

# Statistik\_\_21

*Sebastian Sauer*

*2017-02-03*



# Contents

```
source("../source/libs.R")
```

```
## Loading tidyverse: ggplot2
```

```
## Loading tidyverse: tibble
```

```
## Loading tidyverse: tidyr
```

```
## Loading tidyverse: readr
```

```
## Loading tidyverse: purrr
```

```
## Loading tidyverse: dplyr
```

```
## Conflicts with tidy packages -----
```

```
## filter(): dplyr, stats
```

```
## lag():      dplyr, stats
```



# Vorwort

Es gibt noch kein gutes Buch in deutscher Sprache zu den Grundlagen moderner Statistik, auch “Data Science” genannt. Dieses Buch soll helfen, einen Teil dieser Lücke zu füllen. Die Zielgruppe sind Analysatoren mit praktischem, wirtschaftsnahem Hintergrund. Auf mathematische Hintergründe wird größtenteils verzichtet; Matheliebhaber werden kaum auf ihre Kosten kommen. Im Blick habe ich (hier spricht der Autor) Anwender, die einen Freischwimmer in der modernen Datenanalyse erlernen möchten (oder müssen, liebe Studierende).

Dieses Buch wurde mit dem Paket `bookdown` [bookdown] erstellt, welches wiederum stark auf den Paketen `knitr` (?) und `rmarkdown` (?) beruht. Diese Pakete stellen verblüffende Funktionalität zur Verfügung als freie Software (frei wie in Bier und frei wie in Freiheit).

- Worum geht es in diesem Buch
  - Einführung in moderne Verfahren der Statistik
  - Für Praktiker
  - Betonung liegt auf “modern” und “Praktiker”
- Ziel des Buches
  - Intuitives, grundlegendes Verständnis zu zentralen Konzepten
  - Handwerkszeug zum selber Anwenden
- Unterschied zu anderen Büchern
  - Wenig Formeln
  - Keine/weniger “typischen” klassischen Methoden wie ANOVA, Poweranalyse etc.
  - Aufzeigen von Problemen mit klassischen Verfahren
  - Kritik am Status-Quo
- Didaktik
  - Hands-on
  - R
  - Lernfragen
  - Fallstudien
  - Aktuelle Entwicklungen ausgerichtet

```
library(knitr)
```



# Einführung

## 0.1 Rahmen

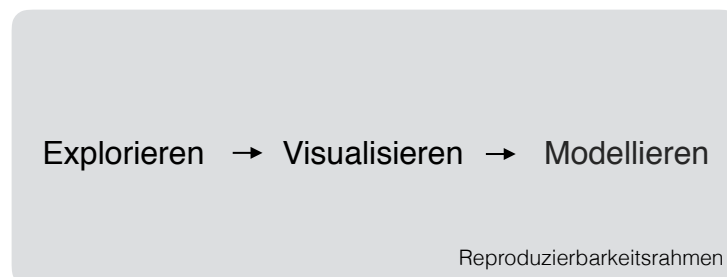
Der “Rahmen” dieses Buches ist der Überblick über wesentliche Schritte der Datenanalyse (aus meiner Sicht). Es gibt viele Ansätze, mit denen der Ablauf von Datenanalyse dargestellt wird. Der im Moment populärste oder bekannteste ist wohl der von Hadley Wickham und Garret Golemund (?). Hadley und Garrett haben einen “technischeren” Fokus als der dieses Buches. Ihr Buch “R for Data Science” ist hervorragend (und frei online verfügbar); nur ist der Schwerpunkt ein anderer; es baut ein viel tieferes Verständnis von R auf. Hier spielen aber statistisch-praktisch und statistisch-philosophische<sup>1</sup> Aspekte eine größere Rolle.

Das Diagramm @ref(“Rahmen”) stellt den Rahmen dieses Buch dar: Die drei Hauptaspekte sind *Umformen*, *Visualisieren* und *Modellieren*. Dies ist vor dem Hintergrund der *Reproduzierbarkeit* eingebettet. Dieser Rahmen spiegelt das hier vertretene Verständnis von Datenanalyse wieder, wobei es sich nicht unbedingt um eine Abfolge von links nach rechts handeln muss. Wilde Sprünge sind erlaubt und nicht unüblich.

Mit *Umformen* ist gemeint, dass Daten in der Praxis häufig nicht so sind, wie man sie gerne hätte. Mal fehlt eine Variable, die den Mittelwert anderer ausdrückt, oder es gibt unschöne “Löcher”, wo starrsinnige Versuchspersonen standhaft keine Antwort geben wollten. Die Zahl an Problemen und (Arten von) Fehlern übersteigt sicherlich die Anzahl der Datensätze. Kurz: Wir sehen uns gezwungen, den Daten einige Einblick abzurufen, und dafür müssen wir sie erst in Form bringen, was man als eine Mischung zwischen Artistik und Judo verstehen kann. Ach ja, die deskriptive Statistik fristet (in diesem Buch) eine untergeordnete Rolle

---

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



**Figure 1:** Rahmen

in diesem Schritt.

Dann die *Visualisierung*. Ein Bild sagt mehr als 1000 Worte, weiß der Volksmund. Für die Datenanalyse gilt dies auch. Ein gutes Diagramm vermittelt eine Fülle an Informationen “auf einen Blick” und erzielt damit eine Syntheseleistung, die digitalen Darbietungsformen, sprich: Zahlen, verwehrt bleibt. Nebenbei sind Diagramme, mit Geschick erstellt, ein Genuss für das Auge, daher kommt der Visualisierung großer Wert zu.

Als letzten, aber wesentlichen Punkt führen wir das *Modellieren* an. Es gibt mehr Definitionen von “Modell” als ich glauben wollte, aber hier ist damit gemeint, dass wir uns eine Geschichte ausdenken, wie die Daten entstanden sind, oder präziser gesagt: welcher Mechanismus hinter den Daten steht. So könnten wir Klausurnoten und Lernzeit von einigen Studenten<sup>2</sup> anschauen, und verkünden, wer mehr lerne, habe auch bessere Noten (ein typischer Dozentenauspruch). Unser Modell postuliert damit einen (vielleicht linearen) Anstieg des Klausurerfolgs bei steigender Vorbereitungszeit. Das schönste an solchen Modellen ist, dass wir Vorhersagen treffen können. Zum Beispiel: “Joachim, du hast 928 Stunden auf die Klausur gelernt; damit solltest du 93% der Punkte erzielen”.

Was ist dann mit dem *Reproduzierbarkeitshintergrund* gemeint? Ihre Arbeiten von Umformen, Visualisieren und Modellieren sollten sich nicht ausschließlich im Arbeitsspeicher Ihres Gehirns stattfinden, auch wenn das bei Ihnen, lieber Leser, vielleicht schneller ginge. Stattdessen soll der Mensch sich Mühe machen, seine Gedanken aufzuschreiben, hier insbesondere die Rechnungen bzw. alles, was den Daten angetan wurde, soll protokolliert werden (auch die Ergebnisse, aber wenn der Weg dorthin klar protokolliert ist, kann man die Ergebnisse ja einfach “nachkochen”). Ein Vorteil dieses Vorgehens ist, dass andere (inklusive Ihres zukünftigen Ich) die Ergebnisse bzw. das Vorgehen einfacher nachvollziehen können.

## 0.2 Was ist Statistik? Wozu ist sie gut?

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

Was ist Statistik? Eine Antwort dazu ist, dass Statistik die Wissenschaft von Sammlung, Analyse, Interpretation und Kommunikation mithilfe mathematischer Verfahren ist und zur Entscheidungshilfe beitragen sollte (??). Damit hätten wir auch den Unterschied zur schnöden Datenanalyse (ein Teil der Statistik) herausgemischt. Statistik wird häufig in die zwei Gebiete *deskriptive* und *inferierende* Statistik eingeteilt. Erstere fasst viele Zahlen zusammen, so dass wir den Wald statt vieler Bäume sehen. Letztere verallgemeinert von den vorliegenden (sog. “Stichproben-”)Daten auf eine zugrunde liegende Grundmenge (Population). Dabei spielt die Wahrscheinlichkeitsrechnung und Zufallsvariablen eine große Rolle.

Auch wenn die gerade genannte Diskussion die häufigste oder eine typische ist, mehren sich doch Stimmen, die Statistik anders akzentuieren. So schreibt Briggs in einem aktuellen Buch (?), dass es in der Statistik darum ginge, die Wahrscheinlichkeit zukünftiger Ereignisse vorherzusagen: “Wie wahrscheinlich ist es, dass - gegeben einem statistischen Modell, allerlei Annahmen und einem Haufen Daten - Kandidat X der neue Präsident wird”<sup>3</sup>? Das schöne an dieser Idee ist, dass das “Endprodukt” etwas sehr Weltliches und damit praktisches ist: Die Wahrscheinlichkeit einer interessanten (und unbekannten) Aussage. Nebenbei ruht diese Idee auf dem sicheren Fundament der Wahrscheinlichkeitstheorie.

Abgesehen von philosophischen Überlegungen zum Wesen der Statistik kann man sagen, dass Vorhersagen von Ereignissen etwas sehr praktisches sind. Sie nehmen daher aus praktischen Überlegungen einen zentralen Platz in diesem Buch an. Die philosophische Relevanz des prädiktiven Ansatzes ist gut bei Briggs (??) nachzulesen.

Traditionell ist die Statistik stark daran interessiert, Parameter von Populationen vorherzusagen. Ein Beispiel dazu wäre die mittlere Größe (Parameter) aller Deutschen (Population). Leider sind Populationen häufig ziemlich abstrakt. Nehmen wir als Beispiel an, ein Dozent der FOM (Prof. S.) wie sich der Lernerfolg

<sup>2</sup><https://www.rstudio.com/resources/cheatsheets/>

<sup>3</sup>zumindest bei den meisten Befehlen.



ändert, wenn die Stoffmenge pro Stunde verdoppelt. Zu seiner Überraschung ist der Lernerfolg geringer als in einem Kontrollkurs. Auf welche Population ist jetzt die Studie bzw. die Daten seiner Stichprobe zu verallgemeinern? Alle Menschen? Alle Studierenden? Alle deutschen Studierenden? Alle Studierenden der FOM? Alle Studierenden aller Zeiten?

- Statistik meint Methoden, die das Ziel haben, Ereignisse präzise vorherzusagen
- Statistik soll sich um Dinge dieser Welt drehen, nicht um Parameter
- Statt einer Frage “ist  $\mu_1$  größer als  $\mu_2$ ?” besser “Wie viel Umsatz erwarte ich von diesem Kunden?”, “Wie viele Saitensprünge hatte er wohl?”, “Wie groß ist die Wahrscheinlichkeit für sie zu überleben?” und dergleichen.
- Der Nutzen von Vorhersagen liegt auf der Hand: Vorhersagen sind praktisch; eine nützliche Angelegenheit (wenn auch schwierig).

## 0.3 Aufbau des Buches

sdkljf

## 0.4 Datensätze

Name des Datensatzes | Quelle | Beschreibung `profiles` | {okcupiddata} | Daten von einer Online-Singlebörse `Wage` | {ISLR} | Gehaltsdaten von US-amerikanischen Männern `inf_test_short` | [https://sebastiansauer.github.io/data/test\\_inf\\_short.csv](https://sebastiansauer.github.io/data/test_inf_short.csv) | Ergebnisse einer Statistiklausur



# Grundlagen

In diesem Kapitel diskutieren wir einige zentrale Begriffe der Wissenschaft bzw. der quantitativen Methodik der Wissenschaft.

## 0.5 Wahrscheinlichkeit

Was ist Wahrscheinlichkeit und (warum) ist sie wichtig? Wo wir schon bei den großen Fragen sind, können wir noch eins drauf setzen: Was ist das Ziel von Wissenschaft? Eine einfache Antwort auf diese Frage ist, die Wahrheit von Aussagen zu bestimmen. Zum Beispiel: “Ein Proton besteht aus siebenundzwanzig Dscharbs” oder “Deutsche Frauen verdienen im Schnitt weniger als Männer” oder “Morgen wird es regnen”. Leider ist es oft nicht möglich, sichere Aussagen über die Natur zu bekommen. In der Logik oder Mathe ist dies einfacher: “Joachim Z. ist eine Mensch und alle Menschen sind sterblich” (A) erlaubt die Ableitung “Joachim Z. ist sterblich” (B). Wir haben soeben eine wahre Aussage abgeleitet. Die Aussage ist sicher wahr, also zu 100%. Die Verneinung dieser Aussage B ist sicher falsch; wir sind uns zu 100% sicher, dass die Verneinung von falsch ist.

Aussagen wie die vom Regen morgen sind nicht sicher, wir sind nicht zu 100% gewiss, dass es morgen regnet. Der kühnste Wetterfrosch auch nicht. Genauso gilt, dass wir nicht zu 0% sicher sind; dies hieße, dass das Regenteil sicher ist. Wir brauchen also eine Methode, *Ungewissheit* auszudrücken. Das ist die Aufgabe der Wahrscheinlichkeit. Sie erlaubt Gewissheitsgrade zwischen 0% und 100%, etwa “ $P(\text{Regen morgen} \mid \text{Daten, Modell, Randbedingungen}) = 84\%$ ”. In Worten: “Die Wahrscheinlichkeit, dass es morgen in einem spezifizierten Gebiet regnet, gegeben meine Daten, mein Modell und sonstige Randbedingungen, die man leicht vergisst, die aber auch wichtig sind, liegt bei 84%”.

Wahrscheinlichkeit ist ein *epistemologisches* Konzept. Wahrscheinlichkeit beschreibt keine (physikalischen) Tatsachen über diese Tatsachen. Wahrscheinlichkeit beschreibt unsere Ungewissheit über Behauptungen.

Ein weiteres Beispiel:  $P(\text{Kopf} \mid \text{Wurf einer fairen Münze}) = 1/2$ . In Worten: “Die Wahrscheinlichkeit, Kopf zu werfen, wenn man eine faire Münze hat, liegt bei 1/2”. Genauer gesagt und etwas pedantisch, müsste man hinzufügen, dass wir stillschweigend vorausgesetzt haben, das Ergebnis des Wurfes nicht zu kennen. Denn:  $P(\text{Kopf} \mid \text{Ich weiß, dass es Kopf ist}) = 1$ . Das Beispiel zeigt, dass die Wahrscheinlichkeit eines Ereignisses von den Prämissen abhängt (das, was nach dem Strich steht). Diese Prämissen können bei Ihnen anders sein als bei; daher ist es plausibel, dass sich unsere Wahrscheinlichkeiten für  $P(\text{Kopf})$  unterscheiden. Das ist rational (nicht subjektiv).

