Einführung in die Statistik mit R

Andreas Handl

Inhaltsverzeichnis

Kapitel 1

Grundbegriffe

> Statistisches Denken wird für den mündigen Bürger eines Tages dieselbe Bedeutung haben wie die Fähigkeit lesen und schreiben zu können.
>
> - H. G. WELLS (1895)

Mehr als 100 Jahre sind vergangen, seitdem H. G. Wells diese Prophezeiung über die Zukunft der Statistik aufgestellt hat. Und man kann sagen, dass er Recht hatte. Es gibt kaum einen Artikel in der Tageszeitung, in dem keine Zahlenangaben in Form von Tabellen oder Grafiken zu finden sind. Und bei vielen Fernsehdiskussionen zwischen Politikern ertrinkt man fast im Zahlenmeer. Es stellt sich aber die Frage, ob statistisches Denken auch Allgemeingut geworden ist. Ziel dieses Skriptes ist es, dass seine Leser nach der Lektüre die Welt mit den Augen eines Statistikers betrachten können.

Daten werden erhoben und ausgewertet. In ihrem hervorragenden Buch überschreibt Utts (1999) die ersten beiden Teile mit *Finding Data in Life* und *Finding Life in Data*. Und auch wir werden zuerst lernen, wie man im Leben Daten findet. Hierfür gibt es unterschiedliche Möglichkeiten. Man spricht von **Erhebungsarten**. Mit diesen werden wir uns im nächsten Abschnitt beschäftigen. Danach suchen wir das Leben in den Daten. Zum einen werden wir eine Vielzahl von Möglichkeiten kennenlernen, Daten darzustellen und zu beschreiben. Man spricht von **deskriptiver** oder **beschreibender Statistik**. Daten werden aber auch benutzt, um von einer Teilgesamtheit auf die Gesamtheit zu schließen, aus der sie stammt, oder um eine Theorie zu überprüfen. Dies ist das Aufgabengebiet der **schließenden Statistik** oder **Inferenzstatistik**. Dieser ist ein großer Teil dieses Skriptes gewidmet.

Schauen wir uns zunächst aber zwei Beispiele an, mit denen wir uns immer wieder beschäftigen werden.

Beispiel 1
Der Dozent einer Weiterbildungsveranstaltung möchte gerne Informationen über seine Teilnehmer gewinnen. Ihn interessiert unter anderem die Geschlechterverteilung und die Altersverteilung. Aus diesem Grund entwirft er einen Fragebogen, den er von den Teilnehmern ausfüllen lässt. Der Fragebogen ist auf Seite 16 zu finden.

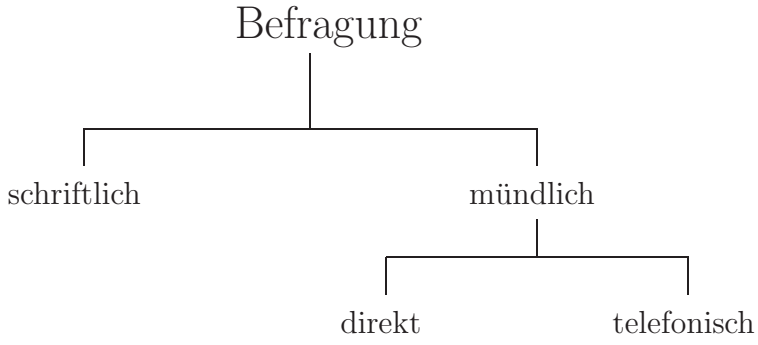
■

Beispiel 2
In einem Projekt zur Betriebsinformatik an der Fakultät für Wirtschaftswissenschaften der Universität Bielefeld wurden die Studierenden nach ihrem Geschlecht, ihrem Alter, dem Alter der Mutter, dem Alter des Vaters und der Anzahl der Geschwister gefragt.

□

1.1 Erhebungsarten

Die Daten der Beispiele 1 und 2 wurden im Rahmen einer *schriftlichen Befragung* erhoben. Neben einer schriftlichen Befragung kann man auch eine **mündliche Befragung** durchführen, wobei man bei der mündlichen Befragung die **direkte mündliche Befragung** und die **telefonische Befragung** unterscheidet. Die folgende Abbildung zeigt noch einmal die unterschiedlichen Befragungsarten.



Bei fast jeder Befragung kann der Befragte die Antwort verweigern. Die einzige Ausnahme sind amtlich statistische Erhebungen. Paragraph 15 BStatG legt fest, dass

> die eine Bundesstatistik anordnende Rechtsvorschrift festzulegen hat, ob und in welchem Umfang die Erhebung mit oder ohne

> Auskunftspflicht erfolgen soll. Ist eine Auskunftspflicht festgelegt, sind alle natürlichen und juristischen Personen des privaten und öffentlichen Rechts, Personenvereinigungen, Behörden des Bundes und der Länder sowie Gemeinden und Gemeindeverbände zur Beantwortung der ordnungsgemäß gestellten Fragen verpflichtet.

Dabei ist

> die Antwort wahrheitsgemäß, vollständig und innerhalb der von den statistischen Ämtern des Bundes und der Länder gesetzten Fristen zu erteilen.

Bei einer Befragung sollte die **Antwortquote** hoch sein. Diese ist das Verhältnis aus der Anzahl der Befragten, die geantwortet haben, und der Anzahl der Personen, die befragt wurden.

$$\text{Antwortquote} = \frac{\text{Anzahl der Befragten, die antworten}}{\text{Anzahl Befragter}}$$

Bei einer mündlichen Befragung kann man mit einer hohen Antwortquote rechnen, da ein guter Interviewer verhindern wird, dass die Befragten das Interview abbrechen. Einen Fragebogen kann man aber in den Papierkorb werfen, ohne dass dies negative Konsequenzen hat. Deshalb ist die Antwortquote bei einer schriftlichen Befragung in der Regel niedrig. Man bietet den Befragten materielle Anreize, wenn sie an der Befragung teilnehmen. Eine mündliche Befragung ist in der Regel sehr zeitaufwändig, während bei einer schriftlichen Befragung der Zeitaufwand gering ist. Die Gefahr von Missverständnissen ist bei mündlichen Befragungen gering, da die Befragten nachfragen können. Dies ist bei schriftlichen Befragungen nicht möglich, sodass man sich nicht darauf verlassen kann, dass jede Frage richtig verstanden wurde. Für eine schriftliche Befragung sprechen aber die im Verhältnis zur mündlichen Befragung niedrigeren Kosten.

Bei den Fragen unterscheidet man **offene** und **geschlossene** Fragen. Bei einer geschlossenen Frage sind die möglichen Antworten vorgegeben. Bei einer offenen Frage ist dies nicht der Fall. Es ist unproblematisch, geschlossene Fragen statistisch auszuwerten, da die Antworten unterschiedlicher Befragter vergleichbar sind. Bei offenen Fragen erhält man oft zu viele unterschiedliche Antworten. Ein Nachteil geschlossener Fragen ist aber, dass der Fragende nicht alle Antwortmöglichkeiten berücksichtigt. Dieses Problem kann man aber dadurch vermeiden, dass man halboffene Fragen stellt, bei denen der Befragte die Möglichkeit besitzt, eine Antwort zu geben, die nicht unter den vorgegebenen Antworten zu finden ist. Wir wollen hier nicht weiter auf Befragungen eingehen. Wertvolle Hinweise zur Frageformulierung und Fragebogengestaltung kann man bei Diekmann (2004) und Hüttner (2002) finden.

Wir haben bisher die Befragung als einzige Erhebungstechnik kennengelernt. Man kann einen Sachverhalt aber auch beobachten, um Daten zu erheben. Man spricht in diesem Fall von einer **Beobachtung**.

Beispiel 3
Sie schlagen am 1. September 1999 die Neue Westfälische auf und suchen alle Einzimmerwohnungen heraus, die explizit in Uninähe liegen. Es sind sechs. Tabelle 1.1 gibt neben der Fläche in m^2 auch die Kaltmiete in DM für jede der sechs Wohnungen an.

Tabelle 1.1: Fläche in m^2 und Kaltmiete in DM von Einzimmerwohnungen

Wohnung	1	2	3	4	5	6
Fläche	55	40	30	23	26	45
Miete	530	520	420	500	440	650

□

Sehr oft ist eine Beobachtung genauer als eine Befragung. Fragt man die Teilnehmer einer Veranstaltung nach ihrem Körpergewicht, so wird man nur einen annähernden Wert erhalten. Viele Teilnehmer werden eine Zahl nennen, die auf 0 oder 5 endet wie 70 kg oder 85 kg. Wiegt man die Personen, so erhält man den exakten Wert. Noch gravierender wäre der Unterschied, wenn man die Teilnehmer nach der Länge ihres Fahrweges von zu Hause zum Veranstaltungsort fragen würde. Wer schaut schon zu Beginn und am Ende einer Fahrt auf die Uhr? Kommt bei beiden Beispielen der Fehler allein durch eine Fehleinschätzung der Befragten zustande, so wird man bei vielen Fragen davon ausgehen müssen, dass nicht die wahre Antwort gegeben wird. Dies wird bei Fragen nach dem Einkommen oder auch dem täglichen Alkoholkonsum der Fall sein. Hier wäre eine Beobachtung viel sinnvoller. Sie ist aber nicht möglich.
Sehr oft wird ein Merkmal zu äquidistanten Zeitpunkten erhoben. Man erhält in diesem Fall eine **Zeitreihe**.

Beispiel 4
Ein Student notiert vom 18.9.2002 bis zum 7.10.2002 die tägliche Höchsttemperatur in Celsius in Bielefeld. Er erhält folgende Werte:

```
17 18 19 17 16 14 15 12 15 15 15 17 20 21 18 17 17 13 11 10
```

□

Werden die Daten durch Beobachtung erhoben, so hat man keinen Einfluss auf die Randbedingungen. Sollen aber zwei oder mehr Verfahren oder Behandlungen verglichen werden, so muss man sicherstellen, dass alle anderen Einflussfaktoren nicht variieren. In diesem Fall sollte man ein **Experiment** durchführen. Schauen wir uns hierzu ein Beispiel an.

Beispiel 5
In der SZ vom 10.8.2005 ist folgender Artikel von Barbara Kerbel zu finden.

> Wenn es am Kopf juckt und kratzt, hat sie wieder zugebissen, die Laus. Alle drei Stunden saugen die Parasiten Blut, der Speichel, den sie dabei verlieren, verursacht heftiges Jucken. Wer sie hat, will sie so schnell wie möglich wieder loswerden - aber wie? Am besten klappt das offenbar, wenn man die nassen Haare mit einem feinen Kamm gründlich kämmt, wie eine Studie von Gesundheitswissenschaftlern der London School of Hygiene zeigt (*British Medical Journal*, Online-Veröffentlichung vom 5. August).
>
> Wie die britische Studie zeigt, eignen sich die Kämme nicht nur zum Aufspüren der Läuse, sondern auch zur Therapie - und zwar wirkungsvoller als die üblichen Insektizide. 133 Probanden zwischen zwei und fünfzehn Jahren wurden untersucht. Nach dem Zufallsprinzip wurden sie auf zwei Behandlungen verteilt: Die Hälfte bekam ein Insektizid, entweder Permethrin oder Malathion, die anderen wurden angewiesen, zwei Wochen lang täglich die nassen, mit Spülung gewaschenen Haare mit dem Nissenkamm zu kämmen. Nach der Behandlung war bei 57 Prozent der Kamm-Gruppe keine Laus mehr zu finden; in der Insektizid-Gruppe waren hingegen nur 13 Prozent der Teilnehmer von den Parasiten befreit.

□

An diesem Beispiel kann man sehr schön erkennen, wodurch sich ein Experiment von einer Beobachtung unterscheidet. Bei einer Beobachtung ist man passiv. Man notiert die Ergebnisse, ohne einzugreifen. Bei einem Experiment hingegen legt man die Rahmenbedingungen fest. Im Text heißt es, dass die Probanden nach dem Zufallsprinzip auf zwei Behandlungen verteilt wurden. Hierdurch soll der Einfluss aller anderen Faktoren ausgeschaltet werden. Warum dies so ist, werden wir an späterer Stelle lernen. Der Begriff *Zufall* wird in diesem Skript noch eine sehr große Rolle spielen.

1.2 Merkmale

Im Rahmen einer Befragung ist jeder Befragte eine **statistische Einheit**. An ihm werden eine Reihe von Merkmalen erhoben. Man nennt ihn deshalb auch einen **Merkmalsträger**. Die Menge aller Merkmalsträger heißt **Gesamtheit**. Werden alle interessierenden Einheiten erfasst, so spricht man einer **Vollerhebung**, ansonsten von einer **Teilerhebung**.

Beispiel 6
Wir schauen uns noch einmal das Beispiel 1 auf Seite 10 an. Wurde die Erhebung nur durchgeführt, um Aussagen über die Population der 25 Teilnehmer zu machen, so handelt es sich um eine Vollerhebung. Dient die Erhebung aber dazu, Aussagen über eine größere Population zu machen, so handelt es sich um eine Teilerhebung.

□

Da Erhebungen teuer sind, sind die meisten Erhebungen Teilerhebungen. Sehr oft ist es auch nicht möglich, eine Vollerhebung durchzuführen. Will man zum Beispiel die Lebensdauer von Glühbirnen untersuchen, so würde eine Totalerhebung zur Zerstörung der gesamten Produktion führen. Man spricht in diesem Fall von **zerstörender Prüfung**.

Auch wenn die meisten Erhebungen Teilerhebungen sind, ist man in der Regel aber nicht an der Teilgesamtheit, sondern an der Gesamtheit interessiert, zu der die Teilgesamtheit gehört. Man spricht auch von der **Grundgesamtheit**. Wie und warum man von einer Teilgesamtheit auf eine Grundgesamtheit schließen kann, werden wir später lernen.

Bei jeder statistischen Einheit werden eine Reihe von **Merkmalen** erhoben. So gehört im Beispiel 1 auf Seite 10 zu jeder Frage ein Merkmal. Jede Frage besitzt mehrere Antwortmöglichkeiten. Diese heißen auf der Ebene der Merkmale **Merkmalsausprägungen**.

Schauen wir uns die Merkmalsausprägungen der einzelnen Merkmale im Fragebogen auf Seite 16 genauer an, so stellen wir fest, dass sich die Merkmale hinsichtlich der Anzahl und der Art der Ausprägungsmöglichkeiten unterscheiden.

Das Merkmal `Geschlecht` hat die Merkmalsausprägungen `weiblich` oder `männlich`. Die Ausprägungen des Merkmals `Geschlecht` sind **Klassen** oder **Kategorien**. Wir können uns vorstellen, dass die Merkmalsausprägungen zwei Kästchen sind. Auf dem einen Kästchen steht `weiblich`, auf dem anderen `männlich`. Jede der Personen passt in genau eines der Kästchen. Es handelt sich um ein **nominalskaliertes Merkmal**. Wir sprechen auch von einem Merkmal mit *nominalem Messniveau*. Wir können die Merkmalsaus-

prägungen nur benennen. Im Fragebogen gibt es noch zwei weitere Fragen, deren Antwortmöglichkeiten ein nominalskaliertes Merkmal erzeugen. Es handelt sich zum einen um die Frage, ob man den Film Titanic gesehen hat. Das zugehörige Merkmal hat die Ausprägungsmöglichkeit `ja`, falls die Person den Film Titanic gesehen hat, und `nein`, falls sie ihn nicht gesehen hat. Am Ende des Fragebogens soll man einen Satz richtig fortsetzen. Das zugehörige Merkmal hat die Ausprägungsmöglichkeit `ja`, falls der Satz richtig ergänzt wurde, und die Ausprägungsmöglichkeit `nein`, falls der Satz falsch ergänzt wurde.

Das Merkmal `Bewertung` sieht auf den Blick genauso wie die bisher betrachteten Merkmale aus. Wir können aber bei diesem Merkmal nicht nur Kategorien bilden, sondern diese Kategorien sind geordnet. Da `sehr gut` besser als `gut` und `gut` besser als `mittelmäßig` ist, sind die Merkmalsausprägungen geordnet. Man spricht von einem **ordinalskalierten** oder auch **ordinalem** Merkmal.

Die Ausprägungsmöglichkeiten des Merkmals `Alter` sind die natürlichen Zahlen $1, 2, 3, \ldots$. Sind bei einem Merkmal die Ausprägungen Zahlen, bei denen Abstände und Verhältnisse sinnvoll interpretiert werden können, so spricht man von einem **metrischen** Merkmal. Das Merkmal `Trinkgeld` ist ebenfalls metrisch.

Im Beispiel 2 auf Seite 10 sind die Merkmale Alter, Alter der Mutter, Alter des Vaters und Anzahl der Geschwister metrisch.

Wir haben also nominalskalierte, ordinalskalierte und metrische Merkmale unterschieden. Man spricht auch vom **Skalenniveau** der Merkmale.

Man kann weiterhin **qualitative** und **quantitative** Merkmale unterscheiden. Nominalskalierte und ordinalskalierte Merkmale sind qualitative Merkmale, metrische Merkmale hingegen quantitative Merkmale. Bei quantitativen Merkmalen unterscheidet man **diskrete** und **stetige** Merkmale. Diskrete Merkmale haben nur endlich oder abzählbar unendlich viele Ausprägungen. So ist das Merkmal `Anzahl der Kunden in einer Warteschlange` diskret. Stetige Merkmale können alle Werte aus einem Intervall annehmen. Das Merkmal `Wartezeit eines Kunden` ist stetig.

Bei einer statistischen Erhebung erhält man nun für jedes Merkmal bei jeder statistischen Einheit eine Ausprägung. Tabelle 1.2 auf Seite 17 zeigt die Ergebnisse der Befragung. In jeder Zeile der Tabelle stehen die Merkmalsausprägungen eines Teilnehmers. So ist der dritte Teilnehmer weiblich, 26 Jahre alt, hat den Film Titanic gesehen, fand ihn gut und würde 1.80 DM Trinkgeld geben. In jeder Spalte stehen die Werte eines Merkmals. Die Daten zur Erhebung aus Beispiel 2 auf Seite 10 sind in Tabelle 1.3 auf Seite 18 zu finden.

Der Fragebogen

GESCHLECHT

w [ ]
m [ ]

ALTER Jahre
HABEN SIE DEN FILM TITANIC GESEHEN?

ja [ ]
nein [ ]

WENN JA, WIE HAT IHNEN DER FILM GEFALLEN?

sehr gut [ ]
gut [ ]
mittelmäßig [ ]
schlecht [ ]
sehr schlecht [ ]

SIE HABEN IN EINEM RESTAURANT EINE RECHNUNG ÜBER 43.20 DM ZU ZAHLEN. WIEVIEL TRINKGELD GEBEN SIE?

gar keines [ ]
0.80 DM [ ]
1.80 DM [ ]
2.80 DM [ ]
3.80 DM [ ]
4.80 DM [ ]
5.80 DM [ ]
6.80 DM [ ]

BITTE ERGÄNZEN SIE DEN FOLGENDEN SATZ:
Zu Risiken und Nebenwirkungen

Tabelle 1.2: Ergebnis einer Befragung unter Teilnehmern einer Weiterbildungsveranstaltung

Person	Geschlecht	Alter	Titanic	Bewertung	Trinkgeld	Satz
1	m	30	n	.	1.80	n
2	w	23	j	g	1.80	n
3	w	26	j	g	1.80	j
4	m	33	n	.	2.80	n
5	m	37	n	.	1.80	n
6	m	28	j	g	2.80	j
7	w	31	j	sg	2.80	n
8	m	23	n	.	0.80	n
9	w	24	j	sg	1.80	j
10	m	26	n	.	1.80	n
11	w	23	j	sg	1.80	j
12	m	32	j	g	1.80	n
13	m	29	j	sg	1.80	j
14	w	25	j	g	1.80	j
15	w	31	j	g	0.80	n
16	m	26	j	g	2.80	n
17	m	37	n	.	3.80	n
18	m	38	j	g	.	n
19	w	29	n	.	3.80	n
20	w	28	j	sg	1.80	n
21	w	28	j	m	2.80	j
22	w	28	j	sg	1.80	j
23	w	38	j	g	2.80	n
24	w	27	j	m	1.80	j
25	m	27	n	.	2.80	j

Tabelle 1.3: Ergebnis einer Befragung unter Teilnehmern eines Projektes zur Betriebsinformatik

Person	Geschlecht	Alter	Alter der Mutter	Alter des Vaters	Anzahl Geschwister
1	0	29	58	61	1
2	1	26	53	54	2
3	0	24	49	55	1
4	1	25	56	63	3
5	1	25	49	53	0
6	1	23	55	55	2
7	0	23	48	54	2
8	0	27	56	58	1
9	0	25	57	59	1
10	0	24	50	54	1
11	1	26	61	65	1
12	0	24	50	52	1
13	0	29	54	56	1
14	0	28	48	51	2
15	1	23	52	52	1
16	0	24	45	57	1
17	1	24	59	63	0
18	1	23	52	55	1
19	0	24	54	61	2
20	1	23	54	55	1

Kapitel 2

Eine kleine Einführung in R

Da die Datensätze in diesem Skript klein sind, kann man alle Beispiele mit Papier, Bleistift und Taschenrechner in vertretbarer Zeit nachvollziehen. Bei größeren Datensätzen sollte man auf den Computern zurückgreifen. Hier kann der Anwender statistischer Verfahren unter einer Vielzahl von Statistikpaketen wählen. Unter diesen werden SAS und SPSS bei der Mehrzahl der professionellen Datenanalysen verwendet. Beide Pakete sind aber sehr teuer und es ist nicht einfach, neue Verfahren zu implementieren. Im Statistik-Paket R sind sehr viele statistische Verfahren vorhanden. Außerdem ist R frei verfügbar.

2.1 R als mächtiger Taschenrechner

R bietet eine **interaktive Umgebung**, den **Befehlsmodus**, in dem man die Daten direkt eingeben und analysieren kann. Durch das **Bereitschaftszeichen** > wird angezeigt, dass eine Eingabe erwartet wird. Der Befehlsmodus ist ein mächtiger Taschenrechner. Wir können hier die Grundrechenarten Addition, Subtraktion, Multiplikation und Division mit den Operatoren +, -, * und / durchführen. Bei Dezimalzahlen verwendet man einen Dezimalpunkt. Nachdem wir einen Befehl mit der Taste `carriage return` abgeschickt haben, gibt R das Ergebnis in der nächsten Zeile aus. Hier sind einige einfache Beispiele:

Befehlsmodus

```
> 2.1+2
[1] 4.1


> 2.1-2
[1] 0.1
```

```
> 2.1*2
[1] 4.2

> 2.1/2
[1] 1.05
```

Zum Potenzieren benutzen wir ^ :

```
> 2.1^2
[1] 4.41
```

Die Quadratwurzel von 2 erhalten wir also durch

```
> 2^0.5
[1] 1.414214
```

Funktion sqrt

Man kann aber auch die **Funktion** sqrt verwenden. Dabei ist sqrt eine Abkürzung für square root, also Quadratwurzel. Namen von Funktionen sind in R unter mnemotechnischen Gesichtspunkten gewählt. Funktionen bieten die Möglichkeit, einen oder mehrere Befehle unter einem Namen abzuspeichern.

Argument

Sie besitzen in der Regel **Argumente**. So muss man der Funktion sqrt mitteilen, von welcher Zahl sie die Quadratwurzel bestimmen soll. Diese Zahl ist Argument der Funktion sqrt. Die Argumente einer Funktion stehen in runden Klammern hinter dem Funktionsnamen und sind durch Kommata voneinander getrennt. Wir rufen die Funktion sqrt also mit dem Argument 2 auf:

```
> sqrt(2)
[1] 1.414214
```

round

R gibt 6 Stellen nach dem Dezimalpunkt aus. Mit weniger Stellen wird das Ergebnis übersichtlicher. Wir sollten also runden und verwenden hierzu die Funktion round. Dabei können wir der Funktion round den Aufruf der Funktion der Funktion sqrt als Argument übergeben, was bei allen Funktionen möglich ist.

```
> round(sqrt(2))
[1] 1
```

help

Jetzt ist das Ergebnis übersichtlich aber ungenau. Wir müssen der Funktion round also noch mitteilen, auf wie viele Stellen nach dem Dezimalpunkt wir runden wollen. Wie wir dies erreichen können, erfahren wir, indem wir die Funktion help mit dem Argument round aufrufen:

```
> help(round)
```

Wir sehen, dass die Funktion folgendermaßen aufgerufen wird:

```
round(x, digits = 0)
```

Neben dem ersten Argument, bei dem es sich um die zu rundende Zahl handelt, gibt es noch das Argument `digits`. Dieses gibt die Anzahl der Stellen nach dem Dezimalpunkt an, auf die gerundet werden soll, und nimmt standardmäßig den Wert 0 an. `digits`

Funktionen in `R` besitzen zwei Typen von Argumenten. Es gibt Argumente, die beim Aufruf der Funktion angegeben werden müssen. Bei der Funktion `round` ist dies das Argument `x`. Es gibt aber auch optionale Argumente, die nicht angegeben werden müssen. In diesem Fall wird ihnen der Wert zugewiesen, der in der Kopfzeile zu finden ist. Das Argument `digits` nimmt also standardmäßig den Wert 0 an.

Wie übergibt man einer Funktion, die mindestens zwei Argumente besitzt, diese? Hierzu gibt es eine Reihe von Möglichkeiten, die wir an Hand der Funktion `round` illustrieren wollen. Kennt man die Reihenfolge der Argumente im Kopf der Funktion, so kann man sie ohne zusätzliche Angaben eingeben.

```
> round(sqrt(2),2)
[1] 1.41
```

Man kann aber auch die Namen der Argumente verwenden, wie sie im Kopf der Funktion stehen.

```
> round(x=sqrt(2),digits=2)
[1] 1.41
```

Verwendet man die Namen, so kann man die Argumente in beliebiger Reihenfolge eingeben.

```
> round(digits=2,x=sqrt(2))
[1] 1.41
```

Man kann die Namen der Argumente abkürzen, wenn sie dadurch eindeutig bleiben. Beginnen zwei Namen zum Beispiel mit `di`, so darf man `di` nicht als Abkürzung verwenden.

```
> round(x=sqrt(2),d=2)
[1] 1.41
```

2.2 Datenstrukturen

Bei statistischen Erhebungen werden bei jedem von n Merkmalsträgern jeweils p Merkmale erhoben. In diesem Kapitel werden wir lernen, wie man die Daten eingibt und unter einem Namen abspeichert, mit dem man auf sie zurückgreifen kann.

Wir gehen zunächst davon aus, dass nur ein Merkmal erhoben wurde. Schauen wir uns ein Beispiel an.

Ein Schallplattensammler im letzten halben Jahr fünf Langspielplatten bei einem amerikanischen Händler gekauft und dafür folgende Preise in US Dollar bezahlt:

```
22  30  16  25  27
```

Vektor

Komponente

`c`

Wir geben die Daten als **Vektor** ein. Ein Vektor ist eine Zusammenfassung von Objekten zu einer endlichen Folge und besteht aus **Komponenten**. Einen Vektor erzeugt man in `R` mit der Funktion `c` . Diese macht aus einer Folge von Zahlen, die durch Kommata getrennt sind, einen Vektor, dessen Komponenten die einzelnen Zahlen sind. Die Zahlen sind die Argumente der Funktion `c`. Wir geben die Daten ein.

```
> c(22,30,16,25,27)
```

Am Bildschirm erhalten wir folgendes Ergebnis:

```
[1] 22 30 16 25 27
```

Die Elemente des Vektors werden ausgegeben. Am Anfang steht `[1]`. Dies zeigt, dass die erste Zahl gleich der ersten Komponente des Vektors ist.

Variable

`<-`

Um mit den Werten weiterhin arbeiten zu können, müssen wir sie in einer **Variablen** speichern. Dies geschieht mit dem **Zuweisungsoperator** `<-` , den man durch die Zeichen `<` und `-` erhält. Auf der linken Seite steht der Name der Variablen, der die Werte zugewiesen werden sollen, auf der rechten Seite steht der Aufruf der Funktion `c`.

Die Namen von Variablen dürfen beliebig lang sein, dürfen aber nur aus Buchstaben, Ziffern und dem Punkt bestehen, wobei das erste Zeichen ein Buchstabe oder der Punkt sein muss. Beginnt ein Name mit einem Punkt, so dürfen nicht alle folgenden Zeichen Ziffern sein. Hierdurch erzeugt man nämlich eine Zahl.

Wir nennen die Variable `lp` und geben ein

```
> lp<-c(22,30,16,25,27)
```

Eine Variable bleibt während der gesamten Sitzung im **Workspace** erhalten, wenn sie nicht mit dem Befehl `rm` gelöscht wird. Beim Verlassen von `R` durch `q()` wird man gefragt, ob man den Workspace sichern will. Antwortet man mit ja, so sind auch bei der nächsten Sitzung alle Variablen vorhanden. Mit der Funktion `ls` kann man durch den Aufruf `ls()` alle Objekte im Workspace auflisten.

`q`

`ls`

```
> ls()
[1] "lp"
```

Den Inhalt einer Variablen kann man sich durch Eingabe des Namens anschauen. Der Aufruf

```
> lp
```

liefert das Ergebnis

```
[1] 22 30 16 25 27
```

`R` unterscheidet Groß- und Kleinschreibung. Die Variablennamen `lp` und `LP` beziehen sich also auf unterschiedliche Objekte.

```
> LP
Fehler: objekt "LP" nicht gefunden
```

Die Preise der Langspielplatten sind in US Dollar. Am 15.5.2006 kostete ein US Dollar 0.774 EURO. Um die Preise in EURO umzurechnen, muss man jeden Preis mit 0.774 multiplizieren. Um alle Preise umzurechnen, multiplizieren den Vektor `lp` mit 0.774

```
> 0.774*lp
[1] 17.028 23.220 12.384 19.350 20.898
```

Um das Ergebnis auf zwei Stellen zu runden, benutzen wir die Funktion `round`:

```
> round(0.774*lp,2)
[1] 17.03 23.22 12.38 19.35 20.90
```

Die Portokosten betragen jeweils 12 US Dollar. Wir addieren zu jeder Komponente von lp die Zahl 12

```
> lp+12
[1] 34 42 28 37 39
```

Indizierung [ ]

Auf Komponenten eines Vektors greift man durch **Indizierung** zu. Hierzu gibt man den Namen des Vektors gefolgt von eckigen Klammern ein, zwischen denen die Nummer der Komponente oder der Vektor mit den Nummern der Komponenten steht, auf die man zugreifen will. Um den Preis der ersten Platte zu erfahren, gibt man ein:

```
> lp[1]
[1] 22
```

length

Um den Preis der Platte zu erhalten, die man zuletzt gekauft hatte, benötigt man die Länge des Vektors `lp`. Diesen liefert die Funktion `length`.

```
> length(lp)
[1] 5
> lp[length(lp)]
[1] 27
```

Wir können auch gleichzeitig auf mehrere Komponenten zugreifen:

```
> lp[c(1,2,3)]
[1] 22 30 16
```

:

Einen Vektor mit aufeinander folgenden natürlichen Zahlen erhält man mit dem Operator `:` . Schauen wir uns einige Beispiele an.

```
> 1:3
[1] 1 2 3
> 4:10
[1]  4  5  6  7  8  9 10
> 3:1
[1] 3 2 1
```

Wir können also auch

```
> lp[1:3]
[1] 22 30 16
```

eingeben.

sum

Schauen wir uns noch einige Funktionen an, mit denen man Informationen aus einem Vektor extrahieren kann. Die Summe aller Werte liefert die Funktion `sum`:

```
> sum(lp)
[1] 120
```

n

Das Minimum erhalten wir mit der Funktion `min`

```
> min(lp)
[1] 16
```

und das Maximum mit der Funktion `max` `max`

```
> max(lp)
[1] 30
```

Die Funktion `sort` sortiert einen Vektor aufsteigend. `sort`

```
> sort(lp)
[1] 16 22 25 27 30
```

Setzt man das Argument `decreasing` auf den Wert `TRUE`, so wird absteigend sortiert.

```
> sort(lp,decreasing=TRUE)
[1] 30 27 25 22 16
```

Wie gibt man die Daten bei einem qualitativen Merkmal ein? Beginnen wir auch hier mit einem Beispiel. Hier ist die Urliste des Geschlechts von 10 Teilnehmern eines Projektes:

```
w m w m w m m m w m
```

Wir geben die Urliste als Vektor ein, dessen Komponenten **Zeichenketten** sind. Eine Zeichenkette ist eine Folge von Zeichen, die in Hochkomma stehen. **Zeichenkette**
So sind `"Berlin"` und `"Bielefeld"` Zeichenketten.
Wir nennen den Vektor `Geschlecht`:

```
> Geschlecht<-c("w","m","w","m","w","m","m","m","w","m")
> Geschlecht
[1] "w" "m" "w" "m" "w" "m" "m" "m" "w" "m"
```

Mit der Funktion `factor` transformieren wir den Vektor `Geschlecht`, dessen Komponenten Zeichenketten sind, in einen Vektor, dessen Komponenten die Ausprägungen eines **Faktors**, also eines qualitativen Merkmals, sind `factor` **Faktor**

```
> Geschlecht<-factor(Geschlecht)
> Geschlecht
[1] w m w m w m m m w m
Levels: m w
```

Wir werden bald sehen, mit welchen Funktionen man Informationen aus Vektoren vom Typ `factor` extrahieren kann. Hier wollen wir nur zeigen, dass man diese wie auch Vektoren, deren Komponenten numerisch sind, indizieren kann.

```
> Geschlecht[2]
[1] m
Levels: m w
> Geschlecht[5:length(Geschlecht)]
[1] w m m m w m
Levels: m w
```

Bisher haben wir nur ein Merkmal betrachtet. Wir wollen nun zeigen, wie man vorgeht, wenn mehrere Merkmale eingegeben werden sollen. Hierbei gehen wir zunächst davon aus, dass alle Merkmale den gleichen Typ besitzen, also entweder alle quantitativ oder alle qualitativ sind. Wir illustrieren die Vorgehensweise an einem Beispiel.
Bei einer Befragung gaben zwei Personen ihr Alter, das Alter ihrer Mutter und das Alter ihres Vaters an. Die Daten sind in Tabelle 2.1 zu finden.

Tabelle 2.1: Alter

Alter	Alter der Mutter	Alter des Vaters
29	58	61
26	53	54

Matrix

Liegen die Daten wie in Tabelle 2.1 vor, so sollte man sie als **Matrix** eingeben. Eine Matrix ist ein rechteckiges Zahlenschema, das aus r Zeilen und s Spalten besteht.

`matrix`

In `R` erzeugt man eine Matrix mit der Funktion `matrix`. Der Aufruf der Funktion `matrix` ist

```
matrix(data,nrow=1,ncol=1,byrow=F)
```

Dabei ist `data` der Vektor mit den Elementen der Matrix. Das Argument `nrow` gibt die Anzahl der Zeilen und das Argument `ncol` die Anzahl der Spalten der Matrix an. Standardmäßig wird eine Matrix spaltenweise eingegeben. Wir geben also ein:

```
> alter<-matrix(c(29,26,58,53,61,54),2,3)
```

```
> alter
     [,1] [,2] [,3]
[1,]   29   58   61
[2,]   26   53   54
```

Sollen die Zeilen aufgefüllt werden, so muss das Argument `byrow` auf den Wert `TRUE` gesetzt werden:

```
> alter<-matrix(c(29,58,61,26,53,54),2,3,TRUE)
> alter
     [,1] [,2] [,3]
[1,]   29   58   61
[2,]   26   53   54
```

Auf Elemente einer Matrix greifen wir wie auf Komponenten eines Vektors durch Indizierung zu, wobei wir die Informationen, die sich auf Zeilen beziehen, von den Informationen, die sich auf Spalten beziehen, durch Komma trennen. Um auf das Element in der ersten Zeile und zweiten Spalte zuzugreifen, geben wir also ein:

```
> alter[1,2]
[1] 58
```

Alle Elemente der ersten Zeile erhalten wir durch

```
> alter[1,]
[1] 29 58 61
```

und alle Elemente der zweiten Spalte durch

```
> alter[,2]
[1] 58 53
```

Die Summe aller Werte erhält man mit der Funktion `sum`:

```
> sum(alter)
[1] 281
```

Oft ist man an der Summe der Werte innerhalb der Zeilen oder Spalten interessiert. Diese liefern die Funktionen `colSums` und `rowSums`.

`colSums`
`rowSums`

```
> rowSums(alter)
[1] 148 133

> colSums(alter)
[1]  55 111 115
```

Man kann aber auch die Funktion `apply` anwenden. Diese wird aufgerufen durch `apply`

```
apply(x,margin,fun)
```

Diese wendet auf die Dimension `margin` der Matrix `x` die Funktion `fun` an. Dabei entspricht die erste Dimension den Zeilen und die zweite Dimension den Spalten. Die Summe der Werte in den Zeilen erhalten wir also durch

```
> apply(alter,1,sum)
[1] 148 133
```

und die Summe der Werte in den Spalten durch

```
> apply(alter,2,sum)
[1]  55 111 115
```

Wir können für `fun` natürlich auch andere Funktionen wie `min` oder `max` verwenden.
Einen Vektor mit den Zeilenminima liefert der Aufruf

```
> apply(alter,1,min)
[1] 29 26
```

und einen Vektor mit den Spaltenmaxima der Aufruf

```
> apply(alter,2,max)
[1] 29 58 61
```

Jetzt schauen wir uns an, wie man Datensätze abspeichert, die sowohl qualitative als auch quantitative Merkmale enthalten. Wir betrachten wieder ein Beispiel.
Bei einer Befragung wurden das Geschlecht und das Alter von drei Personen erhoben. Die Daten sind in Tabelle 2.2 zu finden.

Tabelle 2.2: Alter

Geschlecht	Alter
m	29
w	26
m	24

In R bieten **Datentabellen** die Möglichkeit, die Werte von Merkmalen unterschiedlichen Typs in einer Variablen abzuspeichern. Dabei muss bei jedem Merkmal die gleiche Anzahl von Beobachtungen vorliegen. Eine Datentabelle wird mit dem Befehl `data.frame` erzeugt. Das Beispiel illustriert die Vorgehensweise. `data.frame`

```
> sexage<-data.frame(sex=c("m","w","m"),age=c(29,26,24))
> sexage
  sex age
1   m  29
2   w  26
3   m  24
```

Auf eine Datentabelle kann man wie auf eine Matrix zugreifen.

```
> sexage[2,2]
[1] 26
> sexage[2,]
 sex age
2   w  26
> sexage[,1]
[1] m w m
Levels: m w
```

Der letzte Aufruf zeigt, dass ein Vektor, der aus Zeichenketten besteht, bei der Erzeugung einer Datentabelle automatisch zu einem Faktor wird.

Datentabellen sind **Listen**, die wie Matrizen behandelt werden können. Wir wollen uns hier nicht detailliert mit Listen beschäftigen, sondern nur darauf hinweisen, dass Listen aus Komponenten bestehen, von denen jede einen anderen Typ aufweisen kann. So kann die erste Komponente einer Liste eine Zeichenkette, die zweite ein Vektor und die dritte eine Matrix sein. Auf die Komponenten einer Liste greift man entweder mit einer doppelten eckigen Klammer oder mit Name der Liste Name der Komponente zu. **Liste**

```
> sexage[[1]]
1] m w m
Levels: m w
> sexage$sex
[1] m w m
Levels: m w
> sexage[[2]]
[1] 29 26 24
```

```
> sexage$age
[1] 29 26 24
```

attach

detach

Mit der Funktion `attach` kann man auf die in einer Datentabelle enthaltenen Variablen unter ihrem Namen zugreifen, ohne den Namen der Datentabelle zu verwenden. Mit der Funktion `detach` hebt man diese Zugriffsmöglichkeit auf.

```
> attach(sexage)
> sex
[1] m w m
Levels: m w

> age
[1] 29 26 24

> detach(sexage)
> sex
Fehler: objekt "sex" nicht gefunden
> age
Fehler: objekt "age" nicht gefunden
```

Wir werden in diesem Skript immer wieder mit der Datentabelle `weiterbildung` arbeiten, die Daten aus Tabelle 1.2 auf Seite 17 enthält und folgendermaßen aufgebaut ist:

```
 Geschlecht Alter Film Bewertung Geld Satz
          m    30    n      <NA>  1.8    n
          w    23    j         g  1.8    n
          w    26    j         g  1.8    j
          m    33    n      <NA>  2.8    n
          m    37    n      <NA>  1.8    n
          m    28    j         g  2.8    j
          w    31    j        sg  2.8    n
          m    23    n      <NA>  0.8    n
          w    24    j        sg  1.8    j
          m    26    n      <NA>  1.8    n
          w    23    j        sg  1.8    j
          m    32    j         g  1.8    n
          m    29    j        sg  1.8    j
          w    25    j         g  1.8    j
          w    31    j         g  0.8    n
```

```
m    26    j        g  2.8    n
m    37    n     <NA>  3.8    n
m    38    j        g   NA    n
w    29    n     <NA>  3.8    n
w    28    j       sg  1.8    n
w    28    j        m  2.8    j
w    28    j       sg  1.8    j
w    38    j        g  2.8    n
w    27    j        m  1.8    j
m    27    n     <NA>  2.8    j
```

2.3 Einlesen von Daten aus externen Dateien

Oft liegen die Daten außerhalb von R in einer **Datei** vor. In diesem Fall müssen sie nicht noch einmal eingeben werden, sondern können eingelesen werden. Wir gehen im Folgenden davon aus, dass die Daten aus Tabelle 1.3 auf Seite 18 in einer **ASCII-Datei** gespeichert wurden. Sie sieht folgendermaßen aus

Datei

```
Geschlecht Alter Mutter Vater Geschwister
m          29    58     61    1
w          26    53     54    2
m          24    49     55    1
w          25    56     63    3
w          25    49     53    0
w          23    55     55    2
m          23    48     54    2
m          27    56     58    1
m          25    57     59    1
m          24    50     54    1
w          26    61     65    1
m          24    50     52    1
m          29    54     56    1
m          28    48     51    2
w          23    52     52    1
m          24    45     57    1
w          24    59     63    0
w          23    52     55    1
m          24    54     61    2
w          23    54     55    1
```

read.table

Die Daten mögen auf dem Laufwerk `d:` im Verzeichnis (Ordner) `daten` in der Datei `bidaten.txt` stehen. Wir lesen sie mit der Funktion `read.table` ein. Diese besitzt eine Vielzahl von Argumenten, von denen nur der Dateiname obligatorisch ist. Zu diesem gehört die vollständige Pfadangabe. Dabei müssen für jeden Backslash zwei Backslash eingegeben werden, da in `R` der Backslash in einer Zeichenkette ein Steuerzeichen ist.

Stehen in der Kopfzeile der Datei die Namen der Variablen, so muss das Argument `header` auf den Wert `TRUE` gesetzt werden. Ansonsten wird unterstellt, dass keine Kopfzeile existiert.

Wird bei Dezimalzahlen das Dezimalkomma verwendet, so setzt man das Argument `dec` auf den Wert `","`. Standardmäßig wird der Dezimalpunkt verwendet.

Mit dem Argument `sep` kann man festlegen, durch welches Zeichen Spalten getrennt sind, wobei unterstellt wird, dass das Leerzeichen verwendet wird.

Wir lesen die Daten ein und weisen sie der Variablen `bidaten` zu.

```
> bidaten<-read.table("d:\\daten\\bidaten.txt",header=TRUE)
> bidaten
   Geschlecht Alter Mutter Vater Geschwister
 1          m    29     58    61           1
 2          w    26     53    54           2
 3          m    24     49    55           1
 4          w    25     56    63           3
 5          w    25     49    53           0
 6          w    23     55    55           2
 7          m    23     48    54           2
 8          m    27     56    58           1
 9          m    25     57    59           1
10          m    24     50    54           1
11          w    26     61    65           1
12          m    24     50    52           1
13          m    29     54    56           1
14          m    28     48    51           2
15          w    23     52    52           1
16          m    24     45    57           1
17          w    24     59    63           0
18          w    23     52    55           1
19          m    24     54    61           2
20          w    23     54    55           1
```

Es wird eine Datentabelle erzeugt, auf die wir auf die im letzten Kapitel beschriebene Art und Weise zugreifen können.

```
> attach(bidaten)

     The following object(s) are masked _by_ .GlobalEnv :

      Geschlecht
> Geschlecht
[1] w m w m w m m m w m
Levels: m w
```

Wir sehen, dass wir vorsichtig sein müssen. Auf Seite 25 haben wir eine Variable `Geschlecht` erzeugt. Die Datentabelle `bidaten` enthält eine Variable mit dem gleichen Namen. Nach Eingabe des Befehls `attach(bidaten)` stehen uns unter dem Namen `Geschlecht` die Daten der zuerst erzeugten Variablen zur Verfügung. Wir nennen diese `Ges`. Wenn wir danach noch die Variable `Geschlecht` mit dem Befehl `rm` löschen, können wir auf die Variable `Geschlecht` aus der Datentabelle `bidaten` zugreifen.

```
> Ges<-Geschlecht
> rm(Geschlecht)
> Geschlecht
[1] m w m w w w m m m m w m m m w m w w m w
Levels: m w
```

Man kann die Daten aus der **Zwischenablage** einlesen. Hierzu wählt man als Dateinamen `"clipboard"`. Dies ist vor allem dann sinnvoll, wenn man Datensätze aus dem Internet einliest. Man markiert die Daten und kopiert sie in die Zwischenablage. Mit `read.table("clipboard")` werden sie in `R` eingelesen.

2.4 Selektion unter Bedingungen

Bei der Datenanalyse werden oft Gruppen hinsichtlich eines oder mehrerer Merkmale verglichen. So könnte bei den Daten aus Tabelle 1.3 auf Seite 18 interessieren, ob sich das Alter der Studenten vom Alter der Studentinnen unterscheidet. Um diese Frage beantworten zu können, müssen wir zum einen die Werte des Alters selektieren, bei denen das Merkmal `Geschlecht` den Wert `w` aufweist, und zum anderen die Werte des Merkmals `Alter` selektieren, bei denen das Merkmal Geschlecht den Wert `m` aufweist. Wir müssen also überprüfen, welche Komponenten eines Vektors eine **Bedingung** erfüllen.

Um Bedingungen zu überprüfen, kann man in R die **Operatoren**

==	gleich
!=	ungleich
<	kleiner
<=	kleiner oder gleich
>	größer
>=	größer oder gleich

Mit diesen Operatoren vergleicht man zwei Objekte. Schauen wir uns die Wirkung der Operatoren beim Vergleich von zwei Zahlen an.

```
> 3<4
[1] TRUE
> 3>4
[1] FALSE
```

TRUE
FALSE

Wir sehen, dass der Vergleich den Wert TRUE liefert, wenn die Bedingung wahr ist, ansonsten liefert er den Wert FALSE. Man kann auch Vektoren mit Skalaren vergleichen. Das Ergebnis ist in diesem Fall ein Vektor, dessen Komponenten TRUE sind, bei denen die Bedingung erfüllt ist. Ansonsten sind die Komponenten FALSE.
Wir betrachten die Variable lp von Seite 22.

```
> lp
[1] 22 30 16 25 27
> lp >= 25
[1] FALSE  TRUE FALSE  TRUE  TRUE
```

Man spricht auch von einem **logischen Vektor**. Wenn wir einen gleichlangen Vektor x mit einem logischen Vektor l durch x[l] indizieren, so werden aus x alle Komponenten ausgewählt, die in l den Wert TRUE annehmen. Der Aufruf

```
> lp[lp >= 25]
[1] 30 25 27
```

liefert also die Preise der Langspielplatten, die mindestens 25 US Dollar gekostet haben. Wenn wir wissen wollen, welche dies sind, so geben wir ein

```
> (1:length(lp))[lp >= 25]
[1] 2 4 5
```

ich

Dieses Ergebnis hätten wir auch mit der Funktion `which` erhalten.

```
> which(lp>=25)
[1] 2 4 5
```

Mit den Funktionen `any` und `all` kann man überprüfen, ob mindestens eine Komponente oder alle Komponenten eines Vektors eine Bedingung erfüllen. `any` `all`

```
> any(lp > 30)
[1] FALSE
> all(lp <= 30)
[1] TRUE
```

Zur Überprüfung von mindestens zwei Bedingungen dienen die Operatoren `&` und `|`. Der Operator `&` liefert genau dann das Ergebnis `TRUE`, wenn beide Bedingungen wahr sind, während dies beim Operator `|` der Fall ist, wenn mindestens eine Bedingung wahr ist.

```
> lp[lp < 30 & lp > 25]
[1] 27
> lp[lp < 30 | lp > 25]
[1] 22 30 16 25 27
```

Versuchen wir nun die auf Seite 33 gestellte Aufgabe zu lösen. Wir wollen aus der Datentabelle `bidaten` auf Seite 32 das Alter der Studentinnen und das Alter der Studenten auswählen. Die Daten mögen in `R` in der Datei `bidaten` stehen, die auf Seite 32 zu finden ist. Mit dem bisher gelernten erreichen wir das folgendermaßen:

```
> attach(bidaten)
> alter.w<-Alter[Geschlecht=="w"]
> alter.w
[1] 26 25 25 23 26 23 24 23 23
> alter.m<-Alter[Geschlecht=="m"]
> alter.m
[1] 29 24 23 27 25 24 24 29 28 24 24
```

Mit der Funktion `split` gelangen wir auch zum Ziel. `split`

```
> split(Alter,Geschlecht)
$m
 [1] 29 24 23 27 25 24 24 29 28 24 24

$w
 [1] 26 25 25 23 26 23 24 23 23
```

Die Funktion `split` erstellt eine Liste, deren erste Komponente das Alter der Studenten und deren zweite Komponente das Alter der Studentinnen enthält.

```
> alter.wm<-split(Alter,Geschlecht)
> alter.wm[[1]]
[1] 29 24 23 27 25 24 24 29 28 24 24
> alter.wm[[2]]
[1] 26 25 25 23 26 23 24 23 23
```

Auf die Komponenten dieser Liste können wir mit Hilfe der Funktionen `lapply` und `sapply` Funktionen anwenden.

lapply

Beide Funktionen werden folgendermaßen aufgerufen:

lapply

```
lapply(X,FUN)
sapply(X,FUN)
```

Dabei ist `X` eine Liste und `FUN` eine Funktion wie `min`, `max` oder `sort`.

Das Ergebnis von `lapply` ist eine Liste, deren i-te Komponente das Ergebnis enthält, das man erhält, wenn man die Funktion `FUN` auf die i-te Komponente der Liste `X` anwendet.

Das Ergebnis von `sapply` ist ein Vektor, falls das Ergebnis der Funktion `FUN` ein Skalar ist. Die i-te Komponente dieses Vektors enthält das Ergebnis, das man erhält, wenn man die Funktion `FUN` auf die i-te Komponente der Liste `X` anwendet.

Ist das Ergebnis der Funktion `FUN` ein Vektor mit einer festen Länge, so ist das Ergebnis von `sapply` ist eine Matrix, deren i-te Zeile das Ergebnis enthält, das man erhält, wenn man die Funktion `FUN` auf die i-te Komponente der Liste `X` anwendet.

Ansonsten sind die Ergebnisse der Funktionen `lapply` und `sapply` identisch.

Wollen wir das Minimum des Alters der männlichen und der weiblichen Teilnehmer bestimmen, so geben wir ein

```
> lapply(split(Alter,Geschlecht),min)
$m
[1] 23

$w
 [1] 23

> sapply(split(Alter,Geschlecht),min)
 m  w
23 23
```

Bei den geordneten Datensätzen des Alters der Frauen und Männer liefern `lapply` und `sapply` identische Ergebnisse.

```
> lapply(split(Alter,Geschlecht),sort)
$m
 [1] 23 26 26 27 28 29 30 32 33 37 37 38

$w
 [1] 23 23 24 25 26 27 28 28 28 29 31 31 38

> sapply(split(Alter,Geschlecht),sort)
$m
 [1] 23 26 26 27 28 29 30 32 33 37 37 38

$w
 [1] 23 23 24 25 26 27 28 28 28 29 31 31 38
```

Eine weitere Möglichkeit zur Auswahl von Teilmengen einer Datentabelle bietet der Befehl `subset`. Der Aufruf `subset`

```
subset(x,condition)
```

wählt aus der Datentabelle `x` die Zeilen aus, die die Bedingung `condition` erfüllen. Die Daten aller Studentinnen aus der Datei `bidaten` erhalten wir durch

```
> subset(bidaten,Geschlecht=="w")
   Geschlecht Alter Mutter Vater Geschwister
  2          w    26     53    54           2
  4          w    25     56    63           3
  5          w    25     49    53           0
  6          w    23     55    55           2
 11          w    26     61    65           1
 15          w    23     52    52           1
 17          w    24     59    63           0
 18          w    23     52    55           1
 20          w    23     54    55           1
```

und das Alter der Mütter der Studentinnen durch

```
> subset(bidaten,Geschlecht=="w",select=Mutter)
   Mutter
  2      53
```

```
  4      56
  5      49
  6      55
 11      61
 15      52
 17      59
 18      52
 20      54
```

Man kann natürlich auch mehr als eine Bedingung angeben. Alle Studentinnen, die keine Geschwister haben, erhält man durch

```
> subset(bidaten,Geschlecht=="w" & Geschwister==0)
    Geschlecht Alter Mutter Vater Geschwister
 5           w    25     49    53           0
17           w    24     59    63           0
```

2.5 Grafiken in R

R bietet eine Reihe von Möglichkeiten, eine Grafik zu erstellen, von denen wir in diesem Skript eine Vielzahl kennen lernen werden. Wir wollen hier zunächst eine relativ einfache Grafik erstellen und betrachten folgende Funktion, die auch auf Seite 80 zu finden ist.

$$F_n^*(x) = \begin{cases} 0 & \text{für } x < 20 \\ -0.8 + 0.04 \cdot x & \text{für } 20 \leq x \leq 25 \\ -2.2 + 0.096 \cdot x & \text{für } 25 < x \leq 30 \\ -0.28 + 0.032 \cdot x & \text{für } 30 < x \leq 35 \\ -0.28 + 0.032 \cdot x & \text{für } 35 < x \leq 40 \\ 1 & \text{für } x > 40 \end{cases} \tag{2.1}$$

Diese Funktion ist stückweise linear. Auf jedem Teilintervall müssen wir also eine Strecke zeichnen. Wir betrachten zunächst das Intervall $[20, 25]$. Hier lautet die Gleichung

$$F_n^*(x) = -0.8 + 0.04 \cdot x$$

Um eine Strecke zeichnen zu können, benötigen wir beide Endpunkte. Wir bestimmen $F_n^*(x)$ für $x = 20$ und $x = 25$. Es gilt

$$F_n^*(20) = -0.8 + 0.04 \cdot 20 = 0$$

und

$$F_n^*(25) = -0.8 + 0.04 \cdot 25 = 0.2$$

Wir zeichnen also eine Strecke durch die Punkte $(20, 0)$ und $(25, 0.2)$. Hierzu benutzen wir die Funktion `plot`. Diese benötigt als Argumente die gleich langen Vektoren `x` und `y`. Der Aufruf

```
plot(x,y)
```

zeichnet die Punkte `(x[1],y[1])` und `(x[2],y[2])` in einem kartesischen Koordinatensystem. Wir geben also ein

```
> plot(c(20,25),c(0,0.2))
```

In Abbildung 2.1 auf dieser Seite ist diese Grafik links oben zu finden.

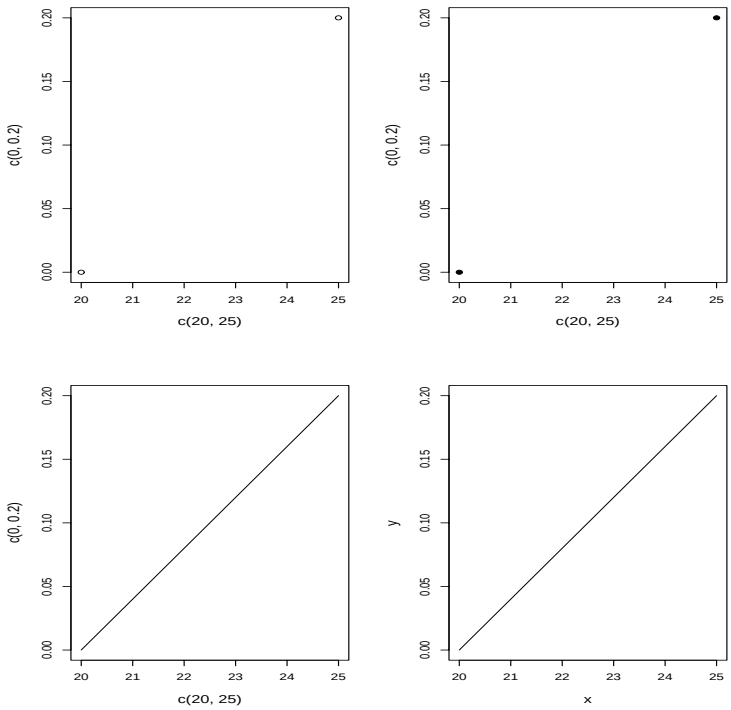


Abbildung 2.1: 4 Grafiken

Vier Bilder in einer Grafik erhält man durch

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

Die Punkte in der Grafik in Abbildung 2.1 auf der vorherigen Seite links oben sind offen. Sollen sie ausgemalt sein, so muss man das Argument `pch` auf den Wert 16 setzen. Dabei steht `pch` für plot character.

pch

```
> plot(c(20,25),c(0,0.2),pch=16)
```

Das Ergebnis ist in der Grafik rechts oben in Abbildung 2.1 auf der vorherigen Seite zu finden.

Die Größe der Achsenbeschriftung legt man mit dem Argument `cex.axis` fest. Dieses nimmt standardmäßig den Wert 1 an.

cex.axis

Sollen nicht die Punkte sondern die Strecke gezeichnet werden, so müssen wir das Argument `type` auf den Wert `"l"` setzen.

type

```
> plot(c(20,25),c(0,0.2),type="l")
```

Diese Grafik ist links unten in Abbildung 2.1 auf der vorherigen Seite zu finden. Der Standardwert von `type` ist `"p"`. Setzt man diesen auf den Wert `"o"`, so werden sowohl die Strecke als auch die Punkte gezeichnet.

Die Beschriftung der Abszisse und Ordinate ist unschön. Mit den Argumenten `xlab` und `ylab` legen wir die gewünschte Beschriftung als Zeichenketten fest.

xlab
ylab

```
> plot(c(20,25),c(0,0.2),type="l",xlab="x",ylab="y")
```

Diese Grafik ist rechts unten in Abbildung 2.1 auf der vorherigen Seite zu finden. Das gleiche Ergebnis können wir auch folgendermaßen erreichen:

```
> x<-c(20,25)
> y<-c(0,0.2)
> plot(x,y,type="l")
```

cex.lab

Die Größe der Buchstaben legt man mit dem Argument `cex.lab` fest. Dieses nimmt standardmäßig den Wert 1 an.

In den USA ist es üblich, dass die Beschriftung der Achsen parallel zu den Achsen gewählt wird. Dies ist auch Standard in R. Soll die Beschriftung der Ordinate orthogonal zu dieser Achse sein, so muss man eingeben

las

```
> par(las=1)
```

Diese Einstellung bleibt während der Sitzung mit R erhalten. Nach Eingabe dieses Befehls erhält man durch

```
> plot(x,y,type="l")
```

die Grafik links oben in Abbildung 2.2 auf dieser Seite.

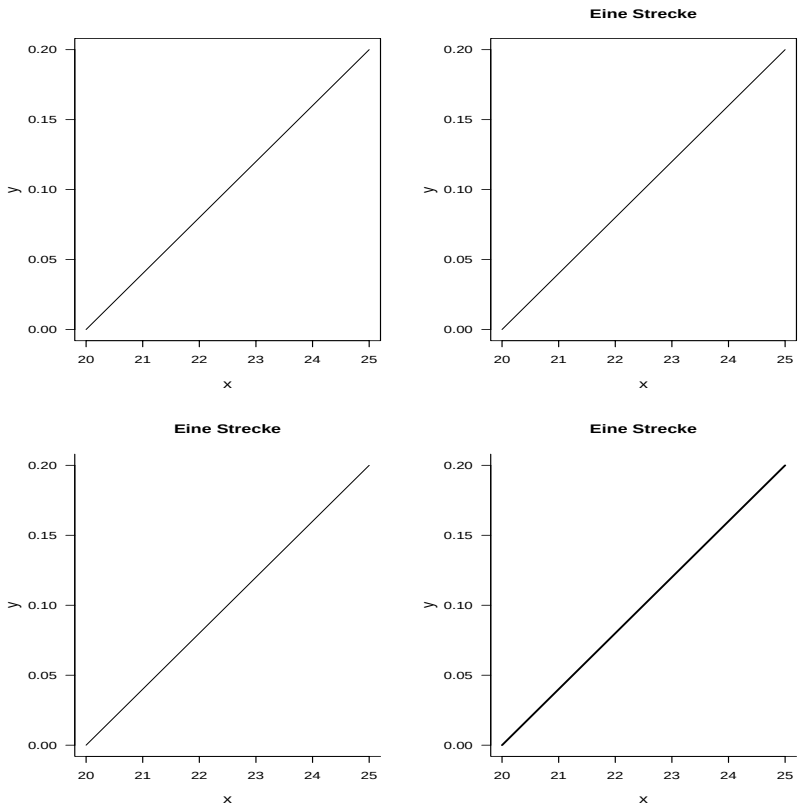


Abbildung 2.2: 4 Grafiken

Wir können über das Argument `main` eine Überschrift als Zeichenkette hinzufügen: `main`

```
> plot(x,y,type="l",main="Eine Strecke")
```

In Abbildung 2.2 auf dieser Seite ist diese Grafik rechts oben zu finden.
Standardmäßig wird um die Grafik eine Box gezeichnet. Soll diese nur auf Höhe der der Abszisse und Ordinate erstellt werden, so muss man das Argument `bty` auf den Wert `"l"` setzen. `bty`

```
> plot(x,y,type="l",main="Eine Strecke",bty="l")
```

Diese Grafik ist in Abbildung 2.2 auf dieser Seite links unten zu finden.
Standardmäßig nimmt `bty` den Wert `"o"` an.

Die Dicke der Linien legt man über das Argument `lwd` fest, das standardmäßig den Wert 1 annimmt. Doppelt so breite Linien erhält man durch: lwd

```
> plot(x,y,type="l",main="Eine Strecke",bty="l",lwd=2)
```

In Abbildung 2.2 auf der vorherigen Seite ist diese Grafik rechts unten zu finden.
Nun wollen wir die Funktion aus Gleichung (2.1) auf Seite 38 im Intervall $[20, 40]$ zeichnen. Die ersten Koordinaten der Punkte sind

$$x_1 = 20 \quad x_2 = 25 \quad x_3 = 30 \quad x_4 = 35 \quad x_5 = 40$$

und die zugehörigen zweiten Koordinaten sind

$$y_1 = 0 \quad y_2 = 0.2 \quad y_3 = 0.68 \quad y_4 = 0.84 \quad y_5 = 1$$

Übergibt man der Funktion `plot` die Vektoren `x` und `y`, die beide n Komponenten besitzen, so werden die Punkte `(x[1],y[1])` und `(x[2],y[2])`, `(x[2],y[2])` und `(x[3],y[3])`... `(x[n-],y[n-1])` und `(x[n],y[n])` durch Geraden verbunden.
Einen Vektor mit den Zahlen $20, 25, 30, 35, 40$ erhalten wir am einfachsten mit der Funktion `seq`. Diese wird folgendermaßen aufgerufen seq

```
seq(from, to, by)
```

Es wird eine Zahlenfolge von `from` bis `to` im Abstand `by` erzeugt. Wir geben also ein

```
> x<-seq(20,40,5)
> x
[1] 20 25 30 35 40
```

Wir erstellen noch den Vektor `y`

```
> y<-c(0,0.2,0.68,0.84,1)
```

und zeichnen die Funktion

```
> plot(x,y,type="l",bty="l")
```

In Abbildung 2.3 auf der nächsten Seite ist diese Grafik links oben zu finden.

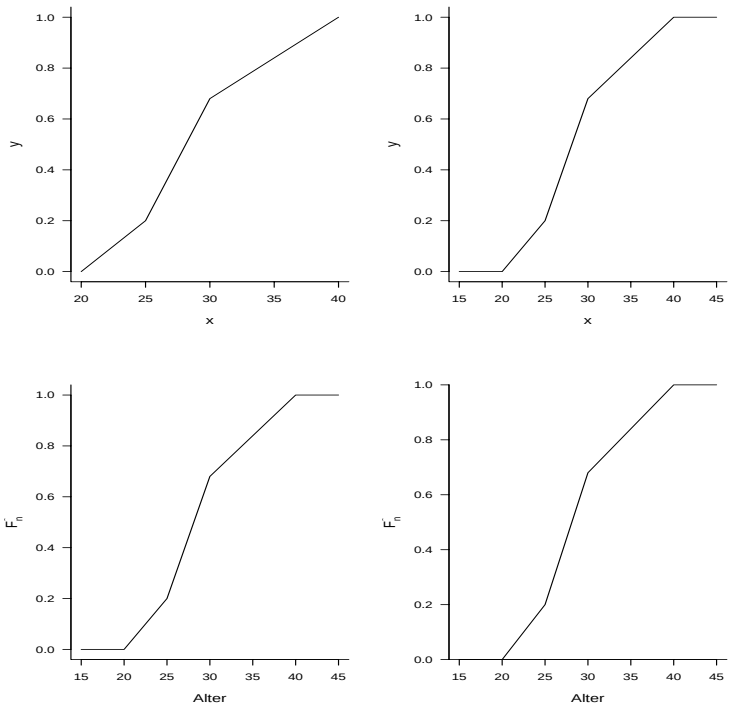


Abbildung 2.3: 4 Grafiken

Nun müssen wir noch den Bereich $x < 20$ und $y > 40$ berücksichtigen. Wir beginnen bei $x = 15$ und enden bei $x = 45$. Wir müssen also noch die Punkte $(15, 0)$ und $(45, 1)$ hinzufügen.

```
> x<-c(15,x,45)
> x
[1] 15 20 25 30 35 40 45
> y<-c(0,y,1)
> y
[1] 0.00 0.00 0.20 0.68 0.84 1.00 1.00
> plot(x,y,type="l",bty="l")
```

Diese Grafik ist rechts oben in Abbildung 2.3 auf dieser Seite zu finden.

Nun beschriften wir noch die Abszisse und die Ordinate mit den Argumenten `xlab` und `ylab`. An die Ordinate schreiben wir F_n^*. Dies ist eine Formel, die wir mit der Funktion `expression` erstellen. Ein tiefer gestelltes Zeichen gewinnt man, indem man es in eckige Klammern setzt, und ein höher gestelltes durch `"^"`. Beispiele für die Erstellung von Formeln erhält man durch

`expression`

help(text).

```
> plot(x,y,type="l",bty="l",xlab="Alter",
      ylab=expression(F[n]^"*"))
```

In Abbildung 2.3 auf der vorherigen Seite ist diese Grafik links unten zu finden.
Standardmäßig wird zwischen der Ordinate und dem Beginn der Kurve ein Zwischenraum gelassen. Diesen entfernen wir, indem wir den Parameter `xaxs` auf den Wert `"i"` setzen. Entsprechend gibt es den Parameter `yaxs`.

xaxs
yaxs

```
> plot(x,y,type="l",bty="l",xlab="Alter",
    ylab=expression(F[n]^"*"),yaxs="i")
```

In Abbildung 2.3 auf der vorherigen Seite ist diese Grafik rechts unten zu finden.
Wir wollen die Abbildung noch um die Gerade durch die Punkte $(20, 0)$ und $(40, 1)$ ergänzen. Hierzu benutzen wir die Funktion `lines`. Setzen wir das Argument `lty` auf den Wert 2, so wird eine gestrichelte Strecke gezeichnet.

lines
lty

```
> lines(c(20,40),c(0,1),lty=2,lwd=2)
```

Diese Gerade ist die Verteilungsfunktion der Gleichverteilung auf dem Intervall $[20, 40]$. Mit der Funktion `legend` fügen wir noch eine Legende hinzu.

legend

```
> legend(15,1,c("Daten","Gleichverteilung"),lty=1:2)
```

In Abbildung 2.4 auf der nächsten Seite ist diese Grafik links oben zu finden.
Schauen wir uns noch einmal die Argumente `xaxs` und `yaxs` an. Die Grafik in Abbildung 2.1 links oben auf Seite 39 zeigt, warum eine Grafik in R nicht bei den Minima der Beobachtungen beginnt und bei den Maxima endet, wenn man eine Punktewolke zeichnet. Diese Punkte werden dann nämlich an den Rand gedrängt. Dies ist in der Grafik rechts oben in Abbildung 2.4 auf der nächsten Seite der Fall, in der wir `xaxs` und `yaxs` auf den Wert `"i"` gesetzt haben:

```
> plot(c(20,25),c(0,0.2),xlab="x",ylab="y",xaxs="i",yaxs="i")
```

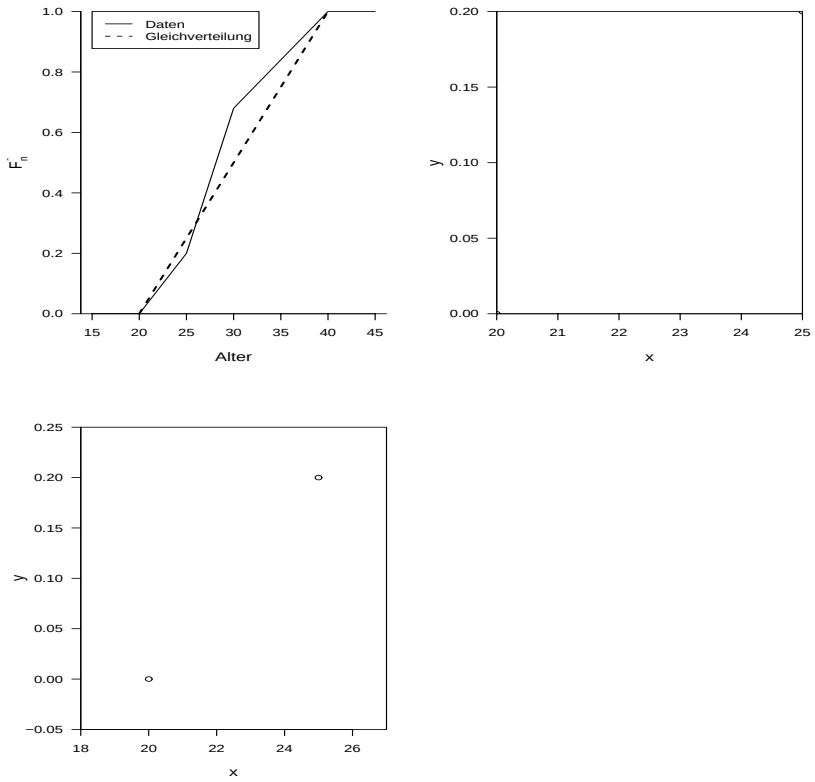


Abbildung 2.4: 3 Grafiken

Wir können den Bereich der Grafik durch die Argumente `xlim` und `ylim` festlegen. Die Grafik links unten in Abbildung 2.4 auf dieser Seite erhalten wir durch

```
> plot(c(20,25),c(0,0.2),xlab="x",ylab="y",
      xaxs="i",yaxs="i",xlim=c(18,27),ylim=c(-0.05,0.25))
```

Will man eine Funktion zeichnen, so kann man durch die Argumente `xaxs` und `yaxs` die Grafik verschönern. Schauen wir uns dies an einem Beispiel an. Wir wollen die folgende Funktion

$$\phi(x) = \frac{1}{\sqrt{2 \cdot \pi}} e^{-0.5 \cdot x^2}$$

im Intervall $[-4, 4]$ zeichnen. Es handelt sich um die Dichtefunktion der Standardnormalverteilung. Mit dieser werden wir uns in Kapitel 10.2.2 beschäftigen.

Die Zahl π erhält man in R durch

```
> pi
[1] 3.141593
```

pi

und die Exponentialfunktion mit der Funktion exp

```
> exp(1)
[1] 2.718282
```

exp

Die Dichtefunktion der Standardnormalverteilung in $x = -2, -1, 0, 1, 2$ erhalten wir also durch

```
> 1/sqrt(2*pi)*exp(-0.5*(-2:2)^2)
[1] 0.05399097 0.24197072 0.39894228 0.24197072 0.05399097
```

curve

Mit der Funktion curve können wir die Dichtefunktion der Standardnormalverteilung folgendermaßen zeichnen

```
> curve(1/sqrt(2*pi)*exp(-0.5*x^2),from=-4,to=4)
```

Die obere Grafik in Abbildung 2.5 auf dieser Seite zeigt das Bild.

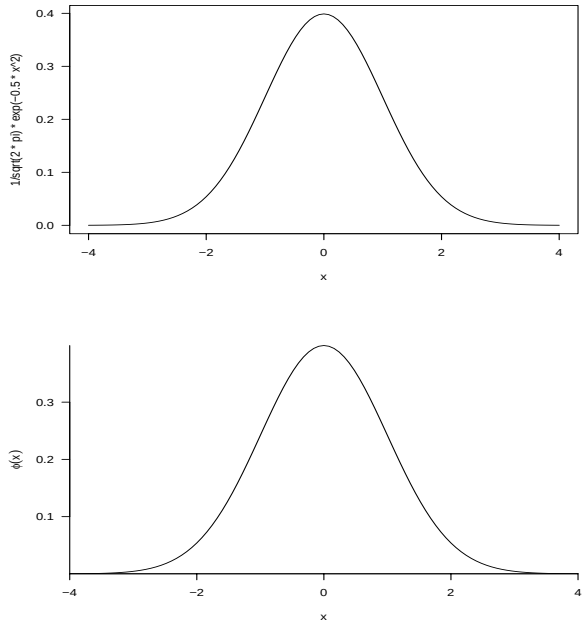


Abbildung 2.5: Dichtefunktion der Standardnormalverteilungf

Hier ist es sinnvoll, `xaxs` und `yaxs` auf den Wert `"i"` zu setzen. Außerdem beschriften wir noch die Ordinate und ändern die Box um die Grafik.

```
> curve(1/sqrt(2*pi)*exp(-0.5*x^2),from=-4,to=4,
       xaxs="i",yaxs="i",bty="l",ylab=expression(phi(x)))
```

Das Ergebnis ist in Abbildung 2.5 auf der vorherigen Seite unten zu finden.
Die Dichtefunktion der Standardnormalverteilung ist in `R` in der Funktion `dnorm` implementiert. Wir können also auch `dnorm`

```
> curve(dnorm,from=-4,to=4,xaxs="i",yaxs="i",
       bty="l",ylab=expression(phi(x)))
```

eingeben und erhalten das gleiche Ergebnis.

2.6 Wie schreibt man eine Funktion?

Im letzten Kapitel haben wir gesehen, mit welcher Befehlsfolge wir die Funktion aus Gleichung (2.1) auf Seite 38 grafisch darstellen können. Es handelt sich um die approximierende empirische Verteilungsfunktion, die wir ab Seite auf Seite 79 näher betrachten werden. Um diese zeichnen zu können, benötigen wir die Klassengrenzen $x_0^*, x_1^*, \ldots, x_k^*$ und die kumulierten relativen Häufigkeiten

$$F_n(x_0^*) = 0 \tag{2.2}$$

$$F_n(x_i^*) = \sum_{j=1}^{i} h_j \tag{2.3}$$

Durch Eingabe der in Kapitel 2.5 beschriebenen Befehlsfolge können wir für jeden Datensatz eine Grafik der approximierenden empirischen Verteilungsfunktion erstellen. Dies ist aber sehr mühselig. Wir können uns die Eingabe der Befehlsfolge ersparen, wenn wir eine Funktion schreiben, die diese Befehlsfolge ausführt. Funktionen bieten die Möglichkeit, eine Befehlsfolge unter einem Namen abzuspeichern und durch Aufruf des Namens der Funktion die für unterschiedliche Werte der Argumente auszuführen.
Eine Funktion wird in `R` durch folgende Befehlsfolge deklariert: `function`

```
fname<-function(Argumente} {
 Koerper der Funktion
 return(Ergebnis)
}
```

Eine Funktion benötigt einen Namen und Argumente. Im Körper der Funktion steht die Befehlsfolge. Der Befehl `return` bewirkt, dass die Funktion verlassen wird und das Argument von `return` als Ergebnis der Funktion zurückgegeben wird.

return

Wir nennen die Funktion `plot.ecdfapprox`. Es liegt nahe, die Koordinaten der Punkte, die durch Geraden verbunden werden sollen, als Argumente der Funktion `plot.ecdfapprox` zu wählen. Das sind die Klassengrenzen $x_0^*, x_1^*, \ldots, x_k^*$ und die kumulierten relativen Häufigkeiten

$$h_1 \quad h_1 + h_2 \quad \ldots \quad h_1 + h_2 + \ldots + h_k$$

Dem Anwender sind aber in der Regel eher die relativen Häufigkeiten

$$h_1, h_2, \ldots, h_k$$

der Klassen bekannt, aus denen er die kumulierten relativen Häufigkeiten bestimmt. Aus diesen bestimmt er dann die kumulierten relativen Häufigkeiten Diese Aufgabe soll die Funktion `plot.ecdfapprox` übernehmen. Sie erhält also als Argumente die Klassengrenzen und die relativen Häufigkeiten der Klassen. Wir nennen diese Argumente `grenzen` und `haeuf`. Wir schauen uns nun die Befehle an und beginnen mit dem Bereich $[x_0^*, x_k^*]$. Wir müssen zuerst die kumulierten relativen Häufigkeiten mit der Funktion `cumsum` bestimmen.

cumsum

```
chaeuf<-cumsum(haeuf)
```

An der Stelle `grenzen[1]` ist die approximierende empirische Verteilungsfunktion gleich 0. Wir ergänzen den Vektor `chaeuf` vorne um den Wert 0.

```
chaeuf<-c(0,chaeuf)
```

Nun müssen wir noch den Bereich vor x_0^* und hinter x_k^* berücksichtigen. Die approximierende empirische Verteilungsfunktion nimmt vor x_0^* den Wert 0 und hinter x_k^* den Wert 1 an.

```
chaeuf<-c(0,chaeuf,1)
```

Bei welchem Werten auf der Abszisse soll die Zeichnung beginnen und enden? Im Beispiel hatten wir die Werte 15 und 45 gewählt. Wir wollen dem Benutzer aber nicht zumuten, dass er diese Werte vor dem Aufruf der Funktion `plot.ecdfapprox` festlegt und der Funktion als Argument übergibt, sondern bestimmen sie innerhalb der Funktion. Wir berechnen die Breite des Intervalls

```
b<-grenzen[length(grenzen)]-grenzen[1]
```

und setzen

```
ug<-grenzen[1]-0.25*b
```

und

```
og<-grenzen[length(grenzen)]+0.25*b
```

Wir erweitern den Vektor `grenzen` um diese Größen

```
grenzen<-c(ug,grenzen,og)
```

und zeichnen mit `bty="l"` die Box nur auf Höhe der Abszisse und Ordinate. Außerdem starten wir die Kurve mit `xaxs="i"` bei der Ordinate und beschriften die Ordinate mit F_n^*. Wir rufen die Funktion `plot` also folgendermaßen auf:

```
plot(grenzen,chaeuf,type="l",bty="l",xaxs="i",
   ylab=expression(F^"*"[n]))
```

Für die Beschriftung der Abszisse wählen wir den Namen des Merkmals. Diesen geben wir der Funktion als Argument.
Schauen wir uns das alles noch für ein Beispiel an.

```
> grenzen
[1] 20 25 30 35 40
> haeuf
[1] 0.20 0.48 0.16 0.16
> chaeuf<-cumsum(haeuf)
> chaeuf
[1] 0.20 0.68 0.84 1.00
> chaeuf<-c(0,chaeuf)
> chaeuf
[1] 0.00 0.20 0.68 0.84 1.00
> chaeuf<-c(0,chaeuf,1)
> chaeuf
[1] 0.00 0.00 0.20 0.68 0.84 1.00 1.00
> b<-grenzen[length(grenzen)]-grenzen[1]
> b
[1] 20
> ug<-grenzen[1]-0.25*b
> ug
[1] 15
```

```
> og<-grenzen[length(grenzen)]+0.25*b
> og
[1] 45
> grenzen<-c(ug,grenzen,og)
> grenzen
[1] 15 20 25 30 35 40 45
> plot(grenzen,chaeuf,type="l",bty="l",xaxs="i",
     ylab=expression(F^"*"[n]))
```

Jetzt können wir die Funktion erstellen.

```
plot.ecdfapprox<-function(grenzen,haeuf,xname) {
chaeuf<-c(0,0,cumsum(haeuf),1)
b<-grenzen[length(grenzen)]-grenzen[1]
ug<-grenzen[1]-0.25*b
og<-grenzen[length(grenzen)]+0.25*b
grenzen<-c(ug,grenzen,og)
plot(grenzen,chaeuf,type="l",bty="l",xaxs="i",xlab=xname,
     ylab=expression(F^"*"[n]))
}
```

Jede Funktion sollte einen Kommentar enthalten, in dem die Argumente und der Output beschrieben werden. Der Kommentar steht hinter dem Zeichen `#`. Wir ergänzen die Funktion um Kommentare.

```
plot.ecdfapprox<-function(grenzen,haeuf,xname)
{ # grafische Darstellung der
  # approximierenden empirischen Verteilungsfunktion
  # Grenzen: Vektor mit den Klassengrenzen
  # haeuf: Vektor mit relativen Häufigkeiten der Klassen
  # xname: Name des Merkmals
  chaeuf<-c(0,0,cumsum(haeuf),1)
  b<-grenzen[length(grenzen)]-grenzen[1]
  ug<-grenzen[1]-0.25*b
  og<-grenzen[length(grenzen)]+0.25*b
  grenzen<-c(ug,grenzen,og)
  plot(grenzen,chaeuf,type="l",bty="l",xaxs="i",xlab=xname,
    ylab=expression(F^"*"[n]))
}
```

In der Funktion `plot.ecdfapprox` wird nicht überprüft, ob die Argumente der Funktion richtig gewählt wurden. So muss der Vektor `grenzen` eine Komponente mehr als der Vektor `haeuf` enthalten. Im Kapitel 2.4 auf

Seite 33 haben wir gelernt, wie man in R Bedingungen überprüft. In unserem Fall geben wir eine Fehlermeldung aus, wenn `length(grenzen)` ungleich `1+length(haeuf)` ist; ansonsten erstellen wir die Grafik. Man spricht auch von einer bedingten Anweisung.
In R setzt man bedingte Anweisungen mit dem Konstrukt

```
if(Bedingung){Befehlsfolge 1} else {Befehlsfolge 2}
```

um.
In unserem Fall enthält die Befehlsfolge 1 die Fehlermeldung, die am Bildschirm erscheinen soll. Wir ergänzen die Funktion um die Befehlsfolge

```
 if(length(grenzen)!=(1+length(haeuf)))
 {return("Fehler: Der Vektor grenzen muss um eine
  Komponente laenger sein als der Vektor haeuf"}
 else
```

Somit sieht die Funktion folgendermaßen aus.

```
plot.ecdfapprox<-function(grenzen,haeuf,xname) { # grafische
Darstellung der
 # approximierenden empirischen Verteilungsfunktion
 # grenzen: Vektor mit den Klassengrenzen
 # haeuf: Vektor mit relativen Häufigkeiten der Klassen
 # xname: Name des Merkmals
if(length(grenzen)!=(1+length(haeuf))) {return(cat("Fehler: Der
Vektor grenzen muss um eine Komponente
            laenger sein als der Vektor haeuf"))}
else { chaeuf<-c(0,0,cumsum(haeuf),1)
  b<-grenzen[length(grenzen)]-grenzen[1]
  ug<-grenzen[1]-0.25*b
  og<-grenzen[length(grenzen)]+0.25*b
  grenzen<-c(ug,grenzen,og)
  plot(grenzen,chaeuf,type="l",bty="l",xaxs="i",xlab=xname,
       ylab=expression(F^"*"[n]))
} }
```

Die Eingabe einer Funktionsdefinition wird in R durch die Funktion `fix` unterstützt. Nach dem Aufruf `fix`

```
fix(Name der Funktion)
```

landet man in einem Editor, der die Eingabe erleichtert.

2.7 Pakete

R ist ein offenes Programm, sodass es durch Funktionen, die von Benutzern erstellt wurden, erweitert werden kann. Diese Funktionen sind in Paketen (packages) enthalten. Um eine Funktion aus einem Paket benutzen zu können, muss man das Paket vcd installieren und laden. Man installiert ein Paket, indem man auf den Schalter

Pakete

und danach auf den Schalter

Installiere Paket(e)

klickt. Es öffnet sich ein Fenster mit einer Liste, in der man auf den Namen des Paketes klickt. Hierauf wird das Paket installiert. Dazu muss natürlich eine Verbindung zum Internet vorhanden sein.

Nachdem man

```
> library(Name des Paketes)
```

eingegeben hat, kann man die Funktionen des Paketes verwenden. Man muss ein Paket nur einmal installieren, muss es aber während jeder Sitzung einmal laden, wenn man es verwenden will.

Kapitel 3

Univariate Analyse

Statistik beschäftigt sich mit Populationen. In der beschreibenden Statistik betrachten wir alle Merkmalsträger einer Population und stellen die Verteilung eines oder mehrerer Merkmale dar. Die Verteilung eines Merkmals zeigt, wie sich die Merkmalsträger auf die einzelnen Merkmalsausprägungen verteilen. Man zählt also, wie oft die Merkmalsausprägungen in der Population vorkommen. Die Statistik wird deshalb auch die Lehre von den Verteilungen genannt.
In diesem Kapitel werden wir jeweils nur *ein* Merkmal betrachten. Man spricht auch von **univariater Datenanalyse**. Einige Aspekte der **multivariaten Datenanalyse** betrachten wir im nächsten Kapitel.

3.1 Darstellung univariater Datensätze

Bei jeder Datenanalyse will man den Datensatz übersichtlich darstellen. Ist die Anzahl der Beobachtungen gering, so reicht es aus, diese aufzuzählen. Bei vielen oder sehr vielen Beobachtungen liefert die Aufzählung keinen Überblick über die Struktur des Datensatzes. Hier sollte man eine Tabelle oder eine Grafik erstellen. Wir werden auf den folgenden Seiten lernen, wie man dabei vorzugehen hat.

3.1.1 Darstellung qualitativer Merkmale

Qualitative Merkmale zeichnen sich dadurch aus, dass die Merkmalsausprägungen Kategorien sind. Dabei sind die Kategorien bei einem nominalskalierten Merkmal ungeordnet und bei einem ordinalskalierten Merkmal geordnet. Nominal- und ordinalskalierte Merkmale werden bis auf eine7 Ausnahme auf

die gleiche Art und Weise dargestellt. Bei ordinalskalierten Merkmalen erlaubt die Ordnungsstruktur eine weitere Darstellung.

Wir betrachten ein qualitatives Merkmal mit k Merkmalsausprägungen, die wir mit $a_1, a_2, \ldots, a_k$ bezeichnen. Wird ein qualitatives Merkmal erhoben, so weist in der Regel jeder Merkmalsträger genau eine der Ausprägungsmöglichkeiten auf. Bei Befragungen werden aber oft Fragen gestellt, bei denen die Befragten mehr als eine der vorgegebenen Antwortmöglichkeiten ankreuzen können. Man spricht von **Mehrfachantworten**.

Bei Fragen mit Mehrfachantworten ordnet man jeder möglichen Antwort ein eigenes Merkmal mit den Merkmalsausprägungen `ja` und `nein` zu. Hierdurch ist sichergestellt, dass jeder Merkmalsträger bei jedem Merkmal genau eine der Merkmalsausprägungen aufweist.

Beispiel 7
Werden Studienanfänger nach Gründen für die Wahl ihres Studienfachs gefragt werden, so könnten folgende Antworten vorgegeben sein:

Gute Berufsaussichten	[ ]
Interesse	[ ]
eigene Fähigkeiten	[ ]

Es ist durchaus möglich, dass mehrere dieser Gründe für einen Studienanfänger ausschlaggebend waren.

Wir definieren die Merkmale `Gute Berufsaussichten`, `Interesse` und `eigene Fähigkeiten` mit den Merkmalsausprägungen `ja` und `nein`.

□

Bei der Datenanalyse interessiert uns, wie viele Merkmalsträger die einzelnen Merkmalsausprägungen aufweisen. Wir sprechen von der **absoluten Häufigkeit** $n(a_i)$ der Merkmalsausprägung a_i, $i = 1, \ldots, k$. Für $n(a_i)$ schreiben wir kurz n_i.

Ob der Wert einer absoluten Häufigkeit klein oder groß ist, hängt von der Anzahl n der Merkmalsträger ab. Sind 8 Personen von 10 Personen weiblich, so ist das viel; sind hingegen 8 Personen von 100 Personen weiblich, so ist dies wenig. Wir beziehen die absolute Häufigkeit einer Merkmalsausprägung auf die Anzahl der Merkmalsträger und erhalten die **relative Häufigkeit** dieser Merkmalsausprägung.

$$\text{relative Häufigkeit} = \frac{\text{absolute Häufigkeit}}{\text{Anzahl Merkmalsträger}}$$

Wir bezeichnen die relative Häufigkeit der i-ten Merkmalsausprägung mit

$f(a_i)$. Hierfür schreiben wir kurz f_i. Es gilt also

$$\boxed{f_i = \frac{n_i}{n}}$$

Die relative Häufigkeit einer Merkmalsausprägung ist genau dann gleich 0, wenn kein Merkmalsträger sie aufweist; sie ist hingegen genau dann gleich 1, wenn alle Merkmalsträger sie besitzen. Da dies die beiden Extremfälle sind, gilt für $i = 1, \ldots, k$:

$$\boxed{0 \leq f_i \leq 1}\,. \tag{3.1}$$

Außderdem gilt

$$f_1 + f_2 + \ldots + f_k = \sum_{i=1}^{k} f_i = \sum_{i=1}^{k} \frac{n_i}{n} = \frac{1}{n} \sum_{i=1}^{k} n_i = \frac{1}{n} \cdot n = 1$$

Wir haben hier das Summenzeichen Σ verwendet. Ab Seite 121 kann man lernen, wie man mit diesem umgeht.
Multiplizieren wir die relativen Häufigkeiten mit 100, so erhalten wir **Prozentangaben**. Die Summe aller Prozentangaben ergibt den Wert 100.

Beispiel 8
Das Merkmal `Satz` in Tabelle 1.2 auf Seite 17 hat nominales Messniveau. Es nimmt die Ausprägung `j` an, wenn eine Person den Satz

Zu Risiken und Nebenwirkungen

richtig fortsetzen konnte. Ansonsten nimmt es die Ausprägung `n` an. Wir bezeichnen die Merkmalsausprägung `j` mit a_1 und die Merkmalsausprägung `n` mit a_2. Die Beobachtungen stehen in der siebenten Spalte der Tabelle 1.2 auf Seite 17. Sie sind

```
n n j n n j n n j n j n j j n n n n n n j j n j j
```

In dieser Form können wir die Struktur nicht erkennen. Wir bestimmen deshalb die relativen Häufigkeiten. Hierzu erstellen wir eine **Strichliste** Strichliste.

```
j  ||||| |||||
n  ||||| ||||| |||||
```

Es gilt also $n_1 = 10$ und $n_2 = 15$. Die relative Häufigkeit der Merkmalsausprägung `j` ist $f_1 = 10/25 = 0.4$ und die relative Häufigkeit der Merkmalsausprägung `n` gleich $f_2 = 15/25 = 0.6$.

□

Wir haben im letzten Beispiel ein qualitatives Merkmal mit zwei Merkmalsausprägungen betrachtet. Man spricht auch von einem **binären** oder **dichotomen** Merkmal. Bei einem binären Merkmal muss man nur die relative Häufigkeit einer der beiden Merkmalsausprägungen angeben, da sich die zweite dadurch ergibt, dass die Summe beider relativer Häufigkeiten gleich 1 ist. Bei einem binären Merkmal ist es also nicht nötig, eine Tabelle oder Grafik zu erstellen. Hier reicht es völlig aus, die relative Häufigkeit in Form einer Prozentangabe in den Fließtext zu integrieren. Man würde das Ergebnis der Befragung also folgendermaßen zusammenfassen:

> Nur 40 Prozent der Teilnehmer konnten den Satz richtig vollenden.

Oft werden die Häufigkeiten von Merkmalen mit mehr als zwei Merkmalsausprägungen in Satzform angegeben. Tufte (2001) weist darauf hin, wie man vorzugehen hat, damit die Informationen vom Leser registriert und verinnerlicht werden. Das folgende Beispiel veranschaulicht die Vorschläge von Tufte.

Beispiel 9
Wollen wir zum Beispiel das Wahlergebnis der großen Parteien bei der Bundestagswahl 2002 im Fließtext darstellen, so könnte dies folgendermaßen aussehen:
Vorschlag 1

> Bei der Bundestagswahl 2002 erhielt die CDU/CSU, die SPD, die FDP und die GRÜNEN 38.5, 38.5, 7.4 und 8.5 Prozent der Zweitstimmen.

In dieser Form ist der Satz schwer zu lesen. Der Leser muss immer hin- und her springen, um die Zahlen mit den jeweiligen Parteien zu verbinden. Besser ist schon der
Vorschlag 2

> Bei der Bundestagswahl 2002 erhielt die CDU/CSU 38.5, die SPD 38.5, die FDP 7.4 und die GRÜNEN 8.5 Prozent der Zweitstimmen.

Hier ist alles schon viel übersichtlicher. Noch besser ist es aber, die Zahlen in einer Tabelle in den Satz zu integrieren.
Vorschlag 3

Bei der Bundestagswahl 2002 erhielten die

CDU/CSU	38.5
SPD	38.5
FDP	7.4
GRÜNEN	8.5

Prozent der Zweitstimmen.

Man kann die Parteien aber auch nach dem Stimmenanteil sortieren und erhält den
Vorschlag 4

Bei der Bundestagswahl 2002 erhielten die

FDP	7.4
GRÜNE	8.5
SPD	38.5
CDU/CSU	38.5

Prozent der Zweitstimmen.

Dies sollte man aber nur dann machen, wenn die Reihenfolge der Merkmalsausprägungen beliebig ist.

□

In der Regel wird man die Tabelle nicht in den Text integrieren. Man wird eine **Häufigkeitstabelle** erstellen. Der allgemeine Aufbau einer Häufigkeitstabelle ist in Tabelle 3.1 zu finden.

Tabelle 3.1: Allgemeiner Aufbau einer Häufigkeitstabelle

Merkmals-ausprägungen	absolute Häufigkeit	relative Häufigkeit (in Prozent)
a_1	n_1	$100 \cdot f_1$
$\vdots$	$\vdots$	$\vdots$
a_k	n_k	$100 \cdot f_k$

Ehrenberg (1981) empfiehlt, die Zahlen in Tabellen auf zwei effektive Stellen zu runden. So erhält man durch Runden auf zwei effektive Stellen aus der Zahl 22317 die Zahl 22000 und aus der Zahl 0.004567 die Zahl 0.0046. Der Leser kann sich die Zahl viel besser einprägen. Da es schwierig ist, Nachkommastellen zu vergleichen, sollte man für relative Häufigkeiten Prozentangaben verwenden.

Beispiel 10
Im Wintersemester 1996/1997 wurden die Erstsemester gefragt, welche Partei sie wählen würden, wenn am nächsten Sonntag Bundestagswahl wäre. Tabelle 3.2 enthält die Häufigkeitsverteilung des Wahlverhaltens der Studentinnen.

Tabelle 3.2: Wahlverhalten von Studienanfängerinnen

Wahl	absolute Häufigkeit	relative Häufigkeit (in Prozent)
CDU	13	20.0
SPD	10	15.4
FDP	3	4.6
GRÜNE	11	16.9
keine	5	7.7
weiss nicht	23	35.4

Um eine einheitliche Darstellung zu erhalten, wurde auch bei Zahlen mit zwei Dezimalstellen vor dem Dezimalpunkt eine Stelle nach dem Dezimalpunkt angegeben.

□

Da ein Bild mehr als 1000 Worte oder Zahlen sagt, stellt man die in einer Häufigkeitstabelle enthaltene Information grafisch dar. Dabei ordnet man den Häufigkeiten Längen von Strecken oder Flächeninhalte zu, die proportional zu ihnen sind. Hierfür gibt es eine Reihe von Möglichkeiten.

Aus Zeitungen und Zeitschriften kennt man das **Kreisdiagramm**, das auch **Tortendiagramm** oder **Kreissektorendiagramm** genannt wird. Bei diesem werden die absoluten oder relativen Häufigkeiten durch Flächen von Kreissegmenten dargestellt. Dabei ist die einer Merkmalsausprägung zugeordnete Fläche proportional zur relativen Häufigkeit.

Beispiel 8 (fortgesetzt von Seite 55)
Abbildung 3.1 zeigt das Kreisdiagramm des Merkmals `Satz` in Tabelle 1.2 auf Seite 17.

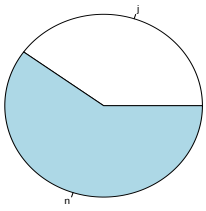


Abbildung 3.1: Kreisdiagramm des Merkmals `Satz`

Wir erkennen sofort, dass der Anteil derjenigen, die den Satz nicht vollenden konnten, höher ist als der Anteil derjenigen, die ihn vollenden konnten. Die genauen Anteil können wir dem Kreisdiagramm nicht entnehmen. Oft werden diese in der Grafik angegeben. Dadurch wird das Bild aber unübersichtlicher. □

Tufte (2001) gibt eine Reihe von Gründen an, warum man Kreisdiagramme nicht verwenden sollte. Das folgende Beispiel illustriert einen dieser Gründe.

Beispiel 10 (fortgesetzt)
Wir schauen uns die Häufigkeiten in Tabelle 3.2 auf Seite 58 an. Abbildung 3.2 zeigt das Kreisdiagramm des Wahlverhaltens der Studentinnen. Aufgrund der vielen Ausprägungsmöglichkeiten des Wahlverhaltens ist es sehr unübersichtlich. Man kann die Häufigkeiten sehr schlecht vergleichen.

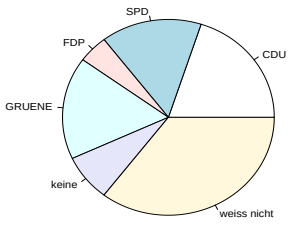


Abbildung 3.2: Kreisdiagramm des Wahlverhaltens der Studentinnen

□

Auch Wainer (1997) weist an Hand von Beispielen auf die Nachteile von Kreisdiagrammen hin und zeigt auf, wie man die in den Kreisdiagrammen enthaltene Information mit einem **Stabdiagramm** oder **Säulendiagramm** beziehungsweise einem **geordneten Säulendiagramm** besser visualisieren kann. Diese beruhen auf der Strichliste, die auf Seite 55 eingeführt wurde. Verzichtet man auf die Unterteilung in Fünferblöcke und wählt den gleichen Abstand zwischen den Strichen, so erhält man folgende Strichliste.

```
j  ||||||||||
n  ||||||||||||||||
```

Stabdiagramm und Säulendiagramm sind Varianten der Strichliste. Bei einem Stabdiagramm und einem Säulendiagramm stehen in einem kartesischen Koordinatensystem auf der Abszisse die Merkmalsausprägungen und auf der Ordinate die relativen oder auch absoluten Häufigkeiten. Wird über jeder Merkmalsausprägung eine senkrechte Linie abgetragen, deren Länge der absoluten oder relativen Häufigkeit der Merkmalsausprägung entspricht, so spricht man von einem Stabdiagramm. Zeichnet man anstatt der Linie eine Säule, so spricht man vom Säulendiagramm. Bei Stab- und Säulendiagrammen kann man die Häufigkeiten sehr gut vergleichen. Außerdem kann man die Werte ablesen. Wir verwenden im Folgenden Säulendiagramme.

Beispiel 10 (fortgesetzt)
Wir schauen uns die Häufigkeiten in Tabelle 3.2 auf Seite 58 an. Abbildung 3.3 zeigt das Säulendiagramm des Wahlverhaltens der Studentinnen.

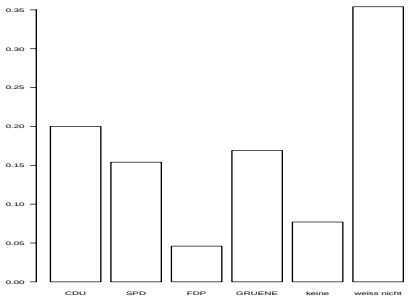


Abbildung 3.3: Säulendiagramm des Wahlverhaltens der Studentinnen

□

Besitzt ein Merkmal sehr viele Ausprägungsmöglichkeiten, so kann das Säulendiagramm sehr unübersichtlich sein, sodass es schwer zu erkennen ist, welche Merkmalsausprägungen die größte relative Häufigkeit besitzen. Es liegt nahe, die Merkmalsausprägungen in der Reihenfolge ihrer Häufigkeit abzutragen. Man spricht von einem *geordneten Säulendiagramm* . Bei einem geordneten Säulendiagramm kann man auf einen Blick erkennen, welche Merkmale am häufigsten auftreten. Ein geordnetes Säulendiagramm sollte man aber nur erstellen, wenn es keine natürliche Ordnung der Merkmalsausprägungen gibt.

Beispiel 10 (fortgesetzt)
Wir schauen uns wieder die Häufigkeiten in Tabelle 3.2 auf Seite 58 an. Abbildung 3.4 zeigt das geordnete Säulendiagramm des Wahlverhaltens der weiblichen Erstsemester.

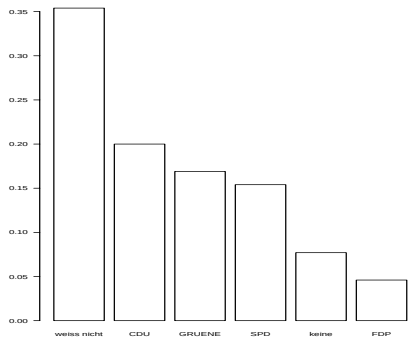


Abbildung 3.4: Geordnetes Säulendiagramm des Wahlverhaltens der weiblichen Erstsemester

□

Schauen wir uns ordinalskalierte Merkmale an. Bei diesen sind die Merkmalsausprägungen $a_1, a_2, \ldots, a_k$ mit $a_1 \leq a_2 \leq \ldots \leq a_k$ geordnet. Deshalb können wir die absoluten und relativen Häufigkeiten kumulieren. Das heißt, dass wir die Summe der ersten beiden absoluten beziehungsweise relativen Häufigkeiten bilden, dann die Summe der ersten drei und so weiter. Wir erhalten die **kumulierten absoluten Häufigkeiten** und **kumulierten relativen Häufigkeiten**. Sind $a_1, a_2, \ldots, a_k$ die geordneten Merkmalsausprägungen, so bestimmen wir die i-te kumulierte absolute Häufigkeit

durch

$$\sum_{j=1}^{i} n_j = n_1 + n_2 + \ldots + n_i \tag{3.2}$$

und die i-te kumulierte relative Häufigkeit durch

$$\sum_{j=1}^{i} f_j = f_1 + f_2 + \ldots + f_i \tag{3.3}$$

Wir ergänzen die Häufigkeitstabelle eines ordinalen Merkmals um eine Spalte, die die kumulierten relativen Häufigkeiten enthält. Der allgemeine Aufbau einer Häufigkeitstabelle eines ordinalskalierten Merkmals ist in Tabelle 3.3 zu finden.

Tabelle 3.3: Allgemeiner Aufbau einer Häufigkeitstabelle eines ordinalskalierten Merkmals

Merkmals-ausprägung	absolute Häufigkeit	relative Häufigkeit (in Prozent)	kumulierte relative Häufigkeit (in Prozent)
a_1	n_1	$100 \cdot f_1$	$100 \cdot f_1$
a_2	n_2	$100 \cdot f_2$	$100 \cdot (f_1 + f_2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_k	n_k	$100 \cdot f_k$	100

Beispiel 11
Die Teilnehmer einer Weiterbildungsveranstaltung wurden gefragt, wie ihnen der Film Titanic gefallen hat. Da `sehr gut` besser als `gut` und `gut` besser als `mittelmäßig` ist, sind die Merkmalsausprägungen geordnet. Die Daten stehen in der fünften Spalte von Tabelle 1.2 auf Seite 17. Es gibt bei diesem Merkmal sehr viele fehlende Beobachtungen, da nur Personen den Film bewerten können, die ihn auch gesehen haben. Wir berücksichtigen die fehlenden Beobachtungen in der Aufzählung nicht:

```
g g g sg sg sg g sg g g g g sg m sg g m
```

Die Häufigkeitstabelle des Merkmals `Bewertung` sieht folgendermaßen aus:

Tabelle 3.4: Häufigkeitstabelle des Merkmals Bewertung

Bewertung	absolute Häufigkeit	relative Häufigkeit	kumulierte relative Häufigkeit
sehr gut	6	0.35	0.35
gut	9	0.53	0.88
mittelmäßig	2	0.12	1.00

Den kumulierten relativen Häufigkeiten können wir unter anderem entnehmen, dass 88 Prozent der Teilnehmer den Film mindestens `gut` finden.

□

Schauen wir uns noch ein Beispiel für die Verwendung von Häufigkeitsverteilungen an.

Beispiel 12
Jede Sprache besitzt eine charakteristische Häufigkeitsverteilung der Buchstaben. Tabelle 3.5 zeigt die Häufigkeitsverteilung der Buchstaben in der deutschen Sprache.

Tabelle 3.5: Häufigkeitsverteilung der Buchstaben in der deutschen Sprache (in Prozent)

a	6.51	e	17.40	i	7.55	m	2.53	q	0.09	u	4.35	y	0.04
b	1.89	f	1.66	j	0.27	n	9.78	r	7.00	v	0.67	z	1.13
c	3.06	g	3.01	k	1.21	o	2.51	s	7.27	w	1.89		
d	5.08	h	4.76	l	3.44	p	0.79	t	6.15	x	0.03		

Quelle: Kippenhahn (1999)
Es ist sinnvoll, die Häufigkeitsverteilung mit einem geordneten Säulendiagramm darzustellen, da man sofort erkennen kann, welche Buchstaben am häufigsten vorkommen. Abbildung 3.5 zeigt das geordnete Säulendiagramm der Buchstaben in der deutschen Sprache. Wir sehen, dass das `e` der häufigste Buchstabe ist. Gefolgt wird es von n, i, s und r. Man kann diese Häufigkeitstabelle benutzen, um Texte zu dekodieren, die nach einem einfachen Verfahren verschlüsselt wurden. Schauen wir uns ein Beispiel an.

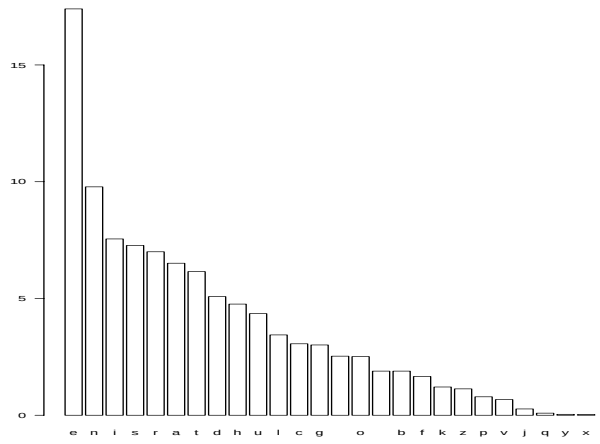


Abbildung 3.5: Geordnetes Säulendiagramm der Buchstaben in der deutschen Sprache

Im Begleitheft der ersten Fünf Freunde CD-ROM finden sich Lösungshinweise, die aber kodiert sind. So findet sich unter *Wie komme ich auf die Insel?* folgender Text:

```
Yq hmi Mrwip fixvixir dy osirrir, fveyglwx hy jspkirhi
Kikirwxeirhi: Imri Ebx yrh hew Wimp eyw hiq Wglyttir yrh
hmiVyhiv eyw hiq Zivwxigo zsr Kisvki. Eywwivhiq qeglx
iwivwx Wmrr, hmi Mrwip dy ivjsvwglir, airr hy hir
Wglexdtper irxhigox lewx.
```

Jedem Buchstaben des Alphabets wurde ein anderer Buchstabe des Alphabets zugeordnet. In der Lösungshilfe steht, dass der Text mit Hilfe einer Cäsar-Verschlüsselung kodiert worden ist. Bei dieser wird jeder Buchstabe durch den Buchstaben verschlüsselt, der eine bestimmte Anzahl von Buchstaben hinter ihm steht. Wird also zum Beispiel das `a` durch das `h` verschlüsselt, so wird das `b` durch das `i` verschlüsselt, das `c` durch das `j` und so weiter. Mit dieser Zusatzinformation ist die Entschlüsselung einfach. Wir wissen, dass das `e` der häufigste Buchstabe in deutschen Texten ist. Wir bestimmen also nur den häufigsten Buchstaben im obigen Text und sind fertig. Der häufigste Buchstabe ist das `i`. Dieses kommt 37-mal vor. Wir ersetzen also jedes `i` durch ein `e`, jedes `j` durch ein `f` und so weiter.
Wir erhalten folgenden Text:

> Um die Insel betreten zu koennen, brauchst du folgende Gegenstaende: eine Axt und das Seil aus dem Schuppen und die Ruder

> aus dem Versteck von George. Ausserdem macht es erst Sinn, die Insel zu erforschen, wenn du den Schatzplan entdeckt hast

Ist eine Botschaft mit dem Cäsar-Verfahren verschlüsselt, so kann man sie leicht entschlüsseln. Kompliziertere Verfahren der Verschlüsselung werden von Beutelspacher, Schwenk & Wolfenstetter (2004) beschrieben. Ein populärwissenschaftliches Buch zur Kryptographie wurde von Singh (2000) geschrieben.

□

3.1.2 Darstellung qualitativer Merkmal in R

Wir betrachten zunächst das Merkmal `Satz` in Tabelle 1.2 auf Seite 17. Wie man die Daten als Zeichenkette eingibt und in einen Faktor transformiert, kann man auf Seite 25 nachlesen. Die Daten stehen im Folgenden in der Variablen `satz`.

```
> satz
 [1] n n j n n j n n j n j n j j n n n n n n j j n j j
Levels: j n
```

Eine Tabelle mit den absoluten Häufigkeiten liefert die Funktion `table`. `table`

```
> table(satz)
satz
 j  n
10 15
```

Die relativen Häufigkeiten erhalten wir durch

```
> h<-table(satz)
> h/sum(h)
satz
  j   n
0.4 0.6
```

und die Prozentangaben durch

```
> 100*h/sum(h)
satz
 j  n
40 60
```

Wir erzeugen die Variable `h` und geben nicht

```
> table(satz)/sum(table(satz))
satz
  j   n
0.4 0.6
```

ein, damit der Befehl `table(satz)` nicht zweimal ausgeführt werden muss.

pie

Mit der Funktion `pie` können wir ein Kreisdiagramm erstellen. Das Kreisdiagramm der Variablen `satz` in Abbildung 3.1 auf Seite 59 gewinnen wir durch

```
> pie(table(satz))
```

Die Erstellung von Säulendiagrammen illustrieren wir an Hand der Daten in Tabelle 3.2 auf Seite 58. Wir geben zunächst die absoluten Häufigkeiten ein.

```
> wahl<-c(13,10,3,11,5,23)
> wahl
[1] 13 10  3 11  5 23
```

Und bestimmen die relativen Häufigkeiten

```
> wahl<-round(wahl/sum(wahl),3)
> wahl
[1] 0.200 0.154 0.046 0.169 0.077 0.354
```

names

Mit der Funktion `names` benennen wir die Komponenten des Vektors `wahl`.

```
> names(wahl)<-c("CDU","SPD","FDP","GRUENE","keine",
                 "weiss nicht")
> wahl
   CDU    SPD    FDP  GRUENE   keine weiss nicht
 0.200  0.154  0.046   0.169   0.077       0.354
```

Ein Säulendiagramm erstellt man mit der Funktion `barplot`. Das Säulendiagramm der Variablen `wahl` in Abbildung 3.3 auf Seite 60 erhält man durch

barplot

```
> par(las=1)
> barplot(wahl,col=0)
```

Der Aufruf `par(las=1)` wird auf Seite 40 beschrieben.
Die Funktion `barplot` besitzt eine Vielzahl fakultativer Argumente. Besitzen die Komponenten des Vektors der relativen Häufigkeiten keine Namen, so kann man diese Namen der Funktion `barplot` als Vektor, der aus Zeichenketten besteht, durch das Argument `names` übergeben. Das Argument `col` gibt die Farbe der Säulen an, wobei 0 weiß entspricht.
Um ein geordnetes Säulendiagramm zu erstellen, muss man den Vektor `wahl` absteigend sortieren. Die Funktion `sort` haben wir auf Seite 25 kennen gelernt.

```
> sort(wahl)
   FDP  keine    SPD  GRUENE    CDU  weiss nicht
 0.046  0.077  0.154   0.169  0.200        0.354
```

Setzt man das Argument `decreasing` auf den Wert `TRUE`, so wird absteigend sortiert.

```
> sort(wahl,decreasing=TRUE)
 weiss nicht    CDU  GRUENE    SPD  keine    FDP
       0.354  0.200   0.169  0.154  0.077  0.046
```

Das geordnete Säulendiagramm des Mermals `wahl` in Abbildung 3.4 auf Seite 61 erhält man also durch

```
> par(las=1)
> barplot(sort(wahl,decreasing=TRUE))
```

3.1.3 Darstellung quantitativer Merkmale

Die Merkmalsausprägungen quantitativer Merkmale sind Zahlen, mit denen man rechnen darf. Im Verhältnis zu qualitativen Merkmalen gibt es deshalb viel mehr Möglichkeiten, die Verteilung quantitativer Merkmale darzustellen und zu beschreiben.
Ausgangspunkt der Analyse quantitativer Merkmale ist die **Urliste**

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

Dabei ist x_1 die Merkmalsausprägung des ersten Merkmalsträgers, x_2 die Merkmalsausprägung des zweiten Merkmalsträgers und so weiter. Allgemein bezeichnen wir die Merkmalsausprägung des i-ten Merkmalsträgers mit x_i. Auch bei quantitativen Merkmalen bestimmen wir absolute Häufigkeiten. Bei diskreten Merkmalen gehen wir genauso vor wie bei ordinalskalierten Merkmalen, während die Natur stetiger Merkmale eine andere Vorgehensweise fordert.

Diskrete Merkmale

Die Anzahl Ausprägungsmöglichkeiten diskreter Merkmale ist endlich oder abzählbar unendlich. In der Praxis fassen wir Merkmale als diskret auf, die wenig Merkmalsausprägungen besitzen. So ist im Beispiel 10 auf Seite 10 das Merkmal `Anzahl Geschwister` ein quantitatives Merkmal mit wenigen Merkmalsausprägungen, also ein diskretes Merkmal.
Da die Urliste unübersichtlich ist, bilden wir den **geordneten Datensatz**

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$$

mit $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$.
Die kleinste Beobachtung bezeichnen wir also mit $x_{(1)}$, die zweitkleinste mit $x_{(2)}$ und die größte mit $x_{(n)}$.

Beispiel 13
Wir betrachten das Merkmal `Anzahl Geschwister`. Die Daten stehen in der sechsten Spalte von Tabelle 1.3 auf Seite 18.
Die Urliste sieht folgendermaßen aus:

```
1 2 1 3 0 2 2 1 1 1 1 1 1 2 1 1 0 1 2 1
```

Sie ist sehr unübersichtlich. Deshalb bilden wir den geordneten Datensatz. Dieser lautet:

```
0 0 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 3
```

An diesem können wir sofort den kleinsten Wert 0 und den größten Wert 3 erkennen.

□

Wie bei einem qualitativen Merkmal bestimmen wir für $i = 1, 2, \ldots, k$ die absoluten Häufigkeiten n_i und die relativen Häufigkeiten f_i der einzelnen Merkmalsausprägungen a_i. Diese stellen wir in der Häufigkeitstabelle zusammen, bei der wir noch die kumulierten relativen Häufigkeiten berücksichtigen. Die Häufigkeitstabelle ist aufgebaut wie Tabelle 3.3 auf Seite 62.

Beispiel 13 (fortgesetzt)
Die Häufigkeitstabelle des Merkmals `Anzahl Geschwister` ist in Tabelle 3.6 auf der nächsten Seite zu finden.
Hier können wir schon sehen, dass mehr als die Hälfte der Studierenden ein Geschwister hat. Noch deutlicher ist dies am Säulendiagramm zu sehen. Dieses ist in Abbildung 3.6 auf der nächsten Seite zu finden.

Tabelle 3.6: Häufigkeitstabelle des Merkmals `Anzahl Geschwister`

Anzahl Geschwister	absolute Häufigkeit	relative Häufigkeit	kumulierte relative Häufigkeit
0	2	0.10	0.10
1	12	0.60	0.70
2	5	0.25	0.95
3	1	0.05	1.00

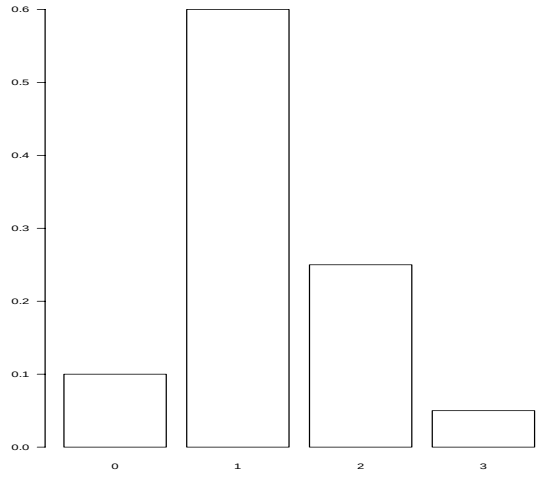


Abbildung 3.6: Säulendiagramm des Merkmals `Anzahl Geschwister`

□

Es gibt bestimmte Muster in Säulendiagrammen, die immer wieder beobachtet werden. Abbildung 3.7 auf der nächsten Seite zeigt einige Beispiele.

Beim Merkmal in der Abbildung links oben werden die relativen Häufigkeiten mit wachsendem x immer kleiner. Kleine Merkmalsausprägungen treten also viel häufiger auf als große. Man spricht von einer **rechtsschiefen** oder auch **linkssteilen** Verteilung. Ein schönes Beispiel hierfür ist die Haushaltsgröße in Deutschland im Jahr 2000. Sehr viele Haushalte bestehen aus wenig Personen, während sehr wenige Haushalte viele Personen enthalten.

Beim Merkmal in der Abbildung rechts oben ist es genau umgekehrt. Große Merkmalsausprägungen treten also viel häufiger auf als kleine. Man spricht von einer **linksschiefen** oder auch **rechtssteilen** Verteilung. Ein schönes

Beispiel hierfür ist die Haushaltsgröße in Deutschland im Jahr 1900. Sehr viele Haushalte bestehen aus vielen Personen, während sehr wenige Haushalte wenige Personen enthalten.
Beim Merkmal in der Abbildung links unten liegen die Säulen nahezu symmetrisch um das Zentrum der Verteilung. Die Merkmalsausprägungen in der Mitte treten häufiger als an den Rändern auf. Man spricht auch von einer **symmetrischen** Verteilung.
In der Abbildung rechts unten gibt es zwei Maxima. Man spricht auch von einer **zweigipfligen** oder auch **bimodalen** Verteilung.

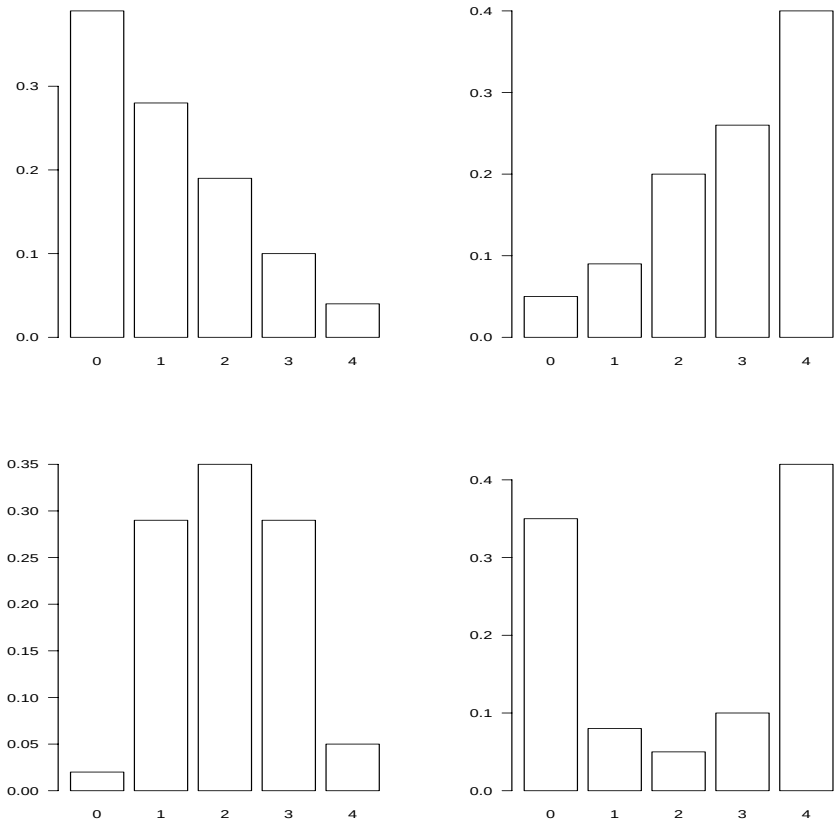


Abbildung 3.7: Unterschiedliche Säulendiagramme

Bisher haben wir die kumulierten relativen Häufigkeiten nur für die Merkmalsausprägungen $a_1, a_2, \ldots, a_k$ bestimmt. Ist das Merkmal quantitativ, so können wir die kumulierten relativen Häufigkeiten für alle reellen Zahlen be-

stimmen. Für jedes $x \in \mathbb{R}$ bezeichnen wir diese mit $f(X \leq x)$. Dies ist der Anteil der Beobachtungen, die kleiner oder gleich x sind. Jeder reellen Zahl x wird also ein Wert $f(X \leq x)$ zugeordnet. Wir nennen die zugehörige Funktion die **empirische Verteilungsfunktion** $F_n(x)$. Sie ist definiert durch

$$F_n : \mathbb{R} \to [0, 1]$$

mit

$$x \mapsto F_n(x) = f(X \leq x)$$

Mit der empirischen Verteilungsfunktion kann man relative Häufigkeiten für Intervalle bestimmen. So erhält man den Anteil $f(X > x)$ der Merkmalsträger, deren Merkmalsausprägung größer als der Wert x ist, durch:

$$f(X > x) = 1 - F_n(x)$$

Den Anteil $f(a < X \leq b)$ der Merkmalsträger, deren Merkmalsausprägung im Intervall $(a, b]$ liegt, bestimmt man durch

$$f(a < X \leq b) = F_n(b) - F_n(a)$$

Beispiel 13 (fortgesetztvon Seite 68)
Wir schauen uns die Häufigkeitstabelle des Merkmals `Anzahl Geschwister` in Tabelle 3.6 auf Seite 69 an. Die empirische Verteilungsfunktion des Merkmals `Anzahl Geschwister` sieht folgendermaßen aus:

$$F_n(x) = \begin{cases} 0 & \text{für } x < 0 \\ 0.1 & \text{für } 0 \leq x < 1 \\ 0.7 & \text{für } 1 \leq x < 2 \\ 0.95 & \text{für } 2 \leq x < 3 \\ 1 & \text{für } x \geq 3 \end{cases}$$

Der Anteil der Studierenden mit mehr als zwei Geschwistern ist also

$$f(X > 2) = 1 - F_n(2) = 1 - 0.95 = 0.05$$

Abbildung 3.8 zeigt die empirische Verteilungsfunktion des Merkmals `Anzahl Geschwister`.

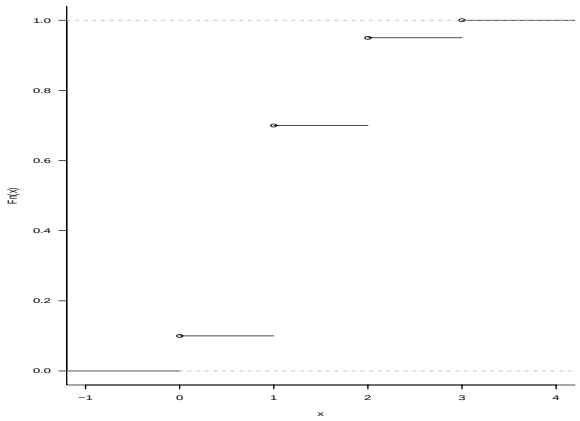


Abbildung 3.8: Empirische Verteilungsfunktion des Merkmals `Anzahl Geschwister`

□

Wir sehen, dass die empirische Verteilungsfunktion eine Treppenfunktion ist. Die Sprungstellen liegen an den Merkmalsausprägungen, die beobachtet wurden. Die Höhe der Sprünge an den Sprungstellen ist gleich den relativen Häufigkeiten der jeweiligen Merkmalsausprägungen.

Stetige Merkmale

Stetige Merkmale können theoretisch alle Werte aus einem Intervall annehmen. In der Praxis behandelt man quantitative Merkmale als stetig, die sehr viele Merkmalsausprägungen besitzen.
Wie bei bei einem diskreten Merkmal bildet die Urliste $x_1, \ldots, x_n$ bei einem stetigen Merkmal den Ausgangspunkt der Analyse. Wir bilden auch hier den geordneten Datensatz $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$.

Beispiel 14
Das Merkmal `Alter` aus Beispiel 1 auf Seite 10 ist stetig. Die Werte dieses Merkmals sind in der dritten Spalte von Tabelle 1.2 auf Seite 17 zu finden. Die Urliste sieht folgendermaßen aus.

```
30 23 26 33 37 28 31 23 24 26 23 32 29
25 31 26 37 38 29 28 28 28 38 27 27
```

Der geordnete Datensatz ist:

```
23 23 23 24 25 26 26 26 27 27 28 28 28
28 29 29 30 31 31 32 33 37 37 38 38
```

□

Ein stetiges Merkmal besitzt sehr viele Merkmalsausprägungen. Eine Häufigkeitstabelle wie im Fall eines diskreten Merkmals wäre sehr unübersichtlich. Aus diesem Grunde bildet man sogenannte **Klassen**. Man fasst also mehrere Werte zusammen. Wir bezeichnen die Untergrenze der i-ten Klasse mit x^*_{i-1} und die Obergrenze mit x^*_i. Bis auf die erste Klasse gehört die Obergrenze zur Klasse, die Untergrenze hingegen nicht. Die erste Klasse ist also $[x^*_0, x^*_1]$, während die i-te Klasse für $i > 1$ von der Form $(x^*_{i-1}, x^*_i]$ ist. Man spricht von einer **links offenen** und **rechts abgeschlossenen** Klasse. Manchmal werden links abgeschlossene und rechts offene Klassen verwendet. Wir werden in der Regel aber links offene und rechts abgeschlossene Klassen betrachten. Wir bestimmen für $i = 1, 2, \ldots, k$ die absolute Häufigkeit n_i und die relative Häufigkeit f_i der i-ten Klasse. Die absoluten und relativen Häufigkeiten stellen wir in der Häufigkeitstabelle zusammen. Der allgemeine Aufbau der Häufigkeitstabelle ist in Tabelle 3.7 zu finden.

Tabelle 3.7: Allgemeiner Aufbau einer Häufigkeitstabelle mit klassierten Beobachtungen

Klasse	Intervall	absolute Häufigkeit	relative Häufigkeit
1	$[x^*_0, x^*_1]$	n_1	f_1
2	$(x^*_1, x^*_2]$	n_2	f_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	$(x^*_{k-1}, x^*_k]$	n_k	f_k

Beispiel 14 (fortgesetzt)
Wir betrachten das Merkmal Alter aus Tabelle 1.2 auf Seite 17 und bilden die vier Klassen $[20, 25]$, $(25, 30]$, $(30, 35]$ und $(35, 40]$. Es gilt $x^*_0 = 20$, $x^*_1 = 25$, $x^*_2 = 30$, $x^*_3 = 35$ und $x^*_4 = 40$.
Von den 25 Teilnehmern sind 5 höchstens 25 Jahre alt, 12 älter 25 aber höchstens 30 Jahre alt, 4 älter als 30 aber höchstens 35 Jahre alt und 4 älter als 35 Jahre.
Wir erstellen die Häufigkeitstabelle mit den absoluten und relativen Häufigkeiten. Diese ist in Tabelle 3.8 auf der nächsten Seite zu finden.

Tabelle 3.8: Die Häufigkeitstabelle des Merkmals Alter

Alter	absolute Häufigkeit	relative Häufigkeit
von 20 bis 25	5	0.20
von 25 bis 30	12	0.48
von 30 bis 35	4	0.16
von 35 bis 40	4	0.16

□

Graphisch stellen wir die relativen Häufigkeiten mit einem **Histogramm** dar. Dabei tragen wir in einem rechtwinkligen Koordiantensystem über jeder Klasse ein Rechteck ab, dessen Fläche gleich der relativen Häufigkeit der Klasse ist. Um dies zu erreichen, wählen wir als Höhe des Rechtecks den Quotienten aus relativer Häufigkeit f_i und Klassenbreite Δ_i.
Die zugehörige Funktion heißt **empirische Dichtefunktion** $\hat{f} : I\!R \to I\!R$ mit

$$\hat{f}(x) = \begin{cases} \dfrac{f_i}{\Delta_i} & \text{für } x^*_{i-1} < x \leq x^*_i, i = 1, \ldots, k \\ 0 & \text{sonst} \end{cases} \tag{3.4}$$

Beispiel 14 (fortgesetzt)
Wir betrachten die Häufigkeitstabelle des Merkmals Alter in Tabelle 3.8 auf Seite 74.
Die Breite Δ_1 des ersten Intervalls ist 5. Außerdem gilt $f_1 = 0.2$. Also nimmt die empirische Dichtefunktion in diesem Intervall den Wert

$$\frac{f_1}{\Delta_1} = \frac{0.2}{5} = 0.04$$

an. Entsprechend erhalten wir die anderen Werte. Die empirische Dichtefunktion lautet

$$\hat{f}(x) = \begin{cases} 0.040 & \text{für } 20 \leq x \leq 25 \\ 0.096 & \text{für } 25 < x \leq 30 \\ 0.032 & \text{für } 30 < x \leq 35 \\ 0.032 & \text{für } 35 < x \leq 40 \\ 0 & \text{sonst} \end{cases}$$

Abbildung 3.9 zeigt das Histogramm des Merkmals Alter.

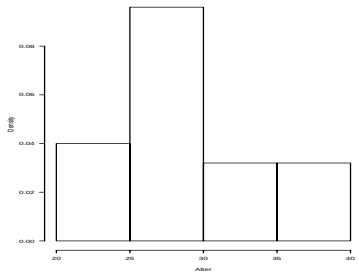


Abbildung 3.9: Histogramm des Merkmals Alter

Wir sehen, dass fast die Hälfte der Teilnehmer zwischen 25 und 30 Jahre alt ist. Die restlichen Klassen sind ungefähr gleich dicht besetzt.

□

Beim Erstellen des Histogramms muss man *die Anzahl der Klassen*, *die Breite der Klassen* und *die Untergrenze der ersten Klasse* wählen. Die Gestalt des Histogramms hängt davon ab, wie man diese Größen wählt. Abbildung 3.10 zeigt zwei Histogramme des Merkmals Alter. Die linke Abbildung zeigt, dass durch zu wenige Klassen Informationen über Details verloren gehen, während bei einer zu hohen Anzahl von Klassen wie in rechten Abbildung die vielen Details die allgemeine Struktur verdecken.

