und für jedes beliebige Ereignis B gilt:

$$\begin{aligned} P(B) &= P(B \mid A_1) \cdot P(A_1) + P(B \mid A_2) \cdot P(A_2) + \ldots + P(B \mid A_k) \cdot P(A_k) \\ &= \sum_{i=1}^{k} P(B \mid A_i) \cdot P(A_i) \end{aligned}$$

Für $k = 2$ gilt:

$$P(B) = P(B \mid A_1) \cdot P(A_1) + P(B \mid A_2) \cdot P(A_2)$$

Und für $k = 3$:

$$P(B) = P(B \mid A_1) \cdot P(A_1) + P(B \mid A_2) \cdot P(A_2) + P(B \mid A_3) \cdot P(A_3)$$

Für diejenigen, die es genauer wissen wollen, hier noch die Herleitung.

Warum gilt diese Formel? Wir betrachten nur den Fall $k = 2$.

In Abbildung 4.7 sind die Mengen A_1, A_2 und B eingezeichnet. Die Mengen A_1 und A_2 bilden eine Partition von Ω.

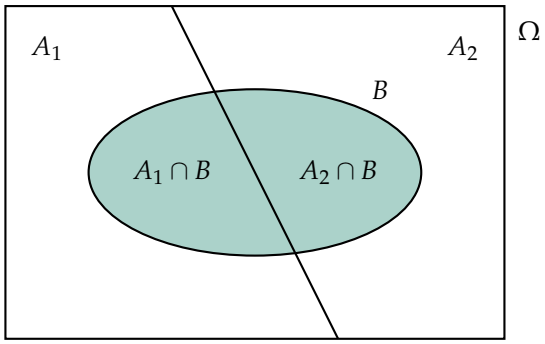


Abbildung 4.7. : Totale Wahrscheinlichkeit für $k = 2$

Es gilt also

$$A_1 \cup A_2 = \Omega \qquad \text{und} \qquad A_1 \cap A_2 = \{\}$$

Die Menge B wird in zwei Teile aufteilt, nämlich $A_1 \cap B$ (die gefärbte Fläche links der Trennungslinie) und $A_2 \cap B$ (die gefärbte Fläche rechts der Trennungslinie). Es gilt also:

$$B = (A_1 \cap B) \cup (A_2 \cap B)$$

Für die entsprechende Wahrscheinlichkeit gilt dann

$$P(B) = P((A_1 \cap B) \cup (A_2 \cap B))$$

Da nun

$$(A_1 \cap B) \cap (A_2 \cap B) = \{\}$$

gilt können wir Axiom A3 anwenden:

$$\begin{aligned} P(B) &= P((A_1 \cap B) \cup (A_2 \cap B)) \\ &= P(A_1 \cap B) + P(A_2 \cap B) \end{aligned}$$

Für die Summanden auf der rechten Seite können wir nun die Definition der bedingten Wahrscheinlichkeiten anwenden:

$$P(B \mid A_1) = \frac{P(A_1 \cap B)}{P(A_1)} \qquad \Rightarrow \qquad P(A_1 \cap B) = P(B \mid A_1)P(A_1)$$

Die entsprechende Formel gilt für $P(A_2 \cap B)$. Dies setzen wir in die Formel vorher ein und erhalten das Gesetz der totalen Wahrscheinlichkeit für $k = 2$:

$$\begin{aligned} P(B) &= P(A_1 \cap B) + P(A_2 \cap B) \\ &= P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) \end{aligned}$$

□

Beispiel 4.5.5

Ich teile meine Emails in drei Kategorien ein:

$$A_1 : \text{„spam”}, \qquad A_2 : \text{„niedrige Priorität”}, \qquad A_3 : \text{„hohe Priorität”}$$

Aus früheren Beobachtungen weiss ich, dass

$$P(A_1) = 0.7, \qquad P(A_2) = 0.2, \quad \text{und} \quad P(A_3) = 0.1$$

Es gilt

$$P(A_1) + P(A_2) + P(A_3) = 1$$

wie es bei einer Partitionierung auch sein sollte.

Sei B das Ereignis, dass das Wort „free” in der Email auftaucht. Dieses Wort kommt sehr oft in Spam-Mails vor, aber auch in den anderen. Von früheren Beobachtungen weiss ich, dass

$$P(B \mid A_1) = 0.9, \qquad P(B \mid A_2) = 0.01, \quad \text{und} \quad P(B \mid A_3) = 0.01$$

Hier ergibt die Summe nicht 1. Dies sind die Wahrscheinlichkeiten, mit denen das Wort „free” in den drei Mailkategorien vorkommt.

Angenommen, wir erhalten eine Email, die das Wort „free" enthält. Wie gross ist die Wahrscheinlichkeit, dass es sich um Spam handelt?

Das Bayes Theorem zusammen mit dem Gesetz der totalen Wahrscheinlichkeit liefert die Lösung.

Wir suchen die Wahrscheinlichkeit $P(A_1 \mid B)$.

Zuerst verwenden wir das Bayes Theorem:

$$P(A_1 \mid B) = \frac{P(B \mid A_1)P(A_1)}{P(B)}$$

Die Wahrscheinlichkeit $P(B)$ ist unbekannt, aber diese können wir mit der totalen Wahrscheinlichkeit berechnen:

$$P(A_1 B|) = \frac{P(B \mid A_1)}{P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + P(B \mid A_3)P(A_3)}$$

Auf der rechten Seite der Gleichung sind nun alle Grössen bekannt. Diese setzen wir ein und berechnen die gesuchte bedingte Wahrscheinlichkeit:

$$\begin{aligned} P(A_1 B|) &= \frac{P(B \mid A_1)P(A_1)}{P(B)} \\ &= \frac{P(B \mid A_1)}{P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + P(B \mid A_3)P(A_3)} \\ &= \frac{0.9 \cdot 0.7}{(0.9 \cdot 0.7) + (0.01 \cdot 0.2) + (0.01 \cdot 0.1)} \\ &= 0.995 \end{aligned}$$

Viele Spamfilter basieren tatsächlich auf diesem Prinzip: Die Mails werden nach Worten wie „free", „credit", etc. durchsucht, die häufig in Spam-Mails vorkommen, in den anderen aber eher nicht. ◀

Beispiel 4.5.6

Some[2] children are born with Down's syndrome. There are tests which can be given to pregnant women to see if their baby may suffer from this condition. A team at the University of Liverpool wanted to see how well test results were interpreted by

[2] From A. Jessop, Let the Evidence Speak, Springer

those involved; the pregnant women, their companions, midwives and obstetricians[3]. Eighty-five people were shown this scenario:

> The serum test screens pregnant women for babies with Down's syndrome. The test is a very good one but not perfect. Roughly 1 % of babies have Down's syndrome. If the baby has Down's syndrome, there is a 90 per cent chance that the result will be positive. If the baby is unaffected, there is still a 1 % chance that the result will be positive. A pregnant woman has been tested and the result is positive. What is the chance that her baby actually has Down's syndrome?

Read this again and write down your answer.

Table 4.11 shows how well the eighty-five people performed.

	correct	overestimate	underestimate	
pregnant women	1	15	6	22
companions	3	10	7	20
midwives	0	10	12	22
obstetricians	1	16	4	21
	5	51	29	85

Tabelle 4.11. : Diagnostic accuracy for Down's syndrome

Only five of the eighty-five gave the correct answer. The health professionals did no better that the pregnant women and their companions.

How about you?

Now, let's denote D as having Down's syndrom and $+$ having a positive test result. Given are the following probabilities:

$$P(D) = 0.01, \qquad P(+ \mid D) = 0.9, \qquad P(+ \mid \overline{D}) = 0.01$$

Applying Bayes's theorem and the law of total probability, it follows

$$\begin{aligned} P(D \mid +) &= \frac{P(+ \mid D) \cdot P(D)}{P(+ \mid \cdot D)P(D) + P(+ \mid \overline{D}) \cdot P(\overline{D})} \\ &= \frac{0.9 \cdot 0.01}{0.9 \cdot 0.01 + 0.01 \cdot 0.99} \\ &= 0.4761905 \end{aligned}$$

[3] Bramwell R, West H, Salmon P (2006) Health professionals' and service users' interpretation of screening test results: experimental study. Br Med J 333:284–286

With a positive test result there is a 48 % chance that the baby has Down's syndrome.

Interestingly, a different group of eighty-one people were shown this alternative scenario:

> The serum test screens pregnant women for babies with Down's syndrome. The test is a very good one but not perfect. Roughly 100 babies out of 10 000 have Down's syndrome. Of these 100 babies with Down's syndrome, 90 will have a positive test result. Of the remaining 9900 unaffected babies 99 will still have a positive test result. How many pregnant women who have a positive result to the test actually have a baby with Down's syndrome?

Table 4.12 shows the result.

	correct	overestimate	underestimate	
pregnant women	3	3	10	21
companions	3	8	9	20
midwives	0	7	13	20
obstetricians	13	3	4	20
	19	26	36	81

Tabelle 4.12. : Diagnostic accuracy for Down's syndrome using reformulated data

This is clearly an improvement. Why?

The reframing of the scenario makes it an easier problem. All that is required is to pick out the two numbers 90 and 99 and so get

$$\frac{90}{90+99} \approx 48\,\%$$

But still only about a quarter of the sample got the right answer. At least the obstetricians fared significantly better. ◀

4.6. Schlussbemerkungen zur Wahrscheinlichkeit

Nachdem wir uns, sehr oberflächlich, mit Wahrscheinlichkeiten beschäftigt haben, hier noch einige Bemerkungen.

Wie in der Einführung 4.1 schon erwähnt, ist es gar nicht so einfach zu sagen oder besser gesagt zu interpretieren, was exakt eine Wahrschlichkeit ist. Es wurden dicke Bücher über diese Frage geschrieben und die Meinungen gehen immer noch auseinander.

In der Einführung haben wir die Aussage „Es regnet morgen mit einer Wahrscheinlichkeit von 80 %" erwähnt. Solche Aussagen sind in Wetterberichten alltäglich geworden. Aber was bedeutet diese eigenlich?

Zunächst ist diese Aussage so schlichtweg sinnlos. Prozentzahlen beziehen sich immer auf irgendeinen Grundwert. Oder anders gefragt: „80 % von *was*?". Eine Prozentzahl nur für sich genommen, macht keinen Sinn. Keinen!

So dann die Frage, worauf sich diese 80 % beziehen . Es gibt nun zwei verschiedene Interpretationen:

- *Frequentistisch*: In 80 % von gleichen (oder ähnlichen) Wetterlagen hatte es morgen geregnet.
- *Bayes'sche*: Wir *glauben*, dass es morgen zu 80 % regnet. Das heisst, wenn wir wetten würden, glauben wir, dass die Gewinnwahrscheinlichkeit 0.8 ist.

Lange Zeit war die frequentistische Interpretation die allgemein akzeptierte, aber dies hat sich in den letzten Jahren geändert. Mittlerweile hat die Bayes'sche Interpretation, vor allem auch in Bezug auf Big Data, massiv an Zustimmung gefunden.

Eine Suchangrage „Bayes" (Feb 2020) auf **`Google`** ergab 39 500 000 hits.

Teil II.

Hypothesentest

Kapitel 5.

Normalverteilung

Everybody believes in the exponential law of errors [i.e., the Normal distribution]: the experimenters, because they think it can be proved by mathematics; and the mathematicians, because they believe it has been established by observation.

(E. T. Whittaker and G. Robinson)

5.1. Stetige Zufallsvariablen und Wahrscheinlichkeitsverteilungen

In vielen Anwendungen haben wir es nicht mit diskreten Zuvallsvariablen zu tun, sondern mit Messdaten, die sogenannt *stetig* sind. Diese können grundsätzlich jeden Wert in einem bestimmten Bereich annehmen, wobei die Genauigkeitsangabe des Messwertes durch die Messgenauigkeit vorgegeben wird.

5.1.1. Diskrete Wahrscheinlichkeitsverteilung

Bevor wir uns mit stetigen Zufallsvariablen und deren zugehörigen Wahrscheinlichkeitsverteilungen beschäftigen, repetieren wir an einem Beispiel eine diskrete Zufallsvariable, die wir dann auf eine stetige Zufallsvariable erweitern können.

Eine Zufallsvariable X ordnet jedem Zufallsexperiment *genau* eine Zahl zu. Wir können somit X auch als *Funktion* auffassen.

Beispiel 5.1.1

Die Zufallsvariable X ordnet einer zufällig ausgewählten, in der Schweiz lebenden Person, die Körpergrösse in cm zu. Die Körpergrösse wird also auf Zentimeter gerundet. Die *Definitionsmenge* dieser Zufallsvariable X ist dann die Menge der in der Schweiz lebenden Personen.

Die Zufallsvariable X kann nur folgende Werte annehmen:

$$W_X = \{0, 1, \ldots, \ldots, 500\}$$

Diese Menge heisst *Wertemenge* oder *Wertebereich*. Der Wertebereich wurde absichtlich zu gross gewählt, damit auch sicher alle vorkommenden Werte dabei sind.

Die Wertemenge dieser Zufallsvariable besteht also nur aus endlich vielen ganzen Zahlen. Eine solche Menge heisst *diskret*. Wichtig ist hier, dass wir *keinen* Wert zwischen zwei Werten der Wertemenge auswählen können, also z.B. 175.25 cm. Die Menge ist „löchrig". Dies kann man (sehr vereinfacht) als Definition von „diskret" auffassen.

Wir wählen nun zufällig (deshalb Zufallsvariable) eine Person aus. Diese Person heisst *Tabea*.

Wir nehmen hier an, dass jeder Name nur genau einmal vorkommen kann, was in der Realistät natürlich nicht der Fall ist. Wir hätten allerdings auch die AHV-Nummer wählen können, die eindeutig ist.

Tabea hat (auf cm gerundet) eine Körpergrösse von 166 cm. Mit der Zufallsvariable X (Funktion) können wir dies wie folgt formulieren

$$X(\text{Tabea}) = 166$$

Nun wählen wir zufällig eine weitere Person aus. Diese Person heisst *Tadeo* und hat eine Körpergrösse von 176 cm hat. Wir schreiben dann

$$X(\text{Tadeo}) = 176$$

Diese Zuordnung können wir mit jeder in der Schweiz lebenden Person machen.

Mit dem Ausdruck

$$X = 174$$

beschreiben wir das *Ereignis*, eine Person ausgesucht zu haben, die eine gerundete Körpergrösse von 174 cm hat. Wir sprechen hier auch von einer *Realisierung*

$$x = 174$$

der Zufallsvariable X.

Beachten Sie den Unterschied zwischen Gross- und Kleinschreibung:

- Der Ausdruck $x = 174$ beschreibt eine *Zahl*.
- Der Ausdruck $X = 174$ beschreibt eine *Menge* (der Personen mit gerundeter Körpergrösse von 174 cm)

Diesem Ereignis $X = 174$ können wir nun eine Wahrscheinlichkeit

$$P(X = 174)$$

zuordnen. Diese berechnet sich hier, indem wir die Anzahl Personen mit gerundeter Körpergrösse von 174 cm durch die Anzahl der in der Schweiz lebenden Personen dividieren. Hat die Schweiz eine Bevölkerung von 8 Mio und hat es 200 000 Personen mit gerundeter Körpergrösse von 174 cm, so ist die Wahrscheinlichkeit, dass eine zufällig ausgewählte Person, diese Körpergrösse hat

$$P(X = 174) = \frac{200\,000}{8\,000\,000} = 0.025$$

Auf diese Weise können wir *alle* Wahrscheinlichkeiten

$$P(X = x)$$

berechnen, wobei x jeden Wert in der Wertemenge annehmen kann. Insbesondere ist auch

$$P(X = 500) = 0$$

da es keine Person mit so einer Körpergrösse gibt (oder zumindest ist es sehr unwahrscheinlich). Aus diesem Grund spielt es auch keine Rolle, wenn wir die Wertemenge zu gross wählen.

Wir können weitere Wahrscheinlichkeiten bestimmen. So ist

$$P(X \leq 170)$$

die Wahrscheinlichkeit, dass eine zufällig ausgewählte Person eine gerundete Körpergrösse von 170 cm *oder weniger* hat. Beachten Sie, dass diese Wahrscheinlichkeit *nicht* der Wahrscheinlichkeit

$$P(X < 170)$$

entspricht. Dies ist die Wahrscheinlichkeit, dass eine zufällig ausgewählte Person eine gerundete Körpergrösse *kleiner als* 170 cm hat. Die Körpergrösse 170 cm gehört hier *nicht* dazu.

Berechnen wir die Wahrscheinlichkeiten für alle x in der Wertemenge, so erhalten wir eine *Wahrscheinlichkeitsverteilung*. Es gilt insbesondere

$$P(X = 0) + P(X = 1) + \ldots + P(X = 499) + P(X = 500) = 1$$

Jede Person *muss* eine Körpergrösse haben. ◀

Bemerkungen:

i. Nochmals, weil die Unterscheidung zwischen X und x sehr wichtig ist:
   - Die Variable X ist eine Funktion.
   - Die Variable x ist ein konkreter Wert (Realisierung) von X

ii. Berechnen wir konkret die Wahrscheinlichkeit

$$P(X = 174)$$

aus dem Beispiel vorher (Anzahl Personen mit dieser Körpergrösse dividiert durch die Anzahl aller Personen), so ändert sich diese leicht von Tag zu Tag, da Personen aus dieser Gruppe sterben, andere kommen hinzu.

Das heisst, diese Wahrscheinlichkeit bleibt zeitlich nicht konstant. Aber wir nehmen an, dass die änderungen so minim sind, dass die Unterschiede, zumindest über einen kurzen Zeitraum, vernachlässigbar sind. ⧫

Beispiel 5.1.2

Wir können nun zufällig[1] 1000 erwachsene Frauen auswählen, deren Körpergrösse messen (hier mit R simuliert) und ein Histogramm erstellen (siehe Abbildung 5.1 links).

Die *Form* dieses Histogrammes ist sehr typisch, sie kommt recht häufig vor. In der Mitte sind die Balken hoch und werden immer kleiner, je weiter sie von der Mitte entfernt sind.

Wir können auch versuchen eine Kurve einzuzeichnen, die dem Histogramm so gut wie möglich folgt (siehe Abbildung 5.1 rechts).

[1] Was das heisst, siehe *Design of Experiment* (DoE).

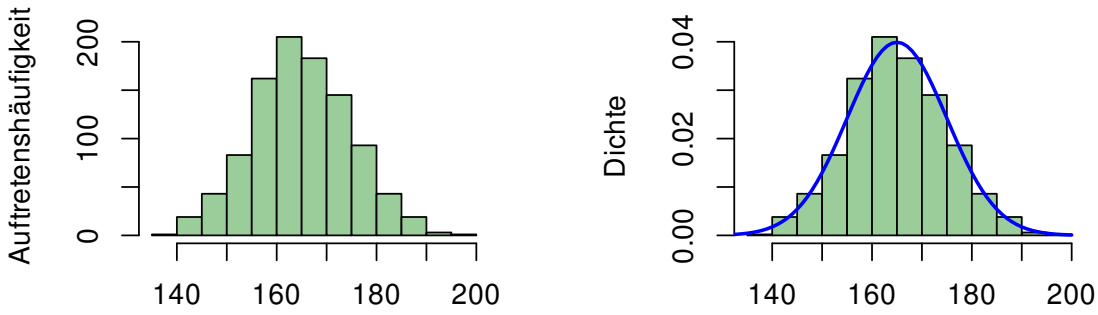


Abbildung 5.1. : Histogramm von der Körpergrösse von 1000 erwachsenen Frauen

Dazu tragen wir auf der vertikalen Achse die Dichten auf, sodass die Fläche des Histogrammes 1 wird.

Die blaue Kurve in Abbildung 5.1 rechts heisst *Normalverteilungskurve* und ist Gegenstand dieses Kapitels. ◀

5.1.2. Von diskreter zu stetiger Wahrscheinlichkeitsverteilung

Die blaue Kurve in Abbildung 5.1 rechts heisst *Wahrscheinlichkeitsdichtefunktion* und spielt für sogenannte stetige Verteilungen eine entscheidende Rolle.

Bevor wir uns mit stetigen Verteilungen befassen, wollen wir an einem Beispiel den übergang von einer diskreten zu einer stetigen Verteilungen illustrieren. Um dies zu erreichen, erweitern wir das Beispiel 5.1.2 und werden die wichtigen Eigenschaften der Dichtefunktion erläutern.

Beispiel 5.1.3

Wir simulieren die Körpergrösse von einer Million Personen mit **R**. Die Zufallsvariable X ordnet wieder jeder Person die Körpergrösse zu.

Zunächst messen wir die Körpergrösse auf 10 cm genau. Die Wertemenge lautet dann

$$\mathbb{W}_X = \{10, 20, \ldots, 490, 500\}$$

und enthält 51 Werte.

Die Daten sind im Histogramm in Abbildung 5.2 links dargestellt. Wir wählen dazu ein *normiertes* Histogramm, damit die Gesamtfläche der Balken gleich 1 ist.

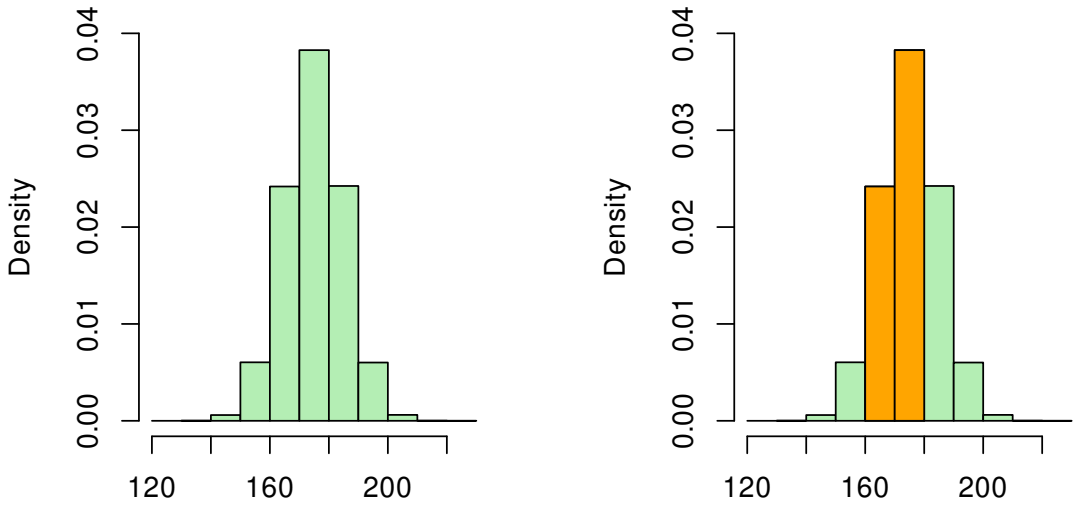


Abbildung 5.2. : Körpergrösse auf 10 cm gerundet

Nun betrachten wir die beiden orangen Balken von 160 cm bis 180 cm in Abbildung 5.2 rechts.

Die *Fläche* der beiden Balken spielt nun eine entscheidende Rolle. Das *Verhältnis* der orangen Fläche zur gesamten grünen Fläche entspricht gerade dem Anteil, dass eine zufällig ausgewählte Person eine Körpergrösse zwischen 160 cm und 180 cm hat.

Da aber die gesamte grüne Fläche 1 ist, ist diese orange Fläche gerade *gleich* dem Anteil, dass eine zufällig ausgewählte Person eine Körpergrösse zwischen 160 cm und 180 cm hat. Dieser Anteil ist aber nichts anderes als die entsprechende Wahrscheinlichkeit.

Wir wählen die Balkenbreite nun 5 cm (siehe Abbildung 5.3), das heisst die Körpergrössen sind auf 5 cm gerundet. Die Wertemenge lautet dann

$$\mathbb{W}_X = \{5, 10, \ldots, 495, 500\}$$

und besteht aus 100 Werten. Verkleinern wir die Balkenbreite, so vergrössern wir die Wertemenge.

Die Fläche der *vier* orangen Balken entspricht immer noch der Wahrscheinlichkeit, dass eine zufällig ausgewählte Person eine Körpergrösse zwischen 160 cm und 180 cm hat.

Die Fläche der vier orangen Balken *zusammen* ist immer noch *annähernd* dieselbe wie der zwei Balken in Abbildung 5.2 rechts, aber die Fläche der *einzelnen* Balken *nimmt ab*. Dies ist aber klar, da in Abbildung 5.2 rechts die Breite der Balken 10 cm war und sich in diesem Bereich mehr Körpergrössen befinden als in einem Bereich von 5 cm.

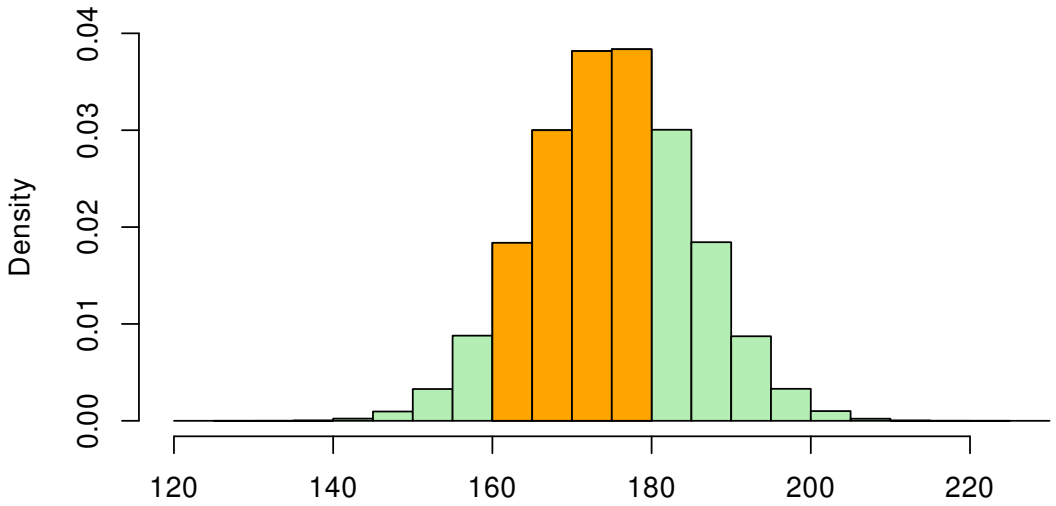


Abbildung 5.3. : Körpergrösse von 160 cm bis 180 cm auf 5 cm genau

Wir wählen die Balkenbreite noch kleiner als vorher. In Abbildung 5.4 ist das Histogramm mit Balkenbreite 0.5 cm dargestellt, das heisst, die Körpergrössen werden auf halbe Zentimeter genau gemessen. Die Wertemenge ist dann

$$\mathbb{W}_X = \{0.5, 1, 1.5, \ldots, 499.5, 500\}$$

und besteht aus 1000 Werten

Es gelten dieselben überlegungen wie für Abbildung 5.3. Die Fläche von den 40 orangen Balken entspricht der Wahrscheinlichkeit, dass eine zufällig ausgewählte Person eine Körpergrösse zwischen 160 cm und 180 cm hat.

Weiter fällt auf, dass die das Histogramm durch eine immer „glattere" Kurve begrenzt wird.

Nun machen wir den letzten Schritt: Wir lassen die Balkenbreite immer kleiner werden, das heisst, wir können die Körpergrösse beliebig genau messen. Schlussendlich erhalten wir eine „Balkenbreite" von 0 cm (siehe Abbildung 5.5) und eine unendlich grossen Wertemenge.

Das „Histogramm" folgt ist einer schön glatten Kurve. Die Interpretation, dass die orange Fläche der Wahrscheinlichkeit entspricht, dass eine zufällig ausgewählte Person eine Körpergrösse zwischen 160 cm und 180 cm hat, bleibt aber immer noch gültig.

Da die Balkenbreite 0 cm ist, ist auch die Fläche eines einzelnen „Balkens" gleich 0. Dieser Umstand können wir wie folgt interpretieren: Die Wahrscheinlichkeit, dass eine zufällig ausgewählte Person *exakt* die Körpergrösse von 173.459 621 24... cm hat, ist 0. Diese „Fläche" ist in Abbildung 5.5 gestrichelt eingezeichnet.

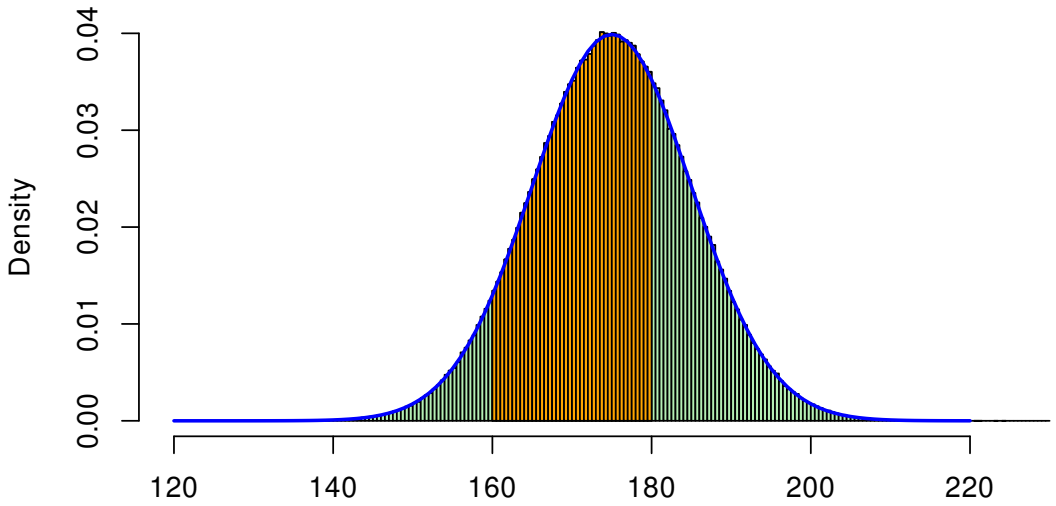


Abbildung 5.4. : Körpergrösse von 160 cm bis 180 cm auf 0.5 cm genau

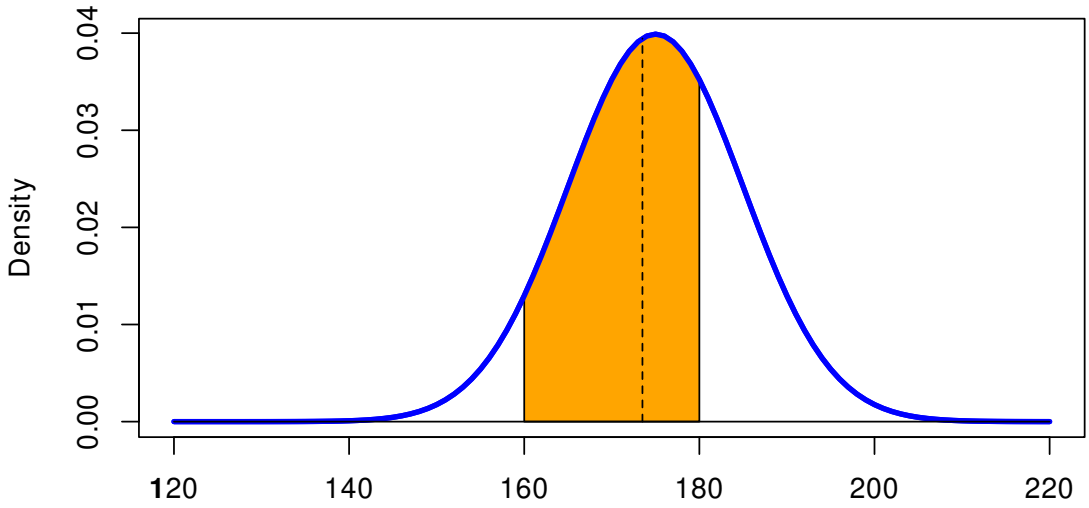


Abbildung 5.5. : Körpergrösse von 160 cm bis 180 cm beliebig genau

Abbildung 5.5 widerspiegelt den Fall einer sogenannten *stetigen* Verteilungsfunktion. Die Annahme ist, dass wir die Körpergrössen beliebig genau messen können. Die Wertemenge besteht aus *allen* Werte im *Bereich* oder *Intervall*

$$\mathbb{W}_X = [0,\ 500]$$

Wichtig ist die Unterscheidung zwischen diskreten und stetigen Wertemengen:

- Die Wertemenge
$$\{5, 10, \ldots, 500\}$$
hat nur endlich viele Elemente, sie ist diskret.

- Die Wertemenge

$$[0,\ 500]$$

besteht aus allen reellen Zahlen (Dezimalbrüchen), zwischen 0 und 500. Eine solche Menge heisst *stetig*.

Zufallsvariablen, deren Wertebereich stetig ist, heisst *stetige Zufallsvariable*. ◀

Bemerkungen:

i. Wir haben am Schluss des vorhergehenden Beispieles 5.1.3 angenommen, dass die wir die Körpergrössen beliebig genau messen können. Diese Annahme können wir praktisch nicht realisieren.

Warum beschäftigen wir uns dann mit stetigen Verteilungen? Es zeigt sich, dass stetige Verteilungen mathematisch viel einfacher zu handhaben sind als diskrete. Können wir *genügend* genau messen, so können wir als *Modell* eine stetige Verteilung anstatt einer diskreten verwenden. Diese Vorgehen erweist sich oft als zweckmässig. ♦

Die blaue Kurve in Abbildung 5.5 heisst *Wahrscheinlichkeitsdichtefunktion* und spielt für stetige Verteilungen eine zentrale Rolle.

Wir werden die überlegungen des Beispieles 5.1.3 nun verallgemeinern.

5.1.3. Stetige Verteilungen

Für den Resten dieses Kapitels beschäftigen wir uns mit Messgrössen, die wir beliebig genau messen können und die *nicht* gerundet werden.

Der *Wertebereich* W_X einer Zufallsvariablen X ist die Menge aller Werte, die X annehmen kann. Eine Zufallsvariable X heisst *stetig*, wenn deren Wertebereich W_X kontinuierlich (stetig) ist. Kontinierlich steht für *zusammenhängend* oder *lückenlos*.

Kontinierliche Wertebereiche sind Ausschnitte aus der Zahlengerade und haben keine „Löcher".

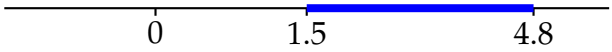


Abbildung 5.6. : Ausschnitte aus der Zahlengeraden

Beispiel 5.1.4

Abbildung 5.6) zeigt das Intervall

$$[1.5,\ 4.8]$$

Intervalle werden mit Klammern bezeichnet: die erste Zahl ist der Anfangspunkt, die zweite der Endpunkt.

◀

Beispiele für wichtige kontinuierliche Wertebereiche sind

$$W_X = \mathbb{R},\ \mathbb{R}^+ \quad \text{oder} \quad [0,1]$$

Dabei ist

- $\mathbb{R}$: Alle reellen Zahlen (Dezimalbrüche, Zahlengerade)
- $\mathbb{R}^+$: Alle positiven reellen Zahlen
- $[0,1]$: Alle reellen Zahlen zwischen 0 und 1, wobei die Grenzen inbegriffen sind

Konvention bei der Klammernotation:

Bei einer runden Klammer liegt der Wert ausserhalb des Intervalls, bei einer eckigen Klammer liegt der Wert innerhalb des Intervalls.

Beispiel 5.1.5

Das Intervall

$$[1.5,\ 4.8]$$

enthält alle reellen Zahlen zwischen 1.5 und 4.8, die beiden Grenzen 1.5 und 4.8 inbegriffen (siehe Abbildung 5.6).

Das Intervall

$$[1.5,\ 4.8)$$

enthält alle reellen Zahlen zwischen 1.5 und 4.8. Die Grenze 1.5 gehört zu diesem Intervall, aber nicht 4.8.

Wie sieht dieses Intervall graphisch aus? Diese Intervall unterscheidet sich vom Intervall $[1.5, 4.8]$ nur um einen einzigen Punkt von unendlich vielen Punkten. Somit sieht $[1.5, 4.8)$ ebenfalls aus wie Abbildung 5.6.

Beim Intervall

$$(1.5,\ 4.8]$$

gehört 1.5 nicht zum Intervall, 4.8 aber schon. Auch dies entspricht der Abbildung 5.6, wobei der Punkt am linken Rand des Intervalles nicht dabei ist.

Beim Intervall

$$(1.5,\ 4.8)$$

gehören beide Grenzen nicht zum Intervall.

Wir werden allerdings sehen, dass diese Unterscheidungen in der Stochastik nicht relevant sind. ◀

In Beispiel 5.1.1 haben wir gesehen, dass die Wahrscheinlichkeitsverteilung einer diskreten Zufallsvariablen beschrieben werden kann, indem wir die sogennannten *Punktwahrscheinlichkeiten*

$$P(X = x)$$

für alle möglichen Werte x im Wertebereich angeben.

Beispiel 5.1.6

In Beispiel 5.1.1 haben wir

$$P(X = 174) = \frac{200\,000}{8\,000\,000} = 0.025$$

berechnet. ◀

Für eine stetige Zufallsvariable X gilt jedoch:

$$P(X = x) = 0$$

für alle $x \in W_X$.

Beispiel 5.1.7

Wir messen wieder die Körpergrösse von Personen, nur gehen wir von nun an davon aus, dass wir die Körpergrösse *beliebig genau* messen können. Die Zufallsvariable X ordnet wieder jeder Person die zugehörige Körpergrösse zu.

Die Wahrscheinlichkeit, eine Körpergrösse von *genau* 182.254 680 895 434 ... cm zu messen, ist gleich 0:

$$P(X = 182.254\,680\,895\,434\ldots) = 0$$

Insbesondere gilt auch im Gegensatz zu Beispiel 5.1.6

$$P(X = 174) = P(X = 174.000\,000\,00\ldots) = 0$$

Diese Punktwahrscheinlichkeiten bringen uns also nichts zur Beschreibung von Wahrscheinlichkeiten für stetige Verteilungen.

Was für Wahrscheinlichkeiten können wir aber im Zusammenhang von Körpergrössen angeben, die nützlich sind?

Wir können die Wahrscheinlichkeit angeben, dass ein Messwert in einem bestimmten *Bereich* liegt, wie z.B. *zwischen* 174 und 175 cm:

$$P(174 < X \leq 175)$$

Diese Wahrscheinlichkeit ist dann *nicht* mehr 0.

Bemerkungen:

i. Wir haben gesehen, dass

$$P(X = 174) = 0 \qquad \text{und} \qquad P(X = 175) = 0$$

Somit spielt es keine Rolle, ob wir diese Grenzen zum Intervall hinzunehmen oder nicht. Es gilt also beispielsweise

$$P(174 < X \leq 175) = P(174 \leq X \leq 175)$$

♦

◀

Wir brauchen für stetige Verteilungen ein neues Konzept anstatt der Punktwahrscheinlichkeit, nämlich die sogenannte *Wahrscheinlichkeitsdichte*, die wir schon in Beispiel 5.1.3 kurz kennengelernt haben.

5.1.4. Wahrscheinlichkeitsdichte

Wahrscheinlichkeitsdichten können unter den unten genannten Einschränkungen fast beliebig aussehen. Wir werden allerdings nur die Normalverteilung und die dazu ver-

wandte t-Verteilung genauer untersuchen.

Eigenschaften Wahrscheinlichkeitsdichte

Für eine Wahrscheinlichkeitsdichte $f(x)$ gelten folgende Eigenschaften (siehe Abbildung 5.7):

1. Es gilt
$$f(x) \geq 0$$
Das heisst, die Kurve liegt auf oder überhalb der x-Achse
2. Die Wahrscheinlichkeit
$$P(a < X \leq b)$$
entspricht der Fläche zwischen a und b unter $f(x)$.
3. Die gesamte Fläche unter der Kurve ist 1.
Dies ist die Wahrscheinlichkeit, dass *irgendein* Wert gemessen wird.

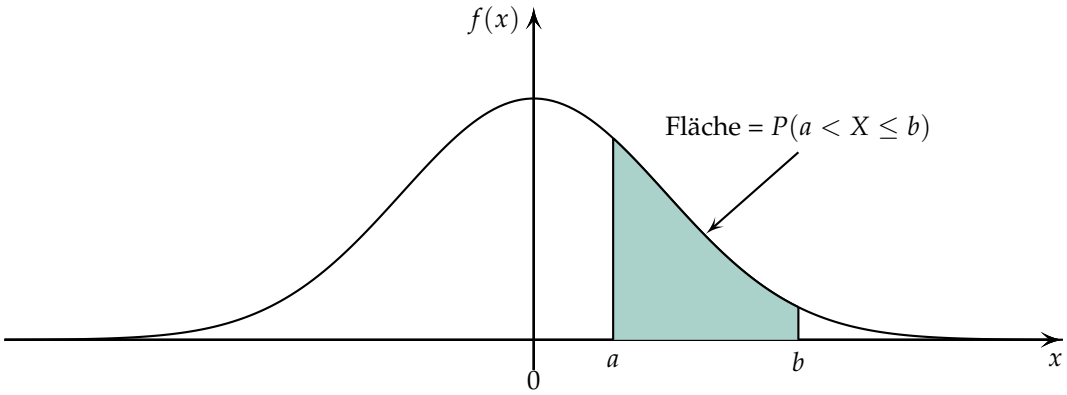


Abbildung 5.7. : Illustration einer Dichte einer Zufallsvariablen und der Wahrscheinlichkeit, in das Intervall $(a, b]$ zu fallen (grüne Fläche).

Für stetige Wahrscheinlichkeitsverteilungen ist der Zusammenhang zwischen Wahrscheinlichkeit und Flächen wichtig:

Merkregel

Für stetige Wahrscheinlichkeitsverteilungen entsprechen Wahrscheinlichkeiten Flächen unter der Dichtefunktion.

Bemerkungen:

i. Wahrscheinlichkeitsdichtefunktionen müssen nicht „schön" aussehen (siehe Abbildung 5.8), aber wir werden es ausschliesslich mit „schönen" Dichtefunktionen zu tun haben.

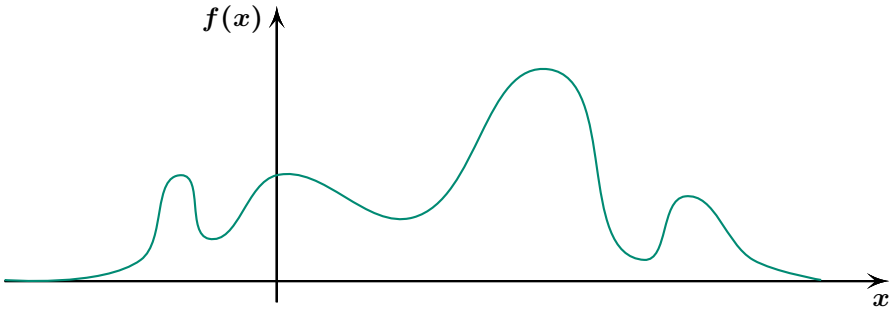


Abbildung 5.8. : Dichtefunktion mit einer „unregelmässigen" Form

♦

5.1.5. Quantile

Bei stetigen Verteilungen ist das α-Quantil q_α derjenige Wert, wo die Fläche (Wahrscheinlichkeit) unter der Dichtefunktion von $-\infty$ bis q_α gerade α entspricht (siehe Abbildung 5.9).

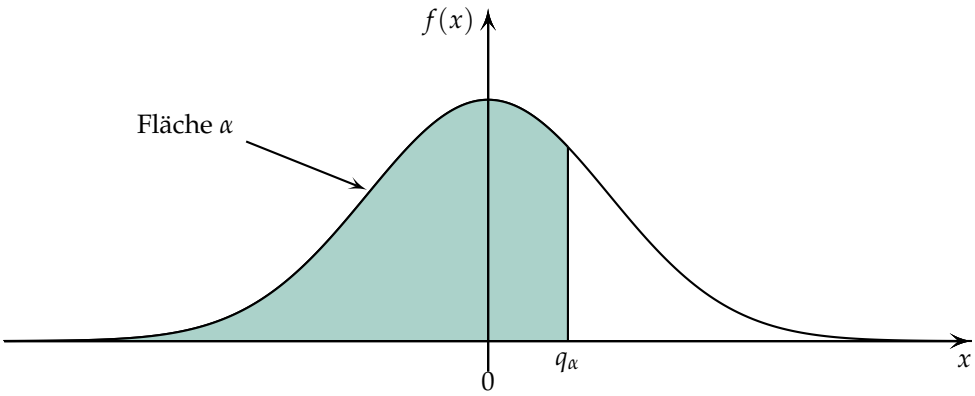


Abbildung 5.9. : Illustration des Quantils q_α anhand der Dichte $f(x)$ für $\alpha = 0.75$.

Das 50 %-Quantil heisst der *Median*.

Beispiel 5.1.8

Wir betrachten wiederum die Verteilung von Körpergrössen. Ist beispielsweise für $\alpha = 0.75$ das zugehörige Quantil gegeben durch

$$q_\alpha = q_{0.75} = 182.5$$

so bedeutet dies, dass 75 % der gemessenen Personen eine Körpergrösse von 182.5 cm oder weniger haben. ◀

5.1.6. Kennzahlen von stetigen Verteilungen

Der *Erwartungswert* $\mathrm{E}(X)$ und die *Standardabweichung* σ_X einer stetigen Zufallsvariablen X werden gleich interpretiert wie im diskreten Fall:

- Der Erwartungswert $\mathrm{E}(X)$ beschreibt die mittlere Lage der Verteilung.
- Die Standardabweichung σ_X beschreibt die Streuung der Verteilung um den Erwartungswert.

Der Erwartungswert $\mathrm{E}(X)$ und σ_X sind ähnlich definiert wie im diskreten Fall, aber mit Integralen anstatt Summen.

Wir werden die exakte Definition für den Erwartungswert und die Standardabweichung einer stetigen Wahrscheinlichkeitsverteilung nicht brauchen, sondern ausschliesslich die Interpretation dieser Werte. Der Vollständigkeit halber sind die Definitionen aufgeführt.

Erwartungswert und Varianz

Erwartungswert und *Varianz* sind wie folgt definiert:

$$\mathrm{E}(X) = \mu_X = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x$$

$$\mathrm{Var}(X) = \sigma_X^2 = \mathrm{E}((X - \mathrm{E}(X))^2) = \int_{-\infty}^{\infty} (x - \mathrm{E}(X))^2 f(x)\,\mathrm{d}x$$

□

5.2. Normalverteilung (Gaussverteilung)

Die *Normalverteilung*, auch *Gauss-Verteilung* genannt[2], ist die häufigste Verteilung für Messwerte und spielt vor allem für Durchschnitte von Messwerten (siehe Zentraler Grenzwertsatz in Unterabschnitt 5.3.3) eine sehr wichtige Rolle. Die Normalverteilung tritt in vielen Anwendungen auf und ist die wichtigste Wahrscheinlichkeitsverteilung in der Statistik überhaupt. Sie hat neben praktischer auch grosse theoretische Bedeutung.

Wir werden die exakte Definition der Normalverteilungsfunktion nicht brauchen, sondern nur deren Graphen, mit der wir Wahrscheinlichkeiten graphisch interpretieren können. Berechnungen werden wir sowieso mit **R** machen.

Man sollte die Definition der Normalverteilung einmal im Leben schon gesehen haben:

Normalverteilung

Eine Zufallsvariable X mit Wertebereich $W_X = \mathbb{R}$ heisst *normalverteilt* mit Parametern $\mu \in \mathbb{R}$ und $\sigma^2 \in \mathbb{R}^+$ falls

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

□

Wir verwenden folgende Schreibweise für die Verteilung einer normalverteilten Zufallsvariable X mit den Parametern μ und σ:

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

Für $X \sim \mathcal{N}(\mu, \sigma^2)$ lauten die Kennzahlen wie folgt (ohne Begründung):

$$\mathrm{E}(X) = \mu$$

$$\mathrm{Var}(X) = \sigma^2$$

$$\sigma_X = \sigma$$

Das heisst, die Parameter μ und σ^2 haben eine natürliche Interpretation als Erwartungswert und Varianz der Verteilung.

[2] Johann Carl Friedrich Gauss (30 April 1777 – 23 February 1855), *sehr* bedeutender deutscher Mathematiker

Beispiel 5.2.1

Am Schluss von Beispiel 5.1.3 haben wir eine Normalverteilungskurve mit $\mu_X = 175$ und $\sigma_X = 10$ gesehen. Sie ist in Abbildung 5.10 nochmals aufgeführt.

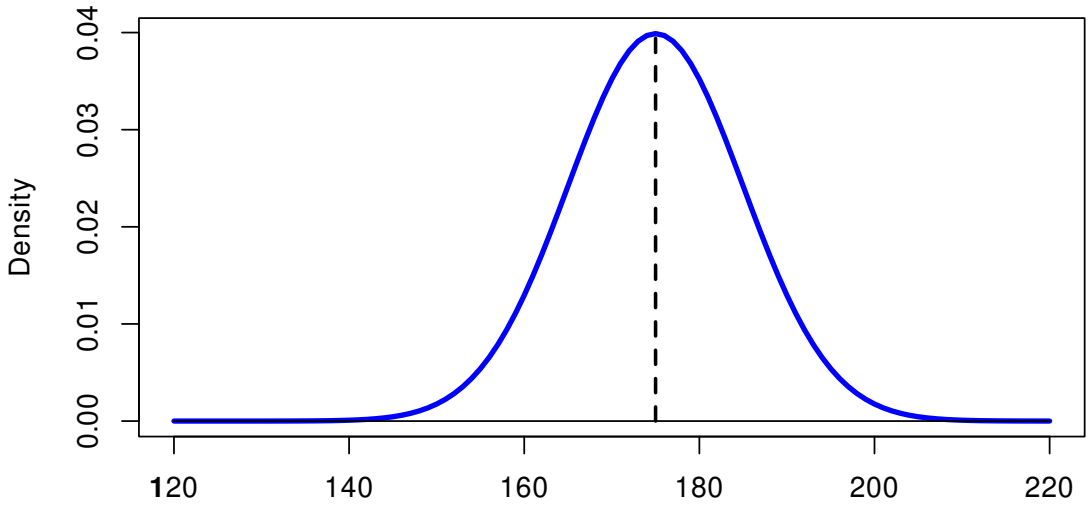


Abbildung 5.10. : Normalverteilungskurve mit $\mu = 175$ und $\sigma = 10$

Die Zufallsvariable X beschreibt hier wieder die Körpergrösse einer zufällig ausgewählten Person und wir schreiben

$$X \sim \mathcal{N}(175, 10^2)$$

Wir werden die Graphen von Normalverteilungskurven gleich genauer untersuchen, aber wir können hier schon jetzt feststellen, dass der „Buckel" der Kurve gerade an der Stelle $\mu_X = 175$ liegt. ◀

5.2.1. Graphische Darstellung der Normalverteilung

In Beispiel 5.2.1 haben wir eine erste Normalverteilungskurve gesehen. Wir wollen nun untersuchen, welche Einfluss die Parameter μ und σ auf die Normalverteilungskurve haben.

Alle Normalverteilungskurven haben folgende Eigenschaften:

- Die Kurven siehen wie eine „Glocken" aus. Wir sprechen deshalb bei Normalverteilungskurven auch von *Glockenkurven* oder *glockenförmigen Kurven*.
- Die Wahrscheinlichkeitsdichtefunktion der Normalverteilung ist symmetrisch um den Erwartungswert μ (siehe Abbildung 5.11).

- Die grüne Kurve in Abbildung 5.11 hat die Parameter $\mu = 0$ und $\sigma = 1$. Mit μ verschieben wir einfach den Graphen dieser Dichtefunktion nach links bzw. rechts. Ist $\mu = -4$, so verschieben wir die grüne Kurve um 4 Einheiten nach links und wir erhalten die blaue Kurve.

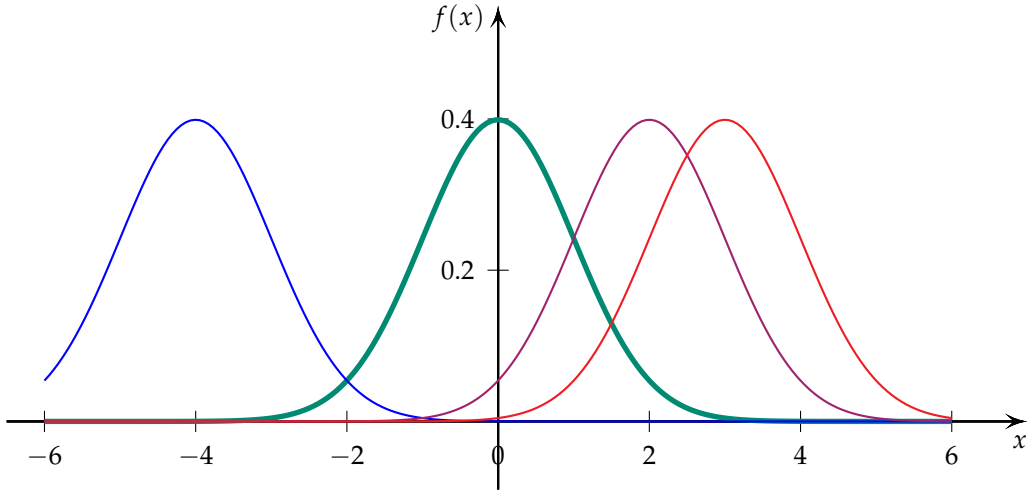


Abbildung 5.11. : Dichten der Normalverteilungen für $\mu = 0$ (grün), $\mu = -4$ (blau), $\mu = 2$, (violett) und $\mu = 3$ (rot). Für alle Kurven ist $\sigma = 1$.

- Je grösser σ, desto flacher oder breiter wird die Dichtekurve. Für ein σ nahe bei Null gibt es einen „schmalen und hohen" Gipfel (siehe Abbildung 5.12).

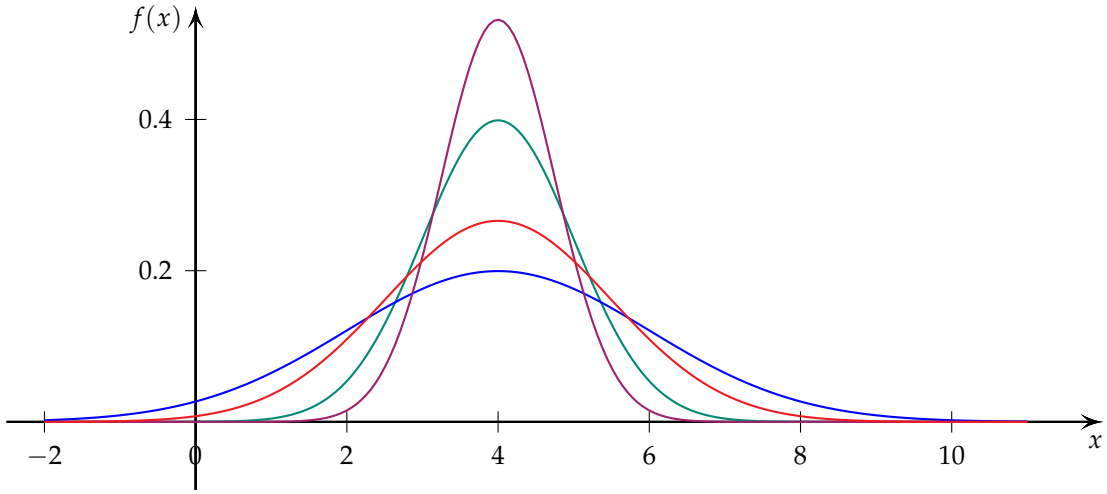


Abbildung 5.12. : Dichten der Normalverteilungen für $\sigma = 1$ (grün), $\sigma = 2$ (blau), $\sigma = 0.75$ (violett) und $\sigma = 1.5$ (rot). Für alle Kurven ist $\mu = 4$.

Warum ist dies so? Die Standardabweichung σ gibt die Streuung um den Erwartungswert μ an. Je grösser σ ist, um so mehr sind die Werte um den Erwartungswert μ verteilt. Die Kurve wird also breiter.

Ist σ nahe bei 0, so weichen die Werte „durchschnittlich" wenig von μ ab. Die Kurve wird also schmaler und höher.

Beispiel 5.2.2

Der Intelligenzquotient (IQ) wird in der Regel mit Intelligenztests ermittelt. Diese IQ-Tests werden so konstruiert, dass die Ergebnisse in etwa einer Normalverteilung mit Mittelwert 100 und Standardabweichung 15 folgen.