Dieses auf der formalen Logik basierende Konzept von Wahrscheinlichkeit (?) besticht mit einer breiten Anwendungsfeld. Was halten Sie von dieser Aussage: “Wenn Hillary Clinton keine EMail-Affäre gehabt hätte, hätte sie die Wahl gewonnen”? Ist es eine sinnvolle Aussage? Es mag schwer oder unmöglich sein, an diese Aussage eine konkrete Wahrscheinlichkeit anzuheften; das ändert aber nichts daran, dass wir über eine solche Frage nachdenken können (ich glaube, viele Leute haben sich diese Frage gestellt). Fragen dieser Art könnten wir als “Rum-ums-Eck-Fragen” bezeichnen. Ein deutscher Politiker sprach in dem Zusammenhang

von “Hätte-Hätte-Fahrradkette”, was zeigt, dass Aussagen *kein empirischer Gehalt* zukommen muss, die Wahrscheinlichkeit dieser Aussagen für Menschen aber von Belang ist.

Definiert man die Wahrscheinlichkeit als relative Häufigkeit, kommt man natürlich bei solchen Rum-ums-Eck-Fragen in die Bredouille, da sie nicht häufig, nämlich überhaupt nicht passiert sind. Die logische Definition hat aber kein Problem mit diesen Fragen an sich.

Oder betrachten Sie dieses Beispiel (angelehnt an ?): “George ist ein Marsianer;  $1/3$  aller Marsianer lieben französischen Weichkäse (speziell Camembert)”. Wie hoch ist die Wahrscheinlichkeit, dass George, ein Marsianer, Weichkäse (speziell Camembert) liebt? Beachten Sie, dass die zu Verfügung stehende Information beachtet werden soll aber sonst keine Information. Die Antwort lautet:  $P = 1/3$ .

Ein weiteres Problem mit Wahrscheinlichkeit, die auf eine unendliche Wiederholung eines Ereignisses aufgebaut ist, ist dass Unendlich kompliziert ist. Es ist unendlich schwierig, sich unendlich viele Dinge vorzustellen oder sich ein unendlich großes Ding vorzustellen. Wenn wir uns aber nicht vorstellen können, was mit einer Aussage gemeint ist, was sagt dann diese Aussage?

Daher ist besser, Wahrscheinlichkeiten als Erweiterung der (formalen) Logik zu begreifen. Wo die Logik sagt, eine Aussage sei richtig oder falsch, gibt die Wahrscheinlichkeit eine Gradierung zwischen diesen beiden Extremen an.

Kurz gesagt: Wahrscheinlichkeit misst den Grad der Gewissheit (oder Ungewissheit) einer Aussage. Eine Aussage mit einer Wahrscheinlichkeit von 100% ist eine sichere Aussage, eine Aussage, die sicher zutrifft (wahr ist); analog ist eine Aussage mit einer Wahrscheinlichkeit von 0% sicher falsch (nicht zutreffend). Die Grade dazwischen markieren die unterschiedlichen Abstufungen von Gewissheit.

## 0.6 Hypothesen

Hypothesen sind Aussagen. Aussagen, bei denen wir nicht sicher sind, dass sie richtig oder falsch (d.h.  $P=0$  oder  $P=1$ ). Man beachte, dass der letzte Satz epistemologisch argumentiert hat (es war eine Aussage über unser Wissen); es war keine Aussage über Tatsachen WIRKLICH???

deren Wahrheitswert nicht extrem ist - die Wahrscheinlichkeit der Richtigkeit der Behauptung ist also größer als 0 aber kleiner als 1.

Aussagen sind Sätze, deren Wahrheitswert überprüfbar ist, zumindest potenziell. Beispiele für solche Aussagen wären “Webseiten mit Bildern sind einfacher zu lesen”, “Power Posing hat einen Effekt auf den Testosteronlevel” und “Es gibt Leben auf den Mars”. Das letzte Beispiel ist interessant, weil es im Moment vielleicht noch nicht im Vermögen der Forschung liegt, diesen Satz zu bestätigen oder zu widerlegen (falsifizieren). Überhaupt sind Sätze der Art “Es gibt...” schwierig zu widerlegen (manchmal geht es). Fruchtbarer sind daher Aussagen mit mehr empirischen Gehalt, die “angreifbarer” weil “gewagter” sind.

Hypothesen haben demnach den Charakter von Wahrscheinlichkeitsaussagen.

## 0.7 Falsifikationismus

Hm.

# Trends

- Big Data
- Open Science
- Computerisierung
- Neue Methoden zur numerischen Vorhersage
- Textmining



# Unbehagen

In diesem Kapitel finden sich einige Probleme, die einigen Wissenschaftlern Bauchschmerzen oder Unbehagen verursacht.

## 0.8 Der p-Wert

Der p-Wert ist die heilige Kuh der Forschung. Das ist nicht normativ, sondern deskriptiv gemeint. Der p-Wert entscheidet (häufig) darüber, was publiziert wird, und damit, was als Wissenschaft sichtbar ist - und damit, was Wissenschaft ist (wiederum deskriptiv, nicht normativ gemeint). Kurz: Dem p-Wert wird viel Bedeutung zugemessen.

Allerdings hat der p-Wert seine Probleme. Vor allem: Er wird missverstanden. Jetzt kann man sagen, dass es dem p-Wert (dem armen) nicht anzulasten, dass andere/ einige ihm missverstehen. Auf der anderen Seite finde ich, dass sich Technologien dem Nutzer anpassen sollten (soweit als möglich) und nicht umgekehrt. Die Definition des p-Werts ist aber auch so kompliziert, man kann sie leicht missverstehen:

Der p-Wert gibt die Wahrscheinlichkeit  $P$  unserer Daten  $D$  an (und noch extremerer), unter der Annahme, dass die getestete Hypothese  $H$  wahr ist (und wenn wir den Versuch unendlich oft wiederholen würden, unter identischen Bedingungen und ansonsten zufällig).  $p = P(D|H)$

Viele Menschen - inkl. Professoren und Statistik-Dozenten - haben Probleme mit dieser Definition (?). Das ist nicht deren Schuld: Die Definition ist kompliziert. Vielleicht denken viele, der p-Wert sage das, was tatsächlich interessant ist: die Wahrscheinlichkeit der (getesteten) Hypothese, gegeben der Tatsache, dass bestimmte Daten vorliegen. Leider ist das *nicht* die Definition des p-Werts. Also:

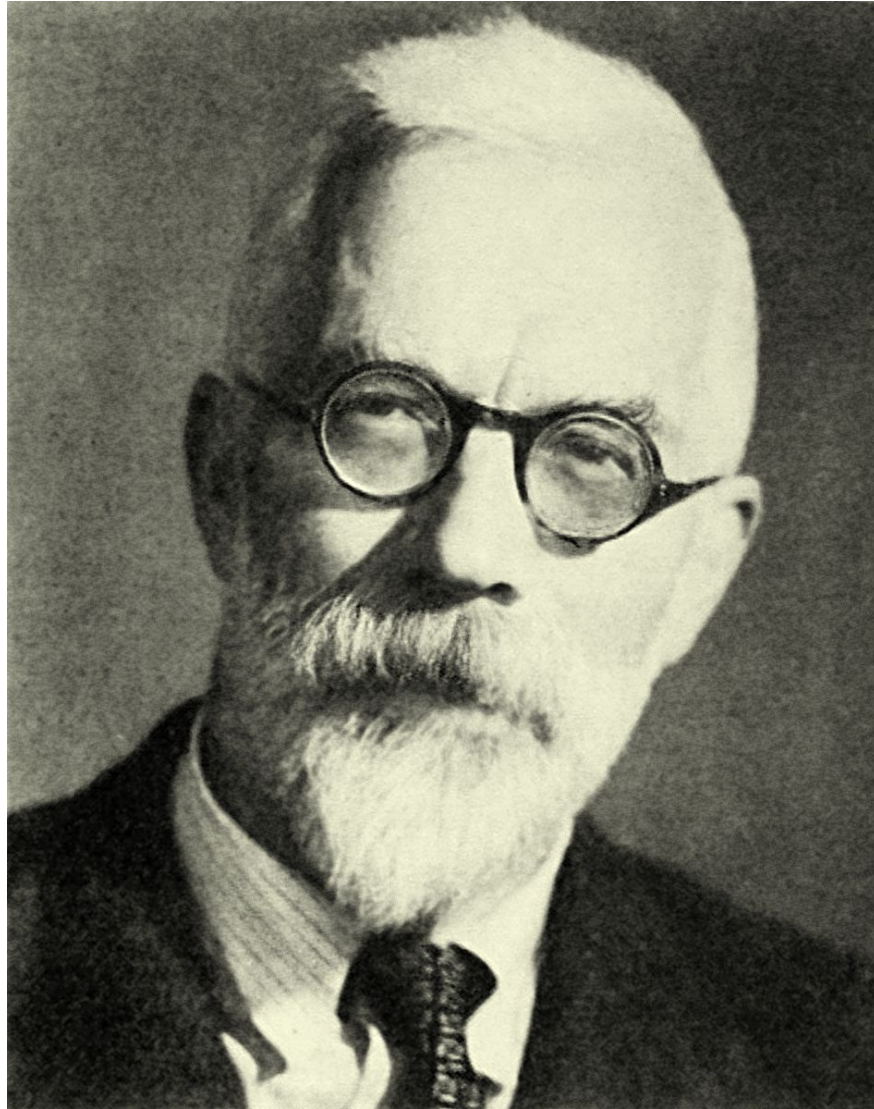
$$P(D|H) \neq P(H|D)$$

Der p-Wert ist für weitere Dinge kritisiert worden (?, ?); z.B. dass die "5%-Hürde" einen zu schwachen Test für die getestete Hypothese bedeutet. Letzterer Kritikpunkt ist aber nicht dem p-Wert anzulasten, denn dieses Kriterium ist beliebig, könnte konservativer gesetzt werden und jegliche mechanisierte Entscheidungsmethode kann ausgenutzt werden. Ähnliches kann man zum Thema "P-Hacking" argumentieren (?, ?); andere statistische Verfahren können auch gehackt werden.

Meine Meinung ist, dass der p-Wert problematisch ist und nicht oder weniger benutzt werden sollte (das ist eine normative Aussage). Da der p-Wert aber immer noch der Platzhirsch auf vielen Forschungsauen ist, führt kein Weg um ihn herum. Er muss genau verstanden werden: Was er sagt und - wichtiger noch - was er nicht sagt.

## 0.9 Wahrscheinlichkeit im Frequentismus

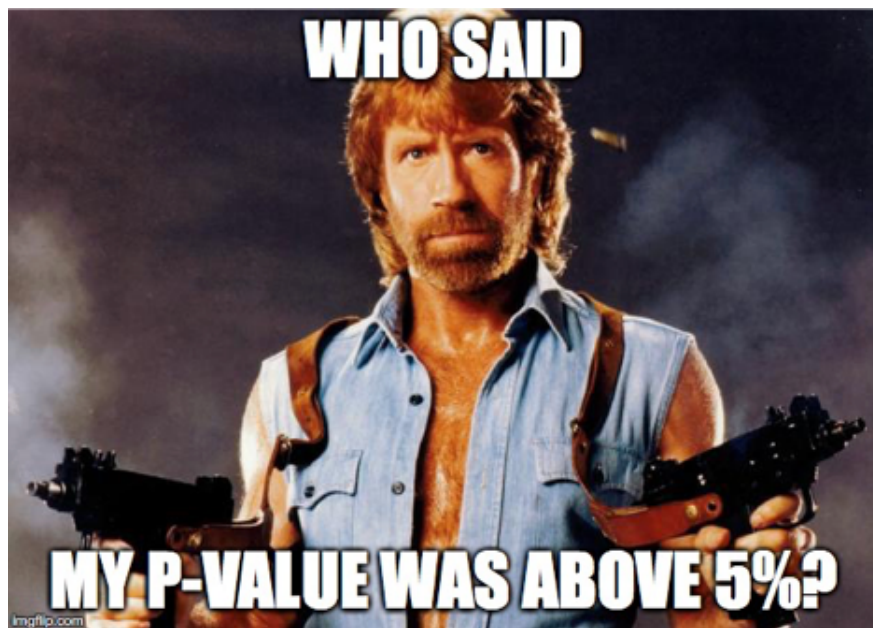
Die Idee von



**Figure 2:** Der größte Statistiker des 20. Jahrhunderts ( $p < .05$ )

- Theorie der Wahrscheinlichkeit im Frequentismus
- Reproduzierbarkeitskrise
- Parameter
- Kausalität
- Übersicherheit





**Figure 3:** Der p-Wert wird oft als wichtig erachtet



# Software

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

## 0.10 R and Friends installieren

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

## 0.11 Hilfe! R tut nicht so wie ich das will

Manntje, Manntje, Timpe Te, Buttje, Buttje inne See, myne Fru de Ilsebill will nich so, as ik wol will. Gebrüder Grimm, Märchen vom Fischer und seiner Frau, [https://de.wikipedia.org/wiki/Vom\\_Fischer\\_und\\_seiner\\_Frau](https://de.wikipedia.org/wiki/Vom_Fischer_und_seiner_Frau)

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

1. Schließen Sie die Augen für eine Minute. Denken Sie gut nach, woran es liegen könnte.
2. Schalten Sie den Rechner aus und probieren Sie es morgen noch einmal.
3. Googeln.

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

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

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

- Sehen Sie eine Fehlermeldung, die von einem fehlenden Paket spricht (z.B. “Package ‘Rcpp’ not available”) oder davon spricht, dass ein Paket nicht installiert werden konnte (z.B. “Package ‘Rcpp’ could

not be installed” oder “es gibt kein Paket namens ‘Rcpp’ ” oder “unable to move temporary installation XXX to YYY”), dann tun Sie folgendes:

- Schließen Sie R und starten Sie es neu.
  - Installieren Sie das oder die angesprochenen Pakete mit `install.packages("name_des_pakets", dependencies = TRUE)` oder mit dem entsprechenden Klick in RStudio.
  - Starten Sie das entsprechende Paket mit `library(paket_name)`.
- Gerade bei Windows 10 scheinen die Schreibrechte für R (und damit RStudio oder RCommander) eingeschränkt zu sein. Ohne Schreibrechte kann R aber nicht die Pakete (“packages”) installieren, die Sie für bestimmte R-Funktionen benötigen. Daher schließen Sie R bzw. RStudio und suchen Sie das Icon von R oder wenn Sie RStudio verwenden von RStudio. Rechtsklicken Sie das Icon und wählen Sie “als Administrator ausführen”. Damit geben Sie dem Programm Schreibrechte. Jetzt können Sie etwaige fehlende Pakete installieren.
  - Ein weiterer Grund, warum R bzw. RStudio die Schreibrechte verwehrt werden könnten (und damit die Installation von Paketen), ist ein Virenschanner. Der Virenschanner sagt, nicht ganz zu Unrecht: “Moment, einfach hier Software zu installieren, das geht nicht, zu gefährlich”. Grundsätzlich gut, in diesem Fall unnötig. Schließen Sie R/RStudio und schalten Sie dann den Virenschanner *komplett* (!) aus. Öffnen Sie dann R/RStudio wieder und versuchen Sie fehlende Pakete zu installieren.
  - Läuft der RCommander unter Mac nicht, dann prüfen Sie, ob Sie X11 (synonym: XQuartz) installiert haben. X11 muss installiert sein, damit der RCommander unter Mac läuft.
  - Die “app nap” Funktion beim Mac kann den RCommander empfindlich ausbremsen. Schalten Sie diese Funktion aus z.B. im RCommander über Tools - Manage Mac OS X app nap for R.app.