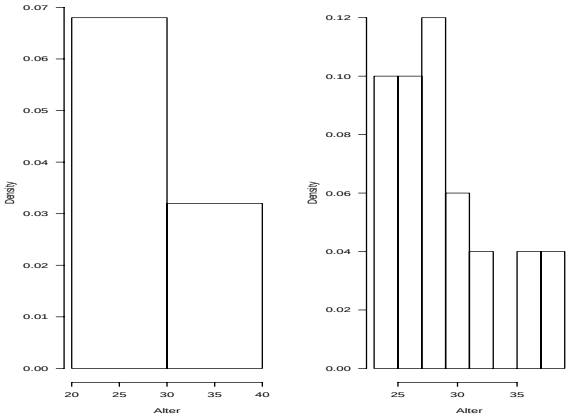


Abbildung 3.10: Zwei Histogramme des Merkmals Alter

Wie soll man die drei Größen wählen? Es gibt eine Reihe von Vorschlägen, von denen sich einige durchgesetzt haben und die in Programmpaketen verwendet werden. Doane (1976) schlägt vor, dass alle Klassen gleichgroß sein sollten. Er wählt also äquidistante Klassen. Die Untergrenze der ersten Klasse sollte eine runde Zahl sein. Im Beispiel war das Minimum 23. Also sollte man als Untergrenze der ersten Klasse den Wert 20 wählen. Für die Anzahl der Klassen gibt es eine Reihe von Vorschlägen, die bei Scott (1992) und Heiler & Michels (1994) zu finden sind, die wir hier aber nicht diskutieren wollen.
Abbildung 3.11 zeigt unterschiedliche Histogramme. Wir können diese so wie die Säulendiagramme in Abbildung 3.7 auf Seite 70 interpretieren.

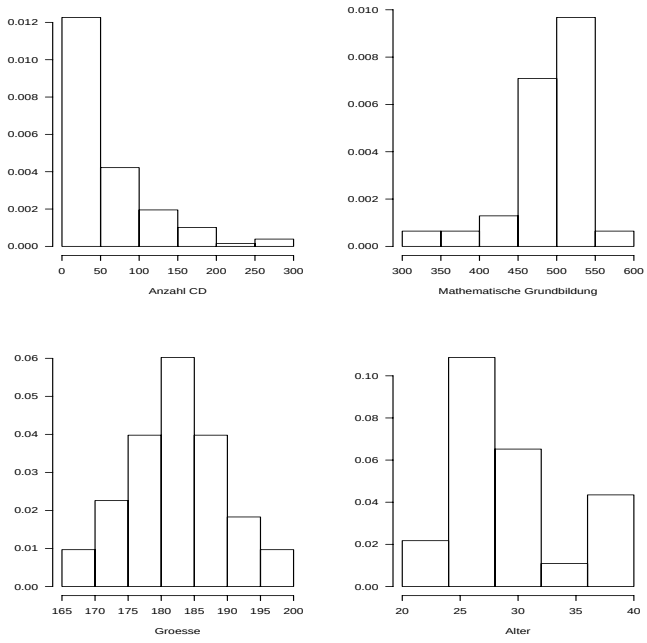


Abbildung 3.11: Unterschiedliche Histogramme

Die Abbildung links oben zeigt das Histogramm der Anzahl CDs von Studenten. Sehr viele Personen haben wenige CDs, während wenige viele CDs besitzen. Es handelt sich um eine **rechtsschiefe linkssteile** Verteilung. Bei der Verteilung in der Abbildung rechts oben liegt der entgegen gesetzte Fall vor. Es handelt sich hier um eine **linksschiefe** (**rechtssteile**) Verteilung. Die Daten stammen aus der PISA-Studie 2000. Wenige Länder hatten hier im Bereich Mathematische Grundbildung einen niedrigen Wert, während viele Länder in diesem Bereich einen hohen Wert erzielten. Die Verteilung in

der Abbildung links unten ist **symmetrisch**. Hier wurde die Körpergröße von männlichen Studienanfängern bestimmt. Die Merkmalsausprägungen im Zentrum der Verteilung treten am häufigsten auf. Je weiter man sich vom Zentrum entfernt, umso seltener werden die Merkmalsausprägungen. In der Abbildung rechts unten ist eine zweigipflige Verteilung zu sehen. Dies deutet darauf hin, dass zwei Gruppen vorliegen. Im Beispiel handelt es sich um das Alter der Teilnehmer einer Weiterbildungsveranstaltung. Hier konnte man zwei Gruppen von Teilnehmern unterscheiden.

Oft besitzt bei klassierten Daten die letzte Klasse keine Obergrenze. Lehn, Müller-Gronbach & Rettig (2000) sprechen von einer offenen Randklasse und weisen darauf hin, dass in diesem Fall kein Histogramm gezeichnet werden kann.

Beispiel 15
Im statistischen Jahrbuch 2004 ist auf Seite 46 eine Häufigkeitstabelle mit dem monatlichen Nettohaushaltseinkommen der Privathaushalte in Deutschland im Jahr 2003 zu finden. Die Daten sind in Tabelle 3.9 wiedergegeben.

Tabelle 3.9: Monatliches Nettohaushaltseinkommen der Privathaushalte in Deutschland im Jahr 2003

Klasse i	$\left(x_{i-1}^*, x_i^*\right]$	n_i	f_i
1	$(0, 500]$	1289	0.035
2	$(500, 900]$	4360	0.117
3	$(900, 1300]$	6315	0.170
4	$(1300, 1500]$	3291	0.089
5	$(1500, 2000]$	6521	0.175
6	$(2000, 2600]$	6038	0.162
7	$(2600, 4500]$	7311	0.197
8	$(4500, -\infty)$	2041	0.055

Wir sehen, dass die letzte Klasse nach oben offen ist. Somit können wir kein Histogramm zeichnen. Man könnte natürlich einen plausiblen Wert für das maximale Haushaltseinkommen vorgeben. Aber welchen soll man wählen? Wählt man zum Beispiel 10000, so weiß man, dass Haushalte diesen Wert überschreiten, und das Histogramm den wahren Tatbestand nicht wiedergibt. Wählt man hingegen 100000, so ist keine Struktur zu erkennen, da das Intervall $(0, 4500]$ im Verhältnis zum Intervall $(0, 100000]$ sehr klein ist.

□

Bei einem diskreten Merkmal haben wir die empirische Verteilungsfunktion $F_n(x)$ betrachtet. Dabei gibt $F_n(x)$ den Anteil der Merkmalsträger an, deren Merkmalsausprägung höchstens x beträgt. Wir wollen diese nun für ein stetiges Merkmal bestimmen. Wir nennen sie in Anlehnung an Burkschat, Cramer & Kamps (2004) **approximierende empirische Verteilungsfunktion** $F_n^*(x)$. Hartung, Elpelt & Klösener (2002) sprechen von der **stetigen empirischen Verteilungsfunktion**.
Den Wert der approximierenden empirischen Verteilungsfunktion an den Obergrenzen der Klassen erhält man durch Summation der relativen Häufigkeiten f_j der Klassen. Für $i = 1, \ldots, k$ gilt also

$$F_n^*(x_i^*) = \sum_{j=1}^{i} f_j$$

Außerdem gilt $F_n^*(x_0^*) = 0$.
Ist x ein Wert innerhalb einer Klasse, so nehmen wir die empirische Dichtefunktion aus Gleichung (3.4) auf Seite 74 als Ausgangspunkt zur Bestimmung von $F_n^*(x)$. Die approximierende empirische Verteilungsfunktion $F_n^*(x)$ an der Stelle x ist gleich der Fläche unter der empirischen Dichtefunktion bis zur Stelle x. Nehmen wir an, der Wert x liegt in der i-ten Klasse mit den Klassengrenzen x_{i-1}^* und x_i^*. Dann erhalten wir den gesuchten Wert, indem wir die Fläche unter dem Histogramm bis zu dieser Stelle bestimmen. Der Wert von $F_n^*(x)$ an der Untergrenze ist $F_n^*(x_{i-1}^*)$. Dazu kommt noch die Fläche innerhalb der Klasse. Diese ist in der Abbildung 3.12 schraffiert dargestellt.

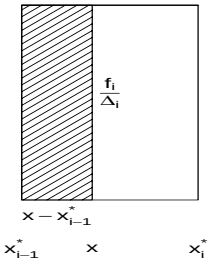


Abbildung 3.12: Bestimmung der approximierenden empirischen Verteilungsfunktion aus dem Histogramm

Die Breite des schraffierten Rechtecks ist $x - x_i^*$ und die Höhe ist $\frac{f_i}{\Delta_i}$. Die Fläche beträgt also

$$(x - x_{i-1}^*) \cdot \frac{f_i}{\Delta_i} \quad = \quad \frac{x - x_{i-1}^*}{\Delta_i} \cdot f_i \,.$$

Also gilt

$$F_n^*(x) = \begin{cases} 0 & \text{für } x \leq x_0^* \\ F_n^*(x_{i-1}^*) + \dfrac{x - x_{i-1}^*}{\Delta_i} \cdot f_i & \text{für } x_{i-1}^* < x \leq x_i^*,\ i = 1, \ldots, k \\ 1 & \text{für } x \geq x_k^* \end{cases} \quad (3.5)$$

Innerhalb der jeder Klasse ist die approximierende empirische Verteilungsfunktion $F_n^*(x)$ eine in x lineare Funktion $a + b \cdot x$.
Es gilt nämlich

$$F_n^*(x) = F_n^*(x_{i-1}^*) + \frac{x - x_{i-1}^*}{\Delta_i} \cdot f_i = \underbrace{F_n^*(x_{i-1}^*) - \frac{x_{i-1}^*}{\Delta_i} \cdot f_i}_{a} + \underbrace{\frac{f_i}{\Delta_i}}_{b} x$$

Sie kann also ganz einfach gezeichnet werden. Wir tragen in einem kartesischen Koordiantensystem an jeder Klassengrenze die kumulierte relative Häufigkeit ein und verbinden je zwei aufeinanderfolgende Punkte durch eine Gerade.

Beispiel 14 (fortgesetzt von Seite 74)
Wir betrachten weiterhin die Häufigkeitsverteilung des Merkmals Alter und schauen wir uns noch einmal Tabelle 3.8 auf Seite 74 an:

Tabelle 3.10: Die Häufigkeitstabelle des Merkmals Alter

i	$\left(x_{i-1}^*, x_i^*\right]$	f_i	Δ_i	$F_n^*(x_{i-1}^*)$	$F_n^*(x_i^*)$
1	$[20, 25]$	0.20	5	0	0.20
2	$(25, 30)$	0.48	5	0.20	0.68
3	$(30, 35)$	0.16	5	0.68	0.84
2	$(35, 40)$	0.16	5	0.84	1.00

In der ersten Klasse gilt

$$a = -\frac{x_0^*}{\Delta_1} \cdot f_1 = -\frac{0.2 \cdot 20}{5} = -0.8$$

und

$$b = \frac{f_1}{\Delta_1} = \frac{0.2}{5} = 0.04$$

Die approximierende empirische Verteilungsfunktion ist

$$F_n^*(x) = \begin{cases} 0 & \text{für } x < 20 \\ -0.8 + 0.04 \cdot x & \text{für } 20 \leq x \leq 25 \\ -2.2 + 0.096 \cdot x & \text{für } 25 < x \leq 30 \\ -0.28 + 0.032 \cdot x & \text{für } 30 < x \leq 35 \\ -0.28 + 0.032 \cdot x & \text{für } 35 < x \leq 40 \\ 1 & \text{für } x > 40 \end{cases} \tag{3.6}$$

Abbildung 3.13 zeigt die approximierende empirische Verteilungsfunktion.

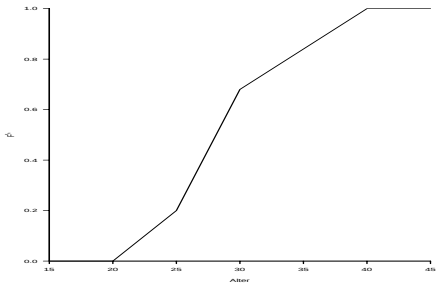


Abbildung 3.13: empirische Verteilungsfunktion des Merkmals Alter

Wir bestimmen den Anteil der Teilnehmer, die höchstens 28 Jahre alt sind. Gesucht ist also $F_n^*(28)$. Wir haben zwei Möglichkeiten. Wenn wir die approximierende empirische Verteilungsfunktion bestimmt haben, dann müssen wir nur das richtige Intervall finden und den vorgegebenen Wert x einsetzen. Der Wert 28 liegt in der zweiten Klasse. Somit gilt

$$F_n^*(28) = -2.2 + 0.096 \cdot 28 = 0.488$$

Wir können aber auch Tabelle 3.10 benutzen. Der Wert 28 liegt in der zweiten Klasse. Mit $F_n^*(25) = 0.20$ und $f_2 = 0.48$ gilt also

$$F_n^*(28) = 0.20 + \frac{28 - 25}{5} \cdot 0.48 = 0.488 \tag{3.7}$$

48.8 Prozent der Teilnehmer sind also höchstens 28 Jahre alt.
Abbildung 3.14 illustriert graphisch die Vorgehensweise:

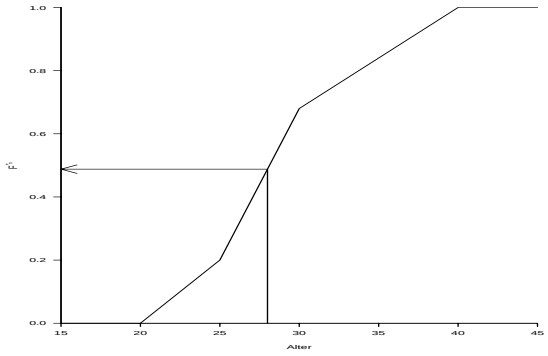


Abbildung 3.14: Bestimmung des Werts der approximierenden empirischen Verteilungfunktion

□

3.1.4 Darstellung quantitativer Merkmal in R

Beginnen wir mit dem Merkmal `Anzahl Geschwister`. Die Daten stehen in der sechsten Spalte von Tabelle 1.3 auf Seite 18. Wir geben die Daten ein, wie wir es auf Seite 22 gelernt haben. Sie mögen in der Variablen `Geschwister` stehen.

```
> Geschwister<-c(1,2,1,3,0,2,2,1,1,1,1,1,1,2,1,1,0,1,2,1)
> Geschwister
 [1] 1 2 1 3 0 2 2 1 1 1 1 1 1 2 1 1 0 1 2 1
```

Den geordneten Datensatz erhalten wir mit der Funktion `sort`: `sort`

```
> sort(Geschwister)
 [1] 0 0 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 3
```

Um die Häufigkeitstabelle und das Säulendiagramm zu ersetellen, gehen wir so vor, wie es bei qualitativen Merkmalen auf den Seiten 65 und 66 gelernt haben.

```
> h<-table(Geschwister)
> h
Geschwister
 0  1  2  3
 2 12  5  1
```

```
> h<-h/sum(h)
> h
Geschwister
   0    1    2    3
0.10 0.60 0.25 0.05
```

cumsum

Die kumulierten relativen Häufigkeiten liefert die Funktion `cumsum`.

```
> cumsum(h)
   0    1    2    3
0.10 0.70 0.95 1.00
```

cbind

Mit der Funktion `cbind` verbinden wir den Vektor der relativen Häufigkeiten mit dem Vektor der kumulierten relativen Häufigkeiten zu einer Matrix.

```
> cbind(h,cumsum(h))
     h
0 0.10 0.10 1 0.60 0.70 2 0.25 0.95 3 0.05 1.00
```

Mit der Funktion `barplot` erstellen wir das Säulendiagramm.

```
> h<-table(Geschwister)
> h<-h/sum(h)
> par(las=1)
> barplot(h)
```

ecdf

ecdf

Die grafische Darstellung der empirischen Verteilungsfunktion aus Abbildung 3.8 auf Seite 72 erhält man durch

```
> plot(ecdf(Geschwister))
```

Schauen wir uns stetige Merkmale an. Wir betrachten das Merkmal `Alter`aus Tabelle 1.2 auf Seite 17. Die Daten der Tabelle mögen in der Datentabelle `weiterbildung` stehen, die auf Seite 30 zu finden ist. Wir greifen auf die Variablen zu, wie auf Seite 30 beschrieben wird.

```
> attach(weiterbildung)
> Alter
 [1] 30 23 26 33 37 28 31 23 24 26 23 32 29  25 31 26 37 38 29
 28 28 28 38 27 27
```

cut

Wir wollen zunächst Klassen bilden und jede Beobachtung klassieren. Hierzu dient die Funktion `cut`. Mit dem Argument `breaks` geben wir die Klassen-

grenzen an. Da wir die Klassen $[20, 25]$, $(25, 30]$, $(30, 35]$ und $(35, 40]$ betrachten, sind die Klassengrenzen 20, 25, 30, 35 und 40. Einen Vektor mit diesen Zahlen erhalten wir am einfachsten mit der Funktion `seq`, die auf Seite 42 beschrieben wird. Wir geben also

```
> seq(20,40,5)
[1] 20 25 30 35 40
```

ein. Neben den Daten `x` müssen wir der Funktion `cut` noch mitteilen, ob die Klassen rechts abgeschlossen oder recht offen sind. Setzen wir das Argument `right` auf den Wert `TRUE`, so sind sie rechts abgeschlossen. Dieser Wert wird auch standardmäßig gewählt. Mit dem Argument `include.lowest` können wir festlegen, ob die erste Klasse links abgeschlossen ist, wenn `right` den Wert `TRUE` annimmt, oder ob die letzte Klasse rechts abgeschlossen ist, wenn `right` den Wert `FALSE` annimmt. Die Funktion `cut` gibt für jede Beobachtung an, zu welcher Klasse sie gehört.
Wir geben also ein

```
> attach(weiterbildung)
> cut(Alter,seq(20,40,by=5),include.lowest=TRUE)
 [1]  (25,30] [20,25] (25,30] (30,35] (35,40] (25,30] (30,35]
 [8]  [20,25] [20,25] (25,30] [20,25] (30,35] (25,30] [20,25]
 [16] (30,35] (25,30] (35,40] (35,40] (25,30] (25,30] (25,30]
 [25] (25,30] (35,40] (25,30] (25,30]
 Levels: [20,25] (25,30] (30,35] (35,40]
```

Wir sehen, dass das Ergebnis der Funktion `cut` vom Typ `factor` ist. Die Häufigkeitstabelle erstellen wir mit Funktion `table`.

```
> h<-table(cut(Alter,seq(20,40,by=5),include.lowest=TRUE))
> h

 [20,25] (25,30] (30,35] (35,40]
       5      12       4       4
```

und bestimmen die relativen Häufigkeiten durch

```
> h<-h/sum(h)
> h

 [20,25] (25,30] (30,35] (35,40]
    0.20    0.48    0.16    0.16
```

hist

Um das Histogramm zu erstellen, verwenden wir die Funktion `hist`. Neben dem Datensatz `x` übergeben wir ihr wie der Funktion `cut` die Klassengrenzen und setzen das Argument `include.lowest` auf den Wert `TRUE`. Außerdem müssen wir das Argument `freq` auf den Wert `FALSE` setzen, um sicherzustellen, dass die Fläche unter dem Histogramm gleich 1 ist. Setzen wir `freq` auf den Wert `TRUE`, so werden auf der ordinate die absoluten Häufigkeiten der Klassen abgetragen. Wir geben also ein

```
> hist(Alter,seq(20,40,5),include.lowest=TRUE,freq=FALSE)
```

Über dem Histogramm steht die Überschrift `Histogram of Alter`. Wir können diese über das Argument `main` ändern, indem wir diesem die gewünschte Überschrift als Zeichenkette zuweisen. Keine Überschrift erhalten wir durch Zuweisung von `""`.

rep

Oft liegen die Daten in Form einer Häufigkeitstabelle mit klassierten Daten vor. Für diesen Fall bietet die Funktion `hist` in `R` keine Argumente an, da sie die Urliste als Argument erwartet. Wir können aber mit der Funktion `rep` auf Basis der Klassengrenzen und der absoluten Häufigkeiten das Histogramm mit der Funktion `hist` erstellen. Die Funktion `rep` wird aufgerufen durch

```
rep(x,times)
```

Ist `x` eine Zahl und `times` eine natürliche Zahl, so ist das Ergebnis von `rep` ein Vektor, der die Zahl `x` `times`-mal enthält.

```
> rep(20,5)
[1] 20 20 20 20 20
```

Ist `x` ein Vektor und `times` eine natürliche Zahl, so ist das Ergebnis von `rep` ein Vektor, der den Vektor `x` `times`-mal enthält.

```
> rep(1:3,4)
 [1] 1 2 3 1 2 3 1 2 3 1 2 3
```

Ist `x` ein Vektor und `times` ein Vektor natürlicher Zahlen, der dieselbe Länge wie tt x besitzt, so ist das Ergebnis von `rep` ein Vektor, der `x[1]` `times[1]`-mal, `x[2]` `times[2]`-mal, u.s.w. enthält.

```
> rep(1:3,c(2,4,3))
[1] 1 1 2 2 2 2 3 3 3
```

Sind also nur die Angaben aus Tabelle 3.8 auf Seite 74 gegeben, so erzeugen wir einen Vektor mit Pseudo-Rohdaten mit der Funktion `rep`, indem wir die Obergrenze der i-ten Klasse n_i-mal wiederholen. Für das Beispiel gilt

```
> rep(c(25,30,35,40),c(5,12,4,4))
 [1] 25 25 25 25 25 30 30 30 30 30 30 30 30 30 30 30 30
 35 35 35 35 40 40 40 40
```

Mit diesem Vektor rufen wir die Funktion `hist` auf und erhalten das gewünschte Histogramm.

```
> hist(rep(seq(25,40,5),c(5,12,4,4)),seq(20,40,5),freq=FALSE)
```

Nun müssen wir nur noch die Abszisse mit dem Argument `xlab` geeignet beschriften und die Überschrift mit dem Argument `main` eliminieren.

```
> hist(rep(seq(25,40,5),c(5,12,4,4)),seq(20,40,5),xlab="Alter",
  main="",freq=FALSE)
```

Zur Bestimmung der approximierenden empirische Verteilungsfunktion gibt es keine Funktion in R. Also erstellen wir eine. Diese sieht folgendermaßen aus:

```
ecdf.approx function(x,grenzen,haeuf)
{ # Wert der approximierenden empirischen Verteilungsfunktion
  # x: Stelle
 # grenzen: Vektor mit den Klassengrenzen
 # haeuf: Vektor mit relativen Häufigkeiten der Klassen
if(x<grenzen[1]){return(0)}
 else if(x>grenzen[length(grenzen)]){return(1)}
       else return(sum(haeuf[1:(wo-1)])+(x-grenzen[wo])/
            (grenzen[wo+1]-grenzen[wo])*haeuf[wo])}
```

Wir bestimmen also $F_n^*(28) = 0.88$.

```
> grenzen<-seq(20,40,5)
> hi<-c(0.2,0.48,0.16,0.16)
> ecdf.approx(28,grenzen,hi)
[1] 0.488
```

Wie man die approximierende empirische Verteilungsfunktion grafisch darstellt, wird auf Seite 43 gezeigt. Die zugehörige Funktion ist auf Seite 51 zu finden. Wir geben also ein

```
> plot.ecdfapprox(grenzen,hi,"Alter")
```

3.2 Beschreibung univariater Datensätze

Wir haben im letzten Kapitel gelernt, dass man die Verteilung eines Merkmals in Form einer Tabelle oder Grafik darstellen kann. Oft will man die Charakteristika einer Verteilung durch eine oder mehrere Maßzahlen zusammenfassen. Dabei sind folgende Charakteristika von Interesse:

1. Welche Merkmalsausprägung tritt am häufigsten auf?
2. Wo liegt das Zentrum der Verteilung?
3. Wie dicht liegen die Beobachtungen um das Zentrum?

Es gibt noch weitere Eigenschaften einer Verteilung, die durch Maßzahlen beschrieben werden sollen. Dazu gehört die Schiefe. Mit dieser wollen wir uns hier aber nicht beschäftigen.

Die Merkmalsausprägung, deren relative Häufigkeit am größten ist, heißt **Modus**. Dieser kann für jedes Messniveau bestimmt werden. Der Modus ist nicht notwendigerweise eindeutig. So umfasst der Modus bei klassierten Daten alle Werte eines oder mehrerer Intervalle.

Beispiel 16
Das Wahlverhalten der Studienanfängerinnen ist ein nominalskaliertes Merkmal. Die Häufigkeitstabelle ist in Tabelle 3.2 auf Seite 58 zu finden. Der Modus ist offensichtlich die Merkmalsausprägung `weiß nicht`.

Die Bewertung des Films Titanic aus Beispiel 1 auf Seite 10 ist ein ordinalskaliertes Merkmal. Die Häufigkeitstabelle ist in Tabelle 3.4 auf Seite 63 zu finden. Der Modus ist die Merkmalsausprägung `gut`.

Die Anzahl der Geschwister aus Beispiel 2 auf Seite 10 ist ein diskretes Merkmal. Die Häufigkeitstabelle ist in Tabelle 3.6 auf Seite 69 zu finden. Der häufigste Wert ist 1.

Das Alter der Teilnehmer aus Beispiel 1 auf Seite 10 ist stetig. Die Häufigkeitstabelle ist in Tabelle 3.8 auf Seite 74 zu finden. In diesem Fall ist der Modus nicht eindeutig. Jeder Wert aus dem Intervall $(25, 30]$ kommt in Frage. □

3.2.1 Maßzahlen für die Lage

Maßzahlen für die Lage geben an, wo das Zentrum der Verteilung liegt. Es gibt eine Reihe von Möglichkeiten, das Zentrum der Verteilung durch eine Maßzahl zu beschreiben.

Der Median

Wir gehen im Folgenden von einem Merkmal aus, das zumindest ordinalskaliert ist. In diesem Fall kann man den geordneten Datensatz $x_{(1)}, \ldots, x_{(n)}$ bilden. Es liegt nahe, für das Zentrum der Verteilung die Mitte des geordneten Datensatzes zu wählen. Man spricht vom **Median** $x_{0.5}$. Wie die folgende Abbildung zeigt, ist der Median für ungerades n eindeutig definiert.

Ist n also ungerade, so gilt

$$x_{0.5} = x_{(k)}$$

mit $k = (n+1)/2$.

Beispiel 17
Das Alter der Väter von 9 Studierenden ist

```
56 54 50 51 59 62 58 53 60
```

Der geordnete Datensatz ist

```
50 51 53 54 56 58 59 60 62
```

Mit $n = 9$ gilt $k = (9+1)/2 = 5$. Also gilt $x_{0.5} = x_{(5)} = 56$.

□

Ist n gerade, so ist der Median nicht eindeutig bestimmt, wie die folgende Abbildung zeigt.

Ist n gerade, so kann für $k = n/2$ jede Beobachtung im Intervall $\big(x_{(k)}, x_{(k+1)}\big)$ als Median gewählt werden. Es ist üblich, den Mittelpunkt des Intervalls $\big(x_{(k)}, x_{(k+1)}\big)$ zu wählen. Mit $k = n/2$ gilt also

$$x_{0.5} = \frac{x_{(k)} + x_{(k+1)}}{2}$$

Beispiel 18
Das Alter der Väter von 10 Studierenden ist

```
51 65 60 59 68 48 49 54 58 62
```

Der geordnete Datensatz ist

```
48 49 51 54 58 59 60 62 65 68
```

Mit $n = 10$ gilt $k = 10/2 = 5$. Also gilt

$$x_{0.5} = \frac{x_{(5)} + x_{(6)}}{2} = \frac{58 + 59}{2} = 58.5$$

□

Der Median ist für ein quantitatives Merkmal folgendermaßen definiert:

$$x_{0.5} = \begin{cases} x_{((n+1)/2)} & \text{falls } n \text{ ungerade ist} \\ \dfrac{x_{(n/2)} + x_{(1+n/2)}}{2} & \text{falls } n \text{ gerade ist} \end{cases} \tag{3.8}$$

Der Median besitzt folgende Eigenschaft:
Mindestens 50 Prozent der Beobachtungen sind kleiner oder gleich und mindestens 50 Prozent der Beobachtungen sind größer oder gleich dem Median.
Diese Eigenschaft des Medians erlaubt es, den Median auch für nichtmetrische ordinalskalierte Merkmale zu bestimmen.

Beispiel 19
Sechs Personen sollten den Film Titanic bewerten. Hier ist der geordnete Datensatz.

```
sg sg g m s s
```

Dabei steht `sg` für sehr gut, `g` für gut, `m` für mittelmäßig und `s` für schlecht. Wir können `g` als Median wählen, da 50 Prozent der Beobachtungen kleiner oder gleich und 66.7 Prozent Beobachtungen größer oder gleich `g` sind. Wir können aber auch `m` als Median wählen, da 66.7 Prozent der Beobachtungen kleiner oder gleich und 50 Prozent Beobachtungen größer oder gleich `m` sind. Wollen wir einen Wert für den Median angeben, so müssen wir uns für einen der beiden Werte entscheiden und können nicht wie bei einem numerischen Merkmal einen Wert wählen, der zwischen beiden liegt.

□

Liegen die Werte eines diskreten oder eines nicht numerischen ordinalskalierten Merkmals in einer Häufigkeitstabelle mit relativen und kumulierten relativen Häufigkeiten vor, so können wir den Median direkt aus der Häufigkeitstabelle bestimmen. Das folgende Beispiel zeigt, wie wir dabei vorzugehen haben.

Beispiel 20
Wir betrachten das Merkmal `Anzahl Geschwister` im Beispiel 2 auf Seite 10. Die Häufigkeitstabelle ist in Tabelle 3.6 auf Seite auf Seite 69 zu finden. In Tabelle 3.11 ist sie noch einmal zu finden.

Tabelle 3.11: Häufigkeitstabelle des Merkmals `Anzahl Geschwister`

Anzahl Geschwister	absolute Häufigkeit	relative Häufigkeit	kumulierte relative Häufigkeit
0	2	0.10	0.10
1	12	0.60	0.70
2	5	0.25	0.95
3	1	0.05	1.00

Wir sehen, dass 70 Prozent der Beobachtungen kleiner gleich 1 und 90 Prozent der Beobachtungen größer gleich 1 sind. Also erfüllt der Wert 1 die Bedingungen, die der Median erfüllt. Die Werte 0, 2 und 3 erfüllen diese Bedingungen nicht.
Schauen wir uns die kumulierten relativen Häufigkeiten an, so sehen wir, dass 1 der erste Wert ist, für den die kumulierte relative Häufigkeit größer als 0.5 ist.
Bei einer anderen Befragung sollten 10 Studierende angeben, wie viele Geschwister sie haben. Tabelle 3.12 enthält die relativen und kumulierten relativen Häufigkeiten.

Tabelle 3.12: Häufigkeitstabelle des Merkmals `Anzahl Geschwister`

Anzahl Geschwister	absolute Häufigkeit	relative Häufigkeit	kumulierte relative Häufigkeit
0	2	0.2	0.2
1	3	0.3	0.5
2	4	0.4	0.9
3	1	0.1	1.0

50 Prozent der Beobachtungen sind kleiner oder gleich 1 und 80 Prozent der Beobachtungen sind größer oder gleich 1. Also erfüllt der Wert 1 die Bedingungen, die wir an einen Median stellen.
90 Prozent der Beobachtungen sind kleiner oder gleich 2 und 50 Prozent der Beobachtungen sind größer oder gleich 2. Also erfüllt der Wert 2 die Bedingungen, die wir an einen Median stellen.
Außerdem gilt für jedes $x \in (1, 2)$, dass 50 Prozent der Beobachtungen kleiner oder gleich x und 50 Prozent der Beobachtungen größer oder gleich x sind.
Also können wir jedes $x \in [1, 2]$ als Median auffassen. Wir wählen den Mittelpunkt des Intervalls, also 1.5.
Schauen wir uns die kumulierten relativen Häufigkeiten an, so sehen wir, dass die kumulierten relativen Häufigkeiten an der Stelle 1 den Wert 0.5 annehmen. Der Median ist der Mittelwert aus 1 und 2.

□

Bei einem quantitativen oder ordinalskalierten Merkmal mit den Ausprägungen $a_1, a_2, \ldots, a_k$ legt das Beispiel folgende Vorgehensweise bei der Bestimmung des Medians aus der empirischen Verteilungsfunktion $F_n(x)$ nahe:
Gibt es eine Merkmalsausprägung a_i mit $F_n(a_i) = 0.5$, so ist der Median gleich

$$\frac{a_i + a_{i+1}}{2}$$

falls das Merkmal quantitativ ist. Ansonsten können wir a_i oder a_{i+1} als Median wählen.
Gibt es keine Merkmalsausprägung a_i mit $F_n(a_i) = 0.5$, so ist der Median die Merkmalsausprägung, bei der die empirische Verteilungsfunktion zum ersten Mal größer als 0.5 ist.
Wir können den Median auch aus der grafischen Darstellung der empirischen Verteilungsfunktion bestimmen. Hierbei zeichnen wir auf Höhe des Wertes 0.5 eine Parallele zur Abszisse. Es können zwei Fälle eintreten.

1. Die Gerade fällt auf einem Intervall mit der empirischen Verteilungsfunktion zusammen. Dann sind alle Punkte aus dem Intervall mögliche Werte des Medians. Ist das Merkmal numerisch, so ist der Median der Mittelpunkt des Intervalls. Ist das Merkmal hingegen nicht numerisch, so können wir einen der beiden Endpunkte des Intervalls als Median wählen.

2. Die Gerade verläuft zwischen zwei Treppenstufen. In diesem Fall ist der Median die Untergrenze der oberen Treppenstufe.

Bei klassierten Daten ist der Median $x_{0.5}$ der Wert x, für den $F_n^*(x) = 0.5$ gilt, denn genau 50 Prozent der Beobachtungen sind kleiner oder gleich x und genau 50 Prozent der Beobachtungen sind größer oder gleich x. Um ihn zu bestimmen, setzen wir in die Gleichung (3.5) auf Seite 79 für x den Median $x_{0.5}$ ein. Dann gilt wegen $F_n^*(x_{0.5}) = 0.5$:

$$0.5 = F_n^*(x_{i-1}^*) + \frac{x_{0.5} - x_{i-1}^*}{\Delta_i} \cdot f_i \,.$$

Lösen wir diese Gleichung nach $x_{0.5}$ auf, so erhalten wir

$$\boxed{x_{0.5} = x_{i-1}^* + \frac{0.5 - \hat{F}(x_{i-1}^*)}{f_i} \cdot \Delta_i} \tag{3.9}$$

falls gilt $F_n^*(x_{i-1}^*) \leq 0.5 < F_n^*(x_i^*)$.

Beispiel 21
Schauen wir uns noch einmal die Häufigkeitstabelle des Merkmals `Alter` aus Tabelle 1.2 auf Seite 17 an.

Tabelle 3.13: Häufigkeitstabelle des Merkmals `Alter`

i	Intervall	f_i	Δ_i	$F_n^*(x_{i-1}^*)$	$F_n^*(x_i^*)$
1	$[20, 25]$	0.20	5	0	0.20
2	$(25, 30]$	0.48	5	0.20	0.68
3	$(30, 35]$	0.16	5	0.68	0.84
2	$(35, 40]$	0.16	5	0.84	1.00

Der Median liegt in der zweiten Klasse, denn es gilt $F_n^*(25) = 0.2$ und $F_n^*(30) = 0.68$. Somit gilt

$$x_{0.5} = 25 + \frac{0.5 - 0.2}{0.48} \cdot 5 = 28.125$$

Aus dem geordneten Datensatz

```
23 23 23 24 25 26 26 26 27 27 28 28 28
28 29 29 30 31 31 32 33 37 37 38 38
```

erhalten wir 28 als Wert des Medians. Die Werte des Medians für den geordneten Datensatz und die Tabelle mit Klassenbildung unterscheiden sich. Dies

ist kein Wunder, da durch die Klassenbildung Information verloren geht. Wir nehmen ja an, dass die Beobachtungen innerhalb jeder Klasse gleichverteilt sind.

□

Oft werden Beobachtungen $x_1, \ldots, x_n$ **linear transformiert**:

$$y_i = a + b \cdot x_i \tag{3.10}$$

Dabei ist $b > 0$. So will man zum Beispiel Preise von Euro in Dollar oder die Temperatur von Celsius in Fahrenheit umrechnen.

Beispiel 22
Im Beispiel 4 auf Seite 12 haben wir die Höchsttemperatur in Celsius betrachtet. Rechnet man die Temperatur von Celsius x in Fahrenheit y um, so bildet man

$$y = 32 + 1.8 \cdot x$$

Der erste Wert der Temperatur in Celsius ist 17. Wir erhalten als Wert in Fahrenheit

$$33 + 1.8 \cdot 17 = 62.6$$

Hier sind sortierten transformierten Werte:

```
50.0 51.8 53.6 55.4 57.2 59.0 59.0 59.0 59.0 60.8
62.6 62.6 62.6 62.6 62.6 64.4 64.4 66.2 68.0 69.8
```

Der Median der Temperatur in Fahrenheit ist

$$\frac{60.8 + 62.6}{2} = 61.7$$

□

Wir haben im Beispiel den Median der linear transformierten Beobachtungen dadurch bestimmt, dass wir jede Beobachtung transformiert haben und den Median der transformierten Werte bestimmt haben. Dies ist aber nicht nötig. Ist $x_{0.5}$ nämlich der Median der Beobachtungen $x_1, \ldots, x_n$, so ist $a + b\,x_{0.5}$ der Median der Beobachtungen $a + b\,x_i$, $i = 1, \ldots, n$.

Beispiel 22 (fortgesetzt)
Der Median der Temperatur in Celsius ist 16.5. Also ist der Median der Temperatur in Fahrenheit

$$32 + 1.8 \cdot 16.5 = 61.7$$

□

Der Mittelwert

Eine andere Maßzahl für die Lage ist der **Mittelwert**. Dieser verteilt die Summe aller Beobachtungen gleichmäßig auf alle Merkmalsträger. Für die Beobachtungen $x_1, x_2, \ldots, x_n$ ist der Mittelwert $\bar{x}$ also folgendermaßen definiert:

$$\boxed{\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i} \tag{3.11}$$

Man bezeichnet den Mittelwert auch als **arithmetisches Mittel**.

Beispiel 20 (fortgesetzt von Seite 89)
Wir betrachten das Merkmal `Anzahl Geschwister` im Beispiel 2 auf Seite 10. Die Daten stehen in der sechsten Spalte von Tabelle 1.3 auf Seite 18. Die Urliste sieht folgendermaßen aus:

```
1 2 1 3 0 2 2 1 1 1 1 1 1 2 1 1 0 1 2 1
```

Es gilt $\bar{x} = 1.25$.

□

Das Beispiel zeigt, dass der Mittelwert einen Wert annehmen kann, der nicht zu den Ausprägungen des Merkmals gehört. Niemand hat 1.25 Geschwister.
Bevor wir uns weitere Eigenschaften des Mittelwerts anschauen, zeigen wir, wie wir den Mittelwert aus einer Häufigkeitstabelle bestimmen können.
Beginnen wir mit diskreten Merkmalen. Es liegen also die Merkmalsausprägungen $a_1, \ldots, a_k$ mit den absoluten Häufigkeiten $n_1, \ldots, n_k$ und den relativen Häufigkeiten $f_1, \ldots, f_k$ vor. Dann ist der Mittelwert $\overline{x}$ gleich:

$$\begin{aligned}
\overline{x} &= \frac{1}{n} \left(\underbrace{a_1 + \ldots + a_1}_{n_1-mal} + \ldots + \underbrace{a_k + \ldots + a_k}_{n_k-mal} \right) \\
&= \frac{1}{n} \left(a_1\, n_1 + \ldots + a_k\, n_k \right) = a_1 \frac{n_1}{n} + \ldots + a_k \frac{n_k}{n} \\
&= a_1\, f_1 + \ldots + a_k\, f_k
\end{aligned}$$

Es gilt also

$$\boxed{\overline{x} = \sum_{i=1}^{k} a_i\, f_i} \tag{3.12}$$

Beispiel 20 (fortgesetzt)
Die Häufigkeitstabelle des Merkmals `Anzahl Geschwister` ist in Tabelle 3.6 auf Seite 69 zu finden. Es gilt

$$\overline{x} = 0 \cdot 0.1 + 1 \cdot 0.6 + 2 \cdot 0.25 + 3 \cdot 0.05 = 1.25$$

□

Schauen wir uns ein stetiges Merkmal an. Hier liegen nur relative Häufigkeiten der Klassen vor. Mit welchem Wert sollen wir die relativen Häufigkeiten multiplizieren? Da die empirische Dichtefunktion in jeder Klasse konstant ist, wählen wir den Mittelpunkt m_i der i-ten Klasse. Es gilt

$$\boxed{m_i = \frac{x^*_{i-1} + x^*_i}{2}}$$

Den Mittelwert bestimmen wir dann durch

$$\boxed{\overline{x} = \sum_{i=1}^{k} m_i \, f_i} \tag{3.13}$$

Beispiel 21 (fortgesetzt von Seite 91)
Wir schauen uns das Alter der Teilnehmer aus Beispiel 10 auf Seite 1 an. Hier ist noch einmal die Häufigkeitstabelle

Tabelle 3.14: Die Häufigkeitstabelle des Merkmals Alter

Alter	relative Häufigkeit
von 20 bis 25	0.20
von 25 bis 30	0.48
von 30 bis 35	0.16
von 35 bis 40	0.16

Die Klassenmitten sind

$$m_1 = \frac{20 + 25}{2} = 22.5 \qquad m_2 = \frac{25 + 30}{2} = 27.5$$

$$m_3 = \frac{30 + 35}{2} = 32.5 \qquad m_4 = \frac{35 + 40}{2} = 37.5$$

Somit gilt

$$\bar{x} = 22.5 \cdot 0.2 + 27.5 \cdot 0.48 + 32.5 \cdot 0.16 + 37.5 \cdot 0.16 = 28.9$$

Wir hätten den Mittelwert auch aus den Originaldaten bestimmen können. Für diese beträgt er 29.08 Jahre. Wir sehen, dass die Werte sich unterscheiden. Dies ist kein Wunder, denn bei der Klassenbildung geht Information verloren.

□

Wie beim Median muss man beim Mittelwert nicht alle Beobachtungen linear transformieren, um den Mittelwert der linear transformierten Beobachtungen zu bestimmen. Man kann ihn einfach aus dem Mittelwert $\bar{x}$ bestimmen. Es gilt

$$\boxed{\overline{y} = a + b\,\overline{x}} \tag{3.14}$$

Die Gültigkeit der Beziehung (3.14) wird auf Seite 126 gezeigt.

Beispiel 22 (fortgesetzt von Seite 92)
Im Beispiel 4 auf Seite 12 haben wir die Höchsttemperatur in Celsius betrachtet. Die Durchschnittstemperatur in Celsius beträgt $\overline{x} = 15.85$. Also beträgt der Mittelwert $\bar{y}$ der Temperatur in Fahrenheit

$$\overline{y} = 32 + 1.8\,\overline{x} = 28.53$$

□

Häufig werden die Beobachtungen $x_1, \ldots, x_n$ zentriert. Dabei subtrahiert von jeder Beobachtung x_i den Mittelwert $\bar{x}$ aller Beobachtungen:

$$\boxed{\tilde{x}_i = x_i - \overline{x}} \tag{3.15}$$

An den **zentrierten Beobachtungen** kann man sofort erkennen, ob die Beobachtungen unter oder über dem Mittelwert liegen:
Sind zentrierte Beobachtungen negativ, so sind die zugehörigen Beobachtungen kleiner als der Mittelwert. Sind zentrierte Beobachtungen hingegen positiv, so sind die zugehörigen Beobachtungen größer als der Mittelwert.

Beispiel 20 (fortgesetzt von Seite 94)
Wir betrachten das Merkmal `Anzahl Geschwister` im Beispiel 2 auf Seite 10. Die Urliste sieht folgendermaßen aus:

```
1 2 1 3 0 2 2 1 1 1 1 1 1 2 1 1 0 1 2 1
```

Der Mittelwert ist $\bar{x} = 1.25$.
Somit erhalten wir folgende zentrierte Beobachtungen:

```
-0.25  0.75 -0.25  1.75 -1.25  0.75  0.75 -0.25 -0.25 -0.25
-0.25 -0.25 -0.25  0.75 -0.25 -0.25 -1.25 -0.25  0.75 -0.25
```

□

Der Mittelwert der zentrierten Beobachtungen ist gleich 0. Es gilt also:

$$\sum_{i=1}^{n} \tilde{x}_i = \sum_{i=1}^{n} (x_i - \overline{x}) = 0 \tag{3.16}$$

Dies wird auf Seite 126 gezeigt.

Vergleich von Mittelwert und Median

Worin unterscheiden sich Mittelwert und Median? Schauen wir uns ein Beispiel an. Wir fragen 5 Personen nach ihrem Nettoeinkommen in Euro und erhalten folgende Werte:

```
1000  1000 1500 2000 7000
```

Der Mittelwert beträgt 2500 Euro und der Median 1500 Euro. Der Median gibt die Lage der Verteilung viel besser wieder als der Mittelwert, da die meisten Beobachtungen um den Wert 1500 Euro liegen. Der Mittelwert wird durch die Beobachtung 7000 Euro beeinflusst. Sie zieht den Mittelwert zu sich. Den Median hingegen beeinflusst sie nicht. Extreme Beobachtungen haben einen großen Einfluss auf den Mittelwert. Der Mittelwert ist **ausreißerempfindlich**, der Median hingegen nicht. Wir sagen auch, dass der Median **robust** ist.
Schauen wir uns die relative Lage von Mittelwert und Median bei symmetrischen und schiefen Verteilungen an.
Beginnen wir mit einer symmetrischen Verteilung. Das Bild links unten in Abbildung 3.11 auf Seite 76 zeigt das Histogramm der Körpergröße von Studenten. Aus Tabelle 3.19 auf Seite 129 können wir den Mittelwert und Median bestimmen. Der Mittelwert ist 183.1 cm und der Median 183 cm. Dies ist das Symmetriezentrum der Verteilung. Bei einer symmetrischen Verteilung nehmen Mittelwert und Median ähnliche Werte an.

Schauen wir uns eine rechtsschiefe Verteilung an. Wir betrachten das Bild links oben in Abbildung 3.11 auf Seite 76. Aus den Daten in Tabelle 3.17 auf Seite 128 erhalten wir für den Mittelwert den Wert 63 und für den Median den Wert 43.5. Dies gilt auch allgemein. Bei einer rechtsschiefen Verteilung ist der Mittelwert immer größer als der Median.

Bei einer linksschiefen Verteilung ist der Mittelwert immer kleiner als der Median. Ein Beispiel ist im Bild rechts oben in Abbildung 3.11 auf Seite 76 zu finden. Aus den Daten in Tabelle 3.18 auf Seite 128 erhalten wir für den Mittelwert den Wert 493.2 und für den Median den Wert 503.

Zwischen dem Mittelwert $\overline{x}$ und dem Median $x_{0.5}$ besteht also in Abhängigkeit von der Schiefe der Verteilung folgende Beziehung:

rechtsschief	$\overline{x} > x_{0.5}$
symmetrisch	$\overline{x} = x_{0.5}$
linksschief	$\overline{x} < x_{0.5}$

Auf Seite 77 haben wir offene Randklassen betrachtet. Eine offene Randklasse besitzt keinen Mittelpunkt. Deshalb kann der Mittelwert in diesem Fall nicht bestimmt werden.

Beispiel 15 (fortgesetzt von Seite 77)

Wir schauen uns noch einmal das monatliche Nettohaushaltseinkommen der Privathaushalte in Deutschland im Jahr 2003 an. Die Daten sind in Tabelle 3.15 auf der nächsten Seite zu finden. Wir wollen zeigen, wie stark der Mittelwert von der Wahl der Obergrenze der letzten Klasse abhängt.

Die Mitten der ersten sieben Klassen sind $m_1 = 250$, $m_2 = 700$, $m_3 = 1100$, $m_4 = 1400$, $m_5 = 1750$, $m_6 = 2300$ und $m_7 = 3550$.

Wir wählen als Obergrenze der letzten Klasse den Wert 10000. Also ist die Mitte dieser Klasse gleich 7250. Es gilt

$$\begin{aligned} \bar{x} &= 250 \cdot 0.035 + 700 \cdot 0.117 + 1100 \cdot 0.17 + 1400 \cdot 0.089 + 1750 \cdot 0.175 \\ &+ 2300 \cdot 0.162 + 3550 \cdot 0.197 + 7250 \cdot 0.055 \\ &= 2179.2 \end{aligned}$$

Mit dem Wert 20000 gilt

$$\bar{x} = 2454.2$$

Wir können also durch die Wahl der Obergrenze der letzten Klasse das Durchschnittseinkommen der Haushalte beliebig vergrößern.

Der Median ist hier eine sinnvollere Maßzahl für die Lage. Tabelle 3.15 auf der nächsten Seite zeigt die neben den relativen Häufigkeiten noch die kumulierten relativen Häufigkeiten.

Tabelle 3.15: Monatliches Nettohaushaltseinkommen der Privathaushalte in Deutschland im Jahr 2003

Klasse i	Intervall	f_i	$F_n^*(x_{i-1}^*)$	$F_n^*(x_i^*)$
1	$(0, 500]$	0.035	0.000	0.035
2	$(500, 900]$	0.117	0.035	0.152
3	$(900, 1300]$	0.170	0.152	0.322
4	$(1300, 1500]$	0.089	0.322	0.411
5	$(1500, 2000]$	0.175	0.411	0.586
6	$(2000, 2600]$	0.162	0.586	0.748
7	$(2600, 4500]$	0.197	0.748	0.945
8	$(4500, -\infty)$	0.055	0.945	1.000

Wegen $F_n^*(1500) = 0.411$ und $F_n^*(2000) = 0.586$ liegt der Median der fünften Klasse. Somit gilt

$$x_{0.5} = 1500 + \frac{0.5 - 0.411}{0.175} \cdot 500 = 1754.29$$

□

Getrimmter Mittelwert

Der Mittelwert ist empfindlich gegenüber extremen Beobachtungen. Maßzahlen, bei denen dies nicht der Fall ist, heißen **robust**. Der Median ist eine robuste Maßzahl der Lage . Er ist ein spezieller **getrimmter Mittelwert**. Bei diesen wird einer fester Anteil α vorgegeben und von beiden Rändern der geordneten Stichprobe $x_{(1)}, \ldots, x_{(n)}$ jeweils $\lfloor n \cdot \alpha \rfloor$ Beobachtungen entfernt. Dabei ist $\lfloor u \rfloor$ der ganzzahlige Anteil von u mit $u \geq 0$, z.B. $\lfloor 2.3 \rfloor = 2$ und $\lfloor 2.8 \rfloor = 2$. Dann wird der Mittelwert der Beobachtungen bestimmt, die nicht entfernt wurden. Also gilt

$$\bar{x}_\alpha = \frac{1}{n - 2\lfloor n\,\alpha \rfloor} \sum_{i=1+\lfloor n\alpha \rfloor}^{n-\lfloor n\alpha \rfloor} x_{(i)} \tag{3.17}$$

Der Mittelwert $\bar{x}$ ist ein spezieller getrimmter Mittelwert mit $\alpha = 0$ und der Median $x_{0.5}$ mit $\alpha = 0.5$.

Beispiel 23
Wir schauen uns das Alter der Teilnehmer aus Beispiel 1 auf Seite 10 an. Hier ist der geordnete Datensatz

```
23 23 23 24 25 26 26 26 27 27 28 28 28
28 29 29 30 31 31 32 33 37 37 38 38
```

Für $\alpha = 0.2$. Mit $n = 25$ gilt $\lfloor 25 \cdot 0.2 \rfloor = 5$. Wir entfernen also 5 Beobachtungen von den Rändern der geordneten Stichprobe. Es bleiben folgende 15 Beobachtungen übrig:

```
26 26 26 27 27 28 28 28 28 29 29 30 31 31 32
```

Der Mittelwert dieser Beobachtungen ist $x_{0.2} = 28.4$. Wir bestimmen noch $x_{0.05}$, $x_{0.1}$ und $x_{0.25}$. Es gilt

$$\begin{aligned} x_{0.05} &= 28.96 \\ x_{0.1} &= 28.81 \\ x_{0.25} &= 28.31 \end{aligned}$$

□

Es stellt sich natürlich die Frage, welcher getrimmte Mittelwert einen Datensatz am besten beschreibt Diese Frage werden wir in Kapitel 13 beantworten.
Der getrimmte Mittelwert ist ein spezieller **gewichteter Mittelwert**. Ein gewichteter Mittelwert $\bar{X}_W$ ist definiert durch

$$\boxed{\bar{X}_W = \sum_{i=1}^{n} w_i \, X_i} \tag{3.18}$$

mit

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

Beim getrimmten Mittelwert gilt

$$w_i = \begin{cases} \dfrac{1}{n - 2\lfloor n\,\alpha \rfloor} & \text{für } 1 + \lfloor n\alpha \rfloor \leq i \leq n - \lfloor n\alpha \rfloor \\ 0 & \text{sonst} \end{cases}$$

und beim Mittelwert

$$w_i = \frac{1}{n}$$

Das folgende Beispiel orientiert sich an Lann & Falk (2005).

Beispiel 24
In einer Schule werden drei weiterführende Kurse angeboten, an denen 6, 9 und 15 Schüler teilnehmen.
Fragt man den Direktor nach dem Mittelwert je Kurs, so erhält man 10 als Antwort. Fragt man hingegen jeden Schüler, wie viele Schüler in seinem Kurs sind, und bildet den Mittelwert dieser Zahlen, so ergibt sich 11.4.
Im ersten Fall bilden wir

$$\frac{6+9+15}{3} = 10$$

Im zweiten Fall geben 6 Schüler die Antwort 6, 9 Schüler die Antwort 9 und 15 Schüler die Antwort 15. Der Mittelwert dieser 30 Zahlen ist

$$\frac{6*6+9*9+15*15}{30} = 11.4$$

Aus Sicht der Lehrer ist die durchschnittliche Kursstärke 10 und aus Sicht der Schüler ist sie 11.4. □

Der im zweiten Fall des Beispiels behandelte Mittelwert heißt **selbstgewichteter** Mittelwert $\bar{x}_{sw}$. Für ihn gilt $w_i = x_i$ und somit

$$\bar{x}_{sw} = \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i} \tag{3.19}$$

Maßzahlen für die Lage in R

Wir betrachten das Merkmal `Alter`aus Tabelle 1.2 auf Seite 17. Die Daten der Tabelle mögen in der Datentabelle `weiterbildung` stehen, auf die wir wie

```
> attach(weiterbildung)
> Alter
 [1] 30 23 26 33 37 28 31 23 24 26 23 32 29 25 31 26 37 38 29
     28 28 28 38 27 27
```

`mean` Den Mittelwert bestimmen wir mit der Funktion `mean` und den Median mit
`median` Funktion `median`.

```
> mean(Alter)
[1] 29.08
> median(Alter)
[1] 28
```

Die Funktionen `mean` und `median` besitzen das optionale Argument `na.rm`. Dieses steuert den Umgang mit fehlenden Beobachtungen im Vektor `x`. Bei fast jeder statistischen Erhebung fehlen Beobachtungen, da Befragte keine Antwort gegeben haben oder Versuche abgebrochen werden mussten. In `R` gibt man fehlende Beobachtungen als `NA` ein. Dies steht für `not available`. Sind im Vektor `x` fehlende Werte, so wird der Mittelwert ohne diese bestimmt, wenn das Argument `na.rm` auf `TRUE` steht. Nimmt es den Wert `FALSE` an, so liefert die Funktion `mean` als Ergebnis den Wert `NA`. Schauen wir uns dies für ein Beispiel an.

```
> z<-c(1,NA,5,3)
> z
[1]  1 NA  5  3
> mean(z)
[1] NA
> mean(z,na.rm=TRUE)
[1] 3
```

Die Funktion `mean` besitzt außerdem noch das optionale Argument `trim`, das standardmäßig den Wert 0 annimmt. Setzt man es auf einen Wert, der größer als 0 aber kleiner als 0.5 ist, so wird der entsprechende getrimmte Mittelwert bestimmt. Den 0.2-getrimmten Mittelwert erhält man durch

```
> mean(Alter,trim=0.2)
[1] 28.4
```

Für die Daten aus Beispiel 24 auf der vorherigen Seite erhalten wir den selbstgewichteten Mittelwert folgendermaßen:

```
> anz<-c(6,9,15)
> sum(anz^2)/sum(anz)
[1] 11.4
```

Liegt eine Häufigkeitstabelle mit klassierten Beobachtungen vor, so wird die Berechnung von Median und Mittelwert durch `R` nicht durch Funktionen unterstützt. Sind die Klassengrenzen und die relativen Häufigkeiten der Klassen gegeben, so kann man Median und Mittelwert mit Hilfe einiger Befehle leicht bestimmen.

Wir betrachten die Tabelle 3.8 auf Seite 74. Die Klassengrenzen sind 20, 25, 30, 35 und 40 und die relativen Häufigkeiten der Klassen 0.2, 0.48, 0.16 und 0.16. Wir müssen die Mitten der Klassen mit ihren relativen Häufigkeiten multiplizieren. Die Summe dieser Produkte ergibt den Mittelwert. Wir geben zuerst die Daten ein.

```
> grenzen<-seq(20,40,5)
> hi<-c(0.2,0.48,0.16,0.16)
```

Die Klassenmitten erhalten wir, indem wir die Summe benachbarter Komponenten durch zwei dividieren. Benachbarte Komponenten erhalten wir dadurch, dass wir zum einen die erste und zum anderen die letzte Komponente des Vektors `grenzen` entfernen

```
> grenzen[-1]
[1] 25 30 35 40
> grenzen[-length(grenzen)]
[1] 20 25 30 35
```

und dann die Summe dieser Vektoren durch 2 dividieren

```
> (grenzen[-length(grenzen)]+grenzen[-1])/2
[1] 22.5 27.5 32.5 37.5
```

Diesen Vektor müssen wir mit dem Vektor der relativen Häufigkeiten multiplizieren und die Summe der Produkte bilden.

```
> sum(((grenzen[-length(grenzen)]+grenzen[-1])/2)*hi)
[1] 28.9
```

Zur Bestimmung des Medians verwenden wir die folgende Funktion `quantil.approx`:

```
quantil.approx<-function(p,grenzen,haeuf)
 { # Quantil bei klassierten Daten
   # p: Vektor der Wahrscheinlichkeiten
   # grenzen: Vektor mit den Klassengrenzen
   # haeuf: Vektor mit relativen Häufigkeiten der Klassen
   # Ergebnis ist Quantil
     if(any(p<=0) | any(p>=1))
     return("Fehlerhafte Eingabe: Es muss gelten 0<p<1")
     cumhaeuf<-cumsum(c(0,haeuf))
     wo<-apply(outer(p,cumhaeuf,">="),1,sum)
     return(grenzen[wo]+(p-cumhaeuf[wo])/haeuf[wo]*
     (grenzen[wo+1]-grenzen[wo]))
 }
```

Für das Merkmal `Alter` gilt

```
> quantil.approx(0.5,grenzen,hi)#
[1] 28.125
```

3.2.2 Quantile

Mindestens 50 Prozent der Beobachtungen sind kleiner oder gleich dem Median $x_{0.5}$ und mindestens 50 Prozent der Beobachtungen sind größer oder gleich dem Median $x_{0.5}$. Ersetzen wir bei dieser Definition die 0.5 durch eine Zahl p mit $0 < p < 1$, so erhalten wir das **Quantil** x_p:
Mindestens $100 \cdot p$ Prozent der Beobachtungen sind kleiner oder gleich dem Quantil x_p und mindestens $100 \cdot (1 - p)$ Prozent der Beobachtungen sind größer oder gleich dem Quantil x_p.
Um x_p aus der geordneten Stichprobe zu bestimmen, schauen wir uns die Vorgehensweise bei der Bestimmung des Medians unter einem Blickwinkel an, der den Wert $p = 0.5$ in den Vordergrund stellt. Wir berechnen $k = n \cdot 0.5$. Ist k eine natürliche Zahl, so ist der Stichprobenumfang gerade. Der Median ist

$$x_{0.5} = \frac{x_{(k)} + x_{(k+1)}}{2}$$

Ist k aber keine natürliche Zahl, dann muss der Stichprobenumfang ungerade sein, und der Median ist

$$x_{0.5} = x_{(\lfloor k \rfloor + 1)}$$

Dabei ist $\lfloor k \rfloor$ der ganzzahlige Anteil von k.
Ersetzen wir bei dieser Vorgehensweise 0.5 durch p, dann haben wir die Vorschrift zur Bestimmung des Quantils x_p gefunden.
Wir bestimmen also $k = n \cdot p$. Ist k eine natürliche Zahl, dann gilt

$$\boxed{x_p = \frac{x_{(k)} + x_{(k+1)}}{2}} \tag{3.20}$$

Ist k aber keine natürliche Zahl, dann muss der Stichprobenumfang ungerade sein. In diesem Fall gilt

$$\boxed{x_p = x_{(\lfloor k \rfloor + 1)}} \tag{3.21}$$

Die Quantile $x_{0.25}$ und $x_{0.75}$ heißen **Quartile**, wobei $x_{0.25}$ das **untere Quartil** und $x_{0.75}$ das **obere Quartil** heißt.

Beispiel 25
Wir betrachten das Alter der Teilnehmer der Weiterbildungsveranstaltung aus Beispiel 1 auf Seite 10 und wollen das untere und das obere Quartil bestimmen. Der geordnete Datensatz ist

```
23 23 23 24 25 26 26 26 27 27 28 28 28
28 29 29 30 31 31 32 33 37 37 38 38
```

Es gilt $k = 25 \cdot 0.25 = 6.25$. Also gilt $x_{0.25} = x_{(7)} = 26$. Außerdem gilt $k = 25 \cdot 0.75 = 18.75$. Also gilt $x_{0.75} = x_{(19)} = 31$.
Wir bestimmen noch $x_{0.2}$. Es gilt $k = 25 \cdot 0.2 = 5$. Also gilt

$$x_{0.2} = \frac{x_{(5)} + x_{(6)}}{2} = \frac{25 + 26}{2} = 25.5$$

□

Ist das Merkmal stetig und liegen die Daten in Form einer Häufigkeitstabelle mit Klassen vor, so gilt $F_n^*(x_p) = p$. Setzen wir in die Gleichung (3.5) auf Seite 79 für x das Quantil x_p ein, so gilt mit $F_n^*(x_p) = p$:

$$p = F_n^*(x_{i-1}^*) + \frac{x_p - x_{i-1}^*}{\Delta_i} \cdot f_i \,.$$

Lösen wir diese Gleichung nach x_p auf, so erhalten wir

$$\boxed{x_p = x_{i-1}^* + \frac{p - F_n^*(x_{i-1}^*)}{f_i} \cdot \Delta_i} \tag{3.22}$$

falls gilt $F_n^*(x_{i-1}^*) \leq p < F_n^*(x_i^*)$.

Beispiel 26
Wir schauen uns das Alter der Teilnehmer der Weiterbildungsveranstaltung aus Beispiel 1 auf Seite 10 an und betrachten jetzt die klassierten Daten. Hier ist die Häufigkeitstabelle

Tabelle 3.16: Häufigkeitstabelle des Merkmals `Alter`

i	Intervall	f_i	Δ_i	$F_n^*(x_{i-1}^*)$	$F_n^*(x_i^*)$
1	$[20, 25]$	0.20	5	0	0.20
2	$(25, 30]$	0.48	5	0.20	0.68
3	$(30, 35]$	0.16	5	0.68	0.84
2	$(35, 40]$	0.16	5	0.84	1.00

Wegen $F_n^*(25) = 0.2$ und $F_n^*(30) = 0.68$ liegt das untere Quartil in der zweiten Klasse. Also gilt

$$x_{0.25} = 25 + \frac{0.25 - 0.2}{0.48} \cdot 5 = 25.52\,.$$

□

Quantile in R

Um die Quantile wie in den Gleichungen (3.20) und (3.21) auf der Seite 103 zu bestimmen, rufen wir die Funktion `quantile`, die folgendermaßen aufgerufen wird:

```
quantile(x,probs=seq(0,1,0.25),na.rm=F,type=7)
```

auf. quantile

Dabei ist `x` der Vektor mit den Daten, `probs` die Werte von p, für die die Quantile bestimmt werden sollen, und `n.rm` steuert wie üblich den Umgang mit fehlenden Beobachtungen. Mit dem Argument `type` können wir steuern, wie die Quantile bestimmt werden. Bei dem im Skript beschriebenen Verfahren, müssen wir `type` auf den Wert 2 setzen.
Den $x_{0.2}$, $x_{0.25}$ und $x_{0.75}$ erhalten wir durch

```
> attach(weiterbildung)
> quantile(Alter,probs=c(0.2,0.25,0.75),type=2)
 20%  25%  75%
25.5 26.0 31.0
```

Zur Bestimmung von x_p verwenden wir die Funktion `quantil.approx`, die auf Seite 102 zu finden ist. Für das Merkmal `Alter` erhalten wir die Quartile durch

```
> grenzen<-seq(20,40,5)
> hi<-c(0.2,0.48,0.16,0.16)
> quantil.approx(c(0.25,0.75),grenzen,hi)
[1] 25.52083 32.18750
```

3.2.3 Maßzahlen für die Variabilität

Neben der Lage einer Verteilung ist noch von Interesse, wie dicht die Beobachtungen um den Lageparameter liegen. Man spricht auch von **Streuung**.

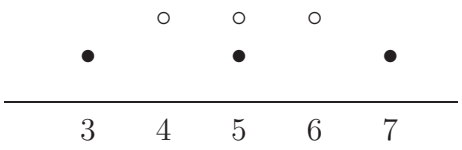
Die Stichprobenvarianz

Um eine Maßzahl für die Streuung zu gewinnen, bestimmen wir den Abstand $|x_i - \bar{x}|$ jeder Beobachtung vom Mittelwert. Es liegt nahe, den Mittelwert dieser Abstände als Maßzahl für die Streuung zu wählen:

$$\boxed{\frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}|} \tag{3.23}$$

Diese heißt **mittlere absolute Abweichung**.

Beispiel 27
Bei einer Befragung wurde eine Gruppe von drei Kindern nach der Höhe ihres monatlichen Taschengelds gefragt. Es beträgt 4, 5 und 6 EURO. Bei einer zweiten Gruppe wurden die Beträge 3, 5 und 7 EURO genannt.
Der Mittelwert des Taschengelds beträgt in beiden Gruppen 5 EURO.
Die folgende Abbildung zeigt, dass die Werte der ersten Gruppe viel dichter um den Mittelwert liegen als die Werte der zweiten Gruppe. Wir sagen auch, dass die Werte in der ersten Gruppe weniger streuen.



In der ersten Gruppe gilt

$$\begin{aligned} |x_1 - \bar{x}| &= |4-5| = 1 \\ |x_2 - \bar{x}| &= |5-5| = 0 \\ |x_3 - \bar{x}| &= |6-5| = 1 \end{aligned}$$

In der zweiten Gruppe gilt

$$\begin{aligned} |x_1 - \bar{x}| &= |3-5| = 2 \\ |x_2 - \bar{x}| &= |5-5| = 0 \\ |x_3 - \bar{x}| &= |7-5| = 2 \end{aligned}$$

In der ersten Gruppe gilt

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}| = \frac{1}{3}(1+0+1) = \frac{2}{3}$$

In der zweiten Gruppe gilt

$$\frac{1}{n}\sum_{i=1}^{n}|x_i-\overline{x}| = \frac{1}{3}(2+0+2) = \frac{4}{3}$$

□

Wählt man anstatt der Abstände die quadrierten Abstände, so erhält man die **mittlere quadratische Abweichung** d^2:

$$\boxed{d^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2} \tag{3.24}$$

Man kann die mittlere quadratische Abweichung auch einfacher berechnen. Es gilt

$$\boxed{d^2 = \overline{x^2} - \overline{x}^2} \tag{3.25}$$

Dabei ist

$$\boxed{\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n}x_i^2}$$

der Mittelwert der quadrierten Beobachtungen. Die Gültigkeit von Gleichung (3.25) wird auf Seite 126 gezeigt.
In der Regel dividiert man die Summe der quadrierten Abweichungen aber nicht durch n, sondern durch $n-1$. In diesem Fall erhält man die **Stichprobenvarianz** s^2.

$$\boxed{s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2} \tag{3.26}$$

Warum es sinnvoller ist, durch $n-1$ und nicht durch n zu dividieren, werden wir später sehen. Für große Werte von n unterscheiden sich d^2 und s^2 kaum. Zwischen der mittleren quadratischen Abweichung d^2 und der Stichprobenvarianz s^2 besteht folgender Zusammenhang:

$$\boxed{s^2 = \frac{n}{n-1}\,d^2} \tag{3.27}$$

Dies sieht man folgendermaßen:

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

Die Stichprobenvarianz weist nicht nicht die Maßeinheiten der Beobachtungen auf. Die Quadratwurzel aus der Stichprobenvarianz heißt **Standardabweichung** und weist die Maßeinheiten der Beobachtungen auf.

$$\boxed{s = \sqrt{s^2}} \tag{3.28}$$

Wir haben auf Seite 95 zentrierte Beobachtungen betrachtet. Teilt man die zentrierten Beobachtungen durch die Standardabweichung, so erhält man **standardisierte Beobachtungen**:

$$\boxed{z_i = \frac{x_i - \bar{x}}{s}} \tag{3.29}$$

Der Mittelwert eines standardisierten Merkmals ist gleich 0 und die Stichprobenvarianz der standardisierten Merkmale ist gleich 1. Dies wird auf Seite 126 gezeigt.

Beispiel 27 (fortgesetzt von Seite auf Seite 106)
Wir schauen uns das Taschengeld der Kinder an. In der ersten Gruppe gilt

$$d^2 = \frac{1}{3} \left[(4-5)^2 + (5-5)^2 + (6-5)^2\right] = \frac{2}{3}.$$

In der zweiten Gruppe gilt

$$d^2 = \frac{1}{3} \left[(3-5)^2 + (5-5)^2 + (7-5)^2\right] = \frac{8}{3}.$$

Wir bestimmen d^2 in der ersten Gruppe mit der Formel in Gleichung (3.25) auf Seite 107. Es gilt $\overline{x} = 5$. Außerdem gilt

$$\overline{x^2} = \frac{1}{3} \left(4^2 + 5^2 + 6^2\right) = \frac{77}{3}$$

Also gilt

$$d^2 = \overline{x^2} - \overline{x}^2 = \frac{77}{3} - 5^2 = \frac{2}{3}$$

Analog erhalten wir den Wert von d^2 in der zweiten Gruppe.
Wir bestimmen die Stichprobenvarianz mit der Formel in Gleichung (3.26) auf Seite 107.
In der ersten Gruppe gilt

$$s^2 = \frac{1}{2}\left[(4-5)^2 + (5-5)^2 + (6-5)^2\right] = 1\,.$$

In der zweiten Gruppe gilt

$$s^2 = \frac{1}{2}\left[(3-5)^2 + (5-5)^2 + (7-5)^2\right] = 4\,.$$

Wir bestimmen die Stichprobenvarianz mit der Formel in Gleichung (3.27) auf Seite 107. In der ersten Gruppe gilt

$$s^2 = \frac{3}{2} \cdot \frac{2}{3} = 1\,.$$

In der zweiten Gruppe gilt

$$s^2 = \frac{3}{2} \cdot \frac{8}{3} = 4\,.$$

Die Stichprobenvarianz des Taschengelds in der ersten Gruppe beträgt 1. Also beträgt die Standardabweichung ebenfalls 1. Die Stichprobenvarianz des Taschengelds in der zweiten Gruppe beträgt 4. Also beträgt die Standardabweichung 2.
Die standardisierten Werte der ersten Gruppe sind

$$z_1 = \frac{4-5}{1} = -1 \qquad z_2 = \frac{5-5}{1} = 0 \qquad z_3 = \frac{6-5}{1} = 1$$

Die standardisierten Werte der zweiten Gruppe sind

$$z_1 = \frac{3-5}{2} = -1 \qquad z_2 = \frac{5-5}{2} = 0 \qquad z_3 = \frac{7-5}{2} = 1$$

□

Schauen wir uns an, wie man die Stichprobenvarianz aus einer Häufigkeitstabelle bestimmt. Hierbei verwendet man am besten Gleichung (3.25) auf Seite 107. Man bestimmt also $\bar{x}$ und $\overline{x^2}$.
Ist das Merkmal diskret mit den Merkmalsausprägungen $a_1, a_2, \ldots, a_k$ und den relativen Häufigkeiten $f_1, f_2, \ldots, f_k$, so berechnen wir

$$\boxed{\overline{x} = \sum_{i=1}^{k} a_i\, f_i}$$

und

$$\boxed{\overline{x^2} = \sum_{i=1}^{k} a_i^2 \, f_i}$$

Die mittlere quadratische Abweichung ist

$$\boxed{d^2 = \overline{x^2} - \overline{x}^2}$$

Beispiel 28
Die Häufigkeitstabelle des Merkmals `Anzahl Geschwister` ist in Tabelle 3.6 auf Seite 69 zu finden. Es gilt

$$\overline{x} = 0 \cdot 0.1 + 1 \cdot 0.6 + 2 \cdot 0.25 + 3 \cdot 0.05 = 1.25$$

und

$$\overline{x^2} = 0^2 \cdot 0.1 + 1^2 \cdot 0.6 + 2^2 \cdot 0.25 + 3^2 \cdot 0.05 = 2.05$$

Somit gilt

$$d^2 = \overline{x^2} - \overline{x}^2 = 2.05 - 1.25^2 = 0.4875$$

Es gilt $n = 20$. Somit gilt

$$s^2 = \frac{20}{19} \cdot 0.4875 = 0.513$$

Und

$$s = \sqrt{0.513} = 0.716$$

□

Bei einem stetigen Merkmal bestimmen wir die Klassenmitten $m_1, m_2, \ldots, m_k$ und berechnen

$$\boxed{\overline{x} = \sum_{i=1}^{k} m_i \, f_i}$$

und

$$\boxed{\overline{x^2} = \sum_{i=1}^{k} m_i^2 \, f_i}$$

Die mittlere quadratische Abweichung ist

$$\boxed{d^2 = \overline{x^2} - \overline{x}^2}$$

Beispiel 29
Die Häufigkeitstabelle des Alters der Teilnehmer aus Beispiel 1 auf Seite 10 ist auf Seite 3.8 zu finden. Die Klassenmitten sind

$$m_1 = \frac{20+25}{2} = 22.5 \qquad m_2 = \frac{25+30}{2} = 27.5$$

$$m_3 = \frac{30+35}{2} = 32.5 \qquad m_4 = \frac{35+40}{2} = 37.5$$

Somit gilt

$$\bar{x} = 22.5 \cdot 0.16 + 27.5 \cdot 0.48 + 32.5 \cdot 0.2 + 37.5 \cdot 0.16 = 29.3$$

und

$$\overline{x^2} = 22.5^2 \cdot 0.16 + 27.5^2 \cdot 0.48 + 32.5^2 \cdot 0.2 + 37.5^2 \cdot 0.16 = 880.25$$

Somit gilt

$$d^2 = \overline{x^2} - \overline{x}^2 = 880.25 - 29.3^2 = 21.76$$

Es gilt $n = 25$. Somit gilt

$$s^2 = \frac{25}{24} \cdot 21.76 = 22.67$$

Und

$$s = \sqrt{22.67} = 4.76$$

□

Schauen wir uns an wie sich die Varianz bei linearen Transformationen der Daten verhält. Transformieren wir die Beobachtungen $x_1, x_2, \ldots, x_n$ zu

$$\boxed{y_i = a + b \cdot x_i}$$

so gilt

$$\boxed{s_y^2 = b^2 \cdot s_x^2} \tag{3.30}$$

Dies wird auf Seite 127 gezeigt.

Beispiel 30
Wir betrachten die Temperatur aus dem Beispiel 4 auf Seite 12. Hier sind die Daten

```
17 18 19 17 16 14 15 12 15 15 15 17 20 21 18 17 17 13 11 10
```

Für die Stichprobenvarianz s_x^2 der Temperatur in Celsius gilt $s_x^2 = 8.24$. Also beträgt die Stichprobenvarianz s_y^2 der Temperatur in Fahrenheit

$$s_y^2 = 1.8^2 \cdot s_x^2 = 1.8^2 \cdot 8.24 = 26.7\,.$$

□

Weitere Maßzahlen für die Streuung

Eine weitere Maßzahl für die Streuung ist die **Spannweite** (**Range**) R, die als Differenz aus der größten und kleinsten Beobachtung definiert ist.

$$\boxed{R = x_{(n)} - x_{(1)}} \tag{3.31}$$

Beispiel 27 (fortgesetzt von Seite 108)
In Gruppe 1 gilt

$$R = 6 - 4 = 2\,.$$

In Gruppe 2 gilt

$$R = 7 - 3 = 4\,.$$

□

Die Spannweite hat den Nachteil, dass sie von den Extremen abhängt. Ist eine Beobachtung ein Ausreißer, so hat dieser einen starken Einfluss auf den Wert der Spannweite. Deshalb betrachtet man in der Regel die Differenz aus dem oberen und dem unteren Quartil. Diese heißt **Interquartilsabstand**(**Inter Quartil Range**) IQR. Es gilt also

$$\boxed{IQR = x_{0.75} - x_{0.25}} \tag{3.32}$$

Beispiel 31
Wir betrachten das Alter der Teilnehmer der Weiterbildungsveranstaltung aus Beispiel 1 auf Seite 10 und wollen den Interquartilsabstand bestimmen. Der geordnete Datensatz ist

```
 23 23 23 24 25 26 26 26 27 27 28 28 28
 28 29 29 30 31 31 32 33 37 37 38 38
```

Es gilt $k = 25 \cdot 0.25 = 6.25$. Also gilt $x_{0.25} = x_{(7)} = 26$. Außerdem gilt $k = 25 \cdot 0.75 = 18.75$. Also gilt $x_{0.75} = x_{(19)} = 31$.
Also gilt

$$IQR = x_{0.75} - x_{0.25} = 31 - 26 = 5$$

□

Maßzahlen für die Variabilität in R

Wir betrachten das Merkmal `Alter` aus Tabelle 1.2 auf Seite 17. Die Daten der Tabelle mögen in der Datentabelle `weiterbildung` stehen.

```
> attach(weiterbildung)
> Alter
 [1] 30 23 26 33 37 28 31 23 24 26 23 32 29 25 31 26 37
     38 29 28 28 28 38 27 27
```

Die Stichprobenvarianz liefert die Funktion `var`. `var`

```
> var(Alter)
[1] 21.32667
```

Für die Spannweite benötigen wir die Funktionen `range` und `diff` . Die Funktion `range` liefert einen Vektor mit dem Minimum und Maximum eines Vektors. `range` `diff`

```
> range(Alter)
[1] 23 38
```

Die Funktion `diff` bildet die Differenz aufeinander folgender Komponenten eines Vektors.

```
> diff(1:4)
[1] 1 1 1
```

Die Spannweite erhalten wir also durch:

```
> diff(range(Alter))
[1] 15
```

und den Interquartilsabstand durch

```
> diff(quantile(Alter,probs=c(0.25,0.75)))
75%
  5
```

Liegt eine Häufigkeitstabelle mit klassierten Beobachtungen vor, so wird die Berechnung der Stichprobenvarianz und des Interquartilsabstandes durch R nicht durch Funktionen unterstützt. Sind die Klassengrenzen und die relativen Häufigkeiten der Klassen gegeben, so kann man diese mit Hilfe einiger Befehle leicht bestimmen.
Wir betrachten die Tabelle 3.8 auf Seite 74. Die Klassengrenzen sind

$$20 \quad 25 \quad 30 \quad 35 \quad 40$$

und die relativen Häufigkeiten der Klassen

$$0.2 \quad 0.48 \quad 0.16 \quad 0.16$$

Den Interquartilsabstand gewinnen wir folgendermaßen mit der Funktion `quantil.approx`, die auf Seite 102 beschrieben wird:

```
> grenzen<-seq(20,40,5)
> hi<-c(0.2,0.48,0.16,0.16)
> diff(quantil.approx(c(0.25,0.75),grenzen,hi))
[1] 6.666667
```

Die Stichprobenvarianz bestimmen wir durch

```
> mi<-(grenzen[-1]+grenzen[-length(grenzen)])/2
> mi
[1] 22.5 27.5 32.5 37.5
> xq<-sum(mi*hi)
> xq
[1] 28.9
> xq2<-sum(mi^2*hi)
> xq2
[1] 858.25
> xq2-xq^2
[1] 23.04
```

3.2.4 Der Boxplot

Der **Boxplot** ist ein effizientes Mittel, um die Verteilung eines quantitativen Merkmals zu visualisieren. Um einen Boxplot zu erstellen, benötigt man die folgenden fünf Zahlen:

das Minimum $x_{(1)}$

das untere Quartil $x_{0.25}$

den Median $x_{0.5}$

das obere Quartil $x_{0.75}$

das Maximum $x_{(n)}$

Man spricht auch von der **Fünf-Zahlen-Zusammenfassung**. Tukey (1977) hat vorgeschlagen, für die Fünf-Zahlen-Zusammenfassung das untere und obere Quartil folgendermaßen zu bestimmen:
Das untere Quartil teilt die untere Hälfte des geordneten Datensatzes in zwei gleich große Hälften, während das obere Quartil die obere Hälfte des geordneten Datensatzes in zwei gleich große Hälften teilt. Somit ist das untere Quartil der Median der unteren Hälfte des geordneten Datensatzes, während das obere Quartil der Median der oberen Hälfte des geordneten Datensatzes ist. Ist der Stichprobenumfang gerade, so ist die untere und obere Hälfte des geordneten Datensatzes eindeutig definiert. Bei einem ungeraden Stichprobenumfang gehört der Median sowohl zur oberen als auch zur unteren Hälfte des geordneten Datensatzes.
Ein Boxplot ist eine graphische Darstellung dieser fünf Zahlen. Wir zeichnen einen Kasten vom unteren bis zum oberen Quartil. In diesem Kasten kennzeichnen wir den Median als Linie. Von den Rändern des Kastens bis zu den Extremen werden Linien gezeichnet, die an sogenannten **Zäunen** enden.

Beispiel 32
Wir schauen uns das Alter der Teilnehmer aus Beispiel 1 auf Seite 10 an. Die Daten sind in Tabelle 1.2 auf Seite 17 zu finden. Schauen wir uns den geordneten Datensatz an.

```
23 23 23 24 25 26 26 26 27 27 28 28 28
28 29 29 30 31 31 32 33 37 37 38 38
```

Die untere Hälfte des geordneten Datensatzes ist

```
23 23 23 24 25 26 26 26 27 27 28 28 28
```

Der Median dieses Datensatzes ist 26. Also ist $x_{0.25} = 26$.
Die obere Hälfte des geordneten Datensatzes ist

```
28 28 29 29 30 31 31 32 33 37 37 38 38
```

Der Median dieses Datensatzes ist 31. Also ist $x_{0.75} = 31$.

Es gilt

$$x_{(1)} = 23 \qquad x_{0.25} = 26 \qquad x_{0.5} = 28 \qquad x_{0.75} = 31 \qquad x_{(n)} = 38$$

Abbildung 3.15 zeigt den Boxplot des Merkmals Alter.

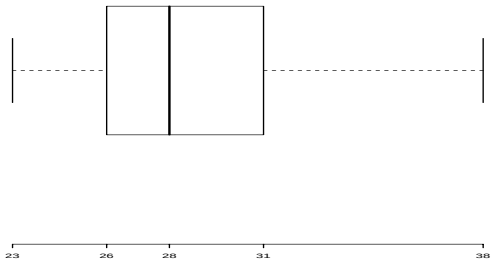


Abbildung 3.15: Boxplot des Merkmals `Alter`

□

Wie interpretiert man einen Boxplot? Durch den Boxplot werden vier Bereiche definiert:

- Minimum bis unteres Quartil
- unteres Quartil bis Median
- Median bis oberes Quartil
- oberes Quartil bis Maximum

Die Interpretation beruht im Wesentlichen darauf, dass jeder der vier Bereiche genau 25 Prozent der Beobachtungen enthält. Als Indikator für die Lage nimmt man den Median und als Indikator für die Streuung die Breite der Box vom unteren Quartil bis zum oberen Quartil. Dies ist gerade der Interquartilsabstand. Die Lage der Medianlinie innerhalb der Box zeigt Symmetrie beziehungsweise Asymmetrie auf. Liegt die Medianlinie ungefähr in der Mitte der Box, so deutet dies auf eine symmetrische Verteilung hin. Liegt sie hingegen näher am unteren als am oberen Quartil, so spricht dies für eine rechtsschiefe Verteilung, während bei einer linksschiefen Verteilung die Medianlinie näher am oberen als am unteren Quartil liegt.

Beispiel 32 (fortgesetzt von Seite 115)
Der Boxplot in Abbildung 3.15 auf Seite 116 deutet auf eine nahezu symmetrische Verteilung hin.

□

Um Ausreißer zu identifizieren, wird der Boxplot modifiziert:
Alle Beobachtungen, die mehr als $1.5 \cdot IQR$ von den Quartilen entfernt sind, werden markiert. Das Minimum aller Beobachtungen, die größer als $x_{0.25} - 1.5 \cdot IQR$ sind, ist der untere Zaun und as Maximum aller Beobachtungen, die kleiner als $x_{0.25} + 1.5 \cdot IQR$ sind, ist der obere Zaun.

Beispiel 33
In einem Projekt zur Betriebinformatik wurden die Studierenden unter anderem nach dem Alter ihres Vaters gefragt. Hier ist die geordnete Stichprobe:

```
44 46 49 50 50 51 51 51 51 51 52 52 53 53 53 54 55 56 57 58 60
```

Es gilt

$$x_{(1)} = 44 \quad x_{0.25} = 51 \quad x_{0.5} = 52 \quad x_{0.75} = 54 \quad x_{(n)} = 60$$

Somit gilt

$$IQR = x_{0.75} - x_{0.25} = 54 - 51 = 3$$

Also gilt

$$x_{0.25} - 1.5 \cdot IQR = 51 - 1.5 \cdot 3 = 46.5$$

und

$$x_{0.75} + 1.5 \cdot IQR = 54 + 1.5 \cdot 3 = 58.5$$

Die Beobachtungen 44 und 46 sind kleiner als 46.5 und die Beobachtung 60ist größer als 58.5. Sie werden markiert. Die kleinste unter den Beobachtungen, die größer als 46.5 sind, ist 49 und größte unter den Beobachtungen, die kleiner als 58.5 sind, ist 58. Also enden die Zäune bei 49 und bei 58. Abbildung 3.16 zeigt den Boxplot.

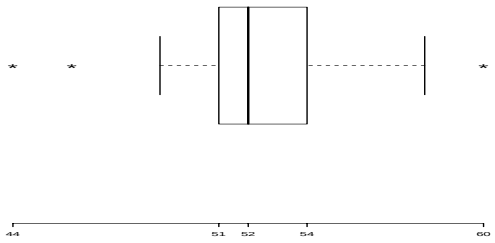


Abbildung 3.16: Boxplot des Alters der Väter

□

Boxplot und Histogramm haben Vor- und Nachteile. In einem Boxplot kann man erkennen, ob eine Verteilung symmetrisch ist, und ob Ausreißer vorliegen. Man kann aber an einem Boxplot nicht erkennen, ob mehrere Gipfel in einem Datensatz vorliegen. Dies erkennt man sofort am Histogramm, in dem man auch Symmetrie und Ausreißer erkennen kann. Das Aussehen des Histogramms hängt aber stark von der Wahl der Klassen ab, die der Benutzer vorgibt. Um die Vorteile beider Darstellungen nutzen zu können, sollte man beide in einer Graphik gemeinsam darstellen. Schauen wir uns zwei Beispiele an.

Abbildung 3.17 zeigt das Histogramm des Merkmals `Mathematische Grundbildung` der PISA-Studie. (siehe Deutsches PISA-Konsortium (Hrsg.) (2001))

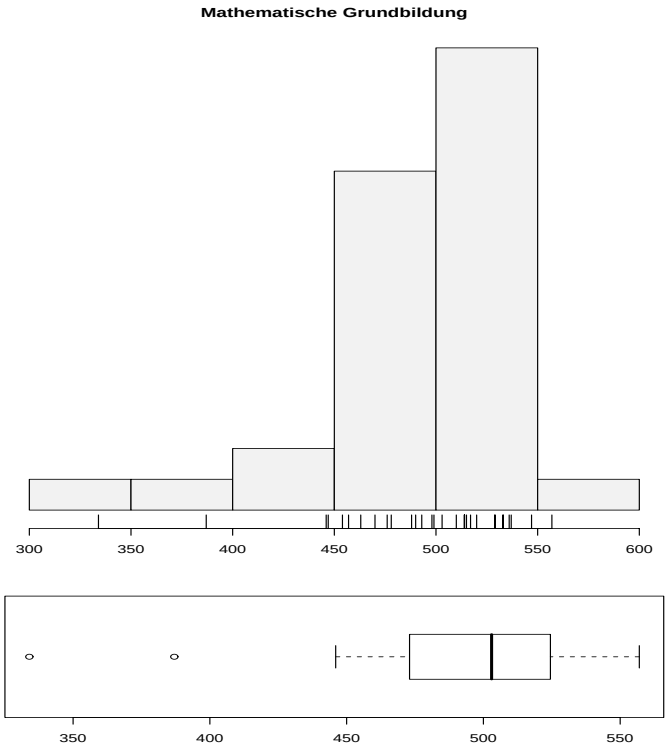


Abbildung 3.17: Histogramm und Boxplot des Merkmals `Mathematische Grundbildung` der PISA-Studie

Das Histogramm deutet auf eine linksschiefe Verteilung hin, während der Boxplot für eine symmetrische Verteilung mit sehr wenig Wahrscheinlichkeitsmasse an den Rändern spricht. Die zwei extremen Ausreißer machen im Histogramm aus dieser symmetrischen Verteilung eine schiefe Verteilung.

Abbildung 3.18 zeigt das Histogramm des Merkmals `Alter` aus Tabelle 1.2 auf Seite 17

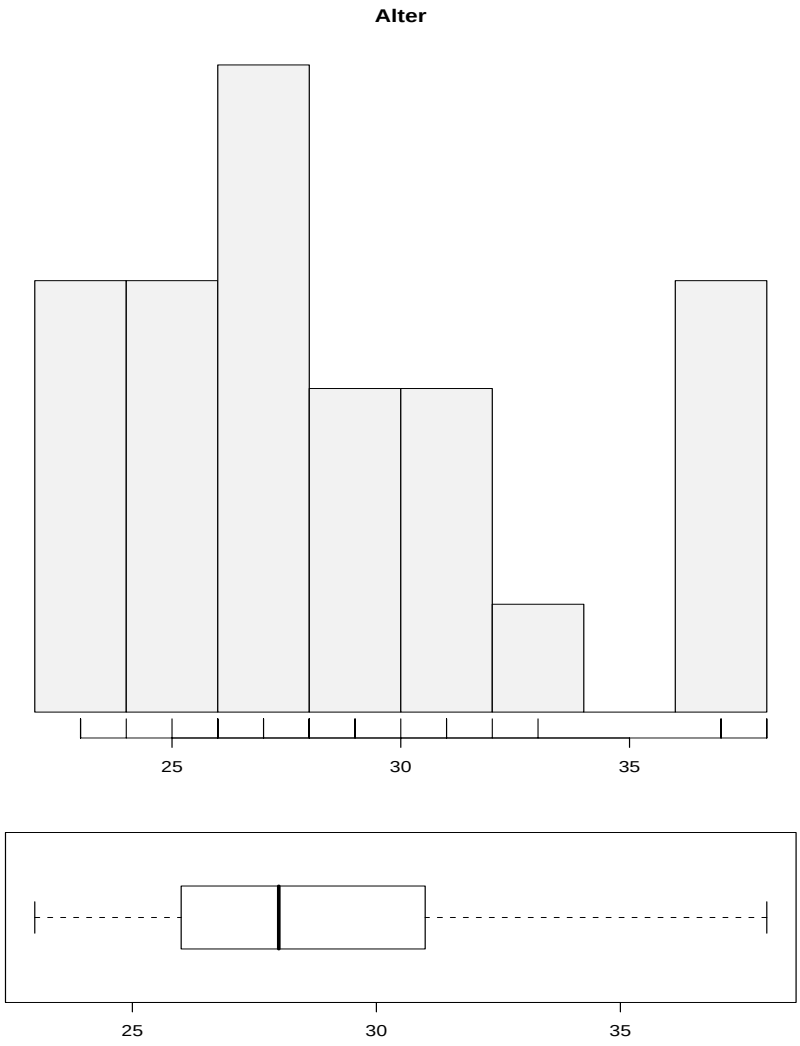


Abbildung 3.18: Histogramm und Boxplot des Merkmals `Alter`

Das Histogramm deutet auf eine zweigipflige Verteilung hin, während der Boxplot keine Zweigipfligkeit zeigen kann.
Die Ausführungen verdeutlichen, dass man sich sowohl das Histogramm als auch den Boxplot anschauen sollte.

Der Boxplot in R

Wir betrachten das Merkmal `Alter` aus Tabelle 1.2 auf Seite 17. Die Daten der Tabelle mögen in der Datentabelle `weiterbildung` stehen.

```
> attach(weiterbildung)
> Alter
 [1] 30 23 26 33 37 28 31 23 24 26 23 32 29 25 31 26 37 38 29
     28 28 28 38 27 27
```

Die Fünf-Zahlen-Zusammenfassung liefert die Funktion `fivenum`.

```
> fivenum(Alter)
 [1] 23 26 28 31 38
```

Die Funktion `summary` liefert neben der Fünf-Zahlen-Zusammenfassung noch den Mittelwert.

```
> summary(Alter)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  23.00   26.00   28.00   29.08   31.00   38.00
```

Einen Boxplot liefert die Funktion `boxplot`. Den Boxplot des Merkmals `Alter` in Abbildung 3.15 auf Seite 116 erhält man durch:

```
> boxplot(Alter,horizontal=TRUE,axes=FALSE)
> axis(1,at=fivenum(Alter))
```

Durch das Argument `horizontal` wird festgelegt, ob der Boxplot horizontal oder vertikal gezeichnet wird. Das Argument `axes` erlaubt es bei Grafiken Achsen zu unterdrücken. Außerdem wird die Box weggelassen, die normalerweise die Grafik einrahmt. Mit dem zweiten Befehl fügen wir die Abszisse hinzu und fordern, dass diese an den fünf Zahlen beschriftet wird.

Einen Boxplot und ein Histogramm in einer Grafik kann man mit der Funktion `simple.hist.and.boxplot` erstellen. Hierzu muss man das Paket `UsingR` installieren und laden. Wie man dabei vorzugehen hat, wird auf Seite 52 beschrieben. Abbildung 3.18 auf der vorherigen Seite erhält man durch folgenden Befehl:

```
> simple.hist.and.boxplot( weiterbildung[,2],main="Alter")
```

3.3 Mathematischer Anhang und Beweise

3.3.1 Das Summenzeichen

Wir betrachten in diesem Skript oft Zahlenfolgen $x_1, x_2, \ldots, x_n$.
So hat ein Schallplattensammler im letzten halben Jahr fünf Langspielplatten bei einem amerikanischen Händler gekauft und dafür folgende Preise in US Dollar bezahlt:

$$x_1 = 22 \qquad x_2 = 30 \qquad x_3 = 16 \qquad x_4 = 25 \qquad x_5 = 27$$

Oft interessieren wir uns für die Summe

$$x_1 + x_2 + \ldots + x_n \tag{3.33}$$

Der Sammler hat insgesamt

$$x_1 + x_2 + x_3 + x_4 + x_5 = 22 + 30 + 16 + 25 + 27 = 120$$

US Dollar gezahlt.
Den Ausdruck in Gleichung (3.33) können wir mit dem Summenzeichen Σ folgendermaßen schreiben

$$\sum_{i=1}^{n} x_i \tag{3.34}$$

Den Ausdruck in Gleichung (3.34) kann man folgendermaßen lesen:
Setze in x_i für i der Reihe nach die Zahlen 1 bis n ein und bilde die Summe dieser Terme.
Der Sammler bildet also

$$\sum_{i=1}^{5} x_i = 120$$

Will man nicht alle Terme aufsummieren, so verändert man den Startwert und Endwert.
Die Ausgaben der ersten vier Käufe erhält der Sammler durch

$$\sum_{i=1}^{4} x_i = 22 + 30 + 16 + 25 = 93$$

Es gibt nun einige Regeln für das Summenzeichen, die einem in Fleisch und Blut übergehen müssen.

Bei jedem Kauf fallen Portokosten in Höhe von 10 US Dollar an. Wie hoch sind die gesamten Portokosten bei fünf Käufen? Offensichtlich sind dies 50 US Dollar.
Wir wollen also n mal die Konstante a aufsummieren. Dies ergibt den Wert $n \cdot a$. Mit dem Summenzeichen stellen wir dies folgendermaßen dar

$$\sum_{i=1}^{n} a = n \cdot a \tag{3.35}$$

Nun schauen wir uns wieder die Preise der Langspielplatten an. Der Sammler will wissen, wie viel er für die Schallplatten in Euro bezahlen muss, wobei der Preis für einen US Dollar 0.80 Euro beträgt.
Es kann den Preis jeder Langspielplatte in Euro umrechnen und erhält

$$\begin{aligned} 0.8 \cdot 22 &= 17.6 \\ 0.8 \cdot 30 &= 24 \\ 0.8 \cdot 16 &= 12.8 \\ 0.8 \cdot 25 &= 20 \\ 0.8 \cdot 27 &= 21.6 \end{aligned}$$

Der Preis aller Langspielplatten in Euro ist

$$17.6 + 24 + 12.8 + 20 + 21.6 = 96$$

Diesen Wert erhält man aber auch, wenn den Preis aller Langspielplatten, der in US Dollar 120 beträgt, mit 0.8 multipliziert. Dies ist kein Zufall.
Ist b eine reelle Zahl, so gilt

$$\sum_{i=1}^{n} b \cdot x_i = b \cdot \sum_{i=1}^{n} x_i \tag{3.36}$$

Der Beweis (3.36) ist ganz einfach:

$$\begin{aligned} \sum_{i=1}^{n} b \cdot x_i &= b \cdot x_1 + b \cdot x_2 + \ldots + b \cdot x_n \\ &= b \cdot (x_1 + x_2 + \ldots + x_n) \\ &= b \cdot \sum_{i=1}^{n} x_i \end{aligned}$$

Oft betrachtet man zwei Folgen $x_1, x_2, \ldots, x_n$ und $y_1, y_2, \ldots, y_n$ der Länge n. So schreibt ein Ehepaar die täglichen Ausgaben auf. Dabei seien x_i die Ausgaben der Frau und y_i die Ausgaben des Mannes am i-ten Tag. Für drei Tage gilt

$$x_1 = 7.6 \qquad x_2 = 9.4 \qquad x_3 = 9$$

und

$$y_1 = 7.7 \qquad y_3 = 6.6 \qquad y_4 = 8.7$$

Die Ausgaben der Frau sind

$$\sum_{i=1}^{5} x_i = 7.6 + 9.4 + 9 = 26$$

und die Ausgaben des Mannes

$$\sum_{i=1}^{5} y_i = 7.7 + 6.6 + 8.7 = 23$$

Wir können die Summe des Ausgaben von beiden auf zwei Arten bestimmen. Wir können zum einen die Summe der Ausgaben für jeden tag bestimmen und diese Werte aufsummieren

$$(7.6 + 7.7) + (9.4 + 6.6) + (9 + 8.7) = 15.3 + 16 + 17.7 = 49$$

Wir können aber auch erst die Summe der Ausgaben der Frau und dann die Summe der Ausgaben des Mannes bestimmen und dann die Summe dieser beiden Ausgabensummen bestimmen:

$$(7.6 + 9.4 + 9) + (7.7 + 6.6 + 8.7) = 26 + 23 = 49$$

In beiden Fällen erhalten wir das gleiche Ergebnis.
Allgemein gilt

$$\sum_{i=1}^{n}(x_i + y_i) \;=\; \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i \tag{3.37}$$

Dies sieht man folgendermaßen:

$$\begin{aligned}
\sum_{i=1}^{n}(x_i + y_i) &= x_1 + y_1 + x_2 + y_2 + \ldots + x_n + y_n \\
&= x_1 + x_2 + \ldots + x_n + y_1 + y_2 + \ldots + y_n \\
&= \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i
\end{aligned}$$

Die Beziehung in Gleichung (3.37) gilt für jede endliche Anzahl von Summanden.
Schauen wir uns noch ein Beispiel an, bei dem man die Regeln (3.35), (3.36) und (3.37) benötigt.

$$\sum_{i=1}^{n}(a + b \cdot x_i) \overset{(3.37)}{=} \sum_{i=1}^{n} a + \sum_{i=1}^{n} b \cdot x_i \overset{\substack{(3.35)\\(3.36)}}{=} n \cdot a + b \cdot \sum_{i=1}^{n} x_i$$

3.3.2 Wie bestimmt man eine Gerade aus zwei Punkten?

Auf Seite 79 müssen wir mehrere Geradengleichungen bestimmen. Schauen wir uns dies etwas genauer an.
Die Menge aller Punkte $(x, y) \in \mathbb{R}$ mit

$$y = a + b \cdot x \tag{3.38}$$

beschreibt eine Gerade.
Für$x = 0$ gilt $y = a$. Man nennt a auch den **Achsenabschnitt**. Der Parameter b ist gleich der **Steigung** der Geraden. Diese gibt an, wie sich y ändert, wenn sich x um 1 erhöht. Abbildung 3.19 veranschaulicht dies.

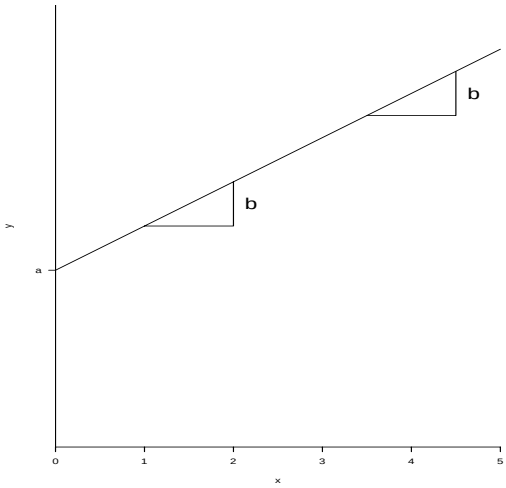


Abbildung 3.19: Eine Gerade

Um die Geradengleichung bestimmen zu können, benötigt man zwei Punkte (x_1, y_1) und (x_2, y_2). Wir werden gleich sehen, welche Bedingung diese

Punkte erfüllen müssen. Wir setzen diese Punkte (x_1, y_1) und (x_2, y_2) in die Gleichung (3.38) ein und erhalten zwei Gleichungen, die linear in den Unbekannten a und b sind:

$$y_1 = a + b \cdot x_1 \tag{3.39}$$

$$y_2 = a + b \cdot x_2 \tag{3.40}$$

Es gibt eine Vielzahl von Möglichkeiten, diese Gleichungen zu lösen. Wir subtrahieren Gleichung (3.39) von Gleichung (3.40):

$$y_2 - y_1 = a + b \cdot x_2 - a - b \cdot x_1 \tag{3.41}$$

Gleichung (3.41) ist äquivalent zu

$$y_2 - y_1 = b \cdot x_2 - b \cdot x_1 \tag{3.42}$$

Gleichung (3.42) ist äquivalent zu

$$y_2 - y_1 = b \cdot (x_2 - x_1) \tag{3.43}$$

Ist $x_1 \neq x_2$, so können wir beide Seiten durch $x_2 - x_1$ dividieren. Somit gilt

$$b = \frac{y_2 - y_1}{x_2 - x_1} \tag{3.44}$$

Setzen wir die rechte Seite von Gleichung (3.44) für b in Gleichung (3.39) ein, so erhalten wir

$$a = y_1 - \frac{y_2 - y_1}{x_2 - x_1} \cdot x_1 \tag{3.45}$$

Schauen wir uns exemplarisch die Gleichung der approximierenden empirischen Verteilungsfunktion in der zweiten Klasse in Gleichung (3.6) auf Seite 80 an. An der Untergrenze 25 der Klasse gilt $F_n^*(25) = 0.2$ und an der Obergrenze 30 der Klasse gilt $F_n^*(30) = 0.68$. Die Gerade läuft also durch die Punkte $(25, 0.2)$ und $(30, 0.68)$. Also gilt

$$b = \frac{y_2 - y_1}{x_2 - x_1} = \frac{0.68 - 0.2}{30 - 25} = 0.096$$

und

$$a = y_1 - \frac{y_2 - y_1}{x_2 - x_1} \cdot x_1 = 0.2 - 0.096 \cdot 25 = -2.2$$

3.3.3 Beweise

Die Gültigkeit der Beziehung (3.14) auf Seite 95 sieht man folgendermaßen:

$$\begin{aligned}
\overline{y} &= \frac{1}{n}\cdot\sum_{i=1}^{n} y_i = \frac{1}{n}\cdot\sum_{i=1}^{n}(a+b\cdot x_i) \overset{(3.37)}{=} \frac{1}{n}\cdot\left(\sum_{i=1}^{n} a + \sum_{i=1}^{n} b\cdot x_i\right) \\
&\overset{\substack{(3.35)\\(3.36)}}{=} \frac{1}{n}\cdot\left(n\cdot a + b\cdot\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\cdot n\cdot a + b\cdot\frac{1}{n}\cdot\sum_{i=1}^{n} x_i \\
&= a + b\cdot\overline{x}
\end{aligned}$$

Die Gültigkeit der Beziehung (3.16) auf Seite 96 sieht man folgendermaßen:

$$\begin{aligned}
\sum_{i=1}^{n}(x_i-\overline{x}) &\overset{(3.37)}{=} \sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\overline{x} \overset{(3.35)}{=} \sum_{i=1}^{n} x_i - n\,\overline{x} \\
&\overset{(3.11)}{=} \sum_{i=1}^{n} x_i - n\,\frac{1}{n}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0
\end{aligned}$$

Die Gültigkeit von Beziehung (3.25) auf Seite 107 sieht man folgendermaßen:

$$\begin{aligned}
d^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2\,y_i\,\overline{x} + \overline{x}^2) \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\sum_{i=1}^{n} 2\,x_i\,\overline{x} + \frac{1}{n}\sum_{i=1}^{n}\overline{x}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\,\overline{x}\,\frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\,n\,\overline{x}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\,\overline{x}\,\overline{x} + \overline{x}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\,\overline{x}^2 + \overline{x}^2 \\
&= \overline{x^2} - \overline{x}^2
\end{aligned}$$

Der Mittelwert eines standardisierten Merkmals ist gleich 0. Dies sieht man folgendermaßen:

$$\overline{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i-\bar{x}}{s} = \frac{1}{n\,s}\sum_{i=1}^{n}(x_i-\bar{x}) = 0\,.$$

Die Stichprobenvarianz der standardisierten Merkmale ist gleich 1. Dies sieht man folgendermaßen:

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^2=\frac{1}{s^2}\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{s^2}\,s^2=1\,.$$

Die Stichprobenvarianz von $y_i=1+b\cdot x_i$ erhält man folgendermaßen:

$$\begin{aligned}
s_y^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2=\frac{1}{n-1}\sum_{i=1}^{n}(a+b\cdot x_i-a-b\cdot\overline{x})^2\\
&= \frac{1}{n-1}\sum_{i=1}^{n}(b\cdot(x_i-\overline{x}))^2=\frac{1}{n-1}\sum_{i=1}^{n}b^2\cdot(x_i-\overline{x})^2\\
&= b^2\cdot\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2=b^2\cdot s_x^2
\end{aligned}$$

3.4 Datensätze

Tabelle 3.17: Die Häufigkeitstabelle des Merkmals Anzahl CDs

Anzahl CDs	absolute Häufigkeit
von 0 bis 50	132
von 50 bis 100	66
von 100 bis 150	27
von 150 bis 200	15
von 200 bis 250	10
von 250 bis 300	6

Tabelle 3.18: Die Häufigkeitstabelle des Merkmals `Mathematische Grundbildung`

Mathematische Grundbildung	absolute Häufigkeit
von 300 bis 350	1
von 350 bis 400	1
von 400 bis 450	2
von 450 bis 500	11
von 500 bis 550	15
von 550 bis 600	1

(Quelle: Deutsches PISA-Konsortium (Hrsg.): PISA 2000. Leske + Budrich, 2001.)

Tabelle 3.19: Die Häufigkeitstabelle der Körpergröße der Männer

Körpergröße	absolute Häufigkeit
von 165 bis 170	5
von 170 bis 175	12
von 175 bis 180	32
von 180 bis 185	65
von 185 bis 190	35
von 190 bis 195	25
von 195 bis 200	12

Tabelle 3.20: Die Häufigkeitstabelle des Merkmals Alter der Teilnehmer einer Weiterbildungsveranstaltung

Alter	absolute Häufigkeit
von 20 bis 24	3
von 24 bis 28	7
von 28 bis 32	9
von 32 bis 36	2
von 36 bis 40	4

Kapitel 4

Multivariate Analyse

In Kapitel 3 haben wir uns jeweils nur ein Merkmal angeschaut und dessen Verteilung dargestellt. Mit Hilfe statistischer Verfahren kann man aber auch untersuchen, ob zwischen mehreren Merkmalen Abhängigkeiten bestehen. Wir wollen hier nur zwei Merkmale betrachten. Ist mindestens eines der beiden Merkmale qualitativ, so kann man auf Basis dieses Merkmals Gruppen bilden, wobei alle Merkmalsträger mit der gleichen Merkmalsausprägung eine Gruppe bilden. Ist zum Beispiel das qualitative Merkmal das Geschlecht, so enthält die eine Gruppe die Frauen und die andere die Männer. Man betrachtet dann die Verteilung des anderen Merkmals in den Gruppen. Beide Merkmale können aber auch quantitativ sein sein. In diesem Fall soll die Abhängigkeitsstruktur zwischen den beiden Merkmalen durch eine geeignete Maßzahl beschrieben werden.

4.1 Zusammenhang zwischen einem qualitativen und einem quantitativen Merkmal

Wir schauen uns die Verteilung eines quantitativen Merkmals in c Gruppen an. Die Gruppen werden durch ein qualitatives Merkmal wie das Merkmal `Geschlecht` gebildet. Wir bezeichnen die Merkmalsausprägung des j-ten Merkmalsträgers in der i-ten Gruppe mit y_{ij}. Dabei ist $i = 1, 2, \ldots, c$ und $j = 1, 2, \ldots, n_i$. Die Anzahl der Beobachtungen in den Gruppen muss also nicht identisch sein. Insgesamt werden die Werte von $N = n_1 + n_2 + \ldots + n_c$ Merkmalsträgern erhoben.

Wir können die Verteilung des Merkmals in den c Gruppen mit Maßzahlen oder grafischen Darstellungen vergleichen.

Beginnen wir mit den Maßzahlen. Die Lage beschreiben wir durch den Mit-

telwert

$$\boxed{\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}} \tag{4.1}$$

und den Median, während wir die Stichprobenvarianz

$$\boxed{s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2} \tag{4.2}$$

benutzen, um die Verteilungen hinsichtlich der Streuung zu vergleichen. Außerdem bestimmen wir für jede der c Gruppen die Fünf-Zahlen-Zusammenfassung, die auf Seite 115 beschrieben wird.
Den besten Überblick über die Unterschiede zwischen den Verteilungen erhält man, wenn man die Boxplots der Gruppen nebeneinander zeichnet.

Beispiel 34
Wir schauen uns das Beispiel 1 auf Seite 10 an und betrachten das nominalskalierte Merkmal `Geschlecht` und das quantitative Merkmal `Alter`. Die Frauen mögen die erste Gruppe bilden. Es gilt $n_1 = 13$ und $n_2 = 12$. Hier ist die Urliste des Alters der Frauen:

```
23 26 31 24 23 25 31 29 28 28 28 38 27
```

Hier ist die Urliste des Alters der Männer:

```
30 33 37 28 23 26 32 29 26 37 38 27
```

Tabelle 4.1 zeigt die Maßzahlen des Alters der Männer und Frauen

Tabelle 4.1: Maßzahlen für die Lage und Streuung des Alters der Männer und Frauen

	Mittelwert	Median	Varianz
Frauen	27.8	28.0	16.5
Männer	30.5	29.5	24.3

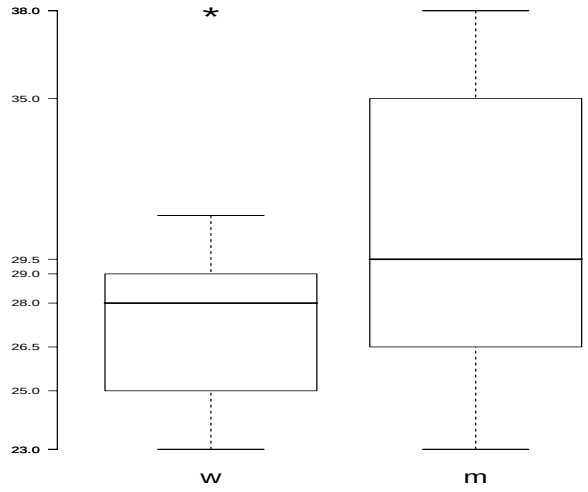
Wir sehen, dass die Männer im Mittel fast drei Jahre älter als die Frauen sind. Außerdem ist die Streuung des Alters der Männern höher als bei den Frauen.
Tabelle 4.2 zeigt die Fünf-Zahlen-Zusammenfassung bei den Frauen und den Männern.

Tabelle 4.2: Fünf-Zahlen-Zusammenfassung des Alters der Männern und Frauen

	Minimum	unteres Quartil	Median	oberes Quartil	Maximum
Frauen	23.0	25.0	28.0	29.0	38.0
Männer	23.0	26.5	29.5	35.0	38.0

Abbildung 4.1 zeigt die Boxplots des Merkmals `Alter` der Frauen und Männern. Die beiden Verteilungen unterscheiden sich bezüglich der Lage. Die Streuung des Alters der Männer ist viel größer als die Streuung des Alters der Frauen. Die Verteilung des Alters der Frauen ist linksschief, die des Alters der Männer rechtsschief. Außerdem gibt es bei den Frauen einen Ausreißer.

Abbildung 4.1: Boxplots des Merkmals `Alter` bei den Frauen und den Männern



□

Im Beispiel gilt $\bar{y}_1 = 27.77$ und $\bar{y}_2 = 30.5$. Will man aus diesen beiden Werten das Durchschnittsalter aller Personen bestimmen, so darf man nicht den Mittelwert

$$\frac{27.77 + 30.5}{2} = 29.135$$

der beiden Mittelwerte bestimmen, sondern man muss die beiden Mittelwerte mit den Anteilen der Stichprobenumfänge gewichten:

$$\bar{y} = \frac{13}{25} \cdot 27.77 + \frac{12}{25} \cdot 30.5 = 29.08$$

Liegen also c Mittelwerte $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_c$ aus Gruppen vor, die $n_1, n_2, \ldots, n_c$ Merkmalsträger enthalten, dann ist der Mittelwert aller $N = n_1 + \ldots n_c$ Beobachtungen gleich:

$$\boxed{\bar{y} = \sum_{i=1}^{c} \frac{n_i}{N} \cdot \bar{y}_i} \tag{4.3}$$

Dies sieht man folgendermaßen

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{c} \sum_{j=1}^{n_i} y_{ij} \overset{(4.1)}{=} \frac{1}{N} \sum_{i=1}^{c} n_i \cdot \bar{y}_i = \sum_{i=1}^{c} \frac{n_i}{N} \cdot \bar{y}_i$$

Liegen die Daten in Form einer Häufigkeitstabelle mit klassierten Daten vor, so sollte man keinen Boxplot zeichnen, da man für das Minimum nur die Untergrenze der ersten Klasse und für das Maximum nur die Obergrenze der letzten Klasse wählen kann. Diese Klassengrenzen werden aber von der Person, die die Daten auswertet, festgelegt und fallen in der Regel nicht mit den Extremwerten der Daten zusammen. Sind Randklassen offen, so gibt es gar keine Information über die Extremwerte der Daten. In beiden Fällen sollte man eine Modifikation des Boxplots verwenden, bei der nur die Box gezeichnet wird, sodass nur die Zentren der Verteilungen verglichen werden.

Beispiel 35
Im Statistischen Jahrbuch 2004 ist auf Seite 46 eine Häufigkeitstabelle mit dem monatlichen Nettohaushaltseinkommen der Privathaushalte in Deutschland im Jahr 2003 zu finden. Dabei wurden die alten und neuen Bundesländer unterschieden. Die Daten sind in Tabelle 4.3 wiedergegeben.

Tabelle 4.3: Monatliches Nettohaushaltseinkommen der Privathaushalte in Deutschland im Jahr 2003 unterschieden nach alten und neuen Bundesländern

Klasse i	$\left(x_{i-1}^*, x_i^*\right]$	alte	neue
1	$(0, 500]$	984	306
2	$(500, 900]$	3295	1066
3	$(900, 1300]$	4853	1462
4	$(1300, 1500]$	2598	693
5	$(1500, 2000]$	5151	1370
6	$(2000, 2600]$	4979	1059
7	$(2600, 4500]$	6369	942
8	$(4500, \infty)$	1885	157

Um die Quartile und den Median zu bestimmen, erstellen wir Tabellen mit den relativen und kumulierten relativen Häufigkeiten. Beginnen wir mit den alten Bundesländern.

Tabelle 4.4: Monatliches Nettohaushaltseinkommen der Privathaushalte in Deutschland im Jahr 2003 in den alten Bundesländern

Klasse i	$\left(x_{i-1}^*, x_i^*\right]$	relative	kumulierte relative
1	$(0, 500]$	0.033	0.033
2	$(500, 900]$	0.109	0.142
3	$(900, 1300]$	0.161	0.303
4	$(1300, 1500]$	0.086	0.389
5	$(1500, 2000]$	0.171	0.560
6	$(2000, 2600]$	0.165	0.725
7	$(2600, 4500]$	0.211	0.936
8	$(4500, \infty)$	0.063	1.000

Also gilt

$$\begin{aligned} x_{0.25} &= 900 + \frac{0.25 - 0.142}{0.161} \cdot 400 = 1168.32 \\ x_{0.5} &= 1500 + \frac{0.5 - 0.389}{0.171} \cdot 500 = 1824.56 \\ x_{0.75} &= 2600 + \frac{0.75 - 0.726}{0.211} \cdot 1900 = 2825.12 \end{aligned}$$

In den neuen Bundesländern erhalten wir folgende Tabelle:

Tabelle 4.5: Monatliches Nettohaushaltseinkommen der Privathaushalte in Deutschland im Jahr 2003 unterschieden in den neuen Bundesländern

Klasse i	$\left(x_{i-1}^*, x_i^*\right]$	relative	kumulierte relative
1	$(0, 500]$	0.043	0.043
2	$(500, 900]$	0.151	0.194
3	$(900, 1300]$	0.207	0.401
4	$(1300, 1500]$	0.098	0.499
5	$(1500, 2000]$	0.194	0.694
6	$(2000, 2600]$	0.150	0.844
7	$(2600, 4500]$	0.134	0.978
8	$(4500, \infty)$	0.022	1.000

Also gilt

$$
\begin{aligned}
x_{0.25} &= 900 + \frac{0.25 - 0.194}{0.207} \cdot 400 = 1008.21 \\
x_{0.5} &= 1500 + \frac{0.5 - 0.499}{0.194} \cdot 500 = 1502.58 \\
x_{0.75} &= 2000 + \frac{0.75 - 0.693}{0.15} \cdot 600 = 2228
\end{aligned}
$$

Abbildung 4.2 zeigt die Zentren der Boxplots in den alten und neuen Bundesländern. Wir sehen, dass das Durchschnittseinkommen gemessen mit dem Median in den alten Bundesländern höher als in den neuen Bundesländern ist. Außerdem streut das Haushaltseinkommen in den alten Bundesländern viel stärker als in den neuen. In den alten und neuen Bundesländern ist die Verteilung des Haushaltseinkommens rechtsschief. Dies ist bei allen Einkommensverteilungen der Fall.

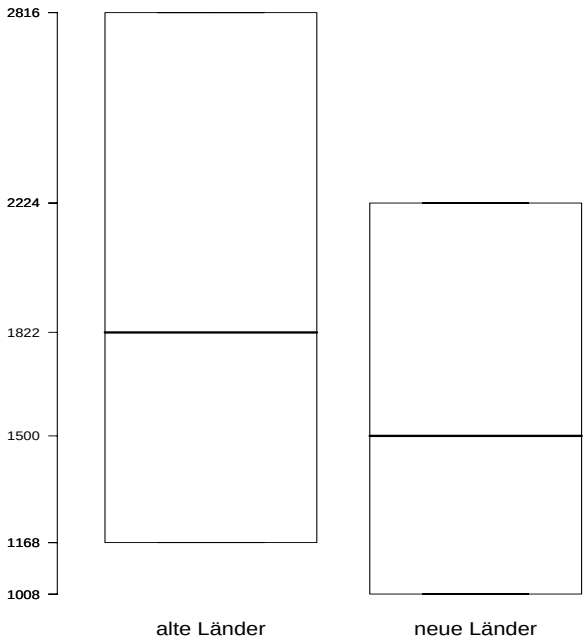


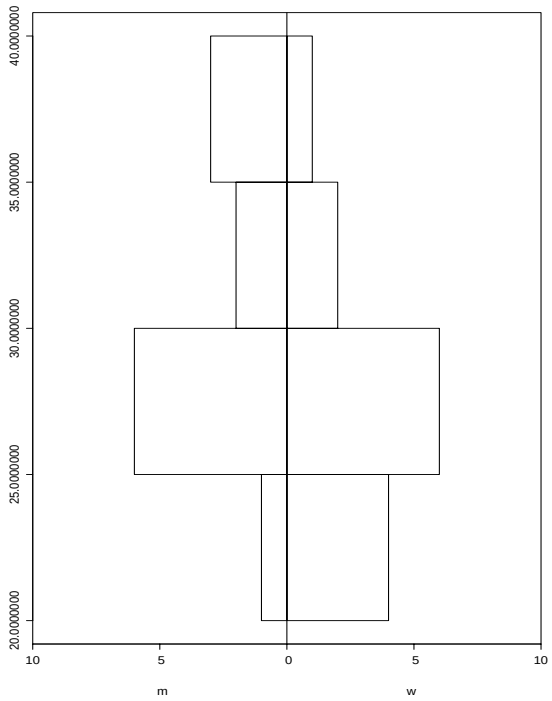
Abbildung 4.2: Boxplot des Zentrums der Einkommensverteilung der Haushalte in den alten und neuen Bundesländern im Jahr 2003

□

Liegen nur zwei Gruppen vor, so kann man die beiden Histogramme der Gruppen gegenüberstellen. Schraffiert man noch die Bereiche, in denen das jeweilige Geschlecht häufiger auftritt, so ist der Vergleich perfekt.

Beispiel 34 (fortgesetzt von Seite 132)
Wir betrachten noch einmal das Alter der Frauen und Männer bei der Fortbildungsveranstaltung. Die Daten sind in Tabelle 1.2 auf Seite 17 zu finden. Abbildung 4.3 zeigt die Histogramme des Alters der Frauen und der Männer mit Schraffur.

Abbildung 4.3: Histogramme des Alters der Frauen und der Männer



Wir sehen, dass in den unteren Altersklassen ein Frauenüberschuss herrscht, während in den oberen Altersklassen die Männer häufiger auftreten.

□

In der Bevölkerungsstatistik heißt diese Darstellung eine **Bevölkerungspyramide**. Hier wird für jedes Alter die Anzahl der Frauen und Männer durch Säulen visualisiert.

4.1.1 Die Analyse mit R

Wir betrachten das Merkmal `Alter` aus Tabelle 1.2 auf Seite 17. Die Daten der Tabelle mögen in der Datentabelle `weiterbildung` stehen, die auf Seite 30 zu finden ist. Wir greifen auf die Variablen zu, wie auf Seite 30 beschrieben wird.

```
> attach(weiterbildung)
> Alter
```

```
 [1] 30 23 26 33 37 28 31 23 24 26 23 32
     29 25 31 26 37 38 29 28 28 28 38 27 27
> Geschlecht
 [1] m w w m m m w m w m w m m w w m m m w w w w w w m
Levels: m w
```

Im Kapitel 2.4 auf Seite 33 haben wir eine Vielzahl von Möglichkeiten kennen gelernt, aus einem Vektor die Komponenten auszuwählen, die eine oder mehrere Bedingungen erfüllen. Mit

```
> alter.w<-Alter[Geschlecht=="w"]
> alter.w
 [1] 23 26 31 24 23 25 31 29 28 28 28 38 27
> alter.m<-Alter[Geschlecht=="m"]
> alter.m
 [1] 30 33 37 28 23 26 32 29 26 37 38 27
```

generieren wir zwei Vektoren `alter.w` und `alter.m`, die das Alter der Frauen und das Alter der Männer enthalten. Auf diese können wir Funktionen wie `mean`, `median` und `var` anwenden, um die jeweiligen Maßzahlen zu bestimmen.

```
> mean(alter.w)
[1] 27.76923
> mean(alter.m)
[1] 30.5
> median(alter.w)
[1] 28
> median(alter.m)
[1] 29.5
> var(alter.w)
[1] 16.52564
> var(alter.m)
[1] 24.27273
```

Durch Eingabe von

```
> boxplot(alter.w,alter.m,names=c("w","m"),outpch="*",
          horizontal=T,cex.axis=2,cex=3)
```

werden die Boxplots der beiden Gruppen nebeneinander gezeichnet.

Wesentlich schneller und eleganter gelangt man zum Ziel, wenn man die auf den Seiten 35 und 36 beschriebenen Funktionen `split`, `lapply` und `sapply` verwendet.

`split`
`lapply`
`sapply`

```
> sapply(split(Alter,Geschlecht),var)
       m        w
24.27273 16.52564
> lapply(split(Alter,Geschlecht),summary)
$m
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  23.00   26.75   29.50   30.50   34.00   38.00

$w
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  23.00   25.00   28.00   27.77   29.00   38.00

> sapply(split(Alter,Geschlecht),summary)
             m     w
 Min.    23.00 23.00
 1st Qu. 26.75 25.00
 Median  29.50 28.00
 Mean    30.50 27.77
 3rd Qu. 34.00 29.00
 Max.    38.00 38.00
```

Den Boxplot erhält man durch

```
> boxplot(split(Alter,Geschlecht),names=c("w","m"),outpch="*",
          horizontal=T,cex.axis=2,cex=3)
```

Wie man den auf Seite 134 beschriebenen modifizierten Boxplot erstellt, illustrieren wir an Hand der Daten in Tabelle 4.3 auf Seite 135. Wir geben also zuerst die Klassengrenzen ein, die in beiden Gruppen identisch sind.

```
> gr<-c(0,5,9,13,15,20,26,45,Inf)
> gr<-100*gr
> gr
[1]    0  500  900 1300 1500 2000 2600 4500  Inf
```

Dabei ist `Inf` gleich ∞.
Dann geben wir die absoluten Häufigkeiten ein

```
> ni.a<-c(984,3295,4853,2598,5151,4979,6369,1885)
> ni.n<-c(306,1066,1462,693,1370,1059,942,157)
```

und bestimmen die relativen Häufigkeiten

```
> hi.a<-ni.a/sum(ni.a)
> hi.n<-ni.n/sum(ni.n)
```

Wir bestimmen die Quartile und den Median mit der Funktion `quantil.approx`, die auf der Seite 102 zu finden ist.

```
> q.a<-quantil.approx(c(0.25,0.5,0.75),gr,hi.a)
> q.a
[1] 1167.834 1822.947 2816.431
> q.n<-quantil.approx(c(0.25,0.5,0.75),gr,hi.n)
> q.n
[1] 1007.182 1500.182 2223.371
```

Bei einem geordneten Datensatz der Länge 5 sind die 5 Beobachtungen gleich der Fünf-Zahlen-Zusammenfassung. Um den modifizierten Boplot zu erstellen, übergeben wir der Funktion `boxplot` also einen Vektor, der zweimal die Quartile und einmal den Median enthält. Außerdem wählen wir eine aussagekräftige Beschriftung der beiden Gruppen.

```
> boxplot(c(q.a[1],q.a,q.a[3]),c(q.n[1],q.n,q.n[3]),
          names=c("alte BL","neue BL"))
```

Die gegenübergestellten Histogramme in Abbildung 4.3 auf Seite 138 kann man mit der Funktion `histbackback` aus dem Paket `Hmisc` erstellen. Dieses Paket muss man zunächst installieren und laden. Wie man dabei vorzugehen hat, wird auf Seite 52 beschrieben.
Folgende Befehlsfolge erstellt die Abbildung 4.3 auf Seite 138:

```
> attach(weiterbildung)
> histbackback(split(Alter,Geschlecht),brks=seq(20,40,5),
               xlim=c(-10,10)
```

Die Funktion `split` wird auf Seite 35 ausführlich diskutiert. Mit dem Argument `brks` legt man die Klassengrenzen fest.

4.2 Zusammenhang zwischen zwei nominalskalierten Merkmalen

Wir wollen nun zwei nominalskalierte Merkmale A und B betrachten. Dabei bezeichnen wir die Merkmalsausprägungen von A mit $A_1, A_2, \ldots, A_I$ und die Merkmalsausprägungen von B mit $B_1, B_2, \ldots, B_J$.

Beispiel 36
Wir schauen uns noch einmal das Beispiel 1 auf Seite 10 an und betrachten die nominalskalierten Merkmale `Geschlecht` und `Titanic`. Die Ausprägungsmöglichkeiten des Geschlechts sind `w` und `m` und die des Merkmals `Titanic` `j` und `n`. Dabei nimmt das Merkmals `Titanic` die Merkmalsausprägung `j` an, wenn die Person den Film `Titanic` gesehen hat.
Es gibt insgesamt vier mögliche Ausprägungen (A_i, B_j), $i = 1, 2$, $j = 1, 2$, wenn wir beide Merkmale gemeinsam betrachten:

Die Person ist weiblich und hat den Film gesehen:	(`w`,`j`).
Die Person ist weiblich und hat den Film nicht gesehen:	(`w`,`n`).
Die Person ist männlich und hat den Film gesehen:	(`m`,`j`).
Die Person ist männlich und hat den Film nicht gesehen:	(`m`,`n`).

□

Wie im univariaten Fall bestimmen wir absolute Häufigkeiten, wobei wir aber die beiden Merkmale gemeinsam betrachten. Wir bezeichnen die Anzahl der Merkmalsträger, die beim Merkmal A die Ausprägung A_i und beim Merkmal B die Ausprägung B_j aufweisen, mit n_{ij}.

Beispiel 36 (fortgesetzt)
Die Daten stehen in Tabelle 1.2 auf Seite 17 in der zweiten und vierten Spalte. Es gilt

$$n_{11} = 12 \quad n_{12} = 1$$

$$n_{21} = 5 \quad n_{22} = 7$$

□

4.2.1 Die Kontingenztabelle

Wir stellen absoluten Häufigkeiten in einer **Kontingenztabelle** tabellarisch dar Eine Kontingenztabelle ist nichts anderes als eine Häufigkeitstabelle mehrerer nominalskalierter Merkmale. Tabelle 4.6 zeigt den allgemeinen Aufbau

einer zweidimensionalen Kontingenztabelle. Die Tabelle besteht aus einem Zentrum und einem Rand.

