Datenerhebung und deskriptive Statistik

Blockpraktikum zur Statistik mit R

27. März 2012

Sören Gröttrup

Institut für Mathematische Statistik
Universität Münster
SS 2012

Gliederung

Gliederung

Literatur


Silke Ahlers
Einführung in die Statistik mit R
http://wwwmath.uni-muenster.de/statistik/lehre/SS12/PrakStat/Skript.pdf


Peter Dalgaard
Introductory Statistics with R
Springer


Fahrmeir, Künstler, Pigeot, Tutz
Statistik. Der Weg zur Datenanalyse
Springer


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

Was ist Statistik?

Wikipedia:

Statistik ist *die Lehre von Methoden zum Umgang mit quantitativen Informationen* (Daten). [...] Sie ist damit unter anderem die Zusammenfassung bestimmter Methoden, um empirische Daten zu analysieren. [...]
Statistik wird einerseits als eigenständige mathematische Disziplin über das *Sammeln, die Analyse, die Interpretation oder Präsentation von Daten* betrachtet, andererseits als Teilgebiet der Mathematik, insbesondere der Stochastik, angesehen.

Aufgaben der Statistik:

- Daten sammeln
- Daten darstellen und analysieren
- Daten interpretieren
- Prognosen und Entscheidungen treffen

Beispiel: Klassenspiegel

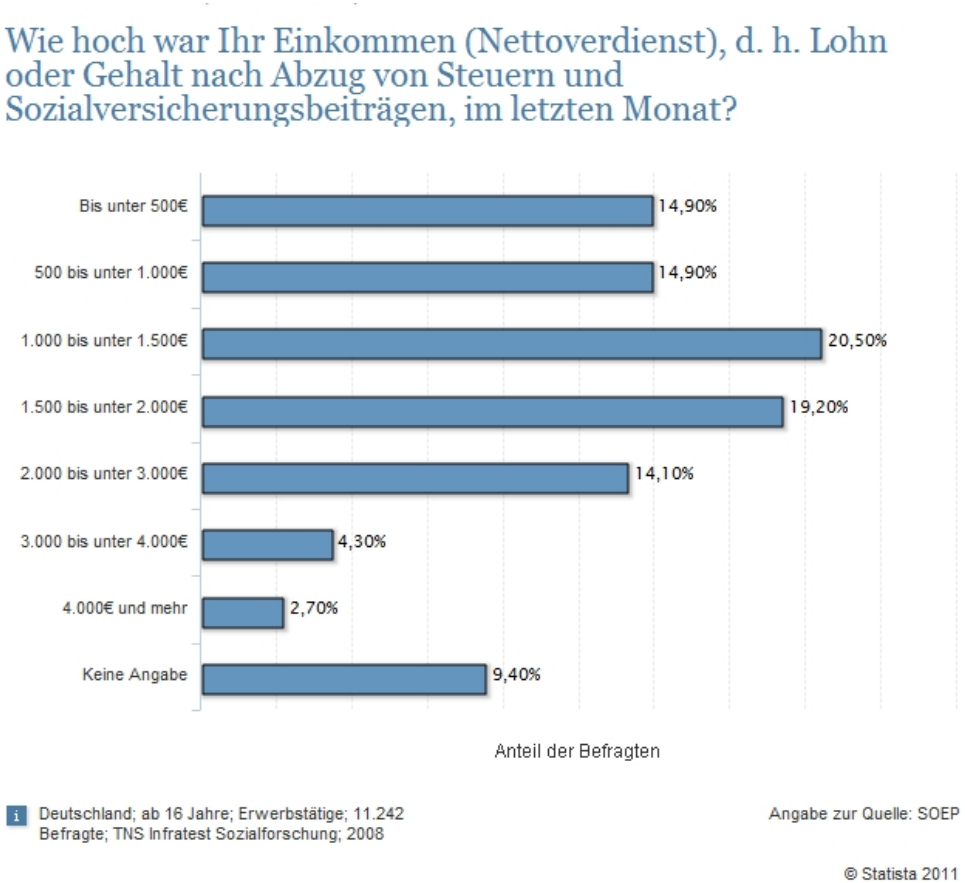
- 30 Schüler bekommen ihre Klausur zurück.
- *Ziel*: Durchschnittsnote berechnen und Notenverteilung skizzieren

Noten	1	2	3	4	5	6
	3	8	11	6	2	0

Durchschnittsnote: 2.866

Beispiel: Einkommensverteilung

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



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

Arten der Datenanalyse

Man unterscheidet zwischen *deskriptiver (beschreibender)*, *explorativer (suchender)* und *induktiver (schließender)* Datenanalyse.

deskriptiv: Beschreiben und Darstellen von Daten & Datenvalidierung

- Grafiken (Tortendiagramm, Histogramm, Boxplot, ...)
- Tabellen und Kenngrößen (Mittelwert, Streuung, ...)
- keine Stochastik

explorativ: Auffinden von Stukturen und Besonderheiten in Daten

- Falls Fragestellung oder stat. Modell nicht genau bekannt
- Ableiten von Fragestellungen und Hypothesen
- keine Stochastik

induktiv: Über Daten hinaus Aussagen über Grundgesamtheit treffen

- sorgfältige Versuchsplanung und deskriptive/explorative Voranalyse
- Testen (Ist Therapie A besser als Therapie B?)
- Schätzen (Wie lange hält eine Glühbirne?)
- braucht Stochastik, Wahrscheilichkeitstheorie

Gliederung

Erhebung von Daten

Arten der Datenerhebung

- Befragung (schriftlich/mündlich/telefonisch; offen/geschlossen)
- Beobachtung
- Experiment

Aufkommende Fragen

- Ziel der Datenerhebung?
- Was will man Fragen/Beobachten?
- Wen will man Befragen/Beobachten?
- Wie will man Daten erheben?
- Datenquelle? (neue Erhebung, ältere Studien, verarbeitete Rohdaten)
- Gesetzliche Grundlagen (Datenschutz, Tierschutz,...)

Erhebung von Daten

Arten der Datenerhebung

- Befragung (schriftlich/mündlich/telefonisch; offen/geschlossen)
- Beobachtung
- Experiment

Aufkommende Fragen

- Ziel der Datenerhebung?
- Was will man Fragen/Beobachten?
- Wen will man Befragen/Beobachten?
- Wie will man Daten erheben?
- Datenquelle? (neue Erhebung, ältere Studien, verarbeitete Rohdaten)
- Gesetzliche Grundlagen (Datenschutz, Tierschutz,...)

Grundlegende Begriffe

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 Teilgesamtheit
Merkmal (Variable):	interessierende Größe
Merkmalsausprägung:	konkreter Wert des Merkmals für eine bestimmte statistische Einheit

Beispiel: Mietspiegel

- Städte und Gemeinden erstellen Mietspiegel
- bieten Mietern und Vermietern eine Marktübersicht zu Miethöhen
- helfen in Mietberatungsstellen
- 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“.*

Beispiel: Mietspiegel

Statistische Einheiten: Wohnungen, an denen die interessierenden Größen erfaßt werden

Grundgesamtheit: Menge aller Wohnungen in Münster

Stichprobe: Wohnungen, deren Daten erfasst wurden

Merkmale: Baujahr, Größe, Preis/qm

Merkmalsausprägungen:

- *Baujahr:* „bis 1929“,..., „2004-2005“
- *Größe:* „21-30 qm“, ..., „151-160qm“
- *Preis/qm:* $x \in (0, \infty)$

Ziel- und Einflussgrößen

- Man unterscheidet Variablen, die beeinflußt werden, „*Zielgrößen*“, und solche, die beeinflussen.
- Die beeinflussenden Variablen werden aufgeteilt in beobachtbare Variablen, „*Einflussgrößen* oder *Faktoren*“, und in nicht beobachtbare Variablen, „*Störgrößen*“.
- Störgrößen kann z.B. mit *randomiesieren* entgegengewirk werden

Beispiel (Mietspiegel)

- *Zielgröße:* Nettomiete/qm
- *Einflussgrößen:* Baujahr, Größe, Badausstattung, Lage, ...
- *Störgrößen:* nicht erhobene Ausstattungsmerkmale, unbekannte Gewohnheiten von Mieter und Vermieter, ...

Stichprobenarten

Man unterscheidet zwischen einer *Vollerhebung* (Erfassung aller statistischen Einheiten einer Grundgesamtheit) und *Teilerhebung* (Ziehen einer *Stichprobe*).

- Vollerhebung nicht immer möglich $\rightsquigarrow$ Stichprobe
- Stichprobenarten: *einfache Zufallsstichprobe*, *geschichtete Zufallsstichprobe*, *Klumpenstichprobe*, *bewußtes Auswahlverfahren*.

