

Blockpraktikum zur Statistik mit R

3. März 2015

Johannes Blank

Gliederung

- 1 Überblick über die Statistik
 - Ziele in der Statistik und der deskriptiven Statistik
- 2 Datenerhebung
 - Grundlegende Begriffe
 - Merkmalstypen
- 3 Funktionsverläufe skizzieren
 - Funktionen plotten
 - Verteilungen und Häufigkeitsverteilungen
- 4 Datendarstellungen in der univariaten Analyse
 - Aufbereitung und grafische Darstellung
 - Beschreibung von Verteilungen (Kenngrößen)
 - Quantile, Boxplots und Normal-Quantil-Plots

Gliederung

- 1 Überblick über die Statistik
 - Ziele in der Statistik und der deskriptiven Statistik
- 2 Datenerhebung
 - Grundlegende Begriffe
 - Merkmalstypen
- 3 Funktionsverläufe skizzieren
 - Funktionen plotten
 - Verteilungen und Häufigkeitsverteilungen
- 4 Datendarstellungen in der univariaten Analyse
 - Aufbereitung und grafische Darstellung
 - Beschreibung von Verteilungen (Kenngrößen)
 - Quantile, Boxplots und Normal-Quantil-Plots

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:

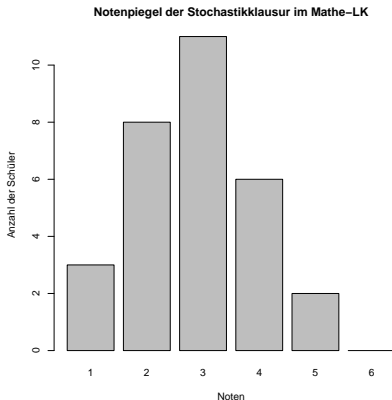
- ▶ Daten sammeln
- ▶ Daten darstellen und analysieren
- ▶ Daten interpretieren
- ▶ Prognosen und Entscheidungen treffen

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

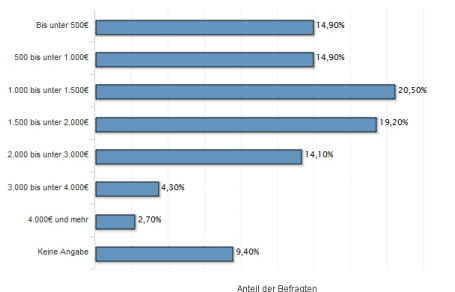
Durchschnittsnote: 2.866



Beispiel: Einkommensverteilung

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

Wie hoch war Ihr Einkommen (Nettoverdienst), d. h. Lohn oder Gehalt nach Abzug von Steuern und Sozialversicherungsbeiträgen, im letzten Monat?



1 Deutschland; ab 16 Jahre; Erwerbstätige; 11.242 Befragte; TNS Infratest Sozialforschung; 2008

Angabe zur Quelle: SOEP

© Statista 2011

Beispiel: Düngemittel

- ▶ *Fragestellung*: Wie stark ist der Zusammenhang zwischen der eingesetzten Menge eines Düngemittels und der Erntemenge?
- ▶ *Ziel*:
- ▶ **Quantifizierung des Einflusses** gewisser Merkmale.

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 Strukturen und Besonderheiten in Daten

- ▶ Ableiten von Fragestellungen und Hypothesen
- ▶ keine Stochastik

induktiv: Über Daten hinaus Aussagen über Grundgesamtheit treffen

- ▶ Testen (Ist Therapie A besser als Therapie B?)
- ▶ Schätzen (Wie lange hält eine Glühbirne?)
- ▶ braucht Stochastik, Wahrscheinlichkeitstheorie

Gliederung

- 1 Überblick über die Statistik
 - Ziele in der Statistik und der deskriptiven Statistik
- 2 **Datenerhebung**
 - Grundlegende Begriffe
 - Merkmalstypen
- 3 Funktionsverläufe skizzieren
 - Funktionen plotten
 - Verteilungen und Häufigkeitsverteilungen
- 4 Datendarstellungen in der univariaten Analyse
 - Aufbereitung und grafische Darstellung
 - Beschreibung von Verteilungen (Kenngrößen)
 - Quantile, Boxplots und Normal-Quantil-Plots