Tabelle 4.6: Allgemeiner Aufbau einer zweidimensionalen Kontingenztabelle

A \ B	B_1	B_2	$\dots$	B_J	
A_1	n_{11}	n_{12}	$\cdots$	n_{1J}	$n_{1.}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2J}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_I	n_{I1}	n_{I2}	$\cdots$	n_{IJ}	$n_{I.}$
	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.J}$	n

Im Zentrum der Tabelle stehen die absoluten Häufigkeiten n_{ij} der Merkmalsausprägungen (A_i, B_j). Am Rand der i-ten Zeile steht die Summe $n_{i.}$ der absoluten Häufigkeiten der i-ten Zeile:

$$n_{i.} = \sum_{j=1}^{J} n_{ij}$$

Dies ist die Anzahl der Merkmalsträger, die die i-te Kategorie des Merkmals A aufweisen.

Am Rand der j-ten Spalte steht die Summe $n_{.j}$ der absoluten Häufigkeiten der j-ten Spalte:

$$n_{.j} = \sum_{i=1}^{I} n_{ij}$$

Dies ist die Anzahl der Merkmalsträger, die die j-te Kategorie des Merkmals B aufweisen.

Man spricht auch von den **Randhäufigkeiten**.

Beim Aufstellen einer zweidimensionalen Kontingenztabelle müssen wir entscheiden, welches der beiden Merkmale wir mit A und welches mit B bezeichnen. Ehrenberg (1981) weist darauf hin, dass Zahlen, die miteinander verglichen werden sollen, übereinander stehen sollten, da man zahlen, die addiert werden sollen, übereinander schreibt. Wir vergleichen deshalb automatisch die Zaheln innerhalb der Spalte einer Tabelle.

Beispiel 36 (fortgesetzt von Seite 142)
Es gilt

$$n_{1\cdot} = 13 \qquad n_{2\cdot} = 12 \qquad n_{\cdot 1} = 17 \qquad n_{\cdot 2} = 8 \qquad n = 25$$

Die Kontingenztabelle ist in Tabelle 4.7 zu finden.

Tabelle 4.7: Kontingenztabelle der Merkmale `Geschlecht` und `Titanic`

Geschlecht \ Titanic	j	n	
w	12	1	13
m	5	7	12
	17	8	25

□

4.2.2 Bedingte relative Häufigkeiten

Durch die Merkmalsausprägungen $A_1, A_2, \ldots, A_I$ des Merkmals A werden I Gruppen gebildet. So gehören zur i-ten Gruppe alle Merkmalsträger, die die Merkmalsausprägung A_i besitzen. In jeder dieser Gruppen schauen wir uns die relativen Häufigkeiten der Ausprägungen des Merkmals B an. Man spricht auch von **bedingten relativen Häufigkeiten**.
Für die bedingte relative Häufigkeit der Merkmalsausprägung B_j unter der Bedingung, dass die Merkmalsausprägung A_i gegeben ist, schreiben wir $h_{B_j|A_i}$.
Es gilt

$$\boxed{h_{B_j|A_i} = \frac{n_{ij}}{n_{i\cdot}}} \tag{4.4}$$

Den allgemeinen Aufbau einer Tabelle mit bedingten relativen Häufigkeiten zeigt Tabelle 4.8. Die Zeilen dieser Tabelle bezeichnet man auch als *Profile*. Die Summe der zahlen in einer Zeile der Tabelle der bedingten relativen Häufigkeiten ist gleich 1. Es macht keinen Sinn, die Spaltensummen zu bestimmen.

Tabelle 4.8: Allgemeiner Aufbau einer Kontingenztabelle mit bedingten relativen Häufigkeiten

B / A	B_1	B_2	$\dots$	B_J
A_1	$h_{B_1\|A_1}$	$h_{B_2\|A_1}$	$\cdots$	$h_{B_J\|A_1}$
A_2	$h_{B_1\|A_2}$	$h_{B_2\|A_2}$	$\cdots$	$h_{B_J\|A_2}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
A_I	$h_{B_1\|A_I}$	$h_{B_2\|A_I}$	$\cdots$	$h_{B_J\|A_I}$

Beispiel 36 (fortgesetzt von Seite 144)
Wir betrachten weiterhin die Merkmale `Geschlecht` und `Titanic` und schauen uns zunächst die Frauen an. Der Tabelle 4.7 auf der vorherigen Seite entnehmen wir, dass von den 13 Frauen 12 den Film Titanic gesehen haben. Also gilt

$$h_{B_1|A_1} = 0.923$$

Also haben 92.3 Prozent der Frauen den Film Titanic gesehen.
Schauen wir uns nun die Männer an. Von den 12 Männern haben 5 den Film Titanic gesehen. Also gilt

$$h_{B_1|A_2} = 0.417$$

Also haben 41.7 Prozent der Männer den Film Titanic gesehen.
Unter den Frauen ist der Anteil derjenigen, die den Film gesehen haben, viel größer als unter den Männern.
Wir erhalten die bedingten relativen Häufigkeiten in Tabelle 4.9.

Tabelle 4.9: Kontingenztabelle der Merkmale `Geschlecht` und `Titanic` mit bedingten relativen Häufigkeiten

`Titanic` / `Geschlecht`	j	n
w	0.923	0.077
m	0.417	0.583

□

Im Beispiel 5 auf Seite 13 sind bedingte relative Häufigkeiten gegeben. Hier wurde untersucht, ob es sinnvoller ist, ein Insektizid oder den Nissenkamm zu benutzen, um die Läuse loszuwerden. Von denen, die einen Kamm benutzt hatten, waren 57 Prozent lausfrei, während von denen, die ein Insektizid benutzt hatten, nur 13 Prozent lausfrei waren. Hier sind genau die bedingten relativen Häufigkeiten gegeben, an denen man interessiert ist. Wir wollen ja wissen, welche der beiden Methoen erfolgreicher ist.

4.2.3 Grafische Darstellung bedingter relativer Häufigkeiten

Wir können die Verteilungen der beiden Gruppen auch grafisch vergleichen. Hierzu gibt es eine Reihe von Möglichkeiten.

Zum einen können wir die `Profile` zeichnen. Hier tragen wir in einem kartesischen Koordinatensystem auf der Abszisse die Ausprägungen des Merkmals B ab. Dann tragen wir auf Höhe dieser Merkmalsausprägungen die bedingten relativen Häufigkeiten ab und verbinden die Punkte, die zu einer Merkmalsausprägung von A gehören.

Beispiel 36 (fortgesetzt von Seite 145)
Wir betrachten die Daten in Tabelle 4.9 auf der vorherigen Seite. Abbildung 4.4 zeigt die Profile.

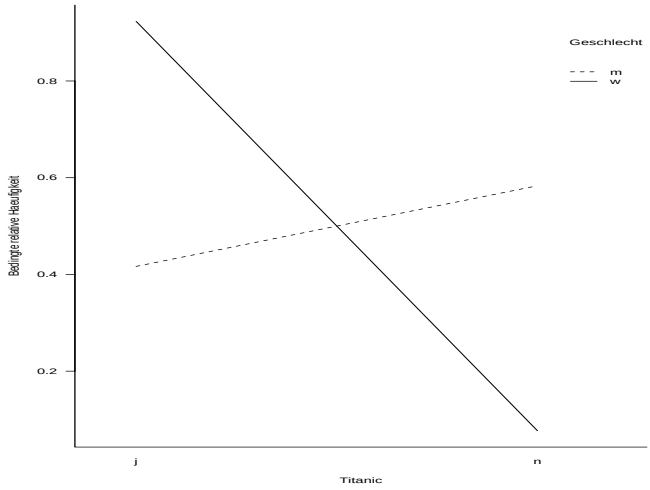


Abbildung 4.4: Profile

□

Eine andere grafische Darstellungsmöglichkeit bietet das **vergleichende Säulendiagramm**. Bei diesem zeichnen wir mehrere Säulendiagramme nebeneinander, wobei das erste Säulendiagramm die Verteilung des Merkmals B enthält, wenn das Merkmal A die Merkmalsausprägung A_1 annimmt. Dabei werden die Merkmalsausprägungen nach ihrer Häufigkeit absteigend sortiert. Das zweite ist ein Säulendiagramm der Häufigkeitsverteilung des Merkmals B, wenn das Merkmal A die Merkmalsausprägung A_2 annimmt. Bei diesem Säulendiagramm werden die Merkmalsausprägungen von B in der gleichen Reihenfolge abgetragen wie beim ersten Säuulendiagramm. Hierdurch ist ein direkter Vergleich möglich. Entsprechend werden für die anderen Merkmalsausprägungen des Merkmals B die Säulendiagramme erstellt.

Beispiel 36 (fortgesetzt von Seite 146)
Wir betrachten die Daten in Tabelle 4.9 auf Seite 145. Abbildung 4.5 zeigt das vergleichende Säulendiagramm.

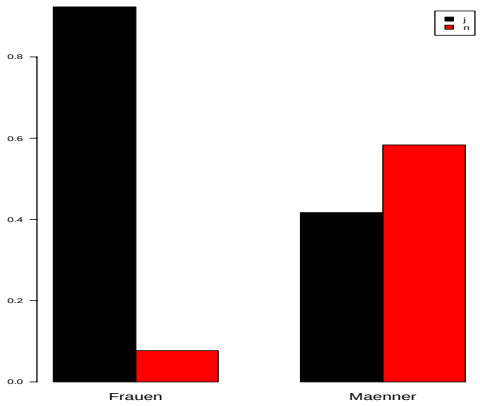


Abbildung 4.5: Vergleichendes Säulendiagramm

□

Bei den bisher betrachteten Grafiken kann man nicht feststellen, wie sich die Besetzungszahlen der Gruppen unterscheiden. Diese Möglichkeit bietet der **Mosaikplot**. Bei diesem wird für jede Zelle der Kontingenztabelle ein Rechteck gezeichnet, dessen Fläche proportional zur absoluten Häufigkeit der Zelle ist. Die Zellen der Zeilen bilden dabei die Spalten des Mosaikplots. Die Summen der vertikalen Seiten der Rechtecke ist in jeder Spalte der Kontingenztabelle konstant. Die Breiten der Rechtecke sind proportional zu den jeweiligen Spaltensummen. An einem Mosaikplot kann man die absoluten Häufigkeiten

der Zeilen erkennen. Diese zeigen sich in den Breiten der Rechtecke. Außerdem kann man die bedingten relativen Häufigkeiten der Merkmalsausprägung einer Zeile an den Längen der vertikalen Seiten der Rechtecke erkennen.

Beispiel 36 (fortgesetzt von Seite 147)
Wir betrachten die Daten in Tabelle 4.9 auf Seite 145. Abbildung 4.6 zeigt den Mosaikplot.

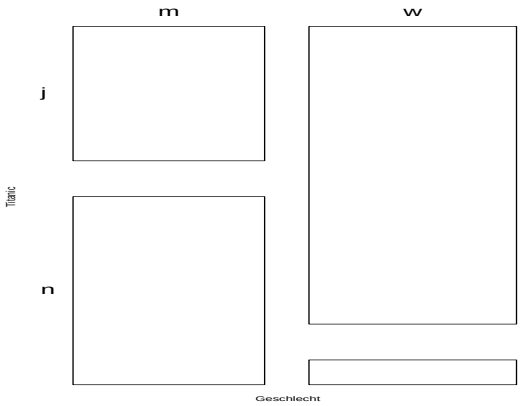


Abbildung 4.6: Mosaikplot

□

Schauen wir uns noch ein Beispiel an.

Beispiel 37
Die Anfangssemester des Wintersemesters 1996/1997 wurden befragt, welche Partei sie wählen würden. Die Werte sind in Tabelle 4.10 zu finden.

Tabelle 4.10: Geschlecht und Wahlverhalten von Studienanfängern

	CDU	SPD	FDP	GRUENE	keine	weiß nicht
weiblich	13	10	3	11	5	23
männlich	55	30	20	24	24	35

Die bedingten relativen Häufigkeiten bei den Frauen und Männern sind in der Tabelle 4.11 zu finden.

Tabelle 4.11: Verteilung des Wahlverhalten weiblicher und männlicher Studienanfänger

	CDU	SPD	FDP	GRUENE	keine	weiß nicht
weiblich	0,20	0,15	0,05	0,17	0,08	0,35
männlich	0,29	0,16	0,11	0,13	0,13	0,19

Die obere Grafik in Abbildung 4.7 zeigt die Profile.

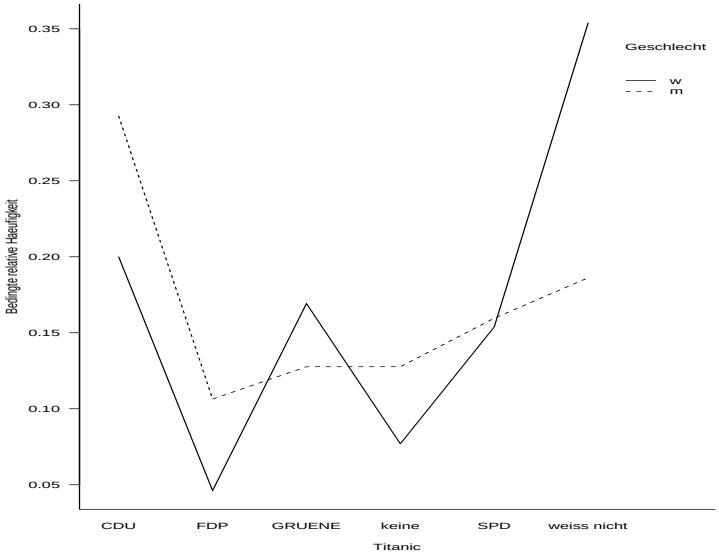


Abbildung 4.7: Profile

Abbildung 4.8 zeigt das vergleichende Säulendiagramm.

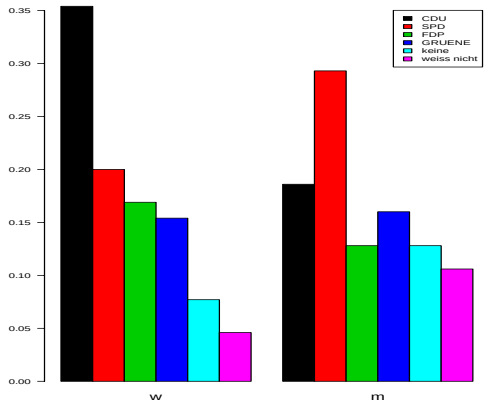


Abbildung 4.8: Vergleichendes Säulendiagramm

Wir sehen nun auf einen Blick, dass das Wahlverhalten der Studentinnen sich beträchtlich von dem der Studenten unterscheidet. Die häufigste Kategorie bei den Frauen ist `weiß nicht`, während bei den Männern die CDU präferiert wird.

Abbildung 4.9 zeigt den Mosaikplot. Wir sehen, dass die Anzahl von Männern und Frauen sich stark unterscheidet. Außerdem sieht man auf einen Blick, dass sich die bedingten Verteilungen stark unterscheiden.

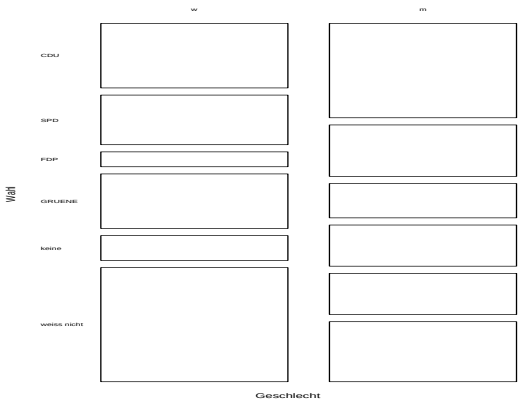


Abbildung 4.9: Mosaikplot

□

4.2.4 Der Kontingenzkoeffizient

Wir wollen nun eine geeignete Maßzahl entwickeln, die den Zusammenhang zwischen zwei nominalskalierten Merkmalen beschreibt. Hierzu überlegen wir uns zunächst, was es bedeutet, dass zwischen zwei nominalskalierten Merkmalen kein Zusammenhang besteht. Schauen wir uns dazu ein Beispiel an.

Beispiel 36 (fortgesetzt von Seite 148)
Wir betrachten Tabelle 4.7 auf Seite 144 an. Wir sehen, dass 92.3 Prozent der Frauen und 41.7 Prozent der Männer den Film Titanic gesehen haben.
Offensichtlich ist der Anteil derjenigen, die den Film gesehen haben, bei den Frauen viel höher als bei den Mänern. Zwischen den beiden Merkmalen besteht anscheinend ein Zusammenhang.
Wenn hingegen zwischen den Merkmalen `Geschlecht` und `Titanic` kein Zusammenhang besteht, dann muss der Anteil derjenigen, die den Film Titanic gesehen haben, sowohl bei den Frauen als auch bei den Männern gleich sein. In unserem Fall müsste er 0.68 betragen, denn von den 25 Teilnehmern haben 17, also 68 Prozent den Film Titanic gesehen.

□

Besteh zwischen zwei nominalskalierten Merkmalen kein Zusammenhang, so muss gelten

$$\frac{n_{11}}{n_{1\cdot}} = \frac{n_{\cdot 1}}{n} \tag{4.5}$$

$$\frac{n_{21}}{n_{2\cdot}} = \frac{n_{\cdot 1}}{n} \tag{4.6}$$

$$\frac{n_{12}}{n_{1\cdot}} = \frac{n_{\cdot 2}}{n} \tag{4.7}$$

und

$$\frac{n_{22}}{n_{2\cdot}} = \frac{n_{\cdot 2}}{n} \tag{4.8}$$

Die Gleichungen (4.5), (4.6), (4.7) und (4.8) können wir folgendermaßen kompakt schreiben:
Für $i = 1, 2$ und $J = 1, 2$ muss also gelten

$$\boxed{\frac{n_{ij}}{n_{i\cdot}} = \frac{n_{\cdot j}}{n}} \tag{4.9}$$

Burkschat et al. (2004) sprechen davon, dass die Merkmale A mit den Merkmalsausprägungen $A_1, A_2, \ldots, A_I$ und B mit den Merkmalsausprägungen $B_1, B_2, \ldots, B_J$ **empirisch unabhängig** sind, wenn die Gleichung (4.9) für $i = 1, 2, \ldots, I$ und $J = 1, 2, \ldots, J$ erfüllt ist.

Beispiel 36 (fortgesetzt von Seite 151)
Die Merkmale `Geschlecht` und `Titanic` in Tabelle 4.7 auf Seite 144 sind nicht empirisch unabhängig, da der Anteil derjenigen, die den Film Titanic gesehen haben, bei den Frauen 0.923 und bei den Männern 0.417 beträgt.
□

Wir wollen nun eine Maßzahl für den Zusammnhang zwischen zwei nominalskalierten Merkmalen betrachten. Hierzu schauen wir uns an, wie stark ihr Zusammenhang von der empirischen Unabhängigkeit abweicht, und bestimmen die absoluten Häufigkeiten unter der Annahme, dass die Merkmale empirisch unabhängig sind. Wir bezeichnen diese mit $\tilde{n}_{ij}$.
Aufgrund von Gleichung (4.9) muss für $\tilde{n}_{ij}$ gelten:

$$\frac{\tilde{n}_{ij}}{n_{i\cdot}} = \frac{n_{\cdot j}}{n} \tag{4.10}$$

Hieraus folgt

$$\tilde{n}_{ij} = \frac{n_{i\cdot}\, n_{\cdot j}}{n} \tag{4.11}$$

Man nennt $\tilde{n}_{ij}$ die **erwartete Häufigkeiten unter empirischer Unabhängigkeit**. Liegt kein Zusammenhang zwischen A und B vor, so stimmen die beobachteten Häufigkeiten n_{ij} mit den unter empirischer Unabhängigkeit erwarteten Häufigkeiten $\tilde{n}_{ij}$ überein. Der Zusammenhang ist um so stärker, je mehr beobachtete und erwartete Häufigkeiten differieren. Wir bilden die Differenz aus beobachteten und erwarteten Häufigkeiten:

$$n_{ij} - \tilde{n}_{ij}$$

Es liegt nun nahe, die Differenzen $n_{ij} - \tilde{n}_{ij}$ zu summieren, um ein Maß für den Zusammenhang zwischen A und B zu gewinnen. Es gilt

$$\sum_{i=1}^{I} \sum_{j=1}^{J} (n_{ij} - \tilde{n}_{ij}) = 0$$

Dies sieht man folgendermaßen:

$$\begin{aligned}\sum_{i=1}^{I}\sum_{j=1}^{J}(n_{ij}-\tilde{n}_{ij}) &= \sum_{i=1}^{I}\sum_{j=1}^{J}n_{ij}-\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{n_{i.}\,n_{.j}}{n}\\ &= n-\frac{1}{n}\sum_{i=1}^{I}n_{i.}\sum_{j=1}^{J}n_{.j}=n-\frac{1}{n}\,n\,n=0\end{aligned}$$

Da sich positive und negative Differenzen im Mittel aufheben, quadrieren wir die Differenzen:

$$(n_{ij}-\tilde{n}_{ij})^2$$

Wir könnten nun die Summe der quadrierten Differenzen als Maß für den Zusammenhang zwischen A und B wählen. Hierbei berücksichtigen wir aber nicht die relative Bedeutung der Differenzen. Eine Abweichung von 5 fällt bei einer erwarteten Häufigkeit von 10 viel stärker ins Gewicht als bei einer erwarteten Häufigkeit von 100. Wir berücksichtigen den Wert einer Abweichung, indem wir die Differenz noch durch die erwartete Häufigkeit dividieren:

$$\frac{(n_{ij}-\tilde{n}_{ij})^2}{\tilde{n}_{ij}}$$

Als Maßzahl für den Zusammenhang zwischen A und B erhalten wir

$$\boxed{X^2=\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij}-\tilde{n}_{ij})^2}{\tilde{n}_{ij}}=\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(n_{ij}-\frac{n_{i.}\,n_{.j}}{n}\right)^2}{\frac{n_{i.}\,n_{.j}}{n}}} \tag{4.12}$$

Beispiel 36 (fortgesetzt von Seite 152)
Es muss also gelten

$$\begin{aligned}\tilde{n}_{11} &= \frac{n_{1.}\,n_{.1}}{n}=\frac{13\cdot 17}{25}=8.84\\ \tilde{n}_{12} &= \frac{n_{1.}\,n_{.2}}{n}=\frac{13\cdot 8}{25}=4.16\\ \tilde{n}_{21} &= \frac{n_{2.}\,n_{.1}}{n}=\frac{12\cdot 17}{25}=8.16\\ \tilde{n}_{22} &= \frac{n_{2.}\,n_{.2}}{n}=\frac{12\cdot 8}{25}=3.84\end{aligned}$$

Es gilt

$$\begin{aligned} n_{11} - \tilde{n}_{11} &= 12 - 8.84 = 3.16 \\ n_{12} - \tilde{n}_{12} &= 1 - 4.16 = -3.16 \\ n_{21} - \tilde{n}_{21} &= 5 - 8.16 = -3.16 \\ n_{22} - \tilde{n}_{22} &= 7 - 3.84 = 3.16 \end{aligned}$$

Es gilt

$$\frac{(12 - 8.84)^2}{8.84} = 1.13 \qquad \frac{(1 - 4.16)^2}{4.16} = 2.4$$
$$\frac{(5 - 8.16)^2}{8.16} = 1.22 \qquad \frac{(7 - 3.84)^2}{3.84} = 2.6$$

Hieraus folgt

$$X^2 = 1.13 + 2.4 + 1.22 + 2.6 = 7.35$$

□

Besitzt eine Kontingenztabelle zwei Zeilen und zwei Spalten, dann gilt

$$\boxed{X^2 = \frac{n \cdot (n_{11} \cdot n_{22} - n_{12} \cdot n_{21})^2}{n_{1.} \cdot n_{2.} \cdot n_{.1} \cdot n_{.2}}} \tag{4.13}$$

Der Beweis dieser Beziehung ist bei Burkschat et al. (2004) auf der Seite 254 zu finden.

Beispiel 36 (fortgesetzt von Seite 153)
Tabelle 4.7 auf Seite 144 besitzt zwei Zeilen und zwei Spalten. Also gilt

$$X^2 = \frac{25 \cdot (12 \cdot 7 - 1 \cdot 5)^2}{13 \cdot 12 \cdot 17 \cdot 8} = 7.35$$

□

X^2 kann nicht negativ werden und ist genau dann gleich 0, wenn die Merkmale A und B empirisch unabhängig sind. Dies wird von Burkschat et al. (2004) auf Seite 252 gezeigt. Werte von X^2 in der Nähe von 0 sprechen also dafür, dass zwischen den Merkmalen A und B kein Zusammenhang besteht. Welche Werte von X^2 deuten aber auf einen Zusammenhang hin? Es gilt

$$\boxed{X^2 \leq n \cdot \min\{I - 1, J - 1\}} \tag{4.14}$$

Der Beweis von (4.14) ist auf Seite 184 zu finden.
Die obere Schranke von X^2 hängt also von n ab.
Pearson hat mit dem **Kontingenzkoeffizienten** eine Maßzahl für den Zusammenhang zwischen zwei nominalskalierten Merkmalen vorgeschlagen, deren obere Schranke nicht von n abhängt. Er ist folgendermaßen definiert:

$$\boxed{K = \sqrt{\frac{X^2}{X^2 + n}}} \tag{4.15}$$

Für den Kontingenzkoeffizienten K gilt

$$\boxed{0 \leq K \leq \sqrt{\frac{\min\{I-1, J-1\}}{\min\{I, J\}}}} \tag{4.16}$$

Der Beweis dieser Beziehung ist auf Seite 186 zu finden.
Dividiert man den Kontingenzkoeffizienten durch $\sqrt{\frac{\min\{I-1,J-1\}}{\min\{I,J\}}}$, so erhält man den **korrigierten Kontingenzkoeffizienten**

$$\boxed{K^* = \frac{K}{\sqrt{\frac{\min\{I-1,J-1\}}{\min\{I,J\}}}}} \tag{4.17}$$

Für den korrigierten Kontingenzkoeffizienten K gilt

$$\boxed{0 \leq K^* \leq 1} \tag{4.18}$$

Liegt der Wert des korrigierten Kontingenzkoeffizienten in der Nähe von 0, so spricht dies dafür, dass zwischen den Merkmalen kein Zusammenhang besteht. Werte in der Nähe von 1 sprechen für einen Zusammenhang zwischen den Merkmalen. Wir kennen aber nicht die Richtung des Zusammenhangs.

Beispiel 36 (fortgesetzt von Seite 154)
Wir haben gesehen, dass für Tabelle 4.7 auf Seite 144 $X^2 = 7.35$ gilt. Somit gilt

$$K = \sqrt{\frac{X^2}{X^2 + n}} = \sqrt{\frac{7.35}{7.35 + 25}} = 0.48$$

Mit $I = 2$ und $J = 2$ gilt

$$K^* = \frac{0.48}{\sqrt{\frac{\min\{1,1\}}{\min\{2,2\}}}} = \frac{0.48}{\sqrt{0.5}} = 0.68$$

Es liegt ein relativ starker Zusammenhang zwischen den beiden Merkmalen vor. Wir können dem Wert des korrigierten Kontingenzkoeffizienten aber nicht entnehmen, ob der Anteil derjenigen, die den Film gesehen haben, bei den Frauen oder bei den Männern höher ist.

□

Schauen wir uns noch ein Beispiel an.

Beispiel 38
Wir schauen uns noch einmal das Beispiel 5 auf Seite 13 an. Hier sind nur die bedingten relativen Häufigkeiten gegeben. Sind die Randhäufigkeiten $n_{i\cdot}$ gegeben, so kann man mit Gleichung (4.4) auf Seite 144 die Häufigkeiten n_{ij} bestimmen. Auf Seite 13 steht, dass die Teilnehmer in zwei gleich große Gruppen aufgeteilt wurden. Dies ist bei 133 Teilnehmern aber schwierig. Deshalb werfen wir einen Blick auf die Homepage des British Medical Journal. Dort steht, dass 126 Personen an der Studie teilnahmen. Von diesen verwendeten 70 das Insektizid, die anderen benutzten den Nissenkamm. Alle anderen Angaben des Artikels in der SZ stimmen. Ist A das Merkmal Behandlung mit den Merkmalsausprägungen A_1 gleich Nissenkamm und A_2 gleich $Insektizid$ und B der Erfolg der Behandlung mit den Merkmalsausprägungen B_1 keine Laus, dann gilt $h_{B_1|A_1} = 0.57$ und $h_{B_1|A_2} = 0.13$. Mit $n_{1\cdot} = 56$ und $n_{2\cdot} = 70$ gilt auf Grund von Gleichung (4.4) auf Seite 144 also

$$\begin{aligned} n_{11} &= h_{B_1|A_1} \cdot n_{1\cdot} = 0.57 \cdot 56 = 31.92 \\ n_{21} &= h_{B_1|A_2} \cdot n_{2\cdot} = 0.13 \cdot 70 = 9.1 \end{aligned}$$

Da die bedingten relativen Häufigkeiten gerundet sind, muss also gelten $n_{11} = 32$ und $n_{21} = 9$. Außerdem gilt

$$\begin{aligned} n_{12} &= n_{1\cdot} - n_{11} = 56 - 32 = 24 \\ n_{22} &= n_{2\cdot} - n_{21} = 70 - 9 = 61 \end{aligned}$$

Nun können wir X^2 bestimmen. Es gilt

$$X^2 = \frac{126 \cdot (32 \cdot 61 - 24 \cdot 9)^2}{56 \cdot 70 \cdot 41 \cdot 85} = 27.8$$

Somit gilt

$$K = \sqrt{\frac{X^2}{X^2 + n}} = \sqrt{\frac{27.8}{27.8 + 126}} = 0.42$$

Mit $I = 2$ und $J = 2$ gilt

$$K^* = \frac{0.42}{\sqrt{\frac{\min\{1,1\}}{\min\{2,2\}}}} = \frac{0.42}{\sqrt{0.5}} = 0.59$$

Der Zusammenhang ist relativ stark.

□

4.2.5 Die Analyse in R

Wir betrachten das Beispiel 1 auf Seite 10 und analysieren die gemeinsame Verteilung der nominalskalierten Merkmale `Geschlecht` und `Titanic`. Die Daten mögen in der Datentabelle `weiterbildung` stehen, die auf Seite 30 zu finden ist. Wir greifen auf die Variablen zu, wie auf Seite 30 beschrieben wird.

```
> Geschlecht
 [1] m w w m m m w m w m w m m w w m m m w w w w w w m
Levels: m w
> Film
 [1] n j j n n j j n j n j j j j j j n j n j j j j j n
Levels: j n
```

Die Kontingenztabelle erzeugen wir folgendermaßen mit der Funktion `table`:

`table`

```
> h.gf<-table(Geschlecht,Film)
> h.gf
           Film
 Geschlecht  j  n
          m  5  7
          w 12  1
```

Dabei bilden die Merkmalsausprägungen der ersten Variablen die Zeilen und die Merkmalsausprägungen der zweiten Variablen die Spalten der Tabelle.
Die Variable `h.gf` ist eine **Matrix**. Die Zeilen und Spalten der Matrix `h.gf` besitzen Namen. Auf diese Namen kann man mit der Funktion `dimnames` zugreifen.

`dimnames`

```
> dimnames(h.gf)
 $Geschlecht
 [1] "m" "w"

 $Film
 [1] "j" "n"
```

Das Ergebnis der Funktion `dimnames` ist eine Liste, deren erste Komponente den Vektor der Namen der Zeilen und deren zweite Komponente den Vektor der Namen der Spalten enthält. Bei den grafischen Darstellungen werden wir auf diese Namen zugreifen.

Oft liegen die Daten schon als Kontingenztabelle vor. Dies ist in Tabelle 4.10 auf Seite 148 der Fall. Wir können die Kontingenztabelle mit der Funktion `matrix` auch direkt als Matrix eingeben. Wir geben also ein

```
> h.gw<-matrix(c(13,55,10,30,3,20,11,24,5,24,23,35),2,6)
> h.gw
      [,1] [,2] [,3] [,4] [,5] [,6]
 [1,]   13   10    3   11    5   23
 [2,]   55   30   20   24   24   35
```

Mit der Funktion `dimnames` können wir den Zeilen und Spalten Namen geben.

```
> wahl<-c("CDU","SPD","FDP","GRUENE","keine","weiss nicht")
> dimnames(h.gw)<-list(Geschlecht=c("w","m"),Wahl=wahl)
> h.gw
           Wahl
 Geschlecht CDU SPD FDP GRUENE keine weiss nicht
          w  13  10   3     11     5          23
          m  55  30  20     24    24          35
```

Wir wollen nun die Randhäufigkeiten bestimmen. Hierzu haben wir zwei Möglichkeiten.

`prop.table` Wir verwenden die Funktion `prop.table`. Diese wird aufgerufen durch

```
prop.table(x, margin=NULL)
```

Dabei ist `x` die Tabelle und `margin` der Rand, auf denen wir die Häufigkeiten besitzen. Sollen die bedingten Verteilungen der Zeilen bestimmt werden, so ist `margin` gleich 1. Sollen die bedingten Verteilungen der Spalten bestimmt werden, so ist `margin` gleich 2. Standardmäßig werden die relativen Häufigkeiten der Tabelle bestimmt.

```
> prop.table(h.gf)
           Film
 Geschlecht    j    n
          m 0.20 0.28
          w 0.48 0.04

> prop.table(h.gf,1)
          Film
 Geschlecht          j          n
          m 0.41666667 0.58333333
          w 0.92307692 0.07692308

> prop.table(h.gf,2)
           Film
 Geschlecht         j         n
          m 0.2941176 0.8750000
          w 0.7058824 0.1250000
```

Zur Bestimmung der bedingten relativen Häufigkeiten können wir aber auch die Funktionen `apply` und `sweep` verwenden. Die Funktion `apply` wird auf Seite 28 beschrieben. Um den Vektor der Zeilensummen von `h.gf` zu erhalten, geben wir ein:

```
> apply(h.gf,1,sum)
  m  w
 12 13
```

Den Vektor der Spaltensummen erhalten wir durch

```
> apply(h.gf,2,sum)
  j  n
 17  8
```

Um die bedingten relativen Häufigkeiten zu erhalten, verwenden wir die Funktion `sweep`. Der Aufruf von `sweep` für eine Matrix `M` ist

`sweep`

```
sweep(M, MARGIN, STATS, FUN)
```

Dabei ist `MARGIN` die Dimension der Matrix, bezüglich der die Funktion angewendet werden soll. Dabei steht 1 für die Zeilen und 2 für die Spalten. Das Argument `STATS` ist ein Vektor, dessen Länge der Größe der Dimension entspricht, die im Argument `MARGIN` gewählt wurde, und das Argument `FUN` ist

der Name der Funktion, die auf `MARGIN` von `M` angewendet werden soll. Standardmäßig wird die Subtraktion gewählt. Die Funktion `sweep` bewirkt, dass die Funktion `FUN` angewendet wird, um die Komponenten des Vektors aus der gewählten Dimension von `M` im wahrsten Sinne des Wortes herauszufegen.
Die Matrix der auf die Zeilen bedingten relativen Häufigkeiten erhält man also durch:

```
> sweep(h.gf,1,apply(h.gf,1,FUN=sum),FUN="/")
           Film
 Geschlecht          j          n
          m 0.41666667 0.58333333
          w 0.92307692 0.07692308
```

Die Matrix der auf die Spalten bedingten relativen Häufigkeiten erhält man analog durch:

```
> sweep(h.gf,2,apply(h.gf,2,FUN=sum),FUN="/")
           Film
 Geschlecht         j         n
          m 0.2941176 0.8750000
          w 0.7058824 0.1250000
```

Schauen wir uns nun die Grafiken an.
Die Profile erzeugen wir folgendermaßen mit der Funktion `interaction.plot`:

`interaction.plot`

Wir erstellen zuerst mit der Funktion `prop.table` die Matrix `h.bz` der bedingten relativen Häufigkeiten und bestimmen die Anzahl `az` der Zeilen und die Anzahl `as` der Spalten dieser Matrix.

```
> h.bz<-prop.table(h.gf,1)
> az<-dim(h.bz)[1]
> as<-dim(h.bz)[2]
```

Zu jeder Zelle der Matrix `h.bz` gehört eine Ausprägung des Faktors A und eine Ausprägung des Faktors B. Wir erzeugen Vektoren `A` und `B`, die dieses Merkmalsausprägungen enthalten. Dabei durchlaufen wir die Matrix spaltenweise.

```
> A<-factor(rep(dimnames(h.bz)[[1]],as))
> A
 [1] m w m w
 Levels: m w
```

```
> B<-factor(rep(dimnames(h.bz)[[2]],rep(az,as)))
> B
 [1] j j n n
 Levels: j n
```

Mit der Funktion `as.vector` bilden wir einen Vektor `v` aus den Spalten der Matrix `h.bz`. `as.vector`

```
> v<-as.vector(h.bz)
> v
[1] 0.41666667 0.92307692 0.58333333 0.07692308
```

Nun rufen wir die Funktion `interaction.plot` auf.

```
> interaction.plot(B,A,v,xlab="Film",trace.label="Geschlecht",
                   ylab="Bedingte relative Haeufigkeit",bty="l")
```

Das vergleichende Säulendiagramm in Abbildung 4.5 auf Seite 147 erhalten wir mit der Funktion `barplot`. Zuerst vertauschen wir die beiden Zeilen der matrix `h.bz`, da wir mit den Frauen beginnen. `barplot`

```
> h.bz<-h.bz[c(2,1),]
```

Nun müssen wir noch die Spalten der Matrix `h.bz` in die richtige Reihenfolge bringen. Hierzu benutzen wir die Funktion `order`. Für einen numerischen Vektor `v` liefert der Aufruf `order`

```
order(v)
```

einen Vektor, dessen i-te Komponente angibt, an der wievielten Stelle das i-t kleinste Element im Vektor `v` steht. Für den Vektor

```
> v
[1] 7 4 6 3 5
```

liefert der Aufruf

```
> order(v)
```

folgendes Ergebnis

```
[1] 4 2 5 3 1
```

Mit der Funktion `order` können wir sowohl die bedingten relativen Häufigkeiten als auch die Namen der Merkmalsausprägungen in die richtige Ordnung bringen. Wir bilden also

```
> o<-order(h.bz[1,],decreasing=T)
> o
[1] 1 2
```

Nun rufen wir die Funktion `barplot` auf.

```
> barplot(t(h.bz),legend.text=dimnames(h.bz)[[2]][o],
          col=0:(dim(h.bz)[1]-1),beside=T,
          names.arg=dimnames(h.bz)[[1]])
```

Schauen wir uns diesen Aufruf genauer an. Die Matrix `m` wird mit der Funktion `t` durch

```
t(m)
```

transponiert. Es werden also Zeilen und Spalten der Matrix vertauscht. Dies ist nötig, da die Funktion `barplot` die bedingten Verteilungen in den Spalten erwartet. Das Argument `legend.text` erstellt eine Legende der Merkmalsausprägungen des interessierenden Merkmals. Durch das Argument `col` werden die Farben für die Säulen festgelegt. Das Argument `beside` zeichnet die Säulen nebeneinander, wenn es auf `TRUE` steht.
Das vergleichende Säulendiagramm in Abbildung 4.8 auf Seite 150 erhalten wir durch folgende Befehlsfolge:

```
>  wahl<-matrix(c(13,55,10,30,3,20,11,24,5,24,23,35),2,6)
>  partei<-c("CDU","SPD","FDP","GRUENE","keine","weiss nicht")
>  dimnames(wahl)<-list(Geschlecht=c("w","m"),Wahl=partei)
>  h<-prop.table(wahl,1)
>  o<-order(h[1,],decreasing=T)
>  par(las=1)
>  barplot(t(h[,o]),legend.text=partei[o],col=1:6,beside=T,
                    names.arg=c("w","m"))
```

`mosaicplot` Einen Mosaikplot erhält man mit der Funktion `mosaicplot`. Abbildung 4.6 auf Seite 148 erhalten wir durch folgende Befehle

```
> attach(weiterbildung)
> h<-table(Geschlecht,Film)
> par(las=1)
> mosaicplot(h,main="")
```

`vcd` Um den Kontingenzkoeffizienten zu bestimmen, müssen wir das Paket `vcd` installieren und laden. Wie man dabei vorzugehen hat, wird auf Seite 52

beschrieben.

socstats Im Paket `vcd` gibt es die Funktion `assocstats`. Mit dieser kann man unter anderem den Kontingenzkoeffizienten bestimmen.

```
> assocstats(h)
                     X^2 df  P(> X^2)
 Likelihood Ratio 7.9919  1 0.0046987
 Pearson          7.3541  1 0.0066909

 Phi-Coefficient   : 0.542
 Contingency Coeff.: 0.477
 Cramer's V        : 0.542
```

Wenn wir nur den Kontingenzkoeffizienten wollen, geben wir

```
> assocstats(h)$cont
[1] 0.4767608
```

oder

```
> assocstats(h)[[4]]
[1] 0.4767608
```

ein.

4.3 Zusammenhang zwischen zwei quantitativen Merkmalen

Wir wollen nun zwei quantitative Merkmale betrachten. Hier kann man mit einem Streudiagramm den Zusammenhang zwischen den Merkmalen grafisch darzustellen. Außerdem gibt es einfache Maßzahlen, die die Stärke des Zusammenhangs zwischen den beiden Merkmalen beschreiben.

4.3.1 Das Streudiagramm

Ausgangspunkt sind die quantitativen Merkmale X und Y, die an jedem von n Merkmalsträgern erhoben wurden. Beim i-ten Merkmalsträger beobachten wir also einen Wert x_i des Merkmals X und einen Wert y_i des Merkmals Y. Wir fassen diese zu einem Vektor (x_i, y_i) zusammen.

Um uns ein Bild vom Zusammenhang zwischen den beiden Merkmalen zu machen, stellen wir die Vektoren in einem **Streudiagramm** dar. Das Merkmal

X ordnen wir der horizontalen Achse, der **Abszisse**, und das Merkmal Y der vertikalen Achse, der **Ordinate**, in einem kartesischen Koordinatensystem zu. Die Werte jedes Merkmalsträgers werden als Punkt in dieses Koordinatensystem eingetragen.

Beispiel 39
In einem Projekt im Hauptstudium wurden die Studierenden unter anderem nach ihren Noten im Abitur und im Vordiplom gefragt. Tabelle 4.12 zeigt die Noten von sechs Studierenden.

Tabelle 4.12: Noten im Abitur und im Vordiplom

Student	Note im Abitur	Note im Vordiplom
1	1.7	2.2
2	2.4	2.4
3	2.0	2.1
4	1.1	1.8
5	2.9	2.7
6	3.1	2.6

Abbildung 4.10 zeigt das Streudiagramm.

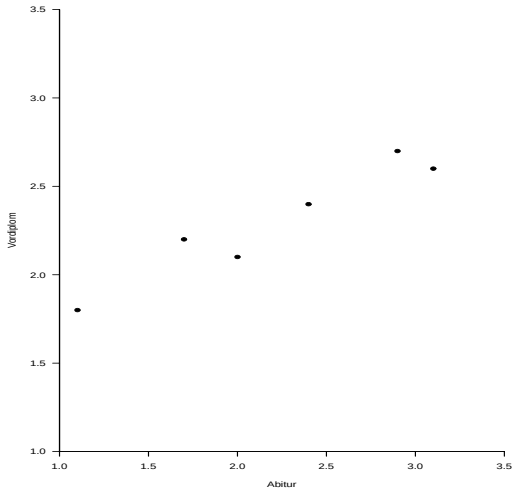


Abbildung 4.10: Streudiagramm der Noten im Abitur und im Vordiplom

Wir sehen, dass zwischen der Note im Abitur und der Note im Vordiplom ein **positiver Zusammenhang** besteht. Wenn ein Studierender eine gute Note im Abitur hat, dann hat er in der Regel auch eine gute Note im Vordiplom. □

Abbildung 4.11 zeigt einige typische Streudiagramme.

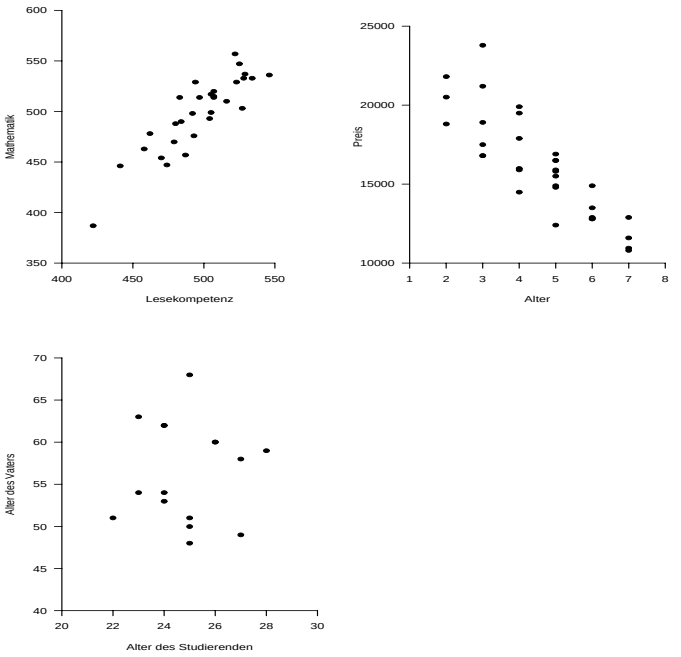


Abbildung 4.11: Beispiele von Streudiagrammen

Das Streudiagramm in Abbildung 4.11 links oben zeigt den Zusammenhang zwischen den Punkten in den Bereichen Lesekompetenz und Mathematische Grundausbildung im Rahmen der PISA-Studie des Jahres 2000 (siehe Deutsches PISA-Konsortium (Hrsg.) (2001)). Je besser ein Land im Bereich Lesekompetenz ist, um so besser ist es in der Regel auch im Bereich Mathematische Grundbildung. Das Streudiagramm deutet auf einen **positiven Zusammenhang** hin.

Das Streudiagramm in Abbildung 4.11 rechts oben stellt Daten ausder Süddeutschen Zeitung vom Ende Juli 1999 dar. Dort wurden im Anzeigenteil 33 VW-Golf 3 angeboten. Die Abbildung zeigt das Streudiagramm zwischen dem Alter eines VW-Golf 3 und dem Angebotspreis in DM. Das Streudiagramm deutet auf einen **negativen Zusammenhang** hin. Je älter ein VW-Golf 3

ist, um so niedriger ist der Angebotspreis.
Das Streudiagramm in Abbildung 4.11 links unten zeigt den Zusammenhang zwischen dem Alter von Studierenden und ihren Vätern. Es deutet weder auf einen positiven noch auf einen negativen Zusammenhang zwischen den beiden Merkmalen hin.
Bisher waren wir bei der Betrachtung von Streudiagrammen an den Merkmalen interessiert. Ein Streudiagramm zeigt uns, ob zwischen den Merkmalen ein Zusammenhang besteht. Wir können aber auch die Merkmalsträger in den Mittelpunkt stellen. Hier suchen wir Gruppen von Merkmalsträgern, sodass die Merkmalsträger innerhalb einer Gruppe sich hinsichtlich der Merkmale ähnlich sind, die Gruppen sich aber unterscheiden.

Beispiel 40
In einem BI-Projekt wurden die 15 Teilnehmer unter anderem nach dem Alter ihrer Eltern gefragt. Abbildung 4.12 zeigt das Streudiagramm der Merkmale `Alter der Mutter` und `Alter des Vaters`.

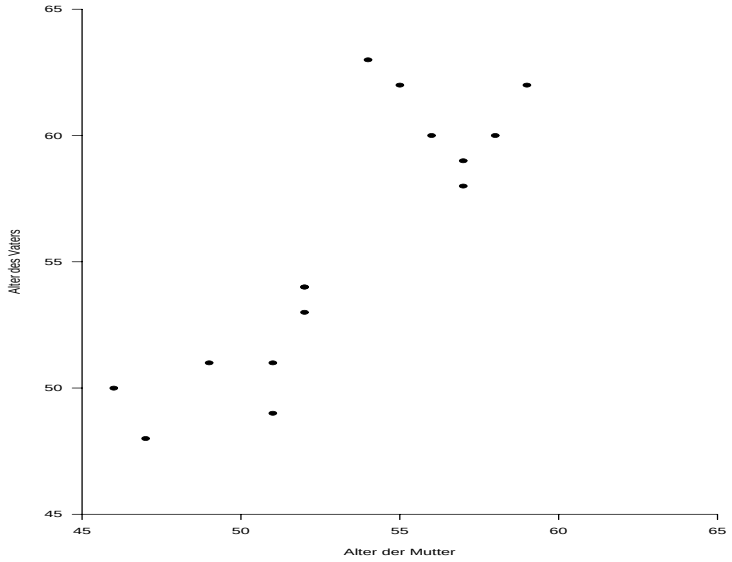


Abbildung 4.12: Streudiagramm der Merkmale `Alter der Mutter` und `Alter des Vaters`

Wir können im Streudiagramm zwei Gruppen identifizieren. In der einen Gruppe sind Studierende, deren Eltern beide relativ jung sind, während in der zweiten Gruppe Studierende mit relativ alten Eltern sind. In der ersten Gruppe ist das Durchschnittsalter der Mütter 50 Jahre und das der Väter

51.2 Jahre, während in der zweiten Gruppe das Durchschnittsalter der Mütter 56.6 Jahre und das der Väter 60.6 Jahre beträgt.

□

4.3.2 Maßzahlen für den Zusammenhang zwischen quantitativen Merkmalen

Bisher haben wir nur an Hand des Streudiagramms entschieden, welcher Zusammenhang zwischen zwei quantitiven Merkmalen vorliegt. Nun wollen wir Maßzahlen betrachten, mit denen wir diesen Zusammenhang beschreiben können. Wir beginnen mit einer Maßzahl für quantitative Merkmale.

Der Korrelationskoeffizient von Bravais-Pearson

Wir suchen eine Maßzahl für den Zusammenhang zwischen zwei quantitativen Merkmalen und schauen wir uns ein Beispiel an.

Beispiel 39 (fortgesetzt von Seite 164)
Wir schauen uns noch einmal das Streudiagramm der Merkmale `Note im Abitur` und `Note im Vordiplom` in Abbildung 4.10 auf Seite 164 an. Wir sehen, dass Studierende mit einer guten Note im Abitur in der Regel auch eine gute Note im Vordiplom haben und Studierende mit einer schlechten Note im Abitur in der Regel auch eine schlechte Note im Vordiplom haben. Liegt die Note im Abitur eines Studierenden also über dem Durchschnitt, so liegt in der Regel auch die Note im Vordiplom über dem Durchschnitt. Dies wird auch am Streudiagramm deutlich, wenn wir die Mittelwerte der beiden Merkmale in diesem berücksichtigen. Hierzu zeichnen wir eine Gerade parallel zur Ordinate in Höhe des Mittelwerts der Note im Abitur und eine Gerade parallel zur Abszisse in Höhe des Mittelwerts der Note im Vordiplom. Abbildung 4.13 auf der nächsten Seite veranschaulicht dies.
Wir erhalten 4 Quadranten, die in der Graphik durchnummeriert sind. Im ersten Quadranten sind die Studierenden, deren Noten in beiden Prüfungen schlechter als der Durchschnitt sind, während sich im dritten Quadranten die Studierenden befinden, deren Noten in beiden Prüfungen besser als der Durchschnitt sind. Im zweiten Quadranten sind die Studierenden, deren Note im Abitur besser als der Durchschnitt und deren Note im Vordiplom schlechter als der Durchschnitt ist, während sich im vierten Quadranten die Studierenden befinden, deren Note im Abitur schlechter als der Durchschnitt und deren Note im Vordiplom besser als der Durchschnitt ist. Besteht ein positiver Zusammenhang zwischen den beiden Merkmalen, so werden wir die meisten Beobachtungen in den Quadranten I und III erwarten, während bei

einem negativen Zusammenhang die meisten Punkte in den Quadranten II und IV liegen. Verteilen sich die Punkte gleichmäßig über die Quadranten, so liegt kein Zusammenhang zwischen den Merkmalen vor.

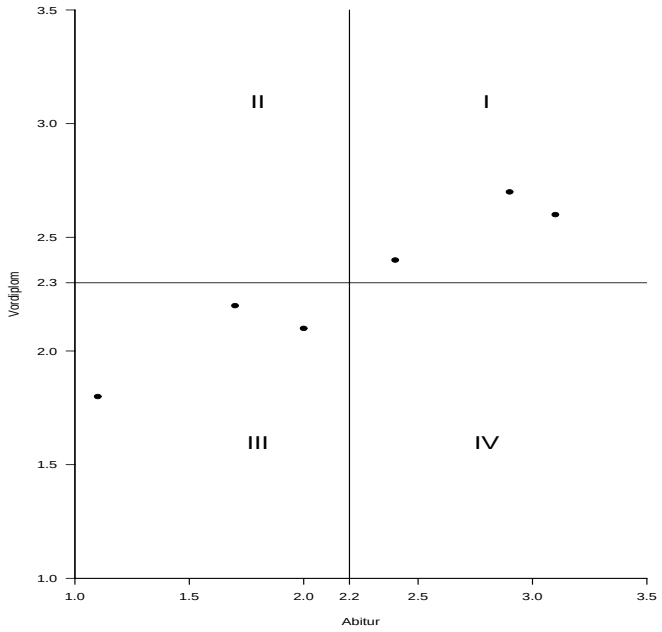


Abbildung 4.13: Streudiagramm der Noten im Abitur und im Vordiplom, aufgeteilt in 4 Quadranten

□

Eine einfache Maßzahl für den Zusammenhang zwischen den beiden Merkmalen erhalten wir, indem wir zählen. Sei n_i die Anzahl der Punkte im i-ten Quadranten. Wir bilden

$$\boxed{\frac{n_1 + n_3 - n_2 - n_4}{n_1 + n_2 + n_3 + n_4}} \tag{4.19}$$

Der Koeffizient in Gleichung (4.19) heißt auch **Korrelationskoeffizient von Fechner**. Er ist positiv, wenn die Mehrzahl der Punkte im ersten und dritten Quadranten liegt, er ist negativ, wenn die Mehrzahl der Punkte im zweiten und vierten Quadranten liegt, und er liegt in der Nähe von 0, wenn sich die Punkte gleichmäßig auf die vier Quadranten verteilen. Er nimmt genau dann

den Wert 1 an, wenn alle Punkte im ersten und dritten Quadranten liegen, und er nimmt genau dann den Wert -1 an, wenn alle Punkte im zweiten und vierten Quadranten liegen. Somit bildet er ein sinnvolles Maß für den Zusammenhang zwischen den Merkmalen, das zudem noch normiert ist.

Beispiel 39 (fortgesetzt von Seite 167)
Wir schauen uns die Abbildung 4.13 an. Es gilt

$$n_1 = 3 \qquad n_2 = 0 \qquad n_3 = 3 \qquad n_4 = 0$$

Also folgt

$$\frac{n_1 + n_3 - n_2 - n_4}{n_1 + n_2 + n_3 + n_4} = \frac{3 + 3 - 0 - 0}{3 + 0 + 3 + 0} = 1$$

□

Der Korrelationskoeffizient von Fechner hat den Nachteil, dass er nicht angibt, wie gut der Zusammenhang durch eine Funktion beschrieben werden kann. Um eine solche Maßzahl zu erhalten, dürfen wir nicht nur zählen, sondern müssen auch die Werte selber berücksichtigen.
Schauen wir uns also noch einmal die vier Quadranten an. Es gilt

Quadrant I: $x_i > \overline{x}$, $y_i > \overline{y}$,
Quadrant II: $x_i < \overline{x}$, $y_i > \overline{y}$,
Quadrant III: $x_i < \overline{x}$, $y_i < \overline{y}$,
Quadrant IV: $x_i > \overline{x}$, $y_i < \overline{y}$.

Wir schauen uns nun die Differenzen aus den Beobachtungen und dem Mittelwert an. Wir bilden also $x_i - \overline{x}$ und $y_i - \overline{y}$ für $i = 1, \ldots, n$.
In den 4 Quadranten gilt:

Quadrant I: $x_i - \overline{x} > 0$, $y_i - \overline{y} > 0$,
Quadrant II: $x_i - \overline{x} < 0$, $y_i - \overline{y} > 0$,
Quadrant III: $x_i - \overline{x} < 0$, $y_i - \overline{y} < 0$,
Quadrant IV: $x_i - \overline{x} > 0$, $y_i - \overline{y} < 0$.

Wir betrachten für $i = 1, \ldots, n$ das Produkt

$$(x_i - \overline{x}) \cdot (y_i - \overline{y}) \,. \tag{4.20}$$

Dieses ist für Punkte im ersten und dritten Quadranten positiv, während es für Punkte im zweiten und vierten Quadranten negativ ist. Addieren wir

die Produkte aus Gleichung (4.20) für alle Punktepaare, so werden wir für die Summe einen positiven Wert erhalten, wenn die meisten Punkte im ersten und dritten Quadranten liegen, einen negativen Wert, wenn die meisten Punkte im zweiten und vierten Quadranten liegen, und ungefähr 0, wenn die Punkte sich gleichmäßig auf die vier Quadranten verteilen.
Wir beziehen diese Summe auf die Anzahl der Beobachtungen und erhalten folgende Maßzahl für den Zusammenhang zwischen den beiden Merkmalen:

$$\boxed{d_{x,y} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x}) \cdot (y_i - \overline{y})} \tag{4.21}$$

Man nennt $d_{x,y}$ die **empirische Kovarianz**.

Beispiel 39 (fortgesetzt von Seite 169)
Wir wollen nun die empirische Kovarianz für die Daten in Tabelle 4.12 auf Seite 164 berechnen. Tabelle 4.13 enthält die relevanten Hilfsgrößen zur Berechnung.

Tabelle 4.13: Hilfstabelle zur Bestimmung der empirischen Kovarianz und des Korrelationskoeffizienten von Bravais-Pearson

i	x_i	y_i	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x}) \cdot (y_i - \overline{y})$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$
1	1.7	2.2	-0.5	-0.1	0.05	0.25	0.01
2	2.4	2.4	0.2	0.1	0.02	0.04	0.01
3	2.0	2.1	-0.2	-0.2	0.04	0.04	0.04
4	1.1	1.8	-1.1	-0.5	0.55	1.21	0.25
5	2.9	2.7	0.7	0.4	0.28	0.49	0.16
6	3.1	2.6	0.9	0.3	0.27	0.81	0.09

Aus der sechsten Spalte erhält man die wichtige Größe durch Addition.

$$\sum_{i=1}^{n} (x_i - \overline{x}) \cdot (y_i - \overline{y}) = 1.21$$

Also gilt

$$d_{x,y} = \frac{1.21}{6} = 0.202$$

□

Man kann die empirische Kovarianz einfacher berechnen. Es gilt

$$\boxed{d_{x,y} = \overline{x\,y} - \overline{x} \cdot \overline{y}} \tag{4.22}$$

mit

$$\overline{x\,y} = \frac{1}{n} \sum_{i=1}^{n} x_i \cdot y_i$$

Der Beweis von Gleichung (4.22) ist auf Seite 186 zu finden.

Beispiel 39 (fortgesetzt von Seite 170)
Wir bestimmen die empirische Kovarianz mit Formel (4.22). Tabelle 4.14 enthält die relevanten Hilfsgrößen zur Berechnung.

Tabelle 4.14: Hilfstabelle zur Bestimmung der empirischen Kovarianz

i	x_i	y_i	$x_i \cdot y_i$
1	1.7	2.2	3.74
2	2.4	2.4	5.76
3	2.0	2.1	4.20
4	1.1	1.8	1.98
5	2.9	2.7	7.83
6	3.1	2.6	8.06

Wir erhalten

$$\overline{x\,y} \;=\; \frac{1}{n} \sum_{i=1}^{n} x_i\,y_i = \frac{1}{6}\,(3.74 + 5.76 + 4.2 + 1.98 + 7.83 + 8.06) = 5.262$$

Mit $\overline{x} = 2.2$ und $\overline{y} = 2.3$ gilt somit

$$d_{x,y} = \overline{x\,y} - \overline{x}\,\overline{y} = 5.262 - 2.2 \cdot 2.3 = 0.202$$

□

Die Wert der empirischen Kovarianz hängt von der Maßeinheit der Merkmale ab. Nehmen wir an, dass wir den Wert der empirischen Kovarianz zwischen der Körpergröße x_i in Metern und dem Körpergewicht y_i in Kilogramm

bestimmt haben. Betrachten wir nun die Körpergröße in Zentimetern, so müssen wir jedes x_i mit 100 multiplizieren. Aufgrund von Gleichung (3.14) auf Seite 95 ist dann der Mittelwert auch 100-mal so groß. Setzen wir diese Größen in Gleichung (4.21) auf Seite 170 ein, so gilt

$$\begin{aligned} d_{100\cdot x,y} &= \frac{1}{n}\sum_{i=1}^{n}(100\cdot x_i - 100\cdot\overline{x})\cdot(y_i-\overline{y}) \\ &= 100\cdot\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})\cdot(y_i-\overline{y}) = 100\cdot d_{x,y} \end{aligned}$$

Die empirische Kovarianz zwischen der Körpergröße in cm und dem Gewicht in kg ist 100-mal so groß wie die empirische Kovarianz zwischen der Körpergröße in m und dem Gewicht in kg. Man kann mit Hilfe der empirischen Kovarianz aber nicht angeben, ob ein Zusammenhang stark oder schwach ist.
Wenn wir die empirische Kovarianz normieren, erhalten den **Korrelationskoeffizienten von Bravais-Pearson**:

$$r_{x,y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})\cdot(y_i-\overline{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\cdot\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2}} = \frac{d_{x,y}}{\sqrt{d_x^2\,d_y^2}} \tag{4.23}$$

Multiplizieren wir Zähler und Nenner des Bruches in Gleichung (4.23) mit n, so gilt:

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

Beispiel 39 (fortgesetzt von Seite 171)
Da im Zähler des Korrelationskoeffizienten von Bravais-Pearson die empirische Kovarianz steht, müssen wir nur die beiden Größen im Nenner bestimmen. Dazu benutzen wir die siebte und achte Spalte der Tabelle 4.13.
Es gilt

$$\sum_{i=1}^{n}(x_i-\overline{x})^2 = 0.473 \qquad \sum_{i=1}^{n}(y_i-\overline{y})^2 = 0.093\,.$$

Also gilt

$$r_{x,y} = \frac{0.2}{\sqrt{0.473 \cdot 0.093}} = 0.95$$

Die beiden Merkmale sind also stark positiv miteinander korreliert.

□

Für den empirischen Korrelationskoeffizienten $r_{x,y}$ gilt:

1. $-1 \leq r_{x,y} \leq 1$,
2. $r_{x,y} = 1$ genau dann, wenn zwischen den beiden Merkmalen ein exakter linearer Zusammenhang mit positiver Steigung besteht,
3. $r_{x,y} = -1$ genau dann, wenn zwischen den beiden Merkmalen ein exakter linearer Zusammenhang mit negativer Steigung besteht.

Der Beweis dieser drei Eigenschaften ist bei Burkschat et al. (2004) auf den seiten 270-271 zu finden.
Die erste Eigenschaft besagt, dass der Korrelationskoeffizient von Bravais-Pearson Werte zwischen -1 und 1 annimmt. Er ist also normiert.
Die beiden anderen Eigenschaften erklären, wie wir die Werte des Korrelationskoeffizienten von Bravais-Pearson zu interpretieren haben. Liegt der Wert des Korrelationskoeffizienten von Bravais-Pearson in der Nähe von 1, so liegt ein positiver **linearer Zusammenhang** zwischen den beiden Merkmalen vor, während ein Wert in der Nähe von -1 auf einen negativen linearen Zusammenhang hindeutet. Ein Wert in der Nähe von 0 spricht dafür, dass kein linearer Zusammenhang zwischen den beiden Merkmalen vorliegt. Dies bedeutet aber nicht notwendigerweise, dass gar kein Zusammenhang zwischen den beiden Merkmalen besteht. Dies verdeutlicht das folgende Beispiel.

Beispiel 41
Tabelle 4.15 zeigt die Realisationen von zwei Merkmalen.

Tabelle 4.15: Werte der Merkmale x und y

i	x_i	y_i	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x}) \cdot (y_i - \overline{y})$
1	-2	4	-2	2	-4
2	-1	1	-1	-1	1
3	0	0	0	-2	0
4	1	1	1	-1	-1
5	2	4	2	2	4

Es gilt

$$d_{x,y} = \frac{1}{5}(-4 + 1 + 0 - 1 + 4) = 0$$

Der Wert des empirischen Kovarianz ist also gleich 0. Somit ist auch der Wert des Korrelationskoeffizienten von Bravais-Pearson gleich 0. Schaut man sich die Werte in der Tabelle genauer an, so stellt man fest, dass $y_i = x_i^2$ gilt. Zwischen den beiden Merkmalen besteht also ein funktionaler Zusammenhang.

□

Ist der Wert des Korrelationskoeffizienten gleich 0, so besteht kein **linearer** Zusammenhang zwischen den Merkmalen. Es kann aber durchaus ein anderer funktionaler Zusammenhang bestehen.

Schauen wir uns noch einmal die Streudiagramme in Abbildung 4.11 auf Seite 165 an. Der Korrelationskoeffizient von Bravais-Pearson nimmt bei diesen folgende Werte an:

Abbildung links oben:	$r_{x,y} = 0.92$
Abbildung rechts oben:	$r_{x,y} = -0.85$
Abbildung links unten:	$r_{x,y} = 0.04$

Beispiel 42

Den Zusammenhang zwischen der Note im Abitur und der Note im Vordiplom haben wir für sechs ausgewählte Studierende dargestellt. Abbildung 4.14 zeigt das Streudiagramm der Noten für 124 Studierende in Projekten zur Betriebsinformatik in den Jahren 2004 und 2005.

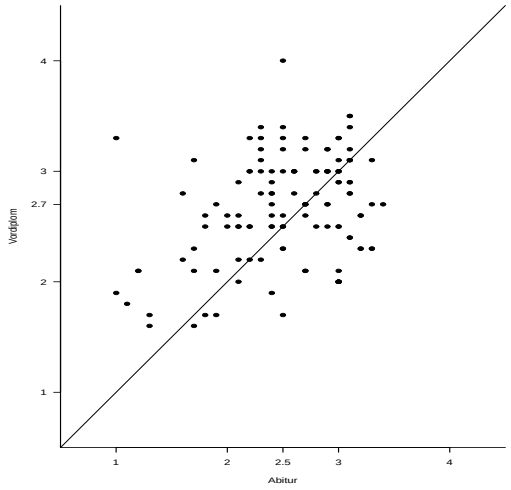


Abbildung 4.14: Streudiagramm der Noten im Abitur und im Vordiplom

An diesem Streudiagramm können wir einiges erkennen. Zwischen den beiden Noten besteht ein positiver linearer Zusammenhang, der aber nicht sehr stark ist, da der Wert des Korrelationskoeffizienten von Bravais-Pearson 0.33 beträgt. Außerdem sticht uns ein Ausreißer ins Auge. Dieser Studierende hat im Abitur die Note 1.0 und im Vordiplom die Note 3.3. Ohne diesen Ausreißer erhöht sich der Wert des Korrelationskoeffizienten von Bravais-Pearson auf 0.38. Der Korrelationskoeffizient von Bravais-Pearson ist also wie der Mittelwert nicht robust.
Im Streudiagramm ist noch die Winkelhalbierende eingezeichnet. Auf dieser liegen alle Studierenden, die in beiden Prüfungen die gleiche Note erzielt haben. Unterhalb der Winkelhalbierenden liegen alle Studierenden, die im Vordiplom besser als im Abitur sind. Entsprechend liegen oberhalb der Winkelhalbierenden alle mit einer besseren Note im Abitur als im Vordiplom. Wir sehen, dass der Anteil derjenigen mit besserer Note im Abitur höher als derjenigen mit besserer Note im Vordiplom ist.

□

Der Rangkorrelationskoeffizient von Spearman

Der Korrelationskoeffizient von Bravais-Pearson ist ein Maß für den linearen Zusammenhang zwischen X und Y. Ein **monotoner Zusammenhang** zwischen X und Y liegt vor, wenn für zwei beliebige Punkte (x_i, y_i) und (x_j, y_j) entweder

$$x_i < x_j \Longleftrightarrow y_i < y_j$$

oder

$$x_i < x_j \Longleftrightarrow y_i > y_j$$

gilt.
Eine Maßzahl für einen monotonen Zusammenhang erhält man, indem man die Beobachtungen jedes Merkmals durch die Ränge ersetzt. Dabei gibt der `Rang` r_i der Beobachtung x_i an, an der wievielten Stelle x_i in der geordneten Stichprobe steht. Wir können auch zählen, wie viele der Beobachtungen kleiner oder gleich x_i sind.

Beispiel 39 (fortgesetzt von Seite 172)
Wir betrachten die Daten aus Tabelle 4.12 auf Seite 164. Schauen wir uns zunächst die Note im Abitur an. Es gilt

$$x_1 = 1.7 \quad x_2 = 2.4 \quad x_3 = 2 \quad x_4 = 1.1 \quad x_5 = 2.9 \quad x_6 = 3.1$$

Die kleinste Beobachtung ist $x_4 = 1.1$. Also gilt $r_4 = 1$. Die zweitkleinste Beobachtung ist $x_1 = 1.7$. Also gilt $r_1 = 2$. Entsprechend erhalten wir

$$r_2 = 4 \quad r_3 = 3 \quad r_5 = 5 \quad r_6 = 6\,.$$

Auch für die y_i können wir die Ränge bestimmen. Wir bezeichnen sie mit s_i und erhalten:

$$s_1 = 3 \quad s_2 = 4 \quad s_3 = 2 \quad s_4 = 1 \quad s_5 = 6 \qquad s_6 = 5\,.$$

□

Sind Beobachtungen identisch, so weisen wir diesen **Durchschnittsränge** zu.

Beispiel 43
Die Daten seien

$$x_1 = 40 \quad x_2 = 40 \quad x_3 = 31 \quad x_4 = 23 \quad x_5 = 31 \quad x_6 = 40$$

Der kleinste Wert ist 23. Dieser tritt nur einmal auf. Also gilt $r_4 = 1$. Der zweitkleinste Wert ist 31. Dieser tritt zweimal auf. Wir würden die Ränge 2 und 3 vergeben. Wir bilden den Durschnittsrang. Also gilt $r_3 = 2.5$ und $r_5 = 2.5$. Der Wert 40 tritt dreimal auf. Wir würden die Ränge 4, 5 und 6 vergeben. Wir bilden auch hier den Durschnittsrang. Also gilt $r_1 = 5$, $r_2 = 5$ und $r_6 = 5$.

□

Ersetzen wir in Gleichung (4.23) auf Seite 172 x_i durch r_i und y_i durch s_i, so erhalten wir den Rangkorrelationskoeffizienten von Spearman:

$$\boxed{r_S = \frac{\sum_{i=1}^{n} (r_i - \overline{r}) \cdot (s_i - \overline{s})}{\sqrt{\sum_{i=1}^{n} (r_i - \overline{r})^2 \cdot \sum_{i=1}^{n} (s_i - \overline{s})^2}}} \tag{4.24}$$

Liegen keine Bindungen vor, so können wir den Rangkorrelationskoeffizienten nach folgender Formel bestimmen:

$$\boxed{r_S = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)}} \tag{4.25}$$

Dabei gilt $d_i = r_i - s_i$. Der Beweis von Beziehung (4.25) ist bei Burkschat et al. (2004) auf Seite 283 zu finden.

Beispiel 39 (fortgesetzt von Seite 175)
Wir betrachten die Daten aus Tabelle 4.12 auf Seite 164 und stellen die Ränge in einer Tabelle zusammen und bestimmen die Differenzen $d_i = r_i - s_i$.

Tabelle 4.16: Hilfstabelle zur Berechnung des Korrelationskoeffizienten von Spearman

Wohnung	1	2	3	4	5	6
r_i	2	4	3	1	5	6
s_i	3	4	2	1	6	5
d_i	-1	0	1	0	-1	1

Es gilt

$$r_S = 1 - \frac{6 \cdot \left[(-1)^2 + 0^2 + 1^2 + 0^2 + (-1)^2 + 1^2\right]}{6 \cdot (6^2 - 1)} = 0.89$$

□

Besteht ein streng monoton wachsender Zusammenhang zwischen den beiden Merkmalen, so gilt $r_i = s_i$ für $i = 1, \ldots, n$ und somit $d_i = 0$. Setzen wir in Gleichung (4.25) für $i = 1, \ldots, n$ $d_i = 0$, so nimmt r_S den Wert 1 an. Ist hingegen $r_S = 1$, so muss gelten

$$\sum_{i=1}^{n} d_i^2$$

Eine Summe nichtnegativer Summanden ist nur dann 0, wenn alle Summanden gleich 0 sind. Also muss für $i = 1, \ldots, n$ $d_i = 0$ gelten. Also gilt $r_i = s_i$ und der Zusammenhang ist monoton.
Ähnliche Überlegungen zeigen, dass r_S genau dann gleich -1 ist, wenn ein streng monoton fallender Zusammenhang vorliegt.
Sehr oft liegen die Werte von zwei Merkmalsträgern bereits als Ränge vor. In diesem Fall kann man mit dem Rangkorrelationskoeffizienten überprüfen, wie sehr die beiden Merkmalsträger in ihrer Bewertung übereinstimmen.

Beispiel 44
Zwei Personen werden gebeten, sechs Paare von Politikern der Ähnlichkeit nach zu ordnen. Dem Paar, bei dem sich die beiden Politiker am ähnlichsten

sind, sollten sie eine 1, dem zweitähnlichsten eine 2,... u.s.w. geben. Die Werte sind in Tabelle 4.17 zu finden.

Tabelle 4.17: Bewertung der Ähnlichkeit von Politikerpaaren durch zwei Personen

Politikerpaar	Person 1	Person 2
Schröder - Fischer	4	1
Schröder - Schäuble	1	3
Schröder - Westerwelle	2	5
Fischer - Schäuble	5	6
Fischer - Westerwelle	6	2
Schäuble - Westerwelle	3	4

Es gilt

$$\begin{aligned} d_1 = 3 \quad & d_2 = -2 \\ d_3 = -3 \quad & d_4 = -1 \\ d_5 = 4 \quad & d_6 = -1 \end{aligned}$$

Also gilt

$$\sum_{i=1}^{n} d_i^2 = 40$$

und es folgt

$$r_S = 1 - \frac{6 \cdot 40}{6 \cdot (6^2 - 1)} = -0.143$$

Wir sehen, dass kein Zusammenhang zwischen den beiden Bewertungen besteht.

□

4.3.3 Zur Interpretation von Korrelation

Ist die Korrelation zwischen zwei Merkmalen groß, so wird oft unterstellt, dass die eine Größe die andere beeinflusst. Bei einer derartigen Interpretation muss man sehr vorsichtig sein. So hängt die Höhe der Geburtenrate sicherlich nicht von der Anzahl der Störche ab, die in einer Region leben.

Beide Merkmale sind aber positiv miteinander korreliert. Diese Korrelation wird aber durch eine dritte Größe bewirkt. Diese ist der Grad der Industralisierung in einem Land. Je höher die Industrialisierung, um so niedriger die Geburtenrate und die Anzahl der Störche. Schauen wir uns noch ein Beispiel an.

Beispiel 45
Bei einer Befragung von Erstsemestern wurden unter anderem die Merkmale `Körpergröße` x, `Körpergewicht` y und `Schuhgröße` z erhoben. Die Werte von 20 Studenten sind in Tabelle 4.18 zu finden.

Tabelle 4.18: Körpergröße, Körpergewicht und Schuhgröße von 20 Studenten

Student i	x_i	y_i	z_i	Student i	x_i	y_i	z_i
1	171	58	40	11	201	93	48
2	180	80	44	12	180	67	42
3	178	80	42	13	183	73	42
4	171	60	41	14	176	65	42
5	182	73	44	15	170	65	41
6	180	70	41	16	182	85	40
7	180	77	43	17	180	80	41
8	170	55	42	18	190	83	44
9	163	50	37	19	180	67	39
10	169	51	38	20	183	75	45

Wir bestimmen die empirische Korrelationsmatrix

$$\mathbf{R} = \begin{pmatrix} 1.000 & 0.882 & 0.796 \\ 0.882 & 1.000 & 0.712 \\ 0.796 & 0.712 & 1.000 \end{pmatrix}. \tag{4.26}$$

□

Zwischen allen Merkmalen in Beispiel 45 besteht eine hohe positive Korrelation. Bei der Korrelation zwischen den Merkmalen `Körpergröße` und `Körpergewicht` wundert uns das nicht. Je größer eine Person ist, umso mehr wird sie auch wiegen. Die starke positive Korrelation zwischen den Merkmalen `Körpergröße` und `Schuhgröße` haben wir auch erwartet. Dass aber die Merkmale `Körpergewicht` und `Schuhgröße` eine starke positive Korrelation aufweisen, ist verwunderlich. Warum sollten schwerere Personen größere

Füße haben? Wir hätten hier eher einen Wert des empirischen Korrelationskoeffizienten in der Nähe von 0 erwartet. Woher kommt dieser hohe positive Wert? Der Zusammenhang zwischen den Merkmalen `Körpergewicht` und `Schuhgröße` kann am Merkmal `Körpergröße` liegen, denn das Merkmal `Körpergröße` bedingt im Regelfall sowohl das Merkmal `Körpergewicht` als auch das Merkmal `Schuhgröße`. Um zu überprüfen, ob das Merkmal `Körpergröße` den Zusammenhang zwischen den Merkmalen `Körpergewicht` und `Schuhgröße` bedingt, müssen wir es kontrollieren. Hierzu haben wir zwei Möglichkeiten:

- Wir betrachten nur Personen, die die gleiche Ausprägung des Merkmals `Körpergröße` besitzen, und bestimmen bei diesen den Zusammenhang zwischen den Merkmalen `Körpergewicht` und `Schuhgröße`. Besteht bei Personen, die die gleiche Ausprägung des Merkmals `Körpergröße` besitzen, kein Zusammenhang zwischen den Merkmalen `Körpergewicht` und `Schuhgröße`, so sollte der Wert des empirischen Korrelationskoeffizienten gleich 0 sein.
- Wir können den Effekt des Merkmals `Körpergröße` auf die Merkmale `Körpergewicht` und `Schuhgröße` statistisch bereinigen und den Zusammenhang zwischen den bereinigten Merkmalen bestimmen.

Bereinigt man die die Korrelation zwischen den Merkmalen Y und Z um den Effekt des Merkmals X, so erhält man **partiellen Korrelationskoeffizienten** $r_{YZ.X}$. Dieser ist folgendermaßen definiert:

$$\boxed{r_{YZ.X} = \frac{r_{YZ} - r_{XY} \cdot r_{XZ}}{\sqrt{(1 - r_{XY}^2) \cdot (1 - r_{XZ}^2)}}} \tag{4.27}$$

Dabei ist r_{YZ} der Korrelationskoeffizient zwischen Y und Z, r_{XY} der Korrelationskoeffizient zwischen X und Y und r_{XZ} der Korrelationskoeffizient zwischen X und Z. Ist der Wert von $r_{YZ.X}$ in der Nähe von 0, so deutet dies darauf hin, dass die Korrelation zwischen Y und Z gleich 0 ist, wenn man beide um den linearen Effekt von X bereinigt.

Beispiel 45 (fortgesetzt)
Mit $r_{XY} = 0.882$, $r_{XZ} = 0.796$ und $r_{YZ} = 0.712$ gilt

$$r_{YZ.X} = \frac{r_{YZ} - r_{XY} \cdot r_{XZ}}{\sqrt{(1 - r_{XY}^2) \cdot (1 - r_{XZ}^2)}} = \frac{0.712 - 0.882 \cdot 0.796}{\sqrt{(1 - 0.882^2) \cdot (1 - 0.796^2)}} = 0.035$$

Diie partielle Korrelation zwischen dem Körpergewicht und der Schuhgröße ist also ungefähr gleich 0. Zwischen dem Körpergewicht und der Schuhgröße

besteht also keine Korrelation, wenn man beide um den linearen Effekt der Körpergrößbereinigt.

□

4.3.4 Die Analyse in R

Schauen wir uns den Zusammenhang zwischen quantitativen Merkmalen an. Wir betrachten die Daten in Tabelle 4.12 auf Seite 164.
Wir geben die Variablen `Abitur` und `Vordiplom` ein:

```
> Abitur<-c(1.7,2.4,2,1.1,2.9,3.1)
> Vordiplom<-c(2.2,2.4,2.1,1.8,2.7,2.6)
```

Das Streudiagramm in Abbildung 4.10 auf Seite 164 erhält man mit der Funktion `plot`. Diese wird detailliert in Kapitel 2.5 beschrieben. In der einfachsten Variante benötigt sie die Vektoren `x` und `y` und zeichnet die Punkte $(x[1], y[1]), \ldots, (x[n], y[n])$, wobei n die Länge der Vektoren `x` und `y` ist. Das Streudiagramm der Noten im Abitur und Vordiplom erhält man durch

```
> plot(Abitur,Vordiplom)
```

Die Abszisse wird mit dem Namen des ersten Vektors und die Ordinate mit dem Namen des zweiten Vektors beschriftet. Will man eine eigene Beschriftung wählen, so muss man die Argumente `xlab` und `ylab` angeben.

```
> plot(Abitur,Vordiplom,xlab="Note im Abitur",
       ylab="Note im Vordiplom")
```

Sollen die Punkte ausgemalt werden, so setzt man den Parameter `pch` auf den Wert 16.

```
> plot(Abitur,Vordiplom,pch=16)
```

Sollen nur die Abszisse und die Ordinate gezeichnet werden, so setzt man den Parameter `bty` auf den Wert `l`.

```
> plot(Abitur,Vordiplom,pch=16,bty="l")
```

Durch die Argumente `xlim` und `ylim` kann man den Bereich der Grafik auf der Abszisse und Ordinate angeben. Sollen beide Achsen im Bereich von 1 bis 4 gezeichnet werden, so gibt man ein

```
> plot(Abitur,Vordiplom,xlim=c(1,4),ylim=c(1,4),bty="l")
```

Hierbei enden die Abszisse und die Ordinate nicht exakt bei der 1 und 4, sondern davor und dahinter wird noch ein wenig Luft gelassen. Sollen sie genau bei der 1 und 4 enden, so setzt man die Argumente `xaxs` und `yaxs` jeweils auf den Wert `"i"`.

```
> plot(Abitur,Vordiplom,xlim=c(1,4),ylim=c(1,4),bty="l",pch=16,
      xaxs="i",yaxs="i")
```

abline

Wir können zum Streudiagramm mit der Funktion `abline` eine Gerade hinzufügen. Das Argument `a` ist der Achsenabschnitt und das Argument `b` die Steigung. Die Winkelhalbierende erhalten wir also durch

```
> abline(a=0,b=1)
```

Um die Abbildung 4.13 auf Seite 168 mit den 4 Quadranten zu erhalten, verwenden wir wiederum die Funktion `abline`. Mit dem Argument `h` zeichnen wir eine Parallele zur Abszisse und mit dem Argument `v` eine Parallele zur Ordinate.

```
> plot(Abitur,Vordiplom,xlim=c(1,4),ylim=c(1,4),bty="l",pch=16,
      xaxs="i",yaxs="i")
> abline(h=mean(Vordiplom))
> abline(v=mean(Abitur))
```

text

Mit der Funktion `text` können wir Text zu einem Streudiagramm hinzufügen. Der Aufruf

```
> text(2.8,3.1,"I",cex=1.8)
```

schreibt im Punkt $(2.8, 3.1)$ die Nummer `I` des ersten Quadranten. Das Argument `cex` steuert die Schriftgröße, wobei der Wert 1.8 bedeutet, dass das 1.8-fache der normalen Schriftgröße gewählt wird.

var

Den Wert der empirischen Kovarianz zwischen den Merkmalen `Abitur` und `Vordiplom` liefert die Funktion `var`:

```
> var(Abitur,Vordiplom)
[1] 0.242
```

`R` verwendet bei der Bestimmung der empirischen Kovarianz folgende Formel:

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

Dies können wir folgendermaßen nachprüfen:

```
> e<-sum((Abitur-mean(Abitur))*(Vordiplom-mean(Vordiplom)))
> e
[1] 1.21
> e/(length(Abitur)-1)
[1] 0.242
```

Den Wert des Korrelationskoeffizienten von Bravais-Pearson zwischen den Merkmalen `Abitur` und `Vordiplom` liefert die Funktion `cor`: `cor`

```
> cor(Abitur,Vordiplom)
[1] 0.9594717
```

Um den Korrelationskoeffizienten von Spearman für die Merkmale `Abitur` und `Vordiplom` bestimmen zu können, benötigt man zuerst die Ränge der Beobachtungen bei beiden Merkmalen. Diese erhält man mit der Funktion `rank`. `rank`

```
> rank(Abitur)
[1] 2 4 3 1 5 6

> rank(Vordiplom)
[1] 3 4 2 1 6 5
```

Wendet man die Funktion `cor` auf die Ränge an, so erhält man den Korrelationskoeffizienten von Spearman:

```
> cor(rank(Abitur),rank(Vordiplom))
[1] 0.8857143
```

Den partiellen Korrelationskoeffizienten können wir mit der Funktion `pcor` aus dem Paket `ggm` bestimmen, das man installieren und laden muss. Wie man dabei vorzugehen hat, wird auf Seite 52 beschrieben. `pcor` `ggm`

Mit folgender Befehlsfolge berechnen wir die partielle Korrelation zwischen der Schuhgröße und dem Gewicht gegeben die Körpergröße.

```
> gewicht<-c(58,80,80,60,73,70,77,55,50,51,93,67,73,65,65,85,
             80,83,67,75)
> groesse<-c(171,180,178,171,182,180,180,170,163,169,201,180,
             183,176,170,182,180,190,180,183)
> schuh<-c(40,44,42,41,44,41,43,42,37,38,48,42,42,42,41,40,41,
           44,39,45)
> pcor(c("schuh","gewicht","groesse"),var(cbind(gewicht,
       groesse,schuh)))
[1] 0.03630280
```

4.4 Beweise

Beweis von Gleichung (4.14) **auf Seite 154**

Wir benötigen zwei Beziehungen für den Beweis von G Für $\tilde{n}_{ij} = \frac{n_{i.}\, n_{.j}}{n}$ gilt

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \tilde{n}_{ij} = n \tag{4.28}$$

Dies sieht man folgendermaßen:

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \tilde{n}_{ij} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{n_{i.}\, n_{.j}}{n} = \frac{1}{n} \sum_{i=1}^{I} n_{i.} \sum_{j=1}^{J} n_{.j} = \frac{n \cdot n}{n} = n$$

Es gilt

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{n_{ij}^2}{\tilde{n}_{ij}} - n \tag{4.29}$$

Dies sieht man folgendermaßen:

$$\begin{aligned}
X^2 &= \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \tilde{n}_{ij})^2}{\tilde{n}_{ij}} \\
&= \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{n_{ij}^2 - 2\, n_{ij}\, \tilde{n}_{ij} + \tilde{n}_{ij}^2}{\tilde{n}_{ij}} \\
&= \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{n_{ij}^2}{\tilde{n}_{ij}} - 2 \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{n_{ij}\, \tilde{n}_{ij}}{\tilde{n}_{ij}} + \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\tilde{n}_{ij}^2}{\tilde{n}_{ij}} \\
&= \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{n_{ij}^2}{\tilde{n}_{ij}} - 2 \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} + \sum_{i=1}^{I} \sum_{j=1}^{J} \tilde{n}_{ij} \\
&\overset{(4.28)}{=} \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{n_{ij}^2}{\tilde{n}_{ij}} - 2 \cdot n + n \\
&= \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{n_{ij}^2}{\tilde{n}_{ij}} - n
\end{aligned}$$

Nun können wir Gleichung (4.14) auf Seite 154 beweisen.

Aus

$$n_{ij} \leq n_{i\cdot}$$

folgt

$$\frac{n_{ij}}{n_{i\cdot}} \leq 1$$

Wir multiplizieren beide Seiten der Ungleichung mit $\frac{n_{ij}}{n_{\cdot j}}$ und summieren über i und j:

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{n_{ij}^2}{n_{i\cdot}\,n_{\cdot j}} \;\leq\; \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{n_{ij}}{n_{\cdot j}}$$

Es gilt

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{n_{ij}}{n_{\cdot j}} = \sum_{j=1}^{J}\frac{1}{n_{\cdot j}}\sum_{i=1}^{I}n_{ij} = \sum_{j=1}^{J}\frac{1}{n_{\cdot j}}\,n_{\cdot j} = \sum_{j=1}^{J}1 = J$$

Also gilt

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{n_{ij}^2}{n_{i\cdot}\,n_{\cdot j}} \;\leq\; J$$

Subtrahieren wir von beiden Seiten dieser Ungleichung 1 und multiplizieren anschließend beide Seiten mit n, so erhalten wir

$$n\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{n_{ij}^2}{n_{i\cdot}\,n_{\cdot j}} - n \;\leq\; n(J-1)$$

Für die linke Seite dieser Ungleichung gilt

$$n\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{n_{ij}^2}{n_{i\cdot}\,n_{\cdot j}} - n = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{n_{ij}^2}{\frac{n_{i\cdot}\,n_{\cdot j}}{n}} - n \overset{(4.11)}{=} \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{n_{ij}^2}{\tilde{n}_{ij}} - n \overset{(4.29)}{=} X^2$$

Also gilt

$$X^2 \leq n(J-1)$$

Starten wir mit

$$n_{ij} \leq n_{\cdot j}$$

so zeigen analoge Berechnungen

$$X^2 \leq n(I-1)$$

Also gilt

$$X^2 \leq n \cdot \min\{I-1, J-1\}$$

Beweis von Gleichung (4.16) **auf Seite 155**

Es gilt

$$
\begin{aligned}
X^2 \leq n \cdot \min\{I-1, J-1\} &\iff \frac{n}{X^2} \geq \frac{1}{\min\{I-1, J-1\}} \\
&\iff \frac{n}{X^2} + 1 \geq \frac{1}{\min\{I-1, J-1\}} + 1 \\
&\iff \frac{X^2+n}{X^2} \geq \frac{\min\{I-1, J-1\}+1}{\min\{I-1, J-1\}} \\
&\iff \frac{X^2+n}{X^2} \geq \frac{\min\{I, J\}}{\min\{I-1, J-1\}} \\
&\iff \frac{X^2}{X^2+n} \leq \frac{\min\{I-1, J-1\}}{\min\{I, J\}} \\
&\iff \sqrt{\frac{X^2}{X^2+n}} \leq \sqrt{\frac{\min\{I-1, J-1\}}{\min\{I, J\}}} \\
&\iff K \leq \sqrt{\frac{\min\{I-1, J-1\}}{\min\{I, J\}}}
\end{aligned}
$$

Beweis von Gleichung (4.22) **auf Seite 171**

Es gilt

$$
\begin{aligned}
d_{x,y} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i\, y_i - x_i\, \overline{y} - \overline{x}\, y_i + \overline{x}\, \overline{y}) \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i\, y_i - \frac{1}{n}\sum_{i=1}^{n} x_i\, \overline{y} - \frac{1}{n}\sum_{i=1}^{n} \overline{x}\, y_i + \frac{1}{n}\sum_{i=1}^{n} \overline{x}\, \overline{y} \\
&= \overline{x\, y} - \overline{y}\,\frac{1}{n}\sum_{i=1}^{n} x_i - \overline{x}\,\frac{1}{n}\sum_{i=1}^{n} y_i + \frac{1}{n}\, n\, \overline{x}\, \overline{y} \\
&= \overline{x\, y} - \overline{y}\, \overline{x} - \overline{x}\, \overline{y} + \overline{x}\, \overline{y} = \overline{x\, y} - \overline{x}\, \overline{y}
\end{aligned}
$$

Kapitel 5

Wahrscheinlichkeitsrechnung

5.1 Zufallsvorgänge und Ereignisse

Der Ausgang der meisten Vorgänge ist unbestimmt. So sind die Brötchen beim morgendlichen Besuch des Bäckerladens manchmal knusperig, manchmal aber leider weich. Manchmal wird man sofort bedient, manchmal muss man warten. Wenn man warten muss, ist die Anzahl der Personen, die vor einem steht, nicht immer gleich. Auch die Wartezeit variiert.

Die betrachteten Vorgänge zeichnen sich dadurch aus, dass sie mit einem von mehreren Ergebnissen enden können, wir aber vor der Durchführung nicht wissen, welches der Ergebnisse eintritt. Wir bezeichnen solche Vorgänge als Zufallsvorgänge.

Die Menge aller möglichen Ergebnisse des Zufallsvorgangs nennt man **Ergebnismenge** Ω.

Beispiel 46
Eine Münze wird einmal geworfen, wobei entweder KOPF K oder ZAHL Z auftritt. Die Ergebnismenge ist somit

$$\Omega = \{K, Z\}\,.$$

□

Beispiel 47
Eine Münze wird zweimal geworfen, wobei bei einem Wurf entweder KOPF oder ZAHL auftritt. Die Ergebnismenge ist somit

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

□

Beispiel 48
Ein Würfel wird einmal geworfen. Die Ergebnismenge ist

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

□

Beispiel 49
Eine Münze wird so oft geworfen, bis zum ersten Mal KOPF auftritt. Die Ergebnismenge ist

$$\Omega = \{K, ZK, ZZK, ZZZK, \ldots\}.$$

□

Beispiel 50
Ein Kunde zählt die Anzahl der Kunden, die vor ihm im Bäckerladen stehen. Die Ergebnismenge ist

$$\Omega = \{0, 1, 2, 3, \ldots\}.$$

□

Beispiel 51
Wir bestimmen die Wartezeit im Bäckerladen, wobei ein Kunde nicht bereit ist, länger als 10 Minuten zu warten. Die Ergebnismenge ist

$$\Omega = \{x | x \in \Re, 0 \leq x \leq 10\} = [0, 10].$$

□

Die Ergebnismengen der ersten drei Beipiele enthalten endlich viele Ergebnisse, die restlichen unendlich viele. Die unendlich vielen Ergebnisse der Beispiele 49 und 50 können abgezählt werden, während dies bei der Ergebnismenge des Beispiels 51 nicht möglich ist.

Ergebnismengen heißen **diskret**, wenn sie endlich viele oder abzählbar unendlich viele Ergbenisse enthalten. Ansonsten heißen sie **stetig**.

Bei Zufallsvorgängen ist man oft an Zusammenfassungen von Ergebnissen interessiert.

Beispiel 47 (fortgesetzt)
Beim zweimaligen Wurf einer Münze ist man daran interessiert, dass genau einmal Kopf eintritt. □

Beispiel 48 (fortgesetzt)
Beim einmaligen Wurf eines Würfels ist von Interesse, ob eine ungerade Augenzahl fällt. □

Eine Zusammenfassung von Ergebnissen aus Ω heißt **Ereignis** A. Man sagt, dass das Ereignis A eingetreten ist, wenn ein Ergebnis ω aus A beobachtet wurde.

Beispiel 47 (fortgesetzt)
Das zugehörige Ereignis ist $A = \{KZ, ZK\}$. □

Beispiel 48 (fortgesetzt)
Das zugehörige Ereignis ist $A = \{1, 3, 5\}$. □

Die Ergebnismenge Ω heißt **sicheres Ereignis**, da sie immer eintritt. Die leere Menge $\emptyset$ heißt **unmögliches Ereignis**, da sie niemals eintritt. Die einelementigen Ereignisse heißen **Elementarereignisse**.

Beispiel 48 (fortgesetzt)
Es gibt die Elementarereignisse

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

□

5.1.1 Operationen zwischen Ereignissen

Da Ereignisse Mengen sind, können wir mit ihnen die üblichen Mengenoperationen durchführen.

Beispiel 48 (fortgesetzt)
Wir betrachten im Folgenden die Ereignisse:

$$A = \{1, 2, 3\} \qquad B = \{1, 3, 5\} \qquad C = \{6\}$$

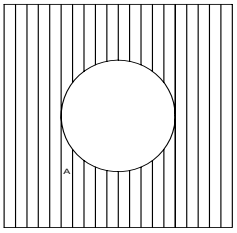
□

Definition 5.1
Das **Komplementärereignis** $\overline{A}$ tritt ein, wenn das Ereignis A nicht eintritt:

$$\boxed{\overline{A} = \{x | x \in \Omega, x \notin A\}\,.} \tag{5.1}$$

Man sagt auch, dass A nicht eintritt. Abbildung 5.1 veranschaulicht das Komplementärereignis. Dabei ist der schraffierte Teil gleich $\overline{A}$.

Abbildung 5.1: Das Komplementärereignis im Venn-Diagramm



Beispiel 48 (fortgesetzt von Seite 189)
Mit $A = \{1, 2, 3\}$, $B = \{1, 3, 5\}$ und $C = \{6\}$ gilt:

$$\overline{A} = \{4, 5, 6\} \qquad \overline{B} = \{2, 4, 6\} \qquad \overline{C} = \{1, 2, 3, 4, 5\}$$

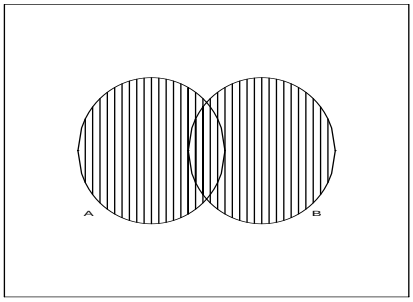
□

Definition 5.2
Sind A und B Ereignisse, dann ist die **Vereinigung** $A \cup B$ der beiden definiert durch

$$A \cup B = \{\omega | \omega \in \Omega, \omega \in A \text{ oder } \omega \in B\}\,. \tag{5.2}$$

Man sagt auch, dass mindestens eines der beiden Ereignisse eintritt. Abbildung 5.2 veranschaulicht die Vereinigung der Ereignisse A und B. Dabei ist der schraffierte Teil gleich $A \cup B$.

Abbildung 5.2: Die Vereinigung der Ereignisse A und B im Venn-Diagramm



Beispiel 48 (fortgesetzt von Seite 190)
Mit $A = \{1, 2, 3\}$, $B = \{1, 3, 5\}$ und $C = \{6\}$ gilt:

$$A \cup B = \{1, 2, 3, 5\} \qquad A \cup C = \{1, 2, 3, 6\} \qquad B \cup C = \{1, 3, 5, 6\}$$

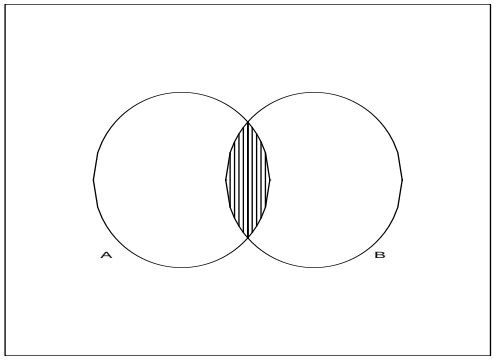
□

Definition 5.3
Sind A und B Ereignisse, dann ist der **Durchschnitt** $A \cap B$ der beiden definiert durch

$$\boxed{A \cap B = \{\omega | \omega \in \Omega, \omega \in A \text{ und } \omega \in B\}\,.} \tag{5.3}$$

Man sagt auch, dass beide Ereignisse gleichzeitig eintreten. Abbildung 5.3 veranschaulicht den Durchschnitt der Ereignisse A und B. Dabei ist der schraffierte Teil gleich $A \cap B$.

Abbildung 5.3: Der Durchschnitt der Ereignisse A und B im Venn-Diagramm



Beispiel 48 (fortgesetzt)
Mit $A = \{1, 2, 3\}$, $B = \{1, 3, 5\}$ und $C = \{6\}$ gilt:

$$A \cap B = \{1, 3\} \qquad A \cap C = \emptyset \qquad B \cap C = \emptyset$$

□

Definition 5.4
Gilt

$$\boxed{A \cap B = \emptyset} \tag{5.4}$$

für zwei Ereignisse A und B, dann heißen A und B **disjunkt** oder **unvereinbar**.

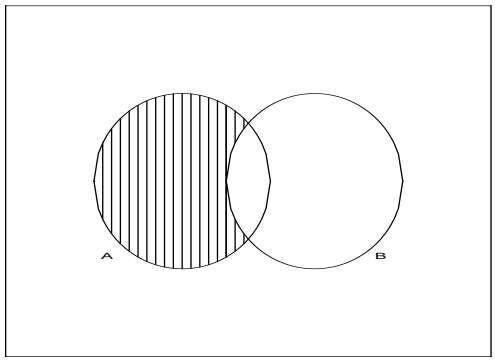
Sind die Ereignisse A und B disjunkt, dann können sie nicht gleichzeitig eintreten.

Definition 5.5
Sind A und B Ereignisse, dann ist die **Differenz** A/B der beiden definiert durch

$$\boxed{A/B = \{\omega | \omega \in \Omega, \omega \in A \text{ und } \omega \notin B\}\,.} \tag{5.5}$$

Man sagt auch, dass nur A eintritt. Abbildung 5.4 veranschaulicht die Differenz der Ereignisse A und B. Dabei ist der schraffierte Teil gleich A/B.

Abbildung 5.4: Die Differenz der Ereignisse A und B im Venn-Diagramm



Beispiel 48 (fortgesetzt von Seite 191)
Mit $A = \{1, 2, 3\}$, $B = \{1, 3, 5\}$ und $C = \{6\}$ gilt:

$$\begin{aligned} A/B &= \{2\} & B/A &= \{5\} \\ A/C &= \{1, 2, 3\} & C/A &= \{6\} \\ B/C &= \{1, 3, 5\} & C/B &= \{6\} \end{aligned}$$

□

Für zwei Ereignisse A und B gelten die de Morganschen Regeln

$$\boxed{\overline{A \cap B} = \overline{A} \cup \overline{B}} \tag{5.6}$$

und

$$\boxed{\overline{A \cup B} = \overline{A} \cap \overline{B}\,.} \tag{5.7}$$

Für drei Ereignisse A, B und C gelten die Distributivgesetze

$$\boxed{A \cap (B \cup C) = (A \cap B) \cup (A \cap C)} \tag{5.8}$$

und

$$\boxed{A \cup (B \cap C) = (A \cup B) \cap (A \cup C)\,.} \tag{5.9}$$

5.2 Wahrscheinlichkeit

Bisher haben wir die bei Zufallsvorgängen interessierenden Ereignisse mit Hilfe von Mengen mathematisch beschrieben. Im täglichen Leben bewerten wir Ereignisse hinsichtlich ihrer Realisierungsmöglichkeit unterschiedlich. Wie können wir dies quantifizieren? Die Wahrscheinlichkeitsrechnung nimmt an, dass diese Zahlen irgendwie gegeben sind und bestimmten trivialen Konsistenzbedingungen (Axiomen) genügen. Es ist der Wahrscheinlichkeitstheorie jedoch völlig gleichgültig, woher man in einem konkreten Fall die Zahlen bekommt. Intuitiv bedeuten Wahrscheinlichkeiten Grade der Realisierungschance von Ereignissen oder Grade des subjektiven Glaubens an Ereignisse. Wir wollen zunächst einige Ansätze zur Formalisierung der intuitiven Vorstellung angeben.

5.2.1 Klassischer Ansatz

Ausgangspunkt ist ein Zufallsvorgang mit endlich vielen Ergebnissen. Die Wahrscheinlichkeit eines Ereignisses A ist definiert durch

$$\boxed{P(A) = \frac{\text{Anzahl der Ergebnisse in A}}{\text{Anzahl der Ergebnisse in}\Omega}\,.} \tag{5.10}$$

Beispiel 49
Wurf eines Würfels. Wir unterstellen, dass der Würfel fair ist. Es gilt also für die Elementarereignisse

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

für $i = 1, 2, 3, 4, 5, 6$.
Für $A = \{1, 2, 3\}$ gilt

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

Für $B = \{1, 3, 5\}$ gilt

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

□

Den klassischen Ansatz kann man nur verwenden, wenn die Ergebnismenge endlich ist und alle Elementarereignisse gleichwahrscheinlich sind.

5.2.2 Frequentistischer Ansatz

Eine Möglichkeit, zu einer Bewertung eines Ereignisses im Rahmen eines Zufallsvorgangs zu gelangen, besteht darin, den Zufallsvorgang mehrmals unter identischen Bedingungen zu beobachten und zu zählen, wie oft das interessierende Ereignis eingetreten ist. Man sammelt also Erfahrungen über die Realisierungsmöglichkeiten eines Ereignisses durch Beobachten und Zählen. Dies ist der klassische Ansatz.
Das Ereignis A sei bei den N Durchführungen des Zufallsvorgangs $n_N(A)$-mal eingetreten. Wir nennen $n_N(A)$ die absolute Häufigkeit des Ereignisses A. Setzen wir $n_N(A)$ in Beziehung zur Anzahl N der Durchführungen, so erhalten wir die relative Häufigkeit

$$h_N(A) = \frac{n_N(A)}{N}\,.$$

Da absolute Häufigkeiten nichtnegativ sind, gilt

$$h_N(A) \geq 0 \quad \text{für jedes Ereignis } A\,. \tag{5.11}$$

Da das sichere Ereignis Ω bei jeder Durchführung des Zufallsvorgangs eintritt, gilt

$$h_N(\Omega) = 1\,. \tag{5.12}$$

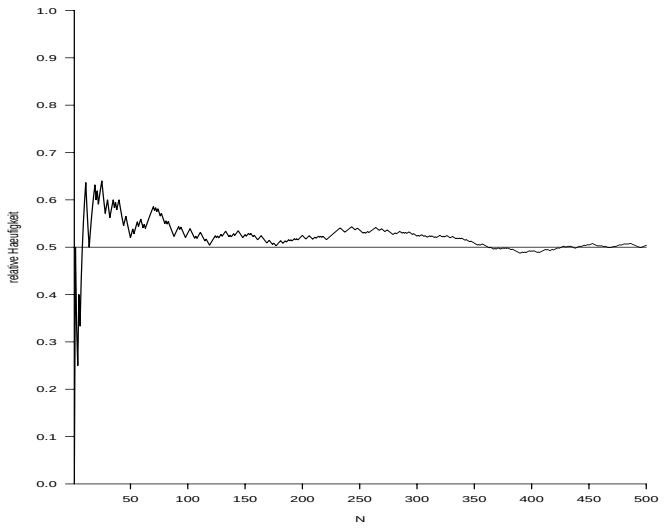
Sind A und B disjunkte Ereignisse, so kann man bei jeder Durchführung des Zufallsvorgangs nur eines der beiden Ereignisse eintreten. Wir können

die absolute Häufigkeit von $A \cup B$ also dadurch bestimmen, dass wir die absolute Häufigkeit von A und die absolute Häufigkeit von B addieren. Es gilt also

$$h_N(A \cup B) = h_N(A) + h_N(B) \quad \text{für disjunkte Ereignisse } A \text{ und } B\,. \tag{5.13}$$

Wiederholen wir einen Zufallsvorgang immer wieder unter den gleichen Bedingungen und bilden für ein beliebiges Ereignis A die Folge der relativen Häufigkeiten $h_N(A)$, $N = 1, 2, \ldots$, dann schwanken die $h_N(A)$ mit wachsendem n immer weniger und scheinen einem Grenzwert zuzustreben. Die Folge der relativen Häufigkeiten zeigt ein konvergenzartiges Verhalten. Abbildung 5.5 zeigt das Verhalten der relativen Häufigkeit von KOPF bei 300 Würfen einer Münze.

Abbildung 5.5: Relativen Häufigkeit von KOPF bei 300 Würfen einer Münze



Es scheint so, als ob gilt

$$\lim_{N\to\infty} h_N(A) = p\,,$$

wobei p eine reelle Zahl ist. Es liegt nun nahe, die Chance der Realisierung von Ereignissen über diesen Grenzwert zu definieren. Die Konvergenz lässt sich aber in der Realität weder verifizieren noch falsifizieren, da wir nur ein Anfangsstück der Reihe beobachten können.

Eine Bewertung der Chance des Eintretens von Ereignissen über den Grenzwert der Folge der relativen Häufigkeiten ist also nicht möglich, da über die Existenz dieses Grenzwerts nichts bekannt ist.

5.2.3 Axiomatische Definition

Bei vielen Zufallsvorgängen ist zu beobachten, dass die relative Häufigkeit eines Ereignisses mit wachsender Zahl der Wiederholungen immer weniger schwankt. Es liegt also nahe, den Grad der Unbestimmtheit des Eintretens eines Ereignisses durch eine Zahl zu charakterisieren. Dabei sollte eine zahlenmäßige Bewertung von Ereignissen die gleichen Eigenschaften wie die relative Häufigkeit besitzen. Sie wird **Wahrscheinlichkeit** P genannt. Kolmogoroff hat 1933 folgende drei Axiome postuliert, die die Wahhrscheinlichkeit P erfüllen muss:
Axiom 1:

$$\boxed{0 \leq P(A) \quad \text{für alle Ereignisse } A} \tag{5.14}$$

Axiom 2:

$$\boxed{P(\Omega) = 1} \tag{5.15}$$

Axiom 3:

$$\boxed{P(A \cup B) = P(A) + P(B) \quad \text{für disjunkte Ereignisse } A \text{ und } B.} \tag{5.16}$$

Schauen wir uns an, welche Konsequenzen wir aus den Axiomen ziehen können.

Satz 5.1
Es gilt

$$\boxed{P(\overline{A}) = 1 - P(A)\,.} \tag{5.17}$$

Beweis:
Es gilt

$$A \cup \overline{A} = \Omega$$

und

$$A \cap \overline{A} = \emptyset\,.$$

Dann gilt wegen Axiom 2 und Axiom 3:

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

Hieraus folgt die Behauptung. □

Beispiel 49 (fortgesetzt von Seite 194)
Mit $A = \{1, 2, 3\}$ und $P(A) = 0.5$ gilt

$$P(\overline{A}) = 1 - P(A) = 1 - 0.5 = 0.5\,.$$

□

Satz 5.2
Es gilt

$$\boxed{P(\emptyset) = 0.} \tag{5.18}$$

Beweis:
Es gilt

$$\overline{\emptyset} = \Omega\,.$$

Somit gilt mit Gleichung 5.17 und Axiom 2:

$$P(\emptyset) = 1 - P(\overline{\emptyset}) = 1 - P(\Omega) = 1 - 1 = 0\,.$$

□

Satz 5.3
Es gilt

$$\boxed{P(A \cap \overline{B}) = P(A) - P(A \cap B)\,.} \tag{5.19}$$

Beweis:
Es gilt

$$A = (A \cap B) \cup (A \cap \overline{B})\,.$$

Die Gültigkeit dieser Beziehung kann man Abbildung 5.6 auf der nächsten Seite entnehmen. Der horizontal schraffierte Bereich ist $A \cap \overline{B}$ und der vertikal schraffierte Bereich $A \cap B$.
Wegen

$$(A \cap B) \cap (A \cap \overline{B}) = \emptyset$$

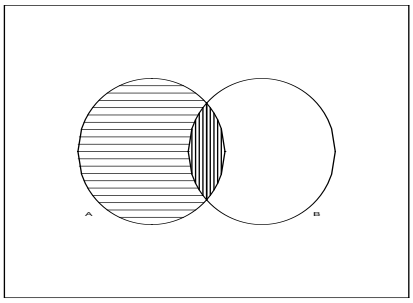
folgt wegen Axiom 3

$$P(A) = P((A \cap B) \cup (A \cap \overline{B})) = P(A \cap B) + P(A \cap \overline{B})\,.$$

Also gilt

$$P(A \cap \overline{B}) = P(A) - P(A \cap B)\,.$$

Abbildung 5.6: Venn-Diagramm zur Erläuterung



□

Beispiel 49 (fortgesetzt von Seite 197)
Es gilt $A \cap B = \{1, 3\}$ und somit

$$P(A \cap B) = \frac{1}{3}\,.$$

Also gilt

$$P(A \cap \overline{B}) = P(A) - P(A \cap B) = \frac{1}{2} - \frac{1}{3} = \frac{1}{6}\,.$$

□

Satz 5.4
Es gilt

$$\boxed{P(A \cup B) = P(A) + P(B) - P(A \cap B)\,.} \tag{5.20}$$

Beweis:
Es gilt

$$A \cup B = (A \cap \overline{B}) \cup B\,.$$

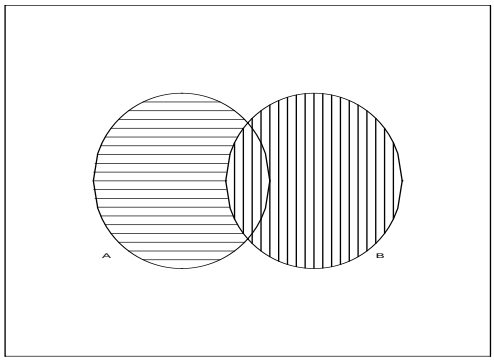
Die Gültigkeit dieser Beziehung kann man Abbildung 5.7 auf der nächsten Seite entnehmen. Der gesamte schraffierte Berich ist $A \cup B$, der horizontal schraffierte Bereich ist $A \cap \overline{B}$ und der vertikal schraffierte Bereich B.
Wegen

$$(A \cap \overline{B}) \cap B = A \cap \overline{B} \cap B = A \cap \emptyset = \emptyset$$

folgt dann wegen Axiom 3 und 5.19

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

Abbildung 5.7: Venn-Diagramm zur Erläuterung



□

Beispiel 49 (fortgesetzt von Seite 198)
Es gilt

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{1}{2} + \frac{1}{2} - \frac{1}{3} = \frac{2}{3}\,.$$

□

Satz 5.5
Seien A und B Ereignisse mit $A \subset B$. Dann gilt

$$P(A) \leq P(B) \tag{5.21}$$

Beweis:
Aus

$$B = (B \cap A) \cup (B \cap \overline{A})$$

und

$$(B \cap A) \cap (B \cap \overline{A}) = \emptyset$$

folgt wegen Axiom 3:

$$P(B) \;=\; P(B \cap A) + P(B \cap \overline{A})\,.$$

Aus

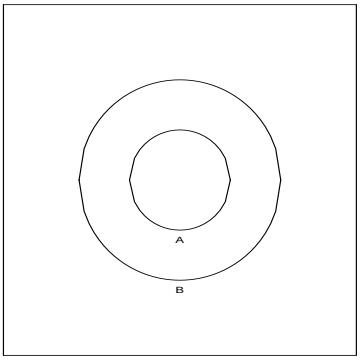
$$A \subset B$$

folgt

$$A \cap B = A\,.$$

Die Gültigkeit dieser Beziehung kann man Abbildung 5.8 entnehmen.

Abbildung 5.8: Venn-Diagramm



Somit gilt

$$P(B) \;=\; P(B \cap A) + P(B \cap \overline{A}) = P(A) + P(B \cap \overline{A})\,.$$

Wegen

$$P(B \cap \overline{A}) \geq 0$$

folgt also

$$P(B) - P(A) \geq 0$$

und somit

$$P(A) \leq P(B)\,.$$

□

Im täglichen Leben wird die Aussage des Satzes beim Vergleich von Wahrscheinlichkeiten oft verletzt. Folgendes Beispiel stammt von Kahneman und Tversky.

Beispiel 50
Linda ist 31 Jahre alt. Sie ist ledig, extrovertiert und intellektuell absolut brillant. Sie hat einen Abschluss in Philosophie. Als Studentin hat sie sich gegen Diskriminierung und soziale Ungerechtigkeit engagiert. Sie hat auch an Demonstrationen gegen Kernkraft teilgenommen.
Was ist wahrscheinlicher:

A: Linda ist Bankangestellte

B: Linda ist Bankangestellte und aktiv in der Frauenbewegung

90 Prozent der Befragten hielten die Alternative B für wahrscheinlicher, obwohl der in Alternative B beschriebene Personenkreis eine Teilmenge des in Alternative A beschriebenen Personenkreises ist.
Tversky und Kahneman erklären diesen Fehlschluss folgendermaßen:
Je detaillierter die Beschreibung eines Sachverhalts ist, um so unwahrscheinlicher ist dieser. Man kann ihn sich aber besser vorstellen, und somit erhöht sich eine Plausibilität. □

Der folgende Satz zeigt, dass Wahrscheinlichkeiten normiert sind. Somit kann man sagen, ob eine Wahrscheinlichkeit groß oder klein ist.

Satz 5.6
Es gilt:

$$\boxed{P(A) \leq 1} \tag{5.22}$$

Beweis:
Für jedes $A \in \mathcal{A}$ gilt

$$A \subset \Omega$$

Aus (5.21) folgt:

$$P(A) \leq P(\Omega)\,.$$

Mit Axiom 2 gilt also

$$P(A) \leq 1\,.$$

□

Beispiel 51
Eine Münze werde einmal geworfen. Es gilt

$$\Omega = \{K, Z\}\,.$$

Wir setzen

$$P(\{K\}) = p$$

mit $0 \leq p \leq 1$.
Aufgrund von Satz 5.2 auf Seite 197 gilt

$$P(\emptyset) = 0\,.$$

Aufgrund von Axiom 3 gilt

$$P(\{K, Z\}) = 1\,.$$

Aufgrund von Satz 5.1 auf Seite 196 gilt

$$P(\{Z\}) = 1 - P(\{K\}) = 1 - p\,.$$

Wir haben jedem Ereignis eine Wahrscheinlichkeit zugeordnet. Wir wissen aber nicht, welcher spezielle Wert für p gewählt werden soll. Sind wir uns sicher, dass die Münze fair ist, so werden wir $p = 0.5$ setzen. Wir werden in der Inferenzstatistik lernen, wie man datengestützt einen geeigneten Wert p finden kann. Außerdem werden wir Verfahren kennenlernen, mit denen man überprüfen kann, ob eine Münze fair ist. □

Das Beispiel zeigt, wie man bei einer endlichen Ergebnismenge Wahrscheinlichkeiten für Ereignisse konstruieren kann.
Sei

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_k\}\,.$$

Wir legen Wahrscheinlichkeiten $P(\{\omega_i\}) = p_i$ für die Elementarereignisse $\{\omega_i\}$, $i = 1, 2, \ldots$, fest.
Dabei muss gelten

$$0 \leq p_i \leq 1$$

und

$$\sum_{i=1}^{k} p_i = 1\,.$$

Für jedes Ereignis A setzen wir dann

$$P(A) = \sum_{\omega_i \in A} P(\{\omega_i\})\,.$$

Die drei Axiome sind erfüllt.
Axiom 1:

$$P(A) = \sum_{\omega_i \in A} P(\{\omega_i\}) \geq 0\,,$$

da alle $P(\{\omega_i\}) \geq 0$.
Axiom 2:

$$P(\Omega) = \sum_{\omega_i \in \Omega} P(\{\omega_i\}) = \sum_{i=1}^{k} p_i = 1\,.$$

Axiom 3:
Seien A und B disjunkte Ereignise.
Dann gilt

$$P(A \cup B) \;=\; \sum_{\omega \in A \cup B} P(\omega) = \sum_{\omega \in A} P(\omega) + \sum_{\omega \in B} P(\omega) = P(A) + P(B)\,.$$

Beispiel 52
Beim Gleichmöglichkeitsmodell gehen wir aus von der Ergebnismenge $\Omega = \{\omega_1, \omega_2, \ldots, \omega_k\}$ und setzen für $i = 1, \ldots, k$:

$$P(\{\omega_i\}) = \frac{1}{k}\,.$$

Die Wahrscheinlichkeit eines beliebigen Ereignisses $A \subset \Omega$ ist dann

$$P(A) = \sum_{\omega_i \in A} P(\{\omega_i\} = \frac{|A|}{|\Omega|}\,,$$

wobei $|M|$ die Mächtigkeit, d.h. die Anzahl der Elemente einer Menge M, angibt. □

5.2.4 Kombinatorik

Oft ist es gar nicht so einfach, die Anzahl der möglichen Ergebnisse zu bestimmen. Wir wollen dies für bestimmte Situationen systematisch angehen. Dabei gehen wir von einem Grundprinzip aus, das wir zunächst an einem Beispiel veranschaulichen.

Beispiel 53
Ein Mann besitzt drei Pullover und zwei Hosen. Die Farben der Pullover sind rot, gelb und blau, während die Farben der Hosen schwarz und weiß sind.
Auf Wie viele Arten kann sich der Mann kleiden?
Er kann den roten Pullover anziehen und zu diesem entweder die schwarze oder die weiße Hose. Er kann den gelben Pullover anziehen und zu diesem entweder die schwarze oder die weiße Hose. Er kann den blauen Pullover anziehen und zu diesem entweder die schwarze oder die weiße Hose.
Es gibt also die folgenden 6 Möglichkeiten:

- roter Pullover und schwarze Hose
- roter Pullover und weiße Hose
- gelber Pullover und schwarze Hose
- gelber Pullover und weiße Hose
- blauer Pullover und schwarze Hose
- blauer Pullover und weiße Hose

Zu jedem der drei Pullover kann man jede der beiden Hosen anziehen. □

Satz 5.7
Seien $A = \{a_1, \ldots, a_m\}$ und $B = \{b_1, \ldots, b_n\}$ Mengen.
Für das kartesische Produkt

$$\boxed{A \times B = \{(a, b) | a \in A, b \in B\}} \tag{5.23}$$

gilt

$$\boxed{|A \times B| = |A| \cdot |B| = m \cdot n\,.} \tag{5.24}$$

Beweis:
Der Beweis ist offensichtlich, wenn man sich die möglichen Ergebnisse folgendermaßen hinschreibt.

	b_1	b_2	$\ldots$	b_j	$\ldots$	b_n
a_1	(a_1, b_1)	(a_1, b_2)	$\ldots$	(a_1, b_j)	$\ldots$	(a_1, b_n)
a_2	(a_2, b_1)	(a_2, b_2)	$\ldots$	(a_2, b_j)	$\ldots$	(a_2, b_n)
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
a_i	(a_i, b_1)	(a_i, b_2)	$\ldots$	(a_i, b_j)	$\ldots$	(a_i, b_n)
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
a_m	(a_m, b_1)	(a_m, b_2)	$\ldots$	(a_m, b_j)	$\ldots$	(a_m, b_n)

□

Hieraus folgt sofort die Verallgemeinerung.

Satz 5.8
Seien $A_1, \ldots, A_r$ endliche Mengen.

Für das kartesische Produkt

$$\boxed{A_1 \times \ldots \times A_r = \{(a_1, \ldots, a_r) | a_i \in A_i, i = 1, \ldots, r\}} \tag{5.25}$$

gilt

$$\boxed{|A_1 \times A_2 \times \ldots \times A_r| = |A_1| \cdot |A_2| \cdot \ldots \cdot |A_r| \,.} \tag{5.26}$$

□

Beispiel 54
Aus $n_1 = 5$ Wegen von A nach B und $n_2 = 4$ Wegen von B nach C kann man $n_1 \cdot n_2 = 5 \cdot 4 = 20$ Reisen von A nach C über B zusammenstellen. □

Schauen wir uns nun die klassischen Fragestellung der Kombinatorik an. Diesen liegt folgendes Urnenmodell zugrunde:

> Ein Urne enthält n Kugeln, die von 1 bis n durchnummeriert sind. Der Urne werden nacheinander k Kugeln entnommen. Nach jedem Zug notiert man die Nummer der gezogenen Kugel und legt die Kugel entweder zurück (Ziehen mit Zurücklegen) oder legt die Kugel zur Seite (Ziehen ohne Zurücklegen).

Außerdem unterscheiden wir noch, ob die Anordnung berücksichtigt werden soll oder nicht.

Beispiel 55
Wir gehen davon aus, dass die Urne vier Kugeln enthält, die von 1 bis 4 durchnummeriert sind.

$$U = \{1, 2, 3, 4\}$$

□

Wir fangen mit dem Ziehen mit Zurücklegen unter Berücksichtigung der Anordnung an.

Beispiel 55 (fortgesetzt)
Vor dem ersten Zug ist der Zustand der Urne

$$U = \{1, 2, 3, 4\}$$

und, da wir mit Zurücklegen ziehen, ist er vor dem zweiten Zug ebenfalls

$$U = \{1, 2, 3, 4\} \,.$$

Die Zahl auf der ersten gezogenen Kugel ist entweder eine 1, 2, 3 oder 4. Nehmen wir an, es ist eine 1. Dann kann die Zahl auf der zweiten gezogenen Kugel entweder eine 1, 2, 3 oder 4 sein.
Wir erhalten also folgende 4 Möglichkeiten

$$(1,1) \quad (1,2) \quad (1,3) \quad (1,4)$$

Zu jeder der anderen 3 ersten Zahlen beim ersten Zug gibt es ebenfalls wieder 4 Möglichkeiten beim zweiten Zug. Somit gibt es folgende $4^2 = 16$ Möglichkeiten:

$$\begin{array}{cccc} (1,1) & (1,2) & (1,3) & (1,4) \\ (2,1) & (2,2) & (2,3) & (2,4) \\ (3,1) & (3,2) & (3,3) & (3,4) \\ (4,1) & (4,2) & (4,3) & (4,4) \end{array}$$

□

Der folgende Satz verallgemeinert das Beispiel.

Satz 5.9
Die Anzahl der geordneten Stichproben vom Umfang k aus einer Menge vom Umfang n beträgt beim Ziehen mit Zurücklegen

$$\boxed{n^k}\,. \tag{5.27}$$

Beweis:
Sei B eine Menge mit n Elementen. Die Menge aller möglichen geordneten Stichproben ist beim Ziehen mit Zurücklegen

$$A = \{(a_1, \ldots, a_k) | a_i \in B, i = 1, \ldots, k\} = B \times B \times \ldots \times B\,.$$

Vor jeder Ziehung enthält die Menge, aus der gezogen wird, also n Elemente. Aufgrund von Satz 5.8 auf Seite 204 gilt also

$$|A| = |B \times B \times \ldots \times B| = \underbrace{|B| \cdot \ldots \cdot |B|}_{k\,Faktoren} = |B|^k = n^k\,.$$

□

Beispiel 56
Mit einer Base lassen sich 4, mit zwei Basen $4 \cdot 4 = 4^2 = 16$ Aminosäuren kodieren. Für 3 Basen haben wir $4 \cdot 4 \cdot 4 = 4^3 = 64 > 20$ Möglichkeiten. Tatsächlich bilden jeweils ein Tripel von Basen die Grundbausteine des genetischen Codes. □

Beispiel 57
Ein Byte besteht aus 8 Bits, wobei ein Bit ein Informationseinheit ist, die zwei unterschiedliche Zustände, zum Beispiel 0 oder 1 annehmen kann. In einem Byte können also $2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 = 2^8 = 256$ unterschiedliche Informationen gespeichert werden. □

Beispiel 58
Eine Menge M vom Umfang n hat 2^n Teilmengen. Diese sieht man folgendermaßen:
Ohne Beschränkung der Allgemeinheit nehmen wir an, dass gilt

$$M = \{1, 2, \ldots, n-1, n\}\,.$$

Wir definieren einen Vektor $v = (v_1, v_2, \ldots, v_n\}$ mit

$$v_i = \begin{cases} 1 & \text{falls i in der Teilmenge} \\ 0 & \text{sonst} \end{cases}$$

Ist zum Beispiel

$$M = \{1, 2, 3\},$$

so liefert der Vektor

$$v = (1, 0, 1)$$

die Teilmenge

$$\{1, 3\},$$

während der Vektor

$$v = (0, 0, 0)$$

die leere Menge $\emptyset$ liefert.
Zu jeder Teilmenge gibt es genau einen Vektor v. Da es 2^n Bitvektoren der Länge n gibt, hat eine Menge vom Umfang n genau 2^n Teilmengen. □

Beispiel 59
Das Beispiel stammt von Tversky und Kahnemann.
Personen wurde jedes der folgenden beiden Muster gezeigt:
Muster 1

```
X X
X X
X X
X X
X X
```

```
X X
X X
X X
X X
```

Muster 2

```
X X X X X X X X X
X X X X X X X X X
X X X X X X X X X
```

Sie sollten sagen, bei welchem der beiden Muster es mehr Pfade von der ersten Zeile bis zur letzten Zeile gibt, wobei man auf einem beliebigen X der ersten Zeile startet und in jeder der darunterliegenden Zeilen auf genau ein Symbol geht.

Für Muster 1 gibt es $2^9 = 512$ unterschiedliche Pfade, da man neun Symbole trifft, und es für jedes dieser Symbole genau 2 Möglichkeiten gibt. Für Muster 2 gibt es $8^3 = 512$ unterschiedliche Pfade, da man drei Symbole trifft, und es für jedes dieser Symbole genau 8 Möglichkeiten gibt. Bei beiden Mustern ist die Anzahl der Pfade also gleich.

Von den befragten Personen fanden 85 Prozent, dass es bei Muster 1 mehr Pfade gibt. Dies liegt daran, dass man sich die unterschiedlichen Pfade bei Muster 1 leichter vorstellen kann. □

Wenden wir uns nun dem Ziehen ohne Zurücklegen mit Berücksichtigung der Anordnung zu.

Beispiel 55 (fortgesetzt von Seite 205)
Vor dem ersten Zug ist der Zustand der Urne

$$U = \{1, 2, 3, 4\}.$$

Da wir ohne Zurücklegen ziehen, hängt er vor dem zweiten Zug vom Ergebnis des ersten Zuges ab. Nehmen wir an, dass die erste gezogenen Kugel eine 1 ist. Dann kann die Zahl auf der zweiten gezogenen Kugel entweder eine 2, 3 oder 4 sein. Wir erhalten also folgende 3 Möglichkeiten

$$(1,2) \quad (1,3) \quad (1,4)$$

Zu jeder der anderen 3 ersten Zahlen beim ersten Zug gibt es ebenfalls wieder 3 Möglichkeiten beim zweiten Zug. Somit gibt es folgende $4 \cdot 3 = 12$ Möglichkeiten:

$$\begin{array}{cccc} & (1,2) & (1,3) & (1,4) \\ (2,1) & & (2,3) & (2,4) \\ (3,1) & (3,2) & & (3,4) \\ (4,1) & (4,2) & (4,3) & \end{array}$$

□

Der folgende Satz verallgemeinert das Beispiel.

Satz 5.10
Die Anzahl der geordneten Stichproben vom Umfang k aus einer Menge vom Umfang n beträgt beim Ziehen ohne Zurücklegen

$$\boxed{(n)_k = n \cdot (n-1) \cdots (n-k+1)\,.} \tag{5.28}$$

Beweis:
Sei B eine Menge mit n Elementen. Die Menge aller möglichen geordneten Stichproben ist beim Ziehen ohne Zurücklegen

$$\begin{aligned} A &= \{(a_1, \ldots, a_k) | a_i \in A_i, |A_i| = n - i + 1, i = 1, \ldots, k\} \\ &= A_1 \times A_2 \times \ldots \times A_k\,. \end{aligned}$$

Bei jeder Ziehung enthält die Menge, aus der gezogen wird, also ein Element weniger, bei der ersten Ziehung n Elemente, bei der zweiten $n-1$ Elemente, u.s.w.. Aufgrund von Satz 5.8 auf Seite 204 gilt also

$$|A| = |A_1 \times A_2 \times \ldots \times A_k| = |A_1| \cdot |A_2| \cdot \ldots \cdot |A_k| = n \cdot (n-1) \cdot \ldots \cdot (n-k+1)$$

□

$(n)_k$ wird gelesen als 'n sub k'.

Satz 5.11
Es gilt

$$\boxed{(n)_k = \frac{n!}{(n-k)!}\,.}$$

Beweis:
Es gilt

$$\begin{aligned} (n)_k &= n \cdot (n-1) \cdots (n-k+1) \\ &= \frac{n \cdot (n-1) \cdots (n-k+1) \cdot (n-k) \cdot (n-k-1) \cdots 2 \cdot 1}{(n-k) \cdot (n-k-1) \cdots 2 \cdot 1} \\ &= \frac{n!}{(n-k)!}\,. \end{aligned}$$

□

Die Menge aller möglichen geordneten Stichproben vom Umfang n aus einer Menge, die n Elemente enthält, ist also:

$$n! = n \cdot (n-1) \cdot \ldots \cdot 2 \cdot 1 \,.$$

Man spricht auch von den Permutationen einer Menge. Wir lesen $n!$ als 'n Fakultät'. Wir setzen $0! = 1$.
Tabelle 5.1 gibt die Werte von $n!$ für $n \leq 10$ an. Wir sehen, dass $n!$ sehr

Tabelle 5.1: Werte von $n!$

n	0	1	2	3	4	5	6	7	8	9	10
$n!$	1	1	2	6	24	120	720	5040	40320	362880	3628800

schnell wächst.

