Praktikum zur Statistik mit R

Till Breuer

Institut für Mathematische Statistik
Universität Münster

5. Oktober 2010

Gliederung

Gliederung

Literatur I

Fahrmeir, Künstler, Pigeot, Tutz
Statistik. Der Weg zur Datenanalyse
Springer-Verlag Berlin · Heidelberg · New York

Ahlers, S.
Einführung in die Statistik mit R
Skript zur Veranstaltung
www.math.uni-muenster.de/statistik/praktika/Statistikpraktikum/SS09/Skript.pdf

Backhaus, Erichsen, Plinke und Weiber
Multivariate Analysemethoden
Springer-Lehrbuch

Gliederung

Was tut man in der Statistik?

- Daten sammeln
- Daten analysieren
- Prognosen und Entscheidungen treffen

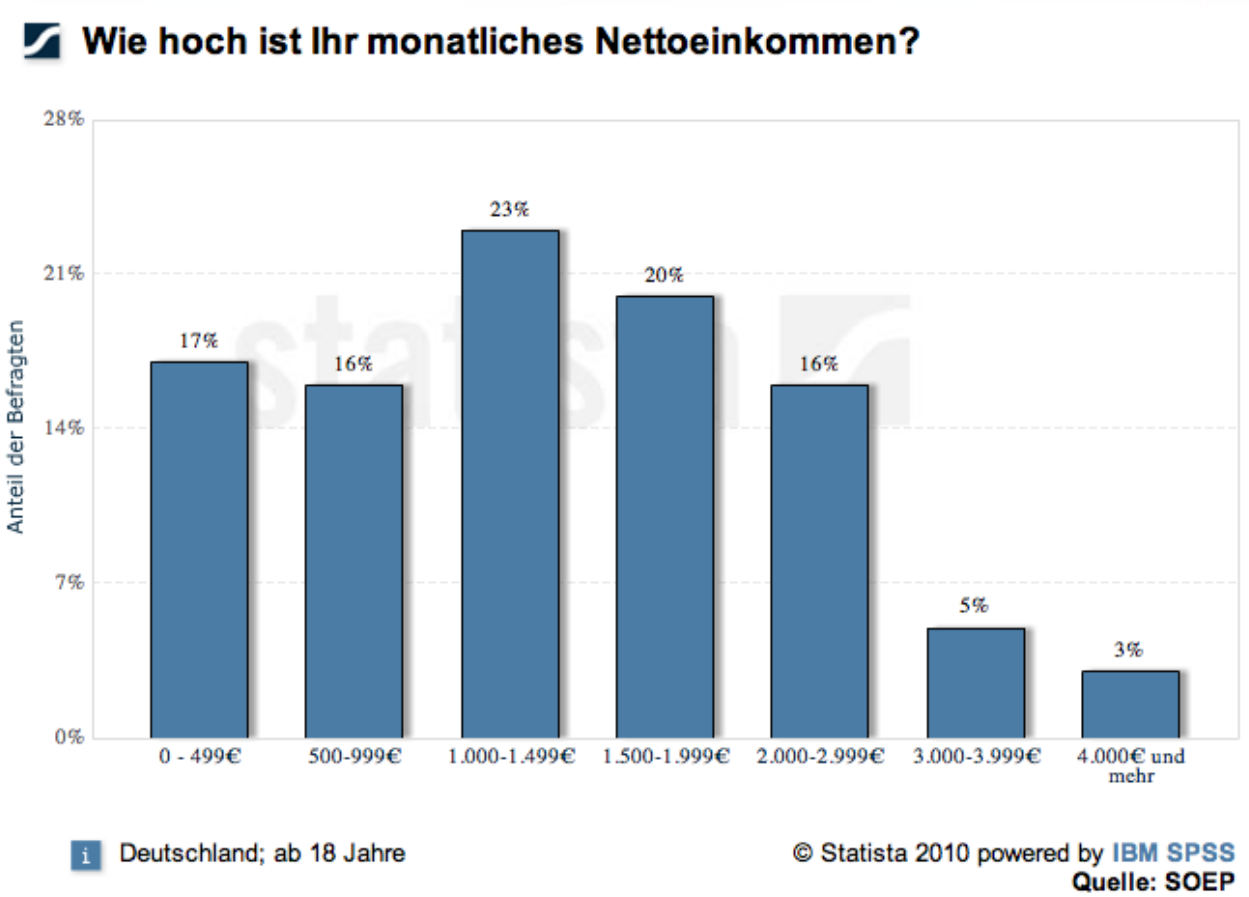
Beispiel: Klassenspiegel

- 30 Schüler bekommen ihre Klausur zurück.
- Ziel: Durchschnittsnote berechnen und Notenverteilung skizzieren

Beispiel: Einkommensverteilung

- *Beispiel*: 10.000 Personen werden zu ihrem Einkommen befragt.
- Ziel: **Darstellung** der Einkommensverteilung, **Lage und Streuung einschätzen**

Beispiel: Einkommensverteilung



Beispiel: Epidemiologische Studie zum Rauchverhalten

- *Fragestellung*: Wie wirkt sich das Merkmal “Rauchverhalten” auf das Lungenkrebsrisiko aus?
- Ziel: **Quantifizierung des Einflusses** gewisser Merkmale und Faktoren.

Beispiel: Düngemittel

- *Fragestellung*: Wie stark ist der Zusammenhang zwischen der eingesetzten Menge eines Düngemittels und der Erntemenge?
- Ziel: **Quantifizierung des Zusammenhanges** zweier Merkmalsausprägungen

Beispiel: Produktionsprozess

- *Fragestellung*: Lohnt sich die Umstellung eines Produktionsprozesses? Wie groß ist das Risiko bei einer Umstellung?
- Ziel: **Treffen und Validieren einer Entscheidung**

Beispiel: Glühbirne

- *Fragestellung*: Wie groß ist die Lebensdauer einer Glühbirne aus einer bestimmten Produktion
- Ziel: **Schätzen** der mittleren Lebensdauer einer Glühbirne

Beispiel: Münzwurf

Ein Schiedsrichter entscheidet über die Wahl der Spielrichtung durch einen Münzwurf.

- *Fragestellung*: Ist die verwendete Münze fair.
- Ziel: **Entscheidung** darüber, ob die Münze fair ist oder nicht.

weitere Beispiele

- Inwieweit sind die Antworten zur Sonntagsfrage, die in einer Umfrage erhalten werden, repräsentativ für alle Wahlberechtigten?
- Ist Therapie A besser als Therapie B?

Drei Arten der Datenanalyse

Bei der Datenanalyse lassen sich drei Grundaufgaben der Statistik unterscheiden:

- Beschreiben (*Deskription*)
- Suchen (*Exploration*)
- Schließen (*Induktion*)

Beschreiben => Deskriptive Statistik

- Beschreibende und graphische Aufbereitung und Komprimierung von Daten, z. B. zur Präsentation umfangreichen Datenmaterials, z.B.
  - ... Beschreiben durch Lage- und Streumaße
  - ... Darstellen durch Gruppierung der Daten
  - ...graphischen Darstellungen durch Balkendiagramme oder Histogramme

Suchen => Explorative Statistik

- Darstellung von Daten
- Suche nach Strukturen und Besonderheiten in den Daten
- verwendet keine Stochastik, dafür häufig rechenaufwendige Methoden
- wird typischerweise eingesetzt, wenn die Fragestellung nicht genau definiert ist oder die Wahl eines geeigneten statistischen Modells unklar ist

Schließen => Induktive (schließende) Statistik

- Zielsetzung ist über die erhobenen Daten hinaus allgemeinere Schlußfolgerungen für umfassendere Grundgesamtheiten zu ziehen.
- Dazu: Einbeziehung von Wahrscheinlichkeitstheorie und Stochastik
- Eine statistisch abgesicherte Beantwortung solcher Fragen erfordert eine sorgfältige Versuchsplanung, vorbereitende deskriptive und explorative Analysen sowie klar definierte stochastische Modelle, um wahrscheinlichkeitstheoretische Rückschlüsse zu ermöglichen.

Gliederung

Definitionen

Statistische Einheiten:	Objekte, an denen interessierende Größen erfasst werden
Grundgesamtheit:	Menge aller für die Fragestellung relevanten statistischen Einheiten
Teilgesamtheit:	Teilmenge der Grundgesamtheit
Stichprobe:	tatsächlich untersuchte Teilmenge der Grundgesamtheit
Merkmal:	interessierende Größe, *Variable*
Merkmalsausprägung:	konkreter Wert des Merkmals für eine bestimmte statistische Einheit

Beispiel: Mietspiegel

- Statistische Einheiten: Wohnungen, an denen die interessierenden Größen erfaßt werden
- Grundgesamtheit: Menge aller Wohnungen in München wie im Gesetz
- Stichprobe: Wohnungen, deren Daten erfasst wurden
- Merkmale: Alter, Größe, Preis/qm
- Merkmalsausprägungen: für das Baujahr gibt es die Ausprägungen „bis 1929“,..., „2004-2005“; für die Wohnfläche die Ausprägungen „21-30 qm“, ..., „151-160qm“, für den Preis/qm die Ausprägungen $x \in \mathbb{R}_{\geq 0}$.