Wir bezeichnen mit X die Zufallsvariable den IQ einer zufällig ausgewählten Person und dann ist X normalverteilt mit $\mu = 100$ und $\sigma = 15$. Wir schreiben

$$X \sim \mathcal{N}(100, 15^2)$$

Es sei nochmals darauf hingewiesen, dass *Wahrscheinlichkeiten für stetige Verteilungen Flächen unter der Dichtekurve entsprechen*. Wir können also bei folgenden Berechnungen, die gesuchten Grössen graphisch darstellen. Dies gibt oft einen Hinweis auf den Lösungsweg.

1. Im Allgemeinen gilt eine Person als hochbegabt, wenn ihr IQ zwei und mehr Standardabweichungen vom Mittelwert nach oben entfernt ist.

   Wir suchen die Wahrscheinlichkeit, dass jemand hochbegabt ist, also einen IQ hat, der 130 oder grösser ist, also

   $$P(X > 130)$$

   Diese Wahrscheinlichkeit können wir wie schon erwähnt als Fläche darstellen, die in Abbildung 5.13 grün eingezeichnet ist.

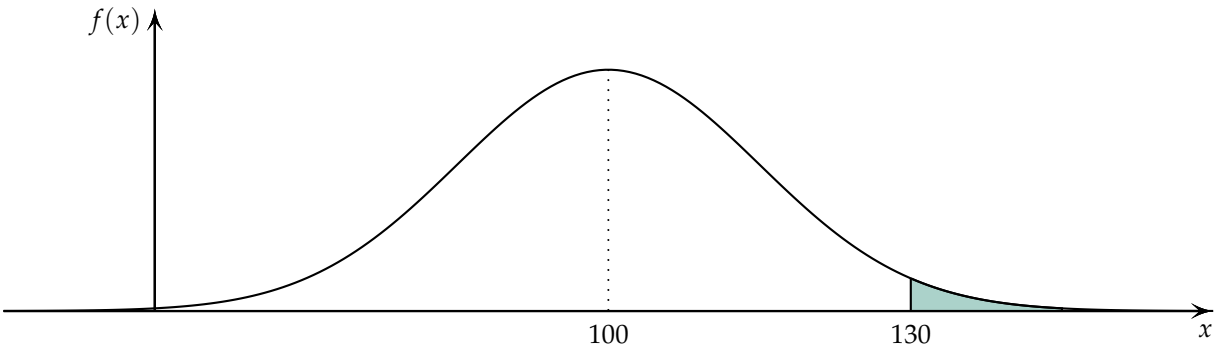


Abbildung 5.13. : Wahrscheinlichkeit $P(X > 130)$

Für die Berechnung von $P(X > 130)$ verwenden wir den R-Befehl **pnorm(...)**. Dieser berechnet *nicht* die gesuchte Wahrscheinlichkeit, aber

$$P(X \leq 130)$$

Beachten Sie die Richtung des Ungleichheitszeichens (siehe Abbildung 5.14)!

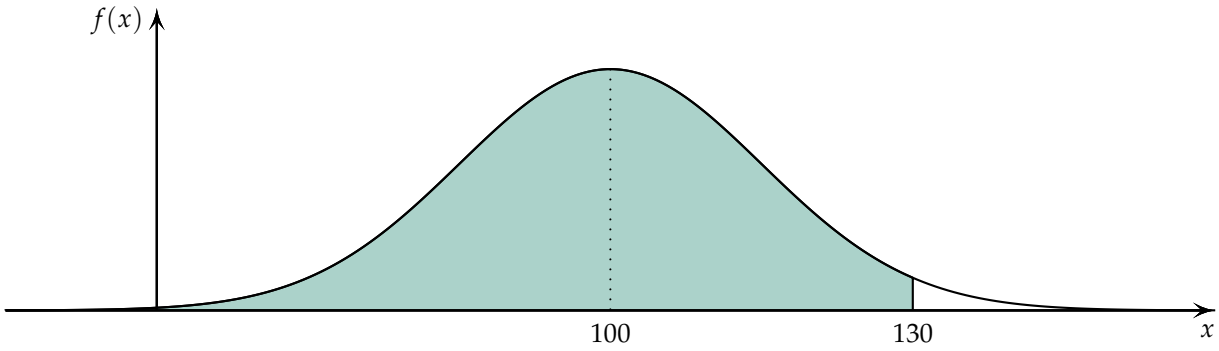


Abbildung 5.14. : Wahrscheinlichkeit $P(X \leq 130)$

```
pnorm(q = 130, mean = 100, sd = 15)

## [1] 0.9772499
```

Dieser Befehl ermittelt die Fläche (Wahrscheinlichkeit) von $-\infty$ bis **q = 130** unter der Normalverteilungskurve mit $\mu = 100$ und $\sigma = 15$.

Dies ist aber *nicht* unsere gesuchte Wahrscheinlichkeit $P(X > 130)$. Aber wir können diese Wahrscheinlichkeit einfach aus $P(X \leq 130)$ berechnen. Da die Gesamtfläche unter der Kurve 1 ergeben muss, können wir unsere gesuchte Wahrscheinlichkeit wie folgt schreiben:

$$P(X > 130) = 1 - P(X \leq 130)$$

Mit R erhalten wir:

```
1 - pnorm(q = 130, mean = 100, sd = 15)

## [1] 0.02275013
```

Also rund 2 % der Bevölkerung ist hochbegabt.

2. Welches Intervall enthält 95 % der IQ's um den Mittelwert $\mu = 100$?

   Auch hier stellen wir diese Wahrscheinlichkeit als Fläche unter der Dichtekurve dar (siehe Abbildung 5.15).

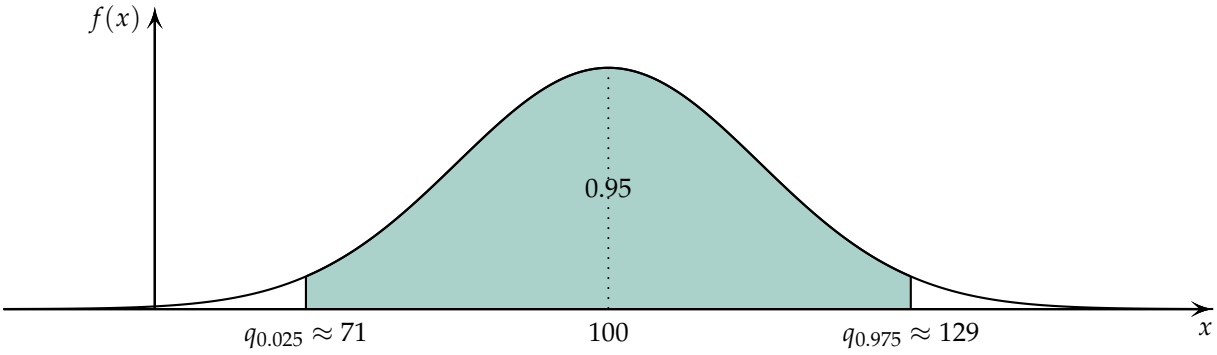


Abbildung 5.15. : Quartile für 95 % der Fläche um 100

Die grüne Fläche in der Mitte der Abbildung 5.15 nimmt 95 % der Gesamtfläche ein. Die kleinen weissen Flächen links und rechts haben jeweils einen Flächeninhalt von 0.025.

Die Wahrscheinlichkeiten sind also gegeben und wir suchen die zugehörigen Werte. Diese entsprechen nichts anderem als der Bestimmung der Quantile $q_{0.025}$ und $q_{0.975}$.

Der **R**-Befehl **qnorm()** bestimmt die Quantile für die Normalverteilung.

```
qnorm(p = 0.025, mean = 100, sd = 15)

## [1] 70.60054

qnorm(p = 0.975, mean = 100, sd = 15)

## [1] 129.3995
```

Oder kürzer:

```
qnorm(p = c(0.025, 0.975), mean = 100, sd = 15)

## [1]  70.60054 129.39946
```

Also haben 95 % der Menschen einen IQ zwischen (ungefähr) 70 und 130. Diese Werte entsprechen einem Abstand von etwa 2 Standardabweichungen vom Mittelwert $\mu = 100$.

3. Wie viel Prozent der Bevölkerung liegen innerhalb *einer* Standardabweichung vom Mittelwert?

   Wir suchen also die Wahrscheinlichkeit

$$P(85 \leq X \leq 115)$$

Auch diese Wahrscheinlichkeit stellen wieder als Fläche unter der Dichtekurve dar (siehe Abbildung 5.16).

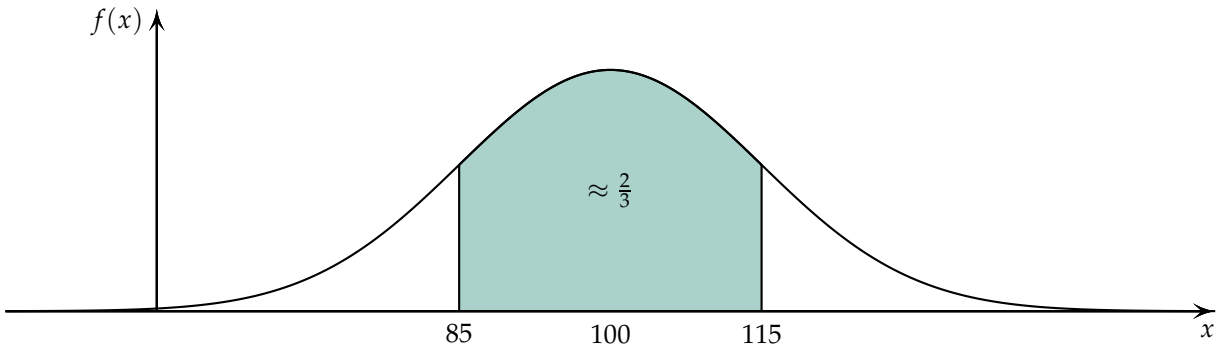


Abbildung 5.16. : Wahrscheinlichkeit für IQ zwischen $\mu \pm \sigma$

Um die Wahrscheinlichkeit $P(85 \leq X \leq 115)$ mit **R** berechnen zu können, müssen wir diese noch umschreiben

$$P(85 \leq X \leq 115) = P(X \leq 115) - P(X \leq 85)$$

Diese Gleichung ist in der Abbildung 5.17 veranschaulicht. Die gesamte Fläche links von 115 ist $P(X \leq 115)$. Verglichen zu Abbildung 5.16 ist die rote Fläche noch zuviel, die wir noch subtrahieren müssen. Die ist $P(X \leq 85)$.

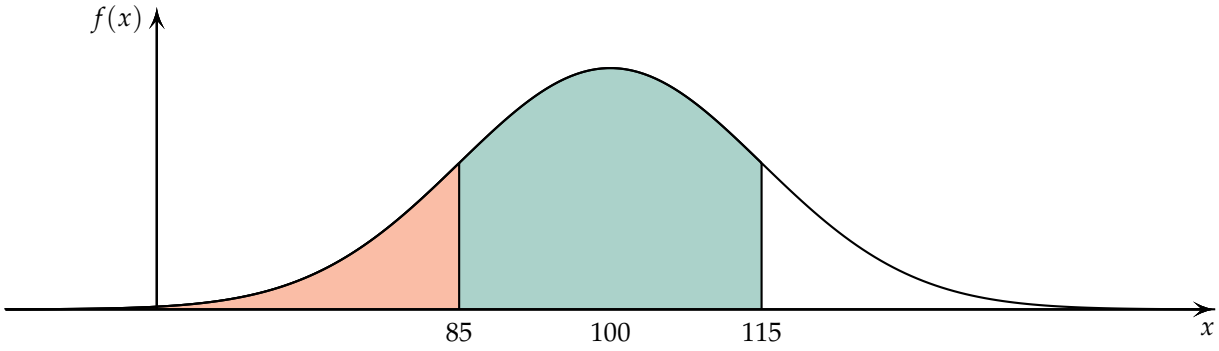


Abbildung 5.17. : Veranschaulichung $P(85 \leq X \leq 115) = P(X \leq 115) - P(X \leq 85)$

Die beiden Wahrscheinlichkeiten $P(X \leq 115)$ und $P(X \leq 85)$ können wir dann mit dem **pnorm()**-Befehl berechnen.

```
pnorm(q = 115, mean = 100, sd = 15) - pnorm(85, 100, 15)

## [1] 0.6826895
```

Das heisst, etwa $^2/_3$ der Bevölkerung hat einen IQ zwischen 85 und 115.

◀

Das letzte Resultat aus dem Beispiel 5.2.2 gilt für *alle* Normalverteilungen $\mathcal{N}(\mu, \sigma^2)$. Die Wahrscheinlichkeit, dass eine Beobachtung höchstens *eine* Standardabweichung vom Erwartungswert abweicht, ist etwa $^2/_3$:

$$P(\mu - \sigma \leq X \leq \mu + \sigma) \approx \frac{2}{3}$$

Nun haben wir mit der Standardabweichung bei der Normalverteilung ein Mass und eine geometrische Interpretation für die mittlere Abweichung vom Erwartungswert:

> Ist eine Zufallsvariable normalverteilt, so liegen etwa zwei Drittel aller Messwerte eine Standardabweichung um den Erwartungswert.

Wir können auch noch die Wahrscheinlichkeit berechnen, dass eine Beobachtung höchstens *zwei* Standardabweichungen vom Erwartungswert abweicht:

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.95$$

Diese Wahrscheinlichkeiten können wir wieder als Flächen interpretieren. Die Fläche der Normalverteilung über dem Intervall

$$[\mu - \sigma,\ \mu + \sigma]$$

ist ca. $^2/_3$.

Die Fläche über dem Intervall

$$[\mu - 2\sigma, \mu + 2\sigma]$$

ist ca. 0.95 (siehe Abbildung 5.18).

5.2.2. Die Standardnormalverteilung

Die Normalverteilung $\mathcal{N}(0, 1)$ mit Mittelwert 0 und Standardabweichung 1 heisst *Standardnormalverteilung*.

Man kann jede normalverteilte Zufallsvariable X standardisieren (ohne Beweis).

Standardisieren einer normalverteilten Zufallsvariaben

Falls $X \sim \mathcal{N}(\mu, \sigma^2)$, so ist die standardisierte Zufallsvariable wieder normalverteilt, hat nun aber

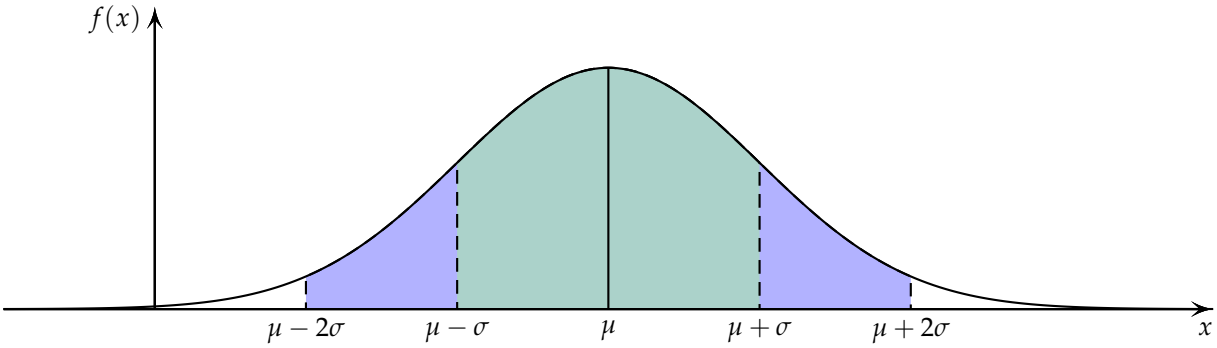


Abbildung 5.18. : Dichte der Normalverteilung. Ca. 68 % der Fläche befindet sich im Intervall $[\mu - \sigma, \mu + \sigma]$, ca. 95 % der Fläche im Intervall $[\mu - 2\sigma, \mu + 2\sigma]$.

Erwartungswert null und Varianz eins. Man erhält also die Standardnormalverteilung:

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

Damit lassen sich Wahrscheinlichkeiten für beliebige Normalverteilungen mit Hilfe der Standardnormalverteilung berechnen.

Die Standardnormalverteilung war lange bis zum Aufkommen von Computern sehr wichtig für Berechnungen. Heute spielt sie für angewandte Statistik keine Rolle mehr. □

5.3. Durchschnitte und Summen von Zufallsvariablen

5.3.1. Einleitung

Im letzten Unterkapitel 5.3.4 haben wir untersucht, wie *eine* Zufallsvariable verteilt ist. In vielen Anwendungen haben wir es aber nicht mit einer, sondern mit *mehreren* Zufallsvariablen zu tun.

Üblicherweise messen wir die *gleiche* Grösse mehrmals wie wir das am Waagenbeispiel 2.1.4 gesehen haben. In diesem Abschnitt untersuchen wir, wie die Summe und vor allem der Durchschnitt *mehrerer* Zufallsvariablen verteilt sind.

Diese mehreren Messungen bezeichnen wir wieder mit

$$x_1, x_2, \ldots, x_n$$

und fassen diese als Realisierungen der Zufallsvariablen

$$X_1, \ldots, X_n$$

auf.

Wir werden im Kapitel 6 beschreiben, wie wir eine bestimmte Angabe überprüfen können. Hier zwei Beispiele, warum die Durchschnitte von Zufallsvariablen so wichtig sind.

Beispiel 5.3.1

Jemand behauptet, dass die durchschnittliche Körpergrösse der erwachsenen Frauen in der Schweiz 185 cm. Nun ist diese Angabe offensichtlich falsch. Aber wie können wir dies überprüfen?

Wir wählen zufällig eine Person aus und messen deren Körpergrösse. Wenn diese in der Nähe von 185 cm liegt, so werden wir der Angabe eher glauben.

Nun ist es aber durchaus möglich, dass wir zufällig eine grosse Frau aussuchen. Oder eine sehr kleine. Diese eine Messung also nicht sehr viel über die wahre durchschnittliche Grösse der Frauen aus.

Die Idee ist jetzt natürlich, dass wir *mehrere* Frauen zufällig auswählen und den Durchschnitt der Körpergrössen messen. Dieser Wert macht eine bessere Aussage über die wahre Körpergrösse als lediglich eine Messung. Der Grund dafür ist, dass es unwahrscheinlich ist (aber nicht unmöglich), dass wir ausschliesslich zufällig beispielsweise 100 sehr grosse Frauen auswählen.

Wenn wir noch die Wahrscheinlichkeitsverteilung des Durchschnittes haben, dann können wir auch die Wahrscheinlichkeit berechnen, dass der Durchschnitt der Körpergrösse der Frauen beispielweise grösser als 185 cm ist. ◀

Beispiel 5.3.2

Wir nehmen eine Eistee-Petflasche auf der steht, dass der Inhalt 500 Milliliter beträgt. Stimmt diese Angabe?

Das Problem ist, dass der Inhalt dieser Pet-Flaschen *nie exakt* 500 ml beträgt. Manchmal kommt von der Abfüllanlage noch ein Tropfen zuviel dazu, manchmal einer weniger. Es gibt immer kleine Abweichungen von dieser Zahl.

Um diese Angabe zu überprüfen (siehe Kapitel 6 für das genaue Vorgehen), messen wir den Inhalt von 100 (oder 200 oder 1000) solcher Pet-Flaschen und nehmen den Durchschnitt dieser Messungen.

Ist der Durchschnitt nun in der „Nähe" von 500 ml, beispielsweise 499.9654 ml, so nehmen wir wohl an, dass die Angabe richtig ist.

Ist der Durchschnitt aber 463.45 ml, so scheint dies doch zuwenig zu sein und wir vermuten, dass die Angabe nicht stimmt.

Konkret machen wir 100 Messungen, die einer bestimmten Wahrscheinlichkeitsverteilung folgen. Jede Messung können durch eine Zufallsvariable X_i beschreiben, wobei i die -te Messung ist.

Wir haben also 100 Zufallsvariablen

$$X_1,\ X_2,\ \ldots,\ X_{100}$$

von denen uns der Durchschnitt interessiert

$$\overline{X}_{100} = \frac{X_1 + X_2 + \ldots + X_{100}}{100} = \frac{1}{100}(X_1 + X_2 + \ldots + X_{100}) = \frac{1}{100}\sum_{i=1}^{100} X_i$$

interessiert.

Wir wollen nun untersuchen, wie sich dieser Durchschnitt verhält. ◀

Haben wir die Messwerte $x_1, x_2, \ldots, x_n$, dann hat der Durchschnitt die Form

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Dieser Ausdruck hat als Input n unabhängige Variablen und eine reelle Zahl als Output. Wenn wir

$$x_1, x_2, \ldots, x_n$$

als Realisierungen der Zufallsvariablen

$$X_1, \ldots, X_n$$

betrachten, dann können wir $\overline{x}_n$ als Realisierung der Zufallsvariablen

$$\overline{X}_n = \frac{1}{n}(X_1 + X_2 + \ldots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

auffassen. Dies ist die Zufallsvariable für das *arithmetisches Mittel*. Das arithmetische Mittel $\overline{x}_n$ der Daten ist also eine Realisierung der Zufallsvariablen $\overline{X}_n$.

Analoges gilt für die *Summe* S_n:

$$S_n = X_1 + \cdots + X_n = \sum_{i=1}^{n} X_i$$

Wir sind an der Verteilung der Zufallsvariablen $\overline{X}_n$ interessiert: Die Kenntnis dieser Verteilung wird uns erlauben, Statistik aufgrund von arithmetischen Mitteln von Daten zu machen.

5.3.2. Unabhängigkeit und i.i.d. Annahme

Oft treffen wir die Annahme, dass die Zufallsvariablen $X_1, \ldots, X_n$ *unabhängig* voneinander sind (siehe auch Unterabschnitt 4.2.9 zur stochastischen Unabhängigkeit). Anschaulich heisst das, es gibt keine gemeinsamen Faktoren, die den Ausgang der verschiedenen Messungen beeinflussen und keine „carry over" Phänomene von einer Messung zur nächsten. Das heisst nichts anderes, dass eine Messung keinen Einfluss hat auf das Resultat der nachfolgenden Messungen.

Zunächst ein Beispiel für Zufallsvariablen, die *nicht* unabhängig voneinander sind.

Beispiel 5.3.3

Wir bezeichnen mit $\overline{X}_i$ die Höchsttemperatur im i-ten Tag in Luzern eines zufällig ausgewählten Jahres.

Diese sind nicht unabhängig voneinander, da diese Werte nicht zufällig um den Jahresdurchschnitt streuen. Benachbarte Sommertage werden die ähnlichen Temperturen und nicht an einem Tag 28 °C und am nächsten −5 °C. ◀

Wenn die Zufallsvariablen $X_1, \ldots, X_n$ unabhängig sind und alle *dieselbe* Verteilung haben, dann schreiben wir das kurz als

$$X_1, \ldots, X_n \quad \text{i.i.d.}$$

Die Abkürzung i.i.d. steht für:

independent, **i**dentically **d**istributed

Wir werden meistens mit dieser i.i.d. Annahme arbeiten, da wir oft von n Durchführungen des gleichen Experimentes ausgehen. Welche Verteilung die X_i's haben, lassen wir offen. Oft handelt es sich um eine Normalverteilung, dies muss aber *nicht* so sein.

Etwas salopp gesprochen:

> Sind Zufallsvariablen i.i.d., so wird dasselbe unter den gleichen Bedingungen gemessen.

Die Unabhängigkeit spielt eine Rolle bei den *Regeln für Erwartungswerte und Varianzen* von Summen. Die Beziehung

$$\mathrm{E}(X_1 + X_2) = \mathrm{E}(X_1) + \mathrm{E}(X_2)$$

gilt immer,

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2)$$

jedoch nur, wenn X_1 und X_2 unabhängig sind. □

5.3.3. Kennzahlen von S_n und $\overline{X}_n$

Wir nehmen in diesem Abschnitt an, dass

$$X_1, \ldots, X_n \quad \text{i.i.d.}$$

Wie gesagt, es wird hier nur verlangt, dass die Verteilungen der Zufallsvariablen gleich sind. Diese können normalverteilt sein mit gleichem μ und σ oder diskret gleichverteilt sein wie in Beispiel 5.3.4.

Wegen dem zweiten „i" in i.i.d. hat *jedes* X_i dieselbe Verteilung und dieselben Kennzahlen:

$$\mathrm{E}(X_i) = \mu \quad \text{und} \quad \mathrm{Var}(X_i) = \sigma_X^2$$

Wir wollen nun untersuchen, welche Kennzahlen (Erwartungswert und Varianz)

$$\mathrm{E}(S_n) \qquad \text{und} \qquad \sigma_{S_n}^2$$

die Summe S_n hat.

Und entsprechend

$$\mathrm{E}(\overline{X}_n) \qquad \text{und} \qquad \sigma_{\overline{X}_n}^2$$

für den Durchschnitt $\overline{X}_n$.

Wir machen hier keine exakte Herleitung, sondern wollen dies an einem Beispiel graphisch untersuchen.

Beispiel 5.3.4

Wir werfen einen fairen Würfel. Die Zufallsvariable X beschreibt die geworfene Augenzahl.

Die Verteilung von X ist in Tabelle 5.1 aufgeführt.

x	1	2	3	4	5	6
$P(X = x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Tabelle 5.1. : Verteilung beim Werfen eines fairen Würfels

Der Erwartungswert ist

$$\mathrm{E}(X) = \mu = \frac{1}{6}(1+2+3+4+5+6) = 3.5$$

und die Varianz ist

$$\begin{aligned}\mathrm{Var}(X) &= \frac{1}{6}\left((1-3.5)^2 + (2-3.5)^2 + (3-3.5)^2 + (4-3.5)^2 + (5-3.5)^2 + (6-3.5)^2\right) \\ &= 2.92\end{aligned}$$

```
x <- c(1, 2, 3, 4, 5, 6)

ave <- mean(x)
ave


## [1] 3.5


var <- mean((x - ave)^2)
var


## [1] 2.916667
```

Nun würfeln wir 10mal. Diese 10 Würfe werden durch die Zufallsvariablen

$$X_1, X_2, \ldots, X_{10} \quad \text{i.i.d.}$$

beschrieben, wobei X_i jeweils die Augenzahl im i-ten Wurf angibt.

Die Verteilung, der Erwartungswert und die Varianz der X_i's entsprechen jeweils den Werten der Zufallsvariable X oben. Also beispielsweise

$$\mathrm{E}(X_1) = \mathrm{E}(X_2) = \ldots = \mathrm{E}(X_{10}) = 3.5$$

Wir notieren uns die Augenzahl dieser 10 Würfe. Dies machen wir 1000 mal und bilden ein Histogramm aller vorkommenden *Augensummen* und *Durchschnitte*. Dann machen wir dasselbe noch mit 40 Würfen. Natürlich simulieren wir dies mit **R** (siehe Abbildung 5.19).

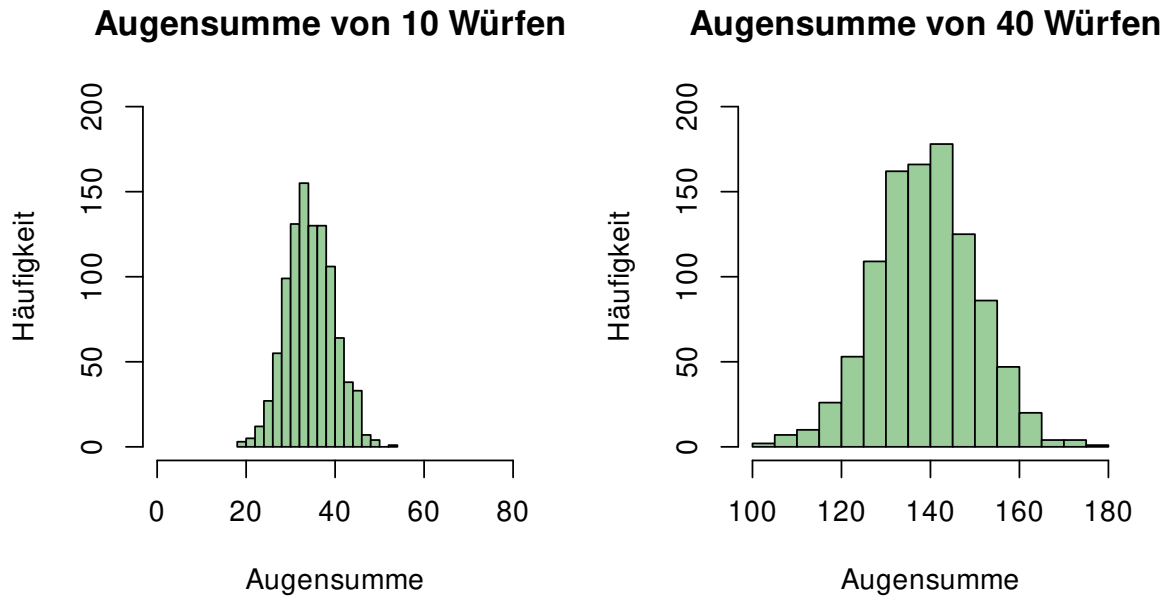


Abbildung 5.19. : Histogramme von Augensummen

Erwartungswert der Summe

Es fällt auf, dass sich die mittlere Augensumme *verschiebt*, wenn mehr Würfe gemacht werden. Dies ist aber nicht sonderlich überraschend, da die Summe grösser wird, wenn wir mehr Würfe machen.

In der Abbildung 5.19 links ist die grösste Häufigkeit bei etwa 35, also

$$10 \cdot 3.5 = 10 \cdot \mu$$

wobei μ hier der Erwartungswert für *einen* Wurf ist.

In der Abbildung 5.19 rechts ist die grösste Häufigkeit bei etwa 140

$$40 \cdot 3.5 = 40 \cdot \mu$$

Wir können hier die Vermutung aufstellen, dass

$$\mathrm{E}(S_n) = n\mu$$

Diese Vermutung kann man auch beweisen, aber wir werden dies hier nicht machen.

Diese Regel ist aber ist aber leicht einsehbar in diesem Beispiel. Nehmen wir den Fall für $n = 40$ Würfe. Der Wertebereich ist dann

$$W = \{40, 41, \ldots, 240\}$$

Der Wert 40 entspricht dem Werfen von 40 Einsen. Dies ist sehr unwahrscheinlich und dies kommt im Histogramm auch sehr selten vor (in diesem Fall gar nicht). Dasselbe gilt auf der anderen Seite mit 40 Sechsen.

Wir erwarten, dass auf die 40 Würfe, die Zahlen 1 bis 6 möglichst gleichmässig verteilt sind. Abweichungen sind natürlich möglich, aber diese werden immer unwahrscheinlicher je extremer das Wurfbild ist.

Der Erwartungswert eines Wurfes ist 3.5 und wenn die Zahlen von 1 bis 6 auf die 40 Würfe gleichmässig verteilt sind, so ist die Summe ungefähr

$$3.5 \cdot 40 = 140$$

Varianz und Standardabweichung der Summe

Es fällt weiter auf, dass die Varianz und Standardabweichung mit zunehmender Anzahl Würfen *zunimmt*. Auch hier können wir dies relativ einfach erklären[3]. Je mehr Würfe wir machen, umso grösser wird der Wertebereich. Das heisst, die Summen verteilen sich auf mehr Zahlen und damit nimmt auch die Streuung zu.

Das Gesetz für die Varianz und Standardabweichung der Summe lautet wie folgt (ohne Beweis):

$$\mathrm{Var}(S_n) = n\,\mathrm{Var}(S_n), \qquad \sigma_{S_n} = \sqrt{n}\sigma_X$$

Erwartungswert des Durchschnittes

Wir können nun noch dasselbe mit dem Durchschnitt $\overline{X}_n$ machen und erhalten die Histogramme in Abbildung 5.20.

Hier fällt auf, dass in beiden Histogrammen die *grösste* Häufigkeit bei 3.5, also bei μ liegt.

[3] Diese Erklärung ist zwar einfach, aber ich nicht ganz richtig. Für unsere Bedürfnisse reicht die Erklärung allerdings.

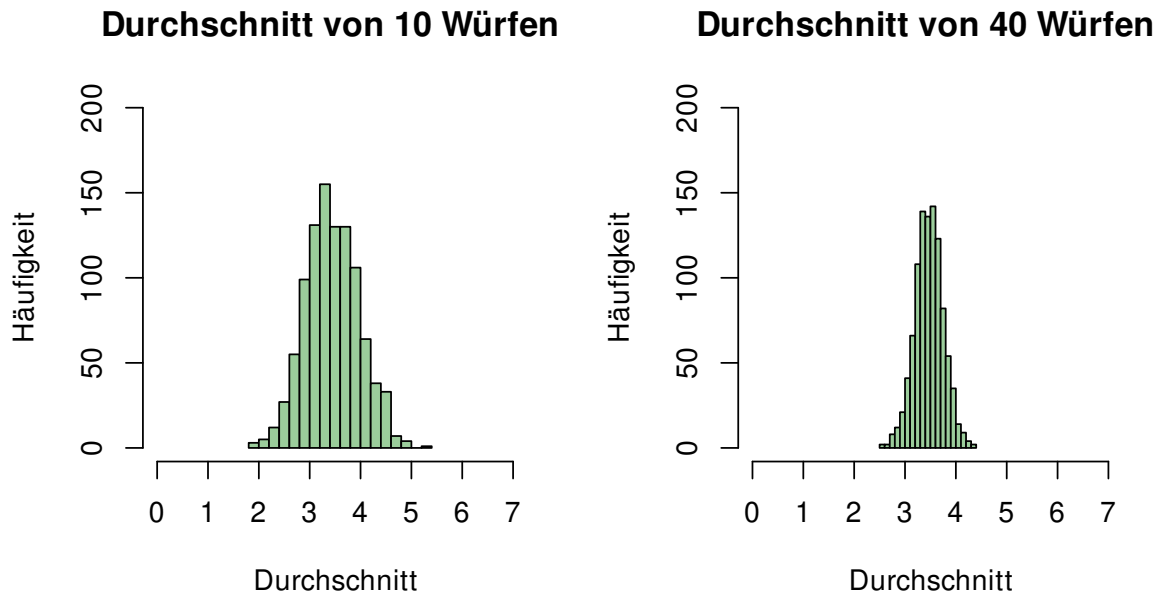


Abbildung 5.20. : Histogramme von Durchschnitten von Würfelwürfen

Auch in diesem Fall ist das leicht erklärbar. Wir erwarten wieder das die Zahlen 1 bis 6 eher gleichmässig auf die beispielweise 40 Würfe verteilt sind und damit entspricht der Durchschnitt in etwa dem Erwartungswert 3.5.

Wir vermuten also (was auch stimmt, aber hier nicht bewiesen wird):

$$\mathrm{E}(\overline{X}_n) = \mu$$

Varianz und Standardabweichung des Durchschnittes

Weiter fällt hier auf, dass die Varianz und damit die Standardabweichung mit zunehmender Anzahl Würfen *abnimmt*.

Auch dies ist relativ leicht einsehbar. Je mehr Würfe wir machen, umso mehr können wir vermuten, dass sich die Zahlen 1 bis 6 immer gleichmässiger auf die Anzahl Würfe verteilen. Das heisst, die extremeren Wurfbilder (beispielsweise sehr viele Einsen) werden immer unwahrscheinlicher und die Durchschnitte „rücken" immer näher an den Erwartungswert 3.5. Die Varianz und damit die Standardabweichung nimmt ab, wenn wir die Anzahl Würfe erhöhen.

Für die Varianz und die Standardabweichung des Durchschnittes gilt folgendes Gesetz (ohne Beweis):

$$\mathrm{Var}(\overline{X}_n) = \frac{\mathrm{Var}(X)}{n}; \qquad \sigma_{\overline{X}_n} = \frac{\sigma_X}{\sqrt{n}}$$

◀

Die im Beispiel oben gemachten Beobachtungen gelten allgemein.

Allgemein

Für

$$X_1, \ldots, X_n \quad \text{i.i.d.}$$

gilt (ohne Herleitung):

Kennzahlen von S_n

$$\begin{aligned} \mathrm{E}(S_n) &= n\mu \\ \mathrm{Var}(S_n) &= n\,\mathrm{Var}(X_i) \\ \sigma(S_n) &= \sqrt{n}\sigma_X \end{aligned}$$

Kennzahlen von $\overline{X}_n$

$$\begin{aligned} \mathrm{E}(\overline{X}_n) &= \mu \\ \mathrm{Var}(\overline{X}_n) &= \frac{\sigma_X^2}{n} \\ \sigma(\overline{X}_n) &= \frac{\sigma_X}{\sqrt{n}} \end{aligned}$$

Die Standardabweichung von $\overline{X}_n$ heisst auch der *Standardfehler* des arithmetischen Mittels.

Die Standardabweichung der Summe wächst also mit wachsendem n, aber langsamer als die Anzahl Beobachtungen n.

Der Erwartungswert von $\overline{X}_n$ ist also gleich demjenigen einer einzelnen Zufallsvariablen X_i, die *Streuung nimmt jedoch mit wachsendem n ab*.

Hier noch die mathematische Schreibweise

Gesetz der grossen Zahlen

Für $n \to \infty$ geht die Streuung gegen null. Es gilt das **Gesetz der grossen Zahlen:** Falls

$X_1, \ldots, X_n$ i.i.d., dann

$$\overline{X}_n \longrightarrow \mu \quad \text{für} \quad n \to \infty$$

□

Standardfehler

Die Standardabweichung des arithmetischen Mittels (*Standardfehler*) ist jedoch *nicht* proportional zu $1/n$, sondern nimmt nur ab mit dem Faktor $1/\sqrt{n}$

$$\sigma_{\overline{X}_n} = \frac{1}{\sqrt{n}} \sigma_X$$

Um den *Standardfehler* zu halbieren, braucht man also *viermal* so viele Beobachtungen. Dies nennt man auch das $\sqrt{n}$*-Gesetz*.

5.3.4. Verteilungen von S_n und $\overline{X}_n$

Wir kennen nun die Kennzahlen von S_n und $\overline{X}_n$, aber wir wissen noch nicht, S_n und $\overline{X}_n$ verteilt sind.

Beispiel 5.3.5

Im Würfelbeispiel 5.3.4 haben wir den Erwartungswert von die Summe S_n und den Durchschnitt $\overline{X}_n$ ermittelt. Aber wir haben noch nicht gesagt, wie die beiden Grössen verteilt sind.

Betrachten wir aber die Histogramme in Abbildung 5.19 und 5.20, so sehen wir, dass die Histogramme annähernd normalverteilt aussehen. Dies ist in der Tat so: S_n *und* $\overline{X}_n$ *sind für grosse n annähernd normalverteilt*.

Dies ist doch eine überraschung, da die Verteilung der X_i (siehe Tabelle 5.1 und Abbildung 5.21) überhaupt nicht normalverteilt ist.

◀

Wir werden sehen, dass die Summe und der Durchschnitt normalverteilt sind. Dies ist die Aussage des *Zentralen Grenzwertsatzes*. Wir wollen dies graphisch an einem weiteren Beispiel veranschaulichen. Dabei untersuchen wir das Verhalten von $\overline{X}_n$.

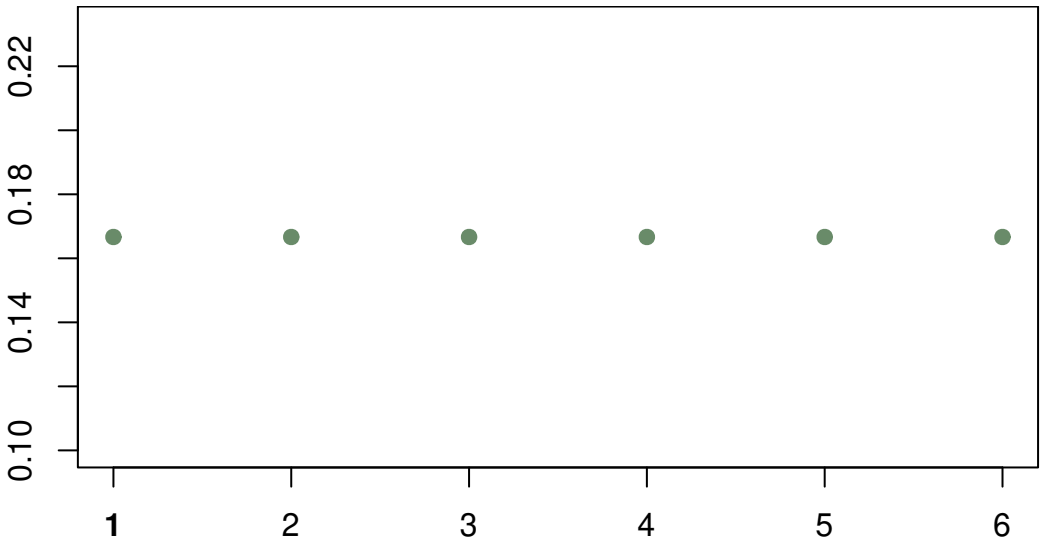


Abbildung 5.21. : Verteilung eines fairen Würfels

Beispiel 5.3.6

Wir haben eine Ergebnismenge

$$\Omega = \{0, 10, 11\}$$

aus der wir eine Zahl ziehen. Die Zufallsvariable X gibt den Wert der gezogenen Zahl an. Zudem gilt

$$P(X = 0) = P(X = 10) = P(X = 11) = \frac{1}{3}$$

Damit gilt für den Erwartungswert von X:

$$\mathrm{E}(X) = \frac{1}{3} \cdot 0 + \frac{1}{3} \cdot 10 + \frac{1}{3} \cdot 11 = 7$$

```
werte <- c(0, 10, 11)

ew <- sum(werte * 1/3)
ew


## [1] 7
```

und für die Varianz gilt dann

$$\mathrm{Var}(X) = \frac{1}{3} \cdot (0-7)^2 + \frac{1}{3} \cdot (10-7)^2 + \frac{1}{3} \cdot (11-7)^2 = 24.6667$$

```
var.X <- sum((werte - ew)^2 * 1/3)

var.X


## [1] 24.66667
```

mit der Standardabweichung

$$\sigma_X = \sqrt{\text{Var}(X)} = 4.9666$$

```
## [1] 4.966555
```

Von nun an soll ein Versuch aus 10 Ziehungen bestehen. An sich müssten wir vielmehr Ziehungen machen, was wir später auch tun werden, aber bei 10 Ziehungen „sieht" man besser, was passiert.

Wir beginnen mit einem Versuch (10 Ziehungen).

```
# Zieht 10-mal aus der Menge {0,10,11} einen Wert mit gleicher
# W'keit
sim <- sample(werte, size = 10, replace = T)

# Vektor mit 10 Werten
sim


##  [1]  0 10 11 11 11 11 10 10  0 10
```

Diese Daten stellen wir noch als Histogramm dar (siehe Abbildung 5.22).

```
# Histogramm mit diesen 10 Werten
hist(sim, col = "darkseagreen3")
```

Wiederholen wir den Versuch, so erhalten wir in der Regel ein leicht anderes Histogramm. In Abbildung 5.23 sind 4 aufgeführt.

Offensichtlich haben wir es hier mit keiner Normalverteilung zu tun. In diesem Versuch kamen nur die Zahlen 0, 10, 11 vor.

Nun können wir zwei solche Versuche (je 10 Ziehungen) hintereinander ausführen und den *Durchschnitt* aus den beiden Versuchen berechnen.

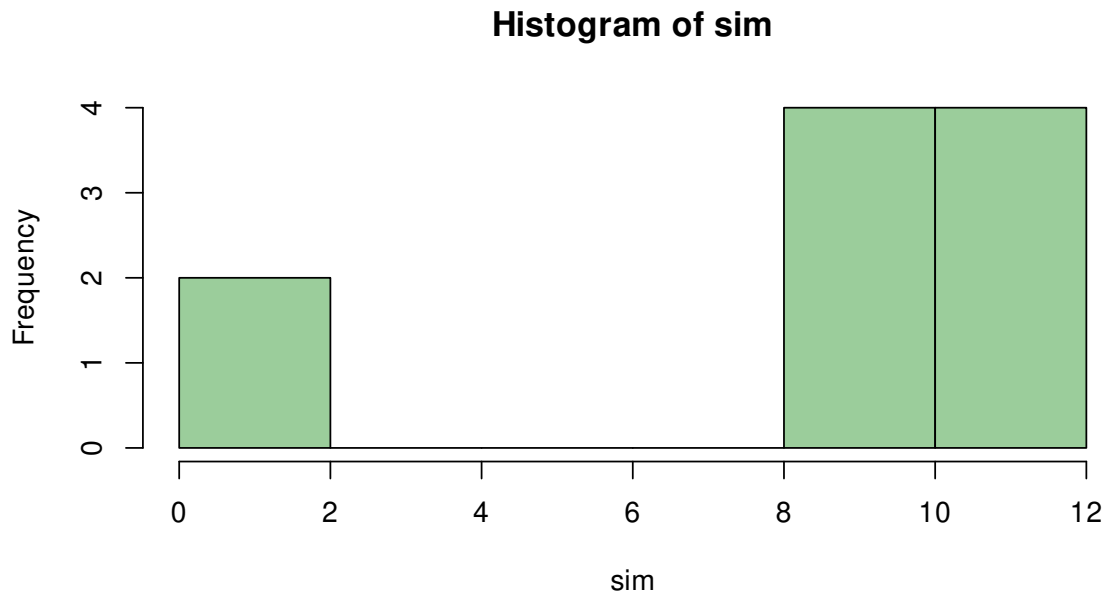


Abbildung 5.22. : Histogramm mit 10 Ziehungen

```
sim.1 <- sample(werte, size = 10, replace = T)
sim.1

##  [1]  0 11  0 10  0 11 11 10 10 11

sim.2 <- sample(werte, size = 10, replace = T)
sim.2

##  [1] 11  0  0  0 10 10 10 10 11  0

sim.mean <- (sim.1 + sim.2)/2
sim.mean

##  [1]  5.5  5.5  0.0  5.0  5.0 10.5 10.5 10.0 10.5  5.5
```

Neben den Zahlen 0, 10, 11 können nun auch noch die Zahlen 5, 5.5 und 10.5 vorkommen.

Diese Werte sind im Histogramm in Abbildung 5.24 dargestellt.

Es zeigt sich auch in diesem Fall, dass das Histogramm mit jedem Versuch anders ausschaut (siehe Abbildung 5.25).

```r
par(mfrow = c(2, 2))
for (i in 1:4) {
    sim <- sample(werte, size = 10, replace = T)
    hist(sim, col = "darkseagreen3")
}
```

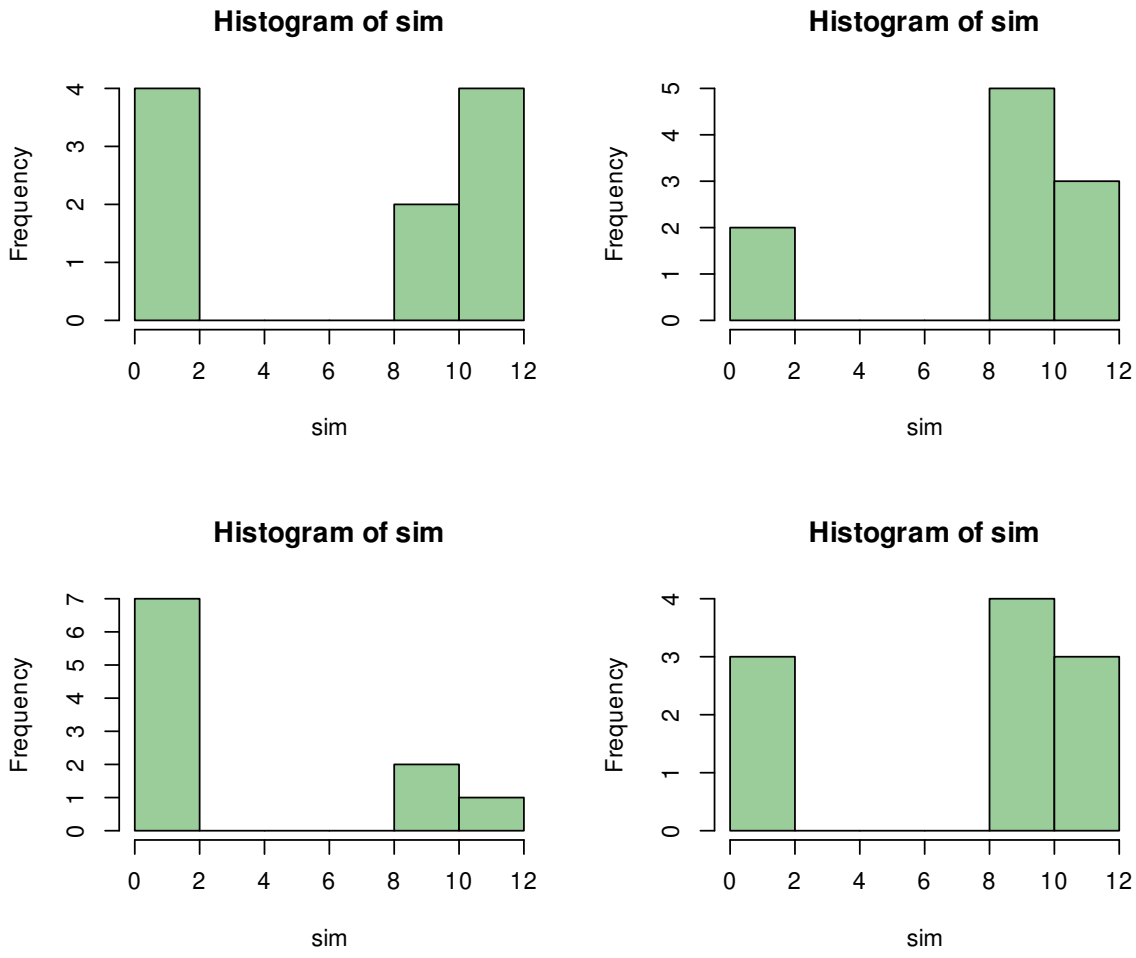


Abbildung 5.23. : 4 Histogramme mit 10 Ziehungen

Obwohl jeder Versuch anders ausschaut, zeichnen sich doch bestimmte Tendenzen ab. Die 0 ist zum Beispiel weniger oft vertreten, da eine doppelte 0 nur mit Wahrscheinlichkeit $\frac{1}{9}$ vorkommt.

Wir können dasselbe für 3 Versuche wiederholen und den Durchschnitt nehmen (siehe Abbildung 5.26). Wir bekommen jetzt noch mehr mögliche Werte für den Durchschnitt.

Wir führen wieder mehrere Versuche durch (siehe Abbildung 5.27).

```
hist(sim.mean, col = "darkseagreen3")
```

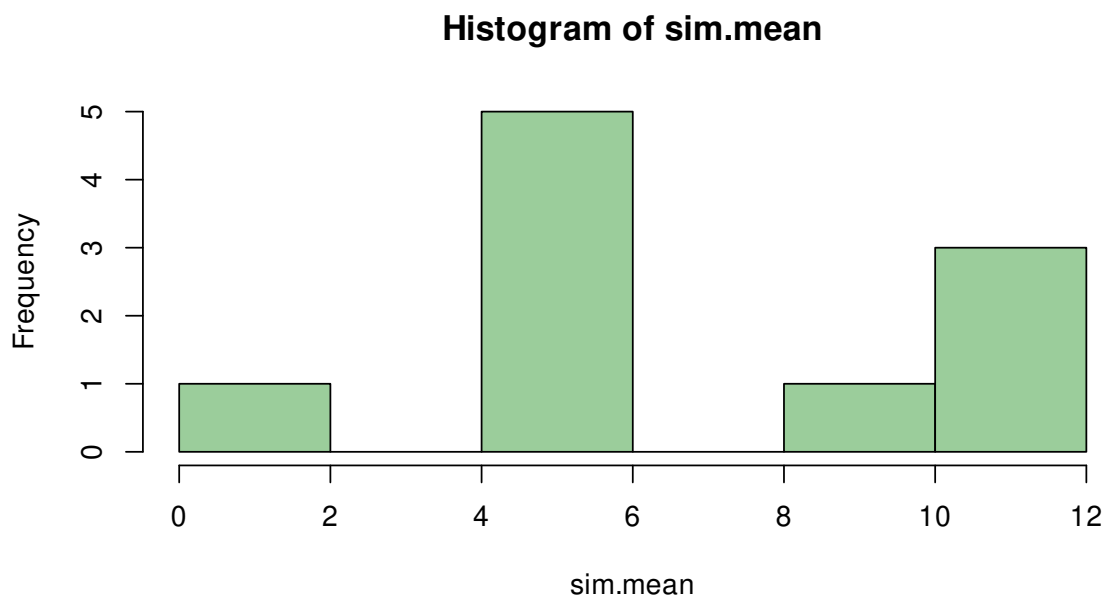


Abbildung 5.24. : Histogramm vom Durchschnitt von zwei Versuchen mit 10 Ziehungen

Es zeigt sich ebenfalls, dass sich die Durchschnitte mehr um den Erwartungswert 7 häufen, je mehr Versuche wir machen.

Wir stellen also fest, dass es beim *Durchschnitt* immer mehr Werte gibt, die sich um den Erwartungswert 7 häufen. Warum ist dies so? Die Zahl 0 im Durchschnitt kommt praktisch nicht mehr vor, da die Wahrscheinlichkeit, dass 3 mal an derselben Stelle eine 0 vorkommt, nur noch $1/27$ ist. Dasselbe ist für die Zahl 11 der Fall.

Für viele Versuche ist dieses Vorgehen ungeeignet und nicht sehr elegant (obwohl man „sieht", was passiert).

Wir wollen nun 16, 64, 256 und 1024 solche Versuche durchführen, aber mit jeweils 1000 Ziehungen. Von denen nehmen wir jeweils den Durchschnitt wie in den Beispielen vorher. Das Resultat ist in der Abbildung 5.28 dargestellt.

Was bei genauerem Hinschauen auffällt:

- die Werte häufen sich um den Erwartungswert 7
- die Standardabweichung wird kleiner, und zwar halbiert sie sich etwa beim Vervierfachen der Anzahl Versuche
- die Histogramme scheinen einer Normalverteilung zu folgen

```
par(mfrow = c(2, 2))
for (i in 1:4) {
    sim.1 <- sample(werte, size = 10, replace = T)
    sim.2 <- sample(werte, size = 10, replace = T)
    sim.mean <- (sim.1 + sim.2)/2
    hist(sim.mean, col = "darkseagreen3")
}
```

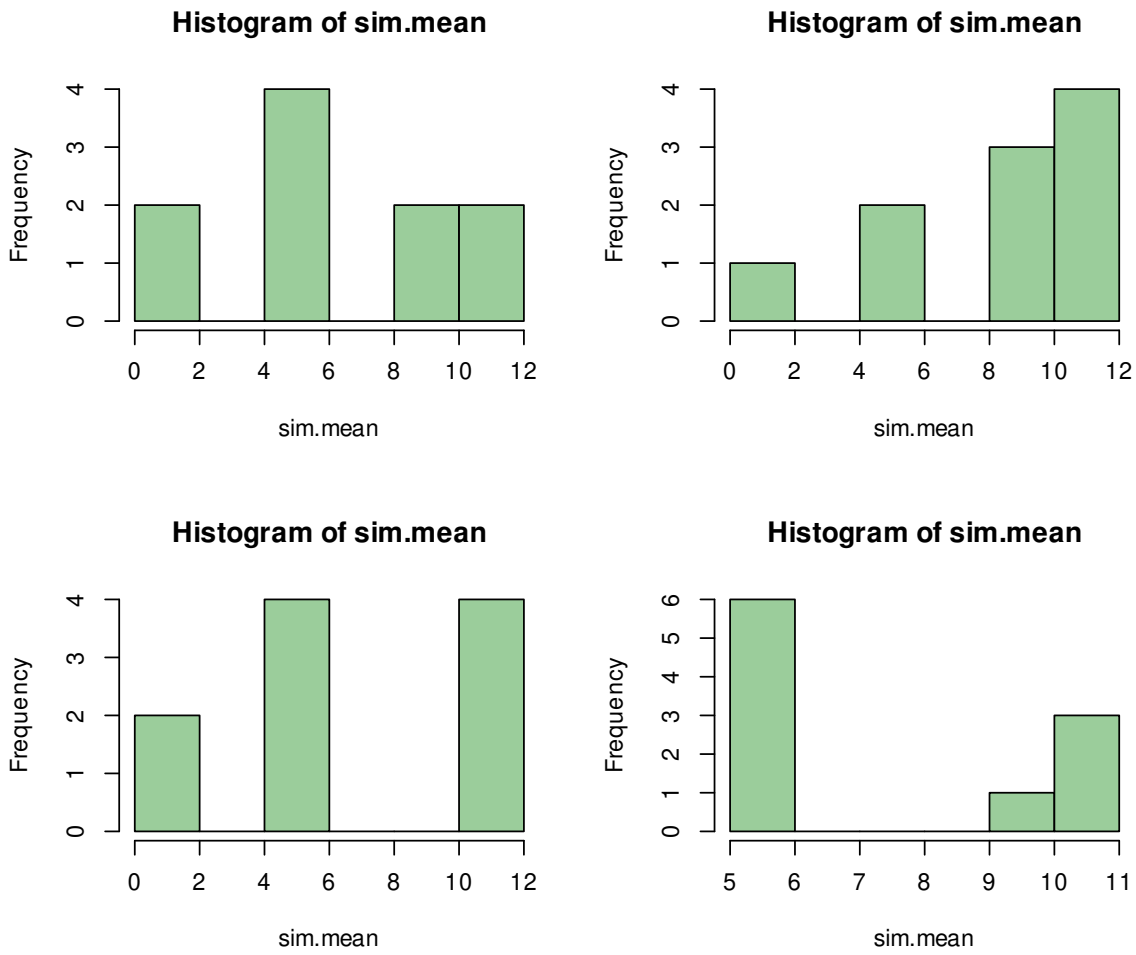


Abbildung 5.25. : Histogramm vom Durchschnitt von drei Versuchen mit 10 Ziehungen

Den dritten Punkt wollen wir noch genauer untersuchen, indem wir die jeweiligen Dichtekurven für

$$\mathcal{N}\left(7, \frac{24.6667}{n}\right)$$

einzeichnen (siehe Abbildung 5.29).