Beispiel 60
Wir betrachten die Menge

$$U = \{1, 2, 3, 4\} \,.$$

Es gibt $4! = 24$ Permutationen der Elemente der Menge U.

$(1,2,3,4)$	$(1,2,4,3)$	$(1,3,2,4)$	$(1,3,4,2)$	$(1,4,2,3)$	$(1,4,3,2)$
$(2,1,3,4)$	$(2,1,4,3)$	$(2,3,1,4)$	$(2,3,4,1)$	$(2,4,1,3)$	$(2,4,3,1)$
$(3,1,2,4)$	$(3,1,4,2)$	$(3,2,1,4)$	$(3,2,4,1)$	$(3,4,1,3)$	$(3,4,3,1)$
$(4,1,2,3)$	$(4,1,3,2)$	$(4,2,1,3)$	$(4,2,3,1)$	$(4,3,1,2)$	$(4,3,2,1)$

□

Beispiel 61
6 Personen können auf $6! = 720$ unterschiedliche Arten auf 6 Stühlen nebeneinander sitzen. □

Wir betrachten nun ein Beispiel mit Ziehen mit und ohne Zurücklegen.

Beispiel 62
Ein fairer Würfel wird sechsmal hintereinander geworfen. Wie groß ist die Wahrscheinlichkeit, dass alle sechs Augenzahlen auftreten?
Wir benutzen das Gleichmöglichkeitsmodell. Die Anzahl der möglichen Ergebnisse ist

$$6^6 = 46656 \,,$$

da wir mit Zurücklegen 6 Zahlen aus der Menge $\{1, 2, 3, 4, 5, 6\}$ ziehen.
Die Anzahl der günstigen Fälle ist

$$6! = 720\,,$$

da wir ohne Zurücklegen 6 Zahlen aus der Menge $\{1, 2, 3, 4, 5, 6\}$ ziehen.
Die gesuchte Wahrscheinlichkeit ist also

$$\frac{6!}{6^6} = 0.0154321\,.$$

□

Beispiel 63
In einem Zimmer befinden sich drei Personen. Wie groß ist die Wahrscheinlichkeit, dass mindestens zwei von ihnen am gleichen Tag des Jahres Geburtstag haben?
Sei
A_3: mindestens zwei der drei Personen haben am gleichen Tag des Jahres Geburtstag.
Es ist einfacher, die Wahrscheinlichkeit des Gegenereignisses zu bestimmen.
Es gilt
$\overline{A_3}$: jede der drei Personen hat an einem anderen Tag des Jahres Geburtstag.
Wir betrachten wieder das Gleichmöglichkeitsmodell.
Die Anzahl der möglichen Fälle ist

$$365^3 = 48627125\,,$$

da wir mit Zurücklegen drei Zahlen aus der Menge $\{1, 2, 3, \ldots, 364, 365\}$ ziehen.
Die Anzahl der günstigen Fälle ist

$$(365)_3 = 48228180\,,$$

da wir ohne Zurücklegen drei Zahlen aus der Menge $\{1, 2, 3, \ldots, 364, 365\}$ ziehen.
Es gilt also

$$P(\overline{A_3}) = \frac{(365)_3}{365^3} = 0.9917958\,.$$

Also gilt

$$P(A_3) = 1 - P(\overline{A_3}) = 0.0082042\,.$$

Wie nicht anders zu erwarten war, ist diese Wahrscheinlichkeit ziemlich klein.

Ab wie vielen Leuten lohnt es sich, darauf zu wetten, dass mindesten zwei am gleichen Tag des Jahres Geburtstag haben?
Sei A_k: mindestens zwei von k Leuten haben am gleichen Tag des Jahres Geburtstag.
Es gilt

$$P(A_k) = 1 - \frac{(365)_k}{365^k}\,.$$

Wir erhalten speziell

$$\begin{aligned} P(A_{20}) &= 0.4114384 \\ P(A_{21}) &= 0.4436883 \\ P(A_{22}) &= 0.4756953 \\ P(A_{23}) &= 0.5072972 \end{aligned}$$

Ab 23 Leuten lohnt es sich, darauf zu wetten, dass mindestens zwei am gleichen Tag des Jahres Geburtstag haben. Diese Anzahl ist überraschend klein.
In einer amerikanischen Talkshow trug ein Teilnehmer dieses Beispiel vor. Der Moderator wollte es gleich überprüfen und fragte:

> Ich habe am 23. Juli Geburtstag. Wer hat noch am 23. Juli Geburtstag?

Keiner der 74 Zuschauer im Raum meldete sich. Die Wahrscheinlichkeit, dass von 75 Personen in einem Raum mindestens zwei am gleichen Tag des Jahres Geburtstag haben, beträgt aber

$$P(A_{75}) = 0.9997199\,.$$

Also spricht sehr viel dafür, dass etwas an der Frage des Moderators falsch sein muss.
Der Moderator hat einen Fehler gemacht. Im Beispiel wurde nach der Wahrscheinlichkeit gefragt, dass zwei Personen an **irgendeinem** Tag des Jahres Geburtstag haben. Der Moderator wollte wissen, ob zwei Personen an einem **bestimmten** Tag des Jahres Geburtstag haben.
Sei B_k: mindestens einer von k Personen hat am gleichen Tag wie der Moderator Geburtstag.
Wir betrachten wiederum das Gegenereignis $\overline{B_k}$: keiner von k Personen hat am gleichen Tag wie der Moderator Geburtstag.
Es gilt

$$P(\overline{B_k}) = \frac{364^k}{365^k}\,.$$

Es gilt

$$P(B_{75}) = 1 - \frac{364^{75}}{365^{75}} = 0.1859728\,.$$

Für $k = 252$ gilt

$$P(B_{252}) = 1 - \frac{364^{252}}{365^{252}} = 0.499$$

und für $k = 253$ gilt

$$P(B_{253}) = 1 - \frac{364^{253}}{365^{253}} = 0.5005\,.$$

Erst ab 253 Personen im Raum lohnt es sich zu wetten, dass mindestens einer am gleichen Tag Geburtstag hat wie der Moderator. □

Nun soll die Anordung nicht wichtig sein. Wir betrachten hier nur das Ziehen ohne Zurücklegen.

Beispiel 55 (fortgesetzt von Seite 208)
Wir betrachten wieder

$$U = \{1, 2, 3, 4\}\,.$$

Wir suchen die Anzahl der Möglichkeiten, aus U zwei Zahlen ohne Zurücklegen zu ziehen, wobei uns die Reihenfolge der Ziehung egal ist.
Die Ziehungen $(1, 2)$ und $(2, 1)$ sind für uns also identisch.
Wir schauen uns also alle Möglichkeiten beim Ziehen ohne Zurücklegen ohne Berücksichtigung der Anordnung an und behalten nur die, bei denen die Zahlen der Größe nach geordnet sind. Die folgende Tabelle zeigt die Vorgehensweise.

$(1, 2)$	$(2, 1)$	$\{1, 2\}$
$(1, 3)$	$(3, 1)$	$\{1, 3\}$
$(1, 4)$	$(4, 1)$	$\{1, 4\}$
$(2, 3)$	$(3, 2)$	$\{2, 3\}$
$(2, 4)$	$(4, 2)$	$\{2, 4\}$
$(3, 4)$	$(4, 3)$	$\{3, 4\}$

□

Der folgende Satz verallgemeinert das Beispiel.

Satz 5.12
Eine Menge vom Umfang n hat

$$\boxed{\frac{(n)_k}{k!}} \tag{5.29}$$

Teilmengen vom Umfang k.

Beweis:

Aus einer Menge vom Umfang n lassen sich $(n)_k$ k-Tupel ohne Wiederholung bilden. Jedes solche k-Tupel entsteht aus einer Teilmenge vom Umfang k durch Permutation der k Elemente dieser Teilmenge. Da es $k!$ Permutationen einer Menge gibt, gibt es $k!$-mal soviele r-Tupel wie Teilmengen vom Umfang k.

Sei x die Anzahl der Teilmengen vom Umfang k. Es gilt also

$$(n)_k = x \cdot k!$$

und somit

$$x = \frac{(n)_k}{k!}\,.$$

□

Wir setzen

$$\boxed{\binom{n}{k} = \frac{(n)_k}{k!}} \tag{5.30}$$

und lesen dies als 'n über k'.

Man spricht auch vom Binomialkoeffizienten.

Schauen wir uns einige Eigenschaften der Binomialkoeffizienten an.

Satz 5.13

Es gilt

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

Beweis:

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

□

Satz 5.14
Es gilt

1. $$\binom{n}{0} = 1$$

2. $$\binom{n}{1} = n$$

3. $$\binom{n}{k} = \binom{n}{n-k}$$

4. $$\binom{n}{k} = \frac{n}{k}\binom{n-1}{k-1}$$

5. $$\binom{n}{k} = \frac{n}{n-k}\binom{n-1}{k}$$

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

Beweis:

1. Es gibt eine Teilmenge vom Umfang 0 aus einer Menge M. Dies ist die leere Menge.

2. Es gibt n Teilmengen vom Umfang 1 aus einer Menge M vom Umfang n.

   Dies sind die einelementigen Mengen.

3. Es gibt $\binom{n}{k}$ Möglichkeiten, aus n Elementen k ohne Zurücklegen zu ziehen, wobei die Reihenfolge keine Rolle spielt. Anstatt die gezogenen Elemente zu betrachten, kann man auch die nicht gezogenen betrachten.

4.
$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!} = \frac{n \cdot (n-1)!}{k \cdot (k-1)!\,(n-k)!} = \frac{n}{k}\binom{n-1}{k-1}$$

5.
$$\binom{n}{k} = \binom{n}{n-k} = \frac{n}{n-k}\binom{n-1}{n-k-1} = \frac{n}{n-k}\binom{n-1}{k}$$

6. Die Auswahl von k Eklementen aus einer Menge von n Elementen kann so erfolgen:

   Wir färben eines der Elemente weiß, die anderen schwarz. Dann gibt es zwei Sorten von Teilmengen vom Umfang k - solche, die das weiße Element enthalten, und solche, die es nicht enthalten. Von der ersten Sorte gibt es $\binom{n-1}{k-1}$, da wir $k-1$ aus den $n-1$ schwarzen auswählen müssen, von der zweiten Sorte gibt es $\binom{n-1}{k}$, da wir k aus den $n-1$ schwarzen aus wählen müssen.

□

Beispiel 64
Beim Lotto werden 6 Kugeln aus 49 Kugeln ohne Zurücklegen gezogen, wobei die Reihenfolge der Ziehung nicht interessiert. Es gibt also

$$\binom{49}{6} = \frac{(49)_6}{6!} = \frac{49 \cdot 48 \cdot 47 \cdot 46 \cdot 45 \cdot 44}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 13983816$$

unterschiedliche Ziehungen. □

Beispiel 65
Wie viele Folgen aus n Kugeln, von denen k schwarz und $n-k$ weiß sind, kann man bilden? Eine Folge ist eindeutig festgelegt, wenn man die Positionen der schwarzen Kugeln kennt. Die Positionen der schwarzen Kugeln erhält man also, wenn man aus den n Positionen k ohne Zurücklegen auswählt, wobei die Reihenfolge der Ziehung irrelevant ist. Es gibt also $\binom{n}{k}$ mögliche Folgen. Für den Fall $n = 4$ und $k = 2$ sind die $\binom{4}{2} = 6$ möglichen Folgen zusammengestellt.

$(1,2)$	ssww
$(1,3)$	swsw
$(1,4)$	swws
$(2,3)$	wssw
$(2,4)$	wsws
$(3,4)$	wwss

□

Wir betrachten noch ein sehr wichtiges Beispiel, auf das wir in diesem Skript sehr oft zurückkomen werden.

Beispiel 66
Eine Urne enthält N Kugeln, die von 1 bis N durchnummeriert sind. Von den Kugeln sind K schwarz und die restlichen $N - K$ weiß. Es werden n Kugeln aus der Urne gezogen.
Wie groß ist die Wahrscheinlichkeit, dass k der gezogenen Kugeln schwarz sind, wenn

1. **mit** Zurücklegen
2. **ohne** Zurücklegen

gezogen wird?
Sei A_k das Ereignis, dass k der gezogenen Kugeln schwarz sind.
Es gilt

$$P(A_k) = \frac{|A_k|}{|\Omega|} .$$

Wir bestimmen zuerst $|\Omega|$:

1. Da aus den N Kugeln n mit Zurücklegen gezogen werden, gilt

$$|\Omega| = N^n .$$

2. Da aus den N Kugeln n ohne Zurücklegen gezogen werden, gilt

$$|\Omega| = (N)_n .$$

Die k schwarzen und $n - k$ weißen Kugeln können auf $\binom{n}{k}$ unterschiedliche Arten angeordnet werden. Zu jeder dieser $\binom{n}{k}$ unterschiedlichen Positionen gibt es beim Ziehen

1. **mit** Zurücklegen
$$K^k \, (N - K)^{(n-k)}$$
2. **ohne** Zurücklegen
$$(K)_k \, (N - K)_{(n-k)}$$

unterscheidbare $n - Tupel$.
Es gilt also beim Ziehen

1. **mit** Zurücklegen

$$\begin{aligned} P(A_k) &= \frac{\binom{n}{k} K^k (N-K)^{n-k}}{N^n} = \frac{\binom{n}{k} K^k (N-K)^{n-k}}{N^k N^{N-K}} \\ &= \binom{n}{k} \left(\frac{K}{N}\right)^k \left(\frac{N-K}{N}\right)^{n-k} = \binom{n}{k} \left(\frac{K}{N}\right)^k \left(1-\frac{K}{N}\right)^{n-k} \\ &= \binom{n}{k} p^k (1-p)^{n-k} \end{aligned}$$

mit

$$p = \frac{K}{N}\,.$$

2. **ohne** Zurücklegen

$$\begin{aligned} P(A_k) &= \binom{n}{k} \frac{K_k (N-K)_{n-k}}{N_n} = \frac{n!}{k!(n-k)!} \frac{K_k (N-K)_{n-k}}{N_n} \\ &= \frac{\frac{(K)_k}{k!} \frac{(N-K)_{n-k}}{(n-k)!}}{\frac{(N)_n}{n!}} = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \end{aligned}$$

□

5.2.5 Bedingte Wahrscheinlichkeit

Vielfach ist bei Zufallsvorgängen bekannt, dass ein Ereignis B eingetreten ist. Gesucht ist dann die Wahrscheinlichkeit, dass auch das Ereignis A eintritt. So ist beispielsweise bekannt, dass eine Person weiblich ist, und sucht die Wahrscheinlichkeit, dass sie eine bestimmte Partei wählt.

Wie kann man die Wahrscheinlichkeit eines Ereignisses A bestimmen, wenn man weiß, dass das Ereignis B eingetreten ist? Schauen wir uns ein motivierendes Beispiel an.

Beispiel 67
Ein fairer Würfel wird einmal geworfen. Die Ergebnismenge ist also

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

Die Wahrscheinlichkeit eines Ereignisses A ergibt sich zu:

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

Sei nun speziell $A = \{1, 2, 3\}$, d.h. die gewürfelte Augenzahl beträgt höchstens 3. Es gilt

$$P(A) = \frac{|\{1, 2, 3\}|}{|\{1, 2, 3, 4, 5, 6\}|} = \frac{3}{6} = \frac{1}{2}\,.$$

Nun seien die Seiten des Würfels, auf denen die ungeraden Augenzahlen stehen, rot gefärbt. Sei B das Ereignis, dass eine rote Seite obenliegt, also $B = \{1, 3, 5\}$. Es gilt

$$P(B) = \frac{|\{1, 3, 5\}|}{|\{1, 2, 3, 4, 5, 6\}|} = \frac{3}{6} = \frac{1}{2}\,.$$

Der Würfel werde einmal geworfen. Man kann erkennen, dass eine rote Seite obenliegt, die Augenzahl ist jedoch nicht zu erkennen.
Wie groß ist die Wahrscheinlichkeit, dass die Augenzahl höchstens 3 beträgt?
Wir wissen also, dass das Ereignis B eingetreten ist und suchen die Wahrscheinlichkeit, dass das Ereignis A eintritt. Dadurch dass das Ereignis B eingetreten ist, ist die Menge der möglichen Ergebnisse nicht mehr Ω, sondern $B = \{1, 3, 5\}$. Die Menge der günstigen Ergebnisse ist dann $\{1, 3\}$. Die Wahrscheinlichkeit, dass die gewürfelte Augenzahl höchstens 3 beträgt, wenn man weiß, dass eine ungerade Augenzahl gewürfelt wurde, beträgt also

$$\frac{|\{1, 3\}|}{|\{1, 3, 5\}|} = \frac{2}{3}\,.$$

□

Man spricht von der bedingten Wahrscheinlichkeit von A unter der Bedingung, dass das Ereignis B eingetreten ist, und schreibt hierfür

$$P(A|B)\,.$$

Überlegen wir uns noch einmal, wie wir diese berechnet haben. Dadurch, dass man weiß, dass das Ereignis B eingetreten ist, ist die Menge der möglichen Ergebnisse nicht mehr Ω, sondern B. Die Wahrscheinlichkeit, dass nun A eintritt, ergibt sich aufgrund des Gleichmöglichkeitsmodells zu

$$\frac{|A \cap B|}{|B|}\,.$$

Wegen

$$\frac{|A \cap B|}{|B|} = \frac{|A \cap B|/|\Omega|}{|B|/|\Omega|} = \frac{P(A \cap B)}{P(B)}$$

hätten wir die bedingte Wahrscheinlichkeit auch über $P(A \cap B)$ und $P(B)$ bestimmen können. Dies legt folgende Definition nahe.

Definition 5.6
Seien A und B Ereignisse. Die bedingte Wahrscheinlichkeit von A gegeben B ist für $P(B) > 0$ definiert durch

$$\boxed{P(A|B) = \frac{P(A \cap B)}{P(B)}}\,. \tag{5.31}$$

Ist $P(B) = 0$, so ist $P(A|B)$ nicht definiert.

$P(A|B)$ erfüllt die drei Axiome von Kolmogoroff.
Axiom 1:
Mit

$$P(A \cap B) \geq 0$$

und

$$P(B) \geq 0$$

folgt

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \geq 0\,.$$

Axiom 2:
Wegen $\Omega \cap B = B$ gilt

$$P(\Omega|B) = \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1\,.$$

Axiom 3:
Für zwei disjunkte Ereignisse A und C haben wir zu zeigen:

$$P(A \cup C|B) = P(A|B) + P(C|B)\,.$$

Wegen $A \cap C = \emptyset$ gilt

$$(A \cap B) \cap (C \cap B) = \emptyset\,.$$

Also folgt

$$\begin{aligned} P(A \cup C|B) &= \frac{P((A \cup C) \cap B)}{P(B)} = \frac{P((A \cap B) \cup (C \cap B))}{P(B)} \\ &= \frac{P(A \cap B) + P(C \cap B)}{P(B)} = \frac{P(A \cap B)}{P(B)} + \frac{P(C \cap B)}{P(B)} \\ &= P(A|B) + P(C|B)\,. \end{aligned}$$

Somit gilt:

$$P(A|B) = 1 - P(\overline{A}|B)\,.$$

Es gilt aber **nicht** notwendigerweise:

$$P(A|B) = 1 - P(A|\overline{B})\,.$$

Im Beispiel 67 gilt $P(A|B) > P(A)$. Das folgende Beispiel zeigt, dass auch $P(A|B) < P(A)$ möglich ist.

Beispiel 67 (fortgesetzt von Seite 218)
Wir betrachten nun das Ereignis $C = \{2\}$.
Es gilt

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

und

$$P(C|B) = \frac{P(B \cap C)}{P(B)} = 0\,.$$

□

Der Fall $P(A|B) = P(A)$ ist so wichtig, dass wir ihm ein eigenes Kapitel widmen.

5.2.6 Multiplikationssätze

Bei vielen Anwendungen ist die Wahrscheinlichkeit $P(A \cap B)$ gesucht. Sind $P(A|B)$ und $P(B)$ bekannt, so erhält man aus

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

durch Multiplikation mit $P(B)$ die Wahrscheinlichkeit:

$$\boxed{P(A \cap B) = P(A|B)\,P(B)\,.} \tag{5.32}$$

Aus

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

folgt

$$\boxed{P(A \cap B) = P(B|A)\, P(A)\,.} \tag{5.33}$$

Beispiel 68
In einer Kiste sind 10 Glühbirnen, von denen 4 defekt sind. Zwei Glühbirnen werden ohne Zurücklegen gezogen. Wie groß ist die Wahrscheinlichkeit, dass beide defekt sind?
Seien

D_1: die erste gezogene Glühbirne ist defekt
D_2: die zweite gezogene Glühbirne ist defekt

Um die Regel

$$P(D_1 \cap D_2) \;=\; P(D_2|D_1)\, P(D_1)$$

anwenden zu können, benötigen wir $P(D_1)$ und $P(D_2|D_1)$.
Vor dem ersten Zug befinden sich 10 Glühbirnen in der Kiste, von denen 4 defekt sind. Es gilt $P(D_1) = 0.4$. Da wir beim ersten Zug eine defekte Glühbirne gezogen haben, befinden sich vor dem zweiten Zug 9 Glühbirnen in der Kiste, von denen 3 defekt sind. Also gilt

$$P(D_2|D_1) = \frac{3}{9}\,.$$

Also gilt

$$P(D_1 \cap D_2) \;=\; P(D_2|D_1)\, P(D_1) = \frac{3}{9} \cdot \frac{4}{10} = \frac{2}{15}\,.$$

Dieses Ergebnis hätten wir auch mit Hilfe der Kombinatorik und dem Gleichmöglichkeitsmodell bestimmen können. Es gibt

$$\binom{10}{2} = 45$$

Möglichkeiten, aus den 10 Glühbirnen 2 ohne Zurücklegen und ohne Berücksichtigung der Anordnung auszuwählen. Um die Anzahl der günstigen Fälle zu bestimmen, schauen wir uns zunächst die defekten Glühbirnen an. Es gibt

$$\binom{4}{2} = 6$$

Möglichkeiten, aus den 4 defekten Glühbirnen 2 ohne Zurücklegen und ohne Berücksichtigung der Anordnung auszuwählen. Zu jeder dieser Möglichkeiten gibt es

$$\binom{6}{0} = 1$$

Möglichkeit, aus den 6 nicht defekten Glühbirnen 0 ohne Zurücklegen und ohne Berücksichtigung der Anordnung auszuwählen. Die gesuchte Wahrscheinlichkeit beträgt also

$$\frac{\binom{4}{2}\binom{6}{0}}{\binom{10}{2}} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 10 \cdot 9} = \frac{2}{15}\,.$$

□

Die Verallgemeinerung auf mehr als zwei Ereignisse liefert der folgende Satz.

Satz 5.15
Seien $A_1, \ldots, A_n$ Ereignisse mit

$$P(A_1 \cap \ldots \cap A_{n-1}) > 0\,.$$

Dann gilt

$P(A_1 \cap \ldots \cap A_n) =$

$$P(A_n|A_1 \cap \ldots \cap A_{n-1}) \cdot P(A_{n-1}|A_1 \cap \ldots \cap A_{n-2}) \cdot \ldots \cdot P(A_2|A_1) \cdot P(A_1)\,.$$

Beweis:
Es gilt

$$(A_1 \cap \ldots \cap A_{n-1}) \subset (A_1 \cap \ldots \cap A_{n-2}) \subset \ldots \subset (A_1 \cap A_2) \subset A_1\,.$$

Wegen

$$P(A_1 \cap \ldots \cap A_{n-1}) > 0.$$

und Satz 5.21 auf Seite 199 folgt

$$0 < P(A_1 \cap \ldots \cap A_{n-1}) \leq P(A_1 \cap \ldots \cap A_{n-2}) \leq \ldots \leq P(A_1)\,.$$

Also gilt

$P(A_1 \cap \ldots \cap A_n) =$

$$\frac{P(A_1 \cap \ldots \cap A_n)}{P(A_1 \cap \ldots \cap A_{n-1})} \frac{P(A_1 \cap \ldots \cap A_{n-1})}{P(A_1 \cap \ldots \cap A_{n-2})} \cdots \frac{P(A_1 \cap A_2)}{P(A_1)} \cdot P(A_1) =$$

$$P(A_n|A_1 \cap \ldots \cap A_{n-1}) \cdot P(A_{n-1}|A_1 \cap \ldots \cap A_{n-2}) \cdot \ldots \cdot P(A_2|A_1) \cdot P(A_1)\,.$$

□

Beispiel 68 (fortgesetzt von Seite 222)
In einer Kiste sind 10 Glühbirnen, von denen 4 defekt sind. Vier Glühbirnen werden ohne Zurücklegen gezogen. Wie groß ist die Wahrscheinlichkeit, dass alle vier defekt sind?
Seien für $i = 1, 2, 3, 4$:

D_i: die i-te gezogene Glühbirne ist defekt

Dann ist

$$P(D_1 \cap D_2 \cap D_3 \cap D_4) =$$

$$P(D_4|D_1 \cap D_2 \cap D_3) \cdot P(D_3|D_1 \cap D_2) \cdot P(D_2|D_1) \cdot P(D_1) =$$

$$\frac{1}{7} \cdot \frac{2}{8} \cdot \frac{3}{9} \cdot \frac{4}{10} =$$

$$\frac{1}{210}.$$

□

5.2.7 Satz von der totalen Wahrscheinlichkeit

Oft betrachten wir in einer Population ein Merkmal. Durch die einzelnen Merkmalsausprägungen wird die Population in disjunkte Teilpopulationen zerlegt. Die folgende Definition überträgt diesen Sachverhalt auf Zufallsvorgänge.

Definition 5.7
Sei Ω eine Ergebnismenge. Die Ereignisse $A_1, A_2, \ldots A_n \subset \Omega$ bilden ein vollständiges System von Ereignissen, wenn gilt

$$A_1 \cup \ldots \cup A_n = \Omega$$

und

$$A_i \cap A_j = \emptyset$$

für $i \neq j$.

Beispiel 69
In einer Population leidet 1 Promille der Bevölkerung an einer Krankheit. Eine Person wird zufällig aus dieser Population gezogen. Die Ergebnismenge dieses Zufallsvorgangs ist $\Omega = \{k, g\}$, wobei k für krank und g für gesund steht. Sei $K = \{k\}$ das Ereignis, dass die gezogene Person an der Krankheit leidet. Dann bilden die Ereignisse K und $\overline{K}$ ein vollständiges System von Ereignissen. Es gilt $P(K) = 0.001$ und $P(\overline{K}) = 0.999$. □

Neben dem Merkmal, das die Population in Teilpopulationen zerlegt, ist ein weiteres Merkmal von Interesse, dessen Wahrscheinlichkeit in jeder der Teilpopulationen bekannt ist. Gesucht ist die Wahrscheinlichkeit des Merkmals in der Population.

Beispiel 69 (fortgesetzt)
Bei der Diagnose wird ein Test verwendet, der in 90 Prozent der Fälle einen Kranken als krank klassifiziert. Weiterhin klassifiziert der Test in 99 Prozent der Fälle einen Gesunden als gesund. Sei A das Ereignis, dass der Test eine Person als krank klassifiziert. Es gilt $P(A|K) = 0.9$ und $P(\overline{A}|\overline{K}) = 0.99$. Also gilt $P(A|\overline{K}) = 0.01$. Gesucht ist $P(A)$. □

Der folgende Satz liefert die Lösung.

Satz 5.16
Sei Ω eine Ergebnismenge. Die Ereignisse $A_1, A_2, \ldots A_n \subset \Omega$ bilden ein vollständiges System von Ereignissen mit $P(A_i) > 0$ für $i = 1, 2, \ldots, n$. Dann gilt für jedes Ereignis $B \subset \Omega$

$$\boxed{P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)\,.}$$

Beweis
Es gilt

$$\Omega = A_1 \cup \ldots \cup A_n$$

und damit wegen Gleichung (5.8)

$$B = B \cap \Omega = B \cap (A_1 \cup \ldots \cup A_n) = (B \cap A_1) \cup \ldots \cup (B \cap A_n)\,.$$

Wegen

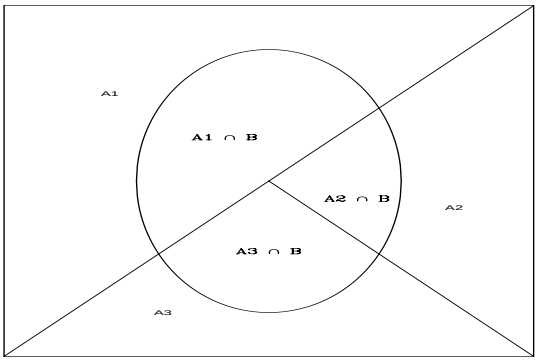
$$A_i \cap A_j = \emptyset$$

für $i \neq j$, gilt auch

$$(B \cap A_i) \cap (B \cap A_j) = \emptyset$$

für $i \neq j$. Abbildung 5.9 veranschaulicht den Sachverhalt. Dabei ist B der Kreis.

Abbildung 5.9: Venn-Diagramm zur Erläuterung



Also folgt

$$\begin{aligned} P(B) &= P((B \cap A_1) \cup \ldots \cup (B \cap A_n)) = P(B \cap A_1) + \ldots + P(B \cap A_n) \\ &= P(B|A_1)\, P(A_1) + \ldots + P(B|A_n)\, P(A_n) = \sum_{i=1}^{n} P(B|A_i) P(A_i)\,. \end{aligned}$$

□

Man spricht vom Satz von der totalen Wahrscheinlichkeit.

Beispiel 69 (fortgesetzt von Seite 225)
Es gilt $P(K) = 0.001$, $P(\overline{K}) = 0.999$, $P(A|K) = 0.9$ und $P(A|\overline{K}) = 0.01$. Also gilt

$$\begin{aligned} P(A) &= P(A|K)P(K) + P(A|\overline{K})P(\overline{K}) = 0.9 \cdot 0.001 + 0.01 \cdot 0.999 \\ &= 0.01089\,. \end{aligned}$$

□

Beispiel 70
Eine Urne enthält N Kugeln, von denen K weiß und die restlichen $N - K$ schwarz sind. Es werden zwei Kugeln nacheinander gezogen.
Wie groß ist die Wahrscheinlichkeit, dass die zweite gezogene Kugel weiß ist?
Seien

W_1: die erste gezogene Kugel ist weiß
W_2: die zweite gezogene Kugel ist weiß

Wir betrachten zunächst das Ziehen **mit** Zurücklegen. Es gilt

$$\begin{aligned} P(W_2) &= P(W_2|W_1)P(W_1) + P(W_2|\overline{W_1})P(\overline{W_1}) \\ &= \frac{K}{N} \cdot \frac{K}{N} + \frac{K}{N} \cdot \frac{N-K}{N} \\ &= \frac{K}{N}\left(\frac{K}{N} + \frac{N-K}{N}\right) \\ &= \frac{K}{N}. \end{aligned}$$

Und nun zum Ziehen **ohne** Zurücklegen. Es gilt

$$\begin{aligned} P(W_2) &= P(W_2|W_1)P(W_1) + P(W_2|\overline{W_1})P(\overline{W_1}) \\ &= \frac{K-1}{N-1} \cdot \frac{K}{N} + \frac{K}{N-1} \cdot \frac{N-K}{N} \\ &= \frac{1}{N(N-1)}\left((K-1)K + K(N-K)\right) \\ &= \frac{1}{N(N-1)}\left(K^2 - K + KN - K^2\right)) \\ &= \frac{1}{N(N-1)}K(N-1) \\ &= \frac{K}{N}. \end{aligned}$$

Wir sehen, dass beim Ziehen mit Zurücklegen und beim Ziehen ohne Zurücklegen die unbedingte Wahrscheinlichkeit für eine weiße Kugel identisch ist, während die bedingten Wahrscheinlichkeiten sich unterscheiden. □

Beispiel 71
Bei vielen Fragen kann man sich nicht sicher sein, dass sie wahrheitsgemäß beantwortet werden. So wird nicht jeder zugeben, dass er Drogen genommen hat oder regelmäßig Alkohol trinkt.
Von Warner wurde ein Verfahren vorgeschlagen, das es erlaubt die Wahrscheinlichkeit einer positiven Antwort zu bestimmen. Dieses Verfahren ist

zweistufig. Wir schauen es uns in der von Hutchinson vorgeschlagenen Form an. Auf der ersten Stufe führt der Befragte ein Zufallsexperiment durch, dessen Ergebnis er nur selber kennt. Wir lassen den Befragten zweimal eine Münze werfen. Fällt beim ersten Wurf Kopf, so soll er die **Frage 1**, ansonsten die **Frage 2** beantworten:

Frage 1: Trinken Sie regelmäßig Alkohol?

Frage 2: Erschien beim zweiten Münzwurf Kopf?

Wir definieren die folgenden Ereignisse:

$F1$	die Frage 1 wird beantwortet
J	die Antwort ist 'ja'

Wir wissen

$$\begin{aligned} P(F1) &= 0.5 \\ P(\overline{F1}) &= 0.5 \\ P(J|\overline{F1}) &= 0.5 \end{aligned}$$

Es gilt

$$\begin{aligned} P(J) &= P(J|F1) \cdot P(F1) + P(J|\overline{F1}) \cdot P(\overline{F1}) \\ &= P(J|F1) \cdot 0.5 + 0.5 \cdot 0.5 \\ &= P(J|F1) \cdot 0.5 + 0.25\,. \end{aligned}$$

Ist $P(J)$ bekannt, so können wir $P(J|F1)$ bestimmen durch

$$P(J|F1) = \frac{P(J) - 0.25}{0.5} = 2 \cdot P(J) - 0.5\,.$$

□

5.2.8 Satz von Bayes

Im Beispiel 69 auf Seite 225 haben wir eine Population in zwei Teilpopulationen zerlegt. In der einen Teilpopulation leiden die Personen an einer bestimmten Krankheit, in der anderen nicht. Außerdem möge ein Test existieren, mit dem man überprüfen kann, ob eine aus der Population ausgewählte Person an der Krankheit leidet. Dieser Test diagnosiziert einen Kranken mit einer

Wahrscheinlichkeit von 0.9 als krank und einen Gesunden mit einer Wahrscheinlichkeit von 0.99 als gesund. In der Praxis ist man an diesen bedingten Wahrscheinlichkeiten aber nicht interessiert. Vielmehr unterzieht sich eine Person dem Test und will aufgrund des Ergebnisses einschätzen, ob sie an der Krankheit leidet oder nicht. Der folgende Satz gibt die Lösung an.

Satz 5.17
Sei Ω eine Ergebnismenge. Die Ereignisse $A_1, A_2, \ldots A_n \subset \Omega$ bilden ein vollständiges System von Ereignissen mit $P(A_i) > 0$ für $i = 1, 2, \ldots, n$. Dann gilt für jedes Ereignis $B \subset \Omega$

$$\boxed{P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum\limits_{i=1}^{n} P(B|A_i)P(A_i)}}\,.$$

Beweis
Wegen

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)}$$

und

$$P(B|A_i) = \frac{P(A_i \cap B)}{P(A_i)}$$

gilt

$$P(A_i|B)P(B) = P(B|A_i)P(A_i)\,.$$

Hieraus folgt

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}\,.$$

Da $A_1, A_2, \ldots, A_n$ ein vollständiges System von Ereignissen bildet, gilt aufgrund von Satz 5.16 auf Seite 225

$$P(B) \;=\; \sum_{i=1}^{n} P(B|A_i)\,P(A_i)$$

und damit

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum\limits_{i=1}^{n} P(B|A_i)P(A_i)}\,.$$

□

Beispiel 69 (fortgesetzt von Seite 226)
Wir wissen

$$\begin{array}{ll} P(K) = 0.001 & P(\overline{K}) = 0.999 \\ P(A|K) = 0.9 & P(A|\overline{K}) = 0.01 \end{array}$$

Gesucht ist $P(K|A)$. Es gilt

$$\begin{aligned} P(K|A) &= \frac{P(A|K)P(K)}{P(A|K)P(K) + P(A|\overline{K})P(\overline{K})} \\ &= \frac{0.9 \cdot 0.001}{0.9 \cdot 0.001 + 0.01 \cdot 0.999} \\ &= 0.083\,. \end{aligned}$$

Diese Wahrscheinlichkeit ist überraschend gering. Woran liegt dies?
Stellen wir uns eine Population von 100000 Personen vor, auf die die obigen Wahrscheinlichkeitsaussagen zutrifft. Somit leiden 100 Personen an der Krankheit, die restlichen 99900 nicht. Von den 100 Personen, die an der Krankheit leiden, werden 90 durch den Test als krank klassifiziert, während von den 99900 Personen, die nicht an der Krankheit leiden, 999 durch den Test als krank klassifiziert werden. Also werden 1089 Personen durch den Test als krank klassifiziert. Unter diesen sind aber nur 99 krank, sodass die gesuchte Wahrscheinlichkeit gleich $90/1089 = 0.083$ ist. Diese Wahrscheinlichkeit ist so gering, da die Krankheit sehr selten ist. □

Das folgende Beispiel stammt wiederum von Kahneman und Tversky.

Beispiel 72
In einer Stadt sind 85 Prozent der Taxis grün. Der Rest ist blau. In einer Nacht ereignet sich ein Unfall mit einem Taxi, wobei der Taxifahrer Fahrerflucht begeht. Ein Zeuge hat den Unfall beobachtet und sagt aus, dass das Taxi blau war. Der Rechtsanwalt der blauen Taxifirma untersucht den Fall genauer. Der Zeuge kann sowohl ein blaues als auch ein grünes Taxi in 80 Prozent der Fälle bei Nacht richtig identifizieren.
Von Interesse ist, wie sicher die Aussage des Zeugen ist. Wir suchen also die Wahrscheinlichkeit, dass das Taxi blau war.
Seien

B: das Taxi ist blau
ZB: der Zeuge stuft die Farbe eines Taxis als blau ein

Gegeben sind:

$$P(B) = 0.15 \qquad P(\overline{B}) = 0.85$$
$$P(ZB|B) = 0.8 \qquad P(\overline{ZB}|\overline{B}) = 0.8$$

Gesucht ist $P(B|ZB)$. Wir wenden den Satz von Bayes an.

$$\begin{aligned} P(B|ZB) &= \frac{P(ZB|B)P(B)}{P(ZB|B)P(B) + P(ZB|\overline{B})P(\overline{B})} \\ &= \frac{0.8 \cdot 0.15}{0.8 \cdot 0.15 + 0.2 \cdot 0.85} \\ &= 0.41\,. \end{aligned}$$

Auch hier können wir so vorgehen, wie bei der Interpretation des Sachverhalts im letzten Beispiel. Stellen wir uns eine Stadt mit 200 Taxis vor. Von diesen sind 30 blau und 170 grün. Von den blauen Taxis werden 24 richtig als blau erkannt, während von den grünen Taxis 34 fälschlich als blau eingestuft werden. Es werden also 58 Taxis für blau gehalten, von denen 24 blau sind. Die gesuchte Wahrscheinlichkeit beträgt somit $24/58 = 0.41$.

□

5.2.9 Unabhängigkeit

Beispiel 73
In einer Vorlesung für Erstsemester sitzen 250 Studierende, von denen 100 weiblich sind. 200 Studierende besitzen einen eigenen PC. Von den Frauen besitzen 80 einen eigenen PC.
Eine Person wird zufällig ausgewählt.
Sei

W: die Person ist weiblich
C: die Person besitzt einen eigenen PC

Es gilt

$$P(C) = \frac{200}{250} = 0.8$$

und

$$P(C|W) = \frac{P(C \cap W}{P(W)} = \frac{0.32}{0.4} = 0.8\,.$$

Das Wissen, dass eine Person weiblich ist, ändert nichts an der Wahrscheinlichkeit, einen eigenen PC zu besitzen. Wir sagen, dass die beiden Merkmale unabhängig sind. □

Definition 5.8
Die Ereignisse A und B heißen unabhängig, wenn gilt

$$\boxed{P(A|B) = P(A)}\,.$$

Satz 5.18
Sind die Ereignisse A und B unabhängig, so gilt

$$\boxed{P(A \cap B) = P(A) \cdot P(B)}\,.$$

Beweis:
Aus

$$P(A|B) = P(A)$$

folgt

$$P(A \cap B) = P(A|B)P(B) = P(A)P(B)\,.$$

□

Sind die Ereignisse A und B unabhängig, so benötigt man nur die Wahrscheinlichkeiten $P(A)$ und $P(B)$, um die Wahrscheinlichkeit zu bestimmen, dass A und B gleichzeitig eintreten.

Satz 5.19
Sind die Ereignisse A und B unabhängig, so sind auch die folgenden Paare von Ereignisse unabhängig:

1. A und $\overline{B}$
2. $\overline{A}$ und B
3. $\overline{A}$ und $\overline{B}$

Beweis:
Wir zeigen nur 1.. Die anderen Beweise verlaufen analog.
Es ist zu zeigen

$$P(A \cap \overline{B}) = P(A)P(\overline{B})\,.$$

Es gilt

$$\begin{aligned} P(A \cap \overline{B}) &= P(A) - P(A \cap B) = P(A) - P(A)P(B) = P(A)(1 - P(B)) \\ &= P(A)P(\overline{B})\,. \end{aligned}$$

□

Der folgende Satz zeigt, wie die Begriffe Disjunktheit und Unabhängigkeit zusammenhängen.

Satz 5.20
Sind A und B disjunkt und gilt $P(A) > 0$ und $P(B) > 0$, so sind A und B nicht unabhängig.
Beweis:
Aus

$$A \cap B = \emptyset$$

folgt

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = 0\,.$$

Da gilt $P(A) > 0$ folgt

$$P(A|B) \neq P(A)\,.$$

□

Die Aussage des Satzes ist auch intuitiv klar:
Sind die Ereignisse disjunkt, so können sie nicht gleichzeitig eintreten, und sind somit im höchsten Maße abhängig. Tritt nämlich A ein, so tritt B nicht ein und umgekehrt.
Wir schauen uns noch ein Beispiel an, das unabhängige Ereignisse verwendet.

Beispiel 74
Oft werden Entscheidungen durch einen Münzwurf getroffen. Hierbei wird unterstellt, dass die verwendete Münze fair ist. Dies muss aber nicht der Fall sein.
John von Neumann hat ein Verfahren vorgeschlagen, bei dem man sich mit gleicher Wahrscheinlichkeit für eine der beiden Alternativen entscheidet.
Die Münze wird zweimal hintereinander geworfen. Fällt beim ersten Mal Kopf und beim zweiten Mal Zahl, so entscheidet man sich für Alternative 1. Fällt beim ersten Mal Zahl und beim zweiten Mal Kopf, so entscheidet man sich für Alternative 2. Fällt bei beiden Würfen das gleiche Symbol, so wird die Münze wiederum zweimal geworfen, und es wird genauso verfahren wie bei der ersten Runde. Die ganze Prozedur wird solange durchgeführt, bis zum ersten Mal zwei unterschiedliche Symbole auftreten.
Inwiefern ist diese Prozedur fair?
Sei K_i das Ereignis, dass beim i-ten Wurf Kopf fällt, i=1,2. Wir unterstellen, dass die Wahrscheinlichkeit für Kopf bei beiden Würfen gleich ist, und dass die Münze kein Gedächtnis hat. Sei $P(K_i) = p$ die Wahrscheinlichkeit für

Kopf beim i-ten Wurf. Dann gilt wegen der Unabhängigkeit

$$P(K_1 \cap \overline{K_2}) = P(K_1) \cdot P(\overline{K_2}) = p \cdot (1-p)$$

und

$$P(\overline{K_1} \cap K_2) = P(\overline{K_1}) \cdot P(K_2) = (1-p) \cdot p\,.$$

Wir sehen, dass die Wahrscheinlichkeiten der beiden Alternativen bei jeder Runde gleich sind.
Ein Problem hat die Prozedur jedoch:
Es kann sehr lange dauern, bis eine Entscheidung fällt. □

Kapitel 6

Univariate Zufallsvariablen

Im ersten Teil dieses Skriptes haben wir uns mit Daten beschäftigt und gezeigt, wie man die Verteilung eines Merkmals beschreiben kann. Ist man nur an der Population interessiert, die zu den Daten gehört, so ist eine solche Analyse unproblematisch. Oft ist man aber an der Verteilung eines Merkmals in einer Grundgesamtheit interessiert und kann aber nicht die gesamte Grundgesamtheit untersuchen. Man wird in diesem Fall eine Teilgesamtheit der Grundgesamtheit untersuchen. Der Schluss von der Teilgesamtheit auf die Grundgesamtheit ist fehlerbehaftet. Die Wahrscheinlichkeitsrechnung erlaubt es nun, eine Aussage über den Fehler zu machen. Hierzu unterstellt man für das Merkmal ein Wahrscheinlichkeitsmodell und spricht von einer Zufallsvariablen. Wir unterscheiden diskrete und stetige Zufallsvariablen.

6.1 Diskrete Zufallsvariablen

Beispiel 75
Wir betrachten Familien mit zwei Kindern, wobei uns die Anzahl der Mädchen interessiert. Offensichtlich kann es in einer Familie mit zwei Kindern 0, 1 oder 2 Mädchen geben. Wir suchen die Wahrscheinlichkeiten dieser drei Ausprägungen. Hierzu fassen wir die Geburt eines Kindes als einen Zufallsvorgang auf. Wir beobachten diesen Zufallsvorgang zweimal. Man spricht von einem verbundenen Zufallsvorgang. Die Ergebnismenge ist

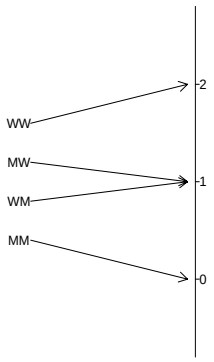
$$\Omega = \{WW, WM, MW, MM\}\,.$$

Dabei steht W für weiblich und M für männlich. Es liegt nahe, ein Gleichmöglichkeitsmodell zu unterstellen. Es gelte also

$$P(\{WW\}) = 0.25 \qquad P(\{WM\}) = 0.25$$
$$P(\{MW\}) = 0.25 \qquad P(\{MM\}) = 0.25$$

Wir ordnen nun jedem Ergebnis die Anzahl der Mädchen zu. So sind 2 Mädchen in der Familie, wenn das Ergebnis WW beobachtet wurde. Abbildung 6.1 illustriert die Zuordnung.

Abbildung 6.1: Illustration einer Zufallsvariablen



□

Definition 6.1
Sei Ω die abzählbare Ergebnismenge eines Zufallvorgangs. Dann heißt die Abbildung $X : \Omega \to I\!R$ diskrete Zufallsvariable X.

Für Zufallsvariablen verwenden wir im Folgenden Großbuchstaben. Die Werte, die die Zufallsvariable annimmt, bezeichnen wir mit Kleinbuchstaben.

Beispiel 75 (fortgesetzt)
Wir betrachten die Zufallsvariable X Anzahl der Mädchen in Familien mit 2 Kindern. X kann die Werte 0, 1 und 2 annehmen. □

Haben wir den Elementarereignissen Wahrscheinlichkeiten zugeordnet, so können wir die Wahrscheinlichkeiten von Ereignissen bestimmen, die über die Zufallsvariable X beschrieben werden. Hier interessieren uns vor allem Ereignisse der Form $\{\omega | X(\omega) = x\}$.

Beispiel 75 (fortgesetzt)
Es gilt

$$\begin{aligned} \{\omega | X(\omega) = 0\} &= \{MM\} \\ \{\omega | X(\omega) = 1\} &= \{MW, WM\} \\ \{\omega | X(\omega) = 2\} &= \{WW\}\,. \end{aligned}$$

□

Die Wahrscheinlichkeiten von Ereignissen der Form $\{\omega|X(\omega) = x\}$ können wir folgendermaßen bestimmen:

$$P(\{\omega|X(\omega) = x\}) = \sum_{\{\omega_i|X(\omega_i)=x\}} P(\{\omega_i\})\,.$$

Für

$$P(\{\omega|X(\omega) = x\})$$

schreiben wir kurz

$$P(X = x).$$

Beispiel 75 (fortgesetzt)
Es gilt

$$\begin{aligned} P(X=0) &= P(\{\omega|X(\omega)=0\}) = P(\{MM\}) = 0.25 \\ P(X=1) &= P(\{\omega|X(\omega)=1\}) = P(\{MW\}) + P(\{WM\}) = 0.5 \\ P(X=2) &= P(\{\omega|X(\omega)=2\}) = P(\{WW\}) = 0.25\,. \end{aligned}$$

□

Definition 6.2
Sei X eine diskrete Zufallsvariable. Dann heißt die Funktion $f_X : \mathbb{R} \to \mathbb{R}$ mit $x \mapsto f_X(x) = P(X = x)$ Wahrscheinlichkeitsfunktion von X.

Beispiel 75 (fortgesetzt)
Es gilt

$$f_X(x) = \begin{cases} 0.25 & \text{für } x = 0 \\ 0.5 & \text{für } x = 1 \\ 0.25 & \text{für } x = 2 \\ 0 & \text{sonst} \end{cases}$$

□

$\mathcal{T}_\mathcal{X} = \{x|f_X(x) > 0\}$ heißt der **Träger** von x.

Beispiel 75 (fortgesetzt)
Es gilt $\mathcal{T}_\mathcal{X} = \{0, 1, 2\}$. □

Die Wahrscheinlichkeitsfunktion $f_X(x)$ besitzt zwei Eigenschaften. Für alle $x \in \mathbb{R}$ gilt

$$\boxed{f_X(x) \geq 0} \,. \tag{6.1}$$

Außerdem gilt

$$\boxed{\sum_{\{x|x\in\mathcal{T}_\mathcal{X}\}} f_X(x) = 1} \tag{6.2}$$

Man muss die Wahrscheinlichkeitsfunktion einer diskreten Zufallsvariablen X nicht notwendigerweise aus einem Zufallsvorgang gewinnen. Jede Funktion, die die Bedingungen in den Gleichungen (6.2) und (6.1) erfüllt, kann als Wahrscheinlichkeitsfunktion einer Zufallsvariablen X aufgefasst werden. Es muss dann natürlich überprüft werden, ob die Wahl sinnvoll ist.

Beispiel 76
Ein Statistiker betrachtet alle Bundesligaspiele, in denen höchstens 5 Tore fallen. Er unterstellt eine Wahrscheinlichkeitsfunktion für die Anzahl der Tore X, die Tabelle 6.1 zu finden ist.

Tabelle 6.1: Wahrscheinlichkeitsfunktion für die Anzahl der Tore in einem Bundesligaspiel

x	0	1	2	3	4	5
$P(X = x)$	0.1	0.2	0.3	0.2	0.1	0.1

Um zu sehen, ob die Wahrscheinlichkeitsfunktion sinnvoll gewählt wurde, sollte man sie mit Daten konfrontieren. Der Statistiker betrachtet alle Spiele der Saison 2001/2002, in denen höchstens 5 Tore gefallen sind, und bestimmt die relative Häufigkeit $h(X = x)$ der Spiele, in denen x Tore gefallen sind. Diese sind in Tabelle 6.2 zu finden. In Statistik II werden wir Verfahren kennen lernen, mit denen wir überprüfen können, ob ein Wahrscheinlichkeitsmodell angemessen ist.

Tabelle 6.2: Häufigkeitstabelle der Anzahl der Tore in einem Bundesligaspiel in der Saison 2001/2002, wobei nur Spiele betrachtet wurden, in denen höchstens 5 Tore fielen

x	0	1	2	3	4	5
$n(X = x)$	23	35	74	69	56	23
$h(X = x)$	0.082	0.125	0.264	0.246	0.200	0.082

□

Beispiel 77
Der Mathematiker Simon Newcomb bemerkte im Jahr 1881, dass in Logarithmentabellen die vorderen Seiten abgegriffener waren als die hinteren. Dies deutet darauf hin, dass Zahlen mit einer niedrigen Anfangsziffer häufiger sind als Zahlen mit einer hohen Anfangsziffer. Newcomb leitete folgende Wahrscheinlichkeit dafür her, dass die erste Ziffer X den Wert x annimmt:

$$P(X = x) = \log(x + 1) - \log(x).$$

Dabei ist log der Logarithmus zur Basis 10.
Tabelle 6.3 zeigt die Verteilung der Anfangsziffern der Einwohnerzahl deutscher Städte zusammen mit den Wahrscheinlichkeiten der 9 Ziffern nach dem Benford-Gesetz.

Tabelle 6.3: Verteilung der der Anfangsziffern der Einwohnerzahl deutscher Städte

x	$n(X = x)$	$h(X = x)$	$P(X = x)$
1	128	0.317	0.3010
2	89	0.220	0.1761
3	59	0.146	0.1249
4	32	0.079	0.0969
5	36	0.089	0.0792
6	18	0.045	0.0669
7	18	0.045	0.0580
8	15	0.037	0.0512
9	9	0.022	0.0458

Wir sehen, dass die empirische Verteilung gut mit der theoretischen übereinstimmt. 60 Jahre nach Newcomb entdeckte der amerikanische Physiker Benford das Phänomen ebenfalls und stellte fest, dass eine Vielzahl von Datensätzen dieses Gesetz erfüllen. Nach ihm heißt es Benford-Gesetz. □

Sehr oft hängt die Wahrscheinlichkeitsfunktion von einer oder mehreren Größen ab, die wir Parameter nennen.

Beispiel 78
Wir betrachten einen Produktionsprozess. Ein Produkt kann entweder defekt oder nicht defekt sein. Dem Produktionsprozess werden zwei Produkte entnommen. Wir sind an der Anzahl X der defekten Produkte interessiert und suchen eine geeignete Wahrscheinlichkeitsfunktion für X.
Wir beobachten zweimal den gleichen Zufallsvorgang. Man spricht von einem verbundenen Zufallsvorgang. Sei D_i, $i = 1, 2$, das Ereignis, dass das i-te entnommene Produkt defekt ist. Wir nehmen an, dass die Wahrscheinlichkeit eines defekten Produktes bei beiden Zufallsvorgängen identisch ist. Es gilt also $P(D_i) = p$ für $i = 1, 2$ mit $0 \leq p \leq 1$. Außerdem seien die beiden Zufallsvorgänge unabhängig. Dies heißt, dass alle Ereignisse des einen Zufallsvorgangs unabhängig von allen Ereignissen des anderen Zufallsvorgangs sind. Es gilt also

$$\begin{aligned}
P(D_1 \cap D_2) &= P(D_1)P(D_2) = p^2 \\
P(D_1 \cap \overline{D_2}) &= P(D_1)P(\overline{D_2}) = p(1-p) \\
P(\overline{D_1} \cap D_2) &= P(\overline{D_1})P(D_2) = (1-p)p \\
P(\overline{D_1} \cap \overline{D_2}) &= P(\overline{D_1})P(\overline{D_2}) = (1-p)^2
\end{aligned}$$

Schauen wir uns nun die Wahrscheinlichkeitsfunktion der Zufallsvariablen X an. Es gilt

$$\begin{aligned}
P(X=0) &= P(\overline{D_1} \cap \overline{D_2}) = (1-p)^2 \\
P(X=1) &= P(D_1 \cap \overline{D_2}) + P(\overline{D_1} \cap D_2) = p(1-p) + (1-p)p \\
&= 2p(1-p) \\
P(X=2) &= P(D_1 \cap D_2) = p^2
\end{aligned}$$

Die Wahrscheinlichkeitsfunktion von X hängt von dem unbekannten Parameter p ab. Im Rahmen der schließenden Statistik werden wir lernen, wie man datengestützt geeignete Werte für p finden kann. Man spricht vom Schätzen von p. □

Neben Ereignissen der Form $\{\omega|X(\omega) = x\}$ betrachten wir noch Ereignisse der Form $\{\omega|X(\omega) \leq x\}$.

Definition 6.3
Sei X eine diskrete Zufallsvariable. Dann heißt

$$\boxed{F_X(x) = P(\{\omega|X(\omega) \leq x\})} \tag{6.3}$$

die Verteilungsfunktion von X.

Für

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

schreiben wir

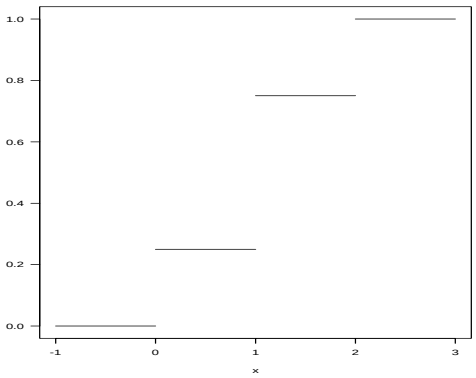
$$\boxed{F_X(x) = P(X \leq x)\,.}$$

Beispiel 75 (fortgesetzt von Seite 237)
Die Verteilungsfunktion $F_X(x)$ von X ist gegeben durch

$$F_X(x) = \begin{cases} 0 & \text{für} \quad x < 0 \\ 0.25 & \text{für} \quad 0 \leq x < 1 \\ 0.75 & \text{für} \quad 1 \leq x < 2 \\ 1 & \text{für} \quad x \geq 2 \end{cases}$$

Abbildung 6.2 zeigt die Verteilungsfunktion.

Abbildung 6.2: Verteilungsfunktion einer diskreten Zufallsvariablen



Wir sehen, dass sie eine Treppenfunktion ist, die monoton wächst. □

Die Verteilungsfunktion $F_X(x)$ einer Zufallsvariablen X besitzt folgende Eigenschaften, die wir ohne Beweis angeben.

$F_X(x)$ ist monoton wachsend.

$F_X(x)$ ist rechtsseitig stetig.

$\lim_{x\to-\infty} F_X(x)=0$

$\lim_{x\to\infty} F_X(x)=1$

Mit Hilfe der Verteilungsfunktion kann man unter anderem folgende Wahrscheinlichkeiten bestimmen.

$$P(X = a) = F_X(a) - \lim_{x\uparrow a} F_X(x)$$

$$P(X \leq a) = F_X(a)$$

$$P(X < a) = F_X(a) - P(X = a)$$

$$P(X > a) = 1 - F_X(a)$$

$$P(a < X \leq b) = F_X(b) - F_X(a)$$

$$P(a \leq X \leq b) = F_X(b) - F_X(a) + P(X = a)$$

$$P(a \leq X < b) = F_X(b) - F_X(a) - P(X = b) + P(X = a)$$

$$P(a < X < b) = F_X(b) - F_X(a) - P(X = b)$$

Beispiel 75 (fortgesetzt)
Es gilt

$$P(X \leq 2) = F_X(2) = 0.6$$

□

Oft ist man an einer Funktion einer Zufallsvariablen interessiert. Das folgende Beispiel stammt aus dem unveröffentlichten Skript zur Vorlesung Statistik nach der Grundausbildung von Bernd Streitberg.

Beispiel 79
Ein Teilchen bewegt sich auf den ganzen Zahlen, wobei es im Nullpunkt startet. Bei jedem Schritt geht es zufällig nach rechts oder links. Das Teilchen möge drei Schritte machen. Uns interessiert die Anzahl der Schritte

nach links. Offensichtlich kann es keinen, einen oder zwei Schritte nach links machen. Die Ergebnismenge ist

$$\Omega = \{LLL, LLR, LRL, RLL, LRR, RLR, RRL, RRR\}$$

Da das Teilchen sich zufällig bewegt, ist jedes der Elementarereignisse gleich wahrscheinlich. Es gilt also

$$\begin{aligned} P(\{LLL\}) &= 0.125 & P(\{LLR\}) &= 0.125 \\ P(\{LRL\}) &= 0.125 & P(\{LRR\}) &= 0.125 \\ P(\{LRR\}) &= 0.125 & P(\{RLR\}) &= 0.125 \\ P(\{RRL\}) &= 0.125 & P(\{RRR\}) &= 0.125 \end{aligned}$$

Wir betrachten nun die Position X des Teilchens nach 3 Schritten. Geht das Teilchen zum Beispiel dreimal nach rechts, so befindet es sich auf der 3. Tabelle 6.4 gibt für jedes Ergebnis $\omega \in \Omega$ den Wert der Zufallsvariablen X an.

Tabelle 6.4: Ergebnisse und zugehörige Werte einer Zufallsvariablen

ω	RRR	RRL	RLR	LRR	RLL	LRL	LLR	LLL
x	3	1	1	1	-1	-1	-1	-3

Da das Teilchen sich zufällig bewegt, ist jedes Elementarereignis gleich wahrscheinlich. Somit erhalten wir folgende Wahrscheinlichkeitsfunktion von X:

x	-3	-1	1	3
$P(X = x)$	0.125	0.375	0.375	0.125

Uns interessiert nun die Verteilung von $Y = |X|$, dem Abstand des Teilchens vom Nullpunkt. Die Zufallsvariable Y kann die Werte 1 und 3 annehmen. Es gilt

$$P(Y = 1) = P(X = -1) + P(X = 1) = 0.75$$

$$P(Y = 3) = P(X = -3) + P(X = 3) = 0.25$$

□

Wie das Beispiel zeigt, kann man die Verteilung einer Funktion $Y = g(X)$ einer diskreten Zufallsvariablen X folgendermaßen bestimmen:

$$P(Y = y) = \sum_{\{x|g(x)=y\}} P(X = x) \tag{6.4}$$

6.2 Stetige Zufallsvariablen

Im Beispiel 14 auf Seite 72 haben wir das stetige Merkmal Alter betrachtet. Bei einem stetigen Merkmal bilden wir Klassen und bestimmen die relativen Häufigkeiten der Klassen. Die Häufigkeitsverteilung stellen wir mit einem Histogramm dar. Abbildung 3.9 auf Seite 75 zeigt das Histogramm des Alters. Das Histogramm ist die graphische Darstellung der empirischen Dichtefunktion $\hat{f} : I\!R \to I\!R$. Für alle $x \in I\!R$ gilt

$$\hat{f}(x) \geq 0\,. \tag{6.5}$$

Außerdem gilt:

$$\int_{-\infty}^{\infty} \hat{f}(x)\,dx = 1\,. \tag{6.6}$$

Die Wert der empirischen Verteilungsfunktion $\hat{F}(x)$ an der Stelle x ist gleich der Fläche unter der empirischen Dichtefunktion bis zur Stelle x. Es gilt also

$$\hat{F}(x) = \int_{-\infty}^{x} \hat{f}(u)\,du\,. \tag{6.7}$$

Wir können nun in Analogie zu dieser Eigenschaft eine stetige Zufallsvariable über die Verteilungsfunktion $F_X(x) = P(X \leq x)$ definieren.

Definition 6.4
Eine Zufallsvariable X heißt stetig, wenn eine Funktion $f_X : I\!R \to I\!R$ existiert, sodass für die Verteilungsfunktion $F_X(x)$ von X gilt:

$$\boxed{F_X(x) = \int_{-\infty}^{x} f(u)\,du} \tag{6.8}$$

Die Funktion $f_X(x)$ heißt Dichtefunktion der Zufallsvariablen X. Für alle $x \in \mathbb{R}$ gilt

$$\boxed{f_X(x) \geq 0}\,. \tag{6.9}$$

Außerdem gilt:

$$\boxed{\int_{-\infty}^{\infty} f_X(x)\,dx = 1}\,. \tag{6.10}$$

Jede Funktion, die die Bedingungen in den Gleichungen (6.9) und (6.10) erfüllt, kann als Dichtefunktion einer stetigen Zufallsvariablen aufgefasst werden.

Beispiel 80
Gegeben sei folgende Funktion $f : \mathbb{R} \to \mathbb{R}$ mit

$$f(x) = \begin{cases} 0.1 & \text{für} \quad 0 \leq x \leq 10 \\ 0 & \text{sonst} \end{cases}$$

Offensichtlich gilt

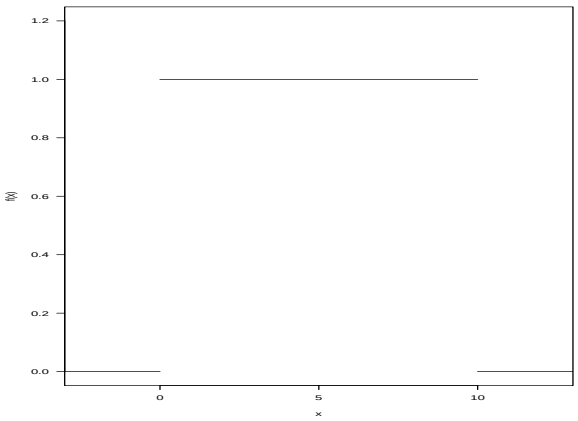
$$f(x) \geq 0 \quad \text{für alle} \quad x \in \mathbb{R}\,.$$

Außerdem gilt

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{0}^{10} 0.1\,dx = \Big[0.1\,u\Big]_0^{10} = 1 - 0 = 1\,.$$

Es handelt sich also um die Dichtefunktion einer Zufallsvariablen. Dies ist die Dichtefunktion einer auf $[0, 10]$ gleichverteilten Zufallsvariablen.
Abbildung 6.3 zeigt die Dichtefunktion.

Abbildung 6.3: Dichtefunktionfunktion der Gleichverteilung auf $[0, 10]$



Für die Verteilungsfunktion $F_X(x)$ gilt $F_X(x) = 0$ für $x < 0$ und $F_X(x) = 1$ für $x > 10$. Für $0 \leq x \leq 10$ gilt:

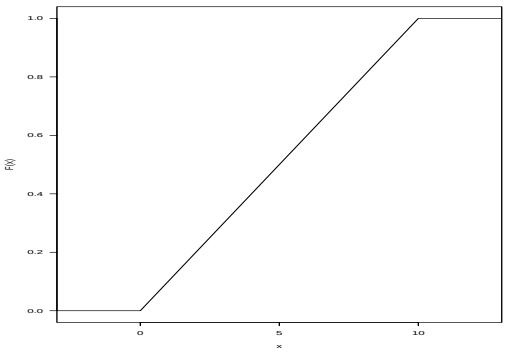
$$F_X(x) \quad = \quad \int_{-\infty}^{x} f(u)\,du = \int_{0}^{x} 0.1\,du = \Big[0.1\,u\Big]_0^x = 0.1\,x\,.$$

Also gilt

$$F_X(x) = \begin{cases} 0 & \text{für } x < 0 \\ 0.1\,x & \text{für } 0 \leq x \leq 10 \\ 1 & \text{für } x > 10 \end{cases}$$

Abbildung 6.4 zeigt die Verteilungsfunktion.

Abbildung 6.4: Verteilungsfunktion der Gleichverteilung



□

Beispiel 81
Gegeben sei folgende Funktion $f : \mathbb{R} \to \mathbb{R}$ mit

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

Offensichtlich gilt

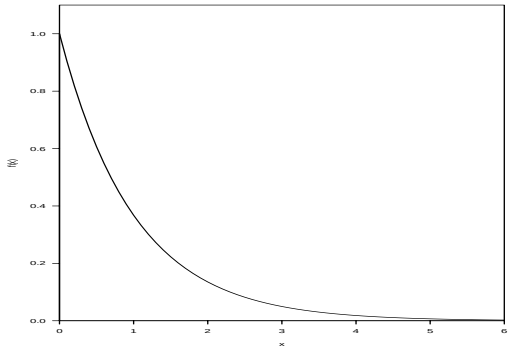
$$f(x) \geq 0 \quad \text{für alle} \quad x \in \mathbb{R}\,.$$

Außerdem gilt

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{0}^{\infty} e^{-x}\,dx = \Big[-e^{-x} \Big]_0^{\infty} = 0 - (-1) = 1\,.$$

Es handelt sich also um die Dichtefunktion einer Zufallsvariablen. Dies ist die Dichtefunktion einer mit Parameter $\lambda = 1$ exponentialverteilten Zufallsvariablen.
Abbildung 6.5 zeigt die Dichtefunktion.

Abbildung 6.5: Dichtefunktionfunktion der Exponentialverteilung mit Parameter $\lambda = 1$



Für die Verteilungsfunktion $F_X(x)$ gilt $F_X(x) = 0$ für $x < 0$. Für $x \geq 0$ gilt:

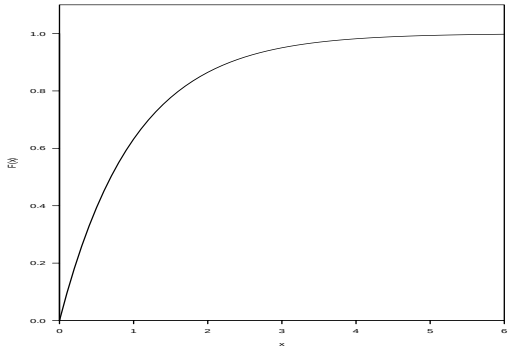
$$F_X(x) \;=\; \int_{-\infty}^{x} f(u)\,du = \int_{0}^{x} e^{-u}\,du = \Big[-e^{-u}\Big]_0^x = -e^{-x} - (-1) = 1 - e^{-x}\,.$$

Also gilt

$$F_X(x) = \begin{cases} 0 & \text{für } x < 0 \\ 1 - e^{-x} & \text{für } x \geq 0 \end{cases}$$

Abbildung 6.6 zeigt die Verteilungsfunktion.

Abbildung 6.6: Verteilungsfunktion der Exponentialverteilung mit $\lambda = 1$



□

Mit Hilfe der Verteilungsfunktion kann man wie bei einer stetigen Zufallsvariablen die relevanten Wahrscheinlichkeiten mit den Formeln auf Seite 242 bestimmen. Dabei ergeben sich einige Vereinfachungen. Die Verteilungsfunktion $F_X(x)$ einer stetigen Zufallsvariablen X ist eine stetige Funktion. Hieraus folgt, dass für eine stetige Zufallsvariabe X für alle $x \in \mathbb{R}$ gilt

$$P(X = x) = 0,$$

denn

$$P(X = x) \quad = \quad F_X(x) - \lim_{u \uparrow x} F_X(u) = F_X(x) - F_X(x) = 0\,.$$

Somit gilt bei einer stetigen Zufallsvariablen:

$$\begin{aligned}
&P(X < a) = F_X(a) - P(X = a)\\
&P(X > a) = 1 - F_X(a)\\
&P(a < X \leq b) = F_X(b) - F_X(a)\\
&P(a \leq X \leq b) = F_X(b) - F_X(a)\\
&P(a \leq X < b) = F_X(b) - F_X(a)\\
&P(a < X < b) = F_X(b) - F_X(a)
\end{aligned}$$

Beispiel 80 (fortgesetzt)
Es gilt

$$P(2 < X < 6) = F_X(6) - F_X(2) = 0.6 - 0.2 = 0.4$$

□

Kapitel 7

Verteilungsparameter

Wie bei einem Merkmal wollen wir nun die Lage und die Streuung der Verteilung einer diskreten Zufallsvariablen durch geeignete Maßzahlen beschreiben. Beginnen wir mit Maßzahlen für die Lage.

7.1 Der Erwartungswert

7.1.1 Diskrete Zufallsvariablen

In der Datenanalyse haben wir den Mittelwert eines diskreten Merkmals mit den Merkmalsausprägungen $a_1, \ldots, a_k$ und zugehörigen relativen Häufigkeiten $h_1, \ldots, h_k$ folgendermaßen berechnet:

$$\bar{x} = \sum_{i=1}^{k} a_i \, h_i \, .$$

Die folgende Definition überträgt dieses Konzept auf eine diskrete Zufallsvariable X mit Wahrscheinlichkeitsfunktion $f_X(x)$.

Definition 7.1
Sei X eine diskrete Zufallsvariable mit Wahrscheinlichkeitsfunktion $f_X(x)$ und Träger $\mathcal{T}_X$. Der Erwartungswert $E(X)$ von X ist definiert durch

$$\boxed{E(X) = \sum_{\{x|x\in\mathcal{T}_X\}} x \, f_X(x)} \tag{7.1}$$

Beispiel 75 (fortgesetzt von Seite 241)
] Wir betrachten die Anzahl X Mädchen in Familien mit zwei Kindern. Die Wahrscheinlichkeitsverteilung ist auf Seite auf Seite 237 zu finden. Sie lautet

$$\begin{aligned} P(X=0) &= 0.25 \\ P(X=1) &= 0.5 \\ P(X=2) &= 0.25\,. \end{aligned}$$

Somit gilt

$$E(X) = 0 \cdot 0.25 + 1 \cdot 0.5 + 2 \cdot 0.25 = 1\,.$$

□

Beispiel 82
Tversky und Kahneman fragten Personen, welche der beiden folgenden Alternativen sie vorzögen.

Alternative A

Man erhält eine sichere Auszahlung von \$ 240.

Alternative B

Mit Wahrscheinlichkeit 0.25 erhält man eine Auszahlung von \$ 1000 und mit Wahrscheinlichkeit 0.75 keine Auszahlung.
Wir wollen zunächst untersuchen, ob eine der beiden Situationen günstiger ist. Die Konsequenz von Alternative A ist klar. Wir erhalten auf jeden Fall \$ 240.
Wir betrachten für Alternative B die Auszahlung X. Es gilt

$$P(X=0) = 0.75 \qquad P(X=1000) = 0.25\,.$$

Es gilt

$$E(X) = 0 \cdot 0.75 + 1000 \cdot 0.25 = 250\,.$$

Bei Alternative A haben wir eine sichere Auszahlung von \$ 240 und bei Alternative B eine erwartete Auszahlung von \$ 250. Obwohl die erwartete Auszahlung bei Alternative B höher ist, entschieden sich 84 Prozent der von Tversky und Kahneman befragten Personen für Alternative A.

7.1.2 Stetige Zufallsvariablen

Definition 7.2
Sei X eine stetige Zufallsvariable mit Dichtefunktion $f_X(x)$. Der Erwartungswert $E(X)$ von X ist definiert durch

$$\boxed{E(X) = \int_{-\infty}^{\infty} x\, f_X(x)\, dx\,.} \tag{7.2}$$

Beispiel 80 (fortgesetzt von Seite 245)
Es gilt

$$E(X) \;=\; \int_{-\infty}^{\infty} x\, f_X(x)\, dx = \int_{0}^{10} x\, 0.1\, dx = \Big[0.1\, \frac{x^2}{2}\Big]_0^{10} = 5\,.$$

□

Beispiel 81 (fortgesetzt von Seite 247)
Es gilt

$$E(X) \;=\; \int_{-\infty}^{\infty} x\, f_X(x)\, dx = \int_{0}^{\infty} x\, e^{-x}\, dx = \Big[-x\, e^{-x}\Big]_0^{\infty} - \int_{0}^{\infty} -e^{-x}\, dx$$

$$= \lim_{x\to\infty} \frac{-x}{e^x} + \Big[-e^{-x}\Big]_0^{\infty} = \lim_{x\to\infty} \frac{-1}{e^x} + (0-(-1)) = 1$$

□

Der Erwartungswert einer Zufallsvariablen muss nicht existieren.

Beispiel 83
Die Zufallsvariable X besitze die Dichtefunktion

$$f_X(x) = \begin{cases} \dfrac{1}{x^2} & \text{für} \quad x > 1 \\[2ex] 0 & \text{sonst} \end{cases}$$

Der Erwartungswert von X existiert nicht. Dies sieht man folgendermaßen:

$$\int_{1}^{\infty} x\, \frac{1}{x^2}\, dx = \Big[\ln x\Big]_1^{\infty} \to \infty$$

7.1.3 Erwartungswerte von Funktionen von Zufallsvariablen

Ist X eine diskrete Zufallsvariable mit Wahrscheinlichkeitsfunktion $f_X(x)$ und $g(X)$ eine Funktion von X. Dann können wir den Erwartungswert von $g(X)$ dadurch bestimmen, dass wir die Wahrscheinlichkeitsfunktion von $g(X)$ und über diese den Erwartungswert von $g(X)$ bestimmen. Wir können den Erwartungswert von $g(X)$ aber auch bestimmen, ohne die Wahrscheinlichkeitsfunktion von $g(X)$ herzuleiten. Es gilt nämlich

$$\boxed{E(g(X)) = \sum_{\{x|x\in\mathcal{T}_X\}} g(x)\, f_X(x)} \tag{7.3}$$

Beispiel 79 (fortgesetzt von Seite 242)
Wir betrachten nun noch einmal die Position X des Teilchens, das sich auf den ganzen Zahlen bewegt. Es gilt

x	-3	-1	1	3
$P(X=x)$	0.125	0.375	0.375	0.125

Hieraus folgt

$$E(|X|) \;=\; |-3|\cdot 0.125 + |-1|\cdot 0.375 + |1|\cdot 0.375 + |3|\cdot 0.125 = 1.5\,.$$

Wir können diesen auch mit der Wahrscheinlichkeitsfunktion von $Y = |X|$ bestimmen. Es gilt $P(Y=1) = 0.75$ und $P(Y=3) = 0.25$. Hieraus folgt

$$E(Y) = 1\cdot 0.75 + 3\cdot 0.25 = 1.5\,.$$

□

Bei einer stetigen Zufallsvariablen X mit Dichtefunktion $f_X(x)$ bestimmen wir den Erwartungswert einer Funktion $g(X)$ folgendermaßen:

$$\boxed{E(g(X)) = \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx\,.} \tag{7.4}$$

Beispiel 80 (fortgesetzt von Seite 245)
Wir suchen den Erwartungswert von X^2. Es gilt

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x)\,dx = \int_0^{10} x^2\, 0.1\,dx = \Big[0.1\frac{x^3}{3}\Big]_0^{10} = \frac{100}{3}\,.$$

□

Beispiel 81 (fortgesetzt von Seite 247)
Es gilt

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x)\,dx = \int_0^{\infty} x^2\, e^{-x}\,dx = \Big[-x^2\, e^{-x}\Big]_0^{\infty} - \int_0^{\infty} 2x(-e^{-x})\,dx$$

$$= \lim_{x\to\infty} -\frac{x^2}{e^x} + 2\int_0^{\infty} xe^{-x}\,dx = \lim_{x\to\infty} -\frac{2x}{e^x} + 2 = \lim_{x\to\infty} -\frac{2}{e^x} + 2 = 2$$

□

7.1.4 Eigenschaften des Erwartungswerts

Der folgende Satz zeigt, wie sich der Erwartungswert unter linearen Transformationen verhält.

Satz 7.1
Sei X eine Zufallsvariable mit Wahrscheinlichkeitsfunktion bzw. Dichtefunktion $f_X(x)$ und a und b reelle Zahlen. Dann gilt

$$\boxed{E(a\,X + b) = a\,E(X) + b}\,. \tag{7.5}$$

Beweis:
Wir zeigen den diskreten Fall.

$$\begin{aligned} E(a\,X+b) &= \sum_{\{x|x\in\mathcal{T}_{\mathcal{X}}\}} (a\,x+b)\,f_X(x) = \sum_{\{x|x\in\mathcal{T}_{\mathcal{X}}\}} (a\,x\,f_X(x)+b\,f_X(x)) \\ &= \sum_{\{x|x\in\mathcal{T}_{\mathcal{X}}\}} a\,x\,f_X(x) + \sum_{\{x|x\in\mathcal{T}_{\mathcal{X}}\}} b\,f_X(x) \\ &= a\sum_{\{x|x\in\mathcal{T}_{\mathcal{X}}\}} x\,f_X(x) + b\sum_{\{x|x\in\mathcal{T}_{\mathcal{X}}\}} f_X(x) \\ &= a\,E(X)+b \end{aligned}$$

□

Mit

$$g(x) = a\,x+b$$

gilt also

$$E(g(X)) = g(E(X)).$$

Dies gilt aber nicht immer, wie das folgende Beispiel zeigt.

Beispiel 80 (fortgesetzt von den Seiten 253 und 255)
Es gilt

$$E(X^2) = \frac{100}{3}$$

und

$$E(X)^2 = 25$$

Nun gilt

$$E(X^2) - E(X)^2 = \frac{100}{3} - \frac{75}{3} = \frac{25}{3}\,.$$

Also ist bei der Gleichverteilung

$$E(X^2) > E(X)^2\,.$$

□

Oft betrachtet man die Zufallsvariable $Y = X - E(X)$. Man sagt auch, dass man die Zufallsvariable X zentriert. Es gilt

$$E(X - E(X)) = 0\,. \tag{7.6}$$

Setzen wir in Gleichung (7.5) $a = 1$ und $b = -E(X)$, so gilt

$$\boxed{E(X - E(X)) = E(X) - E(X) = 0}.$$

Der Erwartungswert einer zentrierten Zufallsvariablen ist also gleich 0. Bei einer zentrierten Zufallsvariablen sieht man also sofort, welche Werte kleiner und welche größer als der Erwartungswert sind.
Die folgende Eigenschaft des Erwartungswertes benötigen wir im nächsten Kapitel.

Satz 7.2
X sei eine Zufallsvariable mit Erwartungswert $E(X)$. Außerdem seien g und h reellwertige Funktionen. Dann gilt

$$\boxed{E(g(X) + h(X)) = E(g(X)) + E(h(X))} \tag{7.7}$$

Beweis:
Wir zeigen den diskreten Fall.

$$\begin{aligned} E(g(X) + E(h(X)) &= \sum_{\{x|x\in\mathcal{T}_\mathcal{X}\}} (g(x) + h(x))\, f_X(x) \\ &= \sum_{\{x|x\in\mathcal{T}_\mathcal{X}\}} (g(x) f_X(x) + h(x) f_X(x)) \\ &= \sum_{\{x|x\in\mathcal{T}_\mathcal{X}\}} g(x) f_X(x) + \sum_{\{x|x\in\mathcal{T}_\mathcal{X}\}} h(x) f_X(x) \\ &= E(g(X)) + E(h(X)) \end{aligned}$$

□

Die Aussage des Satzes gilt auch, wenn wir mehr als zwei Funktionen von X betrachten. Sind also $g_1, \ldots, g_k$ reellwertige Funktionen, so gilt

$$\boxed{E\left[\sum_{i=1}^{k} g_i(X)\right] = \sum_{i=1}^{k} E(g_i(X))} \tag{7.8}$$

7.2 Die Varianz

Der Erwartungswert einer Zufallsvariablen X ist ein Maß für die Lage von X. Neben der Lage der Verteilung einer Zufallsvariablen X ist auch die Streuung von Interesse.
Für eine Urliste $x_1, \ldots, x_n$ ist die mittlere quadratische Abweichung d^2 folgendermaßen definiert:

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

Wir bestimmen hier den Mittelwert der quadratierten Abweichungen der Beobachtungen vom Mittelwert. Ersetzen wir in diesem Satz Mittelwert durch Erwartungswert und Beobachtungen durch Zufallsvariable, so erhalten wir folgende Definition:

Definition 7.3
Sei X eine Zufallsvariable. Die Varianz $Var(X)$ von X ist definiert durch

$$\boxed{Var(X) = E([X - E(X)]^2)} \tag{7.9}$$

Wir berechnen die Varianz im diskreten Fall also durch

$$\boxed{Var(X) = \sum_{\{x|x \in \mathcal{T}_X\}} [x - E(X)]^2 \, f_X(x)} \tag{7.10}$$

und im stetigen Fall durch

$$\boxed{Var(X) = \int_{-\infty}^{\infty} [x - E(X)]^2 \, f_X(x) \, dx} \tag{7.11}$$

Wir wählen für die Varianz oft die Abkürzung σ^2.
Die Varianz besitzt nicht die gleiche Maßeinheit wie X, die Standardabweichung $\sqrt{Var(X)}$ hingegen doch. Wir kürzen im folgenden die Standardabweichung mit σ ab.

Beispiel 84
Im Beispiel 82 auf Seite 252 haben wir zwei Alternativen betrachtet. Bei Alternative A erhalten wir $ 240. Wir bezeichnen die zugehörige Zufallsvariable mit X. Es gilt

$$P(X = 240) \;=\; 1$$

Bei Alternative B erhalten wir mit Wahrscheinlichkeit 0.25 $ 1000 und mit Wahrscheinlichkeit 0.75 nichts. Wir bezeichnen die zugehörige Zufallsvariable mit Y. Es gilt

$$P(Y = 0) = 0.75 \qquad P(Y = 1000) = 0.25$$

Es gilt $E(X) = 240$ und $E(Y) = 250$. Obwohl der Erwartungswert von Alternative B höher ist, entscheiden sich 84 Prozent der Befragten für Alternative A. Dies liegt an der unterschiedlichen Streuung. Offensichtlich gilt

$$Var(X) = 0\,.$$

Für Y gilt

$$Var(Y) \;=\; (0 - 250)^2 \cdot 0.75 + (1000 - 250)^2 \cdot 0.25 = 187500$$

Die zweite Alternative hat die größere Varianz. Tversky und Kahneman stellen fest, dass die meisten Personen in Situationen mit möglichem Gewinn nicht risikofreudig sind. Bei möglichen Verlusten ist dies anders. Tversky und Kahneman fragten Personen, welche der beiden folgenden Alternativen sie vorzögen.

Alternative A

Man hat einen sicheren Verlust von $ 750.

Alternative B

Mit Wahrscheinlichkeit 0.75 verliert man $ 1000 und mit Wahrscheinlichkeit 0.25 nichts.
Sei X der Verlust bei Alternative A und Y der Verlust bei Alternative B. Es gilt

$$P(X = 750) \;=\; 1$$

und

$$P(Y = 1000) = 0.75 \qquad P(Y = 0) = 0.25\,.$$

Somit gilt $E(X) = 750$ und $E(Y) = 750$. In beiden Fällen ist der erwartete Verlust gleich hoch. Die Varianz ist bei Alternative A gleich 0 und bei Alternative B:

$$Var(Y) = (0 - 750)^2 \cdot 0.25 + (1000 - 750)^2 \cdot 0.75 = 187500\,.$$

In diesem Fall entscheiden sich 87 Prozent der Befragten für Alternative B.

Der folgende Satz zeigt, wie man die Varianz einfach berechnen kann.

Satz 7.3
Sei X eine Zufallsvariable mit Varianz $Var(X)$. Dann gilt

$$\boxed{Var(X) = E(X^2) - E(X)^2} \tag{7.12}$$

Beweis: Mit Gleichung (7.8) auf Seite 257 gilt:

$$\begin{aligned} Var(X) &= E([X - E(X)]^2) = E(X^2 - 2XE(X) + E(X)^2) \\ &= E(X^2) - E(2XE(X)) + E(E(X)^2) \\ &= E(X^2) - 2E(X)E(X) + E(X)^2 \\ &= E(X^2) - E(X)^2 \end{aligned}$$

□

Beispiel 84 (fortgesetzt)
Wir berechnen die Varianz der Zufallsvariablen Y mit Hilfe von (7.12). Es gilt $P(Y = 1000) = 0.75$, $P(Y = 0) = 0.25$ und $E(Y) = 750$. Weiterhin gilt

$$E(Y^2) = \sum_y y^2\, P(Y = y) = 0^2 \cdot 0.25 + 1000^2 \cdot 0.75 = 750000\,.$$

Also gilt

$$Var(Y) = E(Y^2) - E(Y)^2 = 750000 - 750^2 = 187500$$

□

Beispiel 80 (fortgesetzt von den Seiten 253 und 255)
Für die Gleichverteilung gilt $E(X) = 5$ und $E(X^2) = 100/3$. Also gilt

$$Var(X) = E(X^2) - E(X)^2 = \frac{100}{3} - 5 = \frac{25}{3}\,.$$

□

Beispiel 81 (fortgesetzt von den Seiten 253 und 255)
Für die Exponentialverteilung mit $\lambda = 1$ gilt $E(X) = 1$ und $E(X^2) = 2$. Also gilt

$$Var(X) = E(X^2) - E(X)^2 = 2 - 1 = 1\,.$$

Schauen wir uns an wie sich die Varianz einer Zufallsvariablen X ändert, wenn X linear transformiert wird. Ist X eine Zufallsvariable und a und b reelle Zahlen, dann gilt:

$$\boxed{Var(a\,X + b) = a^2\,Var(X)}\,. \tag{7.13}$$

Dies sieht man folgendermaßen:

$$\begin{aligned} Var(a\,X + b) &= E\left([a\,X + b - E(a\,X + b)]^2\right) \\ &= E\left([a\,X + b - a\,E(X) - b]^2\right) = E\left([a\,(X - E(X))]^2\right) \\ &= E\left(a^2\,[X - E(X)]^2\right) = a^2\,E\left([X - E(X)]^2\right) \\ &= a^2\,Var(X) \end{aligned}$$

Auf Seite 256 haben wir zentrierte Zufallsvariablen betrachtet. Der Erwartungswert einer zentrierten Zufallsvariablen ist gleich 0. Sei X eine Zufallsvariable mit $E(X) = \mu$ und $Var(X) = \sigma^2$. Wir betrachten die Zufallsvariable

$$\boxed{Z = \frac{X - \mu}{\sigma}}\,. \tag{7.14}$$

Man nennt Z standardisierte Zufallsvariable. Es gilt

$$\boxed{E(Z) = 0} \tag{7.15}$$

und

$$\boxed{Var(Z) = 1}\,. \tag{7.16}$$

Schauen wir uns erst Beziehung (7.15) an:

$$E(Z) = E\left[\frac{X - \mu}{\sigma}\right] = \frac{1}{\sigma}\,E(X - \mu) = 0\,.$$

Beziehung (7.16) gilt wegen (7.13):

$$Var(Z) = Var\left[\frac{X - \mu}{\sigma}\right] = \frac{1}{\sigma^2}\,Var(X - \mu) = \frac{1}{\sigma^2}\,Var(X) = 1\,.$$

7.3 Die Tschebyscheff-Ungleichung

Bisher haben wir die Frage noch nicht beantwortet, wann eine Varianz klein oder groß ist. Dies wollen wir jetzt nachholen. Wir betrachten zunächst eine nichtnegative Zufallsvariable Y. Es gilt also $P(Y = y) = 0$ für $y < 0$.
Ist a eine positive reelle Zahl, so gilt

$$\boxed{P(Y \geq a) \leq \frac{E(Y)}{a}}\,. \tag{7.17}$$

Dies ist die Markow-Ungleichung. Wir beweisen sie für eine stetige Zufallsvariable Y mit Dichtefunktion $f_Y(y)$ auf Seite 264. Vorerst schauen wir uns aber an, welche Folgerungen man aus der Markow-Ungleichung ziehen kann. Ist X eine Zufallsvariable mit $E(X) = \mu$ und $Var(X) = \sigma^2$. Wir betrachten die Zufallsvariable $Y = (X - \mu)^2$. Da diese nichtnegativ ist, können wir die Markow-Ungleichung anwenden. Es gilt also

$$P((X - \mu)^2 \geq a) \leq \frac{E([X - \mu]^2)}{a}\,.$$

Wegen

$$\sqrt{c^2} = |c|$$

und

$$Var(X) = E((X - \mu)^2)$$

ist dies äquivalent zu:

$$P(|X - \mu| \geq a) \leq \frac{Var(X)}{a}\,.$$

Setzen wir $a = k^2\,\sigma^2$ mit $k > 0$, so gilt:

$$\boxed{P(|X - \mu| \geq k\,\sigma) \leq \frac{1}{k^2}}\,. \tag{7.18}$$

Gleichung (7.18) ist die Tschebyscheff-Ungleichung. Diese erlaubt es, die Wahrscheinlichkeit abzuschätzen, dass eine Zufallsvariable X Werte im Intervall $(\mu - k\,\sigma, \mu + k\,\sigma)$ annimmt. Multiplizieren wir beide Seiten von Gleichung (7.18) mit -1 und addieren 1, so erhalten wir folgende Ungleichung:

$$\boxed{1 - P(|X - \mu| \geq k\,\sigma) \geq 1 - \frac{1}{k^2}} \tag{7.19}$$

Auf der linken Seite von Gleichung 7.19 zeigt die Wahrscheinlichkeit des Komplementärereignisses von $|X - \mu| \geq k\,\sigma$. Dieses ist $|X - \mu| < k\,\sigma$. Also gilt

$$P(\mu - k\,\sigma < X < \mu + k\,\sigma) \geq 1 - \frac{1}{k^2}$$

Für $k = 1, 2, 3$ gilt also

$$P(\mu - k\,\sigma < X < \mu + k\,\sigma) \geq \begin{cases} 0 & \text{für } k = 1 \\ \frac{3}{4} & \text{für } k = 2 \\ \frac{8}{9} & \text{für } k = 3 \end{cases}$$

Bei jeder Zufallsvariablen X liegen also mindestens 75 Prozent der Werte im Intervall $(\mu - 2\,\sigma, \mu + 2\,\sigma)$. Bei speziellen Verteilungen kann dieser Wert aber bedeutend größer sein.

Beispiel 80 (fortgesetzt von Seite 260)
Bei der Gleichverteilung auf $[0, 10]$ gilt $\sigma = \sqrt{25/3} = 2.87$. Hieraus folgt

$$\mu - 2\,\sigma = 5 - 2 \cdot 2.87 = -0.74$$

und

$$\mu + 2\,\sigma = 5 + 2 \cdot 2.87 = 10.74\,.$$

Somit liegen alle Werte im Intervall $[\mu - 2\,\sigma < X < \mu + 2\,\sigma]$. □

Beispiel 81 (fortgesetzt von Seite 261)
Bei der Exponentialverteilung mit $\lambda = 1$ gilt $\mu = 1$ und $\sigma = 1$. Hieraus folgt

$$\begin{aligned} P(\mu - 2\,\sigma < X < \mu + 2\,\sigma) &= P(1 - 2 \cdot 1 < X < 1 + 2 \cdot 1) \\ &= P(-1 < X < 3) = F_X(3) - F_X(-1) \\ &= 1 - e^{-3} - 0 = 0.9502 \end{aligned}$$

und

$$\begin{aligned} P(\mu - 3\,\sigma < X < \mu + 3\,\sigma) &= P(1 - 3 \cdot 1 < X < 1 + 3 \cdot 1) \\ &= P(-2 < X < 4) = F_X(4) - F_X(-2) \\ &= 1 - e^{-4} - 0 = 0.9817 \end{aligned}$$

□

Nun noch der Beweis der Markow-Ungleichung. Es gilt

$$\begin{aligned} E(Y) &= \int_0^\infty y\, f_Y(y)dy = \int_0^a y\, f_Y(y)dy + \int_a^\infty y f_Y(y)dy \\ &\geq \int_a^\infty y\, f_Y(y)dy \geq a \int_a^\infty f_Y(y)dy = a\, P(Y \geq a) \end{aligned}$$

Also gilt

$$E(Y) \geq a\, P(Y \geq a)\,.$$

Hieraus folgt

$$P(Y \geq a) \leq \frac{E(Y)}{a}\,.$$

7.4 Quantile

Im Kapitel 3.2.2 haben wir empirische Quantile betrachtet. Das p-Quantil ist der Wert, der von $100 \cdot p$ Prozent der Beobachtungen nicht überschritten wird. Bei einem stetigen Merkmal gilt $\hat{F}(x_p) = p$. Wir wollen Quantile hier nur für stetige Zufallsvariablen betrachten, bei denen die Verteilungsfunktion auf dem Träger streng monoton wachsend ist. Dies ist bei den Verteilungen in den Beispielen 80 und 81 der Fall. In Analogie zur Empirie ist das (theoretische) Quantil x_p der Wert, für den gilt $F_X(x_p) = p$.

Beispiel 80 (fortgesetzt von Seite 245)
Der Träger der Gleichverteilung auf $[0, 10]$ ist das Intervall $[0, 10]$. Hier gilt

$$F_X(x) = 0.1\, x\,. \tag{7.20}$$

Um das Quantil x_p zu bestimmen, setzen wir x_p für x in die Gleichung (7.20) ein und erhalten:

$$F_X(x_p) = p = 0.1\, x_p\,.$$

Lösen wir diese Gleichung nach x_p auf, so erhalten wir

$$x_p = 10 \cdot p$$

So ist das untere Quartil der Gleichverteilung auf $[0, 10]$ gleich $x_{0.25} = 10 \cdot 0.25 = 2.5$. □

Beispiel 81 (fortgesetzt von Seite 247)
Für die Exponentialverteilung mit Parameter $\lambda = 1$ ist der Träger das Intervall $[0, \infty)$. Hier gilt

$$F_X(x) = 1 - e^{-\lambda x}\,. \tag{7.21}$$

Um das Quantil x_p zu bestimmen, setzen wir x_p für x in die Gleichung (7.21) ein und erhalten:

$$F_X(x_p) = p = 1 - e^{-\lambda x_p}\,.$$

Diese Gleichung können wir nach x_p auflösen:

$$\begin{aligned} p = 1 - e^{-\lambda x_p} &\iff e^{-\lambda x_p} = 1 - p \iff -\lambda x_p = \ln(1-p) \\ &\iff x_p = -\frac{1}{\lambda}\ln(1-p) \end{aligned}$$

So ist der Median der Exponentialverteilung mit $\lambda = 1$ gleich $x_{0.5} = \ln 0.5 = -0.693$.

Wir werden später oftmals die standardisierte Zufallsvariable

$$\boxed{Z = \frac{X - \mu}{\sigma}}$$

betrachten. Dabei ist μ der Erwartungswert und σ^2 die Varianz von X. Ist x_p das p-Quantil von X, so gilt für das p-Quantil z_p von Z

$$\boxed{z_p = \frac{x_p - \mu}{\sigma}}\,. \tag{7.22}$$

Wir haben zu zeigen:

$$P(Z \leq \frac{x_p - \mu}{\sigma}) = p$$

Dies sieht man folgendermaßen:

$$\begin{aligned} P(X \leq x_p) = p &\Leftrightarrow P(X - \mu \leq x_p - \mu) = p \Leftrightarrow P(\frac{X-\mu}{\sigma} \leq \frac{x_p - \mu}{\sigma}) = p \\ &\Leftrightarrow P(Z \leq \frac{x_p - \mu}{\sigma}) = p \end{aligned}$$

Wir können z_p natürlich auch aus x_p bestimmen, indem wir Gleichung 7.22 nach z_p auflösen. Wir erhalten

$$\boxed{x_p = \mu + z_p\sigma}\,. \tag{7.23}$$

Kapitel 8

Multivariate Zufallsvariablen

8.1 Diskrete Zufallsvariablen

Bisher haben wir immer nur eine Zufallsvariable betrachtet. Bei vielen Anwendungen sind aber mehrere Zufallsvariablen von Interesse. So besteht ein Fragebogen in der Regel aus mehreren Fragen, mit denen die Ausprägungen von Merkmalen erfragt werden.

Bei einer univariaten diskreten Zufallsvariablen gehen wir von der Ergebnismenge Ω eines Zufallsvorgangs aus und ordnen jedem Ergebnis eine reelle Zahl zu. Eine p-dimensionale Zufallsvariable $(X_1, \ldots, X_p)$ erhalten wir, indem wir jedem Ergebnis p reelle Zahlen zuordnen. Im Folgenden betrachten wir nur bivariate Zufallsvariablen. Dabei bezeichnen wir die Zufallsvariablen mit X und Y und die bivariate Zufallsvariable mit (X, Y). Wie bei einer univariaten diskreten Zufallsvariablen bestimmen wir die Wahrscheinlichkeitsfunktion. Diese ist gegeben durch:

$$f_{X,Y}(x, y) = P(X = x, Y = y)\,. \tag{8.1}$$

Man nennt $f_{X,Y}(x, y)$ auch die gemeinsame Wahrscheinlichkeitsfunktion von X und Y. Wir stellen diese Wahrscheinlichkeiten in einer Tabelle zusammen. Tabelle 8.1 zeigt den Aufbau dieser Tabelle, wenn die Zufallsvariable X die Merkmalsausprägungen $x_1, \ldots, x_k$ und die Zufallsvariable Y die Merkmalsausprägungen $y_1, \ldots, x_l$ besitzt.

Tabelle 8.1: Allgemeiner Aufbau der Wahrscheinlichkeitsfunktion einer zweidimensionalen diskreten Zufallsvariablen

X \ Y	y_1	$\dots$	y_l
x_1	$P(X = x_1, Y = y_1)$	$\dots$	$P(X = x_1, Y = y_l)$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
x_k	$P(X = x_k, Y = y_1)$	$\dots$	$P(X = x_k, Y = y_l)$

Beispiel 85
Eine faire Münze werde dreimal hintereinander geworfen. Wir suchen die Wahrscheinlichkeitsfunktion der bivariaten Zufallsvariablen (X, Y), wobei X die Anzahl ZAHL bei den ersten beiden Würfen und Y die Anzahl ZAHL bei den beiden letzten Würfen ist.
Die Ergebnismenge ist

$$\Omega = \{KKK, KKZ, KZK, ZKK, KZZ, ZKZ, ZZK, ZZZ\}\,.$$

In Tabelle 8.2 sind die Ergebnisse mit den zugehörigen Werten von X und Y zu finden.

Tabelle 8.2: Ergebnisse beim dreimaligen Münzwurf mit zugehörigen Werten der Zufallsvariablen X und Y

ω	x	y
KKK	0	0
KKZ	0	1
KZK	1	1
ZKK	1	0
KZZ	1	2
ZKZ	1	1
ZZK	2	1
ZZZ	2	2

Da die Münze fair ist, sind alle Elementarereignisse gleich wahrscheinlich.

Für die Wahrscheinlichkeitsfunktion von (X, Y) gilt also

$$\begin{aligned} P(X=0, Y=0) &= 0.125 \\ P(X=1, Y=0) &= 0.125 \\ P(X=0, Y=1) &= 0.125 \\ P(X=1, Y=1) &= 0.25 \\ P(X=1, Y=2) &= 0.125 \\ P(X=2, Y=1) &= 0.125 \\ P(X=2, Y=2) &= 0.125 \end{aligned}$$

Wir stellen die Wahrscheinlichkeitsfunktion in einer zweidimensionalen Tabelle dar.

Tabelle 8.3: Wahrscheinlichkeitsfunktion einer zweidimensionalen diskreten Zufallsvariablen

X \ Y	0	1	2
0	0.125	0.125	0
1	0.125	0.250	0.125
2	0	0.125	0.125

□

Aus der Wahrscheinlichkeitsfunktion $P(X = x, Y = y)$ erhält man problemlos die Wahrscheinlichkeitsfunktionen der univariaten Zufallsvariablen X und Y. Es gilt

$$P(X = x) = \sum_y P(X = x, Y = y) \tag{8.2}$$

und

$$P(Y = y) = \sum_x P(X = x, Y = y)\,. \tag{8.3}$$

Man spricht auch von den Randverteilungen, da man diese Wahrscheinlichkeiten durch Summation der Wahrscheinlichkeiten in den Zeilen bzw. Spalten der Tabelle mit der Wahrscheinlichkeitsfunktion erhält. Tabelle 8.4 zeigt dies.

Tabelle 8.4: Allgemeiner Aufbau der Wahrscheinlichkeitsfunktion einer zweidimensionalen diskreten Zufallsvariablen mit Randverteilungen

Y / X	y_1	$\dots$	y_l	
x_1	$P(X=x_1,Y=y_1)$	$\dots$	$P(X=x_1,Y=y_l)$	$P(X=x_1)$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_k	$P(X=x_k,Y=y_1)$	$\dots$	$P(X=x_k,Y=y_l)$	$P(X=x_k)$
	$P(Y=y_1)$	$\dots$	$P(Y=y_k)$	1

Beispiel 85 (fortgesetzt)
Es gilt

$$\begin{aligned} P(X=0) &= P(X=0,Y=0)+P(X=0,Y=1) \\ &= 0.125+0.125=0.25 \\ P(X=1) &= P(X=1,Y=0)+P(X=1,Y=1)+P(X=1,Y=2) \\ &= 0.125+0.25+0.125=0.5 \\ P(X=2) &= P(X=2,Y=1)+P(X=2,Y=2) \\ &= 0.125+0.125=0.25 \end{aligned}$$

und

$$\begin{aligned} P(Y=0) &= P(X=0,Y=0)+P(X=1,Y=0) \\ &= 0.125+0.125=0.25 \\ P(Y=1) &= P(X=0,Y=1)+P(X=1,Y=1)+P(X=2,Y=1) \\ &= 0.125+0.25+0.125=0.5 \\ P(Y=2) &= P(X=1,Y=2)+P(X=2,Y=2) \\ &= 0.125+0.125=0.25 \end{aligned}$$

In Tabelle 8.5 ist die gemeinsame Wahrscheinlichkeitsverteilung mit den Randverteilungen zu finden.

Tabelle 8.5: Wahrscheinlichkeitsfunktion einer zweidimensionalen diskreten Zufallsvariablen mit Randverteilung

X \ Y	0	1	2	
0	0.125	0.125	0	0.25
1	0.125	0.250	0.125	0.50
2	0	0.125	0.125	0.25
	0.250	0.500	0.250	1.00

□

Die Verteilungen der Zufallsvariablen X und Y im Beispiel 85 sind identisch. Wir sprechen auch von identisch verteilten Zufallsvariablen.
Aus der Wahrscheinlichkeitsfunktion $P(X = x, Y = y)$ kann problemlos auf $P(X = x)$ und $P(Y = y)$ geschlossen werden. Von den Wahrscheinlichkeitfunktionen $P(X = x)$ und $P(Y = y)$ kann nicht eindeutig auf die Wahrscheinlichkeitsfunktion $P(X = x, Y = y)$ geschlossen werden.

Beispiel 86
In Tabelle 8.6 findet man die Wahrscheinlichkeitsfunktion $P(V = v, W = w)$, die sich von der Wahrscheinlichkeitsfunktion $P(X = x, Y = y)$ im Beispiel 85 unterscheidet. Die Randverteilungen sind aber identisch.

Tabelle 8.6: Wahrscheinlichkeitsfunktion einer zweidimensionalen diskreten Zufallsvariablen mit Randverteilung

V \ W	0	1	2	
0	0.0625	0.125	0.0625	0.25
1	0.1250	0.250	0.1250	0.50
2	0.0625	0.125	0.0625	0.25
	0.2500	0.500	0.2500	1.00

□

8.2 Stetige Zufallsvariablen

Eine stetige univariate Zufallsvariable haben wir über die Verteilungsfunktion definiert. Schauen wir uns also die Verteilungsfunktion einer zweidimensionalen Zufallsvariablen an. Mit der Verteilungsfunktion $F_{X,Y}(x,y)$ einer zweidimensionalen Zufallsvariablen (X,Y) können wir folgende Wahrscheinlichkeiten bestimmen:

$$F_{X,Y}(x,y) \quad = \quad P(X \leq x, Y \leq y)\,. \tag{8.4}$$

Wir definieren eine stetige zweidimensionale Zufallsvariable wie eine univariate stetige Zufallsvariable über die Verteilungsfunktion.

Definition 8.1
Eine bivariate Zufallsvariable (X,Y) heißt stetig, wenn eine reellwertige Funktion $f_{X,Y} : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ existiert, so dass für die Verteilungsfunktion $F_{X,Y}(x,y)$ von (X,Y) gilt:

$$F_{X,Y}(x,y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(u,v)\,du\,dv$$

Wir nennen $f_{X,Y}(x,y)$ die gemeinsame Dichtefunktion von (X,Y).
Die gemeinsame Dichtefunktion $f_{X,Y}(x,y)$ von (X,Y) erfüllt folgende Bedingungen:

1. $f_{X,Y}(x,y) \geq 0 \quad$ für alle $\quad (x,y) \in \mathbb{R}^2$
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx \ldots\, dy = 1$

Beispiel 87
Wir betrachten folgende Funktion

$$f(x,y) = \begin{cases} 1 & \text{für } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & sonst \end{cases}$$

Offensichtlich gilt

$$f(x,y) \geq 0 \quad \text{für alle} \quad (x,y) \in \mathbb{R}^2\,.$$

Außerdem gilt

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dxy = \int_{0}^{1}\int_{0}^{1} 1\,dx\,dy = \int_{0}^{1}\Big[x\Big]_0^1 dy$$

$$= \int_{0}^{1} 1\,dy = \Big[y\Big]_0^1 = 1$$

Also handelt es sich um eine Dichtefunktion einer zweidimensionalen Zufallsvariablen (X,Y). □

Aus der gemeinsamen Dichtefunktion können wir problemlos die Dichtefunktionen der univariaten Zufallsvariablen gewinnen. Im Fall einer zweidimensionalen stetigen Zufallsvariablen gilt

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy$$

und

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,.$$

Beispiel 87 (fortgesetzt)
Es gilt

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy = \int_{-\infty}^{\infty} 1\,dy = \Big[y\Big]_0^1 = 1$$

und

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx = \int_{-\infty}^{\infty} 1\,dx = \Big[x\Big]_0^1 = 1\,.$$

□

8.3 Unabhängigkeit

Die Ereignisse A und B heißen unabhängig, wenn gilt

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

Das Konzept der Unabhängigkeit lässt sich auch auf Zufallsvariablen übertragen.

Definition 8.2
Gilt für alle $(x, y) \in \mathbb{R}^2$

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y)\,,$$

so heißen die diskreten Zufallsvariablen X und Y unabhängig.

Beispiel 87 (fortgesetzt)
Die Zufallsvariablen Vund W in Tabelle 8.6 sind unabhängig, da gilt

$$\begin{aligned}
P(V=0,W=0) &= 0.0625 = 0.25 \cdot 0.25 = P(V=0) \cdot P(W=0)\\
P(V=0,W=1) &= 0.125 = 0.25 \cdot 0.5 = P(V=0) \cdot P(W=1)\\
P(V=0,W=2) &= 0.0625 = 0.25 \cdot 0.25 = P(V=0) \cdot P(W=2)\\
P(V=1,W=0) &= 0.125 = 0.5 \cdot 0.25 = P(V=1) \cdot P(W=0)\\
P(V=1,W=1) &= 0.25 = 0.5 \cdot 0.5 = P(V=1) \cdot P(W=1)\\
P(V=1,W=2) &= 0.125 = 0.5 \cdot 0.25 = P(V=1) \cdot P(W=2)\\
P(V=2,W=0) &= 0.0625 = 0.25 \cdot 0.25 = P(V=2) \cdot P(W=0)\\
P(V=2,W=1) &= 0.125 = 0.25 \cdot 0.5 = P(V=2) \cdot P(W=1)\\
P(V=2,W=2) &= 0.0625 = 0.25 \cdot 0.25 = P(V=2) \cdot P(W=2)
\end{aligned}$$

□

Sind die Zufallsvariablen X und Y unabhängig, so kann man aus den Wahrscheinlichkeitsfunktionen $P(X = x)$ und $P(Y = y)$ die gemeinsame Wahrscheinlichkeitsfunktion $P(X = x, Y = y)$ bestimmen. Zieht man mehrmals mit Zurücklegen aus einer Urne jeweils eine Kugel oder aus mehreren Urnen jeweils eine Kugel, so sind die einzelnen Ziehungen unabhängig.

Beispiel 88
Ein Urne enthält 10 Kugeln. Von diesen wiegen 4 Kugeln 10 g und die anderen Kugeln 20 g. Es werden zwei Kugeln mit Zurücklegen gezogen. Sei

X das Gewicht der ersten gezogenen Kugel und Y das Gewicht der zweiten gezogenen Kugel. Es gilt

$$P(X = 10) = 0.4 \qquad P(X = 20) = 0.6$$
$$P(Y = 10) = 0.4 \qquad P(Y = 20) = 0.6$$

Da wir ohne Zurücklegen ziehen,sind X und Y unabhängig. Also gilt

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y)\,.$$

Wir erhalten

$$\begin{aligned}
P(X = 10, Y = 10) &= P(X = 10) \cdot P(Y = 10) = 0.4 \cdot 0.4 = 0.16 \\
P(X = 10, Y = 20) &= P(X = 10) \cdot P(Y = 20) = 0.4 \cdot 0.6 = 0.24 \\
P(X = 20, Y = 10) &= P(X = 20) \cdot P(Y = 10) = 0.6 \cdot 0.4 = 0.24 \\
P(X = 20, Y = 20) &= P(X = 20) \cdot P(Y = 20) = 0.6 \cdot 0.6 = 0.36
\end{aligned}$$

Die gemeinsame Wahrscheinlichkeitsfunktion ist in Tabelle 8.7 zu finden.

Tabelle 8.7: Gemeinsame Wahrscheinlichkeitsfunktion

X \ Y	10	20	
10	0.16	0.24	0.40
20	0.24	0.36	0.60
	0.40	0.60	1.00

□

Beispiel 89
Eine Urne enthält 10 Kugeln. Von diesen wiegen 2 Kugeln 10 g und die anderen Kugeln 20 g. Ein zweite Urne enthält ebenfalls 10 Kugeln. Von diesen wiegen 4 Kugeln 10 g und die anderen Kugeln 20 g. Aus jeder Urne wird eine Kugel gzogen. Sei X das Gewicht der ersten gezogenen Kugel und Y das Gewicht der zweiten gezogenen Kugel. Es gilt

$$P(X = 10) = 0.2 \qquad P(X = 20) = 0.8$$
$$P(Y = 10) = 0.4 \qquad P(Y = 20) = 0.6$$

Da wir aus jeder Urne eine Kugel ziehen, sind X und Y unabhängig. Also gilt

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y)\,.$$

Wir erhalten

$$\begin{aligned}
P(X = 10, Y = 10) &= P(X = 10) \cdot P(Y = 10) = 0.2 \cdot 0.4 = 0.08 \\
P(X = 10, Y = 20) &= P(X = 10) \cdot P(Y = 20) = 0.2 \cdot 0.6 = 0.12 \\
P(X = 20, Y = 10) &= P(X = 20) \cdot P(Y = 10) = 0.8 \cdot 0.4 = 0.32 \\
P(X = 20, Y = 20) &= P(X = 20) \cdot P(Y = 20) = 0.8 \cdot 0.6 = 0.48
\end{aligned}$$

Die gemeinsame Wahrscheinlichkeitsfunktion ist Tabelle 8.8 zu finden.

Tabelle 8.8: Gemeinsame Wahrscheinlichkeitsfunktion

X \ Y	10	20	
10	0.08	0.12	0.20
20	0.32	0.48	0.80
	0.40	0.60	0.20

□

Schauen wir uns noch stetige Zufallsvariablen an. Die stetigen Zufallsvariablen X, Y mit gemeinsamer Dichtefunktion $f_{X,Y}(x, y)$ und Randdichtefunktionen $f_X(x)$ und $f_Y(y)$ sind genau dann unabhängig, wenn gilt

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)\,. \tag{8.5}$$

Beispiel 89 (fortgesetzt)
Wir haben gesehen, dass die Randdichtefunktionen $f_X(x)$ und $f_Y(y)$ von

$$f_{X,Y}(x, y) = \begin{cases} 1 & \text{für } 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \textit{sonst} \end{cases}$$

univariate Gleichverteilungen sind. Es gilt also

$$f_X(x) = \begin{cases} 1 & \text{für} \quad 0 \le x \le 1 \\ 0 & \text{sonst} \end{cases}$$

und

$$f_Y(y) = \begin{cases} 1 & \text{für} \quad 0 \leq y \leq 1 \\ 0 & \text{sonst} \end{cases}$$

Offensichtlich gilt

$$f_X(x) \cdot f_Y(y) = \begin{cases} 1 & \text{für } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & sonst \end{cases}$$

Somit sind die Zufallsvariablen Xund Y unabhängig. □

8.4 Funktionen von Zufallsvariablen

In der schließenden Statistik sind Funktionen von Zufallsvariablen von Interesse. Schauen wir uns dies für eine Funktion von zwei diskreten Zufallsvariablen an.
Sind X und Y Zufallsvariablen mit gemeinsamer Wahrscheinlichkeitsfunktion $P(X = x, Y = y)$ und $g : D \subset \mathbb{R}^2 \to \mathbb{R}$ eine Funktion. Dann gilt für die Zufallsvariable $V = g(X, Y)$:

$$P(V = v) = \sum_{\{(x,y)|g(x,y)=v\}} P(X = x, Y = y)$$

Beispiel 90
Wir betrachten die Zufallsvariablen X und Y aus Beispiel 89 auf Seite 275, deren gemeinsame Wahrscheinlichkeitsfunktion in Tabelle 8.8 auf Seite 276 zu finden ist. Wir bestimmen die Wahrscheinlichkeitsfunktion der Zufallsvariablen $V = X + Y$ an. Es gilt

$$\begin{aligned} P(V = 20) &= P(X = 10, Y = 10) = 0.08 \\ P(V = 30) &= P(X = 10, Y = 20) + P(X = 20, Y = 10) = 0.44 \\ P(V = 40) &= P(X = 20, Y = 20) = 0.48 \end{aligned}$$

□

Beispiel 91
Wir betrachten die Zufallsvariablen X und Y aus Beispiel 89 auf Seite 275, deren gemeinsame Wahrscheinlichkeitsfunktion in Tabelle 8.8 auf Seite 276 zu finden ist. Wir bestimmen die Wahrscheinlichkeitsfunktion der Zufallsvariablen $W = X \cdot Y$ an. Es gilt

$$
\begin{aligned}
P(W = 100) &= P(X = 20, Y = 20) = 0.08 \\
P(W = 200) &= P(X = 10, Y = 20) + P(X = 20, Y = 10) = 0.44 \\
P(W = 400) &= P(X = 20, Y = 20) = 0.48
\end{aligned}
$$

□

Kapitel 9

Parameter multivariater Verteilungen

9.1 Erwartungswerte

Wir können auch bei mehrdimensionalen Zufallsvariablen den Erwartungswert betrachten. Dieser ist nichts anderes als der Vektor der Erwartungswerte der univariaten Zufallsvariablen. Es gilt also $E((X,Y)) = (E(X), E(Y))$. Wir müssen also nur den Erwartungswert jeder Komponente bestimmen.

Beispiel 92
Wir betrachten wieder die diskrete Zufallsvariable (X,Y) aus Beispiel 89 auf Seite 275, deren gemeinsame Wahrscheinlichkeitsfunktion in Tabelle 8.8 auf Seite 276 zu finden ist. Es gilt

$$E(X) = 10 \cdot 0.2 + 20 \cdot 0.8 = 18$$

und

$$E(Y) = 10 \cdot 0.4 + 20 \cdot 0.6 = 16\,.$$

Also gilt $E(X,Y) = (18, 16)$. □

In der schließenden Statistik geht man von Zufallsvariablen $X_1, \ldots, X_n$ aus und bestimmt eine reellwertige Funktion $g(X_1, X_2, \ldots, X_k)$. Gesucht ist in der Regel $E(g(X_1, X_2, \ldots, X_k))$. Betrachten wir wieder den Fall $n = 2$ und bezeichnen die Zufallsvariablen mit X und Y.

Ist (X, Y) eine diskrete Zufallsvariable mit gemeinsamer Wahrscheinlichkeitsfunktion $P(X = x, Y = y)$ und $g(x, y)$ eine reellwertige Funktion, so gilt

$$\boxed{E(g(X,Y)) = \sum_x \sum_y g(x,y)\, P(X = x, Y = y)}\,. \tag{9.1}$$

Beispiel 93
Wir betrachten wieder die diskrete Zufallsvariable (X, Y) aus Beispiel 89 auf Seite 275, deren gemeinsame Wahrscheinlichkeitsfunktion in Tabelle 8.8 auf Seite 276 zu finden ist.
Wir betrachten $g(X, Y) = X + Y$ und wenden die Formel (9.1) an:

$$\begin{aligned} E(X+Y) &= (10+10)\cdot 0.08 + (10+20)\cdot 0.12 \\ &+ (20+10)\cdot 0.32 + (20+20)\cdot 0.48 \\ &= 34 \end{aligned}$$

□

Im Beispiel 92 haben wir gezeigt , dass gilt $E(X) = 18$ und $E(Y) = 16$. Wir sehen, dass im Beispiel gilt

$$E(X + Y) = E(X) + E(Y)\,.$$

Wie der folgende Satz zeigt, gilt dies allgemein.

Satz 9.1
Sei (X, Y) eine zweidimensioanle Zufallsvariable mit Wahrscheinlichkeitsfunktion bzw. Dichtefunktion $f_{X,Y}(x, y)$. Es gilt

$$\boxed{E(X + Y) = E(X) + E(Y)}$$

Beweis:

Wir beweisen den diskreten Fall:

$$\begin{aligned}
E(X+Y) &= \sum_x \sum_y (x+y)\, P(X=x, Y=y) = \\
&= \sum_x \sum_y (x\, P(X=x, Y=y) + y\, P(X=x, Y=y)) \\
&= \sum_x \sum_y x\, P(X=x, Y=y) + \sum_x \sum_y y\, P(X=x, Y=y) \\
&= \sum_x x \sum_y P(X=x, Y=y) + \sum_y y \sum_x P(X=x, Y=y) \\
&= \sum_x x\, P(X=x) + \sum_y y\, P(Y=y) = \\
&= E(X) + E(Y)
\end{aligned}$$

Die Aussage des Satzes gilt auch für mehr als zwei Zufallsvariablen. Sind also $X_1, \ldots, X_n$ Zufallsvariablen, so gilt

$$\boxed{E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i)}\,. \tag{9.2}$$

Beispiel 94
Wir betrachten wieder die diskrete Zufallsvariable (X, Y) aus Beispiel 89 auf Seite 275, deren gemeinsame Wahrscheinlichkeitsfunktion in Tabelle 8.8 auf Seite 276 zu finden ist.
Wir betrachten $g(X, Y) = X \cdot Y$ und wenden die Formel (9.1) an:

$$\begin{aligned}
E(X \cdot Y) &= (10 \cdot 10) \cdot 0.08 + (10 \cdot 20) \cdot 0.12 \\
&+ (20 \cdot 10) \cdot 0.32 + (20 \cdot 20) \cdot 0.48 \\
&= 288
\end{aligned}$$

□

Im Beispiel 89 auf Seite 275 haben wir gezeigt, dass X und Y unabhängig sind. Wir sehen, dass im Beispiel 94 gilt $E(X \cdot Y) = E(X) \cdot E(Y)$. Der folgende Satz zeigt, dass dies für unabhängige Zufallsvariablen immer gilt

Satz 9.2
Seien X und Y **unabhängige** Zufallsvariablen. Dann gilt

$$\boxed{E(X\,Y) = E(X)\,E(Y)}$$

Beweis:
Wir beweisen den diskreten Fall:

$$\begin{aligned} E(X\,Y) &= \sum_x \sum_y x\,y\,P(X=x, Y=y) \\ &= \sum_x \sum_y x\,y\,P(X=x)\,P(Y=y) \\ &= \sum_x x\,P(X=x) \sum_y y\,P(Y=y) \\ &= E(X)\,E(Y) \end{aligned}$$

9.2 Kovarianz und Korrelationskoeffizient

Wie wir im letzten Abschnitt gesehen haben, gilt immer

$$E(X+Y) = E(X) + E(Y)\,.$$

Gilt auch immer

$$Var(X+Y) = Var(X) + Var(Y)?$$

Schauen wir uns ein Beispiel an, das von Bernd Streitberg stammt.

Beispiel 95
Ein Teilchen bewegt sich auf den ganzen Zahlen. Bei jedem Schritt geht es entweder nach links oder nach rechts.
Seien

$$X = \begin{cases} 1 & \text{wenn es beim ersten Schritt nach links geht} \\ -1 & \text{wenn es beim ersten Schritt nach rechts geht} \end{cases}$$

und

$$Y = \begin{cases} 1 & \text{wenn es beim zweiten Schritt nach links geht} \\ -1 & \text{wenn es beim zweiten Schritt nach rechts geht} \end{cases}.$$

Von Interesse ist die Position $Z = X + Y$ des Teilchens nach zwei Schritten. Beim ersten Schritt geht das Teilchen zufällig nach links oder rechts. Es gilt also $P(X = -1) = 0.5$ und $P(X = 1) = 0.5$. Also gilt $E(X) = 0$ und $Var(X) = 1$.
Beim zweiten Schritt unterscheiden wir drei Fälle:

1.Fall
Beim zweiten Schritt geht das Teilchen zufällig nach links oder rechts. Tabelle 9.1 zeigt die gemeinsame Wahrscheinlichkeitsfunktion von X und Y.

Tabelle 9.1: Gemeinsame Wahrscheinlichkeitsfunktion von X und Y, wenn das Teilchen beide Schritte zufällig wählt

x \ y	-1	1	
-1	0.25	0.25	0.5
1	0.25	0.25	0.5
	0.50	0.50	1.0

Somit erhalten wir folgende Wahrscheinlichkeitsfunktion von Z:

$$P(Z = z) = \begin{cases} 0.25 & \text{für } z = -2 \\ 0.5 & \text{für } z = 0 \\ 0.25 & \text{für } z = 2 \end{cases}$$

Somit gilt

$$E(Z) = (-2) \cdot 0.25 + 0 \cdot 0.5 + 2 \cdot 0.25 = 0$$

und

$$E(Z^2) = (-2)^2 \cdot 0.25 + 0 \cdot 0.5 + 2^2 \cdot 0.25 = 2$$

Die Varianz von Z ist also

$$Var(Z) = E(Z^2) - E(Z)^2 = 2$$

2.Fall
Beim zweiten Schritt geht das Teilchen in die entgegengesetzte Richtung des ersten Schrittes. Tabelle 9.2 zeigt die gemeinsame Wahrscheinlichkeitsfunktion von X und Y.

Tabelle 9.2: Gemeinsame Wahrscheinlichkeitsfunktion von X und Y, wenn das Teilchen beim zweiten Schritt in die entgegengesetzte Richtung des ersten Schrittes geht

x \ y	-1	1	
-1	0	0.5	0.5
1	0.5	0	0.5
	0.50	0.50	1.0

Da das Teilchen nach 2 Schritten immer im Nullpunkt landet, erhalten wir folgende Wahrscheinlichkeitsfunktion von Z:

$$P(Z = 0) = 1$$

Somit gilt $E(Z) = 0$ und $Var(Z) = 0$.

3.Fall

Beim zweiten Schritt geht das Teilchen in die gleiche Richtung wie beim ersten Schritt. Tabelle 9.3 zeigt die gemeinsame Wahrscheinlichkeitsfunktion von X und Y.

Tabelle 9.3: Gemeinsame Wahrscheinlichkeitsfunktion von X und Y, wenn das Teilchen beim zweiten Schritt in die gleiche Richtung wie beim ersten geht

x \ y	-1	1	
-1	0.5	0	0.5
1	0	0.5	0.5
	0.50	0.50	1.0

Somit erhalten wir folgende Wahrscheinlichkeitsfunktion von Z:

$$P(Z = z) = \begin{cases} 0.5 & \text{für } z = -2 \\ 0.5 & \text{für } z = 2 \end{cases}$$

Somit gilt

$$E(Z) \;=\; (-2)\cdot 0.5 + 2\cdot 0.5 = 0$$

und

$$E(Z^2) \;=\; (-2)^2\cdot 0.5 + 2^2\cdot 0.5 = 4$$

Die Varianz von Z ist also

$$Var(Z) \;=\; E(Z^2) - E(Z)^2 = 4$$

Die Randverteilungen von X und Y sind in allen drei Fällen identisch. Also gilt auch $E(Y) = 0$ und $Var(Y) = 1$.
Beim ersten Fall gilt

$$Var(X+Y) = Var(X) + Var(Y)\,.$$

Beim zweiten Fall gilt

$$Var(X+Y) < Var(X) + Var(Y)\,.$$

Beim dritten Fall gilt

$$Var(X+Y) > Var(X) + Var(Y)\,.$$

□

Das Beispiel zeigt, dass die Größe $Var(X+Y)$ neben $Var(X)$ und $Var(Y)$ offensichtlich noch von einer dritten Größe abhängt. Es gilt

$$\boxed{Var(X+Y) = Var(X) + Var(Y) + 2E[(X-E(X))(Y-E(Y))]}\,.$$

Dies sieht man folgendermaßen:

$$\begin{aligned}
&Var(X+Y) = \\
&\quad E\left[(X+Y-E(X+Y))^2\right] = \\
&\quad E\left[(X+Y-E(X)-E(Y))^2\right] = \\
&\quad E\left[(X-E(X)+Y-E(Y))^2\right] = \\
&\quad E\left[(X-E(X))^2 + (Y-E(Y))^2 + 2\,(X-E(X))(Y-E(Y))\right] = \\
&\quad E\left[(X-E(X))^2\right] + E\left[(Y-E(Y))^2\right] + 2\,E\left[(X-E(X))(Y-E(Y))\right] = \\
&\quad Var(X) + Var(Y) + 2\,E\left[(X-E(X))(Y-E(Y))\right]
\end{aligned}$$

Definition 9.1
Seien X und Y Zufallsvariablen. Dann heißt

$$\boxed{Cov(X,Y) = E\left[(X - E(X))(Y - E(Y))\right]} \tag{9.3}$$

die Kovarianz zwischen X und Y.

Es gilt also

$$\boxed{Var(X+Y) = Var(X) + Var(Y) + 2\,Cov(X,Y)} \tag{9.4}$$

Man kann zeigen, dass für die Zufallsvariablen $X_1, \ldots, X_n$ gilt:

$$\boxed{Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) + 2\sum_{i<j} Cov(X_i, X_j)} \tag{9.5}$$

Schauen wir uns einige Eigenschaften der Kovarianz an. Offensichtlich gilt:

$$\boxed{Cov(X,Y) = Cov(Y,X)}\,. \tag{9.6}$$

Außerdem gilt:

$$\boxed{Cov(X,X) = Var(X)}\,. \tag{9.7}$$

Dies sieht man folgendermaßen:

$$Cov(X,X) \;=\; E\left[(X - E(X))(X - E(X))\right] = E[(X - E(X))^2] = Var(X)\,.$$

Mit der folgenden Formel ist es einfach, die Kovarianz zu berechnen:

$$\boxed{Cov(X,Y) = E(X\,Y) - E(X)\,E(Y)}\,. \tag{9.8}$$

Die Gültigkeit von Gleichung (9.8) sieht man folgendermaßen:

$$\begin{aligned} Cov(X,Y) &= E[(X - E(X))\,(Y - E(Y))] \\ &= E[X\,Y - X\,E(Y) - Y\,E(X) + E(X)\,E(Y)] \\ &= E(X\,Y) - E(X)\,E(Y) - E(Y)\,E(X) + E(X)\,E(Y) \\ &= E[X\,Y] - E(X)\,E(Y) \end{aligned}$$

Das Rechnen mit Kovarianzen ist einfach, wenn man einige Regeln kennt. Es gilt

$$\boxed{Cov(a+X, b+Y) = Cov(X,Y)} \tag{9.9}$$

Dies sieht man folgendermaßen:

$$\begin{aligned} Cov(a+X, b+Y) &= E[(a+X-E(a+X))\,(b+Y-E(b+Y))] \\ &= E[(a+X-a-E(X))\,(b+Y-b-E(Y))] \\ &= E[(X-E(X))\,(Y-E(Y))] \\ &= Cov(X,Y) \end{aligned}$$

Es gilt:

$$\boxed{Cov(a\,X, b\,Y) = a\,b\,Cov(X,Y)} \tag{9.10}$$

Dies sieht man folgendermaßen:

$$\begin{aligned} Cov(a\,X, b\,Y) &= E[(a\,X-E(a\,X))\,(b\,Y-E(b\,Y))] \\ &= E[(a\,X-a\,E(X))\,(b\,Y-b\,E(Y))] \\ &= a\,b\,E[(X-E(X))\,(Y-E(Y))] \\ &= a\,b\,Cov(X,Y) \end{aligned}$$

Es gilt

$$\boxed{Cov(X, Y+Z) = Cov(X,Y) + Cov(X,Z)} \tag{9.11}$$

Dies sieht man folgendermaßen:

$$\begin{aligned} &Cov(X, Y+Z) = \\ &E[(X-E(X))\,(Y+Z-E(Y+Z))] = \\ &E[(X-E(X))\,(Y-E(Y)+Z-E(Z))] = \\ &E[(X-E(X))\,(Y-E(Y)) + (X-E(X))\,(Z-E(Z))] = \\ &E[(X-E(X))\,(Y-E(Y))] + E[(X-E(X))\,(Z-E(Z))] = \\ &Cov(X,Y) + Cov(X,Z) \end{aligned}$$

Der folgende Satz zeigt, dass die Kovarianz bei Unabhängigkeit gleich 0 ist.

