

Praxis der Datenanalyse

Skript zum Modul

*Sebastian Sauer. Mit Beiträgen von Oliver Gansser, Matthias Gehrke,
Karsten Lübke und Norman Markgraf*

06 July, 2017

Inhaltsverzeichnis

Vorwort	xvii
Organisatorisches	xix
0.1 Modulziele	xix
0.2 Themen pro Termin	xix
0.3 Vorerfahrung	xx
0.4 Prüfung	xx
0.4.1 Prüfungshinweise	xx
0.4.2 Klausur	xxi
0.4.3 Datenanalyse	xxi
0.5 Literatur	xxiii
I Grundlagen	1
1 Rahmen	3
1.1 Software installieren	4
1.1.1 R und RStudio installieren	5
1.1.2 Hilfe! R startet nicht!	6
1.1.3 Pakete	7
1.1.4 Vertiefung: Zuordnung von Paketen zu Befehlen	9
1.1.5 Datensätze	11
1.2 ERRRstkontakte	12
1.2.1 R-Skript-Dateien	12
1.2.2 Datentypen in R	12
1.2.3 Hinweise	13
1.2.4 Text und Variablen zuweisen	14
1.2.5 Funktionen aufrufen	14
1.2.6 Das Arbeitsverzeichnis	15
1.3 Hier werden Sie geholfen	16
1.3.1 Wo finde ich Hilfe?	16
1.3.2 Einfache reproduzierbare Beispiele (ERBies)	16
1.4 Was ist Statistik? Wozu ist sie gut?	17
1.5 Aufgaben	19

1.6	Befehlsübersicht	20
1.7	Verweise	21
2	Daten einlesen	23
2.1	Daten in R importieren	24
2.1.1	Excel-Dateien importieren	24
2.1.2	Daten aus R-Paketen importieren	25
2.1.3	Daten im R-Format laden	25
2.1.4	CSV-Dateien importieren	25
2.2	Normalform einer Tabelle	27
2.3	Tabelle in Normalform bringen	28
2.4	Textkodierung	30
2.5	Befehlsübersicht	30
2.6	Aufgaben	31
2.7	Verweise	31
3	Datenjudo	33
3.1	Typische Probleme der Datenaufbereitung	35
3.2	Daten aufbereiten mit <code>dplyr</code>	35
3.2.1	Die zwei Prinzipien von <code>dplyr</code>	35
3.3	Zentrale Bausteine von <code>dplyr</code>	36
3.3.1	Zeilen filtern mit <code>filter</code>	36
3.3.2	Spalten wählen mit <code>select</code>	38
3.3.3	Zeilen sortieren mit <code>arrange</code>	40
3.3.4	Datensatz gruppieren mit <code>group_by</code>	42
3.3.5	Eine Spalte zusammenfassen mit <code>summarise</code>	45
3.3.6	Zeilen zählen mit <code>n</code> und <code>count</code>	47
3.4	Die Pfeife	52
3.4.1	Spalten berechnen mit <code>mutate</code>	54
3.4.2	Aufgaben	55
3.5	Deskriptive Statistik	57
3.6	Befehlsübersicht	58
3.7	Verweise	59
4	Praxisprobleme der Datenaufbereitung	61
4.1	Datenaufbereitung	61
4.1.1	Auf fehlende Werte prüfen	61
4.1.2	Fälle mit fehlenden Werte löschen	62
4.1.3	Fehlende Werte zählen	64
4.1.4	Fehlende Werte ggf. ersetzen	64
4.1.5	Nach Fehlern suchen	64
4.1.6	Ausreißer identifizieren	65
4.1.7	Hochkorrelierte Variablen finden	66
4.1.8	z-Standardisieren	67
4.1.9	Quasi-Konstante finden	68

4.1.10 Auf Normalverteilung prüfen	69
4.1.11 Werte umkodieren und partionieren (“binnen”)	69
4.2 Deskriptive Statistiken berechnen	75
4.2.1 Mittelwerte pro Zeile berechnen	75
4.2.2 Mittelwerte pro Spalte berechnen	75
4.2.3 Korrelationstabellen berechnen	77
4.3 Befehlsübersicht	79
5 Fallstudie ‘movies’	81
5.1 Wie viele Filme gibt es pro Genre?	82
5.2 Welches Genre ist am häufigsten?	82
5.3 Zusammenhang zwischen Budget und Beurteilung	83
5.4 Wurden die Filme im Lauf der Jahre teurer und/oder “besser”?	84
6 Daten visualisieren	85
6.1 Ein Bild sagt mehr als 1000 Worte	86
6.2 Die Anatomie eines Diagramms	87
6.3 Einstieg in ggplot2 - qplot	88
6.4 Häufige Arten von Diagrammen	91
6.4.1 Eine kontinuierliche Variable	91
6.4.2 Zwei kontinuierliche Variablen	95
6.4.3 Eine nominale Variable	99
6.4.4 Zwei nominale Variablen	102
6.4.5 Zusammenfassungen zeigen	103
6.4.6 Überblick zu häufigen Diagrammtypen	107
6.5 Die Gefühlswelt von ggplot2	107
6.6 Aufgaben	108
6.7 Lösungen	109
6.8 Richtig oder Falsch	111
6.9 Befehlsübersicht	111
6.10 Vertiefung: Geome bei ggplot2	112
6.11 Verweise	114
7 Fallstudie zur Visualisierung	115
7.1 Daten einlesen	115
7.2 Daten umstellen	116
7.3 Diagramme für Anteile	117
7.4 Rotierte Balkendiagramme	118
7.5 Text-Labels für die Items	119
7.6 Diagramm mit Häufigkeiten	120
7.7 Farbschemata	121

II Modellieren	123
8 Grundlagen des Modellierens	125
8.1 Was ist ein Modell? Was ist Modellieren?	126
8.2 Ein Beispiel zum Modellieren in der Datenanalyse	128
8.3 Taxonomie der Ziele des Modellierens	131
8.4 Die vier Schritte des statistischen Modellierens	133
8.5 Einfache vs. komplexe Modelle: Unter- vs. Überanpassung	133
8.6 Bias-Varianz-Abwägung	135
8.7 Training- vs. Test-Stichprobe	136
8.8 Wann welches Modell?	137
8.9 Modellgüte	138
8.10 Auswahl von Prädiktoren	138
8.11 Aufgaben	139
8.12 Befehlsübersicht	140
8.13 Verweise	141
9 Der p-Wert, Inferenzstatistik und Alternativen	143
9.1 Der p-Wert sagt nicht das, was viele denken	144
9.1.1 Von Männern und Päpsten	145
9.2 Der p-Wert ist eine Funktion der Stichprobengröße	146
9.3 Mythen zum p-Wert	147
9.4 Wann welcher Inferenztest?	148
9.5 Vertiefung: Beispiele für häufige Inferenztests	148
9.5.1 χ^2 -Test	148
9.5.2 t-Test	149
9.5.3 Varianzanalyse	150
9.5.4 Korrelationen auf Signifikanz prüfen	151
9.5.5 Regression	152
9.5.6 Wilcoxon-Test	152
9.5.7 Kruskal-Wallis-Test	152
9.5.8 Shapiro-Test	153
9.5.9 Logistische Regression	153
9.5.10 Spearmans Korrelation	154
9.6 Zur Philosophie des p-Werts: Frequentismus	154
9.7 Alternativen zum p-Wert	155
9.7.1 Konfidenzintervalle	155
9.7.2 Effektstärke	156
9.7.3 Bayes-Statistik	158
9.8 Aufgaben	161
9.9 Fazit	161
9.10 Verweise	162

III Geleitetes Modellieren	163
10 Lineare Regression	165
10.1 Die Idee der klassischen Regression	166
10.2 Vorhersagegüte	168
10.2.1 Mittlere Quadratfehler	169
10.2.2 R-Quadrat (R^2)	170
10.3 Die Regression an einem Beispiel erläutert	171
10.4 Überprüfung der Annahmen der linearen Regression	173
10.5 Regression mit kategorialen Prädiktoren	175
10.5.1 Aufgaben	177
10.6 Multiple Regression	178
10.7 Interaktionen	179
10.8 Fallstudie zu Overfitting	181
10.9 Aufgaben	182
10.10 Befehlsübersicht	183
11 Klassifizierende Regression	185
11.1 Normale Regression für ein binäres Kriterium	186
11.2 Die logistische Funktion	187
11.3 Die Idee der logistischen Regression	188
11.4 Kein R^2 , dafür AIC	190
11.5 Interpretation der Koeffizienten	190
11.5.1 y-Achsenabschnitt (Intercept) β_0	190
11.5.2 Steigung β_i mit $i = 1, 2, \dots, K$	191
11.5.3 Aufgabe	191
11.6 Kategoriale Prädiktoren	192
11.7 Multiple logistische Regression	192
11.8 Modellgüte	193
11.9 Klassifikationskennzahlen	194
11.9.1 Vier Arten von Ergebnissen einer Klassifikation	194
11.9.2 Klassifikationsgütekennzahlen	195
11.9.3 ROC-Kurven	196
11.10 Aufgaben	198
11.11 Befehlsübersicht	198
12 Fallstudien zum geleiteten Modellieren	201
12.1 Überleben auf der Titanic	201
12.1.1 Daten laden	201
12.1.2 Erster Blick	202
12.1.3 Welche Variablen sind interessant?	202
12.1.4 Univariate Häufigkeiten	202
12.1.5 Bivariate Häufigkeiten	203
12.1.6 Signifikanztest	205
12.1.7 Effektstärke	205

12.1.8 Logististische Regression	206
12.1.9 Effektstärken visualisieren	209
12.1.10 Fazit	211
12.2 Außereheliche Affären	211
12.2.1 Zentrale Statistiken	212
12.2.2 Visualisieren	213
12.2.3 Wer ist zufriedener mit der Partnerschaft: Personen mit Kindern oder ohne?	214
12.2.4 Vertiefung: Wie viele fehlende Werte gibt es?	214
12.2.5 Wer ist glücklicher: Männer oder Frauen?	215
12.2.6 Effektstärken	216
12.2.7 Korrelationen	217
12.2.8 Ehejahre und Affären	218
12.2.9 Ehezufriedenheit als Prädiktor	219
12.2.10 Weitere Prädiktoren der Affärenhäufigkeit	219
12.2.11 Unterschied zwischen den Geschlechtern	220
12.2.12 Kinderlose Ehe vs. Ehen mit Kindern	221
12.2.13 Halodries	221
12.2.14 logistische Regression	222
12.2.15 Zum Abschluss	222
12.3 Befehlsübersicht	223
IV Ungeleitetes Modellieren	225
13 Vertiefung: Clusteranalyse	227
13.1 Grundlagen der Clusteranalyse	228
13.1.1 Intuitive Darstellung der Clusteranalyse	228
13.1.2 Euklidische Distanz	230
13.1.3 k-Means Clusteranalyse	233
13.2 Beispiel für eine einfache Clusteranalyse	234
13.2.1 Distanzmaße berechnen	234
13.2.2 kmeans für den Extraversionsdatensatz	235
13.3 Aufgaben	237
13.4 Befehlsübersicht	238
13.5 Verweise	238
14 Vertiefung: Dimensionsreduktion	241
14.1 Einführung	242
14.2 Warum Datenreduktion wichtig ist	243
14.3 Intuition zur Dimensionsreduktion	244
14.4 Datensatz ‘Werte’	245
14.5 Neuskalierung der Daten	245
14.6 Zusammenhänge in den Daten	246
14.7 Daten mit fehlende Werten	247

14.8 Hauptkomponentenanalyse (PCA)	248
14.8.1 Bestimmung der Anzahl der Hauptkomponenten	248
14.8.2 Scree-Plot	249
14.8.3 Ellbogen-Kriterium	249
14.8.4 Eigenwert-Kriterium	250
14.8.5 Biplot	250
14.8.6 Aufgaben	252
14.8.7 Interpretation der Ergebnisse der PCA	252
14.9 Exploratorische Faktorenanalyse (EFA)	253
14.9.1 Finden einer EFA Lösung	253
14.9.2 Schätzung der EFA	254
14.9.3 Vertiefung: Heatmap mit Ladungen	255
14.9.4 Berechnung der Faktor-Scores	255
14.10 Interne Konsistenz der Skalen	256
14.11 Aufgaben	257
14.12 Befehlsübersicht	258
15 Vertiefung: Grundlagen des Textmining	259
15.1 Zentrale Begriffe	260
15.2 Grundlegende Analyse	261
15.2.1 Tidy Text Dataframes	261
15.2.2 Text-Daten einlesen	263
15.2.3 Worthäufigkeiten auszählen	264
15.2.4 Visualisierung	266
15.3 Aufgaben	268
15.4 Befehlsübersicht	269
15.5 Verweise	269
A Probeklausur	271
B Lösungen	275
C Hinweise	277
D Icons	279
E Voraussetzungen	281
F Zitationen	283
G Lizenz	285
H Autoren	287
I Danke	289
J Zitation dieses Skripts	291

K Kontakt	293
L Technische Details	295
M Sonstiges	297
N Literaturverzeichnis	299

Tabellenverzeichnis

1	Zuordnung von Themen zu Terminen	xx
1.1	Wichtige Datentypen in R	12
1.2	Befehle des Kapitels 'Rahmen'	21
2.1	Befehle des Kapitels 'Daten einlesen'	31
3.1	Befehle des Kapitels 'Datenjudo'	58
4.1	Befehle des Kapitels 'Praxisprobleme'	79
6.1	Häufige Diagrammtypen	107
6.2	Befehle des Kapitels 'Daten visualisieren'	112
8.1	Befehle des Kapitels 'Modellieren'	140
9.1	Überblick über gängige Effektstärkemaße	158
10.1	Befehle des Kapitels 'Regression'	183
11.1	Vier Arten von Ergebnisse von Klassifikationen	194
11.2	Geläufige Kennwerte der Klassifikation	195
11.3	Befehle des Kapitels 'Logistische Regression'	199
12.1	Befehle des Kapitels 'Fallstudien titanic und affairs'	224
13.1	Befehle des Kapitels 'Clusteranalyse'	238
14.1	Befehle des Kapitels 'Dimensionsreduktion'	258
15.1	Die häufigsten Wörter im AfD-Parteidokument	266
15.2	Die häufigsten Wörter im AfD-Parteidokument mit 'stemming'	266
15.3	Befehle des Kapitels 'Textmining'	269

Abbildungsverzeichnis

1.1	Der Prozess der Datenanalyse	4
1.2	RStudio	5
1.3	So installiert man Pakete in RStudio	8
1.4	Hier werden Sie geholfen: Die Dokumentation der R-Pakete	10
1.5	Das Arbeitsverzeichnis mit RStudio auswählen	15
1.6	Sinnbild für die Deskriptiv- und die Inferenzstatistik	18
2.1	Daten sauber einlesen	24
2.2	Daten einlesen (importieren) mit RStudio	24
2.3	Trennzeichen einer CSV-Datei in RStudio einstellen	27
2.4	Schematische Darstellung eines Dataframes in Normalform	28
2.5	Dieselben Daten - einmal breit, einmal lang	28
2.6	Mit 'gather' und 'spread' wechselt man von der breiten Form zur langen Form	29
2.7	Ein Beispiel für eine Abbildung zu einer Normalform-Tabelle	29
3.1	Daten aufbereiten	34
3.2	Lego-Prinzip: Zerlege eine komplexe Struktur in einfache Bausteine	36
3.3	Durchpfeifen: Ein Dataframe wird von Operation zu Operation weitergereicht	36
3.4	Zeilen filtern	37
3.5	Spalten auswählen	39
3.6	Spalten sortieren	41
3.7	Datensätze nach Subgruppen aufteilen	43
3.8	Schematische Darstellung des 'Gruppieren - Zusammenfassen - Kombinieren'	44
3.9	Spalten zu einer Zahl zusammenfassen	45
3.10	Sinnbild für 'count'	49
3.11	Das ist keine Pfeife	52
3.12	Das 'Durchpfeifen'	52
3.13	Sinnbild für mutate	55
4.1	Ausreißer identifizieren	66
4.2	Ein Korrelationsplot	67
4.3	Visuelles Prüfen der Normalverteilung	69
4.4	Sinnbild für Umkodieren	70
4.5	Sinnbild zum 'Binnen'	70

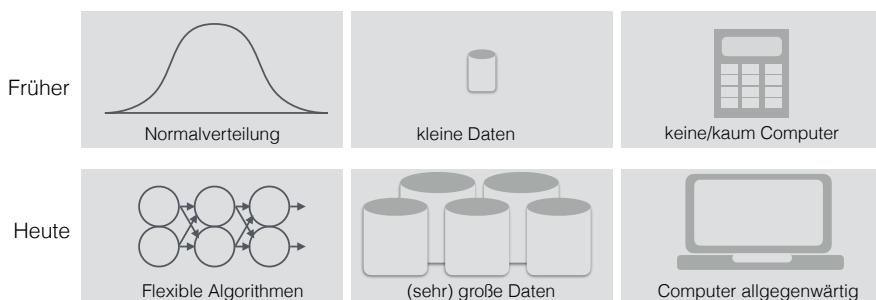
6.1	Das Anscombe-Quartett	86
6.2	Anatomie eines Diagramms	87
6.3	Mittleres Budget pro Jahr	89
6.4	Film-Budgets über die die Jahrzehnte	90
6.5	Verteilung des Budgets von Filmen	91
6.6	Überblick zu häufigen Diagrammtypen	107
6.7	Film-Budgets mit Histogrammen	109
7.1	Relative Häufigkeiten dargestellt anhand von Balkendiagrammen	118
7.2	Rotiertes Balkendiagramm mit Item-Label	119
7.3	... Mit der Brewer-Palette 17	121
8.1	Ein Modell eines VW-Käfers als Prototyp eines Modells	126
8.2	Modellieren	127
8.3	Formaleres Modell des Modellierens	128
8.4	Ein Beispiel für Modellieren	129
8.5	Ein Beispiel für ein Pfadmodell	129
8.6	Ein etwas aufwändigeres Modell	129
8.7	Modelle mit schwarzer Kiste	130
8.8	Die zwei Arten des ungeleiteten Modellierens	132
8.9	Welches Modell (Teil B-D; rot, grün, blau) passt am besten zu den Daten (Teil A) ?	134
8.10	'Mittlere' Komplexität hat die beste Vorhersagegenauigkeit (am wenigsten Fehler) in der Test-Stichprobe	135
8.11	Der Spagat zwischen Verzerrung und Varianz	136
8.12	Bias-Varianz-Abwägung. Links: Wenig Bias, viel Varianz. Rechts: Viel Bias, wenig Varianz.	140
9.1	Der größte Statistiker des 20. Jahrhunderts ($p < .05$)	144
9.2	Der p-Wert wird oft als wichtig erachtet	144
9.3	Mann und Papst zu sein ist nicht das gleiche.	145
9.4	Zwei Haupteinflüsse auf den p-Wert	146
9.5	Anteil von 'Kopf' bei wiederholtem Münzwurf	154
9.6	Die zwei Stufen der Bayes-Statistik in einem einfachen Beispieli	159
10.1	Beispiel für eine Regression	166
10.2	Zwei weitere Beispiele für Regressionen	167
10.3	Geringer (links) vs. hoher (rechts) Vorhersagefehler	169
10.4	Streudiagramm von Lernzeit und Klausurerfolg	172
10.5	Die Residuen verteilen sich hinreichend normal.	174
10.6	Vorhergesagte Werte vs. Residualwerte im Datensatz tips	175
10.7	Eine multivariate Analyse fördert Einsichten zu Tage, die bei einfacheren Analysen verborgen bleiben	179
10.8	Eine Regressionsanalyse mit Interaktionseffekten	181
11.1	Regressionsgerade für das Bestehen-Modell	187

11.2 Die logistische Regression beschreibt eine 's-förmige' Kurve	188
11.3 Modelldiagramm mit logistischer Regression	189
11.4 Eine ROC-Kurve	197
11.5 Beispiel für eine sehr gute (A), gute (B) und schlechte (C) Klassifikation	198
12.1 Überlebensraten auf der Titanic, in Abhängigkeit von der Passagierklasse . .	204
12.2 Logistische Regression zur Überlebensrate nach Passagierklasse	207
12.3 Absolute Überlebenshäufigkeiten	209
12.4 Relative Überlebenshäufigkeiten	210
12.5 Überlebenshäufigkeiten anhand eines Fliesenbildes dargestellt	211
12.6 Affären, mit Jitter	223
12.7 Affären, mit Smooth	224
13.1 Ein Streudiagramm - sehen Sie Gruppen (Cluster) ?	228
13.2 Ein Streudiagramm - mit drei Clustern	229
13.3 Unterschiedliche Anzahlen von Clustern im Vergleich	229
13.4 Die Summe der Varianz within in Abhängigkeit von der Anzahl von Clustern. Ein Screeplot	230
13.5 Distanz zwischen zwei Punkten in der Ebene	231
13.6 Pythagoras in 3D	231
13.7 Pythagoras in Reihe geschaltet	232
13.8 Schematische Darstellung zweier einfacher Clusterlösungen; links: geringe Varianz innerhalb der Cluster; rechts: hohe Varianz innerhalb der Cluster . . .	236
14.1 Der Pfeil ist eindimensional; reduziert also die drei Dimensionen auf eine . .	244
14.2 Screeplot	249
14.3 VSS-Screeplot	250
14.4 Ein Biplot für den Werte-Datensatz	251
14.5 Beispiel für eine rechtwinklige Rotation	254
14.6 Heatmap einer EFA	256
15.1 Illustration eines Tidy Text Dataframe	262

Vorwort



Statistik heute; was ist das? Sicherlich haben sich die Schwerpunkte von “gestern” zu “heute” verschoben. Wenig überraschend spielt der Computer eine immer größere Rolle; die Daten werden vielseitiger und massiger. Entsprechend sind neue Verfahren nötig - und vorhanden, in Teilen - um auf diese neue Situation einzugehen. Einige Verfahren werden daher weniger wichtig, z.B. der p-Wert oder der t-Test. Allerdings wird vielfach, zumeist, noch die Verfahren gelehrt und verwendet, die für die erste Hälfte des 20. Jahrhunderts entwickelt wurden. Eine Zeit, in der kleine Daten, ohne Hilfe von Computern und basierend auf einer kleinen Theoriefamilie im Rampenlicht standen (Cobb 2007). Die Zeiten haben sich geändert!



Zu Themen, die heute zu den dynamischsten Gebieten der Datenanalyse gehören, die aber früher keine große Rolle spielten, gehören (Hardin u. a. 2015):

- Nutzung von Datenbanken und anderen Data Warehouses
- Daten aus dem Internet automatisch einlesen (“scraping”)
- Genanalysen mit Tausenden von Variablen
- Gesichtserkennung

Sie werden in diesem Kurs einige praktische Aspekte der modernen Datenanalyse lernen. Ziel ist es, Sie - in Grundzügen - mit der Art und Weise vertraut zu machen, wie angewandte

Statistik bei führenden Organisationen und Praktikern verwendet wird¹.

Es ist ein Grundlagenkurs; das didaktische Konzept beruht auf einem induktiven, intuitiven Lehr-Lern-Ansatz. Formeln und mathematische Hintergründe sucht man meist vergebens (tja).

Im Gegensatz zu anderen Statistik-Büchern steht hier die Umsetzung mit R stark im Vordergrund. Dies hat pragmatische Gründe: Möchte man Daten einer statistischen Analyse unterziehen, so muss man sie zumeist erst aufbereiten; oft mühselig aufbereiten. Selten kann man den Luxus genießen, einfach “nur”, nach Herzenslust sozusagen, ein Feuerwerk an multivariater Statistik abzubrennen. Zuvor gilt es, die Daten aufzubereiten, umzuformen, zu prüfen und zusammenzufassen. Diesem Teil ist hier recht ausführlich Rechnung getragen.

“Statistical thinking” sollte, so eine verbreitete Idee, im Zentrum oder als Ziel einer Statistik-Ausbildung stehen (Wild und Pfannkuch 1999). Es ist die Hoffnung der Autoren dieses Skripts, dass das praktische Arbeiten (im Gegensatz zu einer theoretischen Fokus) zur Entwicklung einer Kompetenz im statistischen Denken beiträgt.

Außerdem spielt in diesem Kurs die Visualisierung von Daten eine große Rolle. Zum einen könnte der Grund einfach sein, dass Diagramme ansprechen und gefallen (einigen Menschen). Zum anderen bieten Diagramme bei umfangreichen Daten Einsichten, die sonst leicht wortwörtlich überersehen würden.

Dieser Kurs zielt auf die praktischen Aspekte der Analyse von Daten ab: “wie mache ich es?”; mathematische und philosophische Hintergründe werden vernachlässigt bzw. auf einschlägige Literatur verwiesen.

Dieses Skript ist publiziert unter CC-BY-NC-SA 3.0 DE².



Sebastian Sauer

Herausgeber: FOM Hochschule für Oekonomie & Management gemeinnützige GmbH

Dieses Skript dient als Begleitmaterial zum Modul “Praxis der Datenanalyse” des Masterstudiengangs “Wirtschaftspsychologie & Consulting” der FOM Hochschule für Oekonomie & Management.

FOM. Die Hochschule. Für Berufstätige. Die mit bundesweit über 42.000 Studierenden größte private Hochschule Deutschlands führt seit 1993 Studiengang für Berufstätige durch, die einen staatlich und international anerkannten Hochschulabschluss (Bachelor/Master) erlangen wollen. Weitere Informationen finden Sie unter <www.fom.de>

¹Statistiker, die dabei als Vorbild Pate standen sind: Roger D. Peng: <http://www.biostat.jhsph.edu/~rpeng/>, Hadley Wickham: <http://hadley.nz>, Jennifer Bryan: <https://github.com/jennybc>

²<https://creativecommons.org/licenses/by-nc-sa/3.0/de/>

Organisatorisches



0.1 Modulziele



Die Studierenden können nach erfolgreichem Abschluss des Moduls:

- den Ablauf eines Projekts aus der Datenanalyse in wesentlichen Schritten nachvollziehen,
- Daten aufbereiten und ansprechend visualisieren,
- Inferenzstatistik anwenden und kritisch hinterfragen,
- klassische Vorhersagemethoden (Regression) anwenden,
- moderne Methoden der angewandten Datenanalyse anwenden (z.B. Textmining),
- betriebswirtschaftliche Fragestellungen mittels datengetriebener Vorhersagemodelle beantworten.

0.2 Themen pro Termin

Für dieses Modul sind 44 UE für Lehre eingeplant, aufgeteilt in 11 Termine (vgl. 1).

Folgende Abfolge von Themen sind pro Termin vorgeschlagen:

Tabelle 1: Zuordnung von Themen zu Terminen

Termin	Thema (Kapitel)
1	Organisatorisches
1	Einführung
1	Rahmen
1	Daten einlesen
2	Datenjudo
3	Daten visualisieren
4	Fallstudie (z.B. zu 'movies')
5	Daten modellieren
5	Der p-Wert
6	Lineare Regression - metrisch
7	Lineare Regression - kategorial
8	Fallstudie (z.B. zu 'titanic' und 'affairs')
9	Vertiefung 1: Textmining oder Clusteranalyse
10	Vertiefung 2: Dimensionsreduktion
11	Wiederholung

Tabelle 1 ordnet die Themen des Moduls den Terminen (1-11) zu.

0.3 Vorerfahrung

Bei den Studierenden werden folgende Themen als bekannt vorausgesetzt:

- Deskriptive Statistik
- Inferenzstatistik
- Grundlagen R
- Grundlagen der Datenvisualisierung

0.4 Prüfung

0.4.1 Prüfungshinweise

- Die Prüfung besteht aus zwei Teilen
 - einer Klausur (50% der Teilnote)
 - einer Datenanalyse (50% der Teilnote).

Prüfungsrelevant ist der gesamte Stoff aus dem Skript und dem Unterricht mit folgenden Ausnahmen:

- Inhalte/Abschnitte, die als "nicht klausurrelevant" gekennzeichnet sind,

- Inhalte/Abschnitte, die als “Vertiefung” gekennzeichnet sind,
- Fallstudien (nur für Klausuren nicht prüfungsrelevant),
- die Inhalte von Links,
- die Inhalte von Fußnoten,
- die Kapitel *Vorwort*, *Organisatorisches* und *Anhang*.

Alle Hinweise zur Prüfung gelten nur insoweit nicht anders vom Dozenten festgelegt.

0.4.2 Klausur

- Die Klausur besteht fast oder komplett aus Multiple-Choice (MC-)Aufgaben mit mehreren Antwortoptionen (sofern nicht anders vom Dozenten vorgegeben).
- Die (maximale) Anzahl der richtigen Aussagen ist pro Aufgabe angegeben. Werden mehr Aussagen als “richtig” angekreuzt als angegeben, so wird die Aufgabe mit 0 Punkten beurteilt. Ansonsten werden Teipunkte für jede Aufgabe vergeben.
- Jede Aussage gilt ceteris paribus (unter sonst gleichen Umständen). Aussagen der Art “A ist B” (z.B. “Menschen sind sterblich”) sind *nur* dann als richtig auszuwählen, wenn die Aussage *immer* richtig ist.
- Im Zweifel ist eine Aussage auf den Stoff, so wie im Unterricht behandelt, zu beziehen. Werden in Aussagen Zahlen abgefragt, so sind Antworten auch dann richtig, wenn die vorgeschlagene Antwort ab der 1. Dezimale von der wahren Antwort abweicht (einigermaßen genaue Aussagen werden als richtig akzeptiert). Bei Fragen zu R-Syntax spielen Aspekte wie Enter-Taste o.ä. bei der Beantwortung der Frage keine Rolle; diese Aspekte dürfen zu ignorieren.
- Jede Aussage einer MC-Aufgabe ist entweder richtig oder falsch (aber nicht beides oder keines).
- Die MC-Aufgaben sind nur mit Kreuzen zu beantworten; Text wird bei der Korrektur nicht berücksichtigt.
- Bei Nachholklausuren gelten die selben Inhalte (inkl. Schwerpunkte) wie bei der Standard-Klausur, sofern nicht anderweitig angegeben.
- I.d.R. sind nur Klausurpapier und ein nicht-programmierbarer Taschenrechner als Hilfsmittel zulässig.
- Die Musterlösungen zu offenen Fragen sind elektronisch hinterlegt.

0.4.3 Datenanalyse

0.4.3.1 Hinweise

- Wenden Sie die passenden, im Modul eingeführten statistischen Verfahren an.

- Werten Sie die Daten mit R aus; R-Syntax soll verwendet und im Hauptteil dokumentiert werden.
- In der Wahl des Datensatzes sind Sie frei, mit folgender Ausnahme: Im Unterricht besprochene Datensätze dürfen nicht als Prüfungsleistung eingereicht werden (vgl. Abschnitt 1.1.5).
- Der (Original-)Name des Datensatzes (sowie ggf. Link) ist bei der Anmeldung anzugeben.
- Gruppenarbeiten sind nicht zulässig.
- Hat sich jemand schon für einen Datensatz angemeldet, so darf dieser Datensatz nicht mehr gewählt werden (“first come, first serve”).
- Fundorte für Datensätze sind z.B. hier³, hier⁴, hier⁵ und hier⁶; im Internet finden sich viele Datensätze⁷.
- Schreiben Sie Ihre Ergebnisse in einer Ausarbeitung zusammen; der Umfang der Ausarbeitung umfasst ca. 1500 Wörter (nur Hauptteil; d.h. exklusive Deckblatt, Verzeichnisse, Anhang etc.).
- Untersuchen Sie 2-3 Hypothesen.
- Denken Sie daran, Name, Matrikelnummer, Modulname etc. anzugeben (Deckblatt). Bei der Gestaltung des Layout entscheiden Sie selbstständig bitte nach Zweckmäßigkeit (und Ästhetik).
- Fügen Sie keine Erklärungen oder Definitionen von statistischen Verfahren an.

0.4.3.2 Gliederungsvorschlag zur Datenanalyse

1. Datensatz

1. Beschreibung

- Name
- Hintergrund (Themengebiet, Theorien, Relevanz), ca. 100 Wörter
- Dimension (Zeilen*Spalten)
- Zitation (wenn vorhanden)
- sonstige Hinweise (z.B. Datenqualität, Entstehung des Datensatzes)

2. Variablenbeschreibung (nur für Variablen der Hypothese)

- Skalenniveaus
- Kontinuität (nur bei metrischen Variablen)
- R-Datentyp

³<http://www.stat.ufl.edu/~winner/datasets.html>

⁴<http://archive.ics.uci.edu/ml/datasets.html>

⁵<https://www.kaggle.com/datasets>

⁶<http://vincentarelbundock.github.io/Rdatasets/datasets.html>

⁷Googeln Sie mal nach “open datasets” o.ä.

- Anzahl Fälle und fehlende Werte
 - Erläuterung der Variablen
2. Deduktive Analyse
 1. Hypothese(n) Beschreiben Sie die Vermutung(en), die Sie prüfen möchten, möglichst exakt.
 2. Deskriptive Statistiken
 - Berichten Sie deskriptive Statistiken für alle Variablen der Hypothesen.
 - Berichten Sie aber nur univariate Statistiken sowie Subgruppenanalysen dazu.
 - Berichten Sie ggf. Effektstärken.
 3. Diagramme
 - Visualisieren Sie Ihre Hypothese(n) bzw. die Daten dazu, gerne aus mehreren Blickwinkeln.
 4. Signifikanztest
 3. Explorative Analyse
 - Erörtern Sie interessante Einblicke, die über Ihre vorab getroffenen Hypothesen hinausgehen.
 - Diagramme können hier eine zentrale Rolle spielen.
 4. Diskussion
 1. Zentrale Ergebnisse Fassen Sie das zentrale Ergebnisse zusammen.
 2. Interpretation Interpretieren Sie die Ergebnisse: Was bedeuten die Zahlen/Fakten, die die Rechnungen ergeben haben?
 3. Grenzen der Analyse
 - Schildern Sie etwaige Schwachpunkte oder Einschränkungen der Analyse.
 - Geben Sie Anregungen für weiterführende Analysen dieses Datensatzes.

0.5 Literatur

Zum Bestehen der Prüfung ist keine weitere Literatur formal notwendig; allerdings ist es hilfreich, den Stoff aus unterschiedlichen Blickwinkeln aufzuarbeiten. Dazu ist am ehesten das Buch von Wickham und Grolemund (Wickham und Grolemund 2016) hilfreich, obwohl es deutlich tiefer geht als dieses Skript.

Teil I

Grundlagen

Kapitel 1

Rahmen



Lernziele:

- Einen Überblick über die fünf wesentliche Schritte der Datenanalyse gewinnen.
- R und RStudio installieren können.
- Einige häufige technische Probleme zu lösen wissen.
- R-Pakete installieren können.
- Einige grundlegende R-Funktionalitäten verstehen.
- Auf die Frage “Was ist Statistik?” eine Antwort geben können.

In diesem Skript geht es um die Praxis der Datenanalyse. Mit Rahmen ist das “Drumherum” oder der Kontext der eigentlichen Datenanalyse gemeint. Dazu gehören einige praktische Vorbereitungen und ein paar Überlegungen. Zum Beispiel brauchen wir einen Überblick über das Thema. Voilà (Abb. 1.1):

Datenanalyse, praktisch betrachtet, kann man in fünf Schritte einteilen (Wickham und Grolemund 2016). Zuerst muss man die Daten *einlesen*, die Daten also in R (oder einer anderen Software) verfügbar machen (laden). Fügen wir hinzu: In *schöner Form* verfügbar machen; man nennt dies auch *tidy data* (hört sich cooler an). Sobald die Daten in geeigneter

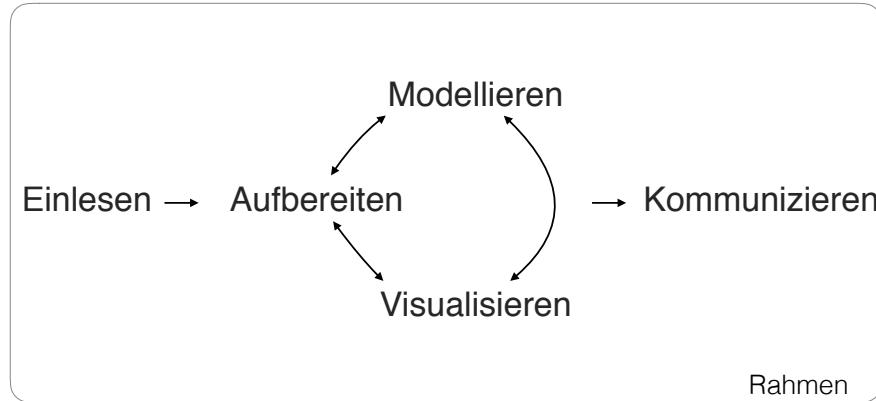


Abbildung 1.1: Der Prozess der Datenanalyse

Form in R geladen sind, folgt das *Aufbereiten*. Das beinhaltet Zusammenfassen, Umformen oder Anreichern je nach Bedarf. Ein nächster wesentlicher Schritt ist das *Visualisieren* der Daten. Ein Bild sagt bekanntlich mehr als viele Worte. Schließlich folgt das *Modellieren* oder das Hypothesen prüfen: Man überlegt sich, wie sich die Daten erklären lassen könnten. Zu beachten ist, dass diese drei Schritte - Aufbereiten, Visualisieren, Modellieren - keine starre Abfolge sind, sondern eher ein munteres Hin-und-Her-Springen, ein aufbauendes Abwechseln. Der letzte Schritt ist das *Kommunizieren* der Ergebnisse der Analyse - nicht der Daten. Niemand ist an Zahlenwüsten interessiert; es gilt, spannende Einblicke zu vermitteln.

Der Prozess der Datenanalyse vollzieht sich nicht im luftleeren Raum, sondern ist in einem *Rahmen* eingebettet. Dieser beinhaltet praktische Aspekte - wie Software, Datensätze - und grundsätzliche Überlegungen - wie Ziele und Grundannahmen.

1.1 Software installieren

Als Haupt-Analysewerkzeug nutzen wir R; daneben wird uns die sog. “Entwicklungsumgebung” RStudio einiges an komfortabler Funktionalität bescheren. Eine Reihe von R-Paketen (“Packages”; d.h. Erweiterungen) werden wir auch nutzen. R ist eine recht alte Sprache; viele Neuerungen finden in Paketen Niederschlag, da der “harte Kern” von R lieber nicht so stark geändert wird. Stellen Sie sich vor: Seit 29 Jahren nutzen Sie eine Befehl, der Ihnen einen Mittelwert ausrechnet, sagen wir die mittlere Anzahl von Tassen Kaffee am Tag. Und auf einmal wird der Mittelwert anders berechnet?! Eine Welt stürzt ein! Naja, vielleicht nicht ganz so tragisch in dem Beispiel, aber grundsätzlich sind Änderungen in viel benutzen Befehlen potenziell problematisch. Das ist wohl ein Grund, warum sich am “R-Kern” nicht so viel ändert. Die Innovationen in R passieren in den Paketen. Und es gibt viele davon; als ich diese Zeilen schreibe, sind es fast schon 10.000! Genauer: 9937 nach dieser Quelle: <https://cran.r-project.org/web/packages/>. Übrigens können R-Pakete auch Daten enthalten.

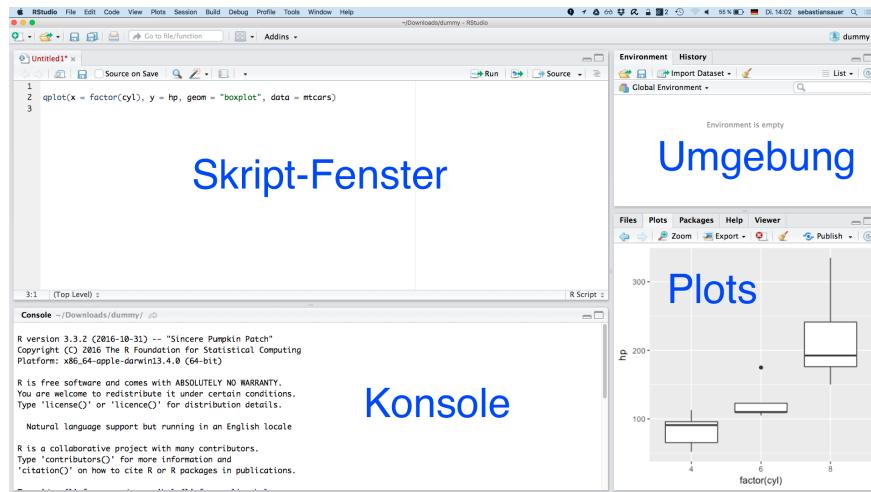


Abbildung 1.2: RStudio

1.1.1 R und RStudio installieren



Sie können R unter <https://cran.r-project.org> herunterladen und installieren (für Windows, Mac oder Linux). RStudio finden Sie auf der gleichnamigen Homepage: <https://www.rstudio.com>; laden Sie die “Desktop-Version” für Ihr Betriebssystem herunter.

RStudio ist “nur” eine Oberfläche (“GUI”) für R, mit einer R von praktischen Zusatzfunktionen. Die eigentlich Arbeit verrichtet das “normale” R, welches automatisch gestartet wird, wenn Sie RStudio starten (sofern R installiert ist).

Die Oberfläche von RStudio sieht (unter allen Betriebssystemen etwa gleich) so aus wie in Abbildung 1.2 dargestellt.

Das *Skript-Fenster* ähnelt einem normalem Text-Editor; praktischerweise finden Sie aber einen Button “run”, der die aktuelle Zeile oder die Auswahl “abschickt”, d.h. in die Konsole gibt, wo die Syntax ausgeführt wird. Wenn Sie ein Skript-Fenster öffnen möchten, so können Sie das Icon klicken (Alternativ: Ctrl-Shift-N oder File > New File > R Script).

Aus dem Fenster der *Konsole* spricht R zu uns bzw. wir mit R. Wird ein Befehl (synonym: *Funktion*) hier eingegeben, so führt R ihn aus. Es ist aber viel praktischer, Befehle in das Skript-Fenster einzugeben, als in die Konsole. Behalten Sie dieses Fenster im Blick, wenn Sie Antwort von R erwarten.

Im Fenster *Umgebung* (engl. “environment”) zeigt R, welche Variablen (Objekte) vorhanden

sind. Stellen Sie sich die Umgebung wie einen Karpfenteich vor, in dem die Datensätze und andere Objekte herumschwimmen. Was nicht in der Umgebung angezeigt wird, existiert nicht für R.

Im Fenster rechts unten werden mehrere Informationen bereit gestellt, z.B. werden Diagramme (Plots) dort ausgegeben. Klicken Sie mal die anderen Reiter im Fenster rechts unten durch.

Wer Shortcuts mag, wird in RStudio überschwänglich beschenkt; der Shortcut für die Shortcuts ist **Shift-Alt-K**.

Wenn Sie RStudio starten, startet R automatisch auch. Starten Sie daher, wenn Sie RStudio gestartet haben, *nicht* noch extra R. Damit hätten Sie sonst zwei Instanzen von R laufen, was zu Verwirrungen (bei R und beim Nutzer) führen kann.

1.1.2 Hilfe! R startet nicht!

Manntje, Manntje, Timpe Te,
 Buttje, Buttje inne See,
 myne Fru de Ilsebill
 will nich so, as ik wol will.

*Gebrüder Grimm, Märchen vom Fischer und seiner Frau*¹

Ihr R startet nicht oder nicht richtig? Die drei wichtigsten Heilmittel sind:

1. Schließen Sie die Augen für eine Minute. Denken Sie an etwas Schönes und was Rs Problem sein könnte.
2. Schalten Sie den Rechner aus und probieren Sie es morgen noch einmal.
3. Googeln.

Sorry für die schnoddrigen Tipps. Aber: Es passiert allzu leicht, dass man *Fehler* wie diese macht:



OH NO:

- `install.packages(dplyr)`
- `install.packages("dliar")`
- `install.packages("derpyler")`
- `install.packages("dplyr") # dependencies vergessen`
- Keine Internet-Verbindung
- `library(dplyr) # ohne vorher zu installieren`

¹https://de.wikipedia.org/wiki/Vom_Fischer_und_seiner_Frau

Wenn R oder RStudio dann immer noch nicht starten oder nicht richtig laufen, probieren Sie dieses:

- Sehen Sie eine Fehlermeldung, die von einem fehlenden Paket spricht (z.B. “Package ‘Rcpp’ not available”) oder davon spricht, dass ein Paket nicht installiert werden konnte (z.B. “Package ‘Rcpp’ could not be installed” oder “es gibt kein Paket namens ‘Rcpp’ ” oder “unable to move temporary installation XXX to YYY”), dann tun Sie folgendes:
 - Schließen Sie R und starten Sie es neu.
 - Installieren Sie das oder die angesprochenen Pakete mit `install.packages("name_des_pakets" dependencies = TRUE)` oder mit dem entsprechenden Klick in RStudio.
 - Starten Sie das entsprechende Paket mit `library(name_des_pakets)`.
- Gerade bei Windows 10 scheinen die Schreibrechte für R (und damit RStudio oder RCommander) eingeschränkt zu sein. Ohne Schreibrechte kann R aber nicht die Pakete (“packages”) installieren, die Sie für bestimmte R-Funktionen benötigen. Daher schließen Sie R bzw. RStudio und suchen Sie das Icon von R oder wenn Sie RStudio verwenden von RStudio. Rechtsklicken Sie das Icon und wählen Sie “als Administrator ausführen”. Damit geben Sie dem Programm Schreibrechte. Jetzt können Sie etwaige fehlende Pakete installieren.
- Ein weiterer Grund, warum R bzw. RStudio die Schreibrechte verwehrt werden könnten (und damit die Installation von Paketen), ist ein VirensScanner. Der VirensScanner sagt, nicht ganz zu Unrecht: “Moment, einfach hier Software zu installieren, das geht nicht, zu gefährlich”. Grundsätzlich gut, in diesem Fall unnötig. Schließen Sie R/RStudio und schalten Sie dann den VirensScanner *komplett* (!) aus. Öffnen Sie dann R/RStudio wieder und versuchen Sie fehlende Pakete zu installieren.

1.1.2.1 I am an outdated model

Verwenden Sie möglichst die neueste Version von R, RStudio und Ihres Betriebssystems. Ältere Versionen führen u.U. zu Problemen; je älter, desto Problem... Update Sie Ihre Packages regelmäßig z.B. mit `update.packages()` oder dem Button “Update” bei RStudio (Reiter **Packages**).

1.1.3 Pakete

Ein Großteil der Neuentwicklungen bei R passiert in sog. ‘Paketen’ (engl. *packages*), das sind Erweiterungen für R. Jeder, der sich berufen fühlt, kann ein R-Paket schreiben und es zum ‘R-Appstore’ (CRAN²) hochladen. Von dort kann es dann frei (frei wie in Bier) heruntergeladen werden.

Am einfachsten installiert man R-Pakete in RStudio über den Button “Install” im Reiter “Packages” (s. Abb. 1.3).

²<https://cran.r-project.org/>

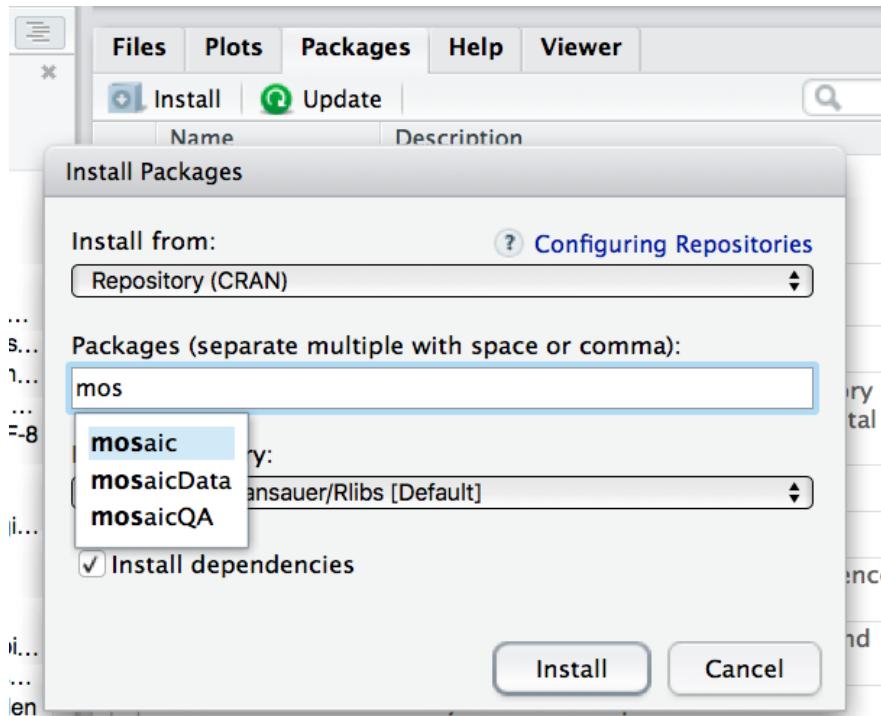


Abbildung 1.3: So installiert man Pakete in RStudio

Ein R-Paket, welches wir gleich benötigen, heißt `devtools`. Bitte installieren Sie es schon einmal (sofern noch nicht geschehen). Sie können auch folgenden Befehl verwenden, um Pakete zu installieren.

```
install.packages("devtools", dependencies = TRUE)
```

Aber einfacher geht es über die RStudio-Oberfläche.

Alle Pakete außer `devtools`, die wir hier benötigen, können über das R-Paket `prada` installiert werden. Sie müssen also nur noch das Paket `pradainstall` installieren. Mit dem Befehl `install_prada_pckgs` (aus dem Paket `prada`) werden dann ggf. eine Reihe weiterer Pakete installiert. Allerdings wohnt `prada` nicht im R-Appstore (CRAN), sondern bei Github³⁴. Um Pakete von Github zu installieren, nutzen wir diesen Befehl (Sie müssen natürlich online sein):

```
devtools::install_github("sebastiansauer/prada")
library(prada)
```

Sofern Sie online sind, sollte das Paket `prada` jetzt installiert sein. Installieren Sie jetzt alle Pakete, die für diesen Kurs benötigt werden mit dem Befehl `install_prada_pckgs()`.

³www.github.com

⁴einer Online-Plattform, auf der man Dateien bereistellen und ihre Veränderungen nachverfolgen kann



Beim Installieren von R-Paketen könnten Sie gefragt werden, welchen “Mirror” Sie verwenden möchten. Das hat folgenden Hintergrund: R-Pakete sind in einer Art “App-Store”, mit Namen CRAN (Comprehensive R Archive Network) gespeichert. Damit nicht ein armer, kleiner Server überlastet wird, wenn alle Studis dieser Welt just gerade beschließen, ein Paket herunterzuladen, gibt es viele Kopien dieses Servers - seine Spiegelbilder (engl. “mirrors”). Suchen Sie sich einfach einen aus, der in der Nähe ist.

Bei der Installation von Paketen mit `install.packages("name_des_pakets")` sollte stets der Parameter `dependencies = TRUE` angefügt werden. Also `install.packages("name_des_pakets", dependencies = TRUE)`. Hintergrund ist: Falls das zu installierende Paket seinerseits Pakete benötigt, die noch nicht installiert sind (gut möglich), dann werden diese sog. “dependencies” gleich mitinstalliert (wenn Sie `dependencies = TRUE` setzen).

Nicht vergessen: Installieren muss man eine Software *nur einmal; starten* (laden) muss man die R-Pakete jedes Mal, wenn man sie vorher geschlossen hat und wieder nutzen möchte.

Wenn Sie R bzw. RStudio schließen, werden alle Pakete ebenfalls geschlossen. Sie müssen die benötigten Pakete beim erneuten Öffnen von RStudio wieder starten.

```
library(dplyr)
```

Der Befehl bedeutet sinngemäß: “Hey R, geh in die Bücherei (library) und hole das Buch (package) dplyr!”.



Wann benutzt man bei R Anführungszeichen? Das ist etwas verwirrend im Detail, aber die Grundregel lautet: wenn man Text anspricht. Im Beispiel oben “library(dplyr)” ist “dplyr” hier erst mal für R nichts Bekanntes, weil noch nicht geladen. Demnach müssten *eigentlich* Anführungsstriche stehen. Allerdings meinte ein Programmierer, dass es doch so bequemer ist. Hat er Recht. Aber bedenken Sie, dass es sich um die Ausnahme einer Regel handelt. Sie können also auch schreiben: `library(“dplyr”)` oder `library(‘dplyr’)`; beides geht.

1.1.4 Vertiefung: Zuordnung von Paketen zu Befehlen

Woher weiß man, welche Befehle (oder auch Daten) in einem Paket enthalten sind?

Eine einfache Möglichkeit ist es, beim Reiter ‘Pakete’ auf den Namen eines der installierten Pakete zu klicken. Daraufhin öffnet sich die Dokumentation des Pakets und man sieht dort alle Befehle und Daten aufgeführt (s. Abbildung 1.4). Übrigens sehen Sie dort auch die Version eines Pakets (vielleicht sagt jemand mal zu Ihnen, “Sie sind ja outdated”, dann schauen Sie mal auf die die Paket-Versionen).

Für geladenen Pakete kann man auch den Befehl `help` nutzen, z.B. `help(ggplot2)`.

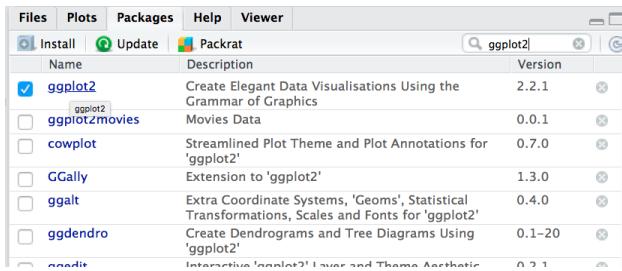


Abbildung 1.4: Hier werden Sie geholfen: Die Dokumentation der R-Pakete

Und umgekehrt, woher weiß ich, in welchem Paket ein Befehl ‘wohnt’?

Probieren Sie den Befehl `help.search("qplot")`, wenn Sie wissen möchten, in welchem Paket `qplot` zuhause ist. `help.search` sucht alle Hilfeseiten von *installierten* Paketen, in der der Suchbegriff irgendwie vorkommt. Um das Paket eines *geladenen* Befehl zu finden, hilft der Befehl `find: find("qplot")`.

Sie können auch den Befehl `find_funs` aus dem Paket `prada` nutzen:

```
prada::find_funs("select")
#> # A tibble: 5 x 3
#>   package_name builtin_package loaded
#>   <chr>          <lgl>    <lgl>
#> 1 dplyr           FALSE    FALSE
#> 2 MASS            TRUE     FALSE
#> 3 plotly          FALSE    FALSE
#> 4 raster          FALSE    FALSE
#> 5 VGAM            FALSE    FALSE
```

In diesem Skript sind am Ende jedes Kapitels die jeweils besprochenen (neuen) Befehle aufgeführt - inklusive ihres Paketes. Falls bei einem Befehl kein Paket angegeben ist, heißt das, dass der Befehl im ‘Standard-R’ wohnt - Sie müssen kein weiteres Paket laden⁵. Also zum Beispiel `ggplot2::qplot`: Der Befehl `qplot` ist im Paket `ggplot2` enthalten. Das Zeichen `::` trennt also Paket von Befehl.



Manche Befehle haben Allerweltsnamen (z.B. ‘filter’). Manchmal gibt es Befehle mit gleichem Namen in verschiedenen Paketen; besonders Befehle mit Allerweltsnamen (wie ‘filter’) sind betroffen (‘mosaic::filter’ vs. ‘dplyr::filter’). Falls Sie von wirre Ausgaben bekommen oder diffuse Fehlermeldung kann es sein, kann es sein, dass R einen Befehl mit dem richtigen Namen aber aus dem ‘falschen’ Paket zieht. Geben Sie im Zweifel lieber den Namen des Pakets vor dem Paketnamen an, z.B. so `dplyr::filter`. Der ‘doppelte Doppelpunkt’ trennt den Paketnamen vom Namen der Funktion.

⁵Eine Liste der Pakete, die beim Standard-R enthalten sind (also bereits installiert sind) finden Sie hier⁶

Außerdem sind zu Beginn jedes Kapitels die in diesem Kapitel benötigten Pakete angegeben. Wenn sie diese Pakete laden, werden alle Befehle dieses Kapitels funktionieren⁷.

Wie weiß ich, ob ein Paket geladen ist?

Wenn der Haken im Reiter ‘Packages’ gesetzt ist (s. Abbildung 1.4), dann ist das Paket geladen. Sonst nicht.

1.1.5 Datensätze

Die folgenden Datensätze sind entweder im Paket `prada` enthalten oder können aus anderen Paketen geladen werden. Um Daten aus einem Paket zu laden, gibt es den Befehl `data: data("name_datenobjekt", package = "Paketname")`. Also zum Beispiel:

```
data("stats_test", package = "prada")
```

Wenn ein bestimmtes Paket geladen ist, können Sie auch auf den Parameter `package = ...` verzichten, wenn ihr Datensatz in jedem Paket wohnt: Geladene Pakete werden vom Befehl `data` automatisch durchsucht.

Alternativ können Sie die Daten auch im Ordner `data` im Github-Repositorium herunterladen. Gehen Sie auf die Github-Seite dieses Kurses⁸, klicken Sie auf den großen grünen Button “Clone or download”, wählen Sie dann “Download ZIP”, um alle Dateien herunterzuladen. Nach dem Entzippen können Sie dann auf alle Dateien, inklusive Daten, zugreifen.

Die Daten dieses Kurses liegen im Ordner ‘data’.

- Datensatz `profiles` aus dem R-Paket `{okcupiddata}` (Kim und Escobedo-Land 2015); es handelt sich um Daten von einer Online-Singlebörsen
- Datensatz `stats_test` aus dem R-Paket `{prada}` (Sauer 2017a); es handelt sich um Ergebnisse einer Statistikklausur (einer Probeklausur)
- Datensatz `flights` aus dem R-Paket `{nycflights13}` (RITA 2013); es handelt sich um Abflüge von den New Yorker Flughäfen
- Datensatz `wo_men`, URL: <https://osf.io/ja9dw> (Sauer 2017b); es handelt sich um Körper- und Schuhgröße von Studierenden
- Datensatz `extra` aus dem R-Paket `{prada}` (Sauer 2016); es handelt sich die Ergebnisse einer Umfrage zu Extraversion
- Datensatz `titanic_train` aus dem Paket `{titanic}` von kaggle⁹; es handelt sich um Überlebensraten vom Titanic-Unglück.
- Datensatz `Affairs` aus dem Paket `{AER}` (Fair 1978); es handelt sich um eine Umfrage zu außerehelichen Affären.

⁷es sei denn, sie tun es nicht

⁸https://github.com/sebastiansauer/Praxis_der_Datenanalyse

⁹<https://www.kaggle.com/c/titanic/data>

Tabelle 1.1: Wichtige Datentypen in R

Name	Beschreibung	Beispiel
Name	Beschreibung	Beispiel
NULL	die leere Menge	NULL
logical	Logische Ausdrücke	TRUE
integer	Ganze Zahl	3
factor	nominale Variablen	"weiblich"
numeric	Reelle Zahl	2.71
character	Text	"Schorsch"

Wie man Daten in R ‘einlädt’ (Studierende sagen gerne ‘ins R hochladen’), besprechen wir im Kapitel 7.1.

1.2 ERRRstkontakt

1.2.1 R-Skript-Dateien

Ein neues *R-Skript* im RStudio können Sie z.B. öffnen mit **File-New File-R Script**. Schreiben Sie dort Ihre R-Befehle; Sie können die Skriptdatei speichern, öffnen, ausdrucken, übers Bett hängen... R-Skripte können Sie speichern (unter **File-Save**) und öffnen. R-Skripte sind einfache Textdateien, die jeder Texteditor verarbeiten kann. Nur statt der Endung **.txt**, sind R-Skripte stolzer Träger der Endung **.R**. Es bleibt aber eine schnöde Textdatei. Geben Sie Ihren R-Skript-Dateien die Endung “**.R**”, damit erkennt RStudio, dass es sich um ein R-Skript handelt und bietet ein paar praktische Funktionen wie den “Run-Button”.

1.2.2 Datentypen in R

Die (für diesen Kurs) wichtigsten Datentypen von R sind in Tabelle 1.1 aufgeführt (vgl. *(Programmieren mit R 2009)*).

All diese Datentypen (mit Ausnahme der leeren Menge) sind als *Vektoren* angelegt, also mehreren Elementen (z.B. Zahlen), die zu einem Ganzen (wie in einer Liste) verknüpft sind. *Faktoren* sind ganz interessant, weil die einzelnen Ausprägungen (*Faktorstufen* genannt) für R als Zahlen gespeichert sind (z.B. “Frau Müller und Herr Schorsch” = 1). Wenn ein Vektor aus 100 Mal diesem Text (“Frau Müller...”) besteht, muss R nur 100 mal 1 speichern und einmal die Zuordnung, was die 1 bedeutet. Spart Speicher. Außerdem kann man definieren, was alles eine Faktorstufe ist (z.B. nur “Mann” und “Frau”). Andere Eingaben sind dann nicht möglich; das kann praktisch sein, wenn man von vornherein nur bestimmte Ausprägungen zulassen möchte. Textvariablen sind da entspannter: Jeglicher Art von Text ist erlaubt. Text ist in R immer in Anführungszeichen (einfach oder doppelt) zu setzen.

Für die praktische Datenanalyse ist der `dataframe` (*Dataframe*; auch Datentabelle oder Datensatz) am wichtigsten. Grob gesagt handelt es sich dabei um eine Tabelle, wie man sie aus Excel kennt. Etwas genauer ist eine Kombination von Vektoren mit gleicher Länge, so dass eine ‘rechteckige’ Datenstruktur entsteht. Alle Spalten (d.h. Vektoren) haben einen Namen, so dass es ‘Spaltenköpfe’ gibt. Eine neuere Variante von Dataframes sind *tibbles* (Tibbles), die *auch* Dataframes sind, aber ein paar praktische Zusatzeigenschaften aufweisen (normale Dataframes können sich manchmal in einfache Vektoren auflösen; Tibbles tun dies nie).

1.2.3 Hinweise

Unser erster Kontakt mit R! Ein paar Anmerkungen vorweg:

- R unterscheidet zwischen Groß- und Kleinbuchstaben, d.h. `Oma` und `oma` sind zwei verschiedene Dinge für R!
- R verwendet den Punkt `.` als Dezimaltrennzeichen.
- Fehlende Werte werden in R durch `NA` kodiert.
- Kommentare werden mit dem Rautezeichen `#` eingeleitet; der Rest der Zeile von von R dann ignoriert.
- *Variablennamen* in R sollten mit Buchstaben beginnen; ansonsten dürfen nur Zahlen und Unterstriche `(_)` enthalten sein. Leerzeichen sollte man meiden. Das gilt auch für Spaltennamen.
- Um den Inhalt einer Variablen auszulesen, geben wir einfach den Namen des Objekts ein (und schicken den Befehl ab).
- Bleiben Sie konsistent, in der Art und Weise, wie Sie Ihre Syntax schreiben. Ein Vorschlag zum ‘Syntax-Stil’ finden Sie hier¹⁰.
- Variablen einen treffenden Namen zu geben, ist nicht immer leicht, aber wichtig. Namen sollten knapp, aber aussagekräftig sein.

```
# so nicht:
var
x
dummy
objekt
dieser_name_ist_etwas_lang_vielleicht

# gut:
tips_mw
lm1
```

¹⁰<http://adv-r.had.co.nz/Style.html>

1.2.4 Text und Variablen zuweisen

Man kann einer Variablen auch Text zuweisen (im Gegensatz zu Zahlen):

```
y <- "Hallo R!"
```

Man kann auch einer Variablen eine andere zuweisen:

```
y <- x
```

Wird jetzt y mit dem Inhalt von x überschrieben oder umgekehrt? Der Zuweisungspfeil `<-` macht die Richtung der Zuweisung ganz klar. Zwar ist in R das Gleichheitszeichen synonym zum Zuweisungspfeil erlaubt, aber der Zuweisungspfeil macht die Sache glasklar und sollte daher bevorzugt werden.

Man kann auch einer Variablen *mehr als* einen Wert zuweisen:

```
x <- c(1, 2, 3)
```

Dieser Befehl erzeugt eine “Spalte” (einen Vektor). Will man einer Variablen *mehr als* einen Wert zuweisen, muss man die Werte erst in einen Vektor “zusammen binden”; das geht mit dem Befehl `c` (vom engl. “*combine*”).

1.2.5 Funktionen aufrufen

Um einen *Befehl* (präziser aber synonym hier: eine Funktion) aufzurufen, geben wir ihren Namen an und definieren sog. *Parameter* in einer runden Klammer, z.B. so:

```
wo_men <- read.csv("data/wo_men.csv")
```

Allgemein gesprochen:

```
funktionsname(parametername1 = wert1, parametername2 = wert2, ...)
```

Die drei Punkte `...` sollen andeuten, dass evtl. weitere Parameter zu übergeben wären. Die Reihenfolge der Parameter ist *egal* - wenn man die Parameternamen anführt. Ansonsten muss man sich an die Standard-Reihenfolge, die eine Funktion vorgibt halten:

```
#ok:
wo_men <- read.csv(file = "data/wo_men.csv", header = TRUE, sep = ",")
wo_men <- read.csv("data/wo_men.csv", TRUE, ",")
wo_men <- read.csv(header = TRUE, sep = ",", file = "data/wo_men.csv")
```

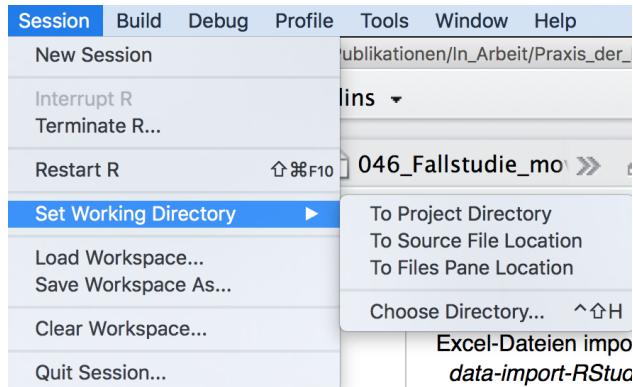


Abbildung 1.5: Das Arbeitsverzeichnis mit RStudio auswählen

```
# ohne:
wo_men <- read.csv(TRUE, "data/wo_men.csv", ",")
```

In der Hilfe zu einem Befehl findet man die Standard-Syntax inklusive der möglichen Parameter, ihrer Reihenfolge und Standardwerten (default values) von Parametern. Zum Beispiel ist beim Befehl `read.csv` der Standardwert für `sep` mit ; voreingestellt (schauen Sie mal in der Hilfe nach). Gibt man einen Parameter nicht an, für den ein Standardwert eingestellt ist, ‘befüllt’ R den Parameter mit diesem Standardwert.

1.2.6 Das Arbeitsverzeichnis

Das aktuelle Verzeichnis (Arbeitsverzeichnis; “working directory”) kann man mit `getwd()` erfragen und mit `setwd()` einstellen. Komfortabler ist es aber, das aktuelle Verzeichnis per Menü zu ändern (vgl. Abb. 1.5. In RStudio: Session > Set Working Directory > Choose Directory ... (oder per Shortcut, der dort angezeigt wird)).

Es ist praktisch, das Arbeitsverzeichnis festzulegen, denn dann kann man z.B. eine Datendatei einlesen, ohne den Pfad eingeben zu müssen:

```
# nicht ausführen:
daten_deutsch <- read.csv("daten_deutsch.csv", sep = ";", dec = ".")
```

R geht dann davon aus, dass sich die Datei `daten_deutsch.csv` im Arbeitsverzeichnis befindet.

Für diesen Kurs ist es sinnvoll, das Arbeitsverzeichnis in einen “Hauptordner” zu legen (z.B. “Praxis_der_Datenanalyse”), in dem Daten und sonstiges Material als Unterordner abgelegt sind.



Übrigens: Wenn Sie keinen Pfad angeben, so geht R davon aus, dass die Daten im aktuellen Verzeichnis (dem *working directory*) liegen.

1.3 Hier werden Sie geholfen

1.3.1 Wo finde ich Hilfe?

Es ist keine Schande, nicht alle Befehle der ca. 10,000 R-Pakete auswendig zu wissen. Schlauer ist, zu wissen, wo man Antworten findet. Hier eine Auswahl:

- Zu diesen Paketen gibt es gute “Spickzettel” (cheatsheets): ggplot2, RMarkdown, dplyr, tidyr. Klicken Sie dazu in RStudio auf *Help > Cheatsheets > ...* oder gehen Sie auf <https://www.rstudio.com/resources/cheatsheets/>.
- In RStudio gibt es eine Reihe (viele) von Tastaturkürzeln (Shortcuts), die Sie hier finden: *Tools > Keyboard Shortcuts Help*.
- Für jeden Befehl aus einem *geladenen* Paket können Sie mit `help()` die Hilfe-Dokumentation anschauen, also z.B. `help("qplot")`.
- Im Internet finden sich zuhauf Tutorials.
- Der Reiter “Help” bei RStudio verweist auf die Hilfe-Seite des jeweiligen Pakets bzw. Befehls.
- Die bekannteste Seite um Fragen rund um R zu diskutieren ist <http://stackoverflow.com>.

1.3.2 Einfache reproduzierbare Beispiele (ERBies)

Sagen wir, Sie haben ein Problem. Mit R. Bevor Sie jemanden bitten, Ihr Problem zu lösen, haben Sie schon drei dreizehn dreißig Minuten recherchiert, ohne Erfolg. Sie entschließen sich, bei Stackoverflow¹¹ Ihr Problem zu posten. Außerdem kann sicher eine Mail zu einem Bekannten, einem Dozenten oder sonstwem, der sich auskennen sollte, nicht schaden. Sie formulieren also Ihr Problem: “Hallo, mein R startet nicht, und wenn es startet, dann macht es nicht, was ich soll, außerdem funktioniert der Befehl ‘mean’ bei mir nicht. Bitte lös mein Problem!”. Seltsamerweise reagieren die Empfänger Ihrer Nachricht nicht alle begeistert. Stattdessen verlangt jemand (dreist) nach einer genauen Beschreibung Ihres Problems, mit dem Hinweis, dass “Ferndiagnosen” schwierig sein. Genauer gesagt möchte ihr potenzieller Helfer ein ‘minimal reproducible example’ (MRE) oder, Deutsch, ein *einfaches reproduzierbares Beispiel* (ERBie).

¹¹ www.stackoverflow.com

Wenn Sie jemanden um R-Hilfe bitten, dann sollten Sie Ihr Problem prägnant beschreiben.

Was sollte alles in einem ERBie enthalten sein?

Ein ERBie besteht aus vier Teilen: Syntax, Daten, Paketen und Infos zum laufenden System (R Version etc.)

Wie sollte so ein ERBie aussehen? Ich empfehle, folgende Eckpunkte zu beachten¹²:

- Syntax: Stellen Sie die R-Syntax bereit, die ein Problem bereit (d.h. die einen Fehler liefert).
- Einfach: Geben Sie sogenig Syntax wie möglich an. Es bereitet Ihrem Helfer nur wenig Spaß, sich durch 2000 Zeilen Code zu wühlen, wenn es 10 Zeilen auch getan hätten.
- Reproduzierbar Geben Sie soviel Syntax wie nötig, um den Fehler zu erzeugen (aber nicht mehr).
- Schreiben Sie Ihre Syntax übersichtlich, verständlich und kommentiert; z.B. sollten die Variablennamen informativ sein.
- Beschreiben Sie den Fehler genau (“läuft nicht” reicht nicht); z.B. ist es hilfreich, den Wortlaut einer Fehlermeldung bereitzustellen.
- Zu Beginn der Syntax sollten die benötigten Pakete geladen werden.
- Zu Ende des ERBie sollte der Output von `sessionInfo()` einkopiert werden; damit werden Informationen zum laufenden System (wie Version von R, Betriebssystem etc.) bereitgestellt.
- Beziehen Sie sich möglichst auf Daten, die in R schon “eingebaut sind” wie die Datensätze `iris` oder `mtcars`.

Natürlich sollte man immer erst selbst nach einer Lösung recherchieren, bevor man jemanden um Hilfe bittet. Viele Fragen wurden schon einmal diskutiert und oft auch gelöst.

1.4 Was ist Statistik? Wozu ist sie gut?

Zwei Fragen bieten sich am Anfang der Beschäftigung mit jedem Thema an: Was ist die Essenz des Themas? Warum ist das Thema (oder die Beschäftigung damit) wichtig?

Was ist Statistik? Eine Antwort dazu ist, dass Statistik die Wissenschaft von Sammlung, Analyse, Interpretation und Kommunikation von Daten ist mithilfe mathematischer Verfahren ist und zur Entscheidungshilfe beitragen solle (*The Oxford Dictionary of Statistical Terms 2006*; Romeijn 2016). Damit hätten wir auch den Unterschied zur schnöden Datenanalyse (ein Teil der Statistik) herausgemiedelt. Statistik wird häufig in die zwei Gebiete *deskriptive* und *inferierende* Statistik eingeteilt (vgl. Abb. 1.6). Erstere fasst viele Zahlen zusammen, so dass wir den Wald statt vieler Bäume sehen. Letztere verallgemeinert von den vorliegenden (sog. “Stichproben-“)Daten auf eine zugrunde liegende Grundmenge (Population). Dabei spielt die Wahrscheinlichkeitsrechnung (Stochastik) eine große Rolle.

¹²Hier finden Sie weitere Hinweise zu ERBies: <https://stackoverflow.com/help/mcve> oder <https://gist.github.com/hadley/270442>

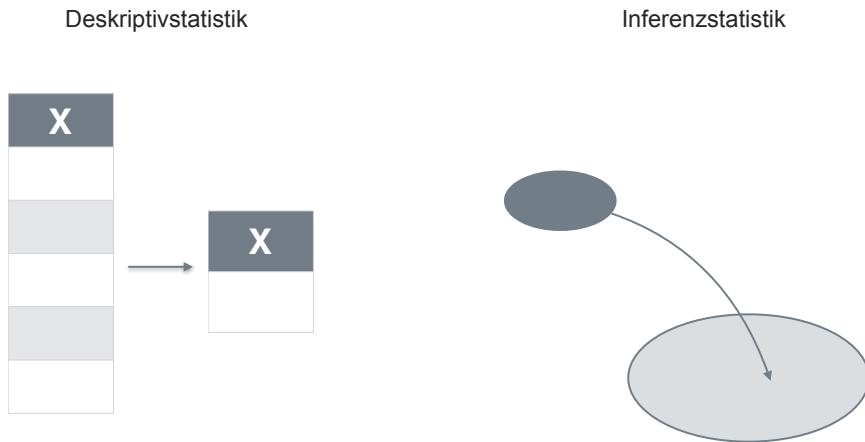


Abbildung 1.6: Sinnbild für die Deskriptiv- und die Inferenzstatistik

Aufgabe der deskriptiven Statistik ist es primär, Daten prägnant zusammenzufassen. Aufgabe der Inferenzstatistik ist es, zu prüfen, ob Daten einer Stichprobe auf eine Grundgesamtheit verallgemeinert werden können.

Dabei lässt sich der Begriff "Statistik" als Überbegriff von "Datenanalyse" verstehen, wenn diese Sicht auch nicht von allen geteilt wird (Gromelund und Wickham 2014). In diesem Buch steht die Aufbereitung, Analyse, Interpretation und Kommunikation von Daten im Vordergrund. Liegt der Schwerpunkt dieser Aktivitäten bei computerintensiven Methoden, so wird auch von *Data Science* gesprochen, wobei der Begriff nicht einheitlich verwendet wird (Wickham und Gromelund 2016; Hardin u. a. 2015)

Daten kann man definieren als *Informationen, die in einem Kontext stehen* (Moore 1990), wobei eine numerische Konnotation mitschwingt.

Modellieren kann man als *zentrale Aufgabe von Statistik* begreifen (Cobb 2007; Gromelund und Wickham 2014). Einfach gesprochen, bedeutet Modellieren in diesem Sinne, ein mathematisches Narrativ ("Geschichte") zu finden, welches als Erklärung für gewisse Muster in den Daten fungiert; vgl. Kap. 8.

Statistisches Modellieren läuft gewöhnlich nach folgendem Muster ab (Gromelund und Wickham 2014):

Prämissen 1: Wenn Modell M wahr ist, dann sollten die Daten das Muster D aufweisen.
 Prämissen 2: Die Daten weisen das Muster D auf.

Konklusion: Daher muss das Modell M wahr sein.

Die Konklusion ist *nicht* zwangsläufig richtig. Es ist falsch zu sagen, dass dieses Argumentationsmuster - Abduktion (Peirce 1955) genannt - wahre, sichere Schlüsse (Konklusionen) liefert. Die Konklusion *kann, muss aber nicht*, zutreffen.

Ein Beispiel: Auf dem Nachhauseweg eines langen Arbeitstags wartet, in einer dunklen Ecke, ein Mann, der sich als Statistik-Professor vorstellt und Sie zu einem Glücksspiel einlädt. Sofort sagen Sie zu. Der Statistiker will 10 Mal eine Münze werfen, er setzt auf Zahl (versteht

sich). Wenn er gewinnt, bekommt er 10€ von Ihnen; gewinnen Sie, bekommen Sie 11€ von ihm. Hört sich gut an, oder? Nun wirft er die Münze zehn Mal. Was passiert? Er gewinnt 10 Mal, natürlich (so will es die Geschichte). Sollten wir glauben, dass er ein Betrüger ist?

Ein Modell, welches wir hier verwenden könnten, lautet: Wenn die Münze gezinkt ist (Modell M zutrifft), dann wäre diese Datenlage D (10 von 10 Treffern) wahrscheinlich - Prämisse 1. Datenlage D ist tatsächlich der Fall; der Statistiker hat 10 von 10 Treffer erzielt - Prämisse 2. Die Daten D "passen" also zum Modell M; man entscheidet sich, dass der Professor ein Falschspieler ist.

Wichtig zu erkennen ist, dass Abduktion mit dem Wörtchen *wenn* beginnt. Also davon *ausgeht*, dass ein Modell M der Fall ist (der Professor also tatsächlich ein Betrüger ist). Das, worüber wir entscheiden wollen, wird bereits vorausgesetzt. Falls M gilt, gehen wir mal davon aus, wie gut passen dann die Daten dazu?

Wie gut passen die Daten D zum Modell M?

Das ist die Frage, die hier tatsächlich gestellt bzw. beantwortet wird.

Natürlich ist es keineswegs sicher, *dass* das Modell gilt. Darüber macht die Abduktion auch keine Aussage. Es könnte also sein, dass ein anderes Modell zutrifft: Der Professor könnte ein Heiliger sein, der uns auf etwas merkwürdige Art versucht, Geld zuzuschanzen... Oder er hat einfach Glück gehabt.

Statistische Modelle beantworten i.d.R. nicht, wie wahrscheinlich es ist, dass ein Modell gilt. Statistische Modelle beurteilen, wie gut Daten zu einem Modell passen.

Häufig trifft ein Modell eine Reihe von Annahmen, die nicht immer explizit gemacht werden, aber die klar sein sollten. Z.B. sind die Münzwürfe unabhängig voneinander? Oder kann es sein, dass sich die Münze "einschießt" auf eine Seite? Dann wären die Münzwürfe nicht unabhängig voneinander. In diesem Fall klingt das reichlich unplausibel; in anderen Fällen kann dies eher der Fall sein¹³. Auch wenn die Münzwürfe unabhängig voneinander sind, ist die Wahrscheinlichkeit für Zahl jedes Mal gleich? Hier ist es wiederum unwahrscheinlich, dass sich die Münze verändert, ihre Masse verlagert, so dass eine Seite Unwucht bekommt. In anderen Situationen können sich Untersuchungsobjekte verändern (Menschen lernen manchmal etwas, sagt man), so dass die Wahrscheinlichkeiten für ein Ereignis unterschiedlich sein können, man dies aber nicht berücksichtigt.

1.5 Aufgaben

1. Öffnen Sie das Cheatsheet für RStudio und machen Sie sich mit dem Cheatsheet vertraut.

¹³Sind z.B. die Prüfungsergebnisse von Schülern unabhängig voneinander? Möglicherweise haben sie von einem "Superschüler" abgeschrieben. Wenn der Superschüler viel weiß, dann zeigen die Abschreiber auch gute Leistung.

2. Sichten Sie kurz die übrigen Cheatsheets; später werden die Ihnen vielleicht von Nutzen sein.
3. Führen Sie diese Syntax aus:

```
meine_coole_variable <- 10
meine_coole_variable
```

Woher röhrt der Fehler?

4. Korrigieren Sie die Syntax:

```
install.packages(dplyr)
```

```
y <- Hallo R!
```

```
Hallo R <- 1
```

```
Hallo_R <- 1
```

Richtig oder Falsch???¹⁴



Richtig oder Falsch!?

1. Statistik wird gemeinhin in zwei Bereiche unterteilt: Deskriktivstatistik und Inferenzstatistik.
2. Unter Deskriktivstatistik versteht man, Daten zu beschreiben. Dazu ist jede Art von Beschreibung sinnvoll, vorausgesetzt es wird eine konsistente Regel eingesetzt.
3. Unter Abduktion versteht man den Schluss vom Allgemeinen auf das Konkrete.
4. Wirft jemand bei 10 von 10 Münzwürfen ‘Kopf’, so muss er ein Betrüger sein.
5. Wirft jemand bei 10 von 10 Münzwürfen ‘Kopf’, so ist die Wahrscheinlichkeit groß, dass er ein Betrüger ist.

1.6 Befehlsübersicht

Tabelle 1.2 stellt die Befehle dieses Kapitels dar.

Diese Befehle “wohnen” alle im Standard-R; es ist für diese Befehle nicht nötig, zusätzliche Pakete zu installieren/ laden.

¹⁴R, F: die Daten müssen sinnvoll zusammengefasst werden, F, F, F: Wenn er ehrlich sein sollte, dann ist das Ereignis ‘10 von 10’ selten

Tabelle 1.2: Befehle des Kapitels 'Rahmen'

Paket::Funktion	Beschreibung
install.packages("x")	Installiert Paket "x" (nicht: Paket "X")
library	lädt ein Paket
<-	Weist einer Variablen einen Wert zu
c	erstellt eine Spalte/ einen Vektor

1.7 Verweise

- Chester Ismay erläutert einige Grundlagen von R und RStudio, die für Datenanalyse hilfreich sind: <https://bookdown.org/chesterismay/rbasics/>.
- Roger Peng und Kollegen bieten hier einen Einstieg in Data Science mit R: <https://bookdown.org/rdpeng/artofdatascience/>
- Wickham und Grolemund (2016) geben einen hervorragenden Überblick in das Thema dieses Buches; ihr Buch ist sehr zu empfehlen.
- Wer einen stärker an der Statistik orientierten Zugang sucht, aber “mathematisch sanft” behandelt werden möchte, wird bei James et al. (2013b) glücklich oder zumindest fündig werden.
- Uwe Ligges (*Programmieren mit R* 2009) ‘Programmieren mit R’ gibt einen tieferen Einstieg in die Grundlagen von R.
- Wer ganz tief ein- und abtauchen möchte in R, dem sei - solide Grundkenntnisse vorausgesetzt - Hadley Wickhams Wickham (2014a) ‘Advanced R’ ans Herz gelegt.

Kapitel 2

Daten einlesen



Lernziele:

- Wissen, auf welchen Wegen man Daten in R hineinbekommt.
- Wissen, was eine CSV-Datei ist.
- Wissen, was UTF-8 bedeutet.
- Erläutern können, was R unter dem “working directory” versteht.
- Erkennen können, ob eine Tabelle in Normalform vorliegt.
- Daten aus R hinauskriegen (exportieren).

In diesem Kapitel werden folgende Pakete benötigt:

```
library(readr) # Daten einlesen  
library(tidyverse) # Datenjudo und Visualisierung
```

Dieses Kapitel beantwortet eine Frage: “Wie kriege ich Daten in vernünftiger Form in R hinein?“.

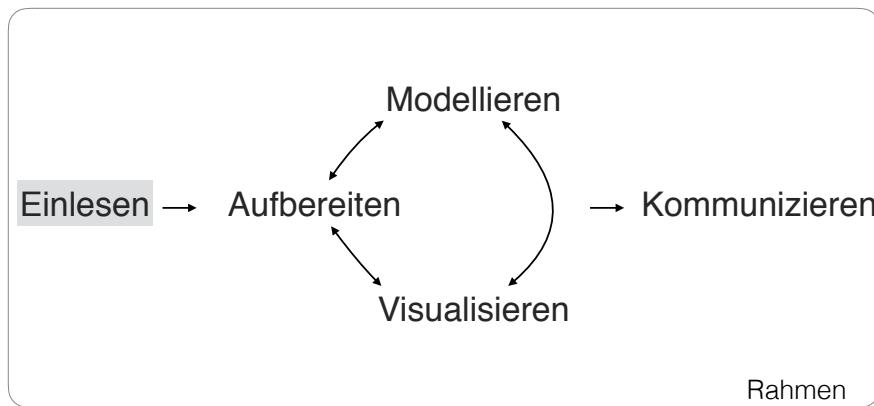


Abbildung 2.1: Daten sauber einlesen

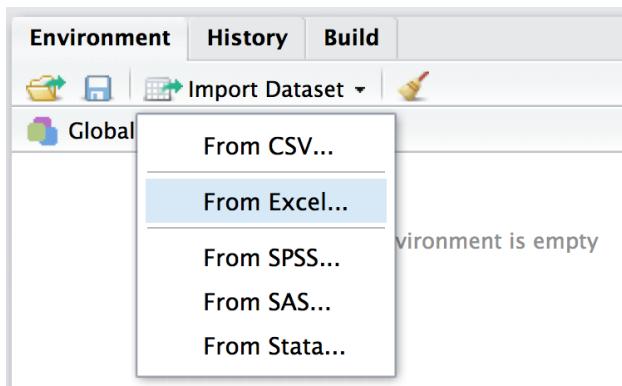


Abbildung 2.2: Daten einlesen (importieren) mit RStudio

2.1 Daten in R importieren

In R kann man ohne Weiteres verschiedene, gebräuchliche (Excel oder CSV) oder weniger gebräuchliche (Feather¹) Datenformate einlesen. In RStudio lässt sich dies z.B. durch einen schnellen Klick auf **Import Dataset** im Reiter **Environment** erledigen².