Es fällt auf, dass die Dichtekurven für grössere n immer besser zu den Histogrammen passen.

```
sim.1 <- sample(werte, size = 10, replace = T)
sim.1

##  [1] 10 10  0 11 11 10 11 10  0 10

sim.2 <- sample(werte, size = 10, replace = T)
sim.2

##  [1]  0 11 11  0 10  0 11 10 11 11

sim.3 <- sample(werte, size = 10, replace = T)
sim.3

##  [1]  0  0 10 10 10  0  0  0 10  0

sim.mean <- (sim.1 + sim.2 + sim.3)/3
round(sim.mean, 2)

##  [1]  3.33  7.00  7.00  7.00 10.33  3.33  7.33  6.67  7.00  7.00

hist(sim.mean, col = "darkseagreen")
```

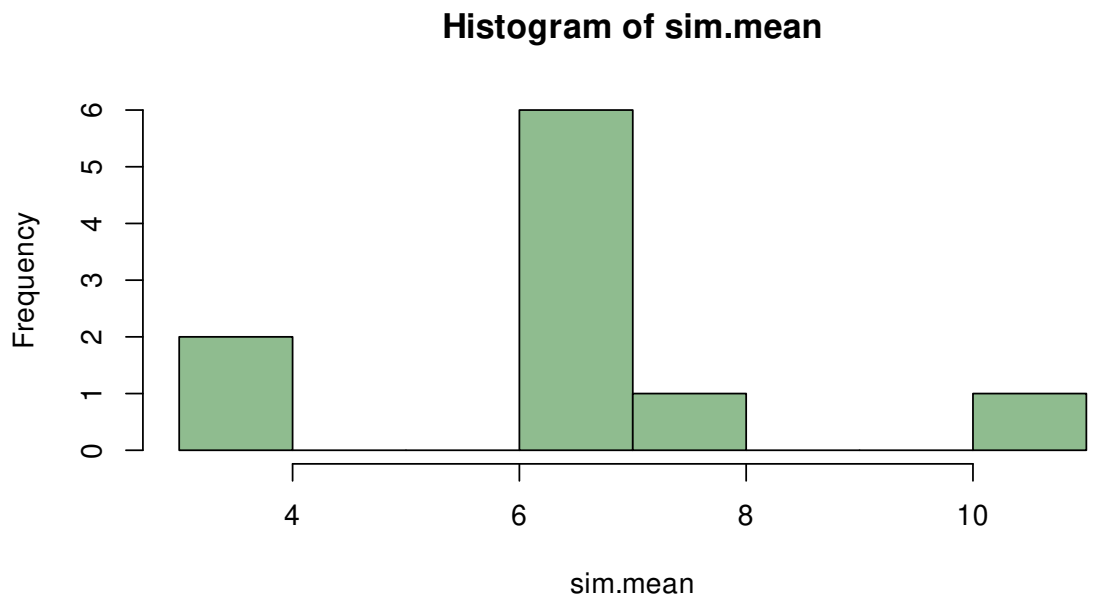


Abbildung 5.26. : Histogramm vom Durchschnitt von drei Versuchen mit 10 Ziehungen

Es sei nochmals erwähnt, dass wir mit einer Verteilung begonnen haben, die *nichts* mit einer Normalverteilung zu tun hat. Aber die Verteilung der *Mittelwerte* $\overline{X}_n$ (oder auch der Summen) nähert sich mit wachsendem n einer Normalverteilung an. ◀

```
par(mfrow = c(2, 2))
for (i in 1:4) {
    sim.1 <- sample(werte, size = 10, replace = T)
    sim.2 <- sample(werte, size = 10, replace = T)
    sim.3 <- sample(werte, size = 10, replace = T)
    sim.mean <- (sim.1 + sim.2 + sim.3)/3
    hist(sim.mean, col = "darkseagreen3")
}
```

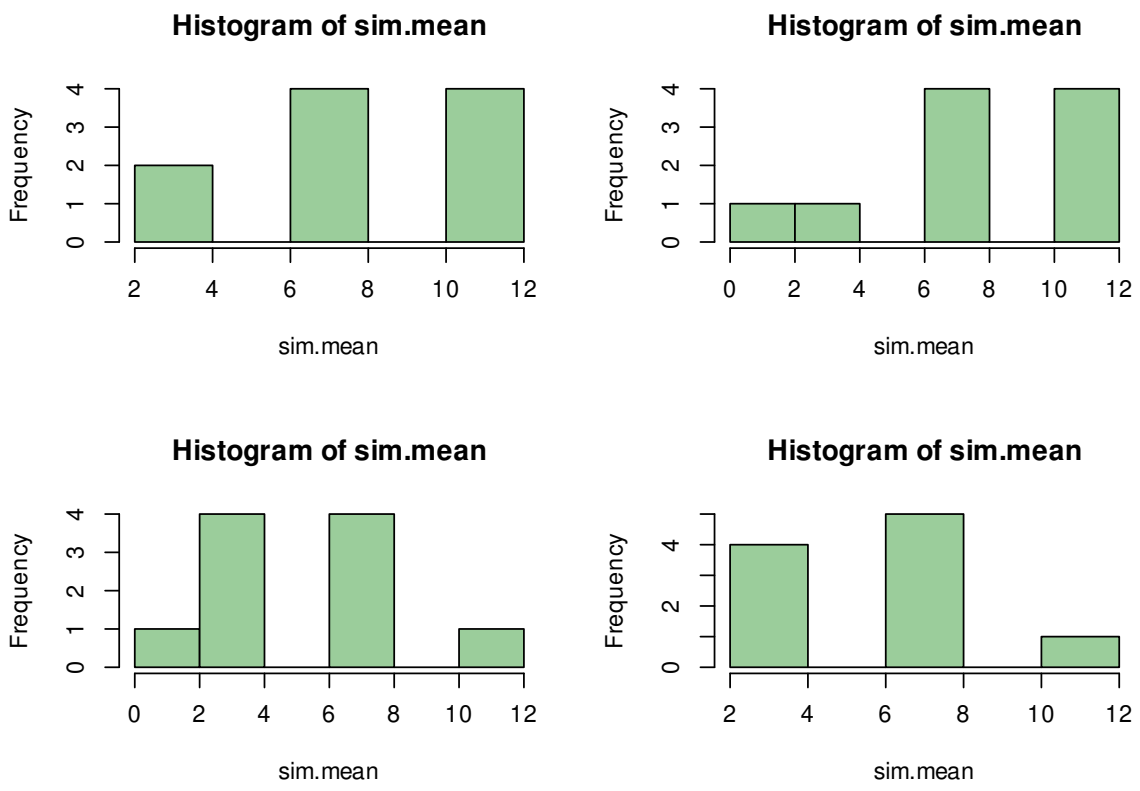


Abbildung 5.27. : Histogramme vom Mittelwert von drei Versuchen von jeweils 10 Ziehungen

Sind die X_i's i.i.d. (nicht notwendig normalverteilt), dann gilt der berühmte Zentrale Grenzwertsatz

Zentraler Grenzwertsatz

Falls $X_1, \ldots, X_n$ i.i.d. mit irgendeiner Verteilung mit Erwartungswert μ und Vari-

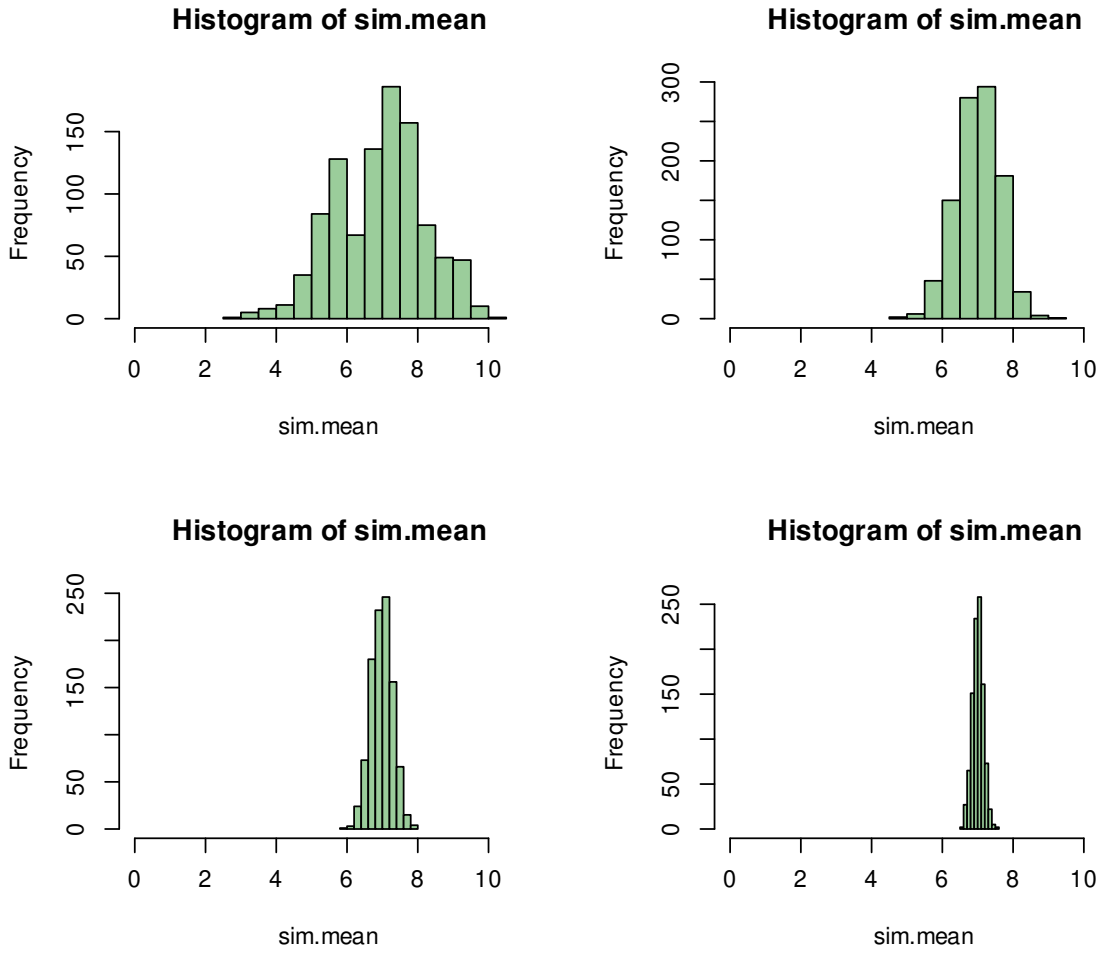


Abbildung 5.28. : 4 Histogramme vom Durchschnitt von 16, 64, 256, 1024 Versuchen mit 1000 Ziehungen

anz σ^2, dann gilt (ohne Beweis)

$$S_n \approx \mathcal{N}(n\mu, n\sigma_X^2)$$

$$\overline{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma_X^2}{n}\right)$$

wobei die Approximation im allgemeinen besser wird mit grösserem n.

Überdies ist auch die Approximation besser, je näher die Verteilung von X_i bei der Normalverteilung $\mathcal{N}(\mu, \sigma_X^2)$ ist.

Selbst wenn wir die Verteilung der X_i nicht kennen, so haben wir eine Ahnung über

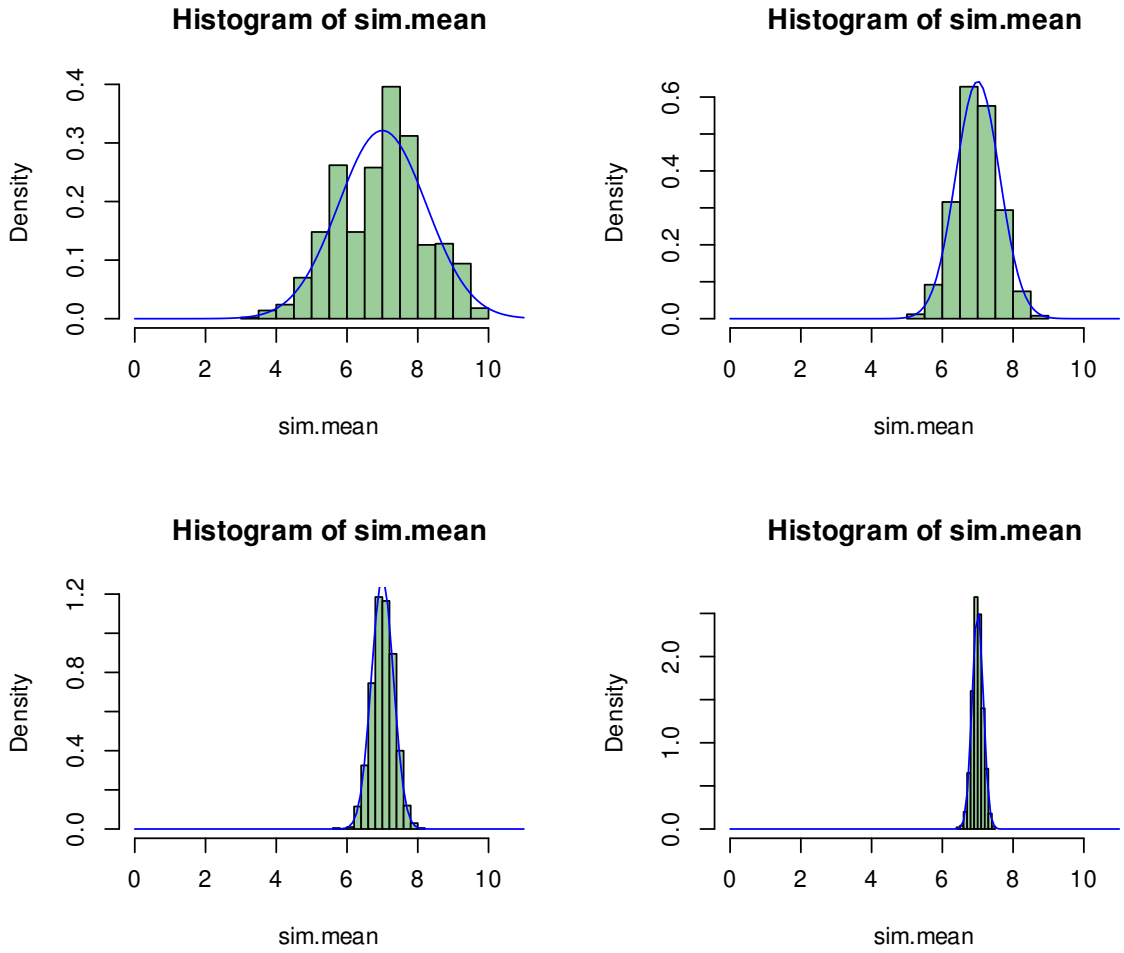


Abbildung 5.29. : 4 Histogramme vom Durchschnitt von 16, 64, 256, 1024 Versuchen mit 1000 Ziehungen mit Dichtekurven

die approximative Verteilung von S_n und X_n. Der Zentrale Grenzwertsatz (ZGWS) ist mitunter ein Grund für die Wichtigkeit der Normalverteilung.

Beispiel 5.3.7

Die Lebensdauer eines bestimmten elektrischen Teils ist durchschnittlich 100 Stunden mit Standardabweichung von 20 Stunden. Wir testen 16 solcher Teile.

Wie gross ist Wahrscheinlichkeit, dass das Stichprobenmittel

1. unter 104 Stunden oder
2. zwischen 98 und 104 Stunden liegt?

Lösung:

X_i ist die Zufallsvariable für die Lebensdauer des Teils i. Es gilt $\mu = 100$ und $\sigma_X = 20$.

Wir betrachten die durchschnittliche Lebensdauer $\overline{X}_{16}$, die annähernd wie folgt verteilt ist:

$$\overline{X}_{16} \approx \mathcal{N}\left(\mu, \frac{\sigma_X^2}{n}\right) = \mathcal{N}\left(100, \frac{20^2}{16}\right) = \mathcal{N}(100, 25)$$

1. Gesucht ist
$$P(\overline{X}_{16} \leq 104) = 0.7881446$$

```
pnorm(q = 104, mean = 100, sd = 20/sqrt(16))

## [1] 0.7881446
```

2. Gesucht ist
$$P(98 \leq \overline{X}_{16} \leq 104) = 0.4435663$$

```
pnorm(q = 104, mean = 100, sd = 20/sqrt(16)) - pnorm(q = 98, mean = 100,
    sd = 20/sqrt(16))

## [1] 0.4435663
```

◀

Beispiel 5.3.8

Das Strassenverkehrsamt hat genug Streusalz gelagert, um mit einem Schneefall von insgesamt 80 cm fertigzuwerden. Täglich fallen im Mittel 1.5 cm mit einer Standardabweichung von 0.3 cm.

Wie gross die Wahrscheinlichkeit, dass das gelagerte Salz für die nächsten 50 Tage ausreicht?

Lösung:

X_i ist die Zufallsvariable für die gefallene Menge Schnee am Tag i. Es gilt $\mu = 1.5$ und $\sigma_X = 0.3$. Zusätzlich nehmen wir an, dass die X_i's i.i.d. sind.

Wir betrachten die Schneemenge (Summe) S_{50} der nächsten 50 Tage und diese soll 80 nicht übersteigen. Es gilt annähernd

$$S_{50} \approx \mathcal{N}\left(50 \cdot \mu, 50 \cdot \sigma_X^2\right) = \mathcal{N}(75, 4.5)$$

Wir suchen

$$P(S_n \leq 80) = 0.9907889$$

```
pnorm(q = 80, mean = 50 * 1.5, sd = sqrt(50) * 0.3)


## [1] 0.9907889
```

In 99 % aller Winter reicht das Streusalz. Das heisst, nur einmal in einem Jahrhundert reicht das Streusalz nicht. ◀

Kapitel 6.

Hypothesentest für Messdaten

> We must be careful not to confuse data with the abstractions we use to analyze them.
>
> *(William James)*

6.1. Einleitung

In diesem Kapitel führen wir Hypothesentests ein. Diese bilden ein wichtiges Hilfsmittel in der Statistik. Worum geht es?

Beispiel 6.1.1

Wir haben schon einige Male das Beispiel mit der Pet-Flasche erwähnt, dessen Inhalt 500 ml betragen soll. Aber stimmt diese Zahl überhaupt?

Um dies zu überprüfen, bestimmen wir den Inhalt von mehreren Pet-Flaschen. Einige werden ein bisschen mehr als 500 ml haben, einige ein bisschen mehr. Liegt der Durchschnitt der Inhalte nicht zu weit von 500 ml entfernt, so glauben wir der Angabe. ◀

Solche Überprüfungen von Angaben werden wir konkret mit einem Hypothesentest gemacht. Die Hypothese ist die Angabe, die stimmt oder halt eben auch nicht.

Hypothesentest sind ein standardisiertes, reproduzierbares Verfahren mit dem Angaben getestet werden. Diese geben auch ein klares (aber nicht eindeutiges) Kriterium an, wann ein Durchschnitt zu weit von einer Angabe entfernt ist.

Es sei hier schon darauf hingewiesen, dass man mit Hypothesentests *nie beweisen* kann, dass eine Angabe stimmt oder nicht, sondern nur dass die Angabe mit einer gewissen Wahrscheinlichkeit stimmt oder nicht.

6.2. Statistische Tests und Vertrauensintervall für eine Stichprobe bei normalverteilten Daten

6.2.1. Problemstellung

Wir veranschaulichen die Thematik dieses Unterkapitels vor allem an folgendem schon bekanntem Beispiel (siehe Beispiel 2.1.4).

Beispiel 6.2.1

Wir betrachten die zwei *Datensätze* des Waagenbeispiels 2.1.4 (siehe Tabelle 6.1).

Waage A	79.98	80.04	80.02	80.04	80.03	80.03	80.04	79.97	80.05
Waage A	80.03	80.02	80.00	80.02					
Waage B	80.02	79.94	79.98	79.97	79.97	80.03	79.95	79.97	

Tabelle 6.1. : Waagenbeispiel

Wie wir im Unterkapitel 5.3 gesehen haben, können wir diese Messungen als Realisierungen von unabhängigen, identisch verteilten Zufallsvariablen X_i betrachten. So ist der zweite Messwert $x_2 = 80.04$ der Waage A eine Realisierung der Zufallsvariable X_2. ◀

Allgemeiner fassen wir nun die Messdaten $x_1, \ldots, x_n$ als Realisierungen von

$$X_1, \ldots, X_n \text{ i.i.d. } \sim \mathcal{N}(\mu, \sigma_X^2)$$

Zwei Kennzahlen der Zufallsvariablen X_i sind:

$$\mathrm{E}(X_i) = \mu \quad \text{und} \quad \mathrm{Var}(X_i) = \sigma_X^2$$

Typischerweise sind diese (und andere) Kennzahlen *unbekannt* und wir möchten Rückschlüsse darüber aus den Daten wie in Beispiel 6.2.1 ziehen. Das heisst, wir wollen aus den Daten eine Aussage über das wahre, aber eben unbekannte μ und σ^2 machen.

Das Ziel ist es, aus den Daten μ und σ^2 anzunähern. Wir sprechen dann auch von einer *Schätzung* der Parameter μ und σ^2.

Geschätzte Werte werden immer mit einem Hut $\widehat{\ }$ bezeichnet. So ist $\widehat{\mu}$ die Annäherung (Schätzung) von μ.

Die (Punkt-) Schätzungen für den Erwartungswert und die Varianz lauten (siehe Unterkapitel 2.2):

$$\widehat{\mu} = \overline{X}_n = \frac{X_1 + \ldots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
$$\widehat{\sigma}_X^2 = \frac{(X_1 - \overline{X}_n)^2 + \ldots + (X_n - \overline{X}_n)^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

Beachten Sie, dass die Schätzer hier als Funktionen der Zufallsvariablen $X_1, \ldots, X_n$ geschrieben sind: insbesondere sind $\widehat{\mu}$ und $\widehat{\sigma}_X^2$ selbst wieder Zufallsvariablen - die Verteilungseigenschaften von $\widehat{\mu}$ wurden im Unterkapitel 5.3 diskutiert. Das heisst, für jede neue Messung sind auch $\widehat{\mu}$ und $\widehat{\sigma}_X^2$ im Allgemeinen anders.

Obwohl im Allgemeinen $\widehat{\mu} \neq \mu$, ist die Hoffnung, dass

$$\widehat{\mu} \approx \mu \qquad \text{und} \qquad \widehat{\sigma}_X^2 \approx \sigma_X^2$$

Das Zeichen $\approx$ steht für „ungefähr" oder „angenähert".

In den folgenden Beispielen werden die Schätzungen für das Beispiel der Waage A aus dem Beispiel 6.2.1 der Gewichtsmessung durchführen und auf die Problem- und Fragestellungen hinweisen.

Beispiel 6.2.2

Für unser Beispiel der Waage A lauten die Schätzungen für den Mittelwert μ und die Varianz σ_X^2

$$\widehat{\mu} = 80.02 \qquad \text{und} \qquad \widehat{\sigma}_X^2 = 0.024^2$$

Diese Werte berechnen wir mit **R**

```
waageA <- c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04, 79.97,
    80.05, 80.03, 80.02, 80, 80.02)

mean(waageA)

## [1] 80.02077

sd(waageA)

## [1] 0.02396579
```

Allerdings stellt sich hier das Problem, dass *andere* Messreihen mit der Waage A normalerweise anders lauten. Somit sind dann diese Schätzwerte praktisch immer leicht (hoffentlich) anders. Dies untersuchen wir im folgenden Beispiel. ◀

Beispiel 6.2.3

Wir wollen neue Messreihen simulieren, die „ähnlich" aussehen, wie die Werte der Waage A in Beispiel 6.2.1.

Dazu machen wir die *Annahme*, dass die Messwerte der Waage A normalverteilt sind mit den wahren Parametern $\mu = 80$ und $\sigma_X^2 = 0.02^2$.

Wir können mit **R** Zufallszahlen generieren, die dieser Verteilung folgen. Dies machen wir mit dem Befehl `rnorm(n = ..., mean = ..., sd = ...)`, wobei `n = ...` die Zahl der generierten Zufallszahlen bedeutet. Wir beschränken uns hier der Übersichtlichkeit halber auf Messreihen der Länge 6. Zudem runden wir die Resultate meist auf zwei Nachkommastellen (`round(..., 2)`).

```
set.seed(1)
waageA_sim1 <- round(rnorm(n = 6, mean = 80, sd = 0.02), 2)

waageA_sim1


## [1] 79.99 80.00 79.98 80.03 80.01 79.98


mean(waageA_sim1)


## [1] 79.99833


sd(waageA_sim1)


## [1] 0.0194079
```

Wir sehen, dass die geschätzten Werte $\widehat{\mu}$ und $\widehat{\sigma}^2$ jeweils (leicht) anders sind, als im Beispiel 6.2.2 vorher.

Führen wir dies fünfmal durch, so sehen die Resultate wie folgt aus:

```
set.seed(9)
for (i in 1:5) {
    waageA_sim1 <- round(rnorm(n = 6, mean = 80, sd = 0.02), 2)
    cat(round(mean(waageA_sim1), 2), "\t", round(sd(waageA_sim1),
        2), "\n")
}
```

```
## 79.99   0.01
## 80   0.02
## 79.99   0.03
## 80.01   0.02
## 80.01   0.02
```

Die Mittelwerte liegen hier alle nahe bei 80, was auch zu erwarten war. Wir haben hier keine Zweifel, dass der wahre Mittelwert $\mu = 80$ sein könnte. Diese Abweichungen sind durchaus zu erwarten. ◀

Bemerkung:

Mit `set.seed(...)` wird erreicht, dass immer dieselben Zufallszahlen erzeugt werden. Dies hat den Vorteil, dass sich die Zahlen mit jeder Erstellung dieses Skriptes nicht ändern. ♦

Beispiel 6.2.4

Im Beispiel 6.2.3 vorher liegen die geschätzten Mittelwerte alle sehr nahe bei $\mu = 80$. Allerdings sind auch folgende Fälle möglich:

```
set.seed(1450070)
waageA_sim2 <- rnorm(n = 6, mean = 80, sd = 0.02)

waageA_sim2


## [1] 80.05403 80.03896 80.03671 80.06336 80.01052 80.04372


mean(waageA_sim2)


## [1] 80.04122


sd(waageA_sim2)


## [1] 0.01804572
```

Der Mittelwert dieser Messreihe ist nach Abschnitt 5.3.4 annähernd verteilt wie

$$\overline{X}_6 \sim \mathcal{N}\left(80, \frac{0.02^2}{6}\right) = \mathcal{N}\left(80, 0.0082^2\right)$$

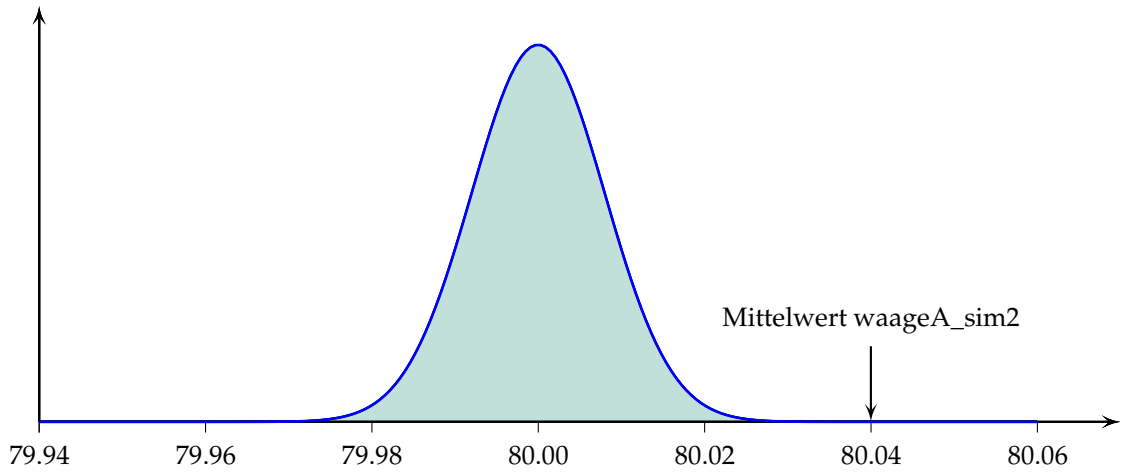


Abbildung 6.1. : Mittelwert, der sehr weit vom erwarteten Wert von $\mu = 80$ entfernt ist

Der Mittelwert dieser Messreihe ist fast 5 Standardabweichungen grösser als 80 (siehe Abbildung 6.1).

Nun kommt der entscheidende Gedanke: Falls der wahre Mittelwert in der Tat $\mu = 80$ ist, so erwarten wir, dass der Mittelwert der Messreihe in der „Nähe" von 80 liegt, wie wir das im Beispiel 6.2.3 gesehen haben.

Aber was heisst hier in der „Nähe"? Wir sehen in Abbildung 6.2 die horizontale Achse in drei Teile aufgeteilt: Der grüne Bereiche in der Mitte und die beiden roten Teile ausserhalb. Wenn nun ein Mittelwert einer Messreihe im „grünen" Bereich liegt, so können wir annehmen, dass das mit dem wahren Mittelwert $\mu = 80$ stimmen dürfte, da der Mittelwert der Messreihe „nahe" bei $\mu = 80$ liegt.

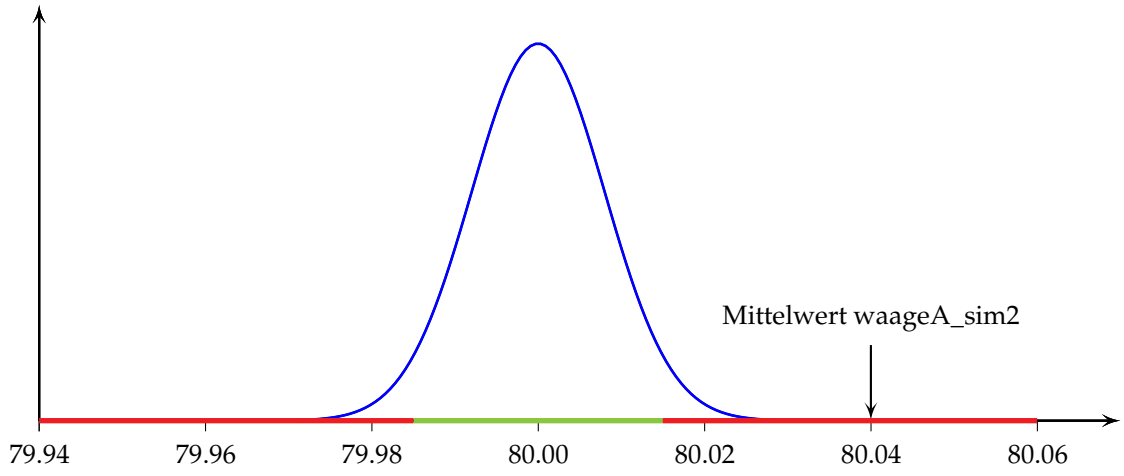


Abbildung 6.2. : Mittelwert, der sehr weit vom erwarteten Wert von $\mu = 80$ entfernt ist

Auf der anderen Seite liegt der Mittelwert der Messreihe **`waageA_sim2`** im roten

Bereich und ist somit zu weit vom Mittelwert entfernt, als dass der wahre Mittelwert vernünftigerweise bei $\mu = 80$ liegt und sagen dann, dass der wahre Mittelwert *nicht*.

Was wir natürlich noch *nicht* gesagt haben, wie gross der grüne Bereich sein soll. Hier machen wir dann eine *Festlegung*, wie gross der grüne Bereich sein soll. ◀

Beispiel 6.2.5

Ein weiteres Beispiel, wo der Mittelwert kleiner als $\mu = 80$ ist:

```
set.seed(505)
waageA_sim3 <- rnorm(n = 6, mean = 80, sd = 0.02)

waageA_sim3

## [1] 79.97758 79.97436 79.95921 79.98125 79.98980 79.99174

mean(waageA_sim3)

## [1] 79.97899

sd(waageA_sim3)

## [1] 0.01182442
```

Hier liegt der Mittelwert etwa 3 Standardabweichungen unter 80 (siehe Abbildung 6.3). Dies ist zwar immer noch weit weg, aber nicht so stark wie in Beispiel 6.2.4.

Wir erwarten allerdings, dass der Durchschnitt der Messreihe in der Nähe vom wahren $\mu = 80$ liegt. Liegt der Durchschnitt weit weg, so beginnen zu zweifeln, ob der wahre Mittelwert tatsächlich 80 ist.

Ist dieser Durchschnitt von 79.98 der Messreihe nun noch gerade im „grünen" oder schon im „roten" Bereich?

Wir werden dies gleich im Abschnitt Hypothesentest 6.2.2 genauer untersuchen.

◀

Die letzten beiden Beispiele führen uns auf die folgenden Fragestellungen:

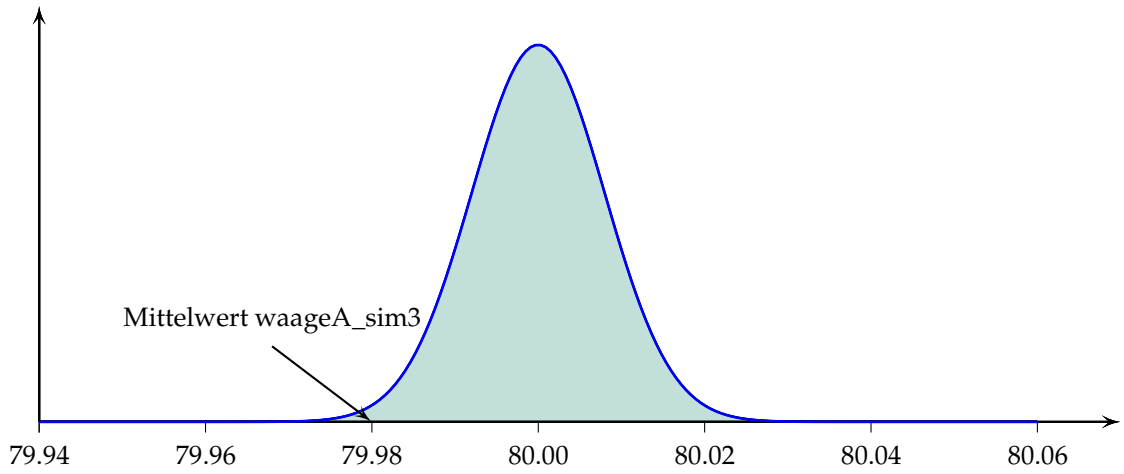


Abbildung 6.3. : Mittelwert, der (zu?) weit vom erwarteten Wert von $\mu = 80$ entfernt ist

- Ist eine Messreihe mit der Annahme $\mu = 80$ noch kompatibel oder müssen an dieser Annahme zweifeln?

  Das heisst: Liegt der Mittelwert der Messreihe in der „Nähe" des wahren Mittelwertes $\mu = 80$ oder liegt er so „weit" entfernt, dass wir an der Angabe des wahren $\mu = 80$ zweifeln müssen?

  Hier stellt sich natürlich die Frage, was „nahe" heisst. Dies werden wir im nächsten Abschnitt beantworten.

- Können wir ein Intervall für $\widehat{\mu}$ angeben, indem sich der wahre Wert von $\mu = 80$ mit einer gewissen Wahrscheinlichkeit befindet?

  Ist also beispielsweise $\widehat{\mu} = 79.98$, so können wir ein Intervall

  $$[79.9667,\ 79.9914]$$

  bestimmen, wo zu 95 % der wahre Mittelwert liegt. Der angenommene wahre Mittelwert von $\mu = 80$ liegt nicht in diesem Intervall und somit ist $\mu = 80$ als wahrer Mittelwert zu unwahrscheinlich.

  Solche Intervalle heissen *Vertrauensintervalle* oder Konfidenzintervalle. wie man diese bestimmt, werden wir im Unterkapitel Vertrauensintervalle 6.5 genauer untersuchen.

Wir wollen nochmals darauf hinweisen, dass der wahre Mittelwert grundsätzlich *nicht* bekannt ist.

Allgemein:

Wann immer wir eine Messreihe neu erheben, werden wir für die betreffende Messreihe unterschiedliche Werte (Realisierungen) von $\widehat{\mu}$ und $\widehat{\sigma}^2_X$ ermitteln.

Es ist also vernünftig, für den Schätzer $\widehat{\mu}$ ein Intervall anzugeben, in welches der wahre Wert von μ mit einer bestimmten Wahrscheinlichkeit fällt. Es handelt sich also darum, für einen geschätzten Parameter ein sogenanntes *Vertrauensintervall* anzugeben.

Weiter kann man sich fragen, ob eine Realisierung von μ kompatibel ist mit einem vermuteten μ_0. Die entsprechende Fragestellung wird durch einen *statistischen Test* beantwortet.

Wir beginnen mit dem *statistischen Test* oder *Hypothesentest*.

6.2.2. Hypothesentest

Nach all den Vorbereitung in den Kapiteln bisher kommen wir zum Hypothesentest, der in vielen Gebieten zahlreiche Anwendungen hat und der auch Big Data oft verwendet wird.

Hypothesentests sind ein wichtiges statistisches Mittel um zu entscheiden, ob eine Messreihe zu einer gewisse Grösse „passt". Hier gehen wir davon, dass wir den wahren Mittelwert *nicht* kennen, aber wir gehen von einem „Idealwert" oder einem vermuteten Wert aus.

Beispiel 6.2.6

Eine Brauerei bestellt eine neue Abfüllmaschine für 500 ml Büchsen. Die Abfüllmaschine füllt *nie genau* 500 ml ab, sondern nur *ungefähr* 500 ml.

Die Brauerei ist daran interessiert, dass die Abfüllmaschine möglichst genau abfüllt. Füllt die Maschine zuviel ab, so ist dies schlecht für die Brauerei, da sie zuviel Bier für denselben Preis verkauft. Füllt sie zuwenig ab, sind die Kunden und der Konsumentenschutz unzufrieden, da diese für den entsprechenden Preis zuwenig Bier bekommen.

Die Herstellerfirma behauptet, dass die Maschine die Büchsen normalverteilt mit $\mu = 500\,\text{ml}$ und $\sigma = 1\,\text{ml}$ abfüllt.

Die Brauerei macht 100 Stichproben. Der Mittelwert dieser Stichproben ist 499.57 ml. Dies ist zwar weniger als 500 ml, aber liegt dies noch „im Rahmen" der Angaben $\mu = 500\,\text{ml}$ und $\sigma = 1\,\text{ml}$ des Herstellers der Abfüllanlage? Wie können wir dies überprüfen? ◀

Beispiel 6.2.7

Wie kann eine Pharmafirma nachweisen, dass ein neues Medikament besser wirkt als das alte? ◀

Beispiel 6.2.8

Eine Anfrage beim Bundesamt für Statistik ergibt, dass die durchschnittliche Körpergrösse der erwachsenen Frauen in der Schweiz bei 180 cm mit einer Standardabweichung von 10 cm liegt.

Diese Angabe für den Mittelwert ist gefühlsmässig wohl falsch, da viel zu hoch. Wie können wir dies aber mathematisch überprüfen und begründen, ohne uns auf unser Gefühl zu verlassen zu müssen? ◀

Ziel dieses Abschnittes ist es, ein standardisiertes, reproduzierbares Verfahren einzuführen, mit dem wir entscheiden können, ob der Mittelwert einer Messreihe zu einem bestimmten „wahren" Mittelwert μ passt oder nicht.

Achtung

Wir wollen hier schon bemerken, dass das kommende Verfahren *niemals einen Beweis* liefert, dass beispielsweise eine Grösse nicht zu einer Messreihe passt.

Wir können mit statistischen Mitteln nur zeigen, dass diese Grösse *mit grosser Wahrscheinlichkeit* nicht zu dieser Messreihe passt.

Lesen Sie in der Zeitung „…mit Statistik bewiesen…", ist das ein Blödsinn!

Wir wollen die wichtigen Begriffe für den Hypothesentest mit Hilfe der Beispiele der Gewichte 6.2.1 und 6.2.5 vertiefen. Zuerst gehen wir informell vor, bevor wir dies mit einem Schema systematisch machen.

Beispiel 6.2.9

Wie in Beispiel 6.2.2 gehen wir davon aus, dass die Daten normalverteilt sind mit $\mu = 80.00$ und $\sigma = 0.02$. Wie können wir überprüfen, ob der Mittelwert $\mu = 80$ auch realistisch ist?

Die Grundidee ist, mit einer Messreihe zu überprüfen, ob *unter dieser Annahme* $\mu = 80$, der Mittelwert dieser Messreihe wahrscheinlich ist oder nicht.

Wir wählen dazu eine Messreihe der Länge 6 aus und gehen von folgendem Modell aus:

Modell

Die 6 Messwerte sind Realisierungen der Zufallsvariablen $X_1, X_2, \dots X_6$, wobei X_i eine kontinuierliche Messgrösse ist. Es soll gelten:

$$X_1, \dots, X_6 \text{ i.i.d. } \sim \mathcal{N}(80, 0.02^2)$$

Diese Modell beschreibt die Voraussetzungen, die wir an die Daten stellen. Diese werden natürlich nicht immer exakt erfüllt sein, aber oft sind diese vereinfachenden Annahmen ausreichend um wichtige Aussagen zu machen.

Wir wollen nun überprüfen, ob die *Annahme* $\mu = 80$ auch gerechfertigt ist. Diese Annahme bezeichnen wir von nun an mit μ_0 und führen folgende Begriffe ein:

Nullhypothese

$$H_0: \quad \mu = \mu_0 = 80$$

Alternativhypothese

$$H_A: \quad \mu \neq \mu_0 = 80 \quad \text{oder „<" oder „>"}$$

Was ist nun der Unterschied zwischen μ und μ_0?

- Der Wert μ ist der wahre, aber unbekannte Mittelwert der Waage A. Dieser ist unbekannt, da wir nicht unendlich viele Messungen durchführen können.
- Der Wert $\mu_0 = 80$ ist eine Annahme über den wahren Mittelwert μ. Die beiden können gleich sein (Nullhypothese) oder eben nicht (Alternativhypothese).
- Diese Annahme wollen wir mit dem Mittelwert $\widehat{\mu}$ einer Messreihe überprüfen, der den wahren Mittelwert, hoffentlich, annähert (schätzt). Also

$$\mu \approx \widehat{\mu}$$

Als Messreihe wählen wir **`waageA.sim3`** aus Beispiel 6.2.5:

```
## [1] 80.00 79.96 79.96 79.96 79.98 79.99
## Mittelwert: 79.975
## Standardabweichung: 0.01760682
```

Der (geschätzte) Mittelwert ist hier $\widehat{\mu} = 79.98$. In Abbildung **??** haben wir die „roten" Bereiche kennengelernt, die angeben, was zu weit weg von der Annahme μ_0 ist. Konkret wird dieser Entscheid aber mit *Wahrscheinlichkeiten* getroffen.

Hier müssen wir zuerst konkretisieren, was es heisst, dass dieser Mittelwert (un)wahrscheinlich ist. Die Wahrscheinlichkeit

$$P(\overline{X}_6 = 79.98)$$

bringt uns hier nicht weiter, da diese 0 ist. Da $\widehat{\mu} < 80$ ist, können wir aber folgende Wahrscheinlichkeit betrachten:

$$P(\overline{X}_6 \leq 79.98)$$

Unter unseren Annahmen $\mu_0 = 80$ und $\sigma = 0.02$ ist $\overline{X}_6$ wie folgt verteilt

$$\overline{X}_6 \sim \mathcal{N}\left(80, \frac{0.02^2}{6}\right)$$

Wir testen mit dieser Verteilung, ob die Annahme $\mu_0 = 80$ gerechtfertig ist.

Teststatistik

Verteilung der Teststatistik T unter der Nullhypothese H_0:

$$T = \overline{X}_6 \sim \mathcal{N}\left(80, \frac{0.02^2}{6}\right)$$

Damit erhalten wir für die Wahrscheinlichkeit

$$P(\overline{X}_6 \leq 79.98) = 0.007$$

```
pnorm(q = 79.98, mean = 80, sd = 0.02/sqrt(6))


## [1] 0.007152939
```

Diese Wahrscheinlichkeit ist klein, 0.7 %. Ist sie aber *zu* klein (siehe Abbildung 6.4)?

Da nirgendwo geschrieben steht, was zu „klein" ist, müssen wir eine eine *Abmachung* machen: Es hat sich als praktisch erwiesen, diese Grenze, was zu klein ist und was nicht bei 2.5 % festzulegen (siehe Abbildung 6.5). Warum dies 2.5 % sind, werden wir gleich sehen.

Gemäss dieser Abmachung ist

$$P(\overline{X}_6 \leq 79.98) < 0.025$$

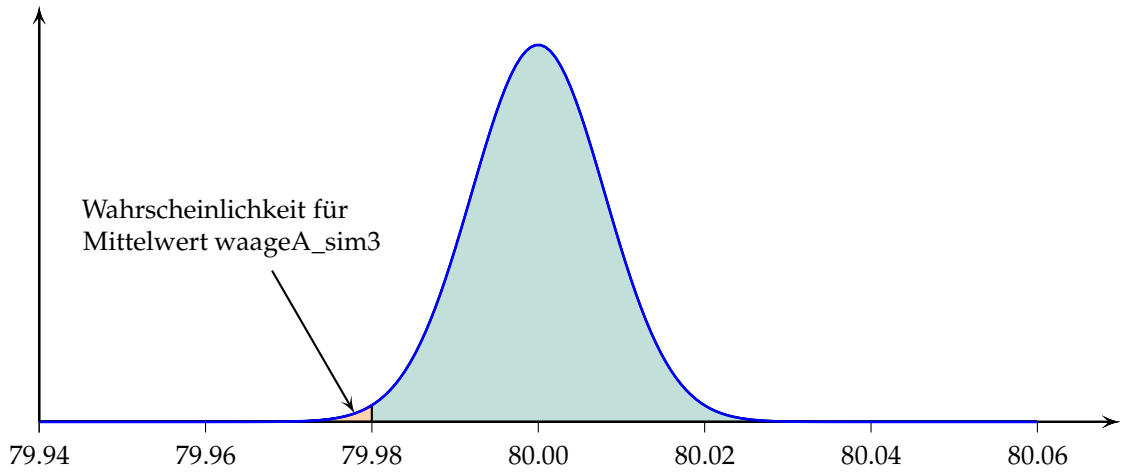


Abbildung 6.4. : Wahrscheinlichkeit für den Mittelwert `waageA_sim3` als Fläche dargestellt

und wir betrachten diesen geschätzen Mittelwert $\widehat{\mu} = 79.98$ als *zu unwahrscheinlich*, als dieser zum Wert $\mu_0 = 80$ passen könnte (siehe auch Abbildung 6.6). *Wir gehen also davon aus, dass der angegebene Mittelwert von* $\mu_0 = 80$ *nicht stimmen kann!*. Also gilt

$$\mu \neq \mu_0$$

Für das bessere Verständnis wollen wir diesen Sachverhalt noch graphisch darstellen. In Abbildung 6.5 teilen wir die Normalverteilungskurve in drei Teile auf:

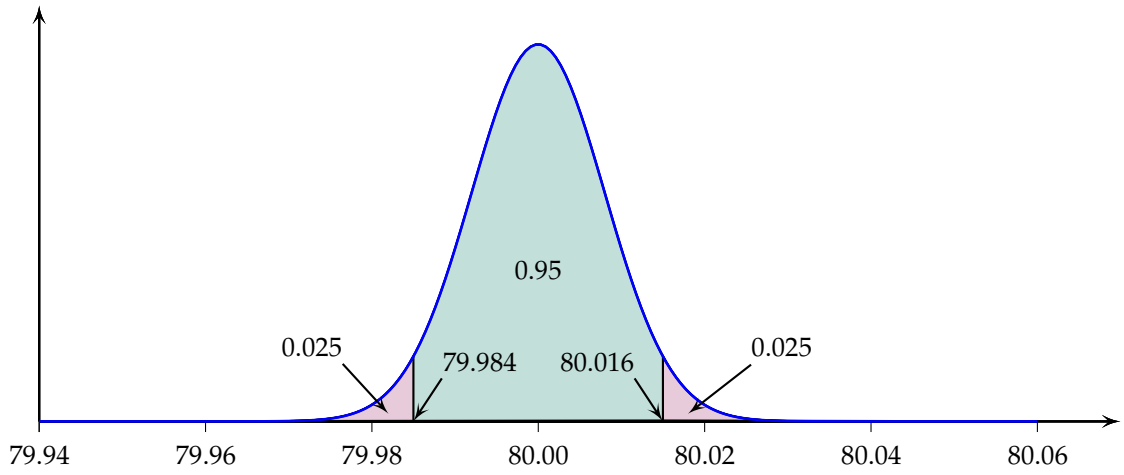


Abbildung 6.5. : Verwerfungsbereich 1

- Der symmetrische Teil um den Mittelwert $\mu_0 = 80$ soll 0.95, also 95 % betragen.
- Die beiden Teilen links und rechts müssen zusammen 0.05 ergeben. Also ergibt sich für jeden Teil 0.025.

- Die Grenzen entsprechen den 0.025- und 0.975-Quantilen.

```
qnorm(p = c(0.025, 0.975), mean = 80, sd = 0.02/sqrt(6))

## [1] 79.984 80.016
```

- Die Fläche 0.05 des gesamten roten Bereiches heisst *Signifikanzniveau*.

Signifikanzniveau α

Das Signifikanzniveau α, gibt an, wie hoch das Risiko ist, das man bereit ist einzugehen, eine falsche Entscheidung zu treffen.

Für die meisten Tests wird ein α-Wert von 0.05 bzw. 0.01 verwendet.

Wir verwenden hier

$$\alpha = 0.05$$

Liegt der gemessene Mittelwert im roten Bereich in Abbildung 6.5, so zweifeln wir an der Nullhypothese

$$H_0: \quad \mu_0 = 80$$

Wir sagen, wir *verwerfen* die Nullhypothese $\mu_0 = 80$. Wir nennen diesen Bereich, wo die Nullhypothese verworfen wird, deshalb auch

Verwerfungsbereich

$$K = (-\infty, 79.984] \cup [80.016, \infty)$$

Wir gehen also davon aus, dass ein Mittelwert einer Messreihe im Verwerfungsbereich so unwahrscheinlich ist, dass wir an der Richtigkeit von $\mu_0 = 80$ zweifeln und annehmen müssen, dass das wahre μ nicht 80 ist.

Nun können wir mit unserer Messreihe überprüfen, ob deren Mittelwert im Verwerfungsbereich liegt oder nicht und machen den sogenannten

Testentscheid

In unserem Beispiel hatten wir (siehe Abbildung 6.6)

$$\overline{X}_6 = 79.98 \in K$$

Dieser Wert liegt im Verwerfungsbereich. Also gehen wir nicht vom wahren $\mu = 80$ aus, da der Mittelwert der Messreihe nicht zu diesem Parameter passt. Das

heisst, dieser Wert ist zu unwahrscheinlich, als dass $\mu_0 = 80$ plausibel ist.

Wir verwerfen also die Nullhypothese und nehmen die Alternativhypothese an, d. h.

$$\mu \neq 80$$

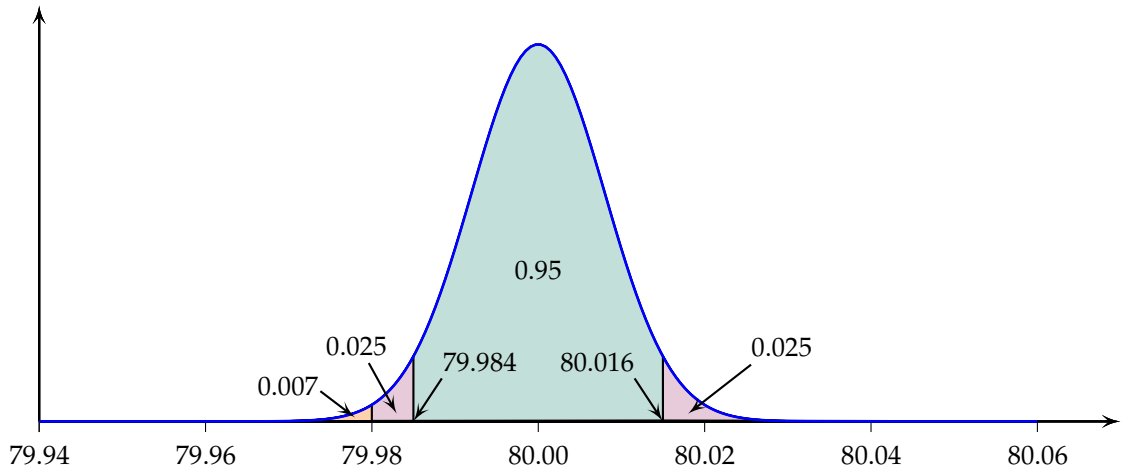


Abbildung 6.6. : Verwerfungsbereich 2

◀

Bemerkungen:

i. Warum haben wir hier den Verwerfungsbereich nach oben und nach unten aufgeteilt, wenn wir schon wissen, dass der gemessene Mittelwert kleiner als $\mu_0 = 80$ ist?

   Nun, das wussten wir *vor* der Messung nicht. Der gemessene Mittelwert hätte also durchaus auch grösser als $\mu_0 = 80$ sein können (siehe Beispiel 6.2.4).

   Wir sprechen in diesem Fall von einem *zweiseitigen Test*.

ii. Es gibt auch *einseitige Tests* (siehe Beispiel 6.2.12).

iii. Wir haben hier eine *Annahme* gemacht, das der gesamte Verwerfungsbereich 5 % betragen soll. Diese Annahme hat sich als praktisch erwiesen, aber wir hätten auch 1 % wählen können, was auch ab und zu gemacht wird.

iv. Im Beispiel 6.2.5 folgt die zufällige Messreihe der Normalverteilung $\mathcal{N}(80, 0.02^2)$ und doch wird der Parameter $\mu_0 = 80$ hier als unwahrscheinlich verworfen. Dies heisst, wir haben hier einen *Fehler* gemacht.

♦

Beispiel 6.2.10

Wählen wir als *andere* Messreihe Beispiel 6.2.4, so ist Modell, Nullhypothese, Alternativhypothese, Teststatistik, Signifikanzniveau und Verwerfungsbereich gleich wie im Beispiel 6.2.9 vorher. Wir müssen also nur noch den Testentscheid durchführen.

```
## Error in eval(expr, envir, enclos): object 'waageA.sim2' not found
## Error in mean(waageA.sim2): object 'waageA.sim2' not found
```

Der geschätzte Mittelwert ist im Verwerfungsbereich und somit wird auch hier die Nullhypothese verworfen.

Es gilt für die Wahrscheinlichkeit

$$P(\overline{X}_6 > 80.04) \approx 5 \cdot 10^{-7}$$

```
1 - pnorm(q = 80.04, mean = 80, sd = 0.02/sqrt(6))

## [1] 4.816785e-07
```

Dieser ist bei weitem kleiner als 0.025 und damit so unwahrscheinlich, dass wir auch auf diese Weise $\mu = 80$ als nicht richtig annehmen (müssen), also

$$\mu \neq \mu_0$$

Wir *verwerfen* die Nullhypothese.

Im Gegensatz zum Verwerfungsbereich, wo nur die Entscheidung gefällt wird, ob der geschätzte Mittelwert im Verwerfungsbereich liegt oder nicht, macht der Wert von $P(\overline{X}_6 > 80.04)$ noch eine Aussage über die *Sicherheit* des Verwerfen. In diesem Fall ist $5 \cdot 10^{-7}$ *sehr viel kleiner* als 0.025 und damit können wir mit grosser Sicherheit davon ausgehen, dass $\mu = 80$ *nicht* gilt. Wir kommen beim p-Wert auf diesen Sachverhalt zurück.