Einfache Zufallsstichprobe:

- zufälliges Ziehen aus der Grundgesamtheit
- technisch schwer Umsetzbar
- Ziehungsmethode kann systematischen Fehler enthalten

Bewußtes Auswahlverfahren:

- Stichprobe wird vom Interviewer ausgewählt
- z.B. *Quotenauswahl*, gleiche %te in Grundgesamtheit und Stichprobe
- Vor- und Nachteil: Kontrolle durch den Interviewer

Stichprobenarten

Man unterscheidet zwischen einer *Vollerhebung* (Erfassung aller statistischen Einheiten einer Grundgesamtheit) und *Teilerhebung* (Ziehen einer *Stichprobe*).

- Vollerhebung nicht immer möglich $\rightsquigarrow$ Stichprobe
- Stichprobenarten: *einfache Zufallsstichprobe*, *geschichtete Zufallsstichprobe*, *Klumpenstichprobe*, *bewußtes Auswahlverfahren*.

Einfache Zufallsstichprobe:

- zufälliges Ziehen aus der Grundgesamtheit
- technisch schwer Umsetzbar
- Ziehungsmethode kann systematischen Fehler enthalten

Bewußtes Auswahlverfahren:

- Stichprobe wird vom Interviewer ausgewählt
- z.B. *Quotenauswahl*, gleiche %te in Grundgesamtheit und Stichprobe
- Vor- und Nachteil: Kontrolle durch den Interviewer

Geschichtete Zufallsstichprobe

- Grundgesamtheit wird in disjunkte Gruppen (Schichten) zerlegt
- Aus jeder Schicht wird eine zufällige Anzahl gezogen
- einfacher umsetzbar und repräsentativer als einfache Zufallsstichprobe

Beispiel (Bundestagswahl)

- Alter, Geschlecht, Bildungsstatus, etc. beeinflussen das Wahlverhalten
- Wahlberechtigten gemäß den Einflussgrößen unterteilen

Klumpenstichprobe

- Grundgesamtheit wird in Gruppen (Klumpen) zerlegt
- zufällige Auswahl ganzer Klumpen $\rightsquigarrow$ Vollerhebung der Klumpen
- Sinnvoll, falls Klumpen „kleines“ Abbild der Grundgesamtheit und untereinander homogen

Beispiel (Einkommensverteilung in Ost- und Westdeutschland)

- Klumpen sind Gemeinden in Ost und West
- In ausgewählten Gemeinden Daten die dortigen Finanzämter untersuchen

Verzerrte Stichproben

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

Verzerrung (Bias)	Ursache und Beispiel
Selektion-Bias	bewusster Ausschluss von Elementen von der Ziehung Bsp: Internet- oder Zeitungsumfrage
Nonresponse-Bias	(unangenehme) Fragen bleiben unbeantwortet Bsp: Fragen zum Sexualverhalten etc.
Selfselection-Bias	Umfragen auf freiwilliger Basis Bsp: Evaluation von Lehrveranstaltungen

Studiendesigns

Studientyp

Querschnittstudie	an einer bestimmten Anzahl von Objekten wird zu einem bestimmten Zeitpunkt ein Merkmal oder mehrere erfasst Bsp: Mietspiegel
Zeitreihe	ein Objekt wird hinsichtlich eines Merkmals über einen ganzen Zeitraum beobachtet Bsp: Verlauf eines Aktienkurses
Längsschnittstudie	eine Gruppe wird hinsichtlich eines Merkmals über einen ganzen Zeitraum beobachtet Bsp: Verlauf eines Aktien-Portfolios

Mermale und Ausprägungen - Mietspiegel

- *Baujahr:* „bis 1929“, „1930-1945“, ..., „2004-2005“, „nach 2006“
- *Größe:* „$\leq$ 20 qm“, „21-30 qm“, ..., „151-160 qm“, „$\geq$ 161 qm“
- *Preis/qm:* $x \in (0, \infty)$
- *Badausstattung:* „mit Badewannde“, „ohne Badewanne“

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

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

Beispiel (Mietspiegel)

- *diskret:* Baujahr, Größe, Badeausstattung
- *stetig:* Preis/qm
- *quasi-stetig:* Preis/qm

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

Beispiel

- *nominalskaliert:* Badeausstattung (Mietspiegel), Geschlecht
- *ordinalskaliert:* Baujahr (Mietspiegel), Schulnoten
- *intervallskaliert:* Temperatur in Celsius
- *verhältnisskaliert:* Preis/qm (Mietspiegel)

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

Qualitative und quantitative Mermale

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

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

Der Abschnitt 5 (Grundlegende Definitionen) des Aufgabenblattes kann jetzt bearbeitet werden.

Gliederung

Punkte plotten

- Mit der Funktion `plot(x, y, type='p')` zeichnet man Punkte mit den x-Werten `x` und y-Werten `y` in ein Koordinatensystem. Dabei müssen die Vektoren `x` und `y` die gleiche Länge haben. Wählt man als Typ `'l'`, wird eine Linie durch die Punkte gezeichnet.
- Mit `points(x, y)` kann man in eine bestehende Grafik weitere Punkte einfügen.
- `lines(x,y)` ist das selbe wie `points(x, y, type='l')`.
- Weitere nützliche Parameter sind unter anderem: `type`, `pch`, `lty`, `cex`, `col`, `main`, `xlab`, `ylab`

Beispiel

```
plot(c(2,6,4), c(1,-3,0))
x <- seq(-10,10,length=30)
plot(x, x, main='Gerade', xlab='x', ylab='y', type='o')
points(x, x^2-4, type='l', col='red')
```

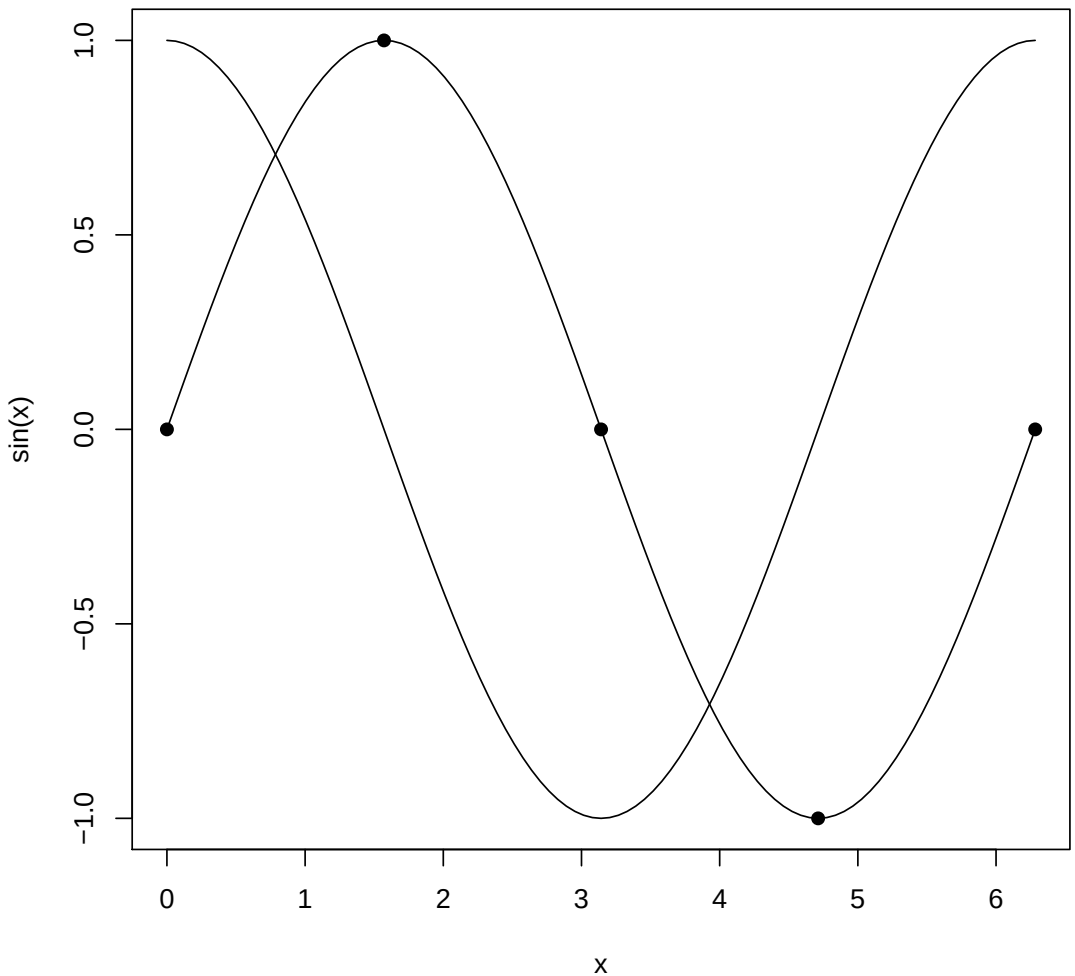
Kurven plotten

- Mit `curve(expr, from, to)` kann man die Funktion `expr` von `from` bis `to` zeichnen. Die Standardvariable der Funktion ist `x`.
- Mit dem Parameter `add=TRUE` können weitere Kurven in eine Grafik gezeichnet werden.
- Weitere Parameter siehe `help(curve)`

