

Einführung in die Statistik für Wirtschafts- und Sozialwissenschaften I–II

Michael Mayer · Dirk Klingbiel

Herbstsemester 2016 und
Frühjahrssemester 2017

28. August 2016



Dank

Wir danken Lutz Dümbgen herzlich für seine hilfreichen Tipps und dafür, dass wir viele Teile von seinem bisherigen Vorlesungsskript übernehmen durften.

Das R-Skript im Anhang entspricht im Wesentlichen einem Software-Skript von Michael Vock. Merci!

Grosser Dank gebührt Martin Bigler für das ergiebige Lektorat und Qiyu Li für die Illustration der Titelseite.

Zudem danken wir etlichen Studierenden für ihre konstruktiven Rückmeldungen.

Inhaltsverzeichnis

1 Überblick	8
1.1 Einleitung	8
1.2 Elementare statistische Begriffe	9
1.3 Variablentypen	10
1.4 Aufbau	10
1.5 Software	11
1.6 Ziele	11
1.7 Zusammenfassung	11
I Univariate Verfahren	12
2 Beschreibung von Daten	13
2.1 Kategoriale Merkmale	13
2.1.1 Quantitative Beschreibung	13
2.1.2 Grafische Darstellung	14
2.2 Numerische Merkmale	15
2.2.1 Grafische Darstellung	15
2.2.2 Quantitative Beschreibung	20
2.3 Datensätze	26
2.4 Zusammenfassung	33
3 Schliessende Statistik und Wahrscheinlichkeitsrechnung	34
3.1 Zufallsvariablen und ihre Verteilungen	35
3.1.1 Zufallsvariablen	35
3.1.2 Wahrscheinlichkeiten	36
3.1.3 (Wahrscheinlichkeits-)Verteilung	37
3.1.4 Zufallsvariablen in der Statistik	37
3.1.5 Charakterisierung von Verteilungen	38
3.2 Binomialverteilung und Verwandtes	55

3.2.1	Binomialverteilung	55
3.2.2	Konfidenzintervalle allgemein und für relative Anteile	57
3.2.3	Tests allgemein und für relative Anteile	62
3.2.4	Konfidenzintervalle für Quantile	69
3.2.5	Poissonverteilung	70
3.3	Normalverteilung und Verwandtes	73
3.3.1	Normalverteilung	73
3.3.2	Konfidenzintervalle und Tests für einen Mittelwert	77
3.3.3	Verfeinerung nach Students Methode	81
3.4	Gammaverteilung und Verwandtes	84
3.4.1	Gammaverteilung	84
3.4.2	Chiquadrat-Verteilung	85
3.4.3	Konfidenzintervalle für eine Standardabweichung	86
3.4.4	Anpassungstest für ein kategorielles Merkmal	87
3.5	Zusammenfassung	89
II	Bivariate Verfahren	92
4	Zwei kategoriale Merkmale	94
4.1	Häufigkeitstabellen	94
4.2	Verdeutlichung des Zusammenhangs	95
4.3	Stärke des Zusammenhangs	99
4.4	Aussagen über die Population	101
4.5	Weitere Beispiele	103
4.6	Vierfeldertafeln und Odds Ratios	108
4.7	Exakter Chiquadrat-Unabhängigkeitstest	112
4.8	Zusammenfassung	112
5	Ein kategorielles, ein numerisches Merkmal	113
5.1	Stratifizierte Beschreibung der Stichprobe	113
5.1.1	Grafische Darstellung	113
5.1.2	Quantitative Auswertung	120
5.2	Vergleich zweier Mittelwerte	121
5.2.1	Aussagen über die Stichprobe	121
5.2.2	Aussagen über die Population	122
5.3	Vergleich mehrerer Mittelwerte	126
5.3.1	Bestimmtheitsmaß	126

5.3.2 Aussagen über die Population	128
5.3.3 Mittelwertunterschiede	133
5.4 Verbundene Stichproben	136
5.4.1 Strukturierung des Datensatzes	137
5.4.2 Naheliegende Analysemöglichkeit im Zweistichprobenfall	138
5.4.3 Weitere Tests bei verbundenen Stichproben	143
5.5 Zusammenfassung	144
6 Zwei numerische Merkmale	145
6.1 Streudiagramm	145
6.2 Lineare Regression	146
6.2.1 Regressionskoeffizienten und Regressionsgerade	146
6.2.2 Leverage-Effekt	148
6.2.3 Vorhersagen	149
6.2.4 Aussagen über die Population	150
6.3 Kovarianz, Bestimmtheitsmass, Korrelationen	153
6.3.1 Kovarianz	153
6.3.2 Bestimmtheitsmass	154
6.3.3 Korrelationskoeffizient nach Pearson	155
6.3.4 Rangkorrelationskoeffizient nach Spearman	156
6.3.5 Aussagen über die Population	158
6.4 Zusammenfassung	164
III Multivariate Verfahren	165
7 Das lineare Modell	166
7.1 Modellstruktur	167
7.2 Aussagen über die Stichprobe	168
7.2.1 Intercept, Effekte, Linearer Prädiktor, Vorhersagen	168
7.2.2 Modellgüte	171
7.3 Modifikationen der Modellgleichung	175
7.3.1 Transformationen	175
7.3.2 Nichtlinearitäten	180
7.3.3 Interaktionen	181
7.4 Aussagen über die Population	182
7.4.1 Modellparameter	182
7.4.2 Vorhersagen	183

7.4.3	Modellgüte	184
7.5	Multikollinearität	191
7.6	Modellvoraussetzungen und deren Überprüfung	192
7.6.1	Passende Modellstruktur	192
7.6.2	Gleiche Varianz	194
7.6.3	Normalverteilung	195
7.6.4	Keine einflussreichen Beobachtungen	195
7.6.5	Unabhängigkeit	195
7.7	Weitere statistische Modelle	199
7.8	Vorbereitung des Modells	200
7.8.1	Daten	200
7.8.2	Wahl der Modellstruktur	200
7.9	Zusammenfassung	202
8	Dimensionsreduktion	203
8.1	Wichtigste Variable auswählen	204
8.2	Summen	204
8.3	Hauptkomponentenanalyse	206
8.4	Clusteranalyse	213
8.5	Zusammenfassung	216
A	R-Skript	217
A.1	Übersicht	217
A.1.1	Fenster	217
A.1.2	Bedienungssprache	217
A.1.3	Dokumentation	218
A.1.4	Zusatzpakete	218
A.1.5	Daten	218
A.1.6	R als Taschenrechner	218
A.2	Programmierung	219
A.2.1	Objekte	220
A.2.2	Zuweisung, Anzeige von Objekten	221
A.2.3	Funktionen	222
A.2.4	Syntax-Regeln	223
A.2.5	Programmfluss	224
A.3	Datenaufbereitung	224
A.3.1	Direkte Eingabe in einen Data Frame	224

A.3.2 Einlesen eines Data Frame aus einer Datei	225
A.3.3 Transformieren eines Data Frame	227
A.3.4 Erzeugen von Daten	229
A.3.5 Sortieren von Daten	232
A.3.6 Namen von Elementen und Faktor-Stufen	232
A.4 Grafiken	232
A.5 Statistische Verfahren	235
A.5.1 Univariate Verfahren	235
A.5.2 Bivariate Verfahren	235
A.5.3 Multivariate Verfahren	236
A.5.4 Zusatz: Weitere Konfidenzintervalle	237
B Formelsammlung	238

Kapitel 1

Überblick

1.1 Einleitung

“Traue keiner Statistik, die du nicht selbst gefälscht hast!”

Dieser häufig gehörte und populäre Ausspruch belegt, dass viele Leute ein falsches Bild von dieser Disziplin haben, nämlich dass Statistiken beliebig manipulierbar sind.

Tatsächlich handelt es sich bei der Statistik um eine präzise Wissenschaft mit starker Anbindung an die Mathematik und Informatik; die Grenzen sind fliessend. Sie wird in verschiedenen naturwissenschaftlichen (biologischen, geographischen, ökologischen), wirtschaftlichen, medizinischen, psychologischen und industriellen sowie amtlichen Gebieten eingesetzt, um aus Daten relevante Informationen zu gewinnen. Dabei nimmt ihre Bedeutung zu, da im Zeitalter des Computers überall Daten gesammelt werden.

Einige Fragestellungen der Statistik:

- Wie nehmen die mittleren Krankheitskosten mit dem Alter zu?
- Wie hoch ist der typische Mietpreis einer 3-Zimmer-Wohnung?
- Verdienen Managerinnen im Schnitt weniger als Manager (bereinigt nach Alter, Position etc.)?
- Wie präzise sind die Wahlprognosen?
- Lohnt es sich bei fallenden Aktienkursen in der Regel, Gold zu kaufen?
- Wie soll ein Wertschriftenportfolio zusammengestellt sein, damit die erwartete Rendite gross und die Volatilität klein ist?
- Welche Kundensegmente verursachen die wenigsten Autoversicherungsschäden?
- Tendieren Schüler und Schülerinnen mit vielen Absenzen zu schlechten Leistungen?

Um solche (und weitere) Fragen zu untersuchen, werden Daten aus

- Datenbanken (z. B. von Krankenkassen, Versicherungen, dem Internet),
- Umfragen (z. B. bei MitarbeiterInnen, potenziellen KundInnen) oder
- Experimenten (z. B. medizinische Studien, Produktevergleiche)

beigezogen bzw. beschafft. Deren Auswertung umfasst typischerweise die Schritte:

1. Daten in das Statistik-Programm einlesen.
2. Daten verstehen, überprüfen, korrigieren und für die Analyse vorbereiten.
3. Eigentliche statistische Analyse (Daten beschreiben, Fragestellungen untersuchen).
4. Ergebnisse publikumsgerecht präsentieren.

Bevor wir den Aufbau der Vorlesung darlegen, führen wir einige elementare statistische Begriffe ein.

1.2 Elementare statistische Begriffe

Daten sind in Tabellen abgelegt. Jede Zeile eines solchen *Datensatzes (Stichprobe)* entspricht einer *Beobachtung* mit *Werten (Ausprägungen, Stichprobenwerten)* von meist mehreren *Variablen (Merken)*. Letztere bilden die Spalten der Tabelle. Die Anzahl der Beobachtungen wird *Stichprobenumfang* genannt.

Beispiel 1.1 (Befragung von Studierenden). In der Vorlesung “Einführung in die Statistik für Wirtschafts- und Sozialwissenschaften I” (Universität Bern, Wintersemester 2003/2004) machten $n = 263$ Studierende Angaben zu folgenden Merkmalen:

Geschlecht	M/W
Alter	Alter in Jahren
GebMonat	Geburtsmonat, kodiert als Zahl von 1 bis 12
Herkunft	Geburtskanton bzw. Geburtsland
Kgroesse	Körpergrösse in cm
Kgewicht	Körpergewicht in kg
MonMiete	Nettomiete pro Monat in CHF
Rauchen	Rauchen (nein = 0, gelegentlich = 1, regelmässig = 2)
ZufZiffer	“Zufallsziffer” von 0 bis 9
AnzGeschw	Anzahl Geschwister
GeschGrDoz	Geschätzte Grösse des Dozenten in cm

Wir werfen einen Blick auf die ersten sechs¹ Beobachtungen dieses Datensatzes.

R Code								
# Eingabe								
head(wiso)								
# Ausgabe								
Geschlecht Alter GebMonat Herkunft Kgroesse Kgewicht MonMiete Rauchen								
1	M	20	3	Luzern	176	68	454	0
2	M	22	11	Luzern	169	62	550	2
3	M	22	9	Bern	178	72	0	0
4	M	22	8	Aargau	170	65	665	2
5	M	21	10	Bern	164	57	0	0
6	W	19	10	Solothurn	165	49	0	0
ZufZiffer AnzGeschw GeschGrDoz								
1	5	2	179					
2	8	1	179					
3	8	2	180					
4	6	2	176					
5	6	1	178					
6	4	1	180					

Z. B. weist die vierte Beobachtung beim Merkmal ‘Herkunft’ den Stichprobenwert ‘Aargau’ auf. ▲

¹Dies wird in der Statistik-Software R mit dem Befehl head angefordert.

1.3 Variablenarten

Wir unterscheiden primär zwischen zwei Typen von Variablen, für die jeweils andere statistische Verfahren eingesetzt werden: *Numerische (quantitative) Variablen* nehmen einen Zahlenwert mit objektiver Bedeutung an, während *kategoriale (qualitative) Variablen* Werte in irgendeinem Bereich annehmen.

Beispiel 1.2 (Befragung von Studierenden, Fortsetzung). In Beispiel 1.1 sind die Merkmale ‘Geschlecht’, ‘Geburtsmonat’, ‘Herkunft’, ‘Rauchen’ und ‘Zufallsziffer’ kategoriall, wohingegen die Variablen ‘Alter’, ‘Körpergrösse’, ‘Körpergewicht’, ‘Monatsmiete’, ‘Anzahl Geschwister’ und ‘geschätzte Grösse des Dozenten’ numerisch sind. Die Variable ‘Rauchen’ ist zwar ebenfalls zahlenkodiert, aber die Ausprägungen wurden willkürlich gewählt. ▲

Zwei spezielle Arten von kategorialen Variablen sind die folgenden: *Ordinalale Variablen* sind kategoriale Variablen, deren Werte in einer natürlichen Reihenfolge stehen mit einem “kleinsten” und einem “grössten” Wert. Solche Variablen sind gerade in der Medizin, in der Psychologie und in den Sozialwissenschaften sehr verbreitet. Man denke beispielsweise an Fragen zur Zufriedenheit mit irgendetwas, bei denen z. B. eine der folgenden Antworten anzukreuzen ist: ‘unzufrieden’, ‘teilweise zufrieden’, ‘überwiegend zufrieden’, ‘rundum zufrieden’. Mitunter entstehen ordinale Variablen aus numerischen Merkmalen durch Einteilung ihres Wertebereichs in Intervalle.

Eine kategoriale Variable, die nur zwei verschiedene Werte annehmen kann, heisst *binär* bzw. *dichotom*. Eine zahlenkodierte binäre Variable gilt stets auch als ordinal und numerisch.

Beispiel 1.3 (Befragung von Studierenden, Fortsetzung). In Beispiel 1.1 ist die Variable ‘Rauchen’ ordinal: 0 (nein) \leq 1 (gelegentlich) \leq 2 (regelmässig). ‘Geschlecht’ ist ein binäres Merkmal. ▲

1.4 Aufbau

In Teil I des Skripts lernen wir, Merkmale eines Datensatzes einzeln, d. h. *univariat*, zu beschreiben, um beispielsweise die Stichprobe zu charakterisieren oder einfache Fragestellungen zu beantworten. Verfahren der *beschreibenden Statistik* ergänzen wir mit Konzepten der *schliessenden Statistik*: Diese ermöglicht anhand der Daten allgemeine Rückschlüsse auf die Grundgesamtheit, aus der die Daten stammen. Dazu benötigen wir einige Grundlagen aus der Wahrscheinlichkeitsrechnung.

Viele statistische Fragestellungen drehen sich um Zusammenhänge zwischen mehreren Merkmalen eines Datensatzes, beispielsweise zwischen ‘Einkommen’ und ‘Geschlecht’. In Teil II lernen wir die Verfahren der *bivariate Statistik* kennen, mit denen Zusammenhänge zwischen zwei Merkmalen analysiert werden.

In Teil III kümmern wir uns schliesslich um die zentralen *multivariaten* Verfahren, mit denen Zusammenhänge zwischen mehr als zwei Merkmalen studiert werden können. Wie auch in den Teilen I und II unterscheiden wir dabei zwischen Verfahren der beschreibenden und schliessenden Statistik.

Der Aufbau orientiert sich generell daran, dass statistische Verfahren anhand folgender Fragen eingeteilt werden können:

- Wie viele Merkmale sind an der Fragestellung beteiligt?
- Welche Variablenarten weisen diese Merkmale auf?
- Soll eine Aussage über die Stichprobe oder die Grundgesamtheit gemacht werden?
- Soll die Antwort in Zahlen oder Bildern erfolgen?

1.5 Software

Für die statistische Auswertung von Daten gibt es zahlreiche Programme. Zu den bekanntesten gehören R, SAS, SPSS und STATA. Zur Illustration der Verfahren arbeiten wir mit R. Es hat sich in den letzten Jahren zu einem der beliebtesten Programme gemausert.

- R ist frei erhältlich,
- kann sowohl $1 + 1$ berechnen als auch komplexe statistische Verfahren durchführen,
- ist beliebig erweiterbar,
- erstellt schöne Grafiken und
- kann unter Windows, Mac OS und Linux installiert werden.

Damit Sie sich mit R vertraut machen können, ist in Anhang A ein kleines Skript angefügt.

1.6 Ziele

Die Hauptziele der Vorlesung lauten folgendermassen:

- Sie verstehen die statistischen Aspekte eines Dokuments.
- Sie können Auswertungen durchführen und deren Ergebnisse sinnvoll präsentieren.
- Sie verfügen über die relevanten theoretischen Grundlagen.

1.7 Zusammenfassung

- Es wurden einige allgemeine Aussagen über Statistik gemacht und wichtige Begriffe eingeführt.
- Wir haben Merkmale anhand ihres Variablentyps unterschieden.
- Schliesslich haben wir den Aufbau des Skripts bzw. der Vorlesung präsentiert und begründet, weshalb die Beispiele mit der Software R illustriert werden. Zudem haben wir die Ziele der Vorlesung festgelegt.

Teil I

Univariate Verfahren

Kapitel 2

Beschreibung von Daten

Dieses Kapitel präsentiert die wichtigsten statistischen Verfahren zur quantitativen und grafischen Beschreibung¹ einzelner Merkmale bzw. der Verteilung ihrer Ausprägungen. Neben der Beschreibung der Stichprobe und der Beantwortung einfacher Fragestellungen (z. B. “Wie gross ist der Raucheranteil unter den Befragten?”) hilft die univariate Beschreibung eines Merkmals, dessen Bedeutung zu verstehen (z. B. ‘Körpergrösse’ in m oder cm?) und Datenfehler zu entdecken (z. B. eine 1.80 cm grosse Person?).

Die Verfahren unterscheiden sich nach VariablenTyp.

2.1 Kategoriale Merkmale

Wir betrachten eine kategoriale Variable X mit L Kategorien x_1, x_2, \dots, x_L . Die Stichprobenwerte dieser Variable seien X_1, X_2, \dots, X_n ; dies sind also die Einträge einer Spalte des Datensatzes.

Beispiel 2.1 (Befragung von Studierenden, Fortsetzung). Die Variable ‘Rauchen’ (X) weist die drei Kategorien $x_1 = 0$ (nein), $x_2 = 1$ (gelegentlich) und $x_3 = 2$ (regelmässig) auf. Die ersten vier Stichprobenwerte betragen $X_1 = 0$, $X_2 = 2$, $X_3 = 0$ und $X_4 = 2$. ▲

Das Merkmal X wird beschrieben, indem man die Häufigkeit jeder Kategorie in Zahlen (quantitativ) oder auch grafisch angibt.

2.1.1 Quantitative Beschreibung

Für $j = 1, 2, \dots, L$ bezeichnen wir mit H_j die *absolute Häufigkeit* der Kategorie x_j in der Stichprobe, also die Anzahl aller Beobachtungen mit Ausprägung x_j . Durch die Angabe aller absoluten Häufigkeiten wird das Merkmal vollständig² beschrieben. Alternativ kann man auch die *relativen Häufigkeiten* $f_j = H_j/n$ betrachten, also die relativen Anteile von Beobachtungen mit Ausprägung x_j .

Hinweise

- Die häufigste Ausprägung wird *Modus* genannt.
- Seltene Kategorien werden manchmal zu einer Kategorie ‘Other’ zusammengefasst.
- Relative Häufigkeiten f_i werden oft in Prozenten ($f_i \cdot 100\%$) ausgedrückt.

¹Die “beschreibende Statistik” wird auch “deskriptive Statistik” genannt.

²Es könnten also daraus die einzelnen Ausprägungen (bis auf ihre Reihenfolge im Datensatz) rekonstruiert werden.

Beispiel 2.2 (Rauchen). Die Frage nach dem Rauchen in Beispiel 1.1 wurde von $n = 261$ Studierenden beantwortet. Ausprägung ‘0’ (nein) wurde $H_1 = 171$ mal genannt, ‘1’ (gelegentlich) $H_2 = 47$ mal und ‘2’ (regelmässig) $H_3 = 43$ mal. Die entsprechenden relativen Häufigkeiten betragen $f_1 = 171/261 \approx 0.655$, $f_2 = 47/261 \approx 0.180$ und $f_3 = 43/261 \approx 0.165$. Die Stichprobe besteht also aus 65.5% NichtraucherInnen, 18% GelegenheitsraucherInnen und 16.5% regelmässigen RaucherInnen.

Die Zahlen können als Tabelle präsentiert werden:

j	1	2	3
x_j	0 (nein)	1 (gelegentl.)	2 (regelm.)
H_j	171	47	43
f_j	0.655	0.180	0.165

Statt von Hand zu zählen, wie häufig jede Kategorie ist, kann die Arbeit der Software überlassen werden¹.

R Code

```
# Eingabe
a.H <- table(wiso$Rauchen)
a.H

# Ausgabe
0   1   2
171  47  43

# Eingabe, Forts.
r.H <- prop.table(a.H)
r.H

# Ausgabe
0       1       2
0.65517 0.18008 0.16475
```



2.1.2 Grafische Darstellung

Die absoluten bzw. relativen Häufigkeiten H_j bzw. f_j werden grafisch als *Balkendiagramm* oder *Stabdiagramm* dargestellt. Dabei wird zu jeder Kategorie x_j ein Balken oder Stab der Höhe H_j bzw. f_j eingezeichnet. Als “tintensparende” Alternative können die Häufigkeiten auch einfach als Punkte eingezeichnet werden. Ein solches Diagramm wird manchmal *Punktediagramm* genannt.

Eine weitere Möglichkeit ist das *Kuchendiagramm*, bei dem die Häufigkeiten durch entsprechend grosse Stücke eines runden Kuchens dargestellt werden.

Beispiel 2.3 (Rauchen, Fortsetzung). Betrachten wir diese Grafiken² für unser Beispiel (Abbildung 2.1).

R Code

```
# Eingabe, Forts.
barplot(a.H)      # Balkendiagramm
plot(r.H)        # Stabdiagramm
dotchart(r.H)    # Punktediagramm
pie(r.H)         # Kuchendiagramm
```



¹Tabellen werden in R mit den Befehlen `table` und `prop.table` erzeugt.

²Die genannten Diagramme werden in R mit den Befehlen `barplot`, `plot`, `dotchart` und `pie` gezeichnet. Um den im Skript gezeigten R-Code übersichtlich zu halten, werden Grafikoptionen wie Änderung der Schriftgrösse etc. generell weggelassen.

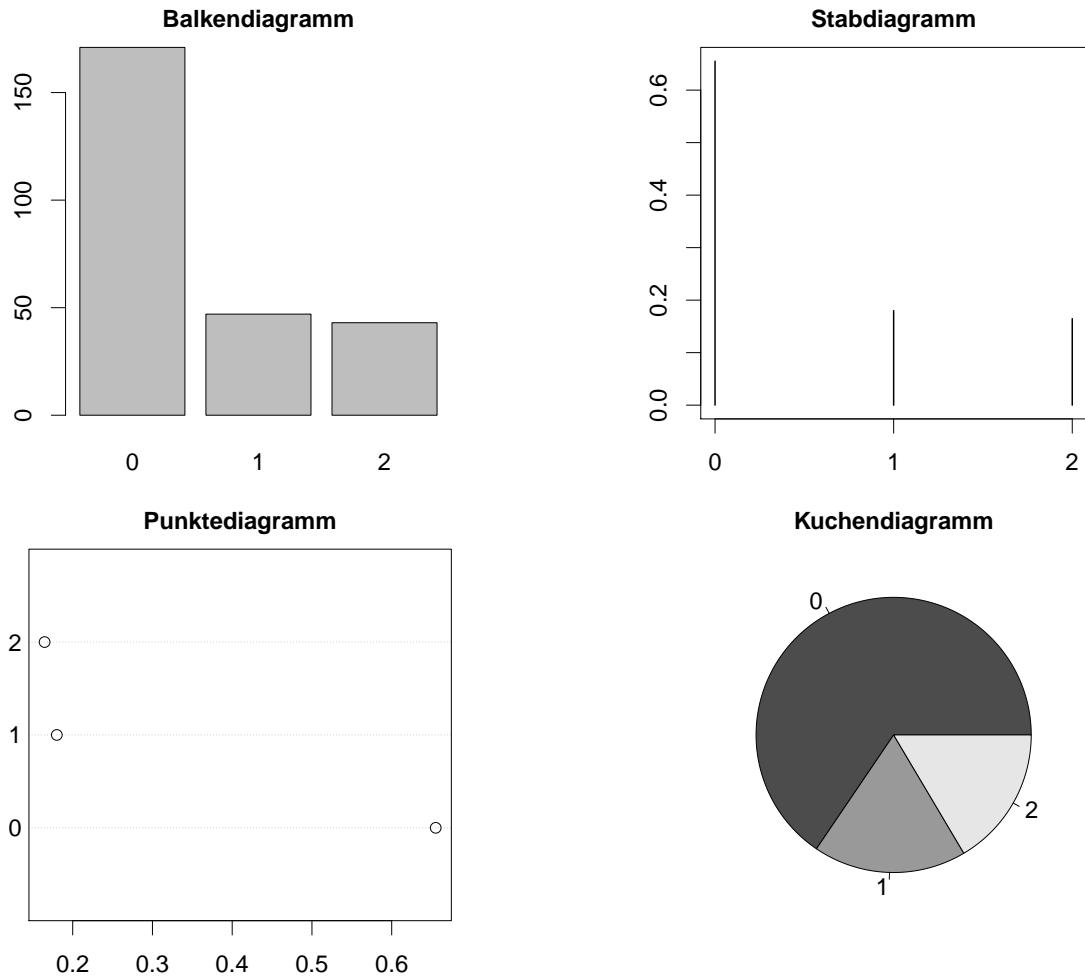


Abbildung 2.1: Verschiedene grafische Darstellungen der absoluten und relativen Häufigkeiten des Merkmals ‘Rauchen’.

2.2 Numerische Merkmale

Nun betrachten wir ein numerisches Merkmal (X) mit den Stichprobenwerten X_1, X_2, \dots, X_n .

Indem man die Werte zu Kategorien zusammenfasst, lässt sich X im Prinzip stets mit den Verfahren für kategoriale Variablen analysieren. Da durch die Kategorisierung in der Regel jedoch wesentliche Information verloren geht, präsentieren wir in diesem Abschnitt Alternativen zu dieser Vorgehensweise.

2.2.1 Grafische Darstellung

Die Verteilung der Stichprobenwerte einer numerischen Variable wird mit ihrer *empirischen Verteilungsfunktion* und/oder ihrem *Histogramm* visualisiert.

Empirische Verteilungsfunktion

Für eine beliebige Schranke r definieren wir

$$F(r) := \text{relativer Anteil von Beobachtungen mit } X \leq r.$$

Dies liefert die sogenannte *empirische Verteilungsfunktion* (kurz: ECDF von engl. *empirical cumulative distribution function*). Von ihrem Graphen kann man ablesen, wie die X -Werte in der Stichprobe verteilt sind (siehe Beispiel 2.5 für konkrete Möglichkeiten).

Aufgrund der Definition gilt:

- Ab dem grössten Stichprobenwert beträgt die ECDF 1, links vom kleinsten ist sie 0.
- Der Graph verläuft horizontal, ausser an den Stellen der Stichprobenwerte. Dort springt er um den relativen Anteil der Beobachtungen mit diesem Wert. So entsteht eine Art Treppe.
- In steilen Abschnitten liegen viele Werte, in flachen wenige.
- Der grösste Sprung identifiziert den Modus, also den häufigsten Stichprobenwert.

Beispiel 2.4 (Konstruktion einer ECDF). Wir betrachten die ersten 10 Altersangaben im Datensatz von Beispiel 1.1 und zeichnen die ECDF davon (Abbildung 2.2). Die sortierten Stichprobenwerte lauten

$$18, 19, 19, 20, 20, 21, 22, 22, 22, 22.$$

Links vom kleinsten Stichprobenwert (18) hat die ECDF den Wert 0. Bei 18 springt sie um $1/10 = 0.1$, bei 19 und 20 je $2/10 = 0.2$, bei 21 um $1/10 = 0.1$ und bei 22 um $4/10 = 0.4$. Ab 22 beträgt sie 1. \blacktriangle

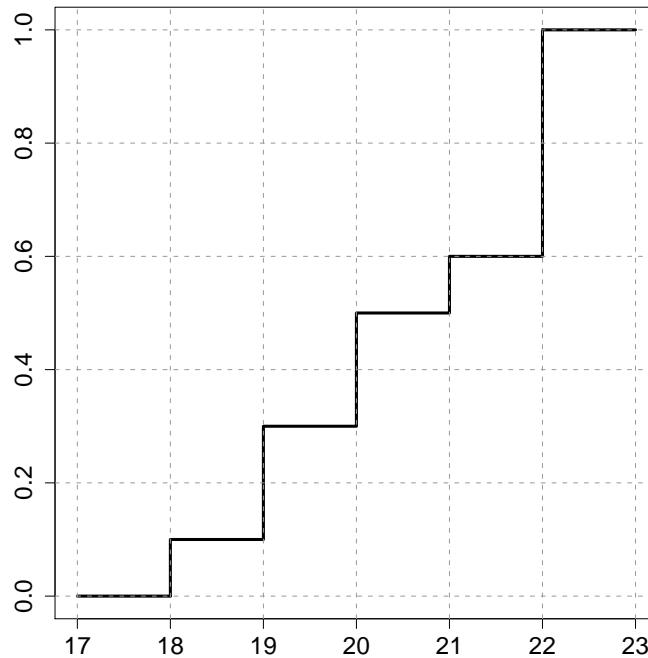


Abbildung 2.2: ECDF der 10 Altersangaben in Beispiel 2.4.

Beispiel 2.5 (Befragung von Studierenden, Fortsetzung). Wir studieren nun die empirischen Verteilungsfunktionen vierer Variablen aus Beispiel 1.1.

Da der Stichprobenumfang (bzw. die jeweilige Anzahl verschiedener Stichprobenwerte) gross ist, lassen wir die ECDFs von der Software¹ zeichnen und konzentrieren uns auf die Beschreibung der Verteilungen in Abbildung 2.3.

¹ECDFs können in R mit der Funktion `Ecdf` gezeichnet werden. Diese ist nicht in der Standardinstallation enthalten, sondern befindet sich im Zusatzpaket `Hmisc` von Frank Harrell. Mit der Option `datadensity = 'rug'` wird zusätzlich ein Stripchart über die x-Achse gelegt. Dessen Strichlein werden hier leicht verzittert eingezeichnet, um gleiche Werte optisch unterscheiden zu können.

Um zu verdeutlichen, in welchen Bereichen viele Stichprobenwerte liegen, werden diese zusätzlich als sogenanntes *Stripchart* über die x -Achse eingezeichnet.

R Code

```
# Eingabe
library(Hmisc)

Ecdf(wiso$Alter, datadensity = 'rug')
Ecdf(wiso$GeschGrDoz, datadensity = 'rug')
Ecdf(wiso$Kgroesse, datadensity = 'rug')
Ecdf(wiso$MonMiete, datadensity = 'rug')
```

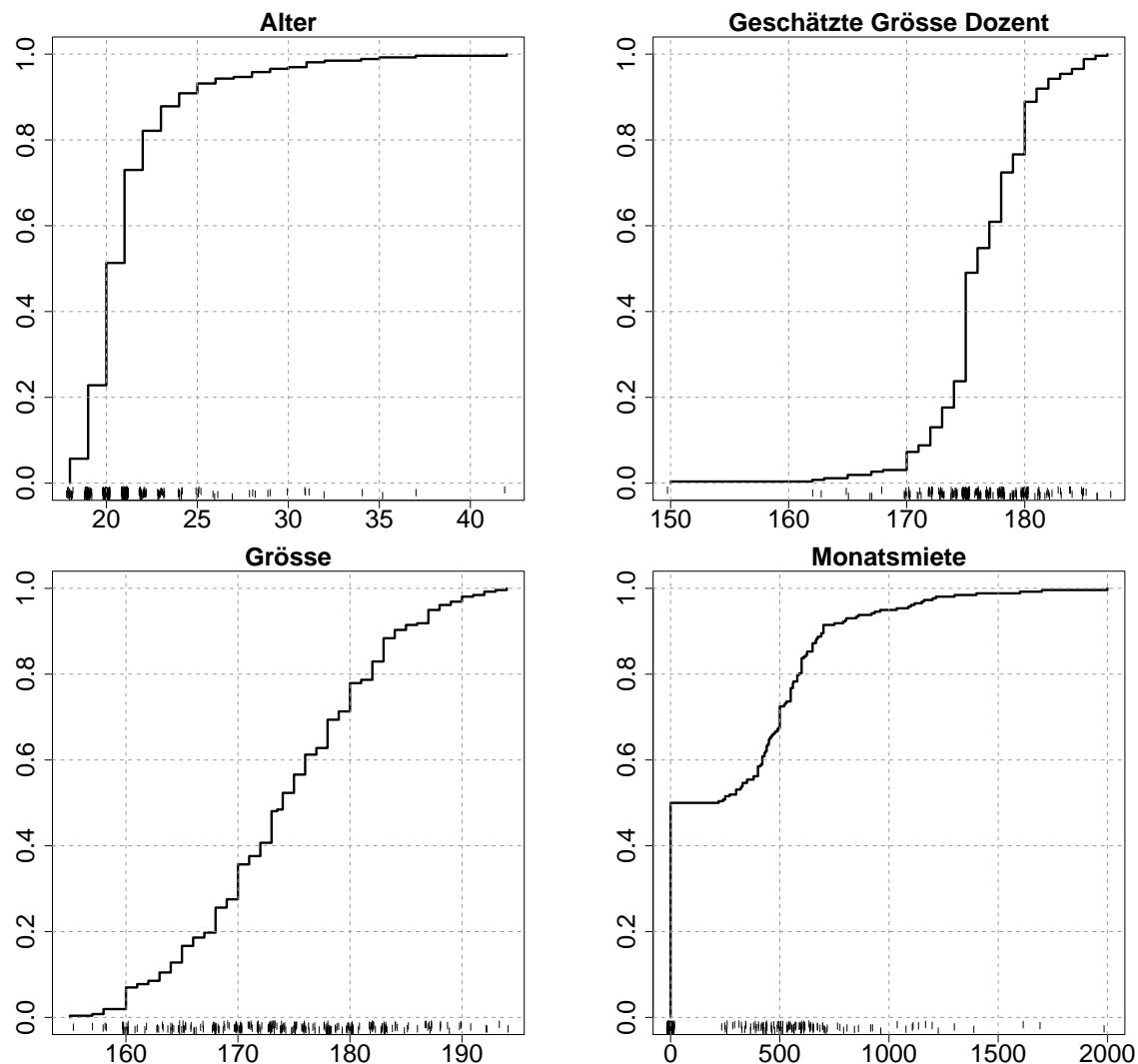


Abbildung 2.3: ECDFs einiger Merkmale in Beispiel 2.5 inkl. Stripcharts.

Kommentare

- ‘Alter’: Man kann erahnen, dass die jüngste Person 18 und die älteste 42 Jahre alt ist. Am stärksten vertreten sind die 20-Jährigen mit einem relativen Anteil von fast 30%. Rund 80% sind höchstens 22 Jahre alt (ab 22 liegt die Kurve über 0.8), entsprechend sind rund 20% älter als 22. Beim Wert 20 springt die Kurve auf knapp über 0.5: Leicht mehr als die Hälfte ist höchstens 20 Jahre alt.

- ‘*Geschätzte Grösse des Dozenten*’: Der kleinste Tipp beträgt 150 cm, der grösste 187. Rund die Hälfte (50%) tippt auf einen Grösse von 175 cm oder weniger (beim Wert 175 springt die Kurve auf etwa 0.5). Der korrekte Wert – 176 cm – wurde nur selten genannt.
- ‘*Körpergrösse*’: Die Werte befinden sich (etwa) im Bereich von 155 bis 195 cm. Man kann abschätzen, dass rund 50% kleiner sind als 175 cm und dass ca. jede vierte Person 180 cm oder grösser ist.
- ‘*Monatsmiete*’: Hier sehen wir, dass etwa die Hälfte der Studierenden keine Miete bezahlt (die Kurve springt bei 0 von 0 auf ca. 0.5), dass niemand zwischen 1 und 200 CHF aufwendet (die Kurve verläuft in diesem Bereich horizontal) und dass rund 10% der Befragten mehr als 700 CHF pro Monat für die Miete ausgeben (ab ca. 700 CHF liegt die Kurve über 0.9). Bei 500 CHF beträgt die ECDF rund 0.7, d. h. etwa 70% der Personen geben höchstens 500 CHF für die Miete aus. ▲

Histogramm

Eine andere, viel populärere Darstellung als die empirische Verteilungsfunktion ist das *Histogramm*. Es liefert einen Eindruck, in welchem Bereich wie viele Stichprobenwerte liegen (“Dichte”) und welche Form die Verteilung der X -Werte hat.

Histogramme sind eng verwandt mit Balkendiagrammen: Die Stichprobenwerte werden zu Kategorien zusammengefasst, deren Häufigkeiten im Wesentlichen als Balkendiagramm visualisiert werden. Die genaue Konstruktion ist die folgende:

1. Der Wertebereich von X wird durch Schranken $a_0 < a_1 < a_2 < \dots < a_L$ in Intervalle unterteilt.
2. Man bestimmt für jedes Intervall¹ $I_j := (a_{j-1}, a_j]$ die absoluten bzw. relativen Häufigkeiten H_j bzw. $f_j = H_j/n$.
3. Für jedes Intervall wird ein Rechteck mit Grundseite I_j und Höhe H_j (Konvention 1) oder $f_j/\text{Länge}(I_j)$ (Konvention 2) gezeichnet.

Bei Konvention 2 entspricht die Fläche des j -ten Rechtecks gerade dem relativen Anteil aller Beobachtungen im Intervall I_j .

Sind alle Intervalle gleich gross, liefern beide Konventionen das gleiche Bild bis auf eine unterschiedliche Beschriftung der vertikalen Achse. Ansonsten sollte man aber unbedingt Konvention 2 verwenden: Einerseits vermeidet man dadurch Verzerrungen durch die unterschiedlich langen Intervalle, da beim Betrachten vor allem die Flächen der Rechtecke wahrgenommen werden. Ausserdem kann man mit Konvention 2 die Histogramme unterschiedlicher (Teil-)Stichproben vergleichen, selbst wenn unterschiedliche Intervalleinteilungen oder Stichprobenumfänge vorliegen.

Beispiel 2.6 (Konstruktion eines Histogramms). Betrachten wir wiederum die ersten 10 (sortierten) Altersangaben 18, 19, 19, 20, 20, 21, 22, 22, 22, 22 von Beispiel 1.1.

Die Intervalleinteilung legen wir durch die Schranken $a_0 = 16$, $a_1 = 18$, $a_2 = 20$ und $a_3 = 22$ fest. So entstehen Intervalle je mit Länge 2. Im ersten Intervall $I_1 := (16, 18]$ liegt ein einzelner Stichprobenwert, im zweiten Intervall $I_2 := (18, 20]$ vier und im dritten Intervall $I_3 := (20, 22]$ schliesslich die restlichen fünf. Nach Konvention 1 ergeben sich damit Rechtecke der Höhen $H_1 = 1$, $H_2 = 4$ und $H_3 = 5$. Dividieren wir diese Werte durch den Stichprobenumfang 10 und die Längen der Intervalle (je 2), so gelangen wir zu den entsprechenden Höhen nach Konvention 2, also zu $1/(2 \cdot 10) = 0.05$, $4/(2 \cdot 10) = 0.2$ und $5/(2 \cdot 10) = 0.25$.

Die beiden Möglichkeiten sind in Abbildung 2.4 ersichtlich. ▲

¹Eine runde Klammer bedeutet “ohne Schranke”, eine eckige “mit Schranke”.

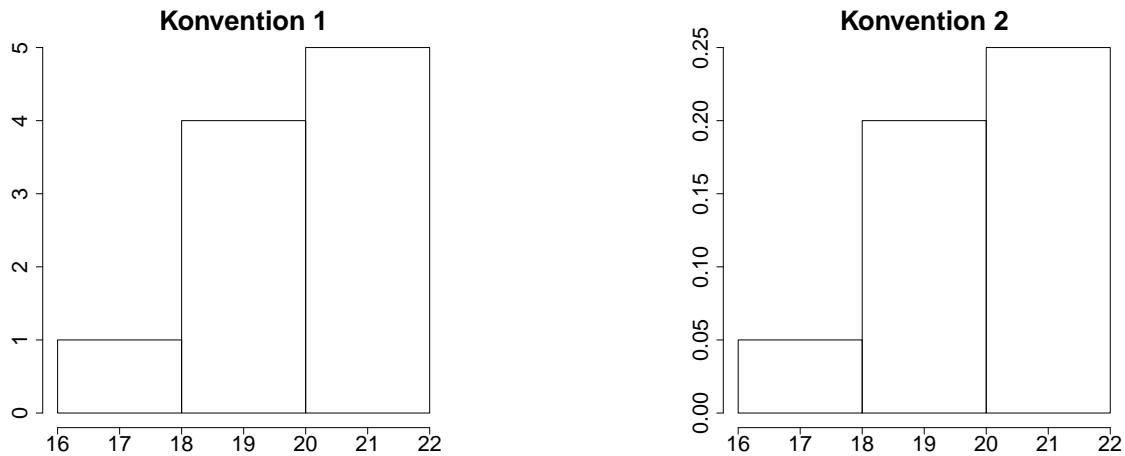


Abbildung 2.4: Histogramm der 10 Altersangaben von Beispiel 2.6, links nach Konvention 1, rechts nach Konvention 2.

Die *Form* der Verteilung der Stichprobenwerte wird u. a. durch die *Schiefe* und die *Anzahl der Höcker* beschrieben. Wir beurteilen die Verteilung

- als *linksschief*, wenn das Histogramm links deutlich weniger steil ist als rechts,
- als *rechtsschief*, wenn es rechts deutlich weniger steil ist als links, oder
- als *symmetrisch*, wenn es ungefähr symmetrisch ist.

Anhand des Histogramms beurteilen wir die Verteilung der Stichprobenwerte

- als *unimodal*, wenn das Histogramm einen deutlichen Höcker aufweist,
- als *bimodal*, wenn das Histogramm zwei deutliche Höcker hat oder
- als *multimodal*, wenn im Histogramm mehrere deutliche Höcker zu erkennen sind.

Abbildung 2.5 zeigt einige verschiedene Formen.

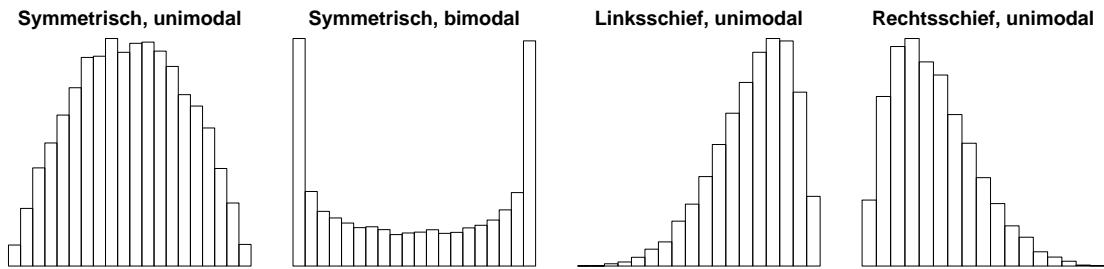


Abbildung 2.5: Einige Histogramme von Verteilungen mit verschiedener Form.

Hinweis (Histogramm versus ECDF). Beide Grafiken visualisieren die Verteilung der Stichprobenwerte. Während an der ECDF wichtige Zahlen abgelesen werden können, eignet sich das Histogramm besser, um Dichte und Form der Verteilung zu beschreiben. So betrachtet man am besten gleich beide Diagramme. Im Gegensatz zu ECDFs sind Histogramme nicht objektiv: Das Bild hängt von der Intervalleinteilung ab. Zudem geht durch die Kategorisierung Information verloren. Dies ist bei der ECDF nicht der Fall: Daraus können die Stichprobenwerte (bis auf ihre Reihenfolge im Datensatz) rekonstruiert werden.

Beispiel 2.7 (Befragung von Studierenden, Fortsetzung). Betrachten wir Histogramme¹ der Merkmale aus Beispiel 2.5.

R Code

```
# Eingabe
hist(wiso$Alter)
hist(wiso$GeschGrDoz)
hist(wiso$Kgroesse)
hist(wiso$MonMiete)
hist(wiso$MonMiete, breaks = c(0, 250, 500, 750, 2000))
hist(wiso$MonMiete, breaks = c(0, 250, 500, 750, 2000), freq = T)
```

Kommentare

- ‘Alter’: Die meisten Personen sind zwischen 18 und 22 Jahre alt, fast niemand ist älter als 26. Das Histogramm zeigt eine unimodale, stark rechtsschiefe Verteilung der Stichprobenwerte.
- ‘Geschätzte Grösse des Dozenten’: Die meisten Personen tippen auf einen Wert zwischen 170 und 180 cm. Das Histogramm weist auf eine unimodale, ungefähr symmetrische Verteilung der Werte hin.
- ‘Körpergrösse’: Ein grosser Teil der Körpergrössen liegt zwischen 165 und 185 cm. Das Histogramm zeigt eine symmetrische unimodale Verteilung der Stichprobenwerte.
- ‘Monatsmiete’: Hier betrachten wir drei Histogramme mit verschiedenen Intervalleinteilungen und Konventionen. Aus allen Bildern folgern wir, dass die meisten Personen weniger als 800 (bzw. 750) CHF pro Monat für das Wohnen ausgeben und dass die Werte rechtsschief verteilt sind. Während jedoch das linke Histogramm eine bimodale Verteilung zeigt, lassen die anderen auf eine unimodale Verteilung schliessen. Dies unterstreicht die Problematik, dass unterschiedliche Intervalleinteilungen zu deutlich unterschiedlichen Bildern führen können. Im Bild rechts wurde gegen die Regeln trotz verschieden breiten Intervallen Konvention 1 verwendet, was das Bild gegenüber der entsprechenden korrekten Darstellung nach Konvention 2 (mittleres Bild) deutlich verzerrt. Hinweis: Beim mittleren Bild zeigt die y-Achse nicht etwa die relativen Anteile, sondern ist so gewählt, dass die gesamte Fläche 1 beträgt.

2.2.2 Quantitative Beschreibung

Im Prinzip können die Zwischenprodukte von ECDF (Sprunghöhen bzw. relative Häufigkeiten der Ausprägungen) oder Histogramm (Häufigkeiten der Kategorien) als quantitative Beschreibung verwendet werden. Viel häufiger jedoch arbeitet man mit Kenngrössen wie Quantile, Lage- und Streuungsmasse, welche wesentliche Aspekte der Verteilung der X -Werte quantifizieren. Eine vollständige quantitative Beschreibung ist nur bei *stark diskret* verteilten Stichprobenwerten (viele gleiche Werte, beispielsweise die Anzahl Geschwister) sinnvoll. Dann kann das Merkmal wie ein kategorielles behandelt werden.

Quantile und Quartile

Eine gute Beschreibung der Verteilung einer Variable liefern die sogenannten *Quantile*. Sie können z. B. mit der empirischen Verteilungsfunktion F bestimmt werden: Für einen Wert $0 \leq \beta \leq 1$ ist das β -Quantil Q_β diejenige Schranke auf der x -Achse, an welcher F den Wert β auf der y -Achse annimmt bzw. überspringt. In Beispiel 2.5 haben wir auf diese Weise bereits einige Quantile bestimmt.

¹Histogramme werden in R mit der Funktion `hist` gezeichnet. Bei gleichlangen Intervallen wird automatisch Konvention 1 verwendet, sonst Konvention 2. Dies kann mit der Option `freq = TRUE/FALSE` geändert werden. Wenn nicht anders via Option `breaks` angefordert, wird die Intervalleinteilung von der Software gewählt.

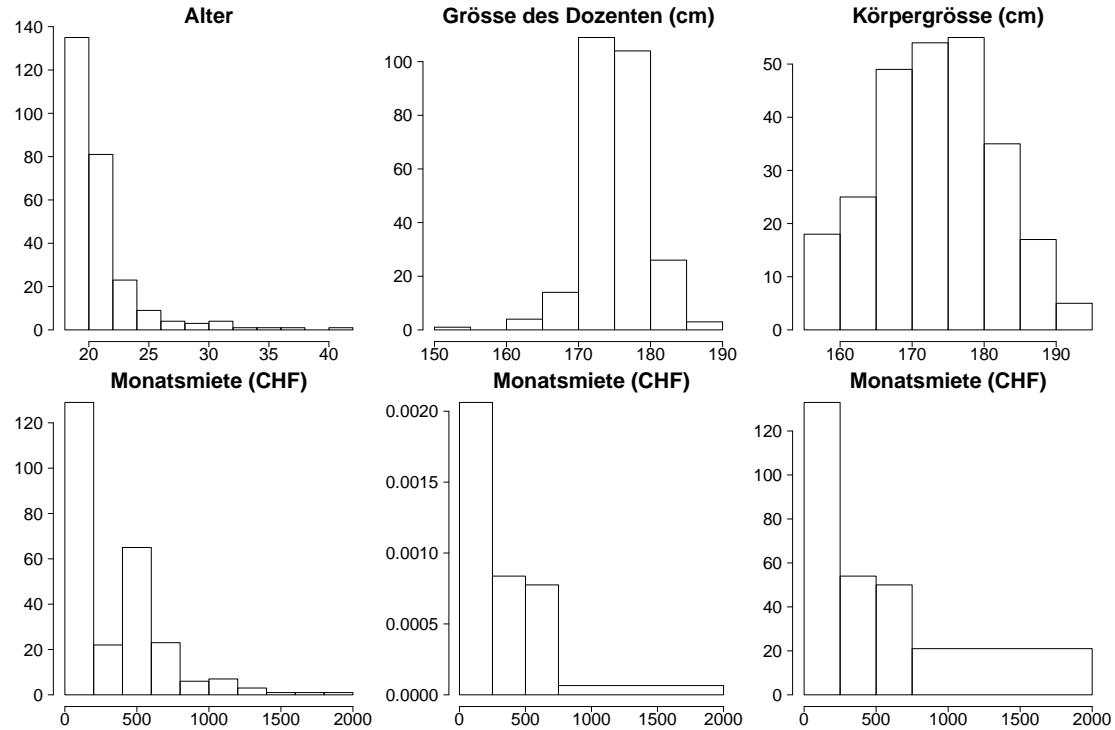


Abbildung 2.6: Histogramme der Merkmale in Beispiel 2.7 nach Konvention 1. Von ‘Monatsmiete’ wurde zudem eine Einteilung in verschiedene lange Intervalle gewählt, wo mit Konvention 2 gearbeitet werden muss (mittleres Bild unten). Konvention 1 vermittelt einen falschen Eindruck (Bild rechts unten).

Grob gesagt unterteilt das β -Quantil den Datensatz im Verhältnis β zu $1 - \beta$ in Beobachtungen mit kleinerem bzw. grösserem X -Wert. (Exakte Berechnungen überlassen wir der Software, da allgemeine Formeln unhandlich sind und von gewissen Konventionen abhängen.)

Beispiel 2.8 (Körpergrösse). Betrachten wir die Variable ‘Körpergrösse’ aus Beispiel 2.5. Anhand der ECDF (Abbildung 2.3 bzw. Abbildung 2.7 inkl. Hilfslinien) können wir abschätzen, dass die kleineren 80% der Personen höchstens 182 cm gross sind. Das 80%-Quantil von ‘Körpergrösse’ beträgt also 182 cm. ▲

Drei spezielle Quantile sind die sogenannten *Quartile*, die den Datensatz anhand der Variable X in vier (möglichst) gleich grosse Teile unterteilen:

- Erstes Quartil $Q_{0.25}$
- Median (zweites Quartil) $Q_{0.5}$
- Drittes Quartil $Q_{0.75}$

Zusammen mit dem *Minimum* (kleinster Wert bzw. Q_0) und dem *Maximum* (grösster Wert bzw. Q_1) bilden die drei Quartile den “five number summary” nach John W. Tukey. Er erlaubt z. B. folgende Angaben:

- Alle Stichprobenwerte liegen zwischen dem Minimum und dem Maximum.
- Etwa die Hälfte der Werte liegen zwischen erstem und drittem Quartil.
- Rund die Hälfte der Werte sind grösser/kleiner als der Median.
- Je rund 25% der Beobachtungen liegen unterhalb des ersten Quartils und über dem dritten Quartil.

Hinweis (Ordinale Merkmale). Quantile (und somit auch Quartile) können nicht nur für numerische, sondern im Prinzip auch für zahlenkodierte ordinale Merkmale angegeben werden, da für die Berechnungen lediglich die Ordnung der Stichprobenwerte relevant ist.

Beispiel 2.9 (Körpergrösse). Wir möchten die Verteilung der Variable ‘Körpergrösse’ aus Beispiel 1.1 mit dem “five number summary”¹ beschreiben.

R Code

```
# Eingabe
summary(wiso$Kgroesse)

# Ausgabe
Min. 1st Qu. Median     Mean 3rd Qu.      Max.    NA's
155.0   168.0   174.0   174.2   180.0   194.0      5.0
```

Kommentare

- *Minimum und Maximum:* Alle Werte liegen zwischen 155 und 194 cm.
- *Median:* Rund die Hälfte der Befragten ist kleiner bzw. grösser als 174 cm.
- *Erstes und drittes Quartil:* Etwa die Hälfte der Befragten ist zwischen 168 und 180 cm gross. Rund 25% sind kleiner als 168, rund 25% grösser als 180 cm.
- *Fehlende Werte:* Fünf Personen haben keine Angabe gemacht (NA = not available).

Abbildung 2.7 zeigt die entsprechende ECDF inklusive Hilfslinien² bei den Quartilen.

R Code

```
# Eingabe (benötigt library(Hmisc))
Ecdf(wiso$Kgroesse, datadensity = 'rug', q = c(0.25, 0.5, 0.75))
```

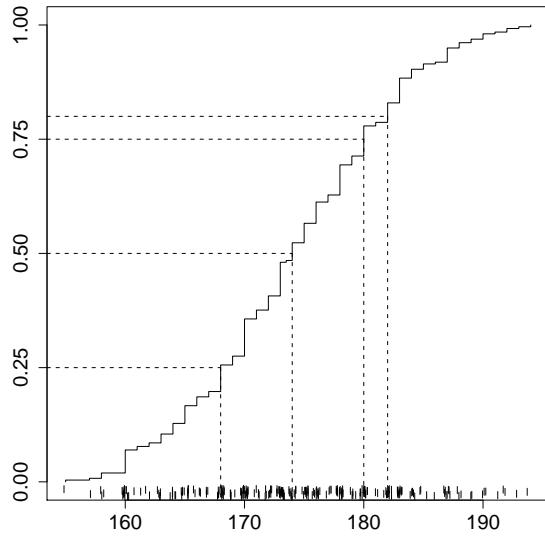


Abbildung 2.7: ECDF des Merkmals ‘Körpergrösse’ inkl. Hilfslinien bei den Quartilen und dem 80%-Quantil.

¹Dazu verwenden wir die R-Funktion `summary`. Beliebige Quantile sind in R via `quantile` verfügbar.

²Option `q` der R-Funktion `Ecdf`.

Lagemasse

Ein *Lagemas* ist eine Zahl, die “möglichst nah” an allen X -Werten liegt bzw. einen typischen Wert der X -Werte angibt. Die bekanntesten Lagemasse heissen Mittelwert und Median.

Der *Mittelwert (Durchschnitt)* ist das arithmetische Mittel der X -Werte, also

$$\bar{X} := \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{1}{n} \sum_{i=1}^n X_i.$$

Der *Median (Zentralwert)* entspricht dem 50%-Quantil bzw. dem zweiten Quartil. Entsprechend trennt er, so gut es geht, die kleinsten 50% der Werte von den grössten. Bei ungerader Stichprobengrösse n entspricht der Median dem mittleren der sortierten Werte, bei geradem n jeder Zahl zwischen den entsprechenden mittleren zwei Werten (z. B. deren Durchschnitt).

Beispiel 2.10 (Einfaches Zahlenbeispiel). Für die $n = 5$ Werte 20, 2, 7, 10, 1 beträgt der Mittelwert

$$\bar{X} = \frac{20 + 2 + 7 + 10 + 1}{5} = \frac{40}{5} = 8.$$

Der Median entspricht dem mittleren der sortierten Werte 1, 2, 7, 10, 20, also 7. Ohne Wert 20 wäre jede Zahl zwischen 2 und 7 ein Median, beispielsweise deren Durchschnitt 4.5.

Bestätigen wir die Ergebnisse mit der Software¹:

Eingabe
R Code

```
# Eingabe
x <- c(20, 2, 7, 10, 1)
summary(x)

# Ausgabe
Min. 1st Qu. Median Mean 3rd Qu. Max.
1       2       7     8     10      20
```



Hinweise

- *Verschiebungen und Skalierungen*: Werden alle Werte des Merkmals um eine fixe Zahl verschoben oder mit einem Faktor multipliziert (eine sogenannte *Skalierung*, beispielsweise eine Änderung in der Einheit), so äussert sich das entsprechend im Mittelwert und Median.
- *Ausreisser, Robustheit, getrimmter Mittelwert*: Der Mittelwert reagiert empfindlich auf *Ausreisser* in den Daten. Dabei verstehen wir Werte, die entweder falsch eingetragen wurden (z. B. durch falsches Setzen von Dezimalpunkten, unsinnige Angaben auf Fragebögen) oder tatsächlich ungewöhnlich gross oder klein sind. Ein einziger extremer Wert kann dafür sorgen, dass der Mittelwert von den meisten Werten sehr weit entfernt ist. Im Gegensatz dazu ist der Median *robust* gegenüber Ausreissern. Eine weitere robuste Alternative zum Mittelwert ist der *getrimmte* Mittelwert: Dieser entspricht dem arithmetischen Mittel ohne die grössten/kleinsten paar Prozent der Werte.

Ein Beispiel ist das monatliche Bruttoeinkommen pro Schweizer Haushalt: Der Mittelwert von rund 8'000 CHF scheint sehr hoch. Dies liegt daran, dass er durch einige Haushalte mit extrem grossen Einkommen deutlich beeinflusst wird. Der Median von ca. 6'500 CHF taugt besser als Wert für das typische Einkommen.

¹Die R-Funktion `summary` gibt u. a. Mittelwert und Median an. Alternativ könnten die beiden Funktionen `mean` und `median` verwendet werden.

Dass gerade Ökonomen lieber mit Mittelwerten als mit Medianen arbeiten, liegt vermutlich daran, dass man mit Mittelwerten leicht rechnen und sie gut extrapoliieren kann: Wenn man beispielsweise schätzt, dass Studierende, die nicht bei Angehörigen wohnen, monatlich und pro Person ca. 600 CHF Miete zahlen, und wenn man davon ausgeht, dass in der Agglomeration Bern ca. 3'500 solche Personen leben, dann beträgt ihr gesamtes Mietaufkommen ca. $3'500 \cdot 600 = 2'100'000$ CHF pro Monat.

- *Symmetrie*: Bei symmetrischer Verteilung entspricht der Median dem Mittelwert.
- *Binäre Merkmale*: Der Mittelwert einer binären (0-1)-Variable entspricht dem relativen Anteil der Ausprägung ‘1’. Der Median entspricht dem Modus, also der häufigeren Ausprägung.

Beispiel 2.11 (Körpergrösse, Fortsetzung). Wir möchten mit Mittelwert und Median einen Eindruck über die typische Körpergrösse der Studierenden gewinnen. Dabei arbeiten wir sowohl in der Originaleinheit “cm” als auch in “m”.

R Code

```
# Eingabe
summary(wiso$Kgroesse)

# Ausgabe
Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
155.0 168.0 174.0 174.2 180.0 194.0 5.0

# Eingabe: Alles in Metern
summary(wiso$Kgroesse/100)

# Ausgabe
Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
1.550 1.680 1.740 1.742 1.800 1.940 5.000
```

Kommentar: Mittelwert und Median sind hier praktisch identisch (Hinweis auf Symmetrie). Der typische Befragte ist etwa 174 cm bzw. 1.74 m gross. Wäre übrigens der erste Wert statt 1.76 m fälschlicherweise 176 m, so bliebe der Median unverändert, während der Mittelwert auf 2.42 m verfälscht würde. ▲

Streuungsmasse

Ein *Streuungsmass* ist eine Zahl, welche die typische Abweichung der X -Werte von ihrem “Zentrum” bzw. die typische Abweichung der X -Werte untereinander quantifiziert. Die wichtigsten Streuungsmasse heissen Spannweite, Interquartilabstand und Standardabweichung.

Die *Spannweite* ist der Abstand zwischen kleinstem und grösstem Stichprobenwert. Sie gibt also die Länge des Bereichs an, in dem alle Werte liegen.

Der *Interquartilabstand* entspricht der Differenz zwischen drittem und erstem Quartil, also

$$\text{IQR} := Q_{0.75} - Q_{0.25}.$$

Mit anderen Worten, es ist die Länge des Intervalls $[Q_{0.25}, Q_{0.75}]$, von dem wir wissen, dass es 50% aller Werte enthält.

Die (*Stichproben-*)*Standardabweichung*, in der Finanzwelt auch *Volatilität* genannt, ist definiert als

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

Diese Zahl gibt, grob gesagt, die typische Abweichung zum Mittelwert an. Die Kenngrösse innerhalb der Quadratwurzel ist die sogenannte (*Stichproben-*)*Varianz*.

Beispiel 2.12 (Einfaches Zahlenbeispiel, Fortsetzung). Betrachten wir wiederum die fünf Werte 20, 2, 7, 10 und 1 von Beispiel 2.10.

Alle Werte befinden sich zwischen 1 und 20. Die Spannweite beträgt somit $20 - 1 = 19$. Der IQR entspricht der Differenz 8 zwischen erstem Quartil (2) und drittem Quartil (10)¹.

R Code																	
# Eingabe																	
<pre>x <- c(20, 2, 7, 10, 1)</pre>																	
summary(x)																	
# Ausgabe																	
<table style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: left; width: 15%;">Min.</th> <th style="text-align: left; width: 15%;">1st Qu.</th> <th style="text-align: left; width: 15%;">Median</th> <th style="text-align: left; width: 15%;">Mean</th> <th style="text-align: left; width: 15%;">3rd Qu.</th> <th style="text-align: left; width: 15%;">Max.</th> </tr> </thead> <tbody> <tr> <td style="text-align: left;">1</td> <td style="text-align: left;">2</td> <td style="text-align: left;">7</td> <td style="text-align: left;">8</td> <td style="text-align: left;">10</td> <td style="text-align: left;">20</td> </tr> </tbody> </table>						Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	1	2	7	8	10	20
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.												
1	2	7	8	10	20												

Zur Berechnung der Varianz S^2 brauchen wir den in Beispiel 2.10 gefundenen Mittelwert 8:

$$S^2 = \frac{1}{5-1} ((20-8)^2 + (2-8)^2 + (7-8)^2 + (10-8)^2 + (1-8)^2) = 58.5.$$

Die Standardabweichung beträgt damit $S = \sqrt{58.5} \approx 7.65$. Mit R erhalten wir das gleiche²:

R Code	
var(x)	# Ergibt 58.5
sd(x)	# Ergibt 7.649



Hinweise

- *Verschiebungen und Skalierungen*: Werden alle Werte des Merkmals um einen fixen Betrag verschoben, so ändert sich deren Streuung nicht. Werden die Werte mit einem positiven Faktor multipliziert, so vervielfacht sich der Wert des betrachteten Streuungsmasses entsprechend.
- *Informativere Darstellungsweise*: Spannweite und IQR geben Längen von Bereichen an. Häufig wird stattdessen der Bereich selbst spezifiziert, also Minimum und Maximum bzw. erstes und drittes Quartil. Dies ist deutlich informativer. Zum einen lassen sich anhand Minimum und Maximum Ausreißer und offensichtlich falsche Stichprobenwerte entdecken. Zum anderen lässt sich mithilfe der Position des Medians zwischen erstem und drittem Quartil ein Eindruck über die Schiefe der Verteilung gewinnen: Bei einer symmetrischen Verteilung liegt der Median in der Mitte des ersten und dritten Quartils, bei einer rechtsschiefen Verteilung liegt er näher beim ersten Quartil und bei einer linksschiefen Verteilung entsprechend näher beim dritten Quartil.
- *Robustheit*: Von den genannten drei Streuungsmassen ist nur der IQR robust gegenüber Ausreißern.
- *Verbindung zu Lagemassen*: Median und Mittelwert unterscheiden sich um höchstens einmal die Standardabweichung.

Beispiel 2.13 (Körpergrösse, Fortsetzung). Wir haben in Beispiel 2.11 festgestellt, dass die typische Körpergrösse der Befragten 174 cm beträgt. Nun möchten wir mit Spannweite, IQR und Standardabweichung³ einen Eindruck über die Streuung der Werte erhalten. Dabei arbeiten wir wiederum sowohl in der Originaleinheit "cm" als auch in "m".

¹Die Berechnung der Quartile überlassen wir der R-Funktion `summary`.

²Die entsprechenden R-Funktionen heissen `var` und `sd`.

³Mit der Option `na.rm = TRUE` teilt man R-Funktionen wie `sd` mit, dass die fehlenden Werte ignoriert werden sollen.

R Code

```
# Eingabe
summary(wiso$Kgroesse)

# Ausgabe
Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
155.0 168.0 174.0 174.2 180.0 194.0 5.0

# Eingabe
sd(wiso$Kgroesse, na.rm = TRUE)      # Ergibt 8.1486

# Eingabe: Alles in Metern
summary(wiso$Kgroesse/100)

# Ausgabe
Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
1.550 1.680 1.740 1.742 1.800 1.940 5.000

# Eingabe
sd(wiso$Kgroesse/100, na.rm = TRUE)    # Ergibt 0.081486
```

Kommentare

- *Spannweite*: Die Stichprobenwerte liegen zwischen 155 und 194 cm, entsprechend ist die Spannweite $194 - 155 = 39$ cm bzw. 0.39 m. Alle Körpergrößen liegen also in einem Bereich der Länge 39 cm.
- *IQR*: Erstes und drittes Quartil betragen 168 und 180 cm, der IQR ist also $180 - 168 = 12$ cm bzw. 0.12 m. Rund 50% der Werte liegen damit in einem Bereich der Länge 12 cm. Der Median 174 liegt hier exakt zwischen den beiden Quartilen, ein weiterer Hinweis auf symmetrisch verteilte Werte.
- *Standardabweichung*: Die Standardabweichung beträgt 8.15 cm bzw. 0.0815 m. Somit beträgt die typische Distanz zum Mittelwert 8.15 cm.
- *Ausreisser*: Wäre der erste Wert statt 1.76 m fälschlicherweise 176 m, so bliebe der IQR unverändert, während die Spannweite $176 - 1.55 = 174.45$ m und die Standardabweichung 10.85 m betragen würden!



2.3 Datensätze

Zu jeder Datenanalyse gehört eine Beschreibung der Stichprobe. Dazu wird jedes relevante Merkmal univariat beschrieben. Zudem werden generelle Informationen wie beispielsweise die Herkunft der Daten, der Stichprobenumfang und die Bedeutung der Merkmale ausgewiesen.

Der Software-Output wird meist in aufbereiteter Form präsentiert: Grafiken werden verschönert und Zahlen zu übersichtlichen Tabellen zusammengefasst. Die Kenngrößen der numerischen Merkmale (z. B. Mittelwert, Standardabweichung, Quartile, Minimum und Maximum) werden dann sinnvollerweise in eine Tabelle geschrieben, jene der kategorialen (absolute und/oder relative Häufigkeiten der Ausprägungen) in eine andere. Stark diskrete numerische Variablen wie Häufigkeiten oder binäre Angaben können sowohl in der einen als auch der anderen Form präsentiert werden.

Üblicherweise werden die wichtigsten Zahlen auch im Fliesstext erwähnt.

Beispiel 2.14 (Befragung von Studierenden, Fortsetzung). Wir beschreiben nun die Stichprobe von Beispiel 1.1 quantitativ¹ und grafisch.

```
R Code
# Eingabe
summary(wiso)

# Ausgabe
Geschlecht      Alter      GebMonat      Herkunft      Kgroesse
M:147   Min.   :18.00   10   : 31   Bern       :108   Min.   :155.0
W:116   1st Qu.:20.00    6   : 28   Luzern     : 30   1st Qu.:168.0
          Median :20.00    3   : 27   Solothurn  : 29   Median :174.0
          Mean   :21.24    8   : 26   Aargau     : 16   Mean   :174.2
          3rd Qu.:22.00    4   : 23   Graubuenden:  9   3rd Qu.:180.0
          Max.   :42.00   (Other):127  (Other)    : 67   Max.   :194.0
          NA's    : 1        NA's     : 4   NA's     : 5.0

Kgewicht      MonMiete      Rauchen      ZufZiffer      AnzGeschw      GeschGrDoz
Min.   :45.00   Min.   : 0.0   0   :171   7   :70   Min.   :0.000   Min.   :150.0
1st Qu.:58.00   1st Qu.: 0.0   1   : 47   8   :41   1st Qu.:1.000   1st Qu.:175.0
Median :65.00   Median :110.0   2   : 43   3   :32   Median :1.000   Median :176.0
Mean   :64.92   Mean   :304.6  NA's: 2   6   :28   Mean   :1.554   Mean   :176.3
3rd Qu.:70.50   3rd Qu.:550.0           4   :25   3rd Qu.:2.000   3rd Qu.:179.0
Max.   :103.00  Max.   :2000.0          (Other):66  Max.   :6.000   Max.   :187.0
NA's   :12.00   NA's   : 5.0           NA's   : 1   NA's   :3.000   NA's   : 2.0

# Eingabe (benötigt library(Hmisc))
Ecdf(wiso, datadensity = 'rug', q = c(0.25, 0.5, 0.75), n.unique = 2)
hist(wiso, rugs = TRUE)
```

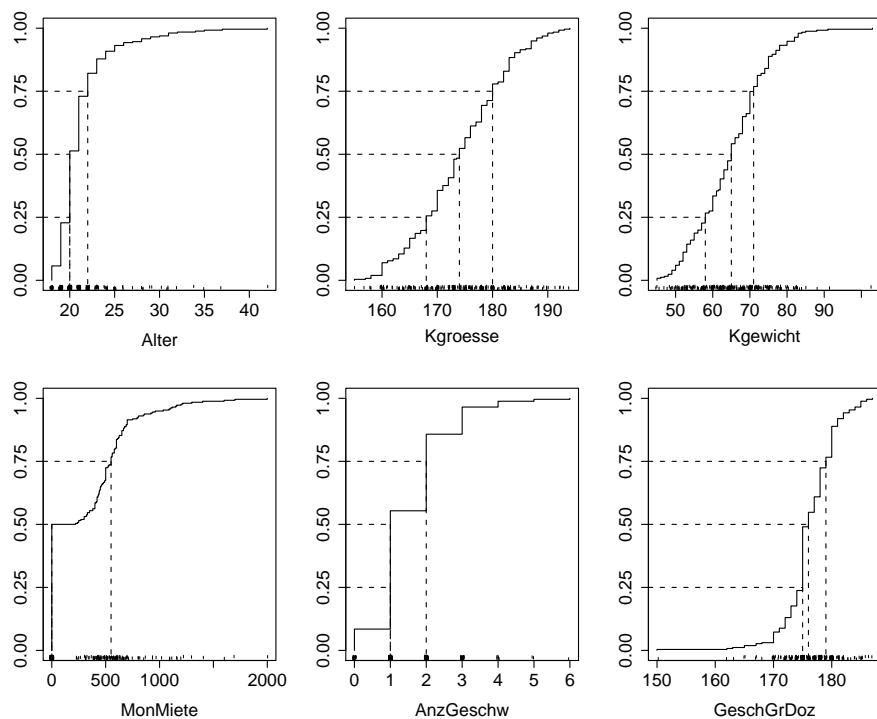


Abbildung 2.8: ECDFs für Beispiel 2.14.

¹Die R-Funktion `summary` für quantitative Beschreibungen (Standardabweichungen via `sd`), die Funktionen `ECDF` und `hist` aus dem `Hmisc`-Paket für grafische.

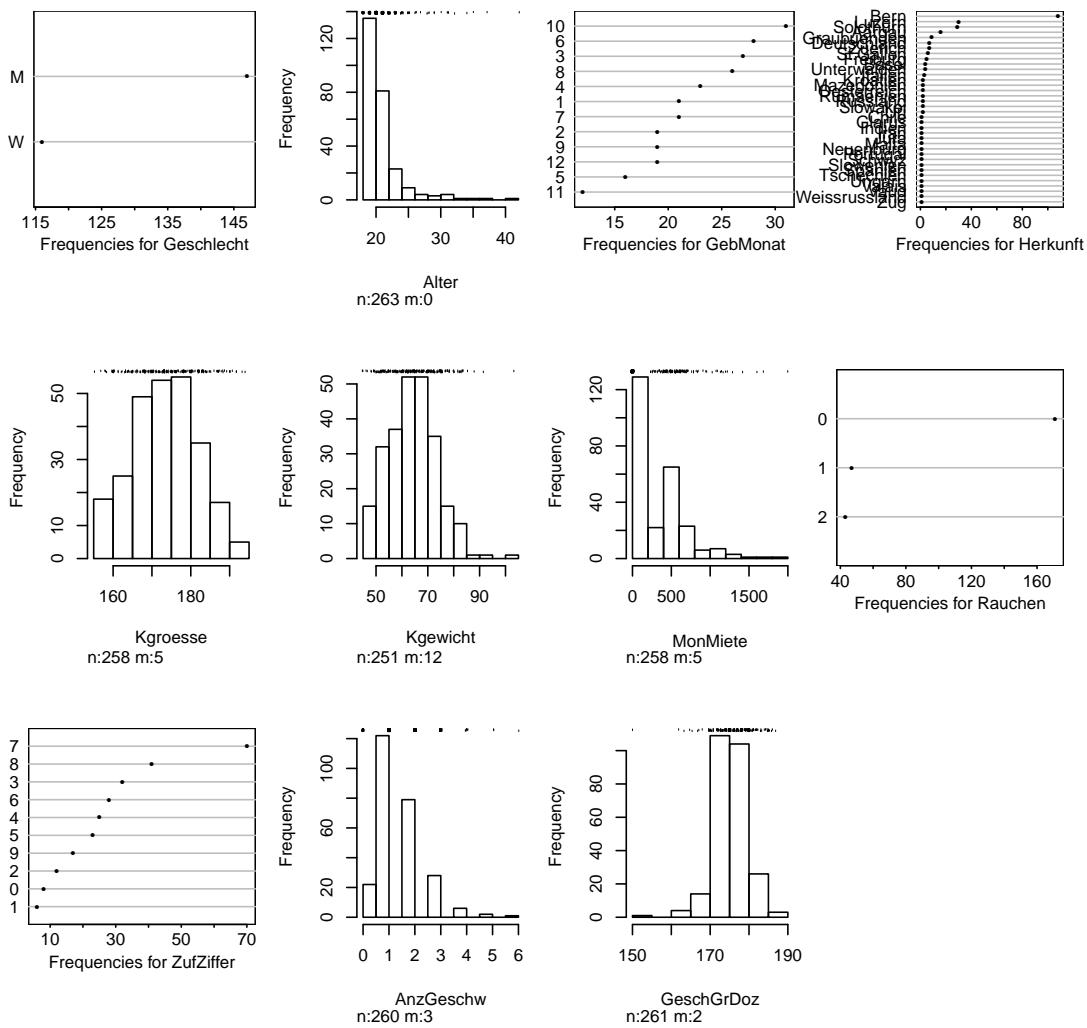


Abbildung 2.9: Histogramme/Punktediagramme der Variablen in Beispiel 2.14.

In einem Bericht über monatliche Mietausgaben von Studierenden könnte also beispielsweise stehen:

“Im Jahr 2003 wurden 263 Studierende zu ihren monatlichen Mietausgaben befragt. Die Stichprobe setzt sich zusammen aus 56% Männern und 44% Frauen, das Durchschnittsalter der Befragten beträgt 21.2 Jahre (Median 20, Spannweite 18 – 42), siehe die folgenden zwei Tabellen sowie Abbildungen 2.8 und 2.9 für die komplette Beschreibung der Stichprobe. Die Verteilung der monatlichen Mietausgaben der 258 gültigen Antworten sieht folgendermassen aus: [...]”

Merkmal	n	Mean	Std	Median	$Q_{0.25}$	$Q_{0.75}$	Min	Max
Alter in Jahren	262	21.20	3.14	20	20	220.0	18	42
Körpergrösse in cm	258	174.00	8.15	174	168	180.0	155	194
Körpergewicht in kg	251	64.90	9.46	65	58	70.5	45	103
Monatsmiete in CHF	258	305.00	367.10	110	0	550.0	0	2000
Anzahl Geschwister	260	1.55	0.97	1	1	2.0	0	6
Grösse des Dozenten	261	176.00	4.25	176	175	179.0	150	187

Merkmal	Ausprägung	Anzahl	%
Geschlecht	M	147	55.9
	W	116	44.1
Geburtsmonat	Januar	21	8.0
	Februar	19	7.3
	März	27	10.3
	April	23	8.8
	Mai	16	6.1
	Juni	28	10.7
	Juli	21	8.0
	August	26	9.9
	September	19	7.3
	Okttober	31	11.8
	November	12	4.6
	Dezember	19	7.3
	NA	1	
Herkunft	Bern	108	41.7
	Luzern	30	11.6
	Solothurn	29	11.2
	Aargau	16	6.2
	Graubünden	9	3.5
	Andere	67	25.9
	NA	4	
Rauchen	nein	171	65.5
	gelegentlich	47	18.0
	regelmässig	43	16.5
	NA	2	
Zufallsziffer	0	8	3.1
	1	6	2.3
	2	12	4.6
	3	32	12.2
	4	25	9.5
	5	23	8.8
	6	28	10.7
	7	70	26.7
	8	41	15.6
	9	17	6.5
	NA	1	

▲

Beispiel 2.15 (Highschool). Nun betrachten wir Daten von 316 Junior High School SchülerInnen an zwei städtischen Schulen in den USA. Der Datensatz enthält folgende fünf Merkmale:

school	Schule (zahlenkodiert als 0 und 1)
male	Männlich (ja = 1, nein = 0)
math	Standardisierte Leistung in Mathematik
langarts	Standardisierte Leistung in Language Arts (Lesen, Schreiben, Sprechen...)
daysabs	Absenztage im Semester (kategorisiert in 0 – 1, 2 – 5, > 5 Tage)

Um einen Eindruck über die Daten zu gewinnen, werfen wir einen Blick auf vier Zeilen dieses Datensatzes.

	R Code				
	school	male	math	langarts	daysabs
158	0	1	43.011	34.440	[0,1]
159	0	1	51.585	43.567	[0,1]
160	1	0	39.557	50.528	(1,5]
161	1	1	53.714	1.007	[0,1]

Dann beschreiben wir die Stichprobe univariat.

R Code

```
# Eingabe (benötigt library(Hmisc))
summary(highschool)
hist(highschool, ruggs = TRUE, n.unique = 2)
Ecdf(highschool, datadensity = 'rug', q = c(0.25, 0.5, 0.75))
```

Ausgabe

	school	male	math	langarts	daysabs
Min.	:0.000	Min. :0.000	Min. : 1.01	Min. : 1.01	[0,1] :108
1st Qu.	:0.000	1st Qu.:0.000	1st Qu.:37.73	1st Qu.:40.15	(1,5] : 99
Median	:0.000	Median :0.000	Median :48.94	Median :50.00	(5,50]:109
Mean	:0.497	Mean :0.487	Mean :48.75	Mean :50.06	
3rd Qu.	:1.000	3rd Qu.:1.000	3rd Qu.:61.04	3rd Qu.:61.04	
Max.	:1.000	Max. :1.000	Max. :98.99	Max. :98.99	

Kommentare: 49.7% der SchülerInnen sind aus Schule 1; 48.7% sind männlich. Die mittlere Leistung in Mathematik beträgt 48.75 Punkte (Median 48.94, Spannweite 1.01 – 98.99), jene in Sprache 50.06 (50.00, 1.01 – 98.99). 34% der SchülerInnen waren höchstens einen Tag abwesend, 31% zwischen zwei und fünf Tagen und 34% länger als fünf Tage. Abbildung 2.10 zeigt Histogramme bzw. ein Punktediagramm der Variablen, Abbildung 2.11 ECDFs der beiden numerischen Merkmale ‘Mathematik’ und ‘Language Arts’.

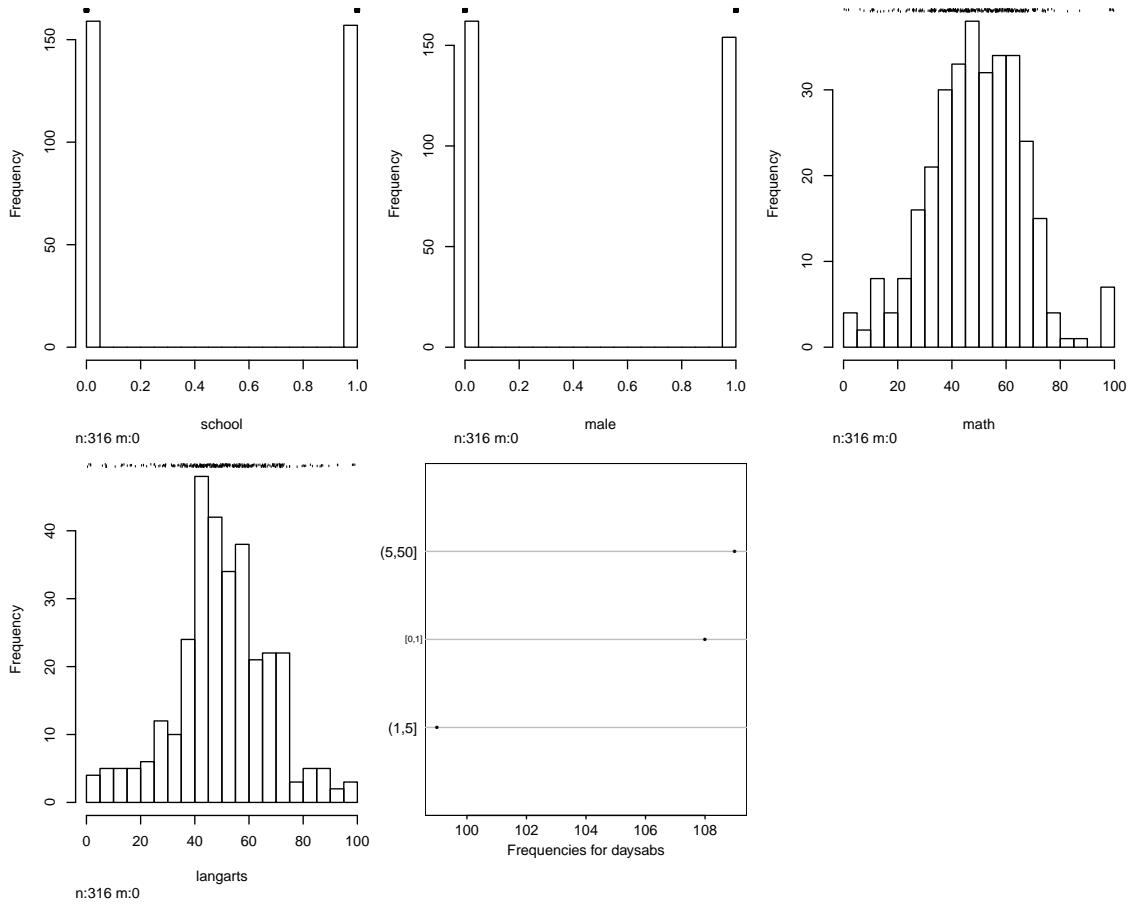


Abbildung 2.10: Univariate grafische Darstellung aller Variablen in Beispiel 2.15.

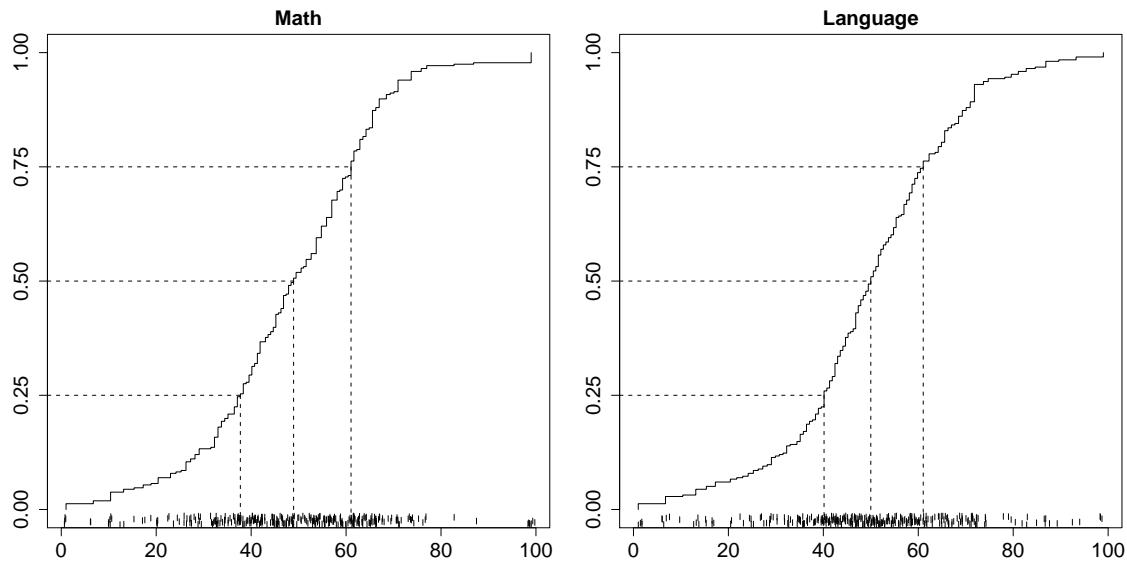


Abbildung 2.11: ECDFs von ‘Mathematik’ und ‘Language Arts’ in Beispiel 2.15.

Beispiel 2.16 (Wohnungen). Hier betrachten wir einen Datensatz mit diversen Angaben zu 76 Mietwohnungen, die an einem bestimmten Tag unter “Immoclick” angeboten wurden.

Er enthält folgende sechs Variablen:

Preis	Miete in CHF pro Monat
Zimmer	Anzahl der Zimmer
Parkett	Parkettboden (ja = 1, nein = 0)
Balkon	Balkon (ja = 1, nein = 0)
Garten	Garten (ja = 1, nein = 0)
Angaben	Anzahl der zusätzlichen Angaben zu jeder einzelnen Wohnung

Um einen Eindruck über die Daten zu erhalten, betrachten wir die ersten sechs Zeilen.

R Code

```
# Eingabe
head(wohnungen)

# Ausgabe
  Preis Zimmer Parkett Balkon Garten Angaben
1 1325     3.5      0     1     0     2
2  617     1.0      0     1     0     5
3 1170     3.0      0     0     0     7
4  887     2.0      1     0     1     4
5 1093     2.0      1     0     0     4
6 1665     3.0      0     1     0     4
```

Dann beschreiben wir die Stichprobe univariat.

R Code

```
# Eingabe (benötigt library(Hmisc))
summary(wohnungen)
hist(wohnungen, rugs = TRUE, n.unique = 2)
Ecdf(wohnungen, datadensity = 'rug', q = c(0.25, 0.5, 0.75))
```

Ausgabe

	Preis	Zimmer	Parkett	Balkon	Garten	Angaben
Min.	: 555	Min. :1.00	Min. :0.00	Min. :0.00	Min. :0.00	Min. :0.0
1st Qu.	: 915	1st Qu.:2.00	1st Qu.:0.00	1st Qu.:0.00	1st Qu.:0.00	1st Qu.:3.0
Median	:1155	Median :3.00	Median :0.00	Median :0.00	Median :0.00	Median :4.0
Mean	:1239	Mean :2.86	Mean :0.28	Mean :0.41	Mean :0.15	Mean :3.9
3rd Qu.	:1415	3rd Qu.:3.50	3rd Qu.:1.00	3rd Qu.:1.00	3rd Qu.:0.00	3rd Qu.:5.0
Max.	:2275	Max. :5.50	Max. :1.00	Max. :1.00	Max. :1.00	Max. :9.0

Kommentare: Die mittlere Monatsmiete in der Stichprobe ($n = 76$) beträgt 1239 CHF (Median 1155, Spannweite 555 – 2275) bei einer mittleren Anzahl Zimmer von 2.86 (3, 1 – 5.5). 27.6% der Wohnungen sind mit Parkettboden, 40.8% mit Balkon und 14.5% mit Garten ausgestattet. Im Schnitt wurden 3.9 (4, 0 – 9) zusätzliche Angaben zu den Wohnungen gemacht. Abbildungen 2.12 und 2.13 zeigen die grafischen Darstellungen der Stichprobenwerte. Eine ausführliche Beschreibung ist durch folgende Tabelle gegeben:

Merkmal	Mean	Median	$Q_{0.25}$	$Q_{0.75}$	Min	Max
Mietpreis in CHF	1239	1155	915	1415	555	2275
Anzahl Zimmer	2.86	3	2	3.5	1	5.5
Parkett vorhanden	0.28	0	0	1	0	1
Balkon vorhanden	0.41	0	0	1	0	1
Garten vorhanden	0.15	0	0	0	0	1
Anzahl zusätzliche Angaben	3.90	4	3	5	0	9

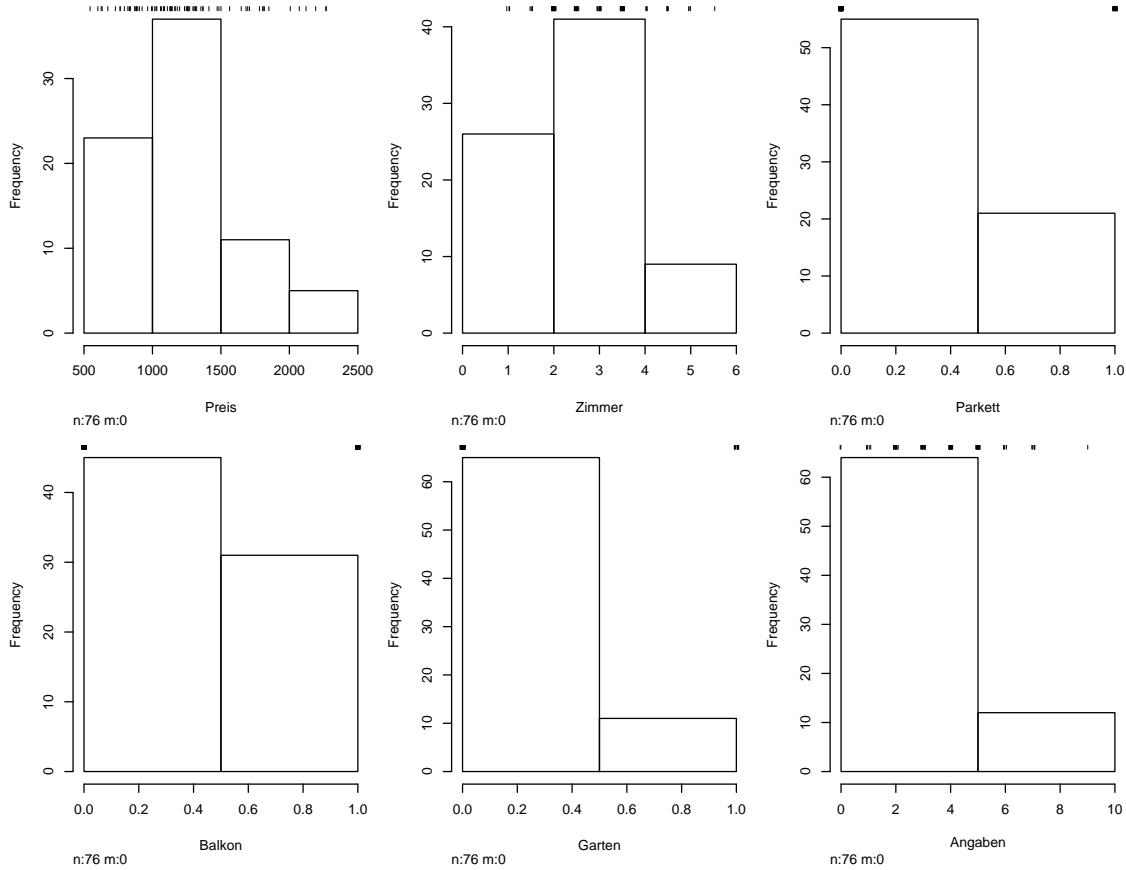


Abbildung 2.12: Univariate grafische Darstellung aller Variablen in Beispiel 2.16.

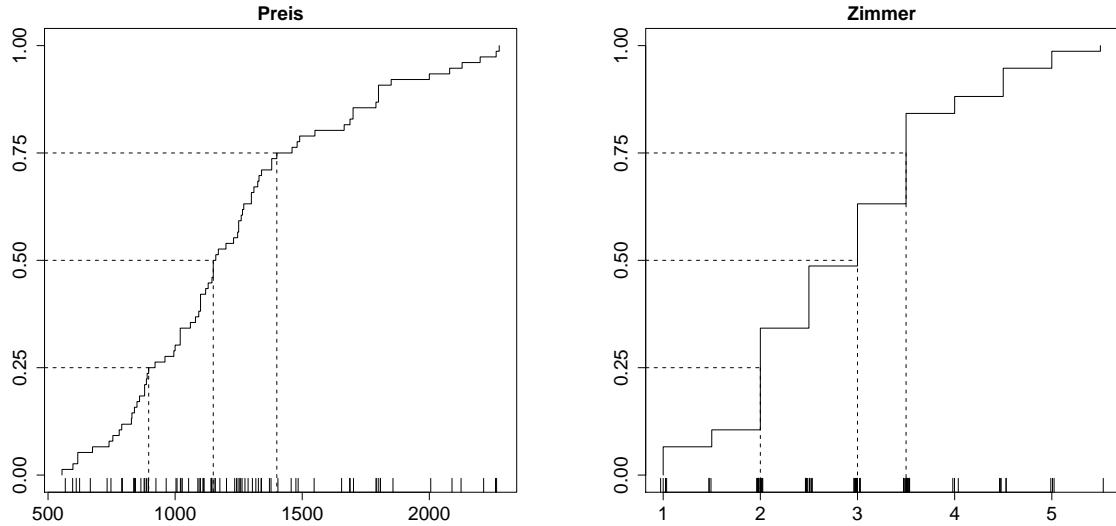


Abbildung 2.13: ECDFs von ‘Mietpreis’ und ‘Anzahl Zimmer’ in Beispiel 2.16.

2.4 Zusammenfassung

- Die univariate Beschreibung eines Merkmals läuft auf eine quantitative und/oder grafische Darstellung der Verteilung ihrer Stichprobenwerte hinaus. So können einfache statistische Fragestellungen beantwortet werden und beispielsweise offensichtliche Datenfehler oder Ausreisser identifiziert werden.
- Für numerische Merkmale werden andere Verfahren eingesetzt als für kategoriale.
- Wir haben gesehen, dass ein kategorielles Merkmal quantitativ durch Angabe der absoluten und/oder relativen Häufigkeiten der Kategorien beschrieben wird. Diese Häufigkeiten können beispielsweise als Balkendiagramm grafisch dargestellt werden.
- Ein numerisches Merkmal wird grafisch mit ECDF und/oder Histogramm dargestellt. Dabei haben wir die Vorzüge beider Möglichkeiten festgehalten. Eine vollständige quantitative Beschreibung ist in der Regel nicht sinnvoll, so dass man sich mithilfe von Kenngrößen wie Quantilen, Lagemassen (Mittelwert, Median) und Streuungsmassen (Spannweite, IQR, Standardabweichung) auf wichtige Aspekte der Verteilung der Stichprobenwerte konzentriert.
- Im Rahmen jeder Datenanalyse wird die Stichprobe beschrieben, indem jedes relevante Merkmal univariat beschrieben wird.

Kapitel 3

Schliessende Statistik und Wahrscheinlichkeitsrechnung

Empirische Daten fasst man in der Regel als zufällig auf. Oft betrachtet man die beobachteten Objekte (z. B. Personen) nämlich als Zufallsstichprobe aus einer grösseren Grundgesamtheit (Population). Dabei ist die Stichprobe an sich gar nicht von Interesse. Vielmehr möchte man mit Hilfe der Stichprobe Rückschlüsse auf die Grundgesamtheit ziehen (siehe Abbildung 3.1). Durch Zufall werden unterschiedliche Stichproben aus der Population unterschiedliche Resultate liefern.

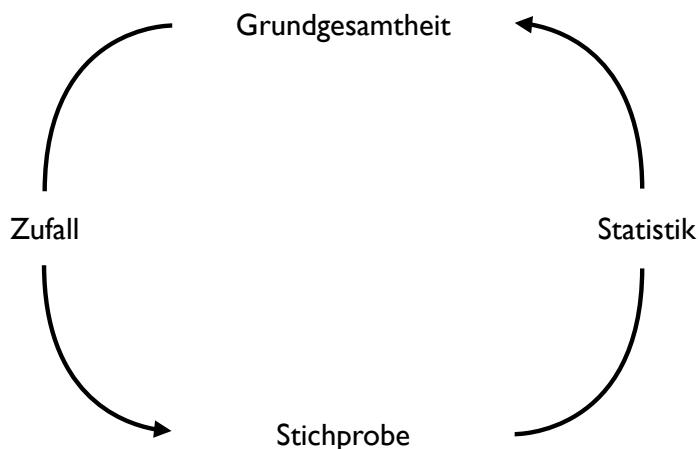


Abbildung 3.1: Diagramm zur Darstellung der Rolle der Statistik: Man benutzt die Information aus der Stichprobe, um auf die Grundgesamtheit zu schliessen, aus der die Stichprobe zufällig gezogen wurde.

Ein weiteres Beispiel sind Datenerhebungen von physikalischen oder chemischen Messungen, bei denen man zufällige Messfehler berücksichtigen muss.

Um Daten trotz solcher Fehlerquellen seriös auszuwerten, werden Werkzeuge der *schliessenden Statistik (Inferenzstatistik)* eingesetzt:

- *Schätzer*: Mit der Stichprobe berechnet man einen Schätzwert für einen unbekannten Parameter θ .
- *Konfidenzintervalle*: Anhand der Stichprobe berechnet man ein Intervall, in welchem θ mit einer gewis-

sen Sicherheit liegen soll.

- *Tests:* Mit Hilfe der Stichprobe soll nachgewiesen werden, dass ein augenscheinlicher Effekt, zum Beispiel die Wirkung eines neuen Medikaments, wirklich vorhanden ist und nicht durch reinen Zufall erklärt werden kann.

Zusammen mit den Grundlagen aus der Wahrscheinlichkeitsrechnung führen wir diese Werkzeuge zunächst für *univariate* Fragestellungen ein.

3.1 Zufallsvariablen und ihre Verteilungen

Ein zufälliger Vorgang (beispielsweise die zufällige Auswahl einer Person bei einer Befragung über das Einkommen) wird generell durch alle mit seinen Ergebnissen verbundenen *Wahrscheinlichkeiten* bzw. durch seine (*Wahrscheinlichkeits-*)*Verteilung* charakterisiert. Da man lieber mit Zahlen statt mit Ergebnissen von zufälligen Vorgängen arbeitet, werden diese meist durch sogenannte *Zufallsvariablen* quantifiziert, also in Zahlen umgewandelt.

Wir gehen nun auf diese Begriffe genauer ein.

3.1.1 Zufallsvariablen

Eine *Zufallsvariable*¹ X quantifiziert das Ergebnis eines zufälligen Vorgangs. Ein konkreter Wert von X heisst *Realisierung* von X . Die Menge \mathcal{X} aller möglichen Realisierungen wird *Wertebereich* von X genannt.

Beispiel 3.1 (Münzwurf). Eine Münze wird geworfen. Der Ausgang dieses zufälligen Vorgangs wird beispielsweise durch die Zufallsvariable

$$X := \begin{cases} 0 & \text{Zahl geworfen,} \\ 1 & \text{Kopf geworfen} \end{cases}$$

mit Wertebereich $\{0, 1\}$ quantifiziert. Werfen wir z. B. “Kopf”, so beträgt die Realisierung von X also 1. ▲

Beispiel 3.2 (Rauchen). Eine Person wird zufällig ausgewählt. Der Ausgang dieses zufälligen Vorgangs wird beispielsweise durch die Zufallsvariable

$$Y := \begin{cases} 0 & \text{Person ist NichtraucherIn,} \\ 1 & \text{Person ist GelegenheitsraucherIn,} \\ 2 & \text{Person ist regelmässige(r) RaucherIn} \end{cases}$$

mit Wertebereich $\{0, 1, 2\}$ quantifiziert. Ist die konkret ausgewählte Person z. B. eine Nichtraucherin, so beträgt die Realisierung der Zufallsvariable 0. ▲

Beispiel 3.3 (Einkommen). Eine erwachsene Person wird zufällig herausgepickt. Der Ausgang dieses zufälligen Vorgangs wird zum Beispiel durch die Zufallsvariable

$$Z := \text{Monatseinkommen in CHF}$$

mit Wertebereich $[0, \infty)$ quantifiziert. Verdient die konkret befragte Person z. B. 8600 CHF pro Monat, so lautet die Realisierung von Z entsprechend 8600. ▲

¹Der Begriff ist unglücklich gewählt: Eine Zufallsvariable ist ein fixer Mechanismus, also weder zufällig noch variabel.

Beispiel 3.4 (Würfelwurf). Ein Würfel wird geworfen. Der Ausgang dieses zufälligen Vorgangs wird sinnvollerweise durch die Zufallsvariable

$$X := \text{Augenzahl}$$

mit Wertebereich $\{1, 2, 3, 4, 5, 6\}$ quantifiziert. Werfen wir z. B. eine Sechs, so beträgt die Realisierung von X entsprechend 6. \blacktriangle

Beispiel 3.5 (Gewinne, Verluste und Renditen). Künftige Gewinne oder Verluste einer Firma bzw. Renditen einer Anlage sind wichtige Zufallsvariablen in der Wirtschaft. Sie quantifizieren das Ergebnis der relevanten Geschäftstätigkeit bzw. Börsentransaktionen. \blacktriangle

3.1.2 Wahrscheinlichkeiten

Mit den Ergebnissen eines zufälligen Vorgangs bzw. einer Zufallsvariable sind bekannte oder unbekannte Wahrscheinlichkeiten verknüpft.