Ziel- und Einflussgrößen

- Merkmale werden auch Variablen genannt.
- Man unterscheidet Variablen, die beeinflußt werden, die sogenannten *Zielgrößen*, und solche, die beeinflussen.
- Die beeinflussenden Variablen werden aufgeteilt in beobachtbare Variablen, die als *Einflussgrößen* oder *Faktoren* bezeichnet werden, und in nicht beobachtbare Variablen, die *Störgrößen*. Störgrößen werden auch als *latente* Faktoren bezeichnet.

Beispiel

In einer epidemiologischen Studie wird der Einfluss des Merkmals *Rauchverhalten* auf das Merkmal *Lungenkrebs* untersucht.

- Das Rauchverhalten ist eine Einflussgröße.
- Das Merkmal Lungenkrebs ist die Zielgröße.
- Als Störgröße tritt z. B. die Prädisposition für Lungenkrebs auf.

Gliederung

Gliederung

Beispiel: Mietspiegel

In vielen Städten und Gemeinden der Bundesrepublik werden sogenannte Mietspiegel erstellt. Sie bieten Mietern und Vermietern eine Marktübersicht zu Miethöhen, helfen in Mietberatungsstellen und werden, neben Sachverständigen, auch zur Entscheidung in Mietstreitprozessen herangezogen.
Nach §558 BGB ist die ortsübliche Vergleichsmiete wie folgt definiert:

> *„Die ortsübliche Vergleichsmiete wird gebildet aus den üblichen Entgelten, die in der Gemeinde oder einer vergleichbaren Gemeinde für Wohnraum vergleichbarer Art, Größe, Ausstattung, Beschaffenheit und Lage in den letzten vier Jahren vereinbart oder, von Erhöhungen nach §560 abgesehen, geändert worden sind“.*

Das Gesetz

- legt die Grundgesamtheiten fest, aus denen die Stichproben für die Erstellung von Mietspiegeln zu ziehen sind.
- gibt einen Hinweis auf die statistische Analysemethode: Sinngemäß bedeutet dies für die Nettomiete, dass ihr Durchschnittswert in Abhängigkeit von Merkmalen wie Art, Größe, Ausstattung, Beschaffenheit und Lage der Wohnung zu bestimmen bzw. zu schätzen ist.

Erstellung des Mietspiegels

- aus der Gesamtheit aller nach dem Mietgesetz relevanten Wohnungen der Stadt wird eine repräsentative Stichprobe gezogen
- die interessierenden Daten werden von Interviewern in Fragebögen eingetragen
- Das mit der Datenerhebung beauftragte Institut, in München Infratest, erstellt daraus eine Datei, die der anschließenden statistischen Beschreibung, Auswertung und Analyse zugrunde liegt.
- Die Präsentation der Ergebnisse erfolgt schließlich in einer Mietspiegelbroschüre bzw. im Internet.

Ausschnitt aus dem Münchener Mietspiegel 2003

Nettomiete/qm			
	Wohnfläche		
Baualter	bis 38 qm	39 bis 80 qm	81qm und mehr
bis 1918	10.96(20)	7.86(189)	7.46(190)
1919 bis 48	8.00(5)	7.07(128)	6.71(53)
1949 bis 65	10.32(64)	8.10(321)	7.68(68)
1966 bis 77	10.43(112)	8.10(364)	7.67(151)
1978 bis 89	11.00(10)	9.41(115)	8.95(42)
ab 1990	11.40(6)	10.19(154)	9.80(59)

Tabelle 1.2: Einfacher Tabellen-Mietspiegel, in Klammern die Anzahl der einbezogenen Wohnungen

Erhebung von Daten

- Befragung
  - schriftlich
  - mündlich
  - offen
  - geschlossen
- Beobachtung
- Experiment

Stichprobenarten

Wann immer man auf eine *Vollerhebung* (d. h. eine Erfassung aller statistischen Einheiten einer Grundgesamtheit) verzichtet, greift man auf die Ziehung einer *Stichprobe* zurück.

Stichprobenart	Bemerkung
einfache Zufalls-stichprobe	stark zufallsabhängig, technisch schwer umsetzbar
systematische Ziehung	kann systematische Fehler haben
geschichtete Zufalls-stichprobe	meistens einfacher umsetzbar und repräsentativer als einf. Zufallsstichprobe
Klumpenstichprobe	erhöhte praktische Umsetzbarkeit, ggf. große Verzerrungen bei Klumpen, die untereinander heterogen sind

Beispiele für Stichprobenarten - Geschichtete Zufallsstichprobe

Beispiel (Bundestagswahl)

- Einflussgrößen wie Alter, Geschlecht, Bildungsstatus, etc. beeinflussen das Wahlverhalten
- Eine geschichtete Zufallsstichprobe ermöglicht bessere Vorhersagen

Beispiele für Stichprobenarten - Klumpenstichprobe

Beispiel

Bei einer soziologischen Befragung in einem bestimmten Beruf werden die Ergebnisse jeweils unternehmensweit zusammengefasst. Die Klumpen sind die einzelnen Unternehmen.

Verzerrte Stichproben

Werden jedoch Elemente der Grundgesamtheit bei der Ziehung nicht berücksichtigt, spricht man von einer *verzerrten Stichprobe*. Mögliche Verzerrungen sind:

Verzerrung	Ursache und Beispiel
Selektions-Bias	bewusster Ausschluss von Elementen von der Ziehung
Beispiel:	Internet- oder Zeitungsumfrage
Nonresponse-Bias	(unangenehme) Fragen bleiben unbeantwortet
Beispiel:	Fragen zum Sexualverhalten etc.
Selfselection-Bias	Umfragen auf freiwilliger Basis
Beispiel:	McKinsey-Studie „Perspektive Deutschland“(2003)

Studiendesigns

Studientyp	
Querschnitt-studie	an einer bestimmten Anzahl von Objekten, wird zu einem bestimmten Zeitpunkt ein Merkmal oder mehrere erfasst
Beispiel:	Absolventenstudie
Zeitreihe	ein Objekt wird hinsichtlich eines Merkmals über einen ganzen Zeitraum beobachtet
Beispiele:	Aktienkurse, DAX
Längsschnitt-studie	eine Gruppe wird hinsichtlich eines Merkmals über einen ganzen Zeitraum beobachtet

Aufgaben zur Datenerhebung

Gliederung

Beispiele: Mermale und Ausprägungen

- Geschlecht

männlich	*weiblich*
43	57

- Schulnoten

1	*2*	*3*	*4*	*5*	*6*
2	4	12	8	2	-

- Körpergröße

$\leq$ *170cm*	*171-190cm*	*> 191cm*
19	65	16

Was lässt sich hinsichtlich Beschaffenheit, Ordnung und Abstand der Merkmalsausprägungen beobachten?

Stetige und diskrete Mermale

diskret:	endlich oder abzählbar unendlich viele Ausprägungen
stetig:	alle Werte eines Intervalls sind mögliche Ausprägungen
quasi-stetig:	diskret messbare, aber fein abgestufte Daten

- Geschlecht, Schulnoten: diskret
- Körpergröße: stetig, diskrete Einteilung
- quasi-stetige Merkmale sind etwa Nettomiete oder Kredithöhe

Skalenarten

nominalskaliert:	Ausprägungen sind Namen, keine Ordnung möglich
ordinalskaliert:	Ausprägungen können geordnet, aber Abstände nicht interpretiert werden
intervallskaliert:	Ausprägungen sind Zahlen, Interpretation der Abstände möglich
verhältnisskaliert:	Ausprägungen besitzen sinnvollen absoluten Nullpunkt

Kriterien für Skalenarten

	sinnvoll interpretierbare Berechnungen			
Skalenart	auszählen	ordnen	Differenzen	Quotienten
nominal	ja	nein	nein	nein
ordinal	ja	ja	nein	nein
intervall	ja	ja	ja	nein
verhältnis	ja	ja	ja	ja

Beispiele

- nominalskaliert: das Merkmal *Zentralheizung* im Mietspiegel mit den möglichen Ausprägungen „ja“ und „nein“
- ordinalskaliert: das Merkmal *Schulnote* mit den Ausprägungen 1 bis 6
- intervallskaliert: das Merkmal *Temperatur in Grad Celsius* mit den möglichen Ausprägungen $x \in \mathbb{R}, x > -273,15$
- verhältnisskaliert: das Merkmal Nettomiete im Mietspiegel mit den Ausprägungen $x \in \mathbb{R}_{\geq 0}$