Grundlegende Begriffe

<i>Statistische Einheiten:</i>	auf bestimmte Merkmale untersuchte Personen bzw. Objekte
<i>Grundgesamtheit:</i>	Menge aller für die Fragestellung relevanten statistischen Einheiten
<i>Teilgesamtheit:</i>	Teilmenge der Grundgesamtheit
<i>Stichprobe:</i>	tatsächlich untersuchte Teilgesamtheit
<i>Merkmal (Variable):</i>	interessierende Größe
<i>Merkmalsausprägung:</i>	konkreter Wert des Merkmals für eine bestimmte statistische Einheit

Beispiel: Mietspiegel

- ▶ Städte und Gemeinden erstellen Mietspiegel
- ▶ bieten Mietern und Vermietern eine Marktübersicht zu Miethöhen
- ▶ 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

<i>Statistische Einheiten:</i>	Wohnungen in Münster
<i>Grundgesamtheit:</i>	Menge aller Wohnungen in Münster
<i>Stichprobe:</i>	Wohnungen, deren Daten erfasst wurden
<i>Merkmale:</i>	Baujahr, Größe, Preis/qm
<i>Merkmalsausprägungen:</i>	<ul style="list-style-type: none">▶ <i>Baujahr:</i> „bis 1929“, ..., „2004-2005“▶ <i>Größe:</i> „21-30 qm“, ..., „151-160qm“▶ <i>Preis/qm:</i> $x \in (0, \infty)$

Stichprobenarten

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

- ▶ Vollerhebung nicht immer möglich \leadsto Stichprobe
- ▶ Stichprobenarten: *einfache Zufallsstichprobe*, *bewußtes Auswahlverfahren* etc..

Einfache Zufallsstichprobe:

- ▶ zufälliges Ziehen aus der Grundgesamtheit
- ▶ technisch häufig schwer umsetzbar
- ▶ Ziehungsmethode kann systematischen Fehler enthalten

Bewußtes Auswahlverfahren:

- ▶ Stichprobe wird vom Interviewer ausgewählt
- ▶ 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 \leadsto Stichprobe
- ▶ Stichprobenarten: *einfache Zufallsstichprobe*, *bewußtes Auswahlverfahren* etc..

Einfache Zufallsstichprobe:

- ▶ zufälliges Ziehen aus der Grundgesamtheit
- ▶ technisch häufig schwer umsetzbar
- ▶ Ziehungsmethode kann systematischen Fehler enthalten

Bewußtes Auswahlverfahren:

- ▶ Stichprobe wird vom Interviewer ausgewählt
- ▶ Vor- und Nachteil: Kontrolle durch den Interviewer

Verzernte Stichproben

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

Verzerrung (Bias)	Ursache und Beispiel
-------------------	----------------------

<i>Selektion-Bias</i>	bewusster Ausschluss von Elementen von der Ziehung Bsp: Internet- oder Zeitungsumfrage; Berkson's Fallacy
<i>Nonresponse-Bias</i>	(unangenehme) Fragen bleiben unbeantwortet Bsp: Fragen zum Sexualverhalten etc.
<i>Selfselection-Bias</i>	Umfragen auf freiwilliger Basis Nur Interessierte nehmen teil

Merkmale und Ausprägungen - Mietspiegel

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

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

Stetige und diskrete Merkmale

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

Beispiel (Mietspiegel)

- ▶ *diskret:* Baujahr, Badeausstattung
- ▶ *stetig:* Preis/qm
- ▶ *quasi-stetig:* Größe, Preis/qm (wird gerundet auf zwei Nachkommastellen)