Beispiel 3.6 (Münzwurf, Fortsetzung). Die Wahrscheinlichkeit, Kopf zu werfen, beträgt

$$P(\text{Kopf geworfen}) = P(X = 1) = 0.5.$$



Beispiel 3.7 (Rauchen, Fortsetzung). Der unbekannte Anteil p_o der NichtraucherInnen entspricht der Wahrscheinlichkeit, eine(n) NichtraucherIn zufällig herauszupicken, also $P(Y = 0)$. \blacktriangle

Bei solchen Überlegungen können die üblichen Regeln für Wahrscheinlichkeiten eingesetzt werden:

Regeln für Wahrscheinlichkeiten im Kontext von Zufallsvariablen

1. Die Zufallsvariable X nimmt nie einen Wert ausserhalb des Wertebereichs \mathcal{X} an, also $P(X \notin \mathcal{X}) = 0$.
2. X nimmt immer einen Wert im Wertebereich an, also $P(X \in \mathcal{X}) = 1$.
3. Falls die Teilmengen A und B von \mathcal{X} keine gemeinsamen Werte haben, lassen sich deren Wahrscheinlichkeiten addieren, also

$$P(X \in A \text{ oder } X \in B) = P(X \in A) + P(X \in B).$$

Falls A und B gemeinsame Werte haben, gilt

$$P(X \in A \text{ oder } X \in B) = P(X \in A) + P(X \in B) - P(X \in A \text{ und } X \in B).$$

4. $P(X \notin A) = 1 - P(X \in A)$ (Gegenwahrscheinlichkeit).

Beispiel 3.8 (Einkommen, Fortsetzung). Angenommen, 20% der Erwachsenen verdienen ≤ 3000 CHF und 50% verdienen ≥ 5000 CHF pro Monat. Dann könnten wir für die Zufallsvariable Z aus Beispiel 3.3 beispielsweise folgende Aussagen machen:

- $P(Z < 0) = 0$
- $P(Z \geq 0) = 1$
- $P(Z > 3000) = 1 - P(Z \leq 3000) = 0.8$

- $P(Z \leq 5000) = 1 - P(Z > 5000) = 0.5$
- $P(Z \leq 3000 \text{ oder } Z > 5000) = P(Z \leq 3000) + P(Z > 5000) = 0.2 + 0.5 = 0.7$
- $P(3000 < Z \leq 5000) = 1 - P(Z \leq 3000 \text{ oder } Z > 5000) = 1 - 0.7 = 0.3$
- $P(Z > 3000 \text{ oder } Z > 5000) = P(Z > 3000) + P(Z > 5000) - P(Z > 5000) = 0.8$ ▲

Beispiel 3.9 (Würfelwurf, Fortsetzung). Beim Würfelwurf gilt

$$P(X = 1) = \dots = P(X = 6) = 1/6$$

und damit z. B.:

- $P(X \leq 6) = 1$
- $P(X > 6) = 0$
- $P(X = 6) = 1/6$
- $P(X \neq 6) = 1 - 1/6 = 5/6$
- $P(X \in \{5, 6\}) = 1/6 + 1/6 = 1/3$ ▲

3.1.3 (Wahrscheinlichkeits-)Verteilung

Eine Zufallsvariable X mit Wertebereich \mathcal{X} ist durch ihre *Verteilung* charakterisiert. Diese gibt für jede Teilmenge B von \mathcal{X} die Wahrscheinlichkeit an, dass X einen Wert in B annimmt, also $P(X \in B)$.

Beispiel 3.10 (Münzwurf, Fortsetzung). Der Wertebereich $\mathcal{X} = \{0, 1\}$ der Zufallsvariable X , die den Münzwurf quantifiziert, hat vier Teilmengen: $\{\}, \{0\}, \{1\}$ und $\{0, 1\}$. Aus Beispiel 3.6 und den Regeln zu Wahrscheinlichkeiten betragen deren Wahrscheinlichkeiten $P(X \in \{\}) = 0$, $P(X \in \{0\}) = 1/2$, $P(X \in \{1\}) = 1/2$ und $P(X \in \{0, 1\}) = 1$. Dies ist die Verteilung von X . ▲

Beispiel 3.11 (Würfelwurf, Fortsetzung). Der Wertebereich hat sechs Elemente und damit $2^6 = 64$ Teilmengen, beispielsweise $\{1\}$, $\{2\}$ oder $\{1, 2, 3\}$. Bereits in diesem einfachen Beispiel wäre es also sehr aufwändig, die Verteilung explizit hinzuschreiben. ▲

Beispiel 3.12 (Einkommen, Fortsetzung). Es gibt beliebig viele Teilmengen des Wertebereichs $[0, \infty)$, beispielsweise $[0, 1000]$, $[1001, 1004]$ etc. Hier lässt sich die Verteilung also nicht explizit aufschreiben. ▲

3.1.4 Zufallsvariablen in der Statistik

Bevor wir uns auf effizientere Charakterisierungen von Verteilungen konzentrieren, gehen wir auf drei Arten von Zufallsvariablen ein, die man in Zufallsstichproben antrifft.

Merkmale

Ein (zahlenkodiertes) Merkmal X quantifiziert das Ergebnis des zufälligen Vorgangs ‘Person zufällig aus der Population ausgewählt’. Die Verteilung dieser Zufallsvariable beschreibt die prozentuale Zusammensetzung der Population in Bezug auf das Merkmal. Sie ist meist unbekannt und muss aus den Daten geschätzt werden.

Beispiel 3.13 (Befragung von Studierenden, Fortsetzung). Fassen wir die Studierenden in Beispiel 1.1 als Zufallsstichprobe aus der Population aller Studierenden auf, so quantifiziert z. B. das Merkmal ‘Körpergewicht’ eine zufällig herausgepickte Person aus dieser Population. ▲

Beobachtungen eines Merkmals

Wird der oben beschriebene Vorgang mehrmals durchgeführt, so lassen sich dessen Ergebnisse durch die Zufallsvariablen X_1 (erste Person zufällig ausgewählt), X_2 (zweite Person zufällig ausgewählt), X_3 (dritte Person zufällig ausgewählt) etc. quantifizieren. Dies sind die Beobachtungen des Merkmals X . Die konkreten Stichprobenwerte sind Realisierungen davon.

Werden die Personen unabhängig voneinander ausgewählt, gelten die Beobachtungen als (stochastisch) *unabhängige* Zufallsvariablen. Zwei Zufallsvariablen X_1 und X_2 sind dann unabhängig, falls für alle Teilmengen A_1 und A_2 aus ihren Wertebereichen gilt, dass $P(X_1 \in A_1 \text{ und } X_2 \in A_2) = P(X_1 \in A_1) \cdot P(X_2 \in A_2)$. Werden die Personen unter den gleichen Bedingungen ausgewählt, sind die Beobachtungen *identisch verteilte* Zufallsvariablen. (Diese haben je die gleiche Verteilung.) Üblicherweise werden die Beobachtungen eines Merkmals sowohl als unabhängig als auch identisch verteilt aufgefasst – eine Eigenschaft, die man mit “i. i. d.” von engl. “independent and identically distributed” abkürzt. Ihre Verteilung entspricht dann jener des betrachteten Merkmals.

Beispiel 3.14 (Befragung von Studierenden, Fortsetzung). Die Stichprobenwerte 68, 62, 72, … von ‘Körpergewicht’ sind Realisierungen von i. i. d. Zufallsvariablen mit der gleichen Verteilung wie die Zufallsvariable ‘Körpergewicht’. ▲

Kenngrößen bzw. Schätzer

Eine Kenngröße (Mittelwert, Varianz, Median, relative Häufigkeit etc.) dient als Schätzer für den tatsächlichen (unbekannten, jedoch fixen) Wert θ in der Population und quantifiziert damit das Ergebnis des zufälligen Vorgangs “Zufallsstichprobe ausgewählt”. Auf der Verteilung dieser Zufallsvariable basiert die Berechnung von Konfidenzintervallen und Tests für den Parameter θ .

Der konkrete Wert in der Stichprobe ist eine Realisierung dieser Zufallsvariable und dient als Schätzwert für θ .

Beispiel 3.15 (Befragung von Studierenden, Fortsetzung). Der Stichprobenmittelwert \bar{X} der Variable ‘Körpergewicht’ ist ein Schätzer für das tatsächliche mittlere Körpergewicht μ in der Population. Die Realisierung 64.9 kg ist ein naheliegender Schätzwert für μ . ▲

3.1.5 Charakterisierung von Verteilungen

Wir haben festgehalten, dass die Verteilung einer Zufallsvariable X durch die Wahrscheinlichkeiten aller Teilmengen ihres Wertebereichs beschrieben wird. Dies ist sehr umständlich und oft sogar unmöglich, deshalb präsentieren wir in diesem Abschnitt effizientere Wege. Dabei unterscheiden wir zwischen diskreten und stetigen Verteilungen, je nachdem, ob X abzählbar viele verschiedene Werte annehmen kann oder nicht.

Wahrscheinlichkeitsfunktion

Die *Wahrscheinlichkeitsfunktion* f gibt für jede mögliche Ausprägung x_1, x_2, x_3, \dots einer diskret verteilten Zufallsvariable X (z. B. eines zahlenkodierten kategorialen Merkmals) deren Wahrscheinlichkeit an, formal

$$f(x_j) := P(X = x_j) := p_j, \quad j = 1, 2, \dots$$

Die (Wahrscheinlichkeits-)Gewichte (Anteile) p_j sind nicht negativ und summieren sich zu eins.

Die Wahrscheinlichkeit einer beliebigen Teilmenge des Wertebereichs $\{x_1, x_2, x_3, \dots\}$ kann bestimmt werden, indem man die Gewichte ihrer Elemente zusammenzählt. Damit charakterisiert die Wahrscheinlichkeitsfunktion die Verteilung komplett.

Beispiel 3.16 (Bernoulliverteilung). Die *Bernoulliverteilung* mit Erfolgswahrscheinlichkeit $0 \leq p \leq 1$, kurz $\text{Bern}(p)$, nimmt die Werte 0 und 1 mit Gewichten

$$\begin{aligned} f(0) = P(X = 0) &= 1 - p \quad \text{und} \\ f(1) = P(X = 1) &= p \end{aligned}$$

an. Sie beschreibt beispielsweise den ‘fairen’ Münzwurf ($p = 0.5$) oder die Verteilung einer binären Variable (p ist der relative Anteil der Ausprägung ‘1’). Das linke Bild von Abbildung 3.2 zeigt die Wahrscheinlichkeitsfunktion von $\text{Bern}(0.4)$ als Balkendiagramm dargestellt. ▲

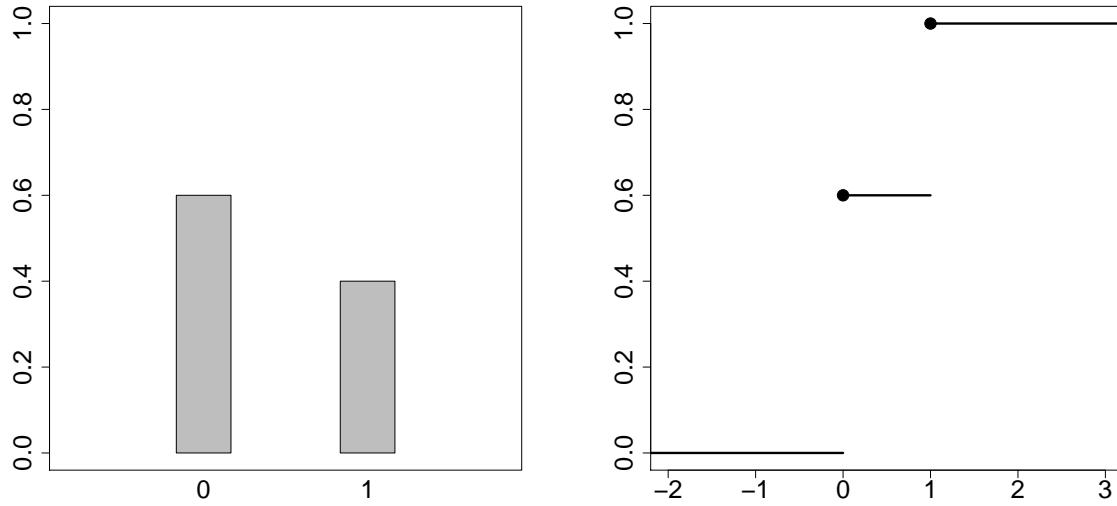


Abbildung 3.2: Wahrscheinlichkeitsfunktion (links) und Verteilungsfunktion (rechts) von $\text{Bern}(0.4)$.

Beispiel 3.17 (Diskrete Uniformverteilung). Betrachten wir eine Zufallsvariable X mit Werten in der Menge $\{x_1, \dots, x_L\}$. Weisen alle das gleiche Gewicht

$$f(x_j) = P(X = x_j) = 1/L$$

auf, so spricht man von der diskreten Uniformverteilung auf $\{x_1, \dots, x_L\}$. Typische Beispiele sind der Münzwurf (zwei verschiedene Ausprägungen mit je einem Gewicht von 0.5) und der Würfelwurf (sechs verschiedene Ausprägungen mit je einem Gewicht von 1/6).

Wahrscheinlichkeiten wie in Beispiel 3.9 (Würfelwurf) erhält man beispielsweise auch mit der Wahrscheinlichkeitsfunktion

$$f(x_j) = 1/6, \quad x_j = 1, \dots, 6.$$

- $P(X \leq 6) = P(X \in \{1, 2, 3, 4, 5, 6\}) = f(1) + \dots + f(6) = 1/6 + \dots + 1/6 = 1$
- $P(X = 6) = f(6) = 1/6$
- $P(X \neq 6) = P(X \in \{1, 2, 3, 4, 5\}) = f(1) + \dots + f(5) = 5/6$
- $P(X \in \{5, 6\}) = f(5) + f(6) = 2/6 = 1/3$

Dabei dürfen natürlich auch die Regeln zu Wahrscheinlichkeiten eingesetzt werden. Beispielsweise ist $P(X \neq 6) = 1 - P(X = 6) = 1 - f(6) = 1 - 1/6 = 5/6$.

Das linke Bild von Abbildung 3.3 illustriert die Wahrscheinlichkeitsfunktion des Würfelwurfs. ▲

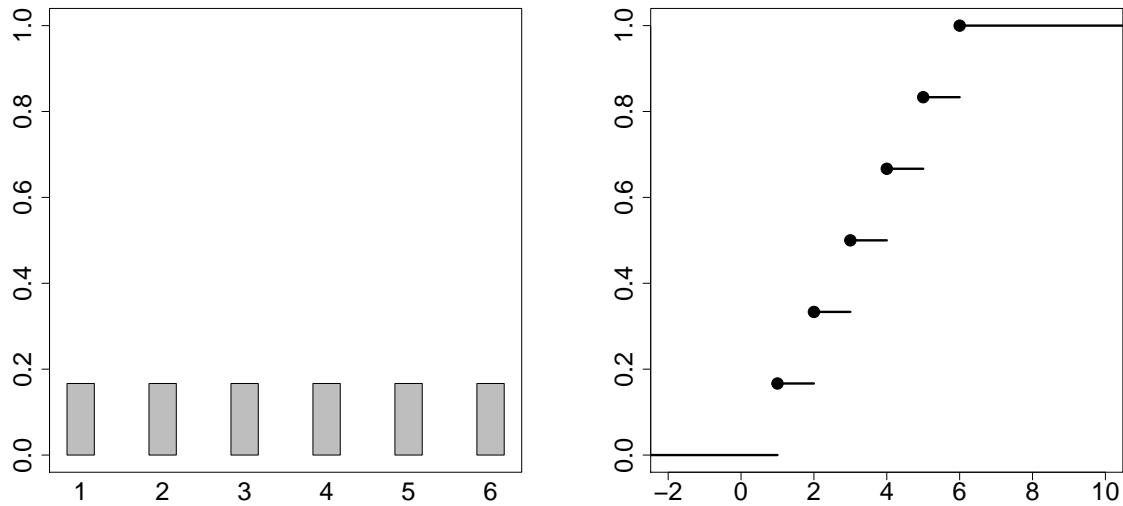


Abbildung 3.3: Wahrscheinlichkeitsfunktion (links) und Verteilungsfunktion (rechts) des Würfelwurfs.

Wahrscheinlichkeitsfunktion und relative Häufigkeiten Mit wachsendem Stichprobenumfang streben relative Häufigkeiten einer diskret verteilten Variable X gegen die tatsächlichen relativen Anteile (bzw. die Wahrscheinlichkeitsfunktion) von X in der Population. Diese sind meist unbekannt und werden durch die relativen Häufigkeiten in der Stichprobe geschätzt. Abbildung 3.4 zeigt diese Verbindung zwischen Praxis und Theorie anhand des Münzwurfs. Wahrscheinlichkeiten können damit als relative Häufigkeiten aufgefasst werden.

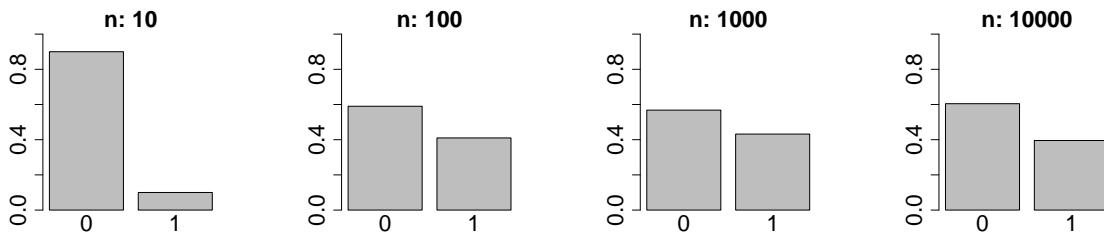


Abbildung 3.4: Eine ‘unfaire’ Münze ergebe ‘Kopf’ mit Wahrscheinlichkeit 0.4. Wir werfen sie $n = 10, 100, 1'000$ und $10'000$ mal und zeichnen je ein Balkendiagramm der relativen Häufigkeiten von ‘Kopf’ (Ausprägung 1) und ‘Zahl’ (0). Je mehr Würfe, je besser stimmen die relativen Häufigkeiten *im Schnitt* mit den tatsächlichen Werten 0.4 und 0.6 überein.

Dichtefunktion

Kann die Zufallsvariable X überabzählbar viele verschiedene Werte annehmen (z. B. ‘Körpergrösse’ bei beliebiger Messgenauigkeit), ist es nicht sinnvoll, mit der Wahrscheinlichkeitsfunktion zu arbeiten, da die Wahrscheinlichkeiten einzelner Werte stets null betragen. Es ist jedoch oft möglich, die Wahrscheinlichkeit, dass eine Zufallsvariable X einen Wert in einem Intervall $[a, b]$ annimmt, als Integral über eine *Dichtefunktion* (kurz: *Dichte*) darzustellen. Mithilfe der Regeln für Wahrscheinlichkeiten lassen sich daraus sämtliche Wahrscheinlichkeiten und damit die Verteilung von X finden. Da Integration aufwändig ist, führen wir hier lediglich das Konzept der Dichtefunktion ein und verschieben konkrete Berechnungen von Wahrscheinlichkeiten für stetige Verteilungen auf den nächsten Abschnitt.

X ist verteilt nach einer Dichtefunktion f , wenn für beliebige Intervalle $[a, b]$, $a < b$, gilt:

$$P(X \in [a, b]) = \int_a^b f(x)dx.$$

Aus der Definition folgen die Aussagen:

- Wahrscheinlichkeiten entsprechen Flächen unter der Dichtefunktion. Die gesamte Fläche unter der Dichtefunktion beträgt demnach eins.
- X neigt zu Werten mit hoher Dichte: Ein Wert von etwa x_1 kommt ungefähr $f(x_1)/f(x_2)$ mal so häufig vor wie ein Wert von etwa x_2 .
- Die Dichtefunktion zeigt die Form der Verteilung (u. a. Schiefe, Anzahl Höcker).
- Achtung: $f(x) \neq P(X = x)$ (Letztere Wahrscheinlichkeit ist stets null.)

Beispiel 3.18 (Grafische Darstellung von Wahrscheinlichkeiten). Bezeichnen wir mit X das Ergebnis eines Intelligenztests einer zufällig ausgewählten Person. Solche Tests sind häufig so konstruiert, dass die Dichtefunktion f von X “glockenförmig” um 100 ist. Das linke Bild in Abbildung 3.5 zeigt den Anteil der Personen mit mindestens 120 Punkten, also

$$P(X \geq 120) = \int_{120}^{\infty} f(x)dx,$$

als schraffierte Fläche unter f , das rechte Bild entsprechend deren Anteil zwischen 80 und 120 Punkten,

$$P(80 \leq X \leq 120) = \int_{80}^{120} f(x)dx.$$

▲

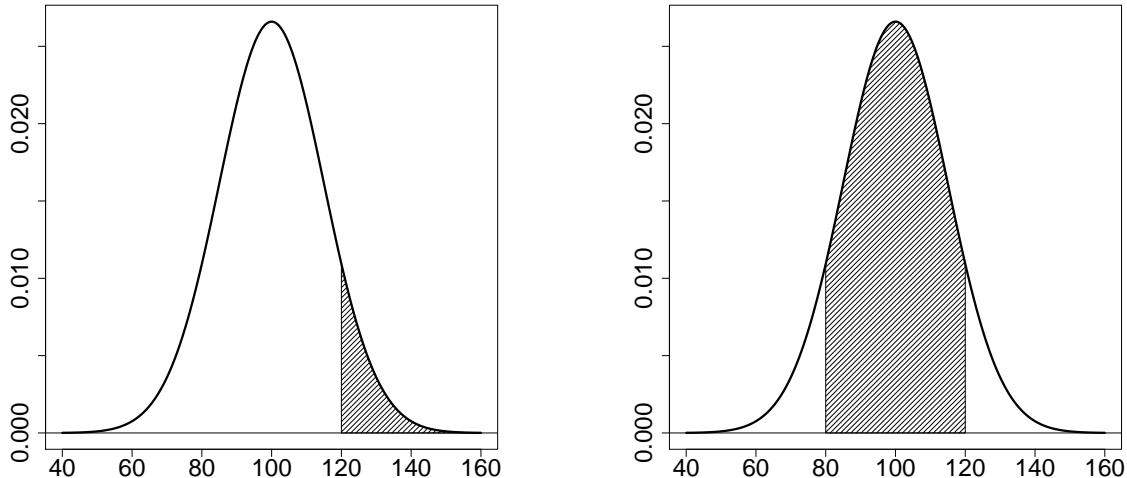


Abbildung 3.5: Anteile der Personen mit Punkten ab 120 Punkten (linkes Bild) und Anteile der Personen zwischen 80 und 120 Punkten (rechtes Bild).

Beispiel 3.19 (Exponentialverteilung). Die Verteilung einer Zeitspanne (Zeit bis Kündigung eines Vertrags (Versicherungspolice), Zeit zwischen zwei Ereignissen (Telefonanrufen), Lebensdauer (eines Produkts)) oder auch von positiven Geldbeträgen wird oft gut durch die *Exponentialverteilung* mit Rate $\lambda > 0$, kurz $\text{Exp}(\lambda)$, beschrieben. Ihre Dichtefunktion ist für $x \geq 0$ definiert und lautet

$$f(x) = \lambda e^{-\lambda x}.$$

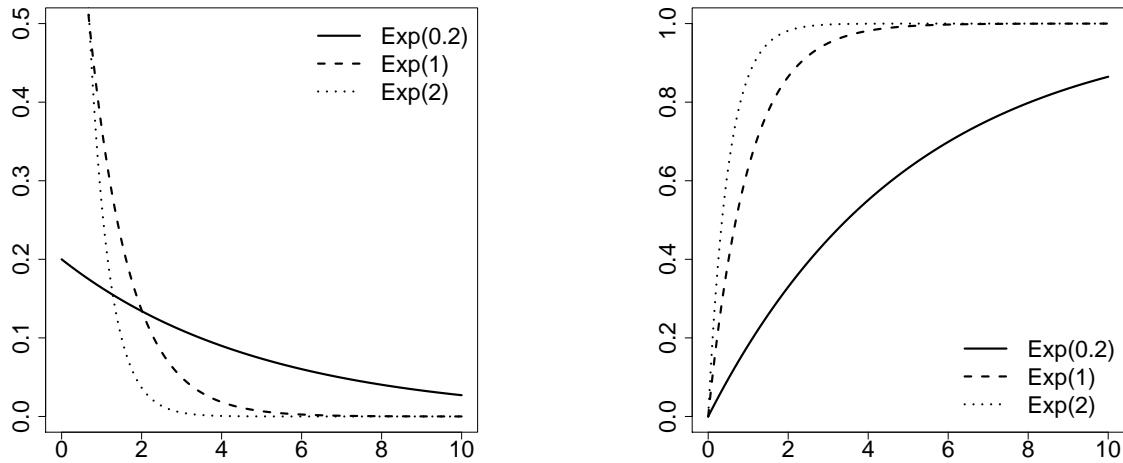


Abbildung 3.6: Dichtefunktionen (links) und Verteilungsfunktionen (rechts) von $\text{Exp}(0.2)$, $\text{Exp}(1)$ und $\text{Exp}(2)$.

Aufgrund dieser Definition sind Exponentialverteilungen rechtsschief mit einzigem Höcker (der Höhe λ) bei 0, siehe linkes Bild von Abbildung 3.6. ▲

Beispiel 3.20 (Uniformverteilung). Weist die Dichtefunktion überall (zwischen den endlichen Grenzen $a < b$) den gleichen Wert $1/(b-a)$ auf, so spricht man von der Uniformverteilung bzw. der Gleichverteilung zwischen a und b , kurz $\text{Unif}(a,b)$. Solche Zufallsvariablen spielen eine wichtige Rolle beim Erzeugen von Zufallszahlen mit vorgegebener Verteilung. Das linke Bild von Abbildung 3.7 zeigt die Dichtefunktion der Standarduniformverteilung, also von $\text{Unif}(0,1)$. ▲

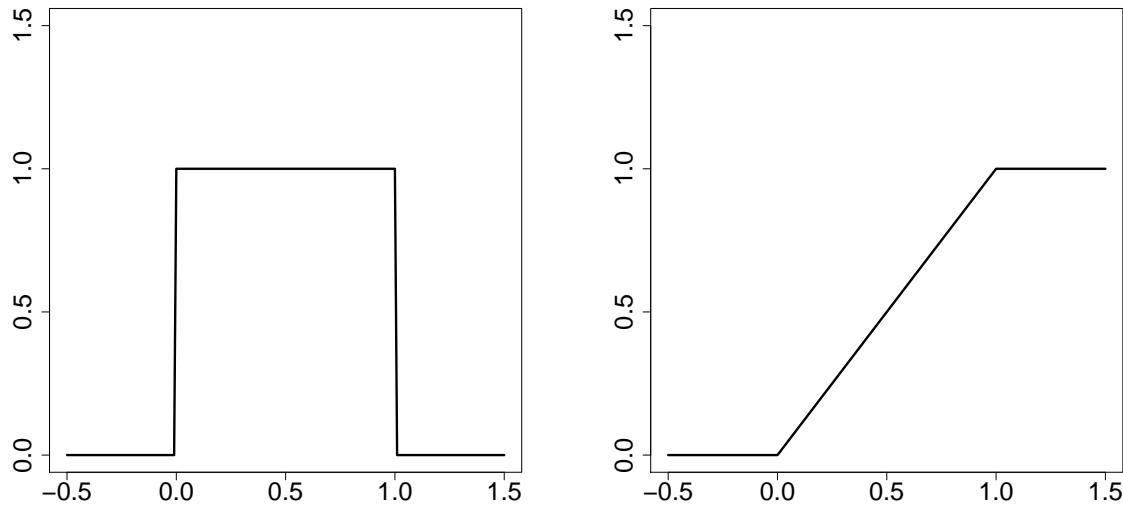


Abbildung 3.7: Dichtefunktion (links) und Verteilungsfunktion (rechts) von $\text{Unif}(0,1)$.

Dichtefunktion und Histogramm Mit wachsendem Stichprobenumfang strebt das Histogramm einer Variable X mit stetiger Verteilung gegen den Graphen der unbekannten tatsächlichen Dichtefunktion von X . Dies illustrieren wir in Abbildung 3.8 anhand der Exponentialverteilung.

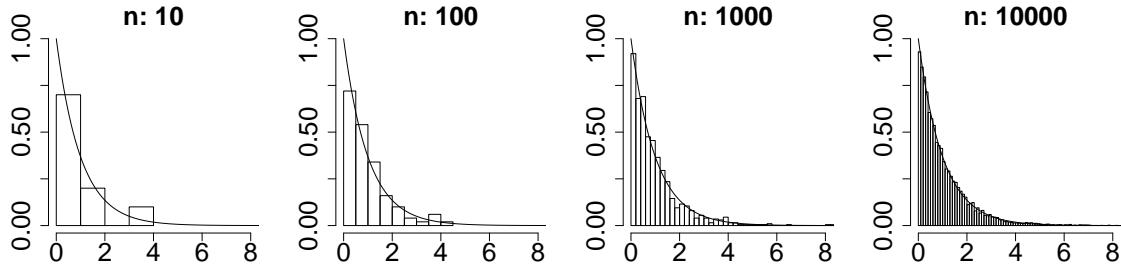


Abbildung 3.8: Wir erzeugen 10, 100, 1'000 und 10'000 Realisierungen von i. i. d. nach $\text{Exp}(1)$ -verteilten Zufallsvariablen und vergleichen deren Histogramm mit der tatsächlichen Dichtefunktion von $\text{Exp}(1)$. Je mehr Realisierungen, desto besser ist die Übereinstimmung im Schnitt.

Verteilungsfunktion

Wir haben gesehen, dass eine Verteilung durch ihre Wahrscheinlichkeits- bzw. Dichtefunktion komplett charakterisiert wird. Eine Alternative dazu ist die *Verteilungsfunktion*, die für eine Zufallsvariable X mit diskreter oder stetiger Verteilung als

$$F(r) := P(X \leq r), \quad r \text{ beliebig},$$

definiert ist. Sie gibt die Wahrscheinlichkeit an, dass die Zufallsvariable X höchstens den Wert r annimmt und stellt damit das theoretische Pendant zur ECDF dar. Daraus lassen sich mit den Regeln für Wahrscheinlichkeiten durch Summen- und Differenzenbildung die Wahrscheinlichkeiten aller Teilmengen des Wertebereichs (und damit die Verteilung) finden.

Beispielsweise gilt für $a \leq b$:

- $P(X \leq a) = F(a)$
- $P(X > b) = 1 - F(b)$
- $P(X \leq a \text{ oder } X > b) = F(a) + (1 - F(b))$
- $P(a < X \leq b) = 1 - P(X \leq a \text{ oder } X > b) = F(b) - F(a).$

Neben der vollständigen und effizienten Charakterisierung einer Verteilung werden Verteilungsfunktion und Wahrscheinlichkeits- bzw. Dichtefunktion auch verwendet, um Fragen der Art ‘Wie gross ist der Anteil von Personen mit einem IQ über 120?’ oder ‘Mit welcher Wahrscheinlichkeit passieren morgen zwei Unfälle am Eigerplatz?’ zu beantworten.

Verteilungsfunktion einer diskreten Verteilung Hat eine diskret verteilte Zufallsvariable X mit Werten in $\{x_1, x_2, \dots\}$ die Wahrscheinlichkeitsfunktion f , so ist ihre Verteilungsfunktion die Treppenfunktion

$$F(r) = \sum_{x_j \leq r} f(x_j).$$

Sie entspricht der ECDF eines Merkmals mit relativen Häufigkeiten wie $f(x_j)$ und gibt die Wahrscheinlichkeit eines Wertes bis und mit r an.

Beispiel 3.21 (Bernoulliverteilung). Die Verteilungsfunktion einer nach $\text{Bern}(0.4)$ -verteilten Zufallsvariable sieht aus wie die ECDF der Werte 0, 0, 0, 1, 1 und kann formal als

$$F(r) = \begin{cases} 0, & \text{wenn } r < 0, \\ 0.6, & \text{wenn } 0 \leq r < 1, \\ 1, & \text{sonst,} \end{cases}$$

geschrieben werden, siehe rechtes Bild von Abbildung 3.2. ▲

Beispiel 3.22 (Würfelwurf). Die Verteilungsfunktion F des Würfelwurfs sieht gleich aus wie die ECDF der Werte 1, 2, 3, 4, 5, 6, siehe rechtes Bild von Abbildung 3.3. Formal¹ ist

$$F(r) = \begin{cases} 0, & \text{wenn } r < 1, \\ \lfloor r \rfloor / 6, & \text{wenn } 1 \leq r \leq 6, \\ 1, & \text{sonst.} \end{cases}$$

Daraus finden wir (eher umständlich) die gleichen Wahrscheinlichkeiten wie in Beispiel 3.17:

- $P(X \leq 6) = F(6) = 1$
- $P(X = 6) = P(5 < X \leq 6) = F(6) - F(5) = 1 - 5/6 = 1/6$
- $P(X \neq 6) = P(X \leq 5) = F(5) = 5/6$
- $P(X \in \{5, 6\}) = P(4 < X \leq 6) = F(6) - F(4) = 1 - 4/6 = 1/3$

Es können im Prinzip beliebige Zahlen in F eingesetzt werden:

- $F(5.5) = \lfloor 5.5 \rfloor / 6 = 5/6$
- $F(-10) = 0$
- $F(100) = 1$ ▲

Verteilungsfunktion einer stetigen Verteilung Hat eine stetig verteilte Zufallsvariable die Dichtefunktion f , so lässt sich die Verteilungsfunktion als Integral darüber schreiben:

$$F(r) = \int_{-\infty}^r f(x) dx.$$

Die Verteilungsfunktion entspricht also der Stammfunktion der Dichte. Umgekehrt ist die Dichtefunktion die Ableitung der Verteilungsfunktion. Verfügt man über die Verteilungsfunktion, so lassen sich Wahrscheinlichkeiten auch bei stetig verteilten Zufallsvariablen ohne Integration finden. Da hier im Gegensatz zu den diskreten Verteilungen $P(X = r)$ stets 0 ist, gilt zudem beispielsweise $P(X \geq r) = P(X > r) = 1 - F(r)$.

Beispiel 3.23 (Exponentialverteilung). Bei der Exponentialverteilung mit Dichte

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

lässt sich die Verteilungsfunktion $F(r)$, $r \geq 0$, mit elementarer Analysis aus der Dichte bestimmen:

$$F(r) = \int_{-\infty}^r f(x) dx = \lambda \int_0^r e^{-\lambda x} dx = -e^{-\lambda x} \Big|_{x=0}^r = -(e^{-\lambda r} - 1) = 1 - e^{-\lambda r}.$$

Das rechte Bild in Abbildung 3.6 zeigt einige solche Verteilungsfunktionen.

Betrachten wir folgende konkrete Situation: Eine grosse Umfrage habe ergeben, dass die Verweildauer X (in Jahren) einer Schweizer Person bei ihrer ersten Arbeitsstelle etwa exponentialverteilt mit $\lambda = 0.2$ ist. Dies erlaubt uns – ohne über die Daten zu verfügen – z. B. folgende Aussagen zu machen:

- Rund 18% verlassen ihre erste Arbeitsstelle bereits im ersten Jahr, denn

$$P(X \leq 1) = F(1) = 1 - e^{-0.2} \approx 0.181.$$

¹Die speziellen Klammern \lfloor und \rfloor bedeuten “abrunden”.

- Rund 15% gehen im zweiten, 12% im dritten, 10% im vierten und 8% im fünften Jahr. Bspw. ist

$$P(2 \leq X \leq 3) = F(3) - F(2) = 1 - e^{-0.2 \cdot 3} - (1 - e^{-0.2 \cdot 2}) = e^{-0.2 \cdot 2} - e^{-0.2 \cdot 3} \approx 0.122.$$

- Rund 37% verweilen länger als 5 Jahre, denn

$$P(X > 5) = 1 - F(5) = e^{-0.2 \cdot 5} \approx 0.368.$$

▲

Beispiel 3.24 (Uniformverteilung). Die Dichtefunktion der Standarduniformverteilung beträgt

$$f(x) = \begin{cases} 1, & \text{wenn } 0 \leq x \leq 1, \\ 0, & \text{sonst.} \end{cases}$$

Entsprechend steigt die Verteilungsfunktion zwischen 0 und 1 gleichmässig von 0 bis 1 an (rechtes Bild von Abbildung 3.7), da in diesem Bereich

$$F(r) = \int_{-\infty}^r f(x) dx = \int_0^r 1 \cdot dx = x \Big|_{x=0}^r = r.$$

Beispielsweise ist die Wahrscheinlichkeit, dass eine so verteilte Zufallsvariable zwischen 0.3 und 0.8 liegt,

$$P(0.3 \leq X \leq 0.8) = F(0.8) - F(0.3) = 0.8 - 0.3 = 0.5$$

bzw. die Länge des Intervalls [0.3, 0.8]. ▲

Verteilungsfunktion und empirische Verteilungsfunktion Mit wachsendem Stichprobenumfang strebt die empirische Verteilungsfunktion einer numerischen Variable X gegen die unbekannte tatsächliche Verteilungsfunktion von X . Dieses Resultat stammt von W. I. Glivenko und F. Cantelli und wird aufgrund seiner grossen Relevanz in der Statistik manchmal *Hauptsatz der Statistik* genannt. Daraus folgt auch die bereits erwähnte Verbindung zwischen Wahrscheinlichkeitsfunktion und relativen Häufigkeiten sowie jene zwischen Dichtefunktion und Histogramm.

Wir illustrieren dieses Resultat in Abbildung 3.9 anhand der Exponentialverteilung.

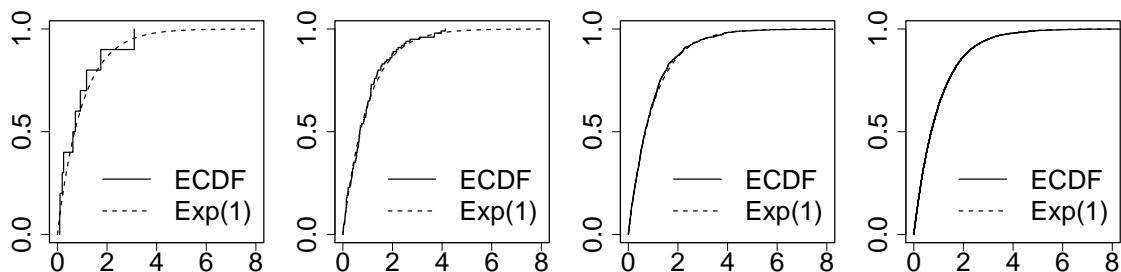


Abbildung 3.9: Wir erzeugen 10, 100, 1'000 und 10'000 Realisierungen von i. i. d. nach $\text{Exp}(1)$ -verteilten Zufallsvariablen und vergleichen deren ECDF mit der tatsächlichen Verteilungsfunktion von $\text{Exp}(1)$. Je mehr Realisierungen, desto besser ist die Übereinstimmung *im Schnitt*.

Quantilfunktion

Stichprobenquantile geben eine gute Beschreibung der Verteilung der Stichprobenwerte eines Merkmals. In Analogie dazu können sowohl stetige als auch diskrete Verteilungen von Zufallsvariablen durch die

sogenannte *Quantilfunktion* vollständig charakterisiert werden. Sie entspricht der (verallgemeinerten) Umkehrfunktion der Verteilungsfunktion F und wird deshalb oft mit F^{-1} gekennzeichnet. Für einen Anteil β zwischen 0 und 1 gibt $F^{-1}(\beta)$ das theoretische β -Quantil an, also den kleinsten Wert, bei dem die Verteilungsfunktion mindestens β beträgt bzw. diejenige Schranke, welche die Zufallsvariable X mit einer Wahrscheinlichkeit von β nicht überschreitet.

Neben der Charakterisierung der Verteilung lassen sich mit der Quantilfunktion Fragen der Art “Wie viel verdienen die 10% ärmsten Haushalte (höchstens)?” oder “Wie lange bleibt die treuere Hälfte der Personen bei ihrem ersten Job (mindestens)?” beantworten.

Das β -Quantil (und damit die Quantilfunktion) bestimmt man genau gleich aus der Verteilungsfunktion F , wie wir Stichprobenquantile grafisch von der ECDF abgelesen haben: Man sucht den kleinsten Wert r auf der x -Achse, bei dem $F(r)$ mindestens β beträgt. Die linke Hälfte von Abbildung 3.10 illustriert dies schematisch.

Solange die Verteilungsfunktion keine horizontal verlaufenden Abschnitte¹ aufweist, kann die Umkehrfunktion F^{-1} von F auch algebraisch bestimmt werden:

Dazu löst man die Gleichung

$$F(r) := \beta$$

nach r auf bzw. fragt sich, für welches r die Verteilungsfunktion $F(r)$ gerade β beträgt. Die Lösung r entspricht dann $F^{-1}(\beta)$.

Beispiel 3.25 (Exponentialverteilung). Wie lautet die Umkehrfunktion von

$$F(r) = 1 - e^{-\lambda r},$$

der Verteilungsfunktion von $\text{Exp}(\lambda)$? Dazu lösen wir die Gleichung $F(r) := \beta$ nach r auf:

$$\begin{aligned} 1 - e^{-\lambda r} &= \beta \\ \Leftrightarrow e^{-\lambda r} &= 1 - \beta \\ \Leftrightarrow -\lambda r &= \ln(1 - \beta) \\ \Leftrightarrow r &= -\ln(1 - \beta)/\lambda. \end{aligned}$$

Die Umkehrfunktion von F bzw. die Quantilfunktion der Exponentialverteilung beträgt also

$$F^{-1}(\beta) = -\ln(1 - \beta)/\lambda.$$

Eine nach $\text{Exp}(\lambda)$ -verteilte Zufallsvariable X hat damit den Median

$$F^{-1}(0.5) = -\ln(1 - 0.5)/\lambda = \ln(2)/\lambda.$$

Für die Situation in Beispiel 3.23 folgt z. B., dass die Hälfte aller Personen kürzer/länger als $\ln(2)/0.2 \approx 3.5$ Jahre bei ihrer ersten Stelle bleiben. Abbildung 3.10 illustriert das Ergebnis auf zwei Arten. ▲

Beispiel 3.26 (Uniformverteilung). Die Verteilungsfunktion der Standarduniformverteilung weist für $0 \leq r \leq 1$ den Wert r auf. Da eine Umkehrfunktion geometrisch eine Spiegelung an der Hauptdiagonale ist, lautet die Quantilfunktion ebenfalls $F^{-1}(\beta) = \beta$. Der Median dieser Verteilung ist also $F^{-1}(0.5) = 0.5$. ▲

Quantile und Stichprobenquantile Mit wachsendem Stichprobenumfang streben die Stichprobenquantile einer numerischen Variable X gegen die unbekannten theoretischen Quantile von X . Dies folgt aus der entsprechenden Verbindung zwischen ECDF und Verteilungsfunktion. Wir illustrieren die Tatsache in Abbildung 3.11 anhand der Exponentialverteilung.

¹Dies gilt für alle wichtigen stetigen Verteilungen, jedoch für keine diskrete Verteilung.

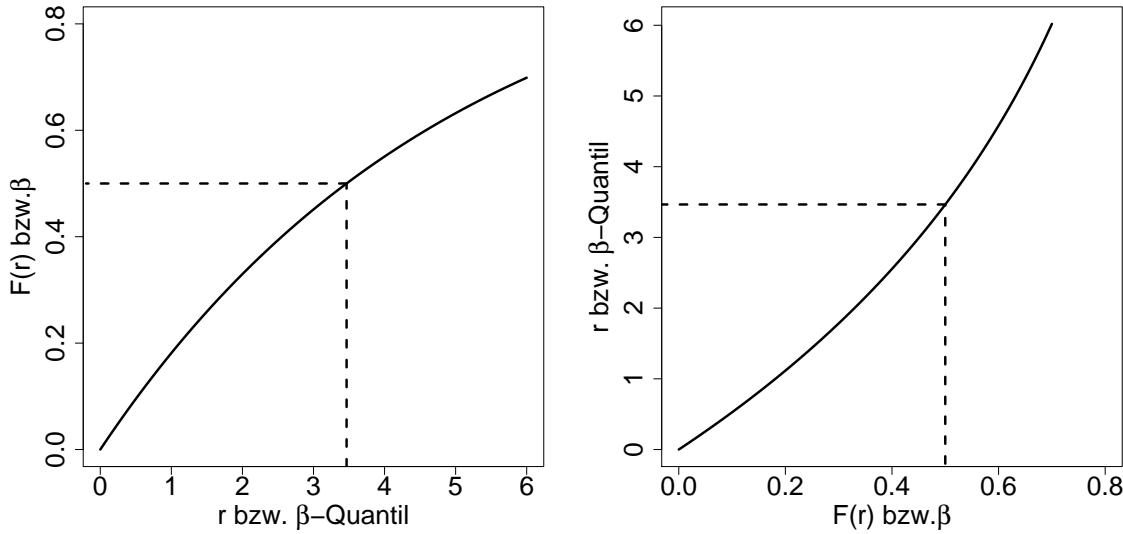


Abbildung 3.10: Verteilungsfunktion (linkes Bild) und Quantilfunktion (rechtes Bild) von $\text{Exp}(0.2)$. Die gepunkteten Linien verbinden den Median mit dem entsprechenden Anteil 0.5.

Erwartungswert

Bis anhin haben wir die theoretischen Pendants zu Histogramm (Dichtefunktion), relativen Häufigkeiten (Wahrscheinlichkeitsfunktion), ECDF (Verteilungsfunktion) und Stichprobenquantilen (Quantilfunktion) kennengelernt und damit je die Verteilung einer Zufallsvariable vollständig charakterisiert. Schliesslich gehen wir auf die Gegenstücke von empirischem Mittelwert und Standardabweichung/Varianz ein, mit denen zwei wichtige Aspekte einer Verteilung quantifiziert werden.

Der *Erwartungswert* einer Zufallsvariable X gibt den durchschnittlichen Wert einer Realisierung von X an. Ist X ein Merkmal, so kann der Erwartungswert von X als Populationsmittelwert aufgefasst werden und ist damit das theoretische Pendant zum Stichprobenmittelwert.

Erwartungswerte werden als Schwerpunkt der Wahrscheinlichkeits- bzw. Dichtefunktion f berechnet – im diskreten Fall also durch

$$E(X) := \sum_j x_j f(x_j)$$

und im stetigen Fall durch

$$E(X) := \int_{-\infty}^{\infty} x f(x) dx.$$

Beispiel 3.27 (Bernoulliverteilung). Mit der Wahrscheinlichkeitfunktion $f(0) = 1 - p$ und $f(1) = p$ von $\text{Bern}(p)$ folgt direkt, dass $E(X) = 0 \cdot f(0) + 1 \cdot f(1) = 0 \cdot (1 - p) + 1 \cdot p = p$.

Die Erfolgswahrscheinlichkeit p kann also als theoretischer Mittelwert von X aufgefasst werden. ▲

Beispiel 3.28 (Würfelwurf). Beim Würfelwurf folgt mit $f(x_j) = 1/6$, $x_j = 1, \dots, 6$, dass

$$E(X) = 1 \cdot f(1) + 2 \cdot f(2) + \dots + 6 \cdot f(6) = (1 + \dots + 6)/6 = 3.5.$$

Die mittlere Augensumme beträgt damit 3.5. ▲

Beispiel 3.29 (Exponentialverteilung). Mit Hilfe einiger Analysis¹ folgt aus der Dichtefunktion

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

¹Für mathematisch Interessierte: Entweder mit partieller Integration und Grenzwertrechnung oder aber mit der Substitution $y := \lambda x$ und dem Gammaintegral $\int_0^\infty y^n e^{-y} dy = n!$.

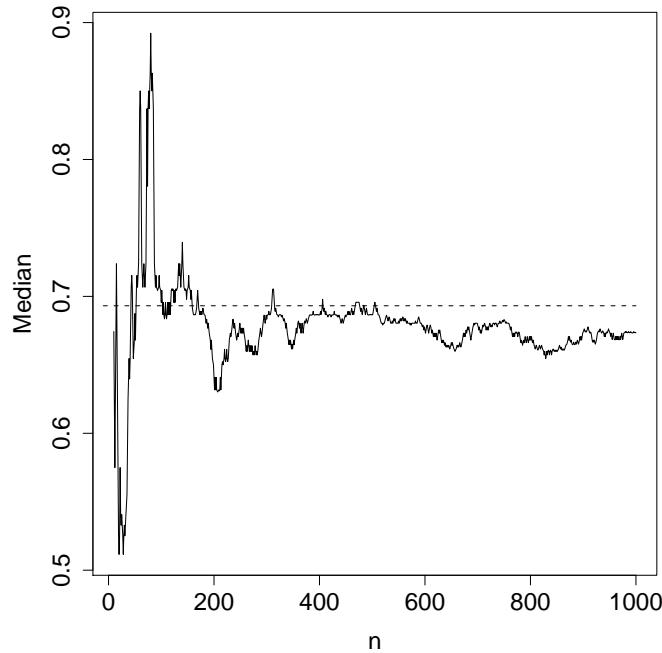


Abbildung 3.11: Der Median von 1'000 Realisierungen von i. i. d. nach $\text{Exp}(1)$ -verteilten Zufallsvariablen wird sukzessive gegen den Stichprobenumfang n aufgetragen und mit dem tatsächlichen Median $\ln(2) \approx 0.693$ verglichen. Im Schnitt wird die Übereinstimmung mit wachsendem n besser.

dass

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \lambda \int_0^{\infty} xe^{-\lambda x}dx = \dots = \frac{1}{\lambda}.$$

Der Parameter λ könnte also als reziproker Wert des Stichprobenmittelwerts aus Daten geschätzt werden.

In Beispiel 3.25 haben wir gezeigt, dass der Median einer $\text{Exp}(\lambda)$ -verteilten Zufallsvariable $\ln(2)/\lambda$ beträgt. Für solche Zufallsvariablen ist der Mittelwert also stets $1/\ln(2) \approx 1.44$ mal so gross wie der Median. ▲

Regeln

1. Der Erwartungswert einer Summe von Zufallsvariablen entspricht der Summe der Erwartungswerte. (Summe und Erwartungswert lassen sich vertauschen.)
2. Der Erwartungswert eines Produktes von *unabhängigen* Zufallsvariablen entspricht dem Produkt der Erwartungswerte. (Produkt und Erwartungswert lassen sich unter Unabhängigkeit vertauschen.)
3. Bei Zufallsvariablen mit achsensymmetrischer Wahrscheinlichkeits- oder Dichtefunktion entspricht der Erwartungswert gerade dem Ort der Symmetrieachse.
4. Nicht alle Zufallsvariablen haben einen (endlichen) Erwartungswert. Das Paradebeispiel einer Verteilung ohne Erwartungswert ist die Cauchyverteilung (Abbildung 3.28).
5. Für beliebige Konstanten a und b gilt $E(a + bX) = a + bE(X)$. (Der Erwartungswert einer linearen Abbildung entspricht der linearen Abbildung des Erwartungswerts.)
6. Der Erwartungswert einer transformierten Zufallsvariable $Y = g(X)$ entspricht nur für lineare g (siehe letzte Regel) dem transformierten Erwartungswert $g(E(X))$. Immerhin kann $E(Y)$ mit Hilfe der Verteilung von X berechnet werden, also ohne erst mit viel Mathematik die Verteilung von Y zu bestimmen:

Hat X Wahrscheinlichkeits- bzw. Dichtefunktion f , gilt im diskreten Fall

$$E(g(X)) = \sum_j g(x_j) f(x_j)$$

und im stetigen Fall

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) f(x) dx.$$

Diese Formeln werden uns bei der Berechnung von Varianzen dienlich sein.

Um diese Regeln zu illustrieren, betrachten wir weitere Beispiele.

Beispiel 3.30 (Stichprobenmittelwert). Eine besonders wichtige Zufallsvariable in der Statistik ist der Stichprobenmittelwert \bar{X} von Beobachtungen X_1, X_2, \dots, X_n je mit Erwartungswert μ . Er dient als Schätzer für den (unbekannten) Populationsmittelwert μ . Aus den Regeln 1 und 5 folgt

$$E(\bar{X}) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} E\left(\sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} \cdot n \cdot E(X_i) = E(X_i) = \mu.$$

Unabhängig vom Stichprobenumfang n entspricht \bar{X} damit *im Schnitt* genau μ .

Der Erwartungswert der relativen Häufigkeit von Ausprägung 1 eines binären (0-1)-Merkmals entspricht also dem tatsächlichen relativen Anteil in der Population. \blacktriangle

Hinweis (Erwartungstreue Schätzer). Neben dem Stichprobenmittelwert und relativen Häufigkeiten entsprechen viele weitere Schätzer im Schnitt dem Parameter θ , den sie schätzen, treffen also (unabhängig vom Stichprobenumfang) im Schnitt ins Schwarze. Sie heißen *erwartungstreu* bzw. *unverfälscht* für θ . Übrigens kann man zeigen, dass die Stichprobenvarianz erwartungstreu für die Varianz in der Population ist. Dies erklärt nachträglich den seltsamen Faktor $1/(n - 1)$ statt $1/n$ in ihrer Definition.

Beispiel 3.31 (Uniformverteilung). Bei der Standarduniformverteilung ergibt sich mit der Dichte

$$f(x) = \begin{cases} 1, & \text{wenn } 0 \leq x \leq 1, \\ 0, & \text{sonst} \end{cases}$$

definitionsgemäß ein Erwartungswert von

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx = \int_0^1 x \cdot 1 \cdot dx = \frac{x^2}{2} \Big|_0^1 = 0.5.$$

Da die Dichtefunktion achsensymmetrisch um 0.5 ist, wären wir dank Regel 3 ohne zu rechnen zum gleichen Ergebnis gekommen. Da die allgemeine Uniformverteilung zwischen $a < b$ achsensymmetrisch um $(a+b)/2$ ist, gilt dort $E(X) = (a+b)/2$. \blacktriangle

Beispiel 3.32 (Portfoliorenditen). Ein Portfolio bestehe zu zwei Dritteln aus Anlage 1 (z. B. ein Aktienpaket) mit Monatsrendite R_1 und zu einem Drittel aus Anlage 2 (z. B. Immobilien) mit Monatsrendite R_2 . Wie hängt die erwartete Portfoliorendite eines künftigen Monats, also der Erwartungswert von

$$T = \frac{2}{3}R_1 + \frac{1}{3}R_2,$$

von $E(R_1)$ und $E(R_2)$ ab? Mit den Regeln 1 und 5 finden wir die Formel

$$E(T) = E\left(\frac{2}{3}R_1 + \frac{1}{3}R_2\right) = E\left(\frac{2}{3}R_1\right) + E\left(\frac{1}{3}R_2\right) = \frac{2}{3}E(R_1) + \frac{1}{3}E(R_2).$$

Anhand der Erfahrungen der letzten Jahre nimmt man an, dass R_1 uniformverteilt zwischen -1.5% und 2% sei und R_2 uniformverteilt zwischen -1% und 0.8% . Aus Beispiel 3.31 folgt, dass

$$E(R_1) = \frac{-1.5\% + 2\%}{2} = 0.25\% \text{ und } E(R_2) = \frac{-1\% + 0.8\%}{2} = -0.1\%.$$

Diese Werte setzen wir in die Formel für $E(T)$ ein und erhalten konkret

$$E(T) = \frac{2}{3} \cdot 0.25\% - \frac{1}{3} \cdot 0.1\% = 4/30\% \approx 0.133\%.$$

Für künftige Monate erwartet man also eine Portfoliorendite von 0.133%.

Es wäre deutlich aufwändiger, mit viel Mathematik die Verteilung von T zu bestimmen und dann damit den Erwartungswert 0.133 zu berechnen. Eine dritte Möglichkeit, $E(T)$ zu finden, beruht auf der *Simulation*. ▲

Simulationen Ist die Verteilung einer oder mehrerer Zufallsvariablen X, Y, \dots bekannt, so ist es mathematisch anspruchsvoll¹ bis unmöglich, die Verteilung einer daraus berechneten neuen Zufallsvariable $T := g(X, Y, \dots)$ zu bestimmen und damit Wahrscheinlichkeiten, Erwartungswerte, Quantile oder weitere Eigenschaften von T zu ermitteln. Eine einfache Alternative stellen sogenannte (Monte-Carlo)-Simulationen dar: Mithilfe einer entsprechenden Software erzeugt man je viele (z. B. eine Million) unabhängige Realisierungen von X, Y, \dots und berechnet damit die entsprechenden Realisierungen von $T = g(X, Y, \dots)$. Wahrscheinlichkeiten, Quantile u. s. w. lassen sich dann approximativ mithilfe der deskriptiven Statistik aus der empirischen Verteilung von T finden.

Das Prinzip der Simulation beruht auf den bisher festgestellten Verbindungen zwischen Empirie und Theorie bei grossen Stichproben. Simulationen bilden also die Brücke zwischen deskriptiver Statistik und Wahrscheinlichkeitsrechnung.

Beispiel 3.33 (Portfoliorenditen, Fortsetzung). Das Management von Beispiel 3.32 interessiert sich nicht nur für die erwartete Portfoliorendite der künftigen Monate. Es möchte generell deren Verteilung kennen, um beispielsweise das Risiko des Portfolios abschätzen zu können. (Der Einfachheit halber trifft es die unrealistische Annahme, dass R_1 und R_2 unabhängig sind.) Als Mass für das Risiko verwendet das Management den sogenannten *5%-value at risk* (kurz: 5%-V@R). Dieser entspricht minus dem 5%-Quantil der Verteilung des Gewinns (bzw. hier der Rendite).

Wir simulieren² je eine Million Monatsrenditen der zwei Anlagen und beantworten die Fragen des Managements anhand der empirischen Verteilung der daraus berechneten Portfoliorenditen.

R Code

```
# Eingabe: Erzeugung der künstlichen Daten
set.seed(1)          # Damit die Ergebnisse reproduzierbar sind
n <- 1000000
R1 <- runif(n, -1.5, 2)
R2 <- runif(n, -1, 0.8)
T <- 2/3*R1 + 1/3*R2

# Verteilungsfunktion von T
Ecdf(T)            # Benötigt Hmisc-Paket

# Mittelwert
mean(T)             # Ergibt 0.13298

# 5%-Quantil
quantile(T, 0.05)   # Ergibt -0.96018
```

Kommentare: Die Simulation liefert fast den gleichen Mittelwert wie in Beispiel 3.32. Von der ECDF in Abbildung 3.12 kann beispielsweise abgelesen werden, dass in etwas weniger als $1 - 0.4 = 60\%$ der Fälle ein Gewinn zu erwarten ist. Das 5%-Quantil der Verteilung beträgt -0.96% , d. h. mit einer Wahrscheinlichkeit von 5% muss mit einem Monatsverlust von mindestens 0.96% gerechnet werden (5%-V@R). ▲

¹Meist sind Mehrfachintegrale involviert.

²Mit der R-Funktion `runif` (von engl. random uniform).

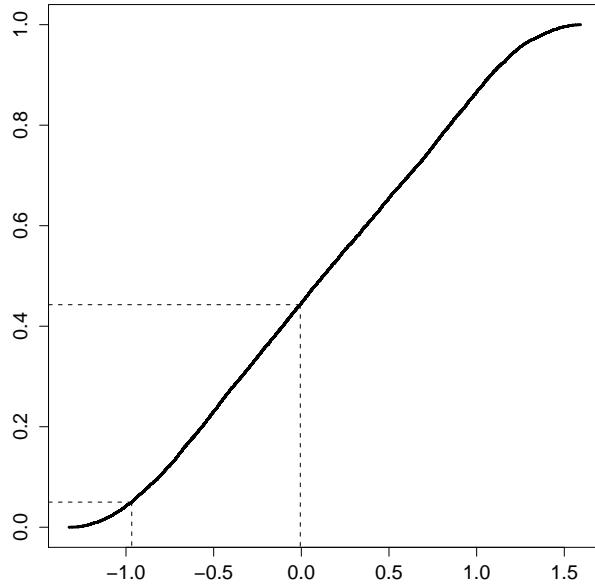


Abbildung 3.12: (Simulierte) Verteilungsfunktion der Portfoliorendite aus Beispiel 3.33.

Varianz und Standardabweichung

Die theoretischen Pendants zu Stichprobenvarianz und -standardabweichung quantifizieren die typischen Abweichungen von $E(X)$.

Die *Varianz* einer Zufallsvariable X ist definiert als die Zahl

$$\text{Var}(X) := E((X - E(X))^2),$$

entspricht also dem Erwartungswert der neuen Zufallsvariable $Y := (X - E(X))^2$. Will man Varianzen aus der Dichte- bzw. Wahrscheinlichkeitsfunktion bestimmen, so verwendet man oft die alternative (äquivalente) Definition

$$\text{Var}(X) := E(X^2) - (E(X))^2,$$

wobei $E(X^2)$ meist mit Regel 6 für Erwartungswerte gefunden wird.

Wie in der deskriptiven Statistik entspricht die *Standardabweichung* der Wurzel aus der Varianz.

Beispiel 3.34 (Bernoulliverteilung). Mit der alternativen Definition der Varianz und Regel 6 für Erwartungswerte folgt für $X \sim \text{Bern}(p)$ mit Wahrscheinlichkeitsfunktion $f(0) = 1 - p$ und $f(1) = p$, dass

$$\text{Var}(X) = E(X^2) - (E(X))^2 = 0^2 \cdot f(0) + 1^2 \cdot f(1) - p^2 = p - p^2 = p(1 - p).$$

Entsprechend gilt $\text{Std}(X) = \sqrt{p(1 - p)}$. Dieser Wert ist übrigens an der Stelle $p = 0.5$ mit einem Wert von 0.5 maximal. ▲

Beispiel 3.35 (Würfelwurf). Beim fairen Würfelwurf finden wir analog, dass

$$\text{Var}(X) = E(X^2) - (E(X))^2 = 1^2/6 + 2^2/6 + \dots + 6^2/6 - 3.5^2 = 91/6 - 12.25 \approx 2.92$$

und damit $\text{Std}(X) = \sqrt{91/6 - 12.25} \approx 1.71$. ▲

Beispiel 3.36 (Exponentialverteilung). Mit einem analytischen Aufwand¹ lässt sich zeigen, dass die Varianz einer $\text{Exp}(\lambda)$ -verteilten Zufallsvariable $1/\lambda^2$ beträgt. Die Standardabweichung ist damit $1/\lambda$. Bei der Exponentialverteilung ist die Standardabweichung also gleich gross wie der Erwartungswert. ▲

Beispiel 3.37 (Uniformverteilung). Für eine standarduniformverteilte Zufallsvariable (Dichte $f(x) = 1$ für $0 \leq x \leq 1$) gilt

$$\text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \int_0^1 x^2 f(x) dx - 0.5^2 = \int_0^1 x^2 dx - 0.5^2 = \frac{x^3}{3} \Big|_{x=0}^1 - 0.5^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12} \approx 0.083.$$

Die Standardabweichung beträgt entsprechend $\text{Std}(X) = \sqrt{1/12} \approx 0.29$.

Für die allgemeine Uniformverteilung zwischen a und b ergäbe sich analog die Varianz $(b-a)^2/12$. ▲

Regeln

1. Für beliebige Konstanten a und b gilt $\text{Var}(a+bX) = b^2 \text{Var}(X)$ und $\text{Std}(a+bX) = |b| \text{Std}(X)$. Insbesondere verändern sich Varianz und Standardabweichung nicht, wenn man die Zufallsvariable um einen festen Wert verschiebt.
2. Die Varianz einer Summe von *unabhängigen* Zufallsvariablen entspricht der Summe ihrer Varianzen.
3. Nicht alle Zufallsvariablen haben eine (endliche) Standardabweichung bzw. Varianz. Ein Beispiel einer Verteilung ohne Varianz ist wiederum die Cauchyverteilung von Abbildung 3.28.

Wenden wir nun die ersten beiden Regeln an Beispielen an.

Beispiel 3.38 (Stichprobenmittelwert). Die Varianz des Mittelwerts \bar{X} von n unabhängigen Beobachtungen X_1, X_2, \dots, X_n je mit $\text{Var}(X_i) = \sigma^2$ und $\mathbb{E}(X_i) = \mu$ beträgt dank Regeln 1 und 2

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{\text{Var}(X_1)}{n} = \frac{\sigma^2}{n}.$$

Entsprechend gilt $\text{Std}(\bar{X}) = \sigma/\sqrt{n}$. Je grösser die Stichprobe, desto weniger streut also der Mittelwert um den wahren Mittelwert μ . Konkret führt eine viermal so grosse Stichprobe zu einer Halbierung der Standardabweichung des Mittelwerts, also zu einer Verdoppelung der Präzision des Schätzwerts für μ .

Wir haben gesehen, dass $\text{Bern}(p)$ -verteilte Beobachtungen je Varianz $\sigma^2 = p(1-p)$ haben. Da die relative Häufigkeit \hat{p} von Ausprägung 1 dem Mittelwert solcher Beobachtungen entspricht, ist $\text{Var}(\hat{p}) = \sigma^2/n = p(1-p)/n$ bzw. $\text{Std}(\hat{p}) = \sqrt{p(1-p)/n}$. ▲

Gesetz der grossen Zahlen Eine Konsequenz aus dem letzten Beispiel ist das *Gesetz der grossen Zahlen*: Der Stichprobenmittelwert strebt bei wachsendem Stichprobenumfang gegen den entsprechenden Populationsmittelwert.

Wir illustrieren dieses Resultat in Abbildung 3.13 anhand der Bernoulliverteilung.

Dank dem Gesetz der grossen Zahlen streben relative Häufigkeiten bei wachsendem Stichprobenumfang gegen die tatsächlichen relativen Anteile.

Hinweis (Konsistente Schätzer). Nicht nur der Mittelwert strebt bei wachsendem n gegen den Parameter, den er schätzt. Alle vernünftigen Schätzer weisen diese Eigenschaft auf. Sie heisst *Konsistenz*.

¹Für mathematisch Interessierte: Am einfachsten mit dem in Beispiel 3.29 erwähnten Weg über Substitution und Gammaintegral.

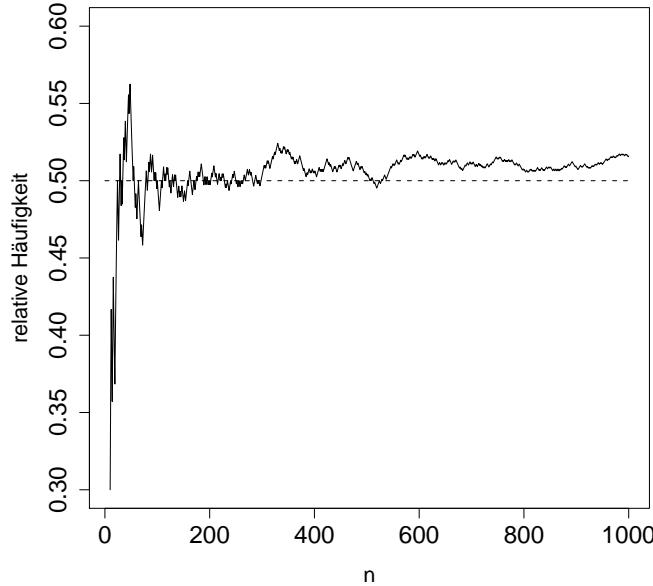


Abbildung 3.13: Die relative Häufigkeit der Ausprägung ‘1’ (also der Mittelwert) von 1’000 unabhängigen Bern(0.5)-Zufallsvariablen wird sukzessive gegen den Stichprobenumfang n aufgetragen und mit dem tatsächlichen relativen Anteil 0.5 verglichen. Dank dem Gesetz der grossen Zahlen wird diese Übereinstimmung *im Schnitt* mit wachsendem n besser.

Beispiel 3.39 (Standardisierung). Manchmal betrachtet man statt die Zufallsvariable X ihre *standardisierte* Version

$$Z := \frac{X - E(X)}{\text{Std}(X)}.$$

Eine solche hat laut Regel 1 die Varianz

$$\text{Var}(Z) = \text{Var}\left(\frac{X - E(X)}{\text{Std}(X)}\right) = \frac{\text{Var}(X)}{\text{Std}(X)^2} = 1.$$

Die entsprechende Regel zu Erwartungswerten liefert zudem

$$E(Z) = E\left(\frac{X - E(X)}{\text{Std}(X)}\right) = \frac{E(X) - E(X)}{\text{Std}(X)} = 0.$$

Standardisierte Zufallsvariablen haben also stets Erwartungswert 0 und Varianz/Standardabweichung 1. ▲

Beispiel 3.40 (Portfoliorenditen). Für die Monatsrendite des Portfolios von Beispiel 3.32 finden wir mit den Regeln 1 und 2 sowie der Annahme, dass die Monatsrenditen R_1 und R_2 unabhängig sind, dass

$$\text{Var}(T) = \text{Var}\left(\frac{2}{3}R_1 + \frac{1}{3}R_2\right) = \text{Var}\left(\frac{2}{3}R_1\right) + \text{Var}\left(\frac{1}{3}R_2\right) = \frac{4}{9}\text{Var}(R_1) + \frac{1}{9}\text{Var}(R_2).$$

Ist R_1 uniformverteilt zwischen -1.5% und 2% und R_2 zwischen -1% und 0.8% , so ist mit Beispiel 3.37 $\text{Var}(R_1) = 3.5^2/12 \approx 1.02$ und $\text{Var}(R_2) = 1.8^2/12 = 0.27$. Damit erhalten wir konkret $\text{Var}(T) \approx 0.48$. (Die Einheit bei diesen Varianzen ist jeweils % im Quadrat.)

Mit der Simulation in Beispiel 3.32 bestätigen wir die Berechnungen:

# Eingabe, Forts.	R Code	
var(T)	# Ergibt 0.4838	

Die Standardabweichung, in der Finanzwelt auch *Volatilität* genannt, beträgt etwa $\sqrt{0.48} = 0.69$ Prozent.

▲

Beispiel 3.41 (Paretoverteilung). Nun illustrieren wir die Verfahren dieses Abschnittes an einer weiteren stetigen Verteilung, der *Paretoverteilung*. Diese rechtsschiefe Verteilung dient in der Ökonomie oftmals als Modell für ungleichverteilte Größen wie Vermögen, Beträge von Versicherungsschäden und Stadtgrößen. Damit ist beispielsweise auch die 80:20-Regel bzw. das Pareto-Prinzip verbunden: 80% der Arbeit lässt sich mit 20% Aufwand erledigen. Die restlichen 20% der Arbeit nehmen 80% des Aufwands in Anspruch.

Die Dichte der Paretoverteilung ist eine Potenzfunktion. Sie lautet

$$f(x) := \frac{k}{x_m} \left(\frac{x_m}{x} \right)^{k+1} = kx_m^k x^{-k-1}$$

für Werte x , die grösser als der Mindestwert x_m sind. Je grösser der Parameter $k > 0$, je steiler ist die Kurve bzw. je ungleichverteilter die betrachtete Grösse.

Bezeichnen wir nun mit X die Anzahl Einwohner einer zufällig herausgepickten Stadt in Europa. Aufgrund

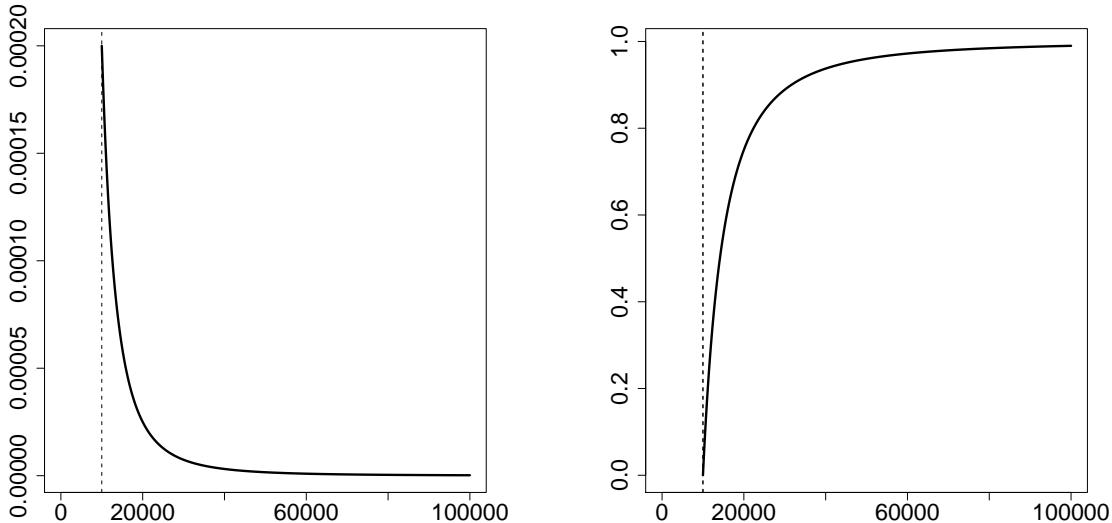


Abbildung 3.14: Dichte (links) und Verteilungsfunktion (rechts) der Paretoverteilung von Beispiel 3.41.

von früheren Betrachtungen wisse man, dass X ungefähr einer Paretoverteilung mit Parametern $k = 2$ und $x_m = 10000$ folgt¹, siehe Abbildung 3.14. Mit der entsprechenden Verteilungsfunktion

$$F(r) = \int_{-\infty}^r f(x) dx = kx_m^k \int_{x_m}^r x^{-k-1} dx = kx_m^k x^{-k} \Big|_{x_m}^r \cdot \frac{1}{-k} = -x_m^k \cdot (r^{-k} - x_m^{-k}) = 1 - \left(\frac{x_m}{r} \right)^k$$

können wir nun, ohne über Daten zu verfügen, z. B. den relativen Anteil der Kleinstädte bis 20'000 Einwohner oder den relativen Anteil der Millionenstädte bestimmen:

$$\begin{aligned} P(X \leq 20000) &= F(20000) = 1 - \left(\frac{10000}{20000} \right)^2 = 3/4 \quad \text{und} \\ P(X \geq 1000000) &= 1 - F(1000000) = \left(\frac{10000}{1000000} \right)^2 = 0.0001. \end{aligned}$$

Bestimmen wir nun die Umkehrfunktion F^{-1} von F , also die Quantilfunktion. Dazu lösen wir die Gleichung $\beta = F(r)$ nach r auf:

$$\beta = 1 - \left(\frac{x_m}{r} \right)^k \Leftrightarrow 1 - \beta = \left(\frac{x_m}{r} \right)^k \Leftrightarrow \sqrt[k]{1 - \beta} = x_m/r \Leftrightarrow r = \frac{x_m}{\sqrt[k]{1 - \beta}}.$$

¹Städte sind Orte mit mindestens 10'000 Einwohnern.

Die Quantilfunktion lautet demnach

$$F^{-1}(\beta) = \frac{x_m}{\sqrt[k]{1-\beta}}.$$

Hiermit können wir beispielsweise berechnen, dass die 10% grössten Städte mindestens

$$F^{-1}(0.9) = \frac{10000}{\sqrt{1-0.9}} = 31'623$$

Einwohner haben oder dass die mediane (typische) Stadt

$$F^{-1}(0.5) = \frac{10000}{\sqrt{1-0.5}} = 14'142$$

Personen umfasst. Die durchschnittliche Einwohnerzahl ist aufgrund der rechtsschiefen Verteilung sicherlich höher. Für eine konkrete Berechnung ermitteln wir den Erwartungswert per Integral:

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = kx_m^k \int_{x_m}^{\infty} x^{-k} dx = kx_m^k x^{-(k-1)} \Big|_{x_m}^{\infty} \cdot \frac{1}{-(k-1)} = x_m^k x_m^{-(k-1)} \frac{k}{k-1} = \frac{x_m k}{k-1}, \quad k > 1.$$

Die mittlere Einwohnerzahl beträgt damit $2x_m = 20'000$. Ähnlich könnte man auch zeigen, dass

$$\text{Var}(X) = E(X^2) - (E(X))^2 = \int_{-\infty}^{\infty} x^2 f(x)dx - \left(\frac{x_m k}{k-1} \right)^2 = \dots = \frac{x_m}{k-1} \sqrt{\frac{k}{k-2}}, \quad k > 2.$$

▲

3.2 Binomialverteilung und Verwandtes

3.2.1 Binomialverteilung

Eine für die Statistik wichtige diskrete Verteilung ist die *Binomialverteilung*. Der Hauptgrund dafür ist die Tatsache, dass mit ihr die Verteilung von absoluten Häufigkeiten beschrieben werden kann.

Die Binomialverteilung mit Parametern $n = 1, 2, \dots$ und $0 \leq p \leq 1$, kurz $\text{Bin}(n, p)$, ist durch ihre Wahrscheinlichkeitsfunktion

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

für $k = 0, 1, \dots, n$ definiert. Abbildung 3.15 zeigt einige Binomialverteilungen.

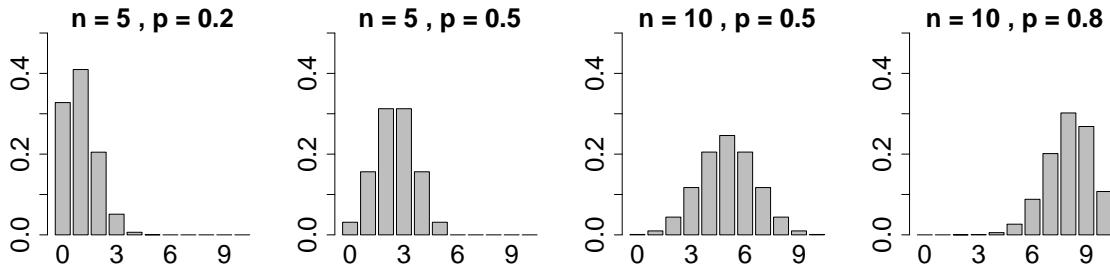


Abbildung 3.15: Wahrscheinlichkeitsfunktionen von $\text{Bin}(5,0.2)$, $\text{Bin}(5,0.5)$, $\text{Bin}(10,0.5)$ und $\text{Bin}(10,0.8)$.

Hinweis. Der *Binomialkoeffizient*

$$\binom{n}{k} := \frac{n!}{(n-k)!k!}$$

wird “ n tief k ” ausgesprochen. Er gibt die Anzahl der Möglichkeiten an, wie man aus n verschiedenen Objekten k Stück auswählen kann, ohne dabei die Reihenfolge zu beachten.

Eigenschaften

1. $\text{Bin}(1, p)$ entspricht $\text{Bern}(p)$.
2. Die Summe von unabhängigen binomialverteilten Zufallsvariablen mit gleichem p ist wiederum binomialverteilt: Für $X \sim \text{Bin}(n, p)$ und $Y \sim \text{Bin}(m, p)$ folgt $X + Y \sim \text{Bin}(n+m, p)$.
3. Aufgrund der ersten beiden Eigenschaften folgt die wichtige Tatsache, dass die Summe X von n unabhängigen nach $\text{Bern}(p)$ -verteilten Zufallsvariablen Z_1, \dots, Z_n binomialverteilt mit Parametern n und p ist. X könnte beispielsweise die absolute Häufigkeit der Ausprägung 1 eines binären Merkmals sein. Daraus folgen Erwartungswert und Varianz von $X \sim \text{Bin}(n, p)$.
4. $E(X) = E(\sum_{i=1}^n Z_i) = \sum_{i=1}^n E(Z_i) = np$.
5. $\text{Var}(X) = \text{Var}(\sum_{i=1}^n Z_i) = \sum_{i=1}^n \text{Var}(Z_i) = np(1-p)$.
6. Die Wahrscheinlichkeitsfunktion f hat einen Höcker um den Erwartungswert.
7. Für $p = 0.5$ ist f symmetrisch, für $p < 0.5$ rechtsschief und für $p > 0.5$ linksschief.

Beispiel 3.42 (Fünffacher Münzwurf). Eine Münze werde fünfmal geworfen. Wir betrachten die Anzahl X von Würfen, bei denen ‘Kopf’ auftritt. Diese Zufallsvariable X ist binomialverteilt mit Parametern $n = 5$ und $p = 1/2$. Also gilt für $k = 0, 1, \dots, 5$:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{5}{k} (1/2)^5 = \binom{5}{k} / 32.$$

Beispielsweise finden wir mit dieser Formel folgende Wahrscheinlichkeiten:

$$\begin{aligned} P(\text{Nie Kopf geworfen}) &= P(X = 0) = \binom{5}{0} / 32 = 1/32 = 0.03125. \\ P(\text{Genau einmal Kopf geworfen}) &= P(X = 1) = \binom{5}{1} / 32 = 5/32 = 0.15625. \\ P(\text{Mindestens zweimal Kopf geworfen}) &= P(X \geq 2) = 1 - P(X = 0) - P(X = 1) \\ &= 1 - 1/32 - 5/32 = 26/32 = 0.8125. \end{aligned}$$

▲

Beispiel 3.43 (Umfrage). Aus einer grossen Population mit gleich vielen Frauen wie Männern ziehen wir eine Zufallsstichprobe vom Umfang 100. Wie gross ist die Wahrscheinlichkeit, dass die Stichprobe höchstens 40 Frauen enthält? Dank Eigenschaft 3 ist die Anzahl X der Frauen $\text{Bin}(100, 0.5)$ -verteilt. Die gesuchte Wahrscheinlichkeit¹ beträgt $P(X \leq 40) = f(0) + f(1) + \dots + f(40) \approx 0.028$. Dabei bezeichnet f die Wahrscheinlichkeitsfunktion von $\text{Bin}(100, 0.5)$. Als Alternative können wir mit der Software die entsprechende Verteilungsfunktion $F(r) = P(X \leq r)$ an der Stelle $r = 40$ bestimmen, vergleiche auch Abbildung 3.16.

<pre>sum(dbinom(0:40, 100, 0.5))</pre>	R Code # Ergibt 0.028444
<pre>pbinom(40, 100, 0.5)</pre>	# Ergibt 0.028444

Ab wie vielen Frauen in der Stichprobe würde die entsprechende Wahrscheinlichkeit mindestens 0.05 betragen? Diese Frage kann man entweder grafisch mit der Verteilungsfunktion (Abbildung 3.16) oder direkt mit der Quantilfunktion der Software an der Stelle 0.05 beantworten.

<pre>qbinom(0.05, 100, 0.5)</pre>	R Code # Ergibt 42
-----------------------------------	-----------------------

¹Da es von Hand mühsam ist, 41 Wahrscheinlichkeiten zu berechnen und diese zu addieren, verwenden wir die Wahrscheinlichkeitsfunktion `dbinom` in R. Die Verteilungsfunktion heisst `pbinom`, die Quantilfunktion `qbinom`.

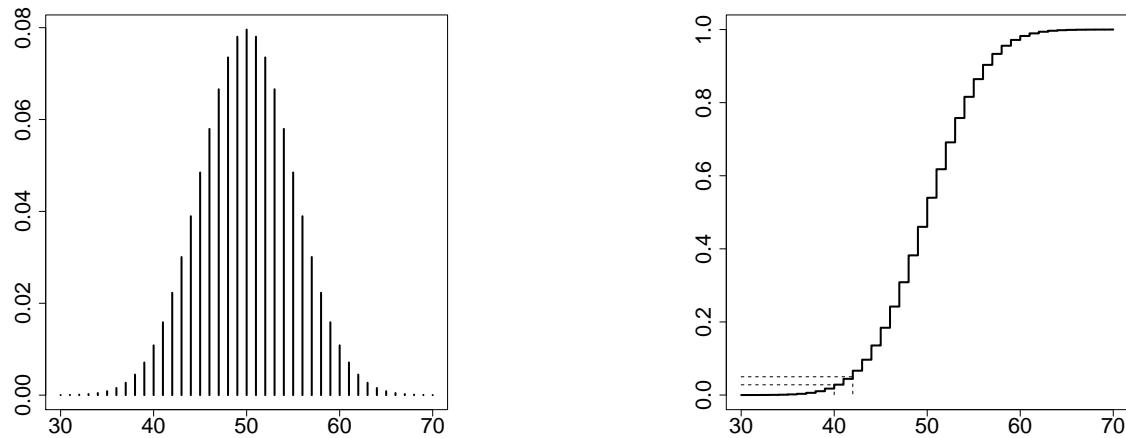


Abbildung 3.16: Wahrscheinlichkeitsfunktion (linkes Bild) und Verteilungsfunktion (rechtes Bild) von $\text{Bin}(100, 0.5)$ aus Beispiel 3.43.

▲

In der Realität ist der Parameter p unbekannt und wird anhand der Stichprobe durch die relative Häufigkeit \hat{p} geschätzt. Da die Verteilung des Schätzers \hat{p} bzw. der absoluten Häufigkeit $n\hat{p}$ (bis auf p) bekannt ist ($n\hat{p} \sim \text{Bin}(n, p)$), lassen sich mit entsprechender Software Konfidenzintervalle und Tests für p berechnen. Wir führen diese allgemeinen Konzepte anhand solcher Binomialkonfidenzintervalle und -Tests ein.

3.2.2 Konfidenzintervalle allgemein und für relative Anteile

Konfidenzintervalle allgemein

Ein *95%-Konfidenzintervall (Vertrauensintervall)* für einen unbekannten Parameter θ ist ein Intervall, in dem sich θ mit einer Sicherheit von (mindestens) 95% befindet.

Hinweise

- Die Software sucht das Intervall anhand Verteilungs- oder Quantilfunktion des Schätzers. Die Berechnung von Hand ist nur für wenige Parameter möglich, beispielsweise für den Median oder den Mittelwert (siehe später).
- Ein Schätzwert für einen wichtigen Parameter θ wird üblicherweise mit einem Konfidenzintervall ausgestattet. Ein solches gibt den möglichen Bereich für θ an.
- Je grösser die Stichprobe, desto präziser der Schätzwert und desto schmäler das Konfidenzintervall. Ein Konfidenzintervall liefert damit einen Eindruck über die Präzision des entsprechenden Schätzwerts.
- Statt ein *Konfidenzniveau* von 95% wird manchmal auch ein anderer Wert verwendet, z. B. 99%. Je grösser dieser Wert, je breiter (vorsichtiger, konservativer) ist das Konfidenzintervall. Allgemein spricht man von einem $(1 - \alpha) \cdot 100\%$ -Konfidenzintervall, wobei α ein kleiner relativer Anteil ist (oft 0.05). Die Aussagen in diesem Abschnitt können entsprechend für verschiedene α formuliert werden.
- Ein Konfidenzintervall für θ enthält stets den Schätzwert dafür. Liegt dieser genau in der Mitte des Intervalls, so spricht man von einem *symmetrischen* Konfidenzintervall.

- Nach Berechnung des 95%-Konfidenzintervalls *behauptet* man, dass sich der Parameter θ darin befindet. Diese Behauptung ist schlichtweg richtig oder falsch. Es würde dann keinen Sinn machen zu sagen: “Mit 95% Wahrscheinlichkeit liegt θ zwischen …”. Stattdessen sagt man: “Mit einer *Sicherheit* von 95% liegt θ zwischen …” oder “Wir können mit einer Sicherheit von 95% behaupten, dass …”.
- “95% Sicherheit” lässt sich folgendermassen interpretieren: Würde man viele hypothetische Stichproben ziehen, so würden die entsprechenden Konfidenzintervalle für θ variieren. Ungefähr 95% dieser Konfidenzintervalle würden den wahren Wert θ enthalten, die restlichen 5% nicht.
- Neben *zweiseitigen* Konfidenzintervallen kann es je nach Fragestellung auch sinnvoll sein, nur eine obere oder untere *Konfidenzschanke* bzw. ein einseitiges Konfidenzintervall für θ anzugeben, beispielsweise um einseitige Behauptungen über θ zu untermauern. Die beim letzten Punkt erwähnten 5% “falschen” Konfidenzintervalle liegen im Fall einer unteren 95%-Konfidenzschanke gänzlich *über* θ und im Fall einer oberen 95%-Konfidenzschanke gänzlich *unter* θ . Im Falle eines zweiseitigen 95%-Konfidenzintervalls sind je die Hälfte davon (also je 2.5%) gänzlich über bzw. unter θ . Deshalb liefert ein zweiseitiges 95%-Konfidenzintervall eine untere und obere 97.5%-Konfidenzschanke für θ . (Ein zweiseitiges 90%-Konfidenzintervall liefert eine untere und obere 95%-Konfidenzschanke.)

Konfidenzintervalle für relative Anteile

Wir haben bereits festgestellt, dass die absolute Häufigkeit $Y = n\hat{p}$ der Ausprägung 1 einer binären (0-1)-Variable in einer Zufallsstichprobe binomialverteilt ist mit Parametern n und p . Dadurch ist die Software in der Lage, *Binomialkonfidenzintervalle* (nach Clopper-Pearson) für den unbekannten relativen Anteil p zu bestimmen.

Beispiel 3.44 (Frauenanteil). Sei p der relative Frauenanteil unter den Studierenden an Schweizer Universitäten. Von $n = 263$ Befragten waren 116 weiblich. Entsprechend schätzen wir p durch die relative Häufigkeit der Frauen $\hat{p} = 116/263 \approx 44\%$. Um zu quantifizieren, wie präzise dieser Schätzwert für p ist bzw. um mögliche Werte von p zu nennen, geben wir ein zweiseitiges 95%-Konfidenzintervall für p an¹.

R Code

```
# Eingabe
binom.test(116, 263)

# Ausgabe
[...]
95 percent confidence interval: 0.3801 0.5034
sample estimates: probability of success 0.4411
```

Kommentar: Mit einer Sicherheit von 95% liegt der wahre Frauenanteil p unter Studierenden an Schweizer Universitäten zwischen 38% und 50%. ▲

Um den Begriff “Sicherheit” zu illustrieren, betrachten wir die Stichprobe von Beispiel 3.44 vorübergehend als Grundgesamtheit ($p = 0.44$) und ziehen daraus viele künstliche Zufallsstichproben je mit Umfang $m = 40$. Die Schätzwerte und damit auch die 95%-Konfidenzintervalle für p variieren von Stichprobe zu Stichprobe. Wir erwarten, dass rund 95% der Intervalle den wahren Wert p enthalten, während je etwa 2.5% zu gross bzw. zu klein sind. Abbildung 3.17 zeigt das Ergebnis bei 100 solchen künstlichen Stichproben.

Beispiel 3.45 (Umfrage vor einer Abstimmung). Sei p der relative Anteil von Befürwortern einer Initiative. Im Vorfeld der Abstimmung werden $n = 300$ Stimmberechtigte befragt, ob sie für die Initiative (Antwort

¹Dazu verwenden wir die R-Funktion `binom.test`.

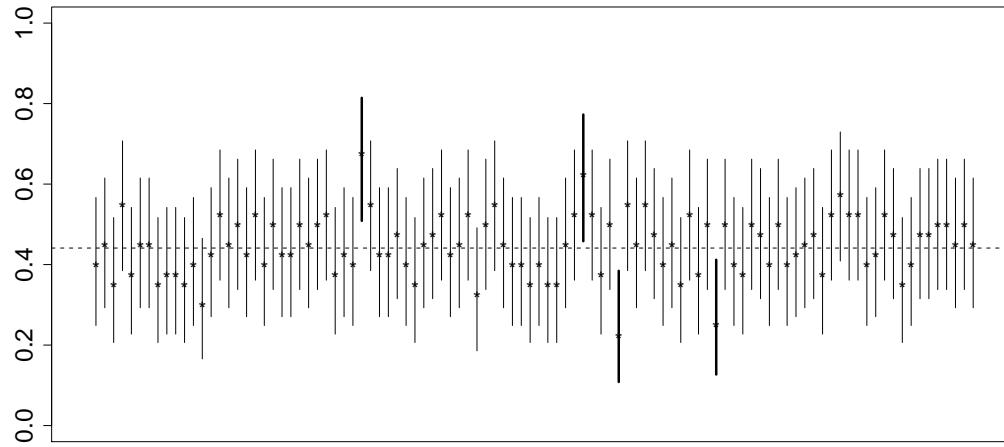


Abbildung 3.17: Aus einer Grundgesamtheit mit (bekanntem) Parameter $p = 0.44$ (gestrichelte Linie) werden 100 künstliche Zufallsstichproben vom Umfang $m = 40$ gezogen. Davon werden je Schätzwert und zweiseitiges 95%-Konfidenzintervall für p eingezeichnet. In der ersten Stichprobe waren also beispielsweise 40% Frauen. Nur in vier Fällen (fett) ist $p = 0.44$ nicht im Konfidenzintervall enthalten. Bei steigender Anzahl künstlicher Stichproben würde sich der Anteil “falscher” Konfidenzintervalle bei 5% einpendeln. Etwa 2.5% wären zu klein, etwa 2.5% zu gross.

‘1’) oder dagegen (Antwort ‘0’) sind. Es haben sich $Y = 171$ für die Initiative ausgesprochen, also

$$\hat{p} = Y/n = 171/300 = 57\%.$$

Nun ergänzen wir diesen Schätzwert mit einem 95%-Konfidenzintervall für p .

R Code

```
# Eingabe
binom.test(171, 300)

# Ausgabe
[...]
95 percent confidence interval: 0.5118521 0.6267516
sample estimates: probability of success 0.57
```

Kommentar: Wir können also mit einer Sicherheit von 95% davon ausgehen, dass zwischen 51.2% und 62.7% der Stimmberechtigten die Initiative befürworten.

Hätte man stattdessen viermal so viele Personen befragt und darunter $4 \cdot 171 = 684$ Befürworter gefunden, was den gleichen Schätzwert $\hat{p} = 57\%$ wie oben ergibt, dann ergäbe sich ein deutlich schmales Konfidenzintervall, da die Präzision des Schätzwerts für p höher wäre:

R Code

```
binom.test(4*171, 4*300) # Ergibt 0.54143 0.59823
```

Dieses Konfidenzintervall ist mit $0.598 - 0.541 = 0.057$ halb so breit wie das obere ($0.627 - 0.512 = 0.115$). Damit bestätigt sich die Feststellung von Beispiel 3.38: Eine Vervierfachung des Stichprobenumfangs führt zu doppelt so präzisen Stichprobenmittelwerten (hier: relative Häufigkeiten).

Bei 16 mal so vielen Befragten und 16 mal so vielen Befürwortern erwarten wir entsprechend ein Konfidenzintervall der Breite $0.115/4 = 0.029$. ▲

Beispiel 3.46 (Rauchen). Sei p der relative Anteil von Rauchern unter männlichen Studenten an Schweizer Universitäten. Von $n = 145$ Befragten haben $Y = 45$ angegeben zu rauchen. Ein Schätzwert für p ist damit durch die relative Häufigkeit $\hat{p} = Y/n \approx 0.31 = 31\%$ geben.

Um zu illustrieren, wie Konfidenzintervalle vom Konfidenzniveau abhängen, berechnen wir ein 90%--, ein 95%- und schliesslich ein 99%-Konfidenzintervall¹ für p .

R Code

```
# 90%-Konfidenzintervall
binom.test(45, 145, conf.level = 0.9)      # Ergibt 0.24708 0.37953

# 95%-Konfidenzintervall
binom.test(45, 145)                         # Ergibt 0.23620 0.39243

# 99%-Konfidenzintervall
binom.test(45, 145, conf.level = 0.99)        # Ergibt 0.21563 0.41787
```

Kommentar: Je höher die Sicherheit, je breiter (vorsichtiger) werden die Konfidenzintervalle. ▲

Beispiel 3.47 (“Mietfreie” Studierende). Sei p der relative Anteil von Studierenden der Universität Bern, die bei Angehörigen umsonst wohnen. Bei einer Befragung von $n = 258$ Studierenden fanden sich $Y = 129$ “mietfreie” Personen. Dies liefert den Schätzwert $\hat{p} = Y/n = 50\%$ für p . Um anhand der Daten den möglichen Bereich für p zu bestimmen, berechnen wir ein zweiseitiges 95%-Konfidenzintervall für p .

R Code

```
binom.test(129, 258)                      # Ergibt 0.43736 0.56264
```

Kommentar: Mit einer Sicherheit von 95% wohnen in Tat und Wahrheit zwischen 43.7% und 56.3% der Berner Studierenden umsonst.

Um zu unterstreichen, dass der Anteil gross ist, berechnen wir eine untere 95%-Konfidenzschranke² für p .

R Code

```
binom.test(129, 258, alternative = 'greater')    # Ergibt 0.44706 1.00000
```

Kommentar: Mit einer Sicherheit von 95% wohnen mehr als 44.7% aller Berner Studierenden bei Angehörigen umsonst. Die Zahl 0.447 ist die *untere Konfidenzschranke*, während der Bereich $[0.447, 1]$ das *einseitige Konfidenzintervall* ist.

Die untere 95%-Konfidenzschranke finden wir alternativ via zweiseitigem 90%-Konfidenzintervall:

R Code

```
binom.test(129, 258, conf.level = 0.9)          # Ergibt 0.44706 0.55294
```

▲

Um den Begriff “Sicherheit” auch in einer einseitigen Situation zu illustrieren, fassen wir die Befragten in Beispiel 3.47 vorübergehend als Grundgesamtheit mit wahrem Anteil $p = 0.5$ auf und ziehen 100 künstliche Zufallsstichproben je mit Umfang $m = 40$ daraus. Wir erwarten, dass etwa fünf Konfidenzintervalle gänzlich über $p = 0.5$ liegen, siehe Abbildung 3.18.

¹Option `conf.level`.

²Option `alternative = 'greater'`.

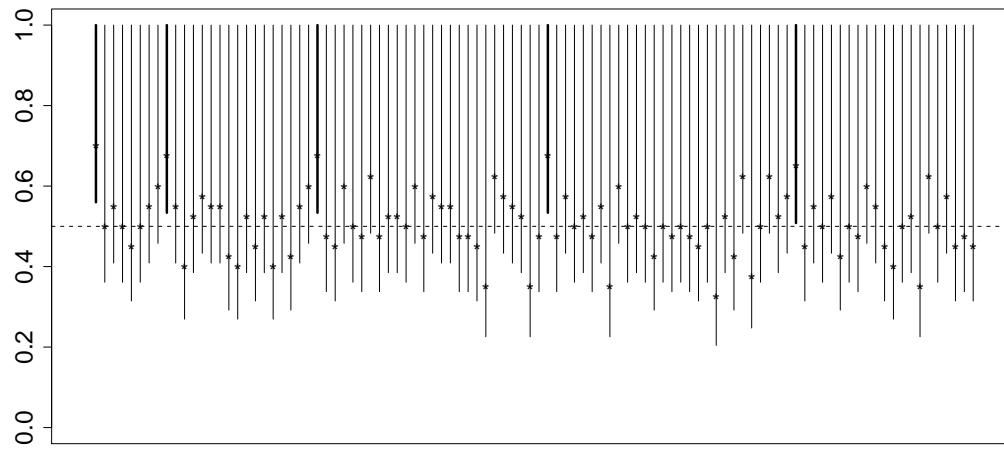


Abbildung 3.18: Aus einer Grundgesamtheit mit bekanntem Parameter $p = 0.5$ (gestrichelte Linie) werden 100 künstliche Zufallsstichproben vom Umfang $m = 40$ gezogen. Die so entstehenden Schätzwerte und einseitigen 95%-Konfidenzintervalle werden pro Stichprobe eingezeichnet. Fünf Intervalle (fett) enthalten den wahren Wert $p = 0.5$ nicht. Dies entspricht der Erwartung.

Beispiel 3.48 (Rauchen, Forts.). Betrachten wir nun das kategoriale Merkmal ‘Rauchen’ aus dem Datensatz von Beispiel 1.1. Was können wir damit über das Rauchverhalten aller Studenten und Studentinnen der Schweiz sagen?

Ein Schätzwert für den Anteil p_1 der NichtraucherInnen ist gegeben durch die relative Häufigkeit $\hat{p}_1 = 171/261 \approx 65.5\%$. Entsprechend erhalten wir für den Anteil p_2 der GelegenheitsraucherInnen den Schätzwert $\hat{p}_2 = 47/261 \approx 18.0\%$. Den wahren Anteil p_3 der regelmässigen RaucherInnen schätzen wir analog auf $\hat{p}_3 = 43/261 \approx 16.5\%$.

Statten wir nun diese Schätzwerte mit 95%-Konfidenzintervallen aus:

R Code

```
# NichtraucherInnen
binom.test(171, 261)                                # Ergibt 0.59408 0.71269

# GelegenheitsraucherInnen
binom.test(47, 261)                                  # Ergibt 0.13542 0.23217

# Regelmässige RaucherInnen
binom.test(43, 261)                                  # Ergibt 0.12187 0.21543
```

Solche Ergebnisse werden manchmal zu einer Tabelle zusammengefasst:

Parameter	Schätzwert	95%-Konfidenzintervall
p_1	0.655	[0.594, 0.713]
p_2	0.180	[0.135, 0.232]
p_3	0.165	[0.122, 0.215]

Wir können für jeden der drei unbekannten Anteile mit einer Sicherheit von 95% behaupten, dass er im entsprechenden Konfidenzintervall liegt. Da sich die Unsicherheiten von je 5% im schlimmsten Fall aufaddieren, kann man insgesamt nur mit einer Sicherheit von $100\% - 3 \cdot 5\% = 85\%$ behaupten, dass *alle* drei Parameter in den angegebenen Konfidenzintervallen liegen. Wollte man letztere Behauptung mit einer Sicherheit von 95% (bzw. einer Unsicherheit von 5%) machen, könnte entsprechend mit Konfidenzniveaus von $100\% - 5\%/3 = 98.333\%$ gearbeitet werden. Mehr dazu sehen wir später bei *multiplen Tests*. ▲

3.2.3 Tests allgemein und für relative Anteile

Während mit Schätzern und Konfidenzintervallen quantitative Fragen über die Population beantwortet werden (“Wie gross ist der Wähleranteil der Partei ABC?”), beantworten Tests Ja/Nein-Fragen (“Ist der Wähleranteil gestiegen?”).

Wir führen das Konzept des Testens schrittweise anhand des Binomialtests ein, mit dem Fragestellungen zu relativen Anteilen beantwortet werden. Er ist eng mit den Binomialkonfidenzintervallen verwandt. Später werden wir weitere Tests kennenlernen.

Schritt 1: Formulierung von Arbeits- und Nullhypothese

Als erstes wird die Fragestellung als Behauptung über die Population formuliert, von der man hofft, dass sie zutrifft (“Der Wähleranteil ist gestiegen”). Neben dieser sogenannten *Arbeitshypothese* (auch Alternativhypothese, H_A oder H_1 genannt) wird die gegenteilige *Nullhypothese* (H_0) formuliert (“Der Wähleranteil ist nicht gestiegen”), von der man hofft, dass sie nicht gilt.