Qualitative und quantitative Mermale

- Qualitative Mermale geben keine Intensität bzw. Ausmaß wieder. Sie besitzen endlich viele Ausprägungen besitzen und sind höchstens ordinalskaliert.
- Quantitative Mermale geben Intensitäten bzw. Ausmaße wieder. Kardinalskalierte (also intervall- / verhältnisskalierte) Merkmale sind stets ebenfalls quantitativ.

qualitativ:	endlich viele Ausprägungen, höchstens Ordinalskala
quantitativ:	Ausprägungen geben Intensität wieder

Zusammenfassung

diskret:	endlich oder abzählbar unendlich viele Ausprägungen
stetig:	alle Werte eines Intervalls sind mögliche Ausprägungen
quasi-stetig:	diskret messbare, aber fein abgestufte Daten
nominalskaliert:	Ausprägungen sind Namen, keine Ordnung möglich
ordinalskaliert:	Ausprägungen können geordnet, aber Abstände nicht interpretiert werden
intervallskaliert:	Ausprägungen sind Zahlen, Interpretation der Abstände möglich
verhältnisskaliert:	Ausprägungen besitzen sinnvollen absoluten Nullpunkt
qualitativ:	endlich viele Ausprägungen, höchstens Ordinalskala
quantitativ:	Ausprägungen geben Intensität wieder

Aufgabe

Diskutieren Sie die im Rahmen des Münchener Mietspiegel erhobenen Mermale *Nettomiete*, *Wohnfläche*, *Baualter*, *Gebäudetyp* (Ausprägungen: *Hochhaus*/*Wohnblock*), *Zentralheizung* (*dezentral betriebene Strom- oder Gasheizungen*/*Einzelöfen*/*keine Heizung*), *Warmwasserversorgung* (*einfache*/*keine*), *Lage der Wohnung* (*einfache*/*durchschnittliche*/*gute*/*beste*) und *Ausstattung des Bads* (*keins*/*nicht gekachelt*/*zweites vollständiges Bad vorhanden*/*besondere Zusatzausstattung*) hinsichtlich ihres jeweiligen Skalenniveaus. Entscheiden Sie zudem, ob es sich um diskrete oder stetige, bzw. quantitative oder qualitative Merkmale handelt.

Gliederung

Gliederung

Histogramme

Hat man einen großen Datensatz mit vielen verschiedenen Merkmalsausprägungen eines quantitativen Merkmals, so werden die obigen Darstellungen häufig unübersichtlich. Man verwendet dann z. B. *Histogramme*.

Definition

Ein *Histogramm* ist ein spezielles Säulendiagramm, bei dem die Merkmalsausprägungen in $k \in \mathbb{N}_{\geq 2}$ Intervalle $[c_0, c_1), \ldots, [c_{k-1}, c_k)$ zusammengefasst sind.
Über dem Intervall $[c_{j-1}, c_j)$ wird ein Rechteck (der Breite $c_j - c_{j-1}$) abgetragen, dessen *Fläche* proportional zur Anzahl der Beobachtungen ist, die in das Intervall fallen.

Histogramme, die die Häufigkeit jedes Wertes skizzieren, zeigen den Verlauf der empirischen Dichtefunktion.

Histogramme II

Histogramme

Zeichne über den Klassen $[c_0, c_1), \ldots, [c_{k-1}, c_k)$ Rechtecke mit

Breite:	$d_j = c_j - c_{j-1}$
Höhe:	proportional zu h_j/d_j bzw. f_j/d_j
Fläche:	proportional zu h_j bzw. f_j

Dabei seien h_j und f_j die absolute bzw. relative Zahl der Beobachtungen in $[c_{j-1}, c_j)$.

Beispiel: Mietspiegel München '03

- Wir lesen die Datentabelle `nettomieten.csv` mit dem `read.table`-Befehl ein.
- Wir erzeugen ein Histogramm mit dem `hist` Befehl aus der Nettomieten-Spalte der Datentabelle `nettomieten.csv`.
- R teilt die x-Achse in gleichgroße Intervalle. Die Anzahl der Intervalle wird automatisch auf ca. $\log_2 n$ festgelegt, wenn n die Anzahl der Beobachtungen ist.
- Es gibt weitere Optionen, z. B. `breaks="Scott"` und `breaks="Freedman-Diaconis"`. Der `breaks` Befehl kann auch mit einem Vektor verwendet werden, der angibt, an welchen Punkten ein neues Rechteck beginnen soll (auf der x-Achse).

Beispiel: Mietspiegel München '03

- Wir lesen die Datentabelle `nettomieten.csv` mit dem `read.table`-Befehl ein.
- Wir erzeugen ein Histogramm mit dem `hist` Befehl aus der Nettomieten-Spalte der Datentabelle `nettomieten.csv`.
- R teilt die x-Achse in gleichgroße Intervalle. Die Anzahl der Intervalle wird automatisch auf ca. $\log_2 n$ festgelegt, wenn n die Anzahl der Beobachtungen ist.
- Es gibt weitere Optionen, z. B. `breaks="Scott"` und `breaks="Freedman-Diaconis"`. Der `breaks` Befehl kann auch mit einem Vektor verwendet werden, der angibt, an welchen Punkten ein neues Rechteck beginnen soll (auf der x-Achse).

Beispiel: Mietspiegel München '03

- Wir lesen die Datentabelle `nettomieten.csv` mit dem `read.table`-Befehl ein.
- Wir erzeugen ein Histogramm mit dem `hist` Befehl aus der Nettomieten-Spalte der Datentabelle `nettomieten.csv`.
- R teilt die x-Achse in gleichgroße Intervalle. Die Anzahl der Intervalle wird automatisch auf ca. $\log_2 n$ festgelegt, wenn n die Anzahl der Beobachtungen ist.
- Es gibt weitere Optionen, z. B. `breaks="Scott"` und `breaks="Freedman-Diaconis"`. Der `breaks` Befehl kann auch mit einem Vektor verwendet werden, der angibt, an welchen Punkten ein neues Rechteck beginnen soll (auf der x-Achse).

Beispiel: Mietspiegel München '03

- Wir lesen die Datentabelle `nettomieten.csv` mit dem `read.table`-Befehl ein.
- Wir erzeugen ein Histogramm mit dem `hist` Befehl aus der Nettomieten-Spalte der Datentabelle `nettomieten.csv`.
- R teilt die x-Achse in gleichgroße Intervalle. Die Anzahl der Intervalle wird automatisch auf ca. $\log_2 n$ festgelegt, wenn n die Anzahl der Beobachtungen ist.
- Es gibt weitere Optionen, z. B. `breaks="Scott"` und `breaks="Freedman-Diaconis"`. Der `breaks` Befehl kann auch mit einem Vektor verwendet werden, der angibt, an welchen Punkten ein neues Rechteck beginnen soll (auf der x-Achse).

Der `hist`-Befehl

`hist(data)`	
`breaks=`	
`"Sturges"`	erzeugt ein Histogramm aus dem Datensatz `data`, wobei $k \approx \log_2(n) + 1$ Säulen mit $d_1 = \ldots = d_k$ verwendet werden
`"Scott"`	wie oben, jedoch mit $k \approx n^{1/3}$
`20`	wie oben, jedoch mit $k = 20$ Säulen
`c(c0,...,ck)`	x-Achse wird unterteilt in Intervalle $[c_0, c_1), \ldots, [c_{k-1}, c_k)$.

Verteilungen in R

Verteilung	*Name in R*	*Parameter in R*	*Parameter*
Binomial	binom	size	n
		prob	p
Hypergeometrisch	hyper	m	M
		n	$N - M$
		k	n
Poisson	pois	lambda	λ
Normal	normal	mean	μ
		sd	σ
Gleich	unif	min	a
		max	b
Exponential	exp	rate	λ
t	t	df	n
Chiquadrat	chisq	df	k
F	f	df1	m
		df2	n

s. S. 32 Skript von S. Ahlers

Aufruf von Verteilungen in R

Beispiel: Normalverteilung

- Dichtefunktion: `dnorm`
- Verteilungsfunktion: `pnorm`
- Quantilsfunktion: `qnorm`
- Zufallsdaten: `rnorm`

Funktionen plottet man in R mit den Befehlen `plot` und `curve`:

```
> curve(dgamma(x, shape = 5), from = 0, to = 20,
+ n = 200, type = "l")
>
> x<-seq(-3,3,length=10000)
> plot(x,dnorm(x),type="l"
```

Anwendung: Gesetz der großen Zahlen

```
> data <- rnorm(n = 100, mean = 0, sd = 1)
> hist(data, prob=T)
> x <- seq(from=par("usr")[1], to=par("usr")[2],
length=100)
> lines(x, dnorm(x, mean=0, sd=1), xpd=T, lwd=2)
```

Mit dem Befehl `lines` können nachträglich Funktionsverläufe in ein Diagramm gezeichnet werden.