Satz 9.3
Sind die Zufallsvariablen X und Y unabhängig, so gilt

$$Cov(X,Y) = 0.$$

Beweis:
Sind X und Y unabhängig, so gilt

$$E(X\,Y) = E(X)\,E(Y).$$

Also gilt

$$Cov(X,Y) = E(X\,Y) - E(X)\,E(Y) = E(X)\,E(Y) - E(X)\,E(Y) = 0$$

Bei Unabhängigkeit kann man die Varianz der Summe von Zufallsvariablen einfach bestimmen.

Satz 9.4
Sind die Zufallsvariablen X und Y **unabhängig**, so gilt

$$\boxed{Var(X+Y) = Var(X) + Var(Y)}$$

Beweis:
Sind X und Y unabhängig, so gilt $Cov(X,Y) = 0$. Also gilt

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

Das folgende Beispiel zeigt, dass aus $Cov(X,Y) = 0$ nicht notwendigerweise die Unabhängigkeit von X und Y folgt.

Beispiel 96
Die Zufallsvariablen X und Y besitzen folgende gemeinsame Wahrscheinlichkeitsfunktion:

x \ y	0	1
-1	0	0.2
0	0.6	0
1	0	0.2

Es gilt

$$E(X) \;=\; (-1)\cdot 0.2 + 0\cdot 0.6 + 1\cdot 0.2 = 0$$

und

$$\begin{aligned} E(X\,Y) &= (-1)\cdot 0\cdot 0 + (-1)\cdot 1\cdot 0.2 \\ &+ 0\cdot 0\cdot 0.6 + 0\cdot 1\cdot 0 \\ &+ 1\cdot 0\cdot 0 + 1\cdot 1\cdot 0.2 \\ &= 0 \end{aligned}$$

Also gilt

$$Cov(X,Y) = 0.$$

Die Zufallsvariablen X und Y sind aber nicht unabhängig, denn

$$P(X=-1, Y=0) = 0 \quad \neq \quad 0.12 = P(X=-1)\,P(Y=0)$$

Vielmehr besteht zwischen X und Y mit Wahrscheinlichkeit 1 ein Zusammenhang.
Es gilt

$$\begin{aligned} P(Y=1\,|\,X=-1) &= 1 \\ P(Y=0\,|\,X=0) &= 1 \\ P(Y=1\,|\,X=1) &= 1 \end{aligned}$$

Also gilt

$$P(Y=X^2) = 1$$

□

Gleichung (9.10) auf Seite 287 zeigt, dass die Kovarianz von den Maßeinheiten der Variablen abhängt. Ist $Cov(X,Y)$ die Kovarianz zwischen dem Körpergewicht X, das in Kilogramm angegeben ist, und der Körpergröße Y, die in Metern bestimmt wurde, so gilt für die Kovarianz zwischen den beiden Variablen, wenn man das Körpergewicht in Gramm und die Körpergröße in Zentimetern bestimmt:

$$Cov(1000\cdot X, 100\cdot Y) = 100000\cdot Cov(X,Y)\,.$$

Wir können also nicht sagen, ob die Kovarianz groß oder klein ist. Dieses Problem kann man beheben, indem man zu den standardisierten Variablen übergeht.

Definition 9.2
Seien X und Y Zufallsvariablen mit Varianzen σ_X^2 und σ_Y^2 und Kovarianz $Cov(X,Y)$. Dann heißt

$$\boxed{\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X\,\sigma_Y}}$$

Korrelationskoeffizient zwischen X und Y.

Wir betrachten die standardisierten Zufallsvariablen

$$\tilde{X} = \frac{X - \mu_X}{\sigma_X}$$

und

$$\tilde{Y} = \frac{Y - \mu_Y}{\sigma_Y}\,,$$

wobei μ_X und μ_Y die Erwartungswerte sind. Der Korrelationskoeffizient $\rho_{X,Y}$ ist die Kovarianz der standardisierten Variablen. Mit den Gleichungen (9.9) und (9.10) auf Seite 287 gilt nämlich:

$$\begin{aligned} Cov(\tilde{X},\tilde{Y}) &= Cov\left(\frac{X-\mu_X}{\sigma_X}\,,\,\frac{Y-\mu_Y}{\sigma_Y}\right) \\ &= \frac{1}{\sigma_X\,\sigma_Y}\,Cov(X-\mu_X, Y-\mu_Y) \\ &= \frac{1}{\sigma_X\,\sigma_Y}\,Cov(X,Y) \\ &= \rho_{X,Y} \end{aligned}$$

Der folgende Satz gibt wichtige Eigenschaften des Korrelationskoeffizienten an.

Satz 9.5
Für den Korrelationskoeffizienten $\rho_{X,Y}$ zwischen den Zufallsvariablen X und Y gilt:

$$\boxed{-1 \le \rho_{X,Y} \le 1}$$

Dabei ist $\rho_{X,Y} = \pm 1$ genau dann, wenn Konstanten a und $b \neq 0$ existieren, so daß gilt

$$P(Y = a \pm b\,X) = 1.$$

Beweis:
Seien $\tilde{X}$ und $\tilde{Y}$ die standardisierten Variablen.
Dann gilt wegen (9.4) und (7.16):

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

Da die Varianz nichtnegativ ist, gilt

$$2 + 2\,\rho_{X,Y} \geq 0$$

und somit

$$\rho_{X,Y} \geq -1$$

Außerdem gilt wegen (9.4), (9.10) und (7.16)

$$Var(\tilde{X}-\tilde{Y}) = Var(\tilde{X}) + Var(\tilde{Y}) - 2\,Cov(\tilde{X},\tilde{Y}) = 2 - 2\,\rho_{X,Y}$$

Hieraus folgt

$$\rho_{X,Y} \leq 1$$

Also gilt

$$-1 \leq \rho_{X,Y} \leq 1$$

Ist $\rho_{X,Y} = 1$, so gilt

$$Var(\tilde{X}-\tilde{Y}) = 0.$$

Somit gilt

$$P(\tilde{X}-\tilde{Y}=0) = 1.$$

Also gilt

$$P(Y = a + b\,X) = 1$$

mit

$$a = \mu_Y - \frac{\sigma}{\sigma_Y}\,\mu$$

und

$$b = \frac{\sigma}{\sigma_Y}.$$

Eine analoge Beziehung erhält man für $\rho_{X,Y} = -1$.

Der Korrelationskoeffizient charakterisiert also den linearen Zusammenhang zwischen X und Y.

Kapitel 10

Verteilungsmodelle

Es gibt eine Reihe von Verteilungsmodellen für univariate diskrete und stetige Zufallsvariablen, die sich in der Praxis bewährt haben. Wir wollen uns von diesen einige anschauen.

10.1 Diskrete Verteilungsmodelle

Es gibt eine Reihe von Zugängen zu Verteilungsmodellen für diskrete Zufallsvariablen. Man kann die Wahrscheinlichkeitsfunktion einer diskreten Zufallsvariablen dadurch gewinnen, dass man den Träger $\mathcal{T}_\mathcal{X}$ vorgibt und für jedes $x \in \mathcal{T}_\mathcal{X}$ die Wahrscheinlichkeit $P(X = x)$ festlegt. Dabei müssen natürlich die Gleichungen 6.2 und 6.1 auf Seite 238 erfüllt sein. Eine andere Möglichkeit besteht darin, auf einem geeigneten Wahrscheinlichkeitsraum eine Zufallsvariable zu definieren. Schauen wir uns zunächst ein Beispiel für die erste Möglichkeit an.

10.1.1 Die Gleichverteilung

Definition 10.1
Die Zufallsvariable X heißt gleichverteilt, wenn gilt

$$\boxed{P(X = x) = \frac{1}{N}} \tag{10.1}$$

für $x = 1, \ldots, N$.

Für den Erwartungswert und die Varianz der diskreten Gleichverteilung gilt

$$\boxed{E(X) = \frac{N+1}{2}} \tag{10.2}$$

und

$$\boxed{Var(X) = \frac{N^2-1}{12}}\,. \tag{10.3}$$

Schauen wir uns zuerst den Erwartungswert an:

$$E(X) = \sum_{x=1}^{N} x\,\frac{1}{N} = \frac{1}{N}\sum_{x=1}^{N} x = \frac{1}{N}\frac{N(N+1)}{2} = \frac{N+1}{2}$$

Für die Varianz bestimmen wir zunächst

$$\begin{aligned} E(X^2) &= \sum_{x=1}^{N} x^2\,\frac{1}{N} = \frac{1}{N}\sum_{x=1}^{N} x^2 = \frac{1}{N}\frac{N(N+1)(2N+1)}{6} \\ &= \frac{(N+1)(2N+1)}{6}\,. \end{aligned}$$

Hierbei haben wir benutzt, das gilt

$$\sum_{x=1}^{N} x^2 = \frac{N(N+1)(2N+1)}{6}\,.$$

Die Varianz ist dann

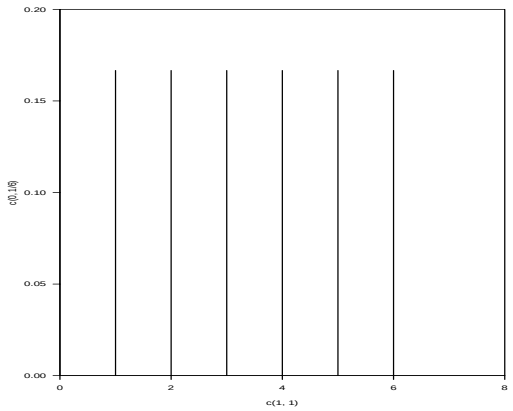
$$\begin{aligned} Var(X) &= E(X^2) - E(X)^2 = \frac{(N+1)(2N+1)}{6} - \frac{(N+1)^2}{4} \\ &= \frac{N+1}{12}[4N+2-3N-3] = \frac{(N+1)(N-1)}{12} = \frac{N^2-1}{12}\,. \end{aligned}$$

Beispiel 97
Die Gleichverteilung ist ein sinnvolles Modell für die Augenzahl beim einmaligen Wurf eines fairen Würfels. Hier gilt

$$P(X = x) = \frac{1}{6}$$

für $x = 1, 2, \ldots, 6$. Es gilt $E(X) = 3.5$ und $Var(X) = 35/12$.
Abbildung 10.1 zeigt die Wahrscheinlichkeitsfunktion.

Abbildung 10.1: Wahrscheinlichkeitsfunktion der Gleichverteilung



10.1.2 Vom Bernoulliprozess abgeleitete Verteilungen

Einer Reihe von Verteilungen liegt der sogenannte **Bernoulliprozess** zu Grunde.

Der Bernoulliprozess

Bei einem Zufallsvorgang sei ein Ereignis A mit $P(A) = p$ von Interesse. Man spricht auch von einem **Bernoullivorgang** und nennt das Ereignis A oft auch **Erfolg**. Einen **Bernoulliprozess** erhält man nun dadurch, dass man einen Bernoullivorgang mehrmals beobachtet, wobei folgende Annahmen getroffen werden:

- Die einzelnen Bernoullivorgänge sind voneinander unabhängig.
- Die Erfolgswahrscheinlichkeit p bleibt konstant.

Beispiel 98
Im Beispiel 79 auf Seite 242 haben wir ein Teilchen betrachtet, dass sich zufällig auf den ganzen Zahlen bewegt, wobei es im Nullpunkt startet. Jeder Schritt des Teilchens ist ein Bernoulliexperiment. Es gibt nur 2 Möglichkeiten. Das Teilchen geht entweder nach links, was wir mit A, oder es geht nach rechts, was wir mit $\bar{A}$ bezeichnen. Da das Teilchen sich zufällig entscheidet,

gilt $P(A) = p = 0.5$. Wenn wir unterstellen, dass die Entscheidung des Teilchens bei einem Schritt unabhängig von den anderen Schritten getroffen wird, und die Erfolgswahrscheinlichkeit p konstant bleibt, beobachten wir bei n Schritten einen Bernoulliprozess der Länge n.

Beim Bernoulliprozess sind für uns zwei Zufallsvariablen von Interesse:

- Anzahl der Erfolge bei n Durchführungen.
- Anzahl der Misserfolge vor dem ersten Erfolg

Schauen wir uns die Verteilungen dieser Zufallsvariablen im Detail an.

Die Binomialverteilung

Definition 10.2
Die Zufallsvariable X heißt binomialverteilt mit den Parametern n und p, wenn ihre Wahrscheinlichkeitsfunktion gegeben ist durch:

$$\boxed{P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}} \tag{10.4}$$

für $x = 0, 1, \ldots, n$.

Die Binomialverteilung erhält man, wenn man bei einem Bernoulliprozess der Länge n die Erfolge zählt. Dies kann man sich folgendermaßen klarmachen: Sei X die Anzahl der Erfolge bei einem Bernoulliprozess der Länge n. Die Zufallsvariable X kann die Werte $0, 1, 2, \ldots, n$ annehmen. Nimmt X den Wert x an, so wurden x Erfolge A und $n-x$ Mißerfolge $\bar{A}$ beobachtet. Aufgrund der Unabhängigkeit der einzelnen Durchführungen beträgt die Wahrscheinlichkeit einer Folge aus x Erfolgen A und $n-x$ Mißerfolgen $\bar{A}$ gleich $p^x (1-p)^{n-x}$. Eine Folge, die x-mal ein A und $n-x$-mal ein $\bar{A}$ enthält, ist eindeutig durch die Positionen der A festgelegt. Wir haben im Beispiel 65 auf Seite 216 gesehen, dass es

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

Möglichkeiten gibt, die Positionen der A zu wählen. Also gilt

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{für} \quad x = 0, 1, \ldots, n.$$

Für eine binomialverteilte Zufallsvariable X gilt

$$\boxed{E(X) = np}$$

und

$$\boxed{Var(X) = np(1-p)}.$$

Beispiel 98 (fortgesetzt)
Sei X die Anzahl der Schritte nach links. Die Zufallsvariable X ist binomialverteilt mit den Parametern $n = 3$ und $p = 0.5$. Es gilt für $x = 0, 1, 2, 3$:

$$P(X = x) = \binom{3}{x} 0.5^x \, (1 - 0.5)^{3-x} = \binom{3}{x} 0.5^3$$

Tabelle 10.1 zeigt die Wahrscheinlichkeitsfunktion.

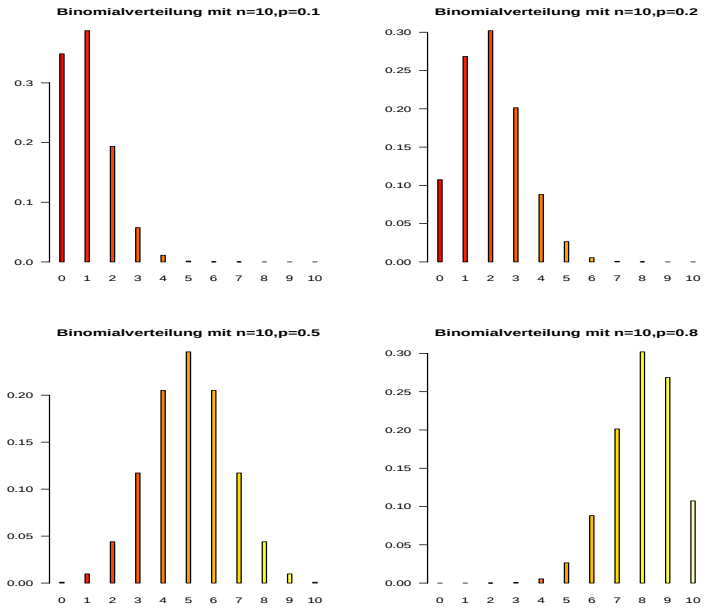
Tabelle 10.1: Wahrscheinlichkeitsfunktion der Binomialverteilung mit $n = 3$ und $p = 0.5$

x	0	1	2	3
$P(X = x)$	0.125	0.375	0.375	0.125

Außerdem gilt $E(X) = 1.5$ und $Var(X) = 0.75$. □

Abbildung 10.2 zeigt die Wahrscheinlichkeitsfunktion der Binomialverteilung für $n = 10$ und $p = 0.1$, $p = 0.2$, $p = 0.5$ und $p = 0.8$.
Wir sehen, dass die Wahrscheinlichkeitsfunktion der Binomialverteilung für $p = 0.5$ symmetrisch ist. Für $p < 0.5$ ist sie rechtsschief und für $p > 0.5$ linksschief.

Abbildung 10.2: Wahrscheinlichkeitsfunktion der Binomialverteilung für $n = 10$ und $p = 0.1$, $p = 0.2$, $p = 0.5$ und $p = 0.8$



Die Binomialverteilung mit $n = 1$ heißt Bernoulliverteilung. Die Wahrscheinlichkeitsfunktion der Bernoulliverteilung ist gegeben durch:

$$\boxed{P(X = x) = p^x \, (1 - p)^{1-x}}$$

für $x = 0, 1$.
Es gilt also

$$P(X = 0) = 1 - p$$

und

$$P(X = 1) = p.$$

Eine bernoulliverteilte Zufallsvariable X zählt, wie oft A bei einem Bernoullivorgang eingetreten ist. Für eine bernoulliverteilte Zufallsvariable X gilt $E(X) = p$ und $Var(X) = p(1 - p)$. Schauen wir uns erst den Erwartungswert an:

$$E(X) = 0 \cdot (1 - p) + 1 \cdot p = p \, .$$

Für die Varianz bestimmen wir

$$E(X^2) = 0 \cdot (1 - p) + 1 \cdot p = p$$

und erhalten

$$Var(X) = E(X^2) - E(X)^2 = p - p^2 = p(1-p)\,.$$

Die geometrische Verteilung

Eine bernoulliverteilte Zufallsvariable zählt die Anzahl der Erfolge bei einem Bernoulliprozess der Länge n. Schauen wir uns die Wahrscheinlichkeitsfunktion der Misserfolge vor dem ersten Erfolg an.

Definition 10.3
Die Zufallsvariable X heißt geometrisch verteilt mit Parameter p, wenn ihre Wahrscheinlichkeitsfunktion gegeben ist durch

$$\boxed{P(X = x) = p\,(1-p)^x \quad \text{für} \quad x = 0, 1, \ldots} \tag{10.5}$$

Die geometrische Verteilung erhält man, wenn man bei einem Bernoulliprozess die Anzahl der Misserfolge vor dem ersten Erfolg zählt. Dies kann man sich folgendermaßen klarmachen:
Sei X die Anzahl der Misserfolge vor dem ersten Erfolg bei einem Bernoulliprozess. Die Zufallsvariable X kann die Werte $0, 1, 2, \ldots$ annehmen. Nimmt X den Wert x an, so wurden x Mißerfolge $\bar{A}$ und genau ein Erfolg A beobachtet, wobei das A das letzte Symbol in der Folge ist. Wir erhalten also

$$\underbrace{\bar{A}\bar{A}\ldots\bar{A}}_{x-mal} A$$

Da die einzelnen Durchführungen unabhängig sind, beträgt die Wahrscheinlichkeit dieser Folge

$$\underbrace{(1-p)(1-p)\cdots(1-p)}_{x-mal}\,p = (1-p)^x p.$$

Ist X geometrisch verteilt mit Prameter p, so gilt

$$\boxed{E(X) = \frac{1-p}{p}}$$

und

$$\boxed{Var(X) = \frac{1-p}{p^2}}\,.$$

Beispiel 98 (fortgesetzt)
Wir warten so lange, bis das Teilchen zum ersten Mal nach links geht, und zählen die Anzahl X der Schritte nach rechts vor dem ersten Schritt nach links. Die Zufallsvariable X ist geometrisch verteilt mit $p = 0.5$. Es gilt

$$P(X = x) = 0.5^{x+1} \quad \text{für} \quad x = 0, 1, 2, ...$$

Außerdem gilt $E(X) = 1$ und $Var(X) = 2$. □

Oft betrachtet man anstatt der Anzahl X der Misserfolge die Anzahl Y der Versuche bis zum ersten Erfolg, wobei man den Erfolg mitzählt. Offensichtlich gilt

$$Y = X + 1 .$$

Die Wahrscheinlichkeitsfunktion von Y ist:

$$\boxed{P(Y = y) = p\,(1 - p)^{y-1} \quad \text{für} \quad y = 1, \dots} . \tag{10.6}$$

Ausßerdem gilt

$$\boxed{E(Y) = \frac{1}{p}}$$

und

$$\boxed{Var(Y) = \frac{1 - p}{p^2}} .$$

10.1.3 Die hypergeometrische Verteilung

Wir haben die Binomialverteilung aus dem Bernoulliprozess hergeleitet. Man kann sie aber auch aus einem einfachen Urnenmodell gewinnen. Hierzu betrachten wir eine Urne, die N Kugeln enthält, von denen M weiß und die restlichen $N - M$ schwarz sind. Der Anteil p der weißen Kugeln ist also M/N. Zieht man n Kugeln **mit** Zurücklegen, so ist die Anzahl X der weißen Kugeln mit den Parametern n und p binomialverteilt. Zieht man **ohne** Zurücklegen aus der Urne, so besitzt die Anzahl X der weißen Kugeln eine hypergeometrische Verteilung. Die Wahrscheinlichkeit $P(X = x)$ erhalten wir in diesem Fall folgendermaßen:
Insgesamt gibt es $\binom{N}{n}$ Möglichkeiten, von den N Kugeln n ohne Zurücklegen auszuwählen. Es gibt $\binom{M}{x}$ Möglichkeiten, von den M weißen Kugeln x Kugeln

auszuwählen, und es gibt $\binom{N-M}{n-x}$ Möglichkeiten, von den $N-M$ schwarzen Kugeln $n-x$ Kugeln auszuwählen. Somit gilt

$$\boxed{P(X=x)=\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}}$$

für $\max\{0, n-(N-M)\} \leq x \leq \min\{n, N\}$.

Definition 10.4
Die Zufallsvariable X heißt hypergeometrisch verteilt mit den Parametern N, M und n, wenn ihre Wahrscheinlichkeitsfunktion gegeben ist durch

$$P(X=x)=\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \tag{10.7}$$

für $\max\{0, n-(N-M)\} \leq x \leq \min\{n, N\}$.

Es gilt

$$\boxed{E(X)=n\,\frac{M}{N}}$$

und

$$\boxed{Var(X)=n\,\frac{M}{N}\,\frac{N-M}{N}\,\frac{N-n}{N-1}}\,.$$

Setzen wir $p=M/N$, so gilt

$$E(X)=n\,p$$

und

$$Var(X)=n\,p(1-p)\,\frac{N-n}{N-1}\,.$$

Der Erwartungswert der Binomialverteilung und der hypergeometrischen Verteilung sind also identisch. Für $n>1$ ist $\frac{N-n}{N-1}$ kleiner als 1. Somit ist für $n>1$ die Varianz der hypergeometrischen Verteilung kleiner als die Varianz der Binomialverteilung.

Beispiel 99
Eine Urne enthält 10 Kugeln, von denen 4 weiß sind. Es werden zuerst 3 Kugeln mit Zurücklegen und dann 3 Kugeln ohne Zurücklegen gezogen. Sei X die Anzahl der weißen Kugeln beim Ziehen mit Zurücklegen und Y die Anzahl der weißen Kugeln beim Ziehen ohne Zurücklegen. Es gilt

$$P(X = x) = \binom{3}{x} 0.4^x \, 0.6^{3-x}$$

und

$$P(Y = y) = \frac{\binom{4}{y}\binom{6}{3-y}}{\binom{10}{3}}.$$

Tabelle 10.2 zeigt die Wahrscheinlichkeitsverteilungen der Zufallsvariablen X und Y.

Tabelle 10.2: Wahrscheinlichkeitsverteilung der hypergeometrischen und der Binomialverteilung

Wert	Binomial-verteilung	Hypergeometrische Verteilung
0	0.216	0.167
1	0.432	0.500
2	0.288	0.300
3	0.064	0.033

Wir sehen an der Tabelle, warum bei der Binomialverteilung die Varianz größer ist als bei der hypergeometrischen Verteilung. Die extremen Werte 0 und 3 sind bei der Binomialverteilung wahrscheinlicher als bei der hypergeometrischen Verteilung.

Was passiert mit der Wahrscheinlichkeitsfunktion der hypergeometrischen Verteilung, wenn N beliebig groß wird, wobei aber $p = \frac{M}{N}$ konstant bleibt?

Es gilt:

$$\begin{aligned} P(X=x) &= \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} = \frac{(M)_x\,(N-M)_{n-x}\,n!}{x!\,(n-x)!\,(N)_n} \\ &= \binom{n}{x}\frac{(M)_x\,(N-M)_{n-x}}{(N)_n} \\ &= \binom{n}{x}\frac{M}{N}\cdots\frac{M-x+1}{N-x+1}\frac{N-M}{N-x}\cdots\frac{N-M-n+x+1}{N-n+1} \end{aligned}$$

Also gilt

$$\begin{aligned} &\lim_{N\to\infty} P(X=x) \\ &= \lim_{N\to\infty}\binom{n}{x}\frac{M}{N}\frac{M-1}{N-1}\cdots\frac{M-x+1}{N-x+1}\frac{N-M}{N-x}\cdots\frac{N-M-n+x+1}{N-n+1} \\ &= \lim_{N\to\infty}\binom{n}{x}\frac{M}{N}\frac{\frac{M}{N}-\frac{1}{N}}{1-\frac{1}{N}}\cdots\frac{\frac{M}{N}-\frac{x-1}{N}}{1-\frac{x-1}{N}}\frac{1-\frac{M}{N}}{1-\frac{x}{N}}\cdots\frac{1-\frac{M}{N}-\frac{n-x-1}{N}}{1-\frac{n-1}{N}} \\ &= \lim_{N\to\infty}\binom{n}{x}p\frac{p-\frac{1}{N}}{1-\frac{1}{N}}\cdots\frac{p-\frac{x-1}{N}}{1-\frac{x-1}{N}}\frac{1-p}{1-\frac{x}{N}}\cdots\frac{1-p-\frac{n-x-1}{N}}{1-\frac{n-1}{N}} \\ &= \binom{n}{x}p^x\,(1-p)^{n-x} \end{aligned}$$

Für große Werte von N können wir also die hypergeometrische Verteilung durch die Binomialverteilung approximieren.

Außerdem bestätigt die Aussage die intuitive Annahme, dass es keinen Unterschied macht, ob man aus einer sehr großen Grundgesamtheit mit oder ohne Zurücklegen zieht.

10.1.4 Die Poissonverteilung

Schauen wir uns an, was mit der Wahrscheinlichkeitsfunktion der Binomialverteilung geschieht, wenn wir n beliebig groß werden lassen. Auch hier müssen wir beim Grenzübergang eine Restriktion berücksichtigen, da sonst die Wahrscheinlichkeit gegen Null geht. Wir fordern, dass $\lambda = np$ konstant bleibt. Mit

$$p = \frac{\lambda}{n}.$$

gilt also

$$\begin{aligned}
\lim_{n\to\infty} P(X = x) &= \lim_{n\to\infty} \binom{n}{x} p^x (1-p)^{n-x} \\
&= \lim_{n\to\infty} \frac{n!}{x!(n-x)!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\
&= \frac{\lambda^x}{x!} \lim_{n\to\infty} \frac{(n)_x}{n^x} \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^{-x} \\
&= \frac{\lambda^x}{x!}\, e^{-\lambda}
\end{aligned}$$

da gilt

$$\begin{aligned}
\lim_{n\to\infty} \frac{(n)_x}{n^x} &= \lim_{n\to\infty} \frac{n}{n} \frac{n-1}{n} \cdots \frac{n-x+1}{n} = 1 \\
\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n &= e^{-\lambda} \\
\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^{-x} &= 1
\end{aligned}$$

Definition 10.5
Die Zufallsvariable X heißt poissonverteilt mit dem Parameter λ, wenn ihre Wahrscheinlichkeitsfunktion gegeben ist durch:

$$\boxed{P(X = x) = \frac{\lambda^x}{x!}\, e^{-\lambda}} \tag{10.8}$$

für $x = 0, 1, \ldots$.

Die Poissonverteilung wird auch die Verteilung der seltenen Ereignisse genannt, da $p = \frac{\lambda}{n}$ mit wachsenem n immer kleiner wird.
Ist X poissonverteilt mit Parameter λ, so gilt

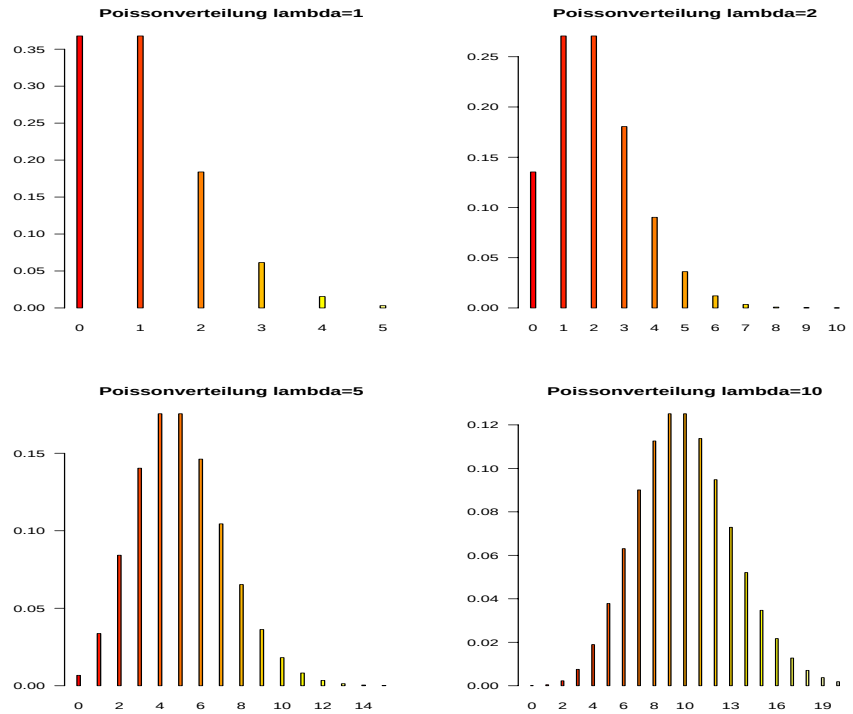
$$\boxed{E(X) = \lambda}$$

und

$$\boxed{Var(X) = \lambda}\ .$$

Abbildung 10.3 zeigt die Wahrscheinlichkeitsfunktion der Poissonverteilung für $\lambda = 1$, $\lambda = 2$, $\lambda = 5$ und $\lambda = 10$.

Abbildung 10.3: Wahrscheinlichkeitsfunktion Poissonverteilung für $\lambda = 1$, $\lambda = 2$, $\lambda = 5$ und $\lambda = 10$



Wir sehen, dass die Wahrscheinlichkeitsfunktion der Poissonverteilung mit wachsendem λ immer symmetrischer wird.

Beispiel 100
Wir haben im Beispiel 76 auf Seite 238 die Anzahl der Tore betrachtet, die in der Saison 2001/2002 in der Fußballbundesliga je Spiel geschossen wurden. Dabei wurden nur Spiele betrachtet, in denen höchstens 5 Tore fielen. Es wurden aber auch mehr Tore geschossen. Tabelle 10.3 zeigt die Häufigkeitsverteilung der Anzahl Tore in allen 306 Spielen.

Tabelle 10.3: Häufigkeitstabelle der Anzahl der Tore in einem Bundesligaspiel in der Saison 2001/2002

x	0	1	2	3	4	5	6	7
$n(X = x)$	23	35	74	69	56	23	19	7
$h(X = x)$	0.075	0.114	0.242	0.225	0.183	0.075	0.062	0.023

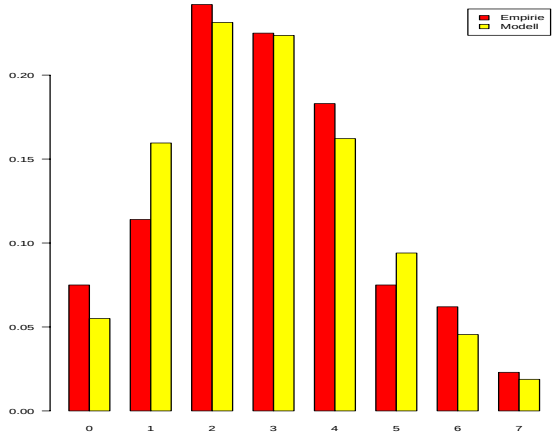
Der Mittelwert der Anzahl Tore beträgt 2.9. Tabelle 10.4 zeigt die Wahrscheinlichkeitsverteilung der Poissonverteilung mit $\lambda = 2.9$.

Tabelle 10.4: Wahrscheinlichkeitsverteilung der Poissonverteilung mit Parameter $\lambda = 2.9$

x	0	1	2	3	4	5	6	7
$P(X = x)$	0.055	0.160	0.231	0.224	0.162	0.094	0.045	0.019

Wir sehen, dass die Wahrscheinlichkeiten und relativen Häufigkeiten gut übereinstimmen. Dies bestätigt auch die Abbildung 10.4, in der beiden Verteilungen gegenübergestellt sind.
Wir werden später lernen, warum es sinnvoll ist, für λ den Mittelwert der Beobachtungen zu wählen.

Abbildung 10.4: Stabdiagramm der Anzahl der Tore mit Wahrscheinlichkeitsfunktion der Poissonverteilung mit Parameter $\lambda = 2.9$



Warum die Poissonverteilung ein sinnvolles Modell für die Anzahl der in einem Spiel geschossenen Tore ist, zeigen die folgenden Überlegungen.
Die Poissonverteilung spielt eine zentrale Rolle bei Ankunftsprozessen. Hier soll die Anzahl $X(a,t)$ der Ankünfte im Intervall $(a, a+t]$ modelliert werden. Dabei geht man von folgenden Annahmen aus:

1. $X(a,t)$ und $X(a^*,t^*)$ sind unabhängig, wenn $(a, a+t]$ und $(a^*, a^*+t^*]$ disjunkt sind.

2. $X(a,h)$ hängt von h, aber nicht von a ab.

3. $$P(X(a,h) = 1) \approx \lambda h\,.$$

4. $$P(X(a,h) > 1) \approx 0\,.$$

Annahme 1 besagt, dass die Häufigkeit des Auftretens von A in einem Intervall unabhängig ist von der Häufigkeit des Auftretens von A in einem dazu disjunkten Intervall.
Annahme 2 besagt, dass die Häufigkeit des Auftretens von A in einem Intervall der Länge h nur von der Länge, aber nicht von der Lage des Intervalls abhängt.

Annahme 3 besagt, dass die Wahrscheinlichkeit, dass A in einem sehr kleinen Intervall der Länge h genau einmal eintritt, proportional zur Länge des Intervalls ist.
Annahme 4 besagt, dass die Wahrscheinlichkeit, dass A in einem sehr kleinen Intervall der Länge h mehr als einmal eintritt, vernachlässigt werden kann.
Schauen wir uns dafür ein Beispiel an:

Beispiel 101
Der Verkehrsstrom einer Fernstraße wird an einem festen Beobachtungspunkt protokolliert, wobei jeweils festgehalten wird, wann ein Fahrzeug den Beobachtungspunkt passiert.
Annahme 1 besagt, dass der Verkehrsstrom völlig regellos ist. Ereignisse in nichtüberlappenden Intervallen sind unabhängig.
Annahme 2 besagt, dass der Verkehrsstrom von konstanter Intensität ist. Die Wahrscheinlichkeit, dass in einem Zeitintervall der Länge h k Fahrzeuge eintreffen, hängt von der Länge des Intervalls, nicht jedoch von seiner Lage ab.
Annahme 4 besagt, dass das Auftreten geschlossener Fahrzeuggruppen vernachlässigt werden kann.

Unter den Annahmen 1 bis 4 ist $X(a,t)$ poissonverteilt mit dem Parameter λt. Es gilt also

$$P(X(a,t)=x)=\frac{(\lambda t)^x}{x!}\,e^{-\lambda t} \quad \text{für} \quad x=0,1,\ldots.$$

Dies sieht man folgendermaßen:
Wir teilen das Intervall $(a,a+t]$ in n gleichgroße Intervalle der Länge $h=\frac{t}{n}$. Aufgrund von Annahme 4 kann A in jedem dieser Intervalle nur einmal oder keinmal eintreten. In jedem dieser Intervalle beträgt die Wahrscheinlichkeit, A genau einmal zu beobachten,

$$p=\lambda h=\frac{\lambda t}{n}$$

Da die Intervalle disjunkt sind, ist das Eintreten von A in den einzelnen Intervallen unabhängig. Außerdem ist jedes Intervall gleich lang, so dass p konstant ist.
Es liegt also ein Bernoulliprozess der Länge n mit Erfolgswahrscheinlichkeit $p=\frac{\lambda t}{n}$ vor. Somit gilt

$$P(X(a,t)=x)=\binom{n}{x}\left(\frac{\lambda t}{n}\right)^x\left(1-\frac{\lambda t}{n}\right)^{n-x}$$

Lassen wir n über alle Grenzen wachsen, so erhalten wir

$$P(X(a,t)=x)=\frac{(\lambda t)^x}{x!}\,e^{-\lambda t}\quad\text{für}\quad x=0,1,\ldots.$$

10.2 Stetige Verteilungsmodelle

Man erhält für eine stetige Zufallsvariable ein Modell, indem man eine Funktion $f_X(x)$vorgibt, die die Gleichungen (6.9)und (6.10) auf Seite (245) erfüllen.

10.2.1 Die Gleichverteilung

Definition 10.6
Die Zufallsvariable X heißt gleichverteilt auf dem Intervall $[a,b]$, wenn ihre Dichtefunktion gegeben ist durch:

$$\boxed{f_X(x)=\begin{cases}\dfrac{1}{b-a} & \text{für}\quad a\le x\le b\\ 0 & \text{sonst}\end{cases}} \tag{10.9}$$

Im Beispiel 80 auf Seite 245 haben wir die Gleichverteilung auf $[0,10]$ näher betrachtet. Die Dichtefunktion ist in Abbildung 6.3 auf Seite 246 zu finden.
Die Verteilungsfunktion $F_X(x)$ ist gegeben durch:

$$\boxed{F_X(x)=\begin{cases}0 & \text{für } x<a\\ \dfrac{x-a}{b-a} & \text{für } a\le x\le b\\ 1 & \text{für } x>b\end{cases}}$$

Die Verteilungsfunktion der Gleichverteilung auf $[0,10]$ ist in Abbildung 6.4 auf Seite 247 zu finden.
Aus der Verteilungsfunktion können wir problemlos die Quantile bestimmen. Es gilt

$$x_p \;=\; a+p(b-a)\,. \tag{10.10}$$

Wir erhalten Gleichung (10.10), indem wir folgende Gleichung nach x_p auflösen:

$$\frac{x_p-a}{b-a}=p\,.$$

Außerdem gilt

$$\boxed{E(X) = \frac{a+b}{2}} \tag{10.11}$$

und

$$\boxed{Var(X) = \frac{(a-b)^2}{12}} \tag{10.12}$$

Dies sieht man folgendermaßen:

$$\begin{aligned} E(X) &= \int_a^b x \, \frac{1}{b-a} \, dx = \frac{1}{b-a} \int_a^b x \, dx = \frac{1}{b-a} \left[\frac{x^2}{2}\right]_a^b \\ &= \frac{1}{b-a} \, \frac{b^2-a^2}{2} = \frac{1}{b-a} \, \frac{(b+a)\,(b-a)}{2} = \frac{a+b}{2} \end{aligned}$$

Für die Varianz bestimmen wir zunächst $E(X^2)$. Es gilt

$$\begin{aligned} E(X^2) &= \int_a^b x^2 \frac{1}{b-a} \, dx = \frac{1}{b-a} \int_a^b x^2 \, dx = \frac{1}{b-a} \left[\frac{x^3}{3}\right]_a^b \\ &= \frac{b^3-a^3}{3(b-a)} = \frac{(b-a)(a^2+ab+b^2)}{3(b-a)} = \frac{a^2+ab+b^2}{3} \end{aligned}$$

Also gilt

$$\begin{aligned} Var(X) &= E(X^2) - E(X)^2 = \frac{a^2+ab+b^2}{3} - \frac{(a+b)^2}{4} \\ &= \frac{4a^2+4ab+4b^2-3a^2-6ab-3b^2}{12} = \frac{a^2-2ab+b^2}{12} \\ &= \frac{(a-b)^2}{12} \end{aligned}$$

Eine zentrale Rolle spielt die Gleichverteilung auf $[0,1]$. Bei dieser gilt $a = 0$ und $b = 1$.

10.2.2 Die Normalverteilung

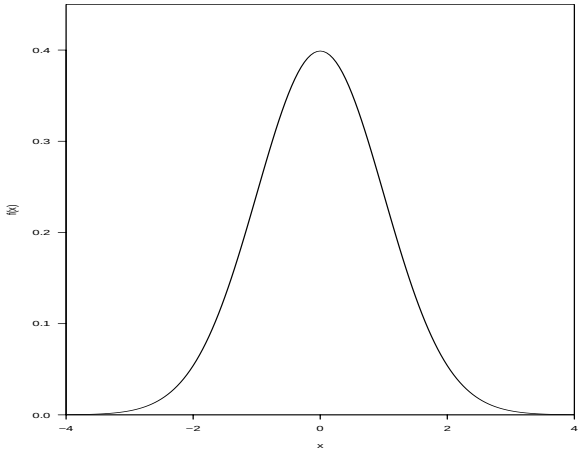
Das wichtigste Verteilungsmodell ist die Normalverteilung, die von Gauss vorgeschlagen wurde.

Definition 10.7
Die Zufallsvariable X heißt normalverteilt mit den Parametern μ und σ^2, wenn ihre Dichtefunktion für $x \in I\!R$ gegeben ist durch:

$$\boxed{f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}} \tag{10.13}$$

Abbildung 10.5 zeigt die Dichtefunktion der Normalverteilung mit $\mu = 0$ und $\sigma = 1$, die **Standardnormalverteilung** heißt.

Abbildung 10.5: Dichtefunktion der Standardnormalverteilung



Für eine mit den Parametern μ und σ^2 normalverteilte Zufallsvariable X gilt

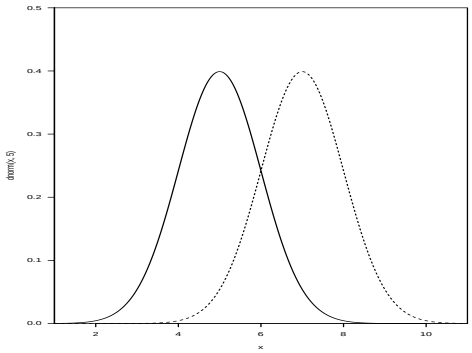
$$\boxed{E(X) = \mu}$$

und

$$\boxed{Var(X) = \sigma^2}$$

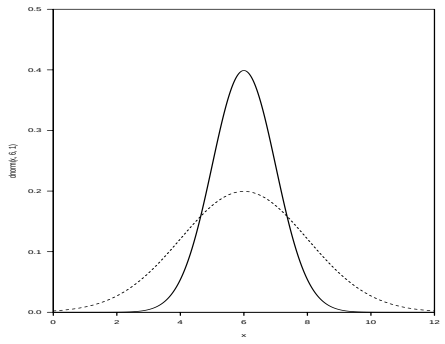
Der Parameter μ beschreibt also die Lage und der Parameter σ^2 die Streuung der Verteilung.
Abbildung 10.6 zeigt die Dichtefunktionen von zwei Normalverteilungen mit unterschiedlichem Erwartungswert und identischen Varianzen.

Abbildung 10.6: Dichtefunktionen der Normalverteilung $\mu = 5$ und $\mu = 6$ und gleichem $\sigma^2 = 1$



Wir sehen, dass die Gestalten der Dichtefunktionen identisch sind und die Dichtefunktion mit dem größeren Erwartungswert nach rechts verschoben ist.
Abbildung 10.7 zeigt die Dichtefunktionen von zwei Normalverteilungen mit identischem Erwartungswert und unterschiedlichen Varianzen.

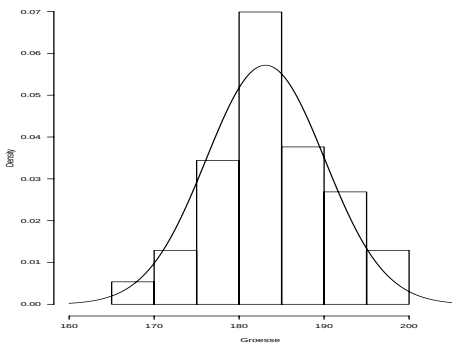
Abbildung 10.7: Dichtefunktionen der Normalverteilung $\sigma^2 = 1$ und $\sigma^2 = 4$ und gleichem $\mu = 5$



Wir sehen, dass die Dichtefunktion mit der kleineren Varianz viel flacher ist.

Beispiel 102
In Tabelle 3.19 auf Seite 129 ist die Verteilung der Körpergröße von männlichen Studienanfängern zu finden. In Abbildung 10.8 ist neben dem Histogramm noch die Dichtefunktion der Normalverteilung mit den Parametern $\mu = 183.1$ und $\sigma^2 = 48.7$ eingezeichnet. Wir sehen, dass die Normalverteilung ein geeignetes Modell für die Körpergröße ist.

Abbildung 10.8: Histogramm der Körpergröße von männlichen Studienanfängern mit Dichtefunktion der Normalverteilung mit den Parametern $\mu = 183.1$ und $\sigma^2 = 48.7$



Die Verteilungsfunktion $F_X(x)$ kann nicht in expliziter Form angegeben werden. Um Wahrscheinlichkeiten zu bestimmen, benötigt man also Tabellen. Es muss aber nicht jede Normalverteilung tabelliert werden, sondern es reicht aus, Tabellen der Standardnormalverteilung zu besitzen.
Ist nämlich X normalverteilt mit den Parametern μ und σ^2, so ist

$$\boxed{Z = \frac{X - \mu}{\sigma}}$$

standardnormalverteilt. Wir wollen diese Beziehung nicht beweisen, sondern ihre Konsequenzen aufzeigen. Dabei bezeichnen wir die Dichtefunktion einer standardnormalverteilten Zufallsvariablen Z mit $\phi(z)$ und dieVerteilungsfunktion mit $\Phi(z)$. Es gilt also

$$\boxed{\phi(z) = \frac{1}{\sqrt{2\pi}}\, e^{-0.5\, z^2}} \tag{10.14}$$

und

$$\boxed{\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}\, e^{-0.5\,u^2}\, du} \tag{10.15}$$

Wir suchen für eine mit den Parametern μ und σ^2 normalverteilte Zufallsvariable folgende Wahrscheinlichkeit:

$$P(X \leq x) = F_X(x)$$

Es gilt

$$\boxed{F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)} \tag{10.16}$$

Dies sieht man folgendermaßen:

$$\begin{aligned} F_X(x) &= P(X \leq x) = P(X - \mu \leq x - \mu) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) \\ &= P\left(Z \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right) \end{aligned}$$

Man muss die Verteilungsfunktion der Standardnormalverteilung nur für positive oder negative Werte von z tabellieren. Es gilt nämlich

$$\boxed{\Phi(z) = 1 - \Phi(-z)}\,. \tag{10.17}$$

Beispiel 103
Die Fahrzeit X eines Studenten zur Universität ist normalverteilt mit Erwartungswert 40 und Varianz 4. Wie groß ist die Wahrscheinlichkeit, dass er höchstens 36 Minuten braucht?
Gesucht ist $P(X \leq 36)$. Es gilt

$$P(X \leq 36) = F_X(36) = \Phi\left(\frac{36-40}{2}\right) = \Phi(-2)\,.$$

Tabelle 20.1 auf Seite 546 entnehmen wir $\Phi(2) = 0.977$. Also gilt

$$P(X \leq 36) = \Phi(-2) = 1 - \Phi(2) = 1 - 0.977 = 0.023\,.$$

Mit Hilfe von Gleichung (7.23) auf Seite 265 können wir das p-Quantil x_p einer mit den Parametern μ und σ^2 normalverteilten Zufallsvariablen aus dem p-Quantil z_p der Standardnormalverteilung folgendermaßen bestimmen:

$$\boxed{x_p = \mu + z_p \sigma}$$

Man muss die Quantile der Standardnormalverteilung nur für Werte von p tabellieren, die kleiner als 0.5 oder größer als 0.5 sind, da gilt

$$\boxed{z_p = -z_{1-p}} \tag{10.18}$$

Beispiel 103 (fortgesetzt)
Welche Fahrzeit wird an 20 Prozent der Tage nicht überschritten? Gesucht ist $x_{0.2}$. Es gilt $z_{0.2} = -z_{0.8}$. Tabelle 20.3 auf Seite 549 entnehmen wir $z_{0.8} = 0.842$. Also gilt $z_{0.2} = -0.842$. Somit erhalten wir

$$x_{0.20} = 40 + z_{0.20} \cdot 2 = 40 - 0.842 \cdot 2 = 38.316\,.$$

□

Bei der Tschebyscheff-Ungleichung haben wir Intervalle der Form

$$\boxed{[\mu - k\,\sigma, \mu + k\,\sigma]} \tag{10.19}$$

betrachtet. Man nennt das Intervall in Gleichung (10.19) auch das k-fache zentrale Schwankungsintervall. Die Wahrscheinlichkeit für das k-fache zentrale Schwankungsintervall bei Normalverteilung ist

$$\boxed{P(\mu - k\,\sigma \leq X \leq \mu + k\,\sigma) = \Phi(k) - \Phi(-k)}\,.$$

In der Tabelle 10.5 sind die Wahrscheinlichkeiten

$$P(\mu - k\,\sigma \leq X \leq \mu + k\,\sigma)$$

bei Normalverteilung den unteren Schranken gegenübergestellt, die sich aus der Tschebyscheff-Ungleichung ergeben.

Tabelle 10.5: Wahrscheinlichkeiten des k-fachen zentralen Schwankungsintervalls bei Normalverteilung und Tschebyscheff

k	$N(\mu, \sigma^2)$	Tschebyscheff
1	0.683	0
2	0.954	0.750
3	0.997	0.889

10.2.3 Die Exponentialverteilung

Kommen wir noch einmal zum Poisson-Prozess zurück. Bei einem Poisson-Prozess im Intervall $(0, t]$ ist die absolute Häufigkeit des Ereignisses A poissonverteilt mit Parameter λt.
Es gilt also

$$P(X = x) = \frac{(\lambda t)^x}{x!}\, e^{-\lambda t} \quad \text{für} \quad x = 0, 1, \ldots .$$

Wir wollen nun einen Poisson-Prozess so lange beobachten, bis A zum ersten Mal eintritt. Gesucht ist die Dichtefunktion $f_T(t)$ und die Verteilungsfunktion $F_X(x)$ der Wartezeit T bis zum ersten Eintreten von A. Für $t < 0$ gilt

$$F_T(t) = 0.$$

Für $t \geq 0$ gilt:

$$F_T(t) \;=\; P(T \leq t) = 1 - P(T > t) = 1 - P(X = 0) = 1 - e^{-\lambda t}$$

Somit gilt:

$$F_T(t) = \begin{cases} 1 - e^{-\lambda t} & \text{für } t > 0 \\ \\ 0 & \text{sonst} \end{cases}$$

Da die erste Ableitung der Verteilungsfunktion gleich der Dichtefunktion ist, erhalten wir folgende

Definition 10.8
Die Zufallsvariable X heißt exponentialverteilt mit Parameter λ, wenn ihre Dichtefunktion gegeben ist durch:

$$\boxed{f_X(x) = \begin{cases} \lambda\, e^{-\lambda x} & \text{für } x > 0 \\ 0 & \text{sonst} \end{cases}} \tag{10.20}$$

Abbildung 6.5 auf Seite 248 zeigt die Dichtefunktion der Exponentialverteilung mit $\lambda = 1$. Die Exponentialverteilung ist eine rechtsschiefe Verteilung. Es gilt

$$\boxed{E(X) = \frac{1}{\lambda}}$$

und

$$\boxed{Var(X) = \frac{1}{\lambda^2}}$$

Für das p-Quantil der Exponentialverteilung gilt

$$\boxed{x_p = -\frac{1}{\lambda} \ln(1-p)}. \tag{10.21}$$

Wir erhalten Gleichung (10.21) folgendermaßen:

$$1 - e^{-\lambda x_p} = p \Leftrightarrow e^{-\lambda x_p} = 1 - p \Leftrightarrow -\lambda x_p = \ln(1-p) \Leftrightarrow x_p = -\frac{1}{\lambda} \ln(1-p)$$

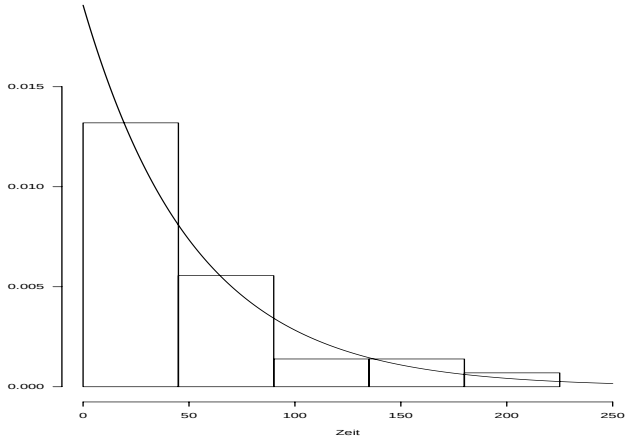
Beispiel 104
Bayer und Jankovic beobachteten im Rahmen eines BI-Projektes 30 Minuten eine Tankstelle und bestimmten die Zeiten zwischen den Ankünften von Kunden. Tabelle 10.6 zeigt die Häufigkeitstabelle.

Tabelle 10.6: Häufigkeitstabelle der Zwischenankunftszeiten an einer Tankstelle

Zeit	absolute Häufigkeit
von 0 bis unter 45	19
von 45 bis unter 90	8
von 90 bis unter 135	2
von 135 bis unter 180	2
von 180 bis unter 225	1

Abbildung 10.9 zeigt das Histogramm mit der Dichtefunktion der Exponentialverteilung mit Parameter $\lambda = 0.019$. Wir sehen, dass die Anpassung gut ist.

Abbildung 10.9: Histogramm der Wartezeit mit der Dichtefunktion der Exponentialveretilung mit Parameter $\lambda = 0.019$



10.2.4 Prüfverteilungen

In der schließenden Statistik verwendet man immer wieder eine Reihe von Verteilungen, die eine Beziehung zur Normalverteilung haben. Schauen wir uns diese an.

Die Chi-Quadrat-Verteilung

Wir wissen, dass $Z = (X - \mu)/\sigma$ standardnormalverteilt ist, wenn X mit den Parametern μ und σ^2 normalverteilt ist. Die Zufallsvariable Z^2 ist in diesem Fall chiquadratverteilt mit $k = 1$ Freiheitsgraden. Man nennt den Parameter der Chi-Quadrat-Verteilung also Freiheitsgrade.

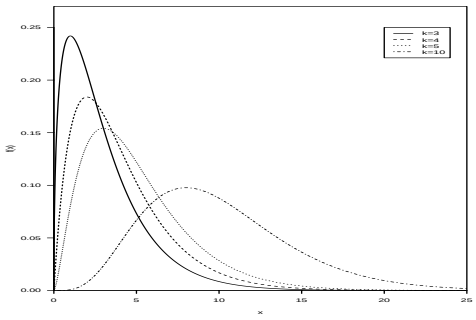
Oft betrachtet man k unabhängige standardnormalverteilte Zufallsvariablen $Z_1, \ldots, Z_k$. In diesem Fall ist

$$\sum_{i=1}^{k} Z_i^2$$

chiquadratverteilt mit k Freiheitsgraden.

Abbildung 10.10 zeigt die Dichtefunktion der Chi-Quadrat-Verteilung mit $k = 3$, $k = 4$, $k = 5$ und $k = 10$ Freiheitsgraden.

Abbildung 10.10: Dichtefunktion der Chi-Quadrat-Verteilung mit $k = 3$, $k = 4$, $k = 5$ und $k = 10$ Freiheitsgraden



Wir sehen, dass die Dichtefunktion der Chi-Quadrat-Verteilung mit wachsender Zahl von Freiheitsgraden immer symmetrischer wird.

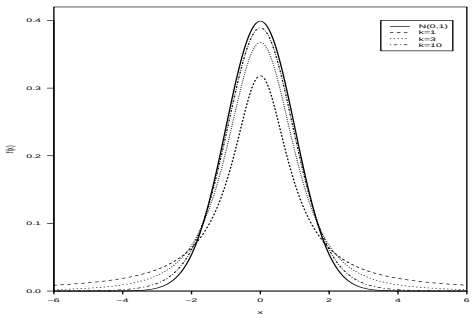
Die t-Verteilung

Die Zufallsvariable Z sei standardnormalverteilt und die Zufallsvariable V chiquadratverteilt mit k Freiheitsgraden. Sind Z und V unabhängig, so ist die Zufallsvariable

$$T = \frac{Z}{\sqrt{V/k}}$$

t-verteilt mit k Freiheitsgraden. Abbildung 10.11 zeigt die Dichtefunktion der t-Verteilung mit $k = 1$, $k = 3$ $k = 10$ Freiheitsgraden. Außerdem ist noch die Dichtefunktion der Standardnormalverteilung eingezeichnet.

Abbildung 10.11: Dichtefunktion der t-Verteillung mit $k = 1$, $k = 3$ und $k = 10$ Freiheitsgraden



Wir sehen, dass die Dichtefunktion der t-Verteilung mit wachsender Zahl von Freiheitsgraden der Dichtefunktion der Standardnormalverteilung immer ähnlicher wird. Die t-Verteilung mit kleiner Anzahl von Freiheitsgraden streut mehr als die Standardnormalverteilung. Dies erkennt man auch an der Varianz der t-Verteilung. Für $k \geq 3$ gilt

$$Var(T) = \frac{n}{n-2}$$

Die Varianz von T konvergiert gegen die Varianz der Standardnormalverteilung.

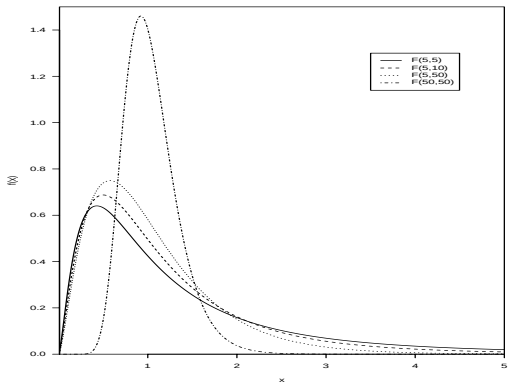
Die F-Verteilung

Bei einer Vielzahl von statistischen Verfahren werden zwei Varianzen verglichen. In diesem Fall kommt die F-Verteilung ins Spiel. Ausgangspunkt sind hier die unabhängigen Zufallsvariablen V und W, wobei V chiquadratverteilt mit m und W chiquadratverteilt mit n Freiheitsgraden ist. In diesem Fall ist die Zufallsvariable

$$F = \frac{V/m}{W/n}$$

F-verteilt mit m und n Freiheitsgraden.
Abbildung 10.12 zeigt die Dichtefunktion der t-Verteilung mit $m = 5, n = 5$, $m = 5, n = 10$, $m = 5, n = 50$ und $m = 50, n = 50$ Freiheitsgraden.

Abbildung 10.12: Dichtefunktion der F-Verteillung mit $m = 5, n = 5$, $m = 5, n = 10$, $m = 5, n = 50$ und $m = 50, n = 50$ Freiheitsgraden



10.3 Spezielle Verteilungen in R

In `R` gibt es Funktionen, mit denen man die Dichte- bzw. Wahrscheinlichkeitsfunktion, die Verteilungsfunktion und die Quantile von Verteilungen bestimmen kann. Jede Verteilung hat einen eigenen Namen. Die Normalverteilung wird mit dem Kürzel `norm` bezeichnet. Durch einen Buchstaben vor dem Namen erhält man nun die oben angesprochenen Funktionen. Ein `d` liefert die Dichte- bzw. Wahrscheinlichkeitsfunktion, ein `p` die Verteilungsfunktion und ein `q` die Quantile. Die Funktionen, die die Dichte- bzw. Wahrscheinlichkeitsfunktion oder die Verteilungsfunktion bestimmen, werden mit der oder den Stellen aufgerufen, an denen die jeweiligen Funktionen berechnet werden sollen. Die Funktion, die die Quantile bestimmt, wird mit dem oder den Werten von p aufgerufen, für die die Quantile berechnet werden sollen. Außerdem besitzt jede Funktion noch Argumente für die Parameter. Bei der Normalverteilung sind dies der Erwartungswert, den man mit dem Argument `mean` und die Standardabweichung, die man mit dem Argument `sd` übergibt. Will man also die Verteilungsfunktion der Standardnormalverteilung an der Stelle 2 bestimmen, so gibt man

```
> pnorm(2,mean=0,sd=1)
```

ein und erhält das Ergebnis

```
[1] 0.9772499 .
```

Wir hätten aber auch nur

```
> pnorm(2)
```

eingeben können, da die Parameter standardmäßig auf 0 und 1 gesetzt sind. Wir können die Verteilungsfunktion auch an mehreren Stellen auswerten:

```
> pnorm(c(-2,0,2))
[1] 0.02275013 0.50000000 0.97724987
```

Den $z_{0.95}$ der Standardnormalverteilung erhalten wir durch

```
> qnorm(0.95)
[1] 1.644854
```

Schauen wir uns die uns bekannten Verteilungen an.

Tabelle 10.7: Verteilungen in R

Verteilung	Name in R	Parameter in R	Parameter
Binomial	`binom`	`size`	n
		`prob`	p
Hypergeometrisch	`hyper`	`m`	M
		`n`	$N - M$
		`k`	n
Poisson	`pois`	`lambda`	λ
Normal	`normal`	`mean`	μ
		`sd`	σ
Gleich	`unif`	`min`	a
		`max`	b
Exponential	`exp`	`rate`	λ
t	`t`	`df`	k
Chiquadrat	`chisq`	`df`	k
F	`f`	`df1`	m
		`df2`	n

Oft will man die Dichtefunktionen oder Verteilungsfunktionen spezieller Verteilungen zeichnen. Bei stetigen Verteilungen sollte man hier auf die auf auf Seite 46 beschriebene Funktion `curve` zurückgreifen. Die Grafik der Dichtefunktion der Standardnormalverteilung erhält man durch

```
> curve(dnorm,from=-3,to=3)
```

und die Verteilungsfunktion der Standardnormalverteilung durch

```
> curve(pnorm,from=-3,to=3)
```

Bi diskreten Verteilungen hilft die in Kapitel 2.5 beschriebene Funktion `plot`. Die Grafik der Wahrscheinlichkeitsfunktion der Binomialverteilung mit den Parametern $n = 5$ und $p = 0.4$ gewinnt man durch

```
> plot(0:5,dbinom(0:5,5,0.4),type="h")
```

und die Verteilungsfunktion der Binomialverteilung mit den Parametern $n = 5$ und $p = 0.4$ durch

```
> plot(0:5,pbinom(0:5,5,0.4),type="S")
```

Mit R ist es aber auch möglich, Zufallszahlen aus den Verteilungen zu erzeugen. Um verstehen zu können, was dies bedeutet, holen wir ein wenig aus. Wir haben bisher ja noch keine Funktion für die diskrete Gleichverteilung kennengelernt. Dieser liegt folgendes Urnenmodell zugrunde. Eine Urne enthält k Kugeln, die von 1 bis k durchnummeriert sind. Wird nun eine dieser Kugeln zufällig ausgewählt, so besitzt die Augenzahl X eine diskrete Gleichverteilung. In R besteht die Möglichkeit, das Ziehen aus der Urne am Computer mit der Funktion `sample` durchzuführen. Die Funktion `sample` besitzt 4 Argumente, von denen zwei fakultativ sind. Beginnen wir zunächt mit den obligatorischen Argumenten. Das erste Argument `x` enthält die Gesamtheit, aus der gezogen werden soll. Man kann für `x` auch eine natürliche Zahl N eingeben. In diesem Fall wird aus den natürlichen Zahlen $1, \ldots, N$ gezogen. Das zweite Argument `size` gibt den Stichprobenumfang an. Der Aufruf

```
> sample(6,size=1)
[1] 6
```

liefert das Ergebnis des Ziehens einer Kugel aus einer Urne, die 6 Kugeln enthält, die von 1 bis 6 durchnummeriert sind. Dies entspricht dem Wurf eines fairen Würfels. Würfeln wir noch einmal, wobei das Argument `size` an seiner Position aufgerufen wird.

```
> sample(6,1)
[1] 3
```

Der Zufallszahlengenerator ist so konstruiert, dass die Zufallszahlen unabhängig sind. Wir können auch mehrere Zufallszahlen erzeugen. Wir ziehen mit Zurücklegen, wenn das Argument `replace` der Funktion `sample` den Wert `TRUE` annimmt. Standardmäßig steht es auf `FALSE`. Es wird also ohne Zurücklegen gezogen.
Um zweimal zu würfeln, geben wir also ein

```
> sample(6,2,replace=TRUE)
[1] 2 5
```

Lottozahlen zieht man durch

```
> sample(49,6)
[1] 32 25 16 19 46  7
```

Man kann den Startwert des Zufallszahlengenerators mit der Funktion `set.seed` setzen. Hierdurch ist es möglich, identische Folgen von Zufallszahlen zu erzeugen.

```
> set.seed(2003)
> sample(49,6)
[1] 46 10 11  2 18 41
> set.seed(2003)
> sample(49,6)
[1] 46 10 11  2 18 41
```

Nun wollen wir aus der Urne aus Beispiel 88 auf Seite 274 ziehen. Diese enthält 4 Kugeln, die 10 g wiegen, und 6 Kugeln, die 20 g wiegen. Wir ziehen also mit Wahrscheinlichkeit 0.4 eine 10 g schwere Kugel und mit Wahrscheinlichkeit 0.6 eine 20 g schwere Kugel. Diese Wahrscheinlichkeiten können wir der Funktion `sample` im Argument `prob`, das an vierter Stelle steht, übergeben. Um zwei Kugeln mit Zurücklegen aus der Urne zu ziehen, geben wir also ein

```
> sample(x=c(10,20),size=2,replace=TRUE,prob=c(0.4,0.6))
[1] 10 20
```

Wir können natürlich auch eingeben

```
> sample(c(10,20),2,TRUE,c(0.4,0.6))
[1] 10 10
```

Wir erhalten im zweiten Fall andere Zufallszahlen, da sich der Startwert des Zufallszahlengenerators nach der ersten Ziehung geändert hat.

Mit dieser Vorgehensweise können wir aus jeder diskreten Verteilung Zufallszahlen ziehen. Bei den Standardverteilungen müssen wir aber nicht `sample` benutzen. Setzen wir ein `r` vor den Namen der Funktion und geben als erstes Argument `n` die Anzahl der Zufallszahlen an, so erhalten wir n unabhängige Zufallszahlen. Schauen wir uns ein Beispiel an.

Beispiel 105
Eine Urne enthält 10 Kugeln, von denen 4 weiß und 6 schwarz sind. Es werden 3 Kugeln gezogen. Von Interesse ist die Anzahl X der weißen Kugeln. Beim Ziehen mit Zurücklegen ist X binomialverteilt mit den Parametern $n = 3$ und $p = 0.4$. Beim Ziehen ohne Zurücklegen ist X hypergeometrisch verteilt mit $N = 10$, $M = 4$ und $n = 3$.

Eine Zufallszahl beim Ziehen mit Zurücklegen erhalten wir durch

```
> rbinom(n=1,size=2,prob=0.4)
[1] 1
```

und eine Zufallszahl beim Ziehen ohne Zurücklegen durch

```
> rhyper(nn=1,m=4,n=6,k=2)
[1] 1
```

□

Bisher haben wir nur diskrete Zufallsvariablen betrachtet. Um Zufallszahlen aus den bekannten stetigen Verteilungen zu ziehen, schreibt man vor den Namen der Funktion den Buchstaben `r`. Die Argumente sind die Anzahl `n` der Zufallszahlen und die Parameter der jeweiligen Verteilung. Die Zufallszahlen sind unabhängig. Fünf Zufallszahlen aus der Standardnormalverteilung erhält man durch

```
> rnorm(5)
[1] -0.2080638 -0.4892996  0.5359943 -0.6403278 -1.7474349
```

Wir werden in den nächsten Kapiteln im Rahmen von Simulationsstudien immer wieder Gebrauch von Zufallszahlen machen, um die Eigenschaften von sogenannten Stichprobenfunktionen zu untersuchen.

Kapitel 11

Stichproben

In der deskriptiven Statistik werden die Charakteristika eines Datensatzes durch Grafiken verdeutlicht und durch Maßzahlen zusammengefasst. In der Regel ist man aber nicht nur an der Verteilung des Merkmals im Datensatz interessiert, sondern man will auf Basis der Daten eine Aussage über die Verteilung des Merkmals in der Grundgesamtheit machen, aus der die Daten gezogen wurden. Man nennt die Teilgesamtheit auch eine **Stichprobe**. So könnte die Durchschnittsnote aller Studenten der Wirtschaftswissenschaften im Abitur und der Anteil der Studenten, die den Mathematik Leistungskurs besucht haben, von Interesse sein. Die Durchschnittsnote ist der Erwartungswert μ. Einen Anteil in einer Grundgesamtheit bezeichnen wir im folgenden mit p. Allgemein bezeichnen wir eine Größe einer Verteilung, an der wir interessiert sind, als Parameter θ. Will man einen oder mehrere Werte für den Parameter angeben, so spricht man vom **Schätzen**. Hierbei unterscheidet man **Punktschätzung** und **Intervallschätzung**. Bei der Punktschätzung bestimmt man aus den Daten einen Wert für den unbekannten Parameter, während man bei der Intervallschätzung ein Intervall angibt. Soll eine Vermutung über den Parameter überprüft werden, so spricht man vom **Testen**. Um verstehen zu können, warum und wann man auf Basis einer Stichprobe Aussagen über eine Grundgesamtheit machen kann, muss man sich Gedanken über Stichproben machen.

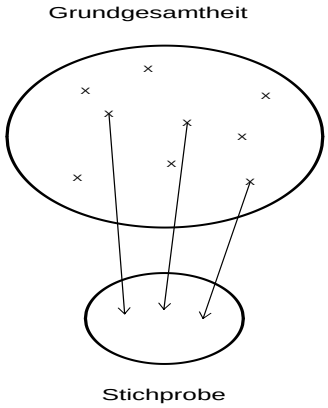
Ausgangspunkt der schließenden Statistik ist eine **Grundgesamtheit**. Dies ist die Menge aller Personen bzw. Objekte, bei denen das oder die interessierenden Merkmale erhoben werden können. So ist die Menge aller Studenten der Wirtschaftswissenschaften in Deutschland eine Grundgesamtheit. Hier könnten der Frauenanteil, die erwartete Dauer des Studiums oder die Durschnittsnote im Diplom von Interesse sein.

Beispiel 106
Eine Grundgesamtheit besteht aus 4 Personen. Die Körpergröße der Personen beträgt

```
168 172 176 180
```

Die durchschnittliche Körpergröße $E(X) = \mu$ aller 4 Personen beträgt 174 und die Varianz $Var(X) = \sigma^2$ der Körpergröße ist 20. □

Es liegt nahe, bei allen Personen bzw. Objekten der Grundgesamtheit die interessienden Merkmale zu erheben. Man spricht in diesem Fall von einer **Vollerhebung**. Ein Beispiel hierfür ist die Volkszählung. Bei dieser werden in regelmäßigen Abständen eine Reihe von Merkmalen von allen Bürgern der Bundesrepublik Deutschland erfragt. In der Regel ist eine Vollerhebung aber zu teuer oder zu aufwendig. Oft ist es auch gar nicht möglich, die Grundgesamtheit vollständig zu untersuchen. Dies ist der Fall, wenn die Untersuchung eines Objekts zu dessen Zerstörung führt. Kennt man die Lebensdauer einer Glühbirne oder eines Autoreifens, so kann man sie nicht mehr gebrauchen. Man spricht von **zerstörender Prüfung**. Da Vollerhebungen eine Reihe Nachteile besitzen, wird man nur einen Teil der Grundgesamtheit, eine sogenannte **Teilgesamtheit** untersuchen. Will man von der Teilgesamtheit sinnvoll auf die Grundgesamtheit schließen, so muss die Teilgesamtheit repräsentativ für die Grundgesamtheit sein. Dies ist unter anderem dann der Fall, wenn jedes Element der Grundgesamtheit die gleiche Chance hat, in die Teilgesamtheit zu gelangen. Man spricht dann von einer **Zufallsstichprobe**. Die folgende Abbildung visualisiert den Ziehungsprozess.



Bezeichnen wir mit x_i den Wert des interessierenden Merkmals beim i-ten Objekt der Teilgesamtheit, so ist $x_1, \ldots, x_n$ die Stichprobe.

Beispiel 106 (fortgesetzt von Seite 328)
Nehmen wir an, dass die Grundgesamtheit nicht vollständig beobachtet wird. Es können nur zwei Personen beobachtet werden. Man zieht also eine Stichprobe (x_1, x_2) vom Umfang $n = 2$. Dabei ist x_1 die Größe der ersten gezogenen Person und x_2 die Größe der zweiten gezogenen Person. Beim Ziehen ohne Zurücklegen gibt es $4 \cdot 3 = 12$ mögliche Sichproben. Sie sind

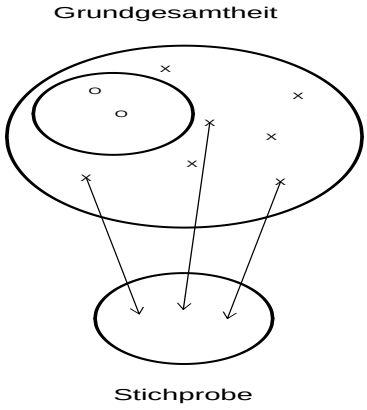
$$\begin{array}{lll} (168,172) & (168,176) & (168,180) \\ (172,168) & (172,176) & (172,180) \\ (176,168) & (176,172) & (176,180) \\ (180,168) & (180,172) & (180,176) \end{array}$$

Beim Ziehen mit Zurücklegen gibt es $4^2 = 16$ mögliche Sichproben. Sie sind

$$\begin{array}{llll} (168,168) & (168,172) & (168,176) & (168,180) \\ (172,168) & (172,172) & (172,176) & (172,180) \\ (176,168) & (176,172) & (176,176) & (176,180) \\ (180,168) & (180,172) & (180,176) & (180,180) \end{array}$$

□

Bei einer Zufallsstichprobe hat jedes Element der Grundgesamtheit die gleiche Chance, in die Stichprobe zu gelangen. In der folgenden Abbildung werden die Objekte in der Ellipse in der Grundgesamtheit bei der Ziehung nicht berücksichtigt. Man spricht von einer **verzerrten Stichprobe**.



Schauen wir uns am Beispiel an, was passiert, wenn bestimmte Elemente der Grundgesamtheit nicht in die Stichprobe gelangen können.

Beispiel 106 (fortgesetzt von Seite 329)
Die ersten beiden Personen sind Frauen und die beiden anderen Männer. Es werden aber nur die Frauen in Betracht gezogen. Ziehen wir mit Zurücklegen, dann gibt es folgende Stichproben

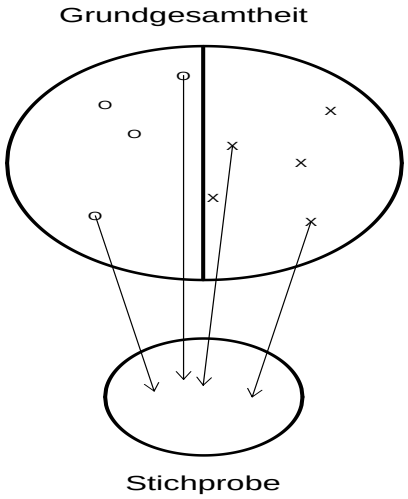
$$(168, 168) \quad (168, 172) \quad (172, 168) \quad (172, 172)$$

Diese liefern alle ein verzerrtes Bild der Grundgesamtheit, da wir die Körpergröße in der Grundgesamtheit auf Basis der Stichprobe zu klein einschätzen.
□

Wie das Beispiel zeigt, liefert eine Stichprobe ein verzerrtes Bild der Grundgesamtheit, wenn Elemente der Grundgesamtheit nicht in die Stichprobe gelangen können. Dies kann dadurch geschehen, dass sie bei der Ziehung der Stichprobe nicht berücksichtigt wurden. Man spricht in diesem Fall vom **Selektions-Bias**. Dieser liegt zum Beispiel bei Befragungen im Internet vor. Hier ist die Stichprobe sicherlich nicht repräsentativ für die Bevölkerung, da nur ein Teil der Bevölkerung Zugang zum Internet besitzt. Eine verzerrte Stichprobe erhält man aber auch dann, wenn Befragte eine Frage nicht beantworten und dieses Verhalten von der gestellten Frage abhängt. Man spricht in diesem Fall vom **Nonresponse-Bias**. Dieser tritt zum Beispiel bei Fragen nach dem Einkommen auf. Hier werden Personen mit sehr niedrigem oder sehr hohem Einkommen aus naheliegenden Gründen diese Frage nicht beantworten. Bei Befragungen auf freiwilliger Basis antworten oft nur die, die bei der Frage besonders involviert sind. Man spricht hier vom **Selfselection-Bias**.

Beispiel 107
Der Literary Digest hatte seit 1916 den Gewinner der Präsidentschaftswahlen in den USA immer richtig prognostiziert. Im Jahr 1936 trat der Kandidat der Republikaner Landon gegen den demokratischen Präsidenten Roosevelt an. Eine Befragung von 2,4 Milionen Amerikanern durch den Literary Digest ergab, dass aber 57 Prozent Landon wählen würden. Bei der Wahl wurde Roosevelt von 62 Prozent der Wähler gewählt. Woran lag die schlechte Prognose des Literary Digest? Der Literary Digest hatte Fragebögen an 10 Millionen Haushalte verschickt. Von diesen haben aber nur 24 Prozent geantwortet. Dies spricht für einen Nonresponse-Bias.
□

Besitzt man keine Informationen über eine Grundgesamtheit, so sollte man eine Zufallsstichprobe ziehen. Liegen jedoch Informationen über die Grundgesamtheit vor, so sollten diese bei der Stichprobenziehung berücksichtigt werden. Ein Beispiel hierfür sind **geschichtete Stichproben**. Bei diesen sind bei jedem Merkmalsträger die Ausprägungen eines oder mehrerer Merkmale bekannt. Auf der Basis dieser Merkmale teilt man die Grundgesamtheit in disjunkte Klassen ein, die man auch **Schichten** nennt. Man zieht aus jeder der Schichten eine Zufallsstichprobe. Die folgende Abbildung visualisiert die Schichtenbildung und den Ziehungsvorgang. Dabei bilden die Kreise die eine und die Kreuze die andere Schicht.



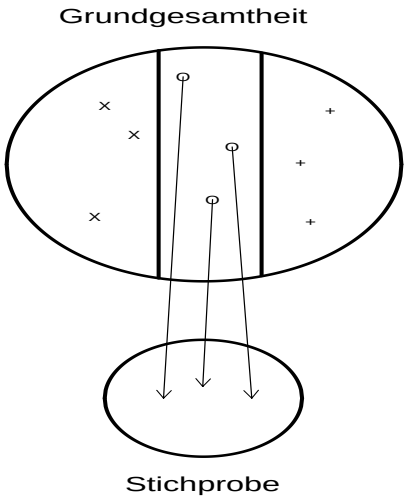
Beispiel 106 (fortgesetzt von Seite 330)
] Die ersten beiden Personen seien Frauen, die beiden anderen Männer. Die erste Schicht besteht aus den Frauen und die zweite aus den Männern. Aus jeder der beiden Schichten wird eine Stichprobe vom Umfang $n = 1$ gezogen. Es gibt also folgende Stichproben:

$$(168, 176) \quad (168, 180) \quad (172, 176) \quad (172, 180)$$

□

Oft werden Personen oder Objekte zu einer Einheit zusammengefasst. So ist es bei mündlichen Befragungen aus Kostengründen sinnvoll, Personen zu

befragen, die nahe beieinander wohnen. Eine Auswahleinheit ist dann nicht die Person, sondern die Gruppe. Man spricht auch von **Klumpen** und dem **Klumpenverfahren**. Die folgende Abbildung illustriert das Klumpenverfahren. Dabei gibt es drei Klumpen. Die Objekte des ersten Klumpen sind durch ein 'X', die des zweiten durch ein 'O' und die des dritten durch ein '+' veranschaulicht.



Beispiel 106 (fortgesetzt von Seite 331)
] Wir fassen die ersten beiden Personen zu einem Klumpen und die beiden anderen Personen zum anderen Klumpen zusammen. Es wird ein Klumpen als Stichprobe ausgewählt. Es gibt also folgende Stichproben:

$$(168, 172) \quad (176, 180)$$

□

Beispiel 108
Der Mikrozensus ist eine Erhebung, bei der jedes Jahr 1 Prozent der Bevölkerung der Bundesrepublik Deutschland befragt wird. Im Mikrozensusgesetz wird die Aufgabe des Mikrozensus beschrieben. Es sollen 'statistische Angaben in tiefer fachlicher Gliederung über die Bevölkerungsstruktur, die wirtschaftliche und soziale Lage der Bevölkerung, der Familien und der Haushalte, den Arbeitsmarkt, die berufliche Gliederung und die Ausbildung der

Erwerbsbevölkerung sowie die Wohnverhältnisse' bereitgestellt werden. Beim Mikrozensus wird das Schichtungsverfahren und das Klumpenverfahren verwendet. Die Klumpen bestehen dabei aus Haushalten, wobei ein Klumpen aus höchstens 9 Wohnungen besteht. Außerdem wird nach mehreren Variablen geschichtet. Es werden zunächst regionale Schichten gebildet, die im Mittel 350000 Einwohner enthalten. So bilden Kreise, Zusammenfassungen von Kreisen oder Großstädte regionale Schichten. Sehr große Städte werden in regionale Schichten geteilt. Als weiteres Schichtungsmerkmal wird die Gebäudegröße betrachtet. Hierbei werden 4 Schichten gebildet. Schicht 1 enthält alle Gebäude, die mindestens eine, aber höchstens vier Wohungen enthalten, Schicht 2 enthält alle Gebäude mit 5 bis 10 Wohnungen, Schicht 3 alle Gebäude mit mindestens 11 Wohnungen und Schicht 4 alle Gebäude, in denen eine Anstalt vermutet wird. Aus jeder Schicht wird ein Prozent der Personen ausgewählt. Die Stadt Bielefeld mit ihren ungefähr 300000 Einwohnern bildet eine eigene regionale Schicht, in der es die vier Schichten der Gebäudegrößen gibt. Aus jeder dieser Schichten wird eine Stichprobe gezogen. □

Kapitel 12

Stichprobenfunktionen

Um eine Aussage über den Wert eines unbekannten Parameters θ zu machen, zieht man eine Zufallsstichprobe vom Umfang n aus der Grundgesamtheit. Das Merkmal wird in diesem Fall an n Personen bzw. Objekten beobachtet. Wir werden im Folgenden von Personen sprechen. Wir erhalten also einen Wert x_1 für die erste Person, einen Wert x_2 für die zweite Person usw. Vor der Ziehung ist der Wert x_i unbekannt und kann somit als Realisation einer Zufallsvariablen X_i aufgefasst werden.

Beispiel 106 (fortgesetzt von Seite 332)
Wir betrachten weiterhin die Grundgesamtheit der vier Personen, deren Körpergröße `168,172,176,180` beträgt. Aus der Grundgesamtheit werden zwei Personen ausgewählt. Sei X_i die Körpergröße der i-ten Person, $i = 1, 2$.
Tabelle 12.1 zeigt die gemeinsame Wahrscheinlichkeitsfunktion und die Randverteilungen von X_1 und X_2 beim Ziehen ohne Zurücklegen.

Tabelle 12.1: Gemeinsame Wahrscheinlichkeitsfunktion und Randverteilungen von X_1 und X_2

x_1 \ x_2	168	172	176	180	
168	0	1/12	1/12	1/12	1/4
172	1/12	0	1/12	1/12	1/4
176	1/12	1/12	0	1/12	1/4
180	1/12	1/12	1/12	0	1/4
	1/4	1/4	1/4	1/4	

Tabelle 12.2 zeigt die gemeinsame Wahrscheinlichkeitsfunktion und die Randverteilungen von X_1 und X_2 beim Ziehen mit Zurücklegen.

Tabelle 12.2: Gemeinsame Wahrscheinlichkeitsfunktion und Randverteilungen von X_1 und X_2

x_1 \ x_2	168	172	176	180	
168	1/16	1/16	1/16	1/16	1/4
172	1/16	1/16	1/16	1/16	1/4
176	1/16	1/16	1/16	1/16	1/4
180	1/16	1/16	1/16	1/16	1/4
	1/4	1/4	1/4	1/4	

□

Im Beispiel sind sowohl beim Ziehen mit Zurücklegen als auch beim Ziehen ohne Zurücklegen die Verteilungen von X_1 und X_2 mit der Verteilung des Merkmals in der Grundgesamtheit identisch. Dies gilt auch allgemein. Jede Zufallsvariable X_i besitzt also die Verteilung der Grundgesamtheit. Die X_i sind also identisch verteilt. Zieht man mit Zurücklegen, so sind $X_1, \ldots, X_n$ auch unabhängig. Dies ist beim Ziehen ohne Zurücklegen nicht der Fall.

Wir suchen in der Regel einen Wert für einen Parameter θ. Da uns n Beobachtungen $x_1, \ldots, x_n$ für einen Wert zur Verfügung stehen, fassen wir diese zu einem Wert zusammen. Wir bilden also eine Funktion $g(x_1, \ldots, x_n)$ der Beobachtungen. Die Werte von $g(x_1, \ldots, x_n)$ hängen von den Realisationen von $X_1, \ldots, X_n$ ab. Somit ist $g(x_1, \ldots, x_n)$ die Realisation einer Zufallsvariablen, die wir mit $g(X_1, \ldots, X_n)$ bezeichnen wollen. Man spricht auch von einer **Stichprobenfunktion**. Die wichtigste Stichprobenfunktion ist $\bar{X}$.

Beispiel 106 (fortgesetzt von Seite 335)
Tabelle 12.3 gibt alle möglichen Stichproben (x_1, x_2) mit den zugehörigen Werten $g(x_1, x_2) = \bar{x}$ von $g(X_1, X_2) = \bar{X}$ an, wenn wir mit Zurücklegen ziehen.

Tabelle 12.3: Stichproben mit zugehörigen Wert der Stichprobenfunktion $\bar{X}$

(x_1, x_2)	$\bar{x}$	(x_1, x_2)	$\bar{x}$	(x_1, x_2)	$\bar{x}$	(x_1, x_2)	$\bar{x}$
$(168, 168)$	168	$(168, 172)$	170	$(168, 176)$	172	$(168, 180)$	174
$(172, 168)$	170	$(172, 172)$	172	$(172, 176)$	174	$(172, 180)$	176
$(176, 168)$	172	$(176, 172)$	174	$(176, 176)$	176	$(176, 180)$	178
$(180, 168)$	174	$(180, 172)$	176	$(180, 176)$	178	$(180, 180)$	180

Da jede der Stichproben gleichwahrscheinlich ist, erhalten wir folgende Wahrscheinlichkeitsfunktion von $\bar{X}$, die in Tabelle 12.4 zu finden ist.

Tabelle 12.4: Wahrscheinlichkeitsfunktion von $\bar{X}$

$\bar{x}$	168	170	172	174	176	178	180
$P(\bar{X} = \bar{x})$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

Schauen wir uns den Erwartungswert und die Varianz von $\bar{X}$ an. Es gilt

$$\begin{aligned} E(\bar{X}) &= 168 \cdot \frac{1}{16} + 170 \cdot \frac{2}{16} + 172 \cdot \frac{3}{16} + 174 \cdot \frac{4}{16} \\ &+ 176 \cdot \frac{3}{16} + 178 \cdot \frac{2}{16} + 180 \cdot \frac{1}{16} = 174\,. \end{aligned}$$

Wir sehen, dass der Erwartungswert der Grundgesamtheit das Zentrum der Verteilung von $\bar{X}$ bildet. Es gilt

$$\begin{aligned} E(\bar{X}^2) &= 168^2 \cdot \frac{1}{16} + 170^2 \cdot \frac{2}{16} + 172^2 \cdot \frac{3}{16} + 174^2 \cdot \frac{4}{16} \\ &+ 176^2 \cdot \frac{3}{16} + 178^2 \cdot \frac{2}{16} + 180^2 \cdot \frac{1}{16} \\ &= 30286\,. \end{aligned}$$

Also gilt

$$Var(\bar{X}) = E(\bar{X}^2) - E(\bar{X})^2 = 30286 - 174^2 = 10$$

Tabelle 12.5 gibt alle möglichen Stichproben (x_1, x_2) mit den zugehörigen Werten $g(x_1, x_2) = \bar{x}$ von $g(X_1, X_2) = \bar{X}$ an, wenn wir ohne Zurücklegen ziehen.

Tabelle 12.5: Stichproben mit zugehörigen Wert der Stichprobenfunktion $\bar{X}$

(x_1, x_2)	$\bar{x}$	(x_1, x_2)	$\bar{x}$	(x_1, x_2)	$\bar{x}$
$(168, 172)$	170	$(168, 176)$	172	$(168, 180)$	174
$(172, 168)$	170	$(172, 176)$	174	$(172, 180)$	176
$(176, 168)$	172	$(176, 172)$	174	$(176, 180)$	178
$(180, 168)$	174	$(180, 172)$	176	$(180, 176)$	178

Da jede der Stichproben gleichwahrscheinlich ist, erhalten wir folgende Wahrscheinlichkeitsfunktion von $\bar{X}$, die in Tabelle 12.6 zu finden ist.

Tabelle 12.6: Wahrscheinlichkeitsfunktion von $\bar{X}$

$\bar{x}$	170	172	174	176	178
$P(\bar{X} = \bar{x})$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Schauen wir uns den Erwartungswert und die Varianz von $\bar{X}$ an. Es gilt

$$E(\bar{X}) = 170 \cdot \frac{1}{6} + 172 \cdot \frac{1}{6} + 174 \cdot \frac{2}{6} + 176 \cdot \frac{1}{6} + 178 \cdot \frac{1}{6} = 174\,.$$

Wir sehen, dass der Erwartungswert der Grundgesamtheit das Zentrum der Verteilung von $\bar{X}$ bildet. Es gilt

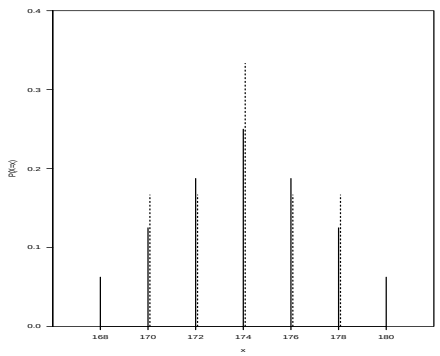
$$E(\bar{X}^2) = 170^2 \cdot \frac{1}{6} + 172^2 \cdot \frac{1}{6} + 174^2 \cdot \frac{2}{6} + 176^2 \cdot \frac{1}{6} + 178^2 \cdot \frac{1}{6} = 30282.67\,.$$

Also gilt

$$Var(\bar{X}) = E(\bar{X}^2) - E(\bar{X})^2 = 30282.67 - 174^2 = 6.67$$

Abbildung 12.1 zeigt die Wahrscheinlichkeitsfunktion von $\bar{X}$ für das Ziehen mit Zurücklegen (durchgezogene Linie) und das Ziehen ohne Zurücklegen (gestrichelte Linie).

Abbildung 12.1: Wahrscheinlichkeitsfunktion von $\bar{X}$



□

An diesem Beispiel kann man gut erkennen, warum $\bar{X}$ für $n = 2$ viel weniger streut als das Merkmal in der Grundgesamtheit. Schauen wir uns zunächst das Ziehen mit Zurücklegen an. Die extrem kleinen oder großen Werte von $\bar{X}$ treten in der Stichprobe viel seltener auf als in der Grundgesamtheit. So besitzt der Wert 168 in der Grundgesamtheit die Wahrscheinlichkeit 1/4, während $\bar{X}$ mit Wahrscheinlichkeit 1/16 diesen Wert annimmt. Außerdem liefert eine Stichprobe mit einem extrem kleinen und einem extrem großen Wert einen Mittelwert in der Nähe des Erwartungswerts. So nimmt $\bar{X}$ für die Stichprobe $(168, 180)$ den Wert 174 an, der gleich dem Erwartungswert ist. Beim Ziehen ohne Zurücklegen ist die Streuung von $\bar{X}$ noch kleiner. Dies liegt unter anderem daran, dass die extremen Werte wie 168 und 180 gar nicht mehr auftreten, da im Gegensatz zum Ziehen mit Zurücklegen die Stichproben $(168, 168)$ und $(180, 180)$ nicht beobachtet werden können. Dies zeigt auch die Abbildung 12.1. Die Varianz von $\bar{X}$ beim Ziehen ohne Zurücklegen beträgt

$$\boxed{Var(\bar{X}) = \frac{\sigma^2}{n} \frac{N-n}{N-1}}. \tag{12.1}$$

dabei ist n der Stichprobenumfang, N der Umfang der Grundgesamtheit und σ^2 die Varianz des Merkmals in der Grundgesamtheit. Da N in der Regel unbekannt ist, kann man die Varianz von $\bar{X}$ beim Ziehen ohne Zurücklegen nicht angeben. Deshalb werden wir im Folgenden immer davon ausgehen, dass wir mit Zurücklegen ziehen. Wie wir noch sehen werden, hängt in diesem Fall die Varianz von $\bar{X}$ nicht von N ab.

Bevor wir uns die Eigenschaften von $\bar{X}$ im nächsten Kapitel genauer anschauen, wenden wir uns noch einmal dem Schichtungsverfahren zu.

Beispiel 106 (fortgesetzt von Seite 336)
Wir gehen wieder davon aus, dass die ersten beiden Personen die erste Schicht und die beiden anderen die zweite Schicht bilden. Tabelle 12.7 zeigt alle Stichproben und den Wert von $\bar{X}$, wenn wir aus jeder Schicht eine Beobachtung ziehen.

Tabelle 12.7: Stichproben mit zugehörigen Wert der Stichprobenfunktion $\bar{X}$ beim Schichtungsverfahren

(x_1, x_2)	$(168, 176)$	$(168, 180)$	$(172, 176)$	$(172, 180)$
$\bar{x}$	172	174	174	176

Tabelle 12.8 zeigt die Verteilung von $\bar{X}$.

Tabelle 12.8: Verteilung von $\bar{X}$ beim Schichtungsverfahren

$\bar{x}$	172	174	176
$P(\bar{X} = \bar{x})$	0.25	0.5	0.25

Es gilt

$$E(\bar{X}) = 172 \cdot 0.25 + 174 \cdot 0.5 + 176 \cdot 0.25 = 174$$

und

$$E(\bar{X}^2) = 172^2 \cdot 0.25 + 174^2 \cdot 0.5 + 176^2 \cdot 0.25 = 30278$$

Also gilt

$$Var(\bar{X}) = E(\bar{X}^2) - E(\bar{X})^2 = 30278 - 174^2 = 2$$

□

Wir sehen, dass im Beispiel $E(\bar{X})$ bei einer Zufallsstichprobe mit Zurücklegen vom Umfang $n = 2$ und bei einer geschichteten Stichprobe mit dem Erwartungswert der Grundgesamtheit zusammenfällt. Die Varianz von $\bar{X}$ ist jedoch bei der geschichteten Stichprobe viel kleiner als die bei der Zufallststichprobe mit Zurücklegen. Man spricht vom **Schichtungseffekt**. Dieser ist gerade dann besonders groß, wenn das Merkmal in den Schichten eine geringe

Streuung besitzt und das Niveau des Merkmals in den einzelnen Schichten unterschiedlich ist. Im Beispiel ist dies gegeben. Die erste Schicht enthält die kleinen Personen und die zweite die großen. Die Streuungen in den beiden Schichten sind klein. Aus jeder der beiden Schichten wird genau eine Person gezogen. Somit enthält die Stichprobe einen kleinen und einen großen Wert. Der Mittelwert der beiden Beobachtungen liegt in der Nähe des Mittelwerts der Grundgesamtheit oder ist sogar mit diesem identisch. Die Zufallsvariable $\bar{X}$ hat in diesem Fall also eine kleine Varianz.

Ist die Streuung in den Schichten hingegen groß, so ist der Schichtungseffekt gering.

Beispiel 106 (fortgesetzt von Seite 340)
Die Schichten seien nun so gewählt, dass die Personen der ersten Schicht 168 und 176 cm groß sind. Die Körpergröße der Personen der zweiten Schicht beträgt also 172 und 180 cm. Aus jeder Schicht wird jeweils eine Person zufällig ausgewählt.
Tabelle 12.9 zeigt alle Stichproben und den Wert von $\bar{X}$, wenn wir aus jeder Schicht eine Beobachtung ziehen.

Tabelle 12.9: Stichproben mit zugehörigen Wert der Stichprobenfunktion $\bar{X}$ beim Schichtungsverfahren

(x_1, x_2)	$(168, 172)$	$(168, 180)$	$(176, 172)$	$(176, 180)$
$\bar{x}$	170	174	174	178

Tabelle 12.10 zeigt die Verteilung von $\bar{X}$.

Tabelle 12.10: Verteilung von $\bar{X}$ beim Schichtungsverfahren

$\bar{x}$	170	174	178
$P(\bar{X} = \bar{x})$	0.25	0.5	0.25

Es gilt

$$E(\bar{X}) = 170 \cdot 0.25 + 174 \cdot 0.5 + 178 \cdot 0.25 = 174$$

und

$$E(\bar{X}^2) = 170^2 \cdot 0.25 + 174^2 \cdot 0.5 + 178^2 \cdot 0.25 = 30284$$

Also gilt

$$Var(\bar{X}) = E(\bar{X}^2) - E(\bar{X})^2 = 30284 - 174^2 = 8$$

□

Das Beispiel zeigt, dass man die Schichtungsvariable so wählen sollte, dass die Streuung des interessierenden Merkmals in den Schichten klein ist und der Lageparameter des interessierenden Merkmals in den Schichten unterschiedliche Werte annimmt. Schauen wir uns nun auch noch das Klumpenverfahren an.

Beispiel 106 (fortgesetzt von Seite 341)
Wir gehen zunächst davon aus, dass die ersten beiden Personen den ersten Klumpen und die beiden anderen den anderen Klumpen bilden. Die Beobachtungen im ersten Klumpen sind also 168 und 172, die Beobachtungen im zweiten Klumpen 176 und 180. Da die Klumpen die Auswahleinheiten bilden, gelangen alle Beobachtungen eines Klumpen in die Stichprobe. Wir wählen einen der beiden Klumpen zufällig aus. Beim ersten nimmt $\bar{X}$ den Wert 170 und beim zweiten den Wert 178 an. Da wir den Klumpen zufällig auswählen, beträgt die Wahrscheinlichkeit für jeden dieser Werte 0.5. Es gilt also

$$P(\bar{X} = 170) = 0.5 \qquad P(\bar{X} = 178) = 0.5\,.$$

Hieraus folgt

$$E(\bar{X}) = 170 \cdot 0.5 + 178 \cdot 0.5 = 174\,.$$

Auch beim Klumpenverfahren bildet der Erwartungswert der Grundgesamtheit das Zentrum der Verteilung von $\bar{X}$.
Weiterhin gilt

$$E(\overline{X^2}) = 170^2 \cdot 0.5 + 178^2 \cdot 0.5 = 30292\,.$$

Hieraus folgt

$$Var(\bar{X}) = E(\overline{X^2}) - E(\bar{X}) = 30292 - 174^2 = 16\,.$$

Wir sehen, dass die Varianz von $\bar{X}$ viel größer ist als beim Ziehen einer Zufallsstichprobe. Dies liegt daran, dass die Klumpen sehr homogen sind. Wenn wir nur einen Klumpen auswählen, gelangt auch nur ein kleiner Teil der Grundgesamtheit in die Stichprobe. Man spricht vom **Klumpeneffekt**. Im Gegensatz zum Schichtungseffekt führt der Klumpeneffekt zu einer Vergrößerung der Varianz. Man will ihn also vermeiden. In der Praxis ist dies aber selten möglich, da man die Klumpen aus pragmatischen Gründen wählt.

So sollen Interviewer Personen in benachbarten Häusern befragen. Diese werden sich in der Regel in vielen Merkmalen ähneln, sodass der Klumpeneffekt auftritt.
Wir haben das Schichtungsverfahren und Klumpenverfahren an einem Datenbeispiel betrachtet, um die Vorteile und Nachteile dieser Verfahren zu illustrieren. Weitere Details zu diesen Verfahren findet man im Buch *Wirtschafts- und Sozialstatistik : Gewinnung von Daten* von *Walter Krug, Martin Nourney und Jürgen Schmidt*. Wir wollen uns im Folgenden nicht weiter mit dem Schichtungsverfahren und dem Klumpenverfahren beschäftigen. □

12.1 Die Stichprobenfunktion $\bar{X}$

Wir gehen also im Folgenden davon aus, dass eine Zufallsstichprobe vorliegt. Das heißt, wir beobachten die Realisationen $x_1, \ldots, x_n$ der unabhängigen und identisch verteilten Zufallsvariablen $X_1, \ldots, X_n$. Speziell gelte $E(X_i) = \mu$ und $Var(X_i) = \sigma^2$ für $i = 1, 2, \ldots, n$. Uns interessiert die Verteilung von

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

12.1.1 Erwartungswert und Varianz von $\bar{X}$

Wir wollen in diesem Abschnitt den Erwartungswert und die Varianz von $\bar{X}$ herleiten. Hierbei benutzen wir zwei Eigenschaften einer Summe von Zufallsvariablen, die wir in Statistik I gezeigt haben.
Sind die Zufallsvariablen $X_1, \ldots, X_n$ identisch verteilt mit $E(X_i) = \mu$, dann gilt

$$E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} \mu = n\,\mu$$

Nun können wir den Erwartungswert von $\bar{X}$ bestimmen. Wir haben am Beispiel gesehen, dass der Erwartungswert von $\bar{X}$ mit dem Erwartungswert der Grundgesamtheit zusammenfiel. Dies gilt allgemein. Es gilt also folgender

Satz 12.1
Die Zufallsvariablen $X_1, \ldots, X_n$ seien identisch verteilt mit $E(X_i) = \mu$ für

$i = 1, 2, \ldots, n$. Dann gilt

$$\boxed{E(\bar{X}) = \mu} \tag{12.2}$$

Beweis:

$$E(\bar{X}) \;=\; E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\,E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\,n\,\mu = \mu$$

□

Der Satz zeigt, dass der Erwartungswert der Grundgesamtheit das Zentrum der Verteilung von $\bar{X}$ bildet. Wenden wir uns nun der Varianz von $\bar{X}$ zu. Sind die Zufallsvariablen $X_1, \ldots, X_n$ unabhängig und identisch verteilt mit $Var(X_i) = \sigma^2$, dann gilt

$$Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i) = \sum_{i=1}^{n} \sigma^2 = n\,\sigma^2$$

Der folgende Satz zeigt, dass $\bar{X}$ mit wachsendem Stichprobenumfang n immer weniger streut.

Satz 12.2
Die Zufallsvariablen $X_1, \ldots, X_n$ seien unabhängig und identisch verteilt mit $E(X_i) = \mu$ und $Var(X_i) = \sigma^2$ für $i = 1, 2, \ldots, n$. Dann gilt

$$\boxed{Var(\bar{X}) = \frac{\sigma^2}{n}} \tag{12.3}$$

Beweis:

$$Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\,Var\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\,n\,\sigma^2 = \frac{\sigma^2}{n}$$

□

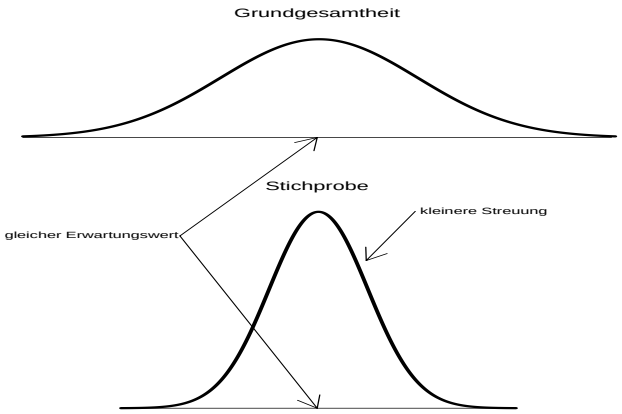
Die Stichprobenfunktion $\bar{X}$ besitzt für eine Zufallsstichprobe vom Umfang n also folgende Eigenschaften:

1. Der Erwartungswert von $\bar{X}$ ist gleich dem Erwartungswert des Merkmals in der Grundgesamtheit.

2. Die Streuung des Merkmals in der Grundgesamtheit ist größer als die Streuung von $\bar{X}$ in der Stichprobe.

Abbildung 12.2 veranschaulicht diesen Zusammenhang.

Abbildung 12.2: Verteilung eines Merkmals in einer Grundgesamtheit und Verteilung von $\bar{X}$ in einer Zufallsstichprobe aus dieser Grundgesamtheit



12.1.2 Normalverteilte Zufallsvariablen

Bisher haben wir nur den Erwartungswert und die Varianz von $\bar{X}$ betrachtet. Kennt man die Verteilung der Grundgesamtheit, so kann man in einigen Fällen Aussagen über die Verteilung von $\bar{X}$ machen. Ist die Grundgesamtheit normalverteilt, so gilt folgender Satz:

Satz 12.3
Seien $X_1, X_2, \ldots, X_n$ unabhängig und identisch mit den Parametern μ und σ^2 normalverteilt.
Sei

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

Dann ist S normalverteilt mit den Parametern $n\mu$ und $n\sigma^2$. □

Beispiel 107
Die Fahrzeit zur Uni sei normalverteilt mit Erwartungswert 30 und Varianz 9. Dann ist die Gesamtzeit S_3 von 3 Tagen normalverteilt mit $\mu = 90$ und $\sigma^2 = 27$. □

Wie der folgende Satz zeigt, ist auch $\bar{X}$ normalverteilt, wenn $X_1, \ldots, X_n$ normalverteilt sind.

Satz 12.4
Die Zufallsvariablen $X_1, X_2, \ldots, X_n$ seien unabhängige und identisch mit den Parametern μ und σ^2 normalverteilte Zufallsvariablen. Dann ist $\bar{X}$ normalverteilt mit den Parametern μ und σ^2/n.
Beweis:
Ist X normalverteilt mit den Parametern μ und σ^2, so ist aX normalverteilt mit den Parametern $a\mu$ und $a^2\sigma^2$. Für $\bar{X}$ gilt $\bar{X} = aS$ mit $a = 1/n$. □

Beispiel 107 (fortgesetzt)
Die mittlere Fahrzeit an drei Tagen ist normalverteilt $\mu = 30$ und $\sigma^2 = 3$. □

12.1.3 Bernoulliverteilte Zufallsvariablen

Oft ist der Anteil p eines Merkmals A in einer Grundgesamtheit von Interesse. Dies könnte zum Beispiel der Anteil der Raucher oder Wähler einer bestimmten Partei sein. In der Regel ist p unbekannt. Um einen Wert für p angeben zu können, wird eine Zufallsstichprobe vom Umfang n gezogen. Wir befragen also n Personen und erhalten eine Folge der Länge n, die aus A und $\bar{A}$ besteht. Unterstellen wir Unabhängigkeit und konstantes $p = P(A)$, so ist die Anzahl S der Personen, die die Eigenschaft A besitzen, binomialverteilt mit den Parametern n und p. Speziell gilt $E(S) = n\,p$ und $Var(S) = n\,p\,(1-p)$. Wir können S auch als Summe von unabhängigen, identisch verteilten Zufallsvariablen $X_1, \ldots, X_n$ darstellen. Hierzu definieren wir für die i-te gezogene Person eine Zufallsvariable X_i mit

$$X_i = \begin{cases} 1 & \text{wenn die Person die Eigenschaft A besitzt} \\ 0 & \text{sonst} \end{cases}$$

Es gilt $P(X_i = 1) = P(A) = p$ und $P(X_i = 0) = 1 - p$. Man nennt X_i eine mit dem Parameter p bernoulliverteilte Zufallsvariable. Es gilt $E(X_i) = p$ und $Var(X_i) = p\,(1-p)$. Also gibt

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

die Anzahl der Personen an, die die Eigenschaft A besitzen. Die Summe von n unabhängigen, identisch mit Parameter p bernoulliverteilten Zufallsvariablen $X_1, \ldots, X_n$ ist also binomialverteilt mit den Parametern n und p. Es gilt also

$$P(S = s) = \binom{n}{s} p^s \, (1-p)^{n-s}$$

Bildet man nun $\bar{X}$, so erhält man die relative Häufigkeit von A. Wir bezeichnen sie mit $\hat{p}$. Es gilt also

$$\hat{p} = \frac{S}{n} \tag{12.4}$$

Es folgt sofort

$$E(\hat{p}) = E(\frac{S}{n}) = \frac{1}{n} \, E(S) = \frac{1}{n} \, n \, p = p \tag{12.5}$$

und

$$Var(\hat{p}) = Var(\frac{S}{n}) = \frac{1}{n^2} \, Var(S) = \frac{1}{n^2} \, n \, p \, (1-p) = \frac{p \, (1-p)}{n} \tag{12.6}$$

Wir sehen, dass p das Zentrum der Verteilung von $\hat{p}$ bildet. Außerdem konzentriert sich die Verteilung von $\hat{p}$ mit wachsendem Stichprobenumfang n immer stärker um p.
Wir können Wahrscheinlichkeitsaussagen über die relative Häufigkeit $\hat{p}$ machen. Wegen Gleichung (12.4) gilt

$$P(\hat{p} \leq x) = P\left(\frac{S}{n} \leq x\right) = P(S \leq n \, x)$$

Beispiel 108
Eine faire Münze wird 10-mal hintereinander geworfen. Wie groß ist die Wahrscheinlichkeit, dass die relative Häufigkeit $\hat{p}$ von KOPF mehr als 0.4 und weniger als 0.6 beträgt? Die Zufallsvariable S sei binomialverteilt mit den Parametern $n = 10$ und $p = 0.5$. Dann gilt

$$P(0.4 < \hat{p} < 0.6) = P(4 < S < 6) = P(S = 5) = \binom{10}{5} 0.5^{10} = 0.246$$

□

12.1.4 Das schwache Gesetz der Großen Zahlen

Wir haben in den beiden letzten Abschnitten gesehen, dass sich die Verteilung von $\bar{X}$ immer stärker um $E(X) = \mu$ konzentriert. Wir wollen diese Aussage weiter formalisieren. Hierzu betrachten wir ein um μ symmetrisches Intervall $(\mu - \epsilon, \mu + \epsilon)$ und fragen uns, wie groß die Wahrscheinlichkeit ist, dass $\bar{X}$ Werte aus diesem Intervall annimmt. Wir suchen also

$$P(\mu - \epsilon < \bar{X} < \mu + \epsilon) \tag{12.7}$$

Dies können wir auch schreiben als

$$P(|\bar{X} - \mu| < \epsilon)$$

Aufgrund der Ungleichung von Tschebyscheff folgt

$$P(|\bar{X} - \mu| < \epsilon) \geq 1 - \frac{\sigma^2}{n \cdot \epsilon^2} \tag{12.8}$$

Wir sehen, dass diese Wahrscheinlichkeit mit wachsendem Stichprobenumfang n immer größer wird. Wenn wir also eine größere Sicherheit haben wollen, dass $\bar{X}$ in einem vorgegebenen Intervall liegt, so müssen wir unseren Stichpobenumfang erhöhen. Schauen wir uns den Grenzwert von (12.8) an. Es gilt

$$\lim_{n\to\infty} P(|\bar{X} - \mu| < \epsilon) \geq \lim_{n\to\infty} (1 - \frac{\sigma^2}{n \cdot \epsilon^2})$$

Hieraus folgt

$$\lim_{n\to\infty} P(|\bar{X} - \mu| < \epsilon) \geq 1$$

Da eine Wahrscheinlichkeit nicht größer als 1 werden kann, gilt also

$$\boxed{\lim_{n\to\infty} P(|\bar{X} - \mu_X| < \epsilon) = 1} \tag{12.9}$$

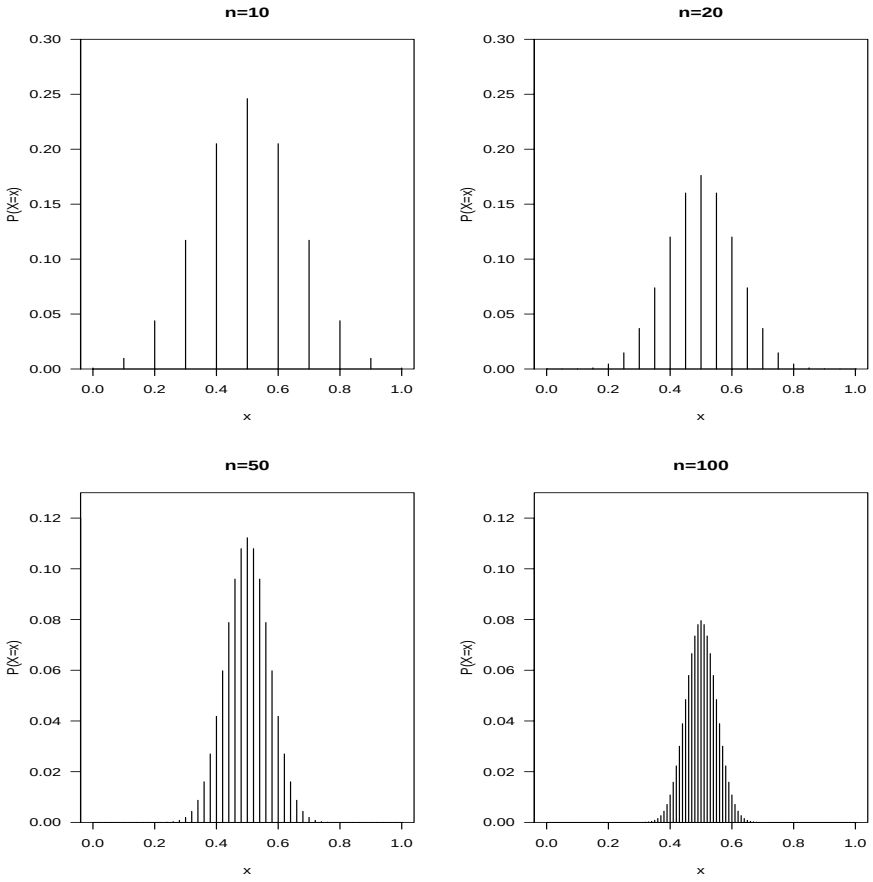
Diese Aussage nennt man auch das **Schwache Gesetz der Großen Zahlen**. Sind die Zufallsvariablen $X_1, \ldots, X_n$ identisch mit dem Parameter p bernoulliverteilt, so ist $\bar{X}$ gleich der relativen Häufigkeit $\hat{p}$ des interessierenden Ereignisses. Auf Grund des schwachen Gesetzes der Großen Zahlen gilt

$$\boxed{\lim_{n\to\infty} P(|\hat{p} - p| < \epsilon) = 1}\,.$$

Mit wachsendem Stichprobenumfang können wir uns also immer sicherer sein, dass die relative Häufigkeit $\hat{p}$ Werte annimmt, die in der Nähe von p liegen.

Abbildung 12.3 veranschaulicht dies für die relative Häufigkeit von KOPF beim n-maligen Wurf einer fairen Münze.

Abbildung 12.3: Wahrscheinlichkeitsfunktion der Anzahl KOPF beim n-maligen Wurf einer fairen Münze

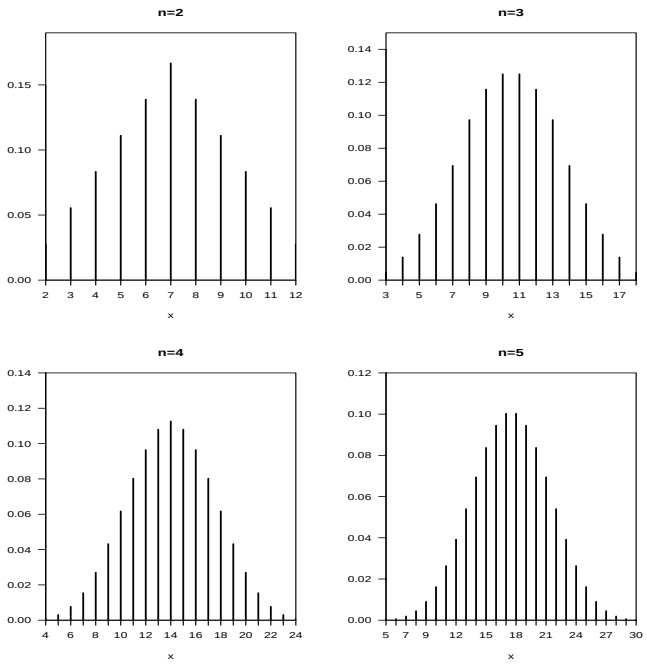


12.1.5 Der Zentrale Grenzwertsatz

Ist die Grundgesamtheit normalverteilt oder bernoulliverteilt, so kann man die Verteilung von $\bar{X}$ exakt angeben. Dies ist bei vielen anderen Verteilungen nicht möglich oder sehr mühselig.

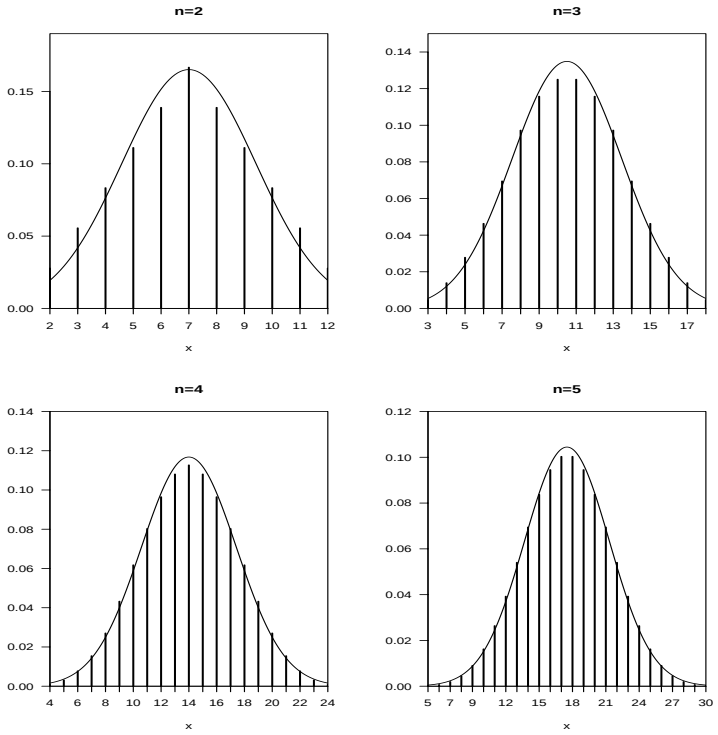
Beispiel 109
Ein Würfel wird n-mal geworfen. Uns interessiert die Verteilung der Summe S der Augenzahlen. Für $n = 2$ und $n = 3$ kann diese auch mit Papier und Bleistift in vetretbarer Zeit bestimmt werden. Für größere Werte von n muss der Computer helfen. In Abbildung 12.4 sind die Wahrscheinlichkeitsfunktionen für $n = 2, 3, 4, 5$ grafisch dargestellt.

Abbildung 12.4: Verteilung der Summe der Augenzahlen beim n-maligen Wurf eines Würfels



Die Stabdiagramme besitzen mit wachsendem Stichprobenumfang n immer mehr die Gestalt der Normalverteilung. In Abbildung 12.5 ist in jeder Grafik die Dichtefunktion der Normalverteilung eingezeichnet. Ist X_i die Augenzahl mit i-ten Wurf, so gilt $E(X_i) = 3.5$ und $Var(X_i) = 35/12$. Für den Erwartungswert der Normalverteilung wählen wir also $3.5n$ und für die Varianz $35n/12$.

Abbildung 12.5: Verteilung der Summe der Augenzahlen beim n-maligen Wurf eines Würfels mit Dichtefunktion der Normalverteilung



□

Eine Begründung für das im Beispiel auftretende Phänomen liefert der folgende Satz.

Satz 12.5
$X_1, X_2, \ldots, X_n, \ldots$ sei eine Folge von unabhängigen, identisch mit $E(X_i) = \mu$ und $Var(X_i) = \sigma^2$ verteilten Zufallsvariablen. Sei

$$S_n = \sum_{i=1}^{n} X_i \,.$$

Dann gilt

$$\lim_{n\to\infty} P\left(\frac{S_n - n\,\mu}{\sqrt{n}\,\sigma} \leq z\right) = \Phi(z)$$

wobei $\Phi(z)$ die Verteilungsfunktion der Standardnormalverteilung an der Stelle z ist. □

Der zentrale Grenzwertsatz sagt aus, dass die Verteilungsfunktion der standardisierten Summe gegen die Verteilungsfunktion der Standardnormalverteilung konvergiert. Von der Verteilung der X_i wird nur gefordert, dass Erwartungswert und Varianz existieren. Die Verteilung kann also diskret oder stetig, schief oder symmetrisch, unimodal oder multimodal sein.

Man kann nun zeigen, dass nicht nur die standardisierte Summe, sondern auch die Summe selbst approximativ normalverteilt ist. Es gilt also, dass die Summe von unabhängigen, identisch mit $E(X_i) = \mu$ und $Var(X_i) = \sigma^2$ verteilten Zufallsvariablen $X - 1, \ldots, X_n$ approximativ mit den Parametern $n\mu$ und $n\sigma^2$ normalverteilt ist. Dabei bedeutet approximativ, dass n groß sein muss. Es gibt eine Vielzahl von Faustregeln. Am häufigsten findet man, dass n bei symmetrischen Verteilungen mindestens 20 und bei schiefen Verteilungen mindestens 30 sein muss.

Auch der Mittelwert $\bar{X}$ von unabhängigen, identisch mit $E(X_i) = \mu$ und $Var(X_i) = \sigma^2$ verteilten Zufallsvariablen ist approximativ normalverteilt und zwar mit den Parametern μ und σ^2/n.

Bevor wir uns eine wichtige Anwendung des zentralen Grenzwertsatzes anschauen, wollen wir uns überlegen, was es eigentlich bedeutet, eine diskrete Verteilung durch eine stetige Verteilung zu approximieren.

Beispiel 110

Wir werfen einen fairen Würfel zweimal. Alle möglichen Ergebnisse sind

$$\begin{array}{cccccc} (1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ (3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\ (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\ (5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6) \end{array}$$

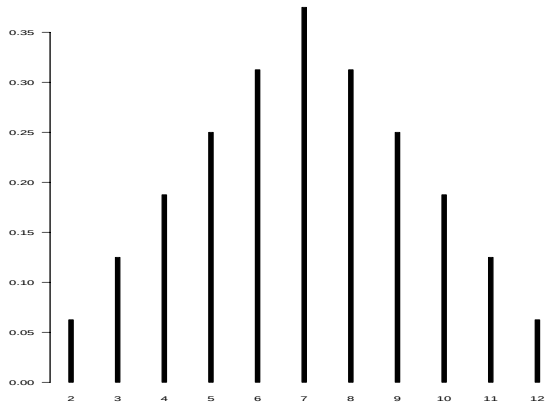
Sei X_1 die Augenzahl beim ersten Wurf und X_2 die Augenzahl beim zweiten Wurf. Uns interessiert die Verteilung der Summe $S = X_1 + X_2$. In Tabelle 12.11 ist die Verteilung von S zu finden.

Tabelle 12.11: Verteilung der Summe der Augenzahlen beim zweimaligen Wurf eines Würfels

s	2	3	4	5	6	7	8	9	10	11	12
$P(S=s)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Abbildung 12.6 zeigt die Wahrscheinlichkeitsfunktion von S mit der Dichtefunktion der Normalverteilung mit den Parametern $\mu = 7$ und $\sigma^2 = 70/12$.

Abbildung 12.6: Wahrscheinlichkeitsfunktion der Augensumme beim zweimaligen Würfeln mit Dichtefunktion der Normalverteilung



□

Approximieren wir die Verteilung einer diskreten Zufallsvariable X durch die Normalverteilung mit den Parametern $\mu = E(X)$ und $\sigma^2 = Var(X)$, so gilt auf Grund des zentralen Grenzwertsatzes approximativ

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

Für kleine Stichprobenumfänge ist diese Approximation schlecht, wie das folgende Beispiel zeigt.

Beispiel 110 (fortgesetzt)
Es gilt exakt

$$P(S \leq 4) = \frac{1}{6} = 0.167$$

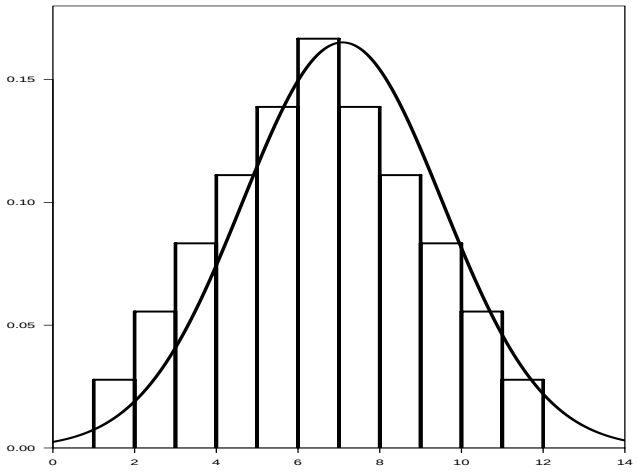
Auf Grund des zentralen Grenzwertsatzes gilt approximativ

$$P(S \leq 4) = \Phi\left(\frac{4-7}{\sqrt{70/12}}\right) = \Phi(-1.242) = 0.107\,.$$

Die Approximation ist schlecht. In Abbildung 12.7 können wir erkennen, woran dies liegt. Die exakte Wahrscheinlichkeit ist gleich der Summe der

Flächen der ersten drei Rechtecke. Diese unterschätzen wir durch die Fläche an der Stelle 4.

Abbildung 12.7: Approximation ohne Stetigkeitskorrektur



□

Das Beispiel zeigt, wie wir die Approximation verbessern können. Wir müssen die Rechtecke um die Realisationsmöglichkeiten von X zentrieren. Dies heißt, dass gilt

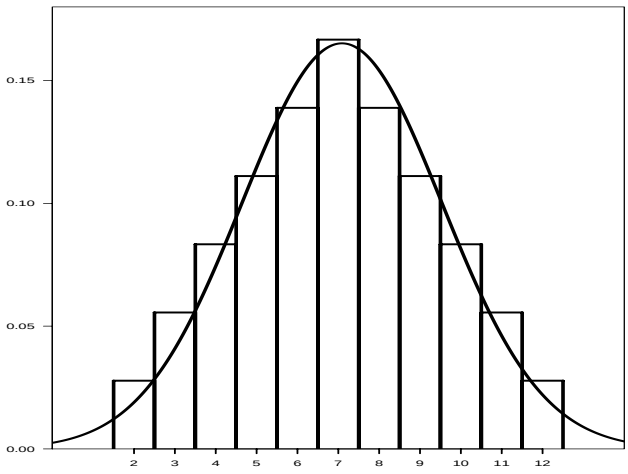
$$P(X \leq x) = \Phi\left(\frac{x + 0.5 - E(X)}{\sqrt{Var(X)}}\right).$$

Beispiel 110 (fortgesetzt)
Abbildung 12.8 auf der nächsten Seite zeigt die Vorgehensweise.
Es gilt approximativ

$$P(S \leq 4) = \Phi\left(\frac{4 + 0.5 - 7}{\sqrt{70/12}}\right) = \Phi(-1.035) = 0.150\,.$$

Wir sehen, dass die Approximation viel besser ist.

Abbildung 12.8: Approximation mit Stetigkeitskorrektur



□

Man spricht in diesem Fall von der **Stetigkeitskorrektur**.
Wie wir auf Seite 347 gesehen haben, können wir eine binomialverteilte Zufallsvariable S als Summe von unabhängigen, identisch mit dem Parameter p bernoulliverteilten Zufallsvariablen $X_1, X_2, \ldots, X_n$ darstellen.
Mit

$$P(X_i = 1) = p \qquad P(X_i = 0) = 1 - p$$

gilt also

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

Wir können also die Ergebnisse des zentralen Grenzwertsatzes verwenden, um Wahrscheinlichkeiten bei der Binomialverteilung für großes n zu bestimmen.
Es gilt approximativ

$$P(S \leq s) = \Phi\left(\frac{s - n\,p}{\sqrt{n\,p\,(1-p)}}\right)$$

Berücksichtigen wir die Stetigkeitskorrektur, so gilt

$$P(S \leq s) = \Phi\left(\frac{s + 0.5 - n\,p}{\sqrt{n\,p\,(1-p)}}\right)$$

Beispiel 111
Eine faire Münze werde 100-mal geworfen. Wie groß ist die Wahrscheinlichkeit, dass höchstens 40-mal KOPF fällt? Die Anzahl S der Würfe, bei denen KOPF fällt, ist auf Grund des zentralen Grenzwertsatzes approximativ normalverteilt mit den Parametern $\mu = n\,p = 50$ und $\sigma^2 = n\,p\,(1-p) = 25$. Ohne Stetigkeitskorrektur erhalten wir

$$P(S \leq 40) = \Phi\left(\frac{40-50}{\sqrt{25}}\right) = \Phi(-2) = 0.0227$$

und mit Stetigkeitskorrektur

$$P(S \leq 40) = \Phi\left(\frac{40+0.5-50}{\sqrt{25}}\right) = \Phi(-1.9) = 0.0287$$

Der exakte Wert ist

$$P(S \leq 40) = 0.0284\,.$$

Die Werte unterscheiden sich nicht sehr stark. Für kleine Werte von n ist der Unterschied jedoch beträchtlich, wie das Beispiel $n = 4$ und $p = 0.5$ zeigt. Wir wollen die Wahrscheinlichkeit bestimmen, höchstens einmal KOPF zu beobachten. Ohne Stetigkeitskorrektur erhalten wir

$$P(S \leq 1) = \Phi\left(\frac{1-4\cdot 0.5}{\sqrt{4\cdot 0.5\cdot 0.5}}\right) = \Phi(-1) = 0.1587$$

und mit Stetigkeitskorrektur

$$P(S \leq 1) = \Phi\left(\frac{1+0.5-4\cdot 0.5}{\sqrt{4\cdot 0.5\cdot 0.5}}\right) = \Phi(-0.5) = 0.3085$$

Der exakte Wert beträgt

$$\begin{aligned} P(S \leq 1) &= P(S=0) + P(S=1) = \binom{4}{0} 0.5^4 + \binom{4}{1} 0.5^4 \\ &= 0.0625 + 4 \cdot 0.0625 = 0.3125 \end{aligned}$$

Wir sehen, dass wir mit der Stetigkeitskorrektur fast den exakten Wert erhalten. □

Wir wissen, dass bei Normalverteilung die Wahrscheinlichkeit für das zweifache zentrale Schwankungsintervall ungefähr 0.95 beträgt. 95 Prozent aller Beobachtungen liegen also im zweifachen zentralen Schwankungsintervall.

Da die relative Häufigkeit $\hat{p}$ approximativ normalverteilt ist mit den Parametern p und $p\,(1-p)/n$, liegt also mit Wahrscheinlichkeit 0.95 die relative Häufigkeit im Intervall

$$\left[p - 2 \cdot \sqrt{\frac{p\,(1-p)}{n}}, p + 2 \cdot \sqrt{\frac{p\,(1-p)}{n}}\right]$$

wenn man einen Bernoulliprozeß der Länge n mit Erfolgswahrscheinlichkeit p beobachtet.