Beispiel 3.49 (Wahlprognosen). Eine politische Partei ABC möchte wissen, ob ihr Wähleranteil gegenüber der letzten Wahl gestiegen ist. Damals war er 20%. Sei also p der relative Anteil von (potenziellen) ABC-Wählern unter allen Wahlberechtigten. Um etwas über p zu erfahren, werden $n = 500$ Wahlberechtigte befragt, ob sie derzeit Partei ABC wählen würden. $Y = 125$ Personen haben die Frage bejaht. Ein naheliegender Schätzwert für p ist dann $\hat{p} = Y/n = 125/500 = 0.25$, also der relative Anteil von ABC-Wählern in der Stichprobe.

Die Partei *hofft*, dass $p > 0.2$; dies ist die Arbeitshypothese H_1 . Das Gegenteil, also $p \leq 0.2$, definiert die Nullhypothese H_0 . ▲

Schritt 2: Evidenz gegen Nullhypothese messen

Als zweites wird mithilfe der Daten die Evidenz gegen die Nullhypothese (und damit indirekt für die Arbeitshypothese) gemessen. Dies geschieht mit der zum Test gehörenden *Teststatistik* T – eine Kenngrösse, deren Wert bei zunehmender Evidenz gegen H_0 grösser bzw. kleiner wird. Bei den meisten Tests ist die Teststatistik eine Funktion des entsprechenden Schätzers, wobei die Funktion so gewählt ist, dass die Verteilung von T unter der Nullhypothese (also falls H_0 stimmt) verfügbar ist.

Ein stets gleich definiertes Mass der Evidenz gegen die Nullhypothese ist der *p-Wert*. Diese “standardisierte” Teststatistik gibt die Wahrscheinlichkeit an, dass unter H_0 (d. h. durch reines Glück) mindestens ebenso viel gegen H_0 spricht wie in den konkret vorliegenden Daten.

Hinweise zum p-Wert

- *p*-Werte werden anhand der Daten und der Verteilung der Teststatistik T unter H_0 berechnet. Nur für wenige Tests gelingt dies von Hand, so dass man auf entsprechende Software angewiesen ist.
- Je kleiner der *p*-Wert, je mehr Evidenz gegen H_0 bzw. für H_1 .
- Der *p*-Wert gibt *nicht* die Wahrscheinlichkeit an, dass die Nullhypothese stimmt.
- Das “p” in “*p*-Wert” hat nichts mit dem Parameter p der Bernoulli- oder Binomialverteilung zu tun.

Beispiel 3.50 (Wahlprognosen, Fortsetzung). Die Teststatistik des Binomialtests ist die absolute Häufigkeit $T = n\hat{p}$, hier also die Anzahl der ABC-Wähler. Je höher ihr konkreter Wert (hier 125) ist, desto mehr spricht

gegen $H_0: p \leq 0.2$. Wir wissen, dass T unter der Nullhypothese $\text{Bin}(500, 0.2)$ -verteilt ist. Der p -Wert ist also definitionsgemäß¹

$$P(T \geq 125) = 1 - F(124) = 0.0037,$$

wobei hier F die Verteilungsfunktion von $\text{Bin}(500, 0.2)$ bezeichnet. Abbildung 3.19 zeigt unter anderem den p -Wert als schraffierte Bereiche der Wahrscheinlichkeitsfunktion von T unter der Nullhypothese.

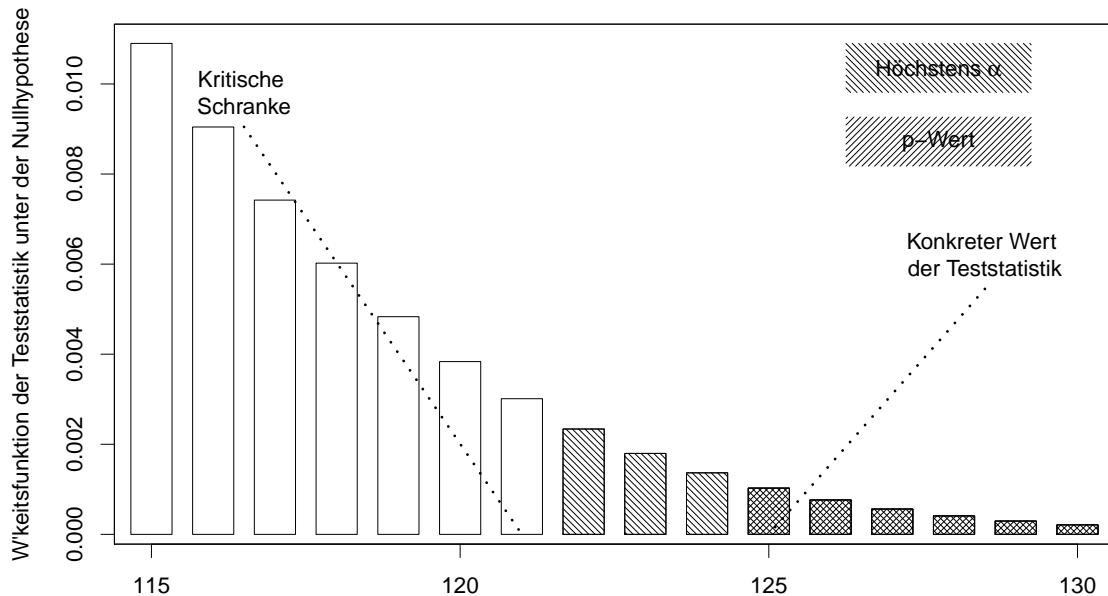


Abbildung 3.19: Verteilung der Teststatistik unter der Nullhypothese aus den Beispielen 3.50 und 3.51 inklusive p -Wert und Signifikanzniveau $\alpha = 0.01$ als schraffierte Bereiche der Wahrscheinlichkeitsfunktion.

Den gleichen Wert erhalten wir, indem wir den Binomialtest anfordern²:

R Code

```
binom.test(125, 500, p = 0.2, alternative = 'greater') # Ergibt p-value = 0.0037
```



Schritt 3: Testentscheid

Im finalen Schritt prüft man, ob genug Evidenz gegen H_0 vorliegt und beantwortet damit die ursprüngliche Fragestellung.

Liegt der p -Wert unterhalb einer vor der Analyse fixierten kleinen Schranke α (oft 5% oder 1%), wird die Nullhypothese auf dem α -(Signifikanz-)Niveau zugunsten der Arbeitshypothese abgelehnt/verworfen. Mit einer Sicherheit von $1 - \alpha$ können wir dann behaupten, dass die Arbeitshypothese stimmt³. Salopp spricht man von einem “signifikanten Ergebnis”.

Liegt der p -Wert nicht unterhalb α , so wird die Nullhypothese auf dem α -Niveau nicht abgelehnt; wir haben keinen Grund, an H_0 zu zweifeln. **Dies bedeutet jedoch nicht, dass H_0 stimmt!** Salopp spricht man dann von einem “nichtsignifikanten Ergebnis”.

¹In R mit $1 - \text{pbinom}(124, 500, 0.2)$.

²Binomialtests werden in R mit der Funktion `binom.test` durchgeführt. Dabei müssen die Hypothesen spezifiziert werden: Um $H_0: p = p_o$ vs. $H_1: p \neq p_o$ zu prüfen, wird $p = p_o$ gesetzt. Für $H_0: p \leq p_o$ vs. $H_1: p > p_o$ wird $p = p_o$ und `alternative = 'greater'` spezifiziert. Für $H_0: p \geq p_o$ vs. $H_1: p < p_o$ entsprechend $p = p_o$ und `alternative = 'less'`.

³Diese Behauptung dürfen wir nur dann machen, wenn die Nullhypothese das genaue Gegenteil der Arbeitshypothese ist. (Es ist grundsätzlich auch möglich, Hypothesen der Art $H_0: p \leq 0.5$ vs. $H_1: p > 0.7$ zu prüfen, obwohl wir dies in der Vorlesung nicht tun.)

Zum gleichen Entscheid gelangt man, indem man prüft, ob die Teststatistik T einen unter der Nullhypothese unwahrscheinlich grossen bzw. kleinen Wert annimmt. Als kritische Schranke dient dann sinnvollerweise ein entsprechend hohes bzw. kleines Quantil der Verteilung von T unter der Nullhypothese.

Beispiel 3.51 (Wahlprognosen, Fortsetzung). Die Partei möchte die Nullhypothese $p \leq 0.2$ versus die Arbeitshypothese $p > 0.2$ auf dem Niveau $\alpha = 0.01$ testen.

Der p -Wert 0.0037 von Beispiel 3.50 ist kleiner als das Niveau 0.01, deshalb verwerfen wir die Nullhypothese zugunsten der Arbeitshypothese: Wir können mit einer Sicherheit von 99% behaupten, dass der tatsächliche Wähleranteil p von ABC grösser ist als 20%.

Zum gleichen Testentscheid gelangen wir mithilfe der Teststatistik, deren Wert 125 grösser als die kritische Schranke 121 ist. Diese entspricht hier dem 99%-Quantil¹ von $\text{Bin}(500, 0.2)$.

Abbildung 3.19 zeigt die Situation schematisch. ▲

Beispiel 3.52 (Qualitätskontrolle). Eine Firma produziert einen Massenartikel und prüft regelmässig, ob die Produktionsanlage zu viel Ausschuss liefert. Bei jedem produzierten Stück besteht eine gewisse Wahrscheinlichkeit, dass es fehlerhaft ist. Wir betrachten die von nun an produzierten Artikel und setzen

$$Z_i := \begin{cases} 1 & \text{wenn der } i\text{-te Artikel fehlerhaft ist,} \\ 0 & \text{sonst.} \end{cases}$$

Wir nehmen an, dass diese Beobachtungen unabhängig sind mit unbekanntem (und hoffentlich kleinem) Parameter $p := P(Z_i = 1)$. Dieser ist also die Ausschusswahrscheinlichkeit für ein einzelnes Teil.

Angenommen, die Kunden des Unternehmens verlangen eine Ausschussrate von höchstens 0.08. Um dies zu verifizieren, überprüft die Firma hin und wieder $n = 100$ Teile. Nun testet man auf dem 5%-Niveau die Nullhypothese, dass $p \geq 0.08$. Wenn diese Nullhypothese verworfen wird, kann die Firma mit einer Sicherheit von 95% davon ausgehen, dass die Produktionsanlage die Qualitätsanforderung der Kunden übertrifft, also dass die Arbeitshypothese $p < 0.08$ stimmt.

Bei der letzten Kontrolle wurden 2 defekte Teile gefunden. Der p -Wert entspricht dann der Wahrscheinlichkeit, dass (obwohl die Nullhypothese stimmt) in einer neuen Stichprobe ebenfalls nur zwei oder sogar noch weniger defekte Teile sind, also dass eine $\text{Bin}(100, 0.08)$ -verteilte Zufallsvariable X höchstens zwei beträgt:

$$\begin{aligned} P(X \leq 2) &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= \binom{100}{0} 0.08^0 \cdot 0.92^{100} + \binom{100}{1} 0.08^1 \cdot 0.92^{99} + \binom{100}{2} 0.08^2 \cdot 0.92^{98} \approx 0.01127. \end{aligned}$$

Der p -Wert 0.011 ist kleiner als das Niveau 0.05, also verwerfen wir die Nullhypothese zugunsten der Arbeitshypothese. Mit einer Sicherheit von 95% können wir behaupten, dass die tatsächliche Ausschussrate kleiner als 8% ist. Zum gleichen Testentscheid kommen wir via Teststatistik: Ihr Wert 2 ist kleiner als das 5%-Quantil 4 von $\text{Bin}(100, 0.08)$, siehe Abbildung 3.20.

Die Software bestätigt unsere Berechnungen:

<pre>binom.test(2, 100, p = 0.08, alternative = 'less')</pre>	R Code	# Ergibt p-value = 0.01127
---	--------	----------------------------

Auf dem 1%-Niveau würde unsere Nullhypothese nicht verworfen, da dann der p -Wert 0.01127 nicht kleiner als 0.01 bzw. der Wert 2 der Teststatistik nicht kleiner als das 1%-Quantil 2 von $\text{Bin}(100, 0.08)$ wäre. ▲

¹In R erhalten wir es mit `qbinom(0.99, 500, 0.2)`.

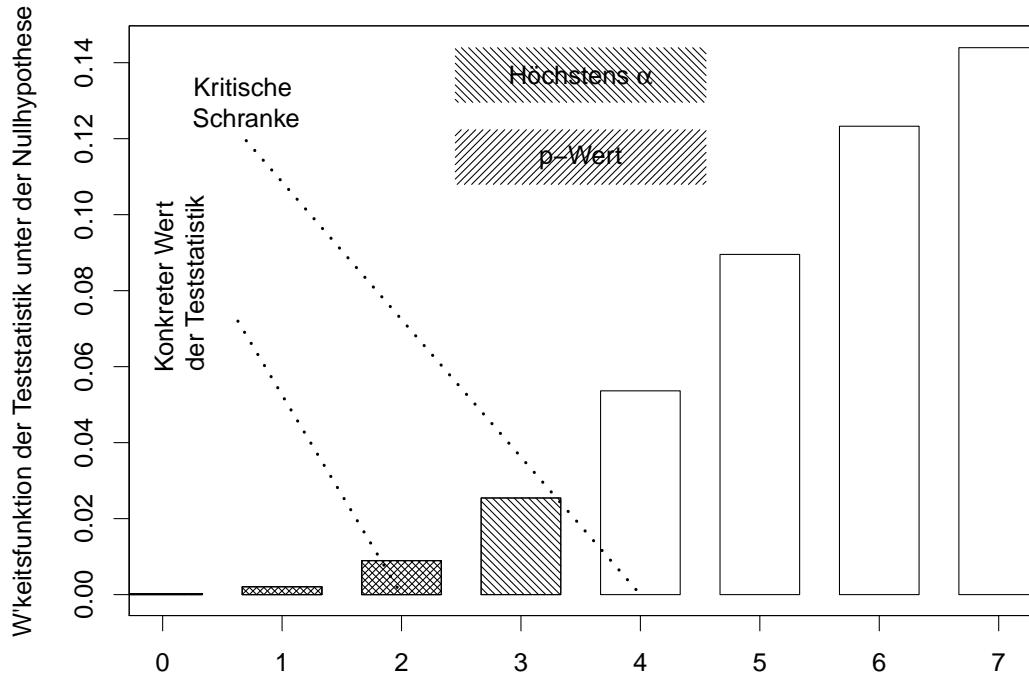


Abbildung 3.20: Verteilung der Teststatistik aus Beispiel 3.52 unter der Nullhypothese inklusive p -Wert und Signifikanzniveau $\alpha = 0.05$ als schraffierte Bereiche der Wahrscheinlichkeitsfunktion.

Ein- und zweiseitige Tests

Je nach Fragestellung wird ein einseitiger ($H_1: \theta < \theta_o$ bzw. $H_1: \theta > \theta_o$) oder zweiseitiger ($H_1: \theta \neq \theta_o$) Test durchgeführt.

Beispiel 3.53 (Wahlprognosen, Fortsetzung). In Beispiel 3.51 hat die Partei mithilfe einer Umfrage auf dem 1%-Niveau gezeigt, dass der tatsächliche aktuelle Wähleranteil p gegenüber der letzten Wahl gestiegen ist: Die Nullhypothese $p \leq 0.2$ wurde zugunsten der Arbeitshypothese $p > 0.2$ verworfen.

Mit den gleichen Daten könnte eine Politologin prüfen, ob sich der Wähleranteil dieser Partei verändert hat: Die Politologin prüft dann die Arbeitshypothese $p \neq 0.2$ versus die Nullhypothese $p = 0.2$.

R Code	# Ergibt p-value = 0.006126
binom.test(125, 500, 0.2)	

Kommentar: Auf dem 1%-Niveau verwirft sie die Nullhypothese zugunsten der Arbeitshypothese. Mit einer Sicherheit von 99% hat sich der wahre Wähleranteil verändert.

Betrachten wir schliesslich eine andere Partei, die sich freut, falls ABC Wähleranteil eingebüsst hat. Sie prüft die Arbeitshypothese $p < 0.2$ versus die Nullhypothese $p \geq 0.2$.

R Code	# Ergibt p-value = 0.9973
binom.test(125, 500, 0.2, alternative = 'less')	

Kommentar: Der p -Wert ist nicht kleiner als 0.01, deshalb kann diese Nullhypothese nicht verworfen werden. ▲

Vorsichtige Aussagen

Gilt die Behauptung der Arbeitshypothese nicht einmal in der Stichprobe, so lässt sich erst recht nicht behaupten, dass sie in der Population gilt. In solchen Situationen sind p -Werte entsprechend gross.

Beispiel 3.54 (Wahlprognosen, Fortsetzung). In Beispiel 3.53 wollte eine andere Partei die Arbeitshypothese nachweisen, dass der wahre Wähleranteil p von Partei ABC unter 20% gesunken ist. Da der Anteil der ABC-Wähler in der Stichprobe mit $\hat{p} = 25\%$ jedoch nicht kleiner als 20% ist, lässt sich erst recht nicht behaupten, dass dies in der Population gilt. ▲

Konfidenzintervallansatz

Da die Antwort auf eine Ja/Nein-Frage ("Ist der Wähleranteil grösser als 20%?") stets aus der Antwort einer entsprechenden quantitativen Frage ("Wie gross ist der Wähleranteil der Partei ABC?") folgt, lassen sich Testentscheide nicht nur mit p -Werten oder Teststatistiken fällen, sondern auch mit passenden Konfidenzintervallen. Aufgrund der quantitativen Aussage des Konfidenzintervalls ist dieser sogenannte *Konfidenzintervallansatz* besonders informativ. Tests können damit als Anwendung von Konfidenzintervallen angesehen werden.

Beispiel 3.55 (Wahlprognosen, Fortsetzung). Treffen wir die Testentscheide von Beispiel 3.53 nun mithilfe des Konfidenzintervallansatzes.

Aus Sicht der Partei ABC: Die Partei möchte auf dem 1%-Niveau die Arbeitshypothese nachweisen, dass der wahre Anteil p grösser als 0.2 ist. Dazu bestimmt sie eine untere 99%-Konfidenzschanke für p . Falls diese grösser als 0.2 ist, wird die Arbeitshypothese angenommen.

	R Code	
	binom.test(125, 500, alternative = 'greater', conf.level = 0.99) # Ergibt 0.2061 1.0000	

Kommentar: Mit einer Sicherheit von 99% ist der wahre Anteil p grösser als 20.6%. Die vorsichtigere Aussage, dass p grösser als 20% ist, gilt damit auch mit (mindestens) 99% Sicherheit.

Aus Sicht der Politologin: Die Politologin möchte auf dem 1%-Niveau die Arbeitshypothese prüfen, dass p ungleich 0.2 ist. Dazu bestimmt sie ein 99%-Konfidenzintervall für p . Enthält dieses den Wert 0.2 nicht, so wird die Arbeitshypothese angenommen.

	R Code	
	binom.test(125, 500, conf.level = 0.99) # Ergibt 0.2017 0.3031	

Kommentar: Mit einer Sicherheit von 99% liegt der wahre Anteil p zwischen 20.17% und 30.3%. Deshalb lässt sich auch mit einer Sicherheit von (mindestens) 99% sagen, dass der Anteil nicht 20% ist.

Aus Sicht einer anderen Partei: Eine andere Partei möchte auf dem 1%-Niveau die Arbeitshypothese prüfen, dass die Partei ABC Wähleranteil verloren hat, also dass $p < 0.2$. Dazu bestimmt sie eine obere 99%-Konfidenzschanke für p . Ist diese kleiner als 0.2, so wird die Arbeitshypothese angenommen.

	R Code	
	binom.test(125, 500, alternative = 'less', conf.level = 0.99) # Ergibt 0.0000 0.2979	

Kommentar: Mit einer Sicherheit von 99% ist p kleiner als 29.8%. Es gibt damit keinen Grund zu behaupten, dass p sogar kleiner als 20% ist.

In allen drei Situationen sind wir zu den gleichen Testentscheiden gelangt wie in Beispiel 3.53. ▲

Fehler erster und zweiter Art

Bei der Durchführung eines Tests riskiert man immer einen Fehler erster Art (man lehnt eine wahre Nullhypothese ab) oder einen Fehler zweiter Art (man lehnt eine falsche Nullhypothese nicht ab).

In einer konkreten Anwendung kann man nicht sagen, ob und welchen Fehler man begangen hat. Wenn man aber in sehr vielen (unabhängigen) Situationen einen Test mit Signifikanzniveau α anwendet, so begeht man in höchstens etwa $\alpha \cdot 100\%$ aller Fälle einen Fehler erster Art. Im Gegensatz dazu ist die Wahrscheinlichkeit eines Fehlers zweiter Art in der Regel unbekannt, sinkt jedoch bei wachsendem Stichprobenumfang. Deren Gegenwahrscheinlichkeit wird *Power (Güte)* des Tests genannt.

Beispiel 3.56 (Qualitätskontrolle, Fortsetzung). Die Firma in Beispiel 3.52 prüft regelmässig, ob die versprochenen Qualitätsanforderungen erfüllt sind. Zu diesem Zweck untersucht sie pro Grosslieferung jeweils $n = 100$ Stück und prüft damit jeweils auf dem 5%-Niveau die Arbeitshypothese, dass die tatsächliche Ausschussrate p in der Lieferung kleiner als 8% ist, versus die Nullhypothese, dass $p \geq 0.08$.

Von den Lieferungen mit zu hoher Ausschussrate werden nur rund 5% per Zufall als gut befunden. ▲

Unter- bzw. überpowerte Tests

Der Stichprobenumfang beeinflusst den Fehler zweiter Art und den Testentscheid massiv: Bei kleinen Stichproben kann es passieren, dass die Daten rein deskriptiv deutlich gegen die Nullhypothese sprechen, diese jedoch aufgrund der hohen Wahrscheinlichkeit eines Fehlers zweiter Art nicht verworfen wird (unterpowerter Test). Auf der anderen Seite wird bei sehr grossen Stichprobenumfängen jede deskriptiv noch so schwache Evidenz gegen die Nullhypothese "signifikant" (überpowerter Test). **Ein signifikantes Resultat kann also ohne Relevanz sein und umgekehrt.**

Konfidenzintervalle weisen diese konzeptuellen Probleme von Tests bzw. von Ja/Nein-Fragen nicht auf.

Beispiel 3.57 (Wahlprognose, Fortsetzung). Angenommen, von 10'000 Befragten würden 21.2% ABC wählen. In der Stichprobe ist die Steigerung des Wähleranteils von 20% auf $\hat{p} = 21.2\%$ nur sehr klein. Ein Test der Arbeitshypothese $H_1 : p > 0.2$ ergibt dennoch ein auf dem 1%-Niveau signifikantes Ergebnis, da der p -Wert mit 0.0015 kleiner als 0.01 ist. Anhand der unteren Konfidenzschanke folgt der gleiche Testentscheid. Man sieht dort jedoch, dass evtl. nur eine minime Steigerung des wahren Wähleranteils vorliegt.

R Code

```
# Eingabe
binom.test(0.212*10000, 10000, 0.2, alternative = 'greater', conf.level = 0.99)

# Ausgabe
[...]
p-value = 0.001497
99 percent confidence interval: 0.20256 1.00000
```

Würden bei einer Miniumfrage fünf von zehn Befragten angeben, ABC zu wählen, liegt deskriptiv eine massive Steigerung des Wähleranteils vor. Der entsprechende Test verwirft die Nullhypothese von keiner Steigerung jedoch nicht.

R Code

```
# Eingabe
binom.test(5, 10, 0.2, alternative = 'greater', conf.level = 0.99)

# Ausgabe
```

```
[...]
p-value = 0.03279
99 percent confidence interval: 0.1504428 1.0000000
```



Multiples Testen

Bei jedem Test kann ein Fehler erster oder zweiter Art passieren. Läuft die Beantwortung einer Fragestellung auf mehrere (m) Tests je auf dem Niveau α hinaus, so kumulieren sich die Fehlerwahrscheinlichkeiten im schlimmsten Fall. Testet man also je auf dem Niveau α , so ist die Wahrscheinlichkeit eines Fehlers erster Art nicht mehr auf α , sondern auf $m\alpha$ beschränkt. Eine einfache Möglichkeit, diese Wahrscheinlichkeit insgesamt auf α zu beschränken, stellt die sogenannte Bonferroni-Korrektur dar: Als Signifikanzniveau wird die strengere Schranke α/m verwendet (oder die p -Werte mal m gerechnet).

Beispiel 3.58 (Rauchen). Betrachten wir die Situation in Beispiel 3.48. Dort haben wir mithilfe von Daten Rückschlüsse auf das tatsächliche Rauchverhalten bei StudentInnen gemacht. Nun wollen wir mit den gleichen Daten auf dem 5%-Niveau prüfen, ob sich deren Rauchverhalten von jenem der Schweizer Gesamtbevölkerung unterscheidet.

Aus einer grossen Erhebung sei bekannt, dass in der Gesamtbevölkerung $p_1^o = 0.59 = 59\%$ nicht rauchen, $p_2^o = 0.21 = 21\%$ gelegentlich rauchen und $p_3^o = 0.2 = 20\%$ regelmässig rauchen. Nun prüfen wir für jeden dieser drei Anteile p_j^o mit einem separaten Binomialtest die Arbeitshypothese, dass er sich vom entsprechenden unbekannten Anteil p_j bei StudentInnen unterscheidet. Um die Wahrscheinlichkeit eines Fehlers erster Art der eigentlichen Fragestellung insgesamt auf 5% zu beschränken, prüfen wir die drei Hypothesen je auf dem strengeren Signifikanzniveau von $0.05/3 = 0.0167$. Sobald einer dieser drei Tests einen p -Wert unterhalb dieser Schranke ergibt, behaupten wir dann mit einer Sicherheit von 95%, dass sich das Rauchverhalten bei StudentInnen tatsächlich von jenem der Gesamtbevölkerung unterscheidet.

R Code

```
# NichtraucherInnen
binom.test(171, 261, p = 0.59) # Ergibt p-value = 0.03245

# GelegenheitsraucherInnen
binom.test(47, 261, p = 0.21) # Ergibt p-value = 0.2547

# Regelmässige RaucherInnen
binom.test(43, 261, p = 0.2) # Ergibt p-value = 0.164
```

Kommentar: Keiner der p -Werte ist kleiner als 0.0167, deshalb können wir auf dem 5%-Niveau nicht behaupten, dass sich das Rauchverhalten bei StudentInnen von jenem der Gesamtbevölkerung unterscheidet. Ohne Bonferroni-Korrektur für multiples Testen wären wir zu einem anderen Testentscheid gekommen. ▲

Äquivalenztests

Mit Tests lassen sich üblicherweise Arbeitshypothesen der Form $\theta > \theta_o$, $\theta < \theta_o$ oder $\theta \neq \theta_o$ prüfen. Möchte man jedoch nachweisen, dass der Parameter θ gleich dem fixen Wert θ_o ist, sind nur indirekte Vorgehensweisen möglich. Beispielsweise könnte man prüfen, ob ein zweiseitiges Konfidenzintervall für θ gänzlich innerhalb eines kleinen (vor der Analyse spezifizierten) Toleranzbereichs um θ_o liegt oder nicht.

Beispiel 3.59 (Wahlprognosen, Fortsetzung). Möchte die Partei zeigen, dass ihr Wähleranteil unverändert auf 20% geblieben ist, wählt sie zuerst einen Toleranzbereich von beispielsweise 18% bis 22%. Anteile p in diesem Bereich gelten als “gleich” wie 20%. Das zweiseitige 99%-Konfidenzintervall für p beträgt:

binom.test(125, 500, conf.level = 0.99)	R Code
	# Ergibt 0.2017 0.3031

Kommentar: Mit einer Sicherheit von 99% können wir also behaupten, dass p zwischen 0.2017 und 0.3031 liegt. Dieses Intervall ist nicht im Toleranzbereich enthalten, somit können wir nicht behaupten, dass der Anteil gleich geblieben ist. \blacktriangle

Hinweis (Häufiger Fehler). In der Praxis ist in diesem Zusammenhang folgendes unsinniges Verfahren häufig anzutreffen: Man spezifiziert $H_1 : \theta \neq \theta_o$ versus $H_o : \theta = \theta_o$. Wird die Nullhypothese nicht verworfen, behauptet man fälschlicherweise, dass sie stimmt.

3.2.4 Konfidenzintervalle für Quantile

Basierend auf der Binomialverteilung lassen sich Konfidenzintervalle für ein Populationsquantil bestimmen. Die Idee beruht auf folgenden Überlegungen:

1. Betrachte n Beobachtungen X_1, \dots, X_n mit unbekanntem β -Quantil q_β .
2. Gesucht sind zwei Stichprobenwerte $X_{(k)}$ und $X_{(\ell)}$, so dass $P(X_{(k)} \leq q_\beta \leq X_{(\ell)}) \geq 1 - \alpha$. Diese Werte bilden die Schranken des $(1 - \alpha) \cdot 100\%$ -Konfidenzintervalls. (Indizes in runden Klammern beziehen sich auf die sortierten Beobachtungen, $X_{(k)}$ ist also die k kleinste Beobachtung.)
3. Aufgrund der Definition eines Quantils nimmt X_i mit Wahrscheinlichkeit β einen Wert von höchstens q_β an. Die Anzahl T von Beobachtungen, die höchstens q_β betragen, ist somit binomialverteilt mit Parametern n und β .
4. Die Ungleichung $X_{(k)} \leq q_\beta \leq X_{(\ell)}$ ist sicher erfüllt, wenn mindestens k , aber höchstens $\ell - 1$ Beobachtungen kleiner oder gleich q_β sind, also wenn $k \leq T \leq \ell - 1$. Die Wahrscheinlichkeit davon beträgt

$$P(X_{(k)} \leq q_\beta \leq X_{(\ell)}) \geq \sum_{j=k}^{\ell-1} P(T = j) = F(\ell - 1) - F(k - 1).$$

F bezeichnet hier die Verteilungsfunktion von $\text{Bin}(n, \beta)$.

5. Mithilfe der entsprechenden Quantilfunktion F^{-1} bestimmt man nun k und ℓ so, dass $F(\ell - 1) - F(k - 1)$ mindestens $1 - \alpha$ beträgt. Für ein zweiseitiges Konfidenzintervall $[X_{(k)}, X_{(\ell)}]$ setzt man dazu $k = F^{-1}(\alpha/2)$ und $\ell = F^{-1}(1 - \alpha/2) + 1$, für eine untere Schranke $k = F^{-1}(\alpha)$ und für eine obere schliesslich $\ell = F^{-1}(1 - \alpha) + 1$. Da k auch den Wert 0 und ℓ auch den Wert $n + 1$ annehmen können, setzen wir vorsichtigerweise $X_{(0)}$ auf minus unendlich und $X_{(n+1)}$ auf plus unendlich.

Mit dem Konfidenzintervallansatz lassen sich Hypothesen über q_β prüfen. Tests für den Median heissen *Medianetest* oder *Vorzeichentest*.

Beispiel 3.60 (Mieterverband). In Beispiel 2.16 haben wir einen Datensatz mit 76 Wohnungen betrachtet. Der Stichprobenmedian 1155 CHF des Mietpreises dient als Schätzwert für den wahren Median $Q_{0.5}$.

Der folgende Code berechnet ein 95%-Konfidenzintervall für den Median:

k <- qbinom(0.025, 76, 0.5) ell <- qbinom(0.975, 76, 0.5) + 1 sort(wohnungen\$Preis)[c(k, ell)]	R Code
	# Ergibt 1093 1270

Kommentar: Mit einer Sicherheit von 95% liegt $Q_{0.5}$ zwischen 1093 und 1270 CHF.

Möchte ein Mieterverband beispielsweise die Arbeitshypothese nachweisen, dass der typische Mietpreis sehr hoch ist, nämlich grösser als 1000 CHF, könnte er eine untere Konfidenzschanke für den wahren Median berechnen und dann prüfen, ob diese den Wert 1000 überschreitet:

R Code

```
k <- qbinom(0.05, 76, 0.5)
sort(wohnungen$Preis) [k] # Ergibt 1100
```

Kommentar: Mit einer Sicherheit von 95% beträgt $Q_{0.5}$ mindestens 1100 CHF. Diese Schranke ist grösser als 1000, somit kann mit einer Sicherheit von 95% behauptet werden, dass die Arbeitshypothese stimmt. ▲

Beispiel 3.61 (Portfoliorenditen). Oft wird das Risiko einer Finanzanlage, also z. B. der 5%-Value at Risk (minus das 5%-Quantil der Verteilung des Gewinns), aus früheren Daten geschätzt und man möchte etwas über das künftige Risiko sagen. Deshalb ist es vernünftig, den Schätzwert mit einer oberen Konfidenzschanke auszustatten. (Also interessieren wir uns für eine untere Konfidenzschanke für das 5%-Quantil der Gewinnverteilung.)

Betrachten wir eine kleine Stichprobe von Portfoliorenditen aus Beispiel 3.33: Die letzten 24 Monatsrenditen des Portfolios wären folgende sortierten Prozentwerte gewesen:

-1.08	-0.75	-0.55	-0.52	-0.37	-0.36	-0.36	-0.34	-0.23	0.01	0.02	0.07
0.20	0.43	0.57	0.75	0.81	0.83	0.90	0.97	1.02	1.16	1.17	1.37

Deren 5%-Quantil beträgt -0.72 , der Value at Risk entsprechend 0.72.

Da das 10%-Quantil von $\text{Bin}(24, 0.05)$ null beträgt, ist die untere 90%-Konfidenzschanke für das 5%-Quantil minus unendlich. Wir können auf dem 10%-Niveau somit nicht ausschliessen, dass der wahre Value at Risk unendlich ist (unendlich riskante Anlage). Dies illustriert, wie gefährlich es sein kann, wichtige Entscheide lediglich basierend auf Schätzwerten von kleinen Stichproben zu fällen. ▲

3.2.5 Poissonverteilung

Wir beschliessen den Teil zur Binomialverteilung mit der eng damit verbundenen *Poissonverteilung*. Diese diskrete Verteilung ist wichtig in der Statistik, da die absolute Häufigkeit von seltenen, unabhängigen Ereignissen etwa poissonverteilt ist. Beispiele von poissonverteilten Zufallsvariablen sind:

- Anzahl Hausbrände, die eine Versicherung pro Jahr bezahlen muss.
- Anzahl Flugzeugabstürze pro Jahr.
- Merkmale, die eine Häufigkeit angeben, beispielsweise die Anzahl Krankenkassenrechnungen pro Person und Jahr oder die Anzahl Autounfälle pro Person und Jahr.

Die Wahrscheinlichkeitsfunktion der Poissonverteilung mit Parameter $\lambda \geq 0$ beträgt

$$f(k) = P(X = k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \dots$$

Diese Verteilung bezeichnen wir kurz mit $\text{Poiss}(\lambda)$.

Eigenschaften

- Die Summe von n unabhängigen $\text{Bern}(p)$ -verteilten Zufallsvariablen ist bekanntlich $\text{Bin}(n, p)$ -verteilt. Ist die Erfolgswahrscheinlichkeit p klein (seltene Ereignisse) und n gross, so entspricht $\text{Bin}(n, p)$ ungefähr $\text{Poiss}(np)$. Diese sogenannten *Poissonapproximation* der Binomialverteilung wird insbesondere dann eingesetzt, wenn statt n und p lediglich deren Produkt $\lambda = np$ bekannt ist, also die mittlere Anzahl Ereignisse pro Zeiteinheit. Abbildung 3.21 zeigt die approximative Poissonverteilung der Binomialverteilung für grosse n und kleine p .

Aus dieser zentralen Eigenschaft folgen die nächsten vier Eigenschaften für $X \sim \text{Poiss}(\lambda)$.

- $E(X) = \lambda$. (Mittelwert, Rate)
- $\text{Var}(X) = \lambda$ bzw. $\text{Std}(X) = \sqrt{\lambda}$. (Varianz gleich Erwartungswert.)
- Die Wahrscheinlichkeitsfunktion f ist rechtsschief und hat einen Höcker bei etwa λ .
- Die Summe von unabhängigen poissonverteilten Zufallsvariablen ist wiederum poissonverteilt mit aufaddierter Rate.
- Die Wahrscheinlichkeit, dass X null ist, beträgt $f(0) = e^{-\lambda} \cdot \frac{\lambda^0}{0!} = e^{-\lambda}$.
- Die Verteilungsfunktion F entspricht definitionsgemäss einer Summe über f und ist entsprechend unhandlich zu verwenden. Quantile lassen sich allenfalls grafisch mithilfe von F finden. Als Faustregel entspricht der Median meist dem abgerundeten Wert von λ .

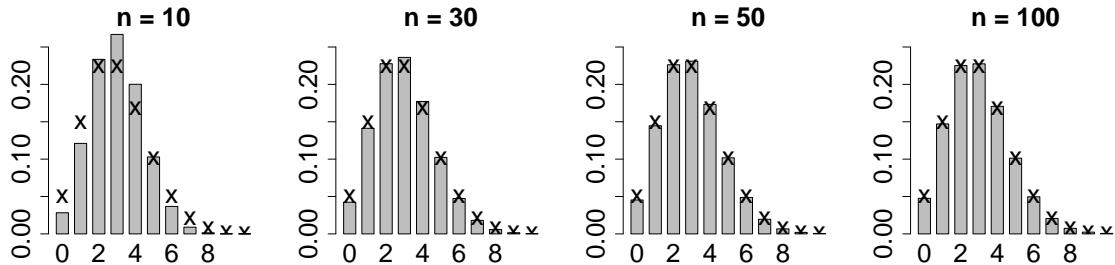


Abbildung 3.21: Für ein konstantes $\lambda = 3$ wird die Wahrscheinlichkeitsfunktion von $\text{Bin}(n, \lambda/n)$ (Balkendiagramm) mit jener von $\text{Poiss}(\lambda)$ (Kreuze) bei wachsendem n verglichen. Bereits für $n = 30$ ist die Übereinstimmung sehr gut.

Beispiel 3.62 (Flugzeugabstürze). Sei X die Anzahl Flugzeugabstürze pro Jahr. Man kann X als Zufallsvariable auffassen und annehmen, dass sie etwa poissonverteilt ist (grosse Anzahl unabhängiger Flüge mit je kleiner Absturzwahrscheinlichkeit). Den unbekannten Parameter λ kann man als mittlere Zahl von Abstürzen pro Jahr auffassen und aus Daten schätzen. Aus der Vergangenheit wisse man, dass pro Jahr im Schnitt $\lambda = 6$ Flugzeuge abstürzen.

Die Wahrscheinlichkeit, dass es nächstes Jahr keinen einzigen Absturz gibt, beträgt somit

$$P(X = 0) = e^{-6} \approx 0.00248.$$

Die Wahrscheinlichkeit, dass es im nächsten Jahr mindestens zwei Abstürze gibt, beträgt

$$P(X \geq 2) = 1 - (P(X = 0) + P(X = 1)) = 1 - e^{-6}(1 + \lambda) = 1 - 7 \cdot e^{-6} \approx 0.98.$$

Die Wahrscheinlichkeit von mindestens zehn Abstürzen ist entsprechend

$$P(X \geq 10) = 1 - P(X \leq 9) = 1 - e^{-6} (1 + 6 + \dots + 6^9 / 9!) \approx 0.084.^1$$

Die unter den Eigenschaften genannte Faustregel liefert den Median 6: In den 50% schlimmsten Jahren stürzen mindestens sechs² Flugzeuge ab. Abbildung 3.22 illustriert die Situation.

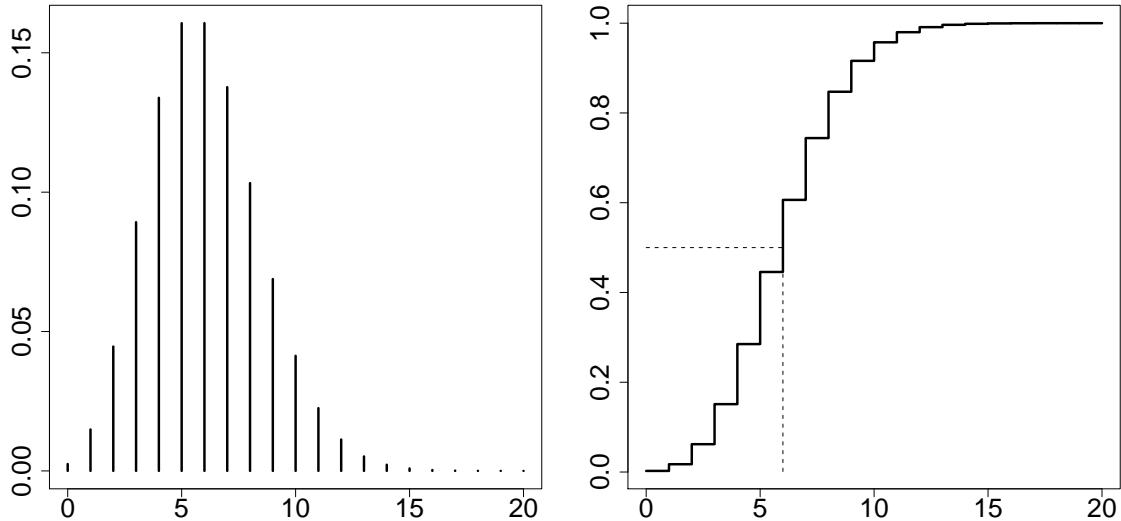


Abbildung 3.22: Wahrscheinlichkeits- und Verteilungsfunktion von $\text{Poiss}(6)$ aus Beispiel 3.62.

Statt die Anzahl Abstürze X pro Jahr könnten wir beispielsweise auch die Anzahl Abstürze Z pro Monat betrachten. Z ist dann entsprechend poissonverteilt mit Rate $6/12 = 0.5$. Die Wahrscheinlichkeit, dass der nächste Monat unfallfrei ist, beträgt $P(Z = 0) = e^{-0.5} \approx 0.6065$.

Für die Berechnungen mussten wir weder die Anzahl der Flüge pro Jahr oder Monat noch die Wahrscheinlichkeit eines Absturzes kennen. ▲

Beispiel 3.63 (Telefonauskunft). Sei X die Anzahl von Anfragen bei einer Telefonauskunftsstelle zwischen 8:00 und 8:15 Uhr am kommenden Freitag. Man kann diese Zahl X als Zufallsvariable betrachten und davon ausgehen, dass sie poissonverteilt ist mit unbekanntem Parameter λ (grosse Anzahl potenzieller Kunden, die selten und unabhängig voneinander anrufen).

Den unbekannten Parameter λ kann man als *mittlere Zahl* von Anfragen, die freitags zwischen 8:00 und 8:15 Uhr eingehen, auffassen und aus Daten schätzen. Angenommen, diese mittlere Anzahl von Anfragen ist gleich $\lambda = 5$. Dann kann man die Wahrscheinlichkeiten für beliebige Auslastungen am kommenden Freitag berechnen.

Beispielsweise ist

$$\begin{aligned} P(\text{keine Anfrage}) &= P(X = 0) = e^{-5} \approx 0.0067, \\ P(\text{genau eine Anfrage}) &= P(X = 1) = 5e^{-5} \approx 0.0337, \\ P(\text{mehr als 10 Anfragen}) &= P(X > 10) = 1 - P(X \leq 10) = 1 - e^{-5}(1 + 5 + \dots + 5^{10} / 10!) \approx 0.0137. \end{aligned}$$

▲

¹Mithilfe der Verteilungsfunktion `ppois(r, λ)` in R erhalten wir dasselbe Ergebnis.

²Dieser Median könnte explizit mit der R-Funktion `qpois(β, λ)` oder grafisch mit der Verteilungsfunktion F gefunden werden: Ab welchem k ist $F(k) = P(X \leq k)$ mindestens 0.5?

Beispiel 3.64 (Feuerwehreinsätze am Heiligen Abend). Sei X die Anzahl der Einsätze für eine bestimmte Feuerwehrstelle zwischen 18 Uhr am kommenden 24. Dezember und 6 Uhr am 25. Dezember. Angenommen, in den vergangenen zwanzig Jahren gab es in der Heiligen Nacht im Mittel 2.5 Einsätze. Nun gehen wir davon aus, dass die Zahl X eine poissonverteilte Zufallsgrösse ist mit Parameter $\lambda = 2.5$ (viele Bäume, die unabhängig voneinander und selten zu einem Feuerwehreinsatz führen).

Dies bedeutet z. B., dass

$$P(X = 0) = e^{-2.5} \approx 0.0821$$

oder

$$P(X > 3) = 1 - P(X \leq 3) = 1 - e^{-2.5}(1 + 2.5 + 2.5^2/2 + 2.5^3/6) \approx 0.2424.$$

▲

3.3 Normalverteilung und Verwandtes

3.3.1 Normalverteilung

Eine für die Statistik zentrale stetige Verteilung ist die *Normalverteilung (Gauss'sche Verteilung)*. Der Grund liegt darin, dass viele etwa symmetrisch verteilte numerische Variablen (z. B. Körpergewicht, Körpergrösse, Intelligenzquotient) sowie die meisten Schätzer (z. B. Mittelwert, Varianz, Quartile) ungefähr normalverteilt sind.

Eine Zufallsvariable X heisst normalverteilt mit Erwartungswert μ und Standardabweichung σ (Varianz σ^2), kurz $X \sim \mathcal{N}(\mu, \sigma^2)$, wenn ihre Dichtefunktion f eine *Gauss'sche Glockenkurve* ist:

$$f(x) := \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad x \in \mathbb{R}.$$

Die Dichtefunktion der *Standardnormalverteilung* ($\mu = 0, \sigma = 1$) wird häufig mit ϕ bezeichnet, deren Verteilungs- und Quantilfunktion mit Φ und Φ^{-1} . Abbildung 3.23 zeigt einige Normalverteilungen.

Eigenschaften

1. f ist unimodal und symmetrisch um μ . Der Parameter μ spezifiziert also die mittlere Lage.
2. σ bestimmt Höhe und Breite der Dichte. Beispielsweise hat f Wendepunkte bei $\mu \pm \sigma$: Innerhalb einer Standardabweichung σ um den Mittelwert μ ist die Kurve gegen unten geöffnet (konkav), außerhalb dieses Bereichs ist sie gegen oben geöffnet (konvex).
3. f ist überall positiv.
4. Sei X eine Zufallsvariable mit Erwartungswert μ und Varianz σ^2 . Dank den Regeln für Erwartungswerte und Varianzen wissen wir, dass die linear abgebildete Zufallsvariable $Y := a + bX$ (a und b beliebig) den Erwartungswert $E(Y) = a + b\mu$ und die Varianz $\text{Var}(Y) = b^2\sigma^2$ aufweist. Ist X normalverteilt, gilt zusätzlich, dass Y ebenfalls normalverteilt ist.
5. Seien X und Y unabhängige Zufallsvariablen. Aus den Regeln zu Erwartungswerten und Varianzen folgt, dass die neue Zufallsvariable $T := X + Y$ den Erwartungswert $E(T) = E(X) + E(Y)$ und die Varianz $\text{Var}(T) = \text{Var}(X) + \text{Var}(Y)$ hat. Sind X und Y normalverteilt, so gilt zudem, dass deren Summe T ebenfalls normalverteilt ist.
6. Dank Symmetrie ist $\Phi(0) = 1/2$, $\Phi(-r) = 1 - \Phi(r)$ und $\Phi^{-1}(\beta) = -\Phi^{-1}(1 - \beta)$, siehe Abbildung 3.24.

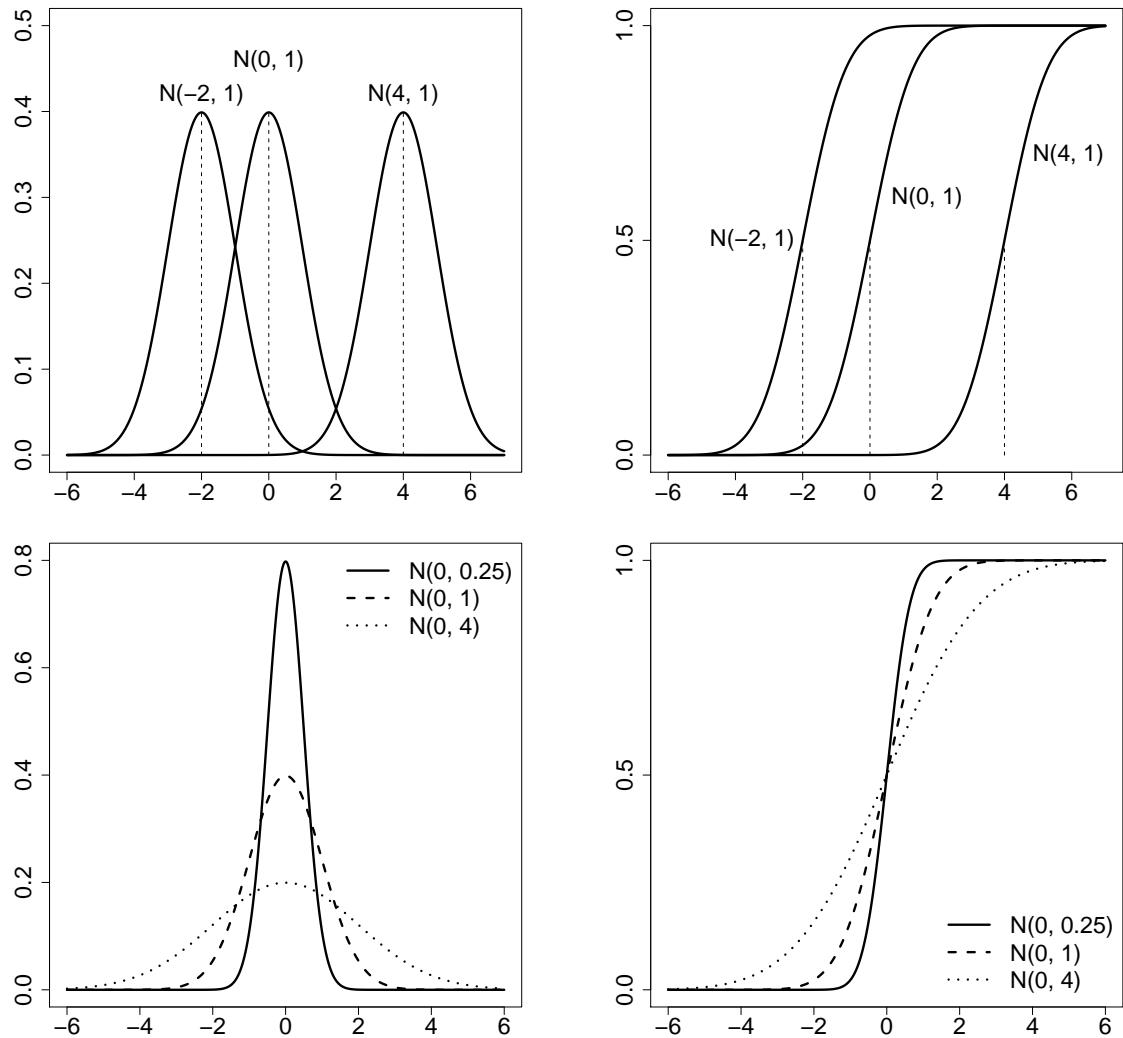


Abbildung 3.23: Die linken Bilder zeigen die Dichtefunktionen, die rechten die entsprechenden Verteilungsfunktionen verschiedener Normalverteilungen. Während in den oberen Bildern die Rolle von μ illustriert wird, zeigen die unteren Bilder die Rolle von σ .

Verteilungs- und Quantilfunktion lassen sich nur mit Integralen darstellen. Entsprechend ist man bei den meisten Berechnungen auf Tabellen oder Software angewiesen.

Beispiel 3.65 (Körpergrösse). In einer bestimmten Population sei die Körpergrösse der Männer näherungsweise normalverteilt mit Mittelwert $\mu = 180\text{cm}$ und Standardabweichung $\sigma = 7\text{cm}$. Wenn beispielsweise ein Eisenbahnunternehmen neue Schlafwaggons in Auftrag gibt, ist vielleicht folgende Frage interessant: Wie gross ist der relative Anteil von Personen mit Körpergrösse 190cm oder mehr? Mit der Körpergrösse X einer zufällig herausgegriffenen Person und der Verteilungsfunktion¹ F von $\mathcal{N}(180, 7^2)$ ist dieser Anteil

$$P(X \geq 190\text{cm}) = 1 - P(X \leq 190\text{cm}) = 1 - F(190) \approx 0.0766.$$

Wenn das Eisenbahnunternehmen sicherstellen will, dass die geplanten Betten für höchstens 3% aller Männer zu kurz sind, bestimmt es das 97%-Quantil $F^{-1}(0.97) \approx 193.17\text{ cm}$. ▲

Beispiel 3.66 (IQ-Test). Intelligenztests werden so konzipiert, dass der Intelligenzquotient (IQ) in der Gesamtbevölkerung normalverteilt ist mit Erwartungswert $\mu = 100$ und Standardabweichung $\sigma = 15$. Be-

¹Mit den R-Funktionen `pnorm` (Verteilungsfunktion) und `qnorm` (Quantilfunktion).

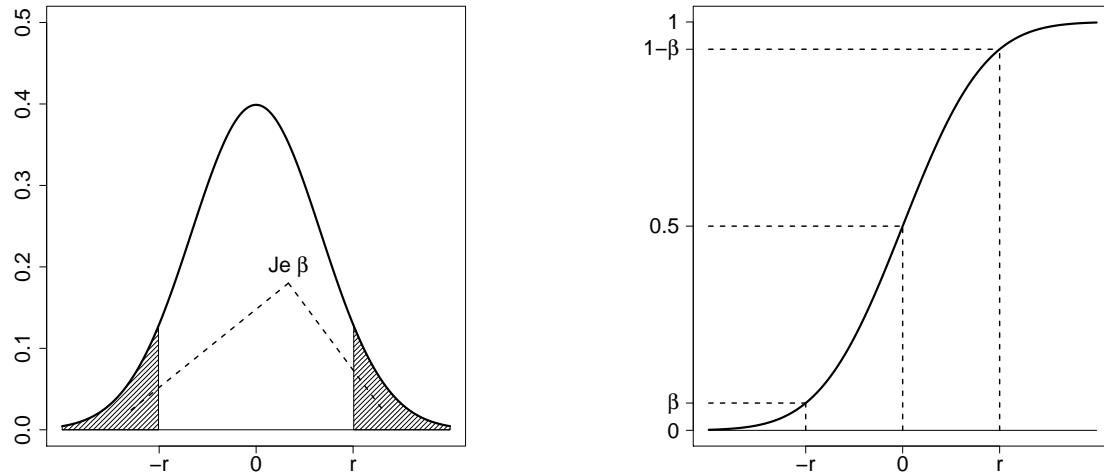


Abbildung 3.24: Illustration der Symmetrieeigenschaften der Standardnormalverteilung (links anhand der Dichte ϕ , rechts anhand der Verteilungsfunktion Φ).

zeichnen wir mit X den IQ einer zufällig herausgegriffenen Person und mit F die Verteilungsfunktion von $\mathcal{N}(100, 15^2)$, dann ist (mithilfe der Software)

- der relative Anteil von Personen mit einem IQ über 130

$$P(X > 130) = 1 - F(130) \approx 2.3\%,$$

- der relative Anteil von Personen mit einem IQ zwischen 85 und 115

$$P(85 \leq X \leq 115) = F(115) - F(85) \approx 68\%,$$

- die Schranke, welche die 20% der Bevölkerung mit höherem IQ von den 80% mit tieferem IQ trennt,
 $F^{-1}(0.80) \approx 113.$

▲

Wir betrachten nun einige wichtigen Anwendungen der Eigenschaften 4, 5 und 6.

Anwendung: Standardisierung In Beispiel 3.39 haben wir gesehen, dass die standardisierte Version

$$\frac{X - E(X)}{\text{Std}(X)}$$

einer Zufallsvariable X den Erwartungswert 0 und die Varianz 1 aufweist. Ist X normalverteilt, so folgt aus Eigenschaft 4 zusätzlich, dass die standardisierte Version von X standardnormalverteilt ist.

Anwendung: Berechnungen von Hand und Interpretation der (Stichproben-)Standardabweichung

Sei k irgendeine positive Zahl und $X \sim \mathcal{N}(\mu, \sigma^2)$. Dank obiger Anwendung und Eigenschaft 6 folgt nun, dass

$$P(\mu - k\sigma \leq X \leq \mu + k\sigma) = P\left(-k \leq \underbrace{\frac{X - \mu}{\sigma}}_{\sim \mathcal{N}(0,1)} \leq k\right) = \Phi(k) - \Phi(-k) = \Phi(k) - (1 - \Phi(k)) = 2\Phi(k) - 1.$$

Setzen wir $k = 1$ in diese Formel ein, so finden wir mit der Software die Wahrscheinlichkeit, dass X einen Wert innerhalb einer Standardabweichung um den Erwartungswert annimmt, nämlich

$$P(\mu - \sigma \leq X \leq \mu + \sigma) = 2\Phi(1) - 1 \approx 0.68.$$

Für $k = 2$ und $k = 3$ folgt entsprechend, dass

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 2\Phi(2) - 1 \approx 0.95$$

und

$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 2\Phi(3) - 1 \approx 0.997.$$

Mit diesen drei Faustregeln können gewisse Wahrscheinlichkeiten von Hand berechnet werden. Aus Gründen der Symmetrie liegt X beispielsweise mit Wahrscheinlichkeit $0.68/2 = 0.34 = 34\%$ zwischen μ und $\mu + \sigma$ und mit Wahrscheinlichkeit $0.34 + 0.95/2 = 0.815 = 81.5\%$ zwischen $\mu - \sigma$ und $\mu + 2\sigma$. Nur ein verschwindend kleiner Anteil nimmt Werte an, die mehr als drei Standardabweichungen vom Mittelwert entfernt sind. Abbildung 3.25 illustriert die Faustregeln grafisch.

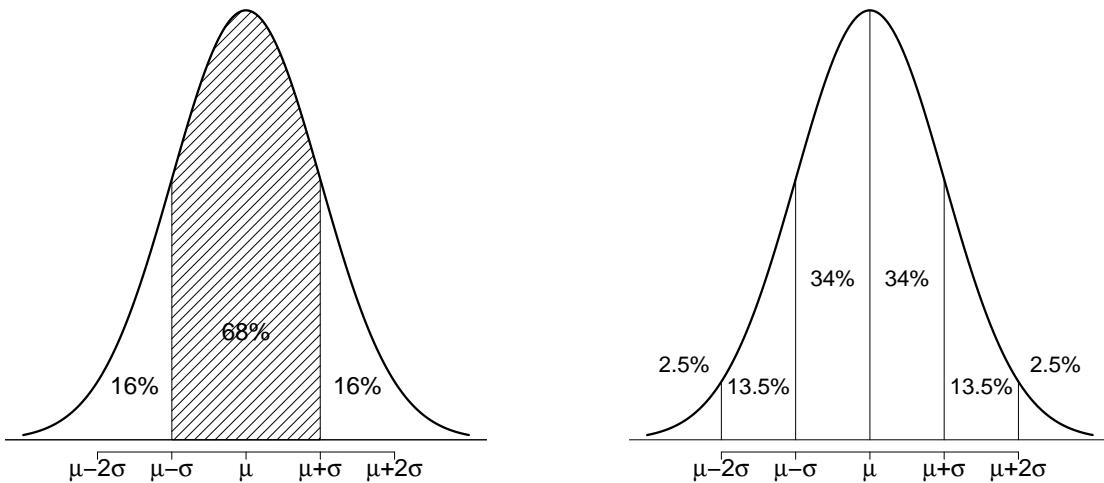


Abbildung 3.25: Illustration der Faustregeln zur Normalverteilung.

Beispiel 3.67 (IQ-Test, Fortsetzung). Der Bereich $[85, 115]$ entspricht dem Intervall $[\mu \pm \sigma]$, also folgt mit den Faustregeln, dass $P(85 \leq X \leq 115) = 0.68$. Die Faustregeln besagen auch, dass X mit Wahrscheinlichkeit 0.05 ausserhalb $[70, 130]$ liegt bzw. dank Symmetrie mit Wahrscheinlichkeit 0.025 über 130. ▲

In der deskriptiven Statistik werden diese Faustregeln oft eingesetzt, um relative Anteile eines ungefähr normalverteilten Merkmals abzuschätzen oder um die Stichprobenstandardabweichung zu interpretieren. Diese entspricht dann der halben Länge des Bereichs, der etwa zwei Drittel der Werte abdeckt.

Beispiel 3.68 (Körpergrösse, Fortsetzung). In den Beispielen 2.11 und 2.13 haben wir u. a. gesehen, dass die Variable ‘Körpergrösse’ den Mittelwert 174.19 cm und die Standardabweichung 8.15 cm hat und etwa symmetrisch verteilte Stichprobenwerte aufweist, die etwa aus einer Normalverteilung stammen könnten.

Anhand obiger Faustregeln können wir z. B. folgende Abschätzungen machen:

- Rund zwei Drittel der Befragten sind zwischen $174.19 - 8.15 = 166.04$ cm und $174.19 + 8.15 = 182.34$ cm gross. (Tatsächlich sind es 64.3%).
- Rund 2.5% sind grösser als $174.19 + 2 \cdot 8.15 = 190.49$ cm. (Tatsächlich sind es 1.9%).
- Ein Bereich der Länge $2 \cdot 8.15$ cm (zwei Standardabweichungen) enthält ca. zwei Drittel der Werte. ▲

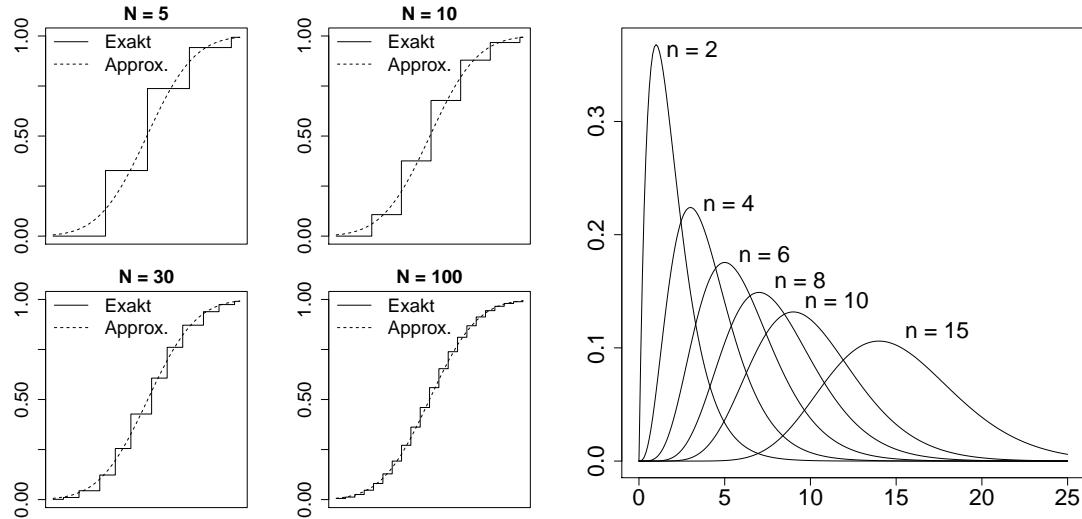


Abbildung 3.26: Illustration des Zentralen Grenzwertsatzes. Kleine Bilder: Die Verteilungsfunktion des standardisierten Mittelwerts von n unabhängigen $\text{Bern}(0.2)$ -verteilten Zufallsvariablen (via Binomialverteilung) wird für verschiedene n mit Φ verglichen. Je grösser n , desto besser die Übereinstimmung. Das grosse Bild zeigt die Dichtefunktionen von Summen von unabhängigen $\text{Exp}(1)$ -verteilten Zufallsvariablen. Je mehr Summanden, desto eher sieht die Verteilung wie eine Normalverteilung aus. In beiden Situationen sind die ursprünglichen Zufallsvariablen überhaupt nicht normalverteilt.

Anwendung: Stichprobenmittelwert Von den Beispielen 3.30 und 3.38 wissen wir, dass der Mittelwert \bar{X} von unabhängigen Zufallsvariablen X_1, \dots, X_n mit je Erwartungswert μ und Varianz σ^2 den Erwartungswert $E(\bar{X}) = \mu$ und die Standardabweichung $\text{Std}(\bar{X}) = \sigma/\sqrt{n}$ aufweist. Für *normalverteilte* Zufallsvariablen ist \bar{X} dank den Regeln 4 und 5 zudem *normalverteilt*. Insbesondere ist der *standardisierte* Mittelwert

$$\frac{\bar{X} - E(\bar{X})}{\text{Std}(\bar{X})} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

von *normalverteilten* Zufallsvariablen *standardnormalverteilt*.

Zentraler Grenzwertsatz Die Summe von nicht allzu wenigen i. d. Zufallsvariablen ist näherungsweise *normalverteilt*. Dieses wichtige Resultat wird in Abbildung 3.26 auf zwei Arten illustriert.

Damit ist beispielsweise der standardisierte Mittelwert bei nicht allzu kleinen Stichprobenumfängen (z. B. 30) auch für *nicht normalverteilte* Zufallsvariablen etwa *standardnormalverteilt*.

Hinweis (Approximative Normalität). Viele Schätzer sind bei nicht allzu kleinen Stichprobenumfängen approximativ *normalverteilt*, beispielsweise der Stichprobenmedian oder die Stichprobenvarianz.

3.3.2 Konfidenzintervalle und Tests für einen Mittelwert

Auf obigen Erkenntnissen zum Mittelwert basiert die Konstruktion von Konfidenzintervallen und Tests für einen unbekannten Mittelwert.

Z-Konfidenzintervalle

Seien X_1, X_2, \dots, X_n zunächst unabhängige, *normalverteilte* Beobachtungen mit unbekanntem Erwartungswert μ und *bekannter* Standardabweichung σ .

Wie könnte man hier z. B. eine untere $(1 - \alpha)$ -Konfidenzschranke für μ bestimmen? Dazu versuchen wir, die Gleichung $P(\bar{X} - c \leq \mu) = 1 - \alpha$ nach c aufzulösen:

$$\begin{aligned} & P(\bar{X} - c \leq \mu) = 1 - \alpha \\ \Leftrightarrow & P(\bar{X} \leq \mu + c) = 1 - \alpha \\ \Leftrightarrow & P\left(\underbrace{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}_{\sim \mathcal{N}(0,1)} \leq \frac{c}{\sigma/\sqrt{n}}\right) = 1 - \alpha \\ \Leftrightarrow & \Phi\left(\frac{c}{\sigma/\sqrt{n}}\right) = 1 - \alpha \\ \Leftrightarrow & \frac{c}{\sigma/\sqrt{n}} = \Phi^{-1}(1 - \alpha) \\ \Leftrightarrow & c = \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1 - \alpha). \end{aligned}$$

Für konkrete Daten liegt dann μ mit einer Sicherheit von $(1 - \alpha) \cdot 100\%$ über der unteren Konfidenzschranke

$$\bar{X} - c = \bar{X} - \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1 - \alpha).$$

Die entsprechende obere Konfidenzschranke sowie das zweiseitige Konfidenzintervall folgen analog.

Bei der Herleitung haben wir die Tatsache verwendet, dass der standardisierte Mittelwert

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

von normalverteilten Beobachtungen standardnormalverteilt ist. Wir haben gesehen, dass dies dank des Zentralen Grenzwertsatzes für grössere n zumindest approximativ auch bei *nicht normalverteilten* Beobachtungen gilt. Da bei grösserem n ausserdem $\sigma \approx S$ ist, ist dann auch der *studentisierte* Mittelwert

$$Z := \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

ungefähr standardnormalverteilt. Sofern also n nicht allzu klein ist, haben wir folgende *Z-Konfidenzintervalle* für den Erwartungswert μ von i. i. d. Beobachtungen X_1, \dots, X_n mit beliebiger Verteilung zur Auswahl, wobei das Konfidenzniveau approximativ gleich $1 - \alpha$ ist:

- Das zweiseitige Konfidenzintervall $\left[\bar{X} \pm \frac{S}{\sqrt{n}}\Phi^{-1}(1 - \alpha/2)\right]$
- Die untere Konfidenzschanke $\bar{X} - \frac{S}{\sqrt{n}}\Phi^{-1}(1 - \alpha)$
- Die obere Konfidenzschanke $\bar{X} + \frac{S}{\sqrt{n}}\Phi^{-1}(1 - \alpha)$

Für $\alpha = 0.05$ lauten die Quantile $\Phi^{-1}(0.975) \approx 1.96$ und $\Phi^{-1}(0.95) \approx 1.645$. Das zweiseitige 95%-Konfidenzintervall beträgt somit $[\bar{X} \pm 1.96 \cdot S/\sqrt{n}]$. Anhand seiner Breite $3.92 \cdot S/\sqrt{n}$ sieht man gut, wie die Präzision von \bar{X} vom Stichprobenumfang und der Streuung der Werte abhängt.

Hinweis (Standardfehler). Der Ausdruck S/\sqrt{n} wird *Standardfehler* des Mittelwerts¹ genannt und darf nicht mit der Standardabweichung S einer Beobachtung verwechselt werden.

Beispiel 3.69 (Mieterverband, Fortsetzung). In Beispiel 2.16 haben wir einen Datensatz mit $n = 76$ Wohnungen betrachtet. Wir verwenden den mittleren Mietpreis 1239.1 CHF in der Stichprobe als Schätzwert für den Mittelwert μ aller Schweizer Wohnungsmieten. Zusammen mit der Standardabweichung

¹ Als Standardfehler bezeichnet man ganz generell einen Schätzer für die Standardabweichung eines Schätzers.

413.9 CHF können wir einen möglichen Bereich für μ angeben: Das approximative, zweiseitige 95%-Konfidenzintervall für μ beträgt $[1239.1 \pm 1.96 \cdot 413.9 / \sqrt{76}] = [1146.0, 1332.2]$. Mit einer Sicherheit von rund 95% liegt μ also zwischen 1146 und 1332 CHF.

Ein Mieterverband möchte nun mithilfe der gleichen Daten nachweisen, dass der mittlere Mietpreis in der Schweiz hoch ist. Zu diesem Zweck betrachtet er eine untere 95%-Konfidenzschranke für μ . Diese beträgt $1239.1 - 1.64 \cdot 413.9 / \sqrt{76} = 1161.2$ CHF. Der Verband kann nun also mit einer Sicherheit von rund 95% behaupten, dass μ mindestens 1161 CHF beträgt. ▲

Beispiel 3.70 (Immobilien, Fortsetzung). Eine deutsche Immobilienfirma möchte ins lukrative Schweizer Geschäft einsteigen. Um einen Eindruck über die typische Ausstattung einer Schweizer Wohnung zu gewinnen, schätzt sie mit den Daten in Beispiel 2.16 u. a. den tatsächlichen relativen Anteil p der Wohnungen mit Balkon und stattet den Schätzwert mit einem 95%-Binomialkonfidenzintervall aus:

R Code

```
# Eingabe
binom.test(sum(wohnungen$Balkon), 76)

# Ausgabe
95 percent confidence interval: 0.2965 0.5267
sample estimates: probability of success 0.40789
```

Mit dem Mittelwert (0.40789) und der Stichprobenstandardabweichung (0.49471) des (0-1)-Merkmals ‘Balkon’ könnten wir im Prinzip stattdessen auch das entsprechende 95%-Z-Konfidenzintervall

$$[0.40789 \pm 1.96 \cdot 0.49471 / \sqrt{76}] = [0.2967, 0.5191]$$

für p angeben, da relative Anteile als Mittelwerte aufgefasst werden können. Das ungenauere Z-Intervall unterscheidet sich dank der nicht allzu kleinen Stichprobe kaum vom exakten Intervall.

Um zu illustrieren, dass Z-Konfidenzintervalle selbst bei kleinen Stichprobenumfängen erstaunlich präzise sind, betrachten wir nun noch die Situation, in der 8 von $n = 20$ Wohnungen einen Balkon hätten: Hier liefert das exakte Binomialverfahren das Intervall [0.191, 0.639], während sich mit dem ungenauerem Z-Verfahren das ähnliche Intervall [0.180, 0.620] ergibt. ▲

Z-Tests

Mit dem Konfidenzintervallansatz lassen sich Arbeitshypothesen der Art $\mu > \mu_o$, $\mu < \mu_o$ oder $\mu \neq \mu_o$ über den wahren Mittelwert μ testen. Die entsprechenden Testentscheide können auch mit der aus dem Schätzer \bar{X} berechneten *Z-Teststatistik*

$$Z_o := \frac{\bar{X} - \mu_o}{S / \sqrt{n}}$$

oder dem darauf beruhenden p -Wert getroffen werden:

- Je grösser Z_o , je mehr spricht gegen $H_o: \mu \leq \mu_o$.
- Je kleiner Z_o , desto mehr spricht gegen $H_o: \mu \geq \mu_o$.
- Je grösster der Absolutbetrag $|Z_o|$, je mehr spricht gegen $H_o: \mu = \mu_o$.

Unter der Nullhypothese ($\mu = \mu_o$) entspricht Z_o dem studentisierten Mittelwert, von dem wir wissen, dass er bei grösserem n approximativ standardnormalverteilt ist. Damit gilt insbesondere

$$P(Z_o > \Phi^{-1}(1 - \alpha)) \approx P(Z_o < \Phi^{-1}(\alpha)) \approx P(|Z_o| > \Phi^{-1}(1 - \alpha/2)) \approx \alpha.$$

Indem man den konkreten Wert z_o der Teststatistik mit der passenden kritischen Schranke ($\Phi^{-1}(1 - \alpha)$ etc.) vergleicht, erhält man deshalb einen Test auf approximativem Niveau α . Zusammengefasst:

- Verwerfe $H_o: \mu \leq \mu_o$ zugunsten $H_1: \mu > \mu_o$, wenn $z_o > \Phi^{-1}(1 - \alpha)$.
- Verwerfe $H_o: \mu \geq \mu_o$ zugunsten $H_1: \mu < \mu_o$, wenn $z_o < \Phi^{-1}(\alpha)$.
- Verwerfe $H_o: \mu = \mu_o$ zugunsten $H_1: \mu \neq \mu_o$, wenn $|z_o| > \Phi^{-1}(1 - \alpha/2)$.

Da ein p -Wert die Wahrscheinlichkeit angibt, dass unter der Nullhypothese mindestens so viel Evidenz gegen H_o wie in der konkret verfügbaren Stichprobe vorliegt, lassen sich die entsprechenden p -Werte folgendermassen bestimmen: (X bezeichnet hier eine standardnormalverteilte Zufallsvariable.)

- $H_o: \mu \leq \mu_o$ vs. $H_1: \mu > \mu_o$: $p\text{-Wert} = P(X \geq z_o) \approx 1 - \Phi(z_o)$
- $H_o: \mu \geq \mu_o$ vs. $H_1: \mu < \mu_o$: $p\text{-Wert} = P(X \leq z_o) \approx \Phi(z_o)$
- $H_o: \mu = \mu_o$ vs. $H_1: \mu \neq \mu_o$:

$$\begin{aligned} p\text{-Wert} &= P(|X| \geq |z_o|) = 1 - P(-|z_o| \leq X \leq |z_o|) \\ &\approx 1 - (\Phi(|z_o|) - (1 - \Phi(|z_o|))) \\ &= 2 - 2\Phi(|z_o|). \end{aligned}$$

Bei approximativen Test wird das Signifikanzniveau nur ungefähr eingehalten. Das bringen wir dadurch zum Ausdruck, dass wir beispielsweise „mit einer Sicherheit von *etwa* 95%“ schreiben.

Beispiel 3.71 (Mieterverband, Fortsetzung). Der Mieterverband von Beispiel 3.69 möchte auf dem 5%-Niveau zeigen, dass der mittlere Mietpreis μ in der Schweiz seit 2009 zugenommen hat. Damals habe der wahre Wert $\mu_o = 1190$ CHF betragen. Die Arbeitshypothese lautet $\mu > 1190$, die Nullhypothese $\mu \leq 1190$.

Mit $\bar{X} = 1239.1$, $S = 413.94$ und $n = 76$ beträgt der konkrete Wert der Z-Teststatistik

$$z_o = \frac{1239.1 - 1190}{413.94/\sqrt{76}} = 1.0341.$$

Dieser Wert ist nicht grösser als die kritische Schranke $\Phi^{-1}(0.95) \approx 1.64$, somit kann der Mieterverband seine Behauptung nicht belegen. Zum gleichen Testentscheid führt auch der p -Wert¹ $1 - \Phi(1.0341) = 0.151 \geq 0.05$ und die untere Konfidenzschranke $1161.2 \leq 1190$ von Beispiel 3.69. Die Testsituation ist im linken Bild von Abbildung 3.27 dargestellt. ▲

Beispiel 3.72 (Wahlprognosen, Fortsetzung). In Beispiel 3.51 haben wir mit einem Binomialtest die Nullhypothese, dass der Wähleranteil p von Partei ABC kleiner oder gleich 20% ist, auf einem Niveau von 1% zugunsten der Arbeitshypothese ($p > 20\%$) verworfen (exakter p -Wert 0.0037). Zum gleichen Testentscheid wären wir mithilfe der exakten unteren 99%-Binomialkonfidenzschanke 0.206 gekommen.

Zu welchen Ergebnissen würde das ungenauere Z-Verfahren führen? Mit den Angaben $n = 500$, $\bar{X} = 125/n = 0.25$ und² $S = \sqrt{\bar{X}(1 - \bar{X})} = 0.433$ finden wir den konkreten Wert der Z-Teststatistik

$$z_o = \frac{0.25 - 0.2}{0.433/\sqrt{500}} = 2.582.$$

¹Mit dem R-Aufruf `pnorm(1.0341)`.

²Via Eigenschaft der Bernoulliverteilung können wir die wahre Standardabweichung auf diese Weise schätzen.

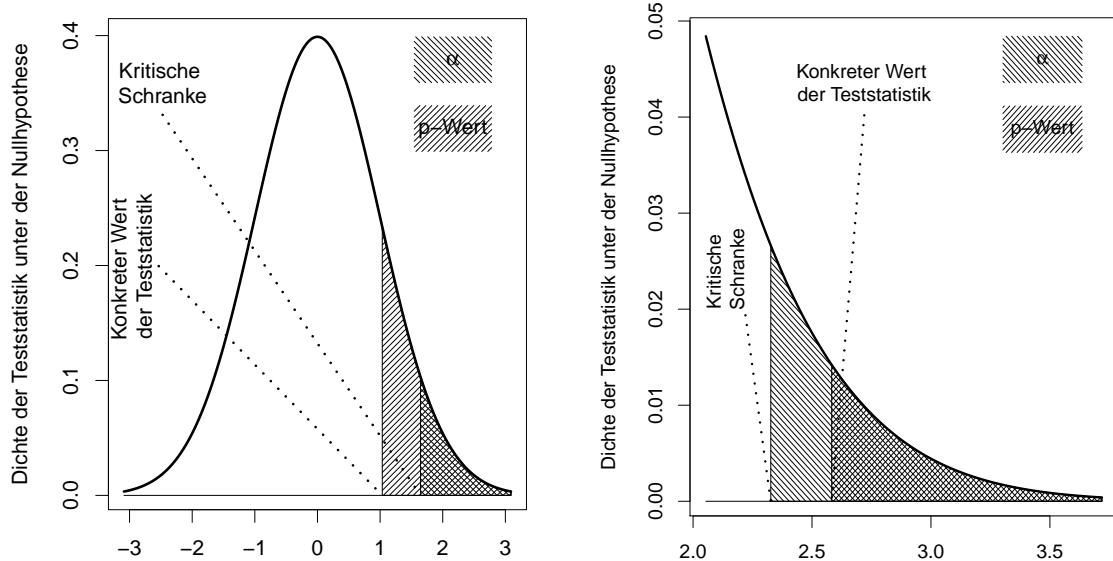


Abbildung 3.27: Illustration der Testsituationen aus den Beispielen 3.71 (links) und 3.72 (rechts).

Da er grösser als die kritische Schranke¹ $\Phi^{-1}(0.99) \approx 2.3263$ ist, verwerfen wir auch hier die Nullhypothese auf dem 1%-Niveau und behaupten mit einer Sicherheit von ungefähr 99%, dass der Wähleranteil grösser als 20% ist. Der approximative p -Wert ist mit $1 - \Phi(2.582) = 0.0049$ zwar anders als der exakte, führt jedoch ebenfalls zum gleichen Testentscheid. Ähnliches gilt bezüglich der approximativen unteren 99%-Z-Konfidenzschanke $\bar{X} - \Phi^{-1}(0.99) \cdot S/\sqrt{500} = 0.25 - 2.3263 \cdot 0.433/22.361 = 0.205$.

Die Situation ist im rechten Bild von Abbildung 3.27 dargestellt. ▲

3.3.3 Verfeinerung nach Students Methode

Die Z-Verfahren basieren auf der approximativen Standardnormalverteilung des *studentisierten* Mittelwerts Z . Im Gegensatz zum *standardisierten* Mittelwert ist Z selbst bei normalverteilten Beobachtungen nicht exakt standardnormalverteilt, sondern hat die sogenannte *Students² t-Verteilung* (*Student-Verteilung*, *t-Verteilung*) mit $n - 1$ *Freiheitsgraden*³, kurz t_{n-1} . Damit lassen sich die Z-Verfahren verfeinern.

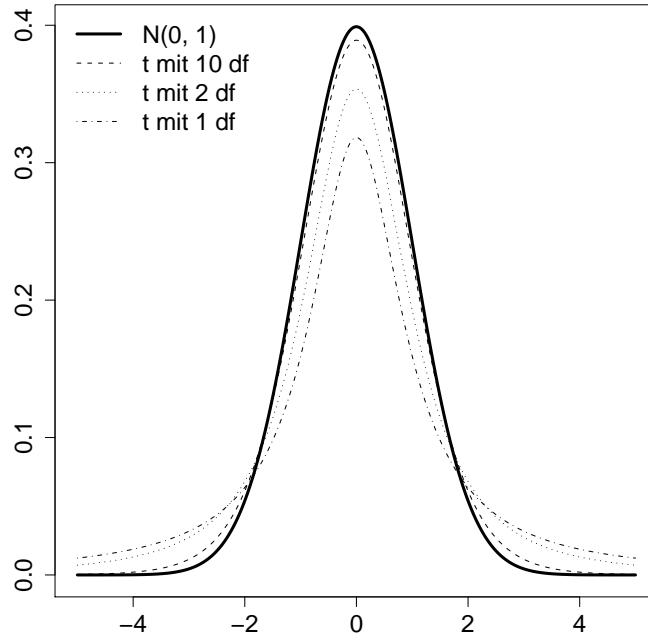
Eigenschaften der Student-Verteilung

- Die Dichtefunktion von t_k ist eine um 0 symmetrische Glockenkurve, die sich mit wachsender Anzahl Freiheitsgraden k immer mehr der Dichte ϕ der Standardnormalverteilung annähert, siehe Abbildung 3.28. Sie hat die Form $c(1+x^2/k)^{-\frac{k+1}{2}}$. (Die Konstante c normiert die Dichte auf Fläche 1.)
- Median und Erwartungswert betragen 0. (t_1 besitzt keinen Erwartungswert.)
- Die Varianz beträgt $k/(k - 2)$. Mit wachsendem k nähert sie sich von oben der Zahl 1 (Varianz der Standardnormalverteilung). (t_1 und t_2 besitzen keine Varianz.)
- Die Student-Verteilung mit einem Freiheitsgrad, t_1 , wird *Cauchyverteilung* genannt, das Paradebeispiel einer Verteilung ohne Erwartungswert und Varianz.

¹Mit dem R-Aufruf `qnorm(0.99)`.

²Dieses wichtige Resultat wurde 1908 von William Gosset publiziert, einem britischen Statistiker. Auf Wunsch seines Arbeitgebers, der Guinness-Brauerei, musste er sein Ergebnis unter dem Pseudonym "Student" publizieren.

³Obwohl der Begriff "Freiheitsgrad" in der Statistik häufig auftaucht, gibt es keine gute Definition dafür.

Abbildung 3.28: Dichtefunktionen von $\mathcal{N}(0, 1)$, t_{10} , t_2 , t_1 .

Verteilungsfunktion und Quantilfunktion sind nur via Tabellen bzw. Software verfügbar. Mit $t_{k;\beta}$ bezeichnen wir hier das β -Quantil, mit F_k die Verteilungsfunktion von t_k .

Student-Konfidenzintervalle

Für unabhängige $\mathcal{N}(\mu, \sigma^2)$ -verteilte Beobachtungen gilt dank der Verfeinerung durch Student exakt, dass

$$\left. \begin{aligned} & P\left(\mu \in \left[\bar{X} \pm \frac{S}{\sqrt{n}} t_{n-1;1-\alpha/2}\right]\right) \\ & P\left(\mu \geq \bar{X} - \frac{S}{\sqrt{n}} t_{n-1;1-\alpha}\right) \\ & P\left(\mu \leq \bar{X} + \frac{S}{\sqrt{n}} t_{n-1;1-\alpha}\right) \end{aligned} \right\} = 1 - \alpha.$$

Daraus ergeben sich exakte $(1 - \alpha)$ -Konfidenzintervalle für μ :

- Das zweiseitige Konfidenzintervall

$$\left[\bar{X} \pm \frac{S}{\sqrt{n}} t_{n-1;1-\alpha/2} \right]$$

- Die untere Konfidenzschanke

$$\bar{X} - \frac{S}{\sqrt{n}} t_{n-1;1-\alpha}$$

- Die obere Konfidenzschanke

$$\bar{X} + \frac{S}{\sqrt{n}} t_{n-1;1-\alpha}$$

Der einzige rechnerische Unterschied zu den Z-Konfidenzintervallen besteht darin, dass hier mit Student-quantilen statt mit Normalquantilen gearbeitet wird.

Student-Konfidenzintervalle gelten dank des Zentralen Grenzwertsatzes bei nicht normalverteilten Beobachtungen immerhin approximativ, sofern die Stichprobe nicht allzu klein ist. Da sie auch dann präziser als die einfacheren Z-Konfidenzintervalle sind, werden sie letzteren stets vorgezogen. Gleiches gilt für die *t*-Tests, auf die wir nachher eingehen.

Weshalb haben wir die Z-Verfahren überhaupt betrachtet? Erstens illustrieren sie den Zentralen Grenzwertsatz, zweitens sind sie die Ausgangslage für die Verfeinerung mit der Student-Verteilung und drittens unterscheiden sich die Ergebnisse dieser beiden Verfahren bei grösserem n kaum.

Beispiel 3.73 (Mieterverband, Fortsetzung). In Beispiel 3.69 haben wir das zweiseitige 95%-Z-Konfidenzintervall $[1146.0, 1332.2]$ für die mittlere Wohnungsmiete μ in der Schweiz bestimmt. Anstelle des 97.5%-Quantils $\Phi^{-1}(0.975) \approx 1.96$ verwenden wir nun das entsprechende Studentquantil¹ $t_{75;0.975} \approx 1.9921$ und erhalten die leicht vorsichtigeren Schranken

$$[1239.1 \pm 1.9921 \cdot 413.9 / \sqrt{76}] \approx [1144.5, 1333.7] \text{ CHF.}$$

Die untere 95%-Student-Konfidenzschranke für μ beträgt mit $t_{75;0.95} = 1.6654$ entsprechend

$$1239.1 - 1.6654 \cdot 413.9 / \sqrt{76} = 1160.0 \text{ CHF}$$

und ist damit fast gleich gross wie die Z-Schranke 1161.2. Verifizieren² wir die Berechnungen nun mit R:

R Code		
# Zweiseitiges t-Konfidenzintervall t.test(wohnungen\$Preis)		# Ergibt 1144.5 1333.6
# Untere Schranke t.test(wohnungen\$Preis, alternative = 'greater')		# Ergibt 1160.0 Inf



Beispiel 3.74 (Immobilien, Fortsetzung). In Beispiel 3.70 haben wir das exakte 95%-Binomialkonfidenzintervall $[0.2965, 0.5267]$ für den tatsächlichen relativen Anteil der Wohnungen mit Balkon mit dem unge naueren Z-Konfidenzintervall $[0.2967, 0.5191]$ verglichen.

Ersetzen wir in letzterem das Normalquantil durch das entsprechende Studentquantil, so erhalten wir das Student-Konfidenzintervall $[0.40789 \pm 1.9921 \cdot 0.49471 / \sqrt{76}] \approx [0.2948, 0.5209]$.

Zum gleichen Ergebnis gelangen wir mit der Software.

R Code		
t.test(wohnungen\$Balkon)		# Ergibt 0.29485 0.52094



(Einstichproben-)t-Tests

Es bietet sich an, Students Verfeinerung auch auf die Z-Tests anzuwenden. Dazu werden in den dortigen Formeln Φ und Φ^{-1} durch die Verteilungs- und Quantilfunktionen der Studentverteilung mit entsprechender Anzahl Freiheitsgraden ersetzt.

Beispiel 3.75 (Mieterverband, Fortsetzung). In Beispiel 3.71 hat ein Mieterverband die Arbeitshypothese “Mittlere Mieten sind angestiegen (bzw. grösser als 1190)” auf dem 5%-Niveau nicht bestätigen können. Die Teststatistik betrug $z_o = 1.0341$, die kritische Schranke $\Phi^{-1}(0.95) \approx 1.645$ und der *p*-Wert $1 - \Phi(1.0341) = 0.151$.

¹In R mit `qt(0.975, 75)`.

²Mit der R-Funktion `t.test`.

Die kritische Schranke¹ des entsprechenden t -Tests beträgt $t_{75;0.95} \approx 1.6654$ und der p -Wert² lautet $1 - F_{75}(1.0341) = 0.152$. Somit folgt der gleiche Testentscheid, zu dem wir auch mit der unteren Konfidenzschranke 1160.0 CHF von Beispiel 3.73 gekommen wären. Das gleiche erhalten wir mit der Software:

	R Code	
	<code>t.test(wohnungen\$Preis, mu = 1190, alternative = 'greater')</code>	# Ergibt p-value = 0.1524



3.4 Gammaverteilung und Verwandtes

3.4.1 Gammaverteilung

Eine flexible Verallgemeinerung der Exponentialverteilung, ebenfalls eine stetige, rechtsschiefe Verteilung, ist die *Gammaverteilung*. Neben Zeitdauern sind häufig auch positive Geldbeträge (beispielsweise Einkommen, Höhen von Autoversicherungsschäden oder Krankenkassenrechnungen) und physikalische Phänomene (z. B. Regenfallmenge, Windstärken) ungefähr gammaverteilt.

Die Gammaverteilung ist zudem wichtig in der Statistik, da ein Spezialfall, die sogenannte *Chi-quadrat-Verteilung*, ein Modell für Stichprobenvarianzen und andere Masse der Abweichung ist.

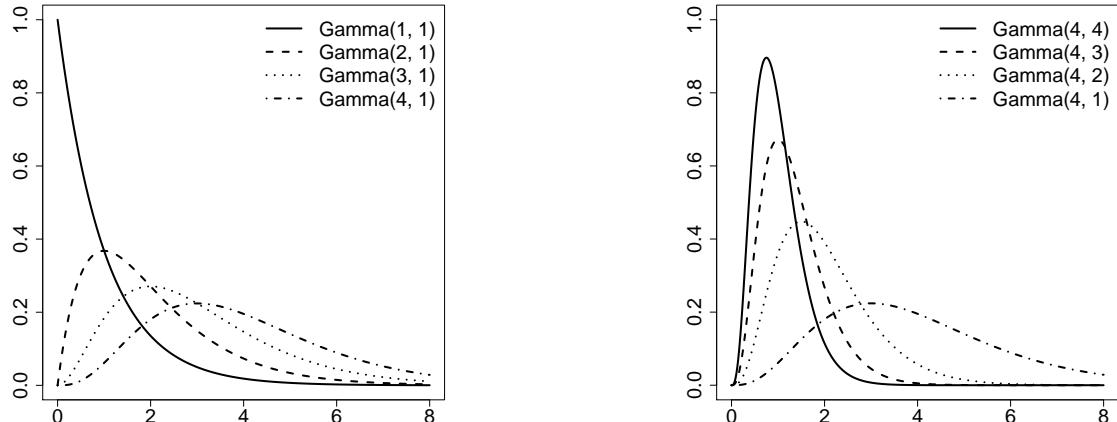


Abbildung 3.29: Dichten einiger Gammaverteilungen. Im linken Bild wird die Rolle des Parameters a (Form), im rechten jene von b (Rate) gezeigt.

Die Gammaverteilung mit Parametern $a > 0$ (Form) und $b > 0$ (Rate), kurz $\text{Gamma}(a, b)$, ist für $x \geq 0$ durch ihre Dichtefunktion

$$f(x) := \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$$

definiert, siehe Abbildung 3.29 für einige Beispiele.

Γ ist die sogenannte *Gammafunktion*

$$\Gamma(a) := \int_0^\infty t^{a-1} e^{-t} dt.$$

Für ganzzahlige a entspricht sie $1 \cdot 2 \cdots (a-1) = (a-1)!$. Die Gammafunktion ist somit eine Verallgemeinerung des Begriffs ‘Fakultät’.

¹Mit der Student-Quantilfunktion `qt` in R.

²Mit der Student-Verteilungsfunktion `pt` in R.

Eigenschaften

1. Die Dichte von $\text{Gamma}(a, b)$ ist rechtsschief und unimodal mit Modus an der Stelle $(a - 1)/b$ (falls $a \geq 1$).
2. Wenn a wächst und b fix bleibt, wird die Dichte symmetrischer.
3. Wenn b wächst und a fix bleibt, wird die Dichte schmäler.
4. $\text{Gamma}(1, b)$ entspricht der Exponentialverteilung mit Parameter b .
5. Die Summe von unabhängigen, gammaverteilten Zufallvariablen mit gleichem b ist auch gammaverteilt:
Für $X \sim \text{Gamma}(a_1, b)$ und $Y \sim \text{Gamma}(a_2, b)$ folgt $X + Y \sim \text{Gamma}(a_1 + a_2, b)$.
6. Dank den letzten zwei Eigenschaften ist die Summe von n unabhängigen $\text{Exp}(\lambda)$ -verteilten Zufallsvariablen gammaverteilt mit Parametern n und λ . Das rechte Bild von Abbildung 3.26 zeigt die Dichten von $n = 2, 4, 6, 8, 10, 15$ aufaddierten, unabhängigen $\text{Exp}(1)$ -verteilten Zufallsvariablen. Diese entsprechen demnach den Dichten von $\text{Gamma}(n, 1)$. Mit dieser Eigenschaft lassen sich auch die Formeln für Erwartungswert und Varianz intuitiv begründen.
7. $E(X) = a/b$.
8. $\text{Var}(X) = a/b^2$.

Verteilungs- und Quantilfunktion F und F^{-1} lassen sich nur für den Spezialfall der Exponentialverteilung explizit angeben. Für die Berechnung von Wahrscheinlichkeiten und Quantilen sind wir ansonsten auf Software angewiesen¹.

Beispiel 3.76 (Einkommen). Aus grossen Studien sei bekannt, dass der monatliche Bruttolohn X bei den ArbeitnehmerInnen in der Schweiz ungefähr gammaverteilt ist mit Parametern $a = 3$ und $b = 0.0005$. Das mittlere Einkommen beträgt somit $a/b = 6'000$ CHF und die Standardabweichung ist $\sqrt{\text{Var}(X)} = \sqrt{a/b} = 3'464.1$ CHF. Mithilfe der Software finden wir beispielsweise:

- Der Anteil der ArbeitnehmerInnen mit mindestens 10'000 CHF Lohn entspricht der Wahrscheinlichkeit, dass eine zufällig ausgewählte Person mindestens 10'000 CHF pro Monat verdient, also

$$P(X \geq 10'000) = 1 - F(10'000) \approx 0.125 = 12.5\%.$$

- 10% der ArbeitnehmerInnen verdienen weniger als $F^{-1}(0.1) = 2'204.1$ CHF.
- 10% der ArbeitnehmerInnen verdienen mehr als $F^{-1}(0.9) = 10'644.6$ CHF.
- Die Hälfte verdient weniger/mehr als $F^{-1}(0.5) = 5'348.1$ CHF. ▲

3.4.2 Chiquadrat-Verteilung

Die Gammaverteilung mit Parametern $a = k/2$ und $b = 1/2$ heisst *Chiquadrat-Verteilung (χ^2 -Verteilung)* mit k Freiheitsgraden, kurz χ_k^2 , mit β -Quantil $\chi_{k,\beta}^2$. Abbildung 3.30 zeigt Dichten und Verteilungsfunktionen einiger Chiquadrat-Verteilungen. Die Chiquadrat-Verteilung „erbt“ ihre Eigenschaften von der Gammaverteilung. Beispielsweise beträgt der Erwartungswert $a/b = k$ und die Varianz $a/b^2 = 2k$.

Für n unabhängige standardnormalverteilte Zufallsvariablen Z_1, Z_2, \dots, Z_n gibt die Chiquadrat-Verteilung mit n Freiheitsgraden die Verteilung von $\sum_{i=1}^n Z_i^2$ an. Auf diesem wichtigen Resultat beruhen einige wichtige Anwendungen der Statistik. Zwei davon zeigen wir in den nächsten beiden Abschnitten.

¹Die Verteilungsfunktion in R heisst `pgamma(x, a, b)`, die Quantilfunktion `qgamma(beta, a, b)`.

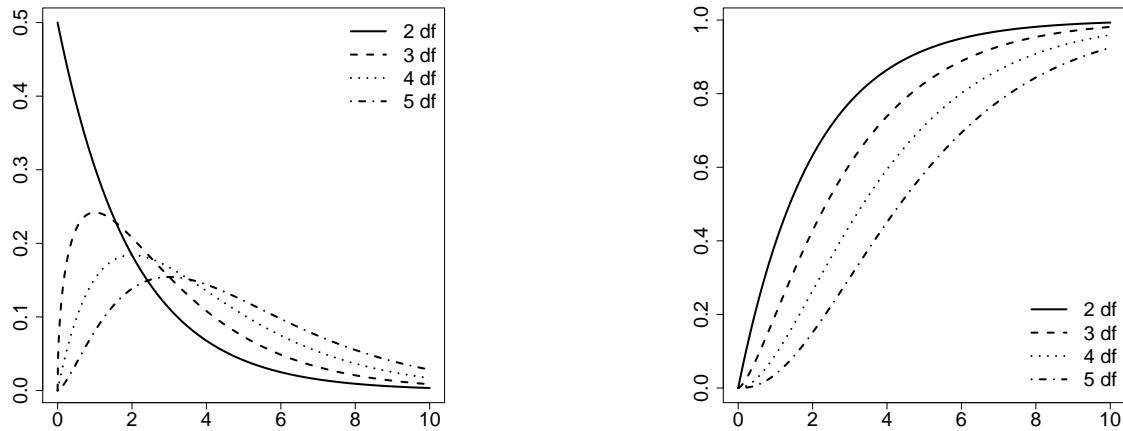


Abbildung 3.30: Das linke Bild zeigt Dichtefunktionen, das rechte entsprechende Verteilungsfunktionen verschiedener Chi-quadrat-Verteilungen.

Hinweise zu obigem Resultat

- Für $n = 1$ besagt es, dass das Quadrat einer standardnormalverteilten Zufallsvariable eine Chi-quadrat-Verteilung mit einem Freiheitsgrad besitzt.
- Für grosse n besagt es zusammen mit dem Zentralen Grenzwertsatz, dass eine Chi-quadrat-verteilte Zufallsvariable mit vielen Freiheitsgraden ungefähr normalverteilt ist.

3.4.3 Konfidenzintervalle für eine Standardabweichung

Die Chi-quadrat-Verteilung kann verwendet werden, um Konfidenzintervalle für eine Standardabweichung zu finden, beispielsweise zur Festlegung der Genauigkeit eines Messinstruments.

Man kann zeigen, dass die normierte Stichprobenvarianz

$$(n-1)S^2/\sigma^2$$

von unabhängigen normalverteilten Beobachtungen mit Varianz σ^2 chi-quadrat-verteilt ist mit $n-1$ Freiheitsgraden.

Daraus folgen die Gleichungen

$$\left. \begin{aligned} & P((n-1)S^2/\sigma^2 \leq \chi_{n-1;1-\alpha}^2) \\ & P((n-1)S^2/\sigma^2 \geq \chi_{n-1;\alpha}^2) \\ & P(\chi_{n-1;\alpha/2}^2 \leq (n-1)S^2/\sigma^2 \leq \chi_{n-1;1-\alpha/2}^2) \end{aligned} \right\} = 1 - \alpha.$$

Die Ungleichungen innerhalb $P(\cdot)$ kann man nach σ auflösen und erhält die folgenden $(1 - \alpha)$ -Konfidenzintervalle für σ , mit denen z. B. auch Hypothesen über σ geprüft werden können:

- Die untere Konfidenzschranke

$$S \sqrt{\frac{n-1}{\chi_{n-1;1-\alpha}^2}}$$

- Die obere Konfidenzschranke

$$S \sqrt{\frac{n-1}{\chi_{n-1;\alpha}^2}}$$

- Das Konfidenzintervall

$$\left[S \sqrt{\frac{n-1}{\chi^2_{n-1;1-\alpha/2}}}, S \sqrt{\frac{n-1}{\chi^2_{n-1;\alpha/2}}} \right]$$

Für grössere n gelten die Ergebnisse approximativ auch für nicht unbedingt normalverteilte Beobachtungen.

Beispiel 3.77 (Weihnachtsguezli). Ein Hersteller von Weihnachtsguezli soll dem Zwischenhändler Säcke à 500 Gramm liefern. Während diese Abmachung *im Schnitt* leicht zu bewerkstelligen ist (z. B. eine halbe Tonne irgendwie auf 1000 Säcke verteilen), ist es nicht möglich, jeden Sack mit exakt 500 g Guezli zu füllen. Deshalb hat der Hersteller versprochen, die Standardabweichung σ des Gewichts der Säcke auf 10 g zu beschränken. (Mit den Faustregeln zur Normalverteilung kann man abschätzen, dass dann etwa 95% der Säcke zwischen 480 und 520 g wiegen würden.)

