Kapitel 1

Statistische Daten und Modelle

1.1 Grundlagen

1.1.1 Begriffe

1.1.1.1 Was ist Statistik?

Das Wort **Statistik** hat mehrere Bedeutungen. Hier die zwei wichtigsten:

Definition 1 (Statistik)

1. *Statistik, eine – (ursprüngliche Bedeutung, Etymologie: Aufstellung, gleiche Wortwurzel wie Staat)*
   Eine Aufstellung über Personen, Werte oder Güter in staatlichen Gebietskörperschaften. z.B. Landwirtschaftsstatistik, Gewerbestatistik, ...

2. **Statistik**, *die (moderne Bedeutung)*
   Die Wissenschaft vom wissenschaftlichen Schließen aus zufallsbeeinflußten Daten.

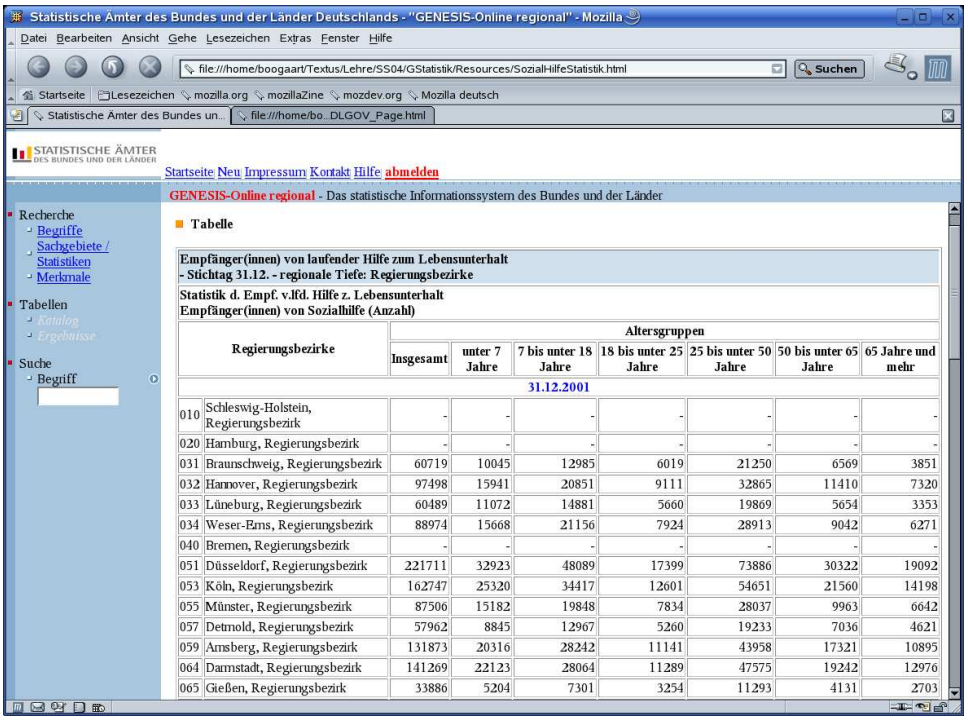
3. *, eine (im Sinne der Statistik 2)*
   eine Vorschrift (oder Formel) zur Berechnung aus Daten von Kenngrößen aus den Daten (z.B. der Mittelwert).

Wir werden uns hier mit Statistik (2) beschäftigen.

1.1.1.2 Was ist Datenanalyse?

Eine statistische Untersuchung findet heutzutage normalerweise in ein paar groben Schritten statt:

1. Formulierung der Fragestellung
2. Planung der Erhebung
3. Durchführung der Datenerhebung
4. Explorative Datenanalyse (Entdecken von Strukturen)
5. Konfirmatorische Datenanalyse (Nachweis der gefundenen Effekte)



Regierungsbezirke		Altersgruppen						
		Insgesamt	unter 7 Jahre	7 bis unter 18 Jahre	18 bis unter 25 Jahre	25 bis unter 50 Jahre	50 bis unter 65 Jahre	65 Jahre und mehr
31.12.2001								
010	Schleswig-Holstein, Regierungsbezirk	-	-	-	-	-	-	-
020	Hamburg, Regierungsbezirk	-	-	-	-	-	-	-
031	Braunschweig, Regierungsbezirk	60719	10045	12985	6019	21250	6569	3851
032	Hannover, Regierungsbezirk	97498	15941	20851	9111	32865	11410	7320
033	Lüneburg, Regierungsbezirk	60489	11072	14881	5660	19869	5654	3353
034	Weser-Ems, Regierungsbezirk	88974	15668	21156	7924	28913	9042	6271
040	Bremen, Regierungsbezirk	-	-	-	-	-	-	-
051	Düsseldorf, Regierungsbezirk	221711	32923	48089	17399	73886	30322	19092
053	Köln, Regierungsbezirk	162747	25320	34417	12601	54651	21560	14198
055	Münster, Regierungsbezirk	87506	15182	19848	7834	28037	9963	6642
057	Detmold, Regierungsbezirk	57962	8845	12967	5260	19233	7036	4621
059	Arnsberg, Regierungsbezirk	131873	20316	28242	11141	43958	17321	10895
064	Darmstadt, Regierungsbezirk	141269	22123	28064	11289	47575	19242	12976
065	Gießen, Regierungsbezirk	33886	5204	7301	3254	11293	4131	2703

Abbildung 1.1: Beispiel für Statistik (1)

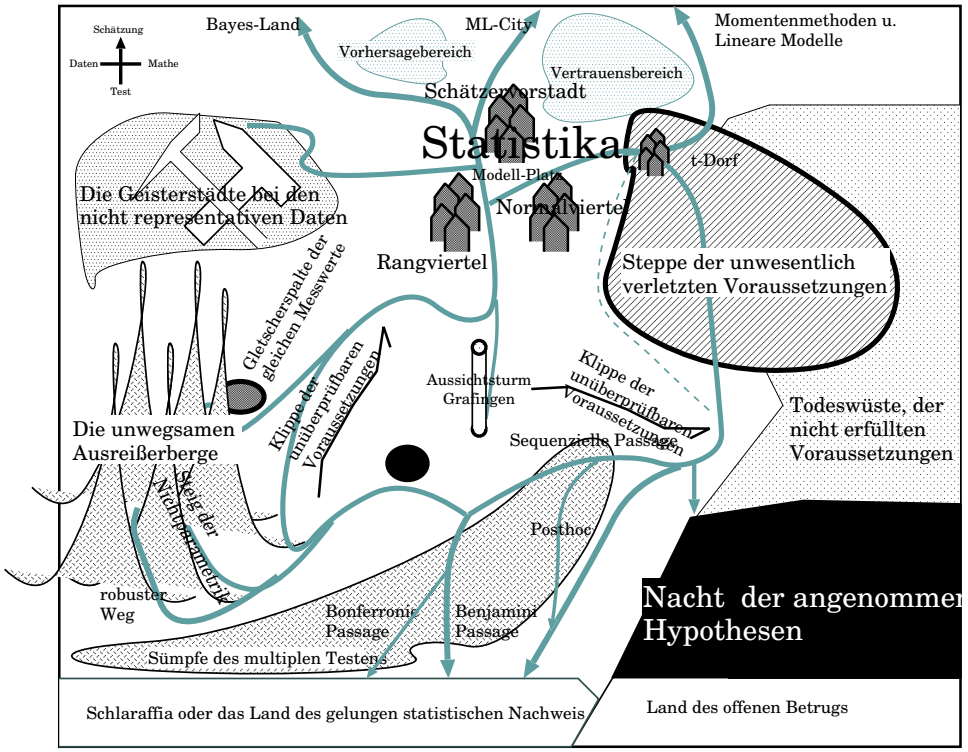


Abbildung 1.2: Eine mentale Landkarte für diese Vorlesung. Der Abschnitt 1.1.2 beschäftigt sich mit dem Ziel des "statistischen Nachweises" als integralem Bestandteil jeder empirischen Wissenschaft.

6. Präsentation der Ergebnisse

In der Zeit vor dem Computer galt eher folgender Ablaufplan:

1. Formulierung der Fragestellung
2. Auswahl der dazu passenden statistischen Auswertungsmethode
3. Durchführung einer geeigneten Datenerhebung
4. Auswertung mit der gewählten statistischen Methode (meist per Hand)

Insofern ist die Datenanalyse ein spezieller Teil der Statistik, der von erhobenen Daten ausgeht und versucht aus diesen so viel wie möglich Informationen über die Wirklichkeit zu erhalten. Die Datenanalyse baut dabei auf den einzelnen statistischen Verfahren auf und kombiniert sie zu einem großen Ganzen.

In einer Vorlesung "Datenanalyse und Statistik" müssen wir also zunächst die statistischen Verfahren verstehen und dann lernen, wann und in welcher Situation man sie anwendet, wie man die Ergebnisse interpretiert und schließlich, wie man sie weiterverwendet.

1.1.1.3 Was ist Stochastik?

In seiner ursprünglichen Wortbedeutung bedeutet Stochastik "Die Kunst (aus Beobachtungen) Schlüsse zu ziehen". In dieser Bedeutung ist das Wort aber von der Statistik (2) abgelöst worden. Heute bezeichnet das Wort Stochastik einfach alles, was mit Zufall zu tun hat. Insbesondere also

- Das Rechnen mit dem Zufall (die Wahrscheinlichkeitstheorie)
- Das Modellieren zufälliger Vorgänge (Stochastische Prozesse)
- Aus zufälligen Beobachtungen zu schließen (Statistik)
- Der Umgang mit zufälligem Risiko (Versicherungsmathematik und die Zuverlässigkeitstheorie)

Eine Vorlesung "Stochastik und Statistik" geht also über die bloße Analyse von Daten insbesondere dadurch hinaus, dass sie versucht, den Zufall auch rechnerisch zu erfassen, die Wirklichkeit durch stochastische Modelle zu beschreiben und daraus wiederum mit formalen mathematischen Mitteln Schlussfolgerungen über Gefährdung und Betriebszuverlässigkeit zu ziehen.

1.1.2 Die Rolle der Statistik in den Natur- und Sozialwissenschaften

Es gibt grundsätzlich verschiedene Ansätze für die Wissenschaft

- **deduktiv**
  Aus bekannten Aussagen wird durch logische Ableitung Neues gefolgert. Anschließend muss diese Behauptung in der Realität belegt werden.

  Das wäre hier die Aufgabe der Statistik.

- **induktiv**
  Aus Beobachtungen werden Zusammenhänge abgeleitet, beschrieben und nachgewiesen.

  Die Aufgabe der Statistik wäre hier, Methoden bereitzustellen, wie man aus Daten Schlüsse ziehen kann. Z.B. ist zu klären, unter welchen Umständen wir

aus 10 Antworten "Bachelor" folgern können, dass die Mehrheit der Studenten in der Veranstaltung einen Bachelor-Studiengang besucht. Z.B. wäre es sicher kein zulässiger Schluss zulässige, würden wir einfach nach der Veranstaltung eine Gruppe zusammenstehender Studenten zu befragen. Vermutlich studieren alle in der Gruppe ohnehin das gleiche, so dass wir nicht mehr Information haben, als wenn wir nur einen einzelnen befragt hätten.

- **falsifizierend**

  Eine These ist wissenschaftlich, wenn sie, so sie denn falsch ist, auch falsifiziert werden könnte, bisher aber nicht falsifiziert worden ist.

  Die Aufgabe der Statistik ist es, Methoden zur Verfügung zu stellen, um Modelle, die Zufallskomponenten enthalten, theoretisch falsifizierbar zu machen und sie so in den Stand wissenschaftlicher Modelle zu erheben.

Beispiel 2 (Deduktiver Ansatz) *Tünen: Landwirtschaftliche Produkte werden am Markt in der Stadt verkauft und müssen vom Produktionsort dorthin transportiert werden. Dadurch entstehen Transportkosten. Diejenigen Produkte, die im Verhältnis zum landwirtschaftlichen Platzverbrauch hohe Transportkosten haben, werden also nahe der Stadt angebaut.*

z.B. Diese Theorie war zu Tünens Zeiten wohl gültig, aber ist sie es heute noch?

Um die Vorhersagen der Theorie in der Realität zu bestätigen, würden wir Daten über dir produktspezifische Transportkosten und die Anbaugebiete sammeln und darin versuchen, den vorhergesagten je-desto-Zusammenhang wiederzufinden. Dabei tritt allerdings ein Problem auf: Auch in der Stadt leben hier und da ein paar schlachtbare Tiere und auch weit draußen haben die Bauern Gemüsebeete. Der je-desto-Zusammenhang gilt also sicher nicht exakt.

Beispiel 3 (Induktiver Ansatz) *Wir machen eine Umfrage in der Vorlesung, was die Anwesenden studieren und schließen daraus, es wäre eine Vorlesung, die hauptsächlich von ... Wissenschaftlern besucht wird.*

Die Daten sind aber zufällig, weil wir nur einen (zufälligen) Teil der Anwesenden befragt haben und die Angaben der Personen unterschiedlich sind. Wir können nicht sicher sein, ob uns nicht gerade die paar Ausnahmen in die Umfrage geraten sind.

1.1.3 Rolle der Stochastik in den Ingenieurwissenschaften

Stochastik und Statistik werden in den Ingenieurwissenschaften in vielfältiger Weise eingesetzt.

1.1.3.1 Themen, die im Rahmen der Vorlesung angesprochen werden

- **Auswertung von Experimenten**
  Im Rahmen der Entwicklung neuer Technologien und im Rahmen der Übertragung derselben in neue Anwendungssituationen werden immer wieder Versuche durchgeführt. Die Statistik ist die Wissenschaft von der Auswertung der so erzeugten Daten.

- **Schätzung**
  Schätzung von Ausfallwahrscheinlichkeiten, durchschnittlichen Lebensdauern, Verarbeitungskapazitäten, Materialkonstanten, Verbrauchszahlen,...

- **Zuverlässigkeitstheorie** Die Berechnung der Betriebszuverlässigkeit einfacher und komplexer Systeme.

- **Belastungsgrenzen: Extremwerttheorie** Sowohl für die Berechnung von über eine längere Betriebsdauer maximal auftretenden Belastungen sowie für die Berechnung der minimalen Belastbarkeitsgrenze über eine größere Anzahl von Belastungsereignissen benötigt man die stochastische Extremwerttheorie. Also z.B. wie hoch muss ich den Damm bauen, damit er die nächsten 100 Jahre hält. Wie haltbar muss das Seil sein, damit in den nächsten 30 Jahren kein Karussellsitz von einem Karussell ihrer Firma abreißt.
- **Fehlerrechnung und Sensitivitätsanalyse**
  Oft sind genaue Daten nicht verfügbar, und es muß mit Schätzgrößen gerechnet werden. Die Fehlerrechnungen und Sensitivitätsanalyse ermöglichen festzustellen, wie stark die berechneten Größen von den dann tatsächlich beobachteten Größen abweichen werden.
- **Nachweis von Unterschieden und Verbesserungen**
  Statistische Tests erlauben zu beweisen, dass eine Technologie im Durchschnitt besser ist als eine andere, selbst, wenn die Verbesserung nicht für jeden Einzelfall garantiert werden kann.

1.1.3.2 Komplexe Themen, die hier leider nicht behandelt werden können

- **Versuchsplanung**
  Wie plant man Experimente so, dass man maximale Informationen daraus ziehen kann.
- **Geostatistik**
  Die Geostatistik beschäftigt sich mit der Vorhersage von Werten an Orten, an denen tatsächlich nicht beobachtet wurde. Das wird beispielsweise bei der Lagerstättenvorratsberechnung und der Abbauplanung in den Geoingeneurwissenschaften benötigt.
- **Stochastische Geometrie, Stereologie und Räumliche Statistik**
  Komplexe zufällige geometrische Strukturen, wie sie im Inneren von Werkstoffen auftreten und die sich aus diesen Strukturen ergebenden makroskopischen Eigenschaften werden systematisch von der Stochastische Geometrie untersucht. Stereologie und räumliche Statistik zeigen, wie dabei wichtige Strukturparameter aus kostengünstig möglichen Beobachtungen geschätzt werden können.
- **Bedienungstheorie**
  Die Warteschlangentheorie beschäftigt sich mit dem Langfristverhalten komplexer Systeme, wie z.B. in Call-Centers, bei Computernetzwerken, oder im Rahmen von betrieblichen Reparaturabläufen zustande kommen.
- **Stochastische Optimierung**
  Die Stochastische Optimierung erlaubt die Optimierung von Systemen unter unvollständiger Information. So z.B. im Rahmen der stochastischen Abbauplanung im Tagebau oder bei der Projektplanung mit unbekannten Bearbeitungszeiten für Einzelvorgänge.
- **Stochastische Simulation**
  Die stochastische Simulation beschäftigt sich mit der Simulation von zufälligen Vorgängen. Mit Hilfe solcher Simulationen kann das Gesamtverhalten komplexer Systeme auch unter extremen Bedingungen untersucht werden, noch ehe überhaupt ein Prototyp gebaut wurde.
- ... und viele andere mehr

1.2 Grundlegende Modelle der Statistik

In der modernen Statistik treffen drei Bereiche aufeinander:

- die Realität (und unser konzeptionelles Modell von ihr)
- die mathematische Statistik (mit ihren Modellen und Formeln)
- die Statistiksoftware (womit man letztlich arbeitet)

Alle drei Bereiche nehmen die selben Dinge (genannt Daten) sehr verschieden wahr. Ziel dieses Abschnitts ist es, eine Gesamtsicht darauf zu entwickeln.

Wir versuchen deshalb, alle Begriffe sofort zur Software in Beziehung zu setzen. Die Vorlesung und das Skript verwenden dazu die Statistiksoftware R.

R ist ein freies Programm-Paket zur Statistik und kann unter http://www.cran.r-project.org/ kostenlos heruntergeladen werden. Es läuft auf allen mir bekannten Betriebssystemen. R ist ein kommandozeilenbasiertes Programm. Man gibt Befehle ein, die dann sofort ausgeführt werden. Das Ergebnis der Berechnung wird wiederum ausgegeben. Auf diese Weise kann man in einem Skript wie diesem sehr leicht zeigen, welcher Befehl was bewirkt.

Gibt man z.B. eine Formel ein, so wird deren Wert ausgerechnet:

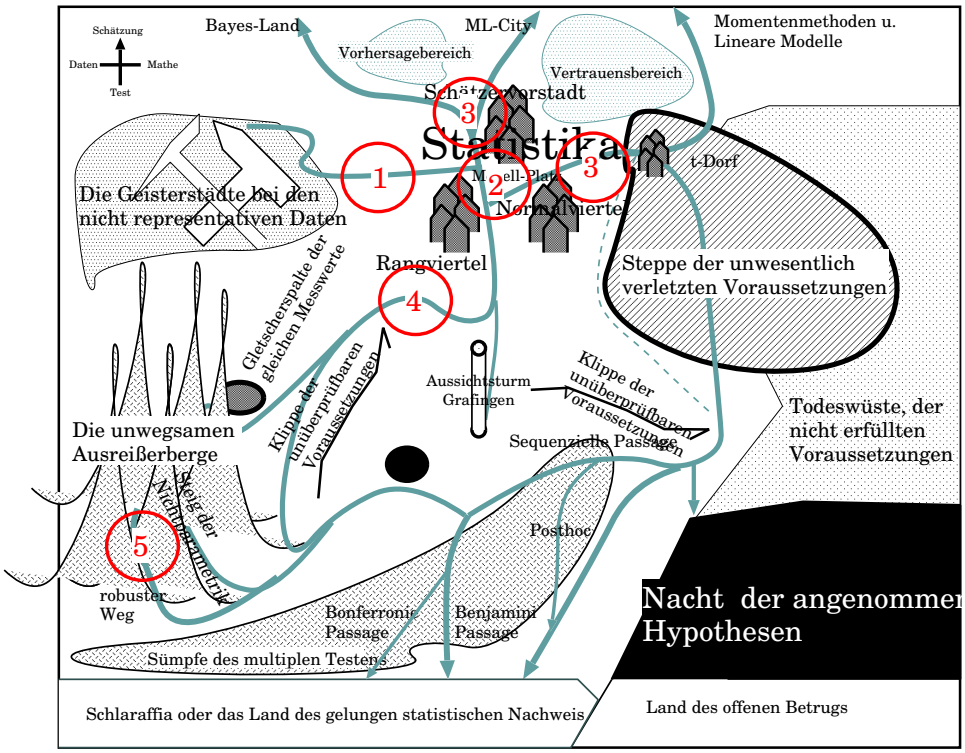
```
R version 2.5.0 beta (2007-04-12 r41139)
Copyright (C) 2007 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R ist freie Software und kommt OHNE JEGLICHE GARANTIE.
Sie sind eingeladen, es unter bestimmten Bedingungen weiter zu verbreiten.
Tippen Sie 'license()' or 'licence()' fÃijr Details dazu.

R ist ein Gemeinschaftsprojekt mit vielen Beitragenden.
Tippen Sie 'contributors()' fÃijr mehr Information und 'citation()',
um zu erfahren, wie R oder R packages in Publikationen zitiert werden kÃ■nnen.

Tippen Sie 'demo()' fÃijr einige Demos, 'help()' fÃijr on-line Hilfe, oder
'help.start()' fÃijr eine HTML Browserschnittstelle zur Hilfe.
Tippen Sie 'q()', um R zu verlassen.

>
NULL

> 6 * 6

[1] 36
```

Sie basiert auf einer Kommandozeile, in die man Befehle eingeben kann. Leider können wir dieses Semester keine computerbasierten Übungen anbieten. Allerdings können Sie anhand des Skripts alle erlernten Dinge zu Hause am Computer nachvollziehen.

1.2.1 Die Grundbegriffe

Typischerweise trifft die Statistik Aussagen über eine oder mehrere Gesamtheiten von Individuen. Obwohl es noch andere Modelle statistischen Schließens gibt, ist dieses das wohl Geläufigste.

Dabei spielen die folgenden Begriffe eine zentrale Rolle:



1 Ohne Repräsentativität geht nichts.

2 Der Ausgangspunkt der Statistik ist eine Modellvorstellung.

3 Parametrische Modelle führen auf die Wege der parametrischen Statistik.

4 Nichtparametrische Modelle führen zu nichtparametrischen Methoden.

5 Außerdem gibt es noch robuste Modelle und Methoden.

Abbildung 1.3: In Abschnitt 1.2 werden die ersten Wegweiser aufgestellt.

- statistisches Individuum
- Grundgesamtheit
- Stichprobe
- repräsentativ oder unabhängig
- Zufallsvariable
- Realisierung

Es gibt dabei zwei grundsätzlich verschiedene Arten der statistischen Erhebung, die allerdings letztlich mit den gleichen Methoden behandelt werden:

1. eine oder mehrere repräsentative Stichproben von Grundgesamtheiten
   Diese Methode wird verwendet, wenn Aussagen über die existierende Natur gemacht werden sollen.

   - Besteht im Untersuchungsgebiet ein Zusammenhang zwischen Gold- und Silber-Vererzung (Diplomarbeit Geologie Freiberg)
   - Woran kann man diese fossilen Käfer unterscheiden (z.B. Paläontologie Freiberg)
   - Ist der Reaktorbereich jetzt strahlungsfrei (z.B. TÜV MV)
   - Ist diese Käferart endemisch? (z.B. Ökologie Greifswald)
   - Erzeugt dieser Produktionsprozess die richtige Textur (Aluminiumindustrie)?
   - ...

2. eine oder mehrere Reihen gleichartiger und unabhängig voneinander durchgeführter Versuche

   - Wie beeinflusst das Maatregime die Vogelpopulation?

3. andere spezielle Versuchsaufbaue

Beispiel 4 (Sonntagsfrage)
Ein Meinungsforschungsinstitut stellt einer Anzahl Wahlberechtigter die Frage: "Welche Partei würden Sie wählen, wenn nächsten Sonntag Bundestagswahl wäre." Das Institut möchte damit Rückschlüsse auf die Gesamtheit aller Wahlberechtigten ziehen.

In diesem Fall bezeichnet die Statistik:

- Die Wahlberechtigten als "statistische Individuen".
- Die Menge aller Wahlberechtigten als die "Grundgesamtheit".
- Die Teilmenge der tatsächlich befragten Wahlberechtigten als "Stichprobe".
- Die Stichprobe als "repräsentativ", wenn sie nach gewissen Regeln erhoben wurde, so dass der obige Rückschluss tatsächlich zulässig ist.

1.2.2 Grundgesamtheit

Definition 5 (Grundgesamtheit) *Zu einer statistischen Analyse legt man zunächst eine Gruppe von gleichartigen Individuen fest, über die Aussagen getroffen werden sollen. Die festgelegte Menge heißt dann "***Grundgesamtheit***". Die Elemente der Menge sind die "***statistischen Individuen***".*

Beispiele für Grundgesamtheiten:

- Die Wahlberechtigten einer bestimmten Wahl.
- Die Einwohner eines Landkreises.
- Die Haushalte eines Landkreises.
- Die Studenten einer Vorlesung.
- Die Pflanzen, die im Rahmen einer Versuchsreihe behandelt wurden.
- Die Kundenkontakte eines Unternehmens.
- Die Käfer auf einer Wiese.
- Die Eichenarten Nordamerikas.
- Die 18 - 26-jährigen wehrtauglichen jungen Männer.
- Die Passanten, die innerhalb eines festgelegten Zeitraums ein spezifisches Einkaufszentrum betreten.
- ...

Verschiedene Personen sind zu verschiedenen Zeiten in unterschiedlichen Rollen und unter verschiedenen Aspekten an verschiedenen solcher Grundgesamtheiten interessiert.

Beispiel 6 *Daten laden:*

```
> help(morley)

morley                package:datasets                R
Documentation

Michaelson-Morley
Speed of Light Data

Description:

     The classical data of Michaelson and Morley on the speed
     of light.  The data consists of five experiments, each
     consisting of 20 consecutive 'runs'.  The response is
     the speed of light measurement, suitably coded.

Usage:

     morley

Format:
```

```
     A data frame contains the following components:

     'Expt' The experiment number, from 1 to 5.

     'Run' The run number within each experiment.

     'Speed' Speed-of-light measurement.

Details:

     The data is here viewed as a randomized block experiment
     with 'experiment' and 'run' as the factors.  'run' may
     also be considered a quantitative variate to account for
     linear (or polynomial) changes in the measurement over
     the course of a single experiment.

Source:

     A. J. Weekes (1986) _A Genstat Primer_. London: Edward
     Arnold.

Examples:

     require(stats) morley£Expt <- factor(morley£Expt)
     morley£Run <- factor(morley£Run) attach(morley) plot(Expt,
     Speed, main = "Speed of Light Data", xlab = "Experiment
     No.")  fm <- aov(Speed ~ Run + Expt, data = morley)
     summary(fm) fm0 <- update(fm, . ~ . - Run) anova(fm0,
     fm) detach(morley)
```

```
> data(morley)
> lightspeeds <- morley$Speed + 299000
```

Alle Messungen der Lichtgeschwindigkeit von Michelson:

```
> lightspeeds
  [1] 299850 299740 299900 300070 299930 299850 299950
  [8] 299980 299980 299880 300000 299980 299930 299650
 [15] 299760 299810 300000 300000 299960 299960 299960
 [22] 299940 299960 299940 299880 299800 299850 299880
 [29] 299900 299840 299830 299790 299810 299880 299880
 [36] 299830 299800 299790 299760 299800 299880 299880
 [43] 299880 299860 299720 299720 299620 299860 299970
 [50] 299950 299880 299910 299850 299870 299840 299840
 [57] 299850 299840 299840 299840 299890 299810 299810
 [64] 299820 299800 299770 299760 299740 299750 299760
 [71] 299910 299920 299890 299860 299880 299720 299840
 [78] 299850 299850 299780 299890 299840 299780 299810
 [85] 299760 299810 299790 299810 299820 299850 299870
 [92] 299870 299810 299740 299810 299940 299950 299800
 [99] 299810 299870
```

Der Datensatz enthält sämtliche mit einer bestimmten Methode durchgeführten Messungen der Lichtgeschwindigkeit. Er kann also als Grundgesamtheit der von Michelson durchgeführten Lichtgeschwindigkeitsmessungen aufgefasst werden.

```
> dotchart(lightspeeds, main = "Michelsons Speed of Light Data")
```

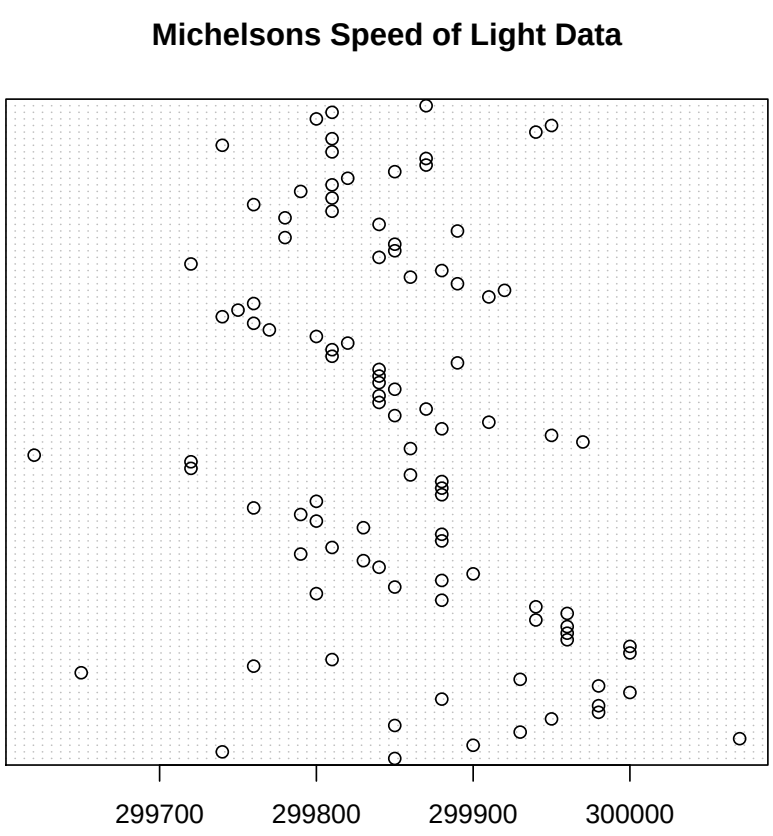


Abbildung 1.4: Ein Punktdiagramm der Lichtgeschwindigkeitsmessungen von Michelson .

1.2.3 Stichprobe

Typischerweise können die relevanten Informationen nicht von allen Mitgliedern einer Grundgesamtheit erhoben werden. Daher erhebt man die Information (z.B. "welche Partei die Person am nächsten Sonntag wählen würde") nur für eine kleinere Teilgruppe. Diese Teilgruppe heißt dann "**Stichprobe**". Die Stichprobe heißt "**repräsentativ**" für eine Grundgesamtheit, wenn jedes statistische Individuum die gleiche Chance hat in die Stichprobe zu gelangen und wenn seine Chancen hineinzugelangen unabhängig davon sind, welche Individuen noch in der Stichprobe sind. Manchmal wird zusätzlich gefordert, dass die Stichprobe hinreichend groß sei, damit von ihr auf die Grundgesamtheit gefolgert werden kann, das ist streng genommen jedoch nicht Teil der Defintion von "repräsentativ" für Stichproben.

Beispiel 7 (Stichprobenwahl im Computer) *Wir suchen aus Michelsons Experimenten einen Teil aus:*

```
> !"10 Messungen zufaellig auswaehlen"
> X <- sample(lightspeeds, 10)
> X

 [1] 299840 299790 300000 299770 299650 299810 299950 299840
 [9] 299980 299840

> !"Jemand anderes haette vielleicht andere 10 Messungen zufaellig auswaehlt"
> X <- sample(lightspeeds, 10)
> X

 [1] 299760 299800 299880 299810 299760 299940 299890 299820
 [9] 299810 299810
```

Der Befehl sample führt dabei eine zufällige Auswahl durch.

Diskussion 8 *Wir haben ein Modell gewählt, in dem die Messungen von Michelson die Grundgesamtheit bilden. In dieser Form können wir Aussagen treffen über die Michelson Experimente. Wir hätten auch alternativ die Menge aller denkbaren richtigen Lichtgeschwindigkeitsmessungsexperimente nehmen können, und so eine Aussage über die Lichtgeschwindigkeit an sich anstreben können. Frage:*

- *Ist der gesamte Datensatz dann eine Stichprobe aus dieser neuen Grundgesamtheit?*
- *Ist der gesamte Datensatz dann eine repräsentative Stichprobe aus dieser Grundgesamtheit?*

Beispiel 9 (Vorlesungsbefragung) *Ich frage eine Reihe von Studenten dieser Vorlesung, welches Fach sie studieren. Ich könnte ein paar Fehler machen, die mich daran hindern, ein realistisches Bild von den Anteilen zu bekommen:*

- *Ich frage einfach nur eine Teilgruppe, z.B. nur die Bachelorstudenten. Offenbar würde sich kein Diplomand in der Stichprobe befinden und ich erhalte ein verzerrtes Ergebnis (Verzerrung durch willkürliche Auswahl).*
- *Ich frage nur die, die zu früh kommen. Auch hier gibt es vermutlich Fachpräferenzen und ich erhalte ein unrealistisches Bild von der Fachverteilung (Verzerrung durch ungleiche Chancen).*
- *Ich frage nur die Studenten einer Sitzreihe. Vermutlich werde ich einen unrelalistisch hohen Anteil einer bestimmten Studienrichtung erhalten, da die Studenten sich oft neben solche setzen, die sie aus ihrem Studiengang näher kennen (Verzerrung durch abhängige Befragung).*

Wenn ich einen dieser Fehler begehe, ist meine Stichprobe nicht repräsentativ. Die Repräsentativität der Stichprobe wird von allen statistischen Verfahren vorausgesetzt. Sie kann nur sehr schwer überprüft werden.

Wenn ich nur einen Studenten befrage, werde ich ebenfalls keine Rückschlüsse ziehen können, einfach weil ich zu wenig Daten habe. Es ist eine wesentliche Aufgabe der Statistik, herauszufinden, ob die Menge der erhobenen Daten für die gewünschten Schlüsse ausreicht.

Ein erster Schritt jeder statistischen Analyse von Daten ist also herauszufinden, für welche Grundgesamtheit die Daten repräsentativ sind.

Der erste Schritt der Planung einer statistischen Erhebung ist immer die Festlegung der intendierten Grundgesamtheit.

Definition 10 *Die spezielle repräsentative Stichprobe, in der jedes Individuum in der Grundgesamtheit enthalten ist, heißt "***Vollerhebung***".*

1.2.4 Nichtrepräsentative statistische Daten

Es gibt viele Daten über gleichartige Individuen, die nicht repräsentativ für eine sinnvolle Grundgesamtheit, außer sich selbst, sind. Z.B. die Patientendaten eines Krankenhauses für einen bestimmten Zeitraum. Solche Daten werden oft im Sinne von Statistik (1) als Statistik bezeichnet, sie werden oft in Datentafeln (z.B. Abb. 1.1) dargestellt.

1.2.5 Statistisches Modell

1.2.5.1 Beobachtungen als Zufallsvariablen und Daten als ihre Realisierung

In der mathematischen Statistik werden die erhobenen Daten $x_1, x_2, \ldots, x_n$ oft durch **Zufallsvariablen** $X_1, X_2, \ldots, X_n$ modelliert.

Definition 11 *Eine* **Zufallsvariable** X *ist ein Modell für das Ergebnis eines Zufallsexperiments. Eine Zufallsvariable hat eine Wahrscheinlichkeitsverteilung* P^X*, die uns aber im allgemeinen nicht bekannt ist.*

Dabei beschreibt die in Großbuchstaben geschriebene Zufallsvariable X_i das Konzept der i-ten Beobachtung als zufälliger Wert und die als kleines x_i geschriebene Realisierung den konkret beobachteten Wert.

In R schreibt man eckige Klammer statt Indizes und verwendet den konkreten Datensatznamen (hier `lightspeeds`) statt X:

```
> lightspeeds

 [1] 299850 299740 299900 300070 299930 299850 299950
 [8] 299980 299980 299880 300000 299980 299930 299650
[15] 299760 299810 300000 300000 299960 299960 299960
[22] 299940 299960 299940 299880 299800 299850 299880
[29] 299900 299840 299830 299790 299810 299880 299880
[36] 299830 299800 299790 299760 299800 299880 299880
[43] 299880 299860 299720 299720 299620 299860 299970
[50] 299950 299880 299910 299850 299870 299840 299840
[57] 299850 299840 299840 299840 299890 299810 299810
[64] 299820 299800 299770 299760 299740 299750 299760
[71] 299910 299920 299890 299860 299880 299720 299840
[78] 299850 299850 299780 299890 299840 299780 299810
```

```
 [85] 299760 299810 299790 299810 299820 299850 299870
 [92] 299870 299810 299740 299810 299940 299950 299800
 [99] 299810 299870
```

```
> lightspeeds[3]
```

```
[1] 299900
```

Die Notation `lightspeeds[3]` bezeichnet also abstrakt den Ausgang X_3 des dritten Experiments der Stichprobe. Dieser ist konkret $x_3 = 299880$. `lightspeeds` und X selbst bezeichnen den Vektor $X = (X_1, X_2, \ldots, X_n)$ aller Beobachtungen.

1.2.5.2 Verteilung

Die Realisierung einer Zufallsvariablen X erfolgt durch eine Zufallsverteilung P^X. Der Zufall kann dabei entstehen durch:

- zufällige Auswahl aus der Grundgesamtheit
- zufällige Messfehler
- zufällige Vorgänge (z.B. biologische Vorgänge)
- chaotische Komponenten in komplexen Systemen (z.B. Wirtschaftsvorgänge)
- nicht berücksichtigte Variation in der Grundgesamtheit (z.B. Einkommensunterschiede)

Oft treten mehrere Zufallsquellen kombiniert auf, und die Zuordnung des Zufalls zu einem Typ mag philosophisch umstritten sein.

Der Zufall wird durch eine Wahrscheinlichkeitsverteilung modelliert. Die für die einfache Statistik wichtigste Verteilungsklasse ist die Normalverteilung.

Die Normalverteilung ist eine Verteilungsklasse mit 2 Parametern und wird mit $N(\mu, \sigma^2)$ abgekürzt. Der erste Parameter μ bzw. `mean` gibt den mittleren Wert an, um den die Werte streuen. Der zweite Parameter, die Standardabweichung σ bzw. `sd` gibt an, wie stark die Daten streuen. Mit dem folgenden Befehl simuliert der Computer unabhängig voneinander $n = 100$ Zufallswerte mit dieser Verteilung.

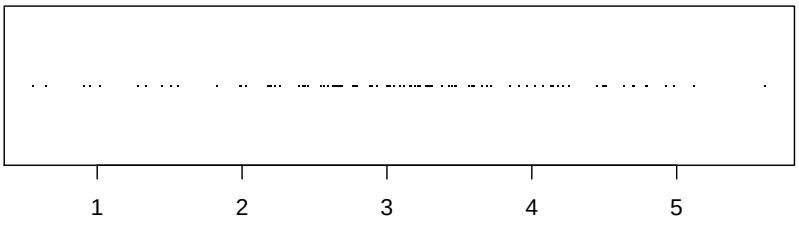
```
> X = rnorm(n = 100, mean = 3, sd = 1)
> X
```

```
  [1] 4.1491 1.9841 2.6828 2.6679 3.8474 4.0797 4.4993
  [8] 3.1608 2.6893 2.0249 2.6621 2.5680 4.9793 3.5981
 [15] 2.2294 2.4549 5.1176 3.2755 4.9255 3.5970 3.0476
 [22] 2.6290 2.1920 4.4940 3.0890 0.9481 1.9926 3.3777
 [29] 2.3940 3.6540 3.4786 3.0496 2.4308 3.2136 4.1334
 [36] 2.1852 3.2896 4.5111 2.1762 4.2551 2.7632 4.1779
 [43] 1.2824 3.9687 2.4228 1.5068 2.2614 3.2316 1.4457
 [50] 2.4288 3.7169 2.8818 4.7063 3.1950 2.4309 0.6451
 [57] 4.4491 2.7772 4.7951 3.2682 1.5603 3.0476 1.0200
 [64] 4.0238 5.6111 3.3103 4.7884 3.3053 2.5468 2.6462
 [71] 1.3351 2.2001 2.6517 3.1667 3.4481 3.0030 3.6886
 [78] 3.4712 3.2891 2.7900 3.5673 3.9110 3.1144 3.4271
 [85] 0.9061 2.6414 3.0215 4.2175 2.5897 2.9331 3.5888
 [92] 3.2696 2.8983 4.6960 4.6967 0.5601 1.8293 4.6348
 [99] 2.6602 3.0175
```

`rnorm` steht dabei für **r**andom **norm**al.

Um das zu verstehen, stellen wir die Daten graphisch dar:

```
> stripchart(X, pch = ".")
```



Obwohl jede Zufallsvariable `X[i]` die gleiche Verteilung hat, ergeben sich doch unterschiedliche Werte. Unser Ziel ist es jedoch, nicht den Einzelnen zu beschreiben, sondern die Gesamtheit. Wir wollen also z.B. aus den Daten auf das μ und σ zurückschließen.

1.2.6 Verteilungsfunktion

Eine Wahrscheinlichkeitsverteilung kann durch eine Verteilungsfunktion oder eine Dichtefunktion beschrieben werden:

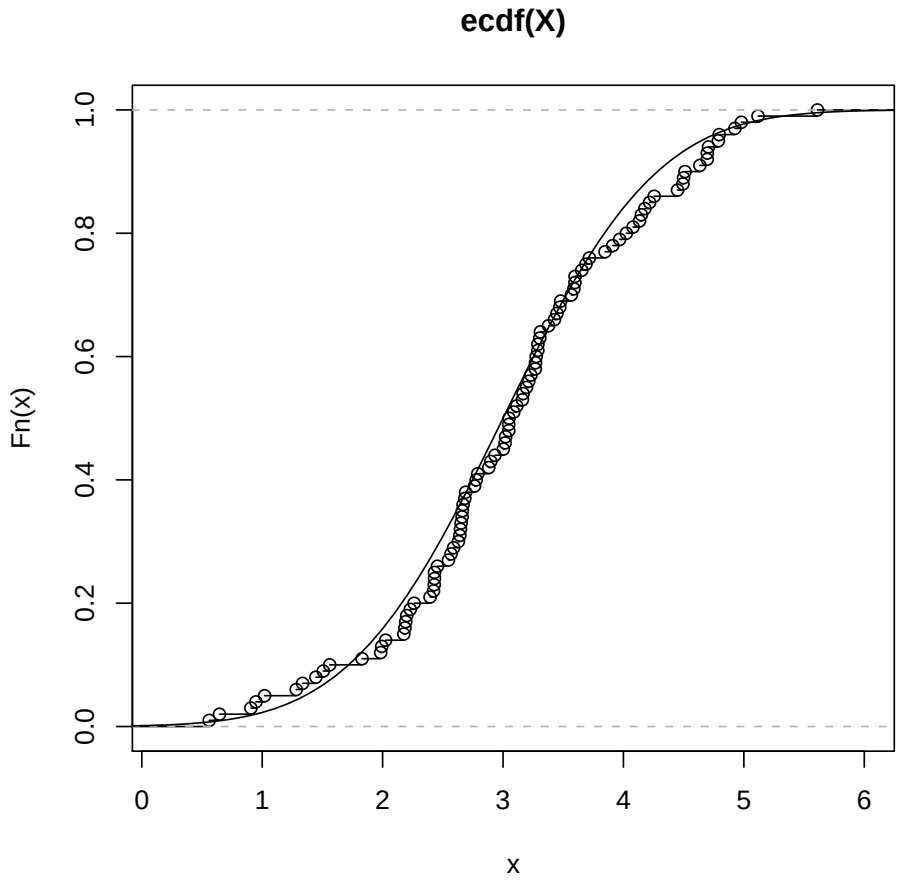
Die **Verteilungsfunktion** $F_X(x)$ ist gegeben durch die Wahrscheinlichkeit, dass X den Wert x nicht übersteigt:

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

```
> plot(ecdf(X))
> x <- seq(-2, 7, by = 0.1)
> x
```

```
 [1] -2.0 -1.9 -1.8 -1.7 -1.6 -1.5 -1.4 -1.3 -1.2 -1.1 -1.0
[12] -0.9 -0.8 -0.7 -0.6 -0.5 -0.4 -0.3 -0.2 -0.1  0.0  0.1
[23]  0.2  0.3  0.4  0.5  0.6  0.7  0.8  0.9  1.0  1.1  1.2
[34]  1.3  1.4  1.5  1.6  1.7  1.8  1.9  2.0  2.1  2.2  2.3
[45]  2.4  2.5  2.6  2.7  2.8  2.9  3.0  3.1  3.2  3.3  3.4
[56]  3.5  3.6  3.7  3.8  3.9  4.0  4.1  4.2  4.3  4.4  4.5
[67]  4.6  4.7  4.8  4.9  5.0  5.1  5.2  5.3  5.4  5.5  5.6
[78]  5.7  5.8  5.9  6.0  6.1  6.2  6.3  6.4  6.5  6.6  6.7
[89]  6.8  6.9  7.0
```

```
> lines(x, pnorm(x, mean = 3, sd = 1))
```



Die Rechtswertachse stellt dabei den Wert x dar und die Hochwertachse den Wert $F(x)$, also die Wahrscheinlichkeit kleiner als dieser Wert zu sein. Der durchgezogene Funktionsgraph stellt die Verteilungsfunktion $F_X(x)$ für die $N(3,1)$ Verteilung dar.[1] Die Stufenlinie stellt den Anteil der Werte, die kleiner als x sind, in der simulierten Stichprobe dar:

$$\hat{F}(x) := \text{Anteil Beobachtungswerte } x_i \text{ kleiner oder gleich x}$$

Wie man sieht, passen beide Kurven relative gut zueinander. Die Stufenfunktion $\hat{F}(x)$ heißt daher auch **empirische Verteilungsfunktion**, weil sie der Verteilungsfunktion sehr ähnlich ist und aus den Daten empirisch ermittelt werden kann.[2] Im Gegensatz dazu bezeichnet man $F_X(x)$ selbst auch oft als **theoretische Verteilungsfunktion**, weil die tatsächliche Verteilung ja normalerweise unbekannt ist und nur theoretisch vorausgesetzt wird.

1.2.6.1 Dichtefunktion

Die Verteilungsfunktion ist eher ein theoretisches Hilfsmittel, da sie immer ähnlich aussieht. Viel wichtiger ist ihre Ableitung, die Dichtefunktion:

$$f_X(x) = F'_X(x)$$

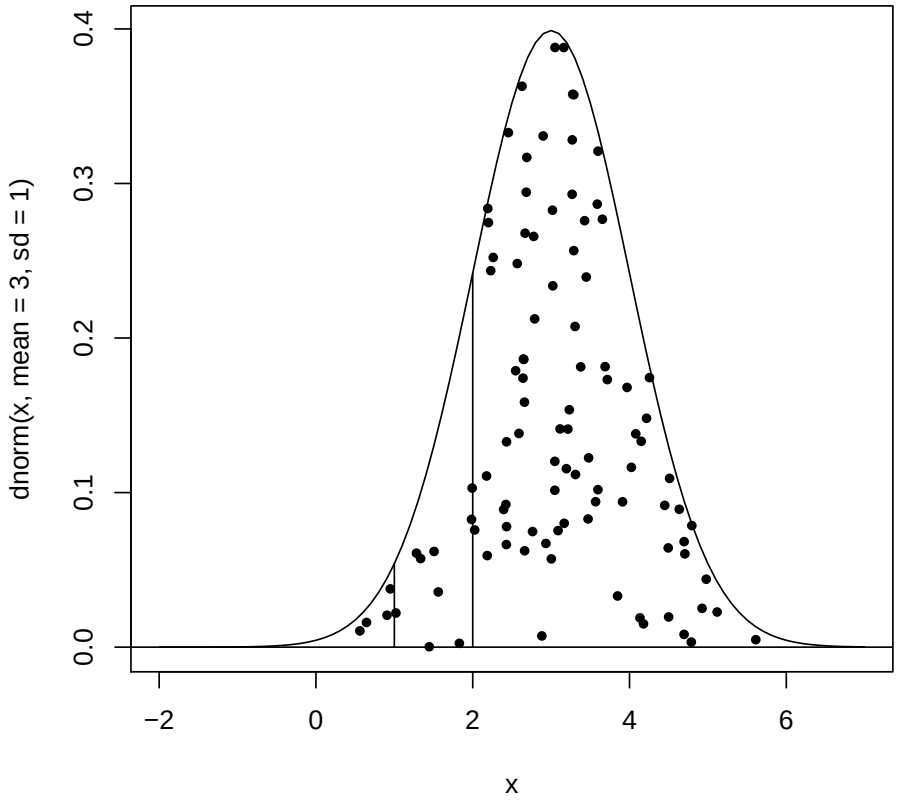
[1] Die Abkürzung `pnorm` steht hier für **p**robabiliy function of **norm**al distribution.

[2] Die Abkürzung `ecdf` steht für **e**mpirical **c**ummulated **d**istributin **f**unction, was die englische Übersetzung von empirische Verteilungsfunktion ist.

```
> plot(x, dnorm(x, mean = 3, sd = 1), type = "l")
```

Um das zu verstehen, zeichnen wir noch einige Hilfslinien ein:

```
> points(X, runif(length(X)) * dnorm(X, mean = 3,
+     sd = 1), pch = 20)
> segments(c(1, 2), c(0, 0), c(1, 2), dnorm(c(1,
+     2), mean = 3, sd = 1))
> abline(h = 0)
```

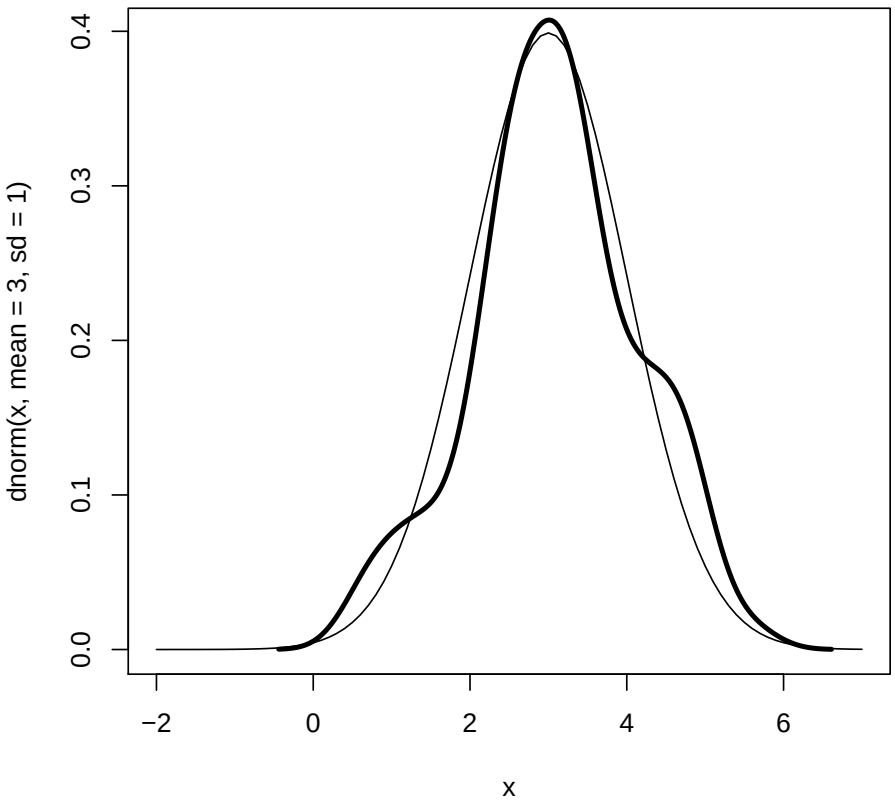


Dabei stellt die Hochwertachse nun, anstelle der Wahrscheinlichkeit, die Dichte (Wahrscheinlichkeit pro Rechtswertachsen-Einheit) dar.

Die Fläche in einem beliebigen senkrechten Streifen (also das Integral), entspricht der Wahrscheinlichkeit einen Wert in diesem Bereich zu realisieren. Man kann sich eine zufällige Simulation dieser Verteilung also so vorstellen: Zunächst wird gleichverteilt auf eine zufällige Stelle, in der Fläche unter der Kurve ein Punkt gestreut und dann wird dessen Rechtswert als Simulationswert verwendet.

Wie die Verteilungsfunktion kann man auch die Dichtefunktion (mit dem Befehl `density`) aus den Daten schätzen lassen:

```
> plot(x, dnorm(x, mean = 3, sd = 1), type = "l")
> lines(density(X), lwd = 3)
```



Allerdings erhält man oft nur eine grobe Annäherung an die tatsächliche Dichtefunktion.

1.2.6.2 Kenngrößen als Realisierung einer Zufallsvariablen

Das erste Ergebnis einer statistischen Analyse sind oft Kenngrößen, wie z.B. der Mittelwert.

Formeln für Kenngrößen, wie z.B. den Mittelwert, können dann in großen oder kleinen Buchstaben angegeben werden. Also:

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

wenn man betonen will, dass der Mittelwert $\bar{X}$ vom zufälligen Ausgang des Experiments abhängt und selbst eine Zufallsgröße ist.

Andererseits schreibt man kleine Buchstaben:

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_n = x_1 + x_2 + \ldots + x_n$$

wenn man betonen möchte, dass eine konkrete Realisierung des Mittelwertes (also eine konkrete Zahl) aus den tatsächlich beobachteten Werten ausgerechnet wird. $\bar{x}$ ist dann also der tatsächlich beobachtete Wert der Zufallsvariable $\bar{X}$.

Eine der Aufgaben der Statistiksoftware ist es, solche Kenngrößen tatsächlich auszurechnen.

```
> mean(lightspeeds)

[1] 299852
```

Unsere Aufgabe ist es, die Ergebnisse richtig zu interpretieren. Dazu kommt es vor allen Dingen darauf an, die zugrunde liegenden Modellvorstellungen zu verstehen und mit der Realität in Beziehung zu setzen.

Dazu müssen wir uns immer im Klaren sein, dass auch diese Kenngrößen vom Zufall abhängen. Der folgende Befehl simuliert 100 solche Datenpunkte und berechnet dann deren Mittelwert:

```
> mean(rnorm(n = 100, mean = 3, sd = 1))

[1] 2.929
```

Der folgende Befehl wiederholt uns diesen Vorgang 20 mal und gibt alle erhaltenen Mittelwerte aus:

```
> replicate(20, mean(rnorm(n = 100, mean = 3, sd = 1)))

 [1] 2.892 2.943 2.935 2.917 3.081 3.068 2.920 3.004 3.187
[10] 2.849 3.006 2.871 3.114 3.118 2.865 2.949 2.983 2.992
[19] 2.896 2.873
```

Zwar sind die erhaltenen Mittelwerte sehr ähnlich, aber doch etwas unterschiedlich und weiterhin zufällig. Auch Kenngrößen sind also zufällig.

1.2.6.3 Wahrscheinlichkeitsmodell

Da wir aus unseren Daten auf die dahinter liegenden Wahrscheinlichkeitsverteilungen schließen wollen, müssen alle X_i auch den gleichen Wahrscheinlichkeitsgesetzen folgen, und diese müssen den zu untersuchenden Gesetzen entsprechen:

$$P^{X_i} = P^{X_j} = P$$

Sind nicht alle P^{X_i} gleich, so wird das Ergebnis beliebig.
Mathematisch bedeutet das, dass die Zufallsvariablen **identisch verteilt** sein müssen.

Beispiel 12 Identische Verteilung

- Das geht nicht!
  Ich messe die Größe von n Erwachsenen und m Kindergartenkindern. Die mittlere Größe sagt weder etwas darüber aus, wie groß Erwachsene sind, noch wie groß Kindergartenkinder sind.
- Aber das geht!
  Ich messe die Größe von $n+m$ zufällig ausgewählten Menschen. Das Ergebnis sagt mir etwas über die mittlere Größe von Menschen.

Entsprechen die Verteilungen nicht der zu untersuchenden Verteilung, so hat man keine Chance, von den Daten auf die Verteilung zu schließen.
Dieser Aspekt wird mathematisch nicht abgebildet, da er ja etwas mit der Identität von Modell und Wirklichkeit zu tun hat.

Beispiel 13 Gleich wahrscheinliche Auswahl

- Das geht nicht! (Repräsentativ für die falsche Grundgesamtheit)
  Sie wollen die Durchschnittsgröße von Menschen bestimmen. Sie untersuchen eine zufällige Auswahl von Greifswaldern. (Menschen aus reichen Ländern sind im Durchschnitt größer.)

- Das geht! (Repräsentativ für eine kleinere Grundgesamtheit)
  Sie wollen die Durchschnittsgröße von Greifswaldern bestimmen. Sie untersuchen eine zufällige Auswahl von Greifswaldern.

- Das geht auch! (Repräsentativ für eine künstliche Grundgesamtheit)
  Sie wollen beweisen, dass ein Schlafmittel wirkt. Dazu beweisen Sie, dass die zufällige Auswahl Ihrer Testpersonen im Durchschnitt eine Stunde länger schläft als die übrigen Testpersonen. Damit ist natürlich die Wirkung erst einmal nur für die Gruppe der Testpersonen bewiesen, aber damit ist man sicher, dass das Mittel zumindest überhaupt wirkt.

Hängen die Zufallseinflüsse voneinander ab, so bekommt man ein beliebiges Ergebnis.
Mathematisch heißt das: die Beobachtungen X_i müssen **stochastisch unabhängig** sein.

Beispiel 14 Unabhängige Auswahl

- Das geht nicht! (Abhängige Auswahl)
  Zunächst eine Vorüberlegung:
  Angenommen, ich habe Leute, die zu einer bestimmten Zeit an einem bestimmten Ort waren, nach ihrem Familienstand befragt: Mittags um 12:00 vor der Mensa oder zur Babyschwimmstunde vor dem Freizeitbad, morgens zu Schulbeginn vor dem Humboldtgymnasium und dann vor dem katholischen Altersheim. Jedes Mal werden wir einen anderen Eindruck von den Anteilen der verschiedenen Gruppen bekommen.
  Und jetzt nehmen wir einfach einmal an, wir suchen uns einen zufälligen Platz aus und befragen dort die Leute. Wir können nicht ahnen, wer warum hier ist und deshalb auch nicht sagen, welche Antwort zu häufig ist. Aber wir können natürlich auch nicht ausschließen, dass wir in genau so eine Situation geraten sind, in der die eine oder andere Antwort bevorzugt wird. Mit einer solchen Studie lassen sich also keine allgemeinen Aussagen treffen.

- Das geht! (Definierte Auswahl)
  Wir erklären einfach die Passanten in einem gewissen Einkaufszentrum zu unserer Grundgesamtheit und untersuchen dann diese Grundgesamtheit. Ob das Sinn macht, hängt natürlich davon ab, ob wir die Wirksamkeit einer Werbekampagne untersuchen oder den Standort für das nächste Altersheim.

- Das wünscht man sich:
  Wir führen tatsächlich unabhängige Wiederholungen eines Zufallsexperiments durch oder wählen aus unserer Grundgesamtheit tatsächlich zufällig und unabhängig aus.

- Die Realität ist leider:
  Man schafft es nicht die Grundgesamtheit zu beproben, die man will, oder die Experimente vollständig unabhängig zu wiederholen. Man versucht es so gut wie möglich hinzubekommen, bleibt sich der Gefahr bewusst und nimmt seine Ergebnisse nicht zu ernst.

Insgesamt kommt man also zu der Vorstellung, dass die X_i unabhängig identisch verteilt sind. Das kürzt man oft mit i.i.d. für **i**ndependent **i**dentically **d**istributed ab.

Die X_i müssen i.i.d. sein.
Dazu muss die Stichprobe repräsentativ sein.
Das ist die Generalvoraussetzung der gesamten "einfachen" Statistik.

- *Für Prüfungsbesteher*
  Vergessen Sie nicht diese Voraussetzung zu erwähnen, wenn in der Prüfung nach Voraussetzungen gefragt wird.
- *Für Wissenschaftler.*
  Vergessen Sie nicht diese Voraussetzung zu schaffen und zu prüfen, wenn Sie etwas statistisch untersuchen.
- *Für BWL-ler*
  Fragen Sie Ihren Statistiker, ob diese Voraussetzung auch erfüllt war. Wenn nein, vergessen Sie die Studie.
- *Für Betrüger*
  Machen Sie doch einfach eine nichtrepräsentative Studie. Das merkt keiner.

Um diese Voraussetzung einzuführen, schreiben wir:

Die X_i seien i.i.d. P verteilt,

wobei P die Verteilung repräsentiert.

Z.B. sind die X_i aus der obigen Simulation i.i.d. $N(3,1)$ verteilt.

1.2.6.4 Das statistische Modell

Das Ziel der Statistik ist es, von den Daten auf die Verteilung zu schließen. Das bedeutet natürlich, dass die Verteilung ganz oder teilweise unbekannt sein muss, da man sich das Ganze ja sonst sparen könnte.

Man betrachtet dabei drei Situationen:

- *Parametrische Statistik*
  In der **parametrischen Statistik** ist die Verteilungsform grundsätzlich bekannt. Nur gewisse **Parameter** der Verteilung sind unbekannt.

  Beispiel 15 (Typisches Modell der parametrischen Statistik) *Die X_i sind i.i.d. $N(\mu, \sigma^2)$ verteilt, aber wir kennen die Parameter μ und σ^2 nicht.*

  Zentrale Begriffe der parametrischen Statistik sind: Normalverteilung, Mittelwert, Varianz, Poissonverteilung, Bernouli-Verteilung, Kontingenztafeln.

  Die parametrische Statistik liefert im Allgemeinen die leistungsfähigeren Verfahren, so dass sie immer dann angewendet werden sollte, wenn es geht, also wenn die Verteilungsform bekannt ist.

- *Die Nichtparametrische Statistik*
  In der **nichtparametrischen Statistik** geht man davon aus, dass nicht einmal die Verteilungsform bekannt ist. Um dann aber überhaupt noch etwas machen zu können, braucht man die Stetigkeit der Verteilungsfunktion.

  Beispiel 16 (Typisches Modell der nichtparametrischen Statistik) *Die X_i sind i.i.d. stetig verteilt.*

Dabei bedeutet stetig verteilt, dass die Verteilungsfunktion $F_X(x)$ stetig ist.

Offenbar ist das eine viel schwächere Voraussetzung und die Verfahren der nichtparametrischen Statistik sind daher um einiges komplizierter, schwieriger zu verstehen und weniger wirksam. Nicht für alle Situationen sind überhaupt nichtparametrische Verfahren bekannt.

- *Die robuste Statistik*
  Die **robuste Statistik** steht zwischen der parametrischen und der nichtparametrischen Statistik.

  Sie ist motiviert durch die Beobachtung, dass in vielen Datensätzen ein Teil der Daten durch Fehler, z.B. durch verunreinigte Proben, Abschreibefehler, fälschlich in die Stichprobe aufgenommene Werte, und auch atypische Werte, die nicht zum gewählten Wahrscheinlichkeitsmodell passen, verfälscht ist. Diese Werte sind insbesondere dann problematisch, wenn sie von den anderen Werten erheblich abweichen. Die robuste Statistik versucht den Einfluss dieser Fehler klein zu halten.

  Einzelne Beobachtungswerte, die von den übrigen Daten erheblich abweichen, bezeichnet man auch als **Ausreißer**. Ein falscher Wert muss jedoch kein Ausreißer sein und ein Ausreißer kein falscher Wert.

  Dazu wird zwar ein parametrisches Modell vorausgesetzt, aber die Verfahren werden so modifiziert, dass ein gewisser Anteil der Beobachtungen von diesem Modell abweichen darf, ohne die Ergebnisse der Analyse zu stark zu verfälschen.

  Beispiel 17 (Typisches Modell der robusten Statistik) *Die X_i sind i.i.d. $N(\mu, \sigma^2)$ verteilt, aber wir kennen die Parameter μ und σ^2 nicht und ein Anteil von $b = 25\%$ der Beobachtungen kann völlig falsch sein.*

  Die Obergrenzen für den Anteil falscher Daten bezeichnet man auch als Bruchpunkt b des Verfahrens. Die robuste Statistik ist noch aufwendiger als die beiden vorhergenannten Verfahren und wird daher in dieser Vorlesung nur am Rande eine Rolle spielen können, wo entsprechende, robuste Verfahren in der Software direkt verfügbar sind und keine eigenen Konzepte benötigen. Als einfacher Anwender kann man oft robuste Verfahren durch nichtparametrische Verfahren ersetzen.

1.3 Datenmatrix

Eine der wichtigsten Darstellungsformen für statistische Daten ist die "Datenmatrix". Die Datenmatrix erlaubt es, mehrere Merkmale am gleichen statistischen Individuum zu beobachten.

Als Beispiel verwenden wir einen einfachen Datensatz:

> This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica.

Der Datensatz enthält 3 Stichproben von jeweils 50 Blumen aus drei Grundgesamtheiten. Die Grundgesamtheiten sind jeweils die im Labor verfügbaren Blumen der Arten "Iris setosa", "Iris versicolor" und "Iris virginica". Zu jeder Blume wurden jeweils die folgenden Informationen erhoben:

- Länge des Kelchblattes ("Sepal.Length")
- Breite des Kelchblattes ("Sepal.Width")
- Länge des Blütenblattes ("Petal.Length")
- Breite des Blütenblattes ("Petal.Width")
- Die Art der Blume ("Species")

Wir laden den Datensatz in "R".

```
> options(width = 70)
```

```
> data(iris)
> iris
```

```
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
4            4.6         3.1          1.5         0.2     setosa
5            5.0         3.6          1.4         0.2     setosa
6            5.4         3.9          1.7         0.4     setosa
7            4.6         3.4          1.4         0.3     setosa
8            5.0         3.4          1.5         0.2     setosa
9            4.4         2.9          1.4         0.2     setosa
10           4.9         3.1          1.5         0.1     setosa
11           5.4         3.7          1.5         0.2     setosa
12           4.8         3.4          1.6         0.2     setosa
13           4.8         3.0          1.4         0.1     setosa
14           4.3         3.0          1.1         0.1     setosa
15           5.8         4.0          1.2         0.2     setosa
16           5.7         4.4          1.5         0.4     setosa
17           5.4         3.9          1.3         0.4     setosa
18           5.1         3.5          1.4         0.3     setosa
19           5.7         3.8          1.7         0.3     setosa
20           5.1         3.8          1.5         0.3     setosa
21           5.4         3.4          1.7         0.2     setosa
22           5.1         3.7          1.5         0.4     setosa
23           4.6         3.6          1.0         0.2     setosa
24           5.1         3.3          1.7         0.5     setosa
25           4.8         3.4          1.9         0.2     setosa
26           5.0         3.0          1.6         0.2     setosa
27           5.0         3.4          1.6         0.4     setosa
28           5.2         3.5          1.5         0.2     setosa
29           5.2         3.4          1.4         0.2     setosa
30           4.7         3.2          1.6         0.2     setosa
31           4.8         3.1          1.6         0.2     setosa
32           5.4         3.4          1.5         0.4     setosa
33           5.2         4.1          1.5         0.1     setosa
34           5.5         4.2          1.4         0.2     setosa
35           4.9         3.1          1.5         0.2     setosa
36           5.0         3.2          1.2         0.2     setosa
37           5.5         3.5          1.3         0.2     setosa
38           4.9         3.6          1.4         0.1     setosa
```

39	4.4	3.0	1.3	0.2	setosa
40	5.1	3.4	1.5	0.2	setosa
41	5.0	3.5	1.3	0.3	setosa
42	4.5	2.3	1.3	0.3	setosa
43	4.4	3.2	1.3	0.2	setosa
44	5.0	3.5	1.6	0.6	setosa
45	5.1	3.8	1.9	0.4	setosa
46	4.8	3.0	1.4	0.3	setosa
47	5.1	3.8	1.6	0.2	setosa
48	4.6	3.2	1.4	0.2	setosa
49	5.3	3.7	1.5	0.2	setosa
50	5.0	3.3	1.4	0.2	setosa
51	7.0	3.2	4.7	1.4	versicolor
52	6.4	3.2	4.5	1.5	versicolor
53	6.9	3.1	4.9	1.5	versicolor
54	5.5	2.3	4.0	1.3	versicolor
55	6.5	2.8	4.6	1.5	versicolor
56	5.7	2.8	4.5	1.3	versicolor
57	6.3	3.3	4.7	1.6	versicolor
58	4.9	2.4	3.3	1.0	versicolor
59	6.6	2.9	4.6	1.3	versicolor
60	5.2	2.7	3.9	1.4	versicolor
61	5.0	2.0	3.5	1.0	versicolor
62	5.9	3.0	4.2	1.5	versicolor
63	6.0	2.2	4.0	1.0	versicolor
64	6.1	2.9	4.7	1.4	versicolor
65	5.6	2.9	3.6	1.3	versicolor
66	6.7	3.1	4.4	1.4	versicolor
67	5.6	3.0	4.5	1.5	versicolor
68	5.8	2.7	4.1	1.0	versicolor
69	6.2	2.2	4.5	1.5	versicolor
70	5.6	2.5	3.9	1.1	versicolor
71	5.9	3.2	4.8	1.8	versicolor
72	6.1	2.8	4.0	1.3	versicolor
73	6.3	2.5	4.9	1.5	versicolor
74	6.1	2.8	4.7	1.2	versicolor
75	6.4	2.9	4.3	1.3	versicolor
76	6.6	3.0	4.4	1.4	versicolor
77	6.8	2.8	4.8	1.4	versicolor
78	6.7	3.0	5.0	1.7	versicolor
79	6.0	2.9	4.5	1.5	versicolor
80	5.7	2.6	3.5	1.0	versicolor
81	5.5	2.4	3.8	1.1	versicolor
82	5.5	2.4	3.7	1.0	versicolor
83	5.8	2.7	3.9	1.2	versicolor
84	6.0	2.7	5.1	1.6	versicolor
85	5.4	3.0	4.5	1.5	versicolor
86	6.0	3.4	4.5	1.6	versicolor
87	6.7	3.1	4.7	1.5	versicolor
88	6.3	2.3	4.4	1.3	versicolor
89	5.6	3.0	4.1	1.3	versicolor
90	5.5	2.5	4.0	1.3	versicolor
91	5.5	2.6	4.4	1.2	versicolor
92	6.1	3.0	4.6	1.4	versicolor

93	5.8	2.6	4.0	1.2	versicolor
94	5.0	2.3	3.3	1.0	versicolor
95	5.6	2.7	4.2	1.3	versicolor
96	5.7	3.0	4.2	1.2	versicolor
97	5.7	2.9	4.2	1.3	versicolor
98	6.2	2.9	4.3	1.3	versicolor
99	5.1	2.5	3.0	1.1	versicolor
100	5.7	2.8	4.1	1.3	versicolor
101	6.3	3.3	6.0	2.5	virginica
102	5.8	2.7	5.1	1.9	virginica
103	7.1	3.0	5.9	2.1	virginica
104	6.3	2.9	5.6	1.8	virginica
105	6.5	3.0	5.8	2.2	virginica
106	7.6	3.0	6.6	2.1	virginica
107	4.9	2.5	4.5	1.7	virginica
108	7.3	2.9	6.3	1.8	virginica
109	6.7	2.5	5.8	1.8	virginica
110	7.2	3.6	6.1	2.5	virginica
111	6.5	3.2	5.1	2.0	virginica
112	6.4	2.7	5.3	1.9	virginica
113	6.8	3.0	5.5	2.1	virginica
114	5.7	2.5	5.0	2.0	virginica
115	5.8	2.8	5.1	2.4	virginica
116	6.4	3.2	5.3	2.3	virginica
117	6.5	3.0	5.5	1.8	virginica
118	7.7	3.8	6.7	2.2	virginica
119	7.7	2.6	6.9	2.3	virginica
120	6.0	2.2	5.0	1.5	virginica
121	6.9	3.2	5.7	2.3	virginica
122	5.6	2.8	4.9	2.0	virginica
123	7.7	2.8	6.7	2.0	virginica
124	6.3	2.7	4.9	1.8	virginica
125	6.7	3.3	5.7	2.1	virginica
126	7.2	3.2	6.0	1.8	virginica
127	6.2	2.8	4.8	1.8	virginica
128	6.1	3.0	4.9	1.8	virginica
129	6.4	2.8	5.6	2.1	virginica
130	7.2	3.0	5.8	1.6	virginica
131	7.4	2.8	6.1	1.9	virginica
132	7.9	3.8	6.4	2.0	virginica
133	6.4	2.8	5.6	2.2	virginica
134	6.3	2.8	5.1	1.5	virginica
135	6.1	2.6	5.6	1.4	virginica
136	7.7	3.0	6.1	2.3	virginica
137	6.3	3.4	5.6	2.4	virginica
138	6.4	3.1	5.5	1.8	virginica
139	6.0	3.0	4.8	1.8	virginica
140	6.9	3.1	5.4	2.1	virginica
141	6.7	3.1	5.6	2.4	virginica
142	6.9	3.1	5.1	2.3	virginica
143	5.8	2.7	5.1	1.9	virginica
144	6.8	3.2	5.9	2.3	virginica
145	6.7	3.3	5.7	2.5	virginica
146	6.7	3.0	5.2	2.3	virginica

```
147          6.3          2.5          5.0          1.9  virginica
148          6.5          3.0          5.2          2.0  virginica
149          6.2          3.4          5.4          2.3  virginica
150          5.9          3.0          5.1          1.8  virginica
```

`data(iris)` lädt den Beispieldatensatz. Schreiben wir nun einfach den Namen des geladenen Datensatzes (`iris`), so wird dieser als Datenmatrix ausgegeben.

Definition 18 (Datenmatrix)
*Eine "***Datenmatrix***" ist eine Darstellung der von den statistischen Individuen einer Stichprobe erhobenen gleichartigen Daten in einer Tabelle mit Zeilen und Spalten. Von jedem Individuum werden die gleichen Merkmale erhoben.*

- *Die Informationen zum gleichen Individuum werden in einer Zeile dargestellt. Die zu einer Zeile gehörenden Individuen bezeichnet man auch als* **Fälle**.
- *Die zum gleichen* **Merkmal** *der Individuen gehörenden Informationen werden in der gleichen Spalte dargestellt, so dass am Schnittpunkt einer Zeile und einer Spalte jeweils die entsprechende Information zum entsprechenden Individuum steht. Die Spalten bezeichnet man auch als* **Variable***; die Spaltenüberschrift als den Namen der Variable.*

In der mathematischen Betrachtung erhält also die Beobachtung X zwei Indizes:

$$X_{ij} = j\text{-te Beobachtung am } i\text{-ten Indiviuum}$$

Der **Zeilenindex** *i bezeichnet also das statistische Individuum und der* **Spaltenindex** *j die erhobene Information.*

Die Spalten $X_{\cdot j}$ der Matrix entsprechen also den Variablen und enthalten die Information zum j-ten Merkmal an allen Individuen bzw. zu allen Fällen.

Die Zeilen $X_{i\cdot}$ der Matrix entsprechen also allen Informationen zum i-ten Individuum oder i-ten Fall.

Im Allgemeinen geht man davon aus, dass Informationen zu verschiedenen statistischen Individuen stochastisch unabhängig sind. Daher geht man davon aus, dass die Informationen in verschiedenen Zeilen der Datenmatrix voneinander unabhängig sind. Andererseits sind die Informationen in der gleichen Zeile natürlich alle voneinander abhängig, da sie vom gleichen Individuum stammen. Die Generalvoraussetzung der Unabhängigkeit bezieht sich also immer auf die Unabhängigkeit der Zeilen.

1.3.0.4.1 R: Zugriff auf Datenmatrizen in R

Informationen zu Datenmatrizen erhält man in R mittels der folgenden Befehle.
Name der Variablen:

```
> names(iris)
```

```
[1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"
[5] "Species"
```

Ist das eine Datenmatrix (`data.frame` in R):

```
> class(iris)
```

```
[1] "data.frame"
```

Zeilen und Spaltenzahl der Datenmatrix:

```
> dim(iris)
```

```
[1] 150   5
```

In "R" kann auf die Information einer einzelnen Tabellenzelle mittels des folgenden Syntax zugegriffen werden:

$$Datensatz[Zeilennummer,Spaltennummer]$$

Also beispielsweise:

```
> iris[1, 3]

[1] 1.4

> iris[1, 5]

[1] setosa
Levels: setosa versicolor virginica

> iris[51, "Species"]

[1] versicolor
Levels: setosa versicolor virginica

> iris["23", "Sepal.Length"]

[1] 4.6
```

Wie man sieht, kann dabei die Nummer auch durch den Namen ersetzt werden, der dann in doppelte Anführungszeichen zu setzen ist.

Eine ganze Zeile erhält man indem man die Angabe der Spalte einfach weglässt:

```
> iris[1, ]

  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa

> iris[51, ]

   Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
51            7         3.2          4.7         1.4 versicolor

> iris["23", ]

   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
23          4.6         3.6            1         0.2  setosa
```

für ganze Spalten gibt es noch einen Sondersyntax:

$$Datensatz\$Variablenname$$

so dass die folgenden Angaben verwendet werden können:

```
> iris$Sepal.Length

  [1] 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 4.4 4.9 5.4 4.8 4.8 4.3 5.8 5.7
 [17] 5.4 5.1 5.7 5.1 5.4 5.1 4.6 5.1 4.8 5.0 5.0 5.2 5.2 4.7 4.8 5.4
 [33] 5.2 5.5 4.9 5.0 5.5 4.9 4.4 5.1 5.0 4.5 4.4 5.0 5.1 4.8 5.1 4.6
 [49] 5.3 5.0 7.0 6.4 6.9 5.5 6.5 5.7 6.3 4.9 6.6 5.2 5.0 5.9 6.0 6.1
 [65] 5.6 6.7 5.6 5.8 6.2 5.6 5.9 6.1 6.3 6.1 6.4 6.6 6.8 6.7 6.0 5.7
 [81] 5.5 5.5 5.8 6.0 5.4 6.0 6.7 6.3 5.6 5.5 5.5 6.1 5.8 5.0 5.6 5.7
 [97] 5.7 6.2 5.1 5.7 6.3 5.8 7.1 6.3 6.5 7.6 4.9 7.3 6.7 7.2 6.5 6.4
[113] 6.8 5.7 5.8 6.4 6.5 7.7 7.7 6.0 6.9 5.6 7.7 6.3 6.7 7.2 6.2 6.1
[129] 6.4 7.2 7.4 7.9 6.4 6.3 6.1 7.7 6.3 6.4 6.0 6.9 6.7 6.9 5.8 6.8
[145] 6.7 6.7 6.3 6.5 6.2 5.9

> iris$Species
```

```
  [1] setosa     setosa     setosa     setosa     setosa
  [6] setosa     setosa     setosa     setosa     setosa
 [11] setosa     setosa     setosa     setosa     setosa
 [16] setosa     setosa     setosa     setosa     setosa
 [21] setosa     setosa     setosa     setosa     setosa
 [26] setosa     setosa     setosa     setosa     setosa
 [31] setosa     setosa     setosa     setosa     setosa
 [36] setosa     setosa     setosa     setosa     setosa
 [41] setosa     setosa     setosa     setosa     setosa
 [46] setosa     setosa     setosa     setosa     setosa
 [51] versicolor versicolor versicolor versicolor versicolor
 [56] versicolor versicolor versicolor versicolor versicolor
 [61] versicolor versicolor versicolor versicolor versicolor
 [66] versicolor versicolor versicolor versicolor versicolor
 [71] versicolor versicolor versicolor versicolor versicolor
 [76] versicolor versicolor versicolor versicolor versicolor
 [81] versicolor versicolor versicolor versicolor versicolor
 [86] versicolor versicolor versicolor versicolor versicolor
 [91] versicolor versicolor versicolor versicolor versicolor
 [96] versicolor versicolor versicolor versicolor versicolor
[101] virginica  virginica  virginica  virginica  virginica
[106] virginica  virginica  virginica  virginica  virginica
[111] virginica  virginica  virginica  virginica  virginica
[116] virginica  virginica  virginica  virginica  virginica
[121] virginica  virginica  virginica  virginica  virginica
[126] virginica  virginica  virginica  virginica  virginica
[131] virginica  virginica  virginica  virginica  virginica
[136] virginica  virginica  virginica  virginica  virginica
[141] virginica  virginica  virginica  virginica  virginica
[146] virginica  virginica  virginica  virginica  virginica
Levels: setosa versicolor virginica
```

```
> iris[, 1]
```

```
  [1] 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 4.4 4.9 5.4 4.8 4.8 4.3 5.8 5.7
 [17] 5.4 5.1 5.7 5.1 5.4 5.1 4.6 5.1 4.8 5.0 5.0 5.2 5.2 4.7 4.8 5.4
 [33] 5.2 5.5 4.9 5.0 5.5 4.9 4.4 5.1 5.0 4.5 4.4 5.0 5.1 4.8 5.1 4.6
 [49] 5.3 5.0 7.0 6.4 6.9 5.5 6.5 5.7 6.3 4.9 6.6 5.2 5.0 5.9 6.0 6.1
 [65] 5.6 6.7 5.6 5.8 6.2 5.6 5.9 6.1 6.3 6.1 6.4 6.6 6.8 6.7 6.0 5.7
 [81] 5.5 5.5 5.8 6.0 5.4 6.0 6.7 6.3 5.6 5.5 5.5 6.1 5.8 5.0 5.6 5.7
 [97] 5.7 6.2 5.1 5.7 6.3 5.8 7.1 6.3 6.5 7.6 4.9 7.3 6.7 7.2 6.5 6.4
[113] 6.8 5.7 5.8 6.4 6.5 7.7 7.7 6.0 6.9 5.6 7.7 6.3 6.7 7.2 6.2 6.1
[129] 6.4 7.2 7.4 7.9 6.4 6.3 6.1 7.7 6.3 6.4 6.0 6.9 6.7 6.9 5.8 6.8
[145] 6.7 6.7 6.3 6.5 6.2 5.9
```

```
> iris[, 5]
```

```
  [1] setosa     setosa     setosa     setosa     setosa
  [6] setosa     setosa     setosa     setosa     setosa
 [11] setosa     setosa     setosa     setosa     setosa
 [16] setosa     setosa     setosa     setosa     setosa
 [21] setosa     setosa     setosa     setosa     setosa
 [26] setosa     setosa     setosa     setosa     setosa
 [31] setosa     setosa     setosa     setosa     setosa
 [36] setosa     setosa     setosa     setosa     setosa
 [41] setosa     setosa     setosa     setosa     setosa
 [46] setosa     setosa     setosa     setosa     setosa
 [51] versicolor versicolor versicolor versicolor versicolor
 [56] versicolor versicolor versicolor versicolor versicolor
 [61] versicolor versicolor versicolor versicolor versicolor
 [66] versicolor versicolor versicolor versicolor versicolor
```

```
 [71] versicolor versicolor versicolor versicolor versicolor
 [76] versicolor versicolor versicolor versicolor versicolor
 [81] versicolor versicolor versicolor versicolor versicolor
 [86] versicolor versicolor versicolor versicolor versicolor
 [91] versicolor versicolor versicolor versicolor versicolor
 [96] versicolor versicolor versicolor versicolor versicolor
[101] virginica  virginica  virginica  virginica  virginica
[106] virginica  virginica  virginica  virginica  virginica
[111] virginica  virginica  virginica  virginica  virginica
[116] virginica  virginica  virginica  virginica  virginica
[121] virginica  virginica  virginica  virginica  virginica
[126] virginica  virginica  virginica  virginica  virginica
[131] virginica  virginica  virginica  virginica  virginica
[136] virginica  virginica  virginica  virginica  virginica
[141] virginica  virginica  virginica  virginica  virginica
[146] virginica  virginica  virginica  virginica  virginica
Levels: setosa versicolor virginica
```

Wünscht man nur einen Teil des Datensatzes, kann man das erreichen, indem man die einzelnen Zeilen- oder Spaltennummern durch eine Bereichsangabe der Form:

$$von\texttt{:}bis$$

oder den Index durch Aufzählung ersetzt:

$$\texttt{c(}name1\texttt{,}name2\texttt{,...)}$$

so dass man z.B. erhält:

```
> iris[1:10, c("Sepal.Length", "Sepal.Width")]
```

```
   Sepal.Length Sepal.Width
1           5.1         3.5
2           4.9         3.0
3           4.7         3.2
4           4.6         3.1
5           5.0         3.6
6           5.4         3.9
7           4.6         3.4
8           5.0         3.4
9           4.4         2.9
10          4.9         3.1
```

Teildatensätze können auch mit einer Bedingung selektiert werden:

```
> iris[iris$Species == "versicolor", c("Sepal.Width", "Species")]
```

```
    Sepal.Width    Species
51          3.2 versicolor
52          3.2 versicolor
53          3.1 versicolor
54          2.3 versicolor
55          2.8 versicolor
56          2.8 versicolor
57          3.3 versicolor
58          2.4 versicolor
59          2.9 versicolor
60          2.7 versicolor
61          2.0 versicolor
62          3.0 versicolor
63          2.2 versicolor
64          2.9 versicolor
65          2.9 versicolor
```

```
66          3.1 versicolor
67          3.0 versicolor
68          2.7 versicolor
69          2.2 versicolor
70          2.5 versicolor
71          3.2 versicolor
72          2.8 versicolor
73          2.5 versicolor
74          2.8 versicolor
75          2.9 versicolor
76          3.0 versicolor
77          2.8 versicolor
78          3.0 versicolor
79          2.9 versicolor
80          2.6 versicolor
81          2.4 versicolor
82          2.4 versicolor
83          2.7 versicolor
84          2.7 versicolor
85          3.0 versicolor
86          3.4 versicolor
87          3.1 versicolor
88          2.3 versicolor
89          3.0 versicolor
90          2.5 versicolor
91          2.6 versicolor
92          3.0 versicolor
93          2.6 versicolor
94          2.3 versicolor
95          2.7 versicolor
96          3.0 versicolor
97          2.9 versicolor
98          2.9 versicolor
99          2.5 versicolor
100         2.8 versicolor
```

Man beachte das doppelte Gleichheitszeichen in Vergleichen.

1.3.1 Abhängigkeit und Unabhängigkeit in der Datenmatrix

Die Datenmatrix unterscheidet sich semantisch von einer Tabelle dadurch, dass man von folgenden Grundannahmen ausgeht:

- Die Zeilen oder Fälle gehören zu den Individuen einer Stichprobe aus einer Grundgesamtheit.
- Alle in einer Spalte enthaltenen Informationen sind gleichartig und beziehen sich jeweils ausschließlich auf das Individuum der jeweiligen Zeile.
- Weiterhin setzt man voraus, dass die Individuen unabhängig voneinander und zufällig ausgewählt wurden.
- Die Reihenfolge der Zeilen ist also für die Auswertung der Tabelle bedeutungslos.

1.4 Skala

1.4.1 Der Begriff der Skala

Offenbar besteht in unserem Datensatz ein wesentlicher Unterschied zwischen den ersten 4 mit Zahlen gefüllten Spalten und der letzten, mit Namen gefüllten, Spalte. Dieser Unterschied ist sehr wichtig für die Auswahl geeigneter Verfahren und wird als die "Skala" der Variable bezeichnet.

Definition 19 (Skala)
Die Skala einer Variablen bezeichnet die Menge der möglichen Werte, zusammen mit den darauf sinnvollen Rechen- und Vergleichsoperationen.

Wir unterscheiden drei grobe Klassen von Skalen, die jeweils weiter unterteilt werden:

- stetige Skalen (z.B. "`Sepal.Length`")
- diskrete Skalen (z.B. "`Species`")
- spezielle Skalen (z.B. Winkel)

Die Einteilung von Variablen bzw. Merkmalen in die richtige Skala wird oft auch **Skalenniveau** genannt

1.4.2 Diskrete Skalen

Eine Skala heißt **diskret**, wenn nur klar voneinander separierte Wert auftreten können. Ist (bis auf Rundung) ein Kontinuum von Werten möglich, so heißt die Skala **stetig**.

1.4.2.1 Kategoriell

Der typische Fall einer diskreten Skala ist die sogenannte "**kategorielle**" Skala. Kategoriell kommt von gr. unterordnen. Dabei gibt es eine vor der Datenerhebung feststehende Einteilung der Grundgesamtheit in endlich viele Klassen. Jeder Fall ist eindeutig in eine der Klassen eingeordnet. In unserem Fall gehört jede Blume zu genau einer Art. Ein Merkmal heißt also **kategoriell**, falls jeder Fall eindeutig in eine feststehende Gruppeneinteilung eingeordnet wird.

Typische kategorielle Merkmale sind:

- Gruppe (z.B. "Geograph", "Geologe", "Politikwissenschaftler", "Lehrer", "Sonstige")
- Status (z.B. "Eigentümer", "Hauptmieter", "Untermieter", "Familienangehöriger")
- Status (z.B. "Azubi", "Geselle", "Meister", "kaufm. Angestellter")
- Familienstand ("ledig", "geschieden", "verheiratet", "verwitwet")
- Behandlung ("Placebo", "Altes Medikament", "Neues Medikament")
- Behandlung ("Wildwuchs", "Mähen", "Gießen")

Die Möglichkeiten heißen dann auch **Kategorien** oder **Stufen** des Merkmals.

Sind nur zwei Kategorien vorhanden, so spricht man auch von "**dichotomen** " (gr. zwei geteilt) Merkmalen: z.B.

- Geschlecht ("männlich", "weiblich")
- Zustimmung ("Ja", "Nein")

1.4.2.2 Ordinal als Sonderform der kategoriellen Skala

Eine kategorielle Skala wird als "**ordinal**" bezeichnet, wenn die Kategorien eine spezifische Anordnung besitzen.

- Höchster Schulabschluss ("Keiner", "Hauptschule", "Mittlere Reife", "Hochschulreife")
- Status (z.B. "Eigentümer", "Hauptmieter", "Untermieter")
- Status (z.B. "Azubi", "Geselle", "Meister")
- Bewertung (z.B. "gut", "mittel", "schlecht")

Aus Sicht der statistischen Verfahren ist ordinal einfach ein Spezialfall von kategoriell und erfordert die Anwendung der gleichen Verfahren. Es ist lediglich darauf zu achten, dass an allen Stellen, an denen die Reihenfolge eine Rolle spielt, dies auch beachtet wird. Dazu kommen in verschiedenen Softwarepaketen verschiedene Techniken zum Einsatz. Z.B. können die Kategorien mit Bezeichnungen versehen werden, deren alphabetische Anordnung ihrer Anordnung als Kategorien entspricht.

1.4.2.3 Nominalskala als fast kategorielle Skala

Eine Sonderform der diskreten Skala ist die **Nominalskala** (Nominal = von Namen her kommend). Ein Merkmal wird als **nominal** bezeichnet, wenn es Namen von Personen, Objekten oder Gruppen beinhaltet, die vor der Erhebung der Daten nicht festgemacht werden können oder sich bedeutungsinhaltlich nicht unterscheiden, außer, dass sie eben verschieden sind. Typische Beispiele nominaler Merkmale sind:

- Name der befragten Person
- Beruf (sofern keine klare Kategorisierung vorgenommen wird)
- Name des Interviewers (dessen der die Befragung durchführt)
- Art (wenn vorher nicht festgelegt wird, welche Arten untersucht werden)

In der Literatur werden "**nominal**" und "**kategoriell**" oft synonym verwendet, meist ohne dass der Unterschied überhaupt erkannt wird. Der Unterschied ist jedoch sehr wichtig, da die meisten bekannten Verfahren tatsächlich kategorielle Daten voraussetzen, bzw. für nominale Daten ganz andere Modelle verwendet werden sollten.

1.4.2.3.1 R: [Darstellung kategorieller Skalen in R] In R werden kategorielle Daten als sogenannte Faktoren (factor) dargestellt:

```
> species = iris$Species
> species
```

```
  [1] setosa     setosa     setosa     setosa     setosa
  [6] setosa     setosa     setosa     setosa     setosa
 [11] setosa     setosa     setosa     setosa     setosa
 [16] setosa     setosa     setosa     setosa     setosa
 [21] setosa     setosa     setosa     setosa     setosa
 [26] setosa     setosa     setosa     setosa     setosa
 [31] setosa     setosa     setosa     setosa     setosa
 [36] setosa     setosa     setosa     setosa     setosa
 [41] setosa     setosa     setosa     setosa     setosa
 [46] setosa     setosa     setosa     setosa     setosa
 [51] versicolor versicolor versicolor versicolor versicolor
```

```
 [56] versicolor versicolor versicolor versicolor versicolor
 [61] versicolor versicolor versicolor versicolor versicolor
 [66] versicolor versicolor versicolor versicolor versicolor
 [71] versicolor versicolor versicolor versicolor versicolor
 [76] versicolor versicolor versicolor versicolor versicolor
 [81] versicolor versicolor versicolor versicolor versicolor
 [86] versicolor versicolor versicolor versicolor versicolor
 [91] versicolor versicolor versicolor versicolor versicolor
 [96] versicolor versicolor versicolor versicolor versicolor
[101] virginica  virginica  virginica  virginica  virginica
[106] virginica  virginica  virginica  virginica  virginica
[111] virginica  virginica  virginica  virginica  virginica
[116] virginica  virginica  virginica  virginica  virginica
[121] virginica  virginica  virginica  virginica  virginica
[126] virginica  virginica  virginica  virginica  virginica
[131] virginica  virginica  virginica  virginica  virginica
[136] virginica  virginica  virginica  virginica  virginica
[141] virginica  virginica  virginica  virginica  virginica
[146] virginica  virginica  virginica  virginica  virginica
Levels: setosa versicolor virginica
```

```
> class(species)
```

```
[1] "factor"
```

```
> levels(species)
```

```
[1] "setosa"     "versicolor" "virginica"
```

Faktoren kann man selbst herstellen durch:

```
> zustimmung = factor(c("Ja", "Ja", "Nein", "Nein", "Ja"))
> zustimmung
```

```
[1] Ja   Ja   Nein Nein Ja
Levels: Ja Nein
```

```
> class(zustimmung)
```

```
[1] "factor"
```

```
> levels(zustimmung)
```

```
[1] "Ja"   "Nein"
```

Um ein ordinales Merkmal festzulegen, ersetzt man `factor` durch `ordered` und gibt die Reihenfolge explizit an:

```
> bewertung = ordered(c("Schlecht", "Schlecht", "Super",
+     "Gut", "Gut"), c("Schlecht", "Gut", "Super"))
> bewertung
```

```
[1] Schlecht Schlecht Super    Gut      Gut
Levels: Schlecht < Gut < Super
```

```
> class(bewertung)
```

```
[1] "ordered" "factor"
```

```
> levels(bewertung)
```

```
[1] "Schlecht" "Gut"      "Super"
```

Die Reihenfolge der Möglichkeiten entspricht dann der inhaltlich vorgegebenen Reihenfolge. In der weiteren Verarbeitung wird dann die Variable auch entsprechend als ordinale Größe verwendet.

1.4.2.4 Intervallskaliert

Eine weitere Spezialisierung der ordinalen Skala ist die **Intervallskalierung**. Dabei wird zusätzlich implizit vorausgesetzt, dass die verschiedenen benachbarten Kategorien den gleichen Abstand voneinander haben. Typische intervallskalierte Merkmale sind:

- Noten (1,2,3,4,5)
- Kategorisierungen in gleichen Abständen ("0-1000", "1000-2000", "2000-3000")

Intervallskalierte Merkmale werden je nach Kontext mal als Zahlen (Nummer der Kategorie, z.B. bei Notendurschnittsbildung) und mal als ordinal (z.B. bei Tabellen, wie oft welche Note vorkam) behandelt.

1.4.2.5 Anzahlen

Eine ähnliche Zwischenposition nehmen **Anzahlen** ein.

- Anzahl der Kinder in der Familie
- Anzahl der Unfälle
- Anzahl der Zimmer in der Wohnung

Hat man es typischerweise mit kleinen Anzahlen (0-5) zu tun, so eignen sich oft die Methoden für ordinale Merkmale. Sind die Anzahlen eher groß (>30), so eignen sich oft die Methoden für stetige Skalen. Oft entstehen Anzahlen aber nicht als Merkmale eines Individuums (z.B. einer Wohnung), sondern als Anzahlen von Individuen. Dann hat man es aber meist mit einer Datentafel und nicht mit einer Datenmatrix zu tun.

Beispiel 20 *Der Datensatz ist ebenfalls ein Beispieldatensatz in R:*

```
This data set gives the number of warp breaks per loom, where a
loom corresponds to a fixed length of yarn.

Usage:

warpbreaks

Format:

A data frame with 54 observations on 3 variables.

'[,1]'  'breaks'   numeric  The number of breaks
'[,2]'  'wool'     factor   The type of wool (A or B)
'[,3]'  'tension'  factor   The level of tension (L, M, H)

There are measurements on 9 looms for each of the six types of
warp ('AL', 'AM', 'AH', 'BL', 'BM', 'BH').

Source:

Tippett, L. H. C. (1950) _Technological Applications of
Statistics_. Wiley.  Page 106.

References:
```

Tukey, J. W. (1977) _Exploratory Data Analysis_. Addison-Wesley.

McNeil, D. R. (1977) _Interactive Data Analysis_. Wiley.

```
> data(warpbreaks)
> warpbreaks
```

	breaks	wool	tension
1	26	A	L
2	30	A	L
3	54	A	L
4	25	A	L
5	70	A	L
6	52	A	L
7	51	A	L
8	26	A	L
9	67	A	L
10	18	A	M
11	21	A	M
12	29	A	M
13	17	A	M
14	12	A	M
15	18	A	M
16	35	A	M
17	30	A	M
18	36	A	M
19	36	A	H
20	21	A	H
21	24	A	H
22	18	A	H
23	10	A	H
24	43	A	H
25	28	A	H
26	15	A	H
27	26	A	H
28	27	B	L
29	14	B	L
30	29	B	L
31	19	B	L
32	29	B	L
33	31	B	L
34	41	B	L
35	20	B	L
36	44	B	L
37	42	B	M
38	26	B	M
39	19	B	M
40	16	B	M
41	39	B	M
42	28	B	M
43	21	B	M

44	39	B	M
45	29	B	M
46	20	B	H
47	21	B	H
48	24	B	H
49	17	B	H
50	13	B	H
51	15	B	H
52	15	B	H
53	16	B	H
54	28	B	H

Der Datensatz besteht aus einem Anzahlmerkmal und zwei kategoriellen Merkmalen. Das statistische Individuum ist ein Web-Versuch aus der Grundgesamtheit der durchgeführten Webversuche.

1.4.3 Stetige Skalen

Ein Skala heißt **stetig**, wenn (bis auf Rundung) ein Kontinuum an Zahlenwerten möglich ist. Ihre beiden wichtigsten Vertreter sind

- reelle Skala
- relative (oder positive) Skala

Wir wollen bei der Einteilung nicht weiter ins Detail gehen und nicht zu streng mit den Merkmalen umgehen.

1.4.3.1 Reelle Skala

Die wichtigsten Merkmale für eine "**reell**" skalierte Größe sind:

- Alle reellen Zahlen sind (im Prinzip) möglich.
- Der Unterschied von x und $x + \alpha$ ist gleich bedeutend, egal wie groß x ist.

Typische reelle Merkmale sind:

- Temperatur in Celsius
- Potentialdifferenz (elektrische oder gravitative, die positiv und negativ sein kann)
- PH-Wert (eigentlich stellt dieser eine ganz eigene Skala dar)
- Logarithmen von relativ skalierten Merkmalen
- Merkmale aller anderen Skalen bei minimaler Variation (z.B. geographische Breite, wenn es nur um ein paar Grad geht), in Ermangelung besserer Methoden
- Differenzen eines Merkmals zum Vorjahr

1.4.3.2 Relative Skala

Die wichtigsten Merkmale für eine "**relativ**" skalierte Größe sind:

- Alle positiven Zahlen sind (im Prinzip) möglich.
- Der Unterschied von x und αx ist gleich bedeutend, egal wie groß x ist.

Typische relative Merkmale sind:

- Temperatur in Kelvin
- Stoffkonzentrationen
- Mengen
- räumliche Abstände
- zeitliche Abstände
- Feldstärken (ungerichtet)
- Anzahlen (wenn sie so groß sind, dass sie nicht mehr diskret behandelt werden sollten und eine Skala sind, für Anzahlen gibt es allerdings oft spezielle Methoden)

Merkmale mit unklarer Zuordnung sind

- Einkommen
- ...

1.4.4 Spezielle Skalen

Oft treten Werte auf, die zwar durch eine oder mehrere Zahlen beschrieben werden können, deren Abstandsbegriffe und Wertebereiche sich jedoch mit einer Interpretation als reelle Skala nicht vereinbaren lassen. Für diese gibt es dann spezielle Methoden, die jedoch den Rahmen dieser Vorlesung bei Weitem übersteigen würden:

- Anteile und Wahrscheinlichkeiten*
  Der Wertebereich ist $(0, 1)$ oder $(0\%, 100\%)$ und der Abstandsbegriff ist für kleine Wahrscheinlichkeiten relativ. Diese Werte treten insbesondere bei Risikoabschätzungen auf.
- Zusammensetzungen*
  Mehrere Anteile, die sich gemeinsam zu 100% addieren oder höchstens zu 100% addieren können. Diese Skala tritt insbesondere bei allen geochemischen Analysen auf.
- Winkel*
  Werte im Bereich $[0^o - 180^o)$ oder $[0^o, 360^o)$, bei denen 179.9^o bzw. 359.9 fast 0^o sind. Dieser Datentyp tritt insbesondere in der Strukturgeologie auf.
- Ausrichtung*
  Die Ausrichtung eines Linear, einer kristallographischen c-Achse, einer Schieferungsrichtung oder einer magnetischen Richtung im Raum. Dieser Datentyp tritt insbesondere in der Strukturgeologie auf.
- Plattenbewegungen*
  Sind eine komplizierte winkelartige Größe, die in der Geotektonik eine zentrale Rolle spielt.

- Kristallorientierungen*
  Die Ausrichtung von Kristallkörnern und Kristallgitterebenen im Raum. Dieser Datentyp tritt insbesondere in den Materialwissenschaften und der Strukturgeologie auf.

- Überlebenszeiten
  Die Überlebenszeit eines Organismus nach einem Eingriff oder die Funktionszeit einer Maschine nach einer Wartung. Oft werden Überlebenszeiten nur unvollständig beobachtet, so dass nur bekannt ist, dass der Patient nach Abschluss der Studie noch gelebt hat. Diese Daten treten insbesondere in der Medizin und der Qualitätssicherung auf.

- Datum/Zeit
  Das Datum scheint auf dem ersten Blick einer reellen Zahl zu ähneln. Sobald man aber Jahreszeiten, Tageszeiten, Wetterlagen, Arbeitswoche, Feiertage und Börsenöffnungszeiten berücksichtigt, wird es richtig kompliziert. Diese Skala tritt insbesondere in der (Finanz-, Betriebs- und Volks)-Wirtschaftslehre, der Klimatologie, der Limnologie und der Ökologie auf.

- Ortskoordinaten*
  Sind ein Dauerthema aller Geowissenschaften. Sie werden in den Veranstaltungen von Frau Prof. Schafmeister und Herrn Prof. Zölitz-Möller sicher noch eine bedeutende Rolle spielen.

- Gebietskörperschaften*
  und ihre gebrochene hierarchische Struktur mit Stadtstaaten und kreisfreien Städten sind eine andauernde Begleiterscheinung jeder amtlichen Statistik und damit tägliches Brot des Anthropogeographen.

1.5 Tafeln

In der Sozialgeographie und der empirischen Sozialforschung spielt die Auswertung von Fragebögen und damit die kategorielle und ordinale Skala die wichtigste Rolle. Die in solchen Datensätzen enthaltenen Informationen lassen sich viel kürzer zusammenfassen, wenn man anstelle der Datenmatrix, die jedes Individuum einzeln auflistet, nur die Anzahl der Individuen mit spezifischen Merkmalskombinationen auflistet.

1.5.1 Häufigkeitstafel

Handelt es sich nur um ein Merkmal, so spricht man von einer **Häufigkeitstafel**.

```
> table(iris$Species)

    setosa versicolor  virginica
        50         50         50
```

Die Häufigkeitstafel gibt für jede Stufe des Merkmals die Anzahl der Individuen in der Stichprobe an, die das Merkmal entsprechend ausweisen. Eine Tafel ist also eine alternative Darstellungsform, die von der Datenmatrix klar zu unterscheiden ist.

1.5.2 Kontingenztafel

Gibt es mehr als ein kategorielles Merkmal, so kann man für jeweils zwei dieser Merkmale eine Tafel mit Zeilen und Spalten aufstellen, in denen die Zeilen jeweils für das eine Merkmal und die Spalten für das andere Merkmal stehen. In jeder Zelle, in der sich Zeilen und Spalten schneiden, steht dann eine Information über den Anteil der Individuen mit dieser speziellen Merkmalskombination.

```
> warpbreaks
```

	breaks	wool	tension
1	26	A	L
2	30	A	L
3	54	A	L
4	25	A	L
5	70	A	L
6	52	A	L
7	51	A	L
8	26	A	L
9	67	A	L
10	18	A	M
11	21	A	M
12	29	A	M
13	17	A	M
14	12	A	M
15	18	A	M
16	35	A	M
17	30	A	M
18	36	A	M
19	36	A	H
20	21	A	H
21	24	A	H
22	18	A	H
23	10	A	H
24	43	A	H
25	28	A	H
26	15	A	H
27	26	A	H
28	27	B	L
29	14	B	L
30	29	B	L
31	19	B	L
32	29	B	L
33	31	B	L
34	41	B	L
35	20	B	L
36	44	B	L
37	42	B	M
38	26	B	M
39	19	B	M
40	16	B	M
41	39	B	M
42	28	B	M
43	21	B	M
44	39	B	M

```
45     29    B       M
46     20    B       H
47     21    B       H
48     24    B       H
49     17    B       H
50     13    B       H
51     15    B       H
52     15    B       H
53     16    B       H
54     28    B       H
```

```
> xtabs(~wool + tension, warpbreaks)

    tension
wool L M H
   A 9 9 9
   B 9 9 9
```

Zeilen und Spalten werden also jeweils Stufen des Merkmals zugeordnet. Eine Verwechslung einer Kontingenztafel mit einer Datenmatrix sollte daher ausgeschlossen sein.

1.5.3 Multivariate Kontingenztafeln

Die Darstellung mehrerer Variablen erfordert, dass in Zeilen und Spalten die Kombinationen mehrerer Merkmale codiert werden.

Unser Beispiel:

```
Survival of passengers on the Titanic

Description:

     This data set provides information on the fate of passengers on
     the fatal maiden voyage of the ocean liner 'Titanic', summarized
     according to economic status (class), sex, age and survival.

Usage:

     Titanic

Format:

     A 4-dimensional array resulting from cross-tabulating 2201
     observations on 4 variables.  The variables and their levels are
     as follows:

       No  Name      Levels
        1  Class     1st, 2nd, 3rd, Crew
        2  Sex       Male, Female
        3  Age       Child, Adult
        4  Survived  No, Yes

Details:

     The sinking of the Titanic is a famous event, and new books are
```

```
     still being published about it.  Many well-known facts-from the
     proportions of first-class passengers to the "women and children
     first" policy, and the fact that that policy was not entirely
     successful in saving the women and children in the third class-are
     reflected in the survival rates for various classes of passenger.

     These data were originally collected by the British Board of Trade
     in their investigation of the sinking.  Note that there is no
     complete agreement among primary sources as to the exact numbers
     on board, rescued, or lost.

     Due in particular to the very successful film 'Titanic', the last
     years saw a rise in public interest in the Titanic.  Very detailed
     data about the passengers is now available on the Internet, at
     sites such as _Encyclopedia Titanica_ (<URL:
     http://www.rmplc.co.uk/eduweb/sites/phind>).

Source:

     Dawson, Robert J. MacG. (1995), The 'Unusual Episode' Data
     Revisited. _Journal of Statistics Education_, *3*. <URL:
     http://www.amstat.org/publications/jse/v3n3/datasets.dawson.html>

     The source provides a data set recording class, sex, age and
     survival status for each person on board of the Titanic, and is
     based on data originally collected by the British Board of Trade
     and reprinted in:

     British Board of Trade (1990), _Report on the Loss of the
     'Titanic' (S.S.)_. British Board of Trade Inquiry Report
     (reprint). Gloucester, UK: Allan Sutton Publishing.
```

Dieser Datensatz wird also bereits als Tabelle angegeben.

```
> data(Titanic)
> ftable(Titanic)

                    Survived  No Yes
Class Sex    Age
1st   Male   Child            0   5
             Adult          118  57
      Female Child            0   1
             Adult            4 140
2nd   Male   Child            0  11
             Adult          154  14
      Female Child            0  13
             Adult           13  80
3rd   Male   Child           35  13
             Adult          387  75
      Female Child           17  14
             Adult           89  76
Crew  Male   Child            0   0
             Adult          670 192
```

```
      Female Child            0   0
             Adult            3  20
```

Es haben also z.B. 20 weibliche erwachsene Crewmitglieder überlebt, während 3 weibliche erwachsene Crewmitglieder die Katastrophe nicht überlebt haben.

1.5.3.0.1 R: [Umgang mit Kontingenztafeln in R] Gehen wir einen Moment davon aus, das wir im warpbreak Datensatz nicht die Experimente, sondern die Brüche als statistische Individuen ansehen, dann gibt der Datensatz

```
> warpbreaks
```

```
   breaks wool tension
1      26    A       L
2      30    A       L
3      54    A       L
4      25    A       L
5      70    A       L
6      52    A       L
7      51    A       L
8      26    A       L
9      67    A       L
10     18    A       M
11     21    A       M
12     29    A       M
13     17    A       M
14     12    A       M
15     18    A       M
16     35    A       M
17     30    A       M
18     36    A       M
19     36    A       H
20     21    A       H
21     24    A       H
22     18    A       H
23     10    A       H
24     43    A       H
25     28    A       H
26     15    A       H
27     26    A       H
28     27    B       L
29     14    B       L
30     29    B       L
31     19    B       L
32     29    B       L
33     31    B       L
34     41    B       L
35     20    B       L
36     44    B       L
37     42    B       M
38     26    B       M
39     19    B       M
40     16    B       M
41     39    B       M
42     28    B       M
43     21    B       M
44     39    B       M
45     29    B       M
46     20    B       H
```

```
47     21    B       H
48     24    B       H
49     17    B       H
50     13    B       H
51     15    B       H
52     15    B       H
53     16    B       H
54     28    B       H
```

die folgende Kontingenztafel vor:

```
> MeineTabelle = xtabs(breaks ~ wool + tension, data = warpbreaks)
> ftable(MeineTabelle)

     tension   L   M   H
wool
A             401 216 221
B             254 259 169
```

diese kann mit as.data.frame(*Tabelle*) in die obige Form zurückverwandelt werden:

```
> as.data.frame(MeineTabelle)

  wool tension Freq
1    A       L  401
2    B       L  254
3    A       M  216
4    B       M  259
5    A       H  221
6    B       H  169
```

Also können wir auch den Titanic Datensatz in die andere Form bringen:

```
> Tit = as.data.frame(Titanic)
> Tit

   Class    Sex   Age Survived Freq
1    1st   Male Child       No    0
2    2nd   Male Child       No    0
3    3rd   Male Child       No   35
4   Crew   Male Child       No    0
5    1st Female Child       No    0
6    2nd Female Child       No    0
7    3rd Female Child       No   17
8   Crew Female Child       No    0
9    1st   Male Adult       No  118
10   2nd   Male Adult       No  154
11   3rd   Male Adult       No  387
12  Crew   Male Adult       No  670
13   1st Female Adult       No    4
14   2nd Female Adult       No   13
15   3rd Female Adult       No   89
16  Crew Female Adult       No    3
17   1st   Male Child      Yes    5
18   2nd   Male Child      Yes   11
19   3rd   Male Child      Yes   13
20  Crew   Male Child      Yes    0
21   1st Female Child      Yes    1
22   2nd Female Child      Yes   13
23   3rd Female Child      Yes   14
24  Crew Female Child      Yes    0
25   1st   Male Adult      Yes   57
26   2nd   Male Adult      Yes   14
```

```
27   3rd   Male Adult      Yes   75
28  Crew   Male Adult      Yes  192
29   1st Female Adult      Yes  140
30   2nd Female Adult      Yes   80
31   3rd Female Adult      Yes   76
32  Crew Female Adult      Yes   20
```

und mit xtabs entsprechend in eine beliebige andere Tabelle umsetzen z.B.

```
> ftable(xtabs(Freq ~ Survived + Class, data = Tit))

         Class 1st 2nd 3rd Crew
Survived
No             122 167 528  673
Yes            203 118 178  212

> ftable(xtabs(Freq ~ Survived + Class + Sex + Age, data = Tit))

                     Age Child Adult
Survived Class Sex
No       1st   Male          0   118
               Female        0     4
         2nd   Male          0   154
               Female        0    13
         3rd   Male         35   387
               Female       17    89
         Crew  Male          0   670
               Female        0     3
Yes      1st   Male          5    57
               Female        1   140
         2nd   Male         11    14
               Female       13    80
         3rd   Male         13    75
               Female       14    76
         Crew  Male          0   192
               Female        0    20
```

welche z.B. weniger Merkmale verwendet oder eine andere Darstellungsreihenfolge wählt.