Beispiel

```
curve(sin(x), from=0, to=2*pi)
points(c(0,pi/2,pi,3/2*pi,2*pi), c(0,1,0,-1,0), pch=19)
curve(cos(x), from=0, to=2*pi, add=TRUE)
```

Sinus- und Cosinus-Kurven



Verteilungen in R

Verteilung	*Name in R*	*Parameter in R*	*Parameter*
$B(n,p)$	binom	size, prob	n, p
$Poisson(\lambda)$	pois	lambda	λ
$N(\mu,\sigma^2)$	norm	mean, sd	μ, σ
$R(a,b)$	unif	min, max	a, b
$Exp(\lambda)$	exp	rate	λ
$\Gamma(n,\lambda)$	gamma	shape, rate	n, λ
t (k Freiheitsgr.)	t	df	k
χ^2_n	chisq	df	n
$F(m,n)$	f	df1, df2	m, n

s. S. 28 Skript von S. Ahlers

Aufruf von Verteilungen in R

Beispiel: Normalverteilung

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

Beispiel

- `rnorm(100, mean=0, sd=2)`
- `curve(dgamma(x, shape = 5), from = 0, to = 20, n = 200)`
- `quant <- c(0.1, 0.25, 0.5, 0.75, 0.99)`
- `qexp(quant, 4)`

Absolute und relative Häufigkeiten

Sei $x_1, ..., x_n$ eine Liste von Beobachtungen des merkmals X und $a_1, ..., a_k$, $k \leq n$ deren Ausprägungen. Im Fall ordinalskalierter Merkmale seien die a_i aufsteigend sortiert. Dann ist:

$h_j := \sum_{i=1}^{n} \mathbf{1}_{\{x_i = a_j\}}$	absolute Häufigkeit von a_j
$f_j := \frac{h_j}{n}$	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*.
- Das Aufführen von absoluten/relativen Häufigkeiten ist nur sinnvoll, falls k deutlich kleiner ist als n.
- Bei (quasi-)stetigen Merkmalen ist es sinnvoll Beobachtungsliste in Gruppen zu unterteilen.

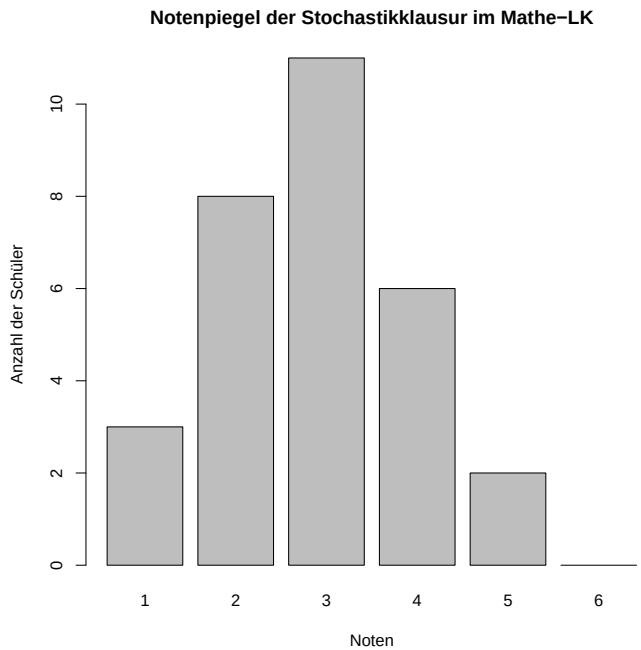
Beispiel: Notenspiegel

30 Studenten haben eine Statistikklausur mitgeschrieben. Student i hat die Punktzahl $x_i \in [0, 80]$ erreicht.

Unterteilung der Punkteskala: $[0, 16]$ mangelhaft, $[17, 32]$ ausreichend, ..., $[64, 80]$ sehr gut

Häufigkeitstabelle: Notenspiegel

sehr gut	3
gut	8
befriedigend	11
ausreichend	6
mangelhaft	2



Kumulierte Häufigkeiten

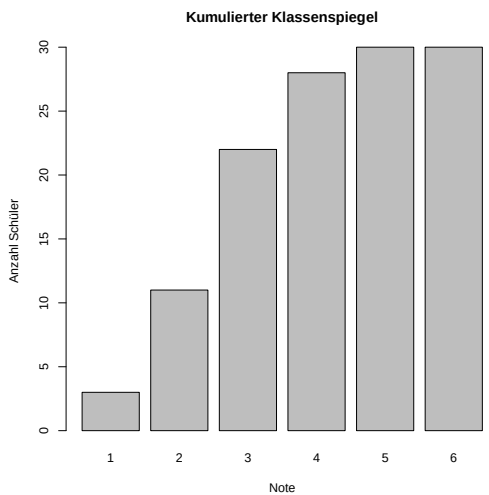
Addiert man die Häufigkeiten sukzessive auf, so spricht man von *kumulierten Häufigkeiten*.

$h_1, \sum_{j=1}^{2} h_j, ..., \sum_{j=1}^{k} h_j$	absolute kumulierte Häufigkeiten
$f_1, \sum_{j=1}^{2} f_j, ..., \sum_{j=1}^{k} f_j$	relative kumulierte Häufigkeiten

- Die Funktion `cumsum(x)` addiert sukzessive die Werte des Vektors `x` auf.

Kumulierte Häufigkeiten

sehr gut	3
gut	11
befriedigend	22
ausreichend	28
mangelhaft	30



Kumulierte Häufigkeitsverteilung

Definition

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

$$H(x) = \sum_{i=1}^{n} \mathbf{1}_{(-\infty,x]}(x_i) = \text{Anzahl der Werte } x_i \text{ mit } x_i \leq x$$

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

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

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.