Ob die Abmachung eingehalten wird, soll mithilfe einer Stichprobe von 100 Säcken auf dem 5%-Niveau verifiziert werden. Die Standardabweichung der $n = 100$ Gewichtsangaben betrage $S = 8$ g. Zusammen mit dem 5%-Quantil $\chi^2_{99;0.05} = 77$ der Chi-quadrat-Verteilung mit 99 Freiheitsgraden ergibt dies eine obere Konfidenzschanke für σ von

$$S \sqrt{\frac{n-1}{\chi^2_{n-1;\alpha}}} = 8 \sqrt{\frac{99}{77}} \approx 9.07 \text{ g.}$$

Damit kann der Hersteller mit einer Sicherheit von ungefähr 95% behaupten, dass σ höchstens 9.07 g beträgt und er insbesondere mit einer Sicherheit von rund 95% die Abmachung ($H_1 : \sigma < 10$) einhält. ▲

3.4.4 Anpassungstest für ein kategorielles Merkmal

Manchmal möchte man anhand von Daten zeigen, dass ein Merkmal anders verteilt ist als vorgegeben bzw. als erwartet, beispielsweise, dass nicht alle Ziffern bei den Leuten gleich beliebt sind. Dies wird mit einem sogenannten *Anpassungstest* gemacht. Wir präsentieren hier den wichtigen *Chi-quadrat-Anpassungstest* für kategoriale Merkmale.

Wir betrachten ein kategorielles Merkmal X mit den Kategorien x_1, \dots, x_K und unbekannten relativen Anteilen p_1, \dots, p_K . Um etwas über die p_j herauszufinden, besorgen wir uns eine Stichprobe von n unabhängigen Beobachtungen X_1, \dots, X_n , die alle wie X verteilt sind. Naheliegende Schätzwerte für die p_j sind dann die relativen Häufigkeiten $\hat{p}_j = H_j/n$.

Will man auf dem α -Niveau die Nullhypothese

$$H_0: p_j = p_j^o \text{ für } j = 1, 2, \dots, K,$$

prüfen, also dass die tatsächlichen Anteile p_j gleich wie gewisse vorgegebene Anteile p_j^o sind, könnte man pro Kategorie einen Binomialtest durchführen, wobei man wegen multiplem Testens das vorsichtigere Niveau α/K verwenden sollte.

Chi-quadrat-Anpassungstest Statt mit mehreren Binomialtests zu arbeiten, kann eine solche Nullhypothese auch mit einem einzigen Test, dem Chi-quadrat-Anpassungstest, geprüft werden. Die beiden Vorgehen führen oft, jedoch nicht immer, zum gleichen Testentscheid.

Für jede Kategorie x_j wird das sogenannte *Pearson-Residuum*

$$e_j := \frac{H_j - np_j^o}{\sqrt{np_j^o}}$$

gebildet. Es gibt an, wie stark die beobachtete von der (unter H_0) erwarteten Häufigkeit abweicht. Zu seltene Kategorien haben negative Residuen, zu häufige positive.

Pearsons *Chi*²-*Teststatistik*

$$T := \sum_{j=1}^K e_j^2,$$

also die Summe der quadrierten Pearson-Residuen, ist dann ein Mass für die gesamte Abweichung und misst die Evidenz gegen H_0 . Dank dem speziellen Nenner der Pearson-Residuen ist T unter der Nullhypothese approximativ nach χ^2_{K-1} verteilt.

Überschreitet T die kritische Schranke $\chi^2_{K-1;1-\alpha}$ bzw. unterschreitet der p -Wert $P(Y \geq T) = 1 - F_{K-1}(T)$ das Niveau α , so verwerfen wir die Nullhypothese und behaupten mit einer Sicherheit von ungefähr $(1 - \alpha) \cdot 100\%$, dass die tatsächliche Verteilung von X von jener der Nullhypothese abweicht.

(F_{K-1} bezeichnet die Verteilungsfunktion von χ^2_{K-1} und Y eine entsprechend verteilte Zufallsvariable.)

Hinweise zum Chi²-Anpassungstest

- Bei nur zwei Kategorien arbeitet man besser mit dem präziseren Binomialtest.
- Wie üblich kann man nicht nachweisen, dass die Nullhypothese stimmt, also dass eine bestimmte Verteilung vorliegt.
- Bei einem numerischen Merkmal kann man den χ^2 -Anpassungstest via Kategorisierung anwenden.

Beispiel 3.78 (Zufallsziffern). Nun wollen wir anhand des Vorlesungsdatensatzes von Beispiel 1.1 auf dem 5%-Niveau nachweisen, dass Personen keine guten Zufallsgeneratoren sind, bzw. dass nicht alle Ziffern mit Wahrscheinlichkeit $p_j^o = 1/10$ genannt werden. Bezeichnen wir die entsprechenden unbekannten Anteile in der Population mit p_1 bis p_{10} , so prüfen wir die Nullhypothese, dass $p_j = p_j^o = 1/10$ für $j = 1, \dots, 10$ versus die Arbeitshypothese, dass nicht alle Anteile $1/10$ betragen.

Falls die Nullhypothese gälte, hätten sich die $n = 262$ Antworten im Idealfall gleichmässig auf die 10 Ziffern verteilt, so dass jede Ziffer $np_j^o = 26.2$ mal genannt worden wäre. Die Realität sieht ganz anders aus:

Ziffer	0	1	2	3	4	5	6	7	8	9
Absolute Häufigkeit	8	6	12	32	25	23	28	70	41	17

Summieren wir die 10 quadrierten Pearson-Residuen (exemplarisch beträgt jenes der Ziffer ‘0’ $e_1 = (8 - 26.2)/\sqrt{26.2} = -3.5557$), finden wir den Wert 122.58 der Teststatistik. Er ist grösser als die kritische Schranke¹ $\chi^2_{9;0.95} = 16.919$, deshalb verwerfen wir die Nullhypothese auf dem 5%-Niveau und behaupten mit einer Sicherheit von rund 95%, dass Personen keine guten Zufallsgeneratoren sind. Zum selben Ergebnis kommen wir mit dem p -Wert² $1 - F_9(122.58) \approx 0$.

Der Chi²-Anpassungstest kann auch direkt mit der Software³ durchgeführt werden:

R Code

```
# Eingabe, Forts.
a.H <- table(wiso$ZufZiffer)
chisq.test(a.H)

# Ausgabe
[...]
X-squared = 122.58, df = 9, p-value < 2.2e-16
```

¹Das 95%-Quantil der Chi²-Verteilung mit 9 Freiheitsgraden ermitteln wir mit dem R-Aufruf `qchisq(0.95, 9)`.

²Hier verwenden wir die Verteilungsfunktion `pcchisq` von R.

³Mit der R-Funktion `chisq.test`; Pearson-Residuen würden mit `residuals` angezeigt.

Beispiel 3.79 (Rauchen, Fortsetzung). In Beispiel 3.58 sind wir der Frage nachgegangen, ob StudentInnen ein anderes Rauchverhalten aufweisen als die Gesamtbevölkerung.

In der Stichprobe waren die relativen Häufigkeiten der Kategorien von ‘Rauchen’ mit $H_1/n = 171/261 = 0.655$ (NichtraucherInnen), $H_2/n = 47/261 = 0.180$ (GelegenheitsraucherInnen) und $H_3/n = 43/261 = 0.165$ (regelmässige RaucherInnen) leicht anders als die entsprechenden Anteile $p_1^o = 0.59$, $p_2^o = 0.21$ und $p_3^o = 0.20$ in der Gesamtbevölkerung. Mithilfe dreier Binomialtests (inkl. Bonferroni-Korrektur für multiples Testen) konnten wir auf dem 5%-Niveau nicht nachweisen, dass sich StudentInnen hinsichtlich des Rauchverhaltens systematisch von der Gesamtbevölkerung unterscheiden. Die tatsächlichen relativen Anteile der Rauchkategorien bei StudentInnen haben wir mit p_1 , p_2 und p_3 bezeichnet.

Nun wollen wir die Nullhypothese

$$p_1 = p_1^o (= 0.59), p_2 = p_2^o (= 0.21), p_3 = p_3^o (= 0.20)$$

auch mit dem Chiquadrat-Anpassungstest prüfen.

Die Pearson-Residuen betragen

$$\begin{aligned} e_1 &= (171 - 261 \cdot 0.59) / \sqrt{261 \cdot 0.59} \approx 1.371, \\ e_2 &= (47 - 261 \cdot 0.21) / \sqrt{261 \cdot 0.21} \approx -1.055 \text{ und} \\ e_3 &= (43 - 261 \cdot 0.20) / \sqrt{261 \cdot 0.20} \approx -1.273. \end{aligned}$$

Anhand des Vorzeichens können wir beispielsweise sagen, dass es im Datensatz mehr NichtraucherInnen gibt, als man unter der Nullhypothese erwarten würde.

Der konkrete Wert der Chiquadrat-Teststatistik beträgt

$$T = e_1^2 + e_2^2 + e_3^2 = 1.371^2 + (-1.055)^2 + (-1.273)^2 \approx 4.613$$

und ist damit nicht grösser als das 95%-Quantil der Chiquadrat-Verteilung mit zwei Freiheitsgraden (5.99). Somit können wir nicht behaupten, dass StudentInnen ein anderes Rauchverhalten an den Tag legen als die Gesamtbevölkerung. Zum gleichen Testentscheid gelangen wir mit dem p -Wert $1 - F_2(4.613) = 0.0996$. Abbildung 3.31 zeigt die Situation grafisch.

Verifizieren wir diese Berechnungen nun noch mit der Software:

R Code

```
# Eingabe
tab.rauchen <- table(wiso$Rauchen)
x <- chisq.test(tab.rauchen, p=c(0.59, 0.21, 0.2))
x

# Ausgabe
[...]
X-squared = 4.6133, df = 2, p-value = 0.0996
```

Mit dem Chiquadrat-Anpassungstest sind wir hier zum gleichen Testentscheid wie mit den drei Binomialtests und der Bonferroni-Korrektur gekommen. ▲

3.5 Zusammenfassung

- In diesem Kapitel haben wir die Grundlagen der Wahrscheinlichkeitsrechnung kennengelernt, mit deren Hilfe wir unter anderem verschiedene konkrete Verfahren der univariaten schliessenden Statistik eingeführt haben. Letztere erlaubt generell, mit Schätzern, Konfidenzintervallen und/oder Tests basierend auf einer Zufallsstichprobe Rückschlüsse auf die Grundgesamtheit zu machen.

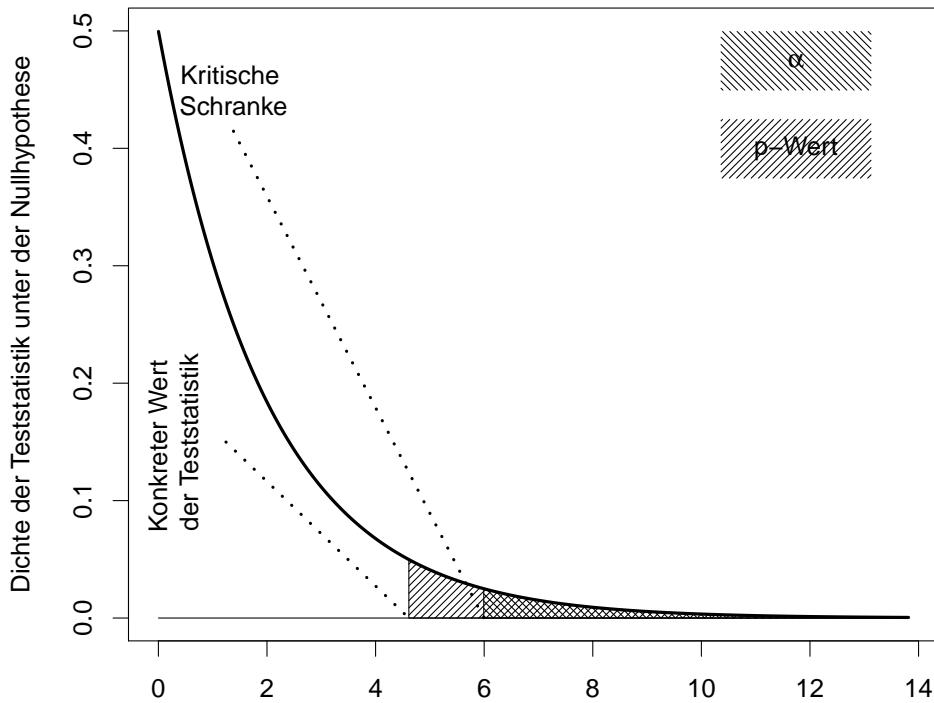


Abbildung 3.31: Dichtefunktion der Chi-quadrat-Verteilung mit 2 Freiheitsgraden inkl. p -Wert und Niveau $\alpha = 0.05$ von Beispiel 3.79 als schraffierte Flächen unter der Dichte.

- Die Wahrscheinlichkeitsrechnung kümmert sich um zufällige Vorgänge. Ein solcher wird durch seine Verteilung charakterisiert, also durch alle mit seinen Ergebnissen verbundenen Wahrscheinlichkeiten. Da man lieber mit Zahlen als mit Ergebnissen von Vorgängen arbeitet, werden letztere in der Regel durch Zufallsvariablen quantifiziert. Zufallsvariablen und ihre Verteilungen sind sehr wichtig in der Statistik, da die Erzeugung einer Stichprobe oft als zufälliger Vorgang gilt. (Zahlenkodierte) Merkmale und deren Beobachtungen sowie Kenngrößen bzw. Schätzer sind dann Zufallsvariablen, die eine solche Zufallsstichprobe quantifizieren.
- Wir haben mit Wahrscheinlichkeits- bzw. Dichtefunktion, Verteilungsfunktion sowie Quantilfunktion drei Möglichkeiten kennengelernt, mit denen man (diskrete und stetige) Verteilungen effizient und vollständig beschreiben kann. Erwartungswert und Varianz bzw. Standardabweichung beschreiben spezifische Aspekte der Verteilung, nämlich deren mittlere Lage und Streuung. Dabei haben wir die enge Verbindung zwischen deskriptiver Statistik und Wahrscheinlichkeitsrechnung herausgestrichen und als zusätzliche Brücke die häufig eingesetzte Methode der Simulation präsentiert.
- Ist die Verteilung eines Merkmals in der Population ungefähr bekannt (evtl. mit aus Daten geschätzten Parametern), so können damit Fragen nach relativen Anteilen, Quantilen, Mittelwerten etc. beantwortet werden, ohne über die Daten zu verfügen. Wir haben einige häufig verwendete Verteilungen kennengelernt, z. B. die Exponentialverteilung.
- Besonderes Augenmerk haben wir auf die Binomialverteilung, die Normalverteilung und die Gamma-verteilung gelegt, da sie zudem die Verteilung von wichtigen Schätzern beschreiben. Dies ist die Basis, um Konfidenzintervalle und Tests zu konstruieren.
- Wir haben festgestellt, dass absolute Häufigkeiten binomialverteilt sind. Darauf beruhen Binomialkonfi-

denzintervalle und -tests für relative Anteile, mit denen wir das allgemeine Konzept des Konfidenzintervalls und des Hypothesentests erläutert haben. Eine andere wichtige Anwendung dieser Verteilung sind Konfidenzintervalle für Quantile. Dank der engen Verbindung zwischen Konfidenzintervallen und Tests lassen sich damit beispielsweise Hypothesen zum Median testen. Gewisse Fragestellungen zu absoluten Häufigkeiten von seltenen Ereignissen lassen sich mit der Poissonverteilung beantworten.

- Ausgehend vom Zentralen Grenzwertsatz haben wir mithilfe der Normalverteilung einfache Konfidenzintervalle und Tests für Populationsmittelwerte eingeführt. In der Praxis werden diesen Z-Verfahren die eng damit verbundenen Verfeinerungen nach Students Methode vorgezogen.
- Die Tatsache, dass standardisierte Varianzen häufig ungefähr chiquadrat-verteilt sind, also eine spezielle Gammaverteilung aufweisen, erlaubt die Konstruktion von Konfidenzintervallen und Tests für Standardabweichungen. Damit verbunden ist auch der Chiquadrat-Anpassungstest, mit dem die Nullhypothese geprüft werden kann, dass ein kategorielles Merkmal eine gewisse Verteilung aufweist.

Teil II

Bivariate Verfahren

Viele statistische Fragestellungen laufen darauf hinaus, den Zusammenhang zwischen zwei Merkmalen X und Y aus dem gleichen Datensatz mit Hilfe geeigneter Grafiken, Kenngrößen und Verfahren der schliessenden Statistik zu untersuchen.

Typische Fragestellungen lauten:

- Wie viel verdienen Frauen im Schnitt weniger als Männer?
- Wie hängt der mittlere Mietpreis von der Wohnungsgröße ab?
- Welchen Effekt hat das Alter auf die mittleren Krankheitskosten?
- Rauchen mehr Männer als Frauen?

Eine zentrale Rolle kommt dabei der *gemeinsamen Verteilung* der X - und Y -Werte zu. Diese gibt – salopp gesagt – an, welche Kombinationen von X - und Y -Werten wie häufig auftreten. Sind gewisse Kombinationen zu selten oder zu häufig, als man unter *Unabhängigkeit* von X und Y (keinerlei Zusammenhang zwischen den Merkmalen) erwarten würde, so liegt ein entsprechender Zusammenhang zwischen den X - und Y -Werten vor. Eng damit verbunden ist folgende alternative Idee: Teile die Daten anhand des einen Merkmals, sagen wir X , in mehrere Teilstichproben auf und vergleiche die Verteilungen der Y -Werte zwischen den Teilstichproben. Unterscheiden sich diese auf X bedingten Verteilungen von Y , so liegt ein entsprechender Zusammenhang zwischen den X - und Y -Werten vor.

Die konkrete Umsetzung dieser Ansätze hängt von den Variabtentypen der zwei Merkmale ab: In Kapitel 4 betrachten wir zwei kategoriale Merkmale, in Kapitel 5 ein kategorielles und ein numerisches Merkmal und schliesslich in Kapitel 6 zwei numerische Merkmale.

Ein statistischer Zusammenhang zwischen zwei Merkmalen heisst noch nicht, dass sich eines der beiden kausal (ursächlich) auf das andere auswirkt. Denkbar wäre beispielsweise, dass es ein drittes Merkmal gibt (z. B. Vorliegen/Nichtvorliegen eines bestimmten genetischen Faktors), welches sich sowohl auf X als auch auf Y auswirkt, wohingegen kein direkter Zusammenhang zwischen den X - und Y -Werten besteht. Diese Abhängigkeit von einem dritten Merkmal nennt man *Confounding*, das zusätzliche Merkmal heisst *Confounder (Störvariable)*. Da man das Vorhandensein von unbekannten Confoundern nie ausschliessen kann, sollten statistische Zusammenhänge nie kausal gewertet werden.

Wird beispielsweise der Zusammenhang zwischen Einkommen und Geschlecht untersucht, so gelten u. a. die Position innerhalb des Betriebs und das Auftreten bei Lohnverhandlungen als bekannte Confounder. Ein anderes – nicht ganz ernst gemeintes – Beispiel stellt der starke Zusammenhang zwischen Anzahl Störchen und Geburten in europäischen Ländern dar. Er entsteht durch den Confounder ‘Fläche des Landes’, der mit beiden Größen stark zusammenhängt.

Ein bekannter (und verfügbarer) Confounder kann berücksichtigt werden, indem der fragliche Zusammenhang pro Ausprägung des Confounders separat untersucht wird. (Numerische Confounder müssen hierfür kategorisiert werden.) Eine Alternative stellt die Verwendung von statistischen Modellen dar. Darauf gehen wir in Kapitel 7 ein.

Kapitel 4

Zwei kategoriale Merkmale

4.1 Häufigkeitstabellen

Sind beide Variablen X und Y kategorial mit den Kategorien x_1, x_2, \dots, x_L und y_1, y_2, \dots, y_M , so kann man die n Paare (X_i, Y_i) zu einer *Häufigkeitstabelle* bzw. *Kontingenztafel* zusammenfassen. Sie gibt an, wie häufig die Kombinationen der X - und Y -Kategorien sind bzw. wie die gemeinsame Verteilung der beiden Merkmale aussieht. Auf ihr beruhen die statistischen Verfahren zur Beschreibung des Zusammenhangs zwischen zwei kategorialen Merkmalen.

Um diese Verfahren sauber einzuführen, verwenden wir folgende Notation für die allgemeine Häufigkeitstabelle:

	y_1	y_2	\dots	y_M
x_1	$H_{1,1}$	$H_{1,2}$	\dots	$H_{1,M}$
x_2	$H_{2,1}$	$H_{2,2}$	\dots	$H_{2,M}$
\vdots	\vdots	\vdots	\ddots	\vdots
x_L	$H_{L,1}$	$H_{L,2}$	\dots	$H_{L,M}$

Dabei ist

$$H_{j,k} = \text{Anzahl aller Beobachtungen mit } X = x_j \text{ und } Y = y_k.$$

Wir illustrieren diese Darstellungsweise anhand eines Beispiels.

Beispiel 4.1 (Geschlecht und Rauchen). Wir möchten die Häufigkeitstabelle¹ der beiden kategorialen Variablen X ('Geschlecht' mit den $L = 2$ Kategorien $x_1 = M$ sowie $x_2 = W$) und Y ('Rauchen' mit den $M = 3$ Kategorien $y_1 = 0$ (nein), $y_2 = 1$ (gelegentlich) sowie $y_3 = 2$ (regelmässig)) von Beispiel 1.1 betrachten.

R Code

```
# Eingabe
a.H <- table(wiso$Geschlecht, wiso$Rauchen)
a.H

# Ausgabe
  0   1   2
M 100  21  24
W  71  26  19
```

Dabei wurden zwei der 263 Beobachtungen wegen fehlender Y -Werte herausgenommen.

Es gibt somit bspw. $H_{1,1} = 100$ Nichtraucher und $H_{L,M} = H_{2,3} = 19$ regelmässige Raucherinnen. ▲

¹Dazu verwenden wir den R-Befehl `table`.

Oftmals ergänzt man eine solche Tabelle noch um die Zeilensummen (univariate Verteilung der X -Werte)

$$H_{j,+} = \sum_{k=1}^M H_{j,k} = \text{Anzahl aller Beobachtungen mit } X = x_j,$$

die Spaltensummen (univariate Verteilung der Y -Werte)

$$H_{+,k} = \sum_{j=1}^L H_{j,k} = \text{Anzahl aller Beobachtungen mit } Y = y_k$$

sowie den Stichprobenumfang n . Dies ergibt die *erweiterte Häufigkeitstabelle*:

	y_1	y_2	\dots	y_M	
x_1	$H_{1,1}$	$H_{1,2}$	\dots	$H_{1,M}$	$H_{1,+}$
x_2	$H_{2,1}$	$H_{2,2}$	\dots	$H_{2,M}$	$H_{2,+}$
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
x_L	$H_{L,1}$	$H_{L,2}$	\dots	$H_{L,M}$	$H_{L,+}$
	$H_{+,1}$	$H_{+,2}$	\dots	$H_{+,M}$	n

Beispiel 4.2 (Geschlecht und Rauchen, Fortsetzung). In unserem Beispiel sieht die erweiterte Häufigkeitstabelle¹ folgendermassen aus:

Eingabe, Forts.
R Code

```

# Eingabe, Forts.
addmargins(a.H)

# Ausgabe
      0   1   2 Sum
M    100  21  24 145
W     71  26  19 116
Sum  171  47  43 261

```

Hier ist also beispielsweise

$$H_{1,+} = H_{1,1} + H_{1,2} + H_{1,3} = 100 + 21 + 24 = 145$$

und

$$H_{+,2} = H_{1,2} + H_{2,2} = 21 + 26 = 47.$$

(Der Stichprobenumfang entspricht der Anzahl der Beobachtungen ohne fehlende Werte bei ‘Rauchen’ und ‘Geschlecht’.) ▲

4.2 Verdeutlichung des Zusammenhangs

Will man nun den Zusammenhang zwischen den X - und Y -Werten verdeutlichen und anschaulich beschreiben, gibt es zwei Möglichkeiten:

- *Zeilennormierung*

Man unterteilt den Datensatz anhand der Variable X in L Teilgruppen und schaut, ob die Y -Werte in diesen Teilgruppen ähnlich oder sehr unterschiedlich verteilt sind. Dazu vergleicht man die relativen Häufigkeiten der Y -Werte zwischen den Teilgruppen, also die auf X bedingten Verteilungen der Y -Werte. Mit anderen Worten, man normiert alle Zeilen der Häufigkeitstabelle auf Summe eins (100%). Dies wird erreicht, indem jeder Tabelleneintrag $H_{j,k}$ durch seine Zeilensumme $H_{j,+}$ geteilt wird.

¹Dazu wird der R-Befehl `addmargins` auf die Tabelle angewendet.

- *Spaltennormierung*

Man unterteilt den Datensatz anhand der Variable Y in M Teilgruppen und schaut, ob die X -Werte in diesen Teilgruppen ähnlich oder sehr unterschiedlich verteilt sind. Dazu vergleicht man die relativen Häufigkeiten der X -Werte zwischen den Teilgruppen, also die auf Y bedingten Verteilungen der X -Werte. Mit anderen Worten, man normiert alle Spalten der Häufigkeitstabelle auf Summe eins (100%). Dies wird erreicht, indem jeder Tabelleneintrag $H_{j,k}$ durch seine Spaltensumme $H_{+,k}$ geteilt wird.

Beispiel 4.3 (Geschlecht und Rauchen, Fortsetzung). Wir vergleichen die relativen Häufigkeiten von ‘Rauchen’ zwischen Männern und Frauen (Zeilennormierung¹).

R Code

```
# Eingabe, Forts.
prop.table(a.H, margin = 1)

# Ausgabe
      0      1      2
M 0.6897 0.1448 0.1655
W 0.6121 0.2241 0.1638
```

Beispielsweise erhält man den Eintrag oben links mit

$$\frac{H_{1,1}}{H_{1,+}} = \frac{100}{145} \approx 0.6897,$$

jenen unten rechts mit

$$\frac{H_{2,3}}{H_{2,+}} = \frac{19}{116} \approx 0.1638.$$

Kommentare: Der Anteil der Nichtraucher ist bei den Männern grösser als bei den Frauen (68.97% vs. 61.21%), jener der Gelegenthsraucher tiefer (14.48% versus 22.41%). Der Anteil der regelmässigen Raucher (16.55%) ist bei den Männern etwa gleich gross wie bei den Frauen (16.38%). Die zwei auf ‘Geschlecht’ bedingten Verteilungen von ‘Rauchen’ unterscheiden sich also leicht – es scheint einen schwachen Zusammenhang zwischen den beiden Variablen zu geben.

Alternativ vergleichen wir die relativen Häufigkeiten der Variable ‘Geschlecht’ zwischen den drei Kategorien von ‘Rauchen’ (Spaltennormierung²).

R Code

```
# Eingabe, Forts.
prop.table(a.H, margin = 2)

# Ausgabe
      0      1      2
M 0.5848 0.4468 0.5581
W 0.4152 0.5532 0.4419
```

Beispielsweise erhält man den Eintrag oben links mit

$$\frac{H_{1,1}}{H_{+,1}} = \frac{100}{171} \approx 0.5848,$$

jenen unten rechts mit

$$\frac{H_{2,3}}{H_{+,3}} = \frac{19}{43} \approx 0.4419.$$

¹Dazu verwenden wir den R-Befehl `prop.table` mit der Option `margin = 1` (weil wir den Datensatz anhand der *ersten* Variable ‘Geschlecht’ unterteilen).

²Via Option `margin = 2` von `prop.table`.

Kommentar: Man sieht beispielsweise, dass der Männeranteil bei den GelegenheitsraucherInnen (44.68%) kleiner ist als bei den anderen Kategorien (58.48% und 55.81%). Die drei Verteilungen von ‘Geschlecht’ unterscheiden sich also leicht, so dass wir von einem schwachen Zusammenhang sprechen könnten. ▲

Hinweis (Gleiche Zeilen- oder Spaltensummen). Liegt eine Häufigkeitstabelle mit gleichen Zeilensummen vor, so lassen sich die entsprechenden bedingten Verteilungen direkt vergleichen – ein Umrechnen von absoluten auf relative Häufigkeiten via Zeilennormierung ist dann nicht nötig bzw. führt zu den gleichen Erkenntnissen. Analoges gilt bei gleichen Spaltensummen für die Spaltennormierung.

Grafische Darstellung Eine Häufigkeitstabelle bzw. deren Zeilennormierung kann grafisch als *Mosaikdiagramm* dargestellt werden. Es besteht aus L gleich hohen vertikalen Balken, deren Breite proportional zu den relativen Häufigkeiten von X sind. Die Balken sind anhand der auf X bedingten Verteilungen von Y unterteilt. Sie zeigen also die Verteilung der Y -Werte in Abhängigkeit von X . Auf diese Weise wird nicht nur die Zeilennormierung visualisiert, sondern auch die Häufigkeitstabelle bzw. die gemeinsame Verteilung selbst, da die Rechtecksflächen proportional zu den Häufigkeiten der Tabelle sind. (Für die Spaltennormierung müssen die Rollen von X und Y vertauscht werden.)

Dieses Diagramm kann oft besser interpretiert werden als die Häufigkeitstabelle selbst.

Beispiel 4.4 (Geschlecht und Rauchen, Fortsetzung). Betrachten wir das Mosaikdiagramm¹ in Abbildung 4.1: Die linke Spalte (Männer) ist breiter als die rechte, weil der Männeranteil in der Stichprobe über 50% liegt. Man sieht sofort, dass es bei den Männern leicht mehr Nichtraucher, etwas weniger Gelegenheitsraucher und etwa gleichviele regelmässige Raucher wie bei den Frauen gibt.

R Code

```
# Eingabe, Forts.  
plot(a.H)
```

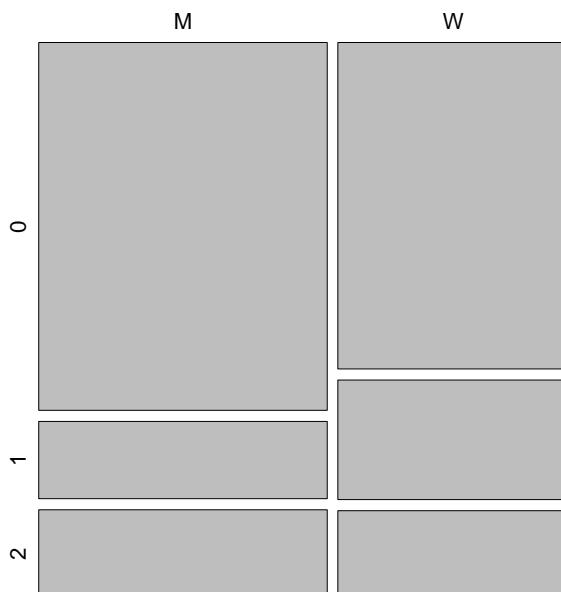


Abbildung 4.1: Mosaikdiagramm der Häufigkeitstabelle von ‘Geschlecht’ und ‘Rauchen’ in Beispiel 4.4.

¹Mit dem R-Befehl `plot`.

Bedingte Wahrscheinlichkeiten Fassen wir X und Y vorübergehend als Zufallsvariablen auf, so können wir die Einträge in der zeilennormierten Tabelle als *bedingte Wahrscheinlichkeiten* $P(Y = y_k | X = x_j)$ auffassen. Eine solche gibt die Wahrscheinlichkeit an, dass $Y = y_k$, falls man bereits weiß, dass $X = x_j$. Es gilt die Beziehung

$$P(Y = y_k | X = x_j) = \frac{P(Y = y_k \text{ und } X = x_j)}{P(X = x_j)}.$$

Ersetzen wir nun die *gemeinsame Wahrscheinlichkeit* im Zähler durch die entsprechende Formel für die spaltennormierten Einträge, so erhalten wir eine Version des sogenannten *Satzes von Bayes*

$$P(Y = y_k | X = x_j) = \frac{P(X = x_j \text{ und } Y = y_k)}{P(X = x_j)} = P(X = x_j | Y = y_k) \cdot \frac{P(Y = y_k)}{P(X = x_j)},$$

mit der wir hier die Spaltennormierung in die Zeilennormierung überführen können und umgekehrt.

Beispiel 4.5 (Geschlecht und Rauchen, Fortsetzung). In Beispiel 4.3 haben wir gesehen, dass 14.48% der Männer gelegentlich rauchen, also dass

$$P(\text{Rauchen} = 1 | \text{Geschlecht} = M) = 0.1448.$$

Tatsächlich finden wir mit

$$P(\text{Rauchen} = 1 \text{ und Geschlecht} = M) = 21/261$$

und

$$P(\text{Geschlecht} = M) = 145/261,$$

dass

$$P(\text{Rauchen} = 1 | \text{Geschlecht} = M) = \frac{P(\text{Rauchen} = 1 \text{ und Geschlecht} = M)}{P(\text{Geschlecht} = M)} = \frac{21/261}{145/261} = 0.1448.$$

Mit dem Satz von Bayes erhalten wir zudem den entsprechenden Eintrag in der spaltennormierten Tabelle:

$$\begin{aligned} & P(\text{Geschlecht} = M | \text{Rauchen} = 1) \\ &= P(\text{Rauchen} = 1 | \text{Geschlecht} = M) \cdot \frac{P(\text{Geschlecht} = M)}{P(\text{Rauchen} = 1)} \\ &= 0.1448 \cdot \frac{145/261}{47/261} = 0.4467. \end{aligned}$$

▲

Das Konzept der bedingten Wahrscheinlichkeit ermöglicht zudem eine neue Charakterisierung von Unabhängigkeit: Falls X und Y unabhängig sind, gilt bekanntlich für jegliche Wahl von x_j und y_k , dass

$$P(Y = y_k \text{ und } X = x_j) = P(Y = y_k)P(X = x_j).$$

Indem wir beide Seiten dieser Gleichung durch $P(X = x_j)$ teilen, folgt

$$\frac{P(Y = y_k \text{ und } X = x_j)}{P(X = x_j)} = P(Y = y_k),$$

also schliesslich

$$P(Y = y_k | X = x_j) = P(Y = y_k).$$

Unter Unabhängigkeit sind also die auf X bedingten Verteilungen von Y alle gleich – insbesondere gleich wie die univariate Verteilung von Y . (Dies gilt auch für vertauschte Rollen von X und Y .)

4.3 Stärke des Zusammenhangs

Wenn es keinerlei Zusammenhang zwischen den X - und Y -Werten gäbe, würde man damit rechnen, dass je zwei Zeilen bzw. Spalten proportional zueinander sind. Tatsächlich kann man zeigen, dass die folgenden drei Aussagen äquivalent sind:

- (a) Die normierten Zeilen der Häufigkeitstabelle sind identisch.
- (b) Die normierten Spalten der Häufigkeitstabelle sind identisch.
- (c) Für jede absolute Häufigkeit gilt

$$H_{j,k} = \frac{H_{j,+} H_{+,k}}{n}.$$

Eine (und damit alle) dieser drei Bedingungen ist selten perfekt erfüllt. Um aber die Stärke des Zusammenhangs zwischen den X - und Y -Werten zu quantifizieren, kann man messen, wie sehr sich die tatsächlichen Tabelleneinträge $H_{j,k}$ von den “idealisierten Werten”

$$\bar{H}_{j,k} := \frac{H_{j,+} H_{+,k}}{n}$$

unterscheiden. Letztere sind im Allgemeinen nicht ganzzahlig. Sie geben die Einträge an, die unter Unabhängigkeit der X - und Y -Werte (also anhand der univariaten Verteilungen der X - und Y -Werte) erwartet würden.

Wie beim Chi-quadrat-Anpassungstest werden die *Pearson-Residuen*

$$e_{j,k} := \frac{H_{j,k} - \bar{H}_{j,k}}{\sqrt{\bar{H}_{j,k}}}$$

als speziell normierte Abweichungen von den idealisierten Werten gebildet. Die Summe

$$\chi^2 = \sum_{j=1}^L \sum_{k=1}^M e_{j,k}^2$$

der quadrierten Abweichungen wird wiederum *Chi-quadrat-Teststatistik* genannt. Sie misst die Stärke des Zusammenhangs und stellt damit ein erstes Beispiel eines *Zusammenhangsmasses* dar, neben Lage- und Streuungsmassen eine weitere Art von Kenngrösse.

Ob ein konkreter Wert der Chi-quadrat-Teststatistik mit einem starken oder schwachen Zusammenhang verbunden ist, lässt sich nicht so leicht sagen, da eine solche Abschätzung vom Stichprobenumfang n und der Anzahl Kategorien L und M der beiden Merkmale abhängt. Deshalb betrachtet man oft eine standardisierte Variante davon, *Cramérs V*¹:

$$V := \sqrt{\frac{\chi^2}{n(\min\{L, M\} - 1)}}$$

Dieses Zusammenhangsmass liegt stets zwischen 0 (kein Zusammenhang) und 1 (“perfekter” Zusammenhang). Es ergänzt die oftmals vage Beurteilung der Stärke des Zusammenhangs mittels Zeilen- oder Spaltennormierung.

Hinweise zu den Pearson-Residuen

- Negative Pearson-Residuen identifizieren zu seltene, positive zu häufige Kombinationen von X - und Y -Ausprägungen, als man unter Unabhängigkeit erwarten würde.

¹Eng mit Cramérs V verbunden ist der sogenannte *Kontingenzkoeffizient C*.

- Je grösser der absolute Betrag, je gewichtiger der Beitrag zur Chiquadrat-Teststatistik.
- Mögliche Faustregel: Absolute Werte bis 1 sind unauffällig, solche ab 2 auffällig gross.
- Pearson-Residuen sind erstens Ausgangslage für die Berechnung der Chiquadrat-Teststatistik bzw. von Cramérs V , zweitens helfen sie, den Zusammenhang zu charakterisieren.

Hinweise zu Cramers V

- Der Ausdruck “ $\min\{L, M\}$ ” entspricht der kleineren Anzahl Kategorien von X und Y . Für eine (2×3) -Häufigkeitstabelle ist $\min\{2, 3\} = 2$, für eine (4×4) -Tabelle $\min\{4, 4\} = 4$.
- Grobe Faustregel: Bei Werten zwischen 0 und 0.1 spricht man manchmal von einem schwachen Zusammenhang, bei Werten ab 0.3 von einem deutlichen bzw. starken. Natürlich hängt eine sinnvolle Einschätzung stets von der jeweils betrachteten Situation ab.

Beispiel 4.6 (Geschlecht und Rauchen, Fortsetzung). Die Tabelle der idealisierten Werte für Beispiel 4.1 lautet

	0	1	2
M	95	26.111	23.889
W	76	20.889	19.111

Den Wert oben links erhalten wir exemplarisch mit

$$\bar{H}_{1,1} = \frac{H_{1,+}H_{+,1}}{261} = \frac{145 \cdot 171}{261} = 95,$$

den Wert unten rechts mit

$$\bar{H}_{2,3} = \frac{H_{2,+}H_{+,3}}{261} = \frac{116 \cdot 43}{261} = 19.111.$$

Daraus findet man die Tabelle der Pearson-Residuen:

	0	1	2
M	0.5130	-1.0002	0.0227
W	-0.5735	1.1183	-0.0254

Den Wert oben links erhält man exemplarisch durch

$$e_{1,1} = \frac{H_{1,1} - \bar{H}_{1,1}}{\sqrt{\bar{H}_{1,1}}} = \frac{100 - 95}{\sqrt{95}} \approx 0.5130,$$

den Wert unten rechts durch

$$e_{2,3} = \frac{H_{2,3} - \bar{H}_{2,3}}{\sqrt{\bar{H}_{2,3}}} = \frac{19 - 19.111}{\sqrt{19.111}} \approx -0.0254.$$

Kommentar: Männliche Gelegenheitsraucher sind also leicht seltener und weibliche Gelegenheitsraucher leicht häufiger, als man anhand der univariaten Verteilungen von ‘Rauchen’ und ‘Geschlecht’ unter Unabhängigkeit erwarten würde. Die Beschreibung des Zusammenhangs anhand Zeilen- oder Spaltennormierung von Beispiel 4.3 könnte durch diese Feststellung ergänzt werden.

Den Wert der Chiquadrat-Teststatistik findet man als Summe der quadrierten Pearson-Residuen, also durch

$$\chi^2 = 0.5130^2 + (-1.0002)^2 + \dots + (-0.0254)^2 = 2.8443.$$

Daraus gelangen wir zu Cramérs V :

$$V = \sqrt{\frac{2.8443}{261 \cdot (2-1)}} = \sqrt{\frac{2.8443}{261}} \approx 0.104.$$

Es weist auf einen schwachen Zusammenhang zwischen den beiden Variablen hin.

Nun verifizieren wir diese Berechnungen mit den entsprechenden R-Aufrufen¹:

R Code

```
# Eingabe, Forts.
zushang <- chisq.test(a.H)

# Eingabe, Forts.: Idealisierte Werte
zushang$expected

Ausgabe
      0       1       2
M 95 26.111 23.889
W 76 20.889 19.111

# Eingabe, Forts.: Pearson-Residuen
residuals(zushang)

# Ausgabe
      0       1       2
M 0.512989 -1.000236 0.022733
W -0.573539 1.118298 -0.025416

# Eingabe, Forts.: Chiquadrat-Teststatistik
zushang$statistic

# Ausgabe
X-squared: 2.8443

# Eingabe, Forts.: Cramérs V
cramers.V(zushang)

# Ausgabe
Cramers.V: 0.10439
```



4.4 Aussagen über die Population

Mit einer Zufallsstichprobe können Aussagen über den Zusammenhang in der Population, also den wahren/echten/tatsächlichen Zusammenhang, gemacht werden. Wie üblich dienen die relevanten Werte in der Stichprobe (z. B. relative Häufigkeiten der Zeilen- oder Spaltennormierung, Cramérs V) als naheliegende Schätzwerte für die entsprechenden Werte in der Population und es können Konfidenzintervalle dafür ausgewiesen werden, beispielsweise die bereits bekannten Binomialkonfidenzintervalle für die wahren Anteile der Zeilen- oder Spaltennormierung.

Besonderes Augenmerk legen wir auf (approximative) Konfidenzintervalle für den Populationswert θ von Cramérs V . Eine untere Konfidenzschanke gibt beispielsweise an, wie stark der Zusammenhang zwischen den betrachteten Merkmalen tatsächlich mit hoher Sicherheit *mindestens* ist. Ist die untere 95%-Konfidenzschanke für θ grösser als null, so lässt sich entsprechend mit einer Sicherheit von rund 95%

¹Idealisierte Werte, Pearson-Residuen und Chiquadrat-Teststatistik lassen sich in R ausgehend vom Befehl `chisq.test` anzeigen. Das Cramérs V finden wir mit der im Anhang aufgeführten R-Funktion `cramers.V`.

behaupten, dass θ grösser als null ist bzw. dass ein echter Zusammenhang vorliegt. Dies ermöglicht den für bivariate Fragestellungen oft zentralen *Test auf Zusammenhang*, bei dem die Arbeitshypothese “Es gibt einen echten Zusammenhang” versus die Nullhypothese “Es gibt keinerlei Zusammenhang” geprüft wird. Bezogen auf das wahre Cramérs V , also θ , lauten die Hypothesen entsprechend $H_1 : \theta > 0$ versus $H_0 : \theta = 0$.

Zum gleichen Testentscheid führt der sogenannte χ^2 -*Unabhängigkeitstest*: Dieser Test beruht auf der Tatsache, dass die ChiQuadrat-Teststatistik unter der Nullhypothese und bei gegebenen Zeilen- und Spaltensummen ungefähr eine ChiQuadrat-Verteilung mit $(L - 1)(M - 1)$ Freiheitsgraden hat¹.

Beispiel 4.7 (Geschlecht und Rauchen, Fortsetzung). Betrachten wir nun die befragten StudentInnen als Zufallsstichprobe aus allen jungen SchweizerInnen. Was können wir über die Stärke des echten Zusammenhangs² zwischen ‘Rauchen’ und ‘Geschlecht’ sagen?

R Code

```
# Eingabe, Forts.: ChiQuadrat-Unabhängigkeitstest
zushang <- chisq.test(a.H)
zushang

# Ausgabe
[...]
X-squared = 2.8443, df = 2, p-value = 0.2412

# Eingabe, Forts.: Cramérs V
cramers.V(zushang)

# Ausgabe
Cramers.V Lower.Limit Upper.Limit
 0.10439      0.00000     0.21269

# Eingabe, Forts.: Untere 95%-Konfidenzschranke
cramers.V(zushang, alternative = 'greater')

# Ausgabe
Cramers.V Lower.Limit
 0.10439      0.00000
```

Kommentare

- *Schätzwert*: Wir schätzen, dass der wahre Wert θ von Cramérs V 0.104 beträgt.
- *95%-Konfidenzintervall*: Mit einer Sicherheit von rund 95% liegt θ zwischen 0 und 0.213.
- *Untere 95%-Konfidenzschranke*: Die untere 95%-Konfidenzschranke für θ beträgt null. Somit können wir nicht behaupten, dass ein echter Zusammenhang zwischen ‘Rauchen’ und ‘Geschlecht’ vorliegt.
- *ChiQuadrat-Unabhängigkeitstest auf 5%-Niveau*: Zum gleichen Schluss wie mit der unteren Konfidenzschranke kommen wir mit dem ChiQuadrat-Unabhängigkeitstest: Der p -Wert 0.24 ist nicht kleiner als 0.05, somit gibt es keinen Grund, an der Nullhypothese (kein Zusammenhang) zu zweifeln. Der Testentscheid kann auch mithilfe der ChiQuadrat-Teststatistik (2.8443) getroffen werden: Ihr Wert ist nicht grösser als die kritische Schranke 5.99 (das 95%-Quantil der ChiQuadrat-Verteilung mit $(L - 1)(M - 1) = 2$ Freiheitsgraden).

¹Sind viele idealisierte Häufigkeiten klein, so ist die Approximation durch die ChiQuadrat-Verteilung nicht sehr gut. Dies kann insbesondere bei kleinen Stichproben und/oder vielen Kategorien passieren. Eine häufig verwendete Faustregel besagt, dass man der Approximation nicht mehr trauen sollte, falls mehr als ein Fünftel der idealisierten Häufigkeiten kleiner als fünf sind.

²In R wird der χ^2 -Unabhängigkeitstest mit der Funktion `chisq.test` berechnet, Konfidenzintervalle für Cramérs V ermitteln wir mit der eigenen Funktion `cramers.V`.

4.5 Weitere Beispiele

Beispiel 4.8 (Absenzen und Schule). In Beispiel 2.15 haben wir einen Datensatz über 316 Junior High School SchülerInnen an zwei städtischen Schulen in den USA beschrieben.

Wir möchten nun den Zusammenhang zwischen der Anzahl Absenzen (Variable ‘daysabs’ mit Kategorien 0 – 1, 2 – 5, > 5) und Schule (Variable ‘school’ mit den Kategorien 0 und 1) untersuchen. Für schliessende Statistik nehmen wir an, dass die SchülerInnen eine Zufallsstichprobe aus den beiden grossen Schulen 0 und 1 darstellen.

Die erweiterte Häufigkeitstabelle sieht folgendermassen aus (Mosaikdiagramm in Abbildung 4.2):

R Code

```
# Eingabe
a.H <- table(highschool$school, highschool$daysabs)
addmargins(a.H)
plot(a.H)

# Ausgabe
  [0,1] (1,5] (5,50] Sum
0      31     50     78 159
1      77     49     31 157
Sum   108    99    109 316
```

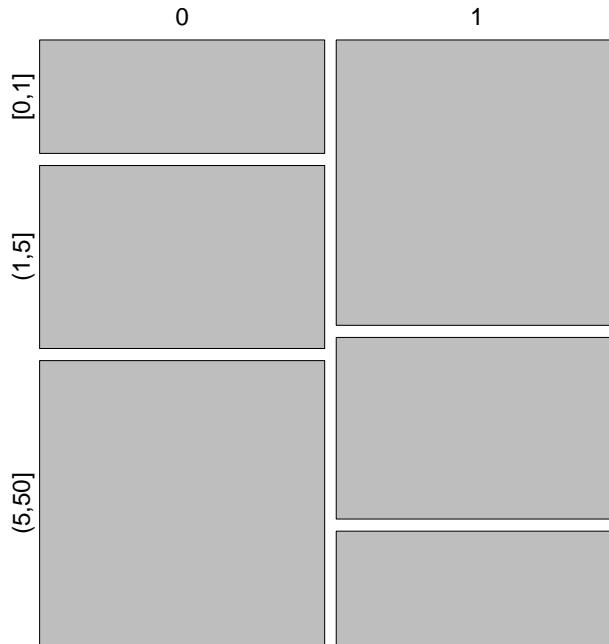


Abbildung 4.2: Mosaikdiagramm der Häufigkeitstabelle von ‘Absenzen’ und ‘Schule’ in Beispiel 4.8.

Anhand der Zeilennormierung erkennen wir, dass sich das Absenzverhalten klar zwischen den beiden Schulen unterscheidet: In Schule 0 ist der Anteil der SchülerInnen mit 0 oder 1 Absenzen deutlich geringer und der Anteil der SchülerInnen mit mehr als 5 Absenzen deutlich höher als in Schule 1.

	0 – 1 Absenzen	2 – 5 Absenzen	> 5 Absenzen	Summe
Schule 0	$31/159 \approx 0.195$	$50/159 \approx 0.314$	$78/159 \approx 0.491$	1
Schule 1	$77/157 \approx 0.490$	$49/157 \approx 0.312$	$31/157 \approx 0.197$	1

Nun berechnen wir die Pearson-Residuen, die ChiQuadrat-Teststatistik sowie Cramérs V. Anhand der univariaten Verteilungen der beiden Merkmale erwarten wir unter Unabhängigkeit folgende (idealisierten) Häufigkeiten:

	0 – 1 Absenzen	2 – 5 Absenzen	> 5 Absenzen
Schule 0	$108 \cdot 159/316 \approx 54.342$	$99 \cdot 159/316 \approx 49.813$	$109 \cdot 159/316 \approx 54.845$
Schule 1	$108 \cdot 157/316 \approx 53.658$	$99 \cdot 157/316 \approx 49.187$	$109 \cdot 157/316 \approx 54.155$

Daraus folgen die Pearson-Residuen:

	0 – 1 Absenzen	2 – 5 Absenzen	> 5 Absenzen
Schule 0	$\frac{31 - 54.342}{\sqrt{54.342}} \approx -3.166$	$\frac{50 - 49.813}{\sqrt{49.813}} \approx 0.026$	$\frac{78 - 54.845}{\sqrt{54.845}} \approx 3.127$
Schule 1	$\frac{77 - 53.658}{\sqrt{53.658}} \approx 3.187$	$\frac{49 - 49.187}{\sqrt{49.187}} \approx -0.027$	$\frac{31 - 54.155}{\sqrt{54.155}} \approx -3.146$

Deren quadrierte Summe liefert den Wert der ChiQuadrat-Teststatistik

$$\chi^2 = (-3.166)^2 + \dots + (-3.146)^2 \approx 39.86$$

und daraus folgt schliesslich Cramers V

$$V = \sqrt{\frac{39.86}{316 \cdot (2-1)}} = \sqrt{\frac{39.86}{316}} \approx 0.355.$$

Kommentare

- Mit einem Cramérs V von 0.355 liegt ein starker Zusammenhang zwischen ‘Schule’ und ‘Absenzverhalten’ vor.
- Anhand der Pearson-Residuen ist klar erkennbar, dass in Schule 0 deutlich zu wenige SchülerInnen 0 – 1 Absenzen haben und deutlich zu viele über fünf Absenzen, als unter Unabhängigkeit zu erwarten wäre. Bei Schule 1 ergibt sich gerade das umgekehrte Bild.

Dies bestätigt und ergänzt die Folgerungen aus der Zeilennormierung und dem Mosaikdiagramm. Die Berechnungen zur schliessenden Statistik überlassen wir der Software¹.

R Code

```
# Eingabe, Forts.: Zeilennormierung
prop.table(a.H, margin = 1)

# Ausgabe
[0,1]  (1,5]  (5,50]
0    0.19497 0.31447 0.49057
1    0.49045 0.31210 0.19745

# Eingabe, Forts.: ChiQuadrat-Unabhängigkeitstest
chisq.result <- chisq.test(a.H)
chisq.result

# Ausgabe
[...]
X-squared = 39.858, df = 2, p-value = 2.213e-09
```

¹Durch einigen zusätzlichen Output verifizieren wir zudem obige Berechnungen.

```
# Eingabe: Untere Konfidenzschranke für Cramérs V
cramers.V(chisq.result, alternative = 'greater')

# Ausgabe
Cramers.V Lower.Limit
0.35515      0.25738

# Eingabe, Forts.: Erwartete Häufigkeiten
chisq.result$expected

# Ausgabe
[0,1]  (1,5] (5,50]
0     54.342 49.813 54.845
1     53.658 49.187 54.155

# Eingabe, Forts.: Pearson-Residuen
residuals(chisq.result)

# Ausgabe
[0,1]      (1,5]      (5,50]
0 -3.166408  0.026454  3.126639
1  3.186513 -0.026622 -3.146491
```

Kommentare

- *Schätzwert:* Wir schätzen, dass der wahre Wert θ von Cramérs V 0.36 beträgt.
- *Untere 95%-Konfidenzschranke:* Mit einer Sicherheit von rund 95% beträgt θ mindestens 0.257. So mit können wir mit etwa 95% Sicherheit behaupten, dass ein recht deutlicher echter Zusammenhang zwischen ‘Schule’ und ‘Absenzverhalten’ vorliegt.
- *ChiQuadrat-Unabhängigkeitstest auf 5%-Niveau:* Der p -Wert des ChiQuadrat-Unabhängigkeitstests ist fast null, also kleiner als 0.05. Somit können wir die Nullhypothese von keinem Zusammenhang auf dem 5%-Niveau zugunsten der Arbeitshypothese (“es gibt einen echten Zusammenhang”) verwerfen. Anstatt mit dem p -Wert oder der unteren Konfidenzschranke für θ lässt sich der Testentscheid auch mit der Teststatistik finden: Der Wert 39.858 ist grösser als die kritische Schranke 5.99 (gleiches Signifikanzniveau und gleiche Anzahl Kategorien wie in Beispiel 4.7), also können wir die Nullhypothese auf dem 5%-Niveau verwerfen. ▲

Beispiel 4.9 (Geschlecht und Beförderung). Im Rahmen einer Fortbildungsveranstaltung nahmen 48 angehende ManagerInnen an einem Experiment teil, ohne dies zu wissen. Jede(r) von ihnen erhielt eine (fiktive) Personalakte und sollte entscheiden, ob die betreffende Person befördert wird oder nicht. Die 48 Personalakten waren identisch bis auf den Namen der Person und wurden rein zufällig verteilt. In 24 Fällen handelte es sich um die Akte eines Mannes, in 24 Fällen um die einer Frau.

Dieses Experiment lieferte einen Datensatz mit $n = 48$ Beobachtungen (ManagerInnen) und den Variablen X (Geschlecht m/w) sowie Y (Beförderung ja/nein). Die (erweiterte) Häufigkeitstabelle fasst die Daten zusammen:

	Beförderung	keine Beförderung	Summe
Kandidat	21	3	24
Kandidatin	14	10	24
Summe	35	13	48

Ein klarer Zusammenhang zwischen ‘Geschlecht’ und ‘Beförderung’ ist augenscheinlich: Während von den 24 Männern nur drei nicht befördert wurden, sind es bei den ebenfalls 24 Frauen zehn, also ganze

$10/3 \approx 3.3$ mal so viele. (Aufgrund des Aufbaus des Experiments sind die Zeilensummen gleich gross. Deshalb lassen sich die auf ‘Geschlecht’ bedingten Verteilungen von ‘Beförderung’ ohne Zeilennormierung direkt vergleichen). Zum gleichen Schluss kommen wir anhand der Zeilennormierung:

	Beförderung	keine Beförderung	Summe
Kandidat	$21/24 = 0.875$	$3/24 = 0.125$	1
Kandidatin	$14/24 \approx 0.583$	$10/24 \approx 0.417$	1

Auch hier ist klar zu erkennen, dass der Anteil der nichtbeförderten Kandidatinnen deutlich grösser ist als jener der nichtbeförderten Kandidaten ($0.417/0.125 \approx 3.3$ mal so gross, siehe oben).

Das Mosaikdiagramm bestätigt diesen Eindruck (Abbildung 4.3).

R Code

```
# Eingabe
manager <- cbind(Ja = c(M = 21, W = 14), Nein = c(3, 10))
manager <- as.table(manager)
plot(manager)
```

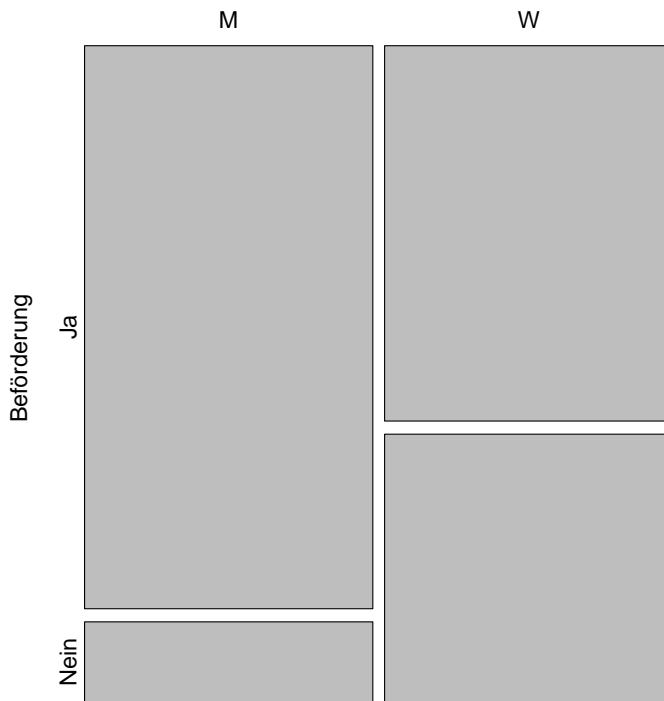


Abbildung 4.3: Mosaikdiagramm von ‘Beförderung’ und ‘Geschlecht’ in Beispiel 4.9.

Nun ergänzen wir diese Feststellungen mit Pearson-Residuen, Chiquadrat-Teststatistik und Cramérs V . Welche Rückschlüsse auf die Population aller Manager und Managerinnen erlauben die Daten?

Anhand der univariaten Verteilungen würde man unter Unabhängigkeit folgende (idealisierten) Häufigkeiten erwarten:

	Beförderung	keine Beförderung
Kandidat	$35 \cdot 24/48 = 17.5$	$13 \cdot 24/48 = 6.5$
Kandidatin	$35 \cdot 24/48 = 17.5$	$13 \cdot 24/48 = 6.5$

Daraus folgen die Pearson-Residuen:

	Beförderung	keine Beförderung
Kandidat	$(21 - 17.5)/\sqrt{17.5} \approx 0.8367$	$(3 - 6.5)/\sqrt{6.5} \approx -1.3728$
Kandidatin	$(14 - 17.5)/\sqrt{17.5} \approx -0.8367$	$(10 - 6.5)/\sqrt{6.5} \approx 1.3728$

Die Summe der quadrierten Pearson-Residuen beträgt

$$\chi^2 = 0.8367^2 + (-0.8367)^2 + (-1.3728)^2 + 1.3728^2 \approx 5.17$$

und daraus finden wir Cramers V :

$$V = \sqrt{\frac{5.17}{48 \cdot (2-1)}} = \sqrt{\frac{5.17}{48}} \approx 0.328.$$

Kommentare: Beförderte Kandidatinnen sind zu selten, nichtbeförderte Kandidatinnen entsprechend zu häufig, als man unter Unabhängigkeit von ‘Beförderung’ und ‘Geschlecht’ erwarten würde. Bei den Kandidaten ergibt sich das umgekehrte Bild. Mit einem Cramérs V von 0.328 liegt ein deutlicher Zusammenhang vor. Dies bestätigt die Folgerungen aus der Zeilennormierung und dem Mosaikdiagramm.

Die Berechnungen zur schliessenden Statistik überlassen wir der Software¹:

R Code

```
# Eingabe, Forts.
zushang <- chisq.test(manager, correct = FALSE)
zushang

# Ausgabe
X-squared = 5.1692, df = 1, p-value = 0.02299

# Eingabe, Forts.
cramers.V(zushang, alternative = 'greater')

# Ausgabe
Cramers.V Lower.Limit
0.32817      0.08795
```

Kommentare

- *Schätzwert:* Ein Schätzwert für das tatsächliche Cramérs V ist gegeben durch 0.328.
- *Untere 95%-Konfidenzschranke:* Mit einer Sicherheit von rund 95% liegt der wahre Wert über 0.087 (mindestens ein schwacher Zusammenhang).
- *Chiquadrat-Unabhängigkeitstest auf 5%-Niveau:* Der p -Wert des Chiquadrat-Unabhängigkeitstests beträgt 0.023. Somit können wir die Nullhypothese von keinem Zusammenhang auf dem 5%-Niveau zugunsten der Arbeitshypothese (“es gibt einen echten Zusammenhang”) verwerfen und mit einer Sicherheit von rund 95% behaupten, dass ein echter Zusammenhang vorliegt. Zum gleichen Testentscheid gelangen wir via kritischer Schranke, da der Wert 5.17 der Chiquadrat-Teststatistik grösser als das 95%-Quantil² 3.84 der Chiquadrat-Verteilung mit $(L-1)(M-1) = 1$ Freiheitsgraden ist. ▲

¹R verwendet bei (2×2) -Tabellen eine leicht andere Definition der Chiquadrat-Teststatistik, mit der in der Regel präzisere p -Werte entstehen. Um die Berechnung der Teststatistik nachvollziehen zu können, unterdrücken wir hier diese sogenannte *Stetigkeitskorrektur* mit der Option `correct = FALSE` jedoch.

²In R mit `qchisq(0.95, 1)`.

4.6 Vierfeldertafeln und Odds Ratios

In Beispiel 4.9 haben wir eine Häufigkeitstabelle von zwei binären Merkmalen betrachtet. Solche Häufigkeitstabellen werden *Vierfeldertafeln* genannt:

$H_{1,1}$	$H_{1,2}$
$H_{2,1}$	$H_{2,2}$

Obwohl die bisher gezeigten Verfahren auch für Vierfeldertafeln eingesetzt werden können, ist es dann einfacher und anschaulicher, die Zahlenverhältnisse innerhalb der Zeilen (oder Spalten) zu berechnen und miteinander zu vergleichen. Die entsprechende Kenngrösse ist die *Odds Ratio (Chancenquotient)*

$$OR := \frac{H_{1,1}/H_{1,2}}{H_{2,1}/H_{2,2}} = \frac{H_{1,1}/H_{2,1}}{H_{1,2}/H_{2,2}}.$$

Die Namen “Odds Ratio” und “Chancenquotient” versteht man am besten anhand eines Beispiels.

Beispiel 4.10 (Geschlecht und Beförderung, Fortsetzung). Interpretieren wir die Vierfeldertafel in Beispiel 4.9 zeilenweise: Die Chancen auf Beförderung sind für die 24 (fiktiven) Männer gleich $21/3 = 7$; für die 24 Frauen gleich $14/10 = 1.4$. Die Chancen auf Beförderung sind bei den Männern also $7/1.4 = 5$ mal so gross wie bei den Frauen.

Nun die spaltenweise Betrachtung: Die Chancen, unter den 35 Beförderten einen Mann anzutreffen, sind gleich $21/14 = 1.5$; bei den 13 Nichtbeförderten sind diese Chancen gleich $3/10 = 0.3$. Der Quotient dieser Chancen beträgt $1.5/0.3 = 5$. Die Chancen, bei den Beförderten einen Mann anzutreffen, sind also 5 mal so gross wie bei den Nichtbeförderten.

Kurzum:

$$OR = \frac{21/3}{14/10} = 5.$$

Interessant wäre hier eine Unterscheidung nach Geschlecht der ManagerInnen: Bevorzugen Frauen Frauen, Männer Männer? Oder werden Frauen von Frauen gleich stark benachteiligt wie von Männern? ▲

Hinweise zur Odds Ratio

- Die Odds Ratio beschreibt den Zusammenhang auf einfache und anschauliche Weise mit einer einzigen Zahl. Sie dient damit u. a. auch als Zusammenhangsmass.
- Die Odds Ratio liegt immer zwischen 0 und unendlich. Ein Wert von 1 entspricht gar keinem Zusammenhang, ein Wert von 0 oder unendlich hingegen einem “perfekten” Zusammenhang. Generell beschreibt eine Odds Ratio von $\theta > 0$ einen gleich starken Zusammenhang wie eine von $1/\theta$, jedoch in umgekehrter Richtung (Vierfeldertafel mit vertauschten Zeilen oder Spalten). Für Beispiel 4.10 könnten wir auch sagen: “Die Chancen auf Beförderung sind für die Frauen $1.4/7 = 0.2 = 1/5$ mal so gross wie für die Männer.”
- Chancen sind nicht zu verwechseln mit Wahrscheinlichkeiten: Die Wahrscheinlichkeit, mit einem fairen Würfel eine Sechs zu würfeln, beträgt $1/6 \approx 0.167$, die entsprechenden Chancen hingegen “Eins zu fünf” (0.2).
- Je nach Situation spricht man eher von *Risiko* statt von Chancen.

Die Charakterisierung des Zusammenhangs zwischen zwei binären Variablen anhand der Odds Ratio ist besonders anschaulich. Deshalb reduziert man grössere Häufigkeitstabellen manchmal durch Zusammenfassen oder Streichen bestimmter Kategorien, so dass sie in eine Vierfeldertafel passen¹.

Beispiel 4.11 (Geschlecht und Rauchen, Fortsetzung). Wir möchten den Zusammenhang zwischen ‘Rauchen’ und ‘Geschlecht’ mit Hilfe der Odds Ratio untersuchen.

Zu diesem Zweck fassen wir die GelegenheitsraucherInnen und die regelmässigen RaucherInnen zur neuen Kategorie “RaucherInnen” zusammen, so dass folgende Vierfeldertafel entsteht:

	NichtraucherIn	RaucherIn
M	100	45
W	71	45

Die Odds Ratio beträgt

$$\frac{100/71}{45/45} \approx 1.4.$$

Zeilenweise betrachtet können wir damit folgendes sagen: Die Chancen, bei den Männern einen Nichtraucher anzutreffen, sind 1.4 mal so gross wie bei den Frauen.

Was würden wir für die äquivalente Vierfeldertafel mit vertauschten Spalten erhalten?

	RaucherIn	NichtraucherIn
M	45	100
W	45	71

Die Odds Ratio entspricht dem Kehrwert von 1.4, nämlich

$$\frac{45/45}{100/71} = 0.71 \approx 1/1.4.$$

Zeilenweise betrachtet würden wir zu folgendem gleichwertigem Schluss gelangen: Die Chancen, bei den Männern einen Raucher anzutreffen, sind 0.71 mal so gross wie bei den Frauen. ▲

In einer Zufallsstichprobe schätzt die Odds Ratio den tatsächlichen Wert θ in der Population. Da die exakte Verteilung der (Stichproben-)Odds Ratio für ein hypothetisches θ bei gegebenen Zeilen- und Spaltensummen bekannt ist, können für θ exakte Konfidenzintervalle angegeben werden und mit *Fishers exaktem Test* ein- und zweiseitige Hypothesen für θ geprüft werden. Mit den Hypothesen $H_1 : \theta \neq 1$ versus $H_0 : \theta = 1$ lässt sich beispielsweise prüfen, ob ein echter Zusammenhang zwischen den beiden binären Merkmalen vorliegt. Via Konfidenzintervallansatz folgen wie üblich identische Testentscheide.

Beispiel 4.12 (Geschlecht und Beförderung, Fortsetzung). Wir illustrieren diese Verfahren für die Situation in Beispiel 4.10. Dabei betrachten wir die Teilnehmenden des Experiments als Zufallsstichprobe aus der Grundgesamtheit aller ManagerInnen.

Was können wir über die Odds Ratio θ in der Grundgesamtheit und damit über den echten Zusammenhang zwischen ‘Geschlecht’ und ‘Beförderung’ sagen? ²

R Code

```
# Eingabe, Forts.
fisher.test(manager)
```

¹Auch numerische Merkmale lassen sich unter Informationsverlust in binäre umwandeln (bspw. “Wert grösser oder kleiner gleich Median?”). Damit lässt sich im Prinzip jede bivariate Situation auf eine Vierfeldertafel zurückführen und mithilfe der Odds Ratio analysieren.

²Odds Ratio, Konfidenzintervall und Fishers exakter Test werden in R mit `fisher.test` berechnet.

```
# Ausgabe
[...]
p-value = 0.04899
95 percent confidence interval: 1.0056 32.2058
sample estimates: odds ratio 4.8312
```

Kommentare

- *Schätzwert:* Der Stichprobenwert 5 aus Beispiel 4.10 dient als Schätzwert für θ .¹
- *95%-Konfidenzintervall:* Mit einer Sicherheit von 95% liegt θ zwischen 1.01 und 32.21. Aufgrund der kleinen Stichprobe liegt also ein sehr unpräziser Schätzwert vor.
- *Test auf Zusammenhang auf 5%-Niveau:* Fishers exakter Test liefert für den Test auf Zusammenhang ($H_1 : \theta \neq 1$ versus $H_0 : \theta = 1$) einen p -Wert von 0.04899, der insbesondere kleiner als das Signifikanzniveau von 5% ist. Damit verwerfen wir die Nullhypothese und behaupten mit einer Sicherheit von 95%, dass θ ungleich 1 ist bzw. dass es einen echten Zusammenhang zwischen ‘Geschlecht’ und ‘Beförderung’ gibt. Zum gleichen Testentscheid gelangen wir mit dem Konfidenzintervallansatz: Der Wert 1 liegt ausserhalb des 95%-Konfidenzintervalls für θ .
- *Vergleich mit dem Test von Beispiel 4.9:* Die p -Werte unterscheiden sich, da der Chiquadrat-Unabhängigkeitstest nur ein approximatives Verfahren ist. Hier folgt jedoch der gleiche Testentscheid. ▲

Beispiel 4.13 (Geschlecht und Beförderung, Fortsetzung). Nun betrachten wir eine gerichtete Fragestellung zu diesem Beispiel, die mit dem Chiquadrat-Unabhängigkeitstest nicht möglich wäre: Werden Männer hinsichtlich Beförderung systematisch bevorzugt? Ist also die Odds Ratio in der Population (θ) grösser als eins?

R Code

```
# Eingabe, Forts.
fisher.test(manager, alternative = 'greater')

# Ausgabe
[...]
p-value = 0.02450
95 percent confidence interval: 1.230224      Inf
sample estimates: odds ratio 4.83119
```

Kommentare

- *Untere 95%-Konfidenzschranke:* Die Odds Ratio θ in der Population aller ManagerInnen liegt mit einer Sicherheit von 95% über 1.23. Die Chancen einer Beförderung sind bei Männern also mit einer Sicherheit von 95% mindestens 1.23 mal so hoch wie bei Frauen (Männer merklich bevorzugt).
- *Einseitiger Test auf 5%-Niveau:* Fishers exakter Test der einseitigen Hypothesen $H_1 : \theta > 1$ versus $H_0 : \theta \leq 1$ ergibt einen p -Wert von 0.0245. Wir verwerfen damit die Nullhypothese und behaupten mit einer Sicherheit von 95%, dass θ grösser als eins ist bzw. dass Männer systematisch bevorzugt werden. Zum gleichen Testentscheid führt der Konfidenzintervallansatz. ▲

Beispiel 4.14 (Geschlecht und Rauchen, Fortsetzung). Was können wir in Beispiel 4.11 über die Odds Ratio θ in der Population bzw. den echten Zusammenhang zwischen ‘Geschlecht’ und dem binären Merkmal Y (NichtraucherIn = 0, RaucherIn = 1) sagen?

¹R verwendet aus theoretischen Gründen einen leicht anderen Schätzer.

R Code

```
# Eingabe
Y <- as.numeric(wiso$Rauchen != 0)
a.H <- table(wiso$Geschlecht, Y)
a.H

# Ausgabe
Y
  0   1
M 100 45
W  71 45

# Eingabe, Forts.
fisher.test(a.H)

# Ausgabe
[...]
p-value = 0.1936
95 percent confidence interval: 0.8158 2.4293
sample estimates: odds ratio 1.407
```

Kommentare

- *Schätzwert*: Wir schätzen den Wert von θ auf 1.41.
- *95%-Konfidenzintervall*: Mit einer Sicherheit von 95% behaupten wir, dass θ zwischen 0.81 und 2.43 liegt. Der Wert 1 (kein Zusammenhang) liegt im möglichen Bereich, somit können wir die Nullhypothese von keinem Zusammenhang nicht verwerfen.
- *Fishers exakter Test auf dem 5%-Niveau*: Der p -Wert von Fishers exaktem Test beträgt 0.19, liegt also nicht unterhalb des 5%-Niveaus. Deshalb haben wir keinen Grund, an der Nullhypothese zu zweifeln. Der p -Wert des Chi-quadrat-Unabhängigkeitstests für die nicht reduzierte Häufigkeitstabelle in Beispiel 4.7 war mit 0.24 ähnlich. ▲

Beispiel 4.15 (Geschlecht und Rauchen, Fortsetzung). Nun betrachten wir auf dem 5%-Niveau die gerichtete Fragestellung, ob die Chancen, dass ein Mann nicht raucht, tatsächlich grösser sind als die Chancen, dass eine Frau nicht raucht (das Chancenverhältnis in der Population bezeichnen wir wiederum mit θ) bzw. ob Männer eher zum Nichtrauchen tendieren als Frauen.

R Code

```
# Eingabe, Forts.
fisher.test(a.H, alternative = 'greater')

# Ausgabe
[...]
p-value = 0.1192
95 percent confidence interval: 0.8857      Inf
```

Kommentare

- *Untere 95%-Konfidenzschanke*: Mit einer Sicherheit von 95% beträgt θ mindestens 0.885; dieser Wert ist nicht grösser als 1. Wir können somit insbesondere nicht behaupten, dass Männer tatsächlich eher zum Nichtrauchen neigen als Frauen.
- *Einseitiger Fishers exakter Test auf dem 5%-Niveau*: Fishers exakter Test der einseitigen Arbeitshypothese $\theta > 1$ versus die Nullhypothese $\theta \leq 1$ liefert einen p -Wert von 0.119. Dieser Wert ist nicht kleiner als 0.05. Somit gibt es keinen Grund, die Nullhypothese anzuzweifeln. ▲

4.7 Exakter Chiquadrat-Unabhängigkeitstest

Eine Möglichkeit, exakte p -Werte für den χ^2 -Unabhängigkeitstest zu erhalten, basiert auf einer Verallgemeinerung von Fishers exaktem Test auf allgemeine Häufigkeitstabellen. Die Verbindung zur Odds Ratio und die Möglichkeit, gerichtete Fragestellungen zu prüfen, gibt es jedoch nur im Fall einer Vierfeldertafel.

Beispiel 4.16 (Geschlecht und Rauchen, Fortsetzung). Mit der Verallgemeinerung von Fishers exaktem Test folgt der gleiche Testentscheid wie in Beispiel 4.7 mit dem approximativen Test (p -Wert 0.24):

R Code	fisher.test(wiso\$Geschlecht, wiso\$Rauchen) # Ergibt p-value = 0.2451
--------	--



Beispiel 4.17 (Absenzen und Schule, Fortsetzung). Dieses Verfahren führt auch in Beispiel 4.8 zum gleichen Schluss.

R Code	fisher.test(highschool\$school, highschool\$daysabs) # Ergibt p-value = 1.388e-09
--------	---



4.8 Zusammenfassung

- Ausgangslage zum Studium des Zusammenhangs zwischen zwei kategorialen Merkmalen ist die Häufigkeitstabelle. Sie gibt für jede Kombination von Ausprägungen ihre Häufigkeit an und stellt damit die gemeinsame Verteilung der beiden Merkmale dar.
- Wir haben den Zusammenhang verdeutlicht, indem wir die Tabelleneinträge pro Zeile bzw. Spalte in relative Häufigkeiten umgerechnet und diese zwischen den Zeilen bzw. Spalten verglichen haben. Unterschiede weisen auf einen Zusammenhang zwischen den Merkmalen hin. Solche Zeilen- bzw. Spalten-normierungen können durch ein Mosaikdiagramm visualisiert werden. Dabei haben wir über bedingte Wahrscheinlichkeiten gesprochen und den Satz von Bayes kennengelernt.
- Die Stärke des Zusammenhangs wird mithilfe der Chiquadrat-Teststatistik bzw. deren standardisierter Version, dem Cramérs V , quantifiziert. Die Berechnung basiert darauf, wie weit die Tabelleneinträge von den unter Unabhängigkeit erwarteten (idealisierten) Werten insgesamt abweichen bzw. wie gross die Pearson-Residuen sind.
- Liegt eine Zufallsstichprobe vor, dient Cramérs V als Schätzwert für den Populationswert und es können approximative Konfidenzintervalle dafür angegeben werden. Der damit verbundene Chiquadrat-Unabhängigkeitstest prüft die Nullhypothese, dass es keinen echten Zusammenhang gibt.
- Sind beide Merkmale binär, so liegt eine Vierfeldertafel vor, die sich einfacher mit der Odds Ratio analysieren lässt. Diese charakterisiert den Zusammenhang durch eine einzige anschauliche Zahl. Grössere Häufigkeitstabellen manchmal auf Vierfeldertafeln reduziert, indem Kategorien zusammengelegt oder gestrichen werden.
- In einer Zufallsstichprobe dient die Odds Ratio als naheliegender Schätzwert für die Odds Ratio in der Population und es können exakte Konfidenzintervalle dafür ermittelt werden. Damit lassen sich ein- oder zweiseitige Hypothesen über die echte Odds Ratio beziehungsweise über den Zusammenhang zwischen den beiden Merkmalen prüfen (Fishers exakter Test).
- Eine Erweiterung von Fishers exaktem Test ermöglicht exakte Chiquadrat-Unabhängigkeitstests für allgemeine Häufigkeitstabellen.

Kapitel 5

Ein kategorielles, ein numerisches Merkmal

Hier untersuchen wir den Zusammenhang zwischen einer numerischen Variable Y und einer kategorialen Variable X mit den Kategorien x_1, x_2, \dots, x_L . Man unterteilt den Datensatz anhand der X -Werte in Teilstichproben und schaut, ob die Y -Werte in diesen Stichproben ähnlich oder sehr unterschiedlich verteilt sind. Unterschiede charakterisieren den Zusammenhang zwischen den X - und Y -Werten.

5.1 Stratifizierte Beschreibung der Stichprobe

Pro Teilstichprobe wird die Verteilung der Y -Werte mit grafischen und/oder quantitativen univariaten Verfahren beschrieben. Man beschreibt also die auf X bedingten Verteilungen von Y bzw. die Verteilung der Y -Werte in Abhängigkeit von X bzw. macht eine nach X stratifizierte Analyse von Y . Danach vergleicht man die Verteilungen zwischen den Teilstichproben hinsichtlich ihrer mittleren Lage, Streuung und evtl. auch Form, um einen allfälligen Zusammenhang zwischen den X - und Y -Werten zu charakterisieren.

5.1.1 Grafische Darstellung

Grafische Vergleiche basieren in der Regel auf ECDFs, Stripcharts oder den sogenannten Boxplots. Im Prinzip können auch Histogramme¹ (nach Konvention 2) verglichen werden.

Empirische Verteilungsfunktionen

Für jede Teilstichprobe zeichnet man die ECDF der entsprechenden Y -Werte in das gleiche Koordinatensystem. Sofern es nur wenig Teilstichproben gibt, also falls X nur wenige Kategorien aufweist, können die Kurven sehr gut verglichen werden. Wie in der univariaten Statistik können pro Teilstichprobe Quantile und Anteile abgelesen werden.

Beispiel 5.1 (Körpergewicht). Betrachten wir nun den Datensatz von Beispiel 1.1. Was können wir mit solchen ECDFs² je über den Zusammenhang zwischen ‘Körpergewicht’ und den drei Merkmalen ‘Geschlecht’, ‘Rauchen’ und ‘Geburtsmonat’ (Abbildung 5.1) sagen?

¹Übersichtlicher als Histogramme sind hier die eng damit verwandten *Häufigkeitspolygone*: Sie entstehen, indem man die oberen Mittelpunkte der Histogrammbalken durch Linien verbindet. Solche Häufigkeitspolygone werden manchmal mit speziellen Verfahren geglättet, so dass sie wie eine typische Dichtefunktion aussehen. Dann spricht man von *Kerndichteschätzern*.

²In R werden stratifizierte ECDFs mit der Funktion `Ecdf` und der Option `group` gezeichnet.

R Code

```
# Eingabe (benötigt library(Hmisc))
Ecdf(wiso$Kgewicht, group = wiso$Geschlecht)
Ecdf(wiso$Kgewicht, group = wiso$Rauchen)
Ecdf(wiso$Kgewicht, group = wiso$GebMonat)
```

Kommentare

- ‘*Körpergewicht*’ in Abhängigkeit von ‘*Geschlecht*’: Die ECDFs und damit die auf ‘*Geschlecht*’ bedingten Verteilungen von ‘*Körpergewicht*’ unterscheiden sich deutlich (auch von der ECDF in Abbildung 2.8 aller Werte von ‘*Körpergewicht*’, die zwischen den beiden Kurven verlaufen würde). Es ist demnach ein deutlicher Zusammenhang (in Form einer Verschiebung) zwischen den beiden Merkmalen ersichtlich. Man kann beispielsweise ablesen, dass der Median bei den Männern etwa 10 kg grösser ist als bei den Frauen oder dass der Anteil der Männer, die höchstens 60 kg wiegen, nur einige Prozente beträgt, während der entsprechende Anteil bei den Frauen rund 70% ausmacht.
- ‘*Körpergewicht*’ in Abhängigkeit von ‘*Rauchen*’: Die Kurven sind sehr ähnlich (auch wie die ECDF in Abbildung 2.8), somit ist kein Zusammenhang zwischen ‘*Körpergewicht*’ und ‘*Rauchen*’ erkennbar.
- ‘*Körpergewicht*’ in Abhängigkeit von ‘*Geburtsmonat*’: Die Darstellung ist aufgrund der vielen Gruppen zu unübersichtlich, um verlässliche Aussagen zu machen.

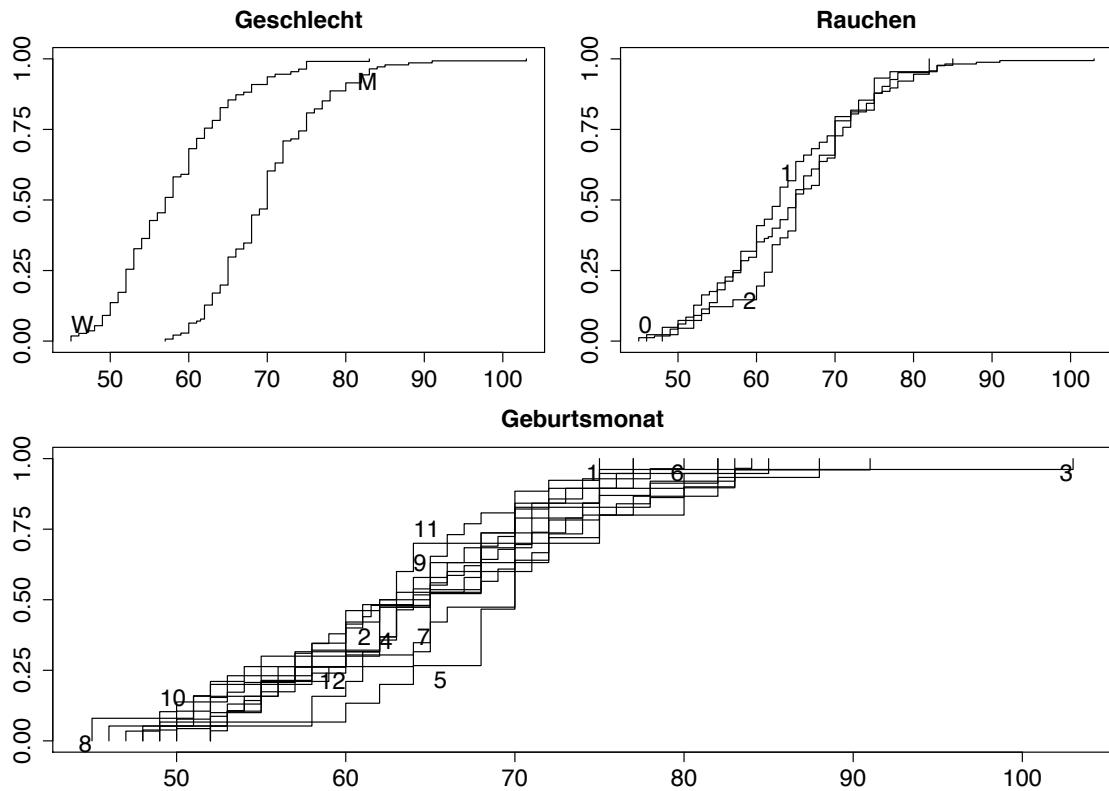


Abbildung 5.1: ECDFs von ‘*Körpergewicht*’ je in Abhängigkeit von ‘*Geschlecht*’, ‘*Rauchen*’ und ‘*Geburtsmonat*’ (Beispiel 5.1).

Manchmal sollen alle/mehrere numerische Variablen eines Datensatzes in Abhängigkeit eines kategorialen Merkmals beschrieben werden, siehe Abbildung 5.2. Nur bei ‘*Körpergrösse*’ und ‘*Körpergewicht*’ sind

deutliche Unterschiede zwischen Männern und Frauen (und damit ein Zusammenhang mit ‘Geschlecht’) zu erkennen.

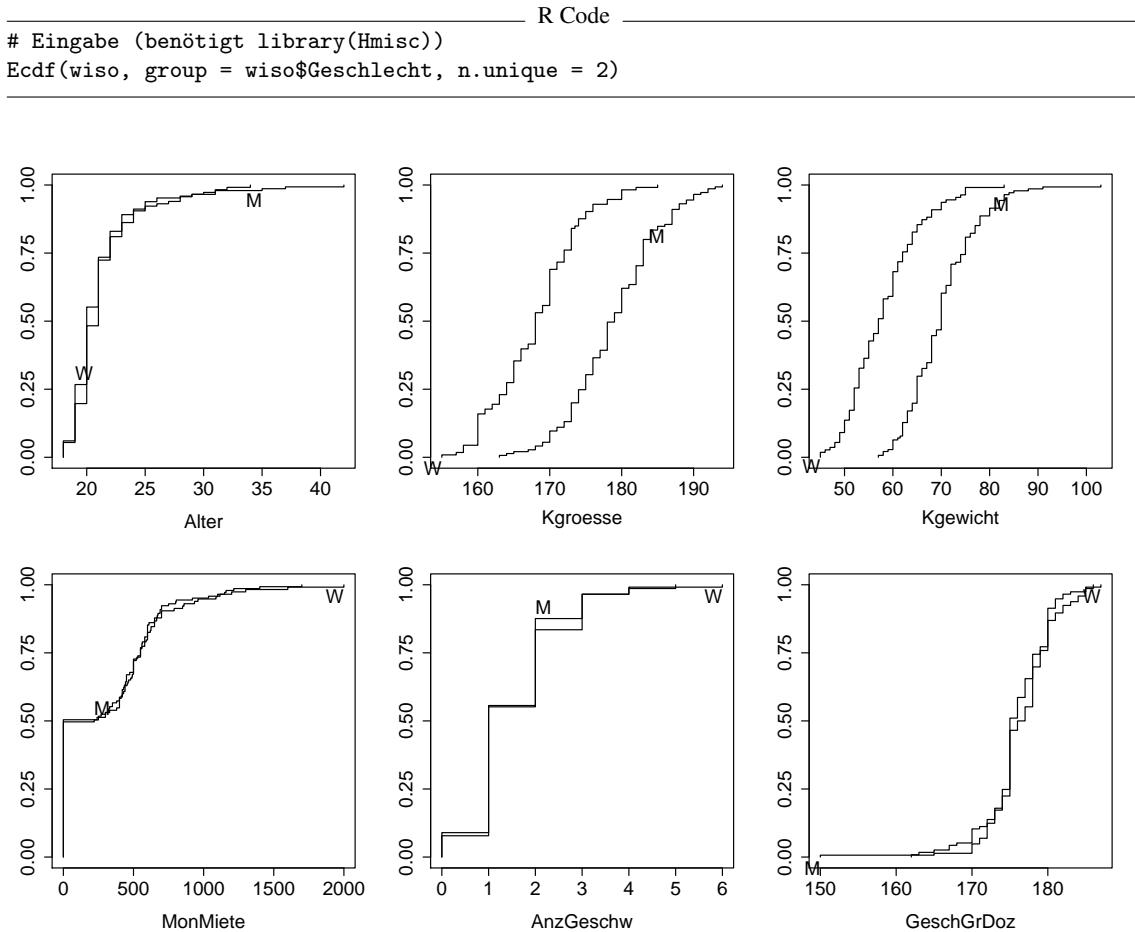


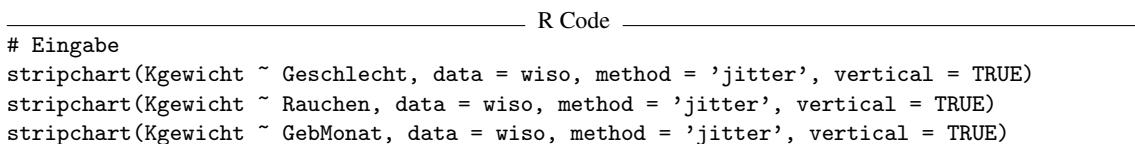
Abbildung 5.2: Nach ‘Geschlecht’ stratifizierte ECDFs aller numerischer Variablen in Beispiel 5.1.



Stripcharts

Eine Alternative zu den ECDFs stellen Stripcharts dar: Pro Ausprägung von X werden dabei die Y -Werte auf einer Linie eingezeichnet. Dies visualisiert sowohl die gemeinsame Verteilung der X - und Y -Werte als auch die auf X bedingten Verteilungen von Y . Im Gegensatz zu den ECDFs können (ausser Minimum und Maximum) keine Kenngrössen direkt abgelesen werden.

Beispiel 5.2 (Körpergewicht, Fortsetzung). Abbildung 5.3 zeigt Stripcharts für die Situation in Beispiel 5.1¹.



¹Dazu verwenden wir die R-Funktion `stripchart`. Bei vielen gleichen Werten bietet es sich an, die Werte mit der Option `method = 'jitter'` leicht verzittert darzustellen. Vertikale Stripcharts können mit `vertical = TRUE` angefordert werden. Die Tilde “~” in sogenannten *R-Formeln* kann generell als “in Abhängigkeit von” interpretiert werden.

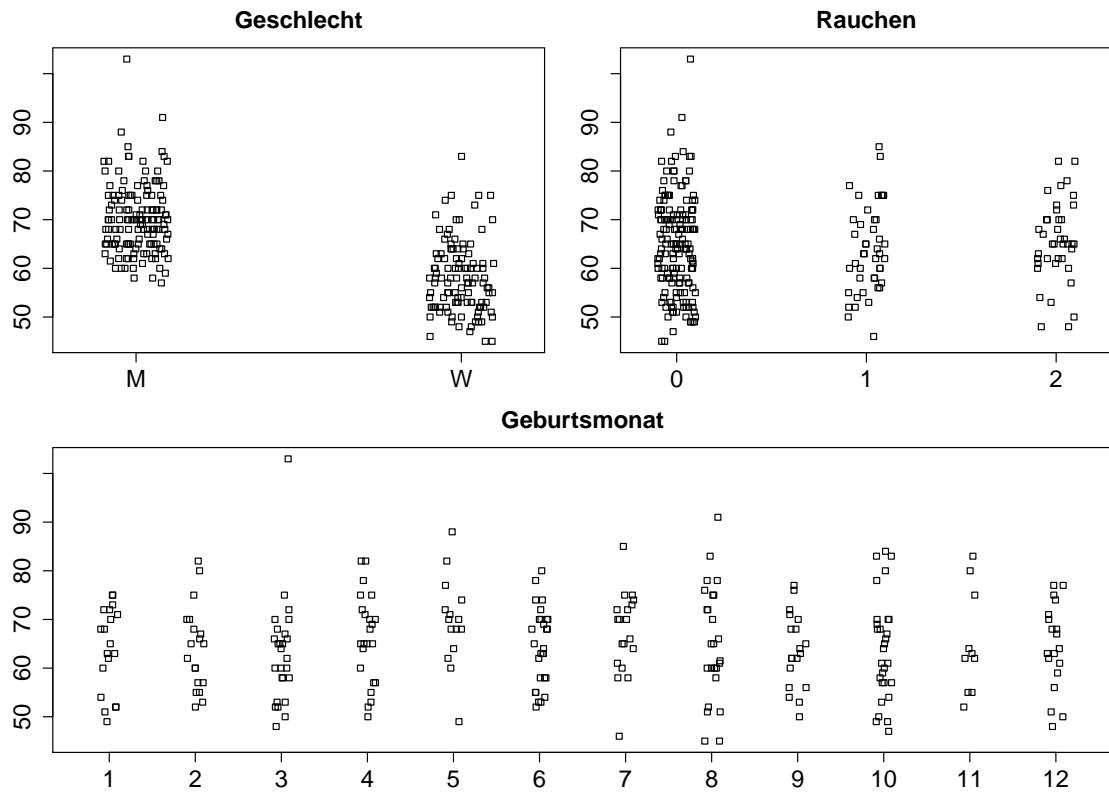


Abbildung 5.3: (Verzitterte) Stripcharts von ‘Körpergewicht’ je in Abhängigkeit von ‘Geschlecht’, ‘Rauchen’ und ‘Geburtsmonat’ (Beispiel 5.2).

Kommentare: Es folgen ähnliche Feststellungen hinsichtlich des Zusammenhangs wie mittels ECDFs. Die Vergleiche nach ‘Geburtsmonat’ sind deutlich übersichtlicher – es ist kein Zusammenhang zu erkennen. ▲

Box-(Whisker-)Plots

Als Alternative zu den obigen Verfahren erfand John W. Tukey die sogenannten *Boxplots* bzw. *Box-Whisker-Plots*. Ein solcher stellt im Wesentlichen die drei Quartile sowie einzelne Ausreisser eines numerischen (seltener auch eines ordinalen) Merkmals grafisch dar.

Konstruktion eines einzelnen Boxplots

- *y-Achse:* Die vertikale Achse entspricht den möglichen *Y*-Werten. (Anstatt vertikale Boxplots sind auch horizontale denkbar.)
- *Box:* Man zeichnet ein Rechteck (Box) mit unterer Kante auf der Höhe des ersten Quartils und oberer Kante auf der Höhe des dritten Quartils. Die Breite der Box ist frei wählbar.
- *Balken:* Auf der Höhe des Medians wird das Rechteck durch eine horizontale Linie (Balken) unterteilt.
- *Ausreisser:* Extrem grosse und kleine Stichprobenwerte werden als Punkte eingezeichnet. Als extrem gross bzw. klein zählt ein Wert, wenn er mehr als $1.5 \cdot \text{IQR}$ oberhalb des dritten Quartils bzw. unterhalb des ersten Quartils liegt¹. Solche Werte bezeichnen wir als Ausreisser.

¹ Manchmal werden auch andere Schranken verwendet, beispielsweise das 10%- und das 90%-Quantil.

- **Whiskers:** Von der Mitte der oberen Kante der Box bis zum grössten Stichprobenwert, der nicht als Ausreisser gilt, wird eine vertikale Linie gezeichnet, die durch einen horizontalen Strich abgeschlossen wird. Dies ist der obere Whisker (*Schnurrhaar*) des Boxplots. Der untere Whisker wird analog eingezeichnet. (Die Whiskers sind in der Regel nicht gleich lang.)

Gibt es keine Ausreisser, so enden die Whiskers dadurch beim kleinsten und grössten Stichprobenwert. Nur in diesem Fall kann ein Boxplot lediglich anhand der Kenngrössen Minimum, erstes Quartil, Median, drittes Quartil und Maximum (“five number summary”) konstruiert werden.

Beispiel 5.3 (Körpergewicht, Fortsetzung). Als Alternative zu den ECDFs in Abbildung 5.1 oder den Stripcharts in Abbildung 5.3 zeigt Abbildung 5.4 die entsprechenden Boxplots¹.

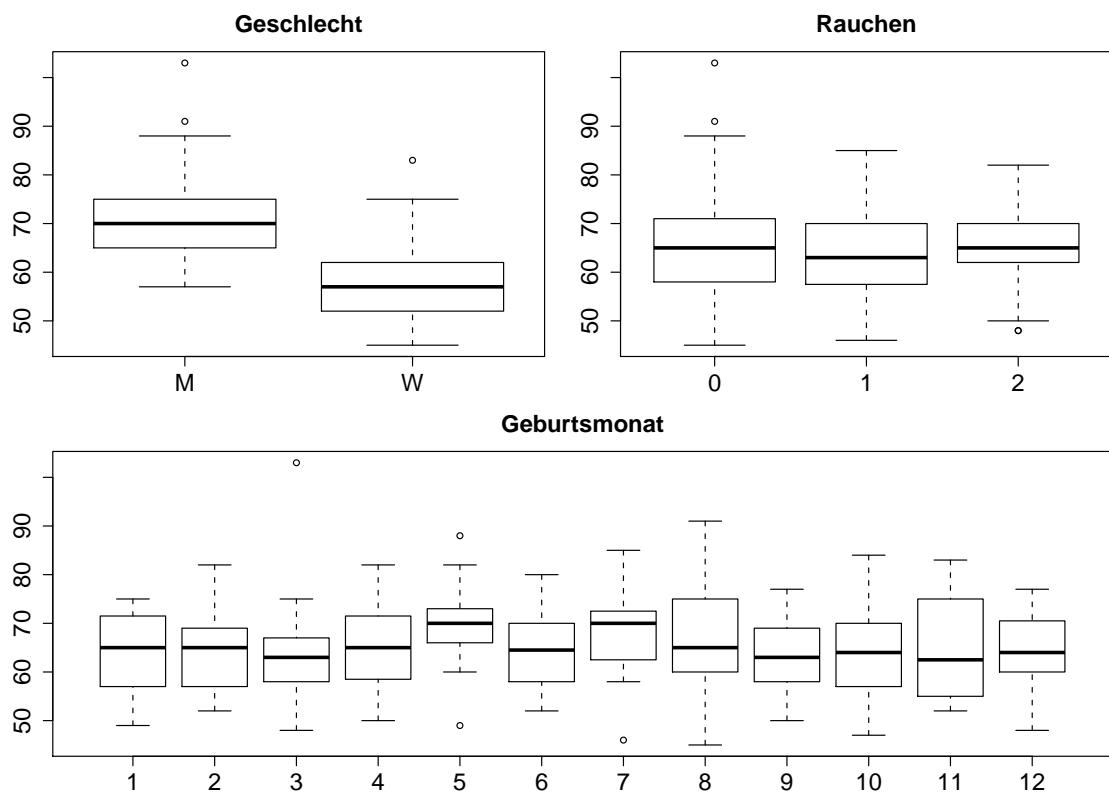


Abbildung 5.4: Boxplots von ‘Körpergewicht’ je in Abhängigkeit von ‘Geschlecht’, ‘Rauchen’ und ‘Geburtsmonat’ (Beispiel 5.3).

R Code

```
# Eingabe
boxplot(Kgewicht ~ Geschlecht, data = wiso)
boxplot(Kgewicht ~ Rauchen, data = wiso)
boxplot(Kgewicht ~ GebMonat, data = wiso)
```

Kommentare: Wir kommen zu den gleichen Schlüssen wie mit ECDFs und Stripcharts. Wie bei den ECDFs können wichtige Kenngrössen abgelesen werden. Im Gegensatz dazu bleiben Boxplots jedoch selbst bei vielen Gruppen übersichtlich.

Wie entsteht beispielsweise der Boxplot aller Frauen? Dazu betrachten wir die Verteilung des Körpergewichts aller Frauen. Die üblichen univariaten Kenngrössen ergänzen wir mit den kleinsten und grössten

¹Boxplots werden in R mit der `boxplot`-Funktion gezeichnet.

Werten 45, 45, 46, 47, 48 kg und 83 75 75 75 74 kg.

R Code

```
# Eingabe
werte <- wiso$Kgewicht[wiso$Geschlecht == "W"]
summary(werte)

# Ausgabe
Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
45.0   52.2   57.0  58.1   62.0  83.0  6.0
```

Die Box reicht von 52.2 bis 62 kg, der Medianbalken liegt auf der Höhe 57 kg. Werte ausserhalb von $1.5 \cdot \text{IQR} = 1.5 \cdot 9.8 = 14.7$ kg der Box, also Werte unterhalb von $52.2 - 14.7 = 37.5$ kg oder über $62 + 14.7 = 76.7$ kg, gelten als Ausreisser und werden einzeln eingezeichnet. Dies betrifft nur den Wert 83 kg. Die Whiskers reichen nun nicht bis zu 37.5 kg bzw. 76.7 kg, sondern bis zu den extremsten Werten in diesem Bereich: Der untere also bis 45 kg, der obere bis 75 kg. ▲

Eigenschaften

- Boxplots stellen univariate Verteilungen kompakt dar und eignen sich deshalb zum Vergleich mehrerer Teilstichproben.
- Wichtige Kenngrössen wie Minimum, Maximum, Quartile, IQR sowie Ausreisser sind sofort erkennbar. Beispielsweise enthält die Box die 50% zentralen Stichprobenwerte und der Balken trennt die 50% grössten von den 50% kleinsten Stichprobenwerten.
- Wir verfügen nun über ein einfaches Kriterium, um Ausreisser zu identifizieren.
- Ein bezüglich des Balkens symmetrischer Boxplot weist auf symmetrisch verteilte Stichprobenwerte hin. Die Anzahl der Höcker hingegen ist nicht erkennbar.
- Bei kleinen Stichproben oder vielen identischen Stichprobenwerten kann der Boxplot auch entarten in dem Sinne, dass z. B. der Balken mit einer Rechteckkante zusammenfällt.
- Bei normalverteilten Stichprobenwerten ist der Boxplot ungefähr symmetrisch und weist nur wenige (und nicht extreme) Ausreisser auf, siehe Abbildung 5.5. Dies erlaubt eine schnelle, jedoch nicht immer sorgfältige Beurteilung, ob die Stichprobenwerte aus einer Normalverteilung stammen können.

Abbildung 5.5 zeigt Boxplots von Stichproben aus einigen bekannten Verteilungen.