Absolute und relative Häufigkeiten

$h(a_j) := \sum_{i=1}^{k} \mathbf{1}_{\{x_i = a_j\}}$ $(=: h_j)$	absolute Häufigkeit von a_j
$f(a_j) := \frac{h_j}{n}$	relative Häufigkeit von a_j
$h_1, \ldots, h_k$	absolute Häufigkeitsverteilung
$f_1, \ldots, f_k$	relative Häufigkeitsverteilung

$a_1, \ldots, a_k$ und $h_1, \ldots, h_k$ heißen Häufigkeitsdaten.

Kumulierte Häufigkeiten I

- Ordinalskaliertes Merkmal: Schulnote
- Häufigkeitstabelle: Klassenspiegel

sehr gut	gut	befriedigend	ausreichend	mangelhaft
4	7	10	8	2

- Kumulierte Häufigkeiten

sehr gut	gut	befriedigend	ausreichend	mangelhaft
4	11	21	29	31

- kumulierte Häufigkeiten: $\sum_{j=1}^{i} n_j$
- kumulierte relative Häufigkeiten: $\sum_{j=1}^{i} f_j$

Kumulierte Häufigkeiten II

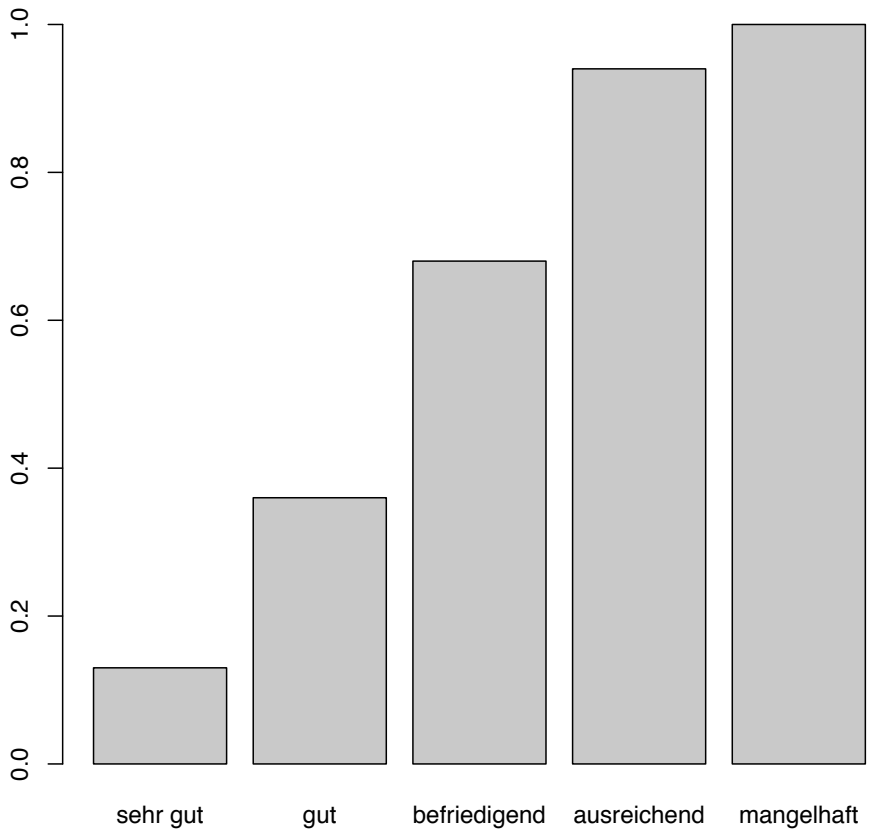
```
> noten <- c(3,4,2,3,4,1,3,1,3,4,4,2,2,2, ...)
> noten_tabelle.summiert <- cumsum(noten_tabelle)
> noten_tabelle.summiert

 sehr gut   gut   befriedigend   ausreichend   mangelhaft
 4          11    21             29            31

> noten_tabelle.relativ.summiert <-
+ cumsum(round(noten_tabelle/sum(noten_tabelle),2))

 sehr gut   gut    befriedigend   ausreichend   mangelhaft
 0.13       0.36   0.68           0.94          1.00
```

Kumulierte Häufigkeiten - Säulendiagramm



Kumulierte Häufigkeitsverteilung

Definition

Die *absolute kumulierte Häufigkeitsverteilung* eines (mindestens ordinalskalierten) Merkmals X ist durch die Funktion H mit

$$H(x) = \text{Anzahl der Werte } x_i \text{ mit } x_i \leq x$$

gegeben.

Bemerkung

Es gilt also auch

$$H(x) = \sum_{i:a_i \leq x} h_i$$

Die empirische Verteilungsfunktion

Definition

Die empirische Verteilungsfunktion F ist definiert durch

$$F(x) := H(x)/n = \sum_{i:a_i \leq x} f_i = n^{-1} \sum_{i=1}^{n} \mathbf{1}_{(-\infty,x]}(x_i).$$

Der Satz von Glivenko und Cantelli

Satz (von Glivenko und Cantelli)

Seien $X_1, X_2, \ldots$ eine Folge u. i. v. Zufallsgrößen mit Werten in $\mathbb{R}$. $F_n(\cdot) = F_n(\cdot, x_1, \ldots, x_n)$ sei die empirische Verteilungsfunktion von $x_1, \ldots, x_n$.
Dann konvergiert $F_n(\cdot, X_1, \ldots, X_n)$ für $n \to \infty$ P-f. s. gleichmäßig in $x \in \mathbb{R}$ gegen die Verteilungsfunktion F von X_1.

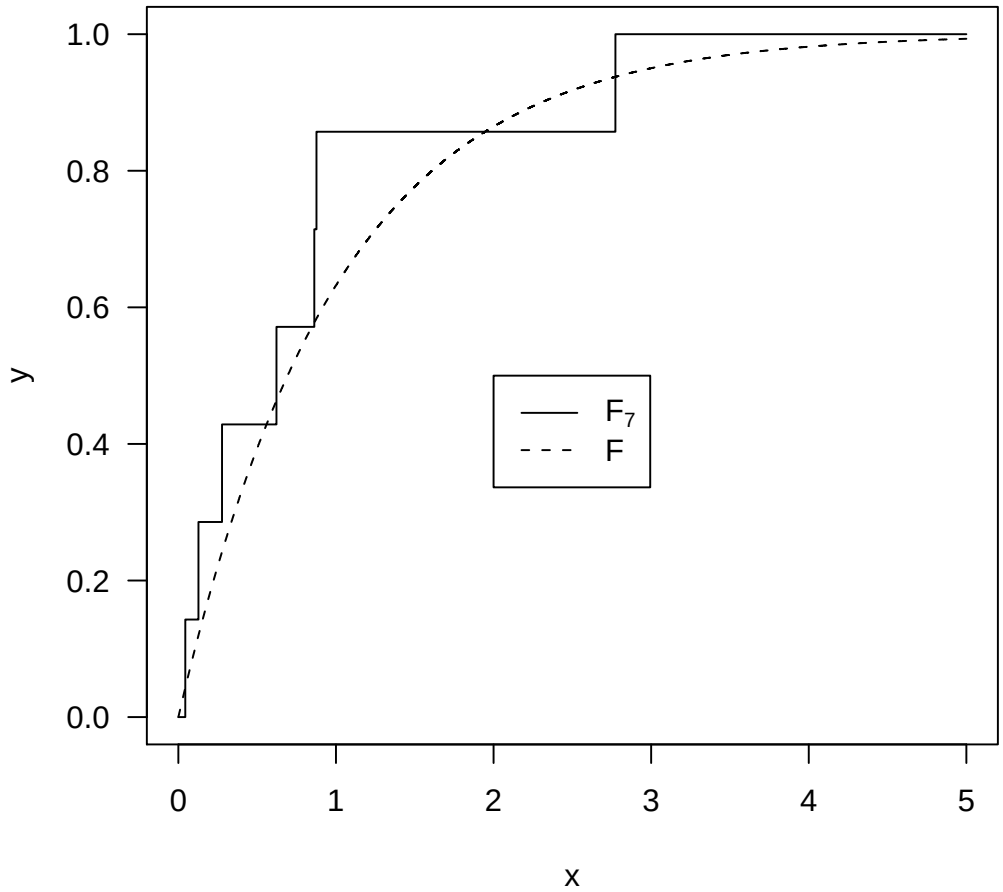


Abbildung: F_7 vs. F

Aufgaben

- Skizziere eine der folgenden Verteilungen:
  - Gamma-Verteilung
  - Exponentialverteilung
  - Binomialverteilung
  - Poissonverteilung