Die empirische Verteilungsfunktion

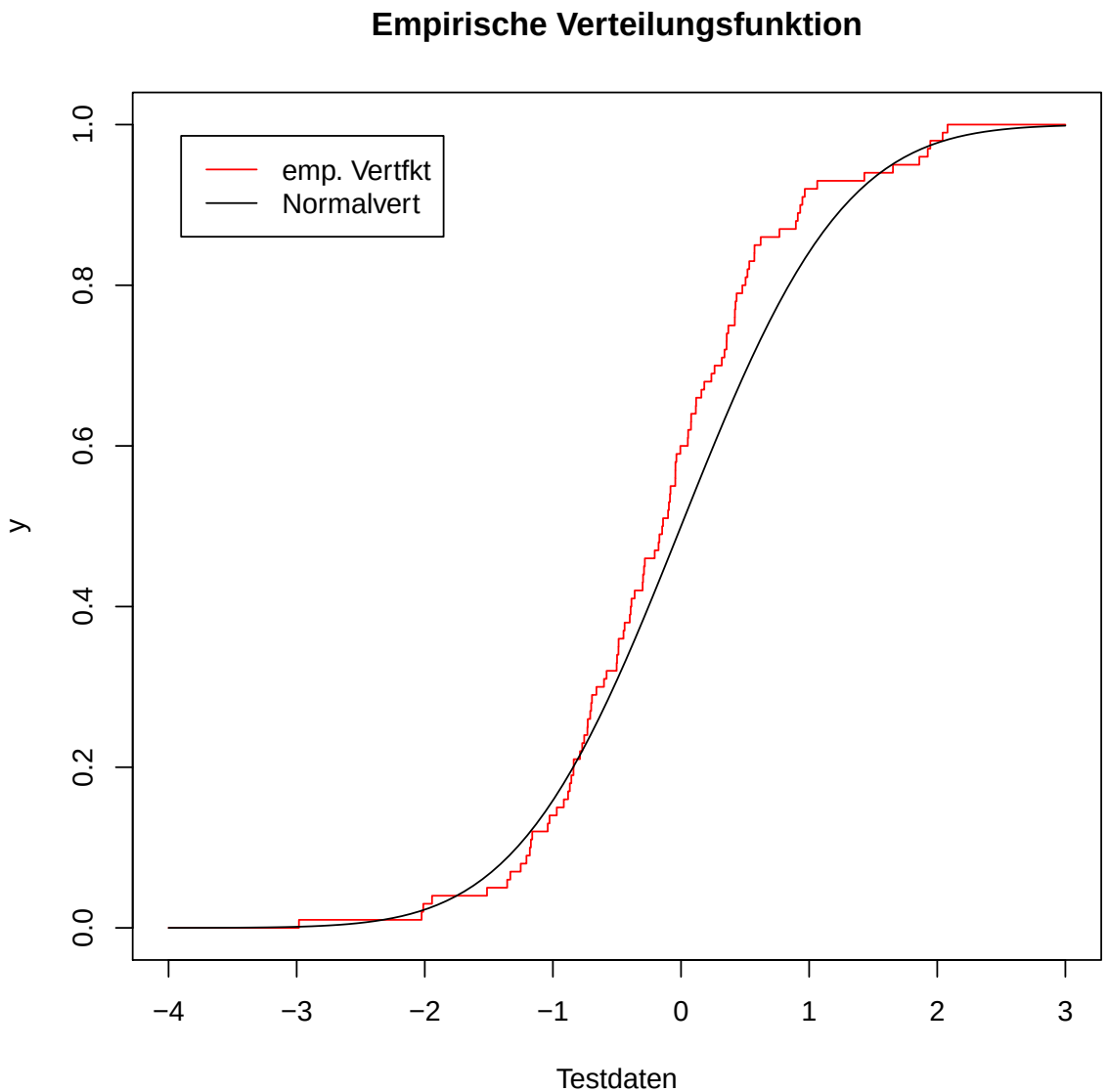
Definition

Die empirische Verteilungsfunktion F ist definiert durch

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

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.


Empirische Verteilungsfunktion
emp. Vertfkt
Normalvert
y
0.0
0.2
0.4
0.6
0.8
1.0
−4
−3
−2
−1
0
1
2
3
Testdaten

Der Abschnitt 6 (Funktionsverläufe) des Aufgabenblattes kann jetzt bearbeitet werden.

Gliederung

Uni- und multivariate Analyse

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

  Darstellungsarten:
  - Tabellen (Häufigkeitstabellen,...)
  - Kenngrößen zur Lage und Streuung (Mittelwert, Median, ...)
  - Diagramme (Histogramm, Tortendiagramm, ...)
  - Boxplot, Quantil-Plot, ...

- *Multivariate Analyse* betrifft die Auswertung der Erhebung *mehrerer* Merkmale
  - *Fragestellung*: Wie stark ist der Zusammenhang zwischen der eingesetzten Menge eines Düngemittels und der Erntemenge?
  - *Beispiel:* Lineare Modelle, ...

Daten- / Häufigkeitstabellen

- *Fahrgastbefragung:* Aus welchem Grund fahren Sie heute mit dem Bus?

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

Häufigkeitstabelle

- Antworten von 1000 befragten Fahrgästen

	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

Grafische 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 das Stabdiagramm nur mit Säulen statt Strichen	`barplot`
Balken-	wie Säulendiagramm, jedoch mit vertauschten Achsen	`barplot(..,horiz=TRUE)`
Torten-	die Flächen der Kreissektoren sind proportional zu den Häufigkeiten: $f_j \cdot 360°$	`pie`

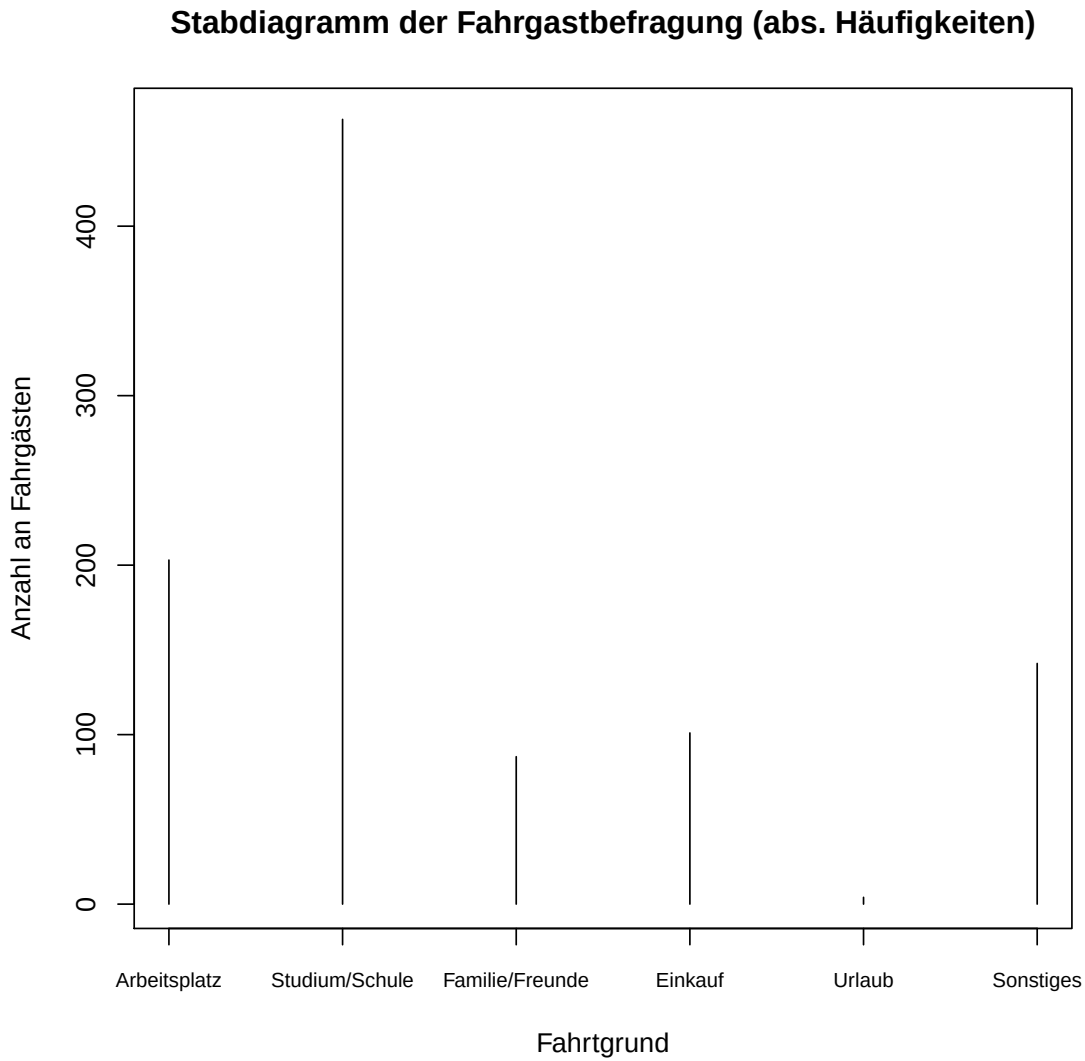
Stabdiagramm

Beispiel

- `x <- c(203, 463, 87, 101, 4, 142)`
- `names(x) <- c("Fahrt zum Arbeitsplatz", ...)`
- `plot(x, type='h', xaxt='n')`
- `axis(1, at=1:length(x), labels=names(x), cex.axis=0.75)`

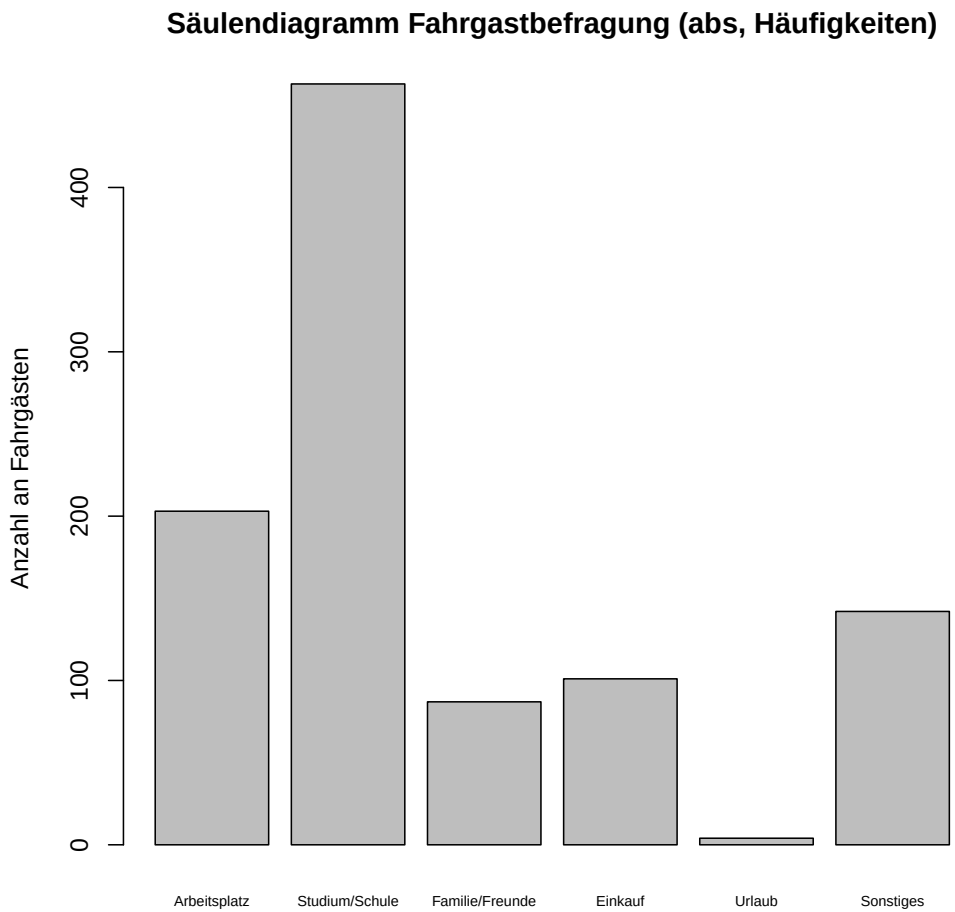
Die Funktion `axis(n, at=.., labels=..)` fügt an der Grafikseite `n` eine Achse hinzu mit Markierungen an den Stellen `at` und Beschriftung `labels`.