Aber auch hier sei nochmals erwähnt, dass diese Messreihe von der wirklichen Verteilung $\mathcal{N}(80.00, 0.02^2)$ stammt. Allerdings ist sie so unwahrscheinlich, dass wir an der Annahme $\mu_0 = 80$ zweifeln müssen. ◀

Beispiel 6.2.11

Wir wollen nun testen, ob die Angabe in Beispiel 6.2.6 mit der Testreihe konform ist.

Die Herstellerfirma behauptet, dass die Maschine die Büchsen normalverteilt mit $\mu = 500\,\text{ml}$ und $\sigma = 1\,\text{ml}$ abfüllt. Brauerei macht 100 Stichproben. Der Mittelwert dieser Stichproben ist 499.84 ml. Wir nehmen an die Messungen sind normalverteilt mit bekanntem $\sigma = 1$.

Modell

X_i: Inhalt der i-ten Büchse

$$X_1, \ldots, X_{100} \text{ i.i.d. } \sim \mathcal{N}(\mu, 1^2)$$

Nullhypothese

$$H_0: \qquad \mu_0 = 500$$

Alternativhypothese

$$H_A: \qquad \mu \neq \mu_0 = 500$$

Teststatistik mit Signifikanzniveau $\alpha = 0.05$:

$$\overline{X}_{100} \sim \mathcal{N}\left(500, \frac{1^2}{100}\right)$$

Verwerfungsbereich

Die Grenze des Verwerfungsbereichs ermitteln wir durch

```
qnorm(p = c(0.025, 0.975), mean = 500, sd = 1/sqrt(100))

## [1] 499.804 500.196
```

Also

$$K = (-\infty, 499.804) \cup (500.196, \infty)$$

Testentscheid

Es gilt

$$499.84 \notin K$$

Somit wird die Nullhypothese nicht verworfen. Wir vertrauen der Angabe des Hersteller der Abfüllanlage. ◀

Beispiel 6.2.12

Wir kommen auf das Beispiel 6.2.8 zurück. Eine Zeitung behauptet, dass die durchschnittliche Körpergrösse der erwachsenen Frauen in der Schweiz bei 180 cm mit einer Standardabweichung von 10 cm liegt.

Unsere Vermutung ist, dass dieser Wert zu gross ist. Hier macht ein zweiseitiger Test wenig Sinn, da wir „wissen", dass dieser Mittelwert zu gross ist. Das heisst, der wahre Wert liegt wohl eher tiefer.

Die Überlegung ist an sich dieselbe wie in Beispiel 6.2.9, wobei wir aber den Verwerfungsbereich nicht auf beide Seiten verteilen, sondern nur nach unten, da wir erwarten, dass der wahre Mittelwert tiefer als $\mu_0 = 180$ ist (siehe Abbildung 6.7). Wir machen einen *einseitigen* Test.

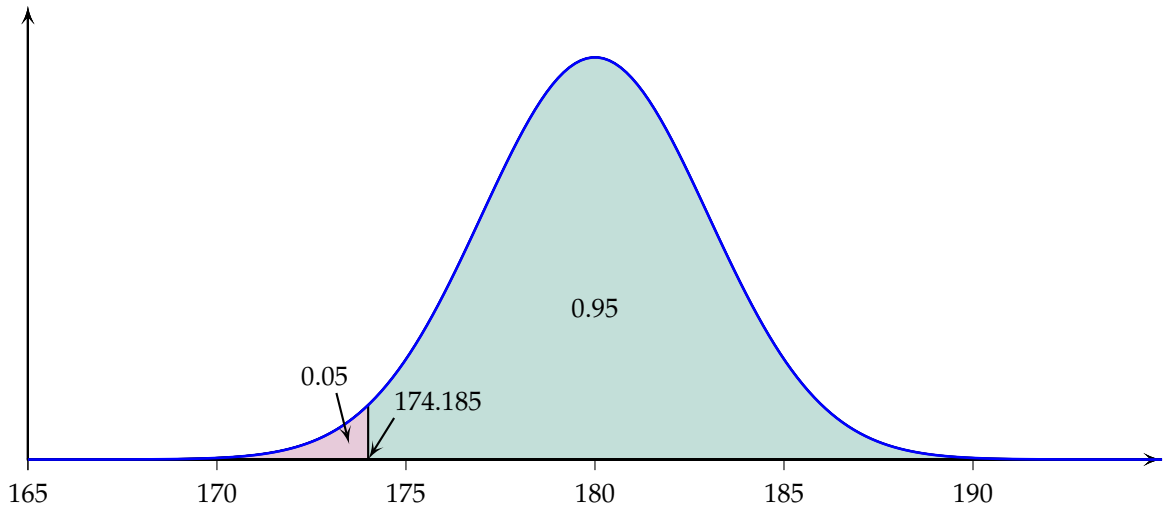


Abbildung 6.7. : Verwerfungsbereich Körpergrösse

Wir wählen zufällig 8 erwachsene Frauen aus, deren durchschnittliche Körpergrösse 171.54 cm beträgt.

Es wird angenommen, dass die Körpergrösse normalverteilt ist mit $\mathcal{N}(\mu, 10^2)$. Wir nehmen weiter an, dass die Standardabweichung dieselbe ist, wie vom Bundesamt angegeben.

Modell:

X_i: Körpergrösse der i–ten Frau. Es gilt

$$X_1, \ldots, X_8 \text{ i.i.d. } \sim \mathcal{N}(\mu, 10^2)$$

Nun gehen wir davon aus, dass der wahre Mittelwert wirklich 180 cm ist.

Nullhypothese

$$H_0: \qquad \mu_0 = 180$$

Alternativhypothese

$$H_A: \qquad \mu < \mu_0 = 180$$

Wir bilden den Mittelwert der 8 Personen und testen, ob jetzt der Wert

$$P(\overline{X}_8 < \overline{x}_8) < 0.05$$

ist oder nicht. Der Verwerfungsbereich ist hier also einseitig nach unten. In Abbildung 6.7) ist der Verwerfungsbereich für $n = 8$ pink eingezeichnet.

Teststatistik mit Signifikanzniveau $\alpha = 0.05$

$$\overline{X}_8 \sim \mathcal{N}\left(180, \frac{10^2}{8}\right)$$

Die Grenze des Verwerfungsbereichs ermitteln wir durch

```
qnorm(p = 0.05, mean = 180, sd = 10/sqrt(8))

## [1] 174.1846
```

Verwerfungsbereich (siehe Abbildung 6.7)

Der Verwerfungsbereich ist also

$$K = (-\infty, 174.185)$$

Dieser Verwerfungsbereich ist natürlich viel zu gross, da wohl kaum Körpergrössen von erwachsenen Frauen unter 50 cm zu erwarten sind. Wir arbeiten hier mit einem *Modell*, das eben nur in einem bestimmten Bereich Sinn macht.

Testentscheid

So ist Wert im Verwerfungsbereich und somit *verwerfen* wir die Nullhypothese, dass das wahre $\mu = 180$ gilt.

Dieser Mittelwert der zufällig ausgewählten acht Frauen erscheint immer noch relativ hoch, aber er reicht schon, damit wir an der Annahme $\mu = 180$ zweifeln müssen.

Der Wert für $P(\overline{X}_6 < 171.54)$ ist (siehe Abbildung 6.8).

$$P(\overline{X}_6 < 171.54) = 0.008$$

```
pnorm(q = 171.54, mean = 180, sd = 10/sqrt(8))


## [1] 0.008359052
```

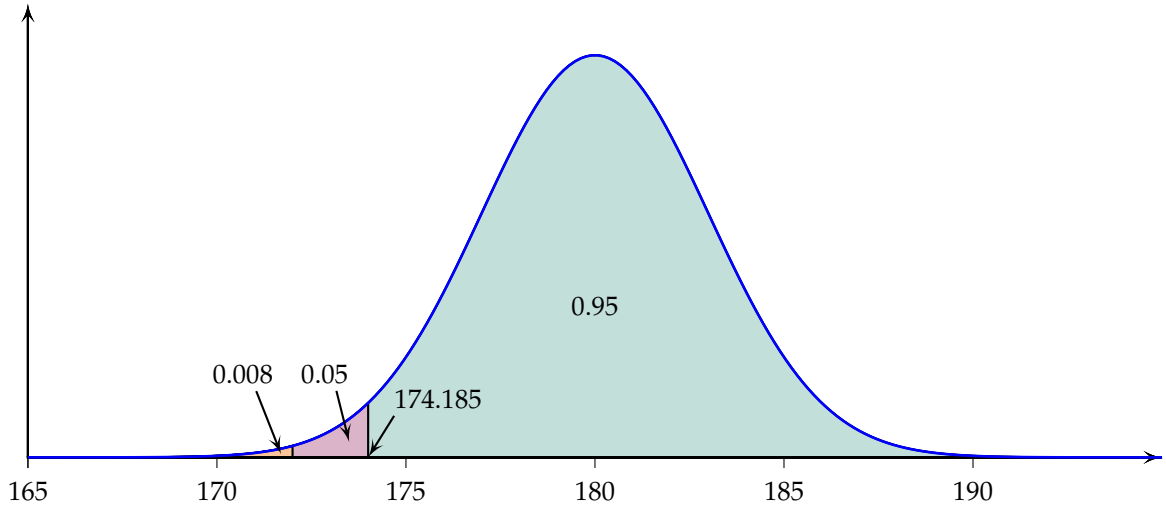


Abbildung 6.8. : Verwerfungsbereich Körpergrösse

Dieser Wert heisst p-Wert und gibt die Sicherheit mit der wir den Testentscheid treffen. Wird die Nullhypothese verworfen, so deutet ein sehr kleiner p-Wert darauf hin, dass die Nullhypothese sicherer verworfen wird, als wenn er in der Nähe des Signifikanzniveaus (hier $\alpha = 0.05$) liegt. ◀

Wir haben in Beispiel 6.2.9 die Grösse $\mu = 80$ verworfen, da die Messreihe einen zu tiefen Mittelwert lieferte. In Beispiel 6.2.13 haben wir die Grösse $\mu = 80$ verworfen, weil der beobachtete Mittelwert (viel) zu gross ist. In beiden Werten Fällen waren die Mittelwerte so unwahrscheinlich, dass wir an dem Mittelwert $\mu = 80$ zweifeln müssen.

Wie sieht es nun aber aus, wenn wir eine neue Messreihe bilden, die aus *beiden* Messreihen **sim_2** und **sim_3** besteht? Diese Messreihe hat dann die Länge 12.

Oder anders gefragt: Welchen Einfluss hat die Anzahl der Messungen auf den Verwerfungsbereich?

Beispiel 6.2.13

Wir wollen dies am Beispiel **??** untersuchen und gehen wie folgt vor: Wir haben Messreihen verschiedener Länge n, die alle den geschätzten Mittelwert $\widehat{\mu} = 79.78$ haben. Dann bestimmen wir für alle Messreihen den Wert

$$P(\overline{X}_n \leq 79.98)$$

mit

$$\overline{X}_n \sim \mathcal{N}\left(80, \frac{0.02^2}{n}\right)$$

Ist dieser Wert grösser als 0.025, dann wird die Nullhypothese nicht verworfen, ansonsten schon.

Für $n = 2$ erhalten wir folgenden Wert für

$$P(\overline{X}_2 \leq 79.98) = 0.079 > 0.025$$

Die Nullhypothese wird also nicht verworfen. Bei 2 Messwerten erachten wir die Abweichung vom wahren Mittelwert als zufällig möglich.

```
pnorm(q = 79.98, mean = 80, sd = 0.02/sqrt(2))
```

```
## [1] 0.0786496
```

Für $n = 4$ erhalten wir

$$P(\overline{X}_4 \leq 79.98) = 0.022 < 0.025$$

Hier wird die Nullhypothese (knapp) verworfen.

```
pnorm(q = 79.98, mean = 80, sd = 0.02/sqrt(4))
```

```
## [1] 0.02275013
```

Für $n = 6$ erhalten wir

$$P(\overline{X}_6 \leq 79.98) = 0.007 < 0.025$$

Die Nullhypothese wird klarer verworfen als für $n = 4$.

```
pnorm(q = 79.98, mean = 80, sd = 0.02/sqrt(6))

## [1] 0.007152939
```

Und schlussendlich noch für $n = 8$:

$$P(\overline{X}_6 \leq 79.98) = 0.002 < 0.025$$

Die Nullhypothese wird noch klarer verworfen, als bei $n = 6$.

```
pnorm(q = 79.98, mean = 80, sd = 0.02/sqrt(8))

## [1] 0.002338867
```

Mit zunehmendem n wird der Wert

$$P(\overline{X}_n \leq 79.98)$$

immer kleiner. Dies liegt daran, dass die Standardabweichung mit grösser werdendem n kleiner wird und damit werden die Normalverteilungskurven schmaler (siehe Abbildung 6.9).

Das heisst, je mehr Messungen wir haben, umso gewichtiger ist eine Abweichung von wahren Mittelwert.

◀

6.2.3. Der p-Wert

Der p-Wert ist ein Wert zwischen 0 und 1, der angibt, wie gut *Nullhypothese* und *Daten* zusammenpassen (0: passt gar nicht; 1: passt sehr gut).

Etwas präziser formuliert, definieren wir den P-Wert als die Wahrscheinlichkeit, unter Gültigkeit der Nullhypothese das erhaltene Ergebnis oder ein *extremeres* zu erhalten (siehe Abbildung 6.10).

Mit dem p-Wert wird also angedeutet, wie extrem das Ergebnis ist: Je kleiner der p-Wert, desto mehr spricht das Ergebnis gegen die Nullhypothese. Werte kleiner als eine im voraus festgesetzte Grenze, wie 5 %, 1 % oder 0.1 % sind Anlass, die Nullhypothese abzulehnen.

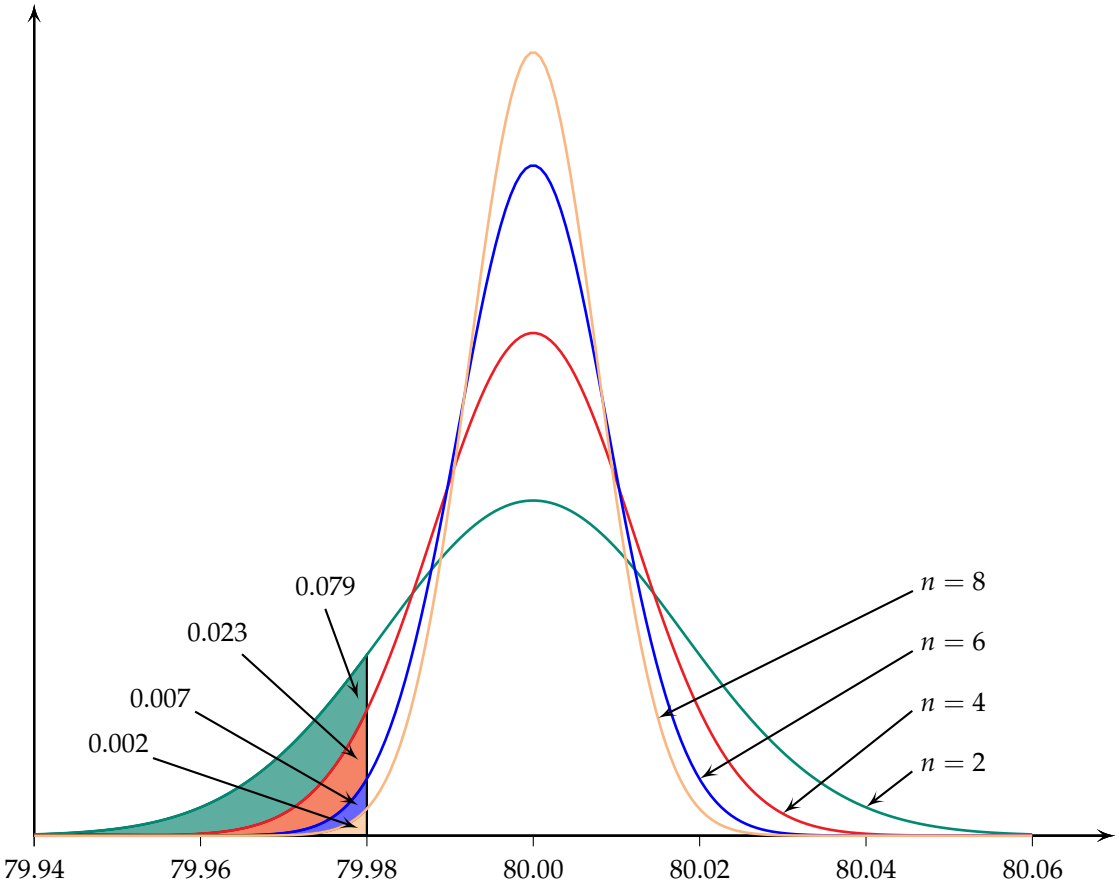


Abbildung 6.9. : Verwerfungsbereich abhängig von der Anzahl Messungen

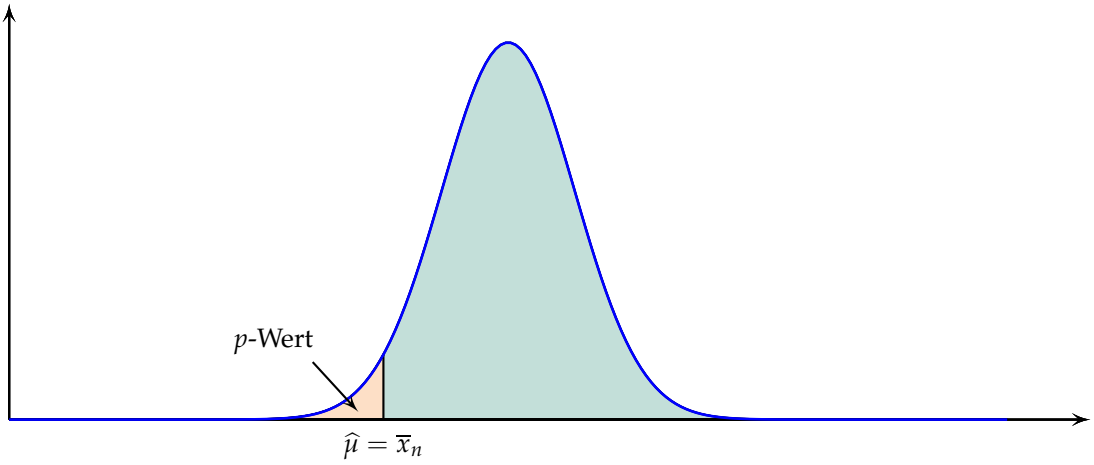


Abbildung 6.10. : p-Wert, einseitig

p-Wert

Der *p-Wert* ist die Wahrscheinlichkeit, unter der Nullhypothese ein mindestens so extremes Ereignis (in Richtung der Alternative) zu beobachten wie das aktuell beobachtete.

Wir können den Testentscheid auch mit Hilfe des p-Wertes durchführen.

p-Wert und Statistischer Test

Man kann anhand des p-Werts direkt den Testentscheid ablesen: Wenn der p-Wert kleiner als das Niveau ist, so verwirft man H_0, ansonsten nicht. Verglichen mit dem reinen Testentscheid enthält der p-Wert aber mehr Information, da man direkt sieht, „wie stark" die Nullhypothese verworfen wird. Bei einem vorgegebenen Signifikanzniveau α (z.B. $\alpha = 0.05$) gilt aufgrund der Definition des p-Werts für einen einseitigen Test:

1. Verwerfe H_0 falls p-Wert $\leq \alpha$
2. Belasse H_0 falls p-Wert $> \alpha$

Viele Computer-Pakete liefern den Testentscheid nur indirekt, indem der p-Wert angegeben wird.

Man kann sich den p-Wert auch als „vollstandardisierte" Teststatistik vorstellen.

Zusätzlich zu dieser Entscheidungsregel quantifiziert der p-Wert, *wie signifikant* eine Alternative ist (d.h. wie gross die Evidenz ist für das Verwerfen von H_0). Manchmal werden sprachliche Formeln oder Symbole anstelle der p-Werte angegeben:

$$\begin{aligned} p\text{-Wert} &\approx 0.05 \;:\; \text{schwach signifikant, ``.''} \\ p\text{-Wert} &\approx 0.01 \;:\; \text{signifikant, ``}*\text{''} \\ p\text{-Wert} &\approx 0.001 \;:\; \text{stark signifikant, ``}**\text{''} \\ p\text{-Wert} &\leq 10^{-4} \;:\; \text{äusserst signifikant, ``}***\text{''} \end{aligned}$$

Achtung

Der p-Wert ist nicht die Wahrscheinlichkeit, dass die Nullhypothese stimmt. Dar-

über können wir hier gar keine Aussagen machen, da die Parameter fix und nicht zufällig sind.

Wir haben den p-Wert für einseitige Tests definiert. Wie sieht nun aber der p-Wert für zweiseitige Tests aus?

Beispiel 6.2.14

In Beispiel 6.2.9 haben wir die Wahrscheinlichkeit

$$P(\overline{X}_6 \leq 79.98) = 0.007$$

der kleiner ist als 0.025. Wir könnten dies als P-Wert betrachten. Da aber das Signifikanzniveau auf $\alpha = 0.05$ liegt, wird die Wahrscheinlichkeit oben auf 5 % umgerechnet, also verdoppelt:

$$p\text{-Wert} = 2 \cdot P(\overline{X}_6 \leq 79.98) = 0.014$$

Dieser p-Wert wird dann mit dem Signifikanzniveau verglichen. ◀

Bemerkungen:

i. Computersoftware gibt den p-Wert *immer* auf Signifikanzniveau an. ♦

6.2.4. Der z-Test (σ_X bekannt)

Wir wollen nun den Hypothesentest aus den Beispielen vorher schematisch festhalten.

Wir nehmen an, dass die Daten $x_1, \ldots, x_n$ Realisierungen von

$$X_1, \ldots, X_n \text{ i.i.d.} \sim \mathcal{N}(\mu, \sigma_X^2)$$

sind. Überdies machen wir die Annahme, dass σ_X^2 bekannt ist. Der *z-Test* für den Parameter μ erfolgt dann wie folgt.

1. *Modell:* X_i ist eine kontinuierliche Messgrösse:

$$X_1, \ldots, X_n \text{ i.i.d.} \sim \mathcal{N}(\mu, \sigma_X^2), \quad \sigma_X \text{ bekannt}$$

2. *Nullhypothese:*

$$H_0: \quad \mu = \mu_0$$

Alternative:

$$H_A: \quad \mu \neq \mu_0 \quad \text{(oder „<" oder „>")}$$

3. *Teststatistik: Verteilung der Teststatistik unter H_0:*

$$T: \overline{X}_n \sim \mathcal{N}\left(\mu_0, \frac{\sigma_X^2}{n}\right)$$

4. *Signifikanzniveau:*

$$\alpha$$

5. *Verwerfungsbereich für die Teststatistik:*

$$K = (-\infty, x_{\frac{\alpha}{2}}] \cup [x_{1-\frac{\alpha}{2}}, \infty) \text{ bei } H_A: \ \mu \neq \mu_0,$$

$$K = (-\infty, x_\alpha] \text{ bei } H_A: \ \mu < \mu_0,$$

$$K = [x_{1-\alpha}, \infty) \text{ bei } H_A: \ \mu > \mu_0$$

wobei

$$x_{\alpha/2}$$

das $\alpha/2$-Quantil.

6. *Testentscheid:*

Überprüfe, ob der beobachtete Wert der Teststatistik im Verwerfungsbereich liegt.

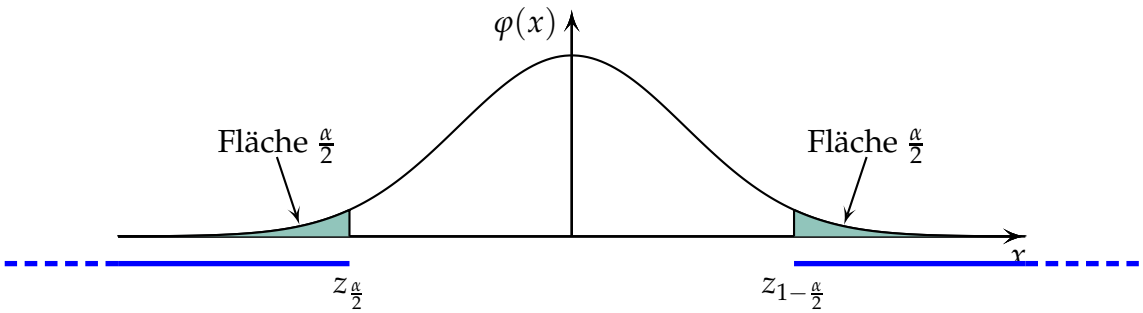


Abbildung 6.11. : Dichtefunktion der Teststatistik z mit Verwerfungsbereich (blau) des zweiseitigen Z-Tests zum Niveau α. Beachte $z_{\alpha/2} = -z_{1-\alpha/2}$, wobei $z_{\alpha/2} = \Phi^{-1}(\alpha/2)$ ist.

Der z-Test basiert auf *mehreren* Beobachtungen. Denn die realisierte Teststatistik z fasst die Beobachtungen in Form des arithmetischen Mittelwertes zusammen.

Beispiel 6.2.15

Bei der Waage A (vgl. Beispiel 6.2.1) scheinen die Gewichte grösser als 80.00 zu sein. Angenommen, wir wissen aus vorhergehenden Studien, dass die Standardabweichung unseres Messinstruments $\sigma_X = 0.01$ ist. Ist es plausibel, dass die Schmelzwärme genau 80.00 cal/g ist? Wir führen dazu einen z-Test durch:

1. *Modell:* X_i ist eine kontinuierliche Messgrösse;

$$X_1, \ldots, X_n \text{ i.i.d.} \sim \mathcal{N}(\mu, \sigma_X^2), \quad \sigma_X = 0.01 \text{ bekannt}, n = 13$$

2. *Nullhypothese:*

$$H_0: \quad \mu = \mu_0 = 80.00$$

Alternative:

$$H_A: \quad \mu \neq \mu_0$$

3. *Teststatistik:*

Der Mittelwert der Messungen

$$T: \quad \overline{X}_n$$

Verteilung der Teststatistik unter H_0:

$$T \sim \mathcal{N}\left(\mu_0, \frac{\sigma_X^2}{n}\right) = \mathcal{N}\left(80, \frac{0.01^2}{13}\right)$$

4. *Signifikanzniveau:*

$$\alpha = 0.05$$

5. *Verwerfungsbereich für die Teststatistik:*

$$K = (-\infty, c_u] \cup [c_o, \infty) \quad \text{bei} \quad H_A: \ \mu \neq \mu_0,$$

Mit **R** erhalten wir mit $\alpha = 0.05$:

```
qnorm(p = 0.025, 80, 0.01/sqrt(13))

## [1] 79.99456
```

```
qnorm(p = 0.975, 80, 0.01/sqrt(13))

## [1] 80.00544
```

oder einfacher

```
qnorm(p = c(0.025, 0.975), 80, 0.01/sqrt(13))

## [1] 79.99456 80.00544
```

Damit erhalten wir den Verwerfungsbereich der Teststatistik:

$$K = (-\infty, 80.00] \cup [80.01, \infty)$$

6. *Testentscheid:*

   Aus den $n = 13$ Daten errechnen wir

   $$\overline{x}_n = 80.02$$

   Der beobachtete Wert liegt im Verwerfungsbereich der Teststatistik. Daher wird die Nullhypothese auf dem 5 % Signifikanzniveau verworfen.

◀

6.3. Schlussbemerkung

Wir haben nun schon einige Male über diese 5 % für das Signifikanzniveau gesprochen. Aber woher kommen diese 5 %

> *In preparing this table we have borne in mind that in practice we do not want to know the exact value of P for any observed χ^2 , but, in the first place, whether or not the observed value is open to suspicion. If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts.* We shall not often be astray if we draw a conventional line at .05 *and consider that higher values of χ^2 indicate a real discrepancy.*
>
> R. A. Fisher, 1925

6.4. t-Test

6.4.1. Der t-Test (σ_X unbekannt)

Das Verfahren im letzten Kapitel heisst z-Test. Es setzt, was wir stillschweigend angenommen haben, voraus, dass die Standardabweichung bekannt ist. Dies ist in der Praxis praktisch nie der Fall. Der folgende sogenannte *t-Test* ist deswegen der wesentlich wichtigere Test, er nimmt keine bekannten Standardabweichung voraus.

Wie vorhin nehmen wir an, dass die Daten Realisierungen von

$$X_1, \ldots, X_n \text{ i.i.d.} \sim \mathcal{N}(\mu, \sigma_X^2)$$

sind. In der Praxis ist die Annahme, dass σ_X bekannt ist, oftmals unrealistisch. In solchen Fällen kann die Teststatistik z (Verfahren mit bekannter Standardabweichung) nicht berechnet werden, weil sie auf σ_X basiert.

Allerdings können wir stattdessen die Schätzung

$$\widehat{\sigma}_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

benutzen. Dies führt aber zu einer zusätzlichen Unsicherheit, was zur Folge hat, dass sich die Verteilung der Teststatistik ändert.

t-Verteilung

Die Verteilung der Teststatistik beim t-Test unter der Nullhypothese

$$H_0: \quad \mu = \mu_0$$

ist gegeben durch[a]

$$T = \overline{X}_n \sim t_{n-1}\left(\mu, \frac{\widehat{\sigma}_X}{\sqrt{n}}\right)$$

wobei t_{n-1} eine *t-Verteilung* mit $n-1$ Freiheitsgraden ist

[a] Die folgende Schreibweise ist nicht Standard, aber sie reicht für unsere Bedürfnisse

Die Normalverteilung wird also durch eine t-Verteilung ersetzt. Was ist aber eine t-Verteilung? Diese ist ähnlich einer Normalverteilung, aber flacher, wegen der grösseren Unsicherheit (unbekannte wahre Standardabweichung). Sie ist auch abhängig

von der Anzahl Beobachtungen. Im Abbildung 6.12 ist die t-Verteilung für $\mu = 0$ und $\sigma \approx 1$ (hängt von n ab) skizziert.

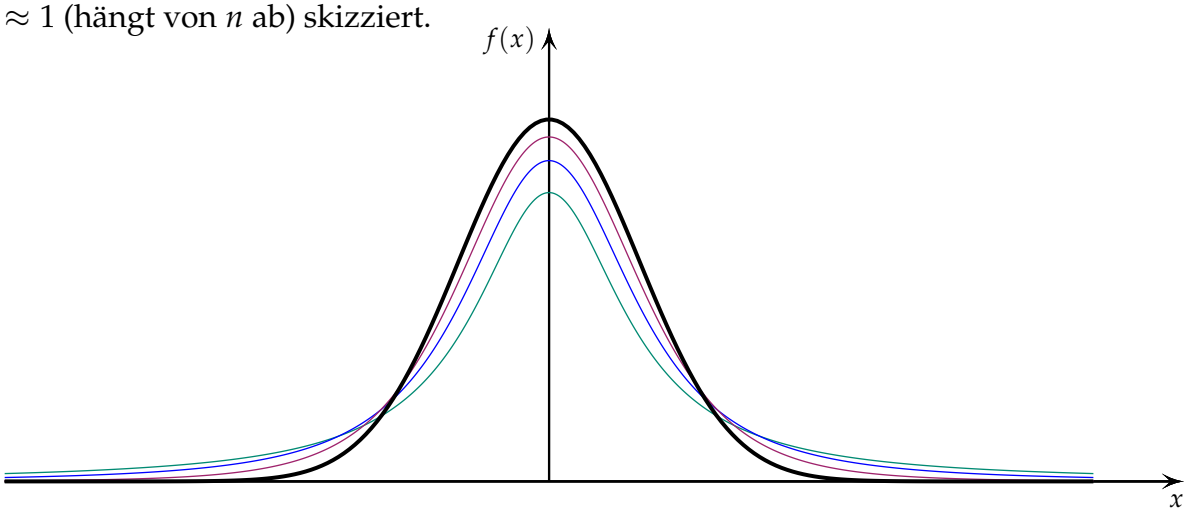


Abbildung 6.12. : Dichten der t-Verteilung mit 1 (grün), 2 (blau) und 5 (violett) Freiheitsgraden. Die schwarze Kurve ist die Dichte der Standardnormalverteilung.

Die t-Verteilung wurde von William Gosset (Chefbrauer Guiness Brauerei) 1908 gefunden.

Wichtig: Für t-Test müssen wir t_{n-1} verwenden.

Alle Begriffe vom z-Test können für den t-Test übernommen werden. Den Verwerfungsbereich bestimmen wir dem `R`-Befehl `qt(...)`[1] anstatt `qnorm(...)` und den p-Wert mit `pt(...)` anstatt `pnorm(...)`.

Da dieser t-Test sehr oft vorkommt, wurde das ganze Verfahren in `R` implementiert. Wir müssen nur noch die Daten in den Befehl `t.test(...)` eingeben und `R` übernimmt uns die Arbeit. Allerdings wird der Verwerfungsbereich nicht ausgegeben, sondern nur der p-Wert ausgegeben, aber der reicht für den Testentscheid völlig aus.

Beispiel 6.4.1

Der folgende Datensatz besteht aus normalverteilten Datenpunkten $x_1, \ldots, x_{20}$

5.9	3.4	6.6	6.3	4.2	2.0	6.0	4.8	4.2	2.1
8.7	4.4	5.1	2.7	8.5	5.8	4.9	5.3	5.5	7.9

[1] Allerdings müssen der Mittelwert hier sogenannt standardisiert werden. Wir werden das allerdings nicht brauchen.

Wir vermuten, dass unsere Daten $x_1, x_2, \ldots, x_{20}$ Realisierungen von

$$X_i \sim \mathcal{N}(5, \sigma_X)$$

sind, wobei wir σ_X nicht kennen. Wir müssen σ_X also aus den Daten schätzen.

```
x <- c(5.9, 3.4, 6.6, 6.3, 4.2, 2, 6, 4.8, 4.2, 2.1, 8.7, 4.4, 5.1,
    2.7, 8.5, 5.8, 4.9, 5.3, 5.5, 7.9)

mean(x)


## [1] 5.215


sd(x)


## [1] 1.883802
```

Die Nullhypothese lautet in diesem Fall $\mu_0 = 5$. Wir wollen wieder testen, ob der Mittelwert 5.215 zum vermuteten Wert μ_0 passt oder nicht.

```
t.test(x, mu = 5)


##
##  One Sample t-test
##
## data:  x
## t = 0.51041, df = 19, p-value = 0.6156
## alternative hypothesis: true mean is not equal to 5
## 95 percent confidence interval:
##  4.333353 6.096647
## sample estimates:
## mean of x
##     5.215
```

Zum **R**-Output:

1. **One Sample t-test**

   Es wird ein Einstichprobentest gemacht.

2. **data: x**

   Datensatz, der verwendet wurde.

3. `t = 0.51041`

   Der t-Wert. Dieser ist für uns uninteressant. Ein grosser t-Wert besagt nur, dass die Nullhypothese verworfen wird. Ist der t-Wert nahe bei 0, so wird die Nullhypothese *nicht* verworfen. Entscheidender ist der P-Wert weiter unten.

4. `df = 19`

   Freiheitsgrad (degree of freedom). Auch dieser Wert in sich uninteressant.

5. `p-value`

   P-Wert. Dies ist *der* entscheidende Wert, denn er entscheidet, ob die Nullhypothese verworfen wird oder nicht. In diesem Fall wird die Nullhypothese auf Signifikanzniveau 5 % nicht verworfen, da der P-Wert grösser als 0.05 ist.

6. `alternative hypothesis: true mean is not equal to 5`

   Hier wird die Alternativhypothese aufgeführt.

7. `95 percent confidence interval: 4.33 6.09`

   Dies ist das Vertrauensintervall, das später eingeführt wird.

8. `mean of x 5.215`

   Hier wird noch der Mittelwert von `x` berechnet.

◀

Beispiel 6.4.2

Wir berechnen nun nochmals das Beispiel 6.2.15. Diesmal schätzen wir allerdings die Standardabweichung σ_X aus den Daten. Die Behauptung ist, dass das wahre $\mu = 80$.

Wir führen also einen t-Test auf dem 5 % Signifikanzniveau durch. Das heisst, wir müssen einen t-Test durchführen.

```
waageA <- c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04, 79.97,
    80.05, 80.03, 80.02, 80, 80.02)

t.test(waageA, mu = 80)
```

```
##
##  One Sample t-test
##
## data:  waageA
## t = 3.1246, df = 12, p-value = 0.008779
## alternative hypothesis: true mean is not equal to 80
## 95 percent confidence interval:
##  80.00629 80.03525
## sample estimates:
## mean of x
##  80.02077
```

Der P-Wert ist hier 0.009. Dieser Wert ist kleiner als das Signifikanzniveau 0.05 und somit wird die Nullhypothese H_0 verworfen. Wir müssen davon ausgehen, dass der wahre Mittelwert *nicht* 80 ist. ◀

Beispiel 6.4.3

Wir wollen nochmals die Aussage des Bundesamtes für Statistik überprüfen, dass die durchschnittliche Körpergrösse der erwachsenen Frauen in der Schweiz 180 cm ist. Unsere Vermutung ist, dass dieser Wert zu gross ist. Wir wollen dies auf einem Signifikanzniveau von 5 % untersuchen.

Wir wählen zufällig 10 Frauen aus und messen deren Körpergrösse (in cm). Die gemessenen Grössen sehen wie folgt aus:

165.7, 156.7, 171.7, 180.3, 163.2, 166.7, 149.9, 170.4, 163.4, 152.5

Da wir vermuten, dass die durchschnittliche Körpergrösse *kleiner* als 180 cm, machen wir einen t-Test nach unten.

```
groesse <- c(165.7, 156.7, 171.7, 180.3, 163.2, 166.7, 149.9, 170.4,
    163.4, 152.5)

t.test(groesse, mu = 180, alternative = "less")


##
##  One Sample t-test
##
## data:  groesse
## t = -5.4836, df = 9, p-value = 0.0001942
## alternative hypothesis: true mean is less than 180
## 95 percent confidence interval:
##     -Inf 169.382
## sample estimates:
```

```
## mean of x
##    164.05
```

Der p-Wert ist 0.0002, also weit unter dem Signifikanzniveau von 0.05. Somit wird die Nullhypothese

$$H_0: \quad \mu_0 = 180$$

verworfen und somit die Alternativhypothese

$$H_A: \quad \mu_0 < 180$$

angenommen. Die Aussage des Bundesamtes für Statistik stimmt also statistisch signifikant (sehr wahrscheinlich) *nicht*. ◀

6.5. Vertrauensintervall für μ

Bei der Punktschätzung für den Mittelwert μ einer Messreihe erhalten wir einen einzigen Zahlwert. Wir wissen allerdings nicht, wie nahe dieser geschätzte Mittelwert beim wahren, aber meist unbekannten, Mittelwert der Verteilung der Messreihe liegt.

Mit dem Vertrauensintervall geben wir hingegen ein Intervall an, wo, grob gesagt, der wahre Mittelwert mit einer bestimmten vorgegebenen Wahrscheinlichkeit liegt.

Am folgenden Beispiel wollen wir die Idee und Interpretation des Vertrauensintervalls kennenlernen.

Beispiel 6.5.1

Bei der Bestimmung des Verwerfungsbereiches beim z-Test, gehen wir vom einem wahren (aber unbekannten) Wert μ aus und einer bekannten Standardabweichung aus. Dann bestimmen wird das Quartil $q_{0.025}$ und $q_{0.975}$ bei einem zweiseitigen Test und Signifikanzniveau $\alpha = 0.05$.

Wir gehen von einer Normalverteilungskurve $X \sim \mathcal{N}(5, 2^2)$ aus. Dann sind die $q_{0.025}$- und $q_{0.975}$-Quantile

```
qnorm(p = c(0.025, 0.975), mean = 5, sd = 2)


## [1] 1.080072 8.919928
```

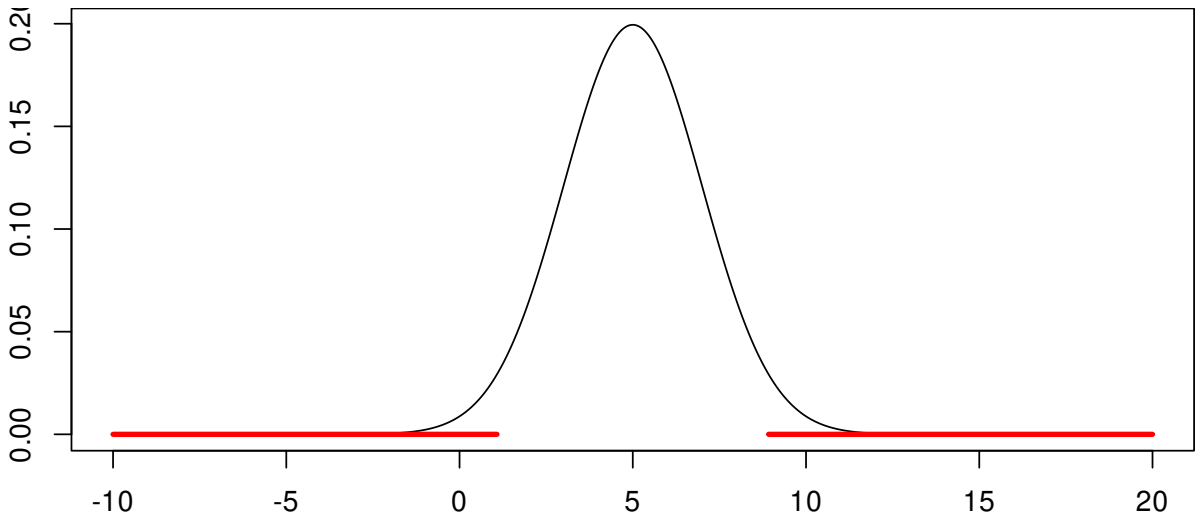


Abbildung 6.13. : Normalverteilungskurve mit Verwerfungsbereich

In Abbildung 6.13 ist die Normalverteilungskurve mit dem Verwerfungsbereich rot eingezeichnet.

Liegt nun $\overline{x}_n$ im Verwerfungsbereich (roter Bereich), dann wird die Nullhypothese H_0 verworfen.

Nun ist das wahre μ_0 praktisch immer unbekannt und für die Bestimmung der Verwerfungsbereiches wurde ein Wert einfach angenommen. Wir können die Frage auch einfach umkehren: Wir kennen $\overline{x}_n$ und fragen uns, für welche μ_0 wird H_0 *nicht* verworfen.

Dies kann man rechnerisch herleiten, wir machen es hier aber graphisch. Wir gehen zur Veranschaulichung wieder von der Annahme $\mu_0 = 5$ aus.

Wir haben nun $\overline{x}_n = 6$ gegeben und zeichnen den Verwerfungsbereich ein. In der Abbildung 6.14 sind:

- Die dicken roten Linien entsprechen dem Verwerfungsbereich für $\overline{x}_n = 6$.
- Die dünnen roten Linien entsprechen dem Verwerfungsbereich für $\mu_0 = 5$.
- Der vertikale schwarze Strich entspricht dem $\overline{x}_n = 6$.
- Der vertikale blaue Strich entspricht dem $\mu_0 = 5$.

Wir stellen fest, dass beide Werte $\overline{x}_n$ und μ_0 nicht innerhalb in einem der beiden Verwerfungsbereiche liegen. Die Idee ist nun, dass wir $\overline{x}_n$ vergrössern und $\mu_0 = 5$ konstant lassen.

Für $\overline{x}_n = 8$ erhalten wir Abbildung 6.15. Auch hier sind $\overline{x}_n$ und μ_0 nicht innerhalb in einem der beiden Verwerfungsbereiche.

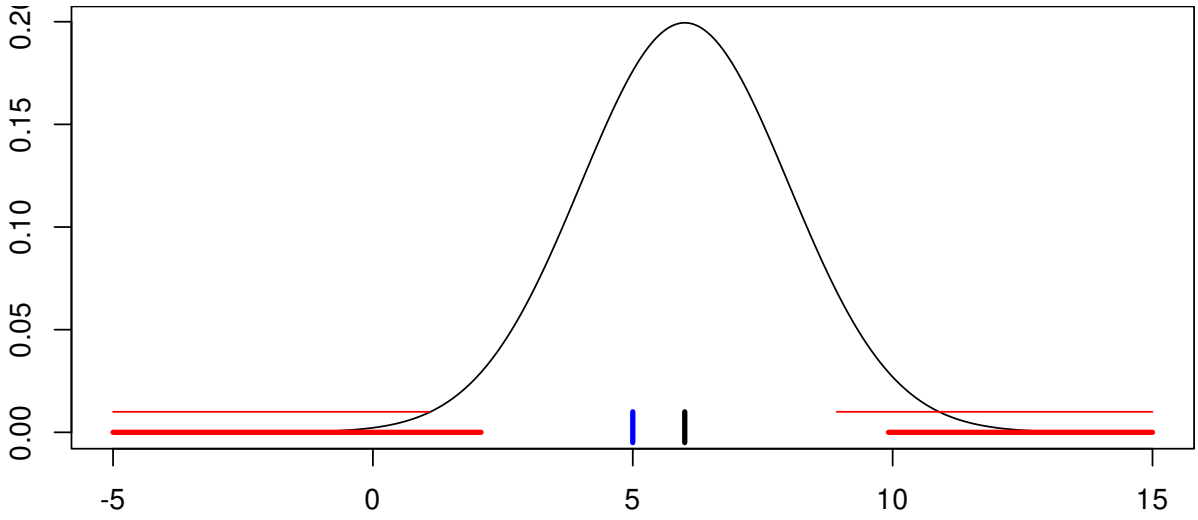


Abbildung 6.14. : Normalverteilungskurve mit zwei Verwerfungsbereichen: $\overline{x}_n = 6$

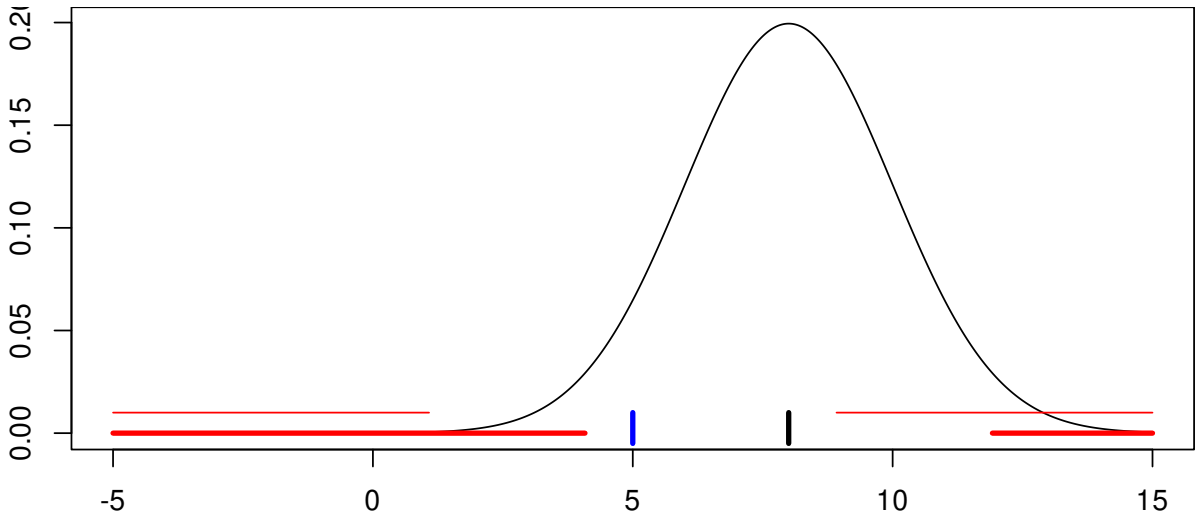


Abbildung 6.15. : Normalverteilungskurve mit zwei Verwerfungsbereichen: $\overline{x}_n = 8$

Für $\overline{x}_n = 9.5$ ändert sich die Situation, da liegt dieser Wert $\overline{x}_n$ (schwarze Linie) im Verwerfungsbereich von $\mu_0 = 5$ (dünne blaue Linien), also wird die Nullhypothese H_0 nun verworfen (siehe Abbildung 6.16). Auf der anderen Seite liegt nun aber $\mu_0 = 5$ im Verwerfungsbereich für $\overline{x}_n = 9.5$.

Wir wollen dies noch anders darstellen. Wir nehmen nicht der Verwerfungsbereich, sondern das was *nicht* zum Verwerfungsbereich gehört. Der ist in Abbildung 6.17 grün eingezeichnet und dieses Intervall heisst *Vertrauensintervall*.

Der Wert 5 liegt nicht im Vertrauensintervall:

```
qnorm(p = c(0.025, 0.975), mean = mean, sd = 2)
```

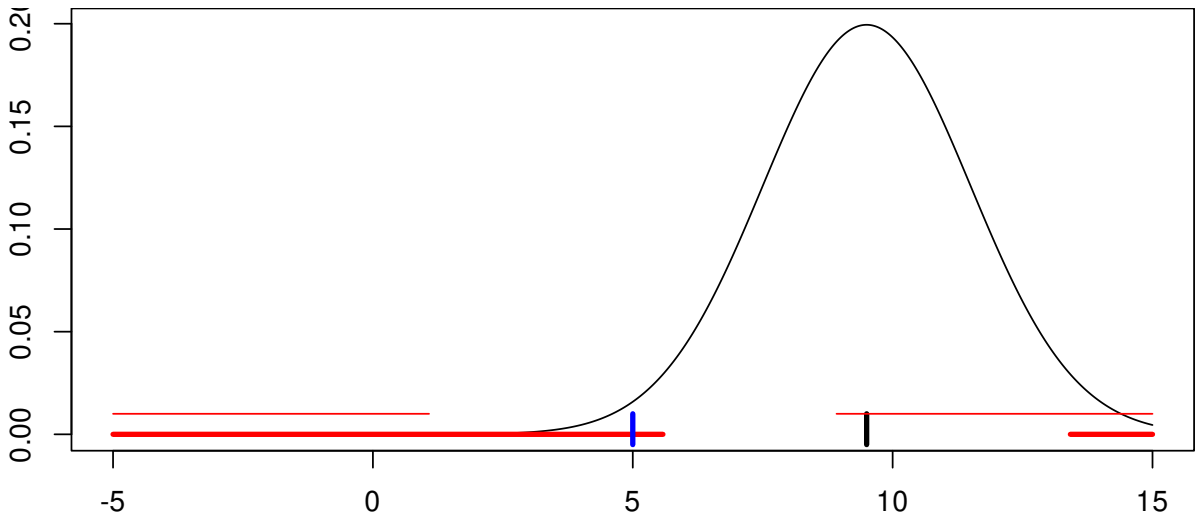


Abbildung 6.16. : Normalverteilungskurve mit zwei Verwerfungsbereichen: $\overline{x}_n = 9.5$

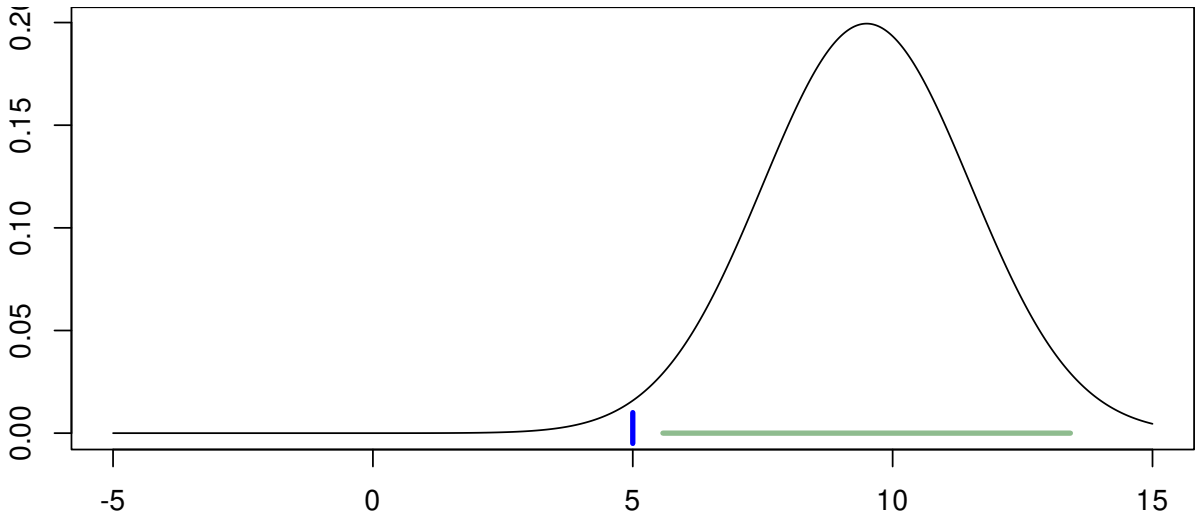


Abbildung 6.17. : Vertrauensintervall

```
## [1]  5.580072 13.419928
```

Wir haben oben gesehen, dass die Nullhypothese dann verworfen wird. Fällt das wahre μ_0 also aus dem Vertrauensintervall, dann wird die Nullhypothese verworfen.

Dies führt uns zu einer weiteren Interpretation des Vertrauensintervalls. Dieses enthält alle μ's für die Nullhypothese *nicht* verworfen wird. Es sagt uns also in welchem Intervall sich das wahre μ_0 befindet. Und wie immer gilt dies nicht absolut, sondern mit einer bestimmten Wahrscheinlichkeit. In diesen Fall liegt das wahre μ_0 zu 95 % im Vertrauensintervall. Wir sprechen deswegen auch von einem 95 %-Vertrauensintervall.

Somit haben wir einen weitere Möglichkeit für einen Testentscheid:

- Liegt μ_0 der Nullhypothese im Vertrauensintervall, so wird die Nullhypothese *nicht* verworfen.
- Liegt μ_0 der Nullhypothese *nicht* im Vertrauensintervall, so wird die Nullhypothese verworfen.

◀

Im Beispiel haben wir das Vertrauensintervall für den z-Test kennengelernt. Das Vertrauensintervall beim t-Test wird analog konstruiert und **R** berechnet dieses und gibt es im Output von `t.test(...)` aus. Die Interpretation ist diesselbe wie beim z-Test.

Der **R**-Output gibt dieses Vertrauensintervall (`confidence interval`) an. Dieses besagt, dass bei einem Signifikanzniveau von 5 % das wahre μ zu 95 % in diesem Intervall liegt.

Mit dem Vertrauensintervall können wir ebenfalls einen Testentscheid durchführen.

Beispiel 6.5.2

Im Beispiel 6.4.2 ist die Nullhypothese

$$H_0: \quad \mu_0 = 80$$

Der **R**-Output gibt ein Vertrauensintervall von

$$[80.00629,\ 80.03525]$$

Zu 95 % liegt das wahre μ in diesem Intervall. Da aber $\mu_0 = 80$ *nicht* in diesem Intervall liegt, so können wir mit 95 % Sicherheit sagen, dass das wahre μ *nicht* 80 ist.

Wir verwerfen also die Nullhypothese und nehmen die Alternativhypothese an. ◀

Beispiel 6.5.3

Im Beispiel 6.4.3 ist die Nullhypothese

$$H_0: \quad \mu_0 = 180$$

Der **R**-Output gibt ein Vertrauensintervall von

$$[-\infty,\ 169.382]$$

Zu 95 % liegt das wahre μ in diesem Intervall. Da aber $\mu_0 = 180$ *nicht* in diesem Intervall liegt, so können wir mit 95 % Sicherheit sagen, dass das wahre μ *nicht* 80 ist.

Wir verwerfen also die Nullhypothese und nehmen die Alternativhypothese an. ◀

Je schmaler das Vertrauensintervall ist, umso sicherer sind wir, wo sich der wahre Mittelwert befindet. Ist das Vertrauensintervall breit, wie

$$[10,\ 1000]$$

so besteht grosse Unsicherheit, wo das wahre μ liegt.

6.6. Statistische Tests bei nicht-normalverteilten Daten

Der t-Test sind optimal, falls die Daten Realisierungen von normalverteilten Zufallsvariablen sind, also

$$X_1, \ldots, X_n \text{ i.i.d. } \sim \mathcal{N}(\mu, \sigma_X^2)$$

Optimalität bedeutet hier, dass der Test die grösste Macht hat. Macht ist die Wahrscheinlichkeit, dass die Nullhypothese richtigerweise verworfen wird.