2.1.1 Excel-Dateien importieren

Am einfachsten ist es, eine Excel-Datei (.xls oder .xlsx) über die RStudio-Oberfläche zu importieren; das ist mit ein paar Klicks geschehen³:

Es ist für bestimmte Zwecke sinnvoll, nicht zu klicken, sondern die Syntax einzutippen. Zum Beispiel: Wenn Sie die komplette Analyse als Syntax in einer Datei haben (eine sog.

¹<https://cran.r-project.org/web/packages/feather/index.html>

²Um CSV-Dateien zu laden wird durch den Klick im Hintergrund das Paket **readr** verwendet (Wickham, Hester, und Francois 2016a); die entsprechende Syntax wird in der Konsole ausgegeben, so dass man sie sich anschauen und weiterverwenden kann

³im Hintergrund wird das Paket **readxl** verwendet

“Skriptdatei”), dann brauchen Sie (in RStudio) nur alles auszuwählen und auf Run zu klicken, und die komplette Analyse läuft durch! Die Erfahrung zeigt, dass das ein praktisches Vorgehen ist.



Daten (CSV, Excel,...) können Sie *nicht* öffnen über File > Open File Dieser Weg ist Skript-Dateien und R-Daten-Objekten vorbehalten.

2.1.2 Daten aus R-Paketen importieren

In R-Paketen wohnen nicht nur Funktionen, sondern auch Daten. Diese Daten kann man mit dem Befehl `data` laden, dem man den Namen des zu ladenen Datensatzes `dataset` und seines Heimatpaket `paket` übergibt: `data(dataset, package = "paket")`. Natürlich muss das Paket installiert sein. Zum Beispiel:

```
data(movies, package = "ggplot2movies")
```

2.1.3 Daten im R-Format laden

Das R-Datenformat erkennt man an der R-Endung `.rda` oder `RData`. Dateien mit diesem Format kann man in RStudio über File > Open File... öffnen. Oder mit dem Befehl `load(file)`, wobei `file` der Dateiname ist, also z.B. `extra.RData`. Mit dem Schwesterbefehl `save` können Sie ein Objekt im R-Datenformat speichern, z.B. `save(stats_test, file = "stats_test.RData")`.

2.1.4 CSV-Dateien importieren

Die gebräuchlichste Form von Daten für statistische Analysen ist wahrscheinlich das CSV-Format. Das ist ein einfaches Format, basierend auf einer Textdatei. Schauen Sie sich mal diesen Auszug aus einer CSV-Datei an.

```
row_number,date_time,study_time,self_eval,interest,score
1,05.01.2017 13:57:01,5,8,5,29
2,05.01.2017 21:07:56,3,7,3,29
3,05.01.2017 23:33:47,5,10,6,40
4,06.01.2017 09:58:05,2,3,2,18
5,06.01.2017 14:13:08,4,8,6,34
6,06.01.2017 14:21:18,NA,NA,NA,39
```

Erkennen Sie das Muster? Die erste Zeile gibt die “Spaltenköpfe” wieder, also die Namen der Variablen. Hier sind es 6 Spalten; die fünft heißt “score” und gibt die Punkte eines Studierenden in einer Statistikklausur wieder. Die Spalten sind offenbar durch Komma ,

voneinander getrennt. Dezimalstellen sind in amerikanischer Manier mit einem Punkt . dargestellt. Die Daten sind “rechteckig”; alle Spalten haben gleich viele Zeilen und umgekehrt alle Spalten gleich viele Zeilen. Man kann sich diese Tabelle gut als Excel-Tabelle mit Zellen vorstellen, in denen z.B. “row_number” (Zelle oben links) oder “39” (Zelle unten rechts) steht.

An einigen Stelle steht NA. Das ist Errisch für “fehlender Wert”. Häufig wird die Zelle auch leer gelassen, um auszudrücken, dass ein Wert hier fehlt (hört sich nicht ganz doof an). Aber man findet alle möglichen Ideen, um fehlende Werte darzustellen. Ich rate von allen anderen ab; führt nur zu Verwirrung.

Lesen wir diese Daten jetzt ein:

```
df <- read.csv("data/stats_test.csv")
```

Übrigens, Sie können die Daten (als CSV) für diesen Kurs auch über diese URL importieren. Z.B. den Datensatz stats_test:

```
prada_stats_test_url <-
  paste0("https://raw.github.com/", # Webseite
         "sebastiansauer/", # Nutzer
         "Praxis_der_Datenanalyse/", # Projekt/Repository
         "master/", # Variante
         "data/stats_test.csv") # Ordner und Dateinamen

stats_test <- read.csv(prada_stats_test_url)
```

Analog gehen Sie für die anderen PraDa-Datensätze vor (vgl. Kapitel 1.1.5).

2.1.4.1 Vorsicht bei nicht-amerikanisch kodierten Textdateien

Der Befehl `read.csv` liest also eine CSV-Datei, was uns jetzt nicht übermäßig überrascht. Aber Achtung: Wenn Sie aus einem Excel mit *deutscher* Einstellung eine CSV-Datei exportieren, wird diese CSV-Datei als Spaltentrennung ; (Strichpunkt) und als Dezimaltrennzeichen , verwenden. Da der Befehl `read.csv` laut amerikanischen Standard mit Komma als Spaltentrennung und Punkt als Dezimaltrennzeichen arbeitet, müssen wir die deutschen Sonderzeichen explizit angeben, z.B. so:

```
# nicht ausführen:
daten_deutsch <- read.csv("daten_deutsch.csv", sep = ";", dec = ".")
```

Dabei steht `sep` (separator) für das Trennzeichen zwischen den Spalten und `dec` für das Dezimaltrennzeichen. R bietet eine Kurzfassung für `read.csv` mit diesen Parametern: `read.csv2("daten_deutsch.csv")`.

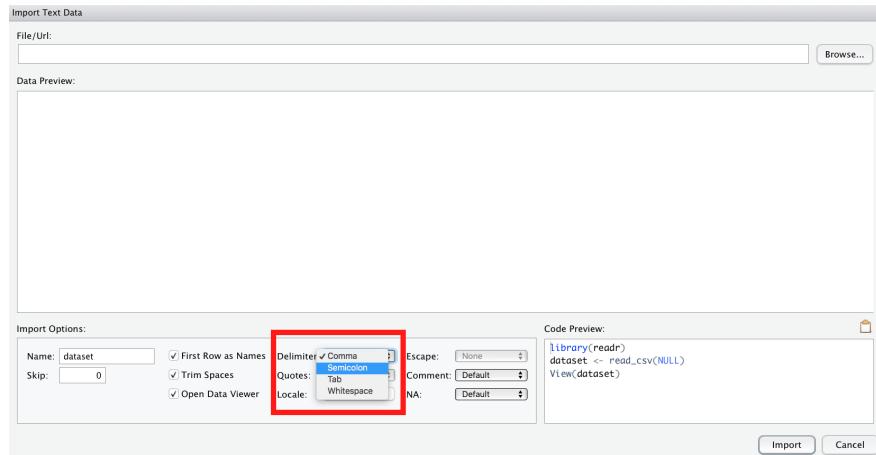


Abbildung 2.3: Trennzeichen einer CSV-Datei in RStudio einstellen

Man kommt hier auch mit “Klicken statt Tippen” zum Ziel; in der Maske von “Import Dataset” (für CSV-Dateien) gibt es den Auswahlpunkt “Delimiter” (Trennzeichen). Dort kann man das Komma durch einen Strichpunkt (oder was auch immer) ersetzen. Es hilft, im Zweifel, die Textdatei vorab mit einem Texteditor zu öffnen.

2.2 Normalform einer Tabelle

Tabellen in R werden als `data frames` (“Dataframe” auf Denglisch; moderner: als `tibble`, Tibble kurz für “Table-df”) bezeichnet. Tabellen sollten in “Normalform” vorliegen (“tidy”), bevor wir weitere Analysen starten. Unter Normalform verstehen sich folgende Punkte:

- Es handelt sich um einen Dataframe, also um eine Tabelle mit Spalten mit Namen und gleicher Länge; eine Datentabelle in rechteckiger Form und die Spalten haben einen Namen.
- In jeder Zeile steht eine Beobachtung, in jeder Spalte eine Variable.
- Fehlende Werte sollten sich in *leeren* Zellen niederschlagen.
- Daten sollten nicht mit Farbmarkierungen o.ä. kodiert werden.
- Es gibt keine Leerzeilen und keine Leerspalten.
- In jeder Zelle steht ein Wert.
- Am besten verwendet man keine Sonderzeichen verwenden und keine Leerzeichen in Variablennamen und -werten, sondern nur Ziffern und Buchstaben und Unterstriche.
- Variablennamen dürfen nicht mit einer Zahl beginnen.

Abbildung 2.4 visualisiert die Bestimmungsstücke eines Dataframes (Wickham und Grolemund 2016):

Der Punkt *Jede Zeile eine Beobachtung, jede Spalte eine Variable, jede Zelle ein Wert* verdient besondere Beachtung. Betrachten Sie das Beispiel in Abbildung 2.5.

In der rechten Tabelle sind die Variablen `Quartal` und `Umsatz` klar getrennt; jede hat ihre

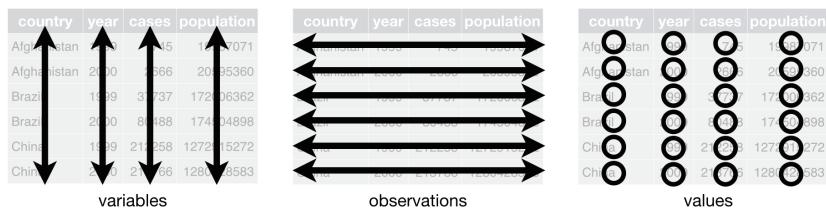


Abbildung 2.4: Schematische Darstellung eines Dataframes in Normalform

ID	Q1	Q2	Q3	Q4
1	123	342	431	675
2	324	342	234	345
3	343	124	456	465
...				

Breit

ID	Quartal	Umsatz
1	Q1	342
2	Q2	342
3	...	124
...	Q1	342
	Q2	342
	Q3	124
	...	

Lang

Abbildung 2.5: Dieselben Daten - einmal breit, einmal lang

eigene Spalte. In der linken Tabelle hingegen sind die beiden Variablen vermischt. Sie haben nicht mehr ihre eigene Spalte, sondern sind über vier Spalten verteilt. Die rechte Tabelle ist ein Beispiel für eine Tabelle in Normalform, die linke nicht.

2.3 Tabelle in Normalform bringen

Eine der ersten Aktionen einer Datenanalyse sollte also die “Normalisierung” Ihrer Tabelle sein. In R bietet sich dazu das Paket `tidyverse` an, mit dem die Tabelle von *Breit- auf Langformat* (und wieder zurück) geschoben werden kann.

Abb. 2.6 zeigt ein Beispiel dazu.

Warum ist es wichtig, von der “breiten” (links in Abb. 2.6) zur “langen” oder “Normalform” (rechts in Abb. 2.6) zu wechseln. Ganz einfach: viele Befehle (allgemeiner: Tätigkeiten) verlangen die Normalform; hin und wieder sind aber die Tabellen von ihrem Schöpfer in breiter Form geschaffen worden. Zum Beispiel erwartet `ggplot2` - und viele andere Diagrammbefehle - dass man *einer* Achse *eine* Spalte (Variable) zuweist, z.B. die Variable “Umsatz” auf die

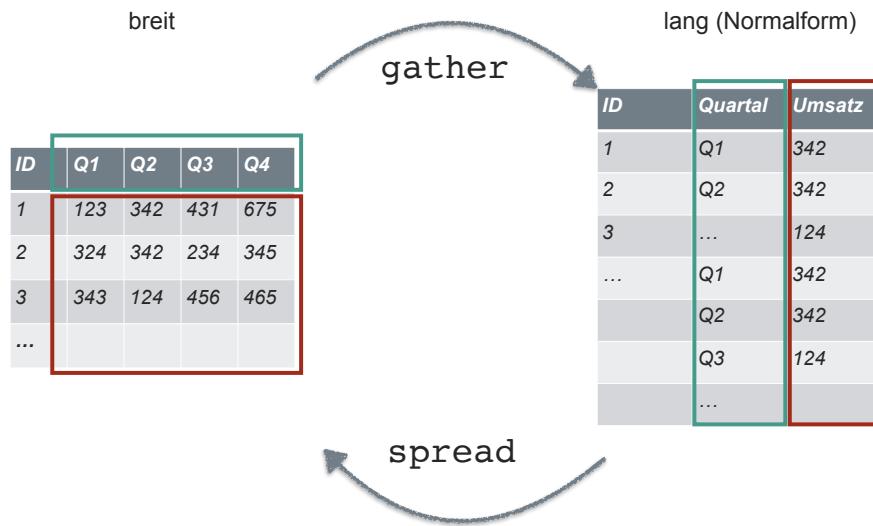


Abbildung 2.6: Mit 'gather' und 'spread' wechselt man von der breiten Form zur langen Form

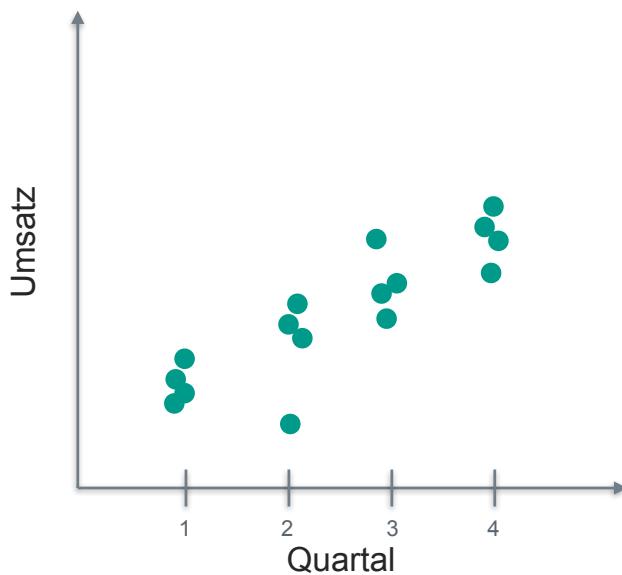


Abbildung 2.7: Ein Beispiel für eine Abbildung zu einer Normalform-Tabelle

Y-Achse. Der X-Achse könnten wir dann z.B. die Variable "Quartal" packen (s. Abb. 2.7).

Um von der breiten Form zur langen Form zu kommen, kann man den Befehl `tidyverse::gather` nehmen. Von der langen Form zur breiten Form gibt es `tidyverse::spread`. Also etwa:

```
library(tidyverse)
df_lang <- gather(df_breit, key = "Quartal", value = "Umsatz")

df_breit <- spread(df_lang, Quartal, Umsatz)
```

Dabei baut `gather` den Dataframe so um, dass nur zwei Spalten übrig bleiben (s. Abb. 2.6).

Eine Spalte nur *Spaltennamen* (“Q1”, “Q2”, ...) enthält; diese Spalte nennt `gather` im Standard `key`. Die zweite Spalte enthält die Werte (z.B. Umsätze), die vormals über mehrere Spalten verstreut waren. Diese Spalte heißt per Default `value`. Im Beispiel oben macht die Spalte ID bei dem Spiel “Aus vielen Spalten werden zwei” nicht mit. Möchte man eine Spalte aussparen, so schreibt man das bei `gather` so:

```
df_lang <- gather(df_breit, key = "Quartal", value = "Umsatz", -ID)
```

In Kapitel 5 werden wir dazu ein Fallstudie einüben.

2.4 Textkodierung

Öffnet man eine Textdatei mit einem Texteditor seiner Wahl, so sieht man... Text und sonst nichts, also keine Formatierung etc. Eine Textdatei besteht aus Text und sonst nichts (daher der Name...). Auch eine R-Skript-Datei (`Coole_Syntax.R`) ist eine Textdatei. Technisch gesprochen werden nur die Textzeichen gespeichert, sonst nichts; im Gegensatz dazu speichert eine Word-Datei noch mehr, z.B. Formatierung. Jetzt steht in der Textdatei der Code “42” für den nächsten Buchstaben. Ja, ist das jetzt ein “A”, oder ein “Ä” oder vielleicht ein griechischer Buchstabe? Woher weiß der Computer das eigentlich? Die Antwort ist: Er braucht eine Art Übersetzungstabelle oder Kodierungstafel. Mehrere solcher Kodierungstafeln existieren. Die gebräuchlichste im Internet heißt *UTF-8*⁴. Leider benutzen unterschiedliche Betriebssysteme unterschiedliche Kodierungstafeln, was zu Verwirrung führt. Ich empfehle, Ihre Textdateien als UTF-8 zu kodieren. RStudio fragt Sie, wie eine Textdatei kodiert werden soll. Sie können auch unter `File > Save with Encoding...` die Kodierung einer Textdatei festlegen.

Speichern Sie R-Textdateien wie Skripte stets mit UTF-8-Kodierung ab.

Wie bekommt man seine Daten wieder aus R raus (“ich will zu Excel zurück!”)?

Eine Möglichkeit bietet die Funktion `write.csv`; sie schreibt eine CSV-Datei:

```
write.csv(name_der_tabelle, "Dateiname.csv")
```

Mit `help(write.csv)` bekommt man mehr Hinweise dazu. Beachten Sie, dass immer in das aktuelle Arbeitsverzeichnis geschrieben wird.

2.5 Befehlsübersicht

Tabelle 2.1 stellt die Befehle dieses Kapitels dar.

⁴<https://de.wikipedia.org/wiki/UTF-8>

Tabelle 2.1: Befehle des Kapitels 'Daten einlesen'

Paket::Funktion	Beschreibung
read.csv	Liest eine CSV-Datei ein.
write.csv	Schreibt einen Dataframe in eine CSV-Datei.
tidyr::gather	Macht aus einem "breiten" Dataframe einen "langen".
tidyr::separate	"Zieht" Spalten auseinander.

2.6 Aufgaben⁵



Richtig oder Falsch!?

1. In CSV-Dateien dürfen Spalten *nie* durch Komma getrennt sein.
2. RStudio bietet die Möglichkeit, CSV-Dateien per Klick zu importieren.
3. RStudio bietet *nicht* die Möglichkeit, CSV-Dateien per Klick zu importieren.
4. "Deutsche" CSV-Dateien verwenden als Spalten-Trennzeichen einen Strichpunkt.
5. In einer Tabelle in Normalform stehen in jeder Zeile eine Beobachtung.
6. In einer Tabelle in Normalform stehen in jeder Spalte eine Variable.
7. R stellt fehlende Werte mit einem Fragezeichen ? dar.
8. Um Excel-Dateien zu importieren, kann man den Befehl `read.csv` verwenden.

2.7 Verweise

- *R for Data Science* bietet umfangreiche Unterstützung zu diesem Thema (Wickham und Grolemund 2016).

⁵F, R, F, R, R, R, F, F

Kapitel 3

Datenjudo



Lernziele:

- Die zentralen Ideen der Datenanalyse mit dplyr verstehen.
- Typische Probleme der Datenanalyse schildern können.
- Zentrale dplyr-Befehle anwenden können.
- dplyr-Befehle kombinieren können.
- Die Pfeife anwenden können.
- Werte umkodieren und “binnen” können.

In diesem Kapitel werden folgende Pakete benötigt:

```
library(tidyverse)  # Datenjudo
library(stringr)   # Texte bearbeiten
library(car)       # für 'recode'
library(desctable) # Statistiken auf einen Streich
library(lsr)        # für Befehl `aad`
```

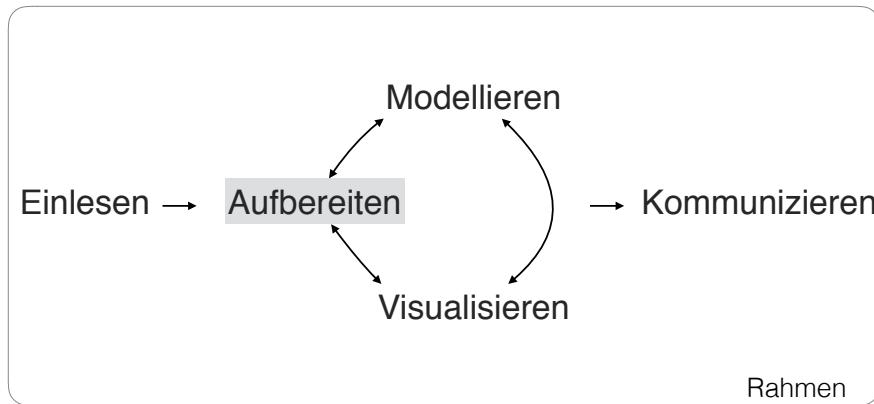


Abbildung 3.1: Daten aufbereiten

Das Paket `tidyverse` lädt `dplyr`, `ggplot2` und weitere Pakete¹. Daher ist es komfortabler, `tidyverse` zu laden, damit spart man sich Tipparbeit. Die eigentliche Funktionalität, die wir in diesem Kapitel nutzen, kommt aus dem Paket `dplyr`.

Mit *Datenjudo* ist gemeint, die Daten für die eigentliche Analyse “aufzubereiten”. Unter *Aufbereiten* ist hier das Umformen, Prüfen, Bereinigen, Gruppieren und Zusammenfassen von Daten gemeint. Die deskriptive Statistik fällt unter die Rubrik Aufbereiten. Kurz gesagt: Alles, was tut, nachdem die Daten “da” sind und bevor man mit anspruchsvoller(er) Modellierung beginnt.

Ist das Aufbereiten von Daten auch nicht statistisch anspruchsvoll, so ist es trotzdem von großer Bedeutung und häufig recht zeitintensiv. Eine Anekdote zur Relevanz der Datenaufbereitung, die (so will es die Geschichte) mir an einer Bar nach einer einschlägigen Konferenz erzählt wurde (daher keine Quellenangabe, Sie verstehen...). Eine Computerwissenschaftlerin aus den USA (deutschen Ursprungs) hatte einen beeindruckenden “Track Record” an Siegen in Wettkämpfen der Datenanalyse. Tatsächlich hatte sie keine besonderen, raffinierten Modellierungstechniken eingesetzt; klassische Regression war ihre Methode der Wahl. Bei einem Wettkampf, bei dem es darum ging, Krebsfälle aus Krankendaten vorherzusagen (z.B. von Röntgenbildern) fand sie nach langem Datenjudo heraus, dass in die “ID-Variablen” Information gesickert war, die dort nicht hingehörte und die sie nutzen konnte für überraschend (aus Sicht der Mitstreiter) gute Vorhersagen zu Krebsfällen. Wie war das möglich? Die Daten stammten aus mehreren Kliniken, jede Klinik verwendete ein anderes System, um IDs für Patienten zu erstellen. Überall waren die IDs stark genug, um die Anonymität der Patienten sicherzustellen, aber gleich wohl konnte man (nach einigem Judo) unterscheiden, welche ID von welcher Klinik stammte. Was das bringt? Einige Kliniken waren reine Screening-Zentren, die die Normalbevölkerung versorgte. Dort sind wenig Krebsfälle zu erwarten. Andere Kliniken jedoch waren Onkologie-Zentren für bereits bekannte Patienten oder für Patienten mit besonderer Risikolage. Wenig überraschen, dass man dann höhere Krebsraten vorhersagen kann. Eigentlich ganz einfach; besondere Mathe steht hier (zumindest in dieser Geschichte) nicht dahinter. Und, wenn man den Trick kennt, ganz einfach. Aber wie so oft ist es nicht leicht, den Trick zu finden. Sorgfältiges Datenjudo hat hier den Schlüssel zum Erfolg gebracht.

¹für eine Liste s. `tidyverse_packages(include_self = TRUE)`

3.1 Typische Probleme der Datenaufbereitung

Bevor man seine Statistik-Trickkiste so richtig schön aufmachen kann, muss man die Daten häufig erst noch in Form bringen. Das ist nicht schwierig in dem Sinne, dass es um komplizierte Mathe ginge. Allerdings braucht es mitunter recht viel Zeit und ein paar (oder viele) handwerkliche Tricks sind hilfreich. Hier soll das folgende Kapitel helfen.

Typische Probleme, die immer wieder auftreten, sind:

- *Fehlende Werte*: Irgend jemand hat auf eine meiner schönen Fragen in der Umfrage nicht geantwortet!
- *Unerwartete Daten*: Auf die Frage, wie viele Facebook-Freunde er oder sie habe, schrieb die Person “I like you a lot”. Was tun???
- *Daten müssen umgeformt werden*: Für jede der beiden Gruppen seiner Studie hat Joachim einen Google-Forms-Fragebogen aufgesetzt. Jetzt hat er zwei Tabellen, die er “verheiraten” möchte. Geht das?
- *Neue Variablen (Spalten) berechnen*: Ein Student fragt nach der Anzahl der richtigen Aufgaben in der Statistik-Probelektionsur. Wir wollen helfen und im entsprechenden Datensatz eine Spalte erzeugen, in der pro Person die Anzahl der richtig beantworteten Fragen steht.

3.2 Daten aufbereiten mit `dplyr`

Willkommen in der Welt von `dplyr`! `dplyr` hat seinen Namen, weil es sich ausschließlich um Dataframes bemüht; es erwartet einen Dataframe als Eingabe und gibt einen Dataframe zurück (zumindest bei den meisten Befehlen).

3.2.1 Die zwei Prinzipien von `dplyr`

Es gibt viele Möglichkeiten, Daten mit R aufzubereiten; `dplyr`² ist ein populäres Paket dafür. `dplyr` basiert auf zwei Ideen:

1. *Lego-Prinzip* Komplexe Datenanalysen in Bausteine zerlegen (vgl. Abb. 3.2).
2. *Durchpfeifen*: Alle Operationen werden nur auf Dataframes angewendet; jede Operation erwartet einen Dataframe als Eingabe und gibt wieder einen Dataframe aus (vgl. Abb. 3.3).

Das *erste Prinzip* von `dplyr` ist, dass es nur ein paar *wenige Grundbausteine* geben sollte, die sich gut kombinieren lassen. Sprich: Wenige grundlegende Funktionen mit eng umgrenzter Funktionalität. Der Autor, Hadley Wickham, sprach einmal in einem Forum (citation needed...), dass diese Befehle wenig können, das Wenige aber gut. Ein Nachteil dieser Konzeption kann sein, dass man recht viele dieser Bausteine kombinieren muss, um zum gewünschten

²<https://cran.r-project.org/web/packages/dplyr/index.html>

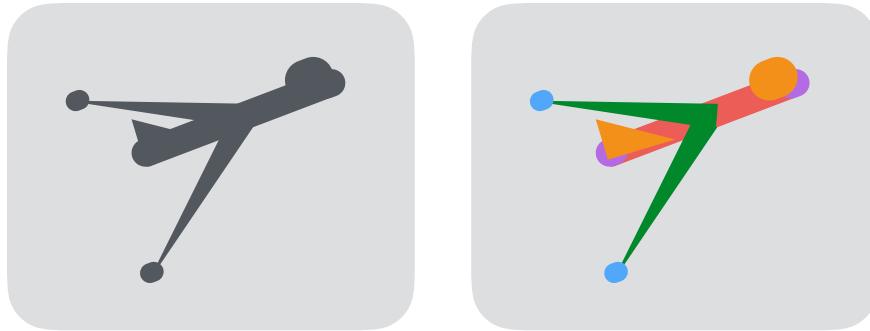


Abbildung 3.2: Lego-Prinzip: Zerlege eine komplexe Struktur in einfache Bausteine

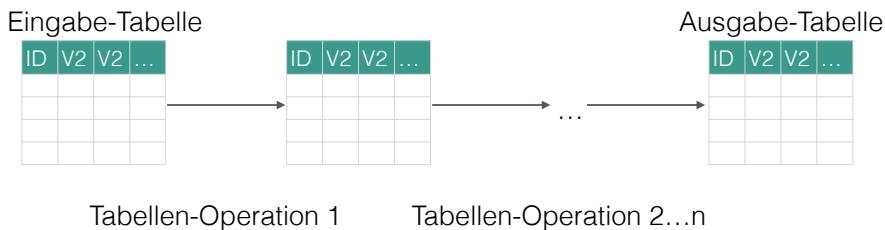


Abbildung 3.3: Durchpfeifen: Ein Dataframe wird von Operation zu Operation weitergereicht

Ergebnis zu kommen. Außerdem muss man die Logik des Baukastens gut verstanden haben - die Lernkurve ist also erstmal steiler. Dafür ist man dann nicht darauf angewiesen, dass es irgendwo "Mrs Right" gibt, die genau das kann, was ich will. Außerdem braucht man sich auch nicht viele Funktionen merken. Es reicht einen kleinen Satz an Funktionen zu kennen (die praktischerweise konsistent in Syntax und Methodik sind). Diese Bausteine sind typische Tätigkeiten im Umgang mit Daten; nichts Überraschendes. Wir schauen wir uns diese Bausteine gleich näher an.

Das *zweite Prinzip* von `dplyr` ist es, einen Dataframe von Operation zu Operation *durchzureichen*. `dplyr` arbeitet also *nur* mit Dataframes. Jeder Arbeitsschritt bei `dplyr` erwartet einen Dataframe als Eingabe und gibt im Gegenzug wieder einen Dataframe aus.

Werfen wir einen Blick auf ein paar typische Bausteine von `dplyr`.

3.3 Zentrale Bausteine von `dplyr`

3.3.1 Zeilen filtern mit `filter`

Häufig will man bestimmte Zeilen aus einer Tabelle filtern; `filter`. Zum Beispiel man arbeitet für die Zigarettenindustrie und ist nur an den Rauchern interessiert (die im Übrigen unser Gesundheitssystem retten (Krämer 2011)), nicht an Nicht-Rauchern; es sollen die nur Umsatzzahlen des letzten Quartals untersucht werden, nicht die vorherigen Quartale; es sollen nur die Daten aus Labor X (nicht Labor Y) ausgewertet werden etc.

ID	Name	Note1
1	Anna	1
2	Anna	1
3	Berta	2
4	Carla	2
5	Carla	2

ID	Name	Note1
1	Anna	1
2	Anna	1

Abbildung 3.4: Zeilen filtern

Abb. 3.4 zeigt ein Sinnbild für `filter`.

Merke:

Die Funktion `filter` filtert Zeilen aus einem Dataframe.

Schauen wir uns einige Beispiel an; zuerst die Daten laden nicht vergessen. Achtung: "Wohnen" die Daten in einem Paket, muss dieses Paket installiert sein, damit man auf die Daten zugreifen kann.

```
data(profiles, package = "okcupiddata") # Das Paket muss installiert sein
```

```
df_frauen <- filter(profiles, sex == "f") # nur die Frauen
df_alt <- filter(profiles, age > 70) # nur die alten Menschen
df_alte_frauen <- filter(profiles, age > 70, sex == "f")
# nur die alten Frauen, d.h. UND-Verknüpfung
df_mittelalt <- filter(profiles, between(age, 35, 60))
# zwischen 35 und 60

df_nosmoke_nodrinks <- filter(profiles, smokes == "no" | drinks == "not at all")
# liefert alle Personen, die Nicht-Raucher *oder* Nicht-Trinker sind
```

Gar nicht so schwer, oder? Allgemeiner gesprochen werden diejenigen Zeilen gefiltert (also behalten bzw. zurückgeliefert), für die das Filterkriterium TRUE ist.

`filter` ist deutlich einfacher (und klarer) als Standard-R. Vergleichen Sie mal:

```
filter(profiles, age > 70, sex == "f", drugs == "sometimes")
```

```
# base-R:
```

```
profiles[df$age > 70 & df$sex == "f" & df$drugs == "sometimes", ]
```



Manche Befehle wie `filter` haben einen Allerweltsnamen; gut möglich, dass ein Befehl mit gleichem Namen in einem anderen (geladenen) Paket existiert. Das kann dann zu Verwirrungen führen - und kryptischen Fehlern. Im Zweifel den Namen des richtigen Pakets ergänzen, und zwar zum Beispiel so: `dplyr::filter(...)`.

3.3.1.1 Aufgaben³



Richtig oder Falsch!?

1. `filter` filtert Spalten.
2. `filter` ist eine Funktion aus dem Paket `dplyr`.
3. `filter` erwartet als ersten Parameter das Filterkriterium.
4. `filter` lässt nur ein Filterkriterium zu.
5. Möchte man aus dem Datensatz `profiles` (`okcupiddata`) die Frauen filtern, so ist folgende Syntax korrekt: `filter(profiles, sex == "f")`.
6. `filter(profiles, age > 35 | age > 60)` filtert die mittelalten Frauen (zwischen 35 und 60)

3.3.2 Spalten wählen mit `select`

Das Gegenstück zu `filter` ist `select`; dieser Befehl liefert die gewählten Spalten zurück. Das ist häufig praktisch, wenn der Datensatz sehr “breit” ist, also viele Spalten enthält. Dann kann es übersichtlicher sein, sich nur die relevanten auszuwählen. Abb. 3.5 zeigt Sinnbild für diesen Befehl:

Merke:

Die Funktion `select` wählt Spalten aus einem Dataframe aus.

Laden wir als ersten einen Datensatz.

```
stats_test <- read.csv("data/stats_test.csv")
```

Dieser Datensatz beinhaltet Daten zu einer Statistikklausur.

³F, R, F, F, R, R

vorher					nachher		
ID	Name	N1	N2	N3	ID	Name	N1
1	Anna	1	2	3	1	Anna	1
2	Berta	1	1	1	2	Berta	1
3	Carla	2	3	4	3	Carla	2
...

Abbildung 3.5: Spalten auswählen

```
select(stats_test, score) # Spalte `score` auswählen
select(stats_test, score, study_time)
# Spalten `score` und `study_time` auswählen

select(stats_test, score:study_time) # dito
select(stats_test, 5:6) # Spalten 5 bis 6 auswählen
```

Tatsächlich ist der Befehl `select` sehr flexibel; es gibt viele Möglichkeiten, Spalten auszuwählen. Im `dplyr`-Cheatsheet⁴ findet sich ein guter Überblick dazu.

3.3.2.1 Aufgaben⁵



Richtig oder Falsch!?

1. `select` wählt *Zeilen* aus.
2. `select` ist eine Funktion aus dem Paket `knitr`.
3. Möchte man zwei Spalten auswählen, so ist folgende Syntax prinzipiell korrekt:
`select(df, spalte1, spalte2)`.
4. Möchte man Spalten 1 bis 10 auswählen, so ist folgende Syntax prinzipiell korrekt:
`'select(df, spalte1:spalte10)`
5. Mit `select` können Spalten nur bei ihrem Namen, aber nicht bei ihrer Nummer aufgerufen werden.

⁴<https://www.rstudio.com/resources/cheatsheets/>

⁵F, F, R, R, F

3.3.3 Zeilen sortieren mit `arrange`

Man kann zwei Arten des Umgangs mit R unterscheiden: Zum einen der “interaktive Gebrauch” und zum anderen “richtiges Programmieren”. Im interaktiven Gebrauch geht es uns darum, die Fragen zum aktuell vorliegenden Datensatz (schnell) zu beantworten. Es geht nicht darum, eine allgemeine Lösung zu entwickeln, die wir in die Welt verschicken können und die dort ein bestimmtes Problem löst, ohne dass der Entwickler (wir) dabei Hilfestellung geben muss. “Richtige” Software, wie ein R-Paket oder Microsoft PowerPoint, muss diese Erwartung erfüllen; “richtiges Programmieren” ist dazu vonnöten. Natürlich sind in diesem Fall die Ansprüche an die Syntax (der “Code”, hört sich cooler an) viel höher. In dem Fall muss man alle Eventualitäten voraussehen und sicherstellen, dass das Programm auch beim merkwürdigsten Nutzer brav seinen Dienst tut. Wir haben hier, beim interaktiven Gebrauch, niedrigere Ansprüche bzw. andere Ziele.

Beim interaktiven Gebrauch von R (oder beliebigen Analyseprogrammen) ist das Sortieren von Zeilen eine recht häufige Tätigkeit. Typisches Beispiel wäre der Lehrer, der eine Tabelle mit Noten hat und wissen will, welche Schüler die schlechtesten oder die besten sind in einem bestimmten Fach. Oder bei der Prüfung der Umsätze nach Filialen möchten wir die umsatzstärksten sowie -schwächsten Niederlassungen kennen.

Ein R-Befehl hierzu ist `arrange`; einige Beispiele zeigen die Funktionsweise am besten:

```
arrange(stats_test, score) # liefert die *schlechtesten* Noten zuerst zurück
arrange(stats_test, -score) # liefert die *besten* Noten zuerst zurück
arrange(stats_test, interest, score)
```

```
#>   row_number      date_time bestanden study_time self_eval interest
#> 1       234 23.01.2017 18:13:15     nein        3       1       1
#> 2       4 06.01.2017 09:58:05     nein        2       3       2
#>   score
#> 1    17
#> 2    18
#>   row_number      date_time bestanden study_time self_eval interest
#> 1       3 05.01.2017 23:33:47     ja         5      10       6
#> 2       7 06.01.2017 14:25:49     ja        NA      NA      NA
#>   score
#> 1    40
#> 2    40
#>   row_number      date_time bestanden study_time self_eval interest
#> 1       234 23.01.2017 18:13:15     nein        3       1       1
#> 2       142 19.01.2017 19:02:12     nein        3       4       1
#>   score
#> 1    17
#> 2    18
```

The diagram shows two data frames side-by-side. The left data frame has columns 'ID', 'Name', and 'Note1'. The right data frame also has columns 'ID', 'Name', and 'Note1'. An arrow points from the left frame to the right frame, with the text 'Gute Noten zuerst!' written vertically next to it. The data in the right frame is sorted by 'Note1' in ascending order.

ID	Name	Note1
1	Anna	1
2	Anna	5
3	Berta	2
4	Carla	4
5	Carla	3

ID	Name	Note1
1	Anna	1
3	Berta	2
5	Carla	3
4	Carla	4
2	Anna	5

Abbildung 3.6: Spalten sortieren

Einige Anmerkungen. Die generelle Syntax lautet `arrange(df, Spalte1, ...)`, wobei `df` den Dataframe bezeichnet und `Spalte1` die erste zu sortierende Spalte; die Punkte `...` geben an, dass man weitere Parameter übergeben kann. Man kann sowohl numerische Spalten als auch Textspalten sortieren. Am wichtigsten ist hier, dass man weitere Spalten übergeben kann. Dazu gleich mehr.

Standardmäßig sortiert `arrange` *aufsteigend* (weil kleine Zahlen im Zahlenstrahl vor den großen Zahlen kommen). Möchte man diese Reihenfolge umdrehen (große Werte zuerst, d.h. *absteigend*), so kann man ein Minuszeichen vor den Namen der Spalte setzen.

Gibt man *zwei oder mehr* Spalten an, so werden pro Wert von `Spalte1` die Werte von `Spalte2` sortiert etc; man betrachte den Output des Beispiels oben dazu. Abbildung 3.6) erläutert die Arbeitsweise von `arrange`.

Merke:

Die Funktion `arrange` sortiert die Zeilen eines Dataframes.

Ein ähnliches Ergebnis erhält mit man `top_n()`, welches die *n größten Ränge* wiedergibt:

```
top_n(stats_test, 3, interest)
#>   row_number      date_time bestanden study_time self_eval interest
#> 1      3 05.01.2017 23:33:47     ja       5      10       6
#> 2      5 06.01.2017 14:13:08     ja       4       8       6
#> 3     43 13.01.2017 14:14:16     ja       4       8       6
#> 4     65 15.01.2017 12:41:27    nein      3       6       6
#> 5    110 18.01.2017 18:53:02     ja       5       8       6
#> 6    136 19.01.2017 18:22:57     ja       3       1       6
#> 7    172 20.01.2017 20:42:46     ja       5      10       6
#> 8    214 22.01.2017 21:57:36     ja       2       6       6
```

```
#> 9      301 27.01.2017 08:17:59      ja      4      8      6
#>   score
#> 1    40
#> 2    34
#> 3    36
#> 4    22
#> 5    37
#> 6    39
#> 7    34
#> 8    31
#> 9    33
```

Gibt man *keine* Spalte an (also nur `top_n(stats_test)`), so bezieht sich `top_n` auf die letzte Spalte im Datensatz.

Wenn sich aber, wie hier, mehrere Objekte den größten Rang (Wert 6) teilen, bekommen wir *nicht* 3 Zeilen zurückgeliefert, sondern entsprechend mehr. dplyr “denkt” sich: “Ok, er will die drei besten Ränge; aber 9 Studenten teilen sich den ersten Rang (Interesse 6), wen sollte ich da ausschließen? Am besten ich liefere alle 9 zurück, sonst wäre es ja ungerecht, weil alle 9 sind ja gleich vom Interesse her”.

3.3.3.1 Aufgaben⁶



Richtig oder Falsch!?

1. `arrange` arrangiert Spalten.
2. `arrange` sortiert im Standard absteigend.
3. `arrange` lässt nur ein Sortierkriterium zu.
4. `arrange` kann numerische Werte, aber nicht Zeichenketten sortieren.
5. `top_n(5)` liefert immer fünf Werte zurück.

3.3.4 Datensatz gruppieren mit `group_by`

Einen Datensatz zu gruppieren ist eine häufige Angelegenheit: Was ist der mittlere Umsatz in Region X im Vergleich zu Region Y? Ist die Reaktionszeit in der Experimentalgruppe kleiner als in der Kontrollgruppe? Können Männer schneller ausparken als Frauen? Man sieht, dass das Gruppieren v.a. in Verbindung mit Mittelwerten oder anderen Zusammenfassungen sinnvoll ist; dazu im nächsten Abschnitt mehr.

Gruppieren meint, einen Datensatz anhand einer diskreten Variablen (z.B. Geschlecht) so aufzuteilen, dass Teil-Datensätze entstehen - pro Gruppe ein Teil-

⁶F, F, F, F, F

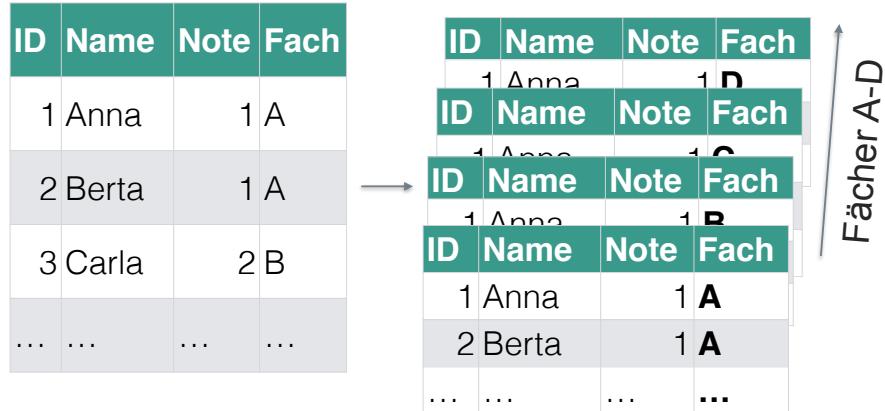


Abbildung 3.7: Datensätze nach Subgruppen aufteilen

Datensatz (z.B. ein Datensatz, in dem nur Männer enthalten sind und einer, in dem nur Frauen enthalten sind).

In Abbildung 3.7 wurde der Datensatz anhand der Spalte (d.h. Variable) Fach in mehrere Gruppen geteilt (Fach A, Fach B...). Wir könnten uns als nächstes z.B. Mittelwerte pro Fach - d.h. pro Gruppe (pro Ausprägung von Fach) - ausgeben lassen; in diesem Fall vier Gruppen (Fach A bis D).

```
test_gruppiert <- group_by(stats_test, interest)
test_gruppiert
#> # A tibble: 306 x 7
#> # Groups:   interest [7]
#>   row_number      date_time bestanden study_time self_eval interest
#>   <int>          <fctr>    <fctr>     <int>     <int>     <int>
#> 1 1 05.01.2017 13:57:01     ja       5        8        5
#> 2 2 05.01.2017 21:07:56     ja       3        7        3
#> 3 3 05.01.2017 23:33:47     ja       5       10        6
#> 4 4 06.01.2017 09:58:05    nein      2        3        2
#> 5 5 06.01.2017 14:13:08     ja       4        8        6
#> 6 6 06.01.2017 14:21:18     ja      NA      NA      NA
#> 7 7 06.01.2017 14:25:49     ja      NA      NA      NA
#> 8 8 06.01.2017 17:24:53    nein      2        5        3
#> 9 9 07.01.2017 10:11:17     ja       2        3        5
#> 10 10 07.01.2017 18:10:05    ja       4        5        5
#> # ... with 296 more rows, and 1 more variables: score <int>
```

Schaut man sich nun den Datensatz an, sieht man erstmal wenig Effekt der Gruppierung. R teilt uns lediglich mit `Groups: interest [7]`, dass es 7 Gruppen gibt, aber es gibt keine extra Spalte oder sonstige Anzeichen der Gruppierung. Aber keine Sorge, wenn wir gleich einen Mittelwert ausrechnen, bekommen wir den Mittelwert pro Gruppe!

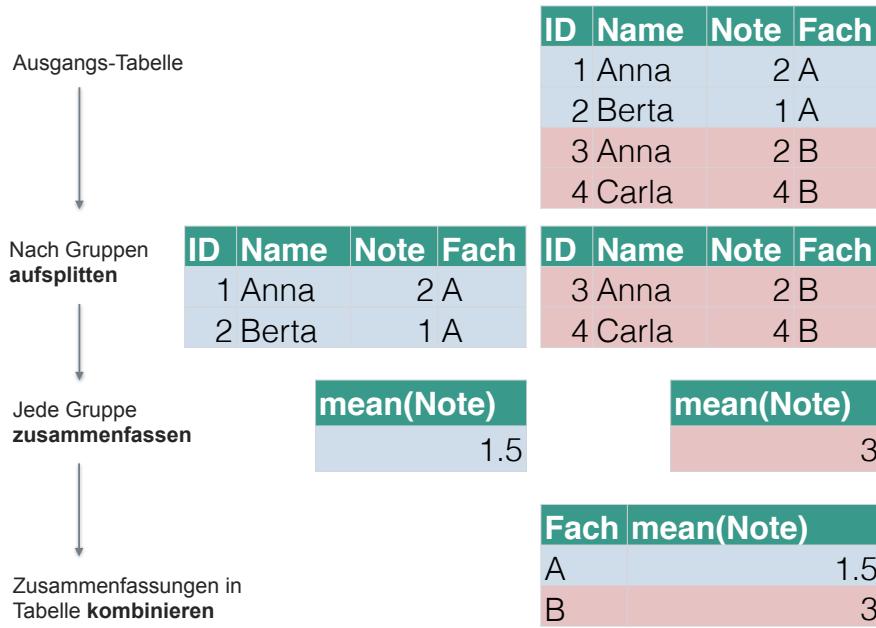


Abbildung 3.8: Schematische Darstellung des 'Gruppieren - Zusammenfassen - Kombinieren'

Ein paar Hinweise: `Source: local data frame [306 x 6]` will sagen, dass die Ausgabe sich auf einen `tibble` bezieht⁷, also eine bestimmte Art von Dataframe. `Groups: interest [7]` zeigt, dass der Tibble in 7 Gruppen - entsprechend der Werte von `interest` aufgeteilt ist.

`group_by` an sich ist nicht wirklich nützlich. Nützlich wird es erst, wenn man weitere Funktionen auf den gruppierten Datensatz anwendet - z.B. Mittelwerte ausrechnet (z.B mit `summarise`, s. unten). Die nachfolgenden Funktionen (wenn sie aus `dplyr` kommen), berücksichtigen nämlich die Gruppierung. So kann man einfach Mittelwerte pro Gruppe ausrechnen. `dplyr` kombiniert dann die Zusammenfassungen (z.B. Mittelwerte) der einzelnen Gruppen in einen Dataframe und gibt diesen dann aus.

Die Idee des "Gruppieren - Zusammenfassen - Kombinieren" ist flexibel; man kann sie häufig brauchen. Es lohnt sich, diese Idee zu lernen (vgl. Abb. 3.8).

3.3.4.1 Aufgaben⁸



Richtig oder Falsch!?

1. Mit `group_by` gruppiert man einen Datensatz.
2. `group_by` lässt nur ein Gruppierungskriterium zu.
3. Die Gruppierung durch `group_by` wird nur von Funktionen aus `dplyr` erkannt.
4. `group_by` ist sinnvoll mit `summarise` zu kombinieren.

⁷<http://stackoverflow.com/questions/29084380/what-is-the-meaning-of-the-local-data-frame-message-from>

⁸R, F, R, R

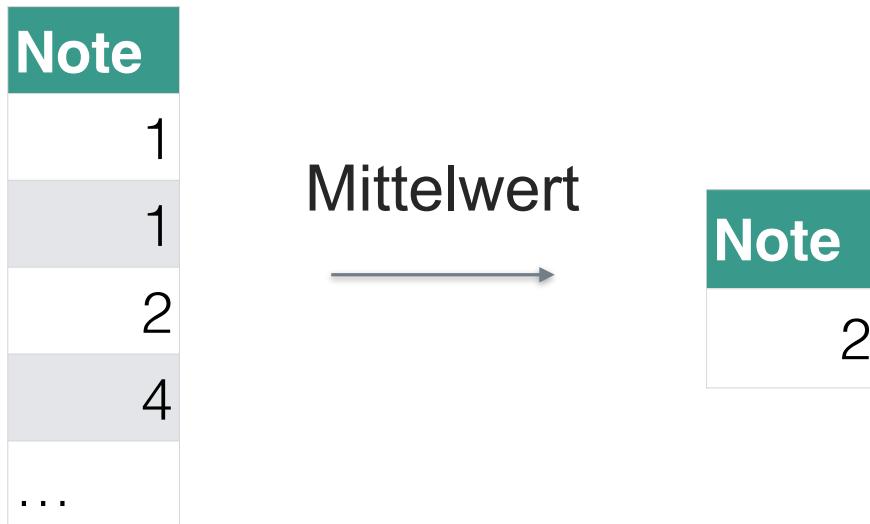


Abbildung 3.9: Spalten zu einer Zahl zusammenfassen

Merke:

Mit `group_by` teilt man einen Datensatz in Gruppen ein, entsprechend der Werte einer mehrerer Spalten.

3.3.5 Eine Spalte zusammenfassen mit `summarise`

Vielleicht die wichtigste oder häufigste Tätigkeit in der Analyse von Daten ist es, eine Spalte zu *einem* Wert zusammenzufassen; `summarise` leistet dies. Anders gesagt: Einen Mittelwert berechnen, den größten (kleinsten) Wert heraussuchen, die Korrelation berechnen oder eine beliebige andere Statistik ausgeben lassen. Die Gemeinsamkeit dieser Operationen ist, dass sie eine Spalte zu einem Wert zusammenfassen, „aus Spalte mach Zahl“, sozusagen. Daher ist der Name des Befehls `summarise` ganz passend. Genauer gesagt fasst dieser Befehl eine Spalte zu einer Zahl zusammen *anhand* einer Funktion wie `mean` oder `max` (vgl. Abb. 3.9). Hierbei ist jede Funktion erlaubt, die eine Spalte als Input verlangt und eine Zahl zurückgibt; andere Funktionen sind bei `summarise` nicht erlaubt.

```
summarise(stats_test, mean(score))
#>   mean(score)
#> 1      31.1
```

Man könnte diesen Befehl so ins Deutsche übersetzen: Fasse aus Tabelle `stats_test` die Spalte `score` anhand des Mittelwerts zusammen. Nicht vergessen, wenn die Spalte `score` fehlende Werte hat, wird der Befehl `mean` standardmäßig dies mit `NA` quittieren. Ergänzt

man den Parameter `nr.rm = TRUE`, so ignoriert R fehlende Werte und der Befehl `mean` liefert ein Ergebnis zurück.

Jetzt können wir auch die Gruppierung nutzen:

```
test_gruppiert <- group_by(stats_test, interest)
summarise(test_gruppiert, mean(score, na.rm = TRUE))
#> # A tibble: 7 x 2
#>   interest `mean(score, na.rm = TRUE)`
#>   <int>          <dbl>
#> 1     1            28.3
#> 2     2            29.7
#> 3     3            30.8
#> 4     4            29.9
#> 5     5            32.5
#> 6     6            34.0
#> 7    NA            33.1
```

Der Befehl `summarise` erkennt also, wenn eine (mit `group_by`) gruppierte Tabelle vorliegt. Jegliche Zusammenfassung, die wir anfordern, wird anhand der Gruppierungsinformation aufgeteilt werden. In dem Beispiel bekommen wir einen Mittelwert für jeden Wert von `interest`. Interessanterweise sehen wir, dass der Mittelwert tendenziell größer wird, je größer `interest` wird.

Alle diese `dplyr`-Befehle geben einen Dataframe zurück, was praktisch ist für weitere Verarbeitung. In diesem Fall heißen die Spalten `interest` und `mean(score)`. Zweiter Name ist nicht so schön, daher ändern wir den wie folgt:

Jetzt können wir auch die Gruppierung nutzen:

```
test_gruppiert <- group_by(stats_test, interest)
summarise(test_gruppiert, mw_pro_gruppe = mean(score, na.rm = TRUE))
#> # A tibble: 7 x 2
#>   interest mw_pro_gruppe
#>   <int>          <dbl>
#> 1     1            28.3
#> 2     2            29.7
#> 3     3            30.8
#> 4     4            29.9
#> 5     5            32.5
#> 6     6            34.0
#> 7    NA            33.1
```

Nun heißt die zweite Spalte `mw_pro_Gruppe`. `na.rm = TRUE` veranlasst, bei fehlenden Werten trotzdem einen Mittelwert zurückzuliefern (die Zeilen mit fehlenden Werten werden in dem Fall ignoriert).

Grundsätzlich ist die Philosophie der `dplyr`-Befehle: "Mach nur eine Sache, aber die dafür gut". Entsprechend kann `summarise` nur *Spalten* zusammenfassen, aber keine *Zeilen*.

Merke:

Mit `summarise` kann man eine Spalte eines Dataframes zu einem Wert zusammenfassen.

3.3.5.1 Aufgaben⁹



Richtig oder Falsch!?

1. Möchte man aus der Tabelle `stats_test` den Mittelwert für die Spalte `score` berechnen, so ist folgende Syntax korrekt: `summarise(stats_test, mean(score))`.
2. `summarise` liefert eine Tabelle, genauer: einen Tibble, zurück.
3. Die Tabelle, die diese Funktion zurückliefert: `summarise(stats_test, mean(score))`, hat eine Spalte mit dem Namen `mean(score)`.
4. `summarise` lässt zu, dass die zu berechnende Spalte einen Namen vom Nutzer zugewiesen bekommt.
5. `summarise` darf nur verwendet werden, wenn eine Spalte zu einem Wert zusammengefasst werden soll.

1. (Fortgeschritten) Bauen Sie einen eigenen Weg, um den mittleren Absolutabstand auszurechnen! Gehen Sie der Einfachheit halber (zuerst) von einem Vektor mit den Werten (1,2,3) aus!

Lösung:

```
x <- c(1, 2, 3)
x_mw <- mean(x)
x_delta <- x - x_mw
x_delta <- abs(x_delta)
mad <- mean(x_delta)
mad
#> [1] 0.667
```

3.3.6 Zeilen zählen mit `n` und `count`

Ebenfalls nützlich ist es, Zeilen zu zählen, also Häufigkeiten zu bestimmen. Im Gegensatz zum Standardbefehl¹⁰ `nrow` versteht der `dplyr`-Befehl `n` auch Gruppierungen. `n` darf im

⁹R, R, R, R, R

¹⁰Standardbefehl meint, dass die Funktion zum Standardrepertoire von R gehört, also nicht über ein Paket extra geladen werden muss

Pfeifen-Workflow nur im Rahmen von `summarise` oder ähnlichen `dplyr`-Befehlen verwendet werden.

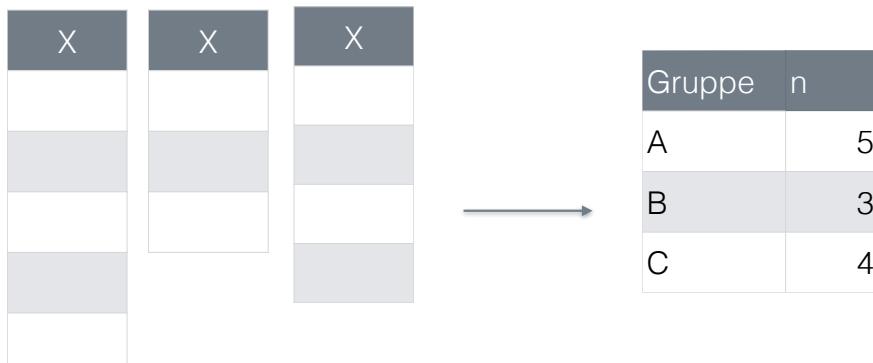
```
summarise(stats_test, n())
#>   n()
#> 1 306
summarise(test_gruppiert, n())
#> # A tibble: 7 x 2
#>   interest `n()`
#>   <int> <int>
#> 1      1    30
#> 2      2    47
#> 3      3    66
#> 4      4    41
#> 5      5    45
#> 6      6     9
#> 7     NA    68
nrow(stats_test)
#> [1] 306
```

Außerhalb von gruppierten Datensätzen ist `nrow` meist praktischer.

Praktischer ist der Befehl `count`, der nichts anderes ist als die Hintereinanderschaltung von `group_by` und `n`. Mit `count` zählen wir die Häufigkeiten nach Gruppen; Gruppen sind hier zumeist die Werte einer auszuzählenden Variablen (oder mehrerer auszuzählender Variablen). Das macht `count` zu einem wichtigen Helfer bei der Analyse von Häufigkeitsdaten.

```
dplyr::count(stats_test, interest)
#> # A tibble: 7 x 2
#>   interest     n
#>   <int> <int>
#> 1      1    30
#> 2      2    47
#> 3      3    66
#> 4      4    41
#> 5      5    45
#> 6      6     9
#> 7     NA    68
dplyr::count(stats_test, study_time)
#> # A tibble: 6 x 2
#>   study_time     n
#>   <int> <int>
#> 1          1    31
#> 2          2    49
```

Gruppe A Gruppe B Gruppe C



5 3 4

Abbildung 3.10: Sinnbild für 'count'

```
#> 3      3    85
#> 4      4    56
#> 5      5    17
#> 6     NA   68
dplyr::count(stats_test, interest, study_time)
#> # A tibble: 29 x 3
#>   interest study_time     n
#>   <int>      <int> <int>
#> 1 1          1        12
#> 2 1          2         7
#> 3 1          3         8
#> 4 1          4         2
#> 5 1          5         1
#> 6 2          1         9
#> 7 2          2        15
#> 8 2          3        16
#> 9 2          4         6
#> 10 2         5         1
#> # ... with 19 more rows
```

Allgemeiner formuliert lautet die Syntax: `count(df, Spalte1, ...)`, wobei `df` der Dataframe ist und `Spalte1` die erste (es können mehrere sein) auszuzählende Spalte. Gibt man z.B. zwei Spalten an, so wird pro Wert der 1. Spalte die Häufigkeiten der 2. Spalte ausgegeben (vgl. Abb. 3.10).

Merke:

`n` und `count` zählen die Anzahl der Zeilen, d.h. die Anzahl der Fälle.

3.3.6.1 Vertiefung zum Zählen von Zeilen: Relative Häufigkeiten

Manchmal ist es praktisch, nicht nur die (absolute) Häufigkeiten von Zeilen zu zählen, sondern ihren Anteil nach (relative Häufigkeit). Klassisches Beispiel: Wie viel Prozent der Fälle sind Frauen, wie viele sind Männer?

In `dplyr` kann man das so umsetzen:

```
stats_test %>%
  count(interest) %>%
  mutate(prop_interest = n / sum(n))
#> # A tibble: 7 x 3
#>   interest     n prop_interest
#>   <int> <int>        <dbl>
#> 1       1     30      0.0980
#> 2       2     47      0.1536
#> 3       3     66      0.2157
#> 4       4     41      0.1340
#> 5       5     45      0.1471
#> 6       6      9      0.0294
#> 7      NA     68      0.2222
```

`prop` steht hier für “Proportion”, also Anteil. `sum(n)` liefert die Summe der Fälle zurück, also 306 in diesem Fall.

Etwas komplexer ist es, wenn man zwei Gruppierungsvariablen hat und dann Anteile auszählen möchte:

```
stats_test$bestanden <- stats_test$score > 25

stats_test %>%
  group_by(interest, bestanden) %>%
  summarise(n = n()) %>%
  mutate(prop_interest = n / sum(n))
#> # A tibble: 14 x 4
#> # Groups:   interest [7]
#>   interest bestanden     n prop_interest
#>   <int>     <lgl> <int>        <dbl>
#> 1       1     FALSE    10      0.333
#> 2       1      TRUE    20      0.667
#> 3       2     FALSE     9      0.191
```

```
#> 4      2      TRUE   38      0.809
#> 5      3      FALSE  14      0.212
#> 6      3      TRUE   52      0.788
#> 7      4      FALSE  9       0.220
#> 8      4      TRUE   32      0.780
#> 9      5      FALSE  6       0.133
#> 10     5      TRUE   39      0.867
#> 11     6      FALSE  1       0.111
#> 12     6      TRUE   8       0.889
#> 13     NA     FALSE  7       0.103
#> 14     NA     TRUE   61      0.897
```

Synonym zur letzten Syntax könnte man auch schreiben:

```
stats_test %>%
  count(interest, bestanden) %>%
  mutate(prop_interest = n / sum(n))
```

3.3.6.2 Aufgaben¹¹



Richtig oder Falsch!?

1. Mit `count` kann man Zeilen zählen.
2. `count` ist ähnlich (oder identisch) zu einer Kombination von `group_by` und `n()`.
3. Mit `count` kann man nur eine Gruppe beim Zählen berücksichtigen.
4. `count` darf nicht bei nominalskalierten Variablen verwendet werden.

1. Bauen Sie sich einen Weg, um den Modus mithilfe von `count` und `arrange` zu bekommen!

```
stats_count <- count(stats_test, score)
stats_count_sortiert <- arrange(stats_count, -n)
head(stats_count_sortiert, 1)
#> # A tibble: 1 x 2
#>   score     n
#>   <int> <int>
#> 1    34     22
```

Ah! Der Score 34 ist der häufigste!

¹¹R, R, F, F



Abbildung 3.11: Das ist keine Pfeife

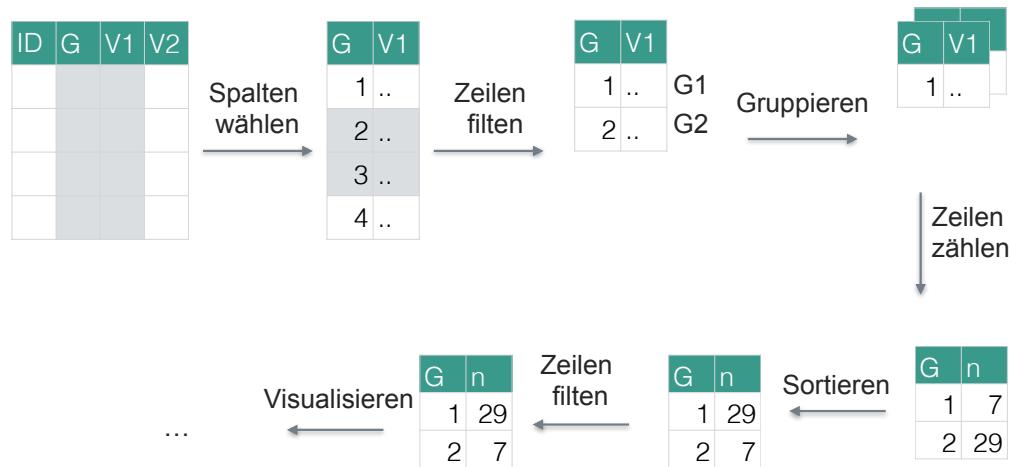


Abbildung 3.12: Das 'Durchpfeifen'

3.4 Die Pfeife

Die zweite zentrale Idee von dplyr kann man salopp als “Durchpfeifen” oder die “Idee der Pfeife” (Durchpfeifen) bezeichnen; ikonographisch mit einem Pfeifen ähnlichen Symbol dargestellt `%>%`. Der Begriff “Durchpfeifen” ist frei vom Englischen “to pipe” übernommen. Das berühmte Bild von René Magritte stand dabei Pate (s. Abb. 3.11; (M7 2004)).

Hierbei ist gemeint, einen Datensatz sozusagen auf ein Fließband zu legen und an jedem Arbeitsplatz einen Arbeitsschritt auszuführen. Der springende Punkt ist, dass ein Dataframe als “Rohstoff” eingegeben wird und jeder Arbeitsschritt seinerseits wieder einen Dataframe ausgibt. Damit kann man sehr schön, einen “Flow” an Verarbeitung erreichen, außerdem spart man sich Tipparbeit und die Syntax wird lesbarer. Damit das Durchpfeifen funktioniert, benötigt man Befehle, die als Eingabe einen Dataframe erwarten und wieder einen Dataframe zurückliefern. Das Schaubild verdeutlicht beispielhaft eine Abfolge des Durchpfeifens (s. Abb. 3.12).

Die sog. “Pfeife” (pipe: `%>%`) in Anspielung an das berühmte Bild von René Magritte, verkettet Befehle hintereinander. Das ist praktisch, da es die Syntax vereinfacht.



Tipp: In RStudio gibt es einen Shortcut für die Pfeife: Strg-Shift-M (auf allen Betriebssystemen).

Vergleichen Sie mal diese Syntax

```
filter(summarise(group_by(filter(stats_test,
    !is.na(score)), interest), mw = mean(score)), mw > 30)
```

mit dieser

```
stats_test %>%
  filter(!is.na(score)) %>%
  group_by(interest) %>%
  summarise(mw = mean(score)) %>%
  filter(mw > 30)
#> # A tibble: 4 x 2
#>   interest     mw
#>   <int> <dbl>
#> 1      3 30.8
#> 2      5 32.5
#> 3      6 34.0
#> 4     NA 33.1
```

Die zweite ist viel einfacher! Lassen Sie uns die “Pfeifen-Syntax” in deutschen Pseudo-Code zu übersetzen.



Nimm die Tabelle “stats_test” UND DANN
 filtere alle nicht-fehlenden Werte UND DANN
 gruppiere die verbleibenden Werte nach “interest” UND DANN
 bilde den Mittelwert (pro Gruppe) für “score” UND DANN
 liefere nur die Werte größer als 30 zurück.

Die zweite Syntax, in “Pfeifenform” ist viel einfacher zu verstehen als die erste! Die erste Syntax ist verschachtelt, man muss sie von innen nach außen lesen. Das ist kompliziert. Die Pfeife in der 2. Syntax macht es viel einfacher, die Syntax zu verstehen, da die Befehle “hintereinander” gestellt (sequenziell organisiert) sind.

Die Pfeife zerlegt die “russische Puppe”, also ineinander verschachtelten Code, in sequenzielle Schritte und zwar in der richtigen Reihenfolge (entsprechend der Abarbeitung). Wir müssen den Code nicht mehr von innen nach außen lesen (wie das bei einer mathematischen Formel der Fall ist), sondern können wie bei einem Kochrezept “erstens …, zweitens .., drittens …” lesen. Die Pfeife macht die Syntax einfacher. Natürlich hätten wir die verschachtelte Syntax in viele einzelne Befehle zerlegen können und jeweils eine Zwischenergebnis speichern

mit dem Zuweisungsoperator `<-` und das Zwischenergebnis dann explizit an den nächsten Befehl weitergeben. Eigentlich macht die Pfeife genau das - nur mit weniger Tipparbeit. Und auch einfacher zu lesen. Flow!



Wenn Sie Befehle verketten mit der Pfeife, sind nur Befehle erlaubt, die einen Datensatz als Eingabe verlangen und einen Datensatz ausgeben. Das ist bei den hier vorgestellten Funktionen der Fall. Viele andere Funktionen erfüllen dieses Kriterium aber nicht; in dem Fall liefert `dplyr` eine Fehlermeldung.

3.4.1 Spalten berechnen mit `mutate`

Wenn man die Pfeife benutzt, ist der Befehl `mutate` ganz praktisch: Er berechnet eine Spalte. Normalerweise kann man einfach eine Spalte berechnen mit dem Zuweisungsoperator:

Zum Beispiel so:

```
df$neue_spalte <- df$spalte1 + df$spalte2
```

Innerhalb einer Pfeifen-Syntax geht das aber nicht (so gut). Da ist man mit der Funktion `mutate` besser beraten; `mutate` leistet just dasselbe wie die Pseudo-Syntax oben:

```
df %>%
  mutate(neue_spalte = spalte1 + spalte2)
```

In Worten:



Nimm die Tabelle “df” UND DANN
bilde eine neue Spalte mit dem Namen `neue_spalte`, die sich berechnet als Summe
von `spalte1` und `spalte2`.

Allerdings berücksichtigt `mutate` auch Gruppierungen, das ist praktisch. Der Hauptvorteil ist die bessere Lesbarkeit durch Auflösen der Verschachtelungen.

Ein konkretes Beispiel:

```
stats_test %>%
  select(bestanden, interest, score) %>%
  mutate(Streber = score > 38) %>%
  head()
#>   bestanden interest score Streber
#> 1     TRUE      5    29 FALSE
#> 2     TRUE      3    29 FALSE
#> 3     TRUE      6    40  TRUE
#> 4    FALSE      2    18 FALSE
```

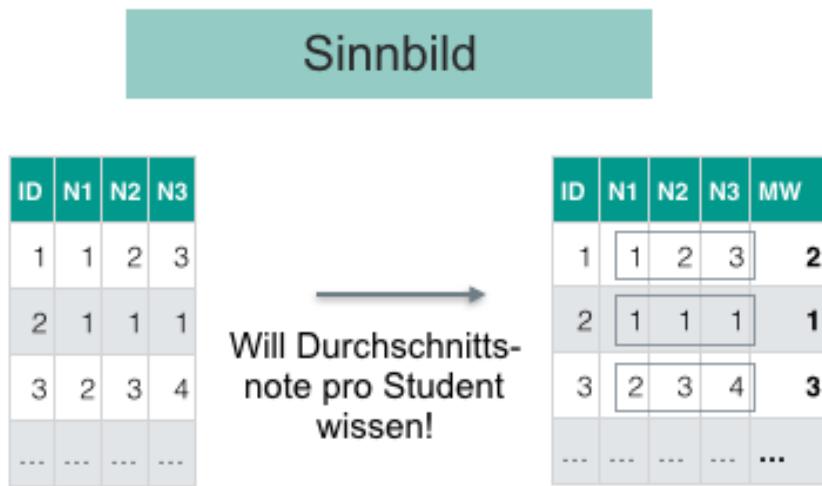


Abbildung 3.13: Sinnbild für mutate

```
#> 5      TRUE      6      34    FALSE
#> 6      TRUE      NA     39    TRUE
```

Diese Syntax erzeugt eine neue Spalte innerhalb von `stats_test`; diese Spalte prüft pro Person, ob `score > 38` ist. Falls ja (TRUE), dann ist `Streber` TRUE, ansonsten ist `Streber` FALSE (tja). `head` zeigt die ersten 6 Zeilen des resultierenden Dataframes an.

Abb. 3.13 zeigt Sinnbild für `mutate`:



`mutate` erwartet als Input *keinen* Dateframe, sondern eine Spalte. Betrachten Sie das Sinnbild von `mutate`. Die Idee ist, eine Spalte umzuwandeln nach dem Motto: "Nimm eine Spalte, mach was damit und liefere die neue Spalte zurück". Die Spalte (und damit jeder einzelne Wert in der Spalte) wird *verändert* ('mutiert', daher 'mutate'). Man kann auch sagen, die Spalte wird *transformiert*.

3.4.2 Aufgaben

1. Entschlüsseln Sie dieses Ungetüm! Übersetzen Sie diese Syntax auf Deutsch:

```
bestanden_gruppen <-
  filter(
    summarise(
      group_by(filter(select(stats_test, -c(row_number, date_time)) , bestanden == "ja"),
      Punkte = mean(score), n = n()))
```

2. Entschlüsseln Sie jetzt diese Syntax bzw. übersetzen Sie sie ins Deutsche:

```
stats_test %>%
  select(-row_number, -date_time) %>%
  filter(bestanden == "ja") %>%
  group_by(interest) %>%
  summarise(Punkte = mean(score),
            n = n())
```

3. Die Pfeife bei im Klausur-Datensatz

- Übersetzen Sie die folgende Pseudo-Syntax ins ERRRische!



Nimm den Datensatz **stats_test** UND DANN...
 Wähle daraus die Spalte **score** UND DANN...
 Berechne den Mittelwert der Spalte UND DANN...
 ziehe vom Mittelwert die Spalte ab UND DANN... quadriere die einzelnen Differenzen
 UND DANN... bilde davon den Mittelwert.

Lösung:

```
stats_test %>%
  select(score) %>%
  mutate(score_delta = score - mean(.score)) %>%
  mutate(score_delta_squared = score_delta^2) %>%
  summarise(score_var = mean(score_delta_squared)) %>%
  summarise(sqrt(score_var))
```

Was sagt uns der Punkt . in der Syntax oben? Der Punkt steht für die Tabelle, wie sie gerade aufbereitet ist (also laut letzter Zeile in der Syntax). Warum müssen wir dem Befehl **mean** sagen, welche Spalte/Variable **score** wir meinen? Ist doch logo, wir meinen natürlich die Spalte **score** im aktuellen, durchgepfiffenen Datensatz! Leider weiß das der Befehl **mean** nicht. **mean** hat keinerlei Idee von Pfeifen, unseren Wünschen und Sorgen. **mean** denkt sich: “Not my job! Sag mir gefälligst *wie immer*, in welchem Dataframe ich die Spalte finde!”. Also sagen wir **mean**, wo er die Spalte findet...

- Berechnen Sie die sd von **score** in **stats_test**! Vergleichen Sie sie mit dem Ergebnis der vorherigen Aufgabe!¹²
- Was hat die Pfeifen-Syntax oben berechnet?¹³

¹²sd(stats_test\$score)

¹³die sd von **score**

3.5 Deskriptive Statistik

`dplyr` kann man gut gebrauchen, um deskriptive Statistik zu berechnen. `summarise` charakterisiert eine Hauptidee der Deskriptivstatistik: Einen Vektor zu einer Zahl zusammenzufassen. `group_by` steht für die Idee, ‘Zahlensäcke’ (Verteilungen) in Subgruppen aufzuteilen. `mutate` transformiert Daten. `n` zählt Häufigkeiten.

Ein weiterer zentraler Gedanken der Deskriptivstatistik ist es, dass es beim Zusammenfassen von Daten nicht reicht, sich auf den Mittelwert oder eine (hoffentlich) ‘repräsentative’ Zahl zu verlassen. Man braucht auch einen Hinweis, wie unterschiedlich die Daten sind. Entsprechend spricht man von zwei Hauptbereichen der deskriptiven Statistik.

Die deskriptive Statistik hat zwei Hauptbereiche: Lagemaße und Streuungsmaße.

Lagemaße geben den “typischen”, “mittleren” oder “repräsentativen” Vertreter der Verteilung an. Bei den Lagemaßen denkt man sofort an das *arithmetische Mittel* (synonym: Mittelwert, arithmetisches Mittel; häufig als \bar{X} abgekürzt; `mean`). Ein Nachteil von Mittelwerten ist, dass sie *nicht robust* gegenüber Extremwerte sind: Schon ein vergleichsweise großer Einzelwert kann den Mittelwert stark verändern und damit die Repräsentativität des Mittelwerts für die Gesamtmenge der Daten in Frage stellen. Eine robuste Variante ist der *Median* (`Md`; `median`). Ist die Anzahl der (unterschiedlichen) Ausprägungen nicht zu groß im Verhältnis zur Fallzahl, so ist der *Modus* eine sinnvolle Statistik; er gibt die häufigste Ausprägung an¹⁴.

Streuungsmaße geben die Unterschiedlichkeit in den Daten wieder; mit anderen Worten: sind die Daten sich ähnlich oder unterscheiden sich die Werte deutlich? Zentrale Statistiken sind der *mittlere Absolutabstand* (`MAA`; engl. mean absolute deviation, `MAD`),¹⁵ die *Standardabweichung* (`sd`; `sd`), die *Varianz* (`Var`; `var`) und der *Interquartilsabstand* (`IQR`; `IQR`). Da nur der `IQR` *nicht* auf dem Mittelwert basiert, ist er robuster als Statistiken, die sich aus dem Mittelwert ergeben. Beliebige Quantile bekommt man mit dem R-Befehl `quantile`. Möchte man z.B. `Q1`, Median und `Q3`, so kann man das so sagen: `quantile(x, probs = c(.25, .50, .75))`, wobei `x` eine Spalte (ein Vektor) ist.

Der Befehl `summarise` eignet sich, um deskriptive Statistiken auszurechnen.

```
summarise(stats_test, mean(score))
#>   mean(score)
#> 1      31.1
summarise(stats_test, sd(score))
#>   sd(score)
#> 1      5.74
```

¹⁴Der *Modus* ist im Standard-R nicht mit einem eigenen Befehl vertreten. Man kann ihn aber leicht von Hand bestimmen; s.u. Es gibt auch einige Pakete, die diese Funktion anbieten: z.B. <https://cran.r-project.org/web/packages/modes/index.html>

¹⁵Der *MAD* ist im Standard-R nicht mit einem eigenen Befehl vertreten. Es gibt einige Pakete, die diese Funktion anbieten: z.B. `lsr::aad` (absolute average deviation from the mean) <https://artax.karlin.mff.cuni.cz/r-help/library/lsr/html/aad.html>

Tabelle 3.1: Befehle des Kapitels 'Datenjudo'

Paket::Funktion	Beschreibung
dplyr::arrange	Sortiert Spalten
dplyr::filter	Filtert Zeilen
dplyr::select	Wählt Spalten
dplyr::group_by	gruppert einen Dataframe
dplyr::n	zählt Zeilen
dplyr::count	zählt Zeilen nach Untergruppen
%>% (dplyr)	verkettet Befehle
dplyr::mutate	erzeugt/berechnet Spalten

```
summarise(stats_test, aad(score)) # aus Paket 'lsr'
#> aad(score)
#> 1      4.84
```

Natürlich könnte man auch einfacher schreiben:

```
mean(stats_test$score)
#> [1] 31.1
median(stats_test$score)
#> [1] 31
aad(stats_test$score)
#> [1] 4.84
```



Viele R-Befehle der deskriptiven Statistik sind im Standard so eingestellt, dass sie `NA` zurückliefern, falls es in den Daten fehlende Werte gibt. Das ist einerseits informativ, aber oft unnötig. Mit dem Parameter `na.rm = TRUE` kann man dieses Verhalten abstellen.

Tipp: Mit dem Befehl `df <- na.omit(df)` entfernen Sie alle fehlenden Werte aus `df`.

`summarise` liefert aber im Unterschied zu `mean` etc. immer einen Dataframe zurück. Da der Dataframe die typische Datenstruktur ist, ist es häufig praktisch, wenn man einen Dataframe zurückbekommt, mit dem man weiterarbeiten kann. Außerdem lassen `mean` etc. keine Gruppierungsoperationen zu; über `group_by` kann man dies aber bei `dplyr` erreichen.

3.6 Befehlsübersicht

Tabelle 3.1 fasst die R-Funktionen dieses Kapitels zusammen.

3.7 Verweise

- Die offizielle Dokumentation von `dplyr` findet sich hier: [`https://cran.r-project.org/web/packages/dplyr/dplyr.pdf`](https://cran.r-project.org/web/packages/dplyr/dplyr.pdf).
- Eine schöne Demonstration wie mächtig `dplyr` ist findet sich hier: [`http://bit.ly/2kX9lvC`](http://bit.ly/2kX9lvC).
- Die GUI “exploratory” ist ein “klickbare” Umsetzung von `dplyr` and friends; mächtig, modern und sieht cool aus: [`https://exploratory.io`](https://exploratory.io).
- *R for Data Science* bietet umfangreiche Unterstützung zu diesem Thema (Wickham und Grolemund 2016).

Kapitel 4

Praxisprobleme der Datenaufbereitung



Lernziele:

- Typische Probleme der Datenaufbereitung kennen.
- Typische Probleme der Datenaufbereitung bearbeiten können.

Laden wir zuerst die benötigten Pakete; v.a. ist das `dplyr` and friends. Das geht mit dem Paket `tidyverse`.

```
library(tidyverse)
library(corr)
library(gridExtra)
library(car)
library(prada) # optional: Daten 'extra' und 'stats_test'
```

Stellen wir einige typische Probleme des Datenjudo (genauer: der Datenaufbereitung) zusammen. Probleme heißt hier nicht, dass es etwas Schlimmes passiert ist, sondern es ist gemeint, wir schauen uns ein paar typische Aufgabenstellungen an, die im Rahmen der Datenaufbereitung häufig anfallen.

4.1 Datenaufbereitung

4.1.1 Auf fehlende Werte prüfen

Das geht recht einfach mit `summary(mein_dataframe)`. Der Befehl liefert für jede Spalte des Dataframe `mein_dataframe` die Anzahl der fehlenden Werte zurück.

```
stats_test <- read.csv("data/stats_test.csv")
summary(stats_test)

#>   row_number           date_time  bestanden   study_time
#> Min.   : 1.0  05.01.2017 13:57:01: 1 ja    :261  Min.   :1.0
#> 1st Qu.: 77.2 05.01.2017 21:07:56: 1 nein: 45  1st Qu.:2.0
#> Median :153.5 05.01.2017 23:33:47: 1                   Median :3.0
#> Mean   :153.5  06.01.2017 09:58:05: 1                   Mean   :2.9
#> 3rd Qu.:229.8  06.01.2017 14:13:08: 1                   3rd Qu.:4.0
#> Max.   :306.0  06.01.2017 14:21:18: 1                   Max.   :5.0
#>          (Other)            :300  NA's    :68
#>
#>   self_eval      interest     score
#> Min.   : 1.0  Min.   :1.0  Min.   :17.0
#> 1st Qu.: 4.0  1st Qu.:2.0  1st Qu.:27.0
#> Median : 5.0  Median :3.0  Median :31.0
#> Mean   : 5.4  Mean   :3.2  Mean   :31.1
#> 3rd Qu.: 7.0  3rd Qu.:4.0  3rd Qu.:36.0
#> Max.   :10.0  Max.   :6.0  Max.   :40.0
#> NA's    :68    NA's    :68
```

4.1.2 Fälle mit fehlenden Werte löschen

Weist eine Variable (Spalte) “wenig” fehlende Werte auf, so kann es schlau sein, nichts zu tun. Eine andere Möglichkeit besteht darin, alle entsprechenden Zeilen zu löschen. Man sollte aber schauen, wie viele Zeilen dadurch verloren gehen.

```
# Ursprünglich Anzahl an Fällen (Zeilen)
nrow(stats_test)
#> [1] 306

# Nach Umwandlung in neuen Dataframe
stats_test %>%
  na.omit -> stats_test_na OMIT
nrow(stats_test_na OMIT)
#> [1] 238

# Nur die Anzahl der bereinigten Daten
stats_test %>%
  na.omit %>%
  nrow
#> [1] 238
```



Bei mit der Pfeife verketteten Befehlen darf man für Funktionen die runden Klammern weglassen, wenn man keinen Parameter schreibt. Also ist `nrow` (ohne Klammern) erlaubt bei `dplyr`, wo es eigentlich `nrow()` heißen müsste. Sie dürfen die Klammern natürlich schreiben, aber sie müssen nicht.

Hier verlieren wir 68 Zeilen, das verschmerzen wir.

Welche Zeilen verlieren wir eigentlich? Lassen wir uns nur die *nicht*-kompletten Fälle anzeigen (und davon nur die ersten paar):

```
stats_test %>%
  filter(!complete.cases(.)) %>%
  head
#>   row_number      date_time bestanden study_time self_eval interest
#> 1       6 06.01.2017 14:21:18      ja      NA      NA      NA
#> 2       7 06.01.2017 14:25:49      ja      NA      NA      NA
#> 3      15 09.01.2017 15:23:15      ja      NA      NA      NA
#> 4      19 10.01.2017 17:16:48     nein      NA      NA      NA
#> 5      42 13.01.2017 14:08:08      ja      NA      NA      NA
#> 6      49 14.01.2017 07:02:39      ja      NA      NA      NA
#>   score
#> 1   39
#> 2   40
#> 3   30
#> 4   22
#> 5   38
#> 6   39
```



Man beachte, dass der Punkt `.` für den Datensatz steht, wie er vom letzten Schritt weitergegeben wurde. Innerhalb einer `dplyr`-Befehls-Kette können wir den Datensatz, wie er im letzten Schritt beschaffen war, stets mit `.` ansprechen; ganz praktisch, weil schnell zu tippen. Natürlich könnten wir diesen Datensatz jetzt als neues Objekt speichern und damit weiter arbeiten. Das Ausrufezeichen `!` steht für logisches “Nicht”. Mit `head` bekommt man nur die ersten paar Fälle (6 im Standard) angezeigt, was oft reicht für einen Überblick.