Stabdiagramm



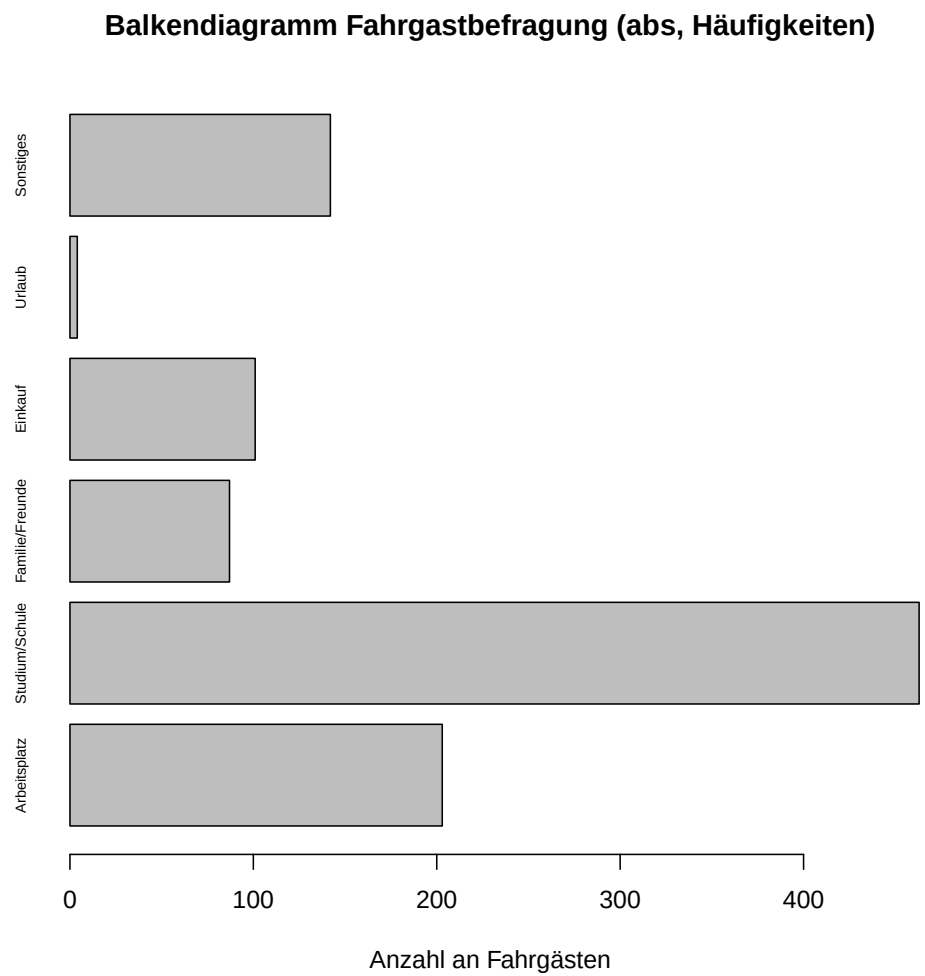
Säulendiagramm

- `barplot(x, cex.names=0.6)`



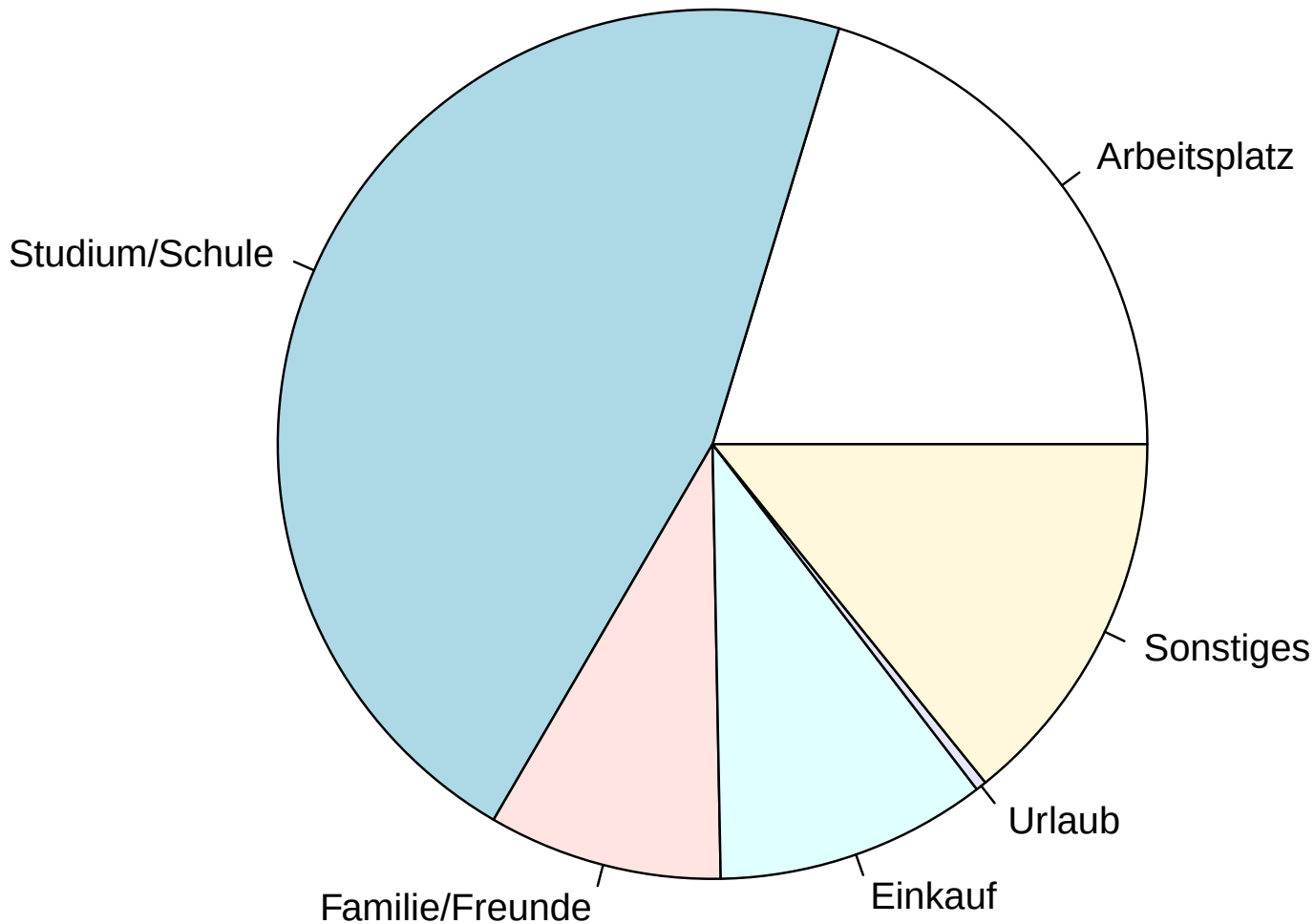
Balkendiagramm

- `barplot(x, horiz=TRUE, cex.names=0.6)`



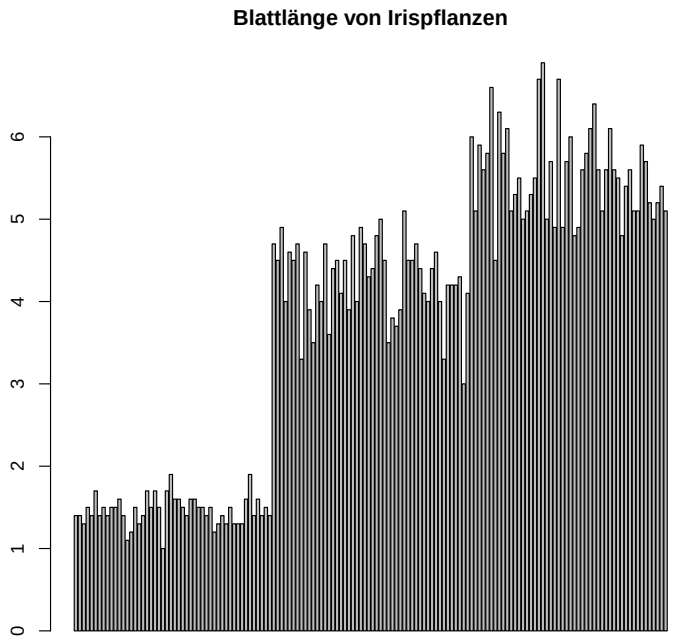
Tortendiagramm

- `pie(x)`



Klasseneinteilung

Ist die Anzahl der Beobachtungen eines Merkmals sehr groß (z.B. bei stetigen/quantitativen Merkmalen), so ist die Darstellung dieses Merkmales in Häufigkeitstabellen oder durch die vorher genannten Diagramme nicht sinnvoll, da diese unübersichtlich werden. (▶ vgl. abs./rel. Häufigkeiten)



⇝ Einteilung der Beobachtungen in Klassen

Der Befehl cut

- In R benutzt man dafür `cut(x, breaks=..)`. Dieser ersetzt die Werte eines Vektors durch die Klasse, innerhalb derer er liegt.
- `breaks` gibt dabei entweder die Bruchpunkte der Klassen oder die Anzahl an Klassen (alle gleiche Länge) an.
- Das Merkmal muss mindestens ordinal skalliert sein.

Beispiel (Blattlänge der Pflanzengattung Iris)

```
bl <- iris$Petal.Length
bl.kl.5 <- cut(bl, breaks=5, include.lowest=TRUE)
bl.kl.ind <- cut(bl, breaks=c(1,2,5,7),
include.lowest=TRUE)
bl.kl.5
```

Histogramme

Ein **Histogramm** ist ein Balkendiagramm, welches die absoluten/relativen Häufigkeiten von Beobachtungen in bestimmten Intervallen angibt.

- Teilt die Merkmalsausprägungen in $k \in \mathbb{N}_{\geq 2}$ Intervalle $[c_0, c_1), \ldots, [c_{k-1}, c_k)$ ein
- Zeichnet 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

- h_j und f_j sind dabei die absolute bzw. relative Zahl der Beobachtungen in $[c_{j-1}, c_j)$.
