

Start:Bayes!

Sebastian Sauer

2022-12-30T00:00:00+01:00

Inhaltsverzeichnis

Hinweise	13
Lernziele	13
Voraussetzungen	14
Software	14
PDF-Version	14
Lernhilfen	15
Videos	15
Online-Zusammenarbeit	15
Modulzeitplan	15
Literatur	15
Technische Details	16
1 Prüfung	18
1.1 Prüfungleistung	18
1.2 Grundsätzliches	18
1.3 Wiederholungsprüfungen	18
1.4 Prüfungsstoff	19
1.5 Bearbeitungshinweise	19
1.5.1 Aufgaben zur Datenanalyse	20
1.6 Teilnahmebedingungen	20
1.7 Organisatorische Hinweise	21
1.8 Open-Book-Prüfungen	22
1.9 Zeitrahmen der Prüfung	22
1.10 Technische und organisatorische Anforderungen einer Open-Book-Prüfung . .	22
1.11 Technische Probleme während der Prüfung	22
1.12 Prüfungsrecht	23
1.13 Bitte formulieren Sie Beanstandungen nachvollziehbar	23
1.14 Datenschutz	23
1.15 Plagiatskontrolle	23
1.16 Typische Fehler in der Prüfung	24
1.17 Hinweise zu Scheimängeln	24
2 Inferenz	26
2.1 Lernsteuerung	26
2.1.1 Lernziele	26
2.1.2 Begleitvideos	26
2.2 Wozu ist Statistik überhaupt da?	27

2.3	Was ist Inferenz?	27
2.3.1	Inferenz als Generalisieren	27
2.4	Stichprobe vs. Population	28
2.4.1	Deskriptiv- vs. Inferenzstatistik	28
2.4.2	Wozu ist die Inferenstatistik gut?	30
2.4.3	Deskriptiv- und Inferenzstatistik gehen Hand in Hand	30
2.4.4	Schätzen von Parametern einer Grundgesamtheit	30
2.4.5	Beispiele für inferenzstatistische Fragestellungen	31
2.5	Modellieren	31
2.5.1	Modellieren als Grundraster des Erkennens	31
2.5.2	Vertiefung	32
2.6	Regression	33
2.6.1	Regression zum Modellieren	33
2.6.2	Viele statistische Verfahren sind Spezialfälle der Regression	34
2.6.3	In voller Pracht	34
2.7	Unsicherheit	36
2.7.1	Inferenz beinhaltet Unsicherheit	36
2.7.2	Beispiele zur Quantifizierung von Ungewissheit	36
2.7.3	Zwei Arten von Ungewissheit	37
2.7.4	Ich weiß, was ich nicht weiß: Ungewissheit angeben	37
2.7.5	Visualisierung von Ungewissheit	38
2.7.6	Konfidenzintervall	39
2.8	Klassische vs. Bayes-Inferenz	40
2.8.1	Klassische Inferenz: Frequentismus	40
2.8.2	Bayesianische Inferenz	40
2.8.3	Vergleich von Wahrscheinlichkeitsaussagen	40
2.8.4	Frequentist und Bayesianer	41
2.8.5	Der p-Wert ist wenig intuitiv	41
2.8.6	Beispiel zum Nutzen von Apriori-Wissen 1	42
2.8.7	Beispiel zum Nutzen von Apriori-Wissen 2	42
2.9	Literatur	42
2.10	Aufgaben	42
3	Wahrscheinlichkeit	43
3.1	Lernsteuerung	43
3.1.1	Lernziele	43
3.1.2	Prüfungsrelevanter Stoff	43
3.1.3	Zentrale Begriffe	43
3.1.4	Begleitvideos	44
3.2	Unterstützung: Wahrscheinlichkeit in Bildern	44
3.2.1	Zufall	44
3.2.2	Relationen von Mengen	45
3.2.3	Additionssatz	47
3.2.4	Bedingte Wahrscheinlichkeit	48
3.2.5	(Un-)Abhängigkeit	49
3.2.6	Multiplikationssatz	50

3.2.7	Totale Wahrscheinlichkeit	53
3.2.8	Bayes	54
3.3	Bayes' Theorem	55
3.3.1	Bayes als bedingte Wahrscheinlichkeit	55
3.3.2	Wozu wird Bayes in der Praxis genutzt?	56
3.3.3	Zusammengesetzte Hypothesen	56
3.4	Aufgaben	57
4	Verteilungen	58
4.1	Lernsteuerung	58
4.1.1	Lernziele	58
4.1.2	Prüfungsrelevanter Stoff	58
4.1.3	Benötigte R-Pakete	58
4.1.4	Zentrale Begriffe	59
4.1.5	Begleitvideos	59
4.2	Verteilungen	59
4.2.1	Häufigkeitsverteilung	59
4.2.2	Wahrscheinlichkeitsverteilung	60
4.3	Gleichverteilung	61
4.3.1	Indifferenz als Grundlage	61
4.3.2	Simulation	61
4.4	Binomialverteilung	64
4.4.1	Veranschaulichung	64
4.4.2	Rechnen mit R	66
4.4.3	Simulieren	67
4.5	Normalverteilung	70
4.5.1	Normal auf dem Fußballfeld	71
4.5.2	Normal durch Addieren	72
4.5.3	Normalverteilung vs. randlastige Verteilungen	73
4.5.4	Beispiele für Normal- und randlastige Verteilungen	73
4.5.5	Formel der Normalverteilung	74
4.5.6	Simulation einer Normalverteilung	75
4.5.7	IQ-Verteilung	76
4.5.8	Quantile der Normalverteilung	78
4.5.9	Standardnormalverteilung	80
4.5.10	Normalverteilung als konservative Wahl	80
4.6	Aufgaben	81
5	Globusversuch	82
5.1	Lernsteuerung	82
5.1.1	Lernziele	82
5.1.2	Benötigte R-Pakete	82
5.1.3	Begleitvideos	82
5.2	Von Welten und Golems	83
5.2.1	Kleine Welt, große Welt	83
5.2.2	Der Golem von Prag	83

5.2.3	Wissenschaftliche Modelle sind wie Golems	84
5.2.4	So denkt unser Bayes-Golem	85
5.3	Ein erster Versuch: Wir werfen den Globus	86
5.3.1	Welcher Anteil der Erdoberfläche ist mit Wasser bedeckt?	86
5.3.2	Wie entstanden die Daten?	86
5.3.3	Ein paar Fachbegriffe	87
5.3.4	Bayes-Updates	87
5.3.5	Likelihood mit Binomialverteilung	88
5.3.6	Binomialverteilung mit R	89
5.3.7	Unser Modell ist geboren	91
5.3.8	Visualisierungen	92
5.4	Zur Erinnerung: Bayes Theorem	92
5.4.1	Herleitung Bayes' Theorem 1/2: Gemeinsame Wahrscheinlichkeit	92
5.4.2	Herleitung Bayes' Theorem 2/2: Posteriori-Wahrscheinlichkeit	93
5.4.3	Bayes' Theorem als Formel	93
5.4.4	Posteriori als Produkt von Priori und Likelihood	94
5.4.5	Wissen updaten: Wir füttern Daten in das Modell	94
5.5	Bayes berechnen mit mit dem Bayes-Gitter	95
5.5.1	Idee	95
5.5.2	Bayes-Gitter in R berechnen	96
5.5.3	Was sagt die Post?	97
5.6	Aufgaben	98
5.7	Abschluss	99
5.7.1	Zusammenfassung	99
5.7.2	Vertiefung	99
5.7.3	Literatur	99
6	Die Post befragen	100
6.1	Lernsteuerung	100
6.1.1	Lernziele	100
6.1.2	Benötigte R-Pakete	100
6.1.3	Begleitvideos	100
6.2	Mit Stichproben die Post-Verteilung zusammenfassen	100
6.2.1	Zur Erinnerung: Gitterwerte in R berechnen	100
6.2.2	Beispiele für Fragen an die Post-Verteilung	101
6.2.3	Bayes-Box für komplexe Modelle	102
6.2.4	Wir arbeiten jetzt mit Häufigkeit, nicht mit Wahrscheinlichkeit	103
6.2.5	Häufigkeiten sind einfacher als Wahrscheinlichkeiten	103
6.2.6	Visualisierung der Stichprobendaten mit $k = 100$ Gitterwerten	105
6.3	Die Post-Verteilung befragen	107
6.3.1	Fragen nach Wahrscheinlichkeiten	108
6.3.2	Fragen nach Parameterwerten	111
6.3.3	Zur Erinnerung: Quantile	112
6.3.4	Den Quantilen unter die Motorhaube geschaut	113
6.3.5	Visualisierung der Intervalle	114

6.4	Schiefe Posteriori-Verteilungen sind möglich	114
6.4.1	50%-Perzentil-Intervall	115
6.4.2	50%-Intervall höchster Dichte	116
6.5	Fazit	117
6.5.1	Intervalle höchster Dichte vs. Perzentilintervalle	117
6.5.2	Zusammenfassung	117
6.6	Aufgaben	118
7	Vorhersage-Verteilung	119
7.1	Lernsteuerung	119
7.1.1	Lernziele	119
7.1.2	Benötigte R-Pakete	119
7.2	Der zwielichtige Dozent: Stichproben-Vert. vs. Post-Vert.	119
7.3	Mit Stichproben neue Beobachtungen simulieren	123
7.3.1	Wir simulieren die Wasserzahl bei Globuswürfen	124
7.3.2	Traue niemals einem Golem (einem Modell)	125
7.4	Mit guten Simulationen kommt man den wahren Werten nahe	125
7.5	Stichprobenverteilung	126
7.6	Die Posterior-Prädiktiv-Verteilung (PPV)	127
7.6.1	Was ist die PPV und wozu ist sie gut?	127
7.6.2	Visualisierung der PPV	128
7.6.3	PPV berechnen	129
7.7	Fazit	132
7.7.1	Vorhersagen sind schwierig	132
7.7.2	Zwei Arten von Ungewissheit in Vorhersagen von Modellen	133
7.7.3	Vergleich der Verteilungen	133
7.7.4	So viele Verteilungen...	134
7.8	Aufgaben	134
8	Gauss-Modelle	135
8.1	Lernsteuerung	135
8.1.1	Lernziele	135
8.1.2	Benötigte R-Pakete	135
8.1.3	Begleitvideos	136
8.2	Wie groß sind die !Kung San?	136
8.2.1	!Kung San	136
8.2.2	!Kung Data	137
8.2.3	Wir gehen apriori von normalverteilter Größe Der !Kung aus	137
8.3	Die Exponentialverteilung	138
8.3.1	Die Apfel-fällt-nicht-weit-vom-Stamm-Verteilung	138
8.3.2	Visualisierung verschiedener Exponentialverteilungen	140
8.4	Unser Gauss-Modell der !Kung	142
8.4.1	Modelldefinition	142
8.4.2	Priori gewichtet mit Likelihood ergibt Posteriori	142
8.4.3	Prioris	143
8.4.4	Fertig!	143

8.5	Zufällige Motivationsseite	143
8.6	Posteriori-Verteilung des Größen-Modells, <code>m41</code>	143
8.6.1	Hallo, Posteriori-Verteilung	145
8.6.2	Posteriori-Stichproben mit <code>stan_glm()</code> berechnen	146
8.6.3	Ausgabe von <code>stan_glm()</code>	147
8.7	Wie tickt <code>stan_glm()</code> ?	148
8.7.1	Schätzwerte zu den Modellparameter	148
8.7.2	Stichproben aus der Posteriori-Verteilung ziehen	148
8.7.3	Standard-Prioriwerte bei <code>stan_glm()</code>	151
8.8	Modell <code>m42</code> : unsere Priori-Werte	152
8.8.1	<code>m42</code>	153
8.8.2	Posteriori-Verteilung und Parameter plotten	153
8.9	Fazit	154
8.10	Wahl der Priori-Werte	155
8.10.1	Welche Beobachtungen sind auf Basis unseres Modells zu erwarten? .	155
8.10.2	Priori-Werte prüfen mit der Priori-Prädiktiv-Verteilung	156
8.10.3	Vorhersagen der Priori-Werte	158
8.10.4	Extrem vage Priori-Verteilung für die Streuung?	158
8.11	Aufgaben	160

9	Lineare Modelle	161
9.1	Lernsteuerung	161
9.1.1	Lernziele	161
9.1.2	Benötigte R-Pakete	161
9.1.3	Begleitvideoso	161
9.2	Post-Verteilung der Regression	162
9.2.1	Einfache Regression	162
9.2.2	Bei jedem Prädiktorwert eine Post-Verteilung für μ	163
9.2.3	Statistiken zum !Kung-Datensatz	163
9.2.4	Etwas mehr EDA	164
9.2.5	Prädiktor zentrieren	167
9.3	Modell <code>m43</code> : zentrierter Prädiktor	168
9.3.1	Modelldefinition von <code>m43</code>	169
9.3.2	Likelihood, <code>m43</code>	169
9.3.3	Regressionsformel, <code>m43</code>	169
9.3.4	Priori-Werte des Modells <code>m43</code>	170
9.4	Vertiefung: Prior-Prädiktiv-Verteilung	170
9.4.1	Moment	170
9.4.2	Priori-Prädiktiv-Verteilung für <code>m43</code>	170
9.4.3	Prior-Prädiktiv-Simulation für <code>m43</code> mit <code>stan_glm()</code>	171
9.4.4	Visualisieren der Prior-Prädiktiv-Verteilung	172
9.4.5	Ein positiver Wert für β ist plausibler	172
9.4.6	Priori-Prädiktiv-Simulation, 2. Versuch	174
9.4.7	Visualisieren der Prior-Prädiktiv-Verteilung, <code>m43a</code>	174
9.4.8	Moment, kann hier jeder machen, was er will?	175
9.4.9	Hier ist unser Modell, <code>m43a</code>	176

9.4.10 Eine Zusammenfassung der Posteriori-Verteilung für $m43a$	176
9.5 Die Post-Verteilung befragen	177
9.5.1 $m43a$	177
9.5.2 Mittelwerte von α und β aus der Post-Verteilung	177
9.5.3 Visualisieren der “mittleren” Regressioneraden	179
9.5.4 Zentrale Statistiken zu den Parametern	181
9.5.5 Streuungsmaße zu den Parametern	185
9.5.6 Ungewissheit von α und β aus der Post-Verteilung visualisiert	185
9.5.7 Fragen zu Quantilen des Achsenabschnitts	189
9.5.8 Fragen zu Wahrscheinlichkeitsmassen des Achsenabschnitts	189
9.5.9 Typischer Bayes-Nutzer in Aktion	190
9.6 Post-Verteilung bedingt auf einen Prädiktorwert	191
9.6.1 Visualisierung	191
9.6.2 Lagemaße und Streuungen	192
9.7 Die PPV befragen	193
9.7.1 Perzentil-Intervalle für verschiedenen Prädiktor-Werte	194
9.7.2 Perzentilintervalle für verschiedenen Prädiktorwerte visualisiert	195
9.7.3 Die PPV visualisiert	195
9.7.4 Fragen an die PPV	197
9.8 Aufgaben	198
10 Forschungsfragen mit metrischer AV	199
10.1 Lernsteuerung	199
10.1.1 R-Pakete	199
10.1.2 Lernziele	199
10.1.3 Begleitvideos	199
10.2 Wissenschaft als Gerechtigkeitsprojekt	200
10.2.1 Meinungen als Grundlage der Konfliktlösung ?	200
10.2.2 Fakten als Grundlage der Konfliktlösung	200
10.2.3 Beispiel Corona: Datenlage spricht zugunsten der Covid19-Impfung	200
10.2.4 Psychologische Intervention zur Erhöhung der Impfquote	201
10.2.5 Was heißt “ist effektiv”?	201
10.3 Arten von Forschungsfragen	202
10.3.1 Nach dem Erkenntnisziel	202
10.3.2 Nach dem Skalenniveau	203
10.4 Eine binäre UV	203
10.4.1 Forschungsfrage	203
10.4.2 IQ von Kindern, binärer Prädiktor	203
10.4.3 Interpretation von $m10.1$	204
10.4.4 $m10.1$ als Mittelwertsdifferenz	205
10.4.5 Antwort auf die Forschungsfrage, $m10.1$	205
10.5 Eine metrische plus eine nominale UV	207
10.5.1 Forschungsfrage	207
10.5.2 1 metrischer Prädiktor	208
10.5.3 Beide Prädiktoren, $m10.3$	210

10.6	Interaktion	212
10.6.1	Interpretation von m10.4	213
10.6.2	Nach der Interpretation von 20 unzentrierten Koeffizienten	214
10.7	Zentrieren von Prädiktoren	214
10.7.1	Interpretation von m10.5	215
10.7.2	Zentrieren ändert nichts an den Vorhersagen	215
10.7.3	Perzentilintervalle aus der Posterori-Verteilung	216
10.7.4	Beantworten der Forschungsfrage	217
10.8	Eine nominale UV mit mehreren Stufen	217
10.8.1	Forschungsfrage	217
10.8.2	Alle Mittelwerte sind gleich, exakt gleich (?)	218
10.8.3	Erster Blick in den Datensatz penguins	218
10.8.4	Visualisierung (EDA)	219
10.8.5	Mittlere Gewichtsunterschiede in der Population	220
10.8.6	Interpretation von m10.6	221
10.8.7	Glauben wir jetzt an Gruppeneffekte?	222
10.9	Priori-Werte	223
10.9.1	Wechsel der Referenzkategorie	224
10.10	Modellgüte mit R-Quadrat bestimmen	225
10.10.1	Modellgüte mit R^2 bestimmen	225
10.10.2	Definition vom "klassischen" R^2	226
10.10.3	Bayes' R^2	227
10.11	Nullhypotesen sind praktisch immer falsch	227
10.11.1	Alternativen zu Nullhypotesen	228
10.11.2	"Praktisch" kein Unterschied: Das Rope-Konzept	228
10.11.3	HDI-Rope-Entscheidungsregel visualisiert	229
10.11.4	Rope berechnen	229
10.11.5	Visualisierung unserer Rope-Werte, m10.6	230
10.11.6	Finetuning des Rope	230
10.11.7	Beantwortung der Forschungsfrage	231
10.12	Mehrere metrische UV	231
10.12.1	Forschungsfrage	231
10.12.2	Was heißt, X hängt mit Y zusammen?	232
10.12.3	Korrelationen zur Forschungsfrage	232
10.12.4	Univariate Regressionen	234
10.12.5	Visualisierung der univariaten Regressionen	234
10.12.6	Multiples Modell (beide Prädiktoren), m10.9	236
10.12.7	3D-Visualisierung eines Modells mit zwei Prädiktoren 1	236
10.12.8	Visualisierung mit Farbe statt 3. Dimension	236
10.12.9	Visualisierung in 10 Dimensionen	238
10.12.10	Relevanz der Prädiktoren	238
10.12.11	ζ -Standardisierung	239
10.12.12	Statistiken zu den z-transformierten Variablen	239
10.13	Modell m10.10	241
10.13.1	95%-PI	241
10.13.2	Was ist ein kleiner, was ein großer Effekt?	242

10.13.3 Vernachlässigbarer Regressionseffekt	243
10.13.4 Modellgüte	243
10.13.5 Priori-Verteilung für <code>m10.10</code> und Modelldefinition	243
10.13.6 Beantwortung der Forschungsfrage	244
10.14 Vertiefung	245
10.14.1 Verwandtheit von Korrelation und Regression	245
10.14.2 Prüfen der Linearitätsannahme	246
10.14.3 Modellprüfung mit der PPV	247
10.14.4 Visualisierung der bereinigten Regressionskoeffizienten	248
10.14.5 Bereinigte Regressionskoeffizienten für <code>mtcars</code>	250
10.14.6 Post: Bayes in fünf Minuten	251
10.15 Ausblick: Binäre AV	251
10.16 Aufgaben	252
11 Kausalinferenz	254
11.1 Lernsteuerung	254
11.1.1 R-Pakete	254
11.1.2 Lernziele	254
11.2 Statistik, was soll ich tun?	254
11.2.1 Studie A: Östrogen	254
11.2.2 Studie B: Blutdruck	256
11.2.3 Studie A und B: Gleiche Daten, unterschiedliches Kausalmodell	257
11.2.4 Sorry, Statistik: Du allein schaffst es nicht	257
11.2.5 Vertiefung	257
11.2.6 Zwischenfazit	259
11.3 Konfundierung	259
11.3.1 Datensatz ‘Hauspreise im Saratoga County’	259
11.3.2 Immobilienpreise in einer schicken Wohngegend vorhersagen	259
11.3.3 Modell 1: Preis als Funktion der Anzahl der Zimmer	260
11.3.4 Don hat eine Idee	261
11.3.5 R-Funktionen, um Beobachtungen vorhersagen	262
11.3.6 Modell 2	263
11.3.7 Die Zimmerzahl ist negativ mit dem Preis korreliert	265
11.4 Kontrollieren von Variablen	265
11.4.1 Das Hinzufügen von Prädiktoren kann die Gewichte der übrigen Prädiktoren ändern	266
11.5 Welches Modell richtig ist, kann die Statistik nicht sagen	266
11.5.1 Kausalmodell für Konfundierung, <code>km1</code>	267
11.5.2 <code>m2</code> kontrolliert die Konfundierungsvariable <code>livingArea</code>	267
11.5.3 Konfundierer kontrollieren	268
11.5.4 <code>m1</code> und <code>m2</code> passen nicht zu den Daten, wenn <code>km1</code> stimmt	268
11.5.5 Kausalmodell 2, <code>km2</code>	269
11.5.6 Dons Kausalmodell, <code>km3</code>	269
11.5.7 Unabhängigkeiten laut der Kausalmodelle	271
11.6 DAGs: Directed Acyclic Graphs	272
11.6.1 Leider passen potenziell viele DAGs zu einer Datenlage	272

11.6.2 Was ist eigentlich eine Ursache?	273
11.6.3 Zwischenfazit	273
11.6.4 Schoki macht Nobelpreis! (?)	274
11.7 Kollision	274
11.7.1 Kein Zusammenhang von Intelligenz und Schönheit (?)	274
11.7.2 Aber Ihre Dates sind entweder schlau oder schön	276
11.8 DAG zur Rettung	276
11.8.1 Was ist eine Kollision?	277
11.8.2 Einfaches Beispiel zur Kollision	277
11.8.3 Noch ein einfaches Beispiel zur Kollision	278
11.8.4 Durch Kontrollieren entsteht eine Verzerrung bei der Kollision	278
11.8.5 IQ, Fleiss und Eignung fürs Studium	278
11.8.6 Schlagzeile "Schlauheit macht Studenten faul!"	279
11.8.7 Kollisionsverzerrung nur bei Stratifizierung	280
11.8.8 Einfluss von Großeltern und Eltern auf Kinder	281
11.9 Vertiefung	282
11.9.1 Der Gespenster-DAG	282
11.10 Die Hintertür schließen	283
11.10.1 Zur Erinnerung: Konfundierung	283
11.10.2 Gute Experimente zeigen den echten kausalen Effekt	284
11.10.3 Hintertür schließen auch ohne Experimente	285
11.10.4 Die vier Atome der Kausalanalyse	285
11.10.5 Mediation	286
11.11 Der Nachfahre	286
11.11.1 Kochrezept zur Analyse von DAGs	286
11.12 Schließen Sie die Hintertür (wenn möglich)!, bsp1	287
11.12.1 Schließen Sie die Hintertür (wenn möglich)!, bsp2	288
11.12.2 Implizierte bedingte Unabhängigkeiten von bsp2	288
11.13 Fazit	289
11.14 Aufgaben	289
12 Abschluss	290
12.1 Lernsteuerung	290
12.1.1 Lernziele	290
12.1.2 Benötigte R-Pakete	290
12.2 Lieblinglingsfehler	290
12.3 Post-Präd-Verteilung (PPV) und Post-Verteilung verwechseln	291
12.4 Quantile und Verteilungsfunktion verwechseln	292
12.4.1 Quantil für p	292
12.4.2 Verteilungsfunktion F	292
12.5 Interaktion falsch interpretieren	292
12.6 Kochrezepte	294
12.6.1 Kochrezept: Forschungsfrage untersuchen	294
12.6.2 Parameter schätzen vs. Hypothesen prüfen	295
12.6.3 Formalisierung von Forschungsfragen	295

12.7	Kerngedanken Bayes	295
12.7.1	Zentraler Kennwert der Bayes-Statistik: Post-Verteilung	295
12.7.2	Posteriori als Produkt von Priori und Likelihood	296
12.8	Beispiele für Prüfungsaufgaben	296
12.8.1	Geben Sie den korrekten Begriff an!	296
12.8.2	DAG mit doppelter Konfundierung	297
12.8.3	DAG mit vielen Variablen	297
12.8.4	Ein Kausalmodell der Schizophrenie, van Kampen (2014)	297
12.8.5	Modelle berechnen	299
12.9	Viel Erfolg bei der Prüfung!	299
	References	300

Hinweise

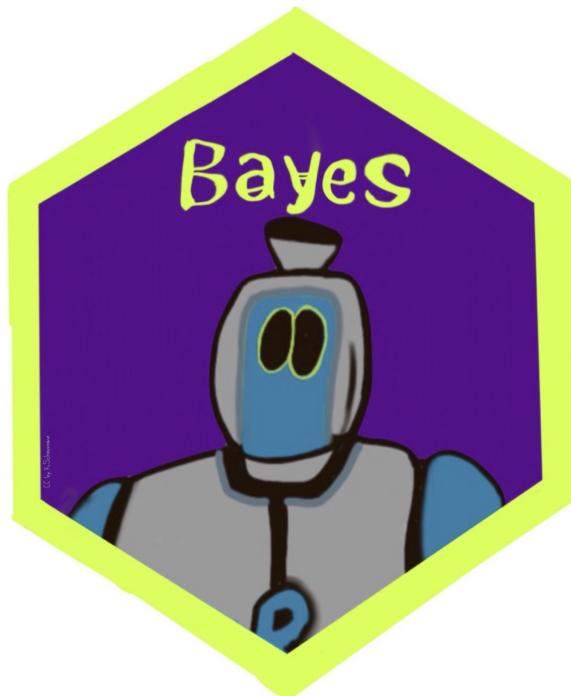


Abbildung 1: Bayes:Start!

Bildquelle: Klara Schaumann

WORK IN PROGRESS

Lernziele

Nach diesem Kurs sollten Sie ...

- grundlegende Konzepte der Inferenzstatistik mit Bayes verstehen und mit R anwenden können

- gängige einschlägige Forschungsfragen in statistische Modelle übersetzen und mit R auswerten können
- kausale Forschungsfragen in statistische Modelle übersetzen und prüfen können
- die Güte und Grenze von statistischen Modellen einschätzen können

Voraussetzungen

Um von diesem Kurs am besten zu profitieren, sollten Sie folgendes Wissen mitbringen:

- grundlegende Kenntnisse im Umgang mit R, möglichst auch mit dem tidyverse
- grundlegende Kenntnisse der deskriptiven Statistik
- grundlegende Kenntnis der Regressionsanalyse

Software

Installieren Sie [R und seine Freunde](#). Für die Bayes-Inferenz brauchen Sie¹ zusätzliche Software, was leider etwas Zusatzaufwand erfordert. Lesen Sie [hier](#) die Hinweise dazu. Installieren Sie die folgende R-Pakete²:

- `tidyverse`
- `rstanarm`
- `easystats`
- weitere Pakete werden im Unterricht bekannt gegeben (es schadet aber nichts, jetzt schon)

[R Syntax aus dem Unterricht](#) findet sich im Github-Repo bzw. Ordner zum jeweiligen Semester.

PDF-Version

– EXPERIMENTAL – EXPERIMENTAL

Von diesem “Webbuch” (HTML-Format) gibt es [hier eine PDF-Version](#). Die PDF-Version eignet sich zum Ausdrucken und zur Offline-Nutzung.

Allerdings wurden die Inhalte *in erster Linie für ein Webbuch-Format* formatiert, die PDF-Ausgabe ist daher nicht ideal. Es ist empfehlenswert, mit der Webbook-Version zu arbeiten. Außerdem wird die PDF-Version nicht ganz aktuell gehalten - die aktuelle Version ist immer die Webbuch-Variante. Prüfen Sie im Zweifel das Datum der Erstellung des Dokuments.

¹nicht gleich zu Beginn, aber nach 2-3 Wochen

²falls Sie die Pakete schon installiert haben, könnten Sie mal in RStudio auf “update.packages” klicken

Lernhilfen

Videos

Auf dem [YouTube-Kanal des Autors](#) finden sich eine Reihe von Videos mit Bezug zum Inhalt dieses Buchs. Besonders [diese Playlist](#) passt zu den Inhalten dieses Buchs.

Online-Zusammenarbeit

Hier finden Sie einige Werkzeuge, die das Online-Zusammenarbeiten vereinfachen:

- [Frag-Jetzt-Raum zum anonymen Fragen stellen während des Unterrichts](#). Der Keycode wird Ihnen bei Bedarf vom Dozenten bereitgestellt.
- [Padlet](#) zum einfachen (und anonymen) Hochladen von Arbeitsergebnissen der Studenten im Unterricht. Wir nutzen es als eine Art Pinwand zum Sammeln von Arbeitsbeiträgen. Die Zugangsdaten stellt Ihnen der Dozent bereit.

Modulzeitplan

Nr	Thema	Datum	Kommentar
1	Was ist Inferenz?	3. - 7. Okt. 2022	Erster Termin: 4. Okt. 22, 14.15-15.00
2	Wahrscheinlichkeit	10. - 14. Okt. 22	NA
3	Verteilungen	17. - 21. Okt. 22	NA
4	Globusversuch	24. - 28. Okt. 22	NA
5	Aufhol-Woche	31. Okt. - 4. Nov. 22	Am Di., 1.11. entfällt die Vorlesung. Am Do., 3. 11. e
6	Frag die Post	7. - 11. Nov. 22	Ab diese Woche benötigen wir rstanarm.
NA	NA	14. - 18. Nov. 22	Blockwoche: Kein regulärer Unterricht
7	Gauss-Modelle	21. - 25. Nov. 22	NA
8	Lineare Modelle	28. Nov. - 2. Dez. 22	NA
9	Metrische AV	5. Dez. - 9. Dez. 22	NA
10	Kausalinferenz 1	12. - 16. Dez. 22	NA
11	Kausalinferenz 2	19. - 23. Dez. 22	NA
NA	NA	NA	Jahreswechsel: Kein Unterricht
12	Abschluss	9. Jan. 23 - 13. Jan. 23	NA

Literatur

Pro Thema wird Literatur ausgewiesen.

Technische Details

Dieses Dokument wurde erzeugt am/um 2022-12-30 17:31:12.

```
## - Session info -----
## setting value
## version R version 4.2.1 (2022-06-23)
## os      macOS Big Sur ... 10.16
## system x86_64, darwin17.0
## ui      X11
## language (EN)
## collate en_US.UTF-8
## ctype   en_US.UTF-8
## tz      Europe/Berlin
## date    2022-12-30
## pandoc  2.19.2 @ /Applications/RStudio.app/Contents/Resources/app/quarto/bin/tools/
##
## - Packages -----
## package     * version date (UTC) lib source
## assertthat  0.2.1   2019-03-21 [1] CRAN (R 4.2.0)
## cellranger  1.1.0   2016-07-27 [1] CRAN (R 4.2.0)
## cli        3.4.1   2022-09-23 [1] CRAN (R 4.2.0)
## colorout    * 1.2-2   2022-06-13 [1] local
## colorspace  2.0-3   2022-02-21 [1] CRAN (R 4.2.0)
## DBI        1.1.3   2022-06-18 [1] CRAN (R 4.2.0)
## digest      0.6.30  2022-10-18 [1] CRAN (R 4.2.0)
## dplyr       1.0.10  2022-09-01 [1] CRAN (R 4.2.0)
## evaluate    0.18    2022-11-07 [1] CRAN (R 4.2.0)
## fansi       1.0.3   2022-03-24 [1] CRAN (R 4.2.0)
## fastmap     1.1.0   2021-01-25 [1] CRAN (R 4.2.0)
## generics    0.1.3   2022-07-05 [1] CRAN (R 4.2.0)
## ggplot2     3.4.0   2022-11-04 [1] CRAN (R 4.2.0)
## glue        1.6.2   2022-02-24 [1] CRAN (R 4.2.0)
## gt          0.8.0   2022-11-16 [1] CRAN (R 4.2.0)
## gtable      0.3.1   2022-09-01 [1] CRAN (R 4.2.0)
## htmltools   0.5.3   2022-07-18 [1] CRAN (R 4.2.0)
## jsonlite    1.8.3   2022-10-21 [1] CRAN (R 4.2.1)
## knitr       1.41    2022-11-18 [1] CRAN (R 4.2.0)
## lifecycle   1.0.3   2022-10-07 [1] CRAN (R 4.2.0)
## magrittr    2.0.3   2022-03-30 [1] CRAN (R 4.2.0)
## munsell     0.5.0   2018-06-12 [1] CRAN (R 4.2.0)
## pillar      1.8.1   2022-08-19 [1] CRAN (R 4.2.0)
## pkgconfig   2.0.3   2019-09-22 [1] CRAN (R 4.2.0)
## R6          2.5.1   2021-08-19 [1] CRAN (R 4.2.0)
## readxl     1.4.1   2022-08-17 [1] CRAN (R 4.2.0)
```

```
## rlang      1.0.6   2022-09-24 [1] CRAN (R 4.2.0)
## rmarkdown   2.18    2022-11-09 [1] CRAN (R 4.2.0)
## rstudioapi  0.14    2022-08-22 [1] CRAN (R 4.2.0)
## scales     1.2.1    2022-08-20 [1] CRAN (R 4.2.0)
## sessioninfo 1.2.2    2021-12-06 [1] CRAN (R 4.2.0)
## stringi     1.7.8    2022-07-11 [1] CRAN (R 4.2.0)
## stringr     1.5.0    2022-12-02 [1] CRAN (R 4.2.0)
## tibble      3.1.8    2022-07-22 [1] CRAN (R 4.2.0)
## tidyselect   1.2.0    2022-10-10 [1] CRAN (R 4.2.0)
## utf8        1.2.2    2021-07-24 [1] CRAN (R 4.2.0)
## vctrs       0.5.1    2022-11-16 [1] CRAN (R 4.2.0)
## withr       2.5.0    2022-03-03 [1] CRAN (R 4.2.0)
## xfun        0.35     2022-11-16 [1] CRAN (R 4.2.0)
## yaml        2.3.6    2022-10-18 [1] CRAN (R 4.2.0)
##
## [1] /Users/sebastiansaueruser/Rlibs
## [2] /Library/Frameworks/R.framework/Versions/4.2/Resources/library
##
## -----
```

1 Prüfung

1.1 Prüfungleistung

Die Prüfungsleistung besteht aus einer *Open-Book-Prüfung*.

1.2 Grundsätzliches

Die folgenden Hinweise gelten *grundsätzlich*, d.h. *soweit nicht anders* in der jeweiligen Prüfung bzw. der jeweiligen Aufgabe angegeben.

Nichtbeachten von Prüfungshinweisen kann zu Punkteabzug oder Nichtbestehen führen.

1.3 Wiederholungsprüfungen

- Wenn Sie bei einer Prüfung durchgefallen sein sollten, so haben Sie grundsätzlich die Möglichkeit, die *Prüfung zu wiederholen*.
- Denken Sie daran, sich *rechtzeitig* für Prüfungsleistungen *anzumelden*; beachten Sie die Fristen.
- Die Termine für die Wiederholungsprüfungen werden stets zusammen/zeitgleich mit denen der regulären Prüfungen bekannt gegeben.
- Wird ein Modul im laufenden Semester nicht angeboten, gibt es eine Wiederholungsprüfung.
- Relevanter Stoff und formale Bedingungen (wie Prüfungsform) sind *grundsätzlich identisch zur letzten abgehaltenen Prüfung des Moduls* (d.h. sofern nicht anders angegeben). Daher sind Wiederholungsprüfungen vom Anspruch vergleichbar wie die reguläre Klausur. Die Prüfungen sollen möglichst gleich vom Anspruch sein, um Fairness zu gewährleisten.
- Beachten Sie immer die *Hinweise*, die für die Wiederholungsprüfung angegeben sind. Im Einzelfall keine eine Wiederholungsprüfung von der vorherigen Prüfung stärker abweichen. Es gelten immer die Regeln, die das Dozenten bei der jeweiligen Wiederholungsprüfung veröffentlicht hat.
- Wird ein *Modul im laufenden Semester* angeboten, so gibt es keine Wiederholungsprüfung. Stattdessen können Sie ggf. an der regulären Klausur des Moduls teilnehmen. Es gilt der aktuelle Stoff bzw. die aktuellen formalen Bedingungen. Es ist möglich, dass der Stoff sich dann substanzial ändert; meist halten sich die Änderungen (im Stoff) aber in Rahmen.

- Sprechen Sie die *Moduldozentis* an für Details zur Prüfung (bzw. lesen Sie vorab auf jeweiligen Modulseite in Moodle nach).
- Manchmal fragen Studenten nach einer *Empfehlung*, ob es besser ist, eine Prüfung zu verschieben, wenn man sich nicht ausreichend vorbereiten konnte. Es ist schwer, eine Empfehlung pauschal abzugeben, es kommt auf den Einzelfall an. Grundsätzlich rate ich aber dazu, Prüfungen nicht zu verschieben: Schließlich könnte in einem folgenden Semester wieder ein unvorhergesehenes Problem auftreten.
- Bei Fragen zum *Prüfungsrecht* sprechen Sie bitte die *Studienberatung* an.

1.4 Prüfungsstoff

Prüfungsrelevanter Stoff ist alles, was im Unterricht behandelt wurde, sofern es nicht explizit (und schriftlich) als “nicht prüfungsrelevant” gekennzeichnet ist.

1.5 Bearbeitungshinweise

1.5.0.1 Allgemeine Hinweise

- Verwenden Sie *Standardwerte* (defaults) der R-Funktionen.
- Findet sich in einer Auswahlliste möglicher Antworten nicht die exakte Lösung, wählen Sie die *am besten passende*.
- *Treffen Sie Annahmen*, wo nötig.
- Die Prüfung besteht auch aus *Single-* bzw. *Multiple-Choice-* (MC-)Aufgaben mit mehreren Antwortoptionen.
- Bei Multiple-Choice-Aufgaben (MC-Aufgaben) ist zumeist genau *eine* Antwortoption auszuwählen aus vier oder fünf Antwortoptionen.
- Im Zweifel ist eine Aussage auf den Stoff, so wie im *Unterricht* behandelt, zu beziehen.
- 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.
- 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.
- Falls Sie bei einer Aufgabe mehrere Antworten finden, aber nur nach einer gefragt ist, geben Sie nur eine an.
- Falls mehrere (widersprüchliche) Antworten gegeben wurden, wird im Zweifel die *erst genannte* gewertet.
- Die Aufgabenstellung in einer Moodle-Prüfung wird erst sichtbar, wenn Sie den *Prüfungsbedingungen* zugestimmt haben und die Prüfungszeit begonnen hat.

1.5.1 Aufgaben zur Datenanalyse

- Je nach Spracheinstellung in Moodle kann es sein, dass Sie als *Dezimaltrennzeichen* ein Komma oder einen Punkt verwenden müssen. Moodle weist Sie, wenn Sie die Aufgabe verlassen, darauf hin, falls eine Zahl nicht als Zahl erkannt wurde.
- Runden Sie ggf. auf *eine Dezimalstelle*.
- Geben Sie keine Prozentzahlen an, sondern *Anteile* (also nicht “50%”, sondern “0.5” bzw. “0,5”).
- Bei Aufgaben, die eine Zahl als Antwort verlangen, ist *nur eine Zahl anzugeben* (nicht etwa Buchstaben).
- Alle Berechnungen, die Zufallszahlen beinhalten, sollen mit fixierten Startwert der Zufallszahlen durchgeführt werden. Es ist die Zahl 42 zu verwenden.
- Wie auch bei den übrigen Hinweisen gelten diese Maßgaben nur soweit in der Prüfung nicht *explizit andere Hinweise* gegeben wurden.
- Wenn *Stichproben simuliert* werden sollen, ziehen Sie 10^3 Zufallsstichproben.
- In einigen Aufgaben kann verlangt sein, dass Sie einen *bestimmten Datensatz in R importieren* sollen. In diesem Fall wird vorausgesetzt, dass Ihnen [diese Bezugsquelle](#) von Datensätzen bekannt ist und dass Sie wissen, wie man einen Datensatz in R importiert.
- Achten Sie darauf, R und R-Pakete sowie R-Studio in *aktueller Version* zu verwenden. Das Verwenden älterer Versionen kann (in seltenen Fällen) zu abweichenden Lösungen führen. Im Zweifel beziehen sich alle Aufgaben auf die jweils aktuellste Version der verwendeten Software.

Speziell zur Bayes-Statistik:

- Verwenden Sie Methoden der *Bayes-Statistik* für inferenzstatistische Analysen.
- Bei Aufgaben zur “*Bayes-Box*” (Erstellung einer Gitterwert-Tabelle) gelten folgende Maßgaben:
 - Handelt es sich um Parameter mit einem begrenzten Wertebereich (wie etwa *Anteile*), so ist der ganze Wertebereich zu modellieren. Es sind 100 verschiedene Parameterwerte zu berechnen.
 - Handelt es sich um Parameter X mit einem unbegrenzten Wertebereich (wie *normalverteilte* Variablen), so ist der Wertebereich $X - 2\sigma \leq X \leq X + 2\sigma$ zu simulieren.
- Nutzen Sie die Software *Stan* in Form des R-Paket `rstanarm` für Regressionsmodelle auf Basis der Bayes-Methode.

1.6 Teilnahmebedingungen

- Die Prüfung ist selbständig, also alleine nur durch Sie, *ohne Hilfe Dritter*, zu absolvieren.
- Die *Bearbeitungszeiten* der Prüfung sind einzuhalten.
- Es dürfen nur die explizit als zulässig gekennzeichneten *Hilfsmittel* verwendet werden.

- Die *zulässigen* Hilfsmittel sind: Notizzettel, Stifte, Taschenrechner, Vorlesungsfolien, Skripte, eigene Notizen, Bücher sowie Quellen aus dem Internet.
- Die Übernahme von Inhalten Dritter muss also solche gekennzeichnet (*zitiert*) werden.
- Eine wörtliche Übernahme (“*copy-paste*”) von Inhalten Dritter (etwa aus einer Quelle aus dem Internet) ist unzulässig.
- Bei *technischen Problemen* ist sofort der Prüfer bzw. die Prüferin zu verständigen und das technische Problem zu dokumentieren. Aus der Dokumentation muss der Fehler erkennbar sein.
- Es ist *untersagt*, die *Prüfung* bzw. Teile daraus (z.B. Prüfungsfragen) zu speichern oder *weiterzugeben*.
- Im Übrigen gelten die allgemeinen *Prüfungsregeln*.
- Es darf *nicht* mit anderen Personen, insbesondere nicht Prüflingen, *kommuniziert* werden während der Prüfung. Dies gilt auch für den Fall von (vermeintlichen) technischen Problemen. Kontaktieren Sie den Prüfer, wenn Sie meinen, es liege ein technisches Problem vor.
- Das Nichtbeachten der Regeln kann zu Notenabzug oder *Nichtbestehen* führen.

1.7 Organisatorische Hinweise

- Etwaige weitere *Stoffeingrenzungen* werden schriftlich bekannt gemacht (auf der Modulseite). Besondere Schwerpunkte gibt es nicht.
- Soweit bestimmte Inhalte nicht explizit ausgeschlossen sind, sind alle Inhalte, die im Rahmen des Moduls bearbeitet wurden, *prüfungsrelevant*.
- Im Übrigen gelten die Hinweise der *offiziellen Regularien* wie SPO, auf dem Modulsteckbrief und der APO. Bitte kontaktieren Sie die Studienberatung für formale oder rechtliche Fragen.
- Während der Prüfung werden nur Fragen beantwortet, die für die Bearbeitung zwingend nötig sind (etwa bei technischen Problemen).
- Es werden keine Fragen der Art “Ist diese Aufgabe klar formuliert?” beantwortet während der Prüfung. Sollten Sie der Meinung sein, eine Frage ist unklar formuliert oder fehlerhaft, so notieren Sie dies bitte (z.B. im Kommentarfeld der Prüfung=). Der Prüfer untersucht im Nachgang die Angelegenheit. Stellt sich eine Frage als fehlerhaft oder unklar formuliert heraus, so wird sie von der Beurteilung herausgenommen.
- Eine Teilnahme an der Prüfung ist nur möglich, wenn Sie den Teilnahmebedingungen der Prüfung zustimmen.
- Die Aufgabenstellung der Prüfung wird nur während des Prüfungszeitraumes angezeigt.
- Beachten Sie eine etwaige Gruppenteilung (zu welcher Gruppe Sie zugewiesen sind).
- Beachten Sie die exakte Prüfungszeit (Beginn, Ende).
- Prüfungszeitraum, Aufgabenstellung und sonstige Materialien können variieren zwischen den Prüflingen etwa aufgrund von Gruppeneinteilungen oder Nachteilsausgleich.
- Die zusätzliche Bearbeitungszeit bei Studenten mit Nachteilsausgleich ist in der Aufgabenstellung bzw. der Prüfung in Moodle hinterlegt. Die Zeit wird automatisch um den jeweiligen Faktor erhöht.

1.8 Open-Book-Prüfungen

Einige Prüfungen werden als “Take-home-Prüfung” im “Open-Book-Format” geschrieben. Was bedeutet dies?

“Take-home-Prüfung”: Sie bearbeiten die Prüfung in Ihrem privaten Umgebung in Moodle oder in Räumlichkeiten der Hochschule. Eine Überwachung per Kamera findet nicht statt.

“Open-Book-Prüfung”: Sie dürfen Hilfsmittel wie Bücher und Folien während der Prüfung nutzen.

Ihre Prüferin/Ihr Prüfer informiert Sie über das Format Ihrer Prüfung.

1.9 Zeitrahmen der Prüfung

Die Prüfung beginnt und endet zu einem festen Zeitpunkt. Sie sind selber verantwortlich, die Prüfung zur korrekten Zeit zu beginnen und zu beenden (einzureichen). Verspätete Abgaben werden u.U. als nicht bestanden gewertet. Die Dauer der Prüfung wird Ihnen von Ihrer Prüferin bzw. Ihrem Prüfer bekannt gegeben.

1.10 Technische und organisatorische Anforderungen einer Open-Book-Prüfung

Um an einer Open-Book-Prüfung teilzunehmen, benötigen Sie IT-Ausstattung sowie Räumlichkeiten. An IT-Ausstattung benötigen Sie einen Computer mit Internetanschluss; ein Smartphone reicht nicht aus. Nutzen Sie Ihr eigenes Gerät (Computer) für die Prüfung; die Hochschule stellt Ihnen keinen Computer zur Verfügung. Sie benötigen keine Webcam und kein Mikrofon. Ein Tablett oder Smartphone reicht nicht für die Prüfung. An Software benötigen Sie Zugang zu Ihrem Moodle-Konto, was einen aktuellen Internet-Browser voraussetzt. Zu den organisatorischen Anforderungen gehören ein Raum, in dem Sie die Prüfung ungestört bearbeiten können sowie ein Internetanschluss zum Bearbeiten der Klausur in Moodle. Bitte benutzen Sie während der Prüfung nicht den Zurück-Button in Ihrem Browser, wenn Sie zu einer vorherigen Frage zurückgehen wollen. Nutzen Sie die in der Prüfung zur Verfügung gestellten Funktionen/Buttons dafür.

1.11 Technische Probleme während der Prüfung

Im Falle eines technischen Problems auf Seiten der *Prüfungsinfrastruktur* ist sofort der Prüfer zu informieren. Ein Beispiel für so ein Problem wäre etwa der Ausfall von Moodle. Der technische Fehler ist zu dokumentieren (z.B. Screenshot) und die Dokumentation ist einzureichen. Bitte beachten Sie, dass der Prüfer bzw. die Hochschule keine Gewähr übernimmt für Probleme mit Ihrer eigenen Ausstattung.

1.12 Prüfungsrecht

Für die Open-Book-Prüfung gilt die aktuelle Prüfungsordnung; die Open-Book-Prüfung fällt *nicht* unter die BayFEV.

1.13 Bitte formulieren Sie Beanstandungen nachvollziehbar

Falls Sie einen Fehler in einer Aufgabenstellung finden (die der Prüfer zu bestanden hat): Freuen Sie sich! In diesem Fall wird zu Ihren Gunsten entschieden.

Damit ich Ihre Beanstandung prüfen kann, ist es nötig, dass ich Ihre *Beanstandung* nachprüfen kann: *Zeigen* Sie mir meinen Fehler. Es reicht *nicht* zu sagen “es hat bei mir nicht funktioniert”. Wenn Sie etwa der Meinung sind, dass es die Variable “year” im Datensatz “penguins” nicht gebe, dann schicken Sie mir den R-Code, der an ansprechender Stelle einen Fehler aufwirft (aber *ansonsten lauffähig* ist). Ein Screenshot kann in einigen Situationen helfen, wenn aber nur ein Teil der Syntax zu sehen ist, ist er nicht ausreichend: Wenn der Befehl `data(penguins)` nicht funktioniert, so ist zu prüfen, ob Sie vorab mit `library(palmerpenguins)` das relevante Paket gestartet haben. Andernfalls kann `data(penguins)` nicht funktionieren und der Fehler läge damit bei Ihnen.

[Hier](#) finden Sie Hinweise für einfache, reproduzierbare Beispiele (ERBies); vgl. Sauer (2019), Kap. 3.8.2 (S. 33).

Die gleiche Messlatte lege ich an mich an: Ich stelle eine Musterlösung (bei der Einsichtnahme) zur Verfügung, die reproduzierbar die Lösung aufzeigt. Sprich: Wenn der R-Code bei mir durchläuft, so wird er auch bei Ihnen durchlaufen.

1.14 Datenschutz

Persönliche Daten werden an eine Stellen übermittelt: Moodle (über bzw. in Ihre Konto). Es findet keine Überwachung statt, weder kamaragestützt, akustisch oder softwaregestützt.

1.15 Plagiatskontrolle

Ihre Prüfungsarbeiten können auf Plagiate hin untersucht werden. Dabei kommen auch automatisierte Verfahren zum Einsatz. Ihre Arbeiten werden dabei nicht online gestellt und auch nicht Dritten zugänglich gemacht. Alle Prüfungen finden auf Rechnern statt, zu denen nur die Prüfer/innen Zugang habe. Es werden keine persönlichen Daten (von Ihnen) weitergegeben.

Bitte beachten: Angenommen in den Projektarbeiten von Studenti A und B werden (substanzielle) Überlappungen gefunden. In dem Fall ist davon auszugehen, dass beide Studentis

getäuscht haben: eine/r hat abgeschrieben, der/die andere hat die eigene Arbeit dafür bereitgestellt. Daher wird in diesem Fall u.U. bei beiden Studenten der Plagiatsfall festgestellt und geahndet (z.B. mit “nicht bestanden” bewertet). Die genauen Konsequenzen legt die Prüfungskommission im Einzelfall fest.

Lassen Sie es auf keinen Fall soweit kommen: Schreiben Sie nicht ab und lassen Sie niemanden von Ihrer Arbeit abschreiben.

Eine faire Prüfung heißt: Gleiche Chancen für alle, und gute Leistung soll belohnt werden, Täuschung nicht.

1.16 Typische Fehler in der Prüfung

- *Rechtschreibfehler* Manchmal muss man genau hinschauen, und leicht vertippt man sich: So heißt der Datensatz vielleicht `tips` und die Spalte, um die es Ihnen geht `tip` (oder war es umgekehrt?). Oder die Spalte heißt `bill_length` aber Sie schreiben `bill_lenght`.
- *Datensatz nicht richtig importiert* Ob ein Datensatz richtig importiert ist, erkennen Sie daran, ob er im Reiter “Environment” angezeigt wird. Außerdem können Sie dort den Datensatz anklicken, um zu einer Tabellenansicht des Datensatzes zu gelangen. Dort können Sie erkennen, ob z.B. die Anzahl der Spalten korrekt ist (und nicht etwa nur eine) oder z.B. ob die Spaltennamen korrekt sind.
- *data(datensatz) ohne vorher das zugehörige R-Paket gestartet zu haben:* Mit `data(datensatz)` können Sie den Datensatz `datensatz` nur dann verfügbar machen, wenn das Paket, in dem `datensatz` “wohnt”, mit `library(paketname)` gestartet worden ist. So “wohnt” z.B. `penguins` im Datensatz `palmerpenguins`. [Hier](#) finden Sie eine Übung (und weitere Erklärung) zum Importieren von Daten in R am Beispiel des Datensatzes `penguins`.
- Verwenden einer *Funktion, ohne das zugehörige R-Paket* vorab gestartet zu haben.
- Das Laden zu vieler R-Pakete, die gar nicht benötigt werden, mit dem Ergebnis, dass es mehrere Funktionen des gleichen Namens gibt (z.B. `filter()`). Das führt dann zu Verwirrung, da dann z.B. nicht die Funktion `filter` aus `tidyverse` (`dplyr`) verwendet wird, wie Sie annehmen, sondern eine Funktion gleichen Namens aus einem anderen Paket, das Sie auch gestartet haben. Tipp: Starten Sie nur die Pakete, die Sie für die Aufgabe benötigen. Zumeist sind das immer die gleichen wenigen Pakete.

1.17 Hinweise zu Scheinmängeln

Immer wieder kommt es vor, dass Studierende Beanstandungen zu einer Prüfung vorbringen. Teilweise sind diese gerechtfertigt, teilweise nicht. Im Folgenden sehen Sie eine Auswahl an *nicht* gerechtfertigten Beanstandungen, also nur scheinbaren Mängeln, keine wirklichen Mängel, in einer Prüfung.

- “Das zu wählende Vorgehen war nicht 100% klar” – Wenn Sie der Meinung sind, dass das zu wählende Vorgehen (zum Lösen der Aufgabe) nicht komplett klar ist, treffen Sie Annahmen und weisen Sie darauf hin, dass Sie Annahmen getroffen haben. Zum anderen halten Sie sich an das Vorgehen aus dem Unterricht (bzw. den Unterlagen und der Literatur, die im Unterricht verwendet wurde). Eine andere Situation läge vor, wenn die Aufgabe nicht lösbar ist ohne weitere Angaben lösbar ist (“Ist ein Effekt bei $n=100$ signifikant?”). Im Falle einer nicht lösbarer Aufgabe liegt Fehler beim Prüfer.
- “Ich sollte einen Punkt (ein Komma) als Dezimaltrennzeichen verwenden, aber Moodle hat ein Komma (einen Punkt) verlangt!” – Je nach Spracheinstellung in Moodle kann es sein, dass Moodle nur einen Punkt als Dezimaltrennzeichen bzw. ein Komma als Dezimaltrennzeichen verwendet. Moodle weist Sie aber darauf hin, wenn eine Zahl nicht als Zahl erkannt wird, und zwar wenn Sie zur nächsten Aufgabe geben. Sie können also ohne Probleme den Fehler korrigieren. Darüber hinaus ist bei den Prüfungshinweisen vorab auf diesen Punkt verwiesen.

2 Inferenz



Abbildung 2.1: Bayes:Start!

2.1 Lernsteuerung

2.1.1 Lernziele

Nach Absolvieren des jeweiligen Kapitels sollen folgende Lernziele erreicht sein.

Sie können ...

- die Definition von Inferenzstatistik sowie Beispiele für inferenzstatistische Fragestellungen nennen
- zentrale Begriffe nennen und in Grundzügen erklären
- den Nutzen von Inferenzstatistik nennen
- erläutern, in welchem Zusammenhang Ungewissheit zur Inferenzstatistik steht
- auch anhand von Beispielen erklären, was ein statistisches Modell ist
- die Grundkonzepte der Regression angeben
- Unterschiede zwischen klassischer und Bayes-Inferenz benennen
- Vor- und Nachteile der klassischen vs. Bayes-Inferenz diskutieren
- Die grundlegende Herangehensweise zur Berechnung des p-Werts informell erklären können

2.1.2 Begleitvideos

- [Video zur Inferenz, Teil 1](#)
- [Video zur Inferenz, Teil 2](#)

2.2 Wozu ist Statistik überhaupt da?

Ja, diese Frage haben Sie sich auch schon mal gestellt?

Abb. Abbildung 2.2 gibt einen Überblick über die Ziele der Statistik.

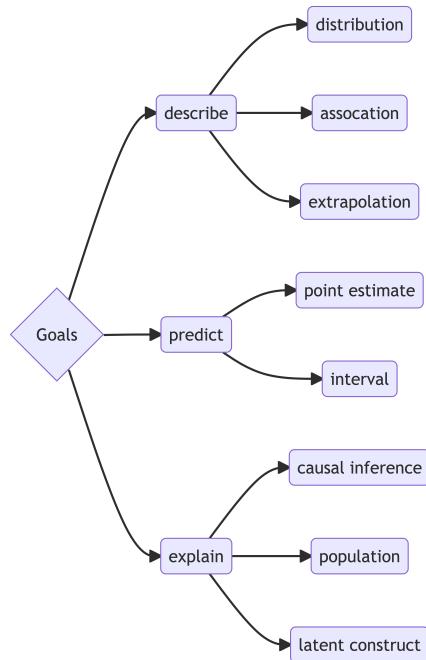


Abbildung 2.2: A taxonomy of statistical goals

i Hinweis

Ziele existieren nicht “in echt” in der Welt. Wir denken sie uns aus. Ziele haben also keine ontologische Wirklichkeit, sie sind epistemologische Dinge (existieren nur in unserem Kopf). Das heißt, dass man sich nach Beliebtem Ziele ausdenken kann. Allerdings hilft es, wenn man andere Menschen vom Nutzen der eigenen Ideen überzeugen kann.

2.3 Was ist Inferenz?

2.3.1 Inferenz als Generalisieren

Statistische Inferenz sieht sich drei “Herausforderungen” gegenüber, laut Gelman, Hill, und Vehtari (2021), Kap. 1.1. Diese betreffen das Schließen (oder Generalisieren) vom Einzelfall auf das Allgemeine:

1. Von der Stichprobe aus die Grundgesamtheit (Population)
2. Von der Experimental- auf die Kontrollgruppe (Kausalinferenz)
3. Von einem Messwert auf das zugrundeliegende Konstrukt

In diesem Kurs beschäftigen wir uns mit den ersten beiden Herausforderungen.

! Wichtig

Statistische Inferenz hat zum Ziel, vom Teil aufs Ganze zu schließen, bzw. vom Konkreten auf das Abstrakte.

2.4 Stichprobe vs. Population

Nehmen wir an, wir möchten herausfinden, wie groß der Anteil der R-Fans an der Population der Studierenden ist. Den Anteil der F-Fans bezeichnen wir der Einfachheit halber hier mit A^1 .

Das *Grundproblem der Inferenzstatistik* ist, dass wir an Aussagen zur Grundgesamtheit interessiert sind, aber nur eine Stichprobe, also einen Ausschnitt oder eine Teilmenge der Grundgesamtheit vorliegen haben.

Wir müssen also den Anteil der R-Fans auf Basis des Anteils in der Stichprobe für die Grundgesamtheit schließen: Wir verallgemeinern oder generalisieren von der Stichprobe auf die Grundgesamtheit, s. Abb. Abbildung 2.3.

Häufig ist das praktische Vorgehen recht simpel: Ah, in unserer Stichprobe sind 42% R-Fans!². Man schreibt: $p = 0.42$ (p wie proportion). Die Stichprobe sei repräsentativ für die Grundgesamtheit aller Studierender. Messerscharf schließen wir: In der Grundgesamtheit ist der Anteil der R-Fans auch 42%, $\pi = 0.42$.

2.4.1 Deskriptiv- vs. Inferenzstatistik

Statistik gibt es in zwei Geschmacksrichtungen, könnte man sagen: Deskriptiv- und Inferenzstatistik, s. Abb. Abbildung 2.4. Einteilungen in Schubladen existieren nicht auf der Welt, sondern in unserem Kopf: Sie besitzen keine ontologische Realität, sondern eine epistemologische. Sie sind frei, sich andere Einteilungen der Statistik auszudenken. Es hilft allerdings, wenn man andere Menschen vom Wert seiner Idee überzeugen kann.

Deskriktivstatistik fasst Stichprobenmerkmale zu Kennzahlen (Statistiken) zusammen.

Inferenzstatistik schließt von Statistiken auf Parameter (Kennzahlen von Grundgesamtheiten).

Schließen Sie die Augen und zeichnen Sie obiges Diagramm!

¹Meistens Manchmal darf man bei der Statistik nicht nach einem tieferen Sinn suchen. Ist Statistik eine Art moderne Kunst?

²Mancheiner hätte mit mehr gerechnet

Population	Sample
AAAABAAAAAA	AAAABAAAAAA
ABAABABABAB	ABAABABAB
AABBAABAA	AABBAABAA
BAABAABABA	BAABAABAA
BABBAAABA	BABBAAABA
ABABAABAA	ABABAABAA
BBBABBABA	BBBABA
AAAABBAAABA	AAAABBAAABA
ABABBABA	ABABBABA
BBAAAABBA	BBAAAABBA

(a) Population

(b) Sample

Abbildung 2.3: Population vs. sample (Image credit: Karsten Luebke)

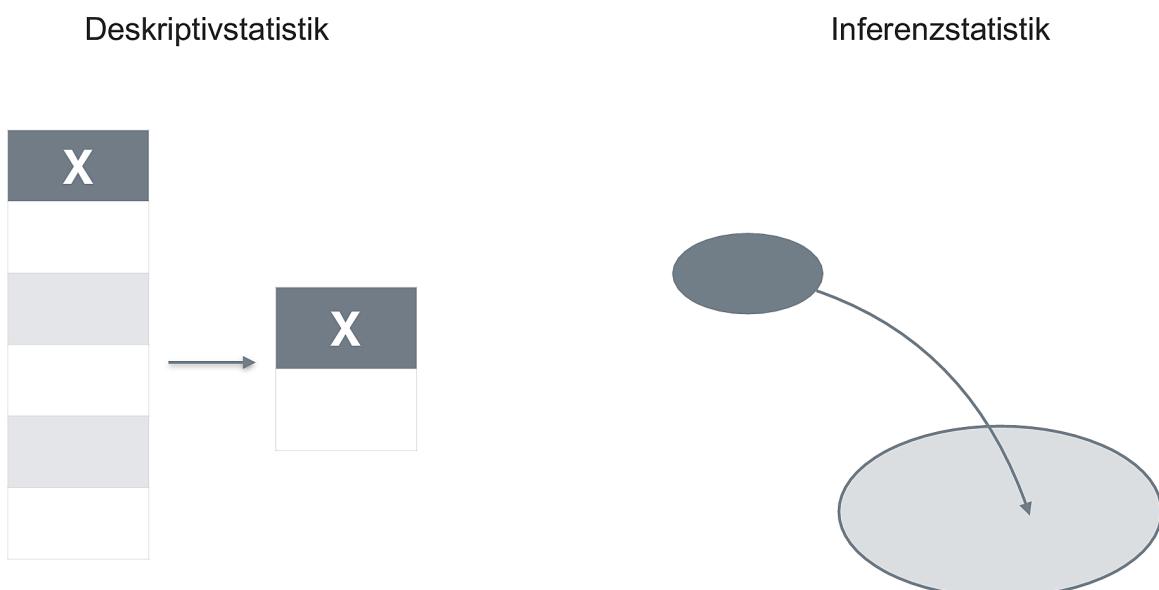


Abbildung 2.4: Deskriptiv- vs. Inferenzstatistik

2.4.2 Wozu ist die Inferenstatistik gut?

i Hinweis

Inferenz bedeutet Schließen; auf Basis von vorliegenden Wissen wird neues Wissen generiert.

Inferenzstatistik ist ein Verfahren, das mathematische Modelle (oft aus der Stochastik) verwendet, um ausgehend von einer bestimmten Datenlage, die eine Stichprobe einer Grundgesamtheit darstellt, allgemeine Schlüsse zu ziehen.

Heute Nacht vor dem Schlafen wiederholen Sie die Definition. Üben Sie jetzt schon mal.

2.4.3 Deskriptiv- und Inferenzstatistik gehen Hand in Hand

Für jede beliebige Statistik (Kennzahl von Stichprobendaten) kann man die Methoden der Inferenzstatistik verwenden, s. Tabelle Tabelle 2.1.

Tabelle 2.1: Bezeichnungen für Kennwerte

Kennwert	Stichprobe	Grundgesamtheit
Mittelwert	\bar{X}	μ
Streuung	sd	σ
Anteil	p	π
Korrelation	r	ρ
Regression	b	β

Für Statistiken (Daten einer Stichprobe) verwendet man *lateinische* Buchstaben; für Parameter (Population) verwendet man *griechische* Buchstaben.

Geben Sie die griechischen Buchstaben für typische Statistiken an!

2.4.4 Schätzen von Parametern einer Grundgesamtheit

Meist begnügt man sich beim Analysieren von Daten nicht mit Aussagen für eine Stichprobe, sondern will auf eine Grundgesamtheit verallgemeinern.

Leider sind die Parameter einer Grundgesamtheit zumeist unbekannt, daher muss man sich mit *Schätzungen* begnügen.

Schätzwerte werden mit einem “Dach” über dem Kennwert gekennzeichnet, z.B.

Kennwert	Stichprobe	Grundgesamtheit	Schätzwert
Mittelwert	\bar{X}	μ	$\hat{\mu}$

Kennwert	Stichprobe	Grundgesamtheit	Schätzwert
Streuung	sd	σ	$\hat{\sigma}$
Anteil	p	π	$\hat{\pi}$
Korrelation	r	ρ	$\hat{\rho}$
Regression	b	β	$\hat{\beta}$

2.4.5 Beispiele für inferenzstatistische Fragestellungen

Sie testen zwei Varianten Ihres Webshops (V1 und V2), die sich im Farbschema unterscheiden und ansonsten identisch sind: Hat das Farbschema einen Einfluss auf den Umsatz?

- Dazu vergleichen Sie den mittleren Umsatz pro Tag von V1 vs. V2, \bar{X}_{V1} und \bar{X}_{V2} .
- Die Mittelwerte unterscheiden sich etwas, $\bar{X}_{V1} > \bar{X}_{V2}$
- Sind diese Unterschiede “zufällig” oder “substanziel”? Gilt also $\mu_{V1} > \mu_{V2}$ oder gilt $\mu_{V1} \leq \mu_{V2}$?
- Wie groß ist die Wahrscheinlichkeit³ $Pr(\mu_{V1} > \mu_{V2})$?

Predictive Maintenance ist ein Anwendungsfeld inferenzstatistischer Modellierung. Lesen Sie dazu S. 3 [dieses Berichts!](#)

2.5 Modellieren

2.5.1 Modellieren als Grundraster des Erkennens

In der Wissenschaft - wie auch oft in der Technik, Wirtschaft oder im Alltag - betrachtet man einen Teil der Welt näher, meist mit dem Ziel, eine Entscheidung zu treffen, was man tun wird oder mit dem Ziel, etwas zu lernen.

Nun ist die Welt ein weites Feld. Jedes Detail zu berücksichtigen ist nicht möglich. Wir müssen die Sache vereinfachen: Alle Informationen ausblenden, die nicht zwingend nötig sind. Aber gleichzeitig die Strukturelemente der wirklichen Welt, die für unsere Fragestellung zentral ist, beibehalten.

Dieses Tun nennt man *Modellieren*: Man erstellt sich ein Modell.

! Wichtig

Ein Modell ist ein vereinfachtes Abbild der Wirklichkeit.

Auf die Statistik bezogen heißt das, dass man einen Datensatz zusammenfasst, dass man das Wesentliche erkennt. Was ist das “Wesentliche”? Meist interessiert man sich für

³oft mit Pr oder p abgekürzt, für *probability*

die Ursachen eines Phänomens? Etwa: "Wie kommt es bloß, dass ich ohne zu lernen die Klausur so gut bestanden habe?"⁴ Noch allgemeiner ist vom häufig am Zusammenhang von X und Y interessiert, s. Abbildung 2.5, linker Teil, die ein Sinnbild eines statistischen Modells wiedergibt.

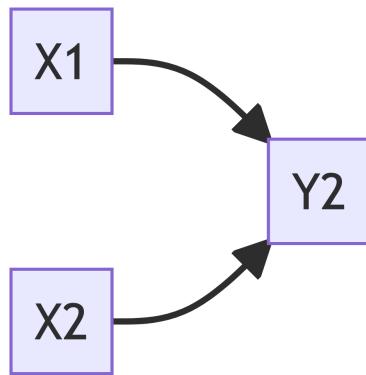


Abbildung 2.5: oben: Sinnbild eines statistischen Modells; unten: Sinnbild eines statistischen Modells, mit zwei Inputvariablen (Ursachen)

Das Diagramm hat Sie nicht so vom Hocker? Okay, ein statistisches Modell kann natürlich komplexer sein, z.B. wie in Abb. Abbildung 2.5, rechter Teil, dargestellt.

Es hört sich zugspitzt an, aber eigentlich ist fast alles Modellieren: Wenn man den Anteil der R-Fans in einer Gruppe Studierender ausrechnet, macht man sich ein Modell: man vereinfacht diesen Ausschnitt der Wirklichkeit anhand einer statistischen Kennzahl, die das forschungsleitende Interesse zusammenfasst.

2.5.2 Vertiefung

Lesen Sie die Einführung zum Thema Modellieren bei Poldrack (2022) (Kap. 5.1).

⁴Das ist natürlich nur ein fiktives, komplett unrealistisches Beispiel, das auch unklaren Ursachen den Weg auf diese Seite gefunden hat.

i Hinweis

Nutzen Sie die Übersetzungsfunktion Ihres Browsers, wenn Sie einen englischen Text lieber auf Deutsch lesen wollen. Oder einen deutschen lieber auf Englisch.

2.6 Regression

Einflussreiche Leute schwören auf die Regressionsanalyse (Abbildung 2.6).



Abbildung 2.6: One regression

2.6.1 Regression zum Modellieren

Die Regression ist eine Art Schweizer Taschenmessen: Für vieles gut einsetzbar.

Anstelle von vielen verschiedenen Verfahren des statistischen Modellierens kann man (fast) immer die Regression verwenden. Das ist nicht nur einfacher, sondern auch schöner. Wir werden im Folgenden stets die Regression zum Modellieren verwenden.

Dann wenden wir die Methoden der Inferenz auf die Kennzahlen der Regression an.

i Hinweis

Regression + Inferenz =

Alternativ zur Regression könnte man sich in den Wald der statistischen Verfahren begeben, wie hier von der Uni Münster als Ausschnitt (!) aufgeführt.

Auf dieser Basis kann man meditieren, welches statistischen Verfahren man für eine bestimmte Fragestellung verwenden sollte, s. Abb. Abbildung 2.7.

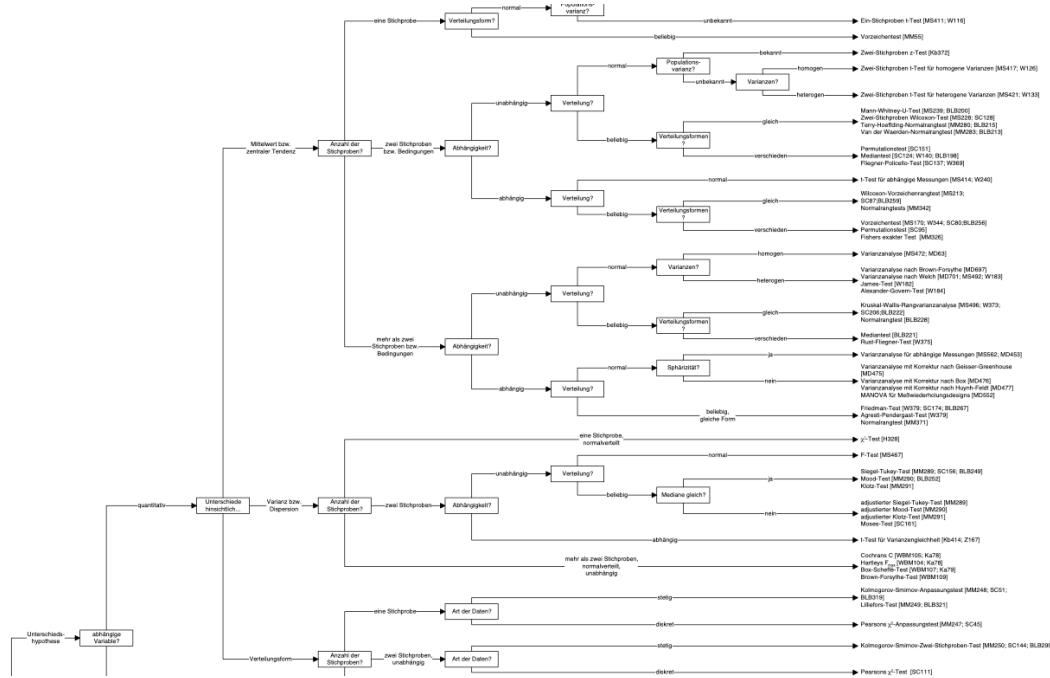


Abbildung 2.7: Wähle deine Statistik mit Bedacht

2.6.2 Viele statistische Verfahren sind Spezialfälle der Regression

Wie Jonas Kristoffer Lindeløv uns erklärt, sind viele statistische Verfahren, wie der sog. t-Test Spezialfälle der Regression, s. Abb. Abbildung 2.8.

2.6.3 In voller Pracht

Hier ist die Regressionsgleichung in voller Pracht; Abb. Abbildung 2.9.

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \epsilon$$

Anhand der Gleichung erkennt man auch, warum man von einem *linearen Modell* spricht: Y wird als gewichteter Mittelwert mehrerer Summanden berechnet. Dabei wird X nicht mit "fortgeschrittenen" Transformationen wie Quadrieren oder Exponentiellieren beglückt, sondern nur mit den Regressionsgewichten multipliziert.

Common statistical tests are linear models

Last updated: 02 April, 2019

See worked examples and more details at the accompanying notebook: <https://lindeloev.github.io/tests-as-linear>

Common name	Built-in function in R	Equivalent linear model in R	Exact?	The linear model in words	Icon	
Simple regression: $\text{Im}(y \sim 1 + x)$	y is independent of x P: One-sample t-test N: Wilcoxon signed-rank	<code>t.test(y)</code> <code>wilcox.test(y)</code>	$\text{Im}(y \sim 1)$ $\text{Im}(\text{signed_rank}(y) \sim 1)$	✓ for N ≥ 14	One number (intercept, i.e., the mean) predicts y. - (Same, but it predicts the signed rank of y.)	
	P: Paired-sample t-test N: Wilcoxon matched pairs	<code>t.test(y1, y2, paired=TRUE)</code> <code>wilcox.test(y1, y2, paired=TRUE)</code>	$\text{Im}(y_2 - y_1 \sim 1)$ $\text{Im}(\text{signed_rank}(y_2 - y_1) \sim 1)$	✓ for N ≥ 14	One intercept predicts the pairwise $y_2 - y_1$ differences. - (Same, but it predicts the signed rank of $y_2 - y_1$.)	
y ~ continuous x P: Pearson correlation N: Spearman correlation		<code>cor.test(x, y, method='Pearson')</code> <code>cor.test(x, y, method='Spearman')</code>	$\text{Im}(y \sim 1 + x)$ $\text{Im}(\text{rank}(y) \sim 1 + \text{rank}(x))$	✓ for N ≥ 10	One intercept plus x multiplied by a number (slope) predicts y. - (Same, but with ranked x and y)	
	P: Two-sample t-test P: Welch's t-test N: Mann-Whitney U	<code>t.test(y1, y2, var.equal=TRUE)</code> <code>t.test(y1, y2, var.equal=FALSE)</code> <code>wilcox.test(y1, y2)</code>	$\text{Im}(y \sim 1 + G_2)^*$ $\text{glm}(y \sim 1 + G_2, \text{weights}=\dots)^*$ $\text{Im}(\text{signed_rank}(y) \sim 1 + G_2)^*$	✓ for N ≥ 11	An intercept for group 1 (plus a difference if group 2) predicts y. - (Same, but with one variance per group instead of one common.) - (Same, but it predicts the signed rank of y.)	
Multiple regression: $\text{Im}(y \sim 1 + x_1 + x_2 + \dots)$	P: One-way ANOVA N: Kruskal-Wallis	<code>aov(y ~ group)</code> <code>kruskal.test(y ~ group)</code>	$\text{Im}(y \sim 1 + G_2 + G_3 + \dots + G_n)^*$ $\text{Im}(\text{rank}(y) \sim 1 + G_2 + G_3 + \dots + G_n)^*$	✓ for N ≥ 11	An intercept for group 1 (plus a difference if group ≠ 1) predicts y. - (Same, but it predicts the rank of y.)	
	P: One-way ANCOVA	<code>aov(y ~ group + x)</code>	$\text{Im}(y \sim 1 + G_2 + G_3 + \dots + G_n + x)^*$	✓	- (Same, but plus a slope on x.) Note: this is discrete AND continuous. ANCOVAs are ANOVAs with a continuous x.	
	P: Two-way ANOVA	<code>aov(y ~ group * sex)</code>	$\text{Im}(y \sim 1 + G_2 + G_3 + \dots + G_n + S_1 + S_2 + \dots + S_k + G_2 * S_1 + G_2 * S_2 + \dots + G_2 * S_k)$	✓	Interaction term: changing sex changes the y ~ group parameters. Similarly for S_{j+k} for sex. The first line (with G) is main effect of group, the second (with S) for sex and the third is the group × sex interaction. For two levels (e.g. male/female), line 2 would just be S_1 ; and line 3 would be S_1 multiplied with each G.	[Coming]
Counts ~ discrete x N: Chi-square test		<code>chisq.test(groupXsex_table)</code>	Equivalent log-linear model <code>glm(y ~ 1 + G_2 + G_3 + \dots + G_n + S_1 + S_2 + \dots + S_k + G_2 * S_1 + G_2 * S_2 + \dots + G_2 * S_k, family=...)^*</code>	✓	Interaction term: (Same as Two-way ANOVA.) Note: Run glm using the following arguments: <code>glm(model, family='poisson()')</code> . As linear-model, the Chi-square test is <code>glm(y ~ log(G) + log(S) + log(G*S), family=poisson())</code> where α and β are proportions. See more info in the accompanying notebook.	Same as Two-way ANOVA
	N: Goodness of fit	<code>chisq.test(y)</code>	<code>glm(y ~ 1 + G_2 + G_3 + \dots + G_n, family=...)^*</code>	✓	(Same as One-way ANOVA and see Chi-Square note.)	1W-ANOVA

List of common parametric (P) non-parametric (N) tests and equivalent linear models. The notation $y \sim 1 + x$ is R shorthand for $y = 1 + b + a \cdot x$ which most of us learned in school. Models in similar colors are highly similar, but really, notice how similar they all are across colors! For non-parametric models, the linear models are reasonable approximations for non-small sample sizes (see "Exact" column and click links to see simulations). Other less accurate approximations exist, e.g., Wilcoxon for the sign test and Goodness-of-fit for the binomial test. The signed rank function is `signed_rank = function(x) sign(x) * rank(abs(x))`. The variables G and S are "dummy coded" indicator variables (either 0 or 1) exploiting the fact that when $\Delta x = 1$ between categories the difference equals the slope. Subscripts (e.g., G_2 or y_1) indicate different columns in data. Im requires long-format data for all non-continuous models. All of this is exposed in greater detail and worked examples at <https://lindeloev.github.io/tests-as-linear>.

* See the note to the two-way ANOVA for explanation of the notation.

** Same model, but with one variance per group: `glm(value ~ 1 + G, weights = varIdent(form = ~1|group), method="ML")`.



Jonas Kristoffer Lindeløv
<https://lindeloev.net>

Abbildung 2.8: Common statistical tests as linear models

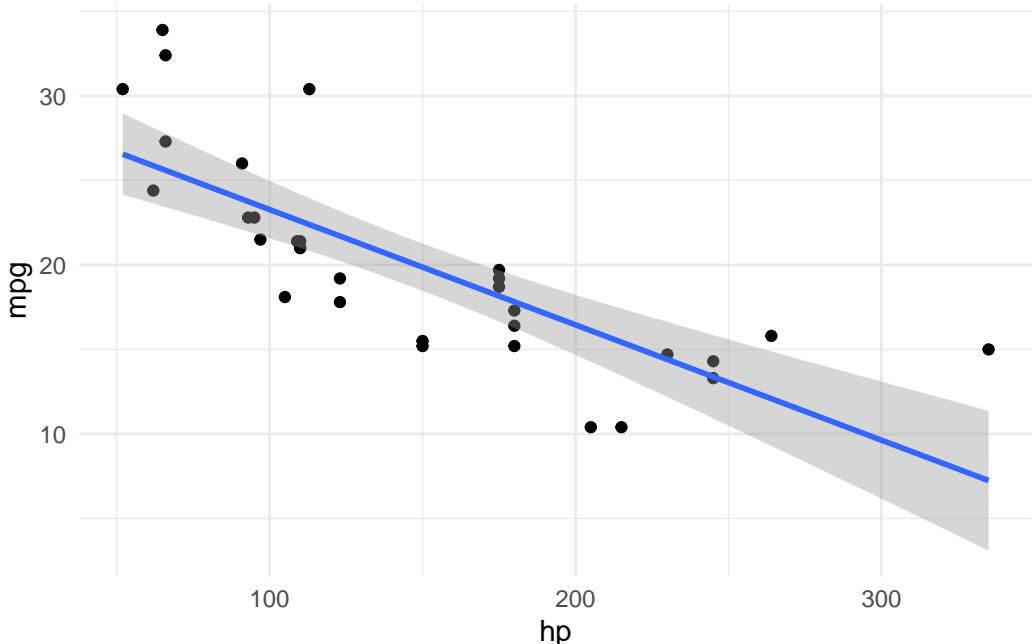


Abbildung 2.9: Die Regressionsgerade in voller Pracht

2.7 Unsicherheit

2.7.1 Inferenz beinhaltet Unsicherheit

Inferenzstatistische Schlüsse sind mit Unsicherheit behaftet: Schließlich kennt man nur einen Teil (die Stichprobe) eines Ganzen (die Population), möchte aber vom Teil auf das Ganze schließen.

! Wichtig

Nichts Genaues weiß man nicht: Schließt man von einem Teil auf das Ganze, so geschieht das unter Unsicherheit. Man spricht von Ungewissheit, da man die Unsicherheit das Wissen über das Ganze betrifft.

Schließt man etwa, dass in einer Grundgesamtheit der Anteil der R-Fans bei 42% liegt, so geschieht das unter Unsicherheit. Man ist sich nicht sicher, dass es wirklich 42% in der Population sind - und nicht etwa etwas mehr oder etwas weniger. Schließlich hat man *nicht* die ganze Population gesehen bzw. vermessen. *Sicher* ist man sich hingegen für die Stichprobe (Messfehler einmal ausgeblendet).

Zur Bemessung der Unsicherheit (Ungewissheit) bedient man sich der Wahrscheinlichkeitsrechnung (wo immer möglich).

Die Wahrscheinlichkeitstheorie bzw. -rechnung wird auch als die Mathematik des Zufalls bezeichnet.

i Hinweis

Unter einem zufälligen Ereignis (random) verstehen wir ein Ereignis, das nicht (komplett) vorherzusehen ist, wie etwa die Augenzahl Ihres nächsten Würfelwurfs. Zufällig bedeutet nicht (zwangsläufig), dass das Ereignisse keine Ursachen besitzt. So gehorchen die Bewegungen eines Würfels den Gesetzen der Physik, nur sind uns diese oder die genauen Randbedingungen nicht (ausreichend) bekannt.

Welche physikalischen Randbedingungen wirken wohl auf einen Münzwurf ein?

2.7.2 Beispiele zur Quantifizierung von Ungewissheit

Aussagen mit Unsicherheit können unterschiedlich präzise formuliert sein.

- Morgen regnet's \Leftrightarrow Morgen wird es hier mehr als 0 mm Niederschlag geben ($p = 97\%$).
- Methode A ist besser als Methode B \Leftrightarrow Mit einer Wahrscheinlichkeit von 57% ist der Mittelwert für Methode A höher als für Methode B .
- Die Maschine fällt demnächst aus \Leftrightarrow Mit einer Wahrscheinlichkeit von 97% wird die Maschine in den nächsten 1-3 Tagen ausfallen, laut unserem Modell.

- Die Investition lohnt sich \Leftrightarrow Die Investition hat einen Erwartungswert von 42 Euro; mit 90% Wahrscheinlichkeit wird der Gewinn zwischen -10000 und 100 Euro.

Geben Sie weitere Beispiele an!

2.7.3 Zwei Arten von Ungewissheit

Im Modellieren im Allgemeinen und in Regressionsmodellen im Besonderen lassen sich (mindestens) zwei Arten von Ungewissheiten angeben, s. auch Abb. Abbildung 2.10.

1. Wie (un)gewiss ist man sich über den Wert des Regressionsgewichts?
2. Wie (un)gewiss ist man sich über den Wert von Y? Schließlich könnte es ja Einflüsse (X) geben, die man nicht berücksichtigt hat.

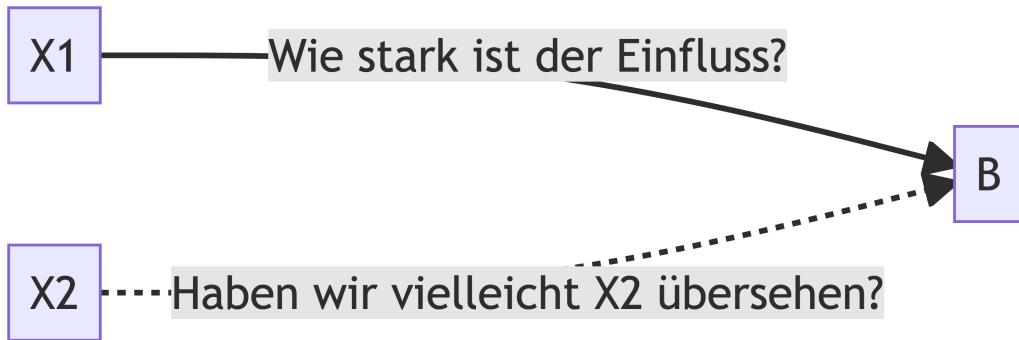


Abbildung 2.10: Zwei Arten der Ungewissheit beim Modellieren

2.7.4 Ich weiß, was ich nicht weiß: Ungewissheit angeben

Streng genommen ist eine Inferenz aus Angabe der Ungewissheit (Genauigkeit der Schätzung) wertlos. Angenommen, jemand sagt, dass sie den Anteil der R-Fans (in der Population) auf 42% schätzt, lässt aber offen wie *sicher* (präzise) die Schätzung ist. Wir wissen also nicht, ob z.B. 2% oder 82% noch erwartbar sind. Oder ob man im Gegenteil mit hoher Sicherheit sagen kann, die Schätzung schließt sogar 41% oder 43% aus.

! Wichtig

Eine Inferenz nennt man auch Schätzung. Es sollte immer die Genauigkeit (Ungewissheit) der Schätzung angegeben werden.

Im Rahmen der Regressionsanalyse schlägt sich die Ungewissheit an zwei Stellen nieder:

1. zur Lage der Regressionsgeraden (β_0, β_1)
2. zu Einflüssen (X), die unser Modell nicht kennt (ϵ, σ)

2.7.5 Visualisierung von Ungewissheit

Gibt man nur einen Punktwert an, wie 42%, als Ergebnis einer Inferenz, spricht man von einem *Punktschätzer*. Punktschäster beinhalten *keine Angabe* der Schätz(un)genauigkeit, s. Abb. Abbildung 2.11, links. Rot markiert: Die Punktschätzung von mpg für hp=200.

Eine Punktschätzung mittels einer Regressionsanalyse ohne (links) bzw. mit Ungewissheitsintervall (rechts, in grau)

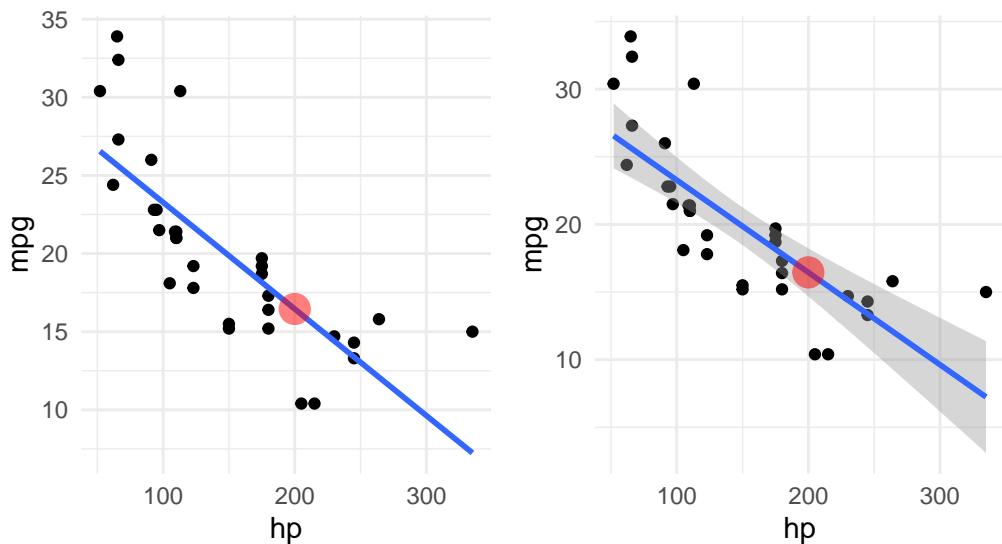


Abbildung 2.11: Eine Punktschätzung und ihre Ungewissheit

In Abb. Abbildung 2.11, rechts, ist die Ungewissheit in den Regressionskoeffizienten visualisiert: Wie sicher sind wir uns zur Stärke des Zusammenhangs von X und Y?

Auch wenn wir uns *sicher* im Hinblick auf die Regressionsgewichte in Abb. Abbildung 2.12 *bliebe eine Restungewissheit*: Unsere Schätzungen wären auch dann nicht sicher, nicht fehlerfrei. Das liegt daran, da das Modell nicht alle Einflüsse auf Y berücksichtigt, sondern nur einen, hier als X bezeichnet.

In Abb. Abbildung 2.12 ist nicht nur die Ungewissheit durch die Regressionsgewichte, sondern auch die "Restungewissheit" dargestellt. In diesem Fall spricht man von einem "Vorhersageintervall", da man nicht nur von "typischen Fällen" auf der Regressioneraden spricht, sondern für echte Fälle Vorhersagen (Schätzungen) tätigt, wo auch die zweite Art von Ungewissheit relevant ist.

Wie man sieht, wird die Ungewissheit *größer*, wenn man beide Arten der Ungewissheit berücksichtigt. Das Vorhersage-Intervall berücksichtigt Ungewissheit in $\beta_0, \beta_1, \epsilon$ bei der Vorhersage von \hat{y}_i .