Skalenarten

<i>nominalskaliert:</i>	Ausprägungen sind Namen, keine Ordnung möglich (=, \neq)
<i>ordinalskaliert:</i>	Ausprägungen können geordnet, aber Abstände nicht interpretiert werden (=, \neq , $<$, $>$)
<i>intervallskaliert:</i>	Ausprägungen sind Zahlen, Interpretation der Abstände möglich (=, \neq , $<$, $>$, $+$, $-$)
<i>verhältnisskaliert:</i>	Quotienten der Ausprägungen können sinnvoll interpretiert werden (=, \neq , $<$, $>$, $+$, $-$, \div , \cdot)
<i>kardinalskaliert:</i>	Oberbegriff für intervallskaliert und verhältnisskaliert; auch: Metrische Skala

Beispiel

- ▶ *nominalskaliert*: Badeausstattung (Mietspiegel), Geschlecht
- ▶ *ordinalskaliert*: Baujahr (Mietspiegel), Zufriedenheitsskala
- ▶ *intervallskaliert*: Temperatur in Celsius
- ▶ *verhältnisskaliert*: Preis/qm (Mietspiegel), Alter der Teilnehmer an einer Umfrage

Qualitative und quantitative Merkmale

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

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

Gliederung

- 1 Überblick über die Statistik
 - Ziele in der Statistik und der deskriptiven Statistik
- 2 Datenerhebung
 - Grundlegende Begriffe
 - Merkmalstypen
- 3 Funktionsverläufe skizzieren
 - Funktionen plotten
 - Verteilungen und Häufigkeitsverteilungen
- 4 Datendarstellungen in der univariaten Analyse
 - Aufbereitung und grafische Darstellung
 - Beschreibung von Verteilungen (Kenngrößen)
 - Quantile, Boxplots und Normal-Quantil-Plots

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 weitere Punkte und mit `lines(x,y)` Linien in eine bestehende Grafik einfügen.
- ▶ Weitere nützliche Parameter sind unter anderem: `type`, `pch`, `lty`, `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')`

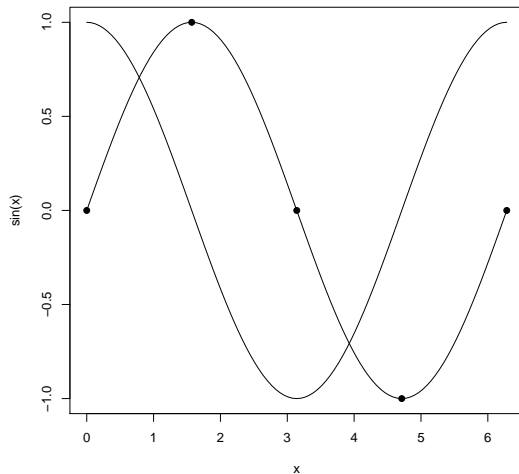
Kurven plotten

- ▶ 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 mehrere Funktionen 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



Hilfreiches

- ▶ Mit `par(mfrow=c(nrow,ncol))` oder `par(mfcol=c(nrow,ncol))` kann man `nrow` x `ncol` Grafiken in ein Fenster zeichnen. Die Grafiken werden dann entweder zeilenweise oder spaltenweise in das Fenster gefüllt.
- ▶ Der Befehl `dev.off()` schließt das Grafikfenster.
- ▶ Mit `pdf(file="name")` schreibt man die bis zum Befehl `dev.off()` folgenden Grafiken in eine PDF-Datei. Das Verzeichnis kann mit `setwd("Pfad")` gewählt werden, ansonsten findet sich die Datei auf dem Desktop wieder.

Verteilungen in R

<i>Verteilung</i>	<i>Name in R</i>	<i>Parameter in R</i>	<i>Parameter</i>
$B(n, p)$	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_n^2	chisq	df	n

s. S. 28 Skript von S. Ahlers

Aufruf von Verteilungen in R

Beispiel: Normalverteilung

- ▶ Dichtefunktion: `dnorm`
- ▶ Verteilungsfunktion: `pnorm`
- ▶ Quantilsfunktion: `qnorm` (z.B. `qnorm(1/2)=0`)
- ▶ 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)`
- ▶ Siehe auch `?stats`

Absolute und relative Häufigkeiten

Sei x_1, \dots, x_n eine Liste von Beobachtungen des Merkmals X und a_1, \dots, a_k , $k \leq n$ deren Ausprägungen.