- Problem, falls die Daten über ein sehr großes Intervall gestreut und nicht beschränkt sind. Dann können die Säulen die Höhe 0 haben.

Der `hist`-Befehl

Histogramme erzeugt man in `R` mit `hist(x, breaks=..)`.

- `x` ist Datenvektor (mindestens ordinalskaliert)
- `breaks` gibt entweder die Bruchpunkte der Klassen oder die Anzahl an Klassen (alle gleiche Länge) an. Es gibt folgende Optionen:

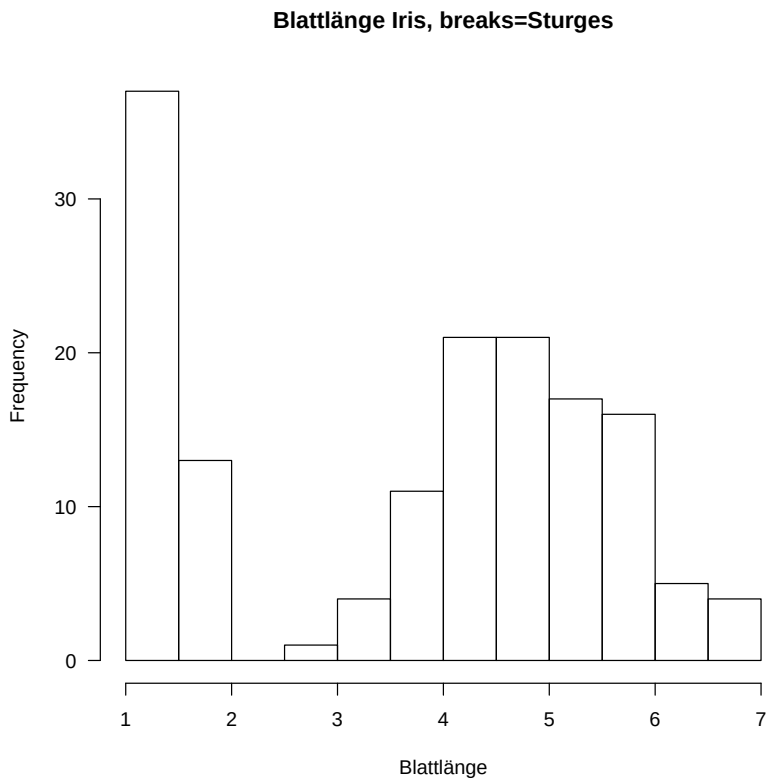
`c(c0,...,ck)`	Intervalle $[c_0, c_1), \ldots, [c_{k-1}, c_k)$.
20	$k = 20$ Intervalle gleicher Länge
`"Sturges"`	(default) $k \approx \log_2(n) + 1$ Intervalle gleicher Länge
`"Scott"`	wie oben, jedoch mit $k \approx n^{1/3}$

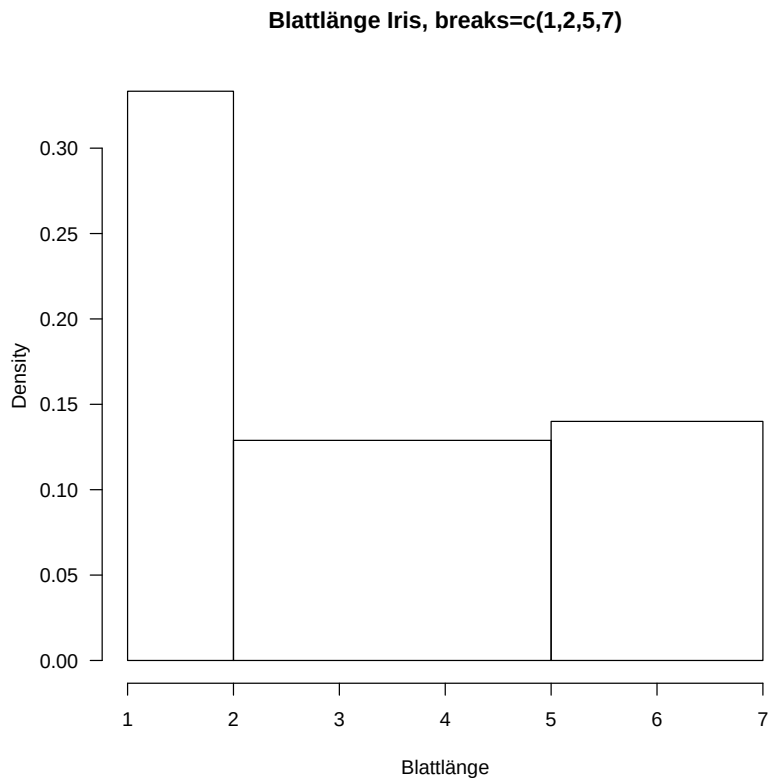
- `freq` gibt an, ob absolute (`TRUE`) oder relative (`FALSE`) Häufigkeiten angezeigt werden sollen.

Beispiel: Blattlänge der Irispflanze

Beispiel

- hist(bl)
- hist(bl, breaks=c(1,2,5,7), include.lowest=TRUE)





Beschreibung von Verteilungen

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

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

Unimodale und multimodale Verteilungen

Viele (empirische) Verteilungen weisen einen oder mehrere Gipfel in deren Dichte auf. Man nennt eine solche Verteilung:

- *unimodel*, falls die Verteilung nur einen Gipfel hat und zu den Randbereichen abfällt ohne dass ein zweiter Gipfel auftritt. (Beispiel: Normalverteilung)
- *bimodal*, falls ein zweiter (und kein weiterer) Gipfel aufritt. (Beispiel: Blattlänge der Irispflanze)
- *multimodal*, falls weitere Nebengipfel auf auftreten.