## 0.12 Allgemeine Hinweise zur Denk- und Gefühlswelt von R

- Wenn Sie RStudio starten, startet R automatisch auch. Starten Sie daher, wenn Sie RStudio gestartet haben, *nicht* noch extra R. Damit hätten Sie sonst zwei Instanzen von R laufen, was zu Verwirrungen (bei R und beim Nutzer) führen kann.
- Ein neues R-Skript im RStudio können Sie z.B. öffnen mit **File-New File-R Script**.
- R-Skripte können Sie speichern (**File-Save**) und öffnen.
- R-Skripte sind einfache Textdateien, die jeder Texteditor verarbeiten kann. Nur statt der Endung `txt`, sind R-Skripte stolzer Träger der Endung `R`. Es bleibt aber eine schnöde Textdatei.
- Bei der Installation von Paketen mit `install.packages("name_des_pakets")` sollte stets der Parameter `dependencies = TRUE` angefügt werden. Also `install.packages("name_des_pakets", dependencies = TRUE)`. Hintergrund ist: Falls das zu installierende Paket seinerseits Pakete benötigt, die noch nicht installiert sind (gut möglich), dann werden diese sog. “dependencies” gleich mitinstalliert (wenn Sie `dependencies = TRUE` setzen).
- Hier finden Sie weitere Hinweise zur Installation des RCommanders: <http://socserv.socsci.mcmaster.ca/jfox/Misc/Rcmdr/installation-notes.html>.
- Sie müssen online sein, um Packages zu installieren.
- Die “app nap” Funktion beim Mac kann den RCommander empfindlich ausbremsen. Schalten Sie diese Funktion aus z.B. im RCommander über Tools - Manage Mac OS X app nap for R.app.

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

R zu lernen kann hart sein. Ich weiß, wovon ich spreche. Wahrscheinlich eine spirituelle Prüfung in Geduld und Hartnäckigkeit... Tolle Gelegenheit, sich in diesen Tugenden zu trainieren :-)

## 0.13 dplyr und andere Pakete installieren

Ein R-Paket, welches für die praktische Datenanalyse praktisch ist, heißt **dplyr**. Wir werden viel mit diesem Paket arbeiten. Bitte installieren Sie es schon einmal, sofern noch nicht geschehen:

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

Übrigens, das `dependencies = TRUE` sagt sinngemäß “Wenn das Funktionieren von dplyr noch von anderen Paketen abhängig ist (es also Abhängigkeiten (dependencies) gibt), dann installiere die gleich mal mit”.

Nicht vergessen: Installieren muss man eine Software *nur einmal*; *starten* muss man sie jedes Mal, wenn man sie vorher geschlossen hat und wieder nutzen möchte:

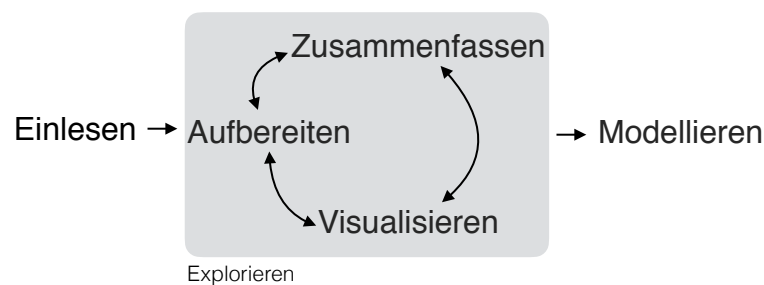
```
library(dplyr)
```

Das Installieren und Starten anderer Pakete läuft genauso ab.



# Daten explorieren

Den Ablauf des Datenexplorierens kann man so darstellen:



Zuerst müssen die Daten für die Analyse(software) verfügbar gemacht werden. Sprich, die Daten müssen *eingelezen* werden. Dann beginnt das eigentliche Explorieren; dieses kann man wiederum in drei Schritte einteilen, die keine Abfolge sind, sondern sich wild abwechseln können. Diese sind: Daten *aufbereiten*, Daten *zusammenfassen* und Daten *visualisieren*.

Unter Daten aufbereiten im engeren Sinne ist gemeint, die Daten einer “Grundreinigung” zu unterziehen, dass sie für weitere Analysen in geeigneter Form sind. Daten zusammenfassen meint die deskriptive Statistik; Daten visualisieren ist das Erstellen von Diagrammen. Im Anschluss kann man die Daten modellieren.

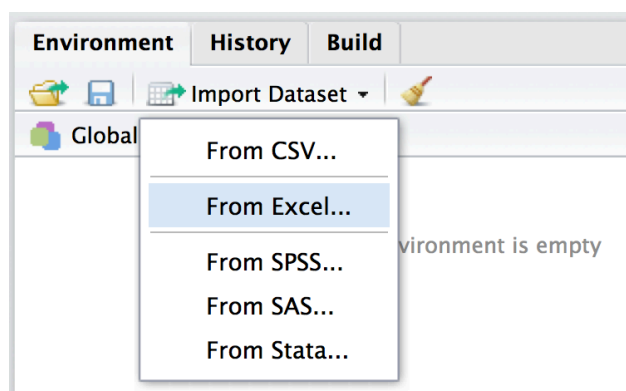
Ist das Explorieren von Daten auch nicht statistisch anspruchsvoll, so ist es trotzdem von großer Bedeutung und häufig recht zeitintensiv, vor allem das Daten aufbereiten. Eine Anekdote zur Relevanz der Exploration, die (so will es die Geschichte) mir an einer Bar nach einer einschlägigen Konferenz erzählt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

überraschen, dass man dann höhere Krebsraten vorhersagen kann. Eigentlich ganz einfach; besondere Mathe steht hier (zumindest in dieser Geschichte) nicht dahinter. Und, wenn man den Trick kennt, ganz einfach. Aber wie so oft ist es nicht leicht, den Trick zu finden. Sorgfältiges Datenjudo hat hier den Schlüssel zum Erfolg gebracht.

## 0.14 Daten einlesen

In R kann man ohne Weiteres verschiedene, gebräuchliche (Excel) oder weniger gebräuchliche (Feather<sup>4</sup>) Datenformate einlesen. In RStudio lässt sich dies z.B. durch einen schnellen Klick auf **Import Dataset** im Reiter **Environment** erledigen. Dabei wird im Hintergrund das Paket **readr** verwendet (?) (die entsprechende Syntax wird in der Konsole ausgegeben, so dass man sie sich anschauen und weiterverwenden kann).

Am einfachsten ist es, eine Excel-Datei über die RStudio-Oberfläche zu importieren; das ist mit ein paar Klicks geschehen:



Es ist für bestimmte Zwecke sinnvoll, nicht zu klicken, sondern die Syntax einzutippen. Zum Beispiel: Wenn Sie die komplette Analyse als Syntax in einer Datei haben (eine sog. “Skriptdatei”), dann brauchen Sie (in RStudio) nur alles auszuwählen und auf **Run** zu klicken, und die komplette Analyse läuft durch! Die Erfahrung zeigt, dass das ein praktisches Vorgehen ist.

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

```
"ID","time","sex","height","shoe_size"
"1","04.10.2016 17:58:51",NA,160.1,40
"2","04.10.2016 17:58:59","woman",171.2,39
"3","04.10.2016 18:00:15","woman",174.2,39
"4","04.10.2016 18:01:17","woman",176.4,40
"5","04.10.2016 18:01:22","man",195.2,46
```

Erkennen Sie das Muster? Die erste Zeile gibt die “Spaltenköpfe” wieder, also die Namen der Variablen. Hier sind es 5 Spalten; die vierte heißt “shoe\_size”. Die Spalten sind offenbar durch Komma , voneinander getrennt. Dezimalstellen sind in amerikanischer Manier mit einem Punkt . dargestellt. Die Daten sind “rechteckig”; alle Spalten haben gleich viele Zeilen und umgekehrt alle Spalten gleich viele Zeilen. Man kann sich diese Tabelle gut als Excel-Tabelle mit Zellen vorstellen, in denen z.B. “ID” (Zelle oben links) oder “46” (Zelle unten rechts) steht.

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

<sup>4</sup><https://cran.r-project.org/web/packages/feather/index.html>



Lesen wir diese Daten jetzt ein:

```
if (!file.exists("./data/wo_men.csv")){
  daten <- read.csv("https://sebastiansauer.github.io/data/wo_men.csv")
} else {
  daten <- read.csv("./data/wo_men.csv")
}
head(daten)
#>   X                time  sex height shoe_size
#> 1 1 04.10.2016 17:58:51 woman   160        40
#> 2 2 04.10.2016 17:58:59 woman   171        39
#> 3 3 04.10.2016 18:00:15 woman   174        39
#> 4 4 04.10.2016 18:01:17 woman   176        40
#> 5 5 04.10.2016 18:01:22  man   195        46
#> 6 6 04.10.2016 18:01:53 woman   157        37
```

Wir haben zuerst geprüft, ob die Datei (`wo_men.csv`) im entsprechenden Ordner existiert oder nicht (das `!`-Zeichen heißt auf Errisch “nicht”). Falls die Datei nicht im Ordner existiert, laden wir sie mit `read.csv` herunter und direkt ins R hinein. Andernfalls (`else`) lesen wir sie direkt ins R hinein.

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

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

Dabei steht `sep` (separator) für das Trennzeichen zwischen den Spalten und `dec` für das Dezimaltrennzeichen.

Übrigens: Wenn Sie keinen Pfad angeben, so geht R davon aus, dass die Daten im aktuellen Verzeichnis liegen. Das aktuelle Verzeichnis kann man mit `getwd()` erfragen und mit `setwd()` einstellen. Komfortabler ist es aber, das aktuelle Verzeichnis per Menü zu ändern. In RStudio: `Session > Set Working Directory > Choose Directory ...` (oder per Shortcut, der dort angezeigt wird).

## 0.15 Datenjudo (Daten aufbereiten)

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

Mit “Datenjudo” (ein Fachbegriff aus der östlichen Zahlentheorie) ist gemeint, die Daten so “umzuformen”, “aufzubereiten”, oder “reinigen”, dass sie passend für statistische Analysen sind.

Typische Probleme, die immer wieder auftreten sind:

- Fehlende Werte: Irgend jemand hat auf eine meiner schönen Fragen in der Umfrage nicht geantwortet!
- Unerwartete Daten: Auf die Frage, wie viele Facebook-Freunde er oder sie habe, schrieb die Person “I like you a lot”. Was tun???
- Daten müssen umgeformt werden: Für jede der beiden Gruppen seiner Studie hat Joachim einen Google-Forms-Fragebogen aufgesetzt. Jetzt hat er zwei Tabellen, die er “verheiraten” möchte. Geht das?

- Neue Spalten berechnen: Ein Student fragt nach der Anzahl der richtigen Aufgaben in der Statistik-Probeklausur. Wir wollen helfen und im entsprechenden Datensatz eine Spalte erzeugen, in der pro Person die Anzahl der richtig beantworteten Fragen steht.

### 0.15.1 Überblick

### 0.15.2 Normalform einer Tabelle

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

- Es handelt sich um einen data frame, also Spalten mit Namen und gleicher Länge; eine Datentabelle in rechteckiger Form
- In jeder Zeile steht eine Beobachtung, in jeder Spalte eine Variable
- Fehlende Werte sollten sich in *leeren* Tabellen niederschlagen
- Daten sollten nicht mit Farbkmarkierungen o.ä. kodiert werden
- keine Leerzeilen, keine Leerspalten
- am besten keine Sonderzeichen verwenden und keine Leerzeichen in Variablennamen und -werten, am besten nur Ziffern und Buchstaben und Unterstriche
- Variablennamen dürfen nicht mit einer Zahl beginnen

Der Punkt “Jede Zeile eine Beobachtung, jede Spalte eine Variable” verdient besondere Beachtung. Betrachten Sie dieses Beispiel:

```
knitr::include_graphics("../images/breit_lang.pdf")
```

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

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

In der rechten Tabelle sind die Variablen **Quartal** und **Umsatz** klar getrennt; jede hat ihre eigene Spalte. In der linken Tabelle hingegen sind die beiden Variablen vermischt. Sie haben nicht mehr ihre eigene Spalte, sondern sind über vier Spalten verteilt. Die rechte Tabelle ist ein Beispiel für eine Tabelle in Normalform, die linke nicht.

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

Ein Beispiel dazu:

```
meindf <- read.csv("http://stanford.edu/~ejdemyr/r-tutorials/data/unicef-u5mr.csv")

df_lang <- gather(meindf, year, u5mr, U5MR.1950:U5MR.2015)

df_lang <- separate(df_lang, year, into = c("U5MR", "year"), sep = ".")
```

- Die erste Zeile liest die Daten aus einer CSV-Datei ein; praktischerweise direkt von einer Webseite.
- Die zweite Zeile formt die Daten von breit nach lang um. Die neuen Spalten, nach der Umformung heißen dann `year` und `u5mr` (Sterblichkeit bei Kindern unter fünf Jahren). In die Umformung werden die Spalten `U5MR 1950` bis `U5MR 2015` einbezogen.
- Die dritte Zeile “entzerzt” die Werte der Spalte `year`; hier stehen die ehemaligen Spaltenköpfe. Man nennt sie auch `key` Spalte daher. Steht in einer Zelle von `year` bspw. `U5MR 1950`, so wird `U5MR` in eine Spalte mit Namen `U5MR` und `1950` in eine Spalte mit Namen `year` geschrieben.

Im Cheatsheet von RStudio zum Thema Datenimport finden sich nützliche Hinweise <sup>5</sup>.