In Pseudo-Syntax liest es sich so:



Nehme den Datensatz `stats_test` UND DANN...
filtere die nicht-kompletten Fälle

4.1.3 Fehlende Werte zählen

Wie viele fehlende Wert weist eine Spalte auf?

```
stats_test$self_eval %>% is.na %>% sum
#> [1] 68
```



Nimm die Spalte `self_eval` aus der Tabelle “`stats_test`” UND DANN finde die fehlenden Werte (NAs) in dieser Spalten UND DANN summiere alle Treffer auf.

4.1.4 Fehlende Werte ggf. ersetzen

Ist die Anzahl der fehlenden Werte zu groß, als dass wir es verkraften könnten, die Zeilen zu löschen, so können wir die fehlenden Werte ersetzen. Allein, das ist ein weites Feld und übersteigt den Anspruch dieses Kurses¹. Eine einfache, aber nicht die beste Möglichkeit, besteht darin, die fehlenden Werte durch einen repräsentativen Wert, z.B. den Mittelwert der Spalte, zu ersetzen.

```
stats_test %>%
  mutate(interest = replace(. $interest,
                            is.na(stats_test$interest),
                            mean(stats_test$interest,
                                  na.rm = TRUE))) -> stats_test

sum(is.na(stats_test$interest))
#> [1] 0
```

`replace`² ersetzt Werte aus dem Vektor `stats_test$interest` alle Werte, für die `is.na(stats_test$interest)` wahr ist, bei Zeilen mit fehlenden Werten in dieser Spalte also. Diese Werte werden durch den Mittelwert der Spalte ersetzt³. Der Punkt `.` ersetzt den Daten der Tabelle (wir hätten aber auch den Namen der Tabelle ausschreiben können).

4.1.5 Nach Fehlern suchen

Leicht schleichen sich Tippfehler oder andere Fehler ein. Man sollte darauf prüfen; so könnte man sich ein Histogramm ausgeben lassen pro Variable, um “ungewöhnliche” Werte gut zu

¹Das sagen Autoren, wenn sie nicht genau wissen, wie etwas funktioniert.

²aus dem “Standard-R”, d.h. Paket “base”.

³Hier findet sich eine ausführlichere Darstellung: https://sebastiansauer.github.io/checklist_data_cleansing/index.html

erkennen. Meist geht das besser als durch das reine Betrachten von Zahlen. Gibt es wenig unterschiedliche Werte, so kann man sich auch die unterschiedlichen Werte ausgeben lassen.

```
stats_test %>%
  count(interest, sort = TRUE) %>% head
#> # A tibble: 6 x 2
#>   interest     n
#>   <dbl> <int>
#> 1     3.21    68
#> 2     3.00    66
#> 3     2.00    47
#> 4     5.00    45
#> 5     4.00    41
#> 6     1.00    30
```

Da in der Umfrage nur ganze Zahlen von 1 bis 5 abgefragt wurden, ist die `3.21...` auf den ersten Blick suspekt. In diesem Fall ist aber alles ok, da wir diesen Wert selber erzeugt haben.

Findet man ‘merkwürdige’ (unplausible) Werte, so kann es sinnvoll sein, diese Werte herauszunehmen (im Detail eine schwierige Entscheidung). Besser als die Zeilen zu löschen, ist es oft, diese Werte in `NA` umzuwandeln. Sagen wir, wir möchten alle Fälle mit `score < 21` entfernen bzw. in `NA` umwandeln.

```
stats_test %>%
  mutate(score_bereinigt = replace(.$score,
                                    .$score < 21,
                                    NA)) -> stats_test
```

4.1.6 Ausreißer identifizieren

Ähnlich zu Fehlern, steht man Ausreißer häufig skeptisch gegenüber. Allerdings kann man nicht pauschal sagen, dass Extremwerte entfernt werden sollen: Vielleicht war jemand in der Stichprobe wirklich nur 1.20m groß? Hier gilt es, begründet und nachvollziehbar im Einzelfall zu entscheiden. Histogramme und Boxplots sind wieder ein geeignetes Mittel, um Ausreißer zu finden (vgl. Abb. 4.1).

```
qplot(x = score, data = stats_test, binwidth = 1)
```

Mit `binwidth = 1` sagen wir, dass jeder Balken (bin) eine Breite (width) von 1 haben soll.

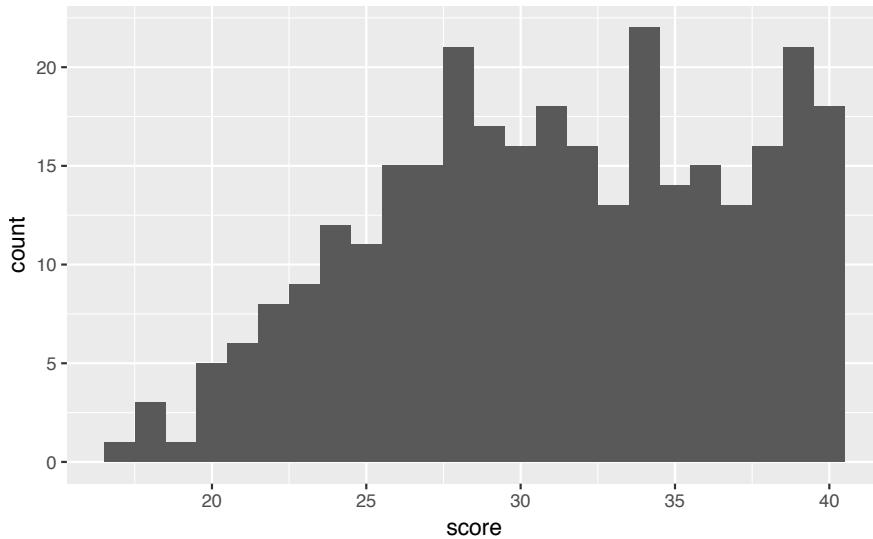


Abbildung 4.1: Ausreißer identifizieren

4.1.7 Hochkorrelierte Variablen finden

Haben zwei Leute die gleiche Meinung, so ist einer von beiden überflüssig - wird behauptet. Ähnlich bei Variablen; sind zwei Variablen sehr hoch korreliert ($>.9$, als grober (!) Richtwert), so bringt die zweite kaum Informationszuwachs zur ersten. Und kann z.B. ausgeschlossen werden.

Nehmen wir dazu den Datensatz `extra` her.

```
extra <- read.csv("data/extradata.csv")
```

```
extra %>%
  select(i01:i10) %>% # Wähle die Variablen von i1 bis i10 aus
  correlate() -> km    # Korrelationsmatrix berechnen
km
#> # A tibble: 10 x 11
#>   rowname   i01    i02r    i03    i04    i05    i06r    i07    i08    i09    i10
#>   <chr>     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
#> 1   i01      NA 0.4676 0.1167 0.433 0.4437 0.450 0.298 0.349 0.363 0.1866
#> 2   i02r     0.468     NA 0.1152 0.349 0.3951 0.516 0.235 0.286 0.245 0.0607
#> 3   i03      0.117 0.1152     NA 0.057 0.0679 0.154 0.122 0.109 0.038 0.0922
#> 4   i04      0.433 0.3495 0.0570     NA 0.6604 0.286 0.440 0.193 0.233 0.3525
#> 5   i05      0.444 0.3951 0.0679 0.660     NA 0.321 0.402 0.270 0.298 0.2988
#> 6   i06r     0.450 0.5159 0.1542 0.286 0.3207     NA 0.150 0.267 0.264 0.1317
#> 7   i07      0.298 0.2345 0.1223 0.440 0.4016 0.150     NA 0.303 0.209 0.3444
#> 8   i08      0.349 0.2863 0.1094 0.193 0.2703 0.267 0.303     NA 0.360 0.1807
#> 9   i09      0.363 0.2447 0.0380 0.233 0.2982 0.264 0.209 0.360     NA 0.1423
```

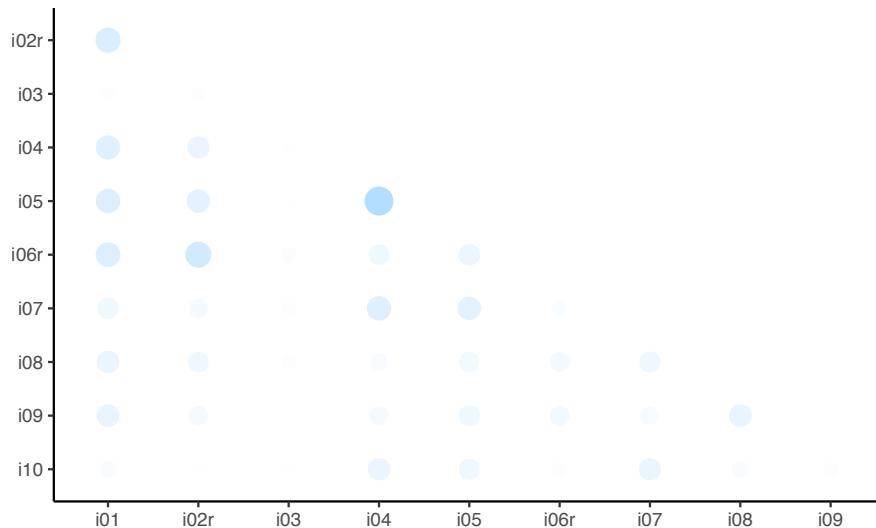


Abbildung 4.2: Ein Korrelationsplot

```
#> 10      i10 0.187 0.0607 0.0922 0.352 0.2988 0.132 0.344 0.181 0.142      NA
```

In diesem Beispiel sind keine Variablen sehr hoch korreliert. Wir leiten keine weiteren Schritte ein, abgesehen von einer Visualisierung.

```
km %>%
  shave() %>% # Oberes Dreieck ist redundant, wird "abgespiert"
  rplot() # Korrelationsplot
```

Die Funktion `correlate` stammt aus dem Paket `corr4`⁴, welches vorher installiert und geladen sein muss. Hier ist die Korrelation nicht zu groß, so dass wir keine weiteren Schritte unternehmen. Hätten wir eine sehr hohe Korrelation gefunden, so hätten wir eine der beiden beteiligten Variablen aus dem Datensatz löschen können.

4.1.8 z-Standardisieren

Für eine Reihe von Analysen ist es wichtig, die Skalierung der Variablen zur vereinheitlichen. Die z-Standardisierung ist ein übliches Vorgehen. Dabei wird der Mittelwert auf 0 transformiert und die SD auf 1; man spricht - im Falle von (hinreichend) normalverteilten Variablen - jetzt von der *Standardnormalverteilung*. Unterscheiden sich zwei Objekte A und B in einer standardnormalverteilten Variablen, so sagt dies nur etwas zur relativen Position von A zu B innerhalb ihrer Verteilung aus - im Gegensatz zu den Rohwerten.

⁴<https://github.com/drsimonj/corr4>

```
extra %>%
  select(i01, i02r) %>%
  scale() %>% # z-standardisieren
  head() # nur die ersten paar Zeilen abdrucken
#>      i01    i02r
#> [1,] -0.519 -0.134
#> [2,] -2.002 -1.383
#> [3,] -0.519  1.115
#> [4,] -0.519 -0.134
#> [5,]  0.964 -0.134
#> [6,] -0.519 -1.383
```

Dieser Befehl liefert z-standardisierte Spalten zurück. Kommoder ist es aber, alle Spalten des Datensatzes zurück zu bekommen, wobei zusätzlich die z-Werte aller numerischen Variablen hinzugekommen sind:

```
extra %>%
  mutate_if(is.numeric, funs("z" = scale)) %>%
  head
```

Der Befehl `mutate` berechnet eine neue Spalte; `mutate_if` tut dies nur, wenn die Spalte numerisch ist. Die neue Spalte wird berechnet als z-Transformierung der alten Spalte; zum Spaltenname wird ein “`_z`” hinzugefügt. Natürlich hätten wir auch mit `select` “händisch” die relevanten Spalten auswählen können.

4.1.9 Quasi-Konstante finden

Hier suchen wir nach Variablen (Spalten), die nur einen Wert oder zumindest nur sehr wenige verschiedene Werte aufweisen. Oder, ähnlich: Wenn 99.9% der Fälle nur von einem Wert bestritten wird. In diesen Fällen kann man die Variable als “Quasi-Konstante” bezeichnen. Quasi-Konstanten sind für die Modellierung von keiner oder nur geringer Bedeutung; sie können in der Regel für weitere Analysen ausgeschlossen werden.

Haben wir z.B. nur Männer im Datensatz, so kann das Geschlecht nicht für Unterschiede im Einkommen verantwortlich sein. Besser ist es, die Variable Geschlecht zu entfernen. Auch hier sind Histogramme oder Boxplots von Nutzen zur Identifikation von (Quasi-)Konstanten. Alternativ kann man sich auch pro die Streuung (numerische Variablen) oder die Anzahl unterschiedlicher Werte (qualitative Variablen) ausgeben lassen:

```
IQR(extra$n_facebook_friends, na.rm = TRUE) # keine Konstante
#> [1] 300
```

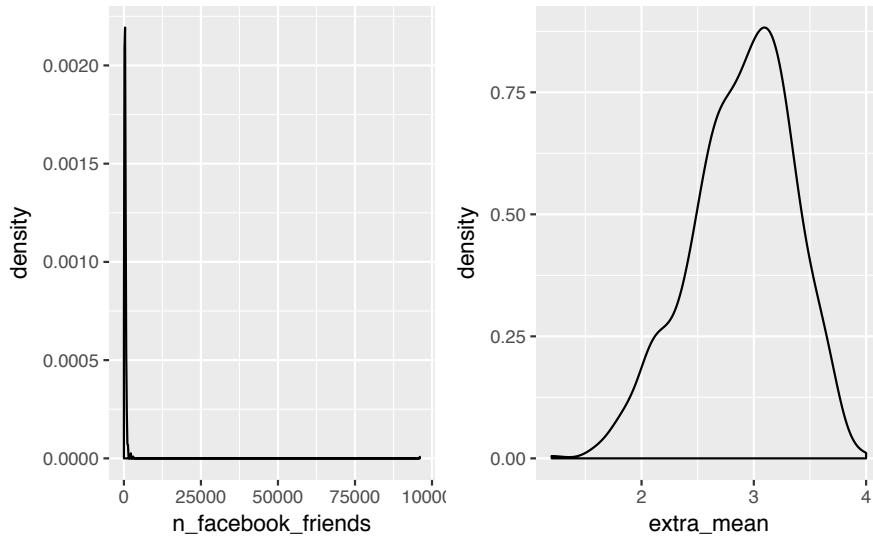


Abbildung 4.3: Visuelles Prüfen der Normalverteilung

```
n_distinct(extra$sex) # es scheint 3 Geschlechter zu geben...
#> [1] 3
```

4.1.10 Auf Normalverteilung prüfen

Einige statistische Verfahren gehen von normalverteilten Variablen aus, daher macht es Sinn, Normalverteilung zu prüfen. *Perfekte* Normalverteilung ist genau so häufig wie *perfekte* Kreise in der Natur. Entsprechend werden Signifikanztests, die ja auf perfekte Normalverteilung prüfen, *immer signifikant* sein, sofern die *Stichprobe groß* genug ist. Daher ist meist zweckmäßiger, einen graphischen “Test” durchzuführen: ein Histogramm, ein QQ-Plot oder ein Dichte-Diagramm als “glatt geschmigelte” Variante des Histogramms bieten sich an (s. Abb. 4.3).

Während die der mittlere Extraversionswert recht gut normalverteilt ist, ist die Anzahl der Facebookfreunde ordentlich (rechts-)schiefl. Bei schießen Verteilung können Transformationen Abhilfe schaffen; ein Thema, auf das wir hier nicht weiter eingehen.

4.1.11 Werte umkodieren und partionieren (“binnen”)

Umkodieren meint, die Werte zu ändern. Man sieht immer mal wieder, dass die Variable “gender” (Geschlecht) mit 1 und 2 kodiert ist. Verwechslungen sind da vorprogrammiert (“Ich bin mir echt ziemlich sicher, dass ich 1 für Männer kodiert habe, wahrscheinlich...”). Besser wäre es, die Ausprägungen `male` und `female` (“Mann”, “Frau”) o.ä. zu verwenden (vgl. Abb. 4.4).

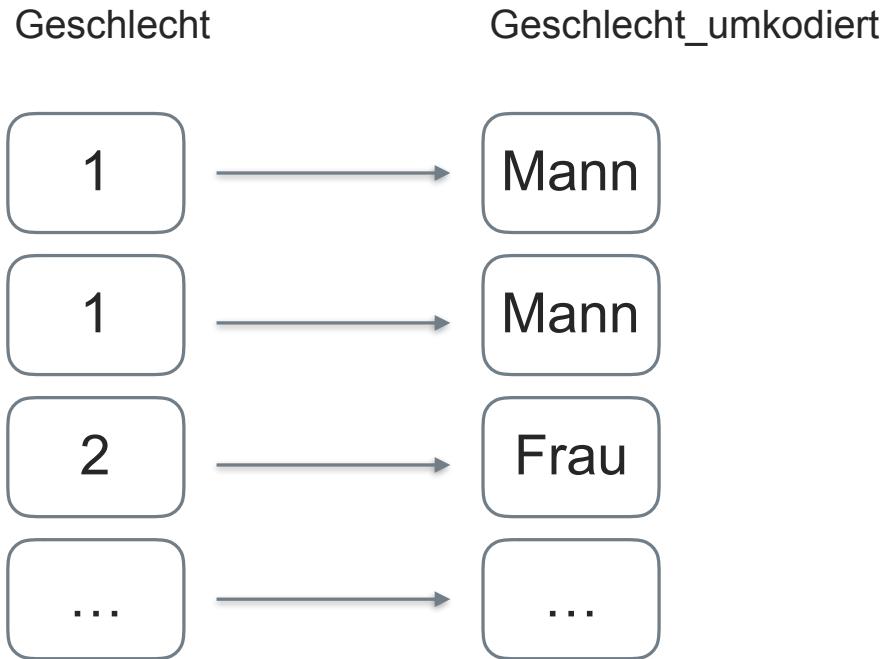


Abbildung 4.4: Sinnbild für Umkodieren

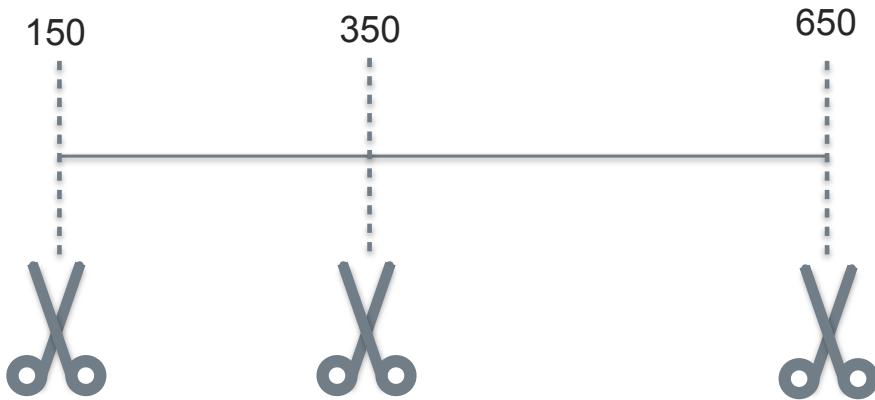


Abbildung 4.5: Sinnbild zum 'Binnen'

Partitionieren) oder ‘*Binnen*’ meint, eine kontinuierliche Variablen in einige Bereiche (mindestens 2) zu zerschneiden. Damit macht man aus einer kontinuierlichen Variablen eine diskrete. Ein Bild erläutert das am einfachsten (vgl. Abb. 4.5).

4.1.11.1 Umkodieren und partitionieren mit `car:::recode`

Manchmal möchte man z.B. negativ gepolte Items umdrehen oder bei kategorialen Variablen kryptische Bezeichnungen in sprechendere umwandeln. Hier gibt es eine Reihe praktischer Befehle, z.B. `recode` aus dem Paket `car`. Schauen wir uns ein paar Beispiele zum Umkodieren an.

```

stats_test$score_fac <- car::recode(stats_test$study_time,
                                      "5 = 'sehr viel'; 2:4 = 'mittel'; 1 = 'wenig'",
                                      as.factor.result = TRUE)
stats_test$score_fac <- car::recode(stats_test$study_time,
                                      "5 = 'sehr viel'; 2:4 = 'mittel'; 1 = 'wenig'",
                                      as.factor.result = FALSE)

stats_test$study_time_2 <- car::recode(stats_test$study_time,
                                         "5 = 'sehr viel'; 4 = 'wenig';
                                         else = 'Hilfe'",
                                         as.factor.result = TRUE)

head(stats_test$study_time_2)
#> [1] sehr viel Hilfe      sehr viel Hilfe      wenig       Hilfe
#> Levels: Hilfe sehr viel wenig

```

Der Befehle `recode` ist praktisch; mit `:` kann man “von bis” ansprechen (das ginge mit `c()` übrigens auch); `else` für “ansonsten” ist möglich und mit `as.factor.result` kann man entweder einen Faktor oder eine Text-Variable zurückgeliefert bekommen. Der ganze “Wechselterm” steht in Anführungsstrichen (`"`). Einzelne Teile des Wechselterms sind mit einem Strichpunkt (`;`) voneinander getrennt.

Das klassische Umkodieren von Items aus Fragebögen kann man so anstellen; sagen wir `interest` soll umkodiert werden:

```

stats_test$no_interest <- car::recode(stats_test$interest,
                                         "1 = 6; 2 = 5; 3 = 4; 4 = 3;
                                         5 = 2; 6 = 1; else = NA")
glimpse(stats_test$no_interest)
#> num [1:306] 2 4 1 5 1 NA NA 4 2 2 ...

```

Bei dem Wechselterm muss man aufpassen, nichts zu verwechseln; die Zahlen sehen alle ähnlich aus...

Testen kann man den Erfolg des Umpolens mit

```

dplyr::count(stats_test, interest)
#> # A tibble: 7 x 2
#>   interest     n
#>   <dbl> <int>
#> 1     1.00    30
#> 2     2.00    47

```

```
#> 3      3.00    66
#> 4      3.21    68
#> 5      4.00    41
#> 6      5.00    45
#> 7      6.00     9
dplyr::count(stats_test, no_interest)
#> # A tibble: 7 x 2
#>   no_interest     n
#>   <dbl> <int>
#> 1      1     9
#> 2      2    45
#> 3      3    41
#> 4      4    66
#> 5      5    47
#> 6      6    30
#> 7      NA   68
```

Scheint zu passen. Noch praktischer ist, dass man so auch numerische Variablen in Bereiche aufteilen kann (“binnen”):

```
stats_test$Ergebnis <- car::recode(stats_test$score,
                                      "1:38 = 'durchgefallen'";
                                      else = 'bestanden'")
```

Natürlich gibt es auch eine Pfeifen kompatible Version, um Variablen umzukodieren bzw. zu binnern: `dplyr::recode`⁵. Die Syntax ist allerdings etwas weniger komfortabel (da strenger), so dass wir an dieser Stelle bei `car::recode` bleiben.

4.1.11.2 Einfaches Umkodieren mit einer Logik-Prüfung

Nehmen wir an, wir möchten die Anzahl der Punkte in einer Statistikklausur (`score`) umkodieren in eine Variable “bestanden” mit den zwei Ausprägungen “ja” und “nein”; der griesgrämige Professor beschließt, dass die Klausur ab 25 Punkten (von 40) bestanden sei. Die Umkodierung ist also von der Art “viele Ausprägungen in zwei Ausprägungen umkodieren”. Das kann man z.B. so erledigen:

```
stats_test$bestanden <- stats_test$score > 24

head(stats_test$bestanden)
#> [1] TRUE  TRUE  TRUE FALSE  TRUE  TRUE
```

⁵<https://blog.rstudio.org/2016/06/27/dplyr-0-5-0/>

Genauso könnte man sich die “Grenzfälle” - die Bemitleidenswerten mit 24 Punkten - anschauen (knapp daneben ist auch vorbei, so der griesgrämige Professor weiter):

```
stats_test$Grenzfall <- stats_test$score == 24

count(stats_test, Grenzfall)
#> # A tibble: 2 x 2
#>   Grenzfall     n
#>   <lgl> <int>
#> 1 FALSE    294
#> 2 TRUE      12
```

Natürlich könnte man auch hier “Durchpfeifen”:

```
stats_test <-
stats_test %>%
  mutate(Grenzfall = score == 24)

count(stats_test, Grenzfall)
#> # A tibble: 2 x 2
#>   Grenzfall     n
#>   <lgl> <int>
#> 1 FALSE    294
#> 2 TRUE      12
```

4.1.11.3 Binnen mit cut

Numerische Werte in Klassen zu gruppieren (“to bin”, denglisch: “binnen”) kann mit dem Befehl `cut` (and friends) besorgt werden.

Es lassen sich drei typische Anwendungsformen unterscheiden:

Eine numerische Variable ...

1. in k gleich große Klassen gruppieren (gleichgroße Intervalle)
2. so in Klassen gruppieren, dass in jeder Klasse n Beobachtungen sind (gleiche Gruppengrößen)
3. in beliebige Klassen gruppieren

4.1.11.3.1 Gleichgroße Intervalle

Nehmen wir an, wir möchten die numerische Variable “Körpergröße” in drei Gruppen einteilen: “klein”, “mittel” und “groß”. Der Range von Körpergröße soll gleichmäßig auf die drei Gruppen

aufgeteilt werden, d.h. der Range (Intervall) der drei Gruppen soll gleich groß sein. Dazu kann man `cut_interval` aus `ggplot2` nehmen⁶.

```
temp <- cut_interval(x = stats_test$score, n = 3)

levels(temp)
#> [1] "[17,24.7]" "(24.7,32.3]" "(32.3,40]"
```

`cut_interval` liefert eine Variable vom Typ `factor` zurück. Hier haben wir das Punktespektrum in drei gleich große Bereiche unterteilt (d.h. mit jeweils gleichem Punkte-Range).

4.1.11.3.2 Gleiche Gruppengrößen

```
temp <- cut_number(stats_test$score, n = 2)
str(temp)
#> Factor w/ 2 levels "[17,31]", "(31,40)": 1 1 2 1 2 2 2 1 1 2 ...
median(stats_test$score)
#> [1] 31
```

Mit `cut_number` (aus `ggplot2`) kann man einen Vektor in `n` Gruppen mit (etwa) gleich viel Observationen einteilen. Hier haben wir `score` am Median geteilt.

Teilt man einen Vektor in zwei gleich große Gruppen, so entspricht das einer Aufteilung am Median (Median-Split).

4.1.11.3.3 In beliebige Klassen gruppieren

```
stats_test$punkte_gruppe <- cut(stats_test$score,
                                breaks = c(-Inf, 25, 29, 33, 37, 40),
                                labels = c("5", "4", "3", "2", "1"))

count(stats_test, punkte_gruppe)
#> # A tibble: 5 x 2
#>   punkte_gruppe     n
#>       <fctr> <int>
#> 1             5    56
#> 2             4    68
#> 3             3    63
```

⁶d.h. `ggplot2` muss geladen sein; wenn man `tidyverse` lädt, wird `ggplot2` automatisch auch geladen

```
#> 4      2      64
#> 5      1      55
```

`cut` ist im Standard-R (Paket “base”) enthalten. Mit `breaks` gibt man die Intervallgrenzen an. Zu beachten ist, dass man eine Unter- bzw. Obergrenze angeben muss. D.h. der kleinste Wert in der Stichprobe wird nicht automatisch als unterste Intervallgrenze herangezogen. Anschaulich gesprochen ist `cut` ein Messer, das ein Seil (die kontinuierliche Variable) mit einem oder mehreren Schnitten zerschneidet (vgl. Abb. 4.5). Wenn wir 6 Schnitte (`breaks`) tun, haben wir 5 Teile, wie Abb. 4.5 zeigt. Darum müssen wir auch nur 5 (6-1) `labels` für die Teile vergeben.

4.2 Deskriptive Statistiken berechnen

4.2.1 Mittelwerte pro Zeile berechnen

4.2.1.1 `rowMeans`

Um Umfragedaten auszuwerten, will man häufig einen Mittelwert *pro Zeile* berechnen. Normalerweise fasst man eine *Spalte* zu einer Zahl zusammen; aber jetzt, fassen wir eine *Zeile* zu einer Zahl zusammen. Der häufigste Fall ist, wie gesagt, einen Mittelwert zu bilden für jede Person. Nehmen wir an, wir haben eine Befragung zur Extraversion durchgeführt und möchten jetzt den mittleren Extraversionswert pro Person (d.h. pro Zeile) berechnen.

```
extra_items <- extra %>%
  select(i01:i10) # `select` ist aus `dplyr`

# oder:
# select(extra_items, i01:i10)

extra$extra_mw <- rowMeans(extra_items)
```

Da der Datensatz über 28 Spalten verfügt, wir aber nur 10 Spalten heranziehen möchten, um Zeilen auf eine Zahl zusammenzufassen, bilden wir als Zwischenschritt einen “schmäleren” Datensatz, `extra_items`. Im Anschluss berechnen wir mit `rowMeans` die Mittelwerte pro Zeile (engl. “row”).

4.2.2 Mittelwerte pro Spalte berechnen

Eine Möglichkeit ist der Befehl `summary` aus `dplyr`.

```
stats_test %>%
  na.omit %>%
  summarise(mean(score),
            sd(score),
            median(score),
            IQR(score))
#>   mean(score) sd(score) median(score) IQR(score)
#> 1      31     5.37      31        8
```

Die Logik von `dplyr` lässt auch einfach Subgruppenanalysen zu. Z.B. können wir eine Teilmenge des Datensatzes mit `filter` erstellen und dann mit `group_by` Gruppen vergleichen:

```
stats_test %>%
  filter(study_time > 1) %>%
  group_by(interest) %>%
  summarise(median(score, na.rm = TRUE))
#> # A tibble: 6 x 2
#>   interest `median(score, na.rm = TRUE)` 
#>   <dbl>                <dbl>
#> 1 1                    28
#> 2 2                    30
#> 3 3                    33
#> 4 4                    31
#> 5 5                    34
#> 6 6                    34
```

Wir können auch Gruppierungskriterien unterwegs erstellen:

```
stats_test %>%
  na.omit %>%
  filter(study_time > 1) %>%
  group_by(intessiert = interest > 3) %>%
  summarise(md_gruppe = median(score))
#> # A tibble: 2 x 2
#>   intessiert md_gruppe
#>   <lgl>      <int>
#> 1 FALSE       30
#> 2 TRUE        32
```

Die beiden Gruppen von `intessiert` sind “ja, interessiert” (`interest > 3` ist `TRUE`) und “nein, nicht interessiert” (`interest > 3` ist `FALSE`). Außerdem haben wir der Spalte, die die Mediane zurückliefert einen ansprechenderen Namen gegeben (`md_gruppe`).

Etwas expliziter wäre es, `mutate` zu verwenden, um die Variable `interessiert` zu erstellen:

```
stats_test %>%
  na.omit %>%
  filter(study_time > 1) %>%
  mutate(interessiert = interest > 3) %>%
  group_by(interessiert) %>%
  summarise(md_gruppe = median(score),
            mw_gruppe = mean(score))
#> # A tibble: 2 x 3
#>   interessiert md_gruppe mw_gruppe
#>   <lgl>        <int>     <dbl>
#> 1 FALSE          30      31.2
#> 2 TRUE           32      31.8
```

Dieses Mal haben wir nicht nur eine Spalte mit den Medianwerten, sondern zusätzlich noch mit Mittelwerten berechnet.



Statistiken, die auf dem Mittelwert (arithmetisches Mittel) beruhen, sind nicht robust gegenüber Ausreißer: Schon wenige Extremwerte können diese Statistiken so verzerren, dass sie erheblich an Aussagekraft verlieren.

Daher: besser robuste Statistiken verwenden. Der Median, der Modus und der IQR bieten sich an.

4.2.3 Korrelationstabellen berechnen

Korrelationen bzw. Korrelationstabellen lassen sich mit dem R-Standardbefehl `cor` berechnen:

```
stats_test %>%
  select(study_time,interest,score) %>%
  cor()
#>             study_time interest score
#> study_time       1       NA      NA
#> interest         NA      1.000  0.196
#> score            NA      0.196  1.000
```

Oh! Lauter NAs! Besser wir löschen Zeilen mit fehlenden Werten bevor wir die Korrelation ausrechnen:

```
stats_test %>%
  select(study_time:score) %>%
  na.omit %>%
  cor()

#>      study_time self_eval interest score
#> study_time     1.000    0.559   0.461 0.441
#> self_eval      0.559    1.000   0.360 0.628
#> interest       0.461    0.360   1.000 0.223
#> score          0.441    0.628   0.223 1.000
```

Alternativ zu `cor` kann man auch `corrr::correlate` verwenden:

```
stats_test %>%
  select(study_time:score) %>%
  correlate

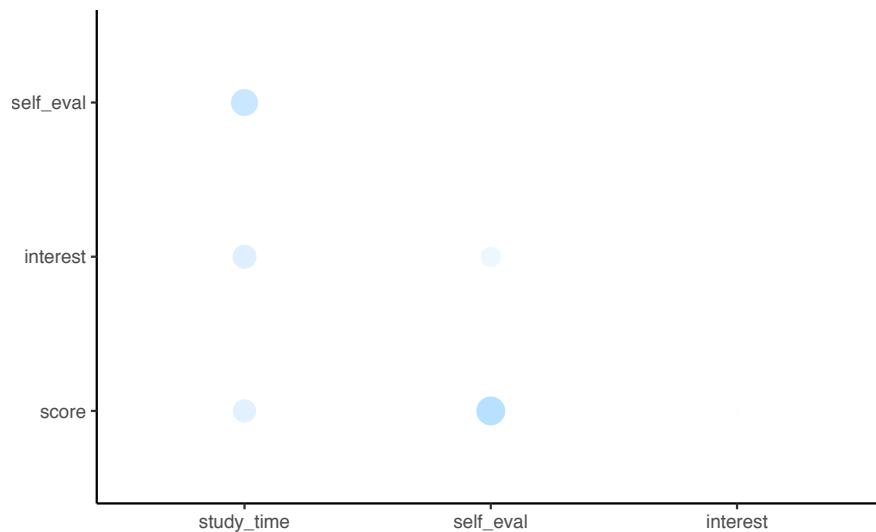
#> # A tibble: 4 x 5
#>   rowname study_time self_eval interest score
#>   <chr>     <dbl>     <dbl>     <dbl> <dbl>
#> 1 study_time     NA     0.559    0.461 0.441
#> 2 self_eval      0.559     NA     0.360 0.628
#> 3 interest       0.461    0.360     NA 0.196
#> 4 score          0.441    0.628    0.196  NA
```

`correlate` hat den Vorteil, dass es bei fehlenden Werten einen Wert ausgibt; die Korrelation wird paarweise mit den verfügbaren (nicht-fehlenden) Werten berechnet. Außerdem wird eine Dataframe (genauer: tibble) zurückgeliefert, was häufig praktischer ist zur Weiterverarbeitung. Wir könnten jetzt die resultierende Korrelationstabelle plotten, vorher “rasieren” wir noch das redundanten obere Dreieck ab (da Korrelationstabellen ja symmetrisch sind):

```
stats_test %>%
  select(study_time:score) %>%
  correlate %>%
  shave %>%
  rplot
```

Tabelle 4.1: Befehle des Kapitels 'Praxisprobleme'

Paket::Funktion	Beschreibung
na.omit	Löscht Zeilen, die fehlende Werte enthalten
nrow	Liefert die Anzahl der Zeilen des Dataframes zurück
complete.cases	Zeigt die Zeilen ohne fehlenden Werte
car::recode	Kodiert Werte um
cut	Schneidet eine kontinuierliche Variable in Wertebereiche
rowMeans	Berechnet Zeilen-Mittelwerte
dplyr::rowwise	Gruppert nach Zeilen
ggplot2::cut_number	Schneidet in n gleich große Bereiche
ggplot2::cut_interval	Schneidet ein Intervalle der Größe k
head	Zeigt nur die ersten Werte eines Objekts an.
scale	z-skaliert eine Variable
dplyr::select_if	Wählt eine Spalte aus, wenn ein Kriterium erfüllt ist
dplyr::glimpse	Gibt einen Überblick über einen Dataframe
dplyr::mutate_if	definiert eine Spalte, wenn eine Kriterium erfüllt ist
:	Definiert einen Bereich von ... bis ...
corr::correlate	Berechnet Korrelationstabelle, liefert einen Dataframe zurück
cor	Berechnet Korrelationstabelle
corr::rplot	Plottet Korrelationsmatrix von correlate
corr::shave	“Rasiert” redundantes Dreieck in Korrelationsmatrix ab



4.3 Befehlsübersicht

Tabelle 4.1 stellt die Befehle dieses Kapitels dar.

Kapitel 5

Fallstudie ‘movies’



Lernziele:

- Grundlegende Funktionen von `dplyr` anwenden können.
- Das Konzept der Pfeife in einem echten Datensatz anwenden können.
- Auch mit relativ großen Daten sicher hantieren können.

Der Datensatz `movies` enthält Bewertungen von Filmen, zusammen mit einigen zusätzlichen Informationen wie Genre, Erscheinungsjahr und Budgethöhe. Wir nutzen diesen Datensatz um uns einige Übung mit Aufbereiten und Zusammenfassen von Daten zu verschaffen.

Für dieses Kapitel werden folgende Pakete benötigt:

```
library(tidyverse) # Datenjudo und Visualisierung  
library(corrr) # Korrelation  
library(ggplot2movies) # Daten
```

Zunächst laden wir die Daten und werfen einen Blick in den Datensatz:

```
data(movies, package = "ggplot2movies")
glimpse(movies)
```

Hier findet man einige Erklärungen zu diesem Datensatz: <http://had.co.nz/data/movies/>.

5.1 Wie viele Filme gibt es pro Genre?

Normalerweise würde man für diese Frage eine Spalte wie “Genre” nehmen und die verschiedenen Werte dieser Spalte auszählen. Das geht sehr bequem mit `dplyr::count`. Hier gibt es allerdings so eine Spalte nicht. Wir müssen uns anders behelfen.

```
movies %>%
  select(Action:Short) %>%
  summarise_all(funs(sum))
#> # A tibble: 1 x 7
#>   Action Animation Comedy Drama Documentary Romance Short
#>   <int>     <int>    <int> <int>       <int>    <int> <int>
#> 1     4688      3690   17271  21811       3472     4744  9458
```

Auf Deutsch heißt diese Syntax



Nimm die Tabelle “movies” UND DANN
 nimm alle Spalten von “Action” bis “Short” UND DANN
 fasse alle Spalten (die wir genommen haben) zusammen und zwar... mit der oder den
 Funktionen “Summe” (sum).

Genau wie der Befehl `summarise` fasst auch `summarise_all` Spalten zu einer Zahl zusammen - nur eben nicht *eine*, sondern *alle* Spalten eines Dataframe. Die Funktion(en), die beim Zusammenfassen verwendet werden sollen, werden mit `funs()` definiert.

5.2 Welches Genre ist am häufigsten?

Bzw. in welchem Genre wurden am meisten Filme gedreht (in unserem Datensatz)?

```
movies %>%
  select(Action:Short) %>%
  summarise_all(funs(sum)) %>%
  gather()
```

```

arrange(-value)
#> # A tibble: 7 x 2
#>   key     value
#>   <chr> <int>
#> 1 Drama  21811
#> 2 Comedy 17271
#> 3 Short   9458
#> 4 Romance  4744
#> 5 Action   4688
#> 6 Animation 3690
#> 7 Documentary 3472

```

Der Befehl `gather` baut einen Dataframe von “breit” nach “lang” um (vgl. Kapitel 2.3). Ah, Schmunzetteln Dramen sind also am häufigsten (wie der Befehl `arrange` dann zeigt). Welcome to Hollywood. :tada:

5.3 Zusammenhang zwischen Budget und Beurteilung

Werden teurere Filme (also Filme mit mehr Budget) besser beurteilt im Schnitt? Das würde man erwarten, denn zum Spaß werden die Investoren wohl nicht ihr Geld raus. Schauen wir es uns an.

```

movies %>%
  select(budget, rating, votes) %>%
  correlate
#> # A tibble: 3 x 4
#>   rowname budget  rating votes
#>   <chr>    <dbl>   <dbl> <dbl>
#> 1 budget      NA -0.0142  0.441
#> 2 rating     -0.0142      NA  0.104
#> 3 votes       0.4413  0.1037    NA

```

Wir haben gerade die drei Spalten `budget`, `rating` und `votes` ausgewählt, dann in der nächsten Zeile die fehlenden Werte eliminiert und schließlich die Korrelation zwischen allen Paaren gebildet. Interessanterweise gibt es keine Korrelation zwischen dem Budget und dem Rating! Teuere Filme sind also mitnichten besser bewertet. Allerdings haben Filme mit mehr Budget eine größere Anzahl an Bewertungen, sind also offenbar bekannter. Vielleicht gehen dann auch entsprechend mehr Leute im Kino - auch wenn diese Filme nicht besser sind. Teurere Filme sind also bekannter, wenn auch nicht besser (beurteilt); so könnte man die Daten lesen.

5.4 Wurden die Filme im Lauf der Jahre teurer und/oder “besser”?

```
movies %>%
  select(year, rating, budget) %>%
  correlate
#> # A tibble: 3 x 4
#>   rowname     year   rating   budget
#>   <chr>      <dbl>    <dbl>    <dbl>
#> 1 year        NA -0.0699  0.2907
#> 2 rating     -0.0699     NA -0.0142
#> 3 budget     0.2907 -0.0142     NA
```

Offenbar wurden die Filme im Lauf der Zeit nicht besser beurteilt: Die Korrelation von `year` und `rating` ist praktisch Null. Wohl wurden sie aber teurer: Die Korrelation von `year` und `budget` ist substanzial.

Kapitel 6

Daten visualisieren



Lernziele:

- An einem Beispiel erläutern können, warum/ wann ein Bild mehr sagt, als 1000 Worte.
- Häufige Arten von Diagrammen erstellen können.
- Diagramme bestimmten Zwecken zuordnen können.

In diesem Kapitel werden folgende Pakete benötigt:

```
library(tidyverse) # Zum Plotten
library(ggplot2movies) # Daten 'movies'
# library(prada) # optional: Daten 'wo_men', 'stats'test'
# library(AER) # optional: Daten 'Affairs'
# library(okcupiddata) # optional: Daten 'profiles'
```

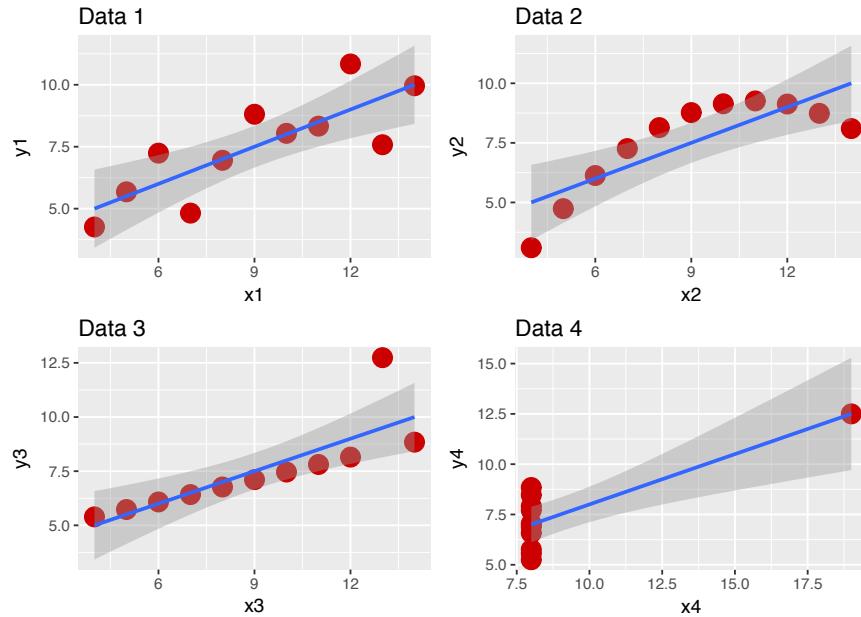
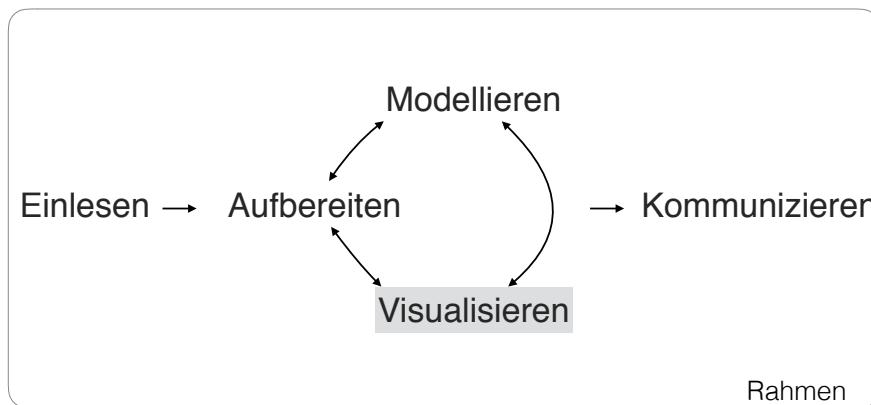


Abbildung 6.1: Das Anscombe-Quartett



Dieses Kapitel erläutert das Daten visualisieren anhand des R-Pakets ggplot2.

6.1 Ein Bild sagt mehr als 1000 Worte

Ein Bild sagt bekanntlich mehr als 1000 Worte. Schauen wir uns zur Verdeutlichung das berühmte Beispiel von Anscombe¹ an. Es geht hier um vier Datensätze mit zwei Variablen (Spalten; X und Y). Offenbar sind die Datensätze praktisch identisch: Alle X haben den gleichen Mittelwert und die gleiche Varianz; dasselbe gilt für die Y. Die Korrelation zwischen X und Y ist in allen vier Datensätzen gleich. Allerdings erzählt eine Visualisierung der vier Datensätze eine ganz andere Geschichte.

Offenbar “passieren” in den vier Datensätzen gänzlich unterschiedliche Dinge. Dies haben die

¹<https://de.wikipedia.org/wiki/Anscombe-Quartett>

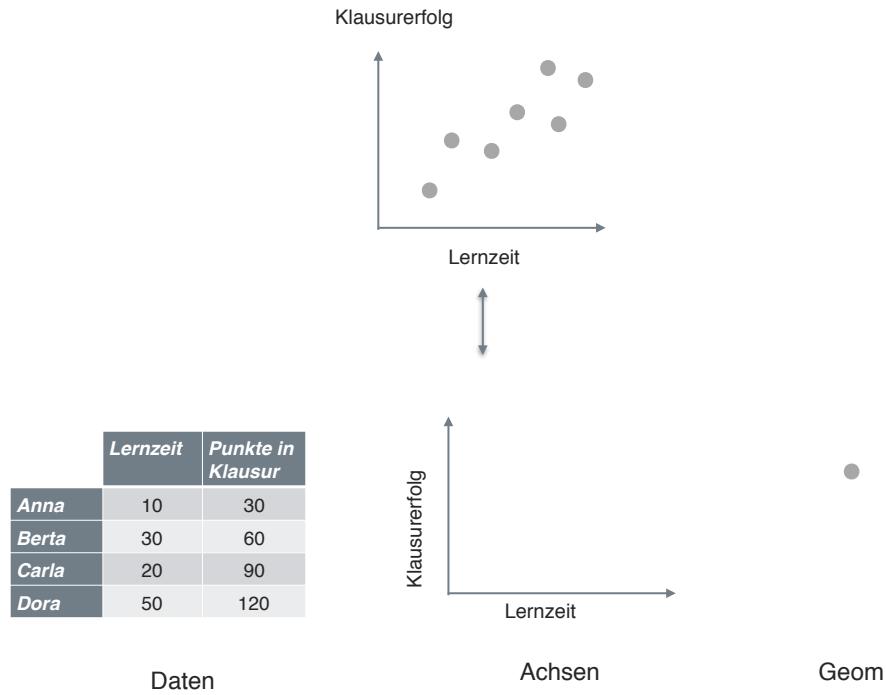


Abbildung 6.2: Anatomie eines Diagramms

Statistiken nicht aufgedeckt; erst die Visualisierung erhellte uns... Kurz: Die Visualisierung ist ein unverzichtbares Werkzeug, um zu verstehen, was in einem Datensatz (und damit in der zugrunde liegenden “Natur”) passiert.

Eine coole Variante mit der gleichen Botschaft findet sich hier² bzw. mit einer Animation hier³; vgl. Matejka und Fitzmaurice (2017).

Es gibt viele Möglichkeiten, Daten zu visualisieren (in R). Wir werden uns hier auf einen Weg bzw. ein Paket konzentrieren, der komfortabel, aber mächtig ist und gut zum Prinzip des Durchpfeifens passt: `ggplot2`⁴.

6.2 Die Anatomie eines Diagramms

`ggplot2` unterscheidet folgende Bestandteile (“Anatomie”) eines Diagramms (vgl. Abb. 6.2):

- Daten
- Abbildende Aspekte (Achsen, Farben, ...)
- Geome (statistische Bilder wie Punkte, Linien, Boxplots, ...)

Bei *Daten* muss ein Dataframe angegeben werden. Zu den *abbildenden Aspekten* (in `ggplot2`

²<https://www.autodeskresearch.com/publications/samestats>

³<https://d2f99xq7vri1nk.cloudfront.net/DinoSequentialSmaller.gif>

⁴“gg” steht für “grammar of graphics” nach einem Buch von Wilkinson(2006); “plot” steht für “to plot”, also ein Diagramm erstellen (“plotten”); vgl. <https://en.wikipedia.org/wiki/Ggplot2>

als “aesthetics” bzw. `aes` bezeichnet) zählen vor allem die Achsen, aber auch Farben u.a. Was ist mit abbildend gemeint? Weist man einer Achse einen Variable zu, so wird jede Ausprägung der Variablen einer Ausprägung der Achse zugeordnet (welcher Wert genau entscheidet `ggplot2` für uns, wenn wir es nicht explizieren). Mit `Geom` ist das eigentlich Art von “Bild” gemeint, wie Punkt, Linie oder Boxplot (vgl. Abschnitt 6.10).

Erstellt `ggplot2` ein Diagramm, so ordnet es Spalten den Bestandteilen des zu erzeugenden Diagramms zu (auch “mapping” genannt).

6.3 Einstieg in `ggplot2` - `qplot`

Los geht's! Laden wir zuerst den Datensatz `movies`.

```
data(movies, package = "ggplot2movies")
```

Betrachten Sie zum Einstieg das Diagramm 6.3.

1. Welche Variable steht auf der X-Achse?
2. Welche Variable steht auf der Y-Achse?
3. Was wird gemalt? Linien, Boxplots, Punkte?
4. Wie heißt der Datensatz, aus dem die Daten gezogen werden?

Der Befehl, der dieses Diagramm erzeugte, heißt `qplot`. Es ist ziemlich genau die Antwort auf die Übungsfragen von gerade eben:

```
qplot(x = year,
      y = budget,
      geom = "point",
      data = movies)
```

Schauen wir uns den Befehl `qplot` etwas näher an. Wie ist er aufgebaut?



`qplot`: Erstelle schnell (q wie quick in `qplot`) mal einen Plot (engl. “plot”: Diagramm).
`x`: Der X-Achse soll die Variable “year” zugeordnet werden.
`y`: Der Y-Achse soll die Variable “budget” zugeordnet werden.
`geom`: (“geometrisches Objekt”) Gemalt werden sollen Punkte und zwar pro Beobachtung (hier: Film) ein Punkt; nicht etwa Linien oder Boxplots. `data`: Als Datensatz bitte `movies` verwenden.

Offenbar geht die Schwere in den Budgets auseinander; außerdem scheint das Budget größer zu werden. Genau kann man es aber schlecht erkennen in diesem Diagramm. Besser ist es vielleicht die Daten pro Jahr zusammenzufassen in einem Geom und dann diese Geome zu vergleichen:

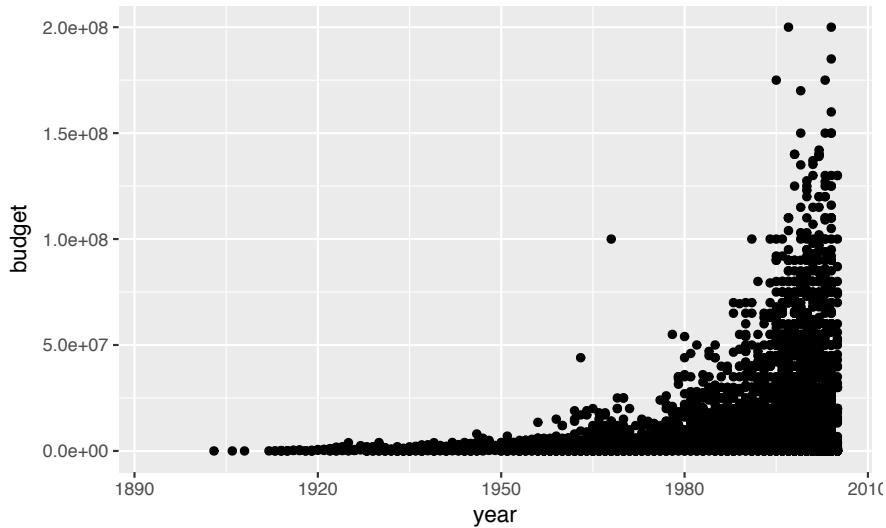
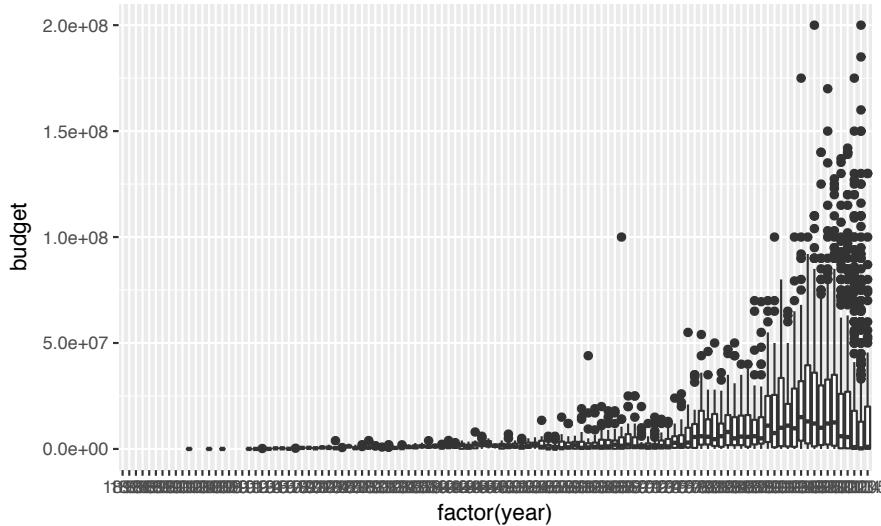


Abbildung 6.3: Mittleres Budget pro Jahr

```
qplot(x = factor(year),
      y = budget,
      geom = "boxplot",
      data = movies)
```



Übrigens: `factor(year)` wird benötigt, um aus `year` eine nominalskalierte Variable zu machen. Nur bei nominalskalierten Variablen auf der X-Achse zeichnet `qplot` mehrere Boxplots nebeneinander. `qplot` bzw. `ggplot2` denkt sich: "Hey, nur wenn es mehrere Gruppen gibt, macht es Sinn, die Gruppen anhand von Boxplots zu vergleichen. Also brauchst du eine Gruppierungsvariable - Faktor oder Text - auf der X-Achse!".

Es sind zu viele Jahre, das macht das Diagramm unübersichtlich. Besser wäre es, Jahrzehnte dazustellen. Ein Jahrzehnt ist so etwas wie eine Jahreszahl, von der die letzte Ziffer abge-

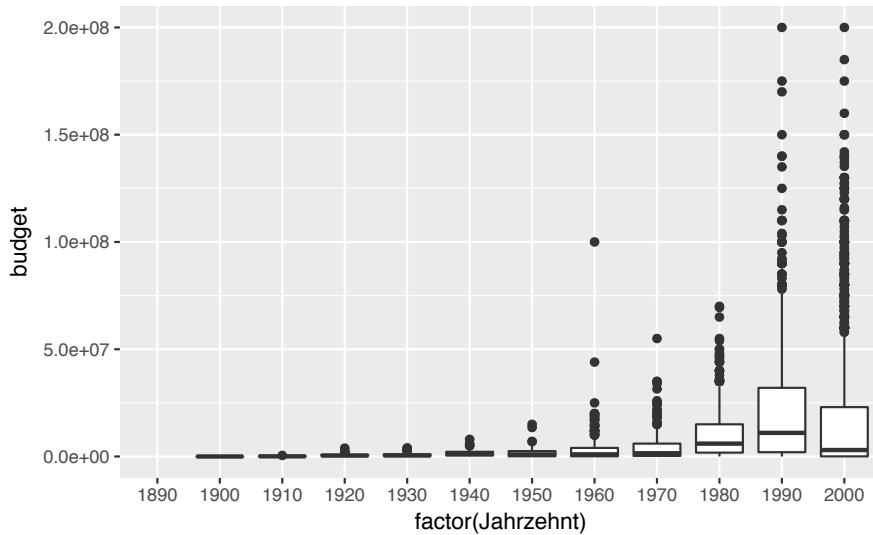


Abbildung 6.4: Film-Budgets über die Jahrzehnte

schnitten (d.h. durch 10 teilen und runden) und dann durch eine Null ersetzt wurde (d.h. mit 10 multiplizieren):

```
movies %>%
  mutate(Jahrzehnt = year / 10) %>%
  mutate(Jahrzehnt = trunc(Jahrzehnt)) %>% # trunkieren, abrunden
  mutate(Jahrzehnt = Jahrzehnt * 10) -> movies
```

Schauen Sie sich die ersten Werte von Jahrzehnt mal an: `movies %>% select(Jahrzehnt) %>% head`.

Ok, auf ein neues Bild (Abb. 6.4):

```
qplot(x = factor(Jahrzehnt),
      y = budget,
      geom = "boxplot",
      data = movies)
```

Aha, gut. Interessanterweise sanken die Budgets gegen Ende unserer Datenreihe; das ist aber vielleicht nur ein Zufallsrauschen.

“q” in `qplot` steht für “quick”. Tatsächlich hat `qplot` einen großen Bruder, `ggplot5`, der deutlich mehr Funktionen aufweist - und daher auch die umfangreichere (komplexere) Syntax. Fangen wir mit `qplot` an.

Diese Syntax des letzten Beispiels ist recht einfach, nämlich:

⁵Achtung: Nicht `qqplot`, nicht `ggplot2`, nicht `gplot...`

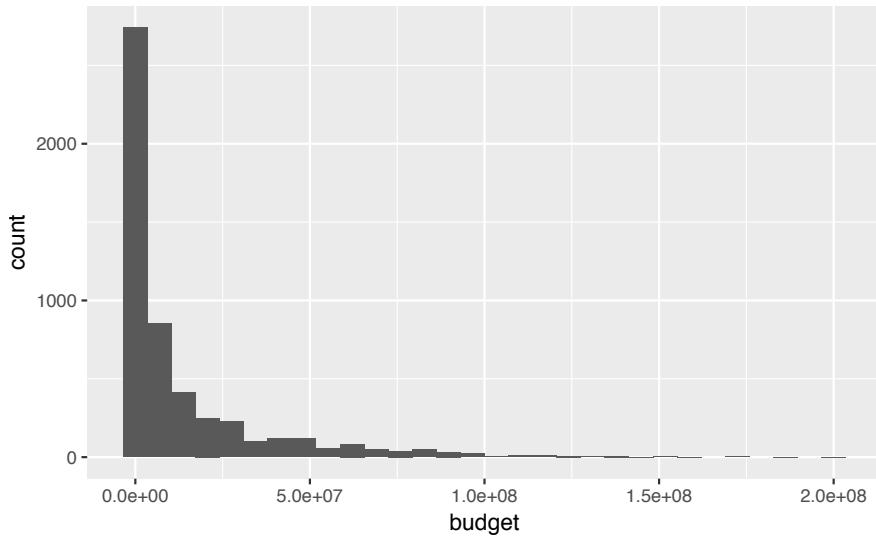


Abbildung 6.5: Verteilung des Budgets von Filmen

```
qplot (x = X_Achse,
       y = Y_Achse,
       data = mein_dataframe,
       geom = "ein_geom")
```

Wir definieren mit `x`, welche Variable der X-Achse des Diagramms zugewiesen werden soll, z.B. `month`; analog mit Y-Achse. Mit `data` sagen wir, in welchem Dataframe die Spalten “wohnen” und als “geom” ist die Art des statistischen “geometrischen Objects” gemeint, also Punkte, Linien, Boxplots, Balken... .

6.4 Häufige Arten von Diagrammen

Unter den vielen Arten von Diagrammen und vielen Arten, diese zu klassifizieren greifen wir uns ein paar häufige Diagramme heraus und schauen uns diese der Reihe nach an.

6.4.1 Eine kontinuierliche Variable

Schauen wir uns die Verteilung von Filmbudgets aus `movies` an (s. Abb. 6.5).

```
qplot(x = budget, data = movies)
```

Weisen wir nur der X-Achse (aber nicht der Y-Achse) eine kontinuierliche Variable zu, so wählt `ggplot2` automatisch als Geom automatisch ein Histogramm; wir müssen daher nicht

explizieren, dass wir ein Histogramm als Geom wünschen (aber wir könnten es hinzufügen).

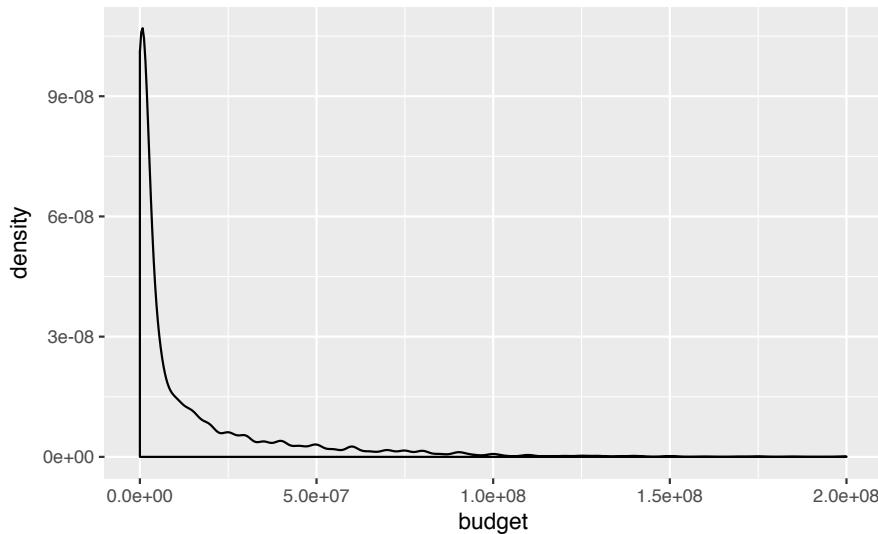


Was heißt das kleine ‘e’, das man bei wissenschaftlichen Zahlen hin und wieder sieht (wie im Diagramm 6.5)?

Zum Beispiel: $5.0\text{e}+07$. Das e sagt, wie viele Stellen im Exponenten (zur Basis 10) stehen: hier 10^7 . Eine große Zahl - eine 1 gefolgt von *sieben* Nullern: 10000000. Die schöne Zahl soll noch mit 5 multipliziert werden: also 50000000. Bei so vielen Nullern kann man schon mal ein Flimmern vor den Augen bekommen... Daher ist die “wissenschaftliche” Notation ganz praktisch, wenn die Zahlen sehr groß (oder sehr klein) werden. Sehr kleine Zahlen werden mit dieser Notation so dargestellt: $5.0\text{e}-07$ heißt $\frac{1}{10^7}$. Eine Zahl sehr nahe bei Null. Das Minuszeichen zeigt hier, dass wir den Kehrwert der großen Zahl nehmen sollen.

Alternativ wäre ein Dichtediagramm hier von Interesse:

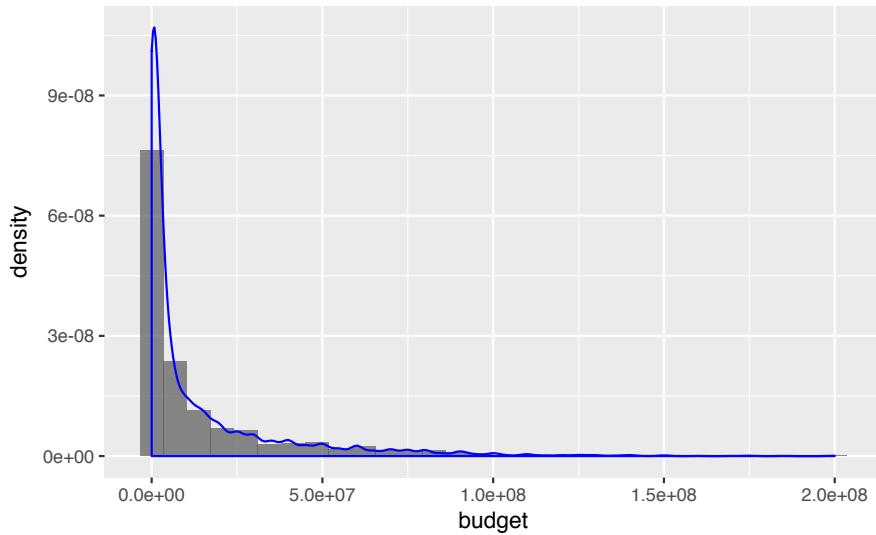
```
qplot(x = budget, data = movies, geom = "density")
```



Was man sich merken muss, ist, dass hier nur das Geom mit Anführungsstrichen zu benennen ist, die übrigen Parameter *ohne*.

Vielleicht wäre es noch schön, beide Geome zu kombinieren in einem Diagramm. Das ist etwas komplizierter; wir müssen zum großen Bruder `ggplot` umsteigen, da `qplot` nicht diese Funktionen anbietet.

```
ggplot(data = movies) +
  aes(x = budget) +
  geom_histogram(aes(y = ..density..), alpha = .7) +
  geom_density(color = "blue")
```



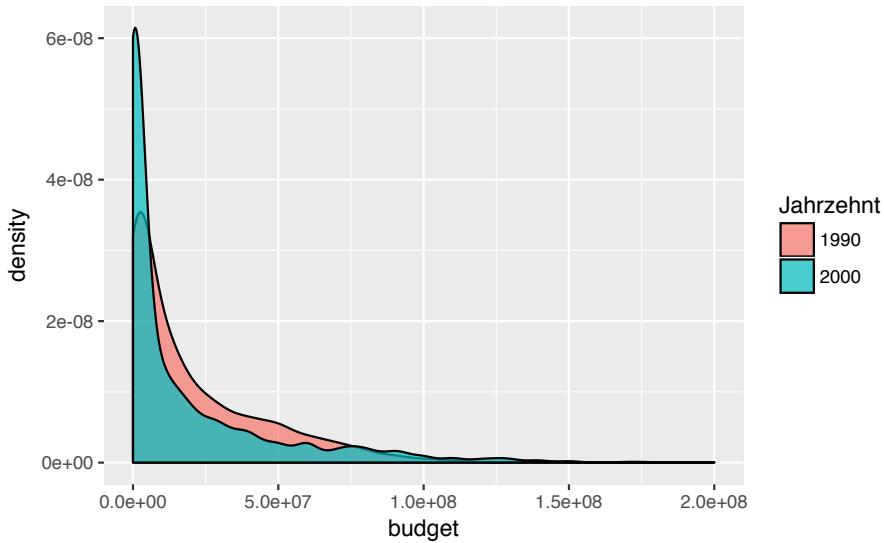
Zuerst haben wir mit dem Parameter `data` den Dataframe benannt. `aes` definiert, welche Variablen welchen Achsen (oder auch z.B. Füllfarben) zugewiesen werden. Hier sagen wir, dass die Schuhgröße auf X-Achse stehen soll. Das `+`-Zeichen trennt die einzelnen Bestandteile des `ggplot`-Aufrufs voneinander. Als nächstes sagen wir, dass wir gerne ein Histogram hätten: `geom_histogram`. Dabei soll aber nicht wie gewöhnlich auf der X-Achse die Häufigkeit stehen, sondern die Dichte. `ggplot` berechnet selbstständig die Dichte und nennt diese Variable `..density..`; die vielen Punkte sollen wohl klar machen, dass es sich nicht um eine “normale” Variable aus dem eigenen Datenframe handelt, sondern um eine “interne” Variable von `ggplot` - die wir aber nichtsdestotrotz verwenden können. `alpha` bestimmt die “Durchsichtigkeit” eines Geoms; spielen Sie mal etwas damit herum. Schließlich malen wir noch ein blaues Dichtediagramm *über* das Histogramm.

Wünsche sind ein Fass ohne Boden... Wäre es nicht interessant, einzelne Zeiträume (Jahrzehnte) zu vergleichen? Schauen wir uns die letzten Jahrzehnte im Vergleich an.

```
movies2 <- filter(movies, Jahrzehnt > 1980)

movies2 %>%
  mutate(Jahrzehnt = factor(.\$Jahrzehnt)) -> movies2

qplot(x = budget,
      data = movies2,
      geom = "density",
      fill = Jahrzehnt,
      alpha = I(.7))
```

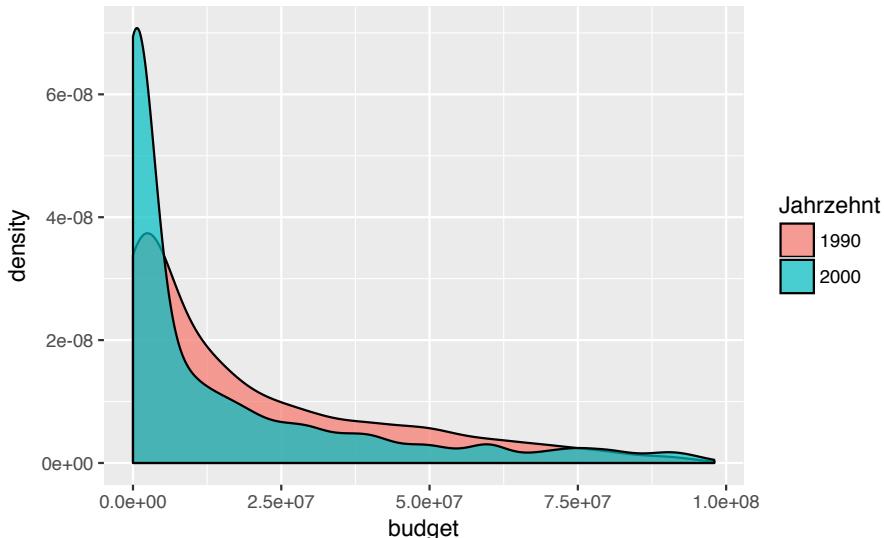


`qplot` erwartet immer *Variablen* als Parameter; wollen wir mal keine Variable, sondern eine fixen Wert, wie 0.7, übergeben, so können wir das mit dem Befehl `I` (wie “identity”) tun.

Hier sollten vielleicht noch die Extremwerte entfernt werden, um den Blick auf das Gros der Werte nicht zu verstauen:

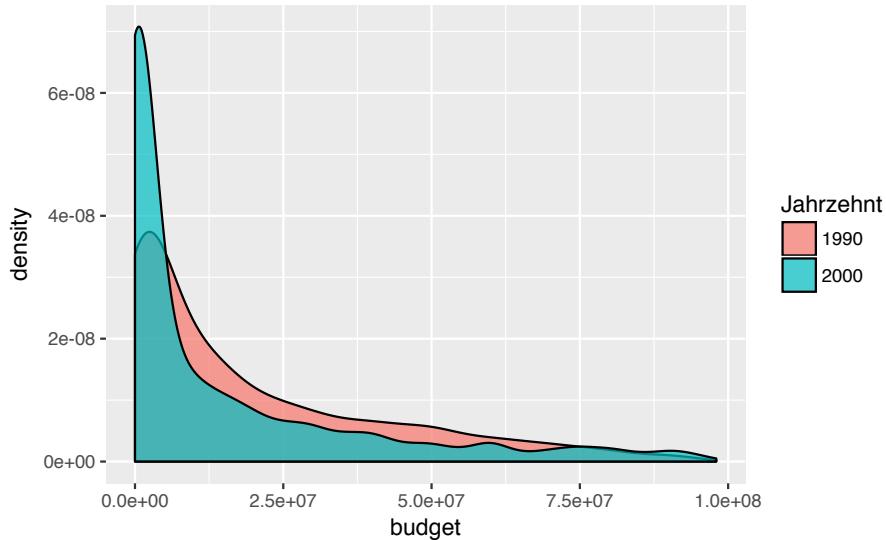
```
movies2 %>%
  filter(budget < 1e08) -> movies2

qplot(x = budget,
      data = movies2,
      geom = "density",
      fill = Jahrzehnt,
      alpha = I(.7))
```



Besser. Man kann das Durchpfeifen auch bis zu `qplot` weiterführen:

```
movies %>%
  filter(budget < 1e+08, Jahrzehnt >= 1990) %>%
  mutate(Jahrzehnt = factor(Jahrzehnt)) %>%
  qplot(x = budget, data = ., geom = "density",
        fill = Jahrzehnt, alpha = I(.7))
```



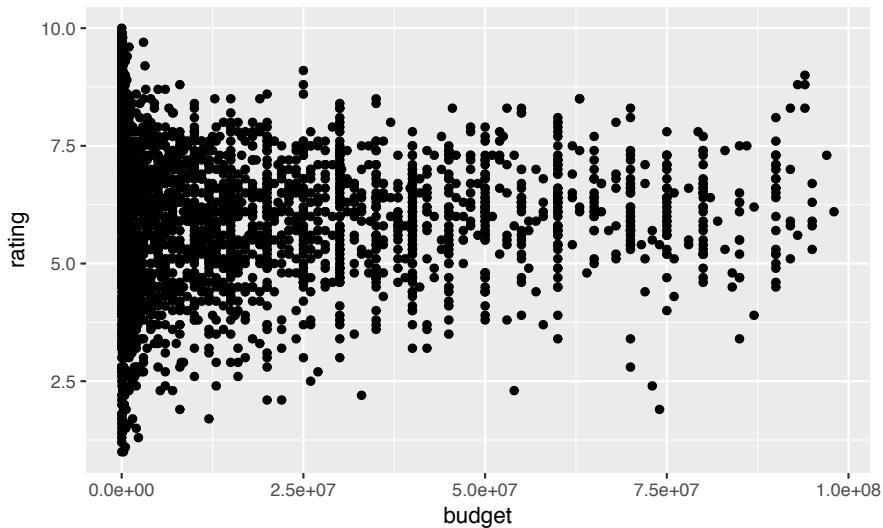
Die Pfeife versucht im Standard, das Endprodukt des letzten Arbeitsschritts an den *ersten* Parameter des nächsten Befehls weiterzugeben. Ein kurzer Blick in die Hilfe von `qplot` zeigt, dass der erste Parameter nicht `data` ist, sondern `x`. Daher müssen wir explizit sagen, an welchen Parameter wir das Endprodukt des letzten Arbeitsschritts geben wollen. Netterweise müssen wir dafür nicht viel tippen: Mit einem schlichten Punkt `.` können wir sagen “nimm den Dataframe, so wie er vom letzten Arbeitsschritt ausgegeben wurde”.

Mit `fill = Jahrzehnt` sagen wir `qplot`, dass er für jedes Jahrzehnt jeweils ein Dichtediagramm erzeugen soll; jedem Dichtediagramm wird dabei eine Farbe zugewiesen (die uns `ggplot2` im Standard voraussucht). Mit anderen Worten: Die Werte von `Jahrzehnt` werden der *Füllfarbe* der Histogramme zugeordnet. Anstelle der Füllfarbe hätten wir auch die Linienfarbe verwenden können; die Syntax wäre dann: `color = sex`. Man beachte, dass die Variable für `fill` oder `color` eine nominale Variable (`factor` oder `character`) sein muss, damit `ggplot2` tut, was will wollen.

6.4.2 Zwei kontinuierliche Variablen

Ein Streudiagramm ist die klassische Art, zwei metrische Variablen darzustellen. Das ist mit `qplot` einfach:

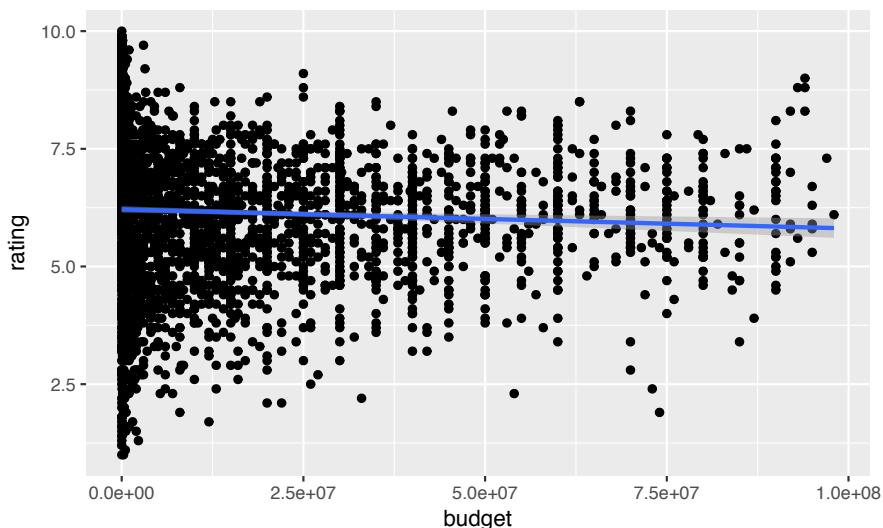
```
p <- qplot(x = budget, y = rating, data = movies2)
p
```



Wir weisen wieder der X-Achse und der Y-Achse eine Variable zu; handelt es sich in beiden Fällen um Zahlen, so wählt ggplot2 automatisch ein Streudiagramm - d.h. Punkte als Geom (geom = "point").

Es ist nicht wirklich ein Trend erkennbar: Teuere Filme sind nicht unbedingt beliebter bzw. besser bewertet. Zeichnen wir eine Trendgerade ein.

```
p + geom_smooth(method = "lm")
```



Synonym könnten wir auch schreiben:

```
wo_men %>%
  filter(height > 150, height < 210, shoe_size < 55) %>%
  ggplot() +
  aes(x = height, y = shoe_size) +
  geom_point() +
  geom_smooth(method = "lm")
```

Da `ggplot` als *ersten* Parameter die Daten erwartet, kann die Pfeife hier problemlos durchgebracht werden. Innerhalb eines `ggplot`-Aufrufs werden die einzelne Teile durch ein Pluszeichen + voneinander getrennt. Nachdem wir den Dataframe benannt haben, definieren wir die Zuweisung der Variablen zu den Achsen mit `aes` (“aes” wie “aesthetics”, also das “Sichtbare” eines Diagramms, die Achsen etc., werden definiert). Ein “Smooth-Geom” ist eine Linie, die sich schön an die Punkte anschmiegt, in diesem Falls als Gerade (lineares Modell, `lm`).

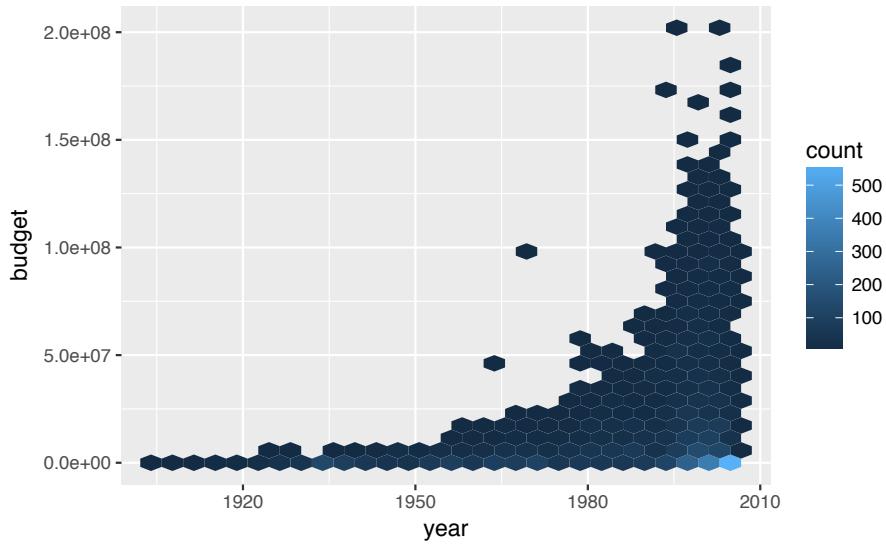
Bei sehr großen Datensätzen, sind Punkte unpraktisch, da sie sich überdecken (“overplotting”). Ein Abhilfe ist es, die Punkte nur “schwach” zu färben. Dazu stellt man die “Füllstärke” der Punkte über `alpha` ein: `geom_point(alpha = 1/100)`. Um einen passablen Alpha-Wert zu finden, bedarf es häufig etwas Probierens. Zu beachten ist, dass es mitunter recht lange dauert, wenn `ggplot` viele (>100.000) Punkte malen soll.

Probieren Sie auch Folgendes aus: Fügen Sie bei `aes` den Parameter `color = sex` hinzu.

Bei noch größeren Datenmengen bietet sich an, den Scatterplot als “Schachbrett” aufzufassen, und das Raster einzufärben, je nach Anzahl der Punkte pro Schachfeld; zwei Geome dafür sind `geom_hex()` und `geom_bin2d()`.

```
nrow(movies) # groß, ein bisschen wenigstens
#> [1] 58788

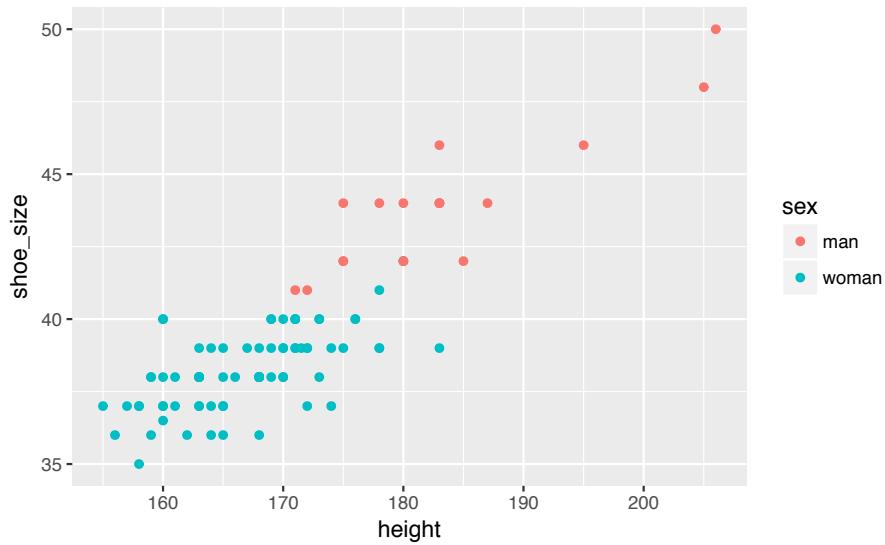
ggplot(movies) +
  aes(x = year, y = budget) +
  geom_hex()
```



Wenn man dies verdaut hat, wächst der Hunger nach einer Aufteilung in Gruppen.

```
wo_men %>%
  dplyr::filter(height > 150, height < 210, shoe_size < 55) -> wo_men2
```

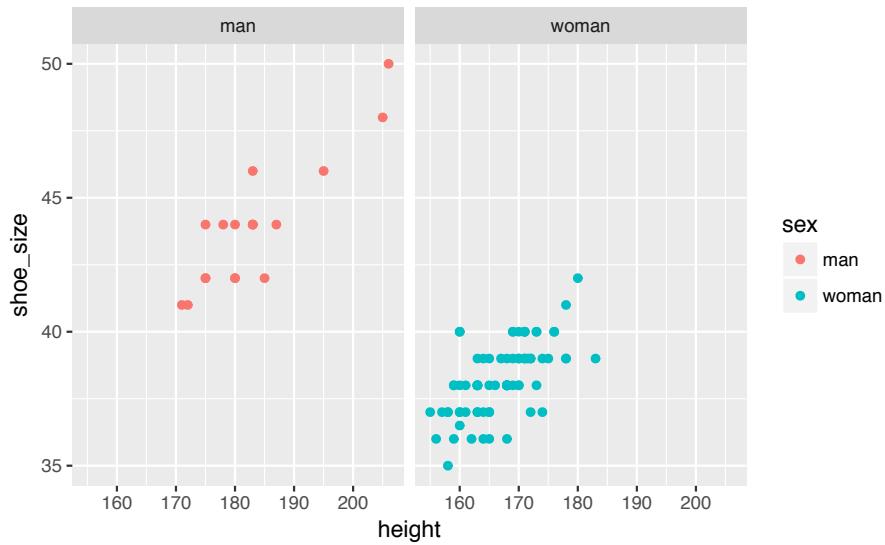
```
wo_men2 %>%
  qplot(x = height, y = shoe_size, color = sex, data = .)
```



Mit `color = sex` sagen wir, dass die Linienfarbe (der Punkte) entsprechend der Stufen von `sex` eingefärbt werden sollen. Die genaue Farbwahl übernimmt `ggplot2` für uns.