Symmetrie

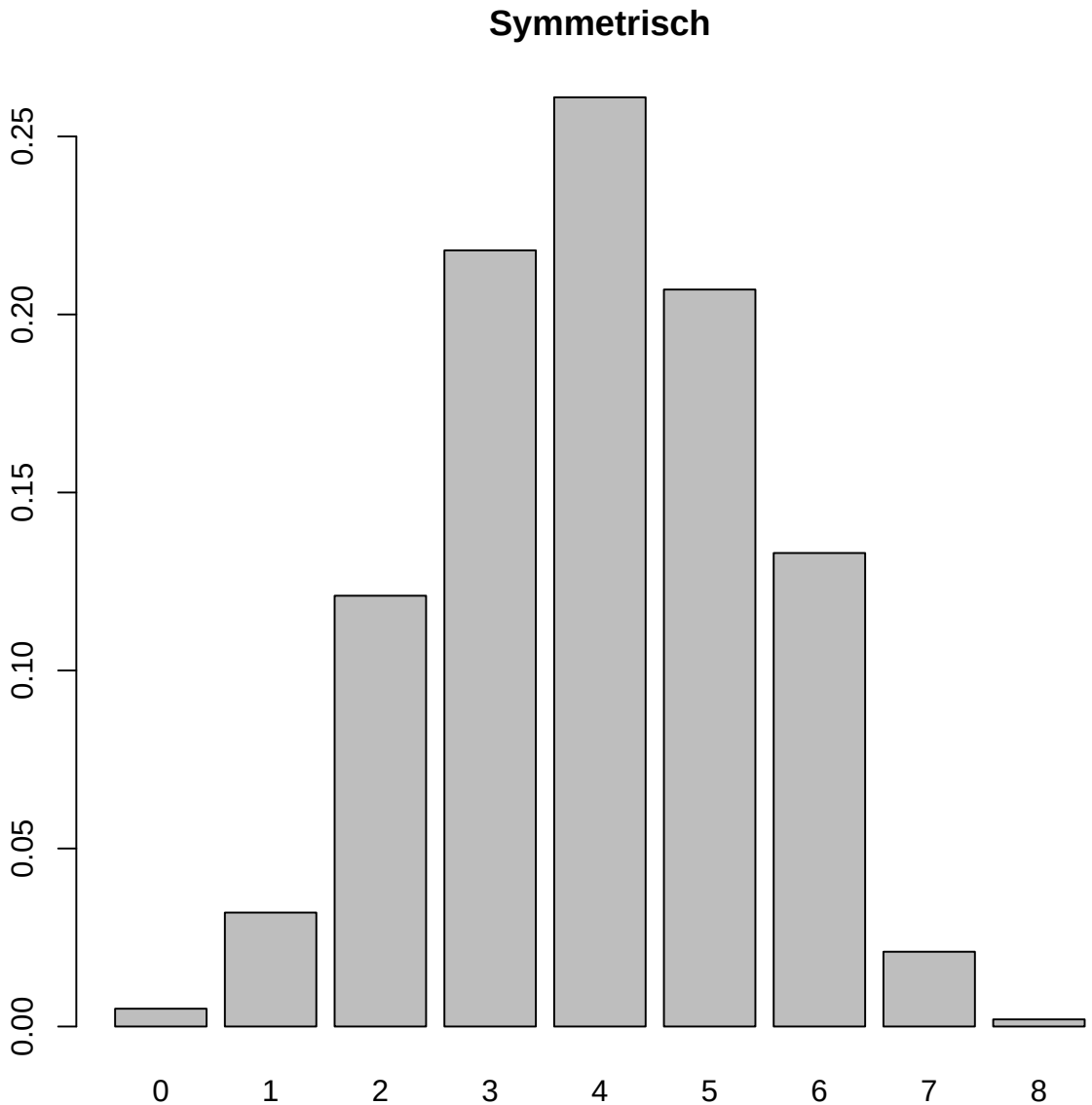
Definition

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

(Beispiel: Normalverteilung, Binomialverteilung, ...)

Bemerkung

Exakte Symmetrie ist bei empirischen Verteilungen selten gegeben.

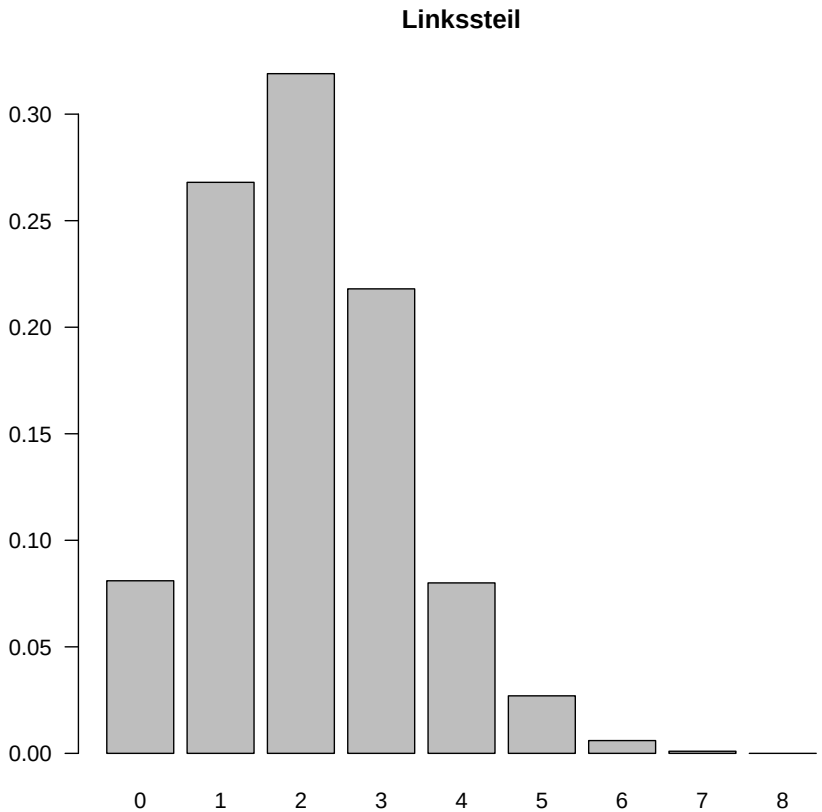

Symmetrisch
0.00
0.05
0.10
0.15
0.20
0.25
0
1
2
3
4
5
6
7
8

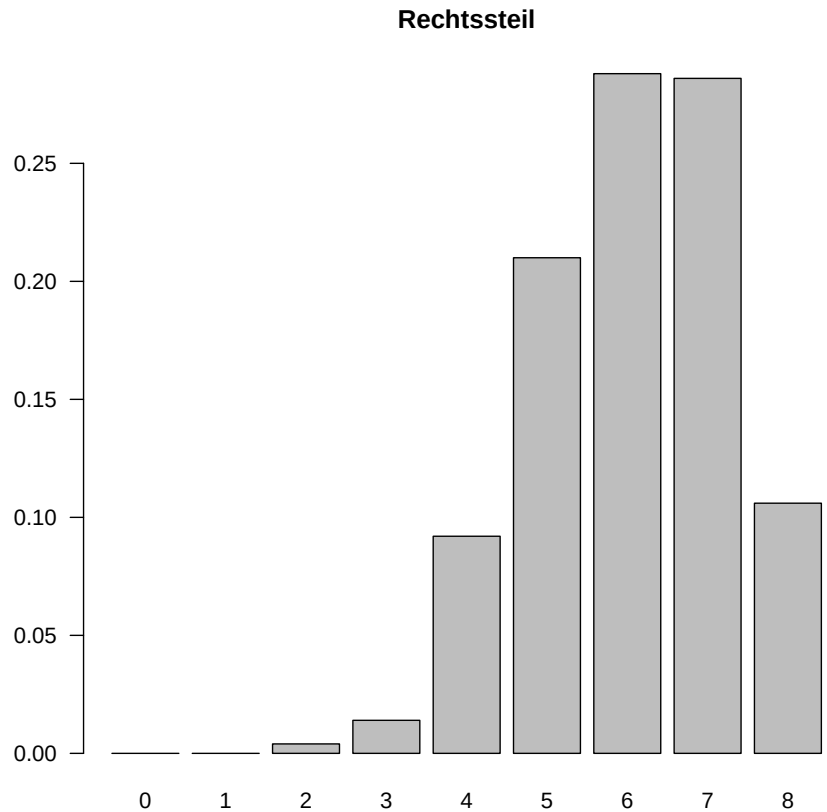
Schiefe

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.


Linkssteil
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
1
2
3
4
5
6
7
8


Rechtssteil
0.25
0.20
0.15
0.10
0.05
0.00
0
1
2
3
4
5
6
7
8

Das arithmetische Mittel

Definition

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

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

berechnet.