### 0.15.3 Daten aufbereiten mit `dplyr`

Es gibt viele Möglichkeiten, Daten mit R aufzubereiten; `dplyr` ist ein populäres Paket dafür. Eine zentrale Idee von `dplyr` ist, dass es nur ein paar wenige Grundbausteine geben sollte, die sich gut kombinieren lassen. Sprich: Wenige grundlegende Funktionen mit eng umgrenzter Funktionalität. Der Autor, Hadley Wickham, sprach einmal in einem Forum (citation needed), dass diese Befehle wenig können, das Wenige aber gut. Ein Nachteil dieser Konzeption kann sein, dass man recht viele dieser Bausteine kombinieren muss, um zum gewünschten Ergebnis zu kommen. Außerdem muss man die Logik des Baukastens gut verstanden haben - die Lernkurve ist also erstmal steiler. Dafür ist man dann nicht darauf angewiesen, dass es irgendwo “Mrs Right” gibt, die genau das kann, so wie ich das will. Außerdem braucht man sich auch nicht viele Funktionen merken. Es reicht einen kleinen Satz an Funktionen zu kennen (die praktischerweise konsistent in Syntax und Methodik sind).

Willkommen in der Welt von `dplyr`! `dplyr` hat seinen Namen, weil es sich ausschließlich um *Dataframes* bemüht; es erwartet einen *Dataframe* als Eingabe und gibt einen *Dataframe* zurück<sup>6</sup>.

Diese Bausteine sind typische Tätigkeiten im Umgang mit Daten; nichts Überraschendes. Schauen wir uns diese Bausteine näher an.

```
library(dplyr) # muss installiert sein
```

#### 0.15.3.1 Zeilen filtern mit `filter`

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

<sup>5</sup><https://www.rstudio.com/resources/cheatsheets/>

<sup>6</sup>zumindest bei den meisten Befehlen.

Ein Sinnbild:

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

→

ID	Name	Note1
1	Anna	1
2	Anna	1

Merke: > Die Funktion `filter` filtert Zeilen aus einem Dataframe.

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

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

```
df_frauen <- filter(profiles, sex == "f") # nur die Frauen
df_alt <- filter(profiles, age > 70) # nur die alten
df_alte_frauen <- filter(profiles, age > 70, sex == "f") # nur die alten Frauen, d.h. UND-Verknüpfung
df_nosmoke_nodrinks <- filter(profiles, smokes == "no" | drinks == "not at all")
# liefert alle Personen, die Nicht-Raucher *oder* Nicht-Trinker sind
```

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



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

Einige fortgeschrittene Beispiele für `filter`:

Man kann alle Elemente (Zeilen) filtern, die zu einer Menge gehören und zwar mit diesem Operator: `%in%`:

```
filter(profiles, body_type %in% c("a little extra", "average"))
```

Besonders Textdaten laden zu einigen Extra-Überlegungen ein; sagen wir, wir wollen alle Personen filtern, die Katzen bei den Haustieren erwähnen. Es soll reichen, wenn `cat` ein Teil des Textes ist; also `likes dogs`

and likes cats wäre OK (soll gefiltert werden). Dazu nutzen wir ein Paket zur Bearbeitung von Strings (Textdaten):

```
library(stringr) # muss installiert sein
filter(profiles, str_detect(pets, "cats"))
```

Ein häufiger Fall ist, Zeilen *ohne* fehlende Werte (NAs) zu filtern. Das geht einfach:

```
profiles_keine_nas <- na.omit(profiles)
```

Aber was ist, wenn wir nur bei bestimmten Spalten wegen fehlender Werte besorgt sind? Sagen wir bei income und bei sex:

```
filter(profiles, !is.na(income) | !is.na(sex))
```

### 0.15.3.2 Spalten wählen mit select

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

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

Merke:

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

```
if (!file.exists("../data/test_inf_short.csv")) {
  stats_test <- read.csv("https://sebastiansauer.github.io/data/test_inf_short.csv")
} else {
  stats_test <- read.csv("../data/test_inf_short.csv")
}
```

Hier haben wir erst geprüft, ob die Datei `test_inf_short.csv` existiert; falls nein, laden wir sie herunter. Andernfalls lesen wir sie aus dem lokalen Verzeichnis.

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

Tatsächlich ist der Befehl `select` sehr flexibel; es gibt viele Möglichkeiten, Spalten auszuwählen. Im `dplyr`-Cheatsheet findet sich ein guter Überblick dazu <sup>7</sup>.

### 0.15.3.3 Zeilen sortieren mit `arrange`

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

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

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

```
arrange(stats_test, score) %>% head() # liefert die *schlechtesten* Noten zurück
#>      X                V_1 study_time self_eval interest score
#> 1 234 23.01.2017 18:13:15      3         1         1      17
#> 2   4 06.01.2017 09:58:05      2         3         2      18
#> 3 131 19.01.2017 18:03:45      2         3         4      18
#> 4 142 19.01.2017 19:02:12      3         4         1      18
#> 5  35 12.01.2017 19:04:43      1         2         3      19
#> 6  71 15.01.2017 15:03:29      3         3         3      20

arrange(stats_test, -score) %>% head() # liefert die *besten* Noten zurück
#>      X                V_1 study_time self_eval interest score
#> 1   3 05.01.2017 23:33:47      5        10         6      40
#> 2   7 06.01.2017 14:25:49     NA         NA        NA      40
#> 3  29 12.01.2017 09:48:16      4        10         3      40
#> 4  41 13.01.2017 12:07:29      4        10         3      40
#> 5  58 14.01.2017 15:43:01      3         8         2      40
#> 6  83 16.01.2017 10:16:52     NA         NA        NA      40

arrange(stats_test, interest, score) %>% head()
#>      X                V_1 study_time self_eval interest score
#> 1 234 23.01.2017 18:13:15      3         1         1      17
#> 2 142 19.01.2017 19:02:12      3         4         1      18
#> 3 221 23.01.2017 11:40:30      1         1         1      23
```

<sup>7</sup><https://www.rstudio.com/resources/cheatsheets/>

```
#> 4 230 23.01.2017 16:27:49      1      1      1      23
#> 5  92 17.01.2017 17:18:55      1      1      1      24
#> 6 107 18.01.2017 16:01:36      3      2      1      24
```

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

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

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

Aber was heißt dieses komisch Symbol: `%>%`? Diese sogenannte “Pfeife” lässt sich mit “und dann” ins Deutsche übersetzen. Also:

`sortiere(diese_Tabelle, nach_dieser_Spalte) UND DANN zeig_die_ersten_Zeilen`

Der Befehl `head` zeigt die ersten paar Zeilen eines Dataframes <sup>8</sup>.

Merke:

Die Funktion `arrange` sortiert die Zeilen eines Dataframes.

Ein Sinnbild zur Verdeutlichung:

```
knitr::include_graphics("./images/arrange.pdf")
```

ID	Name	Note1		ID	Name	Note1
1	Anna	1		1	Anna	<b>1</b>
2	Anna	5		3	Berta	<b>2</b>
3	Berta	2		5	Carla	<b>3</b>
4	Carla	4		4	Carla	<b>4</b>
5	Carla	3		2	Anna	<b>5</b>

Gute Noten zuerst!

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

<sup>8</sup>In der Regel 10 Zeilen, wobei ich irgendwo versteckt gesagt habe, es sollen nur 6 Zeilen am Bildschirm gedruckt werden.

```

top_n(stats_test, 3)
#> Selecting by score
#>      X                V_1 study_time self_eval interest score
#> 1    3 05.01.2017 23:33:47         5        10         6    40
#> 2    7 06.01.2017 14:25:49        NA        NA        NA    40
#> 3   29 12.01.2017 09:48:16         4        10         3    40
#> 4   41 13.01.2017 12:07:29         4        10         3    40
#> 5   58 14.01.2017 15:43:01         3         8         2    40
#> 6   83 16.01.2017 10:16:52        NA        NA        NA    40
#> 7  116 18.01.2017 23:07:32         4         8         5    40
#> 8  119 19.01.2017 09:05:01        NA        NA        NA    40
#> 9  132 19.01.2017 18:22:32        NA        NA        NA    40
#> 10 175 20.01.2017 23:03:36         5        10         5    40
#> 11 179 21.01.2017 07:40:05         5         9         1    40
#> 12 185 21.01.2017 15:01:26         4        10         5    40
#> 13 196 22.01.2017 13:38:56         4        10         5    40
#> 14 197 22.01.2017 14:55:17         4        10         5    40
#> 15 248 24.01.2017 16:29:45         2        10         2    40
#> 16 249 24.01.2017 17:19:54        NA        NA        NA    40
#> 17 257 25.01.2017 10:44:34         2         9         3    40
#> 18 306 27.01.2017 11:29:48         4         9         3    40
top_n(stats_test, 3, interest)
#>      X                V_1 study_time self_eval interest score
#> 1    3 05.01.2017 23:33:47         5        10         6    40
#> 2    5 06.01.2017 14:13:08         4         8         6    34
#> 3   43 13.01.2017 14:14:16         4         8         6    36
#> 4   65 15.01.2017 12:41:27         3         6         6    22
#> 5  110 18.01.2017 18:53:02         5         8         6    37
#> 6  136 19.01.2017 18:22:57         3         1         6    39
#> 7  172 20.01.2017 20:42:46         5        10         6    34
#> 8  214 22.01.2017 21:57:36         2         6         6    31
#> 9  301 27.01.2017 08:17:59         4         8         6    33

```

Gibt man *keine* Spalte an, so bezieht sich `top_n` auf die letzte Spalte im Datensatz.

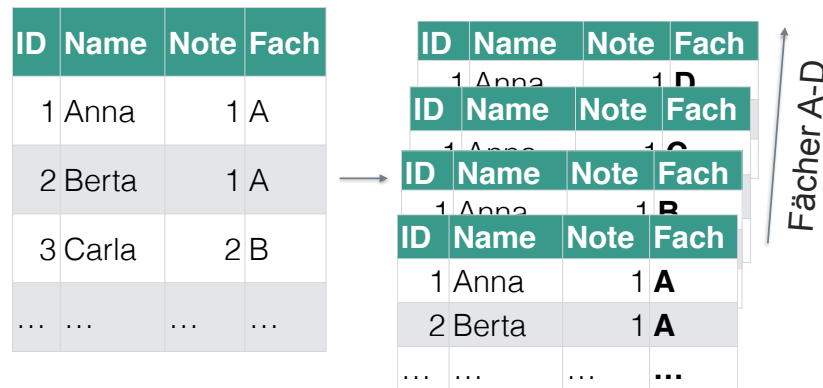
Da sich hier mehrere Personen den größten Rang (Wert 40) teilen, bekommen wir *nicht* 3 Zeilen zurückgeliefert, sondern entsprechend mehr.

#### 0.15.3.4 Datensatz gruppieren mit `group_by`

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

```
knitr::include_graphics("../images/group_by.pdf")
```





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

```
test_gruppiert <- group_by(stats_test, interest)
test_gruppiert
#> Source: local data frame [306 x 6]
#> Groups: interest [7]
#>
#>       X                V_1 study_time self_eval interest score
#>   <int>          <fctr>    <int>      <int>    <int> <int>
#> 1     1 05.01.2017 13:57:01         5         8         5    29
#> 2     2 05.01.2017 21:07:56         3         7         3    29
#> 3     3 05.01.2017 23:33:47         5        10         6    40
#> # ... with 303 more rows
```

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

Merke:

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

### 0.15.3.5 Eine Spalte zusammenfassen mit `summarise`

Vielleicht die wichtigste oder häufigste Tätigkeit in der Analyse von Daten ist es, eine Spalte zu *einem* Wert zusammenzufassen. Anders gesagt: Einen Mittelwert berechnen, den größten (kleinsten) Wert herausuchen, die Korrelation berechnen oder eine beliebige andere Statistik ausgeben lassen. Die Gemeinsamkeit dieser Operationen ist, dass sie eine Spalte zu einem Wert zusammenfassen, "aus Spalte mach Zahl", sozusagen. Daher ist der Name des Befehls `summarise` ganz passend. Genauer gesagt fasst dieser Befehl eine Spalte zu

einer Zahl zusammen *anhand* einer Funktion wie `mean` oder `max`. Hierbei ist jede Funktion erlaubt, die eine Spalte als Input verlangt und eine Zahl zurückgibt; andere Funktionen sind bei `summarise` nicht erlaubt.

```
knitr::include_graphics("images/summarise.pdf")
```



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

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

Jetzt können wir auch die Gruppierung nutzen:

```
test_gruppiert <- group_by(stats_test, interest)
summarise(test_gruppiert, mean(score))
#> # A tibble: 7 × 2
#>   interest `mean(score)`
#>   <int>     <dbl>
#> 1     1         28.3
#> 2     2         29.7
#> 3     3         30.8
#> # ... with 4 more rows
```

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

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

wir den wie folgt:

Jetzt können wir auch die Gruppierung nutzen:

```
test_gruppiert <- group_by(stats_test, interest)
summarise(test_gruppiert, mw_pro_gruppe = mean(score, na.rm = TRUE))
#> # A tibble: 7 × 2
#>   interest mw_pro_gruppe
#>   <int>     <dbl>
#> 1     1         28.3
#> 2     2         29.7
#> 3     3         30.8
#> # ... with 4 more rows
```

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

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

Merke:

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

### 0.15.3.6 Zeilen zählen mit `n` und `count`

Ebenfalls nützlich ist es, Zeilen zu zählen. Im Gegensatz zum Standardbefehle `nrow` versteht der `dplyr`-Befehl `nauch` Gruppierungen. `n` darf nur innerhalb von `summarise` oder ähnlichen `dplyr`-Befehlen verwendet werden.

```
summarise(stats_test, n())
#>   n()
#> 1 306
summarise(test_gruppiert, n())
#> # A tibble: 7 × 2
#>   interest `n()`
#>   <int> <int>
#> 1     1    30
#> 2     2    47
#> 3     3    66
#> # ... with 4 more rows
nrow(stats_test)
#> [1] 306
```

Außerhalb von gruppierten Datensätzen ist `nrow` meist praktischer.

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

```
dplyr::count(stats_test, interest)
#> # A tibble: 7 × 2
#>   interest     n
#>   <int>   <int>
```

```

#>      <int> <int>
#> 1         1    30
#> 2         2    47
#> 3         3    66
#> # ... with 4 more rows
dplyr::count(stats_test, study_time)
#> # A tibble: 6 × 2
#>   study_time     n
#>     <int> <int>
#> 1         1    31
#> 2         2    49
#> 3         3    85
#> 4         4    56
#> 5         5    17
#> 6        NA    68
dplyr::count(stats_test, interest, study_time)
#> Source: local data frame [29 x 3]
#> Groups: interest [?]
#>
#>   interest study_time     n
#>     <int>     <int> <int>
#> 1         1         1    12
#> 2         1         2     7
#> 3         1         3     8
#> # ... with 26 more rows

```

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

Merke:

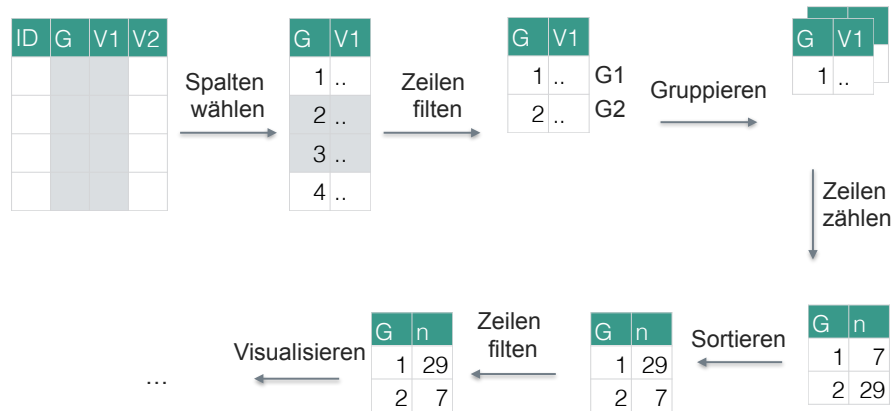
`n` und `count` zählen die Anzahl der Zeilen, d.h. die Anzahl der Fälle.