Beispiel 112
Werfen wir also 100-mal hintereinander eine faire Münze, so wird die relative Häufigkeit von KOPF mit Wahrscheinlichkeit 0.95 zwischen

$$0.5 - 2 \cdot \sqrt{\frac{0.5 \cdot 0.5}{100}} = 0.5 - 2 \cdot 0.05 = 0.4$$

und

$$0.5 + 2 \cdot \sqrt{\frac{0.5 \cdot 0.5}{100}} = 0.5 + 2 \cdot 0.05 = 0.6\,,$$

also zwischen 0.4 und 0.6 liegen. Wird die Münze hingegen 10000-mal geworfen, so wird die relative Häufigkeit KOPF mit Wahrscheinlichkeit 0.95 zwischen 0.49 und 0.51 liegen.

□

Die Aussage des zentralen Grenzwertsatzes ist auch erfüllt, wenn die Zufallsvariablen unterschiedliche Verteilungen besitzen.

Satz 12.6
Sind die Zufallsvariablen $X_1, \ldots, X_n$ unabhängig und mit $E(X_i) = \mu_i$ und $Var(X_i) = \sigma_i^2$ verteilt, dann ist $\sum_{i=1}^n X_i$ normalverteilt mit dem Erwartungswert $\sum_{i=1}^n \mu_i$ und der Varianz $\sum_{i=1}^n \sigma_i^2$. □

Oft wirken viele Einflussfaktoren auf ein Merkmal. So hängt die Fahrzeit zur Uni davon ab, ob die Ampeln rot oder grün sind, ob man an einer Vorfahrtsstraße warten muss oder nicht. Wenn diese Einflüsse additiv wirken, so wird die Fahrzeit auf Grund des oben gesagten approximativ normalverteilt sein.

12.2 Verteilung des Maximums und des Minimums

Eine Kette ist nur so stark wie ihr schwächstes Glied. Deiche werden so geplant, dass sie dem höchsten Pegelstand noch Widerstand leisten können.

Deshalb sind die Verteilung des Maximums und des Minimums der Beobachtungen einer Stichprobe wichtig. Der folgende Satz gibt an, wie das Minimum und Maximum der Beobachtungen in einer Zufallsstichprobe verteilt sind.

Satz 12.7
Die Zufallsvariablen $X_1, \ldots, X_n$ seien uabhängig und identisch mit stetiger Verteilungsfunktion $F_X(x)$ verteilt.
Dann ist die Verteilungsfunktion von $V = min\{X_1, \ldots, X_n\}$ gegeben durch

$$\boxed{F_V(v) = 1 - (1 - F_X(v))^n}$$

und die Verteilungsfunktion von $W = max\{X_1, \ldots, X_n\}$ gegeben durch

$$\boxed{F_W(w) = F_X(w)^n}$$

Beweis:
Wir betrachten nur das Maximum. Der Beweis für das Minimum verläuft analog. Beim Maximum berücksichtigen wir, dass alle Beobachtungen kleiner gleich einer Zahl x sind, wenn das Maximum kleiner gleich x ist. Es gilt also

$$\begin{aligned} F_W(w) &= P(W \leq w) = P(max\{X_1, \ldots, X_n\} \leq w) \\ &= P(X_1 \leq w, \ldots, X_n \leq w) = P(X_1 \leq w) \cdots P(X_n \leq w) \\ &= F_X(w) \cdot \ldots \cdot F_X(w) = F_X(w)^n \end{aligned}$$

□

Beispiel 113
Wir schauen uns noch die Verteilung des Maximums einer Zufallssstichprobe aus einer Gleichverteilung an. Die Zufallsvariablen $X_1, \ldots, X_n$ sind also unabhängig und identisch auf $[0, b]$ gleichverteilt. Es gilt

$$F_{X_i}(x_i) = \begin{cases} 0 & \text{für } x_i < 0 \\ \frac{x_i}{b} & \text{für } 0 \leq x_i \leq b \\ 1 & \text{für } x_i > b \end{cases}$$

Also folgt für W:

$$F_W(w) = \begin{cases} 0 & \text{für } w < 0 \\ \frac{w^n}{b^n} & \text{für } 0 \leq w \leq b \\ 1 & \text{für } w > b \end{cases}$$

Somit gilt für die Dichtefunktion von W:

$$f_W(w) = \begin{cases} \frac{n\, w^{n-1}}{b^n} & \text{für } 0 \leq w \leq b \\ 0 & \text{sonst} \end{cases}$$

□

12.3 Simulation

Wir haben eine Reihe von Verfahren kennengelernt, mit denen man die Verteilung von $\bar{X}$ bestimmen kann. Ist die zugrundeliegende Grundgesamtheit normalverteilt oder bernoulliverteilt, so kennt man die exakte Verteilung von $\bar{X}$. Für große Stichprobenumfänge kann man auf den zentralen Grenzwertsatz zurückgreifen. Es besteht aber auch die Möglichkeit, die Verteilung einer Stichprobenfunktion $g(X_1, \ldots, X_n)$ mit einer **Simulation** approximativ zu bestimmen. Bei dieser erzeugt man eine Anzahl B von Stichproben aus der zugrundeliegenden Verteilung und bestimmt für jede dieser Stichproben den Wert der interessierenden Stichprobenfunktion $g(X_1, \ldots, X_n)$. Die aus der Simulation gewonnene Verteilung der Stichprobenfunktion $g(X_1, \ldots, X_n)$ approximiert dann die theoretische Verteilung von $g(X_1, \ldots, X_n)$. Dies macht man sich folgendermaßen klar. Sei A das Ereignis, dass $g(X_1, \ldots, X_n)$ Werte annimmt, die kleiner oder gleich x sind. Dann ist

$$p = P(A) = P(g(X_1, \ldots, X_n) \leq x)$$

der Wert der Verteilungsfunktion von $g(X_1, \ldots, X_n)$ an der Stelle x. Aus den B Werten von $g(X_1, \ldots, X_n)$, die wir mit der Simulation gewonnen haben, bestimmen wir den Anteil $\hat{p}$, bei denen A eingetreten ist. Wir wissen, dass

$$E(\hat{p}) = p$$

und

$$Var(\hat{p}) = \frac{p\,(1-p)}{B}$$

gilt. Wir sehen, dass mit wachsender Anzahl B der Wiederholungen die Varianz von $\hat{p}$ immer kleiner wird und somit die Genauigkeit von $\hat{p}$ immer größer wird. Wir werden im Kapitel über Schätzung noch genauer auf diesen Sachverhalt eingehen.

Beispiel 114
Wir betrachten die Augensumme S beim zweimaligen Wurf eines fairen Würfels. In der zweiten Spalte von Tabelle 12.12 ist die Verteilung von S zu finden.

Tabelle 12.12: Verteilung der Summe der Augenzahlen beim zweimaligen Wurf eines Würfels

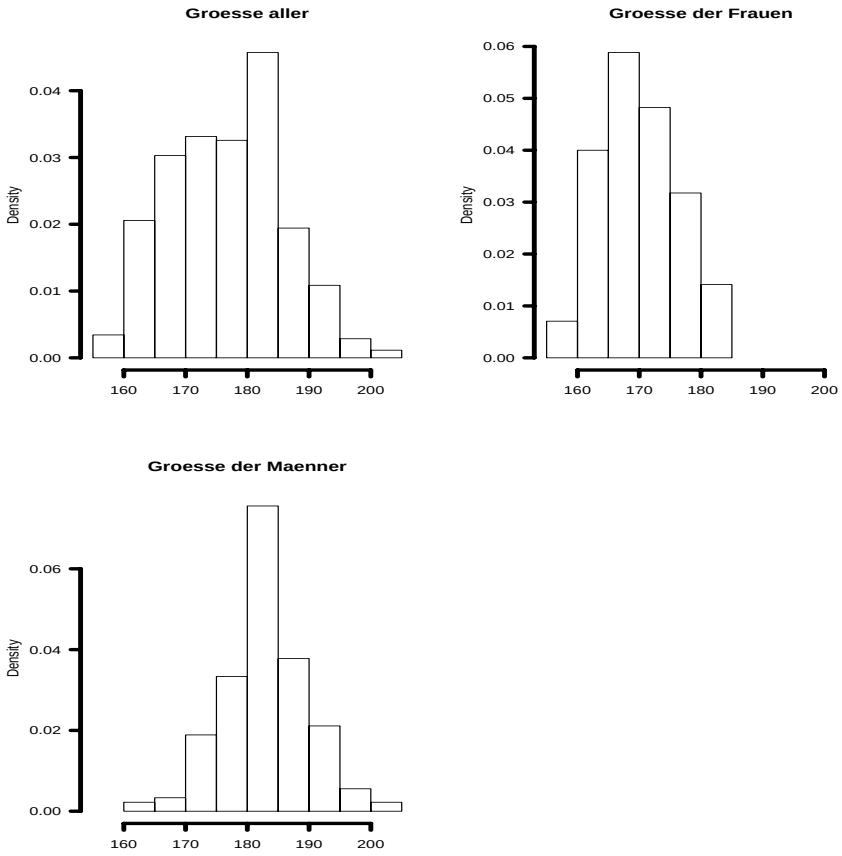
s	$P(S=s)$	$h(S=s)$ $B=5000$	$h(S=s)$ $B=5000$	$h(S=s)$ $B=10000$
2	0.028	0.022	0.032	0.027
3	0.056	0.038	0.038	0.055
4	0.083	0.086	0.074	0.086
5	0.111	0.132	0.106	0.111
6	0.139	0.148	0.142	0.139
7	0.167	0.174	0.162	0.164
8	0.139	0.142	0.164	0.140
9	0.111	0.098	0.116	0.111
10	0.083	0.078	0.092	0.079
11	0.056	0.046	0.038	0.060
12	0.028	0.036	0.036	0.031

Wir wollen diesen Zufallsvorgang aber auch simulieren. Dies ist sehr einfach. Man muss nur den Würfel B-mal zweimal hintereinander werfen und bei jeder Wiederholung die Augensumme notieren. Dies wurde 5000-mal gemacht. Es gilt also $B = 5000$. In der dritten Spalte von Tabelle 12.12 sind die relativen Häufigkeiten der einzelnen Augensummen zu finden. Wir sehen, dass die relativen Häufigkeiten und die Wahrscheinlichkeiten sich unterscheiden. Dies wundert uns aber nicht, da wir mit der Simulation die Verteilung nur approximieren. Das Ergebnis der Simulation hängt natürlich von den Stichproben ab, die wir beobachten. Bei einer weiteren Simulation werden wir ein anderes Ergebnis erhalten. Die vierte Spalte von Tabelle 12.12 bestätigt dies. Die fünfte Spalte von Tabelle 12.12 zeigt das Ergebnis einer Simulation mit 10000 Wiederholungen. □

Man muss nicht einen Würfel werfen, um die Ergebnisse in Tabelle 12.12 zu erhalten. Man kann auch den Computer verwenden. Wie man hierbei in `R` vorzugehen hat, werden wir ab Seite 365 lernen.

Mit Hilfe einer Simulation kann man aber auch die Unterschiede zwischen Verfahren verdeutlichen. Wir betrachten hierzu eine Grundgesamtheit von 350 Studienanfängern, von denen 170 weiblich sind. Abbildung 12.9 zeigt die Histogramme der Körpergröße aller 350 Studierenden, der Frauen und der Männer.

Abbildung 12.9: Histogramm der Körperöße

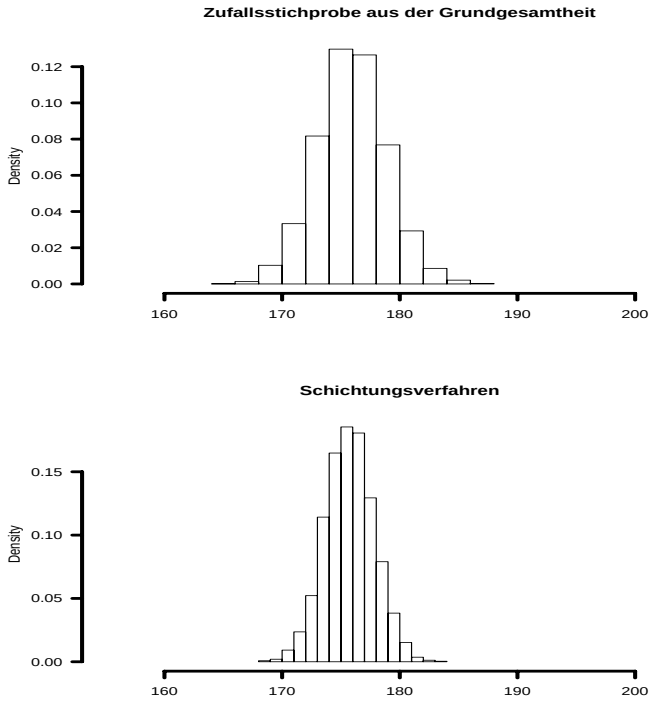


Wir sehen, dass die Verteilung der Körpergröße aller und der Körpergröße der Frauen schief ist, während die Verteilung der Körpergröße der Männer symmetrisch ist. Der Mittelwert der Körpergröße aller Personen beträgt 175.9 cm. Bei den Frauen ist der Mittelwert 169.2 cm und bei den Männern 182.2. Die Varianz der Körpergröße beträgt bei allen Studierenden 84.8, bei den Frauen 37.6 und den Männern 47.2.

Wir wollen uns nun anschauen, wie $\bar{X}$ in einer Stichprobe vom Umfang $n = 10$ verteilt ist. Hierzu betrachten wir zwei Fälle. Im ersten Fall ziehen wir eine Zufallsstichprobe vom Umfang $n = 10$ mit Zurücklegen aus der Grundgesamtheit. Im zweiten Fall bilden wir zwei Schichten, wobei die erste Schicht aus den Frauen und die zweite aus den Männern besteht. Aus jeder der beiden Schichten ziehen wir eine Zufallsstichprobe vom Umfang $n = 5$ mit Zurücklegen. Wir bestimmen dann den Mittelwert aller 10 Beobachtungen.

In beiden Fällen wiederholen wir den beschriebenen Prozess 5000-mal. Abbildung 12.10 zeigt die Histogramme.

Abbildung 12.10: Histogramme von $\bar{X}$

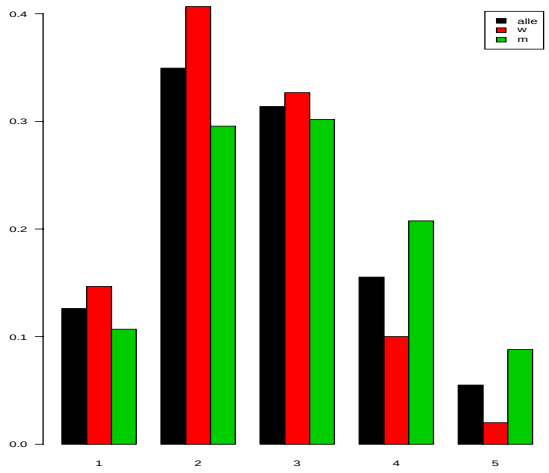


Wir sehen, dass die Streuung von $\bar{X}$ beim Schichtungsverfahren viel kleiner ist als bei einer Zufallsstichprobe aus der Grundgesamtheit. Die Varianz von $\bar{X}$ bei der Zufallsstichprobe beträgt 8.6 und beim Schichtungsverfahren 4.3. Die Varianz ist beim Schichtungsverfahren kleiner, weil die Männer im Mittel 13 cm größer als die Frauen sind, und die Körpergröße in den beiden Schichten weniger streut als in der Grundgesamtheit. Am Histogramm der Zufallsstichprobe aus der Grundgesamtheit sehen wir aber auch, dass der Zentrale Grenzwertsatz für $n = 10$ schon Wirkung zeigt. Obwohl die Verteilung in der Grundgesamtheit schief ist, ähnelt die Verteilung von $\bar{X}$ für $n = 10$ schon der Normalverteilung.

Schauen wir uns noch ein Beispiel an. Wir betrachten eine Grundgesamtheit von 309 Studienanfängern, von denen 150 weiblich sind. Abbildung 12.11 zeigt das Stabdiagramm der Abiturnote in Mathematik aller 309 Studieren-

den. Außerdem sind noch das Stabdiagramm der Frauen und das Stabdiagramm der Männer abgebildet.

Abbildung 12.11: Stabdiagramm der Abiturnote im Abitur bei allen, den weiblichen und den männlichen Studierenden

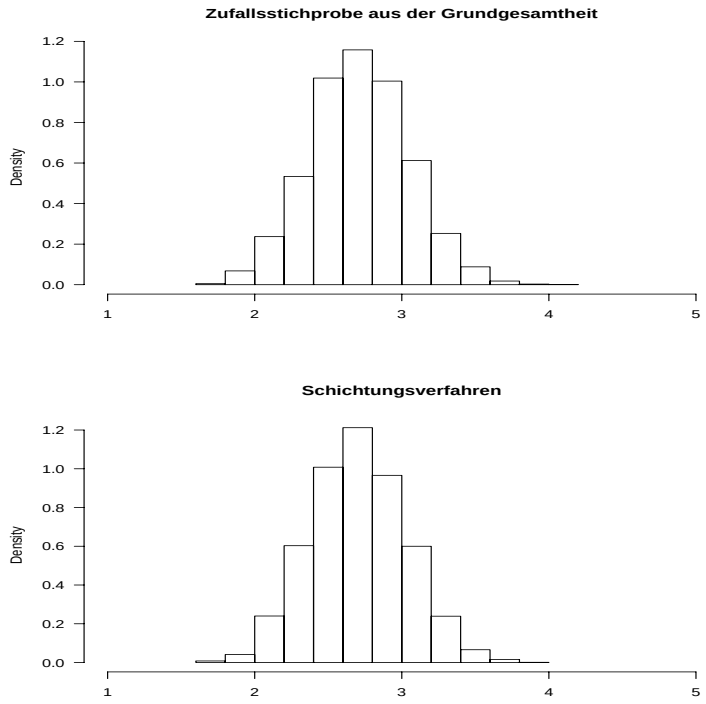


Wir sehen, dass die Verteilung der der Abiturnote in Mathematik in allen drei Gruppen schief ist. Der Mittelwert der Note aller Personen beträgt 2.66. Bei den Frauen ist der Mittelwert 2.44 und bei den Männern 2.87. Die Varianz der Note beträgt bei allen Personen 1.12, bei den Frauen 0.87 und bei den Männern 1.28.

Wir wollen uns wieder anschauen, wie $\bar{X}$ in einer Stichprobe vom Umfang $n = 10$ verteilt ist. Hierzu betrachten wir wieder zwei Fälle. Im ersten Fall ziehen wir eine Zufallsstichprobe vom Umfang $n = 10$ mit Zurücklegen aus der Grundgesamtheit. Im zweiten Fall bilden wir zwei Schichten, wobei die erste Schicht aus den Frauen und die zweite aus den Männern besteht. Aus jeder der beiden Schichten ziehen wir eine Zufallsstichprobe vom Umfang $n = 5$ mit Zurücklegen. Wir bestimmen dann den Mittelwert aller 10 Beobachtungen.

In beiden Fällen wiederholen wir den beschriebenen Prozess 5000-mal. Für $n = 10$ nimmt $\bar{X}$ so viele Werte an, dass wir Klassen bilden und die Histogramme zeichnen. Abbildung 12.12 zeigt diese.

Abbildung 12.12: Histogramme von $\bar{X}$



Wir sehen, dass der Schichtungseffekt viel kleiner ist als beim vorhergehenden Beispiel. Das Histogramm beim Schichtungsverfahren konzentriert sich ein wenig mehr um den Mittelwert als das Histogramm bei der Zufallsauswahl. Die Mittelwerte in den beiden Schichten unterscheiden sich zwar, aber die Varianz der Note bei den Männern ist sogar größer als die Varianz der Note in der Grundgesamtheit.

Die Simulation ist ein wertvolles Werkzeug, mit dem man die Verteilung einer Stichprobenfunktion approximieren kann. Man kann sich mit einer Simulation sehr schnell ein Bild von der Verteilung einer Stichprobenfunktion machen. Schauen wir uns noch ein Beispiel an.

Beispiel 115

Im Total Quality Management benötigt man den Erwartungswert und die Varianz der Spannweite R einer Zufallsstichprobe vom Umfang n aus einer Standardnormalverteilung. Auch diese können wir mit Hilfe einer Simulation schätzen. Wir ziehen $B = 10000$ Stichproben vom Umfang n aus der Standardnormalverteilung und bestimmen für jede den Wert der Spannweite

als Differenz aus Maximum und Minimum der Werte. Den Erwartungswert schätzen wir durch den Mittelwert und die Varianz durch die Stichprobenvarianz der B Spannweiten. Tabelle 12.13 zeigt die Ergebnisse für $n = 2, 3, 4, 5$.

Tabelle 12.13: Schätzung des Erwartungswerts und der Varianz der Spannweite der Bobachtungen in einer Zufallsstichprobe aus der Standardnormalverteilung für $n = 10$

n	$E(R)$	$Var(R)$
2	1.135	0.859
3	1.703	0.884
4	2.054	0.881
5	2.312	0.859

□

12.4 Simulation in R

In `R` können wir die Funktionen zum Erzeugen von Zufallszahlen zum Simulieren verwenden. Wir wollen uns dies für die beispiele im letzten Kapitel anschauen. Damit die Ergebnisse reproduziert werden können, setzen wir zu Beginn den Startwert des Zufallszahlengenerators auf 942003.

```
> set.seed(942003)
```

Beginnen wir mit dem zweimaligen Wurf des Würfels. Hierzu können wir die Funktion `sample` verwenden. Der Aufruf

```
> sample(1:6,1)
[1] 6
```

simuliert das Ergebnis eines einmaligen Wurfs eines Würfels. Um zweimal zu würfeln, erzeugen wir zwei Zufallszahlen, wobei wir berücksichtigen, dass wir mit Zurücklegen ziehen:

```
> z<-sample(1:6,2,replace=T)
> z
[1] 5 5
```

Nun müssen wir nur noch die beiden Komponenten von `z` addieren

```
> sum(z)
[1] 10
```

Wir können natürlich auch alles aufeinmal machen

```
> sum(sample(1:6,2,replace=T))
[1] 2
```

Beachten Sie, dass wir ein anderes Ergebnis erhalten, da der Startwert des Zufallszahlengenerators sich durch die vorherigen Ziehungen verändert hat. Nun wissen wir, wie man den zweimaligen Wurf eines Würfels in `R` simuliert und die Summe der Augenzahlen berechnet. Jetzt müssen wir dies B-mal durchführen. Wir wählen $B = 5000$.

```
> B<-5000
```

Für jede Wiederholung erhalten wir einen Wert der Summe der Augenzahlen. Da wir die Häufigkeitsverteilung dieser Zahlen bestimmen wollen, müssen wir uns alle Zahlen merken. Da wir wissen, wie viele Zahlen wir erhalten, initialisieren wir einen Vektor `erg` der Länge `B`, dessen i-te Komponente den Wert der interessierenden Stichpobenfunktion bei der i-ten Wiederholung enthalten wird.

```
> erg<-rep(0,B)
```

Der Vektor `erg` hat die Länge 5000 und besteht nur aus Nullen.

```
> length(erg)
[1] 5000
> table(erg)
erg
   0
5000
```

Nun müssen wir eine Stichprobe nach der anderen erzeugen, für jede dieser Stichprobe die Summe der Werte bestimmen und diese Summe der entsprechenden Komponente von `erg` zuweisen. Hierzu verwenden wir eine **Iteration**. Bei dieser wird eine Folge von Befehlen mit unterschiedlichen Werten ausgeführt. Kennt man die Anzahl der Wiederholungen, wendet man eine `for`-Schleife an. Diese besitzt folgenden Aufbau:

```
> for(v in Werte} Befehlsfolge
```

Dabei ist `v` eine Variable und `Werte` ein Vektor der Länge `B`. Der Befehl wird so abgearbeitet, dass der Variablen `v` jede Komponente von `Werte` zugewiesen wird und dann die Befehlsfolge ausgeführt wird. Soll eine Befehlsfolge B-mal ausgeführt werden, so wählt man für `v` eine Zählvariable. Typische Namen für Zählvariablen sind `i`, `j` oder `k`. Der Vektor `Werte` enthält dann die natürlichen Zahlen $1, 2, \ldots, B$. Diesen erhalten wir durch

```
> 1:B
```

Wir geben hier das Ergebnis nicht an, da dies sehr lang ist. Schauen wir uns die Iteration für das Würfelbeispiel an, wobei wir den Zufallszahlengenerator auf einen Startwert setzen, um das Ergebnis reproduzieren zu können.

```
> set.seed(942003)
> B<-5000
> erg<-rep(0,B)
> for(i in 1:B) erg[i]<-sum(sample(6,2,replace=T))
```

Mit der Funktion `table` erstellen wir die Tabelle der absoluten Häufigekeiten.

```
> table(erg)
erg
  2   3   4   5   6   7   8   9  10  11  12
154 271 390 556 685 860 664 534 431 290 165
```

Die relativen Häufigkeiten erhalten wir durch

```
> table(erg)
> table(erg)/sum(erg)
erg
       2       3       4       5       6       7
 0.00437 0.00770 0.01108 0.01580 0.01946 0.02443
       8       9      10      11      12
 0.01886 0.01517 0.01224 0.00824 0.00468
```

Wir können nun noch die theoretischen Werte mit den praktischen Werten in einer Grafik vergleichen.

Wenden wir uns dem zweiten Beispiel zu. Hier wollen wir den Erwartungswert und die Varianz der Spannweite einer Zufallsstichprobe vom Umfang n durch eine Simulation schätzen. Die Funktion `range` bestimmt das Minimum und Maximum eines Vektors:

```
> range(1:6)
[1] 1 6
```

Mit folgenden Befehlen erhalten wir die gesuchten Schätzer:

```
set.seed(112003)
m<-matrix(0,4,2)
for (i in 1:4)
 {erg<-rep(0,10000)
   for (j in 1:10000)
   {x<-range(rnorm(i+1)) ; erg[j]<-x[2]-x[1]}
  m[i,1]<-mean(erg) ; m[i,2]<-sd(erg)  }
```

Das Ergebnis steht in `m`.

```
> m
          [,1]      [,2]
[1,] 1.135260 0.8592016
[2,] 1.702873 0.8841447
[3,] 2.053666 0.8809268
[4,] 2.312475 0.8592217
```

Kapitel 13

Schätzung

In der schließenden Statistik will man auf Basis einer Stichprobe $x_1, \ldots, x_n$ Aussagen über die Verteilung eines Merkmals X in einer Grundgesamtheit machen. Ist die Grundgesamtheit die Haushalte der BRD, so könnte das das Durchschnittseinkommen oder das Einkommen, das von 90 Prozent der Haushalte nicht übertroffen wird, von Interesse sein. Die Charakteristika einer Verteilung nennen wir auch **Parameter** und bezeichnen sie mit θ. Ist X das Einkommen eines Haushaltes, so ist $\theta = E(X)$ das Durchschnittseinkommen und $\theta = x_{0.9}$ das 0.9-Quantil. Oft ist man auch daran interessiert, welcher Anteil p der Merkmalsträger in einer Grundgesamtheit eine bestimmte Eigenschaft besitzt. So könnte der Anteil der Haushalte, die mindestens einen PKW besitzen, interessieren.

Gibt man einen Wert für einen Parameter an, so spricht man von **Punktschätzung**. Den Wert nennt man einen **Punktschätzer**. Gibt man hingegen ein Intervall an, so spricht man von **Intervallschätzung**. Das Intervall bezeichnet man als **Konfidenzintervall**.

Manchmal will man aber auch den nächsten Wert x_{n+1} einer Zufallsvariablen auf Basis einer Stichprobe $x_1, \ldots, x_n$ vorhersagen. Man spricht in diesem Fall von Prognose. Wie bei der Schätzung kann man auch bei der Prognose Intervalle für den zukünftigen Wert angeben. Man spricht in diesem Fall von einem **Prognoseintervall**. Außerdem gibt es noch Intervalle, die einen vorgegebenen Anteil der Realisationsmöglichkeiten einer Zufallsvariablen enthalten. Diese heißen **Toleranzintervalle**.

Ein Beispiel soll die unterschiedlichen Fragestellungen verdeutlichen.

Beispiel 116
Ein Arbeitnehmer hat eine neue Stelle angenommen. Er will wissen, welche Charakteristika die Fahrzeit zur Arbeitsstelle besitzt. Deshalb notiert er die Fahrzeit an 10 aufeinander folgenden Tagen. Hier sind die Werte in Sekunden:

```
2318 2457 2282 2428 2376 2439 2272 2626 2255 2577
```

Interessiert ihn, wie lange er im Mittel zur Arbeitsstelle benötigt, so sucht er einen Schätzwert für den Erwartungswert der Fahrzeit. Dies ist ein Punktschätzer. Ein Intervall für den Erwartungswert der Fahrzeit ist ein Konfidenzintervall. Will aber nur wissen, wie lange er bei der nächsten Fahrt unterwegs ist, so wird er eine Prognose bestimmen. Ein Prognoseintervall sagt ihm, zwischen welchen Grenzen die nächste Fahrzeit liegen wird. Dieses Intervall wird er wie auch das Konfidenzintervall mit einer Sicherheit versehen. Schaut er weiter in die Zukunft, so wird er ein Intervall bestimmen, in dem ein vorgegebener Anteil der Fahrzeiten liegt. In Abhängigkeit von der Fragestellungen wird er unterschiedliche Intervalle aufstellen. □

Wir gehen im Folgenden davon aus, dass eine Stichprobe $x_1, \ldots, x_n$ aus einer Grundgesamtheit vorliegt. Dabei wird es sich in der Regel um eine Zufallsstichprobe handeln, die mit Zurücklegen gezogen wurde. Wir beobachten in diesem Fall also die Realisationen $x_1, \ldots, x_n$ der unabhängigen, identisch mit dem Parameter θ verteilten Zufallsvariablen $X_1, \ldots, X_n$. Da wir im Rahmen der Punktschätzung einen Wert für den unbekannten Parameter suchen, fassen wir die Beobachtungen $x_1, \ldots, x_n$ zu einem Wert $g(x_1, \ldots, x_n)$ zusammen. Dieser Wert heißt Schätzwert $t = g(x_1, \ldots, x_n)$. Bei der Intervallschätzung bilden wir $u = g_1(x, \ldots, x_n)$ und $o = g_2(x, \ldots, x_n)$, wobei u die Untergrenze und o die Obergrenze des Intervalls bildet. Das Intervall ist also $[u, o]$. Da die Beobachtungen $x_1, \ldots, x_n$ Realisationen der unabhängigen und identisch verteilten Zufallsvariablen $X_1, \ldots, X_n$ sind, ist t die Realisation der Stichprobenfunktion $T = g(X_1, \ldots, X_n)$. Diese bezeichnen wir im Rahmen der Schätztheorie als **Schätzfunktion**. In der Regel stehen viele Schätzfunktionen zur Schätzung eines Parameters zur Verfügung. Es stellt sich die Frage, welche dieser Schätzfunktionen man verwenden soll. Um dies entscheiden zu können, benötigt man Kriterien, mit denen man Schätzfunktionen beurteilen kann. Mit diesen Kriterien werden wir uns im Kapitel 13.1 beschäftigen. Außerdem werden wir im Kapitel 13.2 Konstruktionsprinzipien betrachten, mit denen man geeignete Schätzfunktionen finden kann. Auch bei einem Konfidenzintervall sind die Grenzen u und o als Funktionen der Beobachtungen $x_1, \ldots, x_n$ Realisationen der Zufallsvariablen $X_1, \ldots, X_n$. Welche Konsequenzen dies für die Interpretation eines Konfidenzintervalls hat, werden wir im Kapitel 13.4.1 aufzeigen. Außerdem werden wir dort zeigen, wie man die Grenzen eines Konfidenzintervalls gewinnt. Anschließend werden wir zeigen, wie man Prognose- und Toleranzintervalle gewinnen kann.

13.1 Eigenschaften von Schätzfunktionen

In der Praxis zieht man eine Stichprobe $x_1, \ldots, x_n$ und bestimmt aus dieser einen Schätzwert $t = g(x_1, \ldots, x_n)$. Der Schätzwert t ist die Realisation der Schätzfunktion T. Da der Wert des Parameters θ unbekannt ist, wissen wir nicht, ob sich der Wert t von T in der Nähe des Wertes des unbekannten Parameters θ befindet. Da wir im Einzelfall nicht wissen, wie nah die Realisation am unbekannten Wert von θ ist, müssen wir durch die Wahl der Schätzfunktion sicherstellen, dass die Wahrscheinlichkeit, dass t stark von θ abweicht gering ist. Wir schauen uns also die Verteilung der Schätzfunktion T an und überprüfen, ob sie bestimmte Eigenschaften besitzt. Zum einen sollte der Wert des Parameters θ das Zentrum der verteilung von T bilden. Hierdurch ist sichergestellt, dass wir mit der Schätzung nicht systematisch daneben liegen. Außerdem sollte die Verteilung von T eine kleine Streuung besitzen. Dann können wir uns sicher sein, dass die Realisation der Schätzfunktion in der Nähe des wahren Wertes des Parameters liegt. Mit diesen beiden Eigenschaften einer Schätzfunktion werden wir uns im Folgenden beschäftigen. Außerdem sollte eine geeignete Schätzfunktion mit wachsendem Stichprobenumfang immer besser werden.

13.1.1 Erwartungstreue

Sind die Zufallsvariablen $X_1, \ldots, X_n$ identisch verteilt mit $E(X_i) = \mu$ für $i = 1, \ldots, n$, so gilt $E(\bar{X}) = \mu$. Wollen wir also den unbekannten Parameter μ auf Basis einer Zufallsstichprobe schätzen, so ist dieser Parameter μ das Zentrum der Verteilung von $\bar{X}$. Man sagt auch, dass $\bar{X}$ eine erwartungstreue Schätzfunktion für μ ist.

Definition 13.1
Eine Schätzfunktion T heißt **erwartungstreu** für den Parameter θ, wenn für alle Werte von θ gilt:

$$\boxed{E(T) = \theta}$$

Man nennt eine erwartungstreue Schätzfunktion auch **unverzerrt**. Im Englischen spricht man von einem **unbiased estimator**.

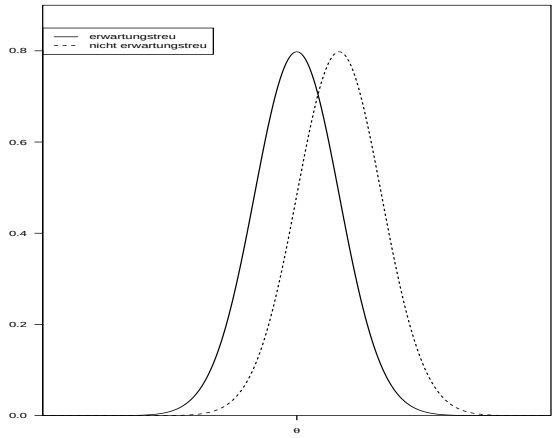
Beispiel 117
Sind $X_1, \ldots, X_n$ unabhängige, identisch mit Parameter p bernoulliverteilte Zufallsvariablen, dann ist $\hat{p} = \bar{X}$ eine erwartungstreue Schätzfunktion für p. Dies haben wir in Gleichung (12.5) auf Seite 347 gezeigt. Im Falle ei-

ner Zufallsstichprobe ist die relative Häufigkeit also eine erwartungstreue Schätzfunktion der Wahrscheinlichkeit des interessierenden Ereignisses.

□

Ist eine Schätzfunktion T nicht erwartungstreu, so streuen die Realisationen von T um den falschen Wert. In Abbildung 13.1 ist die Dichtefunktion einer Schätzfunktion T zu finden, die für den Parameter θ erwartungstreu ist. Außerdem ist die Dichtefunktion einer Schätzfunktion S eingezeichnet, die nicht erwartungstreu ist.

Abbildung 13.1: Dichtefunktion einer erwartungstreuen Schätzfunktion T und einer nicht erwartungstreuen Schätzfunktion S



Wie das folgende Beispiel zeigt, ist nicht jede Schätzfunktion erwartungstreu.

Beispiel 118
Die Zufallsvariablen $X_1, \ldots, X_n$ seien unabhängig und identisch verteilt mit $E(X_i) = \mu$ und $Var(X_i) = \sigma^2$ für $i = 1, \ldots, n$. Die Schätzfunktion

$$D^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

ist nicht erwartungstreu für σ^2. Dies sieht man folgendermaßen:
Wegen

$$Var(X_i) = E(X_i^2) - E(X_i)^2$$

gilt

$$E(X_i^2) = Var(X_i) + E(X_i)^2 = \sigma^2 + \mu^2$$

Wegen

$$Var(\bar{X}) = E(\bar{X}^2) - E(\bar{X})^2$$

gilt

$$E(\bar{X}^2) = Var(\bar{X}) + E(\bar{X})^2 = \frac{\sigma^2}{n} + \mu^2$$

Also folgt

$$\begin{aligned} D^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - 2\,\bar{X}\sum_{i=1}^{n} X_i + n\,\bar{X}^2\right) \\ &= \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - 2\,\bar{X}\,n\,\bar{X} + n\,\bar{X}^2\right) = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - n\,\bar{X}^2\right) \end{aligned}$$

Somit folgt

$$\begin{aligned} E\left(D^2\right) &= E\left(\frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - n\,\bar{X}^2\right)\right) = \frac{1}{n}\left(\sum_{i=1}^{n} E\left(X_i^2\right) - n\,E\left(\bar{X}^2\right)\right) \\ &= \frac{1}{n}\left(n\,\sigma^2 + n\,\mu^2 - n(\frac{\sigma^2}{n} + \mu^2)\right) = \frac{1}{n}\left(n\,\sigma^2 + n\,\mu^2 - \sigma^2 - n\mu^2\right) \\ &= \frac{n-1}{n}\,\sigma^2 \end{aligned}$$

□

Ist eine Schätzfunktion T nicht erwartungstreu für den Parameter θ, so verfehlt die Schätzfunktion systematisch den Wert des Parameters. Die Größe

$$\boxed{bias(T,\theta) = E(T) - \theta} \tag{13.1}$$

heißt **Bias** oder **Verzerrung** der Schätzfunktion. Ist der Bias gleich 0, so ist die Schätzfunktion erwartungstreu. Ist der Bias positiv, so überschätzt die Schätzfunktion den Parameter, ist er negativ, so unterschätzt sie ihn.

Beispiel 118 (fortgesetzt)
Es gilt

$$bias(D^2, \sigma^2) = \frac{n-1}{n}\,\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}$$

Somit unterschätzt D^2 den Wert von σ^2.

Aus D^2 können wir eine erwartungstreue Schätzfunktion für σ^2 gewinnen. Es gilt

$$E\left(\frac{n}{n-1}\,D^2\right) = \sigma^2$$

Es gilt

$$\frac{n}{n-1}\,D^2 = S^2$$

mit

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2.$$

Dies sieht man folgendermaßen:

$$\frac{n}{n-1}\,D^2 = \frac{n}{n-1}\,\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = S^2.$$

Also ist S^2 erwartungstreu für σ^2. Dies erklärt warum wir bei der Stichprobenvarianz durch $n-1$ und nicht durch n dividieren. □

Definition 13.2
Eine Schätzfunktion T heißt asymptotisch erwartungstreu für den Parameter θ, wenn für alle Werte von θ gilt:

$$\boxed{\lim_{n\to\infty} E(T) = \theta}$$

Beispiel 118 (fortgesetzt)
Für die Schätzungsfunktion D^2 gilt:

$$\lim_{n\to\infty} E\left(D^2\right) = \lim_{n\to\infty}\frac{n-1}{n}\,\sigma^2 = \sigma^2\lim_{n\to\infty}\frac{n-1}{n} = \sigma^2\lim_{n\to\infty}\frac{1-1/n}{1} = \sigma^2$$

□

Gilt $E(T) = \theta$, dann gilt auch $\lim_{n\to\infty} E(T) = \theta$. Also ist jede erwartungstreue Schätzfunktion auch asymptotisch erwartungstreu. Die mittlere quadratische Abweichung D^2 zeigt, dass nicht jede asymptotisch erwartungstreue Schätzfunktion auch erwartungstreu ist.

Beispiel 119
Die Zufallsvariablen $X_1, \ldots, X_n$ seien unabhängig, identisch auf $[0, b]$ gleichverteilt. Als Schätzfunktion für b wählen wir das Maximum W der Beobachtungen. Im Beispiel 113 auf Seite 358 haben wir gesehen, dass gilt:

$$f_W(w) = \begin{cases} \frac{n\,w^{n-1}}{b^n} & \text{für} \quad 0 \le w \le 1 \\ 0 & sonst \end{cases}$$

Also gilt

$$\begin{aligned} E(W) &= \int_0^b w \, \frac{n\, w^{n-1}}{b^n} \, dw = \frac{n}{b^n} \int_0^b w^n \, dw = \frac{n}{b^n} \left[\frac{w^{n+1}}{n+1} \right]_0^b = \frac{n}{b^n} \, \frac{b^{n+1}}{n+1} \\ &= \frac{n}{n+1} \, b \end{aligned}$$

Das Maximum ist also keine erwartungstreue Schätzfunktion für b. Der Bias ist

$$bias(W, b) = b - \frac{n}{n+1} \, b = -\frac{1}{n+1} \, b \, .$$

Somit unterschätzt W den Parameter b. Es ist aber eine asymptotisch erwartungstreue Schätzfunktion für b, da gilt

$$\lim_{n\to\infty} \frac{n}{n+1} \, b = \lim_{n\to\infty} \frac{b}{1+1/n} = \frac{\lim\limits_{n\to\infty} b}{\lim\limits_{n\to\infty} (1+1/n)} = b$$

Wegen

$$E(W) = \frac{n}{n+1} \, b$$

gilt

$$E\left(\frac{n+1}{n} \, W \right) = b \, .$$

Also ist $(n+1)/n\, W$ eine erwartungstreue Schätzfunktion für b. □

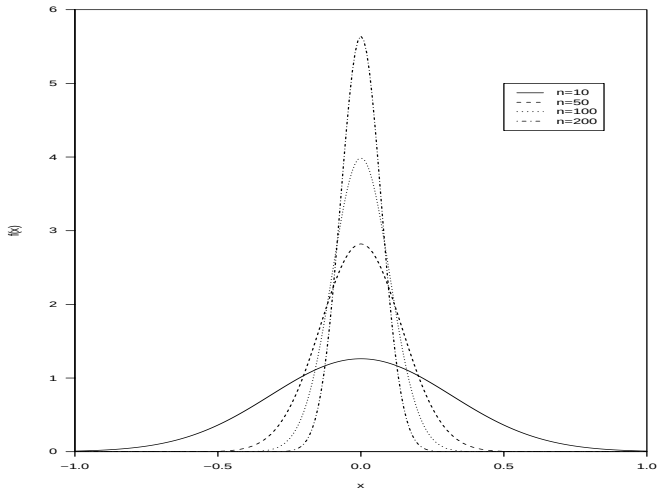
Die Erwartungstreue einer Schätzfunktion stellt sicher, dass das Zentrum der Verteilung der Schätzfunktion mit dem Wert des Parameters zusammenfällt. Bei einer asymptotisch erwartungstreuen Schätzfunktion gilt dies zumindest für große Stichprobenumfänge. Eine erwartungstreue Schätzfunktion nützt einem aber nichts, wenn die Stichprobe verzerrt ist. In diesem Fall ist $\bar{X}$ eine verzerrte Schätzung von μ.

13.1.2 Konsistenz

Da eine Schätzfunktion eine Zufallsvariable ist, kann für eine konkrete Stichprobe der Wert der Schätzfunktion weit entfernt vom wahren Wert des Parameters sein. Viele Schätzfunktionen zeichnen sich aber dadurch aus, dass sich ihre Verteilung mit wachsendem Stichprobenumfang immer stärker um den wahren Wert des Parameters konzentriert. In Abbildung 13.2 sind die Dichtefunktionen von $\bar{X}$ von Zufallsstichproben vom Umfang $n = 10, 50, 100, 200$

aus einer Standardnormalverteilung gezeichnet. Wir sehen, dass die Verteilung von $\bar{X}$ sich immer stärker um den Erwartungswert 0 konzentriert. Da sich die Verteilung von $\bar{X}$ mit wachsendem Stichprobenumfang immer stärker um den wahren Wert des Parameters konzentriert, treten Beobachtungen, die nicht in der Nähe des wahren Wertes des Parameters liegen, für große Stichprobenumfänge selten auf. Ein beobachteter Wert von $\bar{X}$ wird also ziemlich sicher in der Nähe des Wertes des Parameters liegen, wenn der Stichprobenumfang groß ist.

Abbildung 13.2: Dichtefunktion von $\bar{X}$ bei Stichproben aus der Standardnormalverteilung für $n = 10, 50, 100, 200$



Wie können wir diese Eigenschaft einer Schätzfunktion formalisieren? Wir geben eine Zahl $\epsilon > 0$ vor und betrachten die Menge aller Punkte, die von θ weniger als ϵ entfernt sind:

$$\{t|\theta - \epsilon < t < \theta + \epsilon\} = \{t||t - \theta| < \epsilon\}$$

Wir bestimmen die Wahrscheinlichkeit, dass T Werte im Intervall $(\theta - \epsilon < t < \theta + \epsilon)$ annimmt:

$$P(\theta - \epsilon < T < \theta + \epsilon)$$

Wenn sich die Verteilung von T immer stärker um θ konzentriert, so sollte die Wahrscheinlichkeit, dass T Werte im Intervall $(\theta - \epsilon < t < \theta + \epsilon)$ annimmt, mit wachsendem Stichprobenumfang n immer größer werden. Diese Eigenschaft ist nicht leicht zu beweisen. Deshalb wählt man das in der folgenden Definition angesprochene Kriterium:

Definition 13.3
Eine Schätzfunktion T heißt **schwach konsistent** für den Parameter θ, wenn für jedes $\epsilon > 0$ gilt:

$$\lim_{n\to\infty} P\left(|T-\theta| < \epsilon\right) = 1$$

Der Nachweis der schwachen Konsistenz ist nicht sehr einfach. Eine andere Art der Konsistenz ist in der Regel aber leicht nachzuweisen. Hierzu benötigen wir den mittleren quadratischen Fehler. Dieser ist gleich der erwarteten quadrierten Abweichung der Schätzfunktion T vom Parameter θ.

Definition 13.4
Der **mittlere quadratische Fehler** einer Schätzfunktion T bezüglich des Parameters θ ist definiert durch

$$MSE(T,\theta) = E\left[(T-\theta)^2\right]$$

Dabei steht MSE für *mean squared error*. Mit Hilfe des mittleren quadratischen Fehlers erhalten wir folgende Definition der Konsistenz:

Definition 13.5
Eine Schätzfunktion T heißt eine im **quadratischen Mittel konsistente** Schätzfunktion für den Parameter θ, wenn gilt:

$$\lim_{n\to\infty} MSE(T,\theta) = 0$$

Die Konsistenz im quadratischen Mittel kann man sehr schön interpretieren, wenn man den mittleren quadratischen Fehler folgendermaßen umformt:

$$MSE(T,\theta) = Var(T) + \left[(E(T)-\theta)^2\right] \tag{13.2}$$

Der mittlere quadratische Fehler ist also die Summe aus der Varianz $Var(T)$ und dem quadrierten Bias $bias(T,\theta)$ der Schätzfunktion. Bei einer im quadratischen Mittel konsistenten Schätzfunktion verschwindet also mit wachsendem n die Varianz und der Bias. Die Verteilung von T konzentriert sich mit wachsendem Stichprobenumfang n immer mehr um θ.

Beispiel 120
Für $\bar{X}$ gilt $E(\bar{X}) = \mu$ und $Var(\bar{X}) = \sigma^2/n$. Also ist $\bar{X}$ eine im quadratischen Mittel konsistente Schätzfunktion für E(X). □

Um die Gültigkeit von Beziehung (13.2) zeigen zu können, benötigt man folgende Beziehung:

$$\boxed{E(Y^2) = Var(Y) + E(Y)^2} \tag{13.3}$$

Setzen wir in Gleichung (13.3) $Y = T - \theta$, so gilt

$$E((T-\theta)^2) = Var(T-\theta) + E(T-\theta)^2 = Var(T) + (E(T) - \theta)^2 \,. \tag{13.4}$$

Da auf der linken Seite von Gleichung (13.4) $MSE(T, \theta)$ steht, ist die Beziehung (13.2) erfüllt.
Man kann zeigen, dass eine im quadratischen Mittel konsistente Schätzfunktion auch schwach konsistent ist. In Statistik I haben wir die Markow-Ungleichung hergeleitet. Ist Y eine nichtnegative Zufallsvariable und a eine positive reelle Zahl, so gilt

$$P(Y \geq a) \leq \frac{E(Y)}{a} \,. \tag{13.5}$$

Sei nun T eine im quadratischen Mittel konsistente Schätzfunktion des Parameters θ. Wir setzen

$$Y = (T-\theta)^2$$

und

$$a = \epsilon^2$$

Dann gilt aufgrund auf Grund der Markow-Ungleichung

$$P((T-\theta)^2 \geq \epsilon^2) \leq \frac{E[(T-\theta)^2]}{\epsilon^2}$$

und somit

$$P(|T-\theta| \geq \epsilon) \leq \frac{MSE(T,\theta)}{\epsilon^2}$$

Also gilt:

$$P(|T-\theta| < \epsilon) \geq 1 - \frac{MSE(T,\theta)}{\epsilon^2}$$

Da T eine im quadratischen Mittel konsistente Schätzfunktion des Parameters θ ist, gilt

$$\lim_{n\to\infty} MSE(T,\theta) = 0$$

und somit folgt:

$$\lim_{n\to\infty} P(|T-\theta| < \epsilon) \geq 1 - \lim_{n\to\infty} \frac{MSE(T,\theta)}{\epsilon^2}$$

Da ein Wahrscheinlichkeitsmaß kleiner oder gleich 1 ist, folgt

$$\lim_{n\to\infty} P(|T-\theta|<\epsilon) = 1$$

Eine im quadratischen Mittel konsistente Schätzfunktion ist also auch schwach konsistent.

Beispiel 121
Sind $X_1,\ldots,X_n$ unabhängige, identisch mit Parameter p bernoulliverteilte Zufallsvariablen, dann ist der Erwartungswert von $\hat{p}=\bar{X}$ gleich p und die Varianz gleich $p(1-p)/n$. Im Falle einer Zufallsstichprobe ist die relative Häufigkeit also eine konsistente Schätzfunktion der Wahrscheinlichkeit p des interessierenden Ereignisses. □

Beispiel 119 (fortgesetzt von Seite 374)
Die Zufallsvariablen $X_1,\ldots,X_n$ seien unabhängig, identisch auf $[0,b]$ gleichverteilt. Wir wissen, dass $(n+1)/nW$ eine erwartungstreue Schätzfunktion für b ist. Wir bestimmen zunächst die Varianz von W. Es gilt

$$\begin{aligned} E(W^2) &= \int_0^b w^2\,\frac{n\,w^{n-1}}{b^n}\,dw = \frac{n}{b^n}\int_0^b w^{n+1}\,dw = \frac{n}{b^n}\left[\frac{w^{n+2}}{n+2}\right]_0^b \\ &= \frac{n}{b^n}\,\frac{b^{n+2}}{n+2} = \frac{n}{n+2}\,b^2 \end{aligned}$$

Also gilt

$$Var(W) = E(W^2) - E(W)^2 = \frac{n}{n+2}\,b^2 - \frac{n^2}{(n+1)^2}\,b^2 = \frac{n}{(n+2)(n+1)^2}\,b^2$$

Somit gilt

$$\begin{aligned} Var\left(\frac{n+1}{n}W\right) &= \frac{(n+1)^2}{n^2}\,Var(W) = \frac{(n+1)^2}{n^2}\,\frac{n}{(n+2)(n+1)^2}\,b^2 \\ &= \frac{1}{n(n+2)}\,b^2 \end{aligned}$$

Also sind W und $(n+1)/n\,W$ konsistente Schätzfunktionen für b, wobei $(n+1)/n\,W$ erwartungstreu ist. □

13.1.3 Effizienz

Bisher haben wir noch nicht die Frage beantwortet, welche von mehreren Schätzfunktionen man wählen soll, wenn man einen Parameter schätzen will.

Es liegt nahe, die Schätzfunktion zu wählen, die den kleinsten mittleren quadratischen Fehler besitzt. Beschränkt man sich auf die Klasse der erwartungstreuen Schätzfunktionen, so sucht man die Schätzfunktion mit der kleineren Varianz.

Definition 13.6
Seien T_1 und T_2 zwei erwartungstreue Schätzfunktionen des Parameters θ. Die Schätzfunktion T_1 heißt effizienter als die Schätzfunktion T_2, wenn gilt

$$Var(T_1) < Var(T_2)$$

Beispiel 122
Man kann zeigen, dass der Median eine erwartungstreue Schätzfunktion für μ bei Normalverteilung ist. Wir wollen die Varianzen der Stichprobenfunktionen $\bar{X}$ und $X_{0.5}$ einer Zufallsstichprobe vom Umfang $n = 10$ aus einer Standardnormalverteilung vergleichen. Hier gilt $Var(\bar{X}) = 0.1$. Die Varianz des Medians schätzen wir mit einer Simulation. Hierzu ziehen wir $B = 100000$ Stichproben vom Umfang $n = 10$ aus der Standardnormalverteilung und bestimmen für jede dieser Stichproben den Median $X_{0.5}$. Anschließend bestimmen wir die Varianz dieser 100000 Mediane. Wir erhalten den Wert 0.139. Anscheinend ist die Varianz des Medians bei Standardnormalverteilung größer als die Varianz des Mittelwerts. Um zu sehen, ob dies auch für andere Stichprobenumfänge gilt, bestimmen wir die Varianz des Median für $n = 20, 50, 100$. Tabelle 13.1 zeigt das Verhältnis der Varianz des Mittelwerts zur geschätzten Varianz des Medians für diese Stichprobenumfänge.

Tabelle 13.1: Verhältnis der Varianz des Mittelwerts zur geschätzten Varianz des Medians bei Normalverteilung

n	10	20	50	100
$Var(\bar{X})/\widehat{Var}(X_{0.5})$	0.720	0.674	0.640	0.638

Für große Werte von n gilt

$$\frac{Var(\bar{X})}{Var(X_{0.5})} = \frac{2}{\pi} = 0.637$$

Somit ist der Mittelwert eine effizientere Schätzung von μ bei Normalverteilung als der Median $X_{0.5}$. □

Beispiel 119 (fortgesetzt von Seite 379)
Sei $W = \max\{X_1, \ldots, X_n\}$. Wir wissen, dass $T_1 = (n+1)/n\, W$ eine erwartungstreue Schätzfunktion für b ist. Wir betrachten nun noch die Schätzfunktion $T_2 = 2\,\bar{X}$. Wegen $E(X_i) = b/2$ gilt

$$E(T_2) = E(2\,\bar{X}) = 2\,E(\bar{X}) = 2\frac{b}{2} = b$$

Also ist auch T_2 erwartungstreu für b. Wir haben gezeigt, dass gilt

$$Var(T_1) = \frac{1}{n(n+2)}\,b^2\,.$$

Wegen $Var(X_i) = b^2/12$ gilt

$$Var(T_2) = Var(2\,\bar{X}) = 4\,Var(\bar{X}) = 4\,\frac{b^2}{12\,n} = \frac{b^2}{3\,n}\,.$$

Für $n > 1$gilt $n^2 + 2n > 3n$. Dies sieht man folgendermaßen:

$$n > 1 \Longleftrightarrow n^2 > n \Longleftrightarrow n^2 + 2n > n + 2n \Longleftrightarrow n^2 + 2n > 3n$$

Also ist T_1 effizienter als T_2. □

13.2 Konstruktionsprinzipien

Bisher sind wir davon ausgegangen, dass die Schätzfunktionen gegeben sind und haben deren Eigenschaften untersucht. Es gibt aber auch Möglichkeiten, systematisch Schätzfunktionen zu finden. Von diesen wollen wir uns zwei Klassen anschauen. Bei der ersten Klasse benötigen wir nur Informationen über den Erwartungswert und die Varianz der Zufallsvariablen, während wir beim zweiten Verfahren die Verteilung kennen müssen.

13.2.1 Momentenschätzer

Das j-te Moment μ_j einer Zufallsvariablen X ist definiert durch

$$\mu_j \;=\; E(X^j) \tag{13.6}$$

Das erste Moment ist also $\mu_1 = E(X)$ und das zweite Moment $\mu_2 = E(X^2)$.

Beispiel 120
Bei der Poissonverteilung gilt

$$\mu_1 = E(X) = \lambda\,.$$

□

Das theoretische Moment μ_j schätzen wir durch das empirische Moment

$$\hat{\mu}_j \;=\; \frac{1}{n}\sum_{i=1}^{n} X_i^j$$

Speziell gilt $\hat{\mu}_1 = \bar{X}$ und $\hat{\mu}_2 = \overline{X^2}$.
In der Regel hängen die Momente von den Parametern ab. Lösen wir diese Gleichungen nach den Parametern auf und ersetzen die theoretischen Momente durch die empirischen Momente, so erhalten wir die sogenannten **Momentenschätzer** der Parameter.

Beispiel 120 (fortgesetzt)
Es gilt

$$\lambda = \mu_1 \,. \tag{13.7}$$

Wir ersetzen in Gleichung (13.7) μ_1 durch $\bar{X}$ und erhalten den Momentenschätzer $\hat{\lambda} = \bar{X}$. □

Beispiel 121
Bei Normalverteilung gilt $E(X) = \mu$ und $E(X^2) = \mu^2 + \sigma^2$. Die zweite Beziehung gilt wegen $E(X^2) = Var(x) + E(X)^2$.
Wir sehen, dass die Parameter μ und σ^2 von den ersten beiden Momenten $E(X)$ und $E(X^2)$ abhängen. Es gilt $\mu = \mu_1$ und $\sigma^2 = \mu_2 - \mu_1^2$. Wir ersetzen μ_1 durch $\bar{X}$ und μ_2 durch $\overline{X^2}$ und erhalten wir die Momentenschätzer

$$\hat{\mu} = \bar{X}$$

und

$$\widehat{\sigma^2} = \overline{X^2} - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 = D^2$$

Der Momentenschätzer für σ^2 ist nicht erwartungstreu. □

Wir sehen an diesem Beispiel, dass Momentenschätzer nicht immer erwartungstreu sind. Sie müssen aber nicht einmal immer sinnvoll Ergebnisse liefern.

Beispiel 122
Die Wartezeit auf den Bus sei gleichverteilt auf $[0, b]$. Es gilt $E(X) = b/2$ und somit auch $E(\bar{X}) = b/2$. Also ist der Momentenschätzer von b gleich $2\,\bar{X}$. Hier sind die Werte in einer Zufallsstichprobe vom Umfang $n = 5$:

```
3 9 6 19 8
```

Der Mittelwert beträgt 9. Also ist der Momentenschätzer von b gleich 18. Dies ist aber kein sinnvoller Schätzer, da der größte Wert in der Stichprobe 19 beträgt. Die Wartezeit muss also mindestens 19 Minuten betragen. □

13.2.2 Die Maximum-Likelihood-Methode

Ein Verfahren zur Gewinnung einer geeigneten Schätzfunktion ist die **Maximum-Likelihood-Methode**, die wir mit **M-L-Methode** abkürzen. Wir betrachten zunächst eine diskrete Zufallsvariable.

Beispiel 126
Eine Urne enthält 5 Kugeln, wobei es zwei mögliche Zusammensetzungen der Urne gibt:

Zusammensetzung I:	4 schwarze Kugeln und 1 weisse Kugel
Zusammensetzung II:	2 schwarze Kugeln und 3 weisse Kugeln

Auf dem Tisch steht eine Urne. Wir wissen nicht, welche der beiden Zusammensetzungen in ihr vorliegt. Wir dürfen aber eine Kugel ziehen. Die gezogene Kugel sei weiß. Für welche Zusammensetzung der Urne spricht dieses Ergebnis? Bei der ersten Zusammensetzung der Urne beträgt die Wahrscheinlichkeit 0.2, eine weiße Kugel zu ziehen, während diese Wahrscheinlichkeit bei der zweiten Zusammensetzung 0.6 beträgt. Also ist es wahrscheinlicher, aus der zweiten Zusammensetzung eine weiße Kugel zu ziehen. Somit ist es viel plausibler, dass die Urne die zweite Zusammensetzung aufweist, falls die gezogene Kugel weiß ist. Wir entscheiden uns also für die zweite Zusammensetzung der Urne, wenn die gezogene Kugel weiß ist. Ist die gezogene Kugel hingegen schwarz, so entscheiden wir uns für die erste Zusammensetzung. Bei dieser beträgt die Wahrscheinlichkeit 0.8, eine schwarze Kugel zu ziehen, während sie bei der zweiten Zusammensetzung 0.4 beträgt. □

Die im Beispiel verwendete Entscheidungsregel ist die Entscheidungsregel der **Maximum-Likelihood-Methode** (**M-L-Methode**):

> Wir entscheiden uns für den Zustand der Welt, bei dem die beobachtete Stichprobe am wahrscheinlichsten ist.

Beispiel 126 (fortgesetzt von Seite 383)
Versuchen wir nun, diese Vorgehensweise formal darzustellen:
Sei p der Anteil der weißen Kugeln in der Urne. Bei der ersten Zusammensetzung nimmt p den Wert 0.2, bei der zweiten Zusammensetzung den Wert

0.6 an. Unsere Entscheidung über die Zusammensetzung der Urne beruht auf der Farbe der gezogenen Kugel. Wir betrachten die Zufallsvariable X: Anzahl der gezogenen weißen Kugeln. Die Zufallsvariable X kann die Werte 0 und 1 annehmen. Ist die gezogene Kugel weiß, so nimmt sie den Wert 1 an, ansonsten den Wert 0. Die Wahrscheinlichkeitsverteilung von X hängt vom Wert von p ab. Sie ist in Tabelle 13.2 zu finden.

Tabelle 13.2: Wahrscheinlichkeitsverteilung

	p	0.2	0.6
x			
0		0.8	0.4
1		0.2	0.6

Jede Spalte der Tabelle stellt die Wahrscheinlichkeitsverteilung von X in Abhängigkeit von p dar. Eine Zeile der Tabelle ist keine Wahrscheinlichkeitsverteilung. Sie sagt vielmehr aus, wie wahrscheinlich eine Realisation von X unter den verschiedenen Werten des Parameters ist. Die Eintragungen in einer Zeile werden als **Likelihoods** des Parameters gegeben die Beobachtungen bezeichnet. Die gesamte Zeile heißt **Likelihoodfunktion**. In einer Zeile ist der Wert x von X fest. Dies ist die Stichprobe, wenn der Stichprobenumfang n gleich 1 ist. □

Das Maximum-Likelihood-Prinzip besagt nun, denjenigen Wert des Parameters zu wählen, für den die Likelihood am größten ist, für den die Likelihood also ihr Maximum annimmt. Man kann das Maximum-Likelihood-Prinzip auch so beschreiben:

> Wähle den Wert des Parameters, für den die Wahrscheinlichkeit der Stichprobe am größten ist.

Beispiel 126 (fortgesetzt von Seite 383)
Für das Beispiel lautet der M-L-Schätzer:

$$\hat{p}_{ML} = \begin{cases} 0.2 & \text{für} \quad x = 0 \\ 0.6 & \text{für} \quad x = 1 \end{cases}$$

Schauen wir uns an, ob der M-L-Schätzer für das Beispiel erwartungstreu ist. Der M-L-Schätzer $\hat{p}_{ML}$ kann die Werte 0.2 und 0.6 annehmen. Die Wahrscheinlichkeiten dieser Werte hängen davon ab, welchen Wert p annimmt.

Fangen wir mit $p = 0.2$ an. Der M-L-Schätzer $\hat{p}_{ML}$ nimmt den Wert 0.2 an, wenn $x = 0$ ist. Es gilt also

$$P(\hat{p}_{ML} = 0.2) = P(X = 0)$$

Ist $p = 0.2$, so gilt

$$P(\hat{p}_{ML} = 0.2) = 0.8$$

Also gilt, falls $p = 0.2$ ist

$$P(\hat{p}_{ML} = 0.6) = 1 - P(\hat{p}_{ML} = 0.2) = 0.2$$

Ist $p = 0.2$, so gilt

$$E(\hat{p}_{ML}) = 0.2 \cdot 0.8 + 0.6 \cdot 0.2 = 0.28$$

Da $E(\hat{p}_{ML})$ nicht 0.2 ist, ist der M-L-Schätzer $\hat{p}_{ML}$ für p nicht erwartungstreu. □

In der Regel wird man die Entscheidung auf Basis einer Zufallsstichprobe vom Umfang n fällen. Schauen wir uns zunächst an, wie die Entscheidungsregel aussieht, wenn eine Zufallsstichprobe mit Zurücklegen vom Umfang $n = 2$ vorliegt. Sei X_i die Anzahl der beim i-ten Zug gezogenen weißen Kugeln, $i = 1, 2$. Wir bestimmen die folgenden Wahrscheinlichkeiten:

$$P(X_1 = x_1, X_2 = x_2)$$

Liegt die erste Zusammensetzung der Urne vor, so gilt:

$$\begin{aligned}
P(X_1 = 0, X_2 = 0) &= P(X_1 = 0) \cdot P(X_2 = 0) = 0.8 \cdot 0.8 = 0.64 \\
P(X_1 = 0, X_2 = 1) &= P(X_1 = 0) \cdot P(X_2 = 1) = 0.8 \cdot 0.2 = 0.16 \\
P(X_1 = 1, X_2 = 0) &= P(X_1 = 1) \cdot P(X_2 = 0) = 0.2 \cdot 0.8 = 0.16 \\
P(X_1 = 1, X_2 = 1) &= P(X_1 = 1) \cdot P(X_2 = 1) = 0.2 \cdot 0.2 = 0.04
\end{aligned}$$

Liegt die zweite Zusammensetzung der Urne vor, so gilt:

$$\begin{aligned}
P(X_1 = 0, X_2 = 0) &= P(X_1 = 0) \cdot P(X_2 = 0) = 0.4 \cdot 0.4 = 0.16 \\
P(X_1 = 0, X_2 = 1) &= P(X_1 = 0) \cdot P(X_2 = 1) = 0.4 \cdot 0.6 = 0.24 \\
P(X_1 = 1, X_2 = 0) &= P(X_1 = 1) \cdot P(X_2 = 0) = 0.6 \cdot 0.4 = 0.24 \\
P(X_1 = 1, X_2 = 1) &= P(X_1 = 1) \cdot P(X_2 = 1) = 0.6 \cdot 0.6 = 0.36
\end{aligned}$$

Tabelle 13.3 gibt die Wahrscheinlichkeitsverteilung der Stichproben an.

Tabelle 13.3: Wahrscheinlichkeitsverteilung

(x_1, x_2) \ p	0.2	0.6
$(0,0)$	0.64	0.16
$(0,1)$	0.16	0.24
$(1,0)$	0.16	0.24
$(1,1)$	0.04	0.36

Sind beide Kugeln schwarz, beobachten wir also $(0,0)$, so entscheiden wir uns aufgrund des M-L-Prinzips für die erste Zusammensetzung der Urne, also für $p = 0.2$. In allen anderen Fällen nehmen wir an, dass der zweite Zustand vorliegt. Es gilt also

$$\hat{p}_{ML} = \begin{cases} 0.2 & \text{für} \quad (x_1 = 0, x_2 = 0) \\ 0.6 & \text{sonst} \end{cases}$$

Wir können nun die M-L-Methode für den diskreten Fall allgemein formulieren. $X_1, \ldots, X_n$ seien unabhängige, identisch verteilte diskrete Zufallsvariablen, deren Verteilung von einem unbekannten Parameter θ abhängt. Wir wollen θ auf der Basis der Realisationen $x_1, \ldots, x_n$ schätzen. Dann ist

$$P(X_1 = x_1, \ldots, X_n = x_n, \theta)$$

die Wahrscheinlichkeit für das Auftreten der Stichprobe $x_1, \ldots, x_n$ in Abhängigkeit von θ. Diese Wahrscheinlichkeit fassen wir bei gegebenen $x_1, \ldots, x_n$ als Funktion von θ auf und nennen sie **Likelihoodfunktion** $L(\theta)$. Es gilt also

$$\boxed{L(\theta) = P(X_1 = x_1, \ldots, X_n = x_n, \theta)}$$

Der **Maximum-Likelihood-Schätzer** $\hat{\theta}$ (**M-L-Schätzer**) ist nun der Wert von θ, für den die Likelihood am größten ist:

$$\boxed{L(\hat{\theta}) = \max_{\theta} L(\theta)}$$

Aus technischen Gründen betrachtet man oft den Logarithmus der Likelihoodfunktion. Man erhält also die sogenannte **Loglikelihoodfunktion**:

$$l(\theta) = \ln L(\theta)$$

Da der Logarithmus eine monotone Transformation ist, nimmt die Loglikelihoodfunktion ihr Maximum an der gleichen Stelle an wie die Likelihoodfunktion. Schauen wir uns das obige Beispiel für den Stichprobenumfang n an. Außerdem schränken wir die möglichen Werte von p nicht von vornherein ein. Es sind also alle Werte von p im Intervall $(0, 1)$ möglich.

Beispiel 127
$X_1, \ldots, X_n$ seien also unabhängige, identisch mit Parameter p bernoulliverteilte Zufallsvariablen. Für $x_i = 0, 1$ gilt also

$$P(X_i = x_i) = p^{x_i} (1-p)^{1-x_i} .$$

Die Likelihood lautet also

$$\begin{aligned} L(p) &= p^{x_1} (1-p)^{1-x_1} \cdots p^{x_n} (1-p)^{1-x_n} \\ &= p^{\sum x_i} (1-p)^{n-\sum x_i} = p^{n\,\bar{x}} (1-p)^{n\,(1-\bar{x})} . \end{aligned}$$

Die Loglikelihood ist:

$$l(p) = n\,\bar{x}\, \ln p + n\,(1-\bar{x})\, \ln (1-p) .$$

Zur Bestimmung des M-L-Schätzers bilden wir die erste Ableitung:

$$\frac{d}{dp}\, l(p) = \frac{n\,\bar{x}}{p} - \frac{n\,(1-\bar{x})}{1-p} = \frac{n(\bar{x}-p)}{p\,(1-p)}$$

Notwendige Bedingung für einen Extremwert ist, dass die erste Ableitung gleich 0 ist. Es muss also gelten

$$\frac{n(\bar{x}-\hat{p})}{\hat{p}\,(1-\hat{p})} = 0$$

Diese Gleichung wird erfüllt von $\hat{p} = \bar{x}$. Wir überprüfen noch die hinreichende Bedingung. Es gilt

$$\frac{d^2}{dp^2}\, l(p) = -\frac{n\,\bar{x}}{p^2} - \frac{n\,(1-\bar{x})}{(1-p)^2} = -\left(\frac{n\,\bar{x}}{p^2} + \frac{n\,(1-\bar{x})}{(1-p)^2} \right)$$

Aus $0 \leq \bar{x} \leq 1$ folgt

$$\frac{d^2}{dp^2}\, l(p) < 0$$

Es handelt sich also wirklich um ein Maximum. Somit ist $\hat{p} = \bar{X}$ der M-L-Schätzer von p. In disem Fall ist der M-L-Schätzer von p erwartungstreu und konsistent. □

Bei stetigen Zufallsvariablen ist die Likelihoodfunktion die gemeinsame Dichtefunktion der Zufallsvariablen $X_1, \ldots, X_n$:

$$\boxed{L(\theta) = f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)}$$

Wir unterstellen in der Regel, dass die Zufallsvariablen $X_1, \ldots, X_n$ unabhängig sind. In diesem Fall ist die gemeinsame Dichtefunktion das Produkt der einzelnen Dichtefunktionen und die Likelihoodfunktion lautet:

$$\boxed{L(\theta) = \prod_{i=1}^{n} f_{X_i}(x_i, \theta)}$$

und für $l(\theta)$ gilt

$$\boxed{l(\theta) = \sum_{i=1}^{n} \ln f_{X_i}(x_i, \theta)}$$

Beispiel 128
$X_1, \ldots, X_n$ seien also unabhängige, identisch mit den Parametern μ und σ^2 normalverteilte Zufallsvariablen. Gesucht ist der M-L-Schätzer von μ. Wir nehmen also an, dass σ^2 bekannt ist. Es gilt

$$f_{X_i}(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}\,.$$

Hieraus folgt

$$\ln f_{X_i}(x_i) = -\ln\sqrt{2\,\pi} - \frac{1}{2}\,\ln\sigma^2 - \frac{1}{2\sigma^2}\,(x_i-\mu)^2\,.$$

Die Loglikelihoodfunktion von μ bei festem σ^2 lautet somit

$$l(\mu) = -n\,\ln\sqrt{2\,\pi} - \frac{n}{2}\,\ln\sigma^2 - \frac{1}{2\sigma^2}\,\sum_{i=1}^{n}\,(x_i-\mu)^2$$

Notwendige Bedingung für einen Extremwert in $\hat{\mu}$ ist, dass die erste Ableitung an der Stelle $\hat{\mu}$ gleich 0 ist. Die erste Ableitung ist gegeben durch

$$\begin{aligned}\frac{d}{d\mu}\, l(\mu) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = \frac{1}{\sigma^2}\,(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\mu) = \frac{1}{\sigma^2}\,(n\,\bar{x} - n\,\mu)\\ &= \frac{n}{\sigma^2}\,(\bar{x}-\mu)\end{aligned}$$

Für $\hat{\mu}$ muss also gelten

$$\frac{n}{\sigma^2}\,(\bar{x}-\hat{\mu}) = 0$$

Hieraus folgt

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

Die zweite Ableitung lautet

$$\frac{d^2}{d\mu^2}\, l(\mu) = -\frac{n}{\sigma^2}$$

Da σ^2 größer als 0 ist, gilt

$$\frac{d^2}{d\mu^2}\, l(\mu) < 0$$

Der Mittelwert $\bar{X}$ ist also der M-L-Schätzer von μ bei Normalverteilung. □

Wir haben gesehen, dass M-L-Schätzer nicht notwendigerweise erwartungstreu sind. Unter bestimmten Bedingungen sind sie asymptotisch erwartungstreu und auch konsistent. Oft sind M-L-Schätzer asymptotisch normalverteilt. Dabei bedeutet asymptotisch normalverteilt, dass der M-L-Schätzer für große Werte von n approximativ normalverteilt ist. Der M-L-Schätzer im Beispiel 127 auf Seite 387 ist asymptotisch normalverteilt, während der M-L-Schätzer im Beispiel 128 auf Seite 388 exakt normalverteilt ist. Schauen wir uns ein Beispiel eines asymptotisch normalverteilten M-L-Schätzers an.

Beispiel 129
Die Zufallsvariablen $X_1, \ldots, X_n$ seien exponentialverteilt mit dem Parameter λ. Für $i = 1, \ldots, n$ gilt also: ,

$$f_{X_i}(x_i) = \begin{cases}\lambda\, e^{-\lambda x_i} & \text{für } x_i > 0\\ 0 & \text{sonst}\end{cases}$$

Es soll des M-L-Schätzer des Parameters λ bestimmt werden. Es gilt

$$\ln f_{X_i}(x_i) = \ln\lambda - \lambda\, x_i\,.$$

Somit lautet die Loglikelihoodfunktion:

$$l(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^{n} X_i = n \ln \lambda - \lambda n \bar{x}$$

Die erste Ableitung ist:

$$\frac{d}{d\lambda} l(\lambda) = \frac{n}{\lambda} - n \bar{x}$$

Der M-L-Schätzer $\hat{\lambda}$ von λ muss also folgende Bedingung erfüllen:

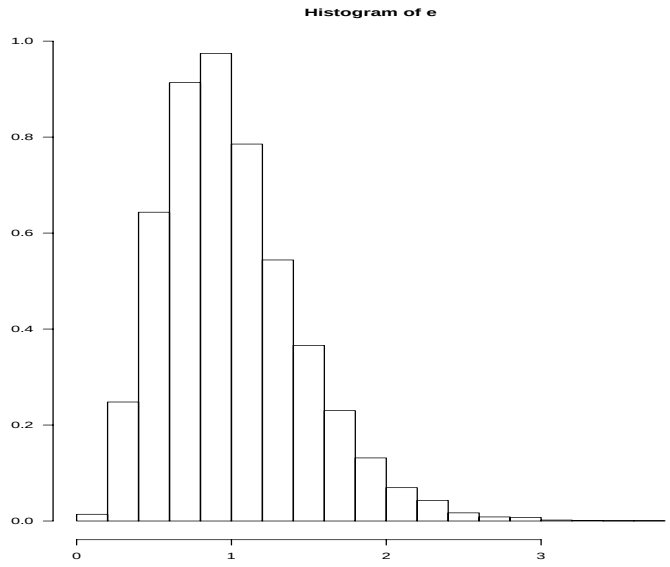
$$\frac{n}{\hat{\lambda}} = n \bar{x}$$

Also gilt

$$\hat{\lambda} = \frac{1}{\bar{x}}$$

Schauen wir uns die Verteilung von $\frac{1}{\bar{X}}$ bei Exponentialverteilung einmal an. Hierzu simulieren wir für $\lambda = 1$ und $n = 5$ und erhalten das Histogramm in Abbildung 13.3.

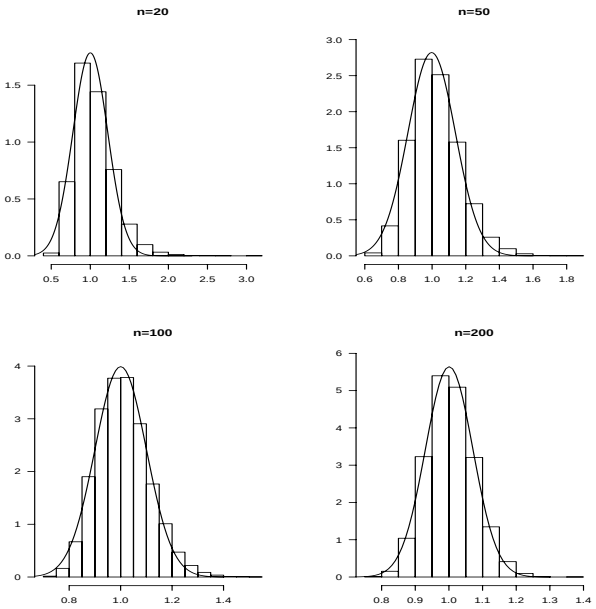
Abbildung 13.3: Mit einer Simulation geschätzte Verteilung des M-L-Schätzers bei Exponentialverteilung mit $\lambda = 1$ für $n = 5$



Wir sehen, dass die Verteilung schief ist. Nun erhöhen wir den Stichprobenumfang. Abbildung 13.4 zeigt die Histogramme für $n = 20, 50, 100, 200$.

Außerdem ist die Dichtefunktion der Normalverteilung eingezeichnet. Wir sehen, dass die Verteilung mit wachsendem n immer mehr der Normalverteilung ähnelt.

Abbildung 13.4: Mit einer Simulation geschätzte Verteilung des M-L-Schätzers bei Exponentialverteilung mit $\lambda = 1$ für $n = 20, 50, 100, 200$



□

13.3 Dichteschätzung

Auf auf Seite 75 haben wir uns mit dem Histogramm beschäftigt. Das Histogramm ist ein Schätzer der Dichtefunktion, der jedoch nicht glatt ist. Rosenblatt (1956) hat als erster sogenannte **Kerndichteschätzer** vorgeschlagen, mit denen wir uns jetzt beschäftigen wollen.
Rosenblatt geht davon aus, dass die Dichtefunktion $f_X(x)$ die Ableitung der Verteilungsfunktion $F_X(x)$ ist.
Es gilt also

$$f_X(x) = \frac{d}{dx} F_X(x) = \lim_{h \to 0} \frac{1}{2h} P(x - h < X < x + h) \tag{13.8}$$

Zu vorgegebenem h können wir $P(x-h<X<x+h)$ durch den Anteil der Beobachtungen schätzen, die in das Intervall $(x-h, x+h)$ fallen:

$$\widehat{P}(x-h<X<x+h) = \frac{1}{n} \cdot (\text{Anzahl von } \; x_1, \ldots, x_n \; \text{ in } (x-h, x+h))$$

Somit erhalten wir als Schätzer der Dichtefunktion an der Stelle x:

$$\hat{f}_X(x) = \frac{1}{2\,h\,n} \cdot (\text{Anzahl von } \; x_1, \ldots, x_n \; \text{ in } (x-h, x+h))$$

Beispiel 130
Gegeben seien die Beobachtungen

$$x_1 = 1.55 \quad x_2 = 1.6 \quad x_3 = 1.9 \quad x_4 = 2 \quad x_5 = 2.$$

Wir wählen $h = 0.25$. Wir wollen die Dichtefunktion an der Stelle $x = 1.7$ schätzen. Hierzu zählen wir die Anzahl der Beobachtungen im Intervall $(1.55, 1.95)$. Es sind drei Beobachtungen.
Also gilt

$$\hat{f}(1.7) = \frac{3}{2 \cdot 0.25 \cdot 5} = 1.$$

Abbildung 13.5 zeigt die Dichteschätzung im Intervall $[1, 3]$.

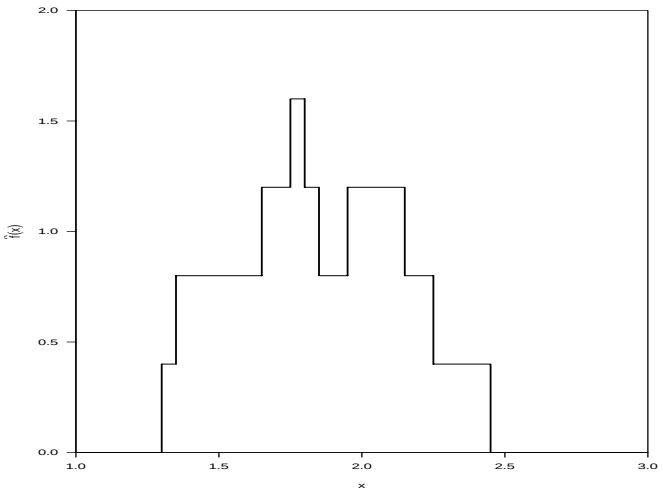


Abbildung 13.5: Dichteschätzung

□

Es gilt

$$-1 < \frac{x - x_i}{h} < 1 \quad \Longleftrightarrow \quad x - h < x_i < x + h$$

Also können wir den Dichteschätzer mit

$$w(x) = \begin{cases} 0.5 & \text{für } -1 < x < 1 \\ 0 & \text{sonst} \end{cases}$$

auch folgendermaßen darstellen:

$$\hat{f}_X(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} w\left(\frac{x - x_i}{h}\right)$$

Mit dieser Darstellung können wir die Schätzung aus einem anderen Blickwinkel betrachten.
Es gilt nämlich auch

$$-1 < \frac{x - x_i}{h} < 1 \quad \Longleftrightarrow \quad x_i - h < x < x_i + h$$

Anstatt also die Anzahl der Beobachtungen $x_1, \ldots, x_n$ im Intervall $(x-h, x+h)$ zu zählen, können wir auch um jede Beobachtung x_i ein Intervall $(x_i - h, x_i + h)$ legen und die Anzahl der Intervalle zählen, in denen x liegt. Jedes der Intervalle $(x_i - h, x_i + h)$ enthält die Punkte, zu denen die Beobachtung x_i einen Beitrag bei der Schätzung der Dichtefunktion liefert.
Abbildung 13.6 veranschaulicht diesen Sachverhalt.

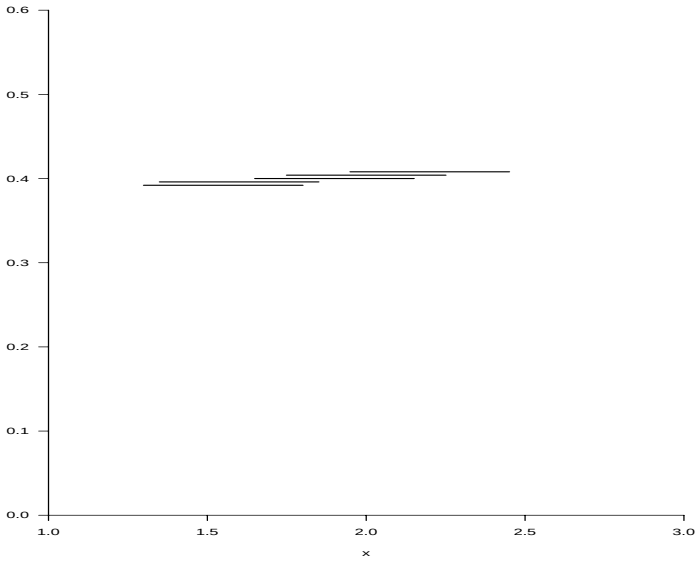


Abbildung 13.6: Veranschaulichung der Dichteschätzung

Der Schätzer ist so konstruiert, dass alle Punkte in der Umgebung einer Beobachtung das gleiche Gewicht erhalten. Eine Beobachtung liefert also für alle Punkte in ihrer Umgebung den gleichen Beitrag zur Dichtefunktion. Dies führt dazu, dass die geschätzte Dichtefunktion nicht glatt ist. Um einen glatteren Verlauf der geschätzten Dichtefunktion zu erhalten, sollte man die Gewichtungsfunktion $w(u)$ so wählen, dass der Beitrag einer Beobachtung zur Dichteschätzung mit wachsendem Abstand von ihr abnimmt. Wählt man dann als Gewichtungsfunktion eine Dichtefunktion, so besitzt auch die Dichteschätzung alle Eigenschaften einer Dichtefunktion. Die Gewichtungsfunktion $w(u)$ heißt auch **Kernfunktion** und der zugehörige Schätzer **Kerndichteschätzer**.
In der Literatur gibt es eine Reihe von Vorschlägen für die Wahl der Kernfunktion. Die klassische Wahl ist der Gauss-Kern:

$$w(t) = \frac{1}{\sqrt{2\pi}} e^{-0.5t^2}$$

Dies ist gerade die Dichtefunktion der Standardnormalverteilung.
Weitere Kernfunktionen sind bei Silverman (1986) und Scott (1992) zu finden.

Beispiel 130 (fortgesetzt von Seite 392)
Schauen wir uns an, was passiert, wenn wir für den obigen Datensatz die Dichtefunktion mit einem Gausskern mit $h = 1.5$ schätzen. Wir legen um jede Beobachtung eine Dichtefunktion der Normalverteilung mit $\sigma = 1.5$ und erhalten Abbildung 13.7.

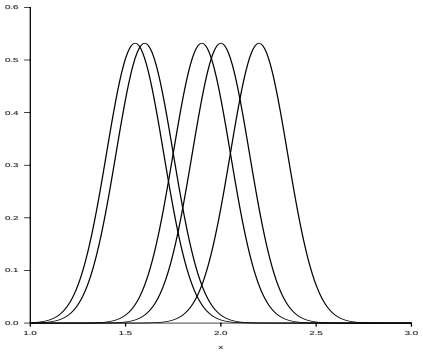


Abbildung 13.7: Gausskerne

Addition der Kernfunktionen liefert die Dichteschätzung in Abbildung 13.8.

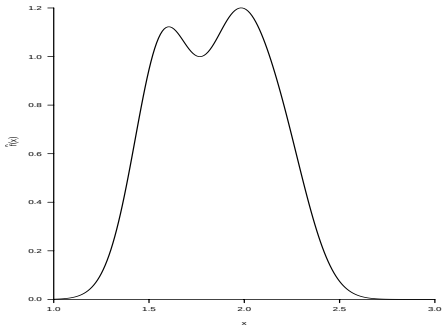


Abbildung 13.8: Dichteschätzung

□

Mit wachsendem h wird die Dichteschätzung immer glatter, dabei gehen aber lokale Informationen verloren. Für $h = 0.2$ erhalten wir fast eine Normalverteilung. Lokale Unterschiede werden für kleine Werte von h gut wiedergegeben, während die Kurve nicht sehr glatt wirkt.
In der Literatur gibt es eine Vielzahl von Vorschlägen für die Wahl von h. So findet man bei Silverman (1986):

$$h = 0.9 \min \left\{ s_x, \frac{IQR}{1.34} \right\} n^{-0.2} \tag{13.9}$$

und bei Scott (1992):

$$h = 1.06 \min(s, \frac{IQR}{1.34}) \, n^{-0.2} \tag{13.10}$$

Dabei s die Standardabweichung und IQR der Interquartilsabstand.

Beispiel 131
Wir betrachten in Tabelle 1.2 auf Seite 17 das Alter der Teilnehmer und bestimmen h mit dem Verfahren von Scott in Gleichung (13.10). Es gilt $s = 4.62$ und $IQR = 5$. Also gilt

$$h = 1.06 \min \left\{ 4.62, \frac{5}{1.34} \right\} 25^{-0.2} = 2.078$$

Abbildung 13.9 zeigt die Dichteschätzung. Sie deutet wie auch das Histogramm in Abbildung 3.18 auf Seite 119 auf eine zweigipflige Verteilung hin.

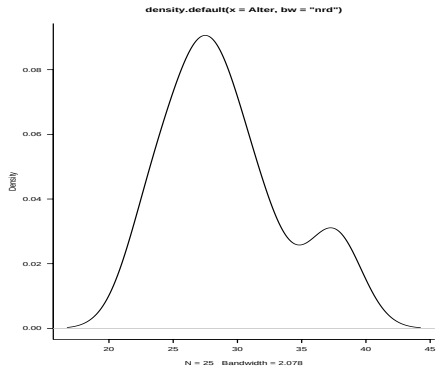


Abbildung 13.9: Dichteschätzung des Merkmals `Alter der Teilnehmer`

□

13.4 Intervallschätzung

Werden die Ergebnisse statistischer Erhebungen veröffentlicht, so werden vielfach Punktschätzer angegeben. So ist beim Umwelt- und Prognose- Institut am 16.01.2004 folgende Aussage zu finden:

> Die durchschnittliche Fahrleistung des Autofahrers liegt seit Jahren stabil bei 12000 Kilometern im Jahr.

Oft wird die Schätzung in Form eines Intervalls angegeben. Dies trägt dem Umstand Rechnung, dass die Schätzung fehlerbehaftet ist. Wir wollen uns im Folgenden unterschiedliche statistische Intervalle anschauen.

13.4.1 Konfidenzintervalle

Da eine Schätzfunktion T eine Zufallsvariable ist, können wir uns für eine spezielle Realisation ziemlich sicher sein, dass sie nicht mit dem wahren Wert des Parameters θ übereinstimmt. Besitzt die Schätzfunktion eine stetige Verteilung, so ist die Wahrscheinlichkeit, den wahren Wert des Parameters zu treffen, sogar 0. Diesem Tatbestand sollte man dadurch Rechnung tragen, dass man den Schätzer mit einer Genauigkeitsangabe versieht. Eine Möglichkeit besteht darin, die Varianz $Var(T)$ des Schätzers anzugeben.

Beispiel 132
Seien $X_1, \ldots, X_n$ unabhängige, identisch mit Erwartungswert $E(X_i) = \mu$ und Varianz $Var(X_i) = \sigma^2$ verteilte Zufallsvariablen. Wir schätzen μ durch $\bar{X}$.

Es gilt:

$$Var(\bar{X}) = \frac{\sigma^2}{n}\,.$$

□

Oft hängt $Var(T)$ von einem oder mehreren unbekannten Parametern ab. Schätzen wir diese und setzen sie in die Formel für $Var(T)$ ein, so erhalten wir die geschätzte Varianz $\widehat{Var}(T)$ von T.

Beispiel 132 (fortgesetzt)
Ist σ^2 unbekannt, so schätzen wir es durch s^2 und erhalten die geschätzte Varianz $\widehat{Var}(\bar{X})$. Es gilt

$$\widehat{Var}(\bar{X}) = \frac{s^2}{n}\,.$$

□

Man nennt $\sqrt{Var(T)}$ auch den Standardfehler von T und schreibt dafür σ_T. Den geschätzten Standardfehler bezeichnen wir entsprechend mit $\hat{\sigma}_T = \sqrt{\widehat{Var}(T)}$.

Beispiel 132 (fortgesetzt)
Der Standardfehler von $\bar{X}$ ist $\sigma/\sqrt{n}$ und der geschätzte Standardfehler $s/\sqrt{n}$.
□

Je größer der Standardfehler ist, desto unsicherer ist die Schätzung. Diese Unsicherheit können wir dadurch ausdrücken, dass wir ein Intervall für den unbekannten Parameter θ angeben, dessen Grenzen vom Standardfehler σ_T bzw. geschätzten Standardfehler $\hat{\sigma}_T$ abhängen. Als Mittelpunkt des Intervalls wählen wir den Wert der Schätzfunktion T. Die Grenzen des Intervalls wählen wir so, dass sie jeweils σ_T bzw. $\hat{\sigma}_T$ von T entfernt sind. Wir erhalten somit das Intervall

$$[T - \sigma_T, T + \sigma_T] \tag{13.11}$$

bzw.

$$[T - \hat{\sigma}_T, T + \hat{\sigma}_T]\,. \tag{13.12}$$

Beispiel 132 (fortgesetzt)
Wir suchen ein Intervall für μ. Dabei gehen wir davon aus, dass σ^2 und somit der Standardfehler $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ bekannt ist. Das Intervall ist

$$\left[\bar{X} - \frac{\sigma}{\sqrt{n}}, \bar{X} + \frac{\sigma}{\sqrt{n}}\right] \tag{13.13}$$

□

Das so gewählte Intervall spiegelt die Ungenauigkeit des Schätzers wider. Je größer σ_T bzw. $\hat{\sigma}_T$ ist, um so breiter ist das Intervall. Unser Ziel ist es, ein Intervall aufzustellen, in dem der Wert des Parameters θ liegt. Wie sicher können wir uns sein, dass dies der Fall ist? Für ein konkretes Intervall gibt es nur zwei Möglichkeiten. Der unbekannte Wert des Parameters liegt in dem Intervall oder er liegt nicht in dem Intervall. Wir wissen nicht, welche der beiden Möglichkeiten zutrifft. Wir können aber die Wahrscheinlichkeit bestimmen, dass wir ein Intervall gefunden haben, das den Wert des Parameters überdeckt. Diese ist für die Intervalle in den Gleichungen (13.11) und (13.12)

$$P(T - \sigma_T \leq \theta \leq T + \sigma_T)$$

bzw.

$$P(T - \hat{\sigma}_T \leq \theta \leq T + \hat{\sigma}_T)$$

Man bezeichnet diese Wahrscheinlichkeit auch als **Konfidenzniveau**. Ist sie groß, so können wir uns ziemlich sicher sein, dass wir ein Intervall gefunden haben, in dem der Wert des Parameters liegt.

Beispiel 132 (fortgesetzt)
Wir betrachten das Intervall in Gleichung (13.13). Um das Konfidenzniveau bestimmen zu können, müssen wir ein spezielles Verteilungsmodell für die X_i unterstellen. Wir unterstellen, dass die X_i mit den Parametern μ und σ^2 normalverteilt sind. In diesem Fall ist $\bar{X}$ normalverteilt mit den Parametern μ und σ^2/n. Dies haben wir in Satz 12.4 auf Seite 346 gezeigt. Also gilt

$$\begin{aligned}
P(\bar{X} - \sigma/\sqrt{n} \leq \mu \leq \bar{X} + \sigma/\sqrt{n}) &= P(-\sigma/\sqrt{n} \leq \mu - \bar{X} \leq \sigma/\sqrt{n}) \\
&= P(-\sigma/\sqrt{n} \leq \bar{X} - \mu \leq \sigma/\sqrt{n}) \\
&= P\left(-1 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1\right) \\
&= \Phi(1) - \Phi(-1) \\
&= 0.6827
\end{aligned}$$

Wir wollen verdeutlichen, wie dieser Wert zu interpretieren ist. Hierzu führen wir eine Simulation durch. Wir unterstellen Standardnormalverteilung und ziehen 20 Stichproben vom Umfang $n = 4$. Für jede dieser Stichproben stellen wir das Konfidenzintervall für μ auf, wobei $\sigma = 1$ bekannt sei. Setzen wir $\sigma = 1$ und $n = 4$ in Gleichung (13.13) auf der vorherigen Seite ein, so erhalten wir das Intervall in Abhängigkeit von $\bar{x}$:

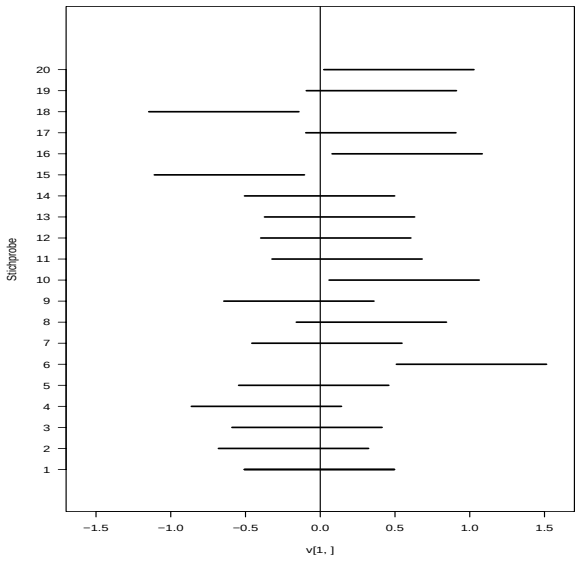
$$\bar{x} - 0.5, \bar{x} + 0.5$$

Nehmen wir an, die erste simulierte Stichprobe lautet

```
0.32412995  0.04917965 -2.67762426  2.28072137
```

Der Mittelwert beträgt -0.006. Das Intervall ist $[-0.506, 0.494]$. Wir kennen den Wert von μ. Dieser ist 0. Das Intervall enthält den Wert von μ. Abbildung 13.10 verdeutlicht dies und zeigt die anderen 19 Konfidenzintervalle.

Abbildung 13.10: 20 Konfidenzintervalle



Wir sehen, dass 14 Konfidenzintervalle den Wert 0 enthalten. □

Wir können die Wahrscheinlichkeit, dass das Intervall den Wert des Parameters überdeckt, dadurch vergrößern bzw. verkleinern, dass wir das Intervall breiter bzw. schmaler machen. Die Grenzen des Intervalls sollten weiterhin vom Standardfehler σ_T bzw. geschätzten Standardfehler $\hat{\sigma}_T$ von T abhängen. Wir setzen an

$$[T - k\,\sigma_T, T + k\,\sigma_T]$$

bzw.

$$[T - k\,\hat{\sigma}_T, T + k\,\hat{\sigma}_T]\ .$$

Für $k = 1$ erhalten wir die Intervalle in Gleichungen (13.11) und (13.12) auf Seite 397. Bisher haben wir das Intervall vorgegeben und das zugehörige

Konfidenzniveau bestimmt. Es ist aber sinnvoller, das Konfidenzniveau vorzugeben und dann das zugehörige Intervall zu bestimmen. Hierbei wählen wir den Wert von k so, dass die Wahrscheinlichkeit gleich α ist, dass das Intervall den Wert des Parameters **nicht** überdeckt. Somit beträgt das Konfidenzniveau $1-\alpha$. Wir nennen das Intervall **Konfidenzintervall**. Der Wert von k hängt vom Konfidenzniveau $1-\alpha$ ab. Schauen wir uns dies für das Beispiel an.

Beispiel 132 (fortgesetzt)
Wir wollen k nun in Abhängigkeit von $1-\alpha$ bestimmen. Es gilt

$$\begin{aligned} P\left(\bar{X} - k\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + k\frac{\sigma}{\sqrt{n}}\right) &= P\left(-k\frac{\sigma}{\sqrt{n}} \leq \mu - \bar{X} \leq k\frac{\sigma}{\sqrt{n}}\right) \\ &= P\left(-k\frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq k\frac{\sigma}{\sqrt{n}}\right) \\ &= P\left(-k \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq k\right) \end{aligned}$$

Da $\bar{X}$ normalverteilt ist mit den Parametern μ und σ^2/n, ist die Größe

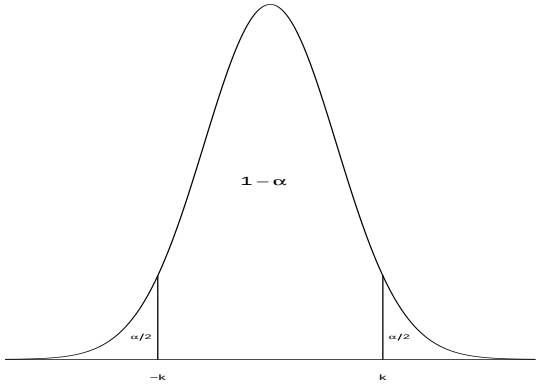
$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

standardnormalverteilt. Die Größe k muss so gewählt werden, dass gilt

$$P(-k \leq Z \leq k) = \Phi(k) - \Phi(-k) = 1-\alpha\,.$$

Die Wahrscheinlichkeit für das Intervall $[-k,k]$ muss also $1-\alpha$ betragen. Außerhalb des Intervalls beträgt die Wahrscheinlichkeit α. Da die Dichtefunktion der Standardnormalverteilung symmetrisch bezüglich 0 ist, muss unterhalb von $-k$ genau so viel Wahrscheinlichkeitsmasse liegen wie über k. Also muss unterhalb von $-k$ die Wahrscheinlichkeitsmasse $\alpha/2$ liegen. Also gilt $-k = z_{\alpha/2}$. Also ist $k = z_{1-\alpha/2}$. Abbildung 13.11 zeigt die Dichtefunktion der Standardnormalverteilung und verdeutlicht den Zusammenhang.

Abbildung 13.11: Veranschaulichung der Konstruktion eines Konfidenzintervalls



Das gesuchte Konfidenzintervall lautet also

$$\boxed{\left[\bar{X} - z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]} \tag{13.14}$$

Man nennt das Intervall in Gleichung (13.14) auch das Konfidenzintervall zum Konfidenzniveau $1-\alpha$ für μ bei Normalverteilung mit bekanntem σ^2. □

Wir wollen uns im Folgenden unterschiedliche Konfidenzintervalle anschauen.

Konfidenzintervall für μ bei Normalverteilung

Wir gehen davon aus, dass die Zufallsvariablen $X_1, \ldots, X_n$ unabhängig und identisch mit den Parametern μ und σ^2 normalverteilt sind. Wir wollen ein Konfidenzintervall für μ aufstellen. Hierbei muss man unterscheiden, ob σ^2 bekannt oder unbekannt ist. Wir betrachten zunächst den Fall mit bekanntem σ^2.

Die Varianz σ^2 ist bekannt

Das Intervall ist in Gleichung (13.14) zu finden. Schauen wir uns dieses Intervall genauer an. Die Länge L des Konfidenzintervalls ist:

$$\boxed{L = 2\, z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}} \tag{13.15}$$

Dies sieht man folgendermaßen:

$$L = \bar{X} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} - (\bar{X} - z_{1-\alpha/2}\, \frac{\sigma}{\sqrt{n}}) = 2\, z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Schauen wir uns Gleichung (13.15) genauer an. Erhöhen wir den Stichprobenumfang n, so wird auch $\sqrt{n}$ größer und somit $1/\sqrt{n}$ kleiner. Mit größerem Stichprobenumfang n wird also das Konfidenzintervall bei festem Konfidenzniveau $1 - \alpha$ kürzer.

Erhöhen wir hingegen das Konfidenzniveau $1 - \alpha$, so wird α und auch $\alpha/2$ kleiner. Wie man Abbildung 13.11 entnehmen kann, wird $z_{1-\alpha/2}$ größer. Also wird das Konfidenzintervall länger. Um eine größere Sicherheit zu erhalten, müssen wir die Länge des Intervalls vergrößern.

Aus Gleichung (13.15) können wir auch herleiten, wie groß n sein muss, damit man bei gegebenem Konfidenzniveau $1 - \alpha$ eine vorgegebene Länge l nicht überschreitet. Es muss also gelten $L \leq l$. Wir setzen die rechte Seite von Gleichung (13.15) in diese Gleichung ein und erhalten:

$$2\, z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \leq l$$

Wir multiplizieren beide Seiten dieser Gleichung mit $\sqrt{n}/l$ und erhalten

$$2\, z_{1-\alpha/2} \frac{\sigma}{l} \leq \sqrt{n}$$

Durch Quadrieren dieser Ungleichung erhalten wir die gesuchte Lösung:

$$\boxed{n \geq \frac{4\, z_{1-\alpha/2}^2 \sigma^2}{l^2}} \tag{13.16}$$

Wollen wir also die Länge des Konfidenzintervalls halbieren, so müssen wir den Stichprobenumfang vervierfachen.

Die Varianz σ^2 ist unbekannt

Schauen wir uns das Konfidenzintervall in Gleichung (13.14) unter praxisrelevanten Gesichtspunkten an. Bei einer Datenanalyse wird σ^2 in der Regel unbekannt sein. Es liegt nahe, σ^2 durch s^2 zu schätzen und diesen Schätzer in Gleichung (13.14) für σ^2 einzusetzen. Das Intervall sieht also folgendermaßen aus:

$$\left[\bar{X} - z_{1-\alpha/2}\,\frac{s}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\frac{s}{\sqrt{n}}\right]. \tag{13.17}$$

Für kleine Stichprobenumfänge gilt aber

$$P\left(\bar{X} - z_{1-\alpha/2}\,\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\,\frac{s}{\sqrt{n}}\right) \neq 1-\alpha\,. \tag{13.18}$$

Eine kleine Simulation zeigt dies. Wir erzeugen 5000 Stichproben vom Umfang 4 aus der Standardnormalverteilung, stellen für jede das Konfidenzintervall in Gleichung (13.17) auf und zählen, wie viele der Konfidenzintervalle den Wert 0 überdecken. Das geschätze Konfidenzniveau beträgt 0.8757. Die Konfidenzintervalle sind also im Mittel zu schmal. Um zu sehen, woran dies liegt, formen wir den Ausdruck in der Klammer auf der linken Seite von Gleichung (13.18) um. Dabei bezeichnen wir $z_{1-\alpha/2}$ mit z.

$$\begin{aligned}
\bar{X} - z\,\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + z\,\frac{s}{\sqrt{n}} \quad &\Longleftrightarrow \quad -z\,\frac{s}{\sqrt{n}} \leq \mu - \bar{X} \leq z\,\frac{s}{\sqrt{n}} \\
&\Longleftrightarrow \quad -z\,\frac{s}{\sqrt{n}} \leq \bar{X} - \mu \leq z\,\frac{s}{\sqrt{n}} \\
&\Longleftrightarrow \quad -z \leq \frac{\bar{X}-\mu}{s/\sqrt{n}} \leq z
\end{aligned}$$

Wir müssen also folgende Wahrscheinlichkeit bestimmen:

$$P\left(-z_{1-\alpha/2} \leq \frac{\bar{X}-\mu}{s/\sqrt{n}} \leq z_{1-\alpha/2}\right)$$

Die Zufallsvariable

$$t = \frac{\bar{X}-\mu}{s/\sqrt{n}}$$

ist nicht standardnormalverteilt, wenn die $X_1, \ldots, X_n$ unabhängig und mit den Parametern μ und σ^2 normalverteilt sind. Die Schätzung von σ^2 führt dazu, dass t stärker streut als die Standardnormalverteilung. Von Gossett wurde 1908 gezeigt, dass t eine t-Verteilung mit $n-1$ Freiheitsgraden besitzt.

Ist $t_{n-1;1-\alpha/2}$ das $1-\alpha/2$-Quantil der t-Verteilung mit $n-1$ Freiheitsgraden, so gilt

$$P\left(-t_{n-1;1-\alpha/2} \leq \frac{\bar{X}-\mu}{s/\sqrt{n}} \leq t_{n-1;1-\alpha/2}\right) = 1-\alpha \tag{13.19}$$

Wir formen den Ausdruck in der Klammer auf der linken Seite von Gleichung (13.19) so um, dass zwischen den Ungleichheitszeichen nur noch μ steht. Es gilt

$$P(\bar{X} - t_{n-1;1-\alpha/2}\,\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}) = 1-\alpha$$

Durch diese Umformung haben wir ein Konfidenzintervall für μ bei Normalverteilung mit unbekannten σ^2 gefunden. Es lautet:

$$\boxed{\left[\bar{X} - t_{n-1;1-\alpha/2}\,\frac{s}{\sqrt{n}}, \bar{X} + t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}\right]}. \tag{13.20}$$

Dabei ist $t_{n-1;1-\alpha/2}$ das $1-\alpha/2$-Quantil der t-Verteilung mit $n-1$ Freiheitsgraden.

Beispiel 116 (fortgesetzt von Seite 369)
Hier sind noch einmal die Werte der Fahrzeit in Sekunden:

```
2318 2457 2282 2428 2376 2439 2272 2626 2255 2577
```

Wir unterstellen, dass die Fahrzeit normalverteilt ist und wollen das Konfidenzintervall für μ zum Konfidenzniveau 0.95 aufstellen. Es gilt $\bar{x} = 2403$ und $s^2 = 16278$. Mit $n = 10$ gilt also $s/\sqrt{n} = 40.346$. Der Tabelle der t-Verteilung auf Seite 547 entnehmen wir $t_{9;0.975} = 2.262$. Mit

$$t_{n-1;1-\alpha/2}\,\frac{s}{\sqrt{n}} = 91.269$$

erhalten wir folgendes Konfidenzintervall

$$[2311.731, 2494.269]$$

□

Schauen wir uns die Länge L des Konfidenzintervalls an. Es gilt

$$\boxed{L = 2\,t_{n-1;1-\alpha/2}\,\frac{s}{\sqrt{n}}} \tag{13.21}$$

Auch hier wird das Intervall breiter, wenn wir das Konfidenzniveau vergrößern. Es wird aber nicht notwendigerweise schmaler, wenn wir den Stichprobenumfang erhöhen. Für eine neue Stichprobe werden wir auch einen anderen Wert von s erhalten, sodass das Intervall größer werden kann. Es ist auch nicht möglich den Mindeststichprobenumfang zu bestimmen, um eine vorgegebene Länge des Intervalls nicht zu überschreiten, da der Stichprobenumfang von s abhängt. Aus $L \leq l$ folgt nämlich

$$n \geq \frac{4\, t^2_{n-1;1-\alpha/2}\, s^2}{l^2}\,.$$

Möglichkeiten zur Konstruktion von Konfidenzintervallen

Wir haben zwei Verfahren kennengelernt, mit denen man Konfidenzintervalle gewinnen kann.
Beim ersten Verfahren geht man aus von einer Schätzfunktion T des Parameters, die entweder exakt oder approximativ normalverteilt ist. Außerdem benötigt man den Standardfehler σ_T oder einen Schätzer $\hat{\sigma}_T$ des Standardfehlers. Ein Konfidenzintervall für θ zum Konfidenzniveau $1-\alpha$ ist dann

$$\left[T - z_{1-\alpha/2}\,\sigma_T, T + z_{1-\alpha/2}\,\sigma_T\right] \tag{13.22}$$

bzw.

$$\left[T - z_{1-\alpha/2}\,\hat{\sigma}_T, T + z_{1-\alpha/2}\,\hat{\sigma}_T\right]\,. \tag{13.23}$$

Dieses Verfahren wendet man bei M-L-Schätzern an, falls diese asymptotisch normalverteilt sind.
Beim zweiten Verfahren ist der Ausgangspunkt eine Stichprobenfunktion $g(T,\theta)$, die von einer Schätzfunktion T und dem Parameter θ abhängt. Weiterhin sei die Verteilung von $g(T,\theta)$ exakt oder approximativ bekannt. Sei q_p das p-Quantil der Verteilung von $g(T,\theta)$. Dann gilt

$$P\left(q_{\alpha/2} \leq g(T,\theta) \leq q_{1-\alpha/2}\right) = 1-\alpha$$

Formen wir den Ausdruck in der Klammer auf der linken Seite der Gleichung so um, dass in der Mitte zwischen den Ungleichheitszeichen nur noch θ steht, so erhalten wir ein Konfidenzintervall für θ.
Manchmal liefern beide Verfahren das gleiche Konfidenzintervall. Im nächsten Kapitel liefern sie unterschiedliche Konfidenzintervalle.

Konfidenzintervall für p

Es soll ein Konfidenzintervall zum Konfidenzintervall $1-\alpha$ für die Wahrscheinlichkeit p eines Ereignisses A aufgestellt werden.

Beispiel 133
Im ersten ZDF-Politbarometer im Februar 2003 wurden 1308 Personen befragt, welche Partei sie wählen würden, wenn am nächsten Sonntag Bundestagswahl wäre. Wir wollen ein Konfidenzintervall für den Anteil p der Wähler der SPD in der Bevölkerung aufstellen. □

Wir gehen von den unabhängigen, identisch mit dem Parameter p bernoulliverteilten Zufallsvariablen $X_1, \ldots, X_n$ aus. Es liegt nahe, den für p erwartungstreuen und konsistenten Schätzer $\hat{p} = \bar{X}$ als Ausgangspunkt bei der Konstruktion des Konfidenzintervalls zu wählen.

Beispiel 133 (fortgesetzt)
Von den 1308 Personen würden 288 SPD wählen. Somit gilt $\hat{p} = 0.22$. □

Beginnen wir mit dem ersten Verfahren. Auf Grund des zentralen Grenzwertsatzes ist $\hat{p}$ approximativ normalverteilt. Es gilt $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$. Da p unbekannt ist, ist auch der Standardfehler $\sigma_{\hat{p}}$ unbekannt. Wir ersetzen p durch $\hat{p}$ und erhalten den geschätzten Standardfehler $\hat{\sigma}_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})/n}$. Setzen wir in Gleichung (13.23) $\hat{p}$ für T und $\sqrt{\hat{p}(1-\hat{p})/n}$ für $\hat{\sigma}_T$ ein, so erhalten wir das Intervall für p:

$$\boxed{\left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]} \tag{13.24}$$

Man nennt dieses auch das **Wald-Konfidenzintervall**.

Beispiel 133 (fortgesetzt)
Wir wählen das Konfidenzniveau 0.95. Mit $n = 1308$ $\hat{p} = 0.22$ und $z_{0.975} = 1.96$ gilt:

$$\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.22 - 1.96\sqrt{\frac{0.22(1-0.22)}{1308}} = 0.198$$

Das Wald-Konfidenzintervall für p zum Konfidenzniveau 0.95 ist somit gegeben durch $[0.198, 0.242]$. □

Mit Hilfe des Wald-Konfidenzintervalls kann man den Stichprobenumfang bestimmen, den man benötigt, um zu vorgegebenem Konfidenzniveau $1-\alpha$

ein Konfidenzintervall mit einer vorgegebenem Höchstlänge zu erhalten. Die Länge des Wald-Konfidenzintervalls ist

$$\boxed{L = 2\, z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \tag{13.25}$$

Aus Gleichung (13.25) können wir auch herleiten, wie groß n sein muss, damit man bei gegebenem Konfidenzniveau $1-\alpha$ eine vorgegebene Länge l nicht überschreitet. Es muss also gelten $L \leq l$. Es muss also gelten

$$2\, z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq l$$

Lösen wir diese Ungleichung nach n auf, so erhalten wir

$$\boxed{n \geq \frac{4\, z_{1-\alpha/2}^2\, \hat{p}(1-\hat{p})}{l^2}} \tag{13.26}$$

Dieser Ausdruck hängt von $\hat{p}$ ab, das wir erst nach der Erhebung der Daten kennen. Um dieses Problem zu lösen, wählen wir den Wert von $\hat{p}$, für den $\hat{p}(1-\hat{p})$ maximal ist. Dies liefert dann den größten Wert von n. Das Maximum von $\hat{p}(1-\hat{p})$ liegt in $\hat{p} = 0.5$. Setzen wir diesen Wert für $\hat{p}$ in (13.26) ein, so erhalten wir

$$\boxed{n \geq \frac{z_{1-\alpha/2}^2}{l^2}}$$

Beispiel 134
Wie viele Personen muss man mindestens befragen, damit die Länge eines Konfidenzintervalls für p zum Konfidnzniveau 0.99 höchsten 0.02 beträgt?
Es gilt $z_{0.995} = 2.576$. Also muss gelten

$$n \geq \frac{2.576^2}{0.02^2} = 16589.44\,.$$

Man muss also mindestens 16590 Personen befragen. □

Schauen wir uns das zweite Verfahren an. Die Zufallsvariable

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

ist approximativ standardnormalverteilt. Es gilt also

$$P\left(-z_{1-\alpha/2} \leq \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{1-\alpha/2}\right) = 1-\alpha$$

Wir formen den Ausdruck in der Klammer so um, dass p in der Mitte der Ungleichheitszeichen alleine steht. Die Herleitung ist im Lehrbuch von Schlittgen zu finden. Wir erhalten folgendes Intervall, bei dem wir aus Gründen der Übersichtlichkeit z für $z_{1-\alpha/2}$ schreiben:

$$\left[\frac{\hat{p}+\frac{z^2}{2n}-z\sqrt{\frac{z^2}{4n^2}+\frac{\hat{p}(1-\hat{p})}{n}}}{1+\frac{z^2}{n}}, \frac{\hat{p}+\frac{z^2}{2n}+z\sqrt{\frac{z^2}{4n^2}+\frac{\hat{p}(1-\hat{p})}{n}}}{1+\frac{z^2}{n}}\right] \tag{13.27}$$

Man nennt dieses Intervall auch das Wilson-Konfidenzintervall.

Beispiel 133 (fortgesetzt von Seite 406)
Wir wählen das Konfidenzniveau 0.95. Mit $n = 1308$, $\hat{p} = 0.22$ und $z_{0.975} = 1.96$ gilt:

$$\begin{aligned}\frac{\hat{p}+\frac{z^2}{2n}-z\sqrt{\frac{z^2}{4n^2}+\frac{\hat{p}(1-\hat{p})}{n}}}{1+\frac{z^2}{n}} &= \frac{0.22+\frac{1.96^2}{2\cdot 1308}-1.96\sqrt{\frac{1.96^2}{4\cdot 1308^2}+\frac{0.22(1-0.22)}{1308}}}{1+\frac{1.96^2}{1308}} \\ &= 0.198\end{aligned}$$

Entsprechend erhalten wir die Obergrenze. Das Wilson-Konfidenzintervall für p zum Konfidenzniveau 0.95 ist [0.198 0.243]. □

13.4.2 Prognose- und Toleranzintervalle

Bisher haben wir nur Intervalle für einen Parameter betrachtet. Oft will aber auf Basis der Beobachtungen $x_1, \ldots, x_n$ ein Intervall für einen zukünftigen Wert x_{n+1} angeben. Dieses Intervall soll mit einer vorgegebenen Wahrscheinlichkeit $1-\alpha$ den zukünftigen Wert überdecken. Man nennt ein solches Intervall ein Pronoseintervall.

Beispiel 116 (fortgesetzt von Seite auf Seite 369)
Der Arbeitnehmer will wissen, wie lange er bei der nächsten Fahrt unterwegs ist. Er wird also den zukünftigen Wert prognostizieren. Ein Intervall für die

nächste Fahrzeit ist ein **Prognoseintervall**. Dieses Intervall wird er wie auch das Konfidenzintervall mit einer Sicherheit versehen.

□

Wir gehen zunächst davon aus, dass die Grundgesamtheit mit den Parametern μ und σ^2 normalverteilt ist. Sind μ und σ^2 bekannt, so liefert das zentrale Schwankungsintervall

$$\boxed{[\mu - z_{1-\alpha/2}\sigma, \mu - z_{1-\alpha/2}\sigma]} \tag{13.28}$$

die Lösung. Es gilt

$$P(\mu - z_{1-\alpha/2}\sigma \leq X \leq \mu - z_{1-\alpha/2}\sigma) = 1 - \alpha$$

In der Regel sind μ und σ^2 aber unbekannt. Es liegt nahe, diese zu schätzen und die Schätzwerte in die Gleichung (13.28) einzusetzen. Dies ist aber nur für sehr große Stichprobenumfänge sinnvoll. Für kleine Stichprobenumfänge hingegen ist das exakte Prognoseintervall bei Normalverteilung gegeben durch

$$\boxed{[\bar{x} - t_{n-1;1-\alpha/2}\, s\, \sqrt{1 + 1/n}, \bar{x} - t_{n-1;1-\alpha/2}\, s\, \sqrt{1 + 1/n}]} \tag{13.29}$$

Schauen wir uns ein Beispiel an, bevor wir zeigen, wie man dieses Intervall gewinnt.

Beispiel 116 (fortgesetzt)
Der Arbeitnehmer sucht ein Prognoseintervall für die Fahrzeit des nächsten Tages zum Niveau 0.95. Es gilt $n = 10$, $\bar{x} = 2403$ und $s = 127.6$. Mit $t_{9;0.975} = 2.262$ erhalten wir folgendes Prognoseintervall

$$[2100.3, 2705.7]$$

□

Um das Intervall herzuleiten, fragen wir uns zunächst, wie wir den Wert von X_{n+1} prognostizieren sollen. Da er möglichst nahe an allen Beobachtungen liegen sollte, bieten sich zwei Kriterien an. Wählt man als Maß für die Nähe, die euklidische Distanz, so erhält man folgendes Kriterium

$$\min \sum_{i=1}^{n} |x_i - x_{n+1}| \tag{13.30}$$

Die quadrierte euklidische Distanz liefert

$$\min \sum_{i=1}^{n} (x_i - x_{n+1})^2 \tag{13.31}$$

Im ersten Fall prognostiziert man x_{n+1} durch den Median, im zweiten Fall durch den Mittelwert der Beobachtungen. Da die Verteilung des Mittelwerts angegeben werden kann, verwenden wir diesen. Als Ausgangspunkt der Konstruktion des Prognoseintervalls wählen wir $\bar{X} - X_{n+1}$. Wir gehen im Folgenden davon aus, dass die Zufallsvariablen $X_1, \ldots, X_n, X_{n+1}$ unabhängig und identisch mit den Parametern μ und σ^2 normalverteilt sind. Unter diesen Annahmen gilt

$$E(\bar{X} - X_{n+1}) = E(\bar{X}) - E(X_{n+1}) = \mu - \mu = 0$$

und

$$Var(\bar{X} - X_{n+1}) = Var(\bar{X}) + Var(X_{n+1}) = \frac{\sigma^2}{n} + \sigma^2 = \sigma^2(1 + 1/n)$$

Außerdem ist $\bar{X} - X_{n+1}$ normalverteilt. Also ist

$$\frac{\bar{X} - X_{n+1}}{\sigma \sqrt{1 + 1/n}} \tag{13.32}$$

standardnormalverteilt. Schätzen wir σ durch s und setzen es in Gleichung (13.32) ein, so erhalten wir folgende mit $n - 1$ Freiheitsgraden t-verteilte Zufallsvariable

$$\frac{\bar{X} - X_{n+1}}{s \sqrt{1 + 1/n}}$$

Es gilt also

$$P\left[-t_{n-1;1-\alpha/2} \leq \frac{\bar{X} - X_{n+1}}{s \sqrt{1 + 1/n}} \leq t_{n-1;1-\alpha/2}\right] = 1 - \alpha \tag{13.33}$$

Formen wir diesen Ausdruck so um, dass zwischen den Ungleichungen X_{n+1} steht, so erhalten wir das Prognoseintervall in Gleichung (13.29).

Die Grenzen eines Prognoseintervalls sind Zufallsvariablen. Dies hat zur Konsequenz, dass die Wahrscheinlichkeit, dass die Beobachtung x_{n+1} im konkreten Prognoseintervall liegt, nicht notwendigerweise $1 - \alpha$ beträgt. Eine kleine Simulation soll dies verdeutlichen. Wir ziehen eine Zufallsstichprobe vom Umfang $n = 9$ aus einer standardnormalverteilten Grundgesamtheit:

```
-0.055 1.419 0.411 -1.252 -0.136 -0.224 0.236 -0.089 0.794
```

Es gilt $\bar{x} = 0.123$ und $s = 0.74$. Mit $t_{8;0.975} = 2.306$ erhalten wir folgendes Prognoseintervall $[-1.676, 1.921]$. Die Wahrscheinlichkeit, dass eine standardnormalverteilte Zufallsvariable einen Wert aus diesem Intervall annimmt, beträgt

$$\phi(1.921) - \phi(-1.676) = 0.926$$

und nicht 0.95. Die Wahrscheinlichkeit $1 - \alpha$ interpretieren wir wie das Konfidenzniveau beim Konfidenzintervall. Die Wahrscheinlichkeit beträgt $1 - \alpha$, dass wir ein Intervall finden, das die Beobachtung x_{n+1} enthält.

Ein wichtiger Unterschied besteht zwischen Konfidenzintervallen und Prognoseintervallen. Ein Konfidenzintervall ist ein Intervall für den Wert eines Parameters, ein Prognoseintervall ein Intervall für eine Realisation einer Zufallsvariablen. Somit ist die Aussage beim Prognoseintervall unsicherer. Dies zeigt sich in der größeren Länge des Prognoseintervalls. Die Länge des Konfidenzintervalls für μ ist

$$L = 2\, t_{n-1;1-\alpha/2}\, s\, \sqrt{1/n}$$

Die Länge des Prognoseintervalls beträgt

$$L = 2\, t_{n-1;1-\alpha/2}\, s\, \sqrt{1 + 1/n}$$

Die Länge des Konfidenzintervalls konvergiert gegen 0, wänhrend die Länge des Prognoseintervalls gegen die Länge des zentralen Schwankungsintervalls konvergiert.

Oft sucht man einseitige Prognoseintervalle. Man will also wissen, welchen Wert die nächste Beobachtung mindestens oder höchsten annimmt.

Bei Normalverteilung gibt es folgende einseitige Prognoseintervalle

$$\boxed{[\bar{x} - t_{n-1;1-\alpha}\, s\, \sqrt{1 + 1/n}, \infty)} \tag{13.34}$$

$$\boxed{(-\infty, \bar{x} - t_{n-1;1-\alpha}\, s\, \sqrt{1 + 1/n}]} \tag{13.35}$$

Beispiel 116 (fortgesetzt)

Der Arbeitnehmer will wissen, welchen Wert seine Fahrzeit am nächsten Tag nicht überschreiten wird. Er stellt ein einseitiges Prognoseintervall zum Niveau 0.95 auf. Es gilt $n = 10$, $\bar{x} = 2403$ und $s = 127.6$. Mit $t_{9;0.95} = 1.8331$ erhalten wir folgendes Prognoseintervall

$$(-\infty, 2648.3]$$

□

Bisher sind wir davon ausgegangen, dass die Grundgesamtheit normalverteilt ist. Kann von dieser Annahme nicht ausgegangen werden, so kann man folgendes zweiseitige Prognoseintervall für x_{n+1} verwenden:

$$\boxed{[x_{(i)}, x_{(j)}]} \tag{13.36}$$

Dabei sind $x_{(i)}$ und $x_{(j)}$ die i-te bzw. j-te Beobachtung in der geordneten Stichprobe. Das Niveau des zweiseitigen Intervalls in Gleichung (13.36) ist

$$1 - \alpha = \frac{j - i}{n + 1}$$

Liegen i und j nahe beieinander, so ist das Niveau klein. Die größte Sicherheit beträgt $(n-1)/(n+1)$. Man erhält sie für das Intervall $[x_{(1)}, x_{(n)}]$.
Analog erhält man einseitige Prognoseintervalle

$$\boxed{(-\infty, x_{(n-i)}]} \tag{13.37}$$

$$\boxed{[x_{(i)}, \infty]} \tag{13.38}$$

Das Niveau des einseitigen Prognoseintervalle in den Gleichungen (13.37) und (13.38) gleich

$$1 - \alpha = \frac{n + 1 - i}{n + 1}$$

Beispiel 116 (fortgesetzt)
Es gilt $x_{(1)} = 2255$ und $x_{(n)} = 2626$. Also ist das zweiseitige Prognoseintervall zum Niveau $(10 - 1)/(10 + 1) = 0.82$ gleich $[2255, 2626]$. Die einseitigen Intervalle besitzen das Niveau 0.91.

□

Hahn (1970) nennt das Prognoseintervall auch das Astronauten-Intervall. Ein Astronaut interessiert sich nur für den nächsten Flug und will wissen, in welchem Bereich er die Werte erwarten kann. Der Hersteller eines Produktes ist aber nicht nur an einer Beobachtung interessiert, sondern an der gesamten Produktion. Er will ein Intervall angeben, in dem sich mindestens der Anteil γ befindet. Auch hier muss die Wahrscheinlichkeit angegeben werden,

da die Grenzen des Intervalls Zufallsvariablen sind. Man sucht also ein Intervall, in dem sich mit Wahrscheinlichkeit $1-\alpha$ mindestens der Anteil γ der Beobachtungen in der Grundgesamtheit befindet. Dieses Intervall heißt **Toleranzintervall**.
Wir gehen zunächst davon aus, dass die Grundgesamtheit normalverteilt ist. Das zweiseitige Toleranzintervall ist gegeben durch

$$\boxed{[\bar{x} - q_{1-\alpha,p,n}\, s, \bar{x} + q_{1-\alpha,p,n}\, s]} \tag{13.39}$$

Tabellen für $q_{1-\alpha,p,n}$ sind auf den Seiten 554 und 555 zu finden.