Alternativ kann man auch zwei "Teil-Bildchen" ("facets") erstellen, eines für Frauen und eines für Männer:

```
wo_men %>%
  dplyr::filter(height > 150, height < 210, shoe_size < 55) %>%
  qplot(x = height, y = shoe_size, facets = "~sex", color = sex, data = .)
```

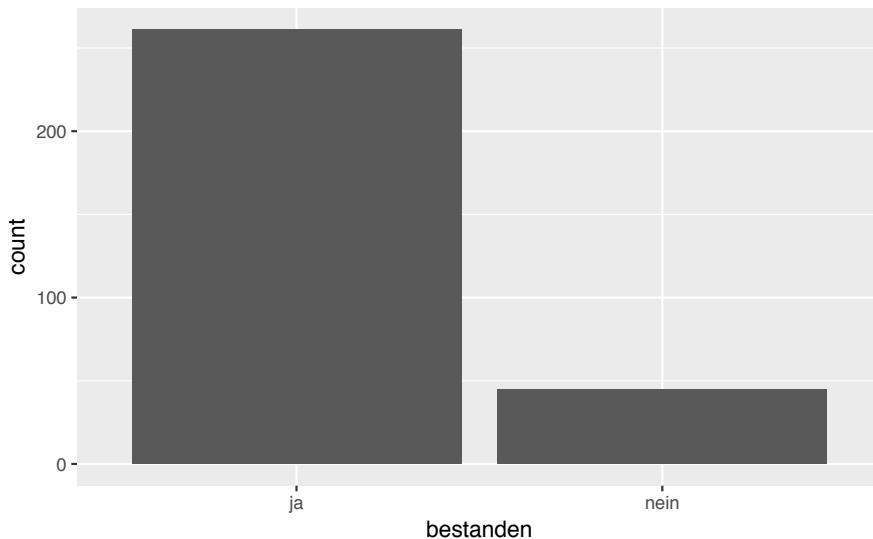


Man beachte die Tilde `~`, die vor die “Gruppierungsvariable” `sex` zu setzen ist.

6.4.3 Eine nominale Variable

Bei nominalen Variablen, geht es in der Regel darum, Häufigkeiten auszuzählen. Ein Klassiker: Wie viele Männer und Frauen finden sich in dem Datensatz? Wie viele Studenten haben (nicht) bestanden?

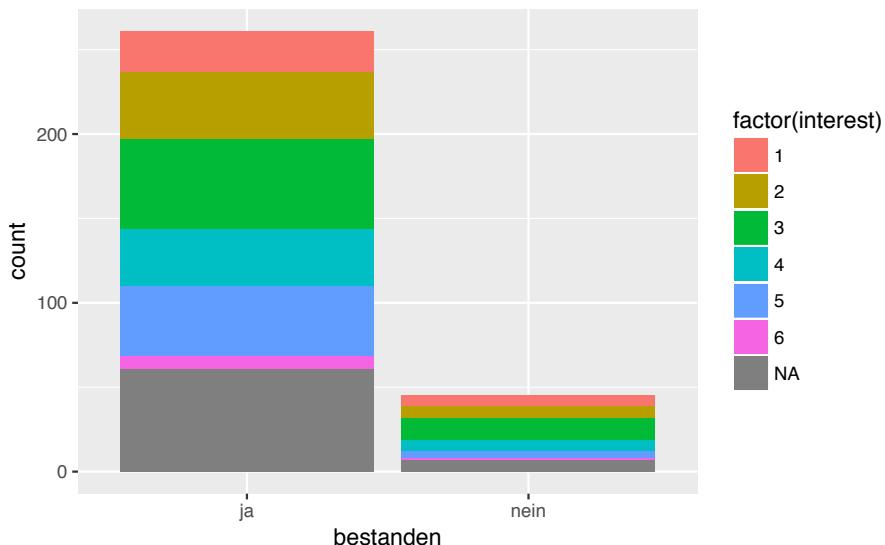
```
stats_test <- read.csv("data/stats_test.csv")
qplot(x = bestanden, data = stats_test)
```



Falls nur die X-Achse definiert ist und dort eine Faktorvariable oder eine Textvariable steht, dann nimmt `qplot` automatisch ein Balkendiagramm als Geom (es steht uns frei, trotzdem `geom = bar` anzugeben).

Wir könnten uns jetzt die Frage stellen, wie viele Nicht-Interessierte und Hoch-Interessierte es in der Gruppe, die bestanden hat (`bestanden == "yes"`) gibt; entsprechend für die Gruppe, die nicht bestanden hat.

```
qplot(x = bestanden, fill = factor(interest), data = stats_test)
```

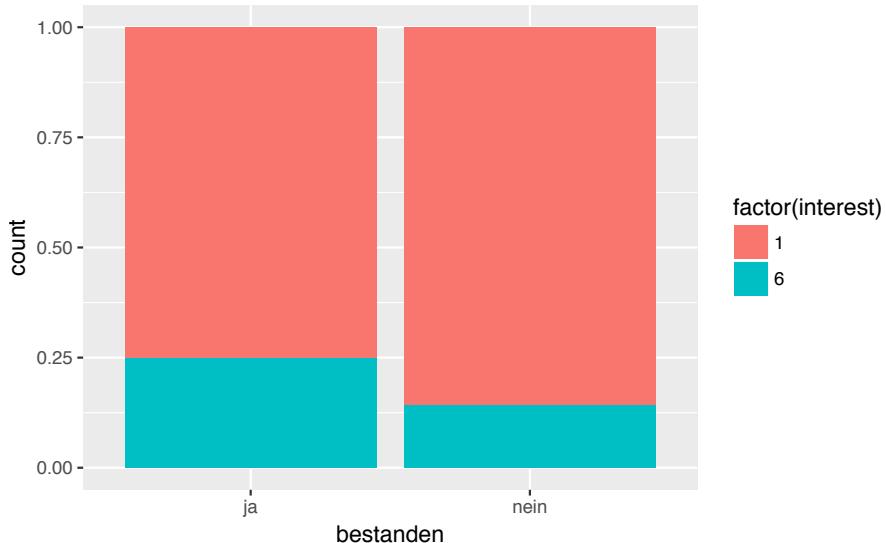


Hier haben wir `qplot` gesagt, dass der die Balken entsprechend der Häufigkeit von `interest` füllen soll. Damit `qplot` (und `ggplot`) sich bequemt, die Füllung umzusetzen, müssen wir aus `interest` eine nominalskalierte Variablen machen - `factor` macht das für uns.

Schön wäre noch, wenn die Balken Anteile (Prozentwerte) angeben würden. Das geht mit

`qplot` (so) nicht; wir schwenken auf `ggplot` um. Und, um die Story zuzuspitzen, schauen wir uns nur die Extremwerte von `interest` an.

```
stats_test %>%
  filter(interest == 1 | interest == 6) %>%
  ggplot() +
  aes(x = bestanden, fill = factor(interest)) +
  geom_bar(position = "fill")
```



Der Lehrer freut sich: In der Gruppe, die bestanden hat, ist der Anteil der `freaks` Hoch-Interessierten größer als bei den Durchfallern.

Schauen wir uns die Struktur des Befehls `ggplot` näher an.



stats_test: Hey R, nimm den Datensatz `stats_test` UND DANN... `ggplot()`: Hey R, male ein Diagramm von Typ `ggplot` (mit dem Datensatz aus dem vorherigen Pfeifen-Schritt, d.h. aus der vorherigen Zeile, also `stats_test`)!

filter: wir wollen nur Zeilen (Studenten), für die gilt `interest == 1` oder `interest == 6`. Der horizontale Strich heißt ‘oder’.

+: Das Pluszeichen grenzt die Teile eines `ggplot`-Befehls voneinander ab.

aes: von “aethetics”, also welche Variablen des Datensatzes den sichtbaren Aspekten (v.a. Achsen, Farben) zugeordnet werden.

x: Der X-Achse (Achtung, `x` wird klein geschrieben hier) wird die Variable `bestanden` zugeordnet.

y: gibt es nicht??? Wenn in einem `ggplot`-Diagramm *keine* Y-Achse definiert wird, wird `ggplot` automatisch ein Histogramm bzw. ein Balkendiagramm erstellen. Bei diesen Arten von Diagrammen steht auf der Y-Achse keine eigene Variable, sondern meist die Häufigkeit des entsprechenden X-Werts (oder eine Funktion der Häufigkeit, wie relative Häufigkeit).

fill Das Diagramm (die Balken) sollen so gefüllt werden, dass sich die Häufigkeit der Werte von `interest` darin widerspiegelt. `geom_XYZ`: Als “Geom” soll ein Balken (“bar”) gezeichnet werden. Ein Geom ist in ggplot2 das zu zeichnende Objekt, also ein Boxplot, ein Balken, Punkte, Linien etc. Entsprechend wird gewünschte Geom mit `geom_bar`, `geom_boxplot`, `geom_pointetc.` gewählt. `position = fill`: `position_fill` sagen, dass die Balken alle eine Höhe von 100% (1) haben, d.h. gleich hoch sind. Die Balken zeigen also nur die Anteile der Werte der `fill`-Variablen.

Die einzige Änderung in den Parametern ist `position = "fill"`. Dieser Parameter weist ggplot an, die Positionierung der Balken auf die Darstellung von Anteilen auszulegen. Damit haben alle Balken die gleiche Höhe, nämlich 100% (1). Aber die “Füllung” der Balken schwankt je nach der Häufigkeit der Werte von `groesse_gruppe` pro Balken (d.h. pro Wert von `sex`).

Wir sehen, dass die Anteile von großen bzw. kleinen Menschen bei den beiden Gruppen (Frauen vs. Männer) *unterschiedlich hoch* ist. Dies spricht für einen *Zusammenhang* der beiden Variablen; man sagt, die Variablen sind *abhängig* (im statistischen Sinne).

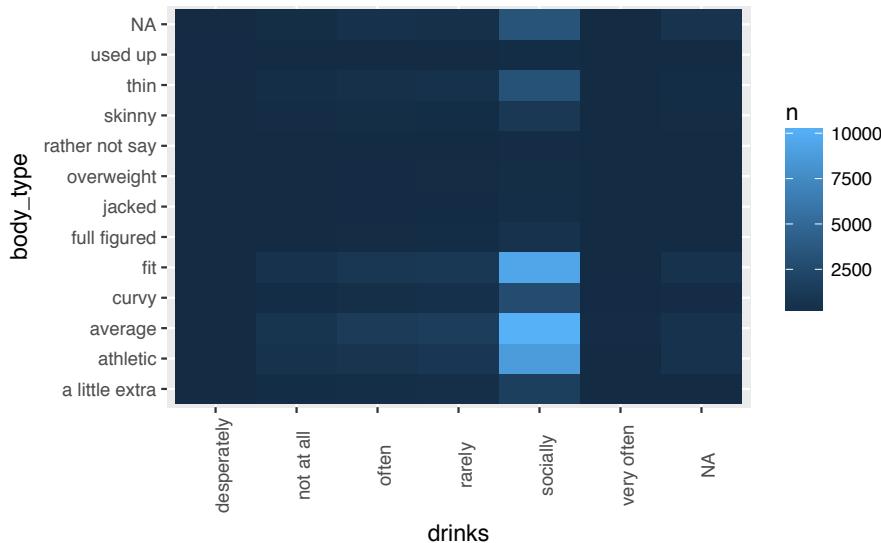
Je unterschiedlicher die “Füllhöhe”, desto stärker sind die Variablen (X-Achse vs. Füllfarbe) voneinander abhängig (bzw. desto stärker der Zusammenhang).

6.4.4 Zwei nominale Variablen

Arbeitet man mit nominalen Variablen, so sind Kontingenztabellen Täglich Brot. Z.B.: Welche Produkte wurden wie häufig an welchem Standort verkauft? Wie viele Narzissen gibt es in welcher Management-Stufe? Wie ist die Verteilung von Alkoholkonsum und Körperperfom bei Menschen einer Single-Börse?. Bleiben wir bei letztem Beispiel.

```
data(profiles, package = "okcupiddata")

profiles %>%
  dplyr::count(drinks, body_type) %>%
  ggplot +
  aes(x = drinks, y = body_type, fill = n) +
  geom_tile() +
  theme(axis.text.x = element_text(angle = 90))
```



Was haben wir gemacht? Also:



Nehme den Datensatz “profiles” UND DANN
 Zähle die Kombinationen von “drinks” und “body_type” UND DANN
 Erstelle ein ggplot-Plot UND DANN
 Weise der X-Achse “drinks” zu, der Y-Achse “body_type” und der Füllfarbe “n” UND DANN
 Male Fliesen UND DANN
 Passe das Thema so an, dass der Winkel für Text der X-Achse auf 90 Grad steht.

Diese Art von Diagramm nennt man auch ‘Mosaicplot’, weil es an ein Mosaic erinnert (wer hätt's gedacht).

Was sofort ins Auge sticht, ist dass “soziales Trinken”, nennen wir es mal so, am häufigsten ist, unabhängig von der Körperform. Ansonsten scheinen die Zusammenhäng nicht sehr stark zu sein.

6.4.5 Zusammenfassungen zeigen

Manchmal möchten wir *nicht* die Rohwerte einer Variablen darstellen, sondern z.B. die Mittelwerte pro Gruppe. Mittelwerte sind eine bestimmte *Zusammenfassung* einer Spalte; also fassen wir zuerst die Körpergröße zum Mittelwert zusammen - gruppiert nach Geschlecht.

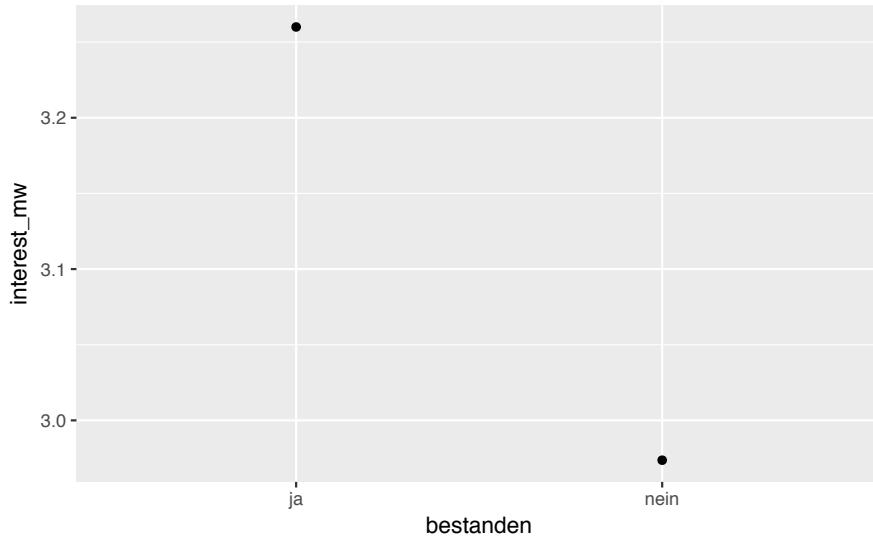
```
stats_test %>%
  group_by(bestanden) %>%
  summarise(interest_mw = mean(interest, na.rm = TRUE)) -> stats_test_summary

stats_test_summary
```

```
#> # A tibble: 2 x 2
#>   bestanden interest_mw
#>   <fctr>      <dbl>
#> 1 ja          3.26
#> 2 nein        2.97
```

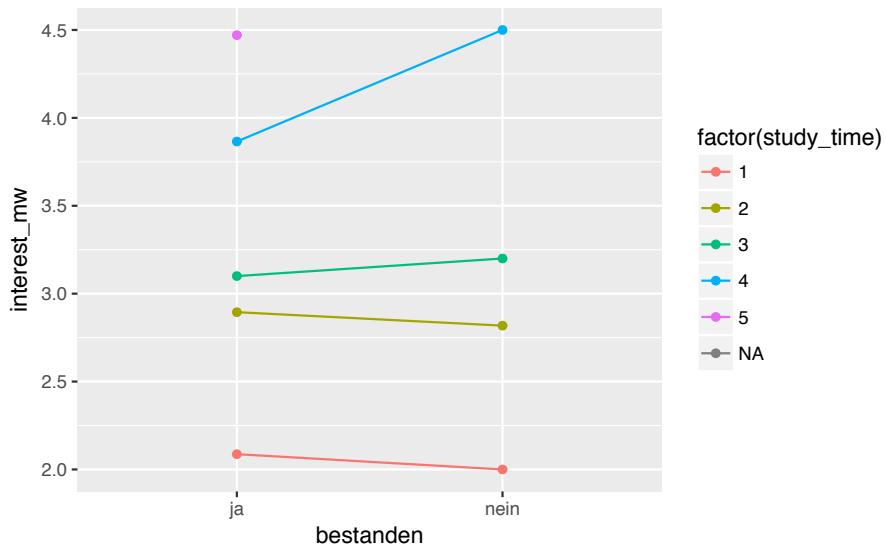
Diese Tabelle schieben wir jetzt in ggplot2; natürlich hätten wir das gleich in einem Rutsch durchpfeifen können.

```
stats_test_summary %>%
  qplot(x = bestanden, y = interest_mw, data = .)
```



Das Diagramm besticht nicht durch die Tiefe und Detaillierung. Bereichern wir das Diagramm um die Frage, wie viel (jeder Student gelernt hat (`study_time`)). Schauen wir uns aber der Einfachheit halber nur die Studenten an, die ganz viel oder ganz wenig gelernt haben.

```
stats_test %>%
  group_by(bestanden, study_time) %>%
  summarise(interest_mw = mean(interest, na.rm = TRUE)) %>%
  qplot(x = bestanden, y = interest_mw, data = ., color = factor(study_time)) +
  geom_line(aes(group = factor(study_time)))
```



In Pseudosyntax:



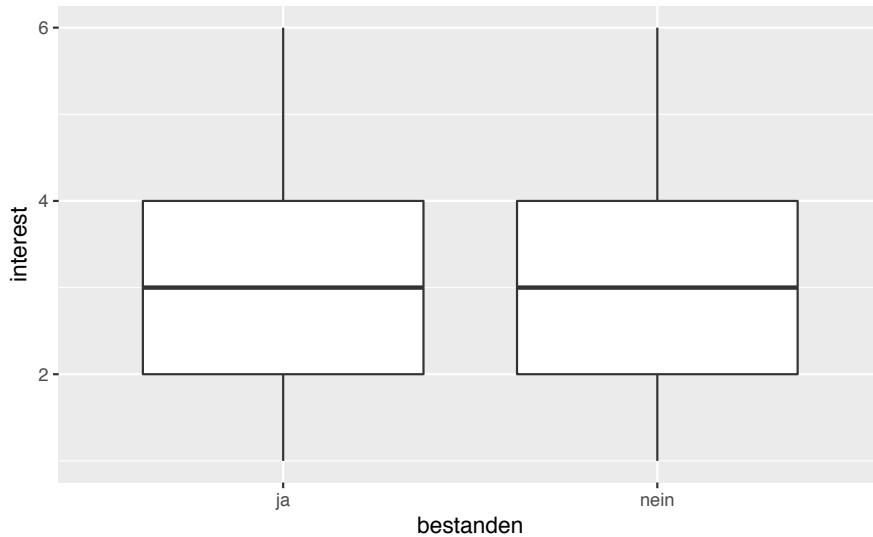
Nehme den Datensatz “stats_test” UND DANN
gruppiere nach den Variablen `bestanden` und `study_time` UND DANN fasse für diese
Gruppen jeweils die Spalte `interest` zum Mittelwert zusammen UND DANN
male einen schnellen Plot mit diesen Daten UND DANN füge ein Liniendiagramm dazu,
wobei jede Stufe von `study_time` eine Gruppe ist. Und Punkte einer Gruppe sollen
verbunden werden.

Warum steht der arme pinkfarbene Punkt bei ‘ja’ und ~4.5 so für sich alleine oder Linie?⁶

Alternativ, und deutlich informationsreicher (besser) sind hier Boxplots.

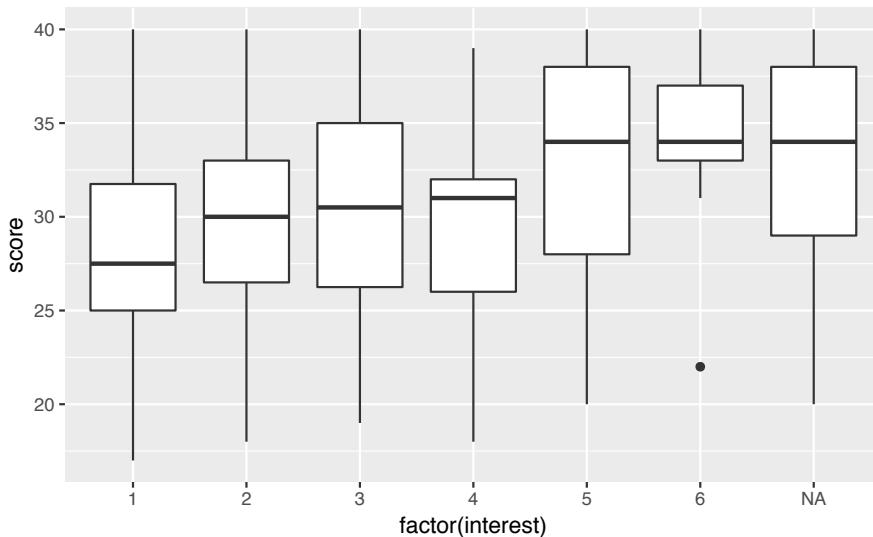
```
qplot(x = bestanden,
      y = interest,
      data = stats_test,
      geom = "boxplot")
```

⁶es gibt kein `study_time == 5` bei den Durchfallen, d.h. bei `bestanden == nein`.



Hm, wie Sie sehen, sehen Sie nix. Kein Unterschied im Median zwischen den Gruppen. Vergleichen wir mal die Punkte zwischen den einzelnen Interessenstufen.

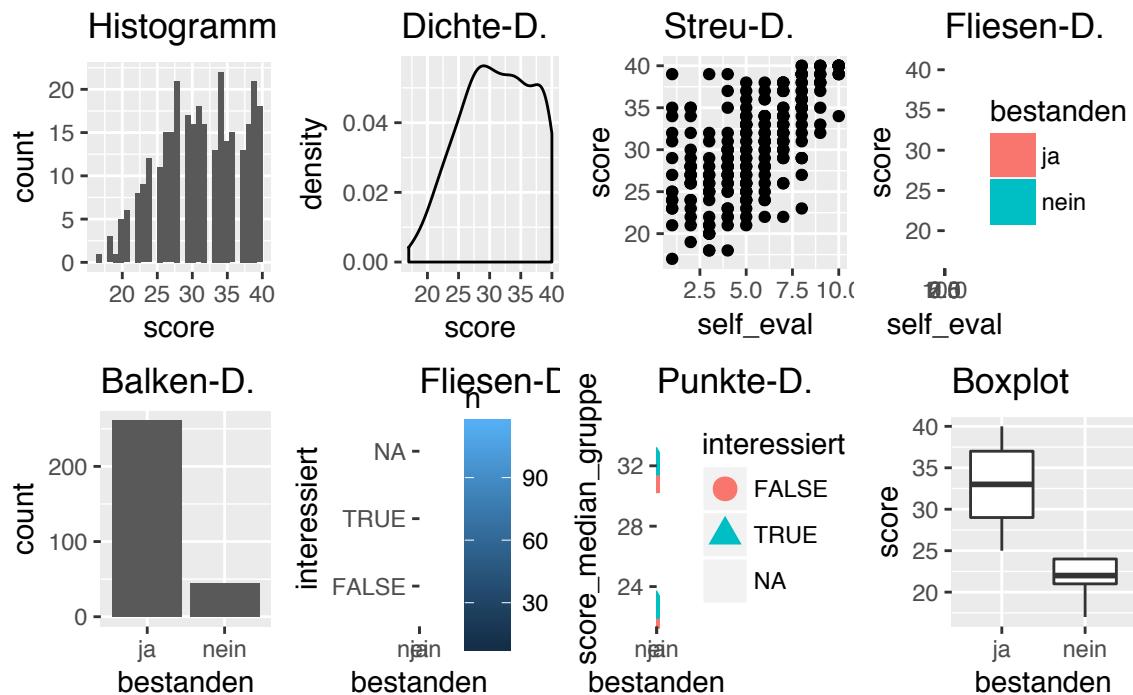
```
qplot(x = factor(interest),
      y = score,
      data = stats_test,
      geom = "boxplot")
```



Das `factor(interest)` brauchen wir, weil `ggplot2` nur dann mehrere Boxplots malt, wenn es Gruppen zum Vergleichen (auf der X-Achse) gibt - sprich wenn auf der X-Achse eine Faktor- oder Textvariable steht.

Tabelle 6.1: Häufige Diagrammtypen

X-Achse	Y-Achse	Diagrammtyp
kontinuierliche Variable	-	Histogramm, Dichtediagramm
kontinuierliche Variable	kontinuierliche Variable	Punkte, Schachbrett-Diagramme
nominale Variable	-	Balkendiagramm
nominale Variable	nominale Variable	Mosaicplot (Fliesen-Diagramm)
nominale Variable	metrische Variable	Punktendiagramm für Zusammenfassungen
nominale Variable	metrische Variable	Boxplots (besser)

**Abbildung 6.6:** Überblick zu häufigen Diagrammtypen

6.4.6 Überblick zu häufigen Diagrammtypen

Die Tabelle 6.1 und Abbildung 6.6 fassen die gerade besprochenen Diagrammtypen zusammen.

6.5 Die Gefühlswelt von ggplot2

- Geben Sie eine *diskrete X-Achse* an und *keine Y-Achse*, so greift `qplot` im Standard auf das Geom `bar` zurück (Balkendiagramm), falls Sie *kein* Geom angeben:

```
qplot(x = score, data = stats_test) # identisch zu
qplot(x = score, data = stats_test, geom = "bar")
```

- Geben Sie eine *kontinuierliche X-Achse* an und *keine Y-Achse*, so greift qplot im Standard auf das Geom `histogram` zurück (Histogramm).

```
qplot(x = score, data = stats_test) # identisch zu
qplot(x = score, data = stats_test, geom = "histogram")
```

- Geben Sie eine *kontinuierliche X-Achse* an und eine *kontinuierliche Y-Achse* an, so greift qplot im Standard auf das Geom `point` zurück (Streudiagramm).

```
qplot(x = score, y = self-eval, data = stats_test) # identisch zu
qplot(x = score, y= self-eval, data = stats_test, geom = "point")
```

- Möchten Sie mehrere Geome für eine Variable darstellen, so muss die Gruppierungs-Variable diskret sein:

```
#oh no:
qplot(x = rating, y = affairs, geom = "boxplot", data = Affairs)

#oh yes:
qplot(x = factor(rating), y = affairs, geom = "boxplot", data = Affairs)

#oh yes:
qplot(x = gender, y = affairs, geom = "boxplot", data = Affairs)
```

6.6 Aufgaben

1. Erzählen Sie einer vertrauenswürdigen Person jeweils eine “Geschichte”, die das Zustandekommen der vier Plots von Anscombe (Abb. 6.1) erklärt!
2. Abb. 6.4 stellt das mittlerer Budget von Filmen dar; als “Geom” wird ein Boxplot verwendet. Andere Geome wären auch möglich - aber wie sinnvoll wären sie?
3. Erstellen Sie ein Diagramm, welches Histogramme der Verspätung verwendet anstelle von Boxplots! Damit das Diagramm nicht so groß wird, nehmen Sie zur Gruppierung nicht `carrier` sondern `origin`.
4. Ist das Histogramm genauso erfolgreich wie der Boxplot, wenn es darum geht, viele Verteilungen vergleichend zu präsentieren? Warum?
5. Erstellen Sie ein sehr grobes und ein sehr feines Histogramm für die Schuhgröße!
6. Vertiefung: Erstellen Sie ein Diagramm, das sowohl eine Zusammenfassung (Mittelwert) der Körpergrößen nach Geschlecht darstellt als auch die einzelnen Werte darstellt!

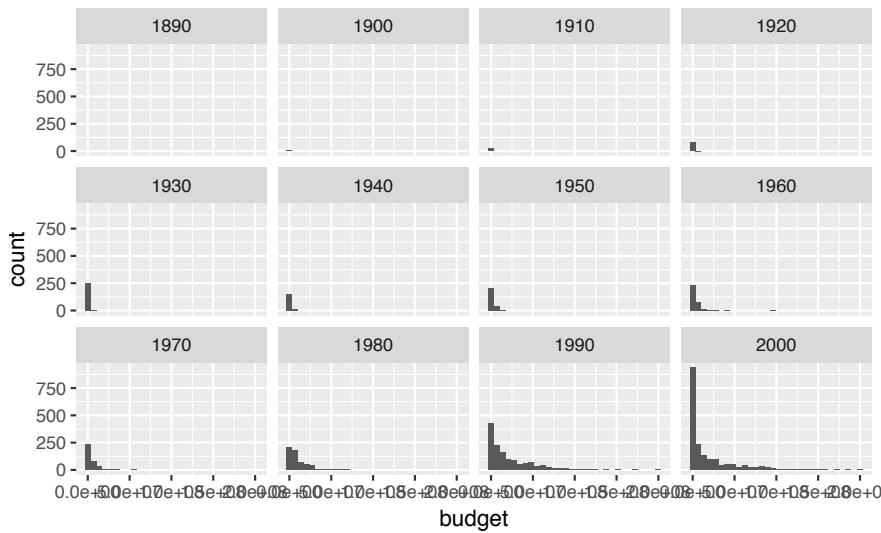


Abbildung 6.7: Film-Budgets mit Histogrammen

6.7 Lösungen

1. :-)

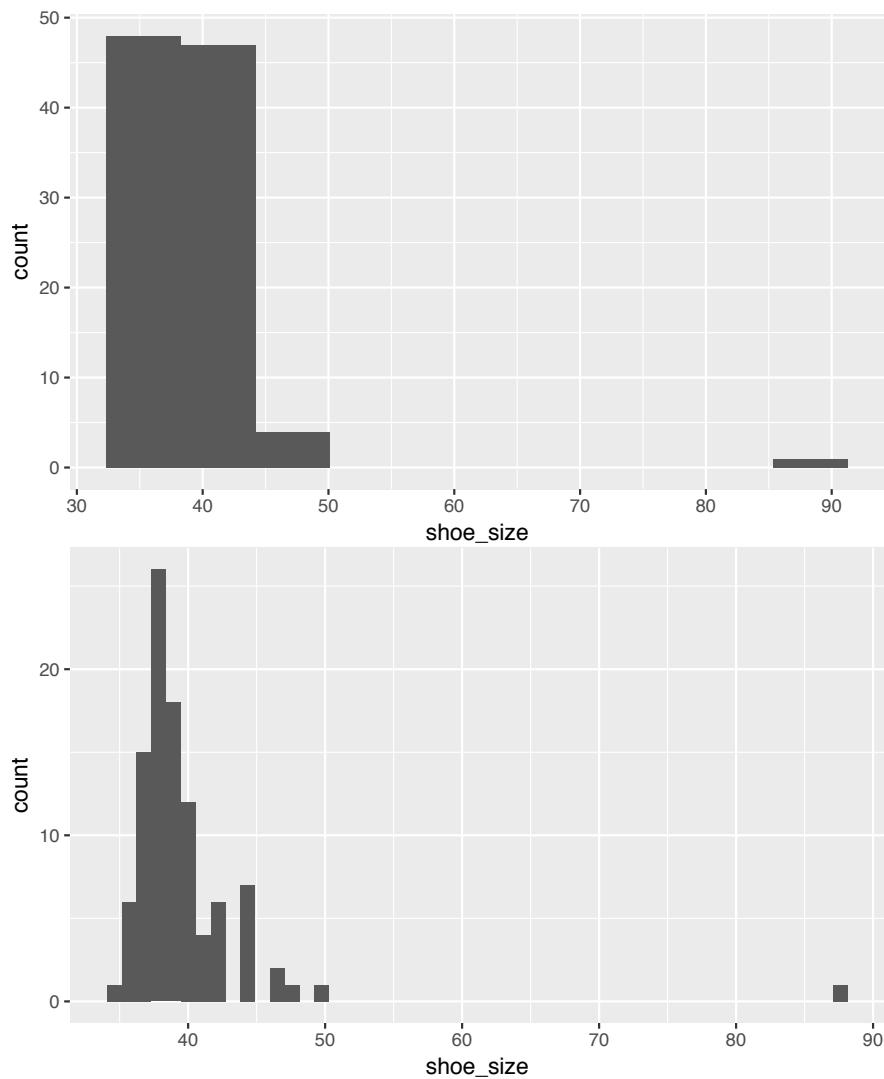
2. :

```
qplot(x = budget, geom = "histogram", data = movies, facets = ~factor(Jahrzehnt))
```

Der Boxplot ist besser geeignet als das Histogramm, um mehrere Verteilungen vergleichend zu präsentieren (vgl. Abb. 6.7). Durch die gleiche Ausrichtung der Boxplots ist es dem Auge viel einfacher, Vergleiche anzustellen im Vergleich zu den Histogrammen. Einen optisch schöneren Effekt könnte man mit `geom_jitter` anstelle von `geom_point` erreichen. Auch die Reihenfolge der beiden Geome könnte man umdrehen. Natürlich ist auch an Form, Größe und Farbe der Geome noch zu feilen.

3. :

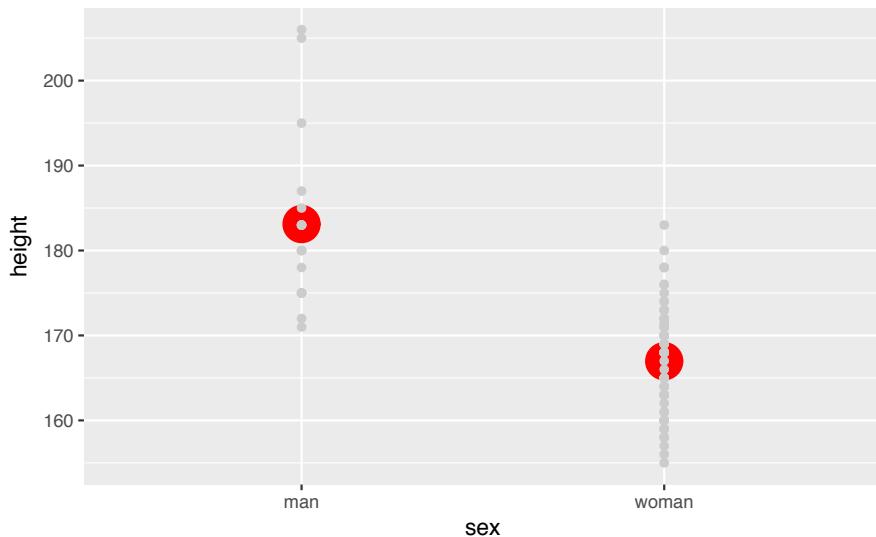
```
qplot(x = shoe_size, data = wo_men, bins = 10)
qplot(x = shoe_size, data = wo_men, bins = 50)
```



4. :

```
wo_men2 %>%
  group_by(sex) %>%
  summarise(height = mean(height)) -> wo_men3
```

```
wo_men3 %>%
  ggplot() +
  aes(x = sex, y = height) +
  geom_point(color = "red", size = 8) +
  geom_point(data = wo_men2, color = "grey80")
```



Der “Trick” ist hier, erst die zusammengefassten Daten in ein Geom zu stecken (`wo_men3`). Dann werden die Rohdaten (`wo_men2`) ebenfalls in ein Geom gepackt. Allerdings muss die Achsen-Beschriftung bei beiden Geomen identisch sein, sonst gibt es eine Fehlermeldung.

6.8 Richtig oder Falsch⁷



Richtig oder Falsch!?

1. Diese Geome gehören zum (Standard-) ggplot2: bar, histogram, point, density, jitter, boxplot.
2. `qplot` ist eine Funktion im Paket `ggplot2`.
3. Mi `aes` definiert man, wie “ästethisch” das Diagramm sein soll (z.B. grauer Hintergrund vs. weißer Hintergrund, Farbe der Achsen etc.).
4. Diese Geome gehören zum (Standard-) ggplot2: smooth, line, boxwhisker, mosaicplot.
5. Möchte man ein Diagramm erstellen, welches auf der X-Achse `total_bill`, auf der Y-Achse `tip` darstellt, als Geom Punkte verwendet und die Daten aus der Tabelle `tips` bezieht, so ist folgende Syntax korrekt: ‘`qplot(x = total_bill, y = tip, geom = “point”, data = tips)`’

6.9 Befehlsübersicht

Tabelle 6.2 fasst die R-Funktionen dieses Kapitels zusammen.

⁷R, R, F, F, R

Tabelle 6.2: Befehle des Kapitels 'Daten visualisieren'

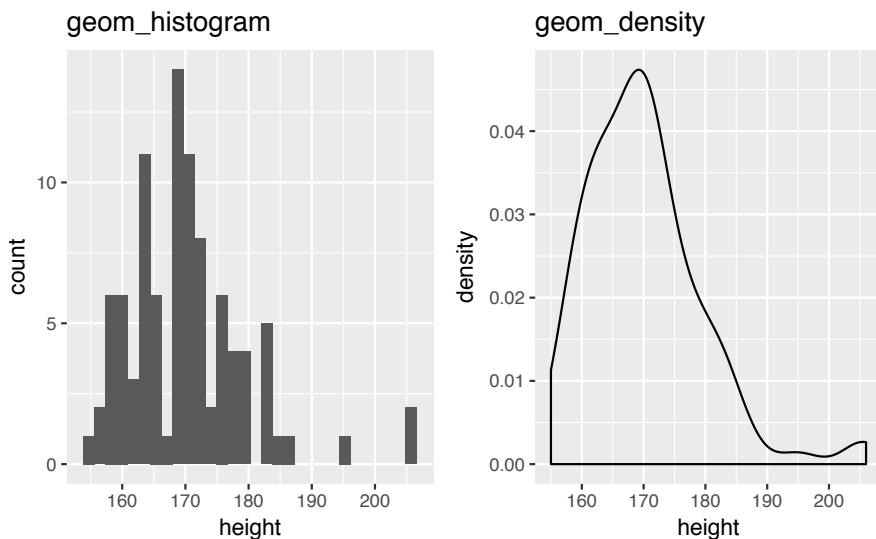
Paket::Funktion	Beschreibung
ggplot2::qplot	Malt schnell mal einen Plot
ggplot2::ggplot	Malt einen Plot
factor	Wandelt einen Vektor in den Typ factor um

6.10 Vertiefung: Geome bei ggplot2

Einen guten Überblick über Geome bietet das Cheatsheet von ggplot2⁸.

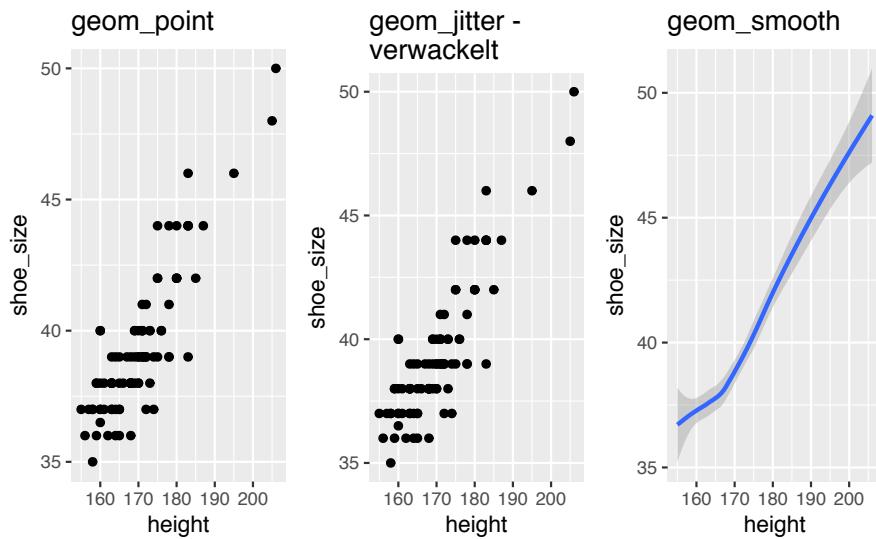
Verschiedenen Taxonomien von statistischen “Bildchen” sind denkbar; eine einfache ist die folgende; es wird nur ein Teil der verfügbaren Geome dargestellt.

1. Eine kontinuierliche Variable



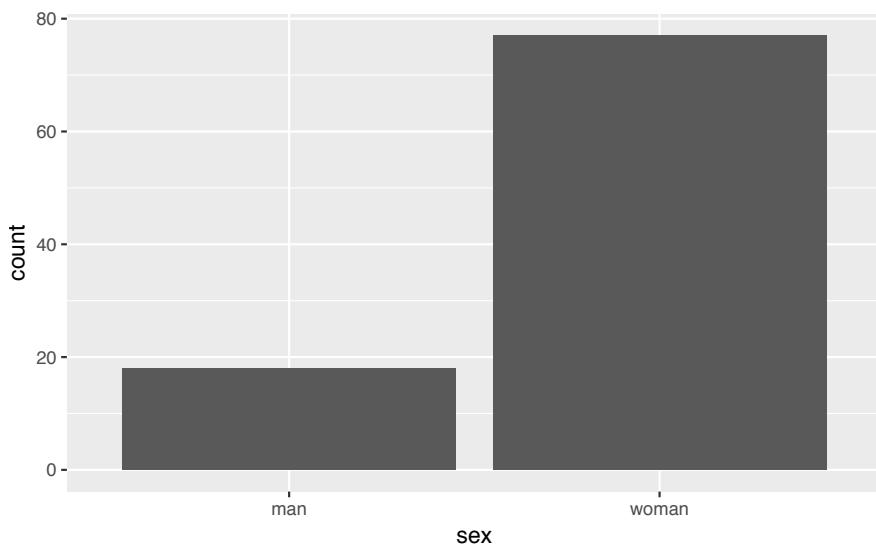
2. Zwei kontinuierliche Variablen

⁸<https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf>

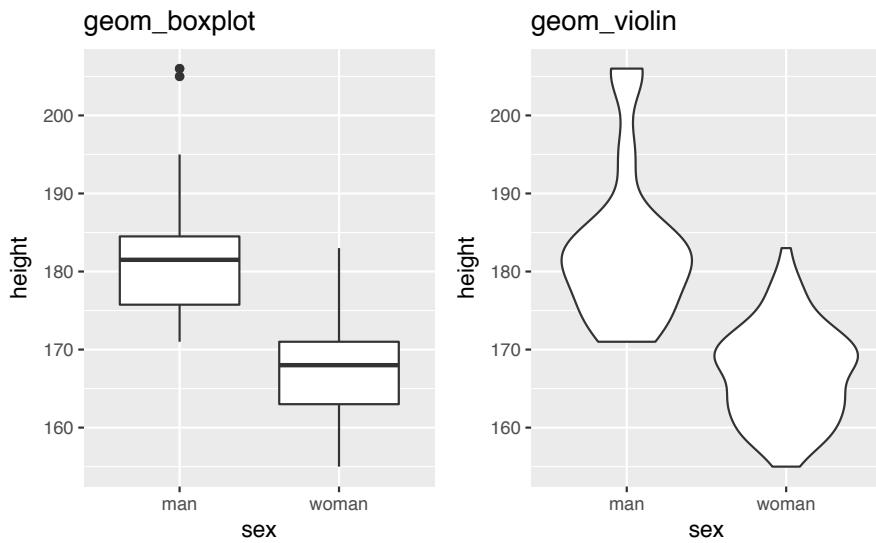


3. Eine diskrete Variable (X-Achse)

```
ggplot(wo_men2) +
  aes(x = sex) +
  geom_bar()
```



4. Eine diskrete Variable auf der X-Achse und eine kontinuierliche Y-Achse



6.11 Verweise

- Einen Befehlsüberblick zu `ggplot2` findet sich hier: <http://ggplot2.tidyverse.org/reference/>.
- Edward Tufte gilt als Grand Seigneur der Datenvisualisierung; er hat mehrere lesenswerte Bücher zu dem Thema geschrieben (Tufte 2001; Tufte 2006; Tufte 1990).
- William Cleveland, ein amerikanischer Statistiker ist bekannt für seine grundlegenden, und weithin akzeptierten Ansätze für Diagramme, die die wesentliche Aussage schnörkellos transportieren (Cleveland 1993).
- Die (graphische) Auswertung von Umfragedaten basiert häufig auf Likert-Skalen. Ob diese metrisches Niveau aufweisen, darf bezweifelt werden. Hier findet sich einige vertiefenden Überlegungen dazu und zur Frage, wie Likert-Daten ausgewertet werden könnten: <https://bookdown.org/Rmadillo/likert/>.
- Es finden sich viele Tutorials online zu `ggplot2`; ein deutschsprachiger Tutorial findet sich hier: <http://md.psych.bio.uni-goettingen.de/mv/unit/ggplot2/ggplot2.html>.

Kapitel 7

Fallstudie zur Visualisierung



Lernziele:

- Diagramme für nominale Variablen erstellen können.
- Balkendiagramme mit Prozentpunkten auf der Y-Achse erstellen können.
- Balkendiagramme drehen können.
- Text-Labels an Balkendiagramme anfügen können.
- Farbschemata von Balkendiagrammen ändern können.

Benötigte Pakete:

```
library(tidyverse)
library(corr)
library(GGally)
```

Eine recht häufige Art von Daten in der Wirtschaft kommen von Umfragen in der Belegschaft. Diese Daten gilt es dann aufzubereiten und graphisch wiederzugeben. Das ist der Gegenstand dieser Fallstudie.

7.1 Daten einlesen

Hier laden wir einen Datensatz von einer Online-Umfrage:

```
extra <- read.csv("data/extracsv")
```

Alternativ können Sie den Datensatz aus dem Paket `prada` laden (mit `data(extra, package = "prada")`) oder aus dem Internet herunterladen:

```
prada_extra_url <-
  paste0("https://raw.github.com/", # Webseite
         "sebastiansauer/", # Nutzer
         "Praxis_der_Datenanalyse/", # Projekt/Repository
         "master/", # Variante
         "data/extrac.csv") # Ordner und Dateinamen

extra <- read.csv(prada_extra_url)
```

Der Datensatz besteht aus 10 Extraversionsitems (B5T nach Satow¹) sowie einigen Verhaltenskorrelaten (zumindest angenommenen). Uns interessieren also hier nur die 10 Extraversionsitems, die zusammen Extraversion als Persönlichkeitseigenschaft messen (sollen). Wir werden die Antworten der Befragten darstellen, aber uns hier keine Gedanken über Messqualität u.a. machen.

Die Umfrage kann hier² eingesehen werden. Schauen wir uns die Daten mal an:

```
glimpse(extra)
```

7.2 Daten umstellen

Wir haben ein Diagramm vor Augen (s.u.), bei dem auf der X-Achse die Items stehen (1,2,...,n) und auf der Y-Achse die Anzahl der Kreuze nach Kategorien.

Viele Grafik-Funktionen sind nun so aufgebaut, dass auf der X-Achsen nur *eine* Variable steht. `ggplot2`, das wir hier verwenden, ist da keine Ausnahme. Wir müssen also die “breite” Tabelle (10 Spalten, pro Item eine) in eine “lange Spalte” umbauen: Eine Spalte heißt dann “Itemnummer” und die zweite “Wert des Items” oder so ähnlich.

Also, los geht’s: Zuerst wählen wir aus der Fülle der Daten, die Spalten, die uns interessieren: Die 10 Extraversionsitems, in diesem Fall (Spalten 3 bis 12).

```
extra_items <- dplyr::select(extra, 3:12)
```

Dann stellen wir die Daten von “breit” nach “lang” um, so dass die Items eine Variable bilden und damit für `ggplot2` gut zu verarbeiten sind.

¹https://www.zpid.de/pub/tests/PT_9006357_B5T_Forschungsbericht.pdf

²https://docs.google.com/forms/d/e/1FAIpQLSfD4wQuhDV_edx1WBfN3Qos7XqoVbe41VpiKLRKtGLeuUD09Q/viewform

```
extra_long <- gather(extra_items, key = items, value = Antwort)

extra_long$Antwort <- factor(extra_long$Antwort)
```

Den Befehl mit **factor** brauchen wir für zum Diagramm erstellen im Folgenden. Dieser Befehl macht aus den Zahlen bei der Variable **Antwort** eine nominale Variable (in R: **factor**) mit Text-Werten “1”, “2” und so weiter. Wozu brauchen wir das? Der Digrammbefehl unten kann nur mit nominalen Variablen Gruppierungen durchführen. Wir werden in dem Diagramm die Anzahl der Antworten darstellen - die Anzahl der Antworten nach Antwort-Gruppe (Gruppe mit Antwort “1” etc.).

Keine Sorge, wenn sich das reichlich ungewöhnlich anhört. Sie müssen es an dieser Stelle nicht erfinden :-)

Man gewöhnt sich daran einerseits; und andererseits ist es vielleicht auch so, dass diese Funktionen nicht perfekt sind, oder nicht aus unserer Sicht oder nur aus Sicht des Menschen, der die Funktion geschrieben hat. Jedenfalls brauchen wir hier eine **factor** Variable zur Gruppierung...

Damit haben wir es schon! Jetzt wird gemalt.

7.3 Diagramme für Anteile

Stellen wir die Anteile der Antworten anhand von farbig gefüllten Balken dar (s. Abbildung 7.1). Beachten Sie, dass die Balken alle auf 1 (100%) skaliert sind; es werden also *relative* Häufigkeiten dargestellt. Absolute Häufigkeiten bleiben hier außen vor.

```
p1 <- ggplot(data = extra_long) +
  aes(x = items) +
  geom_bar(aes(fill = Antwort), position = "fill")

p1
```

Was macht dieser **ggplot** Befehl? Schauen wir es uns in Einzelnen an:

- **ggplot(data = ...)**: Wir sagen “Ich möchte gern die Funktion **ggplot** nutzen, um den Datensatz ... zu plotten”.
- **aes(...)**: Hier definieren wir die “aesthetics” des Diagramms, d.h. alles “Sichtbare”. Wir ordnen in diesem Fall der X-Achse die Variable **items** zu. Per Standardeinstellung geht **ggplot** davon aus, dass sie die Häufigkeiten der X-Werte auf der Y-Achse haben wollen, wenn Sie nichts über die Y-Achse sagen. Jetzt haben wir ein Koordinatensystem definiert (das noch leer ist).

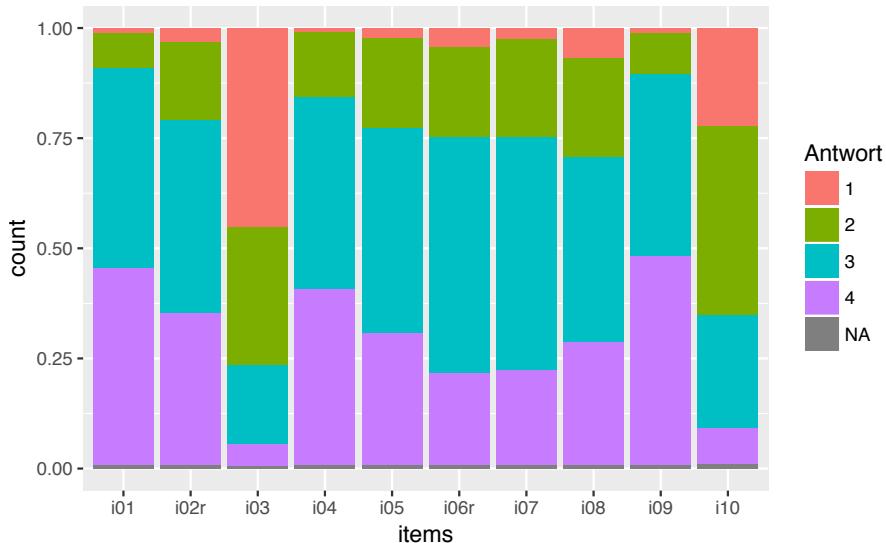


Abbildung 7.1: Relative Häufigkeiten dargestellt anhand von Balkendiagrammen

- `geom_bar()`: “Hey R oder ggplot, jetzt male mal einen barplot in den ansonsten noch leeren plot”.
- `aes(fill = Antwort)`: Genauer gesagt nutzen wir `aes` um einen sichtbaren Aspekte des Diagramms (wie die X-Achse) eine Variable des Datensatzes zuzuordnen. Jetzt sagen wir, dass die Füllung (im Balkendiagramm) durch die Werte von `Antwort` definiert sein sollen (also “1”, “2” etc.).
- `position = "fill"` sagt, dass die Gesamt-Höhe des Balken aufgeteilt werden soll mit den “Teil-Höhen” der Gruppen (Antwort-Kategorien 1 bis 4); wir hätten die Teil-Höhen auch nebeneinander stellen können.

Vielelleicht ist es schöner, die NAs erst zu entfernen.

```
extra_long <- na.omit(extra_long)
```

Plotten Sie das Diagramm dann noch mal:

```
ggplot(data = extra_long) +
  aes(x = items) +
  geom_bar(aes(fill = Antwort), position = "fill")
```

7.4 Rotierte Balkendiagramme

Dazu ergänzen wir die Zeile `+ coord_flip()`; das heißt so viel wie “flippe das Koordinatensystem”.

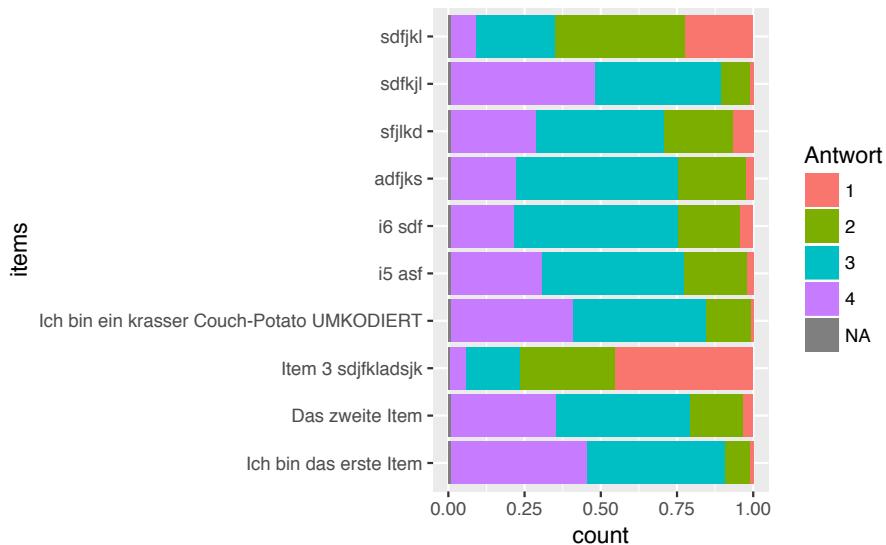


Abbildung 7.2: Rotiertes Balkendiagramm mit Item-Label

```
p1 +
  coord_flip()
```

7.5 Text-Labels für die Items

Wir definieren die Texte ("Labels") für die Items:

```
item_labels <- c("Ich bin das erste Item",
  "Das zweite Item",
  "Item 3 sdjfkladsjk",
  "Ich bin ein krasser Couch-Potato UMKODIERT",
  "i5 asf", "i6 sdf", "adfjks", "sfjlk", "sdfkjl", "sdfjkl")
```

Jetzt hängen wir die Labels an die Items im Diagramm (s. Abbildung 7.2).

```
p1 +
  coord_flip() +
  scale_x_discrete(labels = item_labels)
```

Man kann auch einen Zeilenumbruch in den Item-Labels erzwingen... wobei das führt uns schon recht weit, aber gut, zum Abschluss :-)

```
item_labels <- c("Ich bin das erste Item",
               "Das zweite Item",
               "Item 3 sdjfkladsjk",
               "Ich bin ein krasser \nCouch-Potato***mit Zeilenumbruch***",
               "i5 asf", "i6 sdf", "adfjks", "sfjlkd", "sdfkjl", "sdfjkl")
```

Plotten Sie das dann wieder:

```
ggplot(data = extra_long) +
  aes(x = items) +
  geom_bar(aes(fill = Antwort), position = "fill") +
  coord_flip() +
  scale_x_discrete(labels = item_labels, name = "Extraversionsitems") +
  scale_y_continuous(name = "Anteile")
```

7.6 Diagramm mit Häufigkeiten

Ach so, schön wäre noch die echten Zahlen an der Y-Achse, nicht Anteile. Dafür müssen wir unseren Diagrammtyp ändern, bzw. die Art der Anordnung ändern. Mit `position = "fill"` wird der Anteil (also mit einer Summe von 100%) dargestellt. Wir können auch einfach die Zahlen/Häufigkeiten anzeigen, in dem wir die Kategorien “aufeinander stapeln”. Probieren Sie dazu die folgende Syntax.

```
p2 <- ggplot(data = extra_long) +
  aes(x = items) +
  geom_bar(aes(fill = Antwort), position = "stack") +
  coord_flip() +
  scale_x_discrete(labels = item_labels)
```

p2

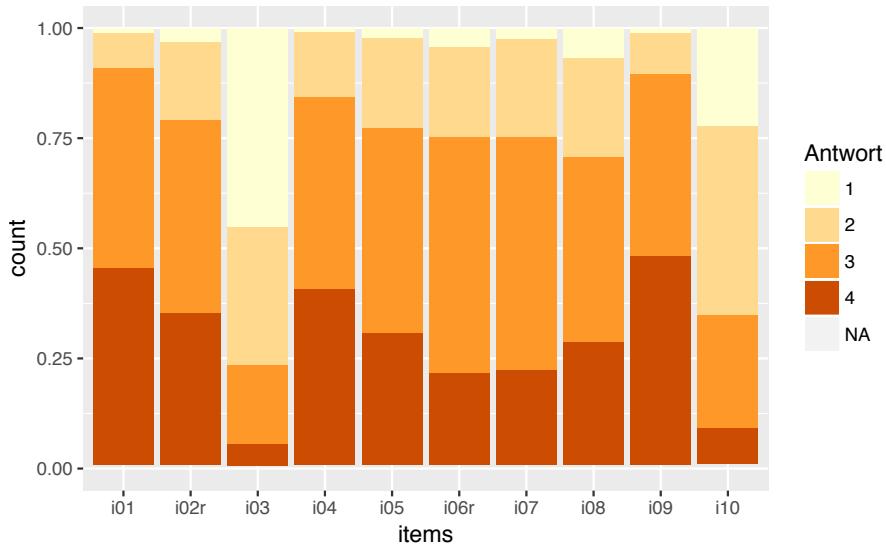
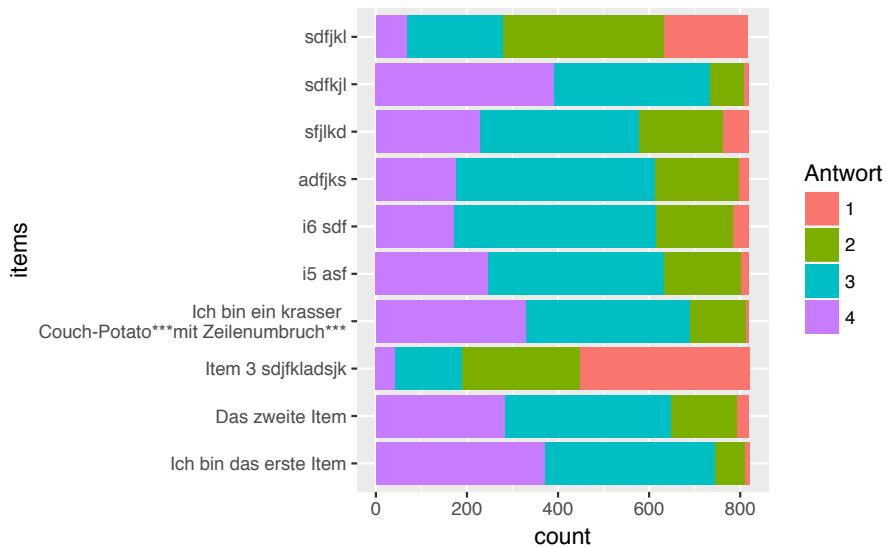


Abbildung 7.3: ... Mit der Brewer-Palette 17



7.7 Farbschemata

Ja, die Wünsche hören nicht auf... Also, noch ein anderes Farbschema (s. Abbildung 7.3).

```
p1 +
  scale_fill_brewer(palette = 17)
```

Das Paket `viridis` hat ein gutes Farbschema. Probieren Sie es mal aus:

```
p1 +  
  scale_fill_viridis(discrete = TRUE)
```

Teil II

Modellieren

Kapitel 8

Grundlagen des Modellierens



Lernziele:

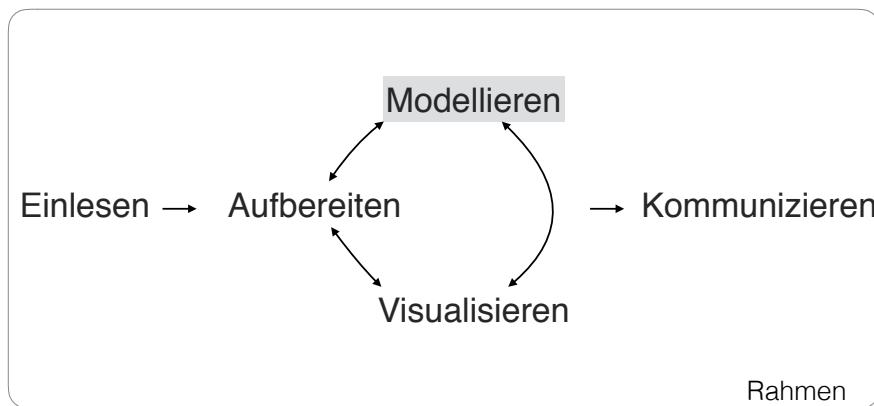
- Erläutern können, was man unter einem Modell versteht.
- Die Ziele des Modellieren aufzählen und erläutern können.
- Die Vor- und Nachteile von einfachen vs. komplexen Modellen vergleichen können.
- Wissen, was man unter “Bias-Varianz-Abwägung” versteht.
- Um die Notwendigkeit von Trainings- und Test-Stichproben wissen.
- Wissen, was man unter Modellgüte versteht.
- Um die Schwierigkeiten der Prädiktorenauswahl wissen.

In diesem Kapitel benötigen wir diese Pakete:



Abbildung 8.1: Ein Modell eines VW-Käfers als Prototyp eines Modells

```
library(tidyverse)
```



8.1 Was ist ein Modell? Was ist Modellieren?

In diesem Kapitel geht es um *Modelle* und *Modellieren*; aber was ist das eigentlich? Seit dem 16. Jahrhundert wird mit dem italienischen Begriff *modelle* ein *verkleinertes Muster*, *Abbild* oder Vorbild für ein Handwerksstück benannt (Gigerenzer 1980). Prototypisch für ein Modell ist - wer hätt's gedacht - ein Modellauto (s. Abb. 8.1; (Spurzem 2017)).

In die Wissenschaft kam der Begriff in der Zeit nach Kant, als man sich klar wurde, dass (physikalische) Theorien nicht die Wirklichkeit als solche zeigen, sondern ein *Modell* davon. Modellieren ist eine grundlegenden Tätigkeit, derer sich Menschen fortlaufend bedienen, um die Welt zu *verstehen*. Denn das Leben ist schwer... oder sagen wir: komplex. Um einen Ausschnitt der Wirklichkeit zu verstehen, erscheint es daher sinnvoll, sich einige als wesentlich erachteten Aspekte "herauszugreifen" bzw. auszusuchen und sich nur noch deren Zusammenspiel näher anzuschauen. Modelle sind häufig vereinfachend: es wird nur ein Ausschnitt der Wirklichkeit in einfacher Form berücksichtigt.

Da wir die Natur bzw. die Wirklichkeit oft nicht komplett erfassen, erschaffen wir uns ein Abbild von der Wirklichkeit, ein Modell.

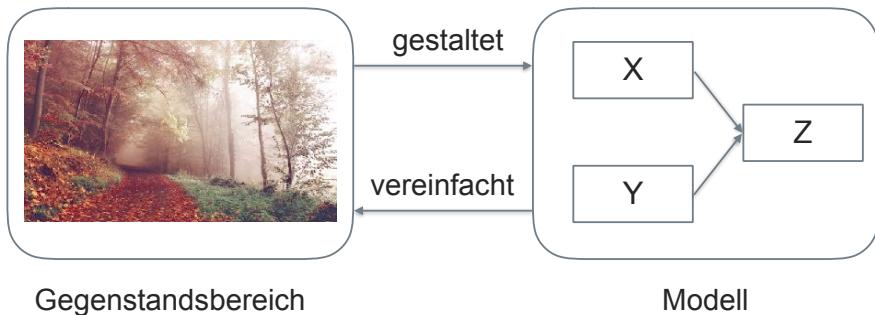


Abbildung 8.2: Modellieren

Manche Aspekte der Wirklichkeit sind wirklicher als andere: Interessiert man sich für den Zusammenhang von Temperatur und Grundwasserspiegel, so sind diese Dinge direkt beobachtbar. Interessiert man sich hingegen für Lebensqualität und Zufriedenheit, so muss man diese Untersuchungsgegenstände erst konstruieren, da Lebensqualität nicht direkt beobachtbar ist. Sprechen wir daher von Wirklichkeit lieber vorsichtiger vom *Gegenstandsbereich*, also den *konstruierten Auszug der Wirklichkeit* für den sich die forschende Person interessiert. Beste-nfalls (er)findet man eine *Annäherung* an die Wirklichkeit, schlechterenfalls eine *verzerrte*, gar *grob falsche* Darstellung. Da keine Wiedergabe der Wirklichkeit perfekt ist, sind streng genommen alle Modelle “falsch” in diesem Sinne.

Gegenstandsbereich und Modelle stehen in einer Beziehung miteinander (vgl. Abb. 8.2, das Foto stammt von Unrau (2017)).

Damit verstehen wir *Modellieren* als eine typische Aktivität von Menschen (Gigerenzer 1980), genauer eines Menschen mit einem bestimmten Ziel. Wir können gar nicht anders, als uns ein Modell unserer Umwelt zu machen; entsprechend kann (und muss man) von *mentalnen Modellen* sprechen. Vielfältige Medien kommen dazu in Frage: Bilder, Geschichten, Logik, Gleichungen. Wir werden uns hier auf eine bestimmte Art formalisierter Modelle, *numerische Modelle*, konzentrieren, weil es dort am einfachsten ist, die Informationen auf präzise Art und Weise herauszuziehen. Allgemein gesprochen ist hier unter Modellieren der Vorgang gemeint, ein Stück Wirklichkeit (“Empirie”) in eine *mathematische Struktur* zu übersetzen.

Wirklichkeit kann dabei als *empirisches System* bezeichnet werden, welches aus einer oder mehr Mengen von Objekten besteht, die zu einander in bestimmten Beziehungen stehen. Ein Beispiel wäre eine Reihe von Personen, die in bestimmten Größe-Relationen zueinander stehen oder eine Reihe von Menschen, bei denen die Füße tendenziell größer werden, je größer die Körpergröße ist.

Mit *mathematische Struktur* ist ein formalisiertes Pendant zum empirischen System gemeint, daher spricht man von einem *numerischen System*. Im Gegensatz zur empirischen System ist das numerische System rein theoretisch, also ausgedacht, nicht empirisch. Auch hier gibt es eine Reihe von Objekten, aber mathematischer Art, also z.B. Zahlen oder Vektoren. Diese mathematischen Objekten stehen wiederum in gewissen Relationen zueinander. Der springende Punkt ist: Im Modell sollen die Beziehungen zwischen den mathematischen Objekten die Beziehungen zwischen den empirischen Objekten widerspiegeln. Was heißt das?



Abbildung 8.3: Formaleres Modell des Modellierens

Stellen wir uns vor, der Klausurerfolg steigt mit der Lernzeit¹. Fragen wir das Modell, welchen Klausurerfolg Alois hat (er hat sehr viel gelernt), so sollte das Modell erwiedern, dass Alois einen hohen Klausurerfolg hat (Modelle geben in diesem Fall gerne eine im Verhältnis große Zahl von sich). Damit würde das Modell korrekt die Empirie widerspiegeln.

Modellieren bedeutet ein Verfahren zu erstellen, welches empirische Sachverhalte adäquat in numerische Sachverhalte umsetzt.

Etwas spitzfindig könnte man behaupten, es gibt keine Modelle - es gibt nur Modelle *von* etwas; dieser Satz soll zeigen, dass zwar ein empirisches System für sich alleine stehen kann, aber ein Modell nicht. Ein Modell verweist immer auf ein empirisches System.

Abb. 8.3 stellt diese formalere Sichtweise des Modellierens dar; das empirische System E wird dem numerischen System Z zugeordnet. Dabei besteht E aus einer Menge von Objekten O sowie einer Menge von Relationen R_E (Relationen meint hier nichts mehr als irgendwelche Beziehungen zwischen den Elementen von O). Analog besteht Z aus einer Menge von numerischen Objekten Z sowie einer Menge von Relationen R_Z (Relationen zwischen den Elementen von Z)².

8.2 Ein Beispiel zum Modellieren in der Datenanalyse

Schauen wir uns ein Beispiel aus der Datenanalyse an; laden Sie dazu zuerst den Datensatz zur Statistikklausur.

Im linken Plot (A; Abb. 8.4) sehen wir - schon übersetzt in eine Datenvisualisierung - den Gegenstandsbereich. Dort sind einige Objekte zusammen mit ihren Relationen abgebildet (Körpergröße und Schuhgröße). Der rechte Plot spezifiziert nun diesen Einfluss: Es wird ein *linearer Zusammenhang* (eine Gerade) zwischen Körpergröße und Schuhgröße unterstellt.

Im rechten Plot (B; Abb. 8.5) sehen wir ein Schema dazu, ein sog. *Pfadmodell*. Noch ist das Modell recht unspezifisch; es wird nur postuliert, dass Körpergröße auf Schuhgröße einen linearen Einfluss habe. Linear heißt hier, dass der Einfluss von Körpergröße auf Schuhgröße immer gleich groß ist, also unabhängig vom Wert der Körpergröße.

Ein etwas aufwändigeres Modell könnte so aussehen (Abb. 8.6):

¹wieder ein typisches Dozentenbeispiel

²Diese Sichtweise des Modellierens basiert auf der Repräsentationstheorie des Messens nach Suppes und Zinnes (1962) zurück; vgl. Gigerenzer (1980)

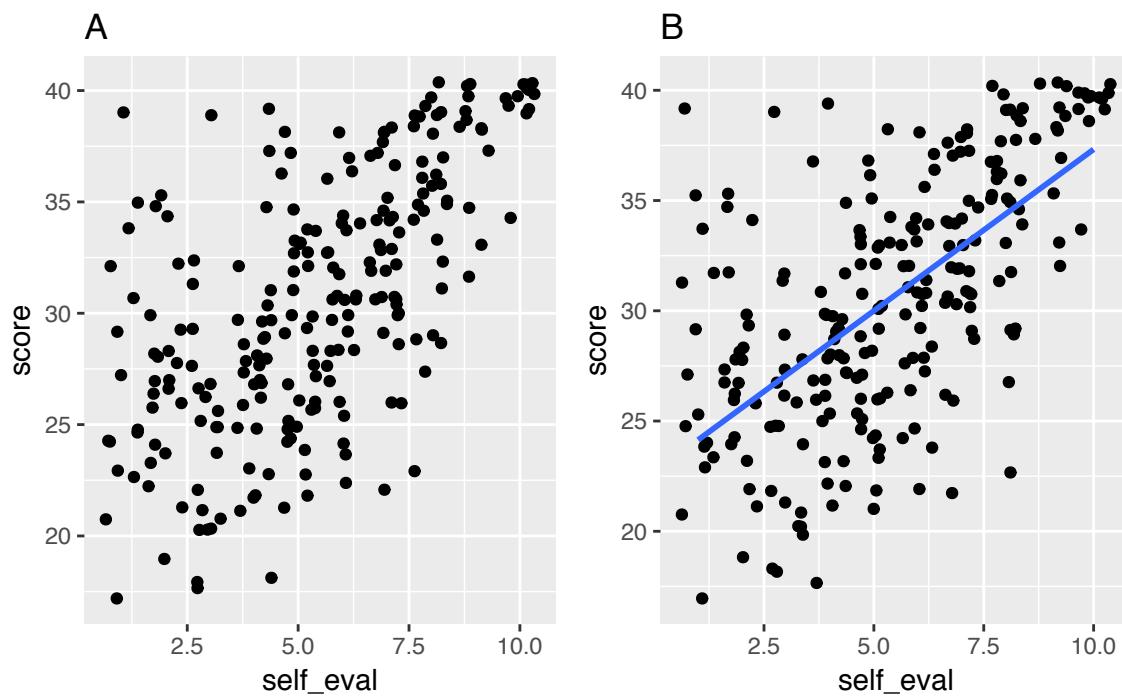


Abbildung 8.4: Ein Beispiel für Modellieren



Abbildung 8.5: Ein Beispiel für ein Pfadmodell

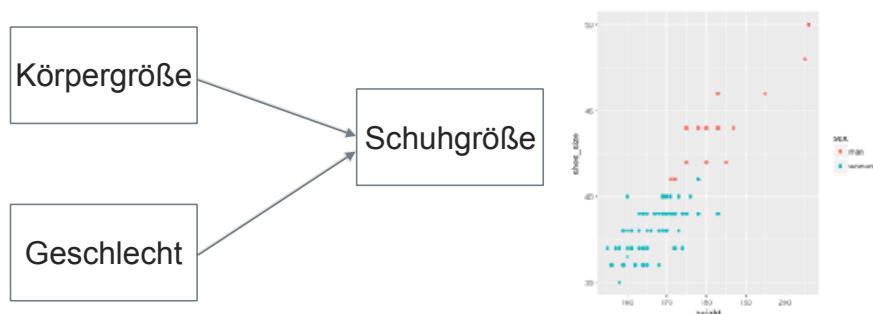


Abbildung 8.6: Ein etwas aufwändigeres Modell

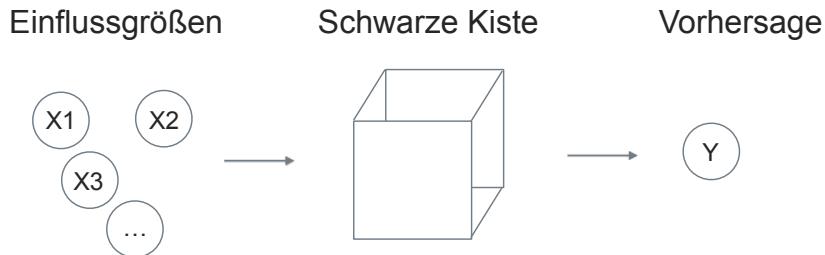


Abbildung 8.7: Modelle mit schwarzer Kiste

Allgemeiner formuliert, haben wir einen oder mehrere *Eingabegrößen* bzw. *Prädiktoren*, von denen wir annehmen, dass sie einen Einfluss haben auf genau eine *Zielgröße* (*Ausgabegröße*) bzw. *Kriterium*.



Einfluss ist hier nicht (notwendig) kausal gemeint, auch wenn es das Wort so vermuten lässt. Stattdessen ist nur ein statistischer Einfluss gemeint; letztlich nichts anderes als ein Zusammenhang. In diesem Sinne könnte man postulieren, dass die Größe des Autos, das man fährt einen “Einfluss” auf das Vermögen des Fahrers habe. Empirisch ist es gut möglich, dass man Belege für dieses Modell findet. Jedoch wird dieser Einfluss nicht kausal sein (man informiere mich, wenn es anders sein sollte).

Modelle, wie wir sie betrachten werden, berechnen eine quantitativen Zusammenhang zwischen diesen beiden Arten von Größen - Prädiktoren und Kriterien. Damit lassen sich unsere Modelle in drei Aspekte gliedern.

Die Einflussgrößen werden in einer “schwarzen Kiste”, die wir hier noch nicht näher benennen, irgendwie verwurstet, will sagen, verrechnet, so dass ein *geschätzter* Wert für das Kriterium, eine *Vorhersage* “hinten bei rauskommt”³. Wir gehen dabei nicht davon aus, dass unsere Modelle perfekt sind, sondern dass Fehler passieren. Mathematischer ausgedrückt:

$$Y = f(X) + \epsilon$$

Hier stehen Y für das Kriterium, X für den oder die Prädiktoren, f für die “schwarze Kiste” und ϵ für den Fehler, den wir bei unserer Vorhersage begehen. Durch den Fehlerterm in der Gleichung ist das Modell *nicht deterministisch*, sondern beinhaltet erstens einen funktionalen Term ($Y = f(x)$) und zweitens einen *stochastischen* Term (ϵ). Die schwarze Kiste könnte man auch als eine *datengenerierende Maschine* oder datengenerierenden Prozess bezeichnen.

Übrigens: Auf das Skalenniveau der Eingabe- bzw. Ausgabegrößen (qualitativ vs. quantitativ) kommt es hier nicht grundsätzlich an; es gibt Modelle für verschiedene Skalenniveaus bzw. Modelle, die recht anspruchslos sind hinsichtlich des Skalenniveaus (sowohl für Eingabe- als auch Ausgabegrößen). Was die Ausgabegröße (das Kriterium) betrifft, so “fühlen” qualitative Variablen von quantitativen Variablen anders an. Ein Beispiel zur Verdeutlichung:

³das ist schließlich entscheidend - frei nach Helmut Kohl

“Gehört Herr Bussi-Ness zur Gruppe der Verweigerer oder der Wichtigmacher?” (qualitatives Kriterium); “Wie hoch ist der Wichtigmacher-Score von Herrn Bussi-Ness?” (quantitatives Kriterium). Ein Modell mit qualitativem Kriterium bezeichnet man auch als *Klassifikation*; ein Modell mit quantitativem Kriterium bezeichnet man auch als *Regression*. Bei letzterem Begriff ist zu beachten, dass er *doppelt* verwendet wird. Neben der gerade genannten Bedeutung steht er auch für ein häufig verwendetes Modell - eigentlich das prototypische Modell - für quantitative Kriterien.

8.3 Taxonomie der Ziele des Modellierens

Modelle kann man auf vielerlei Arten gliedern; für unsere Zwecke ist folgende Taxonomie der Ziele von Modellieren nützlich.

-
- Geleitetes Modellieren
 - Prädiktives Modellieren
 - Explikatives Modellieren
 - Ungeleitetes Modellieren
 - Dimensionsreduzierendes Modellieren
 - Fallreduzierendes Modellieren
-

Betrachten wir diese vier Ziele des Modellierens genauer.

Geleitetes Modellieren ist jede Art des Modellierens, wo die Variablen in Prädiktoren und Kriterien unterteilt werden, z.B. Abb. 8.5. Man könnte diese Modelle einfach darstellen als “X führt zu Y”.

Prädiktives Modellieren könnte man kurz als *Vorhersagen* bezeichnen. Hier ist das Ziel, eine Black Box geschickt zu wählen, so dass der Vohersagefehler möglichst klein ist. Man zielt also darauf ab, möglichst exakte Vorhersagen zu treffen. Sicherlich wird der Vohersagefehler kaum jemals Null sein; aber je präziser, desto besser. Das Innenleben der “schwarzen Kiste” interessiert uns hier *nicht*. Wir machen keine Aussagen über Ursache-Wirkungs-Beziehungen. Ein Beispiel für ein prädiktives Modell: “Je größer das Auto, desto höher das Gehalt”. Dabei werden wir wohl nicht annehmen, dass die Größe des Auto die Ursache für die Höhe des Gehalts ist⁴. Wir sind - in diesem Beispiel - lediglich daran interessiert, das Gehalt möglichst präzise zu schätzen; die Größe des Autos dient uns dabei als Prädiktor, wir verstehen sie nicht als Ursache. Ein altbekanntes Lamento der Statistiklehrer lautet “Korrelation heißt noch nicht Kausation!”. OK.

Explikatives Modellieren oder kurz *Erklären* bedeutet, verstehen zu wollen, *wie* oder *warum* sich ein Kriteriumswert so verändert, wie er es tut. Auf welche Art werden die Prädiktoren verrechnet, so dass eine bestimmter Kriteriumswert resultiert? Welche Prädikatoren sind

⁴bitte mir Bescheid geben, falls ich hier etwas übersehen haben sollte

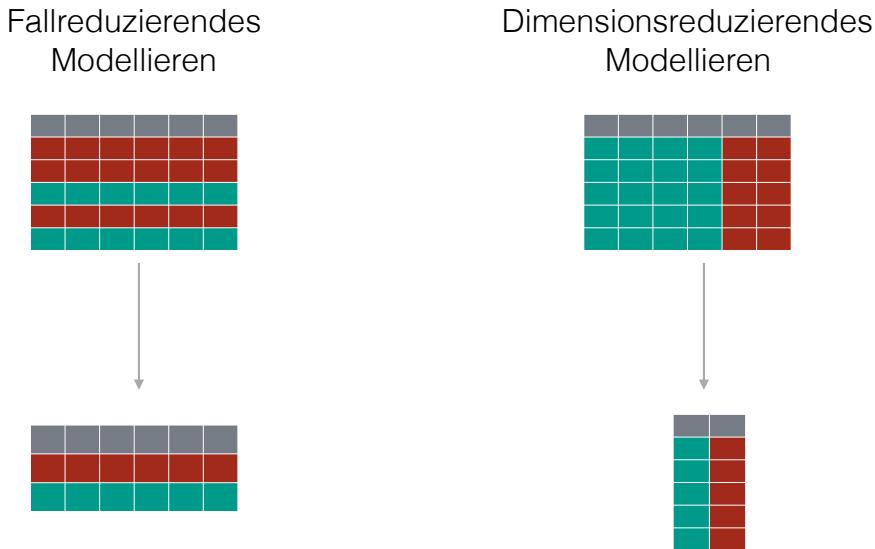


Abbildung 8.8: Die zwei Arten des ungeleiteten Modellierens

dabei (besonders) wichtig? Ist die Art der Verrechnung abhängig von den Werten der Prädiktoren? Hierbei interessiert uns vor allem die *Beschaffenheit* der schwarzen Kiste; die Güte der Vorhersage ist zweitrangig. Oft, aber nicht immer, steht ein Interesse an den Ursache hinter dieser Art der Modellierung. Ursache-Wirkungs-Beziehungen gehören sicherlich zu den interessantesten und wichtigsten Dingen, die man untersuchen kann. Die Wissenschaft ist (bzw. viele Wissenschaftler sind) primär an Fragestellungen zur kausalen Beschaffenheit interessiert (Shmueli 2010). Ein Beispiel für diese Art von Modellierung wäre, ob Achtsamkeit zu weniger intensiven emotionalen Reaktionen führt (Sauer, Walach, und Kohls 2010). Übrigens: Es ist erlaubt, eine kausale Theorie zu vertreten, auch wenn eine Studie solche Schlussfolgerungen nur eingeschränkt oder gar nicht erlaubt (Shmueli 2010). Häufig werden Beobachtungsstudien auf Korrelationsbasis angeführt, um kausale Theorien zu testen. Natürlich ist der Preis für eine einfachere Studie, dass man weniger Evidenz für eine Theorie mit Kausalanspruch einstreichen kann. Aber irgendwo muss man ja anfangen (aber man sollte nicht bei einfachen Studien stehen bleiben).

Vorhersagen und Erklären haben gemein, dass Eingabegrößen genutzt werden, um Aussagen über einen Ausgabegröße zu treffen. Anders gesagt: Es liegt eine Zielgröße mit bekannten Ausprägungen vor, zumindest für eine Reihe von Fällen. Hat man einen Datensatz, so kann man prüfen, *wie gut* das Modell funktioniert, also wie genau man die Ausgabewerte vorhergesagt hat. Das ist also eine Art “Lernen mit Anleitung” oder *angeleitetes Lernen* oder *geleitetes Modellieren* (engl. *supervised learning*). Abbildung 8.7 gibt diesen Fall wieder.

Beim *ungeleiteten Modellieren* entfällt die Unterteilung zwischen Prädiktor und Kriterium. Ungeleitetes Modellieren (*Reduzieren*) meint, dass man die Fülle des Datenmaterials verringert, in dem man ähnliche Dinge zusammenfasst (vgl. Abb. 8.8).

Fasst man Fälle zusammen, so spricht man von *Fallreduzierendem Modellieren*. Zum Beispiel könnte man spektakulärerweise “Britta”, “Carla” und “Dina” zu “Frau” und “Joachim”, “Alois” und “Casper” zu “Mann” zusammen fassen.