  Wie sieht die Verteilung von $X + Y$ aus, wobei X und Y $N(0, 1)$-verteilte Zufallsgrößen seien?
- Erstelle eine Zeichnung, die den Funktionsverlauf der Verteilungsfunktion Standardnormalverteilung zeigt. Füge dieser Zeichnung die empirische Verteilungsfunktion von 100 Zufallsdaten hinzu, die mittels der Standardnormalverteilung generiert sind.
  - Generiere einen Vektor mit den Zufallsdaten und sortiere diesen
  - Bilde die kumulierten Summen
  - Plotte den Vektor mit den kumulierten Daten auf einem geeigneten Intervall
  - Füge die Verteilungsfunktion der Normalverteilung mit ins Diagramm ein

Gliederung

Uni- und multivariate Analyse

- Univariate Analyse betrifft die Auswertung der Erhebung *eines* Merkmals.
- Multivariate Analyse betrifft die Auswertung der Erhebung *mehrerer* Merkmale

Gliederung

Darstellungsarten

- Tabellen, die ein- und mehrdimensionale Häufigkeiten zusammenfassen
- Gruppierung von Daten
- Diagramme
- Verlaufskurven
- Kenngrößen wie zum Beispiel Mittelwert, Median oder Streuung

Daten- / Häufigkeitstabellen

- *Fahrgastbefragung*

	ja	nein
Fahrt zum Arbeitsplatz		
Fahrt zum Studium/Schule		
Besuch von Familie/Freunden		
Einkauf/Shopping		
Urlaub		
Sonstiges		

Häufigkeitstabelle

- 1000 befragte Fahrgäste

	abs. Häufigk.	rel. Häufigk.
Fahrt zum Arbeitsplatz	203	0.2
Fahrt zum Studium/Schule	463	0.46
Besuch von Familie/Freunden	87	0.087
Einkauf/Shopping	101	0.1
Urlaub	4	0.004
Sonstiges	142	0.14

- Werte sind auf zwei Effektive Stellen gerundet
- Die Daten lassen sich grafisch darstellen

Darstellungsmöglichkeiten

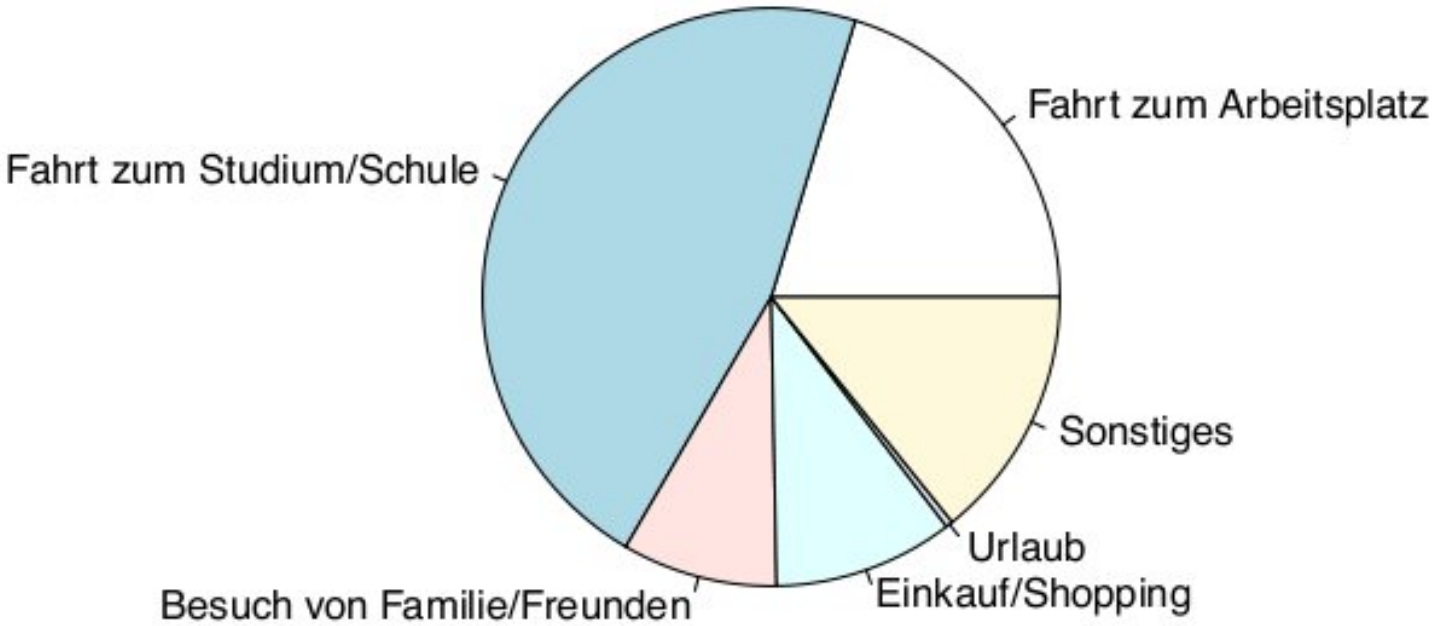
Diagramm	Beschreibung	Befehl in R
Stab-	$a_1, \ldots, a_k$ werden auf der x-Achse abgetragen, orthogonal zur x-Achse wird über a_j ein Strich proportional zu h_j abgetragen	`plot(..., type="h")`
Säulen-	wie Säulendiagramm mit Säulen statt Strichen	`barplot`
Balken-	wie Säulendiagramm, jedoch mit vertauschten Achsen	`barplot(..., horiz=TRUE)`
Kreis-	Flächen der Kreissektoren proportional zu den Häufigkeiten: $f_j \cdot 360°$	`pie`

Kuchendiagramm in R

```
> x <- c(203, 463, 87, 101, 4, 142)
> names(x) <- c("Fahrt zum Arbeitsplatz", ...)
> pie(x, labels = names(x))
```

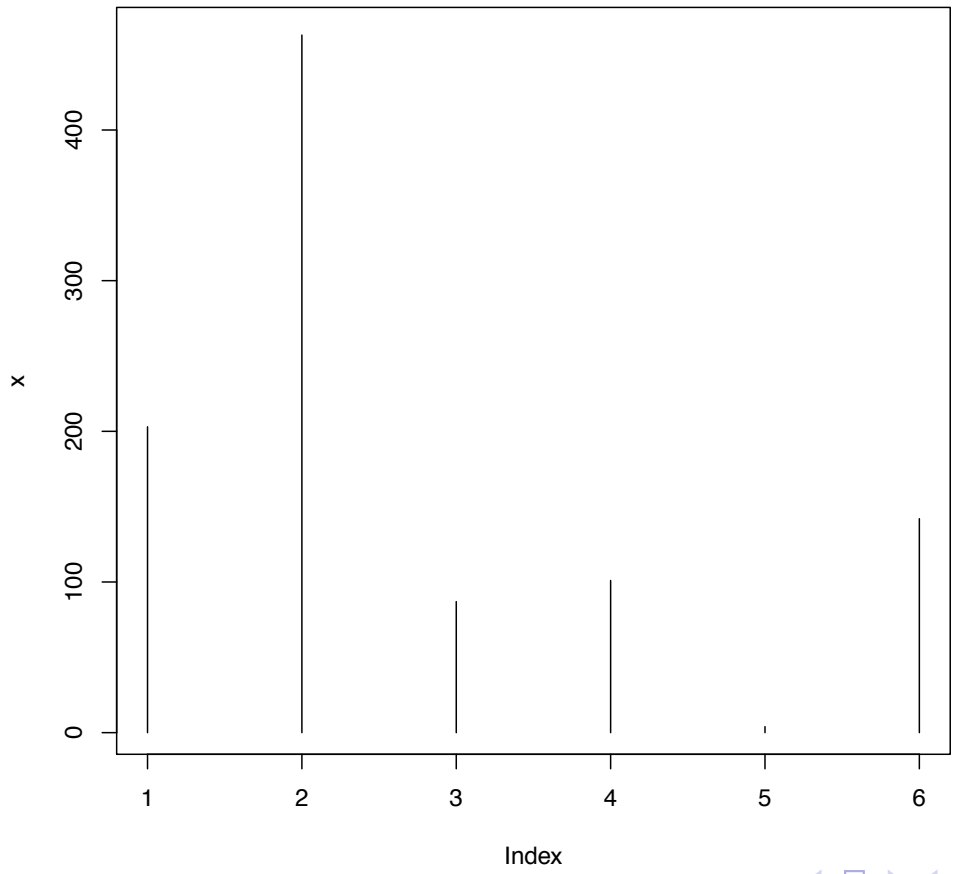
Die gezielte Zuweisung von Farben erfolgt mit dem Parameter "`col`". In R gibt es 9 Standardfarben und weitere Farbpakete, wie z.B. rainbow, heat.colors, terrain.colors, rgb.