Beispiel 116 (fortgesetzt von Seite auf Seite 369)
Der Arbeitnehmer will wissen, in welchem Intervall mindestens 90 Prozent der Fahrzeiten mit einer Wahrscheinlichkeit von 0.95 liegen. Es gilt $n = 10$, $\bar{x} = 2403$ und $s = 127.6$. Der Tabelle 20.10 auf Seite 554 entnehmen wir $q_{0.95,0.9,n} = 2.839$. Wir erhalten folgendes Toleranzintervall

$$[2040.79, 2765.22]$$

□

Für $q_{1-\alpha,p,n}$ in Gleichung (13.39) gilt

$$q_{1-\alpha,p,n} = r\sqrt{\frac{n-1}{\chi^2_{\alpha,n-1}}} \tag{13.40}$$

mit

$$\Phi\left(\frac{1}{\sqrt{n}} + r\right) - \Phi\left(\frac{1}{\sqrt{n}} - r\right) = p \tag{13.41}$$

Dabei ist $\chi^2_{\alpha,n-1}$ das α-Quantil der χ^2-Verteilung mit $n-1$ Freiheitsgraden und $\Phi(z)$ der Wert der Verteilungsfunktion der Standardnormalverteilung an der Stelle z.
Die Gleichungen (13.40) und (13.41) werden von Kendall, Stuart & Ord (1991) auf den Seiten 774-776 hergeleitet.
Wir können $q_{1-\alpha,p,n}$ aber auch numerisch bestimmen. Hierzu müssen wir zuerst den Wert von r bestimmen, der die Gleichung (13.41) erfüllt. Lösen wir die Gleichung (13.41) nach 0 auf, so erhalten wir

$$\Phi\left(\frac{1}{\sqrt{n}} + r\right) - \Phi\left(\frac{1}{\sqrt{n}} - r\right) - p = 0$$

Wir suchen also die Nullstelle von

$$f(r) = \Phi\left(\frac{1}{\sqrt{n}} + r\right) - \Phi\left(\frac{1}{\sqrt{n}} - r\right) - p \tag{13.42}$$

Ein numerisches Verfahren ist das Newton-Raphson-Verfahren. Ist die Nullstelle der Funktion $f(r)$ gesucht, so iteriert man hier beginnend mit dem Startwert r_0

$$r_n = r_{n-1} - \frac{f(r_{n-1})}{f'(r_{n-1})} \tag{13.43}$$

für $n = 1, 1, \ldots$ so lange, bis r_n sich stabilisiert.
Als Startwert r_0 kann man den Wert von r wählen, der für $n \to \infty$ die Gleichung (13.41) erfüllt.
Für gegebenes p wird der Wert r_0 gesucht mit

$$\Phi(r_0) - \Phi(-r_0) = p$$

Wegen

$$\Phi(-r_0) = 1 - \Phi(r_0)$$

muss also gelten

$$2\,\Phi(r_0) - 1 = p$$

Somit folgt

$$r_0 = \Phi^{-1}\left(\frac{p+1}{2}\right)$$

Mit

$$f'(r) = \phi\left(\frac{1}{\sqrt{n}} + r\right) + \phi\left(\frac{1}{\sqrt{n}} - r\right)$$

können wir die Iteration aus Gleichung (13.43) auf Seite 414 also durchführen.
Dabei gilt

$$\phi(z) = \frac{1}{\sqrt{2\,\pi}}\, e^{-0.5\,z^2}$$

Beispiel 116 (fortgesetzt)
Der Arbeitnehmer will wissen, in welchem Intervall mindestens 90 Prozent der Fahrzeiten mit einer Wahrscheinlichkeit von 0.95 liegen.
Es gilt also $p = 0.9$ und $1 - \alpha = 0.95$. Wir bestimmen zunächst r. Es gilt

$$r_0 = \Phi^{-1}\left(\frac{0.9+1}{2}\right) = \Phi^{-1}(0.95) = 1.645$$

Somit gilt

$$\begin{aligned} r_1 &= 1.645 - \frac{\Phi\left(\frac{1}{\sqrt{10}} + 1.645\right) - \Phi\left(\frac{1}{\sqrt{10}} - 1.645\right) - 0.9}{\frac{1}{\sqrt{2\,\pi}}\, e^{-0.5\,(1/\sqrt{(10)}+1.645)^2} + \frac{1}{\sqrt{2\,\pi}}\, e^{-0.5\,(1\sqrt{10}3-1.645)^2}} \\ &= 1.645 - \frac{\Phi(1.96) - \Phi(-1.33) - 0.9}{\frac{1}{\sqrt{2\,\pi}}\, e^{-0.5\,(1/\sqrt{10}+1.645)^2} + \frac{1}{\sqrt{2\,\pi}}\, e^{-0.5\,(1/\sqrt{10}-1.645)^2}} \\ &= 1.645 - \frac{0.975 - 0.092 - 0.9}{0.0583 + 0.165} = 1.645 - \frac{-0.017}{0.2233} = 1.721 \end{aligned}$$

Ausgehend von r_1 bestimmen wir r_2 und erhalten $r_2 = 1.725$. Auch r_3 ist gleich 1.725. Wir können also $q_{0.95,0.9,10}$ bestimmen. Mit $\chi^2_{0.05,9} = 3.325$ gilt

$$q_{0.95,0.9,10} = 1.725 \cdot \sqrt{\frac{9}{3.325}} = 2.838$$

□

Bei einem Toleranzintervall bezieht sich die innere Wahrscheinlichkeitsaussage in Gleichung (13.39) auf Seite 413 auf das Intervall, während sich die äußere Wahrscheinlichkeitsaussage auf die Prozedur bezieht. Wenn wir sehr viele Stichproben $x_1, \ldots, x_n$ und für jede das Toleranzintervall in Gleichung (13.39) bestimmen, so erwarten wir, dass $100 \cdot (1 - \alpha)$ Prozent der Intervalle mindestens die Wahrscheinlichkeitsmasse p enthalten.
Kann keine Normalverteilung unterstellt werden, so kann man das Intervall

$$\boxed{[x_{(1)}, x_{(n)}]}, \tag{13.44}$$

aufstellen. Die Wahrscheinlichkeit $1-\alpha$ ist beim Intervall in Gleichung (13.44) gleich

$$1 - \alpha = 1 - p^n - n(1-p)p^{n-1}$$

Diese Beziehung wird in Mood, Graybill & Boes (1974) auf den Seiten 516-517 hergeleitet.

Beispiel 116 (fortgesetzt)
Der Arbeitnehmer ist an einem Toleranzintervall mit Mindestanteil 0.8 interessiert. Verwendet er das Intervall $[2255, 2626]$ als Toleranzintervall mit Mindestanteil 0.8, so beträgt die Sicherheit

$$1 - \alpha = 1 - 0.8^10 - 10 \cdot 0.2 \cdot 0.8^9 = 0.624$$

□

13.5 Geschichtete Stichproben

Wir sind interessiert an einem Merkmal X in einer Grundgesamtheit vom Umfang N. Wir zerlegen die Grundgesamtheit in die Schichten $G_1, \ldots, G_K$. Der Umfang der i-ten Schicht sei N_i. Der Erwartungswert von X der i-ten Schicht ist μ_i. Für den Erwartungswert μ der Grundgesamtheit gilt

$$\mu = \sum_{i=1}^{K} \frac{N_i}{N} \mu_i \tag{13.45}$$

Um μ zu schätzen, ziehen wir aus jeder der Schichten eine Zufallsstichprobe, wobei der Stichprobenumfang aus der i-ten Schicht n_i beträgt. Sei x_{ij} der Wert, den wir bei der j-ten Ziehung aus der i-ten Schicht beobachten. Es gilt $i = 1, \ldots, K$ und $j = 1, \ldots, n_i$. Der Mittelwert aller Beobachtungen ist

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{K} \sum_{j=1}^{n_i} x_{ij} \tag{13.46}$$

Sei

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} \tag{13.47}$$

der Mittelwert der Stichprobe aus der i-ten Schicht. Dann gilt

$$\bar{x} = \sum_{i=1}^{K} \frac{n_i}{n} \bar{x}_i \tag{13.48}$$

Dies sieht man folgendermaßen:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{K} \sum_{j=1}^{n_i} x_{ij} \overset{(13.47)}{=} \frac{1}{n} \sum_{i=1}^{K} n_i \bar{x}_i = \sum_{i=1}^{K} \frac{n_i}{n} \bar{x}_i$$

Um die Eigenschaften von $\bar{x}$ zu analysieren, fassen wir die x_{ij} als Realisation der Zufallsvariablen X_{ij} auf. Es gilt $E(X_{ij}) = \mu_i$. Wir betrachten die Schätzfunktion

$$\bar{X} = \sum_{i=1}^{K} \frac{n_i}{n} \bar{X}_i$$

Es gilt

$$E(\bar{X}_i) = E(\frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}) = \frac{1}{n_i} \sum_{j=1}^{n_i} E(X_{ij}) \frac{1}{n_i} \sum_{j=1}^{n_i} \mu_i = \mu_i \tag{13.49}$$

Also gilt

$$E(\bar{X}) = E(\sum_{i=1}^{K} \frac{n_i}{n} \bar{X}_i) = \sum_{i=1}^{K} \frac{n_i}{n} E(\bar{X}_i) \overset{(13.49)}{=} \sum_{i=1}^{K} \frac{n_i}{n} \mu_i$$

Ein Vergleich mit Gleichung (13.45) zeigt, dass $\bar{X}$ nicht erwartungstreu ist. Wenn aber gilt

$$\frac{n_i}{n} = \frac{N_i}{N}$$

dann ist $\bar{X}$ erwartungstreu. Gleichung (13.45) zeigt, wie wir eine erwartungstreue Schätzfunktion für μ erhalten können. Wir müssen die $\bar{X}_i$ nicht mit n_i/n, sondern mit N_i/N gewichten:

$$\hat{\mu}_{ST} = \sum_{i=1}^{K} \frac{N_i}{N} \bar{X}_i$$

Die Erwartungstreue von $\hat{\mu}_{ST}$ ist leicht gezeigt:

$$E(\hat{\mu}_{ST}) = E(\sum_{i=1}^{K} \frac{N_i}{N} \bar{X}_i) = \sum_{i=1}^{K} \frac{N_i}{n} E(\bar{X}_i) \overset{(13.49)}{=} = \sum_{i=1}^{K} \frac{N_i}{n} \mu_i \overset{(13.45)}{=} \mu$$

Beim Schichten schätzt man den Erwartungswert also durch einen gewichteten Mittelwert der Mittelwerte der Schichten, wobei die Gewichte die Anteile der Umfänge der Schichten an der Grundgesamtheit sind. Um den Schätzer in Gleichung (13.50) anwenden zu können, benötigt man also die Quoten N_i/N. In *Pokropp: Stichproben: Theorie und Verfahren* wird gezeigt, wie man auch ohne Kenntnis der Quoten schätzen kann.

Schauen wir uns noch einmal das Schichtungsverfahren. Wir gehen wieder aus von einer Grundgesamtheit, die in K Schichten $G_1, \ldots, G_K$ zerlegt wird. Der Erwartungswert des Merkmals X in der i-ten Schicht ist μ_i und die Varianz ist σ_i^2. Wir ziehen aus der i-ten Schicht n_i zufällig mit Zurücklegen und schätzen den Erwartungswert μ der Grundgesamtheit durch

$$\hat{\mu}_{ST} = \sum_{i=1}^{K} \frac{N_i}{N} \bar{X}_i \tag{13.50}$$

Wegen

$$Var(\bar{X}_i) = \frac{\sigma_i^2}{n_i}$$

gilt

$$Var(\hat{\mu}_{ST}) = Var(\sum_{i=1}^{K} \frac{N_i}{N} \bar{X}_i) = \sum_{i=1}^{K} \frac{N_i^2}{N^2} Var(\bar{X}_i) = \sum_{i=1}^{K} \frac{N_i^2}{N^2} \frac{\sigma_i^2}{n_i} \tag{13.51}$$

Die Varianz von $\hat{\mu}_{ST}$ hängt von den Stichprobenumfängen aus den Schichten ab. Diese können wir frei wählen. Wir betrachten im Folgenden zwei unterschiedliche Fälle. Dabei gehen wir davon aus, dass insgesamt n Beobachtungen aus der Grundgesamtheit gezogen wurden. Es gilt also

$$\sum_{i=1}^{K} n_i = n \tag{13.52}$$

Im ersten Fall wählen wir

$$n_i = n \, \frac{N_i}{N} \tag{13.53}$$

Man spricht von proportionaler Aufteilung. In diesem Fall gilt

$$Var(\hat{\mu}_{ST}) = \sum_{i=1}^{K} \frac{N_i^2}{N^2} \, \frac{\sigma_i^2}{n_i} = \sum_{i=1}^{K} \frac{N_i^2}{N^2} \, \frac{N \sigma_i^2}{n \, N_i} = \frac{1}{n} \sum_{i=1}^{K} \frac{N_i}{N} \, \sigma_i^2 \tag{13.54}$$

Im zweiten Fall wählen wir

$$n_i = n \, \frac{N_i \, \sigma_i}{\sum\limits_{i=1}^{K} N_i \, \sigma_i} \tag{13.55}$$

In diesem Fall gilt

$$\begin{aligned}
Var(\hat{\mu}_{ST}) &= \sum_{i=1}^{K} \frac{N_i^2}{N^2} \, \frac{\sigma_i^2}{n_i} \\
&= \sum_{i=1}^{K} \frac{N_i^2}{N^2} \, \frac{\sigma_i^2 \sum_{i=1}^{K} N_i \, \sigma_i}{n \, N_i \, \sigma_i} \\
&= \frac{\sum_{i=1}^{K} N_i \, \sigma_i}{n \, N^2} \sum_{i=1}^{K} N_i \, \sigma_i \\
&= \frac{1}{n} \frac{\left(\sum_{i=1}^{K} N_i \, \sigma_i \right)^2}{N^2} \\
&= \frac{1}{n} \left(\sum_{i=1}^{K} \frac{N_i}{N} \, \sigma_i \right)^2
\end{aligned} \tag{13.56}$$

Wir bilden die Differenz aus der Varianz in Gleichung (13.54) und der Varianz in Gleichung (13.56). Dabei setzen wir $W_k = N_k/N$ und

$$\bar{\sigma} = \sum_{i=1}^{K} W_i \, \sigma_i$$

Es gilt

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^{K} \frac{N_i}{N} \, \sigma_i^2 - \frac{1}{n} \left(\sum_{i=1}^{K} \frac{N_i}{N} \, \sigma_i \right)^2 &= \frac{1}{n} \left(\sum_{i=1}^{K} W_i \, \sigma_i^2 - \bar{\sigma}^2 \right) \\
&= \frac{1}{n} \left(\sum_{i=1}^{K} W_i \, \sigma_i^2 - 2 \, \bar{\sigma}^2 + \bar{\sigma}^2 \right) \\
&= \frac{1}{n} \left(\sum_{i=1}^{K} W_i \, \sigma_i^2 - 2 \, \bar{\sigma} \sum_{i=1}^{K} W_i \, \sigma_i + \bar{\sigma}^2 \sum_{i=1}^{K} W_i \right) \\
&= \frac{1}{n} \left(\sum_{i=1}^{K} W_i \, \sigma_i^2 - \sum_{i=1}^{K} W_i \, 2 \, \bar{\sigma} \, \sigma_i + \sum_{i=1}^{K} W_i \, \bar{\sigma}^2 \right) \\
&= \frac{1}{n} \left(\sum_{i=1}^{K} \left(W_i \, \sigma_i^2 - W_i \, 2 \, \bar{\sigma} \, \sigma_i + W_i \, \bar{\sigma}^2 \right) \right) \\
&= \frac{1}{n} \left(\sum_{i=1}^{K} \left(W_i \left(\sigma_i^2 - 2 \, \bar{\sigma} \, \sigma_i + \bar{\sigma}^2 \right) \right) \right) \\
&= \frac{1}{n} \left(\sum_{i=1}^{K} \left(W_i \left(\sigma_i^2 - \bar{\sigma} \right)^2 \right) \right) \qquad (13.57)
\end{aligned}$$

Wir sehen, dass die Varianz in Gleichung (13.54) größer ist als die Varianz in Gleichung (13.56). An Gleichung (13.57) können wir erkennen, dass die Differenz um so größer wird, je mehr sich die Varianzen in den Schichten unterscheiden. Die Aufteilung in Gleichung (13.55) nennt man auch optimale Aufteilung, da sie die Varianz von $\hat{\mu}$ minimiert. Gleichung (13.55) können wir auch entnehmen, dass wir aus Schichten mit großer Streuung mehr Beobachtungen entnehmen müss en als aus Schichten mit geringer Streuung. Dies ist auch intuitiv klar. Aus Schichten mit geringer Streuung benötigt man wenig Beobachtungen, da sich die Merkmalsträger wenig unterscheiden.

13.6 Schätzen in R

Auf Seite 380 haben wir $Var(\bar{X})$ mit $Var(X_{0.5})$ bei Normalverteilung verglichen. Dabei haben wir $Var(X_{0.5})$ mit einer Simulation geschätzt. Schauen wir uns an, wie man diese Simulation in R für $n = 10$ durchführt.

Wir müssen 100000 Stichproben vom Umfang $n = 10$ aus der Standardnormalverteilung ziehen und den Median jeder Stichprobe bestimmen. Als Schätzer von $Var(X_{0.5})$ dient die Stichprobenvarianz der 100000 Mediane.

Wir initialisieren einen Vektor `me` der Länge 100000, in den wir die Mediane schreiben.

```
> me<-rep(0,100000)
```

Mit einer for-Schleife erzeugen wir mit der Funktion `rnorm` jeweils 10 standardnormalverteilte Zufallszahlen, bestimmen für diese den Median und weisen diesen der i-ten Komponente des Vektors `me` zu.

```
> for (i in 1:100000) me[i]<-median(rnorm(10))
```

Nun berechnen wir mit der Funktion `var` die Stichprobenvarianz der Zahlen im Vektor `me`.

```
> var(me)
[1] 0.1388393
```

In R wird jeder Befehl interpretiert. Deshalb dauert die Iteration relativ lange. Man gelangt mit folgender Lösung wesentlich schneller zum Ziel:

Man erstellt mit der Funktion `matrix` eine $(100000, 10)$-Matrix standardnormalverteilter Zufallszahlen und bestimmt mit der Funktion `apply` den Median jeder Stichprobe. So erhält man ebenfalls einen Vektor mit Medianen, deren Varianz man mit der Funktion `var` bestimmt.

```
> var(apply(matrix(rnorm(1000000),100000,10),1,median))
[1] 0.1387574
```

Wenn wir den Startwert des Zufallszahlengenerators in beiden Fällen gleich wählen, erhalten wir auch das gleiche Ergebnis.

```
> set.seed(14062006)
> me<-rep(0,100000)
> for (i in 1:100000) me[i]<-median(rnorm(10))
> var(me)
[1] 0.1372801
```

```
> set.seed(14062006)
> var(apply(matrix(rnorm(1000000),100000,10,byrow=T),1,median))
[1] 0.1372801
```

Mit der Funktion `density` können wir eine Dichteschätzung durchführen. Das erste Argument ist der Datensatz. Standardmäßig wird eine Gauss-Kern gewählt. Den Wert h übergeben wir im Argument `bw`, wobei `"nrd0"` den Vorschlag von Silverman und `"nrd"` den Vorschlag von Scott enthält. Abbildung 13.9 auf Seite 396 erhalten wir durch

```
> attach(weiterbildung)
> plot(density(Alter,bw="nrd"),bty="l")
```

Schauen wir uns an, wie man Konfidenzintervalle in R bestimmt. Das Konfidenzintervall für μ bei Normalverteilung mit unbekanntem σ^2 wollen wir für die Daten aus Beispiel 116 auf Seite 369 aufstellen. Wir weisen die Daten der Variablen `fahrzeit` zu

```
> fahrzeit<-c(2318,2457,2282,2428,2376,2439,2272,2626,2255,2577)
```

Der Aufruf

```
> t.test(fahrzeit)[[4]]
```

liefert die Grenzen des Konfidenzintervalls für μ zum Konfidenzniveau 0.95:

```
[1] 2311.731 2494.269
attr(,"conf.level")
[1] 0.95
```

Mit dem Argument `conf.level` können wir das Konfidenzniveau festlegen.

```
> t.test(fahrzeit,conf.level=0.99)[[4]]
[1] 2271.882 2534.118
attr(,"conf.level")
[1] 0.99
```

Es gibt noch eine weitere Möglichkeit. Der Aufruf

```
> predict(lm(fahrzeit~1),newdata=data.frame(1),
          interval="confidence")
      fit      lwr      upr
[1,] 2403 2311.731 2494.269
```

liefert ebenfalls das Konfidenzintervalls für μ zum Konfidenzniveau 0.95. Mit dem Parameter `level` können wir das Konfidenzniveau festlegen.

```
> predict(lm(fahrzeit~1),newdata=data.frame(1),
          interval="confidence",level=0.99)
      fit      lwr     upr
 [1,] 420 310.9230 529.077
```

Mit der Funktion `predict` kann man auch das auf Normalverteilung basierende Prognoseintervall aufstellen. Wir betrachten das Beispiel 116 auf Seite 369:

```
> predict(lm(fahrzeit~1),newdata=data.frame(1),
          interval="prediction",level=0.95)
      fit      lwr      upr
[1,] 2403 2271.882 2534.118
```

Toleranzintervalle sind nicht in `R` implementiert. Wir können aber die Werte von $q_{1-\alpha,p,n}$ über die Gleichungen (13.40) und (13.40) auf Seite 413 numerisch bestimmen. Die Nullstelle r der Funktion 13.42 auf Seite 414 erhalten wir folgendermaßen:

```
> fr<-function(r,n,p) pnorm(1/sqrt(n)+r)-pnorm(1/sqrt(n)-r)-p
> n<-10
> p<-0.9
> r<-uniroot(fr,c(-4,4),n=n,p=p)[[1]]
> r
[1] 1.725331
```

Den Wert $q_{1-\alpha,p,n}$ in Gleichung (13.40) auf Seite 413 erhalten wir dann folgendmaßen:

```
> qpn<-r*sqrt((n-1)/qchisq(alpha,n-1))
> qpn
[1] 2.838511
```

Das Toleranzintervall in Gleichung (13.39) auf Seite 413 liefert dann folgende Befehlsfolge:

```
> mean(fahrzeit)-qpn*sd(fahrzeit)
[1] 2040.848
> mean(fahrzeit)+qpn*sd(fahrzeit)
[1] 2765.152
```

Schauen wir uns nun die Konfidenzintervalle für p an. Wir betrachten die Daten aus Beispiel 133 auf Seite 406. Von 1308 Personen würden 288 SPD wählen.

Die Konfidenzintervalle für p kann man mit der Funktion `binconf` aus dem Paket `Hmisc` durchführen. Dieses Paket muss man zunächst installieren und laden. Wie man dabei vorzugehen hat, wird auf Seite 52 beschrieben.

Der Funktion `binconf` übergibt man im ersten Argument `x` die Anzahl der Erfolge, im zweiten Argument `n` den Stichprobenumfang, im dritten Argument `alpha` 1-Konfidenzniveau und im vierten Argument `method` das Verfahren.

Die Grenzen des Wald-Konfidenzintervalls für p zum Konfidenzniveau 0.95 liefert der Aufruf

```
> binconf(288,1308,alpha=0.05,method="exact")
  PointEst     Lower     Upper
 0.2201835 0.1979941 0.2436376
```

und die Grenzen des Wilson-Konfidenzintervalls für p zum Konfidenzniveau 0.95 liefert der Aufruf

```
> binconf(288,1308,alpha=0.05,method="wilson")
  PointEst     Lower     Upper
 0.2201835 0.1985648 0.2434410
```

Kapitel 14

Grundbegriffe statistischer Tests

Oft hat man eine Vermutung über die Verteilung einer Zufallsvariablen X. Diese Vermutung formuliert man als **Hypothese** H_0. So könnte man daran interessiert sein zu überprüfen, ob ein Parameter θ einen speziellen Wert θ_0 annimmt. Diese Hypothese lautet:

$$H_0 : \theta = \theta_0 \tag{14.1}$$

Zu jeder Hypothese H_0 formuliert man eine sogenannte **Gegenhypothese** H_1. Eine Gegenhypothese zur Hypothese in Gleichung (14.1) ist

$$H_1 : \theta \neq \theta_0 \tag{14.2}$$

Beispiel 134
Es soll überprüft werden, ob eine Münze fair ist. Ist die Münze fair, so beträgt die Wahrscheinlichkeit 0.5, dass KOPF fällt. Wir bezeichnen die Wahrscheinlichkeit für KOPF mit p und erhalten folgendes Hypothesenpaar.

$$H_0 : p = 0.5 \quad \text{gegen} \quad H_1 : p \neq 0.5.$$

□

Um mit statistischen Verfahren zu überprüfen, ob die Hypothese oder Gegenhypothese zutrifft, beobachtet man den Zufallsvorgang mehrmals. Dies kann auch bedeuten, dass man eine Stichprobe zieht.

Beispiel 134 (fortgesetzt)
Die Münze wird 5-mal geworfen. Wir bezeichnen KOPF mit `K` und ZAHL mit `Z`. Es ergibt sich folgende Stichprobe:

```
K K K Z K
```

Spricht diese Stichprobe für H_0 oder für H_1? □

Es gibt Stichproben, die für die Hypothese H_0 und Stichproben, die für die Gegenhypothese H_1 sprechen. Um entscheiden zu können, ob die Hypothese oder die Gegenhypothese zutrifft, verdichten wir die Information in der Stichprobe. Wir bestimmen eine Stichprobenfunktion $S = g(X_1, \ldots, X_n)$. Diese Stichprobenfinktion $S = g(X_1, \ldots, X_n)$ nennen wir **Teststatistik** oder **Prüfgröße**.

Beispiel 134 (fortgesetzt)
Die Stichproben `KKKKK` und `ZZZZZ` sprechen dafür, dass die Münze nicht fair ist, während eine Stichprobe wie `ZKKZK` eher für die Hypothese spricht. Als Teststatistik S wählen wir die Anzahl `K` bei den 5 Würfen. Für die Stichprobe `KKKKK` gilt $S = 5$, für die Stichprobe `ZZZZZ` gilt $S = 0$ und für die Stichprobe `ZKKZK` gilt $S = 3$. □

Wir formulieren auf Basis der Teststatistik eine **Entscheidungsregel**. Diese gibt an, bei welchen Werten von S wir uns für H_0 und bei welchen Werten von S wir uns für H_1 entscheiden. Man nennt die Menge der Werte von S, für die man sich für H_1 entscheidet, auch den **kritischen Bereich** oder **Ablehnbereich** C.

Beispiel 134 (fortgesetzt)
Wir sind nicht bereit zu akzeptieren, dass die Münze fair ist, wenn bei allen 5 Würfen immer `K` oder immer `Z` auftritt. Wir erhalten also folgende Entscheidungsregel:
Entscheidung für H_1, wenn $S = 0$ oder $S = 5$ gilt.
Entscheidung für H_0, wenn $1 \leq S \leq 4$ gilt.
Der kritische Bereich ist also $C = \{0, 5\}$. □

Wir werden im Folgenden bei der Formulierung der Entscheidungsregeln immer nur den kritischen Bereich eines Tests angeben.

Beispiel 134 (fortgesetzt)
Auch wenn die Münze fair ist, kann es passieren, dass bei 5 Würfen 5-mal oder 0-mal `K` beobachtet wird. Auf Grund der Entscheidungsregel entscheiden wir uns in diesen Fällen für die Gegenhypothese. Wir entscheiden uns also dafür, dass die Münze nicht fair ist, obwohl sie fair ist. □

Wie das Beispiel zeigt, ist die Entscheidung bei einem Test fehlerbehaftet. Den im Beispiel begangenen Fehler bezeichnen wir als **Fehler 1. Art**. Ein

Fehler 1. Art wird begangen, wenn man sich für H_1 entscheidet, obwohl H_0 zutrifft. Man kann noch einen weiteren Fehler begehen. Der **Fehler 2. Art** wird begangen, wenn man sich für H_0 entscheidet, obwohl H_1 zutrifft. Tabelle 14.1 stellt die Situation dar.

Tabelle 14.1: Die Fehler beim statistischen Test

Realität / Entscheidung	H_0 trifft zu	H_1 trifft zu
für H_0	richtige Entscheidung	Fehler 2.Art
für H_1	Fehler 1.Art	richtige Entscheidung

Beispiel 135
Ein Statistiker muss sich an Tagen, an denen morgens die Sonne scheint, entscheiden, ob er einen Schirm mitnimmt. Er formuliert also folgende Hypothesen:

H_0 : Es wird am Nachmittag regnen
H_1 : Es wird am Nachmittag nicht regnen

Bei seiner Entscheidungsregel orientiert er sich am Wetterbericht. Wird gutes Wetter vorhergesagt, so nimmt er keinen Schirm mit. Wird Regen prognostiziert, so nimmt er einen Schirm mit.
Wenn er am Morgen keinen Schirm mitgenommen hat, es aber am Nachmittag aber regnet, so begeht er einen Fehler 1. Art. Wenn er am Morgen einen Schirm mitgenommen hat, es am Nachmittag aber nicht regnet, so begeht er einen Fehler 2. Art. □

Die Wahrscheinlichkeit des Fehlers 1. Art ist

$$\alpha = P(\text{Entscheidung für } H_1 | H_0 \text{ trifft zu})$$

Die Wahrscheinlichkeit des Fehlers 2. Art ist

$$\beta = P(\text{Entscheidung für } H_0 | H_1 \text{ trifft zu})$$

Um die Wahrscheinlichkeiten der beiden Fehler bestimmen zu können, benötigt man die Verteilung der Teststatistik, wenn H_0 zutrifft und wenn H_1 zutrifft.

Beispiel 134 (fortgesetzt)
Beim fünfmaligen Münzwurf handelt es sich um einen Bernoulliprozess der Länge $n = 5$. Es gilt $p = P(K)$. Die Teststatistik S ist die Anzahl K. Sie ist binomialverteilt mit den Parametern $n = 5$ und p. Es gilt

$$P(S = s) = \binom{5}{s} p^s (1-p)^{5-s}$$

Trifft H_0 zu, so ist die Münze fair und es gilt $p = 0.5$. Tabelle 14.2 enthält die Verteilung von S für diesen Fall.

Tabelle 14.2: Wahrscheinlichkeitsfunktion der Binomialverteilung mit den Parametern $n = 5$ und $p = 0.5$

s	0	1	2	3	4	5
$P(S = s)$	0.03125	0.15625	0.31250	0.31250	0.15625	0.03125

Es gilt

$$\alpha = P(S = 0) + P(S = 5) = 0.0625\,.$$

Die Wahrscheinlichkeit des Fehlers 2. Art können wir nicht so einfach angeben, da p unendlich viele Werte annehmen kann, wenn H_1 zutrifft. Und wir wissen natürlich nicht, welcher der wahre Wert ist. Nehmen wir aber einmal an, dass die Münze mit Wahrscheinlichkeit 0.8 KOPF zeigt. Tabelle 14.3 enthält die Verteilung von S für diesen Fall.

Tabelle 14.3: Wahrscheinlichkeitsfunktion der Binomialverteilung mit den Parametern $n = 5$ und $p = 0.8$

s	0	1	2	3	4	5
$P(S = s)$	0.00032	0.0064	0.0512	0.2048	0.4096	0.32768

Es gilt

$$\beta \;=\; P(S = 1) + P(S = 2) + P(S = 3) + P(S = 4) = 0.672\,.$$

□

Man will natürlich beide Fehler vermeiden. Dies ist aber nicht möglich, da die Wahrscheinlichkeiten der beiden Fehler voneinander abhängen.

Beispiel 134 (fortgesetzt)
Wir ändern die Entscheidungsregel und entscheiden uns für H_1, wenn $S \leq 1$ oder $S \geq 4$ gilt. Der kritische Bereich ist also $C = \{0, 1, 4, 5\}$. Mit den Zahlen aus Tabelle 14.2 auf Seite 428 erhalten wir

$$\alpha \quad = \quad P(S = 0) + P(S = 1) + P(S = 4) + P(S = 5) = 0.375$$

Die Wahrscheinlichkeit für den Fehler 1. Art ist größer, während die Wahrscheinlichkeit des Fehlers 2. Art sinkt. Mit den Zahlen aus Tabelle 14.3 auf Seite 428 erhalten wir nämlich

$$\beta = P(S = 2) + P(S = 3) = 0.256\,.$$

In Tabelle 14.4 sind die Wahrscheinlichkeiten der Fehler und die kritischen Bereiche zusammengestellt.

Tabelle 14.4: Zusammenhang zwischen den Fehlern beim statistischen Test

	C	$\{0, 5\}$	$\{0, 1, 4, 5\}$
α		0.0625	0.375
β		0.6720	0.256

□

Vergrößern wir also die Wahrscheinlichkeit α für den Fehler 1. Art, so werden wir uns häufiger für H_1 und damit seltener für H_0 entscheiden. Also werden wir auch seltener einen Fehler 2. Art begehen. Vergrößern wir hingegen die Wahrscheinlichkeit β für den Fehler 2. Art, so werden wir die uns häufiger für H_0 und damit seltener für H_1 entscheiden. Also werden wir auch seltener einen Fehler 1. Art begehen.
In vielen Programmpaketen wird bei einem statistischen Test die sogenannte **Überschreitungswahrscheinlichkeit** ausgegeben. Man spricht auch vom **p-Wert**. Diese ist das kleinste Signifikanzniveau, zu dem die Hypothese H_0 für den Datensatz abgelehnt wird.

Beispiel 134 (fortgesetzt)
Wir haben den Wert $S = 4$ beobachtet. Wie groß ist die Überschreitungswahrscheinlichkeit? Wir suchen unter allen kritischen Bereichen, in denen der Wert 4 liegt, den mit dem kleinsten Signifikanzniveau.

Wir lehnen H_0 ab, wenn S zu groß oder zu klein ist. Der kleinste kritische Bereich ist also $C = \{0, 5\}$. Bei diesem ist das Signifikanzniveau gleich $0.03125 + 0.03125 = 0.0625$, wie wir Tabelle 14.2 auf Seite 428 entnehmen können. Da 4 aber nicht im kritischen Bereich liegt, lehnen wir zu diesem Signifikanzniveau nicht ab. Wir vergrößern den kritischen Bereich, indem wir 1 und 4 in den kritischen Bereich nehmen. Es gilt also $C = \{0, 1, 4, 5\}$. Bei diesem ist das Signifikanzniveau gleich

$$0.03125 + 0.15625 + 0.15625 + 0.03125 = 0.375\,.$$

Da 4 in diesem kritischen Bereich liegt, ist die Überschreitungswahrscheinlichkeit gleich 0.375. Vergrößern wir nämlich den kritischen Bereich, so lehnen wir H_0 zwar für $S = 4$ ab, das Signifikanzniveau wird aber auch größer. □

In der Regel gibt es mehrere Tests für dasselbe Testproblem. Diese kann man an Hand der Gütefunktion vergleichen. Die Gütefunktion $G(\theta_1)$ an der Stelle θ_1 ist gleich der Wahrscheinlichkeit, die Hypothese H_0 abzulehnen, wenn θ_1 der Wert von θ ist. Die Gütefunktion $G(\theta)$ sollte mit wachsendem Abstand von θ_0 immer größer werden.

Beispiel 134 (fortgesetzt)
Wir betrachten den Test mit kritischem Bereich $\{0, 5\}$. Wir bestimmen $G(0.8)$ und $G(0.9)$. Mit den Wahrscheinlichkeiten in Tabelle 14.3 auf Seite 428 gilt:

$$G(0.8) = P(S = 0) + P(S = 5) = 0.00032 + 0.32768 = 0.328\,.$$

In Tabelle 14.5 ist die Verteilung von S für $p = 0.9$ zu finden.

Tabelle 14.5: Wahrscheinlichkeitsfunktion der Binomialverteilung mit den Parametern $n = 5$ und $p = 0.9$

s	0	1	2	3	4	5
$P(S = s)$	0.00001	0.00045	0.0081	0.0729	0.32805	0.59049

Also gilt

$$G(0.9) = P(S = 0) + P(S = 5) = 0.00001 + 0.59049 = 0.5905$$

Wir sehen, dass die Wahrscheinlichkeit, uns für H_1 zu entscheiden, für $p = 0.9$ größer ist als für $p = 0.8$. □

Wir haben bisher Hypothesen der Form

$$H_0 : \theta = \theta_0 \quad \text{gegen} \quad H_1 : \theta \neq \theta_0 \,.$$

betrachtet. Bei diesen kann der Parameter θ Werte annehmen, die kleiner oder größer als θ_0 sind, wenn H_1 zutrifft. Man spricht von einem zweiseitigen Testproblem. Einseitige Testprobleme sind von der Form

$$H_0 : \theta = \theta_0 \quad \text{gegen} \quad H_1 : \theta > \theta_0$$

oder

$$H_0 : \theta = \theta_0 \quad \text{gegen} \quad H_1 : \theta < \theta_0 \,.$$

Beispiel 136
Eine Partei will überprüfen, ob ihr Wähleranteil mehr als 40 Prozent beträgt. Hierzu befragt sie 10 Personen, von denen 8 die Partei wählen würden.

$$H_0 : p = 0.4 \quad \text{gegen} \quad H_1 : p > 0.4 \,.$$

Wir wählen als Teststatistik S die Anzahl der Wähler der Partei in der Stichprobe. Diese ist binomialverteilt mit den Parametern $n = 10$ und $p = 0.4$, wenn H_0 zutrifft. In Tabelle 14.6 ist die Verteilung von S unter H_0 zu finden.

Tabelle 14.6: Verteilung von S unter H_0

s	$P(S = s)$
0	0.0060
1	0.0403
2	0.1209
3	0.2150
4	0.2508
5	0.2007
6	0.1115
7	0.0425
8	0.0106
9	0.0016
10	0.0001

Wir lehnen H_0 ab, wenn S zu groß ist. Schauen wir uns an, wie die Wahrscheinlichkeit α des Fehlers 1. Art vom kritischen Bereich C abhängt. Tabelle 14.7 zeigt dies.

Tabelle 14.7: α in Abhängigkeit von C

C	α
$C = \{10\}$	0.0001
$C = \{9, 10\}$	0.0017
$C = \{8, 9, 10\}$	0.0123
$C = \{7, 8, 9, 10\}$	0.0548

Wollen wir zum Signifikanzniveau $\alpha = 0.05$ testen, so ist der kritische Bereich $C = \{8, 9, 10\}$. Dies ist nämlich der größte kritische Bereich, bei dem die Wahrscheinlichkeit des Fehlers 1. Art kleiner gleich 0.05 ist. Der kritische Bereich $C = \{7, 8, 9, 10\}$ enthält zwar auch den Wert 8. Aber bei diesem ist die Wahrscheinlichkeit des Fehlers 1. Art größer als 0.05.

Aus Tabelle 14.7 können wir auch die Überschreitungswahrscheinlichkeit bestimmen. Sie beträgt 0.0123. Dies ist nämlich das kleinste Signifikanzniveau, bei dem wir H_0 für den Wert $S = 8$ ablehnen. Der kritische Bereich $C = \{7, 8, 9, 10\}$ enthält zwar auch den Wert 8, aber das Signifikanzniveau 0.0548 ist hier größer. □

Den im Beispiel betrachteten Test nennt man Test auf p. Schauen wir uns an Hand des zweiseitigen Tests auf p noch einmal die Bestandteile eines Tests an.

1. Die Annahmen.

   Beim Test auf p gehen wir davon aus, dass wir n Realisationen eines Bernoulliprozesses beobachten, bei dem ein Ereignis A mit $p = P(A)$ von Interesse ist.

2. Die Hypothesen H_0 und H_1.

   Beim zweiseitigen Test auf p testen wir

   $$H_0 : p = p_0 \quad \text{gegen} \quad H_1 : p \neq p_0 \,.$$

3. Das Signifikanzniveau α, das vom Anwender vorgegeben wird.

4. Die Teststatistik.

   Beim Test auf p bestimmen wir die absolute Häufigkeit S von A bei den n Realisationen des Bernoulliprozesses.

5. Die Entscheidungsregel.

   Beim Test auf p lehnen wir H_0 ab, wenn gilt $S \leq s_{\alpha/2}$ oder $S \geq s_{1-\alpha/2}$. Dabei wählen wir $s_{\alpha/2}$, so dass gilt $P(S \leq s_{\alpha/2}) \leq \alpha/2$ und $P(S \leq 1 + s_{\alpha/2}) > \alpha/2$. Für $s_{1-\alpha/2}$ gilt $P(S \geq s_{1-\alpha/2}) \leq \alpha/2$ und $P(S \leq s_{1-\alpha/2} - 1) > \alpha/2$. Dabei ist S eine mit den Parametern $n$und p_0 binomialverteilte Zufallsvariable.

Kapitel 15

Das Einstichprobenproblem

Im letzten Kapitel haben wir mit dem Test auf p einen Test im Einstichprobenproblem kennen gelernt. Wir wollen uns in diesem Kapitel weitere Tests für das Einstichprobenproblem anschauen. Bei diesen geht man von den unabhängigen, identisch verteilten Zufallsvariablen $X_1, \ldots, X_n$ aus, wobei die Verteilung von X_i von einem oder mehreren Parametern abhängt. Wir werden in diesem Kapitel zwei Testprobleme betrachten. Beim ersten gehen wir davon aus, dass die Verteilung der X_i zumindest am Median stetig ist, und wollen überprüfen, ob ein Lageparameter der Verteilung einen bestimmten Wert annimmt. Man spricht von **Tests auf einen Lageparameter**. Hier werden wir den ***t*-Test**, den **Vorzeichentest** und den **Wilcoxon-Vorzeichen-Rangtest** kennen lernen. Beim zweiten Testproblem wollen wir überprüfen, ob ein Merkmal in der Grundgesamtheit eine spezielle Verteilung besitzt. Hierzu dienen sogenannte **Anpassungstests**. Von diesen betrachten wir den **Chiquadrat-Anpassungstest**.

15.1 Tests auf einen Lageparameter

Wir gehen davon aus, dass die Beobachtungen $x_1, \ldots, x_n$ Realisationen von unabhängigen und identisch verteilten Zufallsvariablen $X_1, \ldots, X_n$ sind. Wir wollen überprüfen, ob ein Lageparameter der X_i, $i = 1, \ldots, n$ einen speziellen Wert annimmt. Dabei kann der Lageparameter der Erwartungswert μ oder der Median M sein. Im zweiseitigen Testproblem lauten die Hypothesen also

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

oder

$$H_0 : M = M_0 \quad \text{gegen} \quad H_1 : M \neq M_0 \,.$$

Beispiel 137
Die Schmuckstücke an den Kleidungsstücken der Schoschonen sind rechteckig. Ein Forscher will nun untersuchen, ob diese Rechtecke nach dem goldenen Schnitt gefertigt wurden. Ein Rechteck weist den goldenen Schnitt auf, wenn gilt

$$\frac{b}{l} = \frac{l}{b+l}, \tag{15.1}$$

wobei b die Länge der kürzeren und l die Länge der längeren Seite ist. Es muss gelten

$$\frac{b}{l} = \frac{\sqrt{5}-1}{2} \approx 0.618.$$

Aus Gleichung (15.1) folgt nämlich $b^2 + bl = l^2$ und nach Division durch l^2 die Gleichung $(b/l)^2 + b/l = 1$. Die Lösungen dieser quadratischen Gleichung sind $-0.5 \pm 0.5\sqrt{5}$.
Die Schoschonen hatten sicherlich eine Vorstellung von einem ästhetischen Verhältnis von Breite zu Länge bei den Rechtecken und wollten dieses Verhältnis auch erreichen. Aufgrund der Unvollkommenheit der Fertigung werden sie das im Einzelfall aber nicht immer erreicht haben. Die einzelnen Rechtecke streuen um diesen Zielwert. Es soll überprüft werden, ob dieser Wert 0.618 ist. Hierzu bestimmt der Forscher von 20 rechteckigen Schmuckstücken der Schoschonen das Verhältnis von b zu l. Es ergaben sich folgende Zahlen:

```
0.693 0.662 0.690 0.606 0.570 0.749 0.672 0.628 0.609 0.844
0.654 0.615 0.668 0.601 0.576 0.670 0.606 0.611 0.553 0.933
```

□

Beispiel 138
Die Firma MFSL stellte bis Anfang des Jahres 2000 audiophile CDs mit Goldbeschichtung her. Nachdem die Firma Konkurs anmelden musste, wurden diese CDs zu begehrten Sammlerstücken. Ein Statistiker beschließt, seine MFSL GOLD-CD von DARK SIDE OF THE MOON von PINK FLOYD im Ebay zu versteigern. Um eine Vorstellung vom realisierbaren Preis zu erhalten, beobachtet er den Markt. In der zweiten Februarwoche des Jahres 2001 wurden 9 CDs zu folgenden Höchstgeboten in Dollar ersteigert:

```
51 56 57 48 45 61 46 53 59
```

Der Statistiker will seine CD nur ins Ebay stellen, wenn der erwartete Höchstpreis mehr als 50 Dollar beträgt. □

In Abhängigkeit von den Annahmen, die man über die Verteilung der Grundgesamtheit macht, erhält man unterschiedliche Tests.

15.1.1 Der t-Test

Wir gehen von folgender Annahme aus:
Die Zufallsvariablen $X_1, \ldots, X_n$ sind unabhängig und identisch mit den Parametern μ und σ^2 normalverteilt.
Im zweiseitigen Testproblem lauten die Hypothesen:

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

Die Teststatistik ist

$$\boxed{t = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s}} \tag{15.2}$$

Dabei sind $\bar{X}$ der Mittelwert und S die Standardabweichung.
Die Entscheidungsregel lautet:
Wir lehnen H_0 zum Signifikanzniveau α ab, wenn gilt $|t| > t_{n-1;1-\alpha/2}$. Dabei ist $t_{n-1;1-\alpha/2}$ das $1-\alpha/2$-Quantil der t-Verteilung mit $n-1$ Freiheitsgraden.
Die Entscheidungsregel ist auch intuitiv einleuchtend. Ist $\bar{X}$ als erwartungstreuer Schätzer des wahren Wertes von μ zu weit vom hypothetischen Wert μ_0 entfernt, so lehnen wir die Hypothese ab, dass μ_0 der wahre Wert von μ ist. Ob $\bar{x}$ weit genug von μ_0 entfernt ist, hängt von α und von s ab.

Beispiel 137 (fortgesetzt von Seite 436)
Wir fassen das Verhältnis von Breite b zu Länge l als Zufallsvariable X auf, wobei wir unterstellen, dass X mit den Parametern μ und σ^2 normalverteilt ist. Wurden die Rechtecke nach dem goldenen Schnitt gefertigt, so ist 0.618 der Wert, den wir für den Erwartungswert des Verhältnisses der Seiten erwarten. Dieser Wert sollte also das Zentrum der Verteilung bilden. Wir wollen also testen

$$H_0 : \mu = 0.618 \quad \text{gegen} \quad H_1 : \mu \neq 0.618$$

Es gilt $\bar{x} = 0.6605$ und $s = 0.0925$. Also gilt

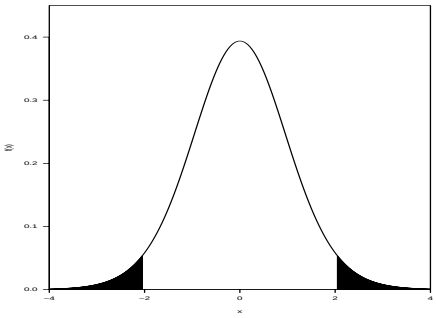
$$t = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s} = \frac{\sqrt{20}(0.6605 - 0.618)}{0.0925} = 2.0545\,.$$

Der Tabelle 20.5 auf Seite 547 entnehmen wir $t_{19;0.975} = 2.093$. Wir lehnen H_0 zum Signifikanzniveau $\alpha = 0.05$ nicht ab, da gilt $|2.0545| < 2.093$. Wir können auch die Überschreitungswahrscheinlichkeit bestimmen. Diese ist das kleinste Signifikanzniveau, zu dem wir für die Daten H_0 ablehnen. Es gilt

$$P(t \leq -2.0545) + P(t \geq 2.0545) = 0.027 + 0.027 = 0.054\,.$$

Dabei ist t eine mit 19 Freiheitsgraden t-verteilte Zufallsvariable. Da die Verteilungsfunktion der t-Verteilung wie die Verteilungsfunktion der Standardnormalverteilung nicht in expliziter Form angegeben werden kann, haben wir R benutzt, um diese Wahrscheinlichkeit zu bestimmen. Abbildung 15.1 veranschaulicht den Wert der Überschreitungswahrscheinlichkeit. Diese ist gleich der Fläche des schraffierten Teils.

Abbildung 15.1: Veranschaulichung der Bestimmung der Überschreitungswahrscheinlichkeit beim zweiseitigen t-Test



□

Schauen wir uns noch einmal die Entscheidungsregel an. Wir lehnen H_0 ab, wenn $|t| > t_{n-1;1-\alpha/2}$ gilt. Also lehnen wir H_0 nicht ab, wenn $|t| \leq t_{n-1;1-\alpha/2}$ gilt. Dieses können wir auch folgendermaßen schreiben:

$$-t_{n-1;1-\alpha/2} \leq t \leq t_{n-1;1-\alpha/2}$$

Es gilt also

$$-t_{n-1;1-\alpha/2} \leq \frac{\sqrt{n}(\bar{x} - \mu_0)}{s} \leq t_{n-1;1-\alpha/2}$$

Multiplizieren wir diese Ungleichung mit $s/\sqrt{n}$, so ergibt sich

$$-t_{n-1;1-\alpha/2} \frac{s}{\sqrt{n}} \leq \bar{x} - \mu_0 \leq t_{n-1;1-\alpha/2} \frac{s}{\sqrt{n}}$$

Wir multiplizieren die Ungleichung mit -1 und erhalten

$$-t_{n-1;1-\alpha/2} \frac{s}{\sqrt{n}} \leq \mu_0 - \bar{x} \leq t_{n-1;1-\alpha/2} \frac{s}{\sqrt{n}}$$

Nun addieren wir $\bar{x}$

$$\bar{x} - t_{n-1;1-\alpha/2} \frac{s}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + t_{n-1;1-\alpha/2} \frac{s}{\sqrt{n}} \,.$$

Wir lehnen also H_0 nicht ab, wenn μ_0 im Intervall

$$\left[\bar{x} - t_{n-1;1-\alpha/2}\,\frac{s}{\sqrt{n}}, \bar{x} + t_{n-1;1-\alpha/2}\,\frac{s}{\sqrt{n}}\right]$$

liegt. Dies ist aber gerade das Konfidenzintervall für μ bei Normalverteilung mit unbekannter Varianz σ^2 zum Konfidenzniveau $1-\alpha$, wie der Vergleich mit Formel (13.20) auf Seite 404 zeigt.
Wir können den zweiseitigen t-Test zum Signifikanzniveau α also auch folgendermaßen durchführen:
Wir stellen das Konfidenzintervall für μ bei Normalverteilung mit unbekannter Varianz σ^2 zum Konfidenzniveau $1-\alpha$ auf. Liegt der hypothetische Wert μ_0 nicht im Konfidenzintervall, so lehnen wir H_0 ab.
Man kann jeden zweiseitigen Test mit einem Konfidenzintervall durchführen. Wir werden dies aber nicht tun, sondern die Entscheidungsregel über eine Teststatistik formulieren.
Bisher haben wir nur den zweiseitigen Test betrachtet. Man kann aber auch einseitige Tests durchführen. Die Hypothesen sind:

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

Die Teststatistik ist

$$t = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s}$$

Die Entscheidungsregel lautet:
Wir lehnen H_0 zum Signifikanzniveau α ab, wenn gilt $t > t_{n-1;1-\alpha}$.
Dabei ist $t_{n-1;1-\alpha}$ das $1-\alpha$-Quantil der t-Verteilung mit $n-1$ Freiheitsgraden.
Auch hier können wir die Entscheidungsregel schön interpretieren. Ist $\bar{x}$ sehr viel größer als μ_0, so lehnen wir H_0 ab.

Beispiel 138 (fortgesetzt von Seite 436)
Er will überprüfen, ob die Daten dafür sprechen, dass der erwartete Höchstpreis mehr als 50 Dollar beträgt. Wir wissen, dass man das, was man zeigen will, als Gegenhypothese H_1 formulieren sollte. Wir erhalten somit folgende Hypothesen:

$$H_0 : \mu = 50 \quad \text{gegen} \quad H_1 : \mu > 50$$

Wir wählen $\alpha = 0.05$. Tabelle 20.5 auf Seite 547 entnehmen wir $t_{8;0.95} = 1.86$.
Aus den Daten bestimmt man $\bar{x} = 52.9$ und $s^2 = 5.78$. Also gilt

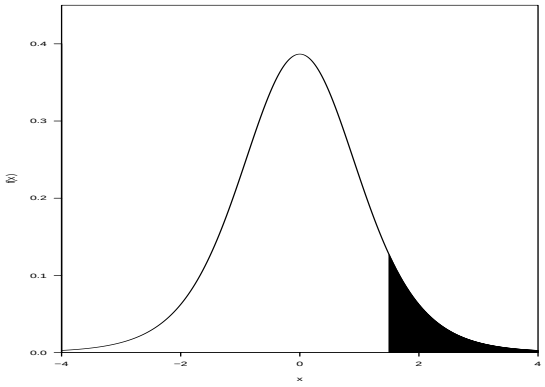
$$t = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s} = \frac{\sqrt{9}(52.9 - 50)}{5.78} = 1.5\,.$$

Wegen $t < t_{8;0.95}$ lehnen wir H_0 nicht ab.
Wir können auch die Überschreitungswahrscheinlichkeit bestimmen. Diese ist

$$P(t \geq 1.5) = 0.086\,.$$

Dabei ist t eine mit 8 Freiheitsgraden t-verteilte Zufallsvariable. In Abbildung 15.2 wird der Wert der Überschreitungswahrscheinlichkeit veranschaulicht. Diese ist gleich der Fläche des schraffierten Teils.

Abbildung 15.2: Veranschaulichung der Bestimmung der Überschreitungswahrscheinlichkeit beim einseitigen t-Test



□

Wir können noch ein anderes einseitiges Testproblem betrachten.
Die Hypothesen sind:

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

Die Teststatistik ist:

$$t = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s}$$

Die Entscheidungsregel lautet:
Wir lehnen H_0 zum Signifikanzniveau α ab, wenn gilt $t < -t_{n-1;1-\alpha}$. Dabei ist $t_{n-1;1-\alpha}$ das $1-\alpha$-Quantil der t-Verteilung mit $n-1$ Freiheitsgraden.
Ist $\bar{x}$ also sehr viel kleiner als μ_0, so lehnen wir H_0 ab.
Man kann auch einseitige Tests mit Konfidenzintervallen durchführen. Hierzu benötigt man aber einseitige Konfidenzintervalle. Mit diesen haben wir uns aber nicht beschäftigt.

15.1.2 Der Vorzeichentest

Der t-Test ist ein Test auf den Erwartungswert einer normalverteilten Grundgesamtheit. Die Annahme der Normalverteilung muss aber nicht immer gerechtfertigt sein. Liegt keine Normalverteilung vor, sollte man einen anderen Test anwenden. Eine Alternative zum t-Test ist der Vorzeichentest. Dieser ist ein Test auf den Median M.

Er beruht auf folgenden Annahmen:

Wir beobachten die Realisationen $x_1, \ldots, x_n$ der Zufallsvariablen $X_1, \ldots, X_n$, die unabhängig und identisch verteilt sind mit Verteilungsfunktion $F_X(x)$, die im Median M stetig ist.

Das Testproblem lautet

$$H_0 : M = M_0 \quad \text{gegen} \quad H_1 : M \neq M_0$$

Wenn M_0 der wahre Wert des Medians M in der Grundgesamtheit ist, so erwarten wir, dass die Hälfte der Beobachtungen größer als M_0 ist. Auf diesem Tatbestand beruht der Vorzeichentest. Wir zählen, wieviele der Beobachtungen größer als M_0 sind. Ist diese Anzahl zu groß oder zu klein, so spricht dies dagegen, dass M_0 der Wert des Medians in der Grundgesamtheit ist. Die folgende Verteilung der Beobachtungen spricht dafür, dass der Median der Grundgesamtheit, aus der die Stichprobe gezogen wurde, gleich 0 ist.

• • • 0 • • •

Die folgende Verteilung der Beobachtungen spricht jedoch dafür, dass der Median der Grundgesamtheit größer als 0 ist.

• 0 • • • • •

Die folgende Verteilung der Beobachtungen spricht dafür, dass der Median der Grundgesamtheit kleiner als 0 ist.

• • • • • 0 •

Die Teststatitik S des Vorzeichentests ist gleich der Anzahl der Beobachtungen, die größer als M_0 sind. Es gilt also

$$\boxed{S = \sum_{i=1}^{n} s(X_i - M_0)}$$

mit

$$s(x) = \begin{cases} 1 & \text{falls } x > 0 \\ 0 & \text{sonst} \end{cases}$$

Um den kritischen Bereich bestimmen zu können, benötigt man die Verteilung von S unter H_0. Beim Vorzeichentest können wir diese leicht herleiten. Auf Grund der Stetigkeit der Verteilungsfunktion in M_0 ist die Wahrscheinlichkeit gleich 0, den Wert M_0 zu beobachten. Also ist eine Beobachtung entweder größer als M_0 oder kleiner als M_0. Wenn $H_0 : M = M_0$ zutrifft, ist die Wahrscheinlichkeit gleich 0.5, dass eine Beobachtung größer als M_0 ist. Auf Grund der Unabhängigkeit der X_i beobachten wir einen Bernoulliprozess der Länge n mit Erfolgswahrscheinlichkeit $p = 0.5$.
Also ist die Anzahl S der Beobachtungen, die größer als M_0 sind, mit den Parametern n und $p = 0.5$ binomialverteilt, wenn H_0 zutrifft. Es gilt also

$$P(S = s) = \binom{n}{s} 0.5^s \, (1 - 0.5)^{n-s} = \binom{n}{s} 0.5^n \, .$$

Die Entscheidungsregel lautet:
Wir lehnen H_0 zum Signifikanzniveau α ab, wenn gilt $S \leq s_{\alpha/2}$ oder $S \geq n - s_{\alpha/2}$.
Dabei gilt $P(S \leq s_{\alpha/2} | H_0) \leq \alpha/2$ und $P(S \leq s_{\alpha/2} + 1 | H_0) > \alpha/2$. Die Werte von s_p für $p = 0.005, 0.01, 0.025, 0.05, 0.1$ sind in Tabelle 20.6 auf Seite 551 zu finden.
Gilt $n > 20$, so können wir die Binomialverteilung durch die Normalverteilung approximieren. Also ist S approximativ normalverteilt mit Erwartungswert $0.5\,n$ und Varianz $0.25\,n$. Wir bilden also folgende standardnormalverteilte Teststatistik:

$$\boxed{Z = \frac{S - 0.5\,n}{0.5\,\sqrt{n}} \, .} \tag{15.3}$$

Wir lehnen H_0 ab, wenn gilt $|Z| \geq z_{1-\alpha/2}$. Dabei ist $z_{1-\alpha/2}$ das $1-\alpha/2$-Quantil der Standardnormalverteilung.
Wir lehnen H_0 also ab, wenn S zu groß oder zu klein ist.

Beispiel 137 (fortgesetzt von Seite 437)
Wir testen

$$H_0 : M = 0.618 \quad \text{gegen} \quad H_1 : M \neq 0.618$$

Die geordnete Stichprobe ist

```
0.553 0.570 0.576 0.601 0.606 0.606 0.609 0.611 0.615 0.628
0.654 0.662 0.668 0.670 0.672 0.690 0.693 0.749 0.844 0.933
```

11 Beobachtungen sind größer als 0.618. Also gilt $S = 11$. Wie wir der Tabelle 20.6 auf Seite 551 entnehmen, gilt $s_{0.025} = 5$. Somit gilt $n - s_{0.025} = 20 - 5 = 15$. Da gilt $5 < 11 < 15$, lehnen wir H_0 also zum Signifikanzniveau $\alpha = 0.05$ nicht ab. □

Es können auch einseitige Tests durchgeführt werden. Im Testproblem

$$H_0 : M = M_0 \quad \text{gegen} \quad H_1 : M > M_0$$

wird H_0 abgelehnt, wenn gilt $S \geq n - s_\alpha$. Die Werte von s_α sind für $\alpha = 0.005, 0.01, 0.025, 0.05, 0.1$ sind in Tabelle 20.6 auf Seite 551 finden.
Für $n > 20$ lehnen wir H_0 ab, wenn für die Teststatistik in Gleichung (15.3) auf Seite 442 $Z \geq z_{1-\alpha}$ gilt. Dabei ist $z_{1-\alpha}$ das $1 - \alpha$-Quantil der Standardnormalverteilung.
Wir lehnen H_0 also ab, wenn S zu groß ist.

Beispiel 138 (fortgesetzt von Seite 439)
Das Testproblem lautet

$$H_0 : M = 50 \quad \text{gegen} \quad H_1 : M > 50$$

Die geordnete Stichprobe ist

```
45 46 48 51 53 56 57 59 61
```

Es gilt $S = 6$. Der Tabelle 20.6 auf Seite 551 entnehmen wir $s_{0.05} = 1$. Also gilt $s_{0.95} = 9 - 1 = 8$. Da gilt $6 < 8$, lehnen wir H_0 zum Signifikanzniveau $\alpha = 0.05$ nicht ab. Wir können aber auch die Überschreitungswahrscheinlichkeit berechnen. Sie beträgt

$$P(S \geq 6) = \binom{9}{6} 0.5^9 + \binom{9}{7} 0.5^9 + \binom{9}{8} 0.5^9 + \binom{9}{9} 0.5^9 = 0.254\,.$$

Da die Überschreitungswahrscheinlichkeit größer als $\alpha = 0.05$ ist, lehnen wir die Nullhypothese zum Signifikanzniveau 0.05 nicht ab. □

Im Testproblem

$$H_0 : M = M_0 \quad \text{gegen} \quad H_1 : M < M_0$$

wird H_0 abgelehnt, wenn gilt $S \leq s_\alpha$. Die Werte von s_α sind für $\alpha = 0.005, 0.01, 0.025, 0.05, 0.1$ sind in Tabelle 20.6 auf Seite 551 finden.

Für $n > 20$ lehnen wir H_0 ab, wenn für die Teststatistik in Gleichung (15.3) auf Seite 442 $Z \leq -z_{1-\alpha}$ gilt. Dabei ist $z_{1-\alpha}$ das $1-\alpha$-Quantil der Standardnormalverteilung.
Wir lehnen H_0 also ab, wenn S zu klein ist.
Die Wahrscheinlichkeit, dass der Wert M_0 in der Stichprobe auftritt, ist auf Grund der Annahme der Stetigkeit der Verteilungsfunktion in M_0 gleich 0. Wird der Wert M_0 aber beobachtet, so sollte dieser Wert aus der Stichprobe entfernt und der Vorzeichentest mit den restlichen Beobachtungen durchgeführt werden. Man spricht in diesem Fall vom **konditionalen Vorzeichentest**.

Beispiel 138 (fortgesetzt von Seite 443)
Der Statistiker will überprüfen, ob der Median des Verkaufspreises mehr als 51 Dollar beträgt. Er testet also

$$H_0 : M = 51 \quad \text{gegen} \quad H_1 : M > 51\,.$$

Der Wert 51 tritt in der Stichprobe auf. Von den restlichen 8 Beobachtungen sind 5 größer als 51. Der Tabelle 20.6 auf Seite 551 entnehmen wir $s_{0.05} = 1$. Somit gilt $n - s_{0.05} = 7$. Wir lehnen die Hypothese zum Signifikanzniveau 0.05 nicht ab. □

15.1.3 Der Wilcoxon-Vorzeichen-Rangtest

Wir betrachten wiederum das Testproblem

$$H_0 : M = M_0 \quad \text{gegen} \quad H_1 : M \neq M_0.$$

Beim Vorzeichentest wird bei jeder Beobachtung nur festgestellt, ob sie größer als M_0 ist. Dabei wird nur unterstellt, dass die Daten aus einer Verteilung stammen, die in M stetig ist. Kann unterstellt werden, dass die Verteilung symmetrisch und stetig ist, so kann man auch die Abstände der Beobachtungen vom hypothetischen Wert M_0 des Medians bei der Entscheidungsfindung berücksichtigen.

Beispiel 139
Es soll getestet werden

$$H_0 : M = 0 \quad \text{gegen} \quad H_1 : M > 0\,.$$

Die Beobachtungen sind

$$x_1 = 0.2 \quad x_2 = 0.5 \quad x_3 = -0.1 \quad x_4 = 0.3\,.$$

Die folgende Abbildung zeigt, dass die Beobachtungen, die größer als 0 sind, weiter von der 0 entfernt sind als die Beobachtung, die kleiner als 0 ist.

□

Wilcoxon hat 1945 vorgeschlagen, die Ränge $R(|x_i - M_0|)$ der Abstände der Beobachtungen von M_0 zu betrachten. Der Rang R_i von x_i gibt an, wie viele Beobachtungen kleiner oder gleich x_i sind.

Beispiel 139 (fortgesetzt von Seite 444)
Die Abstände der Beobachtungen von 0 sind

$$|x_1| = 0.2 \quad |x_2| = 0.5 \quad |x_3| = 0.1 \quad |x_4| = 0.3$$

und die Ränge der Absolutbeträge der Beobachtungen sind

$$R(|x_1|) = 2 \quad R(|x_2|) = 4 \quad R(|x_3|) = 1 \quad R(|x_4|) = 3$$

□

Die Teststatistik W^+ des Wilcoxon-Vorzeichen-Rangtests ist gleich der Summe der Ränge der Absolutbeträge von $x_i - M_0$, bei denen die $x_i - M_0$ postiv sind. Es gilt also

$$\boxed{W^+ = \sum_{i=1}^{n} s(x_i - M_0)\, R(|x_i - M_0|)}$$

Dabei gilt

$$s(x) = \begin{cases} 1 & \text{falls } x > 0 \\ 0 & \text{sonst} \end{cases}$$

Beispiel 139 (fortgesetzt von Seite 445)
Es gilt

$$s(x_1) = 1 \quad s(x_2) = 1 \quad s(x_3) = 0 \quad s(x_4) = 1\,.$$

Also gilt

$$W^+ = 1 \cdot 2 + 1 \cdot 4 + 0 \cdot 1 + 1 \cdot 3 = 9$$

□

Die Verteilung von W^+ unter H_0 kann man für kleine Werte von n durch Auszählen bestimmen. Betrachten wir dazu den Fall $n = 4$. Es gibt $2^4 = 16$ unterschiedlichen Teilmengen der Menge $\{1, 2, 3, 4\}$. Jede dieser Teilmengen beschreibt eine Konfiguration positiver Beobachtungen. So liegt die leere Menge $\emptyset$ vor, wenn keine Beobachtung positiv ist, während $\{2, 3\}$ vorliegt, wenn die zweite und die dritte Beobachtung positiv ist. Alle Möglichkeiten mit dem zugehörigen Wert von W^+ sind in Tabelle 15.1 zu finden.

Tabelle 15.1: Alle Rangkonfigurationen mit dem Wert von W^+ für $n = 4$

Menge	W^+	Menge	W^+	Menge	W^+	Menge	W^+
$\emptyset$	0	$\{4\}$	4	$\{2, 3\}$	5	$\{1, 2, 4\}$	7
$\{1\}$	1	$\{1, 2\}$	3	$\{2, 4\}$	6	$\{1, 3, 4\}$	8
$\{2\}$	2	$\{1, 3\}$	4	$\{3, 4\}$	7	$\{2, 3, 4\}$	9
$\{3\}$	3	$\{1, 4\}$	5	$\{1, 2, 3\}$	6	$\{1, 2, 3, 4\}$	10

Tabelle 15.2 zeigt die Wahrscheinlichkeitsfunktion von W^+ für $n = 4$.

Tabelle 15.2: Wahrscheinlichkeitsfunktion von W^+ für $n = 4$

w	$P(W^+ = w)$
0	0.0625
1	0.0625
2	0.0625
3	0.1250
4	0.1250
5	0.1250
6	0.1250
7	0.1250
8	0.0625
9	0.0625
10	0.0625

Beispiel 139 (fortgesetzt von Seite 445)
Es gilt $W^+ = 9$. Die Überschreitungswahrscheinlichkeit des einseitigen Tests ist also

$$P(W^+ \geq 9) = 0.0625 + 0.0625 = 0.125$$

□

Im Testproblem

$$H_0 : M = M_0 \quad \text{gegen} \quad H_1 : M \neq M_0.$$

lehnen wir H_0 zum Signifikanzniveau α ab, wenn gilt $W^+ \leq w_{\alpha/2}$ oder $W^+ \geq n(n+1)/2 - w_{\alpha/2}$. Die Werte von w_p für $p = 0.005, 0.01, 0.025, 0.05, 0.1$ sind in Tabelle 20.7 auf Seite 552 finden.
Für große Werte von n ist W^+ approximativ normalverteilt mit $E(W^+) = n(n+1)/4$ und $Var(W^+ = n(n+1)(2n+1)/24$. Wir bilden also folgende standardnormalverteilte Teststatistik:

$$\boxed{Z = \frac{S - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}\,.} \tag{15.4}$$

Wir lehnen H_0 ab, wenn gilt $|Z| \geq z_{1-\alpha/2}$. Dabei ist $z_{1-\alpha/2}$ das $1-\alpha/2$-Quantil der Standardnormalverteilung.
Es können auch einseitige Tests durchgeführt werden.
Im Testproblem

$$H_0 : M = M_0 \quad \text{gegen} \quad H_1 : M > M_0$$

wird H_0 zum Signifikanzniveau α abgelehnt, wenn gilt $W^+ \geq n(n+1)/2 - w_\alpha$. Die Werte von w_α sind für $\alpha = 0.005, 0.01, 0.025, 0.05, 0.1$ sind in Tabelle 20.7 auf Seite 552 finden.
Für $n > 20$ lehnen wir H_0 ab, wenn für die Teststatistik in Gleichung (15.4) auf Seite 447 $Z \geq z_{1-\alpha}$ gilt. Dabei ist $z_{1-\alpha}$ das $1-\alpha$-Quantil der Standardnormalverteilung.

Beispiel 138 (fortgesetzt von Seite 444)
Wir testen

$$H_0 : M = 50 \quad \text{gegen} \quad H_1 : M > 50\,.$$

Tabelle 15.3 illustriert die Berechnung.

Tabelle 15.3: Berechnung der Teststatistik des Wilcoxon-Vorzeichen-Rangtests

i	1	2	3	4	5	6	7	8	9
x_i	51	56	57	48	45	61	46	53	59
$x_i - 50$	1	6	7	-2	-5	11	-4	3	9
$\|x_i - 50\|$	1	6	7	2	5	11	4	3	9
$R(\|x_i - 50\|)$	1	6	7	2	5	9	4	3	8
$s(x_i - 50)$	1	1	1	0	0	1	0	1	1

Die Ränge der positiven Beobachtungen sind $1, 6, 7, 9, 3, 8$. Also gilt $W^+ = 34$. Sei $\alpha = 0.05$. In Tabelle 20.7 auf Seite 552 finden wir $w_{0.05} = 8$. Also gilt $n(n+1)/2 - w_{0.05} = 45 - 8 = 39$. Also lehnen wir H_0 nicht ab. □

Im Testproblem

$$H_0 : M = M_0 \quad \text{gegen} \quad H_1 : M < M_0$$

wird H_0 zum Signifikanzniveau α abgelehnt, wenn gilt $W^+ \leq w_\alpha$. Die Werte von w_α sind für $\alpha = 0.005, 0.01, 0.025, 0.05, 0.1$ sind in Tabelle 20.7 auf Seite 552 finden.

Für $n > 20$ lehnen wir H_0 ab, wenn für die Teststatistik in Gleichung (15.4) auf Seite 447 $Z \leq -z_{1-\alpha}$ gilt. Dabei ist $z_{1-\alpha}$ das $1-\alpha$-Quantil der Standardnormalverteilung.

Kommen identische Beobachtungen in der Stichprobe vor, so spricht man von Bindungen.

Beispiel 140

In der Zeit vom 19.3.2003 bis zum 29.3.2003 wurden 10 MFSL CDs von DARK SIDE OF THE MOON im Ebay versteigert. Die Verkaufspreise dieser CDs sind

```
51 64 36 31 31 30 44 44 51 31
```

Es wird übersichtlicher, wenn wir die Stichproben sortieren.

```
30 31 31 31 36 44 44 51 51 64
```

Der Wert 31 kommt dreimal und die Werte 44 und 51 zweimal in der Stichprobe vor. □

Liegen Bindungen vor, so werden Durchschnittsränge bestimmt.

Beispiel 140 (fortgesetzt von Seite 448)
Der Wert 51 erhält den Rang 8.5. Die Ränge aller Beobachtungen sind

```
8.5 10.0 5.0 3.0 3.0 1.0 6.5 6.5 8.5 3.0
```

□

Liegen Bindungen vor, so wird die Entscheidungsregel über die Normalverteilung formuliert. In diesem Fall muss die Varianz von W^+ modifiziert werden. Es gilt

$$Var(W^+) = \frac{n(n+1)(2n+1)}{24} - \frac{1}{48}\sum_{j=1}^{r}\left(b_j^3 - b_j\right)$$

Dabei ist r die Anzahl der Gruppen mit Bindungen und b_j die Anzahl der Beobachtungen in der j-ten Bindungsgruppe.
Wir bilden also folgende standardnormalverteilte Teststatistik:

$$Z = \frac{S - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24 - \frac{1}{48}\sum_{j=1}^{r}\left(b_j^3 - b_j\right)}}. \tag{15.5}$$

Wir lehnen H_0 ab, wenn gilt $|Z| \geq z_{1-\alpha/2}$. Dabei ist $z_{1-\alpha/2}$ das $1-\alpha/2$-Quantil der Standardnormalverteilung. Im Testproblem

$$H_0 : M = M_0 \quad \text{gegen} \quad H_1 : M \neq M_0.$$

lehnen wir H_0 zum Signifikanzniveau α ab, wenn $|Z| \geq z_{1-\alpha/2}$. Dabei ist $z_{1-\alpha/2}$ das $1-\alpha/2$-Quantil der Standardnormalverteilung.
Im Testproblem

$$H_0 : M = M_0 \quad \text{gegen} \quad H_1 : M > M_0.$$

lehnen wir H_0 zum Signifikanzniveau α ab, wenn für die Teststatistik in Gleichung (15.5) $Z \geq z_{1-\alpha}$ gilt. Dabei ist $z_{1-\alpha}$ das $1-\alpha$-Quantil der Standardnormalverteilung.
Im Testproblem

$$H_0 : M = M_0 \quad \text{gegen} \quad H_1 : M < M_0.$$

lehnen wir H_0 zum Signifikanzniveau α ab, wenn für die Teststatistik in Gleichung (15.5) $Z \leq -z_{1-\alpha}$ gilt. Dabei ist $z_{1-\alpha}$ das $1-\alpha$-Quantil der Standardnormalverteilung.

Beispiel 140 (fortgesetzt von Seite 449)
Es soll getestet werden

$$H_0 : M = 50 \quad \text{gegen} \quad H_1 : M > 50\,.$$

Tabelle 15.4 illustriert die Berechnung.

Tabelle 15.4: Berechnung der Teststatistik des Wilcoxon-Vorzeichen-Rangtests

i	1	2	3	4	5	6	7	8	9	10
x_i	51	64	36	31	31	30	44	44	51	31
$x_i - 50$	1	14	−14	−19	−19	−20	−6	−6	1	−19
$\lvert x_i - 50\rvert$	1	14	14	19	19	20	6	6	1	19
$R(\lvert x_i - 50\rvert)$	1.5	5.5	5.5	8	8	10	3.5	3.5	1.5	8
$s(x_i - 50)$	1	1	0	0	0	0	0	0	1	0

Die Ränge der positiven Beobachtungen sind $1.5, 5.5, 1.5$. Also gilt $W^+ = 8.5$. In Tabelle 15.5 sind die Bindungsgruppen der $|x_i - 50|$ mit den Häufigkeiten b_j zu finden.

Tabelle 15.5: Bindungsgruppen

j	1	2	3	4	5
$\lvert x - 50\rvert$	1	6	14	19	20
b_j	2	2	2	3	1

Es gilt

$$\frac{1}{48}\sum_{j=1}^{r}\left(b_j^3 - b_j\right) = \frac{1}{48}\left(2^3 - 2 + 2^3 - 2 + 2^3 - 2 + 3^3 - 3\right) = 0.875$$

Also gilt

$$\sqrt{n(n+1)(2n+1)/24 - \frac{1}{48}\sum_{j=1}^{r}\left(b_j^3 - b_j\right)} = 9.76$$

und somit

$$Z = \frac{8.5 - 27.5}{9.76} = -1.947$$

Wegen $z_{0.975} = 1.96$ lehnen wir H_0 zum Signifikanzniveau $\alpha = 0.05$ nicht ab. □

15.1.4 Praktische Aspekte

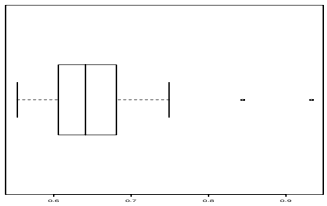
Welchen der drei Tests soll man anwenden?

Ist bekannt, dass die Grundgesamtheit normalverteilt ist, so sollte man auf jeden Fall den t-Test anwenden. Ist die Verteilung symmetrisch, aber nicht normal, so sollte man den Wilcoxon-Vorzeichen-Rangtest anwenden. In allen anderen Fällen kommt der Vorzeichentest zum Zuge.

In der Regel ist nicht bekannt, welche Eigenschaft die Verteilung der Grundgesamtheit besitzt. Hier liefert ein Boxplot wertvolle Hinweise.

Beispiel 137 (fortgesetzt von Seite 442)
Abbildung 15.3 zeigt den Boxplot der Daten.

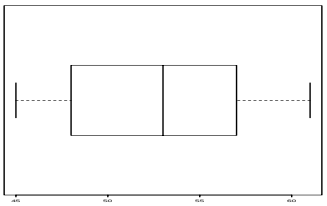
Abbildung 15.3: Boxplot der Daten der Schoschonen



Im Boxplot sind zwei Ausreißer zu erkennen. Also sollte man den Vorzeichentest anwenden. □

Beispiel 138 (fortgesetzt von Seite 447)
Abbildung 15.4 zeigt den Boxplot der Daten.

Abbildung 15.4: Boxplot der Daten aus dem Ebay



Der Boxplot spricht für Normalverteilung. Also sollte man den t-Test anwenden. □

15.2 Anpassungstests

Viele der Tests, die wir betrachtet haben, gehen davon aus, dass in der Grundgesamtheit eine spezielle Verteilung vorliegt. So unterstellen wir beim t-Test im Einstichprobenproblem, dass die Grundgesamtheit normalverteilt ist. Im Einzelfall stellt sich natürlich die Frage, ob diese Annahme erfüllt ist. Wir wollen im Folgenden einen Test betrachten, mit dem man überprüfen kann, ob eine Zufallsstichprobe aus einer speziellen Verteilung stammt. Dies ist der **Chiquadrat-Anpassungstest**.

Beispiel 141
Um herauszufinden, ob ein Würfel fair ist, wirft ein Statistiker ihn 30-mal. Er erhält folgende Stichprobe:

```
5 5 6 3 4 5 1 1 4 5 1 3 1 3 5 4 6 6 4 1 4 3 5 6 5 2 1 5 2 4
```

□

Bei einem Anpassungstest betrachten wir eine Zufallsvariable X und wollen überprüfen, ob diese eine spezielle Verteilung F_0 besitzt. Das Testproblem lautet also

H_0 : Die Zufallsvariable X besitzt die Verteilungsfunktion F_0
H_1 : Die Zufallsvariable X besitzt nicht die Verteilungsfunktion F_0

Beispiel 141 (fortgesetzt von Seite 452)
Sei X die Augenzahl beim einmaligen Wurf eines Würfels. Das Testproblem lautet

$$H_0: \quad P(X=i) = \tfrac{1}{6} \text{ für } i = 1,2,\ldots,6$$
$$H_1: \quad P(X=i) \neq \tfrac{1}{6} \text{ für mindestens ein } i$$

□

Um die Hypothese zu überprüfen, beobachten wir Realisationen $x_1, \ldots, x_n$ der Zufallsvariable X. Ist die Zufallsvariable X diskret, so bestimmen wir die absoluten Häufigkeiten der einzelnen Realisationsmöglichkeiten.

Beispiel 141 (fortgesetzt von Seite 452)
In Tabelle 15.6 sind die Ausprägungsmöglichkeiten und ihre absoluten Häufigkeiten n_i zu finden.

Tabelle 15.6: Ausprägungsmöglichkeiten und ihre absoluten Häufigkeiten n_i

i	1	2	3	4	5	6
n_i	6	2	4	6	8	4

□

Ist die Zufallsvariable stetig, so bilden wir k Klassen und bestimmen die absoluten Häufigkeiten n_i der einzelnen Klassen.

Beispiel 142
Ein Student bestimmt an 25 Tagen die Wartezeit in Sekunden auf die U-Bahn. Er erhält folgende Werte

```
474 513 360  10 405  12 147  89 287 586 524 412  90
 64 355 129 467 186 450 110 325 464 444 342 125
```

Wir bilden 5 gleich große Klassen. Die Untergrenze der i-ten Klasse ist $120 \cdot (i-1)$ und die Obergrenze $120 \cdot i$. In Tabelle 15.7 ist die Häufigkeitsverteilung zu finden.

□

Die absoluten Häufigkeiten n_i sind Realisationen der Zufallsvariablen N_i. Es liegt nahe, die Hypothese H_0 auf Basis dieser absoluten Häufigkeiten zu überprüfen. Womit sollen wir diese vergleichen? Schauen wir uns noch einmal den zweiseitigen t-Test an. Dessen Hypothese lautet $H_0 : \mu = \mu_0$. Um diese Hypothese zu testen, vergleichen wir $\bar{X}$ mit μ_0. Dabei ist μ_0 der Erwartungswert von $\bar{X}$, wenn H_0 zutrifft. Übertragen wir diese Vorgehensweise

Tabelle 15.7: Die Häufigkeitstabelle des Merkmals Wartezeit auf die Straßenbahn

Klasse	1	2	3	4	5
n_i	6	4	4	8	3

auf den Anpassungstest, so müssen wir N_i mit dem Erwartungswert $E(N_i)$ vergleichen, wenn H_0 zutrifft. Ist p_{i0} die Wahrscheinlichkeit, dass X die i-te Merkmalsausprägung bzw. einen Wert in der i-ten Klasse annimmt, wenn H_0 zutrifft, so gilt $E(N_i) = n\,p_{i0}$. Wir bezeichnen diesen mit $\tilde{n}_i$.

Beispiel 141 (fortgesetzt von Seite 453)
Wenn H_0 zutrifft, gilt $p_i = \frac{1}{6}$. Mit $n = 30$ gilt also $\tilde{n}_i = 5$ für $i = 1, \ldots, 6$. □

Beispiel 142 (fortgesetzt von Seite 453)
Wir wollen testen:

H_0 : Die Wartezeit ist im Intervall $[0, 600]$ gleichverteilt.
H_1 : Die Wartezeit ist nicht im Intervall $[0, 600]$ gleichverteilt.

Wenn H_0 zutrifft, gilt $p_i = 0.2$. Mit $n = 25$ gilt also $\tilde{n}_i = 5$ für $i = 1, \ldots, 5$. □

Es liegt nun nahe, für $i = 1, \ldots, k$ die Differenz $n_i - \tilde{n}_i$ aus beobachteter und erwarteter absoluter Häufigkeit zu bilden und diese Differenzen zu summieren. Wir erhalten

$$\begin{aligned}\sum_{i=1}^{k}(n_i - \tilde{n}_i) &= \sum_{i=1}^{k}(n_i - n\,p_{i0}) = \sum_{i=1}^{k} n_i - \sum_{i=1}^{k} n\,p_{i0} \\ &= n - n\sum_{i=1}^{k} p_{i0} = n - n = 0\end{aligned}$$

Diese Statistik liefert also immer den Wert 0. Quadrieren der Differenzen löst dieses Problem:

$$\sum_{i=1}^{k}(n_i - \tilde{n}_i)^2\,.$$

Nun müssen wir diesen Ausdruck geeignet normieren. Warum dies nötig ist, macht man sich folgendemaßen klar. Sei $n_i - \tilde{n}_i = 5$. Ist nun $\tilde{n}_i = 10$, so gilt

$n_i = 15$. Die beobachtete Häufigkeit ist 50 Prozent größer als die erwartete. Für $\tilde{n}_i = 100$ gilt $n_i = 105$. Die beobachtete Häufigkeit ist hier 5 Prozent größer als die erwartete. Für $\tilde{n}_i = 10$ ist die Differenz von 5 also relativ gesehen viel größer als für $\tilde{n}_i = 100$. Diesen Tatbestand berücksichtigen wir, indem wir die Differenzen auf die erwarteten Häufigkeiten beziehen. Wir erhalten folgende Teststatistik:

$$\boxed{X^2 = \sum_{i=1}^{k} \frac{(n_i - \tilde{n}_i)^2 .}{\tilde{n}_i}} . \tag{15.6}$$

Beispiel 141 (fortgesetzt von Seite 454)
Es gilt

$$X^2 = \frac{(6-5)^2}{5} + \frac{(2-5)^2}{5} + \frac{(4-5)^2}{5} + \frac{(6-5)^2}{5} + \frac{(8-5)^2}{5} + \frac{(4-5)^2}{5} = 4.4$$

□

Beispiel 142 (fortgesetzt von Seite 454)
Es gilt

$$X^2 = \frac{(6-5)^2}{5} + \frac{(4-5)^2}{5} + \frac{(4-5)^2}{5} + \frac{(8-5)^2}{5} + \frac{(3-5)^2}{5} = 3.2$$

□

Wenn H_0 nicht zutrifft, so werden sich die beobachteten Häufigkeiten stark von den erwarteten Häufigkeiten unterscheiden. Die quadrierten Differenzen werden also groß sein. Also wird auch der Wert von X^2 groß sein. Wir lehnen also H_0 ab, wenn X^2 zu groß ist. Der kritische Wert hängt wie immer vom Signifikanzniveau α und der Verteilung der Teststatistik unter H_0 ab. Für kleine Stichprobenumfänge ist es sehr mühselig, diese Verteilung zu bestimmen. Gilt aber $\tilde{n}_i \geq 5$, so ist X^2 approximativ chiquadratverteilt mit $k-1$ Freiheitsgraden.
Die Entscheidungsregel ist:
Wir lehnen H_0 also zum Signifikanzniveau α ab, wenn gilt $X^2 \geq \chi^2_{k-1;1-\alpha}$. Dabei ist $\chi^2_{k-1;1-\alpha}$ das $1-\alpha$-Quantil der Chiquadratverteilung mit $k-1$ Freiheitsgraden. Die Quantile der Chiquadratverteilung sind in Tabelle 20.4 auf Seite 550 zu finden.

Beispiel 141 (fortgesetzt von Seite 455)
Es gilt $k = 6$ und $\alpha = 0.05$. Also gilt $\chi^2_{5;0.95} = 11.07$. Wir lehnen H_0 zum Signifikanzniveau $\alpha = 0.05$ nicht ab. □

Beispiel 142 (fortgesetzt von Seite 455)
Es gilt $k = 5$ und $\alpha = 0.05$. Also gilt $\chi^2_{4;0.95} = 9.49$. Wir lehnen H_0 zum Signifikanzniveau $\alpha = 0.05$ nicht ab. □

Bisher sind wir davon ausgegangen, dass alle Parameter der hypothetischen Verteilung bekannt sind. Ist dies nicht der Fall, so schätzt man diese und bestimmt die erwarteten Häufigkeiten, indem man die geschätzten Parameter in die Verteilungsfunktion einsetzt. Man erhält also geschätzte erwartete Häufigkeiten. Bei der Entscheidungsregel wird nun berücksichtigt, dass Parameter geschätzt wurden. Für jeden geschätzten Parameter wird die Anzahl der Freiheitsgrade um 1 vermindert. H_0 wird also abgelehnt, wenn gilt $X^2 \geq \chi^2_{k-1-m;1-\alpha}$, wobei m die Anzahl der geschätzten Parameter ist.

Beispiel 143
Es soll überprüft werden, ob die Körpergröße von männlichen Studienanfängern normalverteilt ist. In der zweiten Spalte von Tabelle 15.8 ist die Verteilung der Körpergröße zu finden.

Tabelle 15.8: Die Häufigkeitstabelle der Körpergröße der Männer

Alter	n_i	$\tilde{n}_i$	$n_i - \tilde{n}_i$	$(n_i - \tilde{n}_i)^2$	$\frac{(n_i-\tilde{n}_i)^2}{\tilde{n}_i}$
von 165 bis unter 170	5	4.24	0.76	0.5776	0.136
von 170 bis unter 175	12	14.51	−2.51	6.3001	0.434
von 175 bis unter 180	32	35.34	−3.34	11.1556	0.316
von 180 bis unter 185	65	51.71	13.29	176.6241	3.416
von 185 bis unter 190	35	45.94	−10.94	119.6836	2.605
von 190 bis unter 195	25	24.55	0.45	0.2025	0.008
von 195 bis unter 200	12	9.67	2.33	5.4289	0.561

Aus den klassierten Daten schätzen wir $\bar{x} = 183.8$ und $s = 6.9$. Die geschätzte Wahrscheinlichkeit der ersten Klasse ist somit

$$\hat{p}_1 \;=\; P(X \leq 170) = \Phi\left(\frac{170 - 183.8}{6.9}\right) = \Phi(-2) = 0.0228$$

Also gilt

$$\tilde{n}_1 = n\,\hat{p}_1 = 186 \cdot 0.0228 = 4.24\,.$$

Analog erhalten wir

$$\hat{p}_2 = 0.078 \quad \hat{p}_3 = 0.190 \quad \hat{p}_4 = 0.278 \quad \hat{p}_5 = 0.247 \quad \hat{p}_6 = 0.132 \quad \hat{p}_7 = 0.052$$

Es gilt $X^2 = 7.476$. Die Anzahl der Freiheitsgrade ist $7 - 1 - 2 = 4$. Wegen $\chi^2_{4;0.95} = 9.49$ lehnen wir H_0 zum Signifikanzniveau $\alpha = 0.05$ nicht ab. □

Beispiel 137 (fortgesetzt von Seite 451)
Wir wollen überprüfen, ob die Rechtecke der Schoschonen aus einer normalverteilten Grundgesamtheit kommen. Die Hypothesen lauten

H_0 : Die Grundgesamtheit ist normalverteilt
H_1 : Die Grundgesamtheit ist nicht normalverteilt

Wir bilden so 4 Klassen, dass wir in jeder Klasse die gleiche Anzahl von Beobachtungen erwarten. Die Klassen sind also

$$(-\infty, x_{0.25}] \quad (x_{0.25}, x_{0.5}] \quad (x_{0.5}, x_{0.75}] \quad (x_{0.25}, \infty)$$

Dabei gilt $x_p = \mu + z_p\,\sigma$. Da die Parameter μ und σ^2 unbekannt sind, schätzen wir sie durch $\bar{x}$ und s^2 aus den Daten. Es gilt $\bar{x} = 0.6605$, $s^2 = 0.0086$ und somit $s = 0.093$. Wir erhalten somit folgende geschätzte Klassengrenzen

$$\begin{aligned} x_{0.25} &= \bar{x} + z_{0.25}\,s = 0.6605 + (-0.6745) \cdot 0.093 = 0.5978 \\ x_{0.50} &= \bar{x} + z_{0.5}\,s = 0.6605 \\ x_{0.75} &= \bar{x} + z_{0.75}\,s = 0.6605 + 0.6745 \cdot 0.093 = 0.7232 \end{aligned}$$

Wir bestimmen aus dem geordneten Datensatz

```
0.553 0.570 0.576 0.601 0.606 0.606 0.609 0.611 0.615 0.628
0.654 0.662 0.668 0.670 0.672 0.690 0.693 0.749 0.844 0.933
```

die absoluten Häufikeiten n_i der Klassen. Es gilt $n_1 = 3$, $n_2 = 8$, $n_3 = 6$ und $n_4 = 3$. Auf Grund der Wahl der Klassen gilt $\tilde{n}_i = 5$ für $i = 1, 2, 3, 4$. Somit gilt

$$X^2 = \frac{(3-5)^2}{5} + \frac{(8-5)^2}{5} + \frac{(6-5)^2}{5} + \frac{(3-5)^2}{5} = 3.6$$

Die Anzahl der Freiheitsgrade ist $4-1-2 = 1$, da wir 2 Parameter geschätzt haben. Wegen $\chi^2_{1;0.95} = 3.84$ lehnen wir H_0 zum Niveau $\alpha = 0.05$ nicht ab. □

15.3 Das Einstichprobenproblem in R

In R können wir alle Tests durchführen, die wir in diesem Kapitel kennengelernt haben. Schauen wir uns zunächst die Tests auf einen Lageparameter an. Wir betrachten die Daten der Schoschonen im Beispiel 137 auf Seite 436 und die Daten aus Ebay im Beispiel 138 auf Seite 436. Wir weisen die Daten der Schoschonen der Variablen `shosho` zu:

```
> shosho<-c(0.693,0.662,0.690,0.606,0.570,0.749,0.672,0.628)
> shosho<-c(shosho,0.609,0.844,0.654,0.615,0.668,0.601)
> shosho<-c(shosho,0.576,0.670,0.606,0.611,0.553,0.933)
```

Wir geben die Daten am Rechner natürlich in einer Zeile ein. Das Papier ist aber leider nicht breit genug, sodass die Eingabe zerlegt wird.
Die Daten aus Ebay weisen wir der Variablen `ebay` zu:

```
> ebay<-c(51,56,57,48,45,61,46,53,59)
```

Mit der Funktion `t.test` kann man einen t-Test durchführen. Sie wird folgendermaßen aufgerufen:

```
t.test(x,y=NULL,alternative=c("two.sided","less", "greater"),
       mu=0,paired=FALSE,var.equal=FALSE,conf.level = 0.95,...)
```

Die Argumente `y`, `paired` und `var.equal` sind im Zweistichprobenproblem relevant. Mit diesem beschäftigen wir uns aber erst im nächsten Kapitel. Das Argument `x` enthält den Datensatz und muss beim Funktionsaufruf angegeben werden. Im Argument `alternative` gibt man an, wie die Gegenhypothese formuliert ist. Dabei steht `"two.sided"` für $\neq$, `"less"` für $<$ und `"greater"` für $>$. Standardmäßig wird ein zweiseitiger Test durchgeführt. Die Funktion `t.test` erstellt auch ein Konfidenzintervall für μ. Mit dem Argument `conf.level` kann man das Konfidenzniveau festlegen. Standardmäßig wird ein Konfidenzintervall für μ bei Normalverteilung mit unbekannter Varianz σ^2 zum Konfidenzniveau 0.95 aufgestellt.
Wir rufen die Funktion `t.test` mit der Variablen `shosho` auf. Außerdem müssen wir dem Argument `mu` den Wert 0.618 zuweisen.

```
> t.test(x=shosho,mu=0.618)

        One Sample t-test

 data:  shosho
```

```
t = 2.0545, df = 19, p-value = 0.05394
alternative hypothesis: true mean is not equal to 0.618
95 percent confidence interval:
 0.6172036 0.7037964
sample estimates: mean of x
  0.6605
```

Wir erhalten den Wert der Teststatistik `t=2.0545`, die Anzahl der Freiheitsgrade `df=19` und noch die Überschreitungswahrscheinlichkeit `p-value = 0.05494`. Außerdem wird noch das Konfidenzintervall $[0.6172036, 0.7037964]$ für μ zum Konfidenzniveau 0.95 und $\bar{x} = 0.6605$ ausgegeben.
Führen wir noch den einseitigen Test für die Daten aus Ebay durch.

```
> t.test(ebay,mu=50,alternative="greater")

        One Sample t-test

data:  ebay
t = 1.5005, df = 8, p-value = 0.08594
alternative hypothesis: true mean is greater than 50
95 percent confidence interval:
49.3087     Inf
sample estimates: mean of x
 52.88889
```

Der Vorzeichentest ist ein spezieller Test auf p. Einen Test auf p kann man in R mit der Funktion `binom.test` durchführen. Sie wird folgendermaßen aufgerufen:

```
binom.test(x,n,p=0.5,alternative=c("two.sided","less",
           "greater"),conf.level=0.95)
```

Dabei ist `x` der Wert der Teststatistik, `n` der Stichprobenumfang und `p` der hypothetischen Wert p_0. Dieser ist standardmäßig 0.5. Außerdem kann man wie beim t-Test die Alternativhypothese mit dem Parameter `alternative` spezifizieren. Es wird auch ein Konfidenzintervall für p aufgestellt. Das Konfidenzniveau übergibt man im Parameter `conf.level`. Es ist standardmäßig 0.95.
Im Beispiel 136 auf Seite 431 sollte überprüft werden, ob der Wähleranteil einer Partei mehr als 0.4 beträgt. Von 10 befragten Personen würden 8 die Partei wählen. Es gilt also `x`=8, `n`=10, `p`=0.4 und `alternative="greater"`. Wir geben also ein

```
> binom.test(8,10,0.4,alternative="greater")
```

und erhalten folgendes Ergebnis

```
        Exact binomial test

data:  8 and 10
number of successes = 8, number of trials = 10,
p-value = 0.01229
alternative hypothesis:
true probability of success is greater than 0.4
95 percent confidence interval:
0.4930987 1.0000000
sample estimates:
probability of success
  0.8
```

Die Überschreitungswahrscheinlichkeit ist 0.0129.
Wir können die Funktion `binom.test` auch für den Vorzeichentest verwenden. Beginnen wir mit den Daten der Schoschonen. Wir wollen überprüfen, ob der Median gleich 0.618 ist. Wir zählen, wieviele Beobachtungen größer also 0.618 sind.

```
> S<-sum(shosho>0.618)
> S
[1] 11
```

und rufen mit diesem Wert die Funktion `binom.test` auf, wobei `n` gleich der Länge von `shosho` und `p` gleich 0.5 ist. Da der Parameter `p` standardmäßig auf 0.5 steht, müssen wir ihn beim Aufruf der Funktion nicht eingeben.

```
> binom.test(S,length(shosho))

        Exact binomial test

data:  S and length(shosho)
number of successes = 11, number of trials = 20,
p-value = 0.8238
alternative hypothesis:
true probability of success is not equal to 0.5
95 percent confidence interval:
0.3152781 0.7694221
```

```
 sample estimates:
 probability of success
                   0.55
```

Die Überschreitungswahrscheinlichkeit ist 0.8238.
Schauen wir uns noch die Daten aus Ebay an. Wir bestimmen den Wert der Teststatistik

```
> S<-sum(ebay>50)
> S
[1] 6
```

und rufen die Funktion `binom.test` auf

```
> binom.test(S,length(ebay),alternative="greater")

        Exact binomial test

 data:  S and length(ebay)
 number of successes = 6, number of trials = 9,
 p-value = 0.2539
 alternative hypothesis:
 true probability of success is greater than 0.5
 95 percent confidence interval:
 0.3449414 1.0000000
 sample estimates:
 probability of success
             0.6666667
```

Um den konditionalen Vorzeichentest auf $M = 51$ im Beispiel 138 auf Seite 444 durchführen zu können, müssen wir den Wert 51 aus der Stichprobe entfernen.

```
> ebayneu<-ebay[ebay!=51]
> ebayneu
[1] 56 57 48 45 61 46 53 59
```

Es folgt die übliche Vorgehensweise.

```
> S<-sum(ebayneu>51)
> S
[1] 5
> binom.test(S,length(ebayneu),alternative="greater")
```

```
        Exact binomial test

 data:  S and length(ebayneu)
 number of successes = 5, number of trials = 8,
 p-value = 0.3633
 alternative hypothesis:
 true probability of success is greater than 0.5 95 percent
 confidence interval:
 0.2892408 1.0000000
 sample estimates:
 probability of success
                 0.625
```

Mit der Funktion `wilcox.test` kann man einen Wilcoxon-Vorzeichen-Rangtest durchführen. Sie wird folgendermaßen aufgerufen:

```
wilcox.test(x,y=NULL,alternative=c("two.sided","less",
            "greater"),mu=0,paired=FALSE,exact=NULL,
            correct=TRUE,conf.int=FALSE,
            conf.level=0.95,...)
```

Die Argumente `y` und `paired` sind im Zweistichprobenproblem relevant. Das Argument `x` enthält den Datensatz und muss beim Funktionsaufruf angegeben werden. Im Argument `alternative` gibt man an, wie die Gegenhypothese formuliert ist. Dabei steht `"two.sided"` für $\neq$, `"less"` für $<$ und `"greater"` für $>$. Standardmäßig wird ein zweiseitiger Test durchgeführt. Mit der Funktion `wilcox.test` kann man auch ein Konfidenzintervall für den Median aufstellen. Hierzu muss man das Argument `conf.int` auf `TRUE` setzen. Mit dem Argument `conf.level` kann man das Konfidenzniveau festlegen. Standardmäßig wird ein Konfidenzintervall zum Konfidenzniveau 0.95 aufgestellt. Wird das Argument `exact` auf `FALSE` gesetzt, so wird nicht die exakte Verteilung von W^+ unter H_0 bestimmt. Wird das Argument `correct` auf `TRUE` gesetzt, so wird mit der Stetigkeitskorrektur gearbeitet.
Wir beginnen mit den Daten aus Ebay, die in `ebay` stehen.

```
> wilcox.test(ebay,mu=50,alternative="greater")

        Wilcoxon signed rank test

 data:  ebay
 V = 34, p-value = 0.1016
 alternative hypothesis: true mu is greater than 50
```

Der Wert von W^+ steht in der Variablen `V`. Die Überschreitungswahrscheinlichkeit ist 0.1016.
Schauen wir uns noch die Daten mit Bindungen im Beispiel 140 auf Seite 448 an. Wir geben sie ein

```
> ebay03<-c(51,64,36,31,31,30,44,44,51,31)
```

und führen den Test durch

```
> wilcox.test(ebay03,mu=50,alternative="greater")

        Wilcoxon signed rank test with continuity correction

 data:  ebay03
 V = 8.5, p-value = 0.977
 alternative hypothesis: true mu is greater than 50

 Warning message: Cannot compute exact p-value with ties in:
 wilcox.test.default(ebay03, mu = 50,
```

Der Wert von W^+ steht in der Variablen `V`. Die Überschreitungswahrscheinlichkeit ist 0.977. `R` weist uns darauf hin, dass nicht die exakte Verteilung benutzt wurde, da Bindungen (ties) vorliegen.
Mit dem von Streitberg und Röhmel entwickelten Shift-Algorithmus ist es möglich, den exakten Wilcoxon-Vorzeichen-Rangtest auch bei Bindungen durchzuführen. Dieser ist in dem von Torsten Hothorn und Kurt Hornik erstellten Paket `exactRankTests` in der Funktion `wilcox.exact` implementiert. Dieses Paket muss man zunächst installieren und laden. Wie man dabei vorzugehen hat, wird auf Seite 52 beschrieben.
Die Funktion `wilcox.exact` wird genauso wie die Funktion `wilcox.test` aufgerufen. Wir geben also ein

```
> wilcox.exact(ebay03,mu=50,alternative="greater")

        Exact Wilcoxon signed rank test

 Data:  ebay03
 V= 8.5, p-value = 0.9785
 alternative hypothesis: true mu is greater than 50
```

Die exakte Überschreitungswahrscheinlichkeit beträgt 0.977 und approximative 0.9785. Die Approximation ist also sehr gut.

Den Chiquadrattest führt man in R mit der Funktion `chisq.test` durch. Die Argumente der Funktion `chisq.test` sind die beobachteten und erwarteten Häufigkeiten. Wir müssen die Daten also entsprechend aufbereiten. Beginnen wir mit dem Beispiel 141 auf Seite 452. Wir geben die Daten ein:

```
> wuerfel<-c(5,5,6,3,4,5,1,1,4,5,1,3,1,3,5,4)
> wuerfel<-c(wuerfel,6,6,4,1,4,3,5,6,5,2,1,5,2,4)
```

Wir erzeugen die absoluten Häufigkeiten n_i mit der Funktion `table`:

```
> ni<-table(wuerfel)
> ni
 wuerfel
 1 2 3 4 5 6
 6 2 4 6 8 4
```

Die erwartete Häufigkeit jeder Merkmalsausprägung ist 5.

```
> chisq.test(ni)

        Chi-squared test for given probabilities

 data:  ni
 X-squared = 4.4, df = 5, p-value = 0.4934
```

Um die Daten der Schoschonen auf Normalverteilung zu testen, gehen wir folgendermaßen vor. Wir bestimmen zunächst die Grenzen der Klassen. Wir benötigen $z_{0.25}, z_{0.5}, z_{0.75}$. Diese erhalten wir durch

```
> z<-qnorm(c(0.25,0.5,0.75))
> z
[1] -0.6744898  0.0000000  0.6744898
```

Die Klassengrenzen sind:

```
> breaks<-mean(shosho)+z*sd(shosho)
> breaks
[1] 0.5981024 0.6605000 0.7228976
```

Wir erweitern diese noch um die Untergrenze der ersten Klasse und die Obergrenze der letzten Klasse. Als Untergrenze der ersten Klasse wählen wir $\min\{x_1, \ldots, x_n\}$. Als Obergrenze der letzten Klasse wählen wir $\max\{x_1, \ldots, x_n\}$.

```
> breaks<-c(min(shosho),breaks,max(shosho))
> breaks
[1] 0.5530000 0.5981024 0.6605000 0.7228976 0.9330000
```

Wir erhalten 4 Klassen und können mit der Funktion `cut` bestimmen, in welche Klassen die Beobachtungen fallen.

```
> > k<-cut(shosho,breaks,labels=1:4,right=FALSE,include.lowest=T)
> k
 [1] 3 3 3 2 1 4 3 2 2 4 2 2 3 2 1 3 2 2 1 4
Levels: 1 2 3 4
```

Die beobachteten Häufigkeiten erhalten wir durch

```
> ni<-table(k)
> ni
 k
 1 2 3 4
 3 8 6 3
```

Die erwarteten Häufigkeiten sind

```
> nis<-rep(5,4)
> nis
[1] 5 5 5 5
```

Der Wert von X^2 ist

```
> x2<-sum(((ni-nis)^2)/nis)
> x2
[1] 3.6
```

Die Anzahl der Freiheitsgrade ist 1. Somit ist die Überschreitungswahrscheinlichkeit gleich

```
> 1-pchisq(x2,1)
[1] 0.05777957
```

Den Test auf Normalverteilung kann man aber auch mit der Funktion `pearson.test` aus dem Paket `nortest` durchführen. Dieses Paket muss man zunächst installieren und laden. Wie man dabei vorzugehen hat, wird auf Seite 52 beschrieben.
Der Funktion `pearson.test` übergibt man im ersten Argument `x` den Datensatz und im zweiten Argument `n.classes` die Anzahl der Klassen, die

gebildet werden sollen. Dabei werden die Klassengrenzen so gewählt, dass die erwartete Anzahl in jeder Klasse gleich groß ist. Wir wenden die Funktion `pearson.test` auf den Datensatz `shosho` an.

```
> pearson.test(shosho,4)

        Pearson chi-square normality test

data:  shosho
P = 3.6, p-value = 0.05778
```

Kapitel 16

Das Zweistichprobenproblem

In vielen Anwendungen will man überprüfen, ob sich zwei oder mehr Verfahren, Behandlungen oder Methoden in ihrer Wirkung auf eine Variable unterscheiden. Wir werden im Folgenden von Verfahren sprechen und nur zwei Verfahren berücksichtigen. Um die Verfahren in ihrer Wirkung auf eine Variable zu vergleichen, wendet man sie bei Personen bzw. Objekten an und bestimmt den Wert der interessierenden Variable. Da man nur an der Wirkung der beiden Verfahren interessiert ist, wird man versuchen, den Einfluss aller anderen Variablen auf die interessierende Variable möglichst auszuschließen. Wenn diese sich nicht ändern, dann sollte sich auch die interessierende Variable nicht ändern. Man wird also versuchen, alle anderen Einflussgrößen konstant zu halten. Dies kann man zum Beispiel dadurch erreichen, dass man beide Verfahren an demselben Objekt bzw. an derselben Person anwendet. Man spricht von **Blockbildung** und bezeichnet die Objekte bzw. Personen als **Blöcke**. Man wendet jedes der beiden Verfahren in jedem Block an und spricht in diesem Fall von einem **verbundenen Zweistichprobenproblem**.

Beispiel 144
Es soll untersucht werden, ob das Zusammensein mit einer anderen Ratte die Herzfrequenz HF (in Schlägen pro Minute) gegenüber dem Alleinsein verändert. Dazu wurde die Herzfrequenz von 10 Ratten bestimmt, während sie allein und während sie mit einer anderen Ratte zusammen waren. Die Werte sind in der Tabelle 16.1 auf der nächsten Seitezu finden.

Es soll nun untersucht werden, ob sich die Herzfrequenz erhöht, wenn die Ratten nicht allein sind. Wir wollen also überprüfen, ob das Zusammensein mit einer anderen Ratte die Herzfrequenz erhöht. Sei X die Herzfrequenz einer Ratte, wenn sie mit einer anderen Ratte zusammen ist, und Y die Herzfrequenz einer Ratte, wenn sie allein ist. Wir testen also

$$H_0 : E(X) = E(Y) \quad \text{gegen} \quad H_1 : E(X) > E(Y)\,.$$

Tabelle 16.1: Herzfrequenz von Ratten

i	1	2	3	4	5	6	7	8	9	10
zusammen	523	494	461	535	476	454	448	408	470	437
allein	463	462	462	456	450	426	418	415	409	402

Quelle: B.Latane & Cappell (1972)

□

Oft ist es nicht möglich, beide Verfahren an derselben Person bzw. demselben Objekt zu betrachten. Will man zum Beispiel zwei Unterrichtsmethoden vergleichen, so kann man nicht eine Person zuerst nach der einen und dann nach der anderen Methode unterrichten. Beim Beginn des Unterrichts nach der zweiten Methode ist die Ausgangssituation nicht die gleiche. In diesem Fall muss man anders vorgehen. Um die Wirkung aller anderen Einflussfaktoren auszuschließen, teilt man $N = m + n$ Personen bzw. Objekte zufällig auf zwei Gruppen auf. Die m Personen bzw. Objekte der ersten Gruppe werden mit dem einen Verfahren, die n Personen bzw. Objekte der zweiten Gruppe mit dem anderen Verfahren behandelt. Durch die zufällige Aufteilung versucht man sicherzustellen, dass sich alle anderen Einflussgrößen gleichmäßig auf die beiden Gruppen verteilen.

Die zufällige Aufteilung auf die beiden Gruppen nennt man auch **Randomisierung**. Eine Verletzung des Prinzips der Randomisierung würde vorliegen, wenn die eine Gruppe nur aus Frauen und die andere nur aus Männern bestehen würde. Wird ein Unterschied zwischen den beiden Gruppen beobachtet, so weiß man nicht, ob dieser an den Verfahren oder am Geschlecht liegt.

Beispiel 145

Wenn man ein neugeborenes Kind so hochhält, dass seine Füße eine flache Oberfläche berühren, so werden die Füße Gehbewegungen machen. Man spricht vom Gehreflex. Wenn die Fußrücken des Neugeborenen gegen den Rand einer flachen Oberfläche gehalten werden, so führt das Kind eine Platzierungsbewegung wie ein junges Kätzchen durch. Man spricht vom Platzierungsreflex. Diese Reflexe verschwinden nach 8 Wochen. Sie können aber durch aktives Einüben beibehalten werden.

Es soll nun untersucht werden, ob dieses Einüben dazu führt, dass die Kinder früher laufen lernen. Hierzu wurde eine Gruppe von 12 männlichen Kleinkindern, die eine Woche alt waren, zufällig auf zwei Gruppen mit jeweils 6 Kleinkindern aufgeteilt. In der ersten Gruppe wurden die Reflexe aktiv eingeübt,

in der zweiten Gruppe nicht. Es handelt sich also um ein unverbundenes Zweistichprobenproblem.
Bei jedem Kind wurde das Alter (in Monaten) bestimmt, in dem es laufen konnte. Die Zeiten in der ersten Gruppe sind:

```
9 9.5 9.75 10 13 9.5
```

Die Zeiten in der zweiten Gruppe sind:

```
11.5 12 9 11.5 13.25 13
```

(Quelle: Zelzano, Zelzano & Kolb (1972))
Wir bezeichnen die Zeit, die ein Kind aus der ersten Gruppe benötigt, um Laufen zu lernen, mit X, und Zeit, die ein Kind aus der zweiten Gruppe benötigt, um Laufen zu lernen, mit Y. Es soll getestet werden

$$H_0 : E(X) = E(Y) \quad \text{gegen} \quad H_1 : E(X) < E(Y)\,.$$

□

Manchmal sind die Gruppen vorgegeben, sodass ein unverbundenes Zweistichprobenproblem auf natürliche Art und Weise entsteht.

Beispiel 146
Im Rahmen der PISA-Studie wurde auch der Zeitaufwand der Schüler für Hausaufgaben erhoben. Dort wird unterschieden zwischen sehr geringem, geringem, mittlerem, großem und sehr großem Aufwand. Wir fassen die Länder mit sehr geringem und geringem Aufwand und die Länder mit großem und sehr großem Aufwand zusammen. Somit liegen drei Gruppen vor. Von diesen betrachten wir zunächst die Länder mit geringem und die Länder mit hohem Zeitaufwand. Wir wollen untersuchen, ob sich die Punkte im Bereich Lesekompetenz zwischen den Ländern mit geringem Aufwand und den Ländern mit hohem Aufwand unterscheiden.
In den Ländern mit geringem Aufwand wurden folgende Punktezahlen erzielt

```
546 522 483 441 507 516 494 492
```

In den Ländern mit hohem Aufwand wurden folgende Punktezahlen erzielt

```
474 523 527 487 458 422 479 462 493 480
```

(Quelle: Deutsches PISA-Konsortium (Hrsg.) (2001))
Sei X die Punktezahl eines Landes mit geringem Aufwand und Y die Punktezahl eines Landes mit hohem Aufwand, so lautet das Testproblem

$$H_0 : E(X) = E(Y) \quad \text{gegen} \quad H_1 : E(X) \neq E(Y)\,.$$

□

Mit der Blockbildung und der Randomisierung haben wir zwei der **drei Prinzipien der Versuchsplanung** kennengelernt. Beide dienen dazu, alle anderen Einflussgrößen unter Kontrolle, also konstant zu halten. Das dritte Prinzip der Versuchsplanung ist die **Wiederholung**. Dieses haben wir immer wieder benutzt, ohne explizit darauf hinzuweisen. Durch die Wiederholung erhält man einen Schätzer für die Streuung der Beobachtungen.

16.1 Verbundene Stichproben

Im verbundenen Zweistichprobenproblem werden bei jeder Person bzw. jedem Objekt beide Verfahren betrachtet. Die Daten fallen also paarweise an. Wir bezeichnen den Wert, den wir bei der i-ten Person bzw. beim i-ten Objekt beim ersten Verfahren beobachten, mit x_i. Den Wert, den wir bei der i-ten Person bzw. beim i-ten Objekt beim zweiten Verfahren beobachten, bezeichnen wir mit y_i. Bei der i-ten Person bzw. beim i-ten Objekt beobachten wir also das Paar (x_i, y_i). Es soll überprüft werden, ob sich die beiden Verfahren unterscheiden. Wir gehen davon aus, dass wir die Realisationen (x_i, y_i) der bivariaten Zufallsvariablen (X_i, Y_i) beobachten, wobei wir davon ausgehen, dass $(X_1, Y_1), \ldots, (X_1, Y_n)$ unabhängig sind. Die Zufallsvariablen X_i und Y_i werden aber in der Regel abhängig sein. Besteht kein Unterschied zwischen den beiden Verfahren, so sollte gelten

$$E(X_i) = E(Y_i)$$

Mit $D_i = X_i - Y_i$ ist dies aber äquivalent zu $E(D_i) = 0$.

$$E(D_i) = E(X_i - Y_i) = E(X_i) - E(Y_i) = 0.$$

Die Verteilung der Differenzen D_i sollte das Zentrum 0 besitzen. Durch die Differenzenbildung machen wir aus zwei Stichproben eine Stichprobe.

Beispiel 144 (fortgesetzt von Seite 467)
Die Differenzen d_i sind:

```
60 32 -1 79 26 28 30 -7 61 35
```

□

Da wir aus zwei Stichproben eine gemacht haben, können wir also die Tests des Einstichprobenproblems verwenden. Die Analyse hängt nun von den Annahmen ab, die über die Differenzen D_i gemacht werden können.

16.1.1 Der t-Test

Können wir davon ausgehen, dass $D_1, \ldots, D_n$ unabhängige, mit den Parametern μ_D und σ_D^2 normalverteilte Zufallsvariablen sind, so wenden wir den t-Test an.
Das zweiseitige Testproblem lautet

$$H_0 : \mu_D = 0 \quad \text{gegen} \quad H_1 : \mu_D \neq 0$$

Die Teststatistik lautet

$$\boxed{t_D = \frac{\sqrt{n}\,\bar{d}}{s_D}}$$

mit

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \bar{x} - \bar{y}$$

und

$$s_D^2 = \frac{1}{n-1}\sum_{i=1}^{n} (d_i - \bar{d})^2 .$$

Die Entscheidungsregel lautet:
H_0 wird abgelehnt, wenn gilt $|t_D| > t_{n-1;1-\alpha/2}$.
Es können natürlich auch einseitige Tests durchgeführt werden.
Im Testproblem

$$H_0 : \mu_D = 0 \quad \text{gegen} \quad H_1 : \mu_D > 0$$

wird H_0 abgelehnt, wenn gilt $t_D > t_{n-1;1-\alpha}$.
Im Testproblem

$$H_0 : \mu_D = 0 \quad \text{gegen} \quad H_1 : \mu_D < 0$$

wird H_0 abgelehnt, wenn gilt $t_D < -t_{n-1;1-\alpha}$.
In allen drei Testproblemen ist t_p das p-Quantil der t-Verteilung mit $n-1$ Freiheitsgraden. Die Quantile sind in Tabelle 20.5 auf Seite 547 zu finden.

Beispiel 144 (fortgesetzt von Seite 470)
Wir testen zum Niveau 0.05:

$$H_0 : \mu_D = 0 \quad \text{gegen} \quad H_1 : \mu_D > 0$$

Es gilt $\bar{d} = 34.3$ und $s_D = 26.78$. Also gilt

$$t_D = \frac{\sqrt{10} \cdot 34.3}{26.78} = 4.05$$

Tabelle 20.5 auf Seite 547 entnehmen wir $t_{9;0.95} = 1.833$. Wegen $t_D = 4.05$ wird H_0 zum Niveau $\alpha = 0.05$ abgelehnt. □

16.1.2 Der Vorzeichentest

Können wir keine Normalverteilung und auch keine andere symmetrische Verteilung unterstellen, so wenden wir den Vorzeichentest an.
Die Hypothesen lauten:

$$H_0 : M_D = 0 \quad \text{gegen} \quad H_1 : M_D \neq 0$$

wobei M_D der Median der Differenzen ist.
Die Teststatistik S ist die Anzahl der positiven Differenzen.
Wir lehnen H_0 zum Signifikanzniveau α ab, wenn gilt $S \leq s_{\alpha/2}$ oder $S \geq n - s_{\alpha/2}$.
Gilt $n > 20$ bilden wir folgende standardnormalverteilte Teststatistik:

$$\boxed{Z = \frac{S - 0.5\,n}{0.5\,\sqrt{n}}}. \tag{16.1}$$

Wir lehnen H_0 ab, wenn gilt $|Z| \geq z_{1-\alpha/2}$.
Es können auch einseitige Tests durchgeführt werden.
Im Testproblem

$$H_0 : M_D = 0 \quad \text{gegen} \quad H_1 : M_D > 0$$

wird H_0 abgelehnt, wenn gilt $S \geq n - s_\alpha$. Für $n > 20$ lehnen wir H_0 ab, wenn für die Teststatistik Z in Gleichung (16.1) $Z \geq z_{1-\alpha}$ gilt.
Im Testproblem

$$H_0 : M_D = 0 \quad \text{gegen} \quad H_1 : M_D < 0$$

wird H_0 abgelehnt, wenn gilt $S \leq s_\alpha$. Für $n > 20$ lehnen wir H_0 ab, wenn für die Teststatistik Z in Gleichung (16.1) $Z \leq -z_{1-\alpha}$ gilt.
In allen drei Testproblemen sind die Werte von s_p in Tabelle 20.6 auf Seite 551 und die Werte von z_p in Tabelle 20.3 auf Seite 549 zu finden.

Beispiel 144 (fortgesetzt von Seite 471)
Wir testen zum Niveau $\alpha = 0.05$

$$H_0 : M_D = 0 \quad \text{gegen} \quad H_1 : M_D > 0$$

Es gilt $S = 8$. Tabelle 20.6 auf Seite 551 entnehmen wir $s_{0.05} = 1$. Also gilt $n - s_{0.05} = 9$ Wir lehnen H_0 also nicht ab. □

16.1.3 Der Wilcoxon-Vorzeichen-Rangtest

Können wir keine Normalverteilung, aber eine andere symmetrische Verteilung unterstellen, so wenden wir den Wilcoxon-Vorzeichen-Rangtest an.
Die Hypothesen lauten im zweiseitigen Testproblem

$$H_0 : M_D = 0 \quad \text{gegen} \quad H_1 : M_D \neq 0$$

wobei M_D der Median der Differenzen ist.
Die Entscheidungsregel ist:
Wir lehnen H_0 zum Signifikanzniveau α ab, wenn gilt $W^+ \leq w_{\alpha/2}$ oder $W^+ \geq n(n+1)/2 - w_{\alpha/2}$.
Für große Werte von n bilden wir folgende standardnormalverteilte Teststatistik:

$$\boxed{Z = \frac{W^+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}}. \tag{16.2}$$

Wir lehnen H_0 ab, wenn gilt $|Z| \geq z_{1-\alpha/2}$.
Im Testproblem

$$H_0 : M = 0 \quad \text{gegen} \quad H_1 : M > 0$$

wird H_0 zum Signifikanzniveau α abgelehnt, wenn gilt $W^+ \geq n(n+1)/2 - w_\alpha$.
Für $n > 20$ lehnen wir H_0 ab, wenn für die Teststatistik in Gleichung (16.2) $Z \geq z_{1-\alpha}$ gilt.
Im Testproblem

$$H_0 : M = 0 \quad \text{gegen} \quad H_1 : M < 0$$

wird H_0 zum Signifikanzniveau α abgelehnt, wenn gilt $W^+ \leq w_\alpha$. Für $n > 20$ lehnen wir H_0 ab, wenn für die Teststatistik in Gleichung (16.2) $Z \leq -z_{1-\alpha}$ gilt.
In allen drei Testproblemen sind die Werte von w_p in Tabelle 20.7 auf Seite 552 und die Werte von z_p in Tabelle 20.3 auf Seite 549 zu finden.

Beispiel 144 (fortgesetzt von Seite 472)
Wir testen

$$H_0 : M_D = 0 \quad \text{gegen} \quad H_1 : M_D > 0$$

Tabelle 16.2 illustriert die Berechnung.

Tabelle 16.2: Berechnung der Teststatistik des Wilcoxon-Vorzeichen-Rangtests

i	1	2	3	4	5	6	7	8	9	10
D_i	60	32	-1	79	26	28	30	-7	61	35
$\lvert D_i\rvert$	60	32	1	79	26	28	30	7	61	35
$R(\lvert D_i\rvert)$	8	6	1	10	3	4	5	2	9	7
$s(D_i)$	1	1	0	1	1	1	1	0	1	1

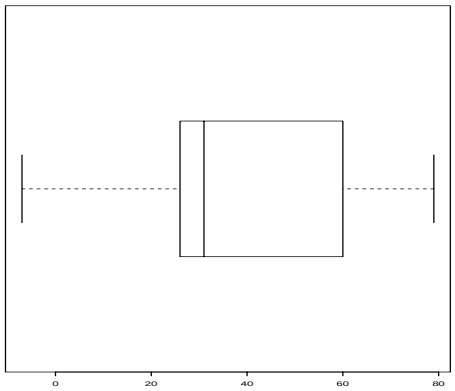
Es gilt $W^+ = 52$. Tabelle 20.7 auf Seite 552 entnehmen wir $w_{0.05} = 10$. Also gilt $n(n+1)/2 - w_{0.05} = 10 \cdot 11/2 - 10 = 55 - 10 = 45$. Also lehnen wir H_0 zum Signifikanzniveau $\alpha = 0.05$ ab. □

16.1.4 Praktische Aspekte

Bei der Auswahl des geeigneten Tests sollte man sich wie auch im Einstichprobenproblem vom Boxplot leiten lassen.

Beispiel 144 (fortgesetzt von Seite 473)
Abbildung 16.1 zeigt den Boxplot der Differenz der Herzfrequenz.

Abbildung 16.1: Boxplot der Differenz



Der Boxplot deutet auf eine schiefe Verteilung hin. Man sollte also den Vorzeichentest anwenden. □

16.2 Unverbundene Stichproben

Im unverbunden Zweistichprobenproblem gehen wir von folgender Situation aus:
Es werden $N = m + n$ Personen bzw. Objekte zufällig ausgewählt und dann zufällig auf eine Gruppe mit m Personen bzw. Objekten und eine Gruppe mit n Personen bzw. Objekten aufgeteilt. Auf die Personen bzw. Objekte der ersten Gruppe wird dann das erste Verfahren und auf die Personen bzw. Objekte der zweiten Gruppe das zweite Verfahren angewendet. Die den Beobachtungen $x_1, \ldots, x_m$ der ersten Stichprobe zugrundeliegenden Zufallsvariablen sind $X_1, \ldots, X_m$ und die den Beobachtungen $y_1, \ldots, y_n$ der zweiten Stichprobe zugrundeliegenden Zufallsvariablen sind $Y_1, \ldots, Y_n$. Wir gehen davon aus, dass die Zufallsvariablen $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ unabhängig sind.

16.2.1 Der t-Test

Die klassische Annahme ist, dass die Zufallsvariablen $X_1, \ldots, X_m$ normalverteilt sind mit den Parametern μ_X und σ_X^2 und die Zufallsvariablen $Y_1, \ldots, Y_n$ normalverteilt mit den Parametern μ_Y und σ_Y^2.
Das Testproblem lautet

$$H_0 : \mu_X = \mu_Y \quad \text{gegen} \quad \mu_X \neq \mu_Y$$

Wir sind nur an einem Lageunterschied interessiert und unterstellen deshalb, dass $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ gilt.
Die Teststatistik ist:

$$\boxed{t = \frac{\bar{x} - \bar{y}}{\hat{\sigma}\sqrt{\frac{1}{m} + \frac{1}{n}}}} \tag{16.3}$$

mit

$$\boxed{\hat{\sigma}^2 = \frac{1}{m+n-2}\left(\sum_{i=1}^{m}(x_i - \bar{x})^2 + \sum_{j=1}^{n}(y_j - \bar{y})^2\right).}$$

Die Entscheidungsregel lautet:
H_0 ablehnen, wenn gilt $|t| > t_{m+n-2;1-\alpha/2}$.
Dabei ist $t_{m+n-2;1-\alpha/2}$ das $1 - \alpha/2$-Quantil der t-Verteilung mit $m + n - 2$ Freiheitsgraden.

Beispiel 146 (fortgesetzt von Seite 469)
Wir wollen überprüfen, ob der Aufwand für Hausaufgaben keinen Einfluss auf die erwartete Punktezahl hat. Wir testen also

$$H_0 : \mu_X = \mu_Y \quad \text{gegen} \quad H_1 : \mu_X \neq \mu_Y$$

Es gilt $\bar{x} = 500.125$ und $\bar{y} = 480.5$. Weiterhin gilt $\sum_{i=1}^{m} (x_i - \bar{x})^2 = 6774.875$ und $\sum_{j=1}^{n} (y_j - \bar{y})^2 = 8482.5$. Also gilt $\hat{\sigma}^2 = 953.586$ und $\hat{\sigma} = 30.88$.
Für die Teststatistik gilt

$$t = \frac{\bar{x} - \bar{y}}{\hat{\sigma}\sqrt{\frac{1}{m} + \frac{1}{n}}} = \frac{500.125 - 480.5}{30.88\sqrt{\frac{1}{8} + \frac{1}{10}}} = 1.34$$

Tabelle 20.5 auf Seite 547 entnehmen wir $t_{16;0.975} = 2.12$. Also lehnen wir H_0 zum Signifikanzniveau 0.05 nicht ab. □

Es können natürlich auch einseitige Tests durchgeführt werden.
Im Testproblem

$$H_0 : \mu_X = \mu_Y \quad \text{gegen} \quad H_1 : \mu_X < \mu_Y$$

wird H_0 abgelehnt, wenn gilt $t < -t_{m+n-2;1-\alpha}$.

Beispiel 145 (fortgesetzt von Seite 468)
Wir wollen überprüfen, ob die Kleinkinder, bei denen die Reflexe eingeübt wurden, im Mittel schneller laufen lernen als die Kleinkinder, bei denen die Reflexe nicht eingeübt wurden. Wir testen also

$$H_0 : \mu_X = \mu_Y \quad \text{gegen} \quad H_1 : \mu_X < \mu_Y$$

Es gilt $\bar{x} = 10.125$ und $\bar{y} = 11.70833$. Außerdem gilt $\sum_{i=1}^{m} (x_i - \bar{x})^2 = 10.46875$ und $\sum_{j=1}^{n} (y_i - \bar{y})^2 = 11.55208$. Also gilt $\hat{\sigma}^2 = 2.2$ und $\hat{\sigma} = 1.48$.
Für die Teststatistik gilt

$$t = \frac{\bar{x} - \bar{y}}{\hat{\sigma}\sqrt{\frac{1}{m} + \frac{1}{n}}} = \frac{10.125 - 11.70833}{1.48\sqrt{\frac{1}{6} + \frac{1}{6}}} = -1.85$$

Tabelle 20.5 auf Seite 547 entnehmen wir $t_{10;0.95} = 1.812$. Wegen $-1.85 < -1.812$ lehnen wir H_0 zum Signifikanzniveau 0.05 ab. □

Im Testproblem

$$H_0 : \mu_X = \mu_Y \quad \text{gegen} \quad H_1 : \mu_X > \mu_Y$$

wird H_0 abgelehnt, wenn gilt $t > t_{m+n-2;1-\alpha}$.
In allen Testproblemen ist $t_{m+n-2;p}$ das p-Quantil der t-Verteilung mit $m + n - 2$ Freiheitsgraden.

16.2.2 Der Wilcoxon Rangsummentest

Der t-Test beruht auf der Annahme der Normalverteilung. Ist diese nicht gerechtfertigt, sollte man den Wilcoxon-Rangsummentest durchführen.

Dieser beruht auf folgenden Annahmen:

Die Zufallsvariablen $X_1, \ldots, X_m$ seien unabhängig und identisch mit stetiger Verteilungsfunktion $F_X(x)$ und die Zufallsvariablen $Y_1, \ldots, Y_n$ seien unabhängig und identisch mit stetiger Verteilungsfunktion $F_Y(y)$ verteilt. Es wird unterstellt, dass sich die Verteilungen nur hinsichtlich der Lage unterscheiden können.

Das zweiseitige Testproblem lautet

$$H_0 : M_X = M_Y \quad \text{gegen} \quad H_1 : M_X \neq M_Y \, .$$

Dabei ist M_X der Median der ersten und M_Y der Median der zweiten Grundgesamtheit.

Wir gehen zunächst davon aus, dass alle Beobachtungen unterschiedlich sind, also keine Bindungen vorliegen. Wenn H_0 zutrifft, kommen alle Beobachtungen aus einer Grundgesamtheit. Dies sollte sich dadurch zeigen, dass die Beobachtungen der beiden Stichproben gut gemischt sind. Es sollten also nicht alle Beobachtungen der einen Stichprobe an dem einen Ende der gemeinsamen geordneten Stichprobe liegen.

Schauen wir uns dazu den Fall m=n=3 an. Die Konfiguration $xyyxxy$ deutet darauf hin, dass die Beobachtungen aus einer Grundgesamtheit kommen. Die Konfiguration $xxxyyy$ und die Konfiguration $yyyxxx$ deuten darauf hin, dass sich die Grundgesamtheiten hinsichtlich der Lage unterscheiden.

Wie können wir diese Muster mit Hilfe einer geeigneten Teststatistik erkennen? Der Wilcoxon Rangsummentest benutzt die Ränge R_i der x_i in der gemeinsamen Stichprobe $x_1, \ldots, x_m, y_1, \ldots, y_n$. Der Rang R_i von x_i gibt an, wieviele von **allen** Beobachtungen kleiner oder gleich x_i sind.

Beispiel 146 (fortgesetzt von Seite 476)
Die gemeinsame Stichprobe ist

```
546 522 483 441 507 516 494 492
474 523 527 487 458 422 479 462 493 480
```

wobei die ersten 8 Beobachtungen aus der ersten Gruppe stammen. Der Rang R_1 von x_1 ist 18, da 546 die größte Beobachtung ist. Für die Ränge der anderen 7 Beobachtungen der ersten Gruppe gilt

$$R_2 = 15 \quad R_3 = 8 \quad R_4 = 2 \quad R_5 = 13 \quad R_6 = 14 \quad R_7 = 12 \quad R_8 = 10$$

□

Wie können wir die Ränge benutzen, um einen Lageunterschied aufzudecken? Für $xyyxxy$ sind die Ränge der x_i gleich $1, 4, 5$, für $xxxyyy$ sind die Ränge der x_i gleich $1, 2, 3$ und für $yyyxxx$ sind die Ränge der x_i gleich $4, 5, 6$. Bildet man nun die Summe der Ränge der x_i, so ist diese im ersten Fall gleich 10, im zweiten Fall gleich 6 und im dritten Fall gleich 15.
Sehr kleine oder sehr große Werte der Summe der Ränge deuten also darauf hin, dass die Beobachtungen aus unterschiedlichen Verteilungen kommen. Auf dieser Idee basiert der Wilcoxon Rangsummentest. Seine Teststatistik lautet:

$$\boxed{W = \sum_{i=1}^{m} R_i} \tag{16.4}$$

Beispiel 146 (fortgesetzt Seite 477)
Es gilt $W = 92$. □

Unter H_0 kann die exakte Verteilung von W für kleine Stichprobenumfänge durch Auszählen einfach hergeleitet werden. Da keine Bindungen vorliegen, werden als Ränge die natürlichen Zahlen $1, 2, \ldots, m + n$ vergeben. Wenn H_0 zutrifft, stammen alle Beobachtungen aus der gleichen Grundgesamtheit, und jede Aufteilung der Ränge auf die beiden Stichproben ist gleichwahrscheinlich. Für jede dieser Rangaufteilungen bestimmen wir den Wert von W.
Wir wollen dies für den Fall $m = n = 3$ durchführen. Es gibt also insgesamt $\binom{6}{3} = 20$ Möglichkeiten, aus der Menge der Ränge $\{1, 2, 3, 4, 5, 6\}$ drei Ränge für die erste Stichprobe auszuwählen. Alle diese Fälle und der zugehörige Wert von W sind in Tabelle 16.3 angegeben.