Analog spricht man von *Dimensionsreduzierendes Modellieren* wenn Variablen zusammengefasst werden. Hat man z.B. einen Fragebogen zur Mitarbeiterzufriedenheit mit den Items “Mein Chef ist fair”, “Mein Chef ist kompetent”, “Meinem Chef ist meine Karriere wichtig”, so könnte man - wenn die Daten dies unterstützen - die Items zu einer Variable “Zufriedenheit mit Chef” zusammenfassen.

Wenn also das Ziel des Modellieren lautet, die Daten zu reduzieren, also z.B. Kunden nach Persönlichkeit zu gruppieren, so ist die Lage anders als beim geleiteten Modellieren: Es gibt keine Zielgröße. Wir wissen nicht, was die “wahre Kundengruppe” ist, zu der Herrn Casper Bussi-Ness gehört. Wir sagen eher, “OK, die drei Typen sind sich irgendwie ähnlich, sie werden wohl zum selben Typen von Kunden gehören”. Wir tappen (noch mehr) in Dunkeln, was die “Wahrheit” ist im Vergleich zum angeleiteten Modellieren. Unser Modell muss ohne Hinweise darauf, was richtig ist auskommen. Man spricht daher in diesem Fall von *Lernen ohne Anleitung* oder *ungeleitetes Modellieren* (engl. *unsupervised learning*).

8.4 Die vier Schritte des statistischen Modellierens

Modellieren ist in der Datenanalyse bzw. in der Statistik eine zentrale Tätigkeit. Modelliert man in der Statistik, so führt man die zwei folgenden Schritte aus:

1. Man wählt eines der vier Ziele des Modellierens (z.B. ein prädiktives Modell).
2. Man wählt ein Modell aus (genauer: eine Modelfamilie), z.B. postuliert man, dass die Körpergröße einen linearen Einfluss auf die Schuhgröße habe.
3. Man bestimmt (berechnet) die Details des Modells anhand der Daten: Wie groß ist die Steigung der Geraden und wo ist der Achsenabschnitt? Man sagt auch, dass man die *Modellparameter* anhand der Daten schätzt (“Modellinstantiierung” oder “Modellanpassung”, engl. “model fitting”).
4. Dann prüft man, wie gut das Modell zu den Daten passt (Modellgüte, engl. “model fit”); wie gut lässt sich die Schuhgröße anhand der Körpergröße vorhersagen bzw. wie groß ist der Vorhersagefehler?

8.5 Einfache vs. komplexe Modelle: Unter- vs. Überanpassung

Je komplexer ein Modell, desto besser passt sie meistens auf den Gegenstandsbereich. Eine grobe, Holzschnitt artige Theorie ist doch schlechter als eine, die feine Nuancen berücksichtigt, oder nicht? Einiges spricht dafür; aber auch einiges dagegen. Schauen wir uns ein Problem mit komplexen Modellen an.

Der Plot A (links) von Abb. 8.9 zeigt den Datensatz ohne Modell; Plot B legt ein lineares Modell (rote Gerade) in die Daten. Plot C zeigt ein Modell, welches die Daten exakt erklärt -

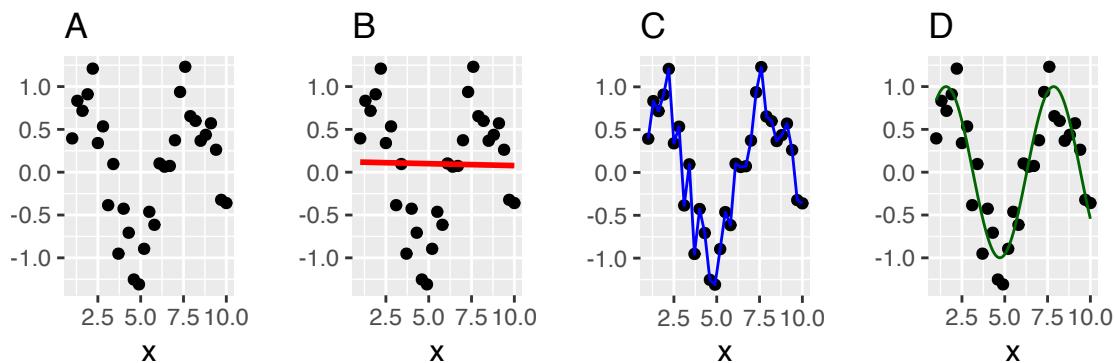


Abbildung 8.9: Welches Modell (Teil B-D; rot, grün, blau) passt am besten zu den Daten (Teil A) ?

die (blaue) Linie geht durch alle Punkte. Der 4. Plot zeigt ein Modell (grüne Linie), welches die Punkte gut beschreibt, aber nicht exakt trifft.

Welchem Modell würden Sie (am meisten) vertrauen? Das “blaue Modell” beschreibt die Daten sehr gut, aber hat das Modell überhaupt eine “Idee” vom Gegenstandsbereich, eine “Ahnung”, wie Y und X zusammenhängen, bzw. wie X einen Einfluss auf Y ausübt? Offenbar nicht. Das Modell ist “übergenau” oder zu komplex. Man spricht von *Überanpassung* (engl. *overfitting*). Das Modell scheint zufälliges, bedeutungsloses Rauschen zu ernst zu nehmen. Das Resultat ist eine zu wackelige Linie - ein schlechtes Modell, da wir wenig Anleitung haben, auf welche Y-Werte wir tippen müssten, wenn wir neue, unbekannte X-Werte bekämen.

Beschreibt ein Modell (wie das blaue Modell hier) eine Stichprobe sehr gut, heißt das noch *nicht*, dass es auch zukünftige (und vergleichbare) Stichproben gut beschreiben wird. Die Güte (Vorhersagegenauigkeit) eines Modells sollte sich daher stets auf eine neue Stichprobe beziehen (Test-Stichprobe), die nicht in der Stichprobe beim Anpassen des Modells (Trainings-Stichprobe) enthalten war.

Was das “blaue Modell” zu detailverliebt ist, ist das “rote Modell” zu simpel. Die Gerade beschreibt die Y-Werte nur sehr schlecht. Man hätte gleich den Mittelwert von Y als Schätzwert für jedes einzelne Y_i hernehmen können. Dieses lineare Modell ist *unterangepasst*, könnte man sagen (engl. *underfitting*). Auch dieses Modell wird uns wenig helfen können, wenn es darum geht, zukünftige Y-Werte vorherzusagen (gegeben jeweils einen bestimmten X-Wert).

Ah! Das *grüne Modell* scheint das Wesentliche, die “Essenz” der “Punktebewegung” zu erfassen. Nicht die Details, die kleinen Abweichungen, aber die “große Linie” scheint gut getroffen. Dieses Modell erscheint geeignet, zukünftige Werte gut zu beschreiben. Das grüne Modell ist damit ein Kompromiss aus Einfachheit und Komplexität und würde besonders passen, wenn es darum gehen sollte, zyklische Veränderungen zu erklären⁵.

Je komplexer ein Modell ist, desto besser beschreibt es einen bekannten Datensatz (Trainings-Stichprobe). Allerdings ist das Modell, welches den Trainings-Datensatz am besten beschreibt, nicht zwangsläufig das Modell, welches neue, unbekannte

⁵Tatsächlich wurden die Y-Werte als Sinus-Funktion plus etwas normalverteiltes Rauschen simuliert.

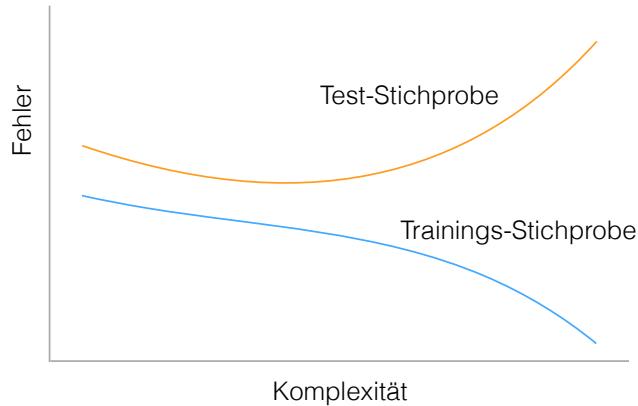


Abbildung 8.10: 'Mittlere' Komplexität hat die beste Vorhersagegenauigkeit (am wenigsten Fehler) in der Test-Stichprobe

Daten am besten beschreibt. Oft im Gegenteil!

Je komplexer das Modell, desto kleiner der Fehler im *Trainings*-Datensatz. Allerdings: Die Fehler-Kurve im *Test*-Datensatz ist *U-förmig*: Mit steigender Komplexität wird der Fehler einige Zeit lang kleiner; ab einer gewissen Komplexität steigt der Fehler im *Test*-Datensatz wieder (vgl. Abb. 8.10)! Eine 'mittlere' Komplexität ist daher am besten; die Frage ist nur, wie viel 'mittel' ist.

8.6 Bias-Varianz-Abwägung

Einfache Modelle bilden (oft) verfehlten oft wesentliche Aspekte des Gegenstandsbereich; die Wirklichkeit ist häufig zu komplex für einfache Modelle. Die resultierende *Verzerrung* in den vorhergesagten Werten nennt man auch *Bias*. Mit anderen Worten: ist ein Modell zu einfach, passt es zu wenig zu den Daten (engl. *underfitting*). Auf der anderen Seite ist das Modell aber *robust* in dem Sinne, dass sich die vorhergesagten Werte kaum ändern, falls sich der Trainings-Datensatz etwas ändert.

Ist das Modell aber zu reichhaltig ("komplex"), bildet es alle Details des Trainings-Datensatzes ab, wird es auch zufällige Variation des Datensatzes vorhersagen; Variation, die nicht relevant ist, der nichts Eigentliches abbildet. Das Modell ist "überangepasst" (engl. *overfitting*); geringfügige Änderungen im Datensatz können das Modell stark verändern. Das Modell ist nicht robust. Auf der positiven Seite werden die Nuancen der Daten gut abgebildet; der Bias ist gering bzw. tendenziell geringer als bei einfachen Modellen.

Einfache Modelle: Viel Bias, wenig Varianz. Komplexe Modelle: Wenig Bias, viel Varianz.

Dieser Sachverhalt ist in folgendem Diagramm dargestellt (vgl. Abb. 8.11; vgl. Kuhn & Johnson (2013)).

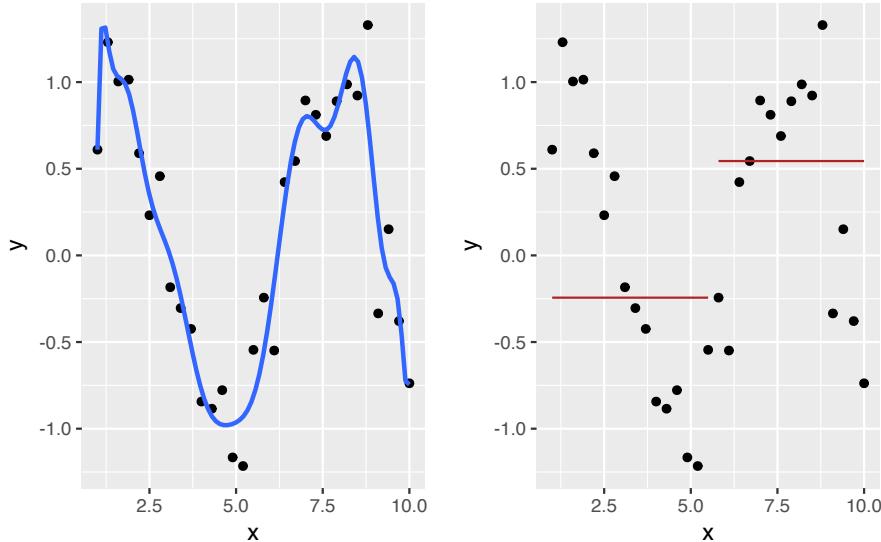


Abbildung 8.11: Der Spagat zwischen Verzerrung und Varianz

Der linke Plot zeigt ein komplexes Modell⁶; das Modell (blaue Linie) erscheint “zitterig”, kleine Änderungen in den Daten können große Auswirkungen auf das Modell (Verlauf der blauen Linie) haben. Darüber hinaus sind einige Details des Modells unplausibel: es gibt viele kleine “Hügel”, die nicht augenscheinlich plausibel sind.

Der Plot auf der rechten Seiten hingegen ist sehr einfach und robust. Änderungen in den Daten werden vergleichsweise wenig Einfluss auf das Modell (die beiden roten Linien) haben.

8.7 Training- vs. Test-Stichprobe

Wie wir gerade gesehen haben, kann man *immer* ein Modell finden, welches die *vorhandenen* Daten sehr gut beschreibt. Das gleicht der Tatsache, dass man im Nachhinein (also bei vorhandenen Daten) leicht eine Erklärung findet. Ob diese Erklärung sich in der Zukunft, bei unbekannten Daten bewahrheitet, steht auf einem ganz anderen Blatt.

Daher sollte man *immer* sein Modell an einer Stichprobe *entwickeln* (“trainieren” oder “üben”) und an einer zweiten Stichprobe *testen*. Die erste Stichprobe nennt man auch *training sample* (Trainings-Stichprobe) und die zweite *test sample* (Test-Stichprobe). Entscheidend ist, dass das Test-Sample beim Entwickeln des Modells unbekannt war bzw. nicht verwendet wurde.

Die Güte des Modells sollte nur anhand eines - bislang nicht verwendeten - Test-Samples überprüft werden. Das Test-Sample darf bis zur Modellüberprüfung nicht analysiert werden.

Die Modellgüte ist im Trainings-Sample meist deutlich besser als im Test-Sample (vgl. die Fallstudie dazu: 10.8).

⁶Genauer gesagt ein Polynom von Grad 5.

Zum Aufteilen verfügbarer Daten in eine Trainings- und eine Test-Stichprobe gibt es mehrere Wege. Einer sieht so aus:

```
train <- slice(stats_test, 1:200)
test <- slice(stats_test, 201:306)
```

`dplyr::slice` schneidet eine ‘Scheibe’ aus einem Datensatz. Mit `anti_join` kann man einen generellen Ansatz wählen: ‘`anti_join`’ vereinigt die *nicht-gleichen* Zeilen zweier Datensätze. Die Gleichheit wird geprüft anhand aller Spalten mit identischem Namen in beiden Datensätzen.

```
train <- stats_test %>%
  sample_frac(.8, replace = FALSE) # Stichprobe von 80%, ohne Zurücklegen

test <- stats_test %>%
  anti_join(train) # Alle Zeilen von "stats_test", die nicht in "train" vorkommen
```

Damit haben wir ein Trainings-Sample (`train`), in dem wir ein oder besser mehrere Modelle entwickeln können.

So schön wie dieses Vorgehen auch ist, es ist nicht perfekt. Ein Nachteil ist, dass unsere Modellgüte wohl *anders* wäre, hätten wir andere Fälle im Test-Sample erwischt. Würden wir also ein neues Trainings-Sample und ein neues Test-Sample aus diesen Datensatz ziehen, so hätten wir wohl andere Ergebnisse. Was wenn diese Ergebnisse nun deutlich von den ersten abweichen? Dann würde unser Vertrauen in die die Modellgüte sinken. Wir bräuchten also noch ein Verfahren, welches *Variabilität* in der Modellgüte widerspiegelt.

8.8 Wann welches Modell?

Tja, mit dieser Frage lässt sich ein Gutteil des Kopfzerbrechens in diesem Metier erfassen. Die einfache Antwort lautet: Es gibt kein “bestes Modell”, aber es mag für *einen bestimmten Gegenstandsbereich, in einem bestimmten (historisch-kulturellen) Kontext, für ein bestimmtes Ziel* und mit *einer bestimmten Stichprobe* ein best mögliches Modell geben. Dazu einige Eckpfeiler:

- Unter sonst gleichen Umständen sind einfachere Modelle den komplexeren vorzuziehen. Gott sei Dank.
- Je nach Ziel der Modellierung ist ein erklärendes Modell oder ein Modell mit reinem Vorhersage-Charakter vorzuziehen.
- Man sollte stets mehrere Modelle vergleichen, um abzuschätzen, welches Modell in der aktuellen Situation geeigneter ist.

8.9 Modellgüte

Wie “gut” ist mein Modell? Modelle bewerten bzw. vergleichend bewerten ist einer der wichtigsten Aufgaben beim Modellieren. Die Frage der Modellgüte hat viele feine technisch-statistische Verästelungen, aber einige wesentlichen Aspekte kann man einfach zusammenfassen.

Kriterium der theoretischen Plausibilität: Ein statistisches Modell sollte theoretisch plausibel sein.

Anstelle “alles mit allem” durchzuprobieren, sollte man sich auf Modelle konzentrieren, die theoretisch plausibel sind. Die Modellwahl ist theoretisch zu begründen.

Kriterium der guten Vorhersage: Die Vorhersagen eines Modells sollen präzise und überraschend sein.

Dass ein Modell die Wirklichkeit präzise vorhersagen soll, liegt auf der Hand. Hier verdient nur der Term *vorhersagen* Beachtung. Es ist einfach, im Nachhinein Fakten (Daten) zu erklären. Jede Nachbesprechung eines Bundesliga-Spiels liefert reichlich Gelegenheit, *posthoc* Erklärungen zu hören. Schwieriger sind Vorhersagen⁷. Die Modellgüte ist also idealerweise an *in der Zukunft liegende* Ereignisse bzw. deren Vorhersage zu messen. Zur Not kann man auch schon in der Vergangenheit angefallene Daten hernehmen. Dann müssen diese Daten aber *für das Modell* neu sein.

Was ist mit überraschend gemeint? Eine Vorhersage, dass die Temperatur morgen in Nürnberg zwischen -30 und +40°C liegen wird, ist sicherlich sehr treffend, aber nicht unbedingt präzise und nicht wirklich überraschend. Die Vorhersage, dass der nächste Chef der Maurer-Innung (wenn es diese geben sollte) ein Mann sein wird, und nicht eine Frau, kann zwar präzise sein, ist aber nicht überraschend. Wir werden also in dem Maße unseren Hut vor dem Modell ziehen, wenn die Vorhersagen sowohl präzise als auch überraschen sind. Dazu später mehr Details.

Kriterium der Situationsangemessenheit: Die Güte des Modells ist auf die konkrete Situation abzustellen.

Ein Klassifikationsmodell muss anders beurteilt werden als ein Regressionsmodell. Reduktionsmodelle müssen wiederum anders beurteilt werden. In den entsprechenden Kapiteln werden diese Unterschiede präzisiert.

8.10 Auswahl von Prädiktoren

Wie oben diskutiert, stellen wir ein (geleitetes) Modell gerne als ein Pfaddiagramm des Typs $X \rightarrow Y$ dar (wobei X ein Vektor sein kann). Nehmen wir an das Kriterium Y als gesetzt an; bleibt die Frage: Welche Prädiktoren (X) wählen wir, um das Kriterium möglichst gut vorherzusagen?

⁷Gerade wenn sie die Zukunft betreffen; ein Bonmot, das Yogi Berra nachgesagt wird.

Eine einfache Frage. Keine leichte Antwort. Es gibt zumindest drei Möglichkeiten, die Prädiktoren zu bestimmen: theoriegeleitet, datengetrieben oder auf gut Glück.

- theoriegeleitet: Eine starke Theorie macht präzise Aussagen, welche Faktoren eine Rolle spielen und welche nicht. Auf dieser Basis wählen wir die Prädiktoren. Diese Situation ist wünschenswert; nicht nur, weil Sie Ihnen das Leben leicht macht, sondern weil es nicht die Gefahr gibt, die Daten zu “overfitten”, “Rauschen als Muster” zu bewerten - kurz: zu optimistisch bei der Interpretation von Statistiken zu sein.
- datengetrieben: Kurz gesagt werden die Prädiktoren ausgewählt, welche das Kriterium am besten vorhersagen. Das ist einerseits stimmig, andererseits birgt es die Gefahr, dass Zufälligkeiten in den Daten für echte Strukturen, die sich auch in zukünftigen Stichproben finden würden, missverstanden werden.
- auf gut Glück: tja, kann man keine Theorie zu Rate ziehen und sind die Daten wenig aussagekräftig oder man nicht willens ist, sie nicht genug zu quälen analysieren, so neigen Menschen dazu, zuerst sich selbst und dann andere von der Plausibilität der Entscheidung zu überzeugen. Keine sehr gute Strategie.

In späteren Kapiteln betrachten wir Wege, um Prädiktoren für bestimmte Modelle auszuwählen.

8.11 Aufgaben

1. Erfolg beim Online-Dating

Lesen Sie diesen⁸ Artikel (Sauer und Wolff 2016). Zeichnen Sie ein Pfaddiagramm zum Modell!⁹.

2. Ziele des Modellierens

Welche drei Ziele des Modellierens kann man unterscheiden?¹⁰

3. Bias-Varianz-Abwägung

Betrachten Sie Abb. 8.12. Welches der beiden Modelle (visualisiert im jeweiligen Plot) ist wahrscheinlich... .

- mehr bzw. weniger robust gegenüber Änderungen im Datensatz?
- mehr oder weniger präzise?

4. Richtig oder falsch?¹¹

⁸https://thewinnow.com/papers///5202-the-effect-of-a-status-symbol-on-success-in-online-dating-/an-experimental-study-data-paper?review_it=true

⁹Status → Erfolg beim Online-Dating

¹⁰8.3

¹¹R, F, F, F, R

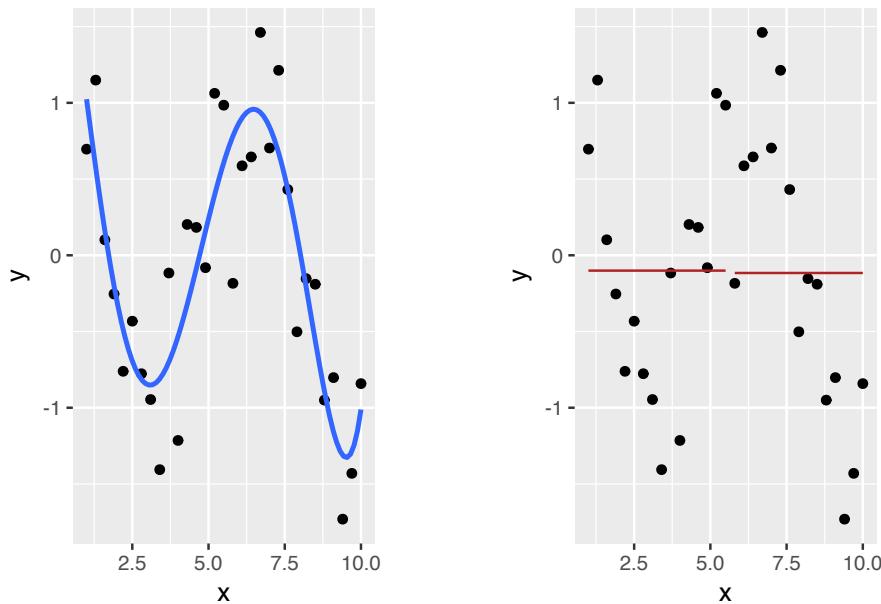


Abbildung 8.12: Bias-Varianz-Abwägung. Links: Wenig Bias, viel Varianz. Rechts: Viel Bias, wenig Varianz.

Tabelle 8.1: Befehle des Kapitels 'Modellieren'

Paket..Funktion	Beschreibung
dplyr::sample_frac	Zieht eine Stichprobe von x% aus einem Dataframe
dplyr::anti_join	Behält alle Zeilen von df1, die *nicht* in df2 vorkommen
dplyr::slice	Schneidet eine 'Scheibe' aus einem Datensatz



Richtig oder Falsch!?

1. Die Aussage "Pro Kilo Schoki steigt der Hüftumfang um einen Zentimeter" kann als Beispiel für ein deterministisches Modell herhalten.
2. Gruppert man Kunden nach ähnlichen Kaufprofilen, so ist man insofern an "Reduzieren" der Datenmenge interessiert.
3. Grundsätzlich gilt: Je komplexer ein Modell, desto besser.
4. Mit "Bias" ist gemeint, dass ein Modell "zittrig" oder "wackelig" ist - sich also bei geringer Änderung der Stichprobendaten massiv in den Vorhersagen ändert.
5. In der Gleichung $Y = f(x) + \epsilon$ steht ϵ für den Teil der Kriteriums, der nicht durch das Modell erklärt wird.

8.12 Befehlsübersicht

Tabelle 8.1 fasst die R-Funktionen dieses Kapitels zusammen.

8.13 Verweise

- Einige Ansatzpunkte zu moderner Statistik (“Data Science”) finden sich bei Peng und Matsui (2015).
- Chester Ismay erläutert einige Grundlagen von R und RStudio, die für Modellierung hilfreich sind: <https://bookdown.org/chesterismay/rbasics/>.
- Eine klassische und sehr gute Einführung findet sich bei James, Witten, Hastie & Tibshirani (James, Witten, Hastie, und Tibshirani 2013b). Dieses Buch bietet ein frei verfügbares PDF¹².

¹²<http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Sixth%20Printing.pdf>

Kapitel 9

Der p-Wert, Inferenzstatistik und Alternativen



Lernziele:

- Den p-Wert erläutern können.
- Den p-Wert kritisieren können.
- Alternativen zum p-Wert kennen.
- Inferenzstatistische Verfahren für häufige Fragestellungen kennen.

In diesem Kapitel werden folgende Pakete benötigt:

```
library(pwr) # Powerberechnung
library(compute.es) # Effektstärken
library(tidyverse) # Datenjudo
library(broom) # Anova-Ergebnisse aufräumen
library(BayesFactor) # Bayes-Faktor berechnen
```

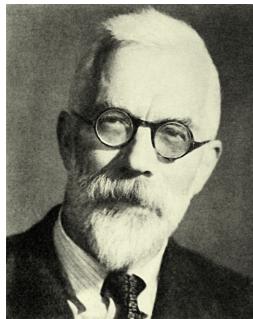


Abbildung 9.1: Der größte Statistiker des 20. Jahrhunderts ($p < .05$)

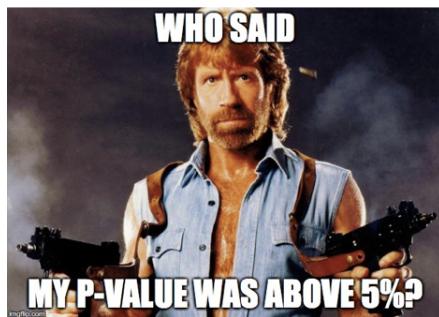


Abbildung 9.2: Der p-Wert wird oft als wichtig erachtet

9.1 Der p-Wert sagt nicht das, was viele denken

Der p-Wert, entwickelt von Sir Ronald Fisher (Abb. 9.1), ist die heilige Kuh der Forschenden. Das ist nicht normativ, sondern deskriptiv gemeint. Der p-Wert entscheidet (häufig) darüber, was publiziert wird, und damit, was als Wissenschaft sichtbar ist - und damit, was Wissenschaft ist (wiederum deskriptiv, nicht normativ gemeint). Kurz: Dem p-Wert kommt viel Bedeutung zu bzw. ihm wird viel Bedeutung zugemessen (vgl. Abb. 9.2).

Der p-Wert ist der tragende Ziegelstein in einem Theoriegebäude, das als *Nullhypothesen-Signifikanztesten* (NHST¹) bezeichnet wird. Oder kurz als ‘Inferenzstatistik’ bezeichnet. Was sagt uns der p-Wert? Eine gute intuitive Definition ist:

Der p-Wert sagt, wie gut die Daten zur Nullhypothese passen.

Die (genaue) Definition des p-Werts ist kompliziert; man kann sie leicht missverstehen:

Der p-Wert - $P(D|H)$ - gibt die Wahrscheinlichkeit P unserer Daten D an (und noch extremerer), unter der Annahme, dass die getestete Hypothese H wahr ist (und wenn wir den Versuch unendlich oft wiederholen würden, unter identischen Bedingungen und ansonsten zufällig).

Mit anderen Worten: Je *größer p*, desto *besser* passen die Daten zur *Nullhypothese*. Mit Nullhypothese (H_0) bezeichnet man die getestete Hypothese. Der Name Nullhypothese röhrt

¹Der Term ‘Signifikanz-Hypothesen-Inferenz-Testen’ hat sich nicht durchgesetzt

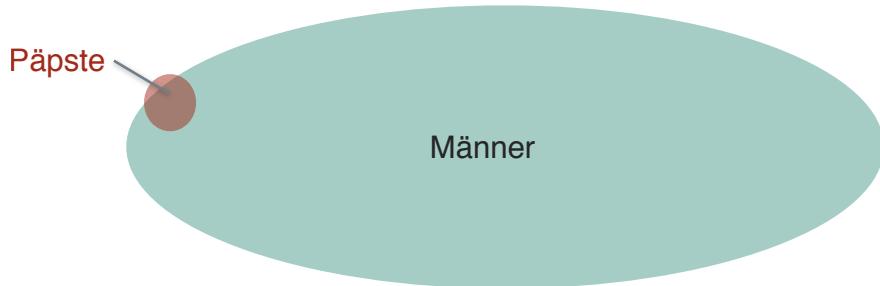


Abbildung 9.3: Mann und Papst zu sein ist nicht das gleiche.

vom Begriff ‘nullifizieren’ (verwerfen) her, da (nach dem Falsifikationismus) eine These immer nur verworfen, nie bestätigt werden kann. Da viele die eigene Hypothese nur ungern verwerfen wollen, wird die ‘gegnerische Hypothese’, die man loswerden will, getestet. Fällt p unter die magische Zahl von 5%, so proklamiert man Erfolg (*Signifikanz*) und verwirft die H_0 .

Der p -Wert ist weit verbreitet. Er bietet die Möglichkeit, relativ objektiv zu quantifizieren, wie gut ein Kennwert, mindestens so extrem wie der aktuell vorliegende, zu einer Hypothese passt. Allerdings hat der p -Wert seine Probleme. Vor allem: Er wird missverstanden. Jetzt kann man sagen, dass es dem p -Wert (dem armen) nicht anzulasten, dass andere/ einige ihn missverstehen. Auf der anderen Seite finde ich, dass sich Technologien dem Nutzer anpassen sollten (soweit als möglich) und nicht umgekehrt.

Viele Menschen - inkl. Professoren und Statistik-Dozenten - haben Probleme mit dieser Definition (Gigerenzer 2004). Das ist nicht deren Schuld: Die Definition ist kompliziert. Vielleicht denken viele, der p -Wert sage das, was tatsächlich interessant ist: die Wahrscheinlichkeit der (getesteten) Hypothese H , gegeben der Tatsache, dass bestimmte Daten D vorliegen. Leider ist das *nicht* die Definition des p -Werts. Also:

$$P(D|H) \neq P(H|D)$$

9.1.1 Von Männern und Päpsten

Formeln haben die merkwürdige Angewohnheit vor dem inneren Auge zu verschwinden; Bilder sind für viele Menschen klarer, scheint's. Übersetzen wir die obige Formel in folgenden Satz:

Wahrscheinlichkeit, Mann zu sein, wenn man Papst ist UNGLEICH zur Wahrscheinlichkeit, Papst zu sein, wenn man Mann ist.

Oder kürzer:

$$P(M|P) \neq P(P|M)$$

Das Bild (Abb. 9.3) zeigt den Anteil der Männer an den Päpsten (sehr hoch). Und es zeigt

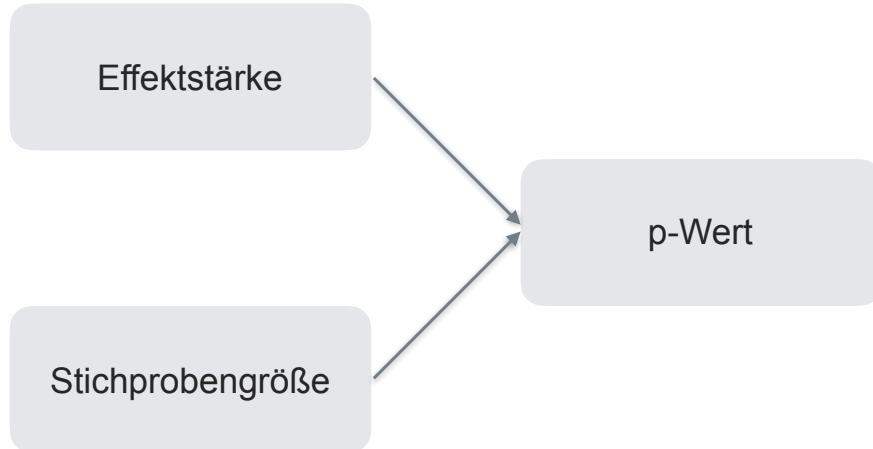


Abbildung 9.4: Zwei Haupeinflüsse auf den p-Wert

den Anteil der Päpsten von allen Männern (sehr gering). Dabei können wir uns Anteil mit Wahrscheinlichkeit übersetzen. Kurz: Die beiden Anteile (Wahrscheinlichkeiten) sind nicht gleich. Man denkt leicht, der p-Wert sei die *Wahrscheinlichkeit, Papst zu sein, wenn man Mann ist*. Das ist falsch. Der p-Wert ist die *Wahrscheinlichkeit, Papst zu sein, wenn man Mann ist*. Ein großer Unterschied.

9.2 Der p-Wert ist eine Funktion der Stichprobengröße

Der p-Wert ist für weitere Dinge kritisiert worden (Wagenmakers (2007), Briggs (2016)); z.B. dass die “5%-Hürde” einen zu schwachen Test für die getestete Hypothese bedeutet. Letzterer Kritikpunkt ist aber nicht dem p-Wert anzulasten, denn dieses Kriterium ist beliebig, könnte konservativer gesetzt werden und jegliche mechanisierte Entscheidungsmethode kann ausgenutzt werden. Ähnliches kann man zum Thema “P-Hacking” argumentieren (Head u. a. (2015), Wicherts u. a. (2016)): andere statistische Verfahren können auch gehackt werden. “Hacken” soll hier sagen, dass man - Kreativität und Wille vorausgesetzt - immer Wege finden kann, um einen Kennwert in die gewünschte Richtung zu drängen.

Ein anderer Anklagepunkt lautet, dass der p-Wert nicht nur eine Funktion der Effektgröße sei, sondern auch der Stichprobengröße. Sprich: Bei großen Stichproben wird jede Hypothese signifikant. Das ist richtig. Das schränkt die praktische Nützlichkeit ein (vgl. Abb. 9.4. Die Details der Simulation, die hinter Abb. 9.4 sind etwas umfangreicher und hier nicht so wichtig, daher nicht angegeben².

Egal wie klein die Effektstärke ist, es existiert eine Stichprobengröße, die diesen Effekt beliebig signifikant werden lässt.

Die Verteidigung argumentiert hier, dass das “kein Bug, sondern ein Feature” sei: Wenn man z.B. die Hypothese prüfe, dass der Gewichtsunterschied zwischen Männern und Frauen

²s. hier für Details: https://sebastiansauer.github.io/pvalue_sample_size/

0,00000000kg sei und man findet 0,000000123kg Unterschied, ist die getestete Hypothese falsch. Punkt. Der p-Wert gibt demnach das korrekte Ergebnis. Meiner Ansicht nach ist die Antwort zwar richtig, geht aber an den Anforderungen der Praxis vorbei.

9.3 Mythen zum p-Wert

Falsche Lehrmeinungen sterben erst aus, wenn die beteiligten Professoren in Rente gehen, heißt es. Jedenfalls halten sich eine Reihe von Mythen hartnäckig; sie sind alle falsch.

Wenn der p-Wert kleiner als 5% ist, dann ist meine Hypothese (H_1) sicher richtig.

Falsch. Richtig ist: "Wenn der p-Wert kleiner ist als 5% (oder allgemeiner: kleiner als α , dann sind die Daten (oder noch extremer) unwahrscheinlich, vorausgesetzt die H_0 gilt".

Wenn der p-Wert kleiner als 5% ist, dann ist meine Hypothese (H_1) höchstwahrscheinlich richtig.

Falsch. Richtig ist: Wenn der p-Wert kleiner ist als α , dann sind die Daten unwahrscheinlich, falls die H_0 gilt. Ansonsten (wenn H_0 nicht gilt) können die Daten sehr wahrscheinlich sein.

Wenn der p-Wert kleiner als 5% ist, dann ist die Wahrscheinlichkeit der H_0 kleiner als 5%.

Falsch. Der p-Wert gibt *nicht* die Wahrscheinlichkeit einer Hypothese an. Richtig ist: Ist der p-Wert kleiner als 5%, dann sind meine Daten (oder noch extremer) unwahrscheinlich (<5%), wenn die H_0 gilt.

Wenn der p-Wert kleiner als 5% ist, dann habe ich die Ursache eines Phänomens gefunden.

Falsch. Richtig ist: Keine Statistik kann für sich genommen eine Ursache erkennen. Bestenfalls kann man sagen: hat man alle konkurrierenden Ursachen ausgeschlossen *und* sprechen die Daten für die Ursache *und* sind die Daten eine plausible Erklärung, so erscheint es der beste Schluss, anzunehmen, dass man *eine* Ursache gefunden hat - im Rahmen des Geltungsbereichs einer Studie.

Wenn der p-Wert kleiner als 5% ist, dann kann ich meine Studie veröffentlichen.

Richtig. Leider entscheidet zu oft (nur) der p-Wert über das Wohl und Wehe einer Studie. Wichtiger wäre zu prüfen, wie "gut" das Modell ist - wie präzise sind die Vorhersagen? Wie theoretisch befriedigend ist das Modell?

Wenn der p-Wert *größer* als 5% ist, dann ist das ein Beleg *für* die H_0 .

Falsch. Richtig ist: Ein großer p-Wert ist ein Beleg, dass die Daten plausibel unter der H_0 sind. Wenn es draußen regnet, ist es plausibel, dass es Herbst ist. Das heißt aber nicht, dass andere Hypothesen nicht auch plausibel sind. Ein großer p-Wert ist also Abwesenheit von klarer Evidenz – *nicht* Anwesenheit von klarer Evidenz zugunsten der H_0 . Schöner ausgedrückt:

“No evidence of effect ist not the same as evidence of no effect”. Für die Wissenschaft ist das insofern ein großes Problem, als dass sich Zeitschriften weigern, nicht-signifikante Studien aufzunehmen: “Das ist eine unklare Befundlage. Kein Mehrwert.” so die Haltung. Das führt dazu, dass die wissenschaftliche Literatur einer großen Verzerrung unterworfen ist.

9.4 Wann welcher Inferenztest?

In der Praxis ist es eine häufige Frage, wann man welchen statistischen Test verwenden soll. Bei Eid, Gollwitzer, und Schmitt (2010) findet man eine umfangreiche Tabelle dazu; auch online wird man schnell fündig (z.B. bei der Methodenberatung der Uni Zürich³ oder beim Ärzteblatt⁴, Prel u. a. (2010)).

Die folgende Auflistung gibt einen *kurzen* Überblick zu gebräuchlichen Verfahren. Entscheidungskriterium ist hier (etwas vereinfacht) das Skalenniveau der Variablen (unterschieden in Input- und Outputvariablen).

1. 2 nominale Variablen: χ^2 -Test - `chisq.test`
2. Output: 1 metrisch, Input: 1 dichotom: t-Test - `t.test`
3. Output: 1 oder mehr metrisch, 1 nominal: Varianzanalyse - `aov`
4. 2 metrische Variablen: Korrelation - `cor.test`
5. Output: 1 metrisch, Input: 1 oder mehr nominal oder metrisch: Regression - `lm`
6. Output: 1 ordinal, Input: 1 dichotom: Wilcoxon (Mann-Whitney-U-Test) - `wilcox.test`
7. Output: 1 ordinal, Input: 1 nominal: Kruskal-Wallis-Test - `kruskal.test`
8. 1 metrisch (Test auf Normalverteilung): Shapiro-Wilk-Test - `shapiro.test`
9. Output: 1 dichotom, Input 1 oder mehr nominal oder metrisch: logistische (klassifikatorische) Regression: `glm(..., family = "binomial")`
10. 2 ordinal: Spearmans Rangkorrelation - `cor.test(x, y, method = "spearman")`

9.5 Vertiefung: Beispiele für häufige Inferenztests

Schauen wir uns für jeden Test aus Kapitel 9.4 ein Anwendungsbeispiel an.

9.5.1 χ^2 -Test

Laden wir den Datensatz `extra`. Ob es wohl einen Zusammenhang gibt zwischen (der Anzahl von) Geschlecht und extremen Alkoholgenuss? Definieren wir ‘extrem’ durch mehr als 10 Kater.

Forschungsfrage: Gibt es einen Zusammenhang zwischen Geschlecht und extremen Alkoholgenuss?

³<http://www.methodenberatung.uzh.ch/de/datenanalyse.html>

⁴<https://www.aerzteblatt.de/archiv/74880/Auswahl-statistischer-Testverfahren>

Synonym wäre zu fragen, ob sich die Stufen von Geschlecht (die Geschlechter, also Mann und Frau) hinsichtlich Säufenextremen Alkoholgenuss unterscheiden.

```
extra <- read.csv("data/extradata.csv")

extra$viel_saeufer <- extra$n_hangover > 10

chisq.test(x = extra$sex, extra$viel_saeufer)
#>
#> Pearson's Chi-squared test with Yates' continuity correction
#>
#> data: extra$sex and extra$viel_saeufer
#> X-squared = 30, df = 1, p-value = 4e-08
```

Achtung, falls Ihre Daten in aggregierter Form vorliegen, müssen Sie sie folgendermaßen übergeben werden:

```
table(x = extra$sex, extra$viel_saeufer) %>% chisq.test
#>
#> Pearson's Chi-squared test with Yates' continuity correction
#>
#> data: .
#> X-squared = 30, df = 1, p-value = 4e-08
```

9.5.2 t-Test

Forschungsfrage: Sind Männer im Schnitt extrovertierter als Frauen?

```
extra %>%
  filter(sex %in% c("Frau", "Mann")) %>%
  mutate(sex = factor($.sex)) %>%
  t.test(extra_mean ~ sex, data = ., alternative = "less")
#>
#> Welch Two Sample t-test
#>
#> data: extra_mean by sex
#> t = 1, df = 500, p-value = 0.9
#> alternative hypothesis: true difference in means is less than 0
#> 95 percent confidence interval:
#> -Inf 0.104
#> sample estimates:
```

```
#> mean in group Frau mean in group Mann
#>           2.91            2.86
```

Auf Deutsch liest sich der letzte Befehlsblock so:



Nimm den Datensatz **extra** UND DANN
 filtere nur Zeilen heraus, in denen bei Geschlecht ‘Mann’ oder ‘Frau’ steht (es gibt Zeilen mit ‘’’’ als Wert) UND DANN
 definiere **sex** als Faktor und zwar so, dass es nur Faktorstufen gibt, die es auch in den Daten gibt (‘Frau oder ’Mann’) UND DANN führe einen gerichteten t-Test durch mit ‘extra_meanals Output-Variable undsex‘ als Gruppierungsvariable.

Der Punkt . meint hier den Datensatz in aktueller Form, so also, wie er aus der letzten (vorherigen) Zeile herausgekommen ist.

Hinweise:

- Der t-Test testet im Standard *ungerichtet*.
- Wird eine Gruppierungsvariable (wie Geschlecht) vom Typ **factor** angegeben, so muss diese 2 Faktorstufen haben. Allein durch filtern wird man zusätzliche Faktorstufen nicht los (im Gegensatz zu Variablen vom Typ **character**, Text). Man muss die Faktorvariable neu als Faktorvariable definieren. Dann werden nur die existierenden Werte als Faktorstufen herangezogen.
- Bei gerichteten Hypothesen sieht **t.test** zwei Möglichkeiten vor: **less** und **greater**. Woher weiß man, welches von beiden man nehmen muss? Die Antwort lautet: Bei Textvariablen sind die Stufen alphabetisch geordnet. R sagt also sozusagen: **Frau < Mann**. Und für ? müssen wir das richtige Ungleichheitszeichen einsetzen (< oder >), so dass es unserer Hypothese entspricht. In diesem Fall glauben wir, dass Frauen weniger (bzw. Männer mehr) trinken, also haben wir **less** gewählt.
- Liegt der Datensatz nicht tidy vor, also gibt es z.B. eine Spalte mit Extraversionswerten für Männer und eine für Frauen, so darf man *nicht* die Formelsyntax (Kringel, Tilde “~”“”) nehmen, sondern benennt die Spalten mit X und Y: **t.text(x = df\$extra_maenner, y = df\$extra_frauen)**.

9.5.3 Varianzanalyse

Forschungsfrage: Unterscheiden sich Menschen mit unterschiedlich viel Kundenkontakt in ihrer Extraversität?

Der Kundenkontakt wurde mit einer Likertskaala gemessen, die mehrere Stufen von “weniger als einmal pro Quartal” bis “im Schnitt mehrfach pro Tag” reichte. Wir gehen nicht davon aus, dass diese Skala Intervallniveau aufweist. Obwohl Ordinalskalenniveau plausibel ist, bleiben wir bei der ANOVA (Varianzanalyse, AOV), die nur nominales Niveau ausschöpft. Beachten Sie, dass die entsprechende Variable **clients_freq** als Ganzzahl in R definiert ist,

obwohl die Abstände nicht sicher gleich sind. Es ist zwar erlaubt, den Stufen einer nominalen Variablen Zahlen zuzuordnen, aber wir sollten nicht vergessen, dass die Zahlen “keine echten Zahlen” sind, also nicht metrisches Niveau aufweisen (zumindest ist das nicht sicher).

Die verschiedenen Stufen einer Variablen kann man sich so ausgeben lassen:

```
extra %>% distinct(clients_freq)
```

Jetzt die ANOVA:

```
aov(extra_mean ~ sex, data = extra) %>% glance
#>   r.squared adj.r.squared sigma statistic p.value df logLik AIC BIC
#> 1  0.00247      0.00125  0.45     2.02   0.156  2  -505 1017 1031
#>   deviance df.residual
#> 1       165         813
```

glance räumt das Ergebnis der ANOVA etwas auf, so dass die Ausgabe ein Dataframe ist und die “Überblick-Koeffizienten” (daher ‘glance’, engl. ‘Blick’) ausgegeben werden. Ganz interessantes Ergebnis: statistisch signifikant ($p < .05$), aber R^2 ist sehr klein. Der F-Wert ist als **statistic** bezeichnet.

9.5.4 Korrelationen auf Signifikanz prüfen

Forschungsfrage: Ist der Extraversion-Mittelwert und die Anzahl der Facebook-Freunde korreliert?

Der Test prüft, ob diese Korrelation 0 ist.

```
cor.test(extra$extra_mean, extra$n_facebook_friends)
#>
#> Pearson's product-moment correlation
#>
#> data: extra$extra_mean and extra$n_facebook_friends
#> t = 1, df = 700, p-value = 0.2
#> alternative hypothesis: true correlation is not equal to 0
#> 95 percent confidence interval:
#> -0.0303 0.1208
#> sample estimates:
#> cor
#> 0.0455
```

Man kann auch - wie beim t-Test - gerichtet testen mit der gleichen Syntax, vgl. `?cor.test`.

9.5.5 Regression

Forschungsfrage: Wie groß ist der Einfluss von der Anzahl von Parties auf die Anzahl der Kater?

```
lm(n_hangover ~ n_party, data = extra) %>% tidy
#>   term estimate std.error statistic p.value
#> 1 (Intercept)  0.159     1.401    0.113 9.10e-01
#> 2 n_party      0.539     0.054    9.983 3.62e-22
lm(n_hangover ~ n_party, data = extra) %>% glance
#>   r.squared adj.r.squared sigma statistic p.value df logLik AIC BIC
#> 1     0.113          0.112   29.2      99.7 3.62e-22  2 -3752 7510 7524
#>   deviance df.residual
#> 1     665751           781
```

`statistic` ist bei dieser Ausgabe übrigens der F-Wert und `sigma` die SD der Residualstreuung.
`estimate` ist die Steigung der Regressionsgeraden.

9.5.6 Wilcoxon-Test

Forschungsfrage: Unterscheiden sich die Geschlechter in ihrer mittleren Extraversion?

Hier nehmen wir nicht an, dass Extraversion metrisch ist, sondern begnügen uns mit der schwächeren Annahme eines ordinalen Niveaus.

```
extra %>%
  filter(sex %in% c("Frau", "Mann")) %>%
  mutate(sex = factor(.sex)) %>%
  wilcox.test(extra_mean ~ sex, data = .)
#>
#> Wilcoxon rank sum test with continuity correction
#>
#> data: extra_mean by sex
#> W = 80000, p-value = 0.4
#> alternative hypothesis: true location shift is not equal to 0
```

9.5.7 Kruskal-Wallis-Test

Forschungsfrage: Unterscheiden sich Menschen mit unterschiedlich viel Kundenkontakt in ihrer Extraversion?

Genau wie beim Wilcoxon-Test gehen wir wieder nur von ordinalem Niveau bei Extraversion aus.

```
extra %>%
  filter(sex %in% c("Frau", "Mann")) %>%
  mutate(sex = factor(. $sex)) %>%
  kruskal.test(extra_mean ~ sex, data = .)
#>
#> Kruskal-Wallis rank sum test
#>
#> data: extra_mean by sex
#> Kruskal-Wallis chi-squared = 0.6, df = 1, p-value = 0.4
```

9.5.8 Shapiro-Test

Forschungsfrage: Ist Extraversion normalverteilt?

Wahrscheinlich ist es sinnvoller, diese Frage mit einem Histogramm (oder QQ-Plot) zu beantworten, weil der Test bei großen Stichproben (zu) schnell signifikant wird. Aber machen wir es mal:

```
shapiro.test(extra$extra_mean)
#>
#> Shapiro-Wilk normality test
#>
#> data: extra$extra_mean
#> W = 1, p-value = 9e-09
```

Signifikant. Die Variable ist also *nicht* (exakt) normalverteilt. Böse Zungen behaupten, die Normalverteilung sei ungefähr so häufig wie Einhörner (Micceri 1989). Trotzdem setzen viele Verfahren sie voraus. Glücklicherweise reicht es häufig, wenn eine Variable *einigermaßen* normalverteilt ist (wobei es hier keine klaren Grenzen gibt).

9.5.9 Logistische Regression

Forschungsfrage: Kann man anhand der Extraversion vorhersagen, ob eine Person Extremtrinker ist?

```
glm(viel_saeufer ~ extra_mean, data = extra, family = "binomial") %>% tidy
#>   term estimate std.error statistic p.value
#> 1 (Intercept) -4.207     0.661      -6.37 1.93e-10
#> 2 extra_mean    0.942     0.218       4.33 1.51e-05
```

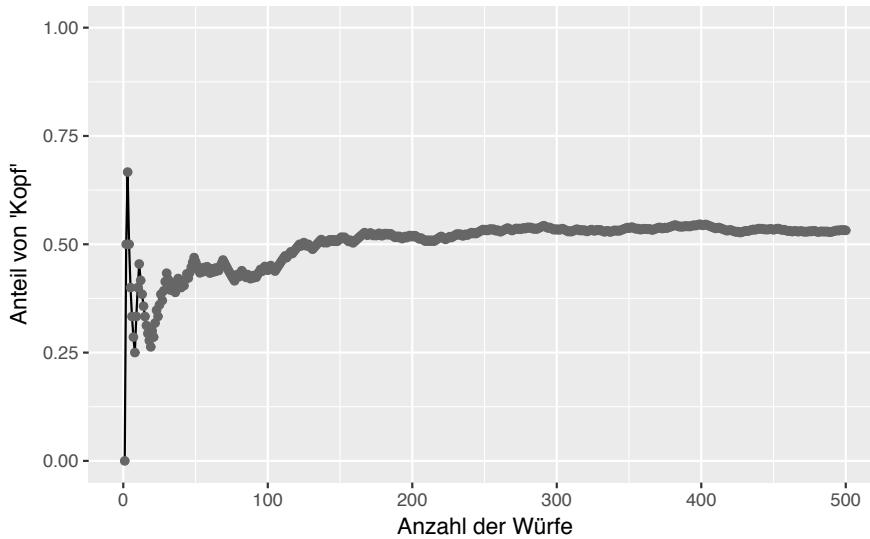


Abbildung 9.5: Anteil von 'Kopf' bei wiederholtem Münzwurf

9.5.10 Spearmans Korrelation

Forschungsfrage: Ist die Extraversion assoziiert mit der Anzahl der Kundenbesuche?

```
cor.test(extra$extra_single_item, extra$clients_freq, method = "spearman")
```

9.6 Zur Philosophie des p-Werts: Frequentismus

Der p-Wert basiert auf der Idee, dass man ein Experiment *unendlich* oft wiederholen könnte (wer die Zeit hat, nicht wahr); und das unter *zufälligen* aber *ansonsten komplett gleichen* Bedingungen; das ist eine Kernidee des sog. ‘Frequentismus’ (Neyman und Pearson 1933). Diese Philosophie betrachtet Wahrscheinlichkeit als der Anteil, der sich bei unendlich häufiger Wiederholung eines Experiments ergibt. Ein Münzwurf hingegen ist das klassische Modell der frequentistischen Idee der Wahrscheinlichkeit (vgl. Abb. 9.5). Wirft man eine faire Münze oft, so nähert sich der relative Anteil von ‘Kopf’ an 50% an.

Ob es im Universum irgendetwas gibt, das unendlich ist, ist streitbar (Rucker 2004, Briggs (2016)). Jedenfalls ist die Vorstellung, das Experiment unendlich oft zu wiederholen, unrealistisch. Inwieweit Zufälligkeit und Vergleichbarkeit hergestellt werden kann, ist auch fragwürdig (Briggs 2016).

Die frequentistische Idee der Wahrscheinlichkeit darf Aussagen wie dieser keine Wahrscheinlichkeit zuweisen: “5 von 10 Marsianer trinken gerne Bier und Schorsch ist Marsianer” (Briggs 2016; Neyman und Pearson 1992; Neyman und Pearson 1933). Häufigkeitsaussagen a la Frequentismus machen hier offenbar wenig Sinn. Trotzdem fühlen sich manche unter uns geneigt, die Wahrscheinlichkeit, dass Schorsch der Marsianer gern Bier trinkt, auf 50% zu

bemessen. Ein anderes, weniger fernes Beispiel: Ich werfe eine Münze hoch, fange sie auf, verdeckt. Wie hoch ist die Wahrscheinlichkeit, dass die Münze mit Kopf nach oben liegt? 50%? Moment, einzelne Ereignisse haben keine Wahrscheinlichkeit, sagt der Frequentismus. Wer sich geneigt fühlt (wie ich), hier doch eine Wahrscheinlichkeit zuzuordnen (50%), der tut dies offenbar nicht auf Basis des Frequentismus. Eine theoretische Position, die Wahrscheinlichkeiten erlaubt, kann man als *epistemologische Wahrscheinlichkeit* bezeichnen (Briggs 2016): Alle möglichen von n Ergebnissen erscheinen uns gleich plausibel. Daher schließen wir, dass die Wahrscheinlichkeit des Ereignisses k 1 durch n ($1/n$) beträgt.

9.7 Alternativen zum p-Wert

Eine Reihe von Alternativen (oder Ergänzungen zum p-Wert) wurden vorgeschlagen.

9.7.1 Konfidenzintervalle

Konfidenzintervalle (Zu) einfach gesagt, gibt ein 95%-Konfidenzintervall an, wie groß der Bereich ist, mit dem der gesuchte Parameter zu 95% Wahrscheinlichkeit liegt (oder allgemeiner das $1 - \alpha$ -Konfidenzintervall). Das kennt man aus dem Wetterbericht, wenn es heißt, dass die Höchsttemperatur morgen zwischen 20 und 24 Grad liegen werde.

Etwas genauer gesagt ist es nach den Urhebern des Konfidenzintervalls, Neyman und Pearson, gar nicht möglich, für ein einzelnes Ereignis eine Wahrscheinlichkeit anzugeben (Clopper und Pearson 1934; Neyman 1935). Wenn ich eine Münze hochwerfe und sie auffange, wie groß ist die Wahrscheinlichkeit, dass sie auf Kopf gelandet ist? 50%? Falsch, sagen ‘Frequentisten’ a la Neyman und Pearson, entweder ist die Münze auf Kopf gelandet, dann kann man höchstens sagen, $p(K) = 1$ oder auf Zahl, dann entsprechend $p(Z) = 1$. Eine Wahrscheinlichkeit macht nur Sinn nach diesem Verständnis, wenn man den Versuch *oft* (unendlich) wiederholt. Daher lautet eine genauere Definition:

Das 95%-Konfidenzintervall ist der Bereich, in dem der Parameter in 95% der Fälle fallen würde bei sehr häufiger Wiederholung des Versuchs.

Mit Parameter ist hier der Mittelwert der Population gemeint (auch bezeichnet als ‘wahrer Mittelwert’). Das Konfidenzintervall macht also Aussagen zur *über ein Verfahren* (einen Bereich berechnen auf Basis von Stichprobendaten), *nicht über den wahren Mittelwert*.

Hier findet sich eine schöne Visualisierung zum Konfidenzintervall⁵.

Genau wie der p-Wert werden Konfidenzintervalle häufig missverstanden (sie sind Blutsbrüder im Geiste). Die Studie von Hoekstra, Morey, Rouder und Wagenmakers (2014) zeigt das auf amüsante Weise. In der Studie legten die Autoren einigen Studenten und Wissenschaftlern sechs Fragen zum Wissens-Konfidenzintervall vor, die beantwortet werden sollten. Es wurde ein Kontext vorgestellt, etwa so “Professor Bumbledorf führt ein Experiment durch. Das

⁵<http://rpsychologist.com/d3/CI/>

Ergebnis fasst er in einem 95%-Konfidenzintervall für den Mittelwert zusammen, welches von 0,1 bis 0,4 reicht". Dann folgten sechs Aussagen, die mit *stimmt* oder *stimmt nicht* zu beantworten waren. Beurteilen auch Sie diese Aussagen⁶.

1. Die Wahrscheinlichkeit, dass der wahre Mittelwert größer als 0 ist, liegt bei mindestens 95%.
 2. Die Wahrscheinlichkeit, dass der wahre Mittelwert gleich 0 ist, ist kleiner als 5%.
 3. Die Nullhypothese, dass der wahre Mittelwert 0 ist, ist wahrscheinlich falsch.
 4. Die Wahrscheinlichkeit, dass der wahre Mittelwert zwischen 0,1 und 0,4 liegt, beträgt 95%.
 5. Wir können zu 95% sicher sein, dass der wahre Mittelwert zwischen 0,1 und 0,4 liegt.
 6. Wenn wir das Experiment immer wieder wiederholen würden, dann würde der wahre Mittelwert in 95% der Fälle zwischen 0,1 und 0,4 fallen.
-

Aussagen 1, 2, 3 und 4 behaupten, der Hypothese bzw. dem Parameter eine Wahrscheinlichkeit zuweisen zu können. Innerhalb des NHST ist das nicht erlaubt, für da Konfidenzintervall sowenig wie für den p-Wert. Aussagen 5 trifft eine Aussage über den wahren Wert, aber Konfidenzintervalle treffen Aussagen über ein Verfahren. Aussage 6 behauptet, dass der wahre Wert variieren könne, tut der aber nicht. Die richtige Aussage, die nicht dabei stand, ist: "Wenn man den Versuch immer wiederholen würden, würden 95% der Intervalle den wahren Mittelwert enthalten". Im Schnitt wurden etwa 3,5 Antworten mit *stimmt* angekreuzt (die Wissenschaftler waren nicht besser als die Studenten).

9.7.2 Effektstärke

Eine weitere Alternative sind Maße der *Effektstärke* (Cohen 1992). Effektstärkemaße geben an, wie sehr sich zwei Parameter unterscheiden: "Deutsche Männer sind im Schnitt 13cm größer als Frauen" (Wikipedia 2017). Oder: "In Deutschland ist die Korrelation von Gewicht und Größe um 0,12 Punkte höher als in den USA" (frei erfunden). Im Gegensatz zu p-Werten wird keine Art von Wahrscheinlichkeitsaussage angestrebt, sondern die Größe von Parameter(unterschieden) quantifiziert. Effektstärken sind, im Gegensatz zum p-Wert, auch nicht abhängig von der Stichprobengröße. Man kann Effektstärken in nicht-standardisierte (wie Unterschiede in der Größe) oder standardisierte (wie Unterschiede in der Korrelation) einteilen.

Nicht-standardisierte Effektstärken haben den Vorteil der Anschaulichkeit. Standardisierte Effektgrößen sind präziser, aber unanschaulicher. Bei Variablen mit unanschaulichen Metriken (wie psychologische Variablen und Umfragen) ist ein standardisiertes Maß häufig nützlicher.

Anschauliche Variablen sind oft mit unstandardisiertes Effektstärken adäquat

⁶alle sechs sind falsch

dargestellt. Variablen mit wenig anschaulichen Metriken profitieren von standariserten Effektstärkemaßen.

Um zwei Mittelwerte zu vergleichen, ist *Cohens d* gebräuchlich. Es gibt den Unterschied der Mittelwert standardisiert an der Standardabweichung an (Cohen 1988). Das ist oft sinnvoll, denn 5\$ Preisunterschied können viel oder weniger sein: Bei Eiskugeln wäre der Unterschied enorm (die Streuung ist viel weniger als 5€), bei Sportwagen wäre der Unterschied gering (die Streuung ist viel höher als 5€).

9.7.2.1 Typische Effektstärkemaße

Zu den typischen Effektstärkemaßen zählen die folgenden (vgl. Eid, Gollwitzer, und Schmitt (2010)):

- d (Cohens d) wird zur Bemessung des Unterschieds der Überlappung zweier Verteilungen verwendet, z.B. um die Effektstärke eines t-Werts zu quantifizieren. d berechnet sich im einfachsten Fall als: $d = \frac{\mu_1 - \mu_2}{sd}$.
- r Pearsons R ist ein Klassiker, um die Stärke des linearen Zusammenhangs zweier metrischen Größen zu quantifizieren. r berechnet sich als: $r = mw(\sum z_x z_y)$, wobei mw für den Mittelwert steht und z für einen z-Wert.
- R^2 , η^2 sind Maße für den Anteil aufgeklärter Varianz; sie finden in der Varianzanalyse oder der Regressionsanalyse Verwendung. R^2 wird u.a. so berechnet: $R^2 = \frac{QS_F}{QS_T}$, wobei QS für die Quadratsummen stehen und QS_F für die Varianz, die auf den Faktor (unabhängige Variable) zurückgeht und QS_T für die Gesamtvarianz.
- f^2 ist ein Maß, dass aus der erklärten Varianz abgeleitet ist. Es gibt das Verhältnis von erklärter zu nicht-erklärter Varianz wieder (auch ‘signal-noise-ratio’ genannt). Es berechnet sich als $f^2 = \frac{R^2}{1-R^2}$.
- ω (Cohens Omega) ist ein Maß für die Stärke des Zusammenhangs zweier nominaler Variablen, abgeleitet vom χ^2 -Test. Es berechnet sich als $\omega = \sqrt{\chi^2}$.
- OR (Odds Ratio) ist ebenfalls ein Maß für die Stärke des Zusammenhangs zweier nominaler Variablen, allerdings *binärer* (zweistufige) Variablen. Es berechnet sich als $OR = \frac{c}{1-c}$, wobei c die Chancen für ein Ereignis E angeben (z.B. 9:1). OR kann aus ω abgeleitet werden.

Tabelle 9.1 gibt einen groben Überblick über Effektstärken (nach Cohen (1988) und Eid, Schmitt und Gollwitzer (2010)). Zu beachten ist, dass die Einschätzung was ein ‘großer’ oder ‘kleiner’ Effekt ist, nicht pauschal übers Knie gebrochen werden sollte. Besser ist es, die Höhe der Effektstärke im eigenen Datensatz mit relevanten anderen Datensätzen zu vergleichen.

Was ein “kleiner” oder “großer” Effekt ist, sollte im Einzelfall entschieden werden.

Mit dem Paket `pwr` kann man sich Cohens Konventionen der Effektstärkehöhen in Erinnerung rufen lassen. Er bietet folgende Optionen:

```
cohen.ES(test = c("p", "t", "r", "anov", "chisq", "f2"),
         size = c("small", "medium", "large"))
```

Tabelle 9.1: Überblick über gängige Effektstärkemaße

Name	kleiner Effekt	mittlerer Effekt	großer Effekt
Cohens d	.2-.5	.5-.8	>.8
r	0.1	0.3	0.5
$\hat{R^2}$, $\hat{\eta^2}$	0.01	0.06	0.14
$\hat{f^2}$	0.02	0.15	0.35
$\hat{\omega^2}$	0.1	0.3	0.5
OR	1.5	3	9

9.7.2.2 Effektstärken berechnen

Möchte man sich Effektstärken berechnen lassen, ist das Paket `compute.es` hilfreich. Im Folgenden sind Effektstärkeberechnungen für gängige Inferenztests vorgestellt, in Fortsetzung zu den Beispielen oben.

- χ^2 -Test: `compute.es::chies(30, n = 826)`
- t-Test: `compute.es::tes(t = 1, n.1 = 529, n.2 = 286)`
- Varianzanalyse (F-Test): `glance` gibt R^2 aus; mit `etaSquared(mein_aov)` ebenfalls. Möchte man f^2 berechnen, so tut man das am besten per Hand, z.B. $0.002 / (1-0.002)$.
- Korrelation: Der Korrelationswert r ist schon ein Maß der Effekstärke. Yeah.
- Regression: Die Steigung der Regressionsgeraden (b) ist ein (unstandardisiertes) Maß für die Stärke des (“Netto”)-Einflusses eines Prädiktors. R^2 hingegen ein Maß für die relative Varianzaufklärung aller Prädiktoren gemeinsam.
- Logistische Regression: Mit `BaylorEdPsych::PseudoR2(mein_glm_objekt)` kann man eine Art R^2 bekommen (s. Kapitel 11.8).
- Wilcoxon-Test/ Mann-Whitney-U-Test: Anteil der paarweisen Vergleiche, die hypothesenkonform sind (vgl. Kerby 2014). Dazu kann man z.B. die Funktion `prop_fav` aus dem Paket `prada` nutzen (vgl. `help(prop_fav)`).

9.7.3 Bayes-Statistik

Bayes' Ansatz verrechnet zwei Komponenten, um die Wahrscheinlichkeit einer Hypothese im Lichte bestimmter Daten zu berechnen. Der Ansatz ist elegant, mathematisch luppenrein und ist überhaupt eine tolle Sache. Bayes' Theorem gibt uns das, was uns eigentlich interessiert: Die Wahrscheinlichkeit der getesteten Hypothese, im Lichte der vorliegenden Daten: $p(H|D)$. Diesen Wert nennt man auch den *Vorhersagewert*. Zur Erinnerung: Der p-Wert gibt die Wahrscheinlichkeit der Daten an, unter Annahme der getesteten Hypothese: $p(D|H)$. Offenbar sind beide Terme nicht identisch.

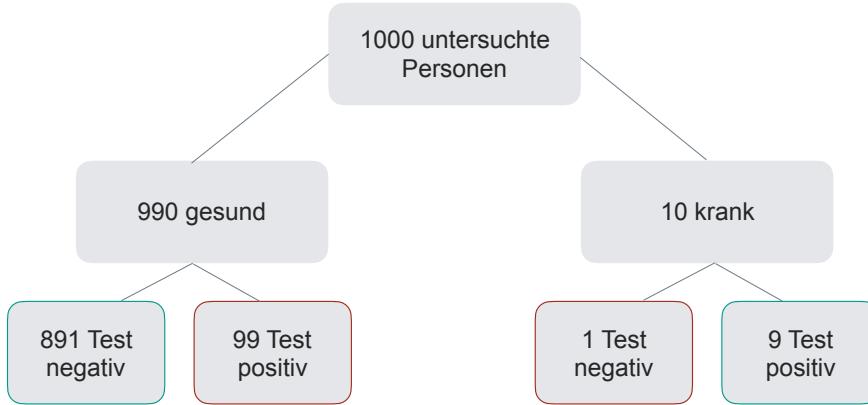


Abbildung 9.6: Die zwei Stufen der Bayes-Statistik in einem einfachen Beispiel

Die Bayes-Statistik zieht zwei Komponenten zur Berechnung von $p(H|D)$ heran. Zum einen die Grundrate einer Hypothese $p(H)$ zum anderen die relative Plausibilität der Daten unter meiner Hypothese im Vergleich zur Plausibilität der Daten unter konkurrierenden Hypothesen. Betrachten wir ein Beispiel. Die Hypothese "Ich bin krank" sei unter Betrachtung (jetzt noch keine vorschnellen Einschätzungen). Die Grundrate der fraglichen Krankheit sei 10 von 1000 (1%). Der Test, der zur Diagnose der Krankheit verwendet wird, habe eine Sicherheit von 90%. Von 100 Kranken wird der Test demnach 90 identifizieren (auch *Sensitivität* genannt) und 10 werden übersehen (ein Überseh- oder *Betafehler* von 10%). Umgekehrt wird der Test von 100 Gesunden wiederum 90 als Gesund, und demnach korrekt diagnostizieren (*Spezifität*); 10 werden fälschlich als krank einschätzen (*Fehlalarm* oder *Alpha-Fehler*).

Jetzt Achtung: Der Test sagt, ich sei krank. Die Gretchen-Frage lautet, wie hoch ist die Wahrscheinlichkeit, dass diese Hypothese, basierend auf den vorliegenden Daten, korrekt ist?

Abbildung 9.6 stellt das Beispiel in Form eines Baumdiagrammes dar.

In der Medizin ist 'positiv' zumeist eine schlechte Nachricht, es soll sagen, dass der Test der Meinung ist, die getestete Person ist krank (das getestete Kriterium trifft zu).

Wie man leicht nachrechnen kann, beträgt die Wahrscheinlichkeit, *in Wirklichkeit krank* zu sein, wenn der positiv ist, $\sim 8\%$: $9/(99 + 9) = \frac{9}{108} \approx 8\%$. Das überrascht auf den ersten Blick, ist doch der Test so überragend zufällig (jedenfalls zu 90%)! Aber die Wahrscheinlichkeit, dass die Hypothese 'krank' zutrifft, ist eben nicht nur abhängig von der Sicherheit des Tests, sondern auch von der Grundrate. Beide Komponenten sind nötig, um den Vorhersagewert zu berechnen. Der p-Wert begnügt sich mit der Aussage, ob der Test positiv oder negativ ist. Die Grundrate wird nicht berücksichtigt.

Die Bayes-Statistik liefert die Wahrscheinlichkeit einer Hypothese H, wenn wir die Daten D (d.h. ein gewisses Stichprobenergebnis) gefunden haben: $p(H|D)$. Damit gibt die Bayes-Statistik die Antwort, die sich die meisten Anwender wünschen.

Fairerweise muss man hinzufügen, dass die Grundrate für die Wissenschaft oft nicht einfach zu bestimmen ist. Wer kennt schon die Grundrate der 'guten Ideen'? Vielleicht der liebe Gott,

aber der hilft uns nicht⁷ (God 2016). Wir werden also eine Einschätzung treffen müssen, die subjektiv sein kann. Diese Subjektivität ist von Kritikern moniert worden.

Auf der anderen Seite kann man diese Subjektivität umgehen, indem man nur angibt, um welchen Faktor die H1 wahrscheinlicher ist als die H0, durch die Daten der Studie. Das wird durch den sog. *Bayes-Faktor BF* ausgedrückt. Liegt *BF* bei 10, so eine gängige Konvention, so ist dies “starke” Evidenz für H1 (da H1 dann 10 mal wahrscheinlicher als die H0); entsprechend stark ist ein *BF* von 0.1 (1/10) - zugunsten H0. Gängige Software (s. Abschnitt 9.10) geben den Bayes-Faktor aus.

Ein t-Test a la Bayes kann z.B. so berechnet werden:

```
extra %>%
  group_by(sex) %>%
  summarise(mean(extra_mean, na.rm = TRUE))
#> # A tibble: 3 x 2
#>   sex `mean(extra_mean, na.rm = TRUE)` 
#>   <fctr>                <dbl>
#> 1 Frau                  2.91
#> 2 Mann                  2.86
#> 3 NA                    2.73

extra %>%
  filter(sex %in% c("Mann", "Frau")) %>%
  mutate(sex = factor(sex)) %>%
  as.data.frame %>% # 'ttestBF' verkraftet nur althergebrachte data.frames!
  ttestBF(formula = extra_mean ~ sex,
          data = .) # 'formula' muss hingeschrieben sein, sonst droht Fehlermeldung
#> Bayes factor analysis
#> -----
#> [1] Alt., r=0.707 : 0.22 ±0%
#>
#> Against denominator:
#>   Null, mu1-mu2 = 0
#> ---
#> Bayes factor type: BFindepSample, JZS
```

Hey, Sie haben gerade einen Bayes-Test gerechnet! Wow! Das Ergebnis zeigt einen *BF* von 0.24; Evidenz *zugunsten* der H0. Nicht stark; sondern schwach. Keine überzeugende Evidenz für H1. Man beachte, dass der Befehl hier “indifferent” gegenüber der H0 und der H1 war. A priori wurden hier beide Hypothesen als gleich wahrscheinlich angesehen. Jetzt ist unsere Überzeugung für die H1 gesunken bzw. für die H0 gestiegen und zwar etwa um den Faktor 1/4 auf 0.24.

⁷<https://twitter.com/TheTweetOfGod/status/688035049187454976>

9.8 Aufgaben⁸



Richtig oder Falsch!?

1. Der p-Wert gibt die Wahrscheinlichkeit der H₀ an unter der Annahme der Daten.
2. $p(D|H) = p(H|D)$
3. Der p-Wert sagt, wie gut die Daten zur Nullhypothese passen.
4. Bei sehr großen Stichproben werden nur sehr große Effekte signifikant.
5. Egal wie klein die Effektstärke ist, es existiert eine Stichprobengröße, die diesen Effekt beliebig signifikant werden lässt.
6. Wenn der p-Wert kleiner als 5% ist, dann ist meine Hypothese (H₁) höchstwahrscheinlich richtig.
7. Wenn der p-Wert größer als 5% ist, dann ist das ein Beleg für die H₀.
8. Der p-Wert basiert auf der Idee, dass man ein Experiment unendlich oft wiederholt; und das unter zufälligen aber ansonsten komplett gleichen Bedingungen.
9. Das 95%-Konfidenzintervall ist der Bereich, in dem der Parameter in 95% der Fälle fallen würde bei sehr häufiger Wiederholung des Versuchs.
10. Der Vorhersagewert ist definiert als $p(H|D)$.

9.9 Fazit

Der p-Wert ist eine häufig verwendete Methode, um datenbasiert zu entscheiden, ob man eine Hypothese annimmt oder nicht. Allerdings hat der p-Wert auch seine Probleme.

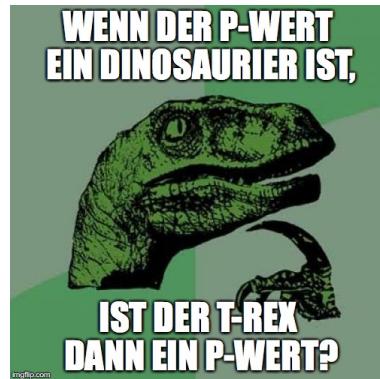
Der p-Wert sollte nicht als einziges Kriterium verwendet werden, um eine Hypothese bzw. ein Modell zu beurteilen.

Da der p-Wert aber immer noch der Platzhirsch auf vielen Forschungsauen ist, führt kein Weg um ihn herum. Er muss genau verstanden werden: Was er sagt und - wichtiger noch - was er nicht sagt.

Alternativen zum p-Wert sind

- Konfidenzintervalle
- Effektstärkemaße inkl. Maße der Vorhersagegenauigkeit
- Bayes-Theorem

⁸F, F, R, F, F, F, F, R, R, R



9.10 Verweise

- Eine Einführung zur Bayes-Statistik findet man z.B. bei Kruschke (2010) oder bei Etz u. a. (2016).
- Eine ausführliche Darstellung der Inferenzstatistik und des p-Werts findet sich z.B. bei Lübke und Vogt (2014) oder Eid, Gollwitzer, und Schmitt (2010).
- Eine vielversprechende, noch recht neue Software ist JASP⁹, die nicht nur schöne Diagramme erstellt, sondern auch auf Mausklick eine Reihe von bayesianischer (und frequentistischer) Tests durchrechnet.

Teil III

Geleitetes Modellieren

Kapitel 10

Lineare Regression



Lernziele:

- Wissen, was man unter Regression versteht.
- Die Annahmen der Regression überprüfen können.
- Regression mit kategorialen Prädiktoren durchführen können.
- Die Modellgüte bei der Regression bestimmen können.
- Interaktionen erkennen und ihre Stärke einschätzen können.

Für dieses Kapitel benötigen Sie folgende Pakete:

```
library(caret) # Modellieren
library(tidyverse) # Datenjudo, Visualisierung, ...
library(gridExtra) # Mehrere Plots kombinieren
library(modelr) # Residuen und Schätzwerte zum Datensatz hinzufügen
library(broom) # Regressionswerte geordnet ausgeben lassen
```

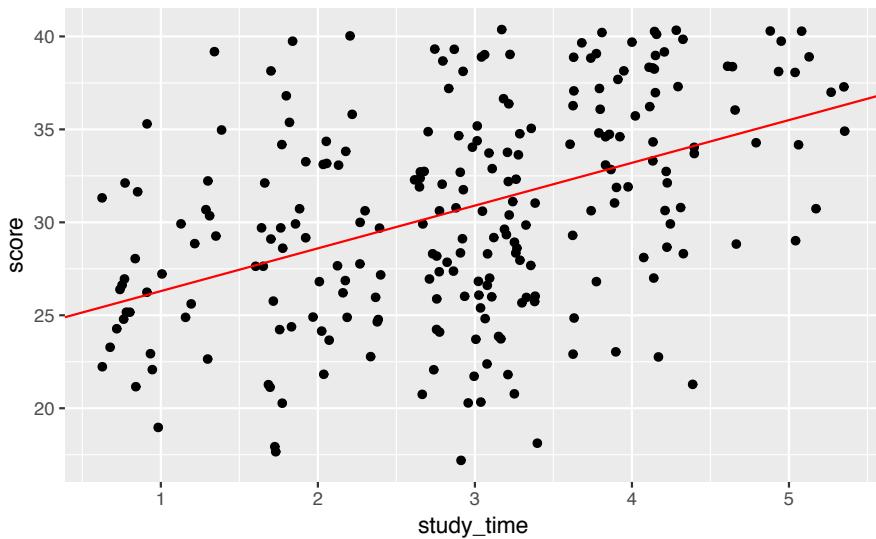


Abbildung 10.1: Beispiel für eine Regression

10.1 Die Idee der klassischen Regression

Regression ist eine bestimmte Art der *Modellierung* von Daten. Wir legen eine Gerade ‘schön mittig’ in die Daten; damit haben wir ein einfaches Modell der Daten (vgl. Abb. 10.1). Die Gerade ‘erklärt’ die Daten: Für jeden X-Wert liefert sie einen Y-Wert als Vorhersage zurück.

```
stats_test <- read.csv("data/stats_test.csv")

stats_test %>%
  ggplot +
  aes(x = study_time, y = score) +
  geom_jitter() +
  geom_abline(intercept = 24,
              slope = 2.3,
              color = "red")
```

Wie wir genau die Regressionsgerade berechnet haben, dazu gleich mehr. Fürs Erste begnügen wir uns mit der etwas groberen Beobachtung, dass die Gerade ‘schön mittig’ in der Punktewolke liegt.

Schauen wir uns zunächst die Syntax genauer an.



Lade die CSV-Datei mit den Daten als `stats_test`.

Nehme `stats_test` UND DANN...
starte ein neues Diagramm mit `ggplot` UND
definiere das Diagramm (X-Achse, Y-Achse) UND DANN

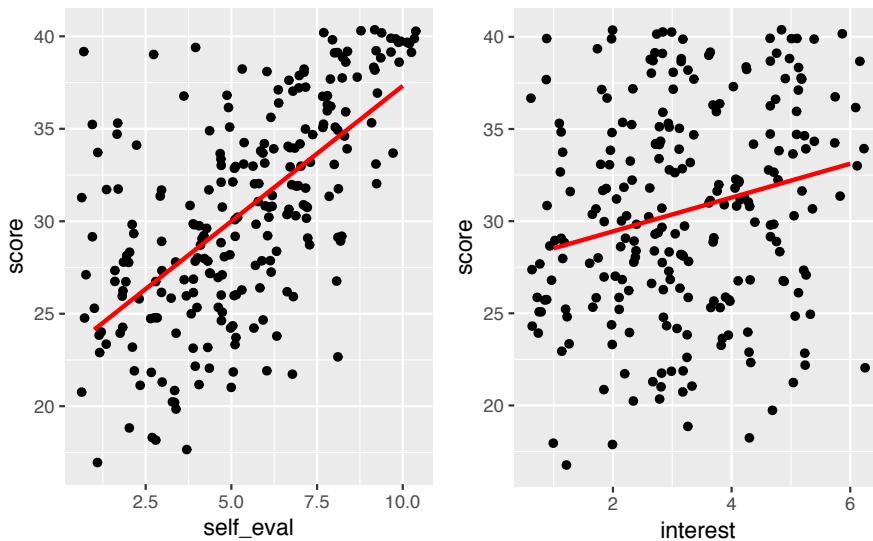


Abbildung 10.2: Zwei weitere Beispiele für Regressionen

zeichne das Geom “Jitter” (verwackeltes Punktediagramm) UND DANN und zeichne danach eine Gerade (“abline” in rot).

Eine Regression zeigt anhand einer Regressionsgeraden einen “Trend” in den Daten an (s. weitere Beispiele in Abb. 10.2).

Eine Regression lädt förmlich dazu ein, Vorhersagen zu treffen: Hat man erstmal eine Gerade, so kann man für jeden X-Wert (“Prädiktor”) eine Vorhersage für den Y-Wert (“Kriterium”) treffen. Anhand des Diagramms kann man also für jede Person (d.h. jeden Wert innerhalb des Wertebereichs von `study_time` oder einem anderen Prädiktor) einen Wert für `score` vorhersagen. Wie gut die Vorhersage ist, steht erstmal auf einen anderen Blatt.

Man beachte, dass eine Gerade über ihre *Steigung* und ihren *Achsenabschnitt* festgelegt ist; in Abb. 10.1 ist die Steigung 2.3 und der Achsenabschnitt 24. Der Achsenabschnitt zeigt also an, wie viele Klausurpunkte man “bekommt”, wenn man gar nicht lernt (Gott bewahre); die Steigung gibt eine Art “Wechselkurs” an: Wie viele Klausurpunkte bekomme ich pro Stunde, die ich lerne.

Unser Modell ist übrigens einfach gehalten: Man könnte argumentieren, dass der Zusatznutzen der 393. Stunde lernen geringer ist als der Zusatznutzen der ersten paar Stunden. Aber dann müssten wir anstelle der Gerade eine andere Funktion nutzen, um die Daten zu modellieren. Lassen wir es erst einmal einfach hier.

Als “Pseudo-R-Formel” ausgedrückt:

```
score = achsenabschnitt + steigung*study_time
```

Die Vorhersage für die Klausurpunkte (`score`) einer Person sind der Wert des Achsenabschnitts plus das Produkt aus der Anzahl der gelernten Stunden mal den Zusatznutzen pro gelernter

Stunde.

Aber wie erkannt man, ob eine Regression “gut” ist - die Vorhersagen also präzise?

In R kann man eine Regression so berechnen:

```
lm(score ~ study_time, data = stats_test)
#>
#> Call:
#> lm(formula = score ~ study_time, data = stats_test)
#>
#> Coefficients:
#> (Intercept)  study_time
#>           23.98          2.26
```

`lm` steht dabei für “lineares Modell”; allgemeiner gesprochen lautet die Rechtschreibung für diesen Befehl:

```
lm(kriterium ~ praediktor, data = meine_datentabelle)
```

Um ausführlichere Informationen über das Regressionsmodell zu bekommen, kann man die Funktion `broom::tidy` nutzen:

```
mein_lm <- lm(kriterium ~ praediktor, data = meine_datentabelle)
tidy(mein_lm)
```

Natürlich kann das auch ~~in der Pfeife rauchen~~ mit der Pfeife darstellen:

```
lm(kriterium ~ praediktor, data = meine_datentabelle) %>%
  summary
```

10.2 Vorhersagegüte

Der einfache Grundsatz lautet: Je geringer die Vorhersagefehler, desto besser; Abb. 10.3 zeigt ein Regressionsmodell mit wenig Vorhersagefehler (links) und ein Regressionsmodell mit viel Vorhersagefehler (rechts).

In einem Regressionsmodell lautet die grundlegenden Überlegung zur Modellgüte damit:

Wie groß ist der Unterschied zwischen Vorhersage und Wirklichkeit?

Die Größe des Unterschieds (Differenz, “Delta”) zwischen vorhergesagten (geschätzten) Wert und Wirklichkeit, bezeichnet man als *Fehler*, *Residuum* oder *Vorhersagefehler*, häufig mit ϵ (griechisches e wie “error”) abgekürzt.