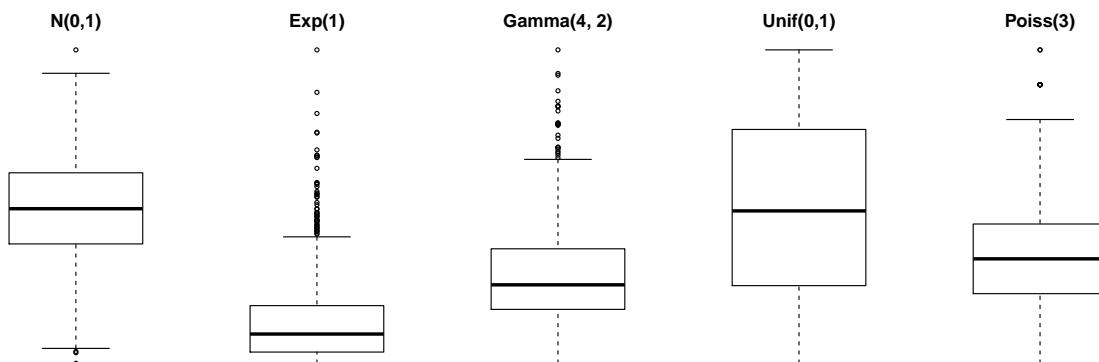


Abbildung 5.5: Boxplots von Stichproben ($n = 1000$) aus einigen bekannten Verteilungen.

Beispiel 5.4 (Leistung in Mathematik). Wir untersuchen nun für Beispiel 2.15 den Zusammenhang zwischen ‘Leistung in Mathematik’ und ‘Geschlecht’ (‘male’: nein = 0, ja = 1) sowie zwischen ‘Leistung in Mathematik’ und ‘Absenzverhalten’ (‘daysabs’: 0 – 1 Tage, 2 – 5 Tage, > 5 Tage) mithilfe von Abbildung 5.6.

R Code

```
# Eingabe
Ecdf(highschool$math, group = highschool$male)
Ecdf(highschool$math, group = highschool$daysabs)

boxplot(math~male, data = highschool)
boxplot(math~daysabs, data = highschool)

stripchart(math~male, data = highschool, method = 'jitter', vertical = TRUE)
stripchart(math~daysabs, data = highschool, method = 'jitter', vertical = TRUE)
```

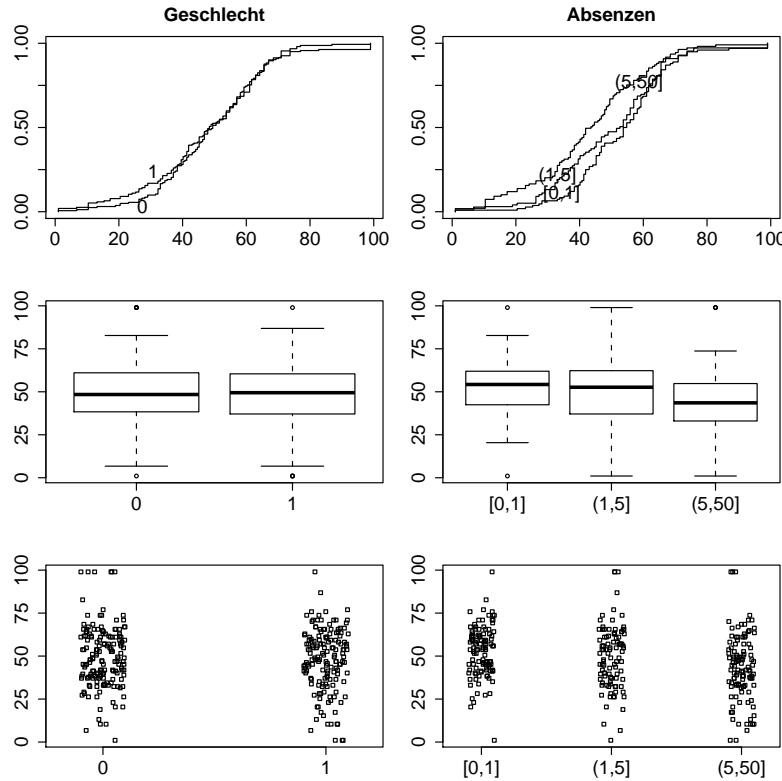


Abbildung 5.6: Grafische Darstellungen des Zusammenhangs zwischen ‘Leistung in Mathematik’ und ‘Geschlecht’ bzw. ‘Absenzen’ (Beispiel 5.4).

Kommentare

- ‘Leistung in Mathematik’ in Abhängigkeit von ‘Geschlecht’: Die Grafiken zeigen keinerlei Zusammenhang zwischen diesen Merkmalen: Die Verteilung von ‘Leistung in Mathematik’ unterscheidet sich praktisch nicht zwischen Männern und Frauen.
- ‘Leistung in Mathematik’ in Abhängigkeit von ‘Absenzen’: Auch zwischen ‘Leistung in Mathematik’ und ‘Absenzen’ ist höchstens ein schwacher Zusammenhang erkennbar: SchülerInnen mit mehr Absenzen neigen zu leicht schlechteren Leistungen. ▲

5.1.2 Quantitative Auswertung

Ausser mit Grafiken können die univariaten Verteilungen auch mithilfe von Kenngrössen (Mittelwert, Quantile, Streuungsmasse) zwischen den Teilstichproben verglichen werden. Häufig konzentriert man sich auf den Vergleich von Mittelwerten oder Medianen. Damit können Zusammenhänge in Form von Lageunterschieden beurteilt werden.

Auf dieser vereinfachten Betrachtungsweise beruhen die meisten Verfahren der schliessenden Statistik, die in solchen Situationen eingesetzt werden. Je nachdem, ob dort zwei oder mehrere Stichproben verglichen werden sollen, werden andere Verfahren verwendet. Deshalb gehen wir auf diese beiden Fälle im Anschluss an die deskriptiven Beispiele separat ein.

Beispiel 5.5 (Leistung in Mathematik, Fortsetzung). Nun ergänzen wir die grafischen Vergleiche von Beispiel 5.4 mit quantitativen Methoden¹.

R Code

```
# Eingabe
by(highschool$math, highschool$male, FUN = summary)

# Ausgabe
highschool$male: 0
  Min. 1st Qu. Median   Mean 3rd Qu.   Max.
  1.01   38.30  48.40  49.70  61.00  99.00
-----
highschool$male: 1
  Min. 1st Qu. Median   Mean 3rd Qu.   Max.
  1.01   37.10  49.50  47.80  60.30  99.00

# Eingabe
by(highschool$math, highschool$daysabs, FUN = summary)

# Ausgabe
highschool$daysabs: [0,1]
  Min. 1st Qu. Median   Mean 3rd Qu.   Max.
  1.01   42.70  54.30  52.90  61.80  99.00
-----
highschool$daysabs: (1,5]
  Min. 1st Qu. Median   Mean 3rd Qu.   Max.
  1.01   37.10  52.60  50.00  62.30  99.00
-----
highschool$daysabs: (5,50]
  Min. 1st Qu. Median   Mean 3rd Qu.   Max.
  1.01   33.00  43.60  43.50  54.80  99.00
```

Kommentare

- ‘Leistung in Mathematik’ in Abhängigkeit von ‘Geschlecht’: Quantile und auch Mittelwerte (49.7 bei den Frauen versus 47.8 bei den Männern) sind praktisch identisch. Somit ist (auch) anhand der quantitativen Angaben kein Zusammenhang erkennbar.
- ‘Leistung in Mathematik’ in Abhängigkeit von ‘Absenzen’: Während die Streuungen (IQR) ähnlich sind, lässt sich bei zunehmender Anzahl Absenzen ein schwacher Trend zu tieferen Leistungen erkennen: Die mittlere Leistung beträgt bei Personen mit höchstens einer Absenz 52.9 Punkte, bei Personen mit zwei bis fünf Absenzen 50 Punkte und bei solchen mit mehr als fünf Absenzen noch 43.5 Punkte. Somit können wir von einem schwachen Zusammenhang in Form eines Lageunterschieds sprechen. ▲

¹Die R-Funktion by wendet die Funktion FUN auf die Daten an, stratifiziert nach den Variablen im zweiten Argument.

Beispiel 5.6 (Körpergewicht und Geschlecht). Ergänzen wir die grafischen Vergleiche auch hier mit quantitativen Angaben.

R Code

```
# Eingabe
by(wiso$Kgewicht, wiso$Geschlecht, FUN = summary)

# Ausgabe
wiso$Geschlecht: M
  Min. 1st Qu. Median   Mean 3rd Qu.   Max.   NA's
  57.0    65.0    70.0    70.2    75.0   103.0    6.0
-----
wiso$Geschlecht: W
  Min. 1st Qu. Median   Mean 3rd Qu.   Max.   NA's
  45.0    52.2    57.0    58.1    62.0    83.0    6.0
```

Kommentar: Die Feststellungen sind ähnlich wie jene der grafischen Darstellungen. Anhand der Mittelwerte bzw. Mediane ist ein Zusammenhang in Form eines Lageunterschieds von rund 12 bzw. 13 kg erkennbar. Die Streuungen (IQR) der beiden Verteilungen sind ähnlich.

Nun konzentrieren wir uns auf die Mittelwerte¹ pro Teilstichprobe.

R Code

```
# Eingabe (benötigt library(Hmisc))
summary(Kgewicht ~ Geschlecht, data = wiso)

# Ausgabe
Kgewicht      N=251, 12 Missing
+-----+-----+
|       | N  |Kgewicht|
+-----+-----+
|M|141|70.244 |
|W|110|58.091 |
+-----+-----+
|Overall | 251|64.918 |
+-----+-----+
```

Kommentar: Das mittlere Gewicht der Männer beträgt 70.24 kg und ist somit um 12.15 kg grösser als jenes der Frauen (58.09 kg). ▲

5.2 Vergleich zweier Mittelwerte

5.2.1 Aussagen über die Stichprobe

Liegt ein binäres Merkmal X mit den Kategorien x_1 und x_2 vor, so lässt sich der Zusammenhang zwischen den X - und Y -Werten vereinfacht mithilfe der beiden Teilstichprobenmittelwerte

$$\bar{Y}_j := \text{Mittelwert der } Y\text{-Werte bei Beobachtungen mit } X\text{-Wert } x_j$$

bzw. ihres Unterschieds $\hat{\beta} := \bar{Y}_2 - \bar{Y}_1$ beschreiben. Der Mittelwertunterschied $\hat{\beta}$ wird manchmal *Effekt* von x_2 bezüglich x_1 auf den Mittelwert von Y genannt (auch “mittlerer Effekt von X auf Y ” oder “Effekt von X auf den Mittelwert von Y ”). Er charakterisiert den Zusammenhang zwischen X und Y durch eine einzelne anschauliche Zahl und spielt unter anderem auch die Rolle eines Zusammenhangsmassen.

¹ Interessieren wir uns lediglich für Mittelwerte, so können diese ebenfalls mit der R-Funktion `by` angezeigt werden. Eine schlankere Alternative ist die Funktion `summary` im Paket `Hmisc`. Sie gibt standardmäßig die Mittelwerte pro Teilstichprobe an sowie den globalen Mittelwert. Mit der Option `fun` können auch andere Kenngrössen verglichen werden.

5.2.2 Aussagen über die Population

In einer Zufallsstichprobe dienen \bar{Y}_1, \bar{Y}_2 bzw. $\hat{\beta}$ als naheliegende Schätzwerte für die tatsächlichen Teilpopulationsmittelwerte μ_1 und μ_2 bzw. den tatsächlichen Mittelwertunterschied (Effekt) β .

Mit den bereits bekannten (Einstichproben-)Studentverfahren sind Konfidenzintervalle und Hypothesentests für die Mittelwerte μ_1 und μ_2 möglich. Für den Mittelwertunterschied β (und damit den Zusammenhang in der Population) sind jedoch neue Überlegungen nötig: Falls die Beobachtungen Y_1, \dots, Y_n unabhängige normalverteilte Zufallsvariablen mit Erwartungswert μ_1 (bei X -Wert x_1) bzw. μ_2 (bei X -Wert x_2) und fixer (meist unbekannter) Standardabweichung sind, so ist der studentisierte Mittelwertunterschied

$$\frac{\hat{\beta} - \beta}{\text{Standardfehler von } \hat{\beta}}$$

studentverteilt mit $n - 2$ Freiheitsgraden. Konfidenzintervalle und Testentscheide für β findet man via Verteilungs- oder Quantilfunktion von t_{n-2} , wobei der Standardfehler von $\hat{\beta}$ mit einer umständlichen Formel aus den Standardabweichungen der Teilstichproben berechnet wird.

Sind die Teilstichproben *je* nicht zu klein und die Standardabweichungen der Y -Werte zwischen den Teilstichproben ähnlich, so gelten die Ergebnisse dieser berühmten Zweistichproben-Studentverfahren dank des Zentralen Grenzwertsatzes immerhin approximativ auch für nicht normalverteilte Beobachtungen.

Mit den Hypothesen $H_1 : \beta \neq 0$ versus $H_0 : \beta = 0$ lässt sich insbesondere prüfen, ob ein echter Zusammenhang (in Form eines Lageunterschieds¹) vorliegt. Es sind auch einseitige Tests (und damit Hypothesen zu “gerichteten” Zusammenhängen) möglich. Der Testentscheid wird wie üblich mit p -Wert, Teststatistik oder Konfidenzintervallansatz (Student-Konfidenzintervall im Zweistichprobenfall) getroffen.

Beispiel 5.7 (Leistung in Mathematik, Fortsetzung). In Beispiel 5.5 haben wir den Zusammenhang zwischen ‘Geschlecht’ und ‘Leistung in Mathematik’ untersucht. Dabei haben wir festgestellt, dass die mittlere Leistung bei den Schülern $\bar{Y}_1 = 47.767$ und bei den Schülerinnen $\bar{Y}_2 = 49.687$ Punkte beträgt. Mit einem mittleren Unterschied von $\hat{\beta} = 49.687 - 47.767 = 1.92$ Punkten zugunsten der Frauen ist dieser Zusammenhang allenfalls schwach. Welche Aussagen können wir mithilfe der Student-Verfahren² über den mittleren Unterschied β in der Population aller US-SchülerInnen bzw. den echten Zusammenhang zwischen den Merkmalen machen?

R Code

```
# Eingabe
t.test(math ~ male, data = highschool)

# Ausgabe
[...]
p-value = 0.3409
95 percent confidence interval: -2.0406  5.8811
sample estimates:
mean in group 0 mean in group 1
        49.687        47.767
```

Kommentare

- **Schätzwert:** Ein Schätzwert für β ist durch den mittleren Unterschied von $\hat{\beta} = 1.92$ Punkten gegeben. Wir schätzen also, dass Schülerinnen im Schnitt tatsächlich 1.92 Punkte besser als Schüler abschneiden.

¹Interessiert man sich für *irgendeinen* Zusammenhang in der Population, arbeitet man beispielsweise mit dem Zweistichprobentest nach Kolmogorov-Smirnov, der auf dem maximalen vertikalen Abstand zwischen den beiden ECDFs beruht.

²Dazu verwenden wir die R-Funktion `t.test`. (Standardmäßig wird die Methode nach Welch durchgeführt. Sie liefert bei ungleicher Varianz zwischen den Teilstichproben präzisere Ergebnisse als das Originalverfahren.)

- **95%-Konfidenzintervall:** Mit einer Sicherheit von rund 95% liegt β zwischen -2.0 und 5.9 Punkten (Zweistichproben-Student-Konfidenzintervall).
- **Test auf Zusammenhang auf 5%-Niveau:** Der p -Wert des t -Tests (Zweistichprobenversion) liegt mit 0.3409 nicht unter 0.05 . Damit wird die Nullhypothese von keinem mittleren Unterschied zwischen Schülern und Schülerinnen hinsichtlich der Leistung in Mathematik bzw. von keinem Zusammenhang zwischen ‘Geschlecht’ und ‘Leistung in Mathematik’ nicht verworfen. Zum gleichen Schluss führt der Konfidenzintervallansatz: Der Wert null liegt im Konfidenzintervall.
- **Präzision der Student-Verfahren:** Die Teilstichproben sind gross und die Streuung von ‘Leistung in Mathematik’ zwischen den Teilstichproben ähnlich (Abbildung 5.6). Somit trauen wir den Ergebnissen.

Betrachten wir nun eine gerichtete Fragestellung: Können wir anhand der Daten auf dem 5%-Niveau behaupten, dass Schülerinnen systematisch überlegen sind, also dass der wahre mittlere Unterschied β grösser als null ist?

R Code

```
# Eingabe
t.test(math ~ male, alternative = 'greater', data = highschool)

# Ausgabe
[...]
p-value = 0.1704
95 percent confidence interval: -1.400738      Inf
```

Kommentar: Die entsprechende Nullhypothese $\beta \leq 0$ kann auf dem 5%-Niveau nicht zugunsten der Arbeitshypothese $\beta > 0$ verworfen werden. Wir können also nicht behaupten, dass Schülerinnen in Mathematik im Schnitt wirklich besser sind als Schüler. Zum gleichen Schluss gelangen wir mit dem Konfidenzintervallansatz: Die untere Schranke für β ist nicht grösser als 0. ▲

Beispiel 5.8 (Körpergewicht und Geschlecht, Fortsetzung). Wir haben in Beispiel 5.6 festgestellt, dass sich das mittlere Gewicht $\bar{Y}_2 = 70.24$ kg der Männer um $\hat{\beta} = \bar{Y}_2 - \bar{Y}_1 = 70.24 - 58.09 = 12.15$ kg von jenem der Frauen ($\bar{Y}_1 = 58.09$ kg) unterscheidet. Der Effekt von Geschlecht auf das mittlere Körpergewicht beträgt in der Stichprobe somit etwa 12 kg zugunsten der Männer. Was können wir über den entsprechenden mittleren Unterschied β in der Population aller Studierenden sagen?

R Code

```
# Eingabe
t.test(Kgewicht ~ Geschlecht, data = wiso)

# Ausgabe
t = 13.068, df = 231.91, p-value < 2.2e-16
95 percent confidence interval: 10.32 13.99
sample estimates:
mean in group M mean in group W
    70.24          58.09
```

Kommentare

- **Schätzwert:** Wir schätzen, dass Männer im Schnitt tatsächlich $\hat{\beta} = 70.24 - 58.09 \approx 12.1$ kg schwerer sind als Frauen, bzw. dass der wahre Mittelwertunterschied $\beta = 12.1$ kg ist.
- **95%-Konfidenzintervall:** Mit einer Sicherheit von ca. 95% liegt β zwischen 10.32 und 13.99 kg (Zweistichproben-Student-Konfidenzintervall).

- *Test auf Zusammenhang auf 5%-Niveau:* Auf einem Niveau von 5% verwerfen wir die Nullhypothese $\beta = 0$ (zweiseitiger p -Wert des Zweistichproben- t -Tests < 0.0001). Mit einer Sicherheit von rund 95% behaupten wir also, dass sich das tatsächliche mittlere Gewicht der Frauen von jenem der Männer unterscheidet bzw. dass ein echter Zusammenhang zwischen ‘Geschlecht’ und ‘Körpergewicht’ besteht. Zum gleichen Testentscheid führt auch der Konfidenzintervallansatz: Das 95%-Konfidenzintervall ist nicht mit der Nullhypothese kompatibel (enthält 0 nicht).
- *Präzision:* Die Teilstichproben sind gross und die Streuungen zwischen den beiden Teilstichproben ähnlich (Abbildung 5.4), deshalb vertrauen wir den Ergebnissen der Student-Methoden.

Betrachten wir nun noch die einseitige Arbeitshypothese, dass Männer wirklich tendentiell schwerer sind als Frauen, also dass $\beta > 0$ versus die Nullhypothese $\beta \leq 0$:

R Code

```
# Eingabe
t.test(Kgewicht ~ Geschlecht, data = wiso, alternative = 'greater')

# Ausgabe
t = 13.068, df = 231.91, p-value < 2.2e-16
95 percent confidence interval:
 10.618     Inf
sample estimates:
mean in group M mean in group W
      70.245        58.091
```

Kommentar: Der p -Wert des einseitigen t -Tests liegt unter dem 5%-Niveau (p -Wert fast null), somit können wir mit einer Sicherheit von rund 95% behaupten, dass Männer im Schnitt tatsächlich schwerer als Frauen sind. Zum gleichen Testentscheid führt der Konfidenzintervallansatz: Der Wert null liegt nicht im Konfidenzintervall von 10.62 bis unendlich. Die Aussage via Konfidenzintervall ist jedoch (wie immer) gewichtiger: Damit können wir nämlich mit einer Sicherheit von rund 95% behaupten, dass Männer im Schnitt tatsächlich mindestens 10.6 kg schwerer sind als Frauen. ▲

Wilcoxons Rangsummentest

Ein anderer wichtiger Test auf Lageunterschied (und damit auf Zusammenhang) ist *Wilcoxons Rangsummentest*, auch *Mann-Whitney U-Test* genannt.

Bei diesem *rangbasierten* Verfahren werden nicht die Y -Werte zwischen den Gruppen verglichen, sondern deren *Ränge*: Der kleinste Y -Wert erhält den Rang 1, der zweitkleinste den Rang 2 etc. Ob sich diese im Schnitt tatsächlich unterscheiden, wird im Wesentlichen mit einem Zweistichproben- t -Test untersucht¹.

Dadurch weist dieser Test folgende Vor- und Nachteile gegenüber dem Student-Verfahren auf:

- + Er kann auf ordinale Y angewendet werden, da eine Rangierung auch dann sinnvoll ist. Dies erlaubt beispielsweise einen Test auf Zusammenhang zwischen einem ordinalen und einem binären Merkmal, z. B. zwischen Geschlecht und einem ordinalen Fragebogenitem. Im Gegensatz zum Chiquadrat-Unabhängigkeitstest, der in solchen Situationen eingesetzt werden könnte, nützt Wilcoxons Rangsummentest die Ordinalität der Y -Werte aus. Im Gegensatz zum t -Test hängt das Ergebnis nicht von der oft willkürlichen Wahl der Zahlenkodierung ab.
- + Wilcoxons Rangsummentest ist robust gegenüber Ausreissern in Y , da die Ränge stets Werte zwischen 1 und dem Stichprobenumfang n annehmen.

¹Bei kleinen Stichproben sind mit kombinatorischen Überlegungen exakte p -Werte möglich.

- + Er liefert auch für kleine, nicht normalverteilte Teilstichproben verlässliche Ergebnisse und führt ansonsten häufig zum gleichen Testentscheid wie das Student-Verfahren.
- + Der Test liefert die gleichen Ergebnisse für monoton steigend transformierte Y -Werte, beispielsweise für logarithmierte¹ Werte, da sich die Rangierung durch eine solche Transformation nicht verändert.
- Die p -Werte passen nicht zu den Student-Konfidenzintervallen.

Wilcoxons Rangsummentest wird deshalb dem t -Test oft vorgezogen.

Beispiel 5.9 (Leistung in Mathematik, Fortsetzung). Für Beispiel 5.7 ergibt sich mit Wilcoxons Rangsummentest² zwar ein anderer p -Wert, jedoch der gleiche Testentscheid:

R Code	wilcox.test(math ~ male, data = highschool)	# Ergibt p-value = 0.8101
--------	---	---------------------------

Zum (fast) gleichen Ergebnis kommen wir, indem wir einen t -Test auf die Ränge der Leistungen anwenden.

R Code	t.test(rank(math) ~ male, data = highschool)	# Ergibt p-value = 0.8102
--------	--	---------------------------

Eine Log-Transformation beeinflusst zwar den p -Wert des t -Tests, nicht aber jenen des Rangsummentests:

R Code	wilcox.test(log(math) ~ male, data = highschool)	# Ergibt p-value = 0.8101
	t.test(math ~ male, data = highschool)	# Ergibt p-value = 0.3409
	t.test(log(math) ~ male, data = highschool)	# Ergibt p-value = 0.1987



Beispiel 5.10 (Absenzen und Schule, Fortsetzung). In Beispiel 4.8 haben wir den Zusammenhang zwischen der kategorisierten Anzahl Absenzen ‘daysabs’ und ‘Schule’ mit Cramérs V (0.36) und dem Chiquadrat-Unabhängigkeitstest (p -Wert < 0.0001) untersucht.

Zum gleichen Testentscheid (jedoch nicht zum gleichen p -Wert) führt Wilcoxons Rangsummentest:

R Code	wilcox.test(as.numeric(daysabs) ~ school, data = highschool) # Ergibt p-value = 2.933e-10
--------	---



Beispiel 5.11 (Körpergewicht und Geschlecht, Fortsetzung). Die Testentscheide in Beispiel 5.8 bleiben gleich:

R Code	# Zweiseitig wilcox.test(Kgewicht ~ Geschlecht, data = wiso)	# Ergibt p-value < 2.2e-16
	# Eingabe: Einseitig wilcox.test(Kgewicht ~ Geschlecht, data = wiso, alternative = 'greater')	# Ergibt p-value < 2.2e-16



¹ Warum dies manchmal gemacht wird, sehen wir später bei den Regressionsmodellen.

² Dieser Test wird in R mit der Funktion `wilcox.test` angefordert.

5.3 Vergleich mehrerer Mittelwerte

Nun betrachten wir ein Merkmal X mit mehreren Kategorien x_1, x_2, \dots, x_L . Der Zusammenhang zwischen den X - und Y -Werten wird *vereinfacht* mithilfe der Teilstichprobenmittelwerte

$$\bar{Y}_j := \text{Mittelwert der } Y\text{-Werte bei Beobachtungen mit } X\text{-Wert } x_j$$

und deren Unterschiede charakterisiert. Im Gegensatz zum Zweistichprobenfall mit einem *einzigem* Mittelwertunterschied gibt es im Mehrstichprobenfall nun deren $\binom{L}{2} = L(L-1)/2$. Bei $L = 3$ Teilstichproben gibt es damit beispielsweise drei Mittelwertunterschiede, bei $L = 12$ (z. B. Monate) unübersichtliche 66.

Damit stellt sich insbesondere die Frage, wie die Stärke des Zusammenhangs durch eine einzige Zahl quantifiziert werden kann. Ein geeignetes Zusammenhangsmass in dieser Situation ist das sogenannte *Bestimmtheitsmass*, das wir nun mithilfe von *Vorhersagen*, *gefitteten Werten* und *Residuen* einführen. Da es gewisse Varianzen (nicht jene von Y in den Teilstichproben!) vergleicht, werden die im weiteren Sinne damit verbundenen Verfahren unter dem Oberbegriff *Varianzanalyse*, kurz: *ANOVA* von engl. *analysis of variance*, zusammengefasst.

5.3.1 Bestimmtheitsmass

Für eine *zukünftige* Beobachtung, von der lediglich bekannt ist, dass ihr X -Wert x_j beträgt, würde man schätzen, dass der entsprechende Y -Wert gleich dem Mittelwert \bar{Y}_j aller Beobachtungen mit gleichem X -Wert (x_j) ist. Man hofft, dass diese *Vorhersage* (*Prognose*, *Prädiktion*, *Tipp*) besser als der einfache Schätzwert \bar{Y} ist.

Für jede Beobachtung (Y_i, X_i) betrachten wir nun den sogenannten *gefitteten (angepassten) Wert* \hat{Y}_i . Er entspricht der Vorhersage von Y_i durch X_i , also dem Mittelwert von Y aller Beobachtungen mit entsprechendem X -Wert. (Bei X -Wert x_j ist also $\hat{Y}_i = \bar{Y}_j$.)

Je besser die gefitteten Werte \hat{Y}_i mit den beobachteten Werten Y_i übereinstimmen bzw. je kleiner die Varianz der *Residuen* (Reste)

$$e_i := Y_i - \hat{Y}_i$$

gegenüber der Varianz der Y_i ist, je stärker ist der Zusammenhang zwischen den X - und Y -Werten.¹ Als Mass für die Stärke des Zusammenhangs bietet sich entsprechend das *Bestimmtheitsmass*

$$R^2 := 1 - \frac{\text{Var}(e)}{\text{Var}(Y)} = \frac{\text{Var}(\hat{Y})}{\text{Var}(Y)}$$

an. Es entspricht dem Prozentsatz der Varianz von Y , der durch X erklärt ist, und liegt deswegen immer zwischen 0 (kein Zusammenhang) und 1 (“perfekter” Zusammenhang: Innerhalb der Teilstichproben sind alle Y -Werte genau gleich).

Hinweise zu Residuen, gefitteten Werten und dem Bestimmtheitsmass

- *Zerlegung (I):* Ausgehend von der Gleichung $e_i = Y_i - \hat{Y}_i$ lassen sich die Y -Werte als Summe aus gefitteten Werten und Residuen schreiben, also $Y_i = \hat{Y}_i + e_i$ für alle $i = 1, \dots, n$. Da die Residuen im Schnitt null betragen, folgt daraus z. B. $\bar{Y} = \bar{\hat{Y}} + \bar{e} = \bar{\hat{Y}}$, also dass die gefitteten Werte den gleichen Mittelwert wie die Y -Werte aufweisen.

¹Sowohl die gefitteten Werte als auch die Residuen können als neue, künstliche Merkmale aufgefasst werden, deren Ausprägungen aus den X - und Y -Werten berechnet werden.

- *Zerlegung (II)*: Aufgrund der Konstruktion¹ lässt sich auch die Varianz der Y -Werte in

$$\text{Var}(Y) = \text{Var}(\hat{Y}) + \text{Var}(e)$$

zerlegen. Dies erklärt die rechte Hälfte der Formel zum Bestimmtheitsmass. Indem man diese Gleichung mal $n - 1$ rechnet, erhält man die oft zitierte Beziehung

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n e_i^2$$

(‘Gesamtvariabilität gleich Variabilität zwischen den Gruppen plus Variabilität innerhalb der Gruppen’).

- *Faustregeln*: Wie beim Cramérs V gibt es keine offiziellen Faustregeln, ab welchen Werten des Bestimmtheitsmasses man von einem deutlichen Zusammenhang sprechen kann – eine gute Einschätzung hängt von der Situation ab. Der Einfachheit halber sprechen wir manchmal bei Werten bis 0.1 allenfalls von einem schwachen Zusammenhang und bei Werten ab 0.3 von einem starken.

Beispiel 5.12 (Körpergewicht und Rauchen). Nun betrachten wir den Datensatz aus Beispiel 1.1 und beschreiben den Zusammenhang² zwischen ‘Körpergewicht’ und ‘Rauchen’ mithilfe einer Varianzanalyse.

R Code

```
# Eingabe
summary(Kgewicht ~ Rauchen, data = wiso)      # library(Hmisc)

# Ausgabe
Kgewicht      N=251, 12 Missing

+-----+-----+---+-----+
|       |       |N   |Kgewicht|
+-----+-----+---+-----+
|Rauchen|0      |165|65.027 |
|       |1      | 44|63.750 |
|       |2      | 41|65.610 |
|       |Missing| 1|70.000 |
+-----+-----+---+-----+
|Overall|      |251|64.918 |
+-----+-----+---+-----+

# Eingabe: Varianzanalyse
fit <- lm(Kgewicht ~ Rauchen, data = wiso)
summary(fit)

# Ausgabe
[...]
Multiple R-squared: 0.00365
```

Kommentar: Die Gruppenmittelwerte sind sehr ähnlich, deshalb ist kein Zusammenhang erkennbar. Diese subjektive Beurteilung wird nun mit dem Bestimmtheitsmass ergänzt: ‘Rauchen’ erklärt nur 0.365% der Varianz von ‘Körpergewicht’, also praktisch nichts. (Siehe auch die Boxplots in Abbildung 5.4.)

Zur Illustration der gefitteten Werte und Residuen betrachten wir nun die ersten Beobachtungen von ‘Körpergewicht’, ‘Rauchen’ und den beiden künstlichen Merkmalen ‘Gefittet’ und ‘Residuen’³.

¹Wir werden im nächsten Kapitel sehen, dass Varianzen und Summen nicht nur (wie bei den Regeln zu Varianzen gesehen) unter Unabhängigkeit vertauscht werden können, sondern auch unter einer leicht schwächeren Voraussetzung, der *Unkorreliertheit*, die hier gegeben ist.

²Das Bestimmtheitsmass bestimmen wir mit der R-Funktion `summary` aus dem Ergebnis von `lm` (für “lineares Modell”).

³Hier wählen wir mit R zuerst die beiden Variablen aus (`na.omit` löscht Beobachtungen mit fehlenden Werten) und ergänzen sie mit den durch `fitted` und `resid` ermittelten neuen Merkmalen.

R Code

```
# Eingabe, Forts.
no.missings <- na.omit(wiso[c("Kgewicht", "Rauchen")])
out <- data.frame(no.missings, Gefittet = fitted(fit), Residuen = resid(fit))
head(out)

# Ausgabe
  Kgewicht Rauchen    Gefittet  Residuen
1   68.0      0  65.02727  2.97273
2   62.0      2  65.60976 -3.60976
3   72.0      0  65.02727  6.97273
4   65.0      2  65.60976 -0.60976
5   57.0      0  65.02727 -8.02727
6   49.0      0  65.02727 -16.02727
```

Die gefitteten Werte entsprechen dem mittleren Gewicht in der entsprechenden Raucherkategorie, deshalb weisen beispielsweise alle NichtraucherInnen einen gefitteten Wert von 65.02727 kg auf. Da der Nichtraucher in der fünften Zeile 57 kg wiegt, beträgt sein Residuum $57 - 65.02727 = -8.02727$ kg.

Mit den Varianzen der drei Merkmale ‘Körpergewicht’ (89.717 kg^2), ‘Gefittet’ (0.327 kg^2) und ‘Residuen’ (89.390 kg^2) ergibt sich

$$R^2 = 1 - 89.390/89.717 = 0.327/89.717 = 0.0036.$$

(Dank der erwähnten Varianzzerlegung folgt die eine Varianz aus den jeweils anderen beiden.) ▲

Das Bestimmtheitsmaß dient als Ergänzung zu den Teilstichprobenmittelwerten, nicht als Ersatz, wie folgendes hypothetische Beispiel zeigt: Angenommen, in einer Stichprobe würde jede Frau exakt 6000 und jeder Mann exakt 7000 CHF pro Monat verdienen. Da man mithilfe des Geschlechts den Lohn genau vorhersagen kann (alle Residuen sind null), ergibt sich ein R^2 von 1. Dasselbe gilt jedoch auch für eine Stichprobe, in der jede Frau 6000 und jeder Mann 6001 CHF verdient.

5.3.2 Aussagen über die Population

Wie bereits beim Zweistichprobenfall dienen die Mittelwerte \bar{Y}_j als naheliegende Schätzwerte für die Teilpopulationsmittelwerte $\mu_j := E(Y | X = x_j)$, also für die Erwartungswerte der Zufallsvariable Y bei Beobachtungen mit X -Wert x_j , und es können univariante Student-Konfidenzintervalle dafür sinnvoll sein.

Das Bestimmtheitsmaß schätzt dessen Wert θ in der Population und es können approximative Konfidenzintervalle für θ ausgewiesen werden. Analog zum Cramérs V ist eine untere $(1 - \alpha) \cdot 100\%$ -Konfidenzschranke besonders nützlich, da sie angibt, wie stark der Zusammenhang tatsächlich mit einer Sicherheit von etwa $(1 - \alpha) \cdot 100\% \text{ mindestens}$ ist. Liegt die Schranke über null ($\theta = 0$ entspricht keinem Zusammenhang in der Population), so können wir mit entsprechender Sicherheit behaupten, es läge ein echter Zusammenhang zwischen X und Y vor.

Zum gleichen Testentscheid gelangt man mit dem *F-Test*. Dieser Test auf Zusammenhang überprüft folgende (äquivalente) Nullhypotesen:

- Das R^2 in der Population ist null, also $\theta = 0$.
- Es gibt keinen echten Zusammenhang zwischen X und Y .
- Die wahren Mittelwerte sind alle gleich, also $\mu_1 = \mu_2 = \dots = \mu_L$.

Die entsprechenden (äquivalenten) Arbeitshypothesen lauten:

- Das R^2 in der Population ist grösser als null, also $\theta > 0$.
- Es gibt einen echten Zusammenhang zwischen X und Y .
- Nicht alle wahren Mittelwerte sind gleich.

Der Test beruht auf der Tatsache, dass die *F-Teststatistik*

$$\frac{n-L}{L-1} \cdot \frac{R^2}{1-R^2}$$

unter der Nullhypothese zumindest für normalverteilte Beobachtungen mit gleicher Varianz eine *F-Verteilung* mit $L-1$ und $n-L$ Freiheitsgraden aufweist. Diese stetige Verteilung mit zwei Parametern ist – wie auch der *F*-Test und Fishers exakter Test für Vierfeldertafeln – nach Ronald A. Fisher benannt und ist eng mit der Chi-quadrat-Verteilung verwandt.

Sind die Teilstichproben *je* nicht zu klein und die Standardabweichungen der Y -Werte zwischen den Teilstichproben ähnlich, so gelten die Ergebnisse des *F*-Tests bzw. des entsprechenden Konfidenzintervalls dank des Zentralen Grenzwertsatzes immerhin approximativ auch für *nicht normalverteilte* Beobachtungen.

Beispiel 5.13 (Körpergewicht und Rauchen, Fortsetzung). Nun betrachten wir den Datensatz aus Beispiel 5.13 als Zufallsstichprobe aus allen Studierenden an Schweizer Universitäten. Was können wir über den echten Zusammenhang zwischen ‘Körpergewicht’ und ‘Rauchen’ sagen?¹

R Code

```
# Eingabe, Forts.
fit <- lm(Kge wicht ~ Rauchen, data = wiso)
summary(fit)

# Ausgabe
[...]
Multiple R-squared: 0.00365
F-statistic: 0.452 on 2 and 247 DF, F-statistic: p-value: 0.637

# Eingabe, Forts.: Konfidenzintervall für R-Quadrat
confint.R2(fit)

# Ausgabe
Lower.Limit Upper.Limit
0.000000 0.025813

#Eingabe, Forts.: Untere Konfidenzschranke
confint.R2(fit, alternative = 'greater')

# Ausgabe
Lower.Limit
0
```

Kommentare

- *Schätzwerte*: Wir schätzen, dass NichtraucherInnen im Schnitt 65.0 kg, GelegenheitsraucherInnen 63.8 kg und regelmässige RaucherInnen 65.6 kg wiegen. Ein Schätzwert für das Bestimmtheitsmass θ in der Population ist gegeben durch 0.00365.
- *Konfidenzintervall-/Schranke für θ* : Mit einer Sicherheit von rund 95% liegt θ zwischen 0 und 0.026, ist also praktisch null. Da die untere 95%-Konfidenzschranke nicht grösser als null ist, können wir auf dem 5%-Niveau nicht von einem Zusammenhang in der Population sprechen.

¹Der *F*-Test gehört zum Output von `summary(lm(...))`. Konfidenzintervalle für das wahre R^2 ermitteln wir mit der eigenen Funktion `confint.R2` im Anhang.

- *Test auf Zusammenhang auf 5%-Niveau:* Zum gleichen Testentscheid führt der F -Test: Mit 0.637 ist der p -Wert nicht kleiner als 0.05. Deshalb gibt es keinen Grund, die Nullhypothese von keinem Zusammenhang ($\theta = 0$) zugunsten der Arbeitshypothese eines echten Zusammenhangs ($\theta > 0$) zu verwerfen.
- *Präzision des F-Tests und der Konfidenzintervalle:* Die Teilstichproben sind gross und die Streuungen von ‘Körpergrösse’ dazwischen ähnlich (Abbildung 5.4), deshalb trauen wir den Ergebnissen. ▲

Hinweis (“Rückwärtskompatibilität”). Im Zweistichprobenfall liefert der F -Test den gleichen p -Wert wie der t -Test auf Zusammenhang, obwohl die beiden Tests auf anderen Zusammenhangsmassen beruhen.

Beispiel 5.14 (Körpergewicht und Geschlecht, Fortsetzung). In den Beispielen 5.8 und 5.11 haben wir festgestellt, dass sich die Männer (70.24 kg) und die Frauen (58.09 kg) hinsichtlich ihres mittleren Gewichts deutlich unterscheiden bzw. dass ein starker Zusammenhang zwischen ‘Geschlecht’ und ‘Körpergewicht’ besteht. Was ergibt hier eine ANOVA?

R Code

```
# Eingabe
fit <- lm(Kgewicht ~ Geschlecht, data = wiso)
summary(fit)

# Ausgabe
[...]
Multiple R-squared: 0.408
F-statistic: 172 on 1 and 249 DF, p-value: <2e-16

# Eingabe, Forts.: Konfidenzintervall für R-Quadrat
confint.R2(fit)

# Ausgabe
Lower.Limit Upper.Limit
0.31809    0.48471

# Eingabe, Forts.: Untere Konfidenzschranke
confint.R2(fit, alternative = 'greater') # Ergibt 0.33261
```

Kommentare

- *Bestimmtheitsmass in der Stichprobe:* Durch ‘Geschlecht’ wird 40.8% der Varianz von ‘Körpergewicht’ erklärt, es liegt also ein deutlicher Zusammenhang zwischen den Merkmalen vor.
- *Bestimmtheitsmass in der Population:* Wir schätzen den wahren Wert θ des Bestimmtheitsmasses auf 0.408. Mit einer Sicherheit von rund 95% liegt θ zwischen 0.318 und 0.485, was auf einen deutlichen Zusammenhang in der Population hindeutet. Die untere 95%-Konfidenzschranke für θ ist mit 0.33 insbesondere grösser als null, somit können wir mit einer Sicherheit von rund 95% behaupten, dass ein echter Zusammenhang vorliegt.
- *Test auf Zusammenhang auf 5%-Niveau:* Zum gleichen Testentscheid führt der F -Test: Sein p -Wert ist kleiner als 0.05 (fast null), deshalb verwerfen wir die Nullhypothese von keinem Zusammenhang ($\theta = 0$) zugunsten der Arbeitshypothese “Es gibt einen Zusammenhang” ($\theta > 0$). Der p -Wert ist gleich wie jener des t -Tests. (In diesem Beispiel sind beide einfach kleiner als 2e-16.)
- *Präzision des F-Tests und der Konfidenzintervalle:* Die Teilstichproben sind gross und die Streuungen von ‘Körpergrösse’ dazwischen ähnlich (Abbildung 5.4), deshalb trauen wir den Ergebnissen. ▲

Das rangbasierte Analogon zum F -Test heisst *Rangsummentest nach Kruskal und Wallis*. Dieser Test kann als Mehrstichprobenversion von Wilcoxons Rangsummentest angesehen werden und weist die gleichen Vorzüge gegenüber dem F -Test auf wie Wilcoxons Rangsummentest gegenüber dem t -Test.

Da der Test auch bei ordinalen Y -Werten angewendet werden kann, können damit zum Beispiel (als Alternative zum Chiquadrat-Unabhängigkeitstest oder allenfalls zum F -Test) Zusammenhänge zwischen einem ordinalen Fragebogenitem und einer kategorialen Information wie ‘Ausbildung’ geprüft werden.

Hinweis (“Rückwärtskompatibilität”). Im Zweistichprobenfall liefert der Rangsummentest nach Kruskal und Wallis den gleichen p -Wert wie Wilcoxons Rangsummentest auf Zusammenhang.

Beispiel 5.15 (Körpergewicht und Rauchen, Fortsetzung). Obwohl der p -Wert nicht gleich ist wie jener des F -Tests in Beispiel 5.13 (p -Wert: 0.64), folgt mit dem Rangsummentest nach Kruskal und Wallis¹ der gleiche Testentscheid.

R Code	kruskal.test(Kgewicht ~ Rauchen, data = wiso) # Ergibt p-value = 0.5395
--------	---

Beispiel 5.16 (Körpergewicht und Geschlecht, Fortsetzung). Auch für Beispiel 5.14 finden wir den gleichen Testentscheid (äquivalentes Ergebnis wie mit Wilcoxons Rangsummentest in Beispiel 5.11):

R Code	kruskal.test(Kgewicht ~ Geschlecht, data = wiso) # Ergibt p-value < 2.2e-16
--------	---

Beispiel 5.17 (Leistung in Mathematik, Fortsetzung). In Beispiel 5.5 haben wir festgestellt, dass Schüler mit mehr Absenzen zu leicht schlechteren Leistungen in Mathematik neigen. Welche zusätzlichen Schlüsse ziehen wir aus einer ANOVA?

(Für die schliessende Statistik fassen wir die Daten als Zufallsstichprobe aus allen US-SchülerInnen auf.)

R Code	# Eingabe: Varianzanalyse fit <- lm(math ~ daysabs, data = highschoo) summary(fit)
--------	--

# Ausgabe [...]	Multiple R-squared: 0.0497 F-statistic: 8.19 on 2 and 313 DF, p-value: 0.000342
--------------------	--

# Eingabe, Forts.: Konfidenzintervall für R-Quadrat confint.R2(fit)	
--	--

# Ausgabe Lower.Limit Upper.Limit 0.011014 0.100061	
---	--

# Eingabe, Forts.: Untere Konfidenzschanke confint.R2(fit, alternative = 'greater')	
--	--

# Ausgabe Lower limit 0.015202	
-----------------------------------	--

# Eingabe: Rangsummentest nach Kruskal-Wallis kruskal.test(math ~ daysabs, data = highschoo) # Ergibt p-value = 0.0001993	
--	--

¹In R wird der Rangsummentest nach Kruskal-Wallis mit dem Befehl `kruskal.test` angefordert.

Kommentare

- *Stärke des Zusammenhangs in Stichprobe:* Das Bestimmtheitsmass beträgt 0.0497. Das Absenzverhalten erklärt also nur etwa 5% der Varianz der Leistung in Mathematik, somit liegt nur ein sehr schwacher Zusammenhang zwischen den beiden Merkmalen vor. Dies konkretisiert die vage Feststellung bezüglich der Stärke des Zusammenhangs in Beispiel 5.5.
- *Schätzwerte für Teilpopulationsmittelwerte:* Anhand der Ergebnisse in Beispiel 5.5 schätzen wir die wahre mittlere Leistung in Mathematik bei SchülerInnen mit 0 – 1 Absenzen auf 52.9, bei solchen mit 2 – 5 Absenzen auf 50 und bei jenen über 5 Absenzen auf 43.5 Punkte.
- *Aussagen über tatsächliches Bestimmtheitsmass θ :* Ein Schätzwert für θ ist gegeben durch den Wert in der Stichprobe, also 0.05. Mit einer Sicherheit von rund 95% wird durch ‘Absenzen’ zwischen 1 und 10% der Varianz von ‘Leistung in Mathematik’ erklärt. Da die untere 95%-Konfidenzschranke grösser als null ist, können wir mit einer Sicherheit von rund 95% behaupten, dass es einen echten Zusammenhang zwischen ‘Leistung in Mathematik’ und ‘Absenzen’ gibt.
- *Test auf Zusammenhang auf 5%-Niveau:* Zum gleichen Testentscheid führt der F -Test: Der p -Wert 0.00034 ist kleiner als 0.05, somit kann die Nullhypothese von keinem Zusammenhang ($\theta = 0$) verworfen werden und man kann mit einer Sicherheit von rund 95% behaupten, dass ein echter Zusammenhang zwischen ‘Absenzen’ und ‘Leistung in Mathematik’ besteht ($\theta > 0$).
- *Präzision:* Wir vertrauen dem p -Wert des F -Tests sowie den Konfidenzintervallen für θ , da die Teilstichproben recht gross sind und die Werte von ‘Leistung in Mathematik’ zwischen den Teilstichproben ähnlich stark streuen (Abbildung 5.6).
- *Rangsummentest nach Kruskal-Wallis auf 5%-Niveau:* Auch der rangbasierte Test führt hier zum selben Testentscheid (p -Wert anders).

Zur Illustration der Konstruktion des Bestimmtheitsmasses betrachten wir einige Zeilen des Datensatzes inkl. gefitteten Werten und Residuen zusammen mit den relevanten Stichprobenvarianzen.

R Code

```
# Eingabe, Forts.: Datenausschnitt
out <- data.frame(highschool[c('math', 'daysabs')], fitted(fit), resid(fit))
out[17:21, ]

# Ausgabe
  math daysabs      fitted  residuals
41.314 [0,1]      52.920   -11.606
41.885 [0,1]      52.920   -11.035
65.560 (1,5]      49.965    15.595
13.131 (5,50]     43.518   -30.387
33.017 (1,5]      49.965   -16.948

# Eingabe, Forts.: Varianzen
var(highschool$math)    # Ergibt 319.72
var(resid(fit))        # Ergibt 303.82
var(fitted(fit))       # Ergibt 15.90
```

Die gefitteten Werte entsprechen den mittleren Leistungen in Mathematik in der jeweiligen Absenzkategorie. SchülerInnen mit 0 – 1 Absenzen (bspw. die ersten beiden Zeilen in diesem Datenausschnitt) weisen deshalb gefittete Werte von 52.9 Punkten auf. Die Residuen messen den Unterschied zwischen beobachteten und gefitteten Werten: Beispielsweise hat die SchülerIn in der ersten Zeile 41.3 Punkte erreicht, also beträgt ihr Residuum $41.3 - 52.9 = -11.6$ Punkte.

Laut Output beträgt die Stichprobenvarianz von ‘Leistung in Mathematik’ 319.72, jene der Residuen 303.82 und jene der gefitteten Werte 15.9 (nämlich 319.72 – 303.82). Somit finden wir

$$R^2 = 1 - 303.82/319.72 = 15.9/319.72 = 0.0497.$$

▲

5.3.3 Mittelwertunterschiede

Im Zweistichprobenfall ist der einzige Mittelwertunterschied ideal, um den Zusammenhang zwischen den Merkmalen zu beschreiben und Rückschlüsse auf den Zusammenhang in der Population zu machen. Auch im Mehrstichprobenfall sind Mittelwertunterschiede eine gute Ergänzung zu den Teilstichprobenmittelwerten und dem Bestimmtheitsmaß. Allerdings stellt sich das bereits angesprochene Problem, dass deren Anzahl

$$\frac{L(L-1)}{2}$$

bei wachsender Anzahl Kategorien L überproportional zunimmt. Falls es eine besonders *wichtige* oder *häufige* Kategorie von X , sagen wir x_1 , gibt, betrachtet man oft nur die $L-1$ Mittelwertunterschiede

$$\hat{\beta}_j = \bar{Y}_j - \bar{Y}_1$$

der anderen Kategorien zu dieser *Referenzkategorie* und fasst sie als Effekte bezüglich der Referenzkategorie auf den Mittelwert von Y auf. Umgekehrt können die Teilstichprobenmittelwerte mit der äquivalenten Gleichung

$$\bar{Y}_j = \bar{Y}_1 + \hat{\beta}_j$$

aus \bar{Y}_1 und den entsprechenden Effekten $\hat{\beta}_j$ gefunden werden (Wir setzen $\hat{\beta}_1 := 0$).

Beispiel 5.18 (Leistung in Mathematik, Fortsetzung). In Beispiel 5.5 haben wir gesehen, dass die mittlere Leistung in Mathematik bei Personen mit höchstens einer Absenz $\bar{Y}_1 = 52.9$ Punkte beträgt, bei Personen mit zwei bis fünf Absenzen $\bar{Y}_2 = 50$ Punkte und bei solchen mit mehr als fünf Absenzen 43.5 Punkte.

Wählen wir nun die Personen mit höchstens einer Absenz als Referenzkategorie, so ist

$$\hat{\beta}_2 = \bar{Y}_2 - \bar{Y}_1 = 50 - 52.9 = -2.9$$

der Effekt von 2–5 Absenzen bezüglich 0–1 Absenzen auf die mittlere Leistung in Mathematik. SchülerInnen mit 2–5 Absenzen erreichen im Schnitt also 2.9 Punkte weniger als solche mit höchstens einer Absenz.

Analog bezeichnet $\hat{\beta}_3 = \bar{Y}_3 - \bar{Y}_1 = 43.5 - 52.9 = -9.4$ den Effekt von mehr als fünf Absenzen bezüglich 0–1 Absenzen auf die mittlere Leistung in Mathematik. SchülerInnen mit mehr als fünf Absenzen erreichen im Schnitt also 9.4 Punkte weniger als solche mit höchstens einer Absenz.

Eine andere Wahl der Referenzkategorie würde zu anderen Effekten führen. ▲

In einer Zufallsstichprobe dienen die Mittelwertunterschiede als Schätzwerte für die entsprechenden Unterschiede in der Population. Im Prinzip kann für jeden Mittelwertunterschied mit einem Zweistichproben- t -Test die Arbeitshypothese geprüft werden, ob ein systematischer Unterschied vorliegt. Dies würde zeigen, welche Kategorien sich hinsichtlich mittlerem Y -Wert tatsächlich unterscheiden. Häufig verwendet man bei solchen *paarweisen t*-Tests wegen multiplem Testen ein entsprechend tieferes Signifikanzniveau.

Die Idee mit den Effekten lässt sich auch auf die Population anwenden: Bezeichnen wir wiederum mit x_1 die Referenzkategorie, so schätzen die $\hat{\beta}_j$ die tatsächlichen Effekte (Mittelwertunterschiede)

$$\beta_j := \mu_j - \mu_1$$

in der entsprechenden Grundgesamtheit. Umgekehrt kann der tatsächliche Mittelwert μ_j von Y bei Beobachtungen mit X -Wert x_j als

$$\mu_j = \mu_1 + \beta_j$$

ausgedrückt werden. Statt die L Teilpopulationsmittelwerte werden bei dieser Darstellung L andere Parameter betrachtet, nämlich der sogenannte *Intercept* μ_1 sowie die Effekte β_2, \dots, β_L . Viele Statistik-Softwares ergänzen den Output einer ANOVA um Schätzwerte für diese Parameter inkl. Student-Konfidenzintervalle (Zweistichprobenversion) und entsprechende t -Tests der Nullhypotesen¹ $\beta_2 = 0, \dots, \beta_L = 0$, um zu prüfen, welche X -Kategorien sich hinsichtlich mittlerem Y -Wert tatsächlich von der Referenzkategorie unterscheiden. Wie bei den angesprochenen paarweisen t -Tests gibt es allenfalls das Problem des multiplen Testens².

Im Zweistichprobenfall sind diese Verfahren identisch zu den üblichen Student-Verfahren.

Beispiel 5.19 (Leistung in Mathematik, Fortsetzung). Was können wir für Beispiel 5.18 auf dem 5%-Niveau (keine Korrektur für multiples Testen) über die Effekte in der Population sagen?³

R Code

```
# Eingabe
fit <- lm(math ~ daysabs, data = highschoo1)
summary(fit)

# Ausgabe
      Estimate Std. Error t value Pr(>|t|)
(Intercept)   52.92     1.68  31.45 < 2e-16 ***
daysabs(1,5]  -2.95     2.43  -1.21    0.23
daysabs(5,50] -9.40     2.37  -3.96 0.000093 ***

# Eingabe, Forts.
confint(fit)

# Ausgabe
      2.5 % 97.5 %
(Intercept) 49.609 56.2304
daysabs(1,5] -7.742 1.8324
daysabs(5,50] -14.073 -4.7306
```

Kommentare

- *Intercept* (μ_1): SchülerInnen mit 0 – 1 Absenzen stellen die Referenzkategorie⁴ dar. Wir schätzen, dass sie im Schnitt $\bar{Y}_1 = 52.92$ Punkte in Mathematik erreichen. Mit einer Sicherheit von rund 95% liegt der wahre Mittelwert μ_1 zwischen 49.6 und 56.2 Punkten. Der t -Test prüft die sinnlose Nullhypothese, dass SchülerInnen mit 0–1 Absenzen im Schnitt null Punkte in Mathematik erreichen, also dass $\mu_1 = 0$.
- *Effekt von 2 – 5 Absenzen* (β_2): Wir schätzen, dass SchülerInnen mit 2 bis 5 Absenztagen im Schnitt 2.95 Punkte weniger erreichen als solche, die höchstens einen Tag gefehlt haben. Mit einer Sicherheit von rund 95% liegt der wahre Effekt β_2 zwischen –7.74 und 1.83 Punkten. Der t -Test der Nullhypothese, dass dieser Effekt null ist, liefert einen p -Wert von 0.23. Somit können wir die Nullhypothese auf dem 5%-Niveau nicht verwirfen. Zum gleichen Testentscheid führt auch der Konfidenzintervallansatz (0 liegt im Konfidenzintervall).

¹Eine weitere mögliche Formulierung der Nullhypothese des F -Tests lautet übrigens $\beta_2 = \beta_3 = \dots = \beta_L = 0$, was jedoch nicht äquivalent zu den hier besprochenen t -Tests ist.

²Die t -Tests zu den Effekten sind (im Wesentlichen) ein Teil aller paarweisen t -Tests.

³Basierend auf dem Ergebnis der Varianzanalyse gibt die R-Funktion `confint` standardmäßig symmetrische 95%-Student-Konfidenzintervalle für die Parameter an. Deren Schätzwerte und t -Tests werden mit der `summary`-Funktion angezeigt.

⁴Da ‘0’ lexikographisch vor ‘1’ und ‘2’ kommt, wurde Ausprägung 0 von R automatisch als Referenzkategorie gewählt.

- *Effekt von > 5 Absenzen (β_3):* Wir schätzen, dass SchülerInnen mit mehr als fünf Absenztagen eine im Schnitt um 9.4 Punkte tiefere Leistung haben als solche mit höchstens einer Absenz. Mit einer Sicherheit von rund 95% liegt der wahre Effekt β_3 zwischen -14.1 und -4.7 Punkten. Der t -Test der Nullhypothese, dass dieser Effekt null ist, liefert einen p -Wert von 0.000093. Somit können wir auf dem 5%-Niveau diese Nullhypothese verwerfen und mit einer Sicherheit von rund 95% behaupten, dass ein wahrer mittlerer Unterschied besteht. Zum gleichen Testentscheid führt auch der Konfidenzintervallansatz (0 liegt nicht im Konfidenzintervall). ▲

Beispiel 5.20 (Körpergewicht und Geschlecht, Fortsetzung). Im Zweistichprobenfall liefern die hier besprochenen Student-Verfahren die gleichen Erkenntnisse wie die üblichen Zweistichprobenverfahren nach Students Methode. Dies zeigen wir hier am Beispiel 5.8.

R Code

```
# Eingabe
fit <- lm(Kgewicht ~ Geschlecht, data = wiso)
summary(fit)

# Ausgabe
[...]
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  70.245     0.614   114.4   <2e-16 ***
GeschlechtW -12.154     0.928   -13.1   <2e-16 ***

# Eingabe, Forts.
confint(fit)

# Ausgabe
      2.5 % 97.5 %
(Intercept) 69.04  71.45
GeschlechtW -13.98 -10.33
```

Kommentare

- *Intercept (μ_1):* Da ‘M’ lexikographisch vor ‘W’ kommt, wählt R die Männer automatisch als Referenzkategorie. Entsprechend schätzen wir das wahre mittlere Gewicht μ_1 der Männer auf $\bar{Y}_1 = 70.245$ kg. Mit einer Sicherheit von rund 95% liegt μ_1 zwischen 69.04 und 71.45 kg.
 - *Effekt der Frauen (β_2):* Wir schätzen, dass Frauen im Schnitt $-\hat{\beta}_2 = 12.154$ kg leichter sind als Männer. Mit einer Sicherheit von rund 95% liegt der echte mittlere Unterschied β_2 (mittleres Gewicht der Frauen in der Population minus jenes Männer) zwischen -13.98 und -10.33 kg, d. h. mit einer Sicherheit von rund 95% weisen Frauen tatsächlich ein um 10.33 bis 13.98 kg tieferes mittleres Gewicht als Männer auf. Der p -Wert des t -Tests beträgt fast null, somit können wir auf dem 5%-Niveau von einem echten Effekt sprechen. Das wahre mittlere Gewicht $\mu_2 = \mu_1 + \beta_2$ der Frauen schätzen wir auf $\bar{Y}_2 = \bar{Y}_1 + \hat{\beta}_2 = 70.245 - 12.154 = 58.091$ kg.
 - *Vergleiche:* Das Student-Konfidenzintervall für den wahren mittleren Unterschied ist bis auf das Vorzeichen gleich wie jenes in Beispiel 5.8. Der ausgewiesene p -Wert ist gleich wie jener des 2-Stichproben- t -Tests in Beispiel 5.8 (und damit auch gleich wie jener des F -Tests von Beispiel 5.14).
- Vorsicht: Beim üblichen t -Test muss man sich stets überlegen, ob Aussagen über den Parameter β_2 oder $-\beta_2$ gemacht werden (in Beispiel 5.8 ist es $-\beta_2$). Dies kann insbesondere bei einseitigen Tests zu falschen Testentscheiden führen. ▲

Beispiel 5.21 (Körpergewicht und Rauchen, Fortsetzung). Schliesslich wollen wir den Zusammenhang zwischen ‘Körpergewicht’ und ‘Rauchen’, den wir in den Beispielen 5.12 und 5.13 mit einer ANOVA untersucht haben, mithilfe der entsprechenden Effekte analysieren.

R Code

```
# Eingabe
fit <- lm(Kgewicht ~ Rauchen, data = wiso)
summary(fit)

# Ausgabe
[...]
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 65.027    0.739   87.99 <2e-16 ***
Rauchen1     -1.277    1.611   -0.79    0.43
Rauchen2      0.582    1.657    0.35    0.73

# Eingabe, Forts.
confint(fit)

# Ausgabe
          2.5 % 97.5 %
(Intercept) 63.57 66.483
Rauchen1     -4.45  1.895
Rauchen2     -2.68  3.845
```

Kommentare

- *Intercept (μ_1):* Der Schätzwert \bar{Y}_1 des Intercepts μ_1 beträgt 65 kg. Er entspricht dem mittleren Gewicht der NichtraucherInnen¹ (Ausprägung 0) in der Stichprobe. Mit einer Sicherheit von rund 95% liegt das wahre mittlere Körpergewicht μ_1 der NichtraucherInnen zwischen 63.6 und 66.5 kg. Der t -Test prüft die sinnlose Nullhypothese “NichtraucherInnen wiegen im Schnitt 0 kg”.
- *Effekt der GelegenheitsraucherInnen (β_2):* Der Schätzwert $\hat{\beta}_2$ von β_2 (wahres mittleres Gewicht aller GelegenheitsraucherInnen minus wahres mittleres Gewicht aller NichtraucherInnen) beträgt $63.750 - 65.027 = -1.277$ kg. Wir schätzen also, dass GelegenheitsraucherInnen im Schnitt 1.277 kg weniger wiegen als NichtraucherInnen. Mit einer Sicherheit von rund 95% liegt β_2 zwischen -4.45 und 1.90 kg. Da 0 im Konfidenzintervall liegt, gibt es keinen Grund, die Nullhypothese $\beta_2 = 0$ auf dem 5%-Niveau zu verwerfen. Entsprechend ist der p -Wert des t -Tests mit 0.43 nicht kleiner als 0.05.
- *Effekt der regelmässigen RaucherInnen (β_3):* Wir schätzen entsprechend, dass regelmässige RaucherInnen im Schnitt $\hat{\beta}_3 = 65.61 - 65.03 = 0.58$ kg mehr wiegen als NichtraucherInnen. Mit einer Sicherheit von rund 95% liegt der wahre mittlere Unterschied β_3 zwischen -2.68 und 3.85 kg. Auch hier liegt der Wert 0 im Konfidenzintervall, deshalb ist der p -Wert des t -Tests mit 0.73 nicht kleiner als 0.05 und wir können die Nullhypothese $\beta_3 = 0$ auf dem 5%-Niveau nicht verwerfen. ▲

5.4 Verbundene Stichproben

Manchmal wird eine numerische Information mehrmals von der gleichen Person bzw. Versuchseinheit erhoben, beispielsweise in einem zeitlichen Abstand (z. B. Jahreseinkommen in mehreren Jahren) oder unter veränderten Bedingungen (z. B. Preise, die jede Person für drei Versionen eines Produktes zahlen würde; “Mit–Ohne”-Fragen: Blutdruck mit/ohne Medikament; “Vorher–Nachher”-Fragen: IQ vor und nach Kurs).

¹Da ‘0’ lexikographisch vor ‘1’ und ‘2’ kommt, wurde Ausprägung 0 von R automatisch als Referenzkategorie gewählt.

Wie kann der Zusammenhang zwischen Zeitpunkt/Bedingung und dieser Information untersucht werden?
Wir starten unsere Überlegungen anhand zwei möglicher Strukturierungen des Datensatzes.

5.4.1 Strukturierung des Datensatzes

Lange Form

Die Information wird als numerisches Merkmal Y erfasst, der Zeitpunkt bzw. die Bedingung als kategorielles Merkmal X . Die Kategorien x_1, x_2, \dots, x_L von X repräsentieren die verschiedenen Zeitpunkte bzw. Bedingungen.

Beispiel 5.22 (Jahreseinkommen). Eine Ökonomin interessiert sich dafür, wie/ob sich das Jahreneinkommen von Familien über die Jahre verändert. Um den Zusammenhang zwischen Einkommen und Zeit zu untersuchen, fragt sie einige Familien nach ihrem Jahreneinkommen in den Jahren 2003, 2004 und 2005 und stellt die Daten folgendermassen dar: Die Variable Y gibt die Einkommen und X die entsprechenden Jahre an. Eine Identifikationsvariable ('Familie') ist sinnvoll, damit die Daten bei einer anderen Sortierung nicht durcheinandergeraten. Der Datensatz sieht schematisch folgendermassen aus.

Familie	X	Y
1	2003	82000
1	2004	84000
1	2005	84000
2	2003	75000
2	2004	76000
2	2005	77000
:	:	:



Diese Strukturierung macht deutlich, dass ein Zusammenhang zwischen einer *kategorialen* und einer *numerischen* Information untersucht werden soll. Die in diesem Kapitel gelernten *schliessenden* Verfahren sind jedoch nicht einsetzbar, da jede Person bzw. Versuchseinheit mehrere Beobachtungen liefert und deshalb keine unabhängigen Beobachtungen vorliegen.

Obwohl spezielle Verfahren der schliessenden Statistik zur Analyse benötigt werden, sind solche Daten erhebungen sehr sinnvoll: Zusammenhänge werden kaum durch Confounding verfälscht und Schätzwerte sind schon bei kleinen Stichproben ziemlich präzise, da Werte der gleichen Person oft deutlich weniger stark streuen als Werte von verschiedenen Personen.

Die Teilstichproben der Y -Werte mit gleichem X -Wert heissen *verbundene (gepaarte) Stichproben*, im Gegensatz zu den bisher behandelten *unverbundenen (unabhängigen) Stichproben*.

Breite Form

In der breiten Form werden die verbundenen Stichproben je als eigenes Merkmal abgespeichert. Pro Person bzw. Versuchseinheit fällt dann lediglich eine Beobachtung an.

Beispiel 5.23 (Jahreseinkommen, Fortsetzung). In unserem Beispiel sieht dies folgendermassen aus:

Familie	$Y^{(2003)}$	$Y^{(2004)}$	$Y^{(2005)}$
1	82000	84000	84000
2	75000	76000	77000
:	:	:	:



Für einfache Fragestellungen wie in diesem Abschnitt werden die Daten praktischerweise in dieser Form abgelegt. Es wird jedoch verschleiert, dass man sich eigentlich für den Zusammenhang zwischen einer numerischen und einer kategorialen Information interessiert.

5.4.2 Naheliegende Analysemöglichkeit im Zweistichprobenfall

Wird die Information zu zwei Zeitpunkten oder unter zwei Bedingungen als Merkmale $Y^{(1)}$ und $Y^{(2)}$ (breite Form) erfasst, lässt sich der Zusammenhang zwischen Information und Zeitpunkt/Bedingung mit den Differenzen $Z_i := Y_i^{(2)} - Y_i^{(1)}$ oder $Z'_i := Y_i^{(1)} - Y_i^{(2)}$ untersuchen. Das neue Merkmal, das pro Beobachtung die Veränderung angibt, kann dann mit den bereits bekannten deskriptiven und schliessenden *univariaten* Verfahren analysiert werden.

Beispiel 5.24 (Vorlesungen als Beruhigungsmittel). In einer Biometrievorlesung mit vielen Studierenden ermittelten $n = 18$ Personen ihre Pulsfrequenz zu Beginn ($Y^{(1)}$) und gegen Ende des Unterrichts ($Y^{(2)}$). Beide Werte sind die Anzahl von Pulsschlägen in einer Minute. Die Arbeitshypothese war, dass die $Y^{(1)}$ -Werte systematisch höher ausfallen würden als die $Y^{(2)}$ -Werte (dass also die Vorlesung beruhigend wirkt). Die Nullhypothese, dass dies nicht der Fall ist, möchten wir auf dem Niveau von $\alpha = 0.05$ testen.

R Code

```
# Dateneingabe
Y.1 <- c(66,78,54,76,80,94,68,64,76,80,64,66,70,80,82,102,74,90)
Y.2 <- c(66,78,56,78,78,90,74,70,70,74,72,58,62,72,72,92,62,78)
Z <- Y.1 - Y.2                                # Reduktion der Pulsfrequenz
Z

# Ausgabe
0 0 -2 -2 2 4 -6 -6 6 6 -8 8 8 8 10 10 12 12

# Eingabe, Forts.: Univariate Beschreibung der Reduktionen
summary(Z)

# Ausgabe
Min. 1st Qu. Median Mean 3rd Qu. Max.
-8.00 -1.50 5.00 3.44 8.00 12.00

# Eingabe, Forts.: Einstichproben-Student-Verfahren
t.test(Z, alternative = 'greater')

# Ausgabe
[...]
t = 2.2734, df = 17, p-value = 0.01813
95 percent confidence interval: 0.80874      Inf
sample estimates: mean of x
3.444444

# Eingabe, Forts.: Der wievielte kleinste Wert ist untere Konfidenzschranke für Median?
qbinom(0.05, 18, 0.5)                         # Ergibt 6

# Eingabe: Falls die beiden Stichproben unverbunden wären
t.test(Y.1, Y.2, alternative = 'greater')

# Ausgabe
t = 0.9612, df = 32.682, p-value = 0.1717
95 percent confidence interval: -2.621541      Inf
sample estimates:
mean of x mean of y
75.77778 72.33333
```

Kommentare

- *Zusammenhang in der Stichprobe:* Die univariate deskriptive Analyse zeigt, dass sich alle Reduktionen zwischen -8 und 12 Schlägen pro Minute bewegen (negative Werte bedeuten einen Anstieg). Rund die Hälfte der Personen hatte eine Reduktion zwischen -1.5 und 8 Schlägen pro Minute. Der Mittelwert der Reduktion beträgt 3.44 , der Median 5 Schläge pro Minute. Somit wirkt die Vorlesung zumindest in der Stichprobe beruhigend. (Ob dies viel ist, müsste ein Mediziner oder Sportwissenschaftler sagen.)
- *Zusammenhang in der Population:* Wir schätzen, dass der beruhigende Effekt $\mu = 3.44$ Schläge pro Minute beträgt. Mit einer Sicherheit von rund 95% ist μ grösser als 0.8 Schläge pro Minute. Anhand des Einstichproben- t -Tests können wir mit einer Sicherheit von rund 95% behaupten, dass die Vorlesung tatsächlich beruhigend wirkt, also dass $H_1 : \mu > 0$ gilt. In einer solchen Situation wird dieser Test auch *t -Test für verbundene Stichproben* genannt.
- *Präzision von Konfidenzschranke und p -Wert:* Da die Stichprobe klein ist, misstrauen wir den Student-Verfahren¹. Wir könnten die Fragestellung alternativ mit den pflegeleichten Verfahren aus Abschnitt 3.2.4 anpacken: Die untere 95% -Konfidenzschranke für den Median² beträgt 0 . Damit könnte also auf dem 5% -Niveau *nicht* behauptet werden, dass die Vorlesung beruhigend wirkt.
- *Unverbundene Stichproben:* Wären die beiden Stichproben übrigens unverbunden (lauter verschiedene Personen untersucht), ergäbe sich zwar die gleiche mittlere Reduktion von $75.78 - 72.33 = 3.44$, dieser Schätzwert wäre jedoch viel unpräziser (untere Konfidenzschranke viel weiter vom Schätzwert entfernt).

▲

Hinweis (Relative Veränderungen). In vielen Situationen ist es sinnvoll, statt absolute Veränderungen *relative* zu betrachten, also das Merkmal $(Y^{(2)} - Y^{(1)})/Y^{(1)}$ bzw. $(Y^{(1)} - Y^{(2)})/Y^{(1)}$.

Beispiel 5.25 (Vorlesungen als Beruhigungsmittel, Fortsetzung). Was würde sich in Beispiel 5.24 ergeben, wenn wir statt absolute Reduktionen relative analysieren würden?

R Code

```
# Eingabe, Forts.
Relative.reduktion <- Z/Y.1
Relative.reduktion

# Ausgabe
0.00  0.00 -0.04 -0.03  0.02  0.04 -0.09 -0.09  0.08  0.08 -0.12  0.12
0.11  0.10  0.12  0.10  0.16  0.13

# Eingabe, Forts.: Univariate Beschreibung
summary(Relative.reduktion)

# Ausgabe
Min. 1st Qu. Median Mean 3rd Qu. Max.
-0.12500 -0.01974 0.05878 0.03901 0.11070 0.16220

# Eingabe, Forts.: Einstichproben-Student-Verfahren
t.test(Relative.reduktion, alternative = 'greater')

# Ausgabe
[...]
t = 1.9112, df = 17, p-value = 0.0365
95 percent confidence interval: 0.0035 Inf
sample estimates: 0.03900812
```

¹Zur Erinnerung: Die Einstichproben-Studentverfahren funktionieren bei kleinen Stichproben nur für ungefähr normalverteilte Werte. Dies ist hier zumindest fraglich, da der Unterschied zwischen Mittelwert und Median auf eine schiefe Verteilung hinweist.

²Das 5% -Quantil der Binomialverteilung mit $n = 18$ und $p = 0.5$ beträgt 6 . Die sechstkleinste Reduktion ist 0 .

Kommentare

- *Zusammenhang in der Stichprobe:* Die univariate deskriptive Analyse zeigt, dass sich alle relativen Reduktionen zwischen -12.5% und 16.2% bewegen (negative Werte bedeuten einen Anstieg). Rund die Hälfte der Personen hatte eine Reduktion zwischen -2.0% und 11.1% . Die mittlere relative Reduktion beträgt 3.9% , die mediane 5.9% . Die Vorlesung wirkt also zumindest in der Stichprobe beruhigend.
- *Zusammenhang in der Population:* Wir schätzen, dass der beruhigende relative Effekt $\mu' = 3.9\%$ beträgt. Mit einer Sicherheit von rund 95% ist μ' grösser als 0.4% . Anhand des t -Tests können wir mit einer Sicherheit von rund 95% behaupten, dass die Vorlesung tatsächlich beruhigend wirkt, also dass $H_1 : \mu' > 0$ gilt. Auch hier gelten die Feststellungen des letzten Beispiels hinsichtlich Präzision. Die verlässlichere untere $95\%-Konfidenzschranke$ für die wahre mediane relative Reduktion beträgt 0 (sechstkleinste relative Veränderung), so dass man auf dem $5\%-Niveau$ nicht von einer systematischen Reduktion der Pulsfrequenz sprechen könnte. ▲

Hinweis (Reduktion auf bivariate Situation). Mit analogem Vorgehen lässt sich der Zusammenhang zwischen drei Merkmalen manchmal auf zwei reduzieren, die dann mit den entsprechenden bivariaten Verfahren analysiert werden können.

Beispiel 5.26 (Numerus-Clausus). Sie möchten den armen angehenden MedizinstudentInnen einen extrem teuren Kurs anbieten, der diese auf die Numerus-Clausus-Prüfung vorbereiten soll. Fairerweise prüfen Sie zuerst mit einem Experiment, ob der Kurs sein Geld wert ist. Dazu lassen Sie 100 zufällig ausgewählte Interessierte je im Abstand von zwei Monaten eine Musterprüfung schreiben. Zwischen den beiden Versuchen übt die eine Hälfte selbstständig (Kontrollgruppe), die andere besucht Ihren Kurs (Testgruppe)¹.

Der Datensatz in *langer Form* enthält folgende Merkmale:

id	Personennummer
kurs	Kurs besucht (ja = 1, nein = 0)
time	Zeitpunkt (0, 2)
score	Erreichte Punktzahl (0 bis 200)

Seine ersten Zeilen sehen folgendermassen aus:

R Code			
id	kurs	time	score
1	0	0	55
1	0	2	68
2	0	0	65
2	0	2	88

Um die eigentliche Fragestellung (“Lohnt sich der Kurs”) zu beantworten, muss der Zusammenhang zwischen den drei Merkmalen ‘kurs’, ‘time’ und ‘score’ untersucht werden. Dabei können die jeweils zwei Beobachtungen der gleichen Person nicht als unabhängig aufgefasst werden.

Um diese *multivariate Situation mit Abhängigkeiten* auf eine bivariate ohne Abhängigkeiten zurückzuführen, betrachten wir die Daten nun in *breiter Form* und kreieren mit den beiden Merkmalen ‘score.0’ (Score vor Training) und ‘score.2’ (Score nach zwei Monaten Training) ein neues Merkmal ‘verbesserung’ (‘score.2’ minus ‘score.0’). Dieses repräsentiert den Zusammenhang zwischen ‘time’ und ‘score’.

Ein kleiner Datenausschnitt sieht folgendermassen aus:

¹Der Einfachheit halber gehen wir davon aus, dass niemand frühzeitig aussteigt (keine “drop-outs”).

R Code																															
# Eingabe																															
head(nc)																															
# Ausgabe																															
<table border="1"> <thead> <tr> <th>kurs</th> <th>score.0</th> <th>score.2</th> <th>verbesserung</th> </tr> </thead> <tbody> <tr><td>0</td><td>55</td><td>68</td><td>13</td></tr> <tr><td>0</td><td>65</td><td>88</td><td>23</td></tr> <tr><td>0</td><td>88</td><td>118</td><td>30</td></tr> <tr><td>0</td><td>73</td><td>77</td><td>4</td></tr> <tr><td>0</td><td>58</td><td>48</td><td>-10</td></tr> <tr><td>0</td><td>131</td><td>151</td><td>20</td></tr> </tbody> </table>				kurs	score.0	score.2	verbesserung	0	55	68	13	0	65	88	23	0	88	118	30	0	73	77	4	0	58	48	-10	0	131	151	20
kurs	score.0	score.2	verbesserung																												
0	55	68	13																												
0	65	88	23																												
0	88	118	30																												
0	73	77	4																												
0	58	48	-10																												
0	131	151	20																												

Bevor wir nun die Fragestellung anpacken, beschreiben wir die Merkmale univariat. Dabei stellen wir fest, dass sich die Personen während den zwei Monaten im Schnitt um 9.38 Punkte verbessert haben:

R Code																																																							
# Eingabe																																																							
summary(nc)																																																							
# Ausgabe																																																							
<table border="1"> <thead> <tr> <th>kurs</th> <th>score.0</th> <th>score.2</th> <th>verbesserung</th> </tr> </thead> <tbody> <tr><td>Min.</td><td>0.0</td><td>53.00</td><td>48.0</td></tr> <tr><td>1st Qu.</td><td>0.0</td><td>78.50</td><td>85.5</td></tr> <tr><td>Median</td><td>0.5</td><td>92.00</td><td>104.5</td></tr> <tr><td>Mean</td><td>0.5</td><td>92.33</td><td>101.7</td></tr> <tr><td>3rd Qu.</td><td>1.0</td><td>103.75</td><td>117.0</td></tr> <tr><td>Max.</td><td>1.0</td><td>142.00</td><td>156.0</td></tr> <tr><td></td><td></td><td></td><td>-20.00</td></tr> <tr><td></td><td></td><td></td><td>1.75</td></tr> <tr><td></td><td></td><td></td><td>9.50</td></tr> <tr><td></td><td></td><td></td><td>9.38</td></tr> <tr><td></td><td></td><td></td><td>17.25</td></tr> <tr><td></td><td></td><td></td><td>31.00</td></tr> </tbody> </table>				kurs	score.0	score.2	verbesserung	Min.	0.0	53.00	48.0	1st Qu.	0.0	78.50	85.5	Median	0.5	92.00	104.5	Mean	0.5	92.33	101.7	3rd Qu.	1.0	103.75	117.0	Max.	1.0	142.00	156.0				-20.00				1.75				9.50				9.38				17.25				31.00
kurs	score.0	score.2	verbesserung																																																				
Min.	0.0	53.00	48.0																																																				
1st Qu.	0.0	78.50	85.5																																																				
Median	0.5	92.00	104.5																																																				
Mean	0.5	92.33	101.7																																																				
3rd Qu.	1.0	103.75	117.0																																																				
Max.	1.0	142.00	156.0																																																				
			-20.00																																																				
			1.75																																																				
			9.50																																																				
			9.38																																																				
			17.25																																																				
			31.00																																																				

Damit sich der Kurs lohnt, müssen nun zwei Fragen mit ‘ja’ beantwortet werden:

- *Frage 1:* Steigen die Leistungen in der Gruppe mit Kurs im Schnitt an?
- *Frage 2:* Ist dieser Anstieg im Schnitt höher als in der Kontrollgruppe?

Um eine allgemeine Aussage über den Nutzen des Kurses zu machen, sind also insbesondere zwei Tests bzw. zwei Konfidenzintervalle nötig. Um den gesamten Fehler erster Art auf 5% zu beschränken, verwenden wir je das Signifikanzniveau von 0.025 bzw. das Konfidenzniveau 0.975 (Bonferroni-Korrektur bei multiplen Tests bzw. multiplen Konfidenzintervallen).

Hier die nötigen Analysen zu Frage 1:

R Code															
# Eingabe: Univariate Beschreibung von ‘verbesserung’ bei Personen mit Kurs															
summary(nc\$verbesserung[nc\$kurs == 1])															
# Ausgabe															
<table border="1"> <thead> <tr> <th>Min.</th> <th>1st Qu.</th> <th>Median</th> <th>Mean</th> <th>3rd Qu.</th> <th>Max.</th> </tr> </thead> <tbody> <tr><td>-11.00</td><td>2.25</td><td>9.00</td><td>8.48</td><td>13.00</td><td>31.00</td></tr> </tbody> </table>				Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	-11.00	2.25	9.00	8.48	13.00	31.00
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.										
-11.00	2.25	9.00	8.48	13.00	31.00										
# Eingabe, Forts.: Untere Student-Konfidenzschranke (Einstichprobenfall)															
t.test(nc\$verbesserung[nc\$kurs==1], alternative = "greater", conf.level = 0.975)															
# Ausgabe															
<table border="1"> <thead> <tr> <th>[...]</th> </tr> </thead> <tbody> <tr><td>sample estimate: 8.48</td></tr> <tr><td>97.5 percent c.i.: 5.730454 Inf</td></tr> </tbody> </table>				[...]	sample estimate: 8.48	97.5 percent c.i.: 5.730454 Inf									
[...]															
sample estimate: 8.48															
97.5 percent c.i.: 5.730454 Inf															

Kommentare: Bei den Personen mit Kurs ist eine deutliche mittlere Verbesserung um 8.48 Punkte sichtbar. Mit einer Sicherheit von rund 97.5% beträgt diese in Wirklichkeit mindestens 5.73 Punkte (untere 97.5%-Konfidenzschranke nach Students Einstichproben-Verfahren). Da die Berechnungen auf immerhin 50 Personen beruhen, trauen wir der Student-Konfidenzschranke.

Schliesslich die relevanten Analysen zu Frage 2. Dazu analysern wir den Effekt zur Referenzkategorie (kein Kurs) via ANOVA, um diese Konstruktion nochmals zu illustrieren.

Eine untere 97.5%-Konfidenzschranke für die wahre mittlere Überlegenheit β des Kurses (wahre mittlere Verbesserung in der Gruppe mit Kurs minus jene in der Gruppe ohne) finden wir via zweiseitigem, symmetrischem 95%-Konfidenzintervall, welches von den meisten Softwares standardmässig angegeben wird.

R Code

```
# Eingabe: Boxplots und Mittelwertvergleich
boxplot(verbesserung~kurs, data = nc)
summary(verbesserung~kurs, data = nc) # library(Hmisc)

# Ausgabe
+-----+---+---+-----+
|       | N  |verbesserung|
+-----+---+---+-----+
|kurs   |No  | 50| 10.28   |
|       |Yes | 50| 8.48    |
+-----+---+---+-----+
|Overall| 100| 9.38    |
+-----+---+---+-----+

# Eingabe: Analyse zum Effekt von Kurs
fit <- lm(verbesserung~kurs, data = nc)
summary(fit)

# Ausgabe
      Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.280     1.492   6.890 5.4e-10 ***
kurs        -1.800     2.110  -0.853   0.396
[...]
Multiple R-squared: 0.007372, Adjusted R-squared: -0.002757
F-statistic: 0.7278 on 1 and 98 DF, p-value: 0.3957

# Eingabe
confint(fit)

# Ausgabe
      2.5 %    97.5 %
(Intercept) 7.319331 13.240669
kurs        -5.987018  2.387018
```

Kommentare

- *Aussagen über Stichprobe:* Bei Personen mit Kurs beträgt die mittlere Verbesserung 8.48 Punkte (siehe Frage 1). Dies sind 1.8 Punkte weniger als bei den Personen ohne Kurs. Auch der Boxplot (Abbildung 5.7) zeigt, dass sich Personen mit Kurs tendenziell weniger stark verbessern als jene ohne.
- *Aussagen über Population auf 2.5%-Niveau:* Wir schätzen β auf -1.8 . Eine untere 97.5%-Konfidenzschranke für β ist gegeben durch den Wert -5.98 . Da dieser Wert nicht grösser als null ist, können wir auf dem 2.5%-Niveau nicht behaupten, dass der Kurs im Schnitt tatsächlich besser wirkt als das individuelle Training¹.

¹Der bei 'kurs' ausgewiesene p -Wert gehört zum *zweiseitigen* Test, der uns hier nicht interessiert.

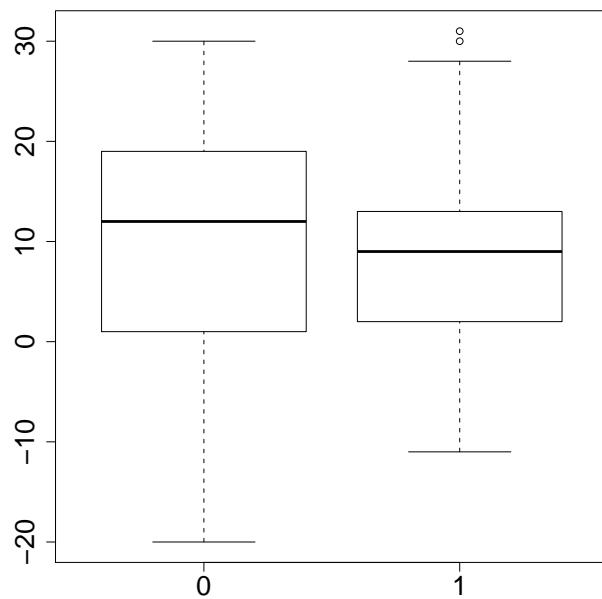


Abbildung 5.7: Boxplots von ‘verbesserung’ in Abhängigkeit von ‘kurs’ (Beispiel 5.26).

- *Präzision:* Die Teilstichproben sind mit je 50 Personen recht gross. Da die Boxplots in Abbildung 5.7 jedoch deutlich verschiedene Streuungen zwischen den Teilstichproben zeigen, trauen wir dem ausgewiesenen Student-Konfidenzintervall nur bedingt.
- *Bemerkungen zur schliessenden Statistik:* Da der Kurs nicht einmal in der Stichprobe besser als das individuelle Training abschneidet, können wir dies sicher auch nicht für die Population behaupten. Deshalb ist der Test hier im Prinzip überflüssig.

Fazit: Personen mit Kurs verbessern sich zwar im Schnitt tatsächlich. Man kann jedoch nicht behaupten, dass sie sich systematisch deutlicher verbessern als Personen, die sich individuell vorbereiten. Der Kurs ist sein Geld damit nicht wert. ▲

5.4.3 Weitere Tests bei verbundenen Stichproben

- *McNemar-Test:* Bei binärer Information zu zwei Zeitpunkten
- *Chiadrat-Test auf Symmetrie:* Bei kategorialer Information zu zwei Zeitpunkten (Verallgemeinerung des McNemar-Tests auf mehrere Kategorien)
- *Wilcoxon's Signed-Rank-Test:* Bei numerischer Information zu zwei Zeitpunkten (rangbasiert)
- *Varianzanalyse mit Messwiederholungen:* Bei numerischer Information zu mehreren Zeitpunkten
- *Cochrancs Q-Test:* Bei binärer Information zu mehreren Zeitpunkten (Verallgemeinerung des McNemar-Tests auf mehrere Zeitpunkte)
- *Friedman-Test, Quade-Test, Page-Test:* Bei ordinaler/numerischer Information zu mehreren Zeitpunkten (rangbasiert)

5.5 Zusammenfassung

- Der Zusammenhang zwischen einem numerischen Merkmal Y und einem kategorialen Merkmal X wird untersucht, indem die univariaten Verteilungen der Y -Werte pro Kategorie von X verglichen werden. Unterschiede charakterisieren den Zusammenhang.
- Die Verteilungen der Y -Werte pro Kategorie von X können grafisch beispielsweise mit ECDFs, Stripcharts oder Boxplots verglichen werden, quantitativ unter anderem mithilfe von Mittelwerten, Standardabweichungen und Quantilen. Häufig konzentriert man sich bei quantitativen Vergleichen auf Mittelwerte (oder Mediane), d. h. man sucht explizit nach Zusammenhängen in Form von Lageunterschieden.
- Auf dieser Vereinfachung beruhen Aussagen über den Zusammenhang in der Population. Dabei unterscheidet man den Zweistichprobenfall (binäres X) und den Mehrstichprobenfall (X mit mehreren Kategorien). Die Verfahren bei letzterem werden unter dem Begriff Varianzanalyse oder ANOVA zusammengefasst.
- Im Zweistichprobenfall wird der Zusammenhang meist durch den einzigen Mittelwertunterschied charakterisiert. Für dessen wahren Wert sind Student-Konfidenzintervalle und t -Tests möglich. Ein mittlerer Unterschied in der Population bedeutet ein echter Zusammenhang zwischen X und Y .
- Im Mehrstichprobenfall wird der Zusammenhang durch mehrere Mittelwerte und deren Unterschiede charakterisiert. Die Stärke des Zusammenhangs wird mit dem Bestimmtheitsmass R^2 quantifiziert, für dessen Wert in der Population Konfidenzintervalle verfügbar sind. Ist der wahre Wert grösser als 0, so liegt ein echter Zusammenhang vor. Hypothesen dieser Art können entweder mit dem Konfidenzintervallansatz oder dem F -Test geprüft werden. Im Zweistichprobenfall liefert der F -Test die gleichen p -Werte wie der t -Test auf Zusammenhang.
- Oft möchte man im Mehrstichprobenfall wissen, welche Kategorien von X sich wirklich unterscheiden. Statt alle möglichen Kategorien hinsichtlich mittlerem Y -Wert mit Zweistichproben- t -Tests zu vergleichen, ist es manchmal sinnvoll, mittlere Unterschiede zu einer wichtigen Referenzkategorie zu studieren und Student-Konfidenzintervalle (und -Tests) für diese Effekte auszuweisen.
- Die genannten Verfahren der schliessenden Statistik liefern vertrauenswürdige bzw. präzise Ergebnisse, falls die Teilstichproben je gross und die Standardabweichungen dazwischen ähnlich sind. Ist dies nicht gegeben oder liegen ordinale Y -Werte vor, so sollte mit den rangbasierten Pendants gearbeitet werden – dem Rangsummentest nach Wilcoxon (2-Stichprobenfall) oder dem Rangsummentest nach Kruskal und Wallis (Mehrstichprobenfall).
- Schliesslich haben wir Situationen betrachtet, bei denen eine numerische Information zu mehreren Zeitpunkten/Bedingungen bei je den gleichen Personen gemessen wird. Die üblichen Verfahren der schliessenden bivariaten Statistik können bei solchen verbundenen Stichproben generell nicht eingesetzt werden, da die Beobachtungen innerhalb der Personen nicht unabhängig sind. Gibt es lediglich zwei Zeitpunkte/Bedingungen, so lässt sich der Zusammenhang zwischen Zeitpunkt/Bedingung und numerischer Information durch Differenzenbildung mit univariaten Methoden analysieren.

Kapitel 6

Zwei numerische Merkmale

Nun betrachten wir den verbleibenden Fall zweier numerischer Variablen. Die gemeinsame Verteilung der X - und Y -Werte wird mit dem *Streudiagramm* visualisiert. Ein allfälliger Zusammenhang zwischen den X - und Y -Werten ist darin gut ersichtlich. Ein solcher liegt dann vor, wenn die (auf X bedingten) Verteilungen der Y -Werte vom X -Wert abhängen (oder umgekehrt), also wenn die Punkte um eine (nicht konstante) Funktion streuen.

Es ist wünschenswert, solche vage Beurteilungen mit objektiven Kenngrößen zu präzisieren: Die *lineare Regression* versucht, die Y -Werte durch eine lineare Funktion der X -Werte zu approximieren. Dazu wird eine Gerade in das Streudiagramm eingezeichnet, welche die Punkte so gut wie möglich repräsentiert. Ihr y -Achsenabschnitt und ihre Steigung quantifizieren die Form des linearen Zusammenhangs. In diesem Kontext sind *Bestimmtheitsmaß*, *Korrelationen* und *Kovarianz* Kenngrößen, die quantifizieren, wie gut diese Approximation funktioniert, also wie stark der entsprechende Zusammenhang ist.

6.1 Streudiagramm

Das Streudiagramm entsteht, indem man jedes Datenpaar (X_i, Y_i) als Punkt in einem zweidimensionalen Koordinatensystem einzeichnet. Abbildung 6.1 zeigt Streudiagramme von vier verschiedenen Datensätzen.

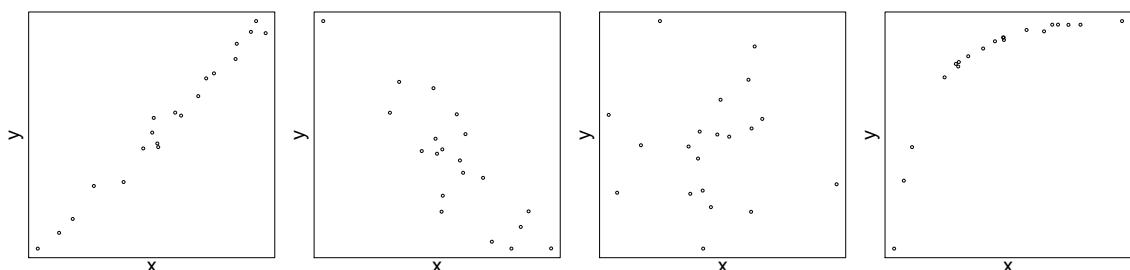


Abbildung 6.1: Vier Streudiagramme von je 20 Beobachtungen. Im Bild ganz links erkennt man einen starken positiven linearen Zusammenhang zwischen den X - und Y -Werten. Das heisst, die Punktpaare liegen in etwa auf einer Geraden mit positiver Steigung. Im zweiten Bild von links sieht man ebenfalls einen linearen Zusammenhang, allerdings etwas schwächer und negativ. Im dritten Bild von links erkennt man keinerlei Zusammenhang zwischen beiden Variablen (Verteilung der Y -Werte hängt nicht vom X -Wert ab bzw. Punkte streuen um eine konstante Funktion, also um eine horizontale Gerade), während das Bild ganz rechts einen starken positiven *nichtlinearen* Zusammenhang zeigt.

Beispiel 6.1 (Körpergrösse und Körpergewicht). Die linke Hälfte von Abbildung 6.2 zeigt das Streudiagramm¹ der Merkmale ‘Körpergrösse’ und ‘Körpergewicht’ im Datensatz von Beispiel 1.1. Dabei ist ein relativ starker positiver linearer Zusammenhang erkennbar.

R Code

```
# Eingabe: Streudiagramm
plot(Kgewicht ~ Kgroesse, data = wiso)

# Eingabe: Streudiagramm-Matrix
plot(wiso[, c('MonMiete', 'Kgroesse', 'Kgewicht')])
```

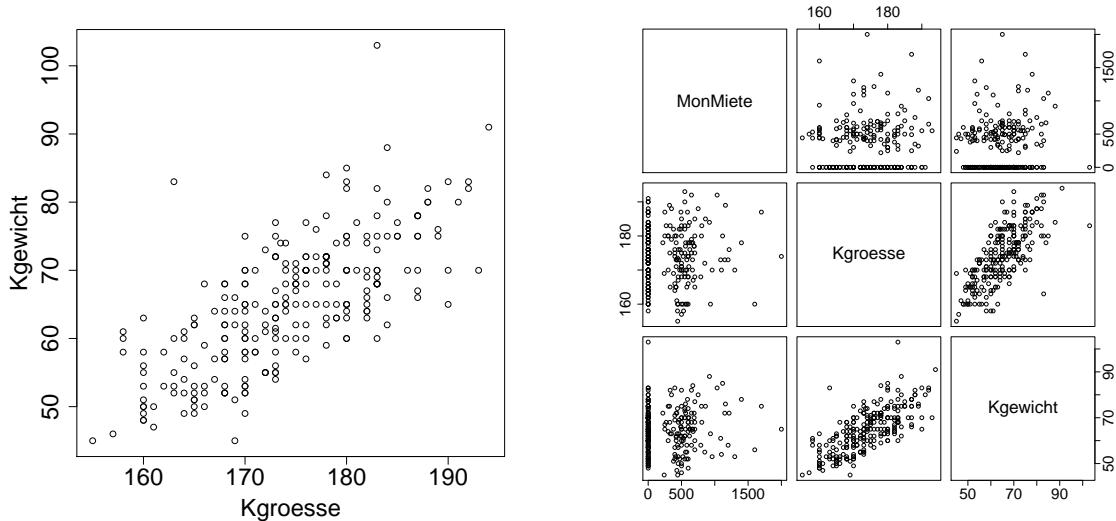


Abbildung 6.2: Linke Hälfte: Streudiagramm von ‘Körpergrösse’ und ‘Körpergewicht’. Rechte Hälfte: Streudiagramm-Matrix dreier Variablen in Beispiel 6.1.

Die rechte Hälfte von Abbildung 6.2 zeigt die sogenannte *Streudiagramm-Matrix* der drei Merkmale ‘Körpergrösse’, ‘Körpergewicht’ und ‘Monatsmiete’, also alle paarweisen Streudiagramme dieser Merkmale. Ein deutlicher Zusammenhang ist nur zwischen ‘Körpergewicht’ und ‘Körpergrösse’ zu erkennen. ▲

6.2 Lineare Regression

6.2.1 Regressionskoeffizienten und Regressionsgerade

Wir möchten untersuchen, inwiefern ein linearer Zusammenhang zwischen den X - und Y -Werten besteht. Dazu versuchen wir, die Y -Werte möglichst gut durch eine lineare Funktion² $f(x) = a + bx$ der X -Werte zu approximieren. (Ebenso könnte man versuchen, die X -Werte durch eine lineare Funktion der Y -Werte zu approximieren.) Genauer gesagt suchen wir zwei Zahlen a und b , so dass die Werte Y_i möglichst gut mit den Werten $f(X_i) = a + bX_i$ übereinstimmen in dem Sinne, dass die Quadratsumme

$$Q(a, b) := \sum_{i=1}^n (Y_i - (a + bX_i))^2$$

¹In R werden Streudiagramme und Streudiagramm-Matrizen mit der `plot`-Funktion erstellt.

²Der Koeffizient a ist der y -Achsenabschnitt (intercept), b ist die Steigung.

möglichst klein wird. Dies ist eine Anwendung der *Kleinste-Quadrate-Methode*, ein wichtiges Optimierungsverfahren. Die optimalen Werte (die *Kleinste-Quadrate-Lösungen*) von b und a sind durch die *Regressionskoeffizienten*

$$\hat{\beta} := \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \quad \text{und}$$

$$\hat{\alpha} := \bar{Y} - \hat{\beta}\bar{X},$$

gegeben¹. Dabei ist

$$\text{Cov}(X, Y) := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

die sogenannte *Stichprobenkovarianz* von X und Y . Auf diese wichtige Hilfsgrösse werden wir später genauer eingehen. Paarweise Kovarianzen von mehreren Variablen werden manchmal analog zur Streudiagramm-Matrix als *Varianz-Kovarianzmatrix* (kurz VC-Matrix) angeordnet².

Die Regressionskoeffizienten definieren die *Regressionsgerade*, also die optimale Gerade durch das Streudiagramm.

Aus statistischer Sicht gibt der Intercept $\hat{\alpha}$ den (interpolierten) Mittelwert von Y bei Beobachtungen mit X -Wert 0 an. Die Steigung $\hat{\beta}$ interpretieren wir folgendermassen: Erhöht sich der X -Wert um eins, so steigt der *mittlere* Y -Wert um $\hat{\beta}$. Man spricht auch vom *linearen Effekt von X auf den Mittelwert von Y*.

Beispiel 6.2 (Körpergrösse und Körpergewicht, Fortsetzung). Wir möchten nun für Beispiel 6.1 die Regressionsgerade berechnen, in das Streudiagramm einzeichnen³ und die Regressionskoeffizienten beurteilen.

R Code

```
# Eingabe: Lineare Regression
fit <- lm(Kgewicht ~ Kgroesse, data = wiso)
fit

# Ausgabe
Coefficients:
(Intercept)      Kgroesse
-80.470          0.835

# Eingabe, Forts.: Streudiagramm inkl. Regressionsgerade
plot(Kgewicht ~ Kgroesse, data = wiso)
abline(fit)
```

Kommentare: Abbildung 6.3 zeigt das Streudiagramm inkl. Regressionsgerade mit y-Achsenabschnitt $\hat{\alpha} = -80.5$ kg und Steigung $\hat{\beta} = 0.84$. Das Gewicht steigt pro cm Körpergrösse im Schnitt also um 0.84 kg an (und eine 0 cm grosse Person wiegt im Schnitt -80.5 kg).

Um diese Zahlen von Hand zu berechnen, benötigen wir die Mittelwerte von ‘Körpergrösse’ (174.25 cm) und ‘Körpergewicht’ (64.95 kg), die Varianz von ‘Körpergrösse’ (66.448 cm^2) sowie die gemeinsame Kovarianz ($55.454 \text{ cm} \cdot \text{kg}$). Daraus ergibt sich

$$\hat{\beta} = \frac{55.454}{66.448} = 0.83455$$

und

$$\hat{\alpha} = 64.95 - 0.83455 \cdot 174.25 \approx -80.47.$$

¹Man gelangt zu ihnen, indem man die Funktion Q einmal nach a und einmal nach b ableitet. Diese beiden Ableitungen setzt man null und löst das Gleichungssystem nach a und b auf.

²Aufgrund der Definition entspricht die Kovarianz zwischen einem Merkmal und sich selbst der Varianz. Dies sind die Diagonaleinträge in der VC-Matrix von oben links nach unten rechts.