$h_j := \sum_{i=1}^n \mathbf{1}_{\{x_i=a_j\}}$	absolute Häufigkeit von a_j
$f_j := \frac{h_j}{n}$	relative Häufigkeit von a_j
h_1, \dots, h_k	absolute Häufigkeitsverteilung
f_1, \dots, f_k	relative Häufigkeitsverteilung

- ▶ Das Aufführen von absoluten/relativen Häufigkeiten ist nur sinnvoll, falls k deutlich kleiner ist als n .
- ▶ Oftmals ist es sinnvoll (z.B. bei (quasi-)stetigen Merkmalen) die Beobachtungsliste in Gruppen zu unterteilen.

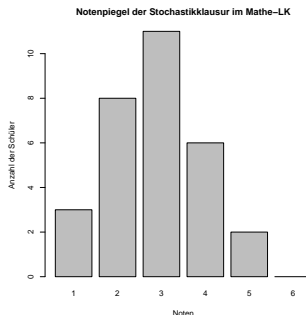
Beispiel: Notenspiegel

30 Studenten haben eine Statistiklausur 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

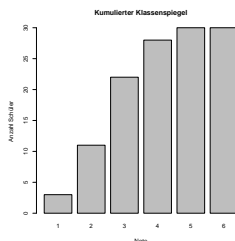
Addiert man die Häufigkeiten sukzessive auf, so spricht man von *kumulierten Häufigkeiten*.

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

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

Satz (von Glivenko und Cantelli)

Seien X_1, X_2, \dots eine Folge u. i. v. Zufallsgrößen mit Werten in \mathbb{R} .

$F_n(\cdot) = F_n(\cdot, x_1, \dots, x_n)$ sei die empirische Verteilungsfunktion von x_1, \dots, x_n .

Dann konvergiert $F_n(\cdot, X_1, \dots, X_n)$ für $n \rightarrow \infty$ 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

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

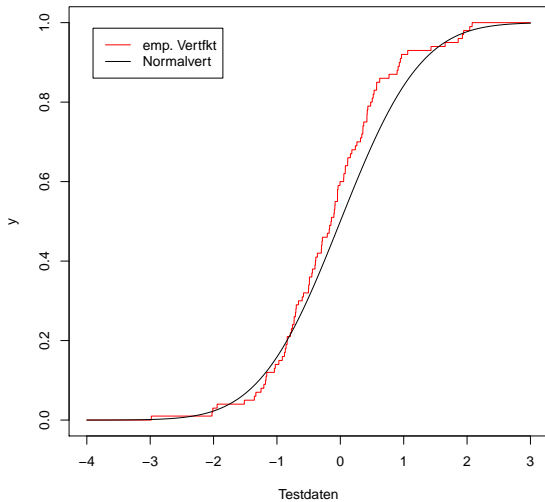
Satz (von Glivenko und Cantelli)

Seien X_1, X_2, \dots eine Folge u. i. v. Zufallsgrößen mit Werten in \mathbb{R} .

$F_n(\cdot) = F_n(\cdot, x_1, \dots, x_n)$ sei die empirische Verteilungsfunktion von x_1, \dots, x_n .

Dann konvergiert $F_n(\cdot, X_1, \dots, X_n)$ für $n \rightarrow \infty$ P -f. s. gleichmäßig in $x \in \mathbb{R}$ gegen die Verteilungsfunktion F von X_1 .