Wir betrachten in diesem Kapitel die allgemeinere Situation, in der die Daten Realisierungen sind von

$$X_1, \ldots, X_n \text{ i.i.d.}$$

wobei X_i einer *beliebigen* Verteilung folgen kann. Wir bezeichnen mit μ einen Lageparameter der Verteilung (z.B. μ = Median der Verteilung von X_i). Die Nullhypothese ist von der Form

$$H_0: \quad \mu = \mu_0$$

6.6.1. Der Wilcoxon-Test

Der Wilcoxon-Test ist eine Alternative zum t-Test, da er weniger voraussetzt als der t-Test.

Er setzt bloss voraus, dass die Verteilung unter der Nullhypothese *symmetrisch* bezüglich μ_0 ist.

Auch hier wird wieder ein Wert berechnet, den V-Wert, der die sogenannte *Rangsumme* repräsentiert. Wir wollen hier auf die Details nicht eingehen, aber die Grundidee ist diesselbe. Ist der V-Wert zu weit weg vom Median in diesem Fall, wird die Nullhypothese verworfen, ansonsten bei beibehalten.

Beispiel 6.6.1

Mit R erfolgt der Wilcoxon-Test für die Waage A folgendermassen:

```
x <- c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04, 79.97, 80.05,
    80.03, 80.02, 80, 80.02)

wilcox.test(x, mu = 80, alternativ = "two.sided")


##
##  Wilcoxon signed rank test with continuity correction
##
## data:  x
## V = 69, p-value = 0.0195
## alternative hypothesis: true location is not equal to 80
```

◀

Wilcoxon-Test

Der Wilcoxon-Test ist in den allermeisten Fällen dem t-Test vorzuziehen: er hat in vielen Situationen oftmals wesentlich grössere Macht (Wahrscheinlichkeit Nullhypothese richtigerweise zu verwerfen.

Selbst in den ungünstigsten Fällen ist er nie viel schlechter.

6.7. Statistische Tests bei zwei Stichproben

Wir besprechen in diesem Kapitel statistische Methoden, um einen Vergleich zwischen zwei Gruppen, Versuchsbedingungen oder Behandlungen hinsichtlich der Lage der Verteilung anzustellen.

6.7.1. Gepaarte Stichproben

Struktur der Daten

Wann immer möglich sollte man eine Versuchseinheit beiden Versuchsbedingungen unterwerfen. Es liegt eine *gepaarte Stichprobe* vor, wenn

- beide Versuchsbedingungen an derselben Versuchseinheit eingesetzt werden

- oder wenn jeder Versuchseinheit aus der einen Gruppe genau eine Versuchseinheit aus der anderen Gruppe zugeordnet werden kann.

Die Daten sind dann von der folgenden Struktur:

$$\begin{aligned} x_1, \ldots, x_n &\text{ unter Versuchsbedingung 1} \\ y_1, \ldots, y_n &\text{ unter Versuchsbedingung 2} \end{aligned}$$

Notwendigerweise ist dann die Stichprobengrösse n für beide Versuchsbedingungen dieselbe. Zudem sind x_i und y_i abhängig, weil die Werte von der gleichen Versuchseinheit kommen.

Beispiel 6.7.1

Wir testen den Muskelzuwachs durch ein Krafttraining. Dazu messen wir die Kraft von 10 Testpersonen zu Beginn des Trainings. Anschliessend durchlaufen alle Testpersonen ein 6-wöchiges Trainingsprogramm. Dann wird die Kraft erneut gemessen.

Für jede Testperson gibt es also zwei Messungen: Vorher und nachher, die Zuordnung ist eindeutig. Somit handelt es sich um gepaarte Stichproben. ◀

Beispiel 6.7.2

Die Wirksamkeit von Augentropfen zur Reduktion des Augeninnendrucks soll untersucht werden. Wir haben 12 Patienten. Bei jedem Patienten wählen wir zufällig ein Auge aus. In das eine Auge kommen die Augentropfen mit dem Wirkstoff. In das andere Auge kommen Tropfen ohne Wirkstoff (Placebo).

Für jede Testperson haben wir also zwei Messungen: Eine für das rechte und eine für das linke Auge; die Zuordnung ist eindeutig. Somit handelt es sich um gepaarte Stichproben. ◀

Beispiel 6.7.3

Wir haben eine Gruppe von 15 eineiigen Zwillingen, die sich für eine Studie für ein Haarwuchsmittel gemeldet haben. Bei jedem Zwillingspaar wird eine Person zufällig ausgewählt und erhält das Medikament. Die andere Person des Zwillingspaares erhält ein Placebo. Nach drei Wochen misst man den Haarwuchs.

Zu jeder Person aus der Gruppe mit Haarwuchsmittel kann man eindeutig eine Person aus der Gruppe ohne Haarwuchsmittel zuordnen. Somit handelt es sich um gepaarte Stichproben. ◀

Statistischer Test für gepaarte Stichproben

Bei der Analyse von gepaarten Vergleichen arbeitet man mit den Differenzen innerhalb der Paare,

$$d_i = x_i - y_i \quad (i = 1, \ldots, n),$$

welche wir als Realisierungen von i.i.d. Zufallsvariablen

$$D_1, \ldots, D_n$$

auffassen. Kein Unterschied zwischen den beiden Versuchsbedingungen heisst dann einfach

$$E[D_i] = 0$$

(oder auch Median(D_i) = 0, je nach Test). Statistische Tests dafür sind in Unterkapitel 6.2 beschrieben: Falls die Daten normalverteilt sind, eignet sich ein t-Test.

Beispiel 6.7.4

Einen t-Test für gepaarte Stichproben wird im **R** mit der Option **paired =TRUE** durchgeführt:

```
vorher <- c(25, 25, 27, 44, 30, 67, 53, 53, 52, 60, 28)

nachher <- c(27, 29, 37, 56, 46, 82, 57, 80, 61, 59, 43)

t.test(nachher, vorher, altervative = "two.sided", mu = 0, paired = TRUE,
    conf.level = 0.95)


##
##  Paired t-test
##
## data:  nachher and vorher
## t = 4.2716, df = 10, p-value = 0.001633
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   4.91431 15.63114
## sample estimates:
## mean of the differences
##                10.27273
```

Die Nullhypothese ist, dass die Behandlung *keine* Wirkung hat.

Der P-Wert ist 0.001633 und damit kleiner als das Signifikanzniveau 0.05. Die Nullhypothese wird somit verworfen. Das heisst, die Behandlung wirkt statistisch signifikant.

Das 95 %-Vertrauensintervall für diese Differenz ist aus dem **R**-Output oben gegeben durch

$$[4.91,\ 15.63]$$

Die Differenz 0 liegt *nicht* in diesem Intervall und somit wird auch hier die Nullhypothese verworfen.

◀

Bemerkungen:

i. Man hätte auch direkt die Differenzen d_i von den gepaarten Stichproben berechnen und einen t-Test für eine Stichprobe durchführen können.

ii. Der Unterschied der Gruppenmittelwerte hat bei einer zweiseitigen Alternative einen P-Wert von 0.0016 und ist somit auf dem 5 % Signifikanzniveau signifikant.

iii. Der Wert der Teststatistik ist 4.27 und folgt unter der Nullhypothese einer t-Verteilung mit `df = 10` Freiheitsgraden.

iv. Der Unterschied *nachher-vorher* erscheint im Funktionsaufruf, indem sich das erste Argument auf das „nachher" und das zweite Argument auf das „vorher" bezieht. ⧫

6.7.2. Ungepaarte Stichproben

Oft ist es nicht möglich, jeder Behandlungseinheit aus der einen Gruppe eine Behandlungseinheit aus der zweiten Gruppe eindeutig zuzuordnen. In diesem Fall ist eine Paarung nicht möglich und man spricht von einer ungepaarten Stichprobe.

Hier muss die Zuordnung zur Behandlungsgruppe durch das Los erfolgen um systematische Fehler zu vermeiden. (vgl. Abschnitt **??** unten).

Struktur der Daten

Bei ungepaarten Stichproben hat man Daten $x_1, \ldots, x_n$ und $y_1, \ldots, y_m$ (siehe Abschnitt 6.7.2), welche wir als Realisierungen der folgenden Zufallsvariablen auffassen:

$$X_1, \ldots, X_n \text{ i.i.d.}$$
$$Y_1, \ldots, Y_m \text{ i.i.d.}$$

wobei auch alle X_i's von allen Y_j's unabhängig sind. Bei einer solchen zufälligen Zuordnung von Versuchseinheiten zu einer von zwei verschiedenen Versuchsbedingungen spricht man von einer ungepaarten Stichprobe. Im Allgemeinen ist in einer ungepaarten Stichprobe $m \neq n$, aber nicht notwendigerweise. Entscheidend ist, dass x_i und y_i zu verschiedenen Versuchseinheiten gehören und als unabhängig angenommen werden können.

Beispiel 6.7.5

Datensätze der beiden Waagen A und B. Wir haben das Gewicht des Metallblockes mit zwei verschiedenen Waagen hintereinander gemessen. Jede Messung ist entweder mit Waage A oder mit Waage B, aber nicht mit beiden gleichzeitig gemacht worden.

Es gibt also keinen eindeutigen Zusammenhang zwischen den Messungen der Waage A und den Messungen der Waage B. Daher sind die beiden Stichproben ungepaart.

Abgesehen davon sind die Messreihen verschieden lang. ◀

Beispiel 6.7.6

Zufällige Zuordnung von 100 Testpatienten zu einer Gruppe der Grösse 50 mit Medikamentenbehandlung und zu einer anderen Gruppe der Grösse 50 mit Placebo-Behandlung. Es gibt keine eindeutige Zuordnung von einem Patienten aus der Medikamentengruppe zu einem Patienten in der Placebo-Gruppe. Daher handelt es sich um ungepaarte Stichproben, obwohl beide Gruppen gleich gross sind. ◀

Zwei-Stichproben t-Test für ungepaarte Stichproben

Nehmen wir an, wir haben einen ungepaarten Datensatz mit

$$X_1, \ldots, X_n \text{ i.i.d. } \sim \mathcal{N}(\mu_X, \sigma_X^2)$$
$$Y_1, \ldots, Y_m \text{ i.i.d. } \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$$

Beispiel 6.7.7

Als Beispiel betrachten wir den Datensatz mit den Waagen und analysieren mit **R**, ob es einen signifikanten Unterschied zwischen den beiden Waagen A und B gibt.

```
x <- c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04, 79.97, 80.05,
    80.03, 80.02, 80, 80.02)

y <- c(80.02, 79.94, 79.98, 79.97, 80.03, 79.95, 79.97)

t.test(x, y, altervative = "two.sided", mu = 0, paired = FALSE, conf.level = 0.95)


##
##  Welch Two Sample t-test
##
## data:  x and y
## t = 2.8399, df = 9.3725, p-value = 0.01866
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.008490037 0.073048425
## sample estimates:
## mean of x mean of y
##  80.02077  79.98000
```

◀

Bemerkungen:

i. Die ersten beiden Argumente enthalten die Daten der beiden Stichproben. Das Argument **alternative** gibt an, ob die Alternative einseitig (und wenn ja in welche Richtung mit **alternative ="greater"** und **alternative ="less"**) oder zweiseitig (mit **alternative = "two.sided"**) ist. Das Argument **mu** gibt an, welcher Unterschied in den Mittelwerten der beiden Gruppen in der Nullhypothese getestet werden soll.

ii. Wenn man testen will, ob die beiden Gruppenmittelwerte gleich sind, ist **mu=0** die richtige Wahl.

iii. `paired = FALSE` gibt an, dass es sich um zwei ungepaarte Stichproben handelt. Mit `conf.level = 0.95` wird ein 95 %-Vertrauensintervall des Unterschieds zwischen den beiden Gruppenmittelwerten ausgegeben.

iv. In der Zeile $t = \ldots$ steht zunächst der beobachtete Wert der Teststatistik: $t = 2.8399$. Unter der Nullhypothese folgt die Teststatistik einer t-Verteilung mit $df = 19$ Freiheitsgraden. Das ergibt bei einer zweiseitigen Alternative (siehe Zeile `alternative hypothesis: ...`) einen P-Wert von 0.01866. Der Unterschied ist also auf dem 5 % Signifikanzniveau signifikant, weil der P-Wert kleiner als 5 % ist.

v. Der Computer berechnet auch das 95 %-Vertrauensintervall des Unterschieds in den Gruppenmittelwerten: Mit 95 % Wahrscheinlichkeit ist der Gruppenmittelwert von `x` um eine Zahl im Bereich $[0.0085, 0.0730]$ grösser als der Gruppenmittelwert von `y`. Die Null ist nicht enthalten, also ist der Unterschied der Mittelwerte signifikant.

vi. In der letzten Zeile werden schliesslich noch die Mittelwerte der beiden Gruppen angegeben. Beachten Sie, dass kein Verwerfungsbereich ausgegeben wird. ⧫

Zwei-Stichproben Wilcoxon-Test (Mann-Whitney-Test)

Die Voraussetzungen für den Zwei-Stichproben Wilcoxon-Test, manchmal auch Mann-Whitney Test genannt, bezüglich

$$\begin{aligned} X_1, \ldots, X_n &\text{ i.i.d.} \\ Y_1, \ldots, Y_m &\text{ i.i.d.} \end{aligned}$$

sind wie folgt:

$$\begin{aligned} X_1, \ldots, X_n \text{ i.i.d.} &\sim F_X \\ Y_1, \ldots, Y_m \text{ i.i.d.} &\sim F_Y \end{aligned}$$

Die Berechnung des P-Werts eines Zwei-Stichproben Wilcoxon-Tests kann mittels Computer erfolgen. Aus den gleichen Gründen wie im Fall einer Stichprobe (siehe Kapitel 6.6) ist der Wilcoxon-Test im Allgemeinen dem t-Test vorzuziehen.

Beispiel 6.7.8

Wir berechnen noch Beispiel 6.7.7 mit dem Wilcoxontest

```
x <- c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04, 79.97, 80.05,
    80.03, 80.02, 80, 80.02)

y <- c(80.02, 79.94, 79.98, 79.97, 80.03, 79.95, 79.97)

wilcox.test(x, y, altervative = "two.sided", mu = 0)


##
##  Wilcoxon rank sum test with continuity correction
##
## data:  x and y
## W = 76.5, p-value = 0.01454
## alternative hypothesis: true location shift is not equal to 0
```

Auch der Wilcoxon-Test verwirft die Nullhypothese $\mu = 0$, und zwar aufgrund des p-Wertes 0.01454, der etwas kleiner ist als im Falle des t-Tests. ◀

Teil III.

Lineare Regression

Kapitel 7.

Einführung in die Regression

7.1. Was ist Regression?

Wir haben in Kapitel 3 gesehen, wie wir die Regressionsgerade für ein Streudiagramm bestimmen. In diesem und den kommenden Kapitel werden diese Überlegungen verallgemeinern und wesentlich erweitern.

Wir beginnen gleich mit einem einfachen einführenden Beispiel, um zu sehen, worum es sich bei der Regression handelt.

Beispiel 7.1.1

Wir wurden als Statistiker von einer Firma mit dem Auftrag betraut, eine Analyse und Strategie auszuarbeiten, wie der Verkauf eines bestimmten Produktes dieser Firma gesteigert werden kann. Die Firma stellt uns die Daten von Werbebudget und Verkauf zur Verfügung. Der Datensatz **Werbung** besteht aus dem **Verkauf** dieses Produktes in 200 verschiedenen Märkten und den Werbebudgets für dieses Produkt in diesen Märkten für die drei verschiedenen Medien **TV**, **Radio** und **Zeitung**.

```
Werbung <- read.csv("../Daten/Werbung.csv")[, -1]
head(Werbung)
```

```
##      TV Radio Zeitung Verkauf
## 1 230.1  37.8    69.2    22.1
## 2  44.5  39.3    45.1    10.4
## 3  17.2  45.9    69.3     9.3
## 4 151.5  41.3    58.5    18.5
## 5 180.8  10.8    58.4    12.9
## 6   8.7  48.9    75.0     7.2
```

Dabei sind **TV**, **Radio** und **Zeitung** die Werbeausgaben in Tausenden CHF und der **Verkauf** die verkauften Einheiten in Tausenden. In Markt 1 wurden beispielsweise CHF 230 100 in TV-, CHF 37 800 in Radio- und CHF 69 200 in Zeitungswerbung investiert. Dabei wurden 22 100 Einheiten des Produktes verkauft.

Wir können nun für jedes einzelne Werbebudget ein Streudiagrammen mit **Verkauf** erstellen (siehe Abbildung 7.1).

```
par(mfrow = c(1, 3))

plot(Verkauf ~ TV, col = "darkcyan", xlab = "TV", ylab = "Verkauf",
    data = Werbung)
plot(Verkauf ~ Radio, col = "darkcyan", xlab = "Radio", ylab = "Verkauf",
    data = Werbung)
plot(Verkauf ~ Zeitung, col = "darkcyan", xlab = "Zeitung", ylab = "Verkauf",
    data = Werbung)
```

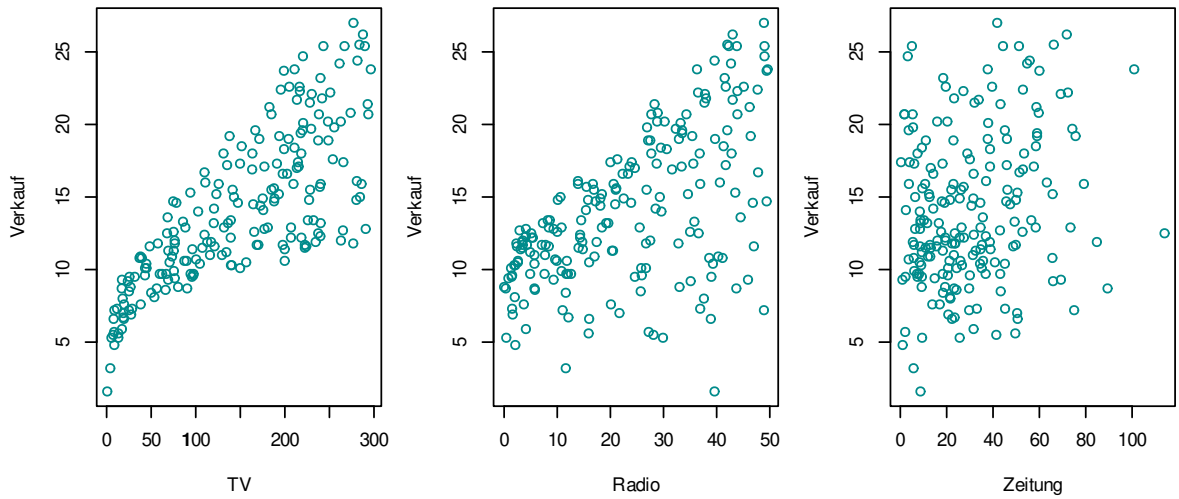


Abbildung 7.1. : **Verkauf** in Abhängigkeit der drei Werbeausgaben **TV**, **Radio** und **Zeitung**

Es ist für die Firma nicht möglich, den Verkauf des Produktes direkt zu erhöhen, aber sie kann die Werbeausgaben in den drei Medien kontrollieren. Können wir einen Zusammenhang zwischen Werbung und Verkauf herstellen, dann können wir der Firma empfehlen, ihre Werbebudgets so anzupassen, dass sie den Verkauf indirekt erhöhen kann. Es ist also unser Ziel, ein möglichst genaues *Modell* zu entwickeln, damit wir auf der Basis der drei Medienbudgets den Verkauf des Produkts *vorhersagen* können. Dieses Modell bildet dann die Basis für die Firma ihr Werbebudget anzupassen um den Verkauf zu erhöhen.

Betrachten wir in Abbildung 7.1 die Graphik links, so sieht man einen deutlichen Zusammenhang zwischen dem Werbebudget von TV und dem Verkauf des Produktes: Je mehr in die Werbung von TV investiert wird, desto grösser die Verkaufszahlen.

Hier kann man sich nun fragen, welche *Form* dieser Zusammenhang hat. Eine Möglichkeit ist, dass die Datenpunkte einer Gerade folgen. Dies werden wir im nächsten Kapitel 8 genauer untersuchen.

In der Graphik rechts in Abbildung 7.1 sieht man überhaupt keinen Zusammenhang. Eine Zunahme der Ausgaben für die Zeitungswerbung hat keinen Einfluss auf den Verkauf und folglich können wir die Werbung in diesem Fall sein lassen. ◀

Aus mathematischen Sichtweise suchen wir eine Funktion f, die aus den drei Werbebudgets X_1 (`TV`), X_2 (`Radio`) und X_3 (`Zeitung`) den Verkauf Y ermittelt:

$$Y \approx f(X_1, X_2, X_3)$$

In dieser Beziehung steht allerdings kein Gleichheitszeichen, da die Streudiagramme keine Graphen einer Funktion darstellen. Deswegen können wir diese Funktion f für den Zusammenhang zwischen X_1, X_2, X_3 und Y nur *approximativ* darstellen.

Bezeichungen: Variablen

In diesem Skript bezeichnen wir:

- Y: *Zielgrösse* oder *Outputvariable*
- X_1, X_2 und X_3: *Prädiktoren*, *Inputvariablen* oder *erklärende Variablen*

Bemerkung:

Die sonst übliche Bezeichnungen von *abhängiger* und *unabhängiger* Variable werden wir hier nicht verwenden, da sie mit stochastischer Abhängigkeit oder Unabhängigkeit verwechselt werden könnten, aber damit nichts zu tun haben. ♦

Im allgemeinen Fall betrachten wir eine quantitative Zielgrösse Y und p verschiedene Prädiktoren $X_1, X_2, \ldots, X_p$. Wir nehmen des weiteren an, dass es *irgendeinen* Zusammenhang zwischen Y und $X_1, X_2, \ldots, X_p$ gibt. Diesen können wir in folgender allgemeiner Form schreiben:

$$Y = f(X_1, X_2, \ldots, X_p) + \varepsilon$$

Hier ist f irgendeine feste, aber *unbekannte* Funktion von $X_1, X_2, \ldots, X_p$. Die Grösse ε ist ein *zufälliger Fehlerterm*, der unabhängig von $X_1, X_2, \ldots, X_p$ ist und Mittelwert 0 hat. Was dieser Fehlerterm ε bedeutet, sehen wir im folgenden Beispiel.

Beispiel 7.1.2

In Abbildung 7.2 betrachten wir zunächst die Graphik auf der linken Seite. Die Zielgrösse **Einkommen** von 30 Individuen ist in Abhängigkeit des Prädiktors **Ausbildungsdauer** (in Jahren) aufgezeichnet. Die Graphik deutet an, dass wir das **Einkommen** aus der **Ausbildungsdauer** berechnen können. Zumindest ist ein Zusammenhang zwischen **Einkommen** und **Ausbildungsdauer** erkennbar: Je länger die Ausbildungsdauer umso grösser ist das jährliche Einkommen.

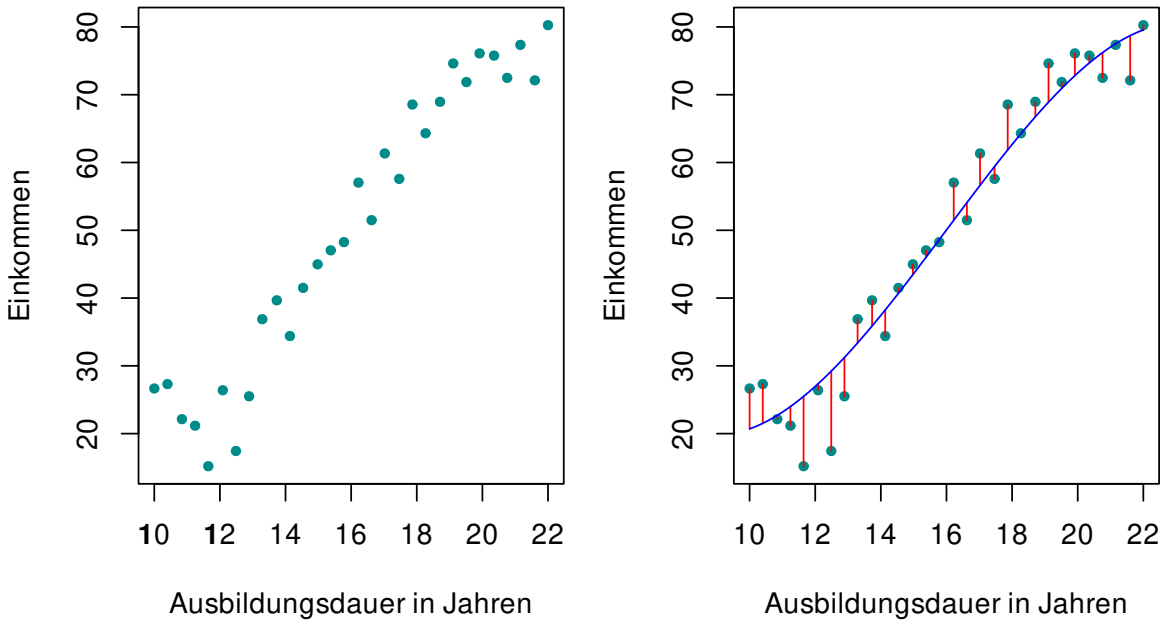


Abbildung 7.2. : Einkommen in Abhängigkeit von der Anzahl Jahre Ausbildung.

Allerdings ist die zugehörige Funktion f, die die Prädiktoren und die Zielgrösse miteinander in Verbindung bringt, in der Regel unbekannt. In dieser Situation müssen wir f aus den Daten annähern oder *schätzen*[1].

In diesem Fall sind die Daten simuliert und die Funktion ist f bekannt. Der Graph von f ist die blaue Kurve in Abbildung 7.2 rechts.

Einige Beobachtungen liegen überhalb, andere unterhalb der blauen Kurve. Die roten vertikalen Linien repräsentieren den Fehlerterm ε, auch Residuen genannt. Insgesamt haben die Fehler einen empirischen Mittelwert, der annähernd 0 ist.

[1] Der Begriff „schätzen" hat in der Stochastik eine andere Bedeutung als in der Umgangssprache. Schätzungen sind in der Stochastik immer *Berechnungen*, die eine gesuchte Grösse annähern. Umgangssprachlich ist es eher eine *gefühlsmässige* Annäherung

Jeder Punkt hat seinen eigenen Wert von ε. Zum Beispiel, für den ersten Datenpunkt (x_1, y_1), falls if

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

dann

$$f(10) \approx 21$$

und

$$\varepsilon_1 \approx 6$$

also

$$y_1 = f(10) + \varepsilon_1 \approx 21 + 6 = 27$$

und somit

$$(x_1, y_1) \approx (10, 27)$$

◀

Im Wesentlichen ist das Ziel der Regression, die Funktion f zu *schätzen*. Wir werden in diesem Kapitel kurz das Vorgehen dieser Schätzung beschreiben.

Bezeichung:

Es wird allgemein eine geschätzte Grösse mit einem Hut ^gekennzeichnet. So ist

- $\widehat{Y}$ die Schätzung der unbekannten Grösse Y oder
- $\widehat{f}$ die Schätzung der unbekannten Funktion f

7.2. Warum soll f geschätzt werden?

Es gibt im Wesentlichen zwei Hauptgründe, warum wir die unbekannte Funktion f schätzen wollen. Der eine liegt darin, Datenpunkte *vorherzusagen* (*Prognose*), und der andere besteht darin, *Rückschlüsse* auf die Funktion selbst zu ziehen.

7.2.1. Prognose

Oft sind die Prädiktoren $X_1, X_2, \ldots, X_p$ einfach verfügbar, aber die Zielgrösse nicht. In so einem Fall können wir Y durch

$$\widehat{Y} = \widehat{f}(X_1, X_2, \ldots, X_p)$$

schätzen, da der Fehlerterm im Mittel 0 ist. Dabei ist $\widehat{f}$ die Schätzung von f und $\widehat{Y}$ repräsentiert die zugehörige Zielgrösse. Hier wird $\widehat{f}$ oft als *black box* behandelt, da uns die exakte Form von $\widehat{f}$ an dieser Stelle nicht speziell interessiert, so lange $\widehat{f}$ möglichst genaue Vorhersagen produziert.

Beispiel 7.2.1

Die Prädiktoren $X_1, X_2, \ldots, X_p$ seien die Werte von verschiedenen Charakteristiken einer Blutentnahme, die der Hausarzt des Patienten in seinem Labor bestimmen kann. Die Zielgrösse Y sei ein Mass für das Risiko, dass der Patient starke Nebenwirkungen bei der Anwendung eines bestimmten Medikamentes erleidet.

Der Arzt möchte bei der Verschreibung eines Medikamentes natürlich Y aufgrund von $X_1, X_2, \ldots, X_p$ vorhersagen können, damit er nicht ein Medikament Patienten verschreibt, die ein hohes Risiko für Nebenwirkungen bei diesem Medikament haben - d.h. bei denen Y gross ist.

Der Arzt ist dabei nicht daran interessiert, wie die Funktion f aussieht, sondern ausschliesslich, dass sie gute Vorhersagen liefert. ◀

Die Genauigkeit von $\widehat{Y}$ als Vorhersage von Y hängt von zwei Grössen ab, die als *reduziblen* und *irreduziblen Fehler* bezeichnen. Im Allgemeinen ist $\widehat{f}$ keine perfekte Schätzung von f und diese Ungenauigkeit führt zu einem Fehler. Hier handelt es sich um einen *reduziblen Fehler*, da wir die Schätzung mit statistischen Methoden verbessern können.

Beispiel 7.2.2

In Abbildung 7.2 rechts von Beispiel 7.1.2 haben wir gesehen, dass der Graph der wahren Funktion f keine lineare Funktion ist.

Allerdings ist das einzige was wir jeweils sehen die Abbildung links. Die Punktwolke ist annähernd linear und so sind wir versucht eine Regressionsgerade zu bestimmen. Damit machen wir allerdings einen Fehler, einen reduziblen Fehler, da die wahre Funktion nicht-linear ist. ◀

Aber auch wenn wir eine perfekte Schätzung für f hätten, so dass die Outputvariable die Form

$$\widehat{Y} = f(X_1, X_2, \ldots, X_p)$$

annimmt, wird unsere Vorhersage Y noch Fehler enthalten. Dies liegt am Fehlerterm ε, der per Definition nicht von $X_1, X_2, \ldots, X_p$ abhängt. Diese Variablilität im Zusammenhang mit ε beeinflusst die Genauigkeit der Vorhersage ebenfalls. Dies ist als *irreduzibler Fehler* bekannt, weil dieser Fehler nicht beeinflusst werden kann, wie gut auch die Schätzung von f ist.

Beispiel 7.2.3

Wir kommen nochmals auf Beispiel 7.1.2 zurück.

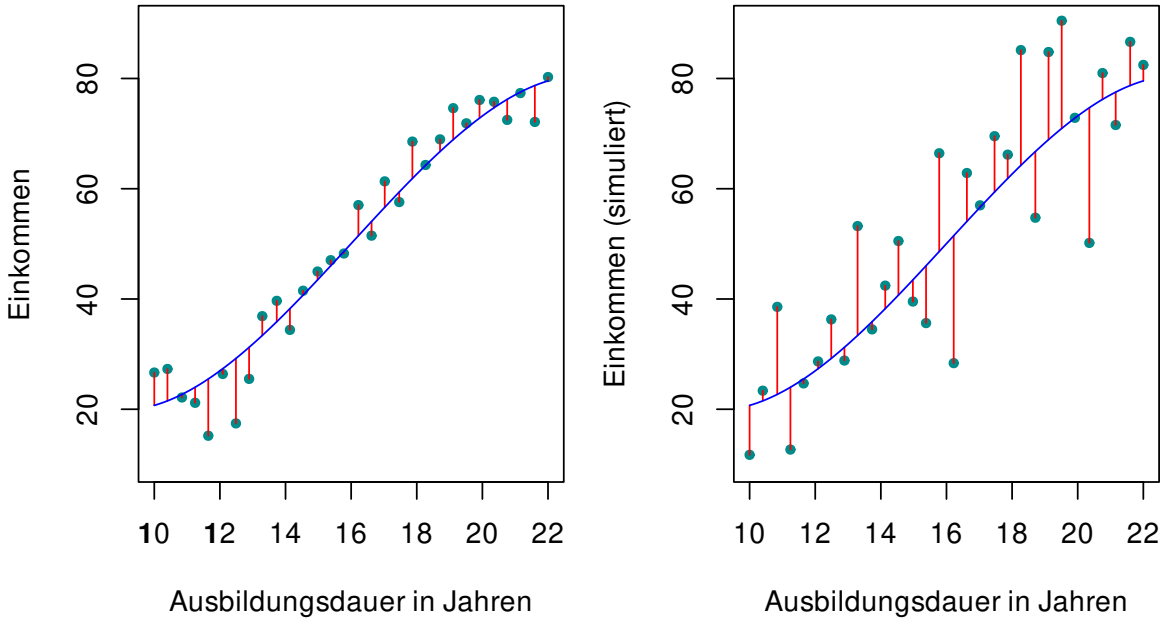


Abbildung 7.3. : Einkommen: Irreduzibler Fehler

In Abbildung 7.3 links sehen wir die ursprüngliche Abbildung von Beispiel 7.1.2. Diese Daten folgen der blauen Kurve, die bekannt ist. Die roten Linien sind der Fehler zu der Kurve. Diesen Fehler nennen wir irreduziblen Fehler.

Dieser Fehler auf der linken Seite der Abbildung 7.3 ist im Durchschnitt „klein". Das heisst, eine Vorhersage mit dem Modell ist relativ genau.

Anders sieht es bei der Abbildung rechts aus. Hier weichen die Punkte doch beträchtlich von der Kurve ab. Eine Prognose mittels der blauen Kurve lässt sich zwar schon machen, aber deren Genauigkeit ist beschränkt, da die Punkte weit von der Kurve abweichen, obwohl die Punkte der Kurve folgen.

Also ist die Prognose genauer, je kleiner der irreduzible Fehler ist. ◀

Woher kommt nun dieser Fehler ε, der ungleich 0 ist? Diese Grösse kann Variablen enthalten, die nicht gemessen wurden, die aber für die Vorhersage von Y wichtig sind. Da diese Variablen nicht gemessen wurden, können wir sie für die Vorhersage auch nicht verwenden. Die Grösse ε kann aber auch nicht messbare Grössen enthalten.

Beispiel 7.2.4

In Beispiel 7.2.1 kann die Stärke der Nebenwirkungen eines Medikamentes abhängig sein von der Tageszeit der Einnahme des Medikamentes oder auch einfach vom allgemeinen Wohlbefinden des Patienten.

Diese Einflüsse auf die Nebenwirkungen sind sehr schwer zu bestimmen und gehören daher zum irreduziblen Fehler. ◀

7.2.2. Rückschlüsse auf f

Im vorhergehenden Abschnitt waren wir nur an der Voraussagekraft von f interessiert, aber nicht an der Form von f. Hier möchten wir verstehen, welchen Einfluss $X_1, X_2, \ldots, X_p$ auf Y haben. Die Schätzung $\widehat{f}$ ist hier keine black box, da wir die genaue Form von $\widehat{f}$ wissen wollen. In diesem Zusammenhang stellen sich folgende Fragen:

- *Welche Inputvariablen werden mit dem Output assoziert?*

  Natürlich alle, denkt man zuerst. Aber oft sind es nur einige wenige Variablen, die auf Y einen substantiellen Einfluss haben. Haben wir sehr viele Inputvariablen, so kann es sehr wichtig sein, die *wichtigen* Inputvariablen zu identifizieren.

  Im Beispiel der **Werbung** in Abbildung 7.1 auf Seite 235 haben die Ausgaben bei der TV-Werbung einen grossen Einfluss auf die Verkaufszahlen, die Zeitungswerbung aber nicht. Also können wir uns auf die TV-Werbung konzentrieren.

  Ziel ist es auch, das Modell so einfach wie möglich zu halten. Wenn das Modell einfach ist, so lässt sich dieses einfach interpretieren. Wir werden auf diesen Punkt noch einige Male zurückkommen.

- *Wie sieht der Zusammenhang zwischen Outputvariable und jeder Inputvariable aus?*

  Einige Inputvariablen haben einen positiven Zusammenhang mit der Outputvariable. Eine Vergrösserung der Inputvariable hat in diesem Fall eine Vergrösserung von Y zur Folge. Andere Inputvariablen haben einen negativen Zusammenhang mit Y. In Abhängigkeit von der Komplexität von f kann der Zusammenhang zwischen der Zielvariablen und einer erklärenden auch von den Werten der anderen erklärenden Variablen abhängen.

- *Kann der Zusammenhang zwischen der Outputvariable und jeder Inputvariable durch eine lineare Gleichung angemessen beschrieben werden oder ist der Zusammenhang komplizierter?*

  Historisch sind die meisten Schätzungen von f linear. Dies hat damit zu tun, dass solche Schätzungen sehr einfach sind, wie wir in Kapitel 8 sehen werden.

  In vielen Situationen ist die Annahme der Linearität ausreichend oder gar wünschenswert. Aber oft ist der wahre Zusammenhang komplizierter und das lineare Modell liefert keinen angemessenen Zusammenhang zwischen Input- und Outputvariablen.

Beispiel 7.2.5

Für **Werbung**-Beispiel 7.1 können wir uns folgende Fragen stellen:

- Welche Medien tragen zum Verkauf des Produktes bei?
- Welche Medien haben den grössten Einfluss auf den Verkauf?
- Welchen Zuwachs im Verkauf hat eine bestimmte Vergrösserung der TV-Werbung zur Folge?

◀

7.3. Wie schätzen wir f?

Es gebt mehrere Verfahren, wie wir f schätzen können. Wir wollen uns aber auf die sogennante *parametrische Methode* beschränken.

Wir gehen wie folgt vor:

1. Wir machen eine *Annahme* über die funktionale Form von f.

   Die einfachste Annahme ist beispielsweise , dass f linear in $X_1, X_2, \ldots, X_p$ ist:

   $$f(X_1, X_2, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

   Dieses lineare Modell werden wir in Kapitel 8 ausführlich behandeln. Haben wir einmal angenommen, dass f linear ist, so vereinfacht sich die Schätzung von f erheblich. Es gilt dann nur noch die $p + 1$ Parameter $\beta_0, \beta_1, \ldots \beta_p$ zu schätzen.

2. Nachdem wir das Modell *gewählt* haben, brauchen wir noch ein Verfahren, das die Daten in das Modell *passt*.

   Im Falle des linearen Modelles müssen wir die Parameter $\beta_0, \beta_1, \ldots \beta_p$ schätzen. Das heisst, wir müssen die Parameter so bestimmen, dass

   $$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

   Die häufigste Methode zur Bestimmung von $\beta_0, \beta_1, \ldots \beta_p$ ist die *Methode der kleinsten Quadrate*. Diese haben wir in Kapitel 3 schon kennengelernt und werden sie in Kapitel 8 vertiefter behandeln.

Beispiel 7.3.1

Im Beispiel **Werbung** sieht das lineare Modell wie folgt aus:

$$\texttt{Verkauf} \approx \beta_0 + \beta_1 \cdot \texttt{TV} + \beta_2 \cdot \texttt{Radio} + \beta_3 \cdot \texttt{Zeitung}$$

◀

Beispiel 7.3.2

Im Beispiel **Einkommen** sieht das lineare Modell wie folgt aus:

$$\texttt{Einkommen} \approx \beta_0 + \beta_1 \cdot \texttt{Ausbildung}$$

◀

Beispiel 7.3.3

Kommen wir nochmals auf den Datensatz **Einkommen** zurück, der nochmals in Abbildung 7.4 abgebildet ist.

Es stellt sich hier die Frage, welches *Modell* wir wählen, oder präziser, welche Form f haben soll.

Betrachten wir die Daten, so kommt ein lineares Modell in Frage, da die Punkte mehr oder weniger einer Geraden folgen:

$$f(X) = \beta_0 + \beta_1 X$$

Die Gerade ist in Abbildung 7.5 links dargestellt.

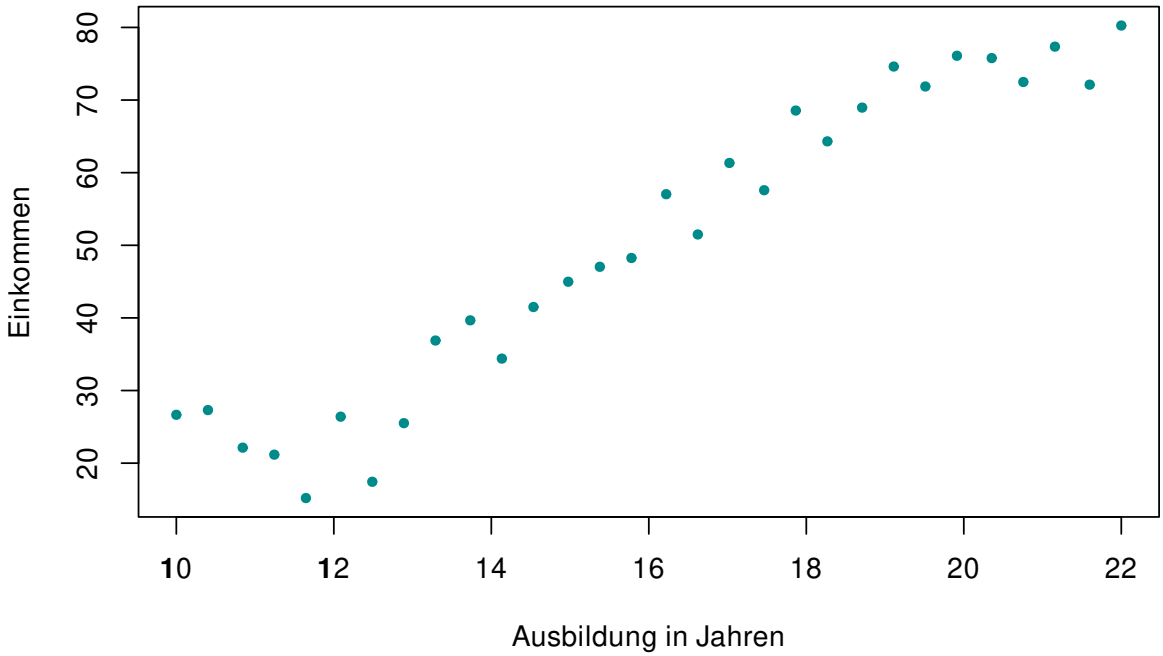


Abbildung 7.4. : Einkommen in Abhängigkeit der Anzahl Jahre Ausbildung

Allerdings könnte man sich auch ein kubisches Modell (Polynom 3. Grades) vorstellen:

$$f(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$$

Der Graph von f ist in Abbildung 7.5 rechts eingezeichnet.

Wir könnten uns noch viele weitere Modelle vorstellen. Aber welches ist nun das „richtige"? Dies lässt sich in dieser Absolutheit nicht entscheiden. Die Funktion f ist im Allgemeinen unbekannt, und so liegt es an uns, das „beste" Modell zu wählen. Die Statistik wird uns in dieser Entscheidungsfindung behilflich sein.

Welches Modell ist in unserem Beispiel das „bessere"? Das kubische Modell scheint besser zu passen, ist aber auch komplizierter.

Das einfachere lineare Modell, das etwas weniger genau passt, hat allerdings einen Vorteil: Die Parameter β_0 und β_1 lassen sich geometrisch interpretieren:

- β_0 ist der y-Achsenabschnitt
- β_1 die Steigung der Geraden

◀

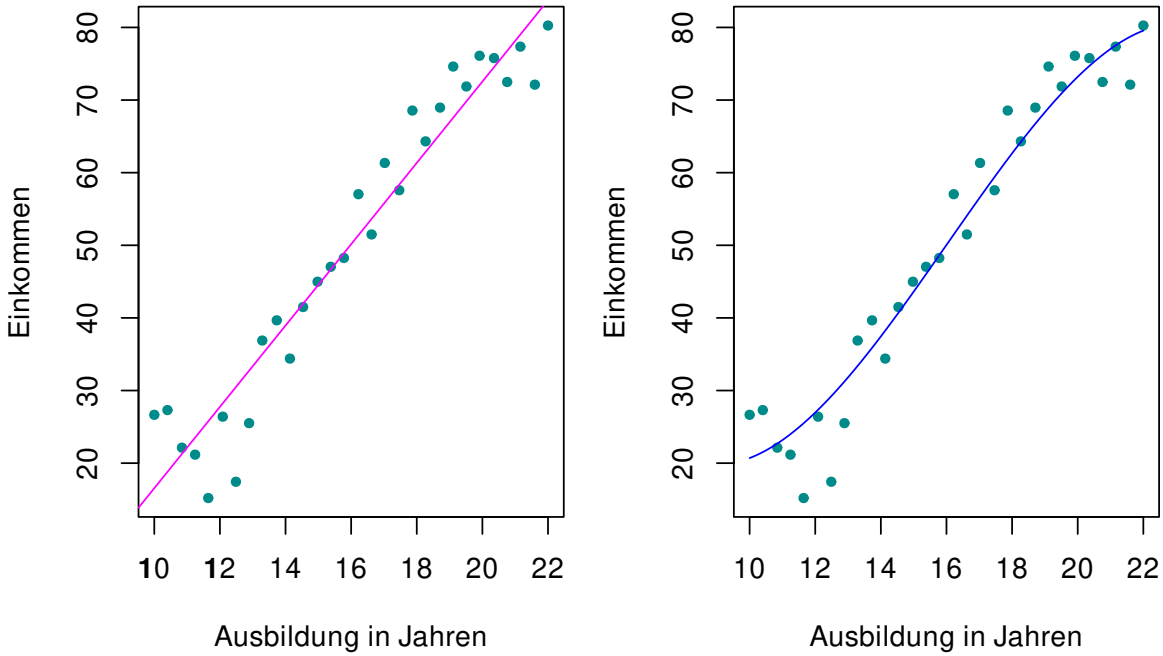


Abbildung 7.5. : Zwei Schätzung von f: linear und kubisch

Bemerkungen:

i. Das kompliziertere Modell muss *nicht* das bessere Modell sein. Sie können zu einem Phänomen führen, das *Overfitting* heisst. Dabei werden Fehler oder Ausreisser zu stark berücksichtigt.

   Wir illustrieren dies in Abbildung 7.6. In der obersten Reihe haben wir zwei Datensätze, die derselben wahren Funktion (Polynom 5. Grades) folgen. Die Unterschiede sind zufällig.

   In der zweiten Reihe wird eine Anpassung der Punkte mit Hilfe eines Polynoms 5. Grades gemacht. Für die beiden Datensätze sind die Anpassung unterschiedlich. Die Methode der kleinsten Quadrate versucht die Punkte so gut wie möglich anzupassen. Dies ist mit einer nichtlinearen Funktion viel besser zu machen. Allerdings passen die Punkte nur für den *einen* Datensatz gut, aber nicht für einen anderen der ähnlich ist. Darum die unterschiedlichen Kurven.

   Wir sprechen in diesem Fall von *Overfitting*: Die Kurve, die für einen Datensatz gut passt, muss für einen ähnlichen nicht mehr gut passen.

   In der dritten Reihe haben wir eine lineare Regression, die sich kaum unterscheiden. Lineare Regression ist viel stabiler gegenüber verschiedenen Datensätzen, die derselben Funktion folgen.

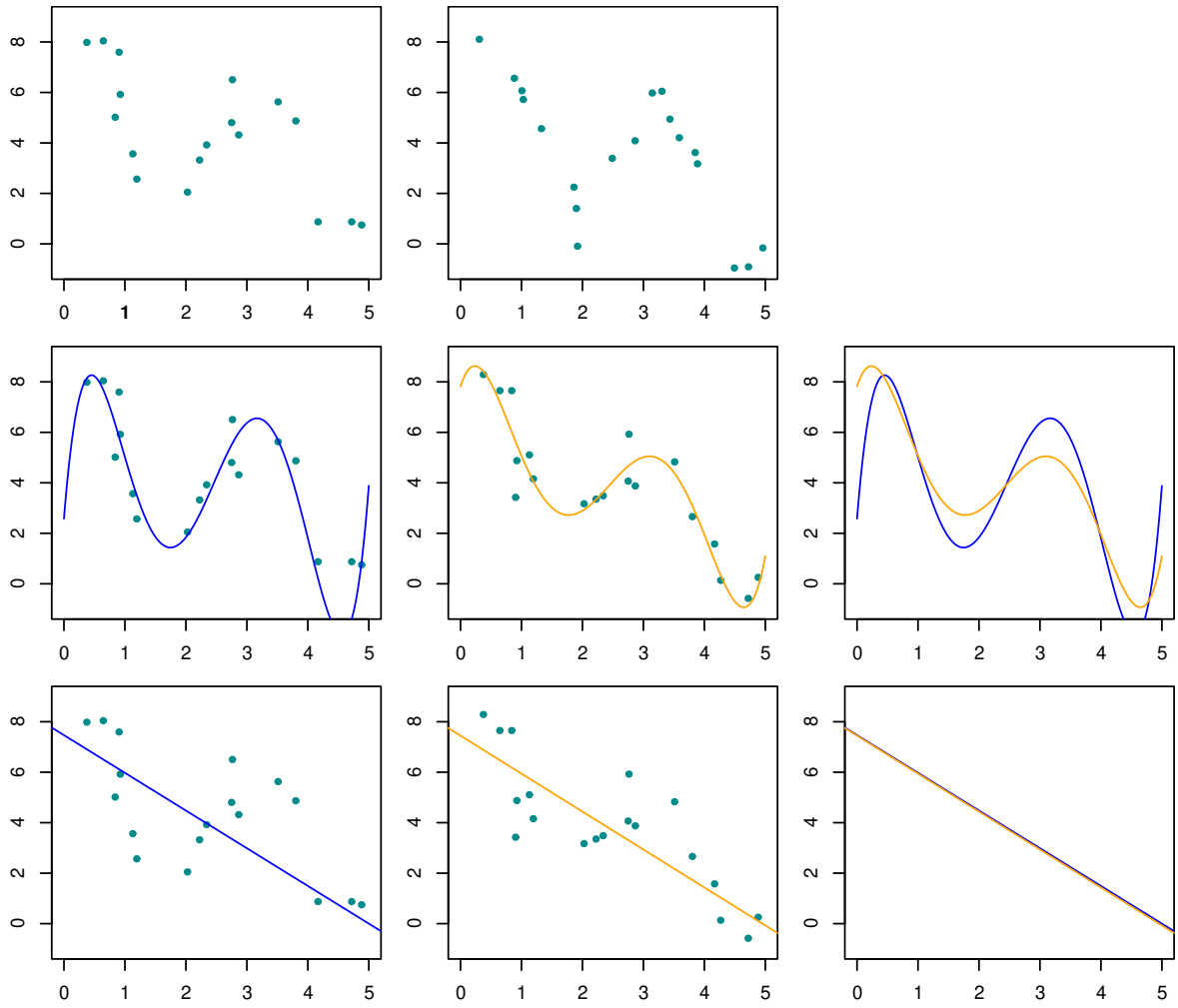


Abbildung 7.6. : Overfitting

ii. In sehr vielen Fällen ist das lineare Modell ausreichend. Das folgende Kapitel handelt davon. ♦

Kapitel 8.

Lineare Regression

8.1. Einleitung

Wir erinnern uns an den Datensatz **Werbung**. Abbildung 7.1 zeigt den **Verkauf** für ein bestimmtes Produkt (in Einheiten von tausend verkauften Produkten) als Funktion von Werbebudgets (in Einheiten von tausend CHF) für **TV**, **Radio** und **Zeitung**. Aufgrund dieser Daten werden wir als Statistiker angefragt, einen Marketingplan zu erstellen, der für das nächste Jahr zu höheren Verkäufen führen soll. Welche Informationen sind nützlich, um solche Empfehlungen auszuarbeiten? Hier sind einige wichtige Fragen, die wir angehen wollen:

1. *Gibt es einen Zusammenhang zwischen dem Werbebudget und dem Verkauf?*

   Unser erstes Ziel ist es zu entscheiden, ob die Daten genügend Hinweise für einen Zusammenhang zwischen Werbebudget und Verkauf liefern. Ist der Hinweis schwach, dann können wir argumentieren, dass auf die Werbung gänzlich verzichtet werden kann.

   Oder anders gefragt: Hat überhaupt eine Variable Einfluss auf den Verkauf?

2. *Wie stark ist der Zusammenhang zwischen Werbebudget und dem Verkauf?*

   Angenommen es gibt einen solchen Zusammenhang zwischen dem Werbebudget und dem Verkauf. Dann möchten wir wissen, wie *stark* dieser Zusammenhang ist. Oder anders ausgedrückt: Können wir für ein gegebenes Werbebudget den Verkauf mit hoher Genauigkeit vorhersagen? In diesem Fall gäbe es einen starken Zusammenhang.

   Oder ist die berechnete Vorhersage nur wenig besser als eine zufällige Vorhersage? In diesem Fall würde ein schwacher Zusammenhang vorliegen.

3. *Welche Medien tragen zum Verkauf bei?*

   Tragen alle drei Medien (TV, Radio, Zeitung) zum Verkauf bei oder ist es nur eines oder zwei?

   Um diese Frage zu beantworten, müssen wir einen Weg finden, den Einfluss jedes einzelnen Mediums auf den Verkauf separat zu ermitteln, auch wenn wir für alle drei Medien Geld ausgegeben haben.

4. *Wie genau können wir den Einfluss jedes einzelnen Mediums auf den Verkauf schätzen?*

   Wie gross ist die Zunahme des Verkaufs für jeden zusätzlichen Franken, den wir für ein spezifisches Medium ausgeben? Wie genau können wir diese Zunahme vorhersagen?

5. *Wie genau können wir zukünftige Verkäufe vorhersagen?*

   Welche Verkäufe können wir für beliebige Werbebudgets für TV, Radio und Zeitung vorhersagen und wie genau ist diese Vorhersage?

6. *Ist der Zusammenhang linear?*

   Ist der Zusammenhang zwischen Werbebudgets für die unterschiedlichen Medien und Verkauf annähernd linear, dann ist Lineare Regression ein angebrachtes Modell. Falls nicht, so kann mit Hilfe von Variablentransformation lineare Regression unter Umständen trotzdem verwendet werden.

7. *Gibt es Synergie zwischen den verschiedenen Medien?*

   Möglicherweise bewirkt CHF 50 000 für TV-Werbung und CHF 50 000 für Radiowerbung mehr Verkäufe als wenn wir CHF 100 000 für das eine oder andere Medium aufgewendet hätten. Im Marketing spricht man in diesem Fall von einem *Synergieeffekt*, in der Statistik von *Interaktionseffekt*.

Es zeigt sich, dass mit linearer Regression alle diese Fragen beantwortet werden können. Wir werden diese erst einmal in einem allgemeinen Kontext beantworten, bevor die Fragen oben konkret beantwortet werden.

8.2. Das einfache Regressionsmodell

8.2.1. Das Modell

Die *einfache lineare Regression* wird ihrem Namen gerecht: es ist ein sehr einfaches Verfahren, um einen quantitativen Output Y auf der Basis einer einzigen Inputvariable

X vorherzusagen. Es wird eine annähernd lineare Beziehung zwischen X und Y angenommen. Mathematisch sieht diese lineare Beziehung wie folgt aus:

$$Y \approx \beta_0 + \beta_1 X$$

Dabei steht „$\approx$" für „ist annähernd modelliert durch".

Beispiel 8.2.1

Im Beispiel **Werbung** stellt X die Grösse **TV** und Y die Grösse **Verkauf** dar. Nach dem linearen Regressionsmodell gilt dann

$$\texttt{Verkauf} \approx \beta_0 + \beta_1 \cdot \texttt{TV}$$

◀

Die Grössen β_0 und β_1 sind unbekannte Konstanten, die den y-Achsenabschnitt und die Steigung des linearen Modells darstellen. Zusammen werden β_0 und β_1 die *Koeffizienten* oder *Parameter* des Modells genannt.

Die Koeffizienten werden aus den gegebenen Daten geschätzt, und wir erhalten die Schätzungen $\widehat{\beta}_0$ und $\widehat{\beta}_1$ für die Modellkoeffizienten (wie das geschieht, sehen wir im nächsten Abschnitt). Sind diese Koeffizienten bekannt, so können wir zukünftige Verkäufe auf der Basis eines bestimmten Werbebudgets für TV vorhersagen. Wir berechnen dies mittels

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

wobei $\widehat{y}$ die Vorhersage von Y auf Basis des Inputs $X = x$ bezeichnet.

Beispiel 8.2.2

Wir werden in Beispiel **??** sehen, dass wir für die TV-Werbung folgende geschätzte Koeffizienten

$$\widehat{\beta}_0 = 7.03 \qquad \text{und} \qquad \widehat{\beta}_1 = 0.047$$

erhalten.