### 0.15.3.7 Die Pfeife

Die zweite Idee kann man salopp als “Durchpfeifen” bezeichnen; ikonographisch mit diesem Symbol dargestellt `%>%`<sup>9</sup>. Der Begriff “Durchpfeifen” ist frei vom Englischen “to pipe” übernommen. Hierbei ist gemeint, einen Datensatz sozusagen auf ein Fließband zu legen und an jedem Arbeitsplatz einen Arbeitsschritt auszuführen. Der springende Punkt ist, dass ein Dataframe als “Rohstoff” eingegeben wird und jeder Arbeitsschritt seinerseits wieder einen Dataframe ausgiebt. Damit kann man sehr schön, einen “Flow” an Verarbeitung erreichen, außerdem spart man sich Tipparbeit und die Syntax wird lesbarer. Damit das Durchpfeifen funktioniert, benötigt man Befehle, die als Eingabe einen Dataframe erwarten und wieder einen Dataframe zurückliefern. Das Schaubild verdeutlicht beispielhaft eine Abfolge des Durchpfeifens.

---

<sup>9</sup>Eine Art Smiley für Nerds.



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

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

mit dieser

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

Es ist hilfreich, diese “Pfeifen-Syntax” in deutschen Pseudo-Code zu übersetzen:

Nimm die Tabelle "stats\_test" UND DANN  
 filtere alle nicht-fehlenden Werte UND DANN  
 gruppieren die verbleibenden Werte nach "interest" UND DANN  
 bilde den Mittelwert (pro Gruppe) für "score" UND DANN  
 liefere nur die Werte größer als 30 zurück.

Die Pfeife zerlegt die “russische Puppe”, also ineinander verschachtelten Code, in sequenzielle Schritte und zwar in der richtigen Reihenfolge (entsprechend der Abarbeitung). Wir müssen den Code nicht mehr

von innen nach außen lesen (wie das bei einer mathematischen Formel der Fall ist), sondern können wie bei einem Kochrezept “erstens ..., zweitens .., drittens ...” lesen. Die Pfeife macht die Syntax einfacher. Natürlich hätten wir die verschachtelte Syntax in viele einzelne Befehle zerlegen können und jeweils eine Zwischenergebnis speichern mit dem Zuweisungspfeil `<-` und das Zwischenergebnis dann explizit an den nächsten Befehl weitergeben. Eigentlich macht die Pfeife genau das - nur mit weniger Tipparbeit. Und auch einfacher zu lesen. Flow!

### 0.15.3.8 Werte umkodieren und “binnen” mit `car::recode`

Manchmal möchte man z.B. negativ gepolte Items umdrehen oder bei kategoriellen Variablen kryptische Bezeichnungen in sprechendere umwandeln (ein Klassiker ist 1 in `maennlich` bzw. 2 in `weiblich` oder umgekehrt, kann sich niemand merken). Hier gibt es eine Reihe praktischer Befehle, z.B. `recode` aus dem Paket `car`. Übrigens: Wenn man explizit angeben möchte, aus welchem Paket ein Befehl stammt (z.B. um Verwechslungen zu vermeiden), gibt man `Paketnamen::Befehlnamen` an. Schauen wir uns ein paar Beispiele zum Umkodieren an.

```
library(car)

stats_test$score_fac <- car::recode(stats_test$study_time, "5 = 'sehr viel'; 2:4 = 'mittel'; 1 = 'wenig'
stats_test$score_fac <- car::recode(stats_test$study_time, "5 = 'sehr viel'; 2:4 = 'mittel'; 1 = 'wenig'

stats_test$study_time <- car::recode(stats_test$study_time, "5 = 'sehr viel'; 4 = 'wenig'; else = 'Hilfe'

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

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

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

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

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

Testen kann man den Erfolg des Umpolens mit

```
dplyr::count(stats_test, interest)
#> # A tibble: 7 × 2
#>   interest     n
#>   <int> <int>
#> 1       1    30
#> 2       2    47
#> 3       3    66
#> # ... with 4 more rows
dplyr::count(stats_test, no_interest)
```

```
#> # A tibble: 7 × 2
#>   no_interest      n
#>   <dbl> <int>
#> 1         1      9
#> 2         2     45
#> 3         3     41
#> # ... with 4 more rows
```

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

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

Natürlich gibt es auch eine Pfeifen kompatible Version, um Variablen umzukodieren bzw. zu binnen: `dplyr::recode`<sup>10</sup>. Die Syntax ist allerdings etwas weniger komfortabel (da strenger), so dass wir an dieser Stelle bei `car::recode` bleiben.

#### 0.15.4 Fallstudie nycflights13

Schauen wir uns einige Beispiele der Datenaufbereitung mittels `dplyr` an. Wir verwenden hier den Datensatz `flights` aus dem Package `nycflights13`. Der Datensatz ist recht groß (~300.000 Zeilen und 19 Spalten); wenn man ihn als Excel importiert, kann eine alte Möhre von Computer schon in die Knie gehen. Beim Import als CSV habe ich noch nie von Problemen gehört; beim Import via Package ebenfalls nicht. Werfen wir einen ersten Blick in die Daten:

```
#install.packages("nycflights13")
library(nycflights13)
data(flights)
glimpse(flights)
#> Observations: 336,776
#> Variables: 19
#> $ year      <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013,...
#> $ month     <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
#> $ day       <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
#> $ dep_time  <int> 517, 533, 542, 544, 554, 554, 555, 557, 557, 55...
#> $ sched_dep_time <int> 515, 529, 540, 545, 600, 558, 600, 600, 600, 60...
#> $ dep_delay <dbl> 2, 4, 2, -1, -6, -4, -5, -3, -3, -2, -2, -2, -2...
#> $ arr_time  <int> 830, 850, 923, 1004, 812, 740, 913, 709, 838, 7...
#> $ sched_arr_time <int> 819, 830, 850, 1022, 837, 728, 854, 723, 846, 7...
#> $ arr_delay <dbl> 11, 20, 33, -18, -25, 12, 19, -14, -8, 8, -2, -...
#> $ carrier   <chr> "UA", "UA", "AA", "B6", "DL", "UA", "B6", "EV",...
#> $ flight    <int> 1545, 1714, 1141, 725, 461, 1696, 507, 5708, 79...
#> $ tailnum   <chr> "N14228", "N24211", "N619AA", "N804JB", "N668DN...
#> $ origin    <chr> "EWR", "LGA", "JFK", "JFK", "LGA", "EWR", "EWR"...
#> $ dest      <chr> "IAH", "IAH", "MIA", "BQN", "ATL", "ORD", "FLL"...
#> $ air_time  <dbl> 227, 227, 160, 183, 116, 150, 158, 53, 140, 138...
#> $ distance  <dbl> 1400, 1416, 1089, 1576, 762, 719, 1065, 229, 94...
#> $ hour      <dbl> 5, 5, 5, 5, 6, 5, 6, 6, 6, 6, 6, 6, 6, 6, 5,...
```

<sup>10</sup><https://blog.rstudio.org/2016/06/27/dplyr-0-5-0/>.

```
#> $ minute      <dbl> 15, 29, 40, 45, 0, 58, 0, 0, 0, 0, 0, 0, 0, ...
#> $ time_hour    <dtm> 2013-01-01 05:00:00, 2013-01-01 05:00:00, 2013...
```

Der Befehl `data` lädt Daten aus einem zuvor gestarteten Paket.

Achtung, Fallstudie. Sie sind der/die Assistent\_in des Chefs der New Yorker Flughäfen. Ihr Chef kommt gut gelaunt ins Büro und sagt, dass er diesen Schnarchnasen einheizen wolle und sagt, sie sollen ihm mal schnell die Flüge mit der größten Verspätung raussuchen. Nix schickes, aber zacki-zacki...

```
flights %>%
  arrange(arr_delay)
#> # A tibble: 336,776 × 19
#>   year month   day dep_time sched_dep_time dep_delay arr_time
#>   <int> <int> <int>   <int>         <int>      <dbl>   <int>
#> 1  2013     5     7    1715           1729      -14    1944
#> 2  2013     5    20     719           735      -16     951
#> 3  2013     5     2    1947           1949       -2    2209
#> # ... with 3.368e+05 more rows, and 12 more variables:
#> #   sched_arr_time <int>, arr_delay <dbl>, carrier <chr>, flight <int>,
#> #   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>,
#> #   distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dtm>
```

Hm, übersichtlicher wäre es wahrscheinlich, wenn wir weniger Spalten anschauen müssten. Am besten neben der Verspätung nur die Information, die wir zur Identifizierung der Schuldigen... will sagen der gesuchten Flüge benötigen

```
flights %>%
  arrange(arr_delay) %>%
  select(arr_delay, carrier, month, day, dep_time, tailnum, flight, dest)
#> # A tibble: 336,776 × 8
#>   arr_delay carrier month   day dep_time tailnum flight dest
#>   <dbl>   <chr> <int> <int>   <int>   <chr> <int> <chr>
#> 1    -86     VX     5     7    1715  N843VA    193  SFO
#> 2    -79     VX     5    20     719  N840VA     11  SFO
#> 3    -75     UA     5     2    1947  N851UA    612  LAX
#> # ... with 3.368e+05 more rows
```

Da Zahlen in ihrer natürlichen Form von klein nach groß sortiert sind, sortiert `arrange` in ebendieser Richtung. Wir können das umdrehen mit einem Minuszeichen vor der zu sortierenden Spalte:

```
flights %>%
  arrange(-arr_delay) %>%
  select(arr_delay, carrier, month, day, dep_time, tailnum, flight, dest)
#> # A tibble: 336,776 × 8
#>   arr_delay carrier month   day dep_time tailnum flight dest
#>   <dbl>   <chr> <int> <int>   <int>   <chr> <int> <chr>
#> 1    1272     HA     1     9     641  N384HA     51  HNL
#> 2    1127     MQ     6    15    1432  N504MQ   3535  CMH
#> 3    1109     MQ     1    10    1121  N517MQ   3695  ORD
#> # ... with 3.368e+05 more rows
```

Eine kleine Zugabe: Mit dem Befehl `knitr::kable` kann man einen Dataframe automatisch in eine (einiger-



maßen) schöne Tabelle ausgeben lassen. Oh halt, wir wollen keine Tabelle mit 300.000 Zeilen (der Chef ist kein Freund von Details). Also begrenzen wir die Ausgabe auf die ersten 10 Plätze.

```
library(knitr) # muss installiert sein

flights %>%
  arrange(-arr_delay) %>%
  select(arr_delay, carrier, month, day, dep_time, tailnum, flight, dest) %>%
  filter(row_number() < 11) %>%
  kable()
```

arr_delay	carrier	month	day	dep_time	tailnum	flight	dest
1272	HA	1	9	641	N384HA	51	HNL
1127	MQ	6	15	1432	N504MQ	3535	CMH
1109	MQ	1	10	1121	N517MQ	3695	ORD
1007	AA	9	20	1139	N338AA	177	SFO
989	MQ	7	22	845	N665MQ	3075	CVG
931	DL	4	10	1100	N959DL	2391	TPA
915	DL	3	17	2321	N927DA	2119	MSP
895	DL	7	22	2257	N6716C	2047	ATL
878	AA	12	5	756	N5DMAA	172	MIA
875	MQ	5	3	1133	N523MQ	3744	ORD

“Geht doch”, war die Antwort des Chefs, als sie die Tabelle rübergeben (er mag auch keine Emails). “Ach ja”, raunt der Chef, als Sie das Zimmer verlassen wollen, “hatte ich erwähnt, dass ich die gleiche Auswertung für jeden Carrier brauche? Reicht bis in einer halben Stunde”.

Wir gruppieren also den Datensatz nach der Fluggesellschaft (**carrier**) und filtern dann die ersten 3 Zeilen (damit die Tabelle für den Chef nicht zu groß wird). Wie jeder **dplyr**-Befehl wird die vorherige Gruppierung berücksichtigt und daher die Top-3-Zeilen *pro Gruppe*, d.h. pro Fluggesellschaft, ausgegeben.

```
flights %>%
  arrange(-arr_delay) %>%
  select(arr_delay, carrier, month, day, dep_time, tailnum, flight, dest) %>%
  group_by(carrier) %>%
  filter(row_number() < 4)
#> Source: local data frame [48 x 8]
#> Groups: carrier [16]
#>
#>   arr_delay carrier month   day dep_time tailnum flight  dest
#>   <dbl>    <chr> <int> <int>   <int>   <chr>   <int> <chr>
#> 1    1272      HA     1     9     641   N384HA     51   HNL
#> 2    1127      MQ     6    15    1432   N504MQ    3535   CMH
#> 3    1109      MQ     1    10    1121   N517MQ    3695   ORD
#> # ... with 45 more rows
```

Vielleicht gefällt dem Chef diese Darstellung (sortiert nach **carrier**) besser:

```
flights %>%
  arrange(-arr_delay) %>%
  select(arr_delay, carrier, month, day, dep_time, tailnum, flight, dest) %>%
  group_by(carrier) %>%
```

```

filter(row_number() < 4) %>%
  arrange(carrier)
#> Source: local data frame [48 x 8]
#> Groups: carrier [16]
#>
#>   arr_delay carrier month   day dep_time tailnum flight dest
#>   <dbl>    <chr> <int> <int>   <int>   <chr>   <int> <chr>
#> 1      744      9E     2    16      757  N8940E   3798  CLT
#> 2      458      9E     7    24     1525  N927XJ   3538  MSP
#> 3      421      9E     7    10     2054  N937XJ   3325  DFW
#> # ... with 45 more rows

```

Da Sie den Chef gut kennen, berechnen Sie gleich noch die durchschnittliche Verspätung pro Fluggesellschaft.

```

flights %>%
  select(arr_delay, carrier, month, day, dep_time, tailnum, flight, dest) %>%
  group_by(carrier) %>%
  summarise(delay_mean = mean(arr_delay, na.rm = TRUE)) %>%
  arrange(-delay_mean) %>%
  kable()

```

carrier	delay_mean
F9	21.921
FL	20.116
EV	15.796
YV	15.557
OO	11.931
MQ	10.775
WN	9.649
B6	9.458
9E	7.380
UA	3.558
US	2.130
VX	1.764
DL	1.644
AA	0.364
HA	-6.915
AS	-9.931

Der Chef ist zufrieden. Sie können sich wieder wichtigeren Aufgaben zuwenden...

### 0.15.5 Weiterführende Hinweise

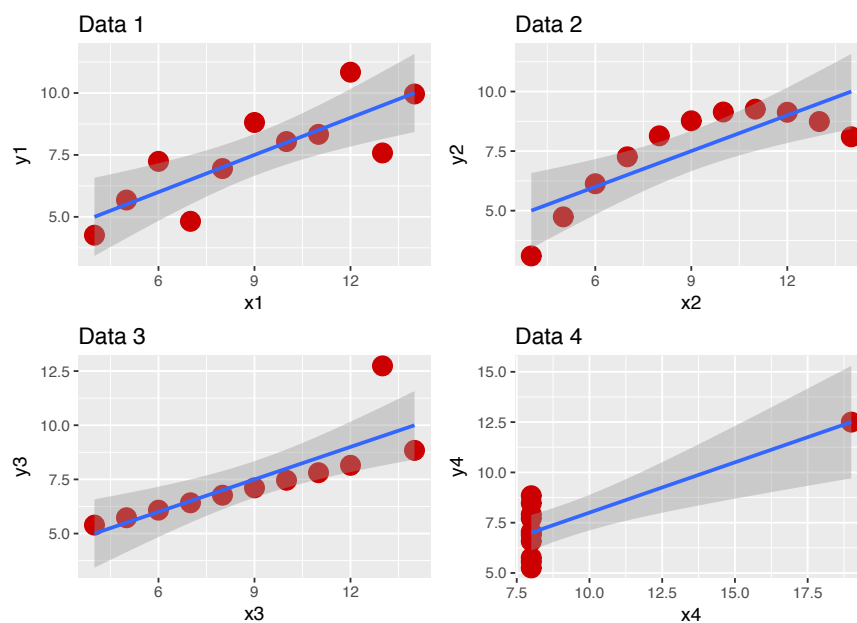
Eine schöne Demonstration der Mächtigkeit von `dplyr` findet sich hier<sup>11</sup>.

<sup>11</sup><http://bit.ly/2kX9lvC>.

## 0.16 Visualisieren

Ein Bild sagt bekanntlich mehr als 1000 Worte. Schauen wir uns zur Verdeutlichung das berühmte Beispiel von Anscombe<sup>12</sup> an. Es geht hier um vier Datensätze mit zwei Variablen (Spalten; X und Y). Offenbar sind die Datensätze praktisch identisch: Alle X haben den gleichen Mittelwert und die gleiche Varianz; dasselbe gilt für die Y. Die Korrelation zwischen X und Y ist in allen vier Datensätzen gleich. Allerdings erzählt eine Visualisierung der vier Datensätze eine ganz andere Geschichte.

```
knitr::include_graphics("images/anscombe.pdf")
```



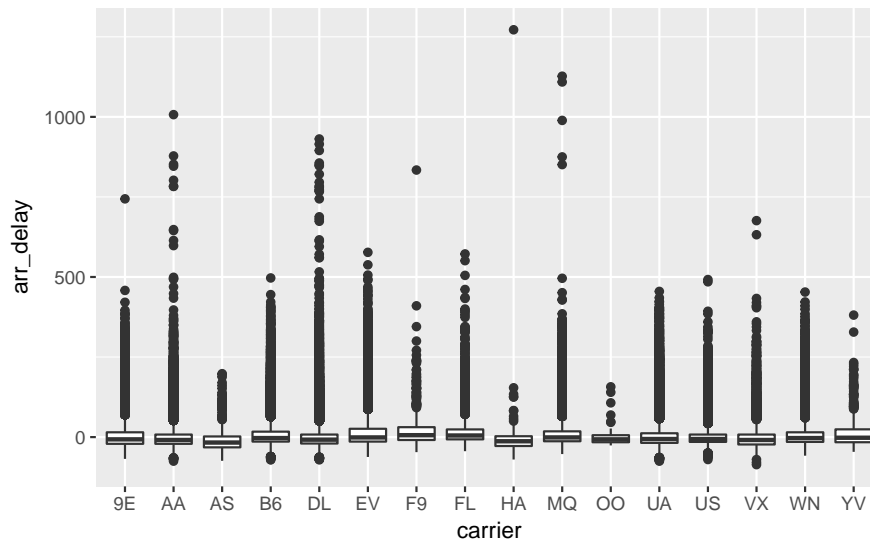
Offenbar “passieren” in den vier Datensätzen gänzlich unterschiedliche Dinge. Dies haben die Statistiken nicht aufgedeckt; erst die Visualisierung erhellte uns... Kurz: Die Visualisierung ist ein unverzichtbares Werkzeug, um zu verstehen, was in einem Datensatz (und damit in der zugrundeliegenden “Natur”) passiert.

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

```
library(ggplot2) # muss installiert sein

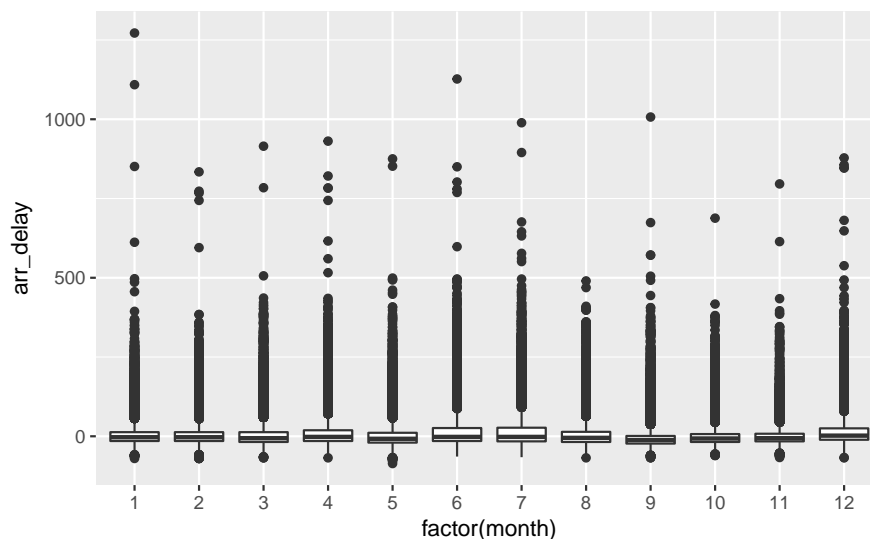
qplot(x = carrier, y = arr_delay, geom = "boxplot", data = flights)
#> Warning: Removed 9430 rows containing non-finite values (stat_boxplot).
```

<sup>12</sup><https://de.wikipedia.org/wiki/Anscombe-Quartett>



Offenbar gibt es viele Extremwerte, was die Verspätung betrifft. Das erscheint mir nicht unplausibel (Schneesturm im Winter, Flugzeug verschwunden...). Vor dem Hintergrund der Extremwerte erscheinen die mittleren Verspätungen (Mediane) in den Boxplots als ähnlich. Vielleicht ist der Unterschied zwischen den Monaten ausgeprägter?

```
qplot(x = factor(month), y = arr_delay, geom = "boxplot", data = flights)
#> Warning: Removed 9430 rows containing non-finite values (stat_boxplot).
```



Kaum Unterschied; das spricht gegen die Schneesturm-Idee als Grund für Verspätung. Aber schauen wir uns zuerst die Syntax von `qplot` näher an. “q” in `qplot` steht für “quick”. Tatsächlich hat `qplot` einen großen Bruder, `ggplot`<sup>13</sup>, der deutlich mehr Funktionen aufweist - und daher auch die umfangreichere (=komplexere) Syntax. Fangen wir mit `qplot` an.

Diese Syntax des letzten Beispiels ist recht einfach, nämlich:

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

<sup>13</sup>Achtung: Nicht `qqplot`, nicht `ggplot2`, nicht `gplot`...

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

### 0.16.0.1 Häufige Arten von Diagrammen

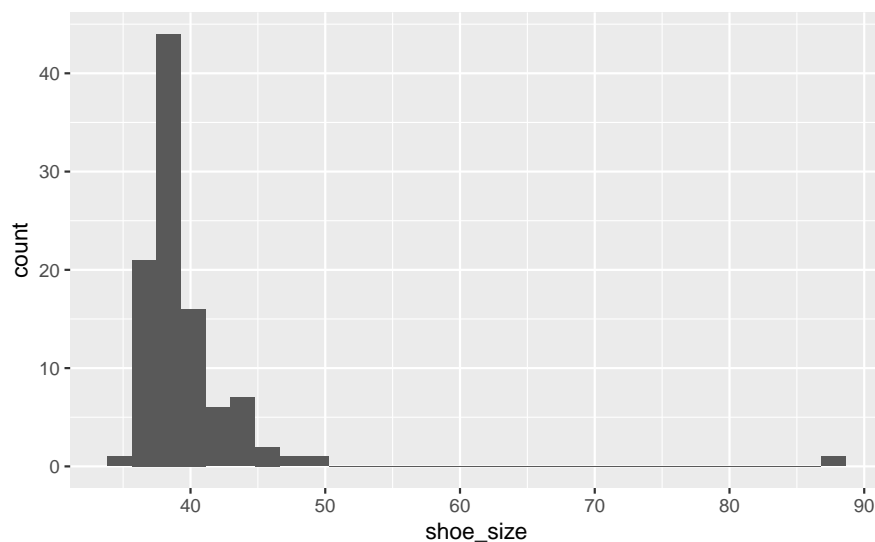
Unter den vielen Arten von Diagrammen und vielen Arten, diese zu klassifizieren greifen wir uns ein paar häufige Diagramme heraus und schauen uns diese der Reihe nach an.

#### 0.16.0.1.1 Eine kontinuierliche Variable

Schauen wir uns die Verteilung der Schuhgrößen von Studierenden an.

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

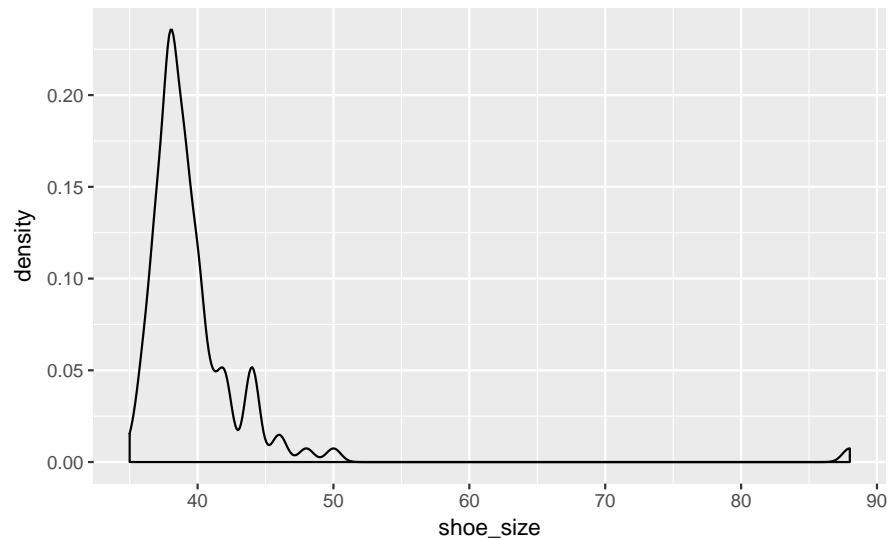
qplot(x = shoe_size, data = wo_men)
#> `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
#> Warning: Removed 1 rows containing non-finite values (stat_bin).
```



Weisen wir nur der X-Achse (aber nicht der Y-Achse) eine kontinuierliche Variable zu, so wählt `ggplot2` automatisch als Geom automatisch ein Histogramm; wir müssen daher nicht explizieren, dass wir ein Histogramm als Geom wünschen (aber wir könnten es hinzufügen). Alternativ wäre ein Dichtediagramm hier von Interesse:

```
# qplot(x = shoe_size, data = wo_men) wie oben

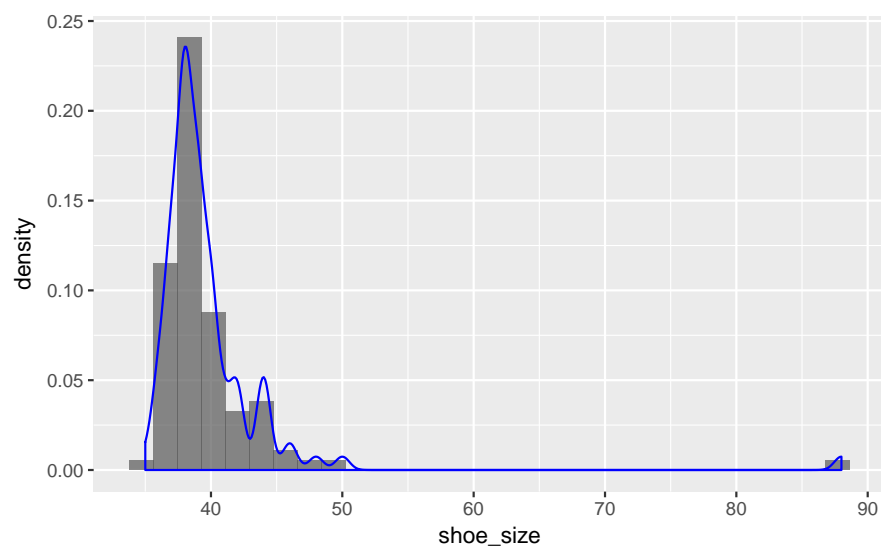
qplot(x = shoe_size, data = wo_men, geom = "density")
#> Warning: Removed 1 rows containing non-finite values (stat_density).
```



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

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

```
ggplot(data = wo_men) +
  aes(x = shoe_size) +
  geom_histogram(aes(y = ..density..), alpha = .7) +
  geom_density(color = "blue")
#> `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
#> Warning: Removed 1 rows containing non-finite values (stat_bin).
#> Warning: Removed 1 rows containing non-finite values (stat_density).
```