³Wie bei der Varianzanalyse ist die R-Funktion `lm` Ausgangslage für diese Berechnungen. Das Streudiagramm wird mit der Funktion `abline` (von ‘a’-‘b’-Linie) um die geschätzte Regressionsgerade erweitert.

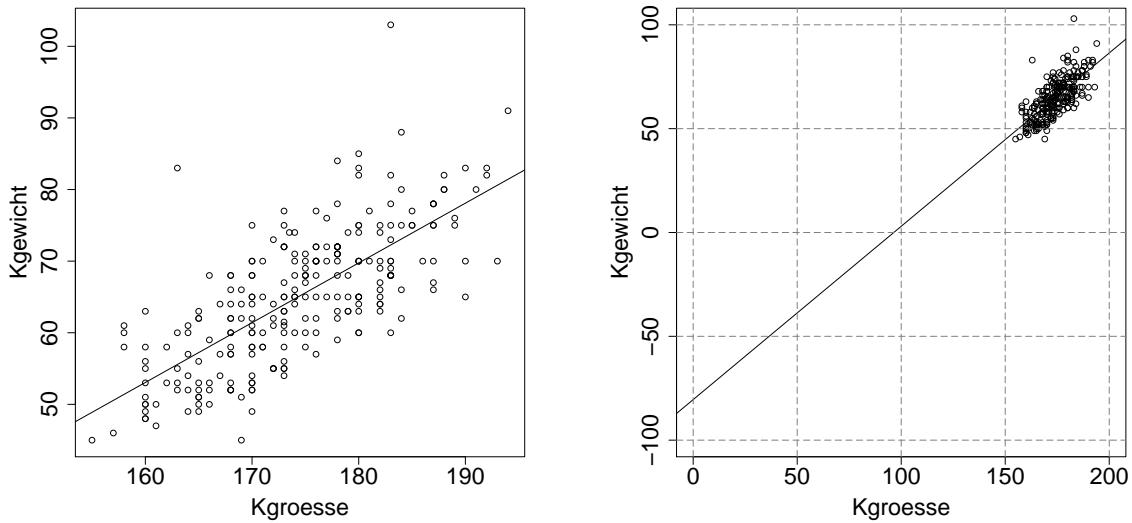


Abbildung 6.3: Zweimal das Streudiagramm von ‘Körpergrösse’ und ‘Körpergewicht’ inkl. Regressionsgerade in Beispiel 6.2. Im Bild rechts wurden die Skalen so gewählt, dass man den y-Achsenabschnitt ablesen kann.

Hinweis: Die benötigte Varianz und Kovarianz wären bspw. aus einer VC-Matrix¹ mit (mindestens) ‘Körpergrösse’ und ‘Körpergewicht’ ersichtlich:

R Code

```
# Eingabe
var(wiso[, c('MonMiete', 'Kgroesse', 'Kgewicht')], use = 'pair')
      MonMiete   Kgroesse   Kgewicht
MonMiete 134762.3159 111.57945 244.62399
Kgroesse  111.5795  66.39982 55.45432
Kgewicht  244.6240  55.45432 89.46230
```



Bevor wir lernen, mit der linearen Regression Vorhersagen zu machen, erwähnen wir ein Problem der Methode der kleinsten Quadrate – den *Leverage-Effekt*.

6.2.2 Leverage-Effekt

Ein extremer Ausreisser in X hat einen unerwünscht grossen Einfluss (“Leverage” bzw. “Hebelwirkung”) auf die Berechnung der Regressionsgerade. Je nach Y -Wert zieht diese Beobachtung die Regressionsgerade stark an und beeinflusst die Regressionskoeffizienten massiv. Abbildung 6.4 illustriert diesen gefährlichen *Leverage-Effekt*.

Extreme Ausreisser in X und damit Beobachtungen mit hoher Leverage werden idealerweise bereits bei der univariaten Analyse identifiziert (Punkte weit ausserhalb der Whiskers eines Boxplots?) und z. B. durch Abschneiden auf hohem Niveau (bspw. beim 99%-Quantil) entschärft.

Beispiel 6.3 (Körpergrösse und Körpergewicht, Fortsetzung). Die univariate Analyse von ‘Körpergrösse’ zeigt keine starken Ausreisser und damit keine besonders einflussreichen Beobachtungen. Alle Werte befinden sich nämlich innerhalb der Whiskers eines “gedachten” Boxplots. Deshalb erwarten wir für Beispiel 6.2 keinen deutlichen Leverage-Effekt.

¹Solche werden mit der R-Funktion var berechnet. Damit die Berechnungen nicht an fehlenden Werten scheitern, setzen wir die Option use = ‘pair’. (Die Varianz des Gewichts weicht deshalb minim von jener ab, die wir oben verwendet haben.)

R Code							
# Eingabe							
summary(wiso\$Kgroesse)							

# Ausgabe							
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's	
155	168	174	174	180	194	5	

▲

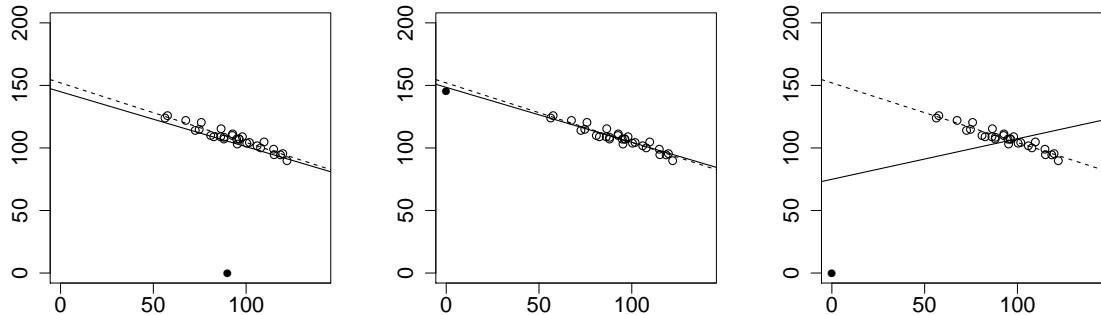


Abbildung 6.4: Illustration des Leverage-Effekts: Die drei Streudiagramme unterscheiden sich lediglich in der Position der fett markierten Beobachtung, die ein Ausreißer in X und/oder Y hat. Linkes Bild: Trotz Ausreißer in Y unterscheidet sich die Regressionsgerade inkl. Beobachtung (durchgezogen) kaum von jener ohne (gestrichelt), da kein Ausreißer in X vorliegt. Die markierte Beobachtung weist keine Hebelwirkung (Leverage) auf. Andere Bilder: Es liegt ein Ausreißer in X vor, deshalb hat die hervorgehobene Beobachtung eine hohe Leverage und beeinflusst somit die Berechnungen stark. Aufgrund der Wahl des entsprechenden Y -Werts tritt ein starker Leverage-Effekt jedoch nur im rechten Bild auf: Die beiden Regressionsgeraden unterscheiden sich massiv.

6.2.3 Vorhersagen

Eine Anwendung der Regression sind *Vorhersagen* analog zur Varianzanalyse. Für eine zukünftige Beobachtung, von der lediglich der X -Wert x bekannt ist, würde man schätzen, dass der entsprechende Y -Wert

$$\hat{\mu}(x) := \hat{\alpha} + \hat{\beta}x$$

beträgt, also den entsprechenden Wert auf der Regressionsgeraden (bzw. den erwarteten Mittelwert von Y bei X -Wert x) annimmt. Man hofft, dass die Vorhersage anhand des *linearen Prädiktors* $\hat{\mu}$ besser ist als der einfache Tipp \bar{Y} .

Da $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$ ist, beträgt die Vorhersage von Y beim Mittelwert \bar{X} übrigens

$$\hat{\mu}(\bar{X}) = \hat{\alpha} + \hat{\beta}\bar{X} = \bar{Y} - \hat{\beta}\bar{X} + \hat{\beta}\bar{X} = \bar{Y},$$

die Regressionsgerade geht also stets durch den Punkt (\bar{X}, \bar{Y}) .

Hinweis (Interpretation der Steigung). Mithilfe von Vorhersagen bestätigen wir die statistische Bedeutung von $\hat{\beta}$ bzw. dem linearen Effekt von X auf den Mittelwert von Y : Dazu vergleichen wir die Vorhersagen für $X = x$ und $X = x + 1$: Ihr Unterschied beträgt $\hat{\mu}(x+1) - \hat{\mu}(x) = \hat{\alpha} + \hat{\beta}(x+1) - (\hat{\alpha} + \hat{\beta}x) = \hat{\beta}$. Somit können wir beispielsweise sagen: ‘‘Objekte mit einem um eins größeren X -Wert haben im Schnitt einen um $\hat{\beta}$ größeren Y -Wert’’ oder ‘‘Erhöht sich X um eins, so erhöht sich Y im Schnitt um $\hat{\beta}$ ’’. Letztere (häufig verwendete) Aussage ist zwar elegant, maskiert jedoch die Tatsache, dass stets *verschiedene Objekte* verglichen werden.

Beispiel 6.4 (Körpergrösse und Körpergewicht, Fortsetzung). Um in Beispiel 6.2 das Gewicht einer 170 cm grossen Person vorherzusagen, setzen wir den Wert 170 in den linearen Prädiktor

$$\hat{\mu}(x) = \hat{\alpha} + \hat{\beta} \cdot x = -80.47 + 0.835 \cdot x$$

ein und erhalten $\hat{\mu}(170) = -80.47 + 0.835 \cdot 170 = 61.48$ kg, also den Wert der Regressionsgerade bei 170. Diese Zahl kann als Mittelwert von 170 cm grossen Personen aufgefasst werden, auch wenn evtl. gar niemand im Datensatz exakt 170 cm gross ist.

Das Gewicht einer 171 cm grossen Person finden wir mit $\hat{\mu}(171) = -80.47 + 0.835 \cdot 171 = 62.315$ kg oder zählen den linearen Effekt $\hat{\beta} = 0.835$ zur Vorhersage bei 170 cm dazu. ▲

6.2.4 Aussagen über die Population

Stammen die X - und Y -Werte aus einer Zufallsstichprobe, postuliert man die lineare Beziehung bzw. die Modellgleichung

$$E(Y | X = x) = \mu(x) := \alpha + \beta x$$

für den wahren Mittelwert $E(Y | X = x)$ von Y bei Beobachtungen mit X -Wert x .

Die Kenngrössen $\hat{\alpha}$, $\hat{\beta}$ und $\hat{\mu}(x)$ schätzen dann die entsprechenden Grössen in der Population und für die wahren Regressionskoeffizienten α und β sind approximative Student-Konfidenzintervalle verfügbar. Zudem kann mit einer Version des (Zweistichproben)- t -Tests die Arbeitshypothese $\beta \neq 0$ versus die Nullhypothese $\beta = 0$ geprüft werden. Da aus einer Steigung ungleich null folgt, dass es einen Zusammenhang zwischen den beiden Merkmalen gibt, stellt dies einen *Test auf Zusammenhang* zwischen zwei numerischen Merkmalen dar. Er kann auch mithilfe des Konfidenzintervallansatzes (liegt 0 im Konfidenzintervall für β ?) durchgeführt werden.

Hinweis (Präzision). Diese Konfidenzintervalle und Tests gelten als präzise, falls die Stichprobe nicht zu klein ist und die meisten Punkte in einem etwa gleichmässig breiten Streifen um die Regressionsgerade liegen.

Beispiel 6.5 (Körpergrösse und Körpergewicht, Fortsetzung). Nun möchten wir die Ergebnisse von Beispiel 6.2 auf die Population übertragen¹. Dabei betrachten wir die Befragten als Zufallsstichprobe aus der Grundgesamtheit aller Studierenden an Schweizer Universitäten.

R Code

```
# Eingabe: Lineare Regression
fit <- lm(Kgewicht ~ Kgroesse, data = wiso)
summary(fit)

# Ausgabe
[...]
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -80.4701    8.9406   -9.0   <2e-16 ***
Kgroesse     0.8345    0.0513   16.3   <2e-16 ***

# Eingabe, Forts.: Konfidenzintervalle
confint(fit)

# Ausgabe
                2.5 %    97.5 %
(Intercept) -98.0793 -62.8608
Kgroesse      0.7336    0.9355
```

¹Die R-Funktionen `summary` und `confint` zeigen die gewünschten Ergebnisse an.

Kommentare

- *Schätzwerte der wahren Regressionskoeffizienten:* Der Schätzwert des tatsächlichen Intercepts α beträgt $\hat{\alpha} = -80.47$ kg, jener der tatsächlichen Steigung β ist $\hat{\beta} = 0.835$. Wir schätzen also, dass sich das tatsächliche mittlere Gewicht pro cm Körpergrösse um 0.835 kg erhöht.
- *95%-Student-Konfidenzintervalle:* Mit einer Sicherheit von je rund 95% liegt der α zwischen -98.1 und -62.9 kg und β zwischen 0.73 und 0.94 .
- *Test auf Zusammenhang auf 5%-Niveau:* Der p -Wert der Nullhypothese, dass $\beta = 0$, ist fast null, also insbesondere kleiner als 5%. Deshalb verwerfen wir die Nullhypothese und behaupten mit einer Sicherheit von rund 95%, dass ‘Körpergrösse’ tatsächlich einen (linearen) Effekt auf den Mittelwert von ‘Körpergewicht’ hat bzw. dass es einen echten Zusammenhang zwischen ‘Körpergrösse’ und ‘Körpergewicht’ gibt. Zum gleichen Testentscheid kommen wir auch via Konfidenzintervallansatz: Der Wert 0 liegt nicht im Konfidenzintervall für β .
- *Präzision:* Wir vertrauen den Ergebnissen der Student-Verfahren: Die Stichprobe ist gross und im Streudiagramm (Abbildung 6.3) liegen die meisten Punkte in einem etwa gleichmässig breiten Streifen um die Gerade. ▲

Beispiel 6.6 (Highschool, Fortsetzung). Nun wollen wir für Beispiel 2.15 den Zusammenhang zwischen ‘Leistung in Mathematik’ und ‘Leistung in Language Arts’ untersuchen. Für die schliessende Statistik betrachten wir den Datensatz als Zufallsstichprobe aus allen US-SchülerInnen.

R Code

```
# Eingabe: Streudiagramm und lineare Regression
fit <- lm(math ~ langarts, data = highschool)
plot(highschool)
abline(fit)
summary(fit)

# Ausgabe
[...]
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.3798     2.1682    6.63  1.4e-10 ***
langarts      0.6865     0.0408   16.84 < 2e-16 ***

# Eingabe, Forts.: Konfidenzintervalle
confint(fit)

# Ausgabe
              2.5 %    97.5 %
(Intercept) 10.11371 18.64595
langarts      0.60631  0.76678
```

Kommentare

- *Streudiagramm:* Das Streudiagramm in Abbildung 6.5 zeigt einen relativ starken positiven linearen Zusammenhang.
- *Regressionskoeffizienten:* Die Steigung $\hat{\beta}$ der Regressionsgerade beträgt 0.6865: Ein zusätzlicher Punkt in Sprache erhöht die Mathleistung also im Schnitt um 0.6865. Der y-Achsenabschnitt $\hat{\alpha}$ beträgt 14.3798: SchülerInnen mit 0 Punkten in Sprache erreichen im Schnitt 14.3798 Punkte in Mathematik.

- *Vorhersage:* Für eine Schülerin mit 70 Punkten in Sprache bzw. für den Mittelwert solcher SchülerInnen erwarten wir eine Mathleistung von

$$\hat{\mu}(70) = 14.3798 + 0.6865 \cdot 70 \approx 62.4.$$

Dies entspricht dem Wert der Regressionsgerade an der Stelle 70.

- *Schätzwerte und 95%-Student-Konfidenzintervalle für die wahren Regressionskoeffizienten:* Wir schätzen, dass die Leistung in Mathematik pro Punkt in Sprache bei amerikanischen SchülerInnen im Schnitt um $\hat{\beta} = 0.69$ Punkte steigt. Mit einer Sicherheit von ca. 95% liegt die wahre Steigung β zwischen [0.61, 0.77]. Den wahren y -Achsenabschnitt α schätzen wir auf $\hat{\alpha} = 14.38$. Mit einer Sicherheit von rund 95% liegt α zwischen 10.11 und 18.65 Punkten.
- *Schätzwert zu Vorhersage:* Wir schätzen, dass die tatsächliche mittlere Punktzahl $\mu(70)$ in Mathematik bei SchülerInnen mit 70 Punkten in Sprache $\hat{\mu}(70) = 62.4$ Punkte beträgt.
- *Test auf Zusammenhang auf 5%-Niveau:* Der t -Test der Nullhypothese, dass die tatsächliche Steigung β null ist, liefert einen p -Wert unter 0.05. Mit einer Sicherheit von rund 95% gibt es also tatsächlich einen Zusammenhang zwischen ‘Leistung in Language Arts’ und ‘Leistung in Mathematik’. Zum gleichen Ergebnis gelangt man mit dem Konfidenzintervallansatz: Null liegt nicht im Konfidenzintervall für β .
- *Präzision:* Wir vertrauen den p -Werten und Konfidenzintervallen, da die Stichprobe gross ist und die meisten Punkte in einem ungefähr gleichmässig breiten Streifen liegen. ▲

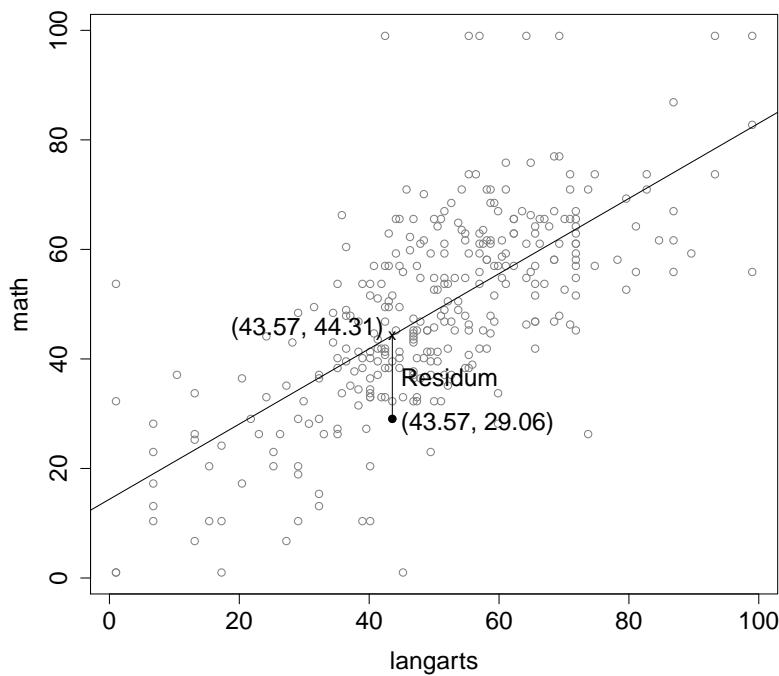


Abbildung 6.5: Streudiagramm von ‘Leistung in Mathematik’ und ‘Leistung in Language Arts’ in den Beispielen 6.6 und 6.8. Der Punkt (43.57, 29.06) der vierten Beobachtung ist hervorgehoben. Der vertikale Abstand zur Regressionsgerade entspricht dem Residuum -15.21 (Punkte unterhalb der Geraden haben negative Residuen). Der y -Wert 44.31 der vertikalen Projektion auf die Gerade entspricht dem gefüllten Wert.

6.3 Kovarianz, Bestimmtheitsmass, Korrelationen

Mithilfe von Kovarianzen, Korrelationen und dem Bestimmtheitsmass gehen wir der Frage nach, wie *stark* ein allfälliger Zusammenhang zwischen den zwei numerischen Merkmalen ist.

6.3.1 Kovarianz

Wir haben die Stichprobenkovarianz

$$\text{Cov}(X, Y) := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

als Hilfsgrösse bei der Regression kennengelernt. Ein positiver bzw. negativer Wert liegt definitionsgemäss dann vor, wenn überdurchschnittliche X -Werte zu über- bzw. unterdurchschnittlichen Y -Werten tendieren. Mit der Kovarianz alleine ist es jedoch schwierig, die Stärke des (linearen) Zusammenhangs zwischen X und Y zu beurteilen, da ihr Wert auch von der Streuung der Merkmale abhängt. Die standardisierte Version der Kovarianz, die Pearson-Korrelation, ist zu diesem Zweck besser geeignet. Wir werden nachher auf sie eingehen.

Kovarianzen dienen uns also lediglich als Hilfsgrösse bei Regression und Korrelation. Da ihr theoretisches Pendant

$$\text{Cov}(X, Y) := E((X - E(X))(Y - E(Y)))$$

jedoch in einigen wichtigen finanzmathematischen Formeln auftaucht, starten wir diesen Abschnitt mit ihren wichtigen Eigenschaften. Diese gelten auch für die Stichprobenkovarianz.

Eigenschaften

1. $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$
2. Sind X und Y unabhängig, ist $E(XY) = E(X)E(Y)$ und die Kovarianz demnach null. Die Umkehrung gilt nicht.
3. Die Kovarianz zwischen X und Y ist gleich wie jene zwischen Y und X .
4. Die Kovarianz zwischen X und X entspricht der Varianz von X .
5. $\text{Cov}(a + bX, Y) = b\text{Cov}(X, Y)$
6. $\text{Cov}(X, Y + Z) = \text{Cov}(X, Y) + \text{Cov}(X, Z)$

Daraus folgt beispielsweise eine berühmte Formel für die Varianz einer Summe von (nicht unabhängigen) Zufallsvariablen:

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

Sie lässt sich auf mehrere Zufallsvariablen erweitern.

Beispiel 6.7 (Portfoliorenditen). In Beispiel 3.40 haben wir zur Berechnung der Portfoliovarianz unrealistischerweise angenommen, dass die Renditen der beiden Anlagen unabhängig sind. Wir verzichten nun auf diese Annahme und bezeichnen die Kovarianz der beiden Renditen mit c .

Mit den Eigenschaften von Varianzen und Kovarianzen finden wir

$$\text{Var}(T) = \text{Var}\left(\frac{2}{3}R_1 + \frac{1}{3}R_2\right) = \frac{4}{9}\text{Var}(R_1) + \frac{1}{9}\text{Var}(R_2) + \frac{4}{9}c.$$



6.3.2 Bestimmtheitsmass

Bevor wir die Kovarianz verwenden, um die berühmte *Korrelation* zu definieren, gehen wir auf einen alternativen Ansatz zur Quantifizierung der Stärke des linearen Zusammenhangs ein, den wir bereits aus der Varianzanalyse kennen: Betrachten wir die *gefitteten Werte*

$$\hat{Y}_i := \hat{\mu}(X_i) = \hat{\alpha} + \hat{\beta} X_i,$$

also die Vorhersagen von Y_i durch X_i . Je besser deren Übereinstimmung mit den Y_i , bzw. je kleiner die Varianz der *Residuen*

$$e_i := Y_i - \hat{Y}_i$$

gegenüber der Varianz von Y ist, je stärker ist der lineare Zusammenhang zwischen den X - und Y -Werten. Als Mass für die Stärke des linearen Zusammenhangs bietet sich entsprechend das *Bestimmtheitsmass*

$$R^2 := 1 - \frac{\text{Var}(e)}{\text{Var}(Y)} = \frac{\text{Var}(\hat{Y})}{\text{Var}(Y)}$$

an, das hier genau gleich definiert und interpretiert wird wie bei der Varianzanalyse. Es gibt auch hier an, wieviel Prozent der Varianz von Y durch X erklärt ist.

Hinweis (Geometrische Bedeutung der Residuen). Die Residuen entsprechen den vertikalen Abständen zwischen den Punkten im Streudiagramm und der Regressionsgerade. Punkte über der Gerade haben ein positives Residuum, Punkte darunter ein negatives. Abbildung 6.5 illustriert die Situation für Beispiel 6.6. Die Methode der kleinsten Quadrate minimiert also die Summe der *quadrierten vertikalen* Abstände zwischen den Punkten und der Gerade.

Beispiel 6.8 (Highschool, Fortsetzung). Für Beispiel 6.6 sehen wir, dass ‘Leistung in Language Arts’ fast die Hälfte der Varianz von ‘Leistung in Mathematik’ erklärt¹. Es liegt also ein starker linearer Zusammenhang vor.

R Code

```
# Eingabe
fit <- lm(math ~ langarts, data = highschool)
summary(fit)

# Ausgabe
[...]
Multiple R-squared: 0.4744
```

Um die Konstruktion des R^2 zu illustrieren, betrachten wir die vierte Beobachtung des Datensatzes: Dieser Schüler hat 29.1 Punkte in Math und 43.6 Punkte in Sprache erreicht. Der gefittete Wert beträgt

$$\hat{Y}_4 = \hat{\mu}(43.6) = \hat{\alpha} + \hat{\beta} \cdot 43.6 = 14.3798 + 0.6865 \cdot 43.6 = 44.3112.$$

Das Residuum entspricht der Differenz $e_4 = 29.1 - 44.3112 = -15.2112$ zwischen beobachtetem und gefittem Wert, siehe Abbildung 6.5 für eine grafische Darstellung. Aus den Varianzen² der Residuen (168.03) und der Leistung in Mathematik (319.72) finden wir $R^2 = 1 - 168.03/319.72 = 0.4744$. ▲

Beispiel 6.9 (Körpergrösse und Körpergewicht, Fortsetzung). Die Situation in Beispiel 6.1 weist ein R^2 von 0.517 auf. Der lineare Zusammenhang zwischen den beiden Merkmalen ist somit stark: Etwa die Hälfte der Varianz von ‘Körpergewicht’ wird durch ‘Körpergrösse’ erklärt.

¹Via R-Funktion `summary`.

²Mit der Software bestimmt.

R Code

```
# Eingabe: Lineare Regression
fit <- lm(Kgewicht ~ Kgroesse, data = wiso)
summary(fit)

# Ausgabe
[...]
Multiple R-squared: 0.517
```



6.3.3 Korrelationskoeffizient nach Pearson

Eng mit dem Bestimmtheitsmass verwandt ist der *Korrelationskoeffizient nach Pearson*¹

$$r := \frac{\text{Cov}(X, Y)}{\text{Std}(X) \text{Std}(Y)},$$

der stets Werte zwischen -1 und 1 annimmt. Das Vorzeichen ist identisch mit dem Vorzeichen von $\hat{\beta}$ und das Quadrat entspricht dem Bestimmtheitsmass, also $r^2 = R^2$. Es ist $r = 1$ bzw. $r = -1$ genau dann, wenn alle Punkte auf einer Geraden mit positiver bzw. negativer Steigung liegen. Je grösser $|r|$ ist, je stärker ist der lineare Zusammenhang zwischen den X - und Y -Werten. Ausserdem bleibt r unverändert, wenn man die Rollen von X und Y vertauscht, zu allen X - oder Y -Werten eine Konstante addiert oder alle X - oder Y -Werte mit einem positiven Wert multipliziert (bei einem negativen Wert ändert sich das Vorzeichen).

Hinweise

- *Statistik-Slang*: Je nach Wert sagt man “ X und Y korrelieren nicht/schwach/stark positiv/negativ”.
- *Faustregeln*: Bei der ANOVA haben wir manchmal bei einem R^2 bis 0.1 von einem höchstens schwachen, bei einem Wert ab 0.3 von einem starken Zusammenhang gesprochen. Entsprechend könnten wir bei $|r| \leq \sqrt{0.1} = 0.316 \approx 0.3$ von einem höchstens schwachen linearen Zusammenhang sprechen, bei $|r| \geq \sqrt{0.3} = 0.5477 \approx 0.5$ von einem starken.
- *Interpretationsmöglichkeit*: Das Bestimmtheitsmass entspricht der quadrierten Pearson-Korrelation zwischen Y und \hat{Y} . Diese nichttriviale Regel gilt generell, also auch bei der Varianzanalyse. Dort kann das R^2 als quadrierte Pearson-Korrelation zwischen den Y -Werten und einem künstlichen Merkmal, das als Ausprägungen die entsprechenden Gruppenmittelwerte enthält (die gefitteten Werte), aufgefasst werden.

Beispiel 6.10 (Highschool, Fortsetzung). Die Leistung in Mathematik korreliert stark positiv mit der Leistung in Sprache²:

R Code

```
cor(highschool$math, highschool$langarts) # Ergibt 0.688792
```

Aus dem R^2 von 0.474 in Beispiel 6.8 und der Tatsache, dass die Steigung der Regressionsgerade positiv ist, finden wir diesen Wert ohne Software durch $r = \sqrt{0.4744} \approx 0.689$.

Die Pearson-Korrelation entspricht definitionsgemäss der Kovarianz dividiert durch die beiden Standardabweichungen. Verfügen wir also über diese Zahlen bzw. über die VC-Matrix, so kann die Korrelation auch auf diese Art bestimmt werden.

¹Die Definition und die meisten genannten Eigenschaften gelten sowohl für die theoretische als auch die empirische Version.

²Dieser Wert kann direkt mit der R-Funktion `cor` gefunden werden.

R Code

```
# Eingabe
var(highschool[, c('math', 'langarts')])

# Ausgabe
      math langarts
math     319.7214 220.9415
langarts 220.9415 321.8153
```

$$\text{Damit finden wir } r = \frac{220.9415}{\sqrt{319.7214} \sqrt{321.8153}} = 0.689.$$

▲

Beispiel 6.11 (Körpergrösse und Körpergewicht, Fortsetzung). Aus dem R^2 von 0.517 in Beispiel 6.9 und der Tatsache, dass die Steigung der Regressionsgerade positiv ist, finden wir die Pearson-Korrelation $r = \sqrt{0.517} \approx 0.719$. Die beiden Merkmale korrelieren damit stark positiv.

Die Software bestätigt diese Berechnung:

R Code

```
cor(wiso$Kgewicht, wiso$Kgroesse, use = 'pair') # Ergibt 0.71881
```

Wie bei einer VC-Matrix werden paarweise Korrelationskoeffizienten von mehreren Merkmalen oft zu einer *Korrelationsmatrix* zusammengefasst. Die Streudiagramm-Matrix in Abbildung 6.2 könnte also durch folgende Korrelationsmatrix ergänzt werden:

R Code

```
# Eingabe
cor(wiso[, c('MonMiete', 'Kgroesse', 'Kgewicht')], use = 'pair')

# Ausgabe
      MonMiete Kgroesse Kgewicht
MonMiete 1.000000 0.037466 0.070462
Kgroesse 0.037466 1.000000 0.718812
Kgewicht 0.070462 0.718812 1.000000
```

Diese liesse sich aus der VC-Matrix aus Beispiel 6.2 ermitteln, indem die Kovarianzen je durch die Standardabweichungen der betreffenden Merkmale dividiert würden.

▲

6.3.4 Rangkorrelationskoeffizient nach Spearman

Wir haben den Korrelationskoeffizienten nach Pearson als Mass für die Stärke des *linearen* Zusammenhangs zwischen zwei numerischen Variablen kennengelernt. Manchmal interessiert man sich jedoch (nur) für die Stärke des *monotonen* Zusammenhangs zwischen den X - und Y -Werten, d. h. man möchte wissen,

- wie stark grössere X -Werte zu grösseren Y -Werten tendieren (positiver monotoner Zusammenhang) bzw.
- wie stark grössere X -Werte zu kleineren Y -Werten tendieren (negativer monotoner Zusammenhang).

Solche Fragestellungen werden mit dem *Rangkorrelationskoeffizienten ρ nach Spearman* beantwortet. Um ihn zu berechnen, ersetzt man die X - und Y -Werte durch ihre Ränge und wendet darauf den Korrelationskoeffizienten nach Pearson an. ρ liegt entsprechend ebenfalls zwischen -1 und 1 , wobei ein Wert von 1 bzw. -1 bedeutet, dass zu einem grösseren X -Wert automatisch ein grösserer bzw. kleinerer Y -Wert gehört.

Der Kenngrösse hat die üblichen angenehmen Eigenschaften von rangbasierten Verfahren, ist also beispielsweise robust gegenüber Ausreissern (im Gegensatz zur Pearson-Korrelation ist hier kein starker Leverage-Effekt möglich, da Ränge keine Ausreisser sein können) und kann auf ordinale¹ X und/oder Y angewendet

¹ Andere Masse für den Zusammenhang zweier ordinaler Merkmale heissen *Kendalls Tau* und *Goodman-Kruskal Gamma*.

werden. Zudem bleibt die Rangkorrelation unverändert, wenn die Merkmale monoton steigend transformiert werden. (Bei einer monoton fallenden Transformation ändert sich das Vorzeichen.)

Hinweise

- *Vergleich mit Pearson-Korrelation:* Ausser bei stark nichtlinearen monotonen Zusammenhängen sind die beiden Korrelationskoeffizienten ähnlich.
- *Fragebögen:* Die Rangkorrelation nach Spearman eignet sich beispielsweise, um die Stärke des Zusammenhangs zwischen einem (zahlenkodierten) ordinalen Fragebogenitem und einer (ordinalen, binären oder numerischen) Angabe wie Geschlecht oder Alter zu quantifizieren.

Beispiel 6.12 (Highschool, Fortsetzung). Der Rangkorrelationskoeffizient nach Spearman¹ ist hier fast gleich gross wie Pearsons Korrelationskoeffizient (0.689):

R Code	cor(highschool\$math, highschool\$langarts, method = 's') # Ergibt 0.69456
--------	--

Er entspricht tatsächlich der Pearson-Korrelation zwischen den rangtransformierten Merkmalen:

R Code	cor(rank(highschool\$math), rank(highschool\$langarts)) # Ergibt 0.69456
--------	--

Folgender Output zeigt die beiden Arten von Korrelationen einmal für linear und einmal für logarithmisch transformierte Leistungen in Mathematik: (Positive) lineare Transformationen haben keinen Einfluss, logarithmische nur auf Pearsons Korrelationskoeffizienten.

R Code	# Eingabe: Pearson-Korrelation mit halbiertter Leistung in Mathematik cor(highschool\$math/2, highschool\$langarts) # Ergibt 0.68879
--------	---

# Eingabe: Pearson-Korrelation mit logarithmierter Leistung in Mathematik cor(log(highschool\$math), highschool\$langarts) # Ergibt 0.60675	
--	--

# Eingabe: Spearman-Rangkorrelation mit halbiertter Leistung in Mathematik cor(highschool\$math/2, highschool\$langarts, method = 's') # Ergibt 0.69456	
--	--

# Eingabe: Spearman-Rangkorrelation mit logarithmierter Leistung in Mathematik cor(log(highschool\$math), highschool\$langarts, method = 's') # Ergibt 0.69456	
---	--



Beispiel 6.13 (Körpergrösse und Körpergewicht, Fortsetzung). Nun wollen wir für die drei Variablen ‘Monatsmiete’, ‘Körpergrösse’ und ‘Körpergewicht’ von Beispiel 6.11 auch noch die paarweisen Rangkorrelationskoeffizienten nach Spearman angeben.

R Code	# Eingabe cor(wiso[, c('MonMiete', 'Kgroesse', 'Kgewicht')], use = 'pair', method = 's')
--------	---

# Ausgabe	MonMiete Kgroesse Kgewicht MonMiete 1.000000 0.030031 0.066535 Kgroesse 0.030031 1.000000 0.730391 Kgewicht 0.066535 0.730391 1.000000
-----------	---



¹Mit der Option `method = 's'`.

6.3.5 Aussagen über die Population

Liegt eine Zufallsstichprobe vor, dienen R^2 , r und ρ als Schätzwerte für die entsprechenden Werte in der Population. Wie üblich sind approximative Konfidenzintervalle verfügbar und können beispielsweise verwendet werden, um ein- oder zweiseitige Hypothesen über die Populationswerte zu prüfen. Softwares bieten in der Regel folgende explizite Tests auf Zusammenhang an:

- *F-Test*: Er prüft die Nullhypothese “tatsächliches R^2 ist null”. Diesen Test haben wir bereits bei den Mittelwertvergleichen verwendet.
- *Test auf Pearson-Korrelation*: Er prüft die Nullhypothese “tatsächliche Pearson-Korrelation ist null”.
- *Test auf Rangkorrelation nach Spearman*: Er prüft die Nullhypothese “tatsächliche Rangkorrelation nach Spearman ist null”.

Liefert ein solcher Test einen p -Wert unterhalb des Signifikanzniveaus, behaupten wir mit entsprechender Sicherheit, dass ein echter Zusammenhang zwischen den beiden Merkmalen vorliegt.

Der *F*-Test, der Test auf Pearson-Korrelation sowie der *t*-Test der linearen Regression liefern die gleichen p -Werte. Ihren Ergebnissen trauen wir unter den gleichen Voraussetzungen wie bei der schliessenden Statistik zur linearen Regression.

Beispiel 6.14 (Körpergrösse und Körpergewicht, Fortsetzung). Nun setzen wir diese Verfahren¹ ein, um Rückschlüsse über den Zusammenhang zwischen ‘Körpergrösse’ und ‘Körpergewicht’ in der Population aller Studierenden zu machen.

Was können wir über den tatsächlichen Wert der Pearson-Korrelation sagen?

R Code

```
# Eingabe
cor.test(wiso$Kgewicht, wiso$Kgroesse, data = wiso)

# Ausgabe
t = 16.283, df = 248, p-value < 2.2e-16
95 percent confidence interval: 0.65298 0.77386
sample estimates: 0.71881
```

Kommentare

- *Schätzwert*: Wir schätzen, dass die wahre Pearson-Korrelation 0.719 beträgt.
- *95%-Konfidenzintervall*: Mit einer Sicherheit von rund 95% liegt die wahre Pearson-Korrelation zwischen 0.65 und 0.77.
- *Test auf Zusammenhang auf 5%-Niveau*: Da der p -Wert kleiner als 0.05 ist, können wir mit einer Sicherheit von rund 95% behaupten, dass es einen echten Zusammenhang zwischen den betrachteten Merkmalen gibt bzw. dass diese tatsächlich korrelieren. Der p -Wert entspricht jenem bei der linearen Regression in Beispiel 6.5. Zum gleichen Testentscheid führt auch der Konfidenzintervallansatz (0 liegt nicht im angegebenen Konfidenzintervall).

Nun betrachten wir die möglichen Aussagen über das wahre Bestimmtheitsmass².

¹ Aussagen über die wahre Pearson-Korrelation erhalten wir mit der R-Funktion `cor.test`.

² Die dafür verwendeten Funktionen kennen wir bereits von den Mittelwertvergleichen.

R Code

```
# Eingabe
fit <- lm(Kgewicht ~ Kgroesse, data = wiso)
summary(fit)
[...]
Multiple R-squared: 0.517
F-statistic: 265 on 1 and 248 DF, p-value: <2e-16

# Eingabe, Forts.
confint.R2(fit)

# Ausgabe
Lower.Limit Upper.Limit
0.43396 0.58310

# Eingabe, Forts.
confint.R2(fit, alternative = 'greater')

# Ausgabe
Lower.Limit
0.44759
```

Kommentare

- *Schätzwert:* Wir schätzen den wahren Wert des Bestimmtheitsmasses auf 0.517.
- *Zweiseitiges 95%-Konfidenzintervall:* Mit einer Sicherheit von rund 95% behaupten wir, der wahre Wert liege zwischen 0.434 und 0.583.
- *Untere 95%-Konfidenzschanke:* Mit einer Sicherheit von rund 95% können wir behaupten, dass mindestens 44.8% der tatsächlichen Varianz von ‘Körpergewicht’ durch ‘Körpergrösse’ erklärt wird.
- *F-Test auf 5%-Niveau:* Der p -Wert ist fast null, somit können wir mit einer Sicherheit von rund 95% behaupten, das R^2 in der Population sei positiv bzw. es gäbe einen echten Zusammenhang zwischen ‘Körpergrösse’ und ‘Körpergewicht’. Zum gleichen Schluss führt auch die untere Konfidenzschanke (größer als null). Der Test liefert den gleichen p -Wert wie jener des obigen Tests auf Pearson-Korrelation und auch wie jener des t -Tests bei der linearen Regression in Beispiel 6.5.

Nun die Aussagen über die wahre Rangkorrelation nach Spearman¹:

R Code

```
# Eingabe
cor.test(~ Kgewicht + Kgroesse, data = wiso, method = 's')

# Ausgabe
S = 702096, p-value < 2.2e-16
sample estimates: 0.73039

# Eingabe: Konfidenzintervalle via Pearson-Korrelation auf Ränge
cor.test(~rank(Kgewicht, na.last = 'keep') + rank(Kgroesse, na.last = 'keep')), data = wiso

# Ausgabe
t = 16.835, df = 248, p-value < 2.2e-16
95 percent confidence interval: 0.66662 0.78338
sample estimates: 0.73029
```

¹Die Option `method = 's'` in der R-Funktion `cor.test` fordert den Test auf Zusammenhang basierend auf Spearmans Rangkorrelation an. Um ein approximatives Konfidenzintervall zu erhalten, muss man den Umweg über die Pearson-Korrelation machen. (Wegen den fehlenden Werten ist der Code umständlich.)

Kommentare

- *Schätzwert*: Wir schätzen, dass die echte Rangkorrelation nach Spearman 0.73 beträgt.
- *95%-Konfidenzintervall*: Mit einer Sicherheit von rund 95% liegt der wahre Wert zwischen 0.67 und 0.78. Der Wert 0 (kein Zusammenhang) liegt nicht im Konfidenzintervall, somit können wir die Nullhypothese von keinem Zusammenhang auf dem 5%-Niveau verwerfen.
- *Test auf 5%-Niveau*: Zum gleichen Testentscheid führt auch der angegebene *p*-Wert. ▲

Beispiel 6.15 (Highschool, Fortsetzung). Nun betrachten wir den Zusammenhang zwischen ‘Leistung in Mathematik’ und ‘Leistung in Sprache’ bei amerikanischen SchülerInnen.

R Code

```
# Eingabe: Pearson Korrelation
cor.test(~ math + langarts, data = highschool)

# Ausgabe
t = 16.836, df = 314, p-value < 2.2e-16
95 percent confidence interval: 0.62604 0.74268
sample estimates: 0.68879

# Eingabe: Spearmans Rangkorrelation
cor.test(~ math + langarts, data = highschool, method = 's')

# Ausgabe
S = 1606322, p-value < 2.2e-16
sample estimates: 0.69456

# Eingabe: Konfidenzintervall durch Umweg
cor.test(~ rank(math) + rank(langarts), data = highschool)

# Ausgabe
[...]
95 percent confidence interval: 0.63271 0.74760
```

Kommentare

- *Schätzwerte*: Wir schätzen, dass die beiden wahren Korrelationen je etwa 0.69 betragen.
- *95%-Konfidenzintervalle*: Mit einer Sicherheit von je rund 95% liegt die echte Pearson-Korrelation zwischen 0.63 und 0.74 und die echte Rangkorrelation nach Spearman zwischen 0.63 und 0.75.
- *Tests auf Zusammenhang auf 5%-Niveau*: Die *p*-Werte beider Tests sind kleiner als 0.05 (fast null). So mit können wir je mit einer Sicherheit von rund 95% behaupten, dass es einen echten Zusammenhang zwischen den beiden Merkmalen gibt. Zum gleichen Schluss gelangt man auch mit dem Konfidenzintervallansatz.

Über das tatsächliche R^2 können wir beispielsweise folgendes aussagen:

R Code

```
# Eingabe
fit <- lm(math ~ langarts, data = highschool)
summary(fit)
[...]
Multiple R-squared: 0.474
F-statistic: 283 on 1 and 314 DF, p-value: <2e-16
```

```
# Eingabe, Forts.
confint.R2(fit, alternative = 'greater')
Lower.Limit
0.41081
```

Kommentare

- *Schätzwert*: Ein Schätzwert für den wahren relativen Anteil der Varianz von ‘Leistung in Mathematik’, der durch ‘Leistung in Language Arts’ erklärt wird, beträgt 0.474.
- *Untere Konfidenzschranke*: Mit einer Sicherheit von rund 95% beträgt der echte Wert mindestens 41%. Da die untere Konfidenzschranke grösser als null ist, können wir mit einer Sicherheit von rund 95% behaupten, dass es einen echten Zusammenhang zwischen den beiden Merkmalen gibt.
- *F-Test auf 5%-Niveau*: Der F-Test führt zum gleichen Testentscheid wie der Konfidenzintervallansatz (*p*-Wert ist fast null, also insbesondere kleiner als 0.05). ▲

Beispiel 6.16 (Immobilien). Hier betrachten wir den Datensatz von Beispiel 2.16 mit diversen Angaben zu Mietwohnungen, die an einem bestimmten Tag unter “Immoclick” angeboten wurden.

Eine Immobilienfirma möchte anhand dieser Zufallsstichprobe den Zusammenhang zwischen Monatsmiete und Anzahl Zimmern bei Mietwohnungen in der Schweiz beschreiben. Sie möchte insbesondere wissen, wieviel Miete im Schnitt pro (zusätzlichem) Zimmer bezahlt wird und wieviel Mietzins im Schnitt für eine 3-Zimmer-Wohnung verlangt wird.

R Code

```
# Eingabe: Lineare Regression und Streudiagramm
fit <- lm(Preis ~ Zimmer, data = wohnungen)
summary(fit)
plot(Preis ~ Zimmer, data = wohnungen)
abline(fit)

# Ausgabe
[...]
      Estimate Std. Error t value Pr(>|t|)
(Intercept) 387.3       91.4     4.24  6.5e-05 ***
Zimmer       298.3      30.1     9.92  3.1e-15 ***
[...]
Multiple R-squared: 0.571
F-statistic: 98.4 on 1 and 74 DF, p-value: 3.13e-15

# Eingabe, Forts.: Konfidenzintervalle für die Regressionskoeffizienten
confint(fit)

# Ausgabe
      2.5 % 97.5 %
(Intercept) 205.12 569.41
Zimmer       238.39 358.25

# Eingabe, Forts.: Konfidenzschanke für das R-Quadrat
confint.R2(fit, alternative = 'greater')

# Ausgabe
Lower.Limit
0.4428

# Eingabe: Pearson-Korrelation
cor.test(~ Preis + Zimmer, data = wohnungen)
```

```

# Ausgabe
t = 9.9181, df = 74, p-value = 3.109e-15
95 percent confidence interval: 0.63878 0.83814
sample estimates: cor 0.75544

# Eingabe: Rangkorrelation nach Spearman
cor.test(~ Preis + Zimmer, data = wohnungen, method = 's')

# Ausgabe
S = 16697, p-value = 3.365e-16
sample estimates: rho 0.77175

# Eingabe: Konfidenzintervall via Umweg
cor.test(~ rank(Preis) + rank(Zimmer), data = wohnungen)

# Ausgabe
[...]
95 percent confidence interval: 0.66136 0.84941
sample estimates: 0.77175

```

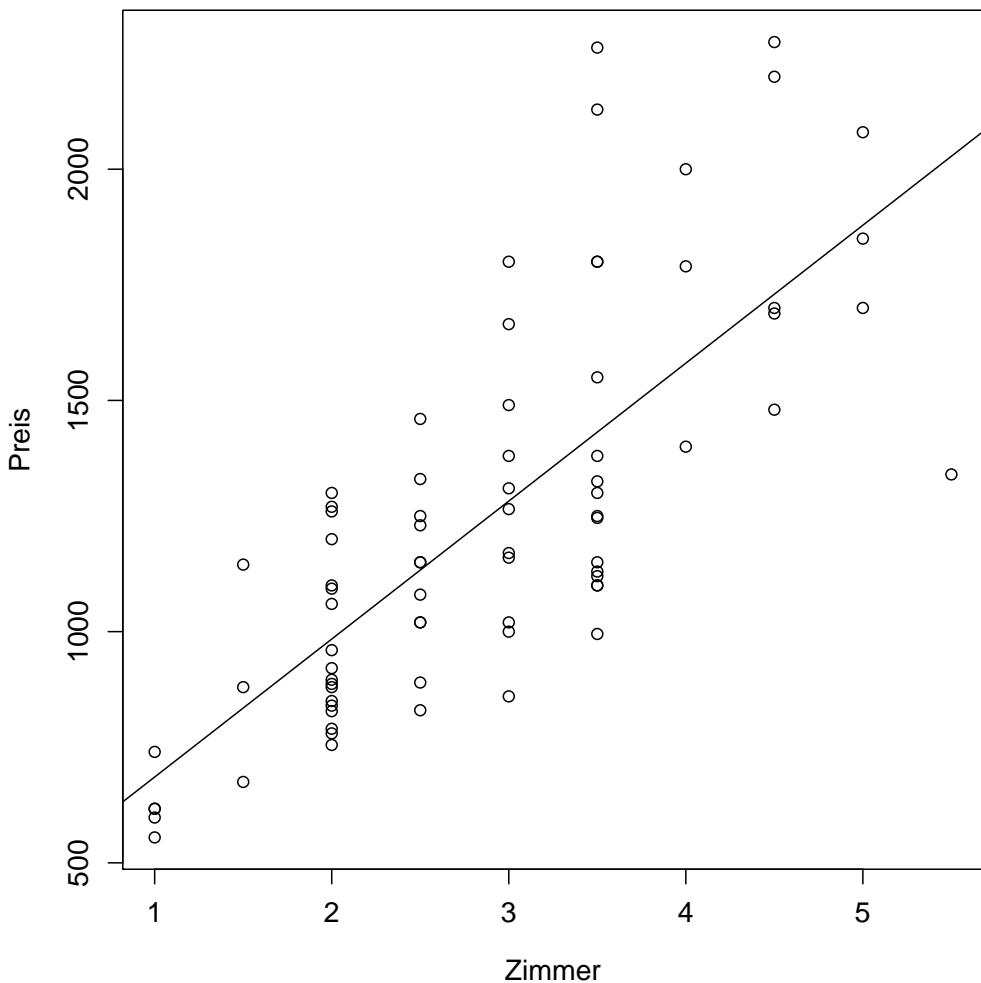


Abbildung 6.6: Streudiagramm und Regressionsgerade von ‘Mietpreis’ und ‘Anzahl Zimmer’ in Beispiel 6.16.

Kommentare

- *Streudiagramm:* Abbildung 6.6 zeigt das Streudiagramm inkl. Regressionsgerade. Es ist ein relativ deutlicher positiver linearer Zusammenhang zu erkennen. Dabei fällt die diskrete Verteilung der Variable ‘Anzahl Zimmer’ auf. Da keine extrem grossen Wohnungen im Datensatz sind, erwarten wir keinen starken Leverage-Effekt.
- *Lineare Regression:* Die Steigung der Regressionsgerade in der Stichprobe beträgt $\hat{\beta} = 298.3$ CHF: Ein zusätzliches Zimmer erhöht den mittleren Mietpreis demnach um rund 300 CHF.

Der y -Achsenabschnitt der Regressionsgerade beträgt $\hat{\alpha} = 387.3$ CHF. Dieser Wert entspricht dem (extrapolierten) mittleren Mietpreis von 0-Zimmerwohnungen.

Der lineare Prädiktor, also die Formel der Regressionsgerade bzw. für Vorhersagen, lautet

$$\hat{\mu}(x) = \hat{\alpha} + \hat{\beta}x = 387.3 + 298.3x$$

Damit erhält man beispielsweise für eine 3-Zimmerwohnung einen erwarteten Mietpreis von

$$\hat{\mu}(3) = 387.3 + 298.3 \cdot 3 = 1282 \text{ CHF}$$

(Da ‘Zimmer’ stark diskret verteilt ist, könnte man alternativ mit den Verfahren aus Kapitel 5 arbeiten und als Vorhersage direkt den mittleren Mietpreis (1283 CHF) der elf 3-Zimmerwohnungen verwenden.)

- *Schliessende Statistik zur linearen Regression:* Ein Schätzwert für die Steigung β in der Population aller Schweizer Mietwohnungen beträgt $\hat{\beta} = 298.3$ CHF. Mit einer Sicherheit von rund 95% liegt β zwischen 238 und 358 CHF. Da dieses Konfidenzintervall den Wert 0 (keine Steigung bzw. kein Zusammenhang) nicht enthält, behaupten wir auf dem 5%-Niveau, dass ein echter Zusammenhang zwischen Anzahl Zimmern und Mietpreis besteht. Zu diesem Schluss gelangen wir auch mit dem t -Test auf Steigung.

Ein Schätzwert für den echten y -Achsenabschnitt α beträgt $\hat{\alpha} = 387.3$ CHF. Mit einer Sicherheit von rund 95% liegt α zwischen 205.1 und 569.4 CHF.

Der tatsächliche mittlere Mietpreis $\mu(3)$ einer 3-Zimmer-Wohnung wird auf $\hat{\mu}(3) = 1282$ CHF geschätzt.

- *Stärke des Zusammenhangs:* Pearsons Korrelationskoeffizient beträgt $r = 0.755$, die beiden Merkmale korrelieren damit deutlich positiv. Spearmans Rangkorrelationskoeffizient ist ähnlich hoch, nämlich 0.772. Das Bestimmtheitsmass $R^2 = r^2 = 0.57$ besagt, dass die Anzahl Zimmer 57% der Varianz des Mietpreises (und umgekehrt) erklärt.
- *Schliessende Statistik zur Stärke des Zusammenhangs:* Ein Schätzwert für die Pearson-Korrelation in der Population beträgt 0.755 (approx. 95% c. i. 0.64 – 0.84), jener für die tatsächliche Rangkorrelation nach Spearman lautet 0.772 (approx. 95% c. i. 0.66 – 0.85). Wir schätzen, dass durch ‘Anzahl Zimmer’ mehr als die Hälfte der Varianz von ‘Monatsmiete’ erklärt wird ($R^2 = 0.571$). Mit einer Sicherheit von rund 95% beträgt der wahre Anteil mindestens 44%. Diese untere Schranke ist grösser als null, somit können wir mit einer Sicherheit von rund 95% behaupten, dass es einen echten Zusammenhang zwischen den beiden Merkmalen gibt. Zu diesem Schluss führt auch der F -Test auf dem 5%-Niveau sowie die Tests auf Korrelation und auch der Test auf positive Steigung.

- *Präzision:* Die meisten Punkte liegen in einem nach rechts oben geöffneten Trichter (und nicht etwa in einem gleichmässig breiten Streifen) – die Streuung der Mietpreise scheint also bei grösseren Wohnungen zu wachsen. Deshalb müssen p -Werte und Konfidenzintervalle vorsichtig beurteilt werden. ▲

6.4 Zusammenfassung

- Der Zusammenhang zwischen zwei numerischen Merkmalen wird grafisch mit dem Streudiagramm beschrieben. Dieses visualisiert die gemeinsame Verteilung der X - und Y -Werte. Paarweise Streudiagramme von mehreren Merkmalen können zu einer Streudiagramm-Matrix zusammengefasst werden.
- Wir haben die lineare Regression kennengelernt: Damit versucht man, den Zusammenhang zwischen den X - und Y -Werten durch eine lineare Funktion zu repräsentieren. Hierfür wird mit der Methode der kleinsten Quadrate eine optimale Gerade durch die Punkte im Streudiagramm gelegt. Der y -Abschnitt und die Steigung dieser Regressionsgerade (die Regressionskoeffizienten) quantifizieren die Form des linearen Zusammenhangs zwischen den Merkmalen. Die Steigung der Regressionsgerade entspricht dem linearen Effekt einer Erhöhung des X -Werts um 1 auf den Mittelwert von Y . Mithilfe der Regressionsgerade bzw. den Regressionskoeffizienten sind Vorhersagen für Y anhand X möglich.
- Wir haben den Leverage-Effekt angesprochen, ein Problem der Methode der kleinsten Quadrate.
- Für die Ermittlung der Regressionsgerade wird unter anderem die Stichprobenkovarianz verwendet, ein Mass für die Stärke des linearen Zusammenhangs zwischen zwei numerischen Merkmalen. Analog einer Streudiagramm-Matrix können paarweise Kovarianzen zwischen mehreren numerischen Merkmalen als Varianz-Kovarianzmatrix präsentiert werden, die auf der Diagonale die Varianzen enthält.
- Mit dem Bestimmtheitsmass R^2 , das wir bereits bei den Mittelwertvergleichen eingeführt haben, wird ebenfalls die Stärke des linearen Zusammenhangs beurteilt. Eng damit verwandt ist die Pearson-Korrelation r , eine Art standardisierte Stichprobenkovarianz. Ihr Vorzeichen gibt die Richtung des Zusammenhangs an. Als Mass für die Stärke des monotonen Zusammenhangs dient die Rangkorrelation ρ nach Spearman. Paarweise Korrelationen zwischen mehreren Merkmalen werden oft zu einer Korrelationsmatrix zusammengefasst.
- Liegt eine Zufallsstichprobe vor, dienen Regressionskoeffizienten sowie R^2 , r und ρ als Schätzwerte für die tatsächlichen Werte in der Population und es können Konfidenzintervalle dafür ausgewiesen werden. Zudem sind Tests auf Zusammenhang verfügbar.

Teil III

Multivariate Verfahren

Kapitel 7

Das lineare Modell

In Teil II haben wir den Zusammenhang zwischen zwei Merkmalen untersucht. Will man dabei Confounder berücksichtigen oder will man generell Zusammenhänge zwischen mehr als zwei Variablen untersuchen, so werden Verfahren der multivariaten Statistik benötigt.

Besonders wichtig sind die sogenannten *Regressionsmodelle*, mit denen man untersucht, wie der Mittelwert einer numerischen Variable Y (der *Zielgröße*) von den Ausprägungen weiterer Merkmale (den *Kovariablen*) abhängt. Damit lassen sich beispielsweise folgende Fragen beantworten:

- Wie sieht der Zusammenhang zwischen Körpergewicht und Größe unter Berücksichtigung des Geschlechts aus?
- Wie hängen die mittleren Mietkosten von der Anzahl Zimmer sowie der Ausstattung ab?
- Verdienen Frauen im Schnitt weniger als Männer unter Berücksichtigung von Alter, Position etc.?

Statt "Zielgröße" und "Kovariable" sind auch folgende Bezeichnungen gebräuchlich:

- Y ist die *abhängige* Variable (AV), die Kovariablen die *unabhängigen* (UV).
- Y ist die *erklärte* Variable, die Kovariablen die *erklärenden*.
- Y ist der *Output*, die Kovariablen der *Input*.
- Weitere Begriffe für die Kovariablen: *Faktoren* oder *Prädiktoren*.

Konkrete Aufgaben von Regressionsmodellen sind die folgenden:

- *Effekte bestimmen*: Es sollen die Effekte der Kovariablen auf den Mittelwert der Zielgröße unter Berücksichtigung der anderen Kovariablen (z. B. potenzielle Confounder) bestimmt werden.
- *Hypothesen prüfen*: Es soll geprüft werden, welche Kovariablen einen echten Zusammenhang mit der Zielgröße aufweisen (unter Berücksichtigung der anderen Kovariablen).
- *Vorhersagen machen*: Anhand der Kovariablen soll der unbekannte Wert der Zielgröße getippt werden.

Es gibt viele Arten von Regressionsmodellen. Wir konzentrieren uns auf das (allgemeine) *lineare Modell*, welches oft kurz *OLS* (von engl. *ordinary least squares*) genannt wird. Dieses zentrale Regressionsmodell ist besonders beliebt, da dessen Ergebnisse vergleichsweise einfach zu interpretieren sind und die Mathematik dahinter vergleichsweise einfach ist. Es umfasst als bivariate Spezialfälle die (einfache) lineare Regression sowie die ANOVA und erweitert und kombiniert diese beiden Verfahren auf natürliche Weise. Zudem ist das lineare Modell Ausgangslage für weitere Arten von Modellen.

7.1 Modellstruktur

Das lineare Modell postuliert für die Population eine lineare Modellstruktur (μ), so dass die Modellgleichung im Wesentlichen folgende Form aufweist:

$$E(Y \mid \text{Kovariablenwerte}) = \mu(\text{Kovariablenwerte}) := \alpha + \text{gewichtete Summe der Kovariablenwerte}.$$

Die Gewichte heissen *Effekte* und bilden zusammen mit dem Intercept α die unbekannten *Modellparameter*, die man mit Daten schätzen möchte.

Da der Mittelwert der Zielgröße durch Veränderungen in den Kovariablenwerten *additiv* beeinflusst wird, liegt eine additive Modellstruktur vor.

In Anlehnung an die Eingabe in die Software R verwenden wir folgende schematische Kurzschreibweise:

$$Y \sim \text{Kovariable 1} + \text{Kovariable 2} + \dots$$

Hinweis (Diskrete Zielgrößen). Raten (bspw. die mittlere Anzahl Krankenkassenrechnungen pro Person und Jahr) und Anteile (bspw. der Befürworteranteil einer Volksinitiative) können als Mittelwerte von numerischen Merkmalen aufgefasst werden. Sie lassen sich deshalb ebenfalls auf diese Weise modellieren.

Beispiel 7.1 (Einfache lineare Regression). Die einfache lineare Regression ist ein OLS mit einer einzigen (numerischen) Kovariable. Ihre Modellgleichung lautet

$$E(Y \mid X = x) = \mu(x) := \alpha + \beta x,$$

schematisch $Y \sim X$. ▲

Beispiel 7.2 (Mittelwertvergleich/ANOVA). Ein Mittelwertvergleich ist ein lineares Modell mit einer einzigen (kategorialen) unabhängigen Variable X mit den Kategorien x_1, \dots, x_L . Bezeichnen wir die entsprechenden Teilpopulationsmittelwerte mit μ_1, \dots, μ_L , so lautet die Modellgleichung

$$E(Y \mid X = x_j) = \mu(x_j) := \mu_j.$$

Legen wir x_1 als Referenzkategorie fest, so kann diese Modellgleichung auch anhand der Mittelwertunterschiede (Effekte) $\beta_j := \mu_j - \mu_1$ durch

$$E(Y \mid X = x_j) = \mu_1 + \beta_j$$

spezifiziert werden. Das Modell lautet schematisch $Y \sim X$. Wie man die rechte Seite der Modellgleichung als lineare Funktion aller Modellparameter $\mu_1, \beta_2, \dots, \beta_L$ schreiben kann, sehen wir im Abschnitt zur *Dummkodierung*. ▲

Beispiel 7.3 (Multiple lineare Regression). Lineare Modelle mit mehreren numerischen Kovariablen (alle in der ersten Potenz) bezeichnet man als *multiple lineare Regressionen*. Die Modellgleichung

$$E(Y \mid X = x, Z = z) = \mu(x, z) := \alpha + \beta x + \gamma z,$$

schematisch $Y \sim X + Z$, beschreibt also eine multiple lineare Regression mit den Kovariablen X und Z . ▲

Beispiel 7.4 (Nichtlineares Modell). Die Modellgleichung

$$E(Y \mid X = x) = \mu(x) := \alpha + x/\beta$$

beschreibt zwar eine lineare Regression (μ ist linear in x), jedoch kein lineares Modell (μ ist nicht linear im Modellparameter β). ▲

Beispiel 7.5 (Einfache quadratische Regression). Mit der Modellgleichung

$$E(Y | X = x) = \mu(x) := \alpha + \beta x + \gamma x^2$$

der einfachen quadratischen Regression, schematisch $Y \sim X + X^2$, lässt sich ein quadratischer (also z. B. "U"-förmiger) Zusammenhang zwischen den numerischen Merkmalen Y und X darstellen. Dieses Modell ist zwar keine lineare Regression (μ ist nicht linear in x), zählt jedoch ebenfalls zu den linearen Modellen, da μ linear in den Modellparametern α , β und γ ist. Zusammen mit der einfachen kubischen Regression (zusätzlich ein x^3) gehört es zu den polynomialen Regressionen. ▲

Beispiel 7.6 (Kovarianzanalyse). Die Kovarianzanalyse¹, kurz ANCOVA, ist ein OLS mit einer kategorialen Kovariablen X mit den Kategorien x_1, \dots, x_L und einer numerischen Kovariablen Z . Sie kombiniert die einfache lineare Regression mit dem Mittelwertvergleich. Ihre Modellgleichung lautet entsprechend

$$E(Y | X = x_j, Z = z) = \mu(x_j, z) := \alpha + \beta_j + \gamma z,$$

schematisch $Y \sim X + Z$, wobei die β_j die Mittelwertunterschiede zur Referenzkategorie x_1 angeben. ▲

7.2 Aussagen über die Stichprobe

7.2.1 Intercept, Effekte, Linearer Prädiktor, Vorhersagen

Intercept und Effekte in der Stichprobe werden analog zur einfachen linearen Regression mit der Methode der kleinsten Quadrate bestimmt. Dies ergibt den *linearen Prädiktor*

$$\hat{\mu}(\text{Kovariablenwerte}) := \hat{\alpha} + \text{gewichtete Summe der Kovariablenwerte},$$

die konkrete Formel für Vorhersagen. Die Gewichte entsprechen den Effekten in der Stichprobe.

Effekte von numerischen Kovariablen werden gleich wie bei der einfachen linearen Regression interpretiert, jene von kategorialen Kovariablen wie bei der ANOVA. Die Ausprägungen der jeweils anderen Kovariablen werden dabei festgehalten. Der Intercept repräsentiert den Mittelwert von Y bei Beobachtungen, deren numerische Kovariablen alle den Wert null haben und deren kategoriale Kovariablen die Referenzkategorie als Ausprägung haben.

Beispiel 7.7 (Körpergewicht, Geschlecht und Grösse). In den Beispielen 5.6 und 6.2 bzw. 6.4 haben wir untersucht, wie das mittlere Körpergewicht von StudentInnen separat vom Geschlecht und der Körpergrösse abhängt. Diese bivariaten Betrachtungen haben ergeben, dass Studentinnen im Schnitt rund 12 kg leichter sind als Studenten (58 versus 70 kg) und dass das mittlere Gewicht pro cm Körpergrösse um 0.835 kg zunimmt (Vorhersage einer 170 cm grossen Person: 61.48 kg).

Nun untersuchen wir mithilfe einer Kovarianzanalyse, wie das mittlere Gewicht *gleichzeitig* von Geschlecht und Körpergrösse abhängt. Bezeichnen wir mit z die Körpergrösse einer konkreten Person, so lautet die Modellgleichung (Männer sind Referenzkategorie)

$$E(\text{Gewicht} | \text{Geschlecht} = \text{Mann}, \text{Grösse} = z) = \alpha + \gamma z$$

bzw.

$$E(\text{Gewicht} | \text{Geschlecht} = \text{Frau}, \text{Grösse} = z) = \alpha + \beta + \gamma z$$

mit den Modellparametern α (Intercept in der Population), β (wahrer Effekt von Frauen ggü. Männern) und γ (tatsächlicher linearer Effekt von 'Körpergrösse').

¹Der Begriff ist eine Kombination aus "Kovariable" und "Varianzanalyse".

Da ‘Geschlecht’ ein binäres Merkmal ist, können wir seine Ausprägungen x auch mit 1 (weiblich) und 0 (männlich; Referenzkategorie) bezeichnen und stattdessen die einheitliche Modellgleichung

$$E(\text{Gewicht} \mid \text{Geschlecht} = x, \text{Grösse} = z) = \mu(x, z) = \alpha + \beta x + \gamma z$$

verwenden, schematisch¹ Körpergewicht \sim Geschlecht + Körpergrösse. Bei Männern fällt der mittlere Summand weg.

R Code

```
# Eingabe
fit <- lm(Kgewicht ~ Geschlecht + Kgroesse, data = wiso)
fit

# Ausgabe
(Intercept) GeschlechtW Kgroesse
-39.4231     -5.5273      0.6128
```

Kommentare

- *Effekte und Intercept:* Frauen sind im Schnitt 5.5 kg leichter als (vergleichbar grosse) Männer. Das mittlere Gewicht steigt (unter Berücksichtigung des Geschlechts) pro cm Körpergrösse um 0.613 kg. Der Intercept gibt das (extrapolierte) Gewicht -39.4 kg von 0 cm kleinen Männern an. Wie bei der einfachen linearen Regression brauchen wir ihn lediglich für die Bestimmung von Vorhersagen.
- *Vorhersagen:* Mithilfe des linearen Prädiktors $\hat{\mu}(x, z) = -39.4231 - 5.5273 \cdot x + 0.6128 \cdot z$ finden wir für einen 170 cm grossen Mann die Vorhersage

$$\hat{\mu}(0, 170) = -39.4231 - 5.5273 \cdot 0 + 0.6128 \cdot 170 = -39.4231 + 0.6128 \cdot 170 = 64.8 \text{ kg},$$

für eine 170 cm grosse Frau entsprechend

$$\hat{\mu}(1, 170) = -39.4231 - 5.5273 \cdot 1 + 0.6128 \cdot 170 = 59.2 \text{ kg}.$$

- *Grafische Darstellung:* Das linke Bild in Abbildung 7.1 zeigt eine grafische Darstellung der Situation. Die Regressionsgerade der Männer unterscheidet sich von jener der Frauen nur um eine vertikale Verschiebung um 5.53 kg, also um den Effekt von Geschlecht.

Zu leicht anderen Ergebnissen kommt man, wenn man für die Männer und die Frauen je separat eine einfache lineare Regression berechnet, da dann der Effekt von ‘Grösse’ zwischen den Männern und den Frauen unterschiedlich sein kann, siehe rechtes Bild in Abbildung 7.1.

R Code

```
# Eingabe: Einfache lineare Regression für die Männer
lm(Kgewicht ~ Kgroesse, data = wiso, subset = Geschlecht == 'M')

# Ausgabe
(Intercept) Kgroesse
-27.792      0.548

# Eingabe: Einfache lineare Regression für die Frauen
lm(Kgewicht ~ Kgroesse, data = wiso, subset = Geschlecht == 'W')

# Ausgabe
(Intercept) Kgroesse
-60.817      0.707
```

¹ Auf diese Art und Weise geschieht auch die Eingabe in die R-Funktion `lm`.

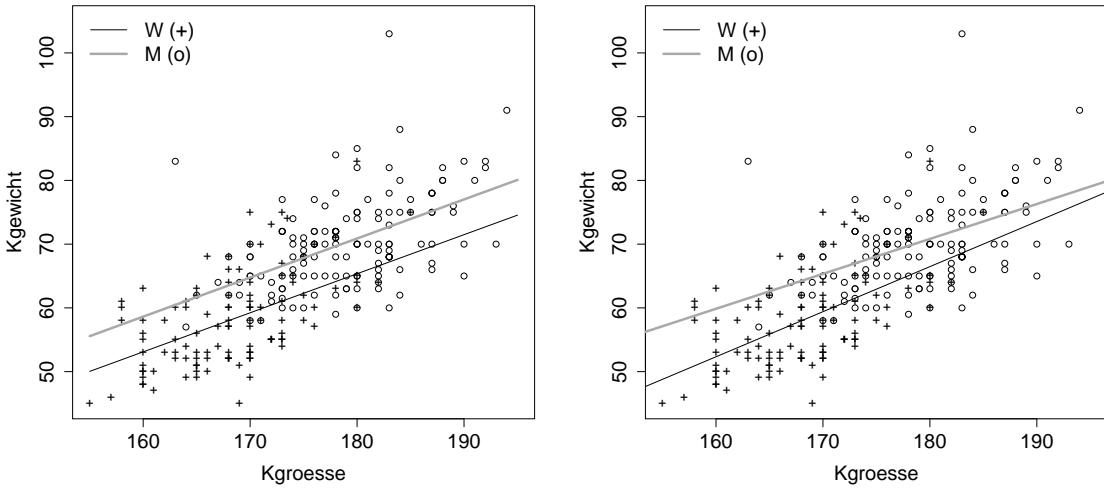


Abbildung 7.1: Grafische Darstellung der Kovarianzanalyse (linkes Bild) sowie der beiden separaten einfachen linearen Regressionen (rechtes Bild) von Beispiel 7.7.

Kommentare

- **Regressionsgeraden:** Männer sind im Schnitt schwerer als (vergleichbar grosse) Frauen. Bei den Frauen ist der lineare Effekt der Körpergrösse auf das mittlere Gewicht grösser ist als bei den Männern.
- **Vorhersagen:** Mit dem linearen Prädiktor der Männer

$$\hat{\mu}_M(x) = -27.792 + 0.548 \cdot x$$

finden wir das erwartete Gewicht eines 170 cm grossen Mannes. Es beträgt

$$\hat{\mu}_M(170) = -27.792 + 0.548 \cdot 170 = 65.4 \text{ kg.}$$

Für eine 170 cm grosse Frau finden wir mit $\hat{\mu}_F(x) = -60.817 + 0.707 \cdot x$ entsprechend

$$\hat{\mu}_F(170) = -60.817 + 0.707 \cdot 170 = 59.4 \text{ kg.}$$

Diese Werte weichen nur leicht von jenen der Kovarianzanalyse ab, da die beiden Geraden im rechten Bild von Abbildung 7.1 eine ähnliche Steigung aufweisen. ▲

Beispiel 7.8 (Highschool, Fortsetzung). In den Beispielen 5.17 und 5.18 haben wir gesehen, dass Personen mit zwei bis fünf Absenzen im Schnitt 2.95 Punkte schlechter in Mathematik abschneiden als solche mit höchstens einer Absenz. Bei mehr als fünf Absenzen beträgt der mittlere Unterschied 9.4 Punkte.

Mithilfe einer Kovarianzanalyse untersuchen wir nun diesen Zusammenhang unter Berücksichtigung des potenziellen Confounders ‘Leistung in Language Arts’. Den Zusammenhang zwischen den beiden Fächern haben wir bereits in den Beispielen 6.6 und 6.8 studiert.

R Code

```
# Eingabe
fit <- lm(math ~ langarts + daysabs, data = highschool)
fit

# Ausgabe
Coefficients:
(Intercept)      langarts    daysabs(1,5]   daysabs(5,50]
15.984          0.673        -0.294        -2.354
```

Kommentare

- *Intercept und Effekte:* SchülerInnen mit 0 Punkten in Sprache und 0–1 Absenzen erreichen im Schnitt 16 Punkte in Mathematik. Pro Punkt in ‘Language Arts’ steigt die mittlere Leistung in Mathematik um 0.673 Punkte. SchülerInnen mit 2 – 5 Absenzen erreichen im Schnitt 0.29 Punkte weniger als solche mit höchstens einer Absenz, während SchülerInnen mit mehr als 5 Absenzen im Schnitt 2.35 Punkte weniger erreichen. Durch den Einbezug des (potenziellen) Confounders ‘Leistung in Language Arts’ sind die Effekte des Absenzverhaltens auf die mittlere Leistung in Mathematik deutlich kleiner geworden.
- *Vorhersage:* Für SchülerInnen mit 10 Absenzen und 70 Punkten in Sprache erwarten wir

$$15.98 + 70 \cdot 0.673 - 2.354 = 60.74 \text{ Punkte}$$

in Mathematik. Zum Vergleich: Die entsprechende Vorhersage ganz ohne Kovariable würde 48.75 Punkte betragen (mittlere Leistung in Mathematik), anhand des Mittelwertvergleichs in Beispiel 5.5 würden wir auf 43.52 Punkte tippen (mittlere Leistung bei > 5 Absenzen), mit der einfachen linearen Regression in Beispiel 6.6 auf 62.44 Punkte. ▲

7.2.2 Modellgüte

Die Stärke des linearen Zusammenhangs zwischen der Zielgröße und den Kovariablen wird durch das (multiple) R^2 bzw. das Bestimmtheitsmaß ausgedrückt¹, welches in diesem Rahmen auch *Modellgüte* genannt wird: Wie bei der einfachen linearen Regression und den Mittelwertvergleichen entspricht es dem Anteil der Varianz von Y , der durch die Kovariablen erklärt ist, bzw. der quadrierten Korrelation zwischen den gefitteten Werten und den Y -Werten. Die Berechnung erfolgt wie bisher durch die Formel

$$R^2 = \frac{\text{Var}(\hat{Y})}{\text{Var}(Y)} = 1 - \frac{\text{Var}(e)}{\text{Var}(Y)},$$

wobei das i -te Residuum

$$e_i := Y_i - \hat{Y}_i$$

auch hier die Differenz zwischen dem beobachteten Wert Y_i und dem gefitteten Wert

$$\hat{Y}_i := \hat{\mu}(\text{Kovariablenwerte der } i\text{-ten Beobachtung})$$

darstellt. Wie bei der einfachen linearen Regression wählt die Kleinst-Quadrat-Methode die Koeffizienten des linearen Prädiktors $\hat{\mu}$ so, dass die Summe der quadrierten Residuen minimal ist.

Beispiel 7.9 (Körpergewicht, Geschlecht und Größe). Die beiden Kovariablen ‘Geschlecht’ und ‘Größe’ im Modell von Beispiel 7.7 erklären zusammen 56% der Varianz von ‘Gewicht’²:

summary(fit)	R Code # Ergibt 0.564
--------------	--------------------------

▲

Beispiel 7.10 (Highschool, Fortsetzung). Für die Kovarianzanalyse von Beispiel 7.8 erhalten wir

summary(fit)	R Code # Ergibt 0.478
--------------	--------------------------

¹Jede zusätzliche Kovariable erhöht das R^2 per Zufall ein bisschen. Eine Variante davon, das *adjustierte R²*, kompensiert dies und ist deshalb kleiner als das R^2 .

²In R gehört die Modellgüte R^2 zur Anzeige von `summary(lm(...))`.

Um die Konstruktion des R^2 zu illustrieren, betrachten wir die vierte Beobachtung des Datensatzes: Diese Schülerin hat 29.1 Punkte in Math und 43.6 Punkte in Sprache erreicht und dreimal gefehlt. Der gefittete Wert beträgt mithilfe des Outputs in Beispiel 7.8 also $\hat{Y}_4 = 15.984 + 43.6 \cdot 0.673 - 0.294 = 45.033$.

Das Residuum entspricht der Differenz $e_4 = 29.1 - 45.033 = -15.933$ zwischen beobachtetem und gefittem Wert. Aus den Varianzen¹ der Residuen (166.97), der gefitteten Werte (152.75) und der Leistung in Mathematik (166.97 + 152.75 = 319.72) finden wir

$$R^2 = 1 - \frac{166.97}{319.72} = \frac{152.75}{319.72} = 0.478.$$

Dieser Wert entspricht übrigens der quadrierten Korrelation zwischen gefitteten und beobachteten Y -Werten:

cor(fitted(fit), highschool\$math)^2	R Code
	# Ergibt 0.478



Um beispielsweise die Stärke des linearen Zusammenhangs zwischen der Zielgröße Y und einer Kovariablen X unter Berücksichtigung von weiteren Kovariablen (z. B. potenziellen Confoundern) zu studieren, vergleicht man die Modellgüte des Modells

$$Y \sim X + \text{weitere Kovariablen}$$

mit jener des Modells

$$Y \sim \text{weitere Kovariablen}$$

ohne X . Man schaut also, um wieviel das R^2 dank X steigt bzw. welcher Varianzanteil der Zielgröße ausschließlich durch X erklärt wird. Ist übrigens dieser Anteil ganz anders als im bivariaten Modell $Y \sim X$, ist dies ein Hinweis auf Confounding durch die weiteren Kovariablen.

Beispiel 7.11 (Highschool, Fortsetzung). Im letzten Beispiel haben wir gesehen, dass die beiden Kovariablen ‘Leistung in Sprache’ und ‘Absenzen’ zusammen 47.8% der Varianz von ‘Leistung in Mathematik’ erklären. Dies ist jedoch nur 0.4% mehr als in der einfachen linearen Regression von Beispiel 6.8 ohne ‘Absenzen’, so dass unter Berücksichtigung von ‘Leistung in Sprache’ kein linearer Zusammenhang zwischen ‘Absenzen’ und ‘Leistung in Math’ festzustellen ist. Im entsprechenden Modell ohne ‘Leistung in Sprache’ (Beispiel 5.17) konnten wir mit einem $R^2 = 0.05$ allenfalls von einem schwachen linearen Zusammenhang zwischen ‘Absenzen’ und ‘Leistung in Math’ sprechen. ▲

Beispiel 7.12 (Körpergewicht und Rauchen, Fortsetzung). In Beispiel 5.12 haben wir praktisch keinen linearen Zusammenhang zwischen ‘Körpergewicht’ und ‘Rauchen’ festgestellt ($R^2 = 0.00364$). Was erhalten wir unter Berücksichtigung der potenziellen Confounder ‘Grösse’ und ‘Geschlecht’?

# Eingabe	R Code
fit <- lm(Kgewicht ~ Rauchen + Geschlecht + Kgroesse, data = wiso)	#
fit	

# Ausgabe	R Code
(Intercept) Rauchen1 Rauchen2 GeschlechtW Kgroesse	summary(fit)
-39.783 -0.549 0.718 -5.472 0.615	

# Eingabe, Forts.	R Code
summary(fit)	[...]
Multiple R-squared: 0.565, Adjusted R-squared: 0.558	

¹Mit der Software bestimmt.

Kommentare

- *Effekte von ‘Rauchen’:* Unter Berücksichtigung der potenziellen Confounder ist das mittlere Gewicht von GelegenheitsraucherInnen 0.55 kg tiefer als jenes von NichtraucherInnen, jenes von regelmässigen RaucherInnen ist 0.72 kg höher.
- *Vorhersagen:* In Beispiel 5.12 haben wir festgestellt, dass NichtraucherInnen im Schnitt 65 kg wiegen. Mit dem hier verfeinerten Modell erhalten wir beispielsweise für einen 170 cm grossen Nichtraucher ein Gewicht von

$$-39.783 + 170 \cdot 0.615 = 64.8 \text{ kg}$$

und für eine 170 cm grosse Nichtraucherin

$$-39.783 - 5.472 + 170 \cdot 0.615 = 59.3 \text{ kg}.$$

Weil die Effekte von ‘Rauchen’ auf das mittlere Gewicht so klein sind, unterscheiden sich diese Vorhersagen kaum von jenen im Modell ohne ‘Rauchen’ (Beispiel 7.7).

- *Modellgüte:* Die Modellgüte ist hier mit 0.565 etwa gleich wie jene des Modells ohne ‘Rauchen’ von Beispiel 7.7 (0.564), so dass der minimale lineare Zusammenhang zwischen ‘Rauchen’ und ‘Gewicht’ von Beispiel 5.12 unter Einbezug von ‘Grösse’ und ‘Geschlecht’ ganz verschwindet. ▲

Beispiel 7.13 (Immobilien, Fortsetzung). In Beispiel 2.16 haben wir einen Datensatz mit Angaben von 76 Wohnungen beschrieben. Dieser wurde in Beispiel 6.16 von einer Immobilienfirma verwendet, um Mietpreise von Schweizer Wohnungen anhand ihrer Anzahl Zimmer einzuschätzen.

Nun soll das Modell mit den binären Variablen ‘Balkon’, ‘Parkett’ und ‘Garten’ verfeinert werden.

R Code

```
# Eingabe
fit <- lm(Preis ~ Zimmer + Balkon + Parkett + Garten, data = wohnungen)
fit

# Ausgabe
(Intercept)      Zimmer      Balkon      Parkett      Garten
406.07        292.67       -36.18       -9.84       102.21

# Eingabe, Forts.
summary(fit)
[...]
Multiple R-squared: 0.582, Adjusted R-squared: 0.559
```

Kommentare

- *Modellgüte:* Mit einem R^2 von 0.58 wird mehr als die Hälfte der Varianz von ‘Mietpreis’ durch die vier Kovariablen erklärt. Wegen den relativ kleinen Effekten von ‘Garten’, ‘Parkett’ und ‘Balkon’ ist die Modellgüte jedoch kaum höher als der entsprechende Wert 0.571 beim einfacheren Modell in Beispiel 6.16.
- *Effekte:* Pro Zimmer erhöht sich der Mietpreis im Schnitt um 293 CHF. Dieser Wert ist sehr ähnlich wie die 298 CHF beim einfacheren Modell in Beispiel 6.16. Bei Wohnungen mit Balkon ist die Miete im Schnitt 36 CHF tiefer als bei solchen ohne (analog für ‘Parkett’ und ‘Garten’).
- *Vorhersage:* Für eine 3-Zimmer-Wohnung mit Balkon (ohne Parkett und Garten) wird ein Mietpreis von

$$406.07 + 3 \cdot 292.67 - 36.18 = 1248 \text{ CHF}$$

erwartet. Dieser Wert ist nahe an der Vorhersage 1282 CHF von Beispiel 6.16 nur anhand ‘Zimmer’. ▲

Dummykodierung Nun wollen wir überlegen, inwiefern eine ANOVA zu den linearen Modellen gehört. Die Idee basiert auf der *Dummykodierung*: Jede kategoriale Kovariable X mit L Kategorien x_1, \dots, x_L kann durch die L binären (0-1)-*Dummyvariablen*

- X_1 : Ausprägung von X ist x_1 (ja = 1, nein = 0)
- X_2 : Ausprägung von X ist x_2 (ja = 1, nein = 0)
- ...
- X_L : Ausprägung von X ist x_L (ja = 1, nein = 0)

ersetzt werden, also durch mehrere numerische Variablen. Da pro Beobachtung genau eine davon den Wert eins annimmt und die anderen null sind, ist eine beliebige Dummyvariable überflüssig und kann ohne Informationsverlust gestrichen werden. Sie spezifiziert die Referenzkategorie.

Die Modellgleichung

$$E(Y | X = x_j) = \mu_1 + \beta_j$$

der ANOVA mit Referenzkategorie x_1 , $E(Y | X = x_1) = \mu_1$ und den Mittelwertunterschieden β_2, \dots, β_L zur Referenzkategorie kann nun mit den Dummyvariablen “kompliziert” als lineare Funktion aller Modellparameter geschrieben werden:

$$E(Y | X_2 = x_2, \dots, X_L = x_L) = \mu_1 + x_2\beta_2 + \dots + x_L\beta_L$$

Dabei ist höchstens einer der Kovariablenwerte¹ x_2, \dots, x_L eins ist (alle anderen sind null). Diese Darstellungsweise rechtfertigt, dass die ANOVA zu den linearen Modellen zählt. Sie ist nichts anderes als eine multiple lineare Regression mit ausschliesslich binären Kovariablen. Tatsächlich werden kategoriale Kovariablen in linearen Modellen von der Software intern stets auf diese Weise behandelt.

Beispiel 7.14 (Körpergewicht und Rauchen, Fortsetzung). Hier berechnen wir eine multiple lineare Regression für die Zielgröße ‚Gewicht‘ mit manuell erzeugten Dummyvariablen für ‚Rauchen‘ und prüfen, ob dies tatsächlich analog zur ANOVA in den Beispielen 5.12 und 5.21 ist:

Zuerst bilden wir die Dummyvariablen ‚Rauchen0‘ (NichtraucherIn: ja = 1, nein = 0), ‚Rauchen1‘ (Gelegentlicher RaucherIn: ja = 1, nein = 0) und ‚Rauchen2‘ (regelmässiger RaucherIn: ja = 1, nein = 0) und betrachten einen Ausschnitt der Daten:

R Code

```
# Eingabe: Dummyvariablen erzeugen
Rauchen0 <- as.numeric(wiso$Rauchen == '0')
Rauchen1 <- as.numeric(wiso$Rauchen == '1')
Rauchen2 <- as.numeric(wiso$Rauchen == '2')

# Eingabe: Erste acht Zeilen zur Illustration
head(data.frame(wiso$Rauchen, Rauchen0, Rauchen1, Rauchen2), 8)

# Ausgabe
Rauchen  Rauchen0  Rauchen1  Rauchen2
0        1        0        0
2        0        0        1
0        1        0        0
2        0        0        1
0        1        0        0
0        1        0        0
0        1        0        0
1        0        1        0
```

¹Die x_2, \dots, x_L bezeichnen hier sowohl die Kategorien von X als auch die Werte der entsprechenden Dummyvariablen.

Indem wir für ‘Grösse’ ein OLS mit ‘Rauchen1’ und ‘Rauchen2’ als Kovariablen berechnen, erhalten wir exakt das gleiche wie mit der ANOVA in den Beispielen 5.12 und 5.21 mit Referenzkategorie ‘0’:

R Code

```
# Eingabe:
fit <- lm(wiso$Kgewicht ~ Rauchen1 + Rauchen2)
fit
summary(fit)

# Ausgabe
Coefficients:
(Intercept)      Rauchen1      Rauchen2
       65.027        -1.277         0.582
[...]
Multiple R-squared:  0.00365
```

Wenn wir mit r_1 und r_2 die Ausprägungen von ‘Rauchen1’ und ‘Rauchen2’ bezeichnen, so lautet der lineare Prädiktor

$$\hat{\mu}(r_1, r_2) = 65.027 - 1.277 \cdot r_1 + 0.582 \cdot r_2.$$

Mit $r_1 = r_2 = 0$ ergibt sich z. B. das mittlere Gewicht 65.027 der NichtraucherInnen. ▲

7.3 Modifikationen der Modellgleichung

Das lineare Modell ist flexibel: Sogenannte

- Transformationen,
- Nichtlinearitäten und
- Interaktionen

in der Modellgleichung ermöglichen manchmal passendere bzw. nützlichere Modelle. Ob solche Modifikationen in einer konkreten Situation nötig bzw. sinnvoll sind, wird idealerweise anhand von Fachwissen oder Literatur *vor* den Berechnungen überlegt. Da sie die Interpretation der Ergebnisse verkomplizieren, werden sie meist nur auf wichtige Variablen angewendet.

Während Transformationen und Nichtlinearitäten auch in bivariaten Situationen eingesetzt werden können, sind Interaktionen nur bei Modellen mit mehreren Kovariablen möglich.

7.3.1 Transformationen

Merkmale werden manchmal *transformiert* als Zielgrösse oder Kovariable verwendet. Hier einige Beispiele:

- Statt Alter als Kovariable wird ein binäres Merkmal ‘Alter über 40 (1 = ja, 0 = nein)’ betrachtet.
- Von der Kovariable ‘Anzahl Zimmer’ werden extrem grosse Werte auf hohem Niveau (z. B. 8) abgeschnitten, um einen allfälligen Leverage-Effekt zu vermindern.
- Stark rechtsschief verteilte Merkmale mit positiven Werten (z. B. Geldbeträge, Zeitdauern) werden oft logarithmiert, z. B. um Werte der gleichen Größenordnung zu erhalten oder Ausreisser zu entschärfen.

Die üblichen Aussagen des Modells (Effekte, Vorhersagen, Modellgüte) beziehen sich dann entsprechend auf die neuen, transformierten Merkmale.

Zumindest bei bijektiven Transformationen stellt sich die Frage, welche Aussagen über die ursprünglichen bzw. untransformierten Merkmale möglich sind. Wir konzentrieren uns auf eine besonders häufige und wichtige Transformation, den *natürlichen Logarithmus*.

Die relevanten Überlegungen werden anhand bivariater Situationen hergeleitet, gelten jedoch allgemein.

Kovariable logarithmiert

Oben wurden bereits Gründe genannt, ein Merkmal X logarithmiert als Kovariable zu verwenden. Hier sind zwei weitere:

- Es soll der Effekt einer *relativen* statt einer absoluten Veränderung von X auf den Mittelwert der Zielgrösse angegeben werden.
- Man erwartet einen logarithmischen Zusammenhang zwischen Kovariable und Zielgrösse.

Im bivariaten Fall lautet die entsprechend modifizierte Modellgleichung

$$E(Y | X = x) = \mu(x) := \alpha + \beta \ln(x),$$

schematisch $Y \sim \ln(X)$. Vorhersagen anhand X werden entsprechend mit dem linearen Prädiktor

$$\hat{\mu}(x) = \hat{\alpha} + \hat{\beta} \ln(x)$$

gemacht.

Welchen Effekt auf den Mittelwert der Zielgrösse hat nun eine Erhöhung des X -Werts um 1%? Dazu vergleichen wir die entsprechenden Mittelwerte für x und $x + x \cdot 1\% = 1.01x$:

$$\begin{aligned} \text{Effekt} &= \mu(1.01x) - \mu(x) \\ &= \alpha + \beta \ln(1.01x) - (\alpha + \beta \ln(x)) \\ &= \beta(\ln(1.01x) - \ln(x)) \\ &= \beta \ln(1.01x/x) \\ &= \beta \ln(1.01) \\ &\approx \beta/100. \end{aligned}$$

Erhöht man den X -Wert um 1%, so erhöht sich der Y -Wert also im Schnitt um etwa $\beta/100$.

Beispiel 7.15 (Körpergrösse und Körpergewicht, Fortsetzung). Was können wir anhand einer einfachen linearen Regression mit ‘Körpergewicht’ als Zielgrösse und ‘ $\ln(\text{Körpergrösse})$ ’ als Kovariable aussagen?

R Code

```
# Eingabe
fit <- lm(Kgewicht ~ log(Kgroesse), data = wiso)
fit
summary(fit)

# Ausgabe
Coefficients:
  (Intercept)  log(Kgroesse)
                -684                 145
[...]
Multiple R-squared: 0.517
```

Kommentare

- *Effekt des neuen Merkmals ‘ $\ln(\text{Körpergrösse})$ ’:* Erhöht sich ‘ $\ln(\text{Körpergrösse})$ ’ um 1, so erhöht sich das mittlere Gewicht um 145 kg.
- *Effekt des ursprünglichen Merkmals ‘Körpergrösse’:* Erhöht sich ‘Körpergrösse’ um 1%, so erhöht sich das mittlere Gewicht um $rund\ 145/100 = 1.45$ kg.
- *Vorhersage für 170 cm grosse Person:* Mit dem linearen Prädiktor

$$\hat{\mu}(x) = -684 + 145 \cdot \ln(x)$$

erhalten wir für eine 170 cm grosse Person die Vorhersage

$$\hat{\mu}(170) = -684 + 145 \cdot \ln(170) = 60.69 \text{ kg.}$$

- *Modellgüte:* Es werden 51.7% der Varianz von ‘Körpergewicht’ durch ‘Körpergrösse’ erklärt.
- *Vergleich mit Modell ohne Ln:* Die Regressionskoeffizienten sowie deren Interpretation waren ganz anders, während die Vorhersage (61.48) und das R^2 (0.517) sehr ähnlich waren. ▲

Zielgrösse logarithmiert

Die Zielgrösse kann aus den beiden bereits oben genannten Gründen logarithmiert werden oder aus einem der folgenden:

- Man ist nicht an absoluten, sondern relativen Effekten auf den typischen Wert der Zielgrösse interessiert, also an einem multiplikativen statt einem additiven Modell.
- Man erwartet einen exponentiellen Zusammenhang zwischen Zielgrösse und Kovariablen.

Indem man die entsprechende (bivariate) additive Modellgleichung

$$E(\ln(Y) | X = x) = \mu(x) := \alpha + \beta x,$$

schematisch $\ln(Y) \sim X$, exponentiiert, entsteht die äquivalente *multiplikative* Modellgleichung

$$e^{E(\ln(Y)|X=x)} = e^{\mu(x)} = e^\alpha \cdot e^{\beta x}.$$

Da man Erwartungswert und nichtlineare Funktionen wie den Logarithmus nicht vertauschen kann, lässt sich die linke Seite leider nicht zu $E(Y | X = x)$ vereinfachen. Es liegt also kein multiplikatives Modell für den arithmetischen Mittelwert, sondern für den sogenannten *geometrischen Mittelwert*

$$e^{E(\ln(Y))}$$

von Y vor. Dessen empirische Version von positiven Werten Y_1, Y_2, \dots, Y_n entspricht dem Ausdruck

$$e^{\frac{1}{n} \sum_{i=1}^n \ln(Y_i)} = e^{\ln(\sqrt[n]{Y_1 \cdot Y_2 \cdots Y_n})} = \sqrt[n]{Y_1 \cdot Y_2 \cdots Y_n},$$

mit dem z. B. in der Finanzwelt durchschnittliche Verzinsungsfaktoren über die Zeit bestimmt werden.

Ein additives Modell für den (arithmetischen) Mittelwert von $\ln(Y)$ liefert also ein multiplikatives Modell für den geometrischen Mittelwert von Y . Vorhersagen für Y findet man, indem man die Werte des linearen Prädiktors $\hat{\mu}$ exponentiiert.

Welchen Effekt hat nun eine Erhöhung des X -Werts um eins? Dazu vereinfachen wir die entsprechende *relative Veränderung* im geometrischen Mittel von Y :

$$\frac{e^{E(\ln Y|X=x+1)} - e^{E(\ln Y|X=x)}}{e^{E(\ln Y|X=x)}} = \frac{e^{E(\ln Y|X=x+1)}}{e^{E(\ln Y|X=x)}} - 1 = \frac{e^{\alpha+\beta(x+1)}}{e^{\alpha+\beta x}} - 1 = e^\beta - 1 \approx \beta,$$

wobei der letzte Schritt nur für kleine β gilt. Erhöht man den X -Wert also um 1, so steigt das geometrische Mittel von Y (oder weniger konkret der *typische Wert* von Y) um *etwa* $\beta \cdot 100\%$ bzw. exakt um $(e^\beta - 1) \cdot 100\%$.

Bei kategorialen Merkmalen wird ebenfalls die relative Veränderung des geometrischen Mittels angegeben.

Beispiel 7.16 (Körpergrösse und Körpergewicht, Fortsetzung). Nun wollen wir ein multiplikatives Modell für den typischen Wert von ‘Körpergewicht’ in Abhängigkeit von ‘Körpergrösse’ finden. Dazu berechnen wir ein lineares Modell für den Mittelwert des logarithmierten Körpergewichts.

R Code

```
# Eingabe
fit <- lm(log(Kgewicht) ~ Kgroesse, data = wiso)
fit
summary(fit)

# Ausgabe
Coefficients:
(Intercept)      Kgroesse
           1.8884        0.0131
[...]
Multiple R-squared:  0.529
```

Kommentare

- *Effekt auf Mittelwert von ‘ $\ln(\text{Körpergewicht})$ ’*: Erhöht sich ‘Körpergrösse’ um 1, so erhöht sich das mittlere logarithmierte Gewicht um 0.013.
- *Effekt auf geometrisches Mittel von ‘Körpergewicht’*: Erhöht sich ‘Körpergrösse’ um 1, so erhöht sich das geometrische Mittel des Gewichts um *rund* $0.0131 \cdot 100\% = 1.31\%$.
- *Vorhersage für 170 cm grosse Person*: Der lineare Prädiktor

$$\hat{\mu}(x) = 1.8884 + 0.0131 \cdot x$$

liefert die Vorhersage für das logarithmierte Gewicht

$$\hat{\mu}(170) = 1.8884 + 0.0131 \cdot 170 = 4.1154.$$

Somit tippen wir auf ein Gewicht von

$$e^{\hat{\mu}(170)} = e^{4.1154} = 61.277 \text{ kg.}$$

Diese Vorhersage lässt sich als typisches Gewicht von 170 cm grossen Personen auffassen.

- R^2 : ‘Körpergrösse’ erklärt 52.9% der Varianz von ‘ $\ln(\text{Körpergewicht})$ ’.
- *Vergleich mit Modell ohne Ln*: Die Regressionskoeffizienten sowie deren Interpretation waren ganz anders, während die Vorhersage (61.48) und das R^2 (0.517) sehr ähnlich waren. ▲

Beispiel 7.17 (Immobilien, Fortsetzung). In Beispiel 7.13 haben wir ein lineares Modell für den mittleren Mietpreis von Wohnungen in der Schweiz bestimmt. Was ergibt ein entsprechendes *multiplikatives* Modell?

R Code

```
# Eingabe
fit <- lm(log(Preis) ~ Zimmer + Balkon + Parkett + Garten, data = wohnungen)
fit

# Ausgabe
Coefficients:
(Intercept)      Zimmer       Balkon       Parkett       Garten
6.3894        0.2420      -0.0526       0.0103      0.0470

# Eingabe, Forts.
[...]
Multiple R-squared: 0.621, Adjusted R-squared: 0.6
```

Kommentare

- *Effekte auf typischen Wert von ‘Preis’*: Pro Zimmer erhöht sich der typische Mietpreis etwa um 24.2%. Bei Wohnungen mit Balkon ist die typische Miete rund 5.3% tiefer als bei solchen ohne.
- *Vorhersage*: Für eine 3-Zimmer-Wohnung mit Balkon (ohne Parkett und Garten) beträgt der lineare Prädiktor $6.3894 + 3 \cdot 0.2420 - 0.0526 = 7.0628$. Somit ergibt sich die Vorhersage $e^{7.0628} = 1167.7$ CHF. Dieser Wert ist tiefer als beim additiven Modell in Beispiel 7.13 (1248 CHF).
- *Modellgüte*: Die Kovariablen erklären 62.1% der Varianz von ‘ $\ln(\text{Preis})$ ’, also leicht mehr als in Beispiel 7.13 (58.2%).

▲

Zielgröße und Kovariable logarithmiert

Je nach Situation wird manchmal sowohl die Zielgröße als auch eine (oder mehrere) numerische Kovariablen log-transformiert. Die Modellgleichung lautet im bivariaten Fall entsprechend

$$E(\ln(Y) | X = x) = \mu(x) := \alpha + \beta \ln(x),$$

bzw. schematisch $\ln(Y) \sim \ln(X)$. Die Interpretation der Ergebnisse eines solchen Modells folgt aus der Kombination der bisherigen Überlegungen. Beispielsweise können wir sagen, dass eine Erhöhung des X -Werts um 1% das geometrische Mittel von Y um *etwa $\beta\%$* erhöht. In der Ökonomie heisst ein solches β übrigens “Elastizität”. Vorhersagen für Y werden gemacht, indem die Werte des entsprechenden linearen Prädiktors $\hat{\mu}$ exponentiiert werden.

Beispiel 7.18 (Körpergrösse und Körpergewicht, Fortsetzung). Was können wir anhand einer einfachen linearen Regression mit ‘ $\ln(\text{Körpergewicht})$ ’ als Zielgröße und ‘ $\ln(\text{Körpergrösse})$ ’ als Kovariable aussagen?

R Code

```
# Eingabe
fit <- lm(log(Kgewicht) ~ log(Kgroesse), data = wiso)
fit
summary(fit)

# Ausgabe
Coefficients:
(Intercept)  log(Kgroesse)
-7.58          2.28
[...]
Multiple R-squared: 0.531
```

Kommentare

- *Effekt auf neue Merkmale bezogen:* Erhöht sich die logarithmierte Körpergrösse um 1, so erhöht sich das mittlere logarithmierte Gewicht um 2.28.
- *Effekt auf ursprüngliche Merkmale bezogen:* Erhöht sich ‘Körpergrösse’ um 1%, so erhöht sich das typische Gewicht um *rund* 2.28 Prozent. Die “Grösse-Gewicht-Elastizität” beträgt also 2.28.
- *Vorhersage für 170 cm grosse Person:* Der lineare Prädiktor

$$\hat{\mu}(x) = -7.58 + 2.28 \cdot \ln(x)$$

liefert den Wert

$$\hat{\mu}(170) = -7.58 + 2.28 \cdot \ln(170) = 4.1296.$$

Somit beträgt die Vorhersage für das Gewicht einer 170 cm grossen Person

$$e^{\hat{\mu}(170)} = e^{4.1296} = 62.153 \text{ kg.}$$

Diesen Wert können wir auch als typisches Gewicht von 170 cm grossen Personen auffassen.