Dann können wir den geschätzten Verkauf $\widehat{y}$ für die TV-Werbeausgaben $x = 120\,000$ berechnen:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x = 7.03 + 0.047 \cdot 120 = 12.67$$

Somit werden für TV-Werbeausgaben von $x = 120\,000$ geschätzt 12 670 Einheiten verkauft. ◀

8.2.2. Schätzung der Parameter

In der Praxis sind β_0 und β_1 unbekannt. Bevor wir das lineare Modell benützen können, müssen wir diese Koeffizienten schätzen. Dabei gehen wir von n Beobachtungspaaren

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

aus. Jedes Paar besteht aus einer Messung von X und einer Messung von Y.

Beispiel 8.2.3

Im Beispiel **Werbung** haben wir $n = 200$ verschiedene Beobachtungspaare (Märkte). Die x-Koordinate besteht aus dem TV-Budget und die zugehörige y-Koordinaten aus den entsprechenden Produktverkäufen (siehe Abbildung 7.1 auf Seite 235 links). ◀

Unser Ziel ist es, $\widehat{\beta}_0$ und $\widehat{\beta}_1$ so zu bestimmen, dass die Gerade

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

möglichst gut zu den Daten passt. Das heisst, dass

$$y_i \approx \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

für alle $i = 1, \ldots, n$. Auf der linken Seite der obigen approximativen Beziehung steht der Messwert, auf der rechten der zugehörige y-Wert auf der Geraden.

Die Frage ist nun, was heisst „möglichst gut"?

Beispiel 8.2.4

In Abbildung 8.1 sind einige Geraden eingezeichnet, die gut zu den Datenpunkten passen. Welche passt „am besten"?

◀

Die Punkte sollten möglichst nahe bei der gesuchten Geraden liegen. Wir müssen also im Beispiel oben $\widehat{\beta}_0$ und $\widehat{\beta}_1$ so bestimmen, dass die resultierende Gerade so nahe wie möglich an den $n = 200$ Datenpunkten entlangläuft.

Was heisst aber so „nahe wie möglich"? Es gibt mehrere Methoden, um *Nähe* zu messen. Die bei weitem gebräuchlichste Methode ist die *Methode der kleinsten Quadrate*, die wir in Kapitel 3 kennengelernt haben.

Dazu sei

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

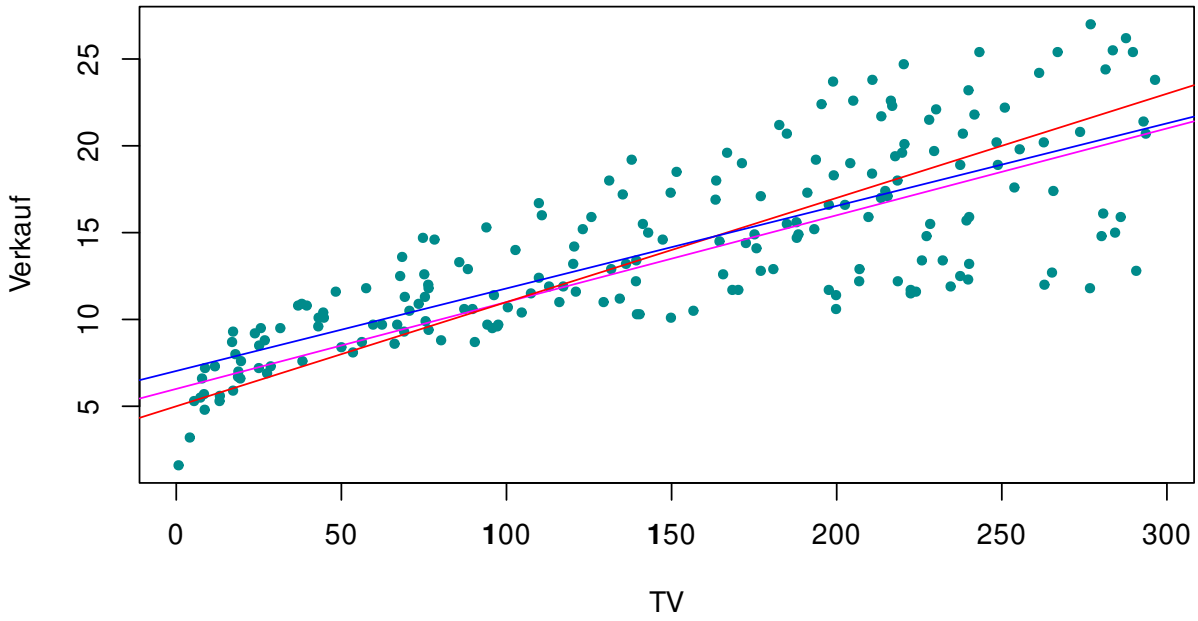


Abbildung 8.1. : Verschiedene Geraden, die zu den Datenpunkten passen könnten. Welche Gerade passt „am besten"?

der vorhergesagte Wert für Y basierend auf dem i-ten Wert von X, also x_i. Dann ist

$$r_i = y_i - \widehat{y}_i$$

das i-te *Residuum*. Dies ist die Differenz zwischen dem i-ten *beobachteten* Wert der Zielgrösse und dem i-ten von unserem linearen Modell *vorhergesagten* Wert der Zielgrösse.

Beispiel 8.2.5

In Abbildung 8.2 sind die Residuen als Strecken rot eingezeichnet. Die Residuen oberhalb der Geraden sind positiv, diejenigen unterhalb der Geraden negativ.

◀

Die *Summe* der Residuen ist kein guter Wert für die Nähe der Punkte zur Geraden, da sich diese aufheben. Anders sieht es mit der Summe der *Quadrate* der Residuen (RSS, residual sum of squares) aus. Es gilt dann

$$\text{RSS} = r_1^2 + r_2^2 + \ldots + r_n^2$$

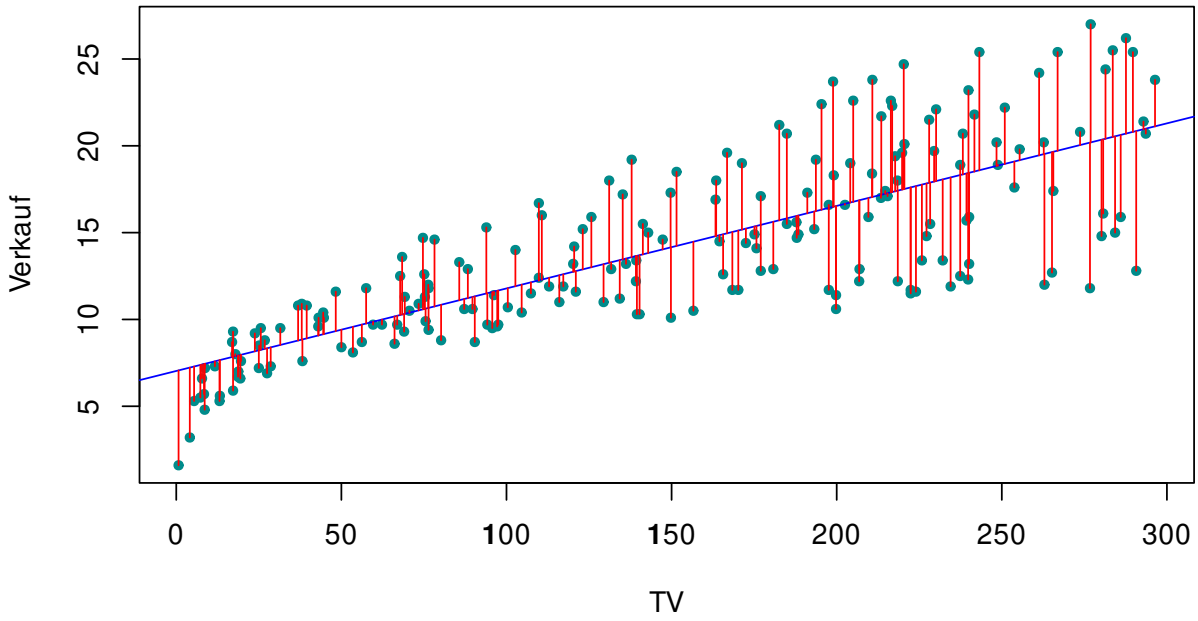


Abbildung 8.2. : Residuen im Beispiel `Werbung`.

oder äquivalent

$$\text{RSS} = (y_1 - \widehat{\beta}_0 - \widehat{\beta}_1 x_1)^2 + (y_2 - \widehat{\beta}_0 - \widehat{\beta}_1 x_2)^2 + \ldots + (y_n - \widehat{\beta}_0 - \widehat{\beta}_1 x_n)^2$$

Bei der Methode der kleinsten Quadrate wird nun $\widehat{\beta}_0$ und $\widehat{\beta}_1$ so gewählt, dass RSS *minimal* wird. Mit Differentialrechnung kann gezeigt werden, dass dann für $\widehat{\beta}_0$ und $\widehat{\beta}_1$ folgendes gilt:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$
$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

mit

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{und} \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Dies sind die mit Hilfe der Methode der kleinsten Quadrate geschätzten Koeffizienten für die einfache lineare Regression. Die praktische Berechnung geschieht immer mit **R**.

Beispiel 8.2.6

Wir wollen $\widehat{\beta}_0$ und $\widehat{\beta}_1$ und die Regressionsgerade für das Beispiel **Werbung** bestimmen.

```
Werbung <- read.csv("../Daten/Werbung.csv")
TV <- Werbung[, 2]
Verkauf <- Werbung[, 5]
lm(Verkauf ~ TV)


##
## Call:
## lm(formula = Verkauf ~ TV)
##
## Coefficients:
## (Intercept)           TV
##     7.03259      0.04754
```

Der Wert unter **Intercept** ist $\widehat{\beta}_0$, also der y-Achsenabschnitt und der Wert unter **TV** ist die Steigung $\widehat{\beta}_1$ der Geraden. Somit erhalten als lineares Modell

$$Y \approx 7.03 + 0.0475X$$

Gemäss dieser Näherung würden wir für zusätzliche CHF 1000 Werbeausgaben 47.5 zusätzliche Einheiten des Produktes verkaufen.

In Abbildung 8.3 ist der **R**-Code und die Regressionsgerade dargestellt.

◀

Bemerkungen:

i. Der **R**-Befehl **lm()** steht für *linear model*.

ii. Man beachte den Unterschied in der Reihenfolge von **x** und **y** in **plot(x,y)** und **lm(y~x)**. ♦

8.2.3. Wie genau sind unsere Schätzungen für die Koeffizienten?

Wir hatten angenommen, dass der *wahre* Zusammenhang von der Form

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

```
par(mar=c(5,4,1,0))
Werbung <- read.csv("../Daten/Werbung.csv")
TV <- Werbung[,2]
Verkauf <- Werbung[,5]
plot(TV,Verkauf,col="darkcyan",xlab="TV",
     ylab="Verkauf",pch=20

)

abline(lm(Verkauf~TV),col="blue")
```

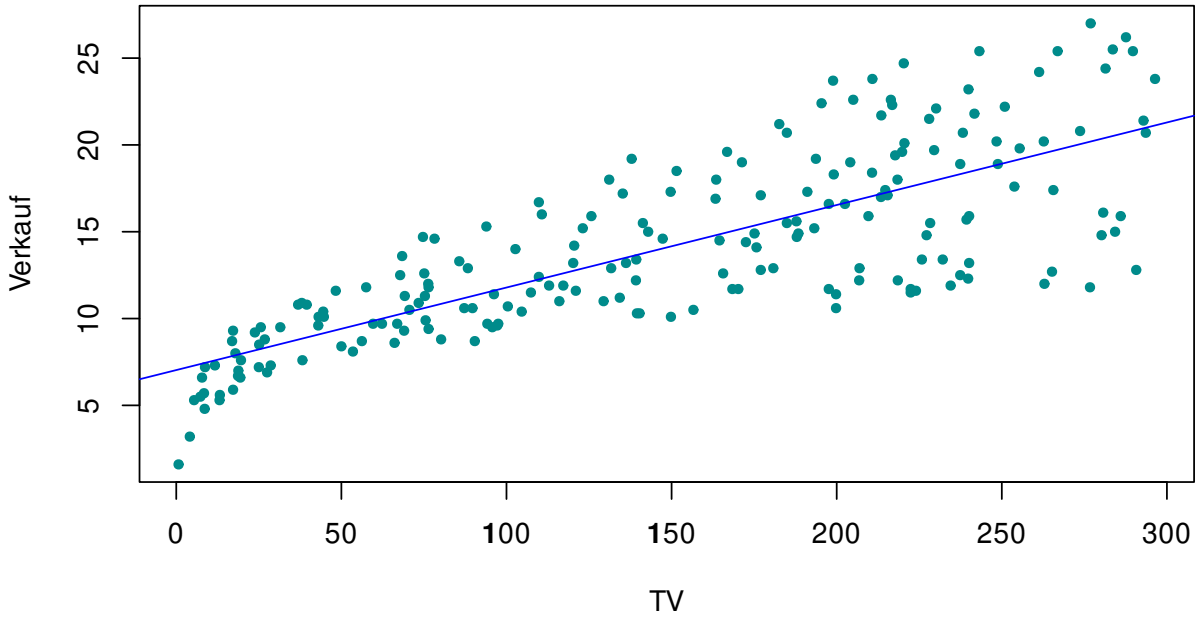


Abbildung 8.3. : `Werbung` mit Regressionsgerade

ist. Dabei ist f eine unbekannte Funktion und ε ist ein zufälliger Fehlerterm mit Mittelwert 0. Wird f durch eine lineare Funktion approximiert, so schreiben wir diesen Zusammenhang als

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Hier ist β_0 der y-Achsenabschnitt mit dem erwarteten Wert für Y, wenn $X = 0$ und β_1 ist die Steigung, also die mittlere Änderung von Y bei einer Zunahme von X um eine Einheit ist. Im Fehlerterm ε ist alles hineingepackt, was wir beim einfachen linearen Modell unterschlagen:

- Der wahre Zusammenhang ist selten linear.
- Es gibt vielleicht noch weitere Variablen, die Y beeinflussen.
- Es gab vielleicht Messfehler.

Für die Summe von diesen zufälligen Variablen darf aufgund des Zentralen Grenzwertsatzes eine Normalverteilung angenommen werden. Weiter nehmen wir in der Regel an, dass der Fehlerterm unabhängig von X ist.

Beispiel 8.2.7

In diesem Beispiel nehmen wir an, dass wir den exakten Zusammenhang zwischen X und Y kennen

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

mit

$$f(X) = 2 + 3X$$

also einer linearen Beziehung. Nun können wir die beobachteten Daten von Y durch

$$Y = 2 + 3X + \varepsilon$$

simulieren, wobei ε normalverteilt mit Mittelwert 0 ist, also

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

Wir erzeugen 100 zufällige Werte von X und die zugehörigen Werte von Y.

```
x <- runif(n = 100, min = -2, max = 2)
y <- 2 + 3 * x + rnorm(n = 100, mean = 0, sd = 4)
```

In Abbildung 8.4 sind 3 solche Simulationen dargestellt.

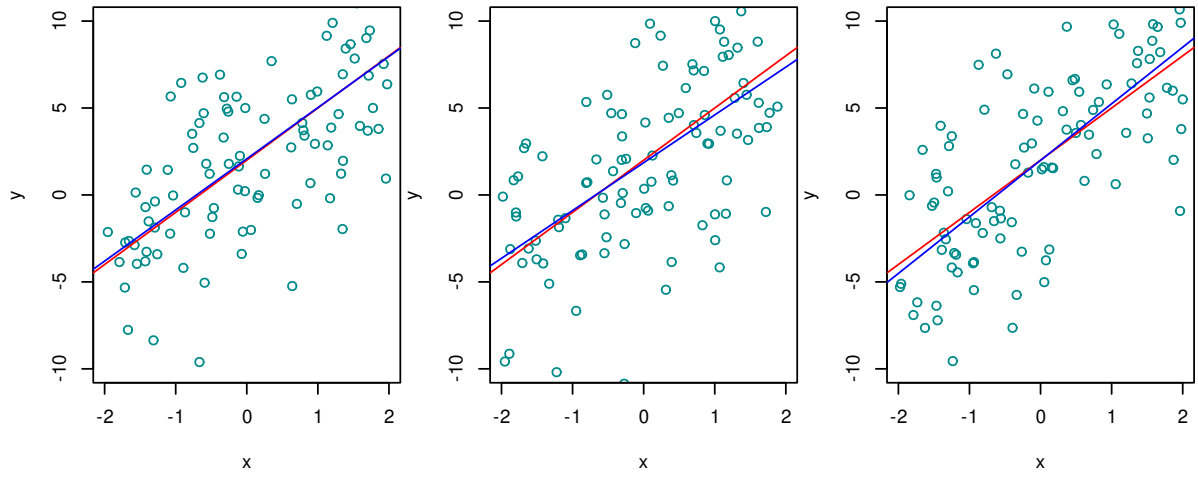


Abbildung 8.4. : Simulationen von $Y = 2 + 3X + \varepsilon$.

Die rote Gerade ist jeweils der Graph der Gleichung $Y = 2 + 3X$, die in allen drei Simulationen gleich bleibt. Die blaue Gerade ist jeweils die Regressionsgerade, die mit Hilfe der Methode der kleinsten Quadrate für die simulierten Datenpunkte bestimmt wurde. Diese ändert sich von Simulation zu Simulation.

In Abbildung 8.5 sind die Regressionsgerade (blau) von 10 Simulationen eingezeichnet. Wir sehen, dass diese der zugrundeliegenden Gerade (rot) ähnlich sind, aber nie gleich.

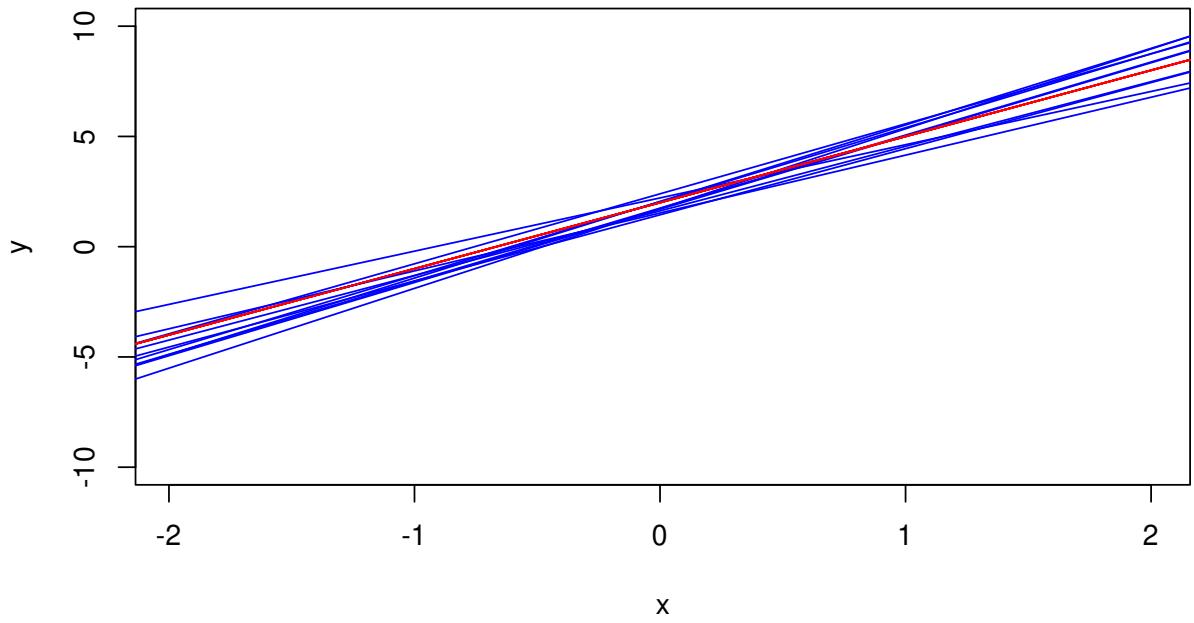


Abbildung 8.5. : Simulationen von $Y = 2 + 3X + \varepsilon$.

◀

Die wahre Beziehung zwischen erklärender Variable X und Zielgrösse Y ist bei reellen Daten im Allgemeinen nie bekannt, dennoch kann die Regressionsgerade immer mit Hilfe der Methode der Kleinsten Quadrate bestimmt werden. In der Anwendung haben wir also Daten, für welche wir die Regressionsgerade bestimmen können. Die wahre lineare Beziehung (sofern sie überhaupt existiert) bleibt uns aber stets unbekannt. Auf das obige Beispiel bezogen, kennen wir die Datensätze und können die blauen Geraden bestimmen, aber wir kennen die rote nicht.

Wir haben also zwei (oder mehr) Geraden, die den Zusammenhang zwischen erklärender und Zielgrössen beschreiben. Bloss kennen wir die Gleichung der wahren (roten) Geraden im Allgemeinen nicht. Wir ziehen also aufgrund eines Datensatzes (blaue Gerade) Rückschlüsse auf den wahren Zusammenhang (rote Gerade), den wir

aber nicht kennen. Das ist allerdings die natürliche Vorgehensweise in der Statistik, wo von Beobachtungen auf die Gesamtheit geschlossen wird.

Beispiel 8.2.8

Wir wollen die durchschnittliche Körpergrösse μ aller 20-Jährigen auf der Erde bestimmen. Nun ist es unmöglich, die Körpergrösse aller 20-Jährigen zu bestimmen. Das heisst, wir müssen μ schätzen

Um aber einen ungefähren Wert $\widehat{\mu}$ für μ zu bekommen, wählen wir eine Gruppe von 1000 und ermitteln von diesen die Körperlänge y_i für $i = 1, \ldots, 1000$ und berechnen den Durchschnitt $\overline{y}$. Eine vernünftige Annahme ist

$$\widehat{\mu} = \overline{y}$$

also

$$\mu \approx \widehat{\mu} = \overline{y}$$

Wählen wir eine andere Gruppe, so wird $\overline{y}$ leicht anders sein. Aber dies ändert nichts an der Tatsache, dass $\mu \approx \overline{y}$. ◀

Vertrauensintervall

Beispiel 8.2.9

Wir erhalten die Vertrauensintervalle für das Beispiel **Werbung** mit **R**:

```
Werbung <- read.csv("../Daten/Werbung.csv")
TV <- Werbung[, 2]
Verkauf <- Werbung[, 5]
confint(lm(Verkauf ~ TV), level = 0.95)


##                 2.5 %     97.5 %
## (Intercept) 6.12971927 7.93546783
## TV          0.04223072 0.05284256
```

Für das 95 %-Vertrauensintervall von β_0 gilt dann

$$[6.130, 7.935]$$

und für β_1

$$[0.042, 0.053]$$

Ohne Werbung liegt der Verkauf also zwischen 6130 und 7935 Einheiten. Auf der anderen Seite verkaufen wir für zusätzliche CHF 1000 für TV-Werbung durchschnittlich irgendwo zwischen 42 und 53 Einheiten mehr. ◀

Hypothesentest

Wir wollen noch untersuchen, ob die geschätzten Parameter $\widehat{\beta}_0$ und vor allem $\widehat{\beta}_1$ statistisch signifikant sind. Dazu machen wir einen Hypothesentest. Der häufigste Hypothesentest besteht aus dem Testen der *Nullhypothese* von

$$H_0: \quad \text{Es gibt } \textit{keinen} \text{ Zusammenhang zwischen } X \text{ und } Y$$

gegen die *Alternativhypothese*

$$H_A: \quad \text{Es gibt } \textit{einen} \text{ Zusammenhang zwischen } X \text{ und } Y$$

Mathematisch entspricht dies

$$H_0: \quad \beta_1 = 0$$

gegen

$$H_A: \quad \beta_1 \neq 0$$

Ist nämlich $\beta_1 = 0$, dann haben wir die Gleichung

$$Y = \beta_0 + \varepsilon$$

und damit hängt Y *nicht* von X ab.

Beispiel 8.2.10

Wir wollen die Null- und Alternativhypothese noch graphisch darstellen (siehe Abbildung **??**).

Abbildung links: Hier ist $\widehat{\beta}_1$ ist praktisch 0. Es gibt keinen Zusammenhang zwischen Variablen x und y.1. Das heisst, egal wie wir x wählen, der zugehörige y-Wert bleibt immer gleich.

Abbildung Mitte: Hier ist $\widehat{\beta}_1$ nicht gleich 0. Es gibt einen Zusammenhang zwischen Variablen x und y.2. Das heisst in diesem Fall, je grösser wir x wählen, umso grösser wird der zugehörige y-Wert.

Abbildung rechts: Hier ist die Gerade zwar leicht steigend, aber es schwer einen Zusammenhang im Streudiagramm zu erkennen. Diese Abweichung kann zufällig sein.

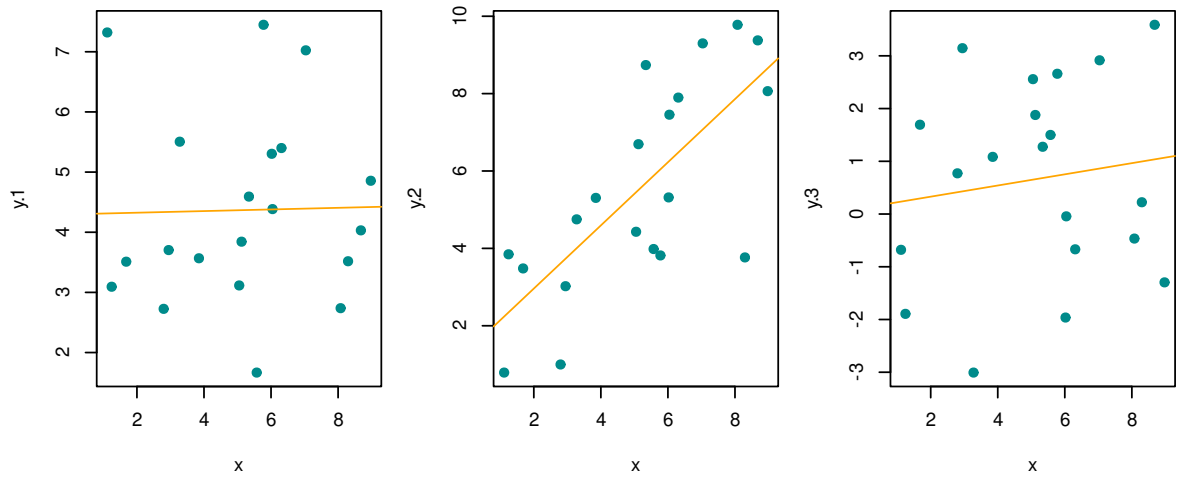


Abbildung 8.6. : Null- und Alternativhypothese graphisch

Oder anders gefragt: Wie stark steigend oder fallend muss die Regressionsgerade sein, damit die Steigung statistisch signifikant ungleich 0 ist? ◀

Um die Nullhypothese zu testen, müssen wir entscheiden, ob $\widehat{\beta}_1$, unsere Schätzung von β_1, genügend weit von 0 weg ist, damit wir sehr sicher sind, dass β_1 nicht 0 ist.

Aber *wie* weit weg von 0 ist genügend weit weg? Dies wird mit einem Hypothesentest entschieden. In diesem Fall handelt es sich um nichts anderes als ein t-Test. Wir lassen allerdings Detail weg und überlassen diese **R**.

Beispiel 8.2.11

Wir wollen den p-Wert von β_1 im Beispiel **Werbung** berechnen.

```
Werbung <- read.csv("../Daten/Werbung.csv")
TV <- Werbung[, 2]
Verkauf <- Werbung[, 5]
summary(lm(Verkauf ~ TV))


##
## Call:
## lm(formula = Verkauf ~ TV)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 7.032594   0.457843   15.36   <2e-16 ***
## TV          0.047537   0.002691   17.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119,Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

Beim Eintrag **Coefficients** unter **Pr(>|t|)** sind die p-Werte $2 \cdot 10^{-16}$ aufgeführt. Diese sind bei weitem kleiner als 0.05. Also können wir die Nullhypothesen $\beta_0 = 0$ und $\beta_1 = 0$ verwerfen, und zwar zu Gunsten der Alternativhypothese $\beta_1 \neq 0$ und $\beta_0 \neq 0$. Wir haben also klare Hinweise für einen Zusammenhang zwischen **TV** und **Verkauf**. ◀

8.2.4. Abschätzung der Genauigkeit des Modells

Sobald wir die Nullhypothese zugunsten der Alternativhypothese verworfen haben, ist es natürlich zu fragen, *in welchem Ausmass das Modell zu den Daten passt*. Die Qualität einer linearen Regression wird typischerweise abgeschätzt durch den *residual standard error* (RSE) und die R^2-Statistik.

Beispiel 8.2.12

Tabelle 8.1 zeigt den RSE, die R^2-Statistik und die F-Statistik (siehe später) für die Anzahl der verkauften Einheiten für das TV Werbebudget.

Menge	Wert
RSE	3.26
R^2	0.612
F-Statistik	312.1

Tabelle 8.1. : RSE, R^2, F-Statistik

Mit **R** erscheinen diese Werte am unteren Ende des Outputs in Beispiel 8.2.11. ◀

Wir werden hier nur auf den R^2-Wert genauer eingehen.

R^2-Statistik

Die R^2-Statistik ist ein Wert zwischen 0 und 1. Sie gibt an, welcher Anteil der Variabilität in Y mit Hilfe des Modells durch X erklärt werden.

Ein Wert nahe bei 1 bedeutet, ein grosser Anteil der Variabiliät durch die Regression erklärt wird. Das Modell beschreibt also die Daten sehr gut.

Ein Wert nahe bei 0 bedeutet, dass die Regression die Variabilität der erklärenden Variablen nicht erklärt.

Wir werden diese Eigenschaften graphisch mittels Beispielen begründen.

Haben wir die Nullhypothese verworfen, das heisst es gibt einen Zusammenhang zwischen Zielvariable und Prädiktor, so stellt sich die Frage in welchem Ausmass das Modell zu den Daten passt.

Beispiel 8.2.13

In Abbildung 8.7 sehen wir zwei Streudiagramme.

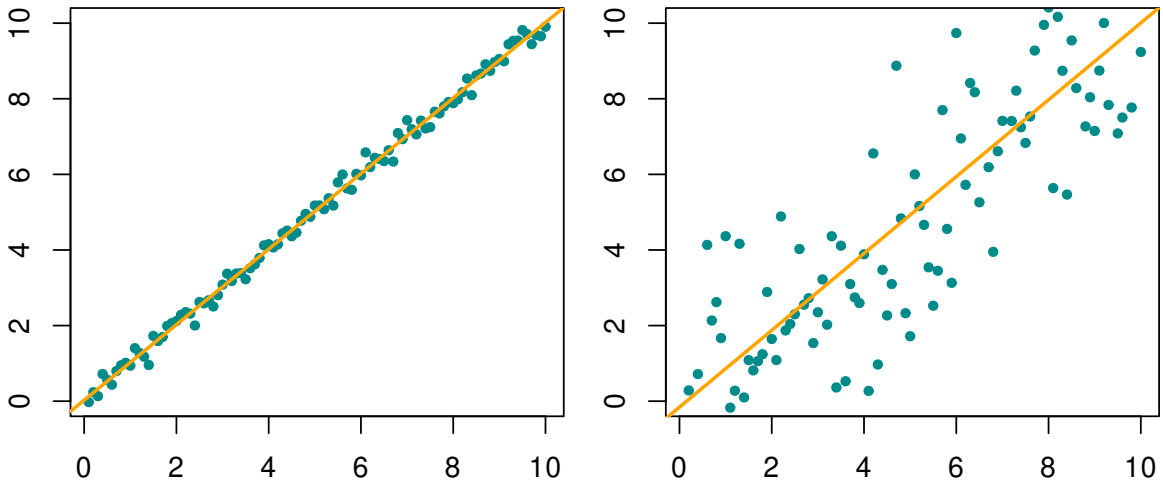


Abbildung 8.7. : Daten, die verschieden gut zur Regressionsgeraden passen

Das Punkte auf der linken Seite passen sehr gut zur Regressionsgeraden, auf der rechten Seite tun sie dies viel schlechter. Die Frage ist, wie wir dies durch einen numerischen Wert zusammenfassen können. ◀

Bemerkung:

In Kapitel 3 haben wir den empirischen Korrelationskoeffizienten kennengelernt, den wir hier anwenden könnten. Er hat aber den Nachteil, dass er nur einfache lineare Regression beschreibt. Wir wollen hier einen Wert kennenlernen, der allgemeiner ist, der R^2-Wert. ♦

Die Qualität einer linearen Regression kann durch den *residual standard error* (RSE) und die R^2-Statistik abgeschätzt werden. Wir behandeln hier nur die R^2-Statistik, da sie einfacher zu interpretieren ist.

Beispiel 8.2.14

Wir beginnen mit einer Punktewolke, die genau einer Geraden folgen (siehe Abbildung 8.8)

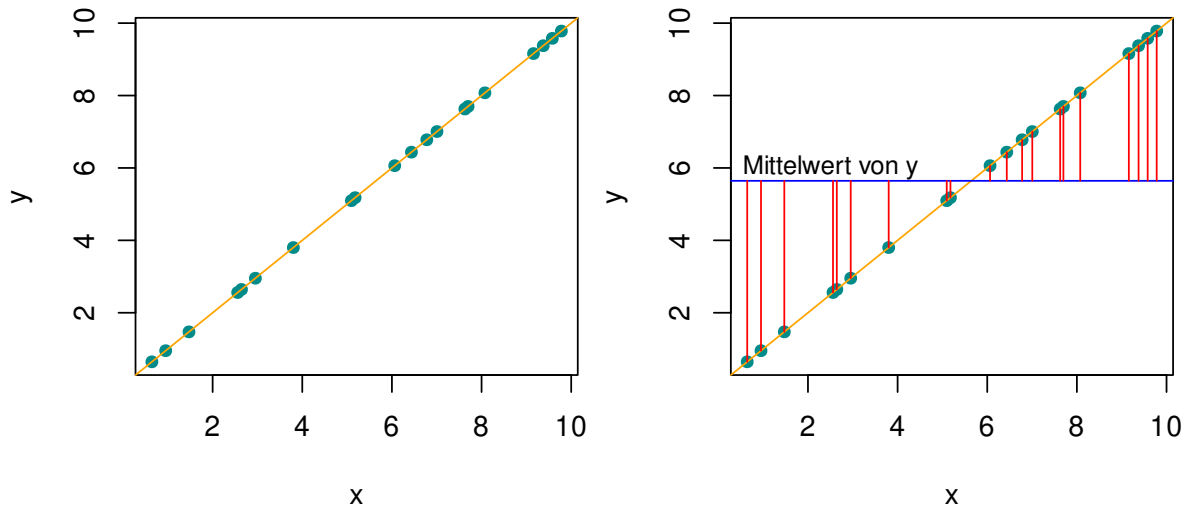


Abbildung 8.8. : Punkte folgen genau einer Geraden und Abstände zum Mittelwert

Der R^2-Wert macht eine Aussage über die Varianz, also wollen wir diese auch noch graphisch darstellen (siehe Abbildung 8.8 rechts).

In Abbildung 8.8 rechts entspricht die blauen Linie dem Mittelwert von y. Die roten Linien sind die Unterschiede der Daten zum Mittelwert. Die Varianz ist dann der „Durchschnitt"[1] der Quadrate der roten Linien.

Wir erhalten für die Varianz

[1] Es ist nicht ganz der Durchschnitt, da in der Definition der Varianz (siehe Unterabschnitt 2.2.2) durch $n-1$ anstatt n dividiert wird. Dieser Unterschied spielt hier allerdings keine Rolle, da sich $n-1$ bez. n bei der Definition von R^2 wegkürzen.

```
var(y)


## [1] 8.998626
```

Nun betrachten wir ein Streudiagramm, wo die mehr oder weniger einer Geraden folgen (siehe Abbildung **??** links).

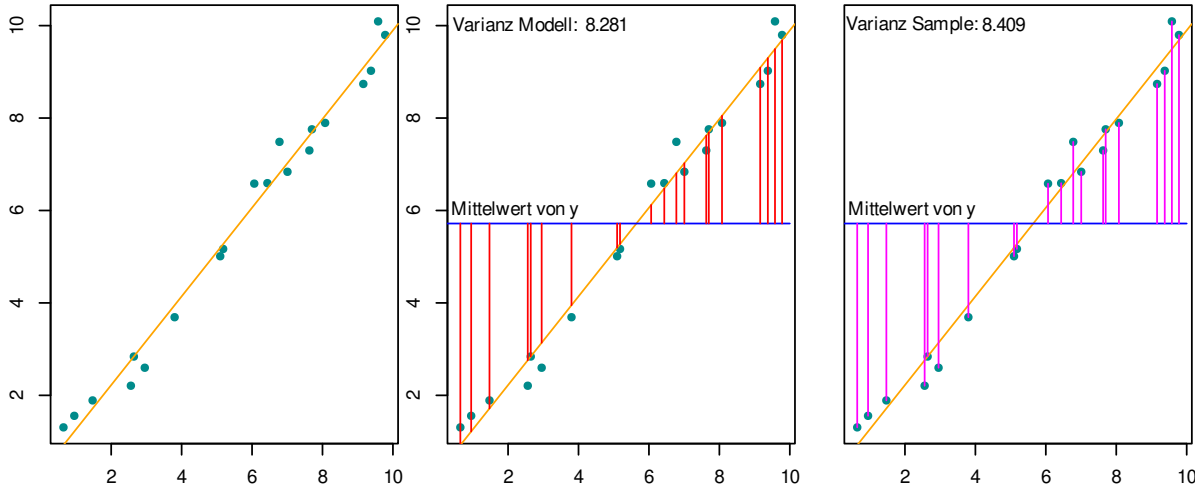


Abbildung 8.9. : Definition von R^2

In diesem Streudiagramm deuten wir zwei verschiedene Varianzen an (siehe Abbildung 8.9 Mitte und rechts).

In Abbildung 8.9 in der Mitte gehen die roten Linien vom Mittelwert von y vertikal zur Regressionsgeraden. Diese Gerade ist unser Modell. Demnach ist der „Durchschnitt" der Quadrate der roten Linien die Varianz des Modelles.

In Abbildung 8.9 rechts sind die pinken Linien die vertikalen Abstände vom Mittelwert zu den Datenpunkten (sample). Somit ist der „Durchschnitt" der Quadrate der pinken Linien die Varianz der Datenpunkte. ◀

Der R^2-Wert wird nun wie folgt definiert:

$$R^2 = \frac{\text{Varianz Modell}}{\text{Varianz Sample}}$$

Beispiel 8.2.15

Für das Streudiagramm in Beispiel 8.2.14 gilt dann

$$R^2 = \frac{8.281}{8.409} = 0.985$$

oder mit **R**

```
summary(lm(y ~ x))$r.squared

## [1] 0.9848312
```

◀

Zur einfacheren Interpretation vom R^2-Wert bringen wir noch eine andere Definition, die aber zu der oben gegebenen äquivalent ist.

Beispiel 8.2.16

Dazu betrachten wir Abbildung 8.10.

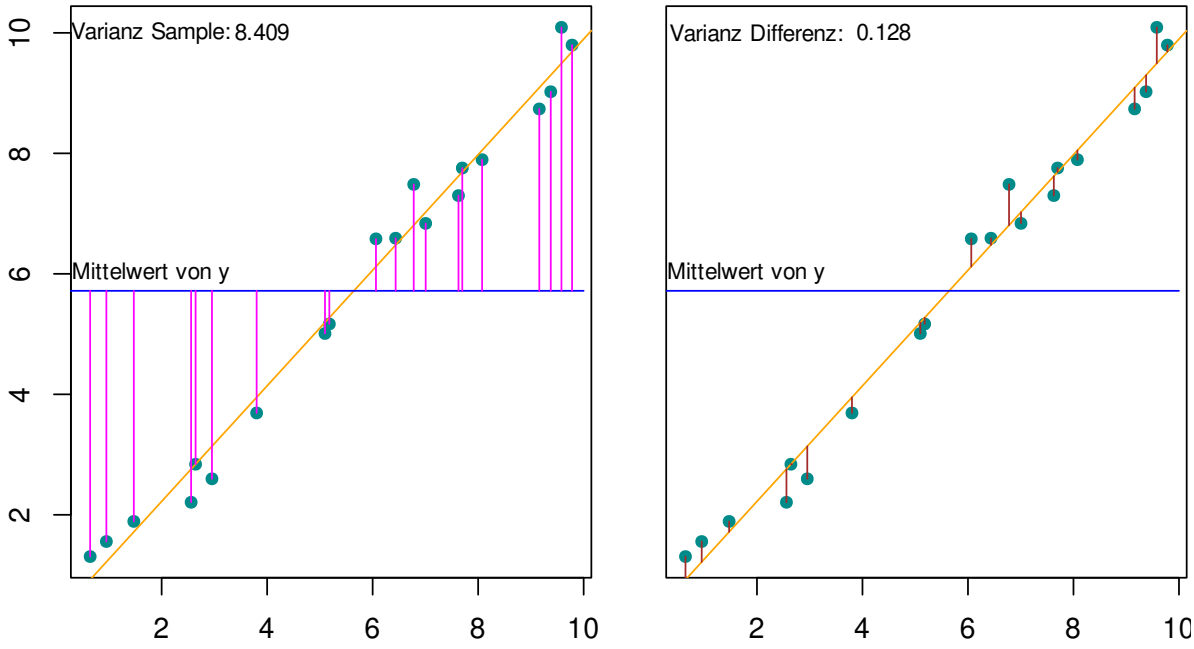


Abbildung 8.10. : Alternative Darstellung von R^2

Auf der linken Seite sehen wir wieder die pinken Linien, deren „Durchschnitt" der Quadrate die Varianz der Daten ergibt.

Auf der rechten Seite sind die brauen Linien der Unterschied der Daten zu dem Modell. Der Durchschnitt der Quadrate der braunen Linien bezeichnen wir mit Varianz der Differenz. ◀

Nun können wir ein weitere Definition von R^2 geben:

$$R^2 = 1 - \frac{\text{Varianz Differenz}}{\text{Varianz Sample}}$$

Beispiel 8.2.17

Für Abbildung 8.10 ergibt dies

$$R^2 = 1 - \frac{\text{Varianz Differenz}}{\text{Varianz Sample}} = 1 - \frac{0.128}{8.409} = 0.985$$

oder mit R:

```
summary(lm(y ~ x))$r.squared

## [1] 0.9848312
```

◀

Mit dieser Definition lässt sich der R^2-Wert einfacher interpretieren:

- Varianz Differenz:

  Dies ist die Varianz des Samples, dass *nicht* durch das Modell erklärt wird.

- $\frac{\text{Varianz Differenz}}{\text{Varianz Sample}}$:

  Anteil der Varianz vom Sample, der *nicht* vom Modell erklärt wird.

- $1 - \frac{\text{Varianz Differenz}}{\text{Varianz Sample}}$:

  Anteil der Varianz vom Sample, der vom Modell erklärt wird.

- R^2:

  Anteil der Varianz vom Sample, der vom Modell erklärt wird.

Beispiel 8.2.18

Somit wird in Abbildung 8.10 98.48% der Varianz von 8.41 wird durch das Modell erklärt.

```
var(y)


## [1] 8.40886
```

Wir wollen nun noch zeigen, dass ein R^2 nahe bei 1 bedeutet, dass die Punkte nahe der Regressionsgeraden liegen. Dazu betrachten wir wieder die Abbildung 8.10, wo dies der Fall ist:

- Varianz Differenz: Hier nahe bei 0
- $\frac{\text{Varianz Differenz}}{\text{Varianz Sample}}$: Nahe bei 0
- $1 - \frac{\text{Varianz Differenz}}{\text{Varianz Sample}}$: Nahe bei 1
- R^2: Passen die Punkte gut zum Modell, dann ist R^2 angenähert 1

◀

Wir betrachten noch ein Beispiel, wo die Punkte nicht gut zu der Geraden passen und zeigen, dass der R^2-Wert nicht mehr nahe bei 1 ist.

Beispiel 8.2.19

Wir betrachten Abbildung 8.11.

Der R^2-Wert ist

$$R^2 = 1 - \frac{22.676}{24.56} = 0.07671$$

oder mit **R**

```
summary(lm(y ~ x))$r.squared


## [1] 0.07670148
```

Es gilt:

- Varianz Differenz: Ähnlich Varianz Sample
- $\frac{\text{Varianz Differenz}}{\text{Varianz Sample}}$: Nahe bei 1
- $1 - \frac{\text{Varianz Differenz}}{\text{Varianz Sample}}$: Nahe bei 0

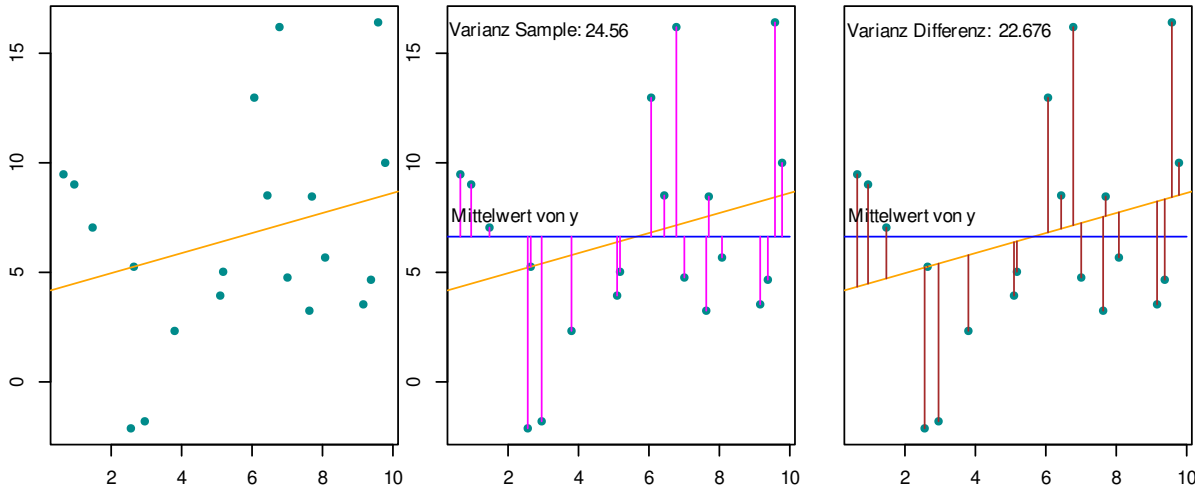


Abbildung 8.11. : Regressionsgerade passt nicht gut zu den Punkten

- R^2: Passen die Punkte nicht gut zum Modell, dann ist R^2 angenähert 0

Wir interpretieren noch diesen R^2-Wert. Nur 7.67% der Varianz von 24.56 wird durch das Modell erklärt

```
var(y)


## [1] 24.55976
```

◀

Der R^2-Wert können wir auch anwenden, wenn das Modell nicht linear ist.

Beispiel 8.2.20

Wir betrachten die Abbildung 8.13, wo die Punkte mehr oder weniger dem Modell folgen.

Wir erhalten einen R^2-Wert von

$$R^2 = 1 - \frac{5.888}{1026.155} = 0.994262$$

oder mit R^2

```
summary(lm(y ~ I(x^2)))$r.squared


## [1] 0.9942619
```

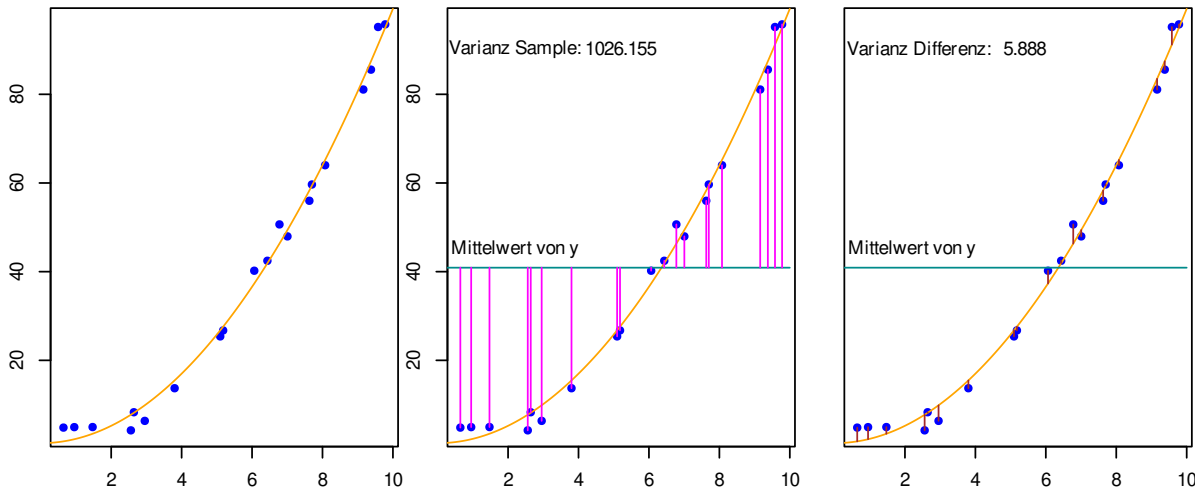


Abbildung 8.12. : Punkte folgen mehr oder weniger einem quadratischen Modell

```
var(y)

## [1] 1026.155
```

Damit werden 99.43% der Varianz von 1026.15 wird durch das Modell erklärt. R^2-Wert nahe bei 1 und somit passen die Daten gut zum Modell. ◀

Beispiel 8.2.21

Wir betrachten Abbildung 8.13, wo die Punkte nicht gut zum Modell passen.

Berechnung R^2:

$$R^2 = 1 - \frac{588.818}{1262.354} = 0.533556$$

oder mit **R**:

```
summary(lm(y ~ I(x^2)))$r.squared

## [1] 0.5335559
```

```
var(y)

## [1] 1262.354
```

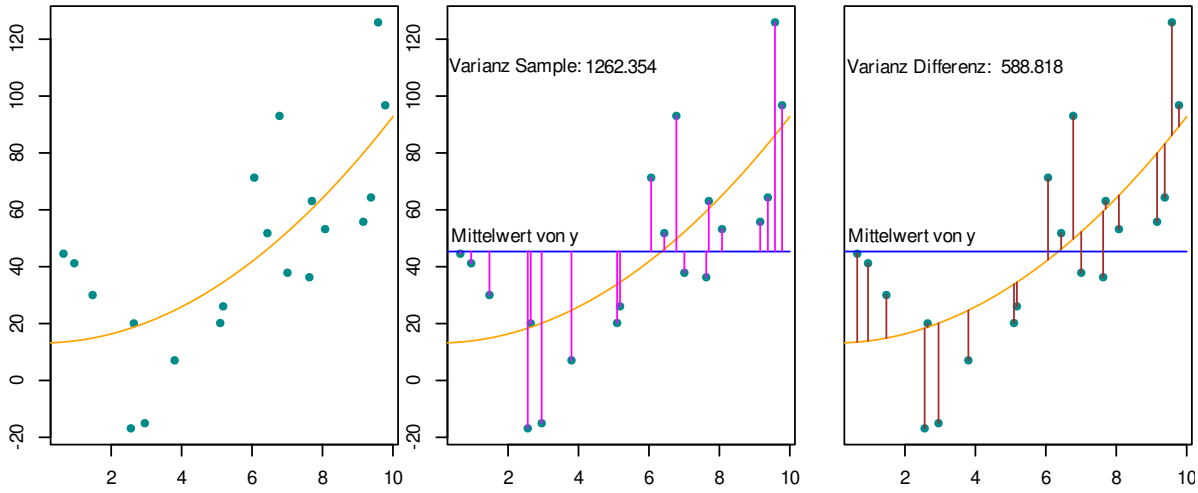


Abbildung 8.13. : Punkte passen nicht gut zum quadratischen Modell

Somit wird 53.36% der Varianz von 1262.35 durch das Modell erklärt. Der R^2-Wert ist nicht so nahe bei 1 und somit passen die Daten nicht so gut zum Modell ◀

Bemerkungen:

i. Die empirische Korrelation r gibt nur die Güte einer *linearen* Regression an

ii. Für eine einfache lineare Regression gilt:

$$r^2 = R^2$$

iii. R^2 kann für jede Regression angewendet werden ♦

Beispiel 8.2.22

Im Beispiel der TV-Werbung war der R^2-Wert 0.61. Somit werden knapp zwei Drittel der Variabilität in `Verkauf` durch `TV` mit linearer Regression erklärt. ◀

8.3. Multiple lineare Regression

8.3.1. Einführung

Einfache lineare Regression ist ein nützliches Vorgehen, um den Output aufgrund einer einzelnen erklärenden Variablen vorherzusagen. Allerdings hängt der Output

in der Praxis oft von mehr als einer erklärenden Variablen ab.

Beispiel 8.3.1

Im Datensatz **Werbung** haben wir den Zusammenhang zwischen TV-Werbung und dem Verkauf untersucht. Wir haben aber auch Daten für die Werbeausgaben für Radio und Zeitung, und wir können uns fragen, ob sich eine oder beide dieser Werbeausgaben auf den Verkauf auswirken. Wie können wir unsere Analyse der Verkaufszahlen erweitern, damit diese beiden zusätzlichen Inputs mitberüchsichtigt werden?

Eine Möglichkeit wäre, für jedes separate Werbebudget eine einfache Regression durchzuführen. In Abbildung 8.14 sind die einzelnen Plots mit Regressionsgeraden (blau) aufgeführt.

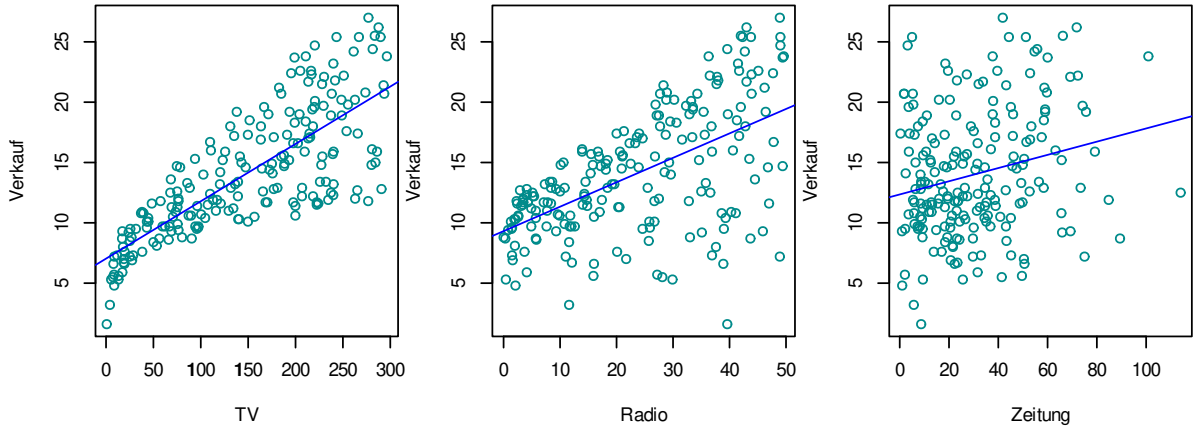


Abbildung 8.14. : Regressionsgeraden in Abbildung 7.1 für **Werbung**

Die Parameter und weitere wichtige Daten sind in Tabelle 8.2 aufgeführt. Diese wurden jeweils wie in Beispiel 8.2.11 auf Seite 259 mit der Methode der kleinsten Quadrate bestimmt.

Wir können beispielsweise mittels einfacher linearer Regression den Verkauf aufgrund der Werbeausgaben für das Medium Radio vorhersagen. Würden wir CHF 1000 mehr in die Werbung für das Radio investieren, so würden wir nach diesem Modell ungefähr 203 Einheiten mehr verkaufen (siehe Tabelle 8.2 Mitte).

Auf der anderen Seite hätte eine Zunahme der Werbeausgaben für die Zeitung um CHF 1000 nur eine Vergrösserung des Verkaufes um ungefähr 55 Einheiten zur Folge (siehe Tabelle 8.2 unten).

Dieser Ansatz, bei der die Anpassung der Daten durch drei separate einfache lineare Regressionen erfolgt, ist allerdings nicht ganz zufriedenstellend. Erstens ist es nicht klar, wie wir für gegebene Werte der drei erklärenden Variablen eine Vorhersage für den Verkauf machen wollen, da jeder Input durch eine andere Regressionsgleichung

Einfache Regression von **Verkauf** auf **TV**

	Koeffizient	Std.fehler	t-Statistik	P-Wert
Intercept	7.033	0.458	15.36	< 0.0001
TV	0.048	0.003	17.67	< 0.0001

Einfache Regression von **Verkauf** auf **Radio**

	Koeffizient	Std.fehler	t-Statistik	P-Wert
Intercept	9.312	0.563	16.54	< 0.0001
Radio	0.203	0.020	9.92	< 0.0001

Einfache Regression von **Verkauf** auf **Zeitung**

	Koeffizient	Std.fehler	t-Statistik	P-Wert
Intercept	12.351	0.621	19.88	< 0.0001
Zeitung	0.055	0.017	3.30	< 0.0001

Tabelle 8.2. : Einfache lineare Regression auf die einzelnen Werbebudgets.

mit dem Verkauf verknüpft ist. Zweitens ignoriert jede der drei Regressionsgleichungen die beiden anderen erklärenden Variablen für die Bestimmung der Regressionskoeffizienten.

Wir werden in Kürze sehen, dass dies zu sehr irreführenden Schätzungen der Wirkung der Werbeausgaben für jedes einzelne Medium auf den Verkauf haben kann, falls die drei erklärenden Variablen miteinander korrelieren.

◀

Anstatt getrennte einfache lineare Regressionen für jede erklärende Variable zu betrachten, ist es besser, das Modell für die einfache lineare Regression

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

so zu erweitern, dass alle erklärenden Variablen direkt mitberücksichtigt werden. Wir können dann jeder erklärenden Variablen einen eigenen Steigungskoeffizienten in *einer* Gleichung zuordnen.

Allgemein gehen wir davon aus, dass wir p verschiedene erklärende Variablen haben. Dann hat das *multiple lineare Regressionsmodell* die Form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

Dabei steht X_j für den j-ten Input und β_j für den Zusammenhang zwischen dieser erklärenden Variablen und der Zielgrösse Y.

Wir können dann β_j als die durchschnittliche Änderung der Zielgrösse bei Änderung von X_j um eine Einheit betrachten, *wenn alle anderen erklärenden Variablen festgehalten werden*.

Beispiel 8.3.2

Das multiple lineare Regressionsmodell für den Datensatz **Werbung** sieht dann wie folgt aus:

$$\texttt{Verkauf} = \beta_0 + \beta_1 \cdot \texttt{TV} + \beta_2 \cdot \texttt{Radio} + \beta_3 \cdot \texttt{Zeitung} + \varepsilon$$

also

$$\texttt{Verkauf} \approx \beta_0 + \beta_1 \cdot \texttt{TV} + \beta_2 \cdot \texttt{Radio} + \beta_3 \cdot \texttt{Zeitung}$$

◀

Da das multiple lineare Modell eine Verallgemeinerung des einfachen linearen Modells ist, bleiben Berechnungen und Interpretationen für das multiple Modell ähnlich, wenn sie auch meist komplizierter sind als beim linearen Modell.

Im Kapitel 8 haben wir ausführlich graphische Methoden zur Analyse des einfachen linearen Regressionsmodells kennengelernt. Graphische Methoden entfallen für das multiple lineare System praktisch vollends. Die Datenpunkte für Beispiel 8.3.2 können graphisch nicht in einem Koordinatensystem dargestellt werden, da wir schon für die erklärenden Variablen drei Achsen brauchen.

Trotzdem ist folgendes Beispiel illustrierend.

Beispiel 8.3.3

Für den Datensatz **Einkommen** war bis jetzt **Ausbildung** die einzige erklärende Variable. Allerdings ist das Einkommen oft auch noch von der **Erfahrung** in Anzahl Berufsjahren abhängig.

Somit haben wir das multiple lineare Modell

$$\texttt{Einkommen} = \beta_0 + \beta_1 \cdot \texttt{Ausbildung} + \beta_2 \cdot \texttt{Erfahrung} + \varepsilon$$

Da wir hier nur zwei erklärende Variablen haben, können wir die Datenpunkte im Raum darstellen (siehe Abbildung 8.15).

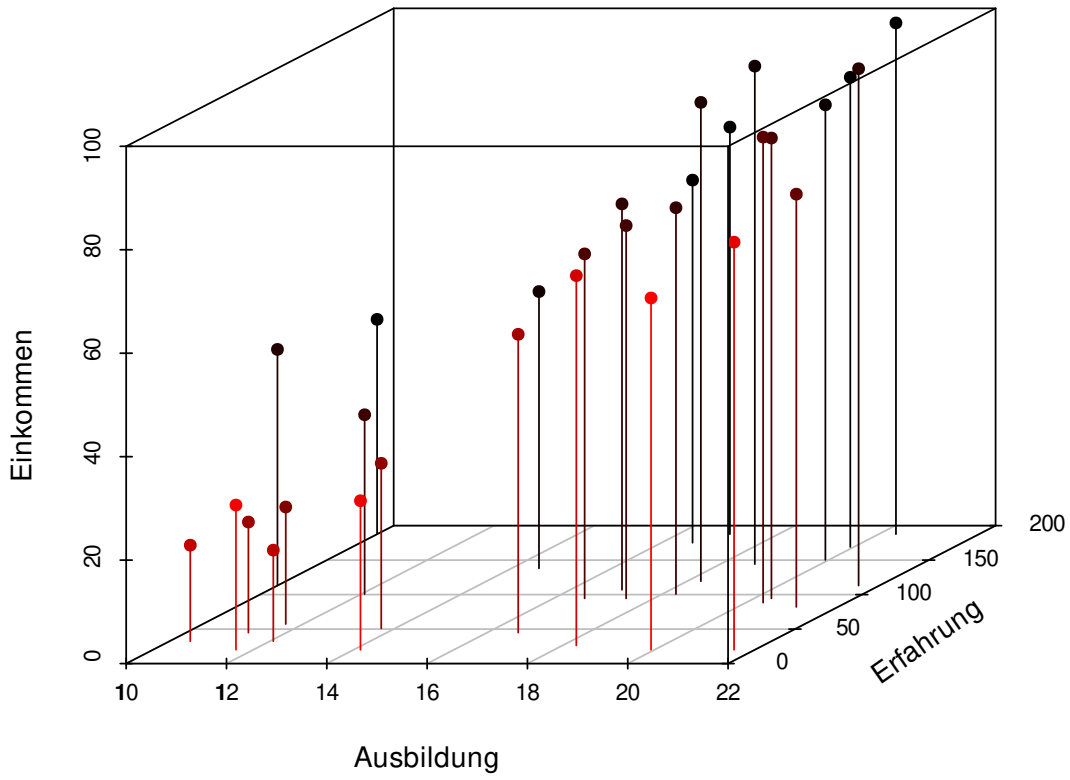


Abbildung 8.15. : Datenpunkte im Raum für den Datensatz **Einkommen**