Empirische Verteilungsfunktion



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

Gliederung

- 1 Überblick über die Statistik
 - Ziele in der Statistik und der deskriptiven Statistik
- 2 Datenerhebung
 - Grundlegende Begriffe
 - Merkmalstypen
- 3 Funktionsverläufe skizzieren
 - Funktionen plotten
 - Verteilungen und Häufigkeitsverteilungen
- 4 **Datendarstellungen in der univariaten Analyse**
 - **Aufbereitung und grafische Darstellung**
 - **Beschreibung von Verteilungen (Kenngrößen)**
 - **Quantile, Boxplots und Normal-Quantil-Plots**

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
<i>Stab-</i>	a_1, \dots, a_k werden auf der x -Achse abgetragen, orthogonal zur x -Achse wird über a_j ein Strich proportional zu h_j abgetragen	<code>plot(..., type="h")</code>
<i>Säulen-</i>	wie das Stabdiagramm nur mit Säulen statt Strichen	<code>barplot</code>
<i>Balken-</i>	wie Säulendiagramm, jedoch mit vertauschten Achsen	<code>barplot(..., horiz=TRUE)</code>
<i>Torten-</i>	die Flächen der Kreissektoren sind proportional zu den Häufigkeiten: $f_j \cdot 360^\circ$	<code>pie</code>

Stabdiagramm

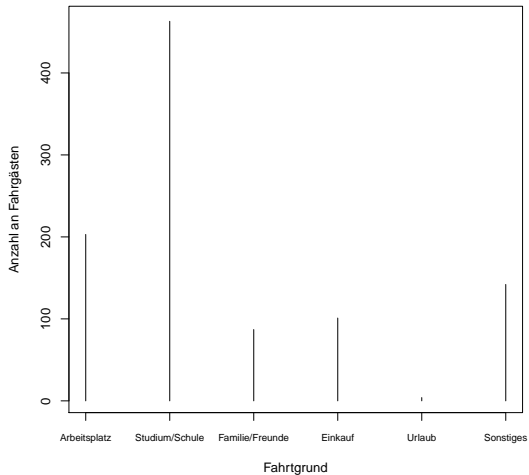
Beispiel

- ▶ `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` (s. `?axis` für genauere Information) eine Achse hinzu mit Markierungen an den Stellen `at` und Beschriftung `labels`.

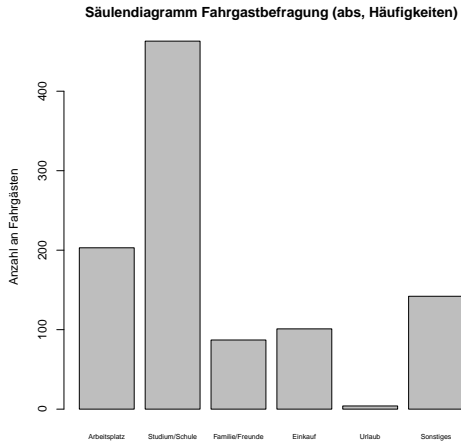
Stabdiagramm

Stabdiagramm der Fahrgastbefragung (abs. Häufigkeiten)