Betrachten Sie die beiden Plots in Abb. 10.3. Die rote Linie gibt die *vorhergesagten* (geschätzten) Werte wieder; die Punkte die *beobachteten* (“echten”) Werte. Je länger die blauen

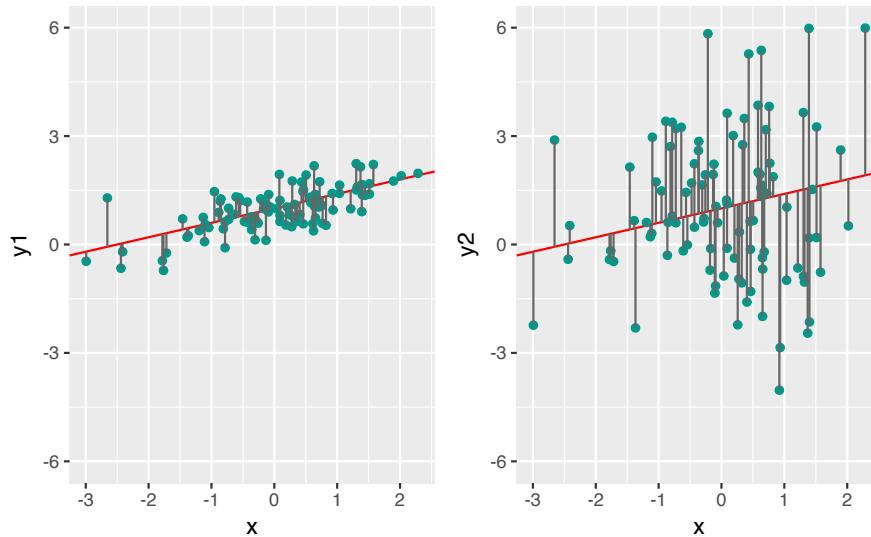


Abbildung 10.3: Geringer (links) vs. hoher (rechts) Vorhersagefehler

Linien, desto größer die Vorhersagefehler. Je größer der Vorhersagefehler, desto schlechter. Und umgekehrt.

Je kürzer die typische “Abweichungslinie”, desto besser die Vorhersage.

Sagt mein Modell voraus, dass Ihre Schuhgröße 49 ist, aber in Wahrheit liegt sie bei 39, so werden Sie dieses Modell als schlecht beurteilen, wahrscheinlich.

Leider ist es nicht immer einfach zu sagen, wie groß der Fehler sein muss, damit das Modell als “gut” bzw. “schlecht” gilt. Man kann argumentieren, dass es keine wissenschaftliche Frage sei, wie viel “viel” oder “genug” ist (Briggs 2016). Das ist zwar plausibel, hilft aber nicht, wenn ich eine Entscheidung treffen muss. Stellen Sie sich vor: Ich zwinge Sie mit der Pistole auf der Brust, meine Schuhgröße zu schätzen.

Eine einfache Lösung ist, das beste Modell unter mehreren Kandidaten zu wählen.

Ein anderer Ansatz ist, die Vorhersage in Bezug zu einem Kriterium zu setzen. Dieses “andere Kriterium” könnte sein “einfach die Schuhgröße raten”. Oder, etwas intelligenter, Sie schätzen meine Schuhgröße auf einen Wert, der eine gewisse Plausibilität hat, also z.B. die durchschnittliche Schuhgröße des deutschen Mannes. Auf dieser Basis kann man dann quantifizieren, ob und wie viel besser man als dieses Referenzkriterium ist.

10.2.1 Mittlere Quadratfehler

Eine der häufigsten Gütekennzahlen ist der *mittlere quadrierte Fehler* (engl. “mean squared error”, MSE), wobei Fehler wieder als Differenz zwischen Vorhersage (`pred`) und beobachtete Wirklichkeit (`obs`, `y`) definiert ist. Dieser berechnet für jede Beobachtung den Fehler, quadriert diesen Fehler und bilden dann den Mittelwert dieser “Quadratfehler”, also einen *mittleren Quadratfehler*. Die englische Abkürzung *MSE* ist auch im Deutschen gebräuchlich.

$$MSE = \frac{1}{n} \sum (pred - obs)^2$$

Konzeptionell ist dieses Maß an die Varianz angelehnt. Zieht man aus diesem Maß die Wurzel, so erhält man den sog. *root mean square error* (RMSE), welchen man sich als die Standardabweichung der Vorhersagefehler vorstellen kann. In Pseudo-R-Syntax:

```
RMSE <- sqrt(mean((df$pred - df$obs)^2))
```

Der RMSE hat die selben Einheiten wie die zu schätzende Variable, also z.B. Schuhgrößen-Nummern.

10.2.2 R-Quadrat (R^2)

R^2 , auch *Bestimmtheitsmaß* oder *Determinationskoeffizient* genannt, setzt die Höhe unseres Vorhersagefehlers im Verhältnis zum Vorhersagefehler eines “Nullmodell”. Das Nullmodell hier würde sagen, wenn es sprechen könnte: “Keine Ahnung, was ich schätzen soll, mich interessieren auch keine Prädiktoren, ich schätzen einfach immer den Mittelwert der Grundgesamtheit!”.

Analog zum Nullmodell-Fehler spricht auch von der Gesamtvarianz oder SS_T (sum of squares total); beim Vorhersagefehler des eigentlichen Modells spricht man auch von SS_M (sum of squares model).

Damit gibt R^2 an, wie gut unsere Vorhersagen im Verhältnis zu den Vorhersagen des Nullmodells sind. Ein R^2 von 25% (0.25) hieße, dass unser Vorhersagefehler 25% kleiner ist als der der Nullmodells. Ein R^2 von 100% (1) heißt also, dass wir den kompletten Fehler reduziert haben (Null Fehler übrig) - eine perfekte Vorhersage. Etwas formaler, kann man R^2 so definieren:

$$R^2 = 1 - \left(\frac{SS_T - SS_M}{SS_T} \right)$$

Präziser, in R-Syntax:

```
R2 <- 1 - sum((df$pred - df$obs)^2) / sum((mean(df$obs) - df$obs)^2)
```

Praktischerweise gibt es einige R-Pakete, z.B. *caret*, die diese Berechnung für uns besorgen:

```
postResample(obs = obs, pred = pred)
```

Hier steht **obs** für beobachtete Werte und **pred** für die vorhergesagten Werte (beides numerische Vektoren). Dieser Befehl gibt sowohl RMSE als auch R^2 wieder. Wir betrachten gleich ein Beispiel an echten Daten.



Verwendet man die Korrelation (r) oder R^2 als Gütekriterium, so sollte man sich über folgenden Punkt klar sein. Bei Skalierung der Variablen ändert sich die Korrelation nicht; das gilt auch für R^2 . Beide Koeffizienten ziehen allein auf das *Muster* der Zusammenhänge ab - nicht die Größe der Abstände. Aber häufig ist die Größe der Abstände zwischen beobachteten und vorhergesagten Werten das, was uns interessiert. In dem Fall wäre der MSE vorzuziehen.

10.3 Die Regression an einem Beispiel erläutert

Schauen wir uns den Datensatz zur Statistikklausur noch einmal an. Welchen Einfluss hat die Lernzeit auf den Klausurerfolg? Wie viel bringt es also zu lernen? Wenn das Lernen keinen Einfluss auf den Klausurerfolg hat, dann kann man es ja gleich sein lassen... Aber umgekehrt, wenn es viel bringt, ok gut, dann könnte man sich die Sache (vielleicht) noch mal überlegen. Aber was heißt "viel bringen" eigentlich?

Wenn für jede Stunde Lernen viele zusätzliche Punkte herausspringen, dann bringt Lernen viel. Allgemeiner: Je größer der Zuwachs im Kriterium ist pro zusätzliche Einheit des Prädiktors, desto größer ist der Einfluss des Prädiktors.

Natürlich könnte jetzt jemand argumentieren, dass die ersten paar Stunden lernen viel bringen, aber dann flacht der Nutzen ab, weil es ja schnell einfach und trivial wird. Aber wir argumentieren (erstmal) so nicht. Wir gehen davon aus, dass jede Stunde Lernen gleich viel (oder wenig) Nutzen bringt.

Geht man davon aus, dass jede Einheit des Prädiktors gleich viel Zuwachs bringt, unabhängig von dem Wert des Prädiktors, so geht man von einem linearen Einfluss aus.

Versuchen wir im ersten Schritt die Stärke des Einfluss an einem Streudiagramm abzuschätzen (s. Abb. 10.1).

Hey R - berechne uns die "Trendlinie"! Dazu nimmt man den Befehl `lm`:

```
mein_lm <- lm(score ~ study_time, data = stats_test)
tidy(mein_lm)
#>   term estimate std.error statistic p.value
#> 1 (Intercept)  23.98     0.934    25.67 2.94e-70
#> 2 study_time    2.26     0.300     7.54 1.02e-12
```

`lm` steht für 'lineares Modell', eben weil eine *Linie* als Modell in die Daten gelegt wird. Aha. Die Steigung der Geraden beträgt 2.3 - das ist der Einfluss des Prädiktors Lernzeit auf das Kriterium Klausurerfolg! Man könnte sagen: Der "Wechselkurs" von Lernzeit auf Klausurpunkte. Für jede Stunde Lernzeit bekommt man offenbar 2.3 Klausurpunkte (natürlich viel zu leicht). Wenn man nichts lernt (`study_time == 0`) hat man 24 Punkte.

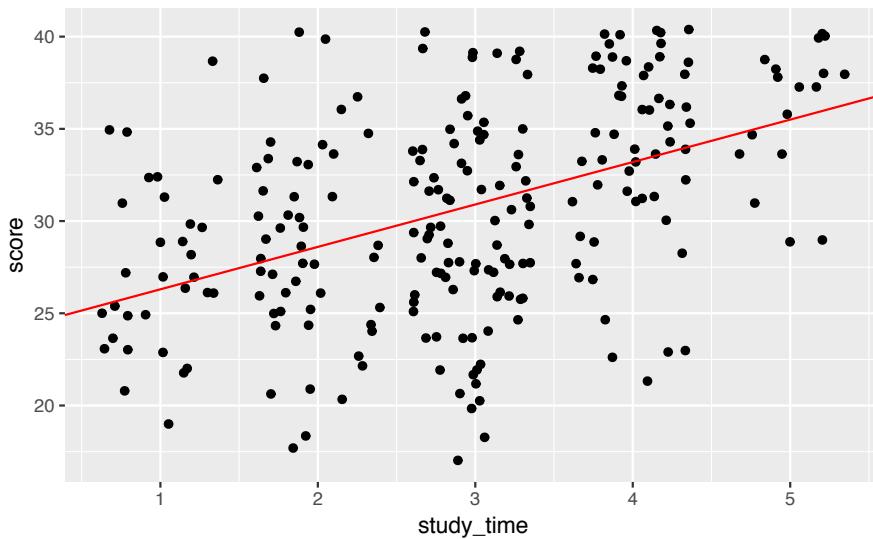


Abbildung 10.4: Streudiagramm von Lernzeit und Klausurerfolg

Der Einfluss des Prädiktors steht unter ‘estimate’. Der Kriteriumswert wenn der Prädiktor Null ist steht unter ‘(Intercept)’.

Malen wir diese Gerade in unser Streudiagramm (Abbildung 10.4).

```
ggplot(data = stats_test) +
  aes(y = score, x = study_time) +
  geom_jitter() +
  geom_abline(slope = 2.3, intercept = 24, color = "red")
```

Jetzt kennen wir die Stärke (und Richtung) des Einflusses der Lernzeit. Ob das viel oder wenig ist, ist am besten im Verhältnis zu einem Referenzwert zu sagen.

Die Gerade wird übrigens so in die Punktewolke gelegt, dass die (quadrierten) Abstände der Punkte zur Geraden minimal sind. Dies wird auch als *Kriterium der Kleinsten Quadrate (Ordinary Least Squares, OLS)* bezeichnet.

Jetzt können wir auch einfach Vorhersagen machen. Sagt uns jemand, ich habe “viel” gelernt (Lernzeit = 4), so können wir den Klausurerfolg grob im Diagramm ablesen.

Genauer geht es natürlich mit dieser Rechnung:

$$y = 4 * 2.3 + 24$$

Oder mit diesem R-Befehl:

```
predict(mein_lm, data.frame(study_time = 4))
#> 1
#> 33
```

Berechnen wir noch die Vorversagegüte des Modells. Dazu kann man den Befehl `summary` nehmen, oder auch `broom::glance`. `glance` gibt Informationen zur Modellgüte zurück und das in Form eines Dateframes. Summary liefert eine Menge Informationen mit einem infomrationen Ausdruck, aber nicht in Form eines Dataframes (sondern in Form einer Liste).

```
glance(mein_lm)
#>   r.squared adj.r.squared sigma statistic p.value df logLik AIC BIC
#> 1     0.194          0.191  5.15      56.8 1.02e-12  2   -727 1459 1470
#>   deviance df.residual
#> 1     6255           236
```

Das Bestimmtheitsmaß R^2 ist mit 0.19 “ok”: 19-% der Varianz des Klausurerfolg wird im Modell ‘erklärt’. ‘Erklärt’ meint hier, dass wenn die Lernzeit konstant wäre, würde die Varianz von Klausurerfolg um diesen Prozentwert sinken.

10.4 Überprüfung der Annahmen der linearen Regression

Aber wie sieht es mit den Annahmen aus?

- Die *Linearität des Zusammenhangs* haben wir zu Beginn mit Hilfe des Scatterplots überprüft. Es schien einigermaßen zu passen.
- Zur Überprüfung der *Normalverteilung der Residuen* zeichnen wir ein Histogramm (s. Abbildung 10.5). Die *Residuen* können über den Befehl `add_residuals` (Paket `modelr`) zum Datensatz hinzugefügt werden. Dann wird eine Spalte mit dem Namen `resid` zum Datensatz hinzugefügt.

Hier scheint es zu passen:

```
stats_test %>%
  add_residuals(mein_lm) %>%
  ggplot +
  aes(x = resid) +
  geom_histogram()
```

Sieht passabel aus. Übrigens kann man das Paket `modelr` auch nutzen, um sich komfortabel die vorhergesagten Werte zum Datensatz hinzufügen zu lassen (Spalte `pred`):

```
stats_test %>%
  add_predictions(mein_lm) %>%
  select(pred) %>%
```

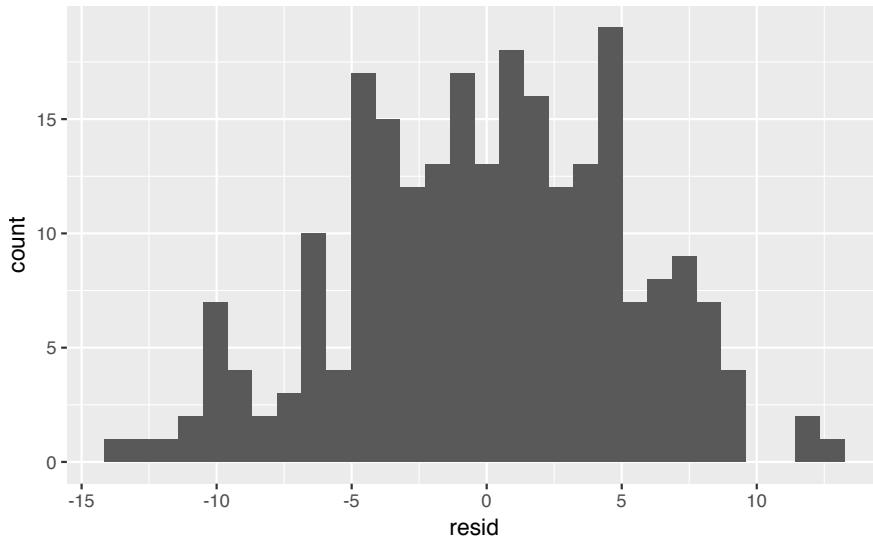


Abbildung 10.5: Die Residuen verteilen sich hinreichend normal.

```
head
#>   pred
#> 1 35.3
#> 2 30.8
#> 3 35.3
#> 4 28.5
#> 5 33.0
#> 6   NA
```

- *Konstante Varianz*: Dies kann z. B. mit einem Scatterplot der Residuen auf der Y-Achse und den vorhergesagten Werten auf der X-Achse überprüft werden. Bei jedem X-Wert sollte die Varianz der Y-Werte (etwa) gleich sein (s. Abbildung 10.6).

Die geschätzten (angepassten) Werte kann man über den Befehl `add_predictions()` aus dem Paket `modelr` bekommen. Die Fehlerwerte entsprechend mit dem Befehl `add_residuals()`.

```
stats_test %>%
  add_predictions(mein_lm) %>%
  add_residuals(mein_lm) %>%
  ggplot() +
  aes(y = resid, x = pred) +
  geom_point()
```

Die Annahme der konstanten Varianz scheint verletzt zu sein: Die sehr großen vorhersagten Werte können recht genau geschätzt werden; aber die mittleren Werte nur ungenau. Die Verletzung dieser Annahme beeinflusst *nicht* die Schätzung der Steigung, sondern die Schätzung

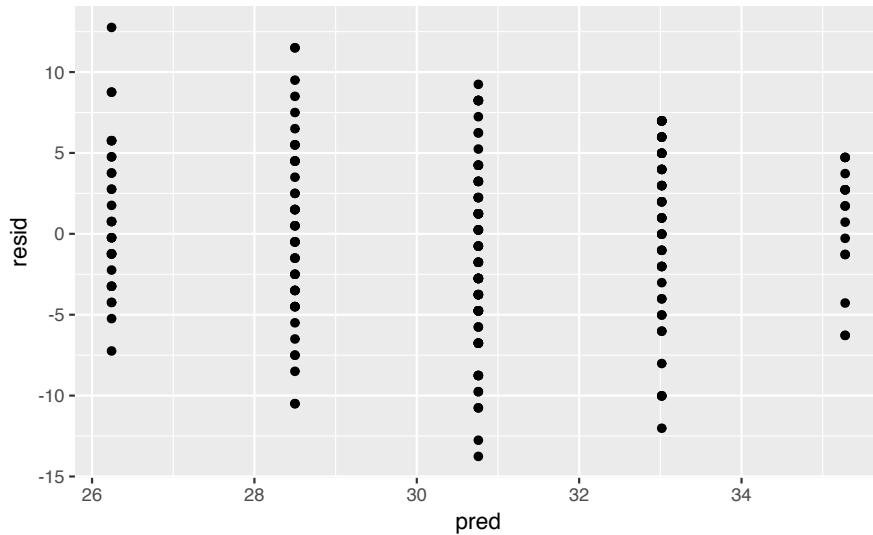


Abbildung 10.6: Vorhergesagte Werte vs. Residualwerte im Datensatz tips

des Standardfehlers, also des p-Wertes der Einflusswerte.

- *Extreme Ausreißer:* Extreme Ausreißer scheint es nicht zu geben.
- *Unabhängigkeit der Beobachtungen:* Wenn die Studenten in Lerngruppen lernen, kann es sein, dass die Beobachtungen nicht unabhängig voneinander sind: Wenn ein Mitglied der Lerngruppe gute Noten hat, ist die Wahrscheinlichkeit für ebenfalls gute Noten bei den anderen Mitgliedern der Lerngruppe erhöht. Böse Zungen behaupten, dass ‘Abschreiben’ eine Gefahr für die Unabhängigkeit der Beobachtungen sei.



1. Wie groß ist der Einfluss des Interessss?
2. Für wie aussagekräftig halten Sie Ihr Ergebnis aus 1.?
3. Welcher Einflussfaktor (in unseren Daten) ist am stärksten?

10.5 Regression mit kategorialen Prädiktoren

Vergleichen wir interessierte und nicht interessierte Studenten. Dazu teilen wir die Variable `interest` in zwei Gruppen (1-3 vs. 4-6) auf:

```
stats_test$interessiert <- stats_test$interest > 3
```

Vergleichen wir die Mittelwerte des Klausurerfolgs zwischen den Interessierten und Nicht-Interessierten:

```
stats_test %>%
  group_by(interessiert) %>%
  summarise(score = mean(score)) -> score_interesse

score_interesse
#> # A tibble: 3 x 2
#>   interessiert score
#>   <lgl>     <dbl>
#> 1 FALSE      29.9
#> 2 TRUE       31.5
#> 3 NA        33.1
```

Aha, die Interessierten haben im Schnitt mehr Punkte; aber nicht viel.

```
stats_test %>%
  na.omit %>%
  ggplot() +
  aes(x = interessiert, y = score) +
  geom_jitter(width = .1) +
  geom_point(data = score_interesse, color = "red", size = 5) +
  geom_line(data = score_interesse, group = 1, color = "red")
```



Mit `group=1` bekommt man eine Linie, die alle Punkte verbindet (im Datensatz `score_interesse` sind es dieser zwei). Wir haben in dem Fall nur zwei Punkte, die entsprechend verbunden werden.

10.5.1 Aufgaben

1. Visualisieren Sie den Gruppenunterschied auch mit einem Boxplot!
2. Berechnen Sie ein lineares Modell dazu!

Lösung:

1. Boxplot: `qplot(x = interessiert, y = score, data = stats_test, geom = "boxplot")`
2. Lineares Modell:

```
lm2 <- lm(score ~ interessiert, data = stats_test)
summary(lm2)
#>
#> Call:
#> lm(formula = score ~ interessiert, data = stats_test)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -13.537  -4.380  -0.537   4.463  10.091
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 29.909     0.475   62.99  <2e-16 ***
#> interessiertTRUE 1.628     0.752    2.17   0.031 *
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 5.68 on 236 degrees of freedom
#> (68 observations deleted due to missingness)
#> Multiple R-squared:  0.0195, Adjusted R-squared:  0.0153
#> F-statistic: 4.69 on 1 and 236 DF,  p-value: 0.0313
```

Der Einfluss von `interessiert` ist statistisch signifikant ($p = .03$). Der Stärke des Einflusses ist im Schnitt 1.6 Klausurpunkte (zugunsten `interessiertTRUE`). Das ist genau, was wir oben herausgefunden haben.



3. Wie ist der Einfluss von `study_time`, auch in zwei Gruppen geteilt?
4. Wie viel % der Variation des Klausurerfolgs können Sie durch das Interesse modellieren?

10.6 Multiple Regression

Aber wie wirken sich mehrere Einflussgrößen *zusammen* auf den Klausurerfolg aus?

```
lm3 <- lm(score ~ study_time + interessiert, data = stats_test)
summary(lm3)
#>
#> Call:
#> lm(formula = score ~ study_time + interessiert, data = stats_test)
#>
#> Residuals:
#>    Min      1Q  Median      3Q     Max
#> -13.896 -3.577  0.418  3.805 13.065
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 23.955     0.938   25.55 < 2e-16 ***
#> study_time    2.314     0.323    7.15  1.1e-11 ***
#> interessiertTRUE -0.333     0.736   -0.45     0.65
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 5.16 on 235 degrees of freedom
#>   (68 observations deleted due to missingness)
#> Multiple R-squared:  0.195, Adjusted R-squared:  0.188
#> F-statistic: 28.4 on 2 and 235 DF, p-value: 8.81e-12
```

Interessant ist das *negative* Vorzeichen vor dem Einfluss von `interessiertTRUE`! Die multiple Regression untersucht den ‘Nettoeinfluss’ jedes Prädiktors. Den Einfluss also, wenn der andere Prädiktor *konstant* gehalten wird. Anders gesagt: Betrachten wir jeden Wert von `study_time` separat, so haben die Interessierten jeweils im Schnitt etwas *weniger* Punkte (jesses). Allerdings ist dieser Unterschied nicht statistisch signifikant.

Die multiple Regression zeigt den ‘Nettoeinfluss’ jedes Prädiktor: Den Einfluss dieses Prädiktor, wenn der andere Prädiktor oder die anderen Prädiktoren konstant gehalten werden.

Hier haben wir übrigens dem Modell aufgezwungen, dass der Einfluss von Lernzeit auf Klausurerfolg bei den beiden Gruppen gleich groß sein soll (d.h. bei Interessierten und Nicht-Interessierten ist die Steigung der Regressionsgeraden gleich). Das illustriert sich am einfachsten in einem Diagramm (s. Abbildung 10.7).

Diese *multivariate* Analyse (mehr als 2 Variablen sind beteiligt) zeigt uns, dass die Regressionsgerade nicht gleich ist in den beiden Gruppen (Interessierte vs. Nicht-Interessierte; s. Abbildung 10.7): Im Teildiagramm A sind die Geraden (leicht) versetzt. Analog zeigt Teildia-

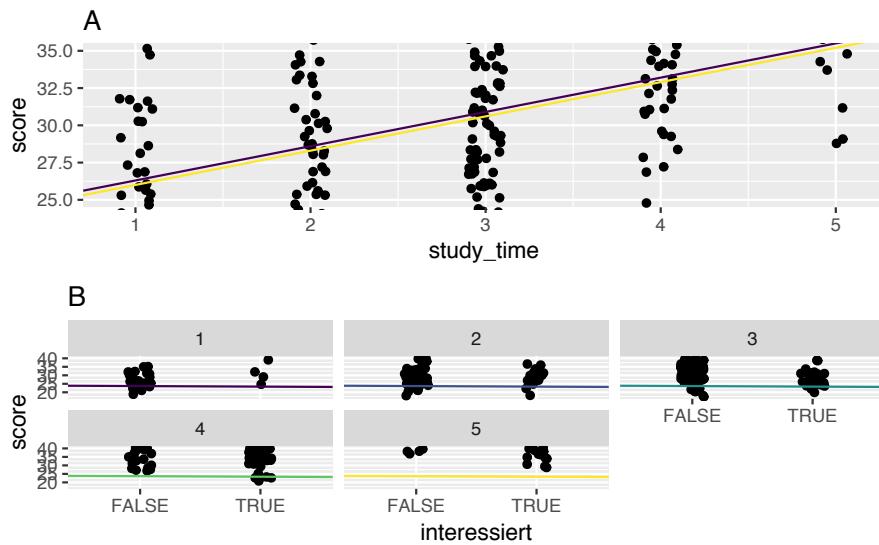


Abbildung 10.7: Eine multivariate Analyse fördert Einsichten zu Tage, die bei einfacheren Analysen verborgen bleiben

gramm B, dass die Interessierten (`interessiert == TRUE`) geringe Punktwerte haben als die Nicht-Interessierten, wenn man die Werte von `study_time` getrennt betrachtet.

Die multivariate Analyse zeigt ein anderes Bild, ein genaueres Bild als die einfache Analyse. Ein Sachverhalt, der für den ganzen Datensatz gilt, kann in Subgruppen anders sein.

Ohne multivariate Analyse hätten wir dies nicht entdeckt. Daher sind multivariate Analysen sinnvoll und sollten gegenüber einfacheren Analysen bevorzugt werden.

Man könnte sich jetzt noch fragen, ob die Regressionssgerade in Abbildung 10.7 parallel sein müssen. Gerade hat unser R-Befehl sie noch gezwungen, parallel zu sein. Gleich lassen wir hier die Zügel locker. Wenn die Regressionsgeraden nicht mehr parallel sind, spricht man von *Interaktionseffekten*.

Das Ergebnis des zugrunde-liegenden F-Tests (vgl. Varianzanalyse) wird in der letzten Zeile angegeben (**F-Statistic**). Hier wird H_0 also verworfen.

10.7 Interaktionen

Es könnte ja sein, dass die Stärke des Einflusses von Lernzeit auf Klausurerfolg in der Gruppe der Interessierten anders ist als in der Gruppe der Nicht-Interessierten. Wenn man nicht interessiert ist, so könnte man argumentieren, dann bringt eine Stunden Lernen weniger als wenn man interessiert ist. Darum müssten die Steigungen der Regressionsgeraden in den beiden Gruppen unterschiedlich sein. Schauen wir uns es an. Um R dazu zu bringen, die Regressionsgeraden frei variieren zu lassen, so dass sie nicht mehr parallel sind, nutzen wir das Symbol *, dass wir zwischen die betreffenden Prädiktoren schreiben:

```

lm4 <- lm(score ~ interessiert*study_time, data = stats_test)
summary(lm4)
#>
#> Call:
#> lm(formula = score ~ interessiert * study_time, data = stats_test)
#>
#> Residuals:
#>    Min      1Q  Median      3Q     Max
#> -13.950  -3.614   0.356   4.020  12.598
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  23.627    1.158   20.40 < 2e-16 ***
#> interessiertTRUE 0.655    2.170   0.30    0.76
#> study_time   2.441    0.418   5.85  1.7e-08 ***
#> interessiertTRUE:study_time -0.321    0.662   -0.48    0.63
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 5.17 on 234 degrees of freedom
#>   (68 observations deleted due to missingness)
#> Multiple R-squared:  0.196, Adjusted R-squared:  0.185
#> F-statistic:  19 on 3 and 234 DF,  p-value: 4.81e-11

```

Interessanterweise zeigen die Interessierten nun wiederum - betrachtet man jede Stufe von `study_time` einzeln - bessere Klausurergebnisse als die Nicht-Interessierten. Ansonsten ist noch die Zeile `interessiertTRUE:study_time` neu. Diese Zeile zeigt die Höhe des *Interaktionseffekts*. Bei den Interessierten ist die Steigung der Geraden um 0.32 Punkte geringer als bei den Nicht-Interessierten. Der Effekt ist klein und nicht statistisch signifikant, so dass wir wahrscheinlich Zufallsrauschen überinterpretieren. Aber die reine Zahl sagt, dass bei den Interessierten jede Lernstunde weniger Klausurerfolg bringt als bei den Nicht-Interessierten. Auch hier ist eine Visualisierung wieder hilfreich.

Wir sehen in Abbildung 10.8, dass der Einfluss von `study_time`` je nach Gruppe (Wert von `interessiert`) unterschiedlich (Teildiagramm A). Analog ist der Einfluss des Interesses (leicht) unterschiedlich, wenn man die fünf Stufen von `study_time`` getrennt betrachtet.

Sind die Regressionsgerade nicht parallel, so liegt ein Interaktionseffekt vor. Andernfalls nicht.

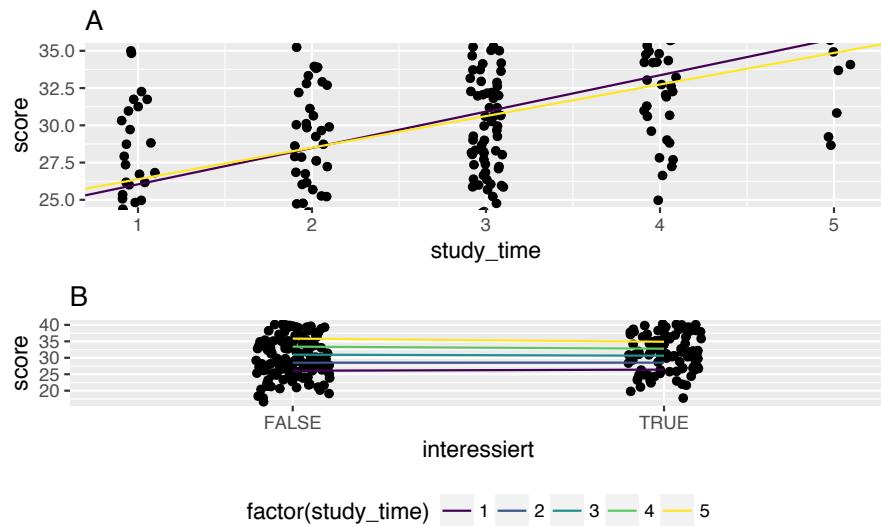


Abbildung 10.8: Eine Regressionsanalyse mit Interaktionseffekten

10.8 Fallstudie zu Overfitting

Vergleichen wir im ersten Schritt eine Regression, die die Modellgüte anhand der *Trainingsstichprobe* schätzt mit einer Regression, bei der die Modellgüte in einer *Test-Stichprobe* überprüft wird.

Betrachten wir nochmal die einfache Regression von oben. Wie lautet das R^2 ?

```
lm1 <- lm(score ~ study_time, data = stats_test)
```

Es lautet `round(summary(lm1)$r.squared, 2)`.

Im zweiten Schritt teilen wir die Stichprobe in eine Trainings- und eine Test-Stichprobe auf. Wir “trainieren” das Modell anhand der Daten aus der Trainings-Stichprobe:

```
train <- stats_test %>%
  sample_frac(.8, replace = FALSE) # Stichprobe von 80%, ohne Zurücklegen

test <- stats_test %>%
  anti_join(train) # Alle Zeilen von "df", die nicht in "train" vorkommen

lm_train <- lm(score ~ study_time, data = train)
```

Dann testen wir (die Modellgüte) anhand der *Test-Stichprobe*. Also los, `lm_train`, mach Deine Vorhersage:

```
lm2_predict <- predict(lm_train, newdata = test)
```

Diese Syntax sagt:



Speichere unter dem Namen “lm2_predict” das Ergebnis folgender Berechnung:
Mache eine Vorhersage (“to predict”) anhand des Modells “lm2”,
wobei frische Daten (“newdata = test”) verwendet werden sollen.

Als Ergebnis bekommen wir einen Vektor, der für jede Beobachtung des Test-Samples den geschätzten (vorhergesagten) Klausurpunktewert speichert.

```
caret::postResample(pred = lm2_predict, obs = test$score)
#>      RMSE Rsquared
#>    4.331    0.345
```

Die Funktion `postResample` aus dem Paket `caret` liefert uns zentrale Gütekennzahlen unser Modell. Wir sehen, dass die Modellgüte im Test-Sample deutlich *schlechter* ist als im Trainings-Sample. Ein typischer Fall, der uns warnt, nicht vorschnell optimistisch zu sein!

Die Modellgüte im in der Test-Stichprobe ist meist schlechter als in der Trainings-Stichprobe. Das warnt uns vor Befunden, die naiv nur die Werte aus der Trainings-Stichprobe berichten.

10.9 Aufgaben¹



Richtig oder Falsch!?

1. X-Wert: Kriterium; Y-Wert: Prädiktor.
2. Der Y-Wert in der einfachen Regression wird berechnet als Achsenabschnitt plus x mal die Geradensteigung.
3. R^2 liefert einen *relativen* Vorhersagefehler und MSE einen *absoluten* (relativ im Sinne eines Anteils).
4. Unter ‘Ordinary Least Squares’ versteht man eine abschätzige Haltung gegenüber Statistik.
5. Zu den Annahmen der Regression gehört Normalverteilung der *Kriteriumswerte*.
6. Die Regression darf nicht bei kategorialen Prädiktoren verwendet werden.

¹F, R, R, F, F, F, F, F, R

Tabelle 10.1: Befehle des Kapitels 'Regression'

Paket..Funktion	Beschreibung
lm	Berechnet eine Regression
sqrt	Zieht die Quadratwurzel
caret::postResample	Berechnet Gütekriterien für das Testsample
summary	Fasst zentrale Informationen zu einem Objekt zusammen
modelr::add_residuals	Erstellt eine Spalte mit Residuen
modelr::add_predictions	Erstellt eine Spalte mit den vorhergesagten Werten
levels	Zeigt oder ändert die Stufen eines Faktors
factor	Erstellt einen Faktor (nominalskalierte Variable)
coef	Zeigt die Koeffizienten eines Objekts an.
step	Führt eine Schrittweise-Rückwärtsselektion durch
sample_frac	Sampelt einen Prozentsatz aus einem Datensatz
anti_join	Fügt Vereint nicht-gleiche Zeilen zweier Datensätze

7. Mehrere bivariate Regressionsanalysen (1 Prädiktor, 1 Kriterium) sind einer multivariaten Regression i.d.R. vorzuziehen.
8. Interaktionen erkennt man daran, dass die Regressionsgeraden *nicht* parallel sind.

10.10 Befehlsübersicht

Tabelle 10.1 stellt die Befehle dieses Kapitels dar.

Kapitel 11

Klassifizierende Regression



Lernziele:

- Die Idee der logistischen Regression verstehen.
- Die Koeffizienten der logistischen Regression interpretieren können.
- Die Modellgüte einer logistischen Regression einschätzen können.
- Klassifikatorische Kennzahlen kennen und beurteilen können.

Für dieses Kapitel benötigen Sie folgende Pakete:

```
library(SDMTools) # Güte von Klassifikationsmodellen
library(pROC) # für ROC- und AUC-Berechnung
library(tidyverse) # Datenjudo
library(BaylorEdPsych) # Pseudo-R-Quadrat
library(broom) # lm-Ergebnisse aufräumen
```

Hilft Lernen, eine Statistikklausur zu bestehen? Kommmt es auf Interesse an? Versuchen wir vorherzusagen, wer eine Statistikklausur besteht. Etwas genauer gesagt, sagen wir ein *binäres*

(dichotomes) Ereignis - Bestehen der Klausur - vorher anhand von einer mehr Variablen mit beliebigen Skalenniveau.

Laden wir die Klausurdaten.

```
stats_test <- read.csv("data/stats_test.csv")
stats_test <- na.omit(stats_test)
```

Um uns das Leben leichter zu machen, haben wir fehlende Werte (NAs) mit `na.omit` gelöscht.

11.1 Normale Regression für ein binäres Kriterium

In gewohnter Manier nutzen wir die normale Regression um das Kriterium ‘Bestehen’ anhand der Vorbreitungszeit vorherzusagen. Mit einem kleinen Trick können wir die binäre Variable `bestanden` in eine Art metrische Variable umwandeln, damit sie wieder in unser Regressions-Handwerk passt: Wenn `bestanden=="ja"` dann sei `bestanden_num = 1`; ansonsten `bestanden_num = 0`. Dieses ‘wenn-dann’ leistet der Befehl `if_else(if_else(bedingung, wenn_erfüllt, ansonsten))`. In unserem Fall sieht das so aus:

```
stats_test %>%
  mutate(bestanden_num = if_else(bestanden == "ja", 1, 0)) -> stats_test
```

Rechnen wir jetzt unsere Regression:

```
lm1 <- lm(bestanden_num ~ study_time, data = stats_test)
tidy(lm1)
#>   term estimate std.error statistic p.value
#> 1 (Intercept)  0.6465    0.0654     9.89 1.61e-19
#> 2 study_time   0.0666    0.0210     3.17 1.70e-03
```

Hm. Stellen wir das Ergebnis grafisch dar (vgl. Abbildung 11.1).

```
ggplot(stats_test) +
  aes(x = study_time, y = bestanden_num) +
  geom_jitter() +
  geom_abline(slope = .07, intercept = .65, color = "red")
```

Ah! Mehr Lernen hilft offenbar: Die Regressionsgerade steigt.

Betrachten Sie diese praktische Eigenschaft der Regression: Obwohl die Kriteriumsvariable (Y-Achse) nur zwei Ausprägungen aufweist (0 und 1), sagt sich die Regression: “Hey, 0 und 1 sind normale reelle Zahlen und zwischen jedem solcher Zahlenpaare gibt es Zahlen

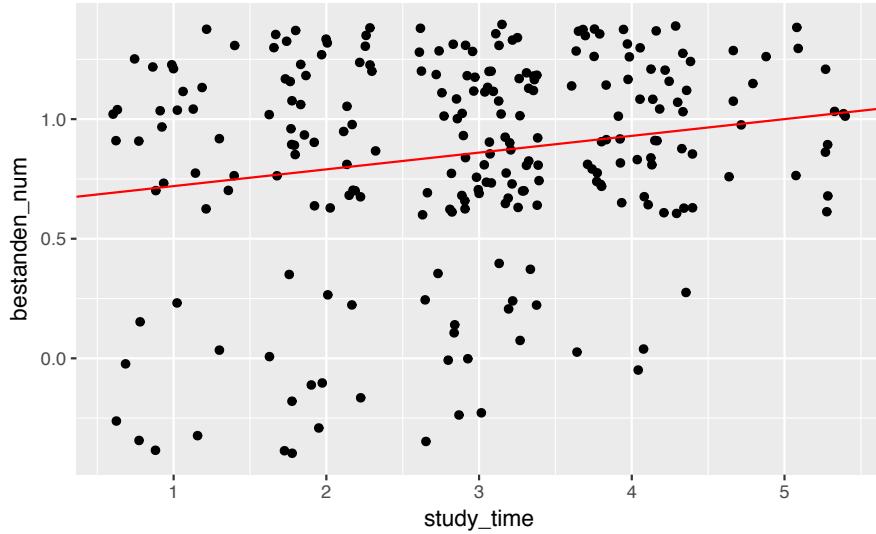


Abbildung 11.1: Regressionsgerade für das Bestehen-Modell

dazwischen. Also kann ich meine Regressionengerade ohne abzusetzen durchmalen". Damit können wir die Werte zwischen 0 und 1 wie Wahrscheinlichkeiten interpretieren: Sagt die Regressionsgerade für bestimmte Prädiktorwerte hohe Kriteriumswerte voraus, so können wir sagen, die Wahrscheinlichkeit, die Klausur zu bestehen, ist hoch.

Soweit, so gut. Aber Moment. Was bedeutet es, wenn die Wahrscheinlichkeit größer 1 ist? Dass der Professor vorher einen eidesstattliche Erklärung für Bestehen geschickt hat? Von so etwas hat man noch nicht gehört... Kurz gesagt: Wahrscheinlichkeiten größer 1 und kleiner 0 sind Quatsch. Wahrscheinlichkeiten müssen zwischen 0 und 1 liegen.

11.2 Die logistische Funktion

Daher brauchen wir eine Funktion, die das Ergebnis einer linearen Regression in einen Bereich von 0 bis 1 "umbiegt" (die sogenannte *Linkfunktion*). Eine häufig dafür verwendete Funktion ist die *logistische Funktion*. Im einfachsten Fall:

$$p(y = 1) = \frac{e^x}{1 + e^x} = \frac{e^x}{e^x(\frac{1}{e^x} + 1)} = \frac{1}{\frac{1}{e^x} + 1} = \frac{1}{e^{-x} + 1}$$

Exemplarisch können wir die logistische Funktion für einen Bereich von $x = -10$ bis $x = +10$ darstellen (vgl. 11.2). Der Graph der logistischen Funktion ähnelt einem langgestreckten S ("Ogive" genannt).

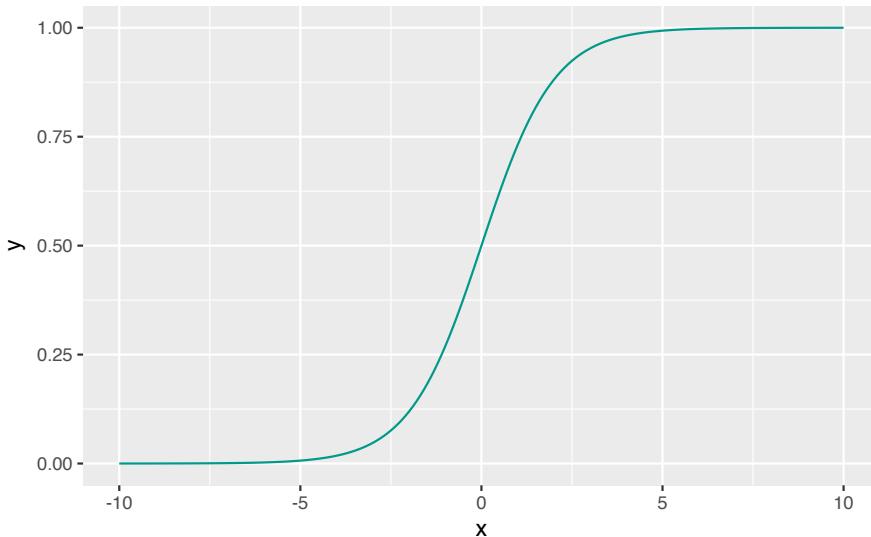


Abbildung 11.2: Die logistische Regression beschreibt eine 's-förmige' Kurve

11.3 Die Idee der logistischen Regression

Die logistische Regression ist eine Anwendung des *Allgemeinen Linearen Modells* (general linear model, GLM). Die Modellgleichung lautet:

$$p(y_i = 1) = L(\beta_0 + \beta_1 \cdot x_{i1} + \cdots + \beta_K \cdot x_{ik}) + \epsilon_i$$

- L ist die Linkfunktion, in unserer Anwendung die logistische Funktion.
- x_{ik} sind die beobachteten Werte der unabhängigen Variablen X_k .
- k sind die unabhängigen Variablen 1 bis K .

Die Funktion `glm` führt die logistische Regression durch.

```
glm1 <- glm(bestanden_num ~ study_time,
              family = "binomial",
              data = stats_test)
```

Wir schauen uns zunächst den Plot an (Abb. 11.3).

Es werden ein Streudiagramm der beobachteten Werte sowie die *Regressionslinie* ausgegeben. Wir können so z. B. ablesen, dass mit einer Lernzeit von 5 die Wahrscheinlichkeit für Bestehen bei knapp 100% liegt; viel zu einfach. . .

Die Zusammenfassung des Modells zeigt folgendes:

```
tidy(glm1)
#>   term estimate std.error statistic p.value
```

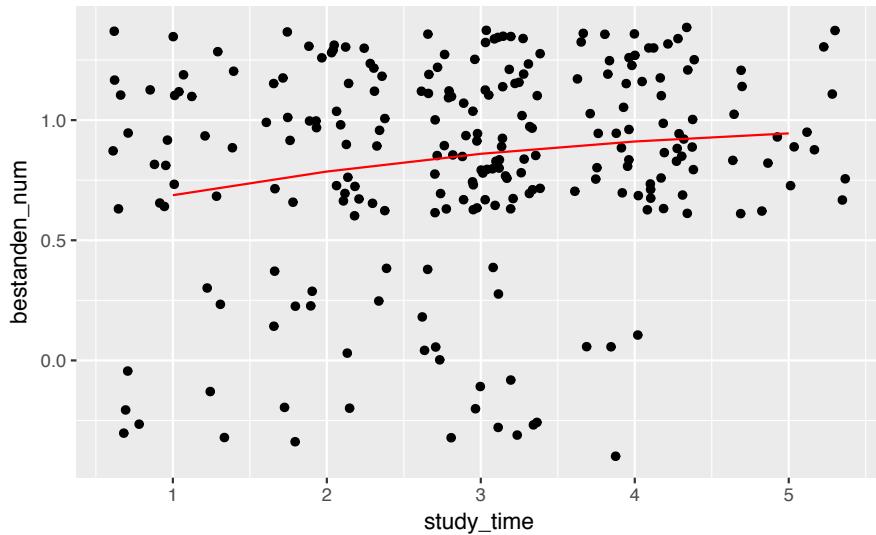


Abbildung 11.3: Modelldiagramm mit logistischer Regression

```
#> 1 (Intercept) 0.276      0.458      0.602 0.54698
#> 2 study_time   0.513      0.168      3.049 0.00229
```

Die p-Werte der Koeffizienten können in der Spalte $\text{Pr}(>|z|)$ abgelesen werden. Der Achsenabschnitt (`intercept`) wird mit 0.28 geschätzt, die Steigung in Richtung `study_time` mit 0.51. Allerdings sind die hier dargestellten Werte sogenannte *Logits* \mathcal{L} ¹:

$$\mathcal{L} = \ln\left(\frac{p}{1-p}\right)$$

Zugeben, dass klingt erstmal opaque. Das Praktische ist, dass wir die Koeffizienten in Logitform in gewohnter Manier verrechnen dürfen. Wollen wir zum Beispiel wissen, wie wahrscheinlich das Ereignis ‘Bestehen’ für eine Person mit einer Lernzeit von 3 ist, können wir einfach rechnen:

`y = intercept + 3*study_time`, also

```
(y <- .27 + 3 * 0.51)
#> [1] 1.8
```

Einfach, oder? Genau wie bei der normalen Regression. Aber beachten Sie, dass das Ergebnis in *Logits* angegeben ist. Was ein *Logit* ist? Naja, das ist der *Logarithmus der Chancen*; unter ‘Chancen’ versteht man den Quotienten von Wahrscheinlichkeit p zur Gegenwahrscheinlichkeit, $1 - p$; die Chancen werden auch *Odds* oder *Wettquotient* genannt.

Um zur ‘normalen’ Wahrscheinlichkeit zu kommen, muss man also erst ‘delogarithmieren’. Delogarithmieren bedeutet, die e-Funktion anzuwenden, `exp` auf Errisch:

¹ein schnödes L wie in Ludwig

```
exp(y)
#> [1] 6.05
```

Jetzt haben wir wir also Chancen. Wie rechnet man Chancen in Wahrscheinlichkeiten um? Ein Beispiel zur Illustration. Bei Prof. Schnaggeldi fallen von 10 Studenten 9 durch. Die Durchfallchance ist also 9:1 oder 9. Die Durchfallwahrscheinlichkeit 9/10 oder .9. Also kann man so umrechnen:

```
wskt = 9 / (9+1) = 9/10 = .9.
```

In unserem Fall sind die Chancen etwa 6:1; also lautet die Umrechnung:

```
(wskt <- 6 / (6+1))
#> [1] 0.857
```

Diesen Ritt kann man sich merklich kommorder bereiten, wenn man diesen Befehl kennt:

```
predict(glm1, newdata = data.frame(study_time = 3), type = "response")
#>     1
#> 0.86
```

11.4 Kein R^2 , dafür AIC

Es gibt kein R^2 im Sinne einer erklärten Streuung der y -Werte, da die beobachteten y -Werte nur 0 oder 1 annehmen können. Das Gütemaß bei der logistischen Regression ist das *Akaike Information Criterion (AIC)*. Hier gilt allerdings: je *kleiner*, desto *besser*. (Anmerkung: es kann ein Pseudo- R^2 berechnet werden – kommt später.) Richtlinien, was ein “guter” AIC-Wert ist, gibt es nicht. Diese Werte helfen nur beim Vergleichen von Modellen.

11.5 Interpretation der Koeffizienten

Ist ein Logit \mathcal{L} größer als 0, so ist die zugehörige Wahrscheinlichkeit größer als 50% (und umgekehrt.)

11.5.1 y-Achsenabschnitt (Intercept) β_0

Für $\beta_0 > 0$ gilt, dass selbst wenn alle anderen unabhängigen Variablen 0 sind, es eine Wahrscheinlichkeit von mehr als 50% gibt, dass das modellierte Ereignis eintritt. Für $\beta_0 < 0$ gilt entsprechend das Umgekehrte.

11.5.2 Steigung β_i mit $i = 1, 2, \dots, K$

Für $\beta_i > 0$ gilt, dass mit zunehmenden x_i die Wahrscheinlichkeit für das modellierte Ereignis steigt. Bei $\beta_i < 0$ nimmt die Wahrscheinlichkeit entsprechend ab.

11.5.3 Aufgabe

Berechnen Sie den Zuwachs an Wahrscheinlichkeit für unser Beispielmodell, wenn sich die `study_time` von 1 auf 2 erhöht. Vergleichen Sie das Ergebnis mit der Punktprognose für `study_time= 7` im Vergleich zu `study_time= 8`.

Lösung:

```
# aus Koeffizient abgeschätzt
wskt1 <- predict(glm1, data.frame(study_time = 1), type = "response")

wskt2 <- predict(glm1, data.frame(study_time = 2), type = "response")

wskt2 - wskt1
#>      1
#> 0.0985
```

Anders gesagt: “Mit jedem Punkt mehr ‘study_time’ steigt der Logit (die logarithmierten Chancen) für Bestehen um 0.513”.

```
# mit dem vollständigen Modell berechnet
predict(glm1, data.frame(study_time = 1),
       type = "response")
#>      1
#> 0.688

predict(glm1, data.frame(study_time = 8),
       type = "response")
#>      1
#> 0.988
```

Bei einer `study_time` von 4 beträgt die Wahrscheinlichkeit für $y = 1$, d.h. für das Ereignis ‘Bestehen’, 0.91. Bei einer `study_time` von 58 liegt diese Wahrscheinlichkeit bei 0.94.

11.6 Kategoriale Prädiktoren

Wie in der linearen Regression können auch in der logistischen Regression kategoriale Variablen als unabhängige Variablen genutzt werden.

Betrachten wir als Beispiel die Frage, ob die kategoriale Variable “Interessiert” (genauer: dichotome Variable) einen Einfluss auf das Bestehen in der Klausur hat, also die Wahrscheinlichkeit für Bestehen erhöht.

```
stats_test$interessiert <- stats_test$interest > 3
```

Erstellen Sie zum Aufwärmen ein passendes Diagramm!

Los geht's, probieren wir die logistische Regression aus:

```
glm2 <- glm(bestanden_num ~ interessiert,
             family = "binomial",
             data = stats_test)
tidy(glm2)
#>   term estimate std.error statistic p.value
#> 1 (Intercept) 1.50     0.217      6.94 4.00e-12
#> 2 interessiertTRUE 0.43     0.377      1.14 2.55e-01
```

Der Einflusswert (die Steigung) von `interessiert` ist positiv: Wenn man interessiert ist, steigt die Wahrscheinlichkeit zu bestehen. Gut. Aber wie groß ist die Wahrscheinlichkeit für jede Gruppe? Am einfachsten lässt man sich das von R ausrechnen:

```
predict(glm2, newdata = data.frame(interessiert = FALSE),
       type = "response")
#> 1
#> 0.818
predict(glm2, newdata = data.frame(interessiert = TRUE),
       type = "response")
#> 1
#> 0.874
```

Also 82% bzw. 87%; kein gewaltig großer Unterschied, aber immerhin...

11.7 Multiple logistische Regression

Können wir unser Model `glm1` mit nur einer erklärenden Variable verbessern, indem weiterer Prädiktoren hinzugefügt werden? Verbessern heißt hier: Können wir die Präzision der

Vorhersage verbessern durch Hinzunahme weiterer Prädiktoren?

```
glm3 <- glm(bestanden_num ~ study_time + interest + self_eval,
             family = binomial,
             data = stats_test)

tidy(glm3)
#>   term estimate std.error statistic p.value
#> 1 (Intercept) -0.202     0.550    -0.367 0.71359
#> 2 study_time    0.115     0.218     0.529 0.59665
#> 3 interest     -0.155     0.155    -0.998 0.31820
#> 4 self_eval      0.447     0.107     4.173 0.00003
```

Hm, die Interessierten schneiden jetzt - unter Konstanthalten anderer Einflussfaktoren - schlechter ab als die Nicht-Interessierten. Als Statistik-Dozent bin ich der Meinung, dieses Ergebnis sollte in der Schubladen verschwinden (wie es geläufige Praxis ist in vielen Laboren...).

11.8 Modellgüte

Aber wie gut ist das Modell? Und welches Modell von beiden ist besser? R hat uns kein R^2 ausgegeben. R hat uns deswegen kein R^2 ausgegeben, weil die Regressionsfunktion nicht über Abweichungsquadrate bestimmt wird. Stattdessen wird das Maximum Likelihood-Verfahren eingesetzt. Man kann also kein R^2 ausrechnen, zumindest nicht ohne Tricks. Einige findige Statistiker haben sich aber Umrechnungswege einfallen lassen, wie man auch ohne Abweichungsquadrate ein R^2 berechnen kann; weil es kein ‘echtes’ R^2 ist, nennt man es auch *Pseudo-R²*. Es gibt ein paar Varianten, wir bleiben bei der Variante von Herrn McFadden (s. Ausgabe).

Eine Reihe von R-Paketen bieten die Berechnung an:

```
library(BaylorEdPsych)
PseudoR2(glm1)
PseudoR2(glm2)
PseudoR2(glm3)
```

Die Ausgabe zeigt uns, dass das erste Modell schon schlecht ist, dass zweite praktisch keinen Erklärungswert und das dritte einen zumindest kleinen bis mittleren Erklärungswert bietet: $f^2 = \frac{R^2}{1-R^2} \approx \frac{1}{.9} = .11$.

Tabelle 11.1: Vier Arten von Ergebnisse von Klassifikationen

Wahrheit	Als negativ (-) vorhergesagt	Als positiv (+) vorhergesagt	Summe
In Wahrheit negativ (-)	Richtig negativ (RN)	Falsch positiv (FP)	N
In Wahrheit positiv (+)	Falsch negativ (FN)	Richtig positiv (RN)	P
Summe	N*	P*	N+P

11.9 Klassifikationskennzahlen

11.9.1 Vier Arten von Ergebnissen einer Klassifikation

Logistische Regressionsmodelle werden häufig zur *Klassifikation* verwendet. Das heißt man versucht, Beobachtungen richtig zu zu Klassen zuzuordnen:

- Ein medizinischer Test soll Kranke als krank und Gesunde als gesund klassifizieren.
- Ein statistischer Test sollte wahre Hypothesen als wahr und falsche Hypothesen als falsch klassifizieren.
- Ein Personaler sollte geeignete Bewerber als geeignet und nicht geeignete Bewerber als nicht geeignet einstufen.

Diese beiden Arten von Klassifikationen können unterschiedlich gut sein. Im Extremfall könnte ein Test alle Menschen als krank ('positiv') einstufen. Mit Sicherheit wurden dann alle Kranken korrekt als krank diagnostiziert. Dummerweise würde der Test 'auf der anderen Seite' viele Fehler machen: Gesunde als gesund ('negativ') zu klassifizieren.

Ein Test, der alle positiven Fälle korrekt als positiv klassifiziert muss deshalb noch lange nicht alle negativen Fälle als negativ klassifizieren. Die beiden Werte können unterschiedlich sein.

Etwas genauer kann man folgende vier Arten von Ergebnisse aus einem Test erwarten (s. Tabelle 11.1, vgl. James, Witten, Hastie, und Tibshirani (2013b)).

Die logistische Regression gibt uns für jeden Fall eine Wahrscheinlichkeit zurück, dass der Fall zum Ereignis 1 gehört. Wir müssen dann einen Schwellenwert (threshold) auswählen. Einen Wert also, der bestimmt, ob der Fall zum Ereignis 1 gehört. Häufig nimmt man 0.5. Liegt die Wahrscheinlichkeit unter dem Schwellenwert, so ordnet man den Fall dem Ereignis 0 zu.

Beispiel: Alois' Wahrscheinlichkeit, die Klausur zu bestehen, wird vom Regressionsmodell auf 51% geschätzt. Unser Schwellenwert sei 50%; wir ordnen Alois der Klasse "bestehen" zu. Alois freut sich. Das Modell sagt also "bestehen" (1) für Alois voraus. Man sagt auch, der 'geschätzte Wert' (*fitted value*) von Alois sei 1.

Die aus dem Modell ermittelten Wahrscheinlichkeiten werden dann in einer sogenannten Konfusionsmatrix (*confusion matrix*) mit den beobachteten Häufigkeiten verglichen:

Tabelle 11.2: Geläufige Kennwerte der Klassifikation

Name	Definition	Synonyme
Falsch-Positiv-Rate (FP-Rate)	FP/N	Alphafehler, Typ-1-Fehler, 1-Spezifität, Fehlalarm
Richtig-Positiv-Rate (RP-Rate)	RP/N	Power, Sensitivität, 1-Betafehler, Recall
Falsch-Negativ-Rate (FN-Rate)	FN/N	Fehlender Alarm, Befafehler
Richtig-Negativ-Rate (RN-Rate)	RN/N	Spezifität, 1-Alphafehler
Positiver Vorhersagewert	RP/P*	Präzision, Relevanz
Negativer Vorhersagewert	RN/N*	Segreganz
Gesamtgenauigkeitsrate	(RP+RN) / (N+P)	Richtigkeit, Korrektklassifikationsrate

```
(cm <- confusion.matrix(stats_test$bestanden_num, glm3$fitted.values))
#>      obs
#> pred  0   1
#>     0   1   1
#>     1 37 199
#> attr(,"class")
#> [1] "confusion.matrix"
```

Dabei stehen `obs` (observed) für die wahren, also tatsächlich beobachteten Werte und `pred` (predicted) für die geschätzten (vorhergesagten) Werte.

Wie häufig hat unser Modell richtig geschätzt? Genauer: Wie viele echte 1 hat unser Modell als 1 vorausgesagt und wie viele echte 0 hat unser Modell als 0 vorausgesagt?

11.9.2 Klassifikationsgütekennzahlen

In der Literatur und Praxis herrscht eine recht wilde Vielfalt an Begriffen dazu, deswegen stellt Tabelle 11.1 einen Überblick vor.

Zu beachten ist, dass die Gesamtgenauigkeit einer Klassifikation an sich wenig aussagekräftig ist: Ist eine Krankheit sehr selten, werde ich durch die einfache Strategie “diagnostiziere alle als gesund” insgesamt kaum Fehler machen. Meine Gesamtgenauigkeit wird beeindruckend genau sein - trotzdem lassen Sie sich davon wohl kaum beeindrucken. Besser ist, die Richtig-Positiv- und die Richtig-Negativ-Raten getrennt zu beurteilen. Aus dieser Kombination leitet sich der *Youden-Index* ab.. Er berechnet sich als: $\text{RP-Rate} + \text{RN-Rate} - 1$.

Sie können die Konfusionsmatrix mit dem Paket `confusion.matrix()` aus dem Paket `SDMTools` berechnen.

```
sensitivity(cm)
#> [1] 0.995
```

```
specificity(cm)
#> [1] 0.0263
```

Unser Modell hat es sich recht leicht gemacht: Es hat immer auf ‘bestanden’ getippt: Damit wurden alle ‘Besteher’ korrekt identifiziert (Sensitivität = 1); allerdings wurden auch alle ‘Nicht-Besteher’ übersehen (Spezifität = 0).

Wir könnten jetzt sagen, dass wir im Zweifel lieber eine Person als Nicht-Besteher einschätzen (um die Lernschwachen noch unterstützen zu können). Dazu würden wir den Schwellenwert (threshold) von 50% auf z.B. 80%\$ heraufsetzen. Erst bei Erreichen des Schwellenwerts klassifizieren wir die Beobachtung als ‘bestanden’ (1):

```
(cm <- confusion.matrix(stats_test$bestanden_num, glm3$fitted.values, threshold = .8))
#>      obs
#> pred  0   1
#>     0 24  47
#>     1 14 153
#> attr(,"class")
#> [1] "confusion.matrix"
sensitivity(cm)
#> [1] 0.765
specificity(cm)
#> [1] 0.632
```

11.9.3 ROC-Kurven

Siehe da! Die Spezifität ist gestiegen, wir haben mehr Nicht-Lerner als solche identifiziert. Unsere liberalere Strategie hat aber mehr Falsch-Negative Fälle produziert (geringere Sensitivität). So können wir jetzt viele verschiedene Schwellenwerte vergleichen.

Ein Test ist dann gut, wenn wir für alle möglichen Schwellenwert insgesamt wenig Fehler produziert.

Hierzu wird der Cutpoint zwischen 0 und 1 variiert und die Richtig-Positiv-Rate (Sensitivität) gegen die Falsch-Positiv-Rate (1–Spezifität) abgetragen. Das Paket pROC hilft uns hier weiter. Zuerst berechnen wir für viele verschiedene Schwellenwerte jeweils die beiden Fehler (Falsch-Positiv-Rate und Falsch-Negativ-Rate). Trägt man diese in ein Diagramm ab, so bekommt man Abbildung 11.4, eine sog. *ROC-Kurve*.

```
lets_roc <- roc(stats_test$bestanden_num, glm3$fitted.values)
```

Da die Sensitivität determiniert ist, wenn die Falsch-Positiv-Rate bekannt ist ($1 - FP\text{-Rate}$), kann man statt Sensitivität auch die FP-Rate abbilden. Für die Spezifität und die Falsch-

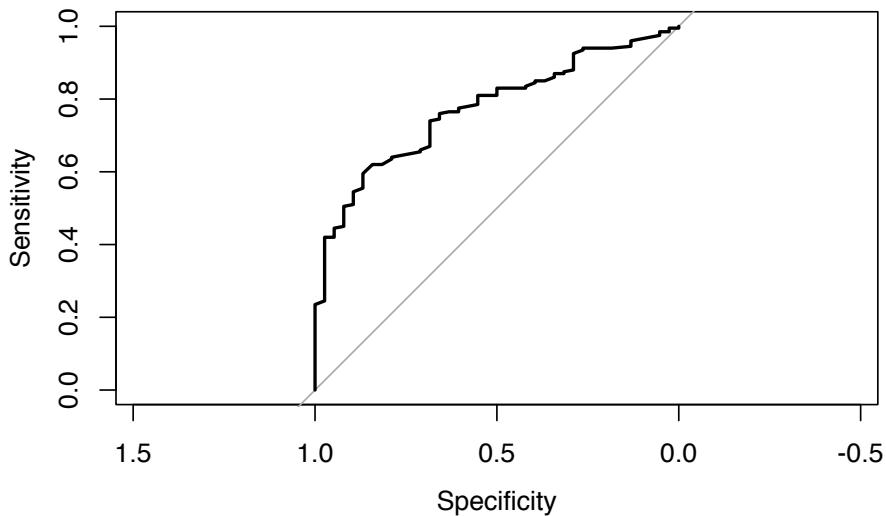


Abbildung 11.4: Eine ROC-Kurve

Negativ-Rate gilt das gleiche. In Abbildung 11.4 steht auf der X-Achse Spezifität, aber die Achse ist ‘rückwärts’ (absteigend) skaliert, so dass die X-Achse identisch ist mit FP-Rate (normal skaliert; d.h. aufsteigend).

```
plot(lets_roc)
```

Die ‘Fläche unter der Kurve’ (area under curve, AUC) ist damit ein Maß für die Güte des Tests. Abbildung 11.5 stellt drei Beispiele von Klassifikationsgüten dar: sehr gute (A), gute (B) und schlechte (C). Ein hohe Klassifikationsgüte zeigt sich daran, dass eine hohe Richtig-Positiv-Rate mit einer kleinen Fehlalarmquote einher geht: Wir finden alle Kranken, aber nur die Kranken. Die AUC-Kurve “hängt oben links an der Decke”. Ein schlechter Klassifikator trifft so gut wie ein Münzwurf: Ist das Ereignis selten, hat er eine hohe Falsch-Positiv-Rate und eine geringe Falsch-Negativ-Rate. Ist das Ereignis hingegen häufig, liegen die Fehlerhöhen genau umgekehrt: Eine hohe Richtig-Positiv-Rate wird mit einer hoher Falsch-Positiv-Rate einher.

Fragt sich noch, wie man den besten Schwellenwert herausfindet. Den besten Schwellenwert kann man als besten Youden-Index-Wert verstehen. Im Paket `pROC` gibt es dafür den Befehl `coords`, der uns im ROC-Diagramm die Koordinaten des besten Schwellenwert und den Wert dieses besten Schwellenwerts liefert:

```
coords(lets_roc, "best")
#>   threshold specificity sensitivity
#>       0.874      0.868      0.595
```

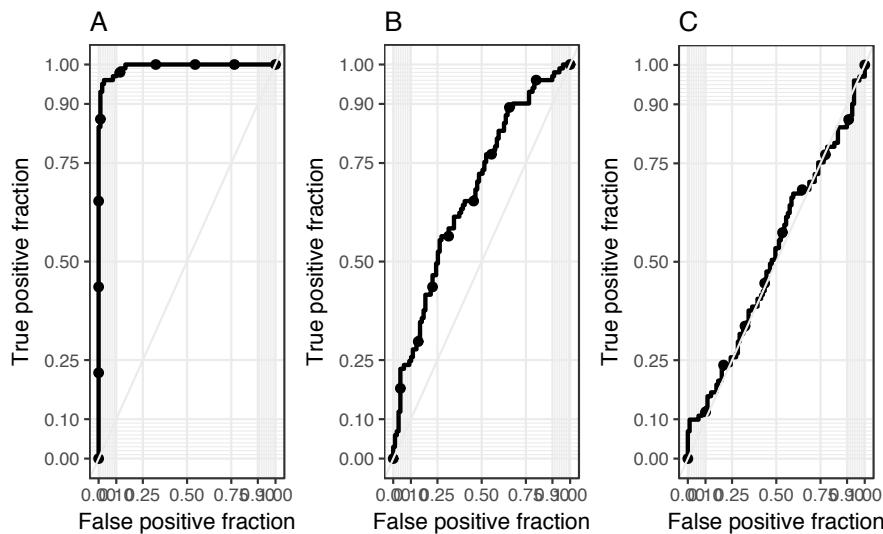


Abbildung 11.5: Beispiel für eine sehr gute (A), gute (B) und schlechte (C) Klassifikation

11.10 Aufgaben²



Richtig oder Falsch!?

1. Die logistische Regression ist eine Regression für dichotome Kriterien.
2. Unter einer OliveOgive versteht man eine eine “s-förmige” Kurve.
3. Berechnet man eine “normale” (OLS-)Regression bei einem dichotomen Kriterium, so kann man Wahrscheinlichkeiten < 0 oder > 1 erhalten, was keinen Sinn macht.
4. Ein Logit ist definiert als der Einfluss eines Prädiktors in der logistischen Regression. Der Koeffizient berechnet sich als Logarithmus des Wettquotienten.
5. Das AIC ein Gütemaß, welches man bei der logistischen Regression generell vermeidet.
6. Eine Klassifikation kann 4 Arten von Ergebnissen bringen - gemessen an der Richtigkeit des Ergebnisses.
7. Der ‘positive Vorhersagewert’ ist definiert als der Anteil aller richtig-positiven Klassifikationen an allen als positiv klassifizierten Objekten.

11.11 Befehlsübersicht

Tabelle 11.3 stellt die Befehle dieses Kapitels dar.

²R, R, R, R, F, R, R

Tabelle 11.3: Befehle des Kapitels 'Logistische Regression'

Paket..Funktion	Beschreibung
ggplot2::geom_abline	Fügt das Geom "abline" (normale Gerade) hinzu
glm	Berechnet eine logistische Regression
exp	Berechnet die e-Funktion
SDMTools::confusion.matrix	Berechnet eine Konfusionsmatrix
SDMTools::sensitivity	Berechnet die Sensitivität eines Klassifikationsmodells
SDMTools::specificity	Berechnet die Spezifität eines Klassifikationsmodells
ROCR::performance	Erstellt Objekte mit Gütekennzahlen von Klassifikationsmodellen
BaylorEdPsych::PseudoR2	Berechnet Pseudo-R-Quadrat-Werte

Kapitel 12

Fallstudien zum geleiteten Modellieren

In diesem Kapitel werden folgende Pakete benötigt.

```
library(tidyverse)  # Datenjudo
library(psych)    # Befehl 'describe'
library(broom)     # lm-Ergebnisse aufpolieren
library(corrplot)  # Korrelationstabellen visualisieren
library(titanic)   # Für Datensatz 'titanic'
library(compute.es) # Effektstärken berechnen
```

12.1 Überleben auf der Titanic

In dieser YACSDA¹ geht es um die beispielhafte Analyse nominaler Daten anhand des “klassischen” Falls zum Untergang der Titanic. Eine Frage, die sich hier aufdrängt, lautet: Kann (konnte) man sich vom Tod freikaufen, etwas polemisch formuliert. Oder neutraler: Hängt die Überlebensquote von der Klasse, in der der Passagiers reist, ab?

12.1.1 Daten laden

Mit dem Befehl `data` kann man Daten aus Paketen laden; lässt man den Parameter `package` weg, so werden alle geladenen Pakete nach diesem Datensatz durchsucht. Benennt man den Parameter, so kann man auch *nicht* geladene Pakete damit ansteuern.

¹Yet-another-case-study-on-data-analysis

```
data(titanic_train, package = "titanic")
titanic_train <- na.omit(titanic_train)
```

12.1.2 Erster Blick

Werfen Sie einen ersten Blick in die Daten mit `glimpse(titanic_train)`. Lassen Sie sich dann einige deskriptive Statistiken ausgeben²

12.1.3 Welche Variablen sind interessant?

Von 12 Variablen des Datensatzes interessieren uns offenbar `Pclass` und `Survived`; Hilfe zum Datensatz kann man übrigens mit `help(titanic_train)` bekommen. Diese beiden Variablen sind kategorial (nicht-metrisch), wobei sie in der Tabelle mit Zahlen kodiert sind. Natürlich ändert die Art der Codierung (hier als Zahl) nichts am eigentlichen Skalenniveau. Genauso könnte man “Mann” mit 1 und “Frau” mit 2 kodieren; ein Mittelwert bliebe genauso (wenig) aussagekräftig. Zu beachten ist hier nur, dass sich manche R-Befehle verunsichern lassen, wenn nominale Variablen mit Zahlen kodiert sind. Daher ist es oft besser, nominale Variablen mit Text-Werten zu benennen (wie “survived” vs. “drowned” etc.). Wir kommen später auf diesen Punkt zurück.

12.1.4 Univariate Häufigkeiten

Bevor wir uns in kompliziertere Fragestellungen stürzen, halten wir fest: Wir untersuchen zwei nominale Variablen. Sprich: wir werden Häufigkeiten auszählen. Häufigkeiten (und relative Häufigkeiten, also Anteile oder Quoten) sind das, was uns hier beschäftigt.

Zählen wir zuerst die univariaten Häufigkeiten aus: Wie viele Passagiere gab es pro Klasse? Wie viele Passagiere gab es pro Wert von `Survived` (also die überlebten bzw. nicht überlebten)?

```
c1 <- dplyr::count(titanic_train, Pclass)
c1
#> # A tibble: 3 x 2
#>   Pclass     n
#>   <int> <int>
#> 1     1    186
#> 2     2    173
#> 3     3    355
```

²z.B. mit `titanic_train %>% count(Survived)` oder `titanic_train %>% summarise(Ticketpreis = mean(Fare, na.rm = TRUE))`



Achtung - Namenskollision! Sowohl im Paket `mosaic` als auch im Paket `dplyr` gibt es einen Befehl `count`. Für `select` gilt Ähnliches - und für eine Reihe anderer Befehle auch. Das arme R weiß nicht, welchen von beiden wir meinen und entscheidet sich im Zweifel für den falschen. Da hilft, zu sagen, aus welchem Paket wir den Befehl beziehen wollen. Das macht der Operator `:::`. Probieren Sie die Funktion `find_funs` aus Kapitel 1.1.4, um herauszufinden, welche Pakete z.B. den Befehl `count` beherbergen.

Aha. Zur besseren Anschaulichkeit können Sie das auch plotten (ein Diagramm dazu malen). Wie?³

Der Befehl `qplot` zeichnet automatisch Punkte, wenn auf beiden Achsen “Zahlen-Variablen” stehen (also Variablen, die keinen “Text”, sondern nur Zahlen beinhalten. In R sind das Variablen vom Typ `int` (integer), also Ganze Zahlen oder vom Typ `num` (numeric), also reelle Zahlen).

```
c2 <- dplyr::count(titanic_train, Survived)
c2
#> # A tibble: 2 x 2
#>   Survived     n
#>   <int> <int>
#> 1     0    424
#> 2     1    290
```

Man beachte, dass der Befehl `count` steht eine Tabelle (data.frame bzw. `tibble`) verlangt und zurückliefert.

12.1.5 Bivariate Häufigkeiten

OK, gut. Jetzt wissen wir die Häufigkeiten pro Wert von `Survived` (dasselbe gilt für `Pclass`). Eigentlich interessiert uns aber die Frage, ob sich die relativen Häufigkeiten der Stufen von `Pclass` innerhalb der Stufen von `Survived` unterscheiden. Einfacher gesagt: Ist der Anteil der Überlebenden in der 1. Klasse größer als in der 3. Klasse?

Zählen wir zuerst die Häufigkeiten für alle Kombinationen von `Survived` und `Pclass`:

```
c3 <- dplyr::count(titanic_train, Survived, Pclass)
c3
#> # A tibble: 6 x 3
#>   Survived Pclass     n
#>   <int> <int> <int>
#> 1     0     1     64
```

³`qplot(x = Pclass, y = n, data = c1)`

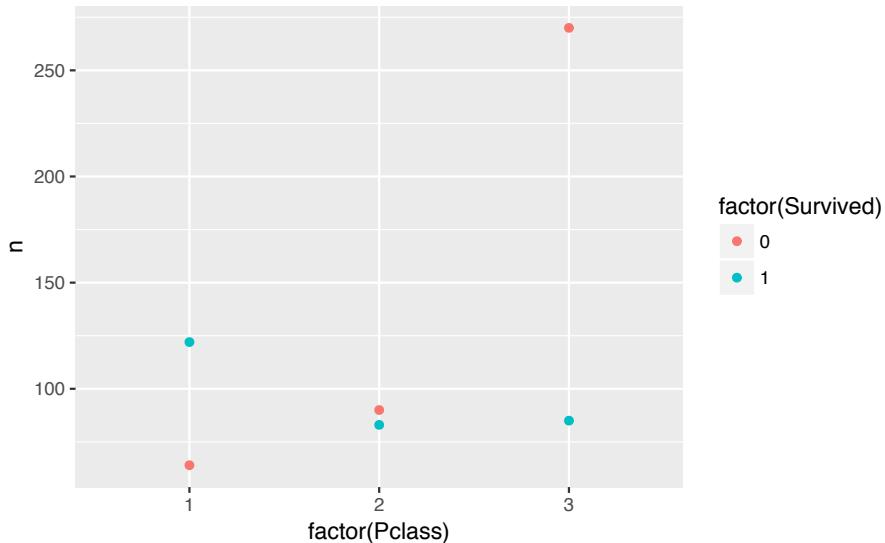


Abbildung 12.1: Überlebensraten auf der Titanic, in Abhängigkeit von der Passagierklasse

```
#> 2      0      2     90
#> 3      0      3    270
#> 4      1      1    122
#> 5      1      2     83
#> 6      1      3     85
```

Da `Pclass` 3 Stufen hat (1., 2. und 3. Klasse) und innerhalb jeder dieser 3 Klassen es die Gruppe der Überlebenden und der Nicht-Überlebenden gibt, haben wir insgesamt $3 \times 2 = 6$ Gruppen.

Es ist hilfreich, sich diese Häufigkeiten wiederum zu plotten; probieren Sie `qplot(x = Pclass, y = n, data = c3)`.

Hm, nicht so hilfreich. Schöner wäre, wenn wir (farblich) erkennen könnten, welcher Punkt für “Überlebt” und welcher Punkt für “Nicht-Überlebt” steht. Mit `qplot` geht das recht einfach: Wir sagen der Funktion `qplot`, dass die Farbe (`color`) der Punkte den Stufen von `Survived` zugeordnet werden sollen: `qplot(x = Pclass, y = n, color = Survived, data = c3)`.

Viel besser. Was noch stört, ist, dass `Survived` als metrische Variable verstanden wird. Das Farbschema lässt Nuancen, feine Farbschattierungen, zu. Für nominale Variablen macht das keinen Sinn; es gibt da keine Zwischentöne. Tot ist tot, lebendig ist lebendig. Wir sollten daher der Funktion sagen, dass es sich um nominale Variablen handelt (s. Abbildung 12.1).

```
qplot(x = factor(Pclass), y = n, color = factor(Survived), data = c3)
```

Viel besser. Jetzt fügen Sie noch ein bisschen Schnickschnack hinzu:

```
qplot(x = factor(Pclass), y = n, color = factor(Survived), data = c3) +
  labs(x = "Klasse",
       title = "Überleben auf der Titanic",
       colour = "Überlebt?")
```

12.1.6 Signifikanztest

Manche Leute mögen Signifikanztests. Ich persönlich stehe ihnen kritisch gegenüber, da ein p-Wert eine Funktion der Stichprobengröße ist und außerdem zumeist missverstanden wird (und er gibt *nicht* die Wahrscheinlichkeit der getesteten Hypothese an, was die Frage aufwirft, warum er mich dann interessieren sollte). Aber seid drüber, berechnen wir mal einen p-Wert. Es gibt mehrere statistische Tests, die sich hier potenziell anbieten und unterschiedliche Ergebnisse liefern können (Briggs 2008a) (was die Frage nach der Objektivität von statistischen Tests in ein ungünstiges Licht rückt). Nehmen wir den χ^2 -Test.

```
chisq.test(titanic_train$Survived, titanic_train$Pclass)
#>
#> Pearson's Chi-squared test
#>
#> data: titanic_train$Survived and titanic_train$Pclass
#> X-squared = 90, df = 2, p-value <2e-16
```

Der p-Wert ist kleiner als 5%, daher entscheiden wir uns, entsprechend der üblichen Gepflogenheit, gegen die H₀ und für die H₁: “Es gibt einen Zusammenhang von Überlebensrate und Passagierklasse”.

12.1.7 Effektstärke

Abgesehen von der Signifikanz, und interessanter, ist die Frage, wie sehr die Variablen zusammenhängen. Für Häufigkeitsanalysen mit 2*2-Feldern bietet sich das “Odds Ratio” (OR), das Chancenverhältnis an. Das Chancen-Verhältnis beantwortet die Frage: “Um welchen Faktor ist die Überlebenschance in der einen Klasse größer als in der anderen Klasse?”. Eine interessante Frage, als schauen wir es uns an.

Das OR ist nur definiert für 2*2-Häufigkeitstabellen, daher müssen wir die Anzahl der Passagierklassen von 3 auf 2 verringern. Nehmen wir nur 1. und 3. Klasse, um den vermuteten Effekt deutlich herauszuschälen:

```
t2 <- filter(titanic_train, Pclass != 2) # "!=" heißt "nicht"
```

Alternativ (synonym) könnten wir auch schreiben:

```
t2 <- filter(titanic_train, Pclass == 1 | Pclass == 3) # "/" heißt "oder"
```

Und dann zählen wir wieder die Häufigkeiten aus pro Gruppe:

```
c4 <- dplyr::count(t2, Pclass)
c4
#> # A tibble: 2 x 2
#>   Pclass     n
#>   <int> <int>
#> 1     1    186
#> 2     3    355
```

Schauen wir nochmal den p-Wert an, da wir jetzt ja mit einer veränderten Datentabelle operieren:

```
chisq.test(t2$Survived, t2$Pclass)
#>
#> Pearson's Chi-squared test with Yates' continuity correction
#>
#> data: t2$Survived and t2$Pclass
#> X-squared = 90, df = 1, p-value <2e-16
```

Ein χ^2 -Wert von ~96 bei einem n von 707.

Dann berechnen wir die Effektstärke (OR) mit dem Paket `compute.es` (muss ebenfalls installiert sein)

```
compute.es::chies(chi.sq = 96, n = 707)
```

Das OR beträgt also etwa 4.21. Die Chance zu überleben ist also in der 1. Klasse mehr als 4 mal so hoch wie in der 3. Klasse. Es scheint: Money buys you life...

12.1.8 Logististische Regression

Berechnen wir noch das Odds Ratio mit Hilfe der logistischen Regression. Zum Einstieg: Ignorieren Sie die folgende Syntax und schauen Sie sich das Diagramm an. Hier sehen wir die (geschätzten) Überlebens-Wahrscheinlichkeiten für Passagiere der 1. Klasse vs. Passagiere der 2. vs. der 3. Klasse.

```
glm1 <- glm(data = titanic_train,
            formula = Survived ~ Pclass,
```

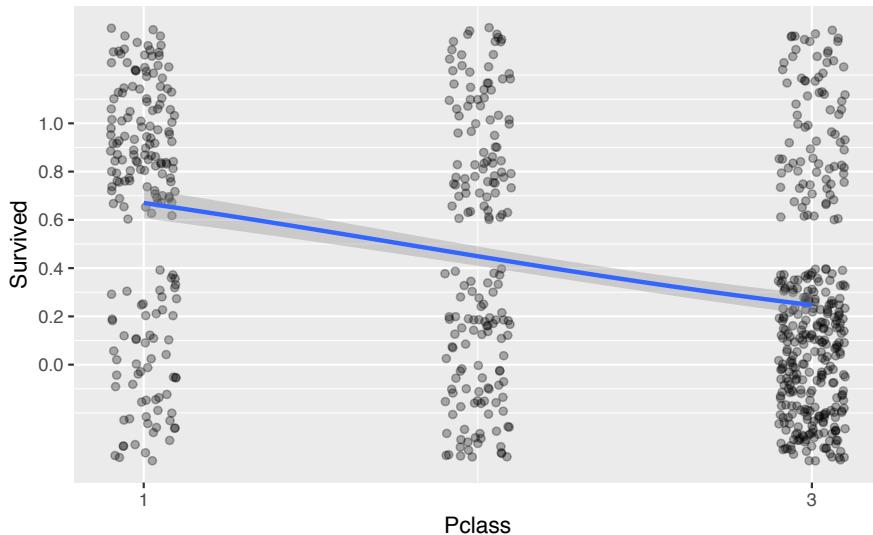


Abbildung 12.2: Logistische Regression zur Überlebensrate nach Passagierklasse

```

family = "binomial")

exp(coef(glm1))
#> (Intercept)      Pclass
#>      5.056       0.402

titanic_train$pred_prob <- predict(glm1, type = "response")

```

Wir sehen, dass die Überlebens-Wahrscheinlichkeit in der 1. Klasse höher ist als in der 3. Klasse. Optisch grob geschätzt, ~60% in der 1. Klasse und ~25% in der 3. Klasse.

Schauen wir uns die logistische Regression an: Zuerst haben wir den Datensatz auf die Zeilen beschränkt, in denen Personen aus der 1. und 3. Klasse vermerkt sind (zwecks Vergleichbarkeit zu oben). Dann haben wir mit `glm` und `family = "binomial"` eine *logistische* Regression angefordert. Man beachte, dass der Befehl sehr ähnlich zur normalen Regression (`lm(...)`) ist.

Da die Koeffizienten in der Logit-Form zurückgegeben werden, haben wir sie mit der Exponential-Funktion in die “normale” Odds-Form gebracht (deologarithmiert, *boa*) mithilfe von `exp(coef)`. Wir sehen, dass sich die Überlebens-*Chance* (Odds; nicht Wahrscheinlichkeit) um den Faktor .4 verringert pro zusätzlicher Stufe der Passagierklasse. Würde jemand in der “nullten” Klasse fahren, wäre seine Überlebenschance ca. 5:1 (5/6, gut 80%). Die Überlebenschance in der 1. Klasse sind demnach etwa: 5 * 0.4, also 2:1, etwa 67%.