Tabelle 16.3: Rangkonfigurationen und Wert von W für $m = n = 3$

Ränge	W	Ränge	W	Ränge	W	Ränge	W
1,2,3	6	1,3,5	9	2,3,4	9	2,5,6	13
1,2,4	7	1,3,6	10	2,3,5	10	3,4,5	12
1,2,5	8	1,4,5	10	2,3,6	11	3,4,6	13
1,2,6	9	1,4,6	11	2,4,5	11	3,5,6	14
1,3,4	8	1,5,6	12	2,4,6	12	4,5,6	15

Durch einfaches Auszählen erhalten wir die Verteilung von W für $m = n = 3$,

die in Tabelle 16.4 zu finden ist.

Tabelle 16.4: Verteilung von W für $m = n = 3$

w	6	7	8	9	10	11	12	13	14	15
$P(W = w)$	0.05	0.05	0.10	0.15	0.15	0.15	0.15	0.10	0.05	0.05

Für $m = n = 3$ gilt also $w_{0.05} = 6$ und $w_{0.10} = 7$.
Die Entscheidungsregel beim zweiseitigen Test lautet:
Entscheidung für H_1, wenn gilt $W \leq w_{\alpha/2}$ oder $W \geq m\,(N+1) - w_{\alpha/2}$. Auf Seite 552 ist eine Tabelle mit Werten von w_p für $m = n$ zu finden.
Sind die Werte von m und n groß, so ist W approximativ normalverteilt mit $E(W) = m\,(N+1)/2$ und $Var(W) = m\,n\,(N+1)/12$. Wir bilden also folgende standardnormalverteilte Teststatistik:

$$\boxed{Z = \frac{W - m(N+1)/2}{\sqrt{m\,n\,(N+1)/12}}}. \tag{16.5}$$

Wir lehnen H_0 ab, wenn gilt $|Z| \geq z_{1-\alpha/2}$. Dabei ist $z_{1-\alpha/2}$ das $1-\alpha/2$-Quantil der Standardnormalverteilung.

Beispiel 146 (fortgesetzt von Seite 478)
Wir wollen überprüfen, ob der Aufwand für Hausaufgaben keinen Einfluss auf den Median der Punktezahl hat. Wir testen also

$$H_0 : M_X = M_Y \quad \text{gegen} \quad H_1 : M_X \neq M_Y$$

Es gilt $W = 92$. Tabelle 20.8 auf Seite 552 enthält Werte von w_p nur für identische Stichprobenumfänge. Eine größere Tabelle ist in Schlittgens Einführung in die Statistik zu finden. Dieser entnehmen wir für $m = 8$ und $n = 10$ den Wert $w_{0.025} = 53$. Somit gilt $m\,(N+1) - w_{0.025} = 152 - 53 = 99$. Also lehnen wir H_0 zum Signifikanzniveau 0.05 nicht ab.
Wir können aber auch die approximativ normalverteilte Teststatistik Z aus Gleichung (16.5) bestimmen. Es gilt

$$Z = \frac{W - m(N+1)/2}{\sqrt{m\,n\,(N+1)/12}} = \frac{92 - 8 \cdot (18+1)/2}{\sqrt{8 \cdot 10 \cdot (18+1)/12}} = 1.42$$

Tabelle 20.3 auf Seite 549 entnehmen wir $z_{0.975} = 1.96$. Also lehnen wir H_0 zum Signifikanzniveau 0.05 nicht ab. □

Es können natürlich auch einseitige Tests durchgeführt werden.
Im Testproblem

$$H_0 : M_X = M_Y \quad \text{gegen} \quad H_1 : M_X > M_Y$$

wird H_0 abgelehnt, wenn gilt $W \geq m\,(N+1) - w_\alpha$. Für große Werte von m und n lehnen wir H_0 ab, wenn für die Teststatistik Z in Gleichung (16.5) $|Z| \geq z_{1-\alpha}$ gilt.
Im Testproblem

$$H_0 : M_X = M_Y \quad \text{gegen} \quad H_1 : M_X < M_Y$$

wird H_0 abgelehnt, wenn gilt $W \leq w_\alpha$. Für große Werte von m und n lehnen wir H_0 ab, wenn für die Teststatistik Z in Gleichung (16.5) $Z \leq -z_{1-\alpha}$ gilt. Werte von w_p sind in Tabelle 20.8 auf Seite 552 und Werte von z_p in Tabelle 20.3 auf Seite 549 zu finden.
In vielen praktischen Anwendungen kommen Bindungen vor.

Beispiel 145 (fortgesetzt von Seite 476)
Die gemeinsame Stichprobe ist:

```
9.0 9.5 9.75 10.0 13.0 9.50 11.5 12.0 9.0 11.5 13.25 13.0
```

Die Werte 9, 9.5, 11.5 und 13 kommen zweimal vor, die restlichen Werte einmal. □

Bei Bindungen werden Durchschnittsränge bestimmt. Die Teststatistik des Wilcoxon-Rangsummentests ist

$$W = \sum_{i=1}^{m} R_i \tag{16.6}$$

Beispiel 145 (fortgesetzt von Seite 480)
Wir bilden Durchschnittsränge. Die Ränge der ersten Stichprobe sind.

```
1.5  3.5 5 6 10.5 3.5
```

Also gilt $W = 30$. □

Wir müssen wie schon im Einstichprobenproblem beim Wilcoxon-Vorzeichen-Rangtest die Varianz der Teststatistik modifizieren. Es gilt

$$Var(W) = \frac{m\,n}{12}\left[N + 1 - \frac{1}{N\,(N-1)}\sum_{j=1}^{r}(b_j^3 - b_j)\right]$$

Dabei ist r die Anzahl der Gruppen mit Bindungen und b_j die Anzahl der Beobachtungen in der j-ten Bindungsgruppe.
Folgende Teststatistik ist approximativ standardnormalverteilt:

$$Z = \frac{W - \frac{m\,(N+1)}{2}}{\sqrt{\frac{m\,n}{12}\left[N+1-\frac{1}{N\,(N-1)}\sum_{j=1}^{r}(b_j^3-b_j)\right]}}$$

Im Testproblem

$$H_0 : M_X = M_Y \quad \text{gegen} \quad H_1 : M_X \neq M_Y$$

wird H_0 abgelehnt, wenn gilt $Z \geq z_{1-\alpha/2}$ gilt.
Im Testproblem

$$H_0 : M_X = M_Y \quad \text{gegen} \quad H_1 : M_X > M_Y$$

wird H_0 abgelehnt, wenn gilt $Z \geq z_{1-\alpha}$ gilt.
Im Testproblem

$$H_0 : M_X = M_Y \quad \text{gegen} \quad H_1 : M_X < M_Y$$

wird H_0 abgelehnt, wenn gilt $Z \leq -z_{1-\alpha}$ gilt.
In allen drei Testproblemen ist z_p das p-Quantil der Standardnormalverteilung.

Beispiel 145 (fortgesetzt von Seite 480)
In Tabelle 16.5 sind die Bindungsgruppen mit ihren Häufigkeiten zu finden.

Tabelle 16.5: Bindungsgruppen

r	1	2	3	4	5	6	7	8
x	9	9.5	9.75	10	11.5	12	13	13.25
b_r	2	2	1	1	2	1	2	1

Es gilt

$$\sum_{j=1}^{r}(b_j^3 - b_j) = 24$$

Also gilt $Z = -1.451338$. Wegen $z_{0.05} = -1.645$ lehnen wir H_0 zum Signifikanzniveau 0.05 nicht ab. □

16.3 Das Zweistichprobenproblem in R

Beginnen wir mit dem verbundenen Zweistichprobenproblem. Wir wollen die Daten aus Beispiel 144 auf Seite 467 in R analysieren. Die Werte der Herzfrequenz der Ratten, die nicht allein sind, weisen wir der Variablen `hf.nicht.allein` zu:

```
> hf.nicht.allein<-c(523,494,461,535,476,454,448,408,470,437)
```

Die Werte der Herzfrequenz der Ratten, die allein sind, weisen wir der Variablen `hf.allein` zu:

```
> hf.allein<-c(463,462,462,456,450,426,418,415,409,402)
```

Nun bilden wir die Differenz `d` aus `hf.nicht.allein` und `hf.allein`:

```
> d<-hf.nicht.allein-hf.allein
> d
 [1] 60 32 -1 79 26 28 30 -7 61 35
```

Auf diese Differenzen können wir nun die Lagetests im Einstichprobenproblem anwenden. Beginnen wir mit dem t-Test.

```
> t.test(d,alternative="greater")

        One Sample t-test

data:  d
t = 4.0498, df = 9, p-value = 0.001443
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
18.77423      Inf
sample estimates:
mean of x
     34.3
```

Die Überschreitungswahrscheinlichkeit beträgt 0.001443. Also lehnen wir H_0 zum Signifikanzniveau 0.05 ab.
Schauen wir uns nun den Vorzeichentest an.

```
> S<-sum(d>0)
> n<-length(d)
> binom.test(S,n,alternative="greater")
```

```
        Exact binomial test

data:  S and n
number of successes = 8, number of trials = 10,
p-value = 0.05469
alternative hypothesis: true probability of
success is greater than 0.5
95 percent confidence interval:
 0.4930987 1.0000000
sample estimates:
probability of success
                  0.8
```

Die Überschreitungswahrscheinlichkeit beträgt 0.05469. Also lehnen wir H_0 zum Signifikanzniveau 0.05 nicht ab.
Und zuletzt führen wir den Wilcoxon-Vorzeichen-Rangtest durch.

```
> wilcox.test(d,alternative="greater")

        Wilcoxon signed rank test

data:  d
V = 52, p-value = 0.004883
alternative hypothesis: true mu is greater than 0
```

Die Überschreitungswahrscheinlichkeit beträgt 0.004883. Also lehnen wir H_0 zum Signifikanzniveau 0.05 nicht ab.
Schauen wir uns nun das unverbundene Zweistichprobenproblem an. Wir wollen die Daten aus Beispiel 146 auf Seite 469 in R analysieren. Die Punkte der Länder mit geringem Aufwand für Hausaufgaben weisen wir der Variablen `ha.wenig` zu:

```
> ha.wenig<-c(546,522,483,441,507,516,494,492)
```

Die Punkte der Länder mit hohem Aufwand für Hausaufgaben weisen wir der Variablen `ha.viel` zu:

```
> ha.viel<-c(474,523,527,487,458,422,479,462,493,480)
```

Für den t-Test können wir die Funktion `t.test` verwenden. Wir rufen sie mit den beiden Variablen auf. Wir müssen außerdem das Argument `var.equal` auf den Wert `TRUE` setzen.

```
> t.test(ha.wenig,ha.viel,var.equal=TRUE)

        Two Sample t-test

data:  ha.wenig and ha.viel
t = 1.3398, df = 16, p-value = 0.1990
alternative hypothesis:
true difference in means is not equal to 0
95 percent confidence interval:
 -11.42686  50.67686
sample estimates:
mean of x mean of y
  500.125   480.500
```

Der Wert der Teststatistik ist 1.3398. Die Überschreitungswahrscheinlichkeit beträgt 0.199. Also lehnen wir H_0 zum Signifikanzniveau 0.05 nicht ab.
Wir wollen die Daten aus Beispiel 145 auf Seite 468 in `R` analysieren. Das Alter der Kleinkinder, bei denen der Reflex eingeübt wurde, weisen wir der Variablen `ueben.ja` zu:

```
> ueben.ja<-c(9,9.5,9.75,10,13,9.5)
```

Das Alter der Kleinkinder, bei denen der Reflex nicht eingeübt wurde, weisen wir der Variablen `ueben.nein` zu:

```
> ueben.nein<-c(11.5,12,9,11.5,13.25,13)
```

Wir rufen den t-Test auf:

```
> t.test(ueben.ja,ueben.nein,alternative="less",var.equal=T)

        Two Sample t-test

data:  ueben.ja and ueben.nein
t = -1.8481, df = 10, p-value = 0.04717
alternative hypothesis:
true difference in means is less than 0
95 percent confidence interval:
        -Inf -0.03049964
sample estimates:
mean of x mean of y
10.12500  11.70833
```

Die Überschreitungswahrscheinlichkeit beträgt 0.04717. Also lehnen wir H_0 zum Signifikanzniveau 0.05 ab.
Für den Wilcoxon-Rangsummentest können wir die Funktion `wilcox.test` verwenden. Wir rufen sie mit den beiden Datensätzen auf. Beginnen wir mit den Daten der PISA-Studie.

```
> wilcox.test(ha.wenig,ha.viel)

        Wilcoxon rank sum test

 data:  ha.wenig and ha.viel
 W = 56, p-value = 0.1728
 alternative hypothesis: true mu is not equal to 0
```

Wir hatten im Beispiel 146 auf Seite 478 einen Wert von 92 für W bestimmt. R gibt den Wert 56 an. Die Diskrepanz zwischen den Werten liegt daran, dass in R die Teststistik

$$W - \frac{m(m+1)}{2}$$

bestimmt wird. Im Beispiel ist $m = 8$ und somit $m(m+1)/2 = 36$. Die Wahl von R führt dazu, dass der kleinste Wert der Teststatistik gleich 0 ist.
Die Überschreitungswahrscheinlichkeit beträgt 0.1728. Also lehnen wir H_0 zum Signifikanzniveau 0.05 nicht ab.
Und nun zu den Kleinkindern.

```
> wilcox.test(ueben.ja,ueben.nein,alternative="less",correct=F)

        Wilcoxon rank sum test

 data:  ueben.ja and ueben.nein
 W = 9, p-value = 0.07334
 alternative hypothesis: true mu is less than 0

 Warning message: Cannot compute exact p-value with ties in:
 wilcox.test.default(ueben.ja,ueben.nein,alternative="less",
 correct=F)
```

Die Überschreitungswahrscheinlichkeit beträgt 0.07334. Also lehnen wir H_0 zum Signifikanzniveau 0.05 nicht ab.
Wie im Einstichprobenproblem ist es auch im Zweistichprobenproblem möglich, mit dem von Streitberg und Röhmel entwickelten Shift-Algorithmus

den exakten Wilcoxon Rangsummentest auch bei Bindungen durchzuführen. Dieser ist in dem von Torsten Hothorn und Kurt Hornik erstellten Paket `exactRankTests` in der Funktion `wilcox.exact` implementiert. Dieses Paket muss man zunächst installieren und laden. Wie man dabei vorzugehen hat, wird auf Seite 52 beschrieben.
Die Funktion `wilcox.exact` wird genauso wie die Funktion `wilcox.test` aufgerufen. Wir geben also ein

```
> wilcox.exact(ueben.ja,ueben.nein,alternative="less")

        Exact Wilcoxon rank sum test

data:  ueben.ja and ueben.nein
W = 9, p-value = 0.08117
alternative hypothesis: true mu is less than 0
```

Die exakte Überschreitungswahrscheinlichkeit beträgt 0.08117 und approximative 0.07334. Die Approximation ist also sehr gut.

Kapitel 17

Einfaktorielle Varianzanalyse

Im Zweistichprobenproblem vergleichen wir zwei Verfahren miteinander. Nun wollen wir mehr als zwei Verfahren betrachten, wobei wir unverbunden vorgehen.

Beispiel 147
Im Rahmen der PISA-Studie wurde auch der Zeitaufwand der Schüler für Hausaufgaben erhoben (Deutsches PISA-Konsortium (Hrsg.) (2001), S.417). Dort wird unterschieden zwischen sehr geringem, geringem, mittlerem, großem und sehr großem Aufwand. Wir fassen die Länder mit sehr geringem und geringem Aufwand und die Länder mit großem und sehr großem Aufwand zusammen. Somit liegen drei Gruppen vor. Die Gruppe der Länder mit wenig Zeitaufwand nennen wir im Folgenden Gruppe 1, die Gruppe der Länder mit mittlerem Zeitaufwand Gruppe 2 und die Gruppe der Länder mit großem Zeitaufwand Gruppe 3. Wir wollen vergleichen, ob sich die Verteilung des Merkmals `Mathematische Grundbildung` in den drei Gruppen unterscheidet. □

Wird untersucht, ob sich die Verteilung eines Merkmals in mehreren Gruppen unterscheidet, so spricht man von **univariater Varianzanalyse**.

17.1 Varianzanalyse bei Normalverteilung

Ausgangspunkt sind die Realisationen y_{ij} der unabhängigen Zufallsvariablen Y_{ij}, $i = 1, \ldots, I$, $j = 1, \ldots, n_i$, die mit Erwartungswert μ_i, $i = 1, \ldots, I$ und Varianz σ^2 normalverteilt sind. Die Erwartungswerte der Gruppen können sich also unterscheiden, während die Varianz identisch sein muss. Dabei bezieht sich der Index i auf die i-te Gruppe, während der Index j sich auf die j-te Beobachtung bezieht. In der i-ten Gruppe liegen also n_i Beobachtungen

vor. Die einzelnen Gruppen können unterschiedlich groß sein. Die Gesamtzahl aller Beobachtungen bezeichnen wir mit N.

Beispiel 147 (fortgesetzt von Seite 487)
Die Beobachtungen in den einzelnen Gruppen sind:

Gruppe 1:	536 557 514 446 515 510 529 498
Gruppe 2:	533 520 334 514 490 517 514 533 547 537 499 454 493
Gruppe 3:	447 529 503 457 463 387 470 478 476 488

□

Es ist zu testen:

$$H_0: \quad \mu_1 = \ldots = \mu_I \tag{17.1}$$

gegen

$$H_1: \quad \mu_i \neq \mu_j \quad \text{für mind. ein Paar } (i,j) \text{ mit } i \neq j\,.$$

Es liegt nahe zur überprüfung von (17.1) die Mittelwerte

$$\boxed{\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}} \tag{17.2}$$

der einzelnen Gruppen zu bestimmen und zu vergleichen.

Beispiel 147 (fortgesetzt von Seite 488)
Es gilt $\bar{y}_1 = 513.125$, $\bar{y}_2 = 498.8462$ und $\bar{y}_3 = 469.8$. Die Mittelwerte unterscheiden sich. □

Der Vergleich von zwei Mittelwerten $\bar{y}_1$ und $\bar{y}_2$ ist einfach. Wir bilden die Differenz $\bar{y}_1 - \bar{y}_2$ der beiden Mittelwerte. Bei mehr als zwei Gruppen können wir alle Paare von Gruppen betrachten und $\bar{y}_i$ mit $\bar{y}_j$ für $i < j$ vergleichen. Hierdurch erhalten wir aber kein globales Maß für den Vergleich aller Gruppen. Um dieses zu erhalten, fassen wir die Mittelwerte $\bar{y}_i$, $i = 1, \ldots, I$ als eine Stichprobe auf und bestimmen, wie stark sie um den Mittelwert

$$\boxed{\bar{y} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} y_{ij}} \tag{17.3}$$

aller Beobachtungen streuen.

Beispiel 147 (fortgesetzt von Seite 488)
Es gilt $\bar{y} = 493.1613$. □

Es liegt nahe, die Streuung der Mittelwerte $\bar{y}_i$ um das Gesamtmittel $\bar{y}$ folgendermaßen zu bestimmen:

$$\sum_{i=1}^{I} (\bar{y}_i - \bar{y})^2.$$

Hierbei wird aber nicht berücksichtigt, dass die Gruppen unterschiedlich groß sein können. Eine große Gruppe sollte ein stärkeres Gewicht erhalten als eine kleine Gruppe. Wir bilden also

$$\boxed{SS_B = \sum_{i=1}^{I} n_i (\bar{y}_i - \bar{y})^2.} \tag{17.4}$$

Man bezeichnet SS_B als **Streuung zwischen den Gruppen**.

Beispiel 147 (fortgesetzt von Seite 489)
Es gilt

$$\begin{aligned} SS_B &= 8(513.125 - 493.1613)^2 + 13(498.8462 - 493.1613)^2 \\ &+ 10(469.8 - 493.1613)^2 = 9066.03. \end{aligned}$$

□

Wie das folgende Beispiel zeigt, ist die Größe SS_B allein aber keine geeignete Teststatistik zur überprüfung der Hypothese (17.1).

Beispiel 148
1. Situation:

Die Werte eines Merkmals in drei Gruppen sind:

Gruppe 1:	47 53 49 50 46
Gruppe 2:	55 54 58 61 52
Gruppe 3:	53 50 51 52 49

Es gilt

$$\bar{y}_1 = 49, \quad \bar{y}_2 = 56, \quad \bar{y}_3 = 51, \quad \bar{y} = 52.$$

2. Situation:

Die Werte eines Merkmals in drei Gruppen sind:

Gruppe 1:	50 42 53 45 55
Gruppe 2:	48 57 65 59 51
Gruppe 3:	57 59 48 46 45

Auch hier gilt

$$\bar{y}_1 = 49, \quad \bar{y}_2 = 56, \quad \bar{y}_3 = 51, \quad \bar{y} = 52.$$

Also ist der Wert von SS_B in beiden Konstellationen identisch. Wie die Abbildungen 17.1 und 17.2 zeigen, unterscheiden sich die beiden Konstellationen aber beträchtlich. Die Boxplots in Abbildung 17.1 verdeutlichen, dass die Streuung innerhalb der Gruppen klein ist, während in Abbildung 17.2 die Streuung innerhalb der Gruppen groß ist. Abbildung 17.1 spricht für einen Lageunterschied zwischen den Gruppen, während die unterschiedlichen Mittelwerte in 17.2 eher durch die hohen Streuungen erklärt werden können.

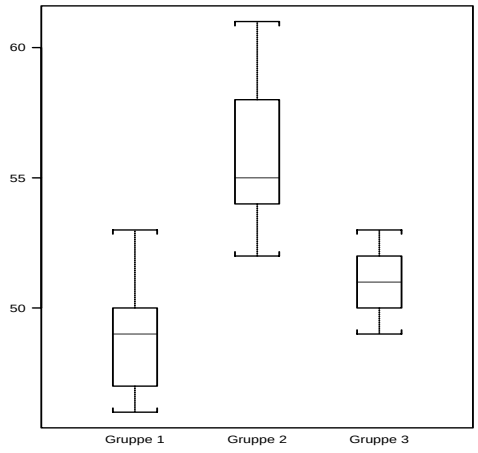


Abbildung 17.1: Boxplot von drei Gruppen mit kleiner Streuung innerhalb der Gruppen

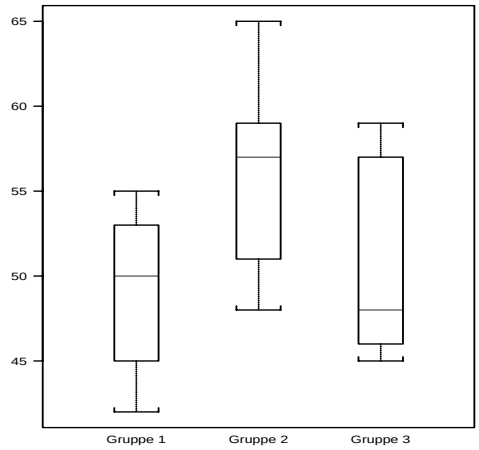


Abbildung 17.2: Boxplot von drei Gruppen mit großer Streuung innerhalb der Gruppen

Die Stichprobenvarianzen in den Gruppen für die erste Konstellation sind

$$s_1^2 = 7.5, \quad s_2^2 = 12.5, \quad s_3^2 = 2.5\,.$$

Für die Gruppen der zweiten Konstellation erhält man folgende Stichprobenvarianzen:

$$s_1^2 = 29.5, \quad s_2^2 = 45.0, \quad s_3^2 = 42.5\,.$$

□

Wir müssen also neben der Streuung zwischen den Gruppen die Streuung innerhalb der Gruppen berücksichtigen. Die Streuung innerhalb der i-ten Gruppe messen wir durch

$$\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2\,. \tag{17.5}$$

Summieren wir (17.5) über alle Gruppen, so erhalten wir

$$\boxed{SS_W = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2\,.} \tag{17.6}$$

Wir nennen SS_W auch **Streuung innerhalb der Gruppen**.

Beispiel 147 (fortgesetzt von Seite 489)
Es gilt $SS_W = 56720.17$. □

Die Gesamtstreuung messen wir durch:

$$\boxed{SS_T = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2} \,. \tag{17.7}$$

Beispiel 147 (fortgesetzt von Seite 492)
Es gilt $SS_T = 65786.2$. □

Im Beispiel gilt

$$SS_T = SS_B + SS_W \,. \tag{17.8}$$

Dies ist kein Zufall. Diese Beziehung gilt allgemein, wie man folgendermaßen sieht:

$$\begin{aligned}
SS_T &= \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i + \bar{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{I} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 + 2 \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)(\bar{y}_i - \bar{y}) \\
&= \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{I} n_i (\bar{y}_i - \bar{y})^2 + 2 \sum_{i=1}^{I} (\bar{y}_i - \bar{y}) \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i) \\
&= \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{I} n_i (\bar{y}_i - \bar{y})^2 \\
&= SS_B + SS_W \,.
\end{aligned}$$

Hierbei haben wir die folgende Beziehung berücksichtigt:

$$\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i) = \sum_{j=1}^{n_i} y_{ij} - \sum_{j=1}^{n_i} \bar{y}_i = n_i \bar{y}_i - n_i \, \bar{y}_i = 0 \,.$$

Eine geeignete Teststatistik erhält man nun, indem man die mittleren Streuungen vergleicht, wobei der Mittelwert unter der Nebenbedingung bestimmt

wird, wie viele der Summanden frei gewählt werden können. Die Streuung zwischen den Stichproben setzt sich aus I Summanden zusammen, von denen aber nur $I-1$ frei gewählt werden können, da sich der Mittelwert der I-ten Stichprobe aus

$$\bar{y}, \bar{y}_1, \ldots, \bar{y}_{I-1}$$

ergibt. Die Streuung innerhalb der Stichproben setzt sich aus n Summanden zusammen. In der i-ten Stichprobe ergibt sich aber y_{in_i} aus der Kenntnis von

$$y_{i1}, \ldots, y_{in_i-1}, \bar{y}_i\,.$$

Somit sind von den N Summanden nur $n-I$ frei wählbar. Wir erhalten also $MSS_B = SS_B/(I-1)$ und $MSS_W = SS_W/(N-I)$.

Beispiel 147 (fortgesetzt von Seite 492)
Es gilt $MSS_B = 4533.013$ und $MSS_W = 2025.72$. □

Die Teststatistik ist

$$F = \frac{MSS_B}{MSS_W} = \frac{\dfrac{1}{I-1}\displaystyle\sum_{i=1}^{I} n_i \left(\bar{Y}_i - \bar{Y}\right)^2}{\dfrac{1}{n-I}\displaystyle\sum_{i=1}^{I}\sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_i\right)^2}. \tag{17.9}$$

Ist die mittlere Streuung zwischen den Stichproben groß im Verhältnis zur mittleren Streuung innerhalb der Stichproben, so wird die Nullhypothese identischer Erwartungswerte abgelehnt. Unter der Nullhypothese ist die Teststatistik in (17.9) F-verteilt mit $I-1$ und $n-I$ Freiheitsgraden.
Wir lehnen die Hypothese (17.1) zum Niveau α ab, wenn gilt $F > F_{I-1,N-I;1-\alpha}$, wobei $F_{I-1,N-I;1-\alpha}$ das $1-\alpha$-Quantil der F-Verteilung mit $I-1$ und $N-I$ Freiheitsgraden ist.

Beispiel 147 (fortgesetzt von Seite 493)
Es gilt

$$F = \frac{4533.013}{2025.72} = 2.2377.$$

Der Tabelle 20.9 auf Seite 553 entnehmen wir $F_{2,28;0.95} = 3.34$. Wir lehnen die Hypothese (17.1) also nicht ab. □

Man spricht auch vom F-Test. Da die Teststatistik das Verhältnis von zwei Schätzern der Varianz σ^2 ist, spricht man von **Varianzanalyse**. Die Ergebnisse einer Varianzanalyse werden in einer ANOVA-Tabelle zusammengestellt. Dabei steht ANOVA für Analysis Of Variance. Tabelle 17.1 zeigt den allgemeinen Aufbau einer ANOVA-Tabelle.

Tabelle 17.1: Allgemeiner Aufbau einer ANOVA-Tabelle

Quelle der Variation	Quadrat-summen	Freiheits-grade	Mittlere Qua-dratsummen	F
zwischen den Gruppen	SS_B	$I-1$	MSS_B	MSS_B/MSS_W
innerhalb der Gruppen	SS_W	$n-I$	MSS_W	
Gesamt	SS_T	$n-1$		

Beispiel 147 (fortgesetzt von Seite 493)
In Tabelle 17.2 ist die ANOVA-Tabelle zu finden.

Tabelle 17.2: ANOVA-Tabelle für den Vergleich des Merkmals Mathematische Grundbildung in den 3 Gruppen

Quelle der Variation	Quadrat-summen	Freiheits-grade	Mittlere Qua-dratsummen	F
zwischen den Gruppen	9066.03	2	4533.013	2.2377
innerhalb der Gruppen	56720.17	28	2025.720	
Gesamt	65786.2	30		

□

17.2 Der Kruskal-Wallis-Test

Ist die Annahme der Normalverteilung nicht gerechtfertigt, so sollte man einen nichtparametrischen Test durchführen. Am bekanntesten ist der *Kruskal-Wallis-Test*. Dieser beruht auf der Annahme, dass die Beobachtungen y_{ij}, $i = 1, \ldots, I$, $j = 1, \ldots, n_i$ Realisationen von unabhängigen Zufallsvariablen Y_{ij}, $i = 1, \ldots, I$, $j = 1, \ldots, n_i$ mit stetiger Verteilungsfunktion sind. Es ist zu testen

$$H_0: \quad \text{Die Verteilungen in allen Gruppen sind identisch} \tag{17.10}$$

gegen

$H_1:$ Mindestens zwei Gruppen unterscheiden sich hinsichtlich der Lage.

Der Kruskal-Wallis-Test beruht auf den *Rängen* R_{ij} der y_{ij}, $i = 1, \ldots, I$, $j = 1 \ldots, n_i$, unter allen Beobachtungen. Dabei ist der Rang R_{ij} gleich der Anzahl der Beobachtungen, die kleiner oder gleich y_{ij} sind. Sind Beobachtungen identisch, so spricht man von *Bindungen*. In diesem Fall vergibt man für die gebundenen Werte *Durchschnittsränge*.

Beispiel 147 (fortgesetzt von Seite 494)
Schauen wir uns noch einmal die Daten an:

Gruppe 1:	536 557 514 446 515 510 529 498
Gruppe 2:	533 520 334 514 490 517 514 533 547 537 499 454 493
Gruppe 3:	447 529 503 457 463 387 470 478 476 488

Die Ränge in den einzelnen Gruppen sind:

Gruppe 1:	28 31 19 3 21 17 24.5 14
Gruppe 2:	26.5 23 1 19 12 22 19 26.5 30 29 15 5 13
Gruppe 3:	4 24.5 16 6 7 2 8 10 9 11

□

Beim Kruskal-Wallis-Test werden nun für $i = 1, \ldots, I$ die Rangsummen R_i in den einzelnen Gruppen bestimmt:

$$R_i \;=\; \sum_{j=1}^{n_i} R_{ij}.$$

Beispiel 147 (fortgesetzt von Seite 495)
Es gilt

$$R_1 = 157.5, \quad R_2 = 241, \quad R_3 = 97.5.$$

□

Diese Rangsummen werden mit ihren Erwartungswerten $E(R_i)$ unter (17.10) verglichen. Wenn keine Bindungen vorliegen, so werden bei n Beobachtungen die Ränge $1, \ldots, n$ vergeben. Trifft (17.10) zu, so ist für eine Beobachtung jeder Rang gleichwahrscheinlich. Es gilt also

$$P(R_{ij} = k) = \frac{1}{n}$$

für $k = 1, \ldots, n$, $i = 1, \ldots, I$ und $j = 1, \ldots, n_i$. Der erwartete Rang $E(R_{ij})$ von Y_{ij} ist dann

$$E(R_{ij}) \;=\; \sum_{k=1}^{n} k \frac{1}{n} = \frac{n(n+1)}{2n} = \frac{n+1}{2}.$$

Die erwartete Rangsumme der i-ten Gruppe ist somit

$$E(R_i) \;=\; E\left(\sum_{j=1}^{n_i} R_{ij}\right) = \sum_{j=1}^{n_i} E(R_{ij}) = \sum_{j=1}^{n_i} \frac{n+1}{2} = \frac{n_i(n+1)}{2}.$$

Beispiel 147 (fortgesetzt von Seite 496)
Mit $n = 31$, $n_1 = 8$, $n_2 = 13$ und $n_3 = 10$ gilt

$$E(R_1) = 128, \quad E(R_2) = 208, \quad E(R_3) = 160.$$

□

Die Teststatistik des Kruskal-Wallis-Tests vergleicht die Rangsummen R_i mit ihren Erwartungswerten $E(R_i)$. Sie lautet:

$$\boxed{H = \frac{12}{n(n+1)} \sum_{i=1}^{I} \frac{1}{n_i} \left(R_i - \frac{n_i(n+1)}{2}\right)^2} \tag{17.11}$$

Beispiel 147 (fortgesetzt von Seite 496)
Es gilt

$$\begin{aligned} H &= \frac{12}{31 \cdot 32} \left[\frac{(157.5 - 128)^2}{8} + \frac{(241 - 208)^2}{13} + \frac{(97.5 - 160)^2}{10}\right] \\ &= 7.054542. \end{aligned}$$

□

Wir lehnen die Hypothese (17.10) ab, wenn gilt $H \geq h_{1-\alpha}$.
Dabei ist $h_{1-\alpha}$ das $1-\alpha$-Quantil der Verteilung von H. Die Verteilung von H ist für kleine Werte von n bei *Büning, Trenkler: Nichtparametrische Statistische Methoden* tabelliert.
Für große Stichprobenumfänge ist H approximativ chiquadratverteilt mit $I-1$ Freiheitsgraden. Wir lehnen (17.10) ab, wenn gilt $H \geq \chi^2_{I-1,1-\alpha}$.
Dabei ist $\chi^2_{I-1,1-\alpha}$ das $1-\alpha$-Quantil der χ^2-Verteilung mit $I-1$ Freiheitsgraden.
Im Beispiel liegen Bindungen vor. In diesem Fall wird H modifiziert zu

$$\boxed{H^* = \frac{H}{1 - \frac{1}{n^3 - n} \sum_{l=1}^{r} (b_l^3 - b_l)}}. \tag{17.12}$$

Dabei ist r die Anzahl der Gruppen mit identischen Beobachtungen und b_l die Anzahl der Beobachtungen in der l-ten Bindungsgruppe. Wir lehnen (17.10) im Fall von Bindungen ab, wenn gilt $H^* \geq \chi^2_{I-1,1-\alpha}$.

Beispiel 147 (fortgesetzt von Seite 496)
Der Wert 514 kommt dreimal und die Werte 529 und 533 kommen jeweils zweimal vor. Somit gibt es 2 Bindungsgruppen mit zwei Beobachtungen und eine Bindungsgruppe mit drei Beobachtungen. Hieraus folgt

$$1 - \frac{1}{n^3 - n} \sum_{l=1}^{r} (b_l^3 - b_l) = 0.99879.$$

Also ist $H^* = 7.0631$. Der Tabelle 20.4 auf Seite 550 entnehmen wir $\chi^2_{2,0.95} = 5.99$. Wir lehnen die Hypothese (17.10) zum Niveau 0.05 also ab. □

17.3 Varianzanalyse in R

Wir wollen die Varianzanalyse für das Beispiel 147 auf Seite 487 durchführen. Im Zweistichprobenproblem haben wir für jede Stichprobe eine Variable erzeugt. Liegen mehr als zwei Stichproben vor, so gehen wir anders vor. Wir weisen alle Werte einer Variablen in der Reihenfolge der Gruppen zu. Diese Variable nennen wir `Punkte`.

```
> Punkte<-c(536,557,514,446,515,510,529,498,533,520)
> Punkte<-c(Punkte,334,514,490,517,514,533,547,537,499,454,493)
> Punkte<-c(Punkte,447,529,503,457,463,387,470,478,476,488)
```

Die ersten 8 Komponenten von `Punkte` gehören zur ersten Gruppe, die nächsten 13 Komponenten zur zweiten Gruppe und die letzte 10 Komponenten zur dritten Gruppe. Wir erzeugen einen Vektor `A`, bei dem die ersten 8 Komponenten gleich 1, die nächsten 13 Komponenten gleich 2 und die letzten 10 Komponenten gleich 3 sind. Hierzu benutzen wir die Funktion `rep`. Der Aufruf

```
rep(x,times)
```

erzeugt einen Vektor, in dem das Argument `x` `times`-mal wiederholt wird:

```
> rep(1,8)
[1] 1 1 1 1 1 1 1 1
```

Dabei können `x` und `times` Vektoren sein. Sind `x` und `times` gleich lange Vektoren, so wird `x[i]` `times[i]`-mal wiederholt.

```
> A<-rep(1:3,c(8,13,10))
> A
 [1] 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3
 [28] 3 3 3 3
```

Nun müssen wir aus A nur noch einen Faktor machen. Dies leistet die Funktion `factor`.

```
> A<-factor(A)
> A
 [1] 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3
 [28] 3 3 3 3
Levels: 1 2 3
```

Um eine Varianzanalyse durchführen zu können, verwenden wir die Funktion `anova`. Die Funktion `anova` hat das Argument `formula`. Mit diesem können wir das Modell durch eine Formel spezifizieren. Wie sieht diese Formel für eine einfaktorielle Varianzanalyse für das Beispiel aus? Wir wollen die Variable `Punkte` durch den Faktor `A` erklären. Hierfür schreiben wire Formel durch `Punkte A`. Auf der linken Seite der Formel steht die zu erklärende Variable. Das Zeichen ~ liest man als 'wird modelliert durch'. Auf der rechten Seite steht die erklärende Variable. Wollen wir also die besprochene Varianzanalyse duchführen, so geben wir ein:

```
> e<-aov(Punkte~A)
```

Die Variable `e` enthält das Ergebnis der Varianzanalyse. Rufen wir die Funktion `summary` mit `e` auf, so erhalten wir die ANOVA-Tabelle.

```
> summary(e)
             Df Sum Sq Mean Sq F value Pr(>F)
 A            2   9066    4533  2.2377 0.1254
 Residuals   28  56720    2026
```

Unter `Pr(>F)` wird die überschreitungswahrscheinlichkeit ausgegeben. Wir lehnen also H_0 zum Niveau 0.05 nicht ab.
Für den Kruskal-Wallis-Test gibt es die Funktion `kruskal.test`, die folgendermaßen aufgerufen wird:

```
kruskal.test(y, groups)
```

Die Daten stehen im Vektor `y`. Die i-te Komponente des Vektors `groups` gibt an, zu welcher Gruppe die i-te Beobachtung gehört. Wir geben also ein

```
> kruskal.test(Punkte,A)
```

und erhalten folgendes Ergebnis:

```
        Kruskal-Wallis rank sum test

 data:  Punkte and A
 Kruskal-Wallis chi-squared = 7.0631, df = 2,
 p-value = 0.02926
```

`R` berücksichtigt das Vorhandensein von Bindungen und bestimmt die Teststatistik H^* in Gleichung (17.12) auf Seite 497. Die überschreitungswahrscheinlichkeit beträgt 0.0293. Somit wird die Nullhypothese (17.10) auf Seite 495 zum Signifikanzniveau $\alpha = 0.05$ abgelehnt.

Kapitel 18

Unabhängigkeit und Homogenität

18.1 Unabhängigkeit

Im Rahmen der Wahrscheinlichkeitsrechnung ist das Konzept der Unabhängigkeit von zentraler Bedeutung. Die Ereignisse A und B sind genau dann unabhängig, wenn gilt

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

Wir können dieses Konzept auf qualitative Merkmale übertragen. Wir betrachten zwei qualitative Merkmale A und B mit den Merkmalsausprägungen $A_1, \ldots, A_r$ und $B_1, \ldots, B_c$.
Sei $p_{ij} = P(A_i, B_j)$ die Wahrscheinlichkeit, dass ein zufällig aus der Grundgesamtheit ausgewähltes Objekt die Merkmalsausprägung A_i beim Merkmal A und die Merkmalsausprägung B_j beim Merkmal B aufweist.
Die Merkmale A und B mit den Merkmalsausprägungen $A_1, \ldots, A_r$ und $B_1, \ldots, B_c$ sind genau dann unabhängig, wenn für $i = 1, \ldots, r$, $j = 1, \ldots, c$ gilt

$$P(A_i \cap B_j) = P(A_i)P(B_j)\,.$$

Mit

$$p_{i\cdot} = P(A_i) = p_{i1} + \ldots + p_{ic} = \sum_{j=1}^{c} p_{ij}$$

und

$$p_{\cdot j} = P(B_j) = p_{1j} + \ldots + p_{rj} = \sum_{i=1}^{r} p_{ij}$$

können wir dies auch schreiben als

$$p_{ij} = p_{i\cdot} p_{\cdot j} \,. \tag{18.1}$$

Dabei ist $p_{i\cdot}$ die Wahrscheinlichkeit, dass das Merkmal A die Merkmalsausprägung A_i und $p_{\cdot j}$ die Wahrscheinlichkeit, dass das Merkmal B die Merkmalsausprägung B_j aufweist.
Wir wollen nun überprüfen, ob die Merkmale A und B unabhängig sind, wenn eine Zufallsstichprobe vorliegt.
Das Testproblem lautet

H_0: Die Merkmale A und B sind unabhängig,

H_1: Die Merkmale A und B sind nicht unabhängig.

Wir beobachten die absoluten Häufigkeiten n_{ij} für das gleichzeitige Auftreten der Merkmalsausprägung A_i des Merkmals A und der Merkmalsausprägung B_j des Merkmals B.
Außerdem ist für $i = 1, \ldots, r$

$$n_{i\cdot} = \sum_{j=1}^{c} n_{ij}$$

und für $j = 1, \ldots, c$

$$n_{\cdot j} = \sum_{i=1}^{r} n_{ij} \,.$$

Dabei ist $n_{i\cdot}$ die absolute Häufigkeit von A_i und $n_{\cdot j}$ die absolute Häufigkeit von B_j.
Diese Informationen stellen wir in einer Kontingenztabelle zusammen. Tabelle 18.1 zeigt den allgemeinen Aufbau einer zweidimensionalen Kontingenztabelle.

Tabelle 18.1: Allgemeiner Aufbau einer zweidimensionalen Kontingenztabelle

A \ B	B_1	B_2	$\ldots$	B_c	
A_1	n_{11}	n_{12}	$\cdots$	n_{1c}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2c}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	$n_{r\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot c}$	n

Beispiel 148
Erstsemester wurden unter anderem gefragt, ob sie nach dem Abitur eine Berufsausbildung abgeschlossen und ob sie den Leistungskurs Mathematik absolviert haben. Wir bezeichnen das Merkmal `Berufsausbildung` mit A und das Merkmal `MatheLK` mit B. Tabelle 18.2 zeigt die Kontingenztabelle.

Tabelle 18.2: Kontingenztabelle der Merkmale `Berufsausbildung` und `MatheLK` bei Studenten

`MatheLK` / `Berufausbildung`	ja	nein	
ja	37	55	92
nein	115	144	259
	152	199	351

Es gilt

$$n_{11} = 37 \qquad n_{12} = 55$$
$$n_{21} = 115 \qquad n_{22} = 144$$

und

$$n_{1\cdot} = 92, \qquad n_{2\cdot} = 259, \qquad n_{\cdot 1} = 152, \qquad n_{\cdot 2} = 199\,.$$

□

Wir können die Hypothese der Unabhängigkeit mit dem Chiquadrattest überprüfen. Bei diesem vergleichen wir die beobachteten Häufigkeiten n_{ij} mit den Häufigkeiten $\tilde{n}_{ij}$, die wir erwarten, wenn die Merkmale A und B unabhängig sind. Es gilt

$$\tilde{n}_{ij} = np_{ij} \overset{(18.1)}{=} n\, p_{i\cdot} p_{\cdot j}\,. \tag{18.2}$$

In Gleichung (18.2) sind die Wahrscheinlichkeiten $p_{i\cdot}$ und $p_{\cdot j}$ unbekannt. Wir schätzen sie durch die entsprechenden relativen Häufigkeiten. Wir schätzen $p_{i\cdot}$ durch $n_{i\cdot}/n$ und $p_{\cdot j}$ durch $n_{\cdot j}/n$. Setzen wir diese Schätzer in (18.2) ein, so erhalten wir die folgenden **geschätzten erwarteten Häufigkeiten**, die wir ebenfalls mit $\tilde{n}_{ij}$ bezeichnen:

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

Dies können wir vereinfachen zu

$$\tilde{n}_{ij} = \frac{n_{i\cdot}\, n_{\cdot j}}{n}\,. \tag{18.3}$$

Beispiel 148 (fortgesetzt von Seite 503)
Die geschätzten erwarteten Häufigkeiten sind

$$\tilde{n}_{11} = \frac{92 \cdot 152}{351} = 39.84 \quad , \quad \tilde{n}_{12} = \frac{92 \cdot 199}{351} = 52.16\,,$$

$$\tilde{n}_{21} = \frac{259 \cdot 152}{351} = 112.16 \quad , \quad \tilde{n}_{22} = \frac{259 \cdot 199}{351} = 146.84\,.$$

□

Die Teststatistik des χ^2-Unabhängigkeitstests lautet

$$\boxed{X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - \tilde{n}_{ij})^2}{\tilde{n}_{ij}}}\,, \,. \tag{18.4}$$

Beispiel 148 (fortgesetzt von Seite 504)
Es gilt

$$\begin{aligned} X^2 &= \frac{(37 - 39.84)^2}{39.84} + \frac{(55 - 52.16)^2}{52.16} + \frac{(115 - 112.16)^2}{112.16} + \frac{(144 - 146.84)^2}{146.84} \\ &= 0.484. \end{aligned}$$

□

Die Entscheidungsregel lautet:
Wir lehnen H_0 ab, wenn gilt $X^2 \geq \chi^2_{(r-1)(c-1);1-\alpha}$.
Dabei ist $\chi^2_{(r-1)(c-1);1-\alpha}$ das $1-\alpha$-Quantil der χ^2-Verteilung mit $(r-1)(c-1)$ Freiheitsgraden.

Beispiel 148 (fortgesetzt von Seite 504)
Sei $\alpha = 0.05$. Der Tabelle 20.4 auf Seite 550 entnehmen wir $\chi^2_{1;0.95} = 3.84$. Wir lehnen H_0 also nicht ab. □

18.2 Homogenität

Man kann den Chiquadrattest auch als Test auf Homogenität verwenden. Hierbei wird die Verteilung eines kategorialen Merkmals Y mit c Kategorien in r Gruppen betrachtet.

Es soll überprüft werden, ob die Verteilung von Y in allen r Gruppen identisch ist.
Da die Merkmale kategorial sind, können wir nur zählen, wie viele Personen bzw. Objekte in die einzelnen Kategorien des Merkmals fallen. Im Folgenden ist n_{ij} die Anzahl der Personen bzw. Objekte, die sich in der i-ten Gruppe in der j-ten Kategorie von Y befinden. Wir können die Daten also folgendermaßen in einer Kontingenztabelle anordnen.

Tabelle 18.3: Kontingenztabelle eines qualitativen Merkmals in r Gruppen

Kategorie / Gruppe	1	2	...	c	
1	n_{11}	n_{12}	$\cdots$	n_{1c}	$n_{1\cdot}$
2	n_{21}	n_{22}	$\cdots$	n_{2c}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	$n_{r\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot c}$	n

Beispiel 149
Wir schauen uns noch einmal das Beispiel 37 auf Seite 148 an. In der ersten Statistik I Vorlesung im WS 96/97 wurden 255 Studenten nach ihrem Wahlverhalten und ihrem Geschlecht befragt. Die absoluten Häufigkeiten sind in Tabelle 18.4 zu finden.

Tabelle 18.4: Wahlverhalten von weiblichen und männlichen Erstsemestern

Wahl / Geschlecht	CDU	SPD	FDP	Grüne	keine	weiß nicht
w	13	10	3	11	5	23
m	55	30	20	26	24	35

Es soll überprüft werden, ob sich das Wahlverhalten der Männer und Frauen unterscheidet. □

Bevor wir den Test durchführen, schauen wir uns die Verteilung des Merkmals Y in den Gruppen an. Wir bilden also in der i-ten Gruppe folgende bedingte

relative Häufigkeiten

$$h_{j|i} = \frac{n_{ij}}{n_{i\cdot}}$$

Beispiel 149 (fortgesetzt von Seite 505)
Wir schauen uns die Verteilung des Wahlverhaltens bei den Frauen und bei den Männern an. Die bedingten relativen Häufigkeiten sind in Tabelle 18.5 zu finden.

Tabelle 18.5: Verteilung des Wahlverhaltens bei weiblichen und männlichen Erstsemestern

Geschlecht \ Wahl	CDU	SPD	FDP	Grüne	keine	weiß nicht
w	0.200	0.154	0.046	0.169	0.077	0.354
m	0.289	0.158	0.105	0.137	0.126	0.184

Wir sehen, dass sich die Verteilungen in den Kategorien `FDP`, `keine` und `weiß nicht` beträchtlich unterscheiden.
Die Profile in Abbildung 4.7 auf Seite 149, das vergleichende Säulendiagramm in Abbildung 4.8 auf Seite 150 und der Mosaikplot in Abbildung 4.9 auf Seite 150 deuten auf einen Unterschied hin.

□

Mit dem Chiquadrat-Unabhängigkeitstest können wir die Homogenität überprüfen. Wir vergleichen die beobachteten absoluten Häufigkeiten n_{ij} mit den absoluten Häufigkeiten, die wir erwarten, wenn Homogenität vorliegt. Liegt Homogenität vor, so sollte die Verteilung von Y nicht von der Gruppe abhängen. Somit ist Homogenität eine spezielle Unabhängigkeit. Die erwarteten Häufigkeiten sind also

$$\tilde{n}_{ij} = \frac{n_{i\cdot}\, n_{\cdot j}}{n}\,. \tag{18.5}$$

Beispiel 149 (fortgesetzt)
Es gilt

$$n_{1\cdot} = 65 \qquad n_{2\cdot} = 190$$

und

$$n_{\cdot 1} = 68 \qquad n_{\cdot 2} = 40 \qquad n_{\cdot 3} = 23 \qquad n_{\cdot 4} = 37 \qquad n_{\cdot 5} = 29 \qquad n_{\cdot 6} = 58$$

Also gilt

$$\tilde{n}_{11} = \frac{65 \cdot 68}{255} = 17.33 \qquad \tilde{n}_{21} = \frac{190 \cdot 68}{255} = 50.67$$

$$\tilde{n}_{12} = \frac{65 \cdot 40}{255} = 10.2 \qquad \tilde{n}_{22} = \frac{190 \cdot 40}{255} = 29.8$$

$$\tilde{n}_{13} = \frac{65 \cdot 23}{255} = 5.86 \qquad \tilde{n}_{23} = \frac{190 \cdot 23}{255} = 17.14$$

$$\tilde{n}_{14} = \frac{65 \cdot 37}{255} = 9.43 \qquad \tilde{n}_{24} = \frac{190 \cdot 37}{255} = 27.57$$

$$\tilde{n}_{15} = \frac{65 \cdot 29}{255} = 7.39 \qquad \tilde{n}_{25} = \frac{190 \cdot 29}{255} = 21.61$$

$$\tilde{n}_{16} = \frac{65 \cdot 58}{255} = 14.78 \qquad \tilde{n}_{26} = \frac{190 \cdot 58}{255} = 43.22$$

□

Die Teststatistik ist

$$\boxed{X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - \tilde{n}_{ij})^2}{\tilde{n}_{ij}}}.$$

Beispiel 149 (fortgesetzt)
Es gilt $X^2 = 10.85$. □

Die Entscheidungsregel lautet:
Wir lehnen H_0 ab, wenn gilt $X^2 \geq \chi^2_{(r-1)(c-1);1-\alpha}$.
Dabei ist $\chi^2_{(r-1)(c-1);1-\alpha}$ das $1-\alpha$-Quantil der χ^2-Verteilung mit $(r-1)(c-1)$ Freiheitsgraden.

Beispiel 149 (fortgesetzt)
Tabelle 20.4 auf Seite 550 entnehmen wir $\chi^2_{5;0.95} = 11.07$. Also lehnen wir H_0 zum Niveau 0.05 nicht ab. □

Besitzt das Merkmal nur zwei Kategorien, so vereinfacht sich die Teststatistik beträchtlich. In diesem Fall sind in der ersten Kategorie alle Personen bzw. Objekte, die eine bestimmte Eigenschaft A besitzen. In der zweiten Kategorie sind alle Personen bzw. Objekte, die diese Eigenschaft nicht besitzen. Man

will überprüfen, ob der Anteil der Personen bzw. Objekte, die die Eigenschaft A aufweisen, in zwei Grundgesamtheiten identisch ist.
Wir stellen die Daten in einer $(2, 2)$-Kontingenztabelle zusammen:

Tabelle 18.6: Kontingenztabelle eines qualitativen Merkmals in r Gruppen

Kategorie / Gruppe	1	2	
1	n_{11}	n_{12}	$n_{1.}$
2	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	n

Beispiel 150
In der *Süddeutschen Zeitung* vom 1.7.2003 wird über eine Studie berichtet, in der die Nebenwirkungen von Hormonbehandlungen untersucht wurden. Hier findet sich folgender Text

> Insgesamt hatten in der Studie 8506 Frauen zwischen 50 und 80 Hormone genommen, weitere 8102 ein Scheinmedikament (Placebo). Nach im Durchschnitt 5.6 Jahren waren 199 Frauen unter der Hormontherapie an aggressivem Brustkrebs erkrankt, von den Frauen der Placebo-Gruppe nur 150.

Die Studie wird im *Journal of the American Medical Association, Bd. 289* beschrieben.
Es soll getestet werden, ob die Wahrscheinlichkeit, an aggressivem Brustkrebs zu erkranken, in beiden Gruppen identisch ist. Wir stellen die Daten in einer Kontingenztabelle zusammen.

Tabelle 18.7: Brustkrebs in Abhängigkeit von Hormonbehandlung

Brustkrebs / Gruppe	ja	nein	
Placebo	150	7952	8102
Hormone	199	8307	8506
	349	16259	16608

□

Die Teststatistik X^2 können wir in diesem Fall vereinfachen zu

$$X^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}}$$

Beispiel 150 (fortgesetzt)
Es gilt

$$X^2 = \frac{16608(150 \cdot 8307 - 7952 \cdot 199)^2}{8102 \cdot 8506 \cdot 349 \cdot 16259} = 4.806$$

□

Wir lehnen H_0 ab, wenn gilt $X^2 \geq \chi^2_{1;1-\alpha}$.
Dabei ist $\chi^2_{1;1-\alpha}$ das $1-\alpha$-Quantil der χ^2-Verteilung mit einem Freiheitsgrad.

Beispiel 150 (fortgesetzt)
Wegen $\chi^2_{1;0.95} = 3.84$ lehnen wir H_0 ab. □

Die Teststatistik ist approximativ chiquadratverteilt. Ein exakter Test für das eben geschilderte Problem ist der **Fisher-Test**. Dieser wird detailliert Büning & Trenkler (1994) beschrieben.

18.3 Unabhängigkeit und Homogenität in R

Wir geben eine Kontingenzabelle als Matrix ein. Eine Matrix ist ein rechteckiges Zahlenschema, das aus m Zeilen und n Spalten besteht. In R können wir mit der Funktion `matrix` eine Matrix eingeben. Die Funktion `matrix` wird aufgerufen durch:

```
matrix(data=NA,nrow=1,ncol=1,byrow=FALSE,dimnames=NULL)
```

Dabei enthält das Element `data` die Elemente der Matrix als Vektor. Die Matrix wird dabei spaltenweise aufgebaut. Wird das Argument `byrow` auf `TRUE` gesetzt, so wird sie zeilenweise aufgebaut. Mit den Argumenten `nrow` und `ncol` gibt man die Anzahl der Zeilen und Spalten an. Das Argument `dimnames` erlaubt die Eingabe von Namen für die Zeilen und Spalten der Matrix. Die Kontingenztabelle in Tabelle 18.2 auf Seite 503 geben wir also folgendermaßen ein:

```
> m<-matrix(c(37,115,55,144),2,2)
> m
```

```
     [,1] [,2]
[1,]   37   55
[2,]  115  144
```

Mit der Funktion `chisq.test` kann man den Chiquadrattest durchführen. Man ruft sie mit der Variablen auf, die die Kontingenztabelle als Matrix enthält. Außerdem muss man das Argument `correct` auf den Wert `FALSE` setzen.

```
> chisq.test(m,correct=FALSE)

        Pearson's Chi-squared test

data:  m
X-squared = 0.4841, df = 1, p-value = 0.4866
```

Schauen wir uns den Homogenitätstest an. Wir weisen die Daten aus Tabelle 18.4 auf Seite 505 der Variablen `wahl` zu

```
> wahl<-matrix(c(13,55,10,30,3,20,11,26,5,24,23,35),2,6)
> wahl
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]   13   10    3   11    5   23
[2,]   55   30   20   26   24   35
```

und rufen die Funktion `chisq.test` auf

```
> chisq.test(wahl,correct=FALSE)

        Pearson's Chi-squared test

data:  wahl
X-squared = 10.8515, df = 5, p-value = 0.0544
```

Wir betrachten noch die Daten in Tabelle 18.7 auf Seite 508, weisen sie der Matrix `h` zu

```
> h<-matrix(c(150,199,7952,8307),2,2)
```

und rufen die Funktion `chisq.test` auf

```
> chisq.test(h,correct=FALSE)
```

```
        Pearson's Chi-squared test

data:  h
X-squared = 4.806, df = 1, p-value = 0.02836
```

Wir können aber auch mit der Funktion `fisher.test` den Fisher-Test durchführen:

```
        Fisher's Exact Test for Count Data

data:  h
p-value = 0.03033
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
0.6312966 0.9805229
sample estimates:
 odds ratio
  0.7874322
```

Wir sehen, dass die Überschreitungswahrscheinlichkeit beim exakten Test ein wenig größer ist. Beide Tests kommen aber zur gleichen Entscheidung.

Kapitel 19

Das lineare Modell

19.1 Das Modell

Wir haben im letzten Kapitel einen Test kennengelernt, mit dem man überprüfen kann, ob zwei qualitative Merkmale abhängig sind. Nun wollen wir den Zusammenhang zwischen zwei stetigen Zufallsvariablen X und Y betrachten. In Statistik I haben wir den Korrelationskoeffizienten von Bravais-Pearson als Maßzahl für den linearen Zusammenhang zwischen zwei stetigen Merkmalen betrachtet. Dieser wurde aus n Punktepaaren $(x_1, y_1), \ldots, (x_n, y_n)$ berechnen. Wir können diese Punktepaare als Realisationen der bivariaten Zufallsvariablen (X, Y) betrachten. Unter diesem Aspekt ist der Korrelationskoeffizienten von Bravais-Pearson ein Schätzer des Korrelationskoeffizienten $\rho_{X,Y}$ zwischen X und Y.
Der Korrelationskoeffizient behandelt beide Zufallsvariablen gleich. Oft ist es aber so, dass eine der beiden Zufallsvariablen von der anderen abhängt. Wir bezeichnen die zu erklärende Zufallsvariable mit Y und die erklärende Zufallsvariable mit X.

Beispiel 151
Es soll untersucht werden wie der Angebotspreis Y eines VW Golfs vom Alter X abhängt. □

Wir können natürlich ansetzen

$$Y = f(X)$$

Dies bedeutet, dass zu jeder Ausprägungsmöglichkeit von X genau eine Ausprägungsmöglichkeit von Y gehört. Dies ist in der Empirie sicherlich fast nie der Fall, da neben X noch viele andere Variablen die Variable Y beeinflus-

sen, sodass man zum gleichen Wert von X unterschiedliche Werte von Y beobachtet.

Beispiel 151 (fortgesetzt)
In der Süddeutschen Zeitung wurden Ende Juli 1999 im Anzeigenteil 33 VW-Golf 3 angeboten. In Tabelle 19.1 sind deren Merkmale `Alter` in Jahren und `Angebotspreis` (in DM) zu finden.

Tabelle 19.1: Alter und Angebotspreis von 33 VW-Golf 3

Alter	Angebots-preis	Alter	Angebots-preis	Alter	Angebots-preis
2	21800	4	15900	5	14900
2	18800	4	17900	5	12400
2	20500	4	19500	6	12800
3	18900	4	16000	6	14900
3	21200	5	16500	6	12900
3	16800	5	15800	6	12800
3	17500	5	15900	6	13500
3	23800	5	16900	7	10950
3	16800	5	14800	7	12900
4	14500	5	15500	7	10800
4	19900	5	16500	7	11600

Wir sehen, dass zu jedem Wert des Alters mehrere Werte des Angebotspreises beobachtet werden. Neben dem Alter haben unter anderem auch der Kilometerstand und der Zustand des Autos einen Einfluss auf den Angebotspreis. □

Zu einem festen Wert x von X ist also die abhängige Variable eine Zufallsvariable, die von x und sicherlich auch anderen Variablen abhängt. Wir fassen alle anderen Einflussgrößen in einer Zufallsvariablen zusammmen, die wir mit ϵ bezeichnen. Dabei unterstellen wir, dass der Einfluss von ϵ auf Y klein ist. Wir postulieren also folgendes Modell für den Zusammenhang zwischen den beiden Variablen

$$Y = f(x) + \epsilon$$

Welche Annahmen sollen wir über ϵ machen? Wir unterstellen, dass die anderen Variablen, die wir in ϵ zusammen, keinen systematischen Einfluss auf

Y ausüben. Wir nehmen also an:

$$E(\epsilon) = 0$$

Hieraus folgt

$$E(Y) = f(x)\,.$$

Dies sieht man folgendermaßen:

$$E(Y) = E(f(x) + \epsilon) = f(x) + E(\epsilon) = f(x)$$

Wie sollen wir $f(x)$ spezifizieren? Am einfachsten zu interpretieren sind Geraden:

$$y = f(x) = \beta_0 + \beta_1 \cdot x$$

Dabei wird β_0 der Achsenabschnitt und β_1 die Steigung der Geraden genannt. Der Achsenabschnitt β_0 gibt den Wert von y an, für den $x = 0$ ist, denn

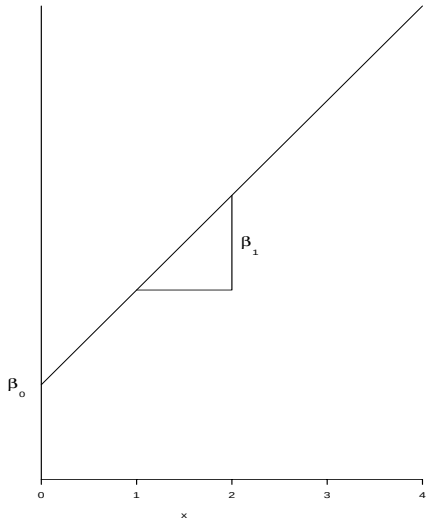
$$f(0) = \beta_0 + \beta_1 \cdot 0 = \beta_0\,.$$

Die Steigung β_1 gibt an, wie sich y ändert, wenn x um den Wert 1 erhöht wird, denn

$$\begin{aligned} f(x+1) - f(x) &= \beta_0 + \beta_1 \cdot (x+1) - (\beta_0 + \beta_1 \cdot x) \\ &= \beta_0 + \beta_1 \cdot x + \beta_1 - \beta_0 - \beta_1 \cdot x = \beta_1 \end{aligned}$$

Abbildung 19.1 verdeutlicht die Bedeutung von β_0 und β_1.

Abbildung 19.1: Was ist eine Gerade?



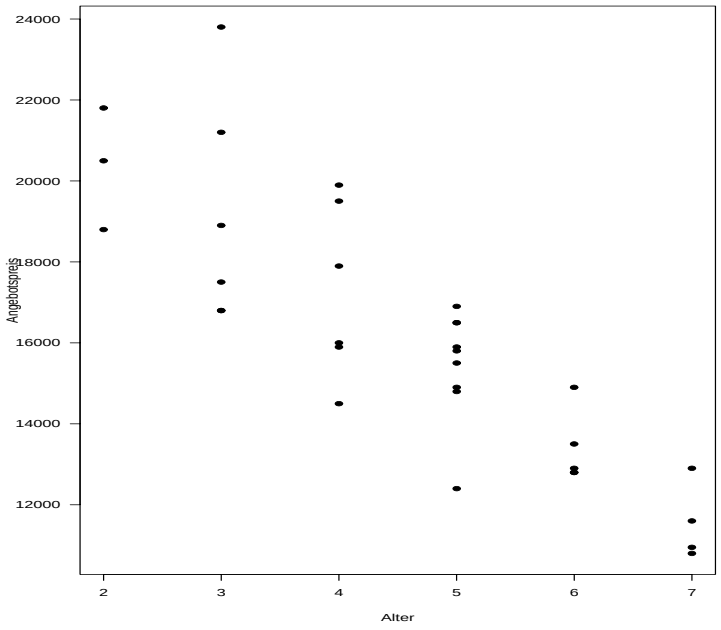
Wir unterstellen also folgendes Modell

$$\boxed{Y = \beta_0 + \beta_1 \cdot x + \epsilon} \tag{19.1}$$

In diesem Modell sind β_0 und β_1 unbekannte Parameter, die wir schätzen wollen. Hierzu beobachten wir n Punktepaare $(x_1, y_1), \ldots, (x_n, y_n)$ und stellen sie in einem Streudiagramm dar.

Beispiel 151 (fortgesetzt von Seite 514)
Abbildung 19.2 zeigt das Streudiagramm der Daten

Abbildung 19.2: Streudiagramm des Alters und des Angebotspreises von 33 Autos vom Typ Golf



Wir sehen, dass mit wachsendem Alter der Angebotspreis sinkt. □

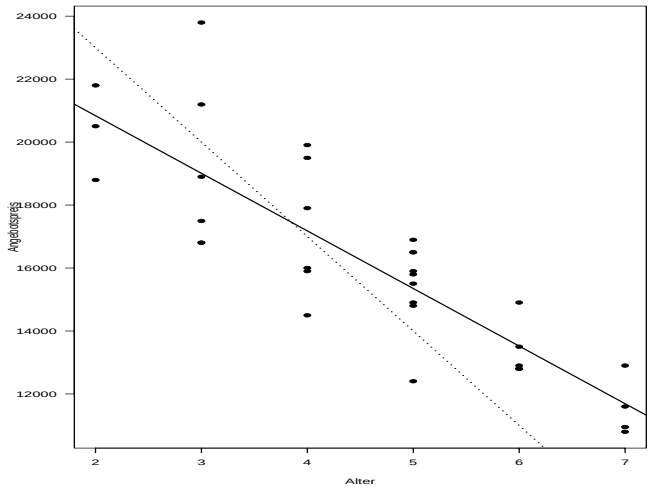
Unser Ziel ist es, durch die Punktewolke eine Gerade legen. Im nächsten Abschnitt werden wir ein Verfahren kennenlernen.

19.2 Die Methode der Kleinsten Quadrate

Wir wollen eine Gerade durch die Punktewolke legen. Liegen alle Punkte auf einer Geraden, so ist dies sehr einfach. In der Regel liegen die Punkte jedoch nicht auf einer Geraden, sodass wir die Wahl zwischen vielen Geraden haben. Die gewählte Gerade sollte die Tendenz in der Punktewolke sehr gut beschreiben. Sie sollte also möglichst nahe an allen Punkten liegen.

Beispiel 151 (fortgesetzt von Seite 516)
In Abbildung 19.3 sind zwei Geraden in das Streudiagramm eingezeichnet.

Abbildung 19.3: Streudiagramm des Alters und des Angebotspreises von 33 Autos vom Typ Golf



Wir sehen, dass die durchgezogenen Gerade die Tendenz viel besser beschreibt als die gestrichelte Gerade. □

Wir finden wir eine geeignete Gerade? Nehmen wir einmal an, dass wir die Gerade gefunden hätten. Den Achsenabschnitt bezeichnen wir mit $\hat{\beta}_0$ und die Steigung mit $\hat{\beta}_1$. Dann lautet die Gerade

$$\hat{y} \quad = \quad \hat{\beta}_0 + \hat{\beta}_1 \cdot x \tag{19.2}$$

Zu einem Wert x_i gibt es einen beobachteten Wert y_i und einen auf der Geraden liegenden Wert $\hat{y}_i$. Der Abstand des Punktes (x_i, y_i) von der Geraden

ist $|e_i|$ mit

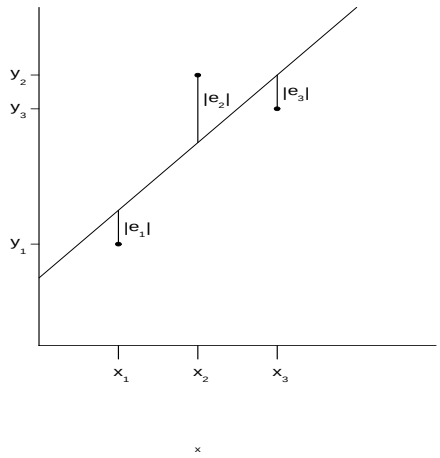
$$\boxed{e_i = y_i - \hat{y}}\,. \tag{19.3}$$

Wir nennen e_i **Residuum**.
Es gilt

$$e_i \;=\; y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\,.$$

In Abbildung 19.4 sind drei Punkte mit einer zugehörigen Geraden eingezeichnet. Die Längen der senkrechten Strecken sind $|e_1|$,$|e_2|$ und $|e_3|$.

Abbildung 19.4: Drei Punkte und eine Gerade



Wir versuchen nun die Gerade so durch die Punktewolke zu legen, dass alle Punkte nah an der Geraden liegen. Es liegt nahe, die Gerade so zu wählen, dass die Summe der Abweichungen $|e_i|$

$$\sum_{i=1}^{n} |e_i|$$

minimal ist. Leider ist es nicht so einfach, diese Gerade zu bestimmen. Einfach ist es jedoch, die Gerade zu bestimmen, bei der die Summe der quadrierten

Abweichungen

$$\sum_{i=1}^{n} e_i^2 \tag{19.4}$$

minimal ist.
Man nennt dies die Gerade nach der **Methode der Kleinsten-Quadrate**.
Schauen wir uns an, wie man diese Gerade gewinnt.
Wir wollen β_0 und β_1 so bestimmen, dass

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

minimal ist.
Wir fassen diesen Ausdruck als Funktion von β_0 und β_1 auf:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{19.5}$$

Notwendig für einen Extremwert ist, dass die partiellen Ableitungen nach β_0 und β_1 von $S(\beta_0, \beta_1)$ gleich Null sind:

$$\frac{\partial}{\partial \beta_0} S(\beta_0, \beta_1) = 0$$

$$\frac{\partial}{\partial \beta_1} S(\beta_0, \beta_1) = 0$$

Die partiellen Ableitungen lauten:

$$\frac{\partial}{\partial \beta_0} S(\beta_0, \beta_1) = (-2) \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \tag{19.6}$$

$$\frac{\partial}{\partial \beta_1} S(\beta_0, \beta_1) = (-2) \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)\, x_i \tag{19.7}$$

Der Achsenabschnitt $\hat{\beta}_0$ und die Steigung $\hat{\beta}_1$ der gesuchten Geraden müssen also folgende Gleichungen erfüllen:

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{19.8}$$

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)\, x_i = 0 \tag{19.9}$$

Mit

$$e_i \;=\; y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

muss also gelten

$$\sum_{i=1}^{n} e_i \;=\; 0 \tag{19.10}$$

und

$$\sum_{i=1}^{n} e_i\, x_i \;=\; 0 \tag{19.11}$$

Man nennt diese Gleichungen auch Normalgleichungen. Die Gleichungen (19.8) und (19.9) sind linear in den Unbekannten $\hat{\beta}_0$ und $\hat{\beta}_1$. Wir bringen $\hat{\beta}_0$ und $\hat{\beta}_1$ in den Gleichungen (19.8) und (19.9) auf eine Seite

$$\sum_{i=1}^{n} y_i \;=\; n\,\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i \tag{19.12}$$

$$\sum_{i=1}^{n} y_i\, x_i \;=\; \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \tag{19.13}$$

Wir dividieren die Gleichungen (19.12) und (19.13) durch n und erhalten:

$$\overline{y} \;=\; \hat{\beta}_0 + \hat{\beta}_1\, \overline{x} \tag{19.14}$$

$$\overline{xy} \;=\; \hat{\beta}_0\, \overline{x} + \hat{\beta}_1\, \overline{x^2} \tag{19.15}$$

mit

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i\, y_i \qquad \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

Aus Gleichung (19.14) folgt

$$\boxed{\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\, \overline{x}} \tag{19.16}$$

Multiplizieren wir Gleichung (19.14) mit $\overline{x}$ und subtrahieren diese Gleichung von Gleichung (19.15), so erhalten wir:

$$\overline{xy} - \overline{x}\,\overline{y} = \hat{\beta}_1\, \overline{x^2} - \hat{\beta}_1\, \overline{x}^2 \tag{19.17}$$

Aus Gleichung (19.17) folgt:

$$\boxed{\hat{\beta}_1 = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2}} \tag{19.18}$$

Wir erhalten also folgende geschätzte Gerade, die wir auch **K-Q-Gerade** nennen.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{19.19}$$

An der Stelle $\bar{x}$ gilt wegen Gleichung (19.16)

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}\,.$$

Also liegt der Punkt $(\bar{x}, \bar{y})$ auf der Geraden.

Beispiel 152
Wir wählen von den 33 Autos aus Tabelle 19.1 auf Seite 514 sechs Autos aus, damit die Berechnung von $\hat{\beta}_0$ und $\hat{\beta}_1$ übersichtlicher ist. Die Daten sind in Tabelle 19.2 zu finden.

Tabelle 19.2: Alter x und Angebotspreis y von 6 VW-Golf 3

i	x_i	y_i	x_i^2	$x_i y_i$
1	2	21800	4	43600
2	3	21200	9	63600
3	4	17900	16	71600
4	5	16900	25	84500
5	6	13500	36	81000
6	7	12900	49	90300
	27	104200	139	434600

In der letzten Zeile von Tabelle 19.2 stehen die Spaltensummen. Hieraus folgt:

$$\bar{x} = 4.5 \qquad \bar{y} = 17366.667 \qquad \overline{x^2} = 23.1667 \qquad \overline{xy} = 72433.333$$

Also gilt

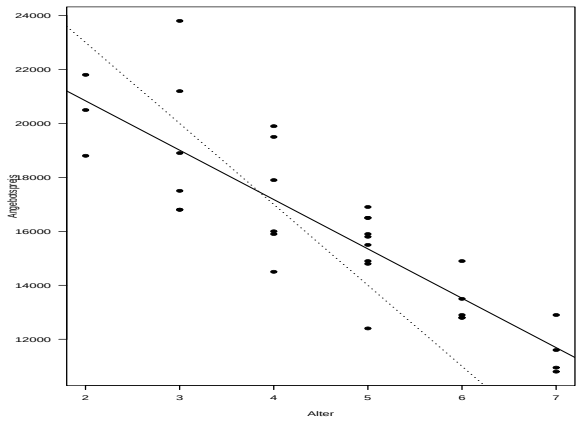
$$\hat{\beta}_1 = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{72433.333 - 4.5 \cdot 17366.667}{23.1667 - 4.5^2} = -1960$$

Mit jedem Jahr vermindert sich der Preis um 1960 DM.
Außerdem gilt

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 17366.667 - (-1960) \cdot 4.5 = 26186.67$$

In Abbildung 19.5 ist das Streudiagramm mit der geschätzten Geraden zu finden.

Abbildung 19.5: Streudiagramm des Alters und des Angebotspreises von 33 Autos vom Typ Golf



□

Beispiel 152 (fortgesetzt von Seite 517)
Es gilt

$$\bar{x} = 4.58 \qquad \bar{y} = 16125.76 \qquad \overline{x^2} = 23.12 \qquad \overline{xy} = 69792.42$$

Also gilt

$$\hat{\beta}_1 = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{69792.42 - 4.58 \cdot 16125.76}{23.12 - 4.58^2} = -1895.7$$

Mit jedem Jahr vermindert sich der Preis um 1895.7 DM.
Außerdem gilt

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 17366.667 - (-1960) \cdot 4.5 = 26186.67$$

□

Wir können den K-Q-Schätzer auch in der Form darstellen:

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}} \tag{19.20}$$

Wir haben in Statistik I gezeigt, dass gilt

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

Außerdem gilt

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y}) = \overline{x\,y} - \overline{y}\,\overline{x} \tag{19.21}$$

Dies sieht man folgendermaßen.

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

$$\frac{1}{n} \sum_{i=1}^{n} (x_i\,y_i - x_i\,\overline{y} - \overline{x}\,y_i + \overline{x}\,\overline{y}) =$$

$$\frac{1}{n} \sum_{i=1}^{n} x_i\,y_i - \overline{y}\,\frac{1}{n} \sum_{i=1}^{n} x_i - \overline{x}\,\frac{1}{n} \sum_{i=1}^{n} y_i + \frac{1}{n} \sum_{i=1}^{n} (\overline{x}\,\overline{y}) =$$

$$\overline{x\,y} - \overline{y}\,\overline{x} - \overline{x}\,\overline{y} + \overline{x}\,\overline{y} =$$

$$\overline{x\,y} - \overline{y}\,\overline{x}$$

Somit gilt die Beziehung in Gleichung (19.20).
Wir wollen noch zwei Spezialfälle betrachten.
Zunächst betrachten wir den Fall $\beta_1 = 0$. Die gesuchte Gerade verläuft also parallel zur Abszisse. Somit hat x keinen Einfluss auf y. Wir suchen den Wert von β_0, für den der folgende Ausdruck minimal ist

$$S(\beta_0) = \sum_{i=1}^{n} (y_i - \beta_0)^2$$

Die notwendige Bedingung für einen Extremwert ist, dass die erste Ableitung gleich Null ist.
Es gilt:

$$\frac{d}{d\,\beta_0} S(\beta_0) = (-2) \sum_{i=1}^{n} (y_i - \beta_0)$$

Für den Achsenabschnitt $\hat{\beta}_0$ der gesuchten Geraden muss also gelten

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0) = 0\,.$$

Dies ist äquivalent zu

$$\sum_{i=1}^{n} y_i = n\,\hat{\beta}_0$$

Hieraus folgt:

$$\boxed{\hat{\beta}_0 = \overline{y}}$$

Der Mittelwert minimiert also die Summe der quadrierten Abweichungen.
Wir betrachten nun noch den Fall $\beta_0 = 0$. Hier geht die Gerade durch den Ursprung.

Beispiel 153
Ein Student sucht am 28. August 1999 aus der Neuen Westfälischen alle Einzimmerwohnungen heraus, die explizit in Uninähe liegen, und schreibt die Fläche und die Höhe der Kaltmiete auf. Die Daten sind in Tabelle 19.3 zu finden.

Tabelle 19.3: Fläche und Höhe der Kaltmiete von Einzimmerwohnungen in Uninähe

Wohnung	Fläche	Miete
1	20	270
2	26	460
3	32	512
4	48	550
5	26	360
6	30	399
7	30	419
8	40	390

Da die Kaltmiete einer Wohnung, deren Fläche gleich 0 ist, sollte β_0 gleich 0 sein. □

Wir suchen den Wert $\hat{\beta}_1$ von β_1, für den der folgende Ausdruck minimal ist

$$S(\beta_1) = \sum_{i=1}^{n} (y_i - \beta_1 x_i)^2 .$$

Die notwendige Bedingung für einen Extremwert ist:

$$\frac{d}{d\beta_1} S(\beta_1) = 0 .$$

Es gilt

$$\frac{d}{d\beta_1} S(\beta_1) = (-2) \sum_{i=1}^{n} (y_i - \beta_1 x_i) x_i .$$

Für die Steigung $\hat{\beta}_1$ der gesuchten Geraden muss also gelten

$$\sum_{i=1}^{n} (x_i y_i - \hat{\beta}_1 x_i^2) = 0 .$$

Hieraus folgt

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \frac{\overline{x\,y}}{\overline{x^2}}} \tag{19.22}$$

Beispiel 153 (fortgesetzt von Seite 524)
Es gilt

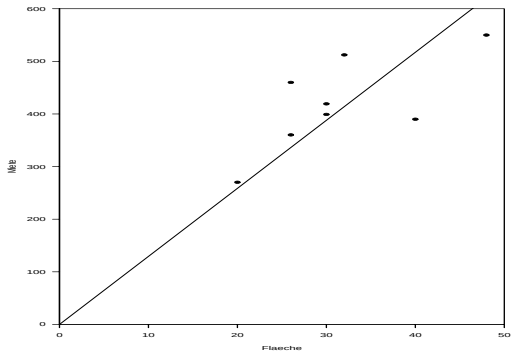
$$\overline{x^2} = 1060 \qquad \overline{xy} = 13705.5$$

Also gilt

$$\hat{\beta}_1 = \frac{\overline{x\,y}}{\overline{x^2}} = \frac{13705.5}{1060} = 12.93$$

In Abbildung 19.6 ist das Streudiagramm mit der geschätzten Geraden zu finden.

Abbildung 19.6: Fläche und Kaltmiete von Einzimmerwohnungen



□

19.3 Die Güte der Anpassung

Nachdem wir eine Gerade durch die Punktewolke gelegt haben, stellt sich die Frage, wie gut die Anpassung ist. Ein Maß für die Güte der Anpassung ist das **Bestimmtheitsmaß**. Dieses beruht auf der folgenden Beziehung, die wir am Ende dieses Kapitels beweisen werden:

$$\boxed{\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \overline{\hat{y}})^2 + \sum_{i=1}^{n} e_i^2} \tag{19.23}$$

Auf der linken Seite der Gleichung (19.23) steht die Streuung der Werte der abhängigen Variablen y. Diese Streuung zerlegen wir in zwei Summanden, die auf der rechten Seite der Gleichung (19.23) stehen.
Der erste Summand

$$\boxed{\sum_{i=1}^{n} (\hat{y}_i - \overline{\hat{y}})^2} \tag{19.24}$$

ist die Streuung der Punkte auf der Geraden. Wir sprechen auch von der **durch die Regression erklärten Streuung**.

Der zweite Summand

$$\boxed{\sum_{i=1}^{n} e_i^2} \tag{19.25}$$

ist die Streuung der Residuen, die wir auch **Reststreuung** nennen. Dass wir die Größe in Gleichung (19.25) als Streuung der Residuen interpretieren können, ist durch Gleichung (19.10) auf Seite 520 gerechtfertigt, die wir hier noch einmal wiedergeben:

$$\sum_{i=1}^{n} e_i = 0\,.$$

Dividieren wir beide Seiten dieser Gleichung durch n, so gilt

$$\bar{e} = 0\,.$$

Gleichung (19.10) erlaubt es auch Gleichung (19.24) zu modifizieren. Mit $e_i = y_i - \hat{y}_i$ gilt

$$\sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$$

und somit

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i\,.$$

Also gilt $\bar{y} = \overline{\hat{y}}$. Somit ist Gleichung (19.24) äquivalent zu

$$\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 \tag{19.26}$$

Wir können Gleichung (19.23) auf Seite 526 also auch folgendermaßen schreiben:

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} e_i^2 \tag{19.27}$$

Als Maß für die Güte der Anpassung betrachtet man nun das Verhältnis aus der der durch die Regression erklärten Streuung zur Streuung der y_i. Dieses

heißt Bestimmtheitsmaß R^2:

$$\boxed{R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}} \tag{19.28}$$

Als Quotient zweier nichtnegativer Größen ist Bestimmtheitsmaß größer oder gleich 0. Es kann aber auch nicht größer als 1 werden, da die Größe im Zähler auf Grund von Gleichung (19.27) ein Teil der Größe im Nenner ist.
Es gilt also

$$\boxed{0 \leq R^2 \leq 1}$$

Dividieren wir beide Seiten von Gleichung (19.27) durch $\sum_{i=1}^{n} (y_i - \overline{y})^2$, so gilt

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

Also gilt

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

Liegen alle Punkte auf einer Geraden, so sind alle Residuen gleich 0. Also ist das Bestimmtheitsmaß R^2 in diesem Fall gleich 1.
Es gilt

$$\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 = \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \overline{x})^2 \tag{19.29}$$

Dies sieht man folgendermaßen:

$$\begin{aligned}\sum_{i=1}^{n}\left(\hat{y}_i-\overline{y}\right)^2 &= \sum_{i=1}^{n}\left(\hat{\beta}_0+\hat{\beta}_1\,x_i-\hat{\beta}_0-\hat{\beta}_1\,\overline{x}\right)^2=\sum_{i=1}^{n}\left(\hat{\beta}_1\left(x_i-\overline{x}\right)\right)^2\\ &= \hat{\beta}_1^2\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2\end{aligned}$$

Ersetzen wir in Gleichung (19.28) den Zähler durch die rechte Seite von Gleichung (19.29), so erhalten wir

$$\boxed{R^2=\frac{\hat{\beta}_1^2\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}{\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2}=\hat{\beta}_1^2\,\frac{\overline{x^2}-\bar{x}^2}{\overline{y^2}-\bar{y}^2}} \tag{19.30}$$

Beispiel 153 (fortgesetzt von Seite 521)
Es gilt $\hat{\beta}_1=-1960$. Außerdem haben wir bereits $\bar{x}=4.5$, $\bar{y}=17366.67$ und $\overline{x^2}=23.1667$ bestimmt. Wir benötigen noch $\sum_{i=1}^{n}(y_i-\bar{y})^2$. Es gilt $\sum_{i=1}^{n}(y_i-\bar{y})^2=69753333$. Also gilt

$$R^2=\frac{(-1960)^2\cdot 6\cdot(23.1667-4.5^2)}{69753333}=0.964$$

□

Beispiel 153 (fortgesetzt von Seite 522)
Hier gilt $R^2=0.7295$. □

Setzen wir in Gleichung (19.30) auf Seite 529 für $\hat{\beta}_1$ den Ausdruck in Gleichung (19.20) auf Seite 523 ein, so gilt

$$R^2 = \frac{\left(\sum_{i=1}^{n}\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right)\right)^2}{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2,\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2}$$

Dies ist aber gerade das Quadrat des Korrelationskoeffizienten von Bravais-Pearson zwischen X und Y.

19.3.1 Beweis von Gleichung (19.23) auf Seite 526

Wir betrachten hierzu zunächst die Residuen, die folgendermaßen definiert sind:

$$e_i = y_i - \hat{y}_i \tag{19.31}$$

Aus Gleichung (19.31) folgt

$$y_i = \hat{y}_i + e_i. \tag{19.32}$$

Für die Mittelwerte der drei Größen in Gleichung (19.32) gilt

1. Der Mittelwert der y_i ist $\bar{y}$.
2. Der Mittelwert der e_i ist 0.
3. Der Mittelwert der $\hat{y}_i$ ist $\bar{y}$.

Nun wenden wir uns Gleichung (19.32) zu. Subtrahieren wir $\overline{y}$ von beiden Seiten von Gleichung (19.32) und berücksichtigen $\overline{\hat{y}} = \overline{y}$, so gilt

$$y_i - \overline{y} = \hat{y}_i - \overline{\hat{y}} + e_i. \tag{19.33}$$

Wir quadrieren nun beide Seiten von Gleichung (19.33) und erhalten

$$(y_i - \overline{y})^2 = (\hat{y}_i - \overline{\hat{y}})^2 + e_i^2 + 2\,(\hat{y}_i - \overline{\hat{y}})e_i. \tag{19.34}$$

Wir summieren nun beide Seiten von Gleichung (19.34) von $i = 1$ bis n und erhalten

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \overline{\hat{y}})^2 + \sum_{i=1}^{n} e_i^2 + 2 \sum_{i=1}^{n} (\hat{y}_i - \overline{\hat{y}})e_i. \tag{19.35}$$

Wenn wir nun noch zeigen, dass der letzte Summand in Gleichung (19.35) gleich 0 ist, so haben wir Gleichung (19.23) bewiesen.
Hierzu schauen wir uns noch einmal die Gleichungen (19.10) und (19.11) an:

$$\sum_{i=1}^{n} e_i \;=\; 0 \tag{19.36}$$

$$\sum_{i=1}^{n} e_i\, x_i \;=\; 0 \tag{19.37}$$

Multiplizieren wir die Gleichung (19.36) mit $\hat{\beta}_0$ und die Gleichung (19.36) mit $\hat{\beta}_1$, so erhalten wir

$$\sum_{i=1}^{n} \hat{\beta}_0 \, e_i \;\; = \;\; 0 \tag{19.38}$$

$$\sum_{i=1}^{n} \hat{\beta}_1 \, x_i \, e_i \;\; = \;\; 0 \tag{19.39}$$

Addieren wir die Gleichungen (19.39) und (19.38), so gilt

$$\sum_{i=1}^{n} \hat{\beta}_0 \, e_i + \sum_{i=1}^{n} \hat{\beta}_1 \, x_i \, e_i \;\; = \;\; 0$$

Somit gilt

$$\sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 \, x_i) \, e_i \;\; = \;\; 0$$

Wegen

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \, x_i$$

gilt

$$\sum_{i=1}^{n} \hat{y}_i \, e_i \;\; = \;\; 0 \tag{19.40}$$

Nun schauen wir uns noch einmal den letzten Summanden in Gleichung (19.35) an:

$$2 \sum_{i=1}^{n} (\hat{y}_i - \overline{\hat{y}}) e_i = 2 \sum_{i=1}^{n} \hat{y}_i \, e_i - 2 \sum_{i=1}^{n} \overline{\hat{y}} e_i \overset{(19.40)}{=} -2 \, \overline{\hat{y}} \sum_{i=1}^{n} e_i \overset{(19.10)}{=} 0$$

Somit ist die Gleichung (19.23) bewiesen.

19.4 Tests und Konfidenzintervalle

Wir haben auf Seite 19.1 folgendes Modell betrachtet:

$$Y = \beta_0 + \beta_1 \cdot x + \epsilon$$

Dabei ist ϵ eine Zufallsvariable. Um in diesem Modell die Parameter β_0 und β_1 zu schätzen, beobachten wir $(x_1, y_1), \ldots, (x_n, y_n)$. Jedes y_i ist eine Realisation einer Zufallsvariablen Y_i. Wir können das Modell 19.1 auch in Abhängigkeit von diesen Zufallsvariablen schreiben. Für $i = 1, \ldots, n$ gilt

$$\boxed{Y_i = \beta_0 + \beta_1 x_i + \epsilon_i} \,. \tag{19.41}$$

Wir unterstellen

$$\boxed{E(\epsilon_i) = 0}$$

und

$$\boxed{Var(\epsilon_i) = \sigma^2}$$

für $i = 1, \ldots, n$. Außerdem nehmen wir noch an, dass die ϵ_i unkorreliert sind. Für $i \neq j$ gilt also

$$\boxed{Cov(\epsilon_i, \epsilon_j) = 0}$$

Da die KQ-Schätzer $\hat{\beta}_0$ und $\hat{\beta}_1$ von den y_i abhängen, sind sie auch Realisationen von Zufallsvariablen, mit denen wir Tests durchführen können und Konfidenzintervalle aufstellen können. Hierzu unterstellen wir, dass die ϵ_i normalverteilt sind. In diesem Fall ist $\hat{\beta}_i$ normalverteilt mit Erwartungwert β_i. Für die Varianzen gilt

$$\boxed{Var(\hat{\beta}_0) = \frac{\sigma^2 \, \overline{x^2}}{n(\overline{x^2} - \bar{x}^2)}}$$

und

$$\boxed{Var(\hat{\beta}_1) = \frac{\sigma^2}{n(\overline{x^2} - \bar{x}^2)}}$$

Die unbekannte Varianz σ^2 schätzen wir durch

$$\boxed{\widehat{\sigma^2} = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2}$$

Beispiel 153 (fortgesetzt von Seite 529)
Es gilt $\hat{\sigma}^2 = 631333.33333$. □

Wir erhalten die geschätzten Varianzen

$$\boxed{\widehat{Var}(\hat{\beta}_0) = \frac{\widehat{\sigma^2}\,\overline{x^2}}{n(\overline{x^2} - \bar{x}^2)}}$$

und

$$\boxed{\widehat{Var}(\hat{\beta}_1) = \frac{\widehat{\sigma^2}}{n(\overline{x^2} - \bar{x}^2)}}$$

Beispiel 153 (fortgesetzt von Seite 533)
Es gilt $\bar{x} = 4.5$ und $\overline{x^2} = 23.16667$. Also gilt

$$\widehat{Var}(\hat{\beta}_0) = \frac{631333.3 \cdot 23.167}{6(23.167 - 20.25)} = 835764.2$$

und

$$\widehat{Var}(\hat{\beta}_1) = \frac{631333.3}{6(23.167 - 20.25)} = 36076.15$$

□

Man kann zeigen, dass

$$\frac{\hat{\beta}_i}{\sqrt{\widehat{Var}(\hat{\beta}_i)}}$$

mit $n-2$ Freiheitsgraden t-verteilt ist. Mit diesem Wissen kann man problemlos Tests durchführen und Konfidenzintervalle aufstellen. Beginnen wir mit den Tests.
Will man überprüfen, ob x einen Einfluss auf y hat, so testet man

$$H_0 : \beta_1 = 0 \quad \text{gegen} \quad H_1 : \beta_1 \neq 0$$

Die Teststatistik ist

$$\boxed{t = \frac{\hat{\beta}_1}{\sqrt{\widehat{Var}(\hat{\beta}_1)}}} \tag{19.42}$$

Wenn H_0 zutrifft, ist t t-verteilt mit $n-2$ Freiheitsgraden.

Beispiel 153 (fortgesetzt von Seite 533)
Es gilt

$$t = \frac{-1960}{\sqrt{36076.15}} = -10.32$$

Wegen $t_{0.975;4} = 2.7764$ lehnen wir H_0 zum Niveau 0.05 ab. □

Wir können natürlich auch testen, ob β_1 einen anderen Wert β_1^0 annimmt, so testen wir

$$H_0 : \beta_1 = \beta_1^0 \quad \text{gegen} \quad H_1 : \beta_1 \neq \beta_1^0$$

Diese Hypothese testen wir mit

$$\boxed{t = \frac{\hat{\beta}_1 - \beta_1^0}{\sqrt{\frac{\widehat{\sigma^2}}{\sum_{t=1}^{T} (x_t - \overline{x})^2}}}} \tag{19.43}$$

Wollen wir testen, ob die Regression durch den Ursprung geht, so lauten die Hypothese

$$H_0 : \beta_0 = 0 \quad \text{gegen} \quad H_1 : \beta_0 \neq 0$$

Wir testen diese Hypotehsen mit

$$\boxed{t = \frac{\hat{\beta}_0}{\sqrt{\frac{\widehat{\sigma^2}\,\overline{x^2}}{\sum_{t=1}^{T} (x_t - \overline{x})^2}}}} \tag{19.44}$$

Wenn H_0 zutrifft, ist t t-verteilt mit $n-2$ Freiheitsgraden.

Beispiel 153 (fortgesetzt von Seite 534)
Es gilt

$$t = \frac{26186.7}{\sqrt{835764.2}} = 28.64$$

Wegen $t_{0.975;4} = 2.7764$ lehnen wir H_0 zum Niveau 0.05 ab. □

Das Konfidenzintvervall für β_i zum Konfidenzniveau $1-\alpha$ ist

$$\boxed{\left[\hat{\beta}_i - t_{n-2;1-\alpha/2} \sqrt{\widehat{Var}(\hat{\beta}_i)}, \hat{\beta}_i + t_{n-2;1-\alpha/2} \sqrt{\widehat{Var}(\hat{\beta}_i)}\right]}$$

Beispiel 153 (fortgesetzt von Seite 534)
Das Konfidenzintvervall für β_0 zum Konfidenzniveau 0.95 ist

$$[23648.51, 28724.89]$$

Das Konfidenzintvervall für β_1 zum Konfidenzniveau 0.95 ist

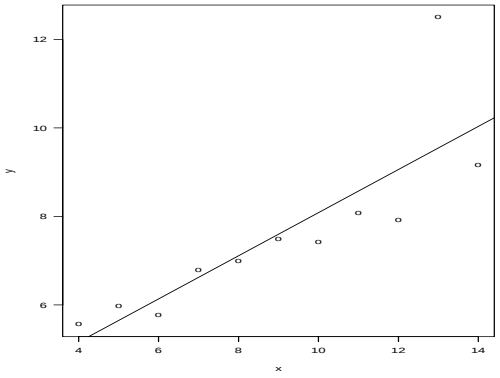
$$[-2487.342, -1432.658]$$

□

19.5 Ausreißer und einflussreiche Beobachtungen

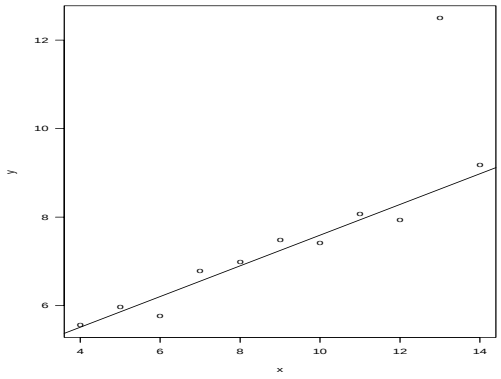
Einzelne Beobachtungen können einen starken Einfluss auf die K-Q-Schätzer haben. Abbildung 19.7 zeigt einen Datensatz, bei dem ein Ausreißer vorliegt.

Abbildung 19.7: Streudiagramm mit Ausreißer



Außerdem ist die K-Q-Gerade eingezeichnet. Abbildung 19.8 zeigt die K-Q-Gerade, die man erhält, wenn man den Ausreißer aus dem Datensatz herausläßt.

Abbildung 19.8: Streudiagramm mit Ausreißer und Regressionsgerade, die ohne den Ausreißer geschätzt wurde

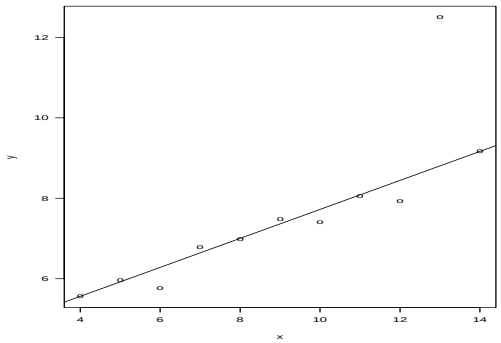


Wir sehen, dass der Ausreißer einen starken Einfluss auf die gewählte Gerade hat. Man spricht von einem Ausreißer im Beobachtungsraum, da er entfernt von der Geraden liegt.
Liegen Ausreißer vor, so sollte man ein Schätzverfahren wählen, das nicht so ausreißerempfindlich ist. Die Minimierung von

$$\sum_{t=1}^{T} |e_t|$$

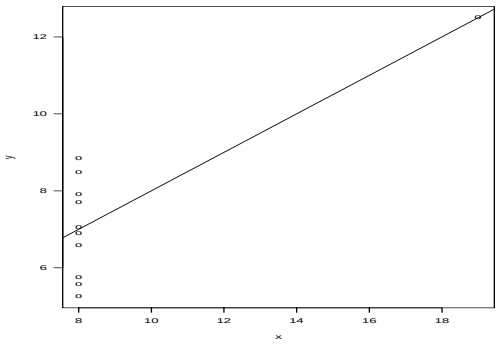
ist robust. Abbildung 19.9 zeigt die mit Hilfe dieses Schätzverfahrens gewonnene Gerade.

Abbildung 19.9: Streudiagramm mit Ausreißer und robuster Regressionsgerade



Die folgende Graphik zeigt einen anderen Typ von Ausreißer.

Abbildung 19.10: Streudiagramm mit Ausreißer



Der Punkt rechts oben im Diagramm steuert die Steigung der Geraden, da die Gerade durch den Punkt geht. Man spricht auch von einem Punkt mit Hebelwirkung. Er hat einen sehr starken Einfluss auf die Schätzung.

19.6 Linearisierbare Zusammenhänge

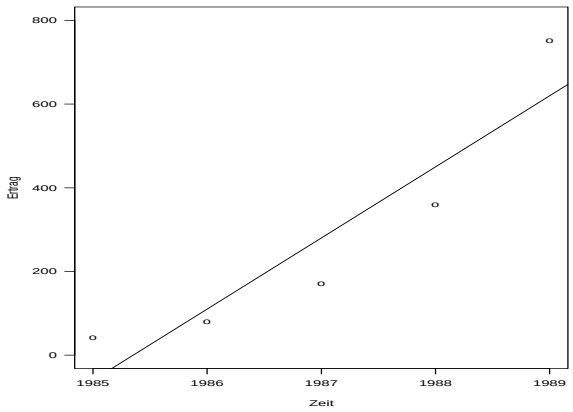
Bisher sind wir davon ausgegangen, dass der Zusammenhang zwischen den Variablen x und y gut durch eine Gerade beschrieben werden kann. Dies ist bei dem nächsten Beispiel nicht der Fall. In Tabelle 19.4 ist der jährliche Ertrag y_t eines Agrarproduktes für die Jahre $x_t = 1985, \ldots, 1989$ zu finden.

Tabelle 19.4: Ertrag eines Agrarprodukts

t	x_t	y_t
1	1985	40
2	1986	80
3	1987	170
4	1988	360
5	1989	750

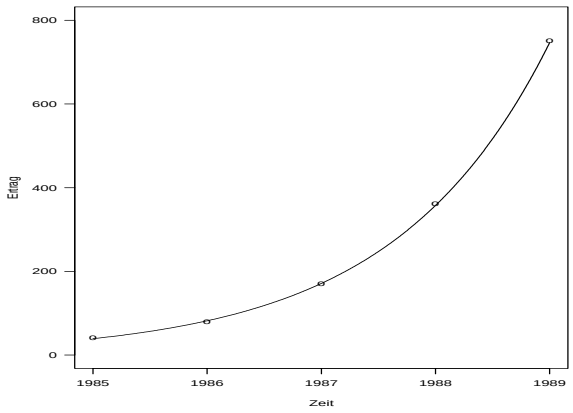
Abbildung 19.11 zeigt die Punktewolke mit der K-Q-Geraden.

Abbildung 19.11: Streudiagramm mit KQ-Gerade



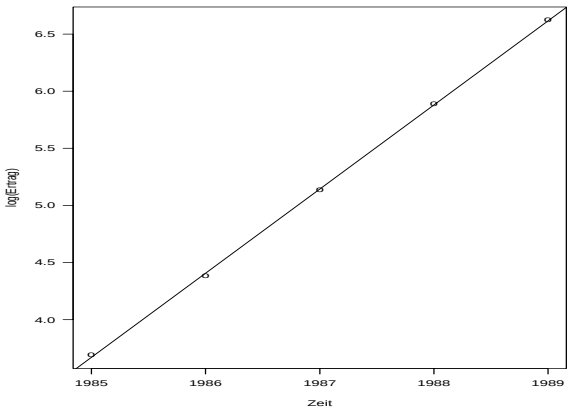
Die Anpassung ist schlecht. Der Zusammenhang wird viel besser durch die in Abbildung 19.12 eingezeichnete Kurve beschrieben.

Abbildung 19.12: Streudiagramm mit nichtlinearer Funktion



Die Anpassung ist hier nahezu perfekt. Wie wurde die Kurve gefunden? Die Gerade wurde nicht durch die Punkte (x, y), sondern durch $(x, \ln y)$ gelegt, wobei $\ln y$ der natürliche Logarithmus von y ist.
Abbildung 19.13 zeigt die zugehörige Punktewolke mit der K-Q-Geraden:

Abbildung 19.13: Streudiagramm



Wir betrachten also nicht y, sondern eine Transformation $g(y)$. Schauen wir uns hierzu ein Beispiel an.
Der Ansatz

$$\begin{aligned} y &= f(x) \\ &= a + b\,x \end{aligned}$$

beschreibt ein lineares Wachstum.
In jedem Zeitintervall der Länge Δx erhöht sich y um den festen Wert $b\,\Delta\,x$, denn

$$\begin{aligned} \Delta\,y &= f(x + \Delta\,x) - f(x) \\ &= a + b\,(x + \Delta\,x) - a - b\,x \\ &= b\,\Delta\,x. \end{aligned}$$

Es gilt also:

$$\frac{\Delta\,y}{\Delta\,x} = b.$$

Lassen wir $\Delta\,x$ gegen 0 streben, so erhalten wir

$$\frac{d\,y}{d\,x} = b$$

Die erste Ableitung der Geraden ist also konstant.
Beim linearen Wachstum ist die Menge, die in einem Zeitintervall hinzukommt, proportional zur Länge des Zeitintervalls. Die Menge, um die die

interessierende Variable in einem Zeitintervall der Länge 2 whächst, ist doppelt so groß wie in einem Zeitintervall der Länge 1. Sehr oft hängt die Menge, um die eine Größe in einem Zeitintervall Δx wächst, nicht nur von Δx ab, sondern auch vom Bestand zu Beginn des Zeitintervalls. Je mehr vorhanden ist, um so mehr kommt auch hinzu. Dies ist zum Beispiel beim Wachstum einer Population in der Regel der Fall. Je mehr Menschen da sind, umso mehr kommen auch hinzu. Die einfachste Annahme ist hier, dass immer ein fester Anteil b hinzukommt. Es muss also

$$\begin{aligned} \Delta y &= f(x + \Delta x) - f(x) \\ &= b\, y\, \Delta x \end{aligned}$$

mit $y = f(x)$ gelten.
Dividieren wir beide Seiten dieser Gleichung durch Δx, so ergibt sich

$$\frac{\Delta y}{\Delta x} = b\, y$$

Der Übergang zum Differentialquotienten liefert:

$$\frac{d\, y}{d\, x} = b\, y$$

Diese Gleichung wird von

$$y = e^{a+b\,x} \tag{19.45}$$

erfüllt, denn

$$\begin{aligned} \frac{d\, y}{d\, x} &= b\, e^{a+b\,x} \\ &= b\, f(x). \end{aligned}$$

Man spricht in diesem Fall von exponentiellem Wachstum.
Bilden wir auf beiden Seiten von Gleichung (37) den natürlichen Logarithmus, so erhalten wir

$$\ln y = a + b\, x \tag{19.46}$$

Die logarithmierten Werte von y hängen also linear von x ab.
Alle Beziehungen der Form

$$g(y) = a + b\, h(x)$$

können wir mit den Verfahren aus Kapitel 3 anpassen. Hierbei sind $g(y)$ und $h(x)$ geeignete Funktionen.

In der Ökonometrie wird sehr oft der natürliche Logarithmus verwendet.

Wir haben bereits y logarithmiert. Da nur y logarithmiert wurde, spricht man von einem semilogarithmischen Zusammenhang. Man nennt eine Beziehung zwischen zwei Variablen x und y semilogarithmisch, wenn nur eine der beiden Variablen logarithmiert wurde.

Wir können natürlich auch nur x logarithmieren.

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

Abbildung 19.14 zeigt eine Punktewolke, bei der dies angebracht ist, zusammen mit der angepaßten Funktion.

Abbildung 19.14: Streudiagramm

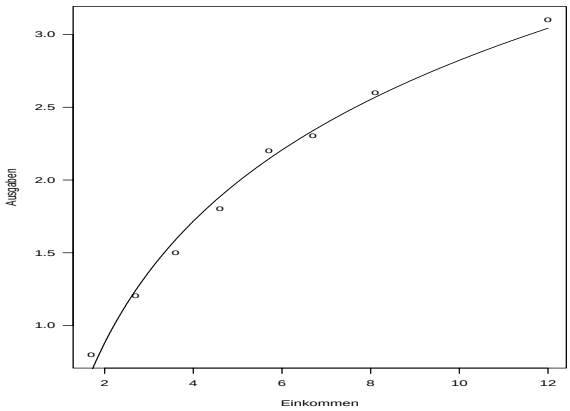
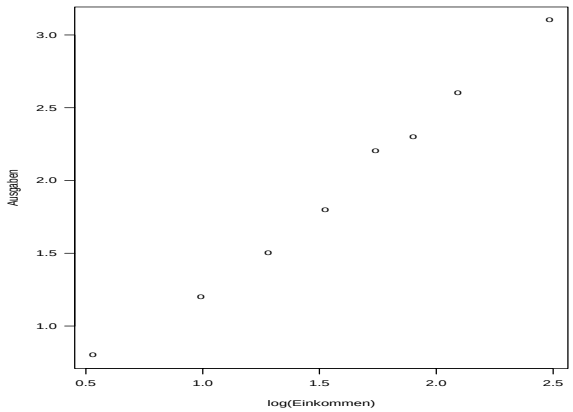


Abbildung 19.15 zeigt, dass der Zusammenhang zwischen $\ln x$ und y linear ist.

Abbildung 19.15: Streudiagramm



Es können natürlich auch beide Seiten logarithmiert werden:

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

Wir delogarithmieren beide Seite und erhalten

$$\begin{aligned} y &= e^{a+b \ln x} \\ &= e^a e^{\ln x^b} \\ &= c x^b \end{aligned}$$

mit $c = e^a$.
Die Ableitung lautet

$$\frac{d y}{d x} = c b x^{b-1}$$

Für die Elastizität

$$\frac{\frac{d y}{y}}{\frac{d x}{x}}$$

der Funktion gilt

$$\begin{aligned} \frac{d y}{d x} \frac{x}{y} &= c b x^{b-1} \frac{x}{c x^b} \\ &= b \end{aligned}$$

Die Elastizität der Funktion ist also konstant. Eine einprozentige änderung in x führt also zu einer b prozentigen Änderung in y.

19.7 Regressionsanalyse in R

Wir schauen uns zunächst das Beispiel 152 auf Seite 521 an. Die Daten sind in Tabelle 19.2 auf Seite 521 zu finden.
Wir erzeugen eine Variable `Alter` und eine Variable `Preis`:

```
> Alter<-c(2,3,4,5,6,7)
> Preis<-c(21800,21200,17900,16900,13500,12900)
```

In `R` kann man eine lineare Regression mit der Funktion `lsfit` durchführen. Wir rufen diese folgendermaßen auf:

```
lsfit(x,y,intercept=TRUE)
```

Dabei ist `x` die erklände Variable und `y` die abhängige Variable. Ist das Argument `intercept` gleich `TRUE`, so schätzen wir

$$y = \beta_0 + \beta_1 x$$

Das Ergebnis von `lsfit` ist eine Liste, deren erste Komponente ein Vektor ist, der $\hat{\beta}_0$ und $\hat{\beta}_1$ enthält.

```
> lsfit(Alter,Preis)[[1]]
Intercept         X
 26186.67  -1960.00
```

Die Funktion `lsfit` liefert nicht den Wert von R^2. Hierzu müssen wir eine andere Funktion verwenden. Man kann aber mit der Funktion `lsfit` folgendermaßen problemlos die Regressionsgerade zum Streudiagramm hinzufügen.

```
> plot(Alter,Preis)
> abline(lsfit(Alter,Preis))
```

Schauen wir uns die Funktion `lm` an. Diese wird folgendermaßen aufgerufen:

```
lm(y~x)
```

Informationen erhalten wir, indem wir die Funktion `summary` auf das Ergebnis von `lm` anwenden.

```
> summary(lm(Preis~Alter))

 Call: lm(formula = Preis ~ Alter)

 Residuals:
     1      2      3      4      5      6
-466.7  893.3 -446.7  513.3 -926.7  433.3

 Coefficients:
             Estimate Std. Error t value Pr(>|t|)
 (Intercept)  26186.7      914.2   28.64 8.84e-06 ***
 Alter        -1960.0      189.9  -10.32 0.000498 ***
 ---
 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Residual standard error: 794.6 on 4 degrees of freedom
 Multiple R-Squared: 0.9638,
 Adjusted R-squared: 0.9547
 F-statistic: 106.5 on 1 and 4 DF,
 p-value: 0.0004976
```

Hier finden wir die Schätzer von $\hat{\beta}_0$ und $\hat{\beta}_1$ unter `Coefficients` und $\sqrt{\widehat{Var}(\hat{\beta}_0)}$ und $\sqrt{\widehat{Var}(\hat{\beta}_1)}$ unter `Std. Error`. Der Wert detr Teststatistik t steht unter `t value`. Den Wert von R^2 finden wir unter `Multiple R-Squared`.

Kapitel 20

Tabellen

Tabelle 20.1: Verteilungsfunktion $\Phi(z)$ der Standardnormalverteilung

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.500	0.504	0.508	0.512	0.516	0.520	0.524	0.528	0.532	0.536
0.1	0.540	0.544	0.548	0.552	0.556	0.560	0.564	0.567	0.571	0.575
0.2	0.579	0.583	0.587	0.591	0.595	0.599	0.603	0.606	0.610	0.614
0.3	0.618	0.622	0.626	0.629	0.633	0.637	0.641	0.644	0.648	0.652
0.4	0.655	0.659	0.663	0.666	0.670	0.674	0.677	0.681	0.684	0.688
0.5	0.691	0.695	0.698	0.702	0.705	0.709	0.712	0.716	0.719	0.722
0.6	0.726	0.729	0.732	0.736	0.739	0.742	0.745	0.749	0.752	0.755
0.7	0.758	0.761	0.764	0.767	0.770	0.773	0.776	0.779	0.782	0.785
0.8	0.788	0.791	0.794	0.797	0.800	0.802	0.805	0.808	0.811	0.813
0.9	0.816	0.819	0.821	0.824	0.826	0.829	0.831	0.834	0.836	0.839
1.0	0.841	0.844	0.846	0.848	0.851	0.853	0.855	0.858	0.860	0.862
1.1	0.864	0.867	0.869	0.871	0.873	0.875	0.877	0.879	0.881	0.883
1.2	0.885	0.887	0.889	0.891	0.893	0.894	0.896	0.898	0.900	0.901
1.3	0.903	0.905	0.907	0.908	0.910	0.911	0.913	0.915	0.916	0.918
1.4	0.919	0.921	0.922	0.924	0.925	0.926	0.928	0.929	0.931	0.932
1.5	0.933	0.934	0.936	0.937	0.938	0.939	0.941	0.942	0.943	0.944
1.6	0.945	0.946	0.947	0.948	0.949	0.951	0.952	0.953	0.954	0.954
1.7	0.955	0.956	0.957	0.958	0.959	0.960	0.961	0.962	0.962	0.963
1.8	0.964	0.965	0.966	0.966	0.967	0.968	0.969	0.969	0.970	0.971
1.9	0.971	0.972	0.973	0.973	0.974	0.974	0.975	0.976	0.976	0.977
2.0	0.977	0.978	0.978	0.979	0.979	0.980	0.980	0.981	0.981	0.982
2.1	0.982	0.983	0.983	0.983	0.984	0.984	0.985	0.985	0.985	0.986
2.2	0.986	0.986	0.987	0.987	0.987	0.988	0.988	0.988	0.989	0.989
2.3	0.989	0.990	0.990	0.990	0.990	0.991	0.991	0.991	0.991	0.992
2.4	0.992	0.992	0.992	0.992	0.993	0.993	0.993	0.993	0.993	0.994
2.5	0.994	0.994	0.994	0.994	0.994	0.995	0.995	0.995	0.995	0.995
2.6	0.995	0.995	0.996	0.996	0.996	0.996	0.996	0.996	0.996	0.996
2.7	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997
2.8	0.997	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998
2.9	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.999	0.999	0.999
3.0	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999

Tabelle 20.5: Quantile der t-Verteilung

ν	$t_{\nu;0.9}$	$t_{\nu;0.95}$	$t_{\nu;0.975}$	$t_{\nu;0.99}$	$t_{\nu;0.995}$
1	3.0777	6.3138	12.7062	31.8205	63.6567
2	1.8856	2.9200	4.3027	6.9646	9.9248
3	1.6377	2.3534	3.1824	4.5407	5.8409
4	1.5332	2.1318	2.7764	3.7469	4.6041
5	1.4759	2.0150	2.5706	3.3649	4.0321
6	1.4398	1.9432	2.4469	3.1427	3.7074
7	1.4149	1.8946	2.3646	2.9980	3.4995
8	1.3968	1.8595	2.3060	2.8965	3.3554
9	1.3830	1.8331	2.2622	2.8214	3.2498
10	1.3722	1.8125	2.2281	2.7638	3.1693
11	1.3634	1.7959	2.2010	2.7181	3.1058
12	1.3562	1.7823	2.1788	2.6810	3.0545
13	1.3502	1.7709	2.1604	2.6503	3.0123
14	1.3450	1.7613	2.1448	2.6245	2.9768
15	1.3406	1.7531	2.1314	2.6025	2.9467
16	1.3368	1.7459	2.1199	2.5835	2.9208
17	1.3334	1.7396	2.1098	2.5669	2.8982
18	1.3304	1.7341	2.1009	2.5524	2.8784
19	1.3277	1.7291	2.0930	2.5395	2.8609
20	1.3253	1.7247	2.0860	2.5280	2.8453
21	1.3232	1.7207	2.0796	2.5176	2.8314
22	1.3212	1.7171	2.0739	2.5083	2.8188
23	1.3195	1.7139	2.0687	2.4999	2.8073
24	1.3178	1.7109	2.0639	2.4922	2.7969
25	1.3163	1.7081	2.0595	2.4851	2.7874

Tabelle 20.2: Quantil z_p der Standardnormalverteilung

p	.000	.001	.002	.003	.004	.005	.006	.007	.008	.009
0.50	0.000	0.002	0.005	0.008	0.010	0.012	0.015	0.018	0.020	0.023
0.51	0.025	0.028	0.030	0.033	0.035	0.038	0.040	0.043	0.045	0.048
0.52	0.050	0.053	0.055	0.058	0.060	0.063	0.065	0.068	0.070	0.073
0.53	0.075	0.078	0.080	0.083	0.085	0.088	0.090	0.093	0.095	0.098
0.54	0.100	0.103	0.106	0.108	0.110	0.113	0.116	0.118	0.121	0.123
0.55	0.126	0.128	0.131	0.133	0.136	0.138	0.141	0.143	0.146	0.148
0.56	0.151	0.154	0.156	0.159	0.161	0.164	0.166	0.169	0.171	0.174
0.57	0.176	0.179	0.182	0.184	0.187	0.189	0.192	0.194	0.197	0.199
0.58	0.202	0.204	0.207	0.210	0.212	0.215	0.217	0.220	0.222	0.225
0.59	0.228	0.230	0.233	0.235	0.238	0.240	0.243	0.246	0.248	0.251
0.60	0.253	0.256	0.258	0.261	0.264	0.266	0.269	0.272	0.274	0.277
0.61	0.279	0.282	0.284	0.287	0.290	0.292	0.295	0.298	0.300	0.303
0.62	0.306	0.308	0.311	0.313	0.316	0.319	0.321	0.324	0.327	0.329
0.63	0.332	0.334	0.337	0.340	0.342	0.345	0.348	0.350	0.353	0.356
0.64	0.358	0.361	0.364	0.366	0.369	0.372	0.374	0.377	0.380	0.383
0.65	0.385	0.388	0.391	0.393	0.396	0.399	0.402	0.404	0.407	0.410
0.66	0.412	0.415	0.418	0.421	0.423	0.426	0.429	0.432	0.434	0.437
0.67	0.440	0.443	0.445	0.448	0.451	0.454	0.456	0.459	0.462	0.465
0.68	0.468	0.470	0.473	0.476	0.479	0.482	0.484	0.487	0.490	0.493
0.69	0.496	0.499	0.501	0.504	0.507	0.510	0.513	0.516	0.519	0.522
0.70	0.524	0.527	0.530	0.533	0.536	0.539	0.542	0.545	0.548	0.550
0.71	0.553	0.556	0.559	0.562	0.565	0.568	0.571	0.574	0.577	0.580
0.72 ´	0.583	0.586	0.589	0.592	0.595	0.598	0.601	0.604	0.607	0.610
0.73	0.613	0.616	0.619	0.622	0.625	0.628	0.631	0.634	0.637	0.640
0.74	0.643	0.646	0.650	0.653	0.656	0.659	0.662	0.665	0.668	0.671

Tabelle 20.3: Quantil z_p der Standardnormalverteilung

p	.000	.001	.002	.003	.004	.005	.006	.007	.008	.009
0.75	0.674	0.678	0.681	0.684	0.687	0.690	0.694	0.697	0.700	0.703
0.76	0.706	0.710	0.713	0.716	0.719	0.722	0.726	0.729	0.732	0.736
0.77	0.739	0.742	0.745	0.749	0.752	0.755	0.759	0.762	0.766	0.769
0.78	0.772	0.776	0.779	0.782	0.786	0.789	0.793	0.796	0.800	0.803
0.79	0.806	0.810	0.813	0.817	0.820	0.824	0.827	0.831	0.834	0.838
0.80	0.842	0.845	0.849	0.852	0.856	0.860	0.863	0.867	0.870	0.874
0.81	0.878	0.882	0.885	0.889	0.893	0.896	0.900	0.904	0.908	0.912
0.82	0.915	0.919	0.923	0.927	0.931	0.935	0.938	0.942	0.946	0.950
0.83	0.954	0.958	0.962	0.966	0.970	0.974	0.978	0.982	0.986	0.990
0.84	0.994	0.999	1.003	1.007	1.011	1.015	1.019	1.024	1.028	1.032
0.85	1.036	1.041	1.045	1.049	1.054	1.058	1.062	1.067	1.071	1.076
0.86	1.080	1.085	1.089	1.094	1.098	1.103	1.108	1.112	1.117	1.122
0.87	1.126	1.131	1.136	1.141	1.146	1.150	1.155	1.160	1.165	1.170
0.88	1.175	1.180	1.185	1.190	1.195	1.200	1.206	1.211	1.216	1.221
0.89	1.226	1.232	1.237	1.243	1.248	1.254	1.259	1.265	1.270	1.276
0.90	1.282	1.287	1.293	1.299	1.305	1.311	1.316	1.322	1.328	1.335
0.91	1.341	1.347	1.353	1.360	1.366	1.372	1.379	1.385	1.392	1.398
0.92	1.405	1.412	1.419	1.426	1.432	1.440	1.447	1.454	1.461	1.468
0.93	1.476	1.483	1.491	1.498	1.506	1.514	1.522	1.530	1.538	1.546
0.94	1.555	1.563	1.572	1.580	1.589	1.598	1.607	1.616	1.626	1.635
0.95	1.645	1.655	1.665	1.675	1.685	1.695	1.706	1.717	1.728	1.739
0.96	1.751	1.762	1.774	1.787	1.799	1.812	1.825	1.838	1.852	1.866
0.97	1.881	1.896	1.911	1.927	1.943	1.960	1.977	1.995	2.014	2.034
0.98	2.054	2.075	2.097	2.120	2.144	2.170	2.197	2.226	2.257	2.290
0.99	2.326	2.366	2.409	2.457	2.512	2.576	2.652	2.748	2.878	3.090

Tabelle 20.4: Quantile der χ^2-Verteilung mit k Freiheitsgraden

k	$\chi^2_{k;0.95}$	$\chi^2_{k;0.975}$	$\chi^2_{k;0.9833}$	$\chi^2_{k;0.9875}$	$\chi^2_{k;0.99}$
1	3.84	5.02	5.73	6.24	6.63
2	5.99	7.38	8.19	8.76	9.21
3	7.81	9.35	10.24	10.86	11.34
4	9.49	11.14	12.09	12.76	13.28
5	11.07	12.83	13.84	14.54	15.09
6	12.59	14.45	15.51	16.24	16.81
7	14.07	16.01	17.12	17.88	18.48
8	15.51	17.53	18.68	19.48	20.09
9	16.92	19.02	20.21	21.03	21.67
10	18.31	20.48	21.71	22.56	23.21
11	19.68	21.92	23.18	24.06	24.72
12	21.03	23.34	24.63	25.53	26.22
13	22.36	24.74	26.06	26.98	27.69
14	23.68	26.12	27.48	28.42	29.14
15	25.00	27.49	28.88	29.84	30.58
16	26.30	28.85	30.27	31.25	32.00
17	27.59	30.19	31.64	32.64	33.41
18	28.87	31.53	33.01	34.03	34.81
19	30.14	32.85	34.36	35.40	36.19
20	31.41	34.17	35.70	36.76	37.57
21	32.67	35.48	37.04	38.11	38.93
22	33.92	36.78	38.37	39.46	40.29
23	35.17	38.08	39.68	40.79	41.64
24	36.42	39.36	41.00	42.12	42.98
25	37.65	40.65	42.30	43.45	44.31

Tabelle 20.6: kritische Werte des Vorzeichentests

n	$s_{0.005}$	$s_{0.01}$	$s_{0.025}$	$s_{0.05}$	$s_{0.10}$
5				0	0
6			0	0	0
7		0	0	0	1
8	0	0	0	1	1
9	0	0	1	1	2
10	0	0	1	1	2
11	0	1	1	2	3
12	1	1	2	2	3
13	1	1	2	3	3
14	1	2	2	3	3
15	2	2	3	3	4
16	2	2	3	4	4
17	2	3	4	4	5
18	3	3	4	5	5
19	3	4	4	5	6
20	3	4	5	5	6

Tabelle 20.7: kritische Werte des Wilcoxon-Vorzeichen-Rangtests

n	$w_{0.005}$	$w_{0.01}$	$w_{0.025}$	$w_{0.05}$	$w_{0.10}$
5				0	2
6		0	0	2	3
7		0	2	3	5
8	0	1	3	5	8
9	1	3	5	8	10
10	3	5	8	10	14
11	5	7	10	13	17
12	7	9	13	17	21
13	9	12	17	21	26
14	12	15	21	25	31
15	15	19	25	30	36
16	19	23	29	35	42
17	23	27	34	41	48
18	27	32	40	47	55
19	32	37	46	53	62
20	37	43	52	60	69

Tabelle 20.8: Quantile der Teststatistik des Wilcoxon-Rangsummentests

$m = n$	$w_{0.005}$	$w_{0.01}$	$w_{0.025}$	$w_{0.05}$	$w_{0.10}$
3				6	7
4			10	11	14
5	15	16	17	19	20
6	23	24	26	28	30
7	32	34	36	39	41
8	43	45	49	51	55
9	56	59	62	66	70
10	71	74	78	82	87

Tabelle 20.9: Das 0.95-Quantil $F_{m,n;0.95}$ der F-Verteilung mit m und n Freiheitsgraden

n \ m	1	2	3	4	5	6	7
1	161.45	199.5	215.71	224.58	230.16	233.99	236.77
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33

Tabelle 20.10: Werte von $q_{0.95,p,n}$ für $p = 0.9, 0.95, 0.99$ und $n = 2, 3, \ldots, 20$

n \ p	0.90	0.95	0.99
2	32.019	37.674	48.430
3	8.380	9.916	12.861
4	5.369	6.370	8.299
5	4.275	5.079	6.634
6	3.712	4.414	5.775
7	3.369	4.007	5.248
8	3.136	3.732	4.891
9	2.967	3.532	4.631
10	2.839	3.379	4.433
11	2.737	3.259	4.277
12	2.655	3.162	4.150
13	2.587	3.081	4.044
14	2.529	3.012	3.955
15	2.480	2.954	3.878
16	2.437	2.903	3.812
17	2.400	2.858	3.754
18	2.366	2.819	3.702
19	2.337	2.784	3.656
20	2.310	2.752	3.615

Tabelle 20.11: Werte von $q_{0.99,p,n}$ für $p = 0.9, 0.95, 0.99$ und $n = 2, 3, \ldots, 20$

n \ p	0.90	0.95	0.99
2	160.194	188.491	242.301
3	18.930	22.401	29.055
4	9.398	11.150	14.527
5	6.612	7.855	10.260
6	5.337	6.345	8.301
7	4.613	5.488	7.187
8	4.147	4.936	6.468
9	3.822	4.550	5.966
10	3.582	4.265	5.594
11	3.397	4.045	5.308
12	3.250	3.870	5.079
13	3.130	3.727	4.893
14	3.029	3.608	4.737
15	2.945	3.507	4.605
16	2.872	3.421	4.492
17	2.808	3.345	4.393
18	2.753	3.279	4.307
19	2.703	3.221	4.230
20	2.659	3.168	4.161

Literaturverzeichnis

Beutelspacher, A., Schwenk, J., Wolfenstetter, K.-D. (2004): Moderne Verfahren der Kryptographie : von RSA zu Zero-Knowledget. Vieweg, Wiesbaden, 5th edition

B.Latane, Cappell, H. (1972): The effects of togetherness on heart rate in rats. Psychonomic Science, **29**, 177–179

Büning, H., Trenkler, G. (1994): Nichtparametrische statistische Methoden. de Gruyter, Berlin, 2nd edition

Burkschat, M., Cramer, E., Kamps, U. (2004): Beschreibende Statistik. Springer, Berlin

Deutsches PISA-Konsortium (Hrsg.) (2001): PISA 2000. Leske + Budrich, Opladen

Diekmann, A. (2004): Empirische Sozialforschung: Grundlagen, Methoden, Anwendungen. Rowohlt, Reinbek, 12th edition

Doane, D. P. (1976): Aesthetic frequency classifications. The American Statistician, **30**, 181–183

Ehrenberg, A. (1981): The Problem of Numeracy. The American Statistician, **35**, 67–71

Hahn, G. J. (1970): Statistical intervals for a normal population: part 1, tables, examples, and applications. Journal of Quality Technology, **2**, 115–125

Hartung, J., Elpelt, B., Klösener, K.-H. (2002): Statistik. Oldenbourg, München, 13th edition

Heiler, S., Michels, P. (1994): Deskriptive und explorative Datenanalyse. Oldenbourg, München

Hüttner, M. (2002): Grundzüge der Marktforschung. Oldenbourg, München, 7th edition

Kendall, M. G., Stuart, A., Ord, J. K. (1991): The advanced theory of statistics., volume 2 Classical inference and relationship. Arnold, London, 5th edition

Kippenhahn, R. (1999): Verschlüsselte Botschaften. Rowohlt, New York, 1st edition

Lann, A., Falk, R. (2005): A Closer Look at a Relatively Neglected Mean. Teaching Statistics, **27**, 76–80

Lehn, J., Müller-Gronbach, T., Rettig, S. (2000): Einführung in die Deskriptive Statistik. Teubner, Leipzig

Mood, A. M., Graybill, F. A., Boes, D. C. (1974): Introduction to the theory of statistics. McGraw-Hill, New York

Rosenblatt, M. (1956): Remarks on some nonparametric estimates of a density function. Ann. Math. Statist., **27**, 832–837

Scott, D. W. (1992): Multivariate Density Estimation. Wiley, New York

Silverman, B. W. (1986): Density Estimation. Chapman & Hall, London

Singh, S. (2000): Geheime Botschaften. Hanser, München

Tufte, E. R. (2001): The visual display of quantitative information. Graphics Press, Cheshire, 2nd edition

Tukey, J. W. (1977): Exploratory data analysis. Addison-Wesley, Reading

Utts, J. M. (1999): Seeing Through Statistics. Duxbury Press, Pacific Grove, 2nd edition

Wainer, H. (1997): Visual revelations : graphical tales of fate and deception from Napoleon Bonaparte to Ross Perot. Copernicus, New York

Zelzano, P., Zelzano, N., Kolb, S. (1972): Walking in the newborn. Science, **176**, 314–315