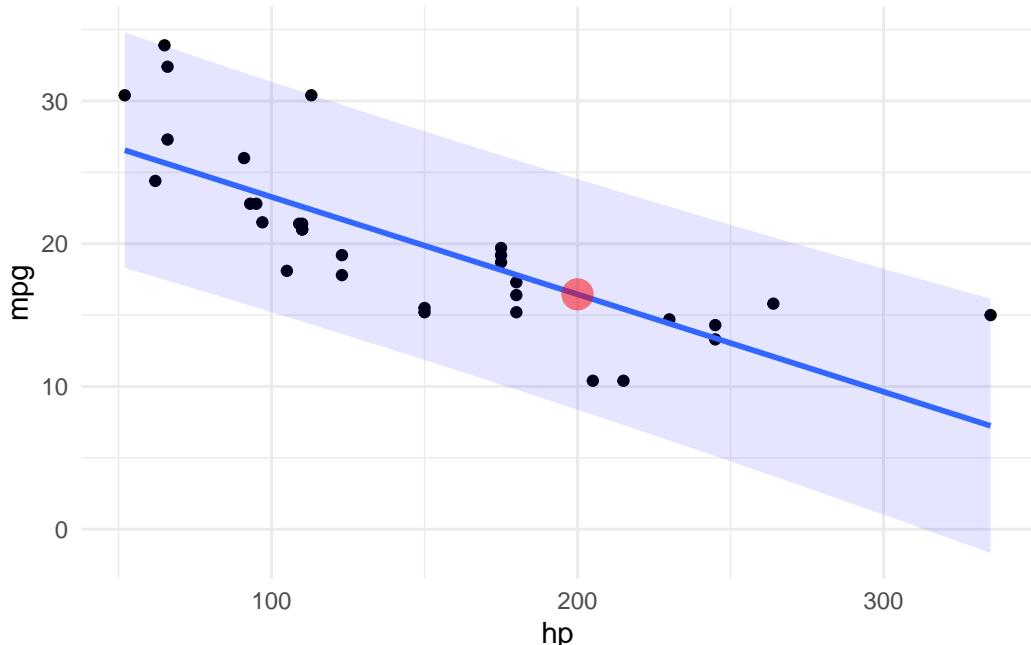


Abbildung 2.12: Zweifache Ungewissheit in den Regressionskoeffizienten - Vorhersageintervall

Geben Sie ein vergleichbares Beispiel an!

2.7.6 Konfidenzintervall

Wir sehen hier, dass ein ‘‘Ungewissheitskorridor’’ angegeben wird. Entsprechend wird nicht ein *Punktschätzer*, sondern ein *Schäzbereich* angegeben. Man spricht auch von einem *Konfidenzintervall* oder *Unsicherheitsbereich*⁵

Ein Konfidenzintervall wird häufig mit 90% oder 95% Genauigkeit angegeben. Im Kontext der Bayes-Analyse ist das einfach zu interpretieren. Sagen wir, wir finden, dass in einem Modell ein 95%-Konfidenzintervall für den Anteil der R-Fans angegeben wird, dass sich von 40 bis 44 Prozent erstreckt. Dieser Befund lässt sich so interpretieren: ‘‘Laut Modell liegt der gesuchte Anteil mit einer Wahrscheinlichkeit von 95% im Bereich von 44 bis 44 Prozentpunkten.’’

! Wichtig

Ein Konfidenzintervall gibt einen Schäzbereich plausibler Werte für den gesuchten Wert in der Population (den Parameter) an.

Interpretieren Sie den Ungewissheitskorridor!

⁵Tatsächlich gibt es mehrere Synonyme oder ähnliche Begriffe für Konfidenzintervall. Wir kommen später darauf detaillierter zu sprechen.

2.8 Klassische vs. Bayes-Inferenz

2.8.1 Klassische Inferenz: Frequentismus

- Die Berücksichtigung von Vorwissen zum Sachgegenstand wird vom Frequentismus als subjektiv zurückgewiesen.
- Nur die Daten selber fliesen in die Ergebnisse ein
- Wahrscheinlichkeit wird über relative Häufigkeiten definiert.
- Es ist nicht möglich, die Wahrscheinlichkeit einer Hypothese anzugeben.
- Stattdessen wird angegeben, wie häufig eine vergleichbare Datenlage zu erwarten ist, wenn die Hypothese gilt und der Versuch sehr häufig wiederholt ist.
- Ein Großteil der Forschung (in den Sozialwissenschaften) verwendet diesen Ansatz.

2.8.2 Bayesianische Inferenz

- Vorwissen (Priori-Wissen) fließt explizit in die Analyse ein (zusammen mit den Daten).
- Wenn das Vorwissen gut ist, wird die Vorhersage genauer, ansonsten ungenauer.
- Die Wahl des Vorwissens muss explizit (kritisierbar) sein.
- In der Bayes-Inferenz sind Wahrscheinlichkeitsaussagen für Hypothesen möglich.
- Die Bayes-Inferenz erfordert mitunter viel Rechenzeit und ist daher erst in den letzten Jahren (für gängige Computer) komfortabel geworden.

2.8.3 Vergleich von Wahrscheinlichkeitsaussagen

2.8.3.1 Frequentismus

Die zentrale Statistik heißt der *p-Wert*

Der p-Wert ist so definiert: "Wie wahrscheinlich ist der Wert der Teststatistik (oder noch extremere Werte), vorausgesetzt die Nullhypothese gilt und man wiederholt den Versuch unendlich oft (mit gleichen Bedingungen, aber zufällig verschieden und auf Basis unseres Modells)?"

Findet man $p < .05$ (oder einen anderen Prozentwert, aber meistens wird 5% hergenommen), so spricht man von "(statistischer) Signifikanz" und nimmt dies als Beleg, dass man einen Effekt gefunden hat, die Hypothese eines Nulleffekts (z.B. kein Zusammenhang von X und Y) also verwerfen kann.

2.8.3.2 Bayes-Statistik

Die zentrale Statistik ist die *Posteriori-Verteilung*.

Die Posteriori-Verteilung beantwortet uns die Frage: "Wie wahrscheinlich ist die Forschungshypothese, jetzt, nachdem wir die Daten kennen, auf Basis unseres Modells?"

Recherchieren Sie eine Definition des p-Werts und lesen Sie sie genau.

In diesem Post wird für Bayes geworben und (vielleicht einseitig) Stellung pro Bayes bezogen.

2.8.4 Frequentist und Bayesianer

Im Cartoon 1132 von [xkcd](#) wird sich über das Nicht-Berücksichtigen von Vorabinformationen (Prior-Verteilung) lustig gemacht, s. Abbildung 2.13.

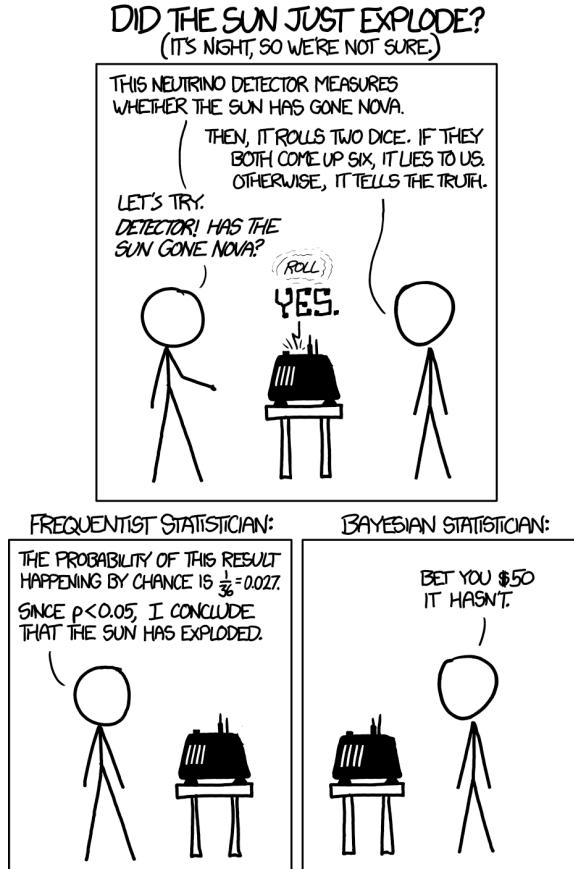


Abbildung 2.13: Frequentist wettet mit Bayesianer

Quelle

2.8.5 Der p-Wert ist wenig intuitiv

from Imgflip Meme Generator

2.8.6 Beispiel zum Nutzen von Apriori-Wissen 1

Ein Betrunkener behauptet, er könne hellsehen. Er wirft eine Münze 10 Mal und sagt jedes Mal korrekt vorher, welche Seite oben landen wird.

Die Wahrscheinlichkeit dieses Ergebnisses ist sehr gering (2^{-10}) unter der Hypothese, dass die Münze fair ist, dass Ergebnis also “zufällig” ist.

Unser Vorwissen lässt uns allerdings trotzdem an der Hellsichtigkeit des Betrunkenen zweifeln, so dass die meisten von uns die Hypothese von der Zufälligkeit des Ergebnisses wohl nicht verwerfen.

2.8.7 Beispiel zum Nutzen von Apriori-Wissen 2

Eine Studie (vgl. Gelman, Hill, und Vehtari (2021)) fand einen “großen Effekt” auf das Einkommen von Babies, eine Stunde pro Woche während zwei Jahren an einem psychosozialen Entwicklungsprogramm teilnahmen (im Vergleich zu einer Kontrollgruppe), $n = 127$.

Nach 20 Jahren war das mittlere Einkommen der Experimentalgruppe um 42% höher (als in der Kontrollgruppe) mit einem Konfidenzintervall von $[+2\%, +98\%]$.

Allerdings lässt uns unser Vorwissen vermuten, dass so ein Treatment das Einkommen nach 20 Jahren kaum verdoppeln lässt. Wir würden den Effekt lieber in einem konservativeren Intervall schätzen (enger um Null).

2.9 Literatur

Bei Gelman, Hill, und Vehtari (2021), Kap. 1 findet sich eine Darstellung ähnlich zu der in diesem Kapitel.

2.10 Aufgaben

1. [Griech-Buchstaben-Inferenz](#)
2. [korr-als-regr](#)
3. [ttest-als-regr](#)
4. [ttest-skalenniveau](#)
5. [adjustieren2](#)
6. [inferenz-fuer-alle](#)
7. [adjustieren1](#)
8. [un gewiss-arten-regr](#)
9. [vorhersageintervall1](#)
10. [lm-standardfehler](#)
11. [punktschaetzer-reicht-nicht](#)

3 Wahrscheinlichkeit

3.1 Lernsteuerung

3.1.1 Lernziele

Nach Absolvieren des jeweiligen Kapitels sollen folgende Lernziele erreicht sein.

Sie können ...

- die Grundbegriffe der Wahrscheinlichkeitsrechnung erläuternd definieren
- die drei Arten der direkten Ermittlung von Wahrscheinlichkeit erläutern
- typische Relationen (Operationen) von Ereignissen anhand von Beispielen veranschaulichen
- mit Wahrscheinlichkeiten rechnen

3.1.2 Prüfungsrelevanter Stoff

Lesen Sie dazu Bourier (2018), Kap. 2-4. Weitere Übungsaufgaben finden Sie im dazugehörigen Übungsbuch, Bourier (2022).

3.1.3 Zentrale Begriffe

3.1.3.1 Grundbegriffe

- Zufallsvorgang (Zufallsexperiment)
- Elementarereignis
- Ereignisraum
- Zufallsereignis (zufälliges Ereignis)
- Sicherer Ereignis
- Unmögliches Ereignis

3.1.3.2 Wahrscheinlichkeitsbegriffe

- Klassische Wahrscheinlichkeit (LaPlace'sche Wahrscheinlichkeit)
- Statistische (empirische) Wahrscheinlichkeitsermittlung
- Subjektive (Bayes) Wahrscheinlichkeitsermittlung

3.1.3.3 Wahrscheinlichkeitsrelationen

- Vereinigung von Ereignissen
- Schnitt(menge) von Ereignissen
- Komplementärereignis
- Vollständiges Ereignissystem
- Kolmogorovs Definition von Wahrscheinlichkeit

3.1.3.4 Wahrscheinlichkeitsrechnung

- Allgemeiner Additionssatz
- Disjunkte Ereignisse
- Additionssatz für disjunkte Ereignisse
- Bedingte Wahrscheinlichkeit
- (Stochastische) Unabhängigkeit
- Baumdiagramm für gemeinsame Wahrscheinlichkeit
- Allgemeiner Multiplikationssatz
- Multiplikationssatz für unabhängige Ereignisse
- Totale Wahrscheinlichkeit
- Satz von Bayes

3.1.4 Begleitvideos

- [Video zum Thema Wahrscheinlichkeit](#)

3.2 Unterstützung: Wahrscheinlichkeit in Bildern

Wahrscheinlichkeit in Bildern: zur einfachen Erschließung des Materials, ein Unterstützungsangebot.

Im Folgenden sind einige Schlüsselbegriffe und -regeln in (ver-)einfach(t)er Form schematisch bzw. visuell dargestellt mit dem Ziel, den Stoff einfacher zu erschließen.

3.2.1 Zufall

Werfen Sie eine Münze!

Diese hier zum Beispiel:



[Quelle: By OpenClipartVectors, CC0](#)

Wiederholen Sie den Versuch 10, nein, 100, nein 1000, nein: 10^6 Mal.

Notieren Sie das Ergebnis!

Oder probieren Sie die [App der Brown University](#).

3.2.2 Relationen von Mengen

Venn-Diagramme eignen sich, um typische Operationen (Relationen) auf Mengen zu visualisieren.

3.2.2.1 Überblick

Die folgenden Diagramme stammen von [Wikipedia \(En\)](#).

Wir gehen von Ereignisraum Ω aus, mit dem Ereignis A als Teilmenge: $A \subset B$.

3.2.2.2 Disjunkte Ereignisse

(Engl. disjoint events)

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

A und B sind disjunkt: ihre Schnittmenge ist leer: $A \cap B = \emptyset$, s. Abbildung [3.6](#)

3.2.2.3 Eselsbrücke zur Vereinigungs- und Schnittmenge

Das Zeichen für eine Vereinigung zweier Mengen kann man leicht mit dem Zeichen für einen Schnitt zweier Mengen leicht verwechseln; daher kommt eine Eselbrücke gelesen, s. Abbildung [3.7](#).

[Quelle: rither.de](#)

3.2.2.4 Animationen

[Animation zu Mengenoperationen](#)

[Animation zur Vereinigung von Mengen](#)

[Quelle](#)

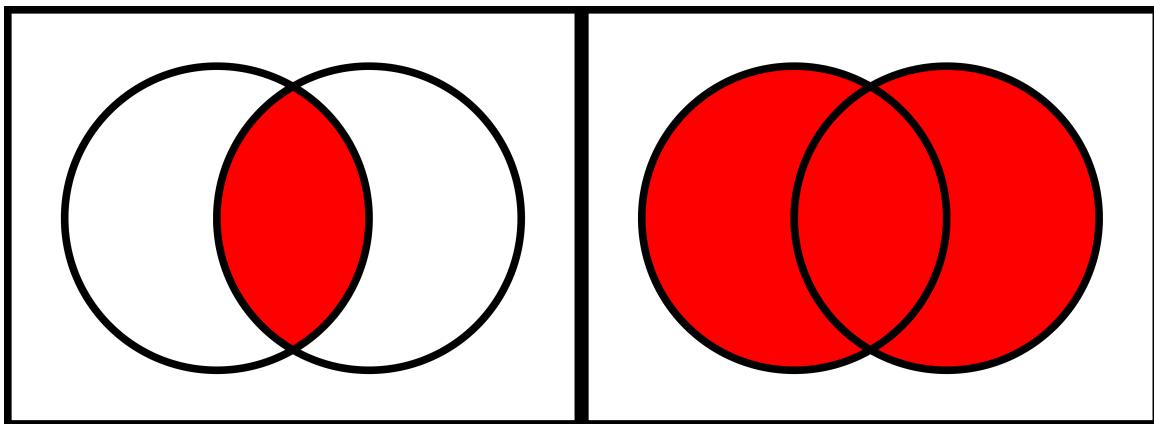


Abbildung 3.1: $A \cap B$

Abbildung 3.2: $A \cup B$

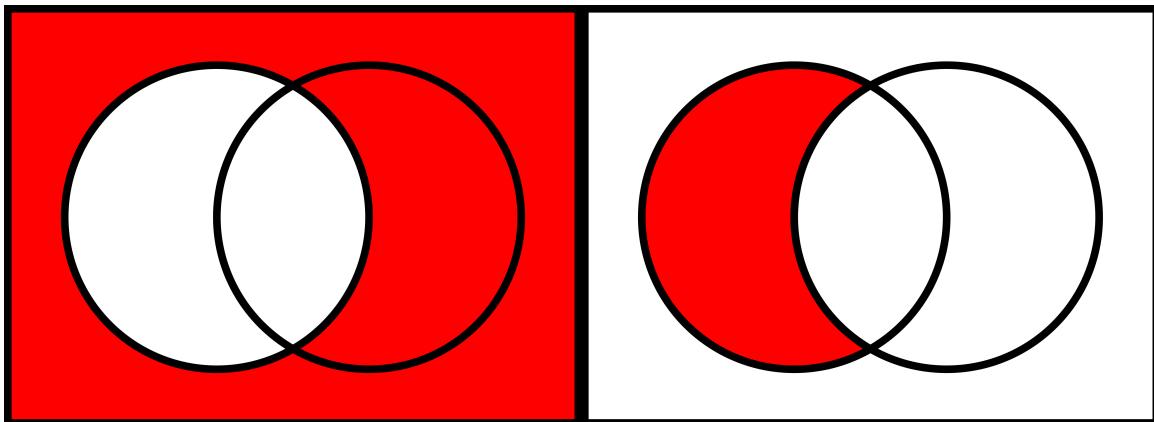


Abbildung 3.3: \bar{A}

Abbildung 3.4: $A \setminus B$

Abbildung 3.5: Typische Mengenoperationen

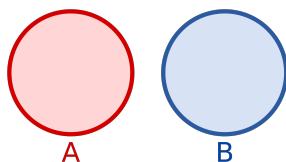


Abbildung 3.6: Zwei disunkte Ereignisse, dargestellt noch überlappungsfreie Kreise

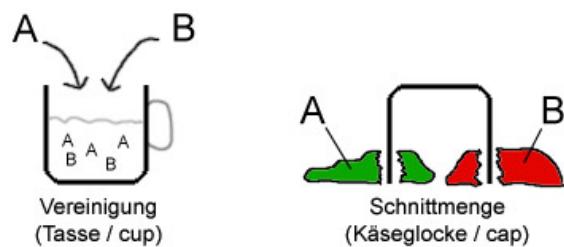


Abbildung 3.7: Eselsbrücke für Vereinigungs- und Schnittmenge

3.2.3 Additionssatz

Der Additionssatz wird verwendet, wenn wir an der Wahrscheinlichkeit interessiert sind, dass *mindestens eines der Ereignisse* eintritt.

3.2.3.1 Diskunkte Ereignisse

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

1 2 3 4 5 6

Gesucht sei die Wahrscheinlichkeit des Ereignis $A = \{1, 2\}$.

1 2 3 4 5 6

$$P(1 \cup 2) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6}$$

3.2.3.2 Allgemein (disjunkt oder nicht disjunkt)

Bei der Addition der Wahrscheinlichkeiten für A und B wird der Schnitt $A \cap B$ doppelt erfasst. Er muss daher noch abgezogen werden (vgl. Abbildung 3.10):

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

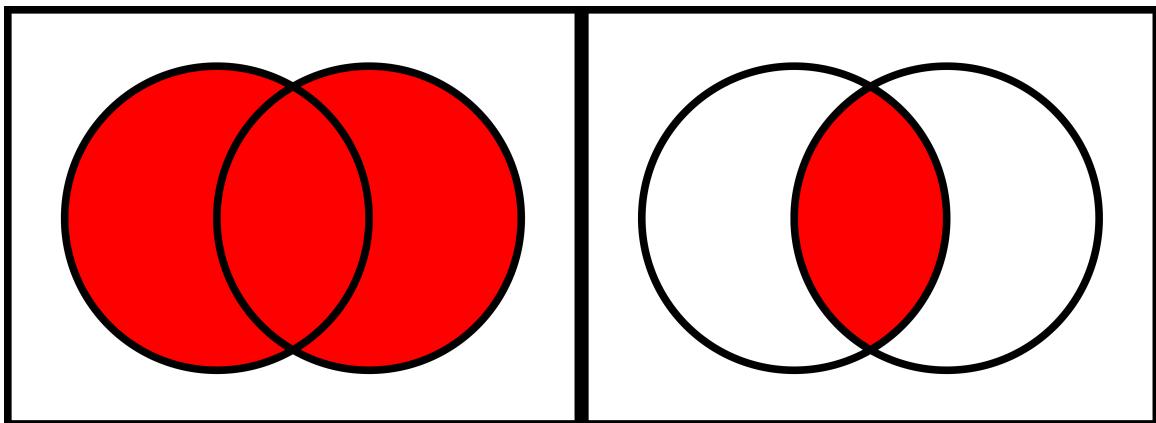


Abbildung 3.8: $A \cup B$

Abbildung 3.9: $A \cap B$

Abbildung 3.10: Die Schnittmenge muss beim Vereinigen abgezogen werden, damit sie nicht doppelt gezählt wird.

3.2.4 Bedingte Wahrscheinlichkeit

3.2.4.1 Animation

Schauen Sie sich mal diese [Wahnsinnsanimation von Victor Powell an](#). Hammer!

3.2.4.2 Schema

Abb. Abbildung 3.11 illustriert gemeinsame Wahrscheinlichkeit, $\Pr(A \cap B)$ und bedingte Wahrscheinlichkeit, $\Pr(A|B)$.

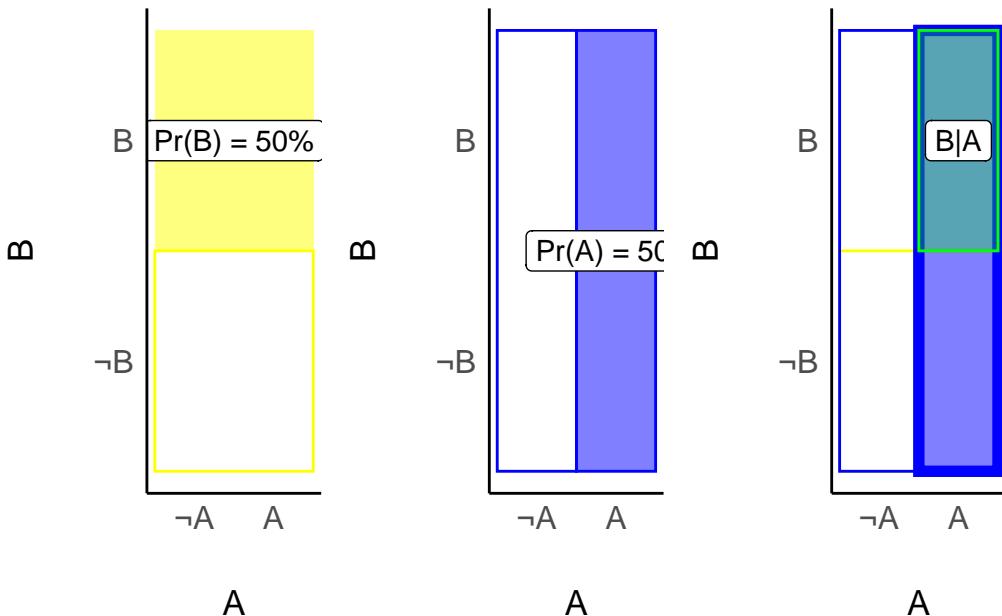


Abbildung 3.11: Illustration von gemeinsamer und bedingter Wahrscheinlichkeit

Bedingte Wahrscheinlichkeit ist vergleichbar zu Filtern einer Tabelle:

```
d <-
  tribble::tribble(
    ~id, ~A, ~B,
    "1", 0L, 0L,
    "2", 0L, 1L,
    "3", 1L, 0L,
    "4", 1L, 1L,
    "SUMME", 2L, 2L
  )
```

Es ergeben sich folgende Wahrscheinlichkeiten:

$$P(A) = 2/4$$

$$P(B) = 2/4$$

$$P(A \cap B) = 1/4$$

$$P(A|B) = 1/2$$

3.2.5 (Un-)Abhangigkeit

Stochastische Unabhangigkeit ist ein Spezialfall von Abhangigkeit: Es gibt sehr viele Auspragungen fur Abhangigkeit, aber nur eine fur Unabhangigkeit. Konnen wir Unabhangigkeit nachweisen, haben wir also eine starke Aussage getatigt.

Abhangig, s. Abbildung 3.12, links: Uberleben auf der Titanic ist offenbar *abhangig* von der Passagierklasse. Auf der anderen Seite: Das Ereignis *Uberleben* auf der Titanic ist *unabhangig* vom Ereignis *Alter ist eine Primzahl*, s. Abbildung 3.12, rechts.

Abhangigkeit zweier Ereignisse

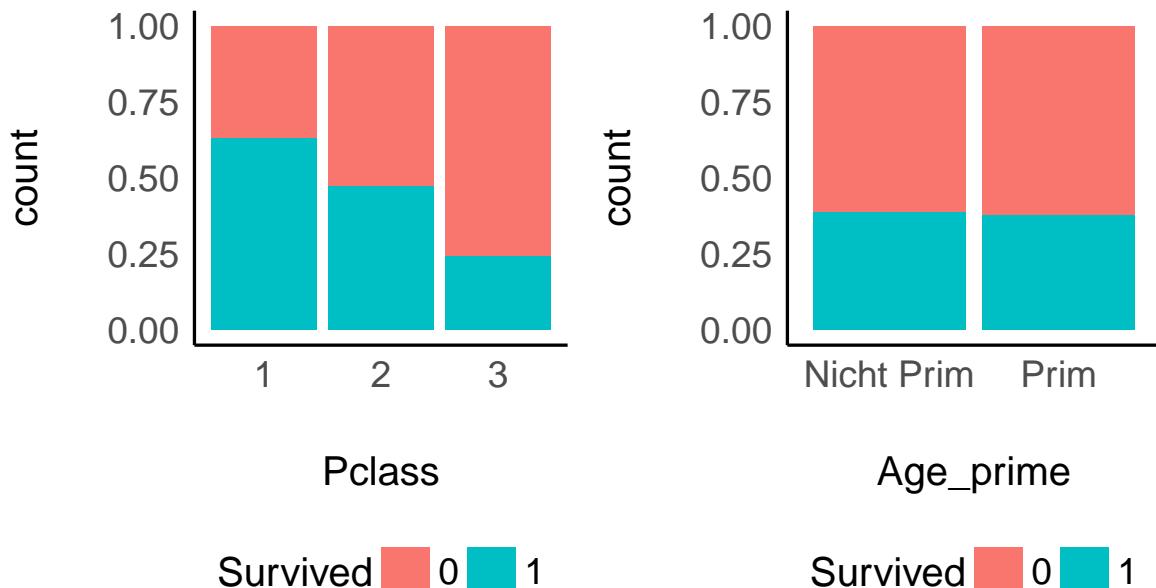


Abbildung 3.12: Abhangigkeit und Unabhangigkeit zweier Ereignisse

Zur Ab- bzw. Un-Abhangigkeit zweier Variablen, an Beispielen illustriert.

Beispiel 3.1 (Zusammenhang von Covidsterblichkeit und Impfquote). Sind die Ereignisse *Tod durch Covid* bzw. *Impfquote (A)* und *Land¹ (B)* voneinander abhangig (Abb. Abbildung 3.13)?

¹hier mit den zwei Auspragungen *DEU* und *USA*

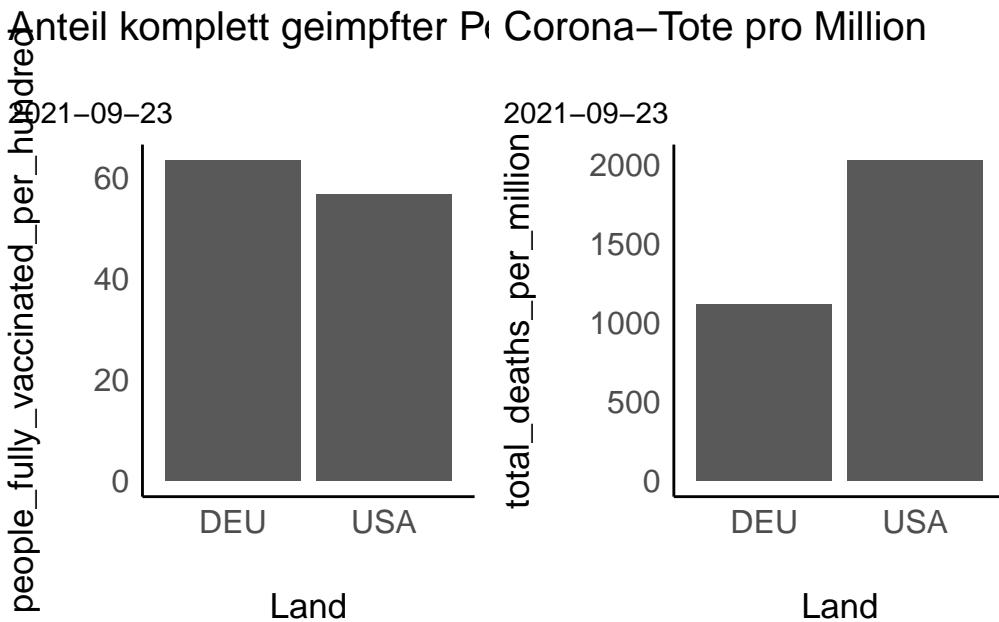


Abbildung 3.13: Impfquote und Sterblichkeit sind voneinander abhängig (bezogen auf Covid, auf Basis vorliegender Daten)

Ja, da in beiden Diagrammen gilt: $P(A|B) \neq Pr(A) \neq Pr(A|\neg B)$.

Daten von [Our World in Data](#).

3.2.6 Multiplikationssatz

Der Multiplikationssatz wird verwendet, wenn wir an der Wahrscheinlichkeit interessiert sind, dass *alle Ereignisse* eintreten.

3.2.6.1 Unabhängige Ereignisse

Wir werfen eine faire Münze *zwei* Mal (Abb. Abbildung 3.14).

Abb. Abbildung 3.14 zeigt ein *Baumdiagramm*. Jeder *Kasten* (Knoten) zeigt ein *Ergebnis*. Die Pfeile (Kanten) symbolisieren die Abfolge des Experiments: Vom "Start" (schwarzer Kreis) führen zwei mögliche Ergebniss ab, jeweils mit Wahrscheinlichkeit 1/2. Die untersten Knoten nennt man auch *Blätter* (Endknoten), sie zeigen das Endresultat des (in diesem Fall) zweifachen Münzwurfs. Der Weg vom Start zu einem bestimmten Blatt nennt man *Pfad*. Die Anzahl der Pfade entspricht der Anzahl der Blätter. In diesen Diagramm gibt es vier Pfade (und Blätter).

```
## # A tibble: 3 x 2
##   Ereignis Pr
##   <chr>    <chr>
```

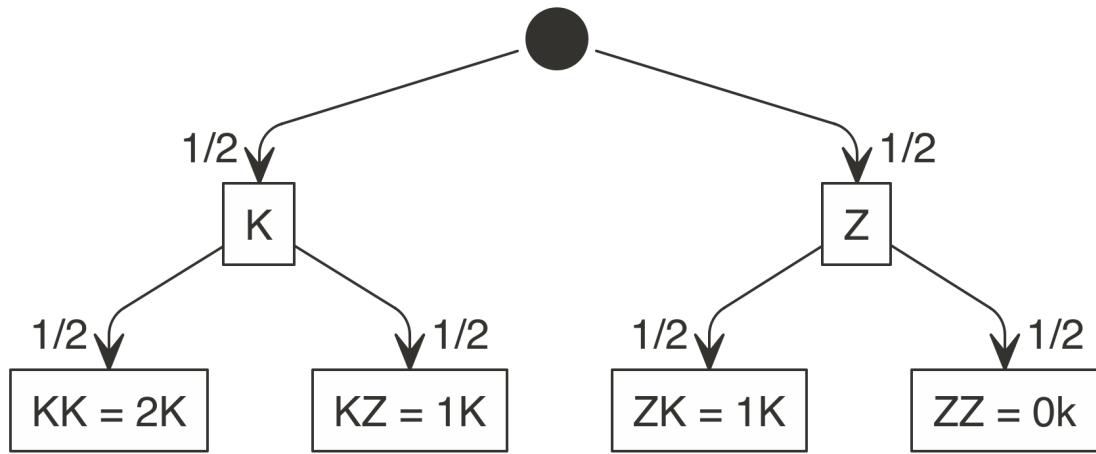


Abbildung 3.14: Wir werfen 2 faire Münzen

```

## 1 OK      1/2 * 1/2 = 1/4
## 2 1K      1/4 + 1/4 = 1/2
## 3 2K      1/2 * 1/2 = 1/4
  
```

Wir werfen eine faire Münze *drei* Mal (Abb. Abbildung 3.15)

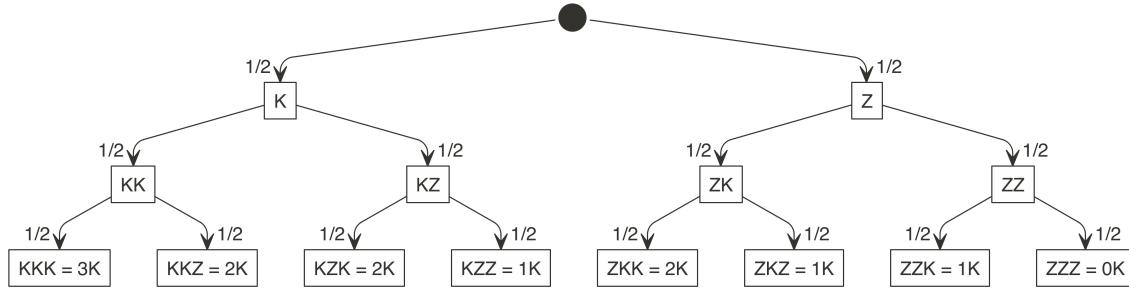


Abbildung 3.15: Wir werfen drei faire Münzen

```

## # A tibble: 4 x 2
##   Ereignis Pr
##   <chr>    <chr>
## 1 OK      1/2 * 1/2 * 1/2 = 1/8
## 2 1K      1/8 + 1/8 + 1/8 = 3/8
## 3 2K      3 * 1/8 = 3/8
## 4 3K      1/2 * 1/2 * 1/2 = 1/8
  
```

$$Pr(AB) = Pr(A) \cdot Pr(B) = 50\% \cdot 50\% = 25\%$$

Abb. Abbildung 3.16 zeigt, dass gilt: $P(A \cap B) = P(A) \cdot P(B) = P(B) \cdot P(A)$.

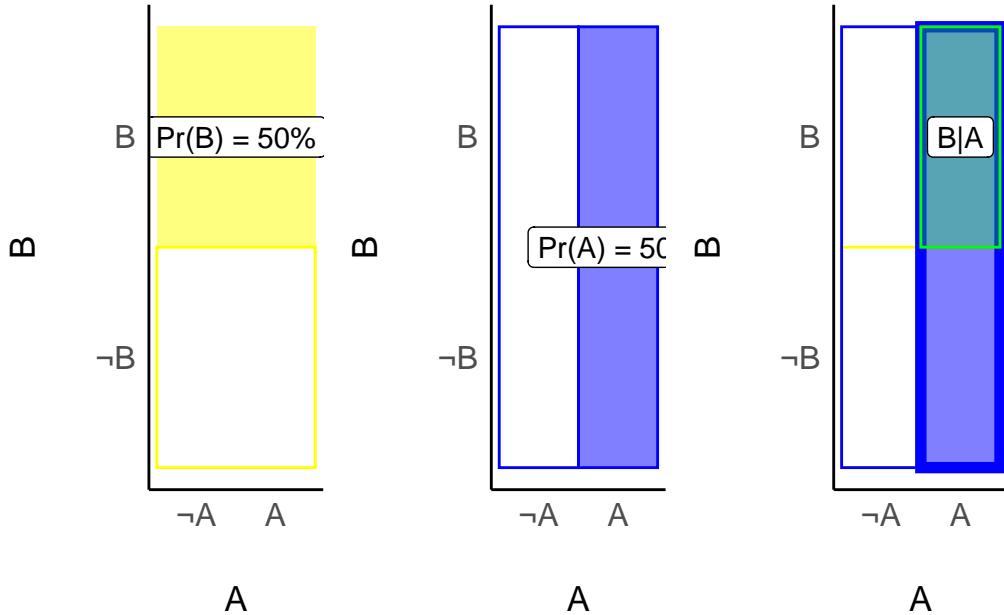


Abbildung 3.16: Unabhängige Ereignisse visualisiert

3.2.6.2 Kalt und Regen

Von McElreath (2020) stammt diese Verdeutlichung der gemeinsamen Wahrscheinlichkeit:

Was ist die Wahrscheinlichkeit für *kalt und Regen* ?

Die Wahrscheinlichkeit für kalt und Regen ist die Wahrscheinlichkeit von *Regen* , wenn's *kalt* ist mal die Wahrscheinlichkeit von *Kälte* .

Ebenfalls gilt:

Die Wahrscheinlichkeit für kalt und Regen ist die Wahrscheinlichkeit von *Kälte* , wenn's *regnet* mal die Wahrscheinlichkeit von *Regen* .

Das Gesagte als Emoji-Gleichung:

$$P(\text{und}) = P(\text{ }) \cdot P(\text{ }) = P(\text{ }) \cdot P(\text{ }) = P(\text{und})$$

Allgemein:

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

Man kann also die “Gleichung drehen”.

3.2.6.3 Abhängige Ereignisse

Ein Baumdiagramm bietet sich zur Visualisierung abhängiger Ereignisse an, s. Abb. Abbildung 3.17. Für unabhängige Ereignisse übrigens auch.

In einer Urne befinden sich fünf Kugeln, von denen vier rot sind und eine blau ist.

wie groß ist die Wahrscheinlichkeit, dass bei zwei Ziehungen ohne Zurücklegen (*ZOZ*) *zwei rote Kugeln* gezogen werden (Bourier 2018), S. 47.

Hier ist unsere Urne:

R, R, R, R, B

Und jetzt ziehen wir. Hier ist das Baumdiagramm, s. Abb. Abbildung 3.17.

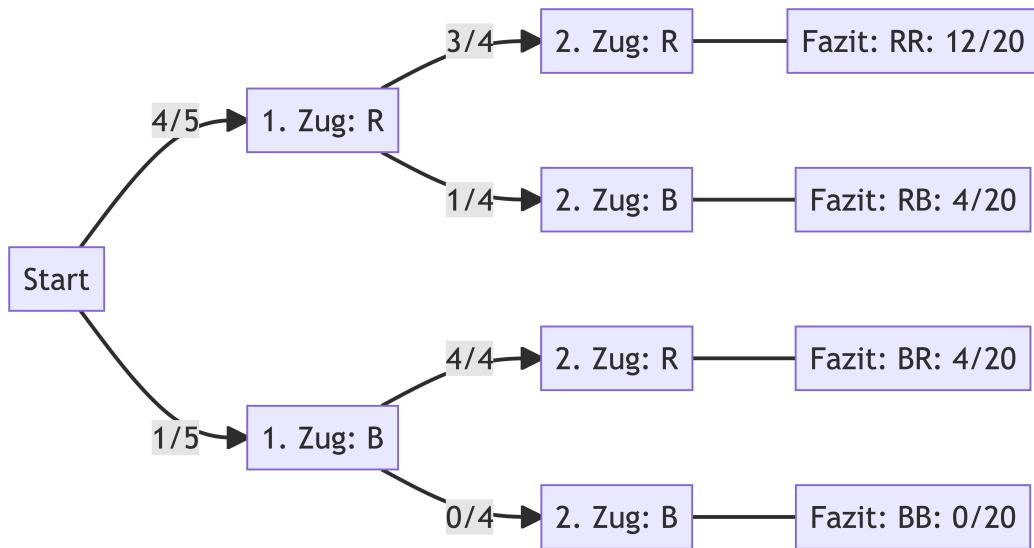


Abbildung 3.17: Baumdiagramm für ein zweistufiges Zufallsereignis, wobei der 2. Zug (Stufe) abhängig ist vom 1. Zug.

Es gilt also: $P(A \cap B) = P(A) \cdot P(B|A)$.

3.2.7 Totale Wahrscheinlichkeit

Abbildung 3.18 zeigt das Baumdiagramm für die Aufgabe Bourier (2018), S. 56.

Gesucht ist die Wahrscheinlichkeit $P(B)$.

Dazu addieren wir die Wahrscheinlichkeiten der relevanten Äste.

```

W_total <- 0.6 * 0.05 + 0.1 * 0.02 + 0.3 * 0.04
W_total
## [1] 0.044
  
```

Die totale Wahrscheinlichkeit beträgt also $P(B) = 4.4\%$.

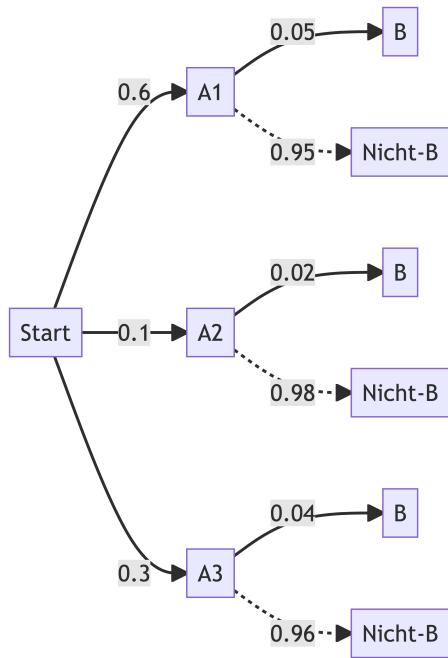


Abbildung 3.18: Totale Wahrscheinlichkeit

Einfacher noch ist es, wenn man anstelle von Wahrscheinlichkeiten absolute Häufigkeiten verwendet.

3.2.8 Bayes

3.2.8.1 Bayes als Baum

Gesucht sei $P(A_1|B)$.

Für Bayes' Formel setzt man die Wahrscheinlichkeit des *günstigen* Ast zur Wahrscheinlichkeit aller relevanten Äste, $P(B)$.

Der günstige Ast ist hier schwarz gedruckt, die übrigen Äste gestrichelt, s. Abbildung 3.19.

$$P(A|B) = \frac{P(A_1 \cap B)}{P(B)} = \frac{0.6 \cdot 0.05}{0.03 + 0.002 + 0.012} = \frac{0.03}{0.044} \approx 0.68$$

$P(A|B)$ beträgt also ca. 68%.

Zur Erinnerung: $P(B)$ ist die totale Wahrscheinlichkeit.

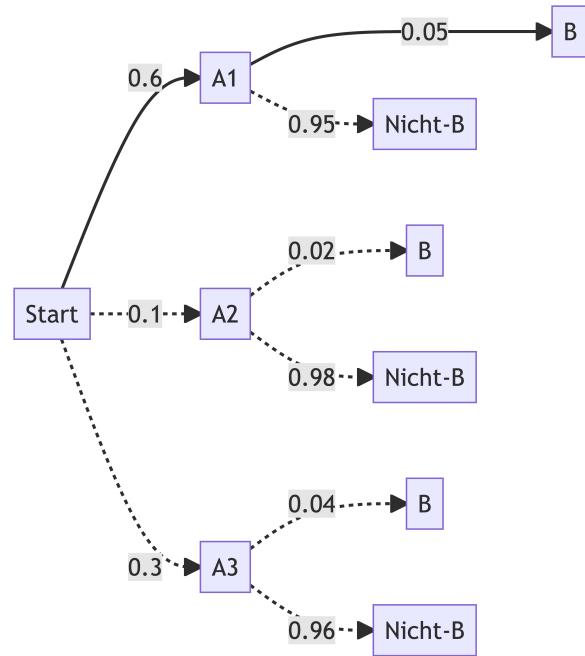


Abbildung 3.19: Günstige Pfade

3.3 Bayes' Theorem

3.3.1 Bayes als bedingte Wahrscheinlichkeit

Bayes' Theorem ist auch nur eine normale bedingte Wahrscheinlichkeit:

$$P(A|B) = \frac{\text{umformen}}{\widehat{P(A \cap B)}} \frac{P(A \cap B)}{P(B)}$$

$P(A \cap B)$ kann man umformen, s. Gleichung 3.1:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A) \cdot P(A)}{P(B)} \quad (3.1)$$

Man kann sich Bayes' Theorem auch wie folgt herleiten:

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

Dann lösen wir nach $P(A|B)$ auf:

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)}$$

3.3.2 Wozu wird Bayes in der Praxis genutzt?

In der Praxis nutzt man Bayes häufig, wenn man Daten zu einer Wirkung W hat, und auf die Ursache U zurückschließen möchte, sinngemäß:

$$W \xrightarrow{\text{Bayes}} U.$$

Dann kann man Gleichung 3.1 so schreiben, s. Gleichung 3.2:

$$P(U|W) = \frac{P(U) \cdot P(W|U)}{P(W)} \quad (3.2)$$

Eine ähnliche Situation, die in der Praxis häufig ist, dass man Daten D hat und auf die Wahrscheinlichkeit einer Hypothese H schließen möchte, s. Gleichung 3.3.

$$D \xrightarrow{\text{Bayes}} H.$$

$$P(H|D) = \frac{P(H) \cdot P(D|H)}{P(D)} \quad (3.3)$$

Gleichung 3.3 fragt nach $P(H|D)$:

Was ist die Wahrscheinlichkeit der Hypothese H , jetzt wo wir die Daten haben (und ein Modell?)

Und antwortet so (Gleichung 3.3):

Diese Wahrscheinlichkeit entspricht der Grundrate (Apriori-Wahrscheinlichkeit) der Hypothese mal der Plausibilität (Likelihood) der Daten unter Annahme (gegeben) der Hypothese. Aus Standardisierungsgründen dividiert man noch die totale Wahrscheinlichkeit der Daten über alle Hypothesen.

3.3.3 Zusammengesetzte Hypothesen

Das ist vielleicht ein bisschen fancy, aber man kann Bayes' Theorem auch nutzen, um die Wahrscheinlichkeit einer *zusammengesetzten Hypothese* zu berechnen: $H = H_1 \cap H_2$. Ein Beispiel wäre: "Was ist die Wahrscheinlichkeit, dass es Regen (R) *und* Blitzes (B) gibt, wenn es kalt (K) ist?".

Das sieht dann so aus, Gleichung 3.4:

$$\begin{aligned} P(R \cap B|K) &= \frac{P(R \cap B) \cdot P(K|R \cap B)}{P(D)} \\ &= \frac{P(R) \cdot P(B) \cdot P(K|R \cap B)}{P(D)} \end{aligned} \quad (3.4)$$

Hier haben wir $P(R \cap B)$ aufgelöst in $P(R) \cdot P(B)$, das ist nur zulässig, wenn R und B unabhängig sind.

3.3.3.1 Bayes-Video von 3b1b

Das [Video zu Bayes von 3b1b](#) verdeutlicht das Vorgehen der Bayes-Methode auf einfache und anschauliche Weise.

3.4 Aufgaben

Zusätzlich zu den Aufgaben im Buch:

- [mtcars-abhaengig](#)
- [voll-normal](#)
- [corona-blutgruppe](#)
- [Bed-Wskt2](#)
- [Gem-Wskt1](#)
- [wuerfel01](#)
- [wuerfel02](#)
- [wuerfel03](#)
- [wuerfel04](#)

4 Verteilungen

4.1 Lernsteuerung

4.1.1 Lernziele

Nach Absolvieren des jeweiligen Kapitels sollen folgende Lernziele erreicht sein.

Sie können ...

- den Begriff der Zufallsvariablen erläutern
- die Begriffe von Wahrscheinlichkeitsdichte und Verteilungsfunktion erläutern und anhand einfacher Beispiele ausrechnen
- den Begriff einer Gleichverteilung erläutern und einfache Fallbeispiele ausrechnen
- die Parameter einer Normalverteilung nennen und erläutern

4.1.2 Prüfungsrelevanter Stoff

Lesen Sie zusätzlich zum Stoff dieses Kapitels noch Bourier (2018), folgende Abschnitte:

- Kap. 6.1 (Zum Begriff Zufallsvariable)
- Kap. 6.3 (Stetige Zufallsvariablen)
- Kap. 7.1.1 (Binomialverteilung)
- Kap. 7.2.1 (Gleichverteilung) und 7.2.3 (Normalverteilung)

Lösen Sie auch die Übungsaufgaben dazu.

Weitere Übungsaufgaben finden Sie im dazugehörigen Übungsbuch, Bourier (2022).

4.1.3 Benötigte R-Pakete

```
library(patchwork)
library(tidyverse)
```

4.1.4 Zentrale Begriffe

4.1.4.1 Eigenschaften von Zufallsvariablen

- Zufallsvariable (random variable)
- Diskret vs. stetig
- Wahrscheinlichkeitsdichte (Dichte, (probability) density, f)
- Wahrscheinlichkeitsfunktion (kumulierte Wahrscheinlichkeit, Wahrscheinlichkeitsmasse)

4.1.4.2 Verteilungen

- Gleichverteilung
- Normalverteilung
- Standardnormalverteilung

4.1.5 Begleitvideos

- [Video 1 zum Thema Verteilungen](#)
- [Video 2 zum Thema Verteilungen](#)

4.2 Verteilungen

! Wichtig

Eine *Verteilung* zeigt, welche Ausprägungen eine Variable aufweist und wie häufig bzw. wahrscheinlich diese sind. Einfach gesprochen, veranschaulicht eine Balken- oder Histogramm eine Verteilung. Man unterscheidet Häufigkeitsverteilungen (s. Abb. Abbildung 4.1) von Wahrscheinlichkeitsverteilungen (Abb. Abbildung 4.2).

4.2.1 Häufigkeitsverteilung

Die Häufigkeitsverteilung eines *diskreten* Merkmals X mit k Ausprägungen zeigt, wie häufig die einzelnen Ausprägungen sind. So hat die Variable *Zylinder* (in einem Datensatz) etwa die Ausprägungen 4,6 und 8.

```
data(mtcars)
mtcars %>%
  count(cyl)
##   cyl   n
## 1    4 11
```

```
## 2   6   7
## 3   8  14
```

Abb. Abbildung 4.1, links, visualisiert die Häufigkeitsverteilung von cyl.

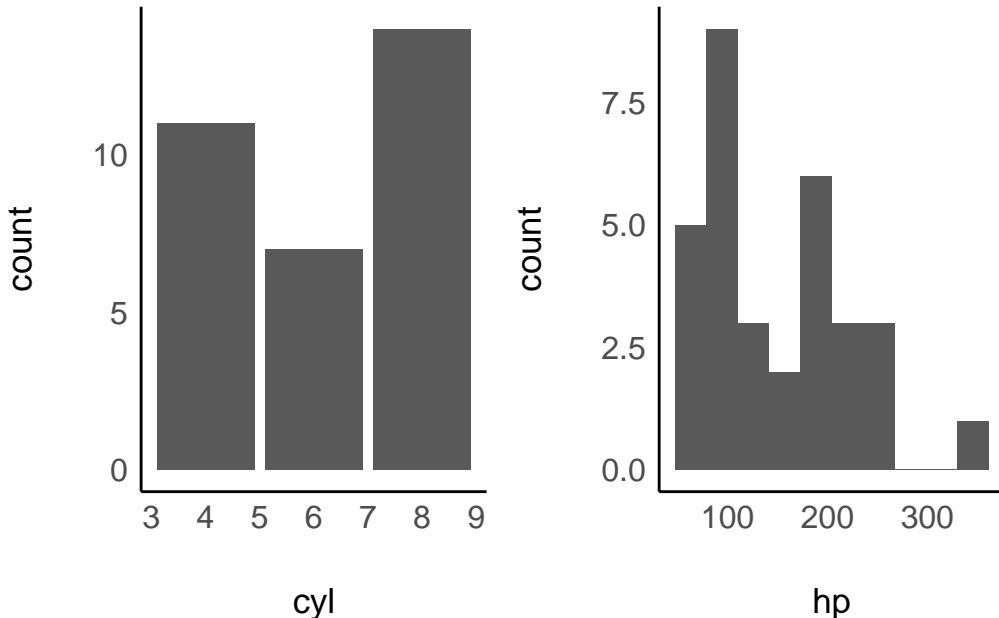


Abbildung 4.1: Häufigkeitsverteilung von cyl und hp (diskretisiert in 10 Körbe oder Gruppen)

...:

Ein *stetiges* Merkmal, wie hp (PS-Zahl), lässt sich durch Klassenbildung diskretisieren, s. Abb. Abbildung 4.1, rechts.

4.2.2 Wahrscheinlichkeitsverteilung

Wahrscheinlichkeitsverteilungen dienen dazu, Ereignissen einer Zufallsvariable eine Wahrscheinlichkeit zuzuordnen.

Eine *diskrete* Wahrscheinlichkeitsverteilung der (diskreten) Zufallsvariablen X ordnet jeder der k Ausprägungen $X = x$ eine Wahrscheinlichkeit p zu. So hat die Variable *Geschlecht eines Babies* die beiden Ausprägungen *Mädchen* und *Junge* mit den Wahrscheinlichkeiten $p_M = 51.2\%$ bzw. $p_J = 48.8\%$ (Gelman, Hill, und Vehtari 2021).

Bei *stetigen* Zufallsvariablen X geht man von unendlich vielen Ausprägungen aus; die Wahrscheinlichkeit einer bestimmten Ausprägung ist (praktisch) Null: $p(X = x_j) = 0$, $j = 1, \dots, +\infty$. So ist die Wahrscheinlichkeit, dass eine Person exakt 166,66666666... cm groß ist, (praktisch) Null. Man gibt stattdessen die *Dichte* der Wahrscheinlichkeit an: Das ist die Wahrscheinlichkeit(smasse) pro Einheit von X .

4.3 Gleichverteilung

4.3.1 Indifferenz als Grundlage

Eine Gleichverteilung nimmt an, dass jeder Wert im Ergebnisraum der zugehörigen Zufallsvariable *gleichwahrscheinlich* ist. Wenn man keinen hinreichenden Grund hat, eine Realisation einer Zufallsvariablen für plausibler als einen anderen zu halten, ist eine Gleichverteilung eine passende Verteilung. Gleichverteilungen gibt es im diskreten und im stetigen Fall.

Abb. Abbildung 4.2 zeigt ein Beispiel für eine (stetige) Gleichverteilung.

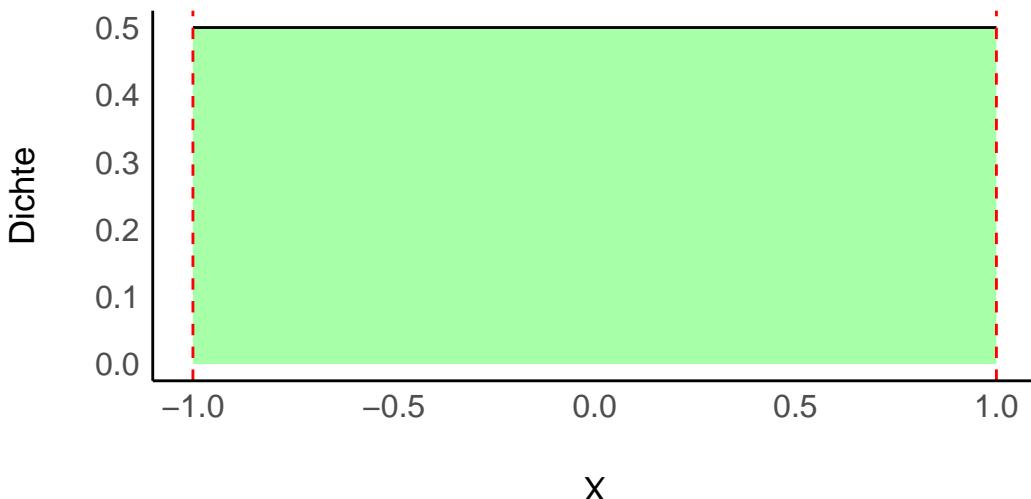


Abbildung 4.2: Gleichverteilung min=-1, max=1

Bei $X = 0$ hat eine Einheit von X die Wahrscheinlichkeitsmasse von 50%.

Definierendes Kennzeichen einer Gleichverteilung ist die *konstante Dichte*.

4.3.2 Simulation

Möchte man die Verteilungsfunktion einer stetigen Zufallsvariablen berechnen, kann das ganz schön kompliziert werden, schließlich muss man Integrale lösen. Aber es gibt einen Trick, wie man die Sache stark vereinfachen kann: man simuliert die Verteilung. Was bedeutet das?

Angenommen, die Wartezeit auf einen Bus ist gleichverteilt (engl. *uniform distribution*); der Bus kommt regelmäßig und pünktlich alle 10 Minuten. Die minimale Wartezeit beträgt also 0 Minuten und die maximale 10 Minuten. Nennen wir die zugehörige Zufallsvariable X , das ist schön kurz.

Dann schreibt man auch:

$$X \sim Unif(0, 10).$$

Ja, das sieht fancy aus, aber wo ist der versprochene Trick zum Vereinfachen? Kommt gleich, Moment.

Eine Frage könnte nun lauten, wie groß ist die Wahrscheinlichkeit, dass man zwischen 3 und 5 Minuten auf den Bus warten muss?

Der Trick ist, dass wir Integralrechnung gegen stumpfes Zählen eintauschen.

Computer (und damit R) haben eingebaut Funktionen, die eine beliebige Zufallszahl ziehen können, zum Beispiel gleichverteilt

Auf Errisch heißt das Zauberwort `runif()`:

```
runif(n = 1, min = 0, max = 10)  
  
## [1] 9.14806
```

Auf Deutsch heißt das: "Hey R, ich hätte gerne eine (daher `n = 1`) Zufallszahl r wie *random*, die gleichverteilt ist (*uniform*) mit `min = 0` und `max = 10`.

(Zu) anschaulich gesprochen: R hat den Bus kommen lassen und es hat gut 9.1 Minuten gedauert, bis er da war.

Achtung, jetzt kommt's: Jetzt lassen wir R mal 10^5 (`1e5` auf Computersprech) Busse vorfahren. R soll jedes Mal notieren, wie lange man auf den Bus warten musste.¹

```
x_simu <- runif(n = 1e5, min = 0, max = 10)
```

Schauen wir uns die Verteilung an, Abbildung 4.3.

```
ggplot(x_simu_df) +  
  aes(x = x_simu) +  
  geom_histogram(bins = 50)
```

¹Machen Sie das mal ohne Computer, wenn Sie ein Wochenende Langeweile haben.

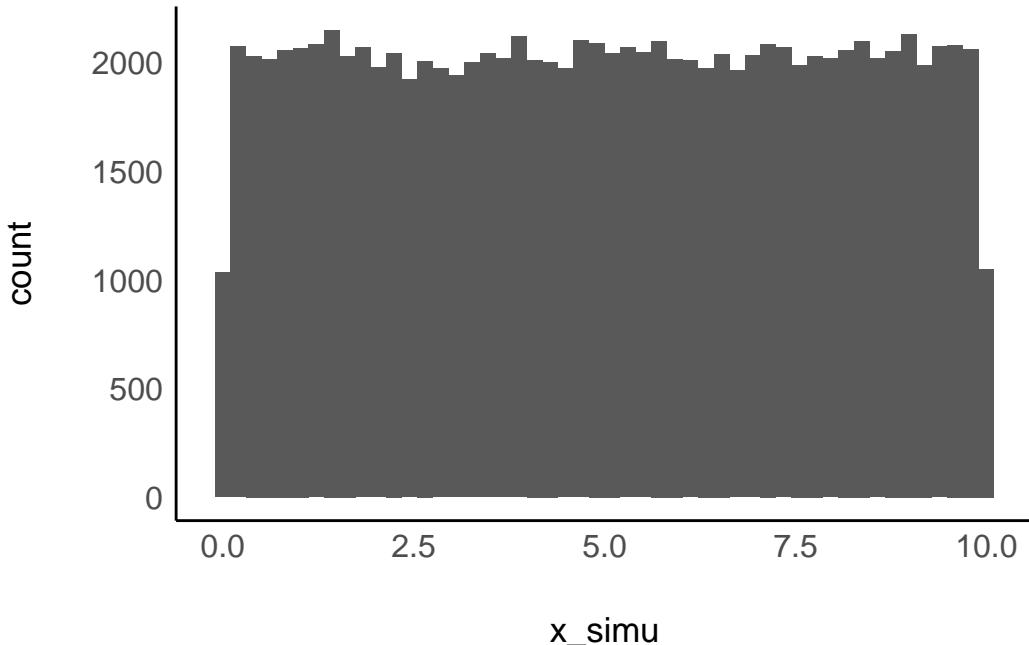


Abbildung 4.3: Simulation einer gleichverteilten Zufallsvariablen

Okay, unsere Verteilung sieht nicht *exakt* gleichverteilt, aber einigermaßen. Gut genug für unsere Zwecke!

So, und jetzt kommt das Ernten. Wir können jetzt nämlich einfach zählen (`count()`), um die Antwort auf unsere Frage (der Wartezeit 3-5 Min.) zu erhalten.

```
x_simu_df %>%
  count(Schnittmenge = x > 3 & x < 5)
## # A tibble: 2 x 2
##   Schnittmenge     n
##   <lgl>      <int>
## 1 FALSE        80072
## 2 TRUE         19928
```

Das Zeichen `&` ist das logische UND, also die Schnittmenge der zwei Mengen $A := \{x|x > 3\}$ und $B := \{x|x < 5\}$, also $A \cap B$.

Wie man sieht, fallen ca. 20% der Stichproben in den entsprechenden Bereich.

Da viele Probleme, wenn sie komplexer werden, kaum noch analytisch (wie Integrieren) ausrechenbar sind, greift man in der modernen (Analyse-)Welt oft lieber auf Simulationsverfahren zurück - Dank sei den schnellen Rechnern. Für uns Menschen ist damit die Aufgabe des Integrierens auf schnödes Zählen zurückgeführt.

4.4 Binomialverteilung

! Wichtig

Die Binomialverteilung dient zur Darstellung der Wahrscheinlichkeit der Ergebnisse eines wiederholten binomialen Zufallsexperiments, eines Zufallsexperiments mit *zwei* Ergebnissen also. Typisches Beispiel ist ein Münzwurf. Bei jeder Wiederholung des Zufallsexperiments bleibt die Wahrscheinlichkeit der Ergebnisse gleich: Die Münze verändert sich nicht durch die Würfe (Ziehen mit Zurücklegen). Außerdem hat ein bestimmtes Ergebnis im ersten Wurf keinen Einfluss auf die Wahrscheinlichkeit eines bestimmten Ergebnisses im zweiten Wurf, etc.

4.4.1 Veranschaulichung

Stellen wir uns eine Kistchen² mit 5 Losen vor, darunter 2 Treffer (Gewinn) und 3 Nieten, s. Abb. Abbildung 4.4. Der Versuch läuft so ab: Wir ziehen ein Los, schauen ob es ein Treffer ist oder nicht, legen es zurück und ziehen erneut.

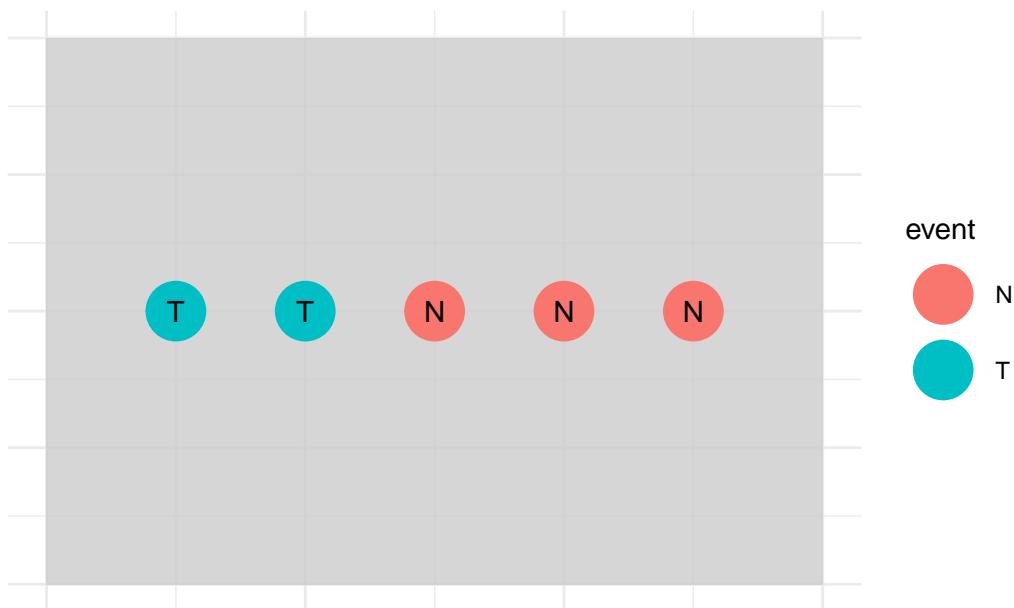


Abbildung 4.4: Ein Kästchen mit 5 Losen, darunter 2 Treffer und 3 Nieten.

²In den Lehrbüchern häufig als Urne bezeichnet, was den bösen Spott von “Friedhofsstatistik” nach sich zog.

! Wichtig

Um die Wahrscheinlichkeitsverteilung einer binomialverteilte Zufallsvariable ausrechnen zu können, muss man zwei Dinge wissen: Erstens die Anzahl der Züge, n (Stichprobengröße) und zweitens die Trefferwahrscheinlichkeit, p .

Wie groß ist die Wahrscheinlichkeit von A' , d.h. bei $n = 4$ Zügen $x = 2$ Treffer zu erzielen, gegeben dass die Trefferwahrscheinlichkeit bei $p = 2/4$ liegt?

Wir könnten jetzt ein Baumdiagramm zeichnen und pro Pfad die Wahrscheinlichkeit ausrechnen (Multiplikationssatz), s. Abbildung 3.15. Die Summe der Wahrscheinlichkeiten der Pfade ist dann die gesuchte Wahrscheinlichkeit, W (Additionssatz). Das ist einfach, dauert aber.

In diesem Fall ist die Wahrscheinlichkeit eines (günstigen) Pfades, A :

$$P(A) = P(T)^2 \cdot P(N)^2 = \left(\frac{2}{5}\right)^2 \cdot \left(\frac{3}{5}\right)^2.$$

```
p_a = (2/5)^2 * (3/5)^2
p_a
## [1] 0.0576
```

Etwas mühevolles Zählen der Pfade würde uns zeigen, dass es $k = 6$ Pfade gibt, die alle die gleiche Wahrscheinlichkeit, $P(A)$, aufweisen.

Damit beträgt die Wahrscheinlichkeit des gesuchten Ereignisses A' (2 Treffer bei 4 Zügen):

$$P(A') = 6 \cdot P(A).$$

```
p_a_strich = 6 * p_a
p_a_strich
## [1] 0.3456
```

Mithilfe der Formel der Binomialverteilung lässt sich das Ergebnis, die Wahrscheinlichkeit von A' schneller ausrechnen. Einfach gesprochen sieht sie so aus:

$$P(A') = k \cdot P(A)$$

Dabei steht k für die Anzahl der günstigen Pfade und $P(A)$ für die Wahrscheinlichkeit eines günstigen Pfades (d.h. 2 Treffer und 2 Nieten) und alle Pfade haben die gleiche Wahrscheinlichkeit.

Die Anzahl der Pfade kann man mit dem *Binomialkoeffizient* ausrechnen, den man so darstellt:

$$\binom{n}{k}$$

Lies: "Wähle aus n möglichen Ereignissen (Pfade im Baum) k günstige Ereignisse (günstige Pfade).

Auf Errisch geht das so:

```
choose(4,2)
## [1] 6
```

4.4.2 Rechnen mit R

Die Binomialverteilung ist in R eingebaut; man kann sich leicht entsprechende Wahrscheinlichkeiten ausrechnen lassen.

Die Wahrscheinlichkeit, bei 4 Zügen 2 Treffer zu erzielen mit $p = 2/5$ unter der Annahme einer Binomialverteilung lässt sich so mit R berechnen:

```
dbinom(x = 2, size = 4, prob = 2/5)
## [1] 0.3456
```

Beispiel 4.1 (Pumpstation-Beispiel zur Binomialverteilung). In einer Pumpstation arbeiten 7 Motoren, die wir als identisch annehmen. Mit einer Wahrscheinlichkeit von 5% fällt ein Motor aus und ist für den Rest des Tages nicht einsatzbereit. Der Betrieb kann aufrecht erhalten werden, solange mindestens 5 Motoren arbeiten. Wie groß ist die Wahrscheinlichkeit, dass die Pumpstation aus dem Betrieb fällt?

$P(X = k)$ (oder kurz: $P(k)$) gibt die Wahrscheinlichkeit (Wahrscheinlichkeitsfunktion) an für das Ereignis, dass k Motoren arbeiten.

Lassen wir R mal $P(X = 5)$ ausrechnen.

```
dbinom(x = 5, size = 7, prob = .95)
## [1] 0.0406235
```

Es gilt also $P(X = 5) \approx .04$. Die Wahrscheinlichkeit, dass (nur) 5 Motoren laufen an einem beliebigen Tag ist relativ gering³.

`dbinom()` steht für die Wahrscheinlichkeitsdichte (im diskreten Fall, also hier, Wahrscheinlichkeitsfunktion genannt) und `binom` für die Binomialverteilung. `x` gibt die Anzahl der Treffer an (das gesuchte Ereignis, hier 5 Motoren arbeiten); `size` gibt die Stichprobengröße an (hier 7 Motoren).

Damit gilt:

$$P(X \geq 5) = P(X = 5) + P(X = 6) + P(X = 7)$$

³wobei “gering” subjektiv ist, die Betreiberfirma findet diese Wahrscheinlichkeit, dass 2 Pumpen ausfallen, wohl viel zu hoch.

```

p_5 <- dbinom(x = 5, size = 7, prob = .95)
p_6 <- dbinom(x = 6, size = 7, prob = .95)
p_7 <- dbinom(x = 7, size = 7, prob = .95)

p_mind_5 <- p_5 + p_6 + p_7

p_mind_5
## [1] 0.996243

```

Die Wahrscheinlichkeit, dass mind. 5 Motoren arbeiten beträgt also 0.9962.

Das Komplement zu diesem Ereignis ist, dass *nicht* mind. 5 Motoren arbeiten, also höchstens 4 und es daher zu einem Ausfall kommt.

Natürlich gilt $P(\bar{X}) = 1 - P(X)$.

```

p_weniger_als_4 <- 1 - p_mind_5
p_weniger_als_4
## [1] 0.003757043

```

Alternativ kann man mit der Verteilungsfunktion rechnen: $P(X \leq 4)$.

In R kann man dafür die Funktion `pbinom()` nutzen (p für (kumulierte) Wahrscheinlichkeit).

```

pbinom(q = 4, size = 7, prob = .95)
## [1] 0.003757043

```

`q = 4` steht für $X \leq 4$, also für höchstens 4 Treffer (arbeitende Motoren); `size = 7` meint die Stichprobengröße, hier 7 Motoren.

! Wichtig

Die Funktion, die die Wahrscheinlichkeit dafür angibt, dass die diskrete Zufallsvariable X eine Realisation annimmt, die kleiner oder gleich (höchstens) einem Wert $X = x$ ist, heißt *Verteilungsfunktion*.

$$F(X = x) = P(X \leq x)$$

4.4.3 Simulieren

Die Binomialverteilung lässt sich gut als “Münzwurf-Verteilung” auffassen.

Werfen wir eine Münze und sehen wir, was passiert.

```

sample(x = c(0, 1), size = 1)
## [1] 1

```

Mit `sample()` ziehen wir eine Stichprobe aus dem Ereignisraum `x`, hier 0 und 1. Dabei vereinbaren wir (willkürlich), dass 0 für “Kopf” steht und 1 für “Zahl”. `size = 1` bedeutet, wir werfen die Münze ein Mal (d.h. Stichprobengröße `size` ist 1).

Okay, noch an Bord? Dann werfen wir die Münze 10 Mal:

```
sample(x = c(0, 1), size = 10, replace = TRUE)
## [1] 0 1 1 1 1 0 0 1 0 1
```

`replace = TRUE` heißt, wir legen die Münze wieder zurück auf den Tisch, wenn wir sie geworfen haben. Oder anders ausgedrückt: *Ziehen mit Zurücklegen*.

R, mach dich bereit, wirf die Münze 1000 ($n = 10^3$ oder `1e3`) Mal⁴:

```
n <- 1e3

muenze_oft <-
  sample(x = c(0, 1), size = n, replace = TRUE)

muenze_oft %>%
  sum()
## [1] 539
```

Mit `sum()` nach dem Pfeifensymbol `%>%` haben wir aus dem Vektor `muenze_oft`, der aus der ersten Zeile resultiert, die Summe ausgerechnet.

Jetzt wissen wir, wie oft die Münze “Zahl” gezeigt hat, nämlich 539 Mal.

i Hinweis

Wenn Sie einen Zufallsversuch wiederholen, muss nicht jedes Mal das gleiche Ergebnis resultieren. Entsprechend wird bei wiederholten Ausführung der Funktion `sample()` nicht immer das gleiche Ergebnis resultieren. Wundern Sie sich also nicht, wenn bei Ihrem Computer eine ähnliche, aber nicht gleiche, Zahl herauskommt.

Visualisieren wir mal unsere Münzwürfe. Dazu erstellen wir zuerst eine geeignete Tabelle:

```
muenz_tab <-
  tibble(
    id = 1:n,
    x = muenze_oft,
    x_cumsum = cumsum(x) / id # gibt Anteil von "Zahl" wieder
  )

head(muenz_tab)
```

⁴R meckert nicht bei langweiligen Aufgaben.

```

## # A tibble: 6 x 3
##       id     x x_cumsum
##   <int> <dbl>    <dbl>
## 1     1     1        1
## 2     2     1        1
## 3     3     0      0.667
## 4     4     0        0.5
## 5     5     1        0.6
## 6     6     1      0.667

```

Und hier der Anteil von “Zahl” im Verlauf unserer Münzwürfe, s. Abbildung 4.5.

```

muenz_tab %>%
  slice_head(n = 1e3) %>%
  ggplot() +
  aes(x = id, y = x_cumsum) +
  geom_line()

```

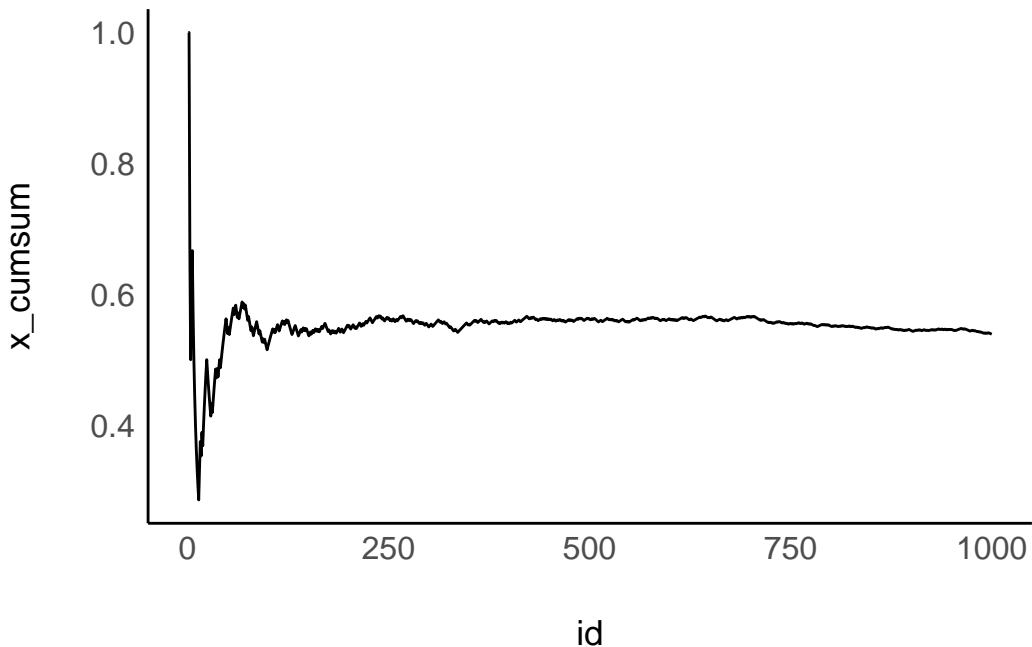


Abbildung 4.5: Das Gesetz der großen Zahl am Beispiel der Stabilisierung des Trefferanteils beim wiederholten Münzwurf

Grob gesagt scheint sich ein Münzwurf nach, naja, vielleicht 500 Würfen “einigermaßen” zu stabilisieren.⁵

⁵Was “einigermaßen” bedeuten soll, ist kein statistischer Begriff, sondern einer, der im echten Leben von den Menschen beantwortet werden muss, die eine Entscheidung zu treffen haben.

! Wichtig

Das Gesetz der großen Zahl

Zieht man (zufällig) immer mehr Werte aus einer Verteilung (mit endlichem Mittelwert), nähert sich der Mittelwert der Stichprobe immer mehr mit dem Mittelwert (oft als Erwartungswert bezeichnet) der Verteilung an

4.5 Normalverteilung

Normalverteilungen haben eine charakteristische Glockenform. Normalverteilungen können sich unterscheiden in ihrem Mittelwert μ und ihrer Streuung, σ . Diese beiden Größen ("Parameter") determinieren den Graphen einer bestimmten Normalverteilungsfunktion, s. Abbildung 4.6. Sind diese beiden Parameter bekannt, so ist die Dichte jedes beliebigen Datenpunkts (aus dieser Normalverteilung) bestimmt.

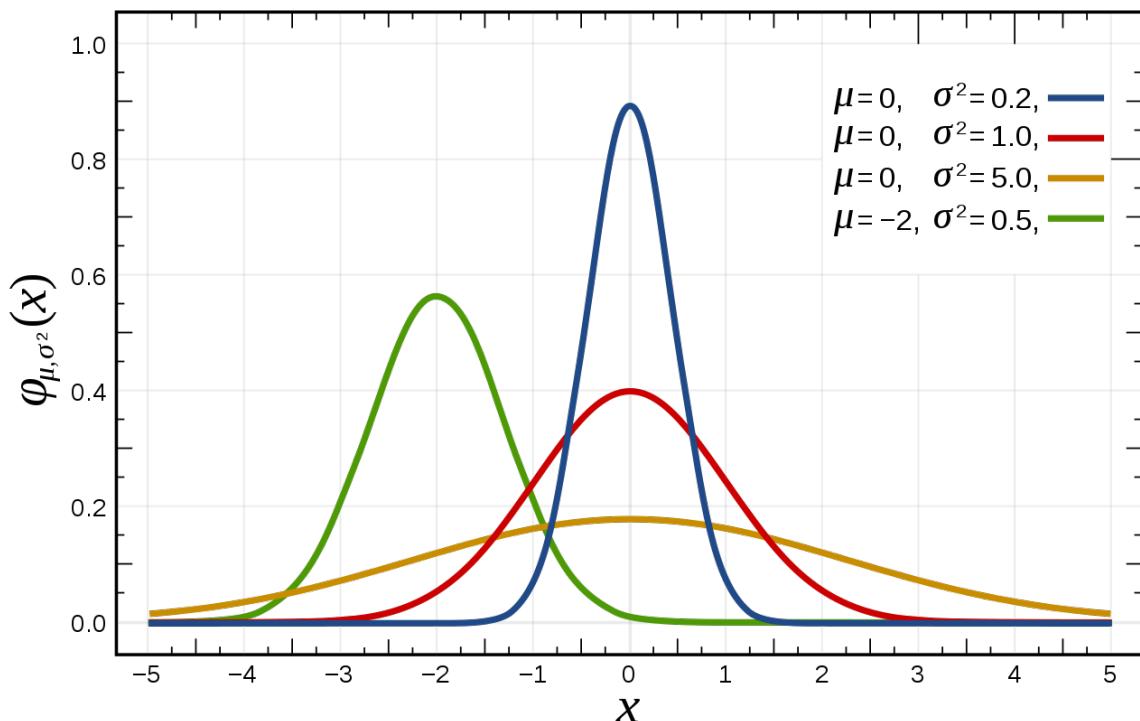


Abbildung 4.6: Beispiele von Normalverteilungen mit verschiedenen Mittelwerten und Streuungen, Quelle: Wikipedia

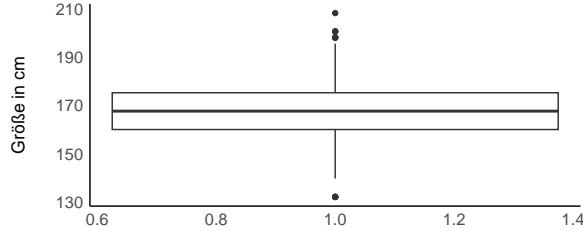
Beispiel: Wie groß sind Studentis ([Quelle des Datensatzes](#))?

Das Quantil von z.B. 25% zeigt die Körpergröße der 25% kleinsten Studentis an, analog für 50%, 75%:

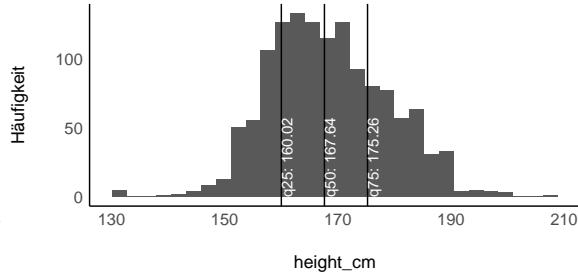
q25	q50	q75
160.02	167.64	175.26

Visualisierung der Quantile:

Die Box zeigt das 25%-, 50%- und 75%-Quantil



Die vertikalen Striche zeigen die Quantile

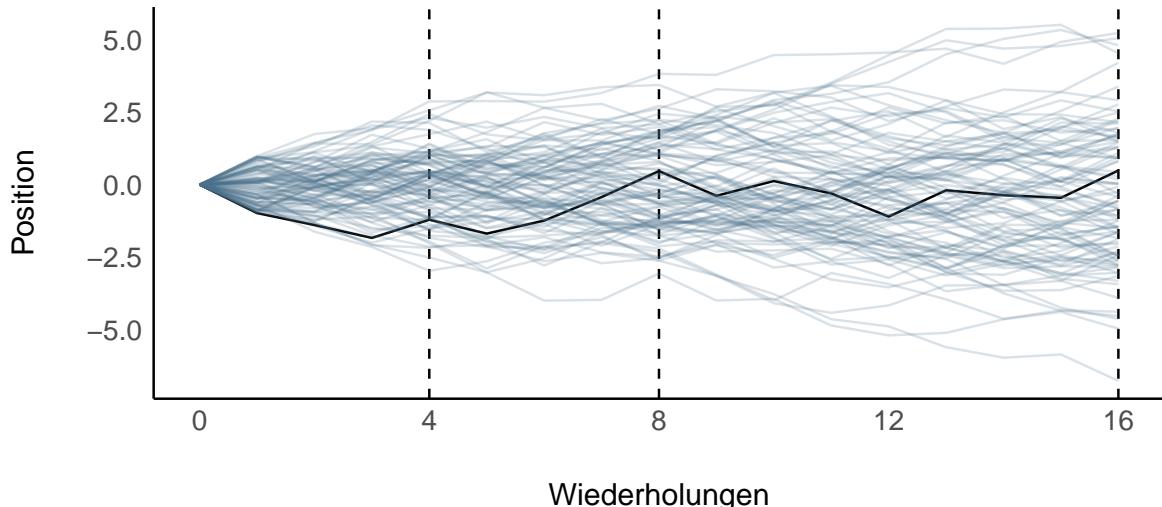


i Hinweis

Das 25%-Quantil nennt man *1. Quartil*, das 50%-Quantil auch *2. Quartil*, das 75%-Quantil das *3. Quartil*, und das 100%-Quantil (Maximalwert) das *4. Quartil*.

4.5.1 Normal auf dem Fußballfeld

Sie und 100 Ihrer besten Freunde stehen auf der Mittellinie eines Fußballfelds. Auf Kommando werfen alle jeweils eine Münze; bei Kopf geht man einen Schritt nach links, bei Zahl nach rechts. Das wird 16 Mal wiederholt. Wie wird die Verteilung der Positionen wohl aussehen?



(McElreath 2020)

4.5.2 Normal durch Addieren

Die Summe vieler (gleich starker) Zufallswerte (aus der gleichen Verteilung) erzeugt eine Normalverteilung; egal aus welcher Verteilung die Zufallswerte kommen (Zentraler Grenzwertsatz), vgl. Abbildung 4.7.

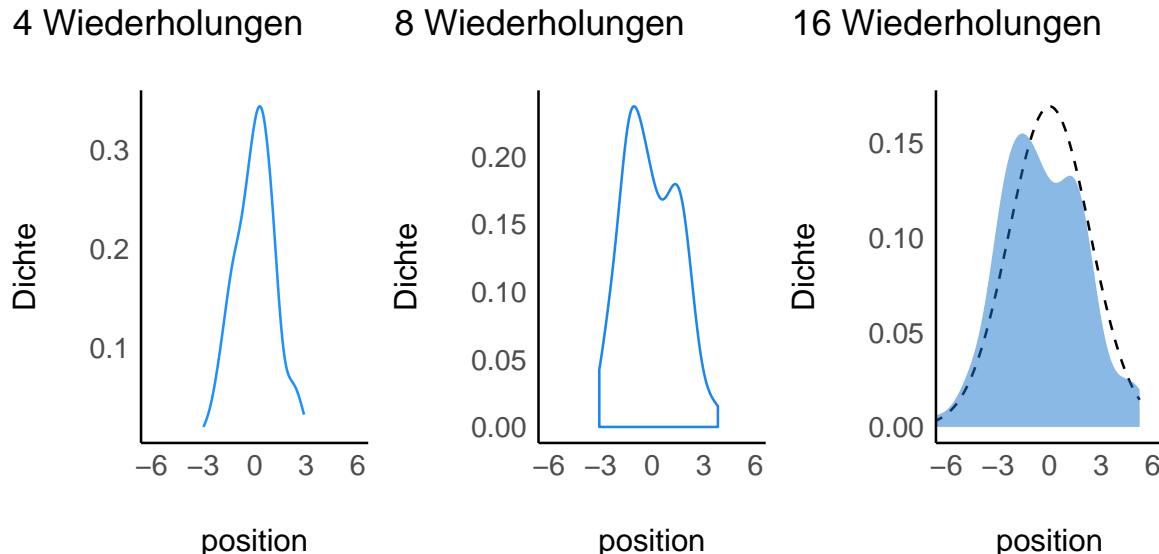
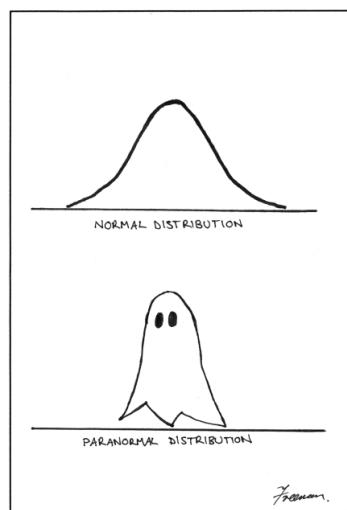
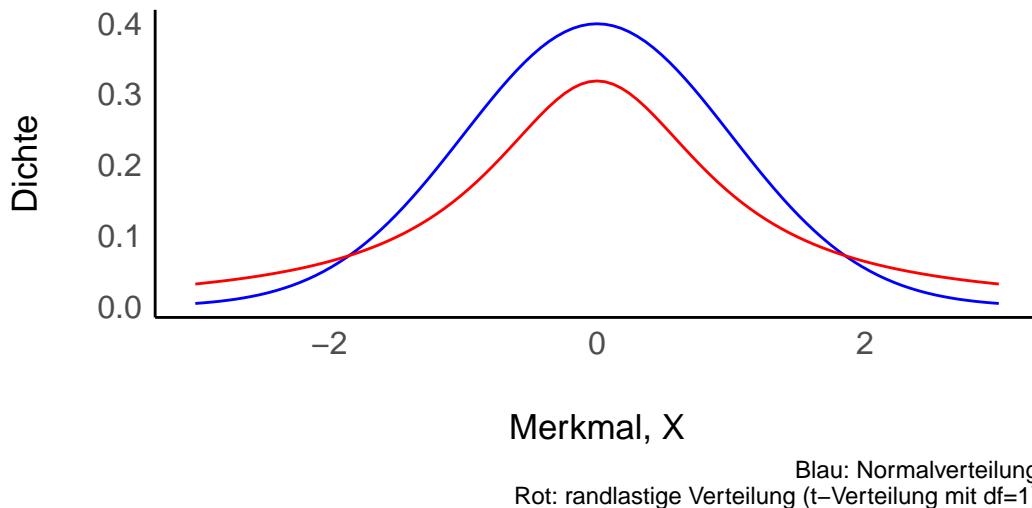


Abbildung 4.7: Entstehen einer Normalverteilung durch Addition vieler unabhängiger Ereignisse

Nicht verwechseln:



4.5.3 Normalverteilung vs. randlastige Verteilungen



Bei randlastigen Verteilungen (“fat tails”) kommen Extremereignisse viel häufiger vor als bei Normalverteilungen. Deshalb ist es wichtig sein, zu wissen, ob eine Normalverteilung oder eine randlastige Verteilung vorliegt. Viele statistische Methoden sind nicht zuverlässig bei (stark) randlastigen Methoden.

4.5.4 Beispiele für Normal- und randlastige Verteilungen

Normal verteilt:

- Größe
- Münzwürfe
- Gewicht
- IQ
- Blutdruck
- Ausschuss einer Maschine

Randlastig verteilt:

- Vermögen
- Verkaufte Bücher
- Ruhm
- Aktienkurse
- Erdbeben
- Pandemien
- Kriege
- Erfolg auf Tinder
- Meteoritengröße
- Stadtgrößen

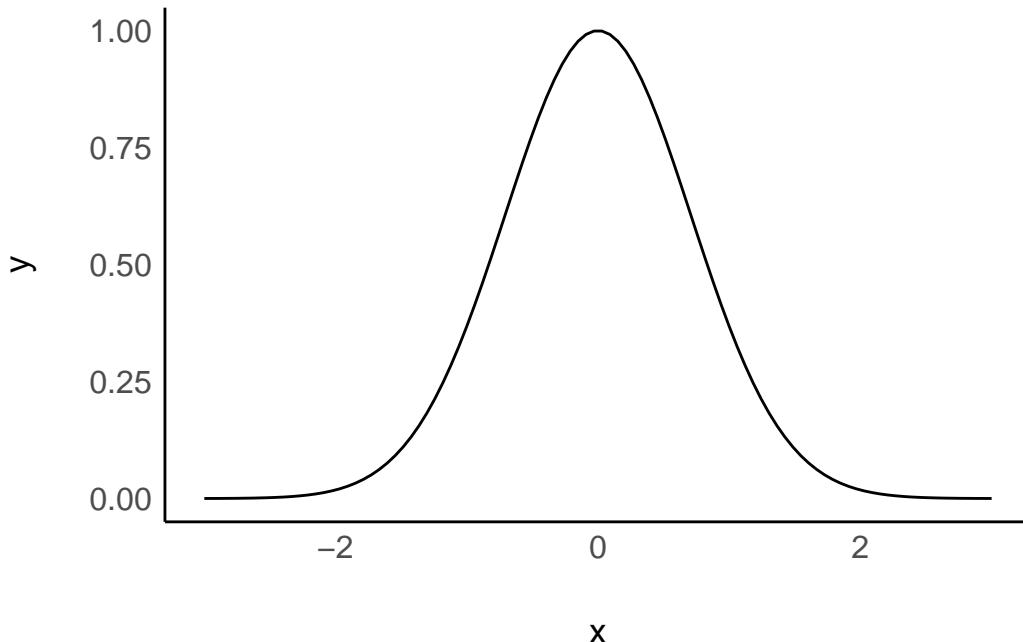
4.5.5 Formel der Normalverteilung

Vereinfacht ausgedrückt lässt die Normalverteilung \mathcal{N} durch Exponenzieren einer Quadratfunktion beschreiben:

$$\mathcal{N} \propto e^{-x^2}$$

mit $e = 2.71\dots$, der Eulerschen Zahl.⁶

```
d <-  
  tibble(  
    x = seq(-3, 3,  
            length.out = 100),  
    y = exp(-x^2)  
)  
  
d %>%  
  ggplot() +  
  aes(x = x, y = y) +  
  geom_line()
```



Eine Normalverteilung mit $\mu = 0$ und $\sigma = 1$ nennt man auch *Standardnormalverteilung* und man schreibt:

$$IQ \sim \mathcal{N}(0, 1)$$

⁶Das Zeichen $y \propto x$ bedeutet "x ist proportional zu y", also $y = mx$.

Die Normalverteilung wird auch *Causs*-Verteilung oder Glockenkurve genannt.

4.5.6 Simulation einer Normalverteilung

R hat eine Funktion eingebaut zur Erzeugung von Zufallszahlen (Zufallszahlengenerator), z.B. `normalverteilt`. Man über gibt dieser Funktion den gewünschten Mittelwert und die gewünschte Streuung und die Funktion zieht dann zufällig Werte aus dieser Verteilung.

Diesen Zufallszahlengenerator kann man mit einem Duschkopf vergleichen, s. Abbildung 4.8. An diesem Duschkopf kann man einen Schwenker einstellen, der den Duschkopf ausrichtet, also steuert, ob die Wassertropfen weit in die eine oder die andere Richtung fallen. Zweitens hat unser Duschkopf noch einen Streuregler, der den Wasserstrahl entweder eng bündelt⁷ oder weit auseinanderfächer. Im ersten Fall fällt der Wasserstrahl eng und schmal aus. Im zweiten Fall fällt der Wasserstrahl breit aus.

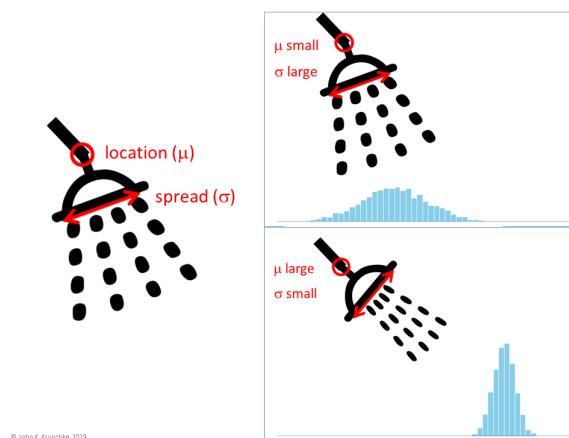


Abbildung 4.8: Zufallszahlengenerator als Duschkopf

Quelle: John Kruschke.

Eine Zufallszahl (*random number*), die *normalverteilt* ist, mit $\mu = 0$ und $\sigma = 1$ kann man in R so erzeugen:

```
rnorm(n = 1, mean = 0, sd = 1)
## [1] 0.2664096
```

Ein Fallbeispiel: Der Inhalt einer Tüte mit Zucker, X , sei normalverteilt mit $\mu = 10002$ g und $\sigma = 1.5$ g. Aus vertragsrechtlichen Gründen darf das Füllgewicht von 1000g nicht unterschritten werden, sonst drohen Konventionalstrafen.

Wie groß ist die Wahrscheinlichkeit, dass 1000g unterschritten werden?

Simulieren wir uns 1e4 Zuckertüten!

⁷Massagedusche, behauptet der Hersteller

```

n <- 1e4
d <-
  tibble(
    id = 1:n,
    x = rnorm(n = n, mean = 1002, sd = 1.5)
  )

head(d)
## # A tibble: 6 x 2
##       id      x
##   <int>  <dbl>
## 1     1 1002.
## 2     2 1001.
## 3     3 1003.
## 4     4 1001.
## 5     5 1001.
## 6     6 1004.

```

Zählen wir, viele der Zuckertüten ein Gewicht von weniger als 1000g aufweisen:

```

d %>%
  count(x < 1000)
## # A tibble: 2 x 2
##   `x < 1000`     n
##   <lgl>     <int>
## 1 FALSE         9013
## 2 TRUE          987

```

Ein ziemlich⁸ kleiner Anteil. Rechnen wir uns noch die Anteile (*proportion*) aus:

```

d %>%
  count(x < 1000) %>%
  mutate(prop = n/1e4)
## # A tibble: 2 x 3
##   `x < 1000`     n   prop
##   <lgl>     <int>  <dbl>
## 1 FALSE         9013 0.901
## 2 TRUE          987  0.0987

```

4.5.7 IQ-Verteilung

Die Verteilung der Zufallsvariablen IQ ist normalverteilt mit einem Mittelwert von 100 und einer Streuung von 15, s. Abbildung 4.9:

⁸“Ziemlich” ist natürlich subjektiv; je nach Situation kann es zu viel oder nicht zu viel sein.

$$IQ \sim \mathcal{N}(100, 15)$$

- Wie schlau muss man sein, um zu den unteren 75%, 50%, 25%, 5%, 1% zu gehören?
- Anders gesagt: Welcher IQ-Wert wird von 75%, 50%, ... der Leute nicht überschritten?

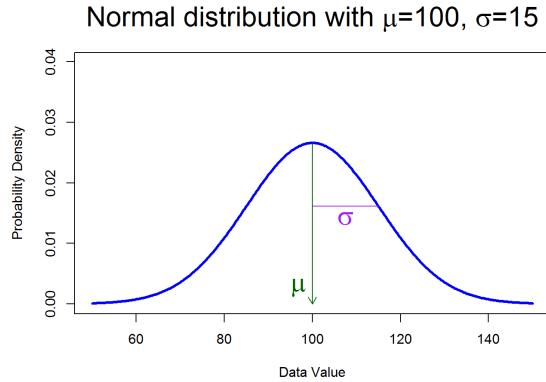


Abbildung 4.9: Visualisierung der theoretischen IQ-Verteilung

Quelle: John Kruschke.

Ziehen wir zufällig 1e4 Stichproben aus $\mathcal{N}(100, 15)$ und berechnen die Quantile:

```
d <-  
  tibble(  
    iq = rnorm(n = 1e4,  
               mean = 100,  
               sd = 15))  
  
probs <- c(0.75,.5,.25,.05,.01)  
  
d_summary <- d %>%  
  summarise(p = probs,  
            q = quantile(iq, probs))
```

p	q
0.75	110
0.50	100
0.25	90
0.05	75
0.01	65

Das *Quantil* q zur kumulierten Wahrscheinlichkeit $p = 75$ ist 110, etc.

Umgekehrt können wir uns auch fragen: Gegeben einer Realisation der Zufallsvariablen (z.B. IQ), was ist die zugehörige Wahrscheinlichkeit (Wert der Verteilungsfunktion?)

- Welcher Anteil der Fläche unter der Kurve p gehört zu den IQ-Werten 75, 100, 115, 130?
- Anders gesagt: Welcher Anteil der Wahrscheinlichkeitsmasse der Verteilung liegt unter $\text{IQ}=75$, $\text{IQ}=100$, etc.?

Ziehen wir Stichproben aus $\mathcal{N}(100, 15)$:

```
d <-  
  tibble(  
    iq = rnorm(1e4,  
               mean = 100,  
               sd = 15)) %>%  
  mutate(iq = round(iq))  
  
qs <- c(75, 100, 115, 130)  
  
d %>%  
  count(p_100 = iq < 100) %>%  
  mutate(prop = n / sum(n))
```

	p_100	n	prop
FALSE	5143	0.51	
TRUE	4857	0.49	

Anstelle von $\text{iq} < 100$ kann man $\text{iq} < 115$ einsetzen, etc.

Die *Verteilungsfunktion* (der Anteil der Wahrscheinlichkeitsmasse), p , für IQ-Werte nicht größer als 100, $\text{IQ} \leq 100$, ist 50%, etc.

4.5.8 Quantile der Normalverteilung

- *Quantile* teilen eine Verteilung so ein, dass ein Anteil p kleiner oder gleich und der andere Teil $1 - p$ größer dem Quantil q ist.
 - *Beispiel:* “50%-Quantil = 100” meint, dass 50% der Elemente der Verteilung einen Wert kleiner oder gleich als 100 haben.
- Die *Verteilungsfunktion* F (für einen Wert x) gibt die Wahrscheinlichkeit an, dass die zugehörige Zufallsvariable X einen Wert höchstens so groß wie x annimmt. Sie zeigt also die kumulierte Wahrscheinlichkeit $[-\infty, q]$.
 - *Beispiel:* “ $F(100) = 50\%$ ” meint: Die Wahrscheinlichkeit für eine Ausprägung von höchstens als 100 beträgt 50%.

50%-Quantil: 100; Verteilung 95%-Quantil: 125; Verteilung

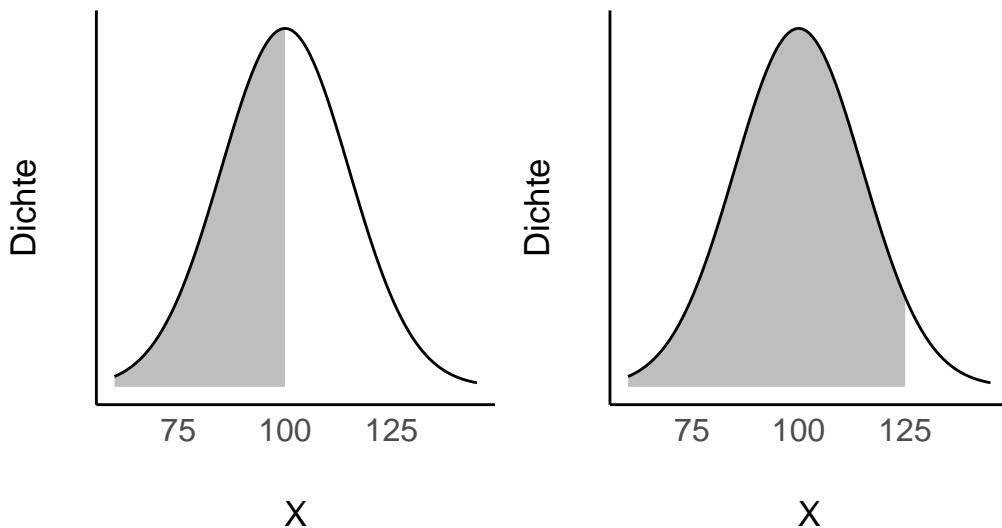


Abbildung 4.10: Quantile der Normalverteilung

Schauen wir uns die Quartile der Normalverteilung einmal näher an. Wir gehen von einer Normalverteilung aus, wie sie zur Beschreibung von Intelligenz (IQ) verwendet wird, s. Abbildung 4.10.

$$IQ \sim \mathcal{N}(100, 15)$$

Mit R kann man sich die beiden Größen komfortabel berechnen lassen:

```
qnorm(.50, mean = 100, sd = 15) # 50%-Quantil
pnorm(100, mean = 100, sd = 15) # Verteilungsfunktion für IQ=100
```

Betrachten wir einige wichtige Quantile, s. Abbildung 4.11.

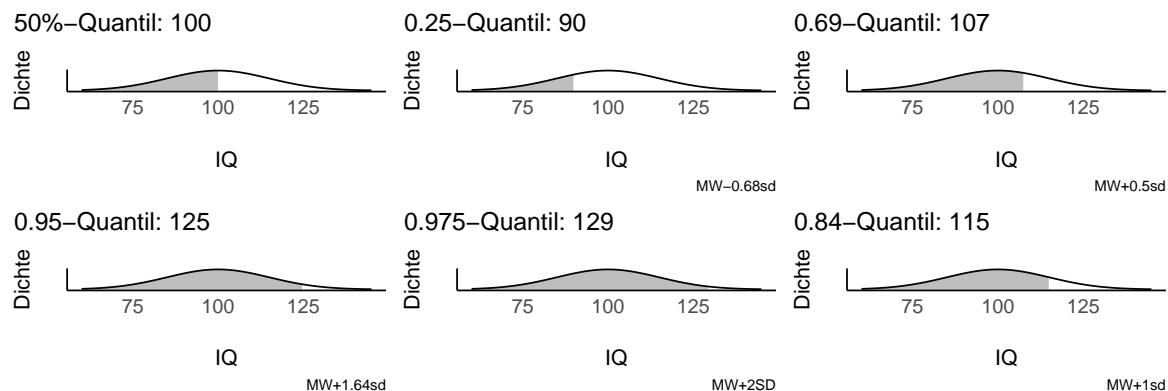
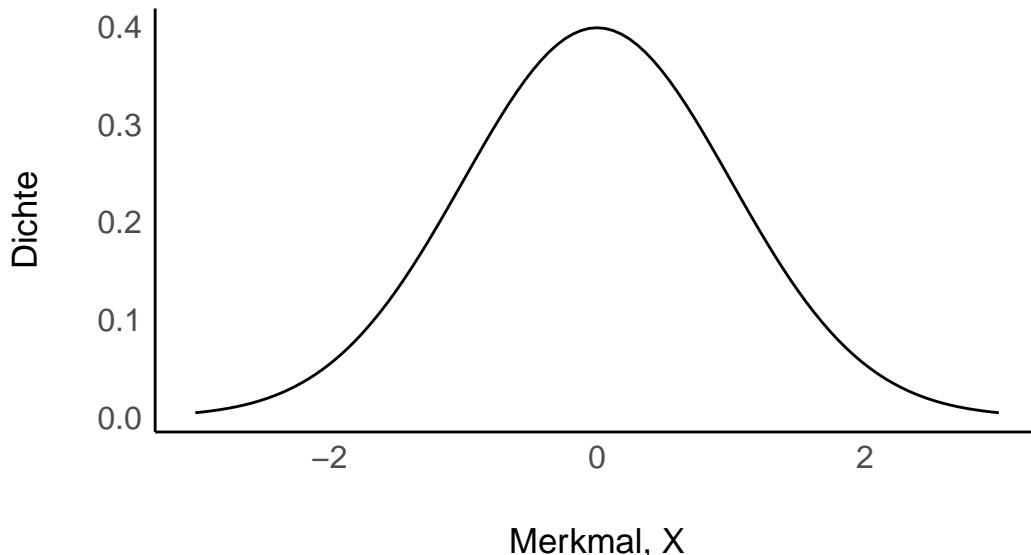


Abbildung 4.11: Verschiedene Quantile der Normalverteilung

4.5.9 Standardnormalverteilung

Normalverteilung mit Mittelwert 0 und SD 1



Bei $X = 0$:

- hat eine Einheit von X die Wahrscheinlichkeitsmasse von 40% (Wahrscheinlichkeitsdichte)
- sind 50% der Wahrscheinlichkeitsmasse (Fläche unter der Kurve) kleiner als dieser Wert (Verteilungsfunktion).

In Summe liegen 100% der Wahrscheinlichkeitsmasse unter der Kurve.

4.5.10 Normalverteilung als konservative Wahl

Dem Mathematiker [Carl Friedrich Gauss](#) (s. Abbildung 4.12) wird die Ehre zuerkannt, die Normalverteilung eingeführt zu haben.



Abbildung 4.12: Zehn-Mark-Geldschein mit Gauss und Normalverteilung

Quelle: Uni Greifswald, Public domain, via Wikimedia Commons

i Hinweis

Ontologische Begründung

- Wirken viele, gleichstarke Einflüsse additiv zusammen, entsteht eine Normalverteilung (McElreath 2020), Kap. 4.1.4.

Epistemologische Begründung

- Wenn wir nur wissen, dass eine Variable über einen endlichen Mittelwert und eine endliche Varianz verfügt und wir keine weiteren Annahmen treffen bzw. über kein weiteres Vorwissen verfügen, dann ist die Normalverteilung die plausibelste Verteilung (maximale Entropie) (McElreath 2020), Kap. 7 und 10.

4.6 Aufgaben

Zusätzlich zu den Aufgaben im Buch:

- [Lose-Nieten-Binomial-Grid](#)
- [Bsp-Binomial](#)

5 Globusversuch



Abbildung 5.1: Bayes:Start

5.1 Lernsteuerung

5.1.1 Lernziele

Nach Absolvieren des jeweiligen Kapitels sollen folgende Lernziele erreicht sein.

Sie können ...

- Unterschiede zwischen Modellen und der Realität erläutern
- die Binomialverteilung heranziehen, um geeignete (einfache) Modelle zu erstellen
- die weite Einsetzbarkeit anhand mehrerer Beispiele exemplifizieren
- Post-Wahrscheinlichkeiten anhand der Gittermethode berechnen

5.1.2 Benötigte R-Pakete

```
library(tidyverse)
```

5.1.3 Begleitvideos

- [Video zum Thema Globusversuch](#)
- [Video zum Thema Übungen zum Globusversuch](#)

5.2 Von Welten und Golems

5.2.1 Kleine Welt, große Welt

Bekanntlich segelte Kolumbus 1492 los, und entdeckte Amerika¹. Das war aber ein glücklicher Zufall, denn auf seinem Globus existierte Amerika gar nicht. Vielleicht sah sein Globus so aus wie der von Behaim, s. Abb Abbildung 5.2.



Abbildung 5.2: Behaims Globus: Kein Amerika

Quelle: Ernst Ravenstein, [Wikimedia](#), Public Domain

Die *kleine Welt des Modells* entsprach hier nicht *der großen Welt, der echten Erdkugel*.

Das ist ein Beispiel, das zeigt, wie Modellieren schiefgehen kann. Es ist aber auch ein Beispiel für, sagen wir, die Komplexität wissenschaftlicher (und sonstiger) Erkenntnis. Einfach gesagt: Glück gehört halt auch dazu.

i Hinweis

Behaims Globus ist nicht gleich der Erde. Die kleine Welt von Behaims Globus ist nicht die große Welt, ist nicht die Erde.

Was in der kleinen Welt funktioniert, muss nicht in der großen Welt funktionieren. Modelle zeigen immer nur die kleine Welt: Vorsicht vor schnellen Schlüssen und vermeintlicher Gewissheit.

Nennen Sie ein Beispiel, in dem ein Modell nicht (exakt) der Wirklichkeit entspricht!

5.2.2 Der Golem von Prag

Quelle

¹wenn auch nicht als Erster



Abbildung 5.3: Der Golem von Prag

Der [Golem von Prag](#), die Legende einer vom Menschen geschaffene Kreatur mit gewaltiger Kraft, die Befehle wörtlich ausführt, s@fig-golem-prag. Die Geschichte besagt, dass ein Rabbi mit Zauberkräften den Golem aus Lehm erschuf, um die jüdische Bevölkerung der Stadt zu schützen. Bei kluger Führung kann ein Golem Nützliches vollbringen. Bei unüberlegter Verwendung wird er jedoch großen Schaden anrichten.

5.2.3 Wissenschaftliche Modelle sind wie Golems

Golem

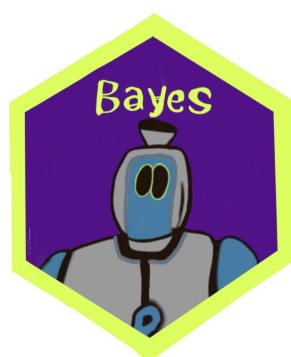


Abbildung 5.4: "Yeah, ich bin ein Golem!" - Bildquelle: Klara Schaumann

Eigenschaften des *Golems*:

- Besteht aus Lehm
- Belebt durch “Wahrheit”
- Mächtig
- dumm
- Führt Befehle wörtlich aus
- Missbrauch leicht möglich
- Märchen

Modell

Eigenschaften eines *Modells*:

- Besteht aus ~~Lehm~~Silikon
- Belebt durch Wahrheit (?)
- Manchmal mächtig
- simpler als die Realität
- Führt Befehle wörtlich aus
- Missbrauch leicht möglich
- Nicht einmal falsch

i Hinweis

Wir bauen Golems.

Abbildung 2.5 stellt ein Sinnbild von Modellen dar.

Vergleichen wir die kleine Welt unserer Modelle, wie Behaims Globus, mit der Großen Welt, die Kolumbus und wir befahren.

Tabelle 5.1: Kleine Welt vs. große Welt

Kleine Welt	Große Welt
Die Welt, wie sie der Golem sieht ist das Modell, aber nicht (zwangsläufig) die Wirklichkeit	Die Welt, wie sie in Wirklichkeit ist entspricht nicht (zwangsläufig) dem Modell
Verwenden wir beim Modellieren	Ist das, was wir modellieren

5.2.4 So denkt unser Bayes-Golem



Abbildung 5.5: So denkt unser Bayes-Golem

Bayes-Inferenz ähnelt dem Lernen von Menschen. Geben Sie ein Beispiel von Lernen bei Menschen, das oben dargestelltem Prozess ähnelt!

5.3 Ein erster Versuch: Wir werfen den Globus

5.3.1 Welcher Anteil der Erdoberfläche ist mit Wasser bedeckt?

Unsere Forschungsfrage lautet, mit welchem Anteil die Erde wohl mit Wasser bedeckt ist (Abbildung 5.6)?



Abbildung 5.6: Die Erde

[Quelle](#) CC 4.0 BY-NC

Analog können wir uns vorstellen, 11 Wissenschaftlis haben jeweils eine andere Hypothese zum Wasseranteil, π , der Erde. Die erste Person hat die Hypothese $\pi_1 = 0$, die zweite Person geht von $\pi_2 = 0.1$ aus ... die 11. Person von $\pi_{11} = 1$.

Um die Forschungsfrage zu beantworten, werfen Sie einen Globus-Ball in die Luft und fangen wieder auf. Sie notieren dann, ob die Stelle unter Ihrem Zeigefinger Wasser zeigt (W) oder Land (L). Den Versuch wiederholen Sie, bis Sie den Globusball insgesamt 9 Mal geworfen haben.

So sah [mein](#)² Ergebnis aus:

W L W W W L W L W

Besorgen Sie sich einen Globus (zur Not eine Münze) und stellen Sie den Versuch nach!

5.3.2 Wie entstanden die Daten?

Der physikalische Prozess, der zur Entstehung der Daten führt, nennt man den *datengenerierenden Prozess*.

In diesem Fall kann man ihn so beschreiben:

1. Der wahre Anteil von Wasser, W , der Erdoberfläche ist p (und $1 - p$ ist der Anteil Land, L).
2. Ein Wurf des Globusballes hat die Wahrscheinlichkeit p , eine W -Beobachtung zu erzeugen.

²Ihr Ergebnis kann anders aussehen, schließlich ist es ja Zufall.

3. Die Würfe des Globusballes sind unabhängig voneinander.
4. Wir haben kein Vorwissen über p ; jeder Wert ist uns gleich wahrscheinlich.

Welche Annahmen würden Sie ändern? Welche könnte man wegnehmen? Welche hinzufügen? Was wären die Konsequenzen?

5.3.3 Ein paar Fachbegriffe

- Für jede Hypothese haben wir ein Vorab-Wissen, das die jeweilige Plausibilität der Hypothese angibt: *Priori-Verteilung*.
- Für jede Hypothese (d.h. jeden *Parameterwert* p) möchten wir wie wahrscheinlich die Daten sind (unter der Annahme, dass die Hypothese richtig ist). Das gibt uns den *Likelihood*.
- Dann gewichten wir den Likelihood mit dem Vorabwissen, so dass wir die *Posteriori-Verteilung*³ bekommen.

5.3.4 Bayes-Updates

Der Golem denkt eigentlich ganz vernünftig: Zuerst hat er ein Vorwissen zum Wasseranteil, die dazugehörige Wahrscheinlichkeitsverteilung nennt man *Priori-Verteilung*. In unserem Beispiel ist das Vorwissen recht bescheiden: Jeder Wasseranteil ist ihm gleich plausibel. Als nächstes beschaut sich der Golem die Daten und überlegt, wie wahrscheinlich die Daten sind, wenn man von einer bestimmten Hypothese ausgeht, z.B. dass der Wasseranteil 10% beträgt. Die zugehörige Wahrscheinlichkeit der Daten unter Annahme einer Hypothese nennt man die⁴ *Likelihood*. Als letztes bildet sich der Golem eine abschließende Meinung zur Wahrscheinlichkeit jeder Hypothese. Diese Wahrscheinlichkeitsverteilung nennt man *Posteriori-Verteilung*. Sie berechnet als Gewichtung des Vorwissen mit den neuen Daten. Anders gesagt: Das Vorwissen wird anhand der Erkenntnisse (der Daten) aktualisiert oder geupdatedet, s. Abbildung 5.7.

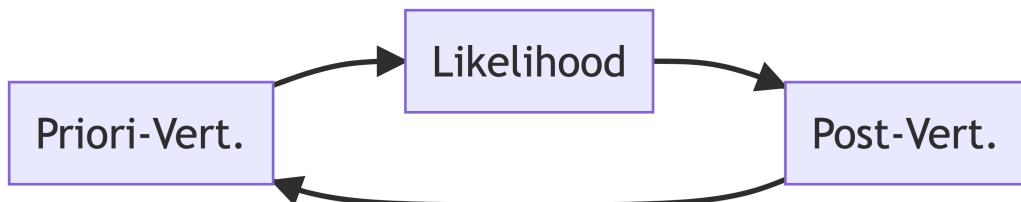


Abbildung 5.7: Updating mit Bayes

³Anstatt von *Priori* liest man auch *Prior*; anstatt *Posteriori* auch *Posterior*

⁴oder den?

5.3.5 Likelihood mit Binomialverteilung

Wie wahrscheinlich ist es, einen bestimmten Wasseranteil, z.B. 6 Treffer, (bei 9) Würfen zu bekommen, wenn man eine bestimmte Hypothese (einen bestimmten Wasseranteil, z.B. 70%) annimmt? Diese Wahrscheinlichkeit hat einen eigenen Namen, weil sie eine wichtige Sache ist. Man nennt sie die *Likelihood*, L^5 .

Geht man von einer Binomialverteilung aus, ist die Likelihood einfach zu berechnen⁶.

Wenn wir eine Binomialverteilung annehmen, dann gehen wir davon aus, dass die Daten unabhängig voneinander entstehen und sich der Parameterwert nicht zwischenzeitlich ändert⁷. Der Wasseranteil der Erde bleibt während des Versuchs gleich (durchaus plausibel).

Lassen Sie uns im Folgenden die Wahrscheinlichkeit (Pr), W mal Wasser und L mal Land zu beobachten, wenn die Wahrscheinlichkeit für Wasser p beträgt, so bezeichnen: $(Pr(W, L|p))$. Diese Wahrscheinlichkeit, $(Pr(W, L|p))$, kann man mit der *Binomialverteilung* berechnen.

Möchte man die Wahrscheinlichkeit ansprechen für das Ereignis “5 mal Wasser und 2 mal Land, wenn wir von einem Wasseranteil von 70% ausgehen”, so würden wir kurz schreiben: $Pr(W = 5, L = 2|p = .7)$.

Die Binomialverteilung zeigt die Verteilung der Häufigkeit (Wahrscheinlichkeit) der Ereignisse (z.B. 2 Mal Kopf) beim wiederholten Münzwurf (und allen vergleichbaren Zufallsexperimenten): “Münzwurfverteilung”, s. Kap. Kapitel 4.4.

Die Formel der Binomialverteilung sieht so aus:

$$Pr(W, L|p) = \frac{(W + L)!}{W!L!} p^W (1 - p)^L = k \cdot P(A) \quad (5.1)$$

Formel Gleichung 5.1 kann wie folgt auf Deutsch übersetzen:

Die Wahrscheinlichkeit für das Ereignis “W,L” gegeben p berechnet als Produkt von zwei Termen. Erstens der Quotient von der Fakultät von W plus L im Zähler und im Nenner das Produkt von erstens der Fakultät von W mit zweitens der Fakultät von L. Der zweite Term ist das Produkt von p hoch W mal der komplementären Wahrscheinlichkeit von p hoch L.

Oder noch kürzer:

Die Wahrscheinlichkeit für das Ereignis “W,L” gegeben p berechnet als Produkt von zwei Termen. Erstens der Anzahl der günstigen Pfade, k und zweitens der Wahrscheinlichkeit für einen günstigen Pfad, $P(A)$.

⁵ L für Freunde alter Schriftarten

⁶ Ein Glück!

⁷ Die sog. “iid-Annahme”, *independently and identically distributed*: Jeder Wurf der Globusballes ist eine Realisation der gleichen Zufallsvariablen. Jeder Wurf ist unabhängig von allen anderen: Das Ergebnis eines Wurfes hat keinen (stochastischen) Einfluss auf ein Ergebnis anderer Würfe. Die Wahrscheinlichkeitsverteilung ist bei jedem Wurf identisch.

Puh, Formeln sind vielleicht doch ganz praktisch, wenn man sich diese lange Übersetzung der Formel in Prosa durchliest. Noch praktischer ist es aber, dass es Rechenmaschinen gibt, die die Formel kennen und für uns ausrechnen. Los, R, mach mal.

5.3.6 Binomialverteilung mit R

Was ist der Anteil der gültigen Pfade in einem Baumdiagramm (Wahrscheinlichkeit), um 2 mal W bei $N = W + L = 3$ Würfen zu bekommen, wenn wir von $p = 1/2$ ausgehen?⁸.

```
dbinom(x = 2, size = 3, prob = 1/2)
## [1] 0.375
```

Von den 8 Endknoten bzw. Pfaden sind 3 günstig. Demnach ist die Wahrscheinlichkeit des gesuchten Ereignis (2 Treffer bei 3 Würfen, binomialverteilt) gleich 3 von 8 (alle Pfade sind gleich wahrscheinlich); 3/8 sind 0.375.

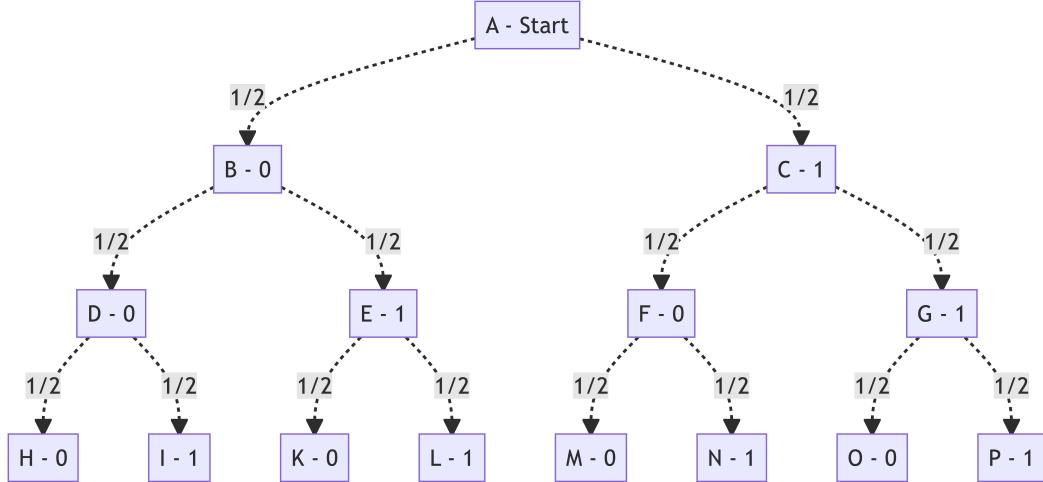


Abbildung 5.8: Wir werfen den Globus (oder eine Münze) 3 Mal

Abb. Abbildung 5.8 stellt einen einfachen Baum für 3 Globuswürfe mit je zwei möglichen Ereignissen (W vs. L) dar. In der ersten (obersten) Zeile (Knoten A; "Start") ist Ausgangspunkt dargestellt: Der Globus ruht wurfbereit in unserer Hand. Jetzt Achtung: Sie werfen den Globusball hoch. Die Pfeile zeigen zu den (zwei) möglichen Ergebnissen. Die zweite Zeile (Knoten B und C) stellt die beiden Ergebnisse des Wurfes dar. Die Ergebnisse sind hier mit 0 und 1 bezeichnet (das eine eine einfache und weiteeinsetzbare Notation). Die dritte Zeile (Knoten D bis G) stellt die Ergebnisse des zweiten Wurfes dar. Die vierte Zeile (Knoten H bis P) stellt die Ergebnisse des dritten Wurfes dar.

⁸Allgemeiner spricht man auch von 2 Treffern bei 3 Würfen (d.h. 1 "Nicht-Treffer", den wir als "Niete" bezeichnen). Treffer werden oft mit 1 und Nieten mit 0 bezeichnet

Für mehr Würfe würde das Diagramm irgendwann unübersichtlich werden.

Was ist der Anteil der gültigen Pfade in einem Baumdiagramm (Wahrscheinlichkeit), um 6 mal W bei $N = W + L = 9$ Würfen zu bekommen, wenn wir von $p = 1/2$ ausgehen?

```
dbinom(x = 6, size = 9, prob = 1/2)
## [1] 0.1640625
```

Abb Abbildung 5.9 ist ein vergeblicher Versuch, so einen großen Baum ($n = 9$) darzustellen.

i Hinweis

Visualisierungen wie Baumdiagramme sind eine praktische Hilfe zum Verständnis, kommen aber bei größeren Daten schnell an ihre Grenze.

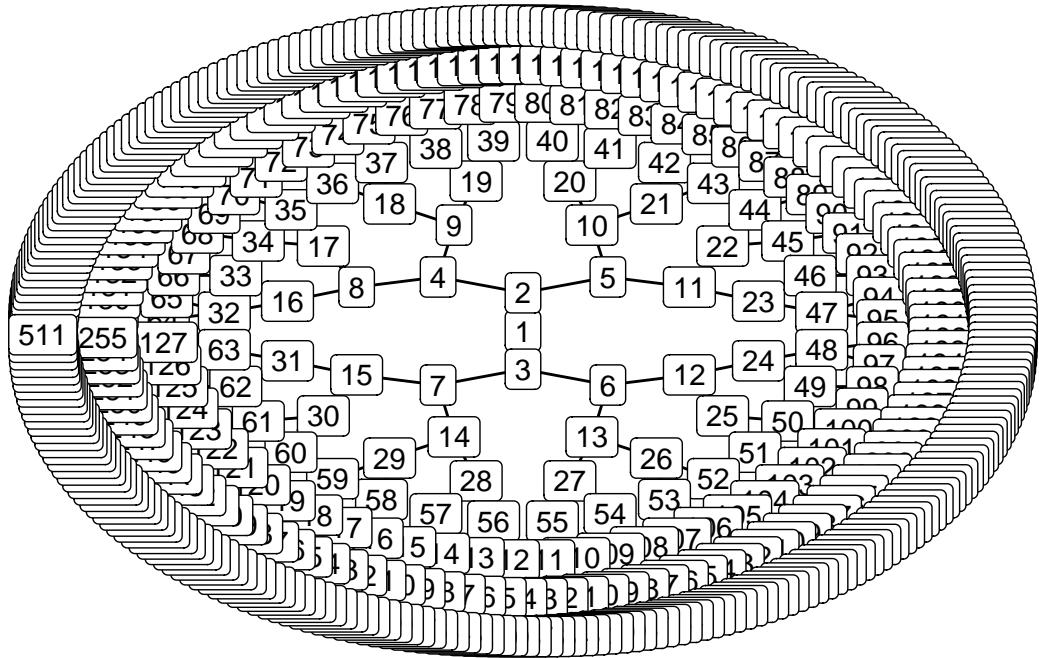


Abbildung 5.9: Wir werfen den Globus (oder eine Münze) 3 Mal

Jetzt folgen einige Beispiele.

Beispiel 5.1 (Globus mit 9 Treffern bei 9 Würfen). Was ist die Wahrscheinlichkeit für $W = 9$ bei $N = 9$ und $p = 1/2$?

```
dbinom(x = 9, size = 9, prob = 1/2)
## [1] 0.001953125
```

Das ist 1 günstiger Pfad von 512 Pfaden.

Beispiel 5.2 (Klausur mit 20-Richtig-Falsch-Fragen). Ein Professi stellt einen Klausur mit 20 Richtig-Falsch-Fragen. Wie groß ist die Wahrscheinlichkeit, durch bloßes Münze werfen genau 15 Fragen richtig zu raten?⁹.

```
dbinom(x = 15, size = 20, prob = .5)
## [1] 0.01478577
```

Um *höchstens* 15 Treffer zu erzielen, müssten wir die Wahrscheinlichkeiten von 0 bis 15 Treffern addieren.

Praktischerweise gibt es einen R-Befehl, der das für uns übernimmt:

```
pbinom(q = 15, size = 20, prob = .5)
## [1] 0.994091
```

Die Wahrscheinlichkeit 0, 1, 2, ... oder 15 Treffer zu erzielen, liegt also bei gut 99%.

Beispiel 5.3 (3 Münzwürfe mit 3 Treffern). Was ist die Wahrscheinlichkeit bei 3 Münzwürfen (genau) 3 Treffer (Kopf) zu erzielen?

Das ist eine Frage an die Binomialverteilung; in R kann man das mit der Funktion `dbinom` beantworten.

```
dbinom(x = 3, size = 3, prob = 1/2)
## [1] 0.125
```

`dbinom` gibt uns die Wahrscheinlichkeit von `x` Treffern, bei `size` Versuchen zurück, wobei eine Binomialverteilung angenommen wird mit Trefferwahrscheinlichkeit `prob`.

5.3.7 Unser Modell ist geboren

Wir fassen das Globusmodell so zusammen:

$$W \sim \text{Bin}(N, p),$$

Lies: "W ist binomial verteilt mit den Parametern N und p ". N gibt die Anzahl der Globuswürfe an: $N = W + L$.

⁹Hey, endlich mal was für echte Leben!

Unser Vorab-Wissen zu p sei, dass uns alle Werte gleich plausibel erscheinen (“uniform”):

$$p \sim \text{Unif}(0, 1).$$

Lies: “ p ist gleich (uniform) verteilt mit der Untergrenze 0 und der Obergrenze 1”.

Man könnte auch sagen: Wir haben praktisch kein Vorwissen, wir sind erstmal (aprior) indifferent, jeder Parameterwert erscheint uns erstmal gleich wahrscheinlich.

5.3.8 Visualisierungen

Abb. Abbildung 5.10 zeigt die Binomialverteilung $X \sim \text{Bin}(9, 1/2)$.

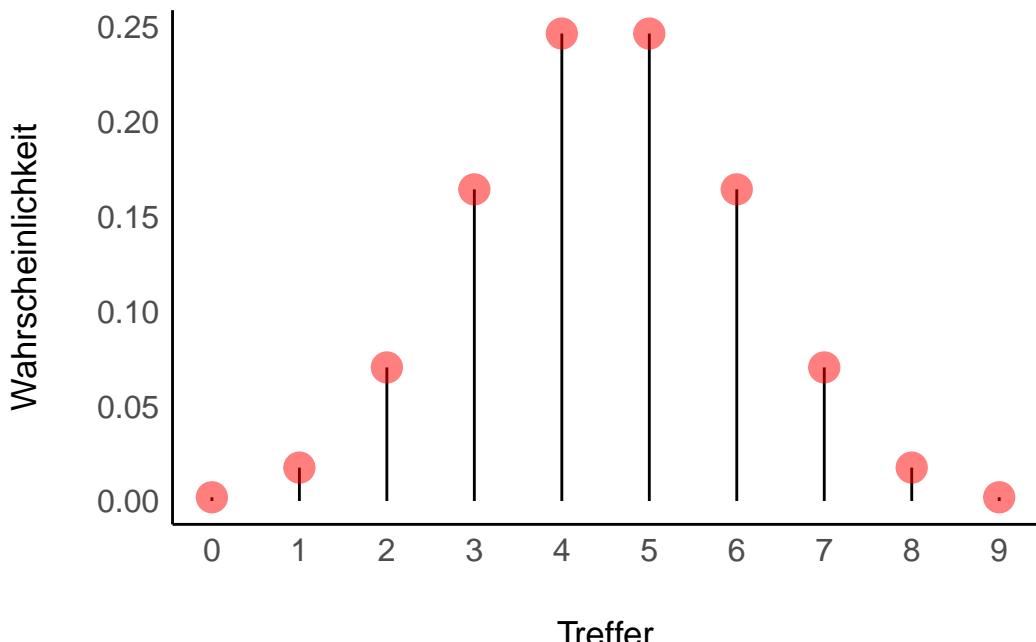


Abbildung 5.10: Ein Beispiel für eine Binomialverteilung mit Parametern $N=9$ und $p=1/2$.

Was fällt Ihnen bei der Binomialverteilung auf? Ist sie symmetrisch? Verändert sich die Wahrscheinlichkeit linear?

5.4 Zur Erinnerung: Bayes Theorem

5.4.1 Herleitung Bayes' Theorem 1/2: Gemeinsame Wahrscheinlichkeit

Die Wahrscheinlichkeit für *Regen* und *kalt* ist gleich der Wahrscheinlichkeit von *Regen, gegeben kalt* mal der Wahrscheinlichkeit von *kalt*. Entsprechend gilt: Die Wahrscheinlichkeit von W, L und p ist das Produkt von $\text{Pr}(W, L|p)$ und der Prior-Wahrscheinlichkeit $\text{Pr}(p)$:

$$Pr(W, L, p) = Pr(W, L|p) \cdot Pr(p)$$

Genauso gilt: Die Wahrscheinlichkeit von *Regen* und *kalt* ist gleich der Wahrscheinlichkeit *kalt, wenn's regnet* mal der Wahrscheinlichkeit von *Regen*:

$$Pr(W, L, p) = Pr(p|W, L) \cdot Pr(W, L)$$

5.4.2 Herleitung Bayes' Theorem 2/2: Posteriori-Wahrscheinlichkeit

Wir setzen die letzten beiden Gleichungen gleich:

$$Pr(W, L|p) \cdot Pr(p) = Pr(p|W, L) \cdot (W, L)$$

Und lösen auf nach der Posteriori-Wahrscheinlichkeit¹⁰, $Pr(p|W, L)$:

$$Pr(p|W, L) = \frac{Pr(W, L|p)Pr(p)}{Pr(W, L)}$$

$Pr(W, L)$ nennt man die *mittlere Wahrscheinlichkeit der Daten* oder *Evidenz*. Die Evidenz berechnet sich als Mittelwert der Likelihoods über alle Werte von p . Die Aufgabe dieser Größe ist nur dafür zu sorgen, dass insgesamt Werte zwischen 0 und 1 herauskommen.

5.4.3 Bayes' Theorem als Formel

$$Pr(H|D) = \frac{Pr(D|H)Pr(H)}{Pr(D)} = \frac{\text{Likelihood} \cdot \text{Priori}}{\text{Evidenz}}$$

- Bestandteile:
 - Posteriori-Wahrscheinlichkeit: $Pr_{Post} := Pr(H|D)$
 - Likelihood: $L := Pr(D|H)$
 - Priori-Wahrscheinlichkeit: $Pr_{Priori} := Pr(H)$
 - Evidenz: $E := Pr(D)$
- Bayes' Theorem gibt die Pr_{Post} an, wenn man die Gleichung mit der Pr_{Priori} und dem L füllt.
- Bayes' Theorem wird häufig verwendet, um die Pr_{Post} zu quantifizieren.
- Die Pr_{Post} ist proportional zu $L \times Pr_{Priori}$.

¹⁰kürzen wir mit Post-Wahrscheinlichkeit or $Pr(Post)$ ab

5.4.4 Posteriori als Produkt von Priori und Likelihood

Die unstandardisierte Post-Wahrscheinlichkeit ist einfach das Produkt von Likelihood und Priori.

Das Standardisieren dient nur dazu, einen Wert zwischen 0 und 1 zu erhalten. Dies erreichen wir, indem wir durch die Summe aller Post-Wahrscheinlichkeiten dividieren. Die Summe der Post-Wahrscheinlichkeiten bezeichnet man (auch) als Evidenz, vgl. Gleichung Gleichung 5.2.

$$\text{Posteriori} = \frac{\text{Likelihood} \times \text{Priori}}{\text{Evidenz}} \quad (5.2)$$

Abb. Abbildung 5.11 visualisiert, dass die Post-Verteilung eine Gewichtung von Priori und Likelihood ist. Mathematisch gesprochen beruht diese Gewichtung auf einer einfachen Multiplikation der beiden genannten Terme.

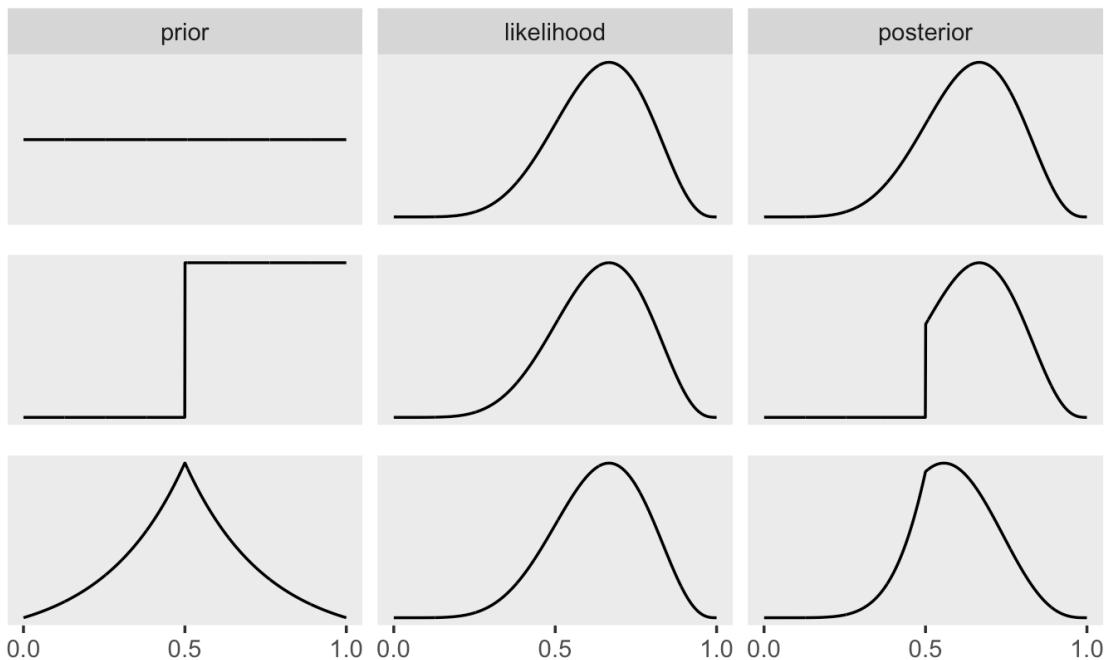


Abbildung 5.11: Prior mal Likelihood = Post

5.4.5 Wissen updaten: Wir füttern Daten in das Modell

Golems können lernen?! Abbildung 5.12 zeigt die Post-Verteilung, nach $n = 1, 2, \dots, n = 9$ Datenpunkten, d.h. Würfen mit dem Globusball. Man sieht: Am Anfang, apriori, also bevor die Daten haben, vor dem ersten Wurf also, ist jeder Parameterwert gleich wahrscheinlich für den Golem (das Modell). Je nach Ergebnis des Wurfes verändert sich die Wahrscheinlichkeit

der Parameterwerte, kurz gesagt, die Post-Verteilung verändert sich in Abhängigkeit von den Daten.

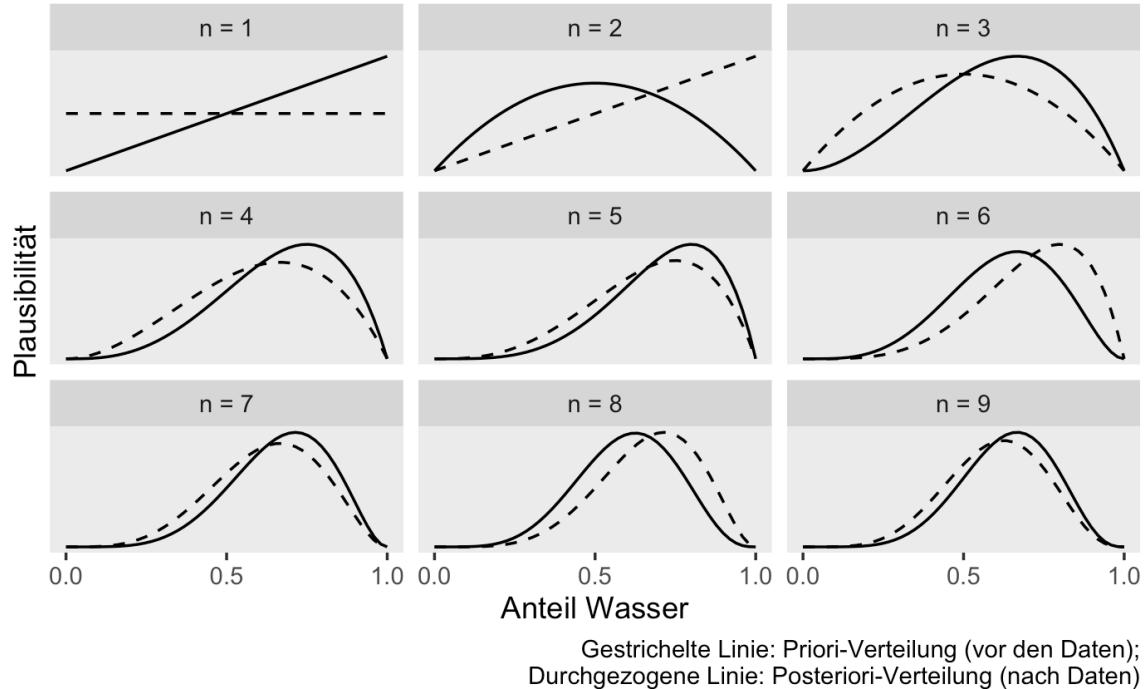


Abbildung 5.12: Unser Golem lernt

Insofern kann man sagen: Unser Golem (das Modell) lernt. Ob das Modell nützlich ist (präzise Vorhersagen liefert), steht auf einem anderen Blatt.

5.5 Bayes berechnen mit mit dem Bayes-Gitter

Wir erstellen uns eine kleine Tabelle, die man “Bayes-Gitter” nennen könnte. Dazu gehen wir so vor:

5.5.1 Idee

1. Teile den Wertebereich des Parameters in ein “Gitter” auf, z.B. 0.1, 0.2, ..., 0.9, 1 (“Gitterwerte”).
2. Wähle den Priori-Wert des Parameters für jeden Gitterwert.
3. Berechne den Likelihood für Gitterwert.
4. Berechne den unstandardisierten Posteriori-Wert für jeden Gitterwert (Produkt von Priori und Likelihood).
5. Standardisiere den Posteriori-Wert durch teilen anhand der Summe aller unstandardisierten Posteriori-Werte.

Für jeden “Gitterwert” berechnen wir eine (Post-)Wahrscheinlichkeit. Ein Gitterwert ist eine mögliche Ausprägung des Parameters. Häufig entspricht eine Hypothese einem Gitterwert, etwa wenn man sagt: “Ich glaube, die Münze ist fair”, was auf einem Parameterwert von 50% herausläuft. Dazu geben wir an, für wie wahrscheinlich wie apriori¹¹ - also bevor wir irgendwelche Daten erheben - jeden einzelnen Gitterwert halten. Wir machen es uns hier einfach und halten jeden Gitterwert für gleich wahrscheinlich. Tatsächlich ist der konkrete Wert hier egal, entscheidend ist das Verhältnis der Apriori-Werte zueinander: Geben wir einigen Gitterwerten den Wert 2, aber anderen den Wert 1, so halten wir Erstere für (apriori) doppelt so plausibel wie Letztere. Der Likelihood wird in diesem Fall mit der Binomialverteilung berechnet. Der Likelihood gibt an, wie wahrscheinlich ein Gitterwert ist gegeben einem bestimmten apriori gewählten Parameterwert. Die “End-Wahrscheinlichkeit”, die unstandardisierte Post-Wahrscheinlichkeit, die “hinten rauskommt” ist das Produkt von Priori-Wert und Likelihood. Anschaulich gesprochen: Die Priori-Werte werden mit den Likelihoodwerten gewichtet¹². Da wir letztlich eine Wahrscheinlichkeitverteilung bekommen möchten teilen wir jeden Posteriori-Wert durch die Summe aller Posteriori-Werte. Dadurch ist garantiert, dass sich die Posteriori-Werte zu eins aufaddieren. Damit haben wir dann die Kolmogorov-Ansprüche an eine Wahrscheinlichkeitsverteilung erfüllt.

5.5.2 Bayes-Gitter in R berechnen

Legen wir uns eine Tabelle mit Gitterwerten an, um deren Posteriori-Wahrscheinlichkeit zu berechnen.

```
d <-  
  tibble(  
    # definiere die Hypothesen (das "Gitter"):  
    p_Gitter = seq(from = 0, to = 1, by = 0.1),  
    # bestimme den Priori-Wert:  
    Priori = 1) %>%  
    mutate(  
      # berechne Likelihood für jeden Gitterwert:  
      Likelihood = dbinom(6, size = 9, prob = p_Gitter),  
      # berechne unstand. Posteriori-Werte:  
      unstd_Post = Likelihood * Priori,  
      # berechne stand. Posteriori-Werte (summiert zu 1):  
      Post = unstd_Post / sum(unstd_Post))
```

Das “Bayes-Gitter” (Tabelle 5.2) zeigt, wie sich die Post-Verteilung berechnet.

¹¹synonym: priori

¹²synonym: Die Likelihoodwerte werden mit den Apriori-Werten gewichtet.

Tabelle 5.2: Die Bayes-Box für den Globusversuch

<code>id</code>	<code>p_Gitter</code>	Priori	Likelihood	unstd_Post	Post
1	0.0	1	0.00	0.00	0.00
2	0.1	1	0.00	0.00	0.00
3	0.2	1	0.00	0.00	0.00
4	0.3	1	0.02	0.02	0.02
5	0.4	1	0.07	0.07	0.07
6	0.5	1	0.16	0.16	0.16
7	0.6	1	0.25	0.25	0.25
8	0.7	1	0.27	0.27	0.27
9	0.8	1	0.18	0.18	0.18
10	0.9	1	0.04	0.04	0.04
11	1.0	1	0.00	0.00	0.00

Für jede Hypothese (Spalte `id`) berechnen wir die unstandardisierte Posteriori-Wahrscheinlichkeit als Produkt von Priori und Likelihood:

$$\text{Post}_{\text{unstand}} = \text{Priori} \cdot \text{Likelihood}$$

Um zur standardisierten Posteriori-Wahrscheinlichkeit zu gelangen, teilen wir in jeder Zeile der Gitterbox (also für jede Hypothese) die unstandardisierte Post-Wahrscheinlichkeit durch die Summe der unstandardisierten Post-Wahrscheinlichkeiten.

Was wohl mit *Post* passiert, wenn wir *Priori* ändern?

5.5.3 Was sagt die Post?

Die Posteriori-Verteilung (Kurz: "Post-Verteilung"), Pr_{Post} , zeigt, wie plausibel wir jeden Wert von p halten.

Abbildung 5.13 zeigt die Post-Wahrscheinlichkeit für 5, 10 und 20 Gitterwerte. Das mittlere Teilbild (10 Gitterwerte) entspricht unserer Tabelle oben.

i Hinweis

Unter sonst gleichen Umständen gilt:

- Mehr Gitterwerte glätten die Annäherung.
- Je größer die Stichprobe (N), desto zuverlässiger wird unsere Berechnung.

! Wichtig

Die Post-Verteilung ist sowas wie das Ziel all Ihrer Träume (falls Sie es noch nicht gewusst haben): Aus der Post-Verteilung können Sie ablesen, wie wahrscheinlich Ihre

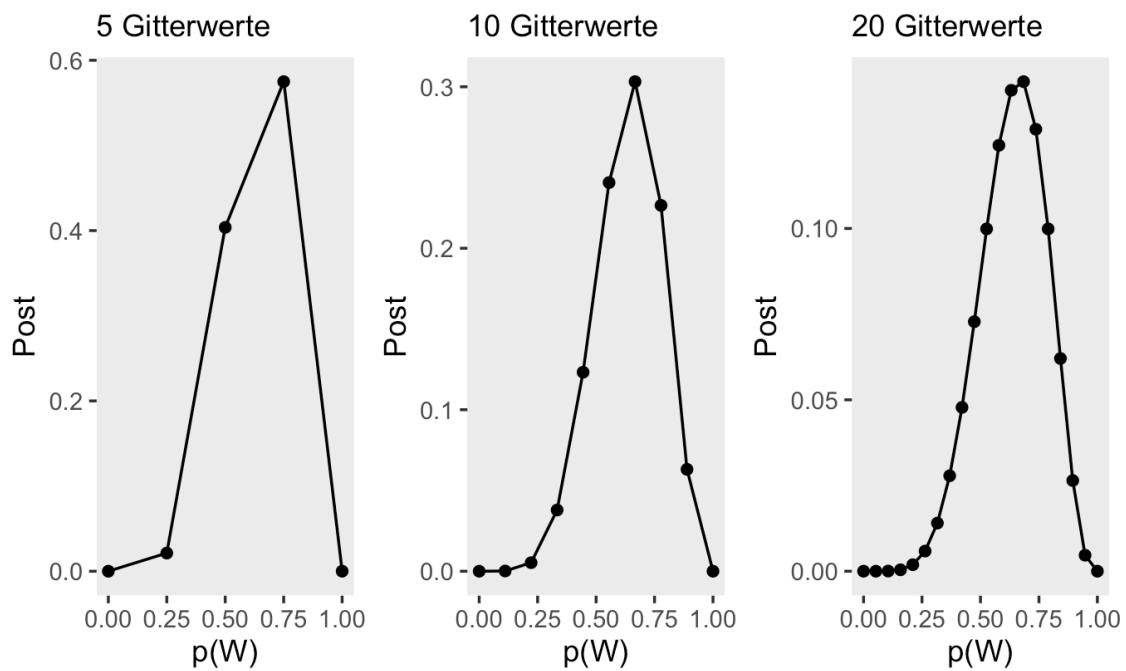


Abbildung 5.13: Je mehr Gitterwerte, desto genauer wird die Verteilung wiedergegeben.

Hypothese (Ihr Lieblings-Parameterwert) ist. Und noch einiges mehr, aber das ist Thema des nächsten Kapitels.

5.6 Aufgaben

- [Rethink_2E4](#)
- [Rethink_2m1](#)
- [Rethink_2m2](#)
- [Rethink_2m3](#)
- [Rethink_2m4](#)
- [Rethink_2m5](#)
- [Rethink_2m6](#)
- [Rethink_2m7](#)
- [kekse01](#)
- [kekse02](#)
- [euro-bayes](#)

5.7 Abschluss

5.7.1 Zusammenfassung

- In unserem Modell haben wir Annahmen zu Pr_{Priori} und L getroffen.
- Auf dieser Basis hat der Golem sein Wissen geupdated zu Pr_{Post} .
- Mit der Gitter-Methode haben wir viele Hypothesen (Parameterwerte) untersucht und jeweils die Pr_{Post} berechnet.
- Unser Modell bildet die kleine Welt ab; ob es in der großen Welt nützlich ist, steht auf einem anderen Blatt.

Wenn Sie auf einen Prozentwert für W tippen müssten, welchen würden Sie nehmen, laut dem Modell (und gegeben der Daten)?

5.7.2 Vertiefung

Das “[Bayes-Paradox-Video](#)” von [3b1b](#) präsentiert eine gut verständliche Darstellung des Bayes-Theorem aus einer zwar nicht gleichen, aber ähnlichen Darstellung wie in diesem Kapitel.

5.7.3 Literatur

Bourier (2018), Kap. 6.2 und 7.1 erläutern einige (grundlegende) theoretische Hintergründe zu diskreten Zufallsvariablen und Wahrscheinlichkeitsverteilungen. Wichtigstes Exemplar ist dabei die Binomialverteilung. McElreath (2020), Kap. 2, stellt das Globusmodell mit mehr Erläuterung und etwas mehr theoretischem Hintergrund vor.

6 Die Post befragen



Abbildung 6.1: Bayes:Start!

6.1 Lernsteuerung

6.1.1 Lernziele

Nach Absolvieren des jeweiligen Kapitels sollen folgende Lernziele erreicht sein.

Sie können ...

- die Post-Verteilung anhand einer Stichprobenverteilung auslesen
- Fragen nach Wahrscheinlichkeitsanteilen der Post-Verteilung anhand der Stichprobenverteilung beantworten
- Fragen nach Quantilen anhand der Stichprobenverteilung beantworten

6.1.2 Benötigte R-Pakete

```
library(tidyverse)
```

6.1.3 Begleitvideos

•

6.2 Mit Stichproben die Post-Verteilung zusammenfassen

6.2.1 Zur Erinnerung: Gitterwerte in R berechnen

Berechnen wir mit der Gittermethode (“Bayes-Box”) die Postverteilung für den Globusversuch.

Die Gittermethode ist ein Weg, die Posteriori-Verteilung zu berechnen. Die Posteriori-Verteilung birgt viele nützliche Informationen.

Modell: $W = 6$ Wasser, $N = 9$ Würfen und $k = 10$ Gitterwerten, also mit 10 Wasseranteils-werten zwischen 0 und 1.

Abb. Abbildung 6.2 zeigt die resultierende Post-Verteilung.

```
n <- 10
n_success <- 6
n_trials <- 9

d <-
  tibble(p_grid = seq(from = 0, to = 1, length.out = n),
         prior = 1) %>%
  mutate(likelihood = dbinom(n_success,
                            size = n_trials,
                            prob = p_grid)) %>%
  mutate(unstand_post = (likelihood * prior),
        post = unstand_post / sum(unstand_post))
```

Voilà, die Post-Verteilung als Tabelle, auch “Bayes-Box” (oder Bayes-Gitter) genannt: s. Tabelle 6.1.

Tabelle 6.1: Postverteilung mit der Gittermethode berechnet

p_grid	prior	likelihood	unstand_post	post
0.00	1	0.00	0.00	0.00
0.11	1	0.00	0.00	0.00
0.22	1	0.00	0.00	0.01
0.33	1	0.03	0.03	0.04
0.44	1	0.11	0.11	0.12
0.56	1	0.22	0.22	0.24
0.67	1	0.27	0.27	0.30
0.78	1	0.20	0.20	0.23
0.89	1	0.06	0.06	0.06
1.00	1	0.00	0.00	0.00

Viele nützliche Fragen (und Antworten) leiten sich ab aus Abb. Abbildung 6.2.

6.2.2 Beispiele für Fragen an die Post-Verteilung

- Mit welcher Wahrscheinlichkeit liegt der Parameter unter einem bestimmten Wert?
- Mit welcher Wahrscheinlichkeit liegt der Parameter zwischen zwei bestimmten Werten?

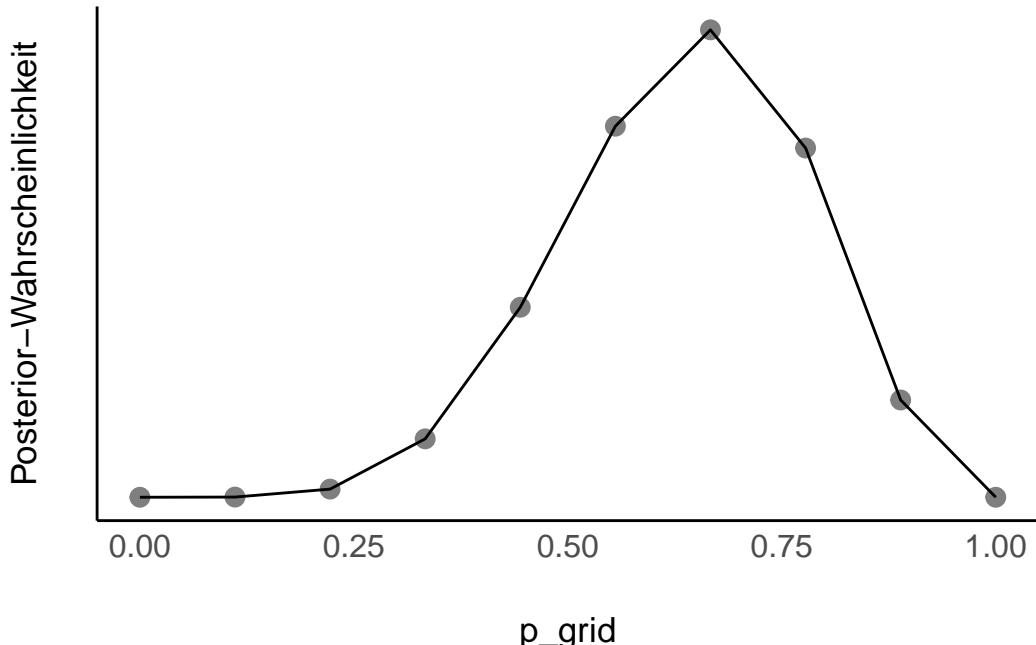


Abbildung 6.2: Die Postverteilung für W=6, N=9, k=10

- Mit 5% Wahrscheinlichkeit liegt der Parameterwert nicht unter welchem Wert?
- Welcher Parameterwert hat die höchste Wahrscheinlichkeit?
- Wie ungewiss ist das Modell über die Parameterwerte?

Solche Fragen kann man in zwei Gruppen aufteilen:

1. Fragen zu Parametern
2. Fragen zu Wahrscheinlichkeiten

6.2.3 Bayes-Box für komplexe Modelle

Bisher, für einfache Fragestellungen hat unsere Bayes-Box, das heißt die Gittermethode bestens funktioniert: einfach, robust, formschön¹. Allerdings: Funktioniert sie auch bei komplexeren Modellen? Schließlich wollen wir ja auch irgendwann Regressionsmodelle berechnen. Angenommen, wir haben ein Regressionsmodell mit 1 Prädiktor, dann haben wir folgende drei Größen² zu schätzen: β_0, β_1, σ . Hört sich gar nicht so viel an. Aber Moment, wir müssten dann z.B. die Frage beantworten, wie wahrscheinlich die Daten aposteriori sind, wenn z.B. $\beta_0 = -3.14$ und $\beta_1 = 2.71$ und $\sigma = 0.70$. Demnach müssen wir alle Ausprägungen (“Gitterwerte”) der Variablen multiplizieren. Puh, das wird eine große Zahl. Wenn wir für die drei Größen jeweils 10 Ausprägungen annehmen, was wenig ist, kämen wir $10 \cdot 10 \cdot 10 = 1000 = 10^3$ Kombinationen. Bei 100 Ausprägungen wären es schon $100^3 = 10^6$ Kombinationen. Das wäre doch eine recht lange Tabelle.

¹naja, nicht unbedingt formschön, aber mir fiel kein dritter Vorzug ein.

²Modellparameter genannt

Bei einer multiplen Regression mit sagen wir 10 Prädiktoren mit jeweils 100 Ausprägungen rechnet das arme R bis zum jüngsten Tag: 10^{100} . Nein, das können wir R nicht zumuten. Wir brauchen eine andere Lösung!

6.2.4 Wir arbeiten jetzt mit Häufigkeit, nicht mit Wahrscheinlichkeit

Kurz gesagt: Komplexere Bayes-Modelle können nicht mehr “einfach mal eben” ausgerechnet werden; die Mathematik wird so umfangreich bzw. zu komplex.

Glücklicherweise gibt es einen Trick, der die Sache nicht nur rechnerisch, sondern auch konzeptionell viel einfacher macht.

Dieser Trick lautet: Wir arbeiten nicht mehr mit Wahrscheinlichkeiten, sondern mit *Häufigkeiten*.

Praktischerweise werden wir in Kürze einen R-Golem kennenlernen, der das für uns erledigt. Dieser Golem liefert uns Stichproben aus der Post-Verteilung zurück.

Lernen wir jetzt also, wie man mit solchen Stichproben umgeht.

! Wichtig

Die Post-Verteilung in Stichprobenform ist viel einfacher zu handhaben als das direkte Arbeiten mit Wahrscheinlichkeiten. Daher sind viele R-Funktionen für Bayes auf Stichproben eingestellt.

Die Grid-Methode ist bei größeren Datensätzen (oder größeren Modellen) zu rechenintensiv. In der Praxis werden daher andere, schnellere Verfahren verwendet, sog. Monte-Carlo-Markov-Ketten (MCMC). Wie diese Verfahren funktionieren sind aber nicht mehr Gegenstand dieses Kurses. Wir wenden Sie einfach an, freuen uns und lassen es damit gut sein (Eine gute Einführung in die Hintergründe findet sich bei McElreath 2020.)

6.2.5 Häufigkeiten sind einfacher als Wahrscheinlichkeiten

Wie gesagt, typische R-Werkzeuge (“R-Golems”) liefern uns die Post-Verteilung in Stichprobenform zurück.

Bevor wir uns aber mit diesen R-Werkzeugen beschäftigen, sollten wir uns vertraut machen mit einer Post-Verteilung in Stichprobenform.

Erstellen wir uns also einen Tabelle mit Stichprobendaten aus der Posteriori-Verteilung (Tabelle `d`), s. Listing 6.1.

Die Wahrscheinlichkeit, einen bestimmten Parameterwert (d.h. aus der Spalte `p_grid`) aus Tabelle `d` zu ziehen, ist proportional zur Posteriori-Wahrscheinlichkeit (`post`) dieses Werts. Ziehen mit Zurücklegen hält die Wahrscheinlichkeiten während des Ziehens konstant. Das Argument `weight_by` legt die Wahrscheinlichkeit fest, mit der eine Zeile gezogen wird.

Listing 6.1 Wir stellen eine Tabelle mit Stichproben aus der Post-Verteilung

```
samples <-
  d %>% # nimmt die Tabelle mit Posteriori-Daten,
  slice_sample( # Ziehe daraus eine Stichprobe,
    n = 1e4, # mit insgesamt n=10000 Zeilen,
    weight_by = post, # Gewichte nach Post-Wskt.,

    replace = T) %>% # Ziehe mit Zurücklegen
  select(p_grid)
```

Wir begnügen uns mit der Spalte mit den Wasseranteilswerten (Parameterwerten), `p_grid`, die anderen Spalten brauchen wir nicht.

Das Ergebnis, Tabelle `samples`, die aus Stichproben aus der Post-Verteilung besteht, ist (in Auszügen) in Tabelle 6.2 dargestellt.

Tabelle 6.2: Die ersten Zeilen der Stichproben aus der Post-Verteilung

p_grid
0.778
0.333
0.778
0.778
0.778

Wenn Sie jetzt denken: "Warum machen wir das jetzt? Brauchen wir doch gar nicht!" - Dann haben Sie Recht. Künftig werden wir aber, wenn wir mit komplexeren Modellen zu tun haben, nur noch mit Post-Verteilungen auf Stichprobenbasis arbeiten, weil es damit viel einfacher ist.

Hier erstmal die ersten 100 gesampelten Gitterwerte (`p_grid`):

```
## [1] 0.78 0.33 0.78 0.78 0.56 0.44 0.67 0.56 0.67 0.78 0.89 0.67 0.56 0.67
## [16] 0.67 0.78 0.78 0.67 0.67 0.67 0.67 0.56 0.56 0.78 0.44 0.67 0.56 0.78
## [31] 0.56 0.67 0.44 0.67 0.56 0.56 0.78 0.78 0.78 0.44 0.67 0.89 0.67 0.67 0.22
## [46] 0.56 0.67 0.67 0.56 0.78 0.44 0.56 0.56 0.33 0.89 0.56 0.56 0.44 0.56 0.44
## [61] 0.67 0.44 0.56 0.78 0.56 0.67 0.78 0.78 0.56 0.56 0.78 0.56 0.89 0.56 0.33
## [76] 0.56 0.78 0.67 0.44 0.78 0.67 0.78 0.67 0.89 0.33 0.56 0.44 0.67 0.67 0.56
## [91] 0.44 0.78 0.89 0.67 0.56 0.78 0.56 0.67 0.78 0.78
```

Wie sieht diese Tabelle wohl als Histogramm³ aus?

³hier als Balkendiagramm, kommt fast aufs selbe raus, sieht aber etwas schöner aus in diesem Fall, da er nur wenige Balken sind

So sieht die Post-Verteilung auf Basis von Stichproben dann aus, s. Abbildung 6.3.

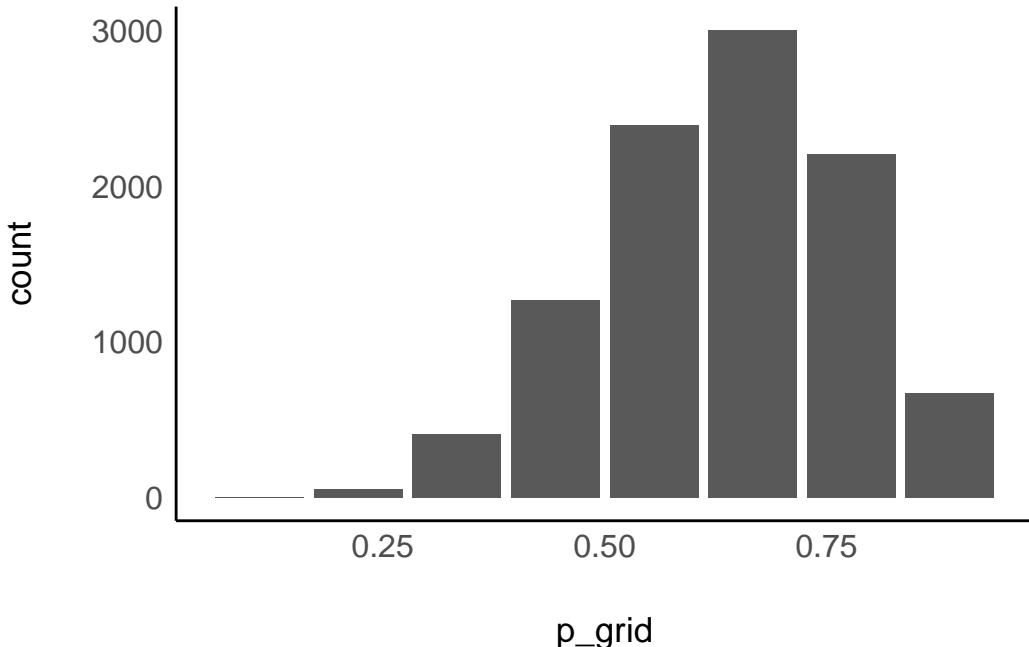


Abbildung 6.3: Stichprobenverteilung auf Basis von Stichproben

Aus Abbildung 6.3 können wir einfach auslesen, wie wahrscheinlich gewisse Parameterwerte sind. So sehen wir, dass das Modell Parameterwerte (Wasseranteil, π) zwischen ca. 50% und 70% für am wahrscheinlichsten hält. Aber auch kleine Anteile wie 25% sind nicht auszuschließen (auf Basis der Daten und der Modellannahmen).

Vergleichen Sie Abbildung 6.3 mit Abbildung 5.13: beide sind sehr ähnlich! Das Stichprobenziehen (Abbildung 6.3) nähert sich recht gut an die exakte Berechnung an (Abbildung 5.13).

6.2.6 Visualisierung der Stichprobendaten mit $k = 100$ Gitterwerten

$k = 10$ Gitterwerte ist ein grobes Raster. Drehen wir mal die Auflösung auf $k = 100$ Gitterwerte (Ausprägungen) nach oben.

```

k <- 100
n_success <- 6
n_trials <- 9

d_k100 <-
  tibble(p_grid = seq(from = 0,
                      to = 1,
                      length.out = k), # 100 Gitterwerte
         prior = 1) %>%

```

```

    mutate(likelihood = dbinom(n_success,
                                size = n_trials,
                                prob = p_grid)) %>%
  mutate(unstand_post = (likelihood * prior),
        post = unstand_post / sum(unstand_post))

```

$d_k|100$ ist eine Bayes-Box mit $W = 6, N = 9, k = 100$.

Und daraus ziehen wir uns $n = 1000$ Stichproben:

```

samples_k100 <-
  d_k100 %>% # nimmt die Tabelle mit Posteriori-Daten,
  slice_sample( # Ziehe daraus eine Stichprobe,
    n = 1000, # mit insgesamt n=1000 Elementen,
    weight_by = post, # Gewichte nach Spalte mit Post-Wskt.,
    replace = T) # Ziehe mit Zurücklegen

```

Abbildung 6.4 zeigt sowohl die exakte Post-Verteilung als auch die Post-Verteilung auf Basis von Stichproben. Im mittleren Teildiagramm sind die Stichproben einzeln als Kreis dargestellt. Im rechten Teildiagramm sind die gleichen Daten als Dichtediagramm dargestellt. In allen Fällen erkennt man gut die zentrale Tendenz: ein Wasseranteil von 70% scheint der “typische” Wert des Modells zu sein. Außerdem erkennt man, dass das Modell durchaus einige Streuung in der Schätzung des Wasseranteils bereithält. Das Modell ist sich nicht sehr sicher, könnte man sagen.

Post-Verteilung ex: Stichproben aus de Stichproben aus c

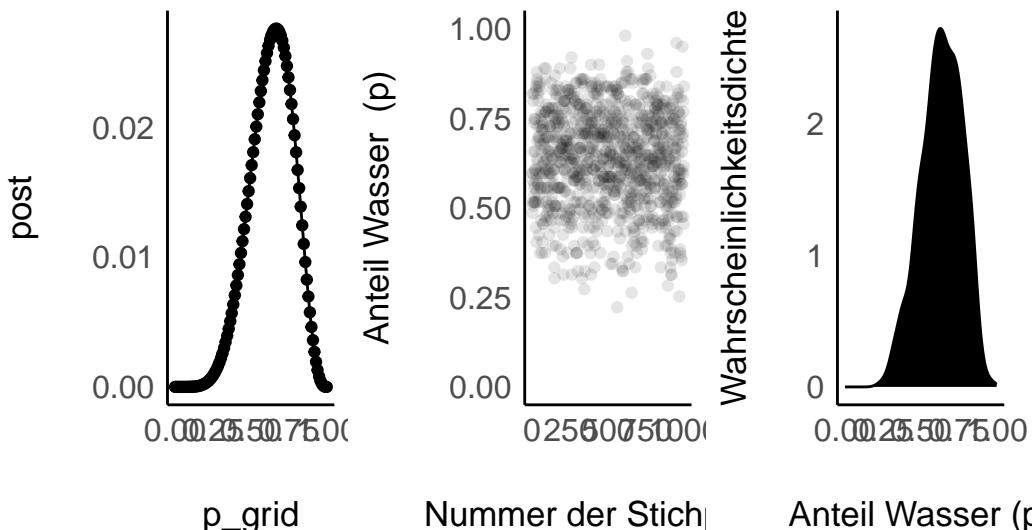


Abbildung 6.4: Post-Verteilung mit 100 Gitterwerten, exakt vs. auf Basis von Stichproben

Die Stichprobendaten nähern sich der “echten” Posteriori-Verteilung an: Die Stichproben-

Post-Verteilung hat jetzt “glattere” Ränder.

i Hinweis

Mehr Stichproben und mehr Gitterwerte glätten die Verteilung.

Jetzt noch mal mit mehr Stichproben: $n = 10^6$ Stichproben bei $k = 100$ Gitterwerten aus der Posteriori-Verteilung, s. Abbildung 6.5.

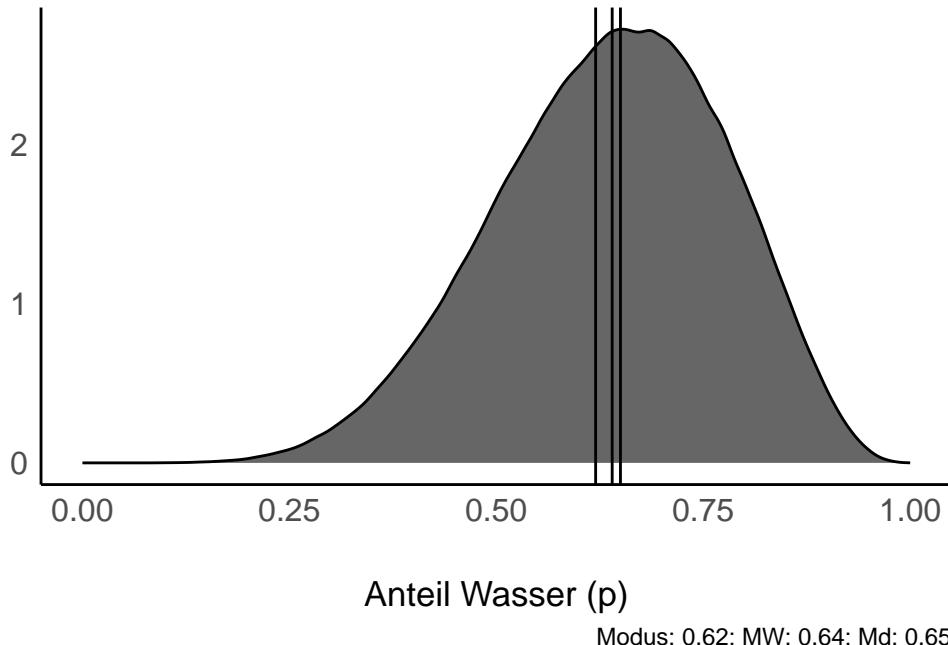


Abbildung 6.5: Post-Verteilung mit vielen Stichproben und vielen Parameterwerten (Gitterwerten): schön ‘glatt’. Mittelwert (MW), Modus und Median (Md) liegen eng nebeneinander, da die Verteilung recht symmetrisch ist.

6.3 Die Post-Verteilung befragen

So, jetzt befragen wir die Post-Verteilung.

! Wichtig

Die Post-Verteilung ist das zentrale Ergebnis einer Bayes-Analyse. Wir können viele nützliche Fragen an sie stellen.

Es gibt zwei Arten von Fragen:

1. nach Wahrscheinlichkeiten (p)
2. nach Parameterwerten (Quantilen, q)

Der Unterschied zwischen beiden Arten von Fragen ist in Abbildung 6.6 illustriert.

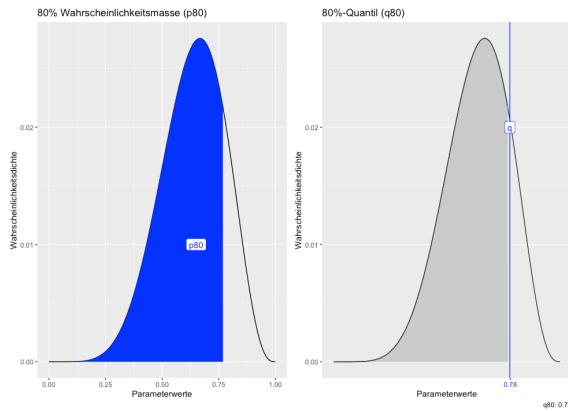


Abbildung 6.6: Fragen nach p vs. Fragen nach q

Im linken Teildiagramm von Abbildung 6.6 fragen wir: "Wie wahrscheinlich ist ein Wasseranteil von höchstens 80%?". Im rechten Teildiagramm fragen wir: "Welcher Wasseranteil wird mit einer Wahrscheinlichkeit von 78% nicht überschritten?".

6.3.1 Fragen nach Wahrscheinlichkeiten

Sagen wir, dass sei unsere Forschungsfrage: *Wie groß ist die Wahrscheinlichkeit, dass der Wasseranteil unter 50% liegt?*

Um diese Frage zu beantworten, zählen wir einfach, wie viele Stichproben die Bedingung erfüllen:

und und summieren die Wahrscheinlichkeiten dieser Stichproben:

Wir zählen (`count`) also die Stichproben, die sich für einen Wasseranteil (`p_grid`) von weniger als 50% aussprechen:

```
samples %>%
  count(p_grid < .5)
## # A tibble: 2 x 2
##   `p_grid < 0.5`     n
##   <lgl>           <int>
## 1 FALSE            8272
## 2 TRUE             1728
```

Da wir insgesamt 10000 (1e4) Stichproben gezogen haben, können wir noch durch diese Zahl teilen, um einen Anteil zu bekommen. Dieser Anteil ist die Antwort auf die Forschungsfrage: Wie Wahrscheinlichkeit (laut Modell) für einen Wasseranteil kleiner als 50%.⁷[Der Befehl `count` macht Folgendes: Er gruppiert die Stichprobe nach dem Prüfkriterium, Wasseranteil höchstens 50%. Dann zählt er in jeder der beiden Teiltabellen die Zeilen und liefert diese zwei Zahlen dann zurück. Man könnte also auch in etwa schreiben:

```
d %>%
  filter(p_grid < .5) %>%
  summarise(sum = sum(post))
## # A tibble: 1 x 1
##      sum
##      <dbl>
## 1 0.167
```

Einfach wie essen.

Beispiel 6.1 (Wasseranteil zwischen 50 und 75%). Noch eine Forschungsfrage: *Mit welcher Wahrscheinlichkeit liegt der Parameter (Wasseranteil) zwischen 0.5 und 0.75?*

```
samples %>%
  count(p_grid > .5 & p_grid < .75)
## # A tibble: 2 x 2
##   `p_grid > 0.5 & p_grid < 0.75`     n
##   <lgl>                           <int>
## 1 FALSE                           4607
## 2 TRUE                            5393

samples %>%
  count(p_grid > .5 & p_grid < .75) %>%
  summarise(Anteil = n / 1e4,
            Prozent = 100 * n / 1e4) # In Prozent
## # A tibble: 2 x 2
##   Anteil Prozent
##   <dbl>    <dbl>
## 1 0.461     46.1
## 2 0.539     53.9
```

Anteile von `count()` könnte man, wenn man möchte, auch `filter()` verwenden:

```
samples %>%
  filter(p_grid > .5 & p_grid < .75) %>%
  summarise(sum      =      n() / 1e4,
            anteil = 100 * n() / 1e4) # In Prozent
## # A tibble: 1 x 2
##      sum      anteil
##      <dbl>    <dbl>
## 1 0.539     53.9
```

Beispiel 6.2 (Wasseranteil zwischen 90 und 100%). Noch ein Beispiel für eine Forschungsfrage: Mit welcher Wahrscheinlichkeit liegt der Parameter zwischen 0.9 und 1?

```
samples %>%
  count(p_grid >= .9 & p_grid <= 1) %>%
  summarise(prop = 100 * n() / 1e4) # prop wie "proportion", Anteil
## # A tibble: 1 x 1
##   prop
##   <dbl>
## 1 0.01
```

Laut unserem Modell ist es also sehr unwahrscheinlich, dass der Wasseranteil der Erde mind. 90% beträgt.

Wir können auch fragen, welcher Parameterwert am wahrscheinlichsten ist; dieser Wert entspricht dem “Gipfel” des Berges, s. Abbildung 6.5.

Für unsere Stichproben-Postverteilung, `samples`, s. Abbildung 6.3, lässt sich der Modus so berechnen:

```
map_estimate(samples$p_grid)
## MAP Estimate: 0.67
```

Dabei steht `map` für **Maximum Aposteriori**, also das Maximum der Post-Verteilung.

Bei der Gelegenheit könnten wir folgende, ähnliche Fragen stellen:

- Was ist der mittlere Schätzwert (Mittelwert) zum Wasseranteil laut Post-Verteilung?
- Was ist der mediane Schätzwert (Median)?

Auf Errisch:

```
samples %>%
  summarise(mean(p_grid),
            median(p_grid))
## # A tibble: 1 x 2
##   `mean(p_grid)` `median(p_grid)`
##   <dbl>           <dbl>
## 1 0.635          0.667
```

6.3.2 Fragen nach Parameterwerten

! Wichtig

Schätzbereiche von Parameterwerten nennt man auch *Konfidenz- oder Vertrauensintervall*^a.

^aTatsächlich gibt es eine Vielzahl an Begriffen, die in der Literatur nicht immer konsistent verwendet werden, etwa Kompatibilitätsintervall, Ungewissheitsintervall, Passungsbereich.

Welcher Parameterwert wird mit 90% Wahrscheinlichkeit nicht überschritten, laut unserem Modell? (Gesucht sind also die unteren 90% der Posteriori-Wahrscheinlichkeit) Wir möchten also ziemlich sicher, was die Obergrenze an Wasser auf diesem Planeten ist⁴.

```
samples %>%
  summarise(quantil90 = quantile(p_grid, p = .9))
## # A tibble: 1 x 1
##   quantil90
##   <dbl>
## 1 0.778
```

Laut unserem Modell können wir zu 90% sicher sein, dass der Wasseranteil kleiner ist als ca. 78%.

Es hilft vielleicht, sich die Post-Verteilung noch einmal vor Augen zu führen, s. fig-post99.

```
samples %>%
  ggplot(aes(x = p_grid)) +
  geom_bar()
```

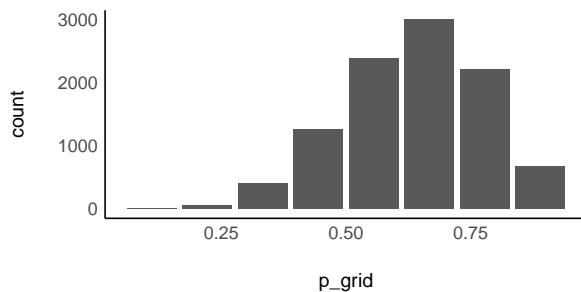


Abbildung 6.7: Die Post-Verteilung im Globusversuch

Was ist das *mittlere* Intervall, das mit 90% Wahrscheinlichkeit den Parameterwert enthält, laut dem Modell?

⁴Vielleicht damit es genug Berge zum Schifahren gibt.

Dafür “schneiden” wir links und rechts die 5% der Stichproben mit den extremsten Werten ab und schauen, bei welchen Parameterwerten wir als Grenzwerte landen:

```
samples %>%
  summarise(
    quant_05 = quantile(p_grid, 0.05),
    quant_95 = quantile(p_grid, 0.95))
## # A tibble: 1 x 2
##   quant_05 quant_95
##       <dbl>     <dbl>
## 1      0.444     0.889
```

Solche Fragen lassen sich also mit Hilfe von *Quantilen* beantworten.

6.3.3 Zur Erinnerung: Quantile

Beispiel: Wie groß sind die Studentis ([Quelle des Datensatzes](#))?

Das Quantil von z.B. 25% zeigt die Körpergröße der 25% kleinsten Studentis an, analog für 50%, 75%, in Inches⁵:

```
speed_gender_height <- read_csv("https://raw.githubusercontent.com/rpruim/OpenIntro/master/data/speed_gender_height.csv")

height_summary <-
  speed_gender_height %>%
  mutate(height_cm = height*2.54) %>%
  select(height_inch = height, height_cm) %>%
  drop_na() %>%
  pivot_longer(everything(), names_to = "Einheit", values_to = "Messwert") %>%
  group_by(Einheit) %>%
  summarise(q25 = quantile(Messwert, prob = .25),
            q50 = quantile(Messwert, prob = .5),
            q75 = quantile(Messwert, prob = .75))

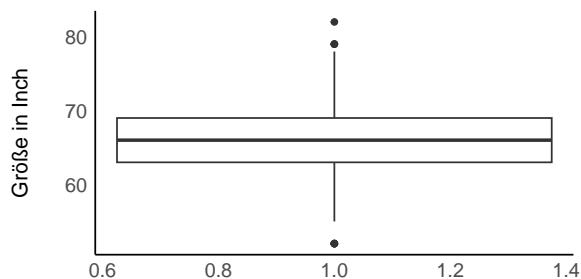
height_summary
## # A tibble: 2 x 4
##   Einheit      q25     q50     q75
##   <chr>      <dbl>    <dbl>    <dbl>
## 1 height_cm  160.    168.    175.
## 2 height_inch 63      66      69
```

Das 25%-Quantil nennt man auch 1. Quartil; das 50%-Quantil (Median) auch 2. Quartil und das 75%-Quantil auch 3. Quartil.

Abbildung [6.8](#) visualisiert die Quantile und die Häufigkeitsverteilung.

⁵1 Inch entspricht 2.54cm

Die Box zeigt das 25%-, 50%- und 75%-Quantil



Die vertikalen Striche zeigen die Quantile

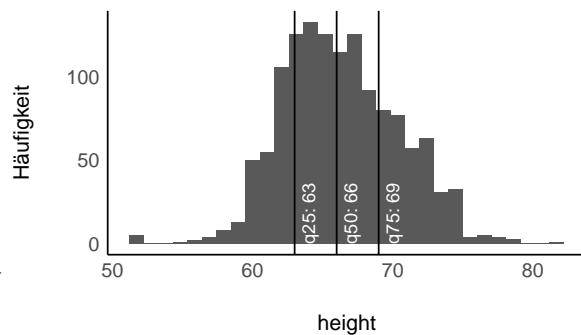


Abbildung 6.8: Größenverteilung von 1325 amerikanischen Studenten

6.3.4 Den Quantilen unter die Motorhaube geschaut

Den R-Befehl `quantile()` kann man sich, wenn man will, einfach nachbauen und entmystifizieren.

Angenommen, wir wollen wissen, welcher Wasseranteil mit 90% Wahrscheinlichkeit nicht überschritten wird. Das können wir mit im Datensatz `samples` so erreichen.

1. Sortiere die Stichproben aufsteigend.
2. Schneide die oberen 10% ab.
3. Schau, was der größte verbleibende Wert ist.

```
samples %>%
  arrange(p_grid) %>%    # sortiere
  slice_head(n = 9000) %>%  # nur die ersten 90000, also die obersten 1000 abschneiden
  summarise(p90 = max(p_grid))
## # A tibble: 1 x 1
##       p90
##   <dbl>
## 1 0.778
```

Das (annähernd) gleiche Ergebnis liefert `quantile()`:

```
samples %>%
  summarise(q90 = quantile(p_grid, .9))
## # A tibble: 1 x 1
##       q90
##   <dbl>
## 1 0.778
```

6.3.5 Visualisierung der Intervalle

Intervalle (Bereiche), die die “abzuschneidende” Wahrscheinlichkeitsmasse hälftig auf die beiden Ränder aufteilen, nennen wir *Perzentilintervalle* oder *Equal-Tails-Intervalle* (ETI), s. Abb. Abbildung 6.9, rechtes Teildiagramm.

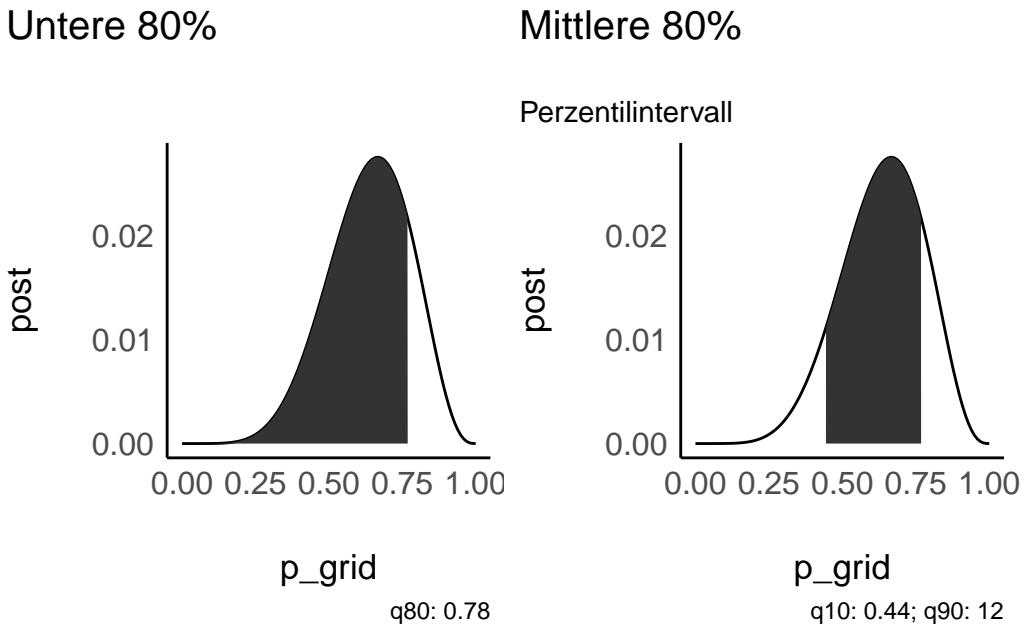


Abbildung 6.9: Perzentilintervalle

6.4 Schiefe Posteriori-Verteilungen sind möglich

Noch einmal zum Globusversuch: Gehen wir von 3 Würfen mit 3 Treffern aus; auf welche Wasseranteile (Parameterwerte) werden wir jetzt schließen?

Vermutlich ziemlich hohe.

Erstellen wir uns dazu mal eine Post-Verteilung (3 Treffer, 3 Würfe):

```
d_33 <-  
  tibble(p_grid = seq(0,1, by = .01),  
         prior = 1) %>%  
  mutate(likelihood = dbinom(3, size = 3, prob = p_grid)) %>%  
  mutate(unstand_post = likelihood * prior) %>%  
  mutate(post_33 = unstand_post / sum(unstand_post))  
  
samples_33 <-  
  d_33 %>%  
  slice_sample(n = 1e4,
```

```
weight_by = post_33,
replace = T)
```

So sehen die ersten paar Zeilen der Post-Verteilung, `samples_33`, aus.

p_grid	prior	likelihood	unstand_post
0.82	1	0.55	0.55
0.97	1	0.91	0.91
0.84	1	0.59	0.59
0.79	1	0.49	0.49
0.82	1	0.55	0.55
0.68	1	0.31	0.31

Mit dieser “schiefen” Post-Verteilung können wir gut die Auswirkungen auf das Perzentil- und das Höchste-Dichte-Intervall anschauen.

6.4.1 50%-Perzentil-Intervall

Hier z.B. ein 50%-Perzentilintervall, s. Abb. Abbildung 6.10.

Ein Perzentilintervall kann, wenn es dumm läuft, den wahr:

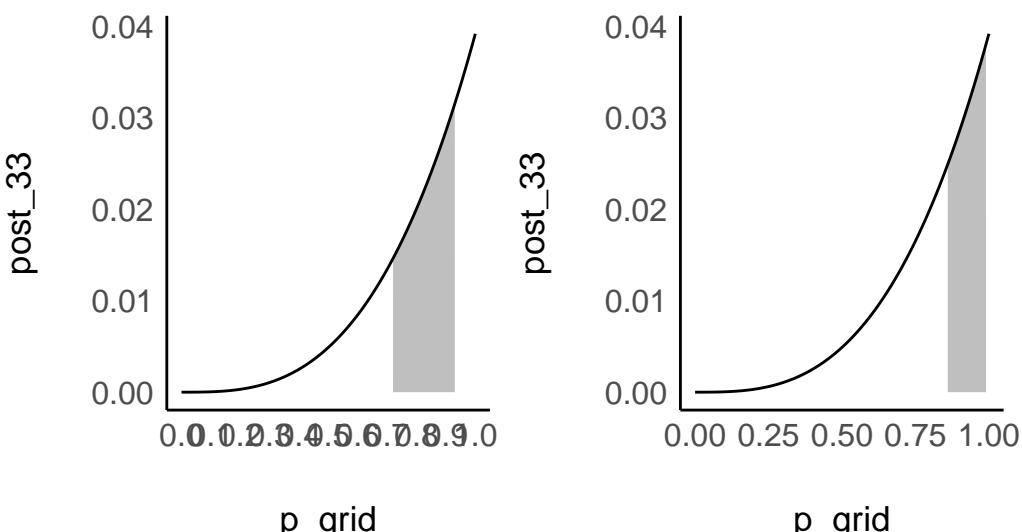


Abbildung 6.10: Schiefe Intervalle

Die Grenzwerte dieses ETI (oder jedes beliebig breiten) kann man sich z.B. so ausgeben lassen:

```

library(easystats)

samples_33 %>%
  select(p_grid) %>%
  eti(ci = .5)
## Equal-Tailed Interval
##
## Parameter /      50% ETI
## -----
## p_grid     / [0.71, 0.94]

```

Der wahrscheinlichste Parameterwert (1) ist *nicht* im Intervall enthalten. Das ist ein Nachteil der ETI.

6.4.2 50%-Intervall höchster Dichte

Intervalle höchster Dichte (Highest density Intervals, HDI oder HDPI) sind definiert als die *schmälsten* Intervalle, die den gesuchten Parameter enthalten.

Der wahrscheinlichste Parameterwert (1) *ist* im Intervall enthalten, was Sinn macht. Bei einem HDI sind die abgeschnittenen Ränder nicht mehr gleich groß, im Sinne von enthalten nicht (zwangsläufig) die gleiche Wahrscheinlichkeitsmasse.

Je symmetrischer die Verteilung, desto näher liegen die Punktschätzer aneinander (und umgekehrt), s. Abb. Abbildung 6.11.

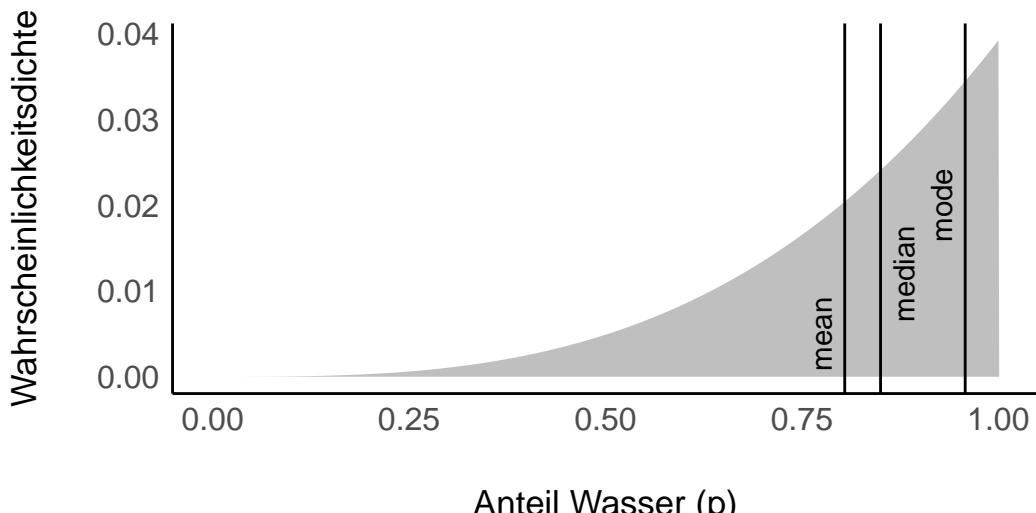


Abbildung 6.11: Visualisierung der Punktschätzer bei einer schiefen Post-Verteilung

So kann man sich die Grenzwerte eines 50%-HDI ausgeben lassen:

```

samples %>%
  select(p_grid) %>%
  bayestestR::hdi(ci = .5) # aus dem Paket `bayestestR`  

## Highest Density Interval  

##  

## Parameter / 50% HDI  

## -----  

## p_grid / [0.67, 0.78]

```

Das Modell ist sich also zu 50% sicher, dass der gesuchte Parameter (der Wasseranteil der Erdoberfläche) sich in diesem Bereich befindet (auf Basis eines HDI).

i Hinweis

Das R-Paket `{bayestestR}` ist Teil des Meta-Pakets `{easystats}`. Es reicht, wenn Sie `easystats` laden, damit wird `bayestestR` automatisch geladen.

6.5 Fazit

6.5.1 Intervalle höchster Dichte vs. Perzentilintervalle

- Bei symmetrischer Posteriori-Verteilung sind beide Intervalle ähnlich
- Perzentilintervalle sind verbreiter
- *Intervalle höchster Dichte* (Highest Density Interval, HDI) sind bei schiefen Post-Verteilungen zu bevorzugen
- Intervalle höchster Dichte sind die *schmalsten* Intervalle für eine gegebene Wahrscheinlichkeitsmasse

6.5.2 Zusammenfassung

Fassen wir zentrale Punkte an einem Beispiel zusammen.

Im Globusversuch, Datendat `samples`, s. Listing 6.1. Sagen wir, wir haben 6 Treffer bei 9 Würfen erzielt.

Lageparamter: Welchen mittleren Wasseranteil muss man annehmen?

```

samples %>%
  summarise(
    mean = mean(p_grid),
    median = median(p_grid))
## # A tibble: 1 x 2
##      mean median
##        <dbl>  <dbl>
## 1 0.670   0.670

```

```
## <dbl> <dbl>
## 1 0.635 0.667
```

Streuungsparameter: Wie unsicher sind wir in der Schätzung des Wasseranteils?

```
samples %>%
  summarise(
    p_sd = sd(p_grid),
    p_iqr = IQR(p_grid),
    p_mad = mad(p_grid))
## # A tibble: 1 x 3
##      p_sd   p_iqr   p_mad
##      <dbl> <dbl> <dbl>
## 1 0.141 0.222 0.165
```

Anstelle der Streuungsparameter ist es aber üblicher, ein HDI oder PI anzugeben.

! Wichtig

Alles Wasser oder was? Im Beispiel dieses Kapitels haben wir unser gefragt, was wohl der Wasseranteil auf dem Planeten Erde ist. Halten Sie sich klar vor Augen: Der Wasseranteil ist ein Beispiel für einen *Parameter*, einer unbekannten Größes eines Modells.

6.6 Aufgaben

1. [iq01](#)
2. [iq02](#)
3. [iq03](#)
4. [iq04](#)
5. [iq05](#)
6. [iq06](#)
7. [iq07](#)
8. [iq08](#)
9. [iq10](#)
10. [fattails1](#)
11. [fattails2](#)
12. [ReThink3e1-7](#)

7 Vorhersage-Verteilung



Abbildung 7.1: Bayes:Start!

7.1 Lernsteuerung

7.1.1 Lernziele

Nach Absolvieren des jeweiligen Kapitels sollen folgende Lernziele erreicht sein.

Sie können ...

- erläutern, was eine Posteriori-Prädiktiv-Verteilung (PPV) ist, und inwiefern Sie vor Übergewissheit schützt
- eine informelle Modellprüfung für das Beispiel aus dem Unterricht anhand der Posteriori-Prädiktiv-Verteilung durchführen

7.1.2 Benötigte R-Pakete

```
library(tidyverse)
```

7.2 Der zwielichtige Dozent: Stichproben-Vert. vs. Post-Vert.

In einer dunklen Gasse fordert Sie ein Statistik-Dozent zu einem Glücksspiel heraus¹. Münzwurf; wenn er gewinnt, müssen Sie 10 Euro zahlen. Gewinnen Sie, bekommen Sie 11 Euro. Klingt nach einer guten Partie, nicht war? Natürlich nehmen Sie sofort an.

Sie spielen also Münzwurf; der Dozent setzt auf Zahl. Sie spielen 10 Runden. Leider gewinnt der Dozent 9 von 10 Mal².

Ist die Münze fair oder zieht der mich über den Tisch?, das ist die Frage, die Ihnen brennend durch den Kopf zieht.

¹Hier bräuchte es ein passendes Meme; Vorschläge bitte an mich.

²was er mit lautem Gelächter quittiert

“Sind 9 von 10 Treffern noch realistisch erwartbar, wenn es mit rechten Dingen zugeht, oder beweist das Ergebnis, dass die Münze gezinkt ist?”

Wütend (und mit leeren Taschen) ziehen Sie von dannen.

Zusammengefasst: Daten: 9 von 10 Treffern beim Münzwurf. Forschungsfrage: Ist die Münze fair?

Schauen wir uns zunächst einmal an, wie wahrscheinlich 9 von 10 Treffern sind, *wenn* die Münze fair ist, s. Abbildung 7.2.

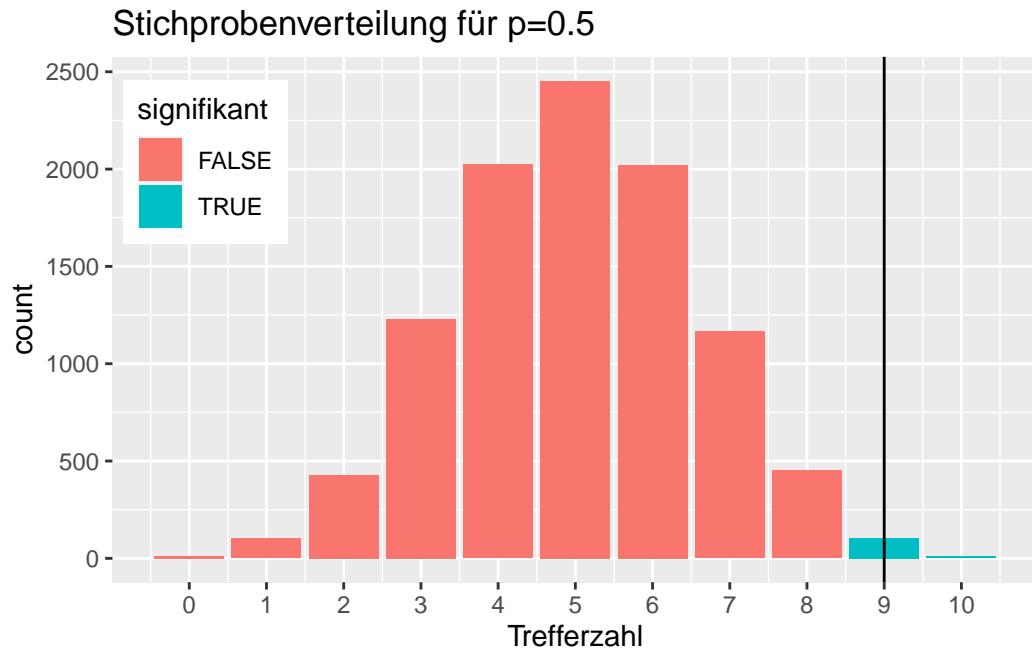


Abbildung 7.2: Stichprobenverteilung einer fairen Münze

Die *Stichprobenverteilung* zeigt, wie wahrscheinlich die empirischen Daten D (z.B. 9 von 10 Treffer) sind, gegeben eines Parameterwerts π (z.B. $p = 0.5$): $Pr(D|\pi)$ ³.

Anders gesagt, die Stichprobenverteilung zeigt die Verteilung der Likelihoods eines bestimmten Parameterwerts.

i Hinweis

Der p -Wert

Der p -Wert ist die zentrale Statistik der Inferenzstatistik. Er wird genutzt, um über die Ablehnung einer Hypothese zu entscheiden. In diesem Fall entspricht der p -Wert dem türkis markierten Flächenanteil in Abbildung 7.2. Ist dieser Anteil kleiner als 5% (der Gesamtfläche im Balkendiagramm), so wird die Hypothese (hier: faire Münze) ver-

³Das griechische kleine p wird “pi” genannt und π geschrieben. Zur Erinnerung: Parameter- oder Populationskennwerte werden in der Statistik häufig mit griechischen Buchstaben benannt, um sie von Stichprobenkennwerten abzugrenzen.

worfen. Allgemeiner gesprochne berechnet sich der p-Wert als Summe der Likelihoods, die mindestens so extrem sind wie das beobachtete empirische Ergebnis.

In der Bayes-Statistik ist die Post-Verteilung Dreh- und Angelpunkt der Entscheidung über eine Hypothese. In Abbildung 7.3 ist die Posteriori-Verteilung für die Daten zum zwielichten Dozent dargestellt.

```
# Post-Verteilung:
d_zwielicht <-
  tibble(
    p_grid = seq( from=0 , to=1 , length.out=100) ,
    prior = 1, # Priori-Gewichte
    likelihood = dbinom(8, size = 10, prob=p_grid) ,
    unstandardisierte_posterior = likelihood * prior ,
    posterior = unstandardisierte_posterior / sum(unstandardisierte_posterior))

# Stichproben ziehen aus der Posteriori-Verteilung:
samples_zwielicht <-
  tibble(
    gewinnchance_muenze = sample(
      d_zwielicht$p_grid ,
      prob=d_zwielicht$posterior,
      size=1e4,
      replace=TRUE)) %>%
  mutate(
    id = row_number())
```

Die *Posteriori-Verteilung* gibt die Wahrscheinlichkeit jedes Parameterwerts p wider, gegeben der empirischen Daten D : $Pr(p|D)$.

Die meisten Forschungsfragen lassen sich mit der Post-Verteilung beantworten, nicht mit der Stichprobenverteilung.

Jetzt können wir wieder die Post-Verteilung auslesen, um die Hypothese zu beantworten. Schauen wir uns einige Beispiel dazu an.

Beispiel 7.1 (Einigermaßen fair?). Wie wahrscheinlich ist es, dass die Münze “einigermaßen” fair ist, sagen wir, eine Trefferwahrscheinlichkeit $0.45 < \pi < 0.55$ ⁴ aufweist?

```
samples_zwielicht %>%
  count(gewinnchance_muenze > 0.45 & gewinnchance_muenze < 0.55) %>%
  mutate(prop = n/sum(n))
## # A tibble: 2 x 3
```

⁴zwischen 45% und 55% mit anderen Worten

Posteriori-Verteilung

Priori: Gleichverteilung; Daten: 9 von 10 Treffern, binomialverteilt

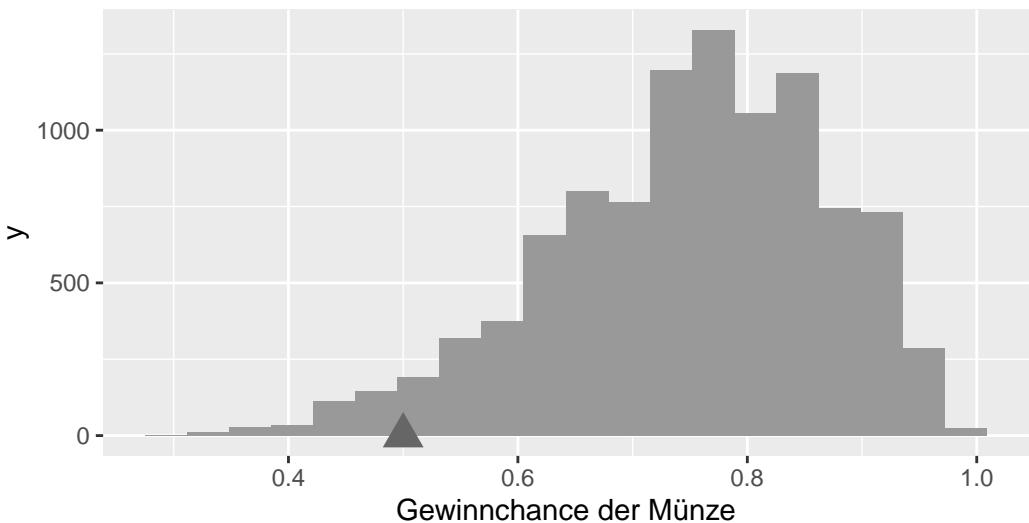


Abbildung 7.3: Post-Verteilung zu den Daten des zwielichten Dozenten (9 von 10 Treffern im wiederholten Münzwurf)

```
## `gewinnchance_muenze > 0.45 & gewinnchance_muenze < 0.55`      n   prop
## <lgl>                                         <int> <dbl>
## 1 FALSE                                         9485 0.948
## 2 TRUE                                         515  0.0515
```

Die Wahrscheinlichkeit für eine “einigermaßen faire” Münze ist klein, etwa 5%!

Beispiel 7.2 (Münze gezinkt?). Schauen wir uns an, wie wahrscheinlich es ist - gegeben der Daten und unserem Modell - dass die Münze massiv gezinkt ist. “Massiv” definieren wir dabei mit “mindestens 70% Trefferwahrscheinlichkeit”⁵, also $\pi \geq .7$ ⁶.

```
samples_zwielicht %>%
  count(gewinnchance_muenze > .7) %>%
  mutate(prop = n / sum(n))
## # A tibble: 2 x 3
##   `gewinnchance_muenze > 0.7`     n   prop
```

⁵ja, das ist subjektiv

⁶Führende Nullen bei Anteilen werden oft weggelassen, man schreibt also oft .7 wenn man 0.7 bzw. 70% meint. Das ist nicht nur kürzer, sondern man weiß auch direkt dass es sich um einen Anteil handelt. Behält man die führende Null bei, etwa 0.77, so würde das signalisieren, dass die Zahl auch größer als Null sein könnte.

```

##   <lgl>          <int> <dbl>
## 1 FALSE           3179  0.318
## 2 TRUE            6821  0.682

```

Wir finden eine recht hohe Wahrscheinlichkeit für eine “massive” Manipulation der Münze.

! Wichtig

Ist es nicht einfach und schön, wie wir mit Hilfe des Stichprobenziehens allerlei Forschungsfragen beantworten können? Eine Post-Verteilung aus Stichproben erlaubt uns, viele Fragen mit einfachen Methoden, nämlich schlichtes Zählen, zu beantworten.

Natürlich könnte (und sollte?) man unser Modell kritisieren. Ist es wirklich sinnvoll, die Trefferwahrscheinlichkeit apriori als gleichverteilt anzunehmen? Das heißt ja, wir glauben, dass eine Trefferwahrscheinlichkeit von 99,99999% genauso wahrscheinlich ist wie 50,55555%. Auf der anderen Seite: Der Charme einer Gleichverteilung ist, dass sie objektiv ist, in dem Sinne, dass wir keinerlei Information einfließen lassen. Wir sind indifferent gegenüber dem Parameter π , der Trefferwahrscheinlichkeit.

i Hinweis

In einem zweiten Versuch könnten wir jetzt unsere Post-Verteilung als Priori-Verteilung nutzen. Das Ergebnis des ersten Versuchs wird dann hergenommen als Ausgangspunkt für einen zweiten Versuch. Damit wird das Wissen der Wissenschaft weitergegeben^a, so wie es sein sollte.

^aübrigens auf mathematisch gesehen ideale Art und Weise.

7.3 Mit Stichproben neue Beobachtungen simulieren

Zur Erinnerung: Der Likelihood (L) zeigt die Wahrscheinlichkeit eine Trefferzahl gegeben eines bestimmten Parameterwerts. In unseren Beispiel könnten wir z.B. die drei Likelihoods für $w = 0, 1, 2$ ausrechnen, gegeben $N = 2$ und $p = 0.5$:

```

L <- dbinom(0:2, size = 2, prob = 0.5)
L
## [1] 0.25 0.50 0.25

```

Ah, die Wahrscheinlichkeit für 0 oder 2 Treffer beträgt 50%, wenn $pi = 1/2$; für 1 Treffer beträgt sie entsprechend 50%⁷.

⁷das sollte uns bekannt vorkommen

7.3.1 Wir simulieren die Wasserzahl bei Globuswürfen

Zurück zu unserem Globusversuch!

Wir könnten uns jetzt Globusbälle basteln mit verschiedenen Wasseranteilen, und diese oft hochwerfen. Damit könnten wir herausfinden, welche Trefferzahlen sich bei verschiedenen Wasseranteilen finden lassen würden.

Wer gerne bastelt, freut sich darauf. Kritischere Geister⁸ würden den Aufwand bemängeln und die Frage nach dem Zweck der Übung stellen⁹.

! Wichtig

Wenn wir wissen, welche Trefferzahlen laut einem Modell zu erwarten sind, können wir die *echten* (beobachteten) Trefferzahlen mit den laut Modell zu erwartenden vergleichen. Damit haben wir eine Methode, mit dem wir ein Modell auf Herz und Nieren prüfen können. Ein schlechtes Modell wird mit seinen Vorhersagen an der Realität scheitern: Erwartung des Modells und beobachtete Daten passen nicht zusammen. Sagt ein Modell etwa $W = 9$ vorher bei $N = 9$, aber wir finden $W = 0$, so wird unser Vertrauen in das Modell erschüttert sein. Simulation von Trefferzahlen sind also ein Modell, um die Glaubwürdigkeit unseres Golems zu prüfen. (Nicht nur) bei Golems gilt: Vertrauen ist gut, Kontrolle ist besser.

Los geht's: Wir simulieren $n = 1$ neuen Globusversuch mit $N = 2, p = 0.7$ und zählen die (Wasser-)Treffer:

```
set.seed(42) # Zufallszahlen festlegen
rbinom(n = 1, size = 2, prob = .7) # 0 Treffer (Wasser)
## [1] 0
```

Das geht wie man sieht mit `rbinom`: *r* wie *random* (zufällig) und *binom* wie binomial verteilt, die Münzwurfverteilung.

Hier sind die Argumente der Funktion `rbinom` noch etwas näher erklärt:

```
rbinom(n = Wie oft soll der Versuch wiederholt werden?,
        size = Wie viele Globuswürfe pro Versuch (Stichprobengröße),
        prob = Wie hoch ist die Wahrscheinlichkeit für Wasser (bzw. für einen Treffer))
```

Weiter: Warum nicht $n = 10$ neue Globusversuche simulieren?

```
rbinom(n = 10, size = 2, prob = 0.7)
## [1] 0 2 1 1 1 1 2 1 1 2
```

⁸oder weniger bastelfreundliche

⁹bravo!

“Simulieren” heißt hier, wir lassen den Computer den Globus werfen, ganz anschaulich gesprochen. Natürlich wirft der Computer nicht in Wirklichkeit einen Globus oder eine Münze, sondern er zieht aus der Menge $\{0,1\}$ eine Zahl, und wir geben die Wahrscheinlichkeit für jedes der beiden Elemente vor, z.B. jeweils 50%.¹⁰.

! Wichtig

Simulationsdaten geben Aufschluss, welche Daten (wie oft Wasser) man bei einem bestimmten Modell, p, N , erwarten kann. Münzwürfe - und analoge Versuche, wie Globuswürfe - kann man in R mit `rbinom` erstellen (simulieren).

7.3.2 Traue niemals einem Golem (einem Modell)



Abbildung 7.4: Never trust a Golem

Quelle: <https://imgflip.com/i/5qmhma>

Immer prüfen und wachsam bleiben:

- (Inwieweit) decken sich die simulierten Daten mit den tatsächlichen Beobachtungen?
- Wie realistisch sind die Modellannahmen?
- Kann man das Modell aus verschiedenen Perspektiven prüfen?

7.4 Mit guten Simulationen kommt man den wahren Werten nahe

Warum nicht $n = 10^6$ neue Globusversuche simulieren¹¹:

¹⁰Übrigens können Computer nicht echten Zufall erzeugen (das kann vermutlich niemand), aber durch gewisse verzwickte Rechnungen sind die Zahlen, die der Computer uns präsentiert, nicht oder kaum vom “Zufall” zu unterscheiden, also z.B. gleichverteilt ohne besondere Muster.

¹¹Wer R nicht mag, ist eingeladen, diesen Versuch von Hand mit selbstgebastelten Globusbällen zu wiederholen.

```

draws <-
  tibble(
    draws = rbinom(1e6, size = 2, prob = .7))

draws %>%
  count(draws) %>%
  mutate(prop = n / sum(n))
## # A tibble: 3 x 3
##   draws     n     prop
##   <int> <int> <dbl>
## 1     0 89770 0.0898
## 2     1 420629 0.421
## 3     2 489601 0.490

```

Diese simulierten Häufigkeiten sind sehr ähnlich zu den theoretisch bestimmten Häufigkeiten mit `dbinom`: Unser Modell liefert plausible Vorhersagen¹².

```

dbinom(0:2, size = 2, prob = .7)
## [1] 0.09 0.42 0.49

```

7.5 Stichprobenverteilung

Wir ziehen viele ($n = 10^6$) Stichproben für unseren typischen Globusversuch: $N = 9$ Globuswürfe mit $p = 0.7$.

Wie viele Wasser (W) erhalten wir wohl typischerweise in diesem Versuch? Die Verteilung der zu erwartenden Treffer ist in Abbildung 7.5 dargestellt.

```

n_draws <- 1e6

draws <-
  tibble(draws = rbinom(n_draws, size = 9, prob = .7))

plot1 <-
  draws %>%
  ggplot(aes(x = draws)) +
  geom_histogram()

```

Die *Stichprobenverteilung* zeigt, welche Stichprobendaten laut unserem Modell (einem bestimmten Parameterwert) zu erwarten sind. Wir können jetzt prüfen, ob die echten Daten zu den Vorhersagen des Modells passen.

¹²Braver Golem!

Stichprobenverteilung für $n=9$ und $p=.7$ (binomial verteilt)

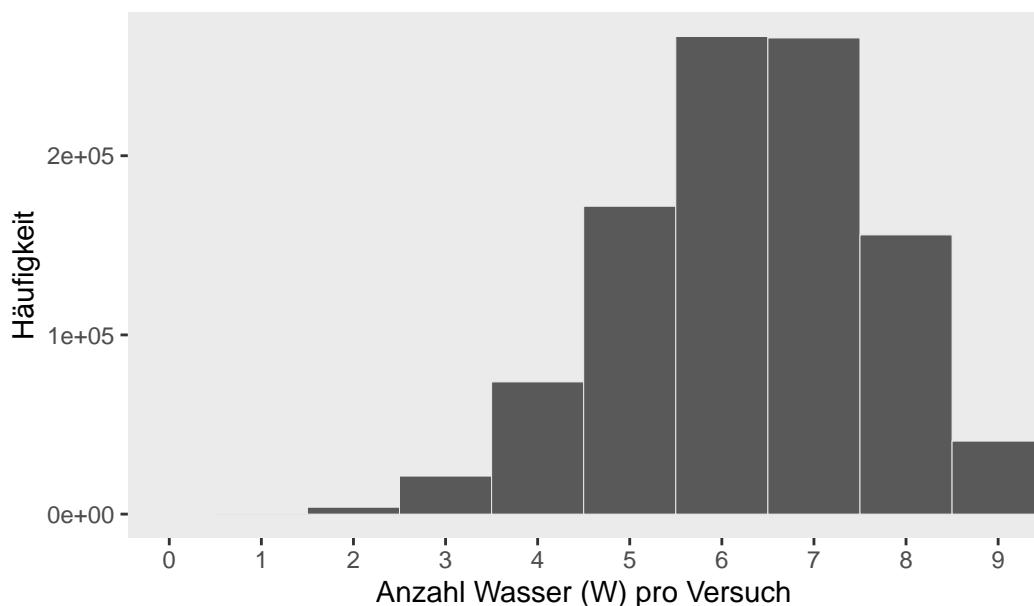


Abbildung 7.5: Anteile (Wahrscheinlichkeit), die man für jede Wasserzahl in unserem Globusversuch erwarten kann

i Hinweis

Die Stichprobenverteilung ist keine empirische Verteilung: Wir führen diese vielen Versuche nicht wirklich durch^a, wir simulieren sie nur am Computer.

^aNur die extremen Bastelfreunde machen das

7.6 Die Posterior-Prädiktiv-Verteilung (PPV)

7.6.1 Was ist die PPV und wozu ist sie gut?

Unsere Stichprobenverteilung zeigt, welche Trefferzahlen bei einem bestimmten Parameterwert, z.B. $\pi = .7$ in welchen Anteilen zu erwarten sind. Allerdings sind wir uns ja nicht sicher, dass der Wasseranteil genau 70% beträgt. Unser (Un-)Wissen über den Wasseranteil wird ja gerade in der Post-Verteilung gespeichert.

Um eine ehrliche(re) Antwort auf die Frage zu erhalten, wie viele Treffer¹³ zu erhalten ist, müssen wir die Post-Verteilung berücksichtigen.

Wir brauchen eine Stichprobenverteilung für *jeden* Wert der Post-Verteilung. Wenn wir dann die resultierenden Stichprobenverteilungen mitteln, haben wir einen ehrlichen Überblick

¹³Im Globusversuch ist Wasser der “Treffer”; in einem Münzwurf-Versuch könnte “Kopf” der Treffer und die Anzahl der geworfenen Köpfe die Trefferzahl sein

über die zu erwartenden Trefferzahlen. Dabei sollten wir natürlich wahrscheinliche Parameterwerte höher gewichtet als unwahrscheinliche. So sollte etwa der (hoch wahrscheinliche) Wasseranteil von 70% ein hohes Gewicht beim Mitteln der Stichprobenverteilung erhalten; der sehr unwahrscheinliche Wasseranteil¹⁴ von 1% Wasser, sollte entsprechend weniger gewichtet werden, beim Zusammenfassen (d.h. Mittelwert bilden) der Stichprobenverteilungen.

Die resultierende Verteilung - gemittelte Stichprobenverteilungen über alle Werte der Post-Verteilungen - nennt man *Posterior-Prädiktiv-Verteilung* (PPV).

! Wichtig

Die PPV entsteht als gewichteter Mittelwert der Stichprobenverteilungen. Die Gewichte sind die Wahrscheinlichkeiten (bzw. Wahrscheinlichkeitsdichten) der Post-Verteilung.

Beispiel 7.3 (Magnus Lagrange braucht die PVV). Im Jahre 10^{99} wird das Universum von Magnus Lagrange regiert. Die Weltraumbehörde, für die Sie arbeiten, ist ihm unterstellt. Der Regent findet Ihre Untersuchungen zwar ganz nett, aber leider versteht er keine Wahrscheinlichkeit. Ist ihm zu abstrakt, sagt er. „Oder können Sie mir mal so eine Wahrscheinlichkeit in die Hand geben? Können Sie sagen, Achtung, da hinten rennt eine Wahrscheinlichkeit, fang sie!“ Magnus ist also ein Freund für des Konkreten. Einige einflussreiche Gruppen an Statistik [unterstützen diese Haltung](#)

Jedenfalls hätte Magnus gerne eine Aussage wie „Vermutlich sehen wir beim nächsten Versuch irgendwas zwischen 4 und 7 Treffern“.

Natürlich haben Sie den Anspruch, eine wissenschaftlich belastbare Aussage zu tätigen.

Was nun? Sie müssen sozusagen die Post-Verteilung in eine Post-Verteilung der Beobachtungen, also der konkreten Werte - in diesem Fall die Anzahl der Wassertreffer - übersetzen. Genau das macht die PPV für Sie!

7.6.2 Visualisierung der PPV

Der Prozess des gewichteten Zusammenfassens der Stichprobenverteilungen ist in Abbildung 7.6 dargestellt.

Quelle: McElreath (2020)

¹⁴zumindest laut unserer Post-Verteilung

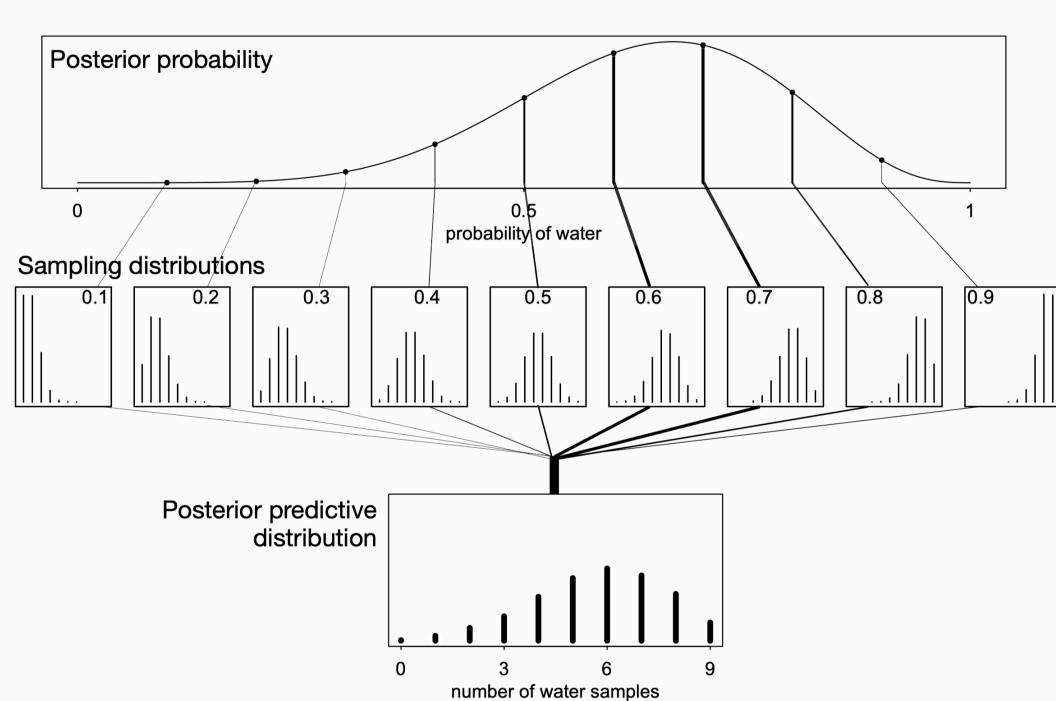


Abbildung 7.6: PPV als gewichtetes Kombinieren der Stichprobenverteilungen

7.6.3 PPV berechnen

Die PPV für unseren Standard-Globusversuch ($N = 9$) berechnen wir so:

Wir berechnen viele (z.B. 10^4) Stichprobenverteilungen. Dabei müssen wir jedes Mal fragen, wie groß die Wahrscheinlichkeit π für Wasser¹⁵ ist. Wasseranteile π , die laut Post-Verteilung wahrscheinlich sind, müssen wir entsprechend *oft* als Parameterwert (π) der Stichprobenverteilung verwenden; umgekehrt dürfen wir nur wenige Stichprobenverteilungen für unwahrscheinliche Parameterwerte erstellen.

Beispielsweise würden wir *viele* Stichprobenverteilungen für $\pi = .7$ erstellen; für $\pi = 0.01$ würden wir *wenige* Stichprobenverteilungen erstellen, s. Abbildung 7.6.

Glücklicherweise spiegelt unsere Stichproben-Postverteilung `samples` wahrscheinlichere Parameterwerte wieder, indem wahrscheinlichere Parameterwerte häufiger vorkommen.

i Hinweis

Wahrscheinliche Parameterwerte kommen in der Stichproben-Postverteilung `samples` häufiger vor. Die Häufigkeit der Parameterwerte spiegelt die Wahrscheinlichkeit der jeweiligen Parameterwerte in der (theoretischen) Postverteilung wider.

¹⁵d.h. einen Treffer

Schauen Sie sich vielleicht zur Erinnerung noch einmal die Definition von `samples` an, s. Listing 6.1. Tabelle `samples`, die aus Stichproben aus der Post-Verteilung besteht, ist (in Auszügen) in Tabelle 6.2 dargestellt. Wie die Post-Verteilung auf Basis von Stichproben dann aussieht sieht man in Abbildung 6.3. Globusversuche kann man mit `rbinom` simulieren, s. Kapitel 7.3.1.

Wir simulieren also viele (z.B 10^4) Globusversuche, jeweils mit $N = 9$ Würfen. Wahrscheinliche Parameterwerte, etwa $\pi = 7$, sollen häufiger verwendet werden (bei unseren vielen Globusversuchen) als unwahrscheinliche.

Praktischerweise sind die Werte in der Spalte `p_grid` in `samples` so häufig vertreten, wie ihre Wahrscheinlichkeit es erwarten lässt. Hier ist ein Auszug aus `samples`:

```
samples %>%
  select(p_grid) %>%
  slice_head(n = 10)
## # A tibble: 10 x 1
##   p_grid
##   <dbl>
## 1 0.778
## 2 0.556
## 3 0.667
## 4 0.667
## 5 0.556
## 6 0.778
## 7 0.889
## 8 0.667
## 9 0.667
## 10 0.667
```

Wie man sieht, sind wahrscheinliche Parameterwerte häufiger vertreten.¹⁶

`p_grid` ist also eine Liste¹⁷ von Parameterwerten, deren Häufigkeit die Wahrscheinlichkeit der Parameterwerte gewichtet.

Auf dieser Basis können wir die PPV erstellen:

```
ppv <-
  rbinom(1e4,
         size = 9,
         prob = samples$p_grid) %>%
  as_tibble()

head(ppv)
```

¹⁶An dieser Stelle sollten Sie sichd die ganze Spalte `p_grid` anschauen, um sich von dieser Behauptung mit eigenen Augen zu überzeugen.

¹⁷technisch in R ein *Vektor*

```
## # A tibble: 6 x 1
##   value
##   <int>
## 1     8
## 2     4
## 3     6
## 4     6
## 5     4
## 6     5
```

Schauen wir uns ein Histogramm aller Trefferzahlen an, s. Abbildung 7.7.¹⁸

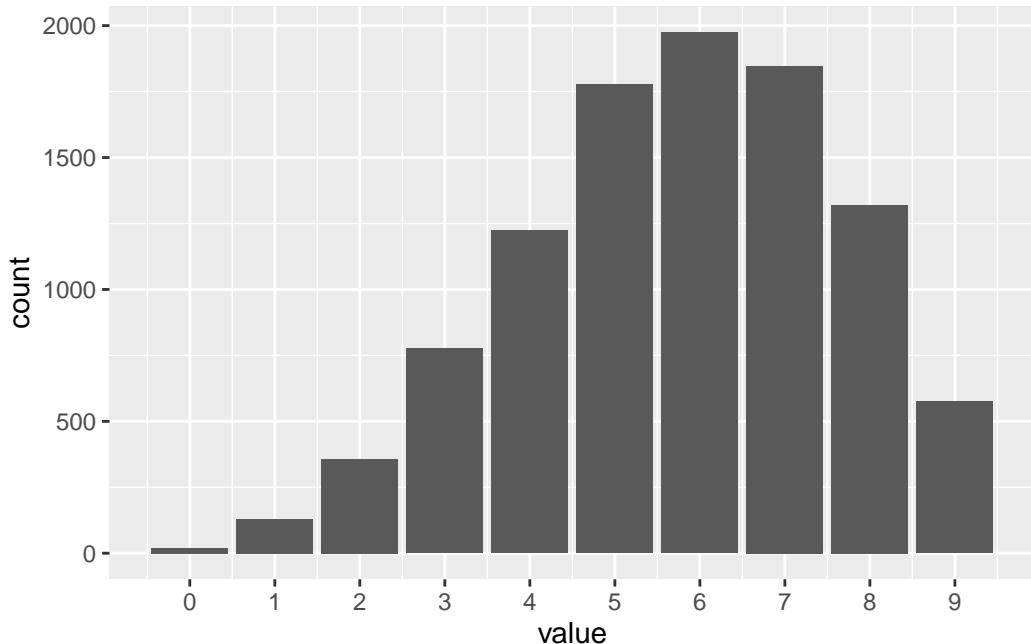


Abbildung 7.7: Die PPV für unseren Standard-Globusversuch (N=9)

Die PPV unseres Modells zeigt uns (Abbildung 7.7), dass wir in künftigen Versuchen zumeist 6 Treffer zu erwarten haben. Aber ein relativer breiter Bereich an Treffern ist ebenfalls gut laut unserer PPV erwartbar.

! Wichtig

Die PPV zeigt, welche Beobachtungen laut unserem Modell häufig und welche selten sind. Die PPV zeigt keine Parameterwerte, sondern welche Daten (Beobachtungen, Wasserzahlen) wir in künftigen Versuchen wie häufig erwarten können.

¹⁸Es kann auch dem Verständnis helfen, dass Sie sich alle Werte der Tabelle ppv selber in Ruhe anschauen, um sich zu überzeugen, welche Wasserzahlen (Trefferzahlen) häufiger und welche seltener vorkommen.

Beispiel 7.4 (Der nächste Planet). Nur zum Spaß spulen wir kurz die Zeit im Universum vor, sagen wir so 10^{99} Jahre. Sie arbeiten bei einer Raumfahrtbehörde, die nach neuen Planeten sucht. Nun wurde ein aussichtsreicher Planet gesichtet. Ihre Behörde hat eine Studie gestartet, im Rahmen derer 9 Sonden zu diesem (weit entfernten) Planeten geschossen sind. Von den 9 Sonden sind 6 im Wasser gelandet, was aus Gründen intergalaktischer Wasserknappheit eine gute Nachricht ist.

“Der nächste Planet wird sicher 6 von 9 Wassertreffer erzielen!”

– Presse-Chefi der intergalaktischer SpaceY Raumfahrtsbehörde

Jetzt plant Ihre Behörde den Versuch zu wiederholen: Wieder sollen 9 Sonden zu diesem Planeten geschossen werden. Dis Presse-Chefi¹⁹ tönt vollmundig: “Ich bin sicher, dass wir wieder 6 von 9 Treffer, also 6 von 9 Mal Wasser, haben werden!”.

Kann man diese Aussage mit (hoher) Sicherheit leisten? Perfekte Sicherheit gibt es bekanntlich nur, was Tod und Steuern betrifft, aber kann diese Aussage mit zumindest hoher Sicherheit geleistet werden?

Nein, die PPV (Abbildung 7.7) zeigt deutlich, dass unser Wissen nicht ausreicht, um präzise Vorhersagen über künftige Ausgänge des Versuchs zu leisten. So sind auch 5 oder 7 Treffer gut möglich. Auch 4 oder 8 Treffer sind nicht so selten. Sogar 9 Treffer sind nicht super selten.

Dis Presse-Chefi Ihrer Behörde sollte also den Mund nicht so voll nehmen.

7.7 Fazit

7.7.1 Vorhersagen sind schwierig

... gerade wenn sie die Zukunft betreffen, so ein Sprichwort.

Das zeigt uns die PPV: Der PPV unseres Modells gelingt es zwar, der theoretisch wahrscheinlichste Parameterwert ist auch der häufigste in unseren Stichproben, aber die Vorhersagen haben eine große Streuung, bergen also recht hohe Ungewissheit. Die PPV zeigt also, welche Beobachtungen laut unserem Modell künftig zu erwarten sind, s. Abbildung 7.7.

Würde man die Vorhersagen nur anhand eines bestimmten Parameterwertes (z.B $p = 0.6$) vornehmen, hätten die Vorhersagen *zu wenig Streuung* in den Vorhersagen, würden also die Ungewissheit nicht ausreichend abbilden. Es würde *Übergewissheit* (Overconfidence, Overfitting) resultieren.

Wir brauchen die PPV. Ohne die PPV können wir nicht seriös abschätzen, wie viel Ungewissheit in unseren Vorhersagen steckt.

¹⁹In der Zeit dieses Beispiels ist es üblich, kein fixes Geschlecht zu haben

7.7.2 Zwei Arten von Ungewissheit in Vorhersagen von Modellen

1. *Ungewissheit innerhalb des Modells* (“intrinsische” Ungewissheit): Auch wenn der (oder die) Modellparameter eines Modells mit Sicherheit bekannt sind, so bleibt Unsicherheit, welche Beobachtung eintreten wird: Auch wenn man sicher weiß, dass $p = 1/4$ Murmeln blau sind, so kann man nicht sicher sagen, welche Farbe die nächste Murmel haben wird (Ausnahme: $p = 1$ oder $p = 0$).
2. *Ungewissheit in den Modellparametern*: Wir sind uns nicht sicher, welchen Wert p (bzw. die Modellparameter) haben. Diese Unsicherheit ist in der Post-Verteilung dargestellt.

Um zu realistischen Vorhersagen zu kommen, möchte man beide Arten von Ungewissheit berücksichtigen: Das macht die *Posteriori-Prädiktiv-Verteilung (PPV)*.

Die PPV zeigt, welche Daten das Modell vorhersagt (prädiktiv) und mit welcher Häufigkeit, basierend auf der Post-Verteilung.

Hinweis

Der Unterschied zwischen der Post-Verteilung und der PPV ist erstmal, dass die PPV *Ausprägungen* in ihrer Wahrscheinlichkeit bemisst, also z.B. wie wahrscheinlich 4 von 9 Wassertreffern sind. Die Post-Verteilung bemisst die Wahrscheinlichkeit von Parameterwerten, also z.B. des Wasseranteils.

Etwas tiefer betrachtet zeigt die PPV zwei Arten von Ungewissheit, die Post-Verteilung nur eine. Die PPV zeigt erstens die Ungewissheit zur Verteilung des Parameters (wie die Post-Verteilung), aber auch noch die intrinsische Ungewissheit. Denn auch wenn wir keine Ungewissheit zum Parameter hätten, bliebe Ungewissheit, welche Beobachtungen sich manifestieren. Insofern ist die PPV “ehrlicher”, sie spiegelt die Ungewissheit zu den Beobachtungen wider.

7.7.3 Vergleich der Verteilungen

?@fig-post-pred-ppv-anim stellt die in diesem Kapitel diskutierten Verteilungen gegenüber:

- Links - *Posterior-Verteilung*: Wahrscheinlichkeiten der Parameterwerte
- Mitte - *Stichprobenverteilung*: Wahrscheinlichkeiten der Beobachtungen gegeben eines bestimmten Parameterwertes
- Rechts - *Posterior-Prädiktiv-Verteilung*: Wahrscheinlichkeiten der Beobachtungen unter Berücksichtigung der Unsicherheit der Posteriori-Verteilung

Quelle: R. McElreath

7.7.4 So viele Verteilungen...

- Die *Posteriori-Verteilung* gibt Aufschluss zur Häufigkeit (Wahrscheinlichkeit) von Parameterwerten:
 - Wie wahrscheinlich ist es, dass “in Wirklichkeit” der Wasseranteil 70% beträgt, also $\pi = .7$
 - In der Wissenschaft ist man meist an den Parametern interessiert.
- Die *PPV* gibt Aufschluss zur Häufigkeit von neuen Beobachtungen:
 - Welche Beobachtungen (wie viele Wasser/Treffer) sind in Zukunft, bei erneuter Durchführung, zu erwarten.
 - Für die Praxis kann das eine interessante Frage sein.
- Der *Likelihood* gibt Aufschluss, wie gut eine bestimmte Hypothese die Datenlage erklärt.
 - Wie gut passt die Hypothese $\pi = 0.7$ auf die Datenlage 6 von 9 Treffern beim Globusversuch?
 - Der Likelihood kann aus der Stichprobenverteilung herausgelesen werden.

7.8 Aufgaben

1. [Zwielichter Dozent-Bayes](#)
2. [Warum Bayes?](#)
3. [subjektiv-Bayes](#)
4. [Likelihood2](#)
5. [Anteil-Apple](#)
6. [ReThink3m1](#)
7. [ReThink3m2](#)
8. [ReThink3m3](#)
9. [ReThink3m4](#)
10. [ReThink3m5](#)
11. [Quiz zu Verteilungen](#)

8 Gauss-Modelle



Abbildung 8.1: BayesStart!

8.1 Lernsteuerung

8.1.1 Lernziele

Nach Absolvieren des jeweiligen Kapitels sollen folgende Lernziele erreicht sein.

Sie können ...

- ein Gaussmodell spezifizieren und in R berechnen
- an Beispielen verdeutlichen, wie sich eine vage bzw. eine informationsreiche Priori-Verteilung auf die Posteriori-Verteilung auswirkt

8.1.2 Benötigte R-Pakete

Für `rstanarm` wird ggf. [weitere Software](#) benötigt.

i Hinweis

Software, und das sind R-Pakete, müssen Sie nur einmalig installieren. Aber bei jedem Start von R bzw. RStudio müssen Sie die (benötigten!) Pakete starten.

```
library(tidyverse)
library(rstanarm)
library(easystats)
```

8.1.3 Begleitvideos

- [Teil 1](#)
- [Teil 2](#)

8.2 Wie groß sind die !Kung San?

Dieser Abschnitt basiert auf McElreath (2020), Kap. 4.3.

8.2.1 !Kung San

In diesem Abschnitt untersuchen wir eine Forschungsfrage in Zusammenhang mit dem Volk der !Kung, s. Abbildung 8.2.

The Kung are one of the San peoples who live mostly on the western edge of the Kalahari desert, Ovamboland (northern Namibia and southern Angola), and Botswana. The names Kung (Xun) and Ju are variant words for ‘people’, preferred by different Kung groups. This band level society used traditional methods of hunting and gathering for subsistence up until the 1970s. Today, the great majority of Kung people live in the villages of Bantu pastoralists and European ranchers.

[Quelle](#)

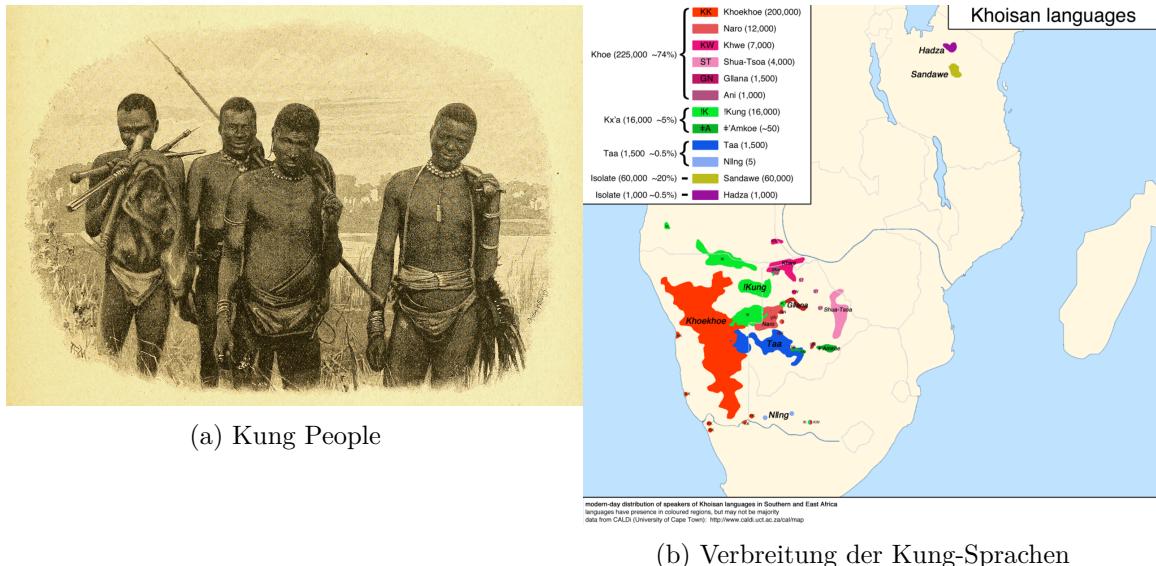


Abbildung 8.2: Die !Kung im südlichen Afrika

Quelle: Internet Archive Book Images, No restrictions, via Wikimedia Commons

Quelle: By Andrewwik.0 - Own work, CC BY-SA 4.0,

8.2.2 !Kung Data

Zuerst laden wir die Daten; Quelle der Daten ist McElreath (2020) mit Bezug auf Howell.

Datenquelle

```
Kung_path <-  
  "https://raw.githubusercontent.com/sebastiansauer/Lehre/main/data/Howell1a.csv"  
  
d <- data_read(Kung_path) # aus dem Paket `easystats`  
  
head(d)  
##   height  weight age male  
## 1 151.765 47.82561 63   1  
## 2 139.700 36.48581 63   0  
## 3 136.525 31.86484 65   0  
## 4 156.845 53.04191 41   1  
## 5 145.415 41.27687 51   0  
## 6 163.830 62.99259 35   1
```

Wir interessieren uns für die Größe der erwachsenen !Kung, also filtern wir die Daten entsprechend und speichern die neue Tabelle als d2.

```
d2 <- d %>%  
  filter(age >= 18)  
  
nrow(d2)  
## [1] 352
```

$N = 352$.

Lassen wir uns einige typische deskriptive Statistiken zum Datensatz ausgeben. `{easystats}` macht das tatsächlich recht easy.

```
describe_distribution(d2)
```

Variable	Mean	SD	IQR	Min	Max	Skewness	Kurtosis	n	n_Missing
height	154.60	7.74	12.06	136.53	179.07	0.15	-0.48	352.00	0
weight	44.99	6.46	9.19	31.07	62.99	0.13	-0.51	352.00	0
age	41.14	15.97	23.00	18.00	88.00	0.67	-0.21	352.00	0
male	0.47	0.50	1.00	0.00	1.00	0.13	-2.00	352.00	0

8.2.3 Wir gehen apriori von normalverteilter Größe Der !Kung aus

Forschungsfrage: Wie groß sind die erwachsenen !Kung im Durchschnitt?

Wir interessieren uns also für den Mittelwert der Körpergröße (erwachsener Kung beider Geschlechter), μ .



Abbildung 8.3: Mensch

Quelle: Own alterations andFile:SVG_Human_With_All_Organs.svg by Madhero88, CC BY-SA, via Wikimedia Commons 3.0

Wir sind uns über diesen Mittelwert nicht sicher¹, und unsere Ungewissheit quantifizieren wir anhand einer Normalverteilung mit Mittelwert von 178cm und Streuung von 20 cm:

$$\mu \sim \mathcal{N}(178, 20)$$

Warum 178 cm? Kein besonderer Grund. Hier wollen wir den Effekt verschiedener Priori-Werte untersuchen.² In einer echten Untersuchung sollte man immer einen inhaltlichen Grund für einen Priori-Wert haben. *Oder* man wählt “schwach informative” Prioris, wie das `{rstanarm}` tut: Damit lässt man kaum Vorab-Information in das Modell einfließen, aber man verhindert extreme Prioris, die meistens unsinnig sind (so wie eine SD von 100 Metern in diesem Fall).

i Hinweis

Wir haben zwar vorab nicht viel Wissen, aber auch nicht gar keines: Eine Gleichverteilung der Körpergrößen kommt nicht in Frage und ein vages Wissen zum Mittelwert haben wir auch. Darüber hinaus ist eine Normalverteilung nicht unplausibel.

8.3 Die Exponentialverteilung

8.3.1 Die Apfel-fällt-nicht-weit-vom-Stamm-Verteilung

Darf ich vorstellen ...

Bevor wir unser Kung-Modell spezifizieren können, sollten wir noch überlegen, welches Vorab-Wissen wir zur Streuung um den Mittelwert herum haben. Da wir uns nicht 100% sicher zur gesuchten Größe sind, müssen wir angeben, wie groß die Streuung um den Mittelwert sein soll. Hier werden wir eingestehen, dass wir uns auch nicht 100% sicher sind, wie groß die Streuung exakt ist. Also geben wir eine Verteilung für die Streuung an.

Etwas Wissen über diese Verteilung haben wir:

¹Darum machen wir hier ja die ganz Show!

²Der Autor des zugrundeliegenden Fachbuchs, Richard McElreath gibt 178cm als seine Körpergröße an.

- Eine Streuung muss positiv sein (es gibt keine negative Streuung).
- Eine Gleichverteilung der Streuung ist vielleicht möglich, aber nicht sehr plausibel.
- Wenn wir der Meinung sind, der Mittelwert betrage “ungefähr 178cm”, so halten wir 180cm für plausibel, aber 18000 cm für unmöglich und schon 200 für sehr unplausibel. Also: Je größer die Abweichung vom Mittelwert desto unplausibler.

Diese Anforderungen³ spiegeln sich in Abbildung 8.4 wider. Außerdem zeigt die Abbilung verschiedene Quantile, wie das 95%-Quantil, das bei 3 liegt; 95% der Werte dieser Verteilung sind also nicht größer als 3.

```
d <-  
  tibble(  
    x = seq(0, 5,.1),  
    y = dexp(x, rate = 1)  
)  
  
d_qs <-  
  tibble(  
    prob = c(0.05, .25, .50, .75, .95),  
    q = qexp(prob)  
)  
  
d %>%  
  ggplot(aes(x,y)) +  
  geom_line() +  
  geom_area(fill = "grey60") +  
  geom_vline(data = d_qs,  
             aes(xintercept = q)) +  
  geom_label(data = d_qs,  
             aes(x = q,  
                  label = prob,  
                  y = prob)) +  
  labs(  
    caption = "Vertikale Striche zeigen die Quantile für 5%, 25%, 50%, 75%, 95%",  
    y = "Dichte")
```

³“Desiderata”

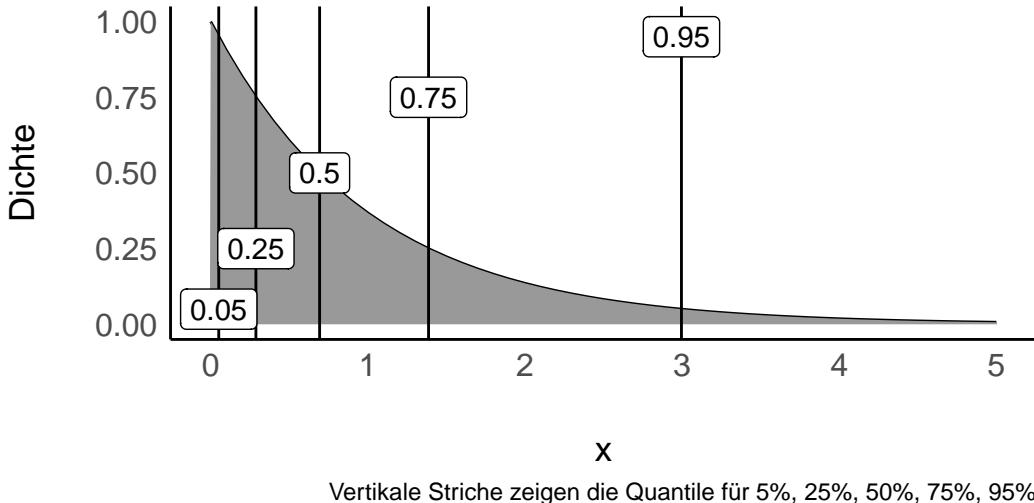


Abbildung 8.4: Die Exponentialverteilung mit einigen ihrer Quantilen

Für eine exponentialverteilte Variable X schreibt man auch:

$$X \sim \text{Exp}(1)$$

Eine Verteilung dieser Form nennt man *Exponentialverteilung*.

- Eine *Exponentialverteilung* ist nur für positive Werte, $x > 0$, definiert.
- Steigt X um eine *Einheit*, so ändert sich Y um einen *konstanten Faktor*.
- Sie hat nur einen Parameter, genannt *Rate* oder λ ("lambda").
- $\frac{1}{\lambda}$ gibt gleichzeitig Mittelwert und Streuung ("Gestrecktheit") der Verteilung an.
- Je größer die Rate λ , desto *kleiner* die Streuung und der Mittelwert der Verteilung.
- Je größer $1/\lambda$, desto *größer* die Streuung und der Mittelwert der Verteilung.

Ohne auf die mathematischen Eigenschaften im Detail einzugehen, halten wir fest, dass der Graph dieser Funktion gut zu unseren Plänen passt.

8.3.2 Visualisierung verschiedener Exponentialverteilungen

Schauen wir uns einige Beispiele von Exponentialverteilungen an. Unterschiede in Exponentiellverteilungen sind rein auf Unterschiede in λ (lambda) zurückzuführen, s. Abbildung 8.5.

Wie wir in Abbildung 8.5 sehen, könnte eine Exponentialverteilung mit $\lambda = 1/8$ grob passen.

i Hinweis

Die "richtigen" Priori-Verteilung zu finden, bzw. die richtigen Parameter für die Priori-Verteilung zu wählen, ist nicht möglich, denn es gibt nicht die eine, richtige Priori-

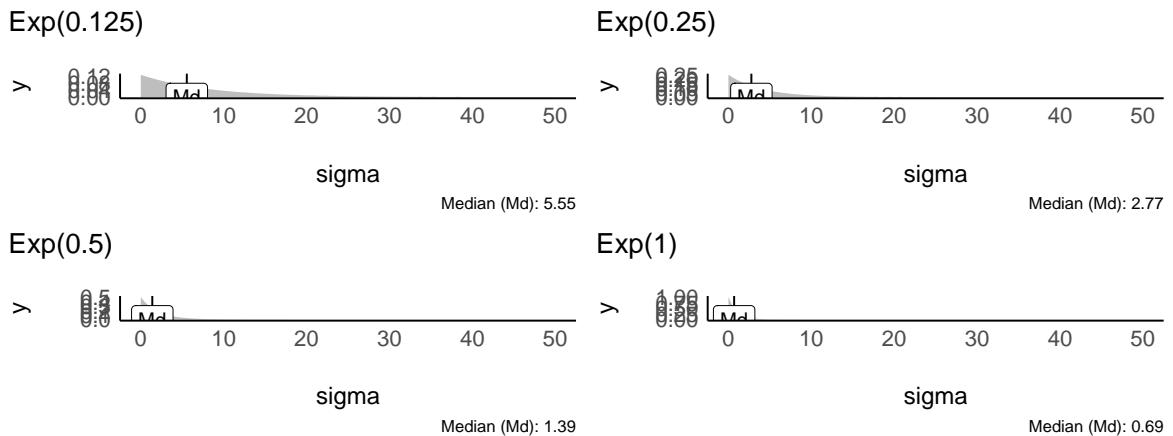


Abbildung 8.5: Beispiele von Exponentialverteilungen mit unterschiedlichem lambda

Verteilung. Eine “gut passende” Verteilung zu finden, ist häufig nicht leicht. Gut beraten ist man mit der Regel, im Zweifel lieber eine liberale Verteilung zu wählen, die einen breiteren Raum an möglichen Werten zulässt. Allerdings sollte man nicht das Baby mit dem Wasser auskippen und extreme Werte, wie mehrere Meter Körpergröße Streuung, erlauben.

Man kann sich die Quantile der Exponentialverteilung mit `qexp` ausgeben lassen, wobei mit `p` man den Wert der Verteilungsfunktion angibt, für den man das Quantil haben möchte. Mit `rate` wird λ bezeichnet.

Dieser Aufruf zum Beispiel:

```
qexp(p = .5, rate = 1/8)
## [1] 5.545177
```

Gibt uns die Verteilungsfunktion einer Exponentialverteilung mit Rate (λ) von $1/8$ zurück, ca. 5.5.

Die Grenzen der inneren 95% dieser Verteilung kann man sich so ausgeben lassen:

```
qexp(p = c(0.025, .975), rate = 1/8)
## [1] 0.2025425 29.5110356
```

Diese Grenzen scheinen hinreichend weit, das wir noch von den Daten überrascht werden können, aber schmal genug, um unsinnige Werte auszuschließen. Ein guter Start! Weiter geht's!

8.4 Unser Gauss-Modell der !Kung

8.4.1 Modelldefinition

Wir nehmen an, dass μ und h_i normalverteilt sind und σ exponentialverteilt (da notwendig positiv) ist:

Likelihood: $h_i \sim \mathcal{N}(\mu, \sigma)$

Prior für μ : $\mu \sim \mathcal{N}(178, 20)$

Prior für σ : $\sigma \sim \mathcal{E}(0, 0.1)$

Daher: 95%KI(μ) : 178 ± 40

In Abbildung 8.6 sind unsere Priori-Verteilungen visualisiert.

`mu ~ dnorm(178, 20)`

`sigma ~ dexp(0.1)`

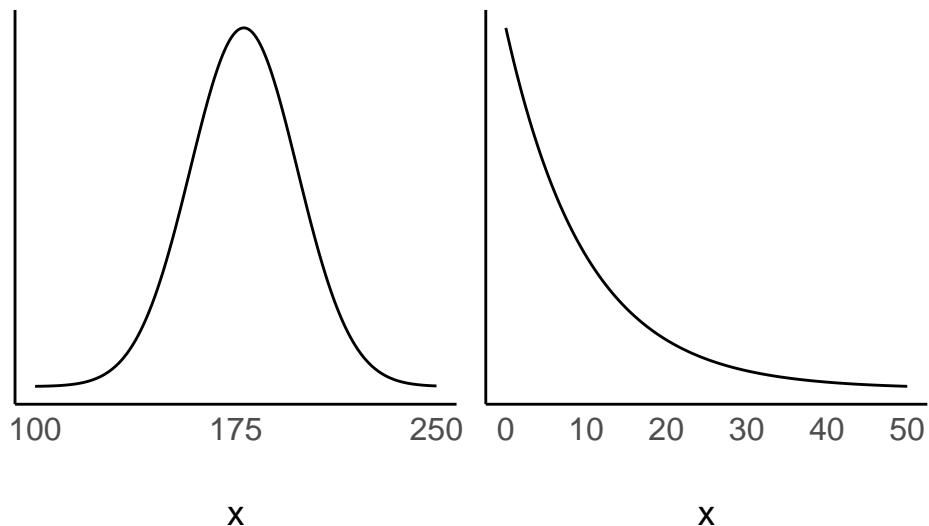


Abbildung 8.6: Unser (erstes) Kung-Modell

8.4.2 Priori gewichtet mit Likelihood ergibt Posteriori

Zu Erinnerung: Die Posteriori-Wahrscheinlichkeit ist das Ergebnis von Priori-Wahrscheinlichkeit und Likelihood.

Die Körpergrößen der einzelnen Personen h_i nehmen wir als normalverteilt an mit Mittelwert μ und Streuung σ :

$$h_i \sim \mathcal{N}(\mu, \sigma)$$

8.4.3 Prioris

Mittelwert der Größe ist normalverteilt mit $\mu = 178$ und $\sigma = 20$:

$$\mu \sim \mathcal{N}(178, 20)$$

Die Streuung σ der Größen ist exponentialverteilt mit $\lambda = 1/8$.

$$\sigma \sim \mathcal{E}(1/8)$$

8.4.4 Fertig!

Jetzt haben wir unser Modell definiert!

Weil es so schön ist, schreiben wir es hier noch einmal auf, Gleichung 8.1.

$$\begin{aligned} h_i &\sim \mathcal{N}(\mu, \sigma) \\ \mu &\sim \mathcal{N}(178, 20) \\ \sigma &\sim \mathcal{E}(1/8) \end{aligned} \tag{8.1}$$

Zur Berechnung nutzen wir jetzt dieses Mal aber *nicht* die Gittermethode (Bayes-Box), sondern lassen R die Arbeit verrichten.

Da gibt es einen neuen Golem, ziemlich kräftig der Bursche, der soll die Arbeit für uns tun. Der Golem hört auf den Namen `rstanarm`⁴.

8.5 Zufällige Motivationsseite

8.6 Posteriori-Verteilung des Größen-Modells, m41

Okay, Golem, an die Arbeit! Berechne uns das Kung-Modell! Nennen wir das Modell `m41`⁵.

```
m41 <- stan_glm(height ~ 1, data = d2, refresh = 0)
m41_post <- as_tibble(m41) # Modellergebnis in Tabelle umwandeln
names(m41_post) <- c("mu", "sigma") # schönere Namen für die Spalten
```

⁴Hey, ich habe ihn diese Namen nicht gegeben.

⁵m wie Modell und 4, weil das Modell in Kapitel 4 von McElreath (2020) in ähnlicher Form berichtet wird, und 1 weil es unsere erste Variante dieses Modells ist.

Das Argument `refresh = 0` verhindert, dass die Details zum Ziehen der Stichproben am Bildschirm ausgegeben werden. Ich finde diese Ausgabe meist nicht informativ, so dass ich sie lieber unterdrücke.

`stan_glm` ist eine Funktion, mit der man Regressionsmodelle berechnen kann. Nun haben wir in diesem Fall kein “richtiges” Regressionsmodell. Man könnte sagen, wir haben eine AV (Körpergröße), aber keine UV (keine Prädiktoren). Glücklicherweise können wir auch solche “armen” Regressionsmodelle formulieren:

`av ~ 1` bzw. in unserem Beispiel `height ~ 1` bedeutet, dass man nur die Verteilung der AV berechnen möchte, aber keine Prädiktoren hat (das soll die 1 symbolisieren).

Für das Modell `m41` haben wir *keine* Prioris spezifiziert. Wir greifen damit auf die Voreinstellung der Prioris von `rstanarm` zurück. Das ist ok, aber wenn Sie Vorab-Wissen haben, sollten Sie das an `rstanarm` weitergeben, weil es ja schade wäre, wenn Sie Wissen haben, das von Ihrem Modell nicht genutzt wird.

Plotten wir mal die Posteriori-Verteilung von `m41`, s. Abbildung 8.7.

```
m41_post %>%
  ggplot() +
  aes(x = mu, y = sigma) %>%
  geom_hex() +
  scale_fill_viridis_c()
```

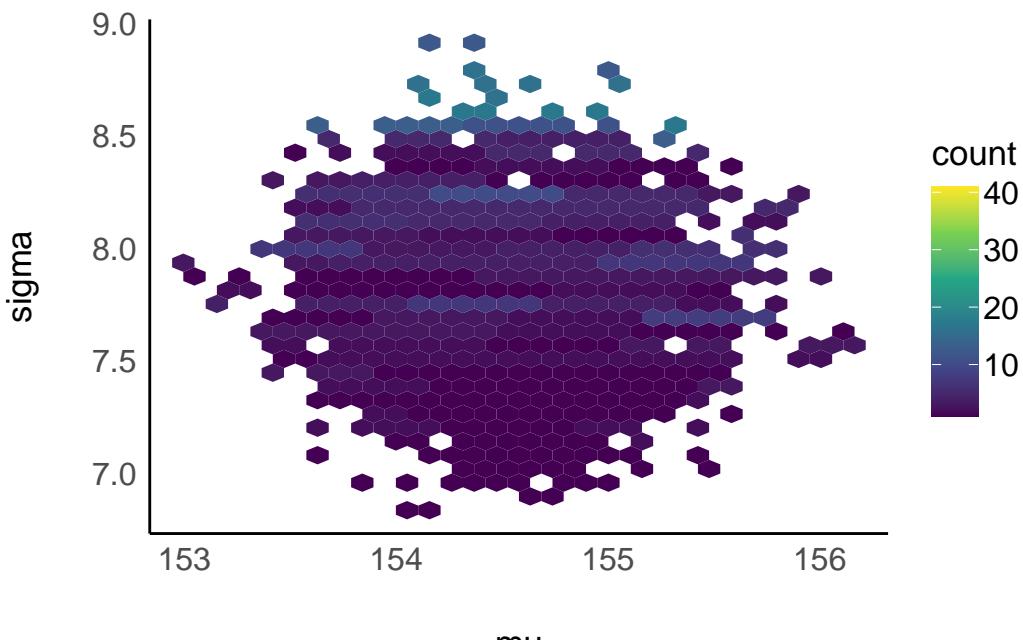


Abbildung 8.7: Die gemeinsame Verteilung von Mittelwert und Streuung.

Da das Modell *zwei* Parameter hat, können wir auch beide gleichzeitig plotten. Wie man

sieht, sind die beiden Parameter unkorreliert. In anderen Modellen können die Parameter korreliert sein.

Abbildung 8.7 erlaubt uns, für jede Kombination von Mittelwert und Streuung zu fragen, wie wahrscheinlich diese bestimmte Kombination ist.

Hier sind noch zwei andere Visualisierungen der Post-Verteilung von $m42$, s. Abbildung 8.8.

Die gemeinsame Verteilung von mu und Sigma als Punkte

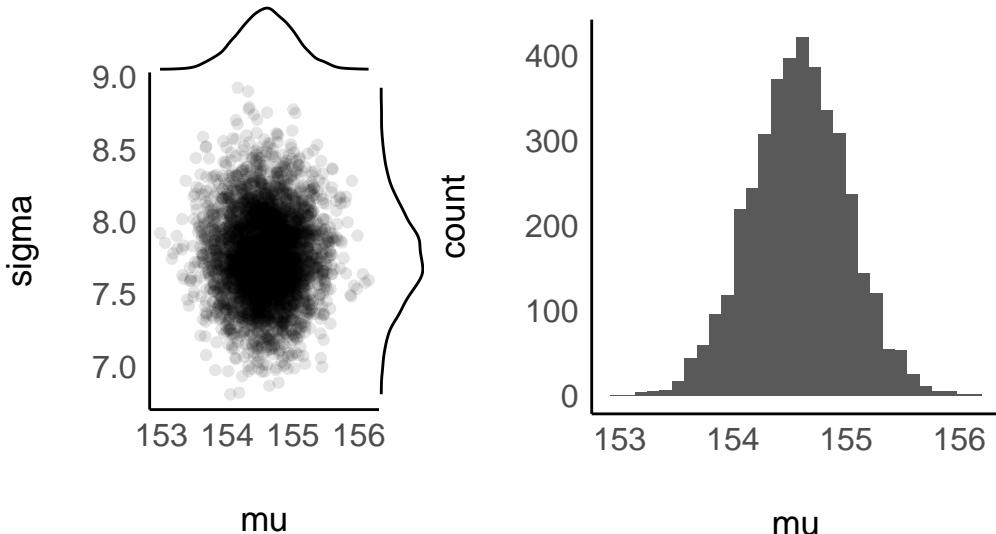


Abbildung 8.8: Die Postverteilung in unterschiedlicher Darstellung

Natürlich können wir auch nur einen Parameter plotten.

Fassen wir die Ergebnisse dieses Modells zusammen:

- Wir bekommen eine Wahrscheinlichkeitsverteilung für μ und eine für σ (bzw. eine zweidimensionale Verteilung, für die μ, σ -Paare).
- Trotz des eher vagen Priors ist die Streuung Posteriori-Werte für μ und σ klein: Die große Stichprobe hat die Priori-Werte überstimmt.
- Ziehen wir Stichproben aus der Posteriori-Verteilung, so können wir interessante Fragen stellen.

8.6.1 Hallo, Posteriori-Verteilung

... wir hätten da mal ein paar Fragen an Sie.

1. Mit welcher Wahrscheinlichkeit ist die *mittlere !Kung-Person* größer als 1,55m?
2. Welche mittlere Körpergröße wird mit 95% Wahrscheinlichkeit nicht überschritten, laut dem Modell?

3. In welchem 90%-PI liegt μ vermutlich?
4. Mit welcher Unsicherheit ist die Schätzung der mittleren Körpergröße behaftet?
5. Was ist der mediane Schätzwert der mittleren Körpergröße, sozusagen der “Best Guess”?

Antworten folgen etwas weiter unten.

Abschließend, eigentlich nur Spielerei, noch eine andere Visualisierung der Post-Verteilung von μ und von σ , Abbildung 8.9.

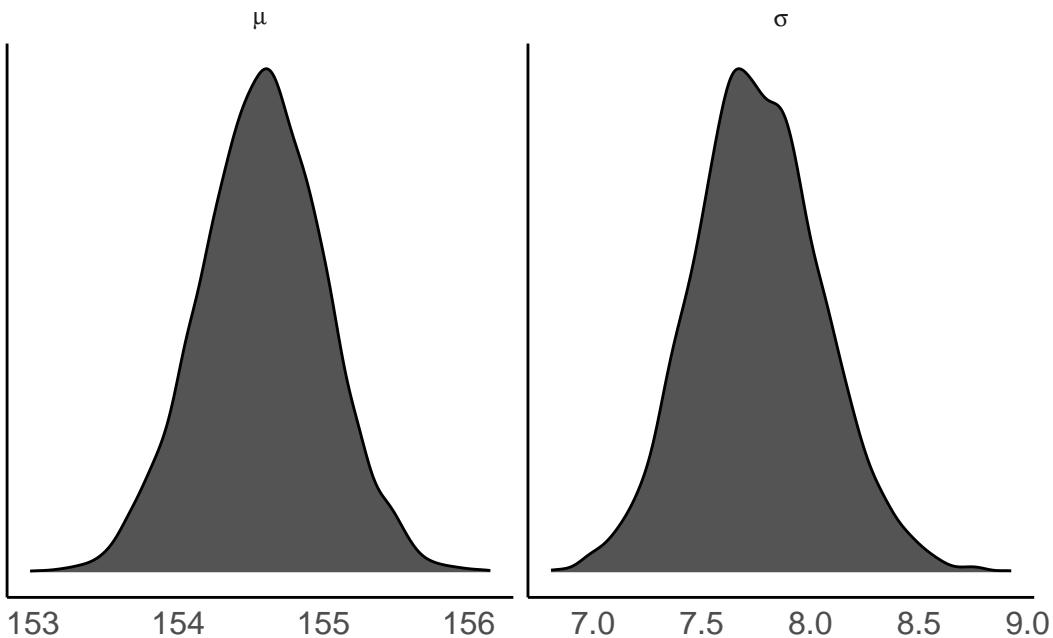


Abbildung 8.9: Die beiden Randverteilungen der Post-Verteilungen, d.h. die Verteilungen für mu und für sigma

8.6.2 Posteriori-Stichproben mit `stan_glm()` berechnen

Mit `stan_glm()` können wir komfortabel die Posteriori-Verteilung berechnen. Die Gittermethode wird nicht verwendet, aber die Ergebnisse sind - in bestimmten Situationen - ähnlich. Es werden aber auch viele Stichproben simuliert (sog. MCMC-Methode). Gibt man keine Priori-Werte an, so greift die Funktion auf Standardwerte zurück.

Grob gesagt berechnen wir die Post-Verteilung mit `stan_glm` so:

```
library(rstanarm) # Paket muss gestartet sein.

# berechnet Post.-Vert.:
stan_glm(
  # modelldefinition:
```

```

AV ~ UV,
# Datensatz:
data = meine_daten
)

```

Modelldefinition:

$h_i \sim \mathcal{N}(\mu, \sigma)$, Likelihood

$\mu \sim \mathcal{N}(155, 19)$, Prior zum Größenmittelwert

$\sigma \sim \mathcal{E}(0.125)$, Prior zur Streuung der Größen

8.6.3 Ausgabe von stan_glm()

Wir können, wie wir es oben getan haben, uns die Stichproben der Post-Verteilung ausgeben lassen, und diese z.B. plotten.

Wir können es aber auch komfortabler haben ... Mit dem Befehl `parameters` kann man sich die geschätzten Parameterwerte einfach ausgeben lassen.

```

m41 <- stan_glm(height ~ 1, data = d2, refresh = 0) # aus Paket rstanarm
parameters(m41) # aus Paket `easystats`
```

Tabelle 8.2: Fixed effects

Parameter	Median	95% CI	pd	Rhat	ESS	Prior
(Intercept)	154.60	(153.82, 155.41)	100%	1.001	2702.00	Normal (154.60 +- 19.36)

Das Wesentliche: Unser Golem schätzt den Größenmittelwert der Kung auf ca. 155cm bzw. auf einen Bereich von etwa 153.8187886 bis 155.408954 schätzt.

Informativ ist vielleicht noch, dass wir den Prior erfahren, der im Modell verwendet wurde. Dazu später mehr.

In dieser Ausgabe sind ein paar Angaben, die wir nicht verstehen, wie pd, Rhat und ESS. Kein Problem: Einfach ignorieren :-)

Wer Näheres wissen will, findet hier einen Anfang. Außerdem sei an McElreath (2020) und Gelman, Hill, und Vehtari (2021) verwiesen.

8.7 Wie tickt `stan_glm()`?



Quelle

Hier ein paar Kernimnfos zu `stan_glm`:

- *Stan* ist eine Software zur Berechnung von Bayesmodellen; das Paket `rstanarm` stellt Stan für uns bereit.
- `stan_glm()` ist für die Berechnung von Regressionsmodellen ausgelegt.
- Will man nur die Verteilung einer Variablen (wie `heights`) schätzen, so hat man man ... eine Regression ohne Prädiktor.
- Eine Regression ohne Prädiktor schreibt man auf Errisch so: $y \sim 1$. Die 1 steht also für die nicht vorhandene UV; y meint die AV (`height`).
- `(Intercept)` (Achsenabschnitt) gibt den Mittelwert an.

Mehr findet sich in der [Dokumentation von RstanArm](#).

8.7.1 Schätzwerte zu den Modellparameter

Die Parameter eines Modells sind die Größen, für die wir eine Priori-Verteilung annehmen sowie einen Likelihood und dann aus den Daten schätzen. Ich sage schätzen um hervorzuheben, dass wir die wahren Werte nicht kennen, sondern nur eine Vermutung haben, unsere Ungewissheit vorab also (wie immer) in der Priori-Verteilung festnageln und unsere Ungewissheit nach Kenntnis der Daten in der Posteriori-Verteilung quantifizieren.

Wie gerade gesehen, lassen sich die Modellparameter (bzw. genauer gesagt deren Schätzungen) einfach mit `parameters(modelname)` auslesen.

8.7.2 Stichproben aus der Posteriori-Verteilung ziehen

Wie wir es vom Globusversuch gewohnt sind, können wir aber auch Stichproben aus der Post-Verteilung ziehen.

Hier die ersten paar Zeilen von `post_m41`:

```
post_m41 <- as_tibble(m41)
head(post_m41)
## # A tibble: 6 x 2
##   `(Intercept)` `sigma`
##       <dbl>     <dbl>
## 1          154.    7.31
```

```

## 2      154. 7.29
## 3      155. 8.06
## 4      155. 7.62
## 5      155. 7.74
## 6      154. 7.64

```

In einer Regression ohne Prädiktoren entspricht der Achsenabschnitt dem Mittelwert der AV, daher gibt uns die Spalte (`Intercept`) Aufschluss über unsere Schätzwerte zu μ (der Körpergröße).

Beispiel 8.1m Mit welcher Wahrscheinlichkeit ist $\mu > 155$?).

```
c("mu", "sigma") # den Namen "(Intercept)" durch "mu" ersetzen, ist prägnanter
```

```

post_m41 %>%
  count(mu > 155) %>%
  mutate(prop = n/sum(n))
## # A tibble: 2 x 3
##   `mu > 155`     n   prop
##   <lgl>     <int> <dbl>
## 1 FALSE        3338 0.834
## 2 TRUE         662  0.166

```

Die Wahrscheinlichkeit ist nicht hoch, aber nicht auszuschließen, dass die Kung im Schnitt größer als 155 cm sind. Wahrscheinlicher ist jedoch, dass sie kleiner als 155 cm sind.

Beispiel 8.2m Mit welcher Wahrscheinlichkeit ist $\mu > 165$?).

```
c("mu", "sigma") # den Namen "(Intercept)" durch "mu" ersetzen, ist prägnanter
```

```

post_m41 %>%
  count(mu > 165) %>%
  mutate(prop = n/sum(n))
## # A tibble: 1 x 3
##   `mu > 165`     n   prop
##   <lgl>     <int> <dbl>
## 1 FALSE        4000    1

```

Oh, diese Hypothese können wir mit an Sicherheit grenzender Wahrscheinlichkeit ausschließen. Aber Achtung: Das war eine Kleine-Welt-Aussage! Die Wahrscheinlichkeit, die Hypothese $\mu > 165$ auszuschließen ist *nur* dann hoch, wenn das Modell gilt! Wenn also der Golem keinen Mist gebaut hat. Und sind wir mal ehrlich, der Golem tut, was sein:e Herr:in und Meister:in ihm befiehlt. Letztlich liegt es an uns, den Golem auf Spur zu kriegen.

Beispiel 8.3(Welche mittlere Körpergröße wird mit 95% Wahrscheinlichkeit nicht überschritten, laut d

```
summarise(q95 = quantile(mu, .95))  
## # A tibble: 1 x 1  
##      q95  
##      <dbl>  
## 1 155.
```

Beispiel 8.4(In welchem 90%-PI liegt μ vermutlich?).

```
eti()  
## Equal-Tailed Interval  
##  
## Parameter / 95% ETI  
## -----  
## mu / [153.82, 155.41]  
## sigma / [ 7.22, 8.35]
```

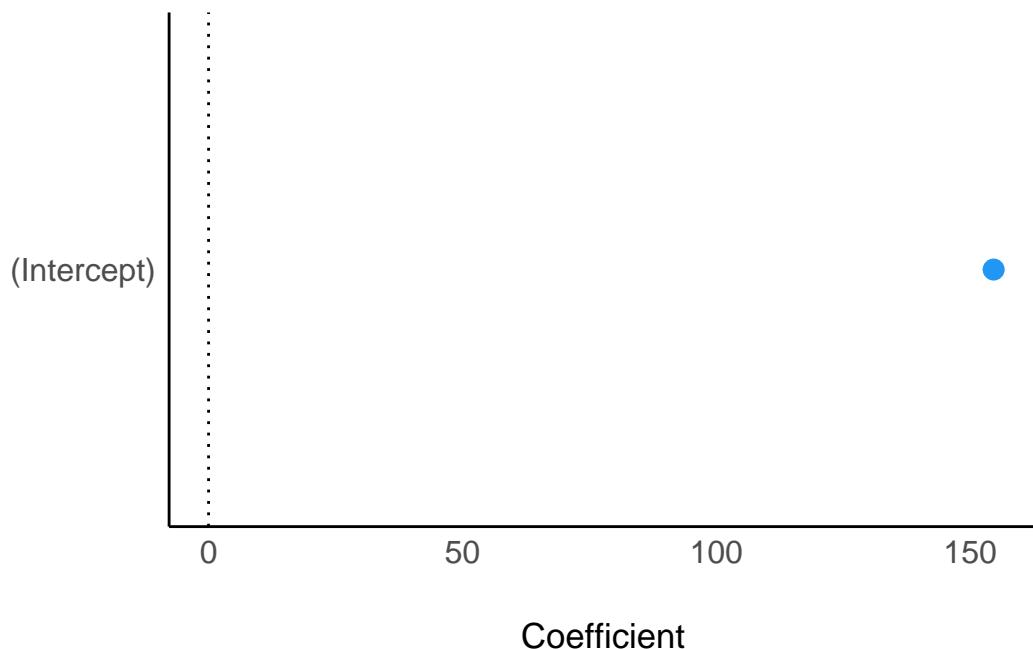
Ein ETI ist synonym zu PI.

Beispiel 8.5 (Mit welcher Unsicherheit ist die Schätzung der mittleren Körpergröße behaftet?).

```
parameters()  
## Parameter / Median / 95% CI / pd / Rhat / ESS /  
## -----  
## (Intercept) / 154.60 / [153.82, 155.41] / 100% / 1.001 / 2702.00 / Normal (154.60 +-
```

Seeing is believing:

```
m41 %>%  
  parameters() %>%  
  plot(show_intercept = TRUE)
```



Das Modell ist sich recht sicher: die Ungewissheit der mittleren Körpergröße liegt bei nicht viel mehr als einem Zentimeter (95%-CI).

Beispiel 8.6 (Was ist der mediane Schätzwert der mittleren Körpergröße, sozusagen der “Best Guess”?). `parameters(m41)` hat uns die Antwort schon gegeben: Ca. 155 cm.

Ähnliche Fragen bleiben als Übung für die Lesis .

8.7.3 Standard-Prioriwerte bei `stan_glm()`

`stan_glm()` nimmt für uns Priori-Wert an. Welche das sind, kann man sich so anzeigen lassen:

```
prior_summary(m41)
## Priors for model 'm41'
## -----
## Intercept (after predictors centered)
##   Specified prior:
##     ~ normal(location = 155, scale = 2.5)
##   Adjusted prior:
##     ~ normal(location = 155, scale = 19)
##
## Auxiliary (sigma)
```

```

##   Specified prior:
##     ~ exponential(rate = 1)
##   Adjusted prior:
##     ~ exponential(rate = 0.13)
## -----
## See help('prior_summary.stanreg') for more details

```

`stan_glm()` verwendet (in der Voreinstellung) *schwach informative* Priori-Werte, die nur wenig Vorabwissen in das Modell geben. Es werden dafür die Stichproben-Daten als Priori-Daten verwendet: Mittelwerte und Streuungen der AV werden als Grundlage für die Priori-Verteilungen herangezogen. Strenggenommen ist das nicht “pures Bayes”, weil die Priori-Werte ja *vorab*, also vor Kenntnis der Daten bestimmt werden sollen. Bitte reichen Sie Ihre Beschwerden bei Andrew Gelman ein.

Man sollte diese Standardwerte als Minimalvorschlag sehen. Kennt man sich im Sachgebiet aus, kann man meist bessere Prioris finden. Die Voreinstellung ist nicht zwingend; andere Werte wären auch denkbar.

- **Intercept:** μ , der Mittelwert der Verteilung Y
 - $\mu \sim \mathcal{N}(\bar{Y}, sd(Y) \cdot 2.5)$
 - als Streuung von μ wird die 2.5-fache Streuung der Stichprobe (für Y) angenommen.
- **Auxiliary (sigma):** σ , die Streuung der Verteilung Y
 - $\sigma \sim \mathcal{E}(\lambda = 1/sd(Y))$
 - als “Streuung”, d.h. λ von h_i wird $\frac{1}{sd(Y)}$ angenommen.

Eine sinnvolle Strategie ist, einen Prior so zu wählen, dass man nicht übergewiss ist, also nicht zu sicher Dinge behauptet, die dann vielleicht doch passieren (also die Ungewissheit zu gering spezifiziert), andererseits sollte man extreme, unplausible Werte ausschließen.

! Wichtig

Bei der Wahl der Prioris gibt es nicht die eine, richtige Wahl. Die beste Entscheidung ist auf transparente Art den Stand der Forschung einfließen zu lassen und eigene Entscheidungen zu begründen. Häufig sind mehrere Entscheidungen möglich. Möchte man lieber vorsichtig sein, weil man wenig über den Gegenstand weiß, dann könnte man z.B. auf die Voreinstellung von `rstanarm` vertrauen, die “schwachinformativ” ist, also nur wenig Priori-Information in das Modell einfließen lässt.

8.8 Modell m42: unsere Priori-Werte

Im Modell m41 haben wir auf die Priori-Werte der Voreinstellung von `rstanarm` vertraut. Jetzt lassen wir mal unsere eigenen Priori-Werte einfließen, in unserem zweiten Kung-Modell, m42.

8.8.1 m42

Dann lassen wir `stan_glm()` unser zweites Modell berechnen. Dieses Mal geben wir die Priori-Werte explizit an.

```
m42 <-  
  stan_glm(height ~ 1,  
            prior_intercept = normal(178, 20), # mu  
            prior_aux = exponential(0.125), # sigma  
            refresh = FALSE, # bitte nicht so viel Ausgabe drucken  
            data = d2)  
parameters(m42)  
## Parameter / Median / 95% CI / pd / Rhat / ESS / Prio  
## -----  
## (Intercept) / 154.60 / [153.80, 155.41] / 100% / 1.001 / 2728.00 / Normal (178 +- 20)
```

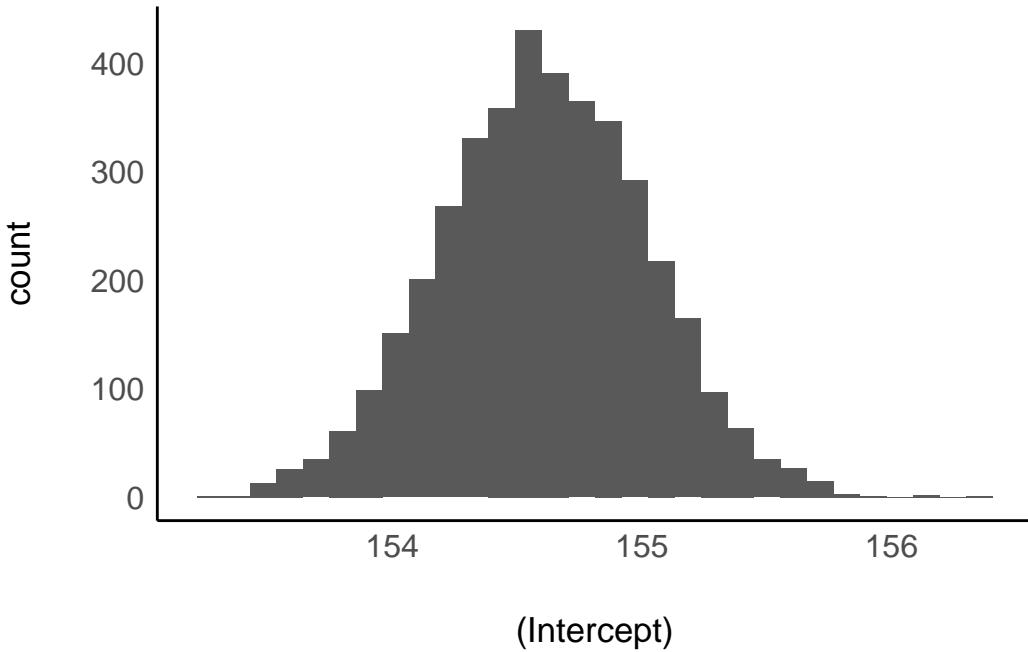
Wir haben noch nicht alle Informationen kennengelernt, die hier ausgegeben werden. Im Zweifel: Einfach ignorieren. Wichtige Fähigkeit im Studium .

! Wichtig

Vergleichen Sie die Parameterwerte von `m41` und `m42`! Was fällt Ihnen auf? Nichts? Gut! Tatsächlich liefern beide Modelle sehr ähnliche Parameterwerte. Die Prioriwerte waren nicht so wichtig, weil wir genug Daten haben. Hat man einigermaßen viele Daten, so fallen Prioriwerte nicht mehr ins Gewicht, zumindest wenn sie moderat gewählt waren.

8.8.2 Posteriori-Verteilung und Parameter plotten

```
m42 %>%  
  as_tibble() %>%  
  ggplot(aes(x = `^`((Intercept)`^`)) +  
  geom_histogram()
```



Ein Vergleich mehrerer Priori-Werte wäre auch nützlich, um ein skeptisches Publikum von der Wahl (bzw. der Indifferenz) der gewählten Priori-Werte zu überzeugen.

8.9 Fazit

Wir haben die Posteriori-Verteilung für ein Gauss-Modell berechnet. Dabei hatten wir ein einfaches Modell mit metrischer Zielvariablen, ohne Prädiktoren, betrachtet. Die Zielvariable, Körpergröße (`height`), haben wir als normalverteilt mit den Parametern μ und σ angenommen. Für μ und σ haben wir jeweils keinen einzelnen (fixen) Wert angenommen, sondern eine Wahrscheinlichkeitsverteilung, der mit der Priori-Verteilung für μ bzw. σ festgelegt ist.



Abbildung 8.10: Bleiben Sie dran!

8.10 Wahl der Priori-Werte

Dieser Abschnitt ist eine VERTIEFUNG und nicht prüfungsrelevant.

8.10.1 Welche Beobachtungen sind auf Basis unseres Modells zu erwarten?

```
n <- 1e4

sim <- tibble(sample_mu =
  rnorm(n,
    mean = 178,
    sd   = 20),
  sample_sigma =
  rexp(n,
    rate = 0.1)) %>%
  mutate(height =
  rnorm(n,
    mean = sample_mu,
    sd   = sample_sigma))

height_sim_sd <-
  sd(sim$height) %>% round()
height_sim_mean <-
  mean(sim$height) %>% round()
```

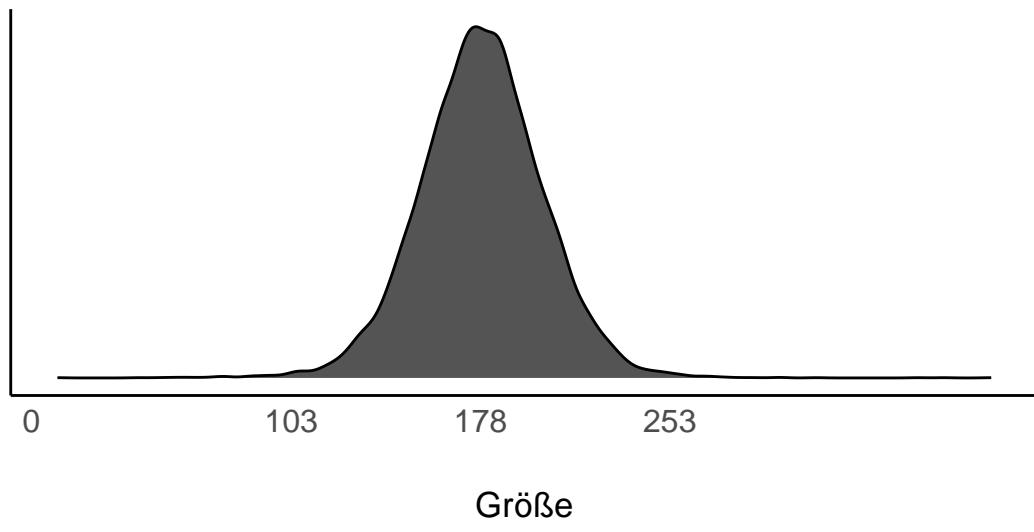
Was denkt der Golem (m41) *apriori* von der Größe der !Kung?

Ziehen wir mal ein paar Stichproben auf Basis des Modells. Voilà:

```
p3 <-
  sim %>%
  ggplot(aes(x = height)) +
  geom_density(fill = "grey33") +
  scale_x_continuous(breaks = c(0, 178-3*height_sim_sd, 178, 178+3*height_sim_sd)) +
  scale_y_continuous(NULL, breaks = NULL) +
  labs(title = "height ~ dnorm(mu, sigma)",
       caption = "X-Achse zeigt MW±3SD",
       x = "Größe") +
  theme(panel.grid = element_blank())
```

p3

`height ~ dnorm(mu, sigma)`



[Quellcode](#)

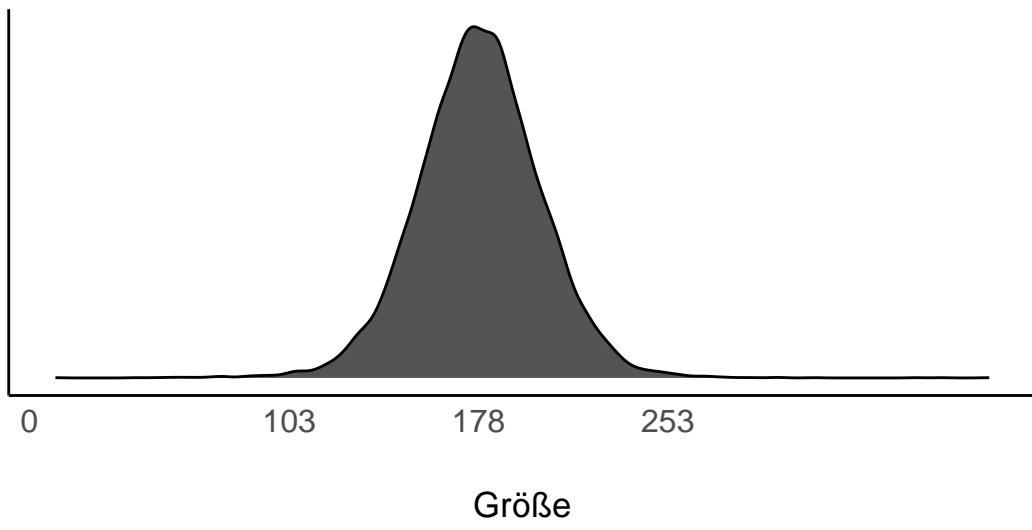
8.10.2 Priori-Werte prüfen mit der Priori-Prädiktiv-Verteilung

- Die Priori-Prädiktiv-Verteilung (`sim`) simuliert Beobachtungen (nur) auf Basis der Priori-Annahmen: $h_i \sim \mathcal{N}(\mu, \sigma)$, $\mu \sim \mathcal{N}(178, 20)$, $\sigma \sim \mathcal{E}(0.1)$
- So können wir prüfen, ob die Priori-Werte vernünftig sind.

Die Priori-Prädiktiv-Verteilung zeigt, dass unsere Priori-Werte ziemlich vage sind, also einen zu breiten Bereich an Größenwerten zulassen:

p3

$\text{height} \sim \text{dnorm}(\mu, \sigma)$



X-Achse zeigt MW±3SD

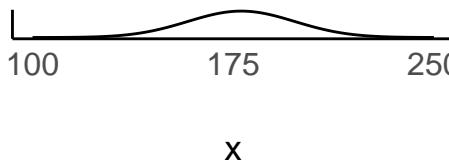
Anteil $h_i > 200$:

```
anteil_großer_kung <-  
  sim %>%  
    count( height > 200 ) %>%  
    mutate(prop = n/sum(n))  
anteil_großer_kung  
## # A tibble: 2 x 3  
##   `height > 200`     n   prop  
##   <lgl>           <int> <dbl>  
## 1 FALSE            8247 0.825  
## 2 TRUE             1753 0.175
```

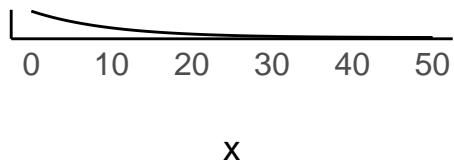
Sehr große Buschleute? 18 Prozent sind größer als 2 Meter. Das ist diskutabel, muss aber nicht zwangsläufig ein schlechter Prior sein.

8.10.3 Vorhersagen der Priori-Werte

$\mu \sim \text{dnorm}(178, 20)$



$\sigma \sim \text{dexp}(0.1)$



$\text{height} \sim \text{dnorm}(\mu, \sigma)$



Größe

X-Achse zeigt MW±3SD

8.10.4 Extrem vage Priori-Verteilung für die Streuung?

$$\sigma \sim \mathcal{E}(\lambda = 0.01)$$

```
# simulate
set.seed(4)

sim2 <-
  tibble(sample_mu      = rnorm(n, mean = 178, sd = 100),
        sample_sigma = rexp(n, rate = .01)) %>%
  mutate(height = rnorm(n, mean = sample_mu, sd = sample_sigma))

# compute the values we'll use to break on our x axis
breaks <-
  c(mean(sim2$height) - 3 * sd(sim2$height), 0, mean(sim2$height), mean(sim2$height) + 3 * sd(sim2$height))

# this is just for aesthetics
text <-
  tibble(height = 272 - 25,
        y       = .0013,
        label   = "größter Mann",
        angle   = 90)
```

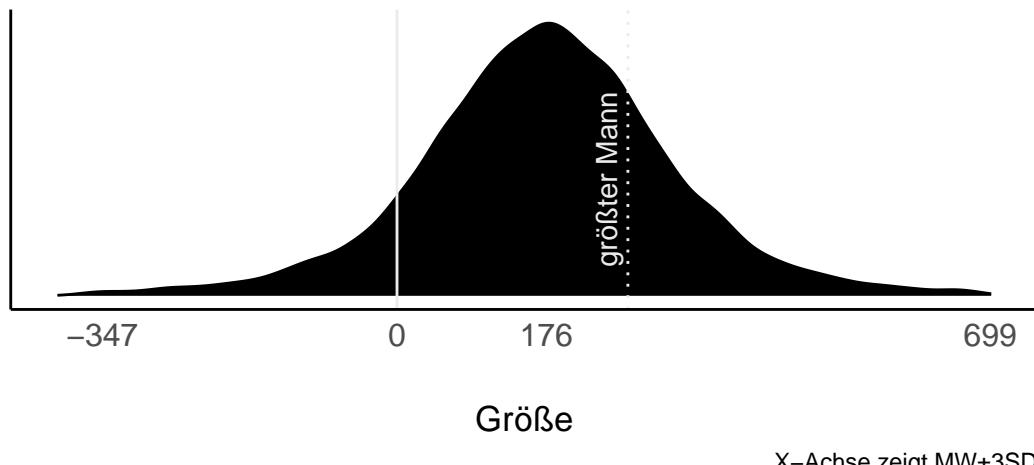
```

# plot
p4 <-
  sim2 %>%
  ggplot(aes(x = height)) +
  geom_density(fill = "black") +
  geom_vline(xintercept = 0, color = "grey92") +
  geom_vline(xintercept = 272, color = "grey92", linetype = 3) +
  geom_text(data = text,
            aes(y = y, label = label, angle = angle),
            color = "grey92") +
  scale_x_continuous(breaks = breaks,
                     limits = c(-400, 700)) +
  scale_y_continuous(NULL, breaks = NULL) +
  labs(title = "height ~ dnorm(mu, sigma)\nmu ~ dnorm(178, 100)\nsigma ~ E(0.01)",
       x = "Größe",
       caption = "X-Achse zeigt MW±3SD") +
  theme(panel.grid = element_blank())

```

p4

$\text{height} \sim \text{dnorm}(\mu, \sigma)$
 $\mu \sim \text{dnorm}(178, 100)$
 $\sigma \sim E(0.01)$



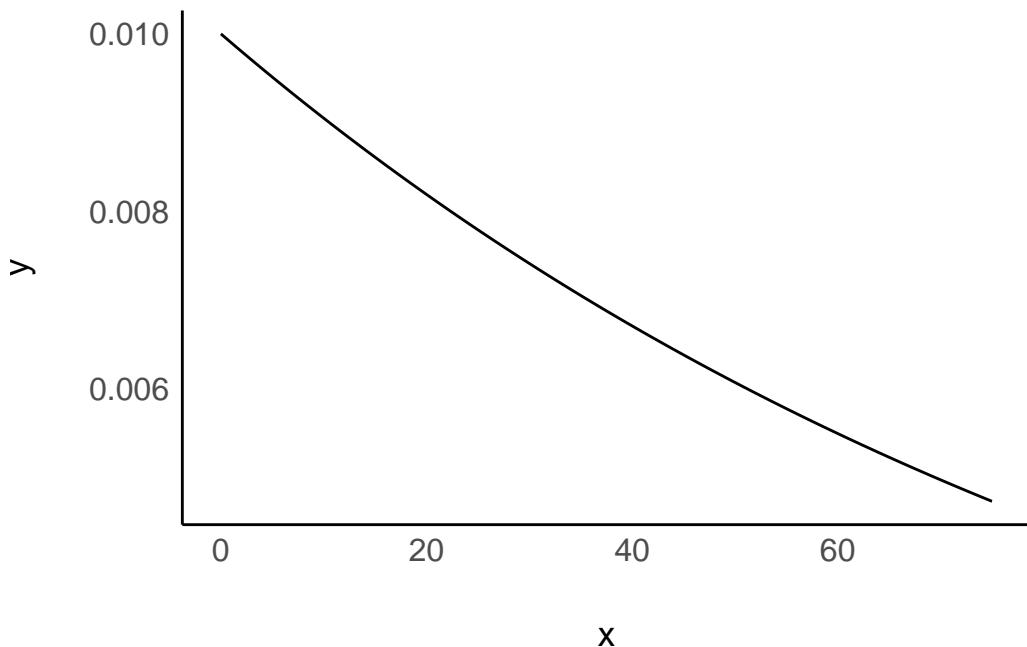
Die Streuung der Größen ist weit:

```

d <-
  tibble(x = seq(0, 75, by = .01),
         y = dexp(x, rate = .01))

```

```
d %>%
  ggplot(aes(x,y)) +
  geom_line()
```



Das Modell geht apriori von ein paar Prozent Menschen mit *negativer* Größe aus. Ein Haufen Riesen werden auch erwartet.

Vage (flache, informationslose, “neutrale”, “objektive”) Priori-Werte machen oft keinen Sinn, weil sie extreme, unplausible Werte zulassen.

8.11 Aufgaben

- stan_glm01
- ReThink4e1
- ReThink4e2
- ReThink4e3
- Kung-height
- Pupil-size
- IQ-Studentis
- Priori-Streuung
- Priorwahl1

9 Lineare Modelle

9.1 Lernsteuerung

9.1.1 Lernziele

Nach Absolvieren des jeweiligen Kapitels sollen folgende Lernziele erreicht sein.

Sie können ...

- die Post-Verteilung für einfache lineare Modelle in R berechnen
- zentrale Informationen zu Modellparametern - wie Lage- oder Streuungsmaße und auch Schätzintervalle - aus der Post-Verteilung herauslesen
- künftige, laut Modell zu erwartende Beobachtungen mit der PPV simulieren

9.1.2 Benötigte R-Pakete

In diesem Kapitel benötigen Sie folgende R-Pakete.

```
library(tidyverse)
library(easystats)
library(rstanarm)
```

Da wir in diesem Kapitel immer mal wieder eine Funktion aus dem R-Paket `{easystats}` verwenden: [Hier](#) finden Sie eine Übersicht aller Funktionen des Pakets.¹

9.1.3 Begleitvideos

- Prädiktoren zentrieren

¹Da es viele Funktionen sind, bietet es sich an mit *Strg-F* auf der Webseite nach Ihrem Lieblingsbefehl zu suchen.

9.2 Post-Verteilung der Regression

9.2.1 Einfache Regression

Die (einfache) Regression prüft, inwieweit zwei Variablen, Y und X linear zusammenhängen. Je mehr sie zusammenhängen, desto besser kann man X nutzen, um Y vorherzusagen (und umgekehrt). Hängen X und Y zusammen, heißt das nicht (unbedingt), dass es einen *kausalen* Zusammenhang zwischen X und Y gibt. *Linear* ist ein Zusammenhang, wenn der Zuwachs in Y relativ zu X konstant ist: wenn X um eine Einheit steigt, steigt Y immer um b Einheiten (nicht kausal, sondern deskriptiv gemeint).

[Datenquelle](#), McElreath (2020).

Laden wir die !Kung-Daten und visualisieren wir uns den Zusammenhang zwischen Gewicht (X) und Größe (Y), Abbildung 9.1.

```
Kung_path <- "https://raw.githubusercontent.com/sebastiansauer/Lehre/main/data/Howell1a.csv"
d <- read_csv(Kung_path)

d2 <-
  d %>%
  filter(age > 18)

d2 %>%
  ggplot(
    aes(x = weight, y = height)) +
  geom_point(alpha = .7) +
  geom_smooth(method = "lm")
```

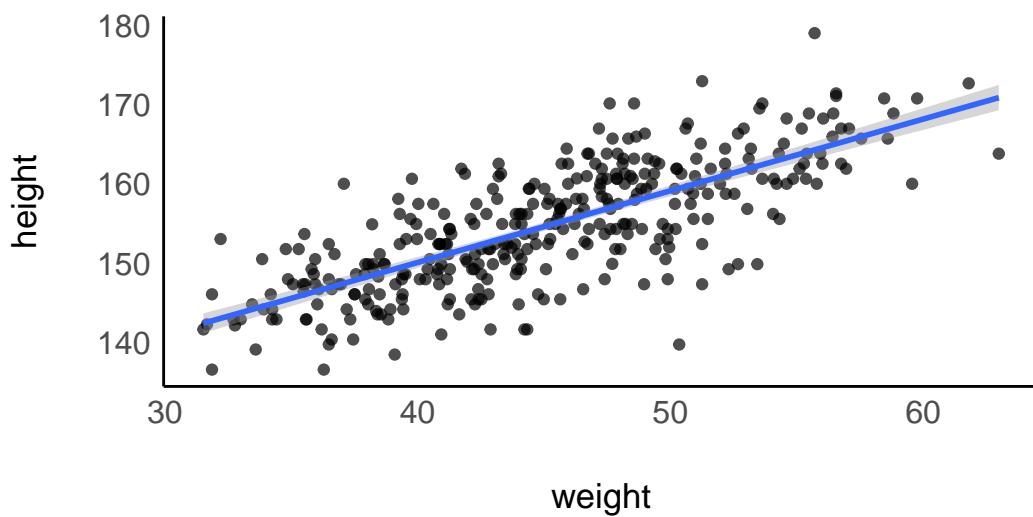


Abbildung 9.1: Der Zusammenhang von Größe und Gewicht im !Kung-Datensatz

9.2.2 Bei jedem Prädiktorwert eine Post-Verteilung für μ

Komfort pur: Unser Modell erlaubt uns für jeden beliebigen Wert des Prädiktors eine Post-Verteilung (von μ) zu berechnen.

Hier am Beispiel von m42, s. Abbildung 9.2.

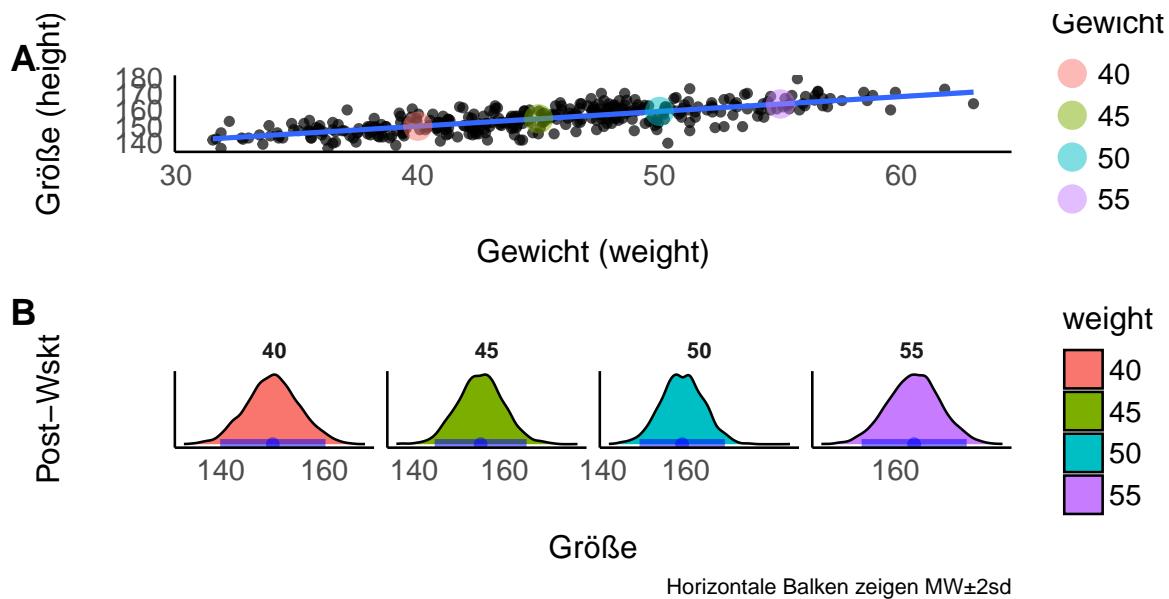


Abbildung 9.2: Für jeden beliebigen Prädiktorwert kann man eine Post-Verteilung bekommen. A: Regressionsmodell mit einigen ausgewählten Gewichtswerten. B: Für jeden beliebigen Gewichtswert bekommt man eine Post-Verteilung

9.2.3 Statistiken zum !Kung-Datensatz

Datenquelle

Hier sind die zentralen deskriptiven Statistiken zum !Kung-Datensatz.

```

Kung_path <- "https://raw.githubusercontent.com/sebastiansauer/Lehre/main/data/Howell1a.csv"
d <- read_csv(Kung_path)

d2 <- d %>% filter(age > 18)

describe_distribution(d2)

## # Variable |      Mean |       SD |      IQR |          Range | Skewness | Kurtosis |      n | n_
## -----
## ## height   |    154.64 |    7.77 |   12.06 | [136.53, 179.07] |     0.14 |    -0.50 |  346 | 346
## ## weight   |     45.05 |    6.46 |    9.14 | [31.52, 62.99] |     0.14 |    -0.53 |  346 | 346

```

```

## age      |    41.54 | 15.81 | 22.00 |    [19.00, 88.00] |    0.68 |    -0.20 | 346 |
## male     |     0.47 |  0.50 |   1.00 |    [0.00, 1.00] |    0.10 |    -2.00 | 346 |
## weight_c | 2.63e-15 |   6.46 |   9.14 | [-13.52, 17.95] |    0.14 |    -0.53 | 346 |

```

Das mittlere Körpergewicht (`weight`) liegt bei ca. 45kg (sd 7 kg).

9.2.4 Etwas mehr EDA

Wir brauchen die EDA hier nicht wirklich, aber es ist praktisch. Das Paket `DataExplorer` hat ein paar nette Hilfen zur explorativen Datenanalyse.

```
library(DataExplorer)
```

9.2.4.1 Gibt es fehlende Werte?

Nein, s. Abb. Abbildung 9.3.

```
d2 %>% plot_missing()
```

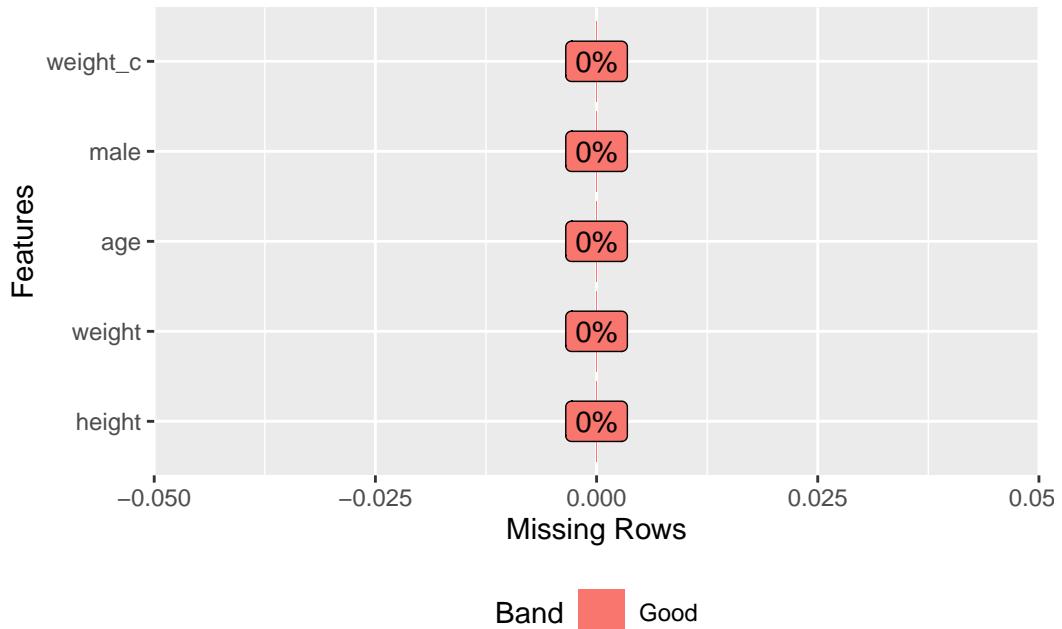


Abbildung 9.3: Fehlende Werte - fehlen.

9.2.4.2 Verteilung der numerischen Variablen

Betrachten wir die Verteilung der *numerischen* Variablen des Datensatzes, s. Abbildung 9.4.

```
d2 %>% plot_histogram()
```

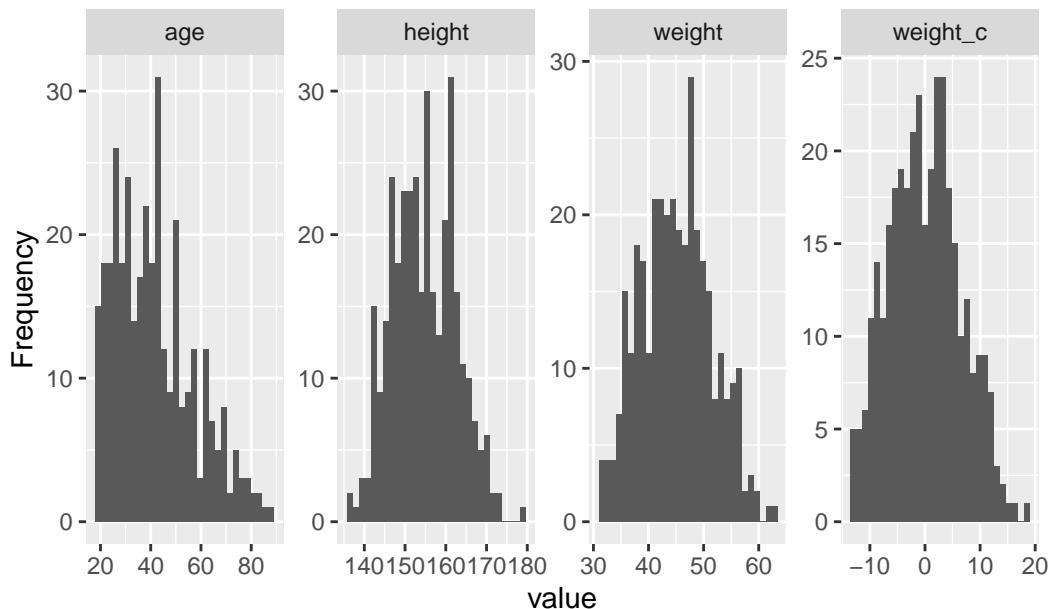


Abbildung 9.4: Verteilung (als Histogramme dargestellt) der numerischen Variablen des Datensatzes

9.2.4.3 Verteilung der kategorialen Variablen

Betrachten wir die Verteilung der *kategorialen* Variablen des Datensatzes, s. Abbildung 9.5.

```
d2 %>% plot_bar()
```

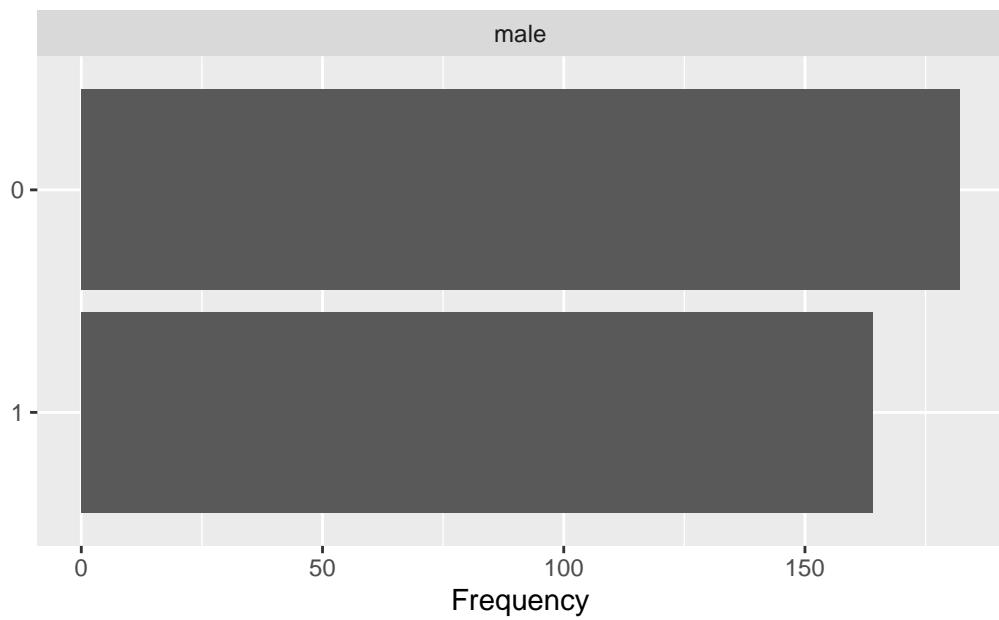


Abbildung 9.5: Verteilung (als Balkendiagramme dargestellt) der kategorialen Variablen des Datensatzes

9.2.4.4 Korrelationen

Die Korrelationen der (numerischen) Variablen sind in Abbildung 9.6 dargestellt.

```
d2 %>% plot_correlation()
```

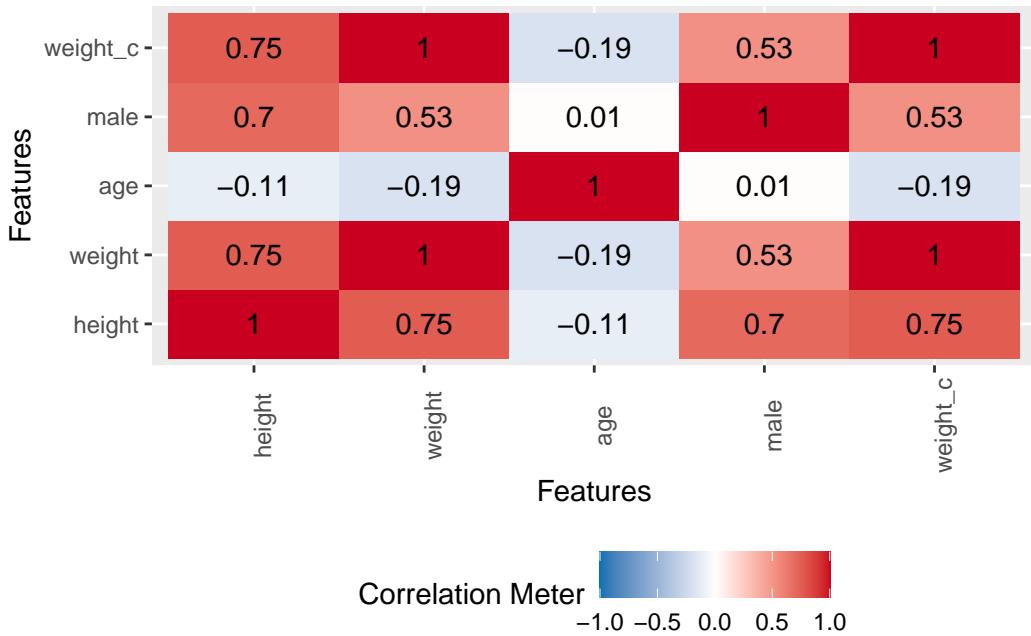


Abbildung 9.6: Korrelationsmatrix

9.2.4.5 Bonus

Probieren Sie mal die folgende Funktion aus, die Ihnen einen Bericht zur EDA erstellt:
`create_report(d2)`.

9.2.5 Prädiktor zentrieren

Zieht man von jedem Gewichtswert den Mittelwert ab, so bekommt man die Abweichung des Gewichts vom Mittelwert (Prädiktor “zentrieren”). Wenn man den Prädiktor (`weight`) zentriert hat, ist der Achsenabschnitt, α , einfacher zu verstehen. In einem Modell mit zentriertem Prädiktor (`weight`) gibt der Achsenabschnitt die Größe einer Person mit durchschnittlichem Gewicht an. Würde man `weight` nicht zentrieren, gibt der Achsenabschnitt die Größe einer Person mit `weight=0` an, was nicht wirklich sinnvoll zu interpretieren ist.

Vgl. Gelman, Hill, und Vehtari (2021), Kap. 10.4, 12.2.

So kann man das Zentrieren bewerkstelligen:

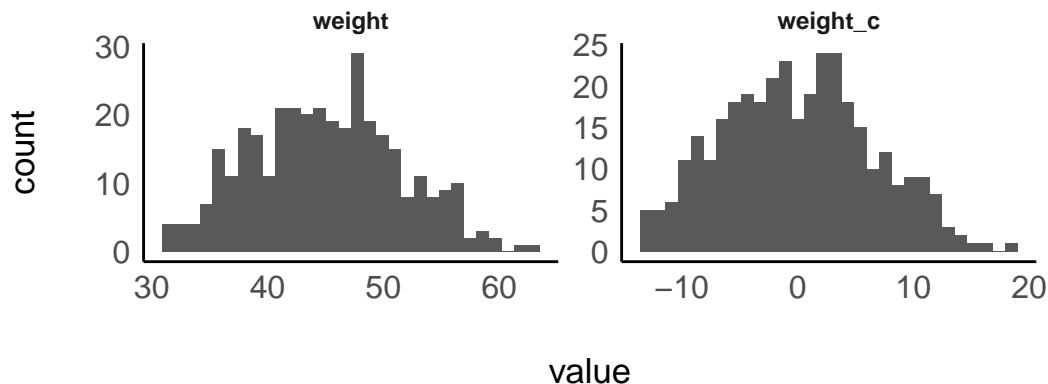
```
d3 <-
  d2 %>%
  center(weight)
```

Oder so, von Hand:

```
d3 <-
d2 %>%
  mutate(weight_c = weight - mean(weight))
```

	height	weight	age	male	weight_c
	152	48	63	1	3
	140	36	63	0	-9
	137	32	65	0	-13

Wie man sieht, ist die Verteilung “zur Seite geschoben”: Der Mittelwert liegt jetzt eben bei 0.



Das schwierigste ist dabei, nicht zu vergessen, dass d3 die Tabelle mit zentriertem Prädiktor ist, nicht d2.

9.3 Modell m43: zentraler Prädiktor

Einige Regressionskoeffizienten, wie der Achsenabschnitt (Intercept) sind schwer zu interpretieren: Bei einem (erwachsenen) Menschen mit *Gewicht 0*, was wäre wohl die Körpergröße? Hm, Philosophie steht heute nicht auf der Tagesordnung.

Da wäre es schön, wenn wir die Daten so umformen könnten, dass der Achsenabschnitt eine sinnvolle Aussage macht. Zum Glück geht das leicht: Wir zentrieren den Prädiktor (Gewicht)!

! Wichtig

Durch Zentrieren kann man die Ergebnisse einer Regression einfacher interpretieren.

9.3.1 Modelldefinition von m43

Für jede Ausprägung des Prädiktors (`weight_centered`), wc_i , wird eine Post-Verteilung für die abhängige Variable (`height`, h_i) berechnet. Der Mittelwert μ für jede Post-Verteilung ergibt sich aus dem linearen Modell (unserer Regressionsformel). Die Post-Verteilung berechnet sich auf Basis der Priori-Werte und des Likelihood (Bayes-Formel). Wir brauchen Priori-Werte für die Steigung β und den Achsenabschnitt α der Regressionsgeraden. Außerdem brauchen wir einen Priori-Wert, der die Streuung σ der Größe (`height`) angibt; dieser Wert wird als exponentialverteilt angenommen. Der Likelihood gibt an, wie wahrscheinlich ein Wert `height` ist, gegeben μ und σ .

$height_i \sim \text{Normal}(\mu_i, \sigma)$	Likelihood
$\mu_i = \alpha + \beta \cdot \text{weightcentered}_i$	Lineares Modell
$\alpha \sim \text{Normal}(178, 20)$	Priori
$\beta \sim \text{Normal}(0, 10)$	Priori
$\sigma \sim \text{Exp}(0.1)$	Priori

i Hinweis

Der Achsenabschnitt (engl. *intercept*) eines Regressionsmodells wird in der Literatur oft mit β_0 bezeichnet, aber manchmal auch mit α . Und manchmal mit noch anderen Buchstaben, das Alphabet ist weit.

9.3.2 Likelihood, m43

$$height_i \sim \text{Normal}(\mu_i, \sigma) \quad \text{Likelihood}$$

Der Likelihood von m43 ist ähnlich zu den vorherigen Modellen (m41, m42). Nur gibt es jetzt ein kleines “Index-i” am μ und am h (h wie `heights`). Es gibt jetzt nicht mehr nur einen Mittelwert μ , sondern für jede Beobachtung (Zeile) einen Mittelwert μ_i . Lies etwa so:

“Die Wahrscheinlichkeit, eine bestimmte Größe bei Person i zu beobachten, gegeben μ und σ ist normalverteilt (mit Mittelwert μ und Streuung σ)”.

9.3.3 Regressionsformel, m43

$$\mu_i = \alpha + \beta \cdot \text{weight}_i \quad \text{Lineares Modell}$$

μ ist jetzt nicht mehr ein Parameter, der (stochastisch) geschätzt werden muss. μ wird jetzt (deterministisch) berechnet. Gegeben α und β ist μ ohne Ungewissheit bekannt. weight_i ist der Prädiktorwert (`weight`) der i ten Beobachtung, also einer !Kung-Person (Zeile i im Datensatz). Lies etwa so:

“Der Mittelwert μ_i der i ten Person berechnet sich als Summe von α und β mal weight_i ”.

μ_i ist eine lineare Funktion von `weight`. β gibt den Unterschied in `height` zweier Beobachtung an, die sich um eine Einheit in `weight` unterscheiden (Steigung der Regressionsgeraden). α gibt an, wie groß μ ist, wenn `weight` Null ist (Achsenabschnitt, engl. intercept).

9.3.4 Priori-Werte des Modells m43

$\alpha \sim \text{Normal}(178, 20)$	Priori Achsenabschnitt
$\beta \sim \text{Normal}(0, 10)$	Priori Regressionsgewicht
$\sigma \sim \text{Exp}(0.1)$	Priori Sigma

- Parameter sind hypothetische Kreaturen: Man kann sie nicht beobachten, sie existieren nicht wirklich. Ihre Verteilungen nennt man Priori-Verteilungen.
- α wurde in m41 als μ bezeichnet, da wir dort eine “Regression ohne Prädiktoren” berechnet haben.
- σ ist uns schon als Parameter bekannt und behält seine Bedeutung aus dem letzten Kapitel.
- Da `height` nicht zentriert ist, der Mittelwert von α bei 178 und nicht 0.
- β fasst unser Vorwissen, ob und wie sehr der Zusammenhang zwischen Gewicht und Größe positiv (gleichsinnig) ist.

9.4 Vertiefung: Priori-Prädiktiv-Verteilung

VERTIEFUNG, nicht prüfungsrelevant

9.4.1 Moment

Moment. Dieser Prior, β in m43 erachtet positive und negative Zusammenhang als gleich wahrscheinlich!?

Sind wir wirklich indifferent, ob der Zusammenhang von Gewicht und Größe positiv oder negativ ist? Nein, sind wir nicht.

9.4.2 Priori-Prädiktiv-Verteilung für m43

Was denkt wir bzw. unser Golem *apriori* über den Zusammenhang von Größe und Gewicht? Um diese Frage zu beantworten ziehen wir Stichproben aus den Priori-Verteilungen des Modells, also für α , β und σ .

```

m43_prior_pred <-
  stan_glm(height ~ weight_c,
            prior = normal(0, 10),
            prior_intercept = normal(178, 20), # mu
            prior_aux = exponential(0.1), # sigma
            refresh = FALSE,
            prior_PD = TRUE, # Schalter für Prior-Pred-Verteilung
            data = d2)

m43_prior_pred_draws <-
  m43_prior_pred %>%
  as_tibble() %>%
  rename(a = `Intercept`),
  b = weight_c) %>%
  slice_sample(n = 50)

```

a	b	sigma
199.3	5.2	6.0
191.6	0.2	4.6
169.3	-24.9	0.7
205.4	2.2	7.8
174.8	7.9	9.4

Jede Zeile definiert eine Regressionsgerade.

9.4.3 Prior-Prädiktiv-Simulation für m43 mit stan_glm()

```

m43_prior_pred <-
  stan_glm(height ~ weight_c,
            prior = normal(0, 10), # beta
            prior_intercept = normal(178, 20), # alpha
            prior_aux = exponential(0.1), # sigma
            refresh = FALSE,
            prior_PD = TRUE, # DIESER Schalter macht's
            data = d2)

m43_prior_pred_draws <-
  m43_prior_pred %>%
  as_tibble() %>%
  rename(a = `Intercept`),
  b = weight_c) %>%

```

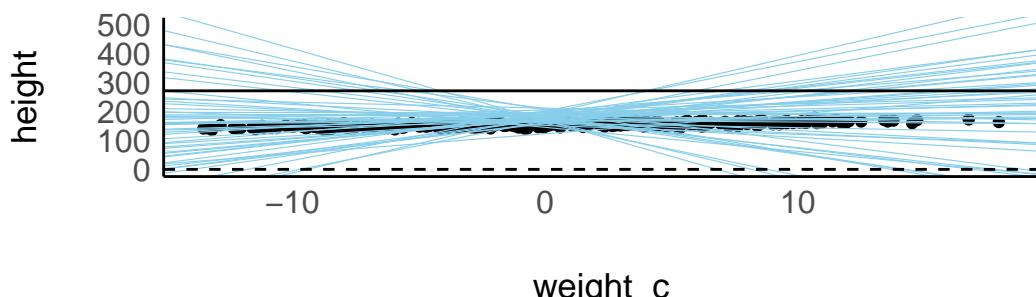
```
slice_sample(n = 50)
```

9.4.4 Visualisieren der Prior-Prädiktiv-Verteilung

```
d2 %>% ggplot() +
  geom_point(aes(x = weight_c, y = height)) +
  geom_abline(data = m43_prior_pred_draws,
  aes(intercept = a, slope = b), color = "skyblue", size = 0.2) +
  scale_y_continuous(limits = c(0, 500)) +
  geom_hline(yintercept = 272, size = .5) +
  geom_hline(yintercept = 0, linetype = "dashed")
```

Einige dieser Regressionsgeraden sind unsinnig!

```
d2 %>% ggplot() +
  geom_point(aes(x = weight_c, y = height)) +
  geom_abline(data = m43_prior_pred_draws,
  aes(intercept = a, slope = b), color = "skyblue", size = 0.2) +
  scale_y_continuous(limits = c(0, 500)) +
  geom_hline(yintercept = 272, size = .5) +
  geom_hline(yintercept = 0, linetype = "dashed")
```



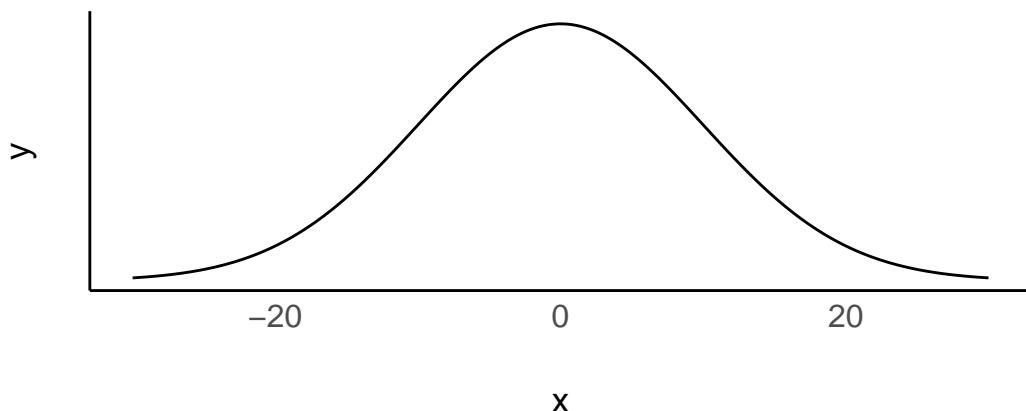
Die durchgezogene horizontale Linie gibt die Größe des [größten Menschen, Robert Pershing Wadlow](#), an.

9.4.5 Ein positiver Wert für β ist plausibler

9.4.5.1 Oh no

Eine Normalverteilung mit viel Streuung:

$\mu=0, s=10$

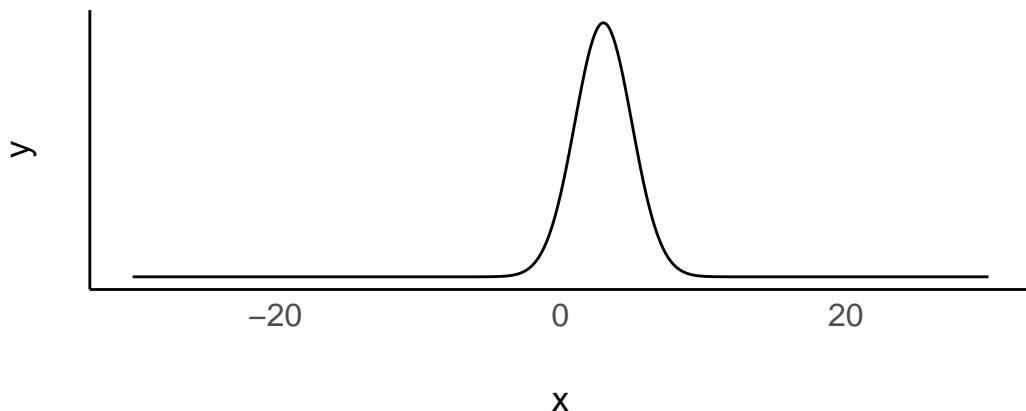


$\beta = -20$ wäre mit diesem Prior gut möglich: Pro kg Gewicht sind Menschen im Schnitt 20cm kleiner, laut dem Modell. Quatsch.

9.4.5.2 Oh yes

Wir bräuchten eher so eine Verteilung, mit mehr Masse auf der positiven Seite ($x>0$):

$\mu=5, sd = 3$



Vermutlich besser: Ein Großteil der Wahrscheinlichkeitsmasse ist $X > 0$. Allerdings gibt's keine Gewähr, dass unser Prior "richtig" ist.

9.4.6 Prior-Prädiktiv-Simulation, 2. Versuch

```
m43a_prior_pred <-
  stan_glm(
    height ~ weight_c,
    prior = normal(2, 2), # Regressionsgewicht
    prior_intercept = normal(178, 20), # mu
    prior_aux = exponential(0.1), # sigma
    refresh = FALSE,
    # Schalter für Prior-Pred-Verteilung:
    prior_PD = TRUE,
    data = d2)

m43a_prior_pred_draws <-
  m43a_prior_pred %>%
  as_tibble() %>%
  # Spaltennamen kürzen:
  rename(a = `Intercept`) %>%
  rename(b = weight_c,
         s = sigma)
```

	a	b	s
188.2	0.8	14.2	
145.9	2.2	15.1	
150.4	2.5	15.0	
202.1	1.5	5.8	
215.5	2.5	6.3	

Das Argument `prior_PD = TRUE` sorgt dafür, dass keine Posteriori-Verteilung, sondern eine Prior-Prädiktiv-Verteilung berechnet wird.

9.4.7 Visualisieren der Prior-Prädiktiv-Verteilung, m43a

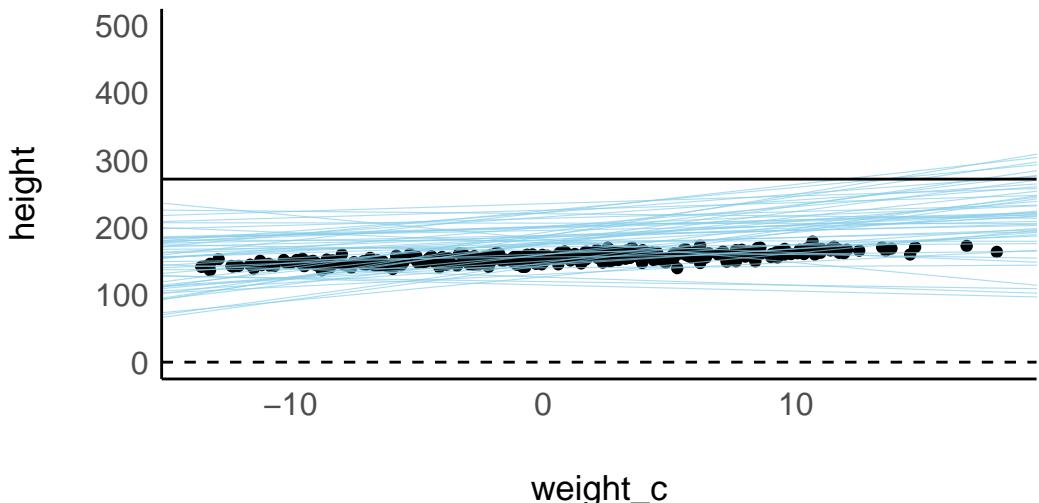
Unsere Priori-Werte scheinen einigermaßen vernünftige Vorhersagen zu tätigen. Allerdings erwartet unser Golem einige Riesen.

```
d2 %>%
  ggplot(aes(x = weight_c, y = height)) +
  geom_point() +
  geom_abline(data = {m43a_prior_pred_draws %>% slice_head(n=50)},
              aes(slope = b,
                  intercept = a),
```

```

color = "skyblue",
size = .2,
alpha = .7) +
geom_hline(yintercept = 272, size = .5) +
geom_hline(yintercept = 0, linetype = "dashed")+
scale_y_continuous(limits = c(0, 500))

```



Die durchgezogene horizontale Linie gibt die Größe des [größten Menschen, Robert Pershing Wadlow](#), an.

9.4.8 Moment, kann hier jeder machen, was er will?

Es doch den einen, richtigen, objektiven Priori-Wert geben?!

Kann denn jeder hier machen, was er will?! Wo kommen wir da hin?!

This is a mistake. There is no more a uniquely correct prior than there is a uniquely correct likelihood. Statistical models are machines for inference. Many machines will work, but some work better than others. Priors can be wrong, but only in the same sense that a kind of hammer can be wrong for building a table.

McElreath (2020), p. 96.

9.4.9 Hier ist unser Modell, m43a

$$\text{height}_i \sim \text{Normal}(\mu_i, \sigma) \quad (9.1)$$

$$\mu_i = \alpha + \beta \cdot \text{weight}_i \quad (9.2)$$

$$\alpha \sim \text{Normal}(178, 20) \quad (9.3)$$

$$\beta \sim \text{Normal}(5, 3) \quad (9.4)$$

$$\sigma \sim \text{Exp}(0.1) \quad (9.5)$$

```
# Zufallszahlen festlegen:
set.seed(42)
# Posteriori-Vert. berechnen:
m43a <-
stan_glm(
  height ~ weight_c, # Regressionsformel
  prior = normal(5, 3), # Regressionsgewicht (beta 1)
  prior_intercept = normal(178, 20), # mu
  prior_aux = exponential(0.1), # sigma
  refresh = 0, # zeig mir keine Details
  data = d2)
```

9.4.10 Eine Zusammenfassung der Posteriori-Verteilung für m43a

```
m43a %>%
parameters()
```

Tabelle 9.4: Fixed effects

Parameter	Median	95% CI	pd	Rhat	ESS	Prior
(Intercept)	154.65	(154.09, 155.21)	100%	0.999	4132.00	Normal (178 +- 20)
weight_c	0.91	(0.83, 0.99)	100%	1.000	3837.00	Normal (5 +- 3)

Unser Modell m43a schätzt die typische Körpergröße einer !Kung-Person *mittleren Gewichts* (`weight_c = 0`) auf knapp 155 cm, und ist sich dieses Werts ziemlich sicher. Pro Kilogramm kommt (laut unserem Modell) ein knapper Zentimeter hinzu, typischerweise; auch hier ist sich das Modell ziemlich sicher, da dass zugehörige 95%-CI keine 20 Zentimeter umfasst.

9.5 Die Post-Verteilung befragen

9.5.1 m43a

Sagen wir, auf Basis gut geprüfter Evidenz haben wir folgendes Modell festgelegt: `height ~ weight_c`, s. Gleichung 9.6.

Prioris:

$$\beta_1 \sim N(5, 3); \beta_0 \sim N(178, 20); \sigma \sim E(0.1) \quad (9.6)$$

Wir nennen das Modell m43a², s. Listing 9.1.

Listing 9.1 Modelldefinition von m43a in R

```
m43a <-  
stan_glm(  
  height ~ weight_c, # Regressionsformel  
  prior = normal(5, 3), # Regressionsgewicht (beta 1)  
  prior_intercept = normal(178, 20), # mu  
  prior_aux = exponential(0.1), # sigma  
  refresh = 0, # zeig mir keine Details  
  seed = 42, # lege die Zufallszahlen fest für Reproduzierbarkeit  
  data = d3)
```

i Hinweis

Mit `seed` kann man die Zufallszahlen fixieren, so dass jedes Mal die gleichen Werte resultieren. So ist die Nachprüfbarkeit der Ergebnisse (“Reproduzierbarkeit”) sichergestellt^a. Welche Wert für `seed` man verwendet, ist egal, solange alle den gleichen verwenden. Der Autor verwendet z.B. oft den Wert 42. Zur Erinnerung: Der Golem zieht Zufallszahlen, damit erstellt er Stichproben, die die Postverteilung schätzen.

^aoder zumindest besser sichergestellt

9.5.2 Mittelwerte von α und β aus der Post-Verteilung

Die ersten paar Zeilen:

id	(Intercept)	weight_c	sigma
1	155.1	0.9	5.0

²Wer ist hier für die Namensgebung zuständig? Besoffen oder was?

2	155.5	0.8	5.1
3	155.5	0.9	5.1

Hier sind die Zusammenfassungen der Stichproben aus der Post-Verteilung, komfortabel zu erhalten mit dem Befehle `parameters`:

```
parameters(m43a)
```

Tabelle 9.6: Fixed effects

Parameter	Median	95% CI	pd	Rhat	ESS	Prior
(Intercept)	154.65	(154.14, 155.19)	100%	0.999	3214.00	Normal (178 +- 20)
weight_c	0.91	(0.82, 0.99)	100%	1.001	4134.00	Normal (5 +- 3)

Die Kennzahl `pd` (propability of direction) gibt die Wahrscheinlichkeit an, dass der Effekt positiv (also größer als Null) oder negativ ist (jenachdem ob der Median des Effekts positiv oder negativ ist). `pd` gibt aber nicht an, wie stark der Effekt ist, nur ob er klar auf einer Seite der Null liegt. Damit ist er so etwas (grob!) Ähnliches wie der p-Wert in der Frequentistischen Statistik (Makowski u. a. 2019).

Am besten das Diagramm dazu anschauen, s Abbildung 9.7.

```
plot(p_direction(m43a))
```

Probability of Direction

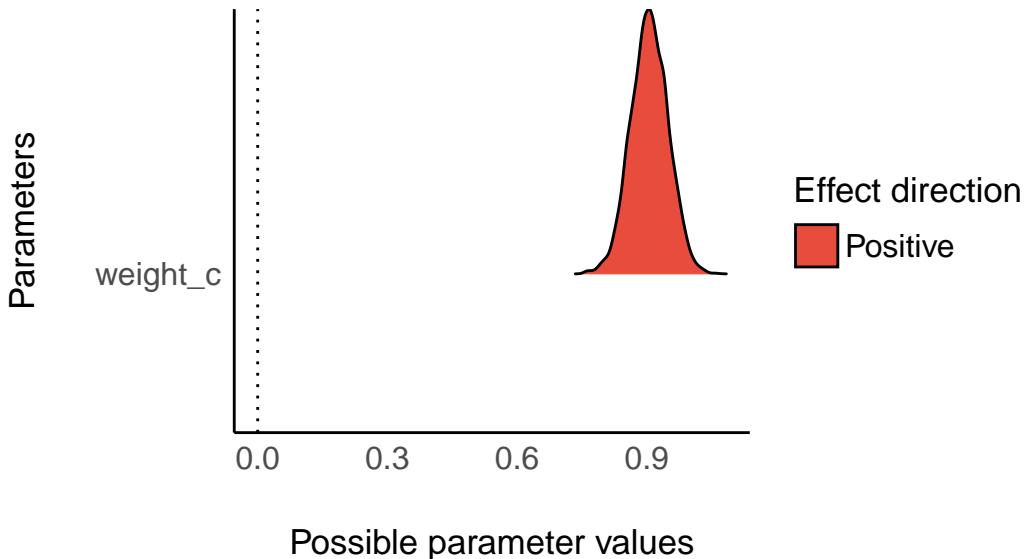


Abbildung 9.7: Diagramm zur Probability of Direction, Modell m43a

Rhat und ESS sind Kennzahlen, die untersuchen, ob mit der Stichprobenziehung im Bayes-Modell alles gut funktioniert hat. Bei einfachen Modellen (die wir hier berechnen) sollte da in der Regel alles in Ordnung sein. Rhat sollte nicht (viel) größer als 1 oder 1,01 sein. ESS (effective sample size) gibt die Anzahl der effektiv nutzbaren Stichproben an (im Standard werden 4000 berechnet). Die Zahl sollte nicht deutlich geringer sein.

Wir werden uns aber mit diesen beiden Kennwerten nicht weiter beschäftigen in diesem Kurs.

9.5.3 Visualisieren der “mittleren” Regressionengeraden

Zur Erinnerung: Die Bayes-Analyse liefert uns viele Stichproben zu den gesuchten Parametern, hier β_0 , β_1 und σ . Überzeugen wir uns mit einem Blick in die Post-Verteilung von m43a:

```
m43a %>%
  as_tibble() %>%
  head()
## # A tibble: 6 x 3
##   `(Intercept)` weight_c sigma
##     <dbl>      <dbl>  <dbl>
## 1       155.    0.858  5.04
## 2       155.    0.835  5.07
## 3       155.    0.855  5.14
```

```

## 4      155.    0.882  5.35
## 5      155.    0.875  5.35
## 6      155.    0.903  5.21

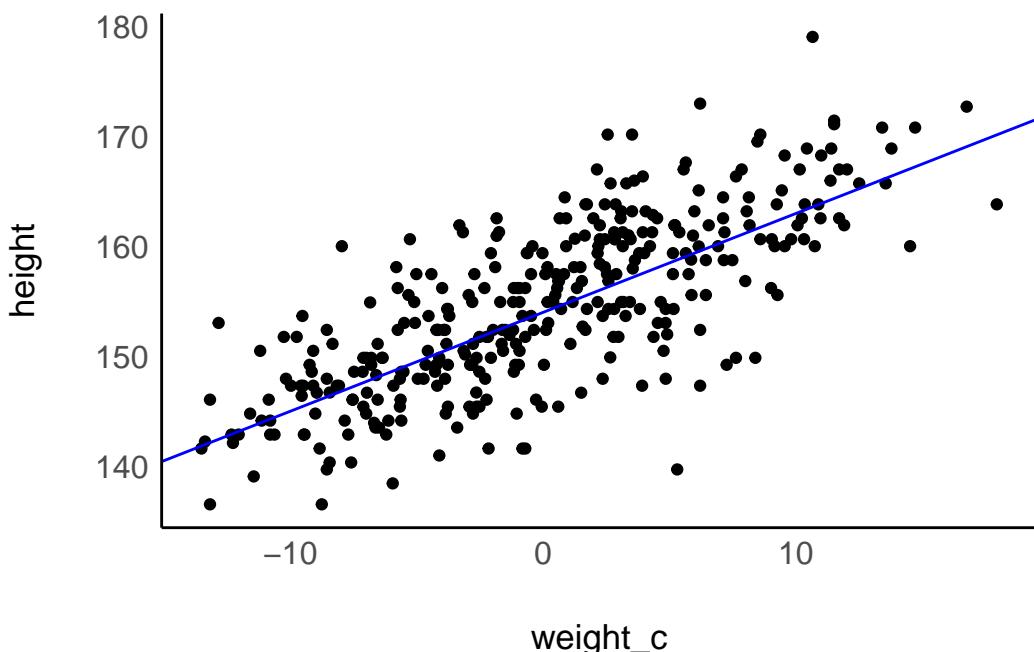
```

Wir können z.B. ein Lagemaß wie den Median hernehmen, um die “mittlere” Regressionsgerade zu betrachten:

```

d2 %>%
  ggplot() +
  aes(x = weight_c, y = height) +
  geom_point() +
  geom_abline(
    slope = 0.9, # Median beta 1
    intercept = 154, # Median beta 0
    color = "blue")

```



Einfacher ist die Syntax vielleicht, wenn man die Funktion `estimate_expectation` benutzt, s. Abbildung 9.8. Mit “expectation” sind hier die erwarteten Werte, also die Regressionsgerade, gemeint.

```

m43_expect <- estimate_expectation(m43a)
plot(m43_expect)

```

Predicted response (height ~ weight_c)

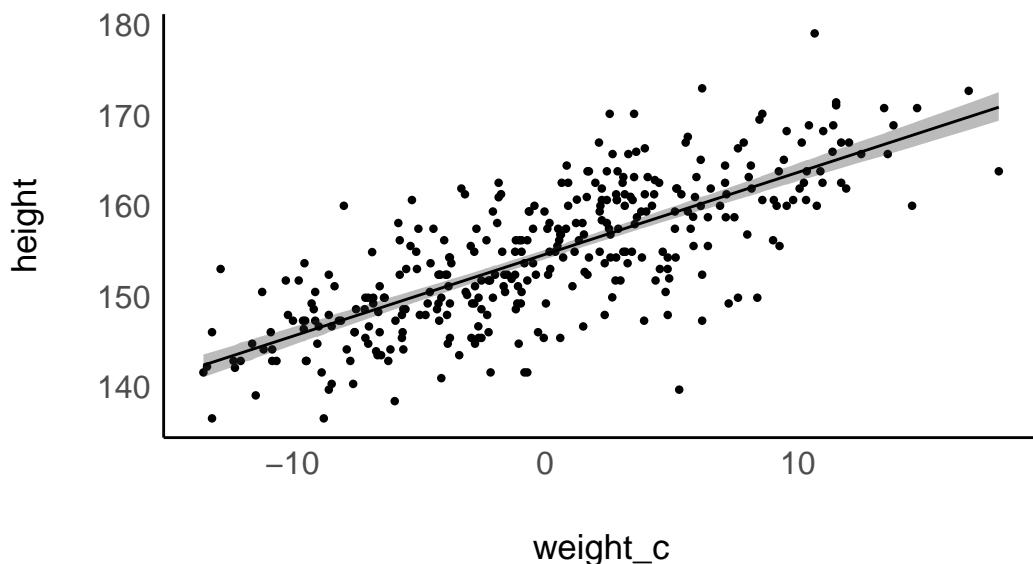


Abbildung 9.8: Erwartete Werte des Modell m43a, sprich, die Regressionsgerade

9.5.4 Zentrale Statistiken zu den Parametern

In diesem Modell gibt es drei Parameter: α, β, σ .

Hier folgen einige Beispiele an Fragen, die wir an unser Modell bzw. die Post-Verteilung stellen können.

9.5.4.1 Lagemaße zu den Parametern

- Was ist die mittlere Größe einer !Kung-Person? (β_0)
- Was ist der Schätzwert für den Zusammenhang von Gewicht und Größe? (β_1)
- Was ist der Schätzwert für Ungewissheit in der Schätzung der Größe? (σ)
- Was ist der wahrscheinlichste Wert für z.B: β_1 ?

Eine nützliche Zusammenfassung der Post-Verteilung bekommt man mit `parameters(modell)`:

```
m43a %>%
  parameters()
```

Tabelle 9.7: Fixed effects

Parameter	Median	95% CI	pd	Rhat	ESS	Prior
(Intercept)	154.65	(154.14, 155.19)	100%	0.999	3214.00	Normal (178 +- 20)
weight_c	0.91	(0.82, 0.99)	100%	1.001	4134.00	Normal (5 +- 3)

Wandelt man das Ausgabe-Objekt der Bayes-Regression, d.h. `m43a`, mit `as_tibble()` in eine Tabelle um, so bekommt man eine Tabelle mit den Stichproben der Post-Verteilung:

```
m43a_post <-
  m43a %>%
  as_tibble()

m43a_post %>%
  head()

## # A tibble: 6 x 3
##   `(Intercept)` weight_c sigma
##       <dbl>     <dbl> <dbl>
## 1      155.    0.858  5.04
## 2      155.    0.835  5.07
## 3      155.    0.855  5.14
## 4      155.    0.882  5.35
## 5      155.    0.875  5.35
## 6      155.    0.903  5.21
```

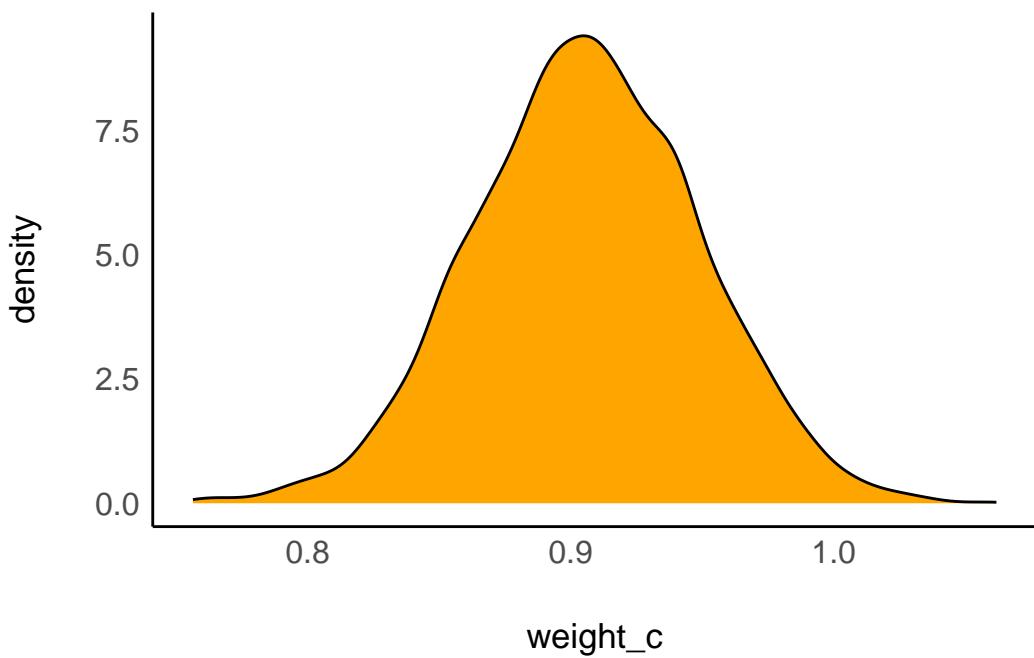
Wie wir gesehen haben, nutzen wir diese Tabelle der Post-Verteilung immer wieder. Speichern wir uns sie also als ein Objekt ab, `m43_post`.

Jetzt haben wir wieder eine schöne Tabelle mit Stichproben aus der Post-Verteilung, die wir wie gewohnt befragen können.

Eine Visualisierung zeigt gut sowohl Lage- als auch Streuungsmaße der Parameter, zumindest grob.,

Oder man erstellt selber ein Diagramm mit `ggplot`.

```
m43a_post %>%
  ggplot(aes(x = weight_c)) +
  geom_density(fill = "orange")
```



Das Diagramm zeigt, dass Mittelwert, Median und Modus eng zusammenliegen. Zur Erinnerung: Der Modus gibt den häufigsten, d.h. hier also den wahrscheinlichsten, Wert an.

Der Modus wird hier auch *Maximum a Posteriori* (MAP) genannt, daher:

```
m43a_post %>%
  summarise(map_b1 = map_estimate(weight_c))
```

Hier ist die Verteilung von σ visualisiert, s. Abbildung 9.9.

```
m43a_post %>%
  ggplot(aes(x = sigma)) +
  geom_density(fill = "orange")
```



Abbildung 9.9: Die Post-Verteilung für den Parameter sigma, m43a

Alternativ kann man sich die Verteilung eines Parameters auch so ausgeben lassen, gleich mit Intervallgrenzen, z.B. 95%, s.@fig-m43a-plot.

```
m43a_hdi <- hdi(m43a_post) # analog mit eti(m43a)  
plot(m43a_hdi)
```

Highest Density Interval (HDI)

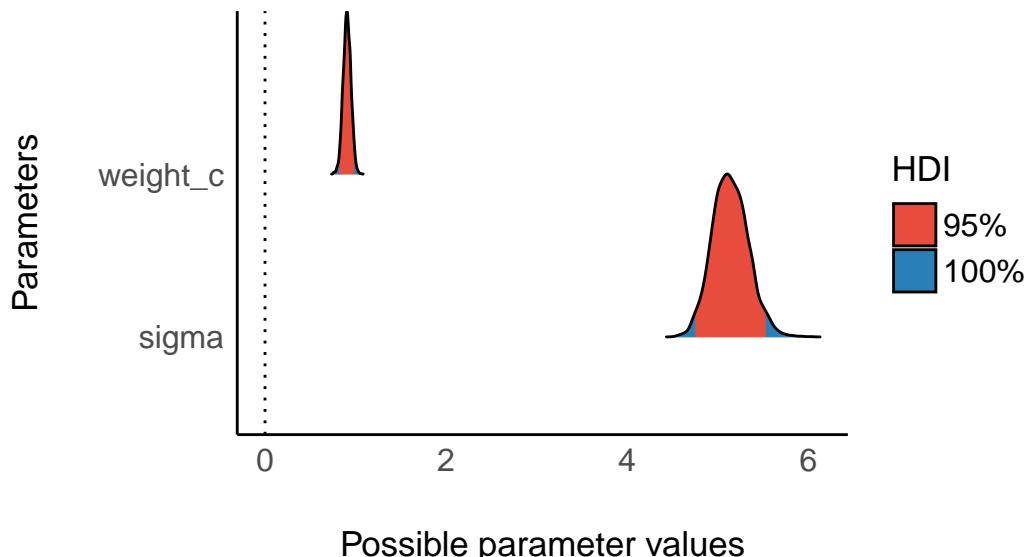


Abbildung 9.10: Die Parameter Gewicht (zentriert) und sigma des Modells m43a

Ergänzt man bei `plot()` noch `show_intercept = TRUE` wird auch der Achsenabschnitt angezeigt.

9.5.5 Streuungsmaße zu den Parametern

- Wie unsicher sind wir uns in den Schätzungen der Parameter?

Diese Frage wird durch die Ungewissheitsintervalle in der Ausgabe beantwortet.

i Hinweis

An einigen Stellen wird empfohlen, anstelle eines (gebräuchlichen) 95%-Intervalls auf ein 90%- oder 89%-Intervall auszuweichen, aufgrund der besseren numerischen Stabilität.

9.5.6 Ungewissheit von α und β aus der Post-Verteilung visualisiert

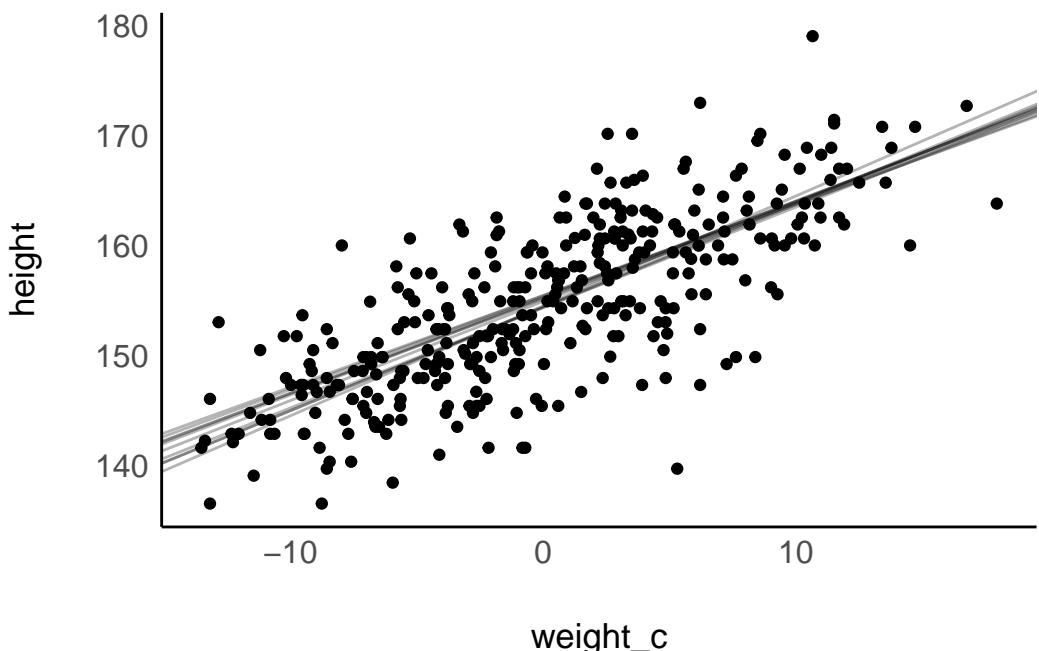
Die ersten 10 Stichproben:

```
d2 %>%
  ggplot(aes(x = weight_c,
             y = height)) +
  geom_point() +
```

```

geom_abline(
  data = m43a_post %>%
    slice_head(n = 10),
  aes(slope = weight_c,
      intercept = `^(Intercept)`),
  alpha = .3)

```

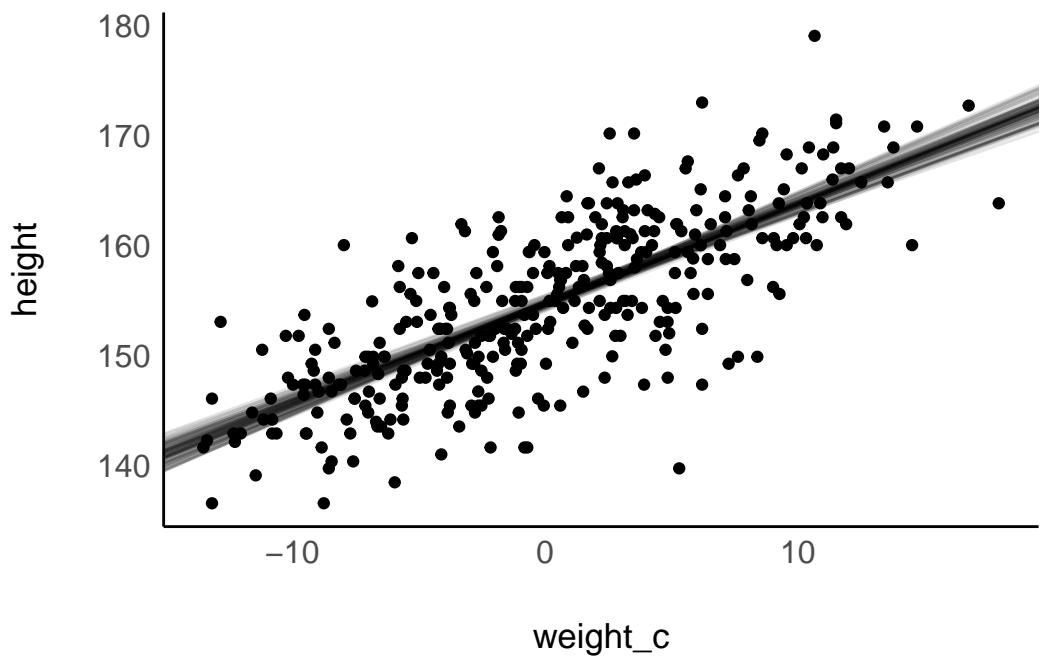


Die ersten 100 Stichproben:

```

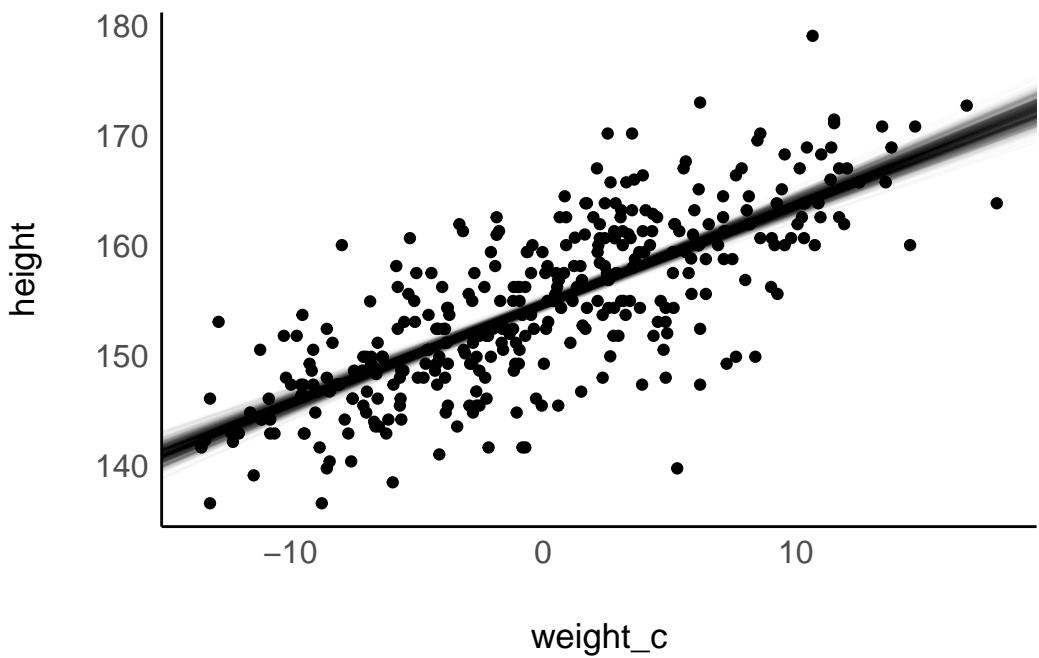
d2 %>%
  ggplot(aes(x = weight_c,
             y = height)) +
  geom_point() +
  geom_abline(
    data = m43a_post %>%
      slice_head(n = 100),
    aes(slope = weight_c,
        intercept = `^(Intercept)`),
    alpha = .1)

```



Die ersten 1e3 Stichproben:

```
d2 %>%
  ggplot(aes(x = weight_c,
             y = height)) +
  geom_point() +
  geom_abline(
    data = m43a_post %>%
      slice_head(n = 1e3),
    aes(slope = weight_c,
        intercept = `^Intercept`^),
    alpha = .01)
```

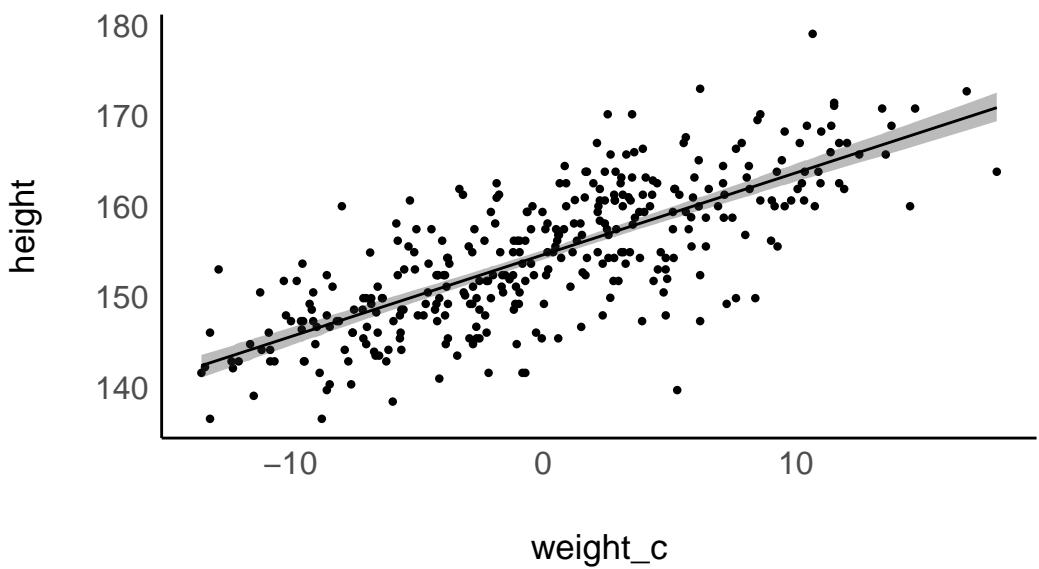


Die ersten 1000000 ... okay, lassen wir es gut sein³.

Einfacher ist die Visualisierung mit `estimate_expectation`:

```
estimate_expectation(m43a) %>% plot()
```

Predicted response (height ~ weight_c)



³Im Standard beschert uns `stan_glm()` 4000 Stichproben.

9.5.7 Fragen zu Quantilen des Achsenabschnitts

i Hinweis

Zur Erinnerung: Bei einem zentrierten Prädiktor misst der Achsenabschnitt die mittlere Größe.

- Welche mittlere Größe wird mit zu 50%, 90% bzw. 95% Wahrscheinlichkeit nicht überschritten?
- Welche mittlere Größe mit zu 95% Wskt. nicht unterschritten?
- Von wo bis wo reicht der innere 50%-Schäzbereich der mittleren Größe?

Quantile:

```
m43a_post %>%
  summarise(
    q_50 = quantile(`(Intercept)` , prob = .5),
    q_90 = quantile(`(Intercept)` , prob = .9),
    q_05 = quantile(`(Intercept)` , prob = .95))
## # A tibble: 1 x 3
##      q_50   q_90   q_05
##      <dbl> <dbl> <dbl>
## 1 155.   155.   155.
```

50%-PI:

```
m43a %>%
  eti(ci = .5)
## Equal-Tailed Interval
##
## Parameter / 50% ETI / Effects / Component
## -----
## (Intercept) / [154.46, 154.84] / fixed / conditional
## weight_c / [ 0.88, 0.94] / fixed / conditional
```

9.5.8 Fragen zu Wahrscheinlichkeitsmassen des Achsenabschnitts

Wie wahrscheinlich ist es, dass die mittlere Größe bei mind. 155 cm liegt?

```
m43a_post %>%
  count(gross = `^(Intercept)` >= 155) %>%
  mutate(prop = n / sum(n))
## # A tibble: 2 x 3
##   gross     n   prop
##   <dbl> <dbl> <dbl>
```

```
## <lgl> <int> <dbl>
## 1 FALSE 3593 0.898
## 2 TRUE 407 0.102
```

Die Wahrscheinlichkeit beträgt 0.1.

Wie wahrscheinlich ist es, dass die mittlere Größe höchstens 154.5 cm beträgt?

```
m43a_post %>%
  count(klein = `^(Intercept)` <= 154.5) %>%
  mutate(prop = n / sum(n))
## # A tibble: 2 x 3
##   klein     n   prop
##   <lgl> <int> <dbl>
## 1 FALSE  2833 0.708
## 2 TRUE   1167 0.292
```

Die Wahrscheinlichkeit beträgt 0.29.

,

9.5.9 Typischer Bayes-Nutzer in Aktion



Abbildung 9.11: Typischer Bayes-Nutzer, der ein Ungewissheitsintervall berechnet. Bildquelle: Easystats, bayestestR

[Quelle](#)

9.6 Post-Verteilung bedingt auf einen Prädiktorwert

9.6.1 Visualisierung

Was ist wohl die Wahrscheinlichkeit der Körpergröße bei einem bestimmten Gewicht?

Angenommen wir wissen, dass das Gewicht bei, sagen wir 45 kg liegt. Welche Körpergröße ist (im Schnitt) zu erwarten? Wie unsicher sind wir uns über diesen Mittelwert?

Etwas formaler ausgedrückt:

$$\mu | \text{weight} = 45$$

45 kg entspricht genau dem Mittelwert von `weight`. Geht man von zentrierten Prädiktorwerten aus, gilt in dem Fall `weight_c = 0`. Erstellen wir uns dazu eine Tabelle:

```
mu_at_45 <-  
  m43a_post %>%  
  mutate(mu_at_45 = `^ (Intercept)`)
```

Und plotten diese, s. Abbildung 9.12.

```
mu_at_45 %>%  
  ggplot(aes(x = mu_at_45)) +  
  geom_density()
```

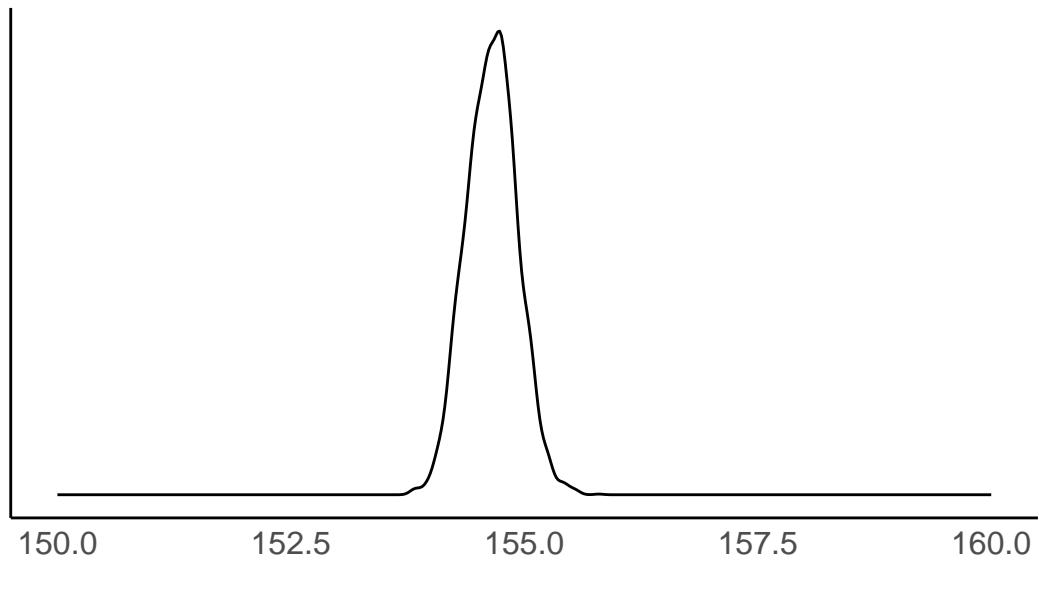


Abbildung 9.12: Post-Verteilung der Größe (laut unserem Modell) bei einem Gewicht von 45kg

Analog können wir fragen, wie groß wohl eine Person mit 50 kg im Mittelwert sein wird und wie (un)gewiss wir uns über diesen Mittelwert sind.

50 kg, das sind 5 über dem Mittelwert, in zentrierten Einheiten ausgedrückt also `weight_c` = 5. Auch dazu erstellen wir uns eine Tabelle.

```
mu_at_50 <-
  mu_at_45 %>%
  mutate(mu_at_50 = `(Intercept)` + 5 * weight_c)

head(mu_at_50)
## # A tibble: 6 x 5
##   `(Intercept)` <dbl> weight_c <dbl> sigma <dbl>
## 1 155.        0.858  5.04    155.     159.
## 2 155.        0.835  5.07    155.     160.
## 3 155.        0.855  5.14    155.     160.
## 4 155.        0.882  5.35    155.     160.
## 5 155.        0.875  5.35    155.     160.
## 6 155.        0.903  5.21    155.     159.
```

Die Verteilung der mittleren Größe bei einem Gewicht von 50kg ist weiter “rechts” (Richtung höhere Größe) zentriert, s. Abbildung 9.13.

```
mu_at_50 %>%
  ggplot(aes(x = mu_at_50)) +
  geom_density()
```

9.6.2 Lagemaße und Streuungen

Befragen wir die bedingte Post-Verteilung. Eine erste Frage zielt nach den typischen deskriptiven Statistiken, also nach Lage und Streuung der Verteilung der Körpergröße.

Was ist das 90% PI für $\mu|w = 50$?

```
mu_at_50 %>%
  eti(mu_at_50, ci = .9)
## Equal-Tailed Interval
##
## Parameter / 90% ETI
## -----
## (Intercept) / [154.21, 155.10]
## weight_c / [ 0.84, 0.98]
## sigma / [ 4.83, 5.48]
## mu_at_45 / [154.21, 155.10]
```

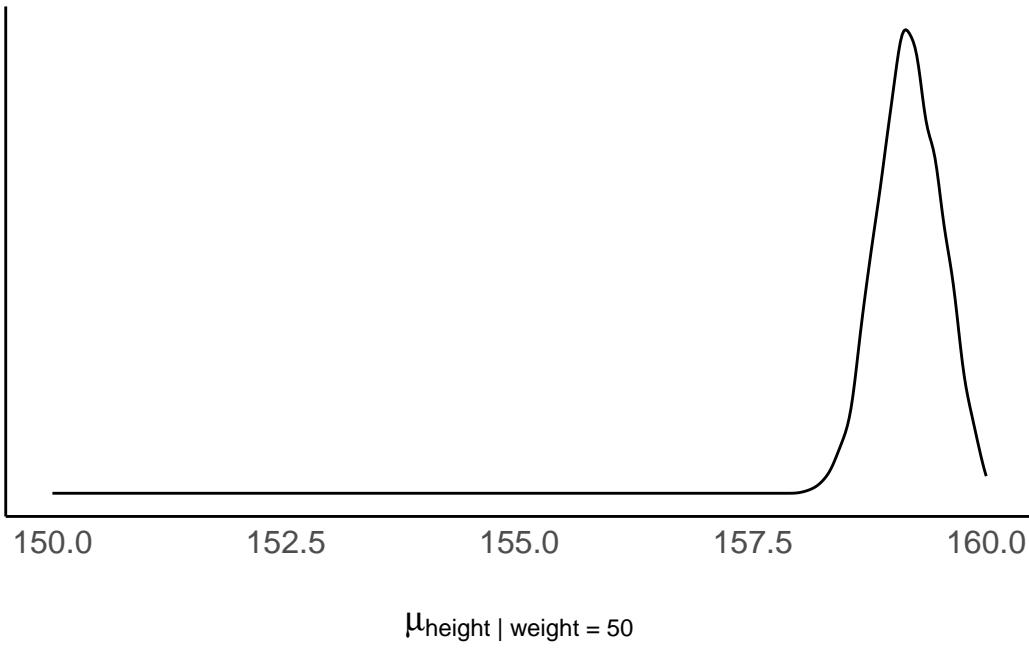


Abbildung 9.13: Post-Verteilung der mittleren Größe (laut unserem Modell) bedingt auf ein Gewicht von 50 kg

```
## mu_at_50      / [158.63, 159.76]
```

Die mittlere Größe - gegeben $w = 50$ - liegt mit 90% Wahrscheinlichkeit zwischen den beiden Werten (ca.) 159cm und 160cm.

Welche mittlere Größe wird mit 95% Wahrscheinlichkeit nicht überschritten, wenn die Person 45kg wiegt?

```
mu_at_45 %>%
  summarise(q_95 = quantile(mu_at_45, prob = .95))
## # A tibble: 1 x 1
##       q_95
##     <dbl>
## 1 155.
```

9.7 Die PPV befragen

Die Posterior-Prädiktiv-Verteilung (PPV) gibt uns die Möglichkeit, nach der Wahrscheinlichkeit *tatsächlicher* Körpergrößen zu fragen - und nicht nur nach *mittleren* Körpergrößen anhand der Post-Verteilung.

! Wichtig

Die Post-Verteilung macht nur Aussagen zur mittleren Körpergröße, denn das ist was wir modellieren wollten. Möchten wir Aussagen zur Wahrscheinlichkeit tatsächlicher Größen treffen, brauchen wir die PPV.

9.7.1 Perzentil-Intervalle für verschiedenen Prädiktor-Werte

Wir erstellen uns eine Sequenz an Prädiktorwerten, die uns interessieren, `weight_df`:

```
weight_df <- tibble(weight_c = seq(-20,20, by = 5))
```

Für diese Werte lassen wir uns dann die Perzentil-Intervalle (PI) ausgeben:

```
mus <-
  predictive_interval(
    m43a,
    seed = 42,
    newdata = weight_df) %>%
  as_tibble() %>%
  bind_cols(weight_df)

head(mus)
## # A tibble: 6 x 3
##   `5%` `95%` weight_c
##   <dbl> <dbl>     <dbl>
## 1 128. 145.     -20
## 2 132. 150.     -15
## 3 137. 154.     -10
## 4 141. 159.      -5
## 5 146. 163.       0
## 6 150. 168.      5
```

Um die Perzentilintervalle zu erstellen, wird von `predictive_interval()` für jeden Prädiktorwert eine PPV erstellt und (in der Voreinstellung) das 5%- sowie 95%-Quantil dafür berechnet. Sie können die Voreinstellung ändern mittels des Arguments `prob`; um ein 89%-PI zu berechnen, würde man z.B. schreiben `prob = .89`.

Um Reproduzierbarkeit sicherzustellen, haben wir mit `seed = 42` die Zufallszahlen fixiert.

Wir sehen etwa, dass wir bei einer Person mittleren Gewichts, eine Körpergröße von ca. 146 cm bis 163 cm zu erwarten haben (95%-KI). Hoppla! Das ist ja viel ungenauer, als die Angaben der Post-Verteilung oben. Ja, denn die Post-Verteilung hat die Ungewissheit zum *Mittelwert* ausgedrückt; die PPV gibt die Ungewissheit *tatsächlicher* beobachtbarer Körpergrößen aus, nicht nur die Ungewissheit zum Mittelwert.

Berechnen wir die PPV für m43a:

```
ppv_m43a <- posterior_predict(  
  m43a,  
  newdata = weight_df,  
  draws = 100) %>%  
  as_tibble() %>%  
  pivot_longer(  
    cols = everything(),  
    names_to = "weight_condition",  
    values_to = "height")
```

Hier ist ein Auszug aus der PPV-Tabelle:

weight_c	5%	95%
-20.0	128.1	145.2
-15.0	132.4	149.6
-10.0	137.0	154.3
-5.0	141.4	158.6
0.0	146.0	163.4
5.0	150.5	167.6

9.7.2 Perzentilintervalle für verschiedenen Prädiktorwerte visualisiert

Abbildung 9.14 visualisiert die Ungewissheit von Vorhersagen laut der PPV. Die Ungewissheit in Abbildung 9.14 ist die Antwort auf die Frage: “Wie sicher sind wir uns, zur Größe einer !Kung-Person, gegeben dass die z.B. 10 kg mehr als der Durchschnitt wiegt?” Eine Vorhersage bezeichnet man auch als “bedingte Verteilung”, da man den Wert einer Verteilung voraussagt, *gegeben* einer Bedingung, z.B. `weight_c = 10`.

Die vertikalen Balken geben die 95%-KI wieder, die wir jeweils zu erwarten haben.

Noch eine andere Visualisierung, s. Abbildung 9.15; je dicker die “Katzenaugen”, desto mehr Stichproben (samples) liegen vor an der Stelle, und umso genauer ist die Schätzung.

Also: Je dicker die Violine, desto wahrscheinlicher $\mu|w_i$.

9.7.3 Die PPV visualisiert

Gerade eben haben wir bedingte PPVen angeschaut: Also eine PPV für einen bestimmten Prädiktorwert, z.B. bei einer Person mittleren Gewichts. Wir können auch den Mittelwert über alle bedingten PPV anschauen, sozusagen die “Master-PPV” oder “unbedingte PPV” oder schlicht PPV. Vergleichen wir die echten Werte für `height`, y , mit den von der PPV simulierten Werten für `height`, y_{rep} , s. Abbildung 9.16.

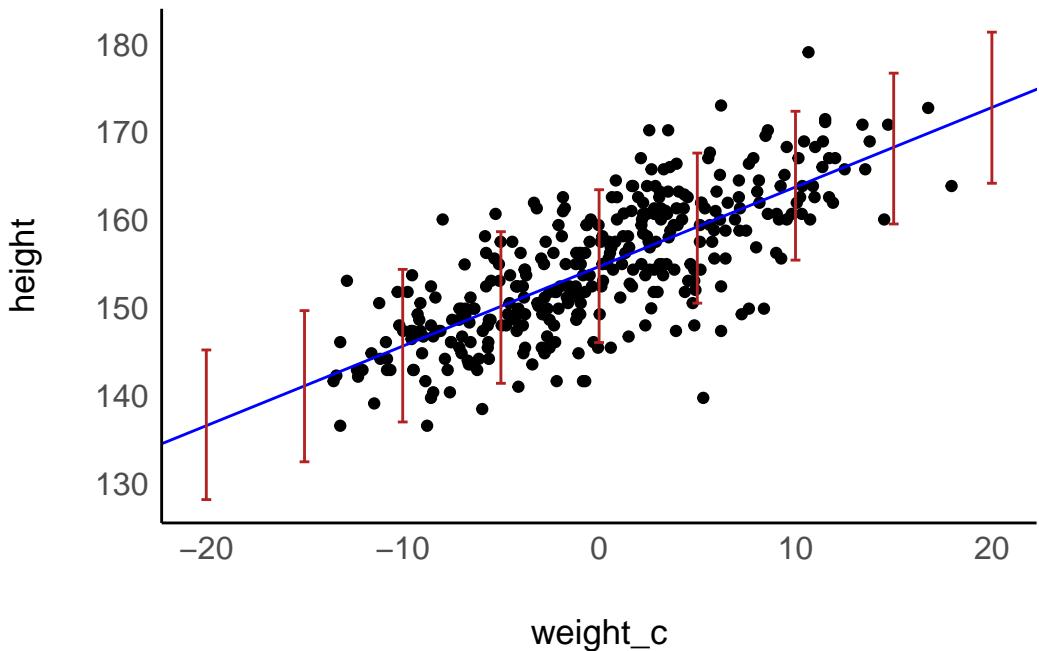


Abbildung 9.14: Visualisierung der Ungewissheit der Vorhersagen laut PPV: Die Vorhersage von Beobachtungen beinhaltet mehr Ungewissheit als die Vorhersage von zu erwartenden Werten, daher sind die Ungewissheitsintervalle der PPV größer als die der Post-Verteilung.

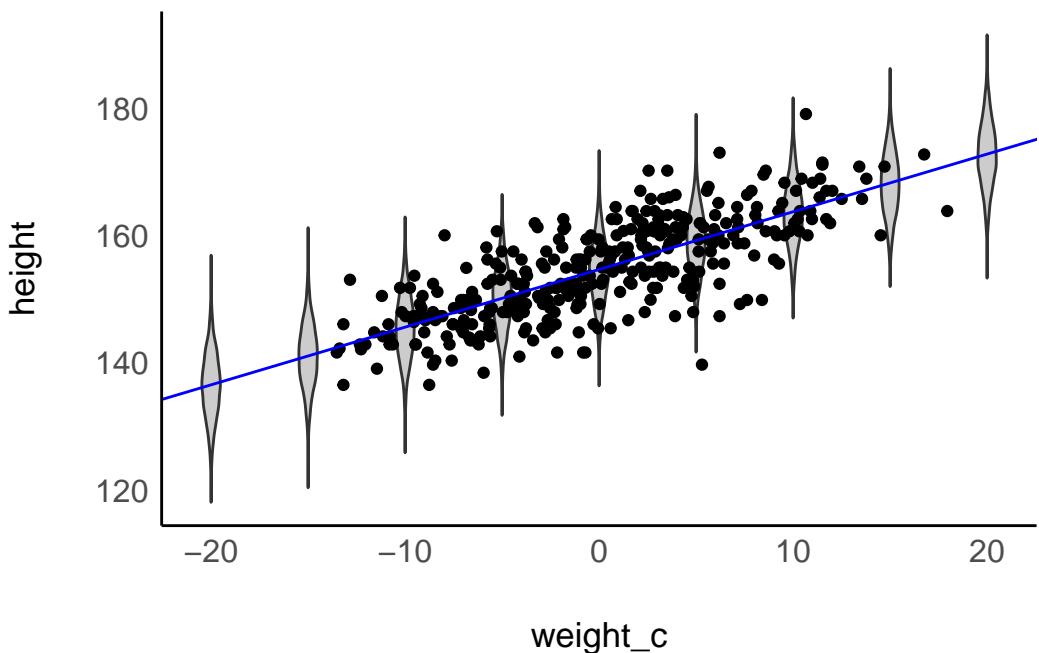


Abbildung 9.15: Die PPV für bestimmte Gewichtswerte w , visualisiert mit Katzenaugendiagrammen

```
check_predictions(m43a)
```

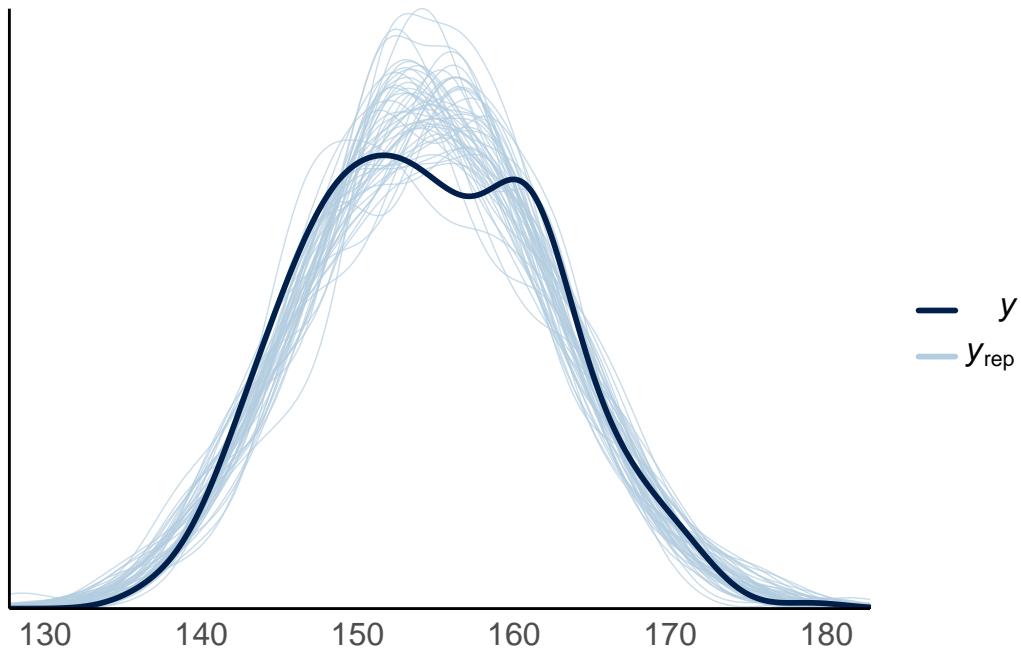


Abbildung 9.16: Vergleich der Vorhersagen für y (leichte, blaue Linien) mit der beobachteten Verteilung von y

?check_predictions zeigt Hilfe für diese Funktion. Die Funktion zeigt die Vorhersagen für die AV laut der Posteriori-Verteilung.

Die zwei Gipfel hat unser Modell nicht mitgekriegt, ansonsten decken sich die Vorhersagen der PPV gut mit den echten Daten.

9.7.4 Fragen an die PPV

- Wie groß sind die !Kung im Schnitt?
- Welche Größe wird von 90% der Personen nicht überschritten?
- Wie groß sind die 10% kleinsten?

```
ppv_m43a %>%
  summarise(
    q_10 = quantile(height, prob = .1),
    height_mean = mean(height),
    q_50 = quantile(height, prob = .5),
    q_90 = quantile(height, prob = .9)
  )
## # A tibble: 1 x 4
```

```
##      q_10 height_mean  q_50   q_90
## <dbl>       <dbl> <dbl> <dbl>
## 1 137.        155. 155. 172.
```

Was ist der 50% Bereich der Körpergröße?

```
ppv_m43a %>%
  eti(ci = .5)
## Equal-Tailed Interval
##
## Parameter /      50% ETI
## -----
## height    / [144.82, 165.24]
```

9.8 Aufgaben

- Bayes-Ziel1
- Bayesmod-bestimmen01
- Likelihood2
- Post-befragen1
- Postvert-Regr-01
- regression1a
- Regression2
- Bed-Post-Wskt1
- Priorwahl1
- Bayesmod-bestimmen02
- Aussagen-einfache-Regr
- Likelihood-identifizieren
- Priorwahl2
- penguins-stan-01

10 Forschungsfragen mit metrischer AV

10.1 Lernsteuerung

10.1.1 R-Pakete

Benötigte R-Pakete für dieses Thema:

```
suppressPackageStartupMessages("rstanarm")
## [1] "rstanarm"
#| message: false
#| results: "hide"
#| warnings: false
library(rstanarm)
library(tidyverse)
library(easystats)
```

10.1.2 Lernziele

Nach Absolvieren des jeweiligen Kapitels sollen folgende Lernziele erreicht sein.

Sie können...

- typische, deskriptive Forschungsfragen spezifizieren als Regression
- Forschungsfragen in Regressionsterme übersetzen
- typische Forschungsfragen auswerten

10.1.3 Begleitvideos

- [Teil 1](#)
- [Teil 2](#)

10.2 Wissenschaft als Gerechtigkeitsprojekt

10.2.1 Meinungen als Grundlage der Konfliktlösung ?

Contra:

- “Ich find Masken doof!”
- “Impfen ist schädlich!”
- “Corona gibt’s gar nicht!”

Pro:

- “Ich find Masken gut!”
- “Impfen ist nützlich!”
- “Corona ist gefährlich!”

Meinungen kennen kein richtig und kein falsch: Meinungen sind keine Fakten. Konflikte können auf Basis von Meinungen nur schwer gelöst werden.

10.2.2 Fakten als Grundlage der Konfliktlösung

Wissenschaft produziert Fakten. Da Fakten universell sind (sein können), ist Wissenschaft potenziell ein Weg zur Konfliktlösung. Warum helfen Fakten bei Konflikten?

Fakten sind neutral gegenüber Personen. Fakten bieten daher eine Chance zur fairen Eini-gung.

Wann ist ein Fakt ein Fakt?

Fakten müssen vor allem nachprüfbar sein (Daten, Analyse und Bericht müssen offen zu-gänglich sein).

10.2.3 Beispiel Corona: Datenlage spricht zugunsten der Covid19-Impfung

The effectiveness of full messenger RNA (mRNA) vaccination (14 days after the second dose) was 89% (95% confidence interval [CI], 87 to 91) against laboratory-confirmed SARS-CoV-2 infection leading to hospitalization, 90% (95% CI, 86 to 93) against infection leading to an ICU admission, and 91% (95% CI, 89 to 93) against infection leading to an emergency department or urgent care clinic visit.

Thompson u. a. (2021); vgl. auch Nasreen u. a. (2021); Pormohammad u. a. (2021)

Drei Anforderungen an die Qualität von Studien:

1. *handwerklich gut*: z.B. vergleichbare Gruppen, genaue Messinstrumente

2. *bescheiden*: die Forschungsfrage wird nur dann selbstbewusst beantwortet, wenn es die handwerkliche Qualität der Studie zulässt. Gibt es eine Vielzahl weiterer Studien mit abweichenden Ergebnissen, wird dies bei der Beantwortung der Forschungsfrage berücksichtigt.
3. *transparent*: Das Vorgehen, die Hintergründe und Ziele werden offengelegt. Das betrifft auch möglich Befangenheit oder Interessenskonflikte der Autoren und Autorinnen

10.2.4 Psychologische Intervention zur Erhöhung der Impfquote

Dai u. a. (2021) zeigen den Effekt einer psychologischen Intervention zur Erhöhung der Impfquote, s. Abbildung 10.1.

Here we present two sequential randomized controlled trials to test the effect of behavioural interventions on the uptake of COVID-19 vaccines. ... We designed text-based reminders that make vaccination salient and easy, and delivered them to participants drawn from a healthcare system one day (first randomized controlled trial) ($n = 93,354$ participants; clinicaltrials number NCT04800965) and eight days (second randomized controlled trial) ($n = 67,092$ individuals; clinicaltrials number NCT04801524) after they received a notification of vaccine eligibility. The first reminder boosted appointment and vaccination rates within the healthcare system by 6.07 (84%) and 3.57 (26%) percentage points, respectively; the second reminder increased those outcomes by 1.65 and 1.06 percentage points, respectively. The first reminder had a greater effect when it was designed to make participants feel ownership of the vaccine dose.

[Quelle/Volltext](#)

10.2.5 Was heißt "ist effektiv"?

Nasreen u. a. (2021) definieren *effectivity*, e , so:

$$e = 1 - C; C = \frac{n_{vacc|pos}}{n_{vacc|neg}}$$

- C nennt man das *Chancenverhältnis (odds ratio)*, es beschreibt einen Bruchterm: $\frac{x}{y}$.
- $n_{vacc|pos}$: Anzahl der geimpften Personen unter allen Personen mit positiver Corona-Diagnose
- $n_{vacc|neg}$: Anzahl der geimpften Personen unter allen Personen mit negativer Corona-Diagnose

Beispiel: Von den 100 Personen mit *positiver* Corona-Diagnose sind 10 geimpft, $n_{vacc|pos} = 10$. Von den 100 Personen mit *negativer* Corona-Diagnose sind 90 geimpft, $n_{vacc|neg} = 90$

$$C = \frac{10}{90} = \frac{1}{9}; e = 1 - \frac{1}{9} = \frac{8}{9} \approx 0.88$$

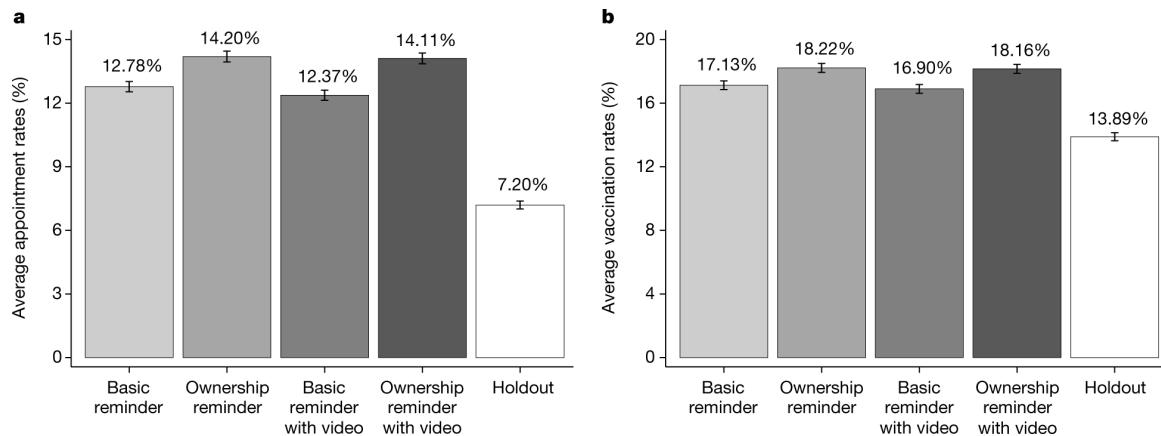


Abbildung 10.1: a, b, Proportion of participants in each condition who scheduled an appointment for the first dose of the COVID-19 vaccine at UCLA Health between 15:00 h on the first reminder date and 23:59 h on the fifth day following the first reminder date (a) and the proportion of participants in each condition who obtained the first dose of the COVID-19 vaccine at UCLA Health within four weeks of the first reminder date (b). Error bars represent ± 1 s.e.m. The number of participants in each condition (from left to right in each panel) is 18,629, 18,592, 18,757, 18,627 and 18,749.

In diesem Beispiel liegt die Effektivität e bei knapp 90%.

10.3 Arten von Forschungsfragen

10.3.1 Nach dem Erkenntnisziel

Deskriptiv (beschreibend)

- Wie stark ist der (lineare) Zusammenhang r von Größe und Gewicht?
- Wie stark ist der (lineare) Zusammenhang b von Lernzeit und Note?
- Bevorzugen unsere Kunden Webshop A oder B?

Prädiktiv (prognostisch, vorhersagend)

- Wie schwer ist ein deutscher Mann der Größe 1,80m im Schnitt?
- Welche Note kann man erwarten, wenn man nichts für die Klausur lernt?
- Wieviel wird ein Kunde ausgeben, wenn er sich in dieser Variante des Webshops aufhält?

Präskriptiv (erklärend, kausal)

- Ist Größe eine Ursache von Gewicht (bei deutschen Männern)?
- Wenn ich 100 Stunden lerne, welche Note schreibe ich dann?
- Hat die Art des Webshops einen Einfluss auf unseren Umsatz?

i Hinweis

Das Erkenntnisziel wissenschaftlicher Studien ist zumeist erklärend. Anhand der verwendeten statistischen Methode (z.B. Regressionsanalyse) kann man nicht feststellen, zu welchem Erkenntnisziel die Studie gehört.

10.3.2 Nach dem Skalenniveau

Wir konzentrieren uns im Folgenden auf Forschungsfragen auf Basis von Regressionsmodellen mit *metrischer AV*. Andere Skalenniveaus bei der AV klammern wir aus.

Für die UV(s) sind nominale und metrische Skalenniveaus erlaubt. Modelle mit mehreren UV (und mehreren Stufen an UV) sind erlaubt.

10.4 Eine binäre UV

10.4.1 Forschungsfrage

Hintergrund:

Eine Psychologin, die im öffentlichen Dienst arbeitet, versucht herauszufinden, warum einige Kinder intelligenter sind als andere. Dazu wurden in einer aufwändigen Studie die Intelligenz vieler Kinder gemessen. Zusätzliche wurden verschiedene Korrelate der Intelligenz erhoben, in der Hoffnung, „Risikofaktoren“ für geringere Intelligenz zu entdecken.

Forschungsfrage:

Unterscheidet sich der mittlere IQ-Wert (`kid_score`) von Kindern in Abhängigkeit davon, ob ihre jeweilige Mutter über einen Schulabschluss (`mom_hs`) verfügt? (ceteris paribus)

10.4.2 IQ von Kindern, binärer Prädiktor

```
data("kidiq") # Paket rstanarm
m10.1 <- stan_glm(
  kid_score ~ mom_hs,
  seed = 42,
  data = kidiq)
```

Alternativ können Sie die Daten [hier](#) herunterladen.

Mit `parameters(m10.1)` bekommt man die Parameter des Modells, s. Tabelle 10.1.

Tabelle 10.1: Parameter des Modells m10.1 (sigma ist nicht dargestellt)

Parameter	Median	95% CI	pd	Rhat	ESS	Prior
(Intercept)	77.56	(73.28, 81.64)	100%	1.001	3917.00	Normal (86.80 +- 51.03)
mom_hs	11.80	(7.18, 16.48)	100%	1.001	3789.00	Normal (0.00 +- 124.21)

In Abbildung 10.2 ist der Unterschied im IQ der Kinder als Funktion des Schlussabschlusses der Mutter dargestellt.

```
ggplot(kidiq) +
  aes(x = mom_hs, y = kid_score) +
  geom_jitter(width = 0.1, alpha = .5) +
  geom_abline(slope = coef(m10.1)[2],
              intercept = coef(m10.1)[1]) +
  scale_x_continuous(breaks = c(0, 1))
```

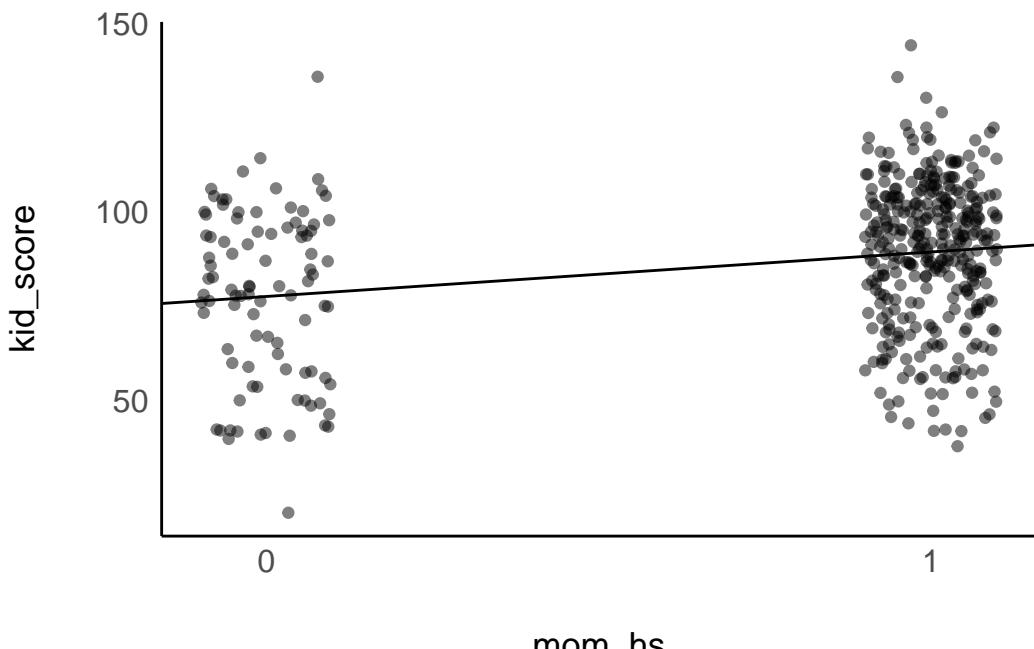


Abbildung 10.2: Kinder, deren Mütter über einen Schulabschluss verfügen, haben im Mittel einen höheren Intelligenztestwert, laut dem vorliegenden Modell

10.4.3 Interpretation von m10.1

```
m10.1: kid_score = 78 + 12*mom_hs + error
```

- Der *Achsensabschnitt* (intercept, β_0 oder auch mit α bezeichnet) ist der mittlere (bzw. vorhergesagte) IQ-Wert von Kindern, deren Mütter über keinen Schulabschluss (`mom_hs = 0`) verfügen:

```
kid_score = 78 + 0*12 + error
```

- Das *Regressionsgewicht* (slope, β) ist der Unterschied im IQ-Wert von Kindern mit Mütter mit Schulabschluss (im Vergleich zum IQ-Wert von Kindern mit Mütter ohne Schulabschluss). Dieser Unterschied entspricht der Steigung der Regressionsgeraden.

```
kid_score = 78 + 1*12 + error = 90 + error
```

- Die Größer der Konfidenzintervalle zeigt, wie genau die Schätzung (Vorhersage) ist bzw. wie stark Prädiktor (UV) und Kriterium (AV) zusammenhängen.

10.4.4 m10.1 als Mittelwertsdifferenz

- UV: binär (zweistufig nominal/kategorial)
- AV: metrisch (quantitativ)

Hey R-Golem! Nimm den Datensatz `kidiq`, gruppieren nach `mom_hs` und fasse zusammen anhand des Mittelwerts. Die resultierende Zahl soll heißen `kid_score_avg`. An die Arbeit!

```
kidiq %>%
  group_by(mom_hs) %>%
  summarise(kid_score_avg =
            mean(kid_score))
## # A tibble: 2 x 2
##   mom_hs kid_score_avg
##   <int>      <dbl>
## 1     0        77.5
## 2     1        89.3
```

In der klassischen Statistik untersucht man diese Datensituation mit einem *t-Test*. Der t-Test ist ein inferenzstatistisches Verfahren, dass prüft, ob die Mittelwertsdifferenz (in der Population) μ_d Null ist: $\mu_d = 0$. In der Bayes-Statistik betrachtet man dazu stattdessen die Posteriori-Verteilung (z.B. mit 95%PI).

Der mittlere (average, avg) IQ-Wert unterscheidet sich um ca. 12 Punkte (89.4-77.6), zu gunsten der Kinder von Müttern mit Abschluss. Allerdings gibt es viel Streuung um die Mittelwerte herum.

10.4.5 Antwort auf die Forschungsfrage, m10.1

Betrachten wir die Ergebnisse von m10.1. Hier sind die ersten paar Zeilen.

```

m10.1_post <-
  m10.1 %>%
  as_tibble()

names(m10.1_post) <- c("Achsenabschnitt", "momhs", "sigma") # schönere Namen

```

Stichprobe aus der Post-Verteilung

Achsenabschnitt	momhs	sigma
76.4	14.1	20.5
80.2	6.8	20.5
74.8	15.1	19.7
80.0	10.2	18.7
78.5	10.0	20.6

Berechnen wir ein 95%-PI von Hand:¹

```

pi_mom_hs <-
  m10.1_post %>%
  summarise(pi_95 = quantile(momhs, c(.025, .975)))

pi_mom_hs
## # A tibble: 2 x 1
##   pi_95
##   <dbl>
## 1 7.18
## 2 16.5

```

Mit 95% Wahrscheinlichkeit liegt der Unterschied im mittleren IQ-Wert zwischen Kindern von Müttern mit bzw. ohne Schulabschluss im Bereich von 7 bis 14 IQ-Punkten, laut unserem Modell: $95\%PI : [7, 16]$. Die Hypothese, dass es keinen Unterschied oder einen Unterschied in die andere Richtung geben sollte, ist vor diesem Hintergrund als unwahrscheinlich abzulehnen.

Visualisieren wir abschließend die Posteriori-Verteilung, s. Abbildung 10.3.

```
plot(eti(m10.1))
```

¹komfortabler geht es mit `eti(m10.1)`.

Credible Interval (CI)

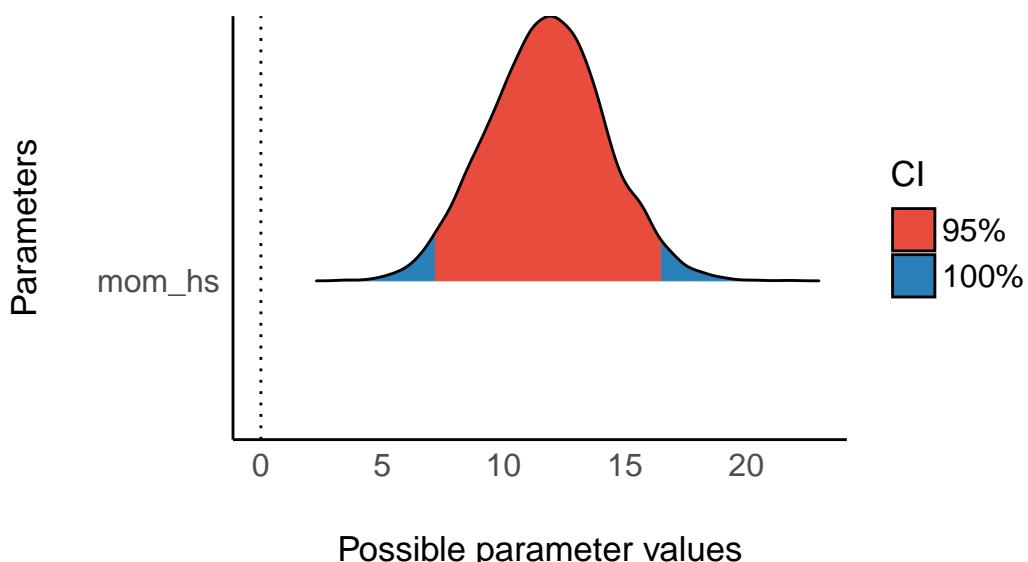


Abbildung 10.3: Das 95% ETI zum (statistischen) Effekt des mütterlichen Schulabschlusses

Zur Einnerung: Korrelation ungleich Kausation. Von einem “Effekt” zu sprechen, lässt in den meisten Köpfen wohl die Assoziation zu einem *kausalen* Effekt entstehen. Ein Kausaleffekt ist eine starke (und sehr interessante und wichtige) Behauptung, die mehr Fundierung bedarf als eine einfache Korrelation bzw. ein einfacher Zusammenhang.

10.5 Eine metrische plus eine nominale UV

10.5.1 Forschungsfrage

Wie stark ist der Zusammenhang von jeweils Schulabschluss der Mutter (`mom_hs`) und IQ der Mutter (`mom_iq`) auf den IQ des Kindes (`kid_score`) ?

Deskriptive Statistiken zum Datensatz sind in Tabelle Tabelle 10.3 dargestellt.

```
data("kidiq") # Paket rstanarm, alternativ über CSV einlesen
describe_distribution(kidiq)
```

Tabelle 10.3: Variablen und ihre Verteilung im Datenatz kidiq

Variable	Mean	SD	IQR	Range	Skewness	Kurtosis	n	n_Missing
kid_score	86.80	20.41	28.00	(20.00, 144.00)	-0.46	-0.16	434	0

Variable	Mean	SD	IQR	Range	Skewness	Kurtosis	n	n_Missing
mom_hs	0.79	0.41	0.00	(0.00, 1.00)	-1.40	-0.05	434	0
mom_iq	100.00	15.00	21.67	(71.04, 138.89)	0.47	-0.57	434	0
mom_age	22.79	2.70	4.00	(17.00, 29.00)	0.18	-0.63	434	0

Datenquelle

10.5.2 1 metrischer Prädiktor

Berechnen wir folgendes Modell: `kid_score ~ mom_iq` (m10.2), s. Tab. Tabelle 10.4.

```
m10.2 <-
  stan_glm(kid_score ~ mom_iq, data = kidiq, seed = 42)

m10.2 %>%
  parameters()
```

Tabelle 10.4: Parameter des Modells m10.2

Parameter	Median	95% CI	pd	Rhat	ESS	Prior
(Intercept)	25.78	(14.04, 36.99)	100%	1.000	3518.00	Normal (86.80 +- 51.03)
mom_iq	0.61	(0.50, 0.73)	100%	1.000	3486.00	Normal (0.00 +- 3.40)

```
kid_score = 26 + 0.6 * mom_iq + error
```

Visualisieren wir uns noch das Modell m10.2, s. Abbildung 10.4.

```
kidiq %>%
  ggplot(aes(x = mom_iq, y = kid_score)) +
  geom_point(alpha = .7) +
  geom_abline(slope = coef(m10.2)[2],
              intercept = coef(m10.2)[1],
              color = "blue")
```

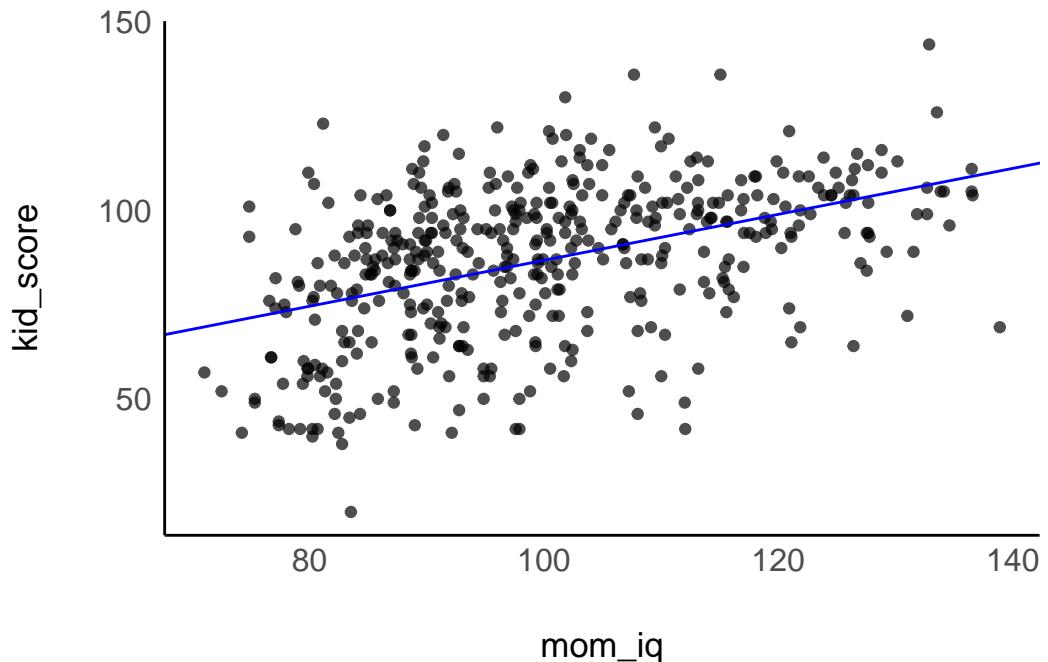


Abbildung 10.4: Die Intelligenz eines Kindes als Funktion der Intelligenz der Mutter (m10.2)

Alternativ kann man sich - einfacher - das Modell (m10.2) so visualisieren, s. Abbildung 10.5.

```
plot(estimate_expectation(m10.2))
```

Predicted response (kid_score ~ mom_iq)

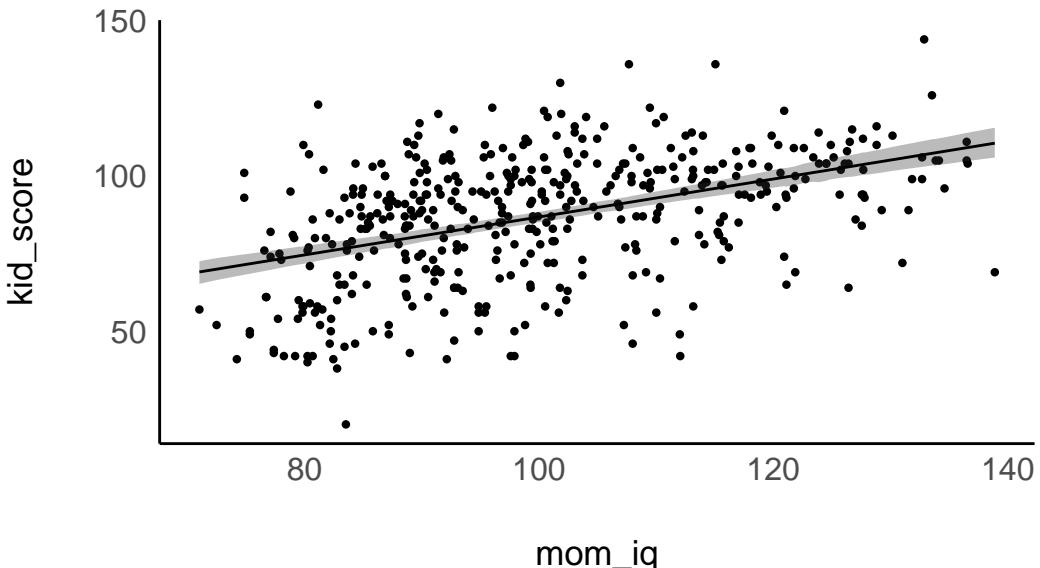


Abbildung 10.5: Die geschätzten Erwartungswerte von m10.2 visualisiert

Die Linie zeigt die vorhergesagten IQ-Werte der Kinder für verschiedene IQ-Werte der Mütter. Vergleicht man Teilpopulationen von Müttern mit mittleren Unterschied von einem IQ-Punkt, so findet man 0.6 IQ-Punkte Unterschied bei ihren Kindern *im Durchschnitt*, laut dem Modell m10.2. Der Achsenabschnitt hilft uns nicht weiter, da es keine Menschen mit einem IQ von 0 gibt.

10.5.3 Beide Prädiktoren, m10.3

Berechnen wir als nächstes ein Modell mit beiden Prädiktoren: `kid_score ~ mom_hs + mom_iq`, s. Tabelle 10.5.

```
m10.3 <-
stan_glm(
  kid_score ~ mom_iq + mom_hs,
  refresh = 0,
  seed = 42,
  data = kidiq)
```

Tabelle 10.5: Parameter des Modells m10.3 (ohne sigma)

Parameter	Median	95% CI	pd	Rhat	ESS	Prior
(Intercept)	25.74	(13.87, 36.76)	100%	1.001	3961.00	Normal (86.80 +- 51.03)

Parameter	Median	95% CI	pd	Rhat	ESS	Prior
mom_iq	0.57	(0.45, 0.69)	100%	1.001	3456.00	Normal (0.00 +- 3.40)
mom_hs	6.04	(1.62, 10.15)	99.60%	0.999	3616.00	Normal (0.00 +- 124.21)

Will man nur schnell die Koeffizienten des Modells (d.h. Punktschätzer der Modellparametern, in diesem Fall den Median) wissen, so kann man anstelle von `parameters(mein_modell)` auch `coef(mein_modell)` schreiben:

```
coef(m10.3)
## (Intercept)      mom_iq      mom_hs
## 25.7447712   0.5654851   6.0376396

m10.3: kid_score = 26 + mom_hs + 0.6*mom_iq + error
```

Möchte man nur z.B. den 3. Wert aus diesem Vektor, so kann man schreiben:

```
coef(m10.3)[3]
## mom_hs
## 6.03764
```

Aber natürlich ist es möglich (und einfacher) anstelle von `coef` den Befehl `parameters` zu verwenden.

Und die Visualisierung des Modells `m10.3`, s. Abbildung 10.6.

```
kidiq2 <-
  kidiq %>%
  mutate(mom_hs = as.factor(mom_hs))

m10.3a <-
  stan_glm(
    kid_score ~ mom_iq + mom_hs,
    refresh = 0,
    seed = 42,
    data = kidiq2)

plot(estimate_expectation(m10.3a))
```

Predicted response ($\text{kid_score} \sim \text{mom_iq} + \text{mom_hs}$)

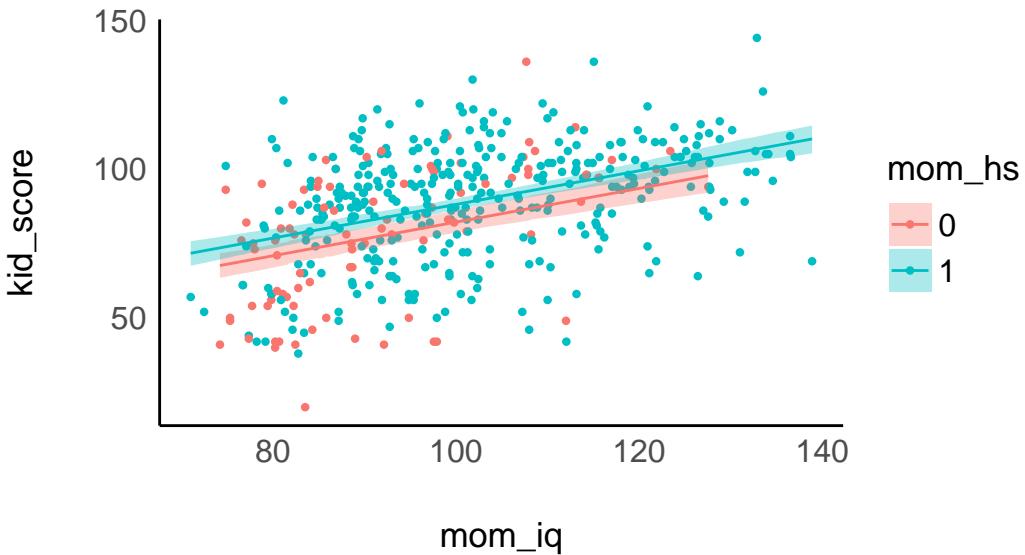


Abbildung 10.6: Der Effekt von sowohl mütterlicher Intelligenz als auch mütterlichem Schulabschluss.

- *Achsenabschnitt:* Hat das Kind eine Mutter mit einem IQ von 0 und ohne Schulabschluss, dann schätzt das Modell den IQ-Wert des Kindes auf 26.
- *Koeffizient zum mütterlichen Schulabschluss:* Vergleicht man Kinder von Müttern gleicher Intelligenz, aber mit Unterschied im Schulabschluss, so sagt das Modell einen Unterschied von 6 Punkten im IQ voraus.
- *Koeffizient zur mütterlichen IQ:* Vergleicht man Kinder von Müttern mit gleichem Wert im Schulabschluss, aber mit 1 IQ-Punkt Unterschied, so sagt das Modell einen Unterschied von 0.6 IQ-Punkten bei den Kindern voraus.

10.6 Interaktion

In m10.3 hat das Modell die Regressionsgeraden gezwungen, parallel zu sein. Betrachtet man das Streudiagramm, so sieht man, dass nicht-parallele Geraden besser passen. Sind die Regressionsgeraden nicht parallel, so spricht man von einer Interaktion (synonym: Interaktionseffekt, Moderation).

! Wichtig

Liegt eine Interaktion vor, so unterscheidet sich die Steigung der Geraden in den Gruppen.

Wir berechnen mit m10.4 folgendes Modell: $\text{kid_score} \sim \text{mom_hs} + \text{mom_iq} + \text{mom_hs}:\text{mom_iq}$, s. Abbildung 10.7 und Tabelle 10.6.

```
m10.4 <-
  stan_glm(kid_score ~ mom_hs + mom_iq + mom_hs:mom_iq,
            seed = 42,
            data = kidiq,
            refresh = 0)
```

Tabelle 10.6: Parameter von m10.4

Parameter	Median	95% CI	pd	Rhat	ESS	Prior
(Intercept)	-10.59	(-37.44, 16.30)	77.62%	1.001	1427.00	Normal (86.80 +- 51.03)
mom_hs	49.76	(21.49, 79.27)	100%	1.003	1401.00	Normal (0.00 +- 124.21)
mom_iq	0.96	(0.67, 1.25)	100%	1.001	1412.00	Normal (0.00 +- 3.40)
mom_hs:mom_iq	-0.47	(-0.78, -0.17)	99.85%	1.003	1367.00	Normal (0.00 +- 1.16)

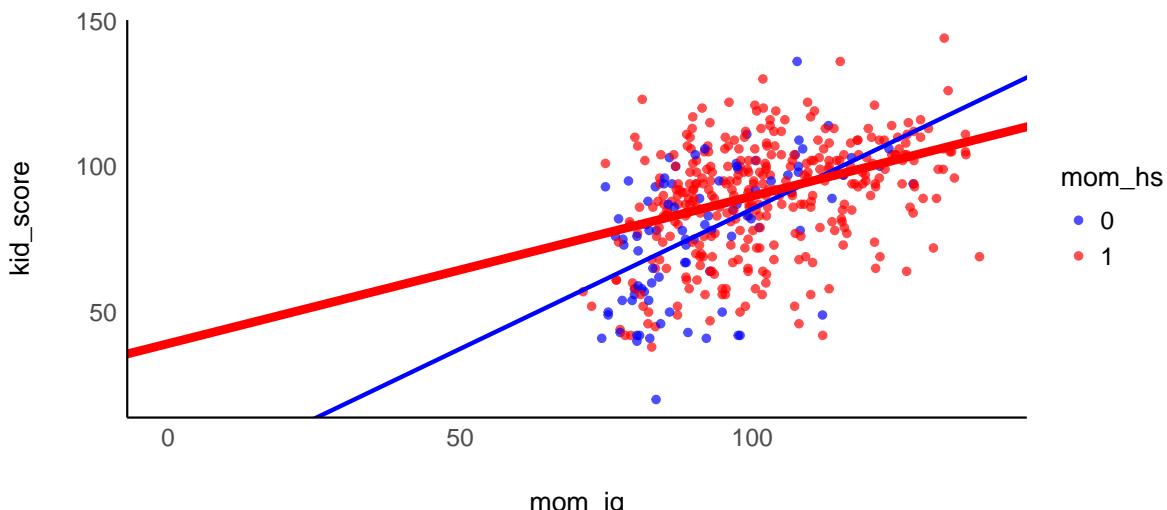


Abbildung 10.7: Wie m10.3, aber mit Interaktionseffekt. Es ist gut zu erkennen, dass der Achsenabschnitt für diese Daten kaum zu interpretieren ist.

10.6.1 Interpretation von m10.4

- *Achsenabschnitt:* IQ-Schätzwerte für Kinder mit Mütter ohne Abschluss und mit einem IQ von 0. Kaum zu interpretieren.
- *mom_hs:* Unterschied der IQ-Schätzwerte zwischen Kindern mit Mutter ohne bzw. mit Schulabschluss und jeweils mit einem IQ von 0. Puh.

- `mom_iq`: Unterschied der IQ-Schätzwerte zwischen Kindern mit Müttern, die sich um einen IQ-Punkt unterscheiden aber jeweils ohne Schulabschluss.
- *Interaktion*: Der Unterschied in den Steigungen der Regressiongeraden, also der Unterschied des Koeffizienten für `mom_iq` zwischen Mütter mit bzw. ohne Schulabschluss.

```
mom_hs=0:
kid_score = -11 + 51*0 + 1.1* mom_iq + 0.5*0*mom_iq
            = -11 + 1.1*mom_iq
```

```
mom_hs=1:
kid_score = -11 + 51*1 + 1.1* mom_iq + 0.5*1*mom_iq
            = 40 + 0.6*mom_iq
```

Gelman, Hill, und Vehtari (2021), Kap. 10.3

10.6.2 Nach der Interpretation von 20 unzentrierten Koeffizienten

via GIPHY

10.7 Zentrieren von Prädiktoren

Unter *Zentrieren* (to center) versteht man das Bilden der Differenz eines Messwerts zu seinem Mittelwert. Zentrierte Werte geben also an, wie weit ein Messwert vom mittleren (typischen) Messwert entfernt ist. Mit zentrierten Werten ist eine Regression einfacher zu interpretieren. Hier zentrieren wir (nur) `mom_iq`.

Man könnte auch `mom_hs` zentrieren, aber für eine einfache Interpretation ist es meist nützlich, nur metrische Prädiktoren zu zentrieren.

```
kidiq <-
  kidiq %>%
  mutate(mom_iq_c = mom_iq - mean(mom_iq))

m10.5 <- stan_glm(kid_score ~ mom_hs + mom_iq_c + mom_hs:mom_iq_c,
                     data = kidiq,
                     seed = 42,
                     refresh = 0)

coef(m10.5)

## NULL
```

10.7.1 Interpretation von m10.5

- Der *Achsenabschnitt (Intercept)* gibt den geschätzten IQ des Kindes an, wenn man eine Mutter *mittlerer* Intelligenz und *ohne* Schulabschluss betrachtet.
- *mom_hs* gibt den Unterschied im geschätzten IQ des Kindes an, wenn man Mütter mittlerer Intelligenz aber mit bzw. ohne Schlusabschluss vergleicht.
- *mom_iq_c* gibt den Unterschied im geschätzten IQ des Kindes an, wenn man Mütter ohne Schlusabschluss aber mit einem IQ-Punkt Unterschied vergleicht.
- *mom_hs:mom_iq_c* gibt den Unterschied in den Koeffizienten für *mom_iq_c* an zwischen den beiden Gruppen von *mom_hs*.

m10.5 ist in Abbildung 10.8 dargestellt.

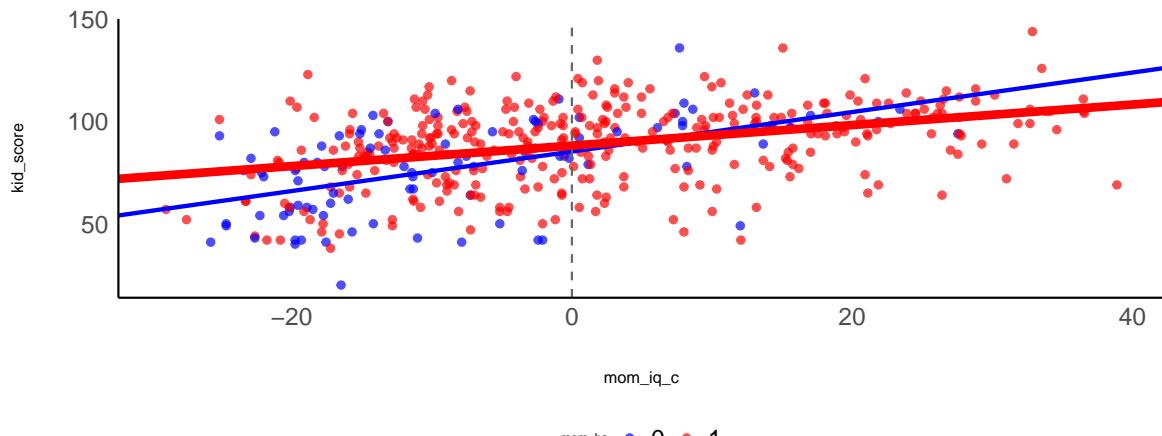


Abbildung 10.8: m10.5: Interaktionsmodell mit zentriertem Prädiktor für mütterlicher Intelligenz

10.7.2 Zentrieren ändert nichts an den Vorhersagen

Betrachten wir die Vorhersagen von m10.4:

```
new <- tibble(mom_hs = 0, mom_iq = mean(kidiq$mom_iq))
pred_new <- posterior_predict(m10.4, newdata = new)
mean(pred_new)
## [1] 85.25181
```

Und vergleichen wir mit diesen die Vorhersagen von m10.5:

```
new <- tibble(mom_hs = 0, mom_iq_c = 0)
pred_new <- posterior_predict(m10.5, newdata = new)
mean(pred_new)
## [1] 85.28128
```

Tabelle 10.7: ?(caption)

(a) Fixed effects

Parameter	Median	95% CI	pd	Rhat	ESS	Prior
(Intercept)	85.31	(80.99, 89.72)	100%	1.001	2610.00	Normal (86.80 +- 51.03)
mom_hs	2.91	(-1.89, 7.69)	88.00%	1.001	2832.00	Normal (0.00 +- 124.21)
mom_iq_c	0.97	(0.67, 1.24)	100%	1.002	1982.00	Normal (0.00 +- 3.40)
mom_hs:mom_iq_c	-0.48	(-0.78, -0.16)	99.78%	1.002	1992.00	Normal (0.00 +- 3.87)

Parameter von m10.5 und ETIs

Wir sehen, die Vorhersagen sind (bis auf Rundungsfehler) identisch.

Auch die Streuungen der vorhergesagten Werte unterscheiden sich nicht (wirklich): $\sigma_{m10.4} = 18$; $\sigma_{m10.5} = 18$.

Das Zentrieren ändert auch *nicht* die Regressionskoeffizienten, da die Streuungen der Prädiktoren nicht verändert wurden.

10.7.3 Perzentilintervalle aus der Posteriori-Verteilung

Tabelle 10.7 zeigt die Punktschätzer der Parameter für m10.5 sowie ihre Perzentilintervalle². Nutzen Sie dafür `parameters(m10.5)`.

Highest Density (Posterior) Intervalle (HDI oder HDPI) kann man sich komfortabel ausgeben lassen mit `hdi(m10.5)` oder mit `parameters(m10.5, ci_method = "hdi")`, s. Tabelle 10.8

```
parameters(m10.5, ci_method = "hdi") %>%
  display()
```

Tabelle 10.8: Parameter von m10.5 und HDIs

Parameter	Median	95% CI	pd	Rhat	ESS	Prior
(Intercept)	85.31	(81.26, 89.88)	100%	1.001	2610.00	Normal (86.80 +- 51.03)
mom_hs	2.91	(-1.89, 7.70)	88.00%	1.001	2832.00	Normal (0.00 +- 124.21)
mom_iq_c	0.97	(0.68, 1.24)	100%	1.002	1982.00	Normal (0.00 +- 3.40)
mom_hs:mom_iq_c	-0.48	(-0.79, -0.17)	99.78%	1.002	1992.00	Normal (0.00 +- 3.87)

²auch ETI genannt

Im Falle symmetrischer Posteriori-Verteilungen (wie hier) kommen beide Arten von Intervallen zu gleichen Ergebnissen.

10.7.4 Beantworten der Forschungsfrage

Das Modell zeigt keine Belege, dass sich die mittlere Intelligenz von Kindern bei Müttern mit bzw. ohne Schulabschluss unterscheidet (95%PI: [-2.0, 7.8]). Hingegen fand sich ein Effekt der mütterlichen Intelligenz; pro Punkt Unterschied in mittlerlichem IQ fand sich ein Unterschied von 0.7 bis 1.3 IQ-Punkte (95%PI). Außerdem fand sich ein Beleg, dass der Zusammenhang des IQ zwischen Mutter und Kind durch den Schulabschluss moderiert wird: Bei Mütter mit Schulabschluss war der Zusammenhang zwischen Mutter-IQ und Kind-IQ geringer (95%PI: [-0.80, -0.17]).

! Wichtig

Das Modell macht *keine* kausalen Aussagen. Es werden lediglich Unterschiede bzw. Zusammenhänge beschrieben. Für kausale Aussagen ist mehr nötig, als einen Zusammenhang festzustellen.

10.8 Eine nominale UV mit mehreren Stufen

10.8.1 Forschungsfrage

Hintergrund:

Nach Ihrem Studium wurden Sie reich als Unternehmensberater:in; Ihre Kompetenz als Wirtschaftspsychologe war heiß begehrte. Von Statistik wollte niemand etwas wissen... Doch nach einiger Zeit kamen Sie in eine Sinnkrise. Sie warfen Ihre Job hin und beschlossen, in die Wissenschaft zu gehen. Kurz entschlossen bewarben Sie sich auf das erste Stellenangebot als Nachwuchswissenschaftler:in.

Ihr Forschungsprojekt führte Sie in die Antarktis... Nun, das war zumindest ein Gegenentwurf zu Ihrem bisherigen Jet-Set-Leben.

Ihre Aufgabe bestand nun darin, Pinguine zu untersuchen. Genauer gesagt ging es um Größenunterschiede zwischen drei Pinguinarten. Ja, stimmt, an so ein Forschungsprojekt hatten Sie vorher nie auch nur nur im Traum gedacht.

Unterscheiden sich die mittleren Körpermassen der drei Pinguinarten?

10.8.2 Alle Mittelwerte sind gleich, exakt gleich (?)

- Formal: $\mu_1 = \mu_2 = \dots = \mu_k$ mit k verschiedenen Gruppen von Pinguinarten.
- Hypothesen, die keinen (Null) Unterschied zwischen Gruppen oder keinen Zusammenhang zwischen Variablen postulieren, kann man als *Nullhypothesen* bezeichnen.
- Moment. Dass sich *alle* Mittelwerte um 0,00000000 unterscheiden, ist wohl nicht zu vermuten. Wer glaubt sowas? Daher ist die bessere Forschungsfrage:

Wie sehr unterscheiden sich mittlere Körpergewichte in Abhängigkeit von der Pinguinart?

Alternativ können wir die Hypothese prüfen, ob die Mittelwerte “praktisch” gleich sind, also sich “kaum” unterscheiden. Der Grenzwert für “praktisch gleich” bzw. “kaum unterschiedlich” ist subjektiv. Dazu in Kapitel 10.11 mehr.

10.8.3 Erster Blick in den Datensatz penguins

Datenquelle, Beschreibung des Datensatzes

Hier ist die Quelle der Daten:

```
penguins_url <- "https://vincentarelbundock.github.io/Rdatasets/csv/palmerpenguins/peng  
#| results: "hide"  
#| message: false  
penguins <-  
  read_csv(penguins_url)
```

Hier ist die Verteilung des Gewichts jeder Spezies im Datensatz:

```
penguins %>%  
  select(body_mass_g, species) %>%  
  group_by(species) %>%  
  describe_distribution(range = FALSE, iqr = FALSE)  
## # species=Adelie  
##  
## Variable / Mean / SD / Skewness / Kurtosis / n / n_Missing  
## -----  
## body_mass_g / 3700.66 / 458.57 / 0.29 / -0.57 / 151 / 1  
##  
## # species=Chinstrap  
##  
## Variable / Mean / SD / Skewness / Kurtosis / n / n_Missing  
## -----
```

```

## body_mass_g / 3733.09 / 384.34 /      0.25 /      0.59 / 68 /      0
##
## # species=Gentoo
##
## Variable      /      Mean /      SD / Skewness / Kurtosis /   n / n_Missing
## -----
## body_mass_g / 5076.02 / 504.12 /      0.07 /     -0.72 / 123 /      1

```

Was fällt Ihnen auf?

10.8.4 Visualisierung (EDA)

Hier kommen die Pinguine! Wie schwer sind die Tiere in unserer Stichprobe, s. Abbildung 10.9, Abbildung 10.10, Abbildung 10.11 ?

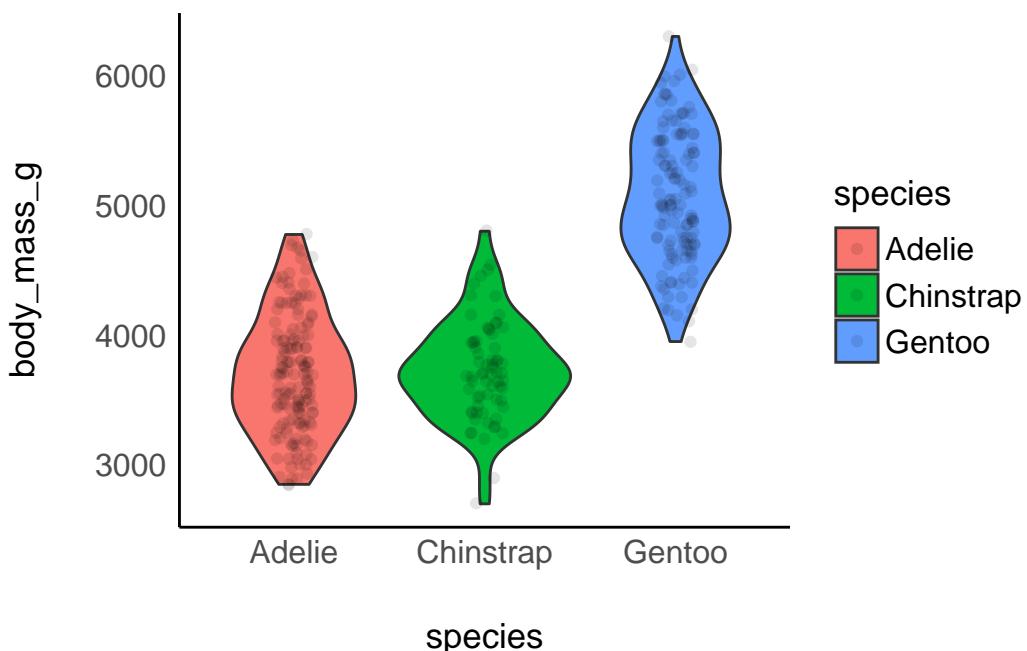


Abbildung 10.9: Verteilung des Körpergewichts dreier Arten von Pinguinen - Geom Violine

Abbildung 10.10 zeigt die Gewichtsverteilung pro Spezies als “Bergrücken” (geom_ridges).

Abbildung 10.11 zeigt die Gewichtsverteilung pro Spezies als “halbe Violinen” (geom_ridges), sozusagen Dichtediagramme um 90 Grad gedreht.

```

penguins %>%
  ggplot(aes(x = species, y = body_mass_g, fill = species)) +
  geom_violindot(fill_dots = "black")

```

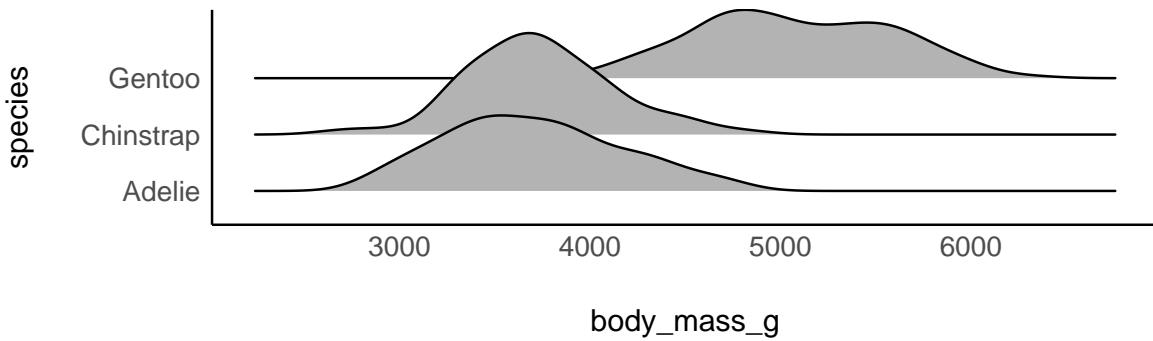


Abbildung 10.10: Verteilung des Körpergewichts dreier Arten von Pinguinen - Geom Ridges

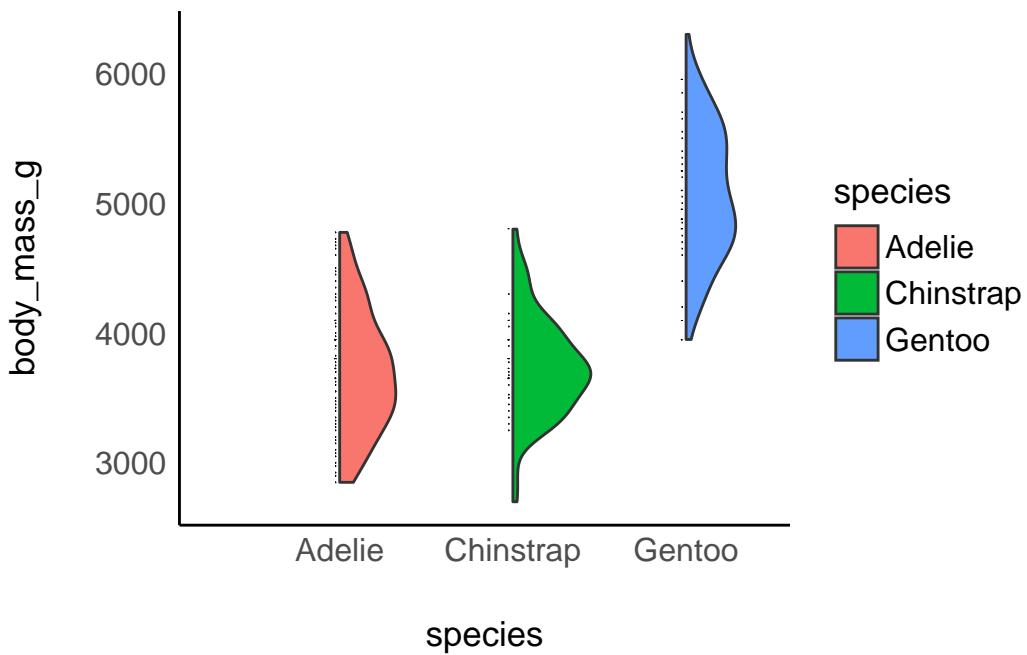


Abbildung 10.11: Verteilung des Körpergewichts dreier Arten von Pinguinen - Geom Violin-dot

10.8.5 Mittlere Gewichtsunterschiede in der Population

Berechnen wir das mittlere Gewicht pro Spezies (Gruppe) der Pinguine, s. m10.6 und Tabelle 10.9.

```
options(mc.cores = parallel::detectCores()) # Turbo einschalten

m10.6 <- stan_glm(body_mass_g ~ species,
                    data = penguins,
                    refresh = 0, # unterdrückt Ausgabe der Posteriori-Stichproben
```

```
seed = 42)
```

```
m10.6 %>%  
parameters()
```

Tabelle 10.9: Parameter des Modells m10.6; neben dem Achsenabschnitt sind die Effekte der Gruppe Adelie und Chinstrap ausgewiesen

Parameter	Median	95% CI	pd	Rhat	ESS	Prior
(Intercept)	3700.35	(3629.70, 3774.26)	100%	1.000	4156.00	Normal (4201.75 +- 2004.89)
speciesChinstrap	33.23	(-101.11, 164.89)	68.08%	1.000	4139.00	Normal (0.00 +- 5015.92)
speciesGentoo	1374.51	(1265.81, 1486.61)	100%	1.000	4201.00	Normal (0.00 +- 4171.63)

10.8.6 Interpretation von m10.6

Die UV hat drei verschiedene Stufen (Werte, Ausprägungen; hier: Spezies), aber es werden in Tabelle 10.9 nur zwei Stufen angezeigt (also eine weniger) zusätzlich zum Achsenabschnitt. Die fehlende Stufe (**Adelie**, nicht ausgegeben) ist die *Vergleichs- oder Referenzkategorie* (baseline) und ist im Achsenabschnitt ausgedrückt (Intercept). Die Koeffizienten für **species** geben jeweils den (vorhergesagten) Unterschied zur Vergleichskategorie wieder. Pinguine der Spezies **Adelie** haben laut Modell ein mittleres Gewicht von ca. 3700g. Pinguine der Spezies Gentoo sind laut Modell im Mittel gut 1000g schwerer als Pinguine der Spezies **Adelie**, etc.

Der Unterschied im mittleren Gewicht von den Gruppen Chinstrap und Gentoo zur Referenzgruppe (Adelie) ist in Abbildung 10.12 verdeutlicht.

```
plot(hdi(m10.6))
```

Highest Density Interval (HDI)

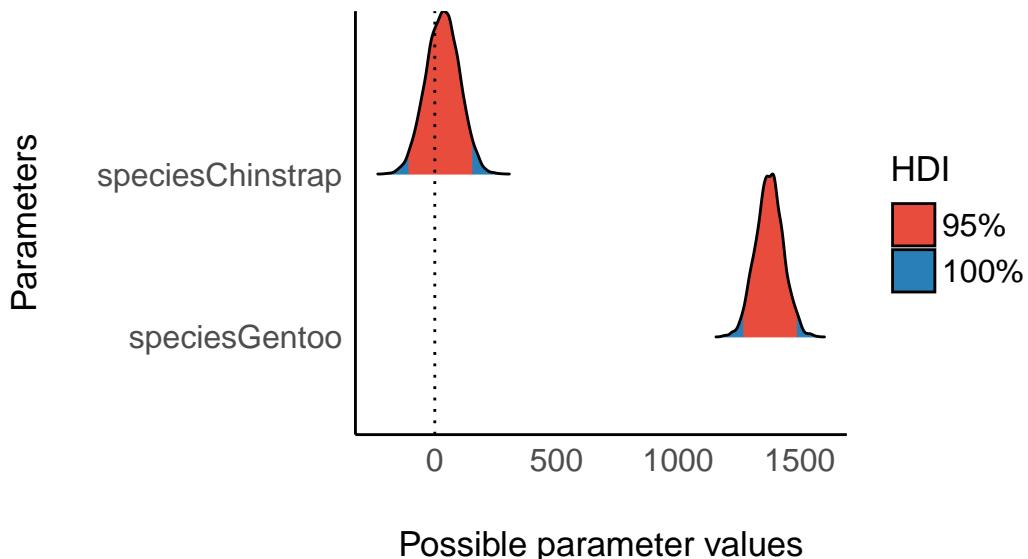


Abbildung 10.12: Effekt der UV: Unterschiede zur Referenzgruppe (95%-HDI)

10.8.7 Glauben wir jetzt an Gruppeneffekte?

Glauben wir jetzt, auf Basis der Modellparameter, an Unterschiede (hinsichtlich der AV) zwischen den Gruppen (UV)?

Es scheinen sich nicht alle Gruppen voneinander zu unterscheiden. So ist der Mittelwert der Gruppe **Gentoo** deutlich höher als der der beiden anderen Gruppen. Umgekehrt sind sich die Pinguinarten **Adelie** und **Chinstrap** in ihren Mittelwerten ziemlich ähnlich.

Wie in Abbildung 10.12 ersichtlich, überlappt sich der Schätzbereich für den Parameter von Gentoo *nicht* mit der Null; hingegen überlappt sich der Schätzbereich des Parameters für Chinstrap deutlich mit der Nulllinie.

Auf Basis unseres Modells schließen wir also (mit hoher Sicherheit) aus, dass *alle* Mittelwerte *exakt* identisch sind.

Ehrlicherweise hätte sowieso (fast) niemand geglaubt, dass die *exakte Nullhypothese* $\mu_1 = \mu_2 = \dots = \mu_k$ bis in die letzte Dezimale gilt. Anders gesagt: Die Wahrscheinlichkeit eines bestimmten Wertes einer stetigen Zufallsvariable ist praktisch Null. Aber: Viele Forschis prüfen gerne die Nullhypothese, daher diskutieren wir den Begriff der (exakten) Nullhypothese. Das Verfahren der Frequentistischen Statistik, um die Nullhypothese $\mu_1 = \mu_2 = \dots = \mu_k$ zu testen, nennt man *Varianzanalyse* (analysis of variance, kurz ANOVA). In der Bayes-Statistik nutzt man - wie immer - primär die Post-Verteilung, um Fragen der Inferenz (z.B. Gruppenunterschiede dieser Art) inferenzstatistisch zu beurteilen.

10.9 Priori-Werte

Unser Modell `m10.6` hat schwach informative (weakly informative) Priors. Für Achsenabschnitt und die Regressionskoeffizienten trifft unser Golem Stan folgende Annahmen in der Voreinstellung:

- Achsenabschnitt und Regressionsgewichte werden als normalverteilt angenommen
- mit Mittelwert entsprechend den Stichprobendaten
- und einer Streuung des Mittelwerts, die der 2.5-fachen der Streuung in der Stichprobe entspricht
- für Sigma wird eine Exponentialverteilung mit Rate $\lambda = 1$ angenommen, skaliert mit der Streuung der AV.

Mehr Infos kann man sich so ausgeben lassen: `prior_summary(modell)`:

```
prior_summary(m10.6)
## Priors for model 'm10.6'
## -----
## Intercept (after predictors centered)
##   Specified prior:
##     ~ normal(location = 4202, scale = 2.5)
##   Adjusted prior:
##     ~ normal(location = 4202, scale = 2005)
##
## Coefficients
##   Specified prior:
##     ~ normal(location = [0,0], scale = [2.5,2.5])
##   Adjusted prior:
##     ~ normal(location = [0,0], scale = [5015.92,4171.63])
##
## Auxiliary (sigma)
##   Specified prior:
##     ~ exponential(rate = 1)
##   Adjusted prior:
##     ~ exponential(rate = 0.0012)
## -----
## See help('prior_summary.stanreg') for more details
```

Wo man man über mehr inhaltliches Wissen verfügt, so wird man die Prioris anpassen wollen, z.B.:

```
m10.6b <- stan_glm(body_mass_g ~ species,
                      data = penguins,
                      refresh = 0,
                      seed = 42,
```

```

prior = normal(location = c(0, 0), # betas, Mittelwert
               scale = c(500, 500)), # betas, Streuung
prior_intercept = normal(3000, 500), # Achsenabschnitt, Mittelwert
prior_aux = exponential(0.001)
)
coef(m10.6b)
##      (Intercept) speciesChinstrap     speciesGentoo
##      3704.89008          22.93895        1356.65099

```

Anstelle von Rohwerten (hier Angabe von Gramm Gewicht) kann man die Streuung auch in z-Werten eingeben, das macht es etwas einfacher. Dazu gibt man bei dem oder den entsprechenden Parametern den Zusatz `autoscale = TRUE` an.

```

m10.6c <- stan_glm(body_mass_g ~ species,
                     data = penguins,
                     refresh = 0,
                     seed = 42,
                     prior = normal(location = c(0, 0), # betas, Mittelwert
                                    scale = c(2.5, 2.5), # betas, Streuung
                                    autoscale = TRUE), # in z-Einheiten
                     prior_intercept = normal(4200, 2.5, # Achsenabschnitt, Mittelwert
                                              autoscale = TRUE),
                     prior_aux = exponential(1, autoscale = TRUE)
)
coef(m10.6c)
##      (Intercept) speciesChinstrap     speciesGentoo
##      3700.99471          32.08176        1374.15711

```

Den Parameter für die Streuung des Modells, σ , kann man sich mit `sigma(modell)` ausgeben lassen:

```

sigma(m10.6b)
## [1] 463.7358

```

Implizit bekommt man die Informationen zu σ mitgeteilt durch die Größe der Konfidenzintervalle.

Übrigens macht es meistens keinen Sinn, extrem weite Prioris zu definieren³.

10.9.1 Wechsel der Referenzkategorie

- `species` ist eine nominale Variable, da passt in R der Typ `factor` (Faktor) am besten. Aktuell ist der Typ noch `character` (Text):

³s. Details [hier](#).

```
penguins <- penguins %>%
  mutate(species = factor(species))
```

Im Standard sortiert R die Faktorstufen alphabetisch, aber man kann die Reihenfolge ändern.

```
levels(penguins$species)
## [1] "Adelie"     "Chinstrap"   "Gentoo"
```

Setzen wir `Gentoo` als Referenzkategorie und lassen die restliche Reihenfolge, wie sie ist:

```
library(forcats)
penguins <- penguins %>%
  mutate(species = factor(species),
        species = fct_relevel(species, "Gentoo"))
```

Beachten Sie, dass dazu das Paket `forcats` verfügbar sein muss.

Jetzt haben wir die Referenzkategorie geändert:

```
levels(penguins$species)
## [1] "Gentoo"      "Adelie"      "Chinstrap"
```

Der Wechsel der Referenzkategorie ändert nichts Wesentliches am Modell:

```
m10.6a <- stan_glm(body_mass_g ~ species, data = penguins, refresh = 0)
hdi(m10.6a)
## Highest Density Interval
##
## Parameter          /          95% HDI
## -----
## (Intercept)       / [ 4999.73,  5158.39]
## speciesAdelie    / [-1480.62, -1265.17]
## speciesChinstrap / [-1472.25, -1194.63]
```

10.10 Modellgüte mit R-Quadrat bestimmen

10.10.1 Modellgüte mit R^2 bestimmen

R^2 gibt den Anteil der Gesamtvarianz (der AV) an, den das Modell erklärt. - Höhere Wert von R^2 bedeuten, dass das Modell die Daten besser erklärt. R^2 wird normalerweise auf Basis eines Punktschätzers definiert. Solch eine Definition lässt aber viel Information - über die Ungewissheit der Schätzung - außen vor. Daher ist es wünschenswert, diese Information in R^2 einfließen zu lassen: *Bayes-R-Quadrat*.

```
r2(m10.6)
## # Bayesian R2 with Compatibility Interval
##
## Conditional R2: 0.667 (95% CI [0.619, 0.712])
```

Möchte man es ausführlicher, und im Komfort einer Bayes-Analyse schwelgen, so kann man sich die Posteriori-Verteilung von R^2 ausgeben lassen, s. Abbildung 10.13.

```
m10.6_r2 <-
m10.6 %>%
  r2_posterior() %>%
  as_tibble()

hdi(m10.6_r2) %>%
  plot()
```

Highest Density Interval (HDI)

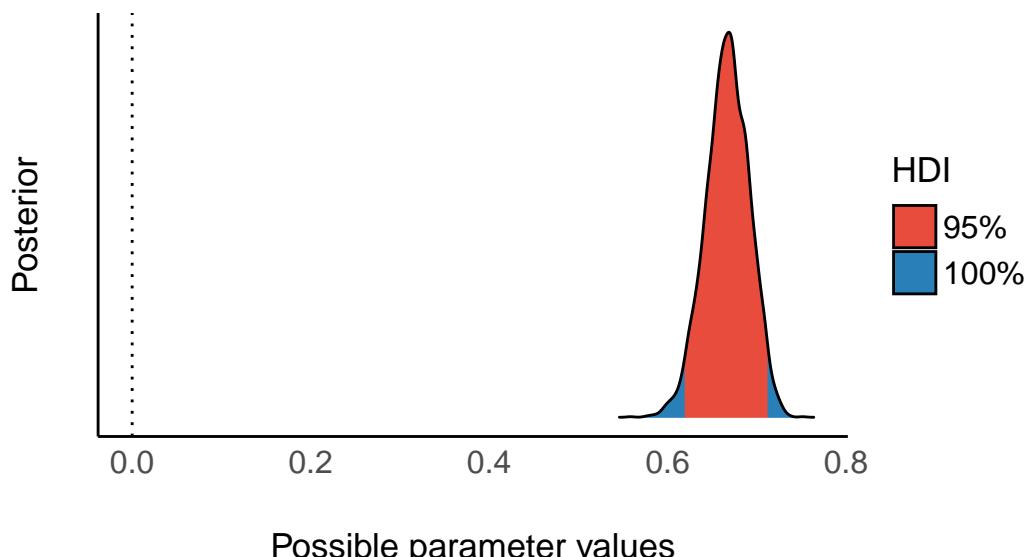


Abbildung 10.13: Die Verteilung von R-Quadrat im Modell m10.6

10.10.2 Definition vom “klassischen” R^2

Wie genau sind die Vorhersagen des Modells? σ (Vorhersagefehler) quantifiziert die Streuung der Residuen $r_i = y_i - X_i \hat{\beta}$, mit $\hat{y}_i = X_i \hat{\beta}$. Anders gesagt: $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots = X \hat{\beta}$. Anders gesagt gibt σ die “typische” Abweichung einer Beobachtung vom vorhergesagten Wert an. Es ist nützlich, σ in Bezug zu setzen zur Streuung der AV, $sd_y = s_y$: $R^2 = 1 - (\hat{\sigma}^2 / s_y^2)$. R^2 gibt damit den Anteil der vom Modell erklärten Varianz, V , an. Berechnet

man das Modell mit der Methode der kleinsten Quadrate (nicht mit Bayes), dann ist der obige Ausdruck äquivalent zu: $R^2 = V_{i=1}^n \hat{y}_i / s_y^2$. Die beiden obigen Ausdrücke nehmen \hat{y}_i als fix (sicher) an und vernachlässigen Ungewissheit; sie sind übergewiss aus Bayes-Sicht.

10.10.3 Bayes' R^2

Besser ist es (aus Bayes-Sicht), die Ungewissheit der Vorhersagen bei der Berechnung der Modellgüte miteinzubeziehen: Bayes $R^2 = \frac{\text{erklärte Varianz}}{\text{Erklärte Varianz} + \text{Residualvarianz}} = \frac{V_{mod}}{V_{mod} + V_{res}}$.

V_{mod} ist die Varianz in der PPV mit $s = 1, \dots, S$ simulierten Stichproben, $V(\hat{y}_i)$ und V_{res} ist die Residualvarianz im Modell. Für jede Stichprobe s berechnet man die vorhergesagten Werte, \hat{y}_i^s , die Residualvarianz σ_s^2 und den Anteil der erklärten Varianz: Bayes $R_s^2 = \frac{V(\hat{y}_i^s)}{V(\hat{y}_i^s + \sigma_s^2)}$, vgl. Gelman u. a. (2019), Gelman, Hill, und Vehtari (2021), Kap. 11.7.

10.11 Nullhypotesen sind praktisch immer falsch

Nullhypotesen sind fast immer falsch, s. Abbildung 10.14.



Abbildung 10.14: Du testest Nullhypotesen?

Quelle: [Imgflip Meme Generator](#)

We do not generally use null hypothesis significance testing in our own work. In the fields in which we work, we do not generally think null hypotheses can be true: in social science and public health, just about every treatment one might consider will have *some* effect, and no comparison or regression coefficient of interest will be exactly zero. We do not find it particularly helpful to formulate and test null hypotheses that we know ahead of time cannot be true.

Gelman, Hill, und Vehtari (2021)

10.11.1 Alternativen zu Nullhypotesen

Nullhypotesen, H_0 , sind z.B.: $\rho = 0$, $\rho_1 = \rho_{\theta_2}$, $\mu_1 = \mu_2$, $\mu = 0$, $\beta_1 = 0$. Nullhypotesen zu testen, ist sehr verbreitet. Ein Grund ist, dass in der Frequentistischen Statistik keine andere Art von Hypothesentest möglich ist⁴

Ein anderer Grund ist vermutlich, ... wir haben es schon immer so gemacht.

Alternativen zum Testen von Nullhypotesen sind:

- Posteriori-Intervalle (PI oder HDI) berichten
- Rope-Konzept, @kruschke_rejecting_2018
- Wahrscheinlichkeit von inhaltlich bedeutsamen Hypothesen quantifizieren.
- Wahrscheinlichkeit quantifizieren, dass der Effekt ein positives bzw. ein negatives Vorzeichen hat.

10.11.2 “Praktisch” kein Unterschied: Das Rope-Konzept

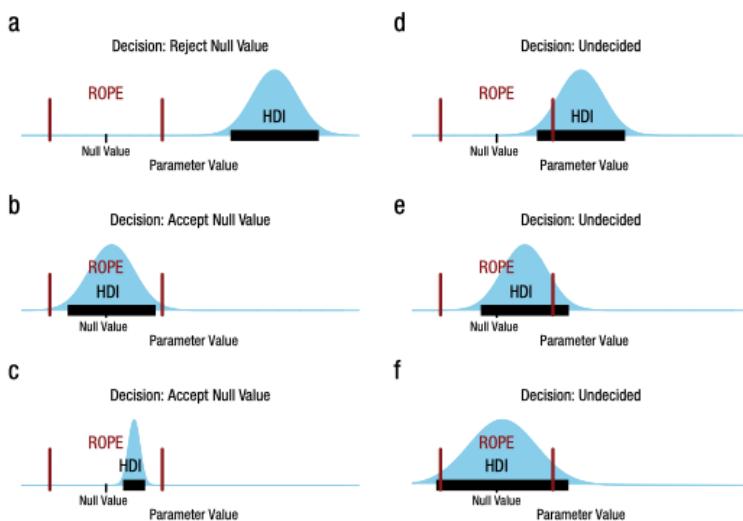
Sagen wir, wenn sich zwei Preismittelwerte um höchstens $d = 100\text{€}$ unterscheiden, gilt dieser Unterschied für uns als “praktisch gleich”, “praktisch kein Unterschied” bzw. vernachlässigbar. Nimmt man (praktisch) keinen Unterschied/Zusammenhang/Effekt an, spricht man von einer *Nullhypothese*: H_0 . Die Wahl von d ist *subjektiv* in dem Sinne als sie von inhaltlichen Überlegungen geleitet sein sollte. Diesen Bereich bezeichnen wir den *Indifferenzbereich* (Äquivalenzzone, Bereich eines vernachlässigbaren Unterschieds oder *Region of practical equivalence*, Rope). Jetzt prüfen wir, ob ein “Großteil” der Posteriori-Stichproben im Rope liegt. Unter “Großteil” wird häufig das 95%-HDI verstanden (das ist auch der Standard der R-Funktion `rope()`, die wir hier nutzen).

Entscheidungsregel nach Kruschke (2018):

- Großteil liegt *innerhalb* von Rope Annahme der Nullhypothese “praktisch kein Effekt”, H_0
- Großteil liegt *außerhalb* von Rope Ablehnung der Nullhypothese “praktisch kein Effekt”, H_0
- Ansonsten keine Entscheidung

⁴Mittlerweile gibt es Ansätze für ein Verfahren ähnlich dem ROPE-Ansatz, der weiter unten vorgestellt wird.

10.11.3 HDI-Rope-Entscheidungsregel visualisiert



10.11.4 Rope berechnen

Den Rope berechnet man mit `rope(model)`.

```
rope(m10.6)
## # Proportion of samples inside the ROPE [-80.20, 80.20]:
##
## Parameter      / inside ROPE
## -----
## (Intercept)    /      0.00 %
## speciesChinstrap / 74.34 %
## speciesGentoo   /      0.00 %
```

Die Faktorstufe `Chinstrap` von `species` hat doch einen beträchtlichen Teil ihrer Wahrscheinlichkeitsmasse der Posteriori-Verteilung im ROPE.

Wir können daher für diese Gruppe das ROPE *nicht* verwerfen.

Aber: `Gentoo` liegt zu 0% im Rope. Für Gentoo können wir das Rope verwerfen.

Das hört sich abstrakt an? Dann lassen Sie uns das lieber visualisieren.

i Hinweis

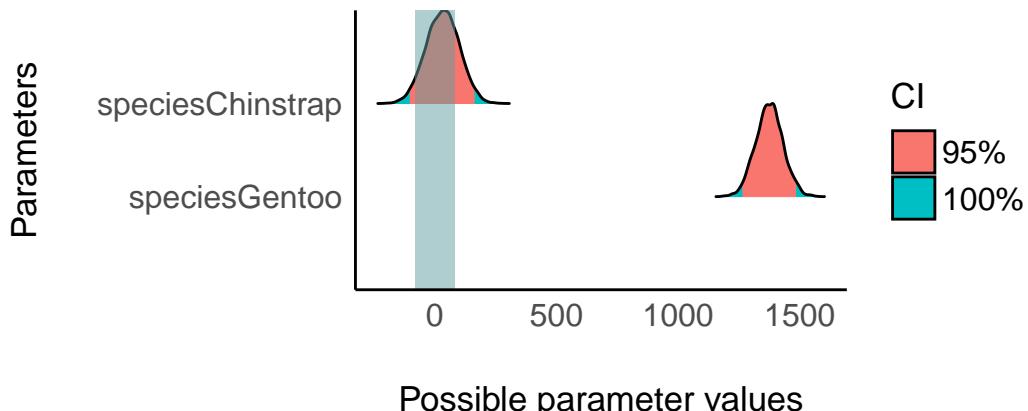
Die angegebenen Prozentwerte beziehen sich nicht auf die 100% der Post-Verteilung, sondern (in der Voreinstellung) auf das 95%-ETI, s. `help(rope)`.

10.11.5 Visualisierung unserer Rope-Werte, m10.6

- Ein Großteil der Posteriori-Masse von m10.6 liegt *nicht* innerhalb des Rope.
- Aber können wir umgekehrt sagen, dass ein Großteil außerhalb liegt? Das erkennt man optisch ganz gut.

```
rope(m10.6) %>% plot()
```

Region of Practical Equivalence (ROPE)



Das ROPE drückt die “Berge” der Posteriori-Verteilung für Chinstrap deutlich. Aber: Das 95%-HDI liegt nicht komplett innerhalb des Rope. Wir können das Rope für Chinstrap *nicht verwerfen*, aber auch *nicht bestätigen*.

Gentoo hingegen wird vom Rope nicht durchkreuzt, es ist weit entfernt vom “blauen Fluss” des Rope: Gentoo liegt außerhalb des Rope. Es gibt einen “substanziellen” Unterschied, größer als das ROPE. Wir verwerfen die “Praktisch-Null-Hypothese” in diesem Fall.

10.11.6 Finetuning des Rope

Wir können festlegen, was wir unter “praktischer Äquivalenz” verstehen, also die Grenzen des Ropes verändern. Sagen wir, 100 Gramm sind unsere Grenze für einen vernachlässigbaren Effekt, s. Abbildung 10.15.

```
rope(m10.6, range = c(-100, 100))  
plot(rope(m10.6, range = c(-100, 100)))
```

Region of Practical Equivalence (ROPE)

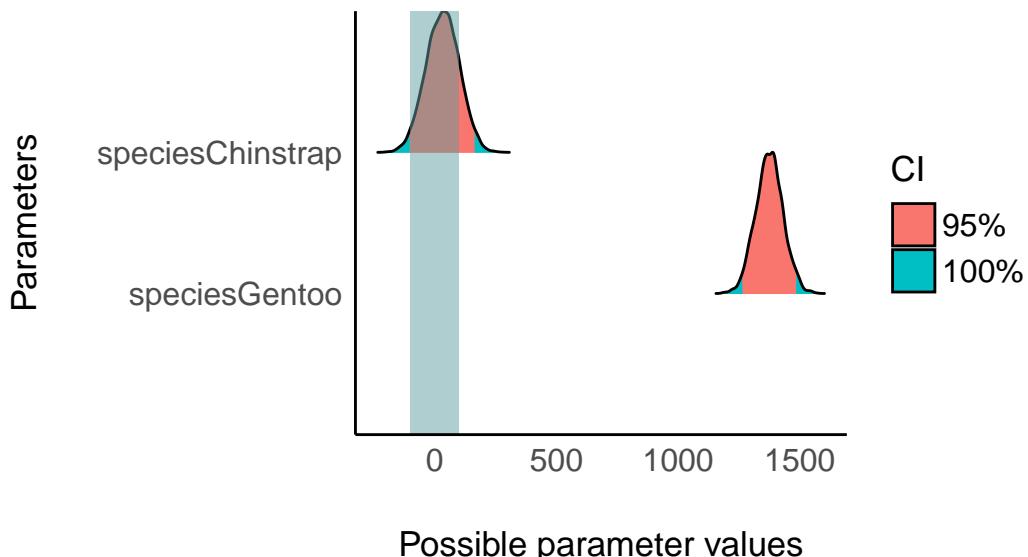


Abbildung 10.15: ROPE mit selber eingestellter Grenze von ± 100 (Gramm)

Im Standard werden 95%-HDI berichtet, das kann man so ändern, wenn man möchte:

```
rope(m10.6, range = c(-100,100), ci = .89, ci_method = "ETI")
```

ETI (equal tails interval) steht für ein PI. Jetzt wird berichtet, welcher Teil eines 89%-CI sich im Rope befindet.

10.11.7 Beantwortung der Forschungsfrage

Für die Spezeis *Gentoo* wurde ein substanzialer Gewichtsunterschied zur Referenzgruppe, Adelie, vom Modell entdeckt. Für *Chinstrap* hingegen ist keine klare inferenzstatistische Aussage hinsichtlich eines Indifferenzbereichs möglich: Es ist plausibel, laut dem Modell, dass es einen praktisch bedeutsamen Unterschied gibt, aber es ist auch plausibel, dass es keinen praktisch bedeutsamen Unterschied gibt.

10.12 Mehrere metrische UV

10.12.1 Forschungsfrage

Stehen sowohl der IQ der Mutter als auch, unabhängig davon, das Alter der Mutter im Zusammenhang mit dem IQ des Kindes?

- Das ist wieder eine *deskriptive* Forschungsfrage. *Keine* Kausalwirkung (etwa “IQ der Mutter ist die Ursache zum IQ des Kindes”) wird impliziert.
- Es geht rein darum, Zusammenhänge in den Daten - bzw. in der Population - aufzuzeigen.
- Viele Forschungsfragen gehen allerdings weiter und haben explizit Kausalwirkungen im Fokus. Für solche Fragen ist eine deskriptive Untersuchung nicht geeignet, sondern eine Kausalanalyse ist nötig.

[Datenquelle](#) als CSV-Datei oder alternativ:

```
library(rstanarm)
data("kidiq")
```

10.12.2 Was heißt, X hängt mit Y zusammen?

- Der Begriff “Zusammenhang” ist nicht exakt.
- Häufig wird er (für metrische Variablen) verstanden als
 - lineare Korrelation ρ bzw. r
 - lineare Regression β , bzw. b
- Der Regressionskoeffizient
 - misst die *Steigung* der Regressionsgerade
 - zeigt, wie groß der vorhergesagte Unterschied in Y, wenn man zwei Personen (Beobachtungseinheiten) vergleicht, die sich um eine Einheit in X unterscheiden
 - wird manchmal mit dem “Effekt von X auf Y” übersetzt. Vorsicht: “Effekt” klingt nach Kausalzusammenhang. Eine Regression ist keine hinreichende Begründung für einen Kausalzusammenhang.
- Der Korrelationskoeffizient
 - misst eine Art der Stärke des linearen Zusammenhangs
 - zeigt, wie klein die Vorhersagefehler der zugehörigen Regrsson im Schnitt sind.
 - [Korrelation ist nicht \(automatisch\) Kausation.](#)

10.12.3 Korrelationen zur Forschungsfrage

```
kidiq %>%
correlation()
```

Tabelle 10.10: Correlation Matrix (pearson-method)

Parameter1	Parameter2	r	95% CI	t(432)	p
kid_score	mom_hs	0.24	(0.15, 0.32)	5.07	< .001***

Parameter1	Parameter2	r	95% CI	t(432)	p
kid_score	mom_iq	0.45	(0.37, 0.52)	10.42	< .001***
kid_score	mom_age	0.09	(-2.15e-03, 0.18)	1.92	0.111
mom_hs	mom_iq	0.28	(0.19, 0.37)	6.13	< .001***
mom_hs	mom_age	0.21	(0.12, 0.30)	4.57	< .001***
mom_iq	mom_age	0.09	(-2.54e-03, 0.18)	1.91	0.111

p-value adjustment method: Holm (1979) Observations: 434

Oder als Korrelationsmatrix:

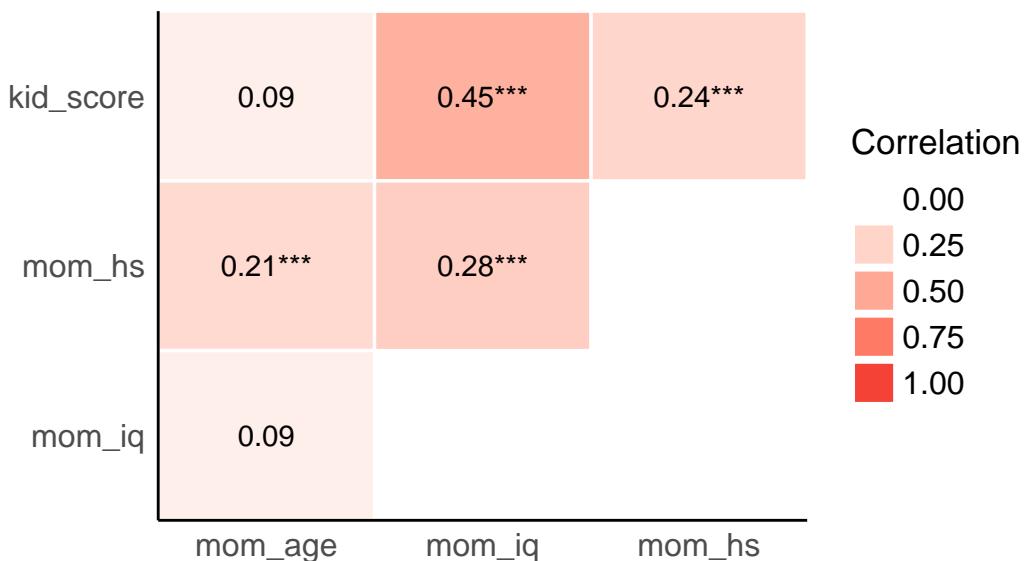
```

kidiq %>%
  correlation() %>%
  summary()
## # Correlation Matrix (pearson-method)
##
## Parameter / mom_age / mom_iq / mom_hs
## -----
## kid_score / 0.09 / 0.45*** / 0.24***
## mom_hs    / 0.21*** / 0.28*** /
## mom_iq    / 0.09 /      /
##
## p-value adjustment method: Holm (1979)

kidiq %>%
  correlation() %>%
  summary() %>%
  plot()

```

Correlation Matrix



10.12.4 Univariate Regressionen

Wir berechnen jeweils eine univariate Regression, pro Prädiktor, also eine für `mom_iq` und eine für `mom_age`.

```
m10.7 <- stan_glm(kid_score ~ mom_iq, data = kidiq, refresh = 0)
m10.8 <- stan_glm(kid_score ~ mom_age, data = kidiq, refresh = 0)
```

Hier die Ergebnisse für `mom_iq`:

```
coef(m10.7)
## (Intercept)      mom_iq
## 25.6540768    0.6117116
```

Hier die Ergebnisse für `mom_age`:

```
coef(m10.8)
## (Intercept)      mom_age
## 71.2317227    0.6783984
```

10.12.5 Visualisierung der univariaten Regressionen

In Abbildung 10.16 ist die univariate Regression mit jeweils einem der beiden Prädiktoren dargestellt.

m10.7: Die Steigung beträgt 0.6. m10.8: Die Steigung beträgt 0.7.

```
p1 <-  
  kidiq %>%  
  ggplot(aes(x = mom_iq, y = kid_score)) +  
  geom_point() +  
  geom_abline(intercept = coef(m10.7)[1],  
              slope = coef(m10.7)[2],  
              color = "blue")  
  
p2 <-  
  kidiq %>%  
  ggplot(aes(x = mom_age, y = kid_score)) +  
  geom_point() +  
  geom_abline(intercept = coef(m10.8)[1],  
              slope = coef(m10.8)[2],  
              color = "blue")  
  
plots(p1, p2,  
      title = c("m10.7: Die univariate Regression mit dem Alter der Mutter als Prädiktor",  
               "m10.8: Die univariate Regression mit dem IQ der Mutter als Prädiktor"))
```

m10.7: Die univariate Regression mit dem Alter der Mutter

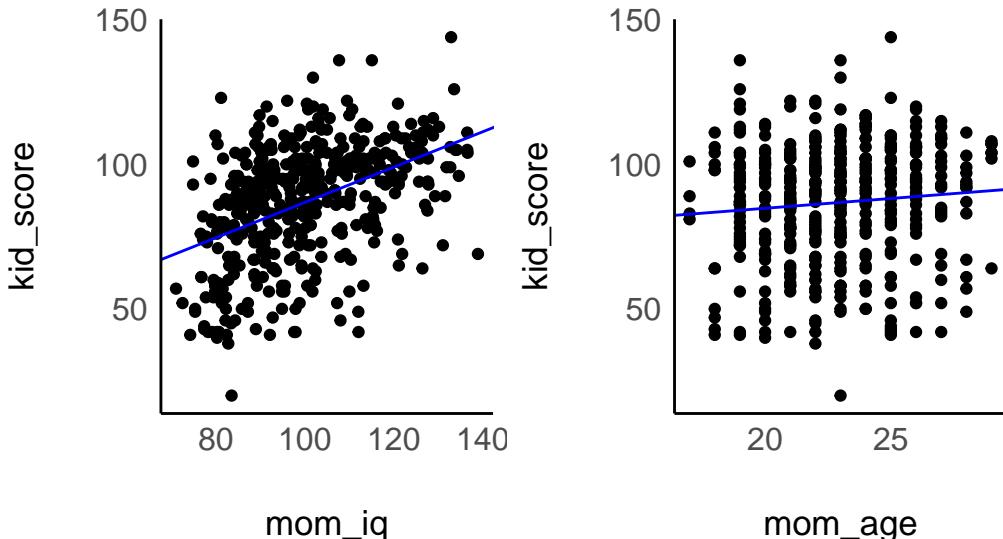


Abbildung 10.16: Zwei univariate Regressionen

Univariate Regressionen

10.12.6 Multiples Modell (beide Prädiktoren), m10.9

m10.9 stellt das multiple Regressionsmodell dar; *multipel* bedeutet in diesem Fall, dass mehr als ein Prädiktor im Modell aufgenommen ist.

```
m10.9 <- stan_glm(kid_score ~ mom_iq + mom_age,  
                     data = kidiq,  
                     refresh = 0)  
  
coef(m10.9)  
## (Intercept)      mom_iq      mom_age  
## 17.4878691    0.6024614   0.3904761
```

! Wichtig

Die Regressionsgewichte unterscheiden sich zu den von den jeweiligen univariaten Regressionen.

- Bei einer multiplen Regression ist ein Regressionsgewicht jeweils “bereinigt” vom Zusammenhang mit dem (oder den) anderen Regressionsgewicht.
- Das bedeutet, man betrachtet den Zusammenhang eines Prädiktors mit der AV, wobei man gleichzeitig den anderen Prädiktor konstant hält.

```
coef(m10.9)  
## (Intercept)      mom_iq      mom_age  
## 17.4878691    0.6024614   0.3904761
```

10.12.7 3D-Visualisierung eines Modells mit zwei Prädiktoren 1

In Abbildung 10.17 ist das Modell m10.9 in 3D dargestellt via [Plotly](#).

10.12.8 Visualisierung mit Farbe statt 3. Dimension

3D-Visualisierungen haben Vorteile, aber auch Nachteile; Abbildung 10.18 zeigt eine alternative Visualisierung, in der die 3. Dimension durch eine Farbschattierung ersetzt ist.

Auf der Achse von mom_iq erkennt man deutlich (anhand der Farbänderung) die Veränderung für die AV (kid_score). Auf der Achse für mom_age sieht man, dass sich die AV kaum ändert, wenn sich mom_age ändert.

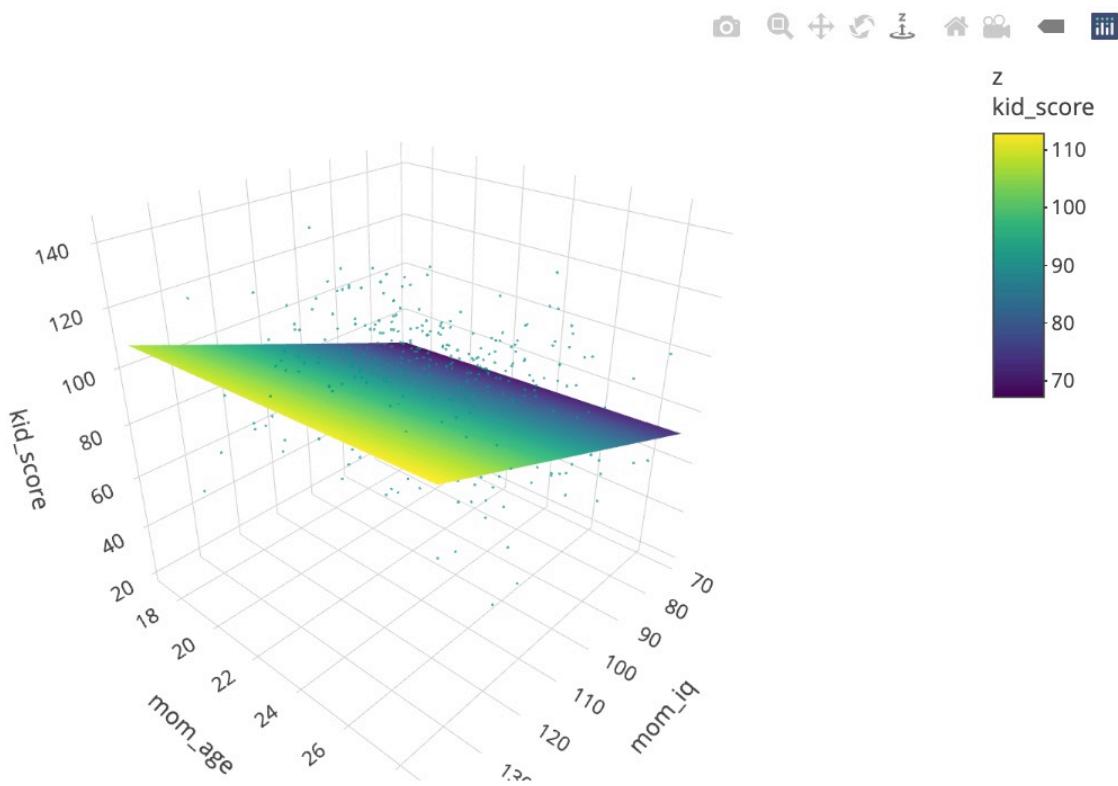


Abbildung 10.17: 3D-Visualisierung von m10.9 (zwei Prädiktoren)

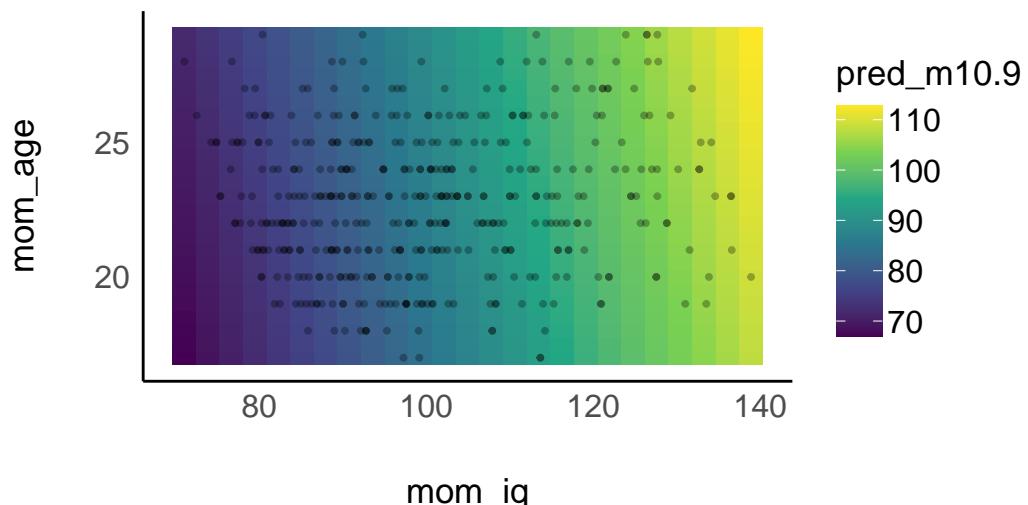


Abbildung 10.18: Modell m10.9; die Farbverläufe zeigen der Wert der abhängigen Variablen



Abbildung 10.19: So sieht der Zusammenhang im 10-dimenionalen Raum aus

10.12.9 Visualisierung in 10 Dimensionen

Abbildung 10.19 visualisiert den Zusammenhang von 10 Variablen untereinander.

Leider macht mein Hirn hier nicht mit. Unsere Schwächen, eine große Zahl an Dimensionen zu visualisieren, ist der Grund, warum wir mathematische Modelle brauchen.

Daher kann man ein Modell verstehen als eine Zusammenfassung eines (ggf. hochdimensionalen) Variablenraums.

10.12.10 Relevanz der Prädiktoren

Woher weiß man, welcher Prädiktor am stärksten mit der AV zusammenhängt? Man könnte auch sagen: Welcher Prädiktor (welche UV) am “wichtigsten” ist oder den “stärksten Einfluss” auf die AV ausübt? Bei solchen kausal konnotierten Ausdrücken muss man vorsichtig sein: Die Regressionsanalyse als solche ist keine Kausalanalyse. Die Regressionsanalyse - wie jede statistische Methoden - kann für sich *nur Muster in den Daten*, also Zusammenhänge bzw. Unterschiede, entdecken, s. Abbildung 10.20.

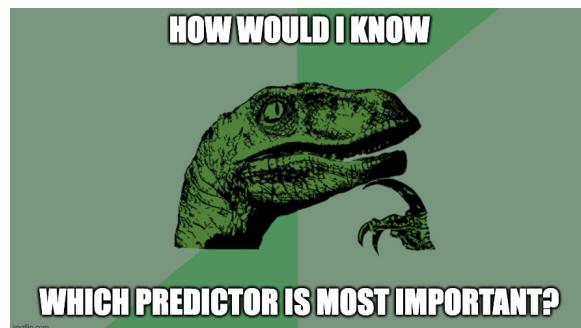


Abbildung 10.20: Made at imgflip.com

Welcher Prädiktor ist nun “wichtiger” oder “stärker” in Bezug auf den Zusammenhang mit der AV, `mom_iq` oder `mom_age` (Modell m10.9)?

- `mom_iq` hat den größeren Koeffizienten.
- `mom_age` hat weniger Streuung.

Um die Relevanz der Prädiktoren vergleichen zu können, müsste man vielleicht die Veränderung von `kid_score` betrachten, wenn man von kleinsten zum größten Prädiktorwert geht. Allerdings sind Extremwerte meist instabil (da sie von einer einzigen Beobachtung bestimmt werden). Sinnvoller ist es daher, die Veränderung in der AV zu betrachten, wenn man den Prädiktor von “unterdurchschnittlich” auf “überdurchschnittlich” ändert. Das kann man mit *z-Standardisierung* erreichen.

10.12.11 z-Standardisierung

z-Standardisierung bedeutet, eine Variable so zu transformieren, dass sie über einen Mittelwert von 0 und eine SD von 1 verfügt:

$$z = \frac{x - \bar{x}}{sd(x)}$$

```
data("kidiq")
kidiq2 <-
  kidiq %>%
  mutate(mom_iq_z = ((mom_iq - mean(mom_iq)) / sd(mom_iq))) %>%
  select(mom_iq, mom_iq_z) %>%
  head()
```

Der Nutzen von Standardisieren (dieser Art) ist die bessere Vergleichbarkeit von Variablen, die (zuvor) verschiedene Mittelwerte und Streuungen hatten⁵. Die Standardisierung ist ähnlich zur Vergabe von Prozenträngen: “Dieser Messwert gehört zu den Top-3-Prozent”. Diese Aussage ist bedeutsam für Variablen mit verschiedenem Mittelwert und Streuung. So werden vergleichende Aussagen für verschiedene Verteilungen möglich.

10.12.12 Statistiken zu den z-transformierten Variablen

Variable	Mean	SD	IQR	Range	Skewness	Kurtosis	n	n_Missing
kid_score	86.80	20.41	28.00	(20.00, 144.00)	-0.46	-0.16	434	0
mom_hs	0.79	0.41	0.00	(0.00, 1.00)	-1.40	-0.05	434	0
mom_iq	100.00	15.00	21.67	(71.04, 138.89)	0.47	-0.57	434	0
mom_age	22.79	2.70	4.00	(17.00, 29.00)	0.18	-0.63	434	0

⁵am nützlichsten ist diese Standardisierung bei normal verteilten Variablen.

Man kann auch die z-Transformation (“Skalierung”) mit `standardize` durchführen:

```
kidiq <-  
  standardize(kidiq, append = TRUE)  
  
head(kidiq)  
##   kid_score mom_hs    mom_iq mom_age kid_score_z mom_hs_z mom_iq_z  
## 1       65     1 121.11753      27 -1.06793237  0.521631  1.4078352  
## 2       98     1  89.36188      25  0.54886757  0.521631 -0.7092079  
## 3       85     1 115.44316      27 -0.08805362  0.521631  1.0295443  
## 4       83     1  99.44964      25 -0.18604150  0.521631 -0.0366907  
## 5      115     1  92.74571      27  1.38176451  0.521631 -0.4836193  
## 6       98     0 107.90184      18  0.54886757 -1.912647  0.5267892  
##   mom_age_z  
## 1  1.5602285  
## 2  0.8197811  
## 3  1.5602285  
## 4  0.8197811  
## 5  1.5602285  
## 6 -1.7717849
```

Der Schalter `append = TRUE` sorgt dafür, dass die ursprünglichen Variablen beim Z-Standardisieren nicht überschrieben werden, sondern angehängt werden (mit einem Suffix `_z`).

Man kann auch nur einzelne Variablen mit `standardize` standardisieren.

```
kidiq %>%  
  standardize(select = c("mom_iq", "mom_age", "kid_score"))
```

Man kann das Standardisieren auch von Hand machen, ohne ein Extra-Paket:

```
#data(kidiq)  
kidiq %>%  
  mutate(mom_iq_z2 = scale(mom_iq),  
         mom_age_z2 = scale(mom_age),  
         kid_score_z2 = scale(kid_score)) %>%  
  head()  
##   kid_score mom_hs    mom_iq mom_age kid_score_z mom_hs_z mom_iq_z  
## 1       65     1 121.11753      27 -1.06793237  0.521631  1.4078352  
## 2       98     1  89.36188      25  0.54886757  0.521631 -0.7092079  
## 3       85     1 115.44316      27 -0.08805362  0.521631  1.0295443  
## 4       83     1  99.44964      25 -0.18604150  0.521631 -0.0366907  
## 5      115     1  92.74571      27  1.38176451  0.521631 -0.4836193  
## 6       98     0 107.90184      18  0.54886757 -1.912647  0.5267892
```

```

##      mom_age_z  mom_iq_z2 mom_age_z2 kid_score_z2
## 1  1.5602285  1.4078352  1.5602285 -1.06793237
## 2  0.8197811 -0.7092079  0.8197811  0.54886757
## 3  1.5602285  1.0295443  1.5602285 -0.08805362
## 4  0.8197811 -0.0366907  0.8197811 -0.18604150
## 5  1.5602285 -0.4836193  1.5602285  1.38176451
## 6 -1.7717849  0.5267892 -1.7717849  0.54886757

```

10.13 Modell m10.10

Im Modell `m10.10` sind die Prädiktoren z-standardisiert. Das Standardisieren der AV, `kid_score` ist *nicht* nötig, um den Effekt der Prädiktoren (UV) auf die AV zu untersuchen. Standardisiert man aber die AV, so liefern die Regressionskoeffizienten (Betas) Aussage darüber, um wie viele *SD*-Einheiten sich die AV verändert, wenn sich ein Prädiktor um eine *SD*-Einheit verändert.

```

m10.10 <- stan_glm(kid_score_z ~ mom_iq_z + mom_age_z,
                     data = kidiq,
                     refresh = 0)
coef(m10.10)
## (Intercept)    mom_iq_z    mom_age_z
## -0.001092363  0.443742017  0.051065031

```

- Der *Achsenabschnitt* gibt den Mittelwert der AV (`kid_score`) an, da `kid_score_z = 0` identisch ist zum Mittelwert von `kid_score`.
- Der Koeffizient für `mom_iq_z` gibt an, um wie viele SD-Einheiten sich `kid_score` (die AV) ändert, wenn sich `mom_iq` um eine SD-Einheit ändert.
- Der Koeffizient für `mom_age_z` gibt an, um wie viele SD-Einheiten sich `kid_score` (die AV) ändert, wenn sich `mom_age` um eine SD-Einheit ändert.

Jetzt sind die Prädiktoren in ihrer Relevanz (Zusammenhang mit der AV) vergleichbar:

- Man sieht, dass die Intelligenz der Mutter *deutlich wichtiger* ist als das Alter der Mutter (im Hinblick auf die Vorhersage bzw. den Zusammenhang mit der AV).

10.13.1 95%-PI

Mit `parameters` können wir uns ein PI für `m10.10` ausgeben lassen, s. Abbildung 10.21; im Standard wird ein ETI berichtet.

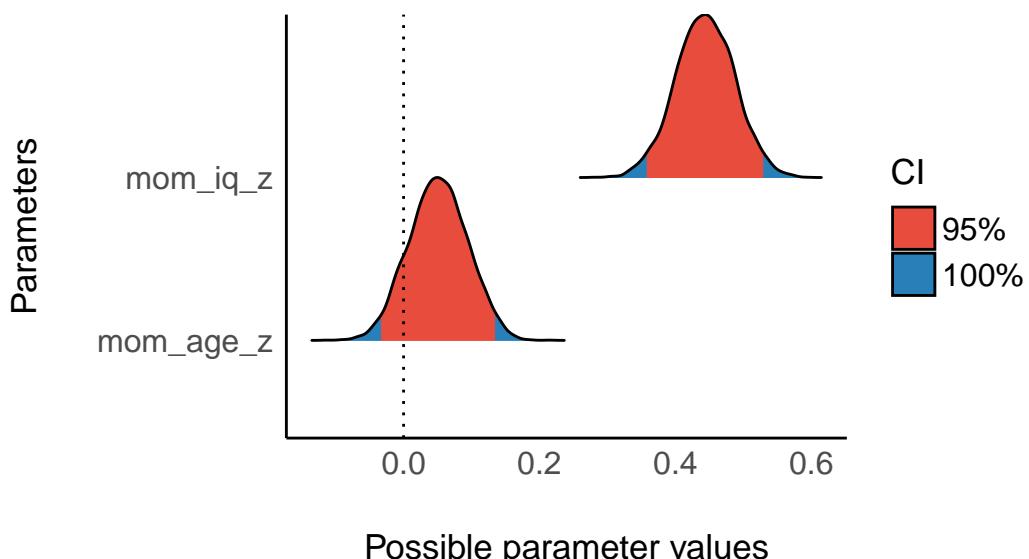
```
parameters(m10.10)
```

Tabelle 10.12: Fixed effects

Parameter	Median	95% CI	pd	Rhat	ESS	Prior
(Intercept)	-1.09e-03	(-0.08, 0.08)	50.95%	0.999	4865.00	Normal (-2.81e-16 +- 2.50)
mom_iq_z	0.44	(0.36, 0.53)	100%	1.000	5118.00	Normal (0.00 +- 2.50)
mom_age_z	0.05	(-0.03, 0.13)	87.83%	1.000	5375.00	Normal (0.00 +- 2.50)

```
plot(eti(m10.10))
```

Credible Interval (CI)



Possible parameter values

Abbildung 10.21: Im Standard wird ein 95%-Intervall gezeigt bzw. berechnet; hier das ETI für m10.10

10.13.2 Was ist ein kleiner, was ein großer Effekt?

Cohen (1988) definiert Effektstärken in Bezug auf Mittelwertsvergleiche anhand von $d = (\mu_1 - \mu_o)/\sigma$. Für kleine, mittlere und große Werte gab er folgende Richtwerte:

- klein: $d \approx 0.2$
- mittel: $d \approx 0.5$
- groß: $d \approx 0.8$

Auf dieser Basis schlägt Kruschke (2018) einen Rope von ± 0.1 vor. Fällt ein Intervall (mit vorab definierter Sicherheit, z.B. 95%) komplett in das Rope, so gilt der Effekt als “praktisch null”. Richtlinien für Effektstärken sind nur Notlösungen, die durch Sachverständigenten ersetzt

werden sollen, wo immer möglich. Man kann Effektstärken ineinander überführen, s. [hier](#), z.B. von Korrelation (r) zu Cohens d oder R^2 .

10.13.3 Vernachlässigbarer Regressionseffekt

Kruschke (2018) schlägt vor, einen Regressionskoeffizienten unter folgenden Umständen als “praktisch Null” zu bezeichnen:

Wenn eine Veränderung über “praktisch den ganzen Wertebereich” von x nur einen vernachlässigbaren Effekt auf y hat. Ein vernachlässigbarer Effekt ist dabei $\hat{y} = \pm 0.1sd_y$. Der “praktisch ganze Wertebereich” von x sei $\bar{x} \pm 2sd_x$. Resultiert der Vergleich von $\bar{x} - 2sd$ mit $\bar{x} + 2sd$ nur eine Veränderung in \hat{y} von $\bar{y} - 0.1sd_y$ auf $\bar{y} + 0.1sd_y$, so ist der Regressionskoeffizient praktisch Null, der Effekt also vernachlässigbar. Das impliziert Rope-Grenzen von $\beta_x = \pm 0.05$ für z-standardisierte Variablen.

10.13.4 Modellgüte

```
r2(m10.10)
## # Bayesian R2 with Compatibility Interval
##
## Conditional R2: 0.204 (95% CI [0.141, 0.267])
```

Ist dieser Wert von R^2 “gut”? Diese Frage ist ähnlich zur Frage “Ist das viel Geld?”; man kann die Frage nur im Kontext beantworten.

Eine einfache Lösung ist immer, Modelle zu vergleichen. Dann kann man angeben, welches Modell die Daten am besten erklärt, z.B. auf Basis von R^2 .

Zu beachten ist, dass das Modell theoretisch fundiert sein sollte. Vergleicht man viele Modelle aufs Geratewohl, so muss man von zufällig hohen Werten der Modellgüte im Einzelfall ausgehen.

Wenn Sie aber unbedingt eine “objektive” Antwort auf die Frage “wie viel ist viel?” haben wollen, ziehen wir Herrn Cohen zu Rate:

```
interpret_r2(0.2) # aus `easystats`
## [1] "moderate"
## (Rules: cohen1988)
```

Danke, Herr Cohen!

10.13.5 Priori-Verteilung für m10.10 und Modelldefinition

Stan hat für uns folgende Prioris ausgesucht:

```

prior_summary(m10.10) # aus rstanarm
## Priors for model 'm10.10'
## -----
## Intercept (after predictors centered)
## ~ normal(location = -2.8e-16, scale = 2.5)
##
## Coefficients
## ~ normal(location = [0,0], scale = [2.5,2.5])
##
## Auxiliary (sigma)
## ~ exponential(rate = 1)
## -----
## See help('prior_summary.stanreg') for more details

```

Wie gesagt, Stan nimmt dafür einfach die empirischen Mittelwerte und Streuungen her⁶.

Stans Ausgabe kann man in Mathe-Sprech so darstellen, s. Gleichung 10.1.

$$\begin{aligned}
\text{kidscore} &\sim \mathcal{N}(0, 2.5) \\
\mu_i &= \alpha + \beta_1 \text{momiq}_i + \beta_2 \text{momage}_i \\
\alpha &\sim \mathcal{N}(0, 2.5) \\
\beta_1 &\sim \mathcal{N}(0, 2.5) \\
\beta_2 &\sim \mathcal{N}(0, 2.5) \\
\sigma &\sim \mathcal{E}(1)
\end{aligned} \tag{10.1}$$

Man beachte, dass der Achsenabschnitt zur Intelligenz der Kinder auf Null festgelegt wird: Bei *mittlerer* Intelligenz und *mittlerem* Alter der Mutter wird *mittlere* Intelligenz des Kindes erwartet in m10.10. Dadurch, dass nicht nur UV, sondern auch AV zentriert (und in der Streuung auf 1 standardisiert) sind, ist der Mittelwert der AV Null.

Schreibt man einen Bericht, so bietet es sich an, die Modelldefinition *zumindest* im Anhang aufzuführen.

10.13.6 Beantwortung der Forschungsfrage

Das Modell spricht sich klar für einen statistischen, linearen Effekt von Intelligenz der Mutter auf die Intelligenz des Kindes aus, wenn das Alter der Mutter statistisch kontrolliert wird (95%PI: [0.38, 0.51]). Hingegen zeigt das Modell, dass das Alter der Mutter statistisch eher keine Rolle spielt (95%PI: [-0.02, 0.12]). Alle Variablen wurden z-transformiert. Insgesamt erklärt das Modell im Median

⁶Nicht unbedingt die feine bayesianische Art, denn die Prioris sollten ja eigentlich apriori, also vor Kenntnis der Daten, bestimmt werden. Auf der anderen Seite behauptet Stan, von uns zur Rede gestellt, dass die empirischen Mittelwerte ja doch gute Schätzer der echten Parameter sein müssten, wenn die Stichprobe, die wir ihm angeschleppt hätten, tatsächlich gut ist...

einen Anteil von ca. 0.2 an der Varianz der Kinderintelligenz. Das Modell griff auf die Standard-Priori-Werte aus dem R-Paket rstanarm (Goodrich u. a. 2022) zurück (s. Anhang für Details).

! Wichtig

Hier wird von einem “statistischen Effekt” gesprochen, um klar zu machen, dass es sich lediglich um assoziative Zusammenhänge, und nicht um kausale Zusammenhänge, handelt. Kausale Zusammenhänge dürfen wir nur verkünden, wenn wir sie a) explizit untersuchen, b) sich in der Literatur Belege dafür finden oder c) wir ein Experiment fachgerecht durchgeführt haben.

10.14 Vertiefung

VERTIEFUNG, nicht prüfungsrelevant

10.14.1 Verwandtheit von Korrelation und Regression

Sind X und Y *z-standardisiert*, so sind Korrelation und Regression identisch.

$$b = r \frac{sd_x}{sd_y}$$

```
m10.11 <-
  stan_glm(kid_score_z ~ mom_iq_z , data = kidiq, refresh = 0)
coef(m10.11)
##   (Intercept)      mom_iq_z
## -0.0003383038  0.4492821953

kidiq %>%
  select(kid_score, mom_iq, kid_score_z, mom_iq_z) %>%
  correlation()
## # Correlation Matrix (pearson-method)
##
## Parameter1 / Parameter2 /   r /      95% CI / t(432) /      p
## -----
## kid_score /      mom_iq / 0.45 / [0.37, 0.52] /  10.42 / < .001 ***
## kid_score / kid_score_z / 1.00 / [1.00, 1.00] /    Inf / < .001 ***
## kid_score /      mom_iq_z / 0.45 / [0.37, 0.52] /  10.42 / < .001 ***
## mom_iq /      kid_score_z / 0.45 / [0.37, 0.52] /  10.42 / < .001 ***
## mom_iq /      mom_iq_z / 1.00 / [1.00, 1.00] /    Inf / < .001 ***
## kid_score_z /      mom_iq_z / 0.45 / [0.37, 0.52] /  10.42 / < .001 ***
```

```

## 
## p-value adjustment method: Holm (1979)
## Observations: 434

```

10.14.2 Prüfen der Linearitätsannahme

Zentrale Annahme: Die AV ist eine *lineare* Funktion der einzelnen Prädiktoren:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots$$

Hingegen ist es weniger, dass die AV (y) normalverteilt ist. Zwar nimmt die Regression häufig normalverteilte Residuen an, aber diese Annahme ist nicht wichtig, wenn es nur darum geht, die Regressionskoeffizienten zu schätzen (Gelman, Hill, und Vehtari 2021).

Ist die Linearitätsannahme erfüllt, so sollte der Residualplot nur *zufällige* Streuung um $y = 0$ herum zeigen, s. Abbildung 10.22.

Ein Residuum e ist der Vorhersagefehler, also die Differenz zwischen vorhergesagtem und tatsächlichem Wert:

$$e_i = y_i - \hat{y}_i$$

```

kidiq <-
  kidiq %>%
  mutate(m10.10_pred = predict(m10.10), # vorhergesagten Werte
        m10.10_resid = resid(m10.10)) # Residuen

kidiq %>%
  ggplot(aes(x = m10.10_pred, y = m10.10_resid)) +
  geom_hline(color="white", yintercept = 0, size = 2) +
  geom_hline(color = "grey40",
             yintercept = c(-1,1),
             size = 1,
             linetype = "dashed") +
  geom_point(alpha = .7) +
  geom_smooth()

```

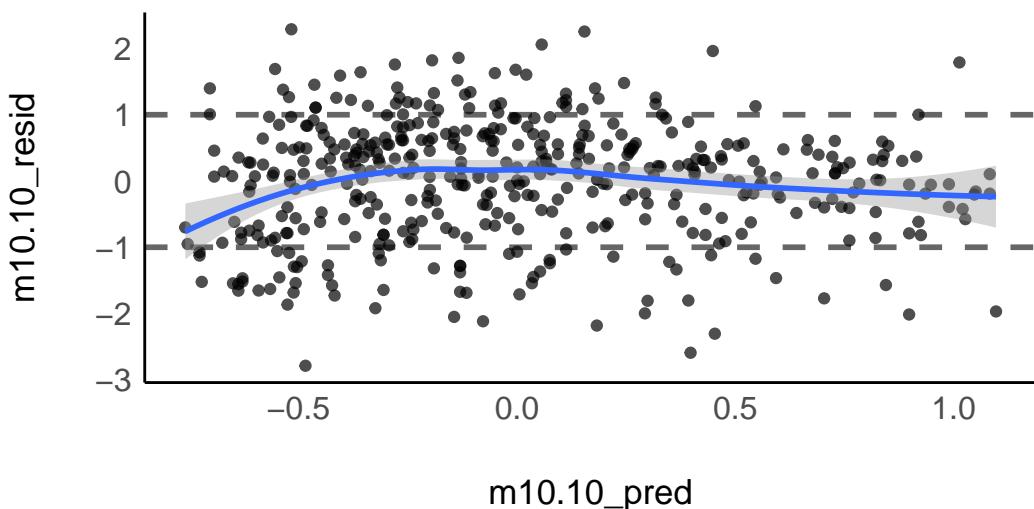
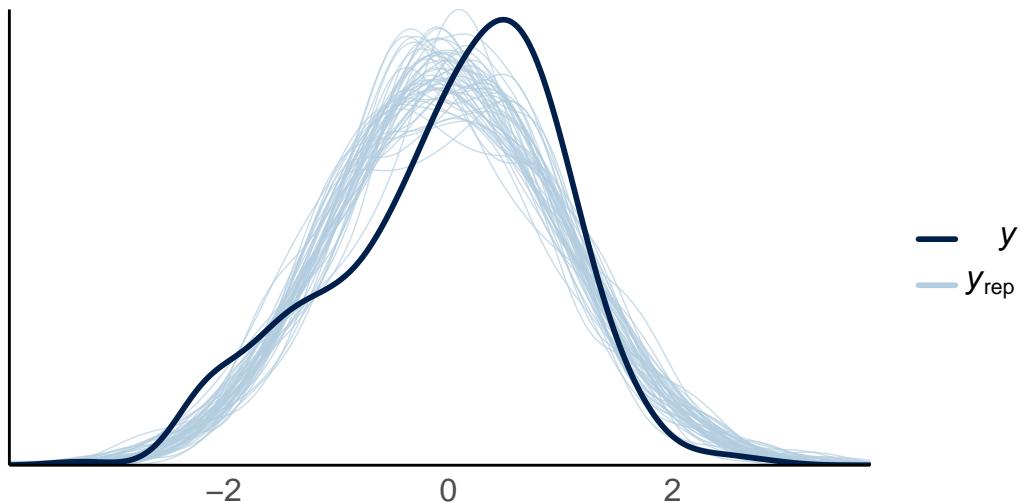


Abbildung 10.22: Die Verteilung der Fehler scheint keinem starken Trend (in Abhangigkeit zum vorhergesagten Wert) zu folgen, was ein gutes Zeichen ist.

Hier erkennt man keine groeren Auffalligkeiten.

10.14.3 Modellprufung mit der PPV

```
pp_check(m10.10)
```



Unser Modell - bzw. die Stichproben unserer Posteriori-Verteilung, y_{rep} verfehlt den Mittelwert von y leider recht haufig.

10.14.4 Visualisierung der bereinigten Regressionskoeffizienten

```
set.seed(42)
data(kidiq)
kidiq3 <-
  kidiq %>%
  standardize	append = TRUE) %>%
  sample_n(size = 300)

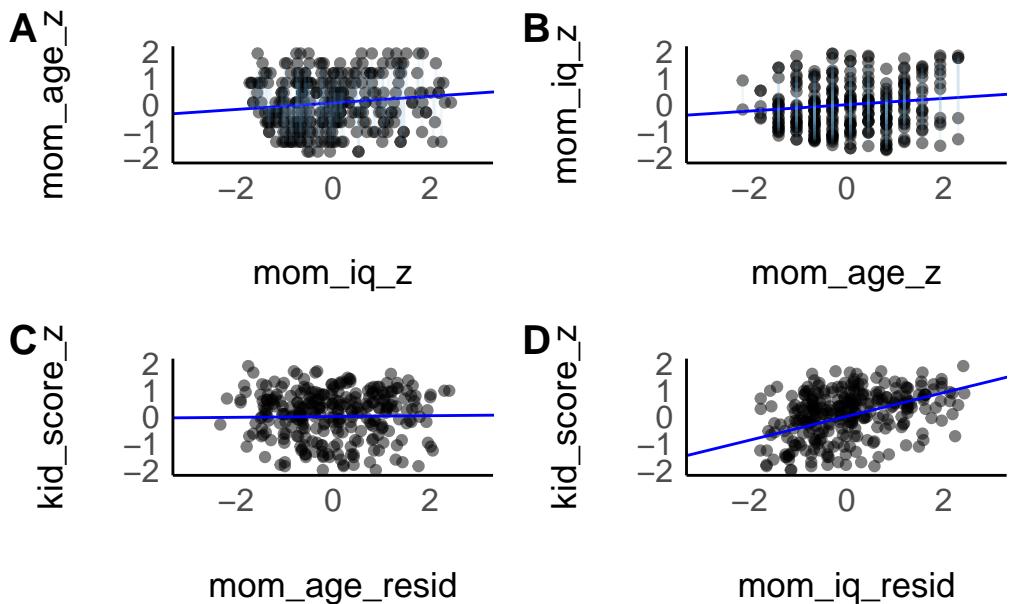
#| results: "hide"
m10.10a <- stan_glm(mom_age_z ~ mom_iq_z, data = kidiq3, refresh = 0, chains = 1)
m10.10b <- stan_glm(mom_iq_z ~ mom_age_z, data = kidiq3, refresh = 0, chains = 1)

kidiq3 <-
  kidiq3 %>%
  mutate(mom_age_resid = resid(m10.10a)) %>%
  mutate(mom_iq_resid = resid(m10.10b))

m10.10c <- stan_glm(kid_score_z ~ mom_age_resid, data = kidiq3, refresh = 0, chains = 1)
m10.10d <- stan_glm(kid_score_z ~ mom_iq_resid, data = kidiq3, refresh = 0, chains = 1)

kidiq3 <-
  kidiq3 %>%
  mutate(m10.10c_resid = resid(m10.10c)) %>%
  mutate(m10.10d_resid = resid(m10.10d))

#(m10.10a_plot + m10.10b_plot) / (m10.10c_plot + m10.10d_plot)
plots(m10.10a_plot, m10.10b_plot, m10.10c_plot, m10.10d_plot,
      n_rows = 2, tags = "A",
      guides = "collect")
```



Die vertikalen Balken zeigen die Residuen.

Obere Reihe: Regression eines Prädiktors auf den anderen Prädiktor. Untere Reihe: Regression der Residuen der oberen Reihe auf die AV, `kid-score_z`. Unten links (C): Die Residuen von `mom_iq_c` sind kaum mit der AV assoziiert. Das heißt, nutzt man den Teil von `mom_age_z`, der nicht mit `mom_iq_z` zusammenhängt, um `kid_score` vorher zusagen, findet man keinen (*kaum*) Zusammenhang. Unten rechts (D): Die Residuen von `mom_age_c` sind *stark* mit der AV assoziiert. Das heißt, nutzt man den Teil von `mom_iq_z`, der nicht mit `mom_age_z` zusammenhängt, um `kid_score` vorher zusagen, findet man einen starken Zusammenhang.

Eine multiple Regression liefert die gleichen Regressionskoeffizienten wie die Modelle aus Teildiagrammen (C) und (D).

Tabelle 10.13: Fixed effects

Parameter	Median	95% CI	pd	Rhat	ESS	Prior
(Intercept)	-1.09e-03	(-0.08, 0.08)	50.95%	0.999	4865.00	Normal (-2.81e-16 +- 2.50)
mom_iq_z	0.44	(0.36, 0.53)	100%	1.000	5118.00	Normal (0.00 +- 2.50)
mom_age_z	0.05	(-0.03, 0.13)	87.83%	1.000	5375.00	Normal (0.00 +- 2.50)

10.14.5 Bereinigte Regressionskoeffizienten für mtcars

```

mtcars2 <-
  mtcars %>%
  standardize()

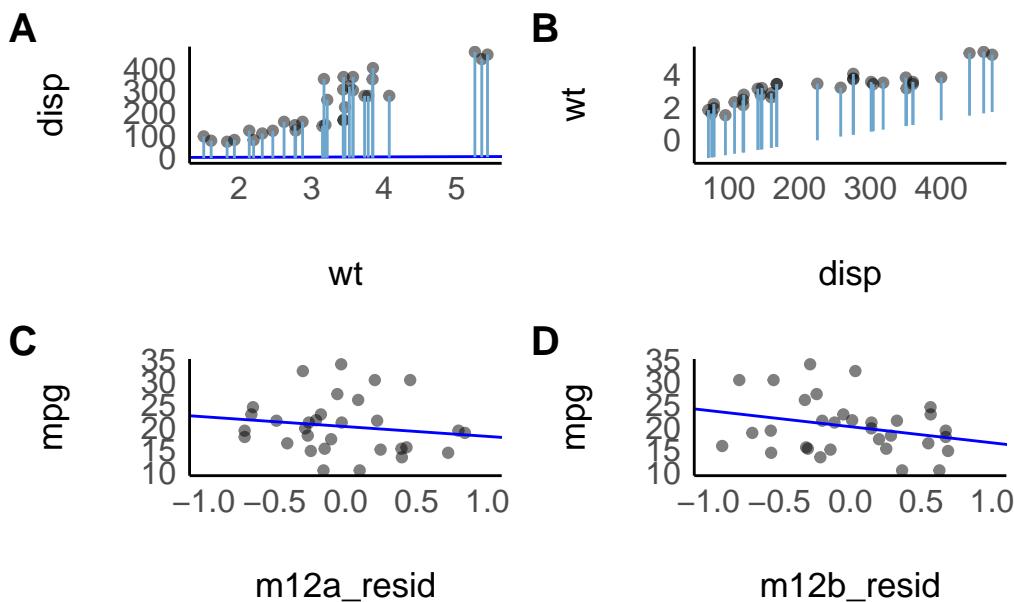
m12a <- stan_glm(disp ~ wt, data = mtcars2, refresh = 0, chains = 1)
m12b <- stan_glm(wt ~ disp, data = mtcars2, refresh = 0, chains = 1)

mtcars2 <-
  mtcars %>%
  mutate(m12a_resid = resid(m12a)) %>%
  mutate(m12b_resid = resid(m12b))

m12c <- stan_glm(mpg ~ m12a_resid, data = mtcars2, refresh = 0, chains = 1)
m12d <- stan_glm(mpg ~ m12b_resid, data = mtcars2, refresh = 0, chains = 1)

mtcars2 <-
  mtcars2 %>%
  mutate(m12c_resid = resid(m12c)) %>%
  mutate(m12d_resid = resid(m12d))

```



Übrigens liefern `stan_glm()` und `lm` oft ähnliche Parameterwerte (bei schwach informativen Prioriwerten):

```

stan_glm(mpg ~ hp + cyl, data = mtcars, refresh = 0) %>% coef()
## (Intercept)          hp          cyl
## 36.84877039 -0.01934222 -2.26124451

lm(mpg ~ hp + cyl, data = mtcars) %>% coef()
## (Intercept)          hp          cyl
## 36.9083305 -0.0191217 -2.2646936

```

! Wichtig

Wenn auch die Parameterwerte eines Frequentistischen und Bayes-Modell numerisch ähnlich sein können, so ist doch die Interpretation grundverschieden. Bayesmodelle erlauben Wahrscheinlichkeitsaussagen zu den Parametern, Frequentistische Modelle nicht.

10.14.6 Post: Bayes in fünf Minuten

Eine Kurzdarstellung des Bayes-Inferenz findet sich [in diesem Post](#) und in [diesem](#).

10.15 Ausblick: Binäre AV

Forschungsfrage: Kann man anhand des Spritverbrauchs vorhersagen, ob ein Auto eine Automatik- bzw. ein manuelle Schaltung hat? Anders gesagt: Hängen Spritverbrauch und Getriebeart? (Datensatz `mtcars`)

Dazu nutzen wir den Datensatz `mtcars`, wobei wir die Variablen z-standardisieren.

```

data(mtcars)
mtcars2 <-
  mtcars %>%
  standardize(append = TRUE)

```

Dann berechnen wir mit Hilfe von Stan ein Regressionsmodell: `m13: am ~ mpg_z`:

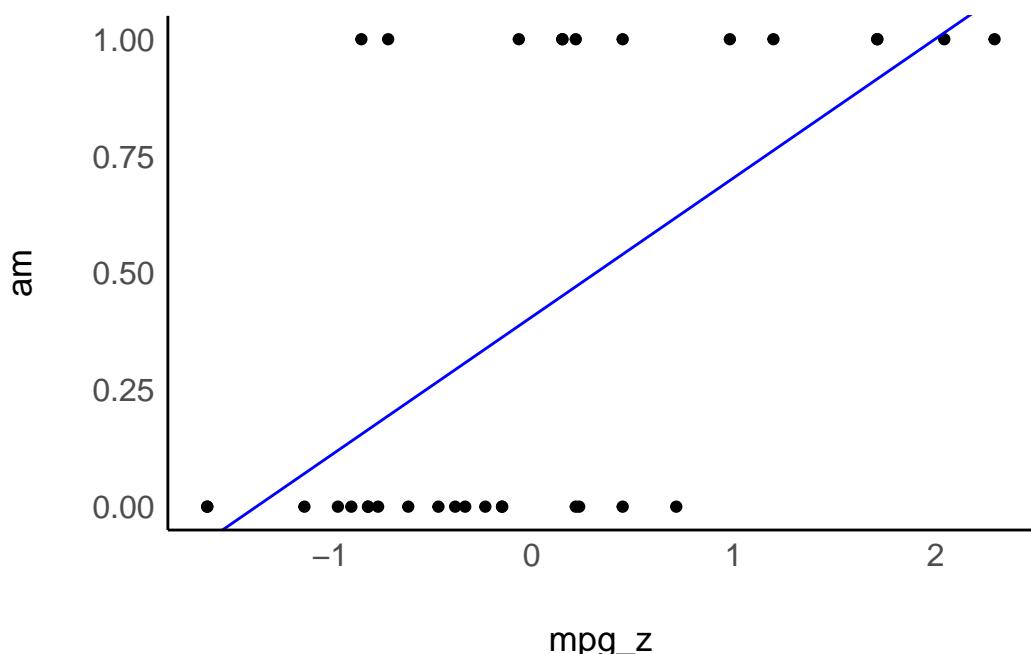
```

m13 <-
  stan_glm(am ~ mpg_z,
            data = mtcars2,
            refresh = 0)
coef(m13)
## (Intercept)      mpg_z
## 0.4060587 0.2975638

```

Ab `mpg_z = 0.41, 0.3` sagt das Modell `am=1` (manuell) vorher. Ganz ok.

```
mtcars2 %>%
  ggplot(aes(x = mpg_z, y = am)) +
  geom_hline(yintercept = 0.5, color = "white", size = 2) +
  geom_point() +
  geom_abline(intercept = coef(m13)[1],
              slope = coef(m13)[2],
              color = "blue")
```



```
neg_am <- predict(m13, newdata = tibble(mpg_z = -1.3))
```

Für kleine Werte von `mpg_z` (<1.3) sagt unser Modell *negative* Werte für `am` voraus. Das macht keinen Sinn: Es gibt keine negativen Werte von `am`, nur 0 und 1. Müssen wir mal bei Gelegenheit besser machen.

Wir waren fleißig ...

[Quelle](#)

Genug für heute.

10.16 Aufgaben

1. Anova-skalenniveau
2. Nullhyp-Beispiel
3. ttest-skalenniveau

4. Griech-Buchstaben-Inferenz
5. Interaktionseffekt1
6. Regression2
7. Regression3
8. diamonds-nullhyp-mws
9. stan_glm_parameterzahl
10. stan_glm_prioriwerte
11. zwert-berechnen
12. Regr-Bayes-interpret
13. Regr-Bayes-interpret03
14. Regr-Bayes-interpret02
15. rope-regr
16. rope1
17. rope2
18. rope3
19. rope4

11 Kausalinferenz

11.1 Lernsteuerung

11.1.1 R-Pakete

Für dieses Kapitel benötigen Sie folgende R-Pakete:

```
library(dagitty)
library(tidyverse)
library(rstanarm)
library(easystats)
```

11.1.2 Lernziele

Nach Absolvieren des jeweiligen Kapitels sollen folgende Lernziele erreicht sein.

Sie können ...

- erklären, wann eine Kausalaussage gegeben eines DAGs berechtigt ist
- die “Atome” der Kausalität eines DAGs benennen
- “kausale Hintertüren” schließen

11.2 Statistik, was soll ich tun?

11.2.1 Studie A: Östrogen

11.2.1.1 Medikament einnehmen?

Mit Blick auf Tabelle 11.1: Was raten Sie dem Arzt? Medikament einnehmen, ja oder nein?

Tabelle 11.1: Daten zur Studie A

Gruppe	Mit Medikament	Ohne Medikament
Männer	81/87 überlebt (93%)	234/270 überlebt (87%)
Frauen	192/263 überlebt (73%)	55/80 überlebt (69%)

Gesamt	273/350 überlebt (78%)	289/350 überlebt (83%)
--------	------------------------	------------------------

Die Daten stammen aus einer (fiktiven) klinischen Studie, $n = 700$, hoher Qualität (Beobachtungsstudie). Bei Männern scheint das Medikament zu helfen; bei Frauen auch. Aber *insgesamt* (Summe von Frauen und Männern) *nicht*?! Was sollen wir den Arzt raten? Soll er das Medikament verschreiben? Vielleicht nur dann, wenn er das Geschlecht kennt (Pearl, Glymour, und Jewell 2016)?

11.2.1.2 Kausalmodell zur Studie A

In Wahrheit sehe die kausale Struktur so aus: Das Geschlecht (Östrogen) hat einen Einfluss (+) auf Einnahme des Medikaments und auf Heilung (-). Das Medikament hat einen Einfluss (+) auf Heilung. Betrachtet man die Gesamt-Daten zur Heilung, so ist der Effekt von Geschlecht (Östrogen) und Medikament *vermengt* (konfundiert, confounded). Die kausale Struktur, also welche Variable beeinflusst bzw. nicht, ist in Abbildung 11.1 dargestellt.

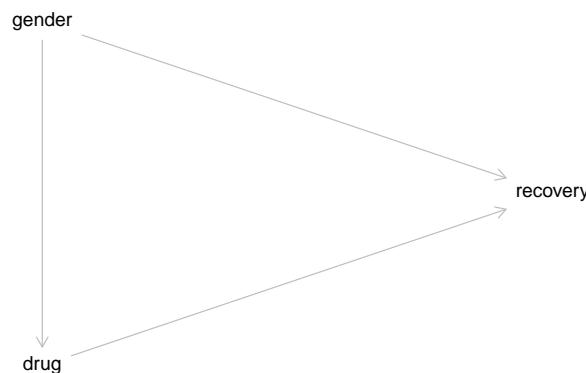


Abbildung 11.1: Zwei direkte Effekte (gender, drug) und ein indirekter Effekt (gender über drug) auf recovery

Betrachtung der Gesamtdaten zeigt in diesem Fall einen *konfundierten* Effekt: Geschlecht konfundiert den Zusammenhang von Medikament und Heilung.

! Wichtig

Betrachtung der Teildaten (d.h. stratifiziert pro Gruppe) zeigt in diesem Fall den wahren, kausalen Effekt. Stratifizieren ist also in diesem Fall der korrekte, richtige Weg. Achtung: Das Stratifizieren ist nicht immer und nicht automatisch die richtige Lösung. Stratifizieren bedeutet, den Gesamtdatensatz in Gruppen oder "Schichten" ("Strata")

11.2.2 Studie B: Blutdruck

11.2.2.1 Medikament einnehmen?

Mit Blick auf Tabelle 11.2: Was raten Sie dem Arzt? Medikament einnehmen, ja oder nein?

Tabelle 11.2: Daten zur Wirksamkeit eines Medikaments (Studie B)

Gruppe	Ohne Medikament	Mit Medikament
geringer Blutdruck	81/87 überlebt (93%)	234/270 überlebt (87%)
hoher Blutdruck	192/263 überlebt (73%)	55/80 überlebt (69%)
Gesamt	273/350 überlebt (78%)	289/350 überlebt (83%)

Die Daten stammen aus einer (fiktiven) klinischen Studie, $n = 700$, hoher Qualität (Beobachtungsstudie). Bei geringem Blutdruck scheint das Medikament zu schaden. Bei hohem Blutdruck scheint das Medikament auch zu schaden. Aber *insgesamt* (Summe über beide Gruppen) *nicht*, da scheint es zu nutzen?! Was sollen wir den Arzt raten? Soll er das Medikament verschreiben? Vielleicht nur dann, wenn er den Blutdruck nicht kennt? Pearl, Glymour, und Jewell (2016)

11.2.2.2 Kausalmodell zur Studie B

Das Medikament hat einen (absenkenden) Einfluss auf den Blutdruck. Gleichzeitig hat das Medikament einen (toxischen) Effekt auf die Heilung. Verringelter Blutdruck hat einen positiven Einfluss auf die Heilung. Sucht man innerhalb der Leute mit gesenktem Blutdruck nach Effekten, findet man nur den toxischen Effekt: Gegeben diesen Blutdruck ist das Medikament schädlich aufgrund des toxischen Effekts. Der positive Effekt der Blutdruck-Senkung ist auf diese Art nicht zu sehen.

Das Kausalmodell ist in Abbildung 11.2 dargestellt.

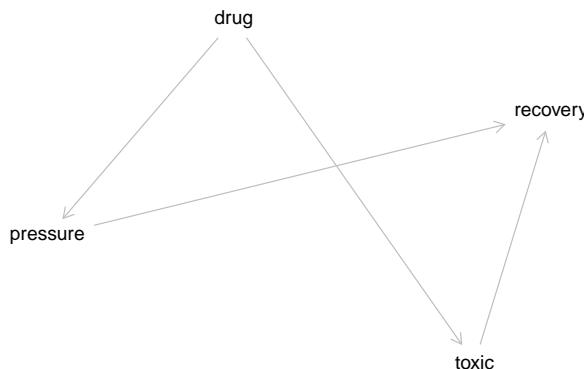


Abbildung 11.2: Drug hat keinen direkten, aber zwei indirekte Effekt auf recovery, einer davon ist heilsam, einer schädlich

Betrachtung der Teildaten zeigt nur den toxischen Effekt des Medikaments, nicht den nützlichen (Reduktion des Blutdrucks).

! Wichtig

Betrachtung der Gesamtdaten zeigt in diesem Fall den wahren, kausalen Effekt. Stratifizieren wäre falsch, da dann nur der toxische Effekt, aber nicht der heilsame Effekt sichtbar wäre.

11.2.3 Studie A und B: Gleiche Daten, unterschiedliches Kausalmodell

Vergleichen Sie die DAGs Abbildung 11.1 und Abbildung 11.2, die die *Kausalmodelle* der Studien A und B darstellen: Sie sind *unterschiedlich*. Aber: Die *Daten* sind *identisch*.

Kausale Interpretation - und damit Entscheidungen für Handlungen - war nur möglich, da das Kausalmodell bekannt ist. Die Daten alleine reichen nicht. Gut merken.

11.2.4 Sorry, Statistik: Du allein schaffst es nicht

Statistik alleine reicht nicht für Kausalschlüsse. Statistik plus Theorie erlaubt Kausalschlüsse.

! Wichtig

Für Entscheidungen (“Was soll ich tun?”) braucht man kausales Wissen. Kausales Wissen basiert auf einer Theorie (Kausalmodell) plus Daten.

11.2.5 Vertiefung¹

11.2.5.1 Studie C: Nierensteine

Nehmen wir an, es gibt zwei Behandlungsvarianten bei Nierensteinen, Behandlung A und B. Ärzte tendieren zu Behandlung A bei großen Steinen (die einen schwereren Verlauf haben); bei kleineren Steinen tendieren die Ärzte zu Behandlung B.

Sollte ein Patient, der nicht weiß, ob sein Nierenstein groß oder klein ist, die Wirksamkeit in der Gesamtpopulation (Gesamtdaten) oder in den stratifizierten Daten (Teildaten nach Steingröße) betrachten, um zu entscheiden, welche Behandlungsvariante er (oder sie) wählt?

¹Dieser Abschnitt ist prüfungsrelevant, birgt aber nichts Neues.

Die Größe der Nierensteine hat einen Einfluss auf die Behandlungsmethode. Die Behandlung hat einen Einfluss auf die Heilung. Damit gibt es eine Mediation (“Kette”) von Größe → Behandlung → Heilung. Darüber hinaus gibt es noch einen Einfluss von Größe der Nierensteine auf die Heilung.

Das Kausalmmodell ist in Abbildung 11.3 dargestellt; Abbildung 11.4 visualisiert alternativ.

Sollte man hier `size` kontrollieren, wenn man den Kausaleffekt von `treatment` schätzen möchte? Oder lieber nicht kontrollieren?

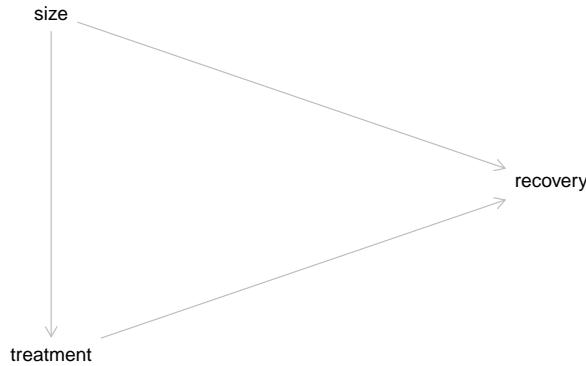


Abbildung 11.3: DAG zur Nierenstein-Studie

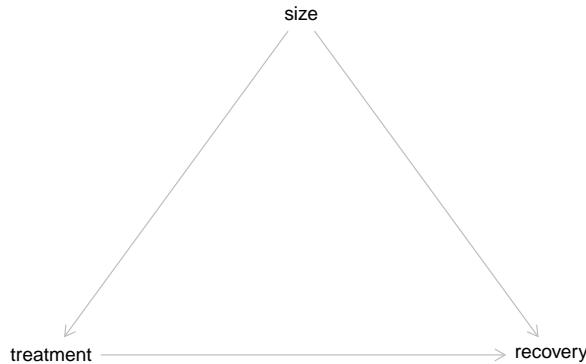


Abbildung 11.4: DAG zur Nierenstein-Studie in zweiter Darstellungsform

Ja: In diesem Fall sollte man `size` kontrollieren, denn man ist am Effekt des `treatments` interessiert. Würde man nicht `size` kontrollieren, bekäme man den “vermengten” Effekt von `size` und `treatment`, also keine (belastbare) Aussage über den Effekt der Behandlung.

11.2.5.2 Mehr Beispiele

Nehmen Sie Bezug zu folgenden Aussagen:

Studien zeigen, dass Einkommen und Heiraten (bzw. verheiratete sein) hoch korrelieren. Daher wird sich dein Einkommen erhöhen, wenn du heiratest.

Tabelle 11.3: ?(caption)

```
##   price livingArea bedrooms waterfront
## 1 132500        906        2       No
## 2 181115       1953        3       No
## 3 109000       1944        4       No
## 4 155000       1944        3       No
## 5  86060        840        2       No
```

Studien zeigen, dass Leute, die sich beeilen, zu spät zu ihrer Besprechung kommen. Daher lieber nicht beeilen, oder du kommst zu spät zu deiner Besprechung.

11.2.6 Zwischenfazit

Bei *Beobachtungsstudien* ist aus den Daten alleine nicht herauszulesen, ob eine Intervention wirksam ist, ob es also einen kausalen Effekt von der Intervention (angenommen Ursache) auf eine AV (Wirkung) gibt. Damit ist auch nicht zu erkennen, welche Entscheidung zu treffen ist. Nur Kenntnis des Kausalmodells zusätzlich zu den Daten erlaubt, eine Entscheidung sinnvoll zu treffen.

Bei *experimentellen Daten* ist die Kenntnis des Kausalmodells nicht nötig (wenn das Experiment handwerklich gut gestaltet ist): Das Randomisieren der Versuchspersonen zu Gruppen und das Kontrollieren der Versuchsbedingungen sorgen dafür, dass es keine Konfundierung gibt.

11.3 Konfundierung

11.3.1 Datensatz ‘Hauspreise im Saratoga County’

Wir nutzen den Datensatz [Saratoga County](#); s. [?@tbl-saratoga](#). Hier gibt es eine [Beschreibung des Datensatzes](#).

```
d_path <- "https://vincentarelbundock.github.io/Rdatasets/csv/mosaicData/SaratogaHouses"
```

11.3.2 Immobilienpreise in einer schicken Wohngegend vorhersagen

Finden Sie den Wert meiner Immobilie heraus! Die muss viel wert sein!”

Das ist Don, Immobilienmogul, Auftraggeber.

Das finde ich heraus. Ich mach das wissenschaftlich.

Das ist Angie, Data Scientistin.

11.3.3 Modell 1: Preis als Funktion der Anzahl der Zimmer

“Hey Don! Mehr Zimmer, mehr Kohle!”

Modell 1 (`m1`) modelliert den Hauspreis als Funktion der Zimmerzahl, s. Abbildung 11.5.

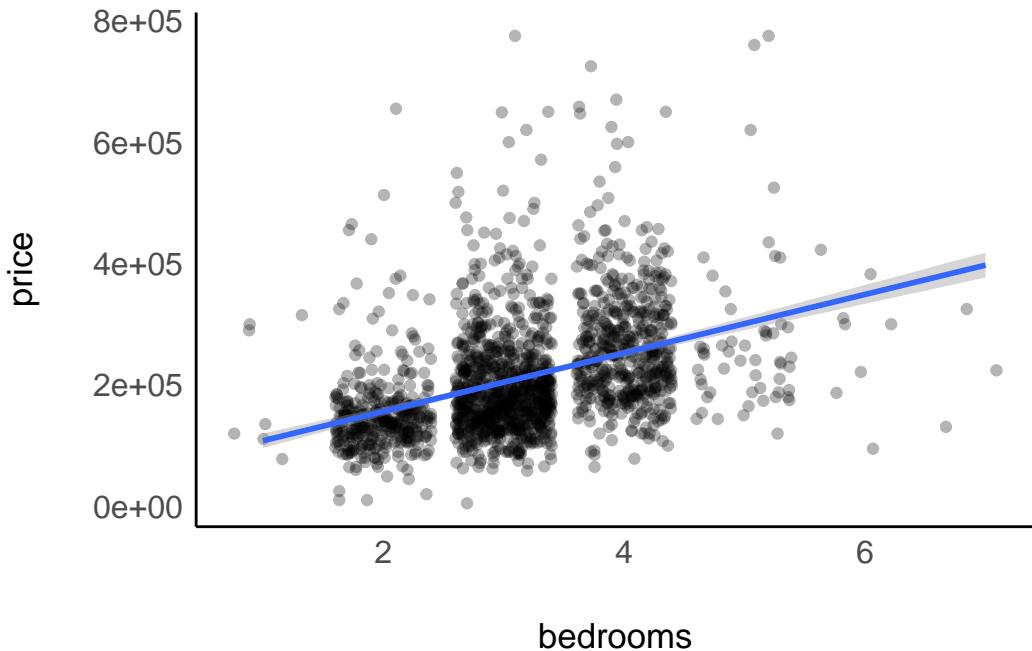


Abbildung 11.5: Modell m1

“Jedes Zimmer mehr ist knapp 50 Tausend wert. Dein Haus hat einen Wert von etwa 150 Tausend Dollar, Don.”

Zu wenig!

Berechnen wir das Modell `m1`; der Punktschätzer des Parameters `bedroom` steht in `?@tbl-m1-hdi`.

```
m1 <- stan_glm(price ~ bedrooms,
                 refresh = 0,
                 seed = 42,
                 data = d)

point_estimate(m1)
## Point Estimate
##
## Parameter / Median / Mean / MAP
```

Tabelle 11.4: ?(caption)

Tabelle 11.5: ?(caption)

```
## -----  
## (Intercept) | 60009.99 | 59849.74 | 60664.24  
## bedrooms    | 48158.27 | 48216.77 | 47909.81
```

`point_estimates(modell)` gibt die Punktschätzer der Parameter eines Modells zurück, aber nicht die Schätzbereiche. Möchten Sie beides, können Sie die Funktion `parameters(modell)` nutzen.²

Mit `estimate_predictions` können wir Vorhersagen berechnen (bzw. schätzen; die Vorhersagen sind ja mit Ungewissheit verbunden, daher ist “schätzen” vielleicht das treffendere Wort). `?@tbl-m1-pred` zeigt den laut `m1` vorhergesagten Hauspreis für ein Haus mit 2 Zimmern.

```
dons_house <- tibble(bedrooms = 2)  
estimate_prediction(m1, data = dons_house)  
## Model-based Prediction  
##  
## bedrooms / Predicted /           SE /           95% CI  
## -----  
## 2.00      /  1.55e+05 / 90704.23 / [-21783.21, 3.32e+05]  
##  
## Variable predicted: price
```

11.3.4 Don hat eine Idee

“Ich bau eine Mauer! Genial! An die Arbeit, Angie!

Don hofft, durch Verdopplung der Zimmerzahl den doppelten Verkaufspreis zu erzielen. Ob das klappt?

Das ist keine gute Idee, Don.”

Berechnen wir die Vorhersagen für Dons neues Haus (mit den durch Mauern halbierten Zimmern), s. `?@tbl-m1-pred2a`.

²In aller Regel macht es mehr Sinn, die Schätzbereiche der Punktschätzer auch zu betrachten. Nur die Punktschätzer zu betrachten vernachlässigt wesentliche Information.

Tabelle 11.6: ?(caption)

```
dons_new_house <- tibble(bedrooms = 4)
estimate_prediction(m1, dons_new_house)
## Model-based Prediction
##
## bedrooms / Predicted /      SE /      95% CI
## -----
## 4.00    /  2.51e+05 / 87947.40 / [86022.85, 4.31e+05]
##
## Variable predicted: price
```

Mit 4 statt 2 Schlafzimmer steigt der Wert auf 250k, laut m1, Abbildung 11.5.

Volltreffer! Jetzt verdien ich 100 Tausend mehr! Ich bin der Größte!

i Hinweis

Zur Erinnerung: “4e+05” ist die Kurzform der wissenschaftlichen Schreibweise und bedeutet: $4 \cdot 100000 = 4 \cdot 10^5 = 400000$

11.3.5 R-Funktionen, um Beobachtungen vorhersagen

`estimate_prediction(m1, dons_new_house)` erstellt *Vorhersageintervalle*, berücksichtigt also *zwei Quellen* von Ungewissheit:

- Ungewissheiten in den Parametern (Modellkoeffizienten, β_0, β_1, \dots)
- Ungewissheit im “Strukturmodell”: Wenn also z.B. in unserem Modell ein wichtiger Prädiktor fehlt, so kann die Vorhersagen nicht präzise sein. Fehler im Strukturmodell schlagen sich in breiten Schätzintervallen (bedingt durch ein großes σ) nieder.

`estimate_expectation(m1, dons_new_house)` erstellt *Konfidenzintervalle*. berücksichtigt also nur *eine Quelle* von Ungewissheit:

- Ungewissheiten in den Parametern (Modellkoeffizienten, β_0, β_1, \dots)

Die Schätzbereiche sind in dem Fall deutlich kleiner:

```
estimate_expectation(m1, dons_new_house)
## Model-based Expectation
##
## bedrooms / Predicted /      SE /      95% CI
## -----
## 4.00    /  2.53e+05 / 3104.56 / [2.47e+05, 2.59e+05]
```

Tabelle 11.7: ?(caption)

Tabelle 11.8: ?(caption)

```
##  
## Variable predicted: price
```

11.3.6 Modell 2

Berechnen wir das Modell `m2`: `price ~ bedrooms + livingArea`, s. ?@tbl-m2.

```
m2 <- stan_glm(price ~ bedrooms + livingArea,  
                 data = d,  
                 seed = 42,  
                 refresh = 0)  
  
point_estimate(m2, centrality = "median")  
## Point Estimate  
##  
## Parameter / Median  
## -----  
## (Intercept) / 36780.69  
## bedrooms / -14238.34  
## livingArea / 125.41
```

Was sind die Vorhersagen des Modells? ?@tbl-m2-pred gibt Aufschluss für den laut `m2` vorhersagten Kaufpreis eines Hauses mit 4 Zimmern und 1200 Quadratfuß Wohnfläche; ?@tbl-m2-pred2 gibt die Schätzung (laut `m2`) für den Preis eines Hauses mit 2 Zimmern (und der gleichen Wohnfläche).

```
estimate_prediction(m2, data = tibble(bedrooms = 4, livingArea = 1200))  
## Model-based Prediction  
##  
## bedrooms / livingArea / Predicted / SE / 95% CI  
## -----  
## 4.00 / 1200.00 / 1.31e+05 / 69459.10 / [-6973.83, 2.63e+05]  
##  
## Variable predicted: price
```

Andere, aber ähnliche Frage: Wieviel kostet ein Haus mit sagen wir 4 Zimmer *gemittelt* über die verschiedenen Größen von `livingArea`? Stellen Sie sich alle Häuser mit 4 Zimmern vor (also mit verschiedenen Wohnflächen). Wir möchten nur wissen, was so ein Haus “im

Tabelle 11.9: ?(caption)

```
## Model-based Prediction
##
##   bedrooms | livingArea | Predicted |      SE |          95% CI
## -----
## 2.00     |    1200.00 | 1.60e+05 | 68590.99 | [27283.79, 2.97e+05]
##
## Variable predicted: price
```

Tabelle 11.10: ?(caption)

Mittel” kostet. Wir möchten also die Mittelwerte pro `bedroom` schätzen, gemittelt für jeden Wert von `bedroom` über `livingArea`. Die Ergebnisse stehen in `?@tbl-m2-preds` und sind in Abbildung 11.6 visualisiert.

```
estimate_means(m2, at = "bedrooms", length = 7)
## Estimated Marginal Means
##
##   bedrooms |      Mean |          95% CI
## -----
## 1.00     | 2.43e+05 | [2.31e+05, 2.54e+05]
## 2.00     | 2.28e+05 | [2.22e+05, 2.35e+05]
## 3.00     | 2.14e+05 | [2.11e+05, 2.17e+05]
## 4.00     | 2.00e+05 | [1.95e+05, 2.05e+05]
## 5.00     | 1.86e+05 | [1.76e+05, 1.96e+05]
## 6.00     | 1.71e+05 | [1.57e+05, 1.86e+05]
## 7.00     | 1.57e+05 | [1.37e+05, 1.77e+05]
##
## Marginal means estimated at bedrooms
```

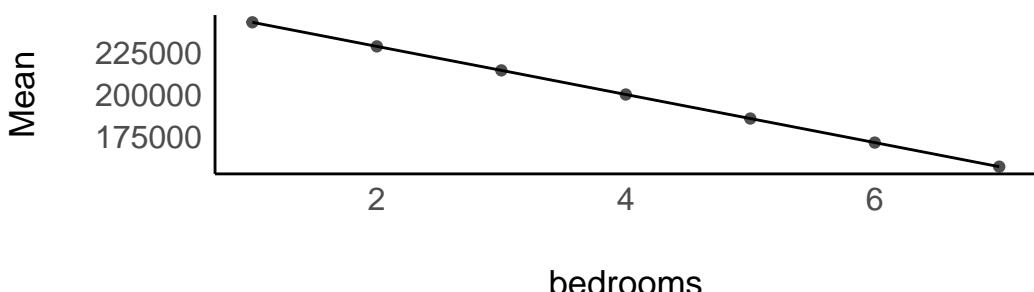


Abbildung 11.6: Hauspreis als Funktion der Zimmerzahl, laut m2

“Die Zimmer zu halbieren, hat den Wert des Hauses *verringert*, Don!”

“Verringert!? Weniger Geld?! Oh nein!”

11.3.7 Die Zimmerzahl ist negativ mit dem Preis korreliert

... wenn man die Wohnfläche (Quadratmeter) kontrolliert, s. Abbildung 11.7.

“Ne-Ga-Tiv!”

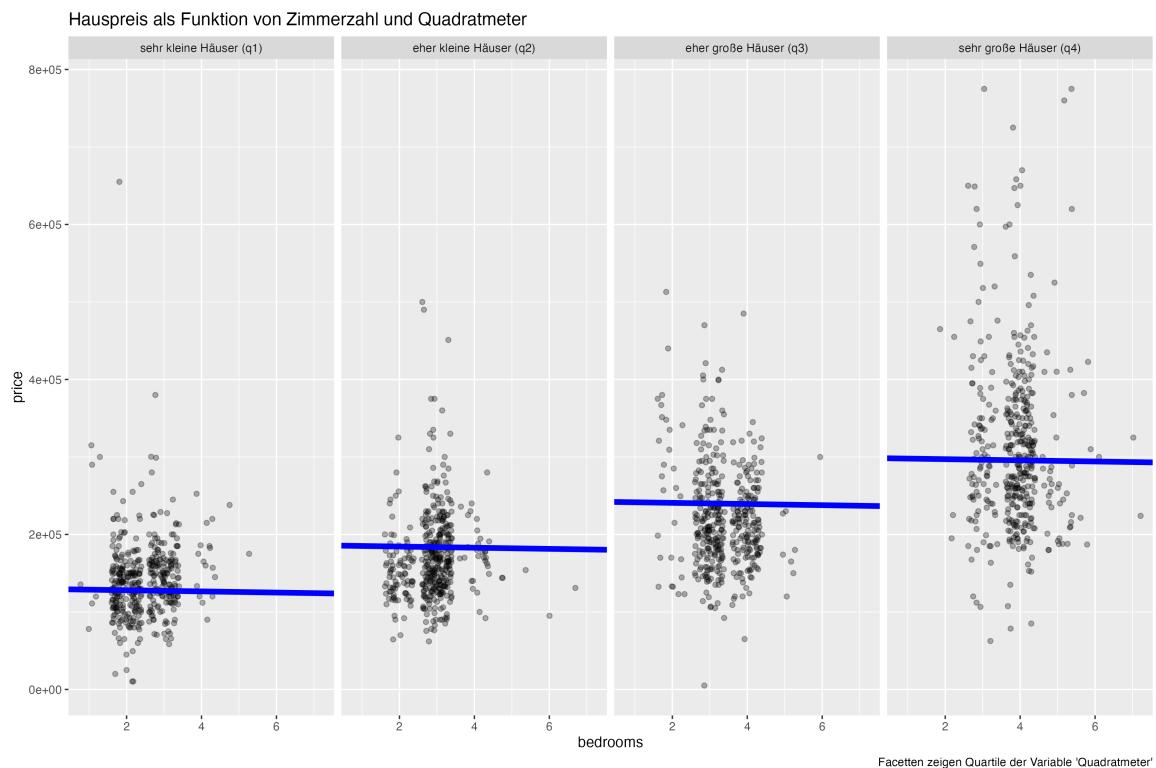


Abbildung 11.7: Hauspreis stratifizieren

[Quellcode](#)

11.4 Kontrollieren von Variablen

Durch das Aufnehmen von Prädiktoren in die multiple Regression werden die Prädiktoren *kontrolliert* (adjustiert, konditioniert):

Die Koeffizienten einer multiplen Regression zeigen den Zusammenhang β des einen Prädiktors mit y , wenn man den (oder die) anderen Prädiktoren statistisch *konstant hält*.

Man nennt die Koeffizienten einer multiplen Regression daher auch *parzielle Regressionskoeffizienten*. Manchmal spricht man, eher umgangssprachlich, auch vom “Netto-Effekt” eines Prädiktors, oder davon, dass ein Prädiktor “bereinigt” wurde vom (linearen) Einfluss der anderen Prädiktoren auf y .

Damit kann man die Regressionskoeffizienten so interpretieren, dass Sie den Effekt des Prädiktors x_1 auf y anzeigen *unabhängig* vom Effekt der anderen Prädiktoren, x_2, x_3, \dots auf y .

Man kann sich dieses Konstanthalten vorstellen als eine Aufteilung in Gruppen: Der Effekt eines Prädiktors x_1 wird für jede Ausprägung (Gruppe) des Prädiktors x_2 berechnet.

11.4.1 Das Hinzufügen von Prädiktoren kann die Gewichte der übrigen Prädiktoren ändern

Aber welche und wie viele Prädiktoren soll ich denn jetzt in mein Modell aufnehmen?! Und welches Modell ist jetzt richtig?!

Leider kann die Statistik keine Antwort darauf geben.

Wozu ist sie dann gut?!

! Wichtig

In Beobachtungsstudien hilft nur ein (korrektes) Kausalmmodell. Ohne Kausalmmodell ist es nutzlos, die Regressionskoeffizienten (oder eine andere Statistik) zur Erklärung der Ursachen heranzuziehen: Die Regressionskoeffizienten können sich wild ändern, wenn man Prädiktoren hinzufügt oder weglässt. Es können sich sogar die Vorzeichen der Regressionsgewichte ändern; in dem Fall spricht man von einem Simpson-Paradox.

11.5 Welches Modell richtig ist, kann die Statistik nicht sagen

Often people want statistical modeling to do things that statical modeling cannot do. For example, we'd like to know wheter an effect is “real” or rather spurious. Unfortunately, modeling merely quantifies uncertainty in the precise way that the model understands the problem. Usually answers to lage world questions about truth and causation depend upon information not included in the model. For example, any observed correlation between an outcome and predictor could be eliminated or reversed once another predictor is added to the model. But if we cannot think of the right variable, we might never notice. Therefore all statical

models are vulnerable to and demand critique, regardless of the precision of their estimates and apparent accuracy of their predictions. Rounds of model criticism and revision embody the real tests of scientific hypotheses. A true hypothesis will pass and fail many statistical “tests” on its way to acceptance.

McElreath (2020), S. 139

11.5.1 Kausalmmodell für Konfundierung, km1

Das Kausalmmodell km1 ist in Abbildung 11.8 dargestellt; vgl. Abbildung 11.7.

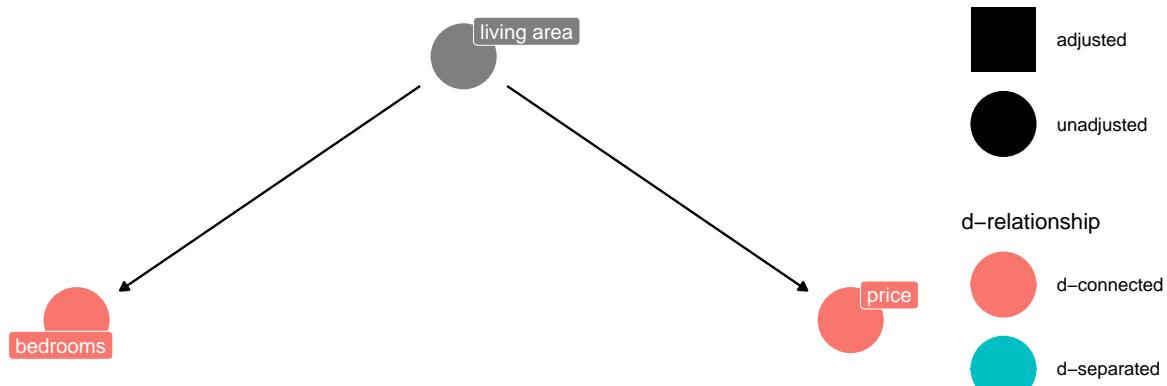


Abbildung 11.8: Kausalmmodell km1 - Eine Erklärung (von mehreren) für m1 bzw. die Daten, die m1 zugrunde liegen

Wenn dieses Kausalmmodell stimmt, findet man eine *Scheinkorrelation* zwischen `price` und `bedrooms`.

Eine Scheinkorrelation ist ein Zusammenhang, der *nicht* auf eine kausalen Einfluss beruht.

`d_connected` heißt, dass die betreffenden Variablen “verbunden” sind durch einen gerichteten (d wie directed) Pfad, durch den die Assoziation (Korrelation) wie durch einen Fluss fließt . `d_separated` heißt, dass sie nicht `d_connected` sind.

11.5.2 m2 kontrolliert die Konfundierungsvariable livingArea

Wenn das Kausalmmodell stimmt, dann zeigt m2 den kausalen Effekt von `livingArea`.

Was tun wir jetzt bloß?! Oh jeh!

Wir müssen die Konfundierungsvariable kontrollieren.

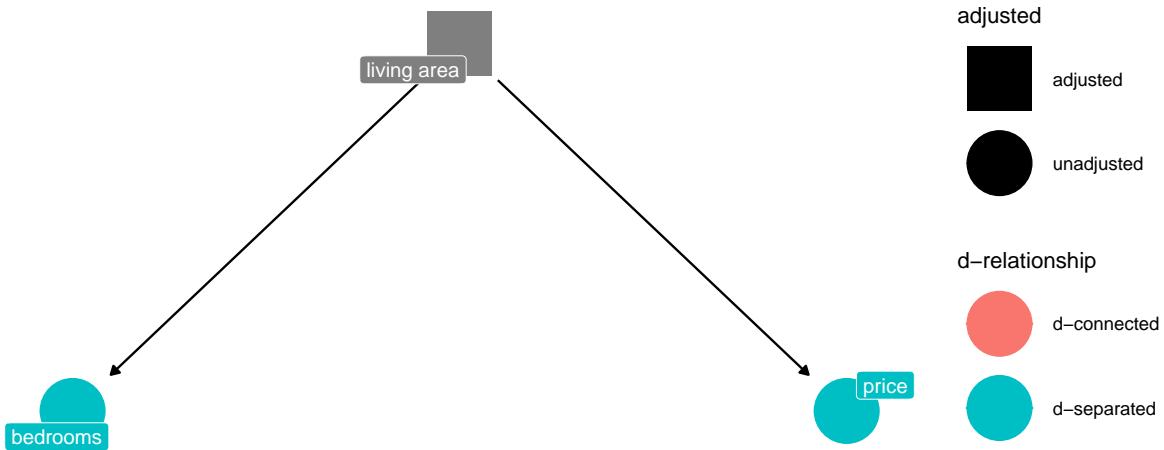


Abbildung 11.9: Durch Kontrolle von living area wird die Assoziation von price und bedrooms aufgehoben.

Abbildung 11.9 zeigt, dass `bedrooms` und `price` *unkorreliert* werden (`d_separated`), wenn man `living area` kontrolliert.

Durch das Kontrollieren (“adjustieren”), sind `bedrooms` und `price` nicht mehr korreliert, nicht mehr `d_connected`, sondern jetzt `d_separated`.

11.5.3 Konfundierer kontrollieren

Gehen wir in diesem Abschnitt davon aus, dass `km1` richtig ist.

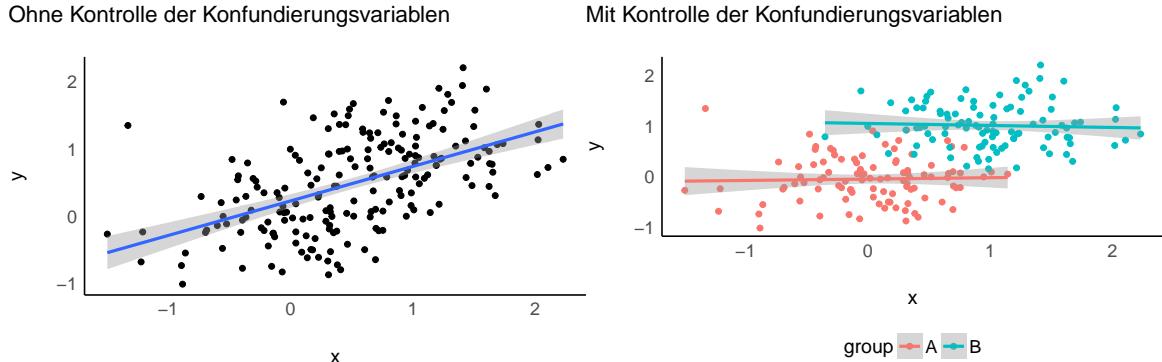
Ohne Kontrollieren der Konfundierungsvariablen: Regressionsmodell $y \sim x$, Abbildung 11.10, links: Es wird (falschlich) eine Korrelation zwischen x und y angezeigt: Scheinkorrelation. *Mit* Kontrollieren der Konfundierungsvariablen: Regressionsmodell $y \sim x + group$, Abbildung 11.10, rechts.

Abbildung 11.10, rechts, zeigt korrekt, dass es keine Korrelation zwischen x und y gibt, wenn `group` kontrolliert wird. Außerdem sieht man im rechten Teildiagramm, dass es ein Kontrollieren der Variable `group` durch Aufnahme als Prädiktor in die Regressionsgleichung einem Stratifizieren entspricht (getrennte Berechnung der Regressionsgerade pro Gruppe).

[Quellcode](#)

11.5.4 m1 und m2 passen nicht zu den Daten, wenn km1 stimmt

Laut `km1` dürfte es keine Assoziation (Korrelation) zwischen `bedrooms` und `price` geben, wenn man `livingArea` kontrolliert, wie in Abbildung 11.8 dargestellt. Es gibt aber noch eine Assoziation zwischen `bedrooms` und `price` geben, wenn man `livingArea` kontrolliert. Daher sind sowohl `m1` und `m2` nicht mit dem Kausalmmodell `km1` vereinbar.



- (a) Ohne Kontrolle der Konfundierungsvariablen:
Konfundierung tritt auf.
- (b) Mit Kontrolle der Konfundierungsvariablen:
Konfundierung tritt nicht auf.

Abbildung 11.10: Konfundierung von y und x!

11.5.5 Kausalmodell 2, km2

Unser Modell `m2` sagt uns, dass beide Prädiktoren jeweils einen eigenen Beitrag zur Erklärung der AV haben.

Daher könnte das folgende Kausalmodell, `km2` besser passen.

In diesem Modell gibt es eine *Wirkkette*: $a \rightarrow b \rightarrow p$.

Insgesamt gibt es zwei Kausaleinflüsse von **a** auf **p**: - $a \rightarrow p$ - $a \rightarrow b \rightarrow p$

Man nennt die mittlere Variable einer Wirkkette auch einen *Mediator* und den Pfad von der UV (**a**) über den Mediator (**b**) zur AV (**p**) auch *Mediation*, s. Abbildung 11.11.

11.5.6 Dons Kausalmodell, km3

So sieht Dons Kausalmodell aus, s. Abbildung 11.12.

Ich glaube aber an mein Kausalmodell. Mein Kausalmodell ist das größte! Alle anderen Kausalmodelle sind ein Disaster!"

"Don, nach deinem Kausalmodell müssten `bedrooms` und `livingArea` unkorreliert sein. Sind sie aber nicht."

Rechne doch selber die Korrelation aus, Don:

Äh, wie ging das nochmal?

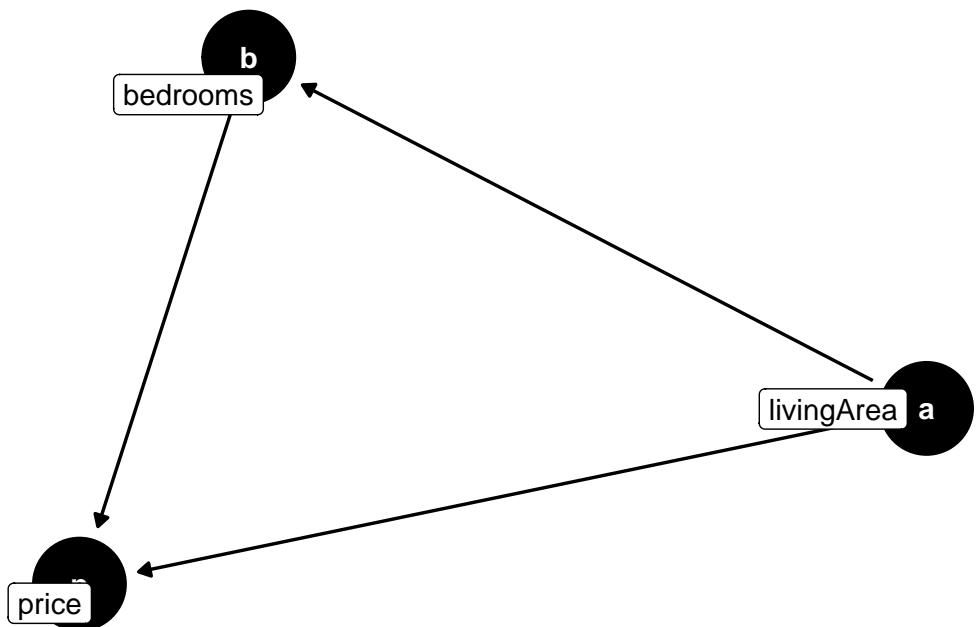


Abbildung 11.11: Der Effekt von livingArea wird über den Mediator bedrooms auf price vermittelt.

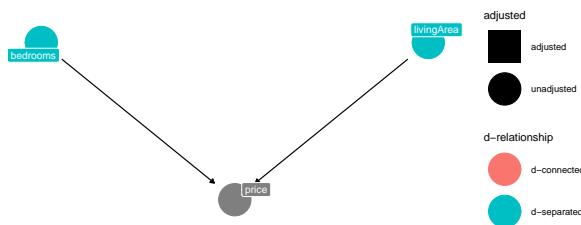


Abbildung 11.12: Dons Kausalmmodell

So könntest du das rechnen, Don: `correlation(d, select = c("bedrooms", "livingArea"))`. Oder z.B. so:

```
dons_r <- d %>%
  summarise(cor(bedrooms, livingArea))
```

Die Korrelation liegt also bei 0.66

Bitte, gerne hab ich dir geholfen, Don.

11.5.7 Unabhängigkeiten laut der Kausalmodelle

`km1`: b: bedrooms, p: price, a area (living area), s. Abbildung 11.8.

Das Kausalmodell `km1` behauptet: $b \perp\!\!\!\perp p | a$: `bedrooms` sind unabhängig von `price`, wenn man `livingArea` kontrolliert.

Kontrollieren einer Variable Z erreicht man auf einfache Art, indem man sie in zusätzlich zur vermuteten Ursache X in die Regressionsgleichung mit aufnimmt, also $y \sim x + z$.

Aber diese behauptete Unabhängigkeit findet sich *nicht* in den Daten wieder, s. `?@tbl-m2`. Also: Passt nicht zu den Daten!

`km2` b: bedrooms, p: price, a area (living area), s. Abbildung 11.11.

Das Kausalmodell `km2` postuliert *keine* Unabhängigkeiten: Laut `km2`sind alle Variablen des Modells miteinander assoziiert (korreliert).

i Hinweis

Ein Modell, in dem alle Variablen miteinander korreliert sind, nennt man auch *saturiert* oder *saturiertes Modell*. So ein Modell ist empirisch *schwach*. Denn: Behauptet ein Modell, dass die Korrelation zwischen zwei Variablen irgendeinen Wert zwischen -1 und +1 beträgt (nur nicht exakt Null), so ist das eine sehr schwache Aussage (und kaum zu falsifizieren). So ein Modell ist wissenschaftlich wenig wert. Das ist so ähnlich wie ein Modell, das voraussagt, dass es morgen irgendeine Temperatur hat zwischen -30 und +30 Grad (nur nicht exakt Null). Trifft diese Temperaturvorhersage ein, so werden wir nicht gerade beeindruckt sein.

Fazit: `km2` passt zu den Daten, aber wir sind nicht gerade beeindruckt vom Modell.

`km3`: b: bedrooms, p: price, a area (living area), s. Abbildung 11.12.

$b \perp\!\!\!\perp a$: `bedrooms` sind unabhängig von `livingArea` (a)

`km3` passt nicht zu den Daten/zum Modell!

11.6 DAGs: Directed Acyclic Graphs

Was sind DAGs? Wir haben in diesem Kapitel schon viele Beispiele gesehen, z.B. Abbildung 11.12.

- DAGs sind eine bestimmte Art von Graphen zur Analyse von Kausalstrukturen.
- Ein *Graph* besteht aus Knoten (Variablen) und Kanten (Linien), die die Knoten verbinden.
- DAGs sind *gerichtet*; die Pfeile zeigen immer in eine Richtung (und zwar von Ursache zu Wirkung).
- DAGs sind *azyklisch*; die Wirkung eines Knoten darf nicht wieder auf ihn zurückführen.
- Ein *Pfad* ist ein Weg durch den DAG, von Knoten zu Knoten über die Kanten, unabhängig von der Pfeilrichtung.

Der DAG von km1 ist in Abbildung 11.8 zu sehen.

11.6.1 Leider passen potenziell viele DAGs zu einer Datenlage

b: bedrooms, p: price, a area (living area)

Ja, der Job der Wissenschaft ist kein Zuckerschlecken. Aber wenn es einfach wäre, die Kausalstruktur der Phänomene zu entdecken, wären sie längst erkannt, und alle Probleme der Menschheit gelöst.

In Abbildung 11.13 sind mögliche Kausalmodelle für Dons Studie dargestellt.

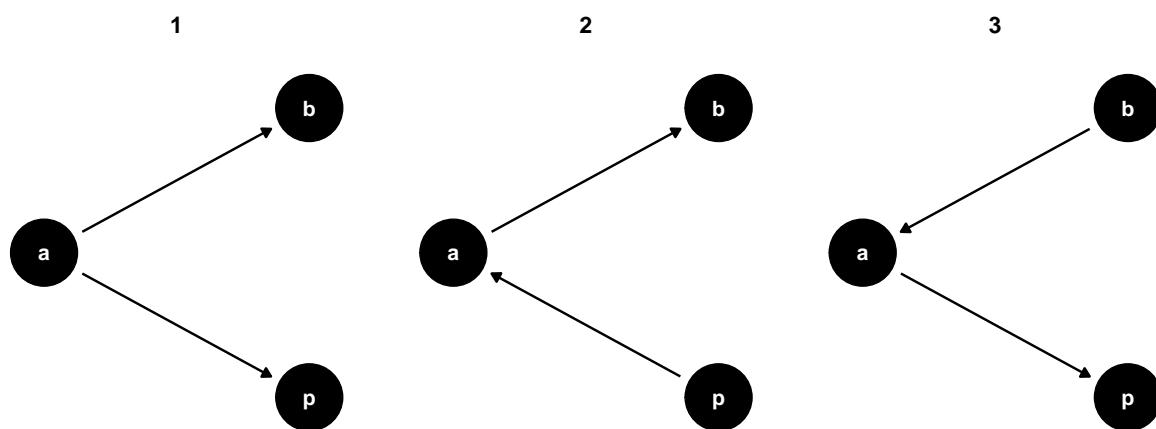


Abbildung 11.13: Kausalmodelle, die potenziell geeignet sind für Dons Fragestellung

Alle diese DAGs in Abbildung 11.8 haben die *gleichen* Implikationen hinsichtlich der (Un-)Abhängigkeiten zwischen den Variablen. Wir können also leider empirisch nicht bestimmen, welcher der DAGs der richtige ist. Um den richtigen DAG zu identifizieren, bräuchten wir z.B. einen reichhaltigeren DAG, also mit mehr Variablen.

11.6.2 Was ist eigentlich eine Ursache?

Etwas verursachen kann man auch (hochtrabend) als “Kausation” bezeichnen.

i Hinweis

Weiß man, was die Wirkung W einer Handlung H (Intervention) ist, so hat man H als Ursache von W erkannt.

McElreath (2020)

Viele Menschen denken - fälschlich - dass Korrelation Kausation bedeuten muss, s. Abbildung 11.14.

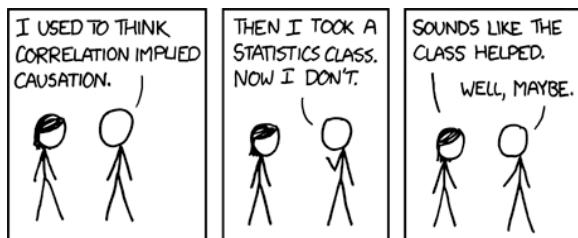


Abbildung 11.14: xkcd zum Thema Kausation

[Quelle](#) und [Erklärung](#)

11.6.3 Zwischenfazit

Sind zwei Variablen korreliert (abhängig, assoziiert), so kann es dafür zwei Gründe geben:

- Kausaler Zusammenhang
- Nichtkausaler Zusammenhang (“Scheinkorrelation”)

Eine mögliche Ursache einer Scheinkorrelation ist Konfundierung.

Konfundierung kann man entdecken, indem man die angenommene Konfundierungsvariable kontrolliert (adjustiert), z.B. indem man ihn als Prädiktor in eine Regression aufnimmt.

Ist die Annahme einer Konfundierung korrekt, so löst sich der Scheinzusammenhang nach dem Adjustieren auf.

Löst sich der Scheinzusammenhang nicht auf, sondern drehen sich die Vorzeichen der Zusammenhänge nach Adjustieren um, so spricht man einem *Simpson-Paradox*.

Die Daten alleine können nie sagen, welches Kausalmodell der Fall ist in einer Beobachtungsstudie. Fachwissen (inhaltliches wissenschaftliches Wissen) ist nötig, um DAGs auszuschließen.

11.6.4 Schoki macht Nobelpreis! (?)

Vertiefung

Eine Studie fand eine starke Korrelation, $r = 0.79$ zwischen (Höhe des) Schokoladenkonsums eines Landes und (Anzahl der) Nobelpreise eines Landes (Messerli 2012), s. Abbildung 11.15.

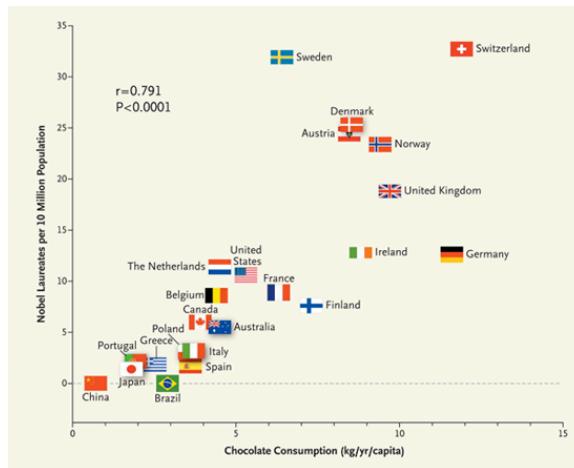


Abbildung 11.15: Je mehr Schoki, desto mehr Nobelpreise

! Wichtig

Korrelation ungleich Kausation! Korrelation kann bedeuten, dass eine Kausation vorliegt, aber es muss auch nicht sein, dass Kausation vorliegt. Liegt Korrelation ohne Kausation vor, so spricht man von einer *Scheinkorrelation*. Um Scheinkorrelation von echter Assoziation (auf Basis von Kausation) abzugrenzen, muss man die Kausalmodelle überprüfen, so wie wir das hier tun.

Der “Schoki-DAG” in Abbildung 11.16 zeigt den DAG für das Schokoladen-Nobelpreis-Modell.

11.7 Kollision

11.7.1 Kein Zusammenhang von Intelligenz und Schönheit (?)

Gott ist gerecht (?)

Zumindest findet sich in folgenden Daten kein Zusammenhang von Intelligenz (**talent**) und Schönheit (**looks**), wie Abbildung 11.17 illustriert. Für geringe Intelligenzwerte gibt es eine breites Spektrum von Schönheitswerten und für hohe Intelligenzwerte sieht es genauso aus.

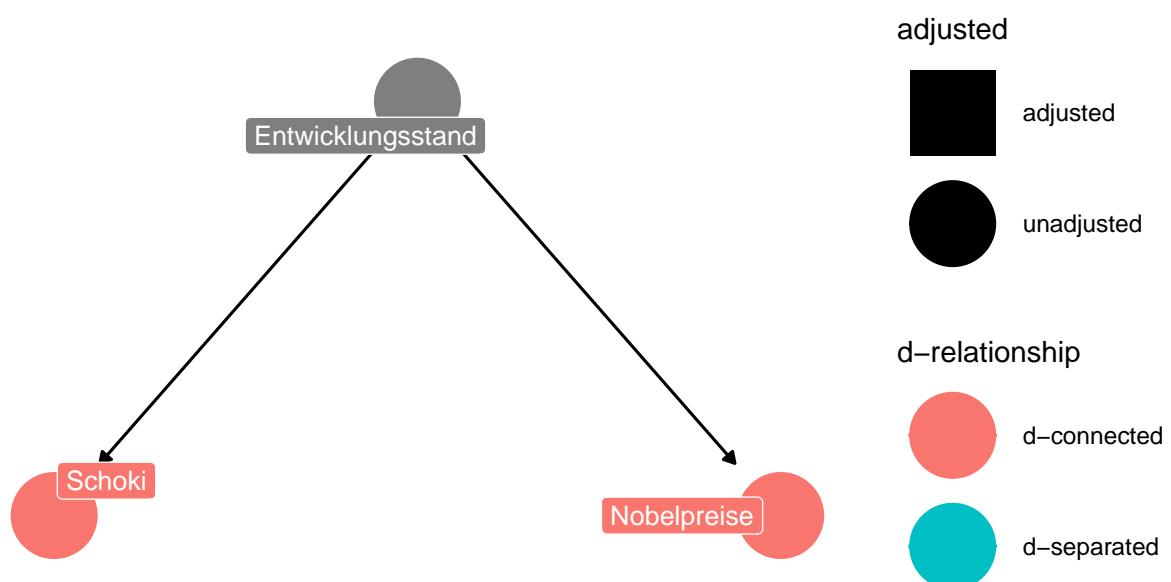


Abbildung 11.16: Macht Schokolade Nobelpreise?

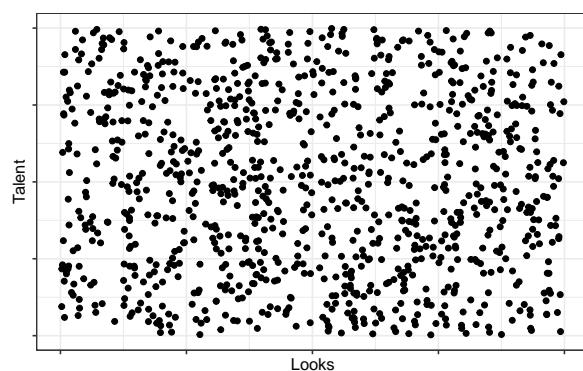


Abbildung 11.17: Kein Zusammenhang von Intelligenz und Schönheit in den Daten

Gott ist gerecht (?)

11.7.2 Aber Ihre Dates sind entweder schlau oder schön

Seltsamerweise beobachten Sie, dass die Menschen, die Sie daten (Ihre Dates), entweder schön sind oder schlau - aber seltens beides gleichzeitig (schade), s. Abbildung 11.18.

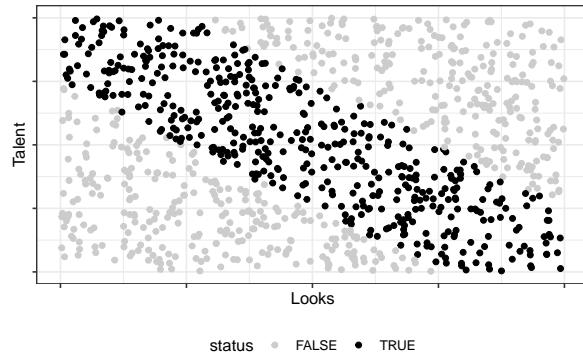


Abbildung 11.18: Ihre Datingpartner sind komischerweise entweder schlau oder schön (aber nicht beides), zumindest in der Tendenz.

Wie kann das sein?

11.8 DAG zur Rettung

Der DAG in Abbildung 11.19 bietet eine rettende Erklärung.

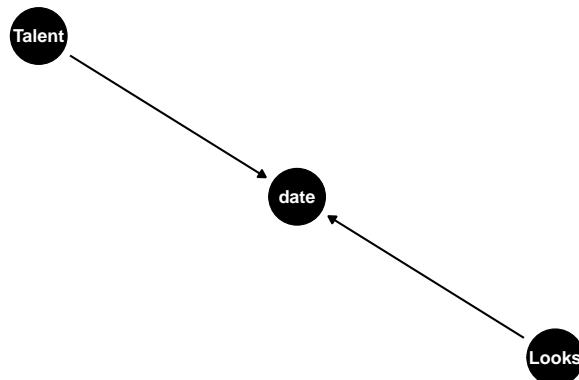


Abbildung 11.19: Date als gemeinsame Wirkung von Schönheit und Intelligenz. Stratifiziert man die gemeinsame Wirkung (dates), so kommt es zu einer Scheinkorrelation zwischen Schönheit und Intelligenz.

Eine ähnliche Visualisierung des gleichen Sachverhalts zeigt Abbildung 11.20.

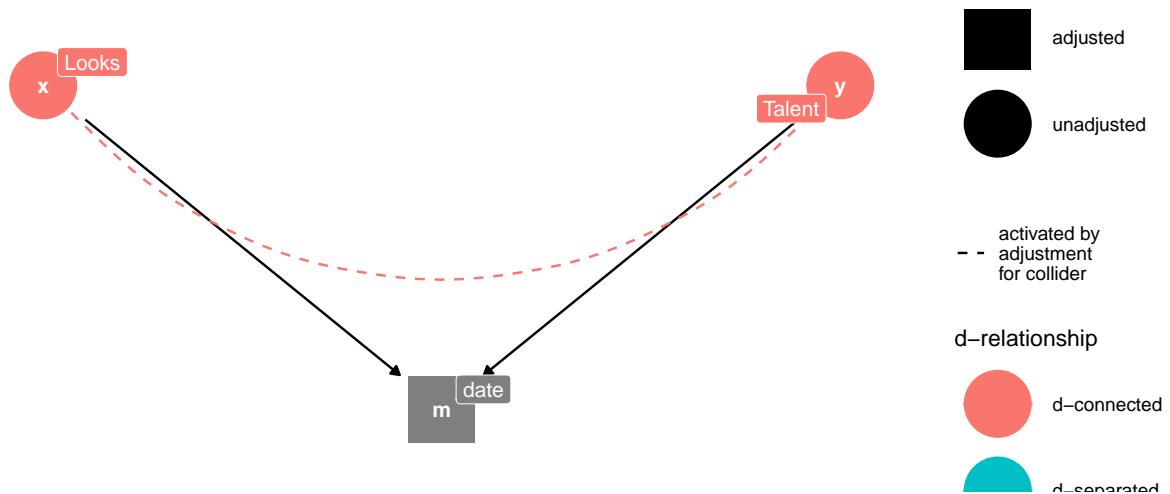


Abbildung 11.20: Durch Kontrolle der gemeinsamen Wirkung entsteht eine Scheinkorrelation zwischen den Ursachen

11.8.1 Was ist eine Kollision?

Als *Kollision* (Kollisionsverzerrung, Auswahlverzerrung, engl. collider) bezeichnet man einen DAG, bei dem eine Wirkung zwei Ursachen hat (eine gemeinsame Wirkung zweier Ursachen). Kontrolliert man die *Wirkung* **m**, so entsteht eine Scheinkorrelation zwischen den Ursachen **x** und **y**. Kontrolliert man die Wirkung nicht, so entsteht keine Scheinkorrelation zwischen den Ursachen, s. Abbildung 11.19, vgl. Rohrer (2018).

! Wichtig

Man kann also zu viele oder falsche Prädiktoren einer Regression hinzufügen, so dass die Koeffizienten nicht die kausalen Effekte zeigen, sondern durch Scheinkorrelation verzerrte Werte.

11.8.2 Einfaches Beispiel zur Kollision

In der Zeitung *Glitzer* werden nur folgende Menschen gezeigt:

- Schöne Menschen
- Reiche Menschen

ehen wir davon aus, dass Schönheit und Reichtum unabhängig voneinander sind.

Wenn ich Ihnen sage, dass Don nicht schön ist, aber in der Glitzer häufig auftaucht, was lernen wir dann über seine finanzielle Situation?³

“Ich bin schön, unglaublich schön, und groß, großartig, tolle Gene!!!”

³Don muss reich sein.

11.8.3 Noch ein einfaches Beispiel zur Kollision

“So langsam check ich’s!”

Sei $Z = X + Y$, wobei X und Y unabhängig sind.

Wenn ich Ihnen sage, $X = 3$, lernen Sie nichts über Y , da die beiden Variablen unabhängig sind. Aber: Wenn ich Ihnen zuerst sage, $Z = 10$, und dann sage, $X = 3$, wissen Sie sofort, was Y ist ($Y = 7$).

Also: X und Y sind abhängig – gegeben Z : $X \perp\!\!\!\perp Y | Z$.

11.8.4 Durch Kontrollieren entsteht eine Verzerrung bei der Kollision

Abbildung 11.19 zeigt: Durch Kontrollieren entsteht eine Kollision, eine Scheinkorrelation zwischen den Ursachen.

Kontrollieren kann z.B. bedeuten:

- *Stratifizieren*: Aufteilen von `date` in zwei Gruppen und dann Analyse des Zusammenhangs von `talent` und `looks` in jeder Teilgruppe von `date`
- *Kontrollieren mit Regression*: Durch Aufnahme von `date` als Prädiktor in eine Regression zusätzlich zu `looks` mit `talent` als Prädiktor

Ohne Kontrolle von `date` entsteht *keine* Scheinkorrelation zwischen `Looks` und `Talent`. Der Pfad (“Fluss”) von `Looks` über `date` nach `Talent` ist blockiert.

Kontrolliert man `date`, so *öffnet* sich der Pfad `Looks` → `date` → `talent` und die Scheinkorrelation entsteht: Der Pfad ist nicht mehr “blockiert”, die Korrelation kann “fließen” – was sie hier nicht soll, denn es handelt sich um Scheinkorrelation.

Das Kontrollieren von `date` geht zumeist durch Bilden einer Auswahl einer Teilgruppe von sich.

11.8.5 IQ, Fleiss und Eignung fürs Studium

Sagen wir, über die *Eignung* für ein Studium würden nur (die individuellen Ausprägungen) von Intelligenz (IQ) und Fleiss entscheiden, s. den DAG in Abbildung 11.21.

Bei positiver `eignung` wird ein Studium aufgenommen (`studium = 1`) ansonsten nicht (`studium = 0`).

Quelle

`eignung` (fürs Studium) sei definiert als die Summe von `iq` und `fleiss`, plus etwas Glück:

```
set.seed(42) # Reproduzierbarkeit
N <- 1e03
```

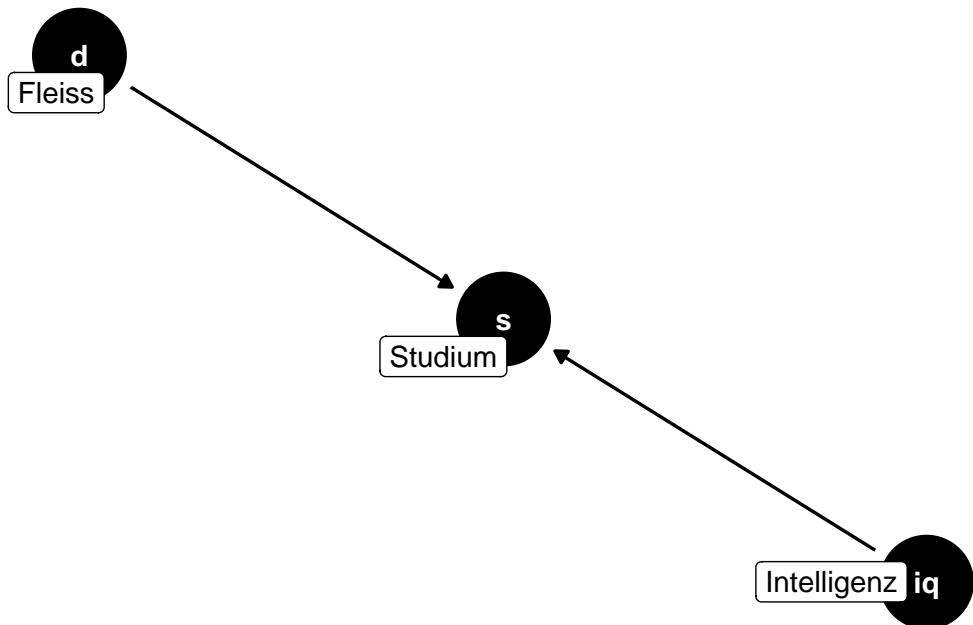


Abbildung 11.21: Kollisionsstruktur im Dag zur Studiumseignung

```

d_eignung <-
tibble(
  iq = rnorm(N), # normalverteilt mit MW=0, sd=1
  fleiss = rnorm(N),
  glueck = rnorm(N, mean = 0, sd = .1),
  eignung = 1/2 * iq + 1/2 * fleiss + glueck,
  # nur wer geeignet ist, studiert (in unserem Modell):
  studium = ifelse(eignung > 0, 1, 0)
)
  
```

Laut unserem Modell setzt sich Eignung zur Hälfte aus Intelligenz und zur Hälfte aus Fleiß zusammen, plus etwas Glück.

11.8.6 Schlagzeile “Schlauheit macht Studenten faul!”

Eine Studie untersucht den Zusammenhang von Intelligenz (iq) und Fleiß (f) bei Studenten (s).

Ergebnis: Ein *negativer* Zusammenhang!?

Berechnen wir das “Eignungsmodell”, aber nur mit Studis (`studium == 1`), s. `?@tbl-m-eignung`.

```

m_eignung <-
  stan_glm(iq ~ fleiss, data = d_eignung %>% filter(studium == 1), refresh = 0)

hdi(m_eignung)
## Highest Density Interval
##
## Parameter / 95% HDI
## -----
## (Intercept) / [ 0.70,  0.86]
## fleiss      / [-0.53, -0.36]

```

Abbildung 11.22 zeigt das Modell und die Daten.

Nativer Zusammenhang von Fleiss und IQ bei Studenten

Macht Fleiss blöd?

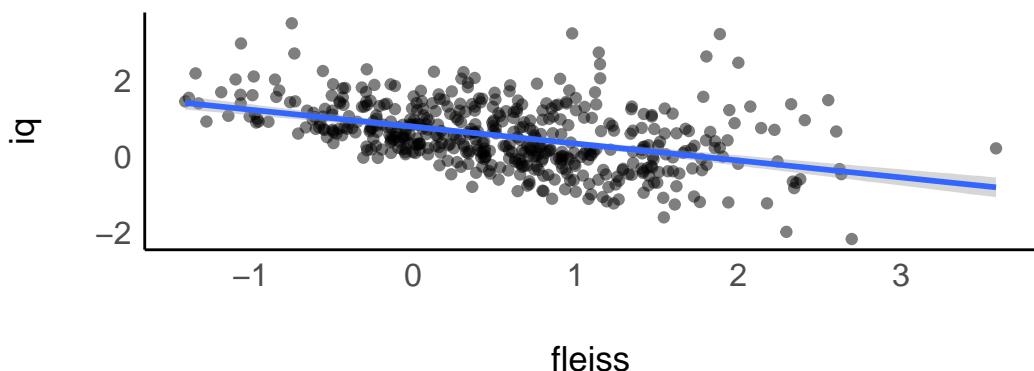


Abbildung 11.22: Der Zusammenhang von Fleiss und IQ

IQ ist *nicht* unabhängig von Fleiß in unseren Daten, sondern abhängig.

Nichtwissenschaftliche Berichte, etwa in einigen Medien, greifen gerne Befunde über Zusammenhänge auf und interpretieren die Zusammenhänge - oft vorschnell - als kausal.⁴

11.8.7 Kollisionsverzerrung nur bei Stratifizierung

Nur durch das Stratifizieren (Aufteilen in Subgruppen, Kontrollieren, Adjustieren) tritt die Scheinkorrelation auf, s. Abbildung 11.23.

i Hinweis

Ohne Stratifizierung tritt keine Scheinkorrelation auf. Mit Stratifizierung tritt Scheinkorrelation auf.

⁴Ehrlicherweise muss man zugeben, dass auch wissenschaftliche Berichte Daten über Zusammenhänge gerne kausal interpretieren, oft vorschnell.

Kein Stratifizierung, keine Scheinkorrelation

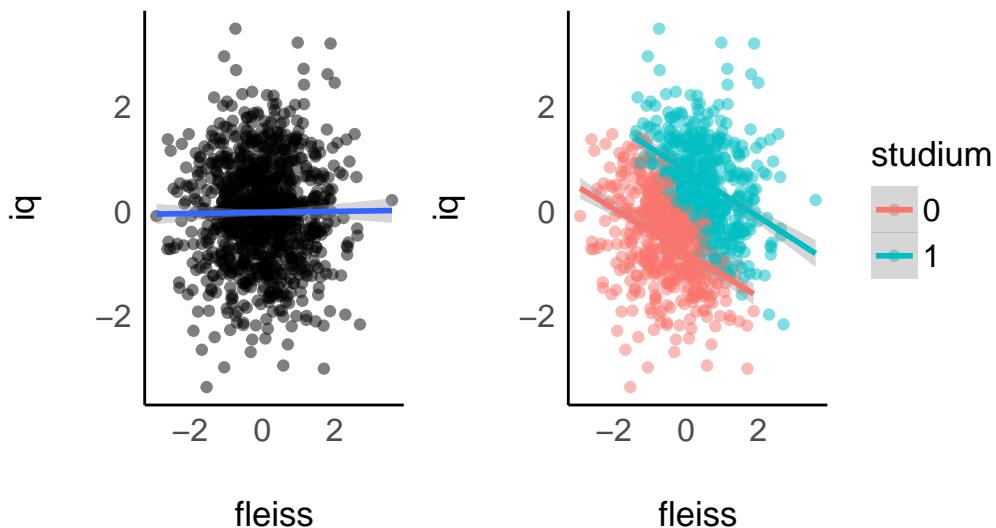


Abbildung 11.23: Stratifizierung und Scheinkorrelation

Wildes Kontrollieren einer Variablen - Aufnehmen in die Regression - kann genauso ut schaden wie nützen.

Nur Kenntnis des DAGs verrät die richtige Entscheidung: ob man eine Variable kontrolliert oder nicht.

i Hinweis

Nimmt man eine Variable als zweiten Prädiktor auf, so “kontrolliert” man diese Variable. Das Regressionsgewicht des ersten Prädiktors wird “bereinigt” um den Einfluss des zweiten Prädiktors; insofern ist der zweite Prädiktor dann “kontrolliert”.

11.8.8 Einfluss von Großeltern und Eltern auf Kinder

Wir wollen hier den (kausalen) Einfluss der Eltern E und Großeltern G auf den *Bildungserfolg* der Kinder K untersuchen.

Wir nehmen folgende Effekte an:

- indirekter Effekt von G auf K: $G \rightarrow E \rightarrow K$
- direkter Effekt von E auf K: $E \rightarrow K$
- direkter Effekt von G auf K: $G \rightarrow K$

Wir sind v.a. interessiert an $G \rightarrow K$, dem *direkten kausalen* Effekt von Großeltern auf ihre Enkel, s. Abbildung 11.24, Kurz (2021).

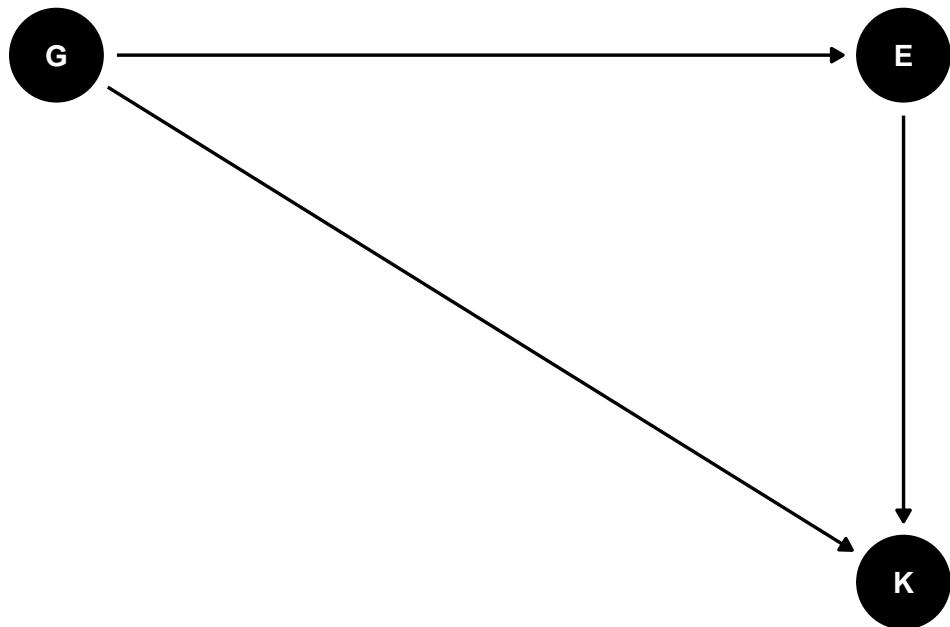


Abbildung 11.24: Der kausale Effekt von Großeltern auf Enkel. Ein verlorener Fall, zumindest was den DAG betrifft

Aber was ist, wenn wir vielleicht eine unbekannte Variable übersehen haben? (S. nächster Abschnitt).

11.9 Vertiefung

VERTIEFUNG - nicht prüfungsrelevant

11.9.1 Der Gespenster-DAG

Es gibt “unheilbare” DAGs, nennen wir sie “Gespenster-DAGs”, in denen es nicht möglich ist, einen (unverzerrten) Kausaleffekt zu bestimmen, s. Abbildung 11.25. Letztlich sagt uns der DAG bzw. unsere Analyse zum DAG: “Deine Theorie ist nicht gut, zurück an den Schreibtisch und denk noch mal gut nach. Oder sammle mehr Daten.”

- U könnte ein ungemessener Einfluss sein, der auf E und K wirkt, etwa *Nachbarschaft*.
- Die Großeltern wohnen woanders (in Spanien), daher wirkt die Nachbarschaft der Eltern und Kinder nicht auf sie.
- E ist sowohl für G als auch für U eine Wirkung, also eine Kollisionsvariable auf diesem Pfad.

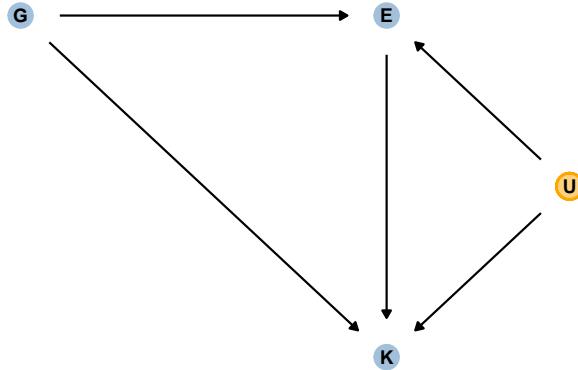


Abbildung 11.25: Der Gespenster-DAG: Eine Identifikation der Kausaleffekt ist nicht (vollständig) möglich.

- Wenn wir E kontrollieren, wird es den Pfad $G \rightarrow K$ verzerrn, auch wenn wir niemals U messen.

Die Sache ist in diesem Fall chancenlos. Wir müssen diesen DAG verloren geben, McElreath (2020), S. 180.

11.10 Die Hintertür schließen

11.10.1 Zur Erinnerung: Konfundierung

Forschungsfrage: Wie groß ist der (kausale) Einfluss der Schlafzimmerzahl auf den Verkaufspreis des Hauses?

a: livingArea, b: bedrooms, p: prize

UV: b, AV: p

Das Kausalmodell ist in Abbildung 11.26 dargestellt.

Im Regressionsmodell $p \sim b$ wird der kausale Effekt verzerrt sein durch die Konfundierung mit a. Der Grund für die Konfundierung sind die zwei Pfade zwischen b und p:

1. $b \rightarrow p$
2. $b \rightarrow a \rightarrow p$

Beide Pfade erzeugen (statistische) Assoziation zwischen b und p. Aber nur der erste Pfad ist kausal; der zweite ist nichtkausal. Gäbe es nur nur den zweiten Pfad und wir würden b ändern, so würde sich p nicht ändern.

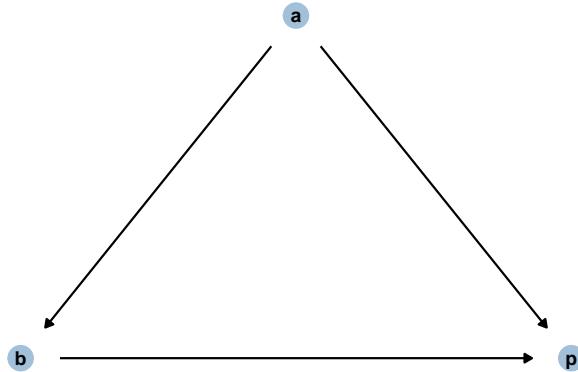


Abbildung 11.26: Der Preis wird sowohl von der Zimmerzahl als auch der Wohnfläche beeinflusst

11.10.2 Gute Experimente zeigen den echten kausalen Effekt

Abbildung 11.27 zeigt eine erfreuliche Situation: Die “Hintertür” zu unserer UV (Zimmerzahl) ist geschlossen!

Ist die Hintertür geschlossen - führen also keine Pfeile in unserer UV - so kann eine Konfondierung ausgeschlossen werden.

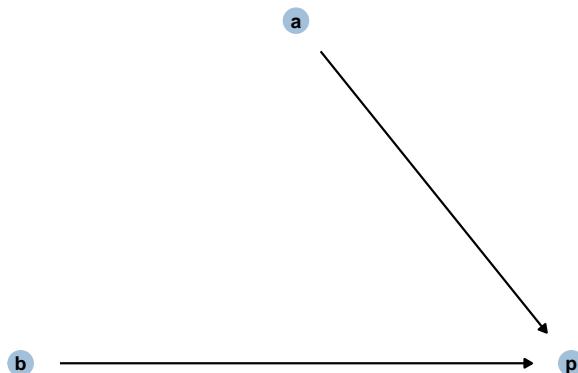


Abbildung 11.27: Unverzerrte Schätzung des kausalen Effekts unserer UV (Zimmerzahl).

Das Regressionsgewicht ist hier der unverzerrte Kausaleffekt. Es spielt keine Rolle, ob der andere Prädiktor im Modell enthalten ist. Da die beiden Prädiktoren unkorreliert sind, hat die Aufnahme des einen Prädiktors keinen Einfluss auf das Regressionsgewicht des anderen.

Die “Hintertür” der UV (b) ist jetzt zu! Der einzige verbleibende, erste Pfad ist der kausale Pfad und die Assoziation zwischen b und p ist jetzt komplett kausal.

Eine berühmte Lösung, den kausalen Pfad zu isolieren, ist ein (randomisiertes, kontrolliertes) Experiment. Wenn wir den Häusern zufällig (randomisiert) eine Anzahl von Schlafzimmern (b) zuweisen könnten (unabhängig von ihrer Quadratmeterzahl, a), würde sich der Graph so ändern. Das Experiment entfernt den Einfluss von a auf b. Wenn wir selber die Werte

von b einstellen im Rahmen des Experiments, so kann a keine Wirkung auf b haben. Damit wird der zweite Pfad, $b \rightarrow a \rightarrow p$ geschlossen (“blockiert”).

11.10.3 Hintertür schließen auch ohne Experimente

Konfundierende Pfade zu blockieren zwischen der UV und der AV nennt man auch *die Hintertür schließen* (backdoor criterion).

Wir wollen die Hintertüre schließen, da wir sonst nicht den wahren, kausalen Effekt bestimmen können.

Zum Glück gibt es neben Experimenten noch andere Wege, die Hintertür zu schließen, wie die Konfundierungsvariable a in eine Regression mit aufzunehmen.

Warum blockt das Kontrollieren von a den Pfad $b \leftarrow a \rightarrow p$? Stellen Sie sich den Pfad als eigenen Modell vor. Sobald Sie a kennen, bringt Ihnen Kenntnis über b kein zusätzliches Wissen über p . Wissen Sie hingegen nichts über a , lernen Sie bei Kenntnis von b auch etwas über p . Konditionieren ist wie “gegeben, dass Sie a schon kennen...”.

$$b \perp\!\!\!\perp p \mid a$$

11.10.4 Die vier Atome der Kausalanalyse

Abbildung 11.28 stellt die vier “Atome” der Kausalinferenz dar. Mehr gibt es nicht! Kennen Sie diese vier Grundbausteine, so können Sie jedes beliebige Kausalsystem (DAG) entschlüsseln.

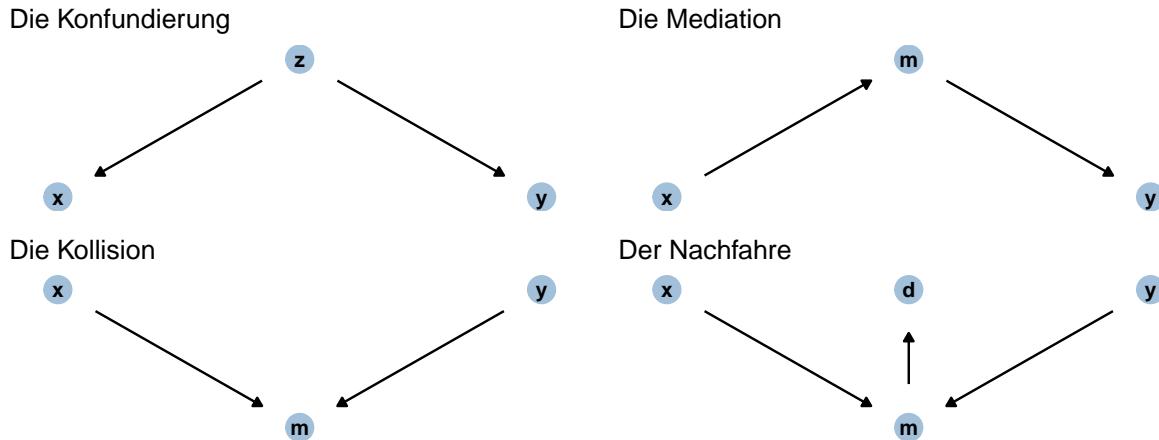


Abbildung 11.28: Die vier Atome der Kausalinferenz

11.10.5 Mediation

Die *Mediation* (Wirkkette, Rohr, Kette, chain) beschreibt Pfade, in der die Kanten gleiche Wirkrichtung haben: $x \rightarrow m \rightarrow y$. Anders gesagt: Eine Mediation, auch "Kette" oder "Wirkkette" genannt, ist eine Kausalanfolge der Art $x \rightarrow m \rightarrow y$, s. Abbildung 11.29. Die Variable in der Mitte m der Kette wird auch *Mediator* genannt, weil sie die Wirkung von X auf Y "vermittelt" oder übeträgt. Die Erforschung von Mediation spielt eine recht wichtige Rolle in einigen Wissenschaften, wie der Psychologie.

Die Mediation

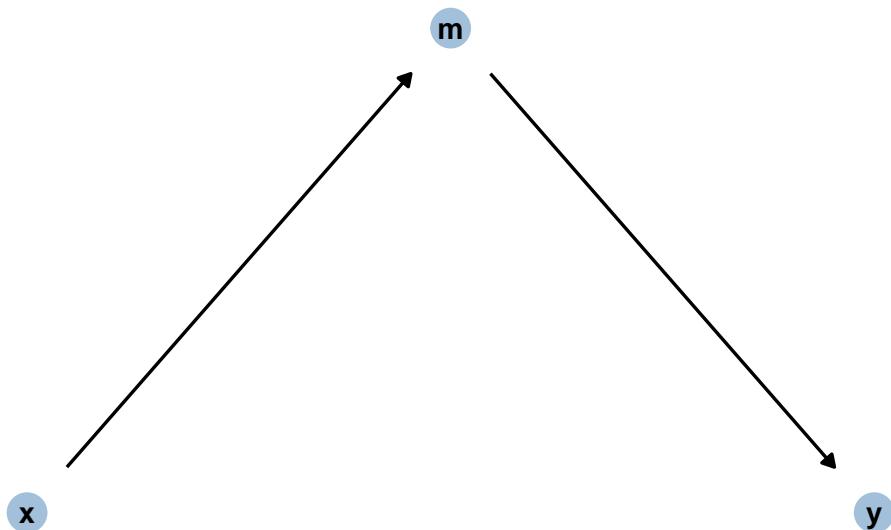


Abbildung 11.29: Das Kausalmodell der Mediation.

Ohne Kontrollieren ist der Pfad offen: Die Assoziation "fließt" den Pfad entlang (in beide Richtungen). Kontrollieren blockt (schließt) die Kette (genau wie bei der Gabel).

11.11 Der Nachfahre

Ein *Nachfahre* (descendent) ist eine Variable die von einer anderen Variable beeinflusst wird, s. fig-dag-nachfahre. Kontrolliert man einen Nachfahren d , so kontrolliert man damit zum Teil den Vorfahren (die Ursache), m . Der Grund ist, dass d Information beinhaltet über m . Hier wird das Kontrollieren von d den Pfad von x nach y teilweise öffnen, da m eine Kollisionsvariable ist.

11.11.1 Kochrezept zur Analyse von DAGs

Wie kompliziert ein DAG auch aussehen mag, er ist immer aus diesen vier Atomen aufgebaut.

Der Nachfahre

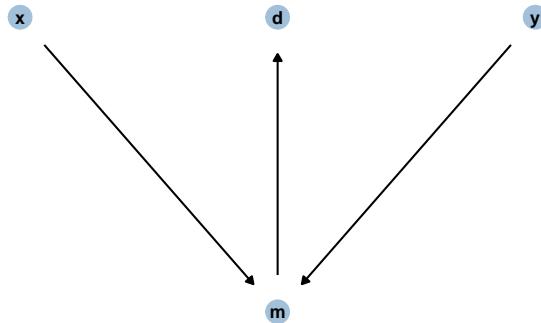


Abbildung 11.30: Ein Nachfahre verhält sich ähnlich wie sein Vorfahre...

Hier ist ein Rezept, das garantiert, dass Sie welche Variablen Sie kontrollieren sollten und welche nicht:

1. Listen Sie alle Pfade von UV (X) zu AV (Y) auf.
2. Beurteilen Sie jeden Pfad, ob er gerade geschlossen oder geöffnet ist.
3. Beurteilen Sie für jeden Pfad, ob er ein Hintertürpfad ist (Hintertürpfade haben einen Pfeil, der zur UV führt).
4. Wenn es geöffnete Hinterpfade gibt, prüfen Sie, welche Variablen man kontrollieren muss, um den Pfad zu schließen (falls möglich).

11.12 Schließen Sie die Hintertür (wenn möglich)!, bsp1

UV: X , AV: Y , drei Covariaten (A, B, C) und ein ungemessene Variable, U

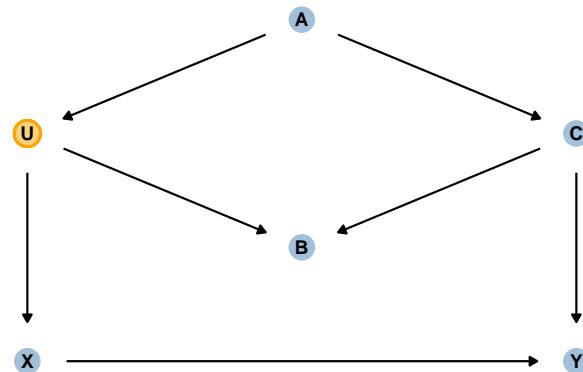


Abbildung 11.31: Puh, ein schon recht komplizierter DAG

Es gibt zwei Hintertürpfade in Abbildung 11.31:

1. $X \leftarrow U \leftarrow A \rightarrow C \rightarrow Y$, offen
2. $X \leftarrow U \rightarrow B \leftarrow C \rightarrow Y$, geschlossen

Kontrollieren von A oder (auch) C schließt die offene Hintertür.

McElreath (2020), Kurz (2021), s.S. 186.

11.12.1 Schließen Sie die Hintertür (wenn möglich)!, bsp2

S. DAG in Abbildung 11.32: UV: W , AV: D

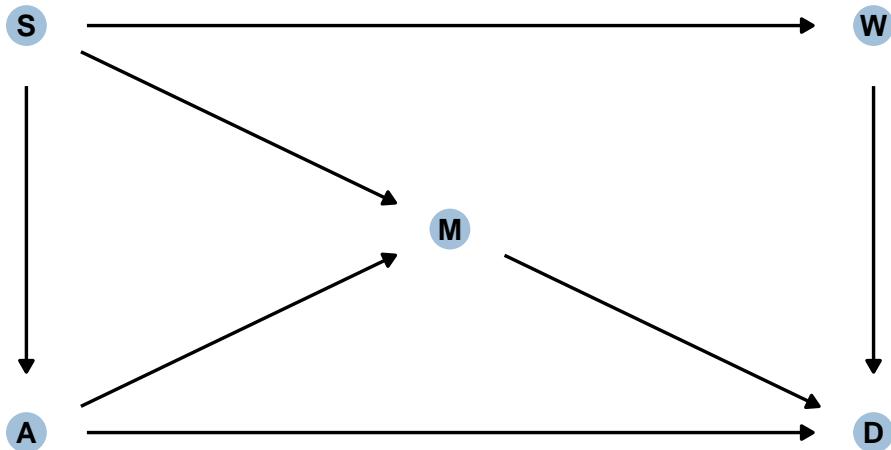


Abbildung 11.32: Welche Variablen muss man kontrollieren, um den Effekt von W auf D zu bestimmen?

Kontrollieren Sie diese Variablen, um die offenen Hintertüren zu schließen:

- entweder A und M
- oder S

[Mehr Infos](#)

Details finden sich bei McElreath (2020) oder Kurz (2021), ,S. 188.

11.12.2 Implizierte bedingte Unabhängigkeiten von bsp2

Ein Graph ohne Us ist eine starke - oft zu starke (unrealistisch optimistische) - Annahme. Auch wenn die Daten nicht sagen können, welcher DAG der richtige ist, können wir zumindest lernen, welcher DAG falsch ist. Die vom Modell implizierten bedingten Unabhängigkeiten geben uns Möglichkeiten, zu prüfen, ob wir einen DAG verwerfen (ausschließen) können. Bedingte Unabhängigkeit zwischen zwei Variablen sind Variablen, die nicht assoziiert (also stochastisch unabhängig) sind, wenn wir eine bestimmte Menge an Drittvariablen kontrollieren.

bsp2 impliziert folgende bedingte Unabhängigkeiten:

```
## A _||_ W | S  
## D _||_ S | A, M, W  
## M _||_ W | S
```

11.13 Fazit

Wie (und sogar ob) Sie statistische Ergebnisse (z.B. eines Regressionsmodells) interpretieren können, hängt von der *epistemologischen Zielrichtung* der Forschungsfrage ab:

- Bei *deskriptiven* Forschungsfragen können die Ergebnisse (z.B. Regressionskoeffizienten) direkt interpretiert werden. Z.B. “Der Unterschied zwischen beiden Gruppen beträgt etwa ...”. Allerdings ist eine kausale Interpretation nicht zulässig.
- Bei *prognostischen* Fragestellungen (Vorhersagen) spielen die Modellkoeffizienten keine Rolle, stattdessen geht es um vorhergesagten Werte, \hat{y}_i , z.B. auf Basis der PPV. Kausalaussagen sind zwar nicht möglich, aber auch nicht von Interesse.
- Bei *kausalen* Forschungsfragen dürfen die Modellkoeffizienten nur auf Basis eines Kausalmodells (DAG) oder eines (gut gemachten) Experiments interpretiert werden.

Modellkoeffizienten ändern sich (oft), wenn man Prädiktoren zum Modell hinzufügt oder wegnimmt. Entgegen der verbreiteten Annahme ist es falsch, möglichst viele Prädiktoren in das Modell aufzunehmen, wenn das Ziel eine Kausalaussage ist. Kenntnis der “kausalen Atome” ist Voraussetzung zur Ableitung von Kausalschlüsse in Beobachtungsstudien.

11.14 Aufgaben

- Sammlung “kausal”

12 Abschluss

12.1 Lernsteuerung

12.1.1 Lernziele

Nach Absolvieren des jeweiligen Kapitels sollen folgende Lernziele erreicht sein.

Sie können ...

- erläutern, wie Sie eine typische, sozialwissenschaftliche Forschungsfrage (quantitativ) untersuchen
- typische “Lieblingsfehler” benennen und Wege aufzeigen, um die Fehler zu umgehen
- zwischen den Grundkonzepten der Frequentistischen Statistik und der Bayes-Statistik übersetzen
- die Grundideen der Bayes-Statistik in eine Gesamtzusammenhang einordnen

12.1.2 Benötigte R-Pakete

In diesem Kapitel benötigen Sie folgende R-Pakete.

```
library(tidyverse)
library(easystats)
library(rstanarm)
```

12.2 Lieblingsfehler

Lieblingsfehler im Überblick :

1. Post-Präd-Verteilung (PPV) und Post-Verteilung verwechseln
2. Quantile und Verteilungsfunktion verwechseln
3. Prädiktoren nicht zentrieren, wenn es einen Interaktionsterm gibt
4. Interaktion falsch interpretieren
5. Regressionskoeffizienten kausal interpretieren, wenn es keine kausale Fundierung gibt

Tabelle 12.1: ?(caption)

12.3 Post-Präd-Verteilung (PPV) und Post-Verteilung verwechseln

Berechnen wir das Standard-mtcars-Modell: `mpg ~ hp`.

```
m1 <- stan_glm(mpg ~ hp, data = mtcars, refresh = 0)
```

Die *Post-Verteilung* zeigt Stichproben zu den Parameterwerten, s. [?@tbl-post-m1](#).

```
post_verteilung <- m1 %>%
  as_tibble()
head(post_verteilung)
## # A tibble: 6 x 3
##   `(Intercept)`    hp sigma
##   <dbl>      <dbl> <dbl>
## 1 29.4     -0.0653  3.68
## 2 30.6     -0.0693  3.90
## 3 29.4     -0.0662  3.96
## 4 30.1     -0.0691  4.46
## 5 30.4     -0.0689  3.37
## 6 29.6     -0.0644  3.34
```

Diese Tabelle kann man hernehmen, um Fragen zu Post-Verteilung zu beantworten. Häufig ist es aber bequemer, z.B. mit `parameters(m1)` Post-Intervalle und Punktschätzer auszulesen.

Die *Posterior-Prädiktiv-Verteilung (PPV)* zeigt die Vorhersagen, also keine Parameterwerte, sondern Beobachtungen.

```
## # A tibble: 5 x 2
##   name          value
##   <chr>        <dbl>
## 1 Mazda RX4     25.2
## 2 Mazda RX4 Wag 17.7
## 3 Datsun 710    23.9
## 4 Hornet 4 Drive 24.1
## 5 Hornet Sportabout 22.2
```

12.4 Quantile und Verteilungsfunktion verwechseln

12.4.1 Quantil für p

Ein p -Quantil teilt eine Verteilung in zwei Teile, und zwar so, dass mind. p kleiner oder gleich dem p -Quantil sind. s. Abbildung 12.1.

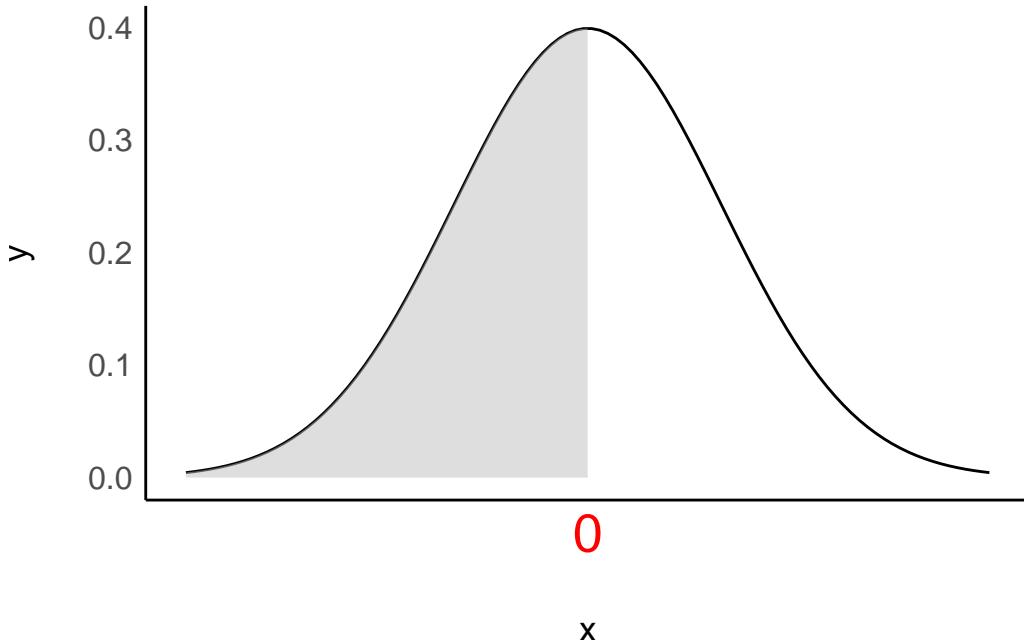


Abbildung 12.1: 50%-Quantil

Das 50%-Quantil (.5-Quantil) beträgt $x = 0$. Mind ein Anteil $1 - p$ ist größer oder gleich p -Quantil.

12.4.2 Verteilungsfunktion F

$F(x)$ gibt die Wahrscheinlichkeit an der Stelle x an, dass X einen Wert kleiner oder gleich x annimmt, s. Abbildung 12.2.

$F(0) = 1/2$, die Wahrscheinlichkeit beträgt hier 50%, dass x nicht größer ist als 0.

12.5 Interaktion falsch interpretieren

Berechnen wir ein einfaches Interaktionsmodell: `mpg ~ hp*vs`.

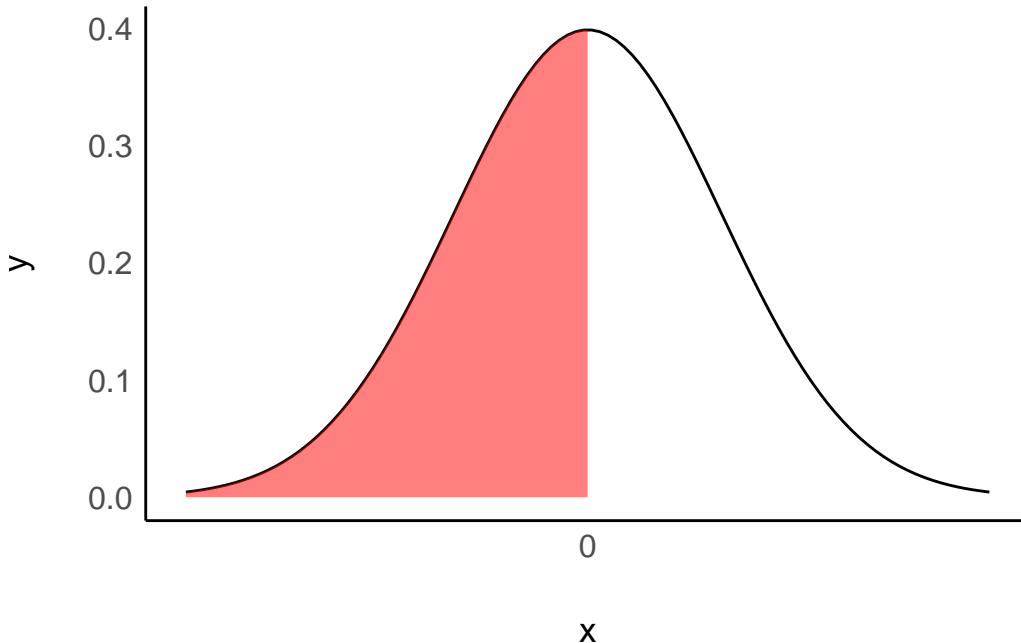


Abbildung 12.2: Verteilungsfunktion $F(x=0)=1/2$

i Hinweis

Zur Erinnerung: `mpg ~ hp*vs` ist synonym zu (aber kürzer als) `mpg ~ hp + vs + hp:vs`.

```
m2 <- stan_glm(mpg ~ hp*vs, data = mtcars) # mit Interaktionseffekt
```

Modellkoeffizienten, s. `?@tbl-m2-params`.

```
parameters(m2)
```

Tabelle 12.2: Parameter von m2

Parameter	Median	95% CI	pd	Rhat	ESS	Prior
(Intercept)	24.70	(19.27, 30.33)	100%	1.003	2363.00	Normal (20.09 +- 15.07)
hp	-0.04	(-0.07, -0.01)	99.88%	1.003	2380.00	Normal (0.00 +- 0.22)
vs	14.08	(4.58, 22.97)	99.85%	1.003	1772.00	Normal (0.00 +- 29.89)
hp:vs	-0.11	(-0.19, -0.03)	99.52%	1.002	1976.00	Normal (0.00 +- 0.31)

`?@tbl-m2-params` zeigt die Visualisierung der Parameter von m2.

```
plot(parameters(m2))
```

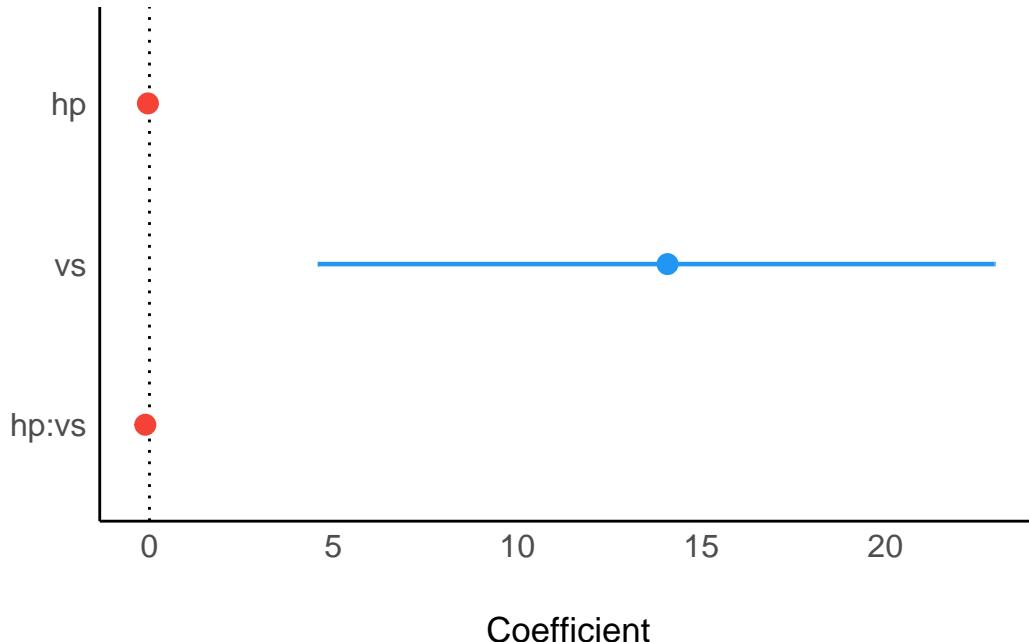


Abbildung 12.3: Parameter von m2 visualisiert

Falsch Der Unterschied im Verbrauch zwischen den beiden Gruppen $vs=0$ und $vs=1$ beträgt ca. -0.11.

Richtig Der Unterschied im Verbrauch zwischen den beiden Gruppen $vs=0$ und $vs=1$ beträgt ca. -0.11 – wenn $hp=0$.

Da $hp=0$ kein realistischer Wert ist, ist das Modell schwer zu interpretieren. Zentrierte Prädiktoren wären hier eine sinnvolle Lösung.

Gelman, Hill, und Vehtari (2021), Kap. 10, McElreath (2020), Kap. 8

12.6 Kochrezepte

12.6.1 Kochrezept: Forschungsfrage untersuchen

Theoretische Phase 1. Staunen über ein Phänomen, y , Kausalfrage finden 2. Literatur wälzen, um mögliche Ursachen x von y zu lernen 3. Forschungsfrage, Hypothese präzisieren 4. Modell präzisieren (DAG(s), Prioris)

Empirische Phase

5. Versuch planen
6. Daten erheben

Analytische Phase

7. Daten aufbereiten
8. Modell berechnen anhand eines oder mehrerer DAGs
9. Modell prüfen/kritisieren
10. Forschungsfrage beantworten

Yeah! Fertig.

12.6.2 Parameter schätzen vs. Hypothesen prüfen

Quantitative Studien haben oft einen von zwei (formalen) Zielen: Hypothesen testen oder Parameter schätzen. Beispiel Hypothesenprüfung: "Frauen parken im Durchschnitt schneller ein als Männer". Beispiel Parameterschätzung: "Wie groß ist der mittlere Unterschied in der Ausparkzeit zwischen Frauen und Männern?"

Je ausgereifter ein Forschungsfeld, desto *kühnere* Hypothesen lassen sich formulieren: - stark ausgereift: - Die nächste totale Sonnenfinsternis in Deutschland wird am 27.7.2082 um 14.47h stattfinden, [Quelle](#) - gering ausgereift: - Die nächste Sonnenfinsternis wird in den nächsten 100 Jahren stattfinden. - Lernen bringt mehr als Nicht-Lernen für den Klausurerfolg. Kühne Hypothesen sind wünschenswert

12.6.3 Formalisierung von Forschungsfragen

Der Mittelwert in Gruppe A ist höher als in Gruppe B (der Unterschied, d , im Mittelwert ist größer als Null):

$$\mu_1 > \mu_2 \Leftrightarrow \mu_1 - \mu_2 > 0 \Leftrightarrow \mu_d > 0$$

12.7 Kerngedanken Bayes

12.7.1 Zentraler Kennwert der Bayes-Statistik: Post-Verteilung

Berechnen wir wieder ein einfaches¹ Modell: `mpg ~ hp`.

```
m3 <- stan_glm(mpg ~ hp, data = mtcars)
```

Und schauen wir uns die Post-Verteilung an, mit eingezeichnetem HDI, s. Abbildung [12.4](#).

Ein Zusammenfassen der Posterior-Verteilung (z.B. zu einem 95%-PI) ist möglich und oft sinnvoll. Verschiedene Arten des Zusammenfassens der Post-Verteilung sind möglich, z.B. zu Mittelwert oder SD oder einem einem HD-Intervall. Allerdings übermittelt nur die gesamte

¹langweiliges

Highest Density Interval (HDI)

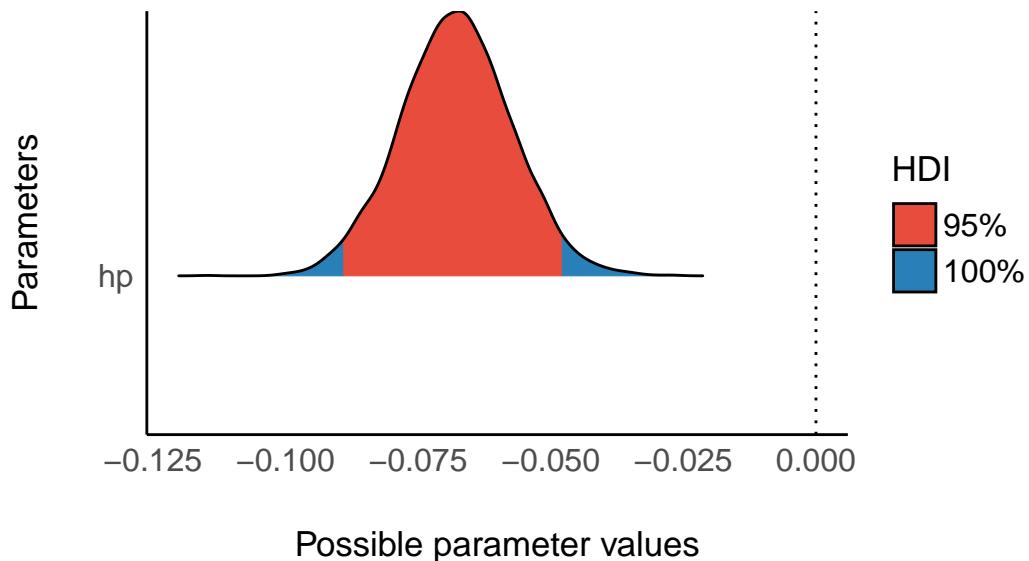


Abbildung 12.4: Post-Verteilung (HDI) von $m3$

Post-Verteilung alle Informationen. Daher empfiehlt es sich (oft), die Post-Verteilung zu visualisieren.

12.7.2 Posteriori als Produkt von Priori und Likelihood

$$\text{Posteriori} = \frac{\text{Likelihood} \times \text{Priori}}{\text{Evidenz}}$$

12.8 Beispiele für Prüfungsaufgaben

12.8.1 Geben Sie den korrekten Begriff an!

Hans Anna Lise

Puh, wie erstelle ich für alle Studis ein anderes Rätsel²?

i Hinweis

In einer Open-Book-Prüfung bekommen alle Studenten eine eigene, jeweils andere Prüfung. Teamarbeit bleibt natürlich trotzdem untersagt.

²Fahr-Hier-Hans-Anna-Lise: Varianzanalyse

12.8.2 DAG mit doppelter Konfundierung

Puh, jetzt kommt ein wilder DAG, s. Abbildung 12.5.

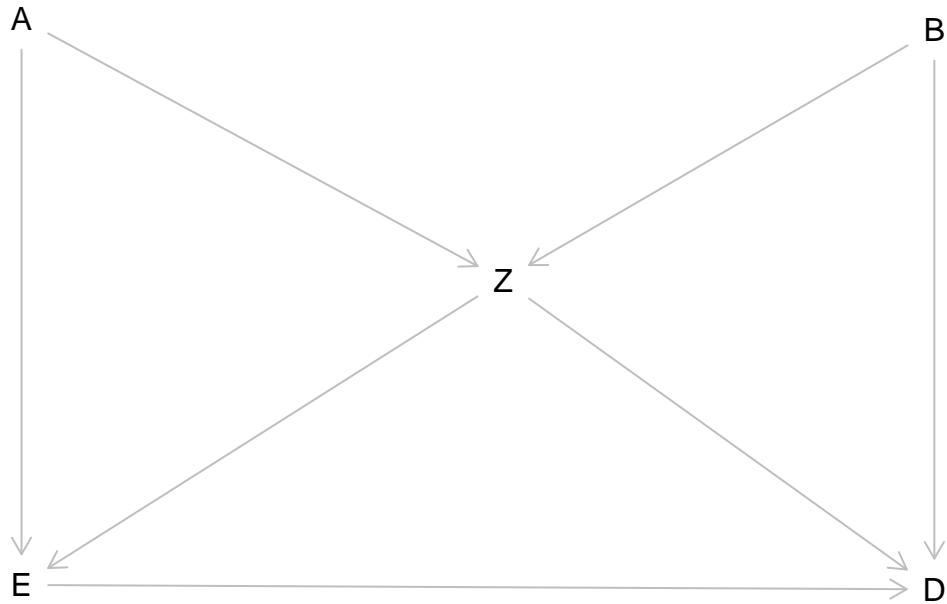


Abbildung 12.5: Ein DAG mit doppelter Konfundierung. Ist hier noch zu helfen?

Geben Sie die minimale Adjustierungsmenge (minimal adjustment set) an, um den totalen (gesamten) Effekt von E auf D zu bestimmen!

Entweder ist die Menge $\{A, Z\}$ zu adjustieren oder die Menge $\{B, Z\}$.

Ja, dem DAG ist zu helfen.

12.8.3 DAG mit vielen Variablen

Je nach dem wie komplex Ihre Theorie ist, ist Ihr DAG auch komplex, s. Abbildung 12.6.

Minimale Adjustierungsmenge, um den Effekt von E auf D zu identifizieren: $\{7\}, \{8\}$.

Trotz der vielen Variablen, ist der kausale Effekt von E auf D recht gut zu identifizieren.

12.8.4 Ein Kausalmodell der Schizophrenie, van Kampen (2014)

The SSQ model of schizophrenic prodromal unfolding revised:

An analysis of its causal chains based on the language of directed graphs

D. van Kampen

Lesen Sie [hier den Abstract](#).

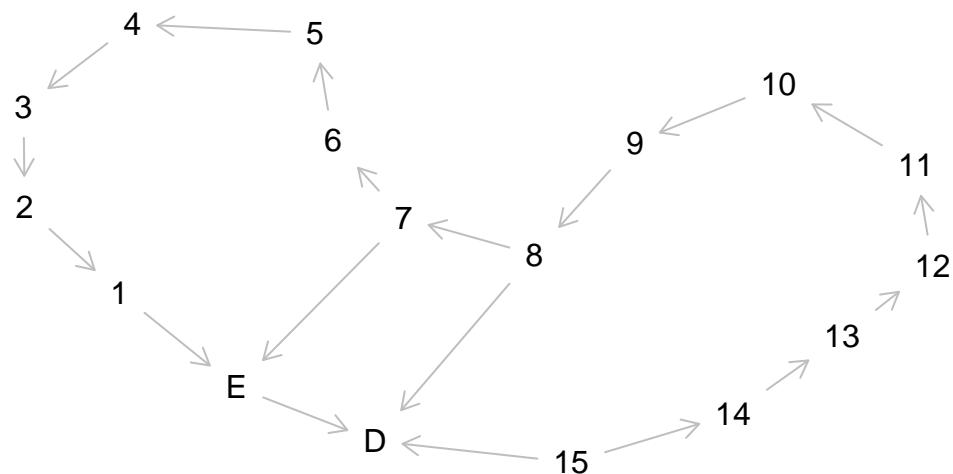


Abbildung 12.6: Ein DAG mit vielen Variablen

Folgende Symptome der Schizophrenie wurden gemessen:

Social Anxiety (*SAN*), Active Isolation (*AIS*), Affective Flattening (*AFF*), Suspiciousness (*SUS*), Egocentrism (*EGC*), Living in a Fantasy World (*FTW*), Alienation (*ALN*), Apathy (*APA*), Hostility (*HOS*), Cognitive Derailment (*CDR*), Perceptual Aberrations (*PER*), and Delusional Thinking (*DET*)

Kampen (2014)

UV: *SUS*, AV: *EGC*

Berechnen Sie die minimale Adjustierungsmenge, um den kausalen Effekt der UV auf die AV zu identifizieren!

Abbildung 12.7 zeigt den DAG von van Kampen (2014) zu den Symptomen der Schizophrenie.

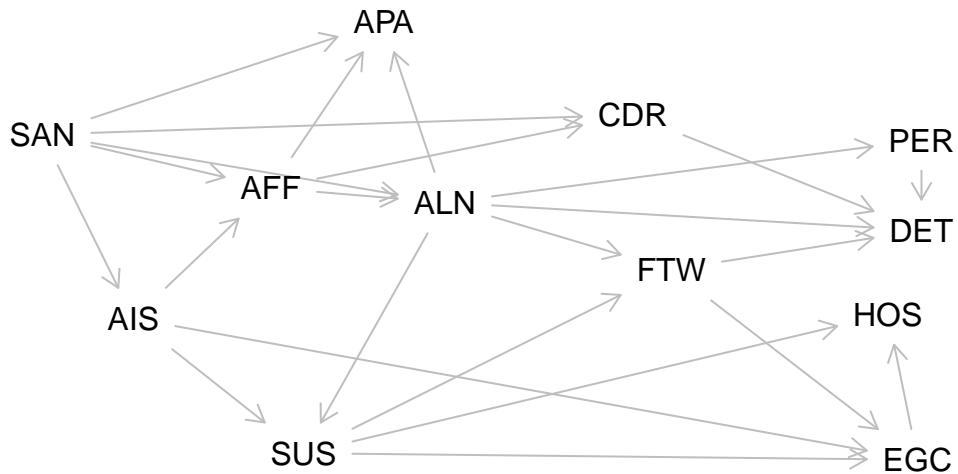


Abbildung 12.7: Ein DAG zu den Symptomen der Schizophrenie

Minimales Adjustment-Set für den totalen Kausaleffekt: {AIS, ALN}

12.8.5 Modelle berechnen

Stellen Sie sich auf Aufgaben ein, in denen Sie Modellparameter berechnen sollen. Orientieren Sie sich an den Aufgaben und Inhalten des Unterrichts.

Prüfungsfragen zu Modellen könnten z.B. sein:

- Geben Sie den Punktschätzer (Median) für den Prädiktor X im Modell Y an!
- Geben Sie ein 89%-HDI für den Parameter X im Modell Y an!
- Geben Sie R-Quadrat an.
- Formulieren Sie ein Interaktionsmodell!
- Welches Modell ist korrekt, um den kausalen Effekt zu modellieren?
- Formulieren Sie ein Modell mit folgenden Prioris ...
- Liegt der Effekt X noch im ROPE ?
- Unterscheidet sich die Breite des CI von der Breite des HDI für den Prädiktor X im Modell Y?
- Was verändert sich an den Parametern, wenn Sie die Prädiktoren zentrieren/z-standardisieren?
- ...

12.9 Viel Erfolg bei der Prüfung!

References

- Bourier, Günther. 2018. *Wahrscheinlichkeitsrechnung und schließende Statistik: praxisorientierte Einführung: mit Aufgaben und Lösungen*. 9., aktualisierte Auflage. Lehrbuch. Wiesbaden [Heidelberg]: Springer Gabler. <https://doi.org/10.1007/978-3-658-07481-4>.
- . 2022. *Statistik-Übungen: Beschreibende Statistik – Wahrscheinlichkeitsrechnung – Schließende Statistik*. 7. Auflage. Wiesbaden: Springer Gabler.
- Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Routledge. <http://dx.doi.org/10.4324/9780203771587>.
- Dai, Hengchen, Silvia Saccardo, Maria A. Han, Lily Roh, Naveen Raja, Sitaram Vangala, Hardikkumar Modi, Shital Pandya, Michael Sloyan, und Daniel M. Croymans. 2021. „Behavioural Nudges Increase COVID-19 Vaccinations“. *Nature* 597 (7876): 404–9. <https://doi.org/10.1038/s41586-021-03843-2>.
- Gelman, Andrew, Ben Goodrich, Jonah Gabry, und Aki Vehtari. 2019. „R-Squared for Bayesian Regression Models“. *The American Statistician* 73 (3): 307–9. <https://doi.org/10.1080/00031305.2018.1549100>.
- Gelman, Andrew, Jennifer Hill, und Aki Vehtari. 2021. *Regression and other stories*. Analytical methods for social research. Cambridge: Cambridge University Press.
- Goodrich, Ben, Jonah Gabry, Imad Ali, und Sam Brilleman. 2022. „rstanarm: Bayesian applied regression modeling via Stan.“ <https://mc-stan.org/rstanarm/>.
- Kampen, D. van. 2014. „The SSQ Model of Schizophrenic Prodromal Unfolding Revised: An Analysis of Its Causal Chains Based on the Language of Directed Graphs“. *European Psychiatry* 29 (7): 437–48. <https://doi.org/10.1016/j.eurpsy.2013.11.001>.
- Kruschke, John K. 2018. „Rejecting or Accepting Parameter Values in Bayesian Estimation“. *Advances in Methods and Practices in Psychological Science* 1 (2): 270–80. <https://doi.org/10.1177/2515245918771304>.
- Kurz, A. Solomon. 2021. *Statistical rethinking with brms, ggplot2, and the tidyverse: Second edition*. <https://bookdown.org/content/4857/>.
- Makowski, Dominique, Mattan S. Ben-Shachar, S. H. Annabel Chen, und Daniel Lüdecke. 2019. „Indices of Effect Existence and Significance in the Bayesian Framework“. *Frontiers in Psychology* 10: 2767. <https://doi.org/10.3389/fpsyg.2019.02767>.
- McElreath, Richard. 2020. *Statistical rethinking: a Bayesian course with examples in R and Stan*. 2. Aufl. CRC texts in statistical science. Boca Raton: Taylor; Francis, CRC Press.
- Messerli, Franz H. 2012. „Chocolate Consumption, Cognitive Function, and Nobel Laureates“. *New England Journal of Medicine* 367 (16): 1562–64. <https://doi.org/10.1056/NEJMMon1211064>.
- Nasreen, Sharifa, Hannah Chung, Siyi He, Kevin A. Brown, Jonathan B. Gubbay, Sarah A. Buchan, Deshayne B. Fell, u. a. 2021. „Effectiveness of mRNA and ChAdOx1 COVID-19 Vaccines Against Symptomatic SARS-CoV-2 Infection and Severe Outcomes with

- “Variants of Concern in Ontario“. <https://doi.org/10.1101/2021.06.28.21259420>.
- Pearl, Judea, Madelyn Glymour, und Nicholas P. Jewell. 2016. *Causal inference in statistics: a primer*. Chichester, West Sussex: Wiley.
- Poldrack, Russell. 2022. *Statistical Thinking for the 21st Century*. <https://statsthinking21.github.io/statsthinking21-core-site/index.html>.
- Pormohammad, Ali, Mohammad Zarei, Saeid Ghorbani, Mehdi Mohammadi, Mohammad Hossein Razizadeh, Diana L. Turner, und Raymond J. Turner. 2021. „Efficacy and Safety of COVID-19 Vaccines: A Systematic Review and Meta-Analysis of Randomized Clinical Trials“. *Vaccines* 9 (5): 467. <https://doi.org/10.3390/vaccines9050467>.
- Rohrer, Julia M. 2018. „Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data“. *Advances in Methods and Practices in Psychological Science* 1 (1): 27–42. <https://doi.org/10.1177/2515245917745629>.
- Thompson, Mark G., Edward Stenehjem, Shaun Grannis, Sarah W. Ball, Allison L. Naleway, Toan C. Ong, Malini B. DeSilva, u. a. 2021. „Effectiveness of Covid-19 Vaccines in Ambulatory and Inpatient Care Settings“. *New England Journal of Medicine* 385 (15): 1355–71. <https://doi.org/10.1056/NEJMoa2110362>.