Komfortabler können wir uns die Überlebens-*Wahrscheinlichkeiten* mit der Funktion `predict` ausgeben lassen.

```
predict(glm1, newdata = data.frame(Pclass = 1), type = "response")
#>     1
#> 0.67
predict(glm1, newdata = data.frame(Pclass = 2), type = "response")
#>     1
#> 0.449
predict(glm1, newdata = data.frame(Pclass = 3), type = "response")
#>     1
#> 0.247
```

Alternativ kann man die tatsächlichen (beobachteten) Häufigkeiten auch noch “per Hand” bestimmen:

```
titanic_train %>%
  filter(Pclass %in% c(1,3)) %>%
  dplyr::select(Survived, Pclass) %>%
  group_by(Pclass, Survived) %>%
  summarise(n = n()) %>%
  mutate(Anteil = n / sum(n))
#> # A tibble: 4 x 4
#> # Groups:   Pclass [2]
#>   Pclass Survived     n Anteil
#>   <int>    <int> <int>  <dbl>
#> 1     1        1     64  0.344
#> 2     1        0    122  0.656
#> 3     3        1     270  0.761
#> 4     3        0     85  0.239
```

Übersetzen wir dies Syntax auf Deutsch:



Nehme den Datensatz “titanic_train” UND DANN
 Filtere nur die 1. und die 3. Klasse heraus UND DANN
 wähle nur die Spalten “Survived” und “Pclass” UND DANN
 gruppiere nach “Pclass” und “Survived” UND DANN
 zähle die Häufigkeiten für jede dieser Gruppen aus UND DANN
 berechne den Anteil an Überlebenden bzw. Nicht-Überlebenden
 für jede der beiden Passagierklassen. FERTIG.

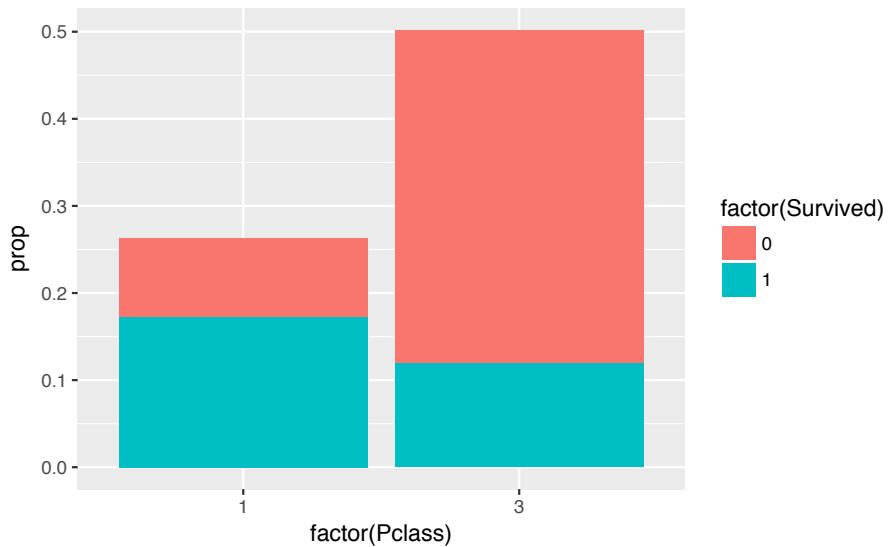


Abbildung 12.3: Absolute Überlebenshäufigkeiten

12.1.9 Effektstärken visualieren

Zum Abschluss schauen wir uns die Stärke des Zusammenhangs noch einmal graphisch an. Wir berechnen dafür die relativen Häufigkeiten pro Gruppe (im Datensatz ohne 2. Klasse, der Einfachheit halber).

```
c5 <- dplyr::count(t2, Pclass, Survived)
c5$prop <- c5$n / 707
c5

#> # A tibble: 4 x 4
#>   Pclass Survived     n    prop
#>   <int>     <int> <int>  <dbl>
#> 1     1         0     64  0.0905
#> 2     1         1    122  0.1726
#> 3     3         0    270  0.3819
#> 4     3         1     85  0.1202
```

Genauer gesagt haben die Häufigkeiten pro Gruppe in Bezug auf die Gesamtzahl aller Passagiere berechnet; die vier Anteile addieren sich also zu 1 auf. Das visualisieren wir wieder, s. Abbildung 12.3.

```
qplot(x = factor(Pclass),
      y = prop,
      fill = factor(Survived),
      data = c5,
      geom = "col")
```

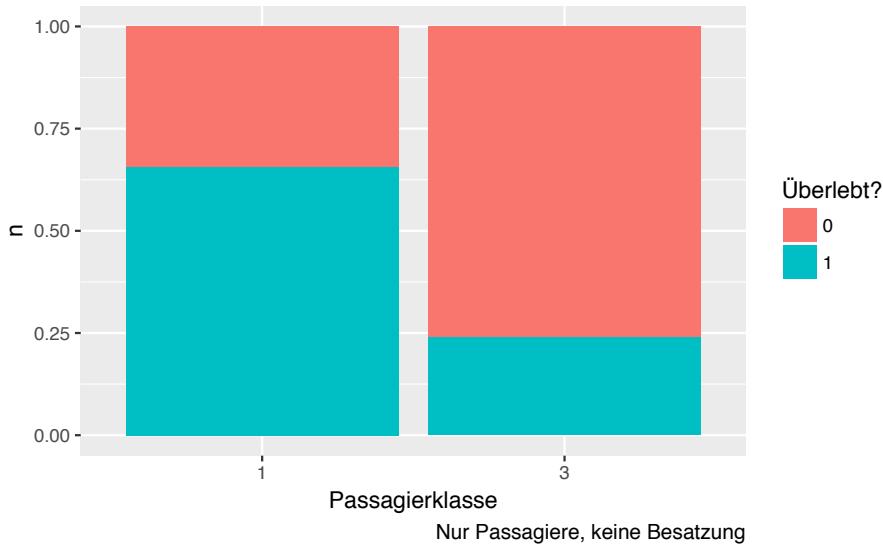


Abbildung 12.4: Relative Überlebenshäufigkeiten

Das `geom = "col"` heißt, dass als “geometrisches Objekt” dieses Mal keine Punkte, sondern Säulen (columns) verwendet werden sollen.

Ganz nett, aber die Häufigkeitsunterschiede von `Survived` zwischen den beiden Werten von `Pclass` stechen noch nicht so ins Auge. Wir sollten es anders darstellen. Hier kommt der Punkt, wo wir von `qplot` auf seinen großen Bruder, `ggplot` wechseln sollten. `qplot` ist in Wirklichkeit nur eine vereinfachte Form von `ggplot`; die Einfachheit wird mit geringeren Möglichkeiten bezahlt. Satteln wir zum Schluss dieser Fallstudie also um, s. Abbildung 12.4.

```
ggplot(data = c5) +
  aes(x = factor(Pclass), y = n, fill = factor(Survived)) +
  geom_col(position = "fill") +
  labs(x = "Passagierklasse",
       fill = "Überlebt?",
       caption = "Nur Passagiere, keine Besatzung")
```

Jeden sehen wir die Häufigkeiten des Überlebens bedingt auf die Passagierklasse besser. Wir sehen auf den ersten Blick, dass sich die Überlebensraten deutlich unterscheiden: Im linken Balken überleben die meisten; im rechten Balken ertrinken die meisten. Mit `labs` haben wir noch die X-Achse (`x`), die Bezeichnung der Füllfarbe (`fill`) sowie die Legende des Diagramms beschrieben. Diese letzte Analyse zeigt schön die Kraft von (Daten-)Visualisierungen auf. Der zu untersuchende Effekt tritt hier am stärken zu Tage; außerdem ist die Analyse relativ einfach.

Eine alternative Darstellung zeigt Abbildung 12.5. Hier werden die vier “Fliesen” gleich groß dargestellt; die Fallzahl wird durch die Füllfarbe besorgt.

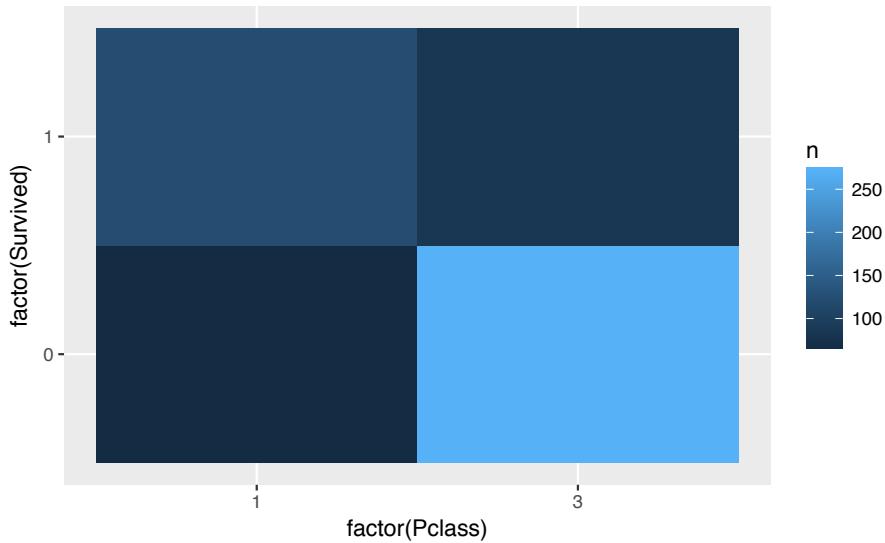


Abbildung 12.5: Überlebenshäufigkeiten anhand eines Fliesenbildes dargestellt

```
c5 %>%
  ggplot +
  aes(x = factor(Pclass), y = factor(Survived), fill = n) +
  geom_tile()
```

12.1.10 Fazit

In der Datenanalyse (mit R) kommt man mit wenigen Befehlen schon sehr weit; `dplyr` und `ggplot2` zählen (zu Recht) zu den am häufigsten verwendeten Paketen. Beide sind flexibel, konsistent und spielen gerne miteinander. Die besten Einblicke haben wir aus deskriptiver bzw. explorativer Analyse (Diagramme) gewonnen. Signifikanztests oder komplizierte Modelle waren nicht zentral. In vielen Studien/Projekten der Datenanalyse gilt ähnliches: Daten umformen und verstehen bzw. „veranschaulichen“ sind zentrale Punkte, die häufig viel Zeit und Wissen fordern. Bei der Analyse von nominalskalierten sind Häufigkeitsauswertungen ideal.

12.2 Außereheliche Affären

Für diese Fallstudie benötigen wir folgende Pakete:

```
library(AER) # Datensatz 'Affairs'
library(psych) # Befehl 'describe'
```

```
library(tidyverse) # Datenjudo
library(broom) # Befehl 'tidy'
```

Wovon ist die Häufigkeit von Affären (Seitensprünge) in Ehen abhängig? Diese Frage soll anhand des Datensatzes `Affairs` untersucht werden. Laden wir als erstes den Datensatz in R.

```
data(Affairs, package = "AER")
```

Verschaffen Sie sich zum Einstieg einen Überblick über die Daten. ... OK, scheint zu passen. Was jetzt?

12.2.1 Zentrale Statistiken

Geben Sie zentrale deskriptive Statistiken an für Affärenhäufigkeit und Ehezufriedenheit!

```
# nicht robust:
mean(Affairs$affairs, na.rm = T)
#> [1] 1.46
sd(Affairs$affairs, na.rm = T)
#> [1] 3.3
# robust:
median(Affairs$Affairs, na.rm = T)
#> NULL
IQR(Affairs$Affairs, na.rm = T)
#> [1] NA
```

Es scheint, die meisten Leute haben keine Affären:

```
count(Affairs, affairs)
#> # A tibble: 6 x 2
#>   affairs     n
#>   <dbl> <int>
#> 1      0    451
#> 2      1     34
#> 3      2     17
#> 4      3     19
#> 5      7     42
#> 6     12     38
```

Man kann sich viele Statistiken mit dem Befehl `describe` aus `psych` ausgeben lassen, das ist etwas praktischer:

```
describe(Affairs$affairs)
#>   vars   n mean  sd median trimmed mad min max range skew kurtosis   se
#> X1     1 601 1.46 3.3      0    0.55   0   0  12    12 2.34      4.19 0.13
describe(Affairs$rating)
#>   vars   n mean  sd median trimmed mad min max range skew kurtosis   se
#> X1     1 601 3.93 1.1      4    4.07 1.48   1   5    4 -0.83     -0.22 0.04
```

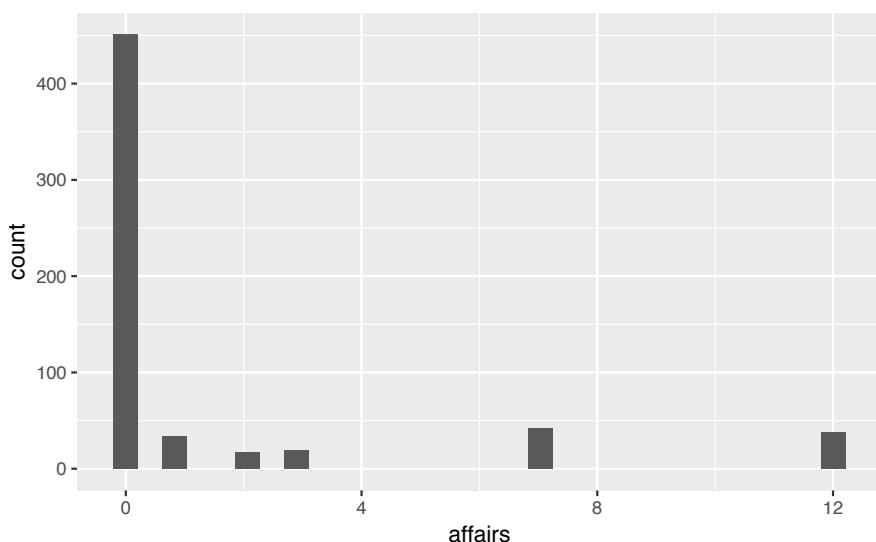
Dazu muss das Paket `psych` natürlich vorher installiert sein. Beachten Sie, dass man ein Paket nur *einmal* installieren muss, aber jedes Mal, wenn Sie R starten, auch starten muss (mit `library`; vgl. Kapitel 1).

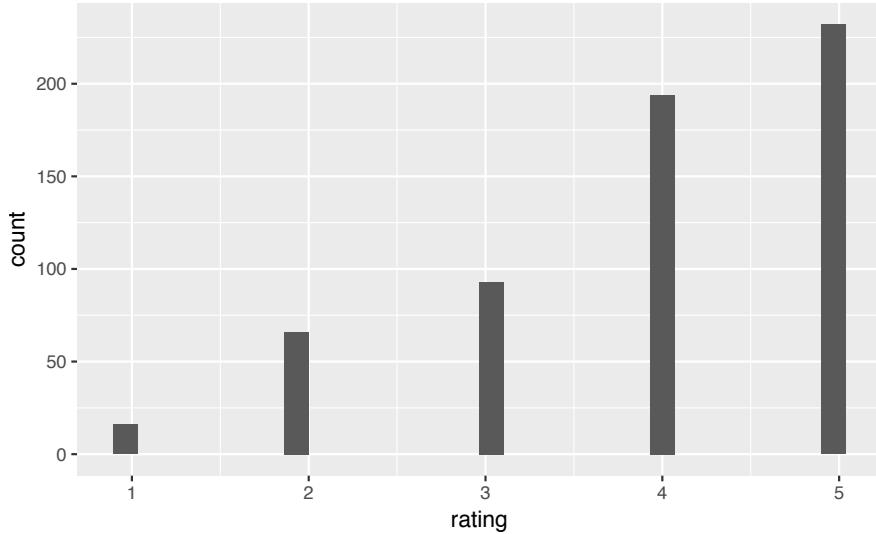
12.2.2 Visualisieren

Visualisieren Sie zentrale Variablen!

Sicherlich sind Diagramme auch hilfreich. Dies geht wiederum mit dem R-Commander oder z.B. mit folgenden Befehlen:

```
qplot(x = affairs, data = Affairs)
qplot(x = rating, data = Affairs)
```





Die meisten Menschen (dieser Stichprobe) scheinen mit Ihrer Beziehung sehr zufrieden zu sein.

12.2.3 Wer ist zufriedener mit der Partnerschaft: Personen mit Kindern oder ohne?

Nehmen wir dazu mal ein paar dplyr-Befehle:

```
library(dplyr)
Affairs %>%
  group_by(children) %>%
  summarise(rating_children =
    mean(rating, na.rm = T))
#> # A tibble: 2 x 2
#>   children rating_children
#>   <fctr>        <dbl>
#> 1 no            4.27
#> 2 yes           3.80
```

Ah! Kinder sind also ein Risikofaktor für eine Partnerschaft! Gut, dass wir das geklärt haben.

12.2.4 Vertiefung: Wie viele fehlende Werte gibt es?

Was machen wir am besten damit?

Diesen Befehl könnten wir für jede Spalte ausführen:

```
sum(is.na(Affairs$affairs))
#> [1] 0
```

Oder lieber alle auf einmal:

```
Affairs %>%
  summarise_all(funs(sum(is.na(.))))
#> affairs gender age yearsmarried children religiousness education
#> 1      0      0    0        0      0      0      0
#> occupation rating
#> 1      0      0
```

Übrigens gibt es ein gutes Cheat Sheet⁴ für dplyr.

Ah, gut, keine fehlenden Werte. Das macht uns das Leben leichter.

12.2.5 Wer ist glücklicher: Männer oder Frauen?

```
Affairs %>%
  group_by(gender) %>%
  summarise(rating_gender = mean(rating))
#> # A tibble: 2 x 2
#>   gender rating_gender
#>   <fctr>     <dbl>
#> 1 female      3.94
#> 2 male        3.92
```

Praktisch kein Unterschied. Heißt das auch, es gibt keinen Unterschied in der Häufigkeit der Affären?

```
Affairs %>%
  group_by(gender) %>%
  summarise(affairs_gender = mean(affairs))
#> # A tibble: 2 x 2
#>   gender affairs_gender
#>   <fctr>     <dbl>
#> 1 female      1.42
#> 2 male        1.50
```

Scheint auch kein Unterschied zu sein...

⁴<https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf>

Und zum Abschluss noch mal etwas genauer: Teilen wir mal nach Geschlecht und nach Kinderstatus auf, also in 4 Gruppen. Theoretisch dürfte es hier auch keine Unterschiede/Zusammenhänge geben. Zum mindesten fällt mir kein sinnvoller Grund ein; zumal die vorherige eindimensionale Analyse keine Unterschiede zu Tage gefördert hat.

```
Affairs %>%
  group_by(gender, children) %>%
  summarise(affairs_mean = mean(affairs),
            rating_mean = mean(rating))
#> # A tibble: 4 x 4
#> # Groups:   gender [?]
#>   gender children affairs_mean rating_mean
#>   <fctr>    <fctr>      <dbl>        <dbl>
#> 1 female     no         0.838       4.40
#> 2 female     yes        1.685       3.73
#> 3 male       no         1.014       4.10
#> 4 male       yes        1.659       3.86

Affairs %>%
  group_by(children, gender) %>%
  summarise(affairs_mean = mean(affairs),
            rating_mean = mean(rating))
#> # A tibble: 4 x 4
#> # Groups:   children [?]
#>   children gender affairs_mean rating_mean
#>   <fctr>    <fctr>      <dbl>        <dbl>
#> 1 no        female    0.838       4.40
#> 2 no        male     1.014       4.10
#> 3 yes       female   1.685       3.73
#> 4 yes       male    1.659       3.86
```

12.2.6 Effektstärken

Berichten Sie eine relevante Effektstärke!

Hm, auch keine gewaltigen Unterschiede. Höchstens für die Zufriedenheit mit der Partnerschaft bei kinderlosen Personen scheinen sich Männer und Frauen etwas zu unterscheiden. Hier stellt sich die Frage nach der Größe des Effekts, z.B. anhand Cohen's d. Dafür müssen wir noch die SD pro Gruppe wissen:

```
Affairs %>%
  group_by(children, gender) %>%
```

```
summarise(rating_mean = mean(rating),
           rating_sd = sd(rating))
#> # A tibble: 4 x 4
#> # Groups:   children [?]
#>   children gender rating_mean rating_sd
#>   <fctr> <fctr>     <dbl>      <dbl>
#> 1 no       female     4.40      0.914
#> 2 no       male      4.10      1.064
#> 3 yes      female    3.73      1.183
#> 4 yes      male      3.86      1.046
```

```
d <- (4.4 - 4.1)/(1)
```

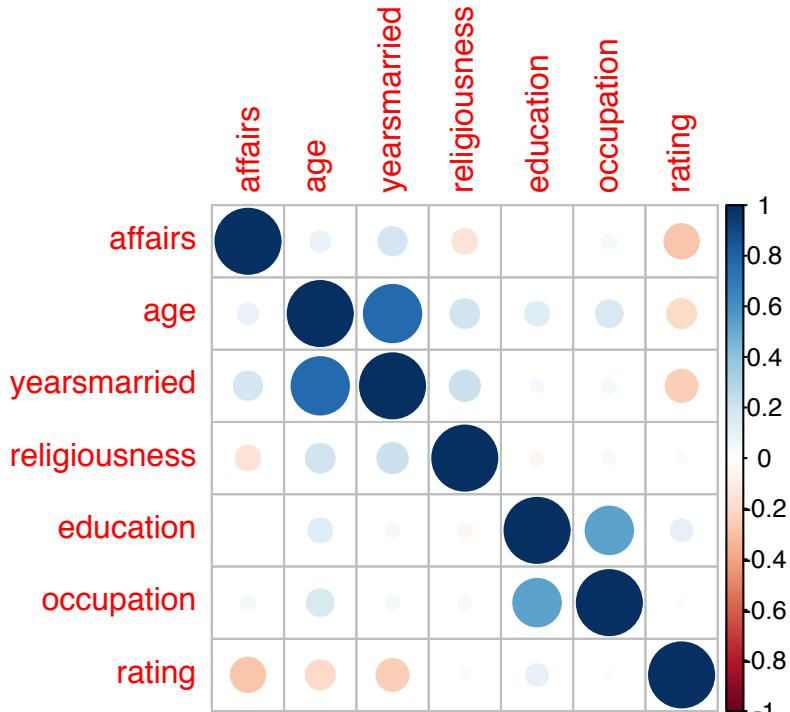
Die Effektstärke beträgt etwa 0.3.

12.2.7 Korrelationen

Berechnen und visualisieren Sie zentrale Korrelationen!

```
Affairs %>%
  select_if(is.numeric) %>%
  cor -> cor_tab

corrplot(cor_tab)
```



12.2.8 Ehejahre und Affären

Wie groß ist der Einfluss (das Einflussgewicht) der Ehejahre bzw. Ehezufriedenheit auf die Anzahl der Affären?

Dazu sagen wir R: "Hey R, rechne mal ein lineares Modell", also eine normale (lineare) Regression. Dazu können wir entweder das entsprechende Menü im R-Commander auswählen, oder folgende R-Befehle ausführen:

```
lm1 <- lm(affairs ~ yearsmarried, data = Affairs)
tidy(lm1) # Ergebnisse der Regression zeigen
#>   term estimate std.error statistic p.value
#> 1 (Intercept)  0.551    0.2351     2.34 0.019378
#> 2 yearsmarried  0.111    0.0238     4.65 0.000004
glance(lm1)
#>   r.squared adj.r.squared sigma statistic p.value df logLik AIC  BIC
#> 1  0.0349      0.0333  3.24      21.7  4e-06  2 -1559 3124 3137
#>   deviance df.residual
#> 1     6301       599
lm2 <- lm(affairs ~ rating, data = Affairs)
tidy(lm2)
#>   term estimate std.error statistic p.value
#> 1 (Intercept)  4.742    0.479     9.90 1.68e-21
#> 2     rating -0.836    0.117    -7.12 3.00e-12
```

```
glance(lm2)
#>   r.squared adj.r.squared sigma statistic p.value df logLik AIC BIC
#> 1    0.0781        0.0766  3.17      50.8   3e-12  2 -1545 3096 3110
#>   deviance df.residual
#> 1      6019         599
```

Also: `yearsmarried` und `rating` sind beide statistisch signifikante Prädiktoren für die Häufigkeit von Affären. Das adjustierte R^2 ist allerdings in beiden Fällen nicht so groß.

12.2.9 Ehezufriedenheit als Prädiktor

Um wie viel erhöht sich die erklärte Varianz (R-Quadrat) von Affärenhäufigkeit wenn man den Prädiktor Ehezufriedenheit zum Prädiktor Ehejahre hinzufügt? (Wie) verändern sich die Einflussgewichte (b)?⁵

```
lm3 <- lm(affairs ~ rating + yearsmarried, data = Affairs)
lm4 <- lm(affairs ~ yearsmarried + rating, data = Affairs)
summary(lm3)
summary(lm4)
```

Ok. Macht eigentlich die Reihenfolge der Prädiktoren in der Regression einen Unterschied? Der Vergleich von Modell 3 vs. Modell 4 beantwortet diese Frage.

Wir sehen, dass beim 1. Regressionsmodell das R^2 0.03 war; beim 2. Modell 0.08 und beim 3. Modell liegt R^2 bei 0.09. Die Differenz zwischen Modell 1 und 3 liegt bei (gerundet) 0.06; wenig.

12.2.10 Weitere Prädiktoren der Affärenhäufigkeit

Welche Prädiktoren würden Sie noch in die Regressionsanalyse aufnehmen?

Hm, diese Frage klingt nicht so, als ob der Dozent die Antwort selber wüsste... Naja, welche Variablen gibt es denn alles:

```
#> [1] "affairs"       "gender"        "age"          "yearsmarried"
#> [5] "children"     "religiousness" "education"    "occupation"
#> [9] "rating"
```

Z.B. wäre doch interessant, ob Ehen mit Kinder mehr oder weniger Seitensprünge aufweisen. Und ob die “Kinderfrage” die anderen Zusammenhänge/Einflussgewichte in der Regression verändert. Probieren wir es auch. Wir können wiederum im R-Commander ein Regressionsmodell anfordern oder es mit der Syntax probieren:

⁵Output im Folgenden nicht abgedruckt.

```
lm5 <- lm(affairs ~ rating + yearsmarried + children, data = Affairs)
summary(lm5)
r2_lm5 <- summary(lm5)$r.squared
```

Das Regressionsgewicht von `childrenyes` ist negativ. Das bedeutet, dass Ehen mit Kindern weniger Affären verbuchen (aber geringe Zufriedenheit, wie wir oben gesehen haben! Hrks!). Allerdings ist der p-Wert nicht signifikant, was wir als Zeichen der Unbedeutsamkeit dieses Prädiktors verstehen können. R^2 lungert immer noch bei mickrigen 0.094 herum. Wir haben bisher kaum verstanden, wie es zu Affären kommt. Oder unsere Daten bergen diese Informationen einfach nicht.

Wir könnten auch einfach mal Prädiktoren, die wir haben, ins Feld schicken. Mal sehen, was dann passiert:

```
lm6 <- lm(affairs ~ ., data = Affairs)
summary(lm6)
```

Der “.” im Befehl `affairs ~ .` oben soll sagen: nimm “alle Variablen, die noch in der Datenmatrix übrig sind”.

Insgesamt bleibt die erklärte Varianz in sehr bescheidenem Rahmen: 0.13. Das zeigt uns, dass es immer noch nur schlecht verstanden ist – im Rahmen dieser Analyse – welche Faktoren die Affärenhäufigkeit erklärt.

12.2.11 Unterschied zwischen den Geschlechtern

Unterscheiden sich die Geschlechter statistisch signifikant? Wie groß ist der Unterschied? Sollte hier lieber das d-Maß oder Rohwerte als Effektmaß angegeben werden?

Hier bietet sich ein t-Test für unabhängige Gruppen an. Die Frage lässt auf eine ungerichtete Hypothese schließen (α sei .05). Mit dem entsprechenden Menüpunkt im R-Commander oder mit folgender Syntax lässt sich diese Analyse angehen:

```
t.test(affairs ~ gender, data = Affairs) -> t1

t1 %>% tidy
#>   estimate estimate1 estimate2 statistic p.value parameter conf.low
#> 1  -0.0775      1.42       1.5     -0.287   0.774      594    -0.607
#>   conf.high               method alternative
#> 1      0.452 Welch Two Sample t-test   two.sided
```

Der p-Wert ist größer als α . Daher wird die H_0 beibehalten. Auf Basis der Stichprobendaten entscheiden wir uns für die H_0 . Entsprechend umschließt das 95%-KI die Null.

Da die Differenz nicht signifikant ist, kann argumentiert werden, dass wir d auf 0 schätzen müssen. Man kann sich den d -Wert auch z.B. von {MBESS} schätzen lassen.

Dafür brauchen wir die Anzahl an Männer und Frauen: 315, 286.

```
library(MBESS)
ci.smd(ncp = t1$statistic,
       n.1 = 315,
       n.2 = 286)
#> $Lower.Conf.Limit.smd
#> [1] -0.184
#>
#> $smd
#>      t
#> -0.0235
#>
#> $Upper.Conf.Limit.smd
#> [1] 0.137
```

Das Konfidenzintervall ist zwar relativ klein (die Schätzung also aufgrund der recht großen Stichprobe relativ präzise), aber der Schätzwert für d `smd` liegt sehr nahe bei Null. Das stärkt unsere Entscheidung, von einer Gleichheit der Populationen (Männer vs. Frauen) auszugehen.

12.2.12 Kinderlose Ehe vs. Ehen mit Kindern

Rechnen Sie die Regressionsanalyse getrennt für kinderlose Ehe und Ehen mit Kindern!

Hier geht es im ersten Schritt darum, die entsprechenden Teil-Mengen der Datenmatrix zu erstellen. Das kann man natürlich mit Excel o.ä. tun. Alternativ könnte man es in R z.B. so machen:

```
Affair4 <- filter(Affairs, children == "yes")
head(Affair4)
```

12.2.13 Halodries

Rechnen Sie die Regression nur für “Halodries”; d.h. für Menschen mit Seitensprüngen. Dafür müssen Sie alle Menschen ohne Affären aus den Datensatz entfernen.

Also, rechnen wir nochmal die Standardregression (`lm1`). Probieren wir den Befehl `filter` dazu nochmal aus:

```
Affair5 <- filter(Affairs, affairs != 0)
lm9 <- lm(affairs ~ rating, data = Affair5)
summary(lm9)
```

12.2.14 logistische Regression

Berechnen Sie für eine logistische Regression mit “Affäre ja vs. nein” als Kriterium, wie stark der Einfluss von Geschlecht, Kinderstatus, Ehezufriedenheit und Ehedauer ist!

```
Affairs %>%
  mutate(affairs_dichotom = affairs == 0) %>%
  glm(affairs_dichotom ~ gender + children + rating + yearsmarried,
      data = .,
      family = "binomial") -> lm10

tidy(lm10)
```

Wenn `if_else` unbekannt ist, lohnt sich ein Blick in die Hilfe mit `?if_else` (`dplyr` muss vorher geladen sein).

Aha, signifikant ist die Ehezufriedenheit: Je größer `rating` desto geringer die Wahrscheinlichkeit für `affairs_dichotom`. Macht Sinn!

Übrigens, die Funktionen `lm`, `glm` und `summary` spucken leider keine brave Tabelle in Normalform aus, was aber schön wäre. Aber man leicht eine Tabelle (data.frame) bekommen mit dem Befehl `tidy` aus `broom`:

```
tidy(lm10)
#>   term estimate std.error statistic p.value
#> 1 (Intercept) -0.0537    0.4299    -0.125 9.01e-01
#> 2 gendermale   -0.2416    0.1966    -1.229 2.19e-01
#> 3 childrenyes  -0.3935    0.2831    -1.390 1.64e-01
#> 4 rating       0.4654    0.0874     5.327 9.97e-08
#> 5 yearsmarried -0.0221    0.0212    -1.040 2.99e-01
```

12.2.15 Zum Abschluss

Visualisieren wir mal was! Ok, wie wäre es mit einem Jitter-Diagramm (vgl. Abbildungen 12.6 und 12.7).

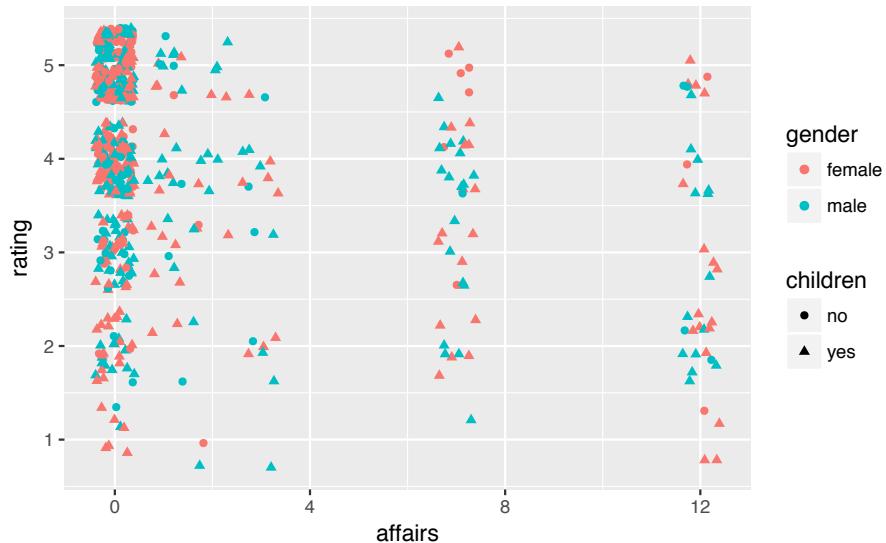


Abbildung 12.6: Affären, mit Jitter

Affairs %>%

```
select(affairs, gender, children, rating) %>%
  ggplot(aes(x = affairs, y = rating)) +
  geom_jitter(aes(color = gender, shape = children))
```

Affairs %>%

```
mutate(rating_dichotom = ntile(rating, 2)) %>%
  ggplot(aes(x = yearsmarried, y = affairs)) +
  geom_jitter(aes(color = gender)) +
  geom_smooth()
```

Puh. Geschafft!

12.3 Befehlsübersicht

Tabelle 12.1 fasst die R-Funktionen dieses Kapitels zusammen.

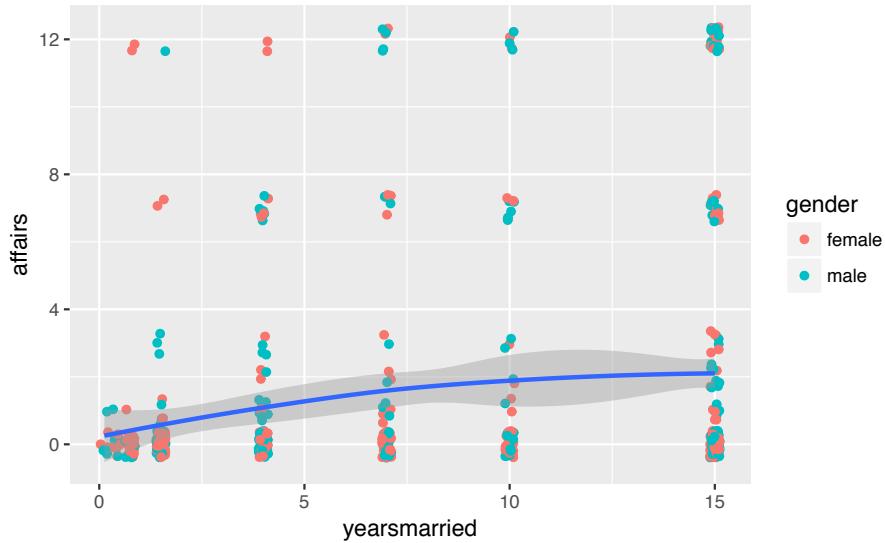


Abbildung 12.7: Affären, mit Smooth

Tabelle 12.1: Befehle des Kapitels 'Fallstudien titanic und affairs'

Paket..Funktion	Beschreibung
data	Lädt Daten aus einem Datensatz
chisq.test	Rechnet einen Chi-Quadrat-Test
compute.es::chies	Liefert Effektstärkemaße für einen Chi-Quadrat-Test
predict	Macht eine Vorhersage
psych::describe	Liefert eine Reihe zentraler Statistiken
is.na	Zeigt an, ob ein Vektor fehlende Werte beinhaltet
dplyr::summarise_each	Führt summarise für jede Spalte aus
t.test	Rechnet einen t-Test
MBESS::ci.smd	Berechnet Cohens d
dplyr::ntile	Teilt einen Vektor in n Teile mit jeweils gleich viel Werten
broom::tidy	Wandelt ein Objekt vom Typ 'lm' in einen Dataframe um.

Teil IV

Ungeleitetes Modellieren

Kapitel 13

Vertiefung: Clusteranalyse



Benötigte Pakete:

```
library(tidyverse)
library(cluster)
```

```
library(broom)
```



Lernziele:

- Das Ziel einer Clusteranalyse erläutern können.
- Das Konzept der euklidischen Abstände verstehen.
- Eine k-Means-Clusteranalyse berechnen und interpretieren können.

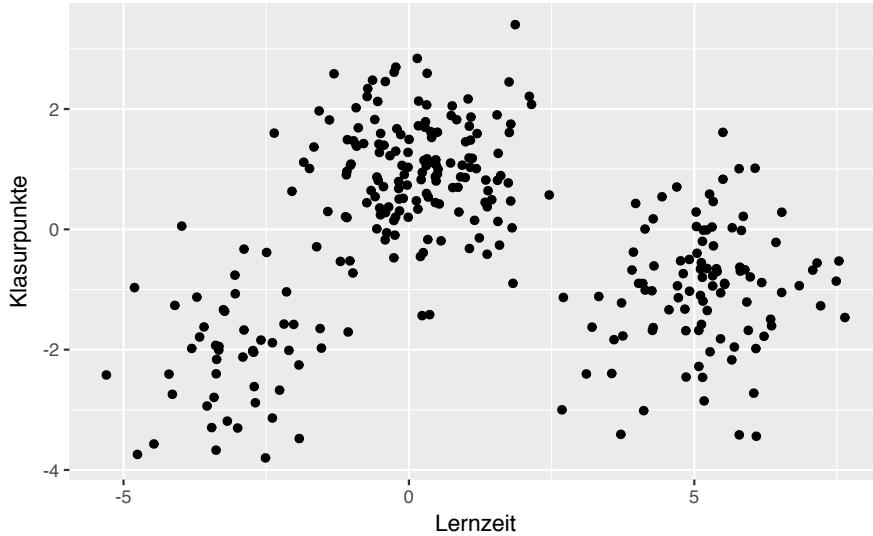


Abbildung 13.1: Ein Streudiagramm - sehen Sie Gruppen (Cluster) ?

13.1 Grundlagen der Clusteranalyse

Das Ziel einer Clusteranalyse ist es, Gruppen von Beobachtungen (d. h. *Cluster*) zu finden, die innerhalb der Cluster möglichst homogen, zwischen den Clustern möglichst heterogen sind. Um die Ähnlichkeit von Beobachtungen zu bestimmen, können verschiedene Distanzmaße herangezogen werden. Für metrische Merkmale wird z. B. häufig die euklidische Metrik verwendet, d. h., Ähnlichkeit und Distanz werden auf Basis des euklidischen Abstands bestimmt. Aber auch andere Abstände wie "Manhattan" oder "Gower" sind möglich. Letztere haben den Vorteil, dass sie nicht nur für metrische Daten sondern auch für gemischte Variablentypen verwendet werden können. Wir werden uns hier auf den euklidischen Abstand konzentrieren.

13.1.1 Intuitive Darstellung der Clusteranalyse

Betrachten Sie das folgende Streudiagramm (die Daten sind frei erfunden; "simuliert", sagt der Statistiker). Es stellt den Zusammenhang von Lernzeit (wie viel ein Student für eine Statistikklausur lernt) und dem Klausurerfolg (wie viele Punkte ein Student in der Klausur erzielt) dar. Sehen Sie Muster? Lassen sich Gruppen von Studierenden mit bloßem Auge abgrenzen (Abb. 13.1)?

Färben wir das Diagramm mal ein (Abb. 13.2).

Nach dieser "Färbung", d.h. nach dieser Aufteilung in drei Gruppen, scheint es folgende "Cluster", "Gruppen" oder "Typen" von Studierenden zu geben:

- "Blaue Gruppe": Fälle dieser Gruppe lernen wenig und haben wenig Erfolg in der Klausur. Tja.
- "Rote Gruppe": Fälle dieser Gruppe lernen viel; der Erfolg ist recht durchwachsen.

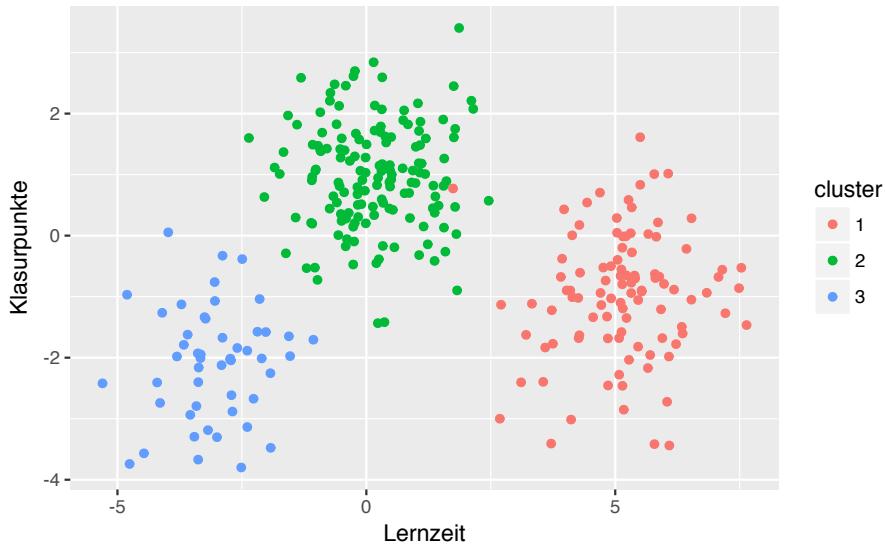


Abbildung 13.2: Ein Streudiagramm - mit drei Clustern

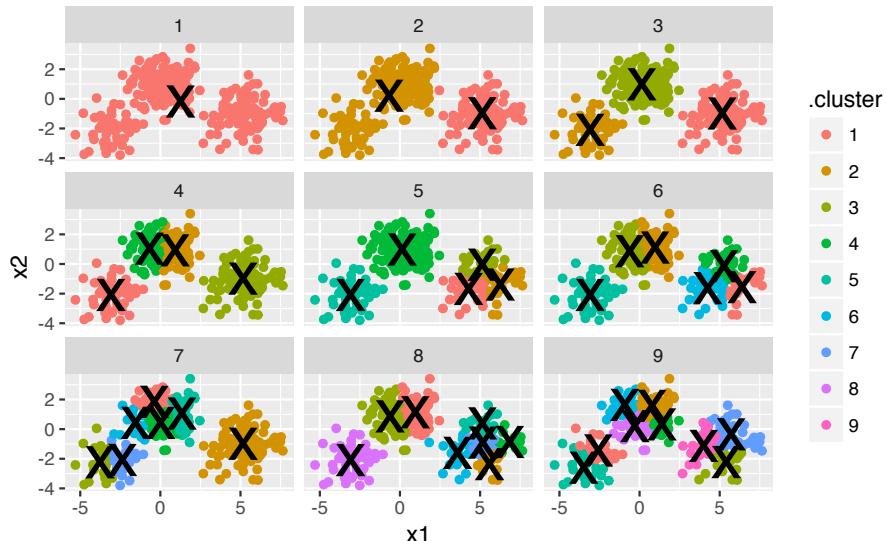


Abbildung 13.3: Unterschiedliche Anzahlen von Clustern im Vergleich

- “Grüne Gruppe”: Fälle dieser Gruppe lernen mittel viel und erreichen einen vergleichsweise großen Erfolg in der Klausur.

Drei Gruppen scheinen ganz gut zu passen. Wir hätten theoretisch auch mehr oder weniger Gruppen unterteilen können. Die Clusteranalyse gibt keine definitive Anzahl an Gruppen vor; vielmehr gilt es, aus theoretischen und statistischen Überlegungen heraus die richtige Anzahl auszuwählen (dazu gleich noch mehr).

Unterteilen wir zur Illustration den Datensatz einmal in bis zu 9 Cluster (Abbildung 13.3).

Das “X” soll den “Mittelpunkt” des Clusters zeigen. Der Mittelpunkt ist so gewählt, dass die Distanz von jedem Punkt zum Mittelpunkt möglichst kurz ist. Dieser Abstand wird auch “Varianz innerhalb des Clusters” oder kurz “Varianz within” bezeichnet. Natürlich wird diese

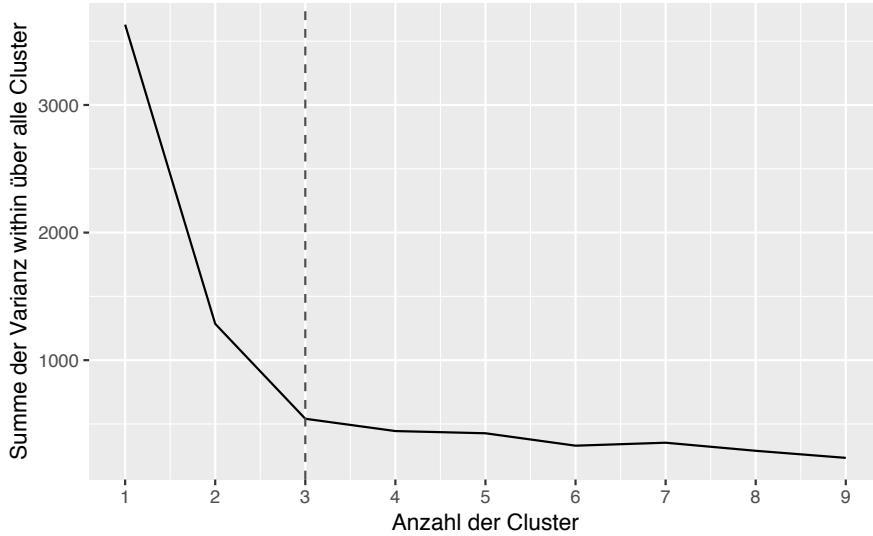


Abbildung 13.4: Die Summe der Varianz within in Abhängigkeit von der Anzahl von Clustern. Ein Screeplot.

Varianz within immer kleiner, je größer die Anzahl der Cluster wird.

Die vertikale gestrichelte Linie zeigt an, wo die Einsparung an Varianz auf einmal “sprunghaft” weniger wird - just an jedem Knick bei $x=3$; dieser “Knick” wird auch “Ellbogen” genannt (da sage einer, Statistiker haben keine Phantasie). Man kann jetzt sagen, dass 3 Cluster eine gute Lösung seien, weil mehr Cluster die Varianz innerhalb der Cluster nur noch wenig verringern. Diese Art von Diagramm wird als “Screeplot” bezeichnet. Fertig!

13.1.2 Euklidische Distanz

Aber wie weit liegen zwei Punkte entfernt? Betrachten wir ein Beispiel. Anna und Berta sind zwei Studentinnen, die eine Statistikklausur geschrieben haben müssen (bedauernswert). Die beiden unterscheiden sich sowohl in Lernzeit als auch in Klausurerfolg. Aber wie sehr unterscheiden sie sich? Wie groß ist der “Abstand” zwischen Anna und Berta (vgl. Abb. 13.5)?

Eine Möglichkeit, die Distanz zwischen zwei Punkten in der Ebene (2D) zu bestimmen, ist der *Satz des Pythagoras* (leise Trompetenfanfare). Generationen von Schülern haben diese Gleichung ähmm... geliebt:

$$c^2 = a^2 + b^2$$

In unserem Beispiel heißt das $c^2 = 3^2 + 4^2 = 25$. Folglich ist $\sqrt{c^2} = \sqrt{25} = 5$. Der Abstand oder der Unterschied zwischen Anna und Berta beträgt also 5 - diese Art von “Abstand” nennt man den *euklidischen Abstand*.

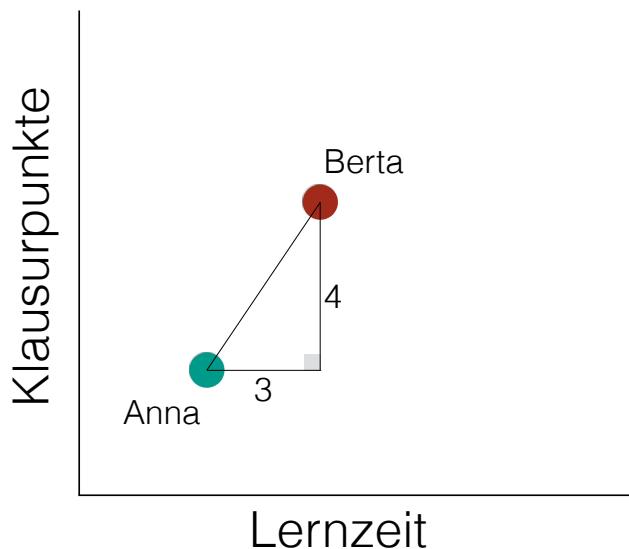


Abbildung 13.5: Distanz zwischen zwei Punkten in der Ebene

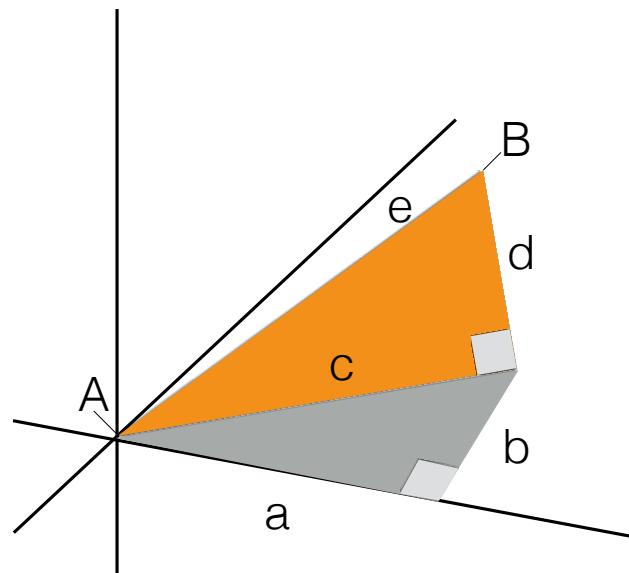


Abbildung 13.6: Pythagoras in 3D

Aber kann man den euklidischen Abstand auch in 3D (Raum) verwenden? Oder gar in Räumen mehr mehr Dimensionen??? Betrachten wir den Versuch, zwei Dreiecke in 3D zu zeichnen. Stellen wir uns vor, zusätzlich zu Lernzeit und Klausurerfolg hätten wir als 3. Merkmal der Studentinnen noch "Statistikliebe" erfasst (Bertas Statistikliebe ist um 2 Punkte höher als Annas).

Sie können sich Punkt A als Ecke eines Zimmers vorstellen; Punkt B schwebt dann in der Luft, in einiger Entfernung zu A.

Wieder suchen wir den Abstand zwischen den Punkten A und B. Wenn wir die Länge e wüssten, dann hätten wir die Lösung; e ist der Abstand zwischen A und B. Im orangenen

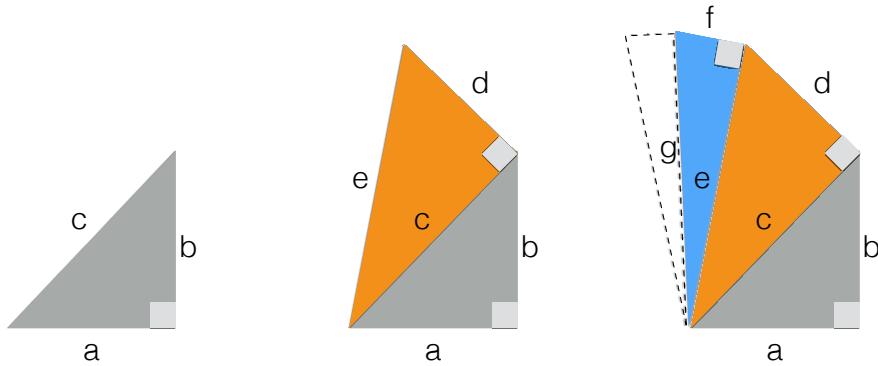


Abbildung 13.7: Pythagoras in Reihe geschaltet

Dreieck gilt wiederum der Satz von Pythagoras: $c^2 + d^2 = e^2$. Wenn wir also c und d wüssten, so könnten wir e berechnen... c haben wir ja gerade berechnet (5) und d ist einfach der Unterschied in Statistikliebe zwischen Anna und Berta (2)! Also

$$\begin{aligned} e^2 &= c^2 + d^2 \\ e^2 &= 5^2 + 2^2 \\ e^2 &= 25 + 4 \end{aligned}$$

$$e = \sqrt{29} \approx 5.4$$

Ah! Der Unterschied zwischen den beiden Studentinnen beträgt also ~ 5.4 !

Intuitiv gesprochen, “schalten wir mehrere Pythagoras-Sätze hintereinander”.

Der euklidische Abstand berechnet sich mit Pythagoras’ Satz!

Das geht nicht nur für “zwei Dreiecke hintereinander”, sondern der Algebra ist es wurscht, wie viele Dreiecke das sind.

Um den Abstand zweier Objekte mit k Merkmalen zu bestimmen, kann der euklidische Abstand berechnet werden mit. Bei $k=3$ Merkmalen lautet die Formel dann $e^2 = a^2 + b^2 + d^2$. Bei mehr als 3 Merkmalen erweitert sich die Formel entsprechend.

Dieser Gedanken ist mächtig! Wir können von allen möglichen Objekten den Unterschied bzw. die (euklidische) Distanz ausrechnen! Betrachten wir drei Professoren, die einschätzen sollten, wie sehr sie bestimmte Filme mögen (1: gar nicht; 10: sehr). Die Filme waren: “Die Sendung mit der Maus”, “Bugs Bunny”, “Rambo Teil 1”, “Vom Winde verweht” und “MacGyver”.

```
profs <- data_frame(
  film1 = c(9, 1, 8),
  film2 = c(8, 2, 7),
```

```

film3 = c(1, 8, 3),
film4 = c(2, 3, 2),
film5 = c(7, 2, 6)
)

```

Betrachten Sie die Film-Vorlieben der drei Professoren. Gibt es ähnliche Professoren hinsichtlich der Vorlieben? Welche Professoren haben einen größeren “Abstand” in ihren Vorlieben?

Wir könnten einen “fünffachen Pythagoras” zu Rate ziehen. Praktischerweise gibt es aber eine R-Funktion, die uns die Rechnerei abnimmt:

```

dist(profs)
#>      1     2
#> 2 13.23
#> 3 2.65 10.77

```

Offenbar ist der (euklidische) Abstand zwischen Prof. 1 und 2 groß (13.2); zwischen Prof 2 und 3 auch recht groß (10.8). Aber der Abstand zwischen Prof. 1 und 3 ist relativ klein! Endlich hätten wir diese Frage auch geklärt. Sprechen Sie Ihre Professoren auf deren Filmvorlieben an...

13.1.3 k-Means Clusteranalyse

Beim k-Means Clusterverfahren handelt es sich um eine bestimmte Form von Clusteranalysen; zahlreiche Alternativen existieren, aber die k-Means Clusteranalyse ist recht verbreitet. Im Gegensatz zur z.B. der hierarchischen Clusteranalyse um ein partitionierendes Verfahren. Die Daten werden in k Cluster aufgeteilt – dabei muss die Anzahl der Cluster im vorhinein feststehen. Ziel ist es, dass die Quadratsumme der Abweichungen der Beobachtungen im Cluster zum Clusterzentrum minimiert wird.

Der Ablauf des Verfahrens ist wie folgt:

1. Zufällige Beobachtungen als Clusterzentrum
2. Zuordnung der Beobachtungen zum nächsten Clusterzentrum (Ähnlichkeit, z. B. über die euklidische Distanz)
3. Neuberechnung der Clusterzentren als Mittelwert der dem Cluster zugeordneten Beobachtungen

Dabei werden die Schritte 2. und 3. solange wiederholt, bis sich keine Änderung der Zuordnung mehr ergibt – oder eine maximale Anzahl an Iterationen erreicht wurde. Aufgrund von (1.) hängt das Ergebnis einer k-Means Clusteranalyse vom Zufall ab. Aus Gründen der Reproduzierbarkeit sollte daher der Zufallszahlengenerator gesetzt werden (mit `set.seed`). Außerdem bietet es sich an verschiedene Startkonfigurationen zu versuchen. In der Funktion `kmeans()` erfolgt dies durch die Option `nstart` =.

13.2 Beispiel für eine einfache Clusteranalyse

Nehmen wir uns noch einmal den Extraversionsdatensatz vor. Kann man die Personen clustern anhand von Ähnlichkeiten wie Facebook-Freunde, Partyfrequenz und Katerhäufigkeit? Probieren wir es aus!

```
extra <- read.csv("data/extra.csv")
```

Verschaffen Sie sich einen Überblick mit der Funktion `glimpse`.

13.2.1 Distanzmaße berechnen

Auf Basis der drei metrischen Merkmale (d. h. `Alter`, `Einkommen` und `Kinder`), die wir hier aufs Geratewohl auswählen, ergeben sich für die ersten sechs Beobachtungen folgende Abstände:

```
extra %>%
  dplyr::select(n_facebook_friends, n_hangover, extra_single_item) %>%
  head %>%
  dist(.)
```

	1	2	3	4	5
#> 2	144.01				
#> 3	35.01	109.00			
#> 4	51.93	95.19	21.24		
#> 5	150.00	6.08	115.00	101.12	
#> 6	126.00	270.00	161.00	176.56	276.00

Sie können erkennen, dass die Beobachtungen 1 und 3 den kleinsten Abstand haben, während 1 und 5 den größten haben.

Allerdings hängen die Abstände von der Skalierung der Variablen ab (`n_facebook_friends` streut stärker als `extra_single_item`). Daher sollten wir die Variablen vor der Analyse zu standardisieren (z. B. über `scale()`).

Mit der Funktion `daisy()` aus dem Paket `cluster` kann man sich auch den Abstand zwischen den Objekten ausgeben lassen. Die Funktion errechnet auch Abstandsmaße, wenn die Objekte aus Variablen mit unterschiedlichen Skalenniveaus bestehen. Allerdings mag `daisy` Variablen vom Typ `chr` nicht, daher sollten wir `sex` zuerst in eine Faktorvariable umwandeln.

```
extra %>%
  dplyr::select(n_facebook_friends, sex, extra_single_item) %>%
  mutate(sex = factor(sex)) %>%
```

```
head %>%
  cluster::daisy(.)
```

13.2.2 kmeans für den Extraversionsdatensatz

Versuchen wir, einige Variablen mit `centers = 4` Clustern mithilfe einer kmeans-Clusteranalyse zu clustern.

```
set.seed(1896)

extra %>%
  mutate(Frau = sex == "Frau") %>%
  dplyr::select(n_facebook_friends, Frau, extra_single_item) %>%
  na.omit %>%
  scale -> extra_cluster

kmeans_extra_4 <- kmeans(extra_cluster, centers = 4, nstart = 10)
```

Lassen Sie sich das Objekt `extra_cluster` ausgeben und betrachten Sie die Ausgabe; auch `str(kmeans_extra_4)` ist interessant. Neben der Anzahl Beobachtungen pro Cluster (z. B. 337 in Cluster 2) werden auch die Clusterzentren ausgegeben. Diese können dann direkt verglichen werden. Schauen wir mal, in welchem Cluster die Anzahl der Facebookfreunde im Schnitt am kleinsten ist:

```
kmeans_extra_4$centers
#>   n_facebook_friends   Frau extra_single_item
#> 1      -0.0237  0.712          1.3792
#> 2      -0.0445  0.712         -0.3909
#> 3      -0.0368 -1.402         -0.0587
#> 4     25.6707 -1.402          1.3792
```

Betrachten Sie auch die Mittelwerte der anderen Variablen, die in die Clusteranalyse eingegangen sind. Wie ‘gut’ ist diese Clusterlösung? Vielleicht wäre ja eine andere Anzahl von Clustern besser? Eine Antwort darauf liefert die Varianz (Streuung) innerhalb der Cluster: Sind die Summen der quadrierten Abweichungen vom Clusterzentrum gering, so ist die Varianz ‘innerhalb’ der Cluster gering; die Cluster sind homogen und die Clusterlösung ist ‘gut’ (vgl. Abbildung 13.8).

Je größer die Varianz innerhalb der Cluster, um schlechter ist die Clusterlösung.

In zwei Dimensionen kann man Cluster gut visualisieren (Abbildung 13.3); in drei Dimensionen wird es schon unübersichtlich. Mehr Dimensionen sind schwierig. Daher ist es oft sinnvoll,

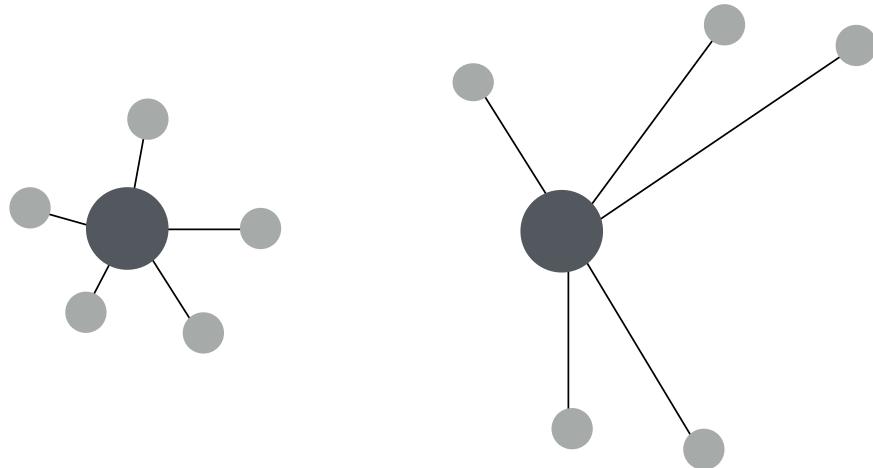


Abbildung 13.8: Schematische Darstellung zweier einfacher Clusterlösungen; links: geringe Varianz innerhalb der Cluster; rechts: hohe Varianz innerhalb der Cluster

die Anzahl der Dimensionen durch Verfahren der Dimensionsreduktion zu verringern. Die Hauptkomponentenanalyse oder die Faktorenanalyse bieten sich dafür an.

Vergleichen wir ein paar verschiedene Lösungen, um zu sehen, welche Lösung am besten zu sein scheint.

```
kmeans_extra_2 <- kmeans(extra_cluster, centers = 2, nstart = 10)
kmeans_extra_3 <- kmeans(extra_cluster, centers = 3, nstart = 10)
kmeans_extra_5 <- kmeans(extra_cluster, centers = 5, nstart = 10)
kmeans_extra_6 <- kmeans(extra_cluster, centers = 6, nstart = 10)
```

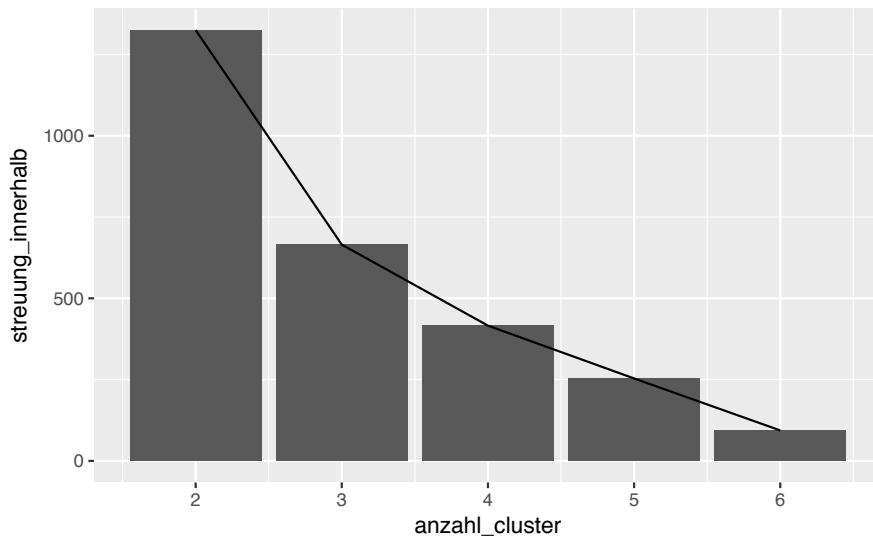
Dann nehmen wir die Gesamtstreuung jeder Lösung und erstellen daraus erst eine Liste und dann einen Dataframe:

```
streuung_innerhalb <- c(kmeans_extra_2$tot.withinss,
                         kmeans_extra_3$tot.withinss,
                         kmeans_extra_4$tot.withinss,
                         kmeans_extra_5$tot.withinss,
                         kmeans_extra_6$tot.withinss)

streuung_df <- data_frame(
  streuung_innerhalb,
  anzahl_cluster = 2:6
)
```

Jetzt plotten wir die Höhe der Streuung pro Clusteranalyse um einen Hinweis zu bekommen, welche Lösung am besten passen könnte.

```
ggplot(streuung_df) +
  aes(x = anzahl_cluster,
      y = streuung_innerhalb) +
  geom_col() +
  geom_line()
```



Nach der Lösung mit 4 Clustern kann man (vage) einen Knick ausmachen: Noch mehr Cluster verbessern die Streuung innerhalb der Cluster (und damit ihre Homogenität) nur noch unwesentlich oder zumindest deutlich weniger. Daher entscheiden wir uns für eine Lösung mit 4 Clustern.

13.3 Aufgaben¹



Richtig oder Falsch!?

1. Die Clusteranalyse wird gemeinhin dazu verwenden, Objekte nach Ähnlichkeit zu Gruppen zusammenzufassen.
2. Die Varianz innerhalb eines Clusters kann als Maß für die Anzahl der zu extrahierenden Cluster herangezogen werden.
3. Unter euklidischer Distanz versteht jedes Maß, welches den Abstand zwischen Punkten in der Ebene misst.
4. Bei der k-means-Clusteranalyse darf man die Anzahl der zu extrahierenden Clustern nicht vorab festlegen.

¹R, R, F, F, R

Tabelle 13.1: Befehle des Kapitels 'Clusteranalyse'

Paket::Funktion	Beschreibung
dist	Berechnet den euklidischen Abstand zwischen Vektoren
dplyr::glimpse	Stellt einen Dataframe im Überblick dar
cluster::daisy	Berechnet verschiedene Abstandsmaße
set.seed	Zufallsgenerator auf bestimmte Zahlen festlegen
cluster::clusplot	Visualisiert eine Clusteranalyse

5. Cluster einer k-means-Clusteranalyse werden so bestimmt, dass die Cluster möglichst homogen sind, d.h. möglichst wenig Streuung aufweisen (m.a.W. möglichst nah am Cluster-Zentrum sind).

Laden Sie den Datensatz **extra** zur Extraversion.

1. Unter Berücksichtigung der 10 Extraversionsitems: Lassen sich die Teilnehmer der Umfrage in eine Gruppe oder in mehrere Gruppen einteilen? Wenn in mehrere Gruppen, wie viele Gruppen passen am besten?
2. Berücksichtigen Sie den Extraversionsmittelwert und einige andere Variablen aus dem Datensatz (aber nicht die Items). Welche Gruppen ergeben sich? Versuchen Sie die Gruppen zu interpretieren!
3. Suchen Sie sich zwei Variablen aus dem Datensatz und führen Sie auf dieser Basis eine Clusteranalyse durch. Visualisieren Sie das Ergebnis anhand eines Streudiagrammes!

13.4 Befehlsübersicht

Tabelle 13.1 fasst die R-Funktionen dieses Kapitels zusammen.

13.5 Verweise

- Diese Übung orientiert sich am Beispiel aus Kapitel 11.3 aus Chapman und Feit (2015) und steht unter der Lizenz Creative Commons Attribution-ShareAlike 3.0 Unported². Der Code steht unter der Apache Lizenz 2.0³
- Der erste Teil dieser Übung basiert auf diesem Skript: <https://cran.r-project.org/web/packages/broom/vignettes/kmeans.html>

²<http://creativecommons.org/licenses/by-sa/3.0>

³<http://www.apache.org/licenses/LICENSE-2.0>

- Eine weiterführende, aber gut verständliche Einführung findet sich bei James, Witten, Hastie, und Tibshirani (2013c).
- Die Intuition zum euklidischen Abstand mit Pythagoras' Satz kann hier im Detail nachgelesen werden: <https://betterexplained.com/articles/measure-any-distance-with-the-pythagorean-theorem/>.

Kapitel 14

Vertiefung: Dimensionsreduktion



Lernziele:

- Den Unterschied zwischen einer Hauptkomponentenanalyse und einer Exploratorischen Faktorenanalyse kennen
- Methoden kennen, um die Anzahl von Dimensionen zu bestimmen
- Methoden der Visualisierung anwenden können
- Umsetzungsmethoden in R anwenden können
- Ergebnisse interpretieren können.

In diesem Kapitel werden folgende Pakete benötigt:

```
library(corrplot) # für `corrplot`  
library(gplots) # für `heatmap.2`  
library(nFactors) # PCA und EFA  
library(tidyverse) # Datenjudo  
library(psych) # für z.B. 'alpha'
```

14.1 Einführung

Häufig möchte man in den Sozialwissenschaften *latente Variablen* messen - z.B. Arbeitszufriedenheit, Extraversion, Schmerz oder Intelligenz. Solche Variablen nennt man *latent*, da man sie nicht direkt messen kann¹.

Konstrukte bezeichnen gedankliche bzw. theoretische Sachverhalt dar, die nicht direkt beobachtbar und damit nicht direkt messbar sind.

Komplementär zu latenten Konstrukten spricht man von manifesten Variablen, wie Schuhgröße oder Körpergewicht; Dinge also, die man in gewohntem Sinne beobachtbar messen kann. Messung von manifesten Variablen bezeichnet man auch als *extensives Messen* (Michell 2000).

Was ist eigentlich *Messen*? Sagen wir, ich finde den Urmeter auf der Straße (Details tun nichts zur Sache). Dann betrachte ich intensiv den Weg von meinem Carport bis zu meiner Haustür. Alsdann schaue ich, wie oft ich den Urmeter hintereinander legen muss, um den Weg von Haustür zu Carport zurückzulegen. Voilà! Die Länge des Weges ist *gemessen*. Allgemein ist Messen - nach diesem Verständnis - also das Vielfache eines Maßstabes in einer Größe (Michell 2000, aber s. Eid, Gollwitzer, und Schmitt (2010) für eine andere, verbreitete Definition).

Nach einer anderen Art von Messdefinition ist Messen alles, was aus manifesten Variablen eine Zahl erzeugt (Michell 2000). Das ist das Verständnis von Messen der meisten Sozialwissenschaftler (doch, im Ernst). Bei Lichte betrachtet "misst" man in den Sozialwissenschaften gerne so:

1. Such ein paar Variablen, die mit dem zu messenden Konstrukt zu tun haben könnten (z.B. Extraversion)
2. Frage ein paar Leute, wie sich selber einschätzen in diesen Variablen
3. Die Antwortskala denkst Du Dir nach eigenem Gusto aus (z.B. von 1 bis 10).
4. Addiere die Punkte aller Variablen auf.
5. Verkündige, dass Du Dein Konstrukt gemessen hast. Fertig.

Natürlich ist das ... nicht ganz richtig. Zumindest kann man nicht sicher sein, dass man Extraversion gemessen hat, oder ausreichend genau gemessen hat.

Da komplexe Phänomene wie Extraversion facettenreich sind, nimmt man häufig mehrere manifeste Variablen und bezeichnet deren Mittelwert dann als Messung von Extraversion. Handelt es sich um Kreuze in einer Befragung, so spricht man von *Items (Indikatoren)*.

Eine notwendige (aber nicht hinreichende) Voraussetzung, dass eine Reihe von Items sagen wir Extraversion messen, ist, dass sie miteinander stark korrelieren. Wenn sie das tun, so kann man sie auf *eine* Variable zusammenfassen, welche dann als Extraversion bezeichnet wird.

In diesem Kapitel betrachten wir zwei gängige Methoden solcher Zusammenfassungsmethoden. Da diese Methoden Variablen zusammenfassen, spricht man *Dimensionsreduktion*. Wir setzen voraus, dass es sich um metrische Variablen handelt (wir prüfen das nicht weiter).

¹Halt, da vorn läuft ein IQ-Punkt. Schnell, fangt ihn!

- Die *Hauptkomponentenanalyse* (engl. principal component analysis, PCA) versucht, unkorrelierte Linearkombinationen zu finden, die die Gesamtvarianz in den Daten erfassen. Die PCA beinhaltet also das Extrahieren von linearen Zusammenhängen der beobachteten Variablen.
- Die *Exploratorische Faktorenanalyse (EFA)* versucht, die Varianz auf Basis einer kleinen Anzahl von Dimensionen zu modellieren, während sie gleichzeitig versucht, die Dimensionen in Bezug auf die ursprünglichen Variablen interpretierbar zu machen. Es wird davon ausgegangen, dass die Daten einem Faktoren Modell entsprechen, bei der die beobachteten Korrelationen auf latente Faktoren zurückführen. Mit der EFA wird *nicht* die gesamte Varianz erklärt.

Die EFA wird oft als *Common Factor Analysis* oder *principal axis analysis (Hauptachsenanalyse)* bezeichnet. Die EFA eröffnet dem Nutzer eine Menge an analytischen Varianten, so dass das Ergebnis, im Gegensatz zu PCA, recht unterschiedlich ausfallen kann. Es gibt also *keine einzige* Lösung bei der EFA. Wichtig ist, genau zu berichten, welche Details man verwendet hat.

Eine einfache Faustregel für die Entscheidung zwischen diesen beiden Methoden:

- Führe die PCA durch, wenn die korrelierten beobachteten Variablen einfach auf einen kleineren Satz von wichtigen unabhängigen zusammengesetzten Variablen reduziert werden soll.
- Führe die EFA durch, wenn ein theoretisches Modell von latenten Faktoren zugrunde liegt, dass die beobachtete Variablen verursacht.

14.2 Warum Datenreduktion wichtig ist

- *Dimensionen reduzieren:* Im technischen Sinne der Dimensionsreduktion können wir statt Variablen-Sets die Faktor-/ Komponentenwerte verwenden (z. B. für Mittelwertvergleiche zwischen Experimental- und Kontrollgruppe, Regressionsanalyse und Clusteranalyse).
- *Unsicherheit verringern:* Wenn wir glauben, dass ein Konstrukt nicht eindeutig messbar ist, dann kann mit einem Variablen-Set die Unsicherheit reduziert werden.
- *Aufwand verringern:* Wir können den Aufwand bei der Datenerfassung vereinfachen, indem wir uns auf Variablen konzentrieren, von denen bekannt ist, dass sie einen hohen Beitrag zum interessierenden Faktor/ Komponente leisten. Wenn wir feststellen, dass einige Variablen für einen Faktor nicht wichtig sind, können wir sie aus dem Datensatz eliminieren. Außerdem werden die statistischen Modelle einfacher, wenn wir statt vieler Ausgangsvariablen einige wenige Komponenten/Faktoren als EingabevARIABLEN verwenden.

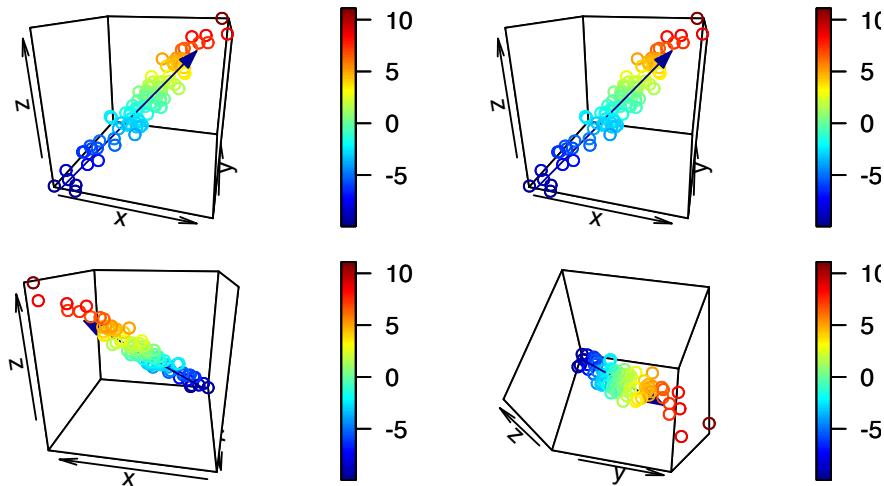


Abbildung 14.1: Der Pfeil ist eindimensional; reduziert also die drei Dimensionen auf eine

14.3 Intuition zur Dimensionsreduktion

Betrachten Sie die Visualisierung eines Datensatzes mit 3 Dimensionen (Spalten) in Abbildung 14.1). Man braucht nicht viel Phantasie, um einen Pfeil (Vektor) in der Punktwolke zu sehen. Um jeden Punkt einigermaßen genau zu bestimmen, reicht es, seine “Pfeil-Koordinate” zu wissen. Praktischerweise geben in Abbildung 14.1 die Farben (in etwa) die Koordinaten auf dem Pfeil an². Damit können wir die Anzahl der Variablen (Dimensionen), die es braucht, um einen Punkt zu beschreiben von 3 auf 1 reduzieren; 2/3 der Komplexität eingespart. Wir verlieren etwas Genauigkeit, aber nicht viel. Dieser Pfeil, der mitten durch den Punkteschwarm geht, nennt man auch die *1. Hauptkomponente*.

Beachten Sie, dass hoch korrelierte Variablen eng an der Regressionsgeraden liegen; entsprechend sind in Abbildung 14.1 die drei Variablen stark korreliert. Sehen Sie auch, dass die Hauptkomponente Varianz erklärt: Jede Variable für sich genommen, hat recht viel Streuung. Die Streuung der Punkte zur Hauptkomponente ist aber relativ gering. Daher sagt man, die Streuung (Varianz) wurde reduziert durch die Hauptkomponente.

Der längste Vektor, den man in die Punktwolke legen kann, bezeichnet man als den *1. Eigenvektor* oder die *1. Hauptkomponente*.

In Abbildung 14.1 ist dieser als Pfeil eingezeichnet³. Weitere Hauptkomponenten kann man nach dem gleichen Muster bestimmen mit der Auflage, dass sie im *rechten Winkel* zu bestehenden Hauptkomponenten liegen. Damit kann man in einer 3D-Raum nicht mehr als 3 Hauptkomponenten bestehen (in einem n -dimensionalen Raum also maximal n Hauptkomponenten).

Hauptkomponenten liegen stets im rechten Winkel zueinander (‘orthogonal’). Das

²genau genommen ist hier die Regressionsgerade gezeichnet, es müsste aber der größte Eigenvektor sein. Geschenkt.

³die Hauptkomponente ist hier ähnlich zur Regressionslinie, aber nicht identisch

bedeutet, dass Werte, die auf verschiedenen Hauptkomponenten liegen, unkorreliert sind.

```
#>      V1     V2     V3
#> V1 1.000 0.933 0.955
#> V2 0.933 1.000 0.833
#> V3 0.955 0.833 1.000
```

Je stärker die Korrelation zwischen Variablen, desto besser kann man sie zusammenfassen.

14.4 Datensatz ‘Werte’

Wir untersuchen die Dimensionalität mittels einer auf 1000 Fälle reduzierten Zufallsauswahl von 15 Variablen zur Messung der grundlegenden Wertorientierungen von Menschen. Die Daten wurden im Sommersemester 2017 von FOM Studierenden im ersten Semester an der FOM bundesweit erhoben. Die Variablen zu Wertorientierungen wurden ursprüngliche aus dem 40-Item-Set des Portraits Value Questionnaire» (PVQ) von Schmidt u. a. (2007) adaptiert und durch Studien an der FOM seit 2014 stufenweise bis auf 15 relevante Variablen reduziert. Alle Variablen wurden auf einer Skala von 1 bis 7 (wobei 1 am wenigsten und 7 am meisten zutrifft) abgefragt.

Laden wir zunächst den Datensatz:

```
Werte <- read.csv("data/Werte.csv")
```

Wir überprüfen zuerst die Struktur des Datensatzes, die ersten 6 Zeilen und die Zusammenfassung. Probieren Sie die folgenden Befehle aus:

```
glimpse(Werte)
```

Wir sehen mit `glimpse`, dass die Bereiche der Bewertungen für jede Variable 1-7 sind. Außerdem sehen wir, dass die Bewertungen als numerisch (Integer, also ganzzahlig) eingelesen wurden. Die Daten sind somit offenbar richtig formatiert.

14.5 Neuskalierung der Daten

In vielen Fällen ist es sinnvoll, Rohdaten neu zu skalieren - auch bei der Dimensionsreduktion. Warum ist das nötig?

Dies wird üblicherweise als *Standardisierung*, *Normierung*, oder *Z-Transformation* bezeichnet. Als Ergebnis ist der Mittelwert aller Variablen über alle Beobachtungen dann 0 und die

Standardabweichung (SD) 1. Da wir hier gleiche Skalenstufen haben, ist ein Skalieren nicht unbedingt notwendig, wir führen es aber trotzdem durch.

Ein einfacher Weg, alle Variablen im Datensatz auf einmal zu skalieren ist der Befehl `scale()`. Da wir die Rohdaten nie ändern wollen, weisen wir die Rohwerte zuerst einem neuen Dataframe `Werte.skaliert` zu und skalieren anschließend die Daten. Wir skalieren in unserem Datensatz alle Variablen.

```
Werte %>% scale %>% as_tibble -> Werte.skaliert
summary(Werte.skaliert)
```



Nimm das Objekt (ein Dataframe) `Werte` UND DANN
z-skaliere das Objekt UND DANN
definiere es als Dataframe (genauer: tibble) UND speichere dies unter dem Namen
`Werte_skaliert`. FERTIG.

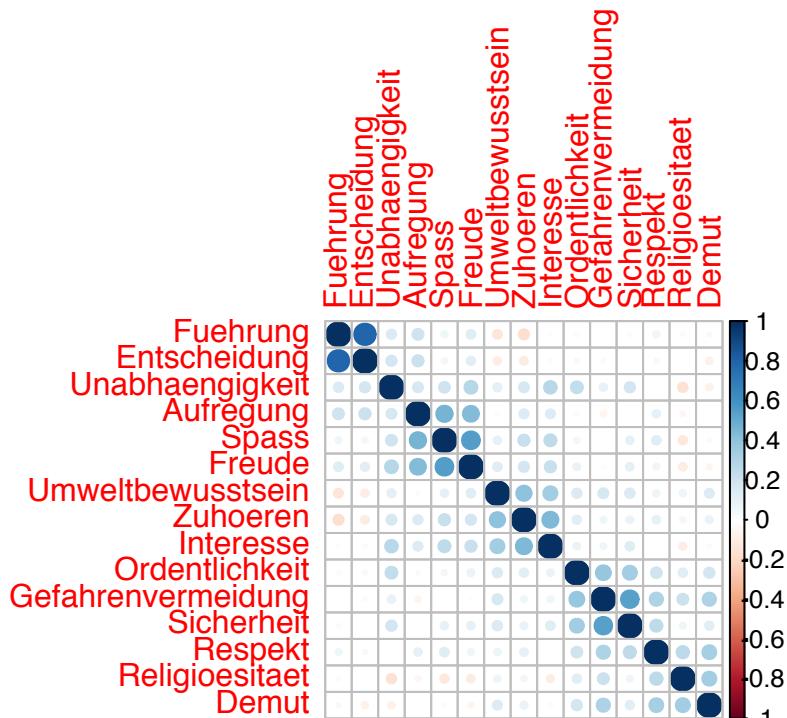
Ach ja, dann zeig noch ein `summary` von diesem Objekt.

Die Daten wurden richtig skaliert, da der Mittelwert aller Variablen über alle Beobachtungen 0 und die sd 1 ist.

14.6 Zusammenhänge in den Daten

Wir verwenden den Befehl `corrplot()` für die Erstinspektion von bivariaten Beziehungen zwischen den Variablen. Das Argument `order = "hclust"` ordnet die Zeilen und Spalten entsprechend der Ähnlichkeit der Variablen in einer hierarchischen Cluster-Lösung der Variablen (mehr dazu im Kapitel 13) neu an.

```
corrplot(cor(Werte.skaliert), order = "hclust")
```



Die Visualisierung der Korrelation der Variablen scheint fünf Cluster zu zeigen:

- (“Führung”, “Entscheidung”)
- (“Aufregung”, “Spaß”, “Freude”)
- (“Umweltbewusstsein”, “Zuhören”, “Interesse”)
- (“Ordnlichkeit”, “Gefahrenvermeidung”, “Sicherheit”)
- (“Respekt”, “Religiösität”, “Demut”)

14.7 Daten mit fehlende Werten

Wenn in den Daten leere Zellen, also fehlende Werte, vorhanden sind, dann kann es bei bestimmten Rechenoperationen zu Fehlermeldungen kommen. Dies betrifft zum Beispiel Korrelationen, PCA und EFA. Der Ansatz besteht deshalb darin, NA-Werte explizit zu entfernen. Dies kann am einfachsten mit dem Befehl `na.omit()` geschehen:

Beispiel:

```
Werte.skaliert <- na.omit(Werte.skaliert)
corrplot(cor(Werte.skaliert), order = "hclust")
```

Da wir in unserem Datensatz vollständige Daten verwenden, gibt es auch keine Leerzellen.

Mit dem Parameter `order` kann man die Reihenfolge (order) der Variablen, wie sie im Diagramm dargestellt werden ändern (vgl `help(corrplot)`). Hier haben wir die Variablen nach Ähnlichkeit aufgereiht: Ähnliche Variablen stehen näher beieinander. Damit können wir

gut erkennen, welche Variablen sich ähnlich sind (hoch korreliert sind) und somit Kandidaten für eine Einsparung (Zusammenfassung zu einer Hauptkomponente bzw. einem Faktor) sind.

14.8 Hauptkomponentenanalyse (PCA)

Die PCA berechnet ein Variablenset (Komponenten) in Form von linearen Gleichungen, die die linearen Beziehungen in den Daten erfassen. Die erste Komponente erfasst so viel Streuung (Varianz) wie möglich von allen Variablen als eine einzige lineare Funktion. Die zweite Komponente erfasst unkorreliert zur ersten Komponente so viel Streuung wie möglich, die nach der ersten Komponente verbleibt. Das geht so lange weiter, bis es so viele Komponenten gibt wie Variablen.

14.8.1 Bestimmung der Anzahl der Hauptkomponenten

Betrachten wir in einem ersten Schritt die wichtigsten Komponenten für die Werte. Wir finden die Komponenten mit `prcomp()`.

```
Werte.pc <- prcomp(Werte.skaliert) # Principal Components berechnen
summary(Werte.pc)
#> Importance of components%>%
#>                 PC1     PC2     PC3     PC4     PC5     PC6     PC7
#> Standard deviation    1.691  1.542  1.384  1.1428  1.0797  0.8855  0.8298
#> Proportion of Variance 0.191  0.159  0.128  0.0871  0.0777  0.0523  0.0459
#> Cumulative Proportion   0.191  0.349  0.477  0.5639  0.6416  0.6939  0.7398
#>                         PC8     PC9     PC10    PC11    PC12    PC13    PC14
#> Standard deviation      0.8078  0.7882  0.7599  0.7413  0.6884  0.648   0.6449
#> Proportion of Variance  0.0435  0.0414  0.0385  0.0366  0.0316  0.028   0.0277
#> Cumulative Proportion   0.7833  0.8247  0.8632  0.8999  0.9315  0.959   0.9872
#>                         PC15
#> Standard deviation      0.4388
#> Proportion of Variance  0.0128
#> Cumulative Proportion   1.0000
```

```
# Berechnung der Gesamtvarianz
Gesamtvarianz <- sum(Werte.pc$sdev^2)

# Bei sum(Werte.pc$sdev^2) wird die Summe aller 15 Standardabweichungen berechnet.

# Varianzanteil der ersten Hauptkomponente
Werte.pc$sdev[1]^2 / Gesamtvarianz
#> [1] 0.191
```

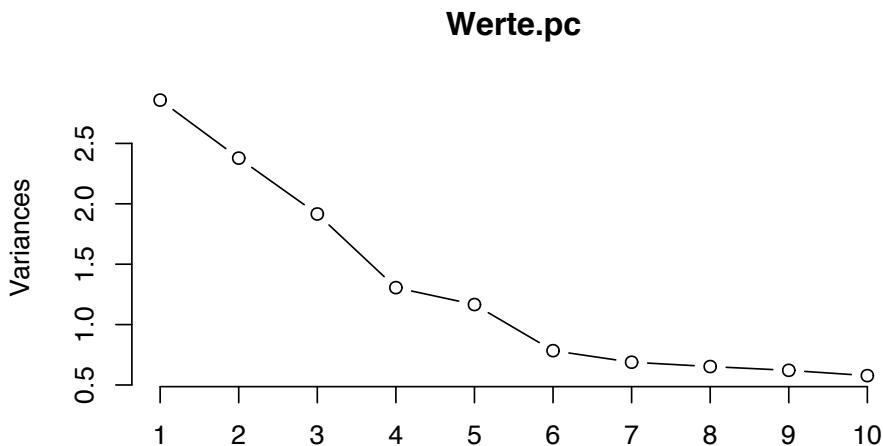


Abbildung 14.2: Screeplot

14.8.2 Scree-Plot

Der Standard-Plot `plot()` für die PCA ist ein *Scree-Plot*⁴. Dieser zeigt uns die jeweils durch eine Hauptkomponente erfasste Streuung (Varianz). Wir plotten ein Liniendiagramm mit dem Argument `type = "l"` (1 für Linie), s. Abb. 14.2).

```
plot(Werte.pc, type="l")
```

Die Höhe der Varianz entspricht der Länge der Pfeile (Eigenvektoren) in Abbildung 14.1: Längere Pfeile bedeuten größere erklärte Varianz. Die Länge der Eigenvektoren bezeichnet man auch als Eigenwert.

Wir sehen in Abb. 14.2, dass bei den Werte-Daten der Anteil der Streuung nach der fünften Komponente nicht mehr wesentlich abnimmt. Es soll die Stelle gefunden werden, ab der die Varianzen der Hauptkomponenten deutlich kleiner sind. Je kleiner die Varianzen, desto weniger Streuung erklärt diese Hauptkomponente.

14.8.3 Ellbogen-Kriterium

Nach dem *Ellbogen-Kriterium* werden alle Hauptkomponenten berücksichtigt, die links von der Knickstelle im Scree-Plot liegen. Gibt es mehrere Knicks, dann werden jene Hauptkomponenten ausgewählt, die links vom rechtenen Knick liegen. Gibt es keinen Knick, dann hilft der Scree-Plot nicht weiter. Bei den Werte-Daten tritt der Ellbogen, je nach Betrachtungsweise,

⁴scree: engl. "Geröll"

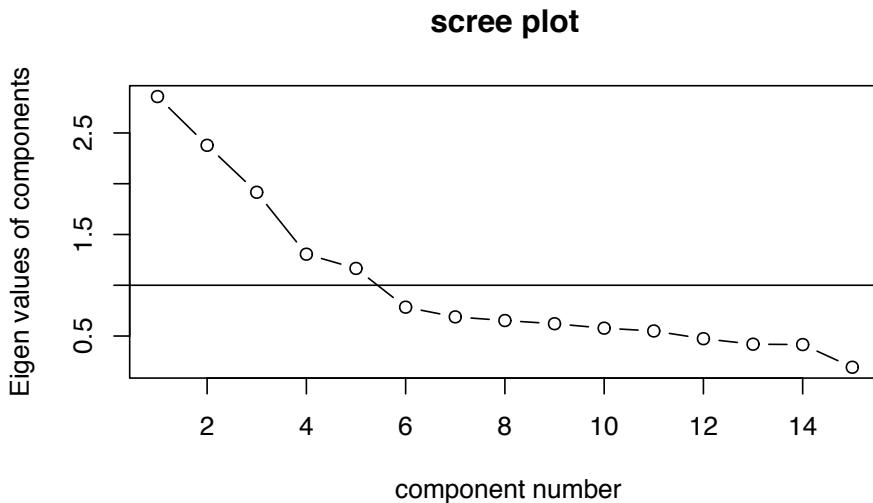


Abbildung 14.3: VSS-Screepplot

entweder bei vier oder sechs Komponenten auf. Dies deutet darauf hin, dass die ersten fünf Komponenten die meiste Streuung in den Werte-Daten erklären.

14.8.4 Eigenwert-Kriterium

Der *Eigenwert* ist eine Metrik für den Anteil der erklärten Varianz pro Hauptkomponente. Die Anzahl Eigenwerte können wir über den Befehl `eigen()` ausgeben.

```
eigen(cor(Werte))
```

Der Eigenwert einer Komponente/ eines Faktors sagt aus, wie viel Varianz dieser Faktor an der Gesamtvarianz aufklärt. Laut dem Eigenwert-Kriterium sollen nur Faktoren mit einem *Eigenwert größer 1* extrahiert werden. Dies sind bei den Werte-Daten fünf Komponenten/ Faktoren, da fünf Eigenwerte größer 1 sind. Der Grund ist, dass Komponenten/ Faktoren mit einem Eigenwert kleiner als 1 weniger Erklärungswert haben als die ursprünglichen Variablen.

Dies kann auch grafisch mit dem `psych::VSS.Scree`⁵ geplottet werden (s. Abb. 14.3).

```
VSS.scree(Werte)
```

14.8.5 Biplot

Eine gute Möglichkeit die Ergebnisse der PCA zu analysieren, besteht darin, die ersten Komponenten zuzuordnen, die es uns ermöglichen, die Daten in einem niedrigdimensionalen

⁵das Paket `psych` wird automatisch vom Paket `nfactors` gestartet, sie müssen es nicht extra starten

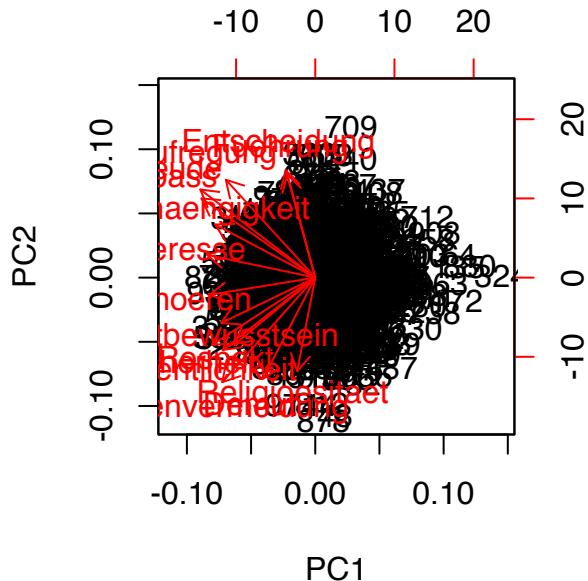


Abbildung 14.4: Ein Biplot für den Werte-Datensatz

Raum zu visualisieren. Eine gemeinsame Visualisierung ist ein *Biplot*. Ein Biplot zeigt die Ausprägungen der Fälle auf den ersten beiden Hauptkomponenten. Häufig sind die beiden ersten Hauptkomponenten schon recht aussagekräftig, vereinen also einen Gutteil der Streuung auf sich. Dazu verwenden wir `biplot()` (s. Abbildung 14.4).

```
biplot(Werte.pc)
```

Die einzelnen Ausgangsvariablen sind in Abbildung Abbildung 14.4 durch rote Pfeile (Vektoren) gekennzeichnet.

Je paralleler der Vektor einer Ausgangsvariable zur X-Achse (1. Hauptkomponente) ist, umso identischer sind sich die entsprechende Variable und die Hauptkomponente (für die Y-Achse gilt entsprechendes). Das hilft uns, die Hauptkomponente inhaltlich zu interpretieren. Hauptkomponenten (oder Faktoren) sollten stets inhaltlich interpretiert werden - auch wenn eine subjektive Komponente mitschwingt.

Die 1. Hauptkomponente wird offenbar stark geprägt durch die Ausgangsvariablen ‘Freude’ und ‘Spaß’. Bei der 2. Hauptkomponente analog durch ‘Demut’ und ‘Gefahrenvermeidung’.

Zusätzlich erhalten wir einen Einblick in die Bewertungscluster (als dichte Bereiche von Beobachtungspunkten): Gruppen von Punkten entsprechen ähnlichen Fällen (ähnlich hinsichtlich ihrer Werte in den ersten zwei Hauptkomponenten). Der Biplot ist hier durch die große Anzahl an Beobachtung allerdings recht unübersichtlich.

14.8.6 Aufgaben



1. Ziehen Sie eine Zufallsstichprobe aus dem Datensatz, berechnen Sie die PCA erneut und betrachten Sie den Biplot. Wie stark ist die Änderung?
2. Erstellen Sie mehrere Streudiagramme und überprüfen Sie die bivariaten Zusammenhänge (die ja zur Dimensionsreduktion führen) visuell.

Am einfachsten lassen sich die Komponenten extrahieren mit dem `principal`-Befehl aus dem Paket `psych`:

```
Werte.pca <- principal(Werte, nfactors = 5, rotate = "none")
print(Werte.pca, cut = 0.5, sort = TRUE, digits = 2)
```

`cut = 0.5` heißt, dass nur Ladungen ab 0.5 angezeigt werden sollen. Mit `rotate = 'none'` sagen wir, dass wir keine Rotation wünschen. Eine Rotation ist

14.8.7 Interpretation der Ergebnisse der PCA

Das Ergebnis sieht sehr gut aus. Es laden immer mehrere Items (Ausgangsvariablen) (mindestens 2) hoch (> 0.5) auf einer Komponente (die mit RC1 bis RC5 bezeichnet werden, *RC* steht für *Rotated Component*). Mit “laden” ist die Parallelität der Ausgangsvariable zur Hauptkomponente gemeint. Vereinfacht gesprochen ist die Ladung die Korrelation der Items mit der jeweiligen Komponente.

Innerhalb einer PCA kann die Interpretierbarkeit über eine **Rotation** erhöht werden. Wenn die Rotation nicht ausgeschlossen wird (mit dem Argument `rotate="none"`), dann ist die Voreinstellung eine **Varimax-Rotation**.

Mit `h2` (Kommunalität) ist der Anteil eines Items bezeichnet, der durch die Komponenten insgesamt erklärt wird. Hier haben die Anzahl der Komponenten auf 5 beschränkt. Daher wird nicht die ganze Varianz des Items erklärt.

Es gibt keine Items die auf mehr als einer Komponente hoch laden. Die Ladungen sind Korrelationskoeffizienten zwischen den Items und den Hauptkomponenten. In der Zeile *SS loadings* finden wir die Eigenwerte der fünf Hauptkomponenten (berechnet als Summe der quadrierten Ladungen). Den Anteil an der Gesamtvarianz, den sie erklären, findet man in der Zeile *Proportion Var.* Aufsummiert sind die Anteile in der Zeile *Cumulative Var.* Insgesamt werden durch die fünf Hauptkomponenten 64% der Gesamtvarianz erklärt. Die starke Hauptkomponente hat einen Eigenwert von 2.08 und erklärt 14% der Varianz.

Einzig das Item “Unabhängigkeit” lädt auf keine der Hauptkomponenten hoch.

Um die inhaltliche Bedeutung der Komponenten zu interpretieren, schauen wir uns die Inhalte der jeweiligen Items an und versuchen hierfür einen inhaltlichen Gesamtbegriff zu finden.

Die Erste Komponenten könnte mit **Genuss**, die zweite mit **Sicherheit**, die dritte mit **Bewusstsein**, die vierte mit **Konformismus** und die fünfte mit **Anerkennung** bezeichnet werden.

Mit der Funktion `fa.diagram` kann das Ergebnis auch grafisch dargestellt werden: `fa.diagram(Werte.pca)`.

14.9 Exploratorische Faktorenanalyse (EFA)

Genau genommen ist der Begriff *Faktorenanalyse (FA)* ein Überbegriff für mehrere Arten von ähnlichen Verfahren der Dimensionsreduktion. Ein Beispiel für eine Art von Faktorenanalyse wäre dann die PCA. Aber der Begriff Faktorenanalyse wird auch verwendet, um eine bestimmte Art von Faktorenanalyse - sozusagen eine Faktorenanalyse im engeren Sinne - zu bezeichnen. Wir halten uns hier an letztere Begriffskonvention.

In diesem Sinne ist die *Exploratorische Faktorenanalyse (EFA)* eine Methode, um die Beziehung von Konstrukten (Konzepten), d. h. Faktoren zu Variablen zu beurteilen. Dabei werden die Faktoren als *latente Variablen* betrachtet, die nicht direkt beobachtet werden können. Stattdessen werden sie empirisch durch mehrere Variablen beobachtet, von denen jede ein Indikator der zugrunde liegenden Faktoren ist. Diese beobachteten Werte werden als *manifeste Variablen* bezeichnet und umfassen Indikatoren. Die EFA versucht den Grad zu bestimmen, in dem Faktoren die beobachtete Streuung der manifesten Variablen berücksichtigen.

Das Ergebnis der EFA ist ähnlich zur PCA: eine Matrix von Faktoren (ähnlich zu den PCA-Komponenten) und ihre Beziehung zu den ursprünglichen Variablen (Ladung der Faktoren auf die Variablen). Im Gegensatz zur PCA versucht die EFA, Lösungen zu finden, die in den *manifesten variablen maximal interpretierbar* sind. Im Allgemeinen versucht sie, Lösungen zu finden, bei denen eine kleine Anzahl von Ladungen für jeden Faktor sehr hoch ist, während andere Ladungen für diesen Faktor gering sind. Wenn dies möglich ist, kann dieser Faktor mit diesem Variablen-Set interpretiert werden.

14.9.1 Finden einer EFA Lösung

Als erstes muss die Anzahl der zu schätzenden Faktoren bestimmt werden. Hierzu verwenden wir wieder das Ellbow-Kriterium und das Eigenwert-Kriterium. Beide Kriterien haben wir schon bei der PCA verwendet, dabei kommen wir auf 5 Faktoren.

Durch das Paket `nFactors` bekommen wir eine ausgedehntere Berechnung der Scree-Plot Lösung mit dem Befehl `nScree()` - es werden noch weitere, sophistiziertere Methoden zur Berechnung der 'richtigen' Anzahl von Faktoren eingesetzt. Wir sparen uns hier die Details.

```
nScree(Werte)
```

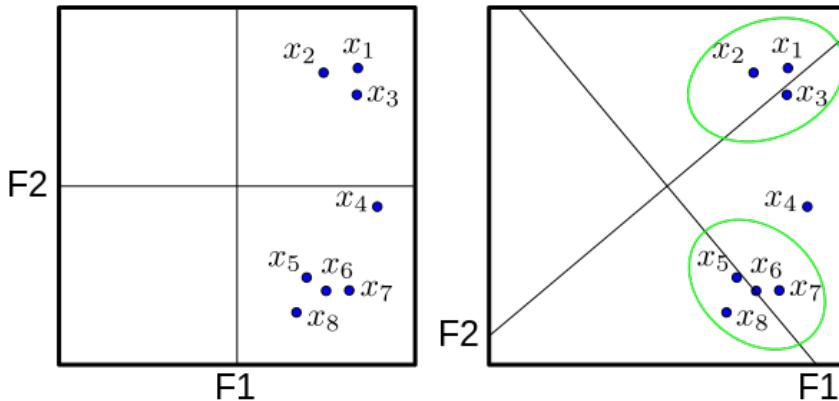


Abbildung 14.5: Beispiel für eine rechtwinklige Rotation

nScree gibt vier methodische Schätzungen für die Anzahl an Faktoren durch den Scree-Plot aus. Wir sehen, dass drei von vier Methoden fünf Faktoren vorschlagen. Nach kurzer Überlegung und Blick aus dem Fenster entscheiden wir uns für 5 Faktoren.

14.9.2 Schätzung der EFA

Eine EFA wird geschätzt mit dem Befehl `factanal(x, factors = k)`, wobei `k` die Anzahl Faktoren angibt und `x` den Datensatz.

```
Werte.fa<-factanal(Werte, factors = 5)
Werte.fa
```

Eine übersichtlichere Ausgabe bekommen wir mit dem `print` Befehl, in dem wir zusätzlich noch die Dezimalstellen kürzen mit `digits = 2`, alle Ladungen kleiner als 0,5 ausblenden mit `cutoff = .4` und die Ladungen mit `sort = TRUE` so sortieren, dass die Ladungen, die auf einen Faktor laden, untereinander stehen.

```
print(Werte.fa, digits = 2, cutoff = .4, sort = TRUE)
```

Standardmäßig wird bei `factanal()` eine *Varimax-Rotation* durchgeführt (das Koordinatensystem der Faktoren wird so rotiert, das eine optimale Zuordnung zu den Variablen erfolgt). Bei Varimax gibt es keine Korrelationen zwischen den Faktoren. Sollen Korrelationen zwischen den Faktoren zugelassen werden, empfiehlt sich die Oblimin-Rotation mit dem Argument `rotation="oblimin"` aus dem Paket `GPArotation`.

Das eine Rotation sinnvoll ist, kann man sich am einfachsten an einem Diagramm verdeutlichen (s. Abbildung 14.5, (Fjälnes 2014)).

Das Rotieren kann man sich als Drehen des Koordinatensystems vorstellen. Durch die Rotation sind die Items ‘näher’ an den Faktoren: Die Faktorladung zu einem Faktor wurde größer,

zum anderen Faktor hingegen geringer. Damit wurde die Ladung, also die Zuordnung der Items zu den Faktoren, insgesamt klarer, besser. Das wollen wir. Übrigens: Der Winkel der Achsen ist beim Rotieren gleich (rechteckig, orthogonal) geblieben. Daher spricht man von einer rechteckigen oder orthogonalen Rotation. Man kann auch die Achsen unterschiedlich rotieren, so dass sie nicht mehr rechteckig sind. Das könnte die Ladung noch klarer machen, führt aber dazu, dass die Faktoren dann korreliert sind. Korrelierte Faktoren sind oft nicht wünschenswert, weil ähnlich.

14.9.3 Vertiefung: Heatmap mit Ladungen

In der obigen Ausgabe werden die Item-to-Faktor-Ladungen angezeigt. Im zurückgegebenen Objekt `Werte.fa` sind diese als `$loadings` vorhanden. Wir können die Item-Faktor-Beziehungen mit einer Heatmap von `$loadings` visualisieren aus dem Paket `gplots`⁶, s. Abb. 14.6:

```
heatmap.2(Werte.fa$loadings,
           dendrogram = "both",
           labRow = NULL,
           labCol = NULL,
           cexRow=1,
           cexCol=1,
           margins = c(7,7),
           trace = "none",
           #lmat = rbind(c(0,0),c(0,1)),
           lhei = c(1,4),
           keysize=0.75,
           key.par = list(cex=0.5)
           )
```

Die Heatmap stellt ähnliche Objekte - hier: Variablen, die hoch auf einer Hauptkomponenten laden - räumlich nahe (nebeneinander) dar. Im Ergebnis zeigt die Heatmap eine deutliche Trennung der Items in 5 Faktoren, die interpretierbar sind als *Anerkennung*, *Genuss*, *Sicherheit*, *Bewusstsein* und *Konformismus*.

14.9.4 Berechnung der Faktor-Scores

Zusätzlich zur Schätzung der Faktorstruktur kann die EFA auch die latenten Faktorwerte für jede Beobachtung schätzen. Die gängige Extraktionsmethode ist die Bartlett-Methode, worauf wir hier nicht weiter eingehen. Kurz gesagt: Jeder Fall (jede Zeile im Datensatz, jede Person) bekommt einen Wert pro Komponente bzw. Faktor, man spricht von Faktor-Scores oder Faktorwerten der Beobachtungen.

⁶bereits automatisch geladen

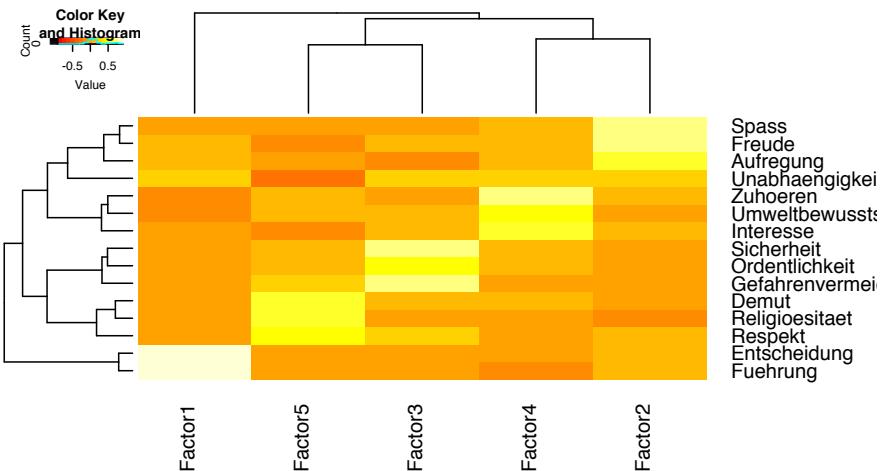


Abbildung 14.6: Heatmap einer EFA

```

Werte.ob <- factanal(Werte, factors = 5, scores = "Bartlett")
Werte.skaliertores <- data.frame(Werte.ob$scores)
names(Werte.skaliertores) <- c("Anerkennung", "Genuss", "Sicherheit", "Bewusstsein", "Konformismus")
head(Werte.skaliertores)

#>   Anerkennung Genuss Sicherheit Bewusstsein Konformismus
#> 1    1.380  0.985    0.563     0.173    -0.106
#> 2   -1.404 -0.721    1.597     1.166     0.695
#> 3    1.532 -0.657    1.672    -2.003     0.239
#> 4   -0.579  2.344   -1.056    -1.120    -0.118
#> 5    0.234 -1.652    1.189    -1.701     0.437
#> 6   -0.130  0.111   -1.053     0.791    -1.003

```

Wir haben nun anstatt der 15 Variablen 5 Faktoren mit Scores. Die Dimensionen wurden um ein Drittel reduziert.

14.10 Interne Konsistenz der Skalen

Das einfachste Maß für die *interne Konsistenz* ist die *Split-Half-Reliabilität*. Die Items werden in zwei Hälften unterteilt und die resultierenden Scores sollten in ihren Kenngrößen ähnlich sein. Hohe Korrelationen zwischen den Hälften deuten auf eine hohe interne Konsistenz hin. Das Problem ist, dass die Ergebnisse davon abhängen, wie die Items aufgeteilt werden. Ein üblicher Ansatz zur Lösung dieses Problems besteht darin, den Koeffizienten *Alpha* (*Cronbachs Alpha*) zu verwenden.

Der Koeffizient *Alpha* ist der Mittelwert aller möglichen Split-Half-Koeffizienten, die sich aus verschiedenen Arten der Aufteilung der Items ergeben. Dieser Koeffizient variiert von 0 bis 1. Inhaltlich ist Alpha eine Art mittlere Korrelation, die sich ergibt wenn man alle Items

(paarweise) miteinander korreliert: I1-I2, I1-I3, ...

Zufriedenstellende Reliabilität wird bei einem Alpha-Wert von 0.7 erreicht. Werte unter 0.5 gelten als nicht akzeptabel, Werte ab 0.8 als gut.

Wir bewerten nun die interne Konsistenz der Items Beispielhaft für das Konstrukt **Sicherheit** und nehmen zur Demonstration das Item **Unabhängigkeit** mit in die Analyse auf.

Werte %>%

```
 dplyr::select(Unabhaengigkeit, Zuhören, Umweltbewusstsein, Interesse) -> df
psych::alpha(df, check.keys = TRUE)
```

Bei dem Konstrukt **Sicherheit** können wir durch Elimination von **Unabhängigkeit** das Cronbachs Alpha von 0,63 auf einen fast akzeptablen Wert von 0,69 erhöhen.

Das Argument `check.keys=TRUE` gibt uns eine Warnung aus, sollte die Ladung eines oder mehrerer Items negativ sein. Dies ist hier nicht der Fall, somit müssen auch keine Items recodiert werden.

14.11 Aufgaben⁷



Richtig oder Falsch!?

1. Addiert man Antwortpunkte einer Reihe von Items zu Aggression, so hat (sicher) man Aggression gemessen.
2. Die Hauptkomponentenanalyse ist eine Methode zur Verringerung der Anzahl der Fälle eines Datensatzes.
3. Hauptkomponenten sind stets orthogonal zueinander (in einem Datensatz).
4. Ein Screeplot ist ein Diagramm, welches die Eigenwerte darstellt.
5. Längere Eigenvektoren sind durch größere Eigenwerte gekennzeichnet.
6. Bei einer rechtwinkligen Rotation bleiben die Faktoren rechtwinklig.
7. Die interne Konsistenz einer Skala ist ein Maß dafür, wie stark die Items miteinander korrelieren.

⁷F, F, R, R, R, R, R

Tabelle 14.1: Befehle des Kapitels 'Dimensionsreduktion'

Paket::Funktion	"Beschreibung"
cor	"Berechnet eine Korrelationsmatrix."
read.csv2	"Liest eine 'deutsche' CSV-Datei ein."
glimpse	"Wirft einen Blick (to glimpse) in den Datensatz."
scale	"führt eine z-Transformation durch"
corrplot::corrplot	"Plottet einen Korrelationsplot."
na.omit	"Schließt Zeilen mit fehlenden Werten von Datensatz aus."
pr.comp	"Berechnet Hauptkomponentenanalyse."
eigen	"Berechnet Eigenwerte."
psych::VSS.scree	"Plottet einen Screeplot."
biplot	"Plottet einen Biplot."
psych::principal	"Berechnet die Statistiken für eine Hauptkomponentenanalyse"
psych::fa.diagram	"Plottet ein Pfaddiagramm für eine Faktorenanalyse"
nFactors::nscreene	"Gibt verschiedenen Vorschläge für die Anzahl der 'richtigen' Faktoren"
factanal	"Berechnet eine Faktorenanalyse"
gplots::heatmap.2	"Plottet ein Heatmap"
factanal	"Berechnet Faktor-Scores"
psych::alpha	"Berechnet Cronbachs Alpha und weitere Statistiken"

14.12 Befehlsübersicht

Tabelle 14.1 fasst die R-Funktionen dieses Kapitels zusammen.

Kapitel 15

Vertiefung: Grundlagen des Textmining



Lernziele:

- Sie kennen zentrale Ziele und Begriffe des Textminings.
- Sie wissen, was ein ‘tidy text dataframe’ ist.
- Sie können Worthäufigkeiten auszählen.
- Sie können Worthäufigkeiten anhand einer Wordcloud visualisieren.

In diesem Kapitel benötigte R-Pakete:

```
library(tidyverse) # Datenjudo
library(stringr) # Textverarbeitung
library(tidytext) # Textmining
library(pdftools) # PDF einlesen
library(downloader) # Daten herunterladen
library(lsa) # Stopwörter
```

```
library(SnowballC) # Wörter trunkieren
library(wordcloud) # Wordcloud anzeigen
```

Ein großer Teil der zur Verfügung stehenden Daten liegt nicht als braves Zahlenmaterial vor, sondern in “unstrukturierter” Form, z.B. in Form von Texten. Im Gegensatz zur Analyse von numerischen Daten ist die Analyse von Texten weniger verbreitet bisher. In Anbetracht der Menge und der Informationsreichhaltigkeit von Text erscheint die Analyse von Text als vielversprechend.

In gewisser Weise ist das Textmining ein alternativer zu klassischen qualitativen Verfahren der Sozialforschung. Geht es in der qualitativen Sozialforschung primär um das Verstehen eines Textes, so kann man für das Textmining ähnliche Ziele formulieren. Allerdings: Das Textmining ist wesentlich schwächer und beschränkter in der Tiefe des Verstehens. Der Computer ist einfach noch (?) wesentlich *dümmer* als ein Mensch, zumindest in dieser Hinsicht. Allerdings ist er auch wesentlich *schneller* als ein Mensch, was das Lesen betrifft. Daher bietet sich das Textmining für das Lesen großer Textmengen an, in denen eine geringe Informationsdichte vermutet wird. Sozusagen maschinelles Sieben im großen Stil. Da fällt viel durch die Maschen, aber es werden Tonnen von Sand bewegt.

In der Regel wird das Textmining als *gemischte* Methode verwendet: sowohl qualitative als auch quantitative Aspekte spielen eine Rolle. Damit vermittelt das Textmining auf konstruktive Art und Weise zwischen den manchmal antagonistischen Schulen der qualitativ-idiographischen und der quantitativ-nomothetischen Sichtweise auf die Welt. Man könnte es auch als qualitative Forschung mit moderner Technik bezeichnen - mit den skizzierten Einschränkungen wohlgernekt.

15.1 Zentrale Begriffe

Die computergestützte Analyse von Texten speiste (und speist) sich reichhaltig aus Quellen der Linguistik; entsprechende Fachtermini finden Verwendung:

- Ein *Corpus* bezeichnet die Menge der zu analysierenden Dokumente; das könnten z.B. alle Reden der Bundeskanzlerin Angela Merkel sein oder alle Tweets von “@realDonaldTrump”.
- Ein *Token* (*Term*) ist ein elementarer Baustein eines Texts, die kleinste Analyseeinheit, häufig ein Wort.
- Unter *tidy text* versteht man einen Dataframe, in dem pro Zeile nur *ein* Token (z.B. Wort) steht (Silge und Robinson 2016). Synonym könnte man von einem “langen” Dataframe sprechen, so wie wir in Kapitel 3 kennen gelernt haben.

15.2 Grundlegende Analyse

15.2.1 Tidy Text Dataframes

Wozu ist es nützlich, einen Text-Dataframe in einen langen Dataframe umzuwandeln? Der Grund ist, dass immer wenn nur ein Wort (allgemeiner: Term) pro Zelle steht, dann können wir die Spalte einfach auszählen. Wir können z.B. `count` nutzen, um zu zählen, wie häufig ein Wort vorkommt. Sprich: Sobald wir einen langen (Text-)Dataframe haben, können wir unsere bekannte Methoden einsetzen.

Basteln wir uns einen *tidy text* Dataframe. Wir gehen dabei von einem Vektor mit mehreren Text-Elementen aus, das ist ein realistischer Startpunkt. Unser Text-Vektor¹ besteht aus 4 Elementen.

```
text <- c("Wir haben die Frauen zu Bett gebracht",
        "als die Männer in Frankreich standen.",
        "Wir hatten uns das viel schöner gedacht.",
        "Wir waren nur Konfirmanden.")
```

Als nächstes machen wir daraus einen Dataframe.

```
text_df <- data_frame(Zeile = 1:4,
                      text = text)
```

Diesen Mini-Datensatz finden Sie auch im Ordner `data` als `Brecht.csv`; nach bekannter Manier können Sie die CSV-Datei importieren:

```
text_df <- read.csv("data/Brecht.csv")
```

Zeile	text
1	Wir haben die Frauen zu Bett gebracht,
2	als die Männer in Frankreich standen.
3	Wir hatten uns das viel schöner gedacht.
4	Wir waren nur Konfirmanden.

Übrigens, falls Sie eine beliebige Textdatei einlesen möchten, können Sie das so tun:

```
text <- read_lines("data/Brecht.txt")
```

Der Befehl `read_lines` (aus `readr`²) liest Zeilen (Zeile für Zeile) aus einer Textdatei.

¹Nach dem Gedicht “Jahrgang 1899” von Erich Kästner

²Teil der Tidyverse-Familie

Breiter Dataframe

Zeile	text
1	Wir haben die Frauen zu Bett gebracht,
2	als die Männer in Frankreich standen.
...	...

Langer Dataframe

Zeile	Wort
1	Wir
1	haben
1	die
1	Frauen
1	zu
...	...

tidy text Dataframe

Abbildung 15.1: Illustration eines Tidy Text Dataframe

Dann “dehnen” wir den Dataframe zu einem *tidy text* Dataframe (s. Abb. 15.1); das besorgt die Funktion `unnest_tokens`. ‘unnest’ heißt dabei so viel wie ‘Entschachteln’, also von breit auf lang dehnen. Mit ‘tokens’ sind hier einfach die Wörter gemeint (es könnten aber auch andere Analyseeinheiten sein, Sätze zum Beispiel).

```
text_df %>%
  unnest_tokens(output = wort, input = text) -> tidytext_df

tidytext_df %>% head
#> # A tibble: 6 x 2
#>   Zeile     wort
#>   <int> <chr>
#> 1     1     wir
#> 2     1     haben
#> 3     1     die
#> 4     1     frauen
#> 5     1     zu
#> 6     1    bett
```

Der Parameter `output` sagt, wie neue ‘saubere’ (lange) Spalte heißen soll; `input` sagt der Funktion, welche Spalte sie als ihr Futter (Input) betrachten soll (welche Spalte in tidy text umgewandelt werden soll).

In einem ‘tidy text Dataframe’ steht in jeder Zeile ein Wort (token) und die Häufigkeit des Worts im Dokument.

Überprüfen Sie, ob das stimmt: Betrachten Sie den Dataframe `tidytext_df`.

Das `unnest_tokens` kann übersetzt werden als “entschachtele” oder “dehne” die Tokens - so dass in *jeder Zeile* nur noch *ein Wort* (genauer: Token) steht. Die Syntax ist `unnest_tokens(Ausgabespalte, Eingabespalte)`. Nebenbei werden übrigens alle Buchstaben auf Kleinschreibung getrimmt.

Als nächstes filtern wir die Satzzeichen heraus, da die Wörter für die Analyse wichtiger (oder zumindest einfacher) sind.

```
text_df %>%
  unnest_tokens(wort, text) %>%
  filter(str_detect(wort, "[a-z]"))
#> # A tibble: 24 x 2
#>   Zeile     wort
#>   <int>     <chr>
#> 1 1         wir
#> 2 1         haben
#> 3 1         die
#> 4 1        frauen
#> 5 1         zu
#> 6 1        bett
#> 7 1         gebracht
#> 8 2         als
#> 9 2         die
#> 10 2        männer
#> # ... with 14 more rows
```

Das "[a-z]" steht für "alle Buchstaben von a-z". In Pseudo-Code heißt dieser Abschnitt:



Nehme den Datensatz "text_df" UND DANN
dehne die einzelnen Elemente der Spalte "text", so dass jedes Element seine eigene Spalte bekommt.
Ach ja: Diese "gedehnte" Spalte soll "Wort" heißen (weil nur einzelne Wörter drinnen stehen).
Ach ja 2: Dieses "dehnen" wandelt automatisch Groß- in Kleinbuchstaben um. UND DANN
filtere die Spalte "wort", so dass nur noch Kleinbuchstaben übrig bleiben. FERTIG.

15.2.2 Text-Daten einlesen

Nun lesen wir Text-Daten ein; das können beliebige Daten sein³. Eine gewisse Reichhaltigkeit ist von Vorteil. Nehmen wir das Parteiprogramm der Partei AfD⁴. Vor dem Hintergrund des Erstarkens des Populismus weltweit und der großen Gefahr, die davon ausgeht - man blicke auf die Geschichte Europas in der ersten Hälfte des 20. Jahrhunderts - verdient erfordert der politische Prozess und speziell Neuentwicklungen darin unsere besondere Beachtung.

³Ggf. benötigen Sie Administrator-Rechte, um Dateien auf Ihre Festplatte zu speichern.

⁴

```
afd_pfad <- "data/afd_programm.pdf"

afd_raw <- pdf_text(afd_pfad)
```

Mit `head(afd_raw)` können Sie sich den Beginn dieses Textvektor anzeigen lassen.

Für uns ist `pdf_text` sehr praktisch, da diese Funktion Text aus einer beliebigen PDF-Datei in einen Text-Vektor einliest. `head(afd_raw, 1)` liest das 1. Element (und nur das erste) aus `afd_raw` aus.

Der Vektor `afd_raw` hat 96 Elemente (entsprechend der Seitenzahl des Dokuments); zählen wir die Gesamtzahl an Wörtern. Dazu wandeln wir den Vektor in einen tidy text Dataframe um. Auch die Stopwörter entfernen wir wieder wie gehabt.

```
afd_df <- data_frame(Zeile = 1:96,
                      afd_raw)
afd_df %>%
  unnest_tokens(output = token, input = afd_raw) %>%
  dplyr::filter(str_detect(token, "[a-z]")) -> afd_df

count(afd_df)
#> # A tibble: 1 x 1
#>       n
#>   <int>
#> 1 26396
```

Eine substanziale Menge von Text. Was wohl die häufigsten Wörter sind?

15.2.3 Worthäufigkeiten auszählen

```
afd_df %>%
  na.omit() %>% # fehlende Werte löschen
  count(token, sort = TRUE)
#> # A tibble: 7,087 x 2
#>   token     n
#>   <chr> <int>
#> 1 die    1151
#> 2 und    1147
#> 3 der    870
#> 4 zu     435
```

```
#> 5 für 392
#> 6 in 392
#> 7 den 271
#> 8 von 257
#> 9 ist 251
#> 10 das 225
#> # ... with 7,077 more rows
```

Die häufigsten Wörter sind inhaltsleere Partikel, Präpositionen, Artikel... Solche sogenannten “Stopwörter” sollten wir besser herausfischen, um zu den inhaltlich tragenden Wörtern zu kommen. Praktischerweise gibt es frei verfügbare Listen von Stopwörtern, z.B. im Paket `lsa`.

```
data(stopwords_de, package = "lsa")

stopwords_de <- data_frame(word = stopwords_de)

stopwords_de <- stopwords_de %>%
  rename(token = word)
# Für das Joinen werden gleiche Spaltennamen benötigt

afd_df %>%
  anti_join(stopwords_de) -> afd_df
```

Unser Datensatz hat jetzt viel weniger Zeilen; wir haben also durch `anti_join` Zeilen gelöscht (herausgefiltert). Das ist die Funktion von `anti_join`: Die Zeilen, die in beiden Dataframes vorkommen, werden herausgefiltert. Es verbleiben also nicht “Nicht-Stopwörter” in unserem Dataframe. Damit wird es schon interessanter, welche Wörter häufig sind.

```
afd_df %>%
  count(token, sort = TRUE) -> afd_count
```

Ganz interessant; aber es gibt mehrere Varianten des Themas “deutsch”. Es ist wohl sinnvoller, diese auf den gemeinsamen Wortstamm zurückzuführen und diesen nur einmal zu zählen. Dieses Verfahren nennt man “stemming” oder “trunkieren”.

```
afd_df %>%
  mutate(token_stem = wordStem(.token, language = "german")) %>%
  count(token_stem, sort = TRUE) -> afd_count

afd_count %>%
  top_n(10) %>%
  knitr::kable(caption = "Die häufigsten Wörter im AfD-Parteiprogramm mit 'stemming'")
```

Tabelle 15.1: Die häufigsten Wörter im AfD-Parteiprogramm

token	n
deutschland	190
afd	171
programm	80
wollen	67
bürger	57
euro	55
dafür	53
eu	53
deutsche	47
deutschen	47

Tabelle 15.2: Die häufigsten Wörter im AfD-Parteiprogramm mit 'stemming'

token_stem	n
deutschland	219
afd	171
deutsch	119
polit	88
staat	85
programm	81
europa	80
woll	67
burg	66
soll	63

Das ist schon informativer. Dem Befehl `SnowballC::wordStem` füttert man einen Vektor an Wörtern ein und gibt die Sprache an (Default ist Englisch). Denken Sie daran, dass `.` bei `dplyr` nur den Datensatz meint, wie er im letzten Schritt definiert war. Mit `.$token` wählen wir also die Variable `token` aus `afd_raw` aus.

15.2.4 Visualisierung

Zum Abschluss noch eine Visualisierung mit einer “Wordcloud” dazu.

```
myword <- na.omit(afd_count$token_stem)
head(myword, 10)
myfreq <- na.omit(afd_count$n)
head(myfreq, 10)
```

```
wordcloud(words = afd_count$token_stem,
          freq = afd_count$n,
          max.words = 100,
          scale = c(2,.5),
          colors=brewer.pal(6, "Dark2"))
```



Man kann die Anzahl der Wörter, Farben und einige weitere Formatierungen der Wortwolke beeinflussen⁵.

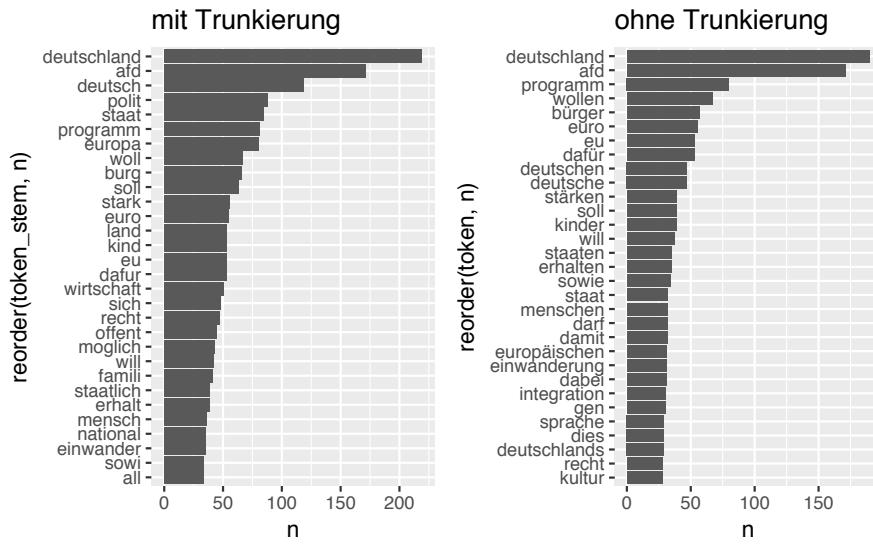
Weniger verspielt ist eine schlichte visualisierte Häufigkeitsauszählung dieser Art, z.B. mit Balkendiagrammen (gedreht).

```
afd_count %>%
  top_n(30) %>%
  ggplot() +
  aes(x = reorder(token_stem, n), y = n) +
  geom_col() +
  labs(title = "mit Trunkierung") +
  coord_flip() -> p1

afd_df %>%
  count(token, sort = TRUE) %>%
  top_n(30) %>%
  ggplot() +
  aes(x = reorder(token, n), y = n) +
  geom_col() +
  labs(title = "ohne Trunkierung") +
  coord_flip() -> p2
```

⁵<https://cran.r-project.org/web/packages/wordcloud/index.html>

```
library(gridExtra)
grid.arrange(p1, p2, ncol = 2)
```



Die beiden Diagramme vergleichen die trunkierten Wörter mit den nicht trunkierten Wörtern. Mit `reorder` ordnen wir die Spalte `token` nach der Spalte `n`. `coord_flip` dreht die Abbildung um 90°, d.h. die Achsen sind vertauscht. `grid.arrange` packt beide Plots in eine Abbildung, welche 2 Spalten (`ncol`) hat.

15.3 Aufgaben⁶



Richtig oder Falsch!?

1. Unter einem Token versteht man die größte Analyseeinheit in einem Text.
2. In einem tidytext Dataframe steht jedes Wort in einer (eigenen) Zeile.
3. Eine hinreichende Bedingung für einen tidytext Dataframe ist es, dass in jeder Zeile ein Wort steht (beziehen Sie sich auf den tidytext Dataframe wie in diesem Kapitel erörtert).
4. Gibt es 'Stop-Wörter' in einem Dataframe, dessen Text analysiert wird, so kommt es - per definitionem - zu einem Stop.
5. Mit dem Befehl `unnest_tokens` kann man einen tidytext Dataframe erstellen.
6. Balkendiagramme sind sinnvolle und auch häufige Diagrammtypen, um die häufigsten Wörter (oder auch Tokens) in einem Corpus darzustellen.
7. In einem 'tidy text Dataframe' steht in jeder Zeile ein Wort (`token`) aber nicht die Häufigkeit des Worts im Dokument.

⁶F, R, F, F, R, R, F, F

Tabelle 15.3: Befehle des Kapitels 'Textmining'

Paket..Befehl	Beschreibung
tidytext::unnest_tokens	Jedes Token (Wort) einer Spalte bekommt eine eigene Zeile in einem Dataframe
stringr::str_detect	Sucht nach einem String (Text)
downloader:: download	lädt eine Datei aus dem Internet herunter
dplyr::rename	Benennt Spalten um
anti_join	Führt Dataframes zusammen, so dass nicht matchende Einträge übernommen werden
wordcloud::wordcloud	Erstellt eine Wordcloud
ggplot2::labs	Fügt Titel oder andere Hinweise einem ggplot2-Objekt hinzu
ggplot2::coord_flip	Dreht die Achsen um 90 Grad

8. Unter 'Stemming' versteht man (bei der Textanalyse), die Etymologie eines Wort (Herkunft) zu erkunden.

15.4 Befehlsübersicht

Tabelle 15.3 fasst die R-Funktionen dieses Kapitels zusammen.

15.5 Verweise

- Das Buch *Tidy Text Mining* (Julia und David 2017) ist eine hervorragende Quelle vertieftem Wissens zum Textmining mit R.

Anhang A

Probeklausur



Aussagen sind entweder als “richtig” oder als “falsch” zu beantworten. Offene Fragen verlangen einen “Text” als Antwort.

1. Bei `install.packages` spielt der Parameter `dependencies = TRUE` in der Praxis keine Rolle.
2. Dateien mit der Endung `.R` sind keine Textdateien.
3. Der Befehl `read.csv` kann auch Dateien einlesen, die nicht lokal, sondern auf einem Server im Internet gespeichert sind.
4. Fehlende Werte werden in R durch `NA` kodiert.
5. Um Variablen einen Wert zuzuweisen, kann man in R den Zuweisungspfeil `<-` verwenden.
6. Die deutsche Version von R verwendet im Standard das Komma als Dezimaltrennzeichen.
7. Statistisches Modellieren verwendet die Abduktion als zentrale Denkfigur.
8. Eine Abduktion führt zu sicheren Schlüssen.
9. Das CSV-Format ist identisch zum Excel-Format, was sich auch darin zeigt, dass Excel CSV-Dateien oft problemlos öffnet.

10. Das Arbeitsverzeichnis (engl. *working directory*) ist der Ort, in dem R eine Datei, die Sie aufrufen, vermutet - sofern kein anderer Pfad angegeben ist.
11. In einer Tabelle in Normalform steht in jeder Zeile eine Variable und in jeder Spalte eine Beobachtung.
12. Die Funktion `filter` filtert Spalten aus einer Tabelle.
13. Die Funktion `select` lässt Spalten sowohl anhand ihres Namens als auch ihrer Nummer (Position in der Tabelle) auswählen.
14. Die Funktion `group_by` gruppert eine Tabelle anhand der Werte einer diskreten Variablen.
15. Die Funktion `group_by` akzeptiert nur Faktorvariablen als Gruppierungsvariablen.
16. Die Funktion `summarise` darf nur für Funktionen verwendet werden, welche genau *einen* Wert zurückliefern.
17. Was sind drei häufige Operationen der Datenaufbereitung?
18. Um Korrelationen mit R zu berechnen, kann man die Funktion `corr::correlate` verwenden.
19. `corr::correlate` liefert stets einen Dataframe zurück.
20. Tibbles sind eine spezielle Art von Dataframes.
21. Was zeigt uns “Anscombes Quartett”?
22. `ggplot` unterscheidet drei Bestandteile eines Diagramms: Daten, Geome und Transformationen.
23. Um eine kontinuierliche Variable zu plotten, wird häufig ein Histogramm verwendet.
24. Das Geom `tile` zeigt drei Variablen.
25. Geleitetes Modellieren kann unterteilt werden in prädiktives und explikatives Modellieren.
26. Der Befehl `scale` verschiebt den Mittelwert einer Verteilung auf 0 und skaliert die sd auf 1.
27. Mit “binnen” im Sinne der Datenanalyse ist gemeint, eine kategoriale Variable in eine kontinuierliche zu überführen.
28. Die Gleichung $y = ax + b$ lässt sich in R darstellen als $y \sim ax + b$.
29. R^2 , auch Bestimmtheitsmaß oder Determinationskoeffizient genannt, gibt die Vorhersagegüte im Verhältnis zu einem “Nullmodell” an.
30. Bei der logistischen Regression gilt: Bei $\beta_0 > 0$ ist die Wahrscheinlichkeit *kleiner* als 50% gibt, dass das modellierte Ereignis eintritt, wenn alle anderen Prädiktoren Null sind.

31. Die logistische Regression sollte *nicht* verwendet werden, wenn die abhängige Variable dichotom ist.
32. Die logistische Regression stellt den Zusammenhang zwischen Prädiktor und Kriterium nicht mit einer Geraden, sondern mit einer “s-förmigen” Kurve dar.
33. Bevor die Koeffizienten der logistischen Regression als Odds Ration interpretiert werden können, müssen sie “deologarithmiert” werden.
34. Unter “deologarithmieren” versteht man, die Umkehrfunktion der e-Funktion auf eine Gleichung anzuwenden.
35. Wendet man die “normale” Regression an, um eine dichotome Variable als Kriterium zu modellieren, so kann man Wahrscheinlichkeiten größer als 1 und kleiner als 0 bekommen.
36. Eine typische Idee der Clusteranalyse lautet, die Varianz innerhalb der Cluster jeweils zu maximieren.
37. Bei einer k-means-Clusteranalyse darf man nicht die Anzahl der Cluster vorab festlegen; vielmehr ermittelt der Algorithmus die richtige Anzahl der Cluster.
38. Für die Wahl der “richtigen” Anzahl der Cluster kann das “Ellbogen-Kriterium” als Entscheidungsgrundlage herangezogen werden.
39. Ein “Screeplot” stellt die Varianz innerhalb der Cluster als Funktion der Anzahl der Cluster dar (im Rahmen der Clusteranalyse).
40. Die euklidische Distanz zwischen zwei Objekten in der Ebene lässt sich mit dem Satz des Pythagoras berechnen.

Anhang B

Lösungen

1. Falsch
2. Falsch
3. Richtig
4. Richtig
5. Richtig
6. Falsch
7. Richtig
8. Falsch
9. Falsch
10. Richtig
11. Falsch
12. Falsch
13. Falsch
14. Richtig
15. Richtig
16. Falsch
17. Richtig
18. Auf fehlende Werte prüfen, Fälle mit fehlenden Werte löschen, Fehlende Werte ggf. ersetzen, Nach Fehlern suche, Ausreiser identifizieren, Hochkorrelierte Variablen finden, z-Standardisieren, Quasi-Konstante finden, Auf Normalverteilung prüfen, Werte umkodieren und “binnen”.
19. Richtig
20. Richtig
21. Richtig
22. Es geht hier um vier Datensätze mit zwei Variablen (Spalten; X und Y). Offenbar sind die Datensätze praktisch identisch: Alle X haben den gleichen Mittelwert und die gleiche Varianz; dasselbe gilt für die Y. Die Korrelation zwischen X und Y ist in allen vier Datensätzen gleich. Allerdings erzählt eine Visualisierung der vier Datensätze eine ganz andere Geschichte.
23. Falsch

- 24. Richtig
- 25. Richtig
- 26. Falsch
- 27. Richtig
- 28. Falsch
- 29. Richtig
- 30. Richtig
- 31. Falsch
- 32. Falsch
- 33. Richtig
- 34. Richtig
- 35. Falsch. Richtig wäre: Die Umkehrfunktion des Logarithmus, also die e-Funktion, auf eine Gleichung anzuwenden.
- 36. Richtig
- 37. Falsch
- 38. Falsch. Richtig wäre: Man gibt die Anzahl der Cluster vor. Dann vergleicht man die Varianz within der verschiedenen Lösungen.
- 39. Richtig
- 40. Richtig
- 41. Richtig

Anhang C

Hinweise

Anhang D

Icons

R spricht zu Ihnen; sie versucht es jedenfalls, mit einigen Items (Icon-Pond 2016).

R-Pseudo-Syntax: R ist (momentan) die führende Umgebung für Datenanalyse. Entsprechend zentral ist R in diesem Kurs. Zugegebenermaßen braucht es etwas Zeit, bis man ein paar Brocken “Errisch” spricht. Um den Einstieg zu erleichtern, ist Errisch auf Deutsch übersetzt an einigen Stellen, wo mir dies besonders hilfreich erschien. Diese Stellen sind mit diesem

Symbol  gekennzeichnet (für R-Pseudo-Syntax).

Achtung, Falle: Schwierige oder fehlerträchtige Stellen sind mit diesem Symbol  markiert.

Übungsaufgaben: Das Skript beinhaltet in jedem Kapitel Übungsaufgaben oder/und Testfragen.

Auf diese wird mit diesem Icon  verwiesen oder die Übungen sind in einem Abschnitt mit einsichtigem Titel zu finden.

Anhang E

Voraussetzungen

Dieses Skript hat einige *Voraussetzungen*, was das Vorwissen der Leser angeht; folgende Themengebiete werden vorausgesetzt:

- Deskriptive Statistik
- Grundlagen der Inferenzstatistik
- Grundlagen der Regressionsanalyse
- Skalenniveaus
- Grundlagen von R

Anhang F

Zitationen

Kunstwerke (Bilder) sind genau wie Standard-Literatur im Text zitiert. Alle Werke (auch Daten und Software) finden sich im Literaturverzeichnis.

Anhang G

Lizenz

Dieses Skript ist publiziert unter CC-BY-NC-SA 3.0 DE¹.



¹<https://creativecommons.org/licenses/by-nc-sa/3.0/de/>

Anhang H

Autoren

Sebastian Sauer schrieb den Hauptteil dieses Skripts. *Oliver Gansser* schrieb das Kapitel zur Dimensionsreduktion. *Karsten Lübke* schrieb den Großteil des Kapitels zur Regression und zur Clusteranalyse sowie Teile des Kapitels ‘Rahmen’. *Matthias Gehrke* schrieb den Großteil des Kapitels zur logistischen Regression.

Anhang I

Danke

Norman Markgraf hat umfangreich Fehler gejagt und Verbesserungen ~~angemahnt~~ vorgenommen. Der Austausch mit den ifes-Mitgliedern hielt die Flamme am Köcheln. Eine Reihe weiterer Kollegen standen mit Rat und Tat zur Seite. Die Hochschulleitung sowie das Dekanat für Wirtschaftspsychologie hat dieses Projekt unterstützt. Die Abteilung Medienentwicklung der FOM hat bei Fragen rund um die Veröffentlichung geholfen. Last but not least: Viele Studierenden wiesen auf Inkonsistenzen, Fehler und Unklarheiten hin. Ihnen allen: Vielen Dank!

Anhang J

Zitation dieses Skripts

Bitte zitieren Sie das Skript so:

Sauer, S. (2017). *Praxis der Datenanalyse*. Skript zum Modul im MSc.-Studiengang “Wirtschaftspsychologie & Consulting” an der FOM. FOM Nürnberg. DOI: 10.5281/zenodo.580649.

Mehr Infos zum DOI hier: <https://zenodo.org/badge/latestdoi/81811975>

Ein Bib-File um dieses Skript zu zitieren finden Sie hier: https://raw.githubusercontent.com/sebastiansauer/Praxis_der_Datenanalyse/master/Praxis_der_Datenanalyse.bib.

Anhang K

Kontakt

Wenn Sie einen Fehler oder Verbesserungshinweise berichten möchten, können Sie unter https://github.com/sebastiansauer/Praxis_der_Datenanalyse/issues einen “Issue” einreichen (Button “New Issue”). Alternativ können Sie Sebastian Sauer und die anderen Autoren über den Online Campus der FOM kontaktieren (eine Nachricht schreiben). Sebastian Sauer können Sie via Twitter folgen (https://twitter.com/sauer_sebastian) oder seinen Blog lesen (<https://sebastiansauer.github.io>).

Anhang L

Technische Details

Dieses Skript wurde mit dem Paket `bookdown` (Xie 2015) erstellt, welches wiederum stark auf den Paketen `knitr` (Xie 2015) und `rmarkdown` (Allaire u. a. 2016a) beruht. Diese Pakete stellen verblüffende Funktionalität zur Verfügung als freie Software (frei wie in Bier und frei wie in Freiheit).

Informationen zu den verwendeten Paketen etc. (`sessionInfo()`) finden Sie hier: https://raw.githubusercontent.com/sebastiansauer/Praxis_der_Datenanalyse/master/includes/sessionInfo_PraDa.html.

Anhang M

Sonstiges

Aus Gründen der Lesbarkeit wird das männliche Generikum verwendet, welches Frauen und Männer in gleichen Maßen ansprechen soll.

Anhang N

Literaturverzeichnis

- Allaire, JJ, Joe Cheng, Yihui Xie, Jonathan McPherson, Winston Chang, Jeff Allen, Hadley Wickham, Aron Atkins, und Rob Hyndman. 2016a. *rmarkdown: Dynamic Documents for R*. <https://CRAN.R-project.org/package=rmarkdown>.
- . 2016b. *rmarkdown: Dynamic Documents for R*. <https://CRAN.R-project.org/package=rmarkdown>.
- Auguie, Baptiste. 2016. *gridExtra: Miscellaneous Functions for „Grid“ Graphics*. <https://CRAN.R-project.org/package=gridExtra>.
- Beaujean, A. Alexander. 2012. *BaylorEdPsych: R Package for Baylor University Educational Psychology Quantitative Courses*. <https://CRAN.R-project.org/package=BaylorEdPsych>.
- Benoit, Kenneth, und Paul Nulty. 2016. *quanteda: Quantitative Analysis of Textual Data*. <https://CRAN.R-project.org/package=quanteda>.
- Bouchet-Valat, Milan. 2014. *SnowballC: Snowball stemmers based on the C libstemmer UTF-8 library*. <https://CRAN.R-project.org/package=SnowballC>.
- Briggs, William M. 2008a. *Breaking the Law of Averages: Real-Life Probability and Statistics in Plain English*. Lulu.com.
- . 2008b. *Breaking the Law of Averages: Real-Life Probability and Statistics in Plain English*. Lulu.com. <https://www.amazon.com/Breaking-Law-Averages-Probability-Statistics/dp/0557019907?SubscriptionId=0JYN1NVW651KCA56C102&tag=techkie-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=0557019907>.
- . 2016. *Uncertainty: The Soul of Modeling, Probability & Statistics*. Springer.
- Bryant, PG, und MA Smith. 1995. „Practical Data Analysis: Case Studies in Business Statistics, Homewood, IL: Richard D“. Irwin Publishing.
- Chang, Winston. 2015. *downloader: Download Files over HTTP and HTTPS*. <https://CRAN.R-project.org/package=downloader>.
- Chapman, Chris, und Elea McDonnell Feit. 2015. *R for Marketing Research and Analytics*.

- New York City: Springer. doi:10.1007/978-3-319-14436-8¹.
- Cleveland, William S. 1993. *Visualizing Data*. Hobart Press.
- Clopper, Charles J, und Egon S Pearson. 1934. „The use of confidence or fiducial limits illustrated in the case of the binomial“. *Biometrika* 26 (4). JSTOR: 404–13.
- Cobb, George W. 2007. „The introductory statistics course: a Ptolemaic curriculum?“ *Technology Innovations in Statistics Education* 1 (1).
- Cohen, J. 1992. „A power primer“. *Psychological Bulletin* 112 (1): 155–59.
- Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Routledge. <http://dx.doi.org/10.4324/9780203771587>.
- Cortez, Paulo, António Cerdeira, Fernando Almeida, Telmo Matos, und José Reis. 2009. „Modeling wine preferences by data mining from physicochemical properties“. *Decision Support Systems* 47 (4). Elsevier: 547–53.
- de Vries, Andrie, und Brian D. Ripley. 2016. *ggdendro: Create Dendograms and Tree Diagrams Using 'ggplot2'*. <https://CRAN.R-project.org/package=ggdendro>.
- Diez, David M, Christopher D Barr, und Mine Cetinkaya-Rundel. 2014. *Introductory Statistics with Randomization and Simulation*. North Charleston, South Carolina: CreateSpace Independent Publishing Platform.
- Eid, Michael, Mario Gollwitzer, und Manfred Schmitt. 2010. *Statistik und Forschungsmethoden*. Göttingen: Hogrefe.
- Etz, Alexander, Quentin Frederik Gronau, Fabian Dablander, Peter Edelsbrunner, und Beth Baribault. 2016. „How to become a Bayesian in eight easy steps: An annotated reading list“. PsyArXiv.
- Fair, Ray C. 1978. „A theory of extramarital affairs“. *Journal of Political Economy* 86 (1). The University of Chicago Press: 45–61.
- Feinerer, Ingo, und Kurt Hornik. 2015. *tm: Text Mining Package*. <https://CRAN.R-project.org/package=tm>.
- Fellows, Ian. 2014. *wordcloud: Word Clouds*. <https://CRAN.R-project.org/package=wordcloud>.
- Fjalnes. 2014. „Orthogonale Faktorrotation“. [https://de.wikipedia.org/wiki/Rotationsverfahren_\(Statistik\)#/media/File:Orthogonale_faktorrotation.svg](https://de.wikipedia.org/wiki/Rotationsverfahren_(Statistik)#/media/File:Orthogonale_faktorrotation.svg).
- Fox, John, und Sanford Weisberg. 2016. *car: Companion to Applied Regression*. <https://CRAN.R-project.org/package=car>.
- Gansser, Oliver. 2017. „Data for Principal Component Analysis and Common Factor Analysis“. Open Science Framework. osf.io/zg89r.
- Gigerenzer, Gerd. 1980. *Messung und Modellbildung in der Psychologie (Uni-Taschenbucher)*.

¹<https://doi.org/10.1007/978-3-319-14436-8>

- Psychologie, Padagogik, Soziologie, Psychiatrie) (German Edition)*. E. Reinhardt.
- . 2004. „Mindless statistics“. *The Journal of Socio-Economics* 33 (5). Elsevier BV: 587–606. doi:10.1016/j.soec.2004.09.033².
- God. 2016. „I don't care about you. Please share this with friends.“ Twitter Tweet. *TheTweetOfGod*. <https://twitter.com/TheTweetOfGod/status/688035049187454976>.
- Gromlund, Garrett, und Hadley Wickham. 2014. „A cognitive interpretation of data analysis“. *International Statistical Review* 82 (2). Wiley Online Library: 184–204.
- Hahsler, Michael, Christian Buchta, Bettina Gruen, und Kurt Hornik. 2016. *arules: Mining Association Rules and Frequent Itemsets*. <https://CRAN.R-project.org/package=arules>.
- Hahsler, Michael, und Sudheer Chelluboina. 2016. *arulesViz: Visualizing Association Rules and Frequent Itemsets*. <https://CRAN.R-project.org/package=arulesViz>.
- Hamerling, Daniel S, und Amy Parker. 2005. „Beauty in the classroom: Instructors' pulchritude and putative pedagogical productivity“. *Economics of Education Review* 24 (4). Elsevier: 369–76.
- Hardin, Johanna, Roger Hoerl, Nicholas J Horton, Deborah Nolan, Ben Baumer, Olaf Hall-Holt, Paul Murrell, u. a. 2015. „Data science in statistics curricula: Preparing students to 'Think with Data'“. *The American Statistician* 69 (4). Taylor & Francis: 343–53.
- Hatzinger, Reinholt, Kurt Hornik, und Herbert Nagel. 2014. *R- Einfuehrung durch angewandte Statistik*. Pearson Studium.
- Head, Megan L., Luke Holman, Rob Lanfear, Andrew T. Kahn, und Michael D. Jennions. 2015. „The Extent and Consequences of P-Hacking in Science“. *PLOS Biology* 13 (3). Public Library of Science (PLoS): e1002106. doi:10.1371/journal.pbio.1002106³.
- Hendricks, Paul. 2015. *titanic: Titanic Passenger Survival Data Set*. <https://CRAN.R-project.org/package=titanic>.
- Hoekstra, Rink, Richard D Morey, Jeffrey N Rouder, und Eric-Jan Wagenmakers. 2014. „Robust misinterpretation of confidence intervals“. *Psychonomic bulletin & review* 21 (5). Springer: 1157–64.
- Hyndman, R.J., und G. Athanasopoulos. 2014. *Forecasting: principles and practice*: OTexts. <https://books.google.de/books?id=gDuRBAAQBAJ>.
- Icon-Pond. 2016. „Education. 35 Icons.“ Flaticon. <http://www.flaticon.com/authors/popcorns-arts>.
- Ingo Feinerer, Kurt Hornik, und David Meyer. 2008. „Text Mining Infrastructure in R“. *Journal of Statistical Software* 25 (5): 1–54. <http://www.jstatsoft.org/v25/i05/>.
- Jackson, Simon. 2016. *corrr: Correlations in R*. <https://CRAN.R-project.org/package=corrr>

²<https://doi.org/10.1016/j.soec.2004.09.033>

³<https://doi.org/10.1371/journal.pbio.1002106>

`corrr`.

James, Gareth, Daniela Witten, Trevor Hastie, und Rob Tibshirani. 2013a. *ISLR: Data for An Introduction to Statistical Learning with Applications in R*. <https://CRAN.R-project.org/package=ISLR>.

James, Gareth, Daniela Witten, Trevor Hastie, und Robert Tibshirani. 2013b. *An introduction to statistical learning*. Bd. 6. Springer.

———. 2013c. *An introduction to statistical learning*. Bd. 6. Springer.

Julia, PhD Silge, und PhD Robinson David. 2017. *Text Mining with R: A tidy approach*. O'Reilly Media.

Kerby, Dave S. 2014. „The Simple Difference Formula: An Approach to Teaching Nonparametric Correlation“. *Comprehensive Psychology* 3: 11.IT.3.1. doi:10.2466/11.IT.3.1⁴.

Kim, Albert Y, und Adriana Escobedo-Land. 2015. „OkCupid Data for Introductory Statistics and Data Science Courses“. *Journal of Statistics Education* 23 (2). Citeseer: n2.

Kim, Albert Y., und Adriana Escobedo-Land. 2016. *okcupiddata: OkCupid Profile Data for Introductory Statistics and Data Science Courses*. <https://CRAN.R-project.org/package=okcupiddata>.

Krämer, W. 2011. *Wie wir uns von falschen Theorien täuschen lassen*. Berlin University Press. <https://books.google.de/books?id=HWUKaAEACAAJ>.

Kruschke, John K. 2010. „Bayesian data analysis“. *Wiley Interdisciplinary Reviews: Cognitive Science* 1 (5). Burlington, MA: Academic Press: 658–76.

Kuhn, Max, und Kjell Johnson. 2013. *Applied predictive modeling*. Bd. 26. Springer.

Ligges, Uwe, Martin Maechler, und Sarah Schnackenberg. 2017. *scatterplot3d: 3D Scatter Plot*. <https://CRAN.R-project.org/package=scatterplot3d>.

Lübke, Karsten, und Martin Vogt. 2014. *Angewandte Wirtschaftsstatistik: Daten und Zufall*. Berlin: Springer.

M7. 2004. „Savinelli's Italian smoking pipe“. https://commons.wikimedia.org/wiki/File:Pipa_savinelli.jpg.

Matejka, Justin, und George Fitzmaurice. 2017. „Same stats, different graphs: Generating datasets with varied appearance and identical statistics through simulated annealing“. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 1290–4. ACM.

Micceri, Theodore. 1989. „The unicorn, the normal curve, and other improbable creatures.“ *Psychological Bulletin* 105 (1): 156–66. doi:10.1037/0033-2909.105.1.156⁵.

Michell, Joel. 2000. „Normal Science, Pathological Science and Psychometrics“. *Theory &*

⁴<https://doi.org/10.2466/11.IT.3.1>

⁵<https://doi.org/10.1037/0033-2909.105.1.156>

Psychology 10 (5): 639–67.

Milborrow, Stephen. 2017. *rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'*. <https://CRAN.R-project.org/package=rpart.plot>.

Moore, David S. 1990. „Uncertainty“. *On the shoulders of giants: New approaches to numeracy*. ERIC, 95–137.

Mullen, Lincoln. 2016. *tokenizers: A Consistent Interface to Tokenize Natural Language Text*. <https://CRAN.R-project.org/package=tokenizers>.

Neuwirth, Erich. 2014. *RColorBrewer: ColorBrewer Palettes*. <https://CRAN.R-project.org/package=RColorBrewer>.

Neyman, J., und E. S. Pearson. 1933. „On the Problem of the Most Efficient Tests of Statistical Hypotheses“. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 231 (694-706): 289–337. doi:10.1098/rsta.1933.0009⁶.

Neyman, Jerzy. 1935. „On the problem of confidence intervals“. *The annals of mathematical statistics* 6 (3). JSTOR: 111–16.

Neyman, Jerzy, und Egon S Pearson. 1992. „On the problem of the most efficient tests of statistical hypotheses“. In *Breakthroughs in statistics*, 73–108. New York City: Springer.

Ooms, Jeroen. 2016. *pdftools: Text Extraction and Rendering of PDF Documents*. <https://CRAN.R-project.org/package=pdftools>.

Peirce, Charles S. 1955. „Abduction and induction“. *Philosophical writings of Peirce* 11. New York.

Peng, Roger D, und Elizabeth Matsui. 2015. „The Art of Data Science“. *A Guide for Anyone Who Works with Data*. Skybrude Consulting 200: 162.

Prel, Jean-Baptist du, Bernd Roehrig, Gerhard Hommel, und Maria Blettner. 2010. *Deutsches Aerzteblatt Online*, Mai. Deutscher Aerzte-Verlag. doi:10.3238/arztebl.2010.0343⁷.

Programmieren mit R. 2009. Springer Berlin Heidelberg. doi:10.1007/978-3-540-79998-6⁸.

Raiche, Gilles, und David Magis. 2011. *nFactors: Parallel Analysis and Non Graphical Solutions to the Cattell Scree Test*. <https://CRAN.R-project.org/package=nFactors>.

Ram, Karthik, und Hadley Wickham. 2015. *wesanderson: A Wes Anderson Palette Generator*. <https://CRAN.R-project.org/package=wesanderson>.

Re, AC Del. 2014. *compute.es: Compute Effect Sizes*. <https://CRAN.R-project.org/package=compute.es>.

Remus, R., U. Quasthoff, und G. Heyer. 2010. „SentiWS – a Publicly Available German-language Resource for Sentiment Analysis“. In *Proceedings of the 7th International Language*

⁶<https://doi.org/10.1098/rsta.1933.0009>

⁷<https://doi.org/10.3238/arztebl.2010.0343>

⁸<https://doi.org/10.1007/978-3-540-79998-6>

Resources and Evaluation (LREC'10), 1168–71.

Ripley, Brian. 2016. *MASS: Support Functions and Datasets for Venables and Ripley's MASS*. <https://CRAN.R-project.org/package=MASS>.

rita, Bureau of transportation statistics. 2013. „nycflights13“. http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236.

Robinson, David. 2016. *gutenbergr: Download and Process Public Domain Works from Project Gutenberg*. <https://cran.rstudio.com/package=gutenbergr>.

Robinson, David, Matthieu Gomez, Boris Demeshev, Dieter Menne, Benjamin Nutter, Luke Johnston, Ben Bolker, Francois Briatte, und Hadley Wickham. 2015. *broom: Convert Statistical Analysis Objects into Tidy Data Frames*. <https://CRAN.R-project.org/package=broom>.

Robinson, David, und Julia Silge. 2016. *tidytext: Text Mining using 'dplyr', 'ggplot2', and Other Tidy Tools*. <https://CRAN.R-project.org/package=tidytext>.

Romeijn, Jan-Willem. 2016. „Philosophy of Statistics“. In *The Stanford Encyclopedia of Philosophy*, herausgegeben von Edward N. Zalta, Winter 2016. <http://plato.stanford.edu/archives/win2016/entries/statistics/>.

Rucker, Rudy. 2004. *Infinity and the Mind*. Princeton: Princeton University Press. <https://books.google.de/books?id=MDOU AwAAQBAJ>.

Sauer, Sebastian. 2016. „Extraversion Dataset“. Open Science Framework. doi:10.17605/OSF.IO/4KGZH⁹.

———. 2017a. „Dataset 'predictors of performance in stats test'“. Open Science Framework. doi:10.17605/OSF.IO/SJHUY¹⁰.

———. 2017b. „Dataset 'Height and shoe size'“. Open Science Framework. doi:10.17605/OSF.IO/JA9DW¹¹.

Sauer, Sebastian, und Alexander Wolff. 2016. „The effect of a status symbol on success in online dating: an experimental study (data paper)“. *The Winnower*, August. doi:10.15200/winn.147241.13309¹².

Sauer, Sebastian, Harald Walach, und Niko Kohls. 2010. „Gray's Behavioural Inhibition System as a mediator of mindfulness towards well-being“. *Personality and Individual Differences* 50 (4). Pergamon: 506–51. doi:10.1016/j.paid.2010.11.019¹³.

Schloerke, Barret, Jason Crowley, Di Cook, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg, und Joseph Larmarange. 2016. *GGally: Extension to 'ggplot2'*. <https://CRAN.R-project.org/package=GGally>.

Schmidt, Peter, Sebastian Bamberg, Eldad Davidov, Johannes Herrmann, und Shalom H. Schwartz. 2007. „Die Messung von Werten mit dem Portraits Value Questionnaire“.

⁹<https://doi.org/10.17605/OSF.IO/4KGZH>

¹⁰<https://doi.org/10.17605/OSF.IO/SJHUY>

¹¹<https://doi.org/10.17605/OSF.IO/JA9DW>

¹²<https://doi.org/10.15200/winn.147241.13309>

¹³<https://doi.org/10.1016/j.paid.2010.11.019>

Zeitschrift für Sozialpsychologie 38 (4). Hogrefe Publishing Group: 261–75. doi:10.1024/0044-3514.38.4.261¹⁴.

Shmueli, Galit. 2010. „To Explain or to Predict?“ *Statistical Science* 25 (3): 289–310. doi:10.1214/10-STS330¹⁵.

Silge, Julia. 2016. *janeaustenr: Jane Austen’s Complete Novels*. <https://CRAN.R-project.org/package=janeaustenr>.

Silge, Julia, David Robinson, und Jim Hester. 2016. „tidytext: Text mining using dplyr, ggplot2, and other tidy tools“. doi:10.5281/zenodo.56714¹⁶.

Silge, Julia, und David Robinson. 2016. „tidytext: Text Mining and Analysis Using Tidy Data Principles in R“. *The Journal of Open Source Software* 1 (3). The Open Journal. doi:10.21105/joss.00037¹⁷.

Spurzem, Lothar. 2017. „VW 1303 von Wiking in 1:87“. [https://de.wikipedia.org/wiki/Modellautomobil#/media/File:Wiking-Modell_VW_1303_\(um_1975\).JPG](https://de.wikipedia.org/wiki/Modellautomobil#/media/File:Wiking-Modell_VW_1303_(um_1975).JPG).

Suppes, Patrick, und Joseph L Zinnes. 1962. *Basic measurement theory*. Institute for mathematical studies in the social sciences.

The Oxford Dictionary of Statistical Terms. 2006. Oxford University Press.

Therneau, Terry, Beth Atkinson, und Brian Ripley. 2015. *rpart: Recursive Partitioning and Regression Trees*. <https://CRAN.R-project.org/package=rpart>.

Tufte, Edward R. 1990. *Envisioning Information*. Graphics Press.

———. 2001. *The Visual Display of Quantitative Information*. Graphics Press.

———. 2006. *Beautiful Evidence*. Graphics Press.

Unrau, Sebastian. 2017. „No Title“. <https://unsplash.com/photos/CoD2Q92UaEg>.

VanDerWal, Jeremy, Lorena Falconi, Stephanie Januchowski, Luke Shoo, und Collin Storlie. 2014. *SDMTools: Species Distribution Modelling Tools: Tools for processing data associated with species distribution modelling exercises*. <https://CRAN.R-project.org/package=SDMTools>.

Wagenmakers, Eric-Jan. 2007. „A practical solution to the pervasive problems of p values“ *Psychonomic Bulletin & Review* 14 (5). Springer Nature: 779–804. doi:10.3758/bf03194105¹⁸.

Warnes, Gregory R., Ben Bolker, Lodewijk Bonebakker, Robert Gentleman, Wolfgang Huber, Andy Liaw, Thomas Lumley, Martin Maechler, u. a. 2016. *gplots: Various R Programming*

¹⁴<https://doi.org/10.1024/0044-3514.38.4.261>

¹⁵<https://doi.org/10.1214/10-STS330>

¹⁶<https://doi.org/10.5281/zenodo.56714>

¹⁷<https://doi.org/10.21105/joss.00037>

¹⁸<https://doi.org/10.3758/bf03194105>

- Tools for Plotting Data.* <https://CRAN.R-project.org/package=gplots>.
- Wei, Taiyun, und Viliam Simko. 2016. *corrplot: Visualization of a Correlation Matrix.* <https://CRAN.R-project.org/package=corrplot>.
- Wicherts, Jelte M., Coosje L. S. Veldkamp, Hilde E. M. Augusteijn, Marjan Bakker, Robbie C. M. van Aert, und Marcel A. L. M. van Assen. 2016. „Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking“. *Frontiers in Psychology* 7 (November). Frontiers Media SA. doi:10.3389/fpsyg.2016.01832¹⁹.
- Wickham, Hadley. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. <http://ggplot2.org>.
- . 2014a. *Advanced R*. Boca Raton, Florida: CRC Press.
- . 2014b. „Tidy Data“. *Journal of Statistical Software* 59 (1): 1–23. doi:10.18637/jss.v059.i10²⁰.
- . 2016a. *reshape2: Flexibly Reshape Data: A Reboot of the Reshape Package.* <https://CRAN.R-project.org/package=reshape2>.
- . 2016b. *tidyR: Easily Tidy Data with ‘spread()’ and ‘gather()’ Functions.* <https://CRAN.R-project.org/package=tidyr>.
- . 2017a. *nycflights13: Flights that Departed NYC in 2013.* <https://CRAN.R-project.org/package=nycflights13>.
- . 2017b. *stringr: Simple, Consistent Wrappers for Common String Operations.* <https://CRAN.R-project.org/package=stringr>.
- . 2017c. *tidyverse: Easily Install and Load ‘Tidyverse’ Packages.* <https://CRAN.R-project.org/package=tidyverse>.
- Wickham, Hadley, Jim Hester, und Romain Francois. 2016a. *readr: Read Tabular Data.* <https://CRAN.R-project.org/package=readr>.
- . 2016b. *readr: Read Tabular Data.* <https://CRAN.R-project.org/package=readr>.
- Wickham, Hadley, und Romain Francois. 2016. *dplyr: A Grammar of Data Manipulation.* <https://CRAN.R-project.org/package=dplyr>.
- Wickham, Hadley, und Garrett Grolemund. 2016. *R for Data Science: Visualize, Model, Transform, Tidy, and Import Data*. O'Reilly Media.
- Wikipedia. 2017. „Körpergröße — Wikipedia, Die freie Enzyklopädie“. <https://de.wikipedia.org/w/index.php?title=K%C3%B6rpergr%C3%B6%C3%9Fe&oldid=165047921>.
- Wild, Chris J, und Maxine Pfannkuch. 1999. „Statistical thinking in empirical enquiry“. *International Statistical Review* 67 (3). Wiley Online Library: 223–48.
- Wild, Fridolin. 2015. *lsa: Latent Semantic Analysis.* <https://CRAN.R-project.org/>

¹⁹<https://doi.org/10.3389/fpsyg.2016.01832>

²⁰<https://doi.org/10.18637/jss.v059.i10>

package=lsa.

Wilkinson, Leland. 2006. *The grammar of graphics*. Springer Science & Business Media.

Xie, Yihui. 2015. *Dynamic Documents with R and knitr*. 2nd Aufl. Boca Raton, Florida: Chapman; Hall/CRC. <http://yihui.name/knitr/>.

———. 2016. *knitr: A General-Purpose Package for Dynamic Report Generation in R*. <https://CRAN.R-project.org/package=knitr>.

Zumel, Nina, John Mount, und Jim Porzak. 2014. *Practical data science with R*. Manning.

Index

- 70
Überanpassung, 134
- Allgemeines Lineares Modells, 188
angeleitetes Lernen, 132
- Befehl, Funktion, 14
Bestimmtheitsmaß, 170
Bias, 135
Binnen, 70
Biplot, 251
- Chancen, 189
Cohens d, 157
Cronbachs Alpha, 256
- Dataframe, 13
datengenerierende Maschine, 130
Datenjudo, 34
Determinationskoeffizient, 170
deterministisch, 130
Dimensionsreduzierendes Modellieren, 133
dplyr::arrange, 40
dplyr::count, 48
dplyr::filter, 36
dplyr::mutate, 54
dplyr::n, 47
dplyr::select, 38
dplyr::summarise, 45
Durchpfeifen, 52
- Effektstärke, 156
Eigenwert, 250
einfaches reproduzierbares Beispiel, 16
Einflussgrößen, 130
Ellbogen-Kriterium, 249
Erklären, 131
euklidischen Abstand, 230
Explikatives Modellieren, 131
- Exploratorische Faktorenanalyse, 243, 253
Faktorenanalyse, 253
Faktorstufen, 12
Fallreduzierendes Modellieren, 132
Funktion, 5
- Geleitetes Modellieren, 131
- Hauptachsenanalyse, 243
Hauptkomponente, 244
Hauptkomponentenanalyse, 243
- Interaktionseffekt, 180
interne Konsistenz, 256
Item, 242
- Klassifikation, 131, 194
Konfidenzintervalle, 155
Konfusionsmatrix, 194
Konsole, 5
Kriterium, 130
Kriterium der Kleinsten Quadrate, 172
- Lagemaße, 57
latente Konstrukte, 242
Lernen ohne Anleitung, 133
logistische Funktion, 187
Logit, 189
- Messen, 242
multivariat, 178
- Nullhypothese, 144
Nullhypotesen-Signifikanztesten, 144
- Odds, 189
Ordinary Least Squares, 172
overfitting, 134
- p-Wert, 144

Parameter eines R-Befehls, 14

PCA, 243

Pfeife, 52

Prädiktives Modellieren, 131

Prädiktoren, 130

R-Skript, 12

Reduzieren, 132

Regression, 166

Relation, 128

Residuen, 173

robust, 135

ROC, 196

Signifikanz, 145

Skript-Fenster, 5

Split-Half-Reliabilität, 256

Standardnormalverteilung, 67

Streuungsmaße, 57

Tibbles, 13

Umgebung, 5

Umkodieren, 69

underfitting, 134

Ungeleiteten Modellieren, 132

unsupervised learning, 133

Unteranpassung, 134

Variablen, 13

Vektoren, 12

Vorhersagefehler, 170

Vorhersagen, 131

Youden-Index, 195