```
In <- read.csv("../Daten/Einkommen2.csv")
Ausbildung <- In[,2]
Erfahrung <- In[,3]
Einkommen <- In[,4]
library(scatterplot3d)
s3d <- scatterplot3d(Ausbildung,Erfahrung, Einkommen,
       highlight.3d=T,type="h",pch=16,angle=30

)
```

Analog zum einfachen linearen Regressionsmodell suchen wir hier eine *Ebene*, die am „besten" zu den Datenpunkten passt (siehe Abbildung 8.16).

Das Vorgehen ist nun analog zur einfachen linearen Regression. Wir bestimmen die Ebene so, dass die Summe der Quadrate der Abstände der Datenpunkte zur Ebene

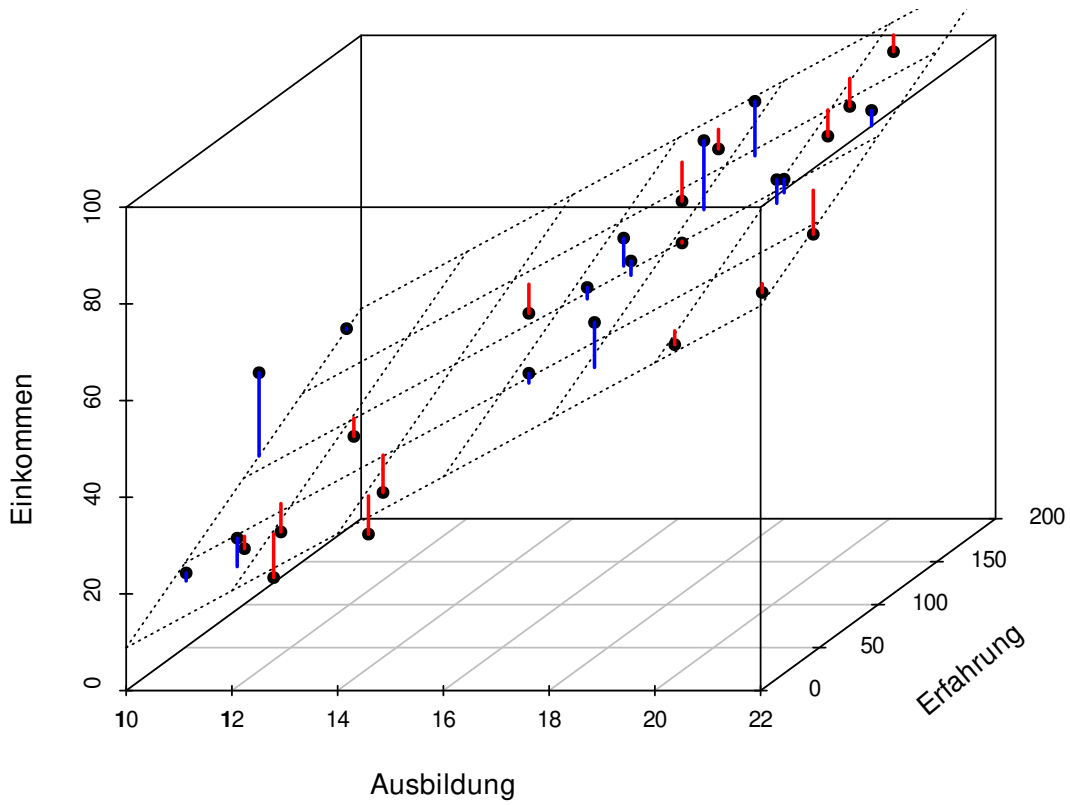


Abbildung 8.16. : Datenpunkte im Raum für den Datensatz `Einkommen`

minimal wird. Die Punkte, die mit blauen Strecken in Abbildung 8.16 mit der Ebene verbunden sind, liegen oberhalb der Ebene, die anderen unterhalb. Diese farbigen Strecken entsprechen den Residuen.

Wir wenden also wieder die Methode der kleinsten Quadrate an. Mit folgendem R-Befehl können wir die Koeffizienten β_0, β_1 und β_2 schätzen und erhalten die Werte

$$\widehat{\beta}_0 = -50.086; \qquad \widehat{\beta}_1 = 5.896; \qquad \widehat{\beta}_2 = 0.173$$

```
In <- read.csv("../Daten/Einkommen2.csv")
Ausbildung <- In[, 2]
Erfahrung <- In[, 3]
Einkommen <- In[, 4]
coef(lm(Einkommen ~ Ausbildung + Erfahrung))
```

```
## (Intercept)  Ausbildung   Erfahrung
## -50.0856388   5.8955560   0.1728555
```

Es gilt dann für dieses multiple lineare Modell

$$\texttt{Einkommen} \approx -50.086 + 5.896 \cdot \texttt{Ausbildung} + 0.173 \cdot \texttt{Erfahrung}$$

◀

Wir werden nun die Überlegungen aus dem letzten Beispiel verallgemeinern.

8.3.2. Schätzung der Regressionskoeffizienten

Wie bei der einfachen linearen Regression sind die Regressionskoeffizienten $\beta_0, \beta_1, \ldots, \beta_p$ im allgemeinen unbekannt, und wir müssen sie schätzen. Aufgrund der Schätzungen $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p$ können wir mit der Formel

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \ldots + \ldots + \widehat{\beta}_p x_p$$

Vorhersagen machen.

Die Parameter werden wieder mit dem Ansatz der kleinsten Quadrate geschätzt, wie wir das im Beispiel 8.3.3 für $p = 2$ schon gesehen haben. Wir wählen $\beta_0, \beta_1 \ldots, \beta_p$ so, dass die Summe der Residuenquadrate (RSS)

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} r_i^2 \\ &= \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \widehat{\beta}_1 x_{i1} - \widehat{\beta}_2 x_{i2} - \ldots - \widehat{\beta}_p x_{ip})^2 \end{aligned}$$

minimiert wird. Dabei ist x_{ij} die i-te Beobachtung der j-ten erklärenden Variable.

Die Werte $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p$ minimieren den RSS. Die Formeln für diese Parameterschätzungen sind etwas verschieden von den Schätzungen im Falle der einfachen linearen Regression in Abschnitt 8.2.3. Das Prinzip ist allerdings dasselbe und deshalb führen wir die Rechnungen hier nicht explizit durch. **R** übernimmt diese Arbeit für uns:

Beispiel 8.3.4

Mit **R** schätzen wir die Koeffizienten für das multiple lineare Regressionsmodell in Beispiel 8.3.2 für den Datensatz **Werbung**.

```
Werbung <- read.csv("../Daten/Werbung.csv")
TV <- Werbung[, 2]
Radio <- Werbung[, 3]
Zeitung <- Werbung[, 4]
Verkauf <- Werbung[, 5]
coef(lm(Verkauf ~ TV + Radio + Zeitung))


##  (Intercept)           TV        Radio      Zeitung
##  2.938889369  0.045764645  0.188530017 -0.001037493
```

Wir können diese Koeffizienten wie folgt interpretieren: Für gegebene Werbeausgaben für TV und Zeitung werden für zusätzliche CHF 1000 Werbeausgaben für das Radio ungefähr 189 Einheiten mehr verkauft.

In Tabelle 8.3 sind weitere wichtige Werte aufgeführt. Diese erhalten wir, indem wir im **R**-Code oben **coef** durch **summary** ersetzen.

	Koeffizient	Std.fehler	t-Statistik	P-Wert
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
Radio	0.189	0.0086	21.89	< 0.0001
Zeitung	-0.001	0.0059	-0.18	0.8599

Tabelle 8.3. : Multiple lineare Regression für den Datensatz **Werbung**.

Betrachten wir die Koeffizienten der separat durchgeführten einfachen linearen Regressionen in Tabelle 8.2, so sehen wir, dass die Steigungskoeffizienten der multiplen linearen Regression für **TV** und **Radio** sehr ähnlich sind.

Allerdings ist der für das einfache lineare Modell geschätzte Regressionskoeffizient $\widehat{\beta}_1$ für die erklärende Variable **Zeitung** verschieden von 0, während er im multiplen linearen Modell praktisch 0 ist. Hinzu kommt, dass der entsprechende p-Wert mit 0.86 bei weitem nicht mehr signifikant ist (siehe Tabellen 8.2 und 8.3). ◀

Dieses Beispiel illustriert, dass die einfachen und multiplen Regressionskoeffizienten sehr verschieden sein können. Der Unterschied erklärt sich aus der Tatsache, dass bei der einfachen Regression die Steigung die Änderung der Zielgrösse **Verkauf** angibt, wenn wir CHF 1000 mehr für die Zeitungswerbung ausgeben, wobei die beiden anderen erklärenden Variablen **TV** und **Radio** *ignoriert* werden.

Bei der multiplen linearen Regression hingegen beschreibt die Steigung für **Zeitung** die Änderung der Zielgrösse **Verkauf**, wenn wir CHF 1000 mehr für die Zeitungswerbung ausgeben, wobei die beiden anderen erklärenden Variablen **TV** und **Radio** *festgehalten* werden.

Macht es Sinn, dass die multiple Regression keinen Zusammenhang zwischen **Verkauf** und **Zeitung** andeutet, aber die einfache Regression das Gegenteil impliziert? Es macht in der Tat Sinn. Betrachten Sie dazu die in Tabelle 8.4 aufgeführten Korrelationskoeffizienten.

	TV	**Radio**	**Zeitung**	**Vekauf**
TV	1.0000	0.0548	0.0567	0.7822
Radio		1.0000	0.3541	0.5762
Zeitung			1.0000	0.2283
Verkauf				1.0000

Tabelle 8.4. : Korrelationskoeffizienten für Datensatz **Werbung**.

```
Werbung <- read.csv("../Daten/Werbung.csv")
TV <- Werbung[, 2]
Radio <- Werbung[, 3]
Zeitung <- Werbung[, 4]
Verkauf <- Werbung[, 5]
cor(data.frame(TV, Radio, Zeitung, Verkauf))


##                 TV      Radio    Zeitung   Verkauf
## TV      1.00000000 0.05480866 0.05664787 0.7822244
## Radio   0.05480866 1.00000000 0.35410375 0.5762226
## Zeitung 0.05664787 0.35410375 1.00000000 0.2282990
## Verkauf 0.78222442 0.57622257 0.22829903 1.0000000
```

Der Korrelationskoeffizient zwischen **Radio** und **Zeitung** beträgt 0.35. Was bedeutet dies? Er zeigt auf, dass es eine Tendenz gibt, bei höheren Werbeausgaben für das Radio auch mehr in die Werbung für Zeitungen zu investieren.

Nehmen wir nun an, dass das multiple Regressionsmodell korrekt sei. Dann haben die Werbeausgaben für Zeitungen keinen direkten Einfluss auf die Zielgrösse **Verkauf**. Allerdings bewirken die Werbeausgaben fürs Radio höhere Verkäufe. Dann sind in Märkten, wo wir mehr in die Werbung fürs Radio investieren, auch die Werbeausgaben für Zeitungen grösser, und zwar aufgrund des Korrelationskoeffizienten von 0.35. In der einfachen linearen Regression betrachten wir ausschliesslich den Zusammenhang zwischen **Zeitung** und **Verkauf**, wobei wir für höhere Werte von **Zeitung** auch höhere Werte für **Verkauf** beobachten. In Tat und Wahrheit beeinflusst aber die Zeitungswerbung die Verkäufe nicht direkt, sondern die höheren Werte für **Zeitung** haben wegen der Korrelation auch grössere Werte für **Radio** zur Folge, und *diese*

Grösse beeinflusst **Verkauf**. **Zeitung** schmückt sich hier mit fremden Lorbeeren, nämlich dem Erfolg von **Radio** auf **Verkauf**.

Dieses Resultat steht in Konflikt mit unserer Intuition. Sie tritt in realen Situationen aber häufig auf.

Beispiel 8.3.5

Hier noch ein eher absurdes Beispiel. Eine einfache Regression wird einen positiven Zusammenhang zwischen Haiattacken auf Menschen und Glaceverkäufen an einem bestimmten Strand in Australien über einen bestimmten Zeitraum aufzeigen. Je grösser die Glaceverkäufe, desto häufiger ereignen sich Haiattacken.

Natürlich kommt niemand auf die Idee, Glaceverkäufe an diesem Strand zu verbieten, damit es keine Haiattacken auf Menschen mehr gibt. Wo liegt aber der Zusammenhang?

In Wirklichkeit kommen bei heissem Wetter mehr Menschen an den Strand, was zu mehr Glaceverkäufen und schliesslich auch zu mehr Haiattacken führt. Ein multiples Regressionsmodell von Haiattacken gegen Glaceverkäufen *und* Temperatur bringt zum Vorschein, dass der Glaceverkauf keinen Einfluss mehr auf die Haiattacken hat, die Lufttemperatur allerdings schon. ◀

8.3.3. Einige wichtige Fragestellungen

Wenn wir eine multiple lineare Regression durchführen, sind wir daran interessiert einige wichtige Fragen zu beantworten:

1. *Ist mindestens eine der erklärenden Variablen $X_1, \ldots, X_p$ nützlich, um die Zielgrösse vorherzusagen?*
2. *Spielen alle erklärenden Variablen $X_1, \ldots, X_p$ für die Vorhersage von Y eine Rolle, oder ist es nur eine Teilmenge der erklärenden Variablen?*
3. *Wie gut passt das Modell zu den Daten?*
4. *Welche Zielgrösse können wir aufgrund konkreter Werte der erklärenden Variablen vorhersagen, und wie genau ist diese Vorhersage?*

Wir werden uns im Folgenden mit diesen Fragen beschäftigen.

Gibt es einen Zusammenhang zwischen den erklärenden Variablen und der Zielgrösse?

Im Abschnitt **??** auf Seite ?? stellten und beantworteten wir die entsprechende Frage für die einfache lineare Regression. Ob es einen Zusammenhang zwischen erklärenden Variablen und Zielgrösse gibt, wurde durch β_1 entschieden. Ist $\beta_1 = 0$, so gibt es keinen Zusammenhang, ansonsten schon.

Bei der multiplen linearen Regression mit p erklärenden Variablen müssen wir uns fragen, ob *alle* Regressionskoeffizienten ausser β_0 Null sind, also

$$\beta_1 = \beta_2 = \ldots = \beta_p = 0$$

Wie bei der einfachen linearen Regression wird dies mit einem Hypothesentest entschieden. Wir testen die Nullhypothese

$$H_0 : \quad \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

gegen die Alternativhypothese

$$H_A : \quad \text{mindestens ein } \beta_i \text{ ist ungleich } 0$$

Dieser Hypothesentest wird mit der Berechnung der *F-Statistik*. Dabei wird ein *F*-Wert berechnet. Ist dieser Wert „weit" weg von 0, so wird die Nullhypothese verworfen, ansonsten beibehalten. Aus diesem *F*-Wert wird dann wieder ein *p*-Wert berechnet mit dem der Testentscheid getroffen. Wir werden auf die Details des *F*-Wertes nicht eingehen, da sie erheblich sind. In Abbildung 8.17 ist eine *F*-Verteilung dargestellt.

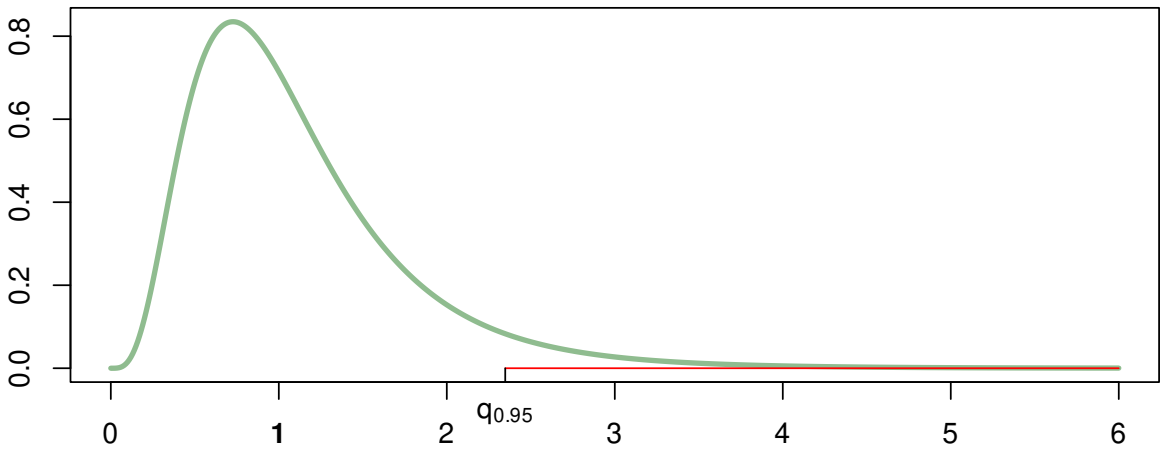


Abbildung 8.17. : *F*-Verteilung mit Verwerfungsbereich

Beispiel 8.3.6

Die F-Statistik für das multiple lineare Modell für den Datensatz **Werbung** ist 570 und ist in der **R**-Ausgabe unter **F-statistic** aufgeführt.

```
Werbung <- read.csv("../Daten/Werbung.csv")
TV <- Werbung[, 2]
Radio <- Werbung[, 3]
Zeitung <- Werbung[, 4]
Verkauf <- Werbung[, 5]
summary(lm(Verkauf ~ TV + Radio + Zeitung))


##
## Call:
## lm(formula = Verkauf ~ TV + Radio + Zeitung)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.938889   0.311908   9.422   <2e-16 ***
## TV           0.045765   0.001395  32.809   <2e-16 ***
## Radio        0.188530   0.008611  21.893   <2e-16 ***
## Zeitung     -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972,Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Dieser Wert ist bei weitem grösser als 1 und liefert überzeugenden Hinweis gegen die Nullhypothese. Das heisst, dass mindestens eine erklärende Variable Einfluss auf die Zielgrösse **Verkauf** hat. ◀

Was aber, wenn der F-Wert näher bei 1 liegt? Wie gross muss der Wert der F-Statistik sein, damit wir H_0 verwerfen und schliessen, dass es mindestens einen Zusammenhang gibt?

Es stellt sich heraus, dass die Antwort von den Werten n und p abhängt. Ist n gross, so kann ein F-Wert wenig grösser als 1 ein genügend starker Hinweis gegen H_0 sein. Im Gegensatz dazu brauchen wir einen grossen F-Wert, wenn n klein ist.

Ist H_0 wahr und sind die Fehler ε_i normalverteilt, so folgt die F-Statistik einer F-Verteilung (mit p und $n-p-1$ Freiheitsgraden). Für jeden Wert von n und p berechnet **R** (wie auch jede andere statistische Software) den p-Wert, also die Wahrscheinlichkeit beruhend auf der F-Verteilung, dass wir einen extremeren F-Wert als den beobachteten messen. Auf der Basis des p-Wertes können wir dann die Nullhypothese H_0 verwerfen oder nicht. Das Vorgehen ist analog zur einfachen linearen Regression, wo allerdings die t-Statistik verwendet wurde.

Beispiel 8.3.7

In der **R**-Ausgabe **p-value** in der Zeile für die F-Statistik im Beispiel 8.3.6 ist der p-Wert für die F-Statistik für das multiple lineare Modell praktisch null. Damit haben wir einen sehr überzeugenden Hinweis, dass mindestens eine erklärende Variable für die Zunahme von **Verkauf** bei vergrösserten Werbeausgaben verantwortlich ist. ◀

Warum müssen wir noch die F-Statistik betrachten, wenn wir schon die p-Werte für die einzelnen Inputs kennen? Es scheint so zu sein, dass für einen kleinen p-Wert *mindestens eine erklärende Variable für die Zielgrösse verantwortlich ist.*

Das muss allerdings nicht sein, wenn die Anzahl p der erklärenden Variablen sehr gross ist. Der p-Wert ist eine Wahrscheinlichkeitsaussage, so dass bei vielen Variablen einige p-Werte zufällig klein sein können.

Ist beispielsweise $p = 100$ und

$$H_0: \quad \beta_1 = \beta_2 = \ldots = \beta_{100} = 0$$

wahr, dann hängt keine Variable wirklich mit der Zielgrösse zusammen. In dieser Situation werden etwa 5 % der einzelnen p-Werte wegen des Zufalls unter 0.05 sein. Wir erwarten also annähernd 5 *kleine* p-Werte, obwohl es keinen Zusammenhang zwischen erklärenden Variablen und Zielgrösse gibt. Die Wahrscheinlichkeit, dass wir zufällig einen p-Wert unter 0.05 beobachten, ist also ziemlich gross. Betrachten wir also nur die t-Statistiken mit den zugehörigen p-Werten der einzelnen erklärenden Variablen, so gibt es eine hohe Wahrscheinlichkeit, dass wir inkorrekt zu Gunsten eines Zusammenhangs entscheiden werden.

Die F-Statistik hat dieses Problem nicht, weil sie alle erklärenden Variablen berücksichtigt. Ist H_0 wahr, dann trifft mit einer Wahrscheinlichkeit von 5 % ein, dass die F-Statistik einen p-Wert kleiner als 0.05 zur Folge hat, und zwar *unabhängig* von der Anzahl erklärender Variablen oder der Anzahl Beobachtungen.

Die Verwendung der F-Statistik, um einen Zusammenhang zwischen erklärenden Variablen und Zielgrösse zu testen, funktioniert, wenn p eher klein ist oder zumindest

klein verglichen mit n. Manchmal haben wir allerdings eine sehr grosse Zahl von erklärenden Variablen: Für $p > n$, also wenn wir mehr Variablen als Beobachtungen haben, dann gibt es mehr β_j als Beobachtungen, aus welchen wir schätzen. In diesem Fall können wir unsere Methoden gar nicht anwenden.

Beispiel 8.3.8

Wir sehen in der folgenden **R**-Simulation den eben besprochenen Effekt.

```
set.seed(4)
v <- 20
d <- 500

df <- matrix(rnorm(n = v * d), nrow = d)

df <- data.frame(df)

Y <- rnorm(n = d)
# Y

df$Y <- Y
head(round(df, 4), 3)


##        X1      X2      X3      X4      X5      X6     X7     X8
## 1  0.2168 -1.6447  0.1766 -0.7407 -1.3944  0.6861 0.8565 0.8157
## 2 -0.5425 -0.8200  1.6890 -0.1870  0.0513 -0.7716 0.1272 0.2769
## 3  0.8911 -1.6782 -1.3473 -0.1343  1.0970  0.4995 0.5711 1.6146
##        X9     X10     X11     X12     X13     X14     X15
## 1 -0.0817 -0.5288  1.1722  0.8675  0.5075  1.0932 -0.7228
## 2 -0.8546 -1.4892  0.4583 -0.5210 -0.3684 -1.1823 -2.7069
## 3 -0.9436  1.0831 -0.7720 -0.2192  0.1560  0.4964 -1.3957
##       X16     X17     X18     X19     X20       Y
## 1  0.8242 -1.0437  0.9288  0.0014 -1.5083 -0.6010
## 2  2.1106  1.0070 -0.0096 -1.1340  1.3256  1.9657
## 3 -0.2661 -1.4833  0.6595  2.5946  0.5704 -0.3115
```

Wir sehen hier ein Dataframe **df**, dass 21 Spalten enthält, 20 Spalten für die Prädiktoren **X1**, **X2**, ..., **X20** und die Zielvariable **Y**. Alle Spalten enthalten normalverteilte Zufallszahlen mit $\mu = 0$ und $\sigma = 1$ (**rnorm(n=...)**). Das heisst, es gilt $\mu_1 = 0$, $\mu_2 = 0$, ..., $\mu_{20} = 0$.

Nun machen wir einen Hypothesentest und betrachten zuerst die p-Werte der einzelnen Variablen.

```
fit <- lm(Y ~ ., data = df)
summary(fit)
```

```
##
## Call:
## lm(formula = Y ~ ., data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.62976 -0.66857  0.00927  0.64462  2.81840
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.029669   0.047272  -0.628   0.5305
## X1          -0.010970   0.048886  -0.224   0.8225
## X2          -0.036943   0.049150  -0.752   0.4526
## X3          -0.005961   0.047734  -0.125   0.9007
## X4          -0.018073   0.047726  -0.379   0.7051
## X5           0.005827   0.048524   0.120   0.9045
## X6          -0.127798   0.049554  -2.579   0.0102 *
## X7          -0.052386   0.049816  -1.052   0.2935
## X8           0.020574   0.048557   0.424   0.6720
## X9          -0.015178   0.047941  -0.317   0.7517
## X10         -0.015107   0.046988  -0.322   0.7480
## X11          0.005580   0.046517   0.120   0.9046
## X12         -0.004676   0.046583  -0.100   0.9201
## X13         -0.021652   0.049114  -0.441   0.6595
## X14         -0.093800   0.046075  -2.036   0.0423 *
## X15          0.019740   0.047451   0.416   0.6776
## X16          0.042796   0.045267   0.945   0.3449
## X17         -0.074511   0.049061  -1.519   0.1295
## X18          0.041733   0.047568   0.877   0.3808
## X19         -0.078238   0.047492  -1.647   0.1001
## X20         -0.057475   0.048156  -1.194   0.2333
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.042 on 479 degrees of freedom
## Multiple R-squared:  0.04522,Adjusted R-squared:  0.005351
## F-statistic: 1.134 on 20 and 479 DF,  p-value: 0.31
```

Wir sehen, dass die p-Werte für **`X6`** und **`X14`** sind statistisch signifikant. Das heisst, dass die geschätzten Werte für $\widehat{\beta}_6$ und $\widehat{\beta}_{14}$ weichen von 0 ab, obwohl in Wahrheit $\beta_6 = 0$ und $\beta_{14} = 0$. Hier wird also eine Abhängigkeit von **`Y`** von **`X6`** und **`X14`** angezeigt, obwohl diese gar nicht vorhanden ist.

Der p-Wert der F-Statistik ist allerdings mit 0.31 nicht signifikant, das heisst es gibt keine Abhängigkeit der Prädiktoren auf die Zielvariable. ◀

Bestimmung der wichtigen erklärenden Variablen

Im letzten Abschnitt haben wir gesehen, dass wir zuerst entscheiden müssen, ob die erklärenden Variablen überhaupt einen Einfluss auf die Zielgrösse haben. Dies geschieht mit Hilfe der F-Statistik und dem zugehörigen p-Wert.

Haben wir auf dieser Basis festgestellt, dass mindestens eine Variable die Zielgrösse beeinflusst, so ist natürlich die nächste Frage, *welche* erklärende Variablen dies sind. Wir können wieder die einzelnen p-Werte wie in Tabelle 8.3 auf Seite 276 betrachten. Aber wir haben gerade gesehen, dass dies bei grossen p zu falschen Schlussfolgerungen führen kann.

Es ist möglich, dass alle erklärenden Variablen die Zielgrösse beeinflussen, aber meist sind es nur einige wenige. Die Aufgabe besteht nun darin, diese Variablen zu bestimmen und dann ein Modell aufzustellen, welches nur diese Variablen enthält. Wir sind an einem möglichst einfachen Modell interessiert, das zu den Daten passt. Dieses Prozedere wird mit *Variablenselektion* bezeichnet. Dies wird in Kapitel 9 ausführlich behandelt.

Wie gut passt das Modell zu den Daten?

Wie gut passt das Modell zu den Daten? Zwei geläufige numerische Grössen, um den Grad der Anpassung des Modells an die Daten zu messen, sind der RSE und das Bestimmtheitsmass R^2 (siehe Abschnitt **??** auf Seite ??). Diese Grössen werden wir nun berechnen und interpretieren, wie wir dies bereits bei der einfachen linearen Regression getan haben.

Ein R^2-Wert nahe bei 1 deutet darauf hin, dass ein grosser Anteil der Varianz in der Zielvariablen durch das Modell erklärt werden kann.

Beispiel 8.3.9

Im **R**-Output in Beispiel 8.3.6 für den Datensatz **Werbung** ist der R^2-Wert 0.8972 aufgeführt. Auf der anderen Seite ist für das Modell, das nur **TV** und **Radio** für den **Verkauf** berücksichtigt, $R^2 = 0.89719$. Das heisst, wir haben einen *sehr kleinen* Zuwachs im Bestimmtheitsmass, wenn wir im Modell **Zeitung** auch noch mitberücksichtigen. Diese Beobachtung deckt sich mit der Feststellung, dass **Zeitung** auf den **Verkauf** statistisch keinen signifikanten Einfluss hat (siehe Tabelle 8.3).

Es zeigt sich, dass sich R^2 immer erhöht, wenn mehr Variablen zum Modell hinzugefügt werden, sogar wenn diese keinen oder kaum Einfluss auf die Zielgrösse haben.

Der Grund dafür ist, dass mit einer weiteren Variable durch die Methode der kleinsten Quadrate die Anpassung an die Daten genauer wird. Damit nimmt R^2 zu, da dieser Wert auf den Daten beruht.

Die Hinzunahme von **Zeitung** zum Modell mit **TV** und **Radio** bewirkt nur eine winzige Zunahme des Wertes von R^2. Dies ist ein weiterer Hinweis dafür, dass für den **Verkauf** die erklärende Variable **Zeitung** kaum eine Rolle spielt und dass wir diese Variable für unser Modell nicht unbedingt berücksichtigen müssen.

Im Gegensatz dazu ist $R^2 = 0.61$ (siehe Tabelle 8.2 oben), wenn wir nur **TV** als erklärende Variable berücksichtigen. Die Hinzunahme von **Radio** zum Modell bewirkt eine deutlich Zunahme von R^2 auf 0.8972. Dies impliziert, dass die Vorhersage von **Verkauf** viel besser durch die erklärenden Variablen **TV** und **Radio** zusammen als durch **TV** alleine beschrieben wird. Wir können dies noch quantifizieren, indem wir den p-Wert von **Radio** im Modell mit **TV** und **Radio** betrachten. ◀

Zusätzlich zum R^2 können wir, wenn möglich, die Daten plotten. Ein graphischer Überblick kann Probleme mit dem Modell aufzeigen, die für die numerischen Werte unsichtbar sind.

Beispiel 8.3.10

In Abbildung 8.18 sehen wir ein dreidimensionales Streudiagramm, in welchem nur **TV** und **Radio** als erklärende Variablen berücksichtigt sind. Gestrichelt ist die Regressionsebene eingezeichnet. Wir beoabachten, dass die Werte der Ebene zu gross sind, wenn die Werbeausgaben ausschliesslich entweder für **TV** oder **Radio** aufgewendet wurden. Hinten links wurde nur Werbung für **Radio** gemacht und vorne rechts nur für **TV**.

Die Werte der Ebene sind hingegen zu tief, wenn die Werbeausgaben gleichmässig auf **TV** und **Radio** verteilt werden.

Dieses nichtlineare Muster kann nicht genau durch eine lineare Regression beschrieben werden. Der Plot deutet auf einen *Interaktion-* oder *Synergieeffekt* hin, der in grösseren Verkäufen mündet, wenn wir die Werbeausgaben aufteilen. Wir werden dies später noch genauer betrachten.

◀

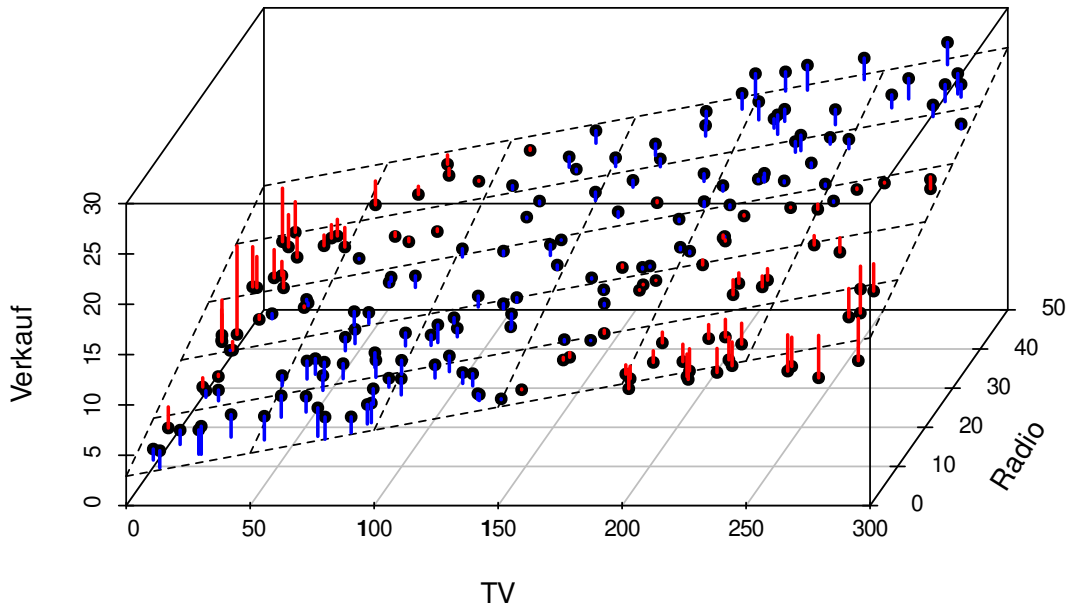


Abbildung 8.18. : Regressionsebene für `TV` und `Radio` als erklärende Variablen.

Vorhersagen

Haben wir einmal das multiple lineare Regressionsmodell aufgestellt, so können wir dies auch einfach anwenden, um die Zielgrösse für Y auf der Basis von $X_1, X_2, \ldots, X_p$ vorherzusagen. Allerdings gibt es drei Arten von *Ungewissheiten*, die mit dieser Vorhersage zusammenhängen.

1. Die Koeffizientenschätzungen $\widehat{\beta}_1, \ldots, \widehat{\beta}_p$ sind nur Abschätzungen für $\beta_1, \ldots, \beta_p$. Das heisst, die Regressionsebene

   $$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \ldots + \widehat{\beta}_p X_p$$

   die wir mit der Methode der kleinsten Quadrate erhalten haben, ist nur eine Abschätzung der *wahren*, aber unbekannten Regressionsebene

   $$f(X_1, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

   Diese Ungenauigkeit hängt mit dem *reduziblen Fehler* aus Kapitel 8 zusammen. Wir können das *Vertrauensintervall* berechnen, um zu bestimmen, wie nahe $\widehat{Y}$

bei $f(X_1, \ldots, X_p)$ liegt.

2. Die Annahme, dass $f(X_1, \ldots, X_p)$ durch ein lineares Modell beschrieben wird, ist beinahe immer nur eine Approximation der Realität, was zu einem weiteren potentiell reduziblen Fehler führt. Dieser wird *Model Bias* genannt.

   Wenn wir also ein lineares Modell verwenden, so passen wir die beste lineare Approximation an die wahre Oberfläche. Wir werden diesen Unterschied ignorieren und so tun, als ob das lineare Modell korrekt ist.

3. Sogar wenn wir $f(X_1, \ldots, X_p)$ kennen, das heisst, sogar wenn die wahren Werte für $\beta_1, \ldots, \beta_p$ bekannt sind, kann die Zielgrösse aufgrund des Fehlerterms ε nicht genau vorhergesagt werden. In Kapitel 8 haben wir dies den *irreduziblen* Fehler genannt.

   Wie stark wird Y von $\widehat{Y}$ abweichen? Die Antwort dazu gibt uns das *Prognoseintervall*. Prognoseintervalle sind immer breiter als Vertrauensintervalle, da diese sowohl den Fehler in der Schätzung von $f(X_1, \ldots, X_p)$ (reduzibler Fehler) als auch die Unsicherheit, wie weit ein einzelner Punkt von der wahren Regressionsebene abweicht (irreduzibler Fehler), beinhaltet.

Beispiel 8.3.11

Wir verwenden das *Vertrauensintervall*, um die Ungewissheit für den *durchschnittlichen* **Verkauf** für eine grosse Zahl von Städten zu quantifizieren. Dabei berücksichtigen wir nur die erklärenden Variablen **TV** und **Radio**, da **Zeitung** für **Verkauf** keinen Einfluss hat.

Wenden wir CHF 100 000 für **TV**-Werbung und CHF 20 000 für **Radio**-Werbung in jeder Stadt auf, so ist das 95 %-Vertrauensintervall

$$[10'985, 11'528]$$

```
Werbung <- read.csv("../Daten/Werbung.csv")
TV <- Werbung[, 2]
Radio <- Werbung[, 3]
Verkauf <- Werbung[, 5]
predict(lm(Verkauf ~ TV + Radio), interval = "confidence", data.frame(TV = 100,
    Radio = 20))


##        fit      lwr      upr
## 1 11.25647 10.98525 11.52768
```

Die Interpretation lautet wie folgt: 95 % aller Intervalle dieser Form enthalten den wahren Wert $f(X_1, X_2)$. Was heisst dies? Sammeln wir eine grosse Menge von Datensätzen wie den **Werbung**-Datensatz, dann können wir für jeden Datensatz jeweils das Vertrauensintervall für den wahren durchschnittlichen **Verkauf** berechnen (bei CHF 100 000 für **TV**-Werbung und CHF 20 000 für **Radio**-Werbung). Dann liegt in 95 % dieser Intervalle der wahre Wert vom durchschnittlichen **Verkauf**.

Für gegebene CHF 100 000 für **TV**-Werbung und CHF 20 000 für **Radio**-Werbung ist in einer *bestimmten* Stadt der *Prognosebereich*

$$[7'930, 14'583]$$

```
Werbung <- read.csv("../Daten/Werbung.csv")
TV <- Werbung[, 2]
Radio <- Werbung[, 3]
Verkauf <- Werbung[, 5]
predict(lm(Verkauf ~ TV + Radio), interval = "prediction", data.frame(TV = 100,
    Radio = 20))


##        fit      lwr      upr
## 1 11.25647 7.929616 14.58332
```

Wir interpretieren dies so, dass in 95 % von allen Intervallen dieser Form eine Beobachtung Y für diese bestimmte Stadt liegt.

Beide Intervalle liegen um 11′256, aber der Prognosebereich ist wesentlich breiter als das Vertrauensintervall. Dies spiegelt die grössere Unsicherheit in Bezug auf **Verkauf** in einer bestimmten Stadt gegenüber dem durchschnittlichen **Verkauf** über viele Städte wider. ◀

8.3.4. Erweiterungen des linearen Modells

Mit dem linearen Standardregressionsmodell

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

haben wir ein Modell, das gut interpretierbare Resultate liefert und für viele reale Probleme ziemlich gut funktioniert. Auf der anderen Seite gehen mit diesem Modell sehr restriktive Einschränkungen einher, die in der Praxis oft nicht gerechtfertigt sind.

Die Beziehung zwischen den erklärenden Variablen und der Zielgrösse beinhaltet im Standardregressionsmodell einschneidende Einschränkungen. Die beiden wichtigsten sind:

1. *Additivität*

   Die Annahme bezüglich Additivität hat zur Folge, dass der Effekt der Änderung, den eine erklärende Variable X_j auf die Zielgrösse Y hat, unabhängig von den Werten der anderen erklärenden Variablen ist.

2. *Linearität*

   Die Annahme bezüglich Linearität besagt, dass bei einer Änderung von X_j um eine Einheit die Änderung in der Zielgrösse Y konstant ist, und zwar unabhängig vom Wert von X_j.

Wir werden im Folgenden einige klassische Ansätze zur Erweiterung des klassischen linearen Ansatzes untersuchen.

Aufhebung der Annahme bezüglich Additivität

Beispiel 8.3.12

In unserer Analyse des Datensatzes **Werbung** sind wir zum Schluss gekommen, dass die erklärenden Zufallsvariablen **TV** und **Werbung** mit der Zielvariablen **Verkauf** zusammenhängen. Das lineare Modell

$$\texttt{Verkauf} = \beta_0 + \beta_1 \cdot \texttt{TV} + \beta_2 \cdot \texttt{Radio} + \varepsilon$$

das die Basis für einige zuvor gezogene Schlussfolgerungen bildete, geht davon aus, dass eine Zunahme von **Verkauf** aufgrund einer Vergrösserung der Werbeausgaben für ein Medium unabhängig von den Werbeausgaben ist, die für das andere Medium aufgewendet wurden. Beispielsweise besagt unser Modell, dass die durchschnittliche Änderung von **Verkauf** immer β_1 ist, wenn wir **TV** um eine Einheit erhöhen, unabhängig davon, wieviel wir für **Radio** investiert haben.

Dass diese Annahme problematisch ist, erkennen wir, wenn wir die Möglichkeit in Betracht ziehen, dass eine Investition in Radiowerbung die Wirkung von TV-Werbung erhöht. Dann müsste mit zunehmenden Werbeausgaben für **Radio** auch die Steigung für **TV** zunehmen. Dieser Effekt wird durch unser Modell nicht beschrieben, da es annimmt, dass die Steigungen konstant bleiben. Gehen wir in dieser Situation von einem festen Budget von CHF 100 000 aus, wobei wir dies je zur Hälfte für **TV** und **Radio** aufwenden. Dann vergrössert sich **Verkauf** *mehr*, als wenn wir den ganzen Betrag für das eine oder andere Medium aufwenden. In Marketing spricht man von einem *Synergieeffekt*, in der Statistik von einem *Interaktionseffekt*.

Abbildung 8.18 auf Seite 286 deutet an, dass beim Datensatz **Werbung** ein solcher Effekt vorzuliegen scheint. Wenn die Ausgaben für **TV** oder für **Radio** tief sind, so ist

Verkauf tiefer, als das lineare Modell vorhergesagt hat. Werden die Werbeausgaben aber aufgeteilt, so ist **Verkauf** grösser als beim linearen Modell. ◀

Wir betrachten das lineare Standardregressionsmodell mit zwei Variablen

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Wenn wir hier X_1 um eine Einheit erhöhen, so vergrössert sich Y im Durchschnitt um β_1. Die Existenz von X_2 hat auf diese Aussage keinen Einfluss. Wie gross auch X_2 ist, die Zunahme von Y ist durch β_1 gegeben, wenn X_1 um eine Einheit vergrössert wird.

Eine Möglichkeit, dieses zu Modell zu erweitern, liegt darin, den Interaktionseffekt mitzuberücksichtigen. Dazu führen wir eine weitere erklärende Variable ein, den *Interaktionsterm*. Dieser besteht aus dem Produkt von X_1 und X_2. Dies resultiert in folgendem Modell

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

Wie lockert die Hinzunahme des Interaktionstermes die Annahme bezüglich Linearität auf? Wir können die Gleichung oben folgendermassen umschreiben

$$\begin{aligned} Y &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \varepsilon \\ &= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \varepsilon \end{aligned}$$

mit

$$\tilde{\beta}_1 = \beta_1 + \beta_3 X_2$$

Da $\tilde{\beta}_1$ mit X_2 ändert, ist die Wirkung von X_1 auf Y nicht mehr konstant. Eine Anpassung von X_2 wird die Wirkung von X_1 auf Y ändern.

Beispiel 8.3.13

Wir sind an der Produktivität einer Fabrik interessiert. Dazu möchten wir die Anzahl produzierter **Einheiten** aufgrund der Produktionslinien **Linien** und der gesamten Anzahl **Arbeiter** vorhersagen. Es ist wahrscheinlich, dass die Wirkung der Vergrösserung der Anzahl der Produktionslinien auch von der Anzahl Arbeiter abhängt. Hat es keine Arbeiter, so werden mehr Produktionslinien keinen Einfluss auf die produzierten Einheiten haben. Damit scheint es angebracht zu sein, einen Interaktionsterm zwischen **Linien** und **Arbeitern** in das lineare Modell für **Einheiten** hinzuzunehmen. Ein solches Modell könnte wie folgt aussehen

$$\begin{aligned} \textbf{Einheiten} &\approx 1.2 + 3.4 \cdot \textbf{Linien} + 0.22 \cdot \textbf{Arbeiter} + 1.4 \cdot (\textbf{Linien} \cdot \textbf{Arbeiter}) \\ &= 1.2 + (3.4 + 1.4 \cdot \textbf{Arbeiter}) \cdot \textbf{Linien} + 0.22 \cdot \textbf{Arbeiter} \end{aligned}$$

Haben wir eine zusätzliche Produktionslinie zur Verfügung wird die Produktion um $3.4 + 1.4 \cdot \texttt{Arbeiter}$ Einheiten erhöht. Je mehr Arbeiter wir haben, umso grösser wird die Wirkung von **Linien** auf **Einheiten** sein.

Bemerkungen:

i. Das Modell ist bestimmten Einschränkungen unterworfen. Nehmen wir an, wir hätten keine Arbeiter, dann würde die Fabrik immer noch

$$\texttt{Einheiten} \approx 1.2 + 3.4 \cdot \texttt{Linien}$$

Einheiten produzieren. Das Problem liegt hier in der *Extrapolation*. Wir machen hier Vorhersagen in einem Bereich, für das das Modell nicht geeignet ist. Die Annahme, dass keine Arbeiter vorhanden sind, ist an einem gewöhnlichen Arbeitstag unrealistisch. ♦

◀

Beispiel 8.3.14

Wir kehren zum Beispiel **Werbung** zurück. Ein lineares Modell, das **TV** und **Radio** zusammen mit einem Interaktionsterm verwendet, um **Verkauf** vorherzusagen, hat die Form

$$\begin{aligned}\texttt{Verkauf} &= \beta_0 + \beta_1 \cdot \texttt{TV} + \beta_2 \cdot \texttt{Radio} + \beta_3 \cdot (\texttt{TV} \cdot \texttt{Radio}) + \varepsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \cdot \texttt{Radio}) \cdot \texttt{TV} + \beta_2 \cdot \texttt{Radio} + \varepsilon\end{aligned}$$

Wir können β_3 als Zunahme der Wirkung der TV-Werbung bei einer Zunahme der Radiowerbung um eine Einheit interpretieren. Die Koeffizienten für dieses Modell sind in der folgenden **R**-Ausgabe aufgeführt.

```
Werbung <- read.csv("../Daten/Werbung.csv")
TV <- Werbung[, 2]
Radio <- Werbung[, 3]
Zeitung <- Werbung[, 4]
Verkauf <- Werbung[, 5]
summary(lm(Verkauf ~ TV + Radio + TV * Radio))


##
## Call:
## lm(formula = Verkauf ~ TV + Radio + TV * Radio)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -6.3366 -0.4028  0.1831  0.5948  1.5246
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.750e+00  2.479e-01  27.233   <2e-16 ***
## TV          1.910e-02  1.504e-03  12.699   <2e-16 ***
## Radio       2.886e-02  8.905e-03   3.241   0.0014 **
## TV:Radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9435 on 196 degrees of freedom
## Multiple R-squared:  0.9678,Adjusted R-squared:  0.9673
## F-statistic:  1963 on 3 and 196 DF,  p-value: < 2.2e-16
```

Die Resultate deuten stark darauf hin, dass dieses Modell dem Modell, welches nur die *Haupteffekte* beinhaltet, überlegen ist. Der p-Wert für **TV** $\cdot$ **Radio** ist extrem klein, was stark darauf hinweist, dass $H_A: \beta_3 \neq 0$ gilt. Der wahre Zusammenhang ist also nicht additiv.

Der R^2-Wert ist 0.968 verglichen mit 0.897 für das Modell, welches nur **TV** und **Radio** ohne Interaktionsterm für die Vorhersage von **Verkauf** verwendet. Das bedeutet, dass

$$\frac{0.968 - 0.897}{1 - 0.897} = 0.69 = 69\,\%$$

der Variabilität von **Verkauf**, die übrigbleibt, wenn wir das additive Modell anpassen, durch den Interaktionsterm erklärt werden kann.

Die Koeffizientenschätzungen in der **R**-Ausgabe oben deutet an, dass eine Zunahme von CHF 1000 der TV-Werbung mit der Zunahme von

$$(\widehat{\beta}_1 + \widehat{\beta}_3 \cdot \textbf{Radio}) \cdot 1.000 = 1.9 + 1.1 \cdot \textbf{Radio}$$

Einheiten von **Verkauf** zusammenhängt.

Eine Zunahme der Radiowerbung um CHF 1000 ist verknüpft mit der Zunahme von

$$(\widehat{\beta}_2 + \widehat{\beta}_3 \cdot \textbf{TV}) \cdot 1.000 = 29 + 1.1 \cdot \textbf{TV}$$

Einheiten von **Verkauf**. ◀

In diesem Beispiel sind die p-Werte, die zu **TV**, **Radio** und dem Interaktionsterm **TV** $\cdot$ **Radio** gehören, statistisch signifikant. Somit scheint es klar zu sein, dass alle diese Variablen im Modell enthalten sein sollten. Es kann allerdings der Fall eintreten, dass

der p-Wert für den Interaktionterm sehr klein ist, aber die p-Werte der Haupteffekte (hier **TV** und **Radio**) sind es nicht.

Das *hierarchische Prinzip* besagt in einem solchen Fall, dass falls ein Interaktionsterm im Modell vorhanden ist, dass dann auch die Haupfeffekte vorhanden sein sollten, auch wenn ihre p-Werte nicht klein sind. Scheint also die Interaktion von X_1 und X_2 wichtig zu sein, dann sollten wir im Modell X_1 und X_2 ebenfalls berücksichtigen, sogar wenn deren Koeffizientenschätzungen sehr grosse p-Werte ergeben.

Die Idee hinter diesem hierarchischen Prinzip ist, dass falls $X_1 \cdot X_2$ mit der Zielgrösse zusammenhängt, es keine Rolle spielt, ob die Koeffizienten von X_1 oder X_2 exakt 0 sind oder nicht. Ebenso korreliert $X_1 \cdot X_2$ mit X_1 und X_2 und damit würde ein Weglassen einer oder beider Variablen die Bedeutung der Interaktion ändern.

Beispiel 8.3.15

Im vorhergehenden Beispiel haben wir die Interaktion zwischen den quantitativen Grössen **TV** und **Radio** betrachtet. Das Konzept der Interaktion können wir aber auch auf qualitative oder auf eine Kombination von quantitativen und qualitativen Grössen anwenden. Die Interaktion zwischen einer qualitativen und einer quantitativen Grösse hat eine besonders nette Interpretation. ◀

8.4. Qualitative erklärende Variablen

In unseren Betrachtungen haben wir bis anhin angenommen, dass alle Variablen in unserem linearen Regressionssystem *quantitativ* sind. Das muss in der Praxis nicht unbedingt der Fall sein, denn oft sind einige der erklärenden Variablen *qualitativ*.

