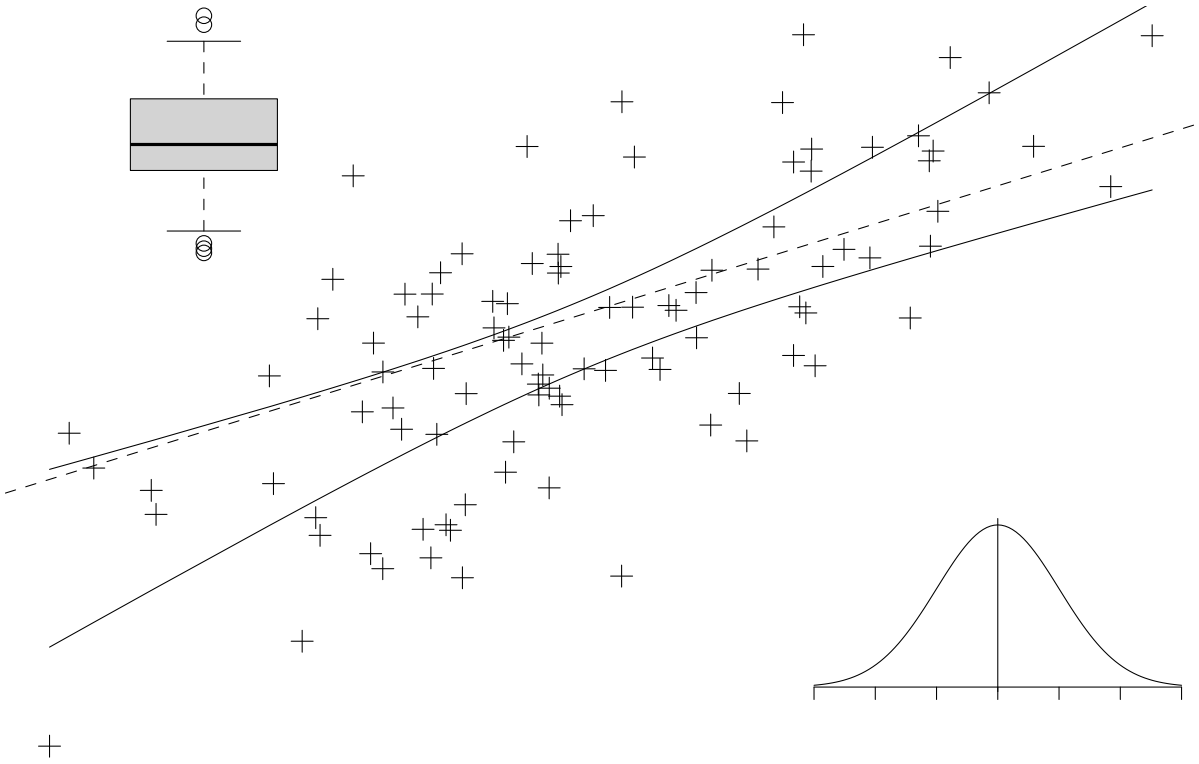
Empirische und Experimentelle Wirtschaftsforschung BW24.1



Oliver Kirchkamp

Inhaltsverzeichnis

1. Einführung — Schätzen von Parametern **8**
1.1. Ein Beispiel zur Motivation . . . 8
1.2. Ziele der Vorlesung . . . 9
1.3. Schätzer . . . 13
1.4. Schätzer als Zufallsvariablen . . . 17
1.5. Einige Punktschätzer . . . 17
1.6. Regression als ein weiterer Punktschätzer . . . 18
1.7. Literatur . . . 22
1.8. Schlüsselbegriffe . . . 23
1.A. Beispiele für die Vorlesung . . . 23
1.B. Übungen . . . 24

2. Wünschenswerte Eigenschaften von Schätzern **25**
2.1. Erwartungstreue (Unverzerrtheit) . . . 25
2.2. Mittlerer quadratischer Fehler . . . 30
2.3. Effizienz (Wirksamkeit) . . . 31
2.4. Konsistenz . . . 31
2.5. Eigenschaften von bekannten Schätzern . . . 32
2.5.1. $\bar{X}$ als Schätzer für den Erwartungswert . . . 32
2.5.2. Median als Schätzer . . . 33
2.5.3. Die Stichprobenvarianz . . . 35
2.6. Literatur . . . 36
2.7. Schlüsselbegriffe . . . 36
2.A. Beispiele für die Vorlesung . . . 36
2.B. Übungen . . . 37

3. Maximum Likelihood und Momentenmethode **40**
3.1. Motivation: Bietverhalten in Englischen Auktionen . . . 40
3.2. Maximum Likelihood Methode . . . 42
3.3. Likelihoodfunktion . . . 43
3.4. Log-Likelihood Funktion . . . 44
3.5. Momentenmethode . . . 46
3.6. Literatur . . . 47
3.7. Schlüsselbegriffe . . . 47
3.A. Beispiele für die Vorlesung . . . 48
3.B. Übungen . . . 49

4. Bayesianische Inferenz **53**
4.1. Einführung . . . 53
4.2. Satz von Bayes – Aggregation von Informationen . . . 56
4.3. Bayesianische Inferenz mit R . . . 58
4.4. Inferenz über das “credible interval” hinaus . . . 63

4.5. Vergleich von Gruppen 64
4.A. Beispiele für die Vorlesung 66
4.B. Übungen 67

5. Frequentistische Inferenz – Tests von Nullhypothesen 72
5.1. Motivation: Wählen Firmen im Oligopol Gleichgewichtsmengen? 72
5.2. Frequentistische Hypothesentests 73
5.3. Fehler 1. und 2. Art 75
5.4. Signifikanzniveau eines Tests 76
5.5. Parametertests — formale Definition 77
5.6. p-Wert eines Tests 78
5.7. Formulierung von Hypothesen 79
5.8. Grenzen 79
5.9. Literatur 81
5.10. Schlüsselbegriffe 82
5.A. Übungen 82

6. Tests für Mittelwerte — parametrisch 87
6.1. Motivation: Ultimatum Verhandlungen 87
6.2. - bei unbekannter Varianz 88
6.2.1. Test des Mittelwerts - p-Wert 88
6.2.2. - bei gegebenem Signifikanzniveau 90
6.3. Vergleich von zwei Stichproben mit unbekannter Varianz 93
6.3.1. Unverbundene Stichproben 93
6.3.2. Paarweise verbundene Stichproben 95
6.4. Literatur 95
6.5. Schlüsselbegriffe 95
6.A. Beispiele für die Vorlesung 96
6.B. Übungen 97

7. Konfidenzintervalle 101
7.1. Motivation: Effizienz von Märkten 101
7.2. Schätzen bei bekannter Varianz 102
7.3. Schätzen bei unbekannter Varianz 107
7.4. Signifikanzniveau / Konfidenzintervall 108
7.5. Simulation 110
7.6. Literatur 111
7.7. Schlüsselbegriffe 111
7.A. Beispiele für die Vorlesung 111
7.B. Übungen 112
7.C. Symmetrie des Konfidenzintervalls 120
7.D. Konfidenzintervall bei Binomialverteilung 123
7.E. Konfidenzintervall für die Varianz 126

8. Nichtparametrische Tests **128**
8.1. Motivation: Ökomische Erwartungen und Guessing Games . . . 128
8.2. Wilcoxon Test für paarweise Stichproben . . . 129
8.3. unverbundene Stichproben . . . 133
8.4. Motivation: Speed dating und mate copying . . . 135
8.5. Vergleich von Häufigkeiten . . . 137
8.6. Abhängigkeit von zwei Merkmalen . . . 138
8.7. Literatur . . . 140
8.8. Schlüsselbegriffe . . . 140
8.A. Beispiele für die Vorlesung . . . 141
8.B. Übungen . . . 143
8.C. Fisher's exakter Test . . . 150

9. Lineare Regression — Einführung **151**
9.1. Motivation . . . 151
9.2. Lineare Regression . . . 152
9.3. Bestimmtheitsmaß R^2 . . . 156
9.4. SER . . . 157
9.5. Die Verteilung des OLS Schätzers . . . 158
9.6. OLS Annahmen . . . 162
9.6.1. Verteilung von $\hat{\beta}_1$ . . . 162
9.6.2. Verteilung von $\hat{\beta}_0$ . . . 162
9.7. Hypothesentests für $\hat{\beta}_1$ . . . 163
9.8. Konfidenzintervalle für β . . . 164
9.9. Verwendung von OLS in der Praxis . . . 166
9.9.1. Darstellung von Schätzergebnissen . . . 166
9.9.2. Bayesianische Schätzung des linearen Modells . . . 167
9.9.3. Eigenschaften von OLS . . . 168
9.9.4. Erweiterte OLS Annahmen . . . 168
9.10. Literatur . . . 169
9.11. Schlüsselbegriffe . . . 169
9.A. Übungen . . . 170
9.B. Herleitung des OLS Schätzers . . . 172
9.C. Unverzerrtheit des Schätzers für $\hat{\beta}$ . . . 172
9.D. Varianz von $\hat{\beta}_1$ . . . 174

10. Multiple Regression **175**
10.1. Motivation . . . 175
10.2. Erweiterung des Beispiels aus Kapitel 9 . . . 176
10.3. Bayesianische Schätzung des linearen Modells . . . 178
10.4. Annahmen . . . 179
10.5. Verteilung des OLS Schätzers . . . 179
10.6. Die Verteilung von $\hat{\beta}$ . . . 180
10.6.1. Varianz von $\hat{\beta}$ . . . 180

10.7. Omitted variable bias . . . 183
10.7.1. Beispiele für Omitted-Variable-Bias: . . . 184
10.7.2. Erweiterung der Schätzgleichung . . . 185
10.8. Multikollinearität . . . 187
10.9. Spezifikationsfehler: Zusammenfassung . . . 189
10.10.Bayesianische Schätzung und Multikollinearität . . . 189
10.11.Literatur . . . 190
10.12.Schlüsselbegriffe . . . 191
10.A. Beispiele für die Vorlesung . . . 191
10.B. Übungen . . . 192
10.C. Omitted variable bias . . . 197
10.D. Details zu Hypothesentest und Konfidenzinterval . . . 197
10.E. Restriktionen mit mehreren Koeffizienten . . . 200
10.E. Modellspezifikation . . . 201
10.E.1. Motivation . . . 201
10.E.2. Skalierung von Variablen und Koeffizienten . . . 203
10.E.3. Messe R^2 . . . 204
10.E.4. Messe Beitrag zum R^2 . . . 205
10.E.5. Informationskriterien . . . 208
10.F. t-Statistik und p-Wert . . . 210
10.F.1. Vergleich von Modellen . . . 210
10.F.2. Diskussion . . . 211
10.F.3. Literatur . . . 211
10.F.4. Übungen . . . 212

11. Kategoriale Variablen in der linearen Regression 218
11.1. Metrische und kategoriale Variablen . . . 218
11.2. Regression mit einer Dummy-Variablen . . . 219
11.2.1. Mehr als zwei Kategorien . . . 221
11.3. Interaktionen . . . 222
11.3.1. Interaktion zwischen binären Variablen . . . 224
11.3.2. Interaktion von diskreten Variablen im Bayesianischen Modell . . . 225
11.3.3. Binäre und stetige Variablen . . . 225
11.3.4. Bayes: Binäre und stetige Variablen . . . 226
11.3.5. Zwei stetige Variablen . . . 227
11.4. Literatur . . . 228
11.5. Schlüsselbegriffe . . . 228
11.A. Beispiele für die Vorlesung . . . 229
11.B. Übungen . . . 230
11.C. Nichtlineare Interaktionsterme . . . 234
11.C.1. Anwendung: Gender gap . . . 236
11.C. Varianzanalyse . . . 237
11.D. Kruskal-Wallis Test . . . 240
11.E. Übungen für Varianzanalyse... . . . 245

11.F. Friedman Test für verbundene Stichproben . . . 247
11.G. Jonckheere-Terpstra . . . 248

12. Nichtlineare Regressionsfunktionen 249
12.1. Motivation . . . 249
12.2. Funktionale Formen . . . 253
12.2.1. Polynome . . . 253
12.2.2. Logarithmische Modelle . . . 257
12.2.3. Logarithmische Modelle - linear-log . . . 258
12.2.4. Logarithmische Modelle - log-linear . . . 260
12.2.5. Logarithmische Modelle - log-log . . . 261
12.2.6. Vergleich der logarithmischen Modelle . . . 263
12.3. Nichtlinearitäten in den Parametern . . . 263
12.4. Schlüsselbegriffe . . . 264
12.A. Beispiele für die Vorlesung . . . 264
12.B. Übungen . . . 265

A. Eine kurze Einführung in R 269
A.1. Installation von R . . . 270
A.2. Datentypen und Zuweisungen . . . 270
A.3. Funktionen . . . 275
A.4. Zufallszahlen . . . 276
A.5. Beispiel-Datensätze . . . 277
A.6. Grafiken . . . 279
A.6.1. Densityplot . . . 280
A.6.2. Boxplot . . . 280
A.6.3. Empirische kumulierte Verteilung . . . 281
A.6.4. Q-Q Normal Plot . . . 281
A.6.5. Mosaicplot . . . 282
A.6.6. Graphen von Funktionen . . . 282
A.6.7. Leere Plots . . . 283
A.6.8. Linientyp . . . 283
A.6.9. Punktstil . . . 284
A.6.10. Legenden . . . 284
A.6.11. Hilfslinien . . . 285
A.6.12. Achsen . . . 286
A.7. Tabellen . . . 287
A.8. Regressionen . . . 287
A.9. Starten und Verlassen von R . . . 288

Homepage: https://www.kirchkamp.de/bw241/

Komponenten zum Lernen:

- Vorlesung

Ich freue mich, wenn Sie sich die Vorlesung anschauen und ich hoffe, dass Ihne die Vorlesung helfen wird, den Stoff zu verstehen. Allerdings gibt es unterschiedliche Lerntypen. Wenn Sie feststellen, dass Sie besser z.B. mit Übungsaufgaben und Handout lernen, dann sollte es Sie beruhigen zu erfahren, dass ich nicht vorhabe, in der Vorlesung »Geheimtipps« zu geben. Dieser Handout soll Ihnen vor allem einen Überblick über den Stoff der Vorlesung verschaffen. Wie Sie sich diesen Stoff aneignen, möchte ich Ihnen überlassen.

- Übung
- Handout
- Tutorien
- Hausaufgaben
- Diskussionsforum
- Online Meeting
- Literatur

Routinen zum Lernen

- Lernen Sie gemeinsam. Bilden Sie eine Lerngruppe.
- Folgen Sie einer Routine. Lernen Sie jede Woche zur gleichen Zeit.

Hausaufgaben

- Problem: Die Techniken, die in dieser Veranstaltung vermittelt werden, kann man nicht »*auswendig lernen*« sondern muss man *verstehen*. Das klappt aber nur mit *Üben*.
- Diese Übungen müssen Sie gleichmäßig über das Semester verteilen.
- Wie bleiben Sie am Ball? Wie kontrollieren Sie Ihren Erfolg?
- Lösung: Regelmäßige (wöchentliche) Aufgaben, regelmäßiges Feedback.
- Hausaufgaben zählen für Gesamtleistung.

Diskussionsforum

- Technische Fragen auf die Sie rasch eine Antwort haben wollen. Bitte verstehen Sie, dass ich alle Fragen zum Stoff so beantworten möchte, dass alle Studierenden die gleichen Informationen haben.

Online Meeting

- Allgemeine Fragen die länger warten können.

Konventionen im Handout Normalen Text werden wir immer so darstellen, wie hier. Berechnungen in Statistikprogrammen schreiben wir in einer Nichtproportionalschrift,so wie im folgenden Beispiel.

```
2+2
```

```
[1] 4
```

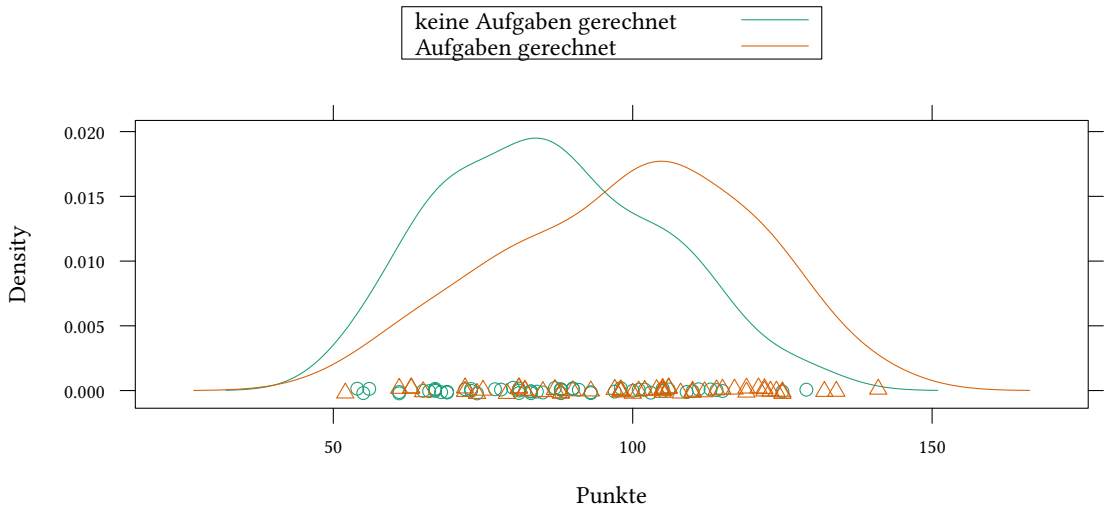
Erklärungen zu diesem Statistikprogramm schreiben wir in einer serifenlosen Schrift.

1. Einführung — Schätzen von Parametern

1.1. Ein Beispiel zur Motivation

Wir begleiten die Vorlesung durch regelmäßige Hausaufgaben. Unsere Hoffnung ist, dass Ihnen die regelmäßigen Hausaufgaben helfen, regelmäßig mitzuarbeiten, und helfen, gut vorbereitet in die Klausur zu gehen. Worauf gründet sich diese Hoffnung?

In den vergangenen Jahren haben wir ein ähnliches, freiwilliges Angebot gemacht. Hier ist die Dichte der Klausurergebnisse aus einem vergangenen Jahr. Wie Sie sehen: Es gibt mehr und weniger erfolgreiche Studierende in beiden Gruppen. Einfach wäre es, wenn alle Leute, die regelmäßig Aufgaben bearbeiten, 120 Punkte erhalten, und alle Leute, die das nicht tun, nur 80 Punkte erhalten. So einfach ist es aber nicht. Es gibt offenbar andere Faktoren, die wir nicht kennen, und die den Erfolg auch beeinflussen. In dieser Vorlesung werden wir darüber reden, wie wir mit dieser Unsicherheit umgehen – Unsicherheit, die entsteht, weil es Einflüsse gibt, die wir nicht kontrollieren können. Wir modellieren diese unkontrollierbaren Einflüsse als „Zufall“.



Vergleichen Sie die folgenden Aussagen:

- Regelmäßiges Rechnen von Aufgaben hat 2009/10 im Durchschnitt den Klausurerfolg verbessert.

- Regelmäßiges Rechnen von Aufgaben hat 2009/10 im Durchschnitt den Klausurerfolg um 12.5 Punkte verbessert.

- Die durchschnittliche Verbesserung des Klausurerfolgs (unabhängig vom Jahr) liegt zwischen 5.7 und 19.3 Punkten.

Alle Aussagen sind, soweit wir es heute beurteilen können, zutreffend — allerdings unterschiedlich präzise.

Die erste Aussage drückt sich um jede Quantifizierung. Gerade wenn wir Kosten und Nutzen vergleichen wollen, ist das wenig hilfreich. Fast alle interessanten ökonomische Maßnahmen haben sowohl positive als auch negative Konsequenzen. Nur wenn wir die Konsequenzen quantifizieren, haben wir eine Chance, Vor- und Nachteile abwägen zu können.

Die zweite Aussage ist sicherlich präzise — allerdings offenbart sie nicht, wie genau oder ungenau unser Wissen zu diesem Zusammenhang ist.

Erst die dritte Aussage klärt, dass wir in diesem Fall — leider — keine allzu genaue Prognose abgeben können.

1.2. Ziele der Vorlesung

Wir werden uns im Laufe der Vorlesung mit drei Dingen beschäftigen:

- *Schätzen* (Quantifizieren) von Parametern

  z.B. wie reagiert …

  - …der Klausurerfolg auf regelmäßiges Bearbeiten von Übungsaufgaben,
  - …das Wirtschaftswachstum auf Bankenregulierung,
  - …Arbeitlosigkeit auf Qualifizierungsmaßnahmen,…

    (Schätzen: → Kapitel 2–4)

    (Schätzen von Effekten/Einflüssen: → Kapitel 9–12.1)

- Bestimmen von *Credible-* und *Konfidenzintervallen* (Quantifizierung der Genauigkeit)

  Wir wissen keine genaue Antwort, aber in welchem Bereich liegt der gesuchte Wert etwa? (z.B. zwischen 5.7 und 19.3 Punkten im Klausurerfolg…)

  (→ Kapitel 4 und 7)

- *Testen* von Hypothesen (Quantifizierung der Signifikanz)

  Wir haben z.B. die Hypothese, regelmäßiges Rechnen von Aufgaben steigert den Klausurerfolg (nicht nur 2009/10, sondern allgemein). Wie können wir eine solche Hypothese testen?

  (→ Kapitel 5–8)

In dieser Vorlesung werden wir Methoden zur Strukturierung *empirische* Zusammenhänge betrachten. Diese Methoden nur *theoretisch* zu diskutieren, ist möglich, allerdings für viele Studierende nicht wirklich praxisrelevant. Ich versuche deshalb, den Stoff mit praktischen Beispielen am Rechner zu illustrieren.

- Theorie $\boxed{\text{z.B.} \frac{\mathrm{d}}{\mathrm{d}\xi} \sum_{i=1}^{n} (X_i - \xi)^2 = \dots}$
- Praxis (Anwendungen am Computer)

Für den praktischen Teil müssen wir uns auf irgendeine Software einigen. Jeder hat seine Lieblingssoftware:

- z.B.…SAS, STATA, SPSS, S, EViews, TSP, …
  - Installation erfordert Lizenz → teuer
- R
  - frei + gratis + sehr leistungsfähig
  - Probieren Sie die Beispiele mit R aus der Vorlesung und Übung möglichst mit Ihrem Computer oder den Computern im Pool der Fakultät aus. Benutzen Sie die Online Hilfe um neue Kommandos zu verstehen.
- Kann man Daten nicht auch in einer Tabellenkalkulation (Libreoffice, Gnumeric, Microsoft Excel…) bearbeiten?

→ Es ist deutlich schwerer, Arbeitsschritte in einer Tabellenkalkulation eindeutig und nachvollziehbar zu beschreiben.

Tabellenkalkulation: Schwer zu dokumentieren, fehleranfällig:

- Wähle im Menü File/Open die Datei `D.csv`.
- Wähle im Dialog eine lange Reihe von Optionen zum Import dieser Datei.
- Wähle das Menü Data/Statistics/Regression
- Wähle im Dialog folgendes:
  - Independent variable range: `$D06.$A$2:$A$41`
  - Dependent variable range: `$D06.$B$2:$B$41`
  - ⋮

Notation wie `$D06.$A$2:$A$41` ist nicht offensichtlich und führt zu Fehlern.

Programm (z.B. R): Leicht zu dokumentieren:

```
lm(y ~ x,data=read.csv("D.csv"))
```

Sie werden (fast) immer wieder zu Ihrer Arbeit zurückkommen:

- Reproduzierbares Arbeiten spart Arbeit.
- Wenn gut dokumentiert ist, was Sie gemacht haben, ist es leicht, an Ihre Arbeit wieder anzuknüpfen.
- Ist es nicht offensichtlich, wie die Analyse von Daten durchzuführen ist?
- Alles, was man braucht, sind Daten und die Fragestellung.

  Die Antwort sollte klar sein — oder?

Leider ist es nicht so einfach:

Beispiel: Silberzahn, R., Uhlman, D.,...Nosek, B. (2018). Many analysts, one data set: making transparent how variations in analytic choices affect results. *Advances in Methods and practices in Psychological Science* (1). 337-356.

Autoren geben die gleichen Daten an 29 Teams von Forschern und Forscherinnen.

- Frage: Werden im europäischen Fußball farbige Fußballspieler benachteiligt?
- Antwort (abhängig von Forschungsteam): farbige Spieler erhalten...
  - 11% weniger gelbe und rote Karten.

    ⋮
  - 2.93 × so viel gelbe und rote Karten.

In dieser Veranstaltung werden wir mit R arbeiten. Für diese Arbeit sollten Sie sich zwei Programme installieren:

- R

  Instruktionen finden Sie auf `https://cran.r-project.org/`
- Ein Frontend für R. Ich verwende hier RStudio.

  `https://rstudio.com/products/rstudio/download/`
- R organisiert seine Funktionen und Daten in »Paketen«. Bevor ein Paket verwendet werden kann, muss man das R zunächst sagen:

  ```
  library(Ecdat)
  ```

  Das klappt nur, wenn das Paket auch vorhanden ist. Wenn es statt dessen eine Fehlermeldung gibt...

```
Error in library(Ecdat): there is no package called 'Ecdat'
```

…müssen wir das Paket installieren (z.B. aus RStudio heraus, rechts unten, Reiter »Packages / Install«).

Ein Frontend (wie RStudio) macht es einfacher, R zu bedienen.

File Edit Code View Plots Session Build Debug Profile Tools Help	
Editor	Environment
	Files Plots Packages
Console	

- Ziel der Vorlesung ist es *nicht*, dass Sie lernen, R perfekt zu beherrschen. Verschwenden Sie nicht zuviel Zeit darauf, sich die genaue Syntax von komplizierten Kommandos einzuprägen.
- Ich gebe hier im Handhout manchmal die genaue Syntax für eine bestimmte Software (R) an, damit Sie es leichter haben, die Beispiele selbst nachzuvollziehen (und damit Sie die Syntax eben nicht auswendig lernen müssen). Die Syntax, die Sie im Handout sehen, wird zuweilen kompliziert, weil ich versuche, die Ergebnisse einigermaßen hübsch darzustellen. Manchmal möchte ich Ihnen auch einen kleinen Trick zeigen. In der Vorlesung werde ich mir das oft sparen. Vielleicht lassen Sie sich ja für die ein oder andere Grafik in Ihrer Seminar- oder Diplomarbeit inspirieren. Für die Klausur sind diese Feinheiten jedenfalls nicht relevant.
- Sie sollen lernen, grundsätzlich Ergebnisse (Text und Grafik) von Statistikprogrammen zu verstehen.
- Die Details stehen dabei nicht im Vordergrund.

  Am Ende der Vorlesung sollten Sie z.B. wissen, wie die folgende Tabelle zu interpretieren ist.

```
data(Caschool,package="Ecdat")
attach(Caschool)
large <- str>20
xtable(est<-lm(testscr ~ str + elpct + expnstu))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	649.5779	15.2057	42.72	0.0000
str	-0.2864	0.4805	-0.60	0.5515
elpct	-0.6560	0.0391	-16.78	0.0000
expnstu	0.0039	0.0014	2.74	0.0064

- Wenn Sie sich jetzt immer noch unsicher sind, wie die Klausur aussieht: Rechnen Sie regelmäßig Aufgaben aus Tutorium und Übung. So haben Sie hinreichend Gelegenheit, realistische Aufgaben zu üben.

1.3. Schätzer

- Die »wahren Zusammenhänge« dieser Welt werden beschrieben durch unbekannte Parameter θ (Zusammenhang zwischen z.B.
  - Arbeitslosenunterstützung↔Arbeitslosigkeit,
  - Übungsaufgaben↔Klausurerfolg,
  - Inflationsrate↔Wirtschaftswachstum,...)
- Wir beobachten einen Ausschnitt dieses Zusammenhangs (*Stichprobe* $X_1, \ldots, X_n$)

  (z.B. Arbeitslose in Thüringen 2020, Klausur 2021,...)
- Wir wollen einen Aussage machen über »den Rest der Welt« (einen anderen Teil der *Population*, z.B. Arbeitsmarkt in Thüringen 2023, Klausurerfolg 2023,...)

		Modell:		**beobachtbar:**
Ereignisraum $\omega \in \Omega$	$\rightarrow$	Zufallsvariable $X \sim F(X\|\theta)$	$\rightarrow$	Stichprobe $X_1, \ldots, X_n$
		Parameter θ	$\leftarrow$	$\hat{\theta} = g(X_1, \ldots, X_n)$

Ereignisraum Ω: Der *Ereignisraum* Ω enhält alle möglichen Ereignisse ω. Um mit Ereignissen rechnen zu können, bilden wir sie in Zahlen X ab. X nennen wir *Zufallsvariable*.

Zufallsvariable X: Wir betrachten eine *Zufallsvariable* $X \sim F(X|\theta)$ mit unbekanntem Parameter θ. (Manchmal werden wir unsere Parameter auch μ, σ, oder β nennen.)

Stichprobe $(X_1, \ldots, X_n)$: Nun beobachten wir eine *Stichprobe*, im einfachsten Fall also n Zahlen.

Diese Zahlen X_i folgen einer (unbekannten) Struktur die durch $X \sim F(X|\theta)$ gegeben ist. Diese Struktur (das θ) versuchen wir zu ergründen.

Stichprobenfunktion g: Eine *Stichprobenfunktion* g berechnet uns aus der Stichprobe $(X_1, \ldots, X_n)$ eine neue Zahl (oder mehrere neue Zahlen).

Schätzer Ein *Schätzer* ist eine Funktion $g(X_1, \ldots, X_n)$ die jeder Realisierung der Stichprobe $X_1, \ldots, X_n$ einen Schätzwert $\hat{\theta}$ für das unbekannte θ zuordnet.

- Schreibweise: $\hat{\theta} = g(X_1, \ldots, X_n)$

Schätzer, die Sie bereits kennen

- Mittelwert: $\hat{\mu}_X = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ (Zufallsvariable) ($\bar{x}$ ist also nicht nur der Mittelwert der Stichprobe, $\bar{x}$ ist auch ein Schätzer für den Mittelwert μ_X der Population)
- Median: $\hat{\mu}_{X,\frac{1}{2}} = \text{median}_{i=1}^{n}(X_i)$ (Zufallsvariable)
- Varianz: $\hat{\sigma}_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ (Zufallsvariable) ($\hat{\sigma}_X^2$ ist also nicht nur die Varianz der Stichprobe, $\hat{\sigma}_X^2$ ist auch ein Schätzer für die Varianz σ_X^2 der Population)
- Schätzer der Standardabweichung: $\hat{\sigma}_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$ (Zufallsvariable) ($\hat{\sigma}_X$ ist also nicht nur die Standardabweichung der Stichprobe, $\hat{\sigma}_X$ ist auch ein Schätzer für die Standardabweichung σ_X der Population)

Die Schätzer für Mittelwert und Median sind sogenannte »plug-in« Schätzer. Die gleiche Funktion, die einen Parameter der Population bestimmt, wird auch als Schätzer für diesen Parameter genommen. Das muss nicht sein. Die Schätzer für Varianz und Standardabweichung führen eine kleine Korrektur ein.

Welchen Schätzer sollte man nehmen? Beispiel: Schätzer für den Mittelwert:

- Nehmen wir an, wir haben eine Stichprobe $X_1 \ldots X_n$ einer Zufallsvariablen $X \sim F(X|\theta)$.
- Angenommen, der unbekannte Parameter θ sei der *Mittelwert* μ. Wie können wir den Mittelwert von X schätzen?

→ Wir nehmen einfach die erste Beobachtung X_1

→ Wir nehmen den Median der Stichprobe $X_1 \ldots X_n$

→ Wir nehmen den Mittelwert $\bar{X}$ der Stichprobe $X_1 \ldots X_n$

→ ⋮

Alle diese Verfahren liefern einen Schätzer für den Mittelwert μ. In Kapitel 2 werden wir Kriterien kennenlernen, die uns helfen, die Qualität dieser Schätzer zu bestimmen.

Beispiel — Ausgaben für Lebensmittel

Im Anhang wird anhand eines einfachen Beispiels erklärt, wie Sie einen Datensatz in R öffnen, wie Sie Daten ansehen, und wie Sie mit den Daten arbeiten können. Hier kürzen wir das Verfahren etwas ab.

Das Kommando `data` öffnet einen Datensatz aus einem Paket das mit `package` spezifiziert wird.

```
data(BudgetFood,package="Ecdat")
```

Das Kommando `dim` sagt und, wie groß der Datensatz ist, also wie viele Zeilen (Stichproben) und Spalten (Variablen) er besitzt.

```
dim(BudgetFood)

[1] 23972     6
```

`names` zeigt die Namen der Variablen in einem Datensatz an:

```
names(BudgetFood)

[1] "wfood"  "totexp" "age"    "size"   "town"   "sex"
```

Die Variable `wfood` enthält den Anteil der Ausgaben spanischer Haushalte für Lebensmittel. Wenn wir uns für den mittleren Anteil interessieren, könnten wir einfach das folgende Kommando verwenden:

Das Kommando `attach` erlaubt es uns, auf die Variablen dieses Datensatzes zuzugreifen, ohne jedes Mal den Namen des Datensatzes dazu schreiben zu müssen. `mean` schließlich berechnet einen Mittelwert.

```
attach(BudgetFood)
mean(wfood)

[1] 0.378321
```

Also 37.83% des verfügbaren Einkommens wird für Lebensmittel ausgegeben. Das ist einfach, wenn wir Zugriff auf den gesamten Datensatz haben. Stellen wir uns vor, die 23972 Mitglieder dieses Datensatzes sind unsere Population (und wir würden uns, im Rahmen dieses Beispiels, nur für diese 23972 interessieren). Stellen wir uns ferner vor, dass es kostspielig ist, Daten zu erheben, und dass wir nur eine kleine Stichprobe nehmen können (dass wir also *nicht* alle 23972 kennen).

Das Kommando `sample` zieht eine Stichprobe aus einem Vektor.

Hier ziehen wir mehrmals eine Stichprobe der Größe 10 aus dem Vektor `wfood` (wir stellen uns vor, dass wir jeweils 10 Haushalte befragen):

```
sample(wfood,10)

 [1] 0.1820983 0.3313547 0.6281348 0.4699777 0.2812468 0.2905913 0.3294758
 [8] 0.2780813 0.3691347 0.5278351
```

```
sample(wfood,10)

 [1] 0.4652740 0.2817307 0.3339667 0.2274102 0.1070396 0.3598388 0.1651700
 [8] 0.3323353 0.7974191 0.6178727
```

```
sample(wfood,10)

 [1] 0.2590431 0.5522304 0.5099138 0.4490395 0.3665509 0.2899829 0.3237376
 [8] 0.3826898 0.3073169 0.1615013
```

Dies sind also jeweils 10 zufällig ausgewählte Haushalte.

Wenn man von einer solchen Stichprobe den Mittelwert nimmt, bekommt man ein Ergebnis. Allerdings bekommt man mit der nächsten Stichprobe auch ein anderes Ergebnis. Hier sind drei Mittelwerte, die auf diese Weise entstanden sind:

```
mean(sample(wfood,10))

[1] 0.4000348
```

```
mean(sample(wfood,10))

[1] 0.3703045
```

```
mean(sample(wfood,10))

[1] 0.4023415
```

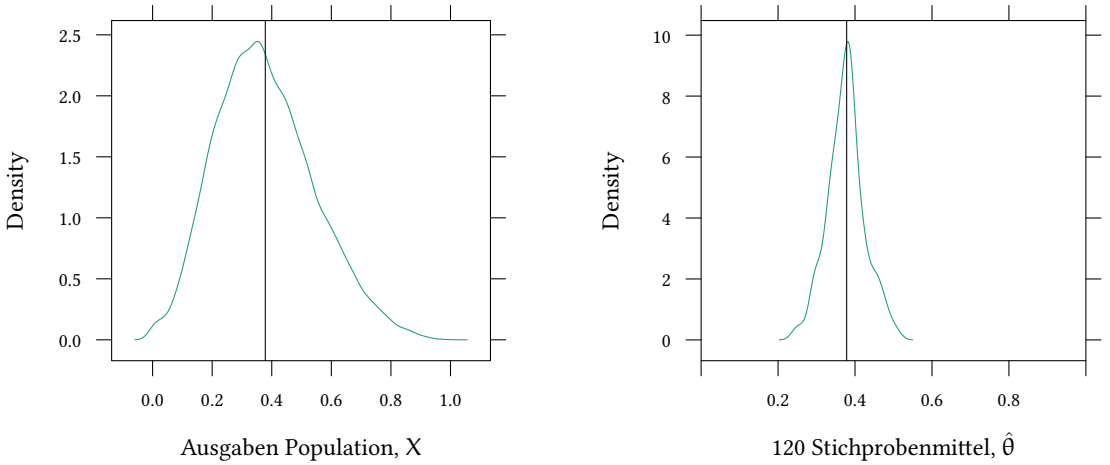
Wenn wir also drei Interviewer zu unserer Population von 23972 Haushalten schicken würden, die jeweils eine Stichprobe der Größe 10 machen würden, dann würden wir drei verschiedene Ergebnisse, drei verschiedene *Schätzungen* für den Mittelwert der Population von 23972 erhalten.

Das passiert uns bei vielen Messungen. Wir befragen nur einen Teil der Wahlberechtigten über ihr Verhalten bei der nächsten Wahl, nur einen Teil der Konsumenten über ihr Kaufverhalten,... In jedem Fall gibt uns unsere Messung ein Ergebnis, aber dieses Ergebnis wird bei jeder neuen Befragung unterschiedlich ausfallen.

Der Schätzer, also der Mittelwert, den wir hier betrachten, ist also wieder eine Zufallsvariable. Wenn wir eine Aussage über die *Genauigkeit* des Schätzers machen wollen, brauchen wir dessen *Verteilung*.

Das Bild links stellt noch einmal die Verteilung der Population dar. Das Bild rechts die Verteilung unseres Schätzers – also die Verteilung, die sich ergeben würden, wenn wir wieder und wieder eine Stichprobe der Population nehmen würden, und für jede dieser Stichproben den Schätzer bestimmen. Die gestrichtelte Linie gibt in beiden Grafiken den Mittelwert der Population von 0.378 an.

`densityplot` berechnet die Dichtefunktion (siehe Abschnitt A.6.1) die mit `plot` als Grafik dargestellt wird.



Auf Basis unserer kleinen Stichproben schätzen wir also zuweilen einen kleineren und zuweilen einen größeren Wert als den wahren Mittelwert (0.378) unserer Population von 23972 Haushalten.

Wir sehen aber auch, dass der Mittelwert (rechts) »genauer« ist, als eine einzelne Beobachtung der Population (links). Die Verteilung rechts ist »schmaler«. Wir erhalten also mit dem Mittelwert einer Stichprobe kein perfektes Ergebnis, aber ein besseres Ergebnis, als wenn wir nur eine Beobachtung nehmen würden.

1.4. Interpretation von Schätzern/Stichprobenfunktionen als Zufallsvariablen

- Die $X_1, \ldots, X_n$ der Stichprobe sind Zufallsvariablen.
- Also sind *Funktionen* $\hat{\theta}(X_1, \ldots, X_n)$ (z.B. der Mittelwert der Stichprobe) ebenfalls Zufallsvariablen.

  Hätten wir eine andere Stichprobe gezogen, dann hätte $\hat{\theta}(X_1, \ldots, X_n)$ (z.B. der Mittelwert) auch einen anderen Wert.
- Die Verteilung von $\hat{\theta}$ über verschiedene mögliche Stichproben nennen wir die *Stichprobenverteilung* von $\hat{\theta}$.
- Mittelwert und Varianz von $\hat{\theta}$ sind Mittelwert und Varianz der Stichprobenverteilung $\mathsf{E}(\hat{\theta})$ und $\mathrm{var}(\hat{\theta})$.

1.5. Einige Punktschätzer

Ein Schätzer für den Mittelwert

$$\hat{\mu}_X = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Varianz Während es vielleicht naheliegend ist, sich für den Mittelwert zu interessieren, ist die Frage nach der Varianz nicht ganz so offensichtlich. Hier ist ein Grund: Wir wollen zuweilen Mittelwerte vergleichen. Entweder den Mittelwert der einen Stichprobe mit dem Mittelwert der anderen Stichprobe, oder einen Mittelwert einer Stichprobe mit einem hypothetischen Wert.

Um (weiter unten) beurteilen zu können, ob eine Abweichung zwischen zwei Mittelwerten oder zwischen einem Mittelwert und einem hypothetischen Wert mehr als nur »zufällig« ist, müssen wir wissen, wie »genau« wir den Mittelwert eigentlich geschätzt haben. Dazu müssen wir wissen, wie »genau« unsere Beobachtungen waren, und das beurteilt die Varianz.

Zuweilen sind wir auch ursächlich an der Varianz in der Population interessiert, etwa wenn eine hohe Varianz auch mit einem hohen Risiko einhergeht.

Ein Schätzer für die Varianz:

$$\hat{\sigma}^2_X = \frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$

Warum sollte man diese Schätzer verwenden, und nicht andere?

In Kapitel 2 werden wir diese Frage beantworten.

1.6. Regression als ein weiterer Punktschätzer

Die Eigenschaften von Mittelwerten allein mag etwas langweilig erscheinen. Wir beginnen mit Mittelwerten, weil es sich um sehr einfache Schätzer handelt. Später werden wir die Idee auf Regressionen verallgemeinern. Damit wir dieses Ziel nicht aus den Augen verlieren, kommt hier bereits eine kleine Einführung. Wir stellen uns vor, dass wir folgenden linearen Zusammenhang schätzen wollen:

Mehrere Variablen X_1,...,X_k haben einen Effekt auf eine abhängige Variable Y:

Wir nehmen an, der zu schätzender Zusammenhang sei linear:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + u$$

Dabei sind

- Y die abhängige Variable.
- $X_1 \ldots X_k$ die unabhängigen Variablen, die Y erklären sollen (Notation: oben waren X_i identisch verteilte Stichprobenbeobachtungen. Wenn wir über Regressionen reden, sind $X_1, X_2, \ldots$ unterschiedliche Variablen).

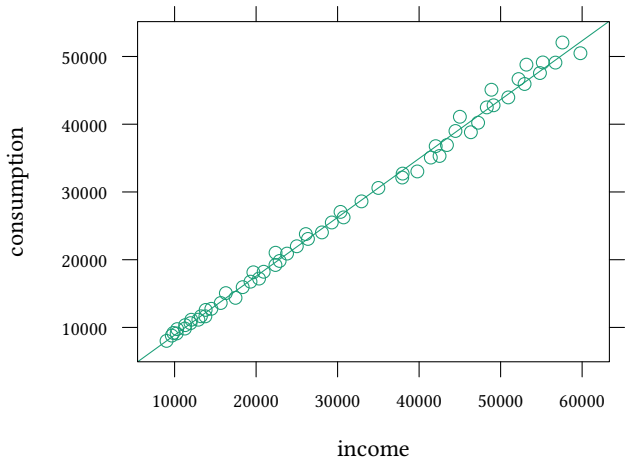
- $\beta_0 \ldots \beta_k$ die Parameter des Modells, die geschätzt werden sollen (genauso wie Mittelwerte)

- u der vom Modell nicht erklärte Rest.

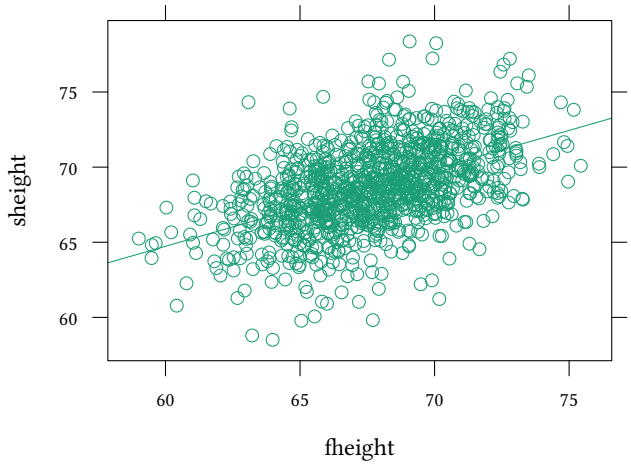
Ziel: Messung der einzelnen Effekte $\beta_1, \ldots, \beta_k$.

Die folgenden drei Grafiken zeigen jeweils den Zusammenhang zwischen zwei Variablen. Zunächst sehen Sie Konsum und Einkommen, in der Mitte die Größe von Söhnen und Vätern, und schließlich die Fruchtbarkeit von Schweizern bezogen auf den Anteil der Katholiken in der Bevölkerung.

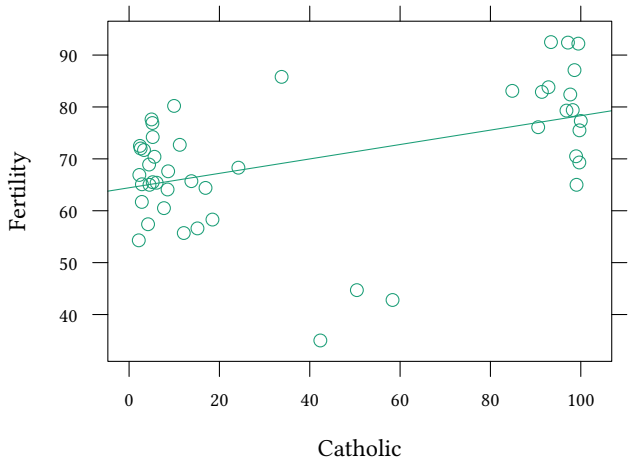
```
library(lattice)
data(IncomeUK,package="Ecdat")
xyplot(consumption ~ income,type=c("r","p"),data=as.data.frame(IncomeUK))
```




```
data(father.son,package="UsingR")
xyplot(sheight ~ fheight,type=c("r","p"),data=father.son)
```




```
data(swiss,package="datasets")
xyplot(Fertility ~ Catholic,type=c("r","p"),data=swiss)
```



Viele Zusammenhänge lassen sich auf diese Art und Weise darstellen. Viele Zusammenhänge sind auch approximativ linear, können also durch eine Linie näherungsweise dargestellt werden. Eine solche Linie finden Sie in allen drei Grafiken. Manchmal ist die Annäherung gut, manchmal weniger gut. Eine solche Linie hilft uns, zu beurteilen, wie sich der Konsum verändert, wenn etwa das Einkommen um einen bestimmten Betrag steigt, oder wie sich die Größe der Söhne ändert, wenn die Väter größer werden, oder welche Auswirkungen auf die Fruchtbarkeit eine Veränderung des Anteils der katholischen Bevölkerung hat.

Eine Möglichkeit, eine solche Gerade zu »schätzen«, ist die OLS Regression. Wir werden in Kapitel 9 ausführlich auf dieses Thema zu sprechen kommen.

Auch wenn wir hier über *Schätzer* im Allgemeinen reden, ist es wichtig, OLS Regressionen

zu erwähnen. Die geschätzten Parameter der Geraden sind ebenfalls Parameterschätzer und haben sehr ähnliche Eigenschaften wie etwa der Mittelwert.

Einstweilen stellen wir uns vor, dass eine Regression folgendes macht:

- lege eine gerade Linie durch zweidimensionale Daten X, Y
- schätze einen kausalen Zusammenhang zwischen Y und X

Linien haben Parameter für *Steigung* und *Achsenabschnitt*
bislang: Schätzung für *Mittelwerte* (nur ein einziger Parameter wird geschätzt)
jetzt: Schätzung für *Steigung* und *Achsenabschnitt* (mehrere Parameter werden geschätzt)

$$\texttt{sheight} = \beta_0 + \beta_1 \cdot \texttt{fheight} + u$$

- β_0 und β_1 sind Parameter der *Population*
- wir kennen sie nicht — also müssen wir sie schätzen (wie $\hat{\theta}$)

Die genauen Verfahren der Schätzung werden wir in Kapitel 9 diskutieren. Einstweilen werden wir das Kommando `lm` verwenden (siehe Abschnitt A.8)
:

Beispiel: Väter und Söhne

```
data(father.son,package='UsingR')
lm(sheight ~ fheight,data=father.son)


Call:
lm(formula = sheight ~ fheight, data = father.son)

Coefficients:
(Intercept)      fheight
    33.8866       0.5141
```

Der geschätzte Koeffizient von `fheight` ist also 0.5141.

$$\texttt{sheight} = \beta_0 + \beta_1 \cdot \texttt{fheight} + u$$

$$\texttt{sheight} = 33.8866 + 0.5141 \cdot \texttt{fheight} + u$$

→ Wenn der Vater um 1 Zoll größer ist, dann ist der Sohn um (etwa) 0.5141 Zoll größer.

Beispiel: Einkommen und Konsum

```
data(IncomeUK,package="Ecdat")
lm(consumption ~ income,data=as.data.frame(IncomeUK))
```

```
Call:
lm(formula = consumption ~ income, data = as.data.frame(IncomeUK))

Coefficients:
(Intercept)       income
    176.848        0.869
```

Der geschätzte Koeffizient von `income` ist also 0.869.

$$\texttt{consumption} = \beta_0 + \beta_1 \cdot \texttt{income} + u$$

$$\texttt{consumption} = 176.848 + 0.869 \cdot \texttt{income} + u$$

→ Wenn das Einkommen um 1 £ wächst, dann steigen die Ausgaben für Konsum um (etwa) 0.869 £.

Beispiel: Fruchtbarkeit und Religion

```
data(swiss,package="datasets")
with(swiss,lm(Fertility ~ Catholic))


Call:
lm(formula = Fertility ~ Catholic)

Coefficients:
(Intercept)     Catholic
    64.4283       0.1389
```

Der geschätzte Koeffizient von `Catholic` ist also 0.1389.

$$\texttt{Fertility} = \beta_0 + \beta_1 \cdot \texttt{Catholic} + u$$

$$\texttt{Fertility} = 64.4283 + 0.1389 \cdot \texttt{Catholic} + u$$

→ Wenn der Anteil der katholischen Bevölkerung um 1% wächst, dann steigt die Fertilität um (etwa) 0.1389.

In allen drei Beispielen können wir zwar eine Steigung β_1 schätzen, damit wird aber keine Aussage über die Kausalität gemacht. Es ist z.B. nicht sicher, dass der Anteil der katholischen Bevölkerung die Fertilität beeinflusst. Denkbar wäre etwa, dass die Nähe zu einer Großstadt einerseits die persönliche Einstellung zum Glauben, und gleichzeitig den Zugang zu oralen Kontrazeptiva beeinflusst.

1.7. Literatur

- Wenn Sie sich in R einlesen wollen: Dolić, Statistik mit R, Kapitel 1-3.
- Mehr zum Begriff der »Stichprobenverteilung« finden Sie in Dolić, Statistik mit R, Kapitel 5.3.

1.8. Schlüsselbegriffe

- Schätzer, Konfidenzintervall, Hypothesentest
- Punktschätzer
- Mittelwert

Anhang 1.A Beispiele für die Vorlesung

Ihre Stichprobe enthält 4 unabhängige und identisch verteilte Beobachtungen: $X_1, \dots, X_4$. Welche Schätzfunktionen für $E(X)$ sind erwartungstreu?

- $\frac{1}{3}X_1 + \frac{2}{3}X_2$
- $X_1 - X_2$
- $\frac{1}{3}X_1 + \frac{1}{3}X_3$
- $\frac{1}{4}\sum_{i=1}^{4}\sqrt[3]{X_i^3}$
- $\frac{1}{4}\sum_{i=1}^{4}\sqrt{X_i^2}$

Betrachten Sie weiter die obige Stichprobe. Die Varianz von X sei 10. Wie groß ist die Varianz von $X_1 + 2X_2 - X_4$?

Ihre Stichprobe enthält 5 unabhängige und identisch verteilte Beobachtungen: $X_1, \dots, X_5$. Welche Schätzfunktionen für $E(X)$ sind erwartungstreu?

- $\sum_{i=1}^{5} X_i$
- X_3
- $\frac{X_2}{2} + \frac{X_3}{2}$
- $X_1 + X_2 - X_3$
- $\frac{1}{5}\sum_{i=1}^{5} X_i$

Betrachten Sie weiter die obige Stichprobe. Die Varianz von X sei 90. Wie groß ist die Varianz von $\frac{1}{3}(X_1 + X_5)$?

Anhang 1.B Übungen

Übung 1.1 *Installieren Sie R auf Ihrem Computer. Das praktische Arbeiten mit dem Computer wird es Ihnen leichter machen, die Inhalte der Vorlesung zu vertiefen.*

- *Wenn Sie Schwierigkeiten bei der Installation haben: Nutzen Sie das Diskussionsforum der Veranstaltung.*
- *Wenn Sie keinen Computer haben: R ist auch im Computerpool der Fakultät installiert.*

Übung 1.3 *Berechnen Sie in R: Die Stichprobe 1 enthält die Beobachtungen (11,12,13), Stichprobe 2 enthält die Beobachtungen (11,12,12).*

- *Bestimmen Sie den Mittelwert der beiden Stichproben und stellen Sie die Mittelwerte graphisch dar.*
- *Berechnen Sie in R:*
  - *Wie viele Beobachtungen enthält Stichprobe 1.*
  - *Wie viele Beobachtungen in Stichprobe 2 sind größer als 11?*
  - *Wie groß ist der Anteil der Beobachtungen in Stichprobe 2 die größer als 11 sind?*

Übung 1.4 *Der Datensatz `Mathlevel` aus der Bibliothek `Ecdat` beschreibt den Erfolg in einem Mathematiktest für Studierende unterschiedlicher Studienfächer.*

- *Betrachten Sie zunächst den gesamten Datensatz. Hängt der SAT-Math-Score davon ab, ob Studierenden eine Fremdsprache sprechen?*
- *Ziehen Sie nun jeweils 10 Studierende ohne und 10 andere mit Fremdsprachenkenntnissen und vergleichen Sie. Wiederholen Sie diesen Vergleich 1000 mal. Wie oft kommen Sie zum gleichen Ergebnis wie oben?*

```
data(Mathlevel,package="Ecdat")
attach(Mathlevel)
populationDiff <- mean(subset(sat,language=="no"))
                  - mean(subset(sat,language=="yes"))
sampledDiff <- replicate(1000,mean(sample(subset(sat,language=="no"),10))-
                               mean(sample(subset(sat,language=="yes"),10)))
mean(sampledDiff<0)
```

Übung 1.5 *Der Erwartungswert von X ist 4. Der Erwartungswert von Y ist 3.*

- *Wie groß ist der Erwartungswert von $X+Y$?*
- *Wie groß ist der Erwartungswert von $X-Y$?*
- *Wie groß ist der Erwartungswert von $2X-3Y$?*

Übung 1.6 *Eine Zufallsvariable X habe die Varianz σ_X^2. Was ist die Varianz von $Y \equiv 5 \cdot X + 2$?*

Übung 1.7 *Die Varianz von X ist 10. Die $X_1, X_2, X_3 \ldots$ sind unabhängig voneinander. Was ist die Varianz von $2X_2 - 3X_3 + 2X_4$?*

2. Wünschenswerte Eigenschaften von Schätzern

In dieser Vorlesung werden wir sehr viel schätzen und testen. Wir werden uns zuweilen fragen, warum ein Schätzer besser ist als ein anderer. In diesem Kapitel werden wir drei wichtige wünschenswerte Eigenschaften von Schätzern betrachten:

- Erwartungstreue (Unverzerrtheit)
- Effizienz
- Konsistenz

2.1. Erwartungstreue (Unverzerrtheit)

Eine Punktschätzung $\hat{\theta}(X_1,\ldots,X_n)$ zum Schätzen von θ nennen wir *erwartungstreuen* Schätzer (oder *unverzerrten* Schätzer) von θ falls

$$\forall\theta : \mathsf{E}\left(\hat{\theta}(X_1,\ldots,X_n)\right) = \theta$$

Wir verlangen also, dass unser Schätzer $\hat{\theta}$ für jeden Parameter θ im Erwartungswert richtig liegt. Eine einzelne Schätzung kann natürlich zu groß oder zu klein sein. Wir wollen aber eine systematische Verzerrung (im Mittel immer etwas zu groß oder im Mittel immer etwas zu klein) vermeiden.

Wenn ein Schätzer nicht erwartungstreu ist, ist er verzerrt.

Die *Verzerrung* von $\hat{\theta}$ ist

$$\text{Bias}(\hat{\theta},\theta) = \mathsf{E}(\hat{\theta}) - \theta$$

Ein verzerrter Schätzer für den Mittelwert wäre etwa

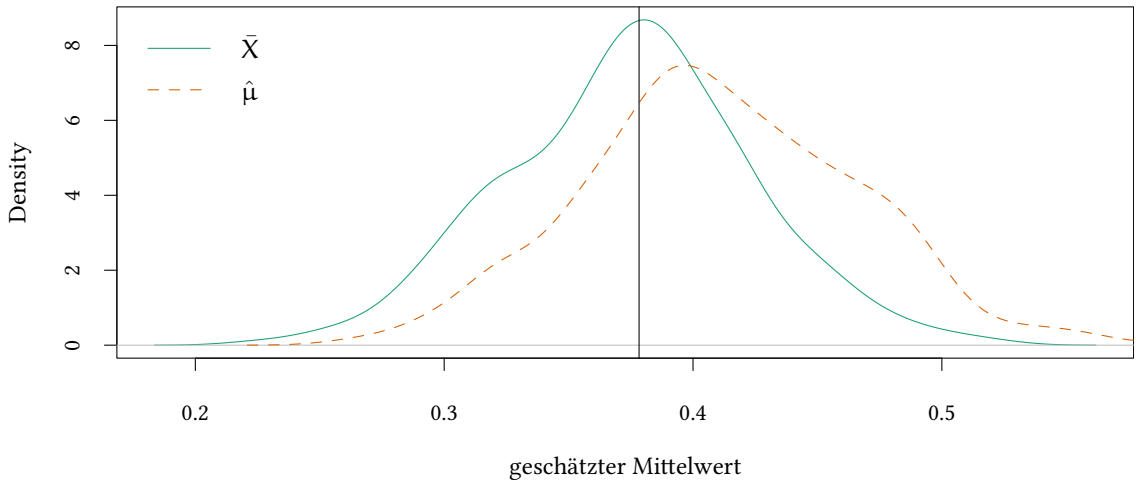
$$\hat{\mu} = \sqrt{\frac{1}{n}\sum X_i^2}$$

Hier vergleichen wir diesen Schätzer mit dem uns vertrauten Schätzer $\bar{X}$ am Beispiel des Datensatzes `BudgetFood` aus dem Paket `Ecdat`:

```
data(BudgetFood,package="Ecdat")
attach(BudgetFood)
sample(wfood,10)

 [1] 0.5494003 0.6511102 0.2639651 0.1514436 0.1957205 0.3504520 0.3765998
 [8] 0.5225546 0.1470028 0.3306543

mean(sample(wfood,10))

[1] 0.3610216

Xquer<-replicate(500,mean(sample(wfood,10)))
Xbias<-replicate(500,sqrt(mean(sample(wfood,10)^2)))
plot(density(Xquer),main="")
lines(density(Xbias),col=2,lty="dashed")
```

Hier sind unsere beiden Schätzer: $\bar{X}$ und $\hat{\mu}$:

$$\bar{X} = \frac{1}{n}\sum X_i \qquad \hat{\mu} = \sqrt{\frac{1}{n}\sum X_i^2}$$



Wir sehen: Beide Schätzer sind manchmal zu groß und manchmal zu klein — aber $\hat{\mu}$ ist sehr oft zu groß und selten zu klein. Deshalb nennen wir einen solchen Schätzer »verzerrt«.

Erwartungstreue des Mittelwerts Ist der Stichprobenmittelwert $\bar{X}$ ein guter Schätzer für den Populationsmittelwert μ_X?

$$\hat{\mu}_X = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Man kann zeigen: $E(\bar{X}) = \mu_X$,

d.h.

$\bar{X}$ ist ein *unverzerrter* Schätzer für μ_X

(Dies gilt unabhängig von der Verteilung, notwendig ist nur Unabhängigkeit der X_i)

Wie sieht man das?

$$
\begin{aligned}
\mathrm{E}(\bar{X}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) &= \frac{1}{n}\cdot \mathrm{E}\left(\sum_{i=1}^{n} X_i\right)\\
&= \frac{1}{n}\cdot \sum_{i=1}^{n} \mathrm{E}(X_i)\\
&= \frac{1}{n}\cdot \sum_{i=1}^{n} \mathrm{E}(X)\\
&= \frac{1}{n}\cdot n\cdot \mathrm{E}(X)\\
&= \mathrm{E}(X) = \mu_X
\end{aligned}
$$

Erwartungstreue der geschätzten Varianz Ähnlich kann man zeigen:

Die Stichprobenvarianz s_X^2 ist ein erwartungstreuer Schätzer für die Populationsvarianz σ_X^2?

$$\hat{\sigma}_X^2 = s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$\mathrm{E}\left(\hat{\sigma}_X^2\right) = \sigma_X^2$

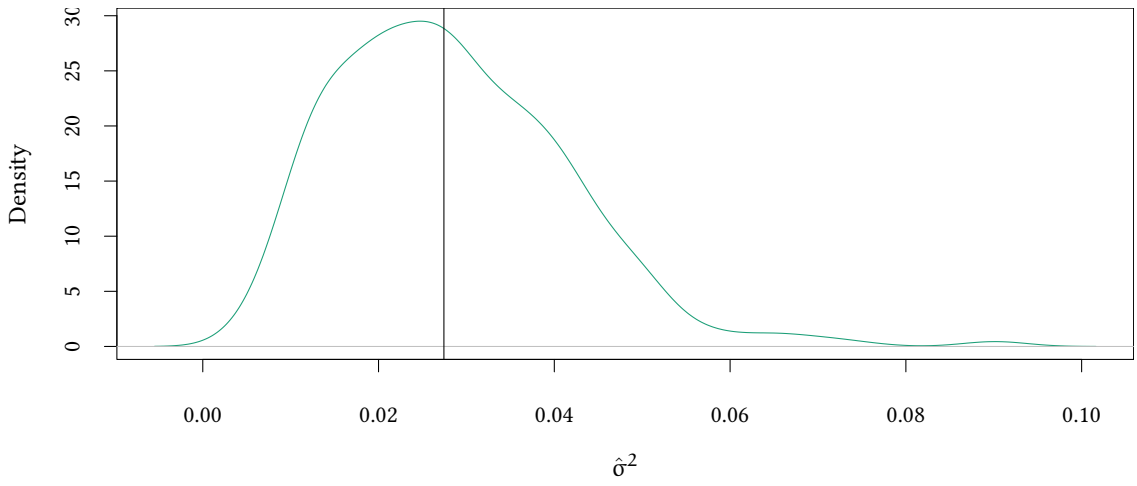
d.h.

s_X^2 ist ein unverzerrter Schätzer für σ_X^2.

(Dies gilt unabhängig von der Verteilung)

```
var(wfood)

[1] 0.02743794

var(sample(wfood,10))

[1] 0.03033003

mean(replicate(500,var(sample(wfood,10))))

[1] 0.02690195
```

```
VAR<-replicate(500,var(sample(wfood,10)))
plot(density(VAR),main="")
```



Erwartungstreue ist nicht alles

Übung 2.1 *Wir nehmen als Schätzer für den Erwartungswert jetzt einfach den Wert der ersten Beobachtung:* $\hat{\mu}_X = X_1$. *Ist dieser Schätzer erwartungstreu?*

— Man sieht leicht:

$$E(\hat{\mu}_X) = E(X_1) = E(X)$$

Allerdings ist dieser Schätzer nicht sehr genau:

$$\operatorname{var}(\hat{\mu}_X) = \operatorname{var}(X_1) = \operatorname{var}(X)$$

Die Varianz des Mittelwertes ist kleiner:

$$\operatorname{var}(\bar{X}) = \operatorname{var}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n^2}\operatorname{var}\left(\sum_{i=1}^{n} X_i\right) = \\ = \frac{1}{n^2} n \operatorname{var}(X) = \frac{1}{n}\operatorname{var}(X)$$

Beide Schätzer, der Mittelwert und die erste Beobachtung, wären also erwartungstreu. Warum nehmen wir also nicht immer einfach die erste Beobachtung? Der Grund ist, dass Erwartungstreue nicht das einzige interessante Kriterium ist.

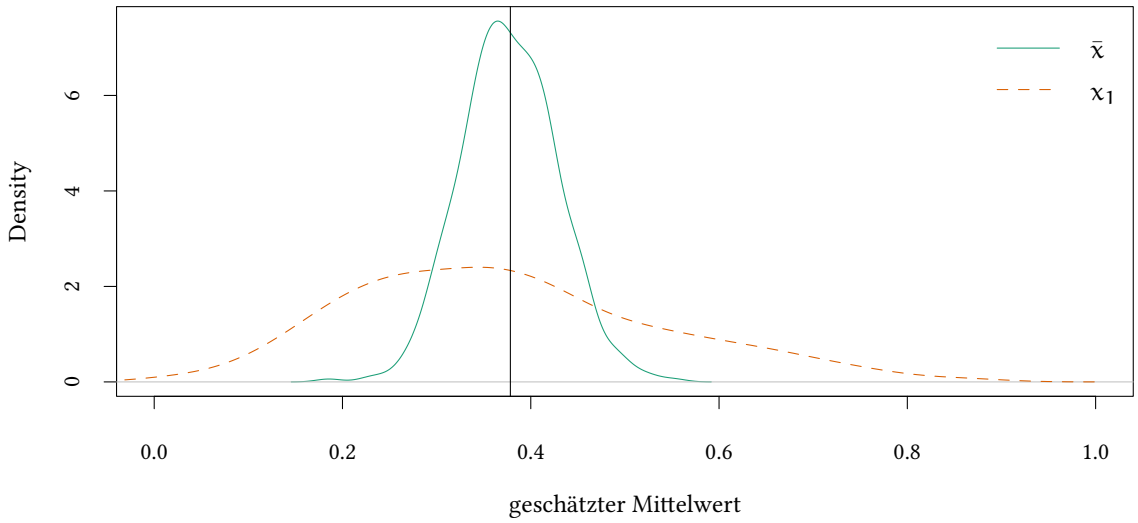
Betrachten wir zur Illustration den Datensatz `wfood` aus der Bibliothek `Ecdat`. In beiden Fällen im nächsten Beispiel nimmt `sample(...,10)` ein Sample der Größe 10 aus dem Datensatz. Im ersten Fall betrachten wir mit `mean(sample(...))` den Mittelwert des Samples, im zweiten Fall betrachten wir mit `sample(...)[1]` nur die erste Beobachtung des Samples.

`replicate(500,...)` sorgt dafür, dass wir dieses Verfahren jeweils 500 mal wiederholen. `density(...)` schätzt eine Dichtefunktion, und `plot(...)` zeichnet diese schließlich. Damit die zweite Dichtefunktion nur als Linie über die erste gezeichnet wird, verwenden wir das Kommando `lines(...)`.

```
data(BudgetFood,package="Ecdat")
attach(BudgetFood)
```

- `Xquer` enthält 500 Mittelwerte von Stichproben (`sample`) der Größe 10.
- `X1` enthält 500 mal jeweils die erste Beobachtung einer Stichprobe.

```
mean(wfood)

[1] 0.378321

Xquer<-replicate(500,mean(sample(wfood,10)))
mean(Xquer)

[1] 0.3772281

X1<-replicate(500,sample(wfood,10)[1])
mean(X1)

[1] 0.3745041

plot(density(Xquer),main="")
lines(density(X1),col=2,lty="dashed")
```



Die Streuung von X_1 ist größer als die Streuung von $\bar{X}$.

Wir sehen, beide Schätzer die wir verwenden, der Mittelwert des Samples und der erste Wert des Samples, sind eingermaßen um den Mittelwert unserer angenommenen Population zentriert. Allerdings hat der erste Wert des Samples eine unangenehm große Streuung.

Diese Streuung messen wir z.B. als mittleren quadratischen Fehler:

2.2. Mittlerer quadratischer Fehler

Ein Schätzer der (wie oben) im Mittel richtig liegt, also erwartungstreu ist, ist schön — Erwartungstreue sagt aber nichts über die Genauigkeit des Schätzers aus. Ein erster Ansatz in dieser Richtung ist der mittlere quadratische Fehler.

Bias eines Schätzers $\hat{\theta}$:

$$\begin{aligned} \text{Bias}(\hat{\theta},\theta) &= \mathsf{E}(\hat{\theta})-\theta \\ -\mathsf{E}(\hat{\theta})+\text{Bias}(\hat{\theta},\theta) &= -\theta \end{aligned}$$

Der *mittlere quadratische Fehler* (MSE) eines Schätzers $\hat{\theta}$ ist

$$\begin{aligned} &\text{MSE}(\hat{\theta}) = \mathsf{E}\left[(\hat{\theta}-\theta)^2\right] \\ &= \mathsf{E}\left[(\hat{\theta}-\mathsf{E}(\hat{\theta})+\text{Bias}(\hat{\theta},\theta))^2\right] \\ &= \mathsf{E}\Big[(\hat{\theta}-\mathsf{E}(\hat{\theta}))^2+2\cdot(\hat{\theta}-\mathsf{E}(\hat{\theta}))\cdot\text{Bias}(\hat{\theta},\theta)+ \\ &\qquad +\text{Bias}(\hat{\theta},\theta)^2\Big] \\ &= \mathsf{E}\left[(\hat{\theta}-\mathsf{E}(\hat{\theta}))^2\right]+2\cdot\mathsf{E}\left[(\hat{\theta}-\mathsf{E}(\hat{\theta}))\right]\cdot\text{Bias}(\hat{\theta},\theta)+ \\ &\qquad +\text{Bias}(\hat{\theta},\theta)^2 \\ &= \text{var}(\hat{\theta})+\text{Bias}(\hat{\theta},\theta)^2 \end{aligned}$$

Bei erwartungstreuen Schätzern: (Bias $= 0$)

$$\text{MSE}(\hat{\theta}) = \mathsf{E}\left[(\hat{\theta}-\theta)^2\right] = \text{var}(\hat{\theta})$$

Wenn wir einen kleinen mittleren quadratischen Fehler (MSE) wollen, dann soll

- $\text{var}(\hat{\theta})$ klein sein (der Schätzer soll eine kleine Varianz haben)
- $\text{Bias}(\hat{\theta},\theta)$ klein sein (der Schätzer soll möglichst unverzerrt sein)

Die Varianz des Stichprobenmittelwertes Wie hängt die Varianz des Stichprobenmittelwerts von der Stichprobengröße n ab?

$$\begin{aligned} \text{var}(\bar{X}) &= \text{var}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n^2}\text{var}\left(\sum_{i=1}^{n}X_i\right) \\ &= \frac{1}{n^2}\left(\sum_{i=1}^{n}\text{var}(X_i)\right)=\frac{1}{n^2}\left(\sum_{i=1}^{n}\text{var}(X)\right)= \\ &= \frac{1}{n^2}\cdot n\cdot\text{var}(X)=\frac{\sigma_X^2}{n} \end{aligned}$$

Je größer die Stichprobe, um so kleiner die Varianz des Stichprobenmittelwertes $\bar{X}$, um so »genauer« ist der Mittelwert also.

2.3. Effizienz (Wirksamkeit)

Es ist mühselig, den genauen mittleren quadratischen Fehler von zwei Schätzern zu vergleichen. Einfacher wäre es, wenn man sich darauf konzentrieren könnte, welcher von zwei Schätzern *immer* genauer ist. Das bringt uns zum Begriff der Effizienz.

Wir vergleichen nun zwei *unverzerrte* Schätzfunktionen $\hat{\theta}_1$, $\hat{\theta}_2$:

$\hat{\theta}_1$ ist effizienter (wirksamer) als $\hat{\theta}_2$ falls

$$\forall\theta : \operatorname{var}(\hat{\theta}_1(X_1,\ldots,X_n)) < \operatorname{var}(\hat{\theta}_2(X_1,\ldots,X_n))$$

Wir sagen auch: $\hat{\theta}_1$ *dominiert* $\hat{\theta}_2$

Der unverzerrte Schätzer $\hat{\theta}$ ist effizienter (wirksamster) Schätzer,
falls für alle unverzerrten Schätzer $\hat{\theta}'$ gilt

$$\forall\theta : \operatorname{var}(\hat{\theta}(X_1,\ldots,X_n)) \leq \operatorname{var}(\hat{\theta}'(X_1,\ldots,X_n))$$

- Mittelwert $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
  $\bar{X}$ ist *effizienter* Schätzer von μ_X
- geschätzte Varianz $\hat{\sigma}_X^2 = s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
  s_X^2 ist *effizienter* Schätzer von σ_X^2.

(Die deutsche Sprache ist hier etwas verwirrend weil »effizienter« sowohl Positiv als auch Komparativ sein kann: Unterscheiden Sie »A ist ein effizienter (wirksamster) Schätzer« (effizienter ist hier Positiv, wirksamer ist Superlativ, es gibt also keinen besseren Schätzer als A) und »B ist effizienter (wirksamer) als C« (effizienter und wirksamer sind hier Komparativ). Im zweiten Fall ist B zwar effizienter als C aber nicht unbedingt »effizient« (am wirksamsten).)

2.4. Konsistenz

Einige Schätzer sind nicht unbedingt erwartungstreu, jedenfalls nicht für kleine Stichproben, wir wissen aber, dass sie bei *größeren* Stichproben immer *besser* werden. Diese Eigenschaft wird durch Konsistenz beschrieben.

Wir betrachten eine Folge von Stichproben. Von Stichprobe zu Stichprobe wird die Größe der Stichprobe n größer. Für jede Stichprobe bestimmen wir nun den Schätzer $\hat{\theta}$

Ein Schätzer ist konsistent falls

$$\forall\epsilon > 0 : \lim_{n\to\infty} \Pr(|\hat{\theta}(X_1,\ldots,X_n) - \theta| < \epsilon) = 1$$

Der Schätzer wird also immer genauer. Diese Definition können wir z.B. auf den Mittelwert anwenden. Dazu brauchen wir das schwache Gesetz der großen Zahl.

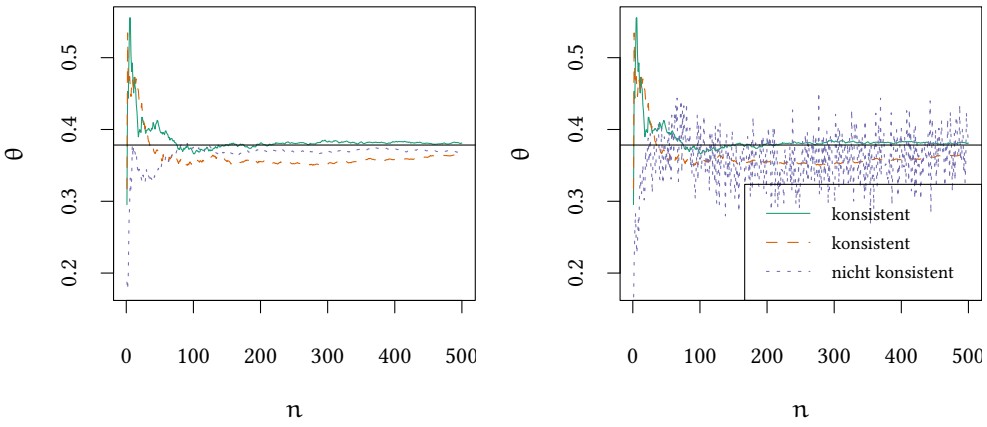
Zur Erinnerung: schwaches Gesetz der großen Zahl:

$$\forall \epsilon > 0 : \lim_{n\to\infty} \Pr\left(\left|\frac{\sum_{i=1}^{n} X_i}{n} - \mu_X\right| < \epsilon\right) = 1$$

→ **der empirische Mittelwert $\bar{X}$ ist ein konsistenter Schätzer für μ_X**
Das kann man wie folgt aufschreiben:

$$\lim_{n\to\infty} \Pr(|\bar{X} - \mu_X| < \epsilon) = 1 \qquad \text{wir sagen auch} \qquad \bar{X} \xrightarrow{p} \mu_X$$

$\xrightarrow{p}$ bedeutet »Konvergenz in Wahrscheinlichkeit«



2.5. Eigenschaften von bekannten Schätzern

2.5.1. Der Stichprobenmittelwert $\bar{X}$ als Schätzer für den Erwartungswert μ_X

Der Stichprobenmittelwert hat viele der oben diskutierten wünschenswerten Eigenschaften:

$$\hat{\mu}_X = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- $\bar{X}$ ist unverzerrt: $E(\bar{X}) = \mu_X$
- $\bar{X}$ ist effizient falls $X \sim N(\mu_X, \sigma_X^2)$: $\text{var}(\bar{X})$ ist klein
- $\bar{X}$ ist konsistent: $\bar{X} \xrightarrow{p} \mu_X$
- $\bar{X}$ ist der »kleinste Quadrate« Schätzer für μ_X, $\bar{X}$ ist die Lösung von

$$\min_{\xi} \sum_{i=1}^{n} (X_i - \xi)^2$$

Das kann man leicht zeigen: Sei ξ eine Zahl, die quadratische Abstände zu unseren Stichprobenbeobachtungen X_i minimiert:

$$\begin{aligned} 0 &\stackrel{!}{=} \frac{d}{d\xi}\sum_{i=1}^{n}(X_i-\xi)^2 = \sum_{i=1}^{n} -2\cdot(X_i-\xi) \\ &= -2\left(\sum_{i=1}^{n} X_i\right) + 2\cdot n\cdot\xi \\ &\Rightarrow \qquad 2\cdot\sum_{i=1}^{n} X_i = 2\cdot n\cdot\xi \qquad \Rightarrow \qquad \frac{\sum_{i=1}^{n} X_i}{n} = \xi \end{aligned}$$

$\bar{X}$ ist also der »kleinste Quadrate« Schätzer für μ_X.

2.5.2. Der Median als Schätzer für den Erwartungswert

Wir haben gesehen: Der Stichprobenmittelwert $\bar{X}$ hat viele schöne Eigenschaften.

- z.B. $\bar{X}$ ist *effizient*, d.h. hat eine kleinere Varianz als alle anderen *unverzerrten* Schätzer.

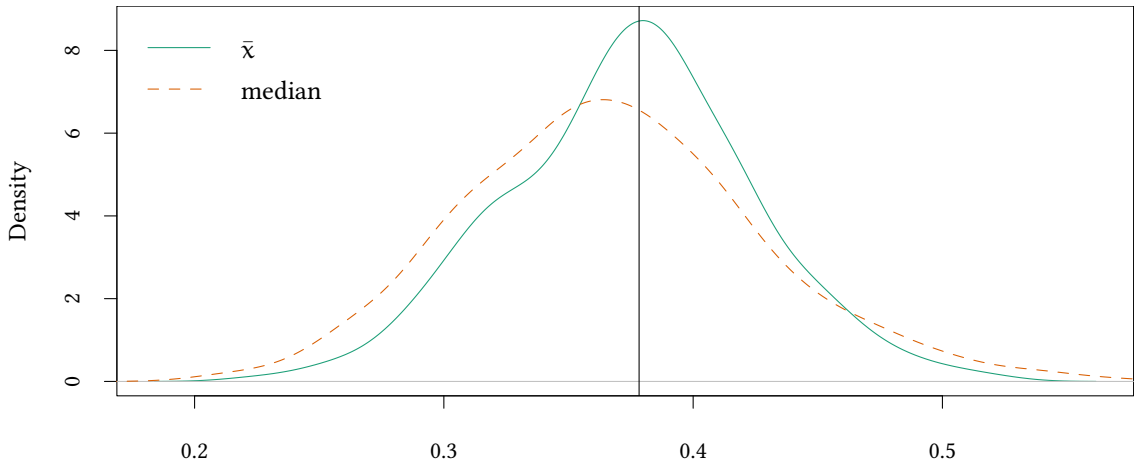
  Für jeden unverzerrten Schätzer $\hat{\mu}_X$ gilt $\text{var}(\bar{X}) \leq \text{var}(\hat{\mu}_X)$

→ Effizienz ist aber nicht alles…

Sei etwa $X \sim N(\mu, 1)$ normalverteilt mit unbekanntem μ.

- der Mittelwert $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ ist erwartungstreu und hat Varianz $\frac{1}{n}$
- der Median ist für große Stichproben etwa normalverteilt mit Mittelwert μ und Varianz $\frac{\pi}{2\cdot n} \approx \frac{1.57}{n}$
- Die Effizienz ist also etwas kleiner als die Effizienz des Mittelwerts (die Varianz ist größer).
- Allerdings ist der Median *robuster*.

```
Xquer<-replicate(500,mean(sample(wfood,10)))
MED<-replicate(500,median(sample(wfood,10)))
```



Die Robustheit des Medians demonstriert das folgende Beispiel:

Übung 2.2 *Wir betrachten wieder die Ausgaben spanischer Haushalte für Lebensmittel*
Was passiert, wenn in jeder Stichprobe bei einer Beobachtung das Komma verrutscht:

Hier ist eine Stichprobe von 9 Beobachtungen:

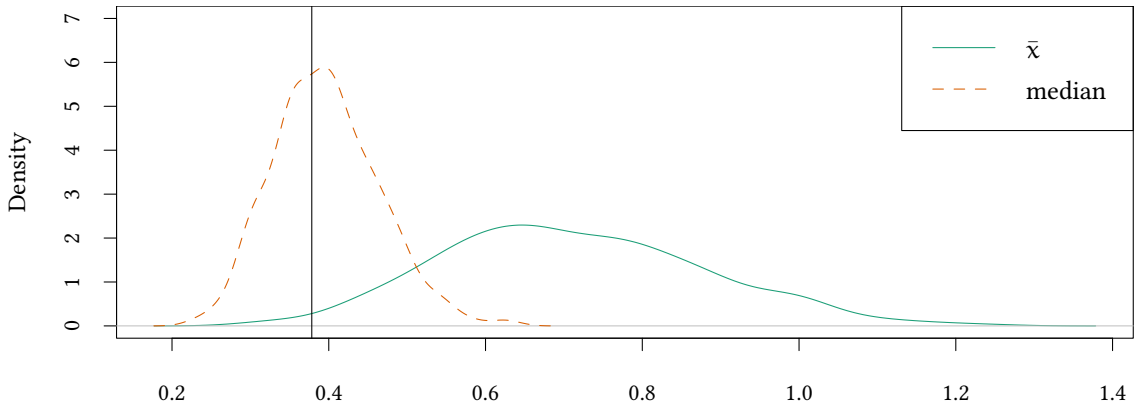
```
[1] 0.0986 0.1476 0.2670 0.3427 0.4003 0.4223 0.4507 0.4883 0.6678
```

($\bar{x} = 0.365$, median=0.4003)
nun wird eine Beobachtung (0.6678) falsch codiert:

```
[1] 0.0986 0.1476 0.2670 0.3427 0.4003 0.4223 0.4507 0.4883 6.6782
```

So verändern sich Mittelwert und Median:
($\bar{x} = 1.0329$, median= 0.4003)
Diese Übung können wir auch mehrmals durchführen. In der folgenden Grafik wird 500 mal eine Stichprobe der Größe gezogen, und jeweils eine Beobachtung »falsch codiert«. Die Grafik zeigt die Verteilung von Mittelwert ($\bar{x}$) und Median:

```
Xquer2<-replicate(500,mean({x<-sample(wfood,10);x[1]<-x[1]*10;x}))
MED<-replicate(500,median({x<-sample(wfood,10);x[1]<-x[1]*10;x}))
```



Wir sehen: Der Mittelwert $\bar{x}$ reagiert sehr sensibel auf Ausreißer (die leicht durch Fehler in den Daten entstehen). Der Median reagiert nicht oder nur unwesentlich.

(Anmerkung: Wenn unsere Zufallsvariable stark asymmetrisch verteilt ist, ist der Median der Stichprobe ein verzerrter Schätzer für den Mittelwert der Population. Der Median der Stichprobe ist aber auch in diesen Fällen ein guter Schätzer für den Median der Population.)

2.5.3. Die Stichprobenvarianz

Auch die Stichprobenvarianz hat viele der oben diskutierten wünschenswerten Eigenschaften:

$$\hat{\sigma}_X^2 = s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\cdot\bar{X}^2\right)$$

$\hat{\sigma}_X^2$ **ist unverzerrt:** $\mathrm{E}(\hat{\sigma}_X^2) = \sigma_X^2$

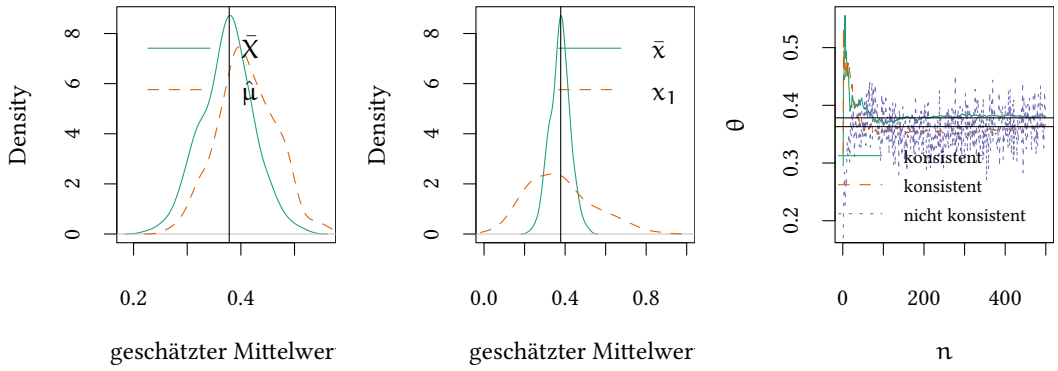
$\hat{\sigma}_X^2$ **ist effizient:** $\mathrm{var}(\hat{\sigma}_X^2) \leq \mathrm{var}(\hat{\sigma}_X^{2\prime})$ für unverzerrte $\hat{\sigma}_X^{2\prime}$

$\hat{\sigma}_X^2$ **ist konsistent:** $\hat{\sigma}_X^2 \xrightarrow{p} \sigma_X^2$ wenn $X_1, \ldots, X_n$ i.i.d. (unabhängig und gleichverteilt) sind und $\mathrm{E}(X^4) < \infty$: (Gesetz der großen Zahl)

Warum gilt hier das Gesetz der großen Zahl?

- $\hat{\sigma}_X^2$ ist ein Stichprobenmittelwert
- Wir fordern hier $\mathrm{E}(X^4) < \infty$ weil der Mittelwert nicht von X_i sondern von X_i^2 gebildet wird.

Übersicht: Wünschenswerte Eigenschaften von Schätzern



Erwartungstreue Effizienz Konsistenz

2.6. Literatur

- Dolić, Statistik mit R, Kapitel 6.1.
- Schira, Statistische Methoden der VWL und BWL-Theorie und Praxis, Kapitel 13.2.

2.7. Schlüsselbegriffe

- Erwartungstreue (Unverzerrtheit)
- Mittlerer quadratischer Fehler
- Effizienz, Wirksamkeit, Dominanz
- Konsistenz

Anhang 2.A Beispiele für die Vorlesung

Welcher Schätzer für $E(X)$ ist effizienter?

- $\frac{X_2}{2} + \frac{X_3}{2}$ ist effizienter als X_3 ?
- X_3 ist effizienter als $X_1 + X_2 - X_3$
- $\frac{1}{5}\sum_{i=1}^{5} X_i$ ist effizienter als $\sum_{i=1}^{5} X_i$
- $\frac{1}{5}\sum_{i=1}^{5} X_i$ ist effizienter als X_3
- $\frac{X_2}{2} + \frac{X_3}{2}$ ist effizienter als $X_3 + X_4 - X_5$

Anhang 2.B Übungen

Übung 2.3 *Vergleichen Sie den MSE des Mittelwerts $\bar{X}$ mit dem MSE des Wertes der ersten Beobachtung X_1?*

Übung 2.4

1. *Ziehen Sie eine Stichprobe mit 100 pseudo-zufälligen Werten.*
2. *Bestimmen Sie den Mittelwert dieser Stichprobe. Wiederholen Sie diese Prozedur.*
3. *Was ist nun der Mittelwert?*
4. *Wiederholen Sie die Prozedur 1000 mal.*
5. *Zeichnen Sie ein Histogramm der Mittelwerte.*
6. *Zeichnen Sie die empirische kumulierte Verteilungsfunktion.*
7. *Vergleichen Sie diese Verteilung mit einer Normalverteilung.*

```
set.seed(123)
runif(100)
mean(runif(100))
replicate(10,mean(runif(100)))
x<-replicate(1000,mean(runif(100)))
hist(x,breaks=20)
plot(ecdf(x))
qqnorm(x)
qqline(x)
```

Übung 2.5 *Wir interessieren uns wie oben für die Ausgaben der spanischen Haushalte für Lebensmittel. Welche Varianz sollten wir in unserem Beispiel für den Mittelwert einer Stichprobe von 10 erwarten? Welche Varianz für den jeweils ersten Wert einer Stichprobe?*

Ziehen Sie jeweils 100 Stichproben und vergleichen Sie.

```
var(wfood)/10
var(replicate(100,mean(sample(wfood,10))))
var(wfood)
var(replicate(100,sample(wfood,10)[1]))
```

Übung 2.6 *Wir ziehen jeweils eine Stichprobe der Größe 20 und nehmen den Median als Schätzer für den Mittelwert der Population. Simulieren Sie dieses Verfahren 100 mal und vergleichen Sie die Varianz dieses Schätzers mit der Varianz des Mittelwerts des Samples.*

Übung 2.7 *Sie entwickeln einen Energiedrink zur Verbesserung des Muskelaufbaus. Den Muskelaufbau pro Person messen wir als X. Sie testen diesen Drink an Versuchspersonen in vier (identischen) Fitness-Studios. Die folgende Tabelle zeigt die Anzahl der Versuchspersonen und den*

durchschnittlichen Muskelaufbau pro Versuchsperson in diesen vier Studios:

	Versuchspersonen	*durchschnittlicher Muskelaufbau*
Studio 1	*10*	$\bar{X}_1$
Studio 2	*20*	$\bar{X}_2$
Studio 3	*30*	$\bar{X}_3$
Studio 4	*30*	$\bar{X}_4$

Nun wollen Sie den Erwartungswert des Muskelaufbaus $E(X)$ *bestimmen (Die* $\bar{X}_1$, $\bar{X}_2$, $\bar{X}_3$, $\bar{X}_4$ *sind Mittelwerte von jeweils unabhängig und identisch verteilten* X*).*

1. *Ist* $\bar{X}_1$ *ein erwartungstreuer Schätzer für den Erwartungswert des Muskelaufbaus* $E(X)$*?*
2. *Hängt Ihre Antwort auf Frage (a) von der Verteilung von* X *ab?*
3. *Was ist ein effizienter (wirksamster) erwartungstreuer Schätzer für den Erwartungswert des Muskelaufbaus pro Person* $E(X)$*.*

Übung 2.9 *Eine Maschine wickelt Toilettenpapier auf Rollen, auf denen es verkauft wird. Die Länge des Papiers, das auf eine Rolle gewickelt wird, ist mit Mittelwert* μ *und Varianz 0,36* m^2 *normalverteilt. Sie stellen fest, dass die Maschine auf die letzten 25 Rollen insgesamt 995 m Papier aufgewickelt hat.*

Wie lautet die beste erwartungstreue Schätzung für μ*?*

Übung 2.11 *Welche Schätzfunktion ist ein erwartungstreuer Schätzer für den Erwartungswert?*

1. $g_1 = \frac{1}{n} \sum_{i=1}^{n} X_i$
2. $g_2 = \frac{1}{n} \sum_{i=1}^{n} \sqrt{X_i^2}$
3. $g_3 = \frac{1}{n} \sum_{i=1}^{n} \sqrt[3]{X_i^3}$
4. $g_4 = \frac{3}{10} \cdot X_1 + \frac{7}{10} \cdot X_n$
5. $g_5 = \frac{1}{10} \cdot X_1 + \frac{8}{10} \cdot X_n$
6. $g_6 = \frac{2}{(n+1)\cdot n} \cdot \sum_{i=1}^{n} X_i \cdot i$

Übung 2.12 *Eine Fast-Food-Kette besitzt 100 Filialen in Deutschland.* $X_1, \ldots, X_{100}$ *seien die durchschnittlichen monatlichen Umsätze der einzelnen Filialen im Jahr 2022 (die Umsätze der einzelnen Filialen sind unabhängig voneinander).* X *sei der monatliche Umsatz einer zufälligen Filiale. Es gilt:* $E(X) = \theta$ *und* $var(X) = \sigma^2$*.*

1. *Welche Schätzfunktionen sind erwartungstreu zum Schätzen von* θ*?*
   a) $g_1 = 10 \cdot X_1 + 90 \cdot X_{99}$

b) $g_2 = \frac{10X_1+90X_{99}}{100}$

c) $g_3 = \frac{1}{100}\sum_{i=1}^{100} X_i$

d) $g_4 = \frac{1}{2}X_5 + \frac{1}{3}X_{20} + \frac{1}{6}X_{37}$

e) $g_5 = \sum_{i=1}^{100} X_i$

2. *Welche der Schätzfunktionen ist am wirksamsten?*

a) $g_1 = 10 \cdot X_1 + 90 \cdot X_{99}$

b) $g_2 = \frac{10X_1+90X_{99}}{100}$

c) $g_3 = \frac{1}{100}\sum_{i=1}^{100} X_i$

d) $g_4 = \frac{1}{2}X_5 + \frac{1}{3}X_{20} + \frac{1}{6}X_{37}$

e) $g_5 = \sum_{i=1}^{100} X_i$

Übung 2.14 *Ein Automobilkonzern ist in zwei Werke aufgeteilt. In Werk I arbeiten 6000 Beschäftigte (= Grundgesamtheit G_1) und in Werk II arbeiten 4000 Beschäftigte (= Grundgesamtheit G_2). Die Vielzahl von neuen Auftragseingängen will die Firmenleitung mit einer neuen Arbeitszeitregelung bewältigen. Der Betriebsrat will die Anteile p_1, p_2 bzw. p der Befürworter der vorgeschlagenen neuen Arbeitszeitregelung in G_1, G_2 bzw. in der Gesamtbelegschaft $G = G_1 \bigcup G_2$ schätzen. Dazu wird in G_1 eine einfache Stichprobe vom Umfang n_1 und in G_2 eine einfache Stichprobe vom Umfang n_2 gezogen, in der jeder Befragte die neue Regelung befürworten kann (= Ergebnis 1) oder nicht (= Ergebnis 0). Es sei $\bar{X}_1$ bzw. $\bar{X}_2$ der zufallsabhängige Anteil der Befürworter in der Stichprobe aus G_1 bzw. G_2.*

1. Welche der folgenden Funktionen sind für beliebige n_1 und n_2 erwartungstreue Schätzer für p, wobei gilt: $p = \frac{6000 \cdot p_1 + 4000 \cdot p_2}{10000}$?

a) $\frac{1}{n_1+n_2} \cdot (n_1 \cdot \bar{X}_1 + n_2 \cdot \bar{X}_2)$

b) $\frac{1}{2} \cdot (\bar{X}_1 + \bar{X}_2)$

c) $\frac{1}{10000} \cdot (6000 \cdot \bar{X}_1 + 4000 \cdot \bar{X}_2)$

2. Von $n_1 = 100$ Befragten aus G_1 waren 40 und von $n_2 = 50$ Befragten aus G_2 waren 30 für die neue Regelung. Was sind effiziente Schätzer für p_1, p_2 und p?

Übung 2.15 *Ein neu entwickeltes Futtermittel soll die Milchproduktion von Kühen anregen. Die Milchproduktion einer Kuh messen wir als X. Es gilt $E(X) = \Theta$ und $var(X) = \sigma^2$. Sie testen das Futtermittel an Kühen in vier (identischen) Milchhöfen. Die folgende Tabelle zeigt die Anzahl der Kühe und die durchschnittliche Milchproduktion pro Kuh in diesen vier Milchhöfen:*

	Anzahl Kühe	*durchschnittliche Milchproduktion*
Milchhof 1	*10*	$\bar{X}_1$
Milchhof 2	*20*	$\bar{X}_2$
Milchhof 3	*30*	$\bar{X}_3$
Milchhof 4	*40*	$\bar{X}_4$

(Die $\bar{X}_1, \bar{X}_2, \bar{X}_3, \bar{X}_4$ sind Mittelwerte von jeweils unabhängig und identischen verteilten X).

1. Welche der folgenden Funktionen sind erwartungstreue Schätzer für den Erwartungswert der Milchproduktion pro Kuh $E(X)$?
   a) $g_1 = 0.7 \cdot \bar{X}_2 + 0.3 \cdot \bar{X}_4$
   b) $g_2 = 0.3 \cdot \bar{X}_1 + 0.9 \cdot \bar{X}_3$
   c) $g_3 = 0.25 \cdot \bar{X}_1 + 0.25 \cdot \bar{X}_2 + 0.5 \cdot \bar{X}_4$
   d) $g_4 = \bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \bar{X}_4$
   e) $g_5 = 0.1 \cdot \bar{X}_1 + 0.2 \cdot \bar{X}_2 + 0.3 \cdot \bar{X}_3 + 0.4 \cdot \bar{X}_4$
   f) $g_6 = \bar{X}_4$

2. Welche der folgenden Funktionen sind erwartungstreue Schätzer für den Erwartungswert der Milchproduktion pro Kuh $E(X)$ *und* sind entweder am wirksamsten oder haben eine Varianz von $\frac{1}{64}\sigma^2$?
   a) $g_1 = 0.7 \cdot \bar{X}_2 + 0.3 \cdot \bar{X}_4$
   b) $g_2 = 0.3 \cdot \bar{X}_1 + 0.9 \cdot \bar{X}_3$
   c) $g_3 = 0.25 \cdot \bar{X}_1 + 0.25 \cdot \bar{X}_2 + 0.5 \cdot \bar{X}_4$
   d) $g_4 = \bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \bar{X}_4$
   e) $g_5 = 0.1 \cdot \bar{X}_1 + 0.2 \cdot \bar{X}_2 + 0.3 \cdot \bar{X}_3 + 0.4 \cdot \bar{X}_4$
   f) $g_6 = \bar{X}_4$

Übung 2.16 *Gegeben ist folgende Schätzfunktion*

$$g(X_1, \ldots, X_n) = \frac{4X_1 + 3X_2 + 2X_3 + X_4}{4 + 3 + 2 + 1}$$

X_i *sind Elemente einer Stichprobe von* X. *Ist die Schätzfunktion* g *erwartungstreu zum Schätzen des Mittelwertes von* X*?*

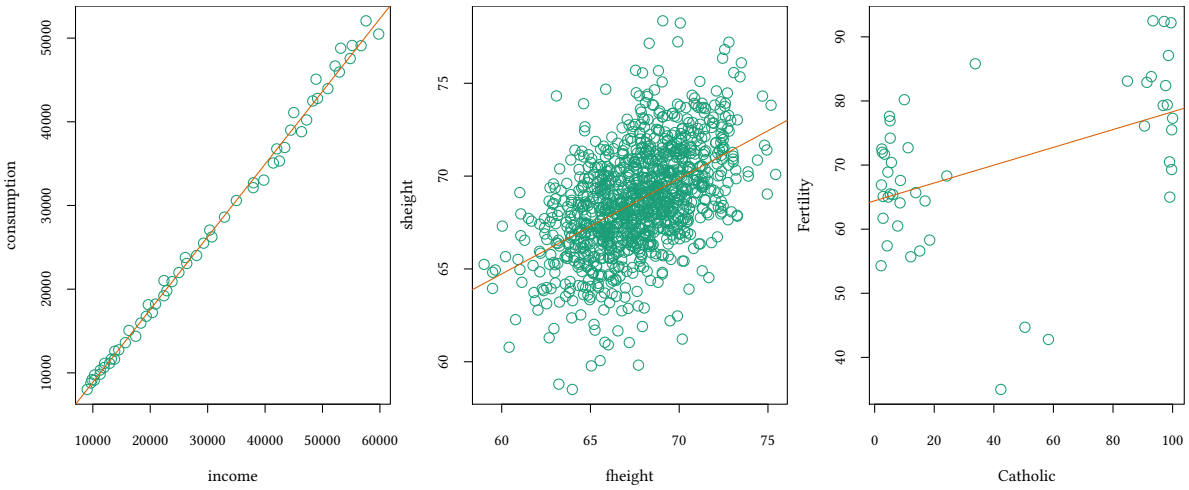
3. Maximum Likelihood und Momentenmethode

3.1. Motivation: Bietverhalten in Englischen Auktionen

In Kapitel 2 haben wir wünschenswerte Eigenschaften von Schätzern diskutiert. In diesem Kapitel stellen wir die Frage, wie man überhaupt einen Schätzer findet. Wir werden zwei Methoden betrachten:

- Maximum Likelihood
- Momentenmethode

Schätzung linearer Zusammenhänge Viele Zusammenhänge lassen sich durch eine gerade Linie approximieren:



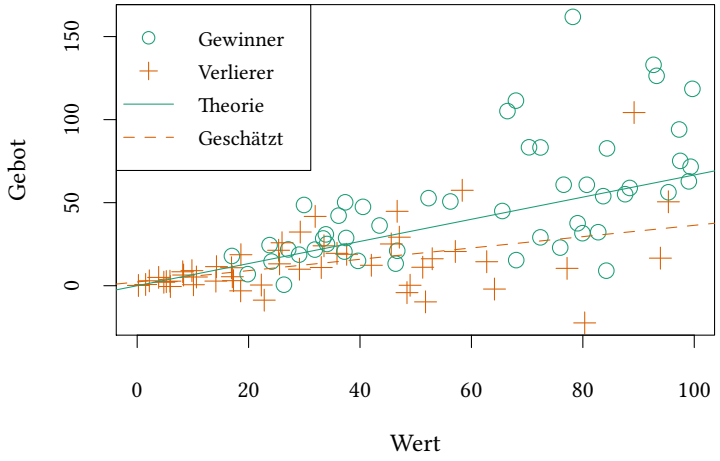
Um den Zusammenhang zu schätzen, nehmen wir an, dass wir alle Ereignisse gleich gut beobachten können. Wir schließen aus, dass es einen *Selektionsbias* gibt.

Manchmal sieht man aber nur einige Beobachtungen. Ein Beispiel sind Gebote in einer Englischen Auktion.

In "An experimental analysis of auctions with interdependent valuation", (Games and Economic Behavior, 2004, Vol. 48/1, pp. 54-85) untersuchen Kirchkamp und Moldovanu, ob Bieter in einer Englischen Auktion von den Geboten ihrer Mitbieter lernen.

Ein Problem bei der Auswertung der Daten ist, dass man die Gebote der Gewinner der Auktion nicht beobachtet. Man sieht zwar, bis zu welchem Preis die *Verlierer* Auktion geboten haben, man sieht aber nicht, bis zu welchem Preis die *Gewinner* geboten hätten. Man weiß nur, dass diese Betrag höher sein muss, als der Betrag, den sie tatsächlich gezahlt haben.

In der folgenden Grafik sind die Maximalbeträge eingezeichnet, bis zu denen einige Spieler bieten wollen. Die Gebote der Verlierer, die man sieht, sind mit einem + gekennzeichnet. Die Gebote der Gewinner, die etwas größer sind, kann man normalerweise nicht beobachten. In der Grafik haben wir sie mit einem ○ gekennzeichnet.



- Frage: Lernen Bieter in Englischen Auktionen von den Geboten ihrer Mitbieter.
- Problem: Man beobachtet nur die Gebote der Verlierer
- Lösung: Likelihoodverfahren: Man verschiebt die zu schätzende Bietfunktion so lange, bis das Ergebnis plausibel ist (und auch mit den nicht-beobachteten Geboten der Gewinner konsistent ist).

Kirchkamp, Moldovanu, Games and Economic Behavior, 2004.

3.2. Maximum Likelihood Methode

- Wir betrachten wieder eine Zufallsvariable $X \sim f(x|\theta)$ mit unbekanntem Parameter θ.
- Nun ziehen wir eine Stichprobe $X_1, \dots, X_n$.
- Die Wahrscheinlichkeit (Likelihood), ein bestimmtes $x_1, \dots, x_n$ zu ziehen, hängt von θ ab.

  Betrachten wir zunächst nur eine Beobachtung X_i:

$$\text{diskret:} \quad P(X_i = x_i|\theta) = \begin{cases} \text{Wahrscheinlichkeit, dass } X_i = x_i \text{ falls } \theta \\ \text{der wahre Parameter ist} \end{cases}$$

$$\text{stetig:} \quad f(x_i|\theta) = \begin{cases} \text{Dichte von } X_i \text{ falls} \\ \theta \text{ der wahre Parameter ist} \end{cases}$$

  Der Einfachheit halber schreiben wir hier immer $f(x_i|\theta)$, auch im diskreten Fall (und nicht $P(X_i = x_i|\theta)$).

Terminologie: $f(x|\theta)$ ist eine Dichtefunktion, keine Wahrscheinlichkeit. Bei einer stetigen Verteilung ist die *Wahrscheinlichkeit*, dass X gerade den Wert x annimmt, Null.

Aber die Wahrscheinlichkeit, dass X einen Wert in der Nähe von x annimmt, ist proportional zu $f(x|\theta)$. Wir werden uns im folgenden damit begnügen, etwas auszurechnen, das nur *proportional* zur Wahrscheinlichkeit ist, unsere Stichprobe $x_1, \ldots, x_n$ zu beobachten. Diesen Ausdruck nennen wir *Likelihood*.

Da Likelihood und Wahrscheinlichkeit proportional zueinander sind, wissen wir, dass, wenn wir (weiter unten) die Likelihood maximieren, wir auch die Wahrscheinlichkeit maximieren.

3.3. Likelihoodfunktion

Likelihood (gegeben θ) ein bestimmtes x_1 zu ziehen: (Hier verwenden wir Kleinbuchstaben für x_i um anzudeuten, dass uns die x_i schon bekannt sind.)

$$f(x_1|\theta)$$

Im nächsten Schritt betrachten wir nicht mehr die Wahrscheinlichkeit eines einzelnen Ereignisses (eines einzelnen x_i, sondern fragen nach der Wahrscheinlichkeit der gesamten Stichprobe:

Likelihood (gegeben θ) eine Stichprobe $x_1, x_2, x_3, \ldots, x_n$ zu ziehen:

$$f(x_1|\theta) \cdot f(x_2|\theta) \cdot f(x_3|\theta) \cdots f(x_n|\theta)$$

Likelihoodfunktion (gegeben eine Stichprobe $x_1, x_2, x_3, \ldots, x_n$):

$$L(x_1, \ldots, x_n|\theta) \equiv f(x_1|\theta) \cdot f(x_2|\theta) \cdot f(x_3|\theta) \cdots f(x_n|\theta)$$

Zwei Interpretationen:

- θ ist gegeben: $L(x_1, \ldots, x_n|\theta)$ ist die Likelihood für das Eintreten von $(x_1, \ldots, x_n)$. Da wir θ normalerweise nicht kennen, hilft uns diese Interpretation nicht.
- $(x_1, \ldots, x_n)$ ist gegeben: $L(x_1, \ldots, x_n|\theta)$ ist die Likelihood, dass sich $(x_1, \ldots, x_n)$ realisiert, falls θ der wahre Parameter ist. Da wir $(x_1, \ldots, x_n)$ kennen, können wir diesen Ansatz nutzen, um θ zu schätzen.

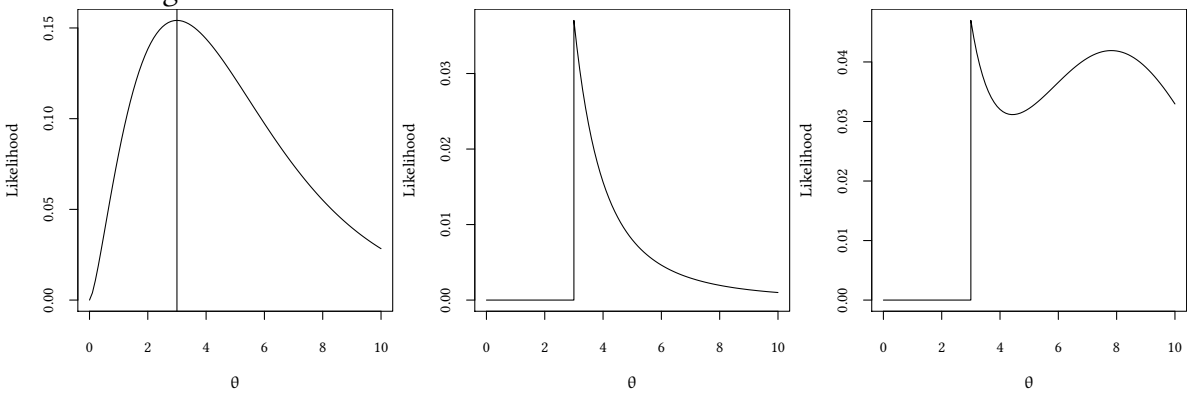
Die Maximum-Likelihood (ML) Methode

- $f(X|\theta)$ Dichtefunktion
- Θ ist die Menge der möglichen Parameter für θ
- Definition: $\hat{\theta}$ ist Maximum-Likelihood (ML) Schätzer genau dann wenn $\forall\theta \in \Theta$:

$$L(x_1, \ldots, x_n|\hat{\theta}) \geq L(x_1, \ldots, x_n|\theta)$$

Wir nehmen als *Schätzwert* für θ den Wert $\hat{\theta}$, für den, gegeben die (angenommene) Verteilungsfunktion $f(X|\theta)$, die Realisierungen $x_1, \ldots, x_n$ am wahrscheinlichsten sind.

Abhängig von der Form von $f(X|\theta)$ kann dieser Schätzerwert mehr oder weniger typisch für die Verteilung von θ sein.



3.4. Log-Likelihood Funktion

Bislang haben wir die Likelihoodfunktion maximiert:

$$L(x_1,\dots,x_n|\theta) \equiv f(x_1|\theta)\cdots f(x_n|\theta)$$

Diese Produkte können sehr lang werden, und dann wird das Ableiten (für die Maximierung) schwierig. Wir maximieren statt dessen die Log-Likelihoodfunktion

$$\begin{aligned}\log L(x_1,\dots,x_n|\theta) &= \log\left(f(x_1|\theta)\cdots f(x_n|\theta)\right) = \\ &\quad \log f(x_1|\theta)+\cdots+\log f(x_n|\theta)\end{aligned}$$

Jetzt leiten wir nur noch eine Summe ab. Das Ableiten der einzelnen Summanden ist oft einfacher.

$$\frac{d}{d\theta}\log L(x_1,\dots,x_n|\theta) = \frac{f'(x_1|\theta)}{f(x_1|\theta)}+\cdots+\frac{f'(x_n|\theta)}{f(x_n|\theta)}$$

Wozu braucht man das Maximum Likelihood Verfahren in der Praxis? Wenn die Annahmen der einfacherern Modelle (Kleinste Quadrate Schätzer) verletzt sind, z.B.:

- Discrete Choice Modelle
  - Logistische Regression (Arbeitslosigkeit, Unternehmensinsolvenz,…)
  - Zähldaten (Anzahl Patente pro Firma, Seitensprünge pro Person, Kinder pro Haushalt,…)
  - Intervalldaten (sozialversicherungspflichtiges Einkommen, Gebote in Auktionen,…)
  - ⋮
- Random Effects Modelle
  - z.B. mehrere Beobachtungen der gleichen Beobachtungseinheit
- Quantilsregression
  - »Ausreißer« im Datensatz.

Beispiel Wir nutzen die Maximum Likelihood Methode um den Mittelwert zu schätzen:

$$X \sim N(\mu, \sigma^2)$$

$$\Pr(\mathbf{X}|\mu,\sigma) = \prod \text{dnorm}(X_i|\mu,\sigma)$$

$$\text{LL} = \sum \log\left(\text{dnorm}(X_i|\mu,\sigma)\right)$$

```
X <- c(1,2,3)
LL <- function(theta)
    -sum(log(dnorm(X,mean=theta[1],sd=theta[2])))
optim(c(0,1),LL)

$par
[1] 1.9999124 0.8164934

$value
[1] 3.648618

$counts
function gradient
      77       NA

$convergence
[1] 0

$message
NULL
```

Alternativ könnte man Stichprobenmittelwert und Standardabweichung verwenden:

```
mean(X)

[1] 2

sd(X)

[1] 1
```

Der ML Schätzer ergibt (fast) das gleiche Ergebnis für μ, aber ein anderes Ergebnis für σ.

Eigenschaften des ML Schätzers (mit Standardannahmen):

- Konsistenz ($\hat{\theta}_{ML} \xrightarrow{p} \theta$)
- Asymptotisch normalverteilt ($\hat{\theta}_{ML} \xrightarrow{p} N(\theta, \frac{1}{n} I^{-1})$)

Ein anderes Verfahren, um Schätzer auszurechnen, ist die Momentenmethode:

3.5. Momentenmethode

Zur Erinnerung: Momente:

$$\begin{aligned} &n\text{-tes } \textit{nicht} \text{ zentriertes Moment von } X: \quad \mu'_n = E(X^n) \\ &n\text{-tes zentriertes Moment von } X: \quad \mu_n = E\left((X-\mu'_1)^n\right) \end{aligned}$$

Wir kennen bereits die ersten vier Momente als Characteristica einer Verteilung (hier nur zur Illustration):

Mittelwert:	$\mu'_1 = E(X)$
Varianz:	$\mu_2 = E\big((X-E(X))^2\big)$
Schiefe:	$\frac{\mu_3}{(\mu_2)^{(3/2)}}$
Wölbung:	$\frac{\mu_4}{(\mu_2)^2}$

Unterscheide:

Momente der Population		Momente der Stichprobe
(unbekannt)	$\Longleftarrow$	(bekannt)
$\mu_1(\theta)$, $\mu_2(\theta)$, $\mu_3(\theta)$, ...		$m_1 = \bar{X}$, m_2, m_3, ...

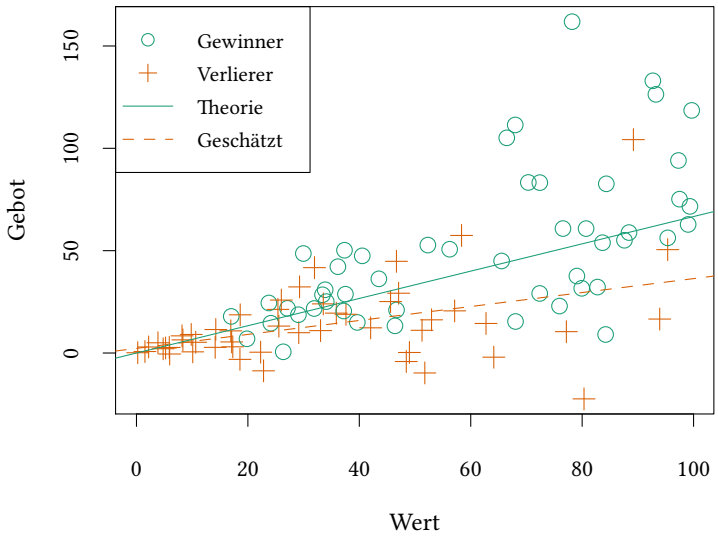
Die *Population* habe eine Dichtefunktion $f(X|\theta)$ mit unbekanntem Parameter θ:

- $f(X|\theta)$ hat Momente $\mu_1(\theta)$, $\mu_2(\theta)$, $\mu_3(\theta)$,...
- Momentenmethode:
  - Bestimme zunächst die Momente der Stichprobe m_1, m_2, m_3...(beginne normalerweise mit dem Mittelwert m_1)
  - Setze die Momente der Stichprobe gleich den Momenten der Population...

    $$\mu_i(\hat{\theta}) = m_i$$

    ...und löse nach $\hat{\theta}$ auf.

Zurück zur Motivation:



Mit einem Likelihood Schätzverfahren kann man zeigen, dass Bieter in einer Englischen Auktion *Informationen* aus den Geboten anderer Bieter *erschließen* und ihre Gebote entsprechend anpassen. Die Allokation unter der Englischen Auktion (bei der andere Gebote sichtbar sind) ist *effizienter* als die Allokation etwa einer Zweitpreisauktion (bei der die anderen Gebote während der Auktion nicht sichtbar sind).

Kirchkamp, Moldovanu, Games and Economic Behavior, 2004.

3.6. Literatur

- Dolić, Statistik mit R, Kapitel 6.2.1.
- Hartung, Statistik, Kapitel III.1.
- Schira, Statistische Methoden der VWL und BWL-Theorie und Praxis, Kapitel 13.3.

3.7. Schlüsselbegriffe

- Likelihoodfunktion eines Parameters θ
- Log-Likelihoodfunktion eines Parameters θ
- Maximum-Likelihood-Methode zur Bestimmung eines Parameters θ
- Momentenmethode zur Bestimmung eines Parameters θ

Anhang 3.A Beispiele für die Vorlesung

Beispiel 1: Eine Zufallsvariable ist wie folgt verteilt: $P(X = -1) = \theta$, $P(X = 1) = \theta$, $P(X = 0) = 1 - 2\theta$. Eine Stichprobe ergibt die Beobachtungen $\{-1, 0\}$:
Was ist der Maximum-Likelihood Schätzer für θ?

Beispiel 2: Die Zufallsvariable X folgt einer Verteilung $\mathcal{X}_\theta$ mit Erwartungswert $E(X) = 1/\theta$ und Varianz θ^2. Es gilt $\theta > 0$. Die Variable x enthält Ihre Stichprobe. Wie berechnen Sie mit R den Momentenschätzer für θ auf Basis des zweiten Moments?

- Keine der folgenden Antworten ist richtig.
- `1/mean(x)`
- `1/sd(x)`
- `sd(x)`
- `sd(x)/2`

Beispiel 3: Eine Zufallsvariable ist wie folgt verteilt: $P(X = 1) = \theta$, $P(X = 2) = \theta$, $P(X = 3) = 1 - 2\theta$ wobei $\theta \in [0, 1/2]$. Eine Stichprobe ergibt die Beobachtungen $\{1, 2, 1, 1, 1, 1\}$.
Was ist der Maximum-Likelihood Schätzer für θ?

Beispiel 4: Eine Zufallsvariable ist gleichverteilt (rechteckverteilt) über dem Intervall $[a, b]$. Ihre Stichprobe enthält drei Beobachtungen: 1, 2 und 3.

1. Bestimmen Sie den Maximum-Likelihood-Schätzer für a und b.
2. Ihre Stichprobe enthält wieder drei Beobachtungen: 1, 3 und 5. Außerdem wissen Sie, dass $b - a = 10$ ist. Was ist der Momentenschätzer für a auf Basis des ersten Moments?

Beispiel 5: X ist gleichverteilt (rechteckverteilt) über dem Intervall $[a, b]$. Wir suchen die Obergrenze dieses Intervalls: b. Das erste Moment der Gleichverteilung ist $E(X) = (a+b)/2$. Die Varianz der Gleichverteilung ist $\text{var}(X) = (b - a)^2/12$ Unsere Stichprobe enthält 9 unabhängige Beobachtungen: $X_1, \ldots, X_9$.

1. Welche Schätzfunktionen für b sind erwartungstreu?
   - X_9
   - $X_1 + X_7 - a$
   - $2X_1 + 2X_7 - 2X_9 - a$
   - $X_1 - X_7 - X_9 + 2 \cdot a$
   - $a + X_1 + X_7 - X_9$
2. Welche Schätzfunktionen für b sind effizienter?

- $X_2 - a + X_3$ ist effizienter als $\frac{1}{2}(4X_1 - 2a)$
- $X_2 - a + X_3$ ist effizienter als $2 \cdot X_1 + X_7 - X_6 - a$
- $\frac{1}{2}(4X_1 - 2a)$ ist effizienter als $2 \cdot X_1 + X_7 - X_6 - a$
- $2 \cdot X_1 - a$ ist effizienter als $2 \cdot X_1 + X_7 - X_6 - a$
- $-a + \frac{2}{9}\sum_i^9 X_i$ ist effizienter als $2 \cdot X_1 + X_7 - X_6 - a$

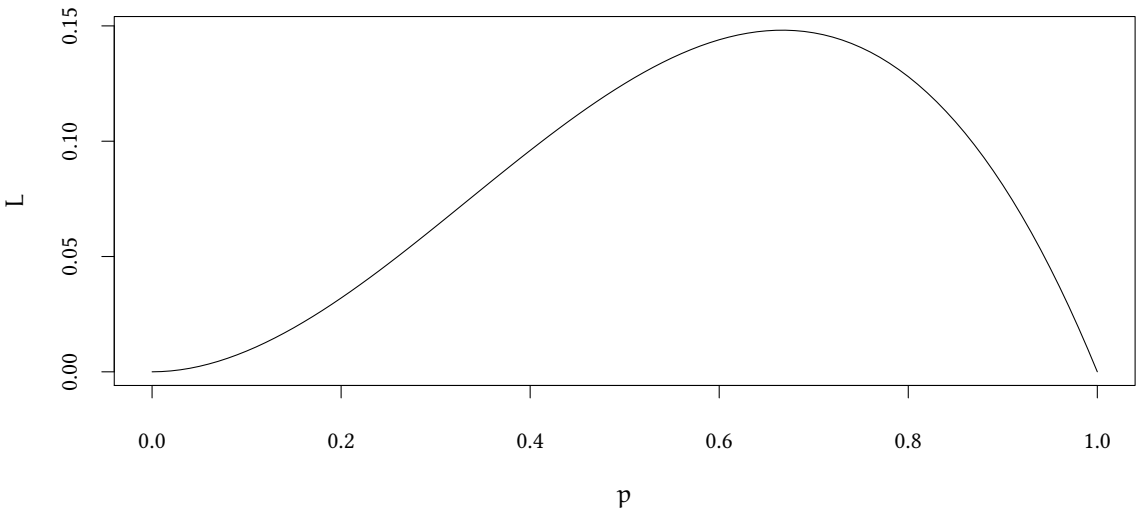
Anhang 3.B Übungen

Die folgenden Hilfestellungen wären auch in der Klausur gegeben:

- Die Exponentialverteilung $X \sim Ex(\lambda)$ hat die Momente $E(x) = \frac{1}{\lambda}$ und $var(x) = \frac{1}{\lambda^2}$.
- Die Binomialverteilung $X \sim B(n, p)$ hat die Momente $E(x) = n \cdot p$ und $var(x) = n \cdot p \cdot (1 - p)$.
- Die Wahrscheinlichkeit, dass bei einer binomialverteilten Zufallsvariable $X \sim B(n, p)$ mit n Versuchen und der Erfolgswahrscheinlichkeit p genau k Erfolge eintreten, ist $P(x = k) = \binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k}$ mit $\binom{n}{k} = \frac{n!}{(n-k)! \cdot k!}$
- Der Erwartungswert einer Poissonverteilten Zufallsvariablen $P(X = x) = \frac{\lambda^x}{x!} \cdot e^{-\lambda}$ ist $E(X) = \lambda$.

Übung 3.1 $X \sim B(1, p)$ *ist binomialverteilt, mit Stichprobengröße 1 und unbekanntem* p. *Drei hintereinander ausgeführte Stichproben ergeben* $X = \{0, 1, 1\}$. *Was ist der ML-Schätzwert für* p*?*

```
curve((1-x)*x^2,from=0,to=1,xlab="$p$",ylab="$L$")
```

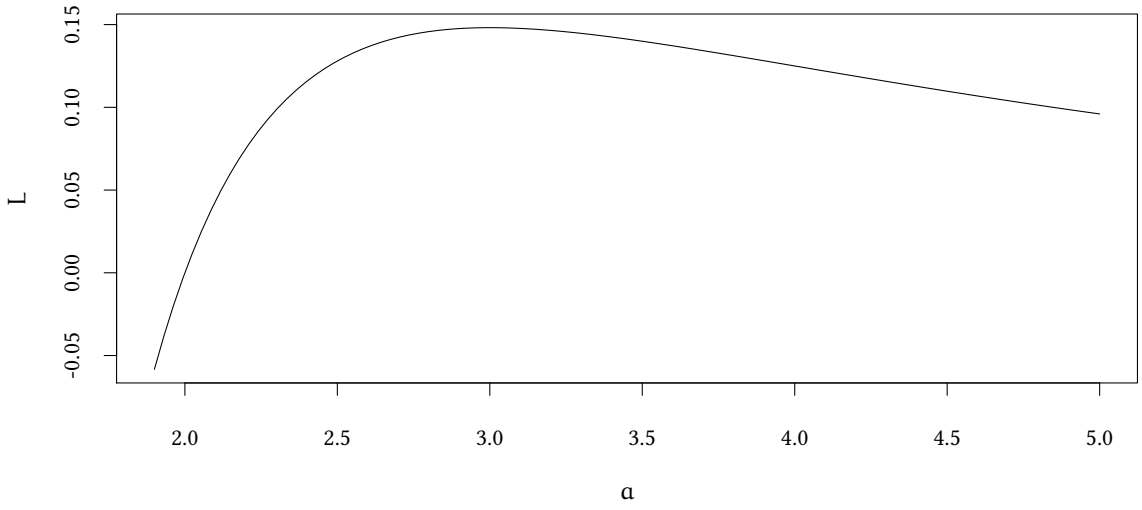


Übung 3.2 $X \sim D(0,a)$ *ist verteilt wie folgt*

$$f(X,a) = \begin{cases} \frac{2}{a} - \frac{2x}{a^2} & \text{falls } x \in [0,a] \\ 0 & \text{sonst} \end{cases}$$

Eine Stichprobe ergibt $X = \{0,2\}$*. Was ist der ML-Schätzwert für* a*?*

```
curve(4*(x-2)/x^3,from=1.9,to=5,xlab="$a$",ylab="$L$")
```



Übung 3.3 *Eine Zufallsvariable ist wie folgt verteilt:* $P(X=1) = \theta, P(X=2) = 1-\theta-\theta^2, P(X=3) = \theta^2$ *mit* $0 \leq \theta \leq \frac{6}{10}$*. Eine Stichprobe ergibt die Werte* $\{1,1,2,3\}$*. Was ist der Maximum-Likelihood-Schätzer für* θ*!*

Übung 3.4 $X \sim B(1,p)$ *ist binomialverteilt, mit Stichprobengröße 1 und unbekanntem* p*. Drei hintereinander ausgeführte Stichproben ergeben* $X = \{0,1,1\}$*. Schätzen Sie den Parameter* p *mit der Momentenmethode. Der Mittelwert der Binomialverteilung ist* $n \cdot p$*.*

Übung 3.5 *Gegeben sei eine binomialverteilte Zufallsvariable mit* $X \sim B(100,\theta)$*. Eine Stichprobe ergab die Werte (20, 35, 14).*

1. *Bestimmen Sie mit R den Schätzwert für* θ *durch die Momentenmethode.*
2. *Wie groß wäre der Schätzer für* θ*, wenn man die Beobachtungen* $(1,19,4)$ *hätte?*

Übung 3.6 *In einer Urne befinden sich schwarze und weiße Kugeln. Um herauszufinden wie groß der Anteil weißer Kugeln ist, zieht man 10 Kugeln mit Zurücklegen.* X *sei die Anzahl gezogener weißer Kugeln* $(X \sim B(10,p))$*. Danach legt man alle Kugeln wieder zurück und zieht nochmals 10 Kugeln. Das macht man insgesamt viermal. Dabei zog man beim ersten Versuch 5, im zweiten 7, im dritten 3 und im vierten 5 weiße Kugeln.*

1. *Schätzen Sie den Parameter p mit der Momentenmethode!*

2. *Berechnen Sie den Wert nun mit der Likelihoodmethode. Ändert sich dabei der Schätzwert für p?*

Übung 3.7 $X \sim D(0,a)$ *ist verteilt wie folgt*

$$f(X,a) = \begin{cases} \frac{2}{a} - \frac{2x}{a^2} & \text{falls } x \in [0,a] \\ 0 & \text{sonst} \end{cases}$$

Eine Stichprobe ergibt $X = \{0,2\}$. Benutzen Sie die Momentenmethode um a zu schätzen. Das erste Moment der obigen Verteilung ist $\mu = \frac{a}{3}$. Das zweite Moment der obigen Verteilung ist $\sigma^2 = \frac{1}{18}a^2$.

Übung 3.8 $X \sim N(\mu,\sigma^2)$ *ist normalverteilt. Die Parameter μ und σ sind unbekannt. Eine Stichprobe ergibt $X = \{0,1,1\}$. Schätzen Sie den Parameter μ mit der Momentenmethode.*

Übung 3.9 *Die Dichtefunktion der Zufallsvariablen X ist*

$$f(x) = \begin{cases} 1 - \frac{x-\theta}{2} & \text{falls } x \in [\theta, \theta+2] \\ 0 & \text{sonst} \end{cases}$$

Der Erwartungswert $E(X) = \frac{2}{3} + \theta$. Der Parameter θ ist unbekannt und soll geschätzt werden.

1. *Bestimmen Sie den Maximum-Likelihood-Schätzer für θ, wenn Ihre Stichprobe die beiden Werte $\{2,3\}$ enthält.*

2. *Bestimmen Sie den Momentenschätzer für θ (auf Basis des ersten Moments), wenn Ihre Stichprobe die beiden Werte $\{2, 3\frac{1}{3}\}$ enthält.*

Übung 3.10 *Für die Verteilungsfunktion einer diskreten Zufallsvariablen X gilt*

$$P(X=x) = \begin{cases} \frac{1}{4} & \text{falls } x = \theta \\ \frac{1}{2} & \text{falls } x = \theta+1 \\ \frac{1}{4} & \text{falls } x = \theta+2 \\ 0 & \text{sonst} \end{cases}$$

Der Parameter θ ist unbekannt und soll geschätzt werden.

1. *Bestimmen Sie den Maximum-Likelihood-Schätzer für θ, wenn Ihre Stichprobe die drei Werte $\{2,2,3\}$ enthält.*

2. *Bestimmen Sie den Momentenschätzer für θ (auf Basis des ersten Moments)*

Übung 3.11 *Eine Zufallsvariable X nimmt die Werte 1,2,3 und 4 an. Dabei gilt*

- $P(X = 1) = \theta^2$
- $P(X = 2) = \theta$
- $P(X = 3) = \theta$
- $P(X = 4) = 1 - 2 \cdot \theta - \theta^2$

mit $\theta \in [0; \frac{2}{5}]$.

Ihnen liegt folgende Stichprobe der Zufallsvariablen vor: (2,1,3,1,4,4,1,1,4,1)
Berechnen Sie den Maximum-Likelihood-Schätzer für θ*!*

Übung 3.12 *Eine Zufallsvariable* X *sei binomialverteilt mit* $n = 150$ *und* $p = \theta$*. Ein Stichprobe fällt folgendermaßen aus: 51, 105, 71, 22, 63. Welche Schätzung erhalten Sie für* θ *mit der Maximum-Likelihood-Methode?*

Übung 3.13 *Gegeben sei* $X \sim B(80, p)$*. Bestimmen sie den Momentenschätzer* θ *für* p *auf Basis des ersten Moments mit Hilfe der folgenden Beobachtung: (1,8,15).*

Hilfe: Der Mittelwert der Binomialverteilung mit Stichprobengröße n *und Erfolgswahrscheinlichkeit* p *ist* $n \cdot p$*. Die Varianz ist* $n \cdot p \cdot (1 - p)$*.*

Übung 3.15 *Eine Zufallsvariable* X *sei exponentialverteilt mit* $X \sim Ex(\lambda)$*. Es liegen folgende Beobachtungen vor: 1, 8, 12.*

1. *Bestimmen Sie in R den Momentenschätzer für* λ *auf Basis des ersten Moments.*
2. *Bestimmen Sie den Momentenschätzer für* λ *auf Basis des zweiten Moments.*

Übung 3.16

1. *Die Zufallsvariable* X *ist Poisson-verteilt. Sie haben folgende Stichprobe erhoben:* $(0, 1, 1, 0, 2)$*. Berechnen Sie den Maximum-Likelihood-Schätzer zum Schätzen des Parameters* λ*.*
2. *Eine andere Stichprobe ist* $(2, 3, 1, 0, 2, 1, 1, 0)$*. Berechnen Sie den Parameter* λ *mit Hilfe der Momentenmethode auf Basis des ersten Moments.*

Übung 3.17 *Sie interessieren sich für den Erwartungswert von* X*. Ihre Stichprobe enthält 9 unabhängige Beobachtungen:* $X_1, \ldots, X_9$*.*

1. *Welche Schätzfunktionen für* $E(X)$ *sind erwartungstreu?*
   - X_9
   - $X_1 + X_7 - X_9$
   - $2X_1 + 2X_7 - 3X_9$
   - $X_1 - X_7 - X_9$
   - $2X_1 + 3X_7 - 2X_9$
2. *Welche Schätzfunktionen für* $E(X)$ *sind effizient?*

- X_9
- $X_1 + X_7 - X_9$
- $2X_1 + 2X_7 - 3X_9$
- $\frac{1}{9}\sum_i^9 X_i$
- $\sum_i^9 X_i$

3. *Welche Schätzfunktionen für* $E(X)$ *sind effizienter?*
   - X_9 *ist effizienter als* $X_1 + X_7 - X_9$
   - $X_1 + X_7 - X_9$ *ist effizienter als* $2 \cdot X_9 - X_8$
   - X_1 *ist effizienter als* $(X_2 + X_3)/2$
   - $\frac{1}{9}\sum_i^9 X_i$ *ist effizienter als* $\frac{1}{7}\sum_i^7 X_i$
   - X_3 *ist effizienter als* $\frac{1}{7}\sum_i^7 X_i$

Übung 3.18 *Eine Zufallsvariable ist wie folgt verteilt:* $P(X = 1) = \theta$, $P(X = 3) = \theta$, $P(X = 5) = 1 - 2\theta$. *Es gilt* $0 \leq \theta \leq \frac{1}{2}$.

1. *Ihre Stichprobe enthält zwei Beobachtungen:* 1 *und* 3. *Was ist der Maximum-Likelihood-Schätzer für* θ*?*

2. *Ihre Stichprobe enthält nun drei Beobachtungen:* 1, 3 *und* 5. *Was ist der Momentenschätzer auf Basis des ersten Moments für* θ*?*

4. Bayesianische Inferenz

4.1. Einführung

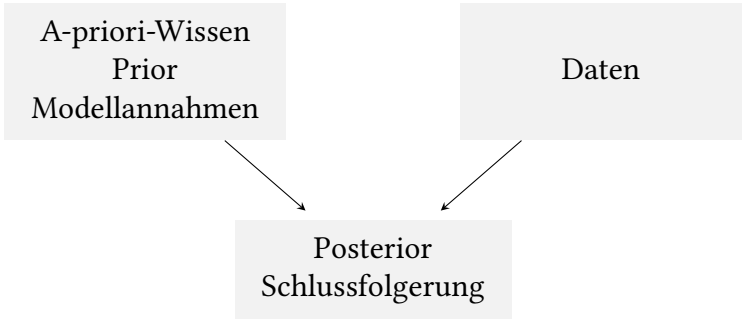
Wir können Inferenz wie folgt zerlegen:

- Wissen, das wir bereits haben. Das nennen wir auch »A-priori-Wissen« oder Prior. Dazu gehört Wissen über Modelle, oder, ganz allgemein, über Parameter.

  Zuweilen ist dieses Wissen unvollständig. Wir wissen bestimmte Dinge nicht genau, und wollen sie genauer wissen.

- Evidenz, die wir beobachten. Oft nennen wir diese Evidenz auch »Daten«. Diese Evidenz hilft uns, unser Wissen zu präziseren.

Das Wissen, das wir nach Zusammenfügen von A-prior-Wissen und Daten erschließen können, nennen wir »Posterior«.




```
library("Ecdat")
data("Caschool")
```

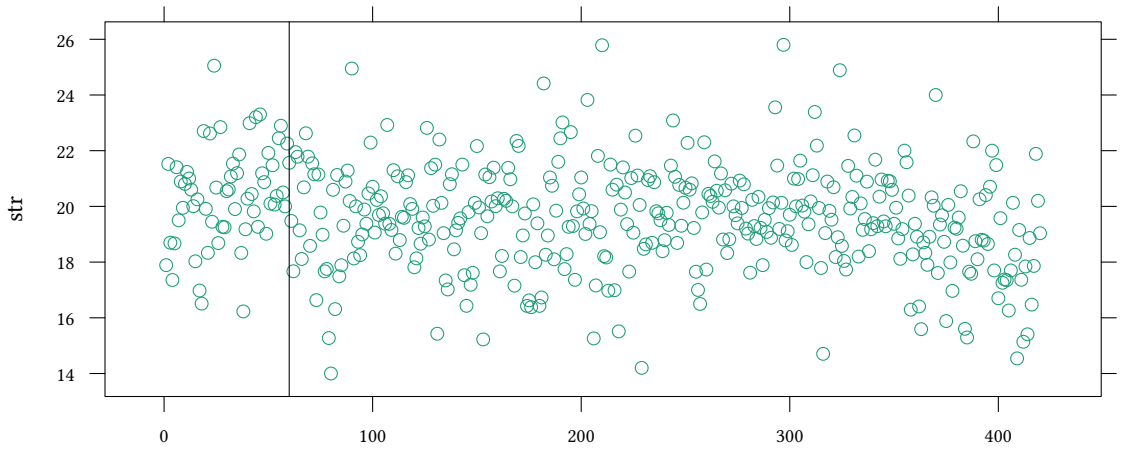
Beispiel Die Datensatz `Caschool` enthält Daten über 420 Bezirke Kaliforniens aus der Zeit 1998/99. Die Variable `str` enthält das »Student/Teacher Ratio«, also wie viele Schüler auf jeden Lehrer entfallen.

Nehmen wir an, wir interessieren uns für den Mittelwert von `str`:

```
mean(Caschool$str)

[1] 19.64043
```

Im Mittel entfallen also auf jeden Lehrer 19.64043 Schüler.

Die folgende Grafik zeigt für jeden der 420 Bezirke den Wert von `str` in diesem Bezirk.



Stellen wir uns vor, wir hätten keine Information für *alle* 420 Bezirke. Statt dessen hätten wir nur eine Stichprobe der ersten 60.

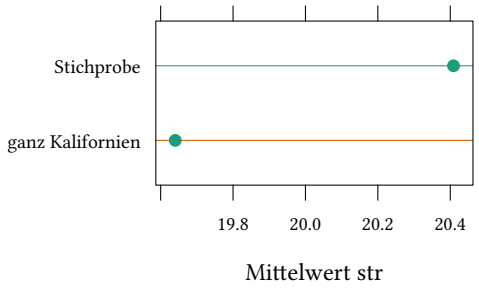
```
sampleSize<-60
Stichprobe<-head(Caschool,sampleSize)
```

Auch hier gibt es einen Mittelwert:

```
mean(Stichprobe$str)

[1] 20.40893
```

Zum Vergleich noch einmal der Mittelwert für alle 420 Bezirke:

```
mean(Caschool$str)

[1] 19.64043
```



Wenn wir keine Information über die übrigen 360 Bezirke haben, können wir uns nicht sicher sein, dass der Mittelwert für ganz Kalifornien 20.408932 beträgt.

Was können wir aus unseren Daten schließen?

Nehmen wir an, wir wissen, bevor wir unsere Stichprobe gezogen haben, eigentlich nichts über die Verteilung von `str`. Gegeben die ersten 60 Bezirke können wir nun folgendes sagen:

> Der Mittelwert in ganz Kalifornien liegt mit Wahrscheinlichkeit 95% im Intervall [19.95, 20.87].

(wie man zu diesen Zahlen kommt, betrachten wir weiter unten).

- Warum ist die Wahrscheinlichkeit interessant?

→ Betrachte den Erwartungswert folgender Lotterie:

  - Wenn der Mittelwert am Ende wirklich im Intervall [19.95, 20.87] liegt, ist der Gewinn 1…
  - …sonst gibt es einen Verlust von 19.

  → Diese Lotterie hätte immer noch einen Erwartungswert von gerade 0.

Obiges Beispiel ist ein bisschen abstrakt. Warum sollte uns jemand gerade eine solche Lotterie anbieten?

Im wirklichen Leben spielen wir gegen die Natur:

Z.B.: Wir schätzen nicht `str` in Kalifornien, sondern die zukünftige Nachfrage für ein Produkt.

- Wenn die Nachfrage im Intervall [...] liegt, rechnen wir mit einem Gewinn von .... Sonst machen wir einen Verlust von ...

- Ist der erwartete Gewinn immer noch positiv?

Aber auch der Bildungspolitiker mag eine Wette mit der »Natur« eingehen.

- Vielleicht verspricht er, dass die Parameter der Bildungspolitik schon in einem bestimmten Intervall liegen. Falls ja, macht er einen Gewinn, ansonsten einen Verlust.

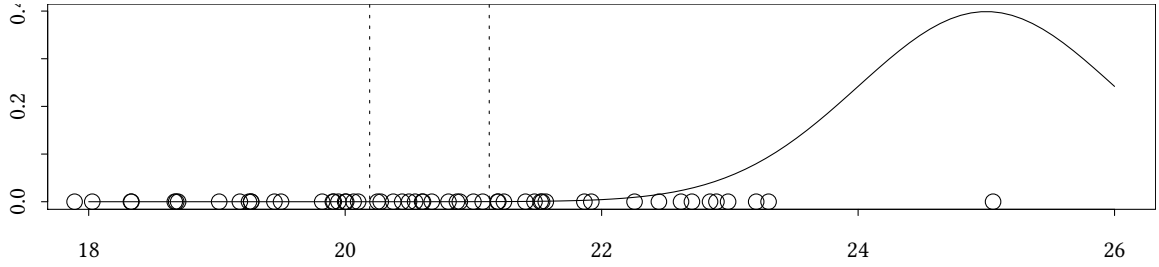
Die Wahrscheinlichkeit (hier 95%) hilft uns also, zusammen mit Auszahlungen, eine Erwartungsauszahlung abzuschätzen.

4.2. Satz von Bayes – Aggregation von Informationen

Nehmen wir an, wir wissen bereits einiges über die Verteilung von `str`.

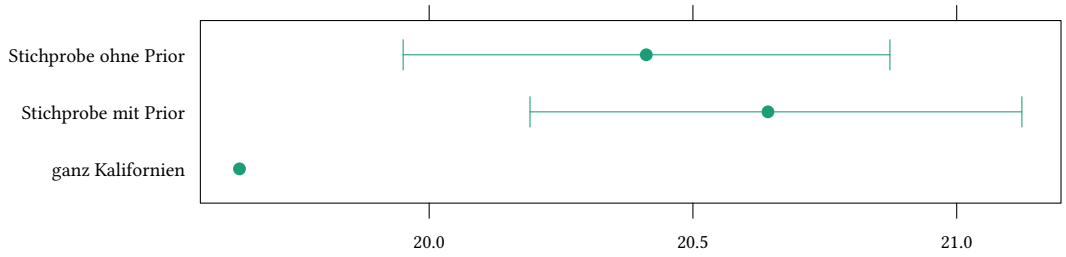
Vielleicht haben wir vorher eine andere Stichprobe gezogen. Auf Basis dieser Stichprobe sind wir zum Ergebnis gekommen, dass der Mittelwert von `str` den Wert 25 hat. Nehmen wir ferner an, dass sich die Genauigkeit unserer Erwartung mit einer Standardabweichung von z.B. 1 beschreiben lässt.

Was würden wir in diesem Fall aus der (neuen) Stichprobe schließen?



→ Der Mittelwert in ganz Kalifornien liegt mit Wahrscheinlichkeit 95% im Intervall [20.19, 21.12].

Auf Basis der ersten 60 Beobachtungen können wir folgendes sagen:



Gegeben unser jeweiliger Prior liegt mit Wahrscheinlichkeit 95% der Mittelwert im jeweiligen Intervall.

Andere Annahmen (Priors) $\longrightarrow$ andere Resultate.

Satz von Bayes Notation:

$P(A)$	Wahrscheinlichkeit, dass A zutrifft
$P(B)$	Wahrscheinlichkeit, dass B zutrifft
$P(A \wedge B)$	Wahrscheinlichkeit, dass A und B gleichzeitig zutreffen
$P(A\|B)$	Vorausgesetzt B trifft zu: Wahrscheinlichkeit, dass A zutrifft
$P(B\|A)$	Vorausgesetzt A trifft zu: Wahrscheinlichkeit, dass B zutrifft

Wie kann man $P(A \wedge B)$ noch ausrechnen?

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

Umstellen ergibt den Satz von Bayes:

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

Nenne $A = \theta$, $B = X$, dann gilt

Satz von Bayes (andere Bezeichnungen, sonst unverändert):

$$\underbrace{P(\theta|X)}_{\text{posterior}} = \underbrace{P(X|\theta)}_{\text{likelihood}} \frac{\overbrace{P(\theta)}^{\text{prior}}}{P(X)}$$

Wie bestimmen wir $P(X)$?

$$P(X) = \int \underbrace{P(\theta)}_{\text{prior}} \underbrace{P(X|\theta)}_{\text{likelihood}} \, d\theta$$

Einsetzen in den Satz von Bayes:

$$\underbrace{P(\theta|X)}_{\text{posterior}} = \underbrace{P(X|\theta)}_{\text{likelihood}} \frac{\overbrace{P(\theta)}^{\text{prior}}}{\int P(\theta)P(X|\theta) \, d\theta}$$

Das sieht nach einfacher Algebra aus – allein, $\int P(\theta)P(X|\theta)\,d\theta$ ist ohne Computer nur schwer zu berechnen.

Dogmengeschichte Statistische Inferenz zwischen 1925 und 1953: *Frequentistisch*
Kommt ohne $\int P(\theta)P(X|\theta)\,d\theta$ aus.
Braucht nur $P(X|\theta)$ (likelihood).

- Kein Computer notwendig.
- Ergebnis bei kleinen Stichproben ungenau.
- Ergebnis schwer zu interpretieren.
- Macht sich keine Gedanken über Prior.

Statistische Inferenz seit 1953: *Bayesianisch*

- Computer erforderlich (normalerweise).
- Ergebnis korrekt.
- Ergebnis leichter zu interpretieren.
- Erlaubt/erfordert Prior $P(\theta)$.

→ Bayesianische Inferenz wird seit 1953 zunehmend populär – es gibt aber immer noch Leute, die frequentistisch denken.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A., Teller, H. (1953). Equation of state calculations by fast computing machines. The Journal of Chemical Physics 21:1087–1092.

4.3. Bayesianische Inferenz mit R

Es gibt für R viele Pakete, die uns unterstützen. Hier verwenden wir das Paket `MCMCpack`.

```
library(MCMCpack)
Stichprobe<-head(Caschool,sampleSize)
summary(MCMCregress(str ~ 1,data=Stichprobe))


Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000
```

```
1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

              Mean     SD Naive SE Time-series SE
(Intercept) 20.411 0.2344 0.002344       0.002381
sigma2       3.193 0.6075 0.006075       0.006180

2. Quantiles for each variable:

              2.5%    25%    50%   75% 97.5%
(Intercept) 19.951 20.256 20.412 20.57 20.87
sigma2       2.212  2.752  3.123  3.55  4.57
```

```
summary(MCMCregress(str ~ 1,data=Stichprobe,b0=25,B0=1))


Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

              Mean     SD Naive SE Time-series SE
(Intercept) 20.646 0.2341 0.002341       0.002463
sigma2       3.251 0.6320 0.006320       0.006539

2. Quantiles for each variable:

              2.5%    25%    50%    75%  97.5%
(Intercept) 20.191 20.491 20.642 20.797 21.124
sigma2       2.242  2.796  3.181  3.613  4.704
```

Notation für Likelihood und Prior Für Bayesianische Inferenz brauchen wir Likelihood und Prior.

$$\underbrace{P(\theta|X)}_{\text{posterior}} = \underbrace{P(X|\theta)}_{\text{likelihood}} \frac{\overbrace{P(\theta)}^{\text{prior}}}{\int P(\theta)P(X|\theta)\,d\theta}$$

- Likelihood $P(\underbrace{X}_{\texttt{str}}|\underbrace{\theta}_{\mu,\sigma})$
  - In unserem Beispiel sind die Daten (X) die Variable `str`.
  - Die Parameter (θ) sind Mittelwert und Standardabweichung von `str`: μ und σ.

    $\texttt{str} \sim N(\mu, \sigma^2)$

  - `MCMCregress` verwendet folgende Notation

    ```
    MCMCregress(str ~ 1,data=Stichprobe)
    ```

- Prior: bislang verwenden wir einen »flachen Prior«.
  - Wir nehmen an, dass wir über μ und σ nichts wissen.
  - Alle Werte von μ sind gleich wahrscheinlich.
  - Notation: $\mu \sim N(0, \infty)$

    (Varianz ist unendlich groß, Mittelwert spielt keine Rolle, wird z.B. gleich 0 gesetzt.)
  - Wenn man mehr Vorwissen (prior knowledge) hat, würde man sagen, dass μ einer bestimmten Verteilung folgt.

    z.B. $\mu \sim N(25, 1)$

Prior für `mu`: Bislang hatten wir (vor der Stichprobenziehung) folgenden Prior für `mu`:

- $\mu \sim N(0, \infty)$

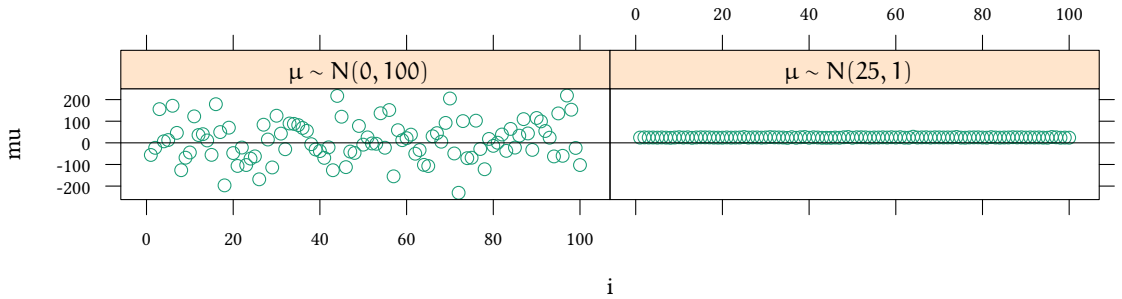
  Also:
  - μ ist normalverteilt um Mittelwert=0 mit Varianz=∞:

    alles ist möglich.

Nehmen wir an, wir hätten (vor der Stichprobenziehung) folgenden Prior:

- $\mu \sim N(25, 1)$

  Also:
  - μ ist normalverteilt um Mittelwert=25 mit Varianz=1.



`MCMCregress` erlaubt uns, einen Prior für μ vorzugeben.
Falls $\mu \sim N(\mu_0, \sigma_0^2)$ dann `b0`=μ_0, `B0`=$1/\sigma^2$.

```
MCMCregress(str ~ 1,data=Stichprobe,b0=...,B0=...)
```

- b0: Mittelwert der Parameter, entsprechend A-priori-Wissen
- B0: Präzision des Wissens

  $1/\sigma^2$ Präzision
  σ^2 Varianz
  σ Standardabweichung

  Es ist egal, ob wie die Streuung mit Präzision, Varianz oder Standardabweichung beschreiben.

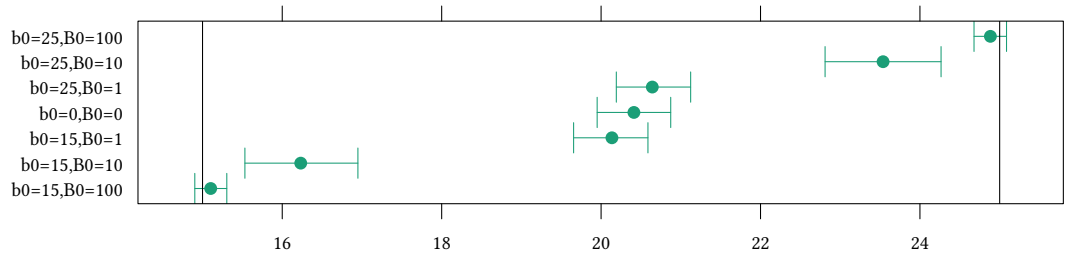
  Für MCMCregress verwenden wir den Begriff der Präzision (Parameter B0).

```
posterior25.1<-MCMCregress(str ~ 1,data=Stichprobe,b0=25,B0=1)
summary(posterior25.1)


Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

              Mean     SD Naive SE Time-series SE
(Intercept) 20.646 0.2341 0.002341       0.002463
sigma2       3.251 0.6320 0.006320       0.006539

2. Quantiles for each variable:

              2.5%    25%    50%    75%  97.5%
(Intercept) 20.191 20.491 20.642 20.797 21.124
sigma2       2.242  2.796  3.181  3.613  4.704
```

Prior → Posterior

Gegeben ein Prior $\mu \sim N(25, 1)$
↓
dann liegt mit 95% Wahrscheinlichkeit
der Posterior für μ zwischen 20.19 und 21.12.

```
posterior25.100<-MCMCregress(str ~ 1,data=Stichprobe,b0=25,B0=100)
posterior25.10 <-MCMCregress(str ~ 1,data=Stichprobe,b0=25,B0=10)
posterior15.1  <-MCMCregress(str ~ 1,data=Stichprobe,b0=15,B0=1)
posterior15.10 <-MCMCregress(str ~ 1,data=Stichprobe,b0=15,B0=10)
posterior15.100<-MCMCregress(str ~ 1,data=Stichprobe,b0=15,B0=100)
```



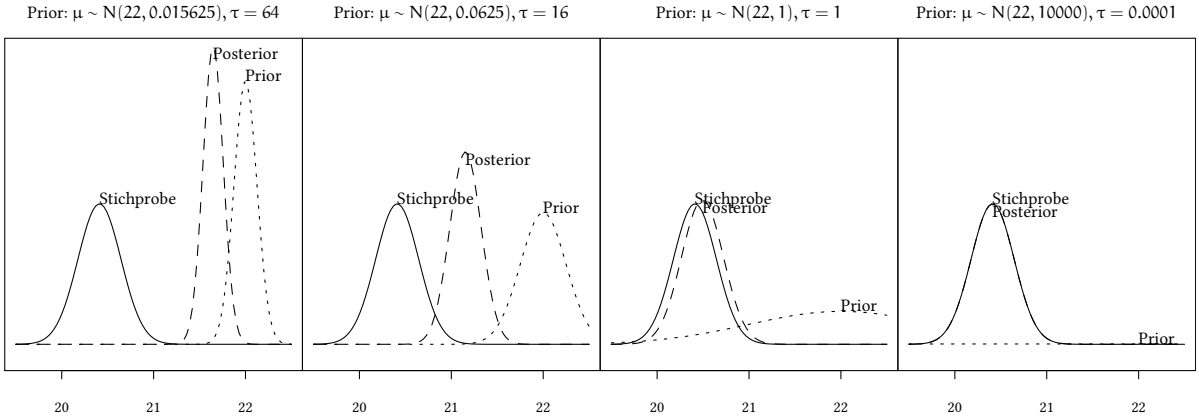
- Uninformierter Prior ($\mu \sim N(0, \infty)$ Präzision des Priors ist Null. Varianz unendlich groß. Kein Vorwissen. Posterior wird nur von Daten bestimmt.
- Vager Prior ($\mu \sim N(0, 100)$ Präzision des Priors ist sehr klein. Varianz sehr groß. Wenig Vorwissen. Posterior wird vor allem von Daten bestimmt.
- Informierter Prior ($\mu \sim N(25, 1)$, Präzision des Priors ist größer. Mehr Vorwissen. Posterior wird stärker vom Prior bestimmt.

Prior funktioniert wie jede andere Modellannahme.

- Wir müssen unsere Zuhörer und Zuhörerinnen von unseren Annahmen überzeugen können. Dazu gehört auch der Prior.
- Oft ist es leichter, einen Zuhörer oder eine Zuhörerin von einem vagen Prior zu überzeugen als von einem informierten Prior (mit dem vagen Prior machen wir keine starke Annahme über die Verteilung der Parameter).

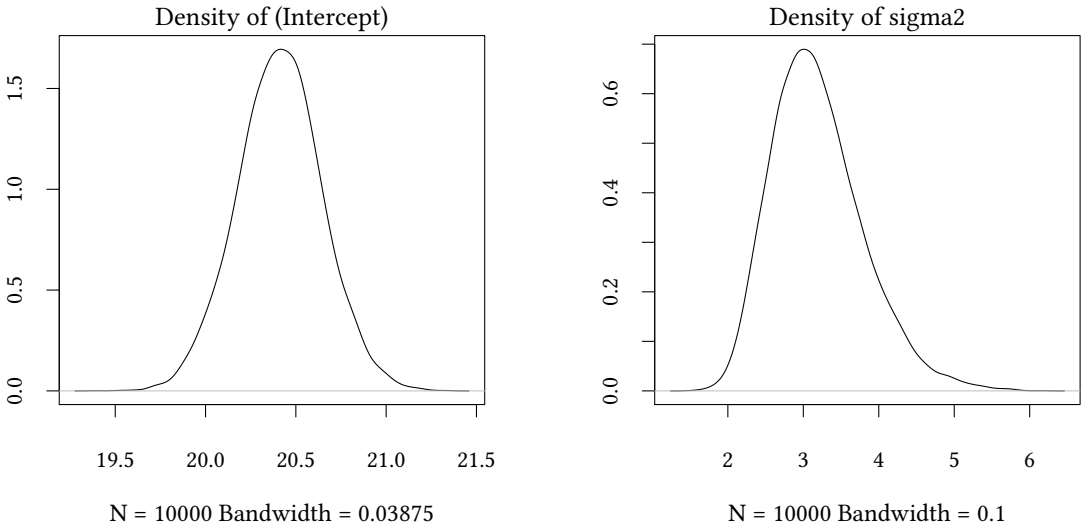
Präzision des Priors → Posterior ↑: Posterior für Stichprobe mit flat Prior: $\mu = 20.4$, $\sigma_\mu = 0.234$.

Vergleiche verschiedene Priors: $\mu = 22$, Präzision $\tau = \frac{1}{\sigma^2}$ ist unterschiedlich.



Grafische Darstellung für Posterior

```
densplot(posterior,show.obs=FALSE)
```



4.4. Inferenz über das “credible interval” hinaus

`MCMCregress` liefert uns das 95%-credible interval, also ein Intervall, in dem die gesuchten Parameter mit Wahrscheinlichkeit 95% sind, frei Haus.

Manchmal interessieren wir uns für andere Fragen, z.B., wie wahrscheinlich es ist, dass ein Parameter in einem bestimmten Bereich ist.

Auch hier hilft uns das Ergebnis von `MCMCregress`.

Unsere `MCMCregress` Schätzung des Posteriors ist eine lange Liste von Samples aus der Verteilung des Posteriors. Hier sind die ersten 6 Samples.

```
head(posterior,5)

Markov Chain Monte Carlo (MCMC) output:
Start = 1001
End = 1006
Thinning interval = 1
     (Intercept)   sigma2
[1,]    20.91196 2.417688
[2,]    20.65073 3.666025
[3,]    20.59721 3.032354
[4,]    20.49083 2.771416
[5,]    19.95767 3.282833
[6,]    20.22015 2.796865
```

Wir können diese Samples verwenden, um andere Fragen zu beantworten. Z.B.: Wie wahrscheinlich ist es, dass der Mittelwert von `str` größer als 20 ist?

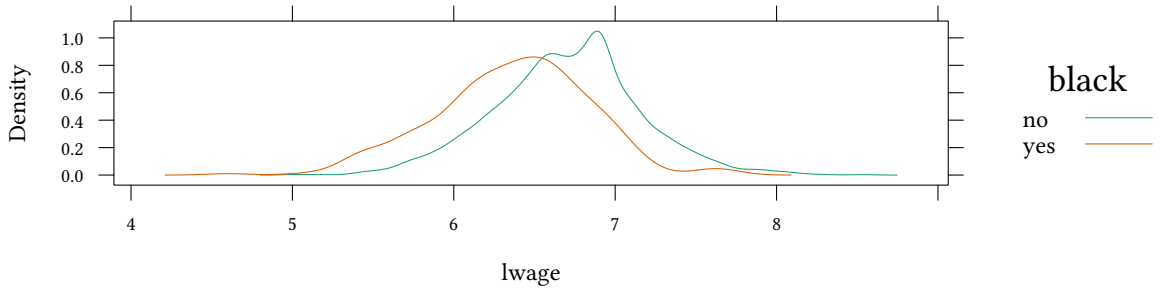
```
mean(posterior[,"(Intercept)"]>20)

[1] 0.9589
```

4.5. Vergleich von Gruppen

Sehr oft wollen wir Eigenschaften von Gruppen vergleichen. Der Datensatz `Wages` enthält 4165 Beobachtungen für Löhne.

Hier vergleichen wir die Gruppen `black` und `white`.

```
data(Wages)
densityplot(~ lwage,group=black, data=Wages, plot.points=FALSE,
            auto.key=list(space="right",title="black"))
```



Modell:

$$\log \texttt{wage} \sim N(\beta_0 + \beta_1 \texttt{black}, \sigma^2)$$

alternative Darstellung:

$$\log \texttt{wage} = \beta_0 + \beta_1 \texttt{black} + \epsilon \qquad \text{wobei } \epsilon \sim N(0, \sigma^2)$$

$$E(\log \texttt{wage}_{\texttt{white}}) = \beta_0 \qquad E(\log \texttt{wage}_{\texttt{black}}) = \beta_0 + \beta_1$$

$$E(\texttt{wage}_{\texttt{white}}) \approx e^{\beta_0} \qquad E(\texttt{wage}_{\texttt{black}}) \approx e^{\beta_0+\beta_1}$$

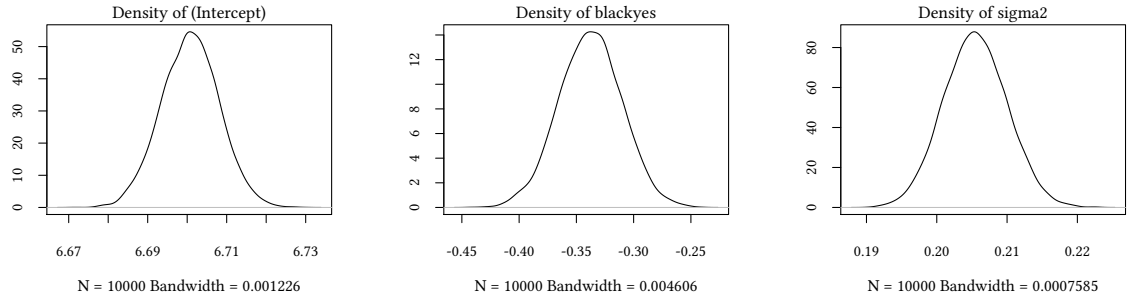
$$\frac{E(\texttt{wage}_{\texttt{black}})}{E(\texttt{wage}_{\texttt{white}})} \approx \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

Die Verteilung in der Stichprobe sieht unterschiedlich aus. Was können wir aus der Stichprobe für die Population schließen?

```
wage.posterior<-MCMCregress(lwage ~ black,data=Wages)
```

Hier ist die Verteilung des Posteriors:

```
densplot(wage.posterior,show.obs=FALSE)
```




```
summary(wage.posterior)


Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

               Mean       SD   Naive SE Time-series SE
(Intercept)  6.7008 0.007297 0.00007297      0.00007297
blackyes    -0.3378 0.027416 0.00027416      0.00027416
sigma2       0.2055 0.004515 0.00004515      0.00004515

2. Quantiles for each variable:

               2.5%     25%     50%     75%   97.5%
(Intercept)  6.6863  6.6959  6.7008  6.7057  6.7151
blackyes    -0.3923 -0.3563 -0.3378 -0.3193 -0.2844
sigma2       0.1968  0.2024  0.2054  0.2085  0.2144
```

Auf Basis der `summary` können wir also z.B. die folgenden Aussagen machen:

- Im Mittel schätzen wir den Unterschied in `lwage` zwischen `black` und `white` auf -0.3378.

  (Der mittlere Lohn von `black` ist nur $e^{-0.3378} = 71.34\%$ des mittleren Lohns von `white`).

- Mit Wahrscheinlichkeit 95% liegt der Unterschied in `lwage` zwischen `black` und `white` zwischen -0.3923 und -0.2844.

  (Mit Wahrscheinlichkeit 95% liegt der mittlere Lohn von `black` zwischen 67.55% und 75.25% des mittleren Lohns von `white`).

- Mit Wahrscheinlichkeit 50% ist der Unterschied in `lwage` zwischen `black` und `white` größer als -0.3378

  (Mit Wahrscheinlichkeit 50% ist der mittlere Lohn von `black` 71.34% des mittleren Lohns von `white` oder kleiner).

So sehen die Samples unseres Posteriors aus:

```
head(wage.posterior)

Markov Chain Monte Carlo (MCMC) output:
Start = 1001
End = 1007
Thinning interval = 1
     (Intercept)   blackyes    sigma2
[1,]    6.703357 -0.3191460 0.2059752
[2,]    6.694306 -0.2948398 0.2126631
[3,]    6.705780 -0.3666009 0.2078127
[4,]    6.706915 -0.4163234 0.2053672
[5,]    6.699219 -0.3796295 0.2054991
[6,]    6.708262 -0.3131989 0.2061392
[7,]    6.710522 -0.3595830 0.2054283
```

Wie oben können wir unseren Posterior auch mit vorgegebenen Werten vergleichen – hier z.B. fragen wir:

- Wie wahrscheinlich ist es, dass der durchschnittliche Lohn in der Gruppe “black” z.B. nur 65% oder weniger der durchschnittlichen Lohns der Gruppe “white” ist?

  ```
  mean(wage.posterior[,"blackyes"]<log(.65))

  [1] 0.0005
  ```

- Wie wahrscheinlich ist es, dass der durchschnittliche Lohn in der Gruppe “black” z.B. nur 80% oder weniger der durchschnittlichen Lohns der Gruppe “white” ist?

  ```
  mean(wage.posterior[,"blackyes"]<log(.80))

  [1] 1
  ```

Anhang 4.A Beispiele für die Vorlesung

Wir betrachten weiter den Datensatz `Wages`.

- Erhalten Frauen und Männer einen unterschiedlichen Lohn?
- Wie groß ist der Unterschied?

- Wie sicher können wir sein, dass es einen Unterschied nicht nur in der Stichprobe, sondern auch in der Population gibt?

- Wie sicher können wir uns sein, dass in der Population Männer im Durchschnitt mehr als 50% mehr Lohn bekommen?

Anhang 4.B Übungen

Übung 4.1 *Betrachten Sie die Zufallsvariable* X.
Gestern haben Sie ein Stichprobe von X *erhalten.*

- *Bestimmen Sie Mittelwert und Standardabweichung.*

Heute ziehen Sie eine neue Stichprobe von X.

- *Verwenden Sie Mittelwert und Standardabweichung von gestern als Prior. Was ist Ihr Posterior für den Mittelwert von* X*?*

- *Was ist Ihr Posterior, wenn Sie die Information über die gestrige Stichprobe nicht hätten?*

Erzeugen Sie die Stichproben wie folgt:

```
N<-100;set.seed(123)
gestern<-rnorm(N)
heute<-rnorm(N,mean=1)
```

Übung 4.2 *Betrachten Sie wieder die Variable* `str` *aus* `Caschool`.

- *Wie bestimmen Sie den Mittelwert der ersten 18 Beobachtungen?*

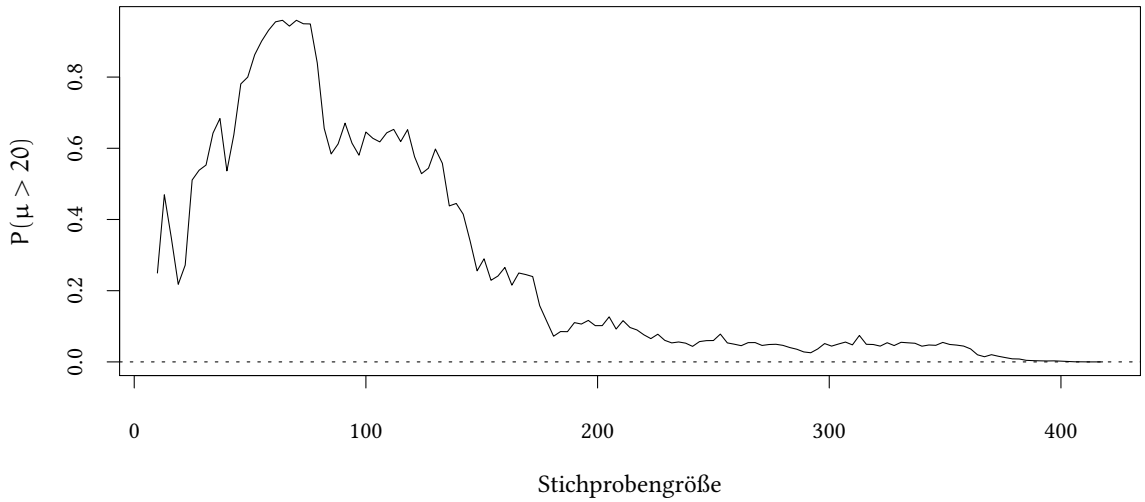
- *Gegeben diese 18 Beobachtungen, wie wahrscheinlich ist es, dass der Mittelwert der Population größer als 20 ist?*

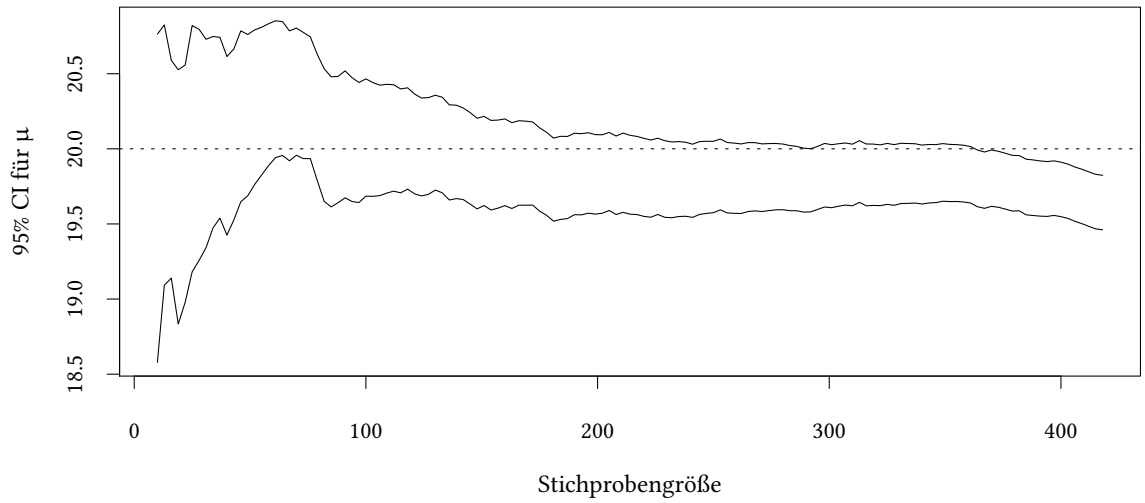
- *Was, wenn Sie die ersten 40 Beobachtungen zu Grunde legen?*

- *Was, wenn Sie die ersten 80 Beobachtungen zu Grunde legen?*

- *Was, wenn Sie die ersten 200 Beobachtungen zu Grunde legen?*

- *Führen Sie die Berechnung für den gesamten Datensatz aus? Wie interpretieren Sie das Ergebnis?*





Übung 4.3 *Betrachten Sie folgenden Output:*

```
p<-MCMCregress(avginc ~ 1,data=head(Caschool,20))
summary(p)


Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:
```

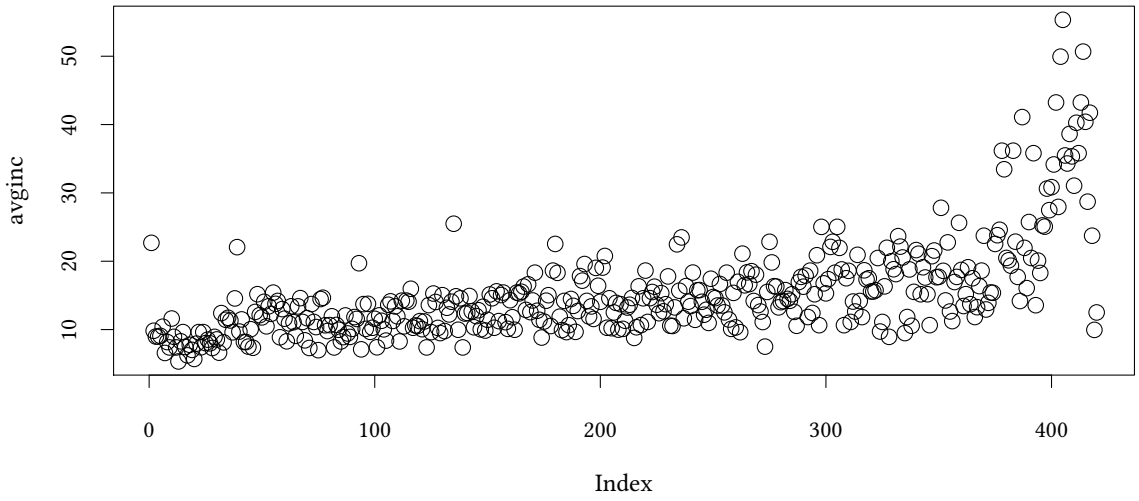
```
              Mean     SD Naive SE Time-series SE
(Intercept)  8.875 0.8635 0.008635       0.008635
sigma2      14.701 5.4314 0.054314       0.057185

2. Quantiles for each variable:

             2.5%    25%    50%   75% 97.5%
(Intercept) 7.146  8.319  8.888  9.43 10.58
sigma2      7.637 10.923 13.625 17.12 28.14
```

- *Wie wahrscheinlich ist es, dass der Mittelwert von `avginc>8.888`?*
- *Wie wahrscheinlich ist es, dass der Mittelwert von `avginc>7.146`?*
- *Wie wahrscheinlich ist es, dass der Mittelwert von `avginc<8.319`?*

Übung 4.4 *Bestimmen Sie nun ein credible interval für den Mittelwert von `avginc` für eine Stichprobengröße von*

- *10*
- *50*
- *100*
- *200*
- *300*
- *400*

```
with(Caschool,plot(avginc))
```



Beobachtung am Rande: Der Datensatz ist so sortiert, dass wir eher mit armen Bezirken anfangen. Am Ende kommen erst die reichen Bezirke.

Wir verwenden besser immer eine zufällige Stichprobe der Bezirke:

Übung 4.5 *Betrachten Sie die Information über* `avginc` *aus 50 zufällig gezogenen Bezirken von* `Caschool`*.*

- *Wie groß ist das 95%-credible Interval für den Mittelwert von* `avginc`*?*
- *Wie wahrscheinlich ist es, dass der Mittelwert kleiner als 13 ist?*

Bevor Sie Ihre Daten erhoben haben, sind sie davon ausgegangen, dass der Mittelwert von `avginc` $\mu_{\texttt{avginc}} \sim N(17, \frac{1}{9})$. *Ihr Prior für* $\mu_{\texttt{avginc}}$ *hat also einen Mittelwert von 17 bei einer Varianz von 1/9.*

- *Was ist nun, auf Basis von Stichprobe und A-priori Wissen, das 95%-credible Interval für den Mittelwert von* `avginc`*?*
- *Wie wahrscheinlich ist es jetzt, dass der Mittelwert kleiner als 13 ist?*

Übung 4.6 *Betrachten Sie den Datensatz* `BudgedFood` *aus dem Paket* `Ecdat`.

- *Vergleichen Sie die Ausgaben für Lebensmittel für Haushalte mit männlichem und weiblichem Haushaltsvorstand.*
- *Geben Sie ein 95% credible interval für den Unterschied zwischen männlichem und weiblichem Haushaltsvorstand an.*
- *Jemand behauptet, der Unterschied sei kleiner als 1% der Gesamtausgaben. Wie wahrscheinlich ist diese Aussage?*
- *Wie wahrscheinlich ist es, dass Haushalte mit einem weiblichen Haushaltsvorstand um mehr als 2% der Gesamtausgaben für Lebensmittel ausgeben als Haushalte mit einem männlichen Haushaltsvorstand?*

Übung 4.7

```
Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

             Mean     SD Naive SE Time-series SE
(Intercept)  11.84  1.856  0.01856         0.0191
sigma2      339.84 49.340  0.49340         0.5054

2. Quantiles for each variable:

              2.5%   25%    50%    75% 97.5%
(Intercept)   8.192  10.6  11.84  13.07  15.5
sigma2      256.386 304.6 335.74 370.00 447.6
```

- *Wie wahrscheinlich ist es, dass* $x > 10.6$*?*
- *Wie wahrscheinlich ist es, dass* $\mu_x > 10.6$*?*
- *Wie wahrscheinlich ist es, dass* $\sigma_x^2 < 370$*?*

Übung 4.8 *Im obigen Beispiel haben wir a-priori Wissen über μ_x. Wir erwarten, dass μ_x normalverteilt ist, mit Mittelwert 10 und Standardabweichung 2. Wie rufen wir `MCMCregress` auf?*

```
summary(MCMCregress(x ~ 1,b0=   ,B0=      ))
```

Übung 4.9

```
Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

              Mean     SD Naive SE Time-series SE
(Intercept)  28.48  2.338  0.02338        0.02338
x           -10.05  0.109  0.00109        0.00109
sigma2      385.26 56.605  0.56605        0.58212

2. Quantiles for each variable:

              2.5%    25%    50%    75%   97.5%
(Intercept)  23.93  26.92  28.51  30.04  33.048
x           -10.27 -10.13 -10.05  -9.98  -9.838
sigma2      290.19 345.48 379.66 419.66 509.134
```

Wir schätzen $Y = \beta_0 + \beta_1 X + u$

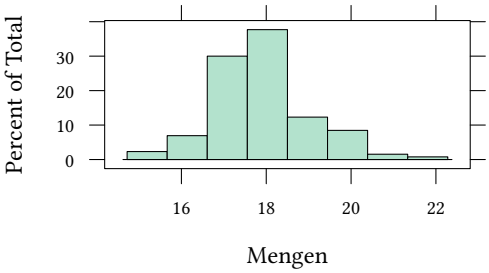
- *Wie wahrscheinlich ist es, dass* $x > -10.13$*?*
- *Wie wahrscheinlich ist es, dass* $\beta_1 > -10.13$*?*
- *Wie wahrscheinlich ist es, dass* $\sigma_x^2 < 379.66$*?*
- *Wie wahrscheinlich ist es, dass* $\sigma_u^2 < 379.66$*?*

5. Frequentistische Inferenz – Tests von Nullhypothesen

5.1. Motivation: Wählen Firmen im Oligopol Gleichgewichtsmengen?

Kann es sein, dass durch stillschweigende Kooperation (*tacit collusion*) Firmen im Oligopol kleinere Mengen als Wettbewerbsmengen anbieten. Dadurch würde der Preis steigen und die Firmen würden einen höheren Gewinn machen. Allerdings würde die Konsumentenrente sinken.

Sauermann und Selten haben in den fünfziger Jahren damit begonnen, Experimente zum Wettbewerb in Oligopolen durchzuführen.



Ist es bemerkenswert, wenn die mittlere Menge im Experiment mit 17.8615385 größer ist als die Cournotmenge von 12?

Sauermann H., Selten R.. Ein Oligopolexperiment. Zeitschrift für die gesamte Staatswissenschaft, 1959.

In Kapitel 4 haben wir "credible intervals" betrachtet, d.h. haben berechnet, in welchem Intervall um unseren geschätzen Parameter $\hat{\theta}$ der wahre Parameter θ wohl liegen könnte.

Manchmal haben wir konkrete Hypothesen, welchen Wert θ annehmen könnte. Diese Hypothesen kann man testen.

Der Mittelwert der *Stichprobe* ist nicht 12, sondern 17.8615385.

Hypothese: Der Mittelwert der *Population* ist 12.

Hier ist das Ergebnis eines t-Tests:

```
t.test(x, mu=12)


 One Sample t-test

data:  x
t = 54.567, df = 129, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 12
95 percent confidence interval:
 17.64901 18.07407
sample estimates:
mean of x
 17.86154
```

5.2. Frequentistische Hypothesentests

Das Schlußverfahren in Kapitel 4, Bayesianische Inferenz, braucht für viele Situationen einen Computer. Mit Papier und Bleistift kommen wir nicht sehr weit. Was haben Leute vor 1953 gemacht?

Ronald Fisher (1925): Idee des Hypothesentests

Frequentistisch

- Wissenschaftler haben Nullhypothesen über fixe Parameter:

  $\theta = \theta_0$

- Wie wahrscheinlich ist es, die Daten der Stichprobe zu finden, wenn die Nullhypothese wahr ist?

- $\rightarrow P(X|\theta_0)$

Diese Frage können wir mit Papier und Bleistift beantworten.

Bayesianisch

- Parameter sind Zufallsvariablen

- Wissenschaftler haben einen mehr oder weniger genauen Prior über die Verteilung des Parameters.

- Daten erlauben uns, die Verteilung des Posteriors zu bestimmen.

- $\rightarrow P(\theta|X) = P(X|\theta)\frac{P(\theta)}{P(X)}$

Diese Frage können wir oft nur mit einem Computer beantworten.

Dogmengeschichte

Hypothesentest: Ronald Fisher (1925).

Interpretation der Ergebnisse dieser Tests ist oft verwirrend, unintuitiv und hat konzeptuelle Schwächen.

Bayesianische Schlussverfahren: Thomas Bayes (1702-1761). Einfachere Alternative.

Seit 1953 gibt es auch effiziente Computeralgorithmen, die es erlauben, eine große Klasse von Problemen zu lösen (Metropolis et al., 1953).

Viele Anwender sind mit Bayesianischen Verfahren noch nicht vertraut. Wir betrachten deshalb in dieser Vorlesung auch den Hypothesentest von Ronald Fisher (1925) und das Konfidenzinterval von Jerzy Neyman (1930).

Einseitige und zweiseitige Tests Beispiel: Test des Mittelwerts der Grundgesamtheit μ:

- $H_0 : \mu = \mu_{X,0}$ versus $H_1 : \mu \neq \mu_{X,0}$ (zweiseitiger Test)
- $H_0 : \mu = \mu_{X,0}$ versus $H_1 : \mu > \mu_{X,0}$ (einseitiger Test)
- $H_0 : \mu = \mu_{X,0}$ versus $H_1 : \mu < \mu_{X,0}$ (einseitiger Test)

5.3. Fehler 1. und 2. Art

Der frequentistische Hypothesentest befasst sich nicht mit der Wahrscheinlichkeit von Parametern. Zur Zeit von Ronald Fisher (1925) war die Berechnung dieser Wahrscheinlichkeit zu kompliziert.

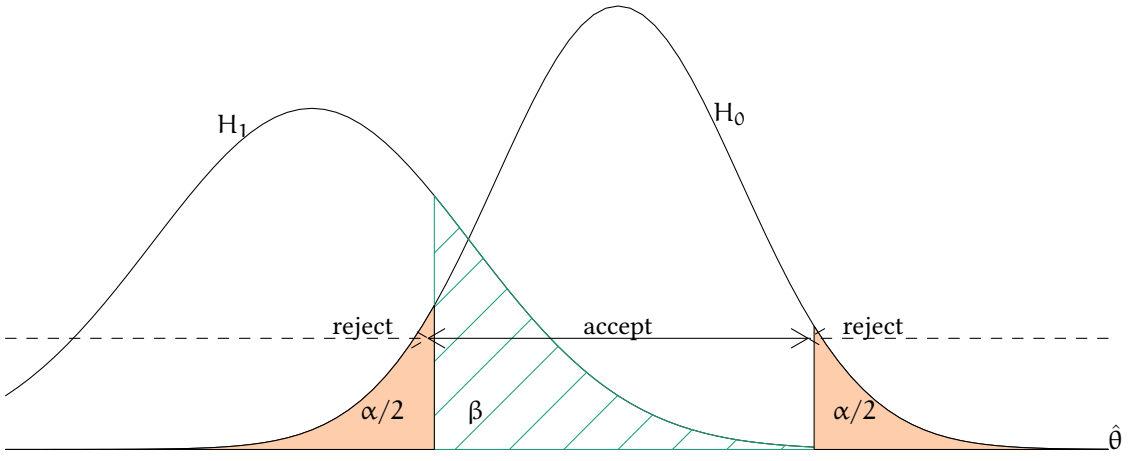
Leichter zu berechnen ist die Wahrscheinlichkeit des »Fehlers 1. Art«. Also stützt sich der größte Teil der frequentistischen Inferenz auf diesen Fehler (bzw. auf Aussagen über Situationen, in denen dieser Fehler klein ist).

Voraussetzung: Wir haben eine Nullhypothese H_0. Die kann abgelehnt oder nicht abgelehnt werden.

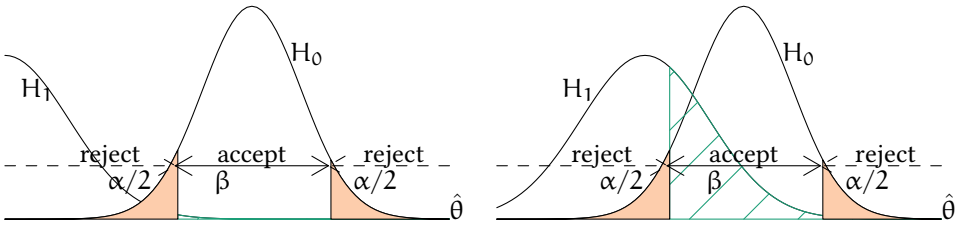
- Fehler 1. Art, α Fehler, falsch positiv: H_0 wird *abgelehnt*, obwohl H_0 *wahr* ist
  - α: false positive rate, Signifikanzniveau, $1-\alpha =$ Spezifität
- Fehler 2. Art, β Fehler, falsch negativ: H_0 wird *nicht abgelehnt*, obwohl H_0 *falsch* ist
  - β: false negative rate, $1-\beta =$ Power des Tests = Sensitivität

Testergebnis	tatsächliche Situation: H_0 falsch	tatsächliche Situation: H_0 wahr
H_0 wird abgelehnt (positiv)	$1-\beta$ Sensitivität Power	α Signifikanzniveau Fehler 1. Art
H_0 wird nicht abgelehnt (negativ)	β Fehler 2. Art	$1-\alpha$ Spezifität

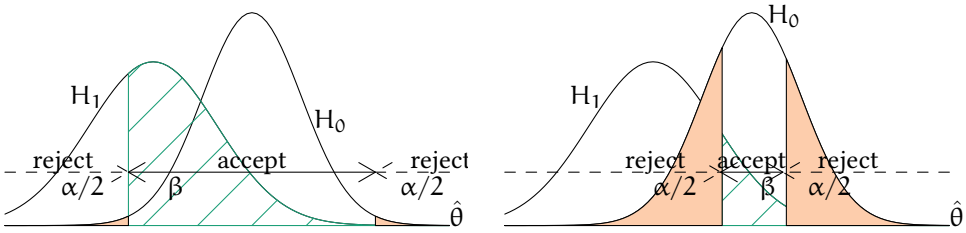
Betrachten wir zunächst nur eine Welt, in der eine Stichprobe nur aus einer von zwei Verteilungen kommen kann. Die eine Verteilung entspricht H_0, die andere H_1. (Oft ist die Welt sehr viel komplizierter):



In der folgenden Grafik sehen Sie links ein einfaches Problem. H_1 ist weit von H_0 weg. Rechts dagegen ist H_1 nahe an H_0. Die Unterscheidung ist schwieriger.



In der nächsten Graphik ist der Abstand zwischen den Verteilungen von H_1 und $H = 0$ gleich. Links ist der Bereich, in dem H_0 nicht abgelehnt wird, groß, rechts ist er klein.



Wir sehen, wir müssen uns entscheiden: Entweder wir nehmen einen großen Fehler erster Art in Kauf, oder einen großen Fehler zweiter Art.

5.4. Signifikanzniveau eines Tests

Signifikanzniveau eines Tests = Vorspezifizierte Wahrscheinlichkeit α die Nullhypothese *fälschlich abzulehnen*, obwohl sie wahr ist.

Beispiel: Der Mittelwert $\bar{x}$ einer Stichprobe sei normalverteilt.

- $\bar{x} \sim N\left(\mu_0, \frac{\sigma_X^2}{n}\right)$
- $\bar{x} - \mu_0 \sim N\left(0, \frac{\sigma_X^2}{n}\right)$
- $\frac{\bar{x}-\mu_0}{\sigma_X/\sqrt{n}} \sim N(0,1)$

Seien $Q(\frac{\alpha}{2})$ und $Q(1-\frac{\alpha}{2})$ die Quantile der Normalverteilung, dann wird H_0 nicht abgelehnt, wenn

$$Q\left(\frac{\alpha}{2}\right) \leq \frac{\bar{x}-\mu_0}{\sigma_X/\sqrt{n}} \leq Q\left(1-\frac{\alpha}{2}\right)$$

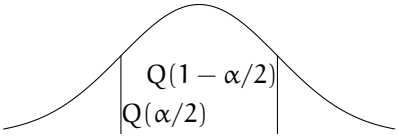
Zweiseitiger Test ($H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$) :

Das heißt, falls eine Stichprobe aus der Grundgesamtheit stammt (H_0 ist wahr), wird H_0 trotzdem mit Wahrscheinlichkeit α abgelehnt (false positive).

H_0 wird um so eher abgelehnt

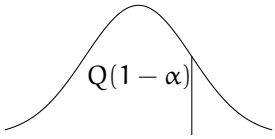
- je größer $|\bar{x} - \mu_0|$ ist
- je kleiner σ_X ist

- je größer n ist
- je größer α ist



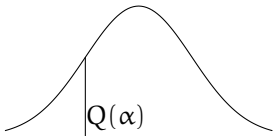
Einseitiger Test ($H_0 : \mu = \mu_0$, $H_1 : \mu > \mu_0$): Sei $Q(1-\alpha)$ das $1-\alpha$-Quantil der Normalverteilung, dann wird H_0 nicht abgelehnt, wenn

$$\frac{\bar{x} - \mu_0}{\sigma_X/\sqrt{n}} \leq Q(1-\alpha)$$



Einseitiger Test ($H_0 : \mu = \mu_0$, $H_1 : \mu < \mu_0$): Sei $Q(\alpha)$ das α-Quantil der Normalverteilung, dann wird H_0 nicht abgelehnt, wenn

$$Q(\alpha) \leq \frac{\bar{x} - \mu_0}{\sigma_X/\sqrt{n}}$$



5.5. Parametertests — formale Definition

- Sei $X \sim f(x|\theta)$ eine Zufallsvariable mit unbekanntem $\theta \in \Theta$
- $H_0 : \theta \in \Theta_0 \subset \Theta$ (Nullhypothese)
- $H_1 : \theta \notin \Theta_0$ (Alternativhypothese)
- $(X_1, \ldots, X_n)$ Stichprobe von X
- $g(X_1, \ldots, X_n)$ Stichprobenfunktion (z.B. t-Statistik)
- $B \subset \mathbb{R}$ Ablehnungsbereich
  - $g((X_1, \ldots, X_n) \in B \quad \Rightarrow \quad H_0$ wird abgelehnt
  - $g((X_1, \ldots, X_n) \notin B \quad \Rightarrow \quad H_0$ wird *nicht* abgelehnt
- Fehler 1. Art: $P(g(X_1, \ldots, X_n) \in B | \theta \in \Theta_0)$
- Fehler 2. Art: $P(g(X_1, \ldots, X_n) \notin B | \theta \notin \Theta_0)$

5.6. p-Wert eines Tests

p-Wert= Wahrscheinlichkeit einer Stichprobe $X_1, \ldots, X_n$ zu ziehen, die wenigstens so advers zu unserer Nullhypothese ist, wie unsere Daten — gegeben, dass unsere Nullhypothese wahr ist.

Sei $\bar{X}$ der Mittelwert irgendeiner Stichprobe. $\bar{x}$ ist der Wert für $\bar{X}$ für unsere Stichprobe.

Zweiseitiger Test ($H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$)

$$\begin{aligned}
\text{p-Wert} &= \Pr_{H_0}\left(\left|\bar{X} - \mu_{X,0}\right| > \left|\bar{x}^{\text{Stp.}} - \mu_{X,0}\right|\right) \\
&= \Pr_{H_0}\left(\left|\frac{\bar{X} - \mu_{X,0}}{\sigma_X/\sqrt{n}}\right| > \left|\frac{\bar{x}^{\text{Stp.}} - \mu_{X,0}}{\sigma_X/\sqrt{n}}\right|\right)
\end{aligned}$$

Um den p-Wert zu berechnen, muss man die Stichprobenverteilung von $\bar{X}$ kennen. Das ist kompliziert, wenn n klein ist und X nicht normalverteilt ist.

Wenn n groß ist, kann man die Stichprobenverteilung von $\bar{X}$ durch die Normalverteilung approximieren

$$\text{p-Wert} \approx 2 \cdot F_N\left(-\left|\frac{\bar{x}^{\text{Stp.}} - \mu_{X,0}}{\sigma_X/\sqrt{n}}\right|\right)$$

Einseitiger Test ($H_0 : \mu = \mu_0, H_1 : \mu > \mu_0$)

$$\begin{aligned}
\text{p-Wert} &= \Pr_{H_0}\left(\bar{X} - \mu_{X,0} > \bar{x}^{\text{Stp.}} - \mu_{X,0}\right) \\
&= \Pr_{H_0}\left(\frac{\bar{X} - \mu_{X,0}}{\sigma_X/\sqrt{n}} > \frac{\bar{x}^{\text{Stp.}} - \mu_{X,0}}{\sigma_X/\sqrt{n}}\right) \\
&\approx F_N\left(-\frac{\bar{x}^{\text{Stp.}} - \mu_{X,0}}{\sigma_X/\sqrt{n}}\right)
\end{aligned}$$

Einseitiger Test ($H_0 : \mu = \mu_0, H_1 : \mu < \mu_0$)

$$\begin{aligned}
\text{p-Wert} &= \Pr_{H_0}\left(\bar{X} - \mu_{X,0} < \bar{x}^{\text{Stp.}} - \mu_{X,0}\right) \\
&= \Pr_{H_0}\left(\frac{\bar{X} - \mu_{X,0}}{\sigma_X/\sqrt{n}} < \frac{\bar{x}^{\text{Stp.}} - \mu_{X,0}}{\sigma_X/\sqrt{n}}\right) \\
&\approx F_N\left(\frac{\bar{x}^{\text{Stp.}} - \mu_{X,0}}{\sigma_X/\sqrt{n}}\right)
\end{aligned}$$

H_0 wird jeweils abgelehnt wenn p-Wert $< \alpha$

5.7. Formulierung von Hypothesen

H_0 ist die Hypothese, die wir ablehnen können. Wenn wir Evidenz *für* einen Zusammenhang sammeln wollen, formulieren wir das Gegenteil als H_0.

Beispiel:

- Wir möchten zeigen, dass die Europäische Währungsunion die Arbeitslosigkeit in der EU verringert. H_0 wäre also: Die Arbeitslosigkeit ist gleich geblieben oder gestiegen. (Da es nie passieren wird, dass die Arbeitslosigkeit *exakt* gleich bleibt, formuliert man H_0 zuweilen kürzer als »die Arbeitslosigkeit ist gestiegen«.)

- Wir möchten zeigen, dass die Europäische Zentralbank das Inflationsziel von 2% verfehlt. H_0 wäre also, die EZB hat das Inflationsziel erreicht.

- Wir möchten zeigen, dass die Steuerpolitik der Bundesregierung das Investitionsniveau fördert. H_0 wäre also, das Investitionsniveau ist kleiner geworden oder gleich geblieben. (Da es nie passieren wird, dass das Investitionsniveau *exakt* gleich bleibt, formuliert man H_0 zuweilen kürzer als »das Investitionsniveau ist kleiner geworden«.)

Wir werden normalerweise *nie* zeigen können, dass ein Parameter einen bestimmten Wert aus einer unendlich großen Menge von Werten hat. Wenn wir z.B. zeigen wollen, dass die EZB erfolgreich das Inflationsziel von 2% erreicht hat, hilft es nicht, als H_0 zu formulieren, die langfristige Inflation sei ungleich 2%. Ziemlich sicher wird jeder Wert, den wir messen, nicht exakt 2% sein, sondern stets ein bisschen größer oder kleiner. Eine solche Nullhypothese können wir also nie ablehnen.

5.8. Grenzen

- p-Werte können zeigen, wie sehr Daten mit einem statistischen Modell inkompatibel sind.

  Sie sagen aber *nicht*, welche Annahmen des Modells vielleicht verletzt sind.

- p-Werte messen nicht die Wahrscheinlichkeit, dass eine Hypothese wahr oder falsch ist.

- p-Werte messen nicht die Größe oder Wichtigkeit eines Effekts.

- Entscheidungen sollten nicht nur davon abhängen, ob ein p-Wert größer oder kleiner als ein bestimmter Wert ist.

Siehe auch: Wasserstein. ASA Statement on Statistical Significance and P-Values. The American Statistician. Vol. 70(2), 2016. 129-133.

Multiples Testen

- Auch wenn es keinen Effekt gibt, werde ich mit einem Signikanzniveau von z.B. 5% in 5% aller Fälle trotzdem einen signifikanten Effekt finden.

- Wenn ich nicht nur eine Hypothese teste, sondern mehrere, habe ich, auch wenn es keinen Effekt gibt, bei jeder neuen Hypothese wieder eine 5% Chance einen signifikanten Effekt zu finden.

→ Frequentistisches Null-Hypothesen Testen verführt dazu, möglichst viele Hypothesen zu testen.

HARKing

Hypothesizing After the Results are Known

(Norbert Herr, 1998, Personality and Social Psychology Review. 2 (3): 196–217.)

Publikationsbias

- Signikante Ergebnisse werden publiziert.
- Nicht-signifikante Ergebnisse werden nicht (so leicht) publiziert, sondern bleiben in der Schublade.

→ Effekte erscheinen größer als sie wirklich sind.

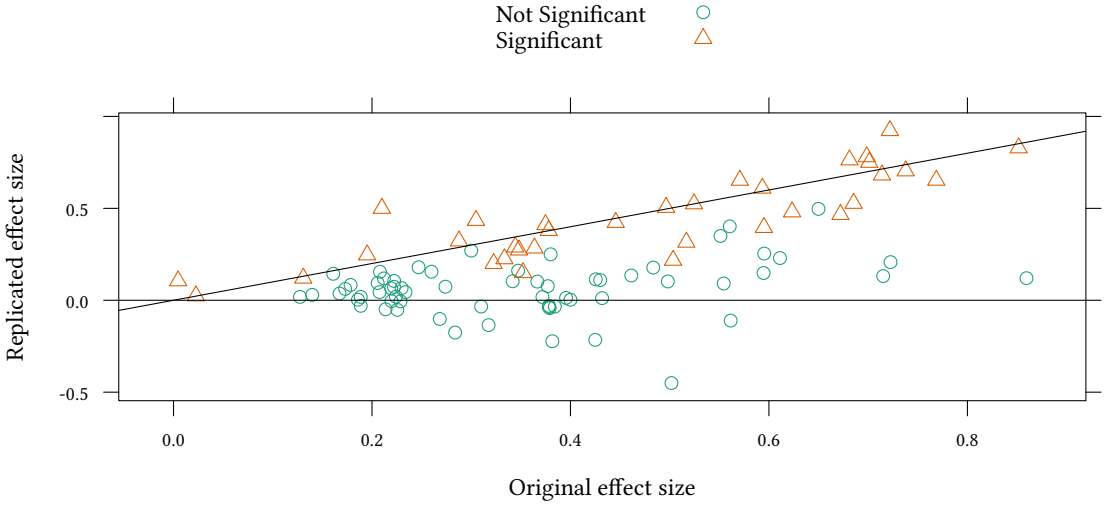
(Theodore D. Sterling, 1959, Publication decisions and their possible effects on inferences drawn from tests of significance–or vice versa. *Journal of the American Statistical Association*. 54 (285): 30–3.)

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*. 349 (6251): aac4716.

- 100 Artikel aus drei Zeitschriften (Psychological Science, Journal of Personality and Social Psychology, and Journal of Experimental Psychology: Learning, Memory, and Cognition.)

- 97 der Artikel haben signifikante Ergebnisse ($p < 0.05$).

- 36 der replizierten Studien haben signifikante Ergebnisse ($p < 0.05$).



Colin F. Camerer, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, Hang Wu. (2016). Evaluating replicability of laboratory experiments in economics. *Science*. 351 (6280): 1433-1436.

- 18 Artikel aus *American Economic Review* and *Quarterly Journal of Economics* zwischen 2011 und 2014.
- 61% der Replikationen finden einen signifikanten Effekt in der gleichen Richtung wie die ursprüngliche Studie.

5.9. Literatur

- Dolić, Statistik mit R, Kapitel 7 - 7.3.
- Hartung, Statistik, Kapitel III.4 - III.6., IV.1.5.
- Schira, Statistische Methoden der VWL und BWL-Theorie und Praxis, Kapitel 15.
- Verzani, Using R for Introductory Statistics, Chapter 8.2, 8.3.
- Greenland, Senn, Rothman, Carlin, Poole, Goodman, Altman. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European Journal of Epidemiology. 2016, Vol. 31(4). 337-350.

5.10. Schlüsselbegriffe

- Nullhypothese / Alternativhypothese
- Einseitiger und zweiseitiger Test
- Fehler 1. und 2. Art, Spezifität, Sensitivität
- Ablehnungsbereich
- Signifikanzniveau
- Parametertest
- p-Wert

- normalverteilte Zufallsvariablen $\rightarrow$ parametrische Tests (exakt)
- große Stichprobe/nicht normalverteilte Zufallsvariablen $\rightarrow$ parametrische Tests (asymptotisch)
  - bekannte Varianz $\rightarrow$ $(\bar{x}-\mu_0)/\sigma_{\bar{x}}$ ist normalverteilt

Anhang 5.A Übungen

Übung 5.1 *In einer Stadt werden regelmäßig Niederschlagsmessungen durchgeführt. Das Jahr 2016 war recht kalt und es hat oft geregnet. Sie vermuten, dass im Durchschnitt höhere Niederschlagsmengen in 2016 als in den vorhergegangenen Jahren gefallen sind.*

Die Variable X *bezeichnet den jährlichen Niederschlag und die Variable* $\mu_{x,0}$ *enthält den langjährigen Mittelwert aus der Vergangenheit.*

Stellen Sie die Null- und Alternativhypothese auf.

Übung 5.2 *Ein Land weist in den letzten Jahren stabile Wirtschaftsstrukturen und stabile Geburtenraten auf. Sie vermuten, dass die Zahl der Arbeitslosen im Durchschnitt konstant geblieben ist.*

Die Variable X *enthält die aktuelle Zahl der Arbeitslosen und die Variable* $\mu_{x,0}$ *enthält den langjährigen Mittelwert aus der Vergangenheit (in der gleichen Einheit). Stellen Sie die Null- und Alternativhypothese auf.*

Übung 5.3 *Um die Unfallzahlen zu senken, wurde ein Gesetz erlassen, das Fahren mit Abblendlicht am Tag vorschreibt. Sie vermuten, dass die Unfallzahlen im Durchschnitt gesunken sind.*

Die Variable X *enthält die Zahl der Unfälle im Monat und die Variable* $\mu_{x,0}$ *enthält den langjährigen Mittelwert aus der Vergangenheit.*

Verwenden Sie für die nächste Aufgabe die folgenden Quantile:

	0.001	0.0025	0.005	0.01	0.025	0.05	0.1
$Q^N(x)$	-3.090	-2.807	-2.576	-2.326	-1.960	-1.645	-1.282
$Q^t_{19}(x)$	-3.579	-3.174	-2.861	-2.539	-2.093	-1.729	-1.328
$Q^t_{20}(x)$	-3.552	-3.153	-2.845	-2.528	-2.086	-1.725	-1.325
$Q^t_{21}(x)$	-3.527	-3.135	-2.831	-2.518	-2.080	-1.721	-1.323

Übung 5.4 *Sie überprüfen die Abfüllmaschine für Duschgel eines Kosmetikunternehmens. Aus einer Stichprobe von 15 Duschgelflaschen bestimmen Sie eine mittlere Füllmenge von 250 ml bei einer Varianz von 24. Sie nehmen an, dass die Füllmenge normalverteilt ist.*

1. *Sie erwarten eigentlich eine Varianz der Füllmenge von 14. Welche Hypothesen stellen Sie auf?*
2. *Sie gehen nun von einer Varianz von 25 aus. Wie groß muss Ihre Stichprobe sein, damit der Bereich, in dem die Nullhypothese nicht abgelehnt wird, für die mittlere Füllmenge zum Signifikanzniveau von 5% eine Breite von nicht mehr als 1.96 ml hat?*

Übung 5.5 *Von einer Abfüllanlage eines Getränkeherstellers werden Flaschen mit Wasser befüllt. Anhand einer Stichprobe soll nachgewiesen werden, ob die Sollfüllmenge von* 1000ml *eingehalten wird.*

1. *Eine Verbraucherschutzorganisation will nachweisen, dass die tatsächliche durchschnittliche Füllmenge kleiner ist. Welche Hypothesen muss sie dazu aufstellen?*
2. *Der Getränkehersteller will nachweisen, dass die tatsächliche durchschnittliche Füllmenge immer eingehalten wird und sogar noch größer ist. Welche Hypothesen muss er dazu aufstellen?*
3. *Die Eichkomission will überprüfen, ob die durchschnittliche Füllmenge genau eingehalten wird. Welche Hypothesen muss sie dazu aufstellen?*

Übung 5.6 *Eine fiktive Vereinbarung der Europäischen Sozialminister sieht vor, dass der durchschnittliche Bruttolohn in allen EU Staaten 11.25 € betragen soll. Verwenden Sie den Datensatz* `Bwages` *aus der Bibliothek* `Ecdat` *um diese Hypothese für Belgien auf einem 5% Signifikanzniveau zu prüfen.*

Übung 5.7 *Eine genauere Lektüre der fiktiven Vereinbarung der Europäischen Sozialminister ergibt, dass der durchschnittliche Bruttolohn in allen EU Staaten* mindestens *11.25 € betragen soll. Testen Sie wieder auf einem 5% Signifikanzniveau.*

Übung 5.8 *Betrachten Sie wieder den Datensatz* `Bwages` *aus der Bibliothek* `Ecdat` *und testen Sie auf einem 5% Signifikanzniveau ob der Lohn im Durchschnitt 11.25€ ist. Bestimmen Sie den* p*-Wert.*

Beachten Sie: Der t-Test, den R (und praktisch alle anderen Statistikprogramme) durchführen, verwendet *nicht* die Standardnormalverteilung, sondern die t-Verteilung mit $n-1$ Freiheitsgraden.

Wir werden weiter unten sehen, unter welchen Umständen das richtig ist.

Übung 5.9 *Testen Sie nun, wie in Beispiel 5.7, ob der durchschnittliche Bruttolohn in Belgien* mindestens *11.25 € beträgt. Bestimmen Sie den* p*-Wert.*

Übung 5.10 *Einer Pressererklärung der Universität Harvard zufolge erhalten 70% der Studierenden in Harvard finanzielle Unterstützung durch die Universität. Ihre Nullhypothese ist, dass diese Zahl zutrifft. Sie befragen 50 Studierende, und stellen fest, dass davon 30 finanzielle Unterstützung durch ihre Universität erhalten. Wenn Sie Ihre Teststatistik als approximativ Normalverteilt annehmen, wie bestimmen Sie einen* p*-Wert um Ihre Nullhypothese zu prüfen?*

Hinweis: Für die Binomialverteilung mit Stichprobengröße n und Erfolgswahrscheinlichkeit p gilt $\mu = n \cdot p$ und $\sigma^2 = n \cdot p \cdot (1-p)$.

Es gibt zwei Möglichkeiten, diese Aufgabe zu lösen. Entweder man betrachtet die Anzahl der Erfolge (Stipendien) im Zähler der Teststatistik, dann muss im Nenner auch die Standardabweichung der Anzahl der Erfolge stehen, oder man betrachtet die Anzahl der Stipendien pro Student, dann steht im Nenner auch die dazu passende Standardabweichung. Beide Ansätze führen zum gleichen Ergebnis.

Übung 5.11 *Max macht ein Praktikum in einer Schokoladenfabrik. Heute hat er den Auftrag, nachzuweisen, dass die Maschine für 100g-Tafeln falsch eingestellt ist und die hergestellten Schokoladentafeln im Mittel weniger als die angegebenen 100g wiegen. Die Maschine hat laut Herstellerangaben eine Standardabweichung von 4g.*

1. *Welche Hypothesen stellt er dazu richtigerweise auf?*

2. *Er nimmt 16 Tafeln aus der laufenden Produktion heraus, wiegt sie sorgfältig und stellt ein durchschnittliches Gewicht von 98 g fest. Nun testet er zum Signifikanzniveau von 1% seine Vermutung. Dabei geht er davon aus, dass das Schokoladentafelgewicht normalverteilt ist. Was stellt er fest wenn er die übliche Testfunktion verwendet?*

   Die folgende Tabelle gibt Quantile der Normalverteilung Q_N*, der* t*-Verteilung mit 15 Freiheitsgraden* $Q_{t_{15}}$*, der* t*-Verteilung mit 99 Freiheitsgraden* $Q_{t_{99}}$*, und der* χ^2 *Verteilung mit 15 Freiheitsgraden* $Q_{\chi^2_{15}}$ *an.*

x	0.95	0.99	0.995
$Q_N(x)$	1.64	2.33	2.58
$Q_{t_{15}}(x)$	1.75	2.60	2.95
$Q_{t_{99}}(x)$	1.66	2.36	2.63
$Q_{\chi^2_{15}}(x)$	25.00	30.58	32.80

3. *Max ist mit den Ergebnis seiner Untersuchungen unzufrieden. Deshalb wiederholt er seinen Test. Diesmal nimmt er gleich 100 Tafeln Schokolade, wiegt sie wieder und stellt wieder ein durchschnittliches Gewicht von 98 g fest. Kann seine Vermutung nun (mit einem Signifikanzniveau von 1%) bestätigt werden?*

4. *Was passiert, wenn der gleiche Sachverhalt zu einem Signifikanzniveau von 5% getestet wird?*

Übung 5.12 *Nach 100 maligem Werfen eines Würfels ist 12 mal die 6 gefallen. Es wird vermutet, dass der Würfel manipuliert ist. Dies soll getestet werden zum Signifikanzniveau $\alpha = 0,05$. Approximieren Sie die Binomialverteilung durch eine Normalverteilung.*

1. *Welches Ergebnis hat die Testfunktion?*
2. *Kann die Vermutung, dass der Würfel manipuliert ist, bestätigt werden?*
3. *Was passiert, wenn die Irrtumswahrscheinlichkeit 10% betragen soll?*
4. *Wie oft darf die 5 bei dem Versuch von oben (100maliges Würfeln, Signifikanzniveau 5%) auftreten, damit die Nullhypothese abgelehnt werden kann?*

Hinweis: Für die Binomialverteilung mit Stichprobengröße n und Erfolgswahrscheinlichkeit p gilt $\mu = n \cdot p$ und $\sigma^2 = n \cdot p \cdot (1-p)$.

Es gibt zwei Möglichkeiten, diese Aufgabe zu lösen. Entweder man betrachtet die Anzahl der Erfolge im Zähler der Teststatistik, dann muss im Nenner auch die Standardabweichung der Anzahl der Erfolge stehen, oder man betrachtet die Anzahl der Erfolge pro Wurf, dann steht im Nenner auch die dazu passende Standardabweichung. Beide Ansätze führen zum gleichen Ergebnis.

Übung 5.13 *Das Gewicht von Orangen der Güteklasse A sei normalverteilt mit einer Varianz von 100. Sie sind Einzelhändler und bekommen eine neue Lieferung. Grundsätzlich nehmen Sie diese nur an, wenn die Orangen mindestens ein durchschnittliches Gewicht von 150 g haben. Sie testen bei jeder Lieferung zum Signifikanzniveau von 10%, ob die Orangen Ihren Anforderungen genügen.*

1. *Wie lauten die Hypothesen?*
2. *Wie jedesmal nehmen Sie eine Stichprobe von 10 Orangen, wiegen sie ab und stellen ein durchschnittliches Gewicht von 156 g fest. Sollen Sie die Orangenlieferung annehmen?*
3. *Wie würde sich der Fehler 2. Art äußern?*
4. *Sie wissen, dass es nur zwei Typen von Orangenlieferungen gibt. Entweder ist das Gewicht in Ordnung (> 150) oder das mittlere Gewicht liegt bei 145. Wie viele Orangen müssten mindestens gewogen werden, um die Wahrscheinlichkeit, einen Fehler 2. Art auf maximal 15% zu begrenzen?*

Übung 5.14 *Einer Presseerklärung des Bundesministeriums für Bildung und Forschung zufolge erhalten in Deutschland 3% aller Studierenden ein Stipendium aus Mitteln dieser Bundesbehörde. Ihre Nullhypothese ist, dass diese Zahl zutrifft. Sie befragen 100 Studierende in Ihrer Universität, und stellen fest, dass davon 2 ein Stipendium aus Mitteln des Bundesministeriums für Bildung und Forschung erhalten. Nehmen Sie, dass Ihre Teststatistik approximativ normalverteilt ist. Der Mittelwert der Binomialverteilung mit Stichprobengröße n und Erfolgswahrscheinlichkeit p ist $\mu = n \cdot p$. Die Varianz einer solchen Zufallsvariablen ist var$(X) = n \cdot p \cdot (1-p)$*

1. *Sollten Sie für Ihre Teststatistik die korrigierte oder die unkorrigierte Standardabweichung verwenden?*

2. *Muss die Standardabweichung noch durch $\sqrt{n}$ geteilt werden?*

3. *Bestimmen Sie in R einen* p*-Wert um Ihre Nullhypothese zu prüfen?*

Übung 5.15 *Sie stellen Tachometer her, die die Geschwindigkeit von Kraftfahrzeugen messen. Ein Automobilclub interessiert sich für die Geschwindigkeit, die Ihre Tachometer im Durchschnitt bei einer tatsächlichen Geschwindigkeit von 50 km/h anzeigen. Sie wissen, dass in diesem Bereich die Standardabweichung der Anzeige Ihrer Tachometer 5 km/h beträgt. Außerdem gehen Sie davon aus, dass der Fehler der Anzeige Ihrer Tachometer normalverteilt ist. Sie nehmen 100 Tachometer aus der laufenden Produktion, und stellen fest, dass diese Tachometer bei einer tatsächlichen Geschwindigkeit von 50 km/h im Mittel 55 km/h anzeigen.*

1. *Geben Sie den Bereich, in dem die Nullhypothese nicht abgelehnt wird, für die durchschnittlich angezeigten Geschwindigkeit bei einem Signifikanzniveau von 1% an. Schreiben Sie Ihr Ergebnis als R Ausdruck auf und verwenden Sie dabei, dass* `qnorm(0.99)` *das 99% Quantil der Normalverteilung,* `qnorm(0.98)` *das 98% Quantil, sowie* `qnorm(0.995)` *das 99.5% Quantil der Normalverteilung ergibt. Verwenden Sie diese Ausdrücke in Ihrer Lösung.*

2. *Oben haben Sie angenommen, dass die Standardabweichung bekannt ist. Wie ändert sich Ihr Ergebnis, wenn Sie mit der empirischen Standardabweichung rechnen. R gibt Ihnen die empirische Standardabweichung als* `sd(x)` *an.*

3. *Eigentlich gehen Sie von einem Erwartungswert für die angezeigte Geschwindigkeit von 56 km/h aus. Stellen Sie eine Nullhypothese auf, berechnen Sie eine Teststatistik für den zweiseitigen Test (als R-Kommando und mit der empirischen Standardabweichung), und rechnen Sie einen* p*-Wert aus (ebenfalls als R-Kommando).*

Übung 5.16 *Welche der folgenden Aussagen sind korrekt?*

1. *Ein Punktschätzer ist erwartungstreu, wenn der erwartete Schätzwert dem wahren zu schätzenden Parameter der Verteilung entspricht.*

2. *Je verzerrter ein Schätzer, desto höher seine Varianz.*

3. *Konfidenzintervalle minimieren die Varianz erwartungstreuer Schätzfunktionen.*

4. *Konfidenzintervalle spannen ein Intervall um einen geschätzen Parameter einer Verteilung.*

5. *Mit zunehmender Varianz verkleinert sich der Wertebreich des Konfidenzintervalls.*

6. *Das α-Niveau eines Hypothesentests begrenzt die Wahrscheinlichkeit die Nullhypothese fäschlicherweise abzulehnen.*

7. *Das α-Niveau eines Hypothesentests begrenzt die Wahrscheinlichkeit die Alternativhypothese fäschlicherweise abzulehnen.*

Übung 5.17 *Ihre Nullhypothese sei H_0, die Alternativhypothese sei H_1. Welche der folgenden Aussagen trifft zu?*

- *Der Fehler 1. Art gibt an, wie häufig H_0 nicht abgelehnt wird, obwohl sie falsch ist.*
- *Der Fehler 1. Art gibt an, wie häufig H_0 abgelehnt wird, obwohl sie wahr ist.*
- *Der Fehler 2. Art gibt an, wie häufig H_0 nicht abgelehnt wird, obwohl sie falsch ist.*
- *Der Fehler 1. Art gibt an, wie häufig H_1 abgelehnt wird, obwohl sie wahr ist.*
- *Der Fehler 2. Art gibt an, wie häufig H_1 nicht abgelehnt wird, obwohl sie falsch ist.*

6. Tests für Mittelwerte — parametrisch

6.1. Motivation: Ultimatum Verhandlungen

Wie verhält sich ein Entscheider in einer Verhandlungssituation, wenn er sehr viel Verhandlungsmacht hat? Im Ultimatumspiel kann ein Entscheider ein Angebot machen. Wenn der Empfänger das Angebot annimmt, wird das Angebot wie vorgeschlagen umgesetzt. Wenn der Empfänger ablehnt, enden die Verhandlungen, und beide bekommen nichts.

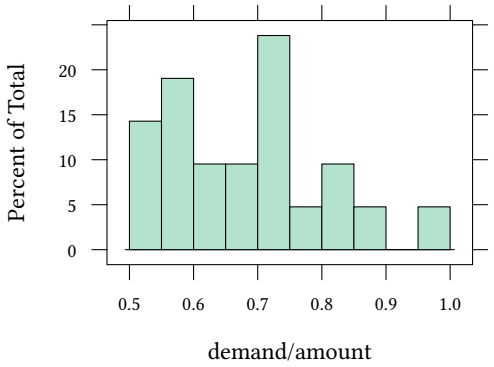
Im Labor implementiert man diese Situation so, dass der Entscheider, der das Angebot macht, einen Geldbetrag aufteilt. Wenn der Empfänger das Angebot akzeptiert, wird der Geldbetrag so aufgeteilt, und beide verlassen das Labor mit ihrem jeweiligen Anteil. Wenn der Empfänger das Angebot ablehnt, bekommen beide nichts.

Ein egoistischer Empfänger wird also jedes Angebot annehmen. Deshalb wird ein egoistischer Entscheider, der um den Egoismus des Empfängers weiß, dem Empfänger nichts oder fast nichts anbieten.

Güth, Schmittberger und Schwarze haben das Verhalten in Ultimatumspielen im Experiment untersucht.

Fragestellungen:

- Wird die theoretische Verhandlungsmacht genutzt?
- Sind die Angebote fair?



29% der Angebote wurden zurückgewiesen.

Güth, Schmittberger, Schwarze. An Experimental Analysis of Ultimatum Bargaining. Journal of Economic Behavior and Organization, 1982.

Wie kann man testen, ob die im Experiment beobachtete Abweichung von der spieltheoretischen Lösung bemerkenswert ist?

6.2. Test für Mittelwerte bei unbekannter Varianz

6.2.1. Test des Mittelwerts - p-Wert

In Kapitel 5 haben wir Hypothesentests betrachtet. Dabei haben wir angenommen, dass die Varianz der Variablen, die wir beobachten, bekannt ist. Das war eine Vereinfachung.

In Kapitel 6 berücksichtigen wir nun, dass die Varianz normalerweise unbekannt ist.

bekannte Varianz	unbekannte Varianz
$g = \frac{\bar{x}^{\text{Stp.}} - \mu_{X,0}}{\sigma_X/\sqrt{n}}$	$g = \frac{\bar{x}^{\text{Stp.}} - \mu_{X,0}}{\hat{\sigma}_X/\sqrt{n}}$
	$\hat{\sigma}_X^2$ ist unser Schätzer für σ_X^2, also $\hat{\sigma}_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}^{\text{Stp.}}\right)^2$

falls X normalverteilt und H_0 wahr ist, dann (vor der Stichprobe)

$g \sim N(0,1)$	$g \sim t_{n-1}$
(Normalverteilung)	(Student t-Verteilung mit $n-1$ Freiheitsgraden)

falls X einer beliebigen Verteilung folgt und H_0 wahr ist und $n \to \infty$ dann:

$$g \sim N(0,1) = t_\infty$$

Wie in Abschnitt 5 können wir diese Teststatistik verwenden, um Hypothesen zu testen.

Zweiseitiger Test ($H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$)

$$\begin{aligned}
\text{p-Wert} &= \Pr_{H_0}\left(\left|\bar{X}-\mu_{X,0}\right| > \left|\bar{x}^{\text{Stp.}}-\mu_{X,0}\right|\right)\\
&= \Pr_{H_0}\left(\left|\frac{\bar{X}-\mu_{X,0}}{\hat{\sigma}_X/\sqrt{n}}\right| > \left|\frac{\bar{x}^{\text{Stp.}}-\mu_{X,0}}{\hat{\sigma}_X/\sqrt{n}}\right|\right)\\
&\approx 2\cdot F_{t,n-1}\left(-\left|\frac{\bar{x}^{\text{Stp.}}-\mu_{X,0}}{\hat{\sigma}_X/\sqrt{n}}\right|\right)
\end{aligned}$$

Einseitiger Test ($H_0:\mu=\mu_0, H_1:\mu>\mu_0$)

$$\begin{aligned}
\text{p-Wert} &= \Pr_{H_0}\left(\bar{X}-\mu_{X,0} > \bar{x}^{\text{Stp.}}-\mu_{X,0}\right)\\
&= \Pr_{H_0}\left(\frac{\bar{X}-\mu_{X,0}}{\hat{\sigma}_X/\sqrt{n}} > \frac{\bar{x}^{\text{Stp.}}-\mu_{X,0}}{\hat{\sigma}_X/\sqrt{n}}\right)\\
&\approx F_{t,n-1}\left(-\frac{\bar{x}^{\text{Stp.}}-\mu_{X,0}}{\hat{\sigma}_X/\sqrt{n}}\right)
\end{aligned}$$

Einseitiger Test ($H_0:\mu=\mu_0, H_1:\mu<\mu_0$)

$$\begin{aligned}
\text{p-Wert} &= \Pr_{H_0}\left(\bar{X}-\mu_{X,0} < \bar{x}^{\text{Stp.}}-\mu_{X,0}\right)\\
&= \Pr_{H_0}\left(\frac{\bar{X}-\mu_{X,0}}{\hat{\sigma}_X/\sqrt{n}} < \frac{\bar{x}^{\text{Stp.}}-\mu_{X,0}}{\hat{\sigma}_X/\sqrt{n}}\right)\\
&\approx F_{t,n-1}\left(\frac{\bar{x}^{\text{Stp.}}-\mu_{X,0}}{\hat{\sigma}_X/\sqrt{n}}\right)
\end{aligned}$$

H_0 wird jeweils abgelehnt wenn p-Wert $< \alpha$

Für große Werte der Freiheitsgrade, bzw. für große n, konvergiert die t-Verteilung gegen die Standardnormalverteilung.

Die t-Verteilung sieht ähnlich aus, wie die Standardnormalverteilung, allerdings braucht die t-Verteilung noch einen Parameter den wir »Freiheitsgrade« nennen. Wenn wir Mittelwerte schätzen, sind die Freiheitsgrade = $n-1$.

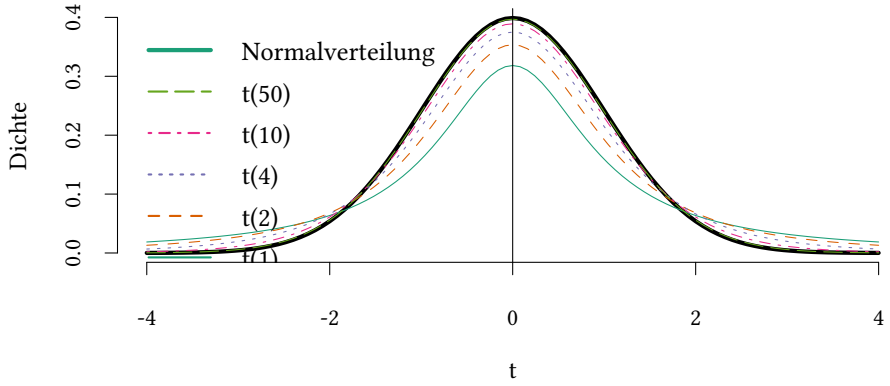
Die Idee zur t-Verteilung stammt von *William Sealy Gosset*. Gosset arbeitete in der Guinness Brauerei und musste zwischen verschiedenen Qualitäten von Gerste zum Brauen von Bier auswählen. Sein Problem war, dass die Anzahl der Beobachtungen (verschiedene Lieferungen von Gerste) klein war. Zu dieser Zeit gab es nur statistische Verfahren für große Stichproben.

→ Entwicklung von statistischen Verfahren für kleine Stichproben.

W. S. Gosset (published as Student). The probable error of a mean. *Biometrica*, 1908.

`dt` bestimmt die Dichte unter der t-Verteilung. Analog bestimmt `qt` ein Quantil der t-Verteilung `pt` berechnet ein Perzentil, und , `rt` berechnet t-verteilte Zufallszahlen.

Hier ist ein Bild der t-Verteilung für verschiedene Freiheitsgrade.



Beobachtung: Wenn n groß ist (d.h. die Anzahl der Freiheitsgrade groß ist), dann wird der Unterschied zwischen t-Verteilung und Normalverteilung sehr klein.

6.2.2. Test des Mittelwerts bei gegebenem Signifikanzniveau α

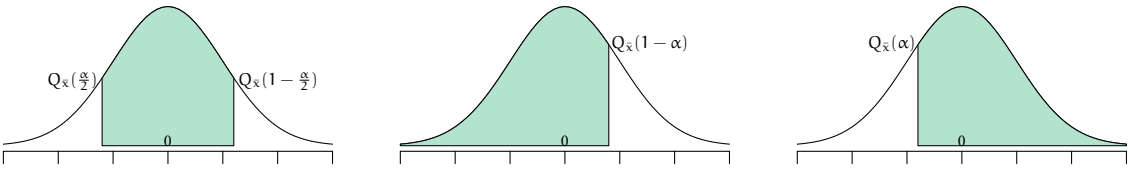
- Berechnen der Teststatistik $g(X_1,\dots,X_n)=\frac{\bar{x}^{\text{Stp.}}-\mu_{X,0}}{\hat{\sigma}_X/\sqrt{n}}$

  (wobei $\hat{\sigma}_X^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}^{\text{Stp.}}\right)^2$)

- falls X normalverteilt und H_0 wahr ist, dann $g(X_1,\dots,X_n)\sim t_{n-1}$ (Student t-Verteilung mit $n-1$ Freiheitsgraden).

- falls X einer beliebigen Verteilung folgt und H_0 wahr ist und $n\to\infty$ dann $g(X_1,\dots,X_n)\sim N(0,1)$.

Wir bestimmen die kritischen Werte $Q(\alpha)$ bzw. $Q(1-\alpha)$ (für einen einseitigen Test) oder $Q(\alpha/2)$ und $Q(1-\alpha/2)$ (für einen zweiseitigen Test).



		ablehnen, falls	
$H_0:\mu=\mu_{X,0}$ vs.	$H_1:\mu\neq\mu_{X,0}$	$g\notin[Q(\frac{\alpha}{2}),Q(1-\frac{\alpha}{2})]$	zweiseitiger Test
$H_0:\mu=\mu_{X,0}$ vs.	$H_1:\mu>\mu_{X,0}$	$g\notin[-\infty,Q(1-\alpha)]$	einseitiger Test
$H_0:\mu=\mu_{X,0}$ vs.	$H_1:\mu<\mu_{X,0}$	$g\notin[Q(\alpha),\infty]$	einseitiger Test

Vergleich: p-Wert und Signifikanzniveau Das Signifikanzniveau α wird vorgegeben. Z.B. wenn das vorgegebene Signifikanzniveau $\alpha=5\%$ ist,...

- ...wird die Nullhypothese verworfen, wenn $|g(X_1,\dots,X_n)|>1.96$ ist,

- ...äquivalent wird die Nullhypothese verworfen, wenn $p < 0.05$ ist.
- Wir nennen den p-Wert auch *marginales Signifikanzniveau.*
- Oft ist es informativer den p-Wert anzugeben, als zu sagen, ob der Test ablehnt oder nicht.

Wann verwenden wir die t-Verteilung, und wann die Normalverteilung?

- Wenn $X_1, \ldots, X_n$ i.i.d. (unabhängig und gleichverteilt) ist und *normalverteilt* entsprechend $N(\mu_X, \sigma_X^2)$, dann folgt die Teststatistik $g(X_1, \ldots, X_n)$ der Student t-Verteilung mit $n - 1$ Freiheitsgraden.

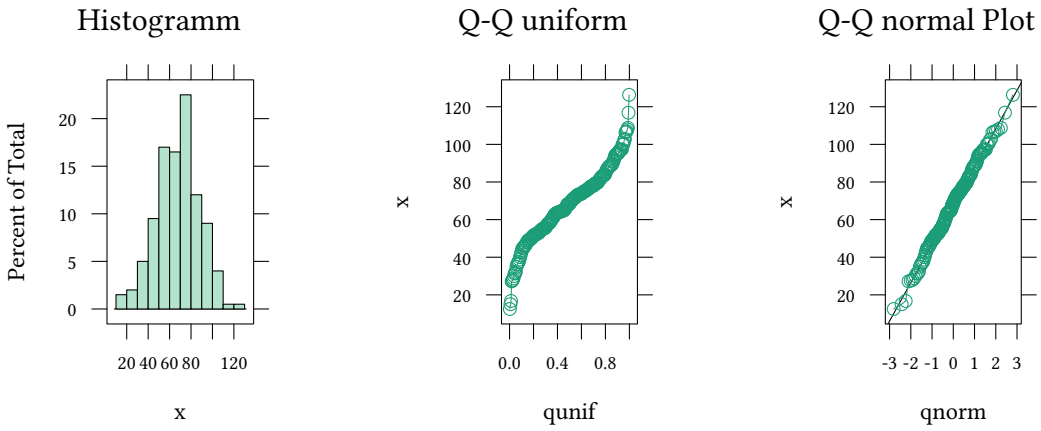
$$g(X_1, \ldots, X_n) \sim t_{n-1}$$

Aber

- Wenn die $X_1, \ldots, X_n$ *nicht normalverteilt* sind (und die Stichprobe *klein* ist), dann nützt uns das alles nichts.

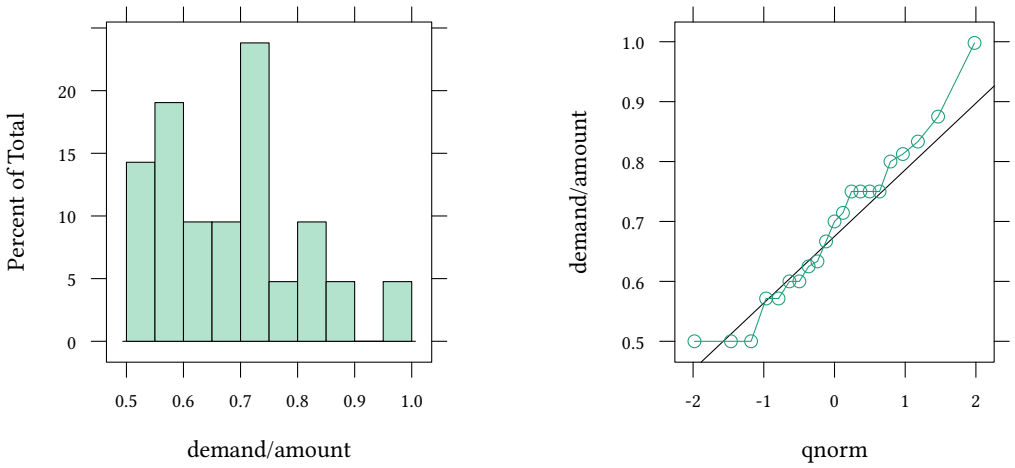
Manche ökonomische Daten sind normalverteilt, andere nicht. Eine einfache Möglichkeit der Überprüfung der Normalverteilungsannahme bietet der Q-Q normal Plot.

Hier ist zunächst eine normalverteilte Variable. Links ein Histogramm. Das ist oft nicht besonders ausagekräftig. In der Mitte ein Q-Q uniform Plot. In diesem Plot wird an der vertikalen Achse die Variable selbst, und an der Horizontalen Achse das Quantil dargestellt. Eine normalverteilte Zufallsvariable liegt in diesem Diagramm auf einer Kurve. Um diese Kurve besser beurteilen zu können, verwendet man im Q-Q normal Plot Quantile der Normalverteilung an der horizontalen Achse. Deshalb sollten nun alle Beobachtungen bei einer normalverteilten Variablen auf einer Linie liegen.



- Es ist schwer, anhand eines Histogramms zu beurteilen, ob eine Stichprobe aus einer Normalverteilung gezogen wurde.
- Stellt man eine normalverteilte Zufallsvariable in einem Q-Q-normal Plot dar, dann liegen alle Beobachtungen auf einer Linie.

Wie sieht es z.B. mit unseren Angeboten im Ultimatumspiel aus? Sind die auch (etwa) normalverteilt?



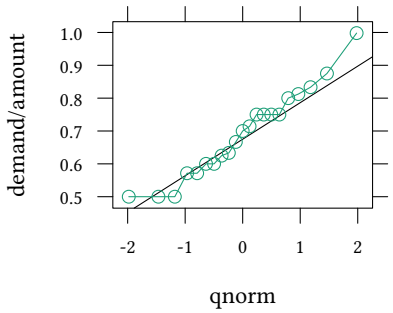
Wir sehen, dass es in diesem Beispiel eine (kleine) Abweichung von der Normalverteilung gibt. Bei einer kleinen Stichprobe ist das lästig. Bei einer großen Stichprobe ist das halb so schlimm (weil die Mittelwerte einer großen Stichprobe immer noch approximativ einer Normalverteilung folgen).

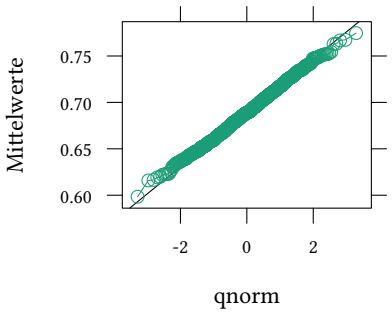
Formale Tests of Normalität Natürlich gibt es auch formale Tests auf Normalität:
H_0: $X \sim N$, H_1: $X \not\sim N$

- Anderson–Darling test
- Cramer–von Mises test
- D'Agostino test
- Jarque–Bera test
- Kolmogorov-Smirnov test
- Lilliefors test
- Pearson χ^2 test
- Shapiro–Francia test
- Shapiro-Wilk's test
- Egal wie X verteilt ist, wenn n groß wird, konvergiert $\bar{X}$ sowieso zur Normalverteilung (zentraler Grenzwertsatz).

$$g(X_1, \ldots, X_n) \sim N(0, 1)$$

Wenn unsere Stichprobe also hinreichend groß ist, haben wir kein Problem und können mit der t-Verteilung oder der Normalverteilung (zwischen den beiden Verteilungen besteht für große n kein Unterschied mehr) weiterarbeiten.





6.3. Vergleich von zwei Stichproben mit unbekannter Varianz

6.3.1. Unverbundene Stichproben

Vergleiche nun $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$ mit $\bar{y} = \frac{1}{m}\sum_{i=1}^{m} Y_i$

$$\hat{\sigma}^2_{\bar{x}} = \frac{\hat{\sigma}^2_X}{n} \qquad \hat{\sigma}^2_{\bar{y}} = \frac{\hat{\sigma}^2_Y}{m}$$

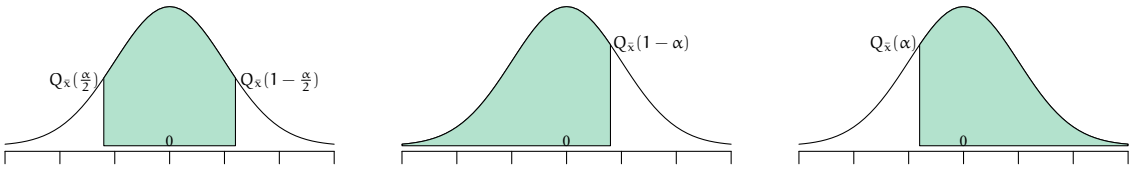
Varianz der Differenz $\bar{x} - \bar{y}$ also

$$\hat{\sigma}^2_{\bar{x}-\bar{y}} = \hat{\sigma}^2_{\bar{x}} + \hat{\sigma}^2_{\bar{y}} = \frac{\hat{\sigma}^2_X}{n} + \frac{\hat{\sigma}^2_Y}{m}$$

Unter H_0 ist unsere Teststatistik

$$g = \frac{\bar{x} - \bar{y}}{\hat{\sigma}_{\bar{x}-\bar{y}}} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\hat{\sigma}^2_X}{n} + \frac{\hat{\sigma}^2_Y}{m}}} \sim t_{n+m-2}$$

Mit dieser Teststatistik können wir wieder p-Werte ausrechnen oder Signifikanztests zu einem vorgegebenen Signifikanzniveau α durchführen.



		ablehnen, falls	
$H_0 : \mu_X = \mu_Y$ vs.	$H_1 : \mu_X \neq \mu_Y$	$g \notin [Q_{t,n+m-2}(\frac{\alpha}{2}), Q_{t,n+m-2}(1-\frac{\alpha}{2})]$	zweiseit. Test
$H_0 : \mu_X = \mu_Y$ vs.	$H_1 : \mu_X > \mu_Y$	$g \notin [-\infty, Q_{t,n+m-2}(1-\alpha)]$	einseit. Test
$H_0 : \mu_X = \mu_Y$ vs.	$H_1 : \mu_X < \mu_Y$	$g \notin [Q_{t,n+m-2}(\alpha), \infty]$	einseit. Test

Zweiseitiger Test ($H_0 : \mu_X = \mu_Y, H_1 : \mu_X \neq \mu_Y$)

$$\begin{aligned}
\text{p-Wert} &= \Pr_{H_0}\left(\left|\bar{X}-\bar{Y}\right| > |\bar{x}-\bar{y}|\right)\\
&= \Pr_{H_0}\left(\left|\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{\hat{\sigma}_X^2}{n}+\frac{\hat{\sigma}_Y^2}{m}}}\right| > \left|\frac{\bar{x}-\bar{y}}{\sqrt{\frac{\hat{\sigma}_X^2}{n}+\frac{\hat{\sigma}_Y^2}{m}}}\right|\right)\\
&\approx 2\cdot F_{t,n+m-2}\left(-\left|\frac{\bar{x}-\bar{y}}{\sqrt{\frac{\hat{\sigma}_X^2}{n}+\frac{\hat{\sigma}_Y^2}{m}}}\right|\right)
\end{aligned}$$

Einseitiger Test ($H_0 : \mu_X = \mu_Y, H_1 : \mu_X > \mu_Y$)

$$\begin{aligned}
\text{p-Wert} &= \Pr_{H_0}\left(\bar{X}-\bar{Y} > \bar{x}-\bar{y}\right)\\
&= \Pr_{H_0}\left(\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{\hat{\sigma}_X^2}{n}+\frac{\hat{\sigma}_Y^2}{m}}} > \frac{\bar{x}-\bar{y}}{\sqrt{\frac{\hat{\sigma}_X^2}{n}+\frac{\hat{\sigma}_Y^2}{m}}}\right)\\
&\approx F_{t,n+m-2}\left(-\frac{\bar{x}-\bar{y}}{\sqrt{\frac{\hat{\sigma}_X^2}{n}+\frac{\hat{\sigma}_Y^2}{m}}}\right)
\end{aligned}$$

Einseitiger Test ($H_0 : \mu_X = \mu_Y, H_1 : \mu_X < \mu_Y$)

$$\begin{aligned}
\text{p-Wert} &= \Pr_{H_0}\left(\bar{X}-\bar{Y} < \bar{x}-\bar{y}\right)\\
&= \Pr_{H_0}\left(\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{\hat{\sigma}_X^2}{n}+\frac{\hat{\sigma}_Y^2}{m}}} < \frac{\bar{x}-\bar{y}}{\sqrt{\frac{\hat{\sigma}_X^2}{n}+\frac{\hat{\sigma}_Y^2}{m}}}\right)\\
&\approx F_{t,n+m-2}\left(\frac{\bar{x}-\bar{y}}{\sqrt{\frac{\hat{\sigma}_X^2}{n}+\frac{\hat{\sigma}_Y^2}{m}}}\right)
\end{aligned}$$

H_0 wird jeweils abgelehnt wenn p-Wert $< \alpha$
Die Teststatistik

$$g = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{\hat{\sigma}_X^2}{n}+\frac{\hat{\sigma}_Y^2}{m}}}$$

ist leider nur dann t-verteilt mit $n+m-2$ Freiheitsgraden, wenn $\sigma_X^2 = \sigma_Y^2$ (der »t-Test« den Sie in Statistikprogrammen finden, macht diese Annahme). Für große Stichproben ($n > 50$) kann man aber ausnutzen, dass g asymptotisch normalverteilt ist.

6.3.2. Paarweise verbundene Stichproben

Wir sind am Mittelwert der Differenz zwischen X_i und Y_i interessiert, betrachten also auf der Ebene der Stichprobe $\Delta_i = X_i - Y_i$. Der Mittelwert wäre $\bar{\Delta} = \bar{x} - \bar{y}$.

$\bar{\Delta}$ ist hier wieder eine übliche Zufallsvariable mit Mittelwert und Standardabweichung. Eine typische Nullhypothese wäre z.B. $H_0 : \bar{\Delta} = 0$. Unsere Teststatistik ist dann

$$g = \frac{\bar{\Delta}}{\hat{\sigma}_{\bar{\Delta}}} \sim t_{n-1}$$

Die Standardabweichung $\hat{\sigma}_{\bar{\Delta}}$ berechnen wir wie jede andere Standardabweichung:

$$\hat{\sigma}_{\Delta} = \sqrt{\frac{\sum_{i=1}^{n}(\Delta_i - \bar{\Delta})^2}{n-1}} \qquad \hat{\sigma}_{\bar{\Delta}} = \frac{1}{\sqrt{n}}\sqrt{\frac{\sum_{i=1}^{n}(\Delta_i - \bar{\Delta})^2}{n-1}}$$

Zurück zur Motivation: Ultimatum-Verhandlungen Werfen wir zum Abschluss noch einen Blick auf die Ultimatum-Verhandlungssituation.

```
t.test(share,mu=1)


 One Sample t-test

data:  share
t = -10.684, df = 20, p-value = 0.000000001029
alternative hypothesis: true mean is not equal to 1
95 percent confidence interval:
 0.6300980 0.7509473
sample estimates:
mean of x
0.6905227
```

6.4. Literatur

- Dolić, Statistik mit R, Kapitel 7 - 7.3.
- Schira, Statistische Methoden der VWL und BWL-Theorie und Praxis, Kapitel 15.
- Verzani, Using R for Introductory Statistics, Chapter 8.2, 8.3.

6.5. Schlüsselbegriffe

- t-Verteilung / Normalverteilung
- Zwei-Stichproben Test

- paarweise verbundene / nicht verbundene Stichprobe
- parametrischer / nicht-parametrischer Test

- normalverteilte Zufallsvariablen → parametrische Tests (exakt)
- große Stichprobe/nicht normalverteilte Zufallsvariablen → parametrische Tests (asymptotisch)
  - bekannte Varianz → t ist normalverteilt
  - unbekannte Varianz → t ist t-verteilt
  - paarweise Tests / Tests für unverbundende Stichproben

Anhang 6.A Beispiele für die Vorlesung

- Die Zufallsvariable X hat Mittelwert μ und Varianz 25. Sie ziehen eine Stichprobe der Größe 100 und stellen fest, dass $\sum_{i=1}^{100} X_i = 500$. Was ist die beste erwartungstreue Schätzung für μ?
- Was ist die Standardabweichung Ihres Schätzers für μ?
- Gehen Sie nun davon aus, dass die Standardabweichung Ihres Schätzers $\hat{\mu}$ den Wert 1 hat. Nehmen Sie ferner an, dass Sie einen Wert von $\hat{\mu} = 8$ erhalten haben. Ihre Nullhypothese ist $\mu_0 = 4$. Was ist der Absolutwert der Teststatistik?
- Die Zufallsvariable X ist normalverteilt mit Mittelwert μ und unbekannter Standardabweichung. Eine Stichprobe von 16 Beobachtungen ergibt einen Stichprobenmittelwert von $\bar{x} = 12$ und eine Standardabweichung der Stichprobenbeobachtungen von 8. Was ist die beste erwartungstreue Schätzung für μ?
- Was ist die Standardabweichung Ihres Schätzers für μ?
- Gehen Sie nun davon aus, dass die Standardabweichung Ihres Schätzers $\hat{\mu}$ den Wert 5 hat. Nehmen Sie ferner an, dass Sie einen Wert von $\hat{\mu} = 10$ erhalten haben. Ihre Nullhypothese ist $\mu_0 = 20$. Was ist der Absolutwert der Teststatistik?
- Wir bleiben bei der oben beschriebenen Situation, gehen aber jetzt davon aus, dass Ihre Stichprobe 8 Beobachtungen enthält, und Ihre Teststatistik den Wert -2 hat. Ihre Nullhypothese ist nach wie vor $\mu_0 = 20$, die Alternativhypothese ist $\mu_0 \neq 20$. Ihr Signifikanzniveau ist 5%.
  1. Wenn die Teststatistik einen Wert zwischen $-\infty$ und $+1.89$ hat, wird H_0 nicht abgelehnt.
  2. Zum Testen sollte man die Normalverteilung verwenden.
  3. Die Nullhypothese wird abgelehnt.

4. Die Alternativhypothese wird abgelehnt.
5. Zum Testen sollte man die t-Verteilung mit 5 Freiheitsgraden verwenden.
6. Wenn die Teststatistik einen Wert zwischen -2.36 und $+2.36$ hat, wird H_0 nicht abgelehnt.

Welche Wahrscheinlichkeit gibt der p-Wert an?

- Keine der folgenden Antworten ist richtig.
- Die Wahrscheinlichkeit eines Fehlers erster Art.
- Die Wahrscheinlichkeit eines Fehlers zweiter Art.
- Die Wahrscheinlichkeit die Alternativhypothese abzulehnen, wenn sie falsch ist.
- Die Wahrscheinlichkeit die Nullhypothese abzulehnen, wenn sie falsch ist.

Anhang 6.B Übungen

Übung 6.1 *Betrachten Sie den Datensatz `Wages` aus der Bibliothek `Ecdat`. Testen Sie, ob Männer im Durchschnitt den gleichen Lohn erhalten wie Frauen.*

Übung 6.2 *Betrachten Sie den Datensatz `Bwages` aus der Bibliothek `Ecdat`. Vergleichen Sie den Brutto-Stundenlohn `wage` für Arbeiter mit geringer Ausbildung `educ==1` mit dem von Arbeitern hoher Ausbildung `educ==5`. Verdienen Arbeiter mit hoher Ausbildung signifikant mehr?*

Wie sieht der Vergleich aus, wenn Sie `educ==1` mit `educ==2` vergleichen?

Vergleichen Sie nun `educ==1` mit `educ==2` mit einem einseitigen Test. Welche Hypothese können Sie mit diesem Test ablehnen?

Übung 6.3 *Sie testen eine Abfüllmaschine für Joghurt. Für eine Stichprobe von 10 Joghurtbechern erwarten Sie eine mittlere Füllmenge von 150 Gramm (H_0). Sie nehmen an, dass die Füllmenge normalverteilt ist.*

1. *Nehmen Sie an, die Varianz der Füllmenge Ihrer Maschine sei langfristig 10. Bestimmen Sie den Ablehnungsbereich für die mittlere Füllmenge bei einem Signifikanzniveau von 5%.*
2. *Gehen Sie nun von einer langfristigen Varianz von 9 aus. Wie groß muß Ihr Stichprobenumfang mindestens sein, damit der Ablehnungsbereich bei einem Signifikanzniveau von 5% höchstens eine Breite von 0,392 Gramm hat*
3. *Eine andere Joghurtmaschine befüllt Becher mit im Mittel 50 Gramm Joghurt. Die Standardabweichung beträgt 10 Gramm. Sie nehmen eine Stichprobe von 100 Bechern. Wie groß ist die Wahrscheinlichkeit, dass die Gesamtmenge Joghurt in diesen 100 Bechern kleiner als 4 800 Gramm ist?*

4. *Eine weitere Joghurtmaschine produziert nach Herstellerspezifikation bei 80 Joghurtbechern im Mittel 4 Becher Ausschuss. Sie nehmen eine Stichprobe von 80 Bechern, und finden darunter 6 Becher Ausschuss. Ist die Qualität dieser Maschine signifikant verschieden von der Herstellerspezifikation?*

   - *Schreiben Sie eine Nullhypothese auf und testen Sie bei einem Signifikanzniveau von 5%.*
   - *Falls Sie »krumme« Zahlen erhalten, verwenden Sie eine möglichst genaue Abschätzung.*
   - *Gehen Sie davon aus, dass Sie die Binomialverteilung durch eine Normalverteilung approximieren können.*

Übung 6.4 *Eine Universität vergleicht die durchschnittliche Performance in einem Sprachtest. Die Stichprobe besteht aus sechs Politologen und sechs Soziologen. Die Punktzahl ergibt sich aus der folgenden Tabelle:*

Fach	*1*	*2*	*3*	*4*	*5*	*6*	*Mittelwert*
Politologie	*45*	*76*	*50*	*51*	*40*	*57*	*53.17*
Soziologie	*59*	*56*	*61*	*66*	*54*	*49*	*57.5*

Sie nehmen an, dass die Punkte jeweils normalverteilt sind. Testen Sie, ob die mittlere Punktzahl für die beiden Fächer gleich ist?

Übung 6.5 *Für die nächste PISA-Studie erprobt eine Schule Methoden zur Leistungsverbesserung ihrer Schüler. Eine davon bezieht sich auf die optimale Raumtemperatur. Der Direktor wählt 10 Schüler aus und teilt sie in zwei gleichstarke Vergleichsgruppen ein. Die eine Gruppe bearbeitet eine Stunde lang knifflige Kopfrechenaufgaben in einem Raum mit 19 Grad, die die andere Gruppe sitzt dazu in einem Raum mit 23 Grad. Hier sind die Ergebnisse (Anzahl richtig gelöster Aufgaben für jeden Schüler):*

19-Grad-Gruppe:	*13, 20, 15, 10, 21*
23-Grad-Gruppe:	*17, 26, 12, 27, 14*

Testen Sie zum Niveau $\alpha = 0,05$ ob bei unterschiedlichen Raumtemperaturen wirklich im Durchschnitt unterschiedlich viele Aufgaben richtig gelöst werden.

1. *Die Teststatistik ist…*

2. *Nehmen Sie an, die Teststatistik hätte den Wert $g = 1,9$. Kann die Hypothese, dass die Leistung der Schüler unabhängig von der Temperatur ist, verworfen werden?*

Verwenden Sie die folgenden Quantile für die nächsten beiden Aufgaben:

	0.001	0.0025	0.005	0.01	0.025	0.05	0.1
$Q^N(x)$	-3.090	-2.807	-2.576	-2.326	-1.960	-1.645	-1.282
$Q^t_4(x)$	-7.173	-5.598	-4.604	-3.747	-2.776	-2.132	-1.533
$Q^t_5(x)$	-5.893	-4.773	-4.032	-3.365	-2.571	-2.015	-1.476
$Q^t_{19}(x)$	-3.579	-3.174	-2.861	-2.539	-2.093	-1.729	-1.328
$Q^t_{20}(x)$	-3.552	-3.153	-2.845	-2.528	-2.086	-1.725	-1.325
$Q^t_{21}(x)$	-3.527	-3.135	-2.831	-2.518	-2.080	-1.721	-1.323
$Q^t_{24}(x)$	-3.467	-3.091	-2.797	-2.492	-2.064	-1.711	-1.318
$Q^t_{25}(x)$	-3.450	-3.078	-2.787	-2.485	-2.060	-1.708	-1.316

Dabei ist Q^N das Quantil der Normalverteilung, und Q^t_k das Quantil der t-Verteilung mit k Freiheitsgraden.

Übung 6.6 *Sie prüfen, ob die Gewichtsangaben von 1 kg auf Lebensmittelverpackungen von Zucker der Wahrheit entsprechen. Dazu kaufen sie in einem Supermarkt 20 Päckchen Zucker und führen anschließend in einem Labor die Kontrollen durch. Die Gewichtsmessungen (in g) ergaben: 987, 1002, 993, 999, 981, 1009, 1013, 995, 1002, 1001, 998, 997, 994, 1005, 1007, 985, 995, 998, 1014, 1003.*

Sie nehmen an, dass die Füllmengen normalverteilt sind und prüfen, ob der Mittelwert des Gewichts der Zuckerpackungen 1 kg entspricht. Das Signifikanzniveau beträgt 5%.

1. *Wie lauten die Null- und die Alternativhypothese, die die Verbraucherorganisation aufstellen muss?*
2. *Wie lautet der Wert der Teststatistik?*
3. *In welchem Bereich wird die Nullhypothese nicht abgelehnt?*
4. *Wird die Nullhypothese abgelehnt?*

Übung 6.7 *Zwei Bauern möchten die Wirksamkeit von Düngemitteln testen. Bauer 1 besitzt 6 Felder und Bauer 2 bewirtschaftet 5 Felder. Es gibt keine systematischen Unterschiede zwischen den Feldern. Die Bauern haben ihre Erträge gewogen (in kg):*

Bauer 1 (X)*: 1025, 937, 1011, 980, 1007, 1120 Bauer 2* (Y)*: 897; 1215; 1106; 1080; 1175*

Die Stichproben sind voneinander unabhängig und $N_X(\mu_X; 400)$ *bzw.* $N_Y(\mu_Y; 500)$ *verteilt. Die Bauern vermuten, dass das Düngemittel, das Bauer 2 verwendet hat, wirksamer ist. Das Signifikanzniveau beträgt 10%.*

1. *Wie lauten die Null- und Alternativhypothese?*
2. *Welchen Wert hat die Teststatistik näherungsweise?*
3. *In welchem Bereich wird* H_0 *nicht abgelehnt?*

Übung 6.9 *Der Betreiber einer Milchabfüllanlage will wissen, ob die Abfüllanlage die Sollfüllmenge von* 1000 *ml genau einhält. Dazu entnimmt er eine Stichprobe von* $n = 25$*. Bei dieser ergibt sich ein Mittelwert* $\bar{x}$ *von* $998,3$ *ml und eine Varianz* $var(x) = 4,2\ ml^2$*. Die Füllmenge sei normalverteilt und das Signifikanzniveau beträgt* 5%*.*

1. *Wie groß ist die Teststatistik?*

2. *In welchem Bereich der Teststatistik wird H_0 abgelehnt?*

3. *Wie interpretieren Sie Ihr Ergebnis?*

4. *Nehmen Sie an, dass $\bar{x}$ und $var(x)$ bei einer größeren Stichprobe sich nicht verändern. Bei welchem n würde man die Nullhypothese annehmen?*

Übung 6.10 *Sie möchten wissen, ob das Ergebnis der Bundestagswahlen Auswirkungen auf ein durchschnittliches Depot hat. Dazu haben Sie von einer Bank Werte von Kundenportfolios vor- und nach der Wahl erhalten. Der Wert des jeweiligen Portfolios vor der Wahl ist in der Variable `vor`, der zugehörige Wert nach der Wahl in der Variable `nach` kodiert:*

```
vor  <- c(356.63, 45983.89, 7279.29, 93245.76, 2854087.73)
nach <- c(585.45, 45683.76, 5876.44, 89345.57, 2540874.65)
```

Sie gehen davon aus, dass beide Zufallsvariablen normalverteilt sind.

1. *Berechnen Sie den Wert der Teststatistik.*

2. *Würden Sie die These, dass die Wahl auf den durchschnittlichen Depotwert Auswirkungen hat, bei einem Signifikanzniveau von 5% annehmen?*

Übung 6.11 *Zur Finanzierung Ihres Studiums geben Ihnen Ihre Eltern jeweils einen Teil ihres (schwankenden) Einkommens ab. Sowohl Ihre Ausgaben während des Studiums als auch die monatliche Unterstützung durch die Eltern können als unabhängige normalverteilte Zufallsvariablen angesehen werden. Die Varianz der Ausgaben beträgt $23.000€^2$ und die Varianz der Einnahmen $9.000€^2$. Unbekannt ist jeweils der Mittelwert. Folgende Daten stehen ihnen zur Verfügung:*

- *Ausgaben:* $X = (470, 500, 720, 600, 550)$
- *Einnahmen:* $Y = (580, 420, 660, 510, 610, 340, 520)$

1. *Mit welchem Test können Sie prüfen, ob die Ausgaben während des Studiums die Unterstüzung der Eltern signifikant übertreffen?*

2. *Wie lauten die zugehörigen Hypothesen?*

3. *Welche Kommandos verwenden Sie in R?*

4. *Bestimmen sie den Bereich der Teststatistik, in dem die Nullhypothese nicht abgelehnt wird (bei einem Signifikanzniveau $\alpha = 5\%$).*

Übung 6.12 *Die Länge von Knipps ist normalverteilt. Ihre Nullhypothese ist, die mittlere Länge von Knipps sei 100. Ihre Alternativhypothese ist, die mittlere Länge von Knipps sei ungleich 100. In Ihrer Stichprobe der Größe n messen Sie einen Mittelwert von $\bar{x}$ und eine Standardabweichung von $\hat{\sigma}_x$. Ihr Signifikanzniveau ist 5%. Sie berechnen eine Teststatistik*

$$g = \sqrt{n} \cdot \frac{\bar{x} - 100}{\hat{\sigma}_x}$$

Sei Q^t_n die Quantilsfunktion der t-Verteilung mit n Freiheitsgraden und Q^N die Quantilsfunktion der Normalverteilung. Wann lehnen Sie die Nullhypothese ab?

Übung 6.13 *Die Länge von Knorz ist normalverteilt. Ihre Nullhypothese ist, die mittlere Länge von Knorz sei kleiner oder gleich 12. Ihre Alternativhypothese ist, die mittlere Länge von Knorz sei größer 12. In Ihrer (eher kleinen) Stichprobe der Größe n messen Sie einen Mittelwert von $\bar{x}$. Für $\bar{x}$ schätzen sie eine Standardabweichung von $\hat{\sigma}_{\bar{x}}$. Sie berechnen eine Teststatistik*

$$g = \frac{12 - \bar{x}}{\hat{\sigma}_{\bar{x}}}$$

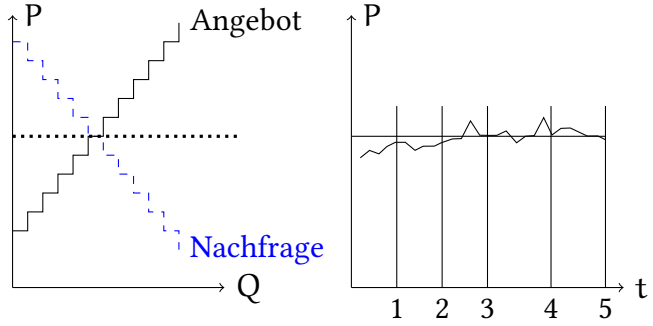
Sei Q^t_n die Quantilsfunktion der t-Verteilung mit n Freiheitsgraden und Q^N die Quantilsfunktion der Normalverteilung.

1. *Ihr Signifikanzniveau ist 10%. Wann lehnen die Nullhypothese ab?*
2. *Jetzt sei Ihr Signifikanzniveau 1%. Wann lehnen Sie die Nullhypothese ab?*
3. *Ihr Signifikanzniveau ist immer noch 1%, allerding ist Ihre Nullhypothese nun, die mittlere Länge von Knorz sei gleich 12. Ihre Alternativhypothese ist nun, die mittlere Länge von Knorz sei ungleich 12. Wann lehnen Sie die Nullhypothese ab?*

7. Konfidenzintervalle

7.1. Motivation: Effizienz von Märkten

Erster Hauptsatz der Wohlfahrtstheorie:
Jedes Walras-Gleichgewicht ist Pareto-Effizient



Die Grafik zeigt Daten eines Experiments von
Vernon Smith, Journal of Political Economy, 1962.

Der Preis, dem im Marktexperiment beobachtet wird, wird immer ein wenig vom Gleichgewichtspreis abweichen. Wie kann man beurteilen, wie »genau« der gemessene Preis ist, und ob z.B. der Gleichgewichtspreis mit dem im Experiment beobachteten Preis vereinbar ist?

Als Maß für die Genauigkeit unserer Schätzung berechnen wir ein *Konfidenzintervall.*

7.2. Schätzung von Konfidenzintervallen bei bekannter Varianz

In Kapitel 3 haben wir uns mit Schätzern beschäftigt. Wenn wir etwas geschätzt haben, wollen wir wissen, wie genau unsere Schätzung eigentlich ist. In Kapitel 4 haben wir dazu das "credible interval" eingeführt. Wir berechnen dieses Intervall mit Hilfe der Stichprobe. Im Intervall liegt dann mit einer vorgegebenen Wahrscheinlichkeit, z.B. 95%, der Parameter der Population. Allerdings brauchen wir Bayesianische Verfahren (d.h. normalerweise einen Computer), um das "credible interval" berechnen zu können.

Was haben Leute vor 1953 gemacht? Können wir die frequentistischen Verfahren aus Kapitel 5 und Kapitel 6 verwenden, um etwas ähnliches, wie ein "credible interval" zu berechnen?

Wir wollen wissen, in welchen Bereich um den geschätzten Parameter $\hat{\theta}$ der wahre Parameter θ liegen könnte.

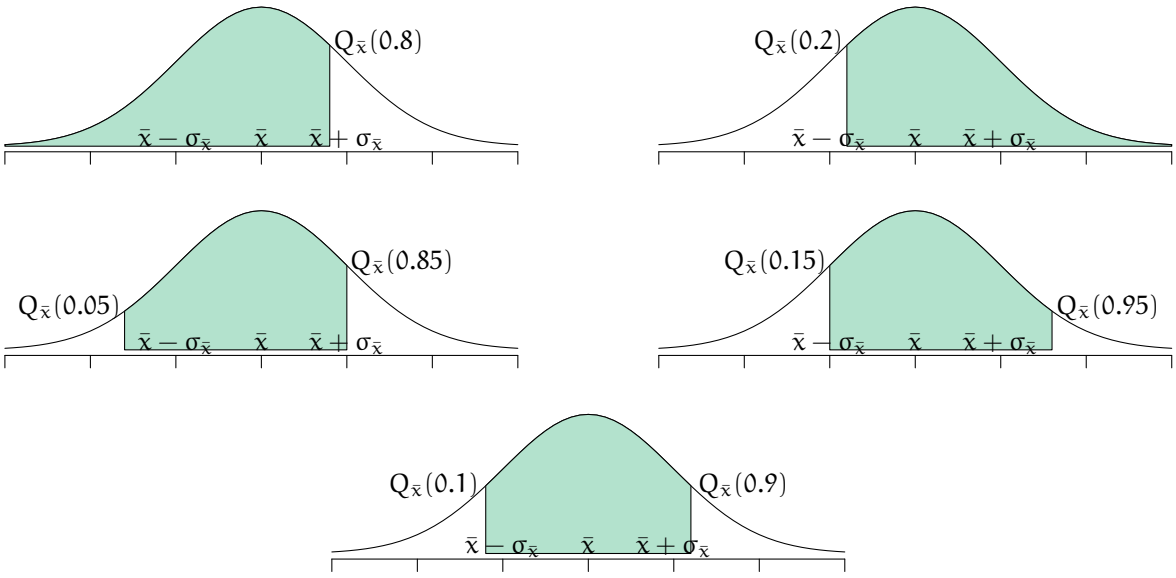
Jerzy Neyman (1937): Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. *Philosophical Transactions of the Royal Society of London. Series A.*

- Population: $X \sim F(x|\theta)$
- Stichprobe: $X_1, \ldots, X_n$
  - Aus der Stichprobe schätzen wir den Populationsparameter $\hat{\theta}$
  - Nun: Finde ein Intervall $[\underline{V}, \bar{V}]$ das den (unbekannten) wahren Parameter θ zu einem vorgegebenen Konfidenzniveau $P = 1 - \alpha$ enthält.
  - Erinnerung: Bei Punktschätzungen war der Schätzer $\hat{\theta}$ eine Zufallsvariable.
  - Genauso bei Intervallschätzungen: $\underline{V}(X_1, \ldots, X_n)$ und $\bar{V}(X_1, \ldots, X_n)$ sind Stichprobenfunktionen, und damit Zufallsvariablen.

Illustration Konfidenzintervall für eine Zufallsvariable mit bekannter Verteilung:

Sei X z.B. normalverteilt $X \sim N(\mu, \sigma^2)$. Suche nach einem Intervall, das X mit Wahrscheinlichkeit z.B. 80% enthält.

Problem: Es gibt etliche solcher Intervalle:

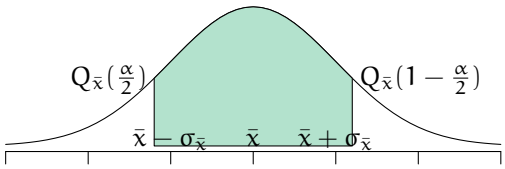


Sei $Q()$ die Quantilsfunktion, dann haben alle Intervalle $[Q(x), Q(x+0.8)]$ für $x \in [0, 0.2]$ die gewünschte »Breite«.

Das Intervall $[Q(0.1), Q(0.9)]$ hat zwei Vorzüge:

- Es ist symmetrisch (rechts und links fehlt jeweils 0.1)
- es hat die kleinste Breite (Eine kurze Erklärung finden Sie im Anhang 7.C zu diesem Kapitel)

Allgemein werden wir symmetrische Intervalle $[Q(\alpha/2), Q(1-\alpha/2)]$ betrachten.



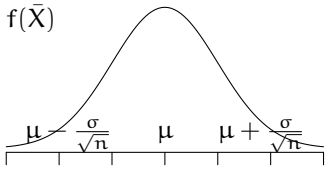
Warum interessieren wir uns überhaupt für Konfidenzintervalle *normalverteilter* Schätzer? Erinnerung: Zentraler Grenzwertsatz

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} X_i}{n} \sim N\left(\mu, \frac{\sigma_X^2}{n}\right)$$

- Unsere Schätzer sind (oftmals) auch normalverteilt (oder wenigstens annähernd normalverteilt) — falls unsere Stichprobe sehr groß ist.
- Unser Schätzer für den Mittelwert ist außerdem normalverteilt, falls die X_i ihrerseits normalverteilt sind — diese Annahme ist allerdings nicht immer gerechtfertigt.

Terminologie Allerdings müssen wir aufpassen — zwar kennen wir den Schätzer $\hat{\theta}$ (oder $\hat{\mu}$) und den wahren Wert θ (oder μ) kennen wir nicht. Dennoch ist (in der frequentistischen Schule) der wahre Wert θ (oder μ) keine Zufallsvariable.

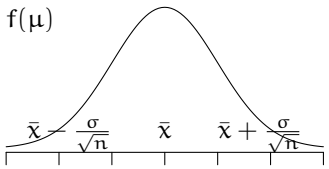
Gedankenexperiment: Wie wäre es, wenn wir das wahre μ kennen würden?



Technisch bedeutet das

- $X \sim N(\mu, \sigma_X^2)$
- $\bar{X} \sim N\left(\mu, \frac{\sigma_X^2}{n}\right)$
- $\bar{X} - \mu \sim N\left(0, \frac{\sigma_X^2}{n}\right)$
- $\frac{\bar{X}-\mu}{\sigma_X/\sqrt{n}} \sim N(0,1)$

Gedankenexperiment 2: unendlich viele Stichproben



Jetzt könnte man auf die Idee kommen, bei bekanntem $\bar{x}$ den Zusammenhang links wie folgt zu schreiben:

- $\mu \sim N\left(\bar{x}, \frac{\sigma_X^2}{n}\right)$

  $\rightarrow$ (frequentistisch) nicht sinnvoll!

Beachte: In der frequentistischen Schule ist μ keine Zufallsvariable, folgt also keiner Verteilung!

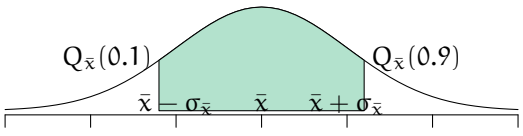
- *Richtig*: Ein 95% Konfidenzintervall für μ_X ist das Intervall das, gegeben diese Stichprobe, in etwa 95% aller wiederholter Stichproben den wahren Wert von μ_X enthält.
- *Auch richtig*: Ein 95% Konfidenzintervall für μ_X ist das Intervall das, gegeben diese Stichprobe, den wahren Wert μ_X bei einem *Konfidenzniveau* von 95% enthält.
- *In der frequentistischen Welt falsch*: ~~Das Konfidenzintervall, das aufgrund dieser Stichprobe ermittelt wurde, enthält den wahren Parameter μ mit Wahrscheinlichkeit 95%~~.

  Der wahre Parameter ist in der frequentistischen Welt keine Zufallsvariable, also können wir über ihn keine Wahrscheinlichkeitsaussage machen.

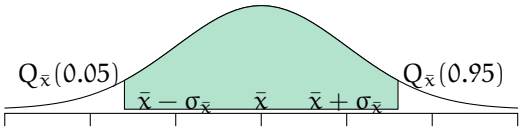
Dogmengeschichte: Die Idee des Konfidenzintervalls wurde 1930 von Jerzy Neyman vorgeschlagen. Die Interpretation ist kompliziert. Das "credible interval" aus der Bayesianischen Statistik hat eine einfachere Interpretation.

Konfidenzintervalle unterschiedlicher Breite

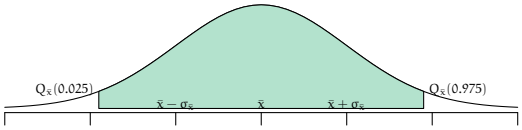
80% Konfidenzintervall ($\alpha = 20\%$):



90% Konfidenzintervall ($\alpha = 10\%$):



95% Konfidenzintervall ($\alpha = 5\%$):



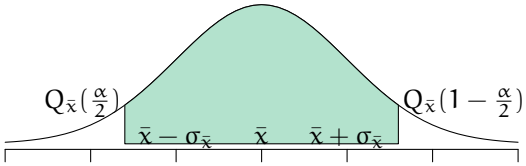
In der Praxis werden üblicherweise 95% Konfidenzintervalle verwendet.

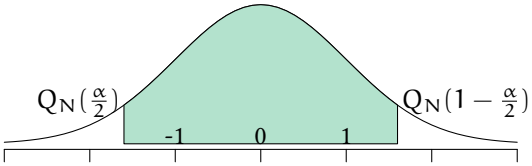
Je breiter unsere Konfidenzintervalle, um so häufiger wird der wahre Parameter μ in unseren Konfidenzintervallen liegen.

Allerdings ist ein »breites« Intervall nicht gerade »genau«.

Wie bestimmen wir ein Konfidenzintervall bei einem Konfidenzniveau P um einen gegebenen Mittelwert $\bar{x}$: (Annahme: Der Mittelwert $\bar{x}$ ist (annähernd) normalverteilt, z.B. weil die Stichprobe sehr groß ist)

Gegeben ein gewünschtes Konfidenzniveau P (z.B. 95%), der Stichprobenmittelwert $\bar{x}$, und entweder die Standardabweichung des Stichprobenmittelwerts $\sigma_{\bar{x}}$ oder die Standardabweichung einer Stichprobenbeobachtung σ_X.

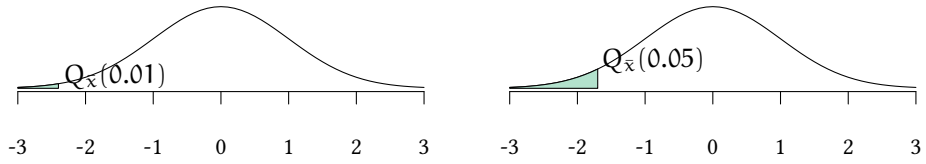




- Bestimme die Fläche unter dem linken/rechten Rand der Verteilung $\frac{\alpha}{2} = \frac{1-P}{2}$
- Bestimme das Quantil der Standardnormalverteilung $Q_N\left(\frac{\alpha}{2}\right)$ oder $Q_N\left(1-\frac{\alpha}{2}\right)$
- Falls nötig, bestimme $\sigma_{\bar{x}} = \frac{\sigma_X}{\sqrt{n}}$
- Das Intervall ist

$$\left[\bar{x}+\sigma_{\bar{x}}\cdot Q_N\left(\frac{\alpha}{2}\right), \bar{x}-\sigma_{\bar{x}}\cdot Q_N\left(\frac{\alpha}{2}\right)\right] = \left[\bar{x}-\sigma_{\bar{x}}\cdot Q_N\left(1-\frac{\alpha}{2}\right), \bar{x}+\sigma_{\bar{x}}\cdot Q_N\left(1-\frac{\alpha}{2}\right)\right]$$

Um diese Intervalle auszurechnen, benötigt man also $\bar{x}$, $\sigma_{\bar{x}}$, sowie $Q_N\left(\frac{\alpha}{2}\right)$. Den Werte von $\bar{x}$ können wir aus unserer Stichprobe ausrechnen. Über $\sigma_{\bar{x}}$ reden wir unten in Abschnitt 7.3. Für $Q_N\left(\frac{\alpha}{2}\right)$ hat man früher Quantilstabellen benutzt. Hier ist eine solche Tabelle:



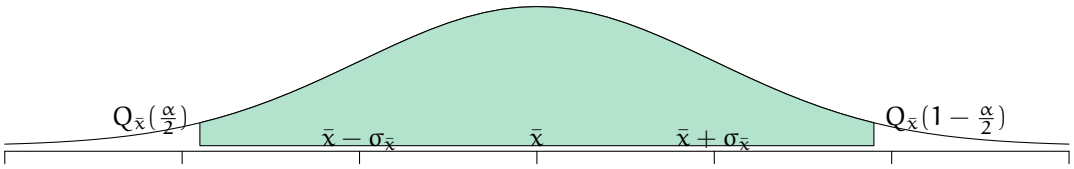
Quantile der Normalverteilung:

x	Q(x)
0.0001	-3.719
0.00025	-3.481
0.0005	-3.291
0.001	-3.090
0.0025	-2.807
0.005	-2.576
0.01	-2.326
0.025	-1.960
0.05	-1.645

in R verwenden wir die Funktion `qnorm`:

```
qnorm(.0001)

[1] -3.719016
```

Länge des Konfidenzintervalls Es wäre schön, wenn unser Konfidenzintervall, also der Bereich, in dem der gesuchte Parameter vermutlich liegt, möglichst klein wären. Welche Faktoren bestimmen die Länge dieses Intervalls?



$$\begin{aligned}
\bar{x}+\sigma_{\bar{x}}\cdot Q_N\left(1-\frac{\alpha}{2}\right)-\left(\bar{x}+\sigma_{\bar{x}}\cdot Q_N\left(\frac{\alpha}{2}\right)\right) &= \\
\bar{x}-\sigma_{\bar{x}}\cdot Q_N\left(\frac{\alpha}{2}\right)-\left(\bar{x}+\sigma_{\bar{x}}\cdot Q_N\left(\frac{\alpha}{2}\right)\right) &= \\
-\sigma_{\bar{x}}\cdot Q_N\left(\frac{\alpha}{2}\right)-\left(\quad\sigma_{\bar{x}}\cdot Q_N\left(\frac{\alpha}{2}\right)\right) &= \\
-2\sigma_{\bar{x}}\cdot Q_N\left(\frac{\alpha}{2}\right)=2\sigma_{\bar{x}}\cdot Q_N\left(1-\frac{\alpha}{2}\right)&
\end{aligned}$$

Das Intervall wird kleiner, wenn…

- $\sigma_{\bar{x}}$ kleiner wird
- n größer wird (weil $\sigma_{\bar{x}} = \frac{\sigma_X}{\sqrt{n}}$)
- α größer wird, bzw. P kleiner wird (weil $\frac{\alpha}{2} = \frac{1-P}{2}$)

7.3. Schätzen von Konfidenzintervallen bei unbekannter Varianz

Bislang haben wir angenommen, dass der Standardfehler $\sigma_{\bar{x}}$ bekannt ist (oder exakt bestimmt werden kann). Das ist normalerweise nicht der Fall. Wir können aber ähnlich mit der geschätzten Varianz $\hat{\sigma}_X$ rechnen:

Für das Marktexperiment von Vernon Smith stellt sich aber das Problem, dass die Standardabweichung der Population σ_X unbekannt ist. Wir kennen nur die Standardabweichung des Samples.

Zur Erinnerung: $\frac{\bar{X}-\mu}{\sigma_{\bar{X}}} = \frac{\bar{X}-\mu}{\sigma_X/\sqrt{n}} \sim N(0,1)$

Bislang benötigen wir also $\sigma_{\bar{X}}$ oder σ_X. Wenn wir weder σ_X noch $\sigma_{\bar{X}}$ kennen, können wir statt σ_X die geschätzte Standardabweichung

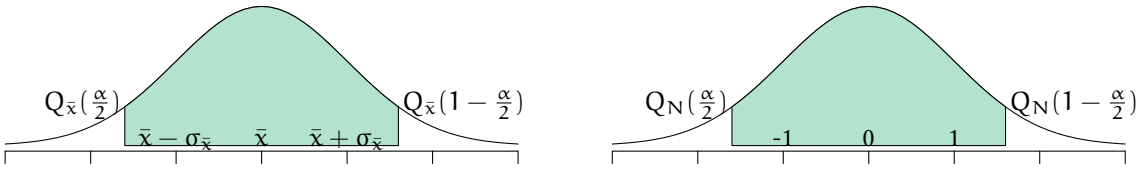
$$\hat{\sigma}_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

verwenden.

bekannte Varianz (bekannte Standardabweichung)	unbekannte Varianz (unbekannte Standardabweichung)
σ_X	$\hat{\sigma}_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$	$\hat{\sigma}_{\bar{X}} = \frac{\hat{\sigma}_X}{\sqrt{n}}$
Falls X normalverteilt ist, dann gilt	
$\frac{\bar{X}-\mu}{\sigma_{\bar{X}}} = \frac{\bar{X}-\mu}{\sigma_X/\sqrt{n}} \sim N(0,1)$	$\frac{\bar{X}-\mu}{\hat{\sigma}_{\bar{X}}} = \frac{\bar{X}-\mu}{\hat{\sigma}_X/\sqrt{n}} \sim t_{n-1}$
	Falls n groß ist, dann gilt approximativ
	$\frac{\bar{X}-\mu}{\hat{\sigma}_{\bar{X}}} = \frac{\bar{X}-\mu}{\hat{\sigma}_X/\sqrt{n}} \sim N(0,1)$

Dabei steht t_{n-1} für die t-Verteilung mit $n-1$ Freiheitsgraden.

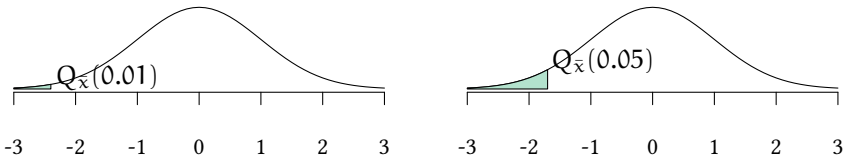
Wie bestimmen wir ein Konfidenzintervall zu einem gegebenen Konfidenzniveau P um einen gegebenen Mittelwert $\bar{x}$ bei geschätzter Varianz $\hat{\sigma}$:



- Bestimme die Fläche unter dem linken/rechten Rand der Verteilung $\frac{\alpha}{2} = \frac{1-P}{2}$
- Freiheitsgrade = $n - 1$
- Bestimme das Quantil der t-Verteilung $Q_t\left(\frac{\alpha}{2}\right)$ oder $Q_t\left(1-\frac{\alpha}{2}\right)$
- Bestimme (falls nötig) $\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}_x}{\sqrt{n}}$
- Das Intervall ist

$$\left[\bar{x} + \hat{\sigma}_{\bar{x}} \cdot Q_t\left(\frac{\alpha}{2}\right), \bar{x} - \hat{\sigma}_{\bar{x}} \cdot Q_t\left(\frac{\alpha}{2}\right)\right] = \left[\bar{x} - \hat{\sigma}_{\bar{x}} \cdot Q_t\left(1-\frac{\alpha}{2}\right), \bar{x} + \hat{\sigma}_{\bar{x}} \cdot Q_t\left(1-\frac{\alpha}{2}\right)\right]$$

Für $Q_t\left(\frac{\alpha}{2}\right)$ hat man früher Quantilstabellen benutzt. Hier ist eine Ausschnitt einer solchen Tabelle:



Quantile der t-Verteilung:

	0.001	0.0025	0.005	0.01	0.025	0.05
$Q(x,1)$	-318.309	-127.321	-63.657	-31.821	-12.706	-6.314
$Q(x,2)$	-22.327	-14.089	-9.925	-6.965	-4.303	-2.920
$Q(x,3)$	-10.215	-7.453	-5.841	-4.541	-3.182	-2.353
$Q(x,5)$	-5.893	-4.773	-4.032	-3.365	-2.571	-2.015
$Q(x,10)$	-4.144	-3.581	-3.169	-2.764	-2.228	-1.812
$Q(x,50)$	-3.261	-2.937	-2.678	-2.403	-2.009	-1.676

in R verwenden wir die Funktion `qt`:

```
qt(.01,5)
```

```
[1] -3.36493
```

7.4. Zusammenhang Signifikanzniveau / Konfidenzintervall

Test bei einem gegebenem Signifikanzniveau α:

Bei einem Konfidenzniveau von $P = 1 - \alpha$ ist das Konfidenzintervall

$$\begin{aligned}
\bar{x} + \hat{\sigma}_{\bar{x}} \cdot Q_t\left(\frac{\alpha}{2}\right) \leq \quad & \mu_0 \quad && \leq \bar{x} + \hat{\sigma}_{\bar{x}} \cdot Q_t\left(1-\frac{\alpha}{2}\right)\\
\hat{\sigma}_{\bar{x}} \cdot Q_t\left(\frac{\alpha}{2}\right) \leq \quad & \mu_0 - \bar{x} && \leq \hat{\sigma}_{\bar{x}} \cdot Q_t\left(1-\frac{\alpha}{2}\right)\\
Q_t\left(\frac{\alpha}{2}\right) \leq \quad & \tfrac{\mu_0-\bar{x}}{\hat{\sigma}_X/\sqrt{n}} && \leq Q_t\left(1-\frac{\alpha}{2}\right)\\
Q_t\left(1-\frac{\alpha}{2}\right) = -Q_t\left(\frac{\alpha}{2}\right) \geq \quad & \tfrac{\bar{x}-\mu_0}{\hat{\sigma}_X/\sqrt{n}} && \geq -Q_t\left(1-\frac{\alpha}{2}\right) = Q_t\left(\frac{\alpha}{2}\right)\\
Q_t\left(\frac{\alpha}{2}\right) \leq \quad & \frac{\bar{x}-\mu_0}{\hat{\sigma}_X/\sqrt{n}} && \leq Q_t\left(1-\frac{\alpha}{2}\right)
\end{aligned}$$

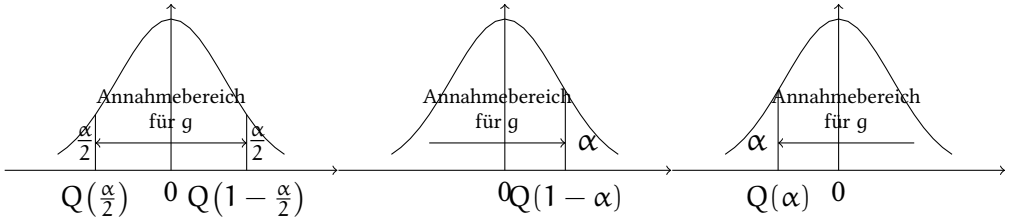
Zweiseitiger Test ($H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$): Seien $Q(\frac{\alpha}{2})$ und $Q(1-\frac{\alpha}{2})$ die Quantile der Normalverteilung, dann wird H_0 nicht abgelehnt, wenn

$$Q_t\left(\frac{\alpha}{2}\right) \leq \frac{\bar{x}-\mu_0}{\hat{\sigma}_X/\sqrt{n}} \leq Q_t\left(1-\frac{\alpha}{2}\right)$$

Genau dann wenn der hypothetische Wert von μ_0 nicht innerhalb des Konfidenzintervalls (mit gegebenem α) liegt, dann verwirft auch der Test H_0

Signifikanztest: Teststatistik $g = \frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}}$

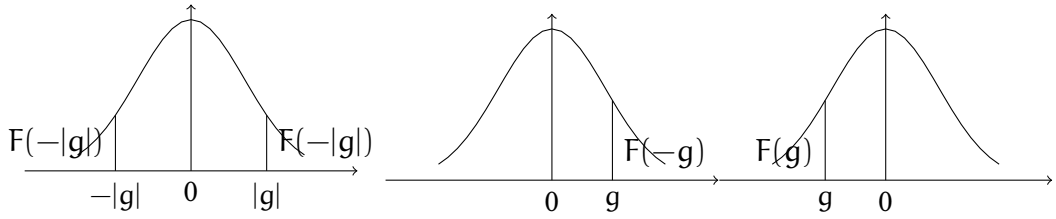
zweiseitig ($H_1 : \mu \neq \mu_0$) einseitig ($H_1 : \mu > \mu_0$) einseitig ($H_1 : \mu < \mu_0$)



H_0 wird abgelehnt, falls g nicht im Annahmebereich liegt.

p-Wert: Teststatistik $g = \frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}}$

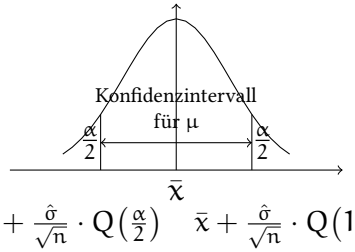
zweiseitig ($H_1 : \mu \neq \mu_0$) einseitig ($H_1 : \mu > \mu_0$) einseitig ($H_1 : \mu < \mu_0$)



$p = 2 \cdot F(-|g|)$ $p = F(-g)$ $p = F(g)$

H_0 wird abgelehnt falls $p < \alpha$

Konfidenzintervall:



$H_0 : \mu = \mu_0$ wird abgelehnt, falls μ_0 nicht im Konfidenzintervall liegt.

7.5. Simulation

Wir beginnen mit einer Verteilung, die wir kennen: Hier ziehen wir 6 Beobachtungen von einer Normalverteilung mit Mittelwert 5:

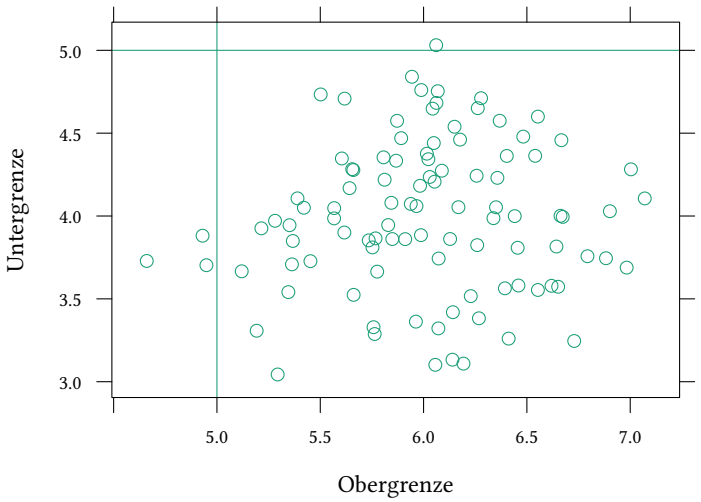
```
rnorm(6,mean=5)

[1] 4.439524 4.769823 6.558708 5.070508 5.129288 6.715065
```

```
t.test(rnorm(6,mean=5))$conf.int

[1] 3.992026 5.890386
attr(,"conf.level")
[1] 0.95
```

Das können wir auch gleich ein paar Mal machen:

```
replicate(5,t.test(rnorm(6,mean=5))$conf.int)

         [,1]     [,2]     [,3]     [,4]     [,5]
[1,] 3.736746 3.840093 3.594711 5.088928 4.378326
[2,] 6.354507 5.222530 6.003657 6.049363 5.255990
```

Hier sind die Grenzen für 100 solcher Konfidenzintervalle:



Wir sehen, in 96 Fällen enthält das Intervall wirklich die 5, in 4 Fällen klappt es aber nicht, das Konfidenzintervall enthält nicht den wahren Wert.

7.6. Literatur

- Dolić, Statistik mit R, Kapitel 6.2.2.
- Hartung, Statistik, Kapitel III.2., IV.1.3.
- Schira, Statistische Methoden der VWL und BWL-Theorie und Praxis, Kapitel 14.
- Verzani, Using R for Introductory Statistics, Chapter 7.

7.7. Schlüsselbegriffe

- Konfidenzniveau
- Konfidenzintervall
- Quantile
- t-Verteilung

- normalverteilte Zufallsvariablen → parametrische Tests (exakt)
- große Stichprobe/nicht normalverteilte Zufallsvariablen → parametrische Tests (asymptotisch)
  - bekannte Varianz → $(\bar{X}-\mu)/\sigma_{\bar{X}}$ ist normalverteilt
  - unbekannte Varianz → $(\bar{X}-\mu)/\hat{\sigma}_{\bar{X}}$ ist t-verteilt

Anhang 7.A Beispiele für die Vorlesung

In Ihrer Stichprobe mit 4 Beobachtungen einer normalverteilten Zufallsvariablen mit unbekannter Varianz ermitteln Sie einen Mittelwert von 10 und eine Varianz von 100. Ihr Konfidenzniveau ist 95%. Wie bestimmen Sie die Breite des Konfidenzintervalls für den Mittelwert?

- Keine der folgenden Antworten ist richtig.
- `qt(10,4)*0.95`
- `qnorm(10)*5`
- `qt(.95,3)*10`
- `qnorm(.95)*10`

In Ihrer Stichprobe mit 25 Beobachtungen einer normalverteilten Zufallsvariablen mit bekannter Varianz 25 bestimmen Sie einen Mittelwert von 10. Ihr Konfidenzniveau ist 95%. Wie bestimmen Sie die Untergrenze des Konfidenzintervalls für den Mittelwert?

- Keine der folgenden Antworten ist richtig.
- `10 + qnorm(0.05)`
- `10 + 5 * qnorm(0.05)`
- `10 + 5 * qnorm(0.025)`
- `10 - qnorm(0.975)`

Anhang 7.B Übungen

Übung 7.1 *Eine Stichprobe ergibt* $X = \{1, 2, 3\}$.

- *Nehmen Sie an,* X *sei normalverteilt und verwenden Sie die* t*-Verteilung um ein 95% Konfidenzintervall für den Mittelwert zu bestimmen.*
- *Bestimmen Sie ein 99% Konfidenzintervall.*
- *Bestimmen Sie ein 95% credible interval.*
- *Bestimmen Sie ein 99% credible interval.*

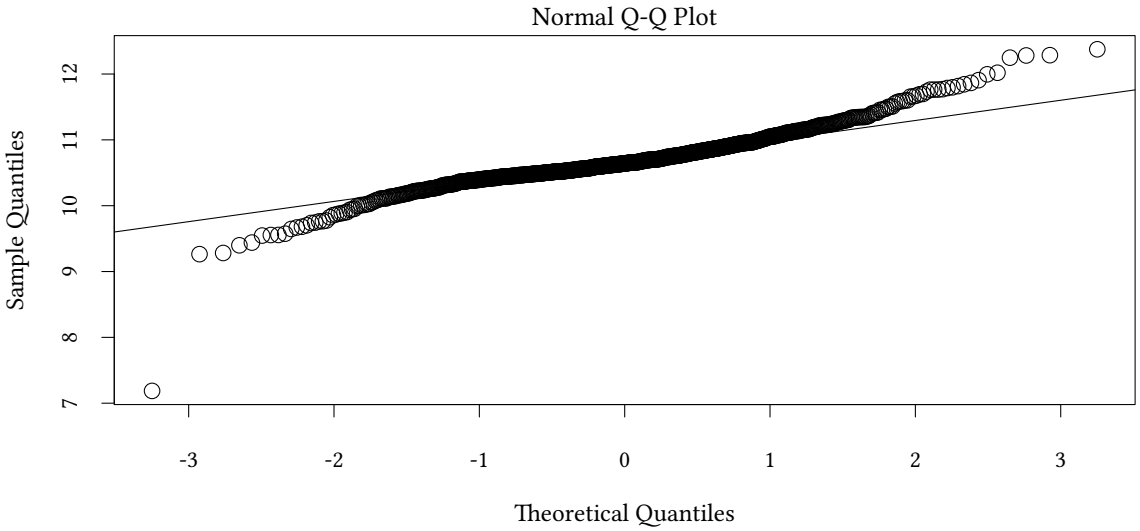
`t.test` kann einen t-Test durchführen (dazu kommen wir später), berechnet aber auch Konfidenzintervalle.

Übung 7.2 *Betrachten Sie den Datensatz* `Participation` *aus der Bibliothek* `Ecdat`.

- *Verwenden Sie die* t*-Verteilung, um ein Konfidenzintervall für den Mittelwert der durchschnittlichen Logarithmus des Nichtarbeitseinkommens* `lnlinc` *zu ermitteln.*
- *Ist diese Verteilungsannahme sinnvoll?*
- *Bestimmen Sie ein credible interval.*

Der t-Test nimmt an, dass unsere Variable X normalverteilt ist. Wie können wir diese Annahme überprüfen? Ein Q-Q Plot (siehe Abschnitt A.6.4) erlaubt uns die Verteilung einer Stichprobe mit einer theoretischen Verteilung (hier der Normalverteilung) zu vergleichen. An der horizontalen Achse sehen wir die theoretischen Quantile, auf der vertikalen Achse die Quantile des Samples. Wenn die Stichproblenverteilung nur eine lineare Transformation der Standardnormalverteilung ist, dann liegen im Idealfall alle Punkte auf einer Geraden.

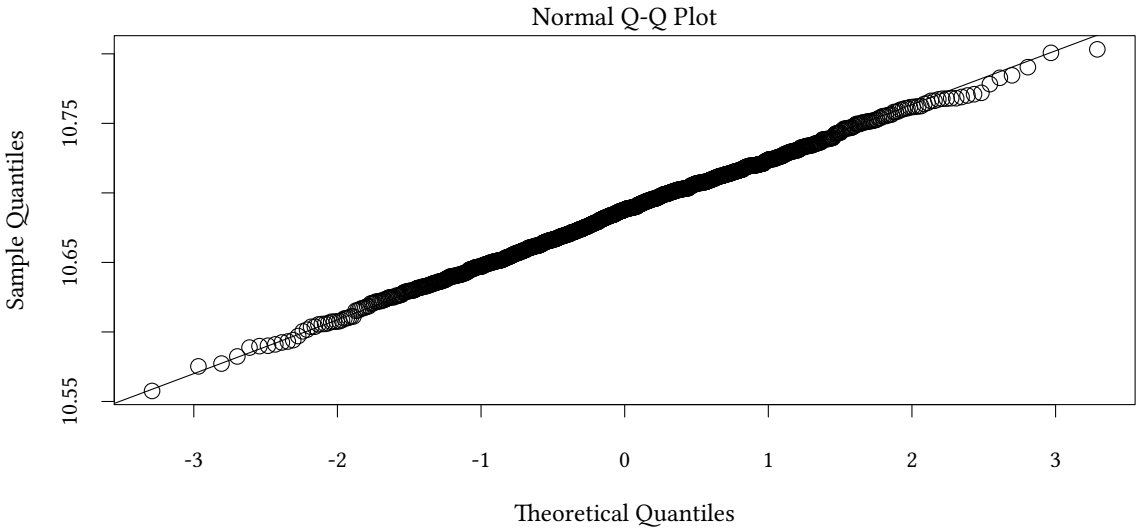
```
qqnorm(lnnlinc)
qqline(lnnlinc)
```



Wir sehen, von einer Normalverteilung ist unsere Stichprobe noch weit entfernt. Die Annahmen des t-Tests sind also nicht erfüllt.

Viel besser sieht es aber für die Mittelwerte aus. Hier ist der Plot für 1000 Mittelwerte von Stichproben aus dieser Verteilung:

```
xx<-replicate(1000,mean(sample(lnnlinc,100)))
qqnorm(xx)
qqline(xx)
```



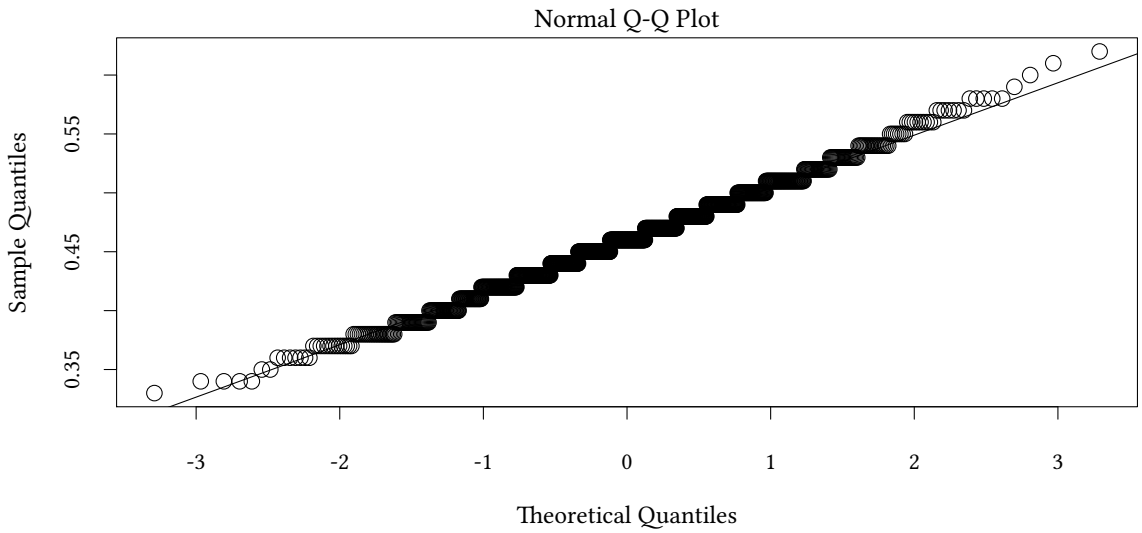
Übung 7.3 *Betrachten Sie den Datensatz `Participation` aus der Bibliothek `Ecdat`.*

- *Verwenden Sie die* t*-Verteilung, um ein Konfidenzintervall für den Mittelwert der durchschnittlichen Beteiligung am Arbeitsmarkt `lfp` zu ermitteln.*

- *Ist diese Verteilungsannahme sinnvoll?*
- *Bestimmen Sie ein credible interval.*

Hier ist wieder der Plot für Mittelwerte:

```
xx<-replicate(1000,mean(sample(lfp,100)=="yes"))
qqnorm(xx)
qqline(xx)
```

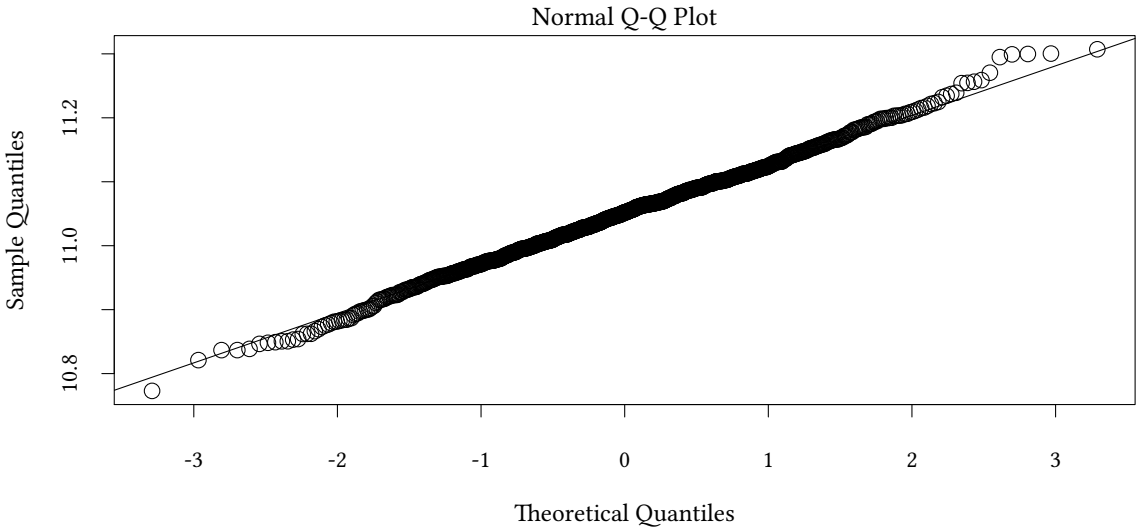


Übung 7.4 *Betrachten Sie den Datensatz `Bwages` aus der Bibliothek `Ecdat`.*

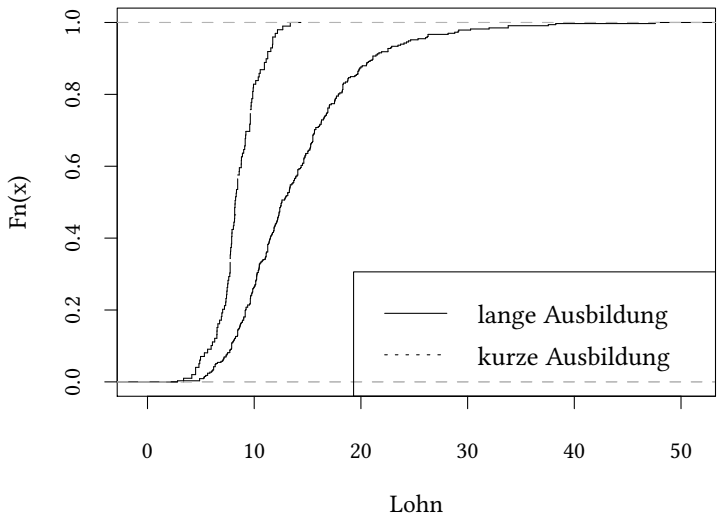
- *Verwenden Sie die t-Verteilung, um ein Konfidenzintervall für den Mittelwert des Brutto-Stundenlohns `wage` zu ermitteln.*
- *Bestimmen Sie ein credible interval.*
- *Ermitteln Sie nun ein Konfidenzintervall für Arbeiter mit geringer Ausbildung `educ==1` und mit hoher Ausbildung `educ==5`.*
- *Vergleichen Sie die Verteilung der Löhne auch graphisch.*

Hier ist wieder der Plot für Mittelwerte:

```
xx<-replicate(1000,mean(sample(wage,1000)))
qqnorm(xx)
qqline(xx)
```




```
plot(ecdf(wage[educ==5]),do.points=FALSE,verticals=TRUE,xlab="Lohn",main="")
plot(ecdf(wage[educ==1]),do.points=FALSE,verticals=TRUE,lty='dashed',add=TRUE)
legend("bottomright",c("lange Ausbildung","kurze Ausbildung"),lty=c(1,3))
```



Übung 7.5 *Eine Maschine wickelt Klopapier auf Rollen, auf denen es verkauft wird. Die Länge des Papiers, das auf eine Rolle gewickelt wird ist mit Mittlewert* μ *und Varianz 0,36* m^2 *normalverteilt. Sie stellen fest, dass die Maschine auf die letzten 25 Rollen insgesamt 995 m Klopapier aufgewickelt hat.*

1. *Wie lautet die beste erwartungstreue Schätzung für den Erwartungswert* μ *der Normalverteilung?*

2. *Bestimmen Sie das 99%-Konfidenzintervall für* μ.

3. *Wie viele Rollen müssten wir abwickeln und exakt nachmessen, um ein 99%-Konfidenzintervall für* μ *aufstellen zu können, das nicht breiter als 0,1 m ist?*

Übung 7.6 *Eine Meinungsumfrage wird jedes Jahr zum gleichen Thema durchgeführt. In den vergangenen Jahren antworteten immer 50 - 70% der Befragten mit Ja, der Rest mit Nein.*

1. *Zur Vorbereitung der diesjährigen Umfrage soll zunächst geklärt werden, wie viele Personen mindestens befragt werden müssen, wenn das Konfidenzintervall für die mittlere Zustimmungsquote auf einem Konfidenzniveau von 99% basieren soll und nicht breiter sein soll als* $\pm 2\%$*. Gehen Sie davon aus, dass wir die Verteilung der empirischen mittleren Zustimmungsquote durch die Normalverteilung approximieren können. Wieviele Teilnehmer braucht man im ungünstigsten Fall?*

2. *Von den Befragten haben leider nur 1300 eine Angabe gemacht. 923 antworteten mit Ja, 377 mit Nein. Bestimmen Sie daraus das 99% Konfidenzintervall für den Anteil der Zustimmung.*

3. *Wie ändert sich das Ergebnis wenn man als Konfidenzniveau statt 99% nun 95% annimmt?*

Übung 7.7 *Für eine Druckerpatrone sollen Aussagen über den Erwartungswert und die Standardabweichung der damit bedruckbaren Blätter (in Fließtext) getroffen werden.*

Bei einer Stichprobe mit 36 solcher Patronen ergab sich, dass durchschnittlich 175 Blätter mit einer Patrone bedruckt werden konnten, wobei die Standardabweichung 17 Blätter und somit die Varianz 289 betrug. Man geht davon aus, dass die Druckkapazität einer Patrone normalverteilt ist.

1. *Welches Konfidenzintervall gilt für den Erwartungswert?*

2. *Ihr Chef findet, die Konfidenzintervalle seien zu groß. Was können Sie tun, um kleinere Intervalle zu erhalten?*

Übung 7.8 *Die Lebensdauer von Batterien ist normalverteilt. Die Standardabweichung der Lebensdauer einer Batterie beträgt 30 Stunden. 100 der Produktion zufällig entnommenen Batterien haben eine Gesamtlebensdauer von 6935,75 Stunden.*

1. *Was ist das 90% Konfidenzintervall für den Mittelwert der Lebensdauer. Verwenden Sie dabei die folgenden Quantile:*

	0.001	*0.0025*	*0.005*	*0.01*	*0.025*	*0.05*	*0.1*
$Q^N(x)$	*-3.090*	*-2.807*	*-2.576*	*-2.326*	*-1.960*	*-1.645*	*-1.282*
$Q^t_{19}(x)$	*-3.579*	*-3.174*	*-2.861*	*-2.539*	*-2.093*	*-1.729*	*-1.328*
$Q^t_{20}(x)$	*-3.552*	*-3.153*	*-2.845*	*-2.528*	*-2.086*	*-1.725*	*-1.325*
$Q^t_{21}(x)$	*-3.527*	*-3.135*	*-2.831*	*-2.518*	*-2.080*	*-1.721*	*-1.323*

2. *Wie groß muss die Stichprobe gewählt werden, damit die Länge des Konfidenzintervalls höchstens 5 (Stunden) beträgt?*

3. *Nun ermitteln Sie auch die Standardabweichung der Lebensdauer aus einer Stichprobe vom Umfang 20: (60.5, 80, 71, 73.7, 65, 68, 64.4, 62.9, 74, 78, 72.9, 74, 67.5, 72.8, 61.9, 71, 58, 61, 72.8, 73). Bestimmen Sie das 90% Konfidenzintervall.*

Übung 7.9 *Die Füllmenge einer Packung Milch sei normalverteilt. In einer Milchabfüllanlage will man herausfinden, wieviel Milch in einer Packung enthalten ist. Dazu wird eine Stichprobe vom Umfang* $n = 10$ *entnommen. Dabei ergaben sich die folgenden Werte (in ml):* $(999, 1000, 997, 1005, 1001, 1000, 998, 999, 1000, 1000)$

1. *Bestimmen Sie mit R ein* 99%*-Konfidenzintervall für die durchschnittliche Füllmenge.*

2. *Welche der folgenden Aussagen ist wahr?*
   a) *Zu* 99% *liegt der wahre Wert im gefundenen Interval.*
   b) *Wenn man dieses Verfahren immer wieder anwendet, wird man in 99% aller Fälle ein Intervall erhalten, das den wahren Wert enthält.*
   c) *Das Intervall beträgt* $[995,6204; 1004,3796]$.
   d) *Die Länge des* 95%*-Konfidenzintervalls ist größer.*
   e) *Die Länge des* 95%*-Konfidenzintervalls ist kleiner.*

Übung 7.10 *Unsere Stichprobe von 87 Beobachtungen ergibt einen Mittelwert von 150. Wir wissen, dass der berechnete Mittelwert eine Standardabweichung von 10 hat. Was ist ein 99% Konfidenzintervall für den Mittelwert? Was ist ein 95% Konfidenzintervall für den Mittelwert?*

Übung 7.11 *Unsere Stichprobe von 100 Beobachtungen ergibt einen Mittelwert von 200. Wir wissen, dass jede Beobachtung eine Standardabweichung von 10 hat. Was ist ein 99% Konfidenzintervall für den Mittelwert? Was ist ein 95% Konfidenzintervall für den Mittelwert?*

Übung 7.12 *Die durchschnittliche Füllmenge eines Sacks Kartoffeln ist 50 kg, die Standardabweichung ist 2 kg. Nehmen Sie an, dass die Füllmenge normalverteilt ist. Sie kaufen 4 Sack Kartoffeln und bestimmen die durchschnittliche Füllmenge* $\bar{x}$ *dieser 4 Säcke. Schreiben Sie jeweils ein R Kommando auf, das die Antworten auf die folgenden Fragen berechnet.*

1. *Wie wahrscheinlich ist es, dass* $\bar{x} > 50$ *kg?*

2. *Wie wahrscheinlich ist es, dass* $\bar{x} > 52$ *kg?*

3. *Bestimmen Sie das Gewicht* x^**, so dass in 99% aller Fälle der Wert* $\bar{x}$ *über diesem Gewicht liegt.*

Übung 7.13 *Ihre Firma stellt Antriebsketten her. Die Länge der einzelnen Kettenglieder ist unabhängig voneinander normalverteilt. Die mittlere Länge eines Kettengliedes ist 10. Ingesamt haben 95% aller Kettenglieder eine Länge zwischen 9 und 11. Jede Kette ist so lang wie die Summe der einzelnen Glieder. Sie interessieren sich für Ketten, die aus 100 Kettengliedern bestehen.*

1. *Wie wahrscheinlich ist es, dass eine solche Kette eine Gesamtlänge von über 1000 hat (verwenden Sie, falls notwendig, einen R-Ausdruck, um die Länge zu bestimmen)?*

2. *Wie wahrscheinlich ist es, dass eine solche Kette eine Gesamtlänge zwischen 900 und 1100 hat (verwenden Sie, falls notwendig, einen R-Ausdruck, um die Länge zu bestimmen)?*

3. *Wie wahrscheinlich ist es, dass eine solche Kette eine Gesamtlänge zwischen 990 und 1010 hat (verwenden Sie, falls notwendig, einen R-Ausdruck, um die Länge zu bestimmen)?*

4. *Wie wahrscheinlich ist es, dass eine solche Kette eine Gesamtlänge zwischen 999 und 1001 hat (verwenden Sie, falls notwendig, einen R-Ausdruck, um die Länge zu bestimmen)?*

Übung 7.14 *Sie kaufen eine neue Maschine. Der Hersteller gibt an, dass man die Maschine im Mittel 10000 Stunden lang nutzen kann, bevor sie verschlissen ist. Die tatsächliche Nutzungsdauer der bisher verkauften Maschinen weicht durchschnittlich um 400 Stunden von diesem Wert ab. Sie nehmen an, dass die Nutzungsdauer der Maschine eine normalverteilte Zufallsgröße* X *ist.*

1. *Welche Varianz legen Sie zu Grunde?*

2. *Mit welcher Wahrscheinlichkeit können Sie eine Maschine länger als 11000 Stunden nutzen?*

3. *Mit welcher Wahrscheinlichkeit liegt die Nutzungsdauer einer Maschine zwischen 9900 und 11000 Stunden?*

4. *Wieviele Maschinen müssten Sie mindestens kaufen um mit 99-prozentiger Sicherheit eine zu haben, die man mindestens 10000 Stunden lang nutzen kann?*

Übung 7.15 *Ein selbsternannter Börsenguru beobachtet die Entwicklungen der Aktienkurse (A, B und C) dreier Firmen. Als ihn eine Börsenzeitschrift um eine Kursprognose für Ende des Jahres bittet, gibt er an, dass A seiner Einschätzung nach bei über 35 Euro, B bei über 90 Euro und C unter 40 Euro notieren wird. Am Ende des Jahres greift die Zeitschrift eines der Papiere zufällig (alle mit der gleichen Wahrscheinlichkeit) heraus und überprüft dafür seine Prognose. Die neuen Werte der Aktienkurse A, B und C sind tatsächlich folgendermaßen verteilt:*

$$A \sim N(40, 225) \qquad B \sim N(100, 324) \qquad C \sim N(50, 100)$$

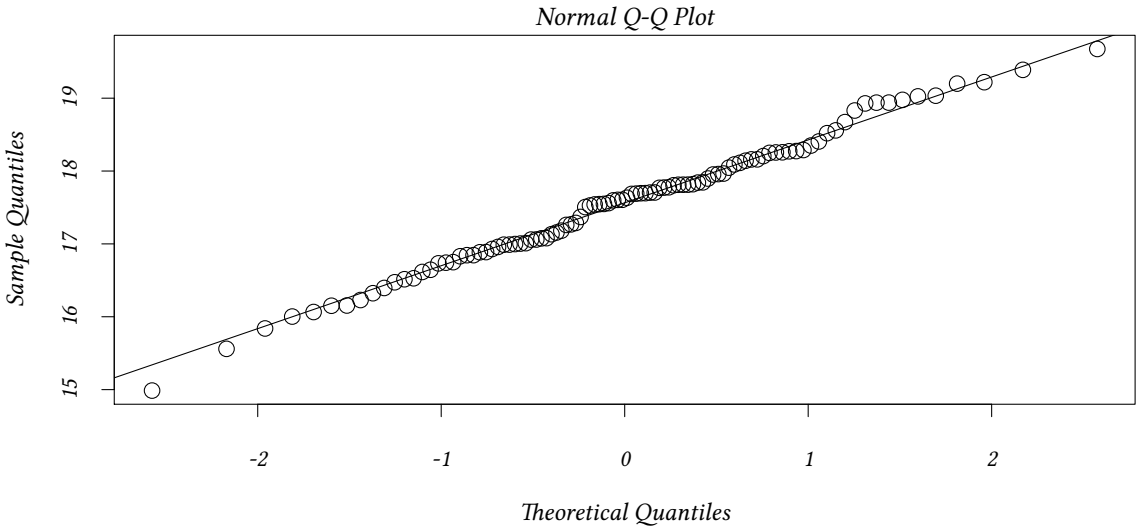
1. *Wie hoch ist die Chance, dass er den Test besteht (also für das zufällig ausgewählte Papier die richtige Prognose abgegeben hat)?*

2. *Gesetzt den Fall, man verrät dem Guru, nachdem er seine Schätzung für C abgegeben hat, den wahren Mittelwert (50) von C. Wird er dann hoffen, dass die Aktie sich volatil (hohe Varianz) oder weniger volatil (niedrige Varianz) entwickelt?*

Übung 7.16 *Bauer Niedermayer hält 100 Milchkühe. Im vergangenen Jahr gab jede im Durchschnitt jeweils 10 000 l Milch bei einer Standardabweichung von 100 l.*

Innerhalb welcher Grenzen kann im nächsten Jahr die Milchleistung jeder Kuh bei gleichen Futter- und Lebensbedingungen wie im Vorjahr erwartetet werden (Konfidenzniveau von 95%)?

Übung 7.17 *Sie messen die Spannungsfestigkeit von 100 Kondensatoren und betrachten den folgenden QQ-Plot:*



1. *Gehen Sie davon aus, dass die Spannungsfestigkeit normalverteilt ist?*

2. *Gehen Sie im Folgenden von der Annahme der Normalverteilung der Spannungsfestigkeit der einzelnen Kondensatoren aus. Welche Verteilung nutzen Sie zum Schätzen des Konfidenzintervalls für die mittlere Spannungsfestigkeit?*

Übung 7.18 *Zur Messung der Geschwindigkeit von Kraftfahrzeugen verwenden Sie ein Radargerät. Die Standardabweichung der Messung Ihres Geräts beträgt 3 km/h. Sie gehen davon aus, dass der Fehler der Anzeige normalverteilt ist. Bei 36 Messungen stellen Sie fest, dass bei einer tatsächlichen Geschwindigkeit von 30 km/h im Mittel 33 km/h angezeigt wird.*

1. *Wie berechnet sich das 95%-Konfidenzintervall für den Erwartungswert?*

2. *In Teilaufgabe 1 haben Sie angenommen, dass die Standardabweichung bekannt ist. Welches 95%-Konfidenzintervall für den Erwartungswert erhalten Sie, wenn Sie mit der empirischen Standardabweichung $\hat{\sigma}_x = 3$ rechnen?*

3. *Es gelten weiterhin die Bedingungen von Teilaufgabe 2, d.h. Sie kennen nur die empirische Standardabweichung. Welches 95%-Konfidenzintervall gilt für die Standardabweichung?*

4. *Sie finden, dass Ihre Konfidenzintervalle zu groß sind. Wie können Sie kleinere Konfidenzintervalle erhalten?*

Übung 7.19 *In einer Mühle wird Getreide gemahlen und das Mehl in Tüten verpackt. Das Gewicht einer Tüte Mehl kann dabei als normalverteilt angenommen werden. Die Varianz des Gewichts ist aus langjähriger Erfahrung bekannt und beträgt $2500\ \text{Gramm}^2$. Eine einfache Stichprobe vom Umfang $n = 25$ ergibt ein Gesamtgewicht von $26\,000$ Gramm.*

1. *Wie groß ist das 95%-Konfidenzintervall für den Mittelwert des Gewichts der Mehltüten?*
2. *Die Länge des Konfidenzintervalls soll nun durch eine größere Stichprobe verkleinert werden. Wie ist n zu wählen, damit das Konfidenzintervall maximal eine Länge von 100 hat?*

Übung 7.20 *Sie wollen mit einem Normal Q-Q Plot überprüfen, ob eine Variable normalverteilt ist. Letzteres ist der Fall, wenn...*

- *alle Punkte im Plot etwa auf einer horizontalen Linie liegen*
- *fast alle Punkte im Plot innerhalb einer Ellipse um den Mittelwert liegen*
- *alle Punkte im Plot etwa auf einer Linie liegen*
- *alle Punkte im Plot etwa auf einer Linie mit Steigung von ca. 1.96 liegen*
- *alle Punkte im Plot gleichmäßig um eine Horizontale streuen*

Übung 7.21 *Die Länge von Knorz ist normalverteilt mit Varianz 100 und unbekanntem Mittelwert. Eine Stichprobe von 100 Stück Knorz ergibt eine mittlere Länge von 200. Wie bestimmen Sie die Untergrenze des 95% Konfidenzintervalls für den Mittelwert?*

Übung 7.22 *Die Länge von Knipps ist normalverteilt mit Standardabweichung 100 und unbekanntem Mittelwert. Eine Stichprobe von 25 Stück Knipps ergibt eine mittlere Länge von 2000.*

1. *Bestimmen Sie die Obergrenze des 99% Konfidenzintervalls für den Mittelwert.*
2. *Wie bestimmen Sie die Untergrenze des 90% Konfidenzintervalls für den Mittelwert?*

Anhang 7.C Warum ist das Konfidenzintervall symmetrisch um den Mittelwert?

Oben haben wir behauptet, das symmetrische Intervall, z.B., $[Q(0.1), Q(0.9)]$ sei das *kleinste* aller möglichen Konfidenzintervalle die 0.8 der Verteilung enthalten.

Um diesen Sachverhalt zu demonstrieren, betrachten wir wieder den Datensatz der Größe von Vätern und Söhnen. Hier schauen wir uns nur die Söhne an. Da viele von uns eher in Zentimetern rechnen, wandeln wir Zoll in Zentimeter um:

```
data(father.son,package="UsingR")
attach(father.son)
scm = sheight * 2.54
```

Ein Intervall in dem 80% der Beobachtungen liegen, wäre folgendes:

```
c(quantile(scm,0),quantile(scm,0.8))
```

```
      0%      80%
148.6080 180.4227
```

Die Breite dieses Intervalls wäre 31.81.

Ein anderes Intervall in dem ebenfalls 80% der Beobachtungen liegen wäre

```
c(quantile(scm,.2),quantile(scm,1))
```

```
     20%     100%
168.8288 199.0466
```

Die Breite dieses Intervalls wäre 30.22.

Wir könnten aber auch ein symmetrisches Intervall nehmen:

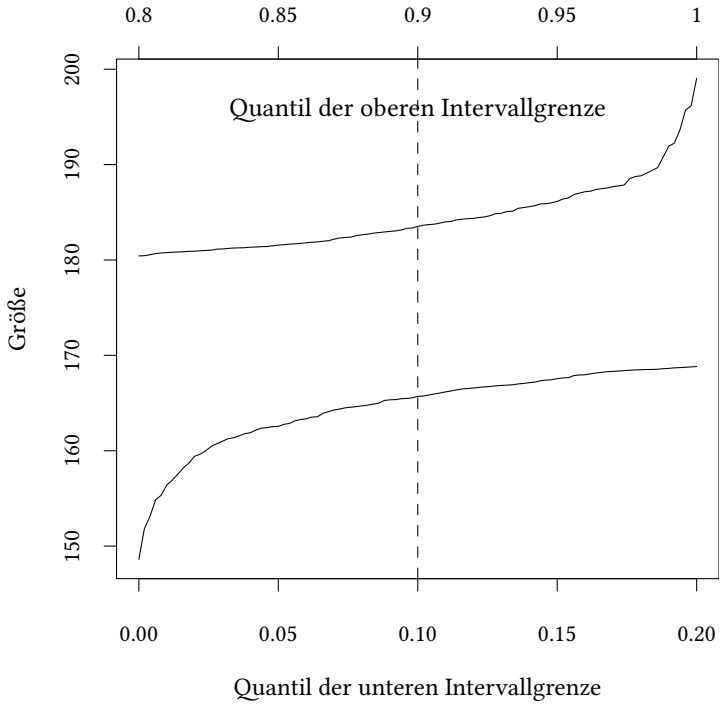
```
c(quantile(scm,.1),quantile(scm,.9))
```

```
     10%      90%
165.6965 183.5202
```

Die Breite dieses Intervalls wäre 17.82. Das ist erheblich weniger als die Breite der asymmetrischen Intervalle.

Die folgende Graphik veranschaulicht, wie die Intervallgrenzen von den der Untergrenze bzw. Obergrenze des Quantils abhängen:

```
plot(function(x) quantile(scm,x),0,.2,ylim=range(scm),ylab="Größe",
     xlab="Quantil der unteren Intervallgrenze")
plot(function(x) quantile(scm,x+.8),0,.2,add=TRUE)
abline(v=.1,lty="dashed")
lower <- seq(0,.2,.05)
upper <- lower + .8
axis(side=3,at=lower,labels=upper)
text(.1,max(scm),pos=1,"Quantil der oberen Intervallgrenze")
```



Wir sehen, dass das für diese (empirische) Verteilung ein Intervall das 80% der Beobachtungen enthält besonders schmal ist, wenn es »in der Mitte«, von 10% bis 90% gebildet wird.

Obige Überlegung bezieht sich allerdings nicht auf den Schätzer.

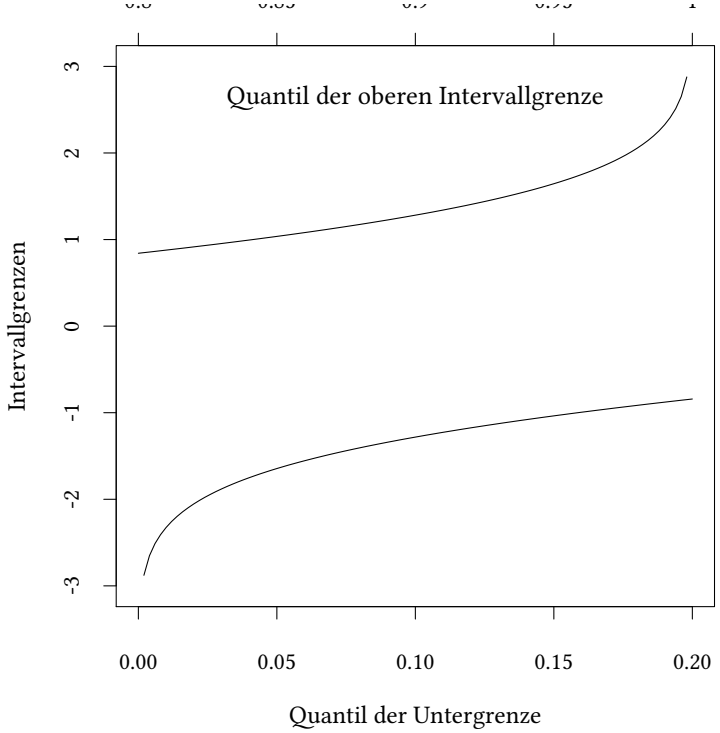
Wenn wir Konfidenzintervalle schätzen, werden wir oft annehmen, dass unser Schätzer $\hat{\theta}$ einer bekannten Verteilung folgt, z.B. der Normalverteilung.

Wenn wir für die Normalerverteilung verschiedene Intervalle bilden, die z.B. alle 80% der Verteilung enthalten, dann ist dieses Intervall bei einem symmetrischen Interval besonders schmal, d.h. wenn die Untergrenze gerade 10% und die Obergrenze gerade 90% ist.

Satz: Sei X eine normalverteilte Zufallsvariable, dann ist das kleinste Intervall das X mit einer vorgegebenen Wahrscheinlichkeit $1-\alpha$ enthält $\left[Q\left(\frac{\alpha}{2}\right), Q\left(1-\frac{\alpha}{2}\right)\right]$

Wir betrachten nochmals verschiedene Konfidenzintervalle, diesmal aber nicht für die empirische Verteilung aus dem Beispiel sondern für die vorgegebene Verteilung, die Normalverteilung:

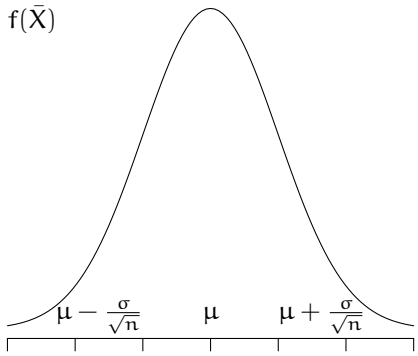
```
plot( function(x) {qnorm(x)},0,.2,ylim=c(-3,3),ylab="Intervallgrenzen",
     xlab="Quantil der Untergrenze")
plot( function(x) {qnorm(x+.8)},0,.2,ylim=c(-3,3),add=TRUE)
axis(side=3,at=lower,labels=upper)
text(.1,3,pos=1,"Quantil der oberen Intervallgrenze")
```



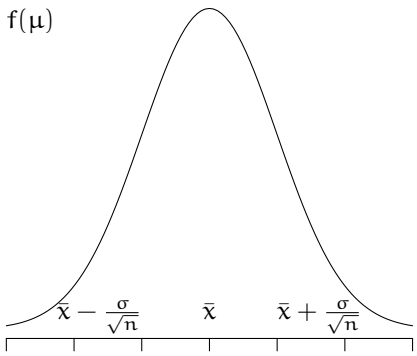
Anhang 7.D Konfidenzintervalle binomialverteilter Zufallsvariablen

zur Erinnerung: Mittelwert normalverteilter Zufallsvariablen:

Gedankenexperiment: Wie wäre es, wenn wir das wahre μ kennen würden?



Gedankenexperiment 2: unendlich viele Stichproben

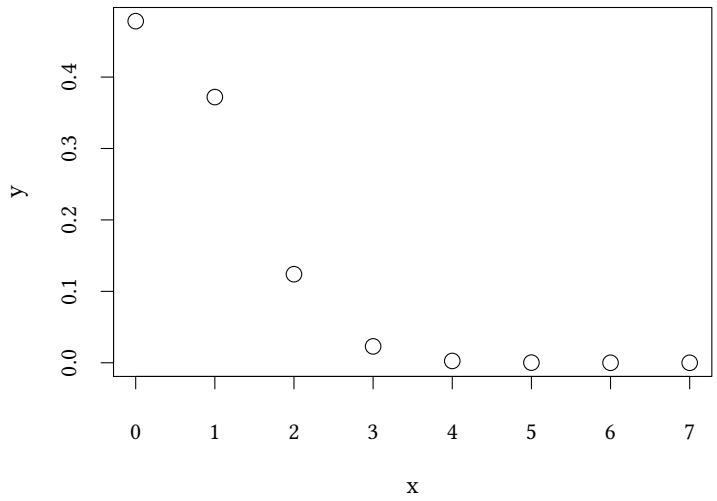


Wir konnten ausnutzen, dass die Normalverteilung symmetrisch ist. Wir sind davon ausgegangen, dass die Standardabweichung sich nicht mit dem Mittelwert ändert.

Es gibt jedoch Zufallsvariablen, die nicht normalverteilt sind, z.B. binomialverteilte Zufallsvariablen.

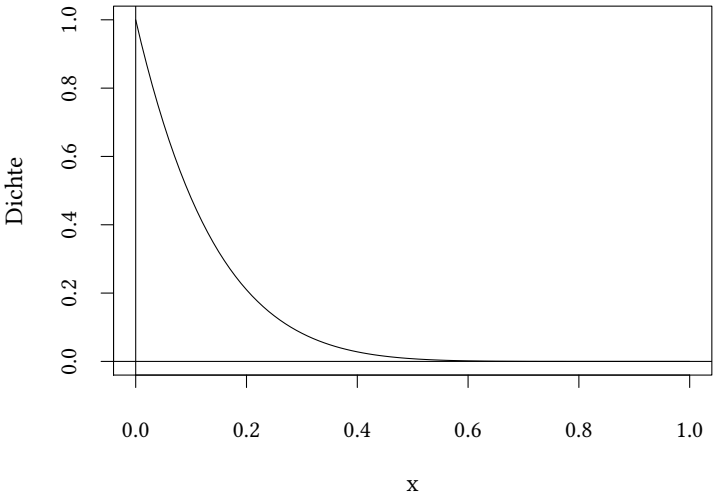
```
x=0:7
y=dbinom(x,size=7,prob=.1)
```

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



Nehmen wir an, im obigen Beispiel hätten wir 0 Erfolge beobakchtet, die a priori Erfolgswahrscheinlichkeit wäre aber unbekannt.

```
plot(function(p)  dbinom(0,size=7,prob=p) ,0,1,ylab="Dichte")
abline(v=0/7)
abline(h=0)
```



`binom.confint` berechnet Konfidenzintervalle für eine binomialverteilte Zufallsvariable und verwendet dazu unterschiedliche Methoden.

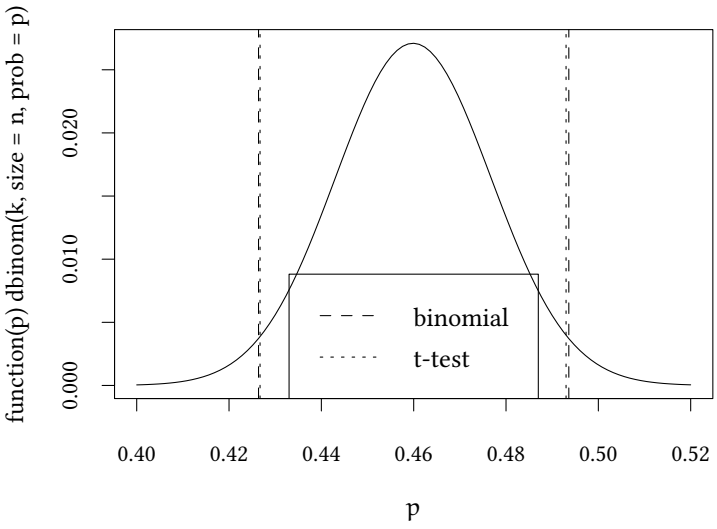
```
library(binom)
binom.confint(0,7,conf.level=.95,method="exact")
```

```
  method x n mean lower     upper
1  exact 0 7    0     0 0.4096164
```

Übung 7.23 *Betrachten Sie wieder den Datensatz* `Participation` *aus der Bibliothek* `Ecdat`. *Verwenden Sie nun die Binomialverteilung, um ein Konfidenzintervall für den Mittelwert der durchschnittlichen Beteiligung am Arbeitsmarkt* `lfp` *zu ermitteln. Vergleichen Sie das Ergebnis mit dem Resultat der* t *Verteilung.*

```
data(Participation,package="Ecdat")
attach(Participation)
(n <- length(lfp))

[1] 872

(k <- sum(lfp=="yes"))

[1] 401

binom.confint(k,n,conf.level=.95,method="exact")

  method   x   n      mean     lower     upper
1  exact 401 872 0.4598624 0.4263934 0.4936035

t.test(lfp=="yes")


 One Sample t-test

data:  lfp == "yes"
t = 27.231, df = 871, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.4267181 0.4930067
sample estimates:
mean of x
0.4598624
```

```
plot(function(p)  dbinom(k,size=n,prob=p) ,0.4,0.52,xlab='$p$')
z<-binom.confint(sum(lfp=="yes"),length(lfp),conf.level=.95,method="exact")
abline(v=z[c("lower","upper")],lty=2)
zt<-t.test(lfp=="yes")
abline(v=zt$conf.int,lty=3)
legend("bottom",c("binomial","t-test"),lty=c(2,3))
```



Anhang 7.E Konfidenzintervall für die Varianz

- Population: $X \sim N(\mu, \sigma^2)$
- Stichprobe: $X_1, \ldots, X_n$
  - Aus der Stichprobe schätzen wir den Populationsparameter $\hat{\sigma}^2$
  - Nun: Finde ein Intervall $[\underline{V}, \bar{V}]$ das den (unbekannten) wahren Parameter σ^2 zu einem vorgegebenen Konfidenzniveau $1 - \alpha$ enthält.
  - Erinnerung: Bei Punktschätzungen war der Schätzer $\hat{\theta}$ eine Zufallsvariable.
  - Genauso bei Intervallschätzungen: $\underline{V}(X_1, \ldots, X_n)$ und $\bar{V}(X_1, \ldots, X_n)$ sind Stichprobenfunktionen, und damit Zufallsvariablen.

$$\sum_{i=1}^{n} X_i \sim N(n \cdot \mu, n \cdot \sigma^2)$$

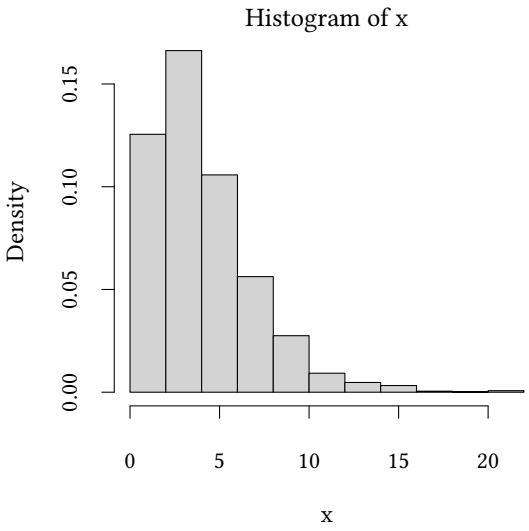
$$\frac{X - \mu}{\sigma} \sim N(0, 1)$$

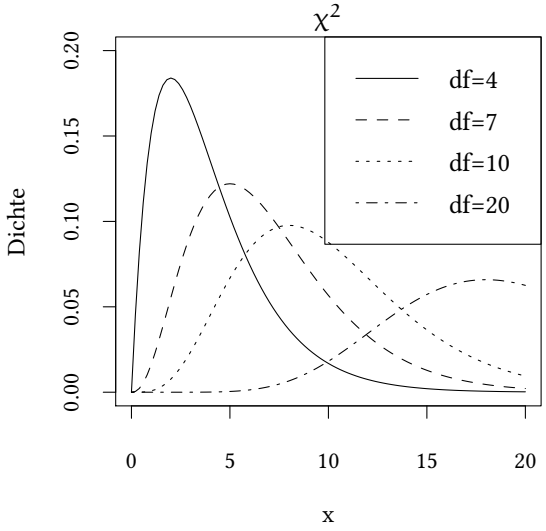
$$\sum_{i=1}^{n} \left(\frac{X - \mu}{\sigma} \right)^2 \sim \chi^2_n$$

Zur Erinnerung: Die Summe von n quadrierten unabhängigen standardnormalverteilten Zufallsvariablen ist χ^2-verteilt mit n Freiheitsgraden.

Wir erzeugen zunächst einige Summen von quadrierten normalverteilten Zufallsvariablen. Deren Verteilung stellt das linke Bild dar. Das rechte Bild zeigt die χ^2 Verteilung für unterschiedliche Freiheitsgrade.

```
x <- sapply(1:2000,function(x) sum(rnorm(4)^2))
```





Betrachte nun die Verteilung der Varianz:

$$\hat{\sigma}_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})^2$$

$$(n-1)\cdot\hat{\sigma}_X^2 = \sum_{i=1}^{n}(X_i - \bar{x})^2$$

$$\frac{(n-1)\cdot\hat{\sigma}_X^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \bar{x}}{\sigma}\right)^2 \sim \chi^2_{n-1}$$

Wir nutzen diesen Zusammenhang, um genauso wie bei der Normalverteilung auch aus den Quantilswerten der χ^2-Verteilung für $\frac{\alpha}{2}$ und $1-\frac{\alpha}{2}$ die Grenzen des Konfidenzintervalls zu bestimmen.

$$\text{definiere } \underline{\chi}^2 = Q_{\chi^2_{n-1}}\left(\frac{\alpha}{2}\right) \qquad \bar{\chi}^2 = Q_{\chi^2_{n-1}}\left(1-\frac{\alpha}{2}\right)$$

$$\frac{(n-1)\cdot\hat{\sigma}_X^2}{\bar{\sigma}^2} = \underline{\chi}^2$$

$$\frac{(n-1)\cdot\hat{\sigma}_X^2}{\underline{\sigma}^2} = \bar{\chi}^2$$

$$\frac{(n-1)\cdot\hat{\sigma}_X^2}{\bar{\chi}^2} \le \sigma^2 \le \frac{(n-1)\cdot\hat{\sigma}_X^2}{\underline{\chi}^2}$$

$$\sqrt{\frac{(n-1)\cdot\hat{\sigma}_X^2}{\bar{\chi}^2}} \le \sigma \le \sqrt{\frac{(n-1)\cdot\hat{\sigma}_X^2}{\underline{\chi}^2}}$$

Übung 7.24 *Geben Sie ein 95% Konfidenzintervall für die Standardabweichung der Größe der Söhne aus dem Datensatz* `father.son` *an.*

Wir schätzen zunächst die Varianz der Größe:

```
## data(father.son,package="UsingR")
## attach(father.son)
(varest = var(sheight))
```

```
[1] 7.922545
```

Die Anzahl der Freiheitsgrade ist

```
(n=length(sheight)-1)
```

```
[1] 1077
```

Also ergeben sich die Quantile der χ^2 Verteilung zu

```
qchisq(c(.975,.025),df=n)
```

```
[1] 1169.8435  987.9446
```

Mithin ergibt sich das Konfindenzintervall zu

```
sqrt(n * varest / qchisq(c(.975,.025),df=n))
```

```
[1] 2.700700 2.938826
```

Übung 7.25 *In einem Unternehmen sei die Anzahl der Krankschreibungen* X *pro Monat unabhängig normalverteilt. Folgende Realisationen von* X *liegen vor: (12,17,35,19,41,22,27,9,38,11).*
Bestimmen Sie das 90%-Konfidenzintervall für die Standardabweichung.

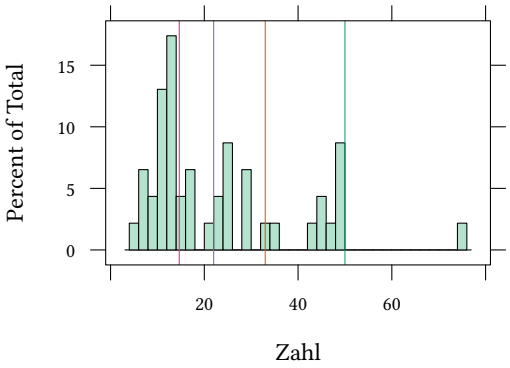
```
## Warning in library(package, lib.loc = lib.loc, character.only = TRUE, logical.
= TRUE, : there is no package called 'binom'
## Warning in library(package, lib.loc = lib.loc, character.only = TRUE, logical.
= TRUE, : there is no package called 'Sleuth2'
## Warning in library(package, lib.loc = lib.loc, character.only = TRUE, logical.
= TRUE, : there is no package called 'UsingR'
## Warning in library(package, lib.loc = lib.loc, character.only = TRUE, logical.
= TRUE, : there is no package called 'calibrate'
## Warning in library(package, lib.loc = lib.loc, character.only = TRUE, logical.
= TRUE, : there is no package called 'relaimpo'
```

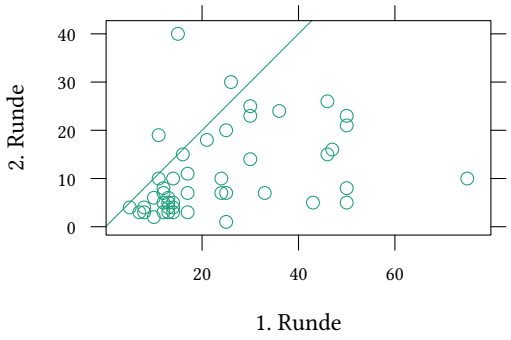
8. Nichtparametrische Tests

8.1. Motivation: Ökomische Erwartungen und Guessing Games

In vielen ökonomischen Situationen ist es für die Entscheider wichtig, Erwartungen über andere Entscheider zu bilden. Sie wollen ein Wertpapier nur dann kaufen, wenn Sie damit rechnen, dass es auch von anderen Leuten gekauft wird, und dann im Preis steigt. Sie wollen verkaufen, bevor alle anderen verkaufen. Diese Situation bildet Rosemarie Nagel in einem einfachen Experiment ab:

Jede Versuchsperson wählt eine Zahl zwischen 0 und 100.
Wir sammeln alle Zahlen, bilden den Mittelwert.
Derjenige, der am nächsten an $\frac{2}{3}$ des Mittelwertes ist, gewinnt einen Preis.
Anwendung: z.B. Finanzmärkte, Währungsspekulationen,...
Frage: Nähern sich die Versuchspersonen bei Wiederholung des Spiels der Gleichgewichtslösung?





Rosemarie Nagel, American Economic Review, 1995.

```
wilcox.test(runde1,runde2,paired=TRUE)


 Wilcoxon signed rank test with continuity correction

data:  runde1 and runde2
V = 1016.5, p-value = 0.0000002019
alternative hypothesis: true location shift is not equal to 0
```

8.2. Wilcoxon signed rank Test für paarweise Stichproben

- kleine Stichprobe (n ist klein)
- Zufallsvariable folgt nicht unbedingt einer bekannten Verteilung (Normalverteilung?)

- Mittelwert folgt nicht unbedingt einer bekannten Verteilung (z.B. Normalverteilung)

$\rightarrow$ kein t-Test möglich.

Statt dessen: Wilcoxon signed rank test.

große Stichprobe oder normalverteilt	sonst
t-Test für unverbundene Stichproben	Mann-Whitney U-Test
t-Test für verbundene Stichproben	Wilcoxon signed rank test.

Wenn »vorher« und »nachher« gleich verteilt wären, sollten wir erwarten, dass gleich viele Beobachtungen oberhalb und unterhalb der 45°-Linie liegen. Es fällt auf, dass fast alle Beobachtungen unter der 45°-Linie liegen, Zahlen in der 1. Runde also größer sind als in der 2. Runde. Wir haben nun zwei Möglichkeiten:

- Wir betrachten einfach das Verhältnis der Beobachtungen oberhalb zu unterhalb der 45°-Linie → Binomialtest. (Weil dieser Test nicht ausnutzt, wie weit die Beobachtungen von der 45°-Line entfernt liegen, hat dieser Test auch weniger Power.)

- Wir nutzen außerdem aus, wie weit die Beobachtungen von der 45°-Linie weg sind → Wilcoxon signed rank test.

Binomialtest Wenn wir davon ausgehen, dass »vorher« und »nachher« gleich verteilt sind, dann sollte etwa die Hälfte der »vorher«-Beobachtungen größer und die andere Hälfte kleiner als die »nachher«-Beobachtungen sein. Der Anteil der größeren »vorher« Beobachtungen sollte also binomialverteilt mit Parameter $p = 1/2$ sein.

```
diff <- runde1 - runde2
table(diff > 0)
```

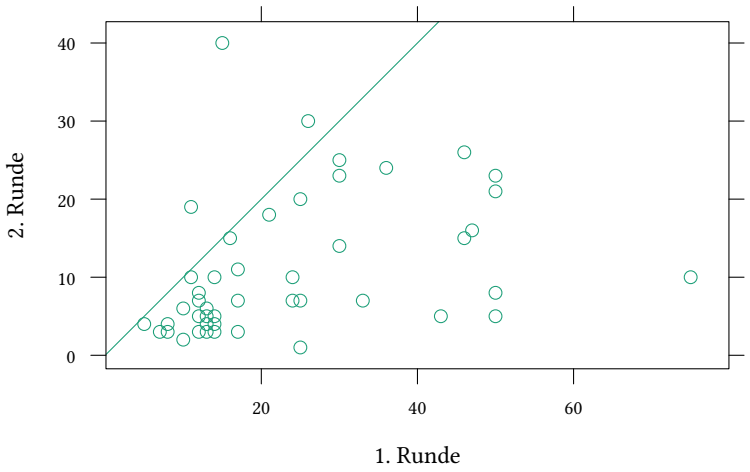
```
FALSE  TRUE
    3    43
```

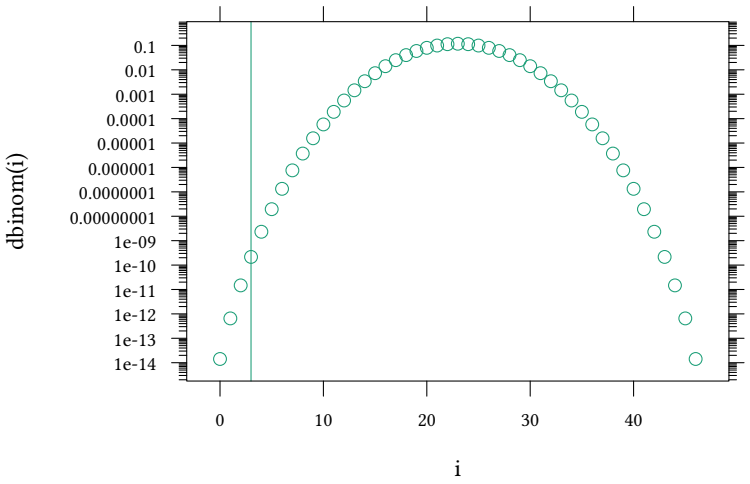
Nur mit den Vorzeichen der Differenzen können wir einen Binomialtest machen. Das heißt, wir rechnen aus, wie wahrscheinlich (gegeben $p = 1/2$) es ist, eine Stichprobe zu beobachten, die zu unserer Hypothese ($p = 1/2$) mindestens so advers ist, wie die Stichprobe, die uns vorliegt.

Der Binomialtest hat aber nur wenig Power — wir erhalten häufiger ein insignifikantes Ergebnis, obwohl der Unterschied zwischen beiden Stichproben mit einem stärkeren Test signifikant ist — in diesem Beispiel klappt es aber:

```
2*pbinom(3,p=.5,size=46)
```

```
[1] 4.621938e-10
```





Vielleicht etwas komfortabler geht es auch so:

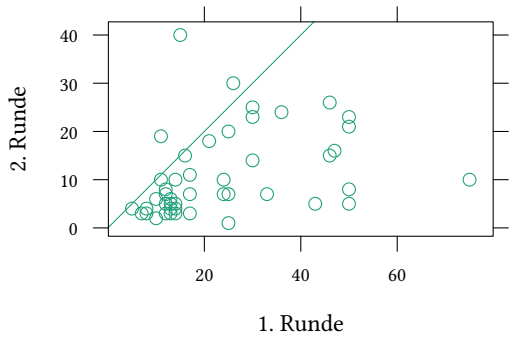
```
table(sign(runde1 - runde2))


-1  1
 3 43

binom.test(table(sign(runde1 - runde2)),p=.5)


 Exact binomial test

data:  table(sign(runde1 - runde2))
number of successes = 3, number of trials = 46, p-value = 4.622e-10
```

```
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.01365677 0.17896439
sample estimates:
probability of success
            0.06521739
```

In beiden Fällen erhalten wir den gleichen p-Wert (4.622×10^{-10}).

Rangsummentest

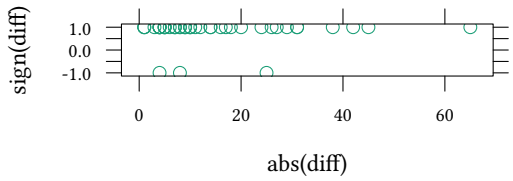
Der Rangsummentest berücksichtigt nicht nur die Vorzeichen, sondern auch die Ränge der Abstände von der 45°-Linie:



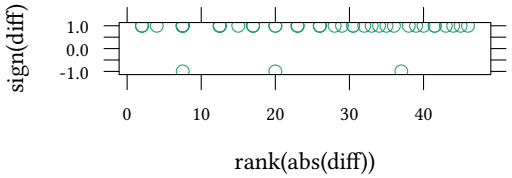

```
sum(rank(abs(diff)) * sign(diff))

[1] 952
```

Hier werden größere Abweichungen von der 45°-Line auch stärker gewichtet. Der Test nutzt die in der Stichprobe vorhandene Information besser aus als der Binomialtest.

Hier sind die Abstände, getrennt für positive und negative Differenzen



Hier sind die Ränge



Es gibt sehr viele (gleichwertige) Möglichkeiten, aus den Rängen mit Vorzeichen eine Statistik zu berechnen. R verwendet nur die positiven Werte:

```
V<-sum(rank(abs(diff))[diff>0])

[1] 1016.5
```

Die Summe der positiven Ränge V hat (unter der Nullhypothese) eine bekannte Verteilung. Diese Verteilung kann durch die Normalverteilung approximiert werden:

$$\frac{V-\frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \sim N(0,1)$$

(n ist die Anzahl der Paare.)

`wilcox.test` führt Wilcoxon Rangsummentests und Mann-Whitney U Tests durch.

```
wilcox.test(runde1,runde2,paired=TRUE,alternative="greater")


 Wilcoxon signed rank test with continuity correction

data:  runde1 and runde2
V = 1016.5, p-value = 0.0000001009
alternative hypothesis: true location shift is greater than 0
```

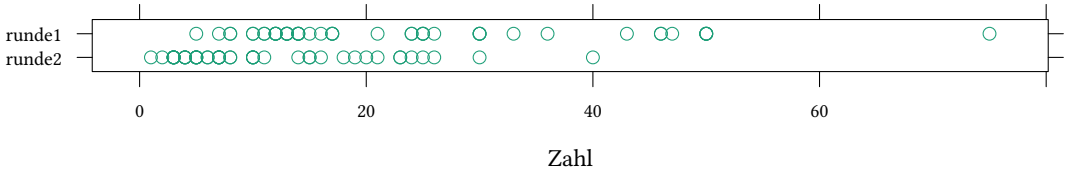
Wenn wir zusätzlich annehmen könnten, dass die Stichprobe aus einer normalverteilten Population gezogen wurde, dann könnten wir auch einen t-Test machen:

```
t.test(runde1,runde2,paired=TRUE,alternative="greater")


 Paired t-test

data:  runde1 and runde2
t = 5.7298, df = 45, p-value = 0.0000003925
alternative hypothesis: true mean difference is greater than 0
95 percent confidence interval:
 9.05135     Inf
sample estimates:
mean difference
       12.80435
```

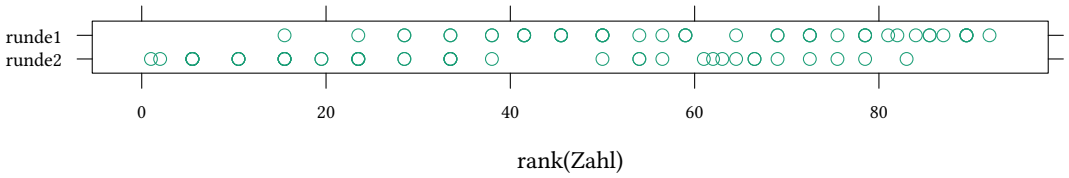
Da beide Tests von unterschiedlichen Annahmen ausgehen, erhalten wir auch unterschiedliche p-Werte: 0.0000001 mit dem Wilcoxon Test und 0.0000004 mit dem t-Test. Welches dieser Ergebnisse wir bevorzugen, hängt von den Annahmen ab, die wir machen wollen.

8.3. Mann-Whitney U Test für unverbundene Stichproben

Übung 8.1 *Wir betrachen wieder die Daten aus dem Guessing Game nehmen aber jetzt an, dass es sich um zwei verschiedene Gruppen von Versuchspersonen handelt. (Wir vergessen sozusagen, welche »runde1« Beobachtung zu welcher »runde2« Beobachtugen gehört.)*



Wir verwenden den gleichen Trick wie bei verbundenen Stichproben: Ränge:



Wir vergleichen nun nicht den Mittelwert der Beobachtungen (wie beim t-Test) sondern den Mittelwert der Ränge. So kann eine seltsame Verteilung nicht stören.

Wir vergleichen zwei Stichproben:

- $X_1, \dots, X_m$
- $Y_1, \dots, Y_n$

Auch hier gibt verschiedene Möglichkeiten, eine normalverteilte Teststatistik zu konstruieren, z.B. folgende:

W ist die Summe der Ränge nur einer der beiden Stichproben, z.B. $(X_1, \dots, X_m)$.

Unsere Teststatistik ist dann

$$\frac{W - \frac{1}{2}n \cdot (m+n+1)}{\sqrt{\frac{1}{12}m \cdot n \cdot (m+n+1)}} \sim N(0,1)$$

```
wilcox.test(runde1,runde2,alternative="greater")
```

```
 Wilcoxon rank sum test with continuity correction

data:  runde1 and runde2
W = 1679.5, p-value = 0.0000006048
alternative hypothesis: true location shift is greater than 0
```

Wenn wir zusätzlich annehmen könnten, dass die Stichprobe aus einer normalverteilten Population gezogen wurde, dann könnten wir auch einen t-Test machen:

```
t.test(runde1,runde2,alternative="greater")
```

```
 Welch Two Sample t-test

data:  runde1 and runde2
t = 4.8071, df = 70.908, p-value = 0.000004157
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 8.36502     Inf
sample estimates:
mean of x mean of y
 23.78261  10.97826
```

Auch hier erhalten wir wieder, da beide Tests von unterschiedlichen Annahmen ausgehen, unterschiedliche p-Werte: 0.0000006 mit dem Mann-Whitney U Test und 0.0000042 mit dem t-Test. Wieder hängt es von den Annahmen ab, die wir machen wollen, welchen Test wir vorziehen.

8.4. Motivation: Speed dating und mate copying

In den Wirtschaftswissenschaften haben wir es oft mit Suchentscheidungen im weitesten Sinne zu tun. Arbeiter suchen eine Beschäftigung, Firmen suchen Arbeiter, Konsumenten suchen Produkte.

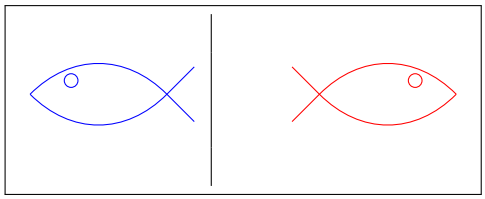
Es ist schwierig, diese Enscheidungen im Labor so zu untersuchen, dass Versuchspersonen auch hohe Anreize haben, gut zu suchen. Eine Situation, für die Anreize vermutlich hoch sind, ist *mate search*, also die Suche nach Partnern für eine romantische Beziehung. Hier gehen wir davon aus, dass Versuchspersonen motiviert und bei der Sache sind.

Eine Hypothese für alle Suchsituationen ist, dass Entscheider die Suchstrategien anderer Entscheider kopieren. Wir nennen dieses Verhalten “mate copying”.

In der Biologie wird mate-copying z.B. bei Fischen beobachtet: Einige weibliche Fische sparen sich die Mühe, einen begehrenswerten männlichen Fisch selbst auszuwählen, und interessieren sich statt dessen für männliche Fische, die bereits das Interesse anderer weiblicher Fische finden (Dugatkin. Sexual selection and imitation: females copy the mate choice of others. Am. Nat. 1992;139:1384–1389.)

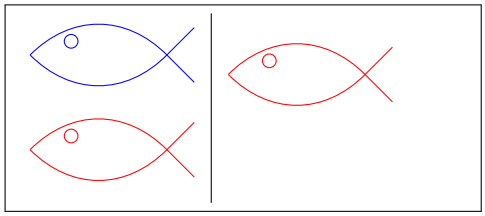
Mate copying findet sich aber auch bei Menschen. Arbeiter wählen die Firma, die auch bei anderen Arbeitern beliebt ist, Firmen stellen Arbeiter ein, die auch bei anderen Firmen erfolgreich sind, Konsumenten wählen Produkte, die auch anderen Konsumenten gefallen, und Männer und Frauen lassen sich bei der Partnerwahl davon leiten, ob andere Männer oder Frauen einen Zielpartner attraktiv finden. Bei der Partnerwahl nennt man dieses Verhalten *mate copying*. Ein effizientes Verfahren um *mate copying* zu beobachten ist *speed dating* (Todd, Penke, Fasolo, Lenton. Proc. Natl Acad. Sci., 2007).

Experiment 1:



Der weiblich Fisch hat kein Interesse, wenn sich kein anderer weibliche Fisch für den männlichen Fisch interessiert

Experiment 2:



Eigentlich interessieren wir uns für ökonomische Suchentscheidungen, z.B. auf dem Arbeitsmarkt.

Speed-Dating ist ein Beispiel für Suchentscheidungen.

Im ersten Schritt des Experiments werden Versuchspersonen beim Speed-Daten gefilmt:

- ca. 20 Männer und 20 Frauen zahlen je ca. 30$ und treffen sich in einem Raum.
- Frauen sitzen jeweils an einem Tisch, Männer rotieren in jeder Runde einen Tisch weiter.
- Jedes »Date« dauert 5 Minuten.
- Beide markieren auf einer Karte, ob sie die andere Person wieder treffen möchten.

Place, S.S., Todd, P.M., Penke, L., Asendorpf, J.B. (2009). The ability to judge the romantic interest of others. Psychological Science, 20(1), 22-26.

Im zweiten Schritt des Experiment sehen andere Versuchspersonen Bilder aus dem ersten Experiment:

1. Speed Dating in Berlin

2. Experiment in Bloomington
   a) Versuchspersonen bewerten ein Bild der Zielperson, auch bezüglich Interesse als kurzfristiger oder langfristiger Partner.
   b) Versuchspersonen sehen Speed-Dating Interaktion und bewerten die Interaktion.
   c) Versuchspersonen bewerten ein Bild der Zielperson erneut.

Die folgende Tabelle zeigt, auf welche Weise sich das eigene Interesse an der Zielperson verändert und wie diese Änderung vom beobachteten Interesse abhängt.

	eigenes Interesse ist…		
	weniger	unverändert	mehr
kein Interesse beobachtet	11	12	17
Interesse beobachtet	5	7	28

```
matechoice

                         weniger unverändert mehr
kein Interesse beobachtet      11          12   17
Interesse beobachtet            5           7   28
```

Hier ist ein Test auf Unabhängigkeit dieser Häufigkeiten:

```
chisq.test(matechoice)


 Pearson's Chi-squared test

data:  matechoice
X-squared = 6.2547, df = 2, p-value = 0.04383
```

Oben haben wir uns in Abschnitt 8.2 und 8.3 gefragt, wie man zwei Gruppen, etwa »runde1« und »runde2«, miteinander vergleichen kann. Der χ^2 Test misst Unterschiede zwischen mehr als zwei Gruppen. Allerdings berücksichtigt er nur Informationen über Häufgkeiten.

8.5. χ^2 Anpassungstest — Vergleich von Häufigkeiten mit einer theoretischen Verteilung

Voraussetzung: Es liegen *Häufigkeiten* für Merkmale in *einer* Dimension vor.

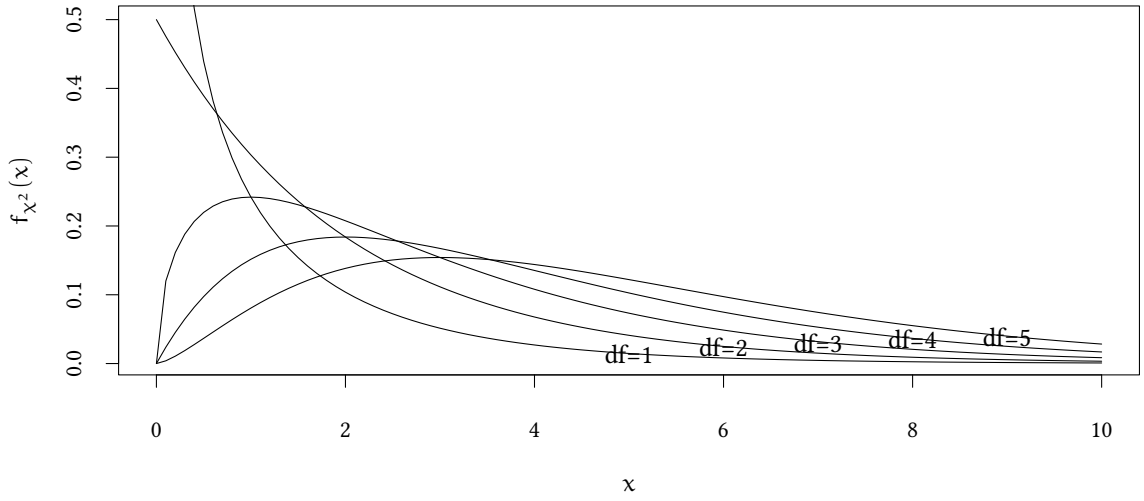
Merkmale	a_1	a_2	$\cdots$	a_k	
theoretisch	$P(a_1)$	$P(a_2)$	$\cdots$	$P(a_k)$	$\sum P(a_i) = 1$
beobachtet	$X(a_1)$	$X(a_2)$	$\cdots$	$X(a_k)$	$\sum X(a_i) = n$

- $a_1, \ldots, a_k$ Merkmale

- $P(a_1), \ldots, P(a_k)$ nullhypothetische Wahrscheinlichkeiten der Merkmale mit $\sum_{i=1}^{k} P(a_i) 1$
- $X(a_1), \ldots, X(a_k)$ empirische Häufigkeiten der Merkmale mit $\sum_{i=1}^{k} X(a_i) = n$
- H_0 : die Stichprobe ist entsprechend $P()$ verteilt.
- H_1 : die Stichprobe ist nicht entsprechend $P()$ verteilt.

Man kann zeigen, dass

$$\sum_{i=1}^{k} \frac{(X(a_i) - n \cdot P(a_i))^2}{n \cdot P(a_i)} \sim \chi^2_{k-1}$$



Beispiel: Wir vermuten, dass die Gruppe der Personen die »kein Interesse beobachtet« haben, einfach mit gleicher Wahrscheinlichkeit (also $\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$) eine Kategorie gewählt haben.

```
matechoice["kein Interesse beobachtet",]

    weniger unverändert        mehr
         11          12          17
```

```
chisq.test( matechoice["kein Interesse beobachtet",] , p=c(1/3, 1/3, 1/3))


 Chi-squared test for given probabilities

data:  matechoice["kein Interesse beobachtet", ]
X-squared = 1.55, df = 2, p-value = 0.4607
```

8.6. Abhängigkeit von zwei Merkmalen – χ^2 Kontingenztest

Abhängigkeit von zwei Merkmalen – Idee Wenn zwei Zufallsvariablen X und Y *unabhängig* sind, dann gilt

$$P(X = x \text{ und } Y = y) = P(X = x) \cdot P(Y = y)$$

Beispiel:

	$X = x_1$	$X = x_2$	
$Y = y_1$	15	35	50
$Y = y_2$	15	35	50
	30	70	100

In unserem Beispiel war die Randverteilung bekannt. Normalerweise ist das nicht der Fall, wir nehmen dann die empirische Randverteilung.

empirische Häufigkeiten:

	$X = x_1$	$X = x_2$
$Y = y_1$	10	40
$Y = y_2$	20	30

$\downarrow$

empirische Randverteilung:

	$X = x_1$	$X = x_2$	
$Y = y_1$			50
$Y = y_2$			50
	30	70	100

$\downarrow$

Häufigkeiten bei Unabhängkeit:

	$X = x_1$	$X = x_2$
$Y = y_1$	15	35
$Y = y_2$	15	35

Abhängigkeit von zwei Merkmalen – χ^2 Kontingenztest Allgemein betrachten wir folgende Situation:

	weniger	unverändert	mehr
kein Interesse beobachtet	11	12	17
Interesse beobachtet	5	7	28

	b_1	b_2	$\cdots$	b_k	
a_1	x_{11}	x_{12}	$\cdots$	x_{1k}	$\sum_{j=1}^{k} x_{1j}$
a_2	x_{21}	x_{22}	$\cdots$	x_{2k}	$\sum_{j=1}^{k} x_{2j}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
a_n	x_{n1}	x_{n2}	$\cdots$	x_{nk}	$\sum_{j=1}^{k} x_{nj}$
	$\sum_{i=1}^{n} x_{i1}$	$\sum_{i=1}^{n} x_{i2}$	$\cdots$	$\sum_{i=1}^{n} x_{ik}$	$\sum_{i=1}^{n} \sum_{j=1}^{k} x_{ij}$

nun definiere $e_{ij} = \frac{\sum_{j=1}^{k} x_{ij} \cdot \sum_{i=1}^{n} x_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}}$
die »erwartete Häufigkeit« falls beide Merkmale nichts miteinander zu tun haben.
dann ist (jedenfalls falls $n + k > 4$)

$$\sum_{i=1}^{n}\sum_{j=1}^{k} \frac{(x_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2_{(n-1)\cdot(k-1)}$$

Hier sind die Zwischenergebnisse für unser Beispiel:

x:

	weniger	unverändert	mehr
kein Interesse beobachtet	11	12	17
Interesse beobachtet	5	7	28

	weniger	unverändert	mehr	Summe
kein Interesse beobachtet	11	12	17	40
Interesse beobachtet	5	7	28	40
Summe	16	19	45	80

e:

	weniger	unverändert	mehr
kein Interesse beobachtet	8.00	9.50	22.50
Interesse beobachtet	8.00	9.50	22.50

$x - e$:

	weniger	unverändert	mehr
kein Interesse beobachtet	3.00	2.50	-5.50
Interesse beobachtet	-3.00	-2.50	5.50

$(x - e)^2$:

	weniger	unverändert	mehr
kein Interesse beobachtet	9.00	6.25	30.25
Interesse beobachtet	9.00	6.25	30.25

$\frac{(x-e)^2}{e}$:

	weniger	unverändert	mehr
kein Interesse beobachtet	1.12	0.66	1.34
Interesse beobachtet	1.12	0.66	1.34

Wenn man diese Zahlen aufsummiert, erhält man die χ^2-Teststatistik.

```
chisq.test(matechoice)


 Pearson's Chi-squared test

data:  matechoice
X-squared = 6.2547, df = 2, p-value = 0.04383
```

8.7. Literatur

- Dolić, Statistik mit R, Kapitel 7.3.2.
- Hartung, Statistik, Kapitel IV.5.4.3.m VII.2.2.2.
- Schira, Statistische Methoden der VWL und BWL-Theorie und Praxis, Kapitel 16.2, 16.3.
- Verzani, Using R for Introductory Statistics, Chapter 8.4.2, 9.1.

8.8. Schlüsselbegriffe

- parametrischer / nicht-parametrischer Test
- Wilcoxon Test (für paarweise Stichproben)
- Mann-Whitney U Test (für unverbundene Stichproben)
- χ^2 Kontingenztest
- normalverteilte Zufallsvariablen → parametrische Tests (exakt)
- große Stichprobe/nicht normalverteilte Zufallsvariablen → parametrische Tests (asymptotisch)
  - bekannte Varianz → t ist normalverteilt
  - unbekannte Varianz → t ist t-verteilt
  - paarweise Tests / Tests für unverbundende Stichproben
- kleine Stichprobe/nicht normalverteilte Zufallsvariablen → nichtparametrische Tests
  - paarweise Tests / unverbundende Stichproben / Kontingenztests

Anhang 8.A Beispiele für die Vorlesung

```
t.test(x1,x2,paired=TRUE)
```

```
 Paired t-test

data:  x1 and x2
t = -5.7127, df = 9, p-value = 0.0002896
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -7.853645 -3.398087
sample estimates:
mean difference
      -5.625866
```

- Es wurde ein paarweiser Test durchgeführt.
- Die Mittelwerte von x1 und x2 sind nicht signifikant voneinander verschieden.
- Die Hypothese, die Differenz zwischen x1 und x2 sei im Erwartungswert −7, kann nicht verworfen werden.
- Die Hypothese, die Differenz zwischen x1 und x2 sei im Erwartungswert −7, wird abgelehnt.
- Bei einer kleinen Stichprobe setzt dieser Test nicht voraus, dass x1 und x2 einer Normalverteilung folgen.

```
wilcox.test(x1,x2,paired=TRUE)
```

```
 Wilcoxon signed rank exact test

data:  x1 and x2
V = 1, p-value = 0.003906
alternative hypothesis: true location shift is not equal to 0
```

- Es wurde ein zweiseitiger Test durchgeführt.
- Die Hypothese, x1 und x2 folgen der gleichen Verteilung, wird nicht abgelehnt.
- Die Hypothese, die Differenz zwischen x1 und x2 sei 1, kann verworfen werden.
- Der Test setzt nicht voraus, dass x1 und x2 einer Normalverteilung folgen.
- Der Test kann auch bei einer kleinen Anzahl an Beobachtungen angewendet werden.

```
set.seed(124)
N<-10
x1<-rnorm(N,5,3)
x2<-x1-rnorm(N,5,3)
```

```
t.test(x1,x2,paired=TRUE)
```

```
 Paired t-test

data:  x1 and x2
t = 4.9339, df = 9, p-value = 0.000809
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 2.395042 6.450793
sample estimates:
mean difference
       4.422917
```

- Die Mittelwerte von x1 und x2 sind signifikant voneinander verschieden.
- Die Hypothese, die Differenz zwischen x1 und x2 sei im Erwartungswert 8, kann verworfen werden.
- Die Hypothese, x1 sei im Erwartungswert um 8 Einheiten größer als x2, wird nicht abgelehnt.
- Bei einer kleinen Stichprobe setzt dieser Test voraus, dass x1 und x2 einer Normalverteilung folgen.
- Es wurde ein paarweiser Test durchgeführt.

```
wilcox.test(x1,x2,paired=TRUE)


 Wilcoxon signed rank exact test

data:  x1 and x2
V = 55, p-value = 0.001953
alternative hypothesis: true location shift is not equal to 0
```

- Es wurde ein einseitiger Test durchgeführt.
- Die Hypothese, x1 und x2 folgen der gleichen Verteilung, wird abgelehnt.
- Die Hypothese, die Differenz zwischen x1 und x2 sei im Mittel 1, kann verworfen werden.
- Der Test setzt voraus, dass x1 und x2 einer Normalverteilung folgen.
- Bei einer kleinen Anzahl an Beobachtungen sollte man besser einen t-Test durchführen.

Anhang 8.B Übungen

Übung 8.2 *Betrachten Sie den Datensatz `HairEyeColor`. Testen Sie die Hypothese, dass alle vier Augenfarben gleichwahrscheinlich sind.*

`apply` wendet eine Funktion entlang einer Dimension eines Arrays an. `chisq.test` führt einen χ^2 Anpassungstest durch.

```
colors <- apply(HairEyeColor,2,sum)
chisq.test(colors)


 Chi-squared test for given probabilities

data:  colors
X-squared = 133.47, df = 3, p-value < 2.2e-16
```

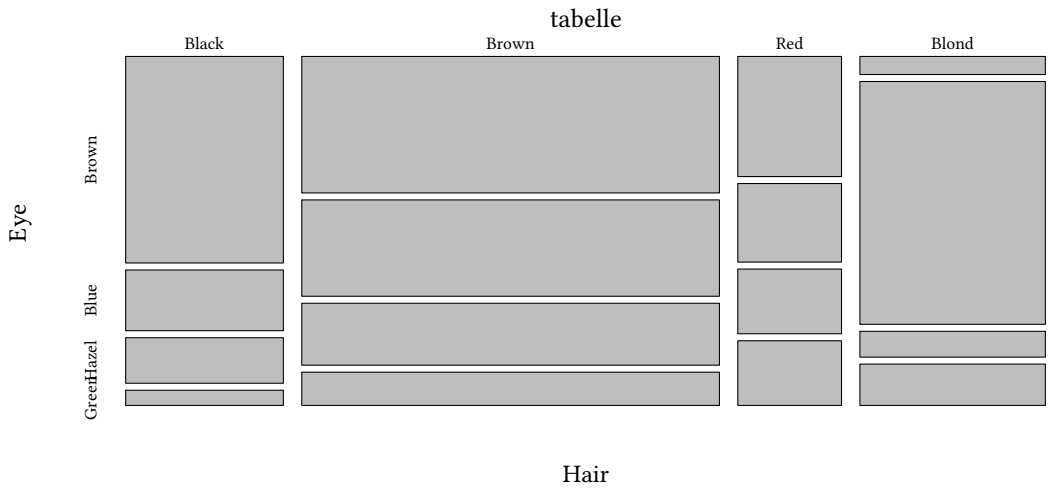
```
chisq.test(colors,p=c(1,1,1,1)/4)


 Chi-squared test for given probabilities

data:  colors
X-squared = 133.47, df = 3, p-value < 2.2e-16
```

Übung 8.3 *Betrachten Sie wieder den Datensatz `HairEyeColor`. Testen Sie die Hypothese, dass die Augenfarbe unabhängig von der Haarfarbe ist.*

```
tabelle <- apply(HairEyeColor,c(1,2),sum)
chisq.test(tabelle)


 Pearson's Chi-squared test

data:  tabelle
X-squared = 138.29, df = 9, p-value < 2.2e-16
```

Ein Mosaicplot ist gut geeignet, Häufigkeitstabellen darzustellen:

```
mosaicplot(tabelle)
```



Übung 8.4 *Betrachten Sie wieder den Datensatz `sleep` und prüfen Sie, ob die beiden Schlafmittel die gleiche Wirksamkeit haben.*

Übung 8.5 *Betrachten Sie wieder den Datensatz `Wages` aus der Bibliothek `Ecdat`. Testen Sie jetzt mit einem Wilcoxon Rangsummentest, ob Männer im Durchschnitt den gleichen Lohn erhalten wie Frauen.*

Übung 8.6 *Verwenden Sie wieder den Datensatz `UCBAdmissions` und testen Sie, ob Männer und Frauen mit gleicher Wahrscheinlichkeit zum Studium an der UCB zugelassen werden.*

Es sieht also so aus, als wäre die Chancen von Frauen schlechter, als die von Männern. Betrachten wir nochmals die Annahmequote, diesmal getrennt nach Departments:

```
par(mar=c(4,4,1,0))
mosaicplot(aperm(UCBAdmissions,c(3,1,2)),dir=c("h","h","v"),las=1,main="",col=gray(c(.7,1)))
```

Da die gewählten Studienfächer von Frauen und Männern sehr unterschiedlich sind, sollten wir vielleicht besser den Test getrennt für jedes Department durchführen.

```
apply(UCBAdmissions,3,chisq.test)
```

Wenn wir uns den Output dieser χ^2 Tests ansehen, stelle wir fest: Es gibt nur ein Department mit einem signifikanten Unterschied. In diesem Department sind die Chancen der Frauen allerdings signifikant besser.

Die oben befürchtete Ungleichbehandlung liegt nur an der unterschiedlichen Verteilung von Frauen und Männern auf verschiedene Studienfächer. Frauen wählen eher Fächer, bei denen die Chancen zugelassen zu werden kleiner sind.

```
UCBAdmissions[,,"A"]

          Gender
Admit      Male Female
  Admitted  512     89
  Rejected  313     19
```

Übung 8.7 *Sie vergleichen zwei Düngemittel für Salatköpfe:* X *und* Y*. In Ihrer Entwicklungsabteilung sind jeweils 30 Salatköpfe mit* X *und 30 Salatköpfe mit* Y *behandelt worden. Das Gewicht der Salatköpfe hat Ihr Assistent bereits in den Variablen* `x` *und* `y` *eingetragen. Ihr Assistent hat ferner zwei Tests mit* `R` *durchgeführt und präsentiert Ihnen den folgenden Output:*

Test 1:

```
> t.test(x,y,paired=TRUE)
        Paired t-test
data:  x and y
t = -2.0558, df = 29, p-value = 0.0489
alternative hypothesis: true difference in
                    means is not equal to 0
95 percent confidence interval:
 -22.02793284  -0.05680469
sample estimates:
mean of the differences
              -11.04237
```

Test 2:

```
> t.test(x,y)
        Welch Two Sample t-test
data:  x and y
t = -2.0203, df = 31.547, p-value = 0.05191
alternative hypothesis: true difference in
                    means is not equal to 0
95 percent confidence interval:
 -22.18182418   0.09708665
sample estimates:
mean of x mean of y
 23.28710  34.32947
```

1. Welche Annahmen werden von beiden obigen Tests vorausgesetzt?

2. Nehmen Sie an, dass diese Annahmen erfüllt sind. Welcher der beiden Tests ist in diesem Fall angemessen? Warum?

3. Ihre Nullhypothese ist, dass die beiden Düngemittel das Gewicht nicht in unterschiedlicher Weise beeinflussen. Ihr Signifikanzniveau ist 5%. Können Sie mit dem Test, den Sie in Teilaufgabe (b) ausgewählt haben, Ihre Nullhypothese ablehnen? Warum?

4. Ihr Assistent präsentiert nun zwei weitere Tests. Sind diese Tests grundsätzlich besser oder weniger gut geeignet als Test 1 oder 2 aus dieser Aufgabe, um Ihre Hypothese zu überprüfen?

   Test 3:

   ```
   > wilcox.test(x,y)
           Wilcoxon rank sum test
   data:  x and y
   W = 220, p-value = 0.0005109
   alternative hypothesis: true location shift
                           is not equal to 0
   ```

   Test 4:

   ```
   > wilcox.test(x,y,paired=TRUE)
           Wilcoxon signed rank test
   data:  x and y
   V = 111, p-value = 0.01130
   alternative hypothesis: true location shift
                           is not equal to 0
   ```

5. Welcher der beiden Tests ist hier angemessen?

6. Ihre Nullhypothese ist, dass die beiden Düngemittel das Gewicht nicht in unterschiedlicher Weise beeinflussen. Ihr Signifikanzniveau ist 5%. Können Sie mit dem Test, den Sie in Teilaufgabe (f) ausgewählt haben, Ihre Nullhypothese ablehnen? Warum?

Übung 8.8 *Wir fragen uns, ob Gummibärchen verschiedener Farben in einer Tüte gleichverteilt sind. Dazu wurde der Inhalt einer Tüte ausgezählt. Die Tüte enthielt 150 Gummibärchen, davon waren 30 weiß, 20 gelb, 40 rot, 30 grün und 30 orange.*

1. *Wie sieht die (theoretische) Verteilung aus, wenn die Bärchen verschiedener Farben tatsächlich gleichverteilt sind?*

2. *Wie lauten die richtigen Hypothesen für das Testproblem?*

3. *Die Testfunktion ergibt für die oben beschriebene Tüte...*

4. *Nehmen Sie an, Sie hätten oben eine Teststatistik* $g = 10$ *erhalten. Kann damit die Hypothese, die Gummibärchen seien gleichverteilt, abgelehnt werden?*

Übung 8.9 *An einer Klausur nahmen* 300 *Studenten teil. Dabei gab es die folgende Notenverteilung:*

Note	*1*	*2*	*3*	*4*	*5*
Anzahl Studenten	*39*	*51*	*73*	*66*	*71*

Wir wollen testen, ob die Noten gleichverteilt sind.

1. *Welche theoretischen Wahrscheinlichkeiten müssen wir für die einzelnen Noten annehmen?*

2. *Wie testen wir zu einem Signifikanzniveau von* 5%*, ob die Noten gleichverteilt sind?*

3. *Eine andere Klausur führt zu folgender Verteilung:*

Note	*1*	*2*	*3*	*4*	*5*
Anzahl Studenten	*44*	*57*	*68*	*65*	*66*

   Wir führen erneut einen Test zur Gleichverteilung der Noten durch. Das Signifikanzniveau ist 5%*. Sind die Noten gleichverteilt?*

Übung 8.10 *Eine Umfrage zu den Essgewohnheiten von 1000 zufällig Befragten ergab folgendes Ergebnis:*

	Regelmäßigkeit der Mahlzeiten			
Gewicht	*regelmäßig*	*leicht unregelmäßig*	*stark unregelmäßig*	$\sum$
Normalgewicht	*350*	*150*	*100*	*600*
Über- und Untergewicht	*100*	*50*	*250*	*400*
$\sum$	*450*	*200*	*350*	*1000*

Es soll zum Signifikanzniveau $\alpha = 10\%$ *getestet werden, ob eine Abhängigkeit zwischen dem Gewicht der Testpersonen und der Regelmäßigkeit der Mahlzeiten vorliegt.*

1. *Welche Hypothesen müssen Sie aufstellen?*

2. *Wie lautet der Wert der Teststatistik?*

3. *Kann die Nullhypothese angenommen werden?*

Übung 8.11 *Es gibt Menschen, die unter Schlafstörungen, d.h. Abweichungen vom gesunden Schlafverhalten, leiden. Eine Form der Schlafstörung ist das vorzeitige Erwachen, ohne danach wieder einschlafen zu können. Sie arbeiten in einem Schlaflabor und erheben für 16 Patienten die Schlafdauer vor* (x) *und nach* (y) *der Einnahme eines Medikaments. Ihr Assistent hat zwei Tests in R durchgeführt und präsentiert ihnen folgende Ergebnisse:*

- ```
  t.test(x, y, paired = TRUE)
  ```

  ```

   Paired t-test

  data:  x and y
  t = -4.3188, df = 15, p-value = 0.0006084
  alternative hypothesis: true mean difference is not equal to 0
  95 percent confidence interval:
   -3.351107 -1.136393
  sample estimates:
  mean difference
         -2.24375
  ```

- ```
  t.test(x,y)
  ```

  ```

   Welch Two Sample t-test

  data:  x and y
  t = -4.3764, df = 24.221, p-value = 0.0001993
  alternative hypothesis: true difference in means is not equal to 0
  95 percent confidence interval:
   -3.301381 -1.186119
  sample estimates:
  mean of x mean of y
    3.53125   5.77500
  ```

1. *Welche Annahmen werden von beiden obigen Tests vorausgesetzt?*
2. *Nehmen Sie an, diese Annahmen sind erfüllt. Welcher der beiden Test ist in diesem Fall angemessen?*
3. *Ihre Nullhpothese ist, dass das Medikament keinen Einfluss auf die Schlafdauer Ihrer Patienten hat. Ihr Signifikanzniveau ist 5%. Welche Antwort ist korrekt?*

```
wilcox.test(x,y,paired=TRUE)
```

```

 Wilcoxon signed rank test with continuity correction

data:  x and y
V = 8, p-value = 0.002079
alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(x,y)


 Wilcoxon rank sum test with continuity correction

data:  x and y
W = 39, p-value = 0.000845
alternative hypothesis: true location shift is not equal to 0
```

Übung 8.12 *Knorz kann auf zwei verschiedene Arten hergestellt werden. Sie haben an 9 Tagen den Output des ersten Verfahrens gemessen. Das Ergebnis (die 9 Zahlen) schreiben Sie in die Variable* x*. Sie haben an 11 weiteren Tagen des Output des zweiten Verfahrens gemessen. Das Ergebnis (die 11 Zahlen) schreiben Sie in die Variable* y*. Nun führen Sie einen Mann-Whitney* U *Test durch. Ihr Signifikanzniveau ist 10%. Sie erhalten folgendes Ergebnis:*

```
 Wilcoxon rank sum test
data:  x and y
W = 13, p-value = 0.9734
alternative hypothesis: true location shift is not equal to 0
```

Was ist Ihre Schlussfolgerung?

- Die Verfahren führen zu einem signifikant verschiedenen Output.
- Bei einer so kleinen Stichprobe sollte man besser einen t-Test durchführen.
- Die gemessenen Unterschiede sind klein und deuten nicht auf einen signifikanten Unterschied hin.
- Mit den Ihnen vorliegenden Daten sollte man besser einen paarweisen Test durchführen.
- Bei einem kleineren Signifikanzniveau würde man einen signifikanten Unterschied finden.

Übung 8.13 *Sie beobachten, dass von 40 Männern 8 Produkt* A *kaufen, weitere 8 kaufen Produkt* B*, und 24 kaufen Produkt* C*. Von 60 Frauen kaufen 12 Produkt* A*, 12 kaufen Produkt* B*, und 36 kaufen Produkt* C*. Ihre Nullhypothese* H_0 *ist, dass Präferenzen für* A*,* B*, und* C *unabhängig vom Geschlecht sind.*

1. Wie groß ist Ihre Teststatistik?
2. Nehmen Sie an, Sie hätten in der obigen Aufgabe eine Teststatistik von 10 berechnet. Wie berechnen Sie den p-Wert für Ihren Test?

Übung 8.14 *Um den Beratungserfolg zu messen, vergleich eine Unternehmensberatung bei 12 Unternehmen den Gewinn jeweils im Jahr vor und nach der Umstrukturierung der von ihr beratenen Unternehmen. Der Gewinn nach der Beratung steht in der Variablen* x*. Der Gewinn vor der Beratung steht in* y*. Nun führen Sie einen einseitigen Wilcoxon Test durch. Ihr Signifikanzniveau* α *ist 1%. Sie erhalten folgendes Ergebnis:*

```
 Wilcoxon signed rank test
data:  x and y
V = 72, p-value = 0.003418
alternative hypothesis: true location shift is greater than 0
```

Was ist Ihre Schlussfolgerung?

- Die Gewinne nach Beratung sind signifikant größer.
- Bei einer so kleinen Stichprobe ist es besser, einen zweiseitigen Test durchzuführen.
- Die gemessenen Unterschiede sind klein und deuten nicht auf einen signifikanten Unterschied hin.
- Wenn die Alternativhypothese ist, der Gewinn sei nach der Beratung größer als vorher, ist es besser, einen zweiseitigen Test durchzuführen.
- Bei einem größeren Signifikanzniveau (α wird also größer, das muss nicht besser sein!) würde man keinen signifikanten Unterschied mehr finden.

Anhang 8.C Fisher's exakter Test

Die oben angegebene Formel

$$\sum_{i=1}^{n}\sum_{j=1}^{k}\frac{(x_{ij}-e_{ij})^2}{e_{ij}} \sim \chi^2_{(n-1)\cdot(k-1)}$$

ist nur für große n hinreichend genau. Bei kleinen n ist der χ^2 Test ungenau und man verwendet deshalb Fisher's exakten Test.

Es heißt, Fisher sei durch das folgende Problem inspiriert worden: Eine englische Dame (Muriel Bristol) behauptet, sie könne schmecken, ob in ihrem Tee die Milch vor oder nach dem Tee eingegossen wurde. Ihr wird 4 mal Tee angeboten, in den Milch zuerst eingegossen wurde, weitere 4 mal erhält sie Tee in den Milch zuletzt eingegossen wurde. In jeweils 3 Fällen liegt sie richtig:

```
Tea <- cbind(c(3,1),c(1,3))
fisher.test(Tea)
```

```
 Fisher's Exact Test for Count Data

data:  Tea
p-value = 0.4857
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
   0.2117329 621.9337505
sample estimates:
odds ratio
  6.408309
```

```
chisq.test(Tea)
```

```
 Pearson's Chi-squared test with Yates' continuity correction

data:  Tea
X-squared = 0.5, df = 1, p-value = 0.4795
```

Wir sehen, die berechneten p-Werte weichen etwas voneinander ab. Fisher's exakter Test kann sogar ein bisschen mehr: Er unterscheidet zwischen einer positiven und einer negativen Korrelation (die Dame könnte ja auch systematisch falsch liegen).

```
fisher.test(Tea,alternative="greater")
```

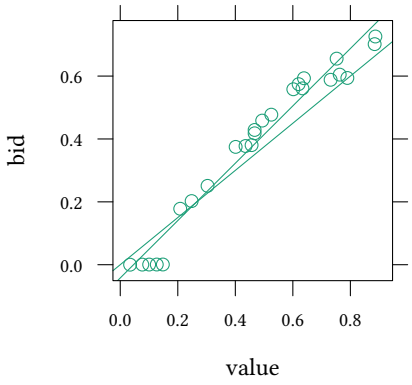
```
 Fisher's Exact Test for Count Data

data:  Tea
p-value = 0.2429
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.3135693       Inf
sample estimates:
odds ratio
  6.408309
```

9. Lineare Regression — Einführung

9.1. Motivation

Praxis Erstpreisauktionen sind Auktionen mit versiegelten Geboten in denen der Bieter mit dem höchsten Gebot die Auktion gewinnt und sein eigenes Gebot zahlt (weil der höchste Preis in einer sortierten Liste der Preise der »erste« ist, nennt man diese Auktion auch »Erstpreisauktion«).

Es ist möglich, für das Bietverhalten in Erstpreisauktionen eine Gleichgewichtsbietfunktion abzuleiten. Die folgende Grafik zeigt das Bietverhalten im Experiment.



Cox, J. C., V. L. Smith, and J. M. Walker, *Journal of Risk and Uncertainty*, 1988. Kirchkamp, Reiß, *Economic Journal*, 2011.

Forschungsfragen:

- Warum sind Gebote für hohe Werte höher als theoretisch vorhergesagt?
- Warum sind Gebote für niedrige Werte niedriger als theoretisch vorhergesagt?

Theorie Bislang:

- Vergleich einer Stichprobe mit einer vorgegebenen Verteilung (t-Test für Mittelwerte, χ^2-Test für Häufigkeiten)
- Vergleich zweier Stichproben (t-Test, Wilcoxon Test, Mann-Whitney U Test)
- Vergleich zweier Merkmale (χ^2-Test für Häufigkeiten).

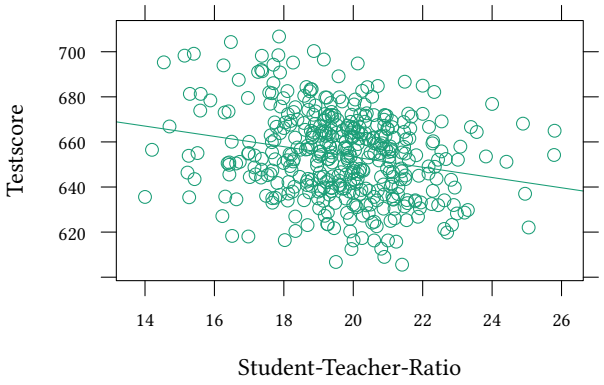
Die Fragestellung war »einfach«. Entweder hatte ein Beobachtung ein bestimmtes Merkmal (bzw. stammte aus einer Stichprobe), oder sie hatte dieses Merkmal nicht (Männer/Frauen, fehlerhaft/nicht-fehlerhaft, erfolgreich/nicht-erfolgreich,…) Jetzt:

- Messe den Einfluss von Faktoren, die viele verschiende Werte annehmen können (wie beeinflusst eine (unabhängige) Variable X eine abhängige Variable Y)

9.2. Lineare Regression

```
data(Caschool,package="Ecdat")
attach(Caschool)
plot(testscr~str)
abline(lm(testscr~str))
```

420 Kalifornische Schulen 1998/99:



$$\texttt{testscr} = 698.9 - 2.28 \cdot \texttt{str} + u$$

Allgemein:

$$Y_i = \beta_0 + \beta_1 \cdot X_i + u_i \qquad i = 1, \ldots, n$$

- Y abhängige Variable (erklärte Variable, endogen)
- X unabhängige Variable (erklärende Variable, exogen, Regressor)
- β_0 Achsenabschnitt
- β_1 Steigung
- u Fehlerterm
  (andere Faktoren die Y beeinflussen)

Wie könnte man β_0 und β_1 schätzen?

Erinnern wir uns: Der Mittelwert $\bar{X}$ war der Kleinste-Quadrate-Schätzer für den Erwartungswert μ_X.

Minimiere ebenfalls quadratische Abstände bei der Bestimmung von β_0 und β_1:

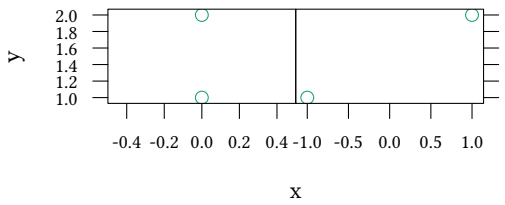
$$KQ = \sum_{i=1}^{n} (Y_i - (\beta_1 X_i + \beta_0))^2 \qquad \min_{\beta_0, \beta_1} KQ$$

Nach erfolgreicher Minimierung (siehe Anhang 9.B zu diesem Kapitel) erhalten wir

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

↑ Kleinste-Quadrate-Schätzer, KQ-Schätzer, OLS

$\hat{\beta}_0$ und $\hat{\beta}_1$ können berechnet werden wenn $\sum_{i=1}^{n} (X_i - \bar{X})^2 \neq 0$, d.h. es gibt mindestens zwei Beobachtungen mit verschiedenen x-Werten (man braucht zwei Punkte um eine Gerade zu beschreiben)



Natürlich muss man diese Werte nicht von Hand ausrechnen. R kann das für uns erledigen `lm` schätzt eine OLS Regression. Das Ergebnis wird in einer Variablen (hier `est`) abgespeichert.

Wenn man das Ergebnis sehen will, muß es angezeigt werden, z.B. mit `summary(est)`.

Das Ergebnis kann aber auch graphisch dargestellt werden, z.B. mit `abline(est)`.

Wir können das Ergebnis einer OLS Schätzung in einer Variablen abspeichern, wie hier:

```
data(Caschool,package="Ecdat")
```

```
est<-lm(testscr~str, data=Caschool)
est
```

```


Call:
lm(formula = testscr ~ str, data = Caschool)

Coefficients:
(Intercept)          str
     698.93        -2.28
```

Die (knappe) Anzeige des Regressionsobjektes zeigt längst nicht alle Details, die R ausgerechnet hat. Wenn wir etwas mehr über die Details der Schätzung erfahren wollen, hilft das Kommando `summary`:

```
est <- lm(testscr~str, data=Caschool)
summary(est)
```

```


Call:
lm(formula = testscr ~ str, data = Caschool)

Residuals:
    Min      1Q  Median      3Q     Max
-47.727 -14.251   0.483  12.822  48.540

Coefficients:
            Estimate Std. Error t value   Pr(>|t|)
(Intercept) 698.9330     9.4675  73.825    < 2e-16 ***
str          -2.2798     0.4798  -4.751 0.00000278 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.58 on 418 degrees of freedom
Multiple R-squared:  0.05124, Adjusted R-squared:  0.04897
F-statistic: 22.58 on 1 and 418 DF,  p-value: 0.000002783
```

```
library(MCMCpack)
summary(MCMCregress(testscr~str, data=Caschool))


Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

               Mean      SD Naive SE Time-series SE
(Intercept) 698.963  9.4739 0.094739       0.094739
str          -2.281  0.4809 0.004809       0.004809
sigma2      346.985 24.2723 0.242723       0.242723

2. Quantiles for each variable:

              2.5%     25%     50%     75%   97.5%
(Intercept) 680.43 692.599 699.020 705.329 717.594
str          -3.22  -2.602  -2.286  -1.957  -1.343
sigma2      302.32 330.080 345.898 362.825 397.550
```

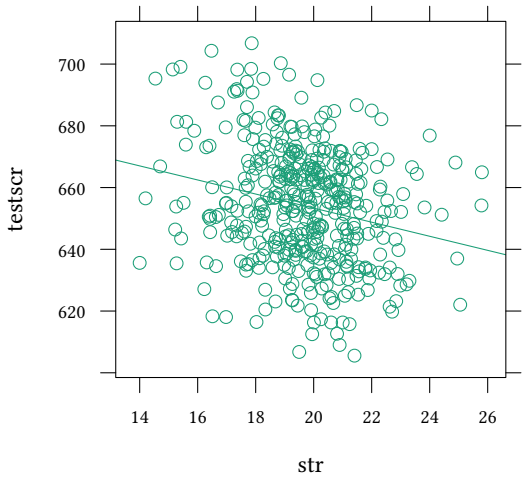
Ergebnis der OLS Schätzung: $\hat{\beta}_0 = 698.93$, $\hat{\beta}_1 = -2.28$.

- Approximation von Y durch $\hat{Y}$:

$$\hat{Y}_i = \hat{\beta}_1 X_i + \hat{\beta}_0 \qquad i = 1, \dots, n$$

- Residuen

$$\hat{u}_i = Y_i - \hat{Y}_i \qquad i = 1, \dots, n$$



In unserem Fall ist die Approximation also

$$\texttt{testsrc} = -2.28 \cdot \texttt{str} + 698.93$$

Damit ergibt sich der marginale Effekt von `str` zu

$$\frac{\Delta \texttt{testsrc}}{\Delta \texttt{str}} = -2.28$$

9.3. Bestimmtheitsmaß R^2

- R^2 relativer Anteil der Varianz von Y, der durch X erklärt wird.

$$Y_i = \hat{Y}_i + \hat{u}_i = \text{OLS Approximation} + \text{OLS Residuen}$$

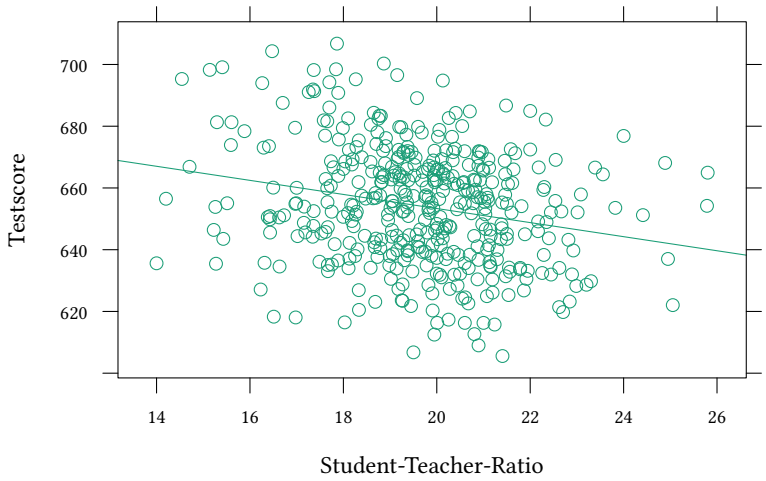
$$\text{var}(Y_i) = \text{var}(\hat{Y}_i) + \text{var}(\hat{u}_i)$$

$$R^2 = \frac{\text{var}(\hat{Y}_i)}{\text{var}(Y_i)}$$

$$0 \leq R^2 \leq 1$$

- Bei Regressionen mit einem einzigen Regressor X ist R^2 der quadrierte Korrelationskoeffizient zwischen X und Y.

Was bedeutet es, wenn R^2 im Beispiel nur 0.05 ist?



Hier sind drei Wege, das R^2 in unserem Beispiel zu finden:

1. Der komfortable: `summary` zeigt uns alles was wir vielleicht wissen wollen über unser Schätzobjekt an:

```
est <- lm(testscr~str, data=Caschool)
summary(est)
```

```
Call:
lm(formula = testscr ~ str, data = Caschool)

Residuals:
    Min      1Q  Median      3Q     Max
-47.727 -14.251   0.483  12.822  48.540

Coefficients:
            Estimate Std. Error t value   Pr(>|t|)
(Intercept) 698.9330     9.4675  73.825    < 2e-16 ***
str          -2.2798     0.4798  -4.751 0.00000278 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.58 on 418 degrees of freedom
Multiple R-squared:  0.05124, Adjusted R-squared:  0.04897
F-statistic: 22.58 on 1 and 418 DF,  p-value: 0.000002783
```

2. Der einfache (klappt aber nur in bestimmten Fällen): `cor` zeigt nur den Korrelationskoeffizienten (und der ist bei einer Regression mit einer Variablen und einem Achsenabschnitt gleich dem R).

```
cor(testscr,str)^2

[1] 0.0512401
```

3. Der spartanische: Schließlich können wir aus dem `summary`-Objekt das R^2 extrahieren:

```
summary(est)$r.squared

[1] 0.0512401
```

9.4. SER

Ein weiteres Maß für die Genauigkeit unserer Approximation ist SER:

- Standardfehler der Residuen $\text{SER} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2}$

```
est <- lm(testscr~str, data=Caschool)
sqrt(with(est,sum(residuals^2)/df.residual))

[1] 18.58097

summary(est)$sigma

[1] 18.58097
```

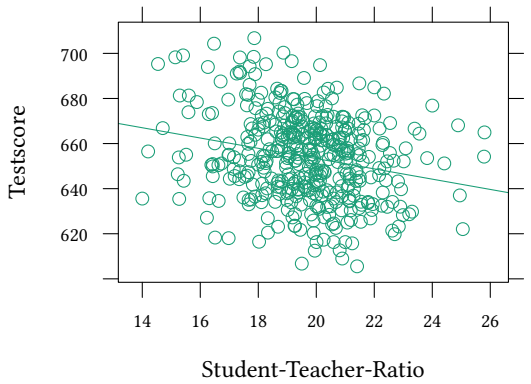
```
est <- lm(testscr~str, data=Caschool)
summary(est)


Call:
lm(formula = testscr ~ str, data = Caschool)

Residuals:
    Min      1Q  Median      3Q     Max
-47.727 -14.251   0.483  12.822  48.540

Coefficients:
            Estimate Std. Error t value   Pr(>|t|)
(Intercept) 698.9330     9.4675  73.825    < 2e-16 ***
str          -2.2798     0.4798  -4.751 0.00000278 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.58 on 418 degrees of freedom
Multiple R-squared:  0.05124, Adjusted R-squared:  0.04897
F-statistic: 22.58 on 1 and 418 DF,  p-value: 0.000002783
```

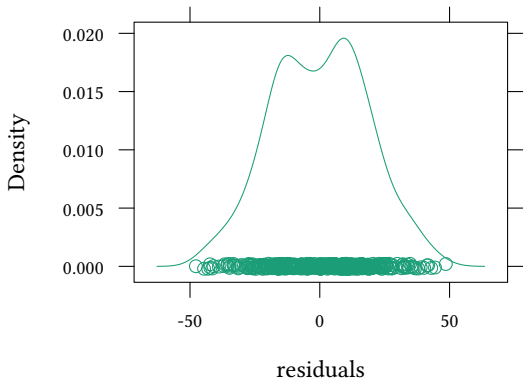
9.5. Die Verteilung des OLS Schätzers

Um mehr über die Verteilung unseres Schätzers zu lernen, wäre es schön, wenn wir einige weitere Stichproben von jeweils 420 Kalifornischen Schulen hätten. Das haben wir leider nicht, aber wir können solche Stichproben simulieren.

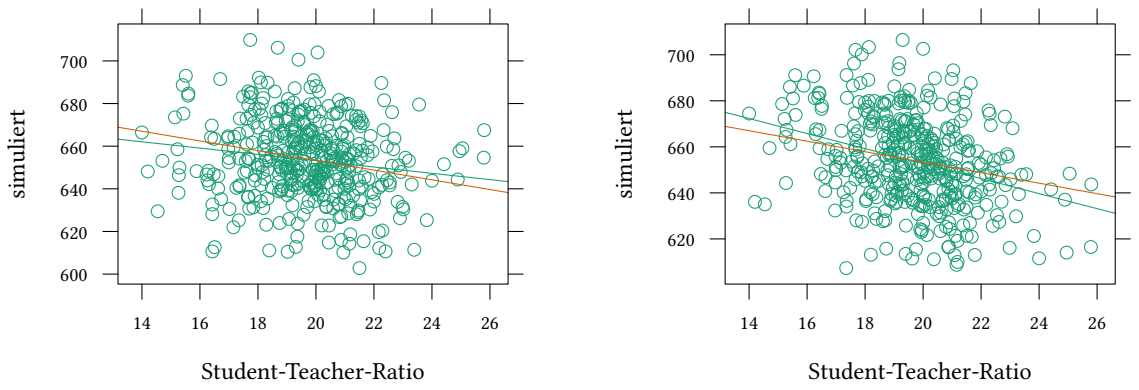
Das machen wir wir folgt: Wir stellen uns vor, der oben geschätzte Zusammenhang wäre der wahre Zusammenhang und die von uns geschätzte Verteilund des Störterm $\hat{u}$ wäre die wahre Verteilung des Störterms. Wir approximieren die Verteilung des Schätzers in unserem Beispiel. Dazu schätzen wir wiederholt mit immer anderen Permutationen des Störterms u.

Annahme: Die geschätzten Residuen bilden die Verteilung der Residuen in der Population einigermaßen gut ab.





Verwende nun die geschätzen Residuen um weitere Stichproben zu simulieren:



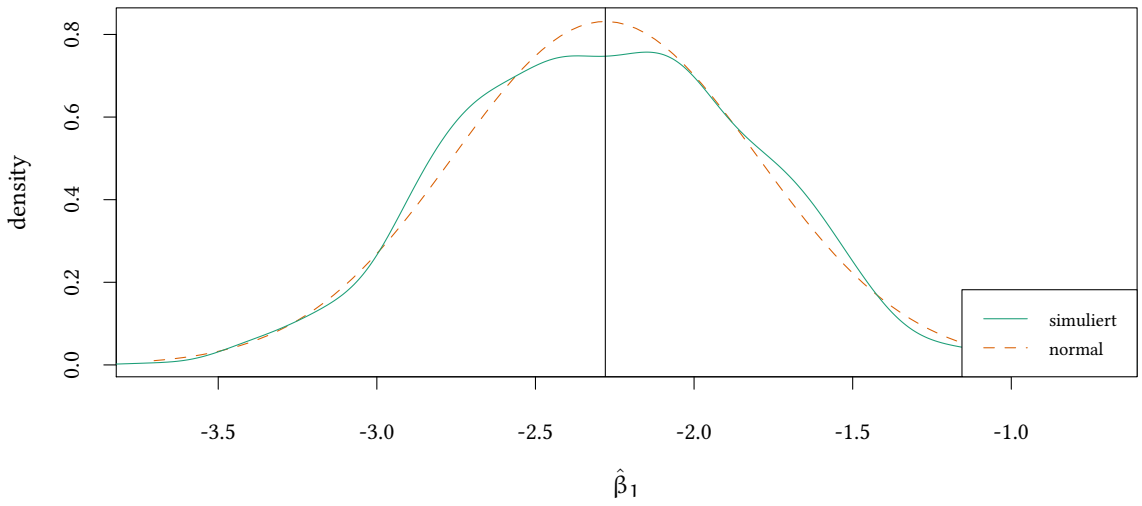
`replicate` führt einen Ausdruck mehrmals aus.

`coef` extrahiert die Koeffizienten aus einer Regression.

`simulate` berechnet unsere abhängige Variable Y auf Basis unserer Schätzung neu. Im Unterschied zu unserer Schätzung werden jetzt die zufälligen Störterme u_i zufällig anders verteilt. Simulate gibt uns also eine Y Variable, die wir »auch hätten ziehen können«.

`density` schätzt eine Dichtefunktion.

```
set.seed(123)
N<-1000
est <- lm(testscr ~ str)
coefdist <- replicate(N,coef(lm (unlist(simulate(est)) ~ str)))
```



- $\hat{\beta}_0$ und $\hat{\beta}_1$ werden mit Hilfe der Stichprobe berechnet. Eine andere Stichprobe (andere Störterme u) ergibt auch andere Werte für $\hat{\beta}_0$ und $\hat{\beta}_1$.

- Genauso wie $\bar{Y}$ gibt es auch für $\hat{\beta}_0$ und $\hat{\beta}_1$ eine Verteilung.
- Eine Beispiel für diese Verteilung sehen wir im Bild oben. Es ist nicht genau eine Normalverteilung. Wenn wir eine größere Anzahl von Simulationen wählen, dann kommen wir einer Normalverteilung näher.

Wir stellen uns nun die folgenden Fragen

- Ist $E(\hat{\beta}_1) = \beta_1$? (OLS ist unverzerrt?)
- Ist $\text{var}(\hat{\beta}_1)$ klein? (OLS ist effizient?)
- Wie testen wir Hypothesen? (z.B. $\beta_1 = 0$)
- Wie bestimmen wir ein Konfidenzintervall für β_0 und β_1?

Mittelwert und Varianz von $\hat{\beta}_1$

Der Mittelwert unserer Verteilung von geschätzten β_1 ist

```
mean(coefdist["str",])
```

```
[1] -2.27974
```

Das ist ziemlich ähnlich wie das β_1 aus unserem Modell:

```
coef(est)["str"]
```

```
      str
-2.279808
```

Ist das Zufall, oder können wir davon ausgehen, dass unsere Schätzung im Mittel immer richtig liegt? In Anhang 9.C zeigen wir:

Erwartungstreue

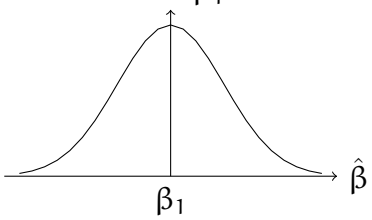
Unter den drei Standard-OLS Annahmen

1. $E(u_i|X_i = x) = 0$
2. (X_i, Y_i) sind i.i.d. (independent identically-distributed)

((X_i, Y_i) ist unabhängig von (X_j, Y_j), identisch verteilt)

3. Große Ausreißer in X und Y sind selten (die vierten Momente von X und Y existieren).

gilt: $\hat{\beta}_1$ ist ein **unverzerrter Schätzer** von β_1




```
summary(est)


Call:
lm(formula = testscr ~ str)

Residuals:
    Min      1Q  Median      3Q     Max
-47.727 -14.251   0.483  12.822  48.540

Coefficients:
            Estimate Std. Error t value   Pr(>|t|)
(Intercept) 698.9330     9.4675  73.825    < 2e-16 ***
str          -2.2798     0.4798  -4.751 0.00000278 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.58 on 418 degrees of freedom
Multiple R-squared:  0.05124, Adjusted R-squared:  0.04897
F-statistic: 22.58 on 1 and 418 DF,  p-value: 0.000002783
```

In Anhang 9.D zeigen wir

$$\operatorname{var}(\hat{\beta}_1) \approx \frac{1}{n} \frac{\operatorname{var}((X_i - \mu_X) \cdot u_i)}{\sigma_X^4}$$

In unserer Schätzung haben wir für $\hat{\beta}_1$ eine Standardabweichung von 0.479826, bzw. eine Varianz von 0.230233 gesehen.

In unserer Simulation hatten die Koeffizienten von `str` eine Varianz von

```
var(coefdist["str",])

[1] 0.219474

sd(coefdist["str",])

[1] 0.4684805
```

Das ist nicht genau dasselbe, aber in etwa die gleiche Größenordnung. Wenn wir mehr Stichproben simulieren, dann kommen wir der geschätzen Varianz näher. Diese Formel wird auch von der Funktion `summary` in der obigen Regression verwendet.

Wir fassen diese Beobachtungen wie folgt zusammen:

9.6. OLS Annahmen

Wenn die drei OLS Annahmen gelten,...

1. $E(u_i|X_i = x) = 0$
2. (X_i, Y_i) sind i.i.d.
3. Große Ausreißer in X und Y sind selten (die vierten Momente von X und Y existieren).

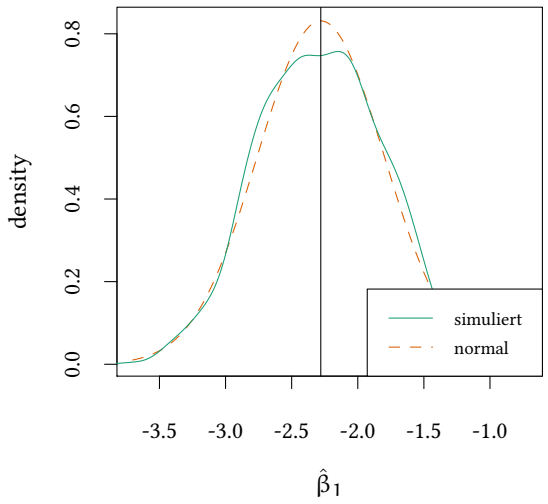
→ ...dann gilt auch...

- $E(\hat{\beta}_1) = \beta_1$ ($\hat{\beta}_1$ ist unverzerrt)
- $\text{var}(\hat{\beta}_1) = \frac{1}{n}\frac{\text{var}((X_i-\mu_X)u_i)}{\sigma_X^4}$

9.6.1. Verteilung von $\hat{\beta}_1$

Wenn n groß ist, dann ist $\hat{\beta}$ etwa normalverteilt (zentraler Grenzwertsatz).

- $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\text{var}((X_i-\mu_X)u_i)}{n\sigma_X^4}\right)$



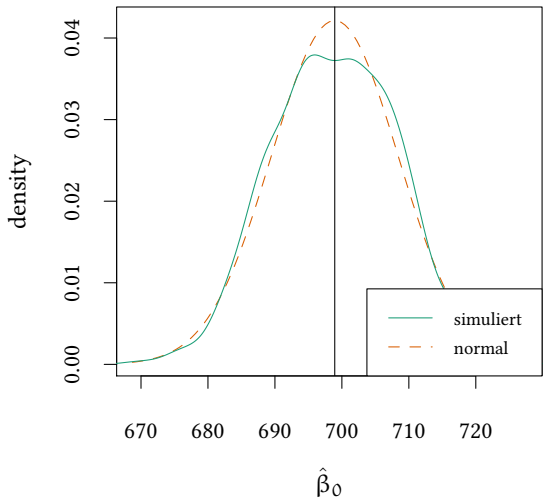
9.6.2. Verteilung von $\hat{\beta}_0$

Auch $\hat{\beta}_0$ ist, bei großem n, normalverteilt mit $\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2_{\hat{\beta}_0}\right)$ wobei

$$\sigma^2_{\hat{\beta}_0} = \frac{1}{n}\frac{\text{var}(H_i u_i)}{\left(E\left(H_i^2\right)\right)^2} \qquad \text{und } H_i = 1 - \left(\frac{\mu_X}{E\left(X_i^2\right)}\right) \cdot X_i$$

↑ Verteilung für β_1 und β_0

→ Hypothesentests und Konfidenzintervalle für β_1 und β_0



9.7. Hypothesentests für $\hat{\beta}_1$

- zweiseitiger Test:

  $H_0 : \beta_1 = \beta_{1,0}$ versus $H_1 : \beta_1 \neq \beta_{1,0}$

- einseitiger Test:

  $H_0 : \beta_1 = \beta_{1,0}$ versus $H_1 : \beta_1 > \beta_{1,0}$

  $H_0 : \beta_1 = \beta_{1,0}$ versus $H_1 : \beta_1 < \beta_{1,0}$

wobei $\beta_{1,0}$ der hypothetische Wert der Nullhypothese ist
Ansatz:

- Konstruiere t-Statistik und bestimme den p-Wert (oder vergleiche die t-Statistik mit dem kritischen Wert aus der Normalverteilung bzw. t-Verteilung.).

- Allgemein:

$$t = \frac{\text{Schätzer} - \text{hypothetischer Wert}}{\text{Standardfehler des Schätzers}}$$

  wobei der Standardfehler des Schätzers aus der geschätzten Varianz des Schätzers hergeleitet wird.

- um den Mittelwert von $\bar{X}$ zu testen:

$$t = \frac{\bar{X} - \mu_{X,0}}{\sigma_X/\sqrt{n}} = \frac{\bar{X} - \mu_{X,0}}{\sigma_{\bar{X}}}$$

- um den Regressionskoeffizienten β_1 zu testen:

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sigma_{\hat{\beta}_1}}$$

Woher nehmen wir die Varianz für den t-Test? Im wirklichen Leben ist uns die theoretische Varianz $\sigma^2_{\hat{\beta}_1}$ nicht bekannt, wir müssen sie schätzen.

Erinnern wir uns an die theoretische Varianz von $\hat{\beta}_1$:

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n}\frac{\operatorname{var}((X_i - \mu_X)u_i)}{\sigma_X^4}$$

Wenn wir σ_X nicht kennen: die geschätzte Varianz:

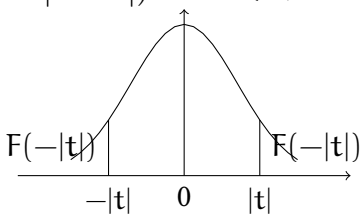
$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n}\frac{\frac{1}{n-2}\sum_{i=1}^{n}\left(\left(X_i - \bar{X}\right)\cdot\hat{u}_i\right)^2}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right)^2}$$

Zusammenfassung: Um die Hypothese $H_0 : \beta_1 = \beta_{1,0}$ versus $H_1 : \beta_1 \neq \beta_{1,0}$ zu testen:

- bestimme die t-Statistik:

$$t^{\text{Stichp.}} = \frac{\hat{\beta}_1 - \beta_{1,0}}{\hat{\sigma}_{\hat{\beta}_1}}$$

- Ablehnung auf 5% falls $|t^{\text{Stichp.}}| > 1.96$
- Der p-Wert ist $p = \Pr\left(|t| > \left|t^{\text{Stichp.}}\right|\right) = 2\cdot F(-|t^{\text{Stichp.}}|)$



- Wir gehen davon aus, dass unsere Teststatistik *normalverteilt* ist. Dazu benötigen wir die Annahme, dass n groß ist ($n = 50$ ist »groß«)

  Viele Statistikprogramme gehen alternativ davon aus, dass unsere Teststatistik t-*verteilt* ist. Dazu muss man voraussetzen, dass die Residuen normalverteilt sind (siehe Abschnitt 9.9.4).

9.8. Konfidenzintervalle für β

Erinnern wir uns an die Formel für Konfidenzintervalle für Mittelwerte:

$$\left[\bar{x} + \hat{\sigma}_{\bar{x}}\cdot Q_t\left(\frac{\alpha}{2}\right), \bar{x} - \hat{\sigma}_{\bar{x}}\cdot Q_t\left(\frac{\alpha}{2}\right)\right] = \left[\bar{x} - \hat{\sigma}_{\bar{x}}\cdot Q_t\left(1-\frac{\alpha}{2}\right), \bar{x} + \hat{\sigma}_{\bar{x}}\cdot Q_t\left(1-\frac{\alpha}{2}\right)\right]$$

Wir gehen genauso vor: $\hat{\beta}_1 = \bar{x}$, $\hat{\sigma}_{\hat{\beta}_1} = \hat{\sigma}_{\bar{x}}$

$$\left[\hat{\beta}_1 + \hat{\sigma}_{\hat{\beta}_1} \cdot Q_t\left(\frac{\alpha}{2}\right), \hat{\beta}_1 - \hat{\sigma}_{\hat{\beta}_1} \cdot Q_t\left(\frac{\alpha}{2}\right)\right] = \left[\hat{\beta}_1 - \hat{\sigma}_{\hat{\beta}_1} \cdot Q_t\left(1-\frac{\alpha}{2}\right), \hat{\beta}_1 + \hat{\sigma}_{\hat{\beta}_1} \cdot Q_t\left(1-\frac{\alpha}{2}\right)\right]$$

`confint` berechnet Konfidenzintervalle für ein geschätztes Modell.

`pnorm` und `pt` berechnen die Verteilungsfunktion der Normalverteilung und der t-Verteilung. `qnorm` und `qt` berechnen die Quantile zu einem gegebenen Verteilungswert.

```
est <- lm(testscr ~ str)
confint(est)

                2.5 %     97.5 %
(Intercept) 680.32313 717.542779
str          -3.22298  -1.336637
```

Dieses *geschätzte* Konfidenzintervall ist den Quantilen aus unserer Simulation sehr ähnlich:

```
quantile(coefdist["str",],c(.025,.975))

     2.5%     97.5%
-3.186735 -1.419616
```

Und es ist auch dem *credible interval* sehr ähnlich:

```
quantile(MCMCregress(testscr ~ str)[,"str"],c(.025,.975))

     2.5%     97.5%
-3.219805 -1.342553
```

Natürlich können wir auch die Konfidenzintervalle von Hand ausrechnen:
`vcov` berechnet die Varianz-Kovarianz Matrix für $\hat{\beta}$ unter der Annahme homoskedastischer Residuen. `diag` nimmt die Diagonale einer Matrix. Bei der Varianz-Kovarianz Matrix sind das die Varianzen der Koeffizienten. `sqrt` berechnet Quadtratwurzeln.

```
coef(est) +  sqrt(diag(vcov(est))) * qnorm(.025)

(Intercept)         str
 680.377010   -3.220249

coef(est) -  sqrt(diag(vcov(est))) * qnorm(.025)

(Intercept)         str
 717.488895   -1.339367
```

Zum Vergleich:

```
confint(est)
```

```
                2.5 %     97.5 %
(Intercept) 680.32313 717.542779
str          -3.22298  -1.336637
```

$t^{\text{Stichp.}} = \frac{\hat{\beta}_1 - \beta_{1,0}}{\hat{\sigma}_{\hat{\beta}_1}}$, $p = 2F(-|t^{\text{Stichp.}}|)$

Auch den p-Wert hatten wir oben mit `summary` schon gesehen. Auch den können wir selbst ausrechnen:

```
2 * pnorm (- abs(coef(est) / sqrt(diag(vcov(est)))))
```

```
  (Intercept)           str
0.000000000000 0.000002020858
```

Gerade haben wir die Approximation durch die Normalverteilung benutzt. R verwendet für `summary` die t-Verteilung. Das können wir natürlich auch:

```
2 * pt (- abs(coef(est) / sqrt(diag(vcov(est)))),est$df.resid)
```

```
  (Intercept)           str
0.000000000000 0.000002783307
```

Wir sehen, die beiden Werte weichen etwas voneinander ab. Beide Methoden konvergieren asymptotisch (wenn die Stichprobe sehr groß wird) zum richtigen Ergebnis. Falls die Störterme i.i.d. normalverteilt sind, liefert die t-Verteilung auch wenn die Stichprobe klein ist das exakte Ergebnis.

Die folgenden beiden Aussagen sind äquivalent:

- Das 95% Konfidenzintervall enthält nicht die Null
- Die Hypothese $\beta_1 = 0$ wird auf dem 5% Niveau abgelehnt

9.9. Verwendung von OLS in der Praxis

9.9.1. Darstellung von OLS Schätzergebnissen:

```
est <- lm(testscr~str, data=Caschool)
summary(est)
```

```

Call:
lm(formula = testscr ~ str, data = Caschool)

Residuals:
    Min      1Q  Median      3Q     Max
-47.727 -14.251   0.483  12.822  48.540
```

```
Coefficients:
            Estimate Std. Error t value   Pr(>|t|)
(Intercept) 698.9330     9.4675  73.825    < 2e-16 ***
str          -2.2798     0.4798  -4.751 0.00000278 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.58 on 418 degrees of freedom
Multiple R-squared:  0.05124, Adjusted R-squared:  0.04897
F-statistic: 22.58 on 1 and 418 DF,  p-value: 0.000002783
```

$$\underset{}{\texttt{testscr} =} \underset{(9.4675)}{698.933} - \underset{(0.4798)}{2.2798} \cdot \texttt{str}, \quad R^2 = 0.05, \text{SER} = 18.58$$

Oft finden wir Standardfehler in Klammern unter den geschätzten Koeffizienten. Diese Darstellung ist praktisch:

- Die geschätzte Regressionsgerade ist $\texttt{testscr} = 698.933 - 2.2798 \cdot \texttt{str}$
- Der Standardfehler von $\beta_0 = 9.4675$
- Der Standardfehler von $\beta_1 = 0.4798$
- Das $R^2 = 0.05$, der Standardfehler der Residuen ist SER $= 18.58$.

Damit hat man (fast) alle für den Hypothesentest und die Bestimmung der Konfidenzintervalle notwendigen Zahlen.

9.9.2. Bayesianische Schätzung des linearen Modells

Die Bayesianische Schätzung des linearen Modells liefert ein sehr ähnliches Ergebnis:

```
library(MCMCpack)
summary(MCMCregress(testscr ~ str,data=Caschool))


Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

               Mean      SD Naive SE Time-series SE
(Intercept) 698.963  9.4739 0.094739       0.094739
str          -2.281  0.4809 0.004809       0.004809
sigma2      346.985 24.2723 0.242723       0.242723

2. Quantiles for each variable:
```

```
              2.5%     25%     50%     75%   97.5%
(Intercept) 680.43 692.599 699.020 705.329 717.594
str          -3.22  -2.602  -2.286  -1.957  -1.343
sigma2      302.32 330.080 345.898 362.825 397.550
```

```
                  2.5%        25%        50%        75%      97.5%
(Intercept) 680.425568 692.598501 699.020321 705.329336 717.593538
str          -3.219805  -2.602326  -2.285552  -1.957015  -1.342553
sigma2      302.316401 330.079696 345.897840 362.824999 397.549961
```

Anders als im Ergebnis von `lm` erhalten wir geschätzte Quantile für unsere Parameter, d.h. wir wissen, in welchem Intervall unsere Parameter mit Wahrscheinlichkeit 95% liegen.

```
confint(est)
```

```
                2.5 %     97.5 %
(Intercept) 680.32313 717.542779
str          -3.22298  -1.336637
```

9.9.3. Eigenschaften von OLS

Was wir über OLS wissen:

- OLS ist unverzerrt
- OLS ist konsistent
- wir können Konfidenzintervalle für $\hat{\beta}$ berechnen
- wir können Hypothesen über $\hat{\beta}$ testen

Ein großer Teil ökonometrischer Auswertung wird in Form von OLS präsentiert. Ein Grund ist, dass sehr viele Leute verstehen, wie OLS funktioniert. Wenn wir einen anderen Schätzer verwenden, kann es sein, dass man uns nicht mehr versteht.

- Reicht das als Begründung um OLS zu verwenden?
- Gibt es bessere Schätzer? Schätzer mit kleinerer Varianz?

Um diese Fragen zu beantworten, werden wir weitere Annahmen machen.

9.9.4. Erweiterte OLS Annahmen

1. $E(u_i|X_i = x) = 0$
2. (X_i, Y_i) sind i.i.d.
3. Große Ausreißer in X und Y sind selten (die vierten Momente von X und Y existieren).

4. u ist homoskedastisch, $\text{var}(u_i|X_i = x)$ ist konstant.

5. u ist normalverteilt, $u_i \sim N(0, \sigma^2)$.

Annahmen 4 und 5 sind restriktiver — lassen sich also seltener rechtfertigen.

Gauss Markov
Unter den Annahmen 1-4 hat $\hat{\beta}_1$ die kleinste Varianz unter *allen linearen Schätzern* (unter allen Schätzern, die lineare Funktionen von Y sind).

Effizienz von OLS-II
Unter den Annahmen 1-5 hat $\hat{\beta}_1$ die kleinste Varianz unter *allen konsistenten Schätzern* wenn $n \to \infty$ (egal ob sie linear oder nichtlinear sind)

Zurück zur Motivation: Bietverhalten in Erstpreisauktionen

- Risikoaversion: höhere Gebote für hohe Werte
- Einfache Bietregeln: niedrigere Gebote für niedrige Werte

Kirchkamp, Reiß, *Economic Journal*, 2011.

9.10. Literatur

- Dolić, Statistik mit R, Kapitel 9.1.
- Hartung, Statistik, Kapitel X.1.
- Schira, Statistische Methoden der VWL und BWL-Theorie und Praxis, Kapitel 17.1, 17.2.
- Stock and Watson. Introduction to Econometrics, Brief Edition, Chapter 4 - 5.
- Verzani, Using R for Introductory Statistics, Chapter 10.1 - 10.2.

9.11. Schlüsselbegriffe

- Regression, KQ-Schätzer, OLS-Schätzer
- abhängige / unabhängige Variable
- Fehlerterm, Residuen
- Bestimmtheitsmaß R^2, SER

Anhang 9.A Übungen

Übung 9.3 *Wir betrachten folgende Stichprobe:*

Ausgaben für Werbung	*900*	*1300*	*1200*	*400*	*700*	*800*	*1000*
Absatz des Produktes	*400*	*700*	*550*	*100*	*250*	*300*	*500*

1. *Schätzen Sie ein lineares Modell um den Zusammenhang zwischen den beiden Variablen zu beschreiben.*
2. *Was ist in diesem Modell die abhängige, was die unabhängige Variable?*
3. *Interpretieren Sie die diagnostischen Plots*
4. *Welchen Wert hat das Bestimmtheitsmaß* R^2*?*

Übung 9.4 *Um den Zusammenhang der Nachfrage nach Schlafanzügen und Nachthemden besser zu verstehen, werden folgende Verkaufszahlen erhoben:*

Saison	*Schlafanzüge* (x)	*Nachthemden* (y)
Frühjahr/Sommer 2018	*600*	*1000*
Herbst/Winter 2018/2019	*1000*	*400*
Frühjahr/Sommer 2019	*600*	*600*
Herbst/Winter 2019/2020	*1000*	*200*
Frühjahr/Sommer 2020	*800*	*800*

1. *Ermitteln sie die optimale Gerade zur Darstellung dieses Zusammenhanges mit einer OLS-Schätzung.*
2. *Was verstehen Sie unter dem Begriff »Residuen«?*
3. *Bestimmen Sie für die optimale Gerade das Bestimmtheitsmaß!*
4. *Ist der Koeffizient* β_1 *signifikant von* 0 *verschieden?*
5. *Ist der Koeffizient* β_1 *signifikant von* -1 *verschieden?*
6. *Geben Sie ein 95%-Konfidenzintervall für* β_1 *an.*
7. *Geben Sie ein 95%-credible interval für* β_1 *an.*

Übung 9.5 *Ein Elektrokonzern will einen neuen DVD-Player auf den Markt bringen. Um den optimalen Preis zu ermitteln, wurde der DVD-Player auf 10 Testmärkten zu unterschiedlichen Preisen angeboten. Dabei gab es das folgende Ergebnis:*

Testmarkt	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Preis in Euro	*50*	*55*	*60*	*65*	*70*	*75*	*80*	*85*	*90*	*95*
Absatzmenge	*100*	*95*	*90*	*85*	*80*	*75*	*70*	*65*	*60*	*55*

1. *Welche ist die abhängige Variable, welche die unabhängige?*

2. *Bestimmen Sie nun die Regressionsgerade!*
3. *Testen Sie, ob β_1 signifikant von 0 verschieden ist.*
4. *Testen Sie, ob β_0 signifikant von 150 verschieden ist.*
5. *Bestimmen Sie ein Konfidenzintervall für β_1.*
6. *Bestimmen Sie ein Konfidenzintervall für β_0.*
7. *Bestimmen Sie ein credible interval.*
8. *Wie groß ist R^2?*

Übung 9.6 *Sie erheben Daten zu einer abhängigen Variablen Y und einer unabhängigen Variablen X. Die Daten sind durch folgende Tabelle gegeben:*

X	0	0	2	2
Y	0	4	0	4

Nun schätzen Sie ein lineares Regressionsmodell ohne Konstante:

$$Y_i = \beta_1 X_i + u_i$$

1. *Wie groß ist der OLS Schätzer $\hat{\beta}_1$?*
2. *Nun erweitern Sie ihr Modell um einen konstanten Term:*

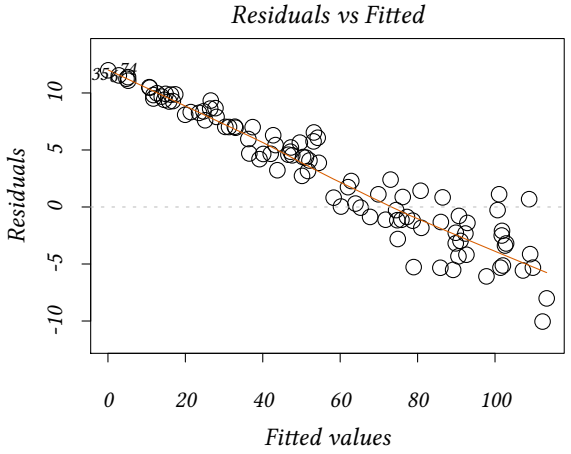
   $$Y_i = \beta_0 + \beta_1 X_i + u_i$$

   Wie groß ist jetzt Ihr OLS Schätzer $\hat{\beta}_1$?
3. *Welches Modell hat ein größeres R^2?*
4. *In welchen Modell ist die Summe der Fehlerquadrate größer?*

Übung 9.7 *Sie schätzen ein lineares Regressionsmodell ohne Konstante*

$$Y_i = \beta_1 X_i + u_i$$

und erhalten den folgenden diagnostischen Plot ihrer Regression.



Welche Schlussfolgerungen ziehen Sie?

- Zwischen unabhängiger und abhängiger Variablen besteht ein fallender Zusammenhang
- Wenn Sie das Modell um eine Konstante erweitern, wird sich der Fit der Regression verbessern
- Die Annahme $E(u_i|X_i = x) = 0$ ist nicht erfüllt
- Die Annahme, $\text{var}(u_i|X_i = x)$ ist konstant, ist nicht erfüllt.
- Die Annahme, große Ausreißer in X und Y seien selten, ist nicht erfüllt.

Anhang 9.B Herleitung des OLS Schätzers

$$\frac{\partial KQ}{\partial b_0} = -2\sum_{i=1}^{n}(Y_i - b_1X_i - b_0)$$

$$\bar{Y} - \hat{\beta}_0 - \hat{\beta}_1\bar{X} = 0$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$$

$$\frac{\partial KQ}{\partial b_1} = -2\sum_{i=1}^{n}(Y_i - b_1X_i - b_0)X_i$$

$$\frac{1}{n}\sum_{i=1}^{n}X_iY_i - \hat{\beta}_0\bar{X} - \hat{\beta}_1\frac{1}{n}\sum_{i=1}^{n}X_i^2 = 0$$

$$\frac{1}{n}\sum_{i=1}^{n}X_iY_i - (\bar{Y} - \hat{\beta}_1\bar{X})\bar{X} - \hat{\beta}_1\frac{1}{n}\sum_{i=1}^{n}X_i^2 = 0$$

$$\frac{1}{n}\sum_{i=1}^{n}X_iY_i - \bar{Y}\bar{X} + \hat{\beta}_1\left(\bar{X}^2 - \frac{1}{n}\sum_{i=1}^{n}X_i^2\right) = 0$$

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}X_iY_i - \bar{X}\bar{Y}}{\frac{1}{n}\sum_{i=1}^{n}X_i^2 - \bar{X}^2}$$

mit

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \bar{X})\cdot(Y_i - \bar{Y}) &= \sum_{i=1}^{n}X_iY_i - \bar{X}\bar{Y}\\ \sum_{i=1}^{n}(X_i - \bar{X})^2 &= \sum_{i=1}^{n}X_i^2 - \bar{X}^2\end{aligned}$$

erhalten wir

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})\cdot(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Anhang 9.C Unverzerrtheit des Schätzers für $\hat{\beta}$

Wir interessieren uns für $\beta_1 - \hat{\beta}_1$. Wir wissen

$$\begin{aligned}Y_i &= \beta_0 + \beta_1X_i + u_i\\ \bar{Y} &= \beta_0 + \beta_1\bar{X} + \bar{u}\\ \text{also } Y_i - \bar{Y} &= \beta_1(X_i - \bar{X}) + (u_i - \bar{u})\end{aligned}$$

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^n (X_i-\bar{X})^2}\\
&= \frac{\sum_{i=1}^n (X_i-\bar{X})\left(\beta_1(X_i-\bar{X})+(u_i-\bar{u})\right)}{\sum_{i=1}^n (X_i-\bar{X})^2}\\
&= \beta_1\frac{\sum_{i=1}^n (X_i-\bar{X})(X_i-\bar{X})}{\sum_{i=1}^n (X_i-\bar{X})^2}\\
&\quad + \frac{\sum_{i=1}^n (X_i-\bar{X})(u_i-\bar{u})}{\sum_{i=1}^n (X_i-\bar{X})^2}\\
\hat{\beta}_1-\beta_1 &= \frac{\sum_{i=1}^n (X_i-\bar{X})(u_i-\bar{u})}{\sum_{i=1}^n (X_i-\bar{X})^2}
\end{aligned}$$

Nun ist

$$\begin{aligned}
\sum_{i=1}^n (X_i-\bar{X})(u_i-\bar{u}) &= \sum_{i=1}^n (X_i-\bar{X})u_i -\\
&\quad \left(\sum_{i=1}^n (X_i-\bar{X})\right)\bar{u}\\
&= \sum_{i=1}^n (X_i-\bar{X})u_i -\\
&\quad \left(\left(\sum_{i=1}^n X_i\right)-n\cdot\bar{X}\right)\bar{u}\\
&= \sum_{i=1}^n (X_i-\bar{X})u_i
\end{aligned}$$

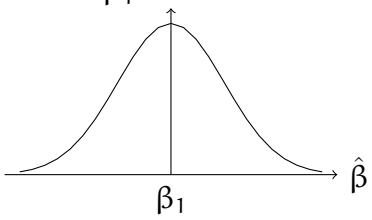
Also

$$\hat{\beta}_1-\beta_1 = \frac{\sum_{i=1}^n (X_i-\bar{X})(u_i-\bar{u})}{\sum_{i=1}^n (X_i-\bar{X})^2} = \frac{\sum_{i=1}^n (X_i-\bar{X})u_i}{\sum_{i=1}^n (X_i-\bar{X})^2}$$

Jetzt können wir $E(\hat{\beta}_1)-\beta_1$ berechnen:

$$\begin{aligned}
E\left(\hat{\beta}_1\right)-\beta_1 &= E\left(\frac{\sum_{i=1}^n (X_i-\bar{X})u_i}{\sum_{i=1}^n (X_i-\bar{X})^2}\right)\\
&= E\left(E\left(\frac{\sum_{i=1}^n (X_i-\bar{X})u_i}{\sum_{i=1}^n (X_i-\bar{X})^2}\right)\middle| X_1,\ldots,X_n\right)\\
&\qquad \text{da nach Annahme 1: } E(u_i|X_i=x)=0\\
E(\hat{\beta}_1)-\beta_1 &= 0
\end{aligned}$$

$\hat{\beta}_1$ ist ein **unverzerrter Schätzer** von β_1



Anhang 9.D Varianz von $\hat{\beta}_1$

Jetzt zur Varianz:

Ein genaueres Studium der obigen Formel für $\bar{\beta}_1$ ergibt

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\begin{aligned}
\text{nenne } (X_i - \bar{X})u_i &= v_i \\
\text{Ausserdem gilt } s_X^2 &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} \\
\hat{\beta}_1 - \beta_1 &= \frac{\sum_{i=1}^{n} v_i}{(n-1)s_X^2} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\frac{n-1}{n}s_X^2}
\end{aligned}$$

für große n gilt $s_x^2 \approx \sigma_x^2$ und $\frac{n-1}{n} \approx 1$, also

$$\begin{aligned}
\hat{\beta}_1 - \beta_1 &\approx \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\sigma_X^2} \\
&\text{nun ist} \\
\operatorname{var}(\hat{\beta}_1) = \operatorname{var}(\hat{\beta}_1 - \beta_1) &\approx \operatorname{var}\left(\frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\sigma_X^2}\right) = \\
&\quad \frac{\operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n} v_i\right)}{(\sigma_X^2)^2} \\
= \frac{\operatorname{var}(v_i)/n}{(\sigma_X^2)^2} &= \frac{1}{n}\frac{\operatorname{var}((X_i - \mu_X)\cdot u_i)}{\sigma_X^4}
\end{aligned}$$

Diese Formel setzt voraus, dass die Varianzen $\operatorname{var}((X_i - \mu_X)\cdot u_i$ und $\operatorname{var}(X_i)$ bekannt sind. Normalerweise ist das nicht der Fall und wir müssen diese Varianzen schätzen.

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n}\frac{\frac{1}{n-2}\sum_{i=1}^{n}\left(\left(X_i - \bar{X}\right)\cdot \hat{u}_i\right)^2}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right)^2}$$

10. Modelle mit mehr als einer unabhängigen Variablen (multiple Regression)

10.1. Motivation

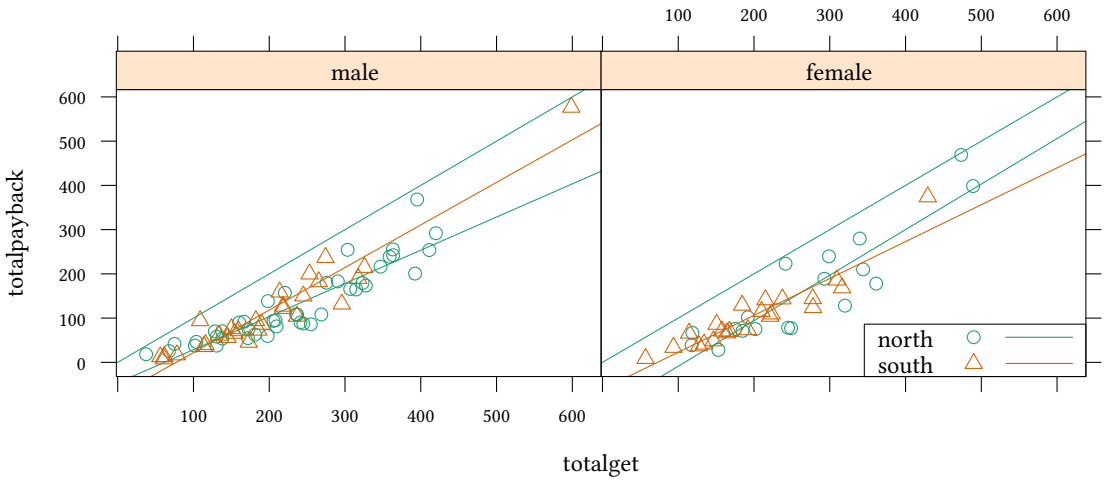
Vertrauensspiele Viele ökonomisch wichtige Transaktionen können nicht oder nicht vollständig mit Verträgen spezifiert werden (z.B. Arbeitsverträge können nicht jeden Handgriff in der Zukunft genau festlegen). Dennoch finden Transaktionen dieser Art statt. Wir modellieren diese Transaktion als »Vertrauensspiel«: Ein Trustor sendet eine Investition an einen Trustee. Der Trustee kann, muss aber nicht, einen Teil zurückschicken.

- Der Trustor ist mit 100 Geldeinheiten ausgestattet und kann einen Teil davon (x) an den Trustee senden.
- Der gesendete Betrag wird beim Trustee mit 3 multipliziert, der Trustee erhält also $3 \cdot x$.
- Nun kann der Trustee einen beliebigen Anteil von $3 \cdot x$ zurücksenden.

→ Frage: Was beeinflusst den Spielerfolg: Nationalität, Alter, Geschlecht, Geschwister...?

Die effiziente Lösung wäre: Der Trustor sendet den vollen Betrag von 100 an den Trustee. Der Trustee sendet einen Betrag zwischen 100 und 300 zurück. Beide erhalten in der Summe 300. Wenn aber der Trustor dem Trustee nicht »vertraut«, wird nichts gesendet und beide erhalten in der Summe nur 100.

Bornhorst, Ichino, Kirchkamp, Schlag, Winter (2010), “Similarities and Differences when Building Trust: the Role of Cultures”, Experimental Economics, Vol. 13/3, pp. 260-283.



In der obigen Grafik sehen wir an der vertikalen Achse den Rückzahlungsbetrag. Der hängt einerseits von dem Betrag ab, den der Trustee bekommen hat und der auf der horizontalen

Achse abgetragen ist (weil in diesem Experiment ein Trustee auch von mehreren Trustoren ausgewählt werden kann, kann ein Trustee auch mehr als 300 erhalten). Andererseits hängt der Rückzahlungsbetrag vielleicht auch vom Geschlecht des Trustees ab. Und schließlich hängt er vielleicht vom kulturellen Hintergrund der Trustees ab, also ob diese aus Nord- oder Südeuropa kommen.

Wie man mehrere erklärende Variablen berücksichtigt, betrachten wir in diesem Kapitel.

10.2. Erweiterung des Beispiels aus Kapitel 9

Wir betrachten wieder den Datensatz `Caschool`. Die Variablen, die wir bislang betrachtet haben, sind:

- `testscr` test score
- `str` student / teacher ratio

```
data(Caschool,package="Ecdat")
attach(Caschool)
```

Bislang hatten wir nur *eine* unabhängige Variable:

$$\texttt{testsrc} = \beta_0 + \beta_1 \texttt{str} + u$$

Vielleicht spielen weitere Faktoren (außer `str`) eine Rolle?
Wie kann man mehrere Faktoren gleichzeitig berücksichtigen?

- Idee: Halte einen Faktor »konstant« indem nur eine kleine Gruppe betrachtet wird (z.B. alle Schüler mit einem sehr ähnlichen `elpct` (english learner percentage))

Um eine geschickte Aufteilung in Gruppen zu bestimmen, betrachten wir zunächst einige Statistiken des english learner percentage `elpct`:

```
summary(elpct)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.941   8.778  15.768  22.970  85.540
```

Hier entscheiden wir uns für eine Aufteilung in drei Gruppen: kleiner als 9, zwischen 9 und 23, und größer als 23.

Nun schätzen wir einen Regression für einzelne Klassen von `elpct`:

Die Option `subset` im Kommando `lm` beschränkt die Schätzung auf einen Teil des Datensatzes.

```
est1 <- lm(testscr ~ str ,subset=(elpct<=9))
est2 <- lm(testscr ~ str ,subset=(elpct>9 & elpct<=23))
est3 <- lm(testscr ~ str ,subset=(elpct>23))
```

```
library(texreg)
```

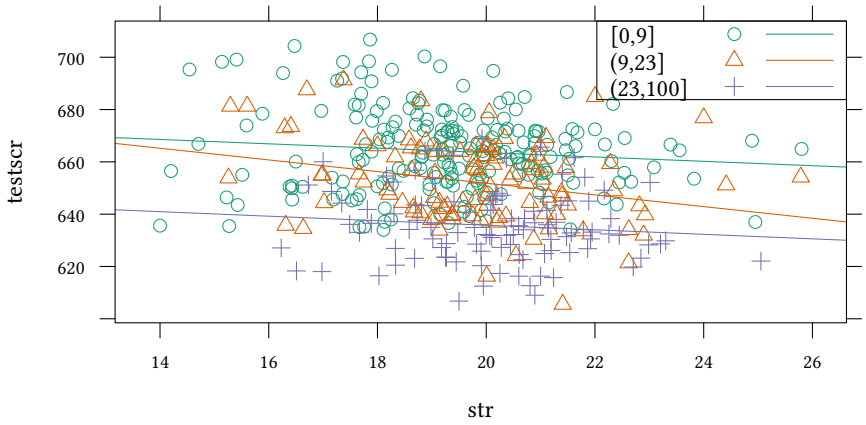
```
texreg(list(`elpct<9`=est1,`9..23`=est2,`>=23`=est3))
```

	elpct ≤ 9	$9 <$ elpct ≤ 23	elpct > 23
(Intercept)	680.25***	696.44***	653.07***
	(10.59)	(15.75)	(16.21)
str	−0.84	−2.23**	−0.87
	(0.55)	(0.79)	(0.80)
R^2	0.01	0.07	0.01
Adj. R^2	0.01	0.06	0.00
Num. obs.	214	101	105

***p < 0.001; **p < 0.01; *p < 0.05

(Zahlen in Klammern unter den geschätzten Koeffizienten sind die geschätzten Standardabweichungen.)

Die folgende Grafik zeigt die drei geschätzten Regressiongeraden für jeweils verschiedene Werte von `elpct`



Abhängig von `elpct` sind die geschätzten Zusammenhänge sehr unterschiedlich

- erweitere das Regressionsmodell

$$\texttt{testsrc} = \beta_1 \texttt{str} + \beta_2 \texttt{elpct} + \beta_0 + u$$

allgemein:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

für jede Beobachtung:

$$\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} + u_1 \\
y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{2k} + u_2 \\
y_3 &= \beta_0 + \beta_1 x_{31} + \beta_2 x_{32} + \cdots + \beta_k x_{3k} + u_3 \\
&\vdots \\
y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} + u_n
\end{aligned}$$

In Matrixschreibweise:

$$\mathbf{y} = X\boldsymbol{\beta} + \mathbf{u}$$

`xtable` aus der Bibliothek `xtable` stellt die Ergebnisse der Schätzung etwas hübscher dar.

	Estimate	Std. Error	t value	Pr(>|t|)	2.5 %	97.5 %
(Intercept)	686.03	7.41	92.57	0.0000	671.46	700.60
str	-1.10	0.38	-2.90	0.0040	-1.85	-0.35
elpct	-0.65	0.04	-16.52	0.0000	-0.73	-0.57

10.3. Bayesianische Schätzung des linearen Modells

Natürlich können wir auch hier mit `MCMCregress` eine Bayesianische Schätzung berechnen:

```
library(MCMCpack)
summary(MCMCregress(testscr ~ str + elpct,data=Caschool))
```

```
Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

                Mean       SD  Naive SE Time-series SE
(Intercept) 686.0129  7.38431 0.0738431      0.0738104
str          -1.1009  0.37906 0.0037906      0.0037927
elpct        -0.6495  0.03969 0.0003969      0.0003969
sigma2      210.0738 14.69920 0.1469920      0.1469920

2. Quantiles for each variable:

               2.5%      25%      50%      75%    97.5%
(Intercept) 671.690 680.9725 685.9850 691.0057 700.5888
str          -1.838  -1.3544  -1.0990  -0.8412  -0.3722
elpct        -0.729  -0.6764  -0.6492  -0.6226  -0.5719
sigma2      183.111 199.8335 209.3915 219.4100 241.1087
```

Quantile der Bayesianischen Schätzung:

	2.5%	25%	50%	75%	97.5%
(Intercept)	671.69	680.97	685.99	691.01	700.59
str	-1.84	-1.35	-1.10	-0.84	-0.37
elpct	-0.73	-0.68	-0.65	-0.62	-0.57
sigma2	183.11	199.83	209.39	219.41	241.11

Konfidenzintervall der OLS Regression:

	2.5 %	97.5 %
(Intercept)	671.46	700.60
str	-1.85	-0.35
elpct	-0.73	-0.57

Wie schon in Abschnitt 9.9.2 ist auch hier das Ergebnis der Bayesianischen Schätzung des linearen Modells sehr ähnlich dem Ergebnis von `lm`. Anders als `lm` erhalten wir aber wieder Quantile für die Verteilung der Parameter, d.h. wir wissen, in welchem Intervall die Parameter mit welcher Wahrscheinlichkeit liegen.

10.4. Annahmen für das multiple Regressionsmodell

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

1. $E(u_i|X_i = x) = 0$
2. (X_i, Y_i) sind i.i.d.
3. Große Ausreißer in X und Y sind selten (die vierten Momente von X und Y existieren).
4. X hat Rang gleich der Anzahl der Spalten (keine Multikollinearität), man kann keine Variable als lineare Funktion der anderen Variablen berechnen.
5. $\text{var}(u|X = \mathbf{x})$ ist konstant, u ist homoskedastisch
6. u ist normalverteilt $u \sim N(0, \sigma^2)$

Annahme 4 ist neu.
Annahmen 5 und 6 sind restriktiver — lassen sich also seltener rechtfertigen.

10.5. Die Verteilung der OLS Schätzer in der multiplen Regression

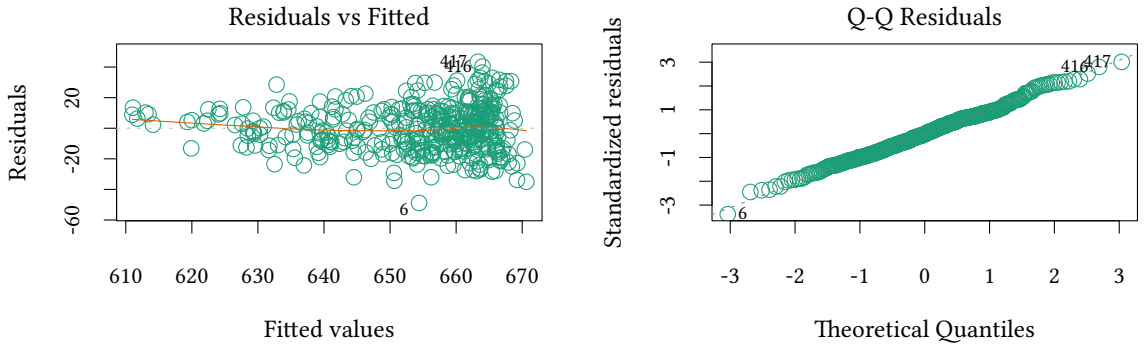
↑ Modell mit einem Regressor: die OLS Schätzer für $\hat{\beta}_0$ und $\hat{\beta}_1$ sind unverzerrte und konsistente Schätzer. Für große Stichproben sind $\hat{\beta}_0$ und $\hat{\beta}_1$ normalverteilt.

- multiple Regression: unter obigen Annahmen 1–4 ist der OLS Schätzer $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y}$ *unverzerrt und konsistent*. Für große Stichproben ist $\hat{\boldsymbol{\beta}}$ multivariat normalverteilt.
- Gauss Markov: Unter den Annahmen 1–5 hat $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y}$ die *kleinste Varianz* unter *allen linearen Schätzern* (unter allen Schätzern, die lineare Funktionen von Y sind).

- Effizienz von OLS: Unter den Annahmen 1–6 hat $\hat{\beta} = (X'X)^{-1}X'y$ die *kleinste Varianz* unter *allen konsistenten Schätzern* wenn $n \to \infty$ (egal ob sie linear oder nichtlinear sind)

Wie kann man überprüfen, ob die Annahmen des Regressionsmodells erfüllt sind? Wenn wir einen konkreten Verdacht haben, dann gibt es Tests. Einen ersten Eindruck erhalten wir bereits mit diagnostischen Plots:

```
est<-lm(testscr ~ str + elpct)
plot(est)
```



Der Plot links stellt die Residuen an der vertikalen Achse und die gefitteten Werte an der horizontalen Achse dar. Die Darstellung erscheint zunächst ungewohnt. Warum nimmt man nicht einfach X an der horizontalen Achse? Der Grund ist, dass X mehrdimensional ist, hier im Beispiel enthält X sowohl `str` als auch `elpct`. Man muss sich also entscheiden. Es wäre durchaus möglich, `str` an der horizontalen Achse darzustellen, aber genauso gut könnte man `elpct` nehmen. Als Kompromiss nehmen wir $\hat{y} = X\hat{\beta}$, also gerade die Linearkombination von X, die wir geschätzt haben.

Der Plot rechts ist ein Q-Q Plot der Residuen gegen die Quantile der Normalverteilung. Wenn die Residuen normalverteilt sind, liegen die Punkte fast auf einer Geraden.

10.6. Die Verteilung von $\hat{\beta}$

10.6.1. Varianz von $\hat{\beta}$

Im Fall der einfachen Regression (zur Erinnerung):

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n}\frac{\frac{1}{n-2}\sum_{i=1}^n \hat{u}_i^2}{\frac{1}{n}\sum_{i=1}^n (X_i-\bar{X})^2} \quad \text{Homoskedastizität}$$

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n}\frac{\frac{1}{n-2}\sum_{i=1}^n \left(X_i-\bar{X}\right)\cdot\hat{u}_i)^2}{\left(\frac{1}{n}\sum_{i=1}^n \left(X_i-\bar{X}\right)^2\right)^2}$$

↑ Heteroskedastizität (immer richtig)

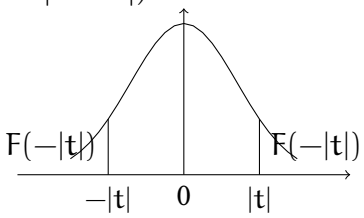
Im Fall der multiplen Regression (in Matrixschreibweise):

$$\begin{aligned}\Sigma_{\hat\beta\hat\beta} &= \hat\sigma_u^2(X'X)^{-1} \quad \text{(bei Homoskedastizität)}\\ \Sigma_{\hat\beta\hat\beta} &= (X'X)^{-1}X'\mathbf{I}u^2X(X'X)^{-1}\\ &\qquad\qquad \uparrow \text{Heteroskedastizität (immer richtig)}\end{aligned}$$

Aus historischen Gründen nehmen Statistikprogramme normalerweise immer Homoskedastizität an. Wir werden im Rahmen dieser Vorlesung *nicht* darauf eingehen, wie man R dazu überredet, mit Heteroskedastizität-robusten Varianz-Kovarianz Matrizen zu rechnen. Falls es Sie interessiert: `hccm` aus der Bibliothek `car` berechnet Heteroskedastizität-robusten Varianz-Kovarianz Matrizen. Die Kommandos zum Testen (wie `linearHypothesis`) die wir unten kennenlernen, kennen die Option `vcov=hccm` um mit robusten Standardfehlern zu testen.

Hypothesentests Um die Hypothese $H_0 : \beta_j = \beta_{j,0}$ versus $H_1 : \beta_j \neq \beta_{j,0}$ zu testen:

- bestimme die t Statistik:
$$t^{\text{Stichp.}} = \frac{\hat\beta_j - \beta_{j,0}}{\hat\sigma_{\hat\beta_j}}$$
- Der p-Wert ist $p = \Pr\left(|t| > \left|t^{\text{Stichp.}}\right|\right) = 2 \cdot F_N(-|t^{\text{Stichp.}}|)$



Wir speichern das Ergebnis unserer Schätzung in der Variablen `est`

```
est <- lm(testscr ~ str + elpct)
```

Mehr Details dazu in Appendix 10.D. Hier ist das Ergebnis, das R uns liefert:

```
xtable(est)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	686.0322	7.4113	92.57	0.0000
str	-1.1013	0.3803	-2.90	0.0040
elpct	-0.6498	0.0393	-16.52	0.0000

Da R annimmt, die Residuen seien normalverteilt, und da die Stichprobe nur endlich groß ist, rechnet R die p-Werte mit der t-Verteilung und nicht mit der Normalverteilung aus. Der Unterschied ist nicht groß.

Konfidenzintervalle Mit der obigen geschätzten Standardabweichung können wir auch Konfidenzintervalle berechnen:

Hier nochmal die bekannte Formel für Konfidenzintervalle für Mittelwerte:

$$\left[\bar{x} + \sigma_{\bar{x}} \cdot Q_N\left(\frac{\alpha}{2}\right), \bar{x} - \sigma_{\bar{x}} \cdot Q_N\left(\frac{\alpha}{2}\right)\right]$$

Bei Regressionskoeffizienten geht es genauso:

$$\left[\hat{\beta} + \hat{\sigma}_{\beta} \cdot Q_N\left(\frac{\alpha}{2}\right), \hat{\beta} - \hat{\sigma}_{\beta} \cdot Q_N\left(\frac{\alpha}{2}\right)\right]$$

bzw., falls wir die t-Verteilung mit $n - k - 1$ Freiheitsgraden verwenden (weil wir (wie R) annehmen, dass die Residuen normalverteilt sind und die Stichprobe nur endlich groß ist):

$$\left[\hat{\beta} + \hat{\sigma}_{\beta} \cdot Q^t_{n-k-1}\left(\frac{\alpha}{2}\right), \hat{\beta} - \hat{\sigma}_{\beta} \cdot Q^t_{n-k-1}\left(\frac{\alpha}{2}\right)\right]$$

Auch für Konfidenzintervalls arbeitet R mit der t-Verteilung, das heißt, wir nehmen an, dass die Residuen normalverteilt sind.

```
confint(est)
```

```
                 2.5 %      97.5 %
(Intercept) 671.4640604 700.6004370
str          -1.8487972  -0.3537945
elpct        -0.7271112  -0.5724423
```

Beachte: Das 95%-Konfidenzintervall ist dem 95%-credible interval aus Abschnitt 10.3 sehr ähnlich.

	2.5%	25%	50%	75%	97.5%
(Intercept)	671.69	680.97	685.99	691.01	700.59
str	-1.84	-1.35	-1.10	-0.84	-0.37
elpct	-0.73	-0.68	-0.65	-0.62	-0.57
sigma2	183.11	199.83	209.39	219.41	241.11

Konfidenzintervall der OLS Regression:

	2.5 %	97.5 %
(Intercept)	671.46	700.60
str	-1.85	-0.35
elpct	-0.73	-0.57

Oft ist nicht nur ein Konfidenzintervall für $\hat{\beta}$ interessant, sondern für ein Vielfaches von $\hat{\beta}$. Was passiert etwa, wenn die Klassengröße z.B. um 2 verkleinert wird?

```
-2*confint(est)

                    2.5 %        97.5 %
(Intercept) -1342.928121 -1401.2008740
str             3.697594     0.7075891
elpct           1.454222     1.1448847
```

Das Konfidenzintervall unserer Prognose für »Was passiert, wenn die Klassengröße um 2 Schüler pro Klasse verkleiner wird« würde also von einer erwarteten Verbesserung des Testscores von 0.71 bis 3.7 reichen.

Diese Aussage ist allerdings problematisch, wie wir im nächsten Abschnitt merken werden.

10.7. Omitted variable bias

Was kann passieren, wenn wir eine Variable in unserem Modell vergessen?

Betrachten wir nochmals unsere einfache Schätzgleichung:

$$\texttt{testsrc} = \beta_1 \texttt{str} + \beta_0 + u$$

Was könnte sonst noch einen Einfluss auf `testscr` haben?

	korr. mit Regressor `str`	beinflusst abh. Var. `testscr`
percent of English learners	x	x
time of day of the test		x
parking lot space per pupil	x	

$$\begin{aligned} \text{wahrer Zusammenhang:} \quad Y &= \beta_0 + \beta_1 X_1 + \underbrace{\beta_2 X_2 + u}_{\tilde{u}} \\ \text{geschätzter Zusammenhang:} \quad Y &= \beta_0 + \beta_1 X_1 + \tilde{u} \end{aligned}$$

Wenn wir eine Variable (z.B. X_2) *nicht* in unsere Schätzgleichung aufnehmen, diese Variable aber

- mit dem Regressor (andere X) korreliert ist
- und die abhängige Variable (y) beeinflusst

dann

→ ist die Annahme $E(u_i|X_i) = 0$ nicht mehr erfüllt

→ ist unsere Schätzung für $\boldsymbol{\beta}$ verzerrt (omitted variable bias)

In Anhang 10.C zeigen wir

$$E(\mathbf{b}_1) = \boldsymbol{\beta}_1 + (X_1'X_1)^{-1}X_1'X_2\boldsymbol{\beta}_2$$

10.7.1. Beispiele für Omitted-Variable-Bias:

- Klassische Musik → Intelligenz von Kindern (Rauscher, Shaw, Ky; Nature; 1993)

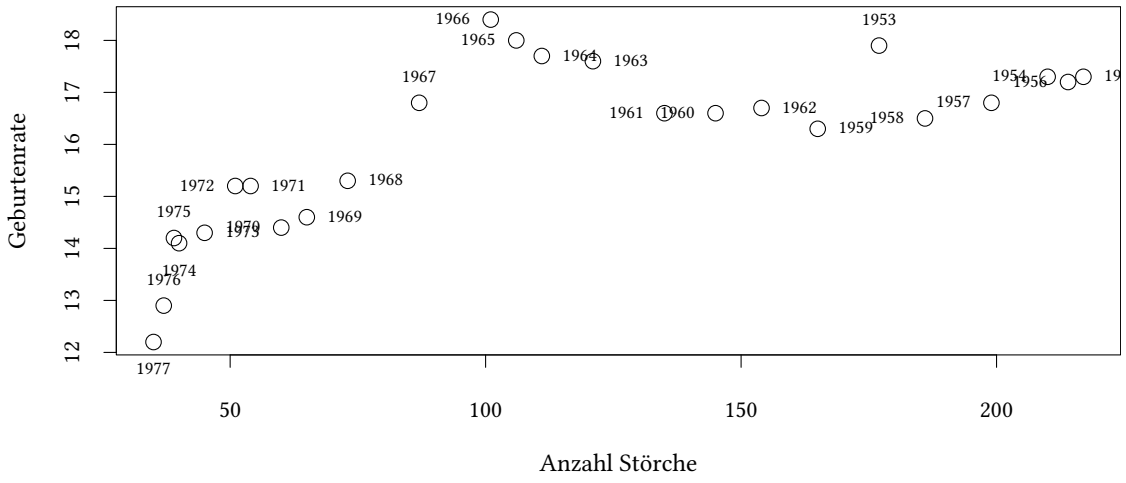
  (fehlende Variable: Einkommen)

- French paradox: Rotwein, Leberpastete → weniger Erkrankungen der Herzkranzgefäße (Samuel Black, 1819)

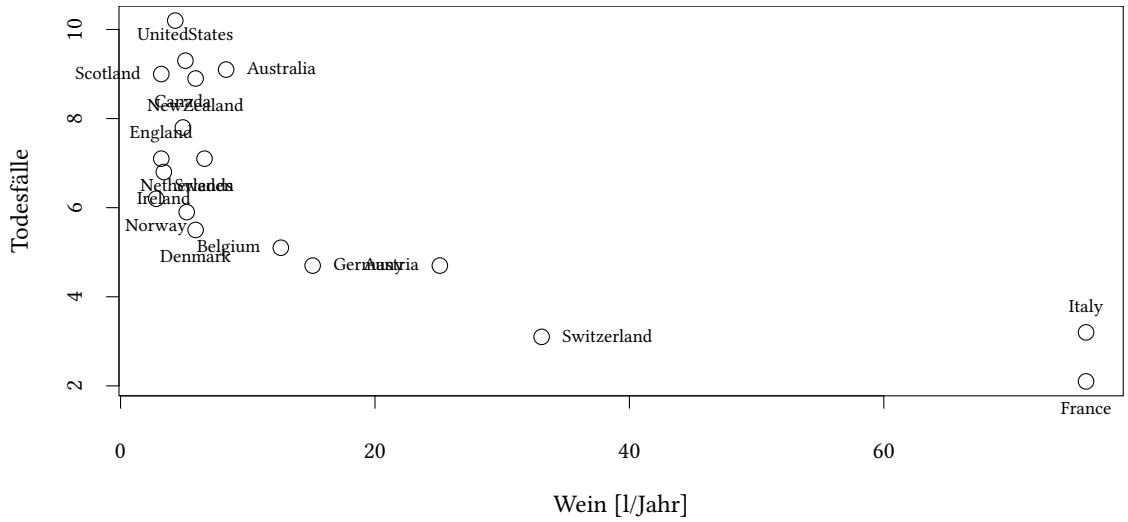
  (fehlende Variable: Fisch und Zucker in der Ernährung,…)

- Störche in Niedersachsen → Geburtenrate

  (fehlende Variable: Industrialisierung)



Gabriel, K. R. and Odoroff, C. L. (1990) Biplots in biomedical research. *Statistics in Medicine* 9(5): pp. 469-485.



Todesfälle infolge koronarer Herzkrankheit (pro 1000 Männer, 55 – 64 Jahre). Daten aus: St. Leger A.S., Cochrane, A.L. and Moore, F. (1979). Factors Associated with Cardiac Mortality in Developed Countries with Particular Reference to the Consumption of Wine, *Lancet*: 1017–1020.

10.7.2. Erweiterung der Schätzgleichung um Ausgaben pro Schüler

Bislang gab es in unserer Regression nur zwei erklärende Variablen: `str` und `elpct`:

Vergleiche:

$$\begin{aligned} \texttt{testscr} &= \beta_0 + \beta_1 \texttt{str} + \beta_2 \texttt{elpct} \\ \texttt{testscr} &= \beta_0 + \beta_1 \texttt{str} + \beta_2 \texttt{elpct} + \beta_3 \texttt{expnstu} \end{aligned}$$

```
est <- lm(testscr ~ str + elpct)
```

```
cbind(xtable(est),xtable(confint(est)))
```

	Estimate	Std. Error	t value	Pr(>|t|)	2.5 %	97.5 %
(Intercept)	686.03	7.41	92.57	0.0000	671.46	700.60
str	-1.10	0.38	-2.90	0.0040	-1.85	-0.35
elpct	-0.65	0.04	-16.52	0.0000	-0.73	-0.57

Nun nehmen wir eine dritte hinzu: `expnstu`

```
est2 <- lm(testscr ~ str + elpct + expnstu)
```

	Estimate	Std. Error	t value	Pr(>|t|)	2.5 %	97.5 %
(Intercept)	649.5779	15.2057	42.72	0.0000	619.6883	679.4676
str	-0.2864	0.4805	-0.60	0.5515	-1.2310	0.6582
elpct	-0.6560	0.0391	-16.78	0.0000	-0.7329	-0.5792
expnstu	0.0039	0.0014	2.74	0.0064	0.0011	0.0066

Wir sehen, der geschätzte Einfluss der Klassengröße `str` ändert sich. Oben war er noch -1.1, jetzt ist er -0.286. Die `Modellspezifikation` hat einen deutlichen Einfluss auf die geschätzten Koeffizienten.

Nicht nur die Koeffizienten ändern sich, auch t-Statistik und p-Werte. Der geschätzte Koeffizient für `str` ist in dieser Spezifikation nicht mehr signifikant von Null verschieden. Auch das Konfidenzintervall unserer Prognose für »Was passiert, wenn die Klassengröße um 2 Schüler pro Klasse verkleiner wird« verändert sich:

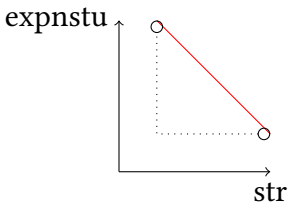
Anstatt, wie oben, eine erwarteten Verbesserung des Testscores von 0.71 bis 3.7 zu erwarten, rechnen wir nun mit eine erwarteten Verbesserung des Testscores von -2.5 bis 1.3.

Was ist passiert?

- In der Schätzgleichung mit `str` und `elpct` haben wir vernachlässigt, dass kleinere Klassen auch mit reicheren Gemeinden, und damit auch höheren Ausgaben pro Schüler einhergehen.

- In Abschnitt 10.6.1 haben wir die Frage »Was passiert…« mit $\rightarrow$ 0.71 bis 3.7 beantwortet.

  Das war eine Antwort auf die Frage: »Was passiert, wenn die Klassen um zwei Schüler kleiner werden *und die Ausgaben pro Schüler auch so steigen*, wie es in entsprechend kleineren Klassen der Fall ist.«



- Wenn man *nur* die Klassengröße ändert, die Ausgaben pro Schüler `expnstu` aber konstant hält, dann bekommt man den Effekt den wir in der erweiterten Schätzgleichung bestimmt haben:

  Eine Änderung des Testscores zwischen -2.5 und 1.3.

Wir erinnern uns:

- Bei einem unterspezifiziertes Modell (ein Regressor β_2, hier `expnstu`, fehlt):

  $\hat{\beta}$ ist nur unverzerrt wenn $\beta_2 = 0$ oder $X_1' X_2 = 0$.

- In unserem Beispiel ist $\beta_2 \neq 0$ und $X_1' X_2 \neq 0$.

- `str` und `expnstu` sind korreliert.

$\rightarrow$ wir erhalten eine verzerrte Schätzung für β_{str} wenn `expnstu` in der Schätzgleichung fehlt.

10.8. Multikollinearität

Oben, in den Abschnitten 10.7 und 10.7.2 haben wir gesehen, dass wir alle wichtigen Variablen in die Regression aufnehmen sollten. Kann man auch zu viele Variablen in die Gleichung aufnehmen?

Beispiel

$$\texttt{testscr} = \beta_1 \texttt{str} + \beta_2 \texttt{elpct} + \beta_0$$

```
lm(testscr ~ str + elpct)
```

	Estimate	Std. Error	t value	Pr(>|t|)	2.5 %	97.5 %
(Intercept)	686.03	7.41	92.57	0.0000	671.46	700.60
str	-1.10	0.38	-2.90	0.0040	-1.85	-0.35
elpct	-0.65	0.04	-16.52	0.0000	-0.73	-0.57

Jetzt erweitern wir das Modell um eine Variable: Anteil der »English learners« `FracEL=elpct/100`:

$$\texttt{testscr} = \beta_1 \texttt{str} + \beta_2 \texttt{elpct} + \beta_3 \texttt{FracEL} + \beta_0$$

```
Call:
lm(formula = testscr ~ str + elpct + FracEL)

Residuals:
    Min      1Q  Median      3Q     Max
-48.845 -10.240  -0.308   9.815  43.461

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 686.03225    7.41131  92.566  < 2e-16 ***
str          -1.10130    0.38028  -2.896  0.00398 **
elpct        -0.64978    0.03934 -16.516  < 2e-16 ***
FracEL             NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.46 on 417 degrees of freedom
Multiple R-squared:  0.4264, Adjusted R-squared:  0.4237
F-statistic:   155 on 2 and 417 DF,  p-value: < 2.2e-16
```

Wir erhalten gar keine Schätzung für den Koeffizienten von `FracEL`. Was ist das Problem? Die neue Variable `FracEL` ist ein Vielfaches von `elpct`, sie ist *kollinear*. Allgemein sprechen wir von *Multikollinearität*.

Im Beispiel sehen wir: R erkennt selbst, dass Multikolliniearität vorliegt, und vereinfacht das Modell entsprechend. In der Zeile `FracEL` erscheinen als Schätzergebnisse nur `NA`.

Das gelingt aber nicht immer.

Wir perturbieren diese Variable nun etwas. Das kann auch unabsichtlich (z.B. durch Rundungsfehler) passieren. Die Variablen sind dann nicht mehr (perfekt) Multikollinear. Wir bekommen deshalb Ergebnisse für alle Koeffizienten. Für jede (kleine) Perturbation ändert sich das Ergebnis erheblich. Die Standardfehler werden sehr groß. Jetzt berechnen wir `FracEL` etwas ungenauer — wir fügen einen kleinen Zufallsterm hinzu:

```
set.seed(123)
FracEL<-elpct/100+rnorm(4)*.001
xtable(lm(testscr ~ str + elpct + FracEL))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	685.9887	7.4172	92.49	0.0000
str	-1.1040	0.3806	-2.90	0.0039
elpct	-6.0027	8.7281	-0.69	0.4920
FracEL	535.1954	872.6427	0.61	0.5400

Große Koeffizienten für `FracEL` werden ausgeglichen durch kleine von `elpct`. Was ist passiert? Nehmen wir an, dies sei der wahre Zusammenhang:

Wir haben $\texttt{FracEL} = \texttt{elpct}/100$

$$\begin{aligned}
\texttt{testscr} &= 686.0322 - 1.1013\texttt{str} - 0.6498\texttt{elpct}\\
\texttt{testscr} &= 686.0322 - 1.1013 \cdot \texttt{str} + (\underline{a} - 0.6498) \cdot \texttt{elpct} - \underline{100a \cdot \texttt{elpct}/100}\\
\texttt{testscr} &= 686.0322 - 1.1013 \cdot \texttt{str} + (a - 0.6498) \cdot \texttt{elpct} - 100a \cdot \texttt{FracEL}
\end{aligned}$$

Koeffizienten β können nicht mehr identifiziert werden.

Technisch: Die Varianz von β wird sehr groß:

$$\Sigma_{\hat\beta\hat\beta} = \hat\sigma_u^2 (X'X)^{-1}$$

Bei perfekter Multikollinearität kann $(X'X)^{-1}$ nicht mehr berechnet werden.

Multikollinearität: Identifikation der verantwortlichen Koeffizienten Wir haben oben gesehen, dass bei Multikollinearität die geschätzte Varianz der geschätzten Koeffzienten β groß wird. Ein Maß dafür ist der *Variance-Inflation-Factor* VIF

Wir schätzen für jede der *unabhängigen Variablen* X_i die folgende Gleichung:

$$X_i = \beta_0 + \sum_{j \neq i} \beta_j X_j + u$$

und bestimmen jeweils den Korrelationskoeffizienten R_i^2. Dann gilt

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

Ein VIF nahe bei 1 ist unproblematisch, ein VIF größer als 10 könnte auf ein Problem hinweisen. Die Funktion `vif` finden wir im Paket `car`.

```
library(car)
vif(lm(testscr ~ str + elpct + FracEL))

         str        elpct       FracEL
    1.036634 50936.875445 50937.840610

vif(lm(testscr ~ str + elpct + expnstu))

     str    elpct  expnstu
1.680787 1.040031 1.629915
```

In unserem Beispiel sehen wir, dass `str` unproblematisch ist, jedoch `elpct` und `FracEL` zu Problemen führen.

10.9. Spezifikationsfehler: Zusammenfassung

- unterspezifiziertes Modell, ein Regressor β_2 fehlt (omitted variable bias):
  - $\hat{\beta}$ ist nur unverzerrt wenn $\beta_2 = 0$ oder $X_1'X_2 = 0$.

    $\rightarrow$ ökonomische Intuition, welche anderen wichtigen Faktoren könnte es geben?

- überspezifiziertes Modell, Regressoren sind kollinear:
  - $\hat{\beta}$ kann nicht geschätzt werden ($X'X$ ist nicht invertiertbar)

    $\rightarrow$ Statistikprogramm entfernt die kollinearen Regressoren selbst.

- überspezifiziertes Modell, Regressoren sind fast kollinear (overfitting):
  - $\hat{\beta}$ kann nur ungenau geschätzt werden

    $\rightarrow$ Identifikation der problematischen Regressoren mit VIF, ökonomische Intuition zur Auswahl der Regressoren.

Ist das ein Widerspruch?

- geringe Korrelation von X_1 und X_2: um omitted variable bias zu vermeiden, sollten beide Variablen aufgenommen werden.

- hohe Korrelation von X_1 und X_2: das Problem des omitted variable bias wird stärker, aber auch das Problem der Multikollinearität:
  - Vielleicht messen X_1 und X_2 eigentlich das gleiche. Dann besser nur eine Variable.
  - Wenn X_1 und X_2 unterschiedliche Dinge messen, und man beide Effekte schätzen will, dann wird die Schätzung halt sehr ungenau.

10.10. Bayesianische Schätzung und Multikollinearität

Das Bayesianische Verfahren kann uns bei Multikollinearität auch nicht weiterhelfen:

```
library(MCMCpack)
FracEL<-elpct/100+rnorm(4)*.001
summary(MCMCregress(testscr ~ str + elpct + FracEL))


Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

                Mean       SD Naive SE Time-series SE
(Intercept)  686.119   7.4279 0.074279       0.074279
str           -1.098   0.3815 0.003815       0.003815
elpct          4.997   6.6393 0.066393       0.066393
FracEL      -564.685 664.1391 6.641391       6.641391
sigma2       210.330  14.6531 0.146531       0.151248

2. Quantiles for each variable:

                 2.5%        25%      50%       75%    97.5%
(Intercept)   671.881   681.0989  686.161  691.0551 701.1505
str            -1.864    -1.3496   -1.096   -0.8388  -0.3727
elpct          -8.088     0.4904    5.066    9.4720  17.8118
FracEL      -1845.470 -1011.7771 -571.650 -113.3976 744.5614
sigma2        183.523   200.0768  209.677  219.8204 240.6606
```

Zurück zur Motivation

- Unterschiede im Spielerfolg lassen sich vor allem durch regionale Zugehörigkeit erklären. Spieler aus dem Norden Europas bauen zu Beginn des Spiels Vertrauen auf, und können das im Laufe des Spiels ausnutzen.

  Andere Faktoren, wie Geschlecht, Alter, Geschwister, etc., tragen nur wenig bei.

Bornhorst, Ichino, Kirchkamp, Schlag, Winter (2010), “Similarities and Differences when Building Trust: the Role of Cultures”, Experimental Economics, Vol. 13/3, pp. 260-283.

10.11. Literatur

- Dolić, Statistik mit R, Kapitel 9.2.
- Verzani, Using R for Introductory Statistics, Chapter 10.3.
- Stock and Watson. Introduction to Econometrics, Brief Edition, Chapter 6.

10.12. Schlüsselbegriffe

- multiple Regression
- Spezifikationsfehler
- Omitted variable bias
- Overfitting, Multikollinearität
- Lineare Hypothesen die mehrere Koeffizienten betreffen

Anhang 10.A Beispiele für die Vorlesung

Für Ihre Schätzung verwenden Sie die Regressionsgleichung

$$Y = \beta_0 + \beta_1 X + u\,.$$

Welche Aussage trifft zu:

1. Wenn β_1 signifikant von 0 verschieden ist, dann beeinflusst die Variable X die Variable Y kausal.
2. Wenn Sie feststellen, dass β_1 signifikant von 0 verschieden ist, dann gibt es einen linearen Zusammenhang zwischen X und Y.
3. Wenn 0 im Konfidenzintervall für β_1 liegt, dann ist β_1 nicht signifikant von 0 verschieden
4. Wenn 0 nicht im Konfidenzintervall für β_1 liegt, dann ist β_1 signifikant von 0 verschieden
5. Die Schätzung für β_1 wird um so genauer, je stärker X und u korreliert sind.

Betrachten Sie das folgende Ergebnis einer Regression. Gehen Sie von einem Signifikanzniveau von 5% aus:

```
Call: lm(formula = y ~ x1 + x2)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.967      8.329   0.836   0.4305
x1             8.222      2.697   3.049   0.0186 *
x2            -3.182      1.453  -2.189   0.0648 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 11.09 on 7 degrees of freedom
Multiple R-squared: 0.5962, Adjusted R-squared: 0.4808
F-statistic: 5.168 on 2 and 7 DF,  p-value: 0.04184
```

1. x1 hat einen signifikanten Einfluss auf y.
2. x2 hat einen signifikanten Einfluss auf y.
3. Die Hypothese, der marginale Effekt von x1 auf y sei −5, wird abgelehnt.
4. Die Hypothese, der marginale Effekt von x2 auf y sei −5, wird abgelehnt.
5. Die Hypothese, der Zusammenhang von x1 auf y sei nicht linear, kann abgelehnt werden.

Nun erweitern Sie die obige Regression um die Variable x3. Sie erhalten folgendes Ergebnis:

```
Call: lm(formula = y ~ x1 + x2 + x3)
Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.967      8.329   0.836   0.4305
x1             8.222      2.697   3.049   0.0186 *
x2            -3.182      1.453  -2.189   0.0648 .
x3                NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 11.09 on 7 degrees of freedom
Multiple R-squared: 0.5962, Adjusted R-squared: 0.4808
F-statistic: 5.168 on 2 and 7 DF,  p-value: 0.04184
```

1. Der Einfluss von x3 auf y ist nicht signifikant.
2. Der marginale Effekt von x3 ist Null.
3. x3 ist eine lineare Funktion von x1 und x2.
4. x3 ist eine nicht-lineare Funktion von x1 und x2.
5. Das Modell erklärt 59.62% der Varianz von y.

Anhang 10.B Übungen

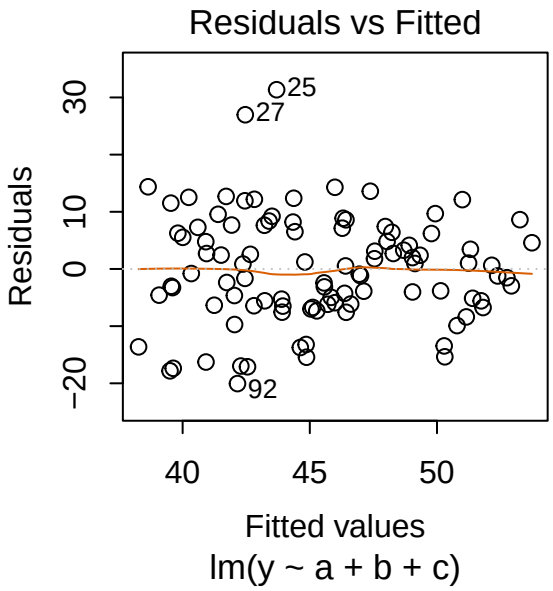
Übung 10.2 *Sie betreiben eine Baumschule und versuchen, das Wachstum Ihrer Bäume mit unterschiedlichen Düngemitteln* A, B, *und* C *zu steigern. Sie düngen 100 Bäume mit unterschiedlichen Kombinationen der drei Düngemittel und messen die Größe nach einem Jahr. In Ihrem Datensatz ist* y *die Größe des Baums, und* a, b *und* c *sind die Mengen der Düngemittel. Sie berechnen in R eine lineare Regression der abhängigen Variablen* y *auf die unabhängigen* a, b, *und* c. *Den Output sehen Sie links, einen diagnostischen Plot der Residuen über die "fitted values" rechts:*

```
Call:
lm(formula = y ~ a + b + c)

Coefficients:
(Intercept)            a            b            c
     52.815        1.513       -1.062      -14.899
```

```
summary(est)
```

```
Call:
lm(formula = y ~ a + b + c)

Residuals:
     Min       1Q   Median       3Q      Max
-20.0487  -6.1479  -0.1334   6.6716  31.3493

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)   52.815      3.033  17.412   < 2e-16 ***
a              1.513      3.360   0.450     0.653
b             -1.062      3.372  -0.315     0.754
c            -14.899      3.633  -4.101 0.0000862 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.571 on 96 degrees of freedom
Multiple R-squared:  0.1533, Adjusted R-squared:  0.1269
F-statistic: 5.796 on 3 and 96 DF,  p-value: 0.001101
```

```
plot(est,which=1)
```



1. Betrachten Sie den diagnostischen Plot. Hilft Ihnen diese Plot zu entscheiden, ob Annahmen des klassichen Regressionsmodells erfüllt oder verletzt sind? Falls ja, welche Annahmen sind erfüllt bzw. verletzt bzw. können anhand der Graphik nicht beurteilt werden?

2. Gehen Sie nun davon aus, dass alle Annahmen des Regressionsmodells erfüllt sind. Um welchen Betrag steigt die Größe Ihrer Bäume etwa, wenn Sie den Anteil des Düngemittels C um eine Einheit erhöhen?

3. Ein befreundeter Gärtner behauptet, die von Ihnen eingesetzten Düngemittel hätten überhaupt keine Wirkung. Falls Sie eine Wirkung beobachtet hätten, sei das purer Zufall. Können Sie dieser Ansicht etwas entgegenhalten? Wenn ja, für welche Düngemittel? Begründen Sie Ihre Antwort.

4. Durch ein Versehen wurden 100 neugepflanzte Bäume mit jeweils 10 Einheiten des Düngemittels B behandelt. Erwarten Sie, dass diese Bäume größer oder kleiner als unbehandelte Bäume werden? Wenn ja, um wieviel werden sie im Mittel größer oder kleiner?

Übung 10.3 *Sie wollen herausfinden, ob Geschlecht oder Staatsangehörigkeit einen Effekt auf die Ergebnisse von Kindern in einem Test von ökonomischen Verständnis haben.*

In Ihrem Regressionsmodell wird das Ergebnis des Ökonomietests S_i *eines Schülers* i *ausgedrückt durch das Testergebnis in einem allgemeinen standardisierten Intelligenztest* I_i *(mit Mittelwert 100 und Standardabweichung 15), durch den Mittelwert der Ausbildungsjahre der Eltern* A_i*, sowie durch die Dummyvariablen* G_i *für das Geschlecht (1 wenn weiblich, 0 sonst) und* D_i *für die Staatsangehörigkeit (1 wenn deutsch, 0 sonst).*

Leider ist der Zugang zum Datensatz nicht mehr möglich. Sie sehen nur die Ergebnisse der durchgeführten Regression (Standardfehler in Klammern):

$$\underset{}{\hat{S}_i} = \underset{}{5.7} - \underset{(0.63)}{0.63} \cdot X_{1i} - \underset{(0.88)}{0.22} \cdot X_{2i} + \underset{(0.08)}{0.16} \cdot X_{3i} + \underset{(0.1)}{1.2} \cdot X_{4i}$$

Die Regression beruht auf 62 Beobachtungen, $R^2 = 0,54$.

1. Versuchen Sie zu bestimmen, welches Schätzergebnis zu welcher Variablen gehört. Begründen Sie Ihre Antwort genau.

2. Bestimmen Sie ausgehend von Ihrer Antwort zu Aufgabe 1 passende Hypothesen für die Variablen. Testen Sie diese Hypothesen zum 5%-Signifikanzniveau.

   a) Welche Teststatistiken berechnen Sie?

   b) Welche Verteilungsfunktion müssen Sie verwenden?

   Falls unterschiedliche Annahmen zu unterschiedlichen Verteilungsfunktionen führen, erklären Sie, welche Annahmen welche Verteilungsfunktion rechtfertigen.

3. Was ist Ihrer Meinung nach der Effekt von Geschlecht und Staatsangehörigkeit auf das Testergebnis S_i in diesem Sample?

Übung 10.5 *Ein Marktforschungsunternehmen untersucht den Zusammenhang zwischen der Nachfrage nach Schokolade, dem verfügbaren Einkommen und der Nachfrage nach Gummitieren. Dabei soll folgender Zusammenhang gelten:*

$$\begin{aligned}\textit{Nachfrage nach Schokolade} =& \beta_1 \cdot \textit{verfügbares Einkommen}\\ &+\beta_2 \cdot \textit{Nachfrage nach Gummitieren}\\ &+\beta_0 + u\end{aligned}$$

Sie analysieren die Daten der Nachfrage nach Schokolade (in Tsd.), das durchschnittliche verfügbare Einkommen (in Euro) und der Nachfrage nach Gummitieren (in Tsd.) der letzten 10 Jahre und erhalten den folgenden Output:

```
Schokolade <- c(1000,2370,1580,960,5670,2090,4650,3020,3650,1960)
Einkommen <- c(1710,1701,1806,1613,2097,1793,1930,1857,1820,1698)
Gummitiere <- c(970,4870,3170,980,3560,2080,5600,4360,1530,910)
```

```
summary(lm(Schokolade~Einkommen+Gummitiere))


Call:
lm(formula = Schokolade ~ Einkommen + Gummitiere)

Residuals:
    Min      1Q  Median      3Q     Max
-1189.6  -416.7   183.9   425.3   933.5

Coefficients:
              Estimate  Std. Error t value Pr(>|t|)
(Intercept) -14663.2577   3477.3113  -4.217  0.00395 **
Einkommen        9.4543      2.0460   4.621  0.00242 **
Gummitiere       0.1130      0.1614   0.700  0.50634
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 720.2 on 7 degrees of freedom
Multiple R-squared:  0.8336, Adjusted R-squared:  0.7861
F-statistic: 17.54 on 2 and 7 DF,  p-value: 0.001878
```

$$\begin{aligned}\text{Nachfrage nach Schokolade} =& \beta_1 \cdot \text{verfügbares Einkommen}\\ &+\beta_2 \cdot \text{Nachfrage nach Gummitieren}\\ &+\beta_0 + u\end{aligned}$$

1. Wie groß sind die Werte von β_0, β_1 und β_2?
2. Wie groß ist das Bestimmtheitsmaß?

3. Hat die Nachfrage nach Gummitieren einen signifikanten Einfluss auf die Nachfrage nach Schokolade?

4. Hat das Einkommen einen signifikanten Einfluss auf die Nachfrage nach Schokolade?

5. Bestimmen Sie ein 95%-Konfidenzintervall für den Einfluss des Einkommens.

6. Prüfen Sie die Hypothese: Die Nachfrage nach Gummitieren und die nach Schokolade verändern sich 1:1, wenn die Nachfrage nach Gummitieren um eine Einheit steigt, dann steigt auch die Nachfrage nach Schokolade um eine Einheit.

Übung 10.6 *Sie schätzen ein lineares Regressionsmodell um den Einfluss von drei Variablen, `x1`, `x2`, und `x3` auf `y` zu ermitteln. Sie erhalten folgendes Ergebnis:*

```
lm(formula = y ~ x1 + x2 + x3)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.2        1.0     -0.2      0.842
x1            0.5        1.0      0.5      0.618
x2            1.5        1.0      1.5      0.137
x3           -1.5        1.5     -1.0      0.320
```

1. *Ihre Nullhypothese ist, dass bei einer Steigerung von `x2` um 10 Einheiten die Variable `y` um 20 Einheiten wächst. Wie groß ist bei einem zweiseitigen Test der absolute Betrag Ihrer Teststatistik, wenn Ihr Datensatz 100 Beobachtungen enthält?*

2. *Welchen* p*-Wert erhalten Sie für den Test der obigen Hypothese?*

Übung 10.7

- *Verwenden Sie den Datensatz `BudgetFood` aus der Bibliothek `Ecdat` um den Anteil der Ausgaben für Lebensmittel `wfood` an den Haushaltsausgaben als Funktion der Gesamtausgaben `totexp`, dem Alter `age` der befragten Person und der Größe des Haushalts `size` zu erklären.*

- *Nun ziehen Sie jeweils eine Stichprobe der Größe 10. Betrachten Sie die gemeinsame Verteilung der Koeffizienten von `totexp` und `age`.*

```
(trueTheta <- with(BudgetFood,coef(lm(wfood ~ totexp + age  + size))))
N <- dim(BudgetFood)[1]
estimates <- t(replicate(200,with(sample(BudgetFood[sample(1:N,10),]),
                                  coef(lm(wfood ~ totexp + age  + size)))))
with(as.data.frame(estimates),plot(totexp ~ age))
abline(h=trueTheta["totexp"],v=trueTheta["age"],col="red")
abline(h=0,v=0,lty="dotted")
```

Anhang 10.C Omitted variable bias

Das wahre Modell sei

$$y = X_1\beta_1 + X_2\beta_2 + u$$

was passiert, wenn wir in der Spezifikation des Modells X_2 vergessen?

$$\hat{\beta} = \underbrace{(X'X)^{-1}X'}_{X^+} y$$

$$\begin{aligned} b_1 &= (X_1'X_1)^{-1}X_1'y \\ &= (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + u) \\ &= (X_1'X_1)^{-1}X_1'X_1\beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'u \\ E(b_1) &= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 \end{aligned}$$

dieser Ausdruck ist $= \beta_1$ nur wenn

- $\beta_2 = 0$
- oder $X_1'X_2 = 0$, d.h. X_1 und X_2 sind orthogonal

Anhang 10.D Details zu Hypothesentest und Konfidenzinterval

Hypothesentests Oben haben wir R verwendet, um einfache Hypothesentests durchzuführen. Hier die Details:

Um die Hypothese $H_0 : \beta_j = \beta_{j,0}$ versus $H_1 : \beta_j \neq \beta_{j,0}$ zu testen:

- bestimme die t Statistik:

$$t^{\text{Stichp.}} = \frac{\hat{\beta}_j - \beta_{j,0}}{\hat{\sigma}_{\hat{\beta}_j}}$$

- Der p-Wert ist $p = \Pr\left(|t| > \left|t^{\text{Stichp.}}\right|\right) = 2 \cdot F_N(-|t^{\text{Stichp.}}|)$

Wir speichern das Ergebnis unserer Schätzung in der Variablen `est`

```
est <- lm(testscr ~ str + elpct)
```

Jetzt rechnen wir die p-Werte selbst aus (H_0 ist $\beta = 0$)

`diag(X)` beschreibt die Diagonalmatrix von `X` falls `X` eine quadratische Matrix ist. Für einen Vektor `x` spannt `diag(x)` die Diagonalmatrix auf. `coef` extrahiert die geschätzten Koeffizienten aus einem Modell.

Homoskedastische Varianz-Kovarianzmatrix von $\hat{\beta}$

$$\Sigma_{\hat{\beta}\hat{\beta}} = \hat{\sigma}_u^2 (X'X)^{-1}$$

```
vcov(est)
```

```
             (Intercept)         str        elpct
(Intercept) 54.92755274 -2.79596671  0.030730824
str         -2.79596671  0.14461160 -0.002807340
elpct        0.03073082 -0.00280734  0.001547836
```

Nur die Varianzen:

```
diag(vcov(est))
```

```
 (Intercept)          str         elpct
54.927552738  0.144611598  0.001547836
```

Aus der Varianz-Kovarianz-Matrix $\Sigma_{\hat\beta\hat\beta}$ extrahieren wir die Standardabweichung von $\boldsymbol{\beta}$:

```
(stddev <- sqrt(diag(vcov(est))))
```

```
(Intercept)         str       elpct
 7.41131248  0.38027832  0.03934255
```

$$\text{zur Erinnerung:}\quad t=\frac{\hat\beta_j-\beta_{j,0}}{\hat\sigma_{\hat\beta_j}}\quad\text{falls}\quad\beta_{j,0}=0\quad\text{dann}\quad t=\frac{\hat\beta_j}{\hat\sigma_{\hat\beta_j}}$$

```
(t <- coef(est) /  stddev)
```

```
(Intercept)         str       elpct
  92.565554   -2.896026  -16.515879
```

Jetzt können wir unsere p-Werte ausrechnen:

$$\text{zur Erinnerung:}\quad p=2\cdot F_N(-|t|)$$

```
round(2*pnorm(- abs(t)),5)
```

```
(Intercept)         str       elpct
    0.00000     0.00378     0.00000
```

Und hier ist das Ergebnis, das R uns liefert:

```
xtable(est)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	686.0322	7.4113	92.57	0.0000
str	-1.1013	0.3803	-2.90	0.0040
elpct	-0.6498	0.0393	-16.52	0.0000

Da R annimmt, die Residuen seien normalverteilt, und da die Stichprobe nur endlich groß ist, rechnet R die p-Werte mit der t-Verteilung und nicht mit der Normalverteilung aus. Der Unterschied ist nicht groß.

Konfidenzintervalle Mit der obigen geschätzten Standardabweichung können wir auch Konfidenzintervalle berechnen:

Hier nochmal die bekannte Formel für Konfidenzintervalle für Mittelwerte:

$$\left[\bar{x} + \sigma_{\bar{x}} \cdot Q_N\left(\frac{\alpha}{2}\right), \bar{x} - \sigma_{\bar{x}} \cdot Q_N\left(\frac{\alpha}{2}\right)\right]$$

Bei Regressionskoeffizienten geht es genauso:

$$\left[\hat{\beta} + \hat{\sigma}_{\beta} \cdot Q_N\left(\frac{\alpha}{2}\right), \hat{\beta} - \hat{\sigma}_{\beta} \cdot Q_N\left(\frac{\alpha}{2}\right)\right]$$

bzw., falls wir die t-Verteilung mit $n - k - 1$ Freiheitsgraden verwenden (weil wir (wie R) annehmen, dass die Residuen normalverteilt sind und die Stichprobe nur endlich groß ist):

$$\left[\hat{\beta} + \hat{\sigma}_{\beta} \cdot Q^t_{n-k-1}\left(\frac{\alpha}{2}\right), \hat{\beta} - \hat{\sigma}_{\beta} \cdot Q^t_{n-k-1}\left(\frac{\alpha}{2}\right)\right]$$

Konfidenzintervall für $\hat{\beta}$

```
coef(est) - qnorm(.975) * stddev

(Intercept)         str       elpct
671.5063431  -1.8466277  -0.7268868

coef(est) + qnorm(.975) * stddev

(Intercept)         str       elpct
700.5581542  -0.3559641  -0.5726668
```

Abhängigkeit der geschätzten Standardabweichung von der Modellspezifikation Nicht nur β, auch die geschätzte Standardabweichung σ_{β} hängen von der Modellspezifikation ab.

Vergleiche die Standardabweichung des Koeffizienten von `str` in den verschiedenen Schätzgleichungen:

```
sqrt(diag(vcov(lm(testscr ~ str ))))["str"]

     str
0.4798256

sqrt(diag(vcov(lm(testscr ~ str + elpct))))["str"]

     str
0.3802783

sqrt(diag(vcov(lm(testscr ~ str + elpct + expnstu))))["str"]

     str
0.4805232
```

Anhang 10.E Restriktionen mit mehreren Koeffizienten

Wenn wir eine Regressionsgleichung mit mehreren Koeffizienten schätzen, dann wollen wir oft auch Hypothesen testen, die mehrere Koeffizienten betreffen. Der Datensatz `RetSchool` erlaubt uns, einen Zusammenhang zwischen dem Lohn und anderen erklärenden Variablen, z.B. dem Ausbildungsniveau der Eltern herzustellen.

```
wage76 ~ grade76 + age76  + black + momed + daded
```

Hypothese: $\beta_{\texttt{momed}} = \beta_{\texttt{daded}}$ (Ausbildung beider Elternteile hat den gleichen Einfluss auf den Lohn)

Lösungsweg 1 (Gleichung umformen, t-test): Allgemein:

$$\begin{aligned} y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u \\ &= \beta_0 + (\beta_1 - \underbrace{\beta_2}) X_1 + \beta_2 (X_2 + \underbrace{X_1}) + u \end{aligned}$$

In der neuen Gleichung ist die Differenz $(\beta_1 - \beta_2)$, die uns interessiert, ein Koeffizient, den wir schätzen können.

Probieren wir es mit unserem Datensatz aus. X_1 entspricht `momed`, X_2 entspricht `daded`. Unsere neue Variable $X_2 + X_1$ ist also `momed+daded`.

```
data(RetSchool, package = "Ecdat")
attach(RetSchool)
momdaded <- momed+daded
```

Zunächst die ursprüngliche Regression:

```
xtable(lm(wage76 ~ grade76 + age76  + black + momed + daded ))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0026	0.0769	0.03	0.9734
grade76	0.0395	0.0030	13.13	0.0000
age76	0.0392	0.0023	17.41	0.0000
black	-0.2183	0.0177	-12.31	0.0000
momed	0.0072	0.0029	2.47	0.0136
daded	0.0005	0.0026	0.18	0.8603

Jetzt die modifizierte Regression:

```
xtable(lm(wage76 ~ grade76 + age76  + black + momed + momdaded ))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0026	0.0769	0.03	0.9734
grade76	0.0395	0.0030	13.13	0.0000
age76	0.0392	0.0023	17.41	0.0000
black	-0.2183	0.0177	-12.31	0.0000
momed	0.0068	0.0047	1.44	0.1510
momdaded	0.0005	0.0026	0.18	0.8603

Die geschätzte Differenz zwischen dem Effekt von momed und daded in der ursprünglichen Gleichung ist in dieser Spezifikation der Koeffizient von momed, also etwa 0.00678 mit einem p-Wert 0.151. Die Differenz ist also nicht signifikant von Null verschieden.

Moderne Statistikprogramme ersparen uns diese Umformulierung. In R schätzen wir einfach wie gewohnt das unveränderte Modell…

```
est <- lm(wage76 ~ grade76 + age76  + black + momed + daded)
```

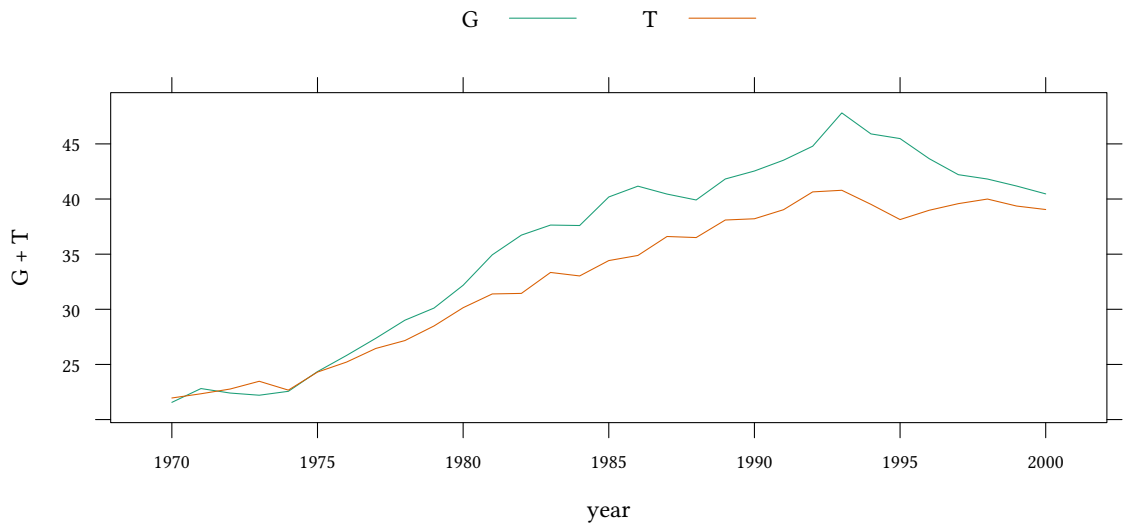
…und wenden dann die Funktion `linearHypothesis` an:

```
linearHypothesis(est,"momed=daded")

Linear hypothesis test

Hypothesis:
momed - daded = 0

Model 1: restricted model
Model 2: wage76 ~ grade76 + age76 + black + momed + daded

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   3054 463.32
2   3053 463.01  1   0.31293 2.0634  0.151
```

Anhang 10.E Modellspezifikation

10.E.1 Motivation

Erwartungsbildung Für wirtschaftsliche Entscheidungen sind Erwartungen über die Zukunft sehr wichtig. In der Fiskalpolitik kann man z.B. annehmen, dass Konsumenten »rationale Erwartungen« bilden, also die Entscheidungen des Staates wenigstens im Mittel richtig vorhersehen.



- Versuchspersonen bilden Erwartungen über zukünftige Steuern (T) und Staatsausgaben (G).

→ Welches Modell beschreibt die Erwartungsbildung unserer Versuchspersonen gut?

Michele Bernasconi, Oliver Kirchkamp, Paolo Paruolo (2009), "Do fiscal variables affect fiscal expectations? Experiments with real world and lab data", Journal of Economic Behavior and Organisation, Vol. 70(1-2), pp. 253-265.

Modellspezifikation In Kapitel 10.2 haben wir gesehen, dass ein Modell weder zu klein, noch zu groß sein sollte:

- Modell zu klein (zu wenige erklärende Variablen): Omitted variable bias
- Modell zu groß (zu viele erklärende Variablen): Overfitting, Multicollinearity

Was also ist die richtige Größe für ein Modell? Welche Variablen sollten wir in die Schätzgleichung aufnehmen, welche nicht?

Leider sind wir nur selten in der Lage, das »richtige« Modell bereits zu kennen. Es gibt aber einige Verfahren, die uns bei der Suche nach einem »vernünftigen« Modell unterstützen.

- `testscr ~ distcod + county + district + grspan + enrltot + teachers + calwpct + mealpct + computer + compstu + expnstu + str + avginc + elpct + readscr + mathscr`

  Dieses Modell ist vermutlich zu groß. Der »Fit« des Modells wird zwar sehr gut sein, aber welcher dieser Koeffizienten ist nun wirklich entscheidend? Viele Koeffizienten messen sehr ähnliche Dinge (z.B. `compstu` und `expnstu` — wir können annehmen, dass in Schulen mit vielen Computern pro Schüler auch die Gesamtausgaben pro Schüler groß sind).

- `testscr ~ str`

- Omitted variable bias Dieses Modell ist zu klein. Wir lassen β_2 weg, obwohl es eine Rolle spielen könnte. Deshalb ist $\hat{\beta}_1$ möglicherweise verzerrt.

$$\mathsf{E}(\mathbf{b}_1) = \boldsymbol{\beta}_1 + (X_1'X_1)^{-1}X_1'X_2\boldsymbol{\beta}_2$$

- Overfitting (Multikollinearität) Unser Modell enthält zu viele Variablen die jeweils sehr ähnliche Dinge messen. Deshalb ist die Schätzung ungenau, die Varianz unserer Schätzer groß.

$$\Sigma_{\hat{\beta}\hat{\beta}} = (X'X)^{-1}X'\mathbf{I}\mathbf{u}^2X(X'X)^{-1}$$

Vorgehen bei der Suche nach dem richtigen Modell

→ beginne mit eine »base specification«

→ darauf aufbauend werden »alternative specifications« entwickelt

wenn sich in einer »alternative specification« Koeffizienten ändern, kann das ein Hinweis auf omitted variable bias sein

- Skaliere die Koeffizienten so, dass die Schätzergebnisse leicht zu lesen und zu interpretieren sind.

10.E.2 Skalierung von Variablen und Koeffizienten

- Der Anteil der »english learners« `elpct` kann entweder als Prozentzahl (zwischen 0 und 100%) oder als relativer Anteil (zwischen 0 und 1) angegeben werden.
- Die Ausgaben pro Schüler `expnstu` können entweder in Dollar, oder in Tausend Dollar angegeben werden.

Welche Darstellung ist besser verständlich?

```
data(Caschool,package="Ecdat")
attach(Caschool)
lm(testscr ~ str + elpct + expnstu)


Call:
lm(formula = testscr ~ str + elpct + expnstu)

Coefficients:
(Intercept)          str        elpct      expnstu
 649.577947    -0.286399    -0.656023     0.003868

elratio <- elpct/100
lm(testscr ~ str + elratio + expnstu)


Call:
lm(formula = testscr ~ str + elratio + expnstu)

Coefficients:
(Intercept)          str      elratio      expnstu
 649.577947    -0.286399   -65.602266     0.003868
```

```
lm(testscr ~ str + elpct + expnstu)
```

```
expnstuTSD <- expnstu/1000
lm(testscr ~ str + elpct + expnstuTSD)
```

```
Call:
lm(formula = testscr ~ str + elpct + expnstuTSD)

Coefficients:
(Intercept)          str        elpct   expnstuTSD
   649.5779      -0.2864      -0.6560       3.8679
```

In 2014	Thüringen	8300€/Schülerin und Schüler
	NRW	5900€ /Schülerin und Schüler

Was bringt eine weitere Variable

- messe R^2
- messe Beitrag zum R^2
- betrachte p-Wert der t-Statistik
- betrachte p-Wert der Varianzanalyse

10.E.3 Messe R^2

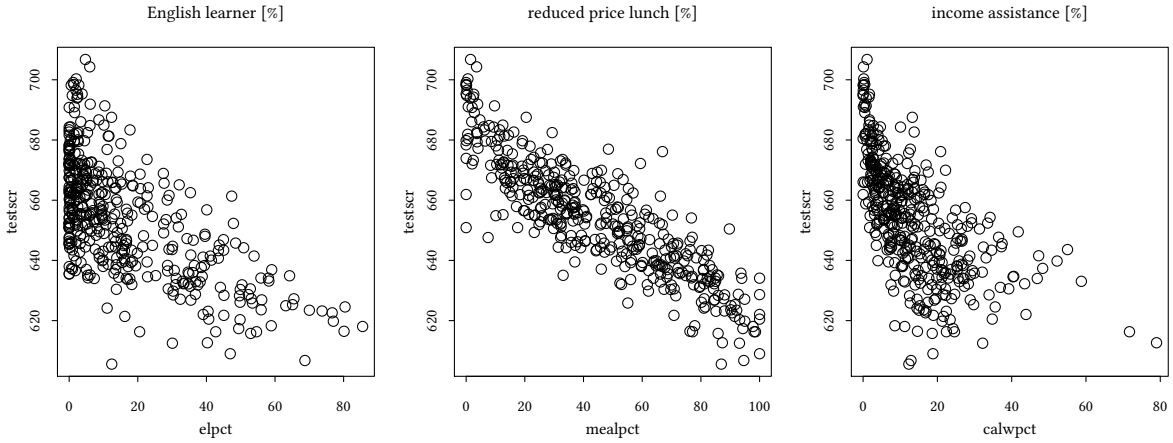
Ist ein Modell mit einem großen R^2 ein »gutes« Modell?

$$R^2 = 1 - \frac{SSR}{TSS} = \frac{\text{erklärte Varianz}}{\text{gesamte Varianz}}$$

- R^2 misst nur den »fit« der Regression
- R^2 misst keine Kausalität (z.B. Parkplatz $\longrightarrow$ testscr)
- R^2 misst nicht die Abwesenheit von Omitted variable bias
- R^2 misst nicht die Korrektheit der Spezifikation

```
par(mfrow=c(1,3),mar=c(4.5,4.5,4,0),mex=.65)
plot(testscr ~ elpct,main="English learner [\\%]")
plot(testscr ~ mealpct,main="reduced price lunch [\\%]")
plot(testscr ~ calwpct,main="income assistance [\\%]")
```

Interpretation von Korrelationen



Können wir diese Korrelation z.B. so interpretieren, dass mehr Geld für Schulspeisung oder mehr Sozialhilfe zu schlechteren Leistungen führt? — Vermutlich nicht!

Der Output der Regression wird uns nicht sagen, ob die »erklärende« Variable wirklich kausal verantwortlich ist, oder ob sie nur mit einer latenten Variablen, die wir nicht beobachten können (hier im Beispiel etwa Einkommen der Eltern), korreliert ist.

Obwohl wir also bei der Interpretation von Korrelationen Vorsicht walten lassen müssen, ist es interessant, sich den Beitrag anzusehen, den eine erklärende Variable zum R^2 leistet. Das machen wir im nächsten Abschnitt.

10.E.4 Messe Beitrag zum R^2

Vergleichen wir zunächst eine Reihe von Schätzungen und ihr jeweiliges R^2:

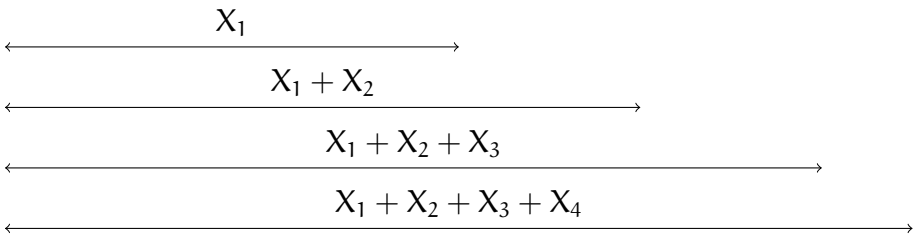
```
summary(lm(testscr ~ str ))$r.squared

[1] 0.0512401

summary(lm(testscr ~ str + elpct ))$r.squared

[1] 0.4264314

summary(lm(testscr ~ str + elpct + mealpct))$r.squared

[1] 0.7745159

summary(lm(testscr ~ str + elpct + mealpct + calwpct ))$r.squared

[1] 0.7748501

summary(lm(testscr ~ str + elpct + mealpct + calwpct + enrltot))$r.squared

[1] 0.775369
```

Während der Beitrag von `mealpct` zum R^2 klein ist, wenn man `mealpct` erst zum Schluss zu einem Modell hinzufügt, ist der Beitrag größer, wenn man mit `mealpct` beginnt:

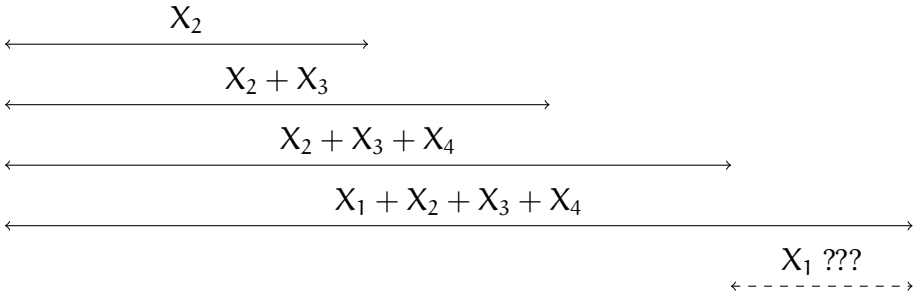
```
summary(lm(testscr ~ mealpct))$r.squared

[1] 0.7547648

summary(lm(testscr ~ mealpct + calwpct ))$r.squared

[1] 0.7552973

summary(lm(testscr ~ mealpct + calwpct + enrltot ))$r.squared

[1] 0.7570513

summary(lm(testscr ~ mealpct + calwpct + enrltot + str ))$r.squared

[1] 0.7670684

summary(lm(testscr ~ mealpct + calwpct + enrltot + str + elpct ))$r.squared

[1] 0.775369
```

Die Länge der Pfeile im folgenden Bild gibt jeweils an, wie groß das R^2 eines Modells ist:

Betrachte eine Folge von Modellen mit unabhängigen Variablen X_1, X_2, X_3, X_4. Wir beginnen mit einem kleinen Modell (eine Variable) und fügen Schritt für Schritt weitere Variablen hinzu. Im ersten Bild beginnen wir mit X_1:



Hier ist eine andere Folge von Modellen. X_1 wird erst am Schluss hinzugefügt:



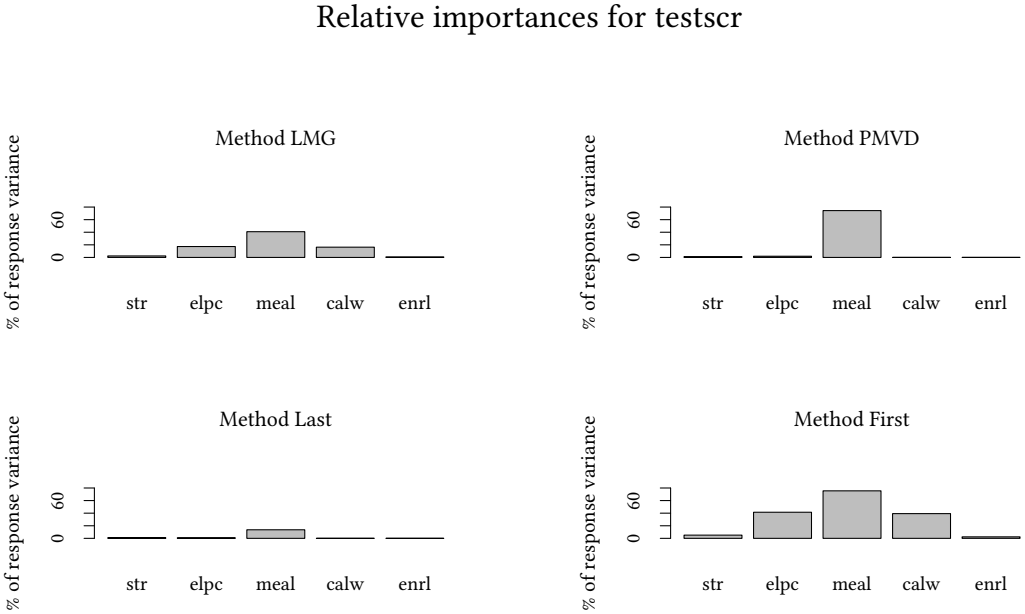
Problem: Der Beitrag einer Variablen zum R^2 hängt davon ab, an welcher Stelle die Variable hinzugefügt wird.

- `"first"`: Beitrag zum R^2, wenn die Variable als einzige (und erste) im Modell ist.

- `"last"`: Beitrag zum R^2, wenn die Variable als letzte hinzugefügt wird.
- `"lmg"`: Mittelwert aus allen denkbaren Reihenfolgen (Lindemann, Merenda, Gold, 1980).
- `"pmvd"` ("proportional marginal variance decomposition", Feldman, 2005). Gewichtung der Reihenfolgen, die unter anderem dafür sorgt, dass (auch bei Korrelation von Faktoren) der Beitrag eines Faktors an der erklären Varianz im Anteil dieses Faktors monoton steigt.

Die Berechnung des Beitrages zum R^2 übernimmt das Kommando `calc.relimp` aus der Bibliothek `relaimpo`.

```
library(relaimpo)
est <- lm(testscr ~ str + elpct + mealpct + calwpct + enrltot)
plot(calc.relimp(est,type=c("first","last","lmg","pmvd")))
```



R^2= 77.54 %, metrics are not normalized.

Die verschiedenen Grafiken, `first`, `last`, `lmg`, und `pmvd`, stellen die Ergebnisse der verschiedenen Methoden dar. Wir haben oben gesehen, dass die Ergebnisse sehr unterschiedlich sein können. `pmvd` stellt einen Kompromiss dar.

Im Diagramm für `pmvd` sehen wir, dass die beiden Variablen `calwpct` und `enrltot` beide weniger als 1% zum R^2 beitragen. Wir verlieren also wenig, wenn wir diese Variablen weglassen.

Falls bei Ihnen `"pmvd"` nicht funktioniert: `"pmvd"` ist nur in der non-US Version von `relaimpo` verfügbar. Diese Version können Sie von der Homepage der Autorin von `relaimpo` herunterladen.

10.E.5 Informationskriterien

Varianzanalyse, RSS und log-Likelihood Anhang 11.C vergleicht die *unerklärte Varianz* RSS zweier Modelle detaillierter als wir das hier tun. Ein kleines Modell mit RSS_1, und ein größeres Modell mit RSS_2. Das größere Modell kann mehr erklären, also ist RSS_2 kleiner. Varianzanalyse fragt, ob RSS_2 *signifikant* kleiner ist.

$$\frac{RSS_1 - RSS_2}{RSS_2}\frac{n-k_2}{k_2-k_1} \sim F_{(k_2-k_1,n-k_2)}$$

Sei L die log-Likelihood des geschätzten Modells.

Man kann zeigen, dass für lineare Modelle gilt

$$-2\cdot L = n\cdot\log\frac{RSS}{n} + C$$

Dann ist aber

$$\begin{aligned}2\cdot(L_2-L_1) = n\cdot\left(\log\frac{RSS_1}{n} - \log\frac{RSS_2}{n}\right) = \\ = n\log\frac{RSS_1}{RSS_2} \sim \chi^2_{k_2-k_1}\end{aligned}$$

Wenn wir zwei Modelle vergleichen wollen, können wir entweder die log-Likelihood vergleichen, oder die RSS.

```
est2 <- lm(testscr ~ str + elpct + mealpct + calwpct)
est1 <- lm(testscr ~ str +  mealpct + calwpct)
```

```
anova(est1,est2,test="Chisq")
```

	Res.Df	RSS	Df	Sum of Sq	Pr(>Chi)
1	416	35450.80			
2	415	34247.46	1	1203.33	0.0001

Alternativ (und das werden in Anhang 11.C machen) kann man die F-Verteilung benutzen. Die Ergebnisse sind sehr ähnlich:

```
anova(est1,est2)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	416	35450.80				
2	415	34247.46	1	1203.33	14.58	0.0002

Hier wird allerdings die F-Verteilung verwendet. Dazu gehört auch eine andere Formel:

$$\frac{RSS_1 - RSS_2}{RSS_2}\frac{n-k_2}{k_2-k_1} \sim F_{k_2-k_1,n-k_2}$$

Ein ähnliches Verfahren benutzt Informationskriterien Ziel: Finde ein Modell, das die Daten gut erklärt, aber möglichst wenige Parameter hat.

- L log-Likelihood des geschätzten Modells
- k Anzahl der Parameter
- n Anzahl der Beobachtungen

Hirotsugo Akaike (1971): An Information Criterion:

$$\mathrm{AIC} = -2 \cdot \mathrm{L} + 2 \cdot \mathrm{k}$$

Ein anderes Informationskriterium ist das BIC:
Gideon E. Schwarz (1978): Bayesian Information Criterion

$$\mathrm{BIC} = -2 \cdot \mathrm{L} + \mathrm{k} \cdot \log \mathrm{n}$$

In beiden Fällen wird das Kriterium, AIC oder BIC, minimiert. Eine große Likelihood wird belohnt, aber ein großes Model wird bestraft.

Die Funktion `extractAIC` bestimmt das AIC einer Regression.

```
extractAIC(lm(testscr ~ str + elpct + mealpct + calwpct + enrltot))

[1]    6.000 1859.498
```

Wir interessiere uns hier nur für die zweite Zahl im Output, das AIC.
Lassen wir nun die Variable `calwpct` weg, dann verkleinert sich das AIC:

```
extractAIC(lm(testscr ~ str + elpct + mealpct + enrltot))

[1]    5.000 1858.359
```

Dieses Modell wäre also vorzuziehen. Das AIC ist etwas kleiner. Nun kann man verschiedene Modelle ausprobieren, jeweils das AIC berechnen, und nach einem möglichst kleinen Wert suchen

Das Ausprobieren übernimmt die Funktion `step` für uns.

```
step(lm(testscr ~ str + elpct + mealpct + calwpct + enrltot))
```

```
Start:  AIC=1859.5
testscr ~ str + elpct + mealpct + calwpct + enrltot

           Df Sum of Sq   RSS    AIC
- calwpct   1      70.1 34239 1858.4
- enrltot   1      78.9 34247 1858.5
<none>                 34169 1859.5
- elpct     1    1262.6 35431 1872.7
```

```
- str      1    1552.8 35721 1876.2
- mealpct  1   20702.3 54871 2056.4

Step:  AIC=1858.36
testscr ~ str + elpct + mealpct + enrltot

          Df Sum of Sq   RSS    AIC
- enrltot  1        60 34298 1857.1
<none>                 34239 1858.4
- elpct    1      1208 35446 1870.9
- str      1      1496 35734 1874.3
- mealpct  1     51150 85388 2240.2

Step:  AIC=1857.09
testscr ~ str + elpct + mealpct

          Df Sum of Sq   RSS    AIC
<none>                 34298 1857.1
- elpct    1      1167 35465 1869.1
- str      1      1441 35740 1872.4
- mealpct  1     52947 87245 2247.2
```

Der kleinste (beste) Wert des AIC ist also 1857.09 mit den erklärenden Variablen `str`, `elpct`, `mealpct`

Anhang 10.F t-Statistik und p-Wert für individuelle Koeffizienten

Alternativ können wir wichtige von unwichtigen Koeffizienten trennen, indem wir die t-Statistik bzw. den jeweiligen p-Wert betreachten.

$$t_i = \frac{\hat{\beta}_i - \beta_{i,0}}{\hat{\sigma}_{\hat{\beta}_i}}$$

Spielt ein Koeffizient keine Rolle, dann ist $\beta_{i,0} = 0$.

```
xtable(lm(testscr ~ str + elpct + mealpct + calwpct + enrltot))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	701.3779	4.8052	145.96	0.0000
str	-1.0804	0.2491	-4.34	0.0000
elpct	-0.1429	0.0365	-3.91	0.0001
mealpct	-0.5218	0.0329	-15.84	0.0000
calwpct	-0.0568	0.0617	-0.92	0.3572
enrltot	0.0001	0.0001	0.98	0.3287

Hier würden wir also `calwpct` und `enrltot` aus dem Modell streichen.

10.F.1 Vergleich von Modellen

```
library(memisc)
```

Nun schätzen wir einige Modelle:

```
est1 <- lm(testscr ~ str)
est2 <- lm(testscr ~ str + elpct)
est3 <- lm(testscr ~ str + elpct + mealpct)
est4 <- lm(testscr ~ str + elpct + calwpct)
est5 <- lm(testscr ~ str + elpct + mealpct + calwpct)
```

`texreg` aus der library `texreg` stellt verschiedene Modelle übersichtlich nebeneinander dar.

```
library(texreg)
texreg(list(est1,est2,est3,est4,est5))
```

	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	698.93***	686.03***	700.15***	698.00***	700.39***
	(9.47)	(7.41)	(4.69)	(6.02)	(4.70)
str	−2.28***	−1.10**	−1.00***	−1.31***	−1.01***
	(0.48)	(0.38)	(0.24)	(0.31)	(0.24)
elpct		−0.65***	−0.12***	−0.49***	−0.13***
		(0.04)	(0.03)	(0.03)	(0.03)
mealpct			−0.55***		−0.53***
			(0.02)		(0.03)
calwpct				−0.79***	−0.05
				(0.05)	(0.06)
R^2	0.05	0.43	0.77	0.63	0.77
Adj. R^2	0.05	0.42	0.77	0.63	0.77
Num. obs.	420	420	420	420	420

***p < 0.001; **p < 0.01; *p < 0.05

10.F.2 Diskussion

- Kontrolle für Schülercharakteristika halbiert den Koeffizienten von `str`
- Schülercharakteristika sind gute Prediktoren
- Das Vorzeichen der Koeffizienten der Schülercharakteristika stimmt mit den Bildern überein
- Kontrollvariablen sind nicht immer signifikant

  `calwpct` erscheint redundant in diesem Kontext

10.F.3 Literatur

- Verzani, Using R for Introductory Statistics, Chapter 10.3.5.
- Stock and Watson. Introduction to Econometrics, Brief Edition, Chapter 7, 9.

10.F.4 Übungen

Der Datensatz `Wages`:

lwage	logarithm of wage
ed	years of education
exp	years of full-time work experience
married	married
south	resides in the south ?
black	is the individual black ?
wks	weeks worked
union	individual's wage set by a union contract ?
ind	works in a manufacturing industry

Betrachten Sie das folgende Schätzergebnis. Abhängige Variable ist `lwage`.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.7657	0.1223	38.97	0.0000
ed	0.0686	0.0049	13.95	0.0000
exp	0.0110	0.0012	9.13	0.0000
marriedyes	0.2302	0.0344	6.69	0.0000
southyes	-0.0802	0.0283	-2.83	0.0048
blackyes	-0.1255	0.0498	-2.52	0.0121
wks	0.0072	0.0020	3.60	0.0004
unionyes	0.1233	0.0279	4.41	0.0000
ind	0.0135	0.0268	0.50	0.6157

R^2 =0.4037808

Ist das Modell gut spezifiziert?

Betrachten Sie den folgenden Beitrag zum R^2 (lmg):

	x
ed	0.1816
exp	0.0767
married	0.0721
south	0.0248
black	0.0187
wks	0.0147
union	0.0128
ind	0.0024

Welche Variablen beeinflussen den Lohn wesentlich?

Ist `married` ein Proxy für eine andere Variable?

```
est2<-lm(lwage ~  ed + exp + married  + south +  black + wks + union + ind)
est1<-lm(lwage ~  ed + exp + married  + south +  black + wks + union)
anova(est1,est2)

Analysis of Variance Table

Model 1: lwage ~ ed + exp + married + south + black + wks + union
Model 2: lwage ~ ed + exp + married + south + black + wks + union + ind
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    587 53.456
2    586 53.433  1  0.023002 0.2523 0.6157

anova(est1,est2,test="Chisq")

Analysis of Variance Table

Model 1: lwage ~ ed + exp + married + south + black + wks + union
Model 2: lwage ~ ed + exp + married + south + black + wks + union + ind
  Res.Df    RSS Df Sum of Sq Pr(>Chi)
1    587 53.456
2    586 53.433  1  0.023002   0.6155
```

```
est2<-lm(lwage ~  ed + exp + married  + south +  black + wks + union)
est1<-lm(lwage ~  ed + exp + married  + south +  wks + union)
anova(est1,est2)

Analysis of Variance Table

Model 1: lwage ~ ed + exp + married + south + wks + union
Model 2: lwage ~ ed + exp + married + south + black + wks + union
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    588 54.045
2    587 53.456  1   0.58909 6.4688 0.01123 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova(est1,est2,test="Chisq")

Analysis of Variance Table

Model 1: lwage ~ ed + exp + married + south + wks + union
Model 2: lwage ~ ed + exp + married + south + black + wks + union
  Res.Df    RSS Df Sum of Sq Pr(>Chi)
1    588 54.045
2    587 53.456  1   0.58909  0.01098 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
step(est,steps=1)

Start:  AIC=-1416.03
```

```
lwage ~ ed + exp + married + south + black + wks + union + ind

          Df Sum of Sq    RSS     AIC
- ind      1    0.0230 53.456 -1417.8
<none>                 53.433 -1416.0
- black    1    0.5781 54.011 -1411.6
- south    1    0.7306 54.163 -1410.0
- wks      1    1.1784 54.611 -1405.0
- union    1    1.7765 55.209 -1398.6
- married  1    4.0773 57.510 -1374.3
- exp      1    7.5962 61.029 -1338.9
- ed       1   17.7403 71.173 -1247.5

Step:  AIC=-1417.77
lwage ~ ed + exp + married + south + black + wks + union

Call:
lm(formula = lwage ~ ed + exp + married + south + black + wks +
    union)

Coefficients:
(Intercept)           ed          exp   marriedyes     southyes     blackyes
    4.77189      0.06818      0.01106      0.23264     -0.08121     -0.12653
        wks     unionyes
    0.00729      0.12435
```

```
Start:  AIC=-1417.77
lwage ~ ed + exp + married + south + black + wks + union

          Df Sum of Sq    RSS     AIC
<none>                 53.456 -1417.8
- black    1    0.5891 54.045 -1413.2
- south    1    0.7527 54.209 -1411.5
- wks      1    1.1959 54.652 -1406.6
- union    1    1.8155 55.271 -1399.9
- married  1    4.2469 57.703 -1374.3
- exp      1    7.7334 61.189 -1339.4
- ed       1   18.1454 71.601 -1245.9

Call:
lm(formula = lwage ~ ed + exp + married + south + black + wks +
    union)

Coefficients:
(Intercept)           ed          exp   marriedyes     southyes     blackyes
    4.77189      0.06818      0.01106      0.23264     -0.08121     -0.12653
        wks     unionyes
    0.00729      0.12435
```

Übung 10.8 *Sie schätzen den Zusammenhang*

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

und erhalten das folgende Ergebnis:

	β	σ	t	p	*pmvd*
(Intercept)	*-3.00*	*12.00*	*-0.25*	*0.80*	*0.00*
X_1	*23.00*	*21.00*	*1.10*	*0.27*	*0.10*
X_2	*7.00*	*6.00*	*1.17*	*0.24*	*0.50*
X_3	*411.00*	*50.00*	*8.22*	*0.00*	*0.40*

Können Sie Ihre Gleichung vereinfachen?

Übung 10.9 *Sie schätzen einen linearen Zusammenhang zwischen einer abhängigen Variablen `y` und vier unabhängigen Variablen `x1`, `x2`, `x3`, `x4`. Sie spezifizieren ein Modell, das alle vier unabhängigen Variablen enthält und erhalten im ersten Schritt einer schrittweisen Suche nach einem Modell mit einem guten Informationskriterium folgenden Output:*

```
Start:  AIC=214.94
y ~ x1 + x2 + x3 + x4

       Df Sum of Sq    RSS    AIC
- x3    1     8.776 785.12 214.07
- x2    1     9.194 785.54 214.12
- x1    1    11.820 788.17 214.45
<none>              776.35 214.94
- x4    1    63.370 839.72 220.79
```

- *Welche Variable sollten Sie als erste aus Ihrer Spezifikation entfernen?*
- *Welche Variable erscheint am wichtigsten?*

Übung 10.10 *Verwenden Sie den Datensatz `Schooling` aus der Bibliothek `Ecdat`. Erklären Sie den Lohn `wage76` als Funktion der Variablen Ausbildung `ed76`, Erfahrung `exp76`, Alter `age76`, Ausbildung des Vaters `daded`, Ausbildung der Mutter `momed`, Hautfarbe `black`, gemessene Intelligenz `iqscore`.*

Verwenden Sie die in diesem Kapitel diskutierten Verfahren um zu entscheiden, welche Variablen in das Modell aufgenommen werden sollten.

Übung 10.11 *Sie interessieren sich für die Lebenserwartung in den USA in den Jahren 1969-71. Ihr fleißiger Assistent hat Ihnen einen Datensatz zusammengestellt, in dem die Variablen wie folgt bezeichnet sind:*

`Population`	*population estimate as of July 1, 1975*
`Income`	*per capita income (1974)*
`Illiteracy`	*illiteracy (1970, percent of population)*
`Life.Exp`	*life expectancy in years (1969-71)*
`Murder`	*murder and non-negligent manslaughter rate per 100,000 population (1976)*
`HS.Grad`	*percent high-school graduates (1970)*
`Frost`	*mean number of days with minimum temperature below freezing (1931-1960) in capital or large city*
`Area`	*land area in square miles*

Auf der Suche nach einem passenden Modell für die Lebenserwartung `Life.Exp` *führen Sie die folgenden Schritte in R durch:*

```
library(Ecdat)
data(state)
statedata <- data.frame(state.x77, row.names=state.abb)
attach(statedata)
est <- lm(Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad + Frost + Area)
step(est)
```

1. Welches Modell bevorzugen Sie auf Grundlage des AIC-Kriteriums?
2. Nehmen Sie an, Sie entscheiden sich für das Modell mit den Variablen `Intercept`, `Illiteracy`, `Income`, `Population`, `Frost`, `HS.Grad` und `Murder`. Welche Bedeutung hat in Ihrem Output `AIC = -24.18`?
3. Die Formeln für AIC und BIC lauten

$$\mathrm{AIC} = -2 \cdot \mathrm{L} + 2 \cdot \mathrm{k}$$

$$\mathrm{BIC} = -2 \cdot \mathrm{L} + \mathrm{k} \cdot \log \mathrm{n}.$$

   L ist die log-Likelihood des geschätzten Modelles. Interpretieren Sie die einzelnen Terme.

Übung 10.12 *1. Welche Nullhypothese stellen Sie auf, um zu überprüfen, ob eine Variable* β_i *in ein Modell gehört?*

2. Was bedeutet es, wenn Sie die Nullhypothese nicht *ablehnen?*

3. Was bedeutet es, wenn Sie die Nullhypothese ablehnen?

4. Welche Statistik verwenden Sie, um diese Hypothese zu testen? Wie berechnet sich diese Statistik?

Übung 10.13 *Sie führen in R folgende Regressionsanalyse durch:*

```
Absatzmenge <- c(300, 250, 100, 400, 600, 800)
Preis <- c(250, 225, 210, 300, 325, 250)
AusgabenWerbung <- c(600, 550, 450, 750, 900, 1100)
est <- lm (Absatzmenge ~ Preis + AusgabenWerbung)
(t_values <- coef(est)/sqrt(diag(vcov(est))))

    (Intercept)           Preis AusgabenWerbung
     -4.0646354      -0.3102413      15.7876720
```

Ihr Signifikanzniveau beträgt 10%. Verwenden Sie die folgende Tabelle um zu ermitteln, welche Variablen von Null signifikant verschieden sind? Q^N *ist das Quantil der Normalverteilung, und* Q^t_k *ist das Quantil der* t*-Verteilung mit* k *Freiheitsgraden.*

x	0.001	0.0025	0.005	0.01	0.025	0.05	0.1
$Q^N(x)$	-3.090	-2.807	-2.576	-2.326	-1.960	-1.645	-1.282
$Q^t_2(x)$	-22.327	-14.089	-9.925	-6.965	-4.303	-2.920	-1.886
$Q^t_3(x)$	-10.215	-7.453	-5.841	-4.541	-3.182	-2.353	-1.638
$Q^t_4(x)$	-7.173	-5.598	-4.604	-3.747	-2.776	-2.132	-1.533
$Q^t_5(x)$	-5.893	-4.773	-4.032	-3.365	-2.571	-2.015	-1.476

Übung 10.14 *Ein Delikatessenhersteller in Deutschland besitzt 5 Filialen in 5 verschiedenen Städten. Das Geschäft läuft gut; es soll eine weitere Filiale eröffnet werden. Um eine richtige Standortwahl treffen zu können, sollen die externen Erfolgsfaktoren des Gewinns identifiziert werden. Dazu stehen ihnen folgende Daten zur Verfügung:*

```
earn <- c(20000,32500,42000,12000,28000) # Gewinn
hab <- c(12000,72000,164000,60000,16000) # Einwohner
slot <- c(600,3600,8200,3000,800) # Parkplaetze
mil <- c(15,19,24,11,17) # Anzahl Millionaere
```

1. Bestimmen Sie zu einem Signifikanzniveau von $\alpha = 5\%$ die Erfolgsfaktoren des Delikatessenhändlers auf Basis der Ihnen zur Verfügung stehenden Daten. Nutzen Sie dazu eine OLS Regression mit jeweils nur einer unabhängigen Variablen. Welche Variablen sind Erfolgsfaktoren?

2. Betrachten Sie nun die folgenden Regressionen mit mehreren unabhängigen Variablen:

```
est1=lm(earn~hab+slot)
est2=lm(earn~hab+mil)
est3=lm(earn~mil+slot)
est4=lm(earn~hab+slot+mil)
mtable(est1, est2, est3, est4, coef.style="all", summary.stats=c("N"))
```

```
Calls:
est1: lm(formula = earn ~ hab + slot)
est2: lm(formula = earn ~ hab + mil)
est3: lm(formula = earn ~ mil + slot)
est4: lm(formula = earn ~ hab + slot + mil)

=====================================================================
                  est1         est2         est3         est4
---------------------------------------------------------------------
  (Intercept)  18650.509   -14873.813   -14873.813   -14873.813
               (6749.754)   (4066.118)   (4066.118)   (4066.118)
                  (2.763)     (-3.658)     (-3.658)     (-3.658)
                  (0.070)      (0.067)      (0.067)      (0.067)
  hab              0.127       -0.011                    -0.011
                  (0.079)      (0.022)                   (0.022)
                  (1.602)     (-0.482)                  (-0.482)
                  (0.207)      (0.677)                   (0.677)
  mil                         2469.044*    2469.044*    2469.044*
```

```
                              (283.160)    (283.160)    (283.160)
                                (8.720)      (8.720)      (8.720)
                                (0.013)      (0.013)      (0.013)
  slot                                       -0.214
                                             (0.444)
                                            (-0.482)
                                             (0.677)
-------------------------------------------------------------------
  N                 5            5            5            5
===================================================================
  Significance: *** = p < 0.001; ** = p < 0.01; * = p < 0.05
```

Welche Variablen sind nun Erfolgsfaktoren?

3. Gibt es Regressoren die kollinear sind?

Übung 10.15 *Betrachten Sie folgenden Datensatz:*

Y	X	Z
1	2	−1
3	6	3
5	10	−5
7	14	7

Sie schätzen $Y_i = \beta_0 + \beta_1 X_i + u_i$ *mit einem OLS-Schätzer.*

1. *Was ist Ihr Schätzer für* β_1*?*
2. *Wie groß ist Ihr* R^2*?*
3. *Was passiert mit* R^2 *und* AIC*, wenn* Z *in das Modell aufgenommen wird?*

11. Kategoriale Variablen in der linearen Regression

11.1. Metrische und kategoriale Variablen

- Metrisch (Abstände haben eine Bedeutung)
  - Bruttosozialprodukt
  - Einkommen in Euro
  - `str` (17.9, 21.5, 18.7, 17.4, 18.7…)
- Kategorial / diskret
  - Geschlecht (F, M, M, F, M, F, F,…)
  - Beruf (Bäcker, Schneider, Tischler, Bäcker,…)
  - Wirtschaftszweig (Fischerei, Baugewerbe, Kunst, Gastronomie, Baugewerbe,…)

  - Einkommen in Kategorien ($< 30\,000$, $30\,000 - 50\,000$, $> 50\,000$,...)

- Binäre Variablen / Dummy-Variablen (Spezialfall kategorialer Variablen)
  - Geschlecht: M/F
  - Einkommen größer als 40 000 Euro: Ja/Nein
  - Arbeitslosigkeit: Ja/Nein
  - Hochschulabschluss: Ja/Nein

- Oft codiert man binären Variablen als 0 und 1, z.B. Nein=0, Ja=1, oder M=0, F=1.

In der Gleichung

$$\texttt{testsrc} = \beta_0 + \beta_1 \texttt{str} + u$$

war die unabhängige Variable `str` metrisch. Was aber, wenn wir über `str` nur binäre Information haben?

$$\texttt{large} = \begin{cases} 1 & \text{falls } \texttt{str>20} \\ 0 & \text{sonst} \end{cases}$$

Schätze nun

$$\texttt{testsrc} = \beta_0 + \beta_1 \texttt{large} + u$$

```
large <- str>20
lm(testscr ~ large)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	657.1846	1.2023	546.62	0.0000
largeTRUE	-7.1851	1.8520	-3.88	0.0001

Interpretation:

$$\begin{array}{llll} \texttt{str} \leq 20 & \texttt{large} = 0 & \texttt{testscr} \approx \beta_0 = 657.1846 \\ \texttt{str} > 20 & \texttt{large} = 1 & \texttt{testscr} \approx \beta_0 + \beta_1 = 649.9994 \end{array}$$

11.2. Regression mit einer Dummy-Variablen

Wie bisher, auch wenn X eine binäre Variable / Dummy-Variable ist, schätzen wir

$$Y_i = \beta_0 + \beta_1 X_i + u_i \,.$$

Interpretation:

Falls $X_i = 0$: $Y_i = \beta_0 + u_i$

Der Mittelwert $\bar{Y} = \beta_0$

$E(Y_i|X_i = 0) = \beta_0$

Falls $X_i = 1$: $Y_i = \beta_0 + \beta_1 + u_i$

Der Mittelwert $\bar{Y} = \beta_0 + \beta_1$

$E(Y_i|X_i = 1) = \beta_0 + \beta_1$

$\beta_1 = E(Y_i|X_i = 1) - E(Y_i|X_i = 0)$ misst die Differenz der Mittelwerte der beiden Gruppen in der Population. Der aus Kapitel 5 bekannte t-Test macht fast das gleiche:

```
t.test(testscr ~ large)


 Welch Two Sample t-test

data:  testscr by large
t = 3.9231, df = 393.72, p-value = 0.0001031
alternative hypothesis: true difference in means between group FALSE and group TRUE is not equa
95 percent confidence interval:
  3.584445 10.785813
sample estimates:
mean in group FALSE  mean in group TRUE
           657.1846            649.9994
```

zum Vergleich nochmal das Ergebnis der Regression:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	657.1846	1.2023	546.62	0.0000
largeTRUE	-7.1851	1.8520	-3.88	0.0001

Die geschätzten Mittelwerte stimmen überein. Die t-Statistik und die Anzahl der Freiheitsgrade berücksichtigt beim t-Test, die unterschiedliche Varianz in beiden Gruppen. Bei der Berechnung der t-Statistik und p-Werte in der OLS-Regression haben wir angenommen, dass $\text{var}(u)$ nicht von X abhängt. Deshalb gibt es hier kleine Abweichungen.

- Es ist egal, ob wir mit einem Student t Test Mittelwerte zwischen *zwei Gruppen* vergleichen, oder eine Regression mit einer *binären Variablen* rechnen.
- Es ist auch egal, ob wir mit einer Varianzanalyse (F-Test) (siehe Abschnitt 11.C) Mittelwerte zwischen *mehr als zwei Gruppen* vergleichen, oder eine Regression mit einer *kategorialen Variablen* rechnen.

Warum macht man überhaupt eine Regression mit Dummy-Variablen (und nicht einfach einen t-Test):

→ Man kann weitere erklärende Variablen einführen (und damit für weitere Effekte kontrollieren)

11.2.1. Mehr als zwei Kategorien

Wenn eine Variable mehr als zwei Werte annehmen kann (Farbe=rot, grün, blau) brauchen wir für jede weitere Kategorie eine weitere Dummy-Variable.

Regression mit konstantem Term β_0:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

	X_1	X_2	
rot	0	0	$E(Y) = \beta_0$
grün	1	0	$E(Y) = \beta_0 + \beta_1$
blau	0	1	$E(Y) = \beta_0 + \beta_2$

Hier ist »rot« die Referenzkategorie.

```
lm(Y ~ X1 + X2)
```

Regression ohne konstanten Term:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

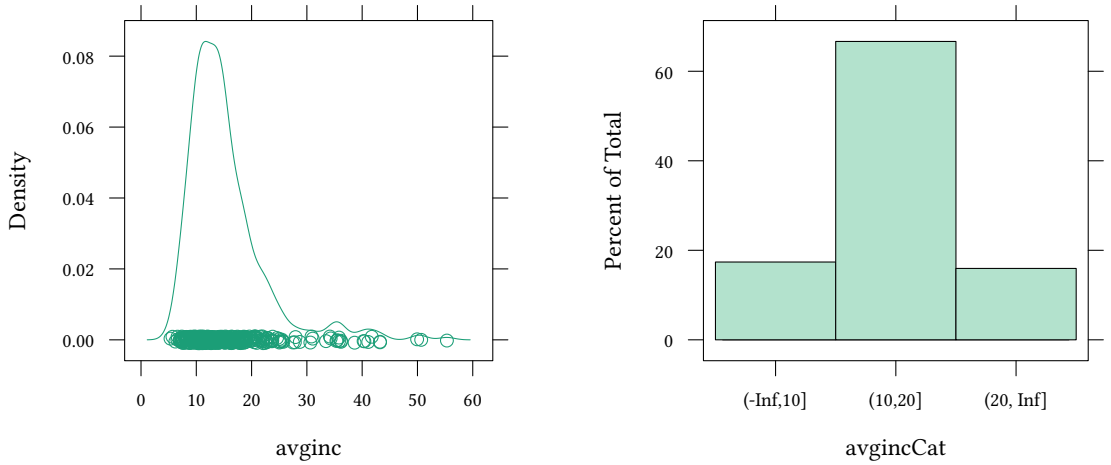
	X_1	X_2	X_3	
rot	1	0	0	$E(Y) = \beta_1$
grün	0	1	0	$E(Y) = \beta_2$
blau	0	0	1	$E(Y) = \beta_3$

Hier gibt es keine Referenzkategorie.

```
lm(Y ~ X1 + X2 + X3  - 1)
```

Beispiel: Zur Illustration schauen wir uns den Einfluss des Durchschnittseinkommens `avginc` auf `testscr` an. Im nächsten Kapitel werden wir das etwas ausführlicher tun, hier vergleichen wir nur drei Kategorien:

```
avgincCat <- cut(avginc,c(-Inf,10,20,Inf))
```




```
coef(lm(testscr ~ avgincCat))

       (Intercept)   avgincCat(10,20] avgincCat(20, Inf]
         635.11028          18.33901          42.75391
```

Die Notation `-1` in der Formel der Regression bedeutet, dass keine Konstante hinzugefügt werden soll.

```
coef(lm(testscr ~ avgincCat - 1))

avgincCat(-Inf,10]   avgincCat(10,20] avgincCat(20, Inf]
          635.1103          653.4493          677.8642
```

Wenn man eine andere Referenzkategorie haben will:

```
avgincCat2 <- relevel(avgincCat,"(20, Inf]")
coef(lm(testscr ~ avgincCat2))

        (Intercept) avgincCat2(-Inf,10]   avgincCat2(10,20]
          677.86419          -42.75391          -24.41490
```

11.3. Interaktionen

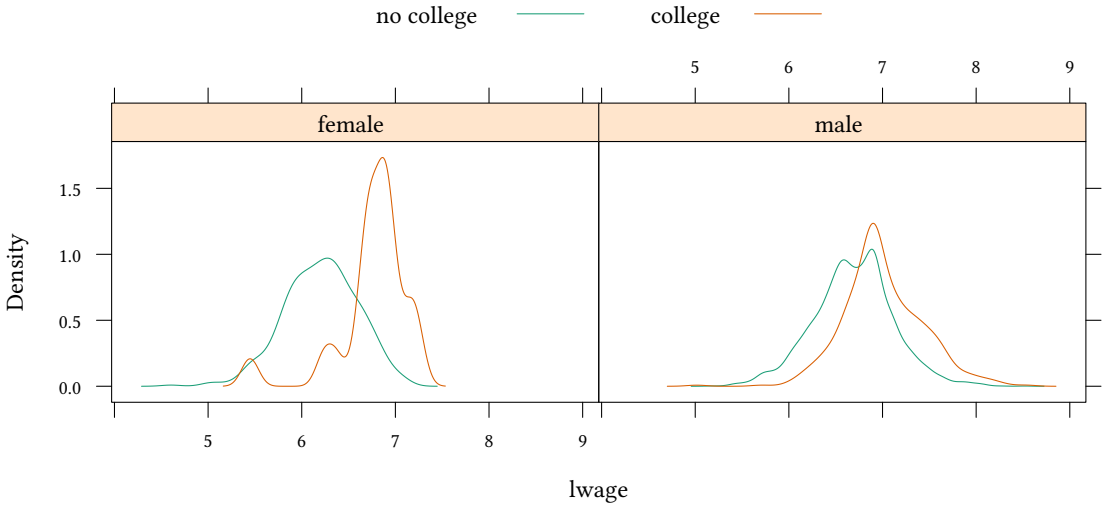
Bislang: $\texttt{testscr} = \beta_0 + \beta_1 \texttt{str} + \beta_2 \texttt{elpct} + \ldots + u$

- Welche Effekte beeinflussen `testscr` ?
- Wie groß sind diese Effekte?
- Annahme: Die Effekte sind konstant, sie beeinflussen sich nicht gegenseitig.

Jetzt:

- Vielleicht hängt der Effekt der Klassengröße auf den Testscore von weiteren Umständen ab?
- Vielleicht sind kleine Klassen gerade bei vielen Ausländern in der Klasse hilfreich, sonst aber nicht?
- $\frac{\partial \texttt{testscr}}{\partial \texttt{str}}$ hängt von `elpct` ab.
- Allgemein: $\frac{\partial Y}{\partial X_1}$ hängt von X_2 ab.
- Wie kann man diese »Interaktion« modellieren?
- Betrachte zunachst binäre X, dann stetige.

Beispiel: College education In der Abbildung sehen wir die geschätzte Dichtefunktion von `lwage` (Logarithmus des Lohns) abhängig vom Geschlecht und von der Ausbildung. Ausbildung hat offensichtlich (im Mittel) einen positiven Effekt auf den Lohn, allerdings scheint die Größe dieses Effekts vom Geschlecht abzuhängen.



Beispiel:

$$\texttt{lwage} = \beta_1 \texttt{college} + \beta_2 \texttt{sex} + \beta_0 + u$$

In diesem Modell ist der Einfluss von `college` unabhängig von `sex`.

Wir laden den Datensatz `Wages` und generieren eine Variable `college`. Die hat den Wert `TRUE` wenn `ed>16`, also wenn die Ausbildungsdauer länger als 16 Jahre ist. R codiert nun `college=TRUE` als 1 und `college=FALSE` als 0. Die Variable `sex` hat die Werte `male` und `female`. R codiert den Wert `sex=male` als 1 und `sex=female` als 0.

```
data(Wages,package="Ecdat")
attach(Wages)
college<-ed>16
```

```
lm(lwage ~ college + sex)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.2254	0.0196	317.56	0.0000
collegeTRUE	0.3340	0.0201	16.62	0.0000
sexmale	0.4626	0.0207	22.31	0.0000

Hier nehmen wir an, das der Effekt von `college` immer gleich ist, egal welchen Wert die Variable `sex` hat.

Diese Annahme geben wir auf, wenn wir eine Interaktion zwischen `sex` und `college` einführen.

11.3.1. Interaktion zwischen binären Variablen

$$\texttt{lwage} = \underbrace{\beta_0}_{6.21} + \underbrace{\beta_1}_{0.55}\texttt{college} + \underbrace{\beta_2}_{0.49}\texttt{sex} + \underbrace{\beta_3}_{-0.24}\texttt{sex}\cdot\texttt{college} + u$$

```
lm(lwage ~ college + sex + sex:college)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.2057	0.0204	303.70	0.0000
collegeTRUE	0.5543	0.0683	8.12	0.0000
sexmale	0.4850	0.0217	22.31	0.0000
collegeTRUE:sexmale	-0.2412	0.0714	-3.38	0.0007

Kürzer kann man auch schreiben:

```
lm(lwage ~ sex*college)
```

Anstelle der Koeffizienten der Regression können wir auch Mittelwerte der einzelnen Kategorien berechnen:

`genTable` berechnet uns gleich mehrere Mittelwerte für mehrere Kategorien.

```
library(memisc)
```

```
genTable(mean(lwage) ~ college + sex)
```

	female	male
no coll.	6.21	6.69
college	6.76	7.00

	female	male
no coll.	β_0	$\beta_0 + \beta_2$
college	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$

- Effekt von College bei Frauen: $\beta_1 = 0.55$
- Effekt von College bei Männern: $\beta_1 + \beta_3 = 0.31$

11.3.2. Interaktion von diskreten Variablen im Bayesianischen Modell

Natürlich können wir auch hier mit `MCMCregress` eine Bayesianische Schätzung berechnen:

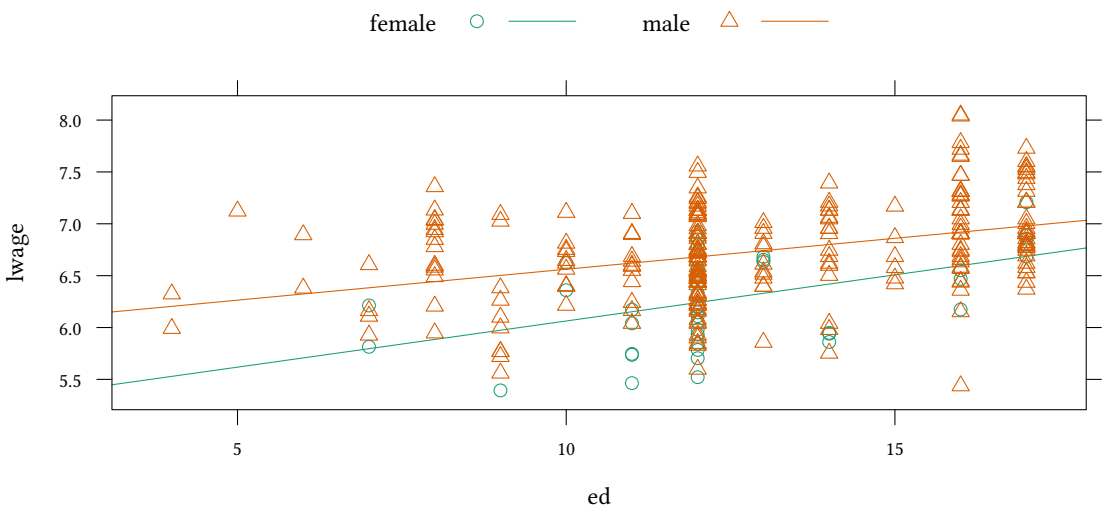
```
library(MCMCpack)
summary(MCMCregress(lwage ~ college * sex,data=Caschool))


Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

                       Mean       SD   Naive SE Time-series SE
(Intercept)          6.2055 0.020412 0.00020412     0.00020412
collegeTRUE          0.5537 0.069120 0.00069120     0.00069120
sexmale              0.4851 0.021797 0.00021797     0.00021797
collegeTRUE:sexmale -0.2403 0.072120 0.00072120     0.00072120
sigma2               0.1783 0.003902 0.00003902     0.00003987

2. Quantiles for each variable:

                       2.5%     25%     50%     75%    97.5%
(Intercept)          6.1661  6.1918  6.2057  6.2190  6.24680
collegeTRUE          0.4171  0.5070  0.5533  0.5997  0.68793
sexmale              0.4414  0.4706  0.4853  0.4996  0.52759
collegeTRUE:sexmale -0.3805 -0.2893 -0.2401 -0.1908 -0.09726
sigma2               0.1709  0.1757  0.1783  0.1809  0.18622
```

Auch hier ist das Ergebnis der Bayesianischen Schätzung des linearen Modells sehr ähnlich dem Ergebnis von `lm`. Anders als `lm` erhalten wir wieder Quantile für die Verteilung der Parameter, d.h. wir wissen, in welchem Intervall die Parameter mit welcher Wahrscheinlichkeit liegen.

11.3.3. Interaktion zwischen einer binären und einer stetigen Variablen

Oben haben wir mit `college` eine diskrete Variable eingeführt. In diesem Abschnitt verwenden wir die ursprüngliche Variable `ed`.

In der Graphik sehen wir, dass der Zusammenhang zwischen Ausbildung und Lohn von `sex` abhängt. Die Steigung der Regressionsgerade für `sex=="female"` ist steiler.



Die Steigung der beiden Regressionsgeraden ist unterschiedlich. Wieder scheint Ausbildung bei Frauen einen stärkeren Effekt zu haben.

$$\texttt{lwage} = \beta_0 + \beta_1 \texttt{sex} + \beta_2 \texttt{ed} + \beta_3 \texttt{sex} \cdot \texttt{ed} + \mathfrak{u}$$

```
lm(lwage ~ sex*ed)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.0421	0.0956	52.72	0.0000
sexmale	0.8885	0.1003	8.85	0.0000
ed	0.0945	0.0073	12.92	0.0000
sexmale:ed	-0.0323	0.0077	-4.21	0.0000

- $\beta_3 = 0$: Regressionsgeraden sind parallel
- $\beta_1 = 0$: Regressionsgeraden haben den gleichen Achsenabschnitt

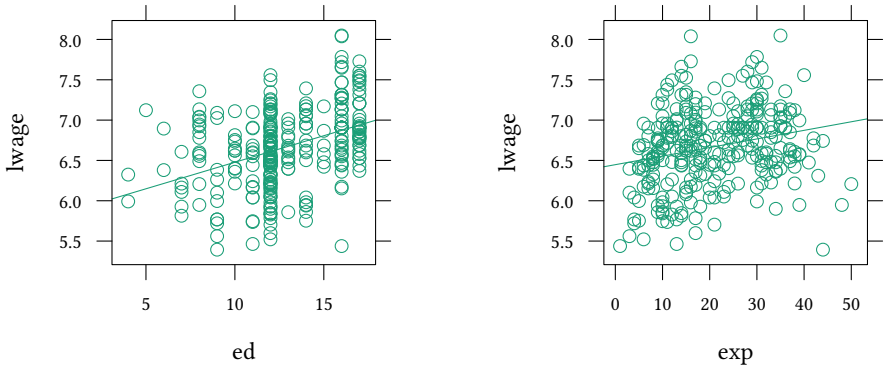
11.3.4. Interaktion zwischen einer binären und einer stetigen Variablen im Bayesianischen Modell

Auch hier das Ergebnis der Bayesianischen Schätzung des linearen Modells sehr ähnlich dem Ergebnis von `lm`:

```
library(MCMCpack)
summary(MCMCregress(lwage ~ sex * ed,data=Caschool))


Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
```

```
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

                Mean       SD   Naive SE Time-series SE
(Intercept)  5.04119 0.095543 0.00095543     0.00095543
sexmale      0.88901 0.100729 0.00100729     0.00100729
ed           0.09458 0.007323 0.00007323     0.00007323
sexmale:ed  -0.03233 0.007713 0.00007713     0.00007713
sigma2       0.15700 0.003436 0.00003436     0.00003511

2. Quantiles for each variable:

                2.5%      25%      50%      75%    97.5%
(Intercept)  4.85674  4.97733  5.04211  5.10427  5.23463
sexmale      0.68990  0.82286  0.88882  0.95753  1.08353
ed           0.07985  0.08974  0.09461  0.09951  0.10884
sexmale:ed  -0.04736 -0.03755 -0.03244 -0.02719 -0.01706
sigma2       0.15042  0.15464  0.15694  0.15928  0.16395
```

11.3.5. Interaktion zwischen zwei stetigen Variablen



Sowohl ed als auch exp beeinflussen lwage positiv. Wird der Einfluss von ed vielleicht kleiner wenn exp groß ist?

Beispiel

$$\texttt{lwage} = \beta_0 + \beta_1 \texttt{ed} + \beta_2 \texttt{exp} + \beta_3 \texttt{ed} \cdot \texttt{exp} + u$$

```
est1 <- lm(lwage ~ ed + exp)
est2 <- lm(lwage ~ ed * exp)
```

	Model 1	Model 2
(Intercept)	5.435861***	5.446193***
	(0.034319)	(0.066655)
ed	0.076403***	0.075608***
	(0.002282)	(0.004955)
exp	0.013048***	0.012589***
	(0.000580)	(0.002607)
ed:exp		0.000036
		(0.000201)
R^2	0.246711	0.246717
Adj. R^2	0.246349	0.246174
Num. obs.	4165	4165

***p < 0.001; **p < 0.01; *p < 0.05

```
detach(Wages)
```

Interaktion zwischen zwei stetigen Variablen, allgemein

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \cdot X_2) + u$$

Marginale Effekte:

$$\frac{\partial Y}{\partial X_1} = \beta_1 + \beta_3 X_2 \qquad \frac{\partial Y}{\partial X_2} = \beta_2 + \beta_3 X_1$$

$$\texttt{testscr} = \beta_0 + \beta_1 \texttt{str} + \beta_2 \texttt{elpct} + \beta_3 \texttt{str} \cdot \texttt{elpct} + u$$

```
lm(testscr ~ str * elpct)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	686.3385	9.4026	72.99	0.0000
str	-1.1170	0.4825	-2.31	0.0211
elpct	-0.6729	0.4380	-1.54	0.1252
str:elpct	0.0012	0.0219	0.05	0.9577

11.4. Literatur

- Stock and Watson. Introduction to Econometrics, Brief Edition, Chapter 7, 9.

11.5. Schlüsselbegriffe

- metrische (stetige) Variablen
- kategoriale (diskrete) Variablen

- binäre Variablen
- Dummies
- Interaktionen
- Interpretation der geschätzten Koeffizienten:
  - Koeffizienten von Dummies
  - Koeffizienten von Interaktionen

Anhang 11.A Beispiele für die Vorlesung

Betrachten Sie das folgende Regressionmodell:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + u$$

X_1 und X_2 sind jeweils Dummyvariablen die entweder den Wert 0 oder 1 haben. Sie messen die folgenden Mittelwerte für Y:

X_1	X_2	$\bar{Y}$
0	0	3
1	0	4
0	1	5
1	1	6

1. Welchen Wert hat $\hat{\beta}_0$?
2. Welchen Wert hat $\hat{\beta}_1$?
3. Welchen Wert hat $\hat{\beta}_2$?
4. Welchen Wert hat $\hat{\beta}_3$?

Betrachten Sie das folgende Ergebnis einer Regression:

```
lm(formula = y ~ x1 * x2)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.0        2.0     0.5  0.61822
x1               7.5        0.5    15.0  < 2e-16
x2              -4.0        4.0    -1.0  0.31982
x1:x2            3.0        1.0     3.0  0.00344
```

1. Die Variable x2 ist eine Dummy-Variable. Sie hat den Wert 0 für Männer und den Wert 1 für Frauen. Wie groß ist der marginale Effekt von x1 für Männer?
2. Wie groß ist der marginale Effekt von x1 für Frauen?
3. Ein Kollege rechnet mit den gleichen Daten, allerdings codiert er x2 anders. Bei ihm hat diese Variable den Wert 1 für Männer und 2 für Frauen. Wie groß ist in diesem Fall der marginale Effekt von x1 für Frauen?

Anhang 11.B Übungen

Übung 11.1 *Sie erklären das Einkommen von Arbeitslosen nach Teilnahme an einer Fördermaßnahme als lineare Funktion eines Dummies (der die Art der Maßnahme beschreibt). Es gibt drei Maßnahmen, codiert als* `MassnA`*,* `MassnB` *und* `MassnC`*. Außerdem kann es sein, dass ein Arbeitsloser an keiner Maßnahme teilnimmt. Dieser Fall wird als* `Massn0` *codiert. Die Basiskategorie ist* `MassnA`*. Sie schätzen die folgenden Koeffizienten:*

```
(Intercept)   7.3
MassnB       -0.3
MassnC        1.2
Massn0       -3.2
```

1. *Welches Einkommen erwarten Sie für einen Absolventen von* `MassnA`*?*
2. *Welches Einkommen erwarten Sie für einen Arbeitslosen, der an keiner Maßnahme teilgenommen hat?*
3. *Die Kosten der jeweiligen Maßnahmen sind für* $C_{MassnA} = 1.8$*,* $C_{MassnB} = 1.6$*,* $C_{MassnC} = 3.3$*. Welche Maßnahme verspricht die größte Differenz zwischen Einkommenssteigerung und Kosten der Fördermaßnahme? Wie groß ist diese Differenz?*

Übung 11.2 *Sie planen, einen Imbissstand in Jena zu eröffnen. Sie wissen allerdings weder, wo Sie den Stand eröffnen wollen, noch welches Produkt Sie verkaufen wollen. Eine erste Studie mit einem mobilen Stand und wechselnden Produkten ergibt den folgenden durchschnittlichen Umsatz (jeweils pro Tag):*

	Weißwurst	*Bratwurst*
Jena-West	*3000 €*	*4000 €*
Jena-Ost	*2500 €*	*4500 €*

Auf Basis des gleichen Datensatzes schätzen Sie auch eine Regression

$$Y = \beta_0 + \beta_1 \cdot d_B + \beta_2 \cdot d_O + \beta_3 \cdot d_B \cdot d_O + u$$

Dabei ist Y *der Umsatz pro Tag,* d_B *ein Dummy, der den Wert Eins annimmt, falls an diesem Stand gerade Bratwurst verkauft wird und sonst Null ist, und* d_O *ein Dummy, der den Wert Eins annimmt, falls sich der Stand in Jena-Ost befindet und sonst Null ist.*

1. *Welche Werte werden Sie für* β_0*,* β_1*,* β_2*, und* β_3 *schätzen?*
2. *Sie führen eine weitere Untersuchung nur in Jena-West durch, diesmal untersuchen Sie allerdings drei verschiedene Speisenangebote: Weißwurst, Bratwurst, und Currywurst. Dazu führen Sie drei Dummyvariablen ein:* d_W *ist nur Eins am Weißwurststand,* d_B *ist nur Eins am Bratwurststand, und* d_C *ist nur Eins am Currywurststand. Ansonsten haben die Dummys den Wert Null. Sie schätzen die folgende Gleichung:*

$$Y = \beta_0 + \beta_1 \cdot d_W + \beta_2 \cdot d_B + \beta_3 \cdot d_C + u$$

a) Welches Problem tritt auf?

b) Was könnte man besser machen?

Übung 11.3 *Betrachten Sie den Datensatz* `Caschool` *und die folgende Schätzgleichung:*

$$\begin{aligned}\texttt{testscr} = &\underbrace{\beta_0}_{686.34} + \underbrace{\beta_1}_{-1.1170}\,\texttt{str} + \underbrace{\beta_2}_{-0.6729}\,\texttt{elpct} + \\ &+ \underbrace{\beta_3}_{0.001162}\,\texttt{str}\cdot\texttt{elpct} + u\end{aligned}$$

1. *Was ist der Effekt der Klassengröße* `str` *für eine Klasse mit Medianausländeranteil?*
2. *Wie ändert sich dieser Effekt für eine Klasse mit Ausländeranteil im 75%-Quantil?*
3. *Ist der Interaktionsterm signifikant?*

Übung 11.4 *Betrachten Sie wieder den Datensatz* `Caschool`*. Ist der Effekt der Klassengröße auf Testscores davon abhängig ob der Anteil der Mutterspachler größer oder kleiner 10% ist?*

Übung 11.5 *Ordnen Sie den Typ der folgenden Variablen zu:*

	stetig	*diskret*	*binär*
Studienabschluss (Diplom, Bachlor, Master)			
Größe in cm			
Ehestatus			

Übung 11.6

1. *Wenn die abhängige Variable eine diskrete Variable ist, welche der Annahmen des Regressionsmodells ist dann verletzt?*
2. *Wenn eine unabhängige Variable eine diskrete Variable ist, ist dann eine der Annahmen des OLS Modells verletzt?*

Übung 11.7 *Sie möchten untersuchen, ob das Geschlecht eines Arbeitnehmers Einfluss auf das Bruttoeinkommen hat. Dazu führen Sie eine lineare Regression durch und führen für das Geschlecht eine Dummy-Variable ein. Die Variable* `Maennlich` *nimmt den Wert* 1 *an, wenn der Arbeitnehmer männlich ist. Ist die Arbeitnehmerin weiblich, nimmt die Dummy-Variable den Wert* 0 *an. Sie haben Ihre Analyse in R durchgeführt und folgendes Ergebnis erhalten:*

```
Bruttoeinkommen<-c(2500,4000,3000,7000,5000,2900,1500)
Maennlich<-c(1,0,0,0,0,1,1)
```

```
summary(lm(Bruttoeinkommen~Maennlich))
```

```
Call:
lm(formula = Bruttoeinkommen ~ Maennlich)

Residuals:
    1     2     3     4     5     6     7
  200  -750 -1750  2250   250   600  -800

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4750.0      699.6   6.789  0.00105 **
Maennlich    -2450.0     1068.7  -2.292  0.07043 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1399 on 5 degrees of freedom
Multiple R-squared:  0.5125, Adjusted R-squared:  0.4149
F-statistic: 5.255 on 1 and 5 DF,  p-value: 0.07043
```

1. *Welche ist die abhängige und welche die unabhängige Variable?*
2. *Wie lautet die Regressionsgleichung?*
3. *Welchen Wert nimmt* R^2 *an?*
4. *Wie groß ist das durchschnittliche Bruttoeinkommen der Männer?*
5. *Wie groß ist das durchschnittliche Bruttoeinkommen der Frauen?*

Übung 11.8 *Sie interessieren sich für die Einkommensunterschiede zwischen verheiraten und unverheirateten Frauen und Männern in Deutschland. In einer Studie erhalten Sie das folgende durchschnittliche monatliche Nettoeinkommen:*

	Verheiratet	*Unverheiratet*
Frau	*2000*	*2500*
Mann	*3500*	*3000*

Auf Basis des gleichen Datensatzes schätzen Sie auch eine Regression:

$$Y = \beta_0 + \beta_1 \cdot d_F + \beta_2 \cdot d_V + \beta_3 \cdot d_F \cdot d_V + u$$

Dabei ist Y *das monatliche Nettoeinkommen,* d_F *ein Dummy der den Wert Eins annimmt, falls die betrachte Person eine Frau ist und bei Männern Null ist, und* d_V *ein Dummy, der den Wert Eins annimmt, falls die betrachtete Person verheiratet ist und sonst Null ist.*

1. *Wie groß sind* β_0, β_1, β_2 *und* β_3*?*
2. *Wie groß ist das durchschnittliche Nettoeinkommen der Männer?*

Übung 11.9 *Sie untersuchen die Wirksamkeit von zwei Marketingmaßnahmen mit Hilfe einer linearen Regression. Ihre abhängige Variable* `y` *ist der Umsatz. Die unabhängigen Variablen* `x1` *und* `x2` *sind jeweils 1 wenn Maßnahme 1 bzw. 2 eingesetzt wurde und 0 sonst. Sie erhalten folgendes Ergebnis:*

```
lm(formula = y ~ x1 * x2)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.5        0.6      2.5     0.014 *
x1            3.0        0.8      3.75    0.000 ***
x2            0.5        0.8      0.625   0.533
x1:x2        -3.2        1.0     -3.2     0.002 ***
```

Welchen Umsatz erwarten Sie, wenn Sie gleichzeitig Maßnahme 1 und Maßnahme 2 einsetzen?

Übung 11.10 *Eine Unternehmensberatung misst die Steigerung des Gewinns nach Beratung von 12 Unternehmen als* `S`*. Sie betrachten zwei mögliche Faktoren, die den Erfolg der Beratung vielleicht beeinflussen: Dauer der Beratung* `L` *in Stunden und Sektor des Unternehmens. Der Sektor ist codiert als* `D=1` *für Dienstleistung und* `D=0` *sonst. Sie erhalten das folgende Ergebnis:*

```
 lm(formula = S ~ L * D)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   10           5      2       0.081 .
L             20           5      4       0.004 ***
D            -10          10     -1       0.347
L:D           -5           1     -5       0.001 ***
```

1. *Ein Unternehmen aus dem Sektor »Dienstleistung« wird 100 Stunden beraten. Welche Steigerung des Gewinns erwarten Sie?*
2. *Ein Unternehmen aus einem anderen Sektor (nicht »Dienstleistung«) wird ebenfalls beraten. Wie groß ist die marginale Steigerung des Gewinns pro Stunde?*
3. *Ihr Signifikanzniveau ist 5%. Welche der folgenden Aussagen treffen zu?*
   - *Die durch Beratung erzielte Steigerung des Gewinns hängt nicht signifikant vom Sektor ab.*
   - *Die Dauer der Beratung hat einen signifikanten Einfluss auf den Gewinn.*
   - *Die* `t`*-Statistik folgt immer einer* `t`*-Verteilung, egal welcher Verteilung die Störterme folgen.*
   - *Die obigen Schätzergebnisse zeigen, dass man den Interaktionsterm zwischen* `L` *und* `D` *besser weglassen sollte.*
   - *Wenn man* `L` *aus der Schätzgleichung weglässt, wird das* R^2 *voraussichtlich sinken.*

Übung 11.11 *Eine Gruppe von Sportlern bereitet sich auf einen Wettkampf vor. Über die Sportler liegen folgende Informationen vor: Alter (*A*), Geschlecht (*G*; 1 falls weiblich, 0 sonst), tägliches Training (*T*, 1 falls ja, 0 sonst), gesunde Ernährung (*E*; 1 falls ja, 0 sonst) und Ranglistenpunkte (*R*). Alter und Geschlecht sind mit den anderen Variablen nicht korreliert. Sie vermuten, dass Sportler nur dann besonders viele Ranglistenpunkte haben, wenn sie täglich trainieren und sich gesund ernähren; ein tägliches Training wirkt nur zusammen mit gesunder Ernährung. Was wären mögliche Spezifikationen des Modells um Ihren Verdacht zu überprüfen. (Hier ist* nicht *nach der »besten« Spezifikation gefragt)*

- $R = \beta_0 + \beta_1 \cdot T + \beta_2 \cdot E + u$
- $T \cdot E = \beta_0 + \beta_1 \cdot R + u$
- $R = \beta_0 + \beta_1 \cdot A + \beta_2 \cdot G + \beta_3 \cdot T + \beta_4 \cdot E + u$
- $R = \beta_0 + \beta_1 \cdot A + \beta_2 \cdot G + \beta_3 \cdot T + \beta_4 \cdot E + \beta_5 \cdot T \cdot E + u$
- $R = \beta_0 + \beta_1 \cdot T + \beta_2 \cdot E + \beta_3 \cdot T \cdot E + u$

Anhang 11.C Nichtlineare Interaktionsterme

```
Hiel=elpct>=10
est1 <- lm(testscr ~ str + elpct + mealpct)
est2 <- lm(testscr ~ str + elpct + mealpct + log(avginc))
est3 <- lm(testscr ~ str * Hiel )
est4 <- lm(testscr ~ str * Hiel + mealpct + log(avginc))
est5 <- lm(testscr ~ str + I(str^2) + I(str^3) + Hiel + mealpct + log(avginc))
est6 <- lm(testscr ~ (str + I(str^2) + I(str^3))*Hiel + mealpct + log(avginc))
est7 <- lm(testscr ~ str + I(str^2) + I(str^3) + elpct + mealpct + log(avginc))
```

```
texreg(list(`(1)`=est1,`(2)`=est2,`(3)`=est3,`(4)`=est4,`(5)`=est5,`(6)`=est6,`(7)`=est7))
```

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(Intercept)	700.15***	658.55***	682.25***	653.67***	252.05	122.35	244.81
	(4.69)	(7.68)	(10.51)	(8.89)	(165.82)	(192.18)	(165.93)
str	−1.00***	−0.73**	−0.97	−0.53	64.34*	83.70**	65.28*
	(0.24)	(0.23)	(0.54)	(0.30)	(25.46)	(29.69)	(25.48)
elpct	−0.12***	−0.18***					−0.17***
	(0.03)	(0.03)					(0.03)
mealpct	−0.55***	−0.40***		−0.41***	−0.42***	−0.42***	−0.40***
	(0.02)	(0.03)		(0.03)	(0.03)	(0.03)	(0.03)
log(avginc)		11.57***		12.12***	11.75***	11.80***	11.51***
		(1.74)		(1.77)	(1.73)	(1.75)	(1.73)
HielTRUE			5.64	5.50	−5.47***	816.07	
			(16.72)	(9.14)	(1.03)	(434.61)	
str:HielTRUE			−1.28	−0.58		−123.28	
			(0.84)	(0.46)		(66.35)	
str^2					−3.42**	−4.38**	−3.47**
					(1.29)	(1.51)	(1.29)
str^3					0.06**	0.07**	0.06**
					(0.02)	(0.03)	(0.02)
str^2:HielTRUE						6.12	
						(3.35)	
str^3:HielTRUE						−0.10	
						(0.06)	
R^2	0.77	0.80	0.31	0.80	0.80	0.80	0.80
Adj. R^2	0.77	0.79	0.31	0.79	0.80	0.80	0.80
Num. obs.	420	420	420	420	420	420	420

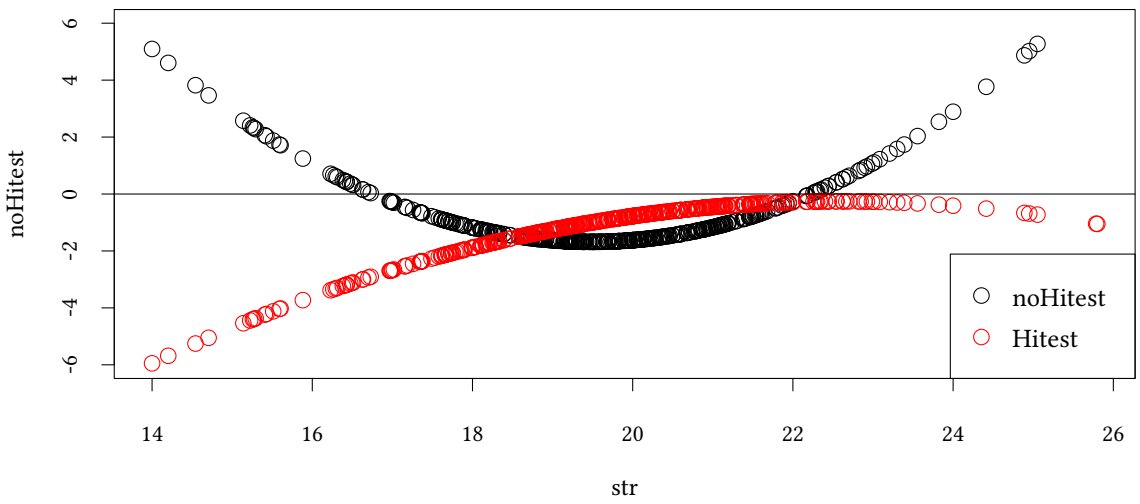
***p < 0.001; **p < 0.01; *p < 0.05

Es sieht so aus, als gäbe es einen nichtlinearen Effect von `str` auf `testscr`. Betrachten wir nochmals Modell 6 und bestimmen den marginalen Effekt von `str`.

```
estC <- coef(est6)
mEffstr <- function (str,Hiel) {
  estC %*% c(0,1,2*str,3*str^2,0,0,0,Hiel,Hiel*2*str,Hiel*3*str^2)
}
mEffstr(20,0)

          [,1]
[1,] -1.622543

mEffstr(20,1)

           [,1]
[1,] -0.7771982
```

sapply wendet eine Funktion auf jedes Element eines Vektors an. In einer Formel sorgt `I()` dafür, dass ein Ausdruck nicht als Interaktion o.ä. nterpretiert wird.

```
noHitest=sapply(str,function(x) {mEffstr(x,0)})
Hitest=sapply(str,function(x) {mEffstr(x,1)})
plot(noHitest ~ str,ylim=c(-6,6))
points(Hitest ~ str,col="red")
abline(h=0)
legend("bottomright",c("noHitest","Hitest"),pch=1,col=c("black","red"))
```



11.C.1 Anwendung: Gender gap

exp	years of full-time work experience
wks	weeks worked
bluecol	blue collar ?
ind	works in a manufacturing industry ?
south	resides in the south ?
smsa	resides in a standard metropolitan statistical area ?
married	married ?
sex	a factor with levels (male,female)
union	individual's wage set by a union contract ?
ed	years of education
black	is the individual black ?
lwage	logarithm of wage

`ifelse` gibt, abhängig vom ersten Argument, entweder das zweite oder dritte Argument zurück. `as.data.frame` wandelt das Argument, z.B. eine Matrix, in einen Dataframe um. Das ist hier nützlich, weil die ausgegebene Struktur Zahlen und Zeichenketten mischt. `colnames` erlaubt es, auf Spaltennamen zuzugreifen (und diese hier zu verändern).

```
#library(lattice)
#data(Wages)
attach(Wages)
library(lmtest)
lmr <- function(...) {
  est<- lm(...)
  print(coeftest(est,vcov=hccm))
  cat("R2=         ",round(summary(est)$r.squared,2),"\n")
}
#$
est1 <- lm(lwage ~ ed)
est2 <- lm(lwage ~ ed + sex)
est3 <- lm(lwage ~ ed * sex)
est4 <- lm(lwage ~ ed * sex + exp + black*sex + south  + married)
```

```
mtable(est1,est2,est3,est4,summary.stats=c("R-squared","N"))
```

	Model 1	Model 2	Model 3	Model 4
(Intercept)	5.84***	5.42***	5.04***	4.93***
	(0.03)	(0.03)	(0.10)	(0.10)
ed	0.07***	0.07***	0.09***	0.09***
	(0.00)	(0.00)	(0.01)	(0.01)
sexmale		0.47***	0.89***	0.64***
		(0.02)	(0.10)	(0.11)
ed:sexmale			−0.03***	−0.02**
			(0.01)	(0.01)
exp				0.01***
				(0.00)
blackyes				−0.10*
				(0.04)
southyes				−0.09***
				(0.01)
marriedyes				0.06**
				(0.02)
sexmale:blackyes				−0.03
				(0.05)
R^2	0.16	0.26	0.26	0.35
Adj. R^2	0.15	0.26	0.26	0.35
Num. obs.	4165	4165	4165	4165

***p < 0.001; **p < 0.01; *p < 0.05

```
detach(Wages)
#unloadNamespace("memisc")
detach(package:memisc)
```

Anhang 11.C Varianzanalyse

Oben haben wir untersucht, wie eine metrische Variable von einer (oder mehreren) kategorialen Variablen abhängt.

(zur Erinnerung: Eine kategoriale Variable hat nur wenige Ausprägungen, z.B. männlich/weiblich, rot/grün/gelb,...)

Wir haben den Einfluss jeder einzelnen Ausprägung geschätzt.

Manchmal interessiert man sich gar nicht für die einzelnen Ausprägungen, sondern will nur wissen, ob die kategoriale einen signifikanten Teil der Varianz erklärt.

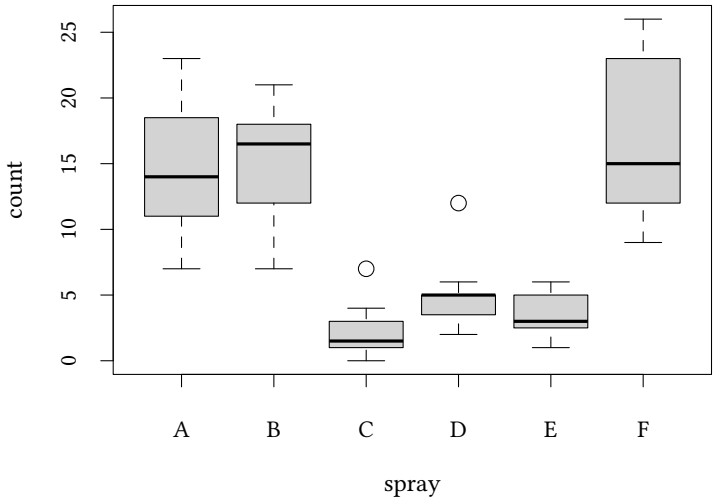
→ Varianzanalyse.

Wir nehmen wieder an, dass unsere Stichprobe aus einer normalverteilten Population gezogen wurde.

Übung 11.12 *Betrachten Sie den Datensatz* `InsectSprays` *und testen Sie, ob die verschiedenen Sprays eine unterschiedliche Auswirkung auf die Überlebensrate der Insekten haben.*

`boxplot` (siehe auch Abschnitt A.6.2) zeichnet einen box-and-whisker plot für Daten die in Gruppen vorliegen.

```
attach(InsectSprays)
boxplot(count ~ spray)
```



Vorüberlegung: Könnte man die 6 Insektensprays nicht einfach mit verschiedenen t-Tests auf unterschiedliche Mittelwerte testen?

- Problem: Welchen Signifikanzwert verlangt man dann? Z.B. 5%. Wenn wir zwei Tests durchführen, dann lehnen wir (fälschlich) etwa in 10% aller Fälle ab, bei drei Tests etwa in 15% aller Fälle...

- Idee: Mache etwas ähnliches wie die t-Statistik für zwei Stichproben:

$$g = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{\hat{\sigma}_A^2}{n_A} + \frac{\hat{\sigma}_B^2}{n_B}}}$$

- Lösung: Varianzanalyse.

$$F = \frac{\text{Variabilität der Gruppenmittelwerte}}{\text{Variabilität innerhalb der Gruppen}}$$

Wir haben r verschiedene Gruppen. Jede Gruppe $k \in \{1, 2, \ldots, r\}$ besteht aus den Beobachtungen X_i mit $i \in I_k$. Dabei gilt $\bigcup_{k=1}^{r} I_k = \{1, \ldots, n\}$

H_0: $\mu_1 = \mu_2 = \ldots = \mu_r$

($\bigcup$ ist das Zeichen für die Vereinigungsmenge, $\bigcup_{k=1}^{r} I_k$ beschreibt die Vereinigung aller Mengen I_k für $k = 1 \ldots r$.) Hier ist ein Beispiel, wie ein Datensatz mit r Gruppen und n Beobachtungen aussehen könnte:

k	Beobachtung	I_k	$\bar{X}_k$
1	$X_1, X_2, X_3, \ldots, X_6$	$\{1, 2, 3, \ldots\}$	$\bar{X}_1$
2	$X_7, X_8, X_9, \ldots, X_{20}$	$\{7, 8, 9, \ldots\}$	$\bar{X}_2$
3	$X_{21}, X_{22}, X_{23}, \ldots$	$\{21, 22, 23, \ldots\}$	$\bar{X}_3$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	$X_{721}, X_{722}, X_{723}, \ldots X_n$	$\{721, 722, 723, \ldots, n\}$	$\bar{X}_r$
			$\bar{\bar{X}}$

Wir berechnen nun Mittelwerte sowohl für jede Gruppe k als auch für alle Beobachtungen:

$$\bar{X}_k = \frac{1}{n_k} \sum_{i \in I_k} X_i \qquad \bar{\bar{X}} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Jetzt vergleichen wir verschiedene Unterschiede:

Unterschiede innerhalb einer Gruppe (eines Treatments) nennen wir SSE (Residuen):

$$\text{SSE} = \sum_{k=1}^{r} \sum_{i \in I_k} (X_i - \bar{X}_k)^2 \quad \text{(sum of squares of errors)}$$

Unterschiede zwischen Gruppen (zwischen Treatments) nennen wir SST:

$$\text{SST} = \sum_{k=1}^{r} n_k (\bar{X}_k - \bar{\bar{X}})^2 \quad \text{(sum of squares of treatments)}$$

Die Abweichungen der einzelnen Beobachtungen X_i zum Gesamtmittel $\bar{\bar{X}}$ nennen wir SSG:

$$\text{SSG} = \sum_{i=1}^{n} (X_i - \bar{\bar{X}})^2 = \text{SSE} + \text{SST} \quad \text{(grand sum of squares)}$$

dann ist (falls alle $X_i \sim N(\mu, \sigma^2)$)

$$\frac{SST}{SSE}\frac{n-r}{r-1} \sim F_{r-1,n-r}$$

Beachten Sie: In der obigen Voraussetzung $X_i \sim N(\mu, \sigma^2)$ stecken drei Dinge

- Unsere Nullhypothese: H_0: $\mu_1 = \mu_2 = \ldots = \mu_r$
- Die Varianz von allen X_i ist gleich, auch wenn sie aus verschiedenen Gruppen stammen.
- Die X_i sind jeweils normalverteilt.

`aggregate` führt eine Funktion (hier z.B. `mean`), getrennt für mehrere Gruppen (hier definiert durch `spray`) aus.

Die Varianzanalyse macht zunächst eine Aussage über die Mittelwerte:

```
aggregate(count ~ spray,FUN=mean)
```

Bei zwei Mittelwerten könnte man einfach die Differenz nehmen und testen, ob diese Differenz groß ist. Hier betrachten wir die quadrierte Summe der Abstände zwischen den einzelnen Mittelwerten und dem Mittelwert aller Beobachtungen.

```
nk<-aggregate(count ~ spray,FUN=length)[,"count"]
Xk<-aggregate(count ~ spray,FUN=mean)[,"count"]
r <- length(Xk)
n<-length(count)
totalMean<-mean(count)
SST <- sum(nk * (Xk - totalMean)^2)
SSG <- sum((count - totalMean)^2)
F <- SST/(SSG-SST) * (n-r)/(r-1)
pf (F,df1=r-1,df2=n-r,lower.tail=FALSE)

[1] 3.182584e-17
```

Man kann das gleiche Ergebnis mit einem einzigen Kommando erzielen:

```
summary(aov(count ~ spray))

            Df Sum Sq Mean Sq F value Pr(>F)
spray        5   2669   533.8    34.7 <2e-16 ***
Residuals   66   1015    15.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Die F-Statistik, die uns die *Varianzanalyse* liefert ist, identisch mit der F-Statistik der *Regression*, die wir weiter unten kennenlernen werden:

```
summary(lm(count ~ spray))
```

```
Call:
lm(formula = count ~ spray)

Residuals:
   Min     1Q Median     3Q    Max
-8.333 -1.958 -0.500  1.667  9.333

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.5000     1.1322  12.807  < 2e-16 ***
sprayB        0.8333     1.6011   0.520    0.604
sprayC      -12.4167     1.6011  -7.755 7.27e-11 ***
sprayD       -9.5833     1.6011  -5.985 9.82e-08 ***
sprayE      -11.0000     1.6011  -6.870 2.75e-09 ***
sprayF        2.1667     1.6011   1.353    0.181
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.922 on 66 degrees of freedom
Multiple R-squared:  0.7244, Adjusted R-squared:  0.7036
F-statistic:  34.7 on 5 and 66 DF,  p-value: < 2.2e-16
```

Zusätzlich gibt uns die Regression noch Informationen über die Wirksamkeit der einzelnen Sprays.

Anhang 11.D Kruskal-Wallis Test für mehr als zwei Stichproben

Oben haben wir mit der *Varianzanalyse* einen Test kennengelernt, der unter der Annahme *normalverteilter* Störgrößen drei oder mehr Stichproben vergleicht. Geht es auch ohne diese Annahme?

Wieder haben wir $r \geq 3$ verschiedene Gruppen. Jede Gruppe $k \in \{1, 2, \dots, r\}$ besteht aus den Beobachtungen X_i mit mit $i \in I_k$.

H_0: Die Verteilung der Beobachtungen ist für alle Gruppen gleich.

Zur Erinnerung, bei der Varianzanalyse:

$$\frac{\text{SST}}{\text{SSE}} \frac{n-r}{r-1} \sim F_{r-1,n-r}$$

Dabei war SST=sum of squares of treatments und SSE=sum of squares of errors

Im Kruskal-Wallis Rangsummen Test berechnen wir zunächst die Ränge R_i der einzelnen Beobachtungen. Dann gilt

$$\left(\frac{12}{n \cdot (n+1)} \cdot \sum_{k=1}^{r} \frac{1}{n_k} \left(\sum_{i \in I_k} R_i \right)^2 \right) - 3(n+1) \sim \chi^2_{r-1}$$

Um diese Formel besser zu verstehen, sehen wir uns zwei Extremfälle an:

Motivation: Wie wäre es, wenn alle Beobachtungen gerade den gleichen »durchschnittlichen Rang« hätten?

$$\left(\frac{12}{n\cdot(n+1)}\cdot\sum_{k=1}^{r}\frac{1}{n_k}\left(\sum_{i\in I_k}\frac{n+1}{2}\right)^2\right)-3(n+1)=$$

$$\left(\frac{12}{n\cdot(n+1)}\cdot\sum_{k=1}^{r}n_k\frac{(n+1)^2}{4}\right)-3(n+1)=$$

$$\left(\frac{12}{n\cdot(n+1)}\cdot\frac{(n+1)^2}{4}\sum_{k=1}^{r}n_k\right)-3(n+1)=$$

$$\left(\frac{3}{(n+1)}\cdot(n+1)^2\right)-3(n+1)=0$$

Motivation 2: Wie wäre es, wenn alle Gruppen gerade eine Beobachtung hätten (also $n_i=1$ und $n=r$) und mit Rängen von $1\dots n$ geordnet wären?

$$\left(\frac{12}{n\cdot(n+1)}\cdot\sum_{k=1}^{n}k^2\right)-3(n+1)=$$

$$\left(\frac{12}{n\cdot(n+1)}\cdot\frac{n\cdot(n+1)(2n+1)}{6}\right)-3(n+1)=$$

$$(2\cdot(2n+1))-3(n+1)=4n+4-3n-3=n-1$$

Zusammenfassung der Motivation:

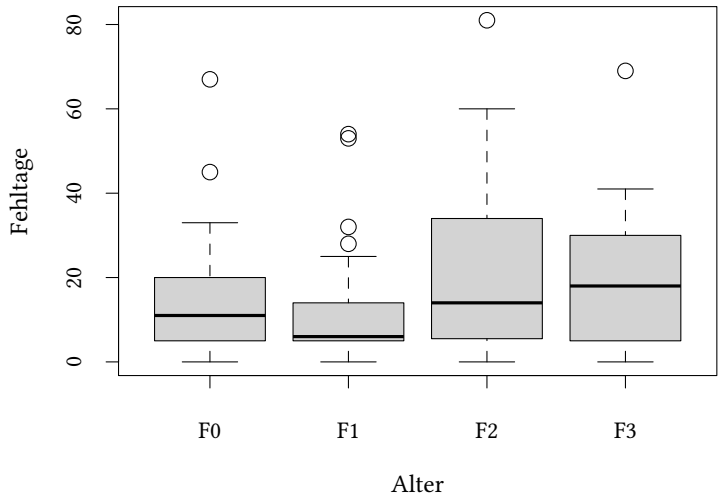
- identische Gruppen: Teststatistik klein ($=0$)
- sehr unterschiedliche Gruppen (perfekt geordnet): Teststatistik groß ($=n-1$)

Unter H_0 (Verteilung in allen Gruppen ist gleich) ist die Teststatistik χ^2_{r-1} verteilt.

Die oben angegebene Formel für die Teststatistik wird auch von R und anderen Statistikprogrammen verwendet. Falls allerdings Beobachtungen gleich sind und damit den gleichen Rang erhalten, führt R noch eine Korrektur ein. Wundern Sie sich also nicht, falls Sie in diesem Fall ein (normalerweise geringfügig) anderes Ergebnis erhalten.

Übung 11.13 *Betrachten Sie den Datensatz `quine` aus dem Paket MASS. Hängt die Anzahl der Fehltage von der Altersgruppe ab? Untersuchen Sie diese Frage getrennt für Jungen und Mädchen.*

```
data(quine,package="MASS")
attach(quine)
boxplot (Days ~ Age,ylab="Fehltage",xlab="Alter")
```

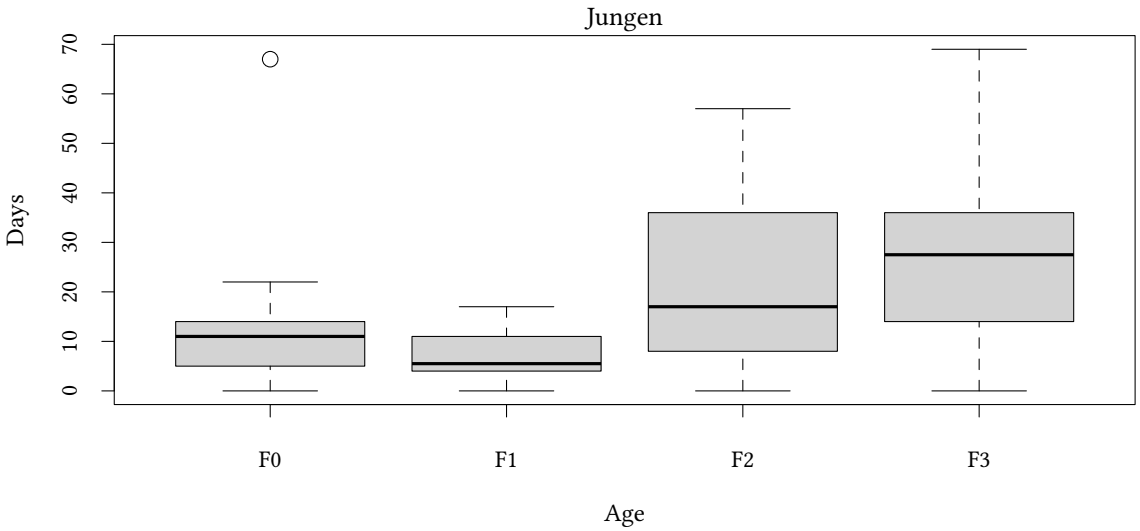



```
n <- length(Days)
r <- length(unique(Age))
nk <- c(table(Age))
Ri <- rank(Days)
g <- 12 / (n * (n+1)) * sum ( (aggregate(Ri ~ Age,FUN=sum)$Ri)^2 / nk) - 3*(n+1)
pchisq(g,df=r-1,lower=FALSE)

[1] 0.0547959
```

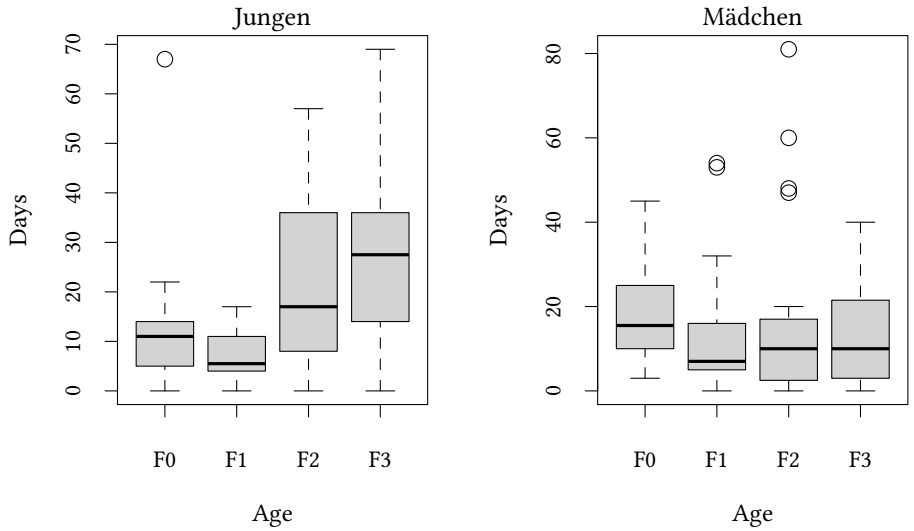
Natürlich gibt es auch hier ein einfaches Kommando (beachten Sie die kleine Abweichung in der Teststatistik; einige Beobachtungen haben den gleichen Rang, deshalb verwendet R noch eine Korrektur):

```
kruskal.test (Days ~ Age)


 Kruskal-Wallis rank sum test

data:  Days by Age
Kruskal-Wallis chi-squared = 7.6347, df = 3, p-value = 0.0542
```

```
boxplot (Days ~ Age, subset = Sex=="M",main="Jungen")
```




```
boxplot (Days ~ Age, subset = Sex=="F",main="Mädchen")
```



Die Grafik zeigt eine klare Altersabhängigkeit bei Jungen, allerdings nicht bei Mädchen. Der Kruskal-Test bestätigt diese Beobachtung.

```
kruskal.test (Days ~ Age, subset = Sex=="M")


 Kruskal-Wallis rank sum test

data:  Days by Age
Kruskal-Wallis chi-squared = 17.823, df = 3, p-value = 0.0004784

kruskal.test (Days ~ Age, subset = Sex=="F")


 Kruskal-Wallis rank sum test

data:  Days by Age
Kruskal-Wallis chi-squared = 1.9249, df = 3, p-value = 0.5881
```

Die Ergebnisse der Varianzanalyse gehen in eine ähnliche Richtung:

```
summary(aov (Days ~ Age, subset = Sex=="M"))

            Df Sum Sq Mean Sq F value Pr(>F)
Age          3   3999  1333.1    5.91 0.0013 **
Residuals   62  13986   225.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(aov (Days ~ Age, subset = Sex=="F"))

            Df Sum Sq Mean Sq F value Pr(>F)
Age          3    506   168.8   0.656  0.582
Residuals   76  19544   257.1
```

Anhang 11.E Übungen für Varianzanalyse und Kruskal-Wallis Test

Übung 11.14 *Betrachten Sie wieder den Datensatz* `InsectSprays` *und testen Sie, ob die verschiedenen Sprays eine unterschiedliche Auswirkung auf die Überlebensrate der Insekten haben. Vergleichen Sie das Ergebnis eines parametrischen mit dem eines nichtparametrischen Tests.*

Übung 11.15 *Eine Universität vergleicht 18 Studenten von insgesamt drei Fächern. Diese Studenten haben den folgenden Erfolg im Test:*

Fach	*1*	*2*	*3*	*4*	*5*	*6*	*Mittelwert*
Politologie	*45*	*76*	*50*	*51*	*40*	*57*	*53.17*
Soziologie	*59*	*56*	*61*	*66*	*54*	*49*	*57.5*
Germanistik	*52*	*60*	*58*	*53*	*72*	*63*	*59.67*

Gehen Sie davon aus, dass

```
xA <- c(45,76,50,51,40,57)
xB <- c(59,56,61,66,54,49)
xC <- c(52,60,58,53,72,63)
treat <- c("A","A","A","A","A","A",
           "B","B","B","B","B","B",
           "C","C","C","C","C","C")
```

- *Wie testen Sie, ob das Studienfach einen Einfluss auf den Testerfolg hat, wenn Sie davon ausgehen, dass die Punkte normalverteilt sind?*
- *Wie testen Sie, wenn Sie diese Annahme nicht machen wollen?*

Übung 11.16 *Wir wollen die durchschnittlichen Semesterwochenstunden der Studendierenden unterschiedlicher Fächer vergleichen. Wir nehmen an, dass die Semesterwochenstunden jeweils normalverteilt sind. Unsere Stichprobe besteht aus 5 Juristen und 5 Physikern. Die Anzahl der Semesterwochenstunden ergibt sich aus folgender Tabelle:*

Fach	*1*	*2*	*3*	*4*	*5*
Jura	*20*	*18*	*22*	*8*	*14*
Physik	*16*	*14*	*24*	*20*	*18*

1. *Wie testen wir, ob die mittlere Anzahl der Semesterwochenstunden für die beiden Fächer gleich ist?*
2. *Unsere Nullhypothese ist, dass die mittlere Anzahl der Semesterwochenstunden für beide Fächer gleich ist. Das Signifikanzniveau ist 5%. Können wir die Nullhypothese ablehnen?*
3. *Nun werden die Semesterwochenstunden von insgesamt 20 Studenten in vier Fächern erhoben. Diese Studenten haben folgende Semesterwochenstunden:*

Fach	*1*	*2*	*3*	*4*	*5*
Jura	*20*	*18*	*22*	*8*	*14*
Physik	*16*	*14*	*24*	*20*	*18*
Medizin	*30*	*20*	*26*	*25*	*22*
Wirtschaftswissenschaften	*26*	*12*	*20*	*18*	*16*

4. *Wie testen wir, ob das Studienfach einen Einfluss auf die Anzahl der Semesterwochenstunden hat, wenn wir davon ausgehen, dass die Anzahl der Semesterwochenstunden normalverteilt sind (mehrere richtige Antworten möglich):*

5. *Die Nullhypothese ist, dass die mittlere Anzahl der Semesterwochenstunden für alle Fächer gleich ist. Das Signifikanzniveau ist 5%. Können wir die Nullhypothese ablehnen?*

6. *Bei welchem α kann die Nullhypothese aus Aufgabe 3 abgelehnt werden?*

Übung 11.17 *Auf dem Weihnachtsmarkt in Jena gibt es 4 Glühweinstände. Die 4 Standbesitzer wurden befragt, wieviel Gläser Glühwein sie jeweils an 7 zufällig ausgewählten Intervallen à 60 Minuten verkauft haben. Diese Intervalle sind für jeden Stand unabhängig voneinander. Dabei gab es das folgende Ergebnis:*

Stand	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$	$t = 7$
1	*49*	*46*	*59*	*51*	*73*	*109*	*141*
2	*55*	*47*	*63*	*41*	*67*	*112*	*130*
3	*37*	*31*	*70*	*48*	*77*	*99*	*156*
4	*42*	*56*	*69*	*57*	*80*	*134*	*98*

Sie sollen nun einen Kruskal-Wallis Rangsummen Test durchführen, um herauszufinden, ob die Anzahl der verkauften Glühweine an den Ständen sich unterscheidet.

Übung 11.18 *Ein neu gegründeter Versandhandel überlegt, mit welcher Spedition er zusammenarbeiten möchte. In Betracht kommen 4 Speditionen der näheren Umgebung. Es wurden Stichproben für die Lieferdauer eines Paketes mit Standardversand für jede Spedition erhoben:*

Spedition	*Lieferdauer*							$\bar{X}_k$
A	*4,5*	*3*	*3,5*	*2*				*3,25*
B	*2*	*4*	*1,5*	*2*	*1*	*1,5*		*2*
C	*1*	*1,5*	*3*	*2,5*	*1,5*	*2,5*	*2*	*2*
D	*1,5*	*2*	*4,5*	*2*				*2,5*

Es wird die Annahme getroffen, dass die zu versendende Ware immer im Lager des Versandhandels zum Zeitpunkt der Bestellung verdsandfähig verpackt und vorrätig ist. Es wurde die Zeit (in Tagen) ab dem Zeitpunkt der Bestellung bis zur Annahme des Paketes durch den Kunden gemessen. Untersuchen Sie zum Signifikanzniveau $\alpha = 5\%$ untersucht werden, ob sich die durchschnittlichen Lieferzeiten der Speditionen unterscheiden.

Übung 11.19 *In der Vorlesung haben wir das Verfahren der Varianzanalyse (AOV) behandelt. Welche der folgenden Aussagen trifft für dieses Verfahren zu?*

- *Es vergleicht die Mediane mehrerer Stichproben*
- *Es ist ein nichtparametrisches Verfahren.*
- *Es lässt sich nur anwenden, wenn die abhängige Variable* F*-verteilt ist*
- *Es betrachtet den Anteil der erklärten Varianz an der nicht erklärten Varianz*
- *Es lässt sich nur anwenden, wenn die erklärende Variable diskret ist*

Anhang 11.F Friedman Test für verbundene Stichproben

Der Kruskal-Wallis Test ist ein Test für *unverbundene* Stichproben. Natürlich gibt es auch für *verbundene* Stichproben einen Test, der r Gruppen miteinander vergleicht. Dieser Test ist der Friedman Test.

- Beim Kruskal-Wallis Test nehmen wir an: Die einzelnen n Beobachtungen X_i sind unabhängig voneinander.
- Beim Friedman Test nehmen wir an: Es gibt t Versuchseinheiten. Jede wird in jedem der r Treatments getestet. Jetzt haben wir $t \times r$ Beobachtungen. Allerdings sind die jeweils r Beobachtungen der gleichen Versuchseinheit abhängig voneinander.

Übung 11.20 *Der Datensatz* `warpbreaks` *beschreibt, wie häufig in einem Webstuhl Fäden aus unterschiedlichem Material und unter unterschiedlicher Belastung reißen.*

Wir berechnen zunächst für jedes Material und für jede Belastung die durchschnittliche Häufgkeit:

```
wb <- with(warpbreaks,aggregate(breaks ~ wool + tension,FUN=mean))
```

Wenn `wool` unsere Versuchseinheit beschreibt, und `tension` das Treatment, rufen wir den Test wie folgt auf:

```
friedman.test(breaks ~ tension | wool,data=wb)


 Friedman rank sum test

data:  breaks and tension and wool
Friedman chi-squared = 1, df = 2, p-value = 0.6065
```

Wir sehen, dass eine unterschiedliche Fadenspannung nicht zu signifikant unterschiedlich häufigem Reissen des Fadens führt.

Übung 11.21 *Betrachten Sie nochmals den Datensatz* `InsectSprays`*. Gehen Sie nun davon aus, dass die 12 Beobachtungen, die Ihnen für jedes Spray vorliegen, von der gleichen Beobachtungseinheit stammen (etwa aus dem gleichen Labor). Führen Sie einen Friedman Test durch.*

Anhang 11.G Jonckheere-Terpstra

Beim Kruskal-Wallis Test testen wir allgemein, ob die Gruppenvariable einen Einfluss auf die Mediane der Stichproben hat. Manchmal vermuten wir, dass die Gruppenvariable *geordnet* ist, und für einen Trend zwischen den Gruppen verantwortlich ist. Testen können wir das mit dem Jonckheere-Terpstra Test.

Wie bei der Varianzanalyse haben wir r verschiedene Gruppen. Jede Gruppe $k \in \{1, 2, \dots, r\}$ besteht aus n_k vielen Beobachtungen X_i mit $i \in I_k$. Dabei gilt $\bigcup_{k=1}^{r} I_k = \{1, \dots, n\}$.

H_0: $\mu_1 = \mu_2 = \dots = \mu_r$

H_1: $\mu_0 \geq \mu_1 \geq \dots \geq \mu_r$ mit mindestens einer strikten Ungleichheit.

Zur Erinnerung, bei der Varianzanalyse:

$$\frac{\text{SST}}{\text{SSE}} \frac{n-r}{r-1} \sim F_{r-1,n-r}$$

Zur Erinnerung (2): Beim Kruskal-Wallis Rangsummen Test berechnen wir die Ränge R_i der einzelnen Beobachtungen. Dann gilt

$$\left(\frac{12}{n \cdot (n+1)} \cdot \sum_{k=1}^{r} \frac{1}{n_k} \left(\sum_{i \in I_k} R_i \right)^2 \right) - 3(n+1) \sim \chi^2_{r-1}$$

Beim Jonckheere-Terpstra Test berechnen wir für jedes Paar von Gruppen k, k' mit $k < k'$

$$\begin{aligned} M_{k,k'} = &\text{Anzahl}(x_{k,i} < x_{k',i'} \text{ mit } i \in I_k \text{ und } i' \in I_{k'}) + \\ &+ \frac{1}{2} \text{Anzahl}(x_{k,i} = x_{k',i'} \text{ mit } i \in I_k \text{ und } i' \in I_{k'}) \end{aligned}$$

Dann ist die Jonckheere-Terpstra Statistik

$$J = \sum_{(k,k') \text{ mit } k<k'} M_{k,k'}$$

Für große Stichproben gilt $J \sim N\left(E_0(J), \text{var}_0(J)\right)$

$$E_0(J) = \frac{1}{4} \left(n^2 - \sum_{k=1}^{r} n_k^2 \right)$$

$$\text{var}_0(J) = \frac{1}{72} \left(n^2(2n+3) - \sum_{k=1}^{r} n_k^2(2n_k+3) \right)$$

Genau rechnet es in jedem Fall der `jonckheere.test` aus.

```
library(clinfun)
data(Bwages,package="Ecdat")
attach(Bwages)
boxplot(wage ~ educ)
jonckheere.test(wage,educ)
```

Nehmen wir einmal an, unser Sample wäre nicht ganz so groß.

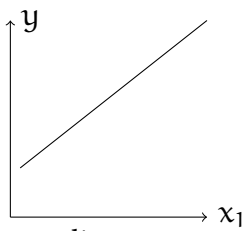
```
detach(Bwages)
(n=dim(Bwages)[1])
set.seed(130)
(mySample = sample(n,40))
with(Bwages[mySample,],boxplot(wage ~ educ))
with(Bwages[mySample,],jonckheere.test(wage,educ))
summary(with(Bwages[mySample,],lm(wage ~ educ)))
```

12. Nichtlineare Regressionsfunktionen

12.1. Motivation

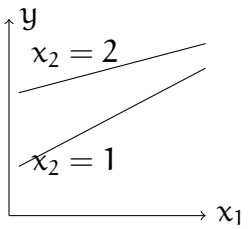
Bislang:

$y = \beta_0 + \beta_1 x_1$



linear

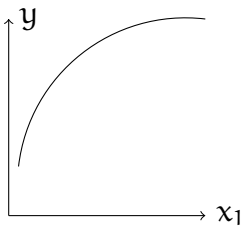
$y = \beta_0 + \beta_1 x_1 \cdot x_2$



Interaktion von x_1 und x_2

In diesem Kapitel:

$y = f(x)$



nicht linear
keine Interaktion

Wenn die Beziehung zwischen Y und X nicht linear ist...

- ist der marginale Effekt von X nicht konstant
- wäre eine einfache lineare Regression falsch spezifiziert

→ der geschätzte Effekt wäre verzerrt

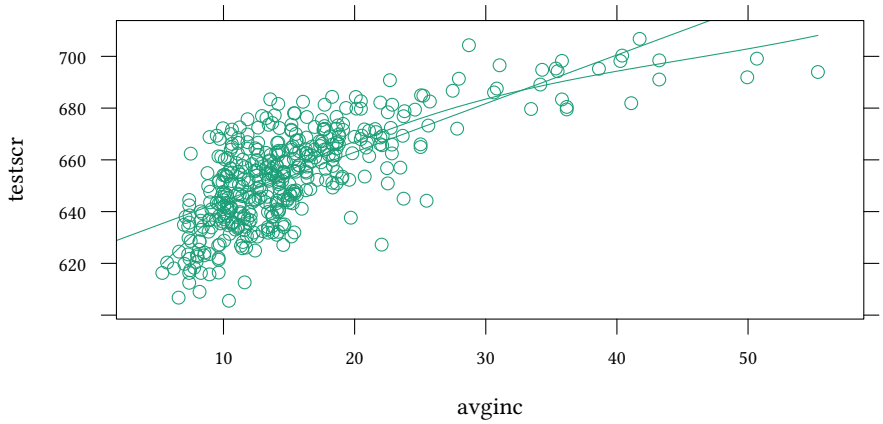
→ deshalb schätzen wir eine nichtlineare Regression in X

Vorgehensweise:

- nichtlinearen Funktionen einer einzelnen unabhängigen Variablen
  - Polynome in X
  - Logarithmische Transformationen

$$\texttt{testscr} = \beta_0 + \beta_1 \texttt{avginc} + u$$

```
data(Caschool,package="Ecdat")
attach(Caschool)
xyplot(testscr ~ avginc,t=c("p","r","smooth"))
```



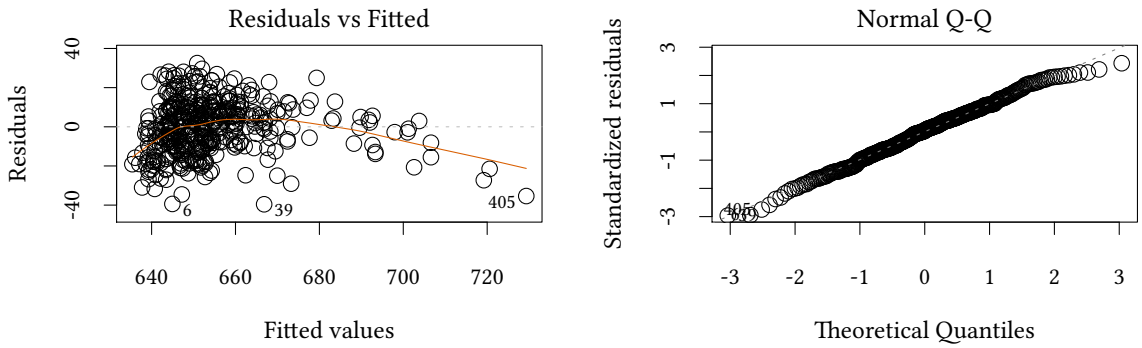
Der lineare Zusammenhang ist nicht befriedigend.

```
est1 <- lm(testscr ~ avginc)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	625.3836	1.5324	408.11	0.0000
avginc	1.8785	0.0905	20.76	0.0000

Der diagnostische Plot bestätigt, dass in diesem linearen Modell die Residuen (u) nicht unabhängig von `avginc` sind: $E(u|X) \neq 0$.

```
plot(est1,which=1:2)
```



Betrachte nun ein quadratisches Modell

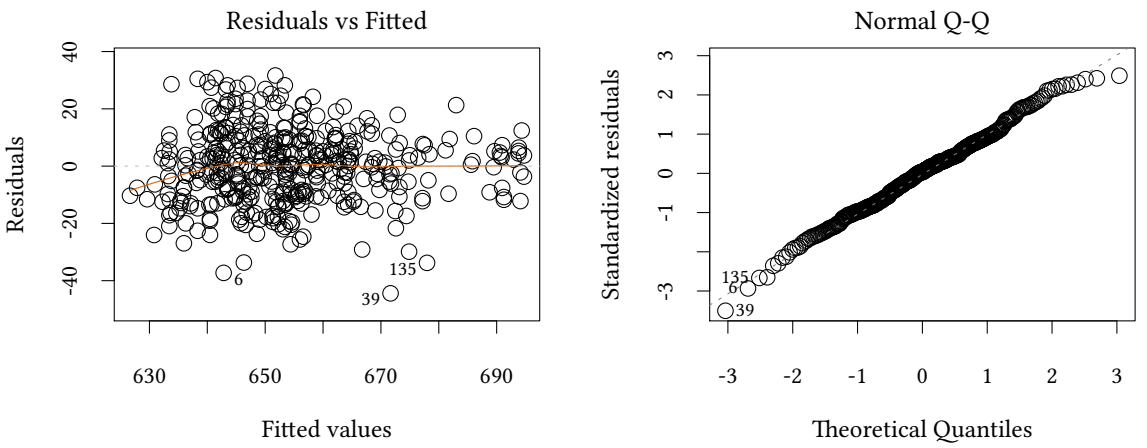
$$\texttt{testscr} = \beta_0 + \beta_1 \texttt{avginc} + \beta_2 \texttt{avginc}^2 + u$$

```
avginc2 <- avginc*avginc
est2 <- lm(testscr ~ avginc + avginc2)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	607.3017	3.0462	199.36	0.0000
avginc	3.8510	0.3043	12.66	0.0000
avginc2	-0.0423	0.0063	-6.76	0.0000

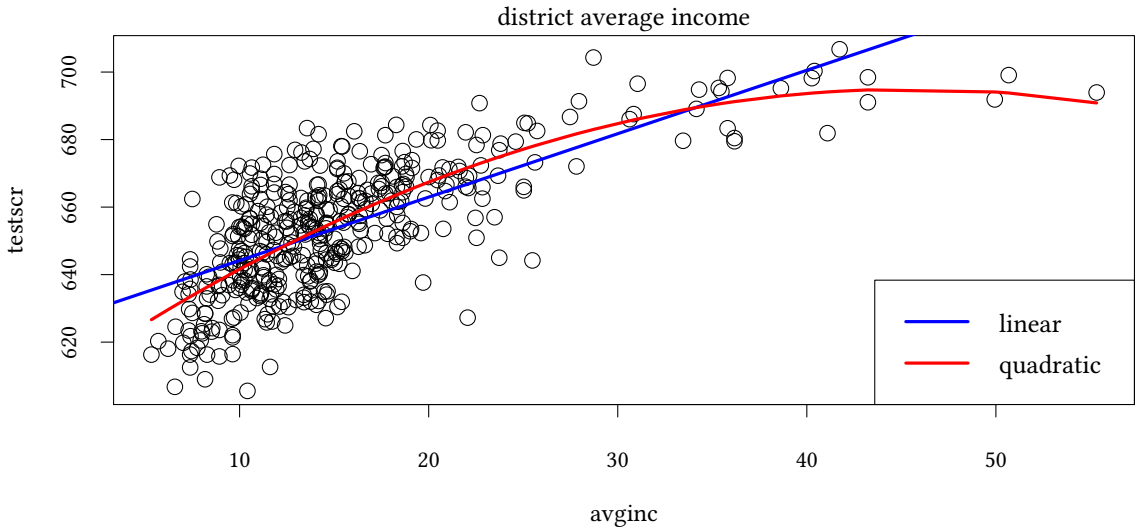
(Die Notation `lm(testscr ~ avginc + avginc*avginc)` funktioniert nicht, weil R `avginc*avginc` in der Modellformel als Interaktion interpretiert). Betrachten wir auch hier wieder den diagnostischen Plot:

```
plot(est2,which=1:2)
```



`order` berechnet einen Vektor von Indizes den man zum Sortieren (dieses Vektors aber auch anderer Vektoren) verwenden kann. `fitted` berechnet zu einer Regression das $\hat{y}$. Wir sortieren hier die Punkte, weil wir eine Kurve zeichnen wollen, und es sehr viel besser aussieht, wenn wir die Punkte einer Kurve in der richtigen Reihenfolge zeichnen.

```
or <- order(avginc)
plot(testscr ~ avginc,main="district average income")
abline(est1,col="blue",lwd=3)
lines(avginc[or],fitted(est2)[or],col="red",lwd=3)
legend("bottomright",c("linear","quadratic"),lwd=3,col=c("blue","red"))
```



Hier vergleichen wir die beiden Modelle, das lineare und das quadratische, nebeneinander:

```
library(texreg)
```

```
texreg(list(est1=est1,est2=est2))
```

Vergleich - das lineare und das quadratische Modell

	est1	est2
(Intercept)	625.38***	607.30***
	(1.53)	(3.05)
avginc	1.88***	3.85***
	(0.09)	(0.30)
avginc2		−0.04***
		(0.01)
R^2	0.51	0.56
Adj. R^2	0.51	0.55
Num. obs.	420	420

***p < 0.001; **p < 0.01; *p < 0.05

- Der Koeffizient von `avginc2` ist signifikant von Null verschieden
- R^2 ist gestiegen

Marginale Effekte im quadratischen Modell

$$\begin{aligned}\texttt{testscr} &= 607.30174 + 3.85100 \cdot \texttt{avginc} - \\ &\quad 0.04231 \cdot \texttt{avginc}^2\end{aligned}$$

Marginaler Effekt einer Änderung von `avginc`?

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u$$

$$\frac{\partial Y_i}{\partial X_i} = \beta_1 + 2\beta_2 X_i$$

Hier berechnen wir den marginalen Effekt für ein Einkommen von 10, 40, und 60:

```
X<-c(10,40,60)
meQuad <- coef(est2)["avginc"] +
          2*X*coef(est2)["avginc2"]
```

	10	40	60
meQuad	3.00	0.47	-1.23

Nichtlineare Modelle Vorgehensweise:

1. theoretische Motivation für nichtlineare Beziehung
2. spezifiziere funktionale Form
3. testen ob nichtlineare Funktion gerechtfertigt ist
   - Diagnostischer Plot der Residuen
   - R^2, AIC, p-Wert der nichtlinearen Koffizienten
   - visueller Test
4. marginale Effekte
   - Bei einfachen linearen Modellen einfach β_i
   - Bei nichtlinearen Modellen etwas komplizierter (ableiten nach X_i)

     β_i hat zuweilen eine andere Interpretation.

12.2. Funktionale Formen

12.2.1. Polynome

Funktionale Formen — Polynome

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \ldots + \beta_r X_i^r + u_i$$

($r = 2$: quadratisches Modell, $r = 3$: kubisches Modell, ...)

Warum bleibt man nicht beim linearen Modell? Kann man testen, ob das Polynom »gerechtfertigt« ist?

- Test der Nullhypothese, die Regressionsfunktion sei linear:

$$H_0 : \beta_2 = 0 \wedge \beta_3 = 0 \wedge \ldots \wedge \beta_r = 0$$

versus

$$H_1 : \text{wenigstens ein } \beta_j \neq 0, j \in \{2, \ldots, r\}$$

- was ist das richtige r?
  - großes r: mehr Flexibilität, besserer Fit
  - kleines r: präzisere Schätzung der einzelnen Koeffizienten

Wahl des »richtigen« Polynoms vom Grad r **Wahl des »richtigen« Polynoms** sequentieller Hypothesentest bei polynomialen Modellen:

1. Wähle den größten sinnvollen Wert von r und schätze eine polynomiale Regression
2. teste $H_0 : \beta_r = 0$. Wenn H_0 abgelehnt wird, dann verwende ein Polynom r-ten Grades
3. Sonst: reduziere r um 1. Weiter bei Schritt 1.

(Problem: eigentlich müsste man für multiples Testen korrigieren. Das Problem ignorieren wir hier.)

Ein Beispiel Wir beginnen mit einem Polynom dritten Grades: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + u$.

```
avginc3 <- avginc*avginc*avginc
est3 <- lm(testscr ~ avginc + avginc2 + avginc3)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	600.0790	5.8296	102.94	0.0000
avginc	5.0187	0.8595	5.84	0.0000
avginc2	-0.0958	0.0374	-2.56	0.0107
avginc3	0.0007	0.0005	1.45	0.1471

Wir sehen, der Koeffizient von `avginc3` ist nicht signifikant. Also fahren wir mit einem kleiner Modell fort:

```
est2 <- lm(testscr ~ avginc + avginc2)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	607.3017	3.0462	199.36	0.0000
avginc	3.8510	0.3043	12.66	0.0000
avginc2	-0.0423	0.0063	-6.76	0.0000

Hier sind alle Koeffizienten signifikant. Also verkleinern wir das Modell nicht weiter.

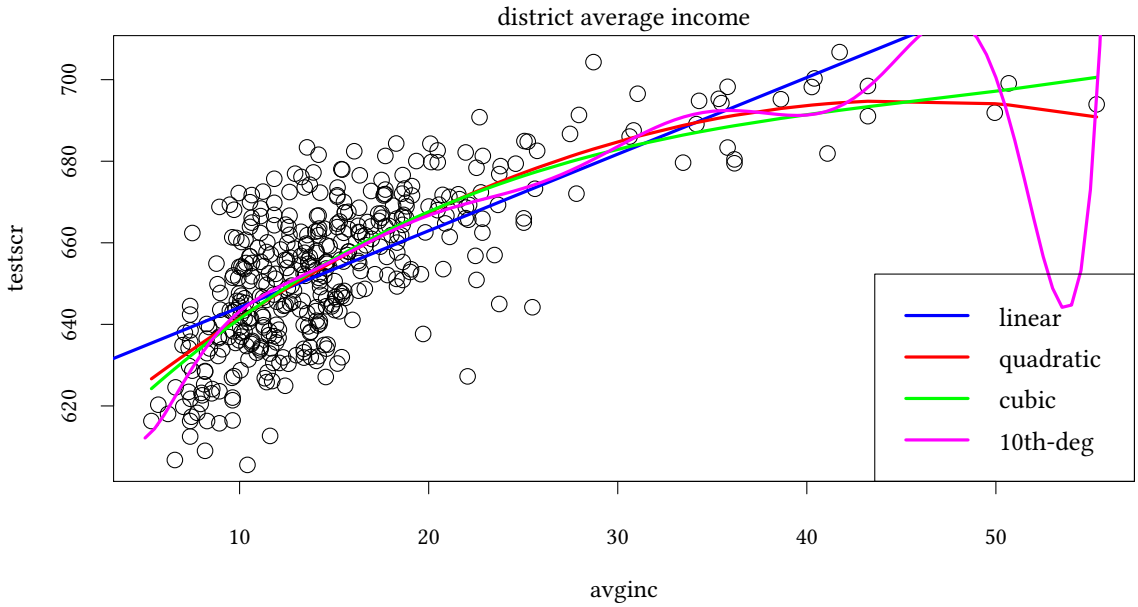
Falls man sich für große `r` interessiert, muss man nicht alle Potenzen von `avginc` von Hand ausrechnen. Hier z.B. betrachten wir ein Polynom zehnten Grades:

```
xtable(estp <- lm(testscr ~ poly(avginc,10)))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	654.1565	0.6181	1058.40	0.0000
poly(avginc, 10)1	277.8568	12.6665	21.94	0.0000
poly(avginc, 10)2	-85.9935	12.6665	-6.79	0.0000
poly(avginc, 10)3	18.4560	12.6665	1.46	0.1459
poly(avginc, 10)4	-28.0133	12.6665	-2.21	0.0275
poly(avginc, 10)5	19.6861	12.6665	1.55	0.1209
poly(avginc, 10)6	-9.0544	12.6665	-0.71	0.4751
poly(avginc, 10)7	3.8626	12.6665	0.30	0.7606
poly(avginc, 10)8	-5.5000	12.6665	-0.43	0.6644
poly(avginc, 10)9	1.7121	12.6665	0.14	0.8925
poly(avginc, 10)10	15.7365	12.6665	1.24	0.2148

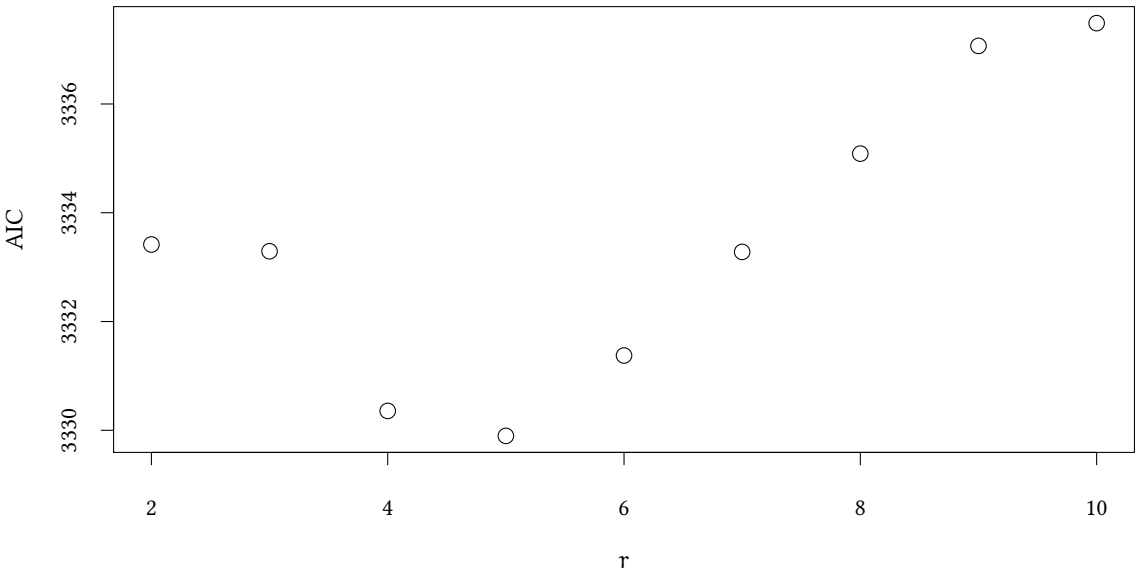
Für unsere Anwendung ist ein solches Polynom nicht besonders sinnvoll. Das sieht man auch in der Graphik in der die verschiedenen Polynome in unserem Datensatz dargestellt werden:

```
estp <- lm(testscr ~ poly(avginc,10))
plot(testscr ~ avginc,main="district average income")
abline(est1,col="blue",lwd=3)
lines(avginc[or],fitted(est2)[or],col="red",lwd=3)
lines(avginc[or],fitted(est3)[or],col="green",lwd=3)
smooth<-list(avginc=seq(5,70,.5))
lines(smooth$avginc,predict(estp,newdata=smooth),col="magenta",lwd=3)
legend("bottomright",c("linear","quadratic","cubic","10th-deg"),lwd=3,
       col=c("blue","red","green","magenta"))
```



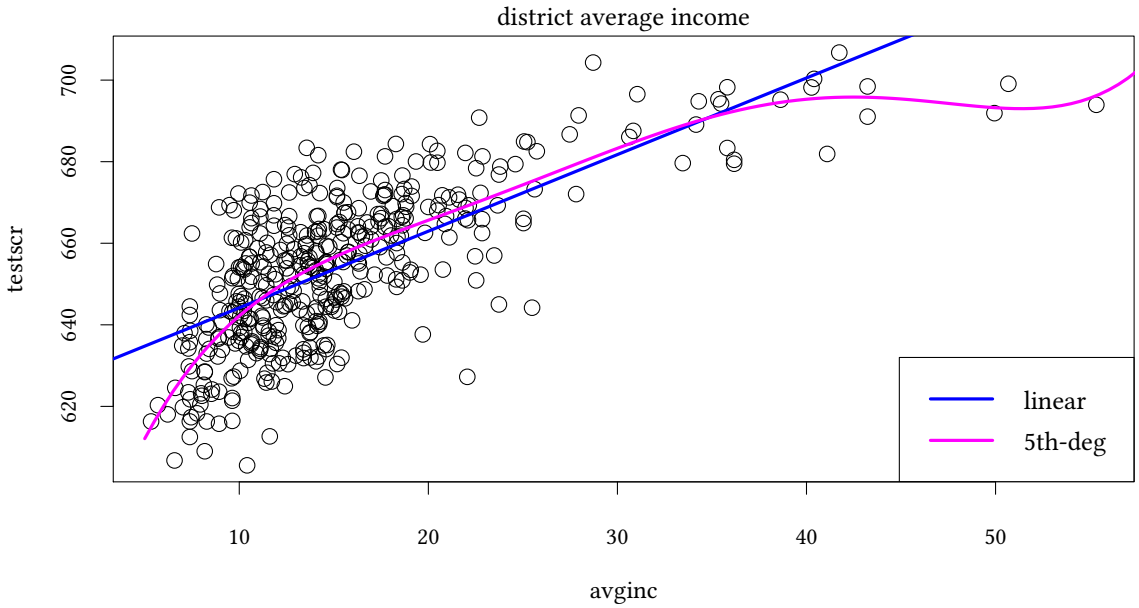
AIC und Größe des Polynoms Mit dem AIC kommen wir zu einem anderen Ergebnis als mit Hilfe von p-Werten. Die folgende Grafik zeigt, wie das AIC von der Größe des Polynoms abhängt.

```
plot(t(sapply(2:10,function(n) c(n,AIC(lm(testscr ~ poly(avginc,n))))),xlab="$r$",ylab="AIC")
```



Auf Basis des AIC (und ohne jede ökonomische Intuition) würde man das Polynom 5. Grades auswählen:

```
estp5 <- lm(testscr ~ poly(avginc,5))
plot(testscr ~ avginc,main="district average income")
abline(est1,col="blue",lwd=3)
smooth<-list(avginc=seq(5,70,.5))
lines(smooth$avginc,predict(estp5,newdata=smooth),col="magenta",lwd=3)
legend("bottomright",c("linear","5th-deg"),lwd=3,
       col=c("blue","magenta"))
```

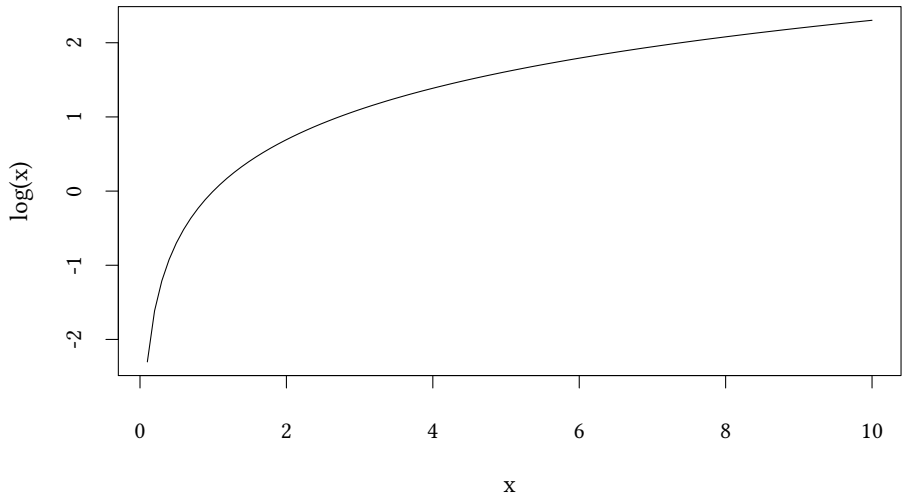



```
lm(testscr ~ poly(avginc,5,raw=TRUE))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	548.1432	22.1189	24.78	0.0000
poly(avginc, 5, raw = TRUE)1	17.8244	5.6132	3.18	0.0016
poly(avginc, 5, raw = TRUE)2	-1.2053	0.5189	-2.32	0.0207
poly(avginc, 5, raw = TRUE)3	0.0432	0.0218	1.98	0.0484
poly(avginc, 5, raw = TRUE)4	-0.0007	0.0004	-1.75	0.0813
poly(avginc, 5, raw = TRUE)5	0.0000	0.0000	1.56	0.1197

12.2.2. Logarithmische Modelle

Oben haben wir gesehen, dass der Zusammenhang zwischen `avginc` und `testscr` eher konkav ist. Eine Funktion, die so ähnlich aussieht, ist die Logarithmusfunktion.



Wir unterscheiden drei Typen logarithmischer Modelle:

- $Y_i = \beta_0 + \beta_1 \cdot \log X_i + u_i$ linear-log
- $\log Y_i = \beta_0 + \beta_1 \cdot X_i + u_i$ log-linear
- $\log Y_i = \beta_0 + \beta_1 \cdot \log X_i + u_i$ log-log

12.2.3. Logarithmische Modelle - linear-log

$$Y_i = \beta_0 + \beta_1 \cdot \log X_i + u_i$$

marginale Effekte:

$$\frac{\partial Y_i}{\partial X_i} = \beta_1 \frac{1}{X_i}$$

$$\frac{\Delta Y_i}{\Delta X_i} \approx \beta_1 \frac{1}{X_i} \quad \rightarrow \quad \Delta Y_i \approx \beta_1 \frac{\Delta X_i}{X_i}$$

wenn sich X_i um 1% ändert ($\Delta X_i = 0.01 \cdot X_i$) ...

$$\Delta Y_i \approx \beta_1 \frac{0.01 X_i}{X_i} = 0.01 \beta_1$$

...dann ändert sich Y_i um $0.01 \cdot \beta_1$

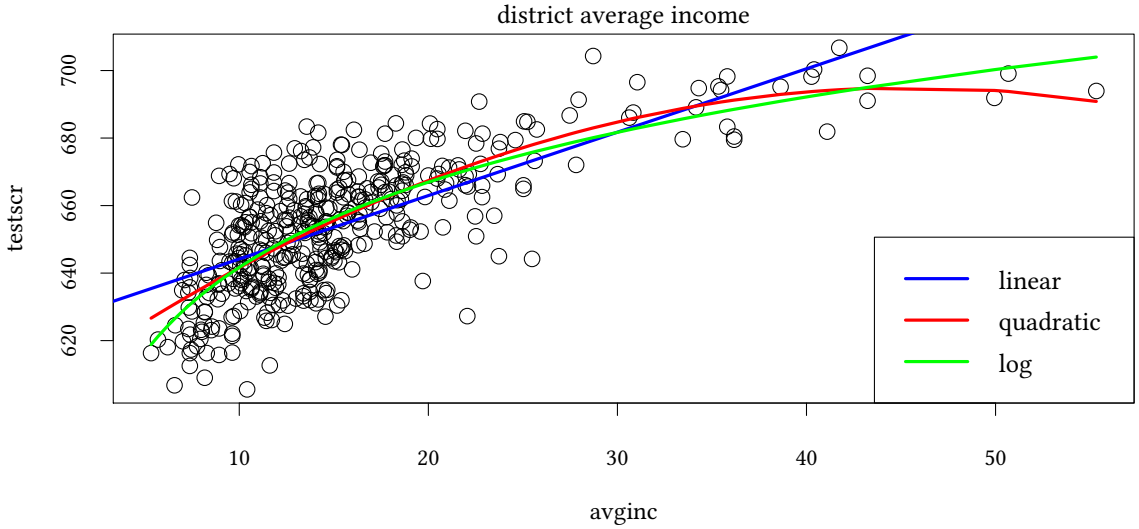
$$\texttt{testscr} = \beta_0 + \beta_1 \log \texttt{avginc} + u$$

```
estL <- lm(testscr ~ log(avginc))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	557.8323	4.2003	132.81	0.0000
log(avginc)	36.4197	1.5710	23.18	0.0000

Wenn sich `avginc` um 1% ändert, dann ändert sich `testscr` um 0.364 .

```
plot(testscr ~ avginc,main="district average income")
abline(est1,col="blue",lwd=3)
lines(avginc[or],fitted(est2)[or],col="red",lwd=3)
lines(avginc[or],fitted(estL)[or],col="green",lwd=3)
legend("bottomright",c("linear","quadratic","log"),lwd=3,
       col=c("blue","red","green"))
```



Marginale Effekte im linear-log Modell allgemein lautet unser Modell:

$$Y_i = \beta_0 + \beta_1 \cdot \log X_i + u_i$$

in diesem Modell haben wir geschätzt:

$$Y_i = 558 + 36.4 \cdot \log X_i + u_i$$

marginaler Effekt:

$$\frac{\Delta Y_i}{\Delta X_i} \approx \beta_1 \frac{1}{X_i} \qquad \Delta Y_i \approx \beta_1 \frac{\Delta X_i}{X_i}$$

Wir berechnen wieder marginale Effekte für verschiedene Werte von `avginc`:

```
X<-c(10,40,60)
meLinLog<-coef(estL)[2]/X
```

	10	40	60
meQuad	3.00	0.47	-1.23
meLinLog	3.64	0.91	0.61

12.2.4. Logarithmische Modelle - log-linear

$$\log Y_i = \beta_0 + \beta_1 \cdot X_i + u_i \quad \rightarrow \quad Y_i = e^{\beta_0+\beta_1 \cdot X_i + u_i}$$

marginale Effekte: $\downarrow$

$$\frac{\partial \log Y_i}{\partial X_i} = \beta_1 \qquad \frac{\partial Y_i}{\partial X_i} = \beta_1 \cdot e^{\beta_0+\beta_1 \cdot X_i}$$

$$\frac{\Delta \log Y_i}{\Delta X_i} \approx \beta_1 \quad \rightarrow \quad \Delta \log Y_i \approx \beta_1 \Delta X_i$$

$$\frac{\Delta \log Y_i}{\Delta Y_i} \approx \frac{1}{Y_i} \quad \rightarrow \quad \Delta \log Y_i \approx \frac{\Delta Y_i}{Y_i} \quad \rightarrow \quad \Delta \log Y_i \approx \frac{\Delta Y_i}{Y_i} \approx \beta_1 \cdot \Delta X_i$$

Eine Änderung von X_i um eine Einheit bedeutet eine Änderung von Y_i um den relativen Anteil β_1 (oder um $100 \cdot \beta_1$ %).

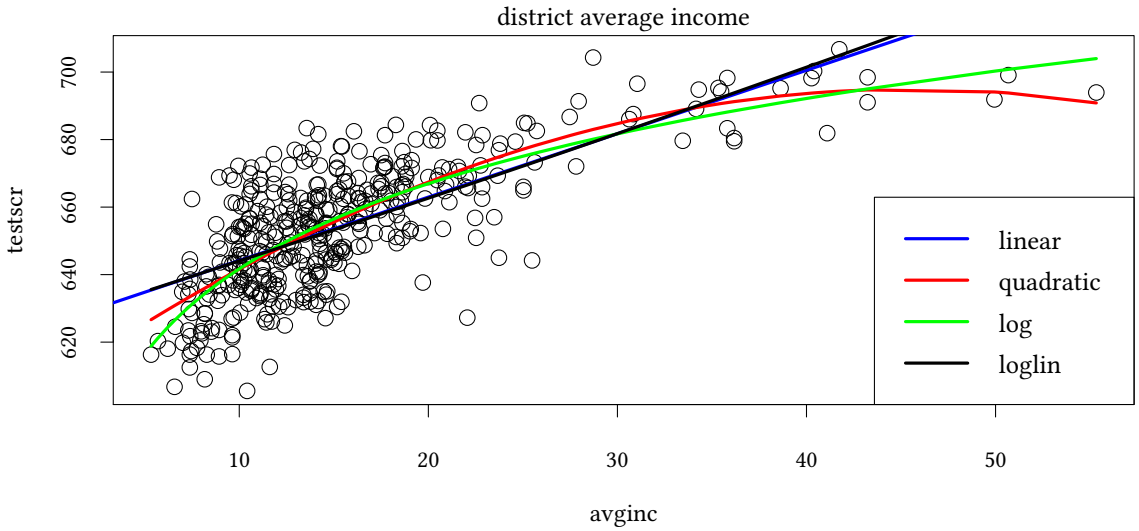
$$\log \texttt{testscr} = \beta_0 + \beta_1 \texttt{avginc} + u$$

```
estLL <- lm(log(testscr) ~ avginc)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.4394	0.0024	2724.16	0.0000
avginc	0.0028	0.0001	20.37	0.0000

Eine Änderung von `avginc` um eine Einheit bedeutet eine Änderung von `testscr` um 0.28%.

```
plot(testscr ~ avginc,main="district average income")
abline(est1,col="blue",lwd=3)
lines(avginc[or],fitted(est2)[or],col="red",lwd=3)
lines(avginc[or],fitted(estL)[or],col="green",lwd=3)
lines(avginc[or],exp(fitted(estLL))[or],col="black",lwd=3)
legend("bottomright",c("linear","quadratic","log","loglin"),lwd=3,
       col=c("blue","red","green","black"))
```



Um die marginalen Effekte einfacher auszurechnen, verwenden wir hier die Funktion `predict`.

$$Y_i = e^{\beta_0+\beta_1\cdot X_i+u_i} \quad \rightarrow \quad \frac{\partial Y_i}{\partial X_i} = \beta_1 \cdot e^{\beta_0+\beta_1\cdot X_i}$$

```
X<-c(10,40,60)
meLogLin<-coef(estLL)[2]*exp(predict(estLL,newdata=list(avginc=X)))
```

	10	40	60
meQuad	3.00	0.47	-1.23
meLinLog	3.64	0.91	0.61
meLogLin	1.83	1.99	2.11

12.2.5. Logarithmische Modelle - log-log

$$\log Y_i = \beta_0 + \beta_1 \cdot \log X_i + u_i$$

alternativ:

$$Y_i = e^{\beta_0} \cdot X_i^{\beta_1} \cdot \underbrace{e^{u_i}}_{\approx 1} \approx e^{\beta_0} \cdot X_i^{\beta_1}$$

marginaler Effekt:

$$\frac{\partial Y_i}{\partial X_i} \approx e^{\beta_0} \cdot \beta_1 X_i^{\beta_1-1} = \beta_1 \frac{Y_i}{X_i}$$

$$\frac{\partial Y_i}{\partial X_i} \cdot \frac{X_i}{Y_i} = \beta_1$$

β_1 gibt die Elastizität von Y_i auf X_i an.

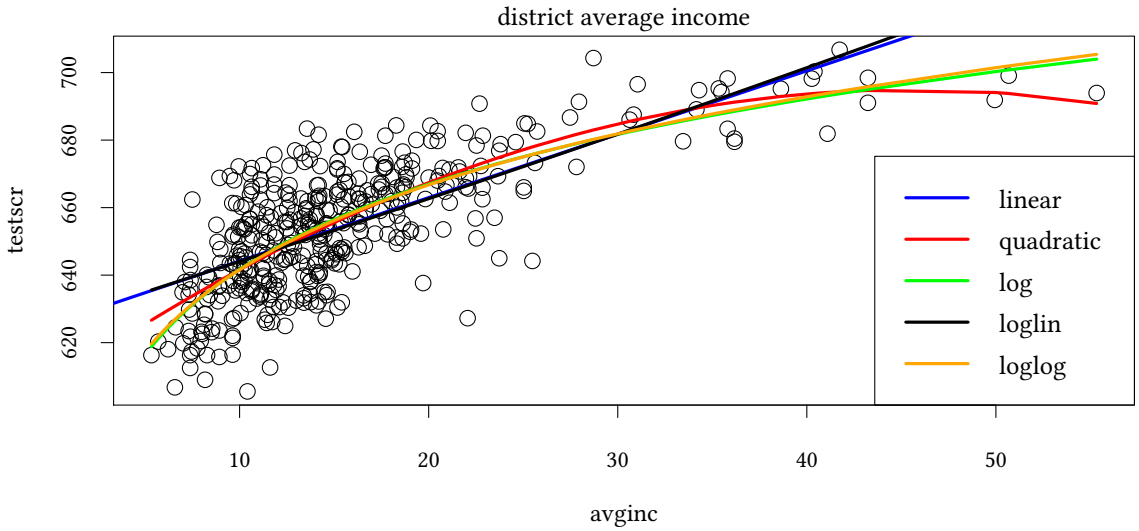
$$\log \texttt{testscr} = \beta_0 + \beta_1 \log \texttt{avginc} + u$$

```
estLLL<- lm(log(testscr) ~ log(avginc))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.3363	0.0065	981.90	0.0000
log(avginc)	0.0554	0.0024	22.96	0.0000

Die Elastizität von `testscr` bezüglich `avginc` ist 0.0554 .

```
plot(testscr ~ avginc,main="district average income")
abline(est1,col="blue",lwd=3)
lines(avginc[or],fitted(est2)[or],col="red",lwd=3)
lines(avginc[or],fitted(estL)[or],col="green",lwd=3)
lines(avginc[or],exp(fitted(estLL))[or],col="black",lwd=3)
lines(avginc[or],exp(fitted(estLLL))[or],col="orange",lwd=3)
legend("bottomright",c("linear","quadratic","log","loglin","loglog"),lwd=3,
       col=c("blue","red","green","black","orange"))
```



Hier sind wieder die marginalen Effekte für verschiedene Werte von `avginc`:

$$\frac{\partial Y_i}{\partial X_i} = e^{\beta_0} \cdot \beta_1 X_i^{\beta_1 - 1} \approx \beta_1 \frac{Y_i}{X_i}$$

```
X<-c(10,40,60)
meLogLog<-exp(coef(estLLL)[1])*coef(estLLL)[2]*X^(coef(estLLL)[2]-1)
```

	10	40	60
meQuad	3.00	0.47	-1.23
meLinLog	3.64	0.91	0.61
meLogLin	1.83	1.99	2.11
meLogLog	3.56	0.96	0.65

12.2.6. Vergleich der logarithmischen Modelle

- X und/oder Y werden jeweils transformiert
- Die Regressionsgleichung ist linear in den transformierten Variablen
- Hypothesentests und Konfizendintervalle können also wie gewohnt bestimmt werden
- Die Interpretation von β ist jeweils unterschiedlich
- R^2 ist geeignet log-log und log-linear zu vergleichen
- R^2 ist geeignet linear-log und lineares Modell zu vergleichen
- Ein Vergleich der Modelle mit $\log Y_i$ und Y_i ist nicht möglich.

$\rightarrow$ ökonomische Theorie ist erforderlich, um eine der Spezifikationen zu motivieren.

```
mtable("linear"=est1,"quadratic"=est2,"linear-log"=estL,"log-linear"=estLL,"log-log"=estLLL,
       summary.stats=c("R-squared","AIC","N"))
```

	linear	quadratic	linear-log	log-linear	log-log
(Intercept)	625.384***	607.302***	557.832***	6.439***	6.336***
	(1.532)	(3.046)	(4.200)	(0.002)	(0.006)
avginc	1.879***	3.851***		0.003***	
	(0.091)	(0.304)		(0.000)	
avginc2		−0.042***			
		(0.006)			
log(avginc)			36.420***		0.055***
			(1.571)		(0.002)
R^2	0.508	0.556	0.563	0.498	0.558
Adj. R^2	0.506	0.554	0.561	0.497	0.557
Num. obs.	420	420	420	420	420

***p < 0.001; **p < 0.01; *p < 0.05

12.3. Nichtlinearitäten in den Parametern

Bislang: Beginne mit einer nichtlinearen Spezifikation. Transformiere diese in eine Spezifikation, die *linear in Parametern ist.*

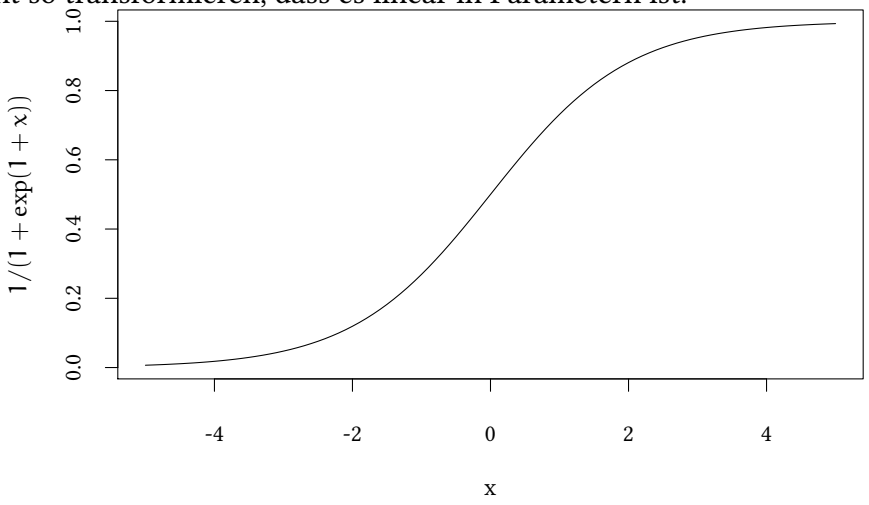
Beispiel:

$$Y_i = \underbrace{e^{\beta_0} \cdot X_i^{\beta_1} \cdot e^{u_i}}_{\text{nicht linear in Parametern}} \quad \rightarrow \quad \log Y_i = \underbrace{\beta_0 + \beta_1 \cdot \log X_i + u_i}_{\text{linear in Parametern}}$$

Das klappt aber nicht immer. Beispiel: Das logistische Modell

$$Y_i = \frac{1}{1 + e^{-\beta_0 + \beta_1 X_i}} + u_i$$

können wir nicht so transformieren, dass es linear in Parametern ist.



- OLS ist nicht mehr möglich
- statt dessen:
  - Nonlinear least squares
  - Maximum-Likelihood Verfahren

12.4. Schlüsselbegriffe

- Nichtlineare Regression
  - Quadratische Regression
  - Polynomiale Regression
  - log-linear / linear-log / log-log Modell
- Marginaler Effekt
- Elastizität

Anhang 12.A Beispiele für die Vorlesung

Betrachten Sie das folgende Regressionsmodell:

$$Y = \beta_0 + \beta_1 \log X_1 + \beta_2 \log X_2 + u$$

Ihre Nullhypothese ist, dass bei einem Wert von $X_1 = 10$ der marginale Effekt von X_1 den Wert -2 hat. Welchen Wert erwarten Sie also für $\hat{\beta}_1$?

Für β_2 schätzen Sie in diesem Modell einen Wert von 12. Was ist der marginale Effekt von X_2 auf Y wenn X_2 den Wert 3 annimmt?

Mit den gleichen Daten schätzen Sie ein neues Regressionmodell und erhalten das folgende Ergebnis:

```
lm(formula = y ~ log(x1) + x2 + x3)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.0        2.0     1.0 0.319962
log(x1)         25.0        2.0    12.5  < 2e-16
x2              -2.0        0.5    -4.0 0.000129
x3               2.5        0.5     5.0 2.77e-06
```

Wie groß ist der marginale Effekt von `x1` an der Stelle `x1` $= 50$?

Wie groß ist der marginale Effekt von `x2` an der Stelle `x2` $= 50$?

Anhang 12.B Übungen

Übung 12.1 *1. Beschreiben Sie die Bedeutung von Nicht-Linearität in den Variablen sowie Nicht-Linearität in den Parametern für die Schätzung mit OLS.*

2. Betrachten Sie folgende Gleichungen und bestimmen Sie, ob diese linear in den Variablen, linear in den Parametern, beides oder keines von beiden sind:

a) $Y_i = \beta_0 + \beta_1 \cdot X_i^3 + \beta_2 \cdot X_i^5 + \epsilon$

b) $Y_i = \beta_0 + \beta_1 \cdot X_i^{\beta_2} + \epsilon$

c) $Y_i = \beta_0 + \beta_1 \cdot \log X_i + \epsilon$

d) $Y_i = \beta_0 - \beta_1 \cdot e^{\beta_2 \cdot X_i} + \epsilon$

e) $Y_i^{\beta_0} = \beta_1 + \beta_2 \cdot X_i^2 + \epsilon$

f) $Y_i = \beta_0 + \beta_1 \cdot \frac{1}{X_i} + \epsilon$

Übung 12.2 *Sie schätzen einen Zusammenhang mit Hilfe unterschiedlicher Modelle. In der folgenden Tabelle sind jeweils die Modelle und die geschätzten Koeffizienten zusammengefasst.*

Modell	β_0	β_1
(1) $Y = \beta_0 + \beta_1 \cdot X + u$	*0.44*	*1.85*
(2) $Y = \beta_0 + \beta_1 \cdot \log X + u$	*1.8*	*0.4*
(3) $\log Y = \beta_0 + \beta_1 \cdot X + u$	*-0.7*	*1.8*
(4) $\log Y = \beta_0 + \beta_1 \cdot \log X + u$	*0.7*	*0.4*

Welches Modell kann ihnen jeweils eine Antwort geben auf die Frage nach...

- *...die marginale Änderung von Y auf eine Änderung von X um ein Prozent?*
- *...die prozentuale Änderung von Y auf eine Änderung von X um ein Prozent?*

- *...die prozentuale Änderung von* Y *auf eine Änderung von* X *um eine Einheit?*
- *...die marginale Änderung von* Y *auf eine Änderung von* X *um eine Einheit?*
- *...die Elastizät von* Y *nach* X*?*

Übung 12.3 *Betrachten Sie den Datensatz Wages und bestimmen Sie den prozentualen Lohnzuwachs für jedes weitere Jahr an Arbeitserfahrung*

Übung 12.4 *Betrachten Sie eine Standard »Cobb-Douglas« Produktionsfunktion*

$$Q = \lambda \cdot K^{\beta_1} \cdot L^{\beta_2} \cdot M^{\beta_3}$$

Dabei ist Q *der produzierte Output,* K *das eingesetzte Kapital,* L *Arbeit, und* M *Verbrauchsmaterial.*

Kann man OLS verwenden, um einen Zusammenhang zwischen diesen Variablen zu schätzen?

Übung 12.5 *Die Nachfrage nach Kaffee wird beschrieben durch*

$$\log Q = \beta_0 + \beta_1 \log P + \beta_2 \dot{Y}$$

Dabei ist Q *die Menge Kaffee,* P *der Preis, und* $\dot{Y}$ *die Wachstumsrate des Volkseinkommens.*

Nehmen Sie an, dass $\beta_0 = 0.7$*,* $\beta_1 = -1$*, und* $\beta_2 = 0.5$*.*

1. *Wie verändert sich* Q*, wenn der Kaffeepreis um 5% steigt?*
2. *Wie verändert sich* Q*, wenn die Wachstumsrate von 5% auf 1% fällt?*

Übung 12.6 *Betrachten Sie nochmals das obige Modell des Kaffeepreises. Neuere Forschungen haben ergeben, dass das Modell den Zusammenhang gut beschreibt, solange* $\dot{Y}$ *positiv ist. Außerhalb dieses Bereichs hat* $\dot{Y}$ *keinen Einfluss auf* Q*, d.h. die Situation* $\dot{Y} = -0.05$ *führt zum gleichen Kaffeekonsum wie* $\dot{Y} = 0$*.*

- *Modellieren Sie diesen Zusammenhang.*

Übung 12.7 *Sie untersuchen wieder, ob das Geschlecht eines Arbeitnehmers Einfluss auf das Bruttoeinkommen hat und führen folgende Regression durch:*

```
Bruttoeinkommen<-c(2500,4000,3000,7000,5000,2900,1500)
Geschlecht<-as.logical(c(1,0,0,0,0,1,1))
summary(lm(log(Bruttoeinkommen)~Geschlecht))
```

```

Call:
lm(formula = log(Bruttoeinkommen) ~ Geschlecht)

Residuals:
       1        2        3        4        5        6        7
 0.12080 -0.12377 -0.41145  0.43585  0.09937  0.26922 -0.39002
```

```
Coefficients:
               Estimate Std. Error t value     Pr(>|t|)
(Intercept)      8.4178     0.1766  47.670 0.0000000767 ***
GeschlechtTRUE  -0.7146     0.2697  -2.649       0.0455 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3532 on 5 degrees of freedom
Multiple R-squared:  0.584, Adjusted R-squared:  0.5007
F-statistic: 7.018 on 1 and 5 DF,  p-value: 0.04547
```

1. *Wie lautet die Regressionsgleichung?*
2. *Welchen Wert nimmt* R^2 *an?*

Übung 12.8 *Ein Autofabrikant untersucht den Zusammenhang zwischen der Anzahl der produzierten Autos (Y) und der Anzahl der eingesetzten Arbeiter (X). Da er nicht weiß, welcher Zusammenhang besteht, vergleicht er das lineare mit dem linear-log Modell mit R. Dabei ergab sich der folgende Output:*

```
x <- c(200,600,800,400,50,1000,500,300)
y <- c(2750,4500,4700,4000,1000,4900,4200,3500)
summary(lm(y~x))


Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-1137.4  -274.6   224.5   444.0   599.5

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1956.9325   452.7692   4.322  0.00497 **
x              3.6090     0.8031   4.494  0.00413 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 667 on 6 degrees of freedom
Multiple R-squared:  0.7709, Adjusted R-squared:  0.7327
F-statistic: 20.19 on 1 and 6 DF,  p-value: 0.004132
```

```
summary(lm(y~log(x)))


Call:
lm(formula = y ~ log(x))
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-169.34  -87.81   15.14   79.78  167.51

Coefficients:
            Estimate Std. Error t value    Pr(>|t|)
(Intercept) -4255.09     313.80  -13.56 0.000009980 ***
log(x)       1349.85      52.69   25.62 0.000000233 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 132.6 on 6 degrees of freedom
Multiple R-squared:  0.9909, Adjusted R-squared:  0.9894
F-statistic: 656.3 on 1 and 6 DF,  p-value: 0.0000002331
```

1. Für welches Modell sollte er sich entscheiden?
2. Sie betrachten nun auch ein log-lineares und log-log Modell.

```
summary(lm(log(y)~x))



Call:
lm(formula = log(y) ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6537 -0.1385  0.1479  0.2144  0.2764

Coefficients:
             Estimate Std. Error t value     Pr(>|t|)
(Intercept) 7.4962744  0.2414628  31.045 0.0000000742 ***
x           0.0013035  0.0004283   3.043       0.0227 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3557 on 6 degrees of freedom
Multiple R-squared:  0.6069, Adjusted R-squared:  0.5413
F-statistic: 9.262 on 1 and 6 DF,  p-value: 0.02271
```

```
summary(lm(log(y)~log(x)))



Call:
lm(formula = log(y) ~ log(x))

Residuals:
     Min       1Q   Median       3Q      Max
-0.17288 -0.11025  0.03017  0.11303  0.13606
```

```
Coefficients:
            Estimate Std. Error t value   Pr(>|t|)
(Intercept)  4.96685    0.32027  15.508 0.00000455 ***
log(x)       0.53607    0.05378   9.968 0.00005897 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1354 on 6 degrees of freedom
Multiple R-squared:  0.9431, Adjusted R-squared:  0.9336
F-statistic: 99.37 on 1 and 6 DF,  p-value: 0.00005897
```

Welches dieser beiden Modelle stellt den Zusammenhang besser dar?

3. Für welches der vier Modelle sollte er sich nun entscheiden?

Übung 12.11 *Sei* P *der Preis und* Q *die nachgefragte Menge,* X *die Sonnenscheindauer an einem Verkaufstag, und* u *der Störterm der Regression. Wir nennen* $\frac{\partial Q}{\partial P}\frac{P}{Q}$ *die Preiselastizität der Nachfrage. Sie schätzen die Gleichung*

$$\log Q = \beta_0 + \beta_P \log P + \beta_X X + u$$

und erhalten folgendes Ergebnis:

	$\hat{\beta}$	$\hat{\sigma}$
β_0	50	10
β_P	5	1
β_X	2	2

Wie groß ist die geschätzte Preiselastizität?

```
## [1] 6.2
```

```
linew

[1] 6.2
```

A. Eine kurze Einführung in R

In diesem Kurs verwenden wir R als ein Beispiel für ein Statistikpaket. Natürlich haben andere Statistikpakete ebenfalls ihre Berechtigung, für diesen Kurs müssen wir uns jedoch für ein Paket entscheiden. R hat den großen Vorteil, dass es Open Source ist. Sie können es ohne Einschränkung auf Ihrem eigenen Rechner installieren. Hinzu kommt, dass R ein sehr leistungsfähiges Paket ist, das mit den kommerziellen Paketen mithalten kann und ihnen in einigen Punkten durchaus überlegen ist.

Auch wenn Sie für Ihre Arbeit später ein anderes Paket verwenden werden, werden Sie feststellen, dass die Konzepte sich in vielen Punkten ähneln.

Auf der Homepage des R Projekts finden Sie zahlreiche Dokumentationen. Probieren Sie aus, welche Ihnen davon gut zusagt.

Für den schnellen Einstieg finden Sie in diesem Kapitel einige Tipps.

Einen freundlichen Zugang zu R bietet die Bibliothek `Rcmdr`. Wir können den `Rcmdr` z.B. mit dem Kommando `library(Rcmdr)` starten.

Im `Rcmdr` sehen wir zwei große Fenster. Das obere ist für Kommandos reserviert, das untere für Ergebnisse. Im oberen Fenster können wir z.B. eingeben

```
2+2
```

und dann auf Submit klicken. Im unteren Fenster erscheint dann das Ergebnis.

Zum Erforschen der Möglichkeiten von R ist der `Rcmdr` recht praktisch. Wenn Sie aber erst einmal wissen, wohin Sie wollen, erscheint er mir eher hinderlich. Aus diesem Grund werde ich ihn in der Vorlesung nicht verwenden. Alle Befehle, die ich hier oder in der Vorlesung verwende, können Sie aber auch im Eingabefenster des `Rcmdr` eintippen. Das gibt Ihnen die Möglichkeit, im Menü nach Alternativen suchen zu können, ohne gleich die Dokumentation Ihrer Wahl zu benutzen.

A.1. Installation von R

Auf der Homepage des R Projekts finden Sie im Menü auf der linken Seite einen Link Download / CRAN. Dieser führt Sie zu einer Auswahl von »Mirrors« auf der ganzen Welt. In Jena liegt möglichweise der GWDG Mirror in Göttingen besonders nahe. Dort finden Sie Anleitungen für die gängigen Betriebssysteme.

Installation von Libraries Wenn das Kommando `library` die gewünschte Bibliothek nicht findet, ist sie vielleicht nicht installiert. In einer vollständigen Installation von R können Sie fehlende Bibliotheken leicht nachinstallieren. Das Kommando

```
install.packages("Ecdat")
```

installiert beispielsweise die Bibliothek `Ecdat`. Bei manchen Installationen gibt es auch ein Menü »Packages« das Ihnen erlaubt, fehlende Libraries nachzuinstallieren. Benutzer von Betriebssystemen der Firma Microsoft finden in der FAQ for Packages hilfreiche Tipps.

A.2. Datentypen und Zuweisungen

R kennt verschiedene Datentypen. Einige davon werden wir in diesem Kapitel kennenlernen. Um eine Zahl (oder einen Vektor, oder ein beliebiges Objekt) einer Variablen zuweisen zu können, verwenden wir den Operator `<-`

```
x <- 4
```

R speichert das Ergebnis dieser Zuweisung als `double`.

```
typeof(x)

[1] "double"
```

Jetzt können wir mit x rechnen:

```
2 * x

[1] 8

sqrt(x)

[1] 2
```

Oft werden wir nicht nur mit einer einzigen Zahl (einem Skalar) rechnen, sondern mit mehreren die in einem Vektor zusammengefasst werden. Mehrere Zahlen werden mit `c` zu einem Vektor zusammengefasst.

```
x <- c(21,22,23,24,25,16,17,18,19,20)
x

 [1] 21 22 23 24 25 16 17 18 19 20
```

Wenn wir eine lange Liste von aufeinanderfolgenden Zahlen brauchen (wie in diesem Beispiel) hilft der Operator `:` oder die Funktion `seq`.

```
21:30

 [1] 21 22 23 24 25 26 27 28 29 30

seq(21,30)

 [1] 21 22 23 24 25 26 27 28 29 30

y <- 21:30
```

Subsets Auf einzelne Elemente eines Vektors können wir mit `[ ]` zugreifen.

```
x[1]

[1] 21
```

Wenn wir auf mehrere Elemente zugreifen wollen, verwenden wir einfach mehrere Indizes (die mit `c` verbunden werden). Das können wir auch verwenden, um die Reihenfolge der Werte zu ändern.

```
x[c(3,2,1)]

[1] 23 22 21

x[3:1]

[1] 23 22 21
```

```
x
```

```
 [1] 21 22 23 24 25 16 17 18 19 20
```

(um einen längeren Vektor zu sortieren, würden wir die Funktion `order` verwenden).

```
order(x)
```

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

```
x[order(x)]
```

```
 [1] 16 17 18 19 20 21 22 23 24 25
```

(`order` berechnet zunächst nur eine »Ordnung«, also eine Reihenfolge, in der die Elemente eines Datensatzes angeordnet werden müssen, um geordnet zu sein. Wir verwenden `x[...]` um das geordnete Ergebnis anzusehen.)

Negative Indizes lassen Elemente weg:

```
x[-1:-3]
```

```
[1] 24 25 16 17 18 19 20
```

Wahrheitswerte Wahrheitswerte können entweder `TRUE` oder `FALSE` sein. Wenn wir einen Vektor mit einer Zahl vergleichen, dann wird jedes Element verglichen (das folgt aus der *recycling Regel*, siehe unten):

```
x
```

```
 [1] 21 22 23 24 25 16 17 18 19 20
```

```
x < 20
```

```
 [1] FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE FALSE
```

```
typeof(x < 20)
```

```
[1] "logical"
```

Wir können auch Wahrheitswerte als Index verwenden:

```
x [ x < 20 ]
```

```
[1] 16 17 18 19
```

Zeichenketten Genauso wie Zahlen, so können wir auch Zeichenketten einer Variablen zuweisen:

```
x <- "Mary"
typeof(x)

[1] "character"
```

Wir können auch mit Vektoren von Zeichenketten arbeiten:

```
x <- c("John","Mary","Jane")
x[2]

[1] "Mary"

x[3]<-"Lucy"
x

[1] "John" "Mary" "Lucy"
```

Faktoren Oftmals ist es umständlich, eine Zeichenkette, die in einem Datensatz immer wieder vorkommt, jedesmal erneut zu speichern. Wir wollen uns z.B. merken, ob eine Beobachtung zu einem Mann oder einer Frau gehört. Effizient merken wir uns z.B. für `male"` eine 2, und für `female"` eine 1.

```
x <- factor(c("male","female",
              "female","male"))
typeof(x)

[1] "integer"

class(x)

[1] "factor"

levels(x)

[1] "female" "male"
```

```
x[2]

[1] female
Levels: female male

as.numeric(x)

[1] 2 1 1 2
```

R behandelt Faktoren auf eine sehr transparente Weise. Der Benutzer kommt normalerweise nur mit den Zeichenketten, nicht mit den Zahlen, in Berührung.

Listen Listen verbinden potenziell verschiedene Datentypen:

```
x <- list(a=123,b="hello world",c=3)
x[[1]]

[1] 123

x[["a"]]

[1] 123

x$a

[1] 123

x$b

[1] "hello world"
```

Natürlich können wir auch Listern verschachteln:

```
y <- list(g=456,h="hello world",i=x)
y$i$c

[1] 3

y[["i"]][["c"]]

[1] 3

typeof(y)

[1] "list"

class(y)

[1] "list"
```

Dataframes Sehr oft arbeiten wir mit »rechteckigen« Datenstrukturen, d.h. Listen, bei denen alle Elemente Vektoren der gleichen Länge sind.

```
x <- data.frame(a=1:3,b=c("a","b","c"))
x

  a b
1 1 a
2 2 b
3 3 c

x$a
```

```
[1] 1 2 3

x$b

[1] "a" "b" "c"
```

```
x[["b"]]

[1] "a" "b" "c"

x[,"b"]

[1] "a" "b" "c"

x[1:2,]

  a b
1 1 a
2 2 b

typeof(x)

[1] "list"
```

A.3. Funktionen

Es gibt eine Menge eingebaute Funktionen:

```
mean(x)
median(x)
max(x)
min(x)
length(x)
unique(c(1,2,3,4,1,1,1))
```

Wenn uns die nicht mehr ausreichen, können wir selbst welche schreiben:

```
square <- function(x) {
  x*x
}
```

Der letzte Wert einer Funktion (hier `x*x`) ist jeweils das Ergebnis (der Rückgabewert). Nun können wir die Funktion verwenden.

```
square(7)

[1] 49
```

Wenn wir eine Funktion auf viele Zahlen anwenden wollen, hilft `sapply`:

```
range <- 1:10
sapply(range,square)

 [1]   1   4   9  16  25  36  49  64  81 100
```

Wir müssen uns keinen extra Namen für die Funktion ausdenken:

```
sapply(range,function(x) x*x)

 [1]   1   4   9  16  25  36  49  64  81 100
```

A.4. Zufallszahlen

Zufallszahlen können für sehr unterschiedliche Verteilungen generiert werden. R berechnet pseudo-Zufallszahlen, d.h. es nimmt Zahlen aus einer langen Liste. Den Startwert dieser Liste legt das Kommando `set.seed` fest.

```
set.seed(123)
```

10 pseudo-normalverteilte Zufallszahlen erhalten wir dann mit

```
rnorm(10)

 [1] -0.56047565 -0.23017749  1.55870831  0.07050839  0.12928774  1.71506499
 [7]  0.46091621 -1.26506123 -0.68685285 -0.44566197
```

Wir erhalten die gleiche Liste, wenn wir die Liste auf den gleichen Startwert setzen:

```
set.seed(123)
rnorm(10)

 [1] -0.56047565 -0.23017749  1.55870831  0.07050839  0.12928774  1.71506499
 [7]  0.46091621 -1.26506123 -0.68685285 -0.44566197
```

Das ist nützlich, wenn wir die gleichen »zufälligen« Ergebnisse zuverlässig reproduzieren wollen.

10 gleichverteilte Zufallszahlen auf dem Intervall $[100, 200]$ gibt es mit

```
runif(10,min=100,max=200)

 [1] 188.9539 169.2803 164.0507 199.4270 165.5706 170.8530 154.4066 159.4142
 [9] 128.9160 114.7114
```

Wir verwenden Zufallszahlen oft, um Prozesse zu simulieren. Um einen Prozess oft zu wiederholen, hilft das Kommando `replicate`.

```
replicate(10,mean(rnorm(100)))

 [1]  0.016749257 -0.024755975  0.061320514 -0.028205903  0.087712299
 [6] -0.025113287 -0.141043824  0.123989920  0.109293109 -0.002743263
```

nimmt beispielsweise 10 mal den Mittelwert von jeweils 100 pseudo-normalverteilten Zufallszahlen.

A.5. Beispiel-Datensätze

Wir haben gerade gesehen, dass das Kommando `c` uns erlaubt, die Elemente eines Vektors einfach einzugeben. Das ist für lange Datensätze jedoch sehr lästig und auch nur selten notwendig. R bringt von sich aus bereits sehr viele Datensätze mit. Diese Datensätze sind, wie auch viele statistische Funktionen, in Bibliotheken (libraries) organisiert. Da der Funktionsumfang aller Bibliotheken sehr groß ist, werden zu Beginn nur wenige Bibliotheken geladen. Weitere Bibliotheken können jedoch jederzeit mit dem Kommando `library` nachgeladen werden.

Wir werden oft die Bibliothek `Ecdat` verwenden. Sie stellt verschiedene ökonometrische Datensätze bereit. Außerdem benutzen wir oft die Bibliothek `car` die uns zu einigen praktischen ökonometrischen Funktionen hilft.

Wenn wir eine bestimmte Funktion suchen, und nicht wissen, welche Bibliothek dafür zuständig ist, hilft das Kommando `RSiteSearch` oder die R Site Search Extension für Firefox.

Der Datensatz `BudgetFood` ist z.B. in der Bibliothek `Ecdat` enthalten.

```
library(Ecdat)
data(BudgetFood)
```

Um die ersten Zeilen des Datensatzes anzusehen, hilft das Kommando `head`:

```
head(BudgetFood)
```

```
      wfood  totexp age size town   sex
1 0.4676991 1290941  43    5    2   man
2 0.3130226 1277978  40    3    2   man
3 0.3764819  845852  28    3    2   man
4 0.4396909  527698  60    1    2 woman
5 0.4036149 1103220  37    5    2   man
6 0.1992503 1768128  35    4    2   man
```

Um mehr über die Struktur zu erfahren, verwenden wir das Kommando `str`:

```
str(BudgetFood)
```

```
'data.frame': 23972 obs. of  6 variables:
 $ wfood : num  0.468 0.313 0.376 0.44 0.404 ...
 $ totexp: num  1290941 1277978 845852 527698 1103220 ...
 $ age   : num  43 40 28 60 37 35 40 68 43 51 ...
 $ size  : num  5 3 3 1 5 4 4 2 9 7 ...
 $ town  : num  2 2 2 2 2 2 2 2 2 2 ...
 $ sex   : Factor w/ 2 levels "man","woman": 1 1 1 2 1 1 1 2 1 1 ...
```

Normalerweise wollen wir aber gerade *nicht* die vielen Zahlen ansehen, sondern strukturiert daraus (übersichtliche) Zahlen ableiten (Parameter oder Konfidenzintervalle, p-Werte, ...)

Die Bedeutung der einzelnen Spalten in diesem Datensatz können wir mit dem Kommando `help` erfahren.

```
help(BudgetFood)
```

Ein wichtiges Kommando um eine Übersicht zu erhalten ist `summary`

```
summary(BudgetFood)
```

Wir können wir nun auf einzelne Spalten unseres Datensatzes zugreifen? Da R mehrere Datensätze gleichzeitig kennen kann, gibt es viele Möglichkeiten. Eine Möglichkeit ist es, an den Namen des Datensatzes (`BudgetFood`) mit einem Dollarzeichen den Namen der Variablen anzuhängen.

```
BudgetFood$age
```

```
 [1] 43 40 28 60 37 35 40 68 43 51 43 48 51 58 61 53 58 64 50 50 47 76 49 44 49
[26] 51 56 63 30 70 29 60 50 56 36 46 43 32 45 34
 [ reached getOption("max.print") -- omitted 23932 entries ]
```

Das ist praktisch, wenn wir mit vielen unterschiedlichen Datensätze gleichzeitig arbeiten wollen.

Wir sehen an diesem Beispiel auch, dass R, um Ihren Bildschirm nicht mit langen Zahlenkolonnen zu überschwemmen, jeweils nur einen Teil des langen Vektors anzeigt und auf diese Unterlassung Rest mit »`omitted ... entries`« hinweist.

Wenn wir aber nur mit einem Datensatz rechnen wollen, hilft das Kommando `attach`.

```
attach(BudgetFood)
age
```

```
 [1] 43 40 28 60 37 35 40 68 43 51 43 48 51 58 61 53 58 64 50 50 47 76 49 44 49
[26] 51 56 63 30 70 29 60 50 56 36 46 43 32 45 34
 [ reached getOption("max.print") -- omitted 23932 entries ]
```

Ab jetzt werden alle Variablen zunächst im Datensatz `BudgetFood` gesucht. Wenn wir das nicht mehr wollen, sagen wir

```
detach(BudgetFood)
```

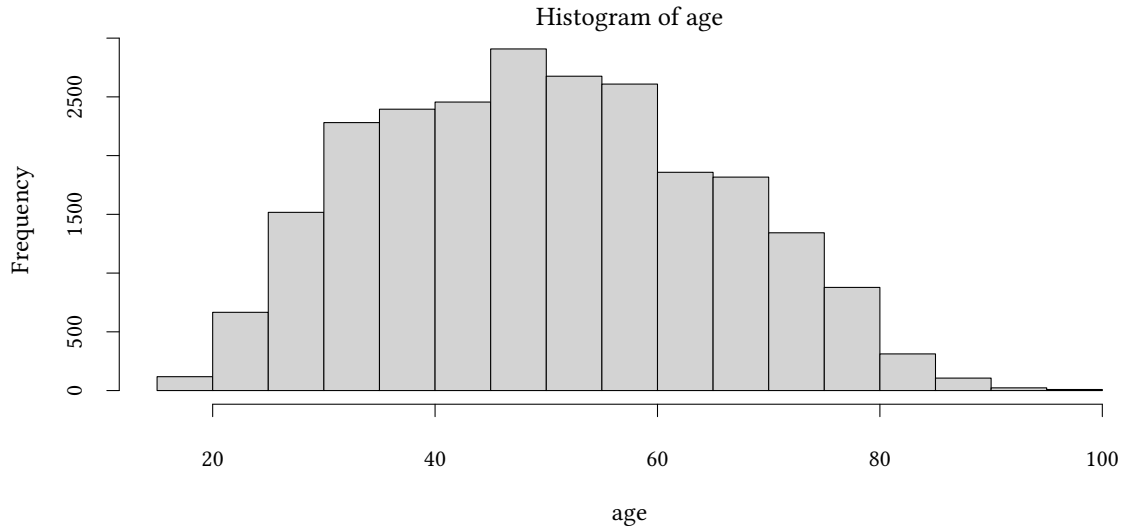
Eine dritte Möglichkeit ist das Kommando `with`:

```
with(BudgetFood,age)
```

```
 [1] 43 40 28 60 37 35 40 68 43 51 43 48 51 58 61 53 58 64 50 50 47 76 49 44 49
[26] 51 56 63 30 70 29 60 50 56 36 46 43 32 45 34
 [ reached getOption("max.print") -- omitted 23932 entries ]
```

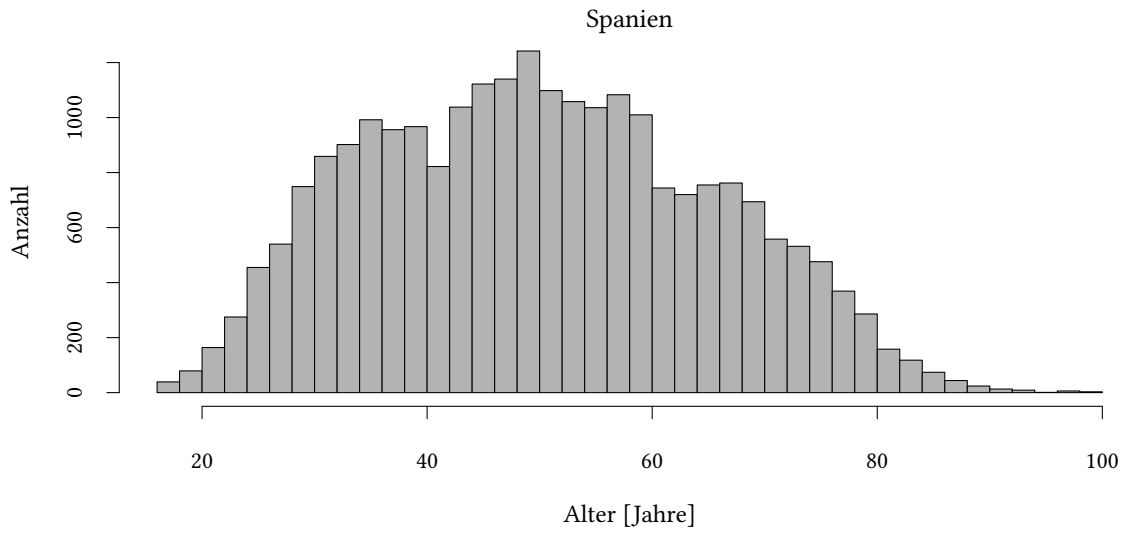
Natürlich können wir uns die Variablen auch grafisch ansehen. `hist` zeichnet z.B. ein Histogramm:

```
with(BudgetFood,hist(age))
```



Die meisten Kommandos haben zahlreichen Optionen, die Ihnen erlauben, das Ergebnis anzupassen. Werfen Sie einmal einen Blick in die Hilfeseite für `hist` die Sie mit `help(hist)` erreichen. Vielleicht gefällt Ihnen die folgende Grafik ja besser:

```
with(BudgetFood,hist(age,breaks=40,xlab="Alter [Jahre]",ylab="Anzahl",col=gray(.7),main="Spanien"))
```



A.6. Grafiken

Es gibt unterschiedliche Möglichkeiten, Zahlen grafisch darzustellen.

Darstellungsweisen für die Verteilung einer Variablen, die wir in der Vorlesung häufig verwenden, sind die folgenden:

```
with(BudgetFood, {
  hist(age)
  plot(density(age))
  boxplot(age ~ sex,main="Boxplot")
})
```

A.6.1. Densityplot

Der Densityplot zeigt eine Information wie das Histogramm. Es gibt zwei wesentliche Unterschiede.

- Die vertikale Achse zeigt die Dichte, d.h. die Häufigkeit/Gesamtanzahl der Beobachtungen.
- Die Linie versucht »glatt« zu sein. Sie zeigt, gegeben die Beobachtungen die wir haben, die *geschätzte* Dichte einer Verteilung, wenn wir denn unendlich viele Beobachtungen haben.

A.6.2. Boxplot

- Die Grenzen der Box sind das erste und dritte Quartil. Der dicke Strich in der Mitte ist der Median. Wir nennen den Abstand vom ersten zum dritten Quartil den Interquartilsabstand (IQR).
- Die gestrichelten Linien reichen vom bis zu dem Punkt der nicht weiter als $1.5 \times$ IQR vom Median entfernt ist.
- Jeder Wert der weiter als $1.5 \times$ IQR vom Median entfernt ist (Ausreißer), wird extra als Punkt markiert.

Warum nimmt man gerade $1.5 \times$ IQR? — Wenn die Daten normalverteilt sind, dann ist das Ende der gestrichelten Linien etwa 2 Standardabweichungen vom Mittelwert entfernt.

```
1.5 * (qnorm(.75) - qnorm(.25))
```

```
[1] 2.023469
```

Das ist recht nahe dem 2.5% und 97.5% Quantil:

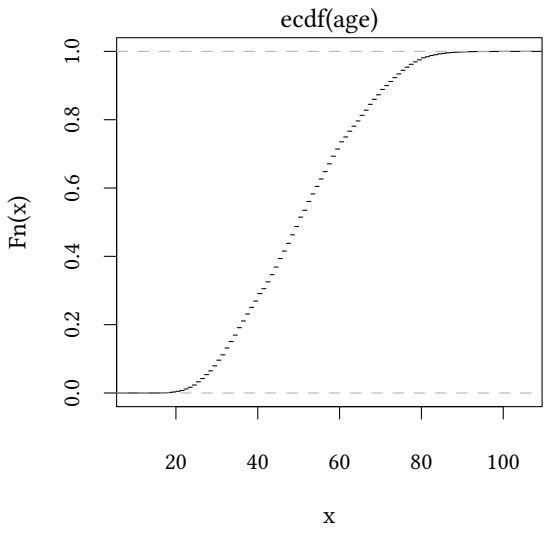
```
qnorm(.975)
```

```
[1] 1.959964
```

das heißt, bei einer Normalverteilung liegen etwa 95% der Werte innerhalb der gestrichelten Linien und nur 5% außerhalb.

A.6.3. Empirische kumulierte Verteilung

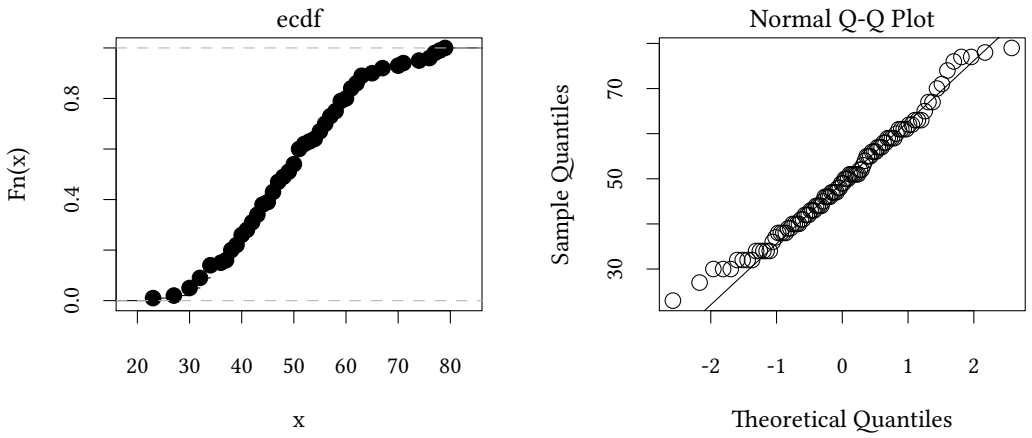
```
plot(ecdf(age),do.points=FALSE)
```



Das Diagramm zeigt auf der horizontan Achse den Wert der Beobachtung, auf der vertikalen Achse die empirische Verteilung, also den Anteil der Werte die kleiner sind als der Wert der jeweiligen Beobachtung.

```
x <- sample(BudgetFood$age,
            100)
plot(ecdf(x),main="ecdf")
```

```
qqnorm(x)
qqline(x)
```



A.6.4. Q-Q Normal Plot

Dieses Diagram (oben rechts) zeigt, ob eine Variable etwa normalverteilt ist (es gibt Varianten für andere Verteilungen). An der vertikalen Achse wird die Variable dargestellt, an der

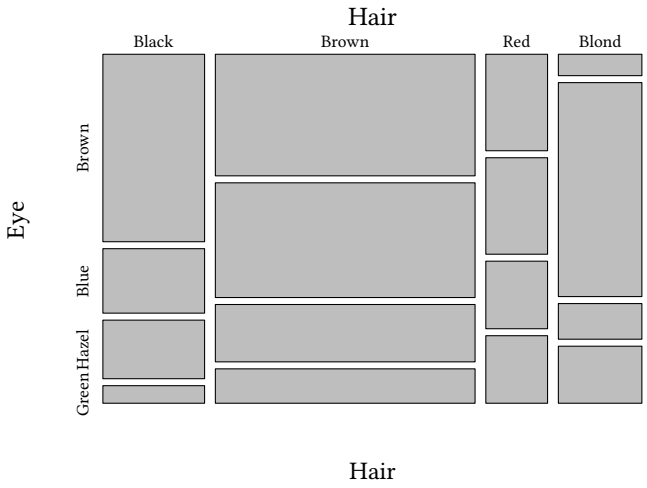
horizontalen Achse die Quantile der Normalverteilung, die der empirischen Verteilung der Variablen entsprechen. Entspricht die Verteilung also einer Normalverteilung, dann liegen alle Punkte auf einer Geraden. `qqline` ist eine Gerade, die durch das erste und dritte Quartil geht.

- Manchmal ist es offensichtlich, wie wir unsere Daten für diese Funktionen vorbereiten. Manchmal ist die Aufbereitung der Daten kompliziert. Dann helfen uns andere Kommandos die ein Objekt berechnen, das gezeichnet werden kann (mit `plot`)
  - `density`, `ecdf`, `xyplot`...
- Einige Kommandos zeichnen unsere aufbereiteten Daten:
  - `plot`, `hist`, `boxplot`, `barplot`, `curve`, `mosaicplot`,...
- Wieder andere Kommandos zeichnen etwas in eine vorhandene Grafik:
  - `points`, `text`, `lines`, `abline`, `qqline`...

A.6.5. Mosaicplot

Der Mosaicplot stellt die gemeinsame Häufigkeitsverteilung von Merkmalen dar. Die Größe der Flächen ist proportional zu den Häufigkeiten. Abhängigkeiten zwischen den Merkmalen fallen leicht als ungleichmäßige Flächen auf.

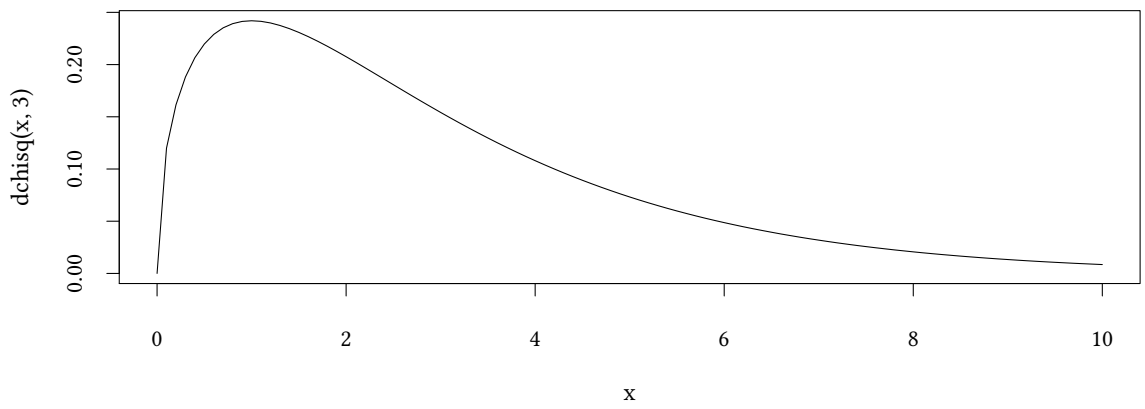
```
mosaicplot(HairEyeColor[,,"Male"],main="Hair")
```



A.6.6. Graphen von Funktionen

`curve` zeichnet Graphen von Funktion von `x`.

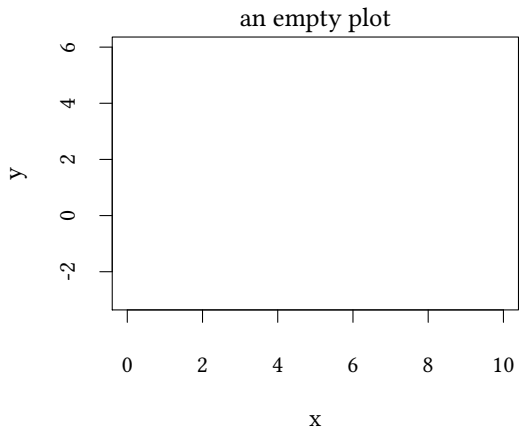
```
curve(dchisq(x,3),from=0,to=10)
```



A.6.7. Leere Plots

Manchmal hilft es, mit einem leeren Plot zu starten. Allerdings müssen wir dann `plot` ein bisschen unter die Arme greifen. Normalerweise kann `plot` aus den Daten schließen, welchen Bereich die Achsen abdecken und wie die Achsen benannt sind. Mit einem leeren Plot müssen wir diese Angaben als Option hinzufügen.

```
plot(NULL,xlim=c(0,10),ylim=c(-3,6),xlab="x",ylab="y",main="an empty plot")
```

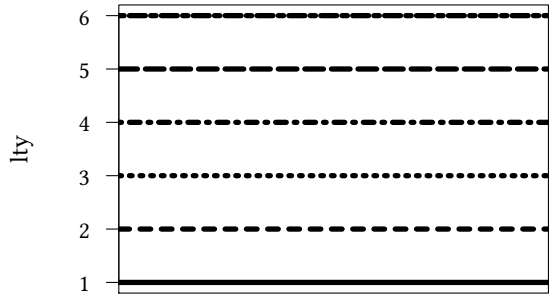


A.6.8. Linientyp

Fast alle Kommandos, die etwas zeichnen, halten sich an die folgenden Konventionen:

- `lty` Linientyp (`"dashed"`, `"dotted"`, oder einfach eine Zahl)

```
plot(NULL,ylim=c(1,6),xlim=c(0,1),xaxt="n",ylab="lty",las=1)
sapply(1:6,function(lty) abline(h=lty,lty=lty,lwd=5))
```

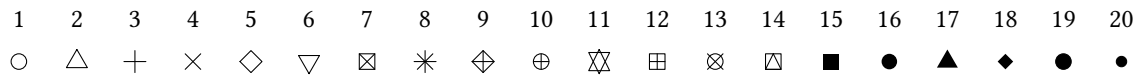


- `lwd` Liniendicke (eine Zahl)
- `col` Farbe (`red"`, `"green"`, `gray(0.5)`)

A.6.9. Punktstil

Die Form von Punkten wird mit `pch` festgelgt.

```
range=1:20
plot(range,range/range,pch=range,frame=FALSE)
text(range,range/range+.2,range)
```

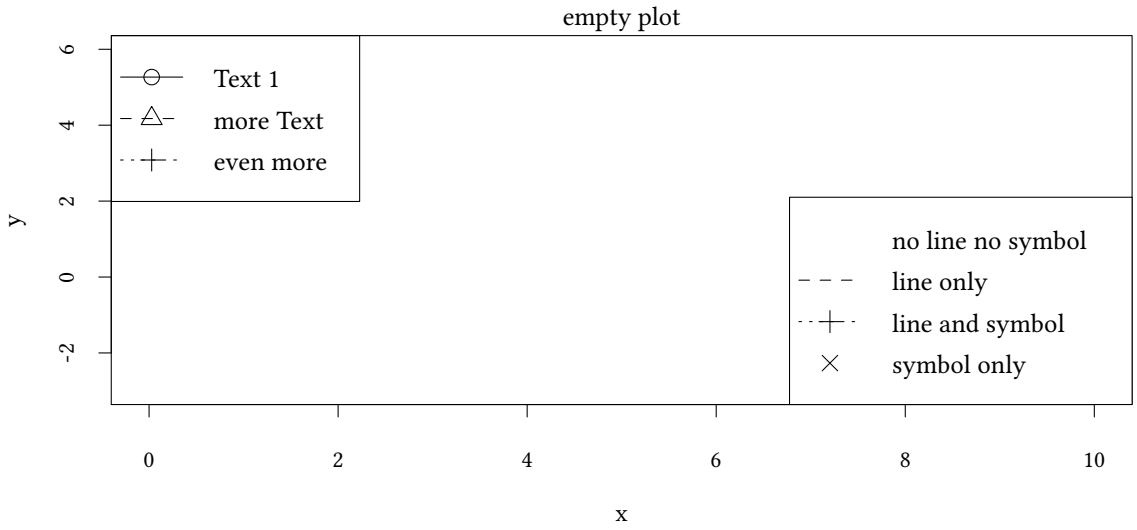


A.6.10. Legenden

Wenn wir mehr als eine Linie oder mehr als ein Symbol in unserem Plot verwenden, dann müssen wir die Bedeutung der Linien und Symbole erklären. Das geschieht in einer Legende

Normalerweise erhält `legend` einen Vektor von Linienstilen `lty` und Symbolen `pch`. Die werden verwendet um eine Beispiellinie und ein Beispielsymbol neben dem Text der Legende zu zeichnen. Wenn `lty` oder `pch` den Wert `NA` haben, dann wird keine Linie oder kein Symbol gezeichnet.

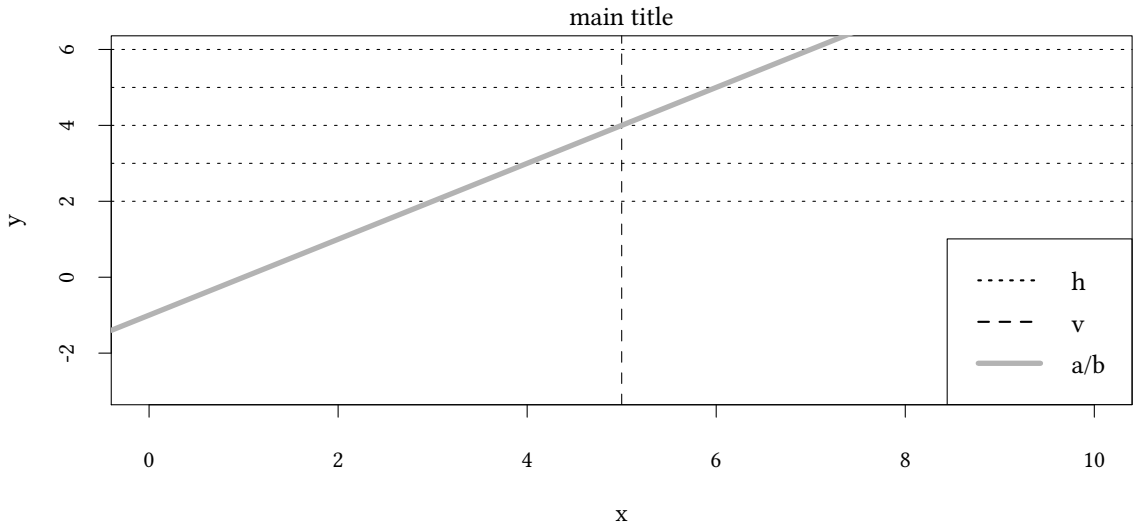
```
plot(NULL,xlim=c(0,10),ylim=c(-3,6),xlab="x",ylab="y",main="empty plot")
legend("topleft",c("Text 1","more Text","even more"),lty=1:3,pch=1:3)
legend("bottomright",c("no line no symbol","line only","line and symbol","symbol only"),
       lty=c(NA,2,3,NA),pch=c(NA,NA,3,4),bg="white")
```



A.6.11. Hilfslinien

Wenn wir in eine Grafik Hilfslinien einzeichnen wollen, hilft das Kommando `abline`

```
plot(NULL,xlim=c(0,10),ylim=c(-3,6),xlab="x",ylab="y",main="main title")
abline(h=2:6,lty="dotted")
abline(v=5,lty="dashed")
abline(a=-1,b=1,lwd=5,col=grey(.7))
legend("bottomright",c("h","v","a/b"),lty=c("dotted","dashed","solid"),col=c("black","black",grey(.7)),lw
```



`abline` kennt die folgenden wichtigen Parameter:

- `h=` für horizontale Linien

- `v=` für vertikale Linien
- `a=..., b=...` für Linien mit Achsenabschnitt `a` und Steigung `b`

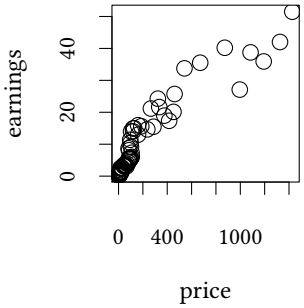
Diese Parameter können auch Vektoren sein, wenn wir mehrere Linien gleichzeitig zeichnen wollen.

A.6.12. Achsen

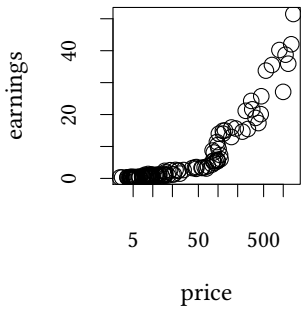
Die Optionen `log='x'`, `log='y'` oder `log='xy'` legen fest, welche Achse logarithmisch dargestellt wird.

```
data(PE,package="Ecdat")
xx<-data.frame(PE)
attach(xx)
```

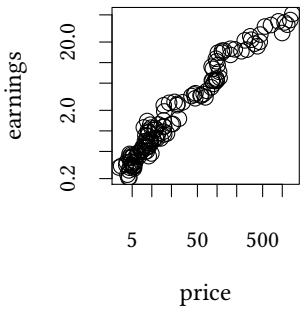
```
plot(price, earnings)
```




```
plot(price, earnings,
     log="x")
```

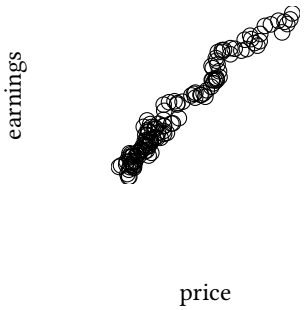



```
plot(price, earnings,
     log="xy")
```

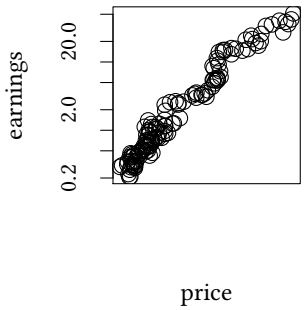


Um mehr Gestaltungsspielraum zu haben, kann man die Achsen entweder ganz entfernen (`axes=FALSE`) oder teilweise unterdrücken (`xaxt="n"` oder `yaxt="n"`) um dann mit `axis` eine neue Achse zu zeichnen.

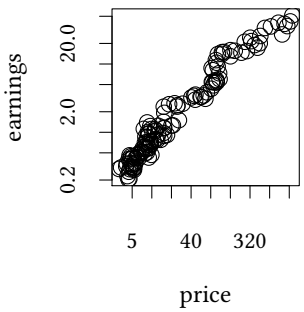
```
plot(price, earnings,
   log="xy",axes=FALSE)
```




```
plot(price, earnings,
     log="xy",xaxt="n")
```




```
plot(price, earnings,
     log="xy",xaxt="n")
axis(1,at=c(5,10,20,40,
     80,160,320,640,1280))
```



Wenn wir viele Labels für die Achsen angeben (wie im obigen Beispiel), wird R nicht alle davon drucken, falls sie überlappen.

A.7. Tabellen

Häufigkeiten Das Kommando `table` berechnet eine Häufkeitstabelle. Hier schauen wir uns nur die ersten 16 Spalten an:

```
with(BudgetFood,table(sex ,age ))[,1:16]

       age
sex      16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31
  man     3   6  21  21  36  37  87 100 132 201 210 248 254 329 367 363
  woman   0   2   7   9  12  21  19  21  22  26  18  28  10  25  28  12
```

Andere Statistiken Das Kommando `aggregate` gruppiert unsere Daten nach dem Wert eines oder mehrerer Faktoren und wendet dann eine Funktion auf jede Gruppe an. Im folgenden Beispiel werden die Gruppen gebildet durch die Variable `sex`, die Funktion ist der Mittelwert `mean` der von der Variablen `age` berechnet wird.

```
with(BudgetFood,aggregate(age ~ sex,FUN=mean))

    sex      age
1   man 49.08985
2 woman 59.47445
```

A.8. Regressionen

Einfache Regressionen können wir mit dem Kommando `lm` schätzen. Der Operator `~` hilft uns, eine Regressionsgleichung anzugeben. Links von `~` steht die abhängige Variable, links die unabhängigen Variablen.

```
lm (wfood ~ totexp,data=BudgetFood)


Call:
lm(formula = wfood ~ totexp, data = BudgetFood)

Coefficients:
  (Intercept)         totexp
 0.4950397225  -0.0000001348
```

Das Ergebnis ist noch etwas trocken. Mehr Details erhalten wir mit dem Kommando `summary`.

```
summary(lm (wfood ~ totexp,data=BudgetFood))
```

```
Call:
lm(formula = wfood ~ totexp, data = BudgetFood)

Residuals:
     Min       1Q   Median       3Q      Max
-0.49307 -0.09374 -0.01002  0.08617  1.06182

Coefficients:
                   Estimate      Std. Error t value Pr(>|t|)
(Intercept)  0.495039722500  0.001561819134  316.96   <2e-16 ***
totexp      -0.000000134849  0.000000001459  -92.41   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1422 on 23970 degrees of freedom
Multiple R-squared:  0.2627, Adjusted R-squared:  0.2626
F-statistic:  8540 on 1 and 23970 DF,  p-value: < 2.2e-16
```

A.9. Starten und Verlassen von R

Wann immer Sie R aufrufen, versucht das Programm zuerst im aktuellen Verzeichnis, dann in Ihrem Home-Verzeichnis eine Datei `.Rprofile` zu lesen und auszuführen. Das ist praktisch, wenn Sie bei jedem Start die gleichen Kommandos ausführen wollen. Z.B. sorgt die Zeile

```
options(browser = "/usr/bin/firefox")
```

in `.Rprofile` dafür, dass das Hilfesystem immer den Browser `firefox` verwendet.
Auch beim Verlassen versucht Ihnen R die Arbeit einfach zu machen. Auf das Kommando

```
q()
```

antwortet R zunächst mit

```
Save workspace image? [y/n/c]:
```

Hier haben Sie die Möglichkeit, alle Daten, mit denen Sie im Moment arbeiten, als Datei `.Rdata` im aktuellen Verzeichnis abzuspeichern. Beim nächsten Start (aus diesem Verzeichnis) liest R diese Datei automatisch ein, und Sie können mit Ihrer Arbeit fortfahren.