Kuchendiagramm



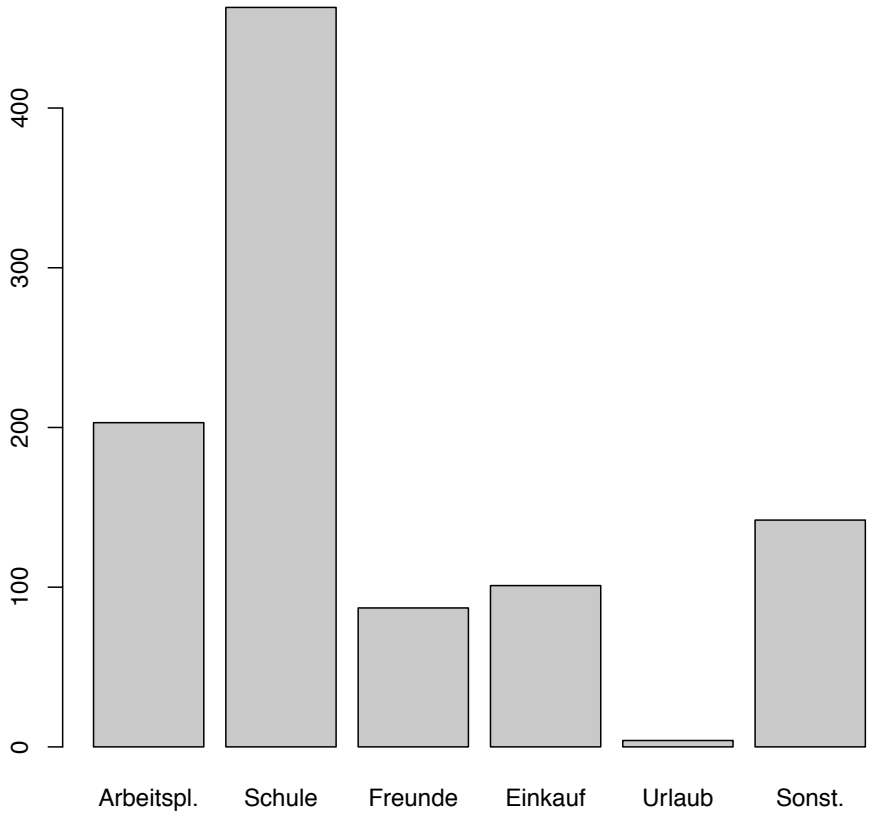
Stabdiagramm

```
> plot(x, type="h")
```



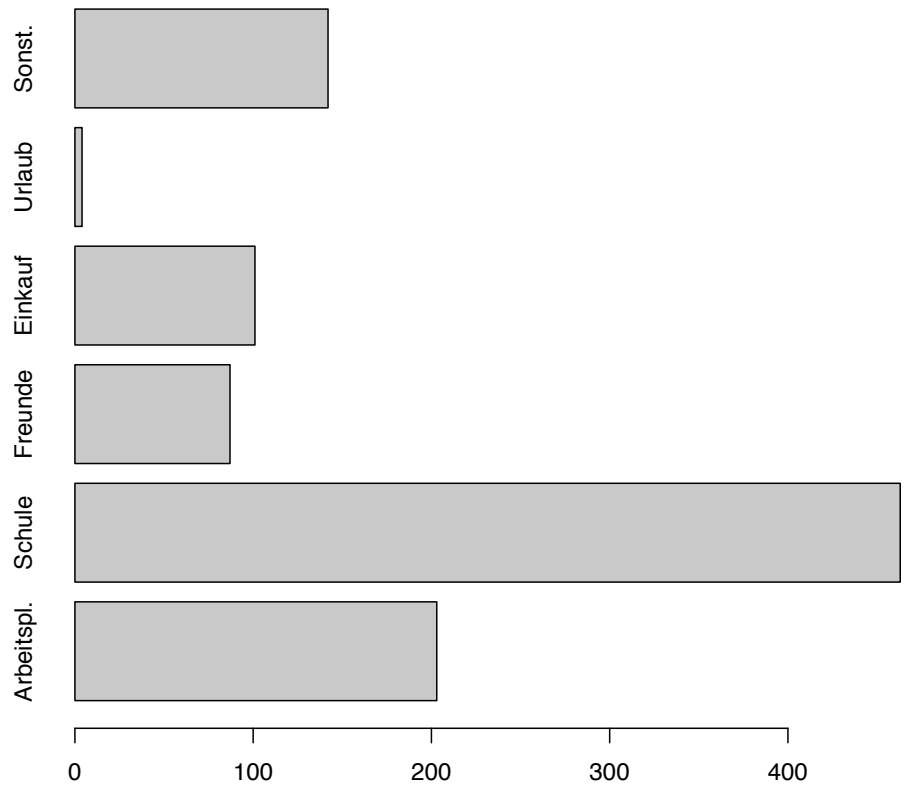
Säulendiagramm

```
> barplot(x)
```



Balkendiagramm

```
> barplot(x, horiz="T")
```



Aufgabe

Bei der letzten Bundestagswahl in Deutschland (im September 2009) ergab sich folgende Stimmverteilung beim Merkmal *Zweitstimme* (bei 44.005.575 Stimmen):

CDU/CSU	SPD	FDP	Die Linke	Grüne	Andere
14.658.515	9.990.488	6.316.080	5.155.933	4.643.272	3.241.287

(a) Geben Sie die Daten als Vektor ein und ordnen Sie den Vektor absteigend. Berechnen Sie die zugehörigen prozentualen Anteile an den abgegebenen (und gültigen) Stimmen auf eine Nachkommastelle genau.

(b) Erzeugen Sie mit den Daten aus (a) ein mit den Parteinamen und den zugehörigen Prozentzahlen beschriftetes Kreissektorendiagramm (in den entsprechenden Parteifarben).

(c) Erstellen Sie ein geordnetes Säulendiagramm in den entsprechenden Parteifarben.

Gliederung

Urliste und sortierte Daten

Von der Urliste $x_1, \ldots, x_n$ können wir zur geordneten Liste $x_{(1)} \leq \ldots \leq x_{(n)}$ übergehen.

Dies geschieht in R mit dem Befehl `sort`.

Klasseneinteilung - Der Befehl `cut`

Stetige Merkmale können in Klassen eingeteilt werden. Dies geschieht in R mit dem Befehl `cut`. Dieser ersetzt die Werte eines Vektors durch die Klasse, innerhalb derer er liegt.

Beispiel: Größenmessung im Kindergarten

```
> groesse <- c(103,105,106, ...)
> klass.groesse <-
+ cut(groesse,c(85,90,95, ...),include.lowest=TRUE)
> klass.groesse
[1] (100, 105] (100, 105] (105, 110] ...
Levels: [85,90] (90,95] (95,100] (100,105] (105,110]
(110,115] (115,120]
```

Für äquidistante Klassen (wie hier) kann der Befehl `seq` genutzt werden.

Grafisch: Häufigkeitsverteilung / Histogramm

Im falle eines *stetigen* quantitativen Merkmals, ist eine Häufigkeitsverteilung nicht mehr aussagekräftig.

In diesem Fall ist ein *Histogramm* hilfreich, das annähernd den Verlauf der empirischen Verteilung skizziert. Die Häufigkeiten werden dabei auf Teilintervallen zusammengefasst. Ein Histogramm erzeugt man mit dem Befehl `hist`.

Probleme bei der Darstellung mittels Histogramm ergeben sich, falls die Daten über ein sehr großes Intervall gestreut sind und nicht beschränkt sind. Dann können die Säulen die Höhe 0 haben.

Unimodale und multimodale Verteilungen

Viele (empirische) Verteilungen weisen eines der folgenden Verhalten auf:

- Im Histogramm gibt es einen Gipfel, von dem aus die Häufigkeiten zu den Randbereichen abfallen, ohne dass ein zweiter Gipfel auftritt. Solche Verteilungen heißen *unimodal*.
- Tritt ein zweiter (und kein weiterer) Gipfel auf, so heißt die Verteilung *bimodal*.
- Treten weitere Nebengipfel auf, so heißt die Verteilung *multimodal*.