Beispiel 8.4.1

Der Datensatz **Credit** wurde in den USA erhoben. Er führt für eine grössere Anzahl Individuen die Zielgrösse **balance** (monatliche Kreditkartenrechnung) wie auch mehrere quantitative erklärende Variablen auf: **age** (Alter), **cards** (Anzahl Kreditkarten), **education** (Anzahl Jahre Ausbildung), **income** (Einkommen in Tausenden Dollars), **limit** (Kreditkartenlimite) und **rating** (Kreditwürdigkeit).

In Abbildung 8.19 sind die Streudiagramme von Paaren von Variablen aufgeführt, deren Identität durch die entsprechenden Spalten- und Zeilenkennzeichnungen gegeben sind. So ist zum Beispiel der Plot direkt rechts des Wortes „Balance" das Streudiagramm der Variablen **age** und **balance**.

```
Credit <- read.csv("../Daten/Credit.csv")
pairs(~Balance + Age + Cards + Education + Income + Limit + Rating,
    Credit, pch = ".", col = "darkcyan")
```

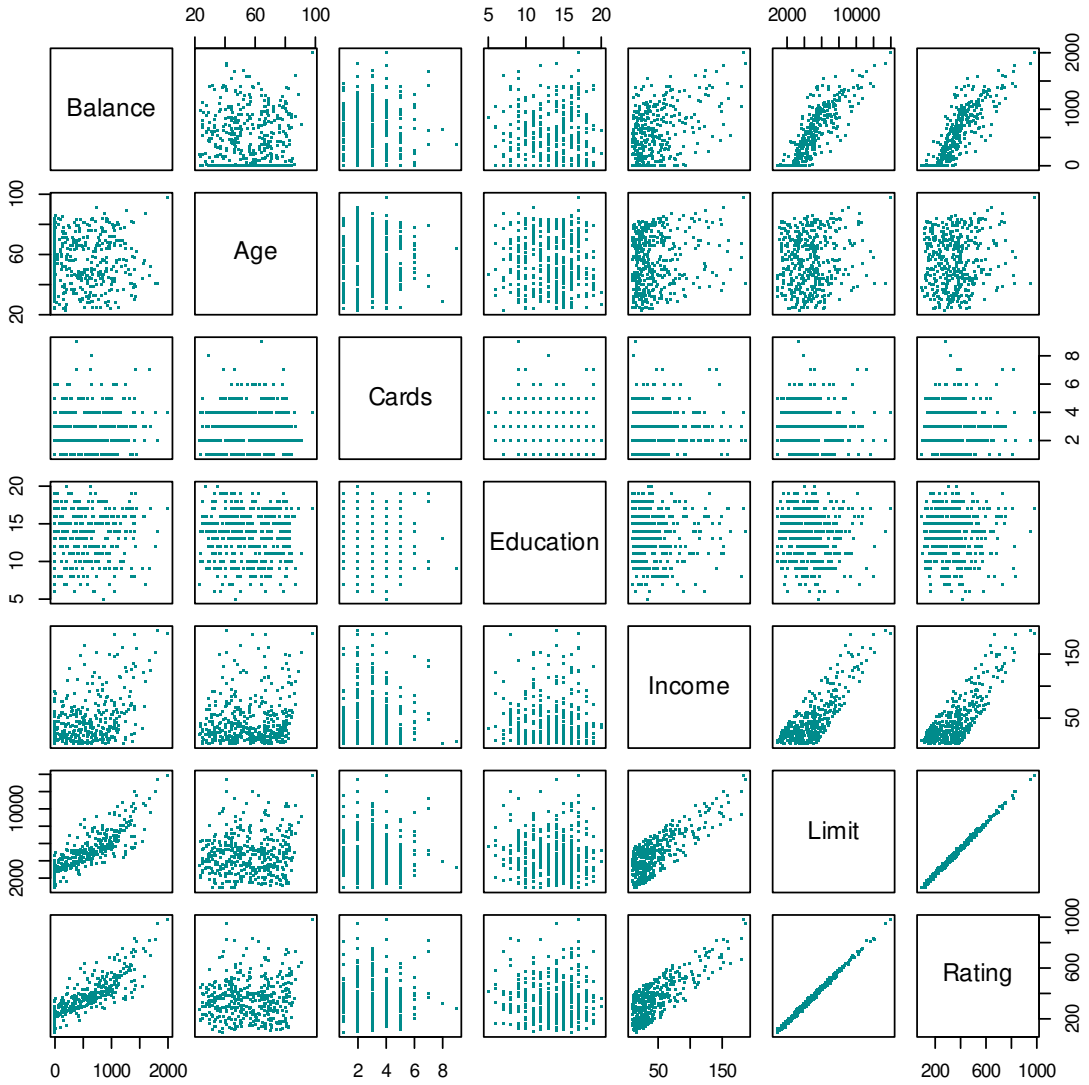


Abbildung 8.19. : Streudiagramme aus dem `Credit`-Datensatz.

Neben diesen quantitativen erklärenden Variablen haben wir noch vier qualitative Variablen: **`gender`** (Geschlecht), **`student`** (Studentenstatus) und **`ethnicity`** (Kaukasier, Afroamerikaner, Asiat). ◀

Qualitative erklärende Variable mit nur zwei Levels

Angenommen wir wollen im vorangehenden Bespiel die Kreditkartenrechnungen (**balance**) zwischen Männern und Frauen untersuchen und die anderen Variablen für den Moment ignorieren. Falls eine qualitative erklärende Variable (auch *Faktor* genannt) nur zwei *Levels* oder mögliche Werte hat, dann ist die Hinzunahme dieser Variable in das Regressionsmodell sehr einfach.

Dazu führen wir einfach eine Indikatorvariable (oder *Dummy-Variable*) ein, die nur zwei mögliche numerische Werte annehmen kann.

Beispiel 8.4.2

Für die Geschlechtervariable **gender** wählen wir beispielsweise eine neue Variable, die folgende Form hat:

$$x_i = \begin{cases} 1 & \text{falls } i\text{-te Person weiblich} \\ 0 & \text{falls } i\text{-te Person männlich} \end{cases}$$

Wir verwenden diese Variable nun als erklärende Variable im Regressionsmodell. Dies resultiert in folgendem Modell

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{falls } i\text{-te Person weiblich} \\ \beta_0 + \varepsilon_1 & \text{falls } i\text{-te Person männlich} \end{cases}$$

Nun interpretieren wir β_0 als die durchschnittlichen Kreditkartenrechnungen unter Männern und $\beta_0 + \beta_1$ als die durchschnittlichen Kreditkartenrechnungen unter Frauen. Die Grösse β_1 alleine entspricht dann dem durchschnittlichen Unterschied der Rechnungen zwischen Männern und Frauen.

Tabelle 8.5 führt unter anderem die Koeffizientenschätzungen für unser Modell auf. Die geschätzten durchschnittlichen Rechnungen für Männer betragen dann \$ 509.80 und der geschätzte Unterschied zu Frauen beträgt \$ 19.73. Dies ergibt für Frauen ein Total von \$ 509.80 + \$ 19.73 = \$ 529.53.

Allerdings ist der p-Wert für die Indikatorvariable β_1 mit 0.6690 so hoch, dass wir nicht von einem statistisch signifikanten Unterschied zwischen Rechnungskosten von Frauen und Rechnungskosten von Männern sprechen können. ◀

Im vorangehenden Beispiel ist die Entscheidung, Frauen mit 1 und Männer mit 0 zu kodieren, völlig willkürlich und hat keinen Einfluss auf den Grad der Anpassung des Regressionsmodells an die Daten. Bei unterschiedlicher Kodierunge ist dann aber die Interpretation der Koeffizienten natürlich unterschiedlich.

```
Credit <- read.csv("../Daten/Credit.csv")
balance <- Credit[, 12]
gender <- Credit[, 8] == "Female"
round(summary(lm(balance ~ gender))$coef, digits = 5)

##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 509.80311   33.12808 15.38885  0.00000
## genderTRUE   19.73312   46.05121  0.42850  0.66852
```

	Koeffizient	Std.fehler	t-Statistik	p-Wert
Intercept	509.80	33.13	15.389	< 0.0001
gender[female]	19.73	46.05	0.429	0.6690

Tabelle 8.5. : Regression von **balance** auf **gender** mit einer Indikatorvariable

Falls wir die Männer mit 1 und die Frauen mit 0 kodieren, so wären die Schätzung für die Parameter β_0 und β_1 \$ 529.53, resp. \$ -19.73 . Dies entspricht wiederum Rechnungen von \$ 529.53 für Frauen und von \$ 529.73 $-$ \$ 19.73 $=$ \$ 509.80 für Männer. Dies ist dasselbe Resultat wie vorher.

Beispiel 8.4.3

Anstatt der 0/1-Kodierung können wir die Indikatorvariable

$$x_i = \begin{cases} 1 & \text{falls } i\text{-te Person weiblich} \\ -1 & \text{falls } i\text{-te Person männlich} \end{cases}$$

verwenden und in das Regressionsmodell aufnehmen. Dieses lautet dann

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{falls } i\text{-te Person weiblich} \\ \beta_0 - \beta_1 + \varepsilon_1 & \text{falls } i\text{-te Person männlich} \end{cases}$$

In diesem Fall kann β_0 als die durchschnittliche Summe der Kreditkartenrechnungen (ohne Berücksichtigung des Geschlechts) interpretiert werden. Die Variable β_1 entspricht dann dem Wert, mit welchem Frauen über dem Durchschnitt liegen und mit welchem Männer unter dem Durchschnitt liegen.

In diesem Fall wird β_0 durch \$ 519.665 geschätzt, was eine durchschnittliche Summe der Kreditkartenrechnungen von \$ 509.80 für Männer und von \$ 529.53 für Frauen ergibt. Die Schätzung für β_1 ist \$ 9.865, was der Hälfte vom Unterschied \$ 19.73 zwischen Männern und Frauen entspricht. ◀

Es sei hier nochmals mit Nachdruck erwähnt, dass die Vorhersagen für die Zielgrösse *nicht* von der Kodierung abhängt. Der einzige Unterschied liegt in der Interpretation

der Koeffizienten.

Qualitative erklärende Variablen mit mehr als zwei Levels

Eine qualitative erklärende Variable kann mehr als zwei Levels haben. Dann reicht aber eine Indikatorvariable für alle möglichen Werte nicht. In dieser Situation fügen wir einfach eine zusätzliche Indikatorvariable hinzu.

Beispiel 8.4.4

Die Variable `ethnicity` hat *drei* mögliche Levels, und somit wählen wir *zwei* verschiedene Indikatorvariablen. Die erste Indikatorvariable wählen wir wie folgt

$$x_{i1} = \begin{cases} 1 & \text{falls } i\text{-te Person asiatisch} \\ 0 & \text{falls } i\text{-te Person nicht asiatisch} \end{cases}$$

und die zweite folgendermassen

$$x_{i2} = \begin{cases} 1 & \text{falls } i\text{-te Person kaukasisch} \\ 0 & \text{falls } i\text{-te Person nicht kaukasisch} \end{cases}$$

Beide Variablen können dann in die Regressionsgleichung aufgenommen werden, und wir erhalten das Modell

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{falls } i\text{-te Person asiatisch} \\ \beta_0 + \beta_2 + \varepsilon_i & \text{falls } i\text{-te Person kaukasisch} \\ \beta_0 + \varepsilon_i & \text{falls } i\text{-te Person afroamerikanisch} \end{cases}$$

Dann beinhaltet β_0 die Information zu den durchschnittlichen Kreditkartenrechnungen von Afroamerikanern. Der Wert β_1 ist die Differenz zwischen den durchschnittlichen Rechnungen von Afroamerikanern und Asiaten und entsprechend β_2 die Differenz zwischen den durchschnittlichen Rechnungen von Afroamerikanern und Kaukasiern.

Bemerkungen:

i. Es gibt immer eine Indikatorvariable weniger, als es Levels hat.

ii. Der Level ohne Indikatorvariable, hier Afroamerikaner, heisst auch *Baseline*.

iii. Die Gleichung

$$y_i = \beta_0 + \beta_1 + \beta_2 + \varepsilon_i$$

macht in diesem Modell keinen Sinn, da dann eine Person asiatisch *und* kaukasisch sein müsste. ♦

Der Tabelle 8.6 entnehmen wir die geschätzte **balance** von $ 531.00 für die Baseline, also für Afroamerikaner.

```
Credit <- read.csv("../Daten/Credit.csv")
balance <- Credit[, 12]
ethnicity <- Credit[, "Ethnicity"]
summary(lm(balance ~ ethnicity))

##
## Call:
## lm(formula = balance ~ ethnicity)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -531.00 -457.08  -63.25  339.25 1480.50
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          531.00      46.32  11.464   <2e-16 ***
## ethnicityAsian       -18.69      65.02  -0.287    0.774
## ethnicityCaucasian   -12.50      56.68  -0.221    0.826
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 460.9 on 397 degrees of freedom
## Multiple R-squared:  0.0002188,Adjusted R-squared:  -0.004818
## F-statistic: 0.04344 on 2 and 397 DF,  p-value: 0.9575
```

	Koeffizient	Std.fehler	t-Statistik	p-Wert
Intercept	531.00	46.32	11.464	< 0.0001
ethnicity[Asian]	-18.69	65.02	-0.287	0.7740
ethnicity[Caucasian]	-12.50	56.68	-0.221	0.8260

Tabelle 8.6. : Regression mit **balance** als Zielgrösse und Faktorvariable **ethnicity** mit zwei Indikatorvariablen (3 Levels).

Die Schätzung für die Kategorie Asiaten ist $ -18.69, und somit sind deren durchschnittliche Rechnungen um diesen Betrag kleiner als die von Afroamerikanern. Die Kaukasier haben um durchschnittlich $ 12.50 kleinere Rechnungen als die Afroamerikaner. Aber auch hier sind die zugehörigen p-Werte so gross, dass wir von zufälligen Abweichungen ausgehen können. Es gibt also keinen signifikanten Unterschied bei den Kreditkartenrechnungen zwischen den Ethnien. Auch hier war der Level, den wir für die Baseline gewählt haben, willkürlich, und die Vorhersage der Zielvariable hängt nicht von der Kodierung ab.

Allerdings hängen die p-Werte von der Kodierung ab, und somit sind wir wieder besser beraten, die F-Statistik zu betrachten. Dazu führen wir wieder einen F-Test durch und testen

$$H_0: \quad \beta_1 = \beta_2 = 0$$

Der p-Wert dieser Statistik hängt *nicht* von der Kodierung ab. In diesem Beispiel beträgt er 0.96 und fällt damit relativ hoch aus, was unsere Vermutung von vorhin bestätigt, dass wir die Nullhypothese *nicht* verwerfen dürfen, und dass es keinen Zusammenhang zwischen **balance** und **ethnicity** gibt. ◀

Dieser auf Indikatorvariablen beruhende Ansatz ermöglicht es uns, auch qualitative und quantitative erklärende Variablen in unser Regressionsmodell zu integrieren. Wir können beispielsweise die Regression von **balance** mit der quantitativen erklärenden Variable **income** und der qualitativen erklärenden Variable **student** durchführen. Dazu kodieren wir **student** mit Hilfe einer Indikatorvariablen und führen wie im vorangehenden Abschnitt eine multiple lineare Regression durch.

Bemerkung:

Es gibt noch weitere Möglichkeiten, qualitative Variablen zu kodieren, als nur mit Hilfe von Indikatorvariablen. Wir gehen an dieser Stelle aber nicht darauf ein. ⧫

Beispiel 8.4.5

Im Datensatz **Credit** wollen wir die Zielgrösse **balance** durch die erklärenden Variablen **income** (quantitativ) und **student** (qualitativ) vorhersagen. Liegt kein Interaktionsterm vor, so hat das Modell die Form

$$\begin{aligned}\texttt{balance}_i &\approx \beta_0 + \beta_1 \cdot \texttt{income}_i + \begin{cases}\beta_2 & \text{falls } i\text{-te Person Student} \\ 0 & \text{falls } i\text{-te Person kein Student}\end{cases} \\ &= \beta_1 \cdot \texttt{income}_i + \begin{cases}\beta_0 + \beta_2 & \text{falls } i\text{-te Person Student} \\ \beta_0 & \text{falls } i\text{-te Person kein Student}\end{cases}\end{aligned}$$

Dieses Modell beschreibt zwei parallele Geraden, eine für Studenten und eine für Nichtstudenten. Die Steigung β_1 ist bei beiden gleich, aber die y-Achsenabschnitte sind verschieden ($\beta_0 + \beta_2$ und β_0). Das ist in Abbildung 8.20 links zu sehen.

Die durchschnittliche Zunahme von **balance** für eine Vergrösserung von **income** um eine Einheit hängt folglich nicht davon ab, ob das entsprechende Individuum ein Student ist oder nicht.

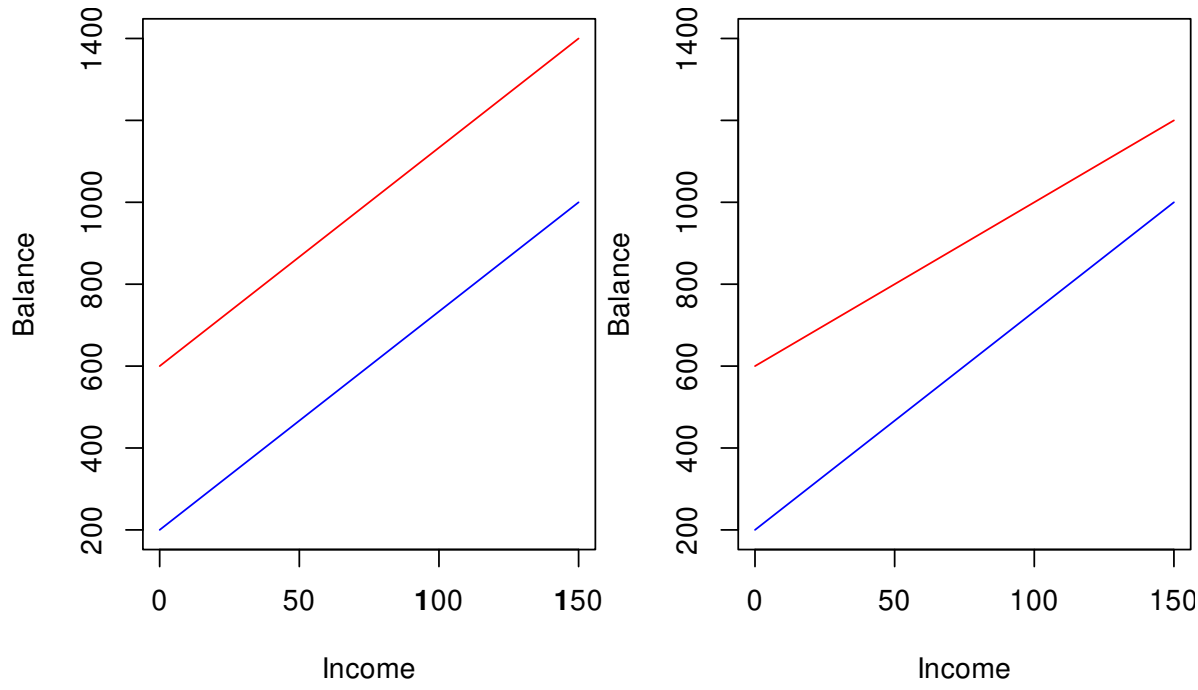


Abbildung 8.20. : Regression der Daten `Credit`. Die roten Geraden sind für Studenten und die blauen für Nichtstudenten.

Dies stellt möglicherweise eine schwerwiegende Einschränkung unseres Modells dar, da eine Änderung in **income** in der Tat eine unterschiedliche Wirkung auf die Kreditkartenrechnungen haben kann, je nachdem, ob jemand Student ist oder nicht. Diese Einschränkung kann wieder mit der Einführung einer Interaktionsvariablen gelockert werden. Dabei multiplizieren wir **income** mit der Indikatorvariablen für **student**. Unser Modell sieht dann wie folgt aus

$$\begin{aligned}\textbf{balance}_i &\approx \beta_0 + \beta_1 \cdot \textbf{income}_i + \begin{cases} \beta_2 + \beta_3 \cdot \textbf{income}_i & \text{falls Student} \\ 0 & \text{falls kein Student} \end{cases} \\ &= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \cdot \textbf{income}_i & \text{falls Student} \\ \beta_0 + \beta_1 \cdot \textbf{income}_i & \text{falls kein Student} \end{cases}\end{aligned}$$

Wir haben wiederum zwei unterschiedliche Regressionsgeraden für Studenten und Nichtstudenten. Aber jetzt haben wir verschiedene Steigungen $\beta_1 + \beta_3$ und β_1 zusätzlich zu den unterschiedlichen y-Achsenabschnitten $\beta_0 + \beta_2$ und β_0 (siehe Abbildung 8.20 rechts).

Dies eröffnet uns die Möglichkeit, die Änderung der Zielgrösse (Kreditkartenrechnungen) aufgrund der Änderungen im Einkommen für Studenten und Nichtstudenten getrennt zu betrachten. Auf der rechten Seite von Abbildung 8.20 sehen wir den geschätzten Zusammenhang zwischen **income** und **balance** für Studenten (rot)

und Nichtstudenten (blau). Die Steigung für Nichtstudenten ist grösser als für Studenten. Das deutet an, dass eine Zunahme im Einkommen eines Studenten eine kleinere Zunahme der Kreditkartenrechnungen zur Folge hat als für Nichtstudenten. ◄

Kapitel 9.

Variablenselektion

9.1. Einleitung

Das lineare Standardregressionsmodell

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p + \varepsilon$$

wird gewöhnlich zur Beschreibung des Zusammenhanges zwischen der Zielvariable Y und den erklärenden Variablen $X_1, X_2, \ldots, X_p$ verwendet. In Kapitel 8 haben wir für die Anpassung des Modells an die Daten die Methode der kleinsten Quadrate benutzt. Diese einfache Methode ist in einer Vielzahl von praktischen Aufgaben komplizierteren Methoden mindestens ebenbürtig, obwohl es mit relativ starken Einschränkungen einhergeht, wie mit der Voraussetzung bezüglich Linearität und Additivität, siehe Kapitel 8.3.4.

Warum sollten wir also noch andere Anpassungsmethoden untersuchen? Wie wir in diesem Kapitel sehen werden, können wir mit alternativen Methoden eine bessere *Vorhersagegenauigkeit* und *Modellinterpretierbarkeit* erreichen.

- *Vorhersagegenauigkeit*

  Ist der wahre Zusammenhang zwischen Zielgrösse Y und erklärenden Variablen $X_1, X_2, \ldots, X_p$ annähernd linear, dann ist die Varianz klein, wenn n sehr viel grösser als p ist. Die Vorhersagen sind dann auch entsprechend gut.

  Ist aber n nicht sehr viel grösser als p, dann wird die Variabilität bei der Methode der kleinsten Quadrate relativ gross, und dementsprechend ist die Vorhersagegenauigkeit schwach. Durch sogenanntes *Schrumpfen* (engl. *Shrinking*) der Koeffizienten kann die Varianz oft erheblich verkleinert werden, was die Vorhersagekraft erhöht.

- *Modellinterpretierbarkeit*

  Es ist oft der Fall, dass in einem multiplen Regressionsmodell einige Variablen keinen Zusammenhang mit der Zielvariable haben. Solche *irrelevanten* Variablen führen zu einer unnötigen Komplexität des resultierenden Modells. Also können wir diese Variable weglassen, wobei wir die entsprechenden Koeffizienten gleich null setzen. Dies führt zu einem Modell, das einfacher interpretierbar ist.

  Beispiel 9.1.1

  Für das Beispiel `Werbung` haben wir gesehen, dass die Variable `Zeitung` keinen Einfluss auf `Verkauf` hat. Zusätzlich zur Einfachheit kommt hier noch der kommerzielle Aspekt hinzu: Wenn die Zeitungswerbung schon nichts bringt, dann können wir auf diese verzichten und Geld einsparen. Haben wir durch Weglassen von `Zeitung` aber tatsächlich das beste Modell gefunden? Können wir noch eine weitere Variable weglassen? ◀

  Es ist aber höchst unwahrscheinlich, dass die Methode der kleinsten Quadrate Koeffizienten liefert, die exakt 0 sind. In diesem Kapitel werden wir Verfahren zur *Variablenselektion* kennenlernen, die automatisch irrelevante Variablen aus einem multiplen Regressionsmodell entfernen.

Im nächsten Kapitel werden wir uns ausführlich mit der Frage beschäftigen, *welche* erklärenden Variablen für die Anpassung an die Daten eine wesentliche Rolle spielen.

9.2. Variablenselektion

9.2.1. Schrittweise Vorwärtsselektion

Die *schrittweise Vorwärtsselektion* ist eine rechnerisch effiziente Methode, um Variablen zu eliminieren. Sie beginnt mit einem Modell, das gar keine erklärenden Variablen enthält. Dann wird schrittweise eine Variable um die andere zum Modell hinzugefügt, bis alle Variablen im Modell sind. Insbesondere wird in jedem Schritt jene Variable ins Modell aufgenommen, die die grösste *zusätzliche* Verbesserung der Anpassung mit sich bringt. Wir wollen das konkrete Vorgehen an einem Beispiel veranschaulichen.

Beispiel 9.2.1

Wir beginnen mit einem graphischen Beispiel und beschreiben die Vorwärtsselektion anhand Beispiel 7.1.2. Es kommen nur zwei Prädiktoren vor und somit können wir das Verfahren graphisch durchführen. In Abbildung 9.1 ist das dreidimensionale Streudiagramm nochmals dargestellt.

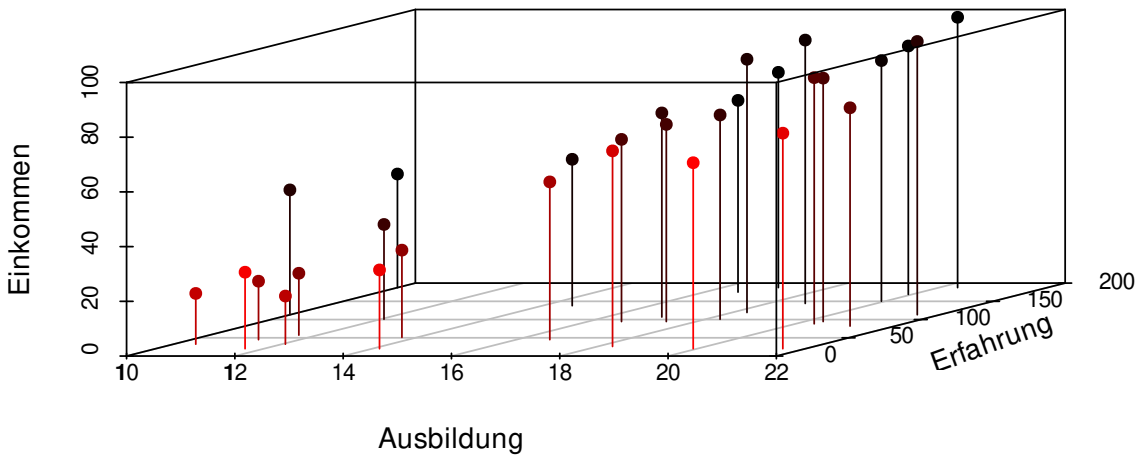


Abbildung 9.1. : Datenpunkte im Raum für den Datensatz `Einkommen`

In diese Streudiagramm legen wir nun eine Ebene, die *nicht* abhängig von den Prädiktoren ist. Das heisst,

$$\texttt{Einkommen} = \beta_0 + \varepsilon$$

Diese Ebene ist in Abbildung 9.2 dargestellt.

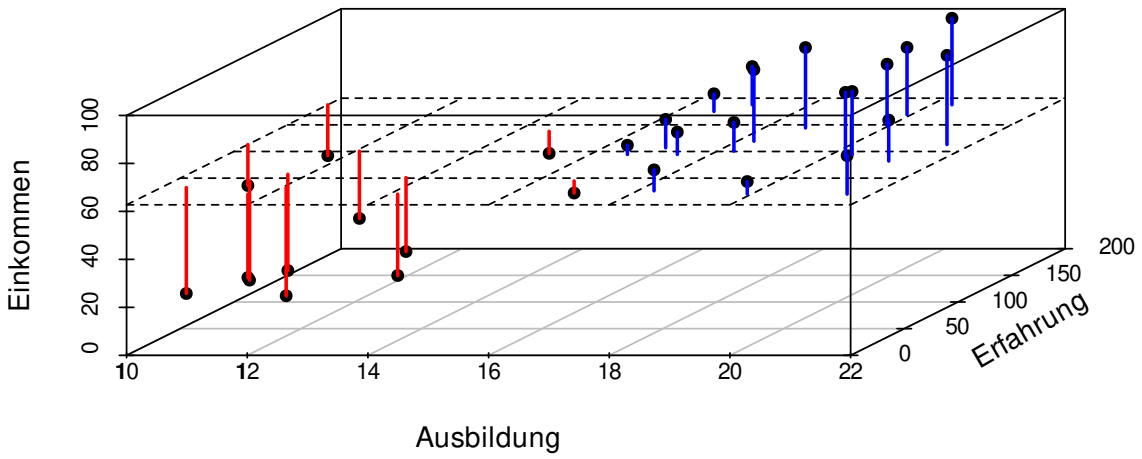


Abbildung 9.2. : Datensatz `Einkommen` ohne erklärende Variablen

Die gefärbten Linien sind wieder die Residuen, die vertikalen Abstände der Punkte zur Ebene. Die Ebene wurde wieder so gewählt, dass die Summe der Quadrate der Residuen (RSS) minimal ist.

Nun stellt sich die Frage, welche Variable wir *zuerst* hinzufügen. Die Idee ist nun, dass wir jeweils eine Variable hinzufügen und die beiden Modelle miteinander vergleichen (siehe Abbildung 9.3).

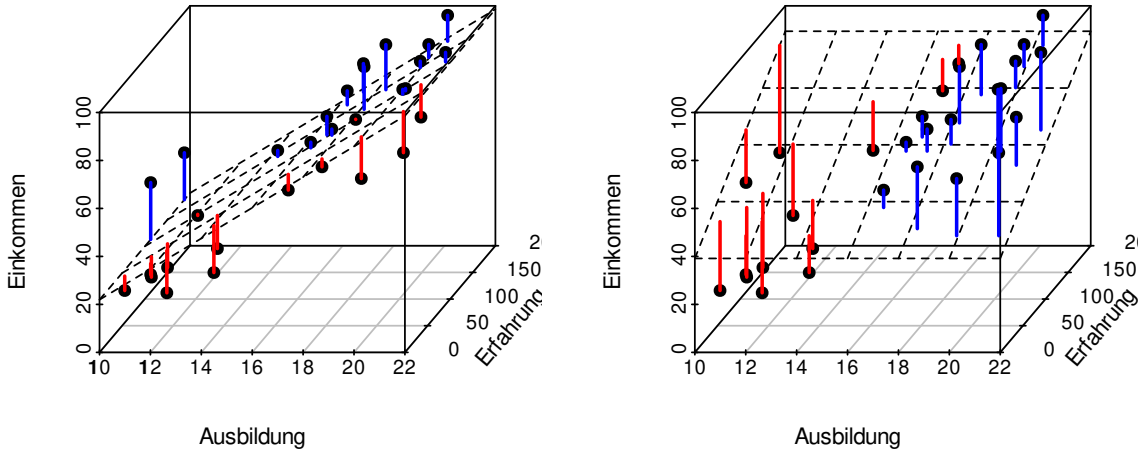


Abbildung 9.3. : Datensatz `Einkommen` mit jeweils *einer* erklärende Variablen

Auf der linken Seite von Abbildung 9.3 wurde **`Ausbildung`** hinzugefügt, auf der rechten **`Erfahrung`**. Nun sehen wir von Auge, dass die Residuen auf der linken Seite kleiner sind, als auf der rechten Seite. Umso kleiner der RSS (residual sum of squares) ist, umso besser passt das Modell zu den Daten. Wir wählen nun diejenige Variable, die besser zu den Punken passt und dies ist **`Ausbildung`**:

$$\texttt{Einkommen} = \texttt{Ausbildung} + \varepsilon$$

Haben wir noch weitere Variablen, so addieren wir diese jeweils *einzeln* zum eben beschriebenen Modell und wählen diejenige Variable aus, die den RSS am meisten verkleinert. ◀

Beispiel 9.2.2

Im Beispiel **`Credit`** beginnen wir mit dem sogenannten *Nullmodell* $\mathcal{M}_0$, das keine erklärenden Variablen enthält:

$$\texttt{Balance} = \beta_0 + \varepsilon$$

Dann fügen wir eine erklärende Variable zum Nullmodell hinzu. Dies machen wir am einfachsten mit dem **R**-Befehl **add1**, der jede vorkommende Variable getrennt addiert.

```
Credit <- read.csv("../Daten/Credit.csv")
f.full <- lm(Balance ~ Income + Limit + Rating + Cards + Age + Education +
    Gender + Student + Married + Ethnicity, data = Credit)
f.empty <- lm(Balance ~ NULL, data = Credit)
add1(f.empty, scope = f.full)


## Single term additions
##
## Model:
## Balance ~ NULL
##           Df Sum of Sq      RSS    AIC
## <none>                 84339912 4905.6
## Income     1  18131167 66208745 4810.7
## Limit      1  62624255 21715657 4364.8
## Rating     1  62904790 21435122 4359.6
## Cards      1    630416 83709496 4904.6
## Age        1       284 84339628 4907.6
## Education  1      5481 84334431 4907.5
## Gender     1     38892 84301020 4907.4
## Student    1   5658372 78681540 4879.8
## Married    1      2715 84337197 4907.5
## Ethnicity  2     18454 84321458 4909.5
```

Nun wählen wir die *beste* Variable aus, also jene, für deren Regressionsmodell sich der kleinste RSS-Wert ergibt. Damit passt diese Variable am besten zu den Daten. In diesem Fall trifft dies auf die Variable **Rating** zu.

Somit haben wir, vorläufig, das Modell $\mathcal{M}_1$

$$\texttt{Balance} = \beta_0 + \beta_1 \cdot \texttt{Rating} + \varepsilon$$

erhalten.

Zu diesem Modell fügen wir nun eine weitere Variable hinzu. Das führen wir am besten zuerst mit dem **update**-Befehl und dann wiederum mit dem **add1**-Befehl aus.

```
f.1 <- update(f.empty, . ~ . + Rating)
add1(f.1, scope = f.full)


## Single term additions
##
```

```
## Model:
## Balance ~ Rating
##           Df Sum of Sq      RSS    AIC
## <none>                 21435122 4359.6
## Income     1  10902581 10532541 4077.4
## Limit      1      7960 21427162 4361.5
## Cards      1    138580 21296542 4359.0
## Age        1    649110 20786012 4349.3
## Education  1     27243 21407879 4361.1
## Gender     1     16065 21419057 4361.3
## Student    1   5735163 15699959 4237.1
## Married    1    118209 21316913 4359.4
## Ethnicity  2     51100 21384022 4362.7
```

Wir wählen wieder diejenige Variable aus, aufgrund welcher das ergänzte Regressionsmodell den kleinsten RSS-Wert hat. Dies ist in diesem Fall **Income**. Somit erhalten wir das Modell $\mathcal{M}_2$:

$$\textbf{Balance} = \beta_0 + \beta_1 \cdot \textbf{Rating} + \beta_2 \cdot \textbf{Income} + \varepsilon$$

Das Verfahren wiederholt sich. Wir fügen jene Variable zum Modell $\mathcal{M}_2$ hinzu, aufgrund welcher das neue Regressionsmodell den kleinsten RSS-Wert hat.

```
f.2 <- update(f.1, . ~ . + Income)
add1(f.2, scope = f.full)


## Single term additions
##
## Model:
## Balance ~ Rating + Income
##           Df Sum of Sq      RSS    AIC
## <none>                 10532541 4077.4
## Limit      1     94545 10437996 4075.8
## Cards      1      2094 10530447 4079.3
## Age        1     90286 10442255 4076.0
## Education  1     20819 10511722 4078.6
## Gender     1       948 10531593 4079.4
## Student    1   6305322  4227219 3714.2
## Married    1     95068 10437473 4075.8
## Ethnicity  2     67040 10465501 4078.9
```

Dies ist hier **Student**, und somit erhalten wir das Modell $\mathcal{M}_3$:

$$\textbf{Balance} = \beta_0 + \beta_1 \cdot \textbf{Rating} + \beta_2 \cdot \textbf{Income} + \beta_3 \cdot \textbf{Student} + \varepsilon$$

Auf diese Weise erhalten wir 11 Modelle $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_{10}$.

Welches ist nun aber das beste unter diesen 11 Modellen? Als Entescheidungskriterium ziehen wir den AIC-Wert hinzu, der in der letzten Spalte aufgeführt ist. Aufgrund von diesem Wert lassen sich verschiedene Modelle miteinander vergleichen. Wir werden im Abschnitt **??** die genauere Bedeutung vom AIC-Wert besprechen.

Das hier aufgeführte Verfahren ist in dieser Form mit wiederholtem **update** und **add1** ziemlich mühsam. Der Befehl **regsubsets** führt das gesamte Verfahren automatisch durch.

```
library(leaps)
Credit <- read.csv("../Daten/Credit.csv")
Credit <- Credit[, -1]
reg <- regsubsets(Balance ~ ., data = Credit, method = "forward",
    nvmax = 11)
reg.sum <- summary(reg)
reg.sum$which
```

```
##    (Intercept) Income Limit Rating Cards   Age Education
## 1         TRUE  FALSE FALSE   TRUE FALSE FALSE     FALSE
## 2         TRUE   TRUE FALSE   TRUE FALSE FALSE     FALSE
## 3         TRUE   TRUE FALSE   TRUE FALSE FALSE     FALSE
## 4         TRUE   TRUE  TRUE   TRUE FALSE FALSE     FALSE
## 5         TRUE   TRUE  TRUE   TRUE  TRUE FALSE     FALSE
## 6         TRUE   TRUE  TRUE   TRUE  TRUE  TRUE     FALSE
## 7         TRUE   TRUE  TRUE   TRUE  TRUE  TRUE     FALSE
## 8         TRUE   TRUE  TRUE   TRUE  TRUE  TRUE     FALSE
## 9         TRUE   TRUE  TRUE   TRUE  TRUE  TRUE     FALSE
## 10        TRUE   TRUE  TRUE   TRUE  TRUE  TRUE     FALSE
## 11        TRUE   TRUE  TRUE   TRUE  TRUE  TRUE      TRUE
##    GenderFemale StudentYes MarriedYes EthnicityAsian
## 1         FALSE      FALSE      FALSE          FALSE
## 2         FALSE      FALSE      FALSE          FALSE
## 3         FALSE       TRUE      FALSE          FALSE
## 4         FALSE       TRUE      FALSE          FALSE
## 5         FALSE       TRUE      FALSE          FALSE
## 6         FALSE       TRUE      FALSE          FALSE
## 7          TRUE       TRUE      FALSE          FALSE
## 8          TRUE       TRUE      FALSE           TRUE
## 9          TRUE       TRUE       TRUE           TRUE
## 10         TRUE       TRUE       TRUE           TRUE
## 11         TRUE       TRUE       TRUE           TRUE
##    EthnicityCaucasian
## 1               FALSE
## 2               FALSE
## 3               FALSE
## 4               FALSE
## 5               FALSE
## 6               FALSE
## 7               FALSE
```

```
## 8                FALSE
## 9                FALSE
## 10                TRUE
## 11                TRUE
```

Überall, wo **TRUE** steht, kommt die entsprechende erklärende Variable vor. So kommen im Modell mit drei erklärenden Variablen also **Income**, **Rating** und **Student** vor.

◀

Formal sieht der Algorithmus wie folgt aus:

Algorithmus: Schrittweise Vorwärtsselektion

1. Sei $\mathcal{M}_0$ das Nullmodell, dass keine erklärende Variablen enthält.
2. Sei nun $k = 0, \ldots, p-1$:
   a) Wir betrachten alle $p-k$ Modelle, die die Anzahl Variablen in $\mathcal{M}_k$ um eine zusätzliche erklärende Variable erhöhen.
   b) Wir wählen das *beste* aus diesen $p-k$ Modellen aus und bezeichnen es mit $\mathcal{M}_{k+1}$. Das „beste" Modell ist dabei jenes mit dem kleinsten RSS oder grösstem R^2.
3. Unter den $\mathcal{M}_0, \ldots, \mathcal{M}_p$ wird jenes ausgesucht, das unter AIC, BIC oder adjusted R^2 am besten abschneidet.

9.2.2. Schrittweise Rückwärtsselektion

Die *schrittweise Rückwärtsselektion* ist rechnerisch ebenfalls effizient und funktioniert ähnlich wie die schrittweise Vorwärtsselektion, allerdings beginnen wir mit dem vollen Modell, das alle erklärenden Variablen enthält. Dann wird schrittweise eine Variable um die andere vom Modell entfernt, bis keine erklärende Variable mehr im Modell vorhanden ist. Insbesondere wird in jedem Schritt jene Variable vom Modell entfernt, die am wenigsten nützlich ist. Wir wollen das konkrete Vorgehen an einem Beispiel veranschaulichen.

Beispiel 9.2.3

Im Beispiel **Credit** beginnen wir mit dem sogenannten *vollen Modell* $\mathcal{M}_{10}$, das alle erklärenden Variablen enthält:

$$\begin{aligned}\texttt{Balance} = \beta_0 &+ \beta_1 \cdot \texttt{Income} + \beta_2 \cdot \texttt{Limit} + \beta_3 \cdot \texttt{Rating} + \beta_4 \cdot \texttt{Cards} \\ &+ \beta_5 \cdot \texttt{Age} + \beta_6 \cdot \texttt{Education} + \beta_7 \cdot \texttt{Gender} + \beta_8 \cdot \texttt{Student} \\ &+ \beta_9 \cdot \texttt{Married} + \beta_{10} \cdot \texttt{Ethnicity} + \varepsilon\end{aligned}$$

Dann entfernen wir eine erklärende Variable vom vollen Modell. Dies machen wir am einfachsten mit dem **R**-Befehl **drop1**, der jede vorkommende Variable getrennt wegnimmt.

```
Credit <- read.csv("../Daten/Credit.csv")
f.full <- lm(Balance ~ Income + Limit + Rating + Cards + Age + Education +
    Gender + Student + Married + Ethnicity, data = Credit)
f.empty <- lm(Balance ~ NULL, data = Credit)
drop1(f.full, scope = f.full)


## Single term deletions
##
## Model:
## Balance ~ Income + Limit + Rating + Cards + Age + Education +
##     Gender + Student + Married + Ethnicity
##           Df Sum of Sq      RSS    AIC
## <none>                  3786730 3686.2
## Income     1  10831162 14617892 4224.5
## Limit      1    331050  4117780 3717.7
## Rating     1     52314  3839044 3689.7
## Cards      1    162702  3949432 3701.0
## Age        1     42558  3829288 3688.7
## Education  1      4615  3791345 3684.7
## Gender     1     11269  3798000 3685.4
## Student    1   6326012 10112742 4077.1
## Married    1      6619  3793349 3684.9
## Ethnicity  2     14084  3800814 3683.7
```

Nun lassen wir die *schlechteste* Variable weg, wobei dies jene Variable ist, die zum reduzierten Regressionsmodell mit dem kleinsten RSS-Wert führt. Damit passt diese Variable am *schlechtesten* zu den Daten, da sie die grösste Verbesserung in Bezug auf den RSS-Werte mit sich bringt. In diesem Fall trifft dies auf die Variable **Education** zu.

Somit haben wir, vorläufig, das Modell $\mathcal{M}_9$

$$\begin{aligned}\texttt{Balance} = \beta_0 &+ \beta_1 \cdot \texttt{Income} + \beta_2 \cdot \texttt{Limit} + \beta_3 \cdot \texttt{Rating} + \beta_4 \cdot \texttt{Cards} \\ &+ \beta_5 \cdot \texttt{Age} + \beta_6 \cdot \texttt{Gender} + \beta_7 \cdot \texttt{Student} + \beta_8 \cdot \texttt{Married} \\ &+ \beta_9 \cdot \texttt{Ethnicity} + \varepsilon\end{aligned}$$

erhalten.

Von diesem Modell entfernen wir nun eine weitere Variable. Das machen wir am besten wieder mit dem **update**-Befehl und führen dann wieder den **drop1**-Befehl aus.

```
f.9 <- update(f.full, . ~ . - Education)
drop1(f.9, scope = f.9)


## Single term deletions
##
## Model:
## Balance ~ Income + Limit + Rating + Cards + Age + Gender + Student +
##     Married + Ethnicity
##           Df Sum of Sq      RSS    AIC
## <none>                  3791345 3684.7
## Income     1  10826551 14617896 4222.5
## Limit      1    326990  4118335 3715.8
## Rating     1     55071  3846417 3688.5
## Cards      1    162683  3954029 3699.5
## Age        1     43114  3834460 3687.2
## Gender     1     11103  3802448 3683.9
## Student    1   6340464 10131810 4075.9
## Married    1      7385  3798730 3683.5
## Ethnicity  2     14226  3805572 3682.2
```

Wir wählen wiederum diejenige Variable aus, die den kleinsten RSS-Wert hat. Dies ist hier **Married**. Somit erhalten wir das Modell $\mathcal{M}_8$:

$$\begin{aligned}\texttt{Balance} = \beta_0 &+ \beta_1 \cdot \texttt{Income} + \beta_2 \cdot \texttt{Limit} + \beta_3 \cdot \texttt{Rating} + \beta_4 \cdot \texttt{Cards} \\ &+ \beta_5 \cdot \texttt{Age} + \beta_6 \cdot \texttt{Gender} + \beta_7 \cdot \texttt{Student} \\ &+ \beta_8 \cdot \texttt{Ethnicity} + \varepsilon\end{aligned}$$

Wir wiederholen diese Schritte, bis keine erklärenden Variable mehr vorhanden ist. Auf diese Weise erhalten wir 11 Modelle $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_{10}$. Diese vergleichen wir wieder miteinander aufgrund ihrer AIC-Werte, welchen wir im Abschnitt **??** besprechen.

Auch hier nimmt uns der **regsubsets**-Befehl die ganze Arbeit ab.

```
library(leaps)
Credit <- read.csv("../Daten/Credit.csv")
Credit <- Credit[, -1]
reg <- regsubsets(Balance ~ ., data = Credit, method = "backward",
    nvmax = 11)
reg.sum <- summary(reg)
reg.sum$which


##    (Intercept) Income Limit Rating Cards   Age Education
## 1         TRUE  FALSE  TRUE  FALSE FALSE FALSE     FALSE
## 2         TRUE   TRUE  TRUE  FALSE FALSE FALSE     FALSE
## 3         TRUE   TRUE  TRUE  FALSE FALSE FALSE     FALSE
## 4         TRUE   TRUE  TRUE  FALSE  TRUE FALSE     FALSE
## 5         TRUE   TRUE  TRUE   TRUE  TRUE FALSE     FALSE
## 6         TRUE   TRUE  TRUE   TRUE  TRUE  TRUE     FALSE
## 7         TRUE   TRUE  TRUE   TRUE  TRUE  TRUE     FALSE
## 8         TRUE   TRUE  TRUE   TRUE  TRUE  TRUE     FALSE
## 9         TRUE   TRUE  TRUE   TRUE  TRUE  TRUE     FALSE
## 10        TRUE   TRUE  TRUE   TRUE  TRUE  TRUE     FALSE
## 11        TRUE   TRUE  TRUE   TRUE  TRUE  TRUE      TRUE
##    GenderFemale StudentYes MarriedYes EthnicityAsian
## 1         FALSE      FALSE      FALSE          FALSE
## 2         FALSE      FALSE      FALSE          FALSE
## 3         FALSE       TRUE      FALSE          FALSE
## 4         FALSE       TRUE      FALSE          FALSE
## 5         FALSE       TRUE      FALSE          FALSE
## 6         FALSE       TRUE      FALSE          FALSE
## 7          TRUE       TRUE      FALSE          FALSE
## 8          TRUE       TRUE      FALSE           TRUE
## 9          TRUE       TRUE       TRUE           TRUE
## 10         TRUE       TRUE       TRUE           TRUE
## 11         TRUE       TRUE       TRUE           TRUE
##    EthnicityCaucasian
## 1               FALSE
## 2               FALSE
## 3               FALSE
## 4               FALSE
## 5               FALSE
## 6               FALSE
## 7               FALSE
## 8               FALSE
## 9               FALSE
## 10               TRUE
## 11               TRUE
```

Im Modell mit drei erklärenden Variablen kommen **`Income`**, **`Limit`** und **`Student`** vor. Dieses Modell unterscheidet sich also vom Modell mit drei Variablen, das durch Vorwärtsselektion gewonnen wurde. Hier kommt **`Rating`** anstelle von **`Limit`** vor.

◀

Formal sieht der Algorithmus wie folgt aus

Algorithmus: Schrittweise Rückwärtsselektion

1. Sei $\mathcal{M}_p$ das volle Modell, das alle erklärende Variablen enthält.
2. Es sei $k = p, p-1, \ldots, 1$:
   a) Wir betrachten alle k Modelle, die die Anzahl Variablen in $\mathcal{M}_k$ um eine erklärende Variable verkleinern.
   b) Wir wählen das *beste* aus diesen k Modellen aus und bezeichnen es mit $\mathcal{M}_{k-1}$. Das „beste" Modell ist dabei jenes mit dem kleinsten RSS oder grösstem R^2.
3. Unter den $\mathcal{M}_0, \ldots, \mathcal{M}_p$ wird jenes ausgesucht, das in Bezug auf AIC, BIC oder adjusted R^2 am besten abschneidet.

Bemerkung:

Es gibt noch weitere Selektionsmethoden, auf die wir hier nicht eingehen wollen. ♦

9.2.3. Anzahl der Variablen

Mit der Vorwärts- und Rückwärtsselektion werden der Reihe nach Variablen hinzugefügt. Allerdings beschreiben sie *nicht*, wieviele Variablen hinzugefügt werden sollen. Der RSS-Wert ist dabei nicht hilfreich, da er mit jeder hinzugefügten Variable kleiner wird. Es gibt hier kein Abbruchkriterium.

Anstatt des RSS hätten wir auch den R^2-Wert wählen können (RSS ist graphisch intuitiver). Aber auch der R^2-Wert bietet kein Abbruchkriterium, da dieser mit jeder hinzugefügten Variable grösser wird.

Wir können den R^2-Wert aber abändern, so dass er auch noch abhängig von der Anzahl Variablen wird. Diese neue Grösse heisst *adjusted* R^2.

adjusted R^2

Für die Methode der kleinsten Quadrate mit p Variablen berechnen wir den adjusted R^2 durch

$$\text{adjusted } R^2 = 1 - (1 - R^2) \cdot \frac{n-1}{n-p-1}$$

Hier spricht ein *grosser* Wert von adjusted R^2 für die Güte des jeweiligen Modells. Um diesen Wert zu maximieren, müssen wir

$$(1 - R^2) \cdot \frac{n-1}{n-p-1} \qquad (*)$$

minimieren. Während $1 - R^2$ immer kleiner wird, je mehr Variablen wir hinzufügen, *kann* der Term $(*)$ grösser oder kleiner werden, weil p im Nenner vorkommt. Die dem adjusted R^2 zugrundeliegende Idee besteht ist die folgende: Beinhaltet ein Modell alle wichtigen Variablen (und keine weiteren), so führt die Hinzunahme einer weiteren Variablen nur zu einer sehr kleinen Abnahme des $1 - R^2$. Die Hinzunahme von einer Variable vergrössert aber p, und dies bewirkt eine *Vergrösserung* von Term $(*)$. Das hat wiederum eine *Abnahme* vom adjusted R^2 zur Folge.

Beispiel 9.2.4

Für den Datensatz **Credit** werden in Abbildung 9.4 graphisch R^2 und adjusted R^2 miteinander verglichen. Von Auge ist kein Unterschied erkennbar.

Vergleichen wir allerdings die Werte von R^2 und adjusted R^2, so stellen wir fest, dass beim R^2 die Werte, wie erwartet, immer zunehmen, während beim adjusted R^2 ein Maximum erreicht wird.

```
library(leaps)
Credit <- read.csv("../Daten/Credit.csv")
Credit <- Credit[, -1]
reg <- regsubsets(Balance ~ ., data = Credit, method = "forward",
    nvmax = 11)
reg.sum <- summary(reg)
round(reg.sum$rsq, 5)


##  [1] 0.74585 0.87512 0.94988 0.95219 0.95416 0.95469 0.95482
##  [8] 0.95489 0.95496 0.95505 0.95510
```

Aus dem **R**-Output ist ersichtlich, dass R^2 mit zunehmender Anzahl erklärender Variablen wächst. Für den adjusted R^2 gilt dies nicht. Dem folgenden **R**-Output entnehmen wir, dass der Wert von adjusted R^2 für sieben erklärende Variablen am grössten ist und danach wieder abnimmt.

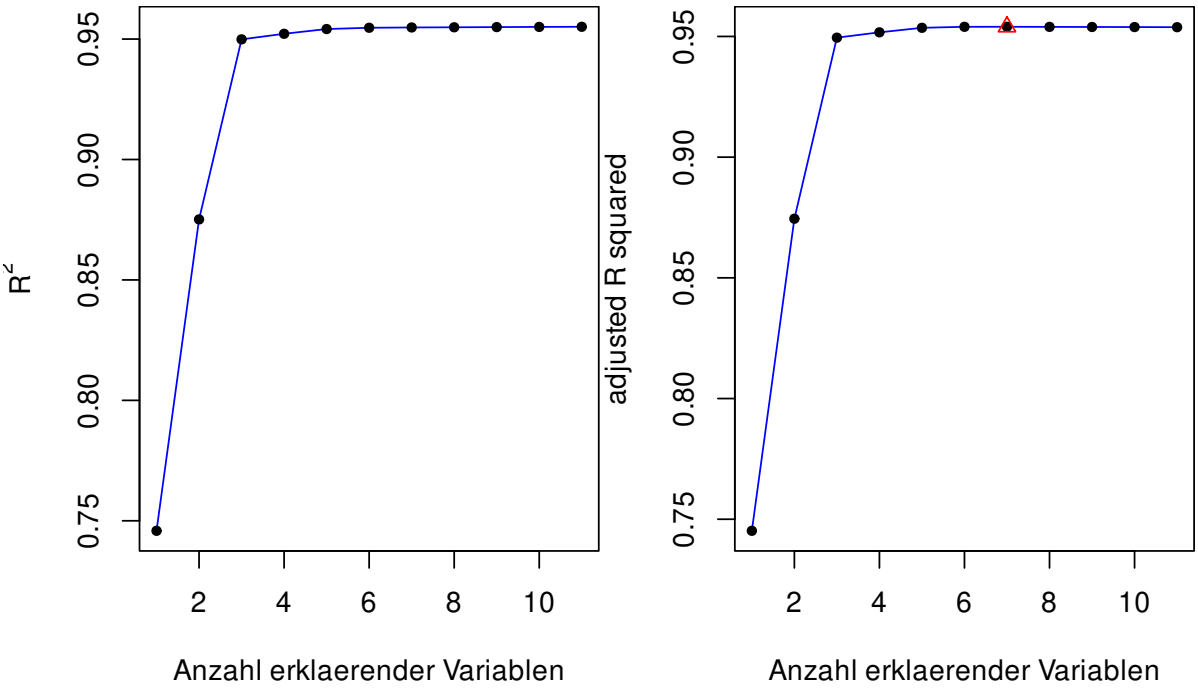


Abbildung 9.4. : Beispieldatensatz `Credits`: Vergleich der Werte für R^2 und adjusted R^2 als Funktion von der Anzahl erklärender Variablen für jedes durch Vorwärtsselektion ermittelte Regressionsmodell.

```
round(reg.sum$adjr2, 5)

##  [1] 0.74521 0.87449 0.94950 0.95170 0.95358 0.95400 0.95401
##  [8] 0.95396 0.95392 0.95389 0.95383

which.max(reg.sum$adjr2)

## [1] 7
```

Wenn wir adjusted R^2 als Entscheidungskriterium zu Hand nehmen, so ist das beste aufgrund von Vorwärtsselektion ermittelte Regressionsmodell also

$$\texttt{Balance} = \beta_0 + \beta_1 \cdot \texttt{Income} + \beta_2 \cdot \texttt{Limit} + \beta_3 \cdot \texttt{Rating} + \beta_4 \cdot \texttt{Cards} + \beta_5 \cdot \texttt{Age} + \beta_6 \cdot \texttt{Gender} + \beta_7 \cdot \texttt{Student} + \varepsilon$$

◀

Der adjusted R^2-Wert ist allerdings nur eines von vielen Gütemassen, um Modelle mit unterschiedlicher Anzahl von erklärenden Variablen miteinander zu vergleichen. Weitere sind AIC, BIC oder Malloy's c_p auf die wir hier nicht weiter eingehen werden.