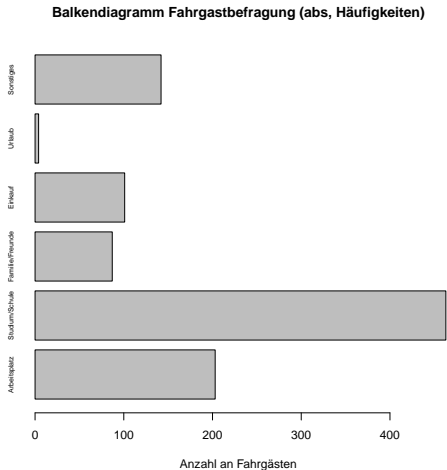
Säulendiagramm

► `barplot(x, cex.names=0.6)`



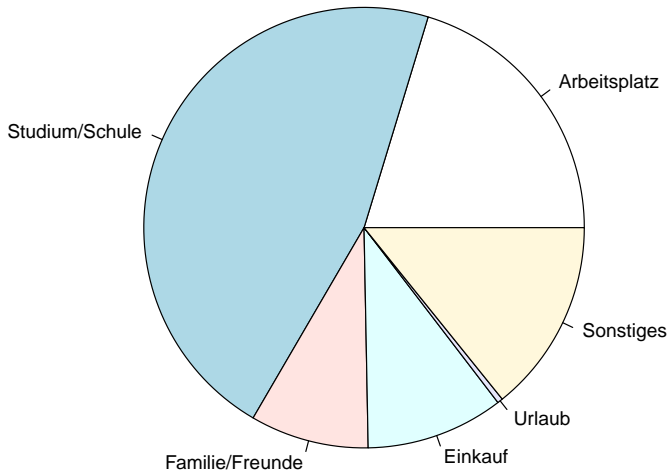
Balkendiagramm

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



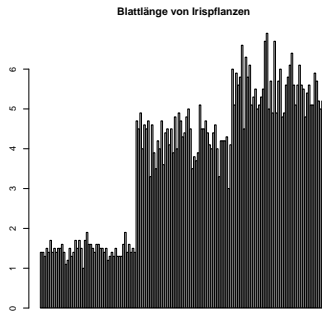
Tortendiagramm

► `pie(x)`



Klasseneinteilung

Ist die Anzahl der unterschiedlichen 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 skaliert 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 $[c_0, c_1), \dots, [c_{k-1}, c_k)$ ein
- ▶ Zeichnet über den Klassen $[c_0, c_1), \dots, [c_{k-1}, c_k)$ 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 ist 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:

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

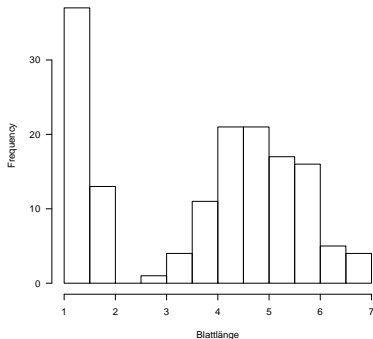
- ▶ `freq` gibt an, ob absolute (TRUE) oder relative (FALSE) Häufigkeiten angezeigt werden sollen.

Beispiel: Blattlänge der Irispflanze

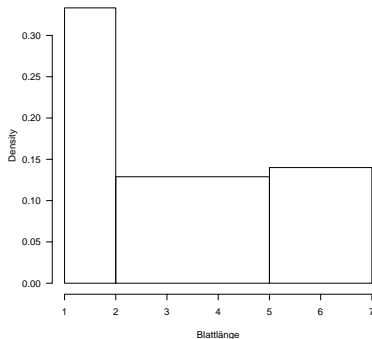
Beispiel

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

Blattlänge Iris, breaks=Sturges



Blattlänge Iris, breaks=c(1,2,5,7)



Beschreibung von Verteilungen

Bei der Datenanalyse 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:

- ▶ *unimodal*, 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 auftritt.
(Beispiel: Blattlänge der Irispflanze)
- ▶ *multimodal*, falls weitere Nebengipfel auftreten.

Symmetrie

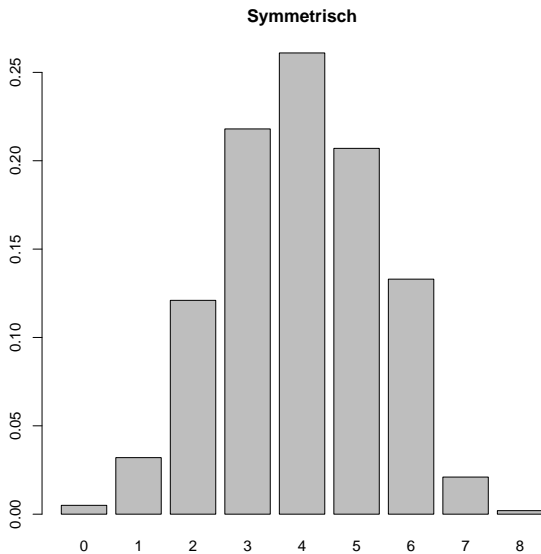
Definition

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, ...)

Bemerkung

Exakte Symmetrie ist bei empirischen Verteilungen selten gegeben.