Symmetrie

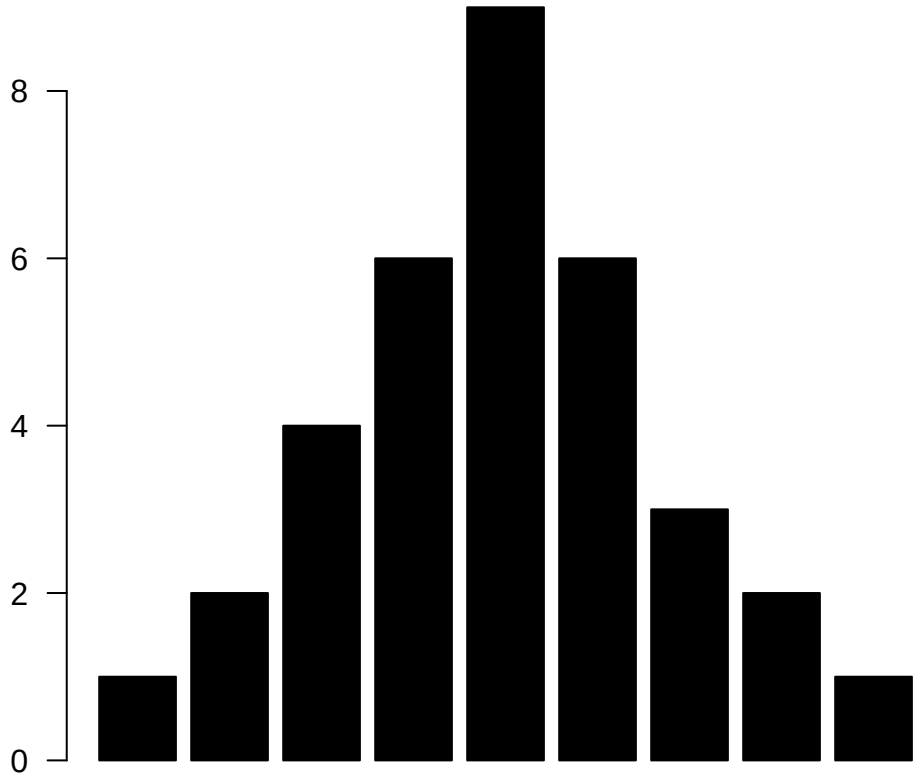
Definition

Eine (empirische) Verteilung heißt *symmetrisch*, wenn es eine Symmetrieachse gibt, so dass die linke und die rechte Hälfte der Verteilung annähernd spiegelbildlich zueinander sind.

Bemerkung

Exakte Symmetrie ist bei empirischen Verteilungen selten gegeben.

Eine symmetrische Datenverteilung



Schiefe

Definition

Eine (empirische) Verteilung heißt *linkssteil* oder *rechtsschief*, wenn der überwiegende Anteil der Daten linksseitig konzentriert ist.
Analog heißt eine (empirische) Verteilung *rechtssteil* oder *linksschief*, wenn der überwiegende Anteil der Daten rechtsseitig konzentriert ist.

- Typische Beispiele für linkssteile Verteilungen sind Einkommensverteilungen.

Schiefe

Definition

Eine (empirische) Verteilung heißt *linkssteil* oder *rechtsschief*, wenn der überwiegende Anteil der Daten linksseitig konzentriert ist. Analog heißt eine (empirische) Verteilung *rechtssteil* oder *linksschief*, wenn der überwiegende Anteil der Daten rechtsseitig konzentriert ist.

- Typische Beispiele für linkssteile Verteilungen sind Einkommensverteilungen.

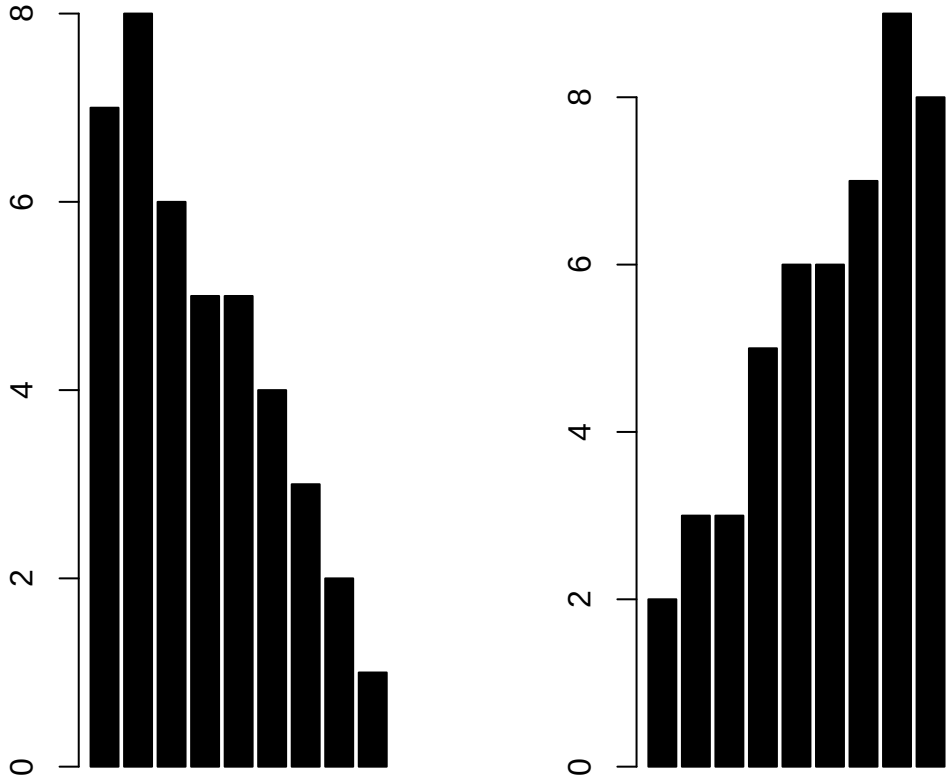


Abbildung: Links- bzw. rechtsschiefe Daten

Aufgaben

- Lesen Sie den Datensatz `nettomieten.csv` ein. Die erste Spalte mit dem Namen `bla` gibt das Mermal “Nettomiete” wieder. Lesen Sie diese in einen Vektor ein.
- Nehmen Sie eine geeignete Klasseneinteilung vor. Zeichnen Sie anschließend ein Säulendiagramm des Datensatzes.
- Listen Sie die absoluten Häufigkeiten auf (Tipp: Nutzen Sie den Befehl `table`)
- Zeichnen Sie ein Histogramm.
- Beurteilen Sie die Schiefe der Verteilung.

Gliederung

Beschreibung von Verteilungen

Bei der Datenanalyse, z. B. der Analyse des Nettomietniveaus in München, ergeben sich häufig Fragen der folgenden Art:

- Wo liegt das Zentrum der Daten?
- Wie stark streuen die Daten um das Zentrum?
- Ist die Verteilung symmetrisch oder schief?
- Gibt es Ausreißer?

Das arithmetische Mittel

Definition

Das *arithmetische Mittel* wird aus der Urliste $x_1, \ldots, x_n$ durch

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

berechnet.

Für Häufigkeitsdaten mit Ausprägungen $a_1, \ldots, a_k$ und relative Häufigkeiten $f_1, \ldots, f_k$ gilt

$$\overline{x} = \sum_{i=1}^{k} f_i a_i.$$

In R lässt sich das arithmetische Mittel eines Vektors x mit dem Befehl `mean(x)` berechnen.

Eigenschaften des arithmetischen Mittels

- Das arithmetische Mittel ist für kardinalskalierte Daten sinnvoll.
- Das arithmetische Mittel besitzt die *Schwerpunkteigenschaft*

$$\sum_{i=1}^{n}(x_i - \overline{x}) \;=\; 0.$$

- Das arithmetische Mittel reagiert empfindlich auf extreme Werte und *Ausreißer*.
  (Man ersetze den größten Wert in der Nettomietenliste durch 20000)
- Das arithmetische Mittel stimmt i. A. mit keiner der möglichen Ausprägungen überein.

Resistente/Robuste Lagemaße

Definition

Ein Lagemaß heißt *resistent* oder *robust*, falls es unempfindlich gegenüber extremen Werten/Ausreißern ist.

Der (Stichproben-)Median

Ein robustes Lagemaß ist der Median. Um ihn zu bilden, betrachtet man die geordnete Liste $x_{(1)}, ..., x_{(n)}$.

Definition

Der *Median* x_{med} von $x_1, \ldots, x_n$ ist durch

$$x_{\text{med}} = \begin{cases} x_{(\frac{n+1}{2})}, & \text{falls } n \text{ ungerade ist,} \\ \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}), & \text{falls } n \text{ gerade ist,} \end{cases}$$

definiert.

Der Median wird in R mit dem Befehl `median` berechnet.

Eigenschaften des Medians

- Der Median ist ab dem Ordinalskalenniveau sinnvoll.
- Der Median x_{med} ist robust gegenüber Ausreißern (Beispiel: Nettomietendatensatz).
- Mindestens 50% der Daten sind $\geq x_{\text{med}}$ und mindestens 50% der Daten sind $\leq x_{\text{med}}$.
- Der Median stimmt i. A. mit keiner der möglichen Ausprägungen überein.

  Wann stimmt der Median mit einer tatsächlichen Ausprägung überein?

Der Modus

Ein weiteres gebräuchliches Lagemaß ist der Modus.

Definition

Ein Modus x_{mod} ist eine Ausprägung mit größter Häufigkeit.

Bemerkung

Der Modus ist eindeutig, falls die Häufigkeitsverteilung ein eindeutiges Maximum besitzt.

Eigenschaften des Modus'

- Der Modus ist bereits auf Nominalskalenniveau sinnvoll.
- Der Modus ist robust.
- Der Modus ist eine Ausprägung des Merkmals.

Lageregeln

Symmetrische Verteilungen:	$\overline{x} \approx x_{\text{med}} \approx x_{\text{mod}}$
Linkssteile Verteilungen:	$x_{\text{mod}} < x_{\text{med}} < \overline{x}$
Rechtssteile Verteilungen:	$\overline{x} < x_{\text{med}} < x_{\text{mod}}$

Gruppierte Lagemaße

Liegen die Daten nicht als Urliste sondern gruppiert vor, so kann man nur Näherungswerte der Lagemaße bilden.

Modus:	Bestimme Modalklasse (Klasse mit der größten Beobachtungszahl) und verwende Klassenmitte als Modus
Median:	Bestimme Einfallsklasse $[c_{i-1}, c_i)$ des Medians und daraus $x_{\text{med, grupp}} = c_{i-1} + \frac{d_i(0{,}5 - F(c_{i-1}))}{f_i}$.
Arithm. Mittel:	$\overline{x}_{\text{grupp}} = \sum_{i=1}^{k} f_i m_i$.

Gruppierte Lagemaße II

Bei der Bildung der gruppierten Lagemaße ergibt sich:

- Der wahre Modus muss nicht einmal in der Modalklasse liegen.
- Der wahre Modus muss nicht mit einem Beobachtungswert zusammenfallen.

Streuung

Folgende Maßzahlen messen die Abweichung quantitativer Daten von ihrem Zentrum:

- Mittlere absolute Abweichung $\frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}|$
- Mittlere quadratische Abweichung $d^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 =: \overline{x^2} - \overline{x}^2$
- Stichprobenvarianz $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{n}{n-1}d^2$
- Spannweite $R(x) = x_{(n)} - x_{(1)}$ (ausreißerempfindlich)
- Interquartilsabstand $IQR(x) = x_{0.75} - x_{0.25}$

Die Stichprobenvarianz eines Vektors `x` wird in R mit dem Befehl `var(x)` berechnet.

Die Spannweite lässt sich durch `diff(range(x))` berechnen.

Aufgabe

Wir betrachten wieder den Datensatz `nettomieten.csv`

- Berechnen Sie arithmetisches Mittel und den Median der Spalte mit den Nettomieten.
- Berechnen Sie das arithmetische Mittel, den Median und den Modus der gruppierten Daten. Was fällt Ihnen auf?
- Berechnen Sie die Varianz und die Spannweite der Daten.

Gliederung

Quantile

Definition

Für $0 < p < 1$ heißt jeder Wert q_p, für den ein Anteil von mindestens p der Daten $\leq q_p$ und mindestens ein Anteil von $1 - p \geq q_p$ ist, *p-Quantil*.

Bemerkung

Für ein p-Quantil gilt

$$\begin{aligned} q_p &= x_{([np]+1)}, && \text{wenn } np \text{ nicht ganzzahlig ist,} \\ q_p &\in [x_{(np)}, x_{(np+1)}], && \text{wenn } np \text{ ganzzahlig ist.} \end{aligned}$$

Quantile in R

- In R werden Quantile mit dem `quantile`-Befehl aufgerufen.
- Man bestimme die Quartile des Nettomieten-Datensatzes.
- Gibt die Lage der Quartile im Vergleich zum Median Aufschluss bzgl. der Schiefe des Datensatzes?

Quantil einer Verteilung

Entsprechend ist die Quantilsfunktion F^{-1} einer Verteilung Q auf $(\mathbb{R}, \mathfrak{B})$ definiert:

Definition (Quantilsfunktion)

$$\begin{aligned} F^{-1}(p) &= \inf\{x \in \mathbb{R} : F(x) \geq p\} \\ &= \inf\{x \in \mathbb{R} : Q((x, \infty)) \leq 1 - p\} \text{ für } p \in (0, 1) \end{aligned}$$

Sie wird auch als Pseudo-Inverses der Verteilungsfunktion oder als $1 - p$-Fraktil bezeichnet. Ihr Aufruf in R erfolgt mittels "`q` + *Name der Verteilung*".

Quartile

Definition

Ein *unteres Quartil* ist ein 25%-Quantil,
ein *oberes Quartil* ein 75%-Quantil.

Fünf-Punkte-Zusammenfassung

Definition

Die *Fünf-Punkte-Zusammenfassung* besteht aus

dem Minimum, dem ersten Quartil, dem Median, dem dritten Quartil und dem Maximum

des Datensatzes.

Die Fünf-Punkte-Zusammenfassung ist in R im Befehl `summary` enthalten.

Boxplots

In einem Boxplot eines Datensatzes $x_1, \ldots, x_n$ werden in ein Koordinatensystem

- ein Rechteck (eine Box) gezeichnet, die auf der y-Achse nach oben gegen das obere Quartil und nach unten gegen das untere Quartil begrenzt ist,
- eine Horizontale auf der Höhe des Medians durch die Box gelegt,
- vertikale Linien eingezeichnet, die sogenannten *Whiskers*, von der Box nach oben und nach unten bis $\min\{q_{3/4} + 3/2(q_{3/4} - q_{1/4}), x_{[n]}\}$ bzw. bis $\max\{q_{1/4} - 3/2(q_{3/4} - q_{1/4}), x_{[1]}\}$, wo die Linien durch kurze horizontale Linien begrenzt werden.

Boxplots II

- Werte jenseits der Whiskers werden in den Boxplot durch $\circ$ oder $\times$ markiert.
- Die Differenz $q_{3/4} - q_{1/4}$ heißt *Interquartilsabstand* (IQR).
- Werte, die jenseits der Whiskers liegen, heißen *Outlyer*, wenn sie im Bereich $[q_{1/4} - 3\mathrm{IQR}, q_{3/4} + 3\mathrm{IQR}]$ liegen.
- Werte außerhalb dieses Bereichs werden *Extremwerte* genannt.

NQ-Plots: Idee

Ein Zweck der Berechnung der empirischen Verteilungsfunktion kann die Überprüfung der Normalverteilungsannahme sein: Ist es statthaft anzunehmen, dass die Daten normalverteilt sind? Diese Fragestellung ist mit einem Normal-Quantil-Plot leichter zugänglich.
Bei diesem Plot trägt man in einem Koordinatensystem die kt kleinste Beobachtung auf der y-Achse gegen die erwartete kt-kleinste Beobachtung eines Vektors mit n standardnormalverteilten Zufallsgrößen ab.
Unabhängig von Erwartungswert und Varianz sollte sich bei normalverteilten Daten eine Gerade abzeichnen.

NQ-Plot

Definition

Sei $x_{(1)}, \ldots, x_{(n)}$ die geordnete Urliste. Für $i = 1, \ldots, n$ werden die $(i - 1/2)/n$-Quantile $z_{(i)}$ der $\mathcal{N}(0, 1)$-Verteilung berechnet. Der *Normal-Quantil-Plot (NQ-Plot)* besteht aus den Punkten

$$(z_{(1)}, x_{(1)}), \ldots, (z_{(n)}, x_{(n)})$$

im z-x-Koordinatensystem.

Bemerkung

Sind die Daten normalverteilt mit Erwartungswert μ und Varianz σ^2, so liegen die Daten in etwa auf der Geraden $x = \mu + \sigma z$.

QQ-Plots in R

Bei einem Quantile-Quantile-Plot werden die Quantile zweier statistischer Variablen gegeneinander abgetragen werden, um ihre Verteilungen zu vergleichen.

- Einen NQ-Plot erhält man mit dem Befehl `qqnorm`.
- Einen QQ-Plot erhält man mit dem Befehl `qqplot`.

Aufgabe

- Zeichnen Sie den NQ-Plot des Nettomietendatensatzes. Zeichnen Sie anschließend zum Vergleich einige NQ-Plots eines mit der Normalverteilung generierten Datensatzes. Wählen Sie dazu Anzahl der Zufallsdaten, sowie Mittelwert und Varianz entsprechend zu den Daten aus dem Nettomietendatensatz.
- Erzeugen Sie einen NQ-Plot jeweils eines Zufallsdatensatzes, der mit der $B(0.1, 10)$-, der $B(0.9, 10)$- bzw. der $\Gamma(5, 1)$-Verteilung generiert werde.
- Man verwende den Befehl `boxplot`, um mit R ein Boxplot des Nettomietendatensatzes zu erzeugen.
- Was lässt sich aus dem NQ-Plot, bzw. dem Boxplot hinsichtlich der Schiefe des Nettomietendatensatzes schließen?