- *Modellgüte:* Es werden 53.1% der Varianz von ‘ $\ln(\text{Körpergewicht})$ ’ durch ‘Körpergrösse’ erklärt. ▲

7.3.2 Nichtlinearitäten

Manchmal werden wichtige numerische Kovariablen (z. B. Alter in einem Modell für mittlere Spitalkosten) oft nicht nur durch einen einzigen Parameter repräsentiert (nämlich die Steigung bzw. den linearen Effekt), sondern durch weitere Parameter, mit denen die nichtlinearen Aspekte des Zusammenhangs erfasst werden. Beispielsweise könnte der Zusammenhang durch ein Polynom höheren Grades beschrieben werden. Neben der Kovariable X werden dann auch ihre Potenzen X^2 , X^3 etc. als zusätzliche Kovariablen ins Modell gesteckt. Spezialfälle von solchen *polynomialen Regressionen* sind die quadratische und die kubische Regression. Im bivariaten Fall lauten ihre Modellgleichungen

$$\begin{aligned} E(Y | X = x) &= \mu(x) := \alpha + \beta x + \gamma x^2 \text{ und} \\ E(Y | X = x) &= \mu(x) := \alpha + \beta x + \gamma x^2 + \delta x^3. \end{aligned}$$

Die Interpretation des Effekts von X gelingt am einfachsten durch systematische Vorhersagen.

Beispiel 7.19 (Immobilien, Fortsetzung). In Beispiel 6.16 haben wir mit einer einfachen linearen Regression den mittleren Mietpreis durch die Anzahl Zimmer erklärt. Obwohl die Punkte im Streudiagramm 6.6 von ‘Preis’ und ‘Zimmer’ ziemlich schön um eine Gerade streuen und damit eine Berücksichtigung von Nichtlinearitäten nicht nötig ist, schauen wir den Output¹ einer entsprechenden kubischen Regression an:

R Code

```
# Eingabe
fit <- lm(Preis ~ Zimmer + I(Zimmer^2) + I(Zimmer^3), data = wohnungen)
fit
summary(fit)

# Ausgabe
(Intercept)      Zimmer   I(Zimmer^2)   I(Zimmer^3)
712.63        -192.03       204.30       -24.67
[...]
Multiple R-squared: 0.5871
```

¹In R-Formeln hat das ‘Hoch’-Zeichen $^$ eine spezielle Bedeutung. Um die Werte eines Merkmals direkt in der Formel zu potenzieren, muss der Ausdruck in ein ‘I’ gepackt werden.

Kommentare: Indem wir die Werte $x = 2, 3, 4$ in den linearen Prädiktor

$$\hat{\mu}(x) = 712.63 - 192.03 \cdot x + 204.3 \cdot x^2 - 24.67 \cdot x^3$$

einsetzen, erhalten wir einen Eindruck über den Effekt von ‘Zimmer’ auf den mittleren Mietpreis: Bei zwei Zimmern ergibt sich ein Preis von 948 CHF, bei drei Zimmern einer von 1309 und bei vier Zimmern schliesslich einer von 1635 CHF. Mit der einfachen linearen Regression von Beispiel 6.16 hätten wir die ähnlichen Vorhersagen 984, 1282 und 1581 CHF erhalten.

Da der Zusammenhang zwischen ‘Preis’ und ‘Zimmer’ ziemlich linear aussieht, haben die nichtlinearen Terme das R^2 nur unwesentlich von 0.571 (Beispiel 6.16) auf 0.587 gehoben, siehe auch Abbildung 7.2. ▲

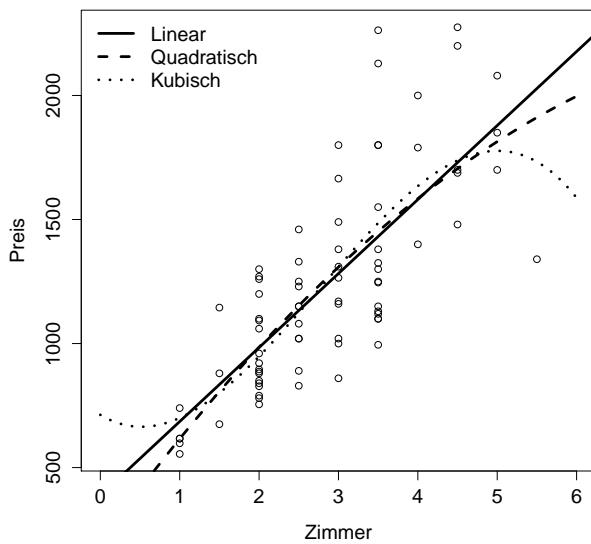


Abbildung 7.2: Streudiagramm von Beispiel 7.19 inkl. Regressionsgerade der einfachen linearen Regression sowie die entsprechenden Polynome der quadratischen und kubischen Regression zum Vergleich. Dabei sieht man u. a., wie gefährlich Extrapolationen sein können.

7.3.3 Interaktionen

Manchmal vermutet man, dass sich die Effekte wichtiger Kovariablen nicht trennen lassen bzw. dass der Effekt der einen Kovariablen deutlich vom Wert einer anderen abhängt. Solche *Interaktionen* bzw. *Wechselwirkungen* zwischen zwei Kovariablen X und Z werden berücksichtigt, indem man die zusätzliche Kovariablen $X \cdot Z$ ins Modell einbaut. Weshalb dies die gewünschte Wirkung hat, sehen wir im folgenden Beispiel.

Beispiel 7.20 (Körpergewicht, Geschlecht und Grösse, Fortsetzung). In Beispiel 7.7 haben wir untersucht, wie das mittlere Gewicht vom Geschlecht und der Körpergrösse abhängt. Dabei haben wir der Einfachheit angenommen, dass der lineare Effekt der Körpergrösse bei den Männern gleich wie bei den Frauen ist (siehe linkes Bild von Abbildung 7.1). Falls wir erwarten, dass der Effekt von ‘Grösse’ bei den Männern deutlich anders sein könnte als bei den Frauen bzw. der Effekt von Geschlecht deutlich von der Grösse abhängen könnte, berechnen wir eine Kovarianzanalyse mit *Interaktion*¹. Die Modellgleichung lautet dann

$$E(\text{Gewicht} | \text{Geschlecht} = x, \text{Grösse} = z) = \mu(x, z) := \alpha + \beta_1 x + \beta_2 z + \beta_3 xz$$

¹Interaktionen zwischen Kovariablen werden in R mit einem * in der Modellgleichung angefordert.

(Dabei fassen wir ‘Geschlecht’ als 0-1-kodiertes binäres Merkmal (1 = weiblich) auf.)

R Code

```
# Eingabe
fit <- lm(Kgewicht ~ Geschlecht * Kgroesse, data = wiso)
fit
summary(fit)

# Ausgabe
Coefficients:
(Intercept)    GeschlechtW   Kgroesse  GeschlechtW:Kgroesse
-27.792        -33.025      0.548       0.159
[...]
Multiple R-squared: 0.567,  Adjusted R-squared: 0.562
```

Kommentare

- *Linearer Prädiktor:* Setzen wir die Koeffizienten des Outputs in die Modellgleichung ein, so erhalten wir den linearen Prädiktor $\hat{\mu}(x,z) = -27.792 - 33.025x + 0.548z + 0.159xz$.
- *Vorhersagen:* Für einen 170 cm grossen Mann erwarten wir damit ein Gewicht von

$$-27.792 + 0.548 \cdot 170 = 65.4 \text{ kg},$$

für eine 170 cm grosse Frau

$$-27.792 - 33.025 + 0.548 \cdot 170 + 0.159 \cdot 170 = -60.817 + 0.707 \cdot 170 = 59.4 \text{ kg}.$$

Die beiden Vorhersagen sind genau gleich wie beim stratifizierten Vorgehen in Beispiel 7.7. Tatsächlich ist eine Kovarianzanalyse mit Interaktion äquivalent zum dort beschriebenen stratifizierten Vorgehen, weist jedoch deutliche Vorteile auf: Beispielsweise muss nur ein einziges Modell berechnet werden und es kann ein gemeinsames R^2 ausgewiesen werden.

- *Interaktionseffekt:* Der lineare Effekt der Grösse ist bei Frauen 0.159 grösser als bei Männern. Dieser Unterschied heisst *Interaktionseffekt*.
- *Modellgüte:* Da die beiden Geraden eine ähnliche Steigung aufweisen, ergibt sich ein nur unwesentlich höheres R^2 von 0.567 als im Modell ohne Interaktion in Beispiel 7.7. Dort war es 0.564. ▲

7.4 Aussagen über die Population

Die Modellgleichung beschreibt den Zusammenhang zwischen den Kovariablen und der Zielgrösse in der Population. Wir überlegen nun, welche Aussagen über Modellparameter, Vorhersagen sowie das R^2 in der Population möglich sind, falls man über eine Zufallsstichprobe verfügt.

7.4.1 Modellparameter

Effekte und Intercept in Stichprobe schätzen die entsprechenden Modellparameter in der Population. Basierend auf dem Studentverfahren sind approximative Konfidenzintervalle und Tests dafür verfügbar, zwischen denen die übliche direkte Verbindung besteht. Der Test prüft die Nullhypothese, dass der betrachtete Modellparameter tatsächlich null ist. Bei einer Kovariable mit nur einem Effekt lässt sich so also prüfen, ob es zwischen ihr und der Zielgrösse einen echten Zusammenhang gibt.

Hinweis (Bivariate Fälle). Bei der einfachen linearen Regression entspricht dieser t -Test dem bereits bekannten Test auf Steigung, bei der ANOVA dem Test auf Mittelwertunterschied zur Referenzkategorie.

Viele Statistiksoftwares fassen die schliessenden Aussagen über alle Modellparameter zur sogenannten *Effekttabelle* zusammen.

Beispiel 7.21 (Immobilien, Fortsetzung). In Beispiel 7.13 haben wir ein Modell für mittlere Wohnungsmieten in Abhängigkeit der Anzahl Zimmer und dem Vorhandensein von Balkon, Parkett und Garten betrachtet. Was können wir damit auf dem 5%-Niveau über die wahren Effekte¹ sagen?

R Code

```
# Eingabe
fit <- lm(Preis ~ Zimmer + Balkon + Parkett + Garten, data = wohnungen)
summary(fit)

# Ausgabe: Effekttabelle
  Estimate Std. Error t value Pr(>|t|)
(Intercept) 406.07     105.55    3.85  0.00026 ***
Zimmer      292.67     30.82    9.50  2.9e-14 ***
Balkon      -36.18     67.25   -0.54  0.59225
Parkett     -9.84     70.89   -0.14  0.88999
Garten       102.21    93.06    1.10  0.27576
[...]

# Eingabe, Forts.: Zusätzlich auch Konfidenzintervalle
confint(fit)

# Ausgabe
            2.5 % 97.5 %
(Intercept) 195.62 616.530
Zimmer      231.21 354.133
Balkon      -170.26 97.904
Parkett     -151.19 131.512
Garten      -83.34 287.760
```

Kommentare (exemplarisch)

- *Zimmer*: Ein Schätzwert für den wahren linearen Effekt β_1 von Zimmer auf den mittleren Mietpreis beträgt $\hat{\beta}_1 = 292.67$ CHF. Mit einer Sicherheit von rund 95% liegt β_1 zwischen 231.21 und 354.13 CHF. Da das Konfidenzintervall den Wert 0 nicht enthält, können wir mit einer Sicherheit von etwa 95% behaupten, dass β_1 nicht null ist bzw. dass es einen echten Zusammenhang zwischen ‘Zimmer’ und ‘Preis’ gibt. Der p -Wert des Tests ist entsprechend kleiner als 0.05.
- *Balkon*: Wir schätzen, dass Wohnungen mit Balkon im Schnitt $-\hat{\beta}_2 = 36.18$ CHF weniger Miete kosten als vergleichbare Wohnungen ohne. Mit einer Sicherheit von rund 95% liegt der wahre Mittelwertunterschied β_2 zwischen -170.26 und 97.90 CHF. Da null in diesem Bereich liegt, können wir auf dem 5%-Niveau nicht behaupten, dass $\beta_2 \neq 0$ bzw. dass es einen echten Zusammenhang zwischen ‘Balkon’ und ‘Preis’ gibt. Deshalb ist der p -Wert des entsprechenden Tests nicht kleiner als 0.05. ▲

7.4.2 Vorhersagen

Der Wert des linearen Prädiktors $\hat{\mu}$ (Kovariablenwerte) kann einerseits als individuelle Prognose des Y -Werts eines Objekts mit diesen Kovariablenwerten aufgefasst werden, andererseits schätzt er den wahren

¹Die Effekttabelle wird in R mit `summary(lm(...))` und `confint(lm(...))` angefordert.

mittleren Y -Wert μ (Kovariablenwerte) aller solcher Objekte in der Population. Solche Mittelwerte können mit approximativen Student-Konfidenzintervallen ausgestattet werden. Will man hingegen die Präzision bzw. den möglichen Bereich einer individuellen Vorhersage angeben, greift man auf die approximativen *Prädiktions-* bzw. *Prognoseintervalle* zurück: Ein 95%-Prädiktionsintervall enthält den wahren Y -Wert eines bestimmten Objekts mit rund 95% Sicherheit, also in rund 95% der Fälle. Auch bei hohem R^2 und grossen Stichproben sind Prädiktionsintervalle in der Regel sehr lang. Dies reflektiert die Tatsache, dass Statistik nicht gut für individuelle Aussagen funktioniert.

Hinweis (Bivariate Fälle). Auch bei den bivariaten linearen Modellen sind beide Arten von Intervallen nützlich.

Beispiel 7.22 (Immobilien, Fortsetzung). In Beispiel 7.13 haben wir den Mietpreis einer 3-Zimmer-Wohnung mit Balkon (ohne Parkett und Garten) auf 1248 CHF getippt. Nun statten wir diese Vorhersage mit einem 95%-Konfidenz- und einem 95%-Prädiktionsintervall¹ aus.

R Code

```

Eingabe: Modell berechnen und neue Daten basteln
fit <- lm(Preis ~ Zimmer + Balkon + Parkett + Garten, data = wohnungen)
new.data <- data.frame(Zimmer = 3, Balkon = 1, Garten = 0, Parkett = 0)

# Eingabe, Forts.: Konfidenzintervall für Populationsmittelwert
predict(fit, new.data, interval = 'c')

# Ausgabe
      fit      lwr      upr
1247.9   1139.8   1356.1

# Eingabe, Forts.: Prognoseintervall für individuelle Prognose
predict(fit, new.data, interval = 'p')

# Ausgabe
      fit      lwr      upr
1247.9    689     1806.8

```

Kommentare

- *Populationsmittelwert*: Die tatsächliche mittlere Miete aller solcher Wohnungen liegt mit einer Sicherheit von etwa 95% im Bereich von 1140 bis 1356 CHF.
- *Individuelle Vorhersage*: Mit einer Sicherheit von rund 95% liegt der Wert einer konkreten solchen Wohnung im viel längeren Intervall von 689 bis 1807 CHF. ▲

7.4.3 Modellgüte

Das R^2 in der Stichprobe schätzt das R^2 in der Population und es sind approximative Konfidenzintervalle verfügbar. Besonders nützlich ist wiederum eine untere Konfidenzschanke: Sie zeigt, wie stark der lineare Zusammenhang zwischen der Zielgrösse und den Kovariablen tatsächlich mit hoher Sicherheit mindestens ist. Ist sie grösser als null, so kann man mit entsprechender Sicherheit behaupten, dass es einen echten Zusammenhang zwischen Zielgrösse und Kovariablen gibt. Letzteres führt zum gleichen Schluss wie der approximative *globale F-Test*. Er prüft die Nullhypothese, dass das wahre R^2 null ist bzw. dass es keinen echten Zusammenhang zwischen der Zielgrösse und den Kovariablen gibt bzw. dass alle Effekte in der Population null sind.

¹In R mit `predict(lm(...), Neue Daten, interval = 'c' bzw. 'p')`.

Hinweis (Bivariate Fälle). Der globale F -Test entspricht bei der einfacher linearer Regression und der ANOVA dem dort üblichen F -Test. Bei der einfachen polynomialen Regression lässt sich damit prüfen, ob es einen echten Zusammenhang zwischen der Kovariablen und der Zielgröße gibt.

Beispiel 7.23 (Immobilien, Fortsetzung). Was können wir anhand des Modells in Beispiel 7.21 über das R^2 in der Population sagen?¹

R Code

```
# Eingabe, Forts.
summary(fit)

# Ausgabe
[...]
Residual standard error: 275 on 71 degrees of freedom
Multiple R-squared: 0.582, Adjusted R-squared: 0.559
F-statistic: 24.7 on 4 and 71 DF, p-value: 7.62e-13

# Eingabe, Forts.: Untere Konfidenzschranke
confint.R2(fit, alternative = 'greater') # Ergibt 0.43161
```

Kommentare

- *Schätzwert*: Wir schätzen das wahre R^2 auf 0.582.
- *Untere 95%-Konfidenzschranke*: Mit einer Sicherheit von etwa 95% erklären die Kovariablen zusammen mindestens 43% der tatsächlichen Varianz von ‘Preis’.
- *Globaler F-Test auf 5%-Niveau*: Mit einer Sicherheit von etwa 95% können wir behaupten, dass es einen echten Zusammenhang zwischen den Kovariablen und dem Mietpreis gibt. Zum selben Schluss kommen wir auch mit der unteren Konfidenzschranke, die grösser als 0 ist. ▲

Mit einem *partiellen F-Test* lässt sich die Nullhypothese prüfen, dass einige bestimmte Modellparameter gleichzeitig null sind bzw. dass diese zusammen das wahre R^2 nicht erhöhen bzw. dass das Modell mit den fraglichen Parametern das gleiche wahre R^2 aufweist wie das Teilmodell ohne diese Parameter.

Hier einige exemplarische Fragestellungen, die mit einem solchen Test beantwortet werden könnten:

- Gibt es einen echten Zusammenhang zwischen der Zielgröße und einer kategorialen Kovariablen?
- Weisen einige Kovariablen gemeinsam einen echten Zusammenhang mit der Zielgröße auf?
- Gibt es einen wahren Zusammenhang zwischen der Zielgröße und einer numerischen Kovariablen mit linearem, quadratischem und kubischem Effekt?
- Verbessern alle Nichtlinearitäten zusammen das wahre R^2 ?
- Gibt es einen Zusammenhang zwischen der Zielgröße und einer Kovariablen mit Interaktion?

Mit partiellen F-Tests lässt sich also u. a. prüfen, welche Kovariablen einen echten Zusammenhang mit der Zielgröße aufweisen. Dies ist insbesondere bei Kovariablen mit mehreren Effekten (Nichtlinearitäten, Interaktionen, kategoriale Kovariablen) nützlich, sonst reicht dazu die Effekttabelle. Bei vielen Statistiksoftwares lassen sich solche *partiellen Tests pro Kovariablen* als sogenannte *Typ-II-ANOVA* anfordern.

¹Schätzwert und globaler F -Test ist in R Teil von `summary(lm(...))`, Konfidenzintervalle sind via Funktion `confint.R2` im Anhang des Skripts verfügbar.

Hinweise zur Typ-II-ANOVA

- Bei einer Kovariable mit nur einem Effekt entspricht der p -Wert des dazugehörigen partiellen F -Tests dem p -Wert des t -Tests in der Effekttabelle.
- Bei einem Modell mit nur einer Kovariable (einfache lineare/polynomiale Regression, Mittelwertvergleich) entspricht der dazugehörige partielle F -Test dem (globalen) F -Test.

Beispiel 7.24 (Körpergewicht und Rauchen, Fortsetzung). Nun soll mit einer Typ-II-ANOVA für das Modell in Beispiel 7.12 geprüft werden, welche der Kovariablen je auf dem 5%-Niveau einen echten Zusammenhang mit der Zielgröße aufweisen¹.

R Code

```
# Eingabe
fit <- lm(Kgewicht ~ Rauchen + Geschlecht + Kgroesse, data = wiso)
summary(fit)

# Ausgabe: Effekttabelle sowie Ergebnisse zum R-Quadrat
[...]
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -39.7833   11.6609  -3.41  0.00076 ***
Rauchen1     -0.5492    1.0744  -0.51  0.60972
Rauchen2      0.7176    1.1000   0.65  0.51479
GeschlechtW  -5.4725    1.0732  -5.10  6.8e-07 ***
Kgroesse      0.6146    0.0651   9.44 < 2e-16 ***
[...]
Residual standard error: 6.3 on 244 degrees of freedom
  (14 observations deleted due to missingness)
Multiple R-squared: 0.565,          Adjusted R-squared: 0.558
F-statistic: 79.4 on 4 and 244 DF, p-value: <2e-16

# Eingabe, Forts.: Typ-II-ANOVA
drop1(fit, test = 'F')

# Ausgabe
[...]
          Df Sum of Sq   RSS AIC F value    Pr(F)
<none>           9680 921
Rauchen     2       34 9714 918    0.43    0.65
Geschlecht  1      1032 10712 945   26.00 6.8e-07 ***
Kgroesse    1      3533 13213 997   89.05 < 2e-16 ***
[...]
```

Kommentare

- Die ausgewiesenen partiellen F -Tests ergeben bei ‘Geschlecht’ und ‘Körpergrösse’ je einen p -Wert unterhalb des 5%-Niveaus. Somit können wir für diese beiden Kovariablen je mit einer Sicherheit von rund 95% behaupten, dass sie einen echten Zusammenhang mit der Zielgröße aufweisen. Der partielle F -Test von ‘Rauchen’ hingegen ist nicht kleiner als 0.05, somit können wir auf dem 5%-Niveau nicht behaupten, dass es einen echten Zusammenhang zwischen ‘Rauchen’ und ‘Gewicht’ gibt.
- Die Typ-II-ANOVA liefert bei den Kovariablen mit nur einem Effekt tatsächlich die gleichen p -Werte wie die Effekttabelle.



¹Eine Typ-II-ANOVA wird in R mit `drop1(lm(...), test = 'F')` angefordert.

Beispiel 7.25 (Immobilien, Fortsetzung). Nun wollen wir auf dem 5%-Niveau prüfen, ob die nichtlinearen Terme der einfachen kubischen Regression von Beispiel 7.19 die Modellgüte tatsächlich verbessern. Den entsprechenden partiellen F -Test erhalten wir, indem wir das Modell mit dem Teilmodell ohne Nichtlinearitäten hinsichtlich dem wahren R^2 vergleichen¹.

R Code

```
# Eingabe: Modell mit Nichtlinearitäten
fit <- lm(Preis ~ Zimmer + I(Zimmer^2) + I(Zimmer^3), data = wohnungen)
summary(fit)

# Ausgabe: Effekttabelle sowie Ergebnisse zum R-Quadrat
# Estimate Std. Error t value Pr(>|t|)
# (Intercept) 712.63 430.45 1.656 0.102
# Zimmer -192.03 486.14 -0.395 0.694
# I(Zimmer^2) 204.30 169.04 1.209 0.231
# I(Zimmer^3) -24.67 18.15 -1.359 0.178
# [...]
# Residual standard error: 271.5 on 72 degrees of freedom
# Multiple R-squared: 0.5871, Adjusted R-squared: 0.5699
# F-statistic: 34.12 on 3 and 72 DF, p-value: 7.841e-14

# Eingabe: Teilmodell ohne Nichtlinearitäten
teil.fit <- lm(Preis ~ Zimmer, data = wohnungen)

# Partiellen F-Test anfordern
anova(fit, teil.fit) # Ergibt p-Wert 0.246
```

Kommentar: Der p -Wert des partiellen F -Tests ist mit 0.246 nicht kleiner als das Signifikanzniveau, somit können wir nicht behaupten, dass die nichtlinearen Terme das wahre R^2 erhöhen.

Anhand des (globalen) F -Tests können wir übrigens auf dem 5%-Niveau behaupten, dass ein echter Zusammenhang zwischen dem Mietpreis und der Anzahl Zimmer (alle Effekte zusammen) besteht. Gäbe es noch weitere Kovariablen im Modell, so könnte dieser Zusammenhang mit einem partiellen F -Test der Nullhypothese ‘Es gibt weder lineare noch nichtlineare Effekte von ‘Zimmer’’ geprüft werden.

Bei der Effekttabelle fällt auf, dass die p -Werte der Effekte alle recht gross sind, obwohl man insgesamt von einem echten Zusammenhang zwischen ‘Zimmer’ und ‘Preis’ sprechen kann. Dieser scheinbare Widerspruch ist eine Konsequenz aus der starken *Multikollinearität* zwischen den verschiedenen Potenzen von ‘Zimmer’ – ein Phänomen, das wir im Anschluss an die nächsten Beispiele besprechen. ▲

Beispiel 7.26 (Immobilien, Fortsetzung). Da die Kovariablen von Beispiel 7.21 je nur einen Effekt aufweisen (keine kategorialen Kovariablen mit mehr als zwei Kategorien, keine Interaktionen, keine Nichtlinearitäten) liefert die Typ-II-ANOVA die gleichen p -Werte wie die Effekttabelle.

R Code

```
# Eingabe, Forts.: Typ-II-ANOVA
drop1(fit, test = 'F')

# Ausgabe
# Df Sum of Sq RSS AIC F value Pr(F)
# <none> 5369832 859
# Zimmer 1 6819043 12188875 919 90.16 2.9e-14 ***
# Balkon 1 21892 5391724 857 0.29 0.59
# Parkett 1 1457 5371289 857 0.02 0.89
# Garten 1 91241 5461073 858 1.21 0.28
```

¹In R wird dies mit `anova(Model, Teilmodell)` gemacht.

Beispiel 7.27 (Immobilien, Fortsetzung). Nun betrachten wir das entsprechende multiplikative Modell, das wir bereits in Beispiel 7.17 beschrieben haben. (Wir verwenden wie üblich das 5%-Niveau.)

R Code

```
# Eingabe
fit <- lm(log(Preis) ~ Zimmer + Balkon + Parkett + Garten, data = wohnungen)
summary(fit)

# Ausgabe: Effekttabelle sowie Ergebnisse zum R-Quadrat
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.3894    0.0803   79.60 < 2e-16 ***
Zimmer      0.2420    0.0234   10.32 8.9e-16 ***
Balkon     -0.0526    0.0511   -1.03    0.31
Parkett     0.0103    0.0539    0.19    0.85
Garten      0.0470    0.0708    0.66    0.51
[...]
Residual standard error: 0.209 on 71 degrees of freedom
Multiple R-squared:  0.621,    Adjusted R-squared:  0.6
F-statistic: 29.1 on 4 and 71 DF,  p-value: 2.43e-14

# Eingabe, Forts.: Untere Konfidenzschranke für das echte R-Quadrat
confint.R2(fit, alternative = 'greater') # Ergibt 0.4805

# Eingabe, Forts.: Konfidenzintervalle für die Modellparameter
confint(fit)

# Ausgabe
            2.5 % 97.5 %
(Intercept) 6.229307 6.549398
Zimmer      0.195228 0.288704
Balkon     -0.154580 0.049353
Parkett     -0.097156 0.117833
Garten      -0.094125 0.188086

# Eingabe, Forts.: Typ-II-ANOVA
drop1(fit, test = 'F')

# Ausgabe
          Df Sum of Sq  RSS AIC F value    Pr(F)
<none>           3.11 -233
Zimmer  1     4.66 7.77 -165 106.56 8.9e-16 ***
Balkon  1     0.05 3.15 -234   1.06    0.31
Parkett 1     0.00 3.11 -235   0.04    0.85
Garten  1     0.02 3.12 -234   0.44    0.51

# Eingabe, Forts.: Vorhersage mit Konfidenzintervall
new.data <- data.frame(Zimmer = 3, Balkon = 1, Parkett = 0, Garten = 0)
predict(fit, new.data, interval = 'c')

# Ausgabe
       fit      lwr      upr
7.0626 6.9804 7.1449

# Eingabe, Forts.: Vorhersage mit Prädiktionsintervall
predict(fit, new.data, interval = 'p')

# Ausgabe
       fit      lwr      upr
7.0626 6.6376 7.4877
```

Kommentare

- *Modellgüte:* Wir schätzen, dass 62.1% der echten Varianz von ‘ln(Preis)’ durch die Kovariablen erklärt werden. Mit einer Sicherheit von rund 95% ist dieser Anteil grösser als 48%. Diese untere Konfidenzschranke ist grösser als null, somit liefert der globale F -Test einen p -Wert unterhalb des 5%-Niveaus. Wir können also mit einer Sicherheit von rund 95% behaupten, dass es einen echten Zusammenhang zwischen den Kovariablen und dem Mietpreis gibt.
- *Typ-II-ANOVA:* Nur ‘Zimmer’ weist einen p -Wert unterhalb des 5%-Niveaus auf. Somit können wir nur für diese Kovariable mit einer Sicherheit von rund 95% behaupten, dass sie einen echten Zusammenhang mit dem Preis aufweist. Da mit jeder Kovariable nur ein Parameter verbunden ist, sind die p -Werte der ausgewiesenen partiellen F -Tests identisch mit jenen der entsprechenden t -Tests.
- *Effekt von ‘Zimmer’:* Wir schätzen, dass sich der typische Preis pro Zimmer um rund 24% erhöht. Mit einer Sicherheit von rund 95% liegt der wahre Wert zwischen ca. 20% und 29%. Dieses Intervall enthält den Wert null nicht. Somit können wir mit einer Sicherheit von rund 95% behaupten, es gäbe einen echten Effekt von ‘Zimmer’ auf den typischen Wert von ‘Preis’ (p -Wert des t -Tests ist kleiner als 0.05).
- *Effekt von ‘Balkon’:* Wir schätzen, dass der typische Preis bei Wohnungen mit Balkon rund 5% tiefer als bei Wohnungen ohne Balkon ist. Mit einer Sicherheit von rund 95% liegt der wahre Effekt zwischen ca. –15% und 5%. Dieses Konfidenzintervall enthält den Wert 0, somit können wir die Nullhypothese von keinem Effekt auf dem 5%-Niveau nicht verwerfen (p -Wert des t -Tests ist ≥ 0.05).
- *Vorhersagen:* Wir schätzen, dass typische 3-Zimmer-Wohnungen mit Balkon (ohne Parkett und Garten) $e^{7.0628} = 1167.7$ CHF Monatsmiete kosten. Mit einer Sicherheit von rund 95% liegt der wahre typische Wert zwischen $e^{6.9804} = 1075.3$ und $e^{7.1449} = 1267.6$ CHF. Das entsprechende Intervall ohne Log-Transformation (Beispiel 7.22) war [1139.8, 1356.1].

Mit einer Sicherheit von rund 95% liegt der Mietpreis einer konkreten solchen Wohnung zwischen $e^{6.6376} = 763.26$ und $e^{7.4877} = 1785.9$ CHF. In Beispiel 7.22 sind wir ohne Log-Transformation zum Intervall [689, 1806.8] gekommen. ▲

Beispiel 7.28 (Highschool, Fortsetzung). In diesem Beispiel betrachten wir ein Modell für die mittlere Leistung in Mathematik in Abhängigkeit der Leistung in Language Arts, des Absenzverhaltens sowie des Geschlechts. Dazu fassen wir die Daten in Beispiel 2.15 als Zufallsstichprobe aus der Population aller US-SchülerInnen auf und möchten auf dem 5%-Niveau mithilfe der Stichprobe Rückschlüsse auf diese Population machen.

R Code

```
# Eingabe
fit <- lm(math ~ langarts + daysabs + male, data = highschool)
summary(fit)

# Ausgabe: Effekttabelle sowie Ergebnisse zum R-Quadrat
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.1280    2.9241   4.83  2.1e-06 ***
langarts     0.6843    0.0428  16.00 < 2e-16 ***
daysabs(1,5]  0.0539    1.8274   0.03    0.98
daysabs(5,50] -2.0228    1.8287  -1.11    0.27
male         2.1441    1.4966   1.43    0.15
[...]
Residual standard error: 13 on 311 degrees of freedom
Multiple R-squared: 0.481,      Adjusted R-squared: 0.475
F-statistic: 72.1 on 4 and 311 DF,  p-value: <2e-16
```

```

# Eingabe, Forts.: Konfidenzintervalle für die Modellparameter
confint(fit)

# Ausgabe
      2.5 %   97.5 %
(Intercept) 8.37443 19.88151
langarts     0.60013  0.76848
daysabs(1,5] -3.54165  3.64952
daysabs(5,50] -5.62086  1.57535
male        -0.80064  5.08889

# Eingabe, Forts.: Untere Konfidenzschranke für das echte R-Quadrat
confint.R2(fit, alternative = 'greater') # Ergibt 0.4122

# Eingabe, Forts.: Typ-II-ANOVA
drop1(fit, test = 'F')

# Ausgabe
    Df Sum of Sq   RSS   AIC F value    Pr(F)
<none>          52252 1624
langarts  1     42989 95241 1812  255.87 <2e-16 ***
daysabs   2      283 52535 1622    0.84   0.43
male      1      345 52597 1624    2.05   0.15

# Eingabe, Forts.: Vorhersage mit Konfidenzintervall
new.data <- data.frame(male = 0, langarts = 50, daysabs = '[0,1]')
predict(fit, new.data, interval = 'c')

# Ausgabe
    fit    lwr    upr
48.343 45.306 51.38

# Eingabe, Forts.: Vorhersage mit Prädiktionsintervall
predict(fit, new.data, interval = 'p')

# Ausgabe
    fit    lwr    upr
48.343 22.659 74.028

```

Kommentare

- *Modellgüte*: Wir schätzen, dass durch die Kovariablen 48.1% der tatsächlichen Varianz der Leistung in Mathematik erklärt werden. Mit einer Sicherheit von rund 95% beträgt der echte Prozentsatz mindestens 41%. Dieser Wert ist grösser als null, somit verwirft der globale F -Test die Nullhypothese “Echtes R^2 ist null” auf dem 5%-Niveau (p -Wert fast null).
- *Typ-II-ANOVA*: Mit einer Sicherheit von rund 95% lässt sich behaupten, dass es einen echten Zusammenhang zwischen ‘Leistung in Language Arts’ und ‘Leistung in Mathematik’ gibt. Für die anderen Merkmale lässt sich dies nicht behaupten.
- *Effekt von ‘Leistung in Language Arts’*: Wir schätzen, dass die wahre Leistung in Mathematik pro Punkt in Language Arts im Schnitt um 0.68 Punkte steigt. Mit einer Sicherheit von rund 95% liegt die tatsächliche Steigerung zwischen 0.60 und 0.77 Punkten. Anhand des t -Tests können wir mit einer Sicherheit von rund 95% behaupten, dass ‘Leistung in Language Arts’ tatsächlich einen Effekt auf die mittlere Leistung in Mathematik hat.
- *Effekt von ‘Geschlecht’*: Wir schätzen, dass Schüler in Mathematik im Schnitt 2.14 Punkte besser ab-

schnieden als vergleichbare Schülerinnen. Mit einer Sicherheit von rund 95% sind sie tatsächlich im Schnitt zwischen -0.80 und 5.1 Punkten „besser“. Da der p -Wert des entsprechenden t -Tests mit 0.15 nicht kleiner als 0.05 ist, kann die Nullhypothese „Es gibt keinen Effekt“ nicht verworfen werden.

- **Effekt von 10 Absenzen:** Wir schätzen, dass Personen mit 10 (bzw. mehr als 5) Absenzen im Schnitt 2.0 Punkte schlechter in Mathematik sind als (vergleichbare) Personen mit höchstens einer Absenz. Mit einer Sicherheit von rund 95% liegt der tatsächliche mittlere Unterschied zwischen -5.6 und 1.6 Punkten. Der Wert 0 befindet sich in diesem 95%-Konfidenzintervall, somit gibt es auf dem 5%-Niveau keinen Grund, die Nullhypothese „Es gibt keinen mittleren Unterschied“ zu verwerfen. Entsprechend ist der p -Wert des t -Tests nicht kleiner als 0.05 .
- **Vorhersagen:** Wir schätzen, dass Schülerinnen ohne Absenzen und 50 Punkten in Sprache im Schnitt $14.128 + 50 \cdot 0.6843 = 48.3$ Punkte in Mathematik erreichen. Mit einer Sicherheit von rund 95% liegt die wahre mittlere Leistung in Mathematik bei solchen Schülerinnen zwischen 45.3 und 51.4 Punkten. Eine konkrete solche Schülerin erreicht mit einer Sicherheit von rund 95% einen Wert zwischen 22.7 und 74.0 Punkten. ▲

7.5 Multikollinearität

Korrierte Kovariablen teilen stets einen gewissen Effekt auf den Mittelwert der Zielgrösse. Dieses Phänomen heisst *Multikollinearität*. Bei einer Gruppe von hochkorrelierten Kovariablen ist deshalb die separate Interpretation der einzelnen Effekte sinnlos. Stattdessen untersucht man ihren gemeinsamen Effekt mit systematischen Vorhersagen und studiert ihren gemeinsamen Beitrag zum R^2 . Gruppen von hochkorrelierten Kovariablen können mithilfe der Korrelationsmatrix der Kovariablen identifiziert werden.

Beispiel 7.29 (Haftpflichtschäden). In Autoversicherungsmodellen wird häufig die mittlere Schadenhöhe u. a. durch hochkorrelierte Autoeigenschaften wie Leistung, Geschwindigkeit, Gewicht und Hubraum modelliert, so dass starke Multikollinearität auftritt. Dies erschwert die Interpretation der Ergebnisse. ▲

Hinweis (Perfekte Multikollinearität). Lässt sich eine Kovariable durch eine lineare Funktion der anderen Kovariablen ausdrücken, liegt perfekte Multikollinearität vor und das Modell kann nicht berechnet werden. Diese Situation tritt beispielsweise auf, wenn man irrtümlicherweise eine Kovariable gleichzeitig in verschiedenen Einheiten (Körpergrösse sowohl in m als auch in cm) verwendet oder wenn eine kategoriale Kovariable mit L Ausprägungen durch alle L statt $L - 1$ Dummyvariablen repräsentiert wird.

Beispiel 7.30 (Immobilien, Fortsetzung). In Beispiel 7.13 ist keine starke Multikollinearität zu erwarten, da es keine hochkorrelierten Kovariablen im Modell gibt:

R Code

```
# Eingabe
round(cor(wohnungen[c('Zimmer', 'Garten', 'Parkett', 'Balkon')]), 2)

# Ausgabe
  Zimmer Garten Parkett Balkon
Zimmer   1.00  0.08 -0.10 -0.15
Garten    0.08  1.00  0.00 -0.27
Parkett   -0.10  0.00  1.00  0.03
Balkon    -0.15 -0.27  0.03  1.00
```

Die negative Korrelation zwischen ‘Garten’ und ‘Balkon’ erklärt übrigens evtl. den negativen Effekt von ‘Balkon’ auf den mittleren Mietpreis. (Warum?) ▲

Beispiel 7.31 (Immobilien, Fortsetzung). In Beispiel 7.19 haben wir den mittleren Mietpreis durch eine kubische Funktion von ‘Zimmer’ erklärt. Wir können ‘Zimmer’, ‘Zimmer²’ und ‘Zimmer³’ als drei Kovariablen auffassen und deren Korrelationsmatrix betrachten:

	R Code		
	Zimmer	Zimmer ²	Zimmer ³
Zimmer	1.000	0.978	0.929
Zimmer ²	0.978	1.000	0.985
Zimmer ³	0.929	0.985	1.000

Kommentar: Wegen den extrem starken Korrelationen tritt im Modell von Beispiel 7.19 extrem starke Multikollinearität auf. Aussagen über einzelne Effekte (deskriptiv und schliessend) sind sinnlos, da beispielsweise eine Veränderung in ‘Zimmer’ automatisch eine Veränderung in ‘Zimmer²’ zur Folge hat. Deshalb arbeitet man mit systematischen Vorhersagen wie in Beispiel 7.19. ▲

7.6 Modellvoraussetzungen und deren Überprüfung

Ein lineares Modell liefert viele Ergebnisse. Welchen man wie gut trauen kann, hängt davon ab, wie gut folgende fünf Voraussetzungen erfüllt sind:

- Passende Modellstruktur
- Gleiche Varianz
- Normalverteilung
- Keine einflussreichen Beobachtungen
- Unabhängigkeit

Wir werden nun überlegen, was diese Voraussetzungen bedeuten, wie man sie überprüft (*Modelldiagnostik*) und was die Konsequenzen sind, wenn Sie klar verletzt sind. Diese Überlegungen gehören grundsätzlich zu jedem linearen Modell.

Bei der einfachen linearen Regression und den Mittelwertvergleichen kann analog wie im allgemeinen Fall vorgegangen werden.

7.6.1 Passende Modellstruktur

Das lineare Modell beruht auf der Annahme, dass die Modellstruktur μ passt bzw. die Modellgleichung

$$E(Y | K) = \mu(K)$$

stimmt. Mit dem Vektor K bezeichnen wir die konkreten Ausprägungen der Kovariablen einer Beobachtung.

Nun wollen wir überlegen, wann diese Annahme erfüllt ist. Dazu zerlegen wir $Y | K$ in den systematischen Teil $\mu(K)$ (Modellstruktur, Signal) und den entsprechenden zufälligen Rest $\varepsilon | K$ (Fehlerterm, Rauschen)

$$Y | K = \mu(K) + \underbrace{Y | K - \mu(K)}_{:= \varepsilon | K} = \mu(K) + \varepsilon | K$$

und bilden auf beiden Seiten der Gleichung den Erwartungswert:

$$E(Y | K) = \mu(K) + E(\varepsilon | K).$$

Die Modellgleichung stimmt also nur, wenn $E(\varepsilon | K)$ für jedes K null ist bzw. nicht von K abhängt¹.

Da man in der Praxis nicht über $E(\varepsilon | K)$ verfügt, lässt sich diese Bedingung nicht direkt überprüfen. Man kann jedoch die Residuen als Realisierungen von $\varepsilon | K$ auffassen und schauen, ob ihr Mittelwert einigermaßen unabhängig von K ist.

Für Modelle mit nur einer Kovariablen kann dazu das Streudiagramm der Residuen und der X -Werte betrachtet werden². Bei mehreren Kovariablen betrachtet man stellvertretend den sogenannten ‘‘Residuals versus Fitted’’-Plot. Dort werden die Residuen den entsprechenden gefitteten Werten (Platzhalter für K) gegenübergestellt. Hängt der Mittelwert der Residuen nicht von den gefitteten Werten ab, so geht man von einer passenden Modellstruktur aus. Abbildungen 7.3 und 7.4 zeigen einige Situationen schematisch.

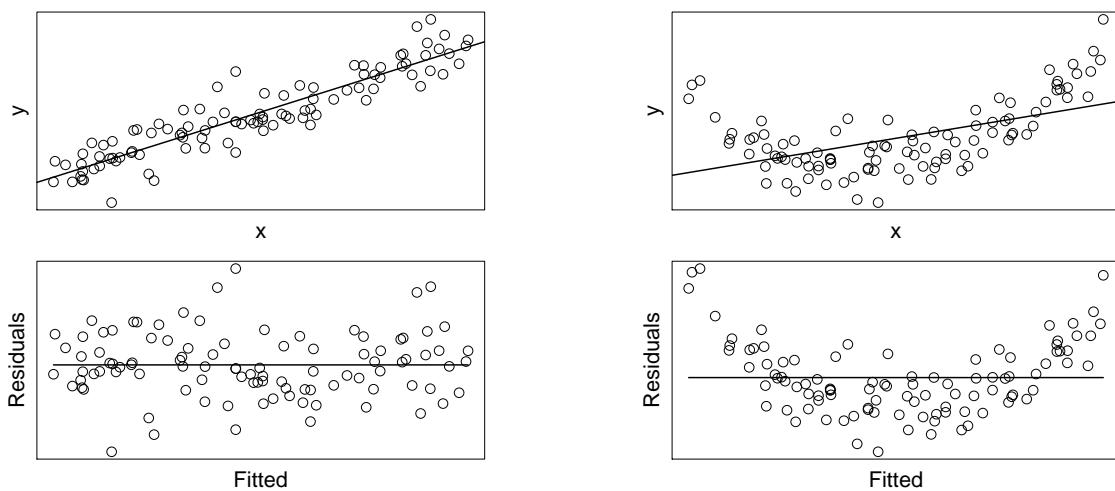


Abbildung 7.3: Streudiagramme inkl. Regressionsgeraden (oben) und entsprechende ‘‘Residuals versus Fitted’’-Plots inkl. Gerade durch null (unten) bei gleicher Varianz. Links eine Situation mit passender, rechts eine mit unpassender Modellstruktur.

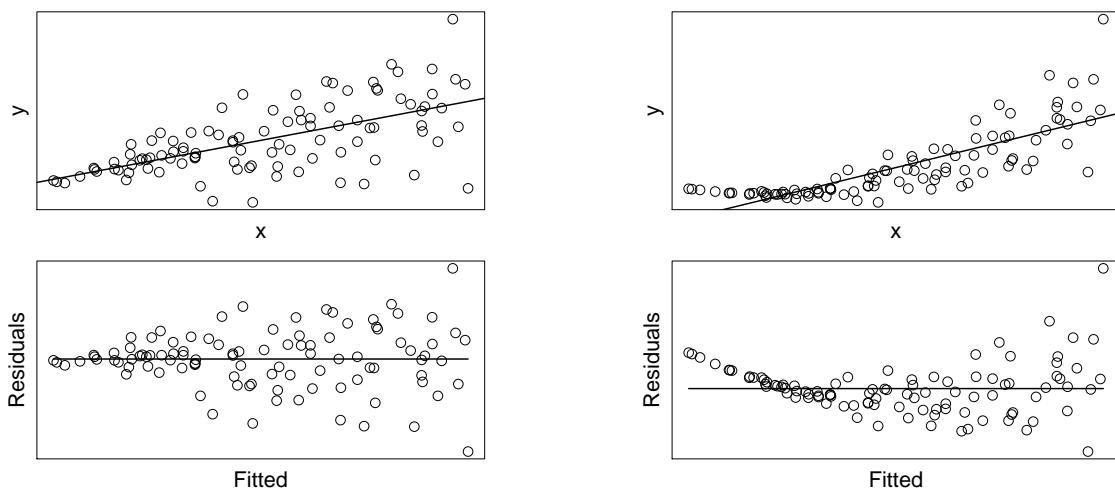


Abbildung 7.4: Streudiagramme inkl. Regressionsgeraden (oben) und entsprechende ‘‘Residuals versus Fitted’’-Plots inkl. Gerade durch null (unten) bei ungleicher Varianz. Links eine Situation mit passender, rechts eine mit unpassender Modellstruktur.

¹Der Intercept wird stets so gewählt, dass $E(\varepsilon) = 0$.

²Bei der einfachen linearen Regression könnte man auch prüfen, ob die Punkte im Streudiagramm der X - und Y -Werte ungefähr um eine Gerade streuen.

Ist die Voraussetzung klar verletzt, so stellt dies sämtliche Ergebnisse in Frage. Beispielsweise sind in den rechten Grafiken von Abbildungen 7.3 und 7.4 die Vorhersagen für zentrale X -Werte systematisch zu hoch, für X -Werte am Rand zu tief.

Manchmal kann das Problem durch eine andere Auswahl der Kovariablen oder durch die Verwendung von Transformationen, Interaktionen oder Nichtlinearitäten verminder werden. Datenbasierte Anpassungen der Modellstruktur sind jedoch heikel, wie wir später sehen werden.

Beispiel 7.32 (Gewicht und Rauchen, Fortsetzung). Im ‘Residuals versus Fitted’-Plot (Abbildung 7.5, linkes Bild) scheint der Mittelwert der Residuen nicht von den gefitteten Werten abzuhängen. Somit erachten wir die Modellstruktur in Beispiel 7.12 als passend. Dort haben wir den Mittelwert von ‘Gewicht’ in Abhängigkeit von ‘Rauchen’, ‘Geschlecht’ und ‘Grösse’ modelliert. ▲

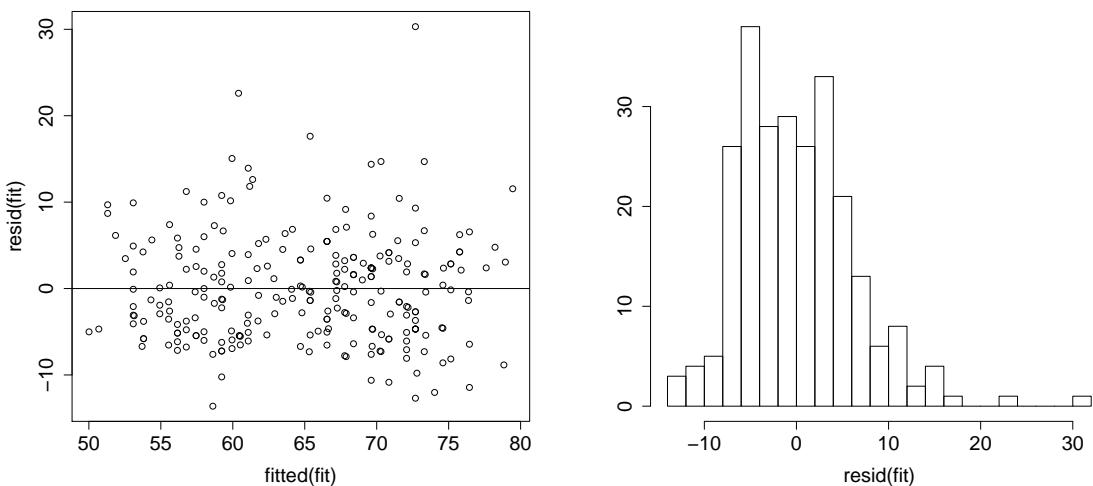


Abbildung 7.5: Residuenplots für Beispiel 7.12 (Gewicht und Rauchen).

7.6.2 Gleiche Varianz

Die Verfahren der schliessenden Statistik beruhen u. a. auf der Voraussetzung, dass die Varianz von $Y | K$ bzw. von $\varepsilon | K$ unabhängig von K , also konstant ist. Diese Forderung heisst *Homoskedastizität* (gleiche Varianz). Ist sie klar verletzt, so spricht man von *Heteroskedastizität* (ungleiche Varianz).

In der Praxis erachtet man die Voraussetzung als erfüllt, wenn die Streuung der Residuen im ‘Residuals versus Fitted’-Plot unabhängig von den gefitteten Werten ist, siehe Abbildungen 7.3 und 7.4.

Ist die Voraussetzung deutlich verletzt, so hängt die Präzision der Vorhersagen und Schätzwerte offensichtlich von den Kovariablenwerten ab. Dies wird bei den entsprechenden Konfidenz- und Prognoseintervallen sowie p -Werten jedoch nicht berücksichtigt, was diese folglich in Frage stellt.

Bilden die Punkte im ‘Residuals versus Fitted’-Plot einen nach rechts geöffneten Trichter, so führt eine Log-Transformation der Zielgrösse manchmal zu einem deutlich besseren Bild.

Hinweis (Bivariate Fälle). Bei der einfachen linearen Regression bedeutet Homoskedastizität, dass die meisten Punkte des Streudiagramms in einem etwa gleich breiten Streifen liegen, bei den Mittelwertvergleichen, dass die Y -Werte pro X -Ausprägung etwa gleich stark streuen.

Beispiel 7.33 (Gewicht und Rauchen, Fortsetzung). Im ‘‘Residuals versus Fitted’’-Plot (Abbildung 7.5, linkes Bild) ist die Streuung der Residuen etwa unabhängig von den gefitteten Werten. Somit zweifeln wir an der Homoskedastizität nicht. ▲

7.6.3 Normalverteilung

Eine weitere notwendige Voraussetzung, damit die Student- und F -Verfahren des linearen Modells exakt sind, ist die Normalverteilung¹ von $Y | K$ bzw. von ε . Dank des Zentralen Grenzwertsatzes stimmen Konfidenzintervalle und p -Werte bei nicht allzu kleinen Stichproben jedoch auch dann approximativ, wenn diese Normalverteilungsvoraussetzung nicht gut erfüllt ist, wenn also z. B. das Histogramm der Residuen nicht normalverteilt aussieht.

Prädiktionsintervalle stimmen hingegen nur, wenn die Voraussetzung erfüllt ist.

Beispiel 7.34 (Gewicht und Rauchen, Fortsetzung). Das Histogramm der Residuen (Abbildung 7.5, rechtes Bild) weist auf eine rechtsschiefe Verteilung hin. Da die Stichprobe jedoch gross ist, würden wir lediglich Prädiktionsintervallen misstrauen. ▲

7.6.4 Keine einflussreichen Beobachtungen

Wie bei der einfachen linearen Regression können einflussreiche Beobachtungen, also solche mit hoher Leverage, den gefürchteten Leverage-Effekt auslösen und damit sämtliche Ergebnisse in Frage stellen.

Einflussreiche Beobachtungen weisen typischerweise starke Ausreisser in den numerischen Kovariablen oder sehr seltene Ausprägungen in kategorialen Kovariablen auf und können deshalb bereits bei der univariaten Analyse identifiziert werden.

Statt eine einflussreiche Beobachtung zu löschen oder die Ergebnisse entsprechend vorsichtig zu beurteilen, werden Ausreisser manchmal durch Logarithmieren der Kovariable oder Abschneiden der Werte auf hohem Niveau entschärft. Seltene Kategorien können zusammengelegt oder zu einer ähnlichen Kategorie dazugeschlagen werden.

Beispiel 7.35 (Gewicht und Rauchen, Fortsetzung). Die univariate Analyse in Beispiel 2.14 zeigt weder Ausreisser in der numerischen Kovariable ‘Körpergrösse’, noch seltene Kategorien bei ‘Rauchen’ oder ‘Geschlecht’. Somit erwarten wir keine einflussreichen Beobachtungen. ▲

7.6.5 Unabhängigkeit

Für die schliessende Statistik ist es stets zentral, dass die Beobachtungen unabhängig sind. Systematische Abhängigkeiten bestehen beispielsweise, wenn pro Objekt mehrere Beobachtungen vorliegen. In solchen Situationen könnte man versuchen, die Abhängigkeiten loszuwerden, indem man alle Beobachtungen des gleichen Objekts zu einer einzelnen Beobachtung ‘komprimiert’ (z. B. durch Differenzen- oder Mittelwertbildung). Ist dies nicht sinnvoll, müssen spezielle Verfahren für abhängige Beobachtungen eingesetzt werden, beispielsweise Zeitreihenmodelle oder Verfahren bei Messwiederholungen.

Beispiel 7.36 (Gewicht und Rauchen, Fortsetzung). Jede Beobachtung repräsentiert eine andere Person, somit gehen wir davon aus, dass die Beobachtungen unabhängig sind. ▲

¹Die Zielgrösse Y muss nicht normalverteilt sein.

Hinweis (Kompakte Formulierung des linearen Modells). Das lineare Modell inkl. Unabhängigkeit, passender Modellstruktur, gleicher Varianz und Normalverteilung kann mathematisch folgendermassen dargestellt werden:

$$Y_i = \mu(\text{Kovariablenwerte von } i\text{-ter Beob.}) + \varepsilon_i,$$

wobei die Fehlerterme $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ unabhängige Zufallsvariablen sind und μ eine lineare Funktion in den Modellparametern ist. Dabei ist jedoch nicht ersichtlich, *inwiefern* die Voraussetzungen wichtig sind. Zudem fehlt die wichtige Voraussetzung, dass es keine einflussreichen Beobachtungen geben sollte.

Beispiel 7.37 (Immobilien, Fortsetzung). Betrachten wir die Situation in Beispiel 7.26. Dort haben wir den mittleren Mietpreis durch die Anzahl Zimmer und drei weitere Eigenschaften modelliert.

- **Passende Modellstruktur:** Erfüllt: Im “Residuals versus Fitted”-Plot (Abbildung 7.6, linkes Bild) hängt der Mittelwert der Residuen nicht von den gefitteten Werten ab.
- **Gleiche Varianz:** Nicht gut erfüllt: Im “Residuals versus Fitted”-Plot (Abbildung 7.6, linkes Bild) ist deutlich zu erkennen, dass die Residuen bei grossen gefitteten Werten klar stärker streuen als bei kleinen (Trichterform). Somit sind p -Werte, Konfidenz- und Prädiktionsintervalle vorsichtig zu verwenden.
- **Normalverteilung:** Erfüllt: Das Histogramm der Residuen (Abbildung 7.6, rechtes Bild) könnte Werte aus einer Normalverteilung zeigen.
- **Keine einflussreichen Beobachtungen:** Erfüllt: Die univariate Beschreibung in Beispiel 2.16 zeigt keine Kovariablen mit Ausreissern oder sehr seltenen Kategorien. Somit erwarten wir keinen starken Leverage-Effekt.
- **Unabhängigkeit:** Erfüllt: Jede Wohnung ist nur einmal im Datensatz vorhanden.

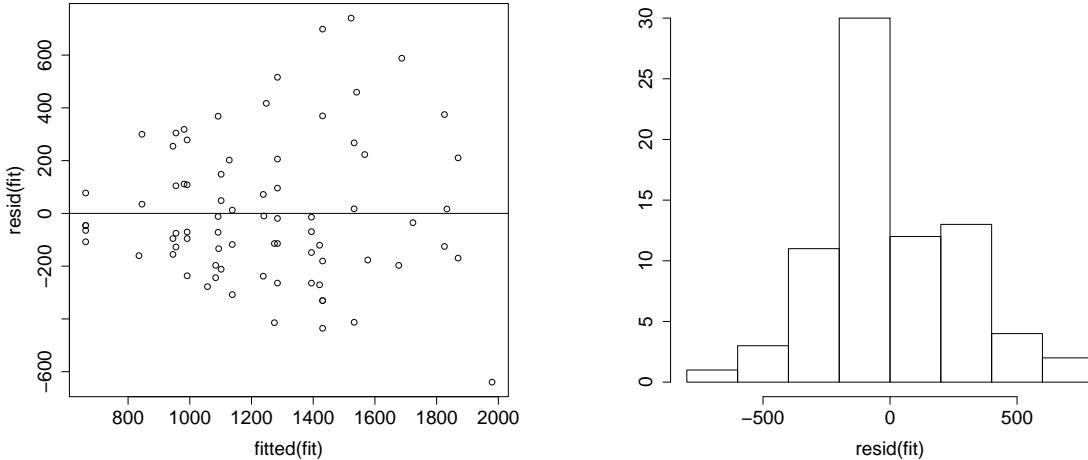


Abbildung 7.6: Residuenplots für Beispiel 7.26 (Immobilien).

Schliesslich betrachten wir die entsprechenden Residuenplots bei logarithmiertem Mietpreis (Beispiel 7.27). (Überlegungen zu Unabhängigkeit und einflussreichen Beobachtungen bleiben gleich.) Im “Residuals versus Fitted”-Plot (Abbildung 7.7, linkes Bild) scheint weder Mittelwert noch Streuung der Residuen von den gefitteten Werten abzuhängen. Damit ist die Modellstruktur passend und die Homoskedastizität gegeben. Das Histogramm (rechtes Bild) könnte ein Hinweis auf leicht rechtsschief verteilte Residuen sein. Da die Stichprobe nicht allzu klein ist, würde dies jedoch allenfalls Prädiktionsintervalle in Frage stellen.

Transformationen sollten jedoch nicht alleine aus diagnostischen Überlegungen eingesetzt werden. ▲

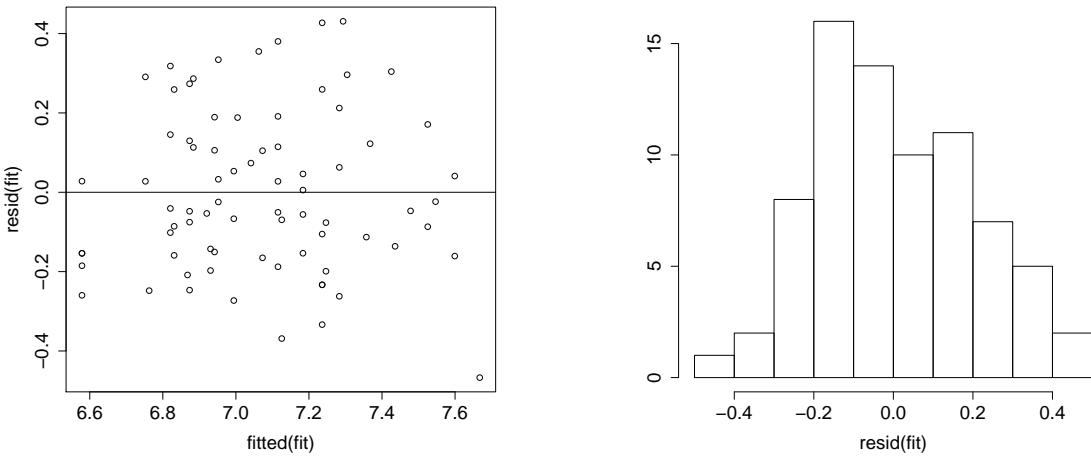


Abbildung 7.7: Residuenplots für Beispiel 7.27 (Immobilien mit logarithmiertem Preis als Zielgröße).

Beispiel 7.38 (Highschool, Fortsetzung). Betrachten wir die Situation in Beispiel 7.28. Dort haben wir die Leistung in Mathematik durch die Leistung in Sprache, das Absenzverhalten sowie das Geschlecht modelliert.

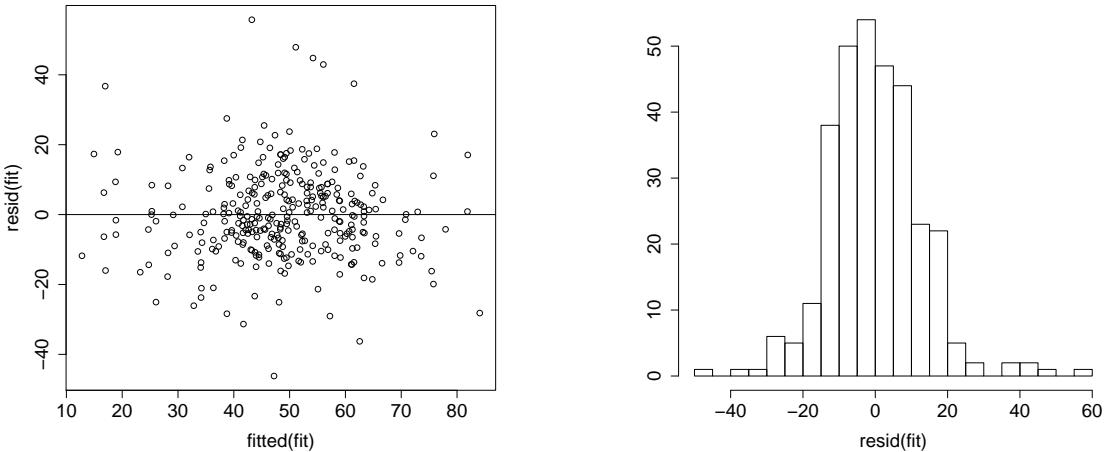


Abbildung 7.8: Residuenplots für Beispiel 7.28 (Highschool).

- *Passende Modellstruktur und gleiche Varianz:* Erfüllt: Im ‘Residuals versus Fitted’-Plot (Abbildung 7.8, linkes Bild) hängt weder Mittelwert noch Streuung der Residuen von den gefitteten Werten ab.
- *Normalverteilung:* Anhand des Histogramms der Residuen (Abbildung 7.8, rechtes Bild) gehen wir davon aus, dass die Normalverteilungsvoraussetzung erfüllt ist.
- *Keine einflussreichen Beobachtungen:* Fast erfüllt: Die univariate Beschreibung in Beispiel 2.15 zeigt keine kategorialen Kovariablen mit seltenen Kategorien. Minimum und Maximum bei ‘Language Arts’ liegen jedoch leicht ausserhalb der Whiskers eines (gedachten) Boxplots. Die entsprechenden Beobachtungen könnten damit einen leichten Leverage-Effekt auslösen.
- *Unabhängigkeit:* Erfüllt: Jede Person ist nur einmal im Datensatz vorhanden. ▲

Zur Normalverteilungsvoraussetzung Schliesslich wollen wir per Simulation illustrieren, dass Regressionskoeffizienten bei nicht allzu kleinem Stichprobenumfang auch für nicht normalverteilte Fehlerterme ungefähr normalverteilt sind und dann die Studentverfahren zumindest approximativ gelten.

Zu diesem Zweck erzeugen wir $B = 1000$ künstliche Datensätze mit je $n = 40$ Beobachtungen einer normalverteilten Kovariablen X und der Zielgröße Y . Die Datensätze unterscheiden sich nur hinsichtlich der Y -Werte: Diese werden als Summe aus wahrem Modell $3 + 0.7 \cdot X$ und Realisierungen eines deutlich rechts-schief verteilten Fehlerterms¹ gebildet. Die Histogramme in Abbildung 7.9 zeigen, dass die Regressionskoeffizienten der 1000 Datensätze ungefähr normalverteilt sind. Zudem sieht man, dass sie um die Werte 3 (wahrer Intercept) und 0.7 (wahre Steigung) streuen.

R Code

```

set.seed(2)                                # Damit Ergebnis reproduzierbar ist

n <- 40                                     # Stichprobenumfang
x <- rnorm(n)                               # Werte der Kovariablen
B <- 1000                                    # Anzahl Simulationen/Datensätze
a <- b <- numeric(B)                         # Regressionskoeffizienten aller Simulationen (leer)

for (i in 1:B)
{
  eps <- rgamma(n, 2, 2)-1      # 40 unabhängige Realisierungen des Fehlerterms
  y <- 3 + 0.7*x + eps          # Y-Werte als Summe aus Modell und Fehlerterm
  fit <- lm(y~x)                # Einfache lineare Regression
  a[i] <- coef(fit)[1]          # Intercept abspeichern
  b[i] <- coef(fit)[2]          # Steigung abspeichern
}

# Histogramme inkl. Hilfslinien
par(mfrow = 2:1)
hist(a)
abline(v = 3)
hist(b)
abline(v = 0.7)

```

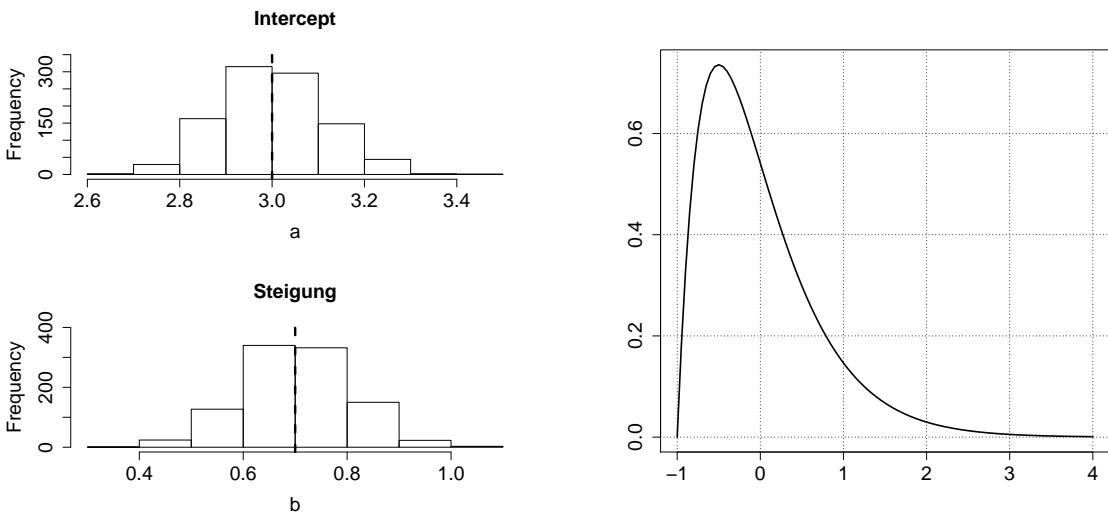


Abbildung 7.9: Linkes Bild: Histogramme der simulierten Regressionskoeffizienten inkl. Hilfslinien bei den wahren Werten. Rechtes Bild: Dichtefunktion des Fehlerterms.

¹Der Fehlerterm ist gammaverteilt mit Erwartungswert 1 und Varianz 0.5 (für unseren Zweck auf Erwartungswert 0 verschoben).

7.7 Weitere statistische Modelle

Wir nennen hier eine (inkomplette) Liste von weiteren statistischen Modellen.

- *Verallgemeinerte lineare Modelle, GLM:* Hier wird eine Funktion – häufig der Logarithmus – des Mittelwerts der Zielgröße modelliert. Diese Klasse von Modellen umfasst u. a. das loglineare Gammamodell (multiplikatives Modell für einen Mittelwert), das binäre Logit-Modell (multiplikatives Modell für die Chancen der Ausprägung ‘1’ eines binären (0-1)-Merkmals) und das binäre Probit-Modell (Modell für Anteile).
- *Quantilregressionen:* Hier wird der Median oder ein anderes Quantil der Zielgröße modelliert. So könnten beispielsweise wichtige Einflussfaktoren für den Value at Risk einer Anlage identifiziert werden. Solche Modelle sind robust/resistent gegenüber Ausreisern in den Kovariablen (kein Leverage-Effekt) und können auch für ordinale Zielgrößen eingesetzt werden.
- *Multivariate Modelle:* Mehrere Zielgrößen werden *gleichzeitig* durch die Kovariablen erklärt. In diese Kategorie fallen neben der multivariaten Kovarianzanalyse (MANCOVA) auch die Vergleiche mehrerer Mittelwerte zwischen zwei Teilstichproben (multivariater *t*-Test bzw. Hotellings *T*-Test) oder zwischen mehreren Teilstichproben (multivariate Varianzanalyse, MANOVA).
- *Zeitreihenmodelle:* Von einem Objekt fallen in zeitlichen Abständen Informationen an, die durch ein Modell erklärt werden sollen, welches die zeitlichen Abhängigkeiten berücksichtigt, z. B. ein Börsenkurs oder die Aarettemperatur.
- *Räumliche Modelle:* Pro Ort (z. B. Gemeinde, Mess-Station etc.) fallen Informationen an, die durch ein Modell erklärt werden sollen, welches die räumlichen Abhängigkeiten berücksichtigt. Liegt zudem ein Zeitaspekt vor, so spricht man von Spatio-Temporal-Models, mit denen beispielsweise Klimadaten analysiert werden.
- *Modelle bei Messwiederholungen oder anderen Abhängigkeiten:* Jeweils mehrere Beobachtungen beziehen sich auf die gleiche Versuchseinheit und sind deswegen nicht unabhängig. Die Daten – oft Paneldaten genannt – bestehen also aus mehreren kurzen Zeitreihen, die mit speziellen Verfahren analysiert werden müssen, beispielsweise mit Mixed-Effects-Modellen.
- *Klassifikationsmodelle:* Modelle für eine kategoriale Zielgröße, z. B. binäre, ordinale oder multinomiale Logit-Modelle oder Diskriminanzanalysen.
- *Strukturgleichungsmodelle, SEM:* Hier liegt ein (Regressions-)Gleichungssystem vor, also mehrere Regressionsmodelle, bei denen die Zielgrößen als Kovariablen in den jeweils anderen Modellen auftreten können.
- *Modelle bei censurierten Zeitdauern:* Bei der Analyse von Zeitdauern tritt oft das Phänomen auf, dass bei einem Teil der Beobachtungen die Zeitdauern zum Zeitpunkt der Analyse noch weiterlaufen (z. B. einige Arbeitsverträge laufen noch). Solche Zeitdauern heißen *censiert*. Entsprechende Modelle tragen z. B. die Namen Cox-Modell oder Survival-Modell.
- *Modelle mit beliebigen nichtlinearen Zusammenhängen:* Anstatt mit einem Polynom hohen Grades können beliebige nichtlineare Zusammenhänge mit speziellen Modellen beschrieben werden, z. B. mit generalized additive models (GAMs), Smoothing Splines oder lokal gewichteten Regressionen.

7.8 Vorbereitung des Modells

Konkrete Fragen mit konkreten Modellen zu beantworten, ist vergleichsweise einfach. Viele reale Fragestellungen sind jedoch oft nur vage formuliert, so dass manchmal nicht klar ist, mit welchem Modell man arbeiten soll und welche Daten man dazu braucht. In diesem Abschnitt beschreiben wir einige wichtige Aspekte dazu.

7.8.1 Daten

Müssen wir die Daten erst beschaffen (z. B. via Umfrage oder einem Auftrag an die Informatikabteilung), so überlegen wir anhand Fragestellung, Literatur und Fachkenntnissen, ob wir an alle wichtigen Merkmale gedacht haben, ob wir sie in der besten Form (bspw. Alter als Zahl, nicht kategorisiert) erhalten werden und ob diese klar definiert und gut zu beschaffen sind (möglichst keine fehlenden Werte).

7.8.2 Wahl der Modellstruktur

Die Kovariablen und allfällige Modifikationen der Modellgleichung sollten *vor* der Modellierung anhand von Fragestellung, Fachkenntnissen, Literatur, univariater und/oder bivariater Analyse der möglichen Kovariablen ausgewählt werden.

Zur Wahl der Kovariablen

Das Modell muss die an der Fragestellung beteiligten Kovariablen sowie allenfalls potenziell wichtige, verfügbare Confounder enthalten.

Manchmal ist die Anzahl der Kovariablen im Verhältnis zu den vorhandenen Daten zu gross. Die Ergebnisse des Modells sind dann zu stark an die konkret vorliegenden Daten gebunden und lassen keine verlässlichen Rückschlüsse auf die Population zu. Dieser unerwünschte Effekt heisst *Overfitting*. Um starkes Overfitting zu vermeiden, dürfen nicht zu viele Parameter geschätzt werden, z. B. nicht mehr als $n/10$. Deshalb muss manchmal die Anzahl der Kovariablen bzw. die Anzahl der zu schätzenden Parameter reduziert werden.

Hier einige Tipps dazu:

- Vermeide fachlich unwichtige Kovariablen.
- Vermeide teuer bzw. schwierig zu erhebende Kovariablen.
- Vermeide Kovariablen mit vielen fehlenden Werten (univariate Analyse beachten).
- Vermeide Kovariablen mit wenig Informationsgehalt (univariate Analyse wichtig), also z. B. kategoriale Merkmale, bei denen fast alle Beobachtungen in der gleichen Kategorie liegen oder numerische Merkmale mit sehr spitzigem Histogramm.
- Fasse seltene Kategorien von kategorialen Kovariablen zu einer einzigen Kategorie zusammen oder schlage sie zur grössten oder ähnlichssten Kategorie dazu (univariate Analyse).
- Ersetze Gruppen von stark korrelierten Kovariablen durch ein einziges Merkmal oder verwende ein anderes Verfahren zur Dimensionsreduktion (siehe Kapitel 8). Hierfür ist eine bivariate Analyse der potenziellen Kovariablen nötig.

In der Praxis werden zum gleichen Zweck häufig leider folgende problematische Verfahren eingesetzt:

- *Manuelle modellbasierte Auswahl*: Die Auswahl der Kovariablen wird anhand der Ergebnisse des Modells immer wieder verändert und das Modell jeweils neu berechnet, bis das Ergebnis schön aussieht.
- *Bivariate “Screening”*: Nur jene Kovariablen, die einen starken bivariaten Zusammenhang mit der Zielgröße aufweisen, werden in das Modell genommen.
- *Backward-Elimination*: Dieses automatische Verfahren berechnet zuerst das Modell mit allen Kovariablen. Danach wird die “unwichtigste” Kovariable eliminiert (z. B. kleinster Beitrag zum R^2) und das Modell ohne dieses Merkmal neu berechnet. Dieser Prozess wird solange wiederholt, bis nur noch “wichtige” Kovariablen übrig bleiben.
- *Forward-Selection*: Diese automatische Methode startet mit einem Modell gänzlich ohne Kovariablen und schliesst dann sukzessive die “besten” Merkmale ein, bis es keine weiteren “wichtigen” mehr gibt.

Da bei diesen Verfahren die Wahl der Kovariablen anhand der Zielgröße getroffen wird, ist das resultierende Modell verfälscht (“man sucht die Eier, die man selbst versteckt hat”), weshalb man solche und ähnliche Verfahren vermeiden sollte – umso mehr, als “statistisch unwichtige” Kovariablen gar nicht stören. Ähnliches gilt, wenn man Modifikationen der Modellgleichung (Transformationen, Interaktionen, Nichtlinearitäten) anhand der Zielgröße auswählt.

Beispiel 7.39 (Backward-Elimination). Wir illustrieren diese Problematik mit einer Backward-Elimination¹ für einen simulierten Datensatz mit 100 Beobachtungen. Die Zielgröße sowie 40 potenzielle Kovariablen enthalten Realisierungen von unabhängigen standardnormalverteilten Zufallsvariablen. Es ist also absolut kein Zusammenhang zwischen den potenziellen Kovariablen und der Zielgröße vorhanden.

R Code

```
# Eingabe
set.seed(10)      # Damit das Ergebnis reproduzierbar ist

Y <- rnorm(100)
X <- data.frame(matrix(rnorm(40*100), ncol = 40))

data <- cbind(Y, X)

fit <- lm(Y ~ ., data = data)
out <- step(fit, direction = 'backward', trace = 0)
summary(out)

# Ergebnis
[...]
  Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.1890    0.0906   -2.09   0.040 *
X6          0.1590    0.0859    1.85   0.067 .
X12         0.2031    0.0822    2.47   0.015 *
X13         0.2086    0.0896    2.33   0.022 *
X18        -0.1367    0.0813   -1.68   0.096 .
X19        -0.2069    0.0934   -2.21   0.029 *
X20         0.1685    0.0909    1.85   0.067 .
X26         0.1339    0.0931    1.44   0.154
X36        -0.1565    0.0856   -1.83   0.071 .

Multiple R-squared:  0.214,    Adjusted R-squared:  0.145
F-statistic: 3.09 on 8 and 91 DF,  p-value: 0.00395
```

¹Mit der R-Funktion `step`.

Kommentar: Obwohl in Tat und Wahrheit überhaupt keine Zusammenhänge bestehen, sieht das Modell vielversprechend aus: Die Variablen ‘X12’, ‘X13’ und ‘X19’ weisen p -Werte unter 0.05 auf und das R^2 beträgt immerhin 0.21. ▲

Eine *seriöse* Möglichkeit, wie man die problematischen Verfahren zur Modellwahl einsetzen kann, ist die folgende: Der Datensatz wird zufällig in zwei Teile geteilt. Mit dem einen Teil wird die optimale Modellstruktur (Wahl von Kovariablen, Transformationen, Interaktionen, Nichtlinearitäten) gesucht, die dann auf den anderen Teil angewendet wird, um valide Ergebnisse zu erhalten. Dieses Vorgehen ist jedoch nur bei grossen Datensätzen möglich. Würden wir im letzten Beispiel ein Modell mit den Kovariablen ‘X12’, ‘X13’ und ‘X19’ auf neue, analog erzeugte Daten anwenden, so wäre das R^2 erwartungsgemäss praktisch null und die drei Kovariablen hätten hohe p -Werte.

7.9 Zusammenfassung

- Sowohl bei der linearen Regression als auch bei den Mittelwertvergleichen haben wir untersucht, wie der Mittelwert eines numerischen Merkmals (Zielgröße) von den Ausprägungen eines anderen Merkmals (Kovariable) abhängt. Regressionsmodelle erweitern und kombinieren diesen Ansatz auf mehrere Kovariablen und zählen entsprechend zu den multivariaten Verfahren. Sie erlauben es, Zusammenhänge zwischen einer Zielgröße und mehreren Kovariablen oder auch den Zusammenhang zwischen zwei Merkmalen unter Berücksichtigung von potenziellen Confoundern zu untersuchen.
- Wir haben uns auf das lineare Modell konzentriert, das wichtigste Regressionsmodell. Dank der linearen Modellstruktur sind dessen Ergebnisse gut interpretierbar. Dieses Modell umfasst wichtige bivariate Spezialfälle und dient zudem als Ausgangslage für weitere Arten von statistischen Modellen, von denen wir einige kurz vorgestellt haben.
- Aussagen über die Stichprobe sind, wie bei den bivariaten Spezialfällen, mit dem R^2 , den Effekten sowie mit Vorhersagen möglich – je nach Zweck der Analyse.
- Die Modellgleichung lässt sich mit Transformationen, Interaktionen und Nichtlinearitäten modifizieren, um manchmal passendere Modelle zu erhalten. Solche Modifikationen erschweren die Interpretation der Ergebnisse zum Teil entscheidend und sollten entsprechend sparsam und vorausblickend eingesetzt werden. Besonderes Augenmerk haben wir auf die wichtige Log-Transformation gelegt, die unter anderem verwendet wird, um multiplikative Modelle zu erhalten.
- Liegt eine Zufallsstichprobe vor, so werden tatsächliche(s) R^2 , Modellparameter und (durch Vorhersagen gefundene) Mittelwerte durch die entsprechenden Werte in der Stichprobe geschätzt und es können Konfidenzintervalle dafür ausgewiesen werden. Die Präzision von individuellen Vorhersagen wird mit Prädiktionsintervallen angegeben. Softwares bieten zudem verschiedene Arten von Hypothesentests an.
- Wir haben die Modellvoraussetzungen und ihre Überprüfung mittels diagnostischer Überlegungen dargelegt und Konsequenzen besprochen, falls sie klar verletzt sind. Die Überprüfung der Modellvoraussetzungen stellt ein wichtiger Teil der Modellierung dar. In diesem Zusammenhang sei auch auf das Phänomen der Multikollinearität verwiesen.
- Im letzten Abschnitt haben wir einige zu beachtende Punkte bei der Vorbereitung des Modells erwähnt. Insbesondere haben wir zwei wichtige Regeln festgehalten: Erstens sollen höchstens $n/10$ Parameter geschätzt werden (vermeidet starkes Overfitting), zweitens sollen Kovariablen nicht anhand der Zielgröße ausgewählt werden (verhindert beschönigte Ergebnisse).

Kapitel 8

Dimensionsreduktion

Zum Leidwesen der StatistikerInnen (und der Befragten) werden auf Fragebögen neben demografischen Angaben wie Geschlecht, Alter, Ausbildung etc. oft sehr viele Items abgefragt, manchmal hunderte. Da statistische Fragestellungen häufig nur vage formuliert sind, sieht man sich bei deren Beantwortung typischerweise mit einer Flut von zu beschreibenden Zusammenhängen konfrontiert (“Unterscheiden sich die Antworten hinsichtlich den demografischen Angaben?”). Ähnliche Items sind meist stark korreliert, was die Interpretation wegen gegenseitigen Confoundings zusätzlich erschwert.

In diesen Situationen helfen Verfahren der *Dimensionsreduktion*: Eine Gruppe von deutlich korrelierten Merkmalen wird durch einige wenige (evtl. neue) Merkmale ersetzt. Anstelle der vielen ursprünglichen Merkmale werden dann diese für die weiteren Analysen verwendet.

Die neuen Merkmale sollten einfach zu interpretieren sein und die ursprünglichen Merkmale möglichst gut repräsentieren (es soll möglichst wenig “Information” verloren gehen).

Eine Dimensionsreduktion ist nicht nur bei der Analyse von Fragebogendaten interessant, sondern kann generell dort nützlich sein, wo Gruppen von deutlich korrelierten Variablen auftreten:

- Man möchte die Effekte von Merkmalen auf die Mittelwerte vieler korrelierter Zielgrößen untersuchen (z. B. eben bei Fragebogendaten).
- Man möchte mit einem Modell die Effekte von hochkorrelierten Kovariablen auf den Mittelwert einer Zielgröße beschreiben. Die damit verbundenen Schwierigkeiten (Interpretation, Multikollinearität) haben wir im letzten Kapitel angesprochen.
- Man möchte die Zusammenhänge zwischen vielen deutlich korrelierten Variablen untersuchen.

In diesem Kapitel präsentieren wir einige klassische Verfahren zur Dimensionsreduktion:

- Wichtigste Variable auswählen
- Summen
- Hauptkomponentenanalyse
- Clusteranalyse

Hinweis (Variablenarten). Die letztgenannten drei Möglichkeiten können nur für numerische Variablen angewendet werden. Werden sie mangels Alternativen dennoch für ordinale Variablen (z. B. Fragebogenitems) eingesetzt, muss klar vermerkt werden, dass die Resultate von der Zahlenkodierung abhängen. Es bietet sich manchmal an, kategoriale Variablen durch Zusammenlegen von Kategorien in binäre (0-1)-Variablen umzuwandeln.

8.1 Wichtigste Variable auswählen

Ein sehr einfaches und naheliegendes Verfahren der Dimensionsreduktion ist das Folgende: Man wählt aus der Gruppe von Variablen eine¹ besonders wichtige (via Literatur, Vorwissen, Fragestellung) Variable aus und repräsentiert damit auch die restlichen Variablen. Die Wahl muss natürlich *vor* der Analyse der Fragestellungen getroffen werden.

Vor-/Nachteile

- + Man muss keine neuen Variablen berechnen.
- + Die Interpretation ist klar.
- + Das Vorgehen funktioniert für beliebige Variablentypen.
- Der Informationsverlust ist schwierig zu quantifizieren.
- Die Auswahl ist häufig willkürlich.

Beispiel 8.1 (Fragebögen). Im Rahmen einer Befragung wurden neben drei demografischen Angaben (Geschlecht, Alter, Einkommen) 50 Aspekte der Zufriedenheit erhoben (Zufriedenheit mit Gesundheit, Regierung, Beruf ...). Anstatt die Zusammenhänge aller $3 \cdot 50 = 150$ Kombinationen von demografischen Merkmalen und Aspekten der Zufriedenheit zu untersuchen, könnte man sich auf die drei Zusammenhänge zwischen den demografischen Merkmalen und dem besonders repräsentativen Item “Wie zufrieden fühlen Sie sich gerade jetzt?” konzentrieren. ▲

Beispiel 8.2 (Autoversicherung). Eine typische statistische Aufgabe bei der Analyse von Autohaftpflichtdaten besteht darin, die Zusammenhänge zwischen Schadenhäufigkeit bzw. Schadenhöhe und Eigenschaften von Auto und LenkerIn zu untersuchen. Prinzipiell könnte man sehr viele Autoeigenschaften als Kovariablen in ein solches Schadenmodell packen. Da Eigenschaften wie Gewicht, Leistung und Hubraum jedoch hochkorreliert sind, wäre die Interpretation der Effekte aufgrund starker Multikollinearität schwierig. Hier ist es sinnvoll, die Autoeigenschaften durch ein einziges Merkmal, z. B. ‘Leistung’, zu vertreten. ▲

8.2 Summen

Ein weiteres naheliegendes Verfahren zur Dimensionsreduktion ist das folgende: Besteht die Variablengruppe aus m vergleichbaren Merkmalen M_1, \dots, M_L , verwendet man manchmal deren (zeilenweise) *Summe* $M_1 + \dots + M_L$ als neues Merkmal für weitere Analysen. Die Ausprägungen werden *Summenscores* genannt.

Beispiele

- Gesamtpunktzahl eines Tests (z. B. IQ-Test, Statistikprüfung)
- Summe von binären (0-1)-Merken (z. B. Anzahl erwähnter Eigenschaften bei Wohnungsinserat)
- Fragebogenitems sind manchmal in mehrere Themengebiete gegliedert. Dann werden die Antworten oft pro Thema zu Teilsummen zusammengezählt.

¹Statt eines einzigen Merkmals können natürlich auch mehrere ausgewählt werden. Einerseits werden die ursprünglichen Merkmale besser repräsentiert bzw. der Informationverlust verkleinert, andererseits verdoppelt sich der Analyseaufwand mindestens.

Vor-/Nachteile

- + Die neue Variable ist einfach zu berechnen.
- + Die Interpretation ist klar.
- Der Informationsverlust ist schwierig zu quantifizieren.
- Die Variablen müssen vergleichbare Skalen aufweisen.
- Alle Variablen werden (vielleicht zu Unrecht) gleich stark gewichtet.

Gewichtete Summen Flexible sind *gewichtete Summen*: Die neue Variable beträgt

$$a_1 \cdot M_1 + \cdots + a_L \cdot M_L,$$

wobei die Gewichte a_1, \dots, a_L fixe Werte darstellen. Mögliche Gewichtungen:

- Alle Gewichte 1: Übliche Summe
- Alle Gewichte $1/m$: Mittelwert der m Merkmale
- Gewichte so, dass die Skalen vergleichbar werden (auch negative Gewichte möglich)
- Gewichte so, dass wichtige Merkmale höher gewichtet werden als unwichtige
- Gewichte anhand Statistik (bspw. Hauptkomponentenanalyse)

Beispiel 8.3 (Highschool, Fortsetzung). Nun verwenden wir den Datensatz von Beispiel 2.15, um die Effekte von ‘Geschlecht’, ‘Schule’ (0 oder 1) und ‘daysabs’ (kategorisierte Anzahl Absenzen) auf die mittlere Leistung zu beschreiben. Dabei betrachten wir folgende verschiedene Zielgrößen:

- ‘math’ (Leistung in Mathematik)
- ‘langarts’ (Leistung in Language Arts)
- ‘summe’ (Gesamtleistung: ‘math’ + ‘langarts’)
- ‘mw’ (Mittlere Leistung: $0.5 \cdot \text{math} + 0.5 \cdot \text{langarts}$ bzw. ‘summe’/2)
- ‘gew.mw’ (Gewichtete Summe bzw. gewichteter Mittelwert: $1/4 \cdot \text{math} + 3/4 \cdot \text{langarts}$)

R Code

```
# Eingabe: Variable erzeugen und Datenausschnitt zeigen
math <- highschool$math
lang <- highschool$langarts
summe <- math + lang
mw <- summe/2
gew.mw <- 0.25*math + 0.75*lang
av <- cbind(math, lang, summe, mw, gew.mw)
head(av)
```

Ausgabe

math	lang	summe	mw	gew.mw
56.989	42.451	99.440	49.720	46.085
37.094	46.821	83.915	41.957	44.389
32.275	43.567	75.842	37.921	40.744
29.057	43.567	72.623	36.312	39.939
6.748	27.248	33.997	16.998	22.123
61.654	48.415	110.069	55.035	51.725

Beispielsweise beträgt der erste Stichprobenwert von ‘summe’

$$56.99 + 42.45 = 99.44,$$

jener von ‘mw’

$$0.5 \cdot 56.99 + 0.5 \cdot 42.45 = 99.44/2 = 49.72$$

und jener von ‘gew.mw’

$$0.25 \cdot 56.99 + 0.75 \cdot 42.45 = 46.09.$$

Nun betrachten wir die fünf entsprechenden Modelle:

R Code

# Eingabe: Fünf separate Modelle berechnen					
fit <- lm(av ~ male + school + daysabs, data = highschool)					
fit					
summary(fit)					
 # Ausgabe					
Coefficients:					
	math	lang	summe	mw	gew.mw
(Intercept)	44.968	49.110	94.078	47.039	48.075
male	-0.807	-5.020	-5.827	-2.914	-3.967
school	11.792	12.123	23.915	11.958	12.040
daysabs(1,5]	-0.497	-2.019	-2.516	-1.258	-1.639
daysabs(5,50]	-4.426	-5.771	-10.197	-5.098	-5.435
[...]					
R-Quadrat	0.148	0.192	0.197	0.197	0.202

Kommentare

- Aufgrund der hohen positiven Korrelation (0.69) zwischen den beiden Leistungen und ihren ähnlichen Verteilungen sind die fünf Modelle sehr ähnlich.
- Weil sich ‘summe’ und ‘mw’ lediglich um einen Faktor unterscheiden, sind deren R^2 genau gleich.
- Die Größenordnung der Effekte bei ‘summe’ ist anders als bei den restlichen Modellen, da sich die Gewichte dort nicht zu eins, sondern zu zwei summieren.
- Aufgrund der additiven Modellstruktur des linearen Modells entspricht der Effekt bei einer gewichteten Summe der gewichteten Summe der Effekte. Beispielsweise ist der Effekt von ‘Geschlecht’ auf den Mittelwert von ‘gew.mw’ $-0.807 \cdot 0.25 - 5.020 \cdot 0.75 = -3.967$. ▲

8.3 Hauptkomponentenanalyse

Die *Hauptkomponentenanalyse* (kurz: *PCA* von engl. *principal component analysis*) ist eines der zentralen Verfahren zur Analyse von mehrdimensionalen Daten, insbesondere zur Dimensionsreduktion von deutlich korrelierten Merkmalen.

Sie beruht auf folgendem Prinzip: m neue Variablen, die sogenannten *Hauptkomponenten* (*PCs*), werden als gewichtete Summen der ursprünglichen – meist auf Mittelwert 0 und Standardabweichung 1 standardisierten – m Variablen gebildet. Die *Gewichte bzw. Ladungen* (engl. *loadings*) sind so gewählt, dass die PCs fallende Varianz (d. h. fallenden Informationsgehalt bzw. fallende Wichtigkeit) haben und unkorreliert sind.

Einige generelle Facts

- Die Bedeutung einer PC ist durch ihre Definition als gewichtete Summe der standardisierten Merkmale gegeben. Die Interpretation geschieht also insbesondere anhand der Loadings.
- Die Loadings werden mit der sogenannten *Eigenwertzerlegung* der Korrelationsmatrix der ursprünglichen Merkmale gefunden – eine Aufgabe, die von Hand nur in ganz einfachen Fällen zu lösen wäre.
- Die Bedeutung einer PC kann mit ihren Stichprobenwerten, den *Scores*, verifiziert werden.
- Die Summe der Varianzen der PCs entspricht der Summe der Varianzen der ursprünglichen (standardisierten) m Merkmale, also m .
- Die PCA ist eng verwandt mit der ebenfalls berühmten *Faktorenanalyse*.

Sind die ursprünglichen Variablen *deutlich* korreliert, so weisen die ersten paar PCs einen Grossteil der ursprünglichen Varianz auf. Dann könnten die weiteren Analysen z. B. auf den ersten paar PCs mit mindestens 75% der ursprünglichen Varianz beruhen.

Vor-/Nachteile von PCA zur Dimensionsreduktion

- + Die Anzahl der neuen Variablen kann objektiv gewählt werden.
- + Der Informationsverlust lässt sich quantifizieren.
- + Die neuen Variablen sind unkorreliert. Die Ergebnisse der weiteren Analysen pro PC sind dadurch leichter zu interpretieren.
- Die Interpretation der Hauptkomponenten ist nicht immer klar und wird durch die Tatsache erschwert, dass es sich um gewichtete Summen von *standardisierten* Merkmalen handelt.

Beispiel 8.4 (Autoversicherung, Fortsetzung). Anhand eines kleinen Datensatzes mit 32 Autos zeigen wir, wie mehrere hochkorrelierte Autoeigenschaften mithilfe einer Hauptkomponentenanalyse optimal durch eine oder zwei neue (unkorrelierte) Merkmale ersetzt werden können. Ein ähnliches Vorgehen ist bei Schadenmodellen, bei denen solche Eigenschaften als potenzielle Kovariablen gelten, sinnvoll. Dadurch wird insbesondere das Problem der starken Multikollinearität gelöst.

Um einen Eindruck über den Datensatz zu erhalten, zeigen wir einige Zeilen davon an. Die Bedeutung der Variablen ist den Namen abzulesen ('Leistung' in kW, 'Hubraum' in ccm).

R Code

```
# Eingabe
auto[c(1, 2, 6, 11, 16, 21, 25, 31, 32),]

# Ausgabe
```

	Zylinder	Leistung	Hubraum	Geschwindigkeit	Tueren	Gewicht
Lotus Omega	6	265	3600	280	4	1885
Mercedes 500 SL	8	240	4973	250	2	1769
Mercedes 300 SE	6	170	3199	220	4	1890
Opel Senator 24V	6	150	2969	235	4	1544
BMW M3	6	210	2956	250	2	1460
Mazda MX-6	6	120	2497	220	2	1195
Toyota Previa	4	96	2438	170	5	1750
Peugeot 106 XSI	4	69	1360	190	3	860
Mini Cooper	4	46	1275	143	2	710

Nun beschreiben wir den Datensatz uni- und bivariat.

R Code

```
# Eingabe
rbind(mean(auto), sd(auto))

# Ausgabe
      Zylinder Leistung Hubraum Geschw. Tueren Gewicht
Mean          5.41    136.09  2678.47   217.78    3.62 1416.06
Standard dev. 1.27     47.08   848.29    27.19    0.94  300.90

# Eingabe
cor(auto)

# Ausgabe
      Zylinder Leistung Hubraum Geschw. Tueren Gewicht
Zylinder        1.00    0.67    0.84    0.50    0.05   0.62
Leistung         0.67    1.00    0.79    0.87    0.08   0.73
Hubraum         0.84    0.79    1.00    0.56    0.05   0.82
Geschwindigkeit 0.50    0.87    0.56    1.00    0.14   0.49
Tueren          0.05    0.08    0.05    0.14    1.00   0.40
Gewicht          0.62    0.73    0.82    0.49    0.40   1.00
```

Die Variablen sind alle deutlich untereinander korreliert (ausser ‘Anzahl Türen’). Wir können somit davon ausgehen, dass Dimensionsreduktion mittels PCA¹ (ohne ‘Türen’) gut funktionieren wird.

Da wir bei der PCA mit standardisierten Merkmalen arbeiten werden, überlegen wir vorher exemplarisch, wie der standardisierte Wert von ‘Leistung’ beim Lotus Omega lautet: Dazu ziehen wir vom Wert 265 die mittlere Leistung 136.09 ab und teilen das Ergebnis durch die Standardabweichung 47.08 der Leistung. Wir erhalten auf diese Weise einen standardisierten Wert von

$$\frac{265 - 136.09}{47.08} = 2.738,$$

was – wie erwartet – auf eine stark überdurchschnittliche Leistung des Lotus Omega hinweist. (Wäre ‘Leistung’ etwa normalverteilt, so würden nur etwa 2.5% der Autos einen standardisierten Wert über 2 haben.)

R Code

```
# Eingabe: PCA ohne Anzahl Türen (fünftes Merkmal)
pc <- princomp(auto[-5], cor = TRUE)
summary(pc, loadings = TRUE)

# Ausgabe
Importance of components:
              Comp.1   Comp.2   Comp.3   Comp.4   Comp.5
Standard deviation   1.94173  0.82459  0.621401  0.315132  0.253564
Proportion of Variance 0.75406  0.13599  0.077228  0.019862  0.012859
Cumulative Proportion  0.75406  0.89005  0.967279  0.987141  1.000000

Loadings:
             Comp.1  Comp.2  Comp.3  Comp.4  Comp.5
Zylinder      -0.432 -0.372  0.684  0.453 -0.039
Leistung       -0.482  0.338 -0.086 -0.120 -0.794
Hubraum        -0.477 -0.338  0.038 -0.768  0.258
Geschwindigkeit -0.403  0.732  0.100  0.111  0.529
Gewicht        -0.436 -0.312 -0.717  0.422  0.146
```

¹In R wird die PCA mit der Funktion `princomp` durchgeführt. Die Standardisierung der Variablen auf Mittelwert 0 und Standardabweichung 1 wird mit der Option `cor = TRUE` erreicht. Das Ergebnis `pc` der Berechnung wird mit `summary(pc, loadings = TRUE)` angezeigt. Die Scores der PCs sind via `pc$scores` verfügbar und können als neue Merkmale an den Datensatz angehängt werden.

Kommentare

- *Wahl der Anzahl PCs bzw. Informationsverlust:* Die Varianz der ersten PC beträgt $1.942^2 = 3.77$. Da die ursprüngliche Varianz fünf beträgt (fünf standardisierte Merkmale mit je Varianz eins), erklärt die erste PC

$$3.77/5 = 0.754 = 75.4\%$$

der ursprünglichen Varianz. Die ersten zwei PCs zusammen erklären sogar

$$(1.942^2 + 0.825^2)/5 = 0.890 = 89\%$$

davon. Wenn wir die fünf technischen Angaben durch die ersten zwei PCs ersetzen, geht also nur wenig Information verloren.

- *Verteilung der PCs:* Ihre Mittelwerte sind null, da sie je gewichtete Summen von *standardisierten* Merkmalen sind. Die Standardabweichungen sind in obigem Output ersichtlich. Aufgrund der Definition sind die PCs unkorreliert.
- *Bedeutung der ersten PC:* Die erste PC entspricht der gewichteten Summe

$$\text{PC1} = -0.432 \cdot \text{Zylinder (standardisiert)} + \dots + (-0.436) \cdot \text{Gewicht (standardisiert)}.$$

Da die Loadings alle ähnlich sind, ist die erste PC etwa proportional zur Summe der standardisierten Merkmale (Pearson-Korrelation zwischen PC1 und Summe der standardisierten Merkmale ist -0.998). Deshalb können wir die erste PC z. B. als ‘Gesamtstärke’ bezeichnen (aufgrund des negativen Vorzeichens der Loadings gehören zu schwachen Autos grosse Scores).

- *Bedeutung der zweiten PC:* Die zweite PC hängt besonders stark von ‘Geschwindigkeit’ ab. Da daneben nur ‘Leistung’ ein Loading mit gleichem Vorzeichen aufweist, könnten wir der zweiten PC z. B. ‘Sportlichkeit’ sagen (je grösser der Score, je sportlicher das Auto).
- *Scores:* Aus den standardisierten Werten und den Loadings können die Scores berechnet werden: Exemplarisch zeigen wir die standardisierten Werte beim Lotus Omega:

R Code				
Zylinder	Leistung	Hubraum	Geschwindigkeit	Gewicht
0.469	2.738	1.086	2.288	1.558

Der Score der ersten PC beträgt beim Lotus Omega

$$-0.432 \cdot 0.469 - 0.482 \cdot 2.738 - 0.477 \cdot 1.086 - 0.403 \cdot 2.288 - 0.436 \cdot 1.558 = -3.64,$$

der Score der zweiten PC entsprechend

$$-0.372 \cdot 0.469 + 0.338 \cdot 2.738 - 0.338 \cdot 1.086 + 0.732 \cdot 2.288 - 0.312 \cdot 1.558 = 1.57.$$

Zur Verifikation der Bedeutung der Hauptkomponenten lassen wir von der Software alle Scores berechnen und betrachten die Beobachtungen mit den extremsten Scores: Der Lotus Omega und der Mercedes 500 SL weisen die kleinsten Scores bei der ersten PC auf, der Mini Cooper und der Peugeot 106 XSI die grössten. Dies passt gut zur Interpretation der ersten PC als ‘Gesamtstärke’ (je stärker, je kleiner der Score).