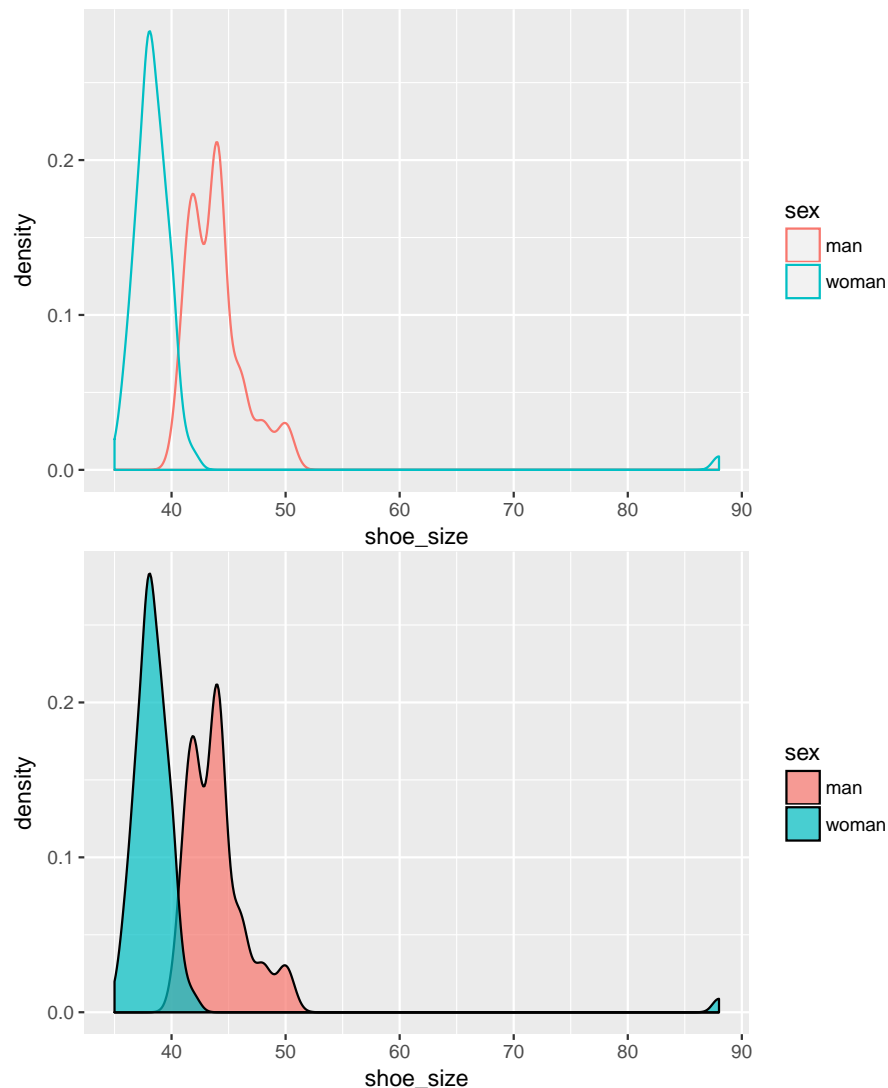


Zuerst haben wir mit dem Parameter `data` den Dataframe benannt. `aes` definiert, welche Variablen welchen Achsen (oder auch z.B. Füllfarben) zugewiesen werden. Hier sagen wir, dass die Schuhgröße auf X-Achse stehen soll. Das `+`-Zeichen trennt die einzelnen Bestandteile des `ggplot`-Aufrufs voneinander. Als nächstes sagen wir, dass wir gerne ein Histogramm hätten: `geom_histogram`. Dabei soll aber nicht wie gewöhnlich auf der X-Achse die Häufigkeit stehen, sondern die Dichte. `ggplot` berechnet selbständig die Dichte und nennt

diese Variabl `..density..`; die vielen Punkte sollen wohl klar machen, dass es sich nicht um eine “normale” Variable aus dem eigenen Dataframe handelt, sondern um eine “interne” Variable von `ggplot` - die wir aber nichtsdestotrotz verwenden können. `alpha` bestimmt die “Durchsichtigkeit” eines Geoms; spielen Sie mal etwas damit herum. Schließlich malen wir noch ein blaues Dichtediagramm über das Histogramm.

Wünsche sind ein Fass ohne Boden... Wäre es nicht schön, ein Diagramm für Männer und eines für Frauen zu haben, um die Verteilungen vergleichen zu können?

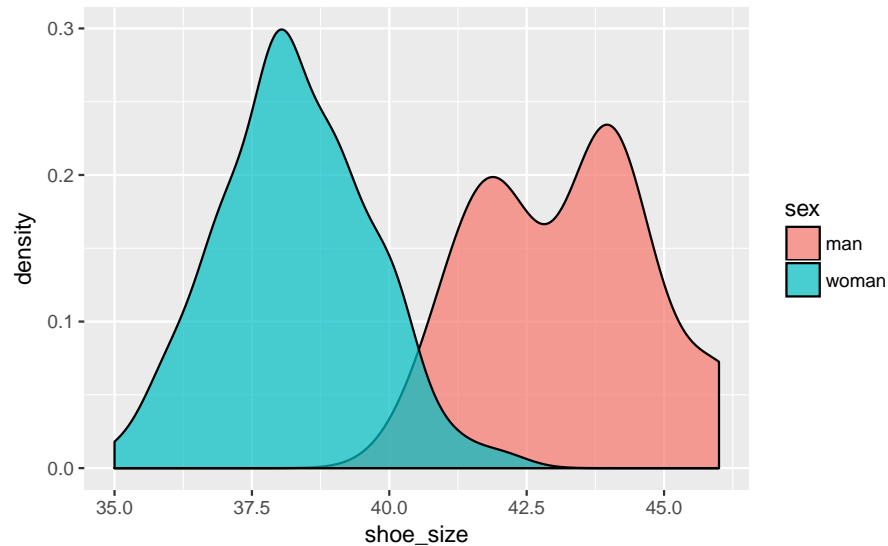
```
qplot(x = shoe_size, data = wo_men, geom = "density", color = sex)
#> Warning: Removed 1 rows containing non-finite values (stat_density).
qplot(x = shoe_size, data = wo_men, geom = "density", fill = sex, alpha = I(.7))
#> Warning: Removed 1 rows containing non-finite values (stat_density).
```



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

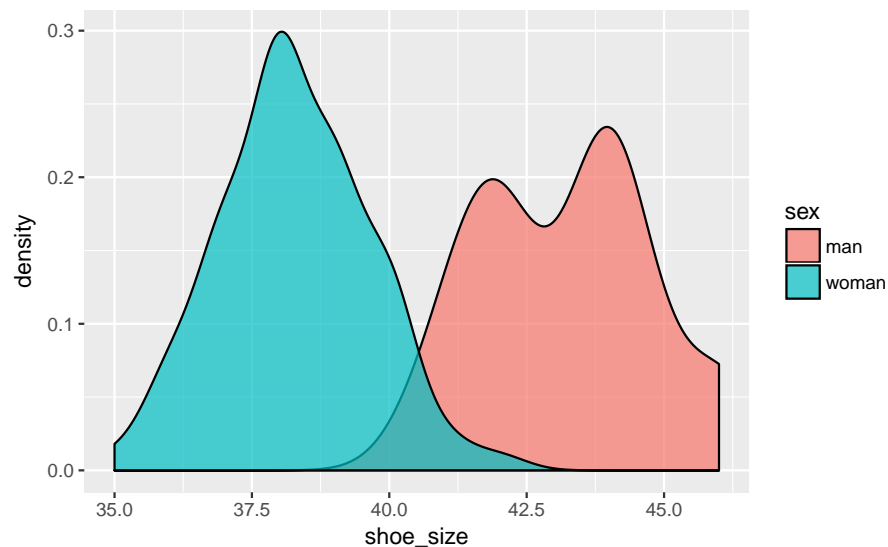
```
wo_men %>%
  filter(shoe_size <= 47) -> wo_men2
```

```
qplot(x = shoe_size, data = wo_men2, geom = "density", fill = sex, alpha = I(.7))
```



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

```
wo_men %>%
  filter(shoe_size <= 47) %>%
  qplot(x = shoe_size, data = ., geom = "density", fill = sex, alpha = I(.7))
```



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

Mit `fill = sex` sagen wir `qplot`, dass er für Männer und Frauen jeweils ein Dichtediagramm erzeugen soll; jedem Dichtediagramm wird dabei eine Farbe zugewiesen (die uns `ggplot2` im Standard voraussucht). Mit anderen Worten: Die Werte von `sex` werden der Füllfarbe der Histogramme zugeordnet. Anstelle der

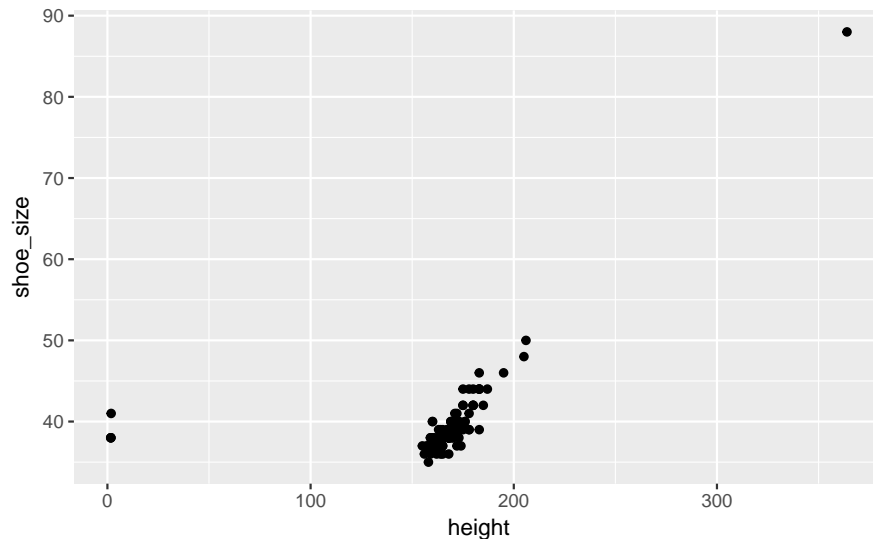


Füllfarbe hätten wir auch die Linienfarbe verwenden können; die Syntax wäre dann: `color = sex`.

## 0.17 Zwei kontinuierliche Variablen

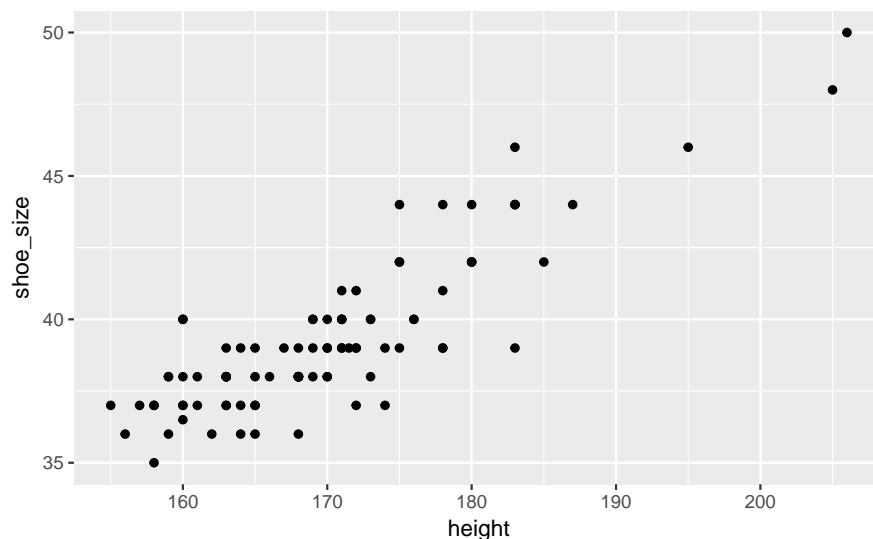
Ein Streudiagramm ist die klassische Art, zwei metrische Variablen darzustellen. Das ist mit `qplot` einfach:

```
qplot(x = height, y = shoe_size, data = wo_men)
#> Warning: Removed 1 rows containing missing values (geom_point).
```



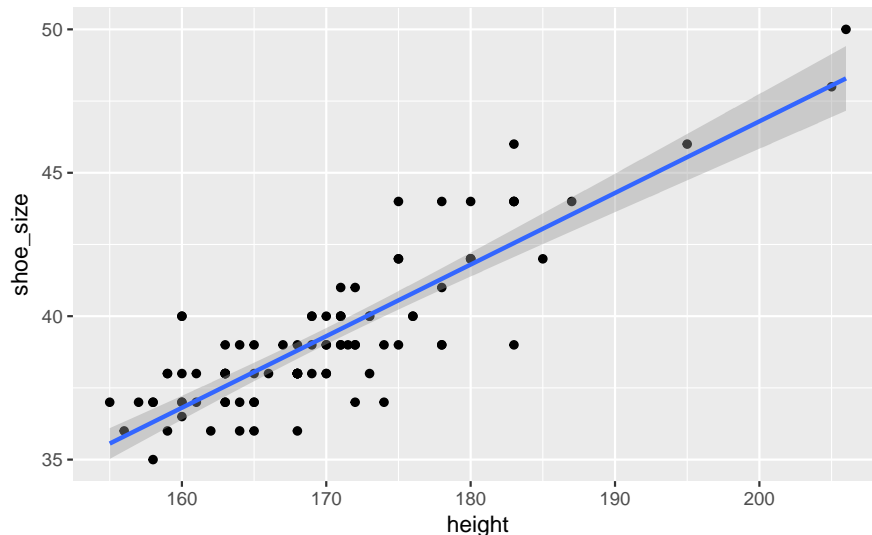
Wir weisen wieder der X-Achse und der Y-Achse eine Variable zu; handelt es sich in beiden Fällen um Zahlen, so wählt `ggplot2` automatisch ein Streudiagramm - d.h. Punkte als Geom (`geom = "point"`). Wir sollten aber noch die Extremwerte herausnehmen:

```
wo_men %>%
  filter(height > 150, height < 210, shoe_size < 55) %>%
  qplot(x = height, y = shoe_size, data = .)
```



Der Trend ist deutlich erkennbar: Je größer die Person, desto länger die Fuß'. Zeichnen wir noch eine Trendgerade ein.

```
wo_men %>%
  filter(height > 150, height < 210, shoe_size < 55) %>%
  qplot(x = height, y = shoe_size, data = .) +
  geom_smooth(method = "lm")
```



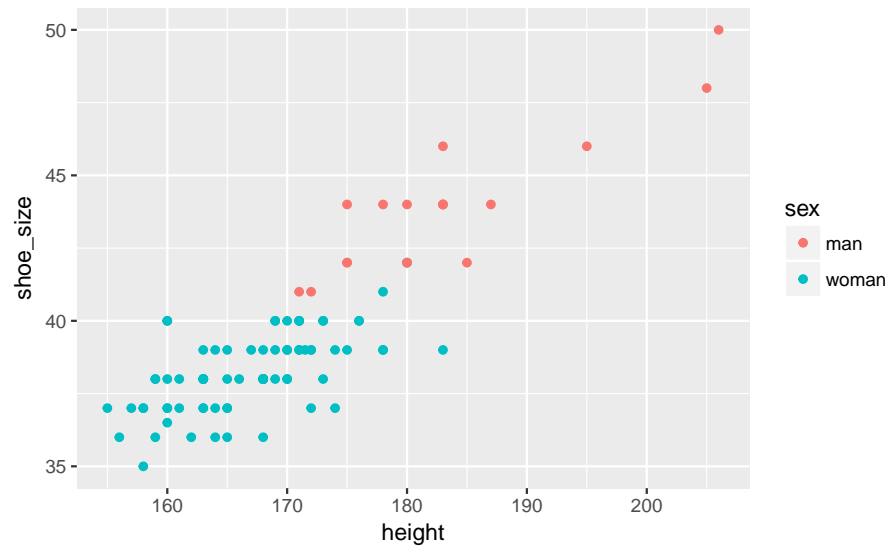
Synonym könnten wir auch schreiben:

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

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

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

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

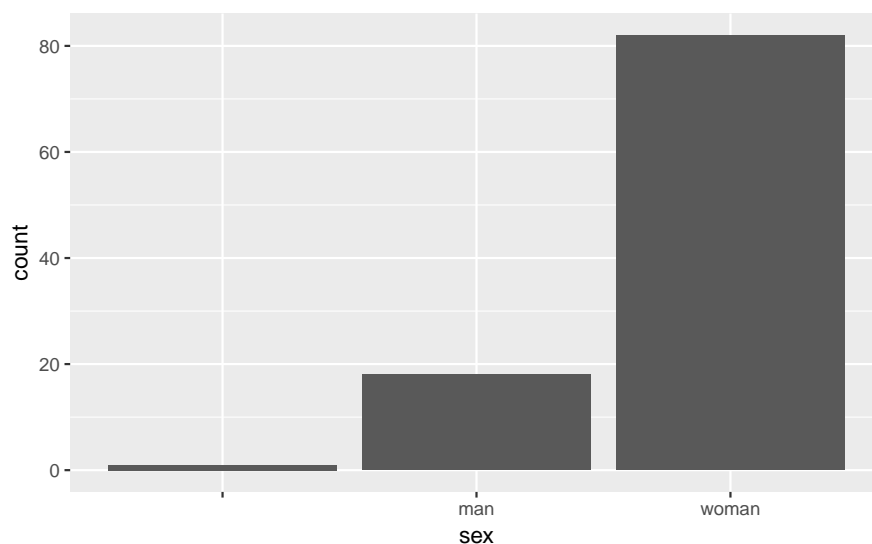


Mit `color = sex` sagen wir, dass die Linienfarbe (der Punkte) entsprechend der Stufen von `sex` eingefärbt werden sollen. Die genaue Farbwahl übernimmt `ggplot2` für uns.

### 0.17.1 Eine diskrete Variable

Bei diskreten Variablen, vor allem nominalen Variablen, geht es in der Regel darum, Häufigkeiten auszuzählen. Wie viele Männer und Frauen sind in dem Datensatz?

```
qplot(x = sex, data = wo_men)
```

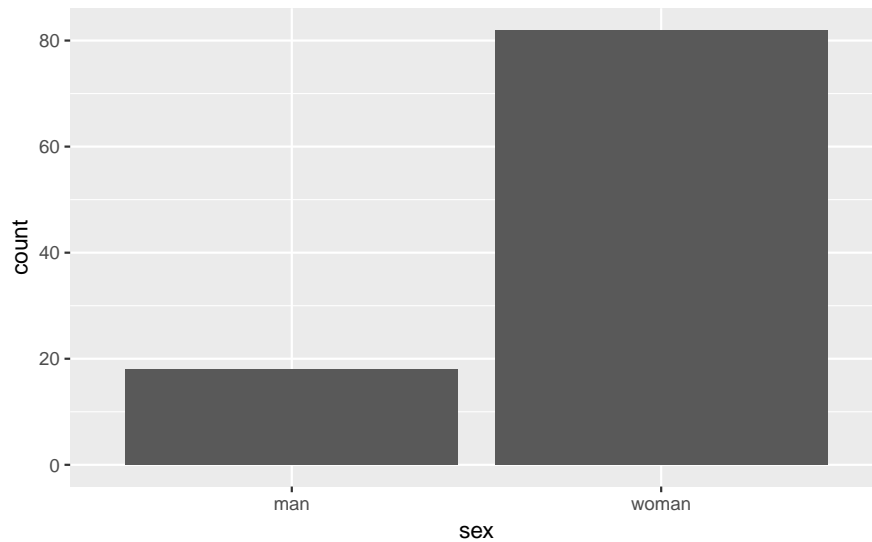


Falls nur die X-Achse definiert ist und dort eine Faktorvariable oder eine Text-Variable steht, dann nimmt `qplot` automatisch ein Balkendiagramm als Geom.

Entfernen wir vorher noch die fehlenden Werte:

```
wo_men %>%  
  na.omit() %>%
```

```
qplot(x = sex, data = .)
```

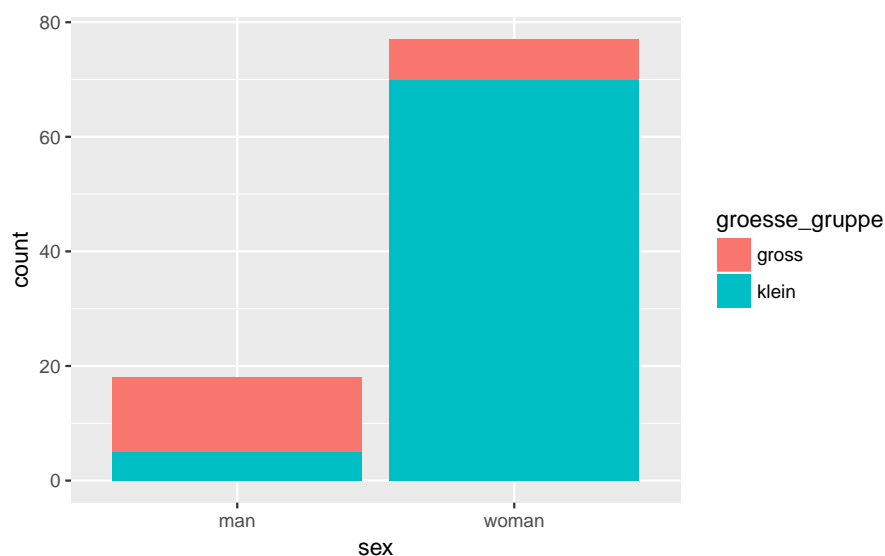


Wir könnten uns jetzt die Frage stellen, wie viele kleine und viele große Menschen es bei Frauen und bei den Männern gibt. Dazu müssen wir zuerst eine Variable wie “Größe gruppiert” erstellen mit zwei Werten: “klein” und “groß”. Nennen wir sie `groesse_gruppe`

```
library(car)
wo_men$groesse_gruppe <- car::recode(wo_men$height, "lo:175 = 'klein'; else = 'gross'")

wo_men %>%
  filter(height > 150, height < 210, shoe_size < 55) %>%
  na.omit -> wo_men2

qplot(x = sex, fill = groesse_gruppe, data = wo_men2)
```

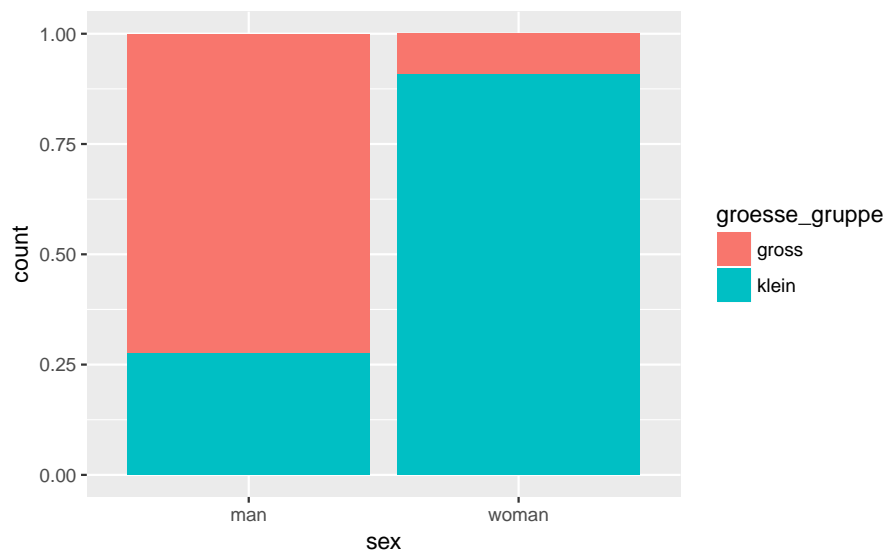


In Worten sagt der `recode`-Befehl hier in etwa: “Kodierte `wo_men$height` um, und zwar vom kleinsten (10) Wert bis 170 soll den Wert `klein` bekommen, ansonsten bekommt eine Größe den Wert `gross`”.

Hier haben wir `qplot` gesagt, dass der die Balken entsprechend der Häufigkeit von `groesse_gruppe` füllen soll. Und bei den Frauen sind bei dieser Variablen die Werte `klein` häufig; bei den Männern hingegen die Werte `gross`.

Schöner wäre noch, wenn die Balken Prozentwerte angeben würden. Das geht mit `qplot` (so) nicht; wir schwenken auf `ggplot` um.

```
wo_men2 %>%
  ggplot() +
  aes(x = sex, fill = groesse_gruppe) +
  geom_bar(position = "fill")
```



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

### 0.17.2 Zusammenfassungen zeigen

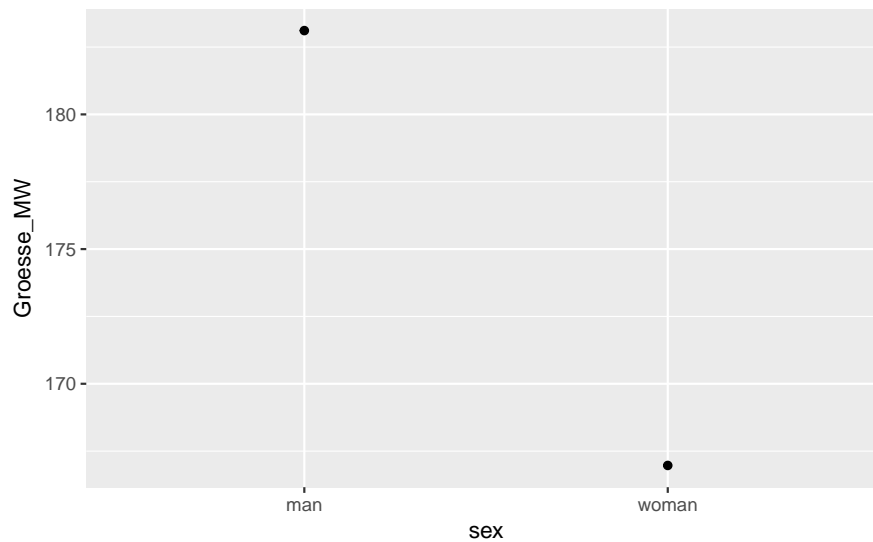
Manchmal möchten wir *nicht* die Rohwerte einer Variablen darstellen, sondern z.B. die Mittelwerte pro Gruppe. Mittelwerte sind eine bestimmte *Zusammenfassung* einer Spalte; also fassen wir zuerst die Körpergröße zum Mittelwert zusammen - gruppiert nach Geschlecht.

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

```
wo_men3
#> # A tibble: 2 × 2
#>   sex Groesse_MW
#>   <fctr>      <dbl>
#> 1   man      183
#> 2  woman      167
```

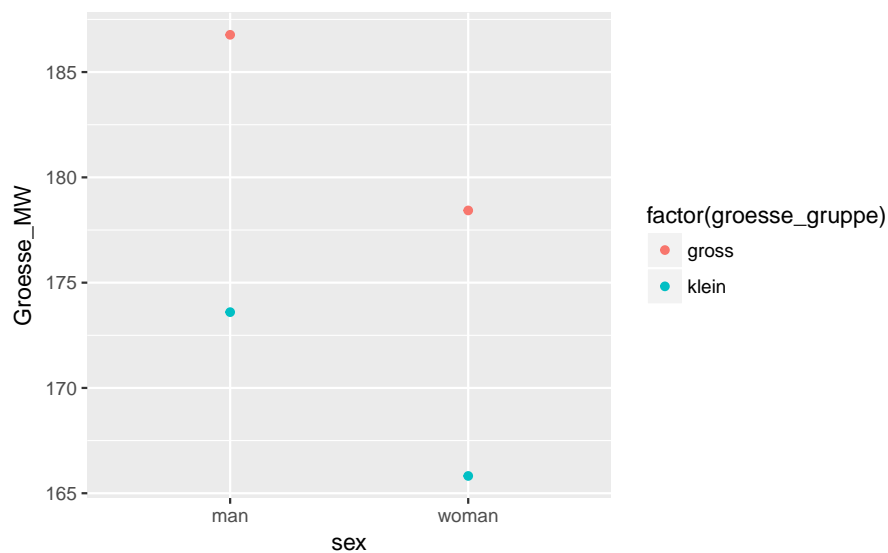
Diese Tabelle schieben wir jetzt in `ggplot2`; natürlich hätten wir das gleich in einem Rutsch durchpfeifen können.

```
wo_men3 %>%  
  qplot(x = sex, y = Groesse_MW, data = .)
```



Das Diagramm besteht nicht durch die Tiefe und Detaillierung. Wenn wir noch zusätzlich die Mittelwerte nach `Groesse_Gruppe` ausweisen, wird das noch überschaubar bleiben.

```
wo_men2 %>%  
  group_by(sex, groesse_gruppe) %>%  
  summarise(Groesse_MW = mean(height)) %>%  
  qplot(x = sex, color = factor(groesse_gruppe), y = Groesse_MW, data = .)
```



# Visualisierung

- Nutzen (Anscombe)
- Prinzipien nach Tufte
- Cleveland
- ggplot2





# Statistisches Modellieren

- Was sind Modelle?
- Überanpassung
- Prädiktion vs. Explanation
- Numerische vs. klassifizierende Modelle
- Geleitete vs. ungeleitete Modelle
- Parametrische vs. nichtparametrische Modelle
- Fehler- vs. Varianzreduktion
- Modellgüte



# Numerische Modelle

- Lineare Regression
  - Grundlagen
  - Multiple Regression
  - Interaktion
  - Eisberge
- Logistische Regression
- Penalisierende Regression
- Baumbasierte Verfahren
- Ausblick



# Klassifizierende Modelle

- Clusteranalyse
- Nächste-Nachbarn-Analyse



# Literaturverzeichnis