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

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

- 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 metrische Daten sinnvoll.
- Das arithmetische Mittel besitzt die *Schwerpunkteigenschaft*

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

- $\bar{x}$ minimiert den quadratischen Abstand, d.h.

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 \;=\; \min_{z\in\mathbb{R}} \sum_{i=1}^{n}(x_i - z)^2$$

- Das arithmetische Mittel reagiert *empfindlich* auf extreme Werte und *Ausreißer*.
- 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 $x_{(1)}, \ldots, x_{(n)}$.

Definition

Der *Median* x_{med} von $x_{(1)} \leq \ldots \leq x_{(n)}$ ist durch

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

definiert.

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

Eigenschaften des (Stichproben-)Medians

- Der Median ist ab dem Ordinalskalenniveau sinnvoll.
- Der Median x_{med} ist robust gegenüber Ausreißern
- Mindestens 50% der Daten sind $\geq x_{\text{med}}$ und mindestens 50% der Daten sind $\leq x_{\text{med}}$.
- Statistisch aussagekräftiger als das arithmetische Mittel
- x_{med} minimiert die absolute Abweichung, d.h.

$$\sum_{i=1}^{n} |x_i - x_{\text{med}}| \;=\; \min_{z \in \mathbb{R}} \sum_{i=1}^{n} |x_i - z|$$

Der Modus

Ein weiteres gebräuchliches Lagemaß ist der Modus.

Definition

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

Eigenschaften des Modus:

- Der Modus ist eindeutig, falls die Häufigkeitsverteilung ein eindeutiges Maximum besitzt.
- Der Modus ist bereits auf Nominalskalenniveau sinnvoll.
- Der Modus ist robust.
- Der Modus ist eine Ausprägung des Merkmals.

Lageregeln

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

- Bei unimodalen Verteilungen gilt sogar $\bar{x} \approx x_{\text{med}} \approx x_{\text{mod}}$

Gruppierte Lagemaße

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

Modus: Bestimme Modalklasse $[c_{i-1}, c_i)$ (Klasse mit der größten Beobachtungszahl) und verwende Klassenmitte $x_{\text{mod.grupp}} = m_i$ als Modus

Median: Bestimme Einfallsklasse $[c_{i-1}, c_i)$ des Medians und daraus $x_{\text{med, grupp}} = c_{i-1} + d_i(0.5 - \sum_{j \le i-1} f_j)/f_i$.

Arithm. Mittel: $\bar{x}_{\text{grupp}} = \sum_{i=1}^{k} f_i m_i$.

$d_i = c_i - c_{i-1}$ Klassenbreite, f_i relative Häufigkeit der Klasse i, $m_i = c_{i-1} + d_i/2$ Klassenmitte.

- Der wahre Modus muss nicht 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	$\frac{1}{n}\sum_{i=1}^{n}\lvert x_i - \overline{x}\rvert$
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	$R(x) = \max(x) - \min(x)$
Interquartilsabstand	$IQR(x) = x_{0.75} - x_{0.25}$

- Stichprobenvarianz eines Vektors `x` wird in `R` mit dem Befehl `var(x)` berechnet.
- Die Spannweite unter anderem mit `diff(range(x))`

Quantile

Definition

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

Bemerkung

- Für ein p-Quantil gilt

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

- $x_{0.5}$ ist der Median.
- $x_{0.25}$ und $x_{0.75}$ heißen *unteres* bzw. *oberes Quartil*.
- In `R` werden Quantile mit dem `quantile`-Befehl aufgerufen.

Quantilfunktion 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*".

Fünf-Punkte-Zusammenfassung, `summary`

Definition

Die *Fünf-Punkte-Zusammenfassung* besteht aus

$$x_{\min},\ x_{0.25},\ x_{\mathrm{med}},\ x_{0.75},\ x_{\max}$$

des Datensatzes.

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

Beispiel

- `summary(bl)`
- `summary(iris)`

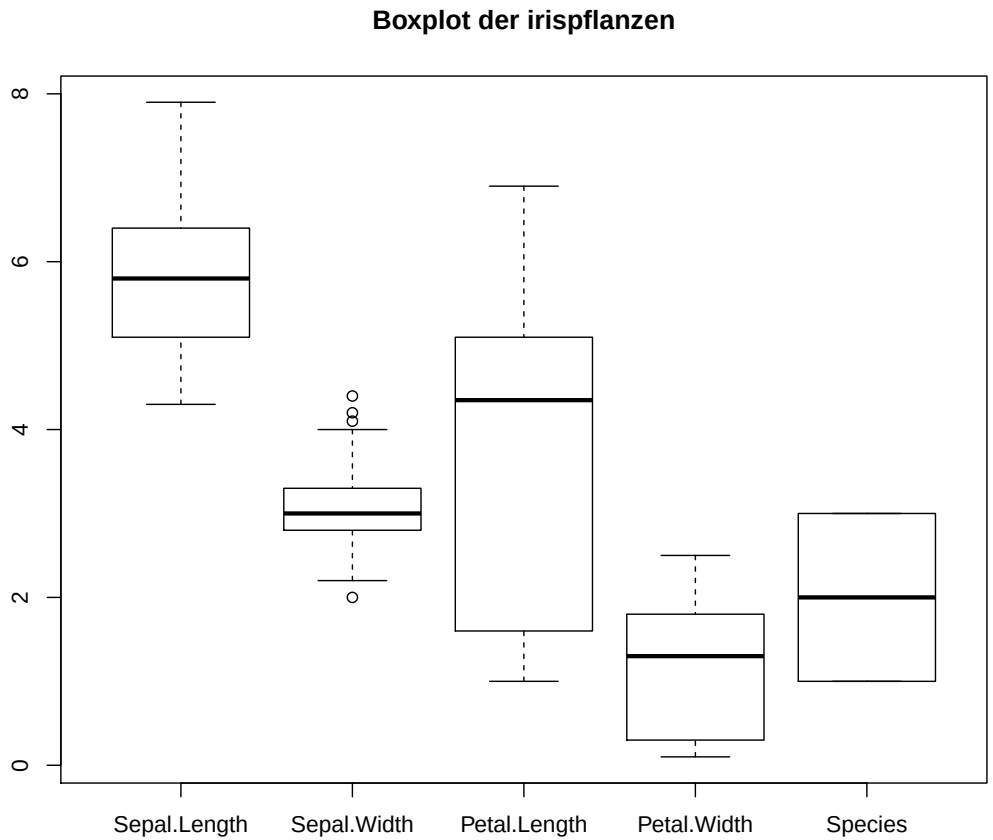
Boxplots

Die Fünf-Punkte-Zusammenfassung eines Datensatzes $x_1, \ldots, x_n$ wird in einem Boxplot visualisiert. Es wird dabei in ein Koordinatensystem gezeichnet:

- Ein Rechteck (eine Box), welches durch das obere Quartil und untere Quartil begrenzt ist.
- Eine Linie auf der Höhe des Medians durch die Box.
- Linien (*Whiskers*) ausgehend von der Box bis $\min\{x_{0.75} + 1.5IQR, x_{\max}\}$ bzw. bis $\max\{x_{0.25} - 1.5IQR, x_{\min}\}$, wo die Linien durch senkrechte Linien begrenzt werden.
- Einzelnen Punkte für Werte jenseits der Whiskers (*Extremwerte*).

Beispiel: Boxplot

- `boxplot(bl, horizontal=TRUE)`
- `boxplot(iris)`



NQ-Plots: Idee

Häufig wird bei Daten angenommen, dass diese normalverteilt sind, da sie dann häufig statistisch einfacher zu behandeln sind.

Frage: 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 k-te kleinste Beobachtung auf der y-Achse gegen die erwartete k-te-kleinste Beobachtung eines Vektors mit n standardnormalverteilten Zufallsgrößen ab.

Ziel: 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 $(i - 0.5)/n$-Quantile $z_{(i)}$ der $N(0,1)$-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 $x = \mu + \sigma z$.
- Einen NQ-Plot erhält man in `R` mit dem Befehl `qqnorm`.

QQ-Plots

Um Daten mit einer beliebigen Verteilung oder einem anderen datensatz visuell zu vergleichen, kann man den *Quantile-Quantile-Plot* (QQ-Plot) benutzen.

- Sind die Verteilungen gleich, so entsteht eine Gerade (wie beim NQ-Plot) .
- Einen QQ-Plot erhält man in `R` mit dem Befehl `qqplot(x,y)`, wobei `x` und `y` Vektoren der zu vergleichenden Daten sind.

Der Abschnitt 7 (Darstellung univariater Daten) des Aufgabenblattes kann jetzt bearbeitet werden.