Bei der zweiten PC weisen der Mini Cooper und der Toyota Previa die kleinsten Scores auf, der Lotus Omega und der BMW M3 die grössten. Dies stützt die Interpretation der zweiten PC als ‘Sportlichkeit’ (je sportlicher, je grösser der Score). ▲

Geometrische Bedeutung der Hauptkomponentenanalyse Wir wollen nun die PCA anhand des Datensatzes aus Beispiel 8.4 aus geometrischer Sicht erklären. Da wir nicht gleichzeitig fünf Merkmale grafisch darstellen können, beschränken wir uns auf die beiden Merkmale ‘Leistung’ und ‘Hubraum’.

Abbildung 8.1 zeigt das Streudiagramm dieser zwei Merkmale einmal für die Originalwerte und einmal für die standardisierten Werte:

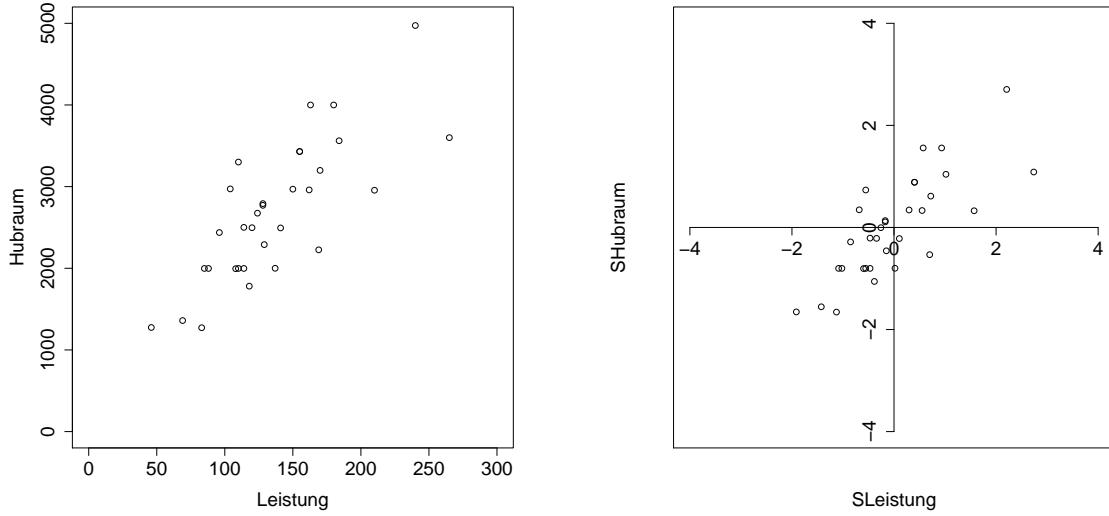


Abbildung 8.1: Streudiagramm für ‘Leistung’ und ‘Hubraum’ (links Originalwerte, rechts standardisierte Werte).

Die PCA sucht nun im Streudiagramm der standardisierten Merkmale die Richtung mit maximaler Streuung. Diese Richtung definiert die erste Hauptkomponente. Die zweite Hauptkomponente steht senkrecht auf die erste. Dies stellt sicher, dass sie unkorreliert mit der ersten ist. Das linke Bild in Abbildung 8.2 zeigt die Richtungen der beiden Hauptkomponenten:

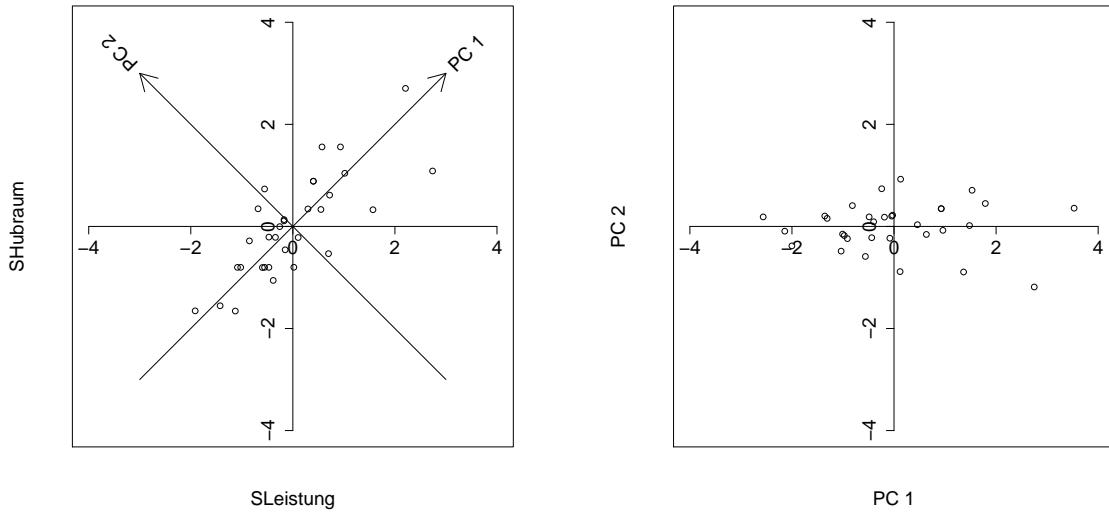


Abbildung 8.2: Richtungen der PCs (links), gedrehtes Koordinatensystem (rechts)

Die Koordinaten der Punkte im (neuen) Koordinatensystem der Hauptkomponenten entsprechen den Scores der Hauptkomponenten. Eine PCA lässt sich also geometrisch als Drehung vorstellen (rechtes Bild von Abbildung 8.2). Die Drehung wird durch die Loadings definiert.

Beispiel 8.5 (Prüfungsergebnisse). 88 amerikanische StudentInnen wurden in fünf verschiedenen technischen Fächern geprüft: In Mechanik, Vektorgeometrie (beide Closed-Book), Algebra, Analysis und Statistik (alle Open-Book). Betrachten wir die Struktur der Daten sowie eine knappe uni- und bivariate Analyse:

R Code					
# Hinsichtlich Gesamtleistung (Summe) beste zwei und schlechteste StudentInnen					
	Mechanics	Vectors	Algebra	Analysis	Statistics
Beste(r)	77	82	67	67	81
Zweitbeste(r)	63	78	80	70	81
Schlechteste(r)	0	40	21	9	14
# Eingabe: Univariate Beschreibung rbind(mean(exam), sd(exam))					
# Ausgabe					
	Mechanics	Vectors	Algebra	Analysis	Statistics
Mean	38.955	50.591	50.602	46.682	42.307
Standard dev.	17.486	13.147	10.625	14.845	17.256
# Eingabe: Bivariate Analyse cor(exam)					
# Ausgabe					
	Mechanics	Vectors	Algebra	Analysis	Statistics
Mechanics	1.00	0.55	0.55	0.41	0.39
Vectors	0.55	1.00	0.61	0.49	0.44
Algebra	0.55	0.61	1.00	0.71	0.66
Analysis	0.41	0.49	0.71	1.00	0.61
Statistics	0.39	0.44	0.66	0.61	1.00

Falls wir mit den Daten Fragen der Art ‘Wie hängen die Leistungen von Eigenschaften wie Geschlecht, Absenzverhalten etc. ab?’ oder ‘Was lässt sich anhand der Leistungen über den künftigen Einstiegslohn sagen?’ beantworten möchten, so bietet es sich an, die Leistungen in den fünf Fächern zu weniger Merkmalen zu komprimieren und dann mit diesen weiterzuarbeiten. Eine solche Dimensionsreduktion erreichen wir beispielsweise, indem wir nur mit dem repräsentativen(?) Merkmal ‘Statistik’, mit der Summe über alle Fächer oder mit den ersten Hauptkomponenten weiterarbeiten. Da die Merkmale untereinander alle deutlich korreliert sind, wäre der Informationsverlust bei allen drei Vorgehensweisen nicht allzu gross.

Konzentrieren wir uns auf das Ergebnis der PCA:

R Code					
# Eingabe pc <- princomp(exam, cor = TRUE) summary(pc, loadings = TRUE)					
# Ausgabe					
Importance of components:					
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	1.7835	0.85998	0.667057	0.622810	0.496579
Proportion of Variance	0.6362	0.14791	0.088993	0.077578	0.049318
Cumulative Proportion	0.6362	0.78411	0.873103	0.950682	1.000000
Loadings:					
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Mechanics	-0.400	0.645	0.621	-0.146	-0.131
Vectors	-0.431	0.442	-0.705	0.298	-0.182
Algebra	-0.503	-0.129	-0.037	-0.109	0.847
Analysis	-0.457	-0.388	-0.136	-0.666	-0.422
Statistics	-0.438	-0.470	0.313	0.659	-0.234

Kommentare

- *Wahl der Anzahl PCs bzw. Informationsverlust:* Die erste PC repräsentiert 64%, die zweite 15% der ursprünglichen Varianz. Da die ersten beiden PCs zusammen einen Grossteil (78%) der ursprünglichen Varianz erklären, ist der Informationsverlust nur gering, wenn wir die fünf Merkmale für weitere Analysen durch die zwei ersten Hauptkomponenten ersetzen.
- *Bedeutung der ersten PC:* Die erste PC hat bei allen Fächern ähnliche Loadings. Sie ist also ungefähr proportional zur Summe der standardisierten Leistungen und – da die Verteilungen der Fächer ähnlich sind – auch ungefähr proportional zur Summe der Leistungen (Pearson-Korrelation zwischen der ersten PC und der Summe ist -0.997). Sie misst also die *Gesamtleistung* (je kleiner der Score, je besser die Leistung). Die standardisierten Werte sowie die PC-Scores der zwei besten und der schlechtesten StudentInnen betragen:

	Mechanics	Vectors	Algebra	Analysis	Statistics	PC1	PC2
Beste(r)	2.176	2.389	1.543	1.369	2.242	-4.310	0.678
Zweitbeste(r)	1.375	2.085	2.767	1.571	2.242	-4.568	-0.215
Schlechteste(r)	-2.228	-0.806	-2.786	-2.538	-1.640	4.545	0.324

Exemplarisch finden wir den Score der ersten PC bei der besten StudentIn mit

$$-0.400 \cdot 2.176 - 0.431 \cdot 2.389 - 0.503 \cdot 1.543 - 0.457 \cdot 1.369 - 0.438 \cdot 2.242 = -4.3,$$

den entsprechenden Score bei der schlechtesten StudentIn mit

$$-0.400 \cdot (-2.228) - 0.431 \cdot (-0.806) - 0.503 \cdot (-2.786) - 0.457 \cdot (-2.538) - 0.438 \cdot (-1.640) = 4.5.$$

Dies bestätigt die zugewiesene Bedeutung der ersten Hauptkomponente.

- *Bedeutung der zweiten PC:* Die zweite PC hat ein positives Loading bei den Closed-Book-Fächern (Mechanics, Vectors) und ein negatives bei den Open-Book-Fächern (Algebra, Analysis, Statistics). Sie könnte also beispielsweise als ‘Leistung in Closed-Book-Prüfungen versus Leistung in Open-Book-Prüfungen’ aufgefasst werden (je höher der Score, je deutlicher überwiegt die standardisierte Leistung in Closed-Book-Prüfungen).

Während die besten zwei StudentInnen erwartungsgemäß einen sehr ähnlichen Score bei der ersten PC haben, unterscheiden sie sich hinsichtlich der zweiten PC deutlich: Tatsächlich ist die beste StudentIn in den Closed-Book-Fächern (Mechanik und Vektorgeometrie) besser und in den Open-Book-Fächern (Algebra, Analysis, Statistik) schlechter als die zweitbeste StudentIn. Dies wiederspiegelt sich in den unterschiedlichen Scores bei der zweiten PC.

- *Zurückrechnen möglich:* Die standardisierten Merkmale lassen sich übrigens als gewichtete Summen der PCs rekonstruieren, wobei (erstaunlicherweise!) wiederum die Loadings als Gewichte dienen. Indem wir die Scores

	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	R Code
	-4.31	0.678	0.124	0.798	-0.517	

aller PCs bei der besten StudentIn mit den entsprechenden Loadings bei ‘Mechanics’ gewichten und zusammenzählen, finden wir (bis auf Rundungsfehler) den standardisierten Wert 2.176 von ‘Mechanics’:

$$-0.400 \cdot (-4.31) + 0.645 \cdot 0.678 + 0.621 \cdot 0.124 - 0.146 \cdot 0.798 - 0.131 \cdot (-0.517) = 2.19$$



8.4 Clusteranalyse

Eine *Clusteranalyse* teilt Beobachtungen in Gruppen bzw. *Cluster* ein. Jeder Cluster enthält Beobachtungen, die hinsichtlich mehrerer (meist standardisierter) numerischer Variablen ähnlich sind. Jeder Cluster wird z. B. durch seinen *Clustermittelwert*, also die Mittelwerte der (standardisierten) Variablen aller Beobachtungen im Cluster, charakterisiert.

Eine solche Einteilung ist in verschiedener Hinsicht interessant: Beispielsweise könnte man Kunden zu Marketingzwecken zu Kundensegmenten gruppieren und dann pro Segment einige Personen zu einem neuen Produkt befragen.

Wir konzentrieren uns hier jedoch auf die *Dimensionsreduktion* via Clusteranalyse: Dazu ersetzen wir die ursprünglichen Merkmale durch die Clustereinteilung, also durch ein neues kategorielles Merkmal.

Vor-/Nachteile zur Dimensionsreduktion

- + Zeigt, welche Beobachtungen ähnlich sind.
- + Die Interpretation der Cluster via Clustermittelwerte ist meist einfach. An letzteren ist beispielsweise zu erkennen, hinsichtlich welcher Variablen sich die Cluster am deutlichsten unterscheiden. Durch Verknüpfung der Clustereinteilung mit den Originaldaten lässt sich die Interpretation verifizieren.
- Das Ersetzen von mehreren numerischen Variablen durch eine kategoriale ist üblicherweise mit einem grossen Informationsverlust verbunden.
- Es gibt sehr viele Arten von Clusteranalysen.

Mithilfe eines Streudiagramms ist es für das menschliche Auge leicht, ähnliche Beobachtungen anhand zweier Merkmale zu gruppieren. Sobald jedoch mehr als zwei Merkmale beteiligt sind, ist dies ein hoffnungsloses Unterfangen und man ist auf entsprechende Software angewiesen. Aus einer Vielfalt von verschiedenen Verfahren präsentieren wir die sogenannte *k-means*-Methode: Die Cluster werden dabei mithilfe der Statistiksoftware durch folgende Schritte gefunden:

1. Der Benutzer legt die Anzahl k der Cluster fest.
2. Das Verfahren startet, indem k Clustermittelwerte willkürlich über den Wertebereich der standardisierten Merkmale verteilt werden.
3. Jede Beobachtung wird dem nächsten Clustermittelwert zugeordnet. Dadurch werden (neue) Cluster gebildet.
4. Die Clustermittelwerte werden neu berechnet.
5. Die Schritte 3 und 4 werden wiederholt, bis die Einteilung stabil ist.

Hinweis (Clusteranalyse mittels PCA). Eine andere, anschauliche Möglichkeit ist die folgende: Mithilfe einer PCA werden die ursprünglichen Merkmale auf zwei neue Merkmale komprimiert. In deren Streudiagramm versucht man von Auge, Gruppen von ähnlichen Beobachtungen zu identifizieren.

Beispiel 8.6 (Prüfungsergebnisse, Fortsetzung). Wir betrachten den Datensatz von Beispiel 8.5 mit Prüfungsergebnissen von 88 SchülerInnen in fünf technischen Fächern. Wir möchten die SchülerInnen mit einer Clusteranalyse¹ anhand ihrer Leistungen in drei Gruppen einteilen und diese Einteilung als neue kategoriale Variable ‘cluster’ im Datensatz speichern. Wir könnten sie zum Beispiel als Kovariable in einem Modell für den Lohn von Amerikanerinnen und Amerikanern 10 Jahre nach Schulabschluss verwenden.

¹Wir benützen zur Berechnung der Clusteranalyse die R-Funktion `kmeans`, die wir auf den standardisierten Datensatz anwenden.

R Code

```
# Eingabe
set.seed(5)      # Damit stets das gleiche Ergebnis entsteht
cl.a <- kmeans(scale(exam), 3)
cl.a

# Ausgabe
K-means clustering with 3 clusters of sizes 20, 45, 23

Cluster means:
  Mechanics    Vectors   Algebra Analysis Statistics
1  0.814667  1.118822  1.1809869  1.02849  1.0949022
2  0.083933 -0.050017  0.0060607  0.12023 -0.0049025
3 -0.872624 -0.875028 -1.0388031 -1.12957 -0.9424969
```

Kommentare: Die 88 Studierenden wurden in drei Gruppen/Cluster vom Umfang 20, 45 und 23 eingeteilt. Der erste Cluster umfasst SchülerInnen mit (im Schnitt) überdurchschnittlichen Leistungen in allen Fächern, der zweite jene mit durchschnittlichen und der dritte schliesslich solche mit unterdurchschnittlichen Leistungen in allen Fächern. Beispielsweise unterscheiden sich Cluster 2 und 3 hinsichtlich allen Fächern ähnlich stark (am schwächsten bei ‘Vectors’, am stärksten bei ‘Analysis’).

Nun hängen wir die Clustereinteilung als neues Merkmal ‘cluster’ an den Datensatz an und zeigen zur Verifikation der Bedeutung der Cluster die Ergebnisse von drei guten, drei mittelmässigen und drei schlechten StudentInnen. Die Clustereinteilung trennt diese drei Gruppen von StudentInnen tatsächlich deutlich:

R Code					
cluster	Mechanics	Vectors	Algebra	Analysis	Statistics
1	77	82	67	67	81
1	63	78	80	70	81
1	75	73	71	66	81
2	40	63	53	54	25
2	23	55	59	53	44
2	48	48	49	51	37
3	12	30	32	35	21
3	5	26	15	20	20
3	0	40	21	9	14

Um die Verwandtschaft zur PCA aufzuzeigen, markieren wir die Punkte im Streudiagramm der ersten zwei PCs von Beispiel 8.5 mit der entsprechenden Clustereinteilung. Im linken Bild von Abbildung 8.3 erkennt man, dass sich die Cluster im Wesentlichen hinsichtlich der ersten PC (Gesamtleistung) unterscheiden. Dies bestätigt zudem die Bedeutung, die wir den Clustern und auch der ersten PC zugewiesen haben.

R Code

```
# Eingabe, Forts.
pc <- princomp(exam, cor = TRUE)
PC1 <- pc$scores[,1]
PC2 <- pc$scores[,2]

plot(PC2 ~ PC1, pch = as.character(cl.a$cluster))
```



Beispiel 8.7 (Autoversicherung, Fortsetzung). Wir betrachten den Datensatz von Beispiel 8.4 mit technischen Angaben von 32 Autos und möchten die Autos in zwei Kategorien einteilen. Diese kategoriale Variable könnte z. B. als binäre Kovariable in einem Schadenmodell verwendet werden. Der Informationsverlust gegenüber der ersten Hauptkomponente oder der Variable ‘Leistung’ wäre jedoch gross.

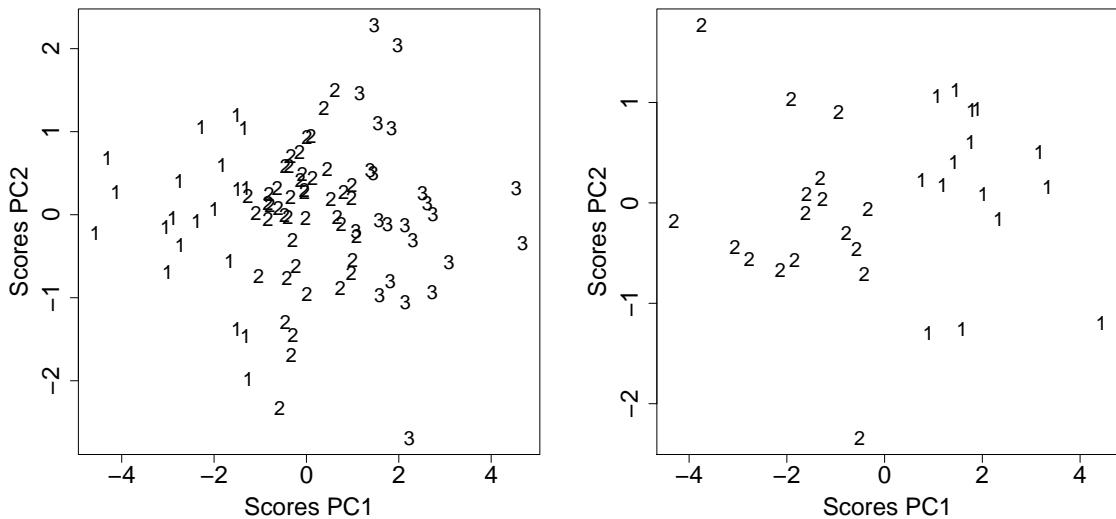


Abbildung 8.3: Linkes Bild: Streudiagramm der ersten beiden Hauptkomponenten von Beispiel 8.5 mit der Clustereinteilung aus Beispiel 8.6 (Prüfungsergebnisse). Rechtes Bild: Streudiagramm der PCs von Beispiel 8.4 inkl. Clustereinteilung von Beispiel 8.7 (Autoversicherung).

R Code

```
# Eingabe
set.seed(10) # Damit stets das gleiche Ergebnis entsteht
cl.a <- kmeans(scale(auto), 2)
cl.a

# Ausgabe
K-means clustering with 2 clusters of sizes 15, 17

Cluster means:
  Zylinder Leistung Hubraum Geschwindigkeit Tueren Gewicht
1 -0.79456 -0.73698 -0.77050      -0.58767 -0.23889 -0.72891
2  0.70108  0.65028  0.67985       0.51853  0.21078  0.64315
```

Kommentare: Die zwei Cluster umfassen 15 bzw. 17 Autos. Anhand der Clustermittelwerte ist zu erkennen, dass der erste Cluster generell die schwächeren Autos enthält, der zweite Cluster die stärkeren. Am wenigsten unterscheiden sich die beiden Teilstichproben in der mittleren Anzahl Türen – dieses Merkmal trennt die Autos also am wenigsten gut, ist also am wenigsten wichtig für die Clusterung. Bei den anderen standardisierten Variablen ist ein deutlicher mittlerer Unterschied festzustellen. (Den grössten bei Zylinder.)

Nun verifizieren wir die Bedeutung der Cluster via Clustereinteilung von vier exemplarischen Autos. Tatsächlich werden die beiden betrachteten schwächeren Autos (Alfa 155 Turbo, Mazda MX-6) dem ersten Cluster zugeordnet, die beiden stärkeren Autos (Mercedes 500 SL, Lexus LS 400) dem zweiten Cluster:

R Code

```
# Eingabe, Forts.
auto$Cluster <- cl.a$cluster
auto[c(2:3, 20:21), ]
  Zylinder Leistung Hubraum Geschwindigkeit Tueren Gewicht Cluster
Mercedes 500 SL     8     240    4973        250      2    1769      2
Lexus LS 400        8     180    4000        245      4    1690      2
Alfa 155 Turbo      4     137    2000        222      4    1210      1
Mazda MX-6          6     120    2497        220      2    1195      1
```

Das rechte Bild von Abbildung 8.3 zeigt das Streudiagramm der ersten beiden PCs von Beispiel 8.4 (ohne ‘Türen’), bei dem die Clustereinteilung hervorheben wird. Man sieht auch hier die enge Verbindung zwischen der PCA und der Clusteranalyse. Zudem bestätigt das Bild die Bedeutungen, die wir den Clustern und auch der ersten PC zugewiesen haben.

R Code

```
# Eingabe, Forts.
pc <- princomp(auto[-5], cor = TRUE)
PC1 <- pc$scores[,1]
PC2 <- pc$scores[,2]

plot(PC2 ~ PC1, pch = as.character(cl.a$cluster))
```



8.5 Zusammenfassung

- Eine typische statistische Aufgabe besteht darin, Zusammenhänge zwischen Fragebogenitems und den erfassten demografischen Angaben zu untersuchen. Da oft viele Items abgefragt werden und diese in der Regel deutlich korrelieren, ist eine solche Analyse mühsam und schwierig zu interpretieren. Dann bietet es sich an, die Items auf wenige (evtl. neue) Merkmale zu reduzieren und dann den Zusammenhang zwischen diesen und den demografischen Angaben zu analysieren. Wir haben einige Verfahren besprochen, die eine solche Dimensionsreduktion ermöglichen. Sie bieten sich generell dann an, wenn Gruppen von hochkorrelierten Merkmalen auftreten, beispielsweise auch bei hochkorrelierten Kovariablen in einem Modell (Multikollinearität).
- Wir haben vier Verfahren zur Dimensionsreduktion kennengelernt: Zwei naheliegende Möglichkeiten (Auswahl der wichtigsten Variable und Summen) sowie zwei rein statistische Verfahren: Die Hauptkomponentenanalyse (PCA) und die Clusteranalyse. Diese zwei multivariaten Verfahren haben verschiedene Einsatzzwecke, wobei wir uns auf die Dimensionsreduktion konzentriert haben.
- Die Hauptkomponentenanalyse ersetzt eine Gruppe von standardisierten Merkmalen durch ebenso viele neue Merkmale, die Hauptkomponenten. Diese sind gewichtete Summen der ursprünglichen (standardisierten) Merkmale. Die Gewichte (Loadings) sind so gewählt, dass die Hauptkomponenten unkorreliert sind und fallende Varianz, d. h. fallende Wichtigkeit, aufweisen. Je deutlicher die ursprünglichen Merkmale korrelieren, desto besser werden sie durch die ersten paar wenigen Hauptkomponenten repräsentiert und desto besser funktioniert die Dimensionsreduktion per PCA. Die Bedeutung der Hauptkomponenten wird anhand der Loadings spezifiziert und kann anhand der Stichprobenwerte (Scores) der Hauptkomponenten verifiziert werden. Die PCA ist eng verwandt mit der Faktorenanalyse.
- Eine Clusteranalyse gruppiert Beobachtungen, die hinsichtlich einiger numerischer Merkmale ähnlich sind, zu Clustern. Die Cluster lassen sich anhand der Clustermittelwerte beschreiben. Clusteranalysen können zur Dimensionsreduktion eingesetzt werden, indem die numerischen Merkmale durch die Clustereinteilung (eine kategoriale Variable) repräsentiert werden. Einer von vielen möglichen Algorithmen ist das k-means-Verfahren, das von einer fixen Anzahl Clustern ausgeht und dann nach einer optimalen Clustereinteilung sucht.

Anhang A

R-Skript

A.1 Übersicht

R wurde von R. Ihaka und R. Gentleman am Statistics Department of the University of Auckland entwickelt und wird vom *R Development Core Team* laufend erweitert. Mittlerweile sind unzählige Zusatzpakete verfügbar, die von Forschenden entwickelt wurden und die Funktionalität von R bei Bedarf ausbauen.

A.1.1 Fenster

Startet man R (genauer: das Programm `Rgui.exe`), so trifft man im Hauptfenster des Programms folgende Fenster an:

- *Console*: Dieses Fenster erscheint unmittelbar beim Start. Es enthält zunächst einige Meldungen von R (z. B. den Copyright-Vermerk) und die Eingabeaufforderung `>`, hinter der man Befehle eintippen kann. Auch der Output erscheint in diesem Fenster; die eingegebenen Befehle und der Output werden farblich unterschieden.
- *Graphics Device*: Grafiken werden in einem eigenen Fenster dargestellt, das sich automatisch öffnet, sobald ein entsprechender Befehl ausgeführt wird.
- *Editor*: Häufig ist es praktischer, die Befehle in ein Programm (ein sogenanntes *Script*) zu schreiben, als sie bei der Eingabeaufforderung einzutippen. Dies ermöglicht es insbesondere, ganze Programmabschnitte nochmals auszuführen, nachdem man z. B. einen Fehler darin korrigiert hat. Ein Editor-Fenster mit einem leeren Programm erhält man mit `File ▶ New script`.

Daneben gibt es noch eigene Fenster für die Hilfe (vgl. Abschnitt A.1.3).

Hinweis. Der Inhalt der Menüs und der Symbolleiste ändert sich je nach aktivem Fenster.

A.1.2 Bedienungssprache

In den neueren Versionen bietet R für die Menüs usw. verschiedene Sprachen an. Standardmäßig wird die Sprache des Betriebssystems verwendet. Leider ist die deutsche Fassung bisher nicht besonders vollständig, so dass zwischendurch immer wieder englische Wörter auftauchen. Indem man beim Start von `Rgui.exe` den Parameter `LANGUAGE=en` angibt (lässt sich auch bei der Verknüpfung auf dem Desktop bzw. im Startmenü ergänzen), erhält man die durchgehend englische Originalversion, auf die sich auch dieses Skript bezieht.

A.1.3 Dokumentation

Die Hilfe zu bestimmten Funktionen erhält man, indem man bei der Eingabeaufforderung `help(funktion)` oder `?funktion` eingibt, oder im Console-Fenster über **Help ▶ R functions (text) . . .**. Dies klappt auch mit Operator-Zeichen, die man allerdings in (einfache oder doppelte) Anführungszeichen setzen muss (z. B. gelangt man mit `help('+')` zur Hilfe zu den arithmetischen Operatoren). Im Help-Menü finden sich unter anderem noch ein zusätzliches Hilfesystem, das in einem Browser geöffnet wird (`Html help`), und diverse Handbücher im PDF-Format (`Manuals (in PDF) ▶ . . .`).

Im Internet gibt es zahlreiche weitere Hilfestellungen unter <http://www.r-project.org> (z. B. “Introduction to R” und “R Reference”). Dort finden sich auch Links zum Herunterladen der Software.

A.1.4 Zusatzpakete

R ist modular aufgebaut. Standardmäßig werden beim Start von R neben dem Grundpaket (`base`) schon einige weitere Packages geladen (z. B. `stats`, `graphics`). Einige weitere sind installiert, müssen aber vor der Verwendung mit `library(paketname)` geladen werden, damit die darin enthaltenen Funktionen zur Verfügung stehen. Zusatz-Packages können aus dem Internet heruntergeladen werden; derart heruntergeladene zip-Dateien kann man (am besten mit Administrator-Rechten) über **Packages ▶ Install package(s) from local zip Files . . .** am richtigen Ort entpacken.

Informationen zu den Packages erhält man mit `library(help = package.name)`.

A.1.5 Daten

R erlaubt einen wesentlich flexibleren Umgang mit Daten als andere Statistikprogramme: Neben Datentabellen, wie sie etwa auch in SPSS zur Verfügung stehen, kann man (wie in Programmiersprachen wie z. B. C++) beinahe beliebige Arten von Daten-Objekten verwenden, also einzelne numerische Werte oder Zeichenketten, Vektoren, Matrizen, Arrays, Listen usw., aber auch selbst definierte Klassen von Objekten. Auf einige dieser Möglichkeiten werden wir zurückkommen.

Der Name einer Datentabelle (eines sogenannten *Data Frame*), aber auch eines beliebigen anderen Objekts, kann aus Buchstaben, Ziffern und dem Punkt bestehen, wobei am Anfang des Namens keine Ziffer erlaubt ist. Gross- und Kleinschreibung wird unterschieden.

Die in R verwendeten Objekte gehen beim Beenden des Programms verloren, wenn man sie nicht explizit abspeichert. Dazu kann man beispielsweise einzelne Datentabellen in Dateien abspeichern. Man kann aber auch über **File ▶ Save Workspace . . .** den aktuellen Zustand von R (d. h. alle vorhandenen Objekte) abspeichern und später wieder laden.

A.1.6 R als Taschenrechner

R lernt man am besten, indem man es benützt – zum Beispiel als Taschenrechner. Tabelle A.1 zeigt einige Rechenoperationen. Beispiele:

R Code
<code>3*4 + 5</code>
<code># ergibt 17</code>
<code>3*(4 + 5)</code>
<code># ergibt 27</code>
<code>exp(1)</code>
<code># ergibt die Eulersche Zahl 2.71...</code>
<code>log(100, 10)</code>
<code># ergibt den Zehnerlogarithmus von 100, also 2</code>
<code>sqrt(abs(-10))</code>
<code># ergibt 3.1623</code>

<code>x + y; x - y; x*y; x/y</code>	Addition; Subtraktion; Multiplikation; Division
<code>sqrt(x); x^y; exp(x)</code>	Wurzel; Potenz; "e hoch"
<code>log(x); log(x, a)</code>	Natürlicher Log; Log zur Basis a
<code>sin(x); cos(x); tan(x)</code>	Trigonometrische Funktionen
<code>asin(x); acos(x); atan(x)</code>	Inverse trigonometrische Funktionen
<code>abs(x); sign(x)</code>	Absolutbetrag; Vorzeichen
<code>factorial(x); choose(x, y)</code>	Fakultät; "tief"
<code>ceiling(x); floor(x); round(x, a)</code>	Aufrunden; Abrunden; auf a Stellen runden
<code>pi; exp(1)</code>	Die Konstanten π und e
<code>gamma(x); beta(x, y)</code>	Gamma- und Betafunktion
<code>x %% y; x %/% y</code>	Modulo; ganzzahlige Division

Tabelle A.1: Einige R-Operatoren/-Funktionen.

Die Berechnungen werden dabei im Konsolenfenster eingetippt und mit der Return-Taste abgeschickt.

Hinweis. Runde Klammern legen die Evaluationsreihenfolge fest.

A.2 Programmierung

R basiert auf der Programmiersprache S. Deshalb kann man in vielen Fällen bis auf kleine Anpassungen die gleichen Programme verwenden wie für den anderen weit verbreiteten Dialekt von S, das kommerzielle Paket S-PLUS. Die Sprache S wiederum gleicht bekannten Programmiersprachen wie C/C++ oder Java.

Befehle kann man einerseits direkt an der Eingabeaufforderung eingeben, so dass man immer gleich sieht, was sie auslösen. (Wenn die eingegebene Anweisung syntaktisch unvollständig ist, wenn also z. B. eine schliessende Klammer fehlt, dann erscheint nach dem Drücken der Enter-Taste statt dem > ein +, und man kann die Anweisung vervollständigen.)

Andererseits kann man Programme in einem Editor-Fenster (File ▶ New script) zusammenstellen und dann als Ganzes ausführen lassen. Dieses Vorgehen ist zu empfehlen, sobald man den Code nicht einfach hinschreiben kann, sondern schrittweise zusammensetzen muss: Beim direkten Arbeiten mit der Kommandozeile kann man zwar die früheren Eingaben mit den Pfeiltasten wieder hervorholen und bearbeiten, aber eben immer nur zeilenweise.

Aus einem Editor-Fenster kann man markierten Code mit Ctrl-R oder F5 (bzw. Edit ▶ Run line or selection) ausführen lassen.

Bevor wir uns systematisch mit der Programmiersprache auseinandersetzen, betrachten wir ein einfaches Beispiel eines R-Programms:

Beispiel A.1. Der folgende Code erzeugt ein Balkendiagramm mit absoluten Häufigkeiten von genannten Zufallsziffern (Abbildung A.1).

R Code	_____
# Anzahl der genannten Zufallsziffern (Befragung von Studierenden)	
<code>zz <- c(0, 1, 2, 3, 4, 5, 6, 7, 8, 9)</code>	
<code>Anzahl <- c(8, 6, 12, 32, 25, 23, 28, 70, 41, 17)</code>	
<code>barplot(Anzahl, names = zz, xlab='Ziffer', main='Anzahl genannte Zufallsziffern')</code>	

Die beiden Zeilen nach dem Kommentar erzeugen ein Objekt namens `zz` mit den Ziffern von 0 bis 9 und ein Objekt namens `Anzahl` mit der entsprechenden Anzahl der Antworten. Anschliessend wird die

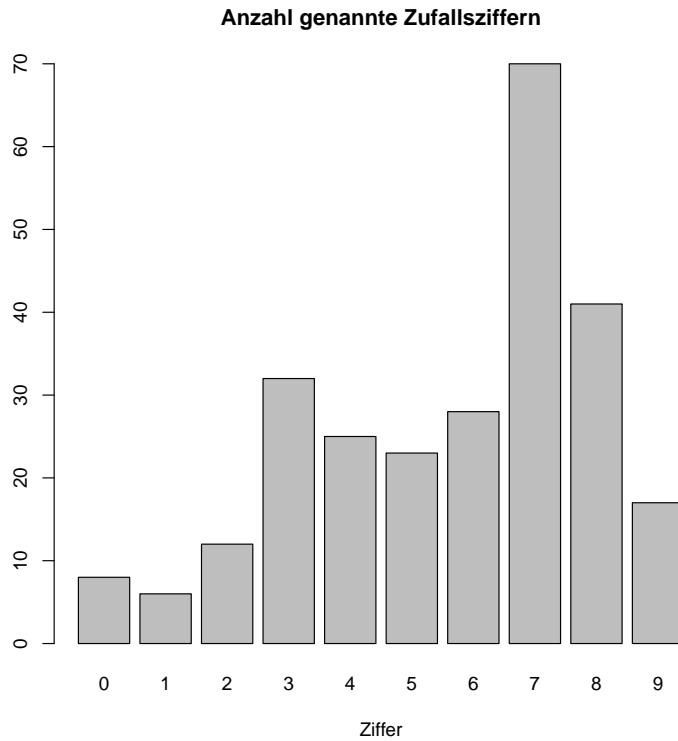


Abbildung A.1: Balkendiagramm zum Beispiel A.1.

Funktion `barplot` aufgerufen, die mit diesen Objekten arbeitet und im Aufruf einige weitere Angaben zur Beschriftung erhält. ▲

A.2.1 Objekte

Im Abschnitt A.1.5 wurde bereits angedeutet, dass man in R Daten nicht nur in Tabellen (Data Frames) ablegen kann, sondern auch allerlei andere Arten von Variablen bzw. Objekten verwenden kann, wie es auch in anderen Programmiersprachen üblich ist. Die Vielfalt dieser Objekte ist nicht leicht zu überblicken; für unsere Bedürfnisse sollte eine Einteilung der Datenobjekte nach den folgenden zwei Kriterien genügen.

Die üblichen Datenobjekte kann man zunächst nach dem Typ der enthaltenen Werte unterscheiden, u.a.:

- `numeric`: Zahlenwerte, als Dezimalbruch bzw. in wissenschaftlicher Schreibweise dargestellt
- `factor`: Werte einer kategorialen Variable
- `ordered`: Werte einer ordinalen Variable. Geordneter Faktor
- `logical`: boolesche Werte (TRUE/FALSE)
- `character`: Zeichenketten

In jedem dieser Typen von Objekten können auch fehlende Werte (NA = not available) enthalten sein. In numerischen Objekten sind ausserdem die speziellen Werte `Inf` (infinity, z. B. $1/0$), `-Inf` (z. B. $-1/0$) und `NaN` (not a number, z. B. $0/0$) möglich.

Als zweites Unterscheidungsmerkmal gibt es verschiedene Formen, wie mehrere Werte zu einem Objekt zusammengesetzt werden können. Die wichtigsten solchen Formen sind:

- Einzelwerte bzw. Vektoren mit mehreren Werten des gleichen Typs
- **matrix**: rechteckige Anordnung von Werten des gleichen Typs
- **array**: Verallgemeinerung von Vektoren/Matrizen auf mehr als zwei Dimensionen
- **data.frame**: rechteckige Tabelle mit Daten eines Typs pro Spalte
- **list**: Sammlung beliebiger Objekte, evtl. unterschiedlichen Typs und/oder unterschiedlicher Länge

A.2.2 Zuweisung, Anzeige von Objekten

Wir müssen Objekten gewisse Inhalte zuweisen können und diese Inhalte auch wieder anzeigen können. Die Zuweisung geschieht mit dem Operator `<-` (bestehend aus einem “kleiner”-Zeichen und einem Bindestrich), der bereits im einführenden Beispiel vorkam. Den Inhalt eines Objekts kann man anzeigen, indem man an der Eingabeaufforderung dessen Namen eingibt.

Beispiel A.2. Erzeugen wir einige Objekte und zeigen sie anschliessend an:

R Code

```
# Eingabe: einzelner numerischer Wert
x <- 1
x

# Ausgabe
[1] 1

# Eingabe: boolescher Vektor
x <- c(TRUE, TRUE, FALSE)
x

# Ausgabe
[1] TRUE TRUE FALSE

# Eingabe: numerische Matrix, aus zwei (Zeilen-)Vektoren zusammengesetzt
X <- rbind(c(1, 2, 3), c(4, 5, 6))
X

# Ausgabe
[,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6

# Eingabe: Data Frame mit boolescher, numerischer und character-Variable
df <- data.frame(b = c(TRUE, FALSE, TRUE), x = c(1, 2, 3),
                  y = c('a', 'bc', 'def'))
df

# Ausgabe
      b   x   y
1  TRUE  1   a
2 FALSE  2   bc
3  TRUE  3 def

# Eingabe, Forts.: Liste mit booleschem Vektor, numerischem Einzelwert
# und der numerischen Matrix X von oben
```

```

li <- list(b = c(TRUE, FALSE), x = 1, X = X)
li

# Ausgabe
$b
[1] TRUE FALSE

$x
[1] 1

$X
 [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6

```



A.2.3 Funktionen

Auch Funktionen sind eine spezielle Art von Objekten – solche, die typischerweise übergebene Parameter verwenden, um ein Resultat zu berechnen. Die übergebenen Parameter können feste Werte, andere Objekte oder auch kompliziertere Ausdrücke sein, z. B. Resultate anderer Berechnungen. Das Resultat einer Funktion ist wiederum ein Objekt.

Funktionen werden ebenfalls mit dem Zuweisungsoperator definiert:

R Code

```

funktionsname <- function(parameter.1 = wert.1, ..., parameter.n = wert.n)
{
  code der funktion
}
```

`parameter.i` sind die Namen der Parameter, die man der Funktion übergeben können will. Oft gibt es für einen Teil der Parameter häufig verwendete “Standardwerte”; gibt man diese als `wert.i` in der Funktionsdefinition an, so braucht man beim Aufruf der Funktion nur davon abweichende Werte anzugeben.

Wenn man an der Eingabeaufforderung nur den Namen einer Funktion eingibt, so wird deren Definition angezeigt. Gibt man dagegen nach dem Namen in Klammern die benötigten Parameter an, so wird die Funktion auf diese Parameter angewendet und das resultierende Objekt (das letzte Resultat im `code der funktion`) angezeigt. Dieses resultierende Objekt kann man auch wiederum mit dem Zuweisungsoperator benennen oder direkt als Parameter für eine weitere Funktion verwenden.

In der R-Hilfe werden die Funktionen wie in der Funktionsdefinition dargestellt. Dort bedeutet also etwa `f(a, b=0)`, dass die Funktion `f` zwei Parameter mit den Namen `a` und `b` hat. Ein Wert für `a` muss beim Aufruf von `f` immer angegeben werden, während `b` einen Standardwert (0) hat – nur davon abweichende Werte muss man angeben.

Beispiel A.3. Der folgende Code und der dazugehörige Output enthalten Beispiele zur Verwendung und Definition einfacher Funktionen.

R Code

```

# Anzeige der Funktionsdefinition: log hat einen zwingend notwendigen
# Parameter x und einen optionalen Parameter base

# Eingabe
log
```

```

# Ausgabe
function (x, base = exp(1)) .Primitive("log")

# Eingabe: Anwendung der Funktion auf feste Werte
log(100)

# Ausgabe
[1] 4.60517

# Eingabe
log(100, 10)

# Ausgabe
[1] 2

# Eingabe: Abspeichern des Resultats
a <- log(100)

# Eingabe: direkte Verwendung des Resultats in einem weiteren Funktionsaufruf
exp(log(100))

# Ausgabe
[1] 100

# Eingabe: Definition und Aufruf einer eigenen Funktion
doppeltes <- function(x)
{
  2 * x
}

doppeltes(10)

# Ausgabe
[1] 20

```



R verwendet bei der Zuordnung der übergebenen Werte zu den Parametern in der Definition einer Funktion sowohl die Reihenfolge der Werte als auch allenfalls mitgegebene Parameternamen. So berechnet `log(100)` das gleiche wie `log(x=100)`, `log(base=exp(1),x=100)` oder gar `log(base=exp(1),100)`. Es ist aber zu empfehlen, in Funktionsaufrufen höchstens bei den ersten Parametern die Namen wegzulassen.

A.2.4 Syntax-Regeln

Der Grundbaustein für R-Programme ist der sogenannte *Ausdruck*. Ein Ausdruck ist eine Anleitung zur Bestimmung eines Werts – meist durch direkte Angabe einer Konstanten, durch Angabe eines Objektnamens oder durch Angabe einer Funktion, die aus ebenfalls angegebenen Argumenten einen Wert berechnet. Wie bei anderen Programmiersprachen kann man überall, wo man einen konstanten Wert einsetzen kann, auch einen beliebigen Ausdruck verwenden, der diesen Wert berechnet. Eine Zuweisung ist immer noch ein Ausdruck – mit dem Wert, der hinter dem Zuweisungsoperator berechnet worden ist: Der Ausdruck `a <- 2` hat z. B. den Wert 2.

Aufrufe von Funktionen sind auch dann Ausdrücke, wenn sie nicht primär einer Berechnung dienen; beispielsweise gibt die Funktion `barplot` einen Vektor mit Koordinatenwerten zurück, die man zur Ergänzung

des Diagramms verwenden kann, die aber nicht angezeigt werden. Gibt eine Funktion tatsächlich “nichts” zurück, dann tut sie dies mit dem Wert `NULL`.

Ausdrücke können mit Semikola voneinander abgetrennt werden. Meist verwendet man dafür aber einfach Zeilenumbrüche.

Zeichenketten können in einfachen oder doppelten *Anführungszeichen* angegeben werden – diese müssen paarweise zusammenpassen.

Innerhalb eines Ausdrucks (aber nicht in einer Zeichenkette) darf man immer dann einen *Zeilenumbruch* einfügen, wenn aus dem davor stehenden Teil klar ist, dass der Ausdruck noch nicht vollständig ist. Dies ist etwa dann der Fall, wenn eine schliessende Klammer fehlt oder am Schluss der Zeile ein Zeichen-Operator steht. *Leerzeichen* darf man ausserhalb von Zeichenketten und Objektnamen praktisch beliebig einfügen, um den Code übersichtlicher zu gestalten.

Zur *Gruppierung von Ausdrücken* dienen geschweifte Klammern `{}`, etwa bei der Definition einer Funktion.

Bei den *Namen von Objekten* (also praktisch überall) wird zwischen Gross- und Kleinschreibung unterschieden. Neben Buchstaben sind nur Ziffern und der Punkt erlaubt, und ein Name kann nicht mit einer Ziffer beginnen. (In aktuellen R-Versionen ist in Namen auch der “`_`” erlaubt, dieser hatte aber in früheren Versionen eine andere Bedeutung.)

Der Rest der Zeile nach einem `#`-Zeichen ist *Kommentar* und bleibt von R unbeachtet.

A.2.5 Programmfluss

In R-Code kann man an beliebigen Stellen Verzweigungs- und Schleifenkonstruktionen verwenden:

- `if (bedingung) ausdruck.1 else ausdruck.2`: Der `ausdruck.1` wird nur ausgewertet, wenn die `bedingung` erfüllt ist. Sonst wird ggf. der `ausdruck.2` ausgewertet (`else ausdruck.2` ist optional). Jeder dieser Ausdrücke kann auch aus mehreren Ausdrücken zusammengesetzt sein – in diesem Fall sind die Ausdrücke mit geschweiften Klammern zusammenzufassen. Unmittelbar vor dem `else` ist kein Zeilenumbruch zulässig, da bereits ein vollständiger Ausdruck gegeben ist.
- `for (index in wertemenge) ausdruck`: Für jedes Element in der `wertemenge` wird nacheinander der `ausdruck` ausgewertet, in welchem das aktuelle Element aus der `wertemenge` mit `index` angesprochen werden kann. Normalerweise ist die `wertemenge` ein Vektor, also etwa `for (i in 1:n){...}`. Der Wert des ganzen `for`-Ausdrucks ist der Wert des `ausdrucks` im letzten Schleifendurchgang.

Manchmal soll mit `n` Schlaufen ein leerer Vektor der Länge `n` gefüllt werden. Er kann vor der Schlaufe mit `Vektor <- numeric(n)`, `Vektor <- character(n)` oder `Vektor <- logical(n)` erzeugt werden – je nach gewünschtem Datentyp. Leere Matrizen, Listen bzw. Data Frames werden auf ähnliche Art und Weise mit `matrix(nrow = n, ncol = m)`, `vector("list", n)` bzw. `data.frame()` kreiert.

A.3 Datenaufbereitung

A.3.1 Direkte Eingabe in einen Data Frame

Wir haben die Möglichkeit, die benötigten Daten direkt im Programmcode zu speichern (statt in einer separaten Datei, die eingelesen wird). Bereits im Beispiel A.2 wurde ein Data Frame erzeugt; die Syntax lautet:

R Code

```
tabellenname <- data.frame(
  var.name.1 = c(wert.11, ..., wert.1n),
  var.name.2 = c(wert.21, ..., wert.2n),
  ...
  var.name.p = c(wert.p1, ..., wert.bn))
```

Natürlich kann man auch bereits bestehende Vektoren zu einem Data Frame zusammenfassen.

Übergibt man zur Bildung eines Data Frame Zeichenketten-Variablen, so werden diese standardmäßig in Faktoren umgewandelt, was insbesondere bei Vergleichen zu Problemen führen kann. Will man dies vermeiden, so muss man den übergebenen Vektor in die Funktion `I()` einschliessen, also z.B.

```
data.frame(name = I(c('Aberegg', 'Abplanalp', 'Aebersold'))).
```

Die Option `stringsAsFactors = TRUE` der Funktion `data.frame` verhindert die Umwandlung in Faktoren für alle Variablen im Datensatz.

A.3.2 Einlesen eines Data Frame aus einer Datei

Zum Einlesen von Daten aus Dateien gibt es u. a. folgende Funktionen:

- `read.delim(Pfad_und_Dateiname)`: Mit dieser Funktion lassen sich verschiedene Text-Dateiformate leicht einlesen, bei denen pro Zeile die Werte der verschiedenen Variablen durch ein bestimmtes Zeichen getrennt sind. Vorgesehen ist die Funktion primär für Tabulator-getrennte Daten (dies entspricht dem Standardwert `sep='\t'`). Das Trennzeichen lässt sich aber umdefinieren: Mit `sep=';'` kann man die csv-Dateien lesen, die Excel mit den üblichen Schweizer Ländereinstellungen erzeugt (d. h. Semikolon als Trennzeichen zwischen den Spalten, Dezimalpunkt).

Vorsicht bei Pfadangaben unter Windows: Statt dem sonst verwendeten Backslash (\) muss man den (Vorwärts-)Schrägstrich (/) als Trennzeichen zwischen den Ordnernamen setzen.

Stehen die Variablennamen nicht in der Datei, so gibt man dies mit `header=FALSE` an. Gibt man in diesem Fall keine Variablennamen vor, so heissen die Variablen V1, V2 usw.

Wie beim direkten Erzeugen eines Data Frame werden Zeichenketten als Faktoren betrachtet. Mit `as.is=TRUE` kann man dies vermeiden. Datums- und Zeitangaben werden in der Regel zunächst als Faktoren eingelesen und müssen anschliessend umgewandelt werden, damit R sie richtig interpretieren kann (vgl. z. B. `?as.Date`).

`read.delim` ruft die Funktion `read.table` auf, die man auch direkt verwenden kann. Die Standardeinstellungen für `read.delim` sind aber meist praktischer.

- Die Funktion `read.fwf` liest Dateien ein, bei der die Werte der gleichen Variable in jeder Zeile genau untereinander stehen, bei denen die Felder also eine feste Breite haben.
- In der Library `foreign` gibt es verschiedene Funktionen, um Dateien anderer Programme zu importieren, z. B. SPSS. Hinweise zum Import von Excel- und vielen weiteren Dateien findet man unter Help ▶ Manuals (in pdf) ▶ R Data Import/Export. Am einfachsten ist es aber häufig, Dateien im Ursprungssprogramm (z. B. Excel) in einem passenden Format (z. B. Tab-getrennt oder csv) abzuspeichern und dann so ins R zu importieren.

<code>min(a.1, ..., a.n)</code>	Minimum
<code>max(a.1, ..., a.n)</code>	Maximum
<code>cummin(a)</code>	Kumulatives Minimum von Vektoren
<code>cummax(a)</code>	Kumulatives Maximum von Vektoren
<code>pmin(a.1, ..., a.n)</code>	komponentenweises Minimum von Vektoren
<code>pmax(a.1, ..., a.n)</code>	komponentenweises Maximum von Vektoren
<code>sum(a); prod(a)</code>	Summe bzw. Produkt der Elemente eines Vektors
<code>cumsum(a); cumprod(a)</code>	Kumulative Summe bzw. Produkt eines Vektors
<code>mean(a); median(a)</code>	Mittelwert bzw. Median eines Vektors
<code>var(a); sd(a)</code>	Varianz bzw. Standardabweichung eines Vektors
<code>length(a)</code>	Anzahl Elemente eines Vektors
<code>dim(a); nrow(a); ncol(a)</code>	Dimension; Anzahl Zeilen bzw. Spalten einer Matrix
<code>paste(A.1, ..., A.m)</code>	Fügt mehrere Objekte zu einer Zeichenkette zusammen
<code>a == b; a != b; a > b</code>	Vergleichsoperatoren
<code>a < b; a >= b; a <= b</code>	
<code>a %in% b</code>	“enthalten in”
<code>a & b; a b; !a</code>	logisches Und; logisches Oder; Negation
<code>is.na(a)</code>	Welche Einträge sind NA?
<code>na.omit(a)</code>	Entfernt fehlende Werte
<code>unique(a)</code>	Alle Einträge nur einmal
<code>union(a); intersect(a,b)</code>	Vereinigung; Schnittmenge
<code>setdiff(a,b)</code>	Elemente in a, die nicht in b enthalten sind
<code>a[i]</code>	$i > 0$: i-tes Element eines Vektors $i < 0$: Vektor ohne das -i-te Element
 	i Vektor: Auswahl/Weglassen mehrerer Elemente
<code>rbind(a.1,...,a.n)</code>	i boolescher Vektor: Auswahl nach Bedingung
<code>cbind(a.1,...,a.n)</code>	Bildung einer Matrix aus untereinander gesetzten Vektoren/Matrizen
<code>a[i, j]</code>	Bildung einer Matrix aus nebeneinander gesetzten Vektoren/Matrizen
<code>a[[i]]</code>	Matrix-Element mit den Koordinaten (i, j) bzw. Untermatrix
<code>a\$name</code>	i-tes Element einer Liste
<code>t(a); a %*% b</code>	benanntes Element einer Liste bzw. Spalte eines Data Frame
<code>det(a); eigen(a); svd(a)</code>	Matrix-Transposition; Matrix-Multiplikation Determinante; Eigenwerte und Eigenvektoren; Singulärwertzerlegung einer Matrix

Tabelle A.2: Weitere R-Operatoren/-Funktionen

A.3.3 Transformieren eines Data Frame

Zum Transformieren von Daten braucht man vor allem die in Tabelle A.2 eingeführten Auswahloperatoren für Matrizen und Listen – ein Data Frame kann nämlich sowohl als Matrix als auch als Liste (deren Einträge die Vektoren der einzelnen Variablen sind) angesprochen werden.

- *Zeilenauswahl (d. h. Auswahl von Beobachtungen)*

```
neue.daten <- alte.daten[i, ]
```

Die Bedeutung von *i* ist gleich, wie dies in Tabelle A.2 für die Auswahl aus einem Vektor erläutert wurde. Man kann also einfach mit der Nummer der Zeile eine Zeile auswählen (bzw. mit einem Vektor von Zeilennummern mehrere Zeilen). Mit einer negativen Zahl schliesst man die entsprechende Zeile aus; wiederum können mehrere negative Zahlen angegeben werden.

Oft noch viel nützlicher ist die Möglichkeit, einen booleschen Vektor anzugeben, d. h. einen Vektor aus den Werten TRUE und FALSE. Die Länge dieses Vektors sollte der Anzahl Beobachtungen entsprechen. In diesem Fall werden die Zeilen ausgewählt, für die der zugehörige Eintrag im Vektor TRUE ist. Natürlich wird man die booleschen Werte in aller Regel nicht direkt eingeben, sondern berechnen lassen.

Man beachte, dass in den eckigen Klammern nach dem Komma nichts angegeben ist – dies bedeutet, dass an die Spalten keine Bedingung gestellt wird, und somit werden alle Spalten ausgewählt.

Als Alternative kann die `subset`-Funktion verwendet werden, die als erstes Argument den Datensatz enthält und als zweites die Auswahl der Zeilen spezifiziert. Mit dem dritten Argument `select` können übrigens auch Spalten ausgewählt werden.

Um die ersten/letzten sechs Zeilen eines Datensatzes (oder eines anderen Objekts) anzuzeigen, kann kurz mit den Befehlen `head` und `tail` gearbeitet werden. Mit dem optionalen zweiten Argument `n` kann die Anzahl der angezeigten Zeilen geändert werden.

- *Spaltenauswahl (d. h. Auswahl von Variablen)*

```
neue.daten <- alte.daten[, j]
```

Die Auswahl von Spalten funktioniert völlig analog wie diejenige von Zeilen. Da man ein Programm oft nochmals braucht, nachdem man neue Variablen eingeführt hat, ist es allerdings etwas gefährlich, die Variablen nach deren Position auszuwählen. Besser ist die Verwendung der Variablennamen: Eine einzelne Spalte kann man mit dem Listen-Auswahloperator als `alte.daten$var.name` ansprechen. Mehrere Variablen bekommt man mit `alte.daten[, c('var.name.1', ..., 'var.name.n')]`.

- *Hinzufügen einer neuen Variable*

```
daten$var.name <- ausdruck
```

Man erweitert den Data Frame um eine weitere Spalte, indem man den Auswahloperator für Listen auf der linken Seite der Zuweisung verwendet. Existiert der Variablenname schon, so wird diese Variable überschrieben.

Der `ausdruck` ist ein Vektor passender Länge und kann natürlich auf anderen Variablen des gleichen Data Frame beruhen. Solche Berechnungen werden dadurch erleichtert, dass viele Operatoren (z. B. die Funktionen in Tabelle A.1 und die `Vergleiche`) sich auch komponentenweise auf Vektoren anwenden lassen.

Mit `transform` können gleichzeitig mehrere neue Spalten erzeugt werden. Das erste Argument ist der ausgehende Datensatz, die weiteren Argumente spezifizieren die neuen Merkmale.

Eine numerische Variable *x* lässt sich mit `cut(x, breaks = c(a.1, ..., a.m))` in einen geordneten Faktor umwandeln. An dieser Stelle sei auf einige Funktionen zur Typenumwandlung verwiesen:

`as.character(x)` wandelt den Vektor `x` (z. B. ein Faktor oder eine zahlenkodierte kategoriale Variable) in einen Vektor von Zeichenketten um, `as.numeric(x)` wandelt `x` (z. B. ein boolscher Vektor) in einen Zahlenvektor um. Mit

```
daten$var.name.b <- as.numeric(daten$var.name > median(daten$var.name))
```

kann beispielsweise eine numerische Variable anhand ihres Medians dichotomisiert werden.

Vektoren werden mit `factor` in Faktoren bzw. mit `ordered` in geordnete Faktoren umgewandelt.

Mit der Funktion `scale(daten$x)` wird der Zahlenvektor `x` auf Mittelwert 0 und Varianz 1 skaliert. Diese Funktion kann mehrere Variablen gleichzeitig verarbeiten.

- *Ersetzen einzelner Werte*

Mit den Auswahloperatoren kann man auch auf der linken Seite der Zuweisung einzelne bestehende Elemente auswählen, die ersetzt werden sollen.

Ein typisches Beispiel ist, dass man in einer Variable in bestimmten Zeilen einen neuen Wert einsetzen will:

```
daten$var.name[i] <- ausdruck
```

`i` kann dabei wiederum Zeilenummern oder boolesche Werte enthalten.

- *Zeilen bzw. Spalten löschen*

Man wählt die beizubehaltenden Zeilen bzw. Spalten aus und rechnet damit weiter. Soll nur eine einzelne Spalte gelöscht werden, wird dies manchmal mit

```
daten$var.name <- NULL
```

erledigt.

- *Zeilen mit fehlenden Werten löschen*

Dies geschieht am einfachsten mit `daten.ohne.NA <- na.omit(daten)`.

- *Kombination von Data Frames*

Data Frames mit den gleichen Variablen lassen sich mit `rbind(df.1, df.2)` untereinander zusammenfügen. Um Data Frames nebeneinander anzuordnen, deren Zeilen “zusammenpassen”, kann man `cbind(df.1, df.2)` verwenden; will man die Zeilen dagegen anhand übereinstimmender Werte einer oder mehrerer gemeinsamer Variablen richtig zusammenfügen, so braucht man die Funktion `merge` – vgl. R-Hilfe.

- *Restrukturierung von Data Frames*

Mit den Funktionen `reshape`, `stack` und `unstack` können verbundene Stichproben vom “breiten” Format ins “lange” überführt werden und umgekehrt. Die Anwendung dieser Funktionen ist nicht ganz einfach zu beschreiben. Am besten konsultiert man die R-Hilfe, sobald man sie einsetzen muss.

- *Aggregierung von Data Frames*

Zur Aggregierung von Daten (d. h. mehrere Zeilen werden zu einer einzelnen verdichtet) gibt es die Funktion `aggregate`.

Der schematische Aufruf

```
aggregate(daten, kat.var, mean, na.rm = TRUE)
```

liefert beispielsweise den Mittelwert der Variablen in `daten` in Abhängigkeit von `kat.var`. Besteht `daten` nur aus einer Spalte, so liefert die Funktion `tapply` manchmal kompaktere Ergebnisse.

Beispiel A.4. Dieses Beispiel zeigt einige der genannten Vorgehensweisen.

R Code

```

# Eingabe
x <- c(10000, 8000, 7200, 6500, 6000)
y <- c("m", "f", "f", "f", "m")
manager <- data.frame(Einkommen = x, gender = I(y))
manager

# Ausgabe
  Einkommen gender
1      10000      m
2       8000      f
3       7200      f
4       6500      f
5       6000      m

# Nur Frauen auswählen
manager.2 <- manager[manager$gender == "f", ]

# Variable "Einkommen" logarithmieren:
manager.2$log.Einkommen <- log(manager.2$Einkommen)

# Variable "Einkommen" dichotomisieren:
manager.2$Einkommen.gross <- as.numeric(manager.2$Einkommen > 7500)
manager.2

# Ausgabe
  Einkommen gender log.Einkommen Einkommen.gross
2       8000      f     8.987197            1
3       7200      f     8.881836            0
4       6500      f     8.779557            0

# Nun das gleiche mit subset und transform:
manager.3 <- subset(manager, gender == "f")
manager.3 <- transform(manager.3, log.Einkommen = log(Einkommen),
                      Einkommen.gross = as.numeric(Einkommen > 7500))
manager.3

# Ausgabe
  Einkommen gender log.Einkommen Einkommen.gross
2       8000      f     8.987197            1
3       7200      f     8.881836            0
4       6500      f     8.779557            0

```



A.3.4 Erzeugen von Daten

Grundsätzlich kann man einen ganzen Data Frame mit einer `for`-Schleife und der Funktion `rbind` zeilenweise erzeugen. Effizienter und oft auch praktischer ist es, wenn man die Spalten auf einmal erzeugen kann. Dazu sind u. a. folgende Funktionen nützlich:

- `a:b`

Vektor mit den Werten `a, a + 1, ..., b`

- `seq(a, b, delta)`

Etwas allgemeiner als mit `a:b` kann man mit dieser Funktion einen Vektor mit den Werten `a, a + delta, ..., b` erzeugen. Statt `delta` (oder `b`) kann mit der Option `length.out = n` die Anzahl `n` der Komponenten angegeben werden.

- `rep(a, times=k)`

Diese Funktion erzeugt einen Vektor, in dem `k` mal die Werte (oder der einzelne Wert) von `a` hintereinander stehen. Ist `a` ein Vektor, so kann auch `k` ein Vektor der gleichen Länge sein, dessen Einträge für jeden Wert von `a` die Anzahl Wiederholungen angeben. Ist `k` eine einzelne Zahl, so kann man mit dem zusätzlichen Argument `each = m` zusätzlich angeben, dass jeder Wert `m` mal wiederholt werden soll, bevor die Werte hintereinander gesetzt werden.

Beispiele:

R Code

```
rep(1, 3)      # [1] 1 1 1
rep(1:3, 3)    # [1] 1 2 3 1 2 3 1 2 3
rep(1:3, 1:3)  # [1] 1 2 2 3 3 3
rep(1:3, each=3) # [1] 1 1 1 2 2 2 3 3 3
rep(1:3, 2, each=3) # [1] 1 1 1 2 2 2 3 3 3 1 1 1 2 2 2 3 3 3
```

- `apply(X, richtung, funktion)`

Will man eine Funktion eines Vektors (z. B. `sum`) zeilen- bzw. spaltenweise auf eine ganze Matrix `X` anwenden, so geht dies mit `apply`, wobei die `richtung` 1 bzw. 2 ist. Die `funktion` muss als ersten Parameter den Vektor erhalten; allfällige weitere Funktionsparameter für die aufzurufende `funktion` werden `apply` als zusätzliche Argumente übergeben. Das Resultat ist in der Regel ein Vektor mit den einzelnen Resultaten.

`apply` kann man auch auf Arrays höherer Dimension anwenden. Weitere ähnliche Funktionen gibt es zur Anwendung auf jedes Element einer Liste (`lapply`, `sapply`). Die Verwendung von `apply`, `lapply` und `sapply` ist teilweise wesentlich effizienter als eine entsprechende `for`-Schleife.

Um spaltenweise Mittelwerte bzw. Summen zu berechnen, gibt es auch die Funktionen `colMeans` und `colSums`. Für zeilenweise Mittelwerte bzw. Summen stehen die Funktionen `rowMeans` bzw. `rowSums` zur Verfügung.

- *Funktionen zum Erzeugen von Zufallszahlen*

Pseudo-Zufallszahlen erhält man je nach gewünschter Verteilung z. B. mit `rnorm(n, mu, sigma)` (Normalverteilung), `runif(n, min, max)` (Uniformverteilung), `rexp(n, rate)` (Exponentialverteilung), `rt(n, df)` (Student-Verteilung), `rgamma(n, shape, rate)` (Gammaverteilung), `rchisq(n, df)` (χ^2 -Verteilung), `rbinom(n, size, prob)` (Binomialverteilung) oder `rpois(n, lambda)` (Poissonverteilung). Dabei ist `n` die gewünschte Anzahl Beobachtungen. Die übrigen Parameter können häufig weggelassen werden, was die jeweilige Standard-Variante der Verteilung ergibt. Man kann auch Vektoren angeben, um Beobachtungen mit unterschiedlichen Parametern zu erzeugen.

Um `n` Stichprobenwerte einer kategorialen Variable mit Ausprägungen `A.1, ..., A.m` und Gewichten `p.1, ..., p.m` zu erzeugen, wird

```
sample(c('A.1', ..., 'A.m'), n, replace = TRUE, prob = c(p.1, ..., p.m))
```

aufgerufen. Sind die Gewichte gleich, so kann das `prob`-Argument weggelassen werden.

Um reproduzierbare Pseudo-Zufallszahlen zu erhalten, kann man den Generator mit `set.seed(n)` initialisieren, bevor man ihn verwendet; `n` ist dabei eine ganze Zahl.

In diesem Zusammenhang sei auch auf die Funktionen hingewiesen, welche die Wahrscheinlichkeitsfunktionen, Dichtefunktionen, Verteilungsfunktionen und Quantilfunktionen bereitstellen. Diese heißen

gleich wie die Funktionen für Zufallszahlen, aber mit einem d, p oder q (statt r) am Anfang – also beispielsweise `dnorm(x, mu, sigma)` für die Normalverteilungsdichte an der Stelle x. Dabei kann x eine Zahl oder ein Vektor sein.

Beispiel A.5. Wir möchten gerne einen Datensatz erzeugen, der für vier Haushalte das Jahreseinkommen von den Jahren 2006, 2007 und 2008 enthält. Die Einkommen simulieren wir mit Hilfe der Gammaverteilung und der Normalverteilung.

R Code

```
# Eingabe
set.seed(1)      # Zur Reproduzierbarkeit

sim1 <- data.frame(zeilen.nr = 1:12, haushalt.nr = rep(1:4, each = 3),
                    jahr = rep(2006:2008, times = 4),
                    einkommen = round(rep(rgamma(4, 8, 0.001), each=3) + rnorm(12, sd=200), -2))

# Ausgabe
zeilen.nr haushalt.nr jahr einkommen
1          1           1 2006    5600
2          2           1 2007    5700
3          3           1 2008    5800
4          4           2 2006   11600
5          5           2 2007  12100
6          6           2 2008  11700
7          7           3 2006  11200
8          8           3 2007  11200
9          9           3 2008  11300
10         10          4 2006  8600
11         11          4 2007  8600
12         12          4 2008  8700
```



Manchmal möchte man alle Kombinationen von Einträgen mehrerer Vektoren als Datensatz zusammenfassen (z. B. für systematische Vorhersagen anhand eines Modells). Dies geschieht folgendermassen:

```
new <- expand.grid(Name1 = Vektor1, Name2 = Vektor2, etc.)
```

Beispiel A.6. Wir möchten alle Kombinationen von Geschlecht und einigen typischen Ausprägungen von Alter als Data Frame zusammenfassen.

R Code

```
# Eingabe
new <- expand.grid(Geschlecht = c("f", "m"), Alter = (2:6)*10)
new

# Ausgabe
Geschlecht Alter
1          f    20
2          m    20
3          f    30
4          m    30
5          f    40
6          m    40
7          f    50
8          m    50
9          f    60
10         m    60
```



A.3.5 Sortieren von Daten

- `sort(x)`

Diese Funktion liefert die aufsteigend sortierte Version von `x`. Mit der Option `decreasing = TRUE` wird absteigend sortiert.

- `rank(x)`

Diese Funktion liefert die Ränge der Einträge innerhalb `x`. Der kleinste Wert erhält Rang 1. Allfällige Bindungen (also gleich grosse Werte) erhalten denselben Rang.

- `order(x.1, ..., x.n)`

`order` liefert die Indizes der Elemente von `x.1`, die zum Ordnen der Elemente verwendet werden können, d.h. das erste Element der Resultats gibt die Position des kleinsten Elements von `x.1` an usw. (`x.1[order(x.1)]` ergibt also den gleichen Vektor wie `sort(x.1)`.) Allfällige Bindungen kann man durch die Angabe weiterer Vektoren (mit der gleichen Anzahl von Elementen wie `x.1`) ordnen lassen. Wiederum kann man mit `decreasing = TRUE` auch die umgekehrte Reihenfolge verlangen.

Die Funktion `order` ermöglicht insbesondere das Sortieren der Zeilen eines Data Frame nach einzelnen Variablen:

```
df [order(df$x.1, ..., df$x.n), ]
```

A.3.6 Namen von Elementen und Faktor-Stufen

Oft ist es nützlich, beispielsweise die Elemente eines Vektors nicht nur aufgrund ihrer Position identifizieren zu können, sondern jedes Element mit einem Namen zu versehen. Auf diese Namen kann man mit der Funktion `names` zugreifen. Dies funktioniert für den Abruf bereits vorhandener Namen, aber auch auf der linken Seite einer Zuweisung.

Ähnlich kann man auch für Matrizen und Data Frames Zeilen- und Spaltennamen festlegen bzw. abrufen. Dies geschieht mit den Funktionen `colnames` und `rownames` (für Data Frames liefert auch `names` die Spaltennamen).

`levels` liefert die Bezeichnungen der Stufen eines Faktors.

A.4 Grafiken

Die Standardfunktion zur Erstellung einer Grafik heisst `plot(x, y)`. Sie erzeugt ein Streudiagramm der Vektoren `x` und `y`.

Einige der unzähligen Optionen von `plot` sind in Tabelle A.3 beschrieben.

Alle statistischen Grafikfunktionen (`barplot`, `hist` etc.) beruhen schliesslich auf `plot`. Deshalb lassen sich dort jeweils die meisten dieser Optionen verwenden.

Mit der Funktion `par` lassen sich zahlreiche Grafik-Parameter einstellen, die für die danach erstellten Grafiken gelten. Eine der besonders nützlichen Optionen dient der Anordnung mehrerer Grafiken neben- bzw. untereinander: Mit `mfrow=c(m,n)` werden die darauf folgenden Plots in `m` Zeilen zu je `n` Grafiken angeordnet. Mit `par` kann man auch andere Standardwerte für einige der bereits erwähnten Grafik-Parameter festlegen – etwa `par(lwd=2)` für etwas dickere Linien. Unter der Hilfe zu `par` findet man ausserdem die genauen Angaben zu weiteren solchen Parametern, die in verschiedenen Grafik-Funktionen verwendet werden können.

<code>xlim = c(lower, upper)</code>	Ausschnitt der x -Achse
<code>ylim = c(lower, upper)</code>	Ausschnitt der y -Achse
<code>xlab = Text; ylab = Text</code>	x - bzw. y -Achsenbeschriftung
<code>main = Text; sub = Text</code>	Titel; Untertitel
<code>xaxt = 'n'; yaxt = 'n'</code>	Keine x - bzw. y -Achse
<code>axes = FALSE</code>	Keine Achsen
<code>bty = Umrahmung</code>	Art der Umrahmung. ' <code>n</code> ' für keine
<code>pch = Symbol</code>	Aussehen der Punkte (Standard ' <code>o</code> ')
<code>type = Typ</code>	' <code>p</code> ' für Punkte, ' <code>l</code> ' für Linien, ' <code>b</code> ' für beides, ' <code>n</code> ' für nichts etc.
<code>lty = Linientyp</code>	1 (durchgezogen), 2 (gepunktet), 3 (gestrichelt), ...
<code>lwd = Breite</code>	Breite der Linie als Zahl (Standard 1)
<code>cex = Textgrösse</code>	Textgrösse als Zahl (Standard 1)
<code>col = Farbe</code>	Farbe als Zahl oder Text, z. B. 1 bzw. ' <code>black</code> ', 2 bzw. ' <code>red</code> ', 12 bzw. ' <code>blue</code> ' etc.

Tabelle A.3: Einige wichtige Optionen von `plot`.

Hinweis. Das aktive Grafikfenster wird entweder via Menü oder mit

```
savePlot(file = Pfad_und_Datenname) gespeichert.
```

Beispiel A.7. Wir zeichnen die Exponentialfunktion e^x im Bereich von 0 bis 5. Abbildung A.2 zeigt das Ergebnis.

R Code

```
# Eingabe
x <- seq(0, 5, by = 0.001)
y <- exp(x)

plot(x, y, type = 'l', lwd = 3, main = 'Exponentialfunktion')
```



Grafiken ergänzen

Die aktive Grafik kann relativ einfach durch zusätzliche Elemente erweitert werden. Eine Auswahl von Funktionen, die jeweils meist viele Optionen aufweisen (siehe R-Hilfe):

- `points(x, y)` bzw. `lines(x, y)`

Mit `points` oder `lines` wird die Grafik um weitere Punkte bzw. Linien ergänzt. Die Syntax entspricht im Wesentlichen derjenigen der Funktion `plot`, jedoch fallen die Optionen zur Beschriftung der Grafik und zur Achsenwahl weg.

- `legend(x, y, legend=beschriftungen, lty=linienart, etc.)`

Eine Legende kann man wie zusätzliche Punkte oder Linien nachträglich zu einer Grafik hinzufügen: `x` und `y` geben die Koordinaten der oberen linken Ecke der Legende an. Alternativ kann man auch z. B. für `x` '`topright`' oder '`topleft`' einsetzen und `y` weglassen, um die Legende in der oberen rechten oder linken Ecke der Grafik zu platzieren.

Die Legende besteht aus den `beschriftungen` sowie einer Identifikation der zu beschriftenden Elemente (z. B. via `lwd` und/oder `lty`).

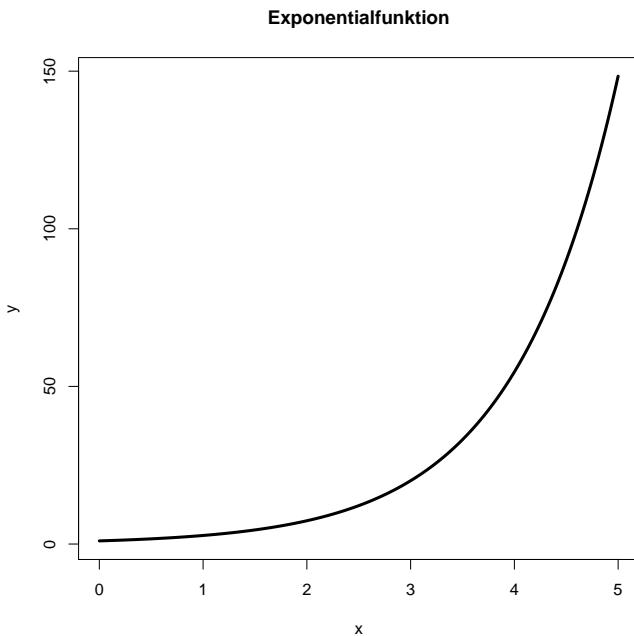


Abbildung A.2: Exponentialfunktion von Beispiel A.7.

- **axis(seite)**

Mit dem Befehl `axis` zeichnet man eine Achse auf einer `seite` der Grafik (1=unten, 2=links, 3=oben, 4=rechts), inkl. Teilstriche und Beschriftungen. Dies kann eine zusätzliche Achse oben oder rechts sein; will man die Achsen unten bzw. links mit selbst gewählten Einteilungen zeichnen, so muss man beim Erstellen der Grafik (z. B. im `plot`-Aufruf) die Option `axes=FALSE` (oder `xaxt = 'n'` bzw. `yaxt = 'n'`) ergänzen, damit die Achsen nicht bereits automatisch gezeichnet werden. Sollen die Positionen der Achsenintervalle nicht automatisch gewählt werden, so kann man diese als Vektor via Option `at` angeben. In diesem Fall kann man mit der Option `labels` auch die Beschriftungen selber wählen (standardmäßig wird die Position als Zahl angeschrieben). Unterdrückt man die automatisch erzeugten Achsen in `plot`, möchte aber trotzdem das übliche Rechteck um den Zeichnungsbereich herum erhalten, so kann man dieses nachträglich mit `box()` hinzufügen.

- **abline(position)**

Mit dieser Funktion kann man gerade Hilfslinien in eine Grafik einzeichnen. `position` kann man mit y-Achsen-Abschnitt und Steigung spezifizieren (z. B. `abline(0, 1)` für die Winkelhalbierende), oder man kann mit `h=y.wert` bzw. `v=x.wert` eine horizontale bzw. vertikale Gerade an der entsprechenden Stelle verlangen.

- **grid()**

Zeichnet anhand der Achsenstriche ein Gitter mit Hilfslinien ein.

- **text(x, y, text)**

Schreibt bei der durch `x` und `y` festgelegten Position `text` hin.

A.5 Statistische Verfahren

Hier zeigen wir eine (inkomplette) Übersicht über die statistischen Funktionen von R.

A.5.1 Univariate Verfahren

Eine kategoriale Variable

- `table`: Absolute Häufigkeiten. Werden mit `prop.table` in relative umgerechnet
- `barplot`, `dotchart`, `pie`: Grafische Darstellungen der Häufigkeiten
- `binom.test`: Binomialkonfidenzintervalle und -tests
- `chisq.test`: Chiquadrat-Anpassungstest

Eine numerische Variable

- `summary`: Mittelwert, Quartile, Minimum, Maximum und Anzahl fehlende Werte
- `quantile`, `mean`, `median`, `min`, `max`, `range`, `IQR`, `sd`, `var`: Einzelne Kenngrößen
- `hist`, `rug`, `Ecdf` (Paket ‘Hmisc’): Grafische Darstellungen
- `t.test`: Studentkonfidenzintervalle und -tests (Einstichprobenfall)
- `wilcox.test`: Wilcoxons Signed-Rank-Test

Beschreibung eines Datensatzes

- `summary`: Beschreibt numerische Merkmale und Faktoren
- `mean`, `sd`: Einzelne Kenngrößen
- `hist`, `Ecdf` (beide in ‘Hmisc’): Grafische Darstellungen

A.5.2 Bivariate Verfahren

Beide Merkmale kategoriall

- `table`: Häufigkeitstabelle. Wird mit `addmargins` um Zeilen- und Spaltensummen erweitert
- `prop.table`: Zeilen- oder Spaltennormierung (Option `margin`)
- `plot`: Mosaikdiagramm
- `chisq.test`: Chiquadrat-Unabhängigkeitstest
- `fisher.test`: Fishers exakter Test; Odds Ratio inkl. Konfidenzintervall
- `mcnemar.test`: McNemars χ^2 -Test auf Symmetrie

Ein Merkmal numerisch, eins kategoriall

- `boxplot`, `stripchart`, `Ccdf` (Paket ‘Hmisc’): Grafische Darstellungen
- `by`, `summary.formula` (Paket ‘Hmisc’): Stratifizierte deskriptive Analysen
- `t.test`: Studentkonfidenzintervalle und -tests im Zweistichprobenfall
- `wilcox.test`: Wilcoxons Rangsummentest
- `lm`: Einweg-ANOVA. Schliessende Statistik mit `summary`
- `kruskal.test`: Rangsummentest nach Kruskal und Wallis
- `friedman.test`, `quade.test`: Mehrstichprobenvergleiche bei verbundenen Stichproben

Beide Merkmale numerisch

- `plot`: Streudiagramm, Streudiagramm-Matrix
- `lm`: Lineare Regression. Schliessende Statistik mit `summary`, `confint`; Regressionsgerade mit `abline`
- `cor`, `cor.test`: Verschiedene Arten von Korrelationen inkl. schliessender Statistik

A.5.3 Multivariate Verfahren

Allgemeines lineares Modell

Das allgemeine lineare Modell wird mit `lm(Formel, data = Data)` aufgerufen, wobei die Formel eine Eingabe der Art `Zielgrösse ~ Kovariablen1 + Kovariablen2 + ...` ist.

Die rechte Seite der Formel kann u. a. zusätzlich folgende Elemente enthalten:

- `+0` bzw. `-1`: Erzwingt einen Intercept von 0 (Regression durch den Ursprung)
- `Kovariablen1 * Kovariablen2`: Kovariablen1 und Kovariablen2 sowie deren Interaktion
- `(Kovariablen1 + Kovariablen2 + ...) ^ 2`: Kovariablen inkl. paarweise Interaktionen
- `log(Kovariablen1)`, `I(Kovariablen1 + Kovariablen2)` etc.: Transformationen der Variablen

Auf das Ergebnis `fit` von `lm` können u. a. folgende Funktionen angewendet werden:

- `summary`: Schätzwerte der Parameter, *t*-Tests, R^2 , globaler *F*-Test
- `drop1`: Typ II-ANOVA
- `anova(fit, subfit)`: Partieller *F*-Test, um `subfit` mit `fit` hinsichtlich echtem R^2 zu vergleichen
- `confint`: Konfidenzintervalle für die geschätzten Parameter
- `predict`: Vorhersagen. Konfidenz- bzw. Prognoseintervalle mit `interval='c'` bzw. `interval='p'`
- `plot`, `hatvalues`, `cooks.distance`, `dfbetas`: Diagnostik
- `coef` (Schätzwerte), `resid` (Residuen), `fitted` (gefittete Werte)

Dimensionsreduktion

- `princomp`: Hauptkomponentenanalyse
- `factanal`: Faktorenanalyse
- `kmeans`: K-means Clusteranalyse

A.5.4 Zusatz: Weitere Konfidenzintervalle

Hier zeigen wir zwei Funktionen, die im Skript zur Berechnung von Konfidenzintervallen für das R^2 und das Cramérs V in der Population verwendet werden.

R Code

```
# Konfidenzintervalle für das echte R-Quadrat. Wird auf lm-Objekt angewendet.
library(MBESS)
confint.R2 <- function(fit, alternative=c('two.sided','less','greater'), conf.level=0.95)
{
  alternative <- match.arg(alternative)
  alpha.lower <- alpha.upper <- 0
  if (alternative == "two.sided")
    alpha.lower <- alpha.upper <- (1 - conf.level)/2
  if (alternative == "greater")
    alpha.lower <- 1 - conf.level
  if (alternative == "less")
    alpha.upper <- 1 - conf.level

  fstat <- summary(fit)$fstatistic
  df.1 <- fstat[2]
  df.2 <- fstat[3]
  Delta <- unlist(conf.limits.ncf(F.value=fstat[1], df.1, df.2, alpha.lower,
    alpha.upper, conf.level = NULL)[c("Lower.Limit", "Upper.Limit")])
  out <- c(Delta/(Delta + df.1 + df.2 + 1))
  out[is.na(out)] <- 0
  out
}

# Cramers V inkl. Konfidenzintervalle. Wird auf chisq.test-Objekt angewendet.
library(MBESS)
cramers.V <- function(z, alternative=c('two.sided','less','greater'), conf.level=0.95)
{
  alternative <- match.arg(alternative)
  alpha.lower <- alpha.upper <- 0
  if (alternative == "two.sided")
    alpha.lower <- alpha.upper <- (1 - conf.level)/2
  if (alternative == "greater")
    alpha.lower <- 1 - conf.level
  if (alternative == "less")
    alpha.upper <- 1 - conf.level

  df <- z$parameter
  chi <- as.numeric(z$statistic)
  n <- sum(z$observed)
  k <- min(dim(z$observed)) - 1
  Delta <- unlist(conf.limits.nc.chisq(chi, conf.level = NULL, df,
    alpha.lower, alpha.upper)[c("Lower.Limit", "Upper.Limit")])
  out <- sqrt(c(Cramers.V = chi, Delta)/(n*k))
  out[is.na(out)] <- 0
  out
}
```

Anhang B

Formelsammlung

Stichproben-Standardabweichung

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

Zu Verteilungen

Verteilung	$\text{Bin}(n, p)$	$\text{Poiss}(\lambda)$	$\mathcal{N}(\mu, \sigma^2)$	$\text{Exp}(\lambda)$	$\text{Gamma}(a, b)$
Parameter	$n \in \mathbb{N}, p \in [0, 1]$	$\lambda \geq 0$	$\mu \in \mathbb{R}, \sigma > 0$	$\lambda > 0$	$a, b > 0$
Wertebereich	$0, 1, \dots, n$	$0, 1, \dots$	\mathbb{R}	$[0, \infty)$	$[0, \infty)$
W'keits-/Dichtefkt.	$\binom{n}{k} p^k (1-p)^{n-k}$	$e^{-\lambda} \cdot \frac{\lambda^k}{k!}$	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma}\right)^2}$	$\lambda e^{-\lambda x}$	$\frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$
Erwartungswert	np	λ	μ	$1/\lambda$	a/b
Varianz	$np(1-p)$	λ	σ^2	$1/\lambda^2$	a/b^2

- Binomialkoeffizient $\binom{n}{k} = \frac{n!}{(n-k)!k!}$
- Für $a = 1, 2, \dots$ ist $\Gamma(a) = (a-1)!$
- Verteilungsfunktion von $\text{Exp}(\lambda)$: $F(r) = 1 - e^{-\lambda r}, \quad r \geq 0$
- Quantilfunktion von $\text{Exp}(\lambda)$: $F^{-1}(\beta) = -\ln(1-\beta)/\lambda, \quad \beta \in [0, 1]$
- Die Chiquadrat-Verteilung mit k Freiheitsgraden entspricht $\text{Gamma}(k/2, 1/2)$
- Die Uniformverteilung zwischen 0 und 1 hat Erwartungswert $\frac{1}{2}$ und Varianz $\frac{1}{12}$.

Regeln zu Erwartungswert und Varianz

Für Zufallsvariablen X und Y und Konstanten a und b gilt

- $E(X + Y) = E(X) + E(Y);$
- $E(a + b \cdot X) = a + b \cdot E(X);$
- $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y), \text{ falls } X \text{ und } Y \text{ unabhängig sind};$
- $\text{Var}(a + b \cdot X) = b^2 \cdot \text{Var}(X).$

Faustregeln für Normalverteilung

Für $X \sim \mathcal{N}(\mu, \sigma^2)$ gilt

- $P(X \in [\mu \pm \sigma]) \approx 0.68$
- $P(X \in [\mu \pm 2\sigma]) \approx 0.95$
- $P(X \in [\mu \pm 3\sigma]) \approx 0.997$

Teststatistik des Z- und t-Tests

$$Z_o = \frac{\bar{X} - \mu_o}{S/\sqrt{n}}$$

Zum Chiquadrat-Anpassungstest

- Kategorielles Merkmal X mit K verschiedenen Ausprägungen x_1, x_2, \dots, x_K
- H_j : Anzahl aller Beobachtungen mit Ausprägung x_j
- p_j^o : Vorgegebene Wahrscheinlichkeit von Ausprägung x_j
- Pearson-Residuum von Ausprägung x_j : $\frac{H_j - np_j^o}{\sqrt{np_j^o}}$
- Pearsons Chiquadrat-Teststatistik: Summe der quadrierten Pearson-Residuen

Zu Häufigkeitstabellen

- Kategorielles Merkmal X mit L verschiedenen Ausprägungen x_1, x_2, \dots, x_L
- Kategorielles Merkmal Y mit M verschiedenen Ausprägungen y_1, y_2, \dots, y_M
- $H_{j,k}$: Anzahl aller Beobachtungen mit $X = x_j$ und $Y = y_k$
- Zeilensummen $H_{j,+} = \sum_{k=1}^M H_{j,k}$ (Anzahl aller Beobachtungen mit $X = x_j$)
- Spaltensummen $H_{+,k} = \sum_{j=1}^L H_{j,k}$ (Anzahl aller Beobachtungen mit $Y = y_k$)
- Idealisierte Häufigkeiten: $\bar{H}_{j,k} = \frac{H_{j,+} H_{+,k}}{n}$
- Pearson-Residuum: $e_{j,k} = \frac{H_{j,k} - \bar{H}_{j,k}}{\sqrt{\bar{H}_{j,k}}}$
- χ^2 -Teststatistik: Summe der quadrierten Pearson-Residuen
- Cramérs V: $\sqrt{\frac{\chi^2\text{-Teststatistik}}{n(\min\{L,M\}-1)}}$

Korrelation und einfache lineare Regression

- Stichproben-Kovarianz: $\text{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$
- Steigung in der Stichprobe: $\hat{\beta} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$
- y -Achsenabschnitt in der Stichprobe: $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$
- Korrelationskoeffizient nach Pearson: $r = \frac{\text{Cov}(X, Y)}{\text{Std}(X)\text{Std}(Y)}$
- Bestimmtheitsmaß: $R^2 = r^2$
- Für Zufallsvariablen X und Y gilt:

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

wobei sich Cov hier auf die theoretische Version

$$\text{Cov}(X, Y) = E((X - E(X)) \cdot (Y - E(Y)))$$

bezieht.

Index

- Äquivalenztest, 68
- Absolute Häufigkeit (count), 13
- Alternativhypothese, 62
- ANCOVA, *siehe* Kovarianzanalyse
- ANOVA, *siehe* Varianzanalyse (analysis of variance)
- Anpassungstest (goodness-of-fit test, gof test), 87
- Approximative Normalität, 77
- Arbeitshypothese (working hypothesis), 62
- Ausprägung, 9
- Ausreisser (outlier), 23, 116
- AV, 166
- Backward-Elimination, 201
- Balkendiagramm (bar chart), 14
- Bedingte Verteilungen (conditional distributions), 93
- Bedingte Wahrscheinlichkeit, 98
- Beobachtung (observation, case), 9
- Bernoulliverteilung, 39
- Bestimmtheitsmaß
- Lineare Regression, 154
 - Mittelwertvergleich, 126
- Binomialkoeffizient, 55
- Binomialkonfidenzintervall, 58
- Binomialtest, 62
- Binomialverteilung, 55
- Bonferroni-Korrektur, 68
- Box-(Whisker-)Plots, 116–119
- Breite Form, 137–138
- Cauchyverteilung, 48, 81
- Chancen (odds), 108
- Chancenquotient (odds ratio), 108
- Chiqrat-Anpassungstest, 87
- Chiqrat-Teststatistik, 88, 99
- Chiqrat-Unabhängigkeitstest (chi-square test of independence), 102
- Chiqrat-Verteilung, 85
- Cluster, 213
- Clusteranalyse (cluster analysis), 213–216
- Clustermittelwert (cluster mean), 213
- Cochrands Q-Test, 143
- Confounder, 93
- Cramérs V, 99
- Datensatz (data set), 9
- Dichtefunktion ((probability) density function, pdf), 40
- Dimensionsreduktion, 203–216
- Diskriminanzanalyse, 199
- Drop-out, 140
- Dummycodierung, 175
- Dummykodierung, 174
- Dummyvariable, 174
- ECDF, 16, 113–115
- Effekttabelle, 183
- Eigenwertzerlegung, 207
- Einflussreiche Beobachtung, 148, 195
- Elastizität, 179
- Empirische Verteilungsfunktion (empirical cumulative distribution function), *siehe* ECDF
- Erwartungstreue, *siehe* Unverfälscht (unbiased)
- Erwartungswert (mean, expectation), 47
- Exponentialverteilung, 41
- F-Test
- Globaler, 184
 - Lineare Regression, 158
 - Mittelwertvergleich, 128
 - Partieller, 185
- F-Teststatistik, 129
- F-Verteilung, 129
- Faktorenanalyse, 207
- Fehler erster Art, 67
- Fehler zweiter Art, 67
- Fehlerterm, 192
- Fishers exakter Test, 109
- Five number summary, 21
- Forward-Selection, 201
- Freiheitsgrad (degrees of freedom, df), 81, 85, 102
- Friedman-Test, 143
- GAM, 199
- Gammafunktion, 84
- Gammaverteilung, 84
- Gauss'sche Verteilung (gaussian distribution), 73
- Gefittete Werte (fitted values)
- Lineare Regression, 154
 - Lineares Modell, 171
- Gefittert Wert (fitted value)
- Mittelwertvergleich, 126
- Gemeinsame Verteilung (joint distribution), 93
- Gemeinsame Wahrscheinlichkeit, 98
- Gesetz der grossen Zahlen (law of large numbers), 52
- Getrimmter Mittelwert, 23
- GLM, 199
- Goodman-Kruskal Gamma, 156
- Grundgesamtheit (Population), 34
- Häufigkeitspolygon (frequency polygon), 113
- Häufigkeitstabelle (frequency table), 94
- Erweiterte, 95
- Höcker (peak), 19
- Hauptkomponente (principal component), *siehe* PC
- Hauptkomponentenanalyse (principal component analysis), *siehe* PCA

- Hauptsatz der Statistik, 45
 Heteroskedastizität, 194
 Histogramm (histogram), 18
 Homoskedastizität, 194
 Hotellings T-Test, 199
- Identisch verteilt, 38
 Inferenzstatistik, 34
 Input, 166
 Interaktionen, 175, 181–182
 Intercept
 Lineare Regression, 146
 Lineares Modell, 167
 Mittelwertvergleich, 134
 Interquartilabstand, 24
 IQR, *siehe* Interquartilabstand
- k-means-Methode, 213
 Kendalls Tau, 156
 Kerndichteschätzer (kernel density estimator), 113
 Kleinstes-Quadrat (least-squares), 147, 168
 Kollinearität, *siehe* Multikollinearität
 Kolmogorov-Smirnov-Unabhängigkeitstest, 122
 Konfidenzintervall (confidence interval), 57
 Konfidenzniveau (coverage probability), 57
 Konsistenz, 52
 Kontingenzkoeffizient C, 99
 Kontingenztafel (contingency table), 94
 Korrelationskoeffizient nach Pearson, 155–156
 Korrelationsmatrix, 156
 Kovariablen (covariate), 166
 Kovarianz (covariance), 147, 153
 Kovarianzanalyse (analysis of covariance), 168
 Kruskal-Wallis-Rangsummentest, 131
 Kuchendiagramm (pie chart), 14
- Ladung (loading)
 PCA, 206
 Lagemass (measure of location, center), 23
 Lange Form, 137
 Leverage, 148, 195
 Leverage-Effekt, 148–149, 195
 Lineare Regression (linear regression), 146–152
 Linearer Prädiktor (linear predictor)
 Lineare Regression, 149
 Lineares Modell, 168
- MANCOVA, 199
 Mann-Whitney U-Test, *siehe* Wilcoxons Rangsummentest
 MANOVA, 199
 Maximum, 21
 McNemar-Test, 143
 Median, 23
 Mediantest, 69
 Merkmal (variable), 9
 Minimum, 21
 Mittelwert (mean), 23
 Modell
 Additives, 167
 Cox-, 199
 Gamma-, 199
 Generalized additive, 199
- Klassifikations-, 199
 Lineares, 166
 Logit-, 199
 Mixed-Effects-, 199
 Multinomiales, 199
 Multivariates, 199
 Ordinales, 199
 Probit-, 199
 Räumliches, 199
 Resistentes, 199
 Strukturgleichungs-, 199
 Survival, 199
 Verallgemeinertes lineares, 199
 Zeitreihen-, 195, 199
- Modelldiagnostik, 192–198
 Modellgüte, 171
 Modellparameter, 167
 Modellvoraussetzungen, 192–198
 Modus (mode), 13
 Monte-Carlo-Simulation, *siehe* Simulation
 Mosaikdiagramm (mosaic plot), 97
 Multikollinearität, 191, 203
 Multiples Testen, 68
- Nichtlinearitäten, 175, 180
 Normalverteilung (normal distribution), 73
 Nullhypothese (null hypothesis), 62
- Odds Ratio, 108
 OLS, 166
 Output, 166
 Overfitting, 200
- p-Wert (p value), 62
 Page-Test, 143
 Paretoverteilung, 54
 PC, 206
 PCA, 206–212
 Pearson-Residuen (pearson residuals), 87, 99
 Poissonverteilung, 70
 Power (Güte), 67
 Prädiktionsintervall, 184
 Prädiktor (predictor), 166
 Prognoseintervall, *siehe* Prädiktionsintervall
 Punktediagramm (dot chart), 14
- Quade-Test, 143
 Quantil, 20
 Quantifunktion, 46
 Quantilregression, 199
 Quartil, 21
- R-Quadrat
 Adjustiertes (adjusted), 171
 Lineare Regression, 154
 Lineares Modell, 171
 Mittelwertvergleich, 126
- Rangkorrelationskoeffizient nach Spearman, 156–157
 Realisierung, 35
 Referenzkategorie, 133
 Regression
 Kubische, 168, 180

- Lineare, 146–152
- Multiple lineare, 167
- Polynomiale, 168, 180
- Quadratische, 168, 180
- Regressionsgerade (regression line), 147
- Regressionskoeffizienten, 147
- Regressionsmodell, 166
- Relative Häufigkeit (proportion), 13
- Residuals versus Fitted, 193
- Residuen (residuals)
 - Lineare Regression, 154
 - Lineares Modell, 171
 - Mittelwertvergleich, 126
- Risiko (risk), 108
- Robust (resistant), 23
- Satz von Bayes, 98
- Schätzer (estimator), 38
- Schätzwert (estimate), 38
- Schiefe (skewness), 19
- Score
 - PCA, 207
 - Summen-, 204
- Signifikanzniveau, 63
- Simulation, 50
- Smoothing Splines, 199
- Spaltennormierung, 96
- Spannweite (range), 24
- Spearmans Rangkorrelation, 156
- Störvariable, *siehe* Confounder
- Stabdiagramm, 14
- Standardabweichung (standard deviation), 24, 51
- Standardfehler, 78
- Standardisierter Mittelwert, 77
- Standardnormalverteilung, 73
- Steigung (slope), 146
- Stichprobe (sample), 9
- Stichprobenumfang (sample size), 9
- Stichprobenwert, 9
- Stratifizierte Analyse, 113
- Streudiagramm (scatter plot), 145–146
- Streudiagramm-Matrix, 146
- Streuungsmass (measure of spread/variation), 24
- Stripchart, 17, 115–116
- Student-Konfidenzintervall
 - Lineare Regression, 150
 - Lineares Modell, 182
 - Mittelwertvergleich, 122, 134
 - Univariates, 82
- Student-Verteilung, 81
- Studentisierter Mittelwert, 78
- Summen, 204–206
 - Gewichtete (weighted), 205
- t-Test
 - für verbundene Stichproben, 139
 - Lineare Regression, 150
 - Lineares Modell, 182
 - Mittelwertvergleich, 122, 134
 - Univariater, 83
- t-Verteilung, *siehe* Student-Verteilung
- Test, 62
- Teststatistik, 62
- Transformationen, 175
 - Lineares Modell, 175–180
- Typ-II-Anova, 185
- Uniformverteilung, 42
- Unverbundene Stichproben (independent samples), 137
- Unverfälscht (unbiased), 49
- UV, 166
- Value at risk, 50
- Variable, 9
 - Abhängige (dependent), 166
 - Erklärende, 166
 - Erklärte, 166
 - Unabhängige (independent), 166
- Variabletypen (types of variables), 10
 - binär (binary), 10
 - dichotom (dichotomous), 10
 - kategorial (categorical), 10
 - numerisch (numeric), 10
 - ordinal (ordinal), 10
 - qualitativ, 10
 - quantitativ, 10
- Varianz, 24, 51
- Varianz-Kovarianzmatrix, 147
- Varianzanalyse (analysis of variance), 126
- VC-Matrix, *siehe* Varianz-Kovarianzmatrix
- Verbundene Stichproben (paired samples), 137
- Verteilung (distribution), 37
- Verteilungsfunktion ((cumulative) distribution function, cdf), 43
- Vierfeldertafel (2-by-2 table), 108
- Volatilität, 24, 54
- Vorher–Nachher, 136
- Vorhersage (prediction)
 - Lineare Regression, 149
 - Lineares Modell, 168
 - Mittelwertvergleich, 126
- Vorzeichentest (sign test), 69
- Wahrscheinlichkeit, 36
- Wahrscheinlichkeitsfunktion (probability (mass) function, pmf), 38
- Wahrscheinlichkeitsverteilung, *siehe* Verteilung
- Wert (value), 9
- Wertebereich, 35
- Wilcoxons Rangsummentest, 124–125
- Wilcoxons Signed-Rank-Test, 143
- y-Achsenabschnitt (intercept), 146
- Z-Konfidenzintervall, 77
- Z-Test, 79
- Z-Teststatistik, 79
- Zeilennormierung, 95
- Zensierte Beobachtungen, 199
- Zentraler Grenzwertsatz (Central Limit Theorem), 77
- Zentralwert, *siehe* Median
- Zielgrösse (response, outcome), 166
- Zufälliger Vorgang, 35
- Zufallsvariable (random variable), 35
- Zusammenhangsmass (measure of dependence), 99