Schiefe

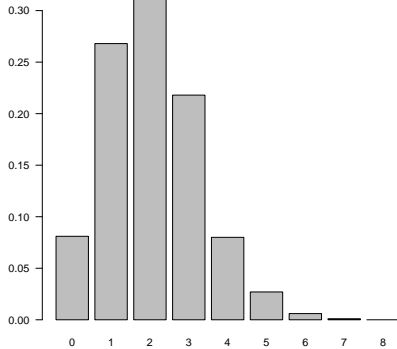
Definition

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

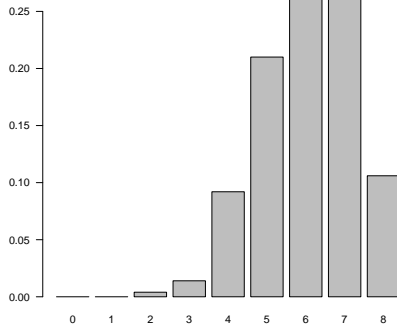
Analog heißt eine (empirische) Verteilung *rechtssteil* oder *linksschief*, wenn der überwiegende Anteil der Daten rechtsseitig konzentriert ist.

- ▶ Typische Beispiele für linkssteile Verteilungen sind Einkommensverteilungen.

Linkssteil



Rechtssteil



Das arithmetische Mittel

Definition

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

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

berechnet.

- ▶ Für Häufigkeitsdaten mit Ausprägungen a_1, \dots, a_k und relative Häufigkeiten f_1, \dots, 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 intervallskalierte 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.

Der (Stichproben-)Median

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

Definition

Der *Median* x_{med} von $x_{(1)} \leq \dots \leq x_{(n)}$ 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 (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 $\leq x_{\text{med}}$.
- ▶ 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 gegenüber Ausreißern.
- ▶ Der Modus ist eine Ausprägung des Merkmals.

Streuung

Folgende Maßzahlen messen die Abweichung quantitativer Daten von ihrem Zentrum:

Mittlere absolute Abweichung $\frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$

Mittlere quadratische Abweichung $d^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$

Stichprobenvarianz $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{n}{n-1} d^2$

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

- ▶ $x_{0.5}$ ist der Median.
- ▶ $x_{0.25}$ und $x_{0.75}$ heißen *unteres* bzw. *oberes Quartil*.
- ▶ In R werden Quantile mit dem `quantile`-Befehl aufgerufen, z.B. `quantile(x, probs=c(0.25,0.75))`.

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 $(1 - p)$ -Fraktile bezeichnet.
- ▶ Ihr Aufruf in R erfolgt mittels “q + *Name der Verteilung*” (z.B. `qnorm`).

Fünf-Punkte-Zusammenfassung, summary

Definition

Die *Fünf-Punkte-Zusammenfassung* besteht aus

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

des Datensatzes.

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

Beispiel

- ▶ `summary(b1)`
- ▶ `summary(iris)`

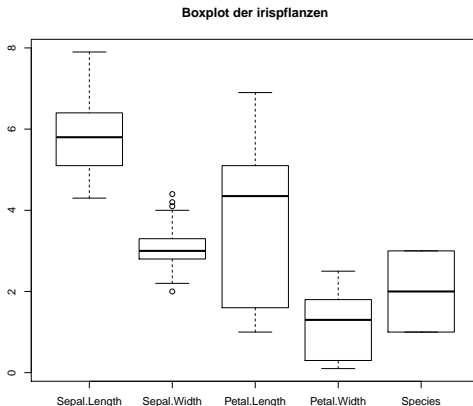
Boxplots

Die Fünf-Punkte-Zusammenfassung eines Datensatzes x_1, \dots, 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 $\min\{x_{0.75} + 1.5/IQR, x_{\max}\}$ bzw. bis $\max\{x_{0.25} - 1.5/IQR, x_{\min}\}$, wo die Linien durch senkrechte Linien begrenzt werden.
- ▶ Einzelnen Punkte für Werte jenseits der Whiskers (*Extremwerte*).

Beispiel: Boxplot

- ▶ `boxplot(bl, horizontal=TRUE)`
- ▶ `boxplot(iris, main="Boxplot der irispflanzen",
cex.axis=0.75)`



NQ-Plots: Idee

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

Frage: Ist es sinnvoll 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 $x_{(1)}, \dots, x_{(n)}$ die geordnete Urliste. Für $i = 1, \dots, n